National Cancer Institute

Portfolio Analysis Partnerships

IC Collaborations to Customize Tools for Portfolio Analysis

National Cancer Institute

NCI and CIT: Creating a Knowledge Management Toolkit

NCI: Lisa Krueger, Michele Vos, Maria Bukowski
CIT: Calvin Johnson, William Lau, Krishna Collie

National Cancer Institute

NCI-CIT Partnership as described in the trans-NIH Collaborations Report:

Knowledge Management tool-kit to facilitate research portfolio analysis

This project is a collaboration between NCI and CIT to develop tools to assist with portfolio analysis. The overarching goal is to create opportunities for users to interact with sophisticated information retrieval tools, supported by visualization software to steer the machine learning process as well as refine the classification output. Success will conserve users' time as well as maximize the efficient integration of users' judgment and expertise. The focus is on the evaluation and optimization of currently available tools and resources.

National Cancer Institute

The Knowledge Management and Special Projects (KMSP) Branch in the NCI/OD/CSSI

- Serves as the lead for the NCI on NIH scientific reporting efforts that utilize IT-assisted approaches (also called knowledge management systems and tools).
- Participates in collaborative relationships across organizational lines and between ICs to further knowledge management efforts within the NCI and throughout the NIH.
- KMSP Branch participants in this workshop:
  - Presenters: Lisa Krueger and Michele Vos
  - Discussant: Maria Bukowski

National Cancer Institute

NCI-CIT Use Cases for Discussion

Use Case 1: An ensemble classification system for intramural research categorization and decision support

Use Case 2: One-sided Classifier training

Use Case 3: Matching NIH grants to ClinicalTrials.gov protocols

NIAID-CIT Partnership
Text Mining in Portfolio Analysis

Marie Parker, Dolan Ghosh-Das, Peter Choi, NIAID/OSPIDA
Calvin Johnson, William Lau, Krishna Collie, CIT

NIAID's Office of Strategic Planning, Initiative Development and Analysis(OSPIDA)

- **Strategic Planning and Evaluation:** Create Institute-wide plans for carrying out the NIAID mission and achieving NIAID research priorities; Support and guide evaluation of NIAID programs.
- **Initiative Development:** Manage planning, design, development, review, and quality control of NIAID research grant and contract initiatives.
- **Program Analysis and Reporting**: Classify scientific content of NIAID-funded projects and provide periodic and *ad hoc reports* to senior leadership, Program staff, US Congress and others.

Why Collaborate?

- NIAID already has numerous processes, tools and information resources for portfolio analysis and reporting.

- However, categorization and reporting efforts can be difficult and labor-intensive when:
  - Portfolios are initially being defined
  - Portfolios are constantly changing
  - Existing categorization methods or coding systems do not accurately capture the science area
  - Extensive manual review of projects is required for large portfolios

Goal of NIAID-CIT Collaboration

Explore and develop novel portfolio building approaches to facilitate the creation and maintenance of portfolios of research projects for further analysis and comparison.

NIAID Use Cases for Discussion

- Use Case 4: Mapping NIAID/DAIDS projects to science priorities and objectives.

- Use Case 5: Update NIAID research portfolio, “B-Cell Mediated Vaccines for HIV.”

PORTFOLIO ANALYSIS PARTNERSHIPS

IC Collaborations to Customize Tools for Portfolio Analysis

Calvin A. Johnson, William Lau, and Krishna Collie

CIT
Center for Information Technology

DCB

Division of Computational Bioscience
Center for Information Technology | National Institutes of Health

HPCIO

CIT DIVISION OF COMPUTATIONAL BIOSCIENCE

Division of Computational Bioscience

Center for Information Technology | National Institutes of Health

HPCIO

The Division of Computational Bioscience (DCB) of CIT is a research and development organization that provides scientific and technical expertise in applying state-of-the-art technologies to support the NIH Intramural Research Program (IRP). Working with NIH's Institutes and Centers (ICs), DCB develops leading-edge computational methods and tools to solve complex biomedical laboratory and clinical research problems.

The High Performance Computing and Informatics Office responds to critical demands to develop capabilities in new areas, including:

- Portfolio analysis
- Text mining and analytics
- Machine learning
- Knowledge-based systems
- High-end computing
- Scientific visualization

TOOLS FOR PORTFOLIO ANALYSIS

- **Machine Learning** – Algorithm that learns from experience with respect to some task, based on some performance measure.
- **Clustering** – Unsupervised learning
- **Classification** – Supervised learning
- **Information Retrieval** – Can be supervised or unsupervised. Often involves ranking
- Measuring Similarity between documents or corpora
- Natural language processing
- Decision support systems
- Visualization, exploratory analysis, dimensionality reduction

ASSERTIONS AND ASSUMPTIONS

- Machines should not supplant expert knowledge but rather should assist and amplify experts.
- Separation of responsibility: informaticians should not annotate
- Experts are good at making decisions, although perhaps not quickly.
- Machines are good at handling large amounts of data quickly and consistently but need to be told what to do (i.e., machines are fast but stupid).
- Effective paradigm – tools to assist subject matter experts.
- There exists a tradeoff between flexibility and ease of use.

WHAT TO EXPECT

- Define your problem: what questions are you trying to answer?
- Do the data match the problem/question?
  - Can exploratory analyses help understand the data and refine the problem?
- How would you like to categorize your data?
- Can you identify examples of each category (both positive and negative)?
- What are inclusion or exclusion criteria for the query?
- Do you need to expand your dataset (e.g., literature, citations)?

ITERATIVE REFINEMENT

- Exploratory data analysis (clustering, visualization)
- Define categories as well as examples of each category (positive and negative)
- Perform query of candidates based on inclusion/exclusion criteria
- Train initial model from set of exemplars and unknowns
- Repeat until performance is satisfactory:
  - Perform validation (measure recall and precision) on a subset of the examples.
  - Deploy classifier for prediction/detection/retrieval task
  - Analyze results of retrieval task
    - (Optional) Clustering, anomoly detection, visualization
  - Annotation (identify additional exemplars and/or negatives)
  - Retrain model if necessary.

METHODOLOGY CONSIDERATIONS

- Possible feature-space elements
  - Scientific thesaurus concepts
  - OAR codes
  - Problem-dependent measures of fit
- Training regimes
  - Binary classifier
  - Multi-class classifier
  - “One sided” classifier
  - Negatively weighted one-sided classifier
  - Ensemble of classifiers
  - Semi-supervised clustering
  - Outlier detection

National Cancer Institute

Use Case 1: An Ensemble Classification System for Intramural Research Categorization and Decision Support

Rationale

- NCI intramural scientists and staff categorize the intramural research portfolio.
- Scientific categorization of research portfolios can be expedited by utilization of decision-support tools.
- Most NIH tools require that users have permission to access the IMPAC II database. IMPACII is an extramural database not accessible to NIH intramural staff.
- To support staff efforts, support vector machine (SVM) ensemble classifiers developed by CIT are being optimized for use by NCI staff.

National Cancer Institute

Use Case 1: Methods and Results

- The SVM classifiers were trained using exemplars from several cancer categories. Training included true negatives as well as true positives.
- CIT further customized the classification tool for NCI staff by including functionality that allows the user to select from multiple thesaurus options, including the NCI Metathesaurus.
- Recall and Precision for the LibSVM ensemble in eleven cancer categories was calculated by CIT after the initial training. Recall weighted average: 77% and Precision weighted average: 87%
- We presented progress to intramural staff at 2011 NCI retreat and made tool available for staff to test.

USE CASE 1: DECISION SUPPORT

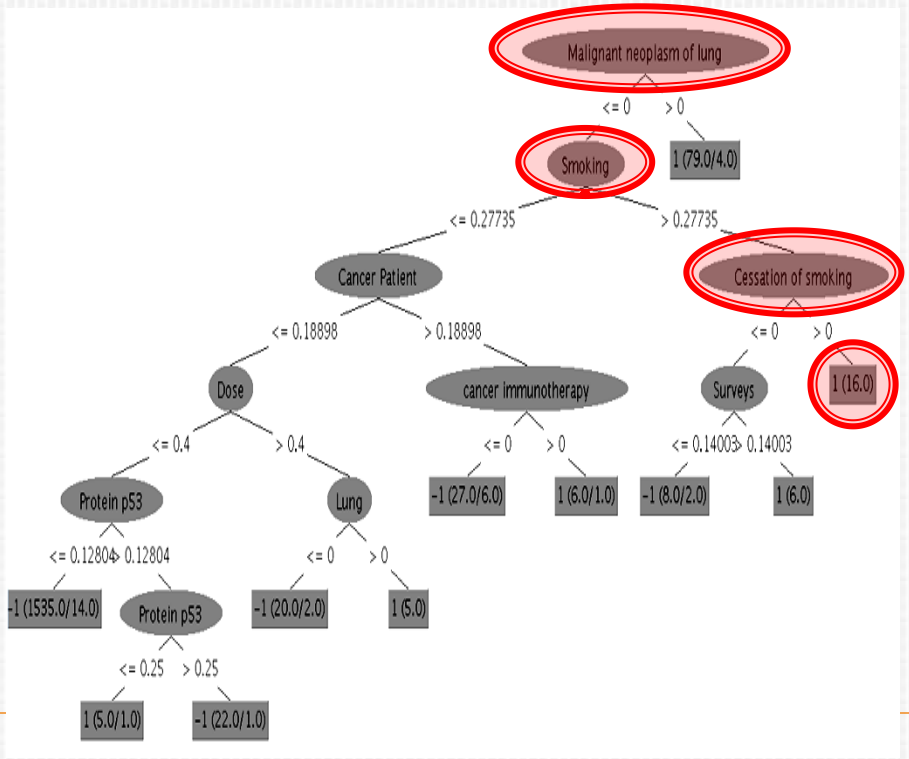
- A decision tree is trained to simulate the reasoning of the primary classifier
- The goal is to provide evidence for the machine classification

Title:
Project Title: Proactive Smoking Cessation for Adolescents

Abstract:
DESCRIPTION (provided by applicant): The overall aim of this renewal application is to complete the Hutchinson Study of High School Smoking, a group-randomized trial in adolescent smoking cessation. This randomized trial is motivated by (1) the current unacceptably high smoking prevalence among youth, and its serious public health consequences, (2) the importance of identifying via rigorous intervention trials effective programs for helping youth who smoke to quit, and (3) the considerable potential of telephone counseling using motivational interviewing to be effective with youth. Now nearly two-thirds complete, this ongoing, 2-arm trial includes as participants 2,886 high school seniors (all the smokers and a sample of nonsmokers, identified via their baseline survey responses) from 50 Washington high schools. The adolescent smoking cessation intervention, implemented to participants in experimental high schools during their senior year, is a proactive, individually tailored telephone counseling intervention that incorporates both Motivational Interviewing and cognitive behavioral techniques. Participants are followed to endpoint, approximately 6 months post-high school, to assess the intervention's impact on cessation status, number of quit attempts, change in readiness to quit, and reduction in frequency and level of smoking. Major activities in years 06-07 covered by this application include completion of tracking and outcome data collection, and statistical analyses and reporting of results. It is clear from previous studies that a majority of teen smokers want to quit and try to do so, but with little success. The primary goal of this randomized trial is to develop and evaluate an innovative smoking cessation intervention to help teens succeed in quitting. A positive finding would have significant implications for reducing youth smoking and, ultimately, for improving the nation's health.

Class: Lung Cancer

Explanation: Even though *Malignant neoplasm of lung* does not appear since *Smoking* and *Cessation of smoking* appear frequently, this document is classified as Lung Cancer.



National Cancer Institute

Use Case 1: Next Steps

- Create new training and test datasets in order to update existing categories in the classification tool and for new categories in the tool.
- Explore the possibility of training a classifier with only ‘true positives’ in order to streamline the creation of training and test datasets
- Add ability to batch load by unique project IDs.

National Cancer Institute

Use Case 1: Discussion

- High recall classifiers may be helpful in identifying potential false negatives.
- Once concordance is confirmed on a larger dataset, these tools may be useful as part of annual QC process.
- Since training can be done using intramural exemplars, these tools may be useful for performing targeted analyses of the intramural portfolio.

National Cancer Institute

Use Case 2: Training of the 1-Sided Classifier

Rationale:

- Traditional supervised classification systems require training of the classifier using ‘positive’ and ‘negative’ examples
- In many situations, such as text classification, ‘negative’ examples may be unavailable or hard to determine
- The use of one-sided classification techniques will be explored as a solution to overcoming situations when it is difficult to compile training datasets of ‘negative’ examples

National Cancer Institute

Use Case 2: Methods

KMSP provided CIT with:

- A project list of “Exemplars” for breast (1200 Applids) and lung cancer (515 Appl_ids) for training the classifier
- All projects had an assigned percent relevance that ranged from 20 – 100% for each cancer type
- All exemplars are FY2010 funded grants
- The training datasets are comprised of the following activity codes: F’s, K’s, R’s, SC’s and U01’s

National Cancer Institute

Use Case 2: Results

1-Sided Classifier results included extramural grants for fiscal years 2006 – 2010

DATA OUTPUT FOR THE CIT 1-SIDED CLASSIFIER

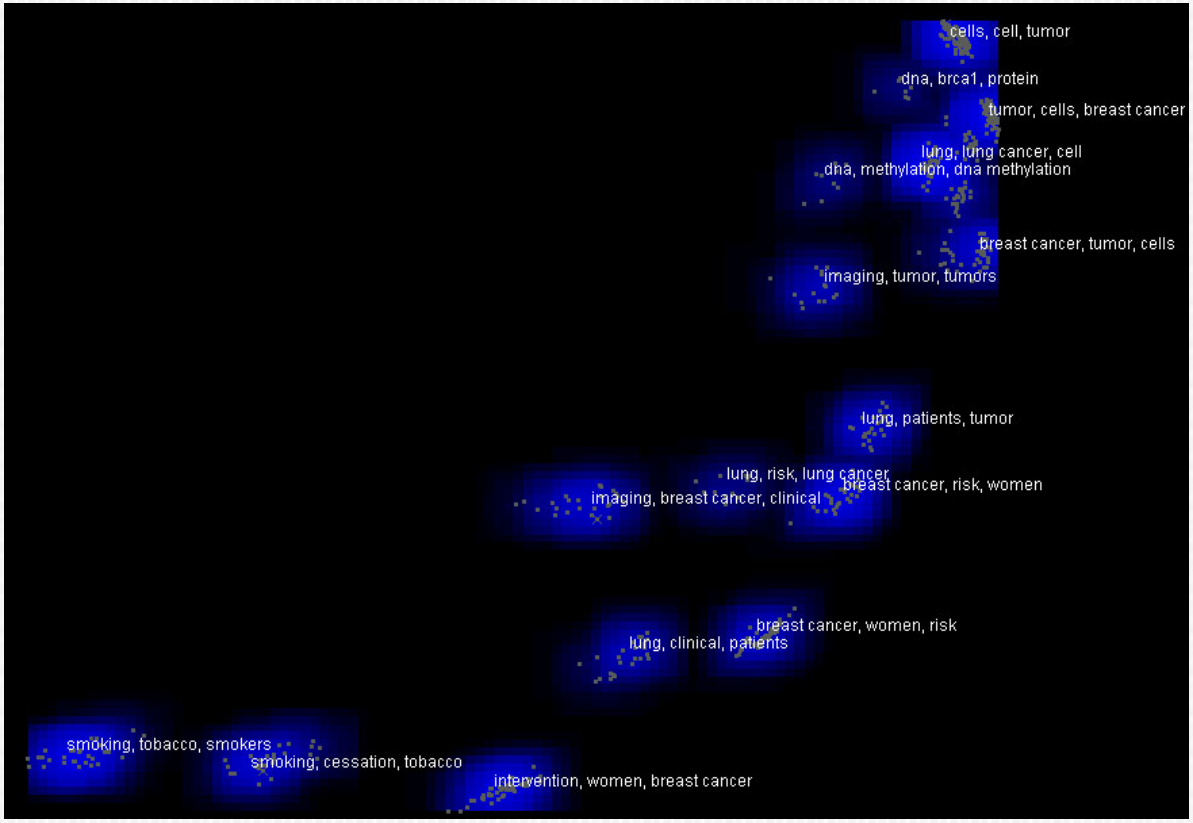
FY	BREAST CANCER	LUNG CANCER
2010	2552	697
2009	3185	771
2008	2567	901
2007	2427	1198
2006	1767	1036

Preliminary analysis of the 2010 BCA classifier results revealed that the trained classifier is able to identify 'True Positives' with a range of project relevancies

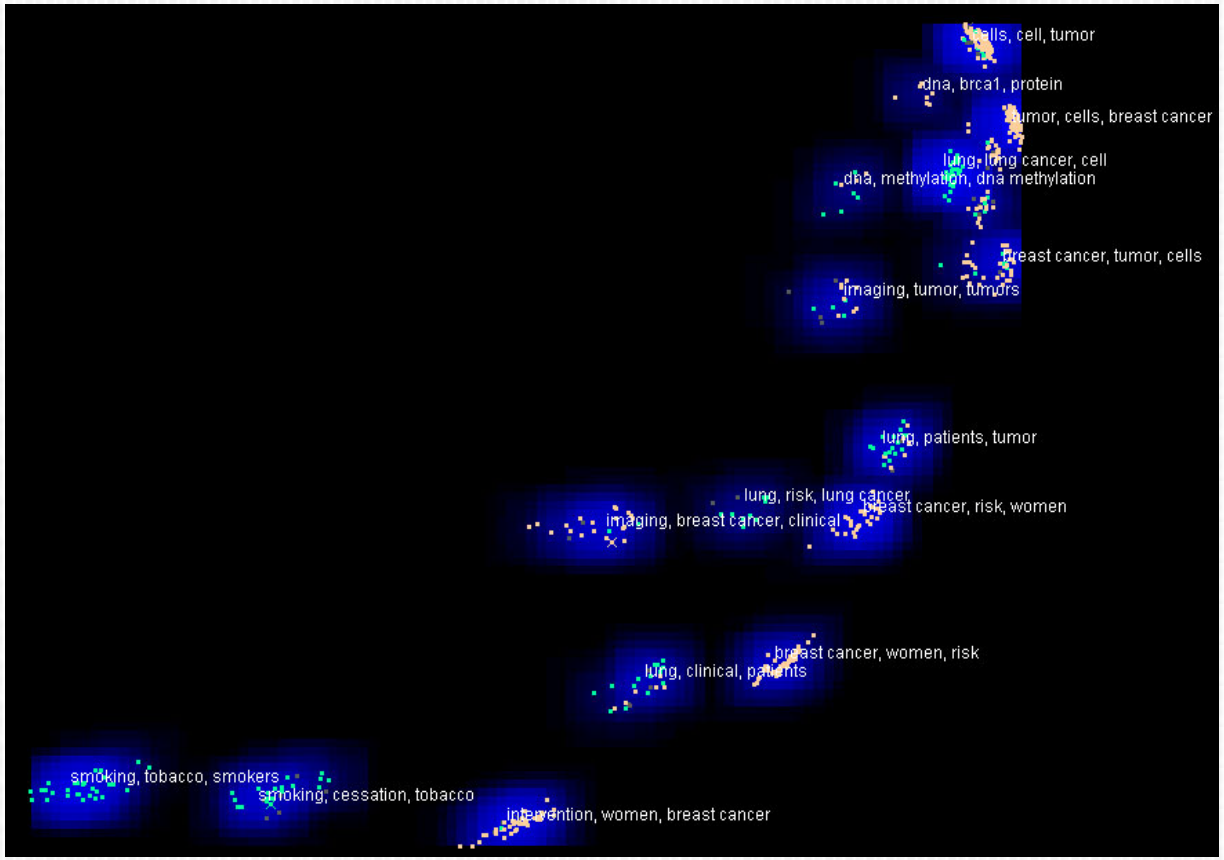
CONFIDENCE SCORES 1-SIDED CLASSIFIER	PROJECT PERCENT RELEVANCE RANGE
0.500 - 0.625	<= 20 - 100%
0.630 - 0.750	33 - 100%
0.760 - 0.875	<= 20 - 100%
0.883 - 1.00	<= 20 - 100%

USE CASE 2: RESULTS

Can we separate the lung cancer grants from the breast cancer grants in this clustering of 567 documents?



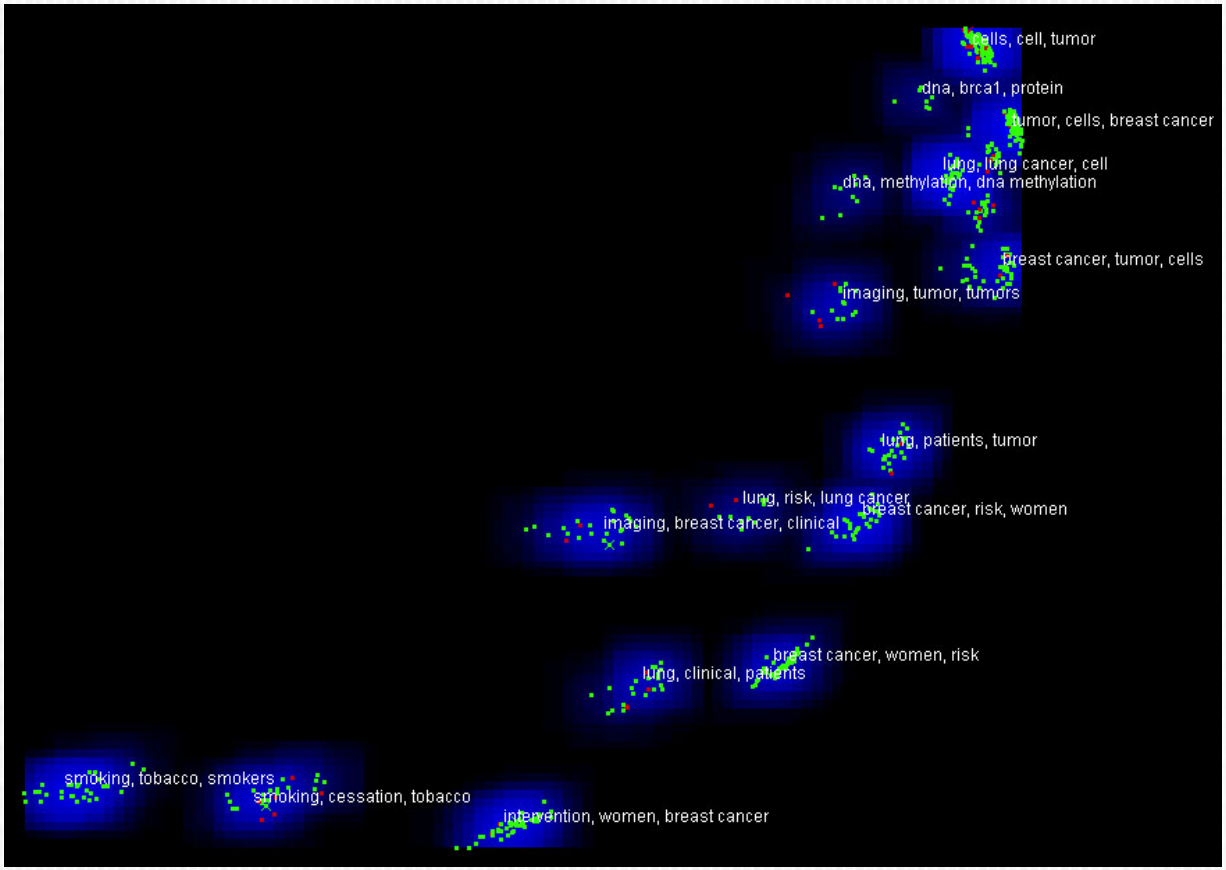
USE CASE 2: CLUSTERING VS. CLASSIFICATION



A classifier was trained to identify lung and breast cancer documents using a set of exemplars.

Lung cancer – Aqua
Breast cancer - Peach

USE CASE 2: COMPARING MACHINE TO EXPERT ASSIGNMENT



Over 90% agreement between the machine classification and expert annotations.

Agree – Green
Disagree - Red

National Cancer Institute

Use Case 2: Next Steps

- Validation of the fiscal year project lists returned by the trained classifier
    - o use Carrot^2, the CIT customized clustering engine to facilitate project validation

- Provide feedback to CIT concerning the validation results and how they can be used to further refine the one-sided classifier

National Cancer Institute

Use Case 3: Matching ClinicalTrials.gov Protocols to NIH grants

Rationale

ClinicalTrials.gov is a database of federally and privately supported clinical studies that can be used to evaluate investments in clinical research.

- The ClinicalTrials.gov database: In a sampling of NIH-sponsored clinical trials, the majority of the CT.gov records were not associated with a NIH grant number.

- The NIH IMPAC II database:
  - No direct association of CT.gov Identifiers with grant, contract or intramural records
  - Contains a set of population tracking tables containing “IC_protocol_IDs” that may be associated with the ‘Study IDs’ contained in CT.gov

National Cancer Institute

Use Case 3: Methods

Phase I: Devised a system to automatically match the IMPAC II projects with CT.gov records. A pilot was performed testing the ability of a text mining tool to match clinical trials with a NIH grant number.

Phase II: CIT continues to improve the system's ability to match records using text mining. NCI is investigating associations between certain data elements in both databases to supplement matching potential and facilitate data validation.

- Data pulled from CT.gov with the filters 'NIH-funded' and keyword=Cancer.
- We are working with only a subset of the data (761 records).

ClinicalTrials.gov
A service of the U.S. National Institutes of Health
Developed by the National Library of Medicine

NCT_IDs
associated with the following data fields

PI_NAME	3701
SECONDARY_ID	5203
RP_NAME	945
LOC_NAME	13403
PROTOCOL_TITLE	3454
ORG_STUDY_ID	3454
LOC_ORG	2817
PI_ORG	3700
RP_ORG	944

→ NATIONAL CANCER INSTITUTE **KMSP**

↓ Org_study_id, secondary_id

IMPAC II
National Institutes of Health

↓

POP_TRACKING_MV

↓ POP_APPL_ID

PROJECTS_VERSIONS_T

↓

Identifies the NCT_ID_NUM
that is associated with POP_Track_Protocol
ID and PROJ_NUM

↓

Data Validation

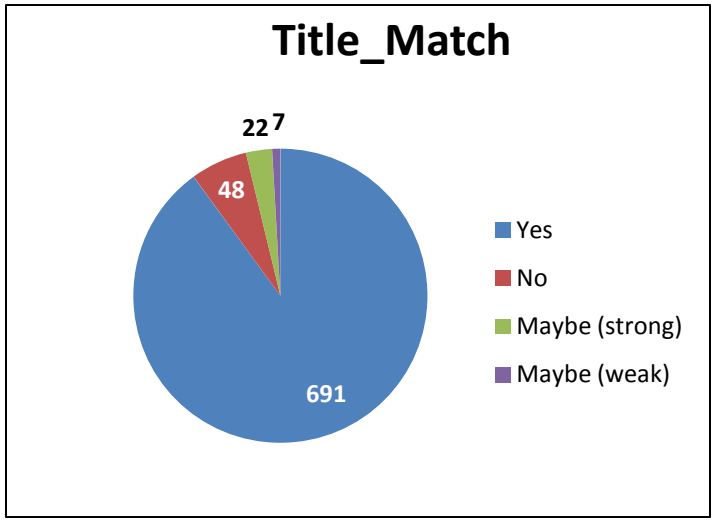
Allows us to develop best practices for CT.gov and
IMPACII Data mapping and matching

National Cancer Institute

Use Case 3: Results: Title matching of CT.gov official title and protocol title

Title Match Methods and Results

Human annotation	SQL code string match	Number of records
Yes	CT_POP_EXACT	408
Yes	CT_POP_EXACT_25%	59
Yes	CT_POP_EXACT_50%	113
Yes	CT_POP_SOUNDEX	60
Yes	CT_PROJ_EXACT	2
Yes	CT_PROJ_SOUNDEX	1
Yes	NO MATCH	48
No	CT_POP_SOUNDEX	13
No	NO MATCH	35
Maybe (strong)	CT_POP_SOUNDEX	2
Maybe (strong)	NO MATCH	20
Maybe (weak)	CT_POP_SOUNDEX	1
Maybe (weak)	NO MATCH	6



Title Match?		Details
Yes	691	47 = N01; 16 = P01; 25= P50; 15= R01; 2 = R21; 15 = U01; 319 = U10; 249 = Z; 3 = U19
No	48	20 = N01; 12 = P's; 6 = U01; 2 = R01; 8 = Z01
Maybe (strong)	22	7 = N01; 1 = P01; 6 = P50; 8 = Z01
Maybe (weak)	7	1 = P50; 6 = Z01

USE CASE 3: CONCEPT

ClinicalTrials.gov
A service of the U.S. National Institutes of Health

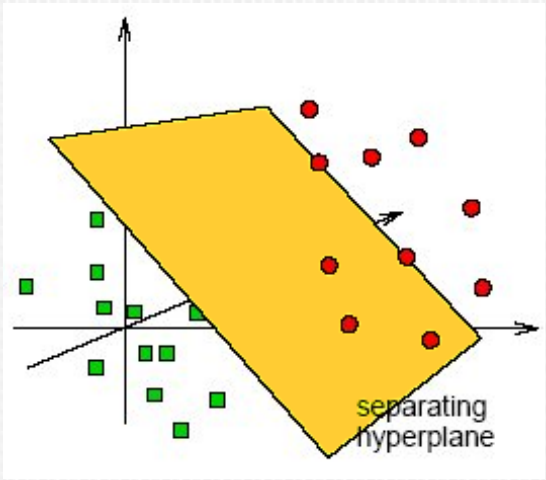
IMPAC II
National Institutes of Health
> Information for Management, Planning, Analysis, and Coordination

Clinical Trials

Search IMPACII for Project

Grants

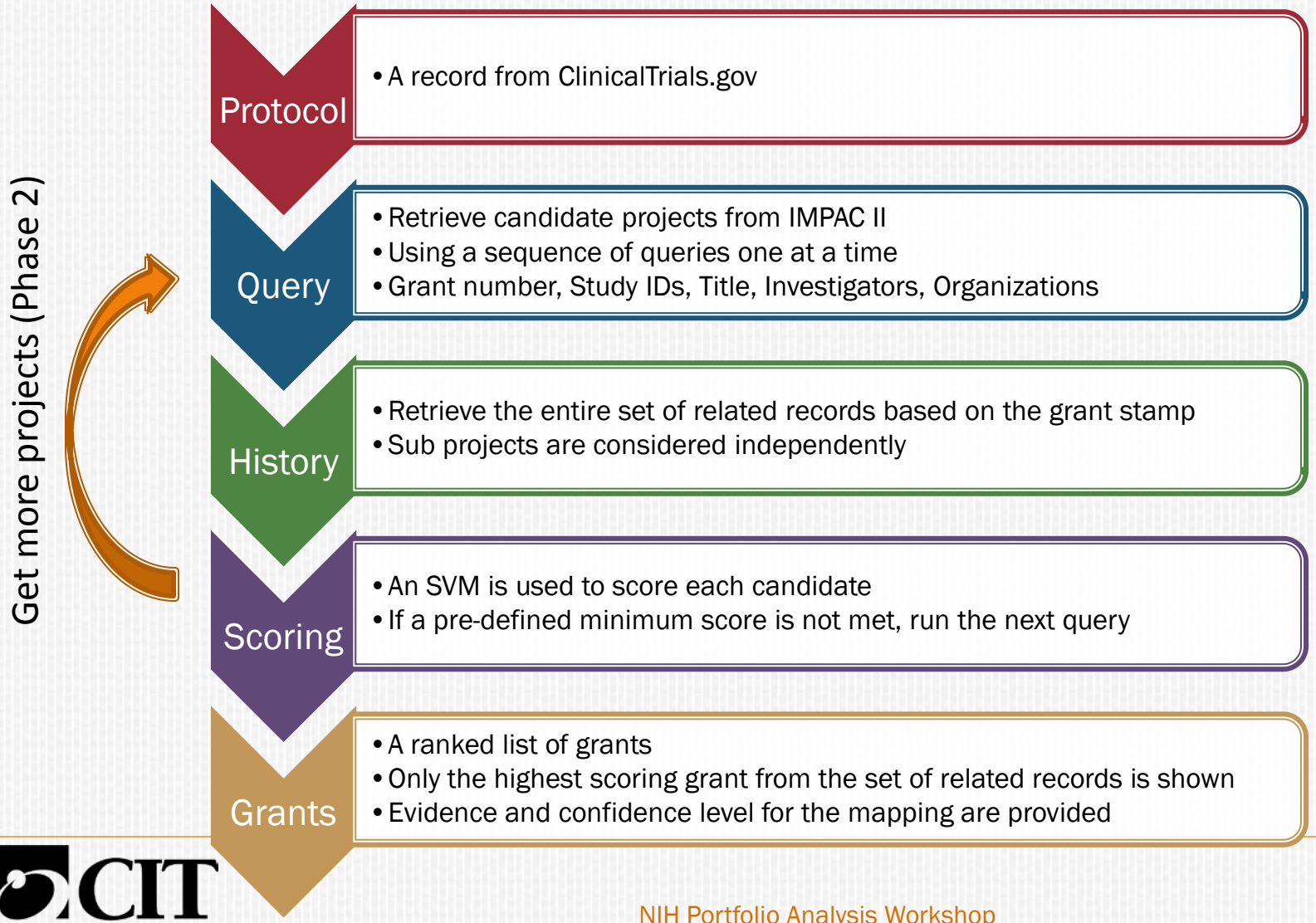
Candidates Found



Ranked IMPAC II projects returned

Candidates scored by a classifier

USE CASE 3: WORKFLOW



USE CASE 3: RANKING STRATEGIES

- Phase 1 features:
  - Edit distance between the protocol and project titles
  - Similarity between the concept fingerprints
  - Min. edit distance between organization names
  - RCDC score of the project for the Clinical Trials category
- Potential phase 2 features:
  - Number of investigator name match
  - Amount of overlap between protocol and project study period
  - Whether the grant number is mentioned in the protocol

USE CASE 3: RESULTS

- In an evaluation study of 30 protocols, 13 matched grants were confirmed to be the actual funding source, with additional 13 (mostly P30s and U10s) having similar scientific contents but would require further review to confirm linkage.
- Among those highly scored mappings (n=8), only one protocol has the grant number listed.
- The tool can be used to automatically recover (with high confidence) the missing linkage between the clinical trials and their funding for those easy-to-match protocols.

National Cancer Institute

Use Case 3: Next Steps

Phase II

- Continue data validation on expanded dataset
- Investigate matching other data elements such as PI and Organization
- Improve systems ability to match records using text mining

Phase III:

- Expansion of the project in order to test and validate the mapping on other ICs protocols in ClinicalTrials.gov
- With additional NIH collaborators, create use cases of trans-NIH interest

Use Case 4: Mapping DAIDS Projects

Goal: Identify projects within NIAID's Division of AIDS (DAIDS) that support it's long-range science priorities and objectives.

Use Case 4: Mapping DAIDS Projects Process outline

1. NIAID/OSPIDA provides CIT with
   - DAIDS scientific priorities and objectives
   - List of FY 10 DAIDS projects with associated NIAID assigned scientific codes
   - A set of exemplar projects for each objective
2. CIT devises algorithm/search method and maps projects to each objective.
3. OSPIDA reviews results and annotates projects.
4. CIT refines results based on annotations and works towards establishing an automated system that minimizes manual review.

Use Case 4: Mapping DAIDS Projects Results

- First round of mapping completed by CIT and annotated by OSPIDA.
- Results vary for different objectives and range in accuracy from 28%-98%. For 5/9 objectives >80% projects were correctly mapped to the objective.
- Algorithm is currently being refined by CIT to improve accuracy.

USE CASE 4: MAPPING NIAID/DAIDS PROJECTS TO SCIENCE PRIORITIES AND OBJECTIVES.

- Recall validation experiment. Compare feature space: NIAID scientific codes vs. biomedical thesaurus. Leave-one-out.

Objective	Positive Exemplars	RPAB recall	Thesaurus recall
1	28	0.96	0.89
2	12	0.42	0.33
3	21	0.95	0.91
4	9	0.89	0.44
5	26	0.89	0.89
6	9	0.89	0.67
7	22	0.91	0.82
8	8	1.00	0.50
9	6	0.33	0.00

USE CASE 4: MAPPING NIAID/DAIDS PROJECTS TO SCIENCE PRIORITIES AND OBJECTIVES

- Prediction experiment. NIAID scientific codes used as feature space.

Obj.	Exemplars	Strong Detections	Total Detections	Total New	False Detections	Ambiguous
1	28	88	161	133	23	0
3	21	29	100	71	2	0
4	9	11	37	28	6	0
5	26	101	165	139	0	3
6	9	12	20	11	0	5
7	22	46	91	69	3	2
8	8	8	32	24	23	0
9	6	7	32	26	6	6

USE CASE 4: MAPPING NIAID/DAIDS PROJECTS TO SCIENCE PRIORITIES AND OBJECTIVES

- NIAID scientific codes used as feature space. Precision determined from ((total new) – (false + ambiguous))/(total new)

Obj.	Total Detections	Total New Detections	False + Ambiguous	Precision	Recall	F-score
1	161	133	23	0.83	0.96	0.89
3	100	71	2	0.97	0.95	0.96
4	37	28	6	0.79	0.89	0.84
5	165	139	3	0.98	0.89	0.93
6	20	11	5	0.55	0.89	0.68
7	91	69	5	0.93	0.91	0.92
8	32	24	23	0.04	1.00	0.08
9	32	26	12	0.54	0.33	0.41

Use Case 5:
Update NIAID research portfolio

Goal: Update NIAID research portfolio "B-cell mediated vaccines for HIV." This portfolio was created by manual review of projects.

Use Case 5: Update NIAID research portfolio Process outline

1. OSPIDA provides CIT with
   - FY 2008 exemplars (positives and negatives) identified by Subject Matter Experts (SMEs).
   - Business rules for inclusion/exclusion.
2. CIT provides success rate for validation process for FY 2008 projects using various algorithms.
3. SMEs review results and annotate projects.
4. CIT refines results based on annotations and extends algorithm to capture projects from other Fiscal Years.

USE CASE 5: UPDATE NIAID RESEARCH PORTFOLIO, "B-CELL MEDIATED VACCINES FOR HIV."

- Develop classifier from annotated training and validation set (2008):
  - 119 exemplars (strong positives)
  - 370 relevant projects but not exemplars
  - 94 non-relevant but close projects
  - 384 total non-relevant projects, including the close but non-relevant projects
- Compare binary SVM model training with training SVM on positives and unknowns including negatively weighted unknowns
- Train with positive exemplars, test on other relevant.

USE CASE 5: UPDATE NIAID RESEARCH PORTFOLIO, "B-CELL MEDIATED VACCINES FOR HIV."

- HIV B-CELL Vaccines Classifier Results
- Test conducted on portfolio of 5986 grants from 2008.

Model	Threshold	Recall	Precision	F-score	Detections
Binary	0.0	1.00	0.65	0.78	2348
1.5-sided	0.4	0.42	0.98	0.59	122
1.5-sided	0.2	0.48	0.98	0.64	140
1.5-sided	0	0.49	0.97	0.65	146
1.5-sided	-1.0	0.80	0.83	0.82	279
1.5-sided	-1.1	0.83	0.82	0.84	658

- Annotation of the detected project list is in progress.

Use Case 5: Update NIAID research portfolio Next Steps

OSPIDA is currently reviewing projects and CIT will use that information to improve accuracy.

Questions?

- This presentation can be found at http://dcb.cit.nih.gov/PortfolioPartnership.pdf