02 INFORMATION ABOUT PRINCIPAL INVESTIGATORS/PROJECT DIRECTORS(PI/PD) and co-PRINCIPAL INVESTIGATORS/co-PROJECT DIRECTORS

Submit only ONE copy of this form **for each PI/PD and co-PI/PD** identified on the proposal. The form(s) should be attached to the original proposal as specified in GPG Section II.C.a. Submission of this information is voluntary and is not a precondition of award. This information will not be disclosed to external peer reviewers. ***DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPROMISE THE CONFIDENTIALITY OF THE INFORMATION.***

PI/PD Name: Timothy J Menzies

Gender: ☒ Male ☐ Female

Ethnicity: (Choose one response) ☐ Hispanic or Latino ☒ Not Hispanic or Latino

Race: (Select one or more)
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Black or African American
- ☐ Native Hawaiian or Other Pacific Islander
- ☒ White

Disability Status: (Select one or more)
- ☐ Hearing Impairment
- ☐ Visual Impairment
- ☐ Mobility/Orthopedic Impairment
- ☐ Other
- ☒ None

Citizenship: (Choose one) ☐ U.S. Citizen ☒ Permanent Resident ☐ Other non-U.S. Citizen

Check here if you do not wish to provide any or all of the above information (excluding PI/PD name): ☐

REQUIRED: Check here if you are currently serving (or have previously served) as a PI, co-PI or PD on any federally funded project ☒

Ethnicity Definition:
Hispanic or Latino. A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.

Race Definitions:
American Indian or Alaska Native. A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
Black or African American. A person having origins in any of the black racial groups of Africa.
Native Hawaiian or Other Pacific Islander. A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.
White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

WHY THIS INFORMATION IS BEING REQUESTED:

The Federal Government has a continuing commitment to monitor the operation of its review and award processes to identify and address any inequities based on gender, race, ethnicity, or disability of its proposed PIs/PDs. To gather information needed for this important task, the proposer should submit a single copy of this form for each identified PI/PD with each proposal. Submission of the requested information is voluntary and will not affect the organization's eligibility for an award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information recieved from others. Any individual not wishing to submit some or all the information should check the box provided for this purpose. (The exceptions are the PI/PD name and the information about prior Federal support, the last question above.)

Collection of this information is authorized by the NSF Act of 1950, as amended, 42 U.S.C. 1861, et seq. Demographic data allows NSF to gauge whether our programs and other opportunities in science and technology are fairly reaching and benefiting everyone regardless of demographic category; to ensure that those in under-represented groups have the same knowledge of and access to programs and other research and educational oppurtunities; and to assess involvement of international investigators in work supported by NSF. The information may be disclosed to government contractors, experts, volunteers and researchers to complete assigned work; and to other government agencies in order to coordinate and assess programs. The information may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records", 63 Federal Register 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records", 63 Federal Register 268 (January 5, 1998).

02 INFORMATION ABOUT PRINCIPAL INVESTIGATORS/PROJECT DIRECTORS(PI/PD) and co-PRINCIPAL INVESTIGATORS/co-PROJECT DIRECTORS

Submit only ONE copy of this form **for each PI/PD and co-PI/PD** identified on the proposal. The form(s) should be attached to the original proposal as specified in GPG Section II.C.a. Submission of this information is voluntary and is not a precondition of award. This information will not be disclosed to external peer reviewers. ***DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPROMISE THE CONFIDENTIALITY OF THE INFORMATION.***

PI/PD Name: Bonnie W Morris

Gender: ☐ Male ☒ Female

Ethnicity: (Choose one response) ☐ Hispanic or Latino ☒ Not Hispanic or Latino

Race: (Select one or more)
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Black or African American
- ☐ Native Hawaiian or Other Pacific Islander
- ☐ White

Disability Status: (Select one or more)
- ☐ Hearing Impairment
- ☐ Visual Impairment
- ☐ Mobility/Orthopedic Impairment
- ☐ Other
- ☒ None

Citizenship: (Choose one) ☒ U.S. Citizen ☐ Permanent Resident ☐ Other non-U.S. Citizen

Check here if you do not wish to provide any or all of the above information (excluding PI/PD name): ☐

REQUIRED: Check here if you are currently serving (or have previously served) as a PI, co-PI or PD on any federally funded project ☐

Ethnicity Definition:
Hispanic or Latino. A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.

Race Definitions:
American Indian or Alaska Native. A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
Black or African American. A person having origins in any of the black racial groups of Africa.
Native Hawaiian or Other Pacific Islander. A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.
White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

WHY THIS INFORMATION IS BEING REQUESTED:

The Federal Government has a continuing commitment to monitor the operation of its review and award processes to identify and address any inequities based on gender, race, ethnicity, or disability of its proposed PIs/PDs. To gather information needed for this important task, the proposer should submit a single copy of this form for each identified PI/PD with each proposal. Submission of the requested information is voluntary and will not affect the organization's eligibility for an award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information recieved from others. Any individual not wishing to submit some or all the information should check the box provided for this purpose. (The exceptions are the PI/PD name and the information about prior Federal support, the last question above.)

Collection of this information is authorized by the NSF Act of 1950, as amended, 42 U.S.C. 1861, et seq. Demographic data allows NSF to gauge whether our programs and other opportunities in science and technology are fairly reaching and benefiting everyone regardless of demographic category; to ensure that those in under-represented groups have the same knowledge of and access to programs and other research and educational oppurtunities; and to assess involvement of international investigators in work supported by NSF. The information may be disclosed to government contractors, experts, volunteers and researchers to complete assigned work; and to other government agencies in order to coordinate and assess programs. The information may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records", 63 Federal Register 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records", 63 Federal Register 268 (January 5, 1998).

02 INFORMATION ABOUT PRINCIPAL INVESTIGATORS/PROJECT DIRECTORS(PI/PD) and co-PRINCIPAL INVESTIGATORS/co-PROJECT DIRECTORS

Submit only ONE copy of this form **for each PI/PD and co-PI/PD** identified on the proposal. The form(s) should be attached to the original proposal as specified in GPG Section II.C.a. Submission of this information is voluntary and is not a precondition of award. This information will not be disclosed to external peer reviewers. ***DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPROMISE THE CONFIDENTIALITY OF THE INFORMATION.***

PI/PD Name: Cecil Pollard

Gender: ☒ Male ☐ Female

Ethnicity: (Choose one response) ☐ Hispanic or Latino ☒ Not Hispanic or Latino

Race: (Select one or more)
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Black or African American
- ☐ Native Hawaiian or Other Pacific Islander
- ☒ White

Disability Status: (Select one or more)
- ☐ Hearing Impairment
- ☐ Visual Impairment
- ☐ Mobility/Orthopedic Impairment
- ☐ Other
- ☐ None

Citizenship: (Choose one) ☒ U.S. Citizen ☐ Permanent Resident ☐ Other non-U.S. Citizen

Check here if you do not wish to provide any or all of the above information (excluding PI/PD name): ☒

REQUIRED: Check here if you are currently serving (or have previously served) as a PI, co-PI or PD on any federally funded project ☐

Ethnicity Definition:
Hispanic or Latino. A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.

Race Definitions:
American Indian or Alaska Native. A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
Black or African American. A person having origins in any of the black racial groups of Africa.
Native Hawaiian or Other Pacific Islander. A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.
White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

WHY THIS INFORMATION IS BEING REQUESTED:

The Federal Government has a continuing commitment to monitor the operation of its review and award processes to identify and address any inequities based on gender, race, ethnicity, or disability of its proposed PIs/PDs. To gather information needed for this important task, the proposer should submit a single copy of this form for each identified PI/PD with each proposal. Submission of the requested information is voluntary and will not affect the organization's eligibility for an award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information recieved from others. Any individual not wishing to submit some or all the information should check the box provided for this purpose. (The exceptions are the PI/PD name and the information about prior Federal support, the last question above.)

Collection of this information is authorized by the NSF Act of 1950, as amended, 42 U.S.C. 1861, et seq. Demographic data allows NSF to gauge whether our programs and other opportunities in science and technology are fairly reaching and benefiting everyone regardless of demographic category; to ensure that those in under-represented groups have the same knowledge of and access to programs and other research and educational oppurtunities; and to assess involvement of international investigators in work supported by NSF. The information may be disclosed to government contractors, experts, volunteers and researchers to complete assigned work; and to other government agencies in order to coordinate and assess programs. The information may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records", 63 Federal Register 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records", 63 Federal Register 268 (January 5, 1998).

02 INFORMATION ABOUT PRINCIPAL INVESTIGATORS/PROJECT DIRECTORS(PI/PD) and co-PRINCIPAL INVESTIGATORS/co-PROJECT DIRECTORS

Submit only ONE copy of this form **for each PI/PD and co-PI/PD** identified on the proposal. The form(s) should be attached to the original proposal as specified in GPG Section II.C.a. Submission of this information is voluntary and is not a precondition of award. This information will not be disclosed to external peer reviewers. ***DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPROMISE THE CONFIDENTIALITY OF THE INFORMATION.***

PI/PD Name: Cynthia Tanner

Gender: ☐ Male ☒ Female

Ethnicity: (Choose one response) ☐ Hispanic or Latino ☒ Not Hispanic or Latino

Race: (Select one or more)
☐ American Indian or Alaska Native
☐ Asian
☐ Black or African American
☐ Native Hawaiian or Other Pacific Islander
☒ White

Disability Status: (Select one or more)
☐ Hearing Impairment
☐ Visual Impairment
☐ Mobility/Orthopedic Impairment
☐ Other
☒ None

Citizenship: (Choose one) ☒ U.S. Citizen ☐ Permanent Resident ☐ Other non-U.S. Citizen

Check here if you do not wish to provide any or all of the above information (excluding PI/PD name): ☐

REQUIRED: Check here if you are currently serving (or have previously served) as a PI, co-PI or PD on any federally funded project ☐

Ethnicity Definition:
Hispanic or Latino. A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.
Race Definitions:
American Indian or Alaska Native. A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
Black or African American. A person having origins in any of the black racial groups of Africa.
Native Hawaiian or Other Pacific Islander. A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.
White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

WHY THIS INFORMATION IS BEING REQUESTED:

The Federal Government has a continuing commitment to monitor the operation of its review and award processes to identify and address any inequities based on gender, race, ethnicity, or disability of its proposed PIs/PDs. To gather information needed for this important task, the proposer should submit a single copy of this form for each identified PI/PD with each proposal. Submission of the requested information is voluntary and will not affect the organization's eligibility for an award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information recieved from others. Any individual not wishing to submit some or all the information should check the box provided for this purpose. (The exceptions are the PI/PD name and the information about prior Federal support, the last question above.)

Collection of this information is authorized by the NSF Act of 1950, as amended, 42 U.S.C. 1861, et seq. Demographic data allows NSF to gauge whether our programs and other opportunities in science and technology are fairly reaching and benefiting everyone regardless of demographic category; to ensure that those in under-represented groups have the same knowledge of and access to programs and other research and educational oppurtunities; and to assess involvement of international investigators in work supported by NSF. The information may be disclosed to government contractors, experts, volunteers and researchers to complete assigned work; and to other government agencies in order to coordinate and assess programs. The information may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records", 63 Federal Register 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records", 63 Federal Register 268 (January 5, 1998).

02 INFORMATION ABOUT PRINCIPAL INVESTIGATORS/PROJECT DIRECTORS(PI/PD) and co-PRINCIPAL INVESTIGATORS/co-PROJECT DIRECTORS

Submit only ONE copy of this form **for each PI/PD and co-PI/PD** identified on the proposal. The form(s) should be attached to the original proposal as specified in GPG Section II.C.a. Submission of this information is voluntary and is not a precondition of award. This information will not be disclosed to external peer reviewers. ***DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPROMISE THE CONFIDENTIALITY OF THE INFORMATION.***

PI/PD Name: Forrest Shull

Gender: ☐ Male ☐ Female

Ethnicity: (Choose one response) ☐ Hispanic or Latino ☐ Not Hispanic or Latino

Race: (Select one or more)
- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Black or African American
- ☐ Native Hawaiian or Other Pacific Islander
- ☐ White

Disability Status: (Select one or more)
- ☐ Hearing Impairment
- ☐ Visual Impairment
- ☐ Mobility/Orthopedic Impairment
- ☐ Other
- ☐ None

Citizenship: (Choose one) ☒ U.S. Citizen ☐ Permanent Resident ☐ Other non-U.S. Citizen

Check here if you do not wish to provide any or all of the above information (excluding PI/PD name): ☒

REQUIRED: Check here if you are currently serving (or have previously served) as a PI, co-PI or PD on any federally funded project ☒

Ethnicity Definition:
Hispanic or Latino. A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.
Race Definitions:
American Indian or Alaska Native. A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
Black or African American. A person having origins in any of the black racial groups of Africa.
Native Hawaiian or Other Pacific Islander. A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.
White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

WHY THIS INFORMATION IS BEING REQUESTED:

The Federal Government has a continuing commitment to monitor the operation of its review and award processes to identify and address any inequities based on gender, race, ethnicity, or disability of its proposed PIs/PDs. To gather information needed for this important task, the proposer should submit a single copy of this form for each identified PI/PD with each proposal. Submission of the requested information is voluntary and will not affect the organization's eligibility for an award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information recieved from others. Any individual not wishing to submit some or all the information should check the box provided for this purpose. (The exceptions are the PI/PD name and the information about prior Federal support, the last question above.)

Collection of this information is authorized by the NSF Act of 1950, as amended, 42 U.S.C. 1861, et seq. Demographic data allows NSF to gauge whether our programs and other opportunities in science and technology are fairly reaching and benefiting everyone regardless of demographic category; to ensure that those in under-represented groups have the same knowledge of and access to programs and other research and educational oppurtunities; and to assess involvement of international investigators in work supported by NSF. The information may be disclosed to government contractors, experts, volunteers and researchers to complete assigned work; and to other government agencies in order to coordinate and assess programs. The information may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records", 63 Federal Register 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records", 63 Federal Register 268 (January 5, 1998).

List of Suggested Reviewers or Reviewers Not To Include (optional)

SUGGESTED REVIEWERS:
Not Listed

REVIEWERS NOT TO INCLUDE:
Not Listed

List of Suggested Reviewers or Reviewers Not To Include (optional)

SUGGESTED REVIEWERS:
Not Listed

REVIEWERS NOT TO INCLUDE:
Not Listed

COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE/if not in response to a program announcement/solicitation enter NSF 10-1

NSF 10-575 **11/28/10**

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)

CNS - TRUSTWORTHY COMPUTING

FOR NSF USE ONLY

NSF PROPOSAL NUMBER

1111012

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
11/23/2010	2	05050000 CNS	7795	191510239	11/23/2010 4:07pm S

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)

550665758

SHOW PREVIOUS AWARD NO. IF THIS IS
☐ A RENEWAL
☐ AN ACCOMPLISHMENT-BASED RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE

West Virginia University Research Corporation

AWARDEE ORGANIZATION CODE (IF KNOWN)

0038273001

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE

West Virginia University Research Corporation
P.O. Box 6845
Morgantown, WV. 265066845

NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE

West Virginia University

PERFORMING ORGANIZATION CODE (IF KNOWN)

0038273000

ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE

West Virginia University
Morgantown, WV 26506

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions)
☐ SMALL BUSINESS ☐ MINORITY BUSINESS
☐ FOR-PROFIT ORGANIZATION ☐ WOMAN-OWNED BUSINESS

☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **NSF 10-575 TC:Large: Collaborative Research: The Price of Privacy**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ 1,249,111	48 months	07/01/11	

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

☐ BEGINNING INVESTIGATOR (GPG I.G.2)
☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C.1.e)
☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.D, II.C.1.d)
☐ HISTORIC PLACES (GPG II.C.2.j)
☐ EAGER* (GPG II.D.2) ☐ RAPID** (GPG II.D.1)
☐ VERTEBRATE ANIMALS (GPG II.D.6) IACUC App. Date _____
PHS Animal Welfare Assurance Number _____

☐ HUMAN SUBJECTS (GPG II.D.7) Human Subjects Assurance Number _____
Exemption Subsection _____ or IRB App. Date _____
☐ INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)
_____
☐ HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.G.1)

PI/PD DEPARTMENT
Lane Dept of CSEE

PI/PD FAX NUMBER

PI/PD POSTAL ADDRESS
P.O Box 6109
Morgantown, WV 26506
United States

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address
PI/PD NAME **Timothy J Menzies**	**PhD**	**1995**	**503-901-1543**	**tim@menzies.us**
CO-PI/PD **Bonnie W Morris**	**PhD**	**1992**	**304-293-3998**	**bonnie.morris@mail.wvu.edu**
CO-PI/PD **Cecil Pollard**	**MA**	**1982**	**304-293-1080**	**cpollard@hsc.wvu.edu**
CO-PI/PD **Cynthia Tanner**	**MS**	**1979**	**304-293-9138**	**Cindy.Tanner@mail.wvu.edu**
CO-PI/PD				

Electronic Signature

CERTIFICATION PAGE

Certification for Authorized Organizational Representative or Individual Applicant:

By signing and submitting this proposal, the Authorized Organizational Representative or Individual Applicant is: (1) certifying that statements made herein are true and complete to the best of his/her knowledge; and (2) agreeing to accept the obligation to comply with NSF award terms and conditions if an award is made as a result of this application. Further, the applicant is hereby providing certifications regarding debarment and suspension, drug-free workplace, lobbying activities (see below), responsible conduct of research, nondiscrimination, and flood hazard insurance (when applicable) as set forth in the NSF Proposal & Award Policies & Procedures Guide, Part I: the Grant Proposal Guide (GPG) (NSF 10-1). Willful provision of false information in this application and its supporting documents or in reports required under an ensuing award is a criminal offense (U. S. Code, Title 18, Section 1001).

Conflict of Interest Certification

In addition, if the applicant institution employs more than fifty persons, by electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative of the applicant institution is certifying that the institution has implemented a written and enforced conflict of interest policy that is consistent with the provisions of the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.A; that to the best of his/her knowledge, all financial disclosures required by that conflict of interest policy have been made; and that all identified conflicts of interest will have been satisfactorily managed, reduced or eliminated prior to the institution's expenditure of any funds under the award, in accordance with the institution's conflict of interest policy. Conflicts which cannot be satisfactorily managed, reduced or eliminated must be disclosed to NSF.

Drug Free Work Place Certification

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Drug Free Work Place Certification contained in Exhibit II-3 of the Grant Proposal Guide.

Debarment and Suspension Certification

(If answer "yes", please provide explanation.)

Is the organization or its principals presently debarred, suspended, proposed for debarment, declared ineligible, or voluntarily excluded from covered transactions by any Federal department or agency? Yes ☐ No ☒

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Debarment and Suspension Certification contained in Exhibit II-4 of the Grant Proposal Guide.

Certification Regarding Lobbying

The following certification is required for an award of a Federal contract, grant, or cooperative agreement exceeding $100,000 and for an award of a Federal loan or a commitment providing for the United States to insure or guarantee a loan exceeding $150,000.

Certification for Contracts, Grants, Loans and Cooperative Agreements

The undersigned certifies, to the best of his or her knowledge and belief, that:
(1) No federal appropriated funds have been paid or will be paid, by or on behalf of the undersigned, to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with the awarding of any federal contract, the making of any Federal grant, the making of any Federal loan, the entering into of any cooperative agreement, and the extension, continuation, renewal, amendment, or modification of any Federal contract, grant, loan, or cooperative agreement.
(2) If any funds other than Federal appropriated funds have been paid or will be paid to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with this Federal contract, grant, loan, or cooperative agreement, the undersigned shall complete and submit Standard Form-LLL, "Disclosure of Lobbying Activities," in accordance with its instructions.
(3) The undersigned shall require that the language of this certification be included in the award documents for all subawards at all tiers including subcontracts, subgrants, and contracts under grants, loans, and cooperative agreements and that all subrecipients shall certify and disclose accordingly.

This certification is a material representation of fact upon which reliance was placed when this transaction was made or entered into. Submission of this certification is a prerequisite for making or entering into this transaction imposed by section 1352, Title 31, U.S. Code. Any person who fails to file the required certification shall be subject to a civil penalty of not less than $10,000 and not more than $100,000 for each such failure.

Certification Regarding Nondiscrimination

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative is providing the Certification Regarding Nondiscrimination contained in Exhibit II-6 of the Grant Proposal Guide.

Certification Regarding Flood Hazard Insurance

Two sections of the National Flood Insurance Act of 1968 (42 USC §4012a and §4106) bar Federal agencies from giving financial assistance for acquisition or construction purposes in any area identified by the Federal Emergency Management Agency (FEMA) as having special flood hazards unless the:
(1) community in which that area is located participates in the national flood insurance program; and
(2) building (and any related equipment) is covered by adequate flood insurance.

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant located in FEMA-designated special flood hazard areas is certifying that adequate flood insurance has been or will be obtained in the following situations:
(1) for NSF grants for the construction of a building or facility, regardless of the dollar amount of the grant; and
(2) for other NSF Grants when more than $25,000 has been budgeted in the proposal for repair, alteration or improvement (construction) of a building or facility.

Certification Regarding Responsible Conduct of Research (RCR)
(This certification is not applicable to proposals for conferences, symposia, and workshops.)

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative of the applicant institution is certifying that, in accordance with the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.B., the institution has a plan in place to provide appropriate training and oversight in the responsible and ethical conduct of research to undergraduates, graduate students and postdoctoral researchers who will be supported by NSF to conduct research. The undersigned shall require that the language of this certification be included in any award documents for all subawards at all tiers.

AUTHORIZED ORGANIZATIONAL REPRESENTATIVE	SIGNATURE	DATE
NAME **Alan B Martin**	**Electronic Signature**	**Nov 23 2010 1:47PM**

TELEPHONE NUMBER	ELECTRONIC MAIL ADDRESS	FAX NUMBER
304-293-3998	**Alan.Martin@mail.wvu.edu**	**304-293-7435**

* EAGER - EArly-concept Grants for Exploratory Research
** RAPID - Grants for Rapid Response Research

COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE/if not in response to a program announcement/solicitation enter NSF 10-1

NSF 10-575 **11/28/10**

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)

CNS - TRUSTWORTHY COMPUTING

FOR NSF USE ONLY

NSF PROPOSAL NUMBER

1110986

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
11/22/2010	**2**	**05050000 CNS**	**7795**		**11/23/2010 4:07pm S**

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)

383203030

SHOW PREVIOUS AWARD NO. IF THIS IS
☐ A RENEWAL
☐ AN ACCOMPLISHMENT-BASED RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE

Fraunhofer Center for Experimental Software Engineering

AWARDEE ORGANIZATION CODE (IF KNOWN)

5300016026

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE

Fraunhofer Center for Experimental Software Engineering
5825 University Research Court
College Park, MD. 207403823

NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE

PERFORMING ORGANIZATION CODE (IF KNOWN)

ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions)
☐ SMALL BUSINESS ☐ MINORITY BUSINESS
☐ FOR-PROFIT ORGANIZATION ☐ WOMAN-OWNED BUSINESS

☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **NSF 10-575 TC:Large: Collaborative Research :The Price of Privacy**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ **808,000**	**48** months	**07/01/11**	

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

☐ BEGINNING INVESTIGATOR (GPG I.G.2)
☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C.1.e)
☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.D, II.C.1.d)
☐ HISTORIC PLACES (GPG II.C.2.j)
☐ EAGER* (GPG II.D.2) ☐ RAPID** (GPG II.D.1)
☐ VERTEBRATE ANIMALS (GPG II.D.6) IACUC App. Date _____
PHS Animal Welfare Assurance Number _____

☐ HUMAN SUBJECTS (GPG II.D.7) Human Subjects Assurance Number _____
Exemption Subsection _____ or IRB App. Date _____
☐ INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)
_____
☐ HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.G.1)

PI/PD DEPARTMENT
Measurement and Knowledge Management

PI/PD FAX NUMBER
240-487-2960

PI/PD POSTAL ADDRESS
5825 University Research Court
Suite 1300
College Park, MD 20740
United States

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address
PI/PD NAME **Forrest Shull**	**PhD**	**1998**	**240-487-2904**	**fshull@fc-md.umd.edu**
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				

Electronic Signature

CERTIFICATION PAGE

Certification for Authorized Organizational Representative or Individual Applicant:

By signing and submitting this proposal, the Authorized Organizational Representative or Individual Applicant is: (1) certifying that statements made herein are true and complete to the best of his/her knowledge; and (2) agreeing to accept the obligation to comply with NSF award terms and conditions if an award is made as a result of this application. Further, the applicant is hereby providing certifications regarding debarment and suspension, drug-free workplace, lobbying activities (see below), responsible conduct of research, nondiscrimination, and flood hazard insurance (when applicable) as set forth in the NSF Proposal & Award Policies & Procedures Guide, Part I: the Grant Proposal Guide (GPG) (NSF 10-1). Willful provision of false information in this application and its supporting documents or in reports required under an ensuing award is a criminal offense (U. S. Code, Title 18, Section 1001).

Conflict of Interest Certification

In addition, if the applicant institution employs more than fifty persons, by electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative of the applicant institution is certifying that the institution has implemented a written and enforced conflict of interest policy that is consistent with the provisions of the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.A; that to the best of his/her knowledge, all financial disclosures required by that conflict of interest policy have been made; and that all identified conflicts of interest will have been satisfactorily managed, reduced or eliminated prior to the institution's expenditure of any funds under the award, in accordance with the institution's conflict of interest policy. Conflicts which cannot be satisfactorily managed, reduced or eliminated must be disclosed to NSF.

Drug Free Work Place Certification

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Drug Free Work Place Certification contained in Exhibit II-3 of the Grant Proposal Guide.

Debarment and Suspension Certification

(If answer "yes", please provide explanation.)

Is the organization or its principals presently debarred, suspended, proposed for debarment, declared ineligible, or voluntarily excluded from covered transactions by any Federal department or agency? Yes ☐ No ☒

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Debarment and Suspension Certification contained in Exhibit II-4 of the Grant Proposal Guide.

Certification Regarding Lobbying

The following certification is required for an award of a Federal contract, grant, or cooperative agreement exceeding $100,000 and for an award of a Federal loan or a commitment providing for the United States to insure or guarantee a loan exceeding $150,000.

Certification for Contracts, Grants, Loans and Cooperative Agreements

The undersigned certifies, to the best of his or her knowledge and belief, that:
(1) No federal appropriated funds have been paid or will be paid, by or on behalf of the undersigned, to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with the awarding of any federal contract, the making of any Federal grant, the making of any Federal loan, the entering into of any cooperative agreement, and the extension, continuation, renewal, amendment, or modification of any Federal contract, grant, loan, or cooperative agreement.
(2) If any funds other than Federal appropriated funds have been paid or will be paid to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with this Federal contract, grant, loan, or cooperative agreement, the undersigned shall complete and submit Standard Form-LLL, "Disclosure of Lobbying Activities," in accordance with its instructions.
(3) The undersigned shall require that the language of this certification be included in the award documents for all subawards at all tiers including subcontracts, subgrants, and contracts under grants, loans, and cooperative agreements and that all subrecipients shall certify and disclose accordingly.

This certification is a material representation of fact upon which reliance was placed when this transaction was made or entered into. Submission of this certification is a prerequisite for making or entering into this transaction imposed by section 1352, Title 31, U.S. Code. Any person who fails to file the required certification shall be subject to a civil penalty of not less than $10,000 and not more than $100,000 for each such failure.

Certification Regarding Nondiscrimination

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative is providing the Certification Regarding Nondiscrimination contained in Exhibit II-6 of the Grant Proposal Guide.

Certification Regarding Flood Hazard Insurance

Two sections of the National Flood Insurance Act of 1968 (42 USC §4012a and §4106) bar Federal agencies from giving financial assistance for acquisition or construction purposes in any area identified by the Federal Emergency Management Agency (FEMA) as having special flood hazards unless the:
(1) community in which that area is located participates in the national flood insurance program; and
(2) building (and any related equipment) is covered by adequate flood insurance.

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant located in FEMA-designated special flood hazard areas is certifying that adequate flood insurance has been or will be obtained in the following situations:
(1) for NSF grants for the construction of a building or facility, regardless of the dollar amount of the grant; and
(2) for other NSF Grants when more than $25,000 has been budgeted in the proposal for repair, alteration or improvement (construction) of a building or facility.

Certification Regarding Responsible Conduct of Research (RCR)
(This certification is not applicable to proposals for conferences, symposia, and workshops.)

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative of the applicant institution is certifying that, in accordance with the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.B., the institution has a plan in place to provide appropriate training and oversight in the responsible and ethical conduct of research to undergraduates, graduate students and postdoctoral researchers who will be supported by NSF to conduct research.
The undersigned shall require that the language of this certification be included in any award documents for all subawards at all tiers.

AUTHORIZED ORGANIZATIONAL REPRESENTATIVE	SIGNATURE	DATE
NAME **Forrest Shull**	**Electronic Signature**	**Nov 22 2010 6:10PM**

TELEPHONE NUMBER	ELECTRONIC MAIL ADDRESS	FAX NUMBER
240-487-2904	**fshull@fc-md.umd.edu**	**240-487-2960**

* EAGER - EArly-concept Grants for Exploratory Research
** RAPID - Grants for Rapid Response Research

NSF 10-575 TC:Large: Collaborative Research :The Price of Privacy

Forrest Shull (Fraunhofer, USA); Tim Menzies, Bonnie Morris, Cindy Tanner , Cecil Pollard (WVU)

A strange feature of the 21st century is that..

- while there is *much we can learn* from each other,
- there is *little we dare to share*.

The increasing use of electronic records goes hand in hand with an increasing awareness of the security and privacy concerns about that data. So how can we enable more effective data sharing to provide a quantitative and qualitative better data set to improve learning for all?

Prior research work on privacy and data mining has to a large degree focused on anonymizing data so that it can be safely shared and analyzed. However, data sharing is still inhibited by (a) privacy statutes; (b) regulations that limit the distribution of non-public personal information; and (c) by organizational fears of disclosing confidential, sensitive, or proprietary information. Rather than cajole organizations to expose their data, we need to build trust by refining data mining techniques that can work in the real world, by allowing data to stay protected, behind firewalls, under full control of the owners, while at the same time building common knowledge, which is also beneficial for each single organization.

To do so, we are proposing research to quantify the **tradeoffs amongst two trustworthy computing properties**; i.e. improving privacy restrictions and decreasing data mining efficacy. We will apply our tools to several areas of important economic and social benefit: (1)Software cost estimation, (2)software inspection control, and (3)disease patterns in communities. These tasks are exemplars of a wide class of activities where groups engaged in similar activities cannot share data due to institutional or legislative or social considerations. In order to achieve these goals we will:

1. Design and implement trusted enclaves in multiple data domains as testbeds that allow experimentation with different privacy policies. Our test domains will be community medicine and software engineering.
2. Implement privacy-based (local) data mining algorithms that exploit such trusted enclaves to create knowledge without exposing private data.
3. Improve the state of the art in software cost estimation, quality inspection control and recognizing patterns in chronic disease management.

Intellectual Merit: While searching for general models, we always remember that locally learned lessons give different, and better, results than general conclusions. Hence, we need better ways to apply and compare the results of data mining from numerous (local) results. Such comparisons are impossible unless some degree of access is permitted and not blocked by security or privacy considerations.

Once we know to effectively learn from multiple data sources, then this will **transform research.** Currently, scientists and engineers mostly analyze the data that they have collected themselves. If we succeed in the work proposed here, much of this current approach will change, Science will become a world-wide crowd sourcing activity in which large communities quickly discover the nuances and insights within shared data sources.

Broader Impacts: We will make public our tools (under an open source license) and all the collected data that our participating industrial partners will allow us to distribute. While our case studies are from software developers and public hospitals, however, our algorithms could be applied to any group of organizations. If we could automate the data collection from multiple organizations, and check if patterns at one organization apply to another while maintaining the privacy and security restrictions of those organizations then we would offer a significant boost to any multi-organizational data sharing initiative.

Our work would be used to develop new course work for under-graduate and graduate software engineering subjects at UMCP (which has associations with Fraunhofer, USA) and WVU. We also aim at maturing existing course work, in particular the senior year project undertaken by most fourth year undergraduate computer science students

Finally, part of these funds will be used to train students from traditionally under-represented areas in computer science. The PIs have an exemplary record in graduating female masters and Ph.D.s, as well as other minority groups.

Keywords: Security. Privacy. Data Mining. Trust.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)	1	
Table of Contents	1	
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) **(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**	15	
References Cited	7	
Biographical Sketches (Not to exceed 2 pages each)	7	
Budget (Plus up to 3 pages of budget justification)	7	
Current and Pending Support	4	
Facilities, Equipment and Other Resources	2	
Special Information/Other Supplementary Docs/Mentoring Plan	1	
Appendix (List below.) **(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**		
Appendix Items:		

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)		
Table of Contents	1	
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) **(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**	0	
References Cited		
Biographical Sketches (Not to exceed 2 pages each)	3	
Budget (Plus up to 3 pages of budget justification)	6	
Current and Pending Support	2	
Facilities, Equipment and Other Resources	2	
Special Information/Other Supplementary Docs/Mentoring Plan	1	
Appendix (List below.) **(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**		
Appendix Items:		

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

2. PROJECT DESCRIPTION: A strange feature of the 21st century is that while there is *much we can learn* from each other, there is *little we dare to share*. The increasing use of data mining has led to an increasing awareness of the privacy concerns about data. For example, the Real-Time Outbreak Detection System (at U. Pitt.) seeks disease out-break patterns in data from healthcare providers. While that data is de-identified in accordance with HIPAA safe-harbor rules, privacy concerns make some participants hide information vital to tracking disease patterns; e.g., number of visits by ZIP code [Cli04].

We argue that is a mistake to try and cajole organizations to expose their data. Rather, we need to build trust by *refining data mining techniques* that can *work in the real world*, by allowing *data to stay protected, behind firewalls*, under *full control of the owners*, while at the same time *building common knowledge,* which is also *beneficial for each single organization*.

2.1a Research goals & anticipated results: We seek to understand *the trade-offs between two properties of trustworthy computing*; i.e. *increasing privacy constraints* and *decreasing data mining efficacy*. We will:

- In years 1 & 2, implement and test a new *distributed privacy-aware data miner* on real-world domains.
- In year 3, use the experience of years 1&2 to package and make a public release of this data miner.
- In year 4, we will stress-test our tools by applying them to large private patient databases found in community health databases across the USA and the various islands of the South Pacific.
- In all years, we will study *how changing privacy* rules affects what can be learned by the data miner.

The real-world domains we will explore represent areas of great important economic and social benefit: 1) Software development planning and prediction, and 2) disease patterns in communities. These tasks are exemplars of a wide class of activities where groups cannot share data due to institutional, legislative, and / or social considerations, yet where progress requires analysis across datasets.

1) Software has become a critical enabling technology for realization of functions central to our society. Yet software engineering lags in its ability to deliver sufficient *quality* on *schedule*. The best defect- and cost-prediction models require data sharing from many sites; rarely does an organization have enough data to calibrate their own models.
2) *Chronic disease management is complicated by conditions that are infrequent at one site, but occur at repeated sites*. By using data from many sites, we can (a) learn best practices; (b) check if particular sites are not following best practices; and (c) detect if current best practices are losing efficacy (due to changes in the disease patterns).

We expect that our work will result in specific improvements within these areas, and accelerate the broader adoption of our distributed data mining methods. For example, cancer registries, injury control programs, and medical surveillance programs such as national programs for tracking influenza outbreaks would all benefit from our research.

This is the right time for this research. In the near future, privacy regulations will be extensively restructured. Already, there is much debate on problems with the 1996 HIPA regulations [Ness07, Ness10]. Also, responses to electronic terrorism will require granting controlled government access to more data sets (see http://goo.gl/gvpr). In order to best redesign that legislation, we must understand the trade-offs between aspects of trustworthy computing such as privacy and data mining efficacy.

This is the right team to perform this research. Our team has had extensive experience in learning patterns from data (see [Men03a, Men06, Men07a, Men09a, Men10b, Mor09, Mor10, Pol09a, Pol09b, Bas99, Shu00a, Shu00b, Shu02a, Shu02b, Shu07]). Also, the PIs come from diverse backgrounds[1] and hence have the experience required to interpret and critique the patterns found by our data miners. Further, through our industrial and research contacts, we already have access to the medical and software engineering data required for this work[2]. For example,

- PI Pollard administrates the databases used for our studies on medical applications.
- PI Menzies runs an open source repository containing over a hundred data sets from software engineering projects- see http://promisedata.org/data.

---

1 *Pollard*: community health; *Shull*: info.sys; *Menzies* & *Tanner*: CS; *Morris*: business & economics.

2 Lest we understate the effort required for this proposal, we add the following. While we have immediate access the data needed by this study, this proposal still requires extensive domain engineering to determine the range of privacy restrictions for that data. See Section 2.5 for more details.

- PI Shull has access to extensive databases describing multiple facets of the development of large software systems in the aerospace domain.

2.1b Motivations and Differences to Existing Research: A repeated issue faced by data mining researchers is obtaining access to data. For example, since 2006, we have tried to obtain permission to apply PI Menzies' data mining methods to PI Shull's software inspection data. The reply from our business partners has always been the same:

- The data cannot leave the firewalls to travel to Menzies' lab at WVU;
- But if the data miners could work inside the local firewalls, and if the contributors of the data could *audit and censor* the results before they are distributed, then that would be permissible.

Formally, these business partners are requesting (i) a *distributed data mining solution* where (ii) the learned models from the data miners are in *some human-readable* (and hence, *human-auditable*) format. Clearly, one major issue with such an architecture is that if business users can censor the data mining results, will we lose data mining efficacy? This research proposal was designed to address this question.

If we can show it is possible to build privacy-aware distributed data miners, and that the conclusions of those data miners are not unduly damaged by privacy restrictions, then this would usher in a new age of trust where data owners understand they can retain control of their data while still coordinating and sharing with other groups.

Our proposal differs from other research in three ways:

1) ***No data exposure:*** Based on decades of work in data mining and data sharing [Men03a, Men06, Men07a, Men09a, Men10b, Mor09, Mor10, Bas99, Shu00a, Shu00b, Shu02a, Shu02b, Shu07], we assert that it is *very unlikely that organizations will expose their data.* However, some communities might form *trusted enclaves* [Mor09, Mor10] of data providers which, under strictly controlled conditions, will grant limited access to other enclave members. For example, in our approach, if a data miner is dispatched from a sender to a receiver then before that data miner returns to sender, the receiver imposes their privacy restrictions to expunge conclusions they wish to keep private.
2) ***Extensive verification:*** To be convincing, privacy methods need to be assessed on records from multiple sources. Hence, we test our approach using (a) numerous data sets from the medical and software engineering domain; and (b) the privacy restrictions associated with those data sets.
3) ***Determining the price of privacy*****:** We will selectively increase the privacy restrictions on our test data until our data miners stop working. In this way we will report the price of privacy; i.e., *how much can we protect our data* before losing the ability to *make useful conclusions*. Such an understanding can lead to clearer guidelines that organizations can use to make better determinations about costs and benefits of excluding data sharing, and help minimize inadvertent or over-cautious exclusions.

To see one difference of our work from other research, note that most of that other work proposes exposing data after anonymization by (e.g.) adding random noise [Bec80, Agr00, Vai04]; or generalizing specific data [Swe02a, Mas07]. That research suffers from limited verification. Fung et al. report that one data set (the 48,842 records of ADULT; see http://goo.gl/1XZT7) is the "de facto benchmark for testing anonymization algorithms" and list 13 papers that use it as the *only test case* for their algorithms [Fun08].

Such limited verification is extremely troubling. Brickell and Shmatikov report experiments where to achieve privacy using standard methods like k-anonymity and ℓ-diversity "requires almost complete destruction of the data-mining utility" [Bri08]. Their conclusions, based on just one data set (again, the ADULT data set) may not be externally valid. However, such disturbing results clearly demand more verification of privacy methods, ideally on more data sets from more sources (e.g. our test domains).

2.2 TEST DOMAINS: This research will impose increasingly onerous levels of privacy restrictions on a distributed data miner executing in two domains: (1) software developing organizations; (2) healthcare providers. At first glance, these domains seem different. However, in terms of data sharing, they are very similar. Both need to share data while at the same time, *retaining privacy*. Both can be modeled as a *nested trusted enclave*; i.e., a tree of information sources in which parents can only see their children; and where all nodes strive to retain privacy. The rest of this section describes these test domains.

2.2.a Sharing, Privacy and Learning in Software Engineering: Software engineering (SE) as a discipline lags in terms of its ability to provide engineering processes which deliver artifacts of predictable quality within the required time frame. One approach to this problem is empirical software engineering which strives to find the patterns of success and the patterns of failures seen in real-world data from software

projects. Advanced AI techniques such as data mining can find patterns in project data that predict for some quality variable such as number and/or location of bugs, development time, etc. Recent results from this work include the following empirical discoveries which have found patterns:

- That reduce the effort of inspecting code by 71% [Tos10];
- That reduce development effort, without incurring the penalty of greater defects [Men09b];
- To predict defect locations that are 1.5 to 3 times better than industrial practice [Men10b].

Sometimes, these empirical patterns apply only to a particular suite of software or an organization [Kit07]. However, recent results suggest that some of these empirical patterns might even apply to multiple software suites or organizations [Tur09, Koc10]. That is, if software organizations dared to share data, they could use each other's data to manage new kinds of projects that had not been attempted locally, but which had been tried elsewhere.

There is a problem, however, with such data sharing. Extracting project data from organizations is very difficult due to the business sensitivity associated with the data. Recently, open source code repositories have become a rich source of software *product data*. However, software *process data* (e.g., describing which development approaches lead to how much effort) is still very hard to obtain:

- Boehm (personnel communication) was able to collect very few project records relating to development effort despite 30 years of work with many companies in the USA and China.
- In our own work, after two years we were only able to add 7 records to our NASA-wide software cost metrics repository [Gre09b].

We diagnose the problem as a lack of trust. Software organizations cannot trust each other to share data, lest it gets used against them (say) during competitive bidding. The goal of our distributed mining is to refine methods and technologies that could enable a consortium of companies to share data, without any of them revealing critical information.

The research undertaken in this proposal would be initially performed on existing databases of software process and product data. Some of this data is already collected by the Fraunhofer Center over years of research for NASA and project support in the aerospace domain. For example, specific processes (like application of software inspections) are described in data across many projects and multiple NASA Centers in a database of thousands of records [Shu10]. Processes in-the-large are described in data collected for progress monitoring of large-scale aerospace software systems over multiple years. Other data comes from the aforementioned PROMISE open repository.

2.2.b Sharing, Privacy and Learning in Medicine: The ability to effectively share information, process it, and use the results in clinical decision support, while respecting patient privacy and ethical regulations in the entire process, can have significant impact on the quality of care offered [Xia09]. Mining of existing records related to chronic conditions such as diabetes, obesity, and cardiovascular disease is expected to lead to more effective treatment and prevention [Rak10]. Diseases develop and evolve over time so modeling the treatment sequentially is also expected to be beneficial [Rak10]. According to Epstein [Eps10], previously unexplained variations in patient care are really indicative of the lack of knowledge of best practices. Clearly, such regional variations motivate the wider sharing of information, as well as ways to actually use the information for improving care.

Although the necessity of collaborative sharing and learning has been recognized, there is little systematic knowledge sharing of clinical intervention outcomes [Xia09, Gre06]. Privacy concerns are a mitigating factor impeding the sharing of data between entities. Competent health care depends on accurate and complete information. The collection and use of information relies on trust between the provider and the recipient and the belief by the provider that his privacy will not be compromised. The potential costs when the provider feels a lack of privacy include: misdiagnosis or errors in care in the medical arena; missed opportunities and severe monetary penalties in the private sector.

In our discussions with hospital administrators, it has become clear that if we give those administrators the same *audit-and-censor* functionality requested by the software engineering managers, then that would increase the willingness for receiver sites to accept and execute someone else's data miner (since they could audit and censor any out-going results, before anyone else sees them).

Recent legislation has required health care providers to express their records in a uniform manner. A key outcome of Health Information Technology for Economic and Clinical Health (HITECH) initiative is the mandated move to the use of International Classification of Diseases v.10 (ICD-10) coding scheme (see

http://bit.ly/gk3BO). Such data uniformity greatly simplifies distributed data mining. For example, this proposal will use the WVU Chronic Disease EMS tool administrated by PI Pollard. His software is currently collecting data in 15 community health clinics in West Virginia. Installations of his tools are underway in two dozen health clinics across the country (see Figure 1). Other sites are planned in the USA Pacific Territories (Marshall Islands, Micronesia, and Palau have confirmed their participation while American Samoa and Guam are still evaluating the software). This data will be an excellent test bed within which we can test our proposed privacy-preserving distributed data mining scheme.

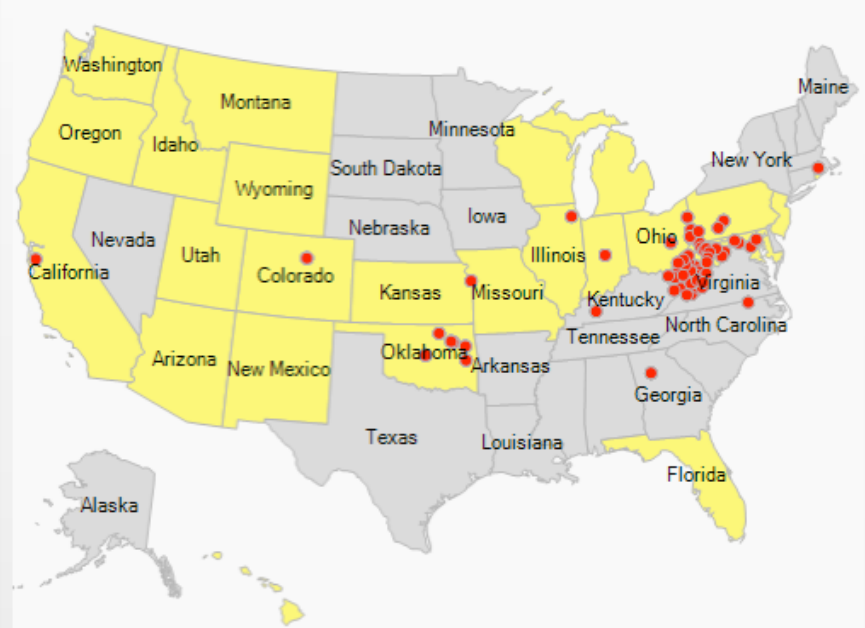

Figure 1: *Installation sites of PI Pollard's CDEMS software (chronic disease electronic management system).Red dots are currently collecting data; yellow states have expressed interest. Note that the sites in West Virginia are currently collecting data while the other sites will come on-line by 2012.*

Not shown on this map are the planned installations in the USA Pacific Territories, to come on-line by 2013.

For more on CDEMS, see http://wvuohsr.org.

2.2c Trusted Nested Enclaves: Our thesis is that hospitals and software engineering companies can both be modeled as a *nested trusted enclave*; i.e., a tree of information sources in which parents can only see their children; and where all nodes strive to retain privacy. Such enclaves form an acyclic network where data miners of some parent node (at level N) can only access the data of their child nodes (at level N+1). Each child node also contains data miners that reflect on grandchild nodes (recursively).

It is simple enough to demonstrate the nested structure of information sources in hospitals and software engineering organizations. Figure 2 details a possible recursive refinement into several levels for the software engineering and medical domains. In general, such refinement levels will help align the implementations of our enclave learners with any hierarchical structure in other organizations and domain, such as finances, intelligent services, or even universities.

Level	Software Engineering Domain	Medical Domain
1	Research network (e.g., International Software Engineering Research Network – ISERN: http://isern.iese.de)	State (e.g., WV, MD, DC)
2	Research group (e.g., FC-MD, SEI, IESE: http://www.iese.fraunhofer.de/index.jsp)	Hospital Groups
3	Application Context (e.g., aerospace, education, financial)	Single hospital / unit
4	Projects / programs (e.g., NASA, DoD)	Departments (e.g., ICU, pediatrics, radiology, …)
5	Sub-groups (e.g., according to location, code packages, application context, …)	Private practice / remote clinics / physician specialties

Figure 2: *Refinement example of research enclaves for the software engineering and medical domains.*

Formally, we say that enclaves are sets of *semi-honest adversaries* collaborating to try to learn from each other, without revealing too much about themselves (a semi-honest party follows the rules of the protocol using its correct input, but is free to later use what it sees during execution of the protocol to compromise security). In our work, we explore data sharing in trusted enclaves (e.g., see [Mor09, Mor10]). A trusted enclave runs agents that offer Continuous Compliance Assurance (CCA), where "compliance" is assessed with respect to some query describing the information that must not be revealed.

Based on our work with government institutions, we assume the following ontology for our compliance restrictions: They are combinations of "and, or, not" around attribute range queries; e.g., *Age<21 or organization="nsf".* The xpath query of Figure 3, for example, shows compliance restrictions in that ontology. A DOD secrecy rule is shown that demands that the weight of its consignment cannot be revealed (heavy DOD consignments may be nuclear materials with heavy shielding). Note that our ontology assumption is hardly controversial: We can find in the literature similar assumptions about privacy restrictions [Agr03]. However, as discussed below, it is an important assumption for when we come to selecting data mining technology.

One way to perform privacy preserving data mining in an enclave is via a data miner passed around a ring of enclave nodes. At the end of the ring, the data miner returns to where it started from to report its final conclusions. As it passes over nodes in the ring, local administrators can restrain the data mining by:

- *HIDING THE DATA:* The locals can run their own local version of the data miner to find data rows that lead to rules they want to censor. The locals can then decide, at their discretion, to hide those rows.
- *PRUNING THE RULES:* Before a data miner leaves a node, the locals can prune any parts of the learned rules that overlap with the compliance assurance (e.g., Figure 3).

Techniques for data pruning and rule pruning are discussed later in this proposal.

```
// U.S. Privacy Act of 1974
not(boolean(//RAAR/CrewDiscrepancies/Cre
wDiscrepancy/CrewPersons/CrewPerson
[CitizenshipCode = "US" and
(boolean(SID) or boolean(DateOfBirth) or
boolean(PlaceOfBirth)or
boolean(Height)or boolean(Weight) or
boolean(HairColor) or boolean(EyeColor)
or boolean(DistinguishingMarks) or
boolean(Sex) or boolean(VesselName)or
boolean(Status) or
boolean(CurrentLocation))]))

// Non-disclosure of DOD cargo
not(boolean(//ShipManifest/CargoManifest
/NonContainerItems/NonContainerItem[
(Consignee = "DoD" or Owner = "DoD" or
MarksAndNumbersMarkAndNumber/Mark = "DoD
Restricted") and boolean(Weight)]))
```

***Figure 3: Top:** A privacy restriction. **Bottom:** A secrecy requirement.*

2.3 ALTERNATIVES TO ENCLAVES: Before going any further, we need to explain our preference for enclave-based data mining. Hence, this section discusses privacy methods that *do not* use enclaves.

Data mining is the process of finding patterns in data. Traditionally, data miners assume access to a global database of all the information. This is worrisome if the data miner includes in their learned model some pattern that should remain confidential. Fung et al. [Fun10] distinguish two classes of research in this area: *privacy-preserving data publishing* (PPDP) and *enclave methods* that offer query control on multilevel secure databases (the difference being that, with enclaves, the data remains private and but the enclaves publish their queries[3] while with PPDP, the data is published and any inference on that data remains private).

Research in *statistical databases* implements PPDP by allowing statistical information (sum, count, average, maximum, minimum, nth-percentile, etc) to be accessible, without revealing sensitive individual information. PPDP techniques include *query restriction, data perturbation*, and *anonymization*. The *query restriction* methods includes restricting the size of query results [Den79], controlling overlap in successive queries [Dob79], keeping audit trails of answered queries (to check for possible compromises) [Chi82], avoiding data cells of small size [Cox80], and clustering entities into mutually exclusive atomic populations [Yu77]. *Data perturbation* techniques include swapping values between records [Den82], replacing the original database by a sample from the same distribution [Rei84], sampling the result of a

[3] Brickell and Shmatikov discuss an extension of enclaves, which we do not explore, where all nodes can access a global data dictionary, but not the attributes used in a query [Bri09]. Such an extension is not appropriate for our purposes since our users want to browse ranges to check that they do not violate compliance assurances.

query [Den82], and adding noise to the values in the database [War65] or the results of a query [Bec80]. *Anonymization* methods range from (a) the naïve approach (just remove all solo identifiers) to information from the data release; to (b) k-*anonymity* [Swe02a] which ensures that all queries return rows where any individual is indistinguishable from k-1 others; to more sophisticated techniques such as (c) *ℓ-diversity* [Mas07] or (d) the dozens of other methods reviewed by [Fun10].

The effectiveness of all these methods is an open issue. Repeated patterns in databases allows for the simple identification of individuals, even after sanitization [Swe02b]. Adding noise hides the details of individuals, but can confuse learners that (say) try to find the best place to discretize numeric data [Vai04]. Agrawal and Srikant [Agr00] offer one solution where Bayes' rules are used to reconstruct the original distributions using knowledge of the distributions used to add the noise- which then means that the reconstruction gives us information about the original data values, thus violating privacy [Zha07]. K-anonymity does not ensure privacy in the case of attackers using background knowledge on the groups returned by a query [Fun10]. Other PPDP methods have their limitations; e.g. ℓ-diversity implicitly assumes that each sensitive attribute takes values uniformly over its domain; i.e., that the frequencies of the various values of a confidential attribute are similar.

Worse still, Brickell and Shmatikov [Bri08] report that simplistic trivial sanitization provides equivalent utility and better privacy results than supposedly better methods such as k-anonymity or ℓ-diversity. As discussed above, their results are hardly conclusive (since they are based on a single data set). However, given all the problems discussed above with PPDP, we agree with Vaidya and Clifton [Vai04] that rather than struggle to secure a public data set, it might be better not to publish that data set in the first place. Hence, for the rest of this proposal, we discuss distributed data mining methods over trusted enclaves.

2.4 DISTRIBUTED DATA MINING OVER TRUSTED ENCLAVES: In our enclaves, parents can see their children but not their grandchildren. How can a parent gain insight into grandchild nodes? To address this problem of recursive insight, we distinguish between *top-down data mining* and *bottom-up rule fusion*.

TOP-DOWN DATA MINING: Parents collect statistics from their children, while maintaining the privacy of each child. One way to achieve this is secure multipart computation (SMC). As described by [Vai04], SMC is a conversation between *N* parties, none of which want to display their data to another. For a simple example of SMC, consider the "*no collusion*" case where a parent node in an enclave wants to sum some value *k* across its *N* children. A random number *R* is passed to any child at random. That child adds its local *k* value and passes the sum to another child, selected at random. When all children are visited, the parent receives back the sum, removes *R*, thus accessing the actual sum. Note that no child can infer what are the actual values of *k* in the other children since those values are masked by *R*. Two open issues with SMC are *collusion* and the *runtime cost*:

- *Collusion* between children can make SMC computation insecure. If children *z-1* and *z+1* compare the values of the running sum, they can compute the exact value for *k* in child *z*. Assuming more than three children, we can fix this problem by passing the sum around in random order amongst the children (so child *z+1* never knows who was child *z- 1*).
- As to *runtime cost*, Vaidya and Clifton [Vai04] report that SMC can be remarkably slow. In one test case using SMC on a multi-node network that required some joins between different tables, it took 29 hours to build a 408-node decision tree from 1,728 examples. Clearly, SMC is not recommended when nodes in a network must engage in high bandwidth communication to achieve some results.

BOTTOM-UP RULE FUSION: After a parent data mines their children, they broadcast the rules upward (after first deleting any rules that violate compliance). Grandparent nodes collect and fuse the rules from the parents. In this way, insights gained deep in the enclave can bubble upwards (but only if they are supported by multiple children). A rule fused from a few children needs to be tested across all children. If (a,b,c,d) are a classification rule's true negative rate, false negative rate, false positive rate, and true positive rates, then that rule's *precision*= d/(c+d); *accuracy*=(a+d)/(a+b+c+d); *recall*=d/(b+d); *false alarm rate* = c/(a+c) and *"f" measure*= 2*prec*pd/(pd+pf). If a parent collects these a,b,c,d values by passing a rule over its children using SMC, then the value of a fused rule can be computed without violating privacy.

From the above, we can deduce seven essential aspects of the design of our distributed learners:

1) *RULE FUSION:* There must be some way to combine rules from multiple sources.
2) *COLLUSION AVOIDANCE:* In order to avoid collusion during SMC, each enclave must include a transaction manager whose task is to pass a computation around a ring of nodes in a random order.

3) *SMC MANAGEMNENT:* When a data miner is initialized at the start of a query, its internal frequency counts are distorted by a random amount known only to the creator of the miner.
4) *BATCH OPTIMIZATION:* To avoid the computational overhead of unconstrained SMC, our data miners must not dispatch thousands of queries across an enclave. Rather, they should be one-pass learners that can make their conclusion after a single round-robin traversal of a set of nodes.
5) *AUDITABILITY:* In order for locals to recognize a compliance violation, the output of the learner must match the ontology for our compliance described above. Hence, whatever learner we use, it should produce user-readable high-level rules and not some arcane incomprehensible internal format.
6) *RULE PRUNING:* In order to let the locals censor rules that violate compliance, the learner's model should contain parts, any one of which can be deleted.
7) *DATA HIDING:* If the locals are to remove the data that lead to rules that violate compliance, it must be possible to track backwards from any rule to the data that generated it. This data should then be hidden from any incoming SMC requests.

Of the above criteria, items (3) and (5) lets us quickly rule out many data mining technologies. Those technologies include (a) discrete learners that find either association rules that report frequent patterns of attribute ranges that occur together [Agr93], or classification rules/decision trees that find frequent patterns between independent attributes and one dependent "class" attribute [Bre84, Qui92]; or (b) probabilistic methods such as fuzzy learners or Bayes classifiers [Wit05] or the EM clustering algorithm [Dem77] that represent different classes as distributions; or (c) methods that use probability distribution propagation over a directed graph like Bayes nets, neural nets [Hin92], distributed Kalman filters [Olf07], or non-parametric belief propagation (NBP) [Tse04]; or (d) instance-based learners that reason about examples nearest some test instance [Aha91]. In order to support *AUDITABILITY*, the ontology of the learnt model must match the compliance restriction. Hence, instance-based learners, which generate no model, are inappropriate for our work.

The discrete learners are most suited to our task since the and-or-not nature of discrete rule conditions are closest to the ontology of our compliance restrictions. However, many discrete learners are not suitable. Decision tree learners like C4.5 [Qui92] recursively divide the data set and call themselves on each subset of the data. This means repeated inspection of subsections of the data- which fails the *BATCH OPTIMIZATION* criteria. Recall from the above that using SMC took 29 hours to build a 408-node decision tree. Similar issues exist with association rule learners like APRIORI [Agr94]. This algorithm finds frequent item sets of increasing size and, for each such larger set, it conducts a repeated search of the data to count the occurrences of that set.

At first glance, the probabilistic models are inappropriate since our users require categorical rejection rather than some partial probabilistic pruning. This is unfortunate since probabilistic methods such as Bayes classifiers have some of the properties that we desire such as one-pass incremental learning. However, such classifiers build a single model of the data expressed in a format that is quite alien to the compliance ontology.

Recently, we have had success with rule generation from Bayes classifiers [Cla05, Gay10, Mil08]. TAR5 grows sets of interesting ranges (given discretized data, the attribute data falls into a finite number of ranges). The ranges are sorted on a stack according to how well they selected for a preferred class (see Figure 4[4]). TAR5 combines ranges at random to form rules (favoring ranges that appeared higher in the stack). TAR5 runs one stack per classification. Each stack finds a rule set that selects for that classification.

Round0	Round1
Top of stack	**Top of stack**
78 if sex=female	78 if sex=female
71 if class=1st	74 if class=1st and sex=female
68 if age= child	71 if class=2nd and sex=female
65 if class=2nd	72 if class=1st
	68 if age=child and class=1st
	68 if age=child
	68 if age=child and sex=female
	65 if class=2nd

Figure 4: *Rules found by TAR5. Left-hand-side numbers are accuracies predicting for survival from the Titanic. TAR5 sorting ranges from the last round, combining the better ones (selected stochastically, favoring those nearer top of stack), then scoring and sorting the new combinations into a new stack.*

The stack is initialized by passing all

[4] Since we cannot show real data from the confidential databases of our clients, we must resort to examples based on publically available information. Hence, the example of Figure 3 is based on survival data from the S.S. Titanic.

ranges through the scoring scheme of Figure 5. TAR5 repeatedly selects *R=2* items from the stack (favoring items with higher scores). Each selection is combined into a conjunction, scored, and sorted back into the stack. If it scores worse than existing items, it sorts lower on the stack (becoming less likely to be used in future). Otherwise, the new item moves up the stack, making it available for future selects. As TAR5 runs, items can grow in size as more useful conjunctions are discovered and combined (Figure 4 shows TAR5's rule growth using data on who survived the loss of the Titanic). TAR5 terminates when the score of the rule on top-of-stack stabilizes, at which point, TAR5 returns the top item as the best selector for some class.

When applied to the task of defect prediction for software modules, TAR5 out-performed standard learners such as Naive Bayes or decision-tree learners [Mil08]. It has been used at NASA to tune the settings of complex guidance, navigation and control flight systems [Gun08, Gun09]. In comparisons with state-of-the-art optimizers (a Quasi-Newton method that incremental updates a Hessian approximation), our method ran 40 times faster, and found better solutions [Gay10].

TAR5 was an experiment with one-pass learning. 80% of the algorithm's runtime arises from the repeated checking of the rules against examples of data. The algorithm removes that runtime by replacing that check with a Bayesian evaluation heuristic. After one pass of the data, the algorithm computes just enough information to allow for the fast ranking of different rules without needing to pass again through the examples. In practice, the algorithm ran two orders of magnitude faster than earlier versions, and used an order of magnitude less memory [Cla05]. Further, as described in Figure 5, this Bayesian heuristic has proven to be remarkably accurate.

TAR5 sorts all ranges into its stack as follows. A table of observations containing *N* observations has labels $L_1, L_2, ...$ appearing in $N_1, N_2, ...$ examples (so $N = \sum_i N_i$). For any label L_i, we say Re*st* is all the other labels (i.e., $L_i \notin$ Re*st*). If a range r appears at frequency $f(r)$ in all examples, and $f(r \mid L_1)$ in the rows labeled L_i, then r appears outside of the L_i rows at frequency $f(r \mid \mathrm{Re}st) = f(r) - f(r \mid L_i)$. The likelihood of r being in label L_i, or in Re*st*, is $\frac{f(r \mid L_i)}{N_i} * \frac{N_i}{N}$ or $\frac{f(r \mid \mathrm{Re}st)}{(N - N_i)} * \frac{(N - N_i)}{N}$ respectively. According to Bayes' theorem, the probability $P(L_i \mid \mathrm{Re}st)$ that r occurs in L_i is $like(r \mid L_i)$ normalized by the sum of the other likelihood that this range appears in other classes. To this normalizing fraction, we include a *support* term that favors ranges that occur with a high frequency (this stops over-fitting to the data [Cla05]). Note that likelihood increases as the frequency of a range increases; i.e., *like* can also serve as *support*. Combining all this, we compute a score that dictates how much to focus on r as a predictor for L_i, and not other labels:

$$score(r \mid L_i) = prob(r \mid L_i) * \sup port(r \mid L_i) = \frac{like(r \mid L_i)^2}{like(r \mid L_i) + like(r \mid \mathrm{Re}st)} \quad [1]$$

Equation 1 requires only a count of feature ranges r (which can be collected in linear time). It also computes a heuristic score for generated rules. To compute the likelihood of a conjunction $like(r_1 \wedge r_2 \wedge ... \mid L_i)$ we use $\frac{like(r \mid L_i)^2}{like(r \mid L_i) + like(r \mid \mathrm{Re}st)}$ where $like(r \mid L_i) = \left(\prod_i \frac{f(r \mid L_i)}{N_i} \right) * \frac{N_i}{N}$

This approximation is fast to calculate from frequency counts collected in linear time. Also, it is quite accurate at ranking rules. Clarke [Cla05] generated 10,000 randomly generated conjunctions. Each conjunction was ranked using (a) the fast approximation discussed here, or (b) a laborious search that over all the test data collecting actual observations selected by the conjunction. If score(a) < score(b) then that was an *agreement*. In 10,000 conjunctions generated from UCI data sets [Bla98], agreement was seen in 95% of comparisons. We hypothesize that the support term of Equation 1 sends the reasoning to well sampled regions where approximations make few mistakes.

Figure 5: *TAR5's Bayesian scoring method.*

2.5. RESEARCH TASKS: While a promising start, the current version of TAR5 is unsuitable for data mining in trusted nested enclaves. As described in this section, we need to extend TAR5 to address certain research challenges (described below). This extended system, called TAR6, is our proposed distributed data miner for privacy-aware learning over enclaves. In order to commission and test TAR6, we will perform the following tasks, in the following order.

2.5a. From multiple sites, collect real world data & their compliance requirements (Task 1). The introduction of this proposal lamented the poor state of the art in verification of privacy algorithms. Many papers assess their work on theoretical grounds or using a single data set. Clearly, this is not an ideal verification method. Real world data is notoriously quirky (what is true in one data set may be irrelevant in another [Men06, Men07d]. Privacy tool needs to be tested on dozens of real-world examples, ideally from multiple organizations.

These real world examples will be collected and explored during the first, second and fourth years of the project. In year one, we will explore SE data collected by PI Shull in proprietary databases of software metrics from industrial software development projects in the aerospace domain maintained at the Fraunhofer, and/or from the PROMISE repository of open SE data (http://promisedata.org). In year two, we will explore the CDEMS data.

Figure 6 shows a set of research questions and possible privacy restrictions that could be relevant to this SE data. Note that this figure is only preliminary. Extensive domain engineering would be required to ensure that this table realistically reflected the concerns of the managers of aerospace SE development projects. That domain engineering would refine/add/reject rows in this table.

Research questions and nodes	Enclave privacy restrictions
What is a typical effort distribution over development phases? How much effort does it take to correct a major / minor error? Nodes will be the different classes of a concrete NASA project. The enclave refinement for this case study is similar to levels 4+5 described in ***Figure 2***	Names of people involved in developing the class Owner of the class Center (i.e., name) developing the class Program or mission the class is/was developed for Used programming language • All restrictions focusing on the protection of specific people on the development team so they continue to share their data without having to fear that this will be held against them at any later point in time. Additional restrictions focusing on the protection of the integrity of a concrete Center or program
How are defects distributed over different project phases? What kind of defects are typically for each development phase? As nodes we will use the different NASA Centers. The enclave refinement for this case study is similar to levels 4+5 described in ***Figure 2***	Program or mission the class is/was developed for Center (i.e., name) reporting Size of development team (i.e., # of people) Period of code development Used programming language • Restrictions focusing on the protection of the integrity of a concrete Center or program. Using team size and/or development year could allow to indirectly identifying the program/mission.
How does reuse of code influence the defect density of a program? As nodes we will use the different (aerospace) projects. The enclave refinement for this case study is similar to levels 3-5 described in ***Figure 2***	Size of developed code (e.g., LoC) Size of project team (i.e., # of people) Program(s) (i.e., names) from which code was reused Current project phase (e.g., requirements, testing, finished/in use) Used programming language • All restrictions focusing on the protection of the integrity of a concrete program. Using the different information could indirectly allow 3rd parties to identify the original source of the data.

Figure 6: *Possible research questions vs. privacy restrictions for the software engineering domain in context of aerospace applications*

Figure 7 lists the research questions and privacy restrictions that might apply to our medical data. As with Figure 6, domain engineering is required to improve this table.

Before we can process Figures 6 and 7 (and their associated data), we need to build the infrastructure needed for our rule learners. This requirement takes us to our next task (task 2).

research questions and nodes	Enclave privacy restrictions
How do treatments for a particular diagnosis vary across locations? What is the typical treatment for a particular diagnosis? Nodes will be community health clinics. The enclave refinement for this case study is similar to levels 2-5 described in ***Figure 2***	Provider Information (e.g. name, hospital name) • Restrictions focusing on the protection of Healthcare providers, including both the specific clinic as well as the physicians and other healthcare workers. Patient Information (e.g. name, data of birth, zip code) • Restrictions on all patient identifiers in compliance with HIPAA including the general geographic location of the patient's home.
How does the presence of chronic conditions affect the choice of treatment? What are the characteristics of patients for whom standard treatment did not work? (chronic conditions such as diabetes, obesity, cardiovascular disease) Nodes will be patient sub categories based on chronic conditions. The enclave refinement for this case study is similar to levels 3-5 described in ***Figure 2***	Provider Information (e.g. name, hospital name) Patient Information (e.g. name, zip code, data of birth) • Restrictions focusing on the protection of patient data, physician identity, hospital unit providing the service and patient identifiers.
How is the treatment of a patient referred for specialized care affected by the source and timing of referrals? What impact does the stage of progression of the disease As nodes we would look at referring physician groups, and specialists to whom patients were referred. The enclave refinement for this case study is similar to levels 4-5 described in ***Figure 2***	Provider Information (e.g. name, hospital name) Patient Information (e.g. name, zip code, data of birth) • Restrictions on the identity of the referring the patient, all patient identifying information, including the geographic location of the referred patient.
Figure 7: *Possible research questions vs. privacy restrictions for the medical domain in context of the Chronic Health Disease Database*	

2.5b. Add rule fusion to our learner (Task2): Fusion means combining ranges from different rules from different sources. In the context of TAR6, a grandparent node in the enclave can fuse rules as follows. If P parents offer N rules (learned from their children) to their grandparent, then there are N*P rules to fuse. Each rule, in isolation, will have some score (from Figure 5). The grandparent could sort the rules on that score and try stochastic combinations of those N*P rules, favoring those with higher scores. This is almost the same \ rule growth procedure described above for TAR5, with one important variation. When a combined rule is scored, the grandparent sends the new rule back to each parent and asks them to score it from their children. Any combined rule that scores well will float to the top of the grandparent stack. Similarly, rules that worked well at one parent, but not on all, will float to the bottom.

Note that this fusion strategy is analogous to gossip networks [Dim06] where nodes compute some joint value using randomly sampled neighbors. However, where as gossip networks assume continuous distributions, TAR6 assumes that the knowledge to be combined is discrete rules.

Rule fusion enables combining rules learned in multiple parents. This, in turn, requires that some rules exist in the first place. In order to use SMC to sample child data to build parent rules, we must:

2.5c. Add collusion avoidance (Task 3): This is a systems engineering task. Some special node in the enclave will be declared "the manager". If some client node wants to poll N others, then it sends that list of nodes to the transaction manager. This manager orchestrates the SMC queries across that set of nodes such that no visited node knows who was visited before or after. Finally, the manager returns the

results of the SMC computation. The client knows that the results come from certain other nodes, but not which particular nodes offered which particular data items.

2.5d. Implement SMC management and Batch optimization (Task4): The high cost of SMC was noted above. Our computations must avoid too many low-level queries. Recalling Figure 5, we already have much support for believing that our learning can be done in one pass of the enclave members. Hence, in theory, we can avoid the high computational cost of standard SMC as follows. Rather than have a centralized data miner in the parent that runs multiple slow SMC queries through the children, we would instead pass the data miner around the children. That is, each data miner would process all the data inside a node in one batch operation. In theory, since this batch process reduces inter-node communication, it will remove the overheads of SMC (and this theoretical prediction must be tested via experimentation).

The next task addresses two concerns. First, in order to detect rules that violate privacy concerns, the ontology of the compliance regulation must match the generated model. Second, if some rule violates compliance, it should be possible to delete it without harming the rest of the learned model:

2.5e. Auditability and rule pruning (Task5): TAR6, since it is based on the TAR5 rule growth algorithm, will generate rules in the same ontology as Figure 3. Hence, auditability (recognizing if a rule matches a compliance restriction) will be a simple matter. Similarly, rule pruning is a simple matter in TAR6. The algorithms contain a stack of separate rules, all struggling to extend themselves in order to float towards top of stack. Pruning any rule from the stack (if it violates the compliance restrictions) will not stop the TAR6 algorithm.

While rule pruning will not stop TAR6, it may reduce the efficacy of the learned rules. Hence, task 10 (discussed below) is very important to this research (task 10 explores the tradeoff between the trustworthy computing properties of privacy vs data mining efficacy).

2.5f. Data pruning (Task6): Recall from the above that data pruning is the process of hiding data from an incoming SMC request such that it is difficult/impossible for that request to learn a pattern that the locals wish to keep private.

To meet this challenge, TAR6 reasons over the dependency graph of what rules lead to other rules. Before the locals at any enclave accept any SMC requests, they could run TAR6 to generate, e.g., the graph of Figure 8. Suppose the locals wanted to hide, say, (a) the information that 2nd-class women receive nearly the same preferential treatment as 1st-class women; or (b) advertise that female children have a better-than-most chance of survival. Figure 8 can be used to select rows that lead to some observations, but not others:

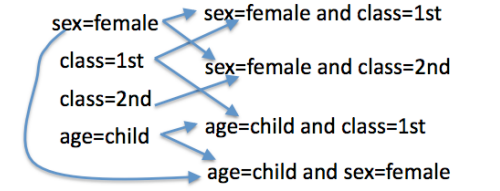

Figure 8*: Connections of smaller rules to larger rules from Figure 4.*

- Take each row of the data and present it to the left-hand-side of the dependency graph.
- If any range in the row is found in the root of the graph, then mark that row as *suspect*.
- Move the suspects right across the dependency graph, awarding points to suspects that match to preferred rules and subtracting points to rows that match to undesired rules.
- Hide the rows with H% lowest points.

Note that this algorithm is tunable: increasing H hides more detail while setting H=0 exposes all data to any incoming data miner. Once the above tasks are meet, we would have a working distributed data miner over nested enclaves. With that in hand, we turn to the remaining tasks.

2.5g. Efficacy of the data mining (Task 7): This whole proposal assumes that our data miners can find interesting patterns in data from software engineering and medical companies. Using historical data of known past patterns, this assumption must be tested. Hence, in this task, we will take analysis conclusions found in the past on our data sets, then check if our learners can find those known results.

We have extensive background information on what we expect to find in that data. For example, the NASA inspection database maintained at Fraunhofer has already been analyzed regarding influencing factors on the efficacy of inspections [Sea08, Shu10], and will serve as our oracle of what results we will first check for in the data. Since this dataset contains info from multiple NASA Centers, we will then check if those same, or different, results are to be found in the subset of data for each Center. After that,

there is much other SE data available for our analysis (for example, Fraunhofer datasets tracking progress of multi-year projects in the aerospace industry). Some of these projects have tuned their own estimation models already, and for others we will use industry "best practice" models such as COCOMO to yield an initial hypothesized set of influencing factors. Similarly, for the medical data, we will check if our learners can find any of the conclusions made previously with that data.

2.5h. Anomaly detection (Task 8): Once we have a working data miner, it would be possible for local administrators to test if rules learned elsewhere are working as expected at their local site. For example, in the medical domain, the locals would suspect an anomaly if a rule predicting for low post-operative infections (that is published by many other nodes) does not work well at this local site. Such anomalies could be an early warning sign for some previously undetected change in local conditions.

2.5i. Model revision (Task 9): In all the domains we study here, it is unlikely that once a model is learned, that it will remain constant for all time thereafter. For example:

- In software engineering domains, new technologies are constantly appearing.
- In medicine, health patterns change along with the seasons or as new drugs /diseases appear.

Hence, just as important as learning an effective model is knowing when to change an existing model. TAR6 implements theory revision as follows. Recall that TAR5 executes by selecting and combining R = 2 items from its internal stack. TAR6 modifies this selection policy as follows: select R ≥ 1 items from the stack. There are two interesting cases:

- If R > 1 then TAR6 is trying to combine existing items on the stack into larger conjunctions. That is, when R > 1, TAR6 is *building bigger rules*.
- Also, if R = 1, TAR6 is selecting one existing item from the stack, and restoring it. That is, when R=1, TAR6 is *reviewing old rules* (and possibly discarding them).

During this second case (review of old rule), it is possible to update TAR6's rules if changes to the observations have changed. If the Figure 5 scoring for a rule has altered, then TAR6 can demote (or promote) a rule according to its effectiveness on the latest observations.

2.5j. Assessing the price of privacy (Task 10): With all the above machinery in place, we will be able to perform distributed data mining where the results of that mining are audited and censored by compliance assurance requirements. The major question of this research can now be addressed: how much does compliance assurance hinder data mining? To test this, we take the privacy restrictions of Figures 6 and 7, express them as XPATH queries (e.g. as done in Figure 3), then randomly add/delete restrictions. This will generate a spectrum of queries ranging from least to most restrictive. We will then re-perform tasks 7, 8, 9 for each member of this space of queries. The result will be an envelope of privacy restrictions within which data mining efficacy is not damaged. Our research goal would be then to offer design guidelines for maintaining privacy while also maintaining data mining efficacy.

2.5k. Toolkit Generation (Task 11): Based on those lessons learned from tasks 1 to 11, we plan to devote year 3 to rewriting, generalizing, optimizing and TAR6. The result will be a software package suitable for distribution to other researchers (to ensure wide access, the code will be open sourced).

2.5l. Stress Test (Task 12): In year 4, we plan to test the toolkit built in year3 using a massive large scale-up of the size and geographical locations of our enclave. Our medical data will offer a ready test bed for such a scale up task. PI Pollard's data collection software is currently being deployed across the USA and USA Pacific Territories. By the fourth year of this project, this software will contain hundreds of thousands of records, expressed in the same database schema, describing disease patterns from across the planet.

2.6. LEADERSHIP AND COLLABORATION PLAN: This team has extensive experience in managing large multi-institutional research projects. For example,

- In his role as director of a research division at Fraunhofer, PI Shull has administrated collaborative research projects across the country and with Europe.
- In his role as SE research chair at NASA (2002-2008), Menzies ran data mining research projects with collaborators from NASA centers all over the USA.
- One reason for the widespread use of PI Pollard's software is the infrastructure support he offers along with his software (e.g. PI Pollard runs a national user group for those that use his software).

For the life of the project, we will:

- Conduct teleconferences every two weeks
- Run physical meetings every three months (rotating between Fraunhofer in Maryland and WVU).
- Maintain a website for this project where anyone can download nightly builds of our tools, as well as all our tech reports and publications.

From a management viewpoint, the project divides into four teams, each lead by different PIs:

	1. Admin	2. Systems	3. SE data analysis	4. Medical data analysis
Lead	Shull & Menzies	Menzies, Tanner	Shull	Morris, Pollard
Where	WVU + Fraunhofer	WVU	Fraunhofer	WVU

In the following gantt chart, the **admin** team leads the work shown in gray (i.e. tasks 0, 13, 14). These tasks include organizing the year1 set up, the annual reports, and regular face-to-face team meetings.

Task		Year 1				Year 2				Year 3				Year 4			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
0	Hiring and Set up																
1	Data/ compliance req,													5	5		
2	Implement rule fusion			1			2										
3	Add collusion avoidance																
4	SMC manage & batch opt																
5	Audit + prune rules				1			2			3		4		6	6	
6	Data pruning				1			2			3		4		6	6	
7	Data mining efficacy				1			2			3		4		6	6	
8	Anomaly detection					1			2		3		4		6	6	
9	Model revision					1			2		3		4		6	6	
10	Price of privacy					1			2		3		4		6	6	
11	Develop Toolkit																
12	Stress test																
13	Annual reports																
14	Intra-team meetings (at WV) Intra-team meetings (at Fraunhofer)																

The data **analysis** teams leads the work shown in green (on the medical data) and red (on the SE data):

- In years1&2, for task 1, they will take the data that we already have in hand and augment it with (a) its associated security restrictions; (b) its known baseline results. For the SE data, we have numerous baseline results in our prior publications (e.g. [Sea08, Shu10, Men10b], and others). For the medical data, we will examine the subset of Pollard's CDEMS data that comes from WV health clinics. This data will test the tools from the systems team (in the cells labeled “1” and”2”) as well as the generalized tool kit built in year 3 (in the cells labeled “3” and “4”).
- In year 4, we will stress test the toolkit written in year3 with large quantities of data from databases distributed around the world. For the medical data, this will be data from the CDEMS clinics across American and the USA Pacific Territories. For the SE data, we will also conduct tests on additional data (we expect to use data from additional NASA Centers). These stress tests will involve some initial modeling (see cells labeled ”5”) followed by the application of the tools on these larger data sets (in the cells labeled “6”).

The **systems** team will lead the work shown in black; i.e. construction of the software infrastructure used by the data analysis teams. Once built, the systems team will work with the other teams to test the software on the SE and medical data (see the cells labeled “1” and “2”). Also, in task 11, the **systems** team will spend Year 3 rewriting all the core tools using the lessons learned in years 1 & 2. During that rewrite, all year 1 & 2 experiments will be run again (see cells labeled “3” and “4”).

Note that the **systems** team will be very active, even when they are not being "lead" on any tasks. For example, during the activities in the cells labeled "1, 2, 3, 4, 6", the SE data and medical teams will (a) bring their domain knowledge to fashion the inputs to the tools then (b) critique the output of the tools. During that time, the systems team will be busy in a support role, running the tools, handling network issues, writing patches for bugs, etc. The systems team will be particularly busy in year 4, porting and maintaining our distributed data miner on all the distributed sites used as part of that stress test.

2.7. EVALUATION CRITERIA

Our goals for this work are twofold: 1) to learn important conclusions for our specific testbed areas (software effort estimation, software defect prediction / inspection deployment, chronic disease monitoring), and 2) while at the same time identifying general principles for the design of privacy restrictions such that they do not unnecessary hinder data mining.

Since our team is so large and diverse, the intermediary stepping stones in the following table have been identified to help us chart progress toward that goal.

Task		Milestone/ deliverable	Evaluation criteria
0	Hiring/Set up	Creation of web site	
1	Data/ compliance req,	Data sets, with baseline results extracted from prior historical analysis, compliance regulations mapped in XPATH	Deliverables exist for (a) the SE data; (b) for the medical data from the WV CDEMS sites; (c) In year 4, for the CDEM sites outside of WV
2	Implement rule fusion	Working prototype.	Business users, or leads of the SE/medical data teams can browse the fused rules and assert that they are reasonable inferences.
3	Add collusion avoidance	SMC working	SMC working
4	SMC manage & batch opt	Data mining queries via SMC can pass over the enclave	The SMC queries avoid the massive slowdown reported previously in the literature [Vai04]
5	Audit + prune rules	Working prototype.	Compliance assurance statements from task 1 can automatically prune rules
6	Data pruning	Working prototype	It can be demonstrated that administrators within an enclave can find and hide rows from the arriving SMC queries
7	Data mining efficacy	Working prototype of data miner.	The learner can reproduce (some) of the known historical baseline results that other researchers have extracted from this literature.
8	Anomaly detection	Working prototype of anomaly detector.	If business users or the leads of the SE/medical data teams seed rare but significant events, then these can be detected in the ensembles.
9	Model revision	Working prototype of model revision	If business users or the leads of the SE/medical data teams change the data such that old rules no longer apply, it is possible to update the rules to handle the new situations.
10	Price of privacy	Report on how increasing / decreasing privacy restrictions effect the efficacy on the data mining.	It is possible to characterize the space of privacy changes within which data mining efficacy is not damaged
11	Develop Toolkit	Download, installable software	
12	Stress test	Report on the results	The toolkit from task 11 scales to very large data sets distributed across the planet.
13	Annual reports	Reports	Reports registered in Fastlane.
14	Intra-team meetings	Meetings held	Updates to website describing current progress

2.8. EDUCATION / BROADER IMPACT:

Benefits to society at large: Our specific research goals focus on issues of tremendous economic or social importance (better control of software projects; better understanding of disease patterns in society).

Now more than ever, researchers are tackling ambitious research questions with vast potential for societal impact. Yet this research has often been hampered by the sensitivity of data from within commercial and government organizations – the data that is needed if we are to produce high-fidelity research results that are representative of real contexts. If successful, our work will result in concrete outputs (such as guidelines that enable data providers to organize themselves as trusted enclaves, and data miners that work within such enclaves) that enable other researchers to perform analyses across multiple data sources with many fewer restrictions due to data privacy. Science will become a world-wide crowd-sourcing activity where large communities quickly discover insights in shared data sources.

How will individuals at underserved institutions benefit from this grant? By funding this work at WVU, NSF will be promoting higher education in a region which, currently, lags far behind the rest of the country in terms of its population starting, or completing, a University degree. Appalachia, which includes West Virginia, is one of the most economically depressed regions in the United States. This economic condition greatly impacts the number of young people choosing to attend college. In the 1990's, the gap between Appalachia and the rest of the U.S. in the percent of the adult population who are college graduates increased from only 6.1% to 6.6% [Haa04]. In addition, the number of students seeking science and engineering degrees lags national averages: West Virginia and the rest of Appalachia are in the bottom quartile in the percent of science and engineering degrees awarded [Nsb06].

Also, by funding this work at the Fraunhofer Center at the University of Maryland, NSF will be promoting higher female involvement in engineering. The Fraunhofer has a strong track record of including women in software engineering research projects. 40% of the Fraunhofer technical staff are women, far above the latest numbers for women's representation in U.S. computer science degree programs [Dea07]. Fraunhofer's support of women in science is particularly important since, at this time, the rates of increase in computer science by men is far greater than by women [Dea07].

How will this research be integrated into teaching? Much of the research in this project will also be integrated into a classroom environment. PI Menzies teaches graduate data mining and all the tools will be used in that subject. PI Menzies also places all of his teaching materials on the web which means that any other data mining lecturer will be able to access tutorials, assignments, and lectures.

Dissemination of knowledge: We will make the developed tools and underlying technology for setting up enclaves freely available as open source system. Also, the PIs on this grant frequently publish in numerous research forums (IEEE TSE, IEEE Computer, IEEE Software, ASE, ICSE, etc). We hence anticipate numerous publications from this work.

Also, we have a goal of making some data from this work freely available for other researchers. The authors have an exemplary reputation of placing their data on-line (Shull ran the CeBASE repository [Bas01] while Menzies is the webmaster of the PROMISE repositories: http://promisedata.org). However, with regard to the data studied in this work, the practicality of this goal will be assessed with respect to the business or federal government restrictions on the data.

2.9. RESULTS FROM PRIOR NSF WORK: Dr. Forrest Shull was co-PI of NSF Science of Design collaborative grant CCF0438933 and CCF0438923, $1M from 2/1/05 to 1/31/09 "Collaborative Research: Flexible High Quality Design for Software". This project applied an empirical approach to investigate effective indicators for assessing the flexibility of software architecture, and V&V techniques which aim at improving flexibility under various conditions. The work produced a number of laboratory packages used to replicate studies and analyses at various collaborating sites. This grant provided support for 9 graduate students over its existence. It produced 7 journal publications [Bas05, Can07, Gup10, Lin05, Lin07, Shu05, Wil08], 18 conference publications [Ack06, Ack07, Ack09, Can05, Car05, Gup08, He09, Kar07, Kno05, Lin06, Mor07, Nak05, Old09, Sar08, Sar09a, Sar09b, Wil07, Zaz09], and 2 Master theses.

Dr. Menzies was awarded in July 2008 an NSF grant (CCF-0810879, $350,000, end date June 30 2011) on "Automatic Quality Assessment: Exploiting Knowledge of the Business Case", together with Dr. Bojan Cukic (Co-PI) from WVU. This research explores the inner loop of software data miners improves the learning of defect predictors. This project has produced 3 journal papers [Jia08a, Men10b, Tur09] and 11 conference papers [Cuk09, Gay09, Gre09b, Jia09, Jia08b, Kal09a, Kal09b, Men09b, Men08, Orr09, Cuk08]. Also, it has fully supported to completion one female Ph.D., one female masters and four others masters students to completion. It is also currently supporting one Ph.D. and one female masters student.

On August 15, 2010, Menzies started a new NSF grant, CCF-1017330: "Better Comprehension of Software Engineering Data" ($500,000; with Andrian Marcus at Wayne State).

References Cited

[Ack06] C. Ackermann and M. Lindvall, “Understanding change requests to predict software impact,” in SEW ’06: Proceedings of the 30th Annual IEEE/NASA Software Engineering Workshop, IEEE Computer Society, Washington, DC, USA, 2006, pp. 66–75.

[Ack07] C. Ackermann, F. Shull, R. Carbon, C. Denger and M. Lindvall, “Assessing the quality impact of design inspections,” in ESEM ’07: Proceedings of the First International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, Washington, DC, USA, 2007, pp. 470–472.

[Ack09] C. Ackermann, M. Lindvall, and G. Dennis, “Redesign for flexibility and maintainability: A case study,” in CSMR ’09: Proceedings of the 2009 European Conference on Software Maintenance and Reengineering, IEEE Computer Society, Washington, DC, USA, 2009, pp. 259–262.

[Agr94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

[Agr00] R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining,” in Proc. ACM SIGMOD Conf. Management of Data, ACM Press, 2000, pp. 439–450.

[Agr03] R. Agrawal, J. Kiernan, R. Srikant and Y. Xu., “Implementing P3P Using Database Technology,” ICDE, 2003.

[Aha91] D. W. Aha, D. Kibler and M. K. Albert, “Instance-Based Learning Algorithms,” Mach. Learn., vol. 6, no. 1, pp. 37-66, Jan. 1991.

[Bas99] V.R. Basili, F. Shull, and F. Lanubile, “Building knowledge through families of experiments,” IEEE Trans. Softw. Eng., vol. 25, no. 4, pp. 456–473, 1999.

[Bas01] V. Basili, R. Tesoriero, P. Costa, M. Lindvall, I. Rus, F. Shull, and M. Zelkowitz. Building an experience base for software engineering: A report on the first cebase eworkshop. In in Profes (Product Focused Software Process Improvement, pages 110–125, 2001.

[Bas05] V. Basili and F. Shull. "Evolving Defect 'Folklore': A Cross-Study Analysis of Software Defect Behavior," Lecture Notes in Computer Science, vol. 3840, Berlin: Springer, pp. 1-9, 2005.

[Bec80] L.L. Beck, “A security mechanism for statistical databases,” ACM Transactions on Database Systems, vol. 5, no. 3, pp. 316–338, Sep. 1980.

[Bla98] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. URL: http://www.ics.uci.edu/˜ mlearn/MLRepository.html.

[Bre84] L. Breiman et al., “Classification and Regression Trees,” Wadsworth Int’l, Tech. Rep., 1984.

[Bri08] J. Brickell and V. Shmatikov, “The cost of privacy: destruction of data-mining utility in anonymized data publishing,” in Proceeding of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, Aug. 24 - 27, 2008.

[Can05] A. Betin-Can, T. Bultan, M. Lindvall, B. Lux, and S. Topp, “Application of design for verification with concurrency controllers to air traffic control software,” in ASE ’05: Proceedings of the 20th IEEE/ACM international Conference on Automated Software Engineering, ACM, New York, NY, USA, 2005, pp. 14–23.

[Can07] Aysu Betin Can, Tevfik Bultan, Mikael Lindvall, Benjamin Lux and Stefan Topp, “Eliminating synchronization faults in air traffic control software via design for verification with concurrency controllers,” Automated Software Engineering, vol. 14, no. 2, pp. 129–178, 2007.

[Car05] J. Carver and K. Lemon, “Architecture reading techniques: A feasibility study,” in Proceedings of 2005 International Symposium on Empirical Software Engineering (Late Breaking Research Track), Noosa, Australia, 2005, pp. 17–20.

[Chi82] F. Chin and G. Ozsoyoglu, “Auditing and infrence control in statistical databases,” IEEE Trans. on Software Eng., vol. 8 no. 6, pp. 113–139, Apr. 1982.

[Cla05] R. Clark, “Faster treatment learning,” in Computer Science, Portland State University. Master’s thesis, 2005.

[Cli04] C. Clifton, M. Kantarcioğlu, A. Doan, G. Schadow, J. Vaidya, A. Elmagarmid, and D. Suciu, “Privacy-preserving data integration and sharing,” in Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Paris/France, DMKD '04, 2004.

[Cox80] L. Cox, “Suppression methodology and statistical disclosure control,” J. Am. Stat. Assoc., vol. 75, no. 370, pp. 377–395, Apr. 1980.

[Cuk08] B. Cukic, Y. Jiang and T. Menzies, “Cost curve evaluation of fault prediction models,” in Proceedings, ISSRE’08, 2008.
(Available online at http://menzies.us/pdf/08costcurves.pdf)

[Cuk09] B. Cukic, T. Menzies, and Y. Jiang. Variance analysis in software fault prediction models. In IEEE ISSRE’09, 2009. Available from http://menzies.us/pdf/09irrf.pdf.

[Dea07] Cornelia Dean, “Computer science takes steps to bring women to the fold,” New York Times, Apr. 2007.

[Dem77] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, pp. 1–38, 1977.

[Den79] D. Denning, P. Denning, and M. Schwartz, “The tracker: A threat to statistical database security,” ACM Transactions on Database Systems, vol. 4, no. 1, pp. 76–96, Mar 1979.

[Den82] D. Denning, “Cryptography and Data Security,” Addison-Wesley, 1982.

[Dim06] A.G. Dimakis, A.D. Sarwate, and M.J. Wainwright, “Geographic gossip: Efficient aggregation for sensor networks,” in PSN06, Nashville, Tennessee, Apr. 1921, 2006, pp. 69–76.

[Dob79] D. Dobkin, A. Jones, and R. Lipton, “Secure databases: Protection against user influence,” ACM Transactions on Database Systems, vol. 4, no. 1, pp. 97–106, Mar. 1979.

[Eps10] Epstein, Arnold M. Geographic Variation in Medicare Spending. New England Journal of Medicine 2010; 363:85-86 July 1, 2010

[Fun08] B. C. M. Fung, K. Wang, L. Wang, and C. K. Hung Patrick, „Privacy-Preserving Data Publishing for Cluster Analysis,” in Systems Engineering, Dec. 2008, pp. 1-43.

[Fun10] B.C.M. Fung, K. Wang, R. Chen, “Privacy-Preserving Data Publishing: A Survey of Recent Developments,” ACM Computing Surveys, vol. 42, no. 4, article 14, Jun. 2010.

[Gay09] G. Gay, T. Menzies, B. Cukic, and Burak Turhan, “How to build repeatable experiments,” PROMISE’09, 2009.
(Available online at http://menzies.us/pdf/09ourmine.pdf)

[Gay10] Gregory Gay, Tim Menzies, Misty Davies and Karen Gundy-Burlet, “Automatically finding the control variables for complex system behavior,” Automated Software Engineering, vol. 1 / 1994 - Volume 17 / 2010.

[Gre06] Jan Greene, "Under-Mined. Hospitals & Health Networks,” ABI/INFORM Global, vol. 80, no. 12, pp. 38-40, 42, 44, first retrieved Apr. 23, 2010, (Document ID: 1192850141), Dec. 2006.

[Gre09b] P. Green, T. Menzies, S. Williams, and O. El-waras, “Understanding the value of software engineering technologies,” IEEE ASE’09, 2009.

(Available from http://menzies.us/pdf/09value.pdf)

[Gun08] K. Gundy-Burlet, J. Schumann, T. Menzies, and T. Barrett, “Parametric analysis of antares re-entry guidance algorithms using advanced test generation and data analysis,” in 9th International Sym- posium on Artifical Intelligence, Robotics and Automation in Space, 2008.

[Gun09] K. Gundy-Burlet, J. Schumann, T. Menzies, and T. Barrett, “Parametric analysis of a hover test vehicle using advanced test generation and data analysis,” in AIAA Aerospace 2009.

[Gup08] A. Gupta, R. Conradi, F. Shull, D. Cruzes, C. Ackermann, H. Rønneberg, and E. Landre, “Experience report on the effect of software development characteristics on change distribution,” in PROFES ’08: Proceedings of the 9th international conference on Product-Focused Software Process Improvement, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 158–173.

[Gup10] A. Gupta, D. Cruzes, F. Shull, R. Conradi, H. Rønneberg, and E. Landre, “An examination of change profiles in reusable and non-reusable software systems,” Journal of Software Maintenance and Evolution: Research and Practice, vol. 22, no. 5, pp. 359-380, August 2010. Available at: http://onlinelibrary.wiley.com/doi/10.1002/smr.459/abstract

[Haa04] J. Haaga, Appalachian Regional Commission, and Population Reference Bureau. Educational attainment in appalachia. page 24 p., 2004. [electronic resource] / by John Haaga. digital, PDF file. Demographic and socioeconomic change in Appalachia. Title from title screen (viewed on Aug. 14, 2008) ”May 2004.” Mode of access: internet from the ARC web site. Address as of 8/14/08: http://www.arc.gov/images/reports/prbeducation/attainment.pdf ; current access via PURL.

[He09] L. He and J. Carver, “Modifiability measurement from a task complexity perspective: A feasibility study,” in ESEM ’09: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, Washington, DC, USA, 2009, pp. 430–434.

[Hin92] G.E. Hinton, “How neural networks learn from experience,” Scientific American, pp. 144–151, Sep. 1992.

[Jia08a] Y. Jiang, B. Cukic, and Y. Ma. Techniques for evaluating fault prediction models. Empirical Software Engineering, pages 561–595, October 2008.

[Jia08b] Y. Jiang, B. Cukic, and T. Menzies. Can data transformation help in the detection of fault-prone modules. In DEFECTS’08, 2008.

[Jia09] Y. Jiang and B. Cukic. Misclassification cost-sensitive fault prediction models. In PROMISE 2009, 2009.

[Kal09a] N. Kalka, N. Bartlow, and B. Cukic. An automated method for predicting iris segmentation failures. In IEEE 3rd Int’l Conference on Biometrics Theory, Applications and Systems (BTAS 2009), 2009.

[Kal09b] N. Kalka, N. Bartlow, and B. Cukic. Decision dependability and its application to identity management. In Cyber Security and Information Intelligence Research Workshop, Oak Ridge, TN, 2009.

[Kar07] K. Karppinen and M. Lindvall. Software architecture-driven detection of security vulnerabilities. In Proc. World Congress in Computer Science, Computer Engineering, & Applied Computing, June 25-28, 2007. Las Vegas, NV, USA, 2007.

[Kit07] B. A. Kitchenham, E. Mendes, and G. H. Travassos, “Cross- vs. within-company cost estimation studies: A systematic review,” IEEE Transactions on Software Engineering, pp. 316–329, May 2007.

[Koc10] Ekrem Kocaguneli, Gregory Gay, Tim Menzies, Ye Yang, and Jacky W. Keung, “When to use data from other projects for effort estimation,” IEEE ASE’10, 2010.

[Kno05] J. Knodel, M. Lindvall and D. Muthig, “Static evaluation of software architectures - a short summary,” in WICSA ’05: Proceedings of the 5th Working IEEE/IFIP Conference on Software Architecture, IEEE Computer Society, Washington, DC, USA, 2005, pp. 237–238.

[Lin05] M. Lindvall, I. Rus, F. Shull, M. V. Zelkowitz, P. Donzelli, A. Memon, V. R. Basili, P. Costa, R. T. Tvedt, L. Hochstein, S. Asgari, C. Ackermann, and D. Pech. An evolutionary testbed for software technology evolution. Innovations in Systems and Software Engineering - A NASA Journal, 1:3–11, 2005.

[Lin06] M. Lindvall, D. Muthig, J. Knodel, and M. Naab, “Static evaluation of software architectures,” in Conference on Software Maintenance and Reengineering, 2006.

[Lin07] M. Lindvall, I. Rus, P. Donzelli, A. Memon, M. Zelkowitz, A. Betin-Can, T. Bultan, C. Ackermann, B. Anders, S. Asgari, V. Basili, L. Hochstein, J. Fellmann, F. Shull, R. Tvedt, D. Pech, and D. Hirschbach. Experimenting with software testbeds for evaluating new technologies. Empirical Software Engineering, 12(4):417–444, 2007.

[Mas07] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, “L-diversity: Privacy beyond k-anonymity,” TKDD, vol. 1, no. 1, 2007.

[Men03a] T. Menzies and J.S. Di Stefano, “More success and failure factors in software reuse,” IEEE Transactions Softw. Eng., May 2003.

[Men06] T. Menzies, Z. Chen, J. Hihn, and K. Lum, ”Selecting best practices for effort estimation,” IEEE Trans. Softw. Eng., Nov. 2006.

[Men07a] T. Menzies, A. Dekhtyar, J. Distefano, and J. Greenwald, “Problems with precision,” IEEE Trans. Softw. Eng., Sep. 2007.

[Men07d] T. Menzies, J. Greenwald and A. Frank, "Data Mining Static Code Attributes to Learn Defect Predictors,” IEEE Trans. Softw. Eng., (Available from http://menzies.us/pdf/06learnPredict.pdf), Jan. 2007.

[Men08] T. Menzies, S. Williams, O. El-rawas, B. Boehm, and J. Hihn. How to avoid drastic software process change (using stochastic statbility). In ICSE’09, 2009. Available from http://menzies.us/pdf/08drastic.pdf.

[Men09a] T. Menzies, S. Williams, O. El-rawas, B. Boehm, and J. Hihn, “How to avoid drastic software process change (using stochastic stability),” ICSE’09, 2009.

[Men09b] T. Menzies, O. El-Rawas, J. Hihn and B. Boehm, “Can we build software faster and better and cheaper?,” PROMISE’09, 2009.
(Available online at http://menzies.us/pdf/09bfc.pdf)

[Men10a] G. Gay, T. Menzies, M. Davies and K. Gundy-Burlet, “Automatically finding the control variables for complex system behavior,” Automated Software Engineering, vol. 17, no. 4, pp. 439-468.

[Men10b] T.J. Menzies, Z. Multon, B. Turhan, B. Cukic, Y. Jiang, and A. Bener. Defect prediction from static code features: Current results, limitations, new approaches. Automated Software Engineering, 2010.

[Mil08] Z.A. Milton, “Which rules,” in Master’s thesis, Lane Department of Computer Science and Electrical Engineering, 2008.

[Mor07] I. Morschhauser and M. Lindvall, “Model-based validation & verification integrated with SW architecture analysis: A feasibility study,” Proc. 2007 IEEE Aerospace Conference, p. 1–18, Big Sky, MT, USA, Mar. 3-10, 2007.

[Mor09] Joseph D'Alessandro, Cynthia Tanner, Bonnie Morris, Tim Menzies, “Is Continuous Compliance Assurance Possible?,” in International Conference on Information Technology: New Generations, 2009.

[Mor10] Bonnie Morris, Cynthia Tanner, Joseph D'Alessandro, "Enabling Trust through Continuous Compliance Assurance," in Information Technology: New Generations,

Third International Conference on, 2010 Seventh International Conference on Information Technology, 2010, pp. 708-713.

[Nak05] Taiga Nakamura and Victor R. Basili, “Metrics of software architecture changes based on structural distance,” METRICS ’05: Proceedings of the 11th IEEE International Software Metrics Symposium, IEEE Computer Society, p. 8, Washington, DC, USA, 2005.

[Ness07] R.B. Ness, “Influence of the HIPAA Privacy Rule on health research,” Journal of the American Medical Association, vol. 298, no. 18, pp. 2196-8, Nov. 2007.

[Ness10] R.B. Ness, “Population-based human subjects research in the era of enhanced privacy regulation,” American Journal Epidemiology, vol. 172, no. 6, pp. 648-50, discussion 651-2, Sep. 15, 2010.

[Nsb06] National Science Board (U.S.). Science and engineering indicators 2006, —2006—. [electronic resource] digital, PDF file. Title from title screen (viewed on Sep. 18, 2006) Mode of access: Internet from the NSF web site. Address as of 9/18/2006: http://www.nsf.gov/statistics/seind06/pdf/volume1.pdf; http://www.nsf.gov/statistics/seind06/pdf/volume2.pdf; current access available via PURL.

[Olb09] Steffen Olbrich, Daniela S. Cruzes, Victor Basili, and Nico Zazworka, “The evolution and impact of code smells: A case study of two open source systems,” ESEM ’09: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, p. 390–400, Washington, DC, USA, 2009.

[Olf07] R. Olfati-Saber, “Distributed kalman filtering for sensor networks,” in Decision and Control, 2007 46th IEEE Conference on, Dec. 2007, pp. 5492 –5498.

[Orr09] A. Orrego, T. Menzies, and O. El-Rawas. On the relative merits of software reuse. In International Conference on Software Process, 2009. Available from http://menzies.us/pdf/09reuse.pdf.

[Pol09a] C. Pollard with S. Frisbee, AP Brooks, A Maher, P. Flensborg, S. Arnold, T. Fletcher, K. Steenland, A. Shankar, S. Knox, J. Halverson, V. Vieira, C. Jin, L. Leyden, “The C8 Health Project: design, methods, and participants,” Environ Health Perspect, Environmental Health Perspectives, vol. 117, no. 12, Dec. 2009.

[Pol09b] C. Pollard, A. Kelly, M.A. Bailey, T. Petitte, A. Baus, M. Swim, M. Hendryx, “Electronic Patient Registries Improve Diabetes Care and Clinical Outcomes in Rural Community Health Centers,” The Journal of Rural Health, Winter 2009.

[Qui92] R. Quinlan, “C4.5: Programs for Machine Learning,” Morgan Kaufmann, 1992.

[Ram10] Ramakrishnan, Naren, David A. Hanauer, Benjamin J. Keller. Mining Electronic Health Records. Computer October 2010: 77-81

[Rei84] S. P. Reiss, “Practical data-swapping: The first steps,” ACM Trans. on Database Systems, vol. 9, no. 1, pp. 20–37, 1984.

[Sar08] S.A. Sarcia, G. Cantone, and V.R. Basili. Adopting curvilinear component analysis to improve software cost estimation accuracy: Model, application strategy, and an experimental verification. In Evaluation and Assessment in Software Engineering (EASE 2008), University of Bari, Italy, June 2008.

[Sar09a] S. A. Sarcia, V. R. Basili, and G. Cantone, “Scope error detection and handling concerning software estimation models,” ESEM ’09: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, p. 123–132, Washington, DC, USA, 2009.

[Sar09b] S. A. Sarcia, V. R. Basili, and G. Cantone, “Using uncertainty as a model selection and comparison criterion,” PROMISE ’09: Proceedings of the 5th International Conference on Predictor Models in Software Engineering, ACM, p. 1–9, New York, NY, USA, 2009.

[Sea08] C. Seaman, F. Shull, M. Regardie, D. Elbert, R. L. Feldmann, Y. Guo, S. Godfrey, "Defect Categorization: Making Use of a Decade of Widely Varying Historical Data," in Proceedings of the International Symposium on Empirical Software Engineering and Measurement (ESEM08), Kaiserslautern, Germany, October 2008.

[Shu00a] F. Shull, F. Lanubile, and V.R. Basili, "Investigating reading techniques for object-oriented framework learning," IEEE Trans. Softw. Eng., vol. 26, no. 11, pp. 1101–1118, 2000.

[Shu00b] F. Shull, I. Rus, and V.R. Basili, "How perspective-based reading can improve requirements inspections," IEEE Computer, vol. 33, no. 7, pp. 73–79, 2000.

[Shu02a] Forrest Shull, Victor Basili, Jeffrey Carver, Jose' C. Maldonado, Guilherme Horta Travassos, Manoel Mendonca, and Sandra Fabbri, "Replicating software engineering experiments: Addressing the tacit knowledge problem," ISESE '02: Proceedings of the 2002 International Symposium on Empirical Software Engineering, IEEE Computer Society, p. 7, Washington, DC, USA, 2002.

[Shu02b] F. Shull, V.R. Basili and B. Boehm, A.W. Brown, P. Costa, M. Lindvall, D. Port, I. Rus, R. Tesoriero and M.V. Zelkowitz, "What we have learned about fighting defects," in Proceedings of 8th International Software Metrics Symposium, Ottawa, Canada, 2002, pp. 249–258.
(Available online at http://fc-md.umd.edu/fcmd/Papers/shull_defects.ps)

[Shu05] F. Shull, D. Cruzes, V. Basili and M. Mendonca, "Simulating families of studies to build confidence in defect hypotheses," Inf. Softw. Technol., vol. 47, no. 15, pp. 1019–1032, 2005.

[Shu07] F. Shull, J. Singer, and D. I. K. Sjøberg, (eds.) Guide to Advanced Empirical Software Engineering, London: Springer, 2007.

[Shu10] F. Shull., R. Feldmann, C. Seaman, M. Regardie, and S. Godfrey, "Fully Employing Software Inspections Data," Innovations in Systems and Software Engineering - a NASA Journal, 2010. Available at: http://dx.doi.org/10.1007/s11334-010-0132-1

[Swe02a] L. Sweeney, "K-anonymity: a model for protecting privacy," International journal on uncertainty, Fuzziness and knowldege based systems, vol. 10, no. 5, pp. 557–570.

[Tos10] A. Tosun, A. Bener, R. Kale, AI-Based Software Defect Predictors: Applications and Benefits in a Case Study (Deployed) , IAAI 2010.

[Tse04] Shin-Mu Tseng and Ching-Fu Tsui, "Mining multilevel and location-aware service patterns in mo- bile web environments. Systems, Man, and Cybernetics, Part B: Cybernetics," IEEE Trans., vol. 34, no. 6, pp. 2480 –2485, Dec. 2004.

[Tur09] B. Turhan, T. Menzies, A. Bener, and J. Distefano, "On the relative value of cross-company and within-company data for defect prediction," Empirical Software Engineering, vol. 68, no. 2, pp. 278–290, 2009.
(Available online at http://menzies.us/pdf/08ccwc.pdf)

[Vai04] J. Vaidya, C. Clifton, "Privacy-Preserving Data Mining: Why, How, and When," IEEE Security and Privacy, pp. 19-27, November-December, 2004

[War65] S. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," J. Am. Stat. Assoc., vol. 60, no. 309, pp. 63–69, Mar. 1965.

[Wil07] B. J. Williams and J. C. Carver, "Characterizing software architecture changes: An initial study," in ESEM '07: Proceedings of the First International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, Washington, DC, USA, 2007, pp. 410–419.

[Wil08] B. Williams and J. Carver, "Characterizing software architecture changes: A systematic review," in Tech. Rep., Department of Computer Science and Engineering, Mississippi State University Technical Report MSU-081216, Dec. 2008.

[Witt05] I. H. Witten and E. Frank, "Data Mining, Second Edition," Morgan Kaufmann, Los Altos, US, 2005.

[Xia09] Liang Xiao, Bo Hu, L. Hederman, P. Lewis, B. D. Dimitrov and T. Fahey, "Towards knowledge sharing and patient privacy in a clinical decision support system," in Information Technology Interfaces 09 - Proceedings of the ITI 2009 31st International Conference, Jun. 2009, pp. 22-25, 99-104.

[Yu77] C. Yu and F. Chin, “A study on the protection of statistical databases,” in Proc. of the ACM SIGMOD Conf. on Management of Data, 1977, pp. 169–181.

[Zaz09] N. Zazworka, V. R. Basili and F. Shull, “Tool supported detection and judgment of nonconformance in process execution,” in ESEM ’09: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, Washington, DC, USA, 2009, pp. 312–323.

[Zha07] N. Zhang and W. Zhao, “Privacy-Preserving Data Mining Systems,” Computer, vol. 40, no. 4, pp. 52-58, Apr. 2007.

Biographical Sketch – Tim Menzies

Professional Preparation

Undergraduate Institutions

University New South Wales, Computer Science, BSc 1985

Graduate Institutions

University of NSW, Austrlia, MCogSc (Cognitive Science) 1989

University of NSW, Austrlia, Ph.D. (AI) 1996 (advisor: Paul Compton)

Appointments

Academic

Assoc/Professor of CSEE, West Virginia University, 2006--

NASA Software Engineering Research Chair, 2001-2003

Research Assist/Professor of CSEE, West Virginia University, 1998 – 2005

Industrial

Commercial object-oriented and expert system consultant: 1986 – 1995

Publications

Related Publications

1. "Defect prediction from static code features: current results, limitations, new approaches" Tim Menzies, with Zach Milton, Burak Turhan, Bojan Cukic, Yue Jiang and Ayşe Bener **Journal of Automated Software Engineering**, December 2010 p375-407.
2. "Stable rankings for different effort models Tim Menzies, Omid Jalali, Jairus Hihn, Dan Baker and Karen Lum , **Journal of Automated Software Engineering**, December 2010 p 409-437
3. "On the Relative Value of Cross-Company and Within-Company Data for Defect ,Prediction" by T. Menzies with B. Turhan , A. Bener and J. Distefano. **Empirical Software Engineering** 2009 . Available from http://menzies.us/pdf/08ccwc.pdf .
4. Tim Menzies, Zhihao Chen, Jairus Hihn, and Karen Lum. Selecting best practices for effort estimation. **IEEE Transactions on Software Engineering**, November 2006. Available from http://menzies.us/pdf/06coseekmo.pdf .
5. Tim Menzies, Jeremy Greenwald, and Art Frank. Data mining static code attributes to learn defect predictors. **IEEE Transactions on Software Engineering,** January 2007. http://menzies.us/pdf/06learnPredict.pdf.

Other Publications

6. Genetic Algorithms for Randomized Unit Testing Tim Menzies with James H. Andrews, Felix C.H. Li, **IEEE Transactions on Software Engineering**, 25 Mar. 2010
7. Tim Menzies, Alex Dekhtyar, Justin Distefano, and Jeremy Greenwald. Problems with precision. **IEEE Transactions on Software Engineering,** September 2007. http://menzies.us/pdf/07precision.pdf .
8. "More Success and Failure Factors in Software Reuse" by T. Menzies and J.S. Di Stefano. **IEEE Transactions on Software Engineering** May 2003 . http://menzies.us/pdf/02sereuse.pdf
9. T. Menzies, O. Elwaras, J. Hihn, Feathear and B. Boehm M, and R. Madachy. The business case for automated software engineerng. In **IEEE ASE, 2007**. Available from http://menzies.us/pdf/07casease-v0.pdf .
10. "Better reasoning about software engineering activities" by T. Menzies and J.D. Kiper. **IEEE ASE-2001** 2001 . Available from http://menzies.us/pdf/01ase.pdf ..

Synergistic Activities

Innovations in teaching and training :

Teaching graduate level data mining, WVU 2002-2009

Development and/or refinement of research tools:

Creation of an open source data mining tool kit (http://unbox.org/wisp);
Creator of the treatment learning tools

Development of databases to support research and education:

Curator of the PROMISE on-line repository on SE data. Currently, 90+ datasets, growing 40% each year. See http://promisedata.org

Service to the scientific and engineering community outside of the individual's immediate organization:

Editorial board member: Journal of Empirical SE; Journal of Vis Lang & Computing.
PC member: ESEM'10, INFOS'10, MSR'10, ASE'09, ISSRE'09.
Steering committee member: PROMISE workshop;
Scientific advisor to NASA on their national software assurance research program

Collaborators & Other Affiliations

Collaborators and Co-Editors:

- D. Allen (CS, Portland U.), Barry W. Boehm (CS, USC), Zhihao Chen (CS, USC), S.. Cornford (JPL), B. Cukic (CS WVU), Alex Dekhtyar (CS UK), O. Elwaras (CS WVU), Martin S. Feather (JPL), Marcus S. Fisher (NASA), J. Gao (CS, U. Minnesota), Art Frank (CS, Portland U.), M. Heimdahl (CS, U. Minnesota), Jairus Hihn (JPL), Y. Hu (EE, UBC), Y. Jiang (CS, WVU), James D. Kiper (CS, U. Miami), Karen T. Lum (JPL), R. Madachy (CS, USC), A. Orrego. (CS, WVU), David Owen (CS, WVU), Charles Pecheur (NASA), Daniel Port (CS, U. Hawaii), David Raffo (CS, Portland U.), Julian Richardson (NASA), Justin S. Di Stefano (CS, WVU)

Graduate Advisors and Postdoctoral Sponsors:

- Paul Compton, Claude Sammut, UNSW

Graduate Students and Advisees:

- Ph.D.: D. Owen 2007; W. Abdelmoez 2006; K. Cooper 2001;
- Masters: Z. Milton 2008, O. Jalali, 2008, D. Baker 2008, B. Taylor, 2008, D. Boland 2007, R. Clark 2006; G. Greenwald 2006; E. Chaing 2003; Y. Hu 2003; D. Owen 2003;

Bonnie W. Morris, Ph.D.

Go-Mart Professor of Accounting Information Systems, Division of Accounting
College of Business and Economics, West Virginia University
Morgantown, WV 26506-6025
Phone: (304) 293-7851 Fax: (304) 293-0635 E-mail: bmorris@wvu.edu

Professional Preparation

1992	Ph.D.	Accounting and Artificial Intelligence Joseph M. Katz Graduate School of Business University of Pittsburgh, Pittsburgh, PA
1991	M.B.A.	Joseph M. Katz Graduate School of Business University of Pittsburgh, Pittsburgh, PA
1974	B.A.	Mathematics, West Virginia University

Appointments

2007-present	GoMart Professor of Accounting Information Systems, West Virginia University
1997-present	Associate Professor, Department of Accounting, West Virginia University
1998-2000	Interim Director of the Center for Electronic Commerce, College of Business and Economics, West Virginia University
1990- 1997	Assistant Professor, Department of Accounting, West Virginia University
1980-1986	Private practice-Public accounting with a concentration in consulting in small business accounting information systems implementation, Pittsburgh, PA
1977-1980	Director, Internal Auditing, Duquesne University, Pittsburgh, PA
1974-1977	Senior Auditor, Arthur Andersen & Co., Pittsburgh, PA

Recent Published Refereed Journal Articles and Proceedings

Related Publications

Morris, B., Tanner, C., D'Alessandro, J. (2010). "Enabling Trust Through Continuous Compliance Assurance." In Shahram Latifi (Ed.), *International Conference on Information Technologies - New Generation.* Washington DC: IEEE Computer Society/CPS (IEEE Conference Publishing Services*)*

D'Alessandro, J., Tanner, C., Morris, B., Menzies, T. (2009). "Is Continuous Compliance Assurance Possible?" Published Abstract in Proceedings of the ITNG 2009, 6th International Conference on Information Technology : New Generations, Las Vegas, Conference, Technology, IEEE Computer Society.

Morris, B., Shaw, G., Tanner, C.,Trapp, G. "Continuous Compliance Assurance for Trusted Information Sharing: A Research Framework," Fourteenth World Continuous Auditing and Reporting Symposium, Rutgers University, Newark, NJ. (November 3, 2007).

Morris, B., Sinha, A. (2005). "Applicability of Case-Based Reasoning for Business Problems: A Study of Three Systems." Artificial Intelligence in Accounting and Auditing, 6, 214-221.

Other Recent Publications

Kuhn, J.R., J. F.Courtney, and B.W. Morris, "Agent-Based Analysis and Simulation of the Consumer Airline Market Share for Frontier Airlines" (forthcoming 2010) *Journal of Knowledge-Based Systems.*

Kleist, V., Morris, B., Denton, J. (2009). "Information Systems Security Assurance Management at Municipal Software Solutions, Inc" *International Journal of Information Security and Privacy*, 3(2), pp. 1-9.

Santucci, J. M., Cervone, D. P., Morris, B., Neidermeyer, P., Fleming, S. (2010). In Tracy Tuten (Ed.), *Accounting in the Clouds: How Web 2.0, Cloud Computing, and SaaS are impacting the Accounting Profession* (vol. 1). Greenwich. CT: Praeger Publishing.

Synergistic Activities

Recent Funded Research

"Information Fusion Networks For Intelligence and Security (InfoNets)," Sponsored by WVEPSCOR, $1,796,211. (2007 - Present). Co-Principal Investigators: Arun Ross, Donald Adjeroh, Robert Duval, Elaine Eschen, Eddie Fuller, Bonnie Morris, Jason Thomas, and CQ Zhang (all at WVU).

"Continuous Policy Compliance Auditing," as a sub-contractor to the University of Oklahoma as part of the US Department of Transportation's Intermodal Containerized Freight Security Project, $187,689 (2006-2010). Co-Principal Investigators: Bonnie Morris, Cynthia Tanner, and George Trapp (all at WVU)

"RTCE-Continuous Auditing," Morris, Bonnie (Co-Principal).Sponsored by Lockheed Martin Corporation IRAD Program, Private, $543,000. (2003-2004).

Teaching

Graduate: IT Auditing, Fraud Data Analysis
Undergraduate: Accounting Systems

Editorial Board Member

Journal of Information Systems,
Journal of Emerging Technologies in Accounting,
International Journal of Digital Accounting Research

Cecil R. Pollard
Research Assistant Professor, Department of Community Medicine

a. Professional Preparation

- Florida Atlantic University, Boca Raton, Fl., BS in Education and Sociology, 1972
- West Virginia University, Morgantown, WV, MA in Sociology and Anthropology, 1982

b. Appointments

- Research Assistant Professor, Dept. of Community Medicine, WVU 1989-present
- Director, WVU, Dept. of Community Medicine, Office of Health Services Research 1984-present
- Project Manager, WVU, Dept. of Community Medicine, Office of Health Services Research 1983-1984
- Research Technologist II, WVU, Dept. of Community Medicine, Office of Health Services Research 1982-1983
- Research Technologist I, WVU, Dept. of Community Medicine, Office of Health Services Research 1982

c. **Publication**

- Pollard, C., Bailey, K. A., Petitte, T., Baus, A., Swim, M., Hendryx, M. "Electronic Patient Registries Improve Diabetes Care and Clinical Outcomes in Rural Community Health Centers." Journal of Rural Health, Winter 2009

- Charumathi S, Shankar A, Li J, Pollard C, Ducatman A. Serum gamma-glutamyl transferase level and diabetes mellitus among US adults. Eur J Epidemiol 2009; 24:369-73.

- Gainor RE, Fitch, D, Pollard D. Maternal Diabetes and Perinatal Outcomes in West Virginia Medicaid Enrollees. *West Virginia Medical Journal.* January/February 2006. 102: 314-316.

- Goins RT, Gainor SJ, Pollard C, Spencer SM. Geriatric knowledge and educational needs among rural health care professionals. *Educational Gerontology. 2003.* 29(3): 261-272.

d. Synergistic Activities

- Conceived and implemented National CDEMS Support Group 2008
- Member CDC Expert Panel on Electronic Medical Records 2006
- Consulting Pacific Island Tuberculosis Control Association and Pacific Island Chronic Disease Programs
- Consultant to Oklahoma Health Department Cardiovascular Health Program and the Oklahoma Primary Care Association Chronic Disease Registry Development
- Consultant to Johns Hopkins University Diabetes Center's Trinidad-Tobago Diabetes Project

e. Collaborators & Other Affiliations

Collaborators:
Chris Saudek (Johns Hopkins University, Comprehensive Diabetes Center), Henry Taylor (Johns Hopkins University, School of Public Health)

Graduate Students and Advisees:

Cynthia D. Tanner
Program Coordinator/Lecturer
C0-Director of the WVU Continuous Compliance Assurance Laboratory
West Virginia University
Lane Department of Computer Science and Electrical Engineering
951 Engineering Sciences Building, Morgantown, WV 26506-6109
Phone: 304-293-9138 Fax 304-293-8602 Email: Cindy.Tanner@mail.wvu.edu

Professional Preparation

West Virginia University	MS	Computer Science	1979
Northwestern University	BSE	Mathematics Education	1975

Appointments

2004-present	Computer Science Undergraduate Coordinator Lane Department of Computer Science and Electrical Engineering
1999-2004	Graduate Coordinator of Software Engineering Lane Department of Computer Science and Electrical Engineering
1993-present	Lecturer/Program Coordinator Lane Department of Computer Science and Electrical Engineering
1981-1990	Lecturer Department of Statistics and Computer Science
1978-1980	Programmer/Analyst Health Sciences Systems, West Virginia University
1977-1978	System Analyst Administrative Information Group, West Virginia University

Synergistic Activities

Recent External Funding

"Continuous Policy Compliance Auditing," as a sub-contractor to the University of Oklahoma as part of the US Department of Transportation's Intermodal Containerized Freight Security Project, $187,689 (2006-2010). Co-Principal Investigators: Bonnie Morris, Cynthia Tanner, and George Trapp (all at WVU)

"RTCE-Continuous Auditing," Morris, Bonnie (Co-Principal).Sponsored by Lockheed Martin Corporation IRAD Program, Private, $543,000. (2003-2004).

"National Biometric Security Project" Phase 1. 2004.

"EOUSA Automated Litigation Support Project" Sponsored by: Executive Office of the United States Attorney 2003.

Recent Published Refereed Journal Articles and Proceedings

Related Publications

Morris, B., Tanner, C., D'Alessandro, J. (2010). "Enabling Trust Through Continuous Compliance Assurance." In Shahram Latifi (Ed.), *International Conference on Information Technologies - New Generation.* Washington DC: IEEE Computer Society/CPS (IEEE Conference Publishing Services*)*

D'Alessandro, J., Tanner, C., Morris, B., Menzies, T. (2009). "Is Continuous Compliance Assurance Possible?" Published Abstract in Proceedings of the ITNG 2009, 6th International Conference on Information Technology : New Generations, Las Vegas, Conference, Technology, IEEE Computer Society.

Morris, B., Shaw, G., Tanner, C., Trapp, G. "Continuous Compliance Assurance for Trusted Information Sharing: A Research Framework," Fourteenth World Continuous Auditing and Reporting Symposium, Rutgers University, Newark, NJ. (November 3, 2007).

Other Selected Publications

Tanner, C. D., Roberts, B., Flash Cards an Object Oriented Approach to an Old Favorite, Presented at Killer Examples Workshop, OOPSLA 2008.

Hislop, Ellis, MacNeil, Tanner, Challenges in Teaching Software engineering Professionals Online, 33rd ASEE/IEEE Frontiers in Education Conference, November 5-8, 2003.F4G-1.

Ellis, H.J.C., Mead, N.R., Moreno, A., Tanner, C.D., Ramsey, D, Characteristics of Successful Collaborations to Produce Educated Software Engineering Professionals, Computer Science Education, Vol 12, Numbers 1-2, pp 119-140, 2002.

Honor Societies

Upsilon Pi Epsilon
Golden Key

Forrest Shull
Division Director, Measurement and Knowledge Management Division
Fraunhofer USA Inc.
Center for Experimental Software Engineering, Maryland
5825 University Research Court, Suite 1300, College Park, MD 20740
Phone: +1 240 487 2904 - Email: fshull@fc-md.umd.edu

Professional preparation

Loyola College in Maryland	Computer Science	B.S.	1994
University of Maryland, College Park	Computer Science	M.S.	1996
University of Maryland, College Park	Computer Science	Ph.D.	1998

Appointments

2008-present	Associate Adjunct Professor, Dept. of Comp. Sci., University of Maryland College Park
2006-present	Division Director, Fraunhofer USA Inc., Ctr for Experimental Software Engineering, Maryland
1999-2006	Scientist, Fraunhofer USA Inc., Center for Experimental Software Engineering, Maryland
1999	Faculty Research Associate, University of Maryland, College Park.

Publications

a) Most Closely Related to Proposed Project:

1. Shull., F., Feldmann, R., Seaman, C., Regardie, M., and Godfrey, S., “Fully Employing Software Inspections Data,” Innovations in Systems and Software Engineering - a NASA Journal, 2010. Available at: http://dx.doi.org/10.1007/s11334-010-0132-1
2. Seaman, C., Shull, F., Regardie, M., Elbert, D., Feldmann, R., Guo, Y., and Godfrey, S., “Defect Categorization: Making Use of a Decade of Widely Varying Historical Data,” Proc. International Symposium on Empirical Software Engineering and Measurement (ESEM). Kaiserslautern, Germany, Oct. 9-10, 2008.
3. Lindvall, M., Rus, I., Shull, F., Zelkowitz, M. V., Donzelli, P., Memon, A., Basili, V. R., Costa, P., Tvedt, R. T., Hochstein, L., Asgari, S., Ackermann, C., and Pech, D., “An Evolutionary Testbed for Software Technology Evaluation,” *Innovations in Systems and Software Engineering - a NASA Journal*, vol. 1, no. 1, pp. 3-11, 2005.
4. F. Shull, V. Basili, B. Boehm, A. W. Brown, P. Costa, M. Lindvall, D. Port, I. Rus, R. Tesoriero, and M. Zelkowitz, "What We Have Learned About Fighting Defects", In *Proceedings of the 8th International Software Metrics Symposium*, Ottawa, Canada, 2002, pp. 249-258.
5. V. Basili, F. Shull, and F. Lanubile. "Building Knowledge through Families of Experiments." *IEEE Transactions on Software Engineering*, 25(4): 456-473, July 1999.

b) Other Significant Publications

1. Carver, J., Juristo, N., Shull, F., and Vegas, S., “The Role of Replications in Empirical Software Engineering,” *Empirical Software Engineering: An International Journal*, vol. 13, no. 2, pp. 211-218, April 2008.
2. Gupta, A., Shull, F., Cruzes, D., Ackermann, C., Rønneberg, H., and Landre, E., “Experience Report on the Effect of Software Development Characteristics on Change Distribution,” Proc. Product Focused Software Process Improvement Conference (PROFES08). Rome, Italy, June 23-25, 2008.
3. Shull, F., Cruzes, D., Basili, V. R., and Mendonca, M., “Simulating Families of Studies to Build Confidence in Defect Hypotheses,” Information and Software Technology, vol. 47, no. 15, pp. 1019-1032, December 2005.
4. F. Shull, F. Lanubile, and V. Basili. "Investigating Reading Techniques for Object-Oriented Framework Learning." *IEEE Trans. on Software Engineering*, Vol. 26, No. 11, November 2000.
5. F. Shull, J. Carver, and G. H. Travassos. "An Empirical Methodology for Introducing Software Processes." In *Proceedings of Joint European Software Engineering Conference and Symposium on Foundations of Software Engineering (ESEC/FSE)*. Vienna, Austria, Sept. 10-14, 2001. p. 288-296.

Synergistic activities

- Editor in Chief, *IEEE Software*, beginning January 2011

- Program Co-Chair, Internat'l Symp. on Empirical Software Engineering and Measurement, 2011
- Program Chair: ISERN International Advanced School for Empirical Software Engineering (IASESE), September 2006.
- Program Chair of Experience Track, ICSE 2006
- Editorial Board Member, *Journal of Empirical Software Engineering*, Kluwer, 2002 - present

Collaborators and other affiliations

Collaborators & Co-Editors

- Kathleen Dangle, Madeline Diep, Raimund Feldmann, Lucas Layman, Mikael Lindvall, Myrna Regardie, Michele Shaw, Marv Zelkowitz: Fraunhofer Center – Maryland
- Muhammed Ali Babar, University of Copenhagen, Denmark
- Barry Boehm, University of Southern California
- Jeffrey Carver, University of Alabama
- Reidar Conradi, Norwegian University of Science and Technology, Norway
- Daniela Cruzes, Norwegian University of Science and Technology, Norway
- Christian Denger, Siemens, Germany
- Oscar Dieste, Universidad Politécnica de Madrid, Spain
- Torgeir Dingsøyr, Norwegian University of Science and Technology, Norway
- Tore Dyba, University of Oslo, Norway
- Hakan Erdogmus, Kalemun Research, Canada
- Sandra Fabbri, Federal University of Sao Carlos, Brazil
- Sally Godfrey, NASA Goddard Space Flight Center
- Anita Gupta, Cap Gemini, Norway
- Jo Hannay, Simula Research Labs, Norway
- Lorin Hochstein, ISI
- Jeff Hollingsworth, University of Maryland College Park
- Martin Höst, Lund University, Sweden
- Letizia Jaccheri, Norwegian University of Science and Technology, Norway
- Ross Jeffery, University of New South Wales, Australia
- Philip Johnson, University of Hawaii
- Natalia Juristo, Universidad Politécnica de Madrid, Spain
- Christin Lindholm, Lund University, Sweden
- Jose' Carlos Maldonado, Universidade de Sao Paulo, Brazil
- Walcelio Melo, Model Driven Solutions, USA
- Manoel Mendonca, Salvador University, Brazil
- Sandro Morasca, Università degli Studi dell'Insubria, Italy
- Ana Moreno, Universidad Politécnica de Madrid, Spain
- Taiga Nakamura, IBM, Japan
- Rafael Prikladnicki, Pontificia Universidade Catolica do Rio Grande do Sul, Brazil
- Lynn Reid, University of Chicago
- Dieter Rombach, Fraunhofer IESE, Germany
- Kurt Schneider, University of Hannover, Germany
- Carolyn Seaman, University of Maryland Baltimore County
- Janice Singer, National Research Council, Canada
- Dag Sjøberg, Simula Research Labs, Norway
- Guilherme Horta Travassos, Federal University of Rio de Janeiro, Brazil
- Burak Turhan, University of Oulu, Finland
- Rich Turner, Stevens Institute of Technology
- Sira Vegas, Universidad Politécnica de Madrid, Spain
- Christiane Gresse von Wangenheim, Federal University of Santa Catarina (UFSC)

Graduate and Postdoctoral Advisors

- Vic Basili, University of Maryland College Park

Madeline Diep
Research Scientist, Measurement and Knowledge Management Division
Fraunhofer USA Inc.
Center for Experimental Software Engineering, Maryland
5825 University Research Court, Suite 1300, College Park, MD 20740
Phone: +1 240 487 2904 - Email: fshull@fc-md.umd.edu

Professional preparation

University of Nebraska - Lincoln	Computer Science & Mathematics	B.S.	2001
University of Nebraska - Lincoln	Computer Science	M.S.	2004
University of Nebraska - Lincoln	Computer Science	Ph.D.	2009

Appointments

2009-present	Adjunct Professor, Dept. of Comp. Sci., University of Maryland, College Park
2009-present	Research Scientist, Fraunhofer USA Inc., Ctr for Experimental Software Engineering, Maryland
2005-2009	Graduate Research Assistant, University of Nebraska - Lincoln

Publications

1. V. Basili, M. Zelkowitz, L. Layman, K. Dangle, and M.Diep, *Obtaining Valid Safety Data for Software Safety Measurement and Process Improvement*, International Symposium on Empirical Software Engineering and Measurement, 2010.
2. M. Diep, M. Dwyer, and S. Elbaum, *Lattice-based Sampling for Path Property Monitoring*, Transactions on Software Engineering and Methodology, to appear.
3. M. Diep, S. Elbaum, and M. Dwyer, *Trace Normalization*, International Symposium of Software Reliability Engineering, pp. 67-76, November 2008.
4. M. Hardojo (Diep), M. Cohen, and S. Elbaum, *Probe Distribution Techniques to Profile Events in Deployed Software*, International Symposium of Software Reliability Engineering, pp. 395-406, November 2006.
5. S. Elbaum and M. Diep, *Profiling Deployed Software: Assessing Strategies and Testing Opportunities*, IEEE Transactions on Software Engineering, 31(4):312-327, April 2005.

Collaborators and other affiliations

Collaborators

- Forrest Shull, Raimund Feldmann, Lucas Layman, Vic Basili, Marv Zelkowitz, Kathleen Dangle: Fraunhofer Center – Maryland
- Hakan Erdogmus, Kalemun Research, Canada
- Sally Godfrey, NASA Goddard Space Flight Center
- Carolyn Seaman, University of Maryland Baltimore County
- Burak Turhan, University of Oulu, Finland
- Matthew Dwyer, University of Nebraska – Lincoln
- Myra Cohen, University of Nebraska - Lincoln

Graduate Advisor

- Sebastian Elbaum, University of Nebraska - Lincoln

Current and Pending Support

(See GPG Section II.C.2.h for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Timothy Menzies	Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title: Automatic Quality Assessment: Exploiting Knowledge of the Business Case

Source of Support: nsf
Total Award Amount: $ 350,000 Total Award Period Covered: 07/01/08 - 06/30/11
Location of Project: Morgantown WV
Person-Months Per Year Committed to the Project. Cal: 1.00 Acad: 1.00 Sumr: 1.00

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title: Better Comprehension of Software Engineering Data

Source of Support: NSF
Total Award Amount: $ 24,000 Total Award Period Covered: 08/15/10 - 08/14/13
Location of Project: Morgantown , West Virginia
Person-Months Per Year Committed to the Project. Cal: 0.00 Acad: 0.00 Sumr: 1.00

Support: ☐ Current ☒ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title: Empirical Software Engineering , Version 2.0

Source of Support: NSF
Total Award Amount: $ 523,222 Total Award Period Covered: 02/28/11 - 02/27/13
Location of Project: Morgantown , West Virginia
Person-Months Per Year Committed to the Project. Cal: 0.00 Acad: 0.00 Sumr: 1.00

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

Current and Pending Support

(See GPG Section II.C.2.h for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Bonnie Morris

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title: Information Fusion Networks For Intelligence and Security (InfoNets)

Source of Support: WVEPSCOR

Total Award Amount: $ 1,796,211 Total Award Period Covered: 01/01/07 - 01/01/12

Location of Project: Morgantown, West Virginia

Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 0.60

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

Current and Pending Support

(See GPG Section II.C.2.h for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Cecil Pollard	Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title: Comprehensive Diabetes Project ? Diabetes Prevention and Control Program

Source of Support: West Virginia Bureau for Public Health
Total Award Amount: $ 225,000 Total Award Period Covered: 01/01/10 - 12/31/10
Location of Project: Morgantown , West Virginia
Person-Months Per Year Committed to the Project. Cal:4.20 Acad: 0.00 Sumr: 0.00

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title: OHSR Cardiovascular Health

Source of Support: West Virginia Bureau for Public Health
Total Award Amount: $ 151,226 Total Award Period Covered: 01/01/10 - 12/31/10
Location of Project: Morgantown , West Virginia
Person-Months Per Year Committed to the Project. Cal:5.40 Acad: 0.00 Sumr: 0.00

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title: Asthma Education and Prevention Program

Source of Support: West Virginia Bureau for Public Health
Total Award Amount: $ 51,773 Total Award Period Covered: 01/01/10 - 12/31/10
Location of Project: Morgantown , West Virginia
Person-Months Per Year Committed to the Project. Cal:2.40 Acad: 0.00 Sumr: 0.00

Support: ☐ Current ☒ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title: Minnie Hamilton Health System Project Proposal

Source of Support: Minnie Hamilton Health System Project
Total Award Amount: $ 36,000 Total Award Period Covered: 12/31/10 - 12/31/13
Location of Project: Morgantown , West Virginia
Person-Months Per Year Committed to the Project. Cal:2.00 Acad: 0.00 Sumr: 0.00

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

Current and Pending Support

(See GPG Section II.C.2.h for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Cynthia Tanner

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title: Continuous Policy Compliance Auditing

Source of Support: sub-contractor to the University of Oklahoma, ICFS of DOT
Total Award Amount: $ 187,689 Total Award Period Covered: 06/12/06 - 08/15/11
Location of Project: WVU
Person-Months Per Year Committed to the Project. Cal:0.00 Acad: 0.00 Sumr: 1.00

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support
Project/Proposal Title:

Source of Support:
Total Award Amount: $ Total Award Period Covered:
Location of Project:
Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

 USE ADDITIONAL SHEETS AS NECESSARY

Forrest Shull

Pending Support

Project/Proposal Title: SHF:Large:Collaborative Research: The Price of Privacy
Source of Funding: NSF
Amount: $808,000
Dates: 6/1/11 to 5/31/15
Location of Project: Fraunhofer Center - Maryland
PI effort/year: 3.0 months

Project/Proposal Title: II-EN:Empirical Software Engineering, Version 2.0
Source of Funding: NSF
Amount: $750,000
Dates: 1/1/11 to 12/31/13
Location of Project: Fraunhofer Center - Maryland
PI effort/year: 1.0 months

Current Support

Project/Proposal Title: Measuring and Monitoring Technical Debt
Source of Funding: NSF
Amount: $464,538
Dates: 7/1/09 to 6/30/11
Location of Project: Fraunhofer Center - Maryland
PI effort/year: 1.0 months

Project/Proposal Title: Inspections for Systems and Software
Source of Funding: NASA
Amount: $410K
Dates: 10/1/08 to 9/30/11
Location of Project: Fraunhofer Center - Maryland
PI effort /year: 3 months

Project/Proposal Title: Professional Master of Engineering Courses: Software Engineering
Source of Funding: University of Maryland
Amount: $14K
Dates: 9/1/10 to 8/31/11
Location of Project: Fraunhofer Center - Maryland
PI effort /year: 0.5 months

Madeline Diep

Pending Support

Project/Proposal Title: SHF:Large:Collaborative Research: The Price of Privacy
Source of Funding: NSF
Amount: $808,000
Dates: 6/1/11 to 5/31/15
Location of Project: Fraunhofer Center - Maryland
PI effort/year: 3.0 months

Current Support

Project/Proposal Title: Inspections for Systems and Software
Source of Funding: NASA
Amount: $410K
Dates: 10/1/08 to 9/30/11
Location of Project: Fraunhofer Center - Maryland
PI effort /year: 3 months

Project/Proposal Title: Professional Master of Engineering Courses: Software Engineering
Source of Funding: University of Maryland
Amount: $14K
Dates: 9/1/10 to 8/31/11
Location of Project: Fraunhofer Center - Maryland
PI effort /year: 1.5 months

Project/Proposal Title: ESR Process Improvement
Source of Funding: ESR
Amount: $40K
Dates: 1/1/11 to 12/31/11
Location of Project: Fraunhofer Center - Maryland
PI effort /year: 2.5 months

FACILITIES, EQUIPMENT & OTHER RESOURCES

FACILITIES: Identify the facilities to be used at each performance site listed and, as appropriate, indicate their capacities, pertinent capabilities, relative proximity, and extent of availability to the project. Use "Other" to describe the facilities at any other performance sites listed and at sites for field studies. USE additional pages as necessary.

Laboratory: Dr. Menzies maintain research labs at West Virginia University College of Engineering and Mineral Resources. Each lab is equipped with cubicle office seating for about 10 undergraduate and/or graduate students, high performance workstations, laser printers and a large

Clinical:

Animal:

Computer: As described above, labs maintained by the investigators have a dozen high performance workstations.

Office: PI maintain offices in the Lane Department of Computer Science and Electrical Engineering, West Virginia University.

Other:

MAJOR EQUIPMENT: List the most important items available for this project and, as appropriate identifying the location and pertinent capabilities of each.

OTHER RESOURCES: Provide any information describing the other resources available for the project. Identify support services such as consultant, secretarial, machine shop, and electronics shop, and the extent to which they will be available for the project. Include an explanation of any consortium/contractual arrangements with other organizations.

FACILITIES, EQUIPMENT & OTHER RESOURCES

Continuation Page:

LABORATORY FACILITIES (continued):

Dr. Morris and Ms. Tanner maintain a research lab at West Virginia University in the College of Engineering and Mineral Resources which contains a distributed computing environment on an internal 100 Mbps network switch. The environment includes 8 high performance workstations, one equipped with six core processors and another equipped with a dual core processor. The distributed environment includes both a solid state disk drive and a TB 7200 RPM disk drive. This highly configurable set of computing resources will be used to create the simulated enclaves for the testbed studies. The lab also contains cubicle office seating for undergraduate and graduate students each with a high performance workstation. The lab is also equipped with a printer and two white boards.

Fraunhofer Center Maryland
Fraunhofer Center Maryland (FC-MD) is a not-for-profit organization that began operations in 1998 as the only Fraunhofer USA center to specialize in software and related engineering fields. FC-MD is affiliated with the Computer Science Department at the University of Maryland College Park. This affiliation gives it a unique insight into the latest research results and technologies in computer science. The Center grew out of 25 years of successful software research in collaboration with NASA Goddard resulting in numerous awards and thereby utilizes proven experimental approaches to introduce innovative techniques into industry.

FC-MD has been a partner in multiple research projects funded by the NSF, NASA, DoD, and commercial companies such as ABB, Boeing, DaimlerChrysler, Motorola, and Nokia, and has collaborated with universities such as the University of Southern California, Carnegie Mellon University, Massachusetts Institute of Technology, University of Washington, University of Alabama, and Mississippi State University.

FC-MD's mission is to be a nationally and internationally recognized center for software engineering competence and the preferred partner with industry in:

- Establishing software improvement programs,
- Transferring innovative technologies,
- Engaging in cooperative research,
- Performing technology studies, and
- Evaluating business processes and organizations.

FC-MD is collaborating with US and European companies as well as US and Maryland state government organizations.

FC-MD employs about 20 full and part time scientists and staff. Ensuring a close relationship between FC-MD and the University of Maryland:

- Executive Director Dr. Rance Cleaveland is also a professor at UMD;
- Senior Research Fellows Dr. Victor R. Basili and Dr. Marvin V. Zelkowitz are professors emeriti;
- Division Directors Dr. Forrest Shull and Dr. Mikael Lindvall are associate adjunct professors.

Equipment:
Fraunhofer Center Maryland (FC-MD) is located in College Park, Maryland, in the close vicinity of the University of Maryland campus and with access to university resources, e.g. libraries and laboratories.

FC-MD's facilities are located in M Square, the University of Maryland Research Park. Facilities include offices, computer server room, a secure room for protected data, and meeting rooms each equipped with computers and an electronic projector appropriate for conference calls. FC-MD has a secure wireless network linking Windows desktop and laptop computers. The wireless network uses a Cisco 1800 Secure Router with firewall and VPN, and has two Linksys and one Dell access points. The network is based on one primary domain controller with 200 Gigabyte of disk space and a full backup computer. It has two secondary domain controllers with 38 and 76 Gigabyte of disk space respectively. It has one e-mail server with 80 Gigabyte of disk space, one BlackBerry email server, and one Web server with 70 Gigabyte of disk space. In addition, it has one Experience Base Server for experience management with 70 Gigabyte of disk space. Information is backed up daily using disk and tape on a SureStore DAT24*6 Tape Switcher.

Library Resources:
FC-MD has, through the University of Maryland's library system, access to a wide range of Web-based information services including Web-based reference materials, and literature databases that include the IEEE and ACM digital computer science libraries.

List of all PIs, Co-PIs, Senior Personnel, paid Consultants, Collaborators and Postdocs to be involved in the project

Forrest Shull; Fraunhofer Center for Experimental Software Engineering, PI

Tim Menzies; West Virginia University; PI

Bonnie Morris ; West Virginia University; PI

Cynthia Tanner ; West Virginia University; PI

Cecil Pollard ; West Virginia University; PI