SRI International

Machine-Learning Based Classification of Research Grant Award Records

Jeffrey Alexander, Ph.D.

Christina Freyman, Ph.D.

Center for Science, Technology & Economic Development

John Byrnes, Ph.D.

SRI Advanced Analytics

5th Annual Global TechMining Conference

16 September 2015

Acknowledgements

- This research was supported under contract no. NSFDACS09C1350 issued by the National Center for Science & Engineering Statistics of the National Science Foundation

- The authors gratefully acknowledge the substantial intellectual guidance and contributions of Jeri Mulrow and Darius Singpurwalla of NCSES, and Patrick Lambe of Straits Knowledge, Singapore

Project Provenance

- NCSES is the Federal Statistical Agency tasked with data collection & analysis regarding the U.S. science & engineering enterprise
- Among its surveys, NCSES administers the annual Survey of Federal Funds for Research & Development
- Classifying R&D by "Field of Science & Engineering" (FOSE) is very problematic for many agencies

H. R. 5116—26

SEC. 505. NATIONAL CENTER FOR SCIENCE AND ENGINEERING STATISTICS.

(a) ESTABLISHMENT .—There is established within the Foundation a National Center for Science and Engineering Statistics that shall serve as a central Federal clearinghouse for the collection, interpretation, analysis, and dissemination of objective data on science, engineering, technology, and research and development.

(b) DUTIES .—In carrying out subsection (a) of this section, the Director, acting through the Center shall—

(1) collect, acquire, analyze, report, and disseminate statistical data related to the science and engineering enterprise in the United States and other nations that is relevant and useful to practitioners, researchers, policymakers, and the public, including statistical data on—

(A) research and development trends;

(B) the science and engineering workforce;

(C) United States competitiveness in science, engineering, technology, and research and development;

and

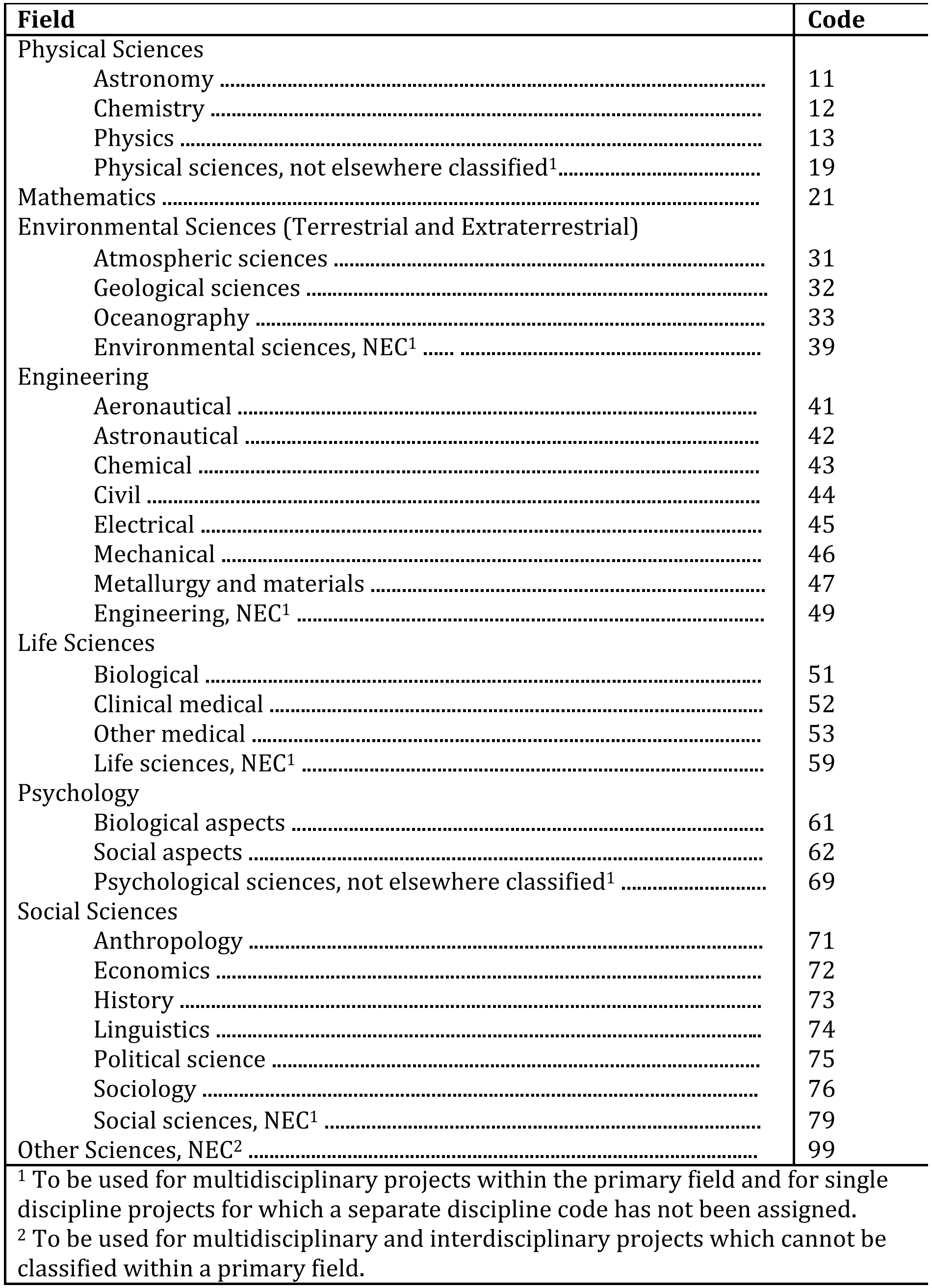
(D) the condition and progress of United States STEM education;

(2) support research using the data it collects, and on methodologies in areas related to the work of the Center; and

(3) support the education and training of researchers in the use of large-scale, nationally representative data sets.

Project Provenance

- FOSE is mandated by OMB Directive 16, issued in 1978
- Directive 16 has never been revised or reissued
- Most agencies do not use FOSE as an internal classification
- Therefore, reporting by FOSE is done using labor-intensive, bespoke processes that may be inconsistent across agencies and time periods



Field	**Code**
Physical Sciences	
Astronomy	11
Chemistry	12
Physics	13
Physical sciences, not elsewhere classified[1]	19
Mathematics	21
Environmental Sciences (Terrestrial and Extraterrestrial)	
Atmospheric sciences	31
Geological sciences	32
Oceanography	33
Environmental sciences, NEC[1]	39
Engineering	
Aeronautical	41
Astronautical	42
Chemical	43
Civil	44
Electrical	45
Mechanical	46
Metallurgy and materials	47
Engineering, NEC[1]	49
Life Sciences	
Biological	51
Clinical medical	52
Other medical	53
Life sciences, NEC[1]	59
Psychology	
Biological aspects	61
Social aspects	62
Psychological sciences, not elsewhere classified[1]	69
Social Sciences	
Anthropology	71
Economics	72
History	73
Linguistics	74
Political science	75
Sociology	76
Social sciences, NEC[1]	79
Other Sciences, NEC[2]	99

[1] To be used for multidisciplinary projects within the primary field and for single discipline projects for which a separate discipline code has not been assigned.

[2] To be used for multidisciplinary and interdisciplinary projects which cannot be classified within a primary field.

Project Provenance

- Is there a better way to collect these data that will:
  - Reduce respondent burden
  - Use existing administrative data records
  - Be implemented more consistently
  - Reflect to some extent changes in the nature of fields
  - Connect FOSE to outcomes of interest to policy

- Thus, two objectives:
  - Improve classification using FOSE
  - Explore potential to create a replacement for FOSE (motivated by 2010 CNSTAT report recommendation on "tagging")

Data on Federal Research and Development Investments

A PATHWAY TO MODERNIZATION

Proposed Solution: Machine Classification based on Abstracts

- Key constraints of *statistical* data collection
  - Consistency
  - Comparability
  - Maintain time series

- Weaknesses of current classification methods
  - Classifying by organization/mission
  - Classifying by performer's discipline
  - Keyword searching
  - SME-driven classification

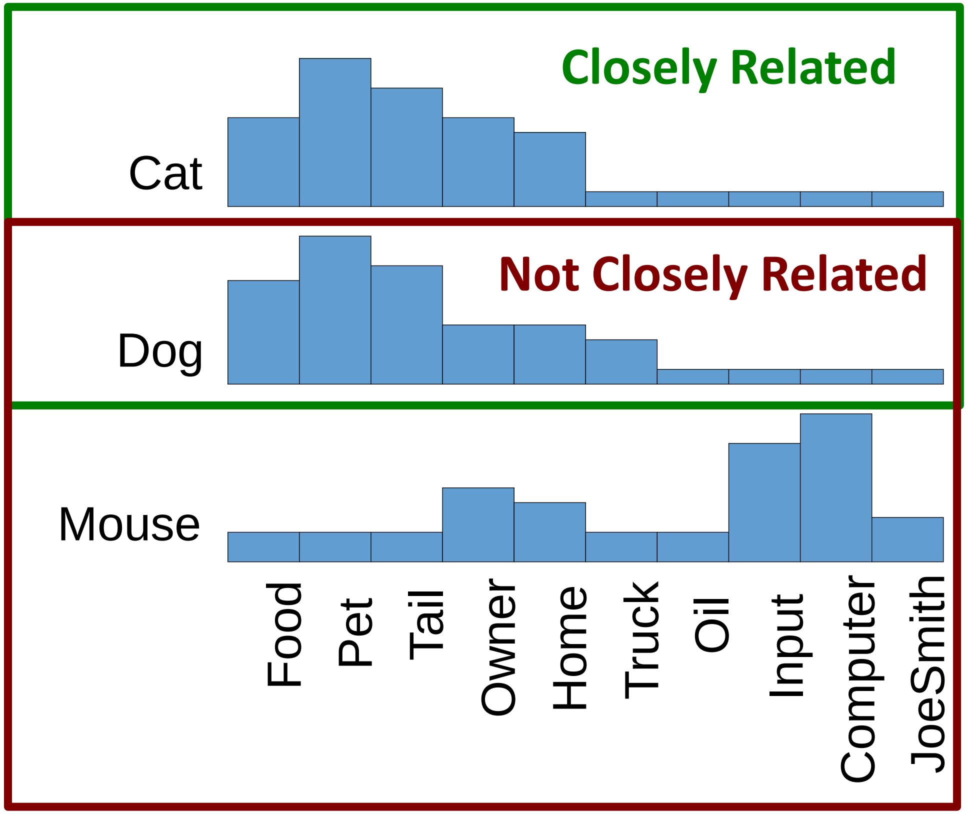
- Objective: a useful means of portfolio analysis

Machine-learning via text analytics: alternatives

- Topic modeling (Blei et al., 2003)
  - Statistical assignment of document clusters to topics derived from document text
  - Polyhierarchical (same document appears in multiple topics)
- Topic co-clustering (Ilgen & Rowher, 1998)
  - "Forces" documents into a single topic cluster
  - Documents and terms are assigned to topics separately
- Common complaints
  - Variable outputs
  - Topics not human-interpretable
  - No clear alignment with external taxonomies

Association-Grounded Semantics: Concept

- The basic tenet: the *meaning* of a data object is based on the *associations* in which the object participates.
- Histograms count the number of times each of a given set of keywords was found within in some fixed proximity to the target words.
- Similarity of meaning is captured by similarity of probability distribution (information geometric divergence measures – Kullback-Lieber divergence).
  - K-L is used to determine how similar two papers are.



$$\breve{x} \equiv P(Y \mid X = x)$$

$$\mathrm{KL}[x_1, x_2] = \sum_k P(y_k \mid x_1) \log \frac{P(y_k \mid x_1)}{P(y_k \mid x_2)}$$

For more information see: Brynes and Rohwer. “Test Modeling for Real-Time Document Categorization.” IEEEAC paper #1375. 2004

Association-Grounded Semantics: Process

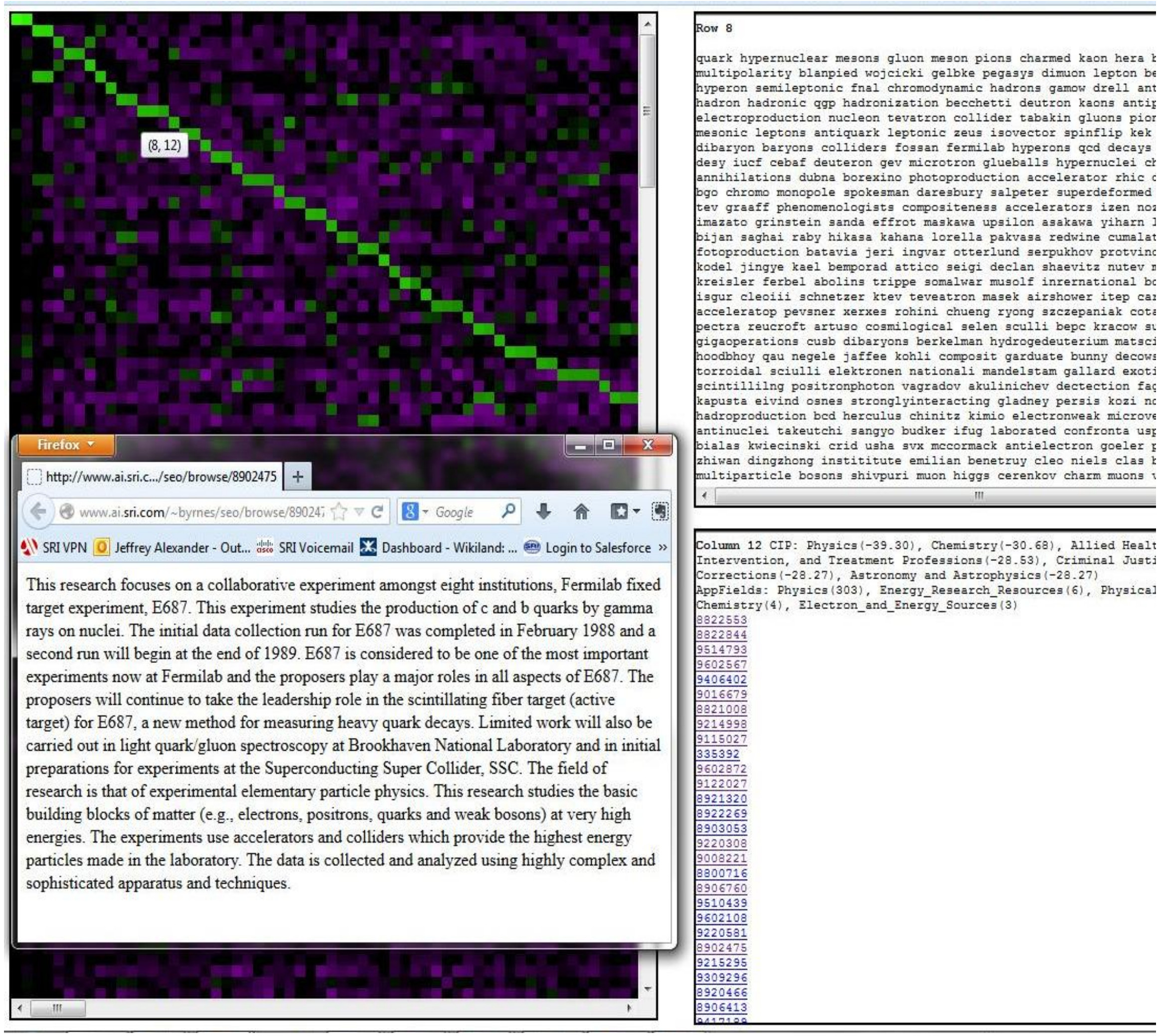
- Select an external taxonomy with standardized classifications

- For each term, build a “language model” with an existing corpus that describes each classification and its associations

- Use topic co-clustering to place abstracts in a single topic cluster

- Measure the statistical similarity between terms found in clusters in the test set and terms found in each language model

AGM: Illustration


(8, 12)
Firefox
http://www.ai.sri.c.../seo/browse/8902475
+
www.ai.sri.com/~byrnes/seo/browse/890247
Google
SRI VPN
Jeffrey Alexander - Out...
SRI Voicemail
Dashboard - Wikiland: ...
Login to Salesforce
»

This research focuses on a collaborative experiment amongst eight institutions, Fermilab fixed target experiment, E687. This experiment studies the production of c and b quarks by gamma rays on nuclei. The initial data collection run for E687 was completed in February 1988 and a second run will begin at the end of 1989. E687 is considered to be one of the most important experiments now at Fermilab and the proposers play a major roles in all aspects of E687. The proposers will continue to take the leadership role in the scintillating fiber target (active target) for E687, a new method for measuring heavy quark decays. Limited work will also be carried out in light quark/gluon spectroscopy at Brookhaven National Laboratory and in initial preparations for experiments at the Superconducting Super Collider, SSC. The field of research is that of experimental elementary particle physics. This research studies the basic building blocks of matter (e.g., electrons, positrons, quarks and weak bosons) at very high energies. The experiments use accelerators and colliders which provide the highest energy particles made in the laboratory. The data is collected and analyzed using highly complex and sophisticated apparatus and techniques.

```
Row 8

quark hypernuclear mesons gluon meson pions charmed kaon hera b
multipolarity blanpied wojcicki gelbke pegasys dimuon lepton be
hyperon semileptonic fnal chromodynamic hadrons gamow drell ant
hadron hadronic qgp hadronization becchetti deutron kaons antip
electroproduction nucleon tevatron collider tabakin gluons pion
mesonic leptons antiquark leptonic zeus isovector spinflip kek
dibaryon baryons colliders fossan fermilab hyperons qcd decays
desy iucf cebaf deuteron gev microtron glueballs hypernuclei ch
annihilations dubna borexino photoproduction accelerator rhic c
bgo chromo monopole spokesman daresbury salpeter superdeformed
tev graaff phenomenologists compositeness accelerators izen noz
imazato grinstein sanda effrot maskawa upsilon asakawa yiharn l
bijan saghai raby hikasa kahana lorella pakvasa redwine cumalat
fotoproduction batavia jeri ingvar otterlund serpukhov protvino
kodel jingye kael bemporad attico seigi declan shaevitz nutev m
kreisler ferbel abolins trippe somalwar musolf inrernational bo
isgur cleoiii schnetzer ktev teveatron masek airshower itep car
acceleratop pevsner xerxes rohini chueng ryong szczepaniak cota
pectra reucroft artuso cosmilogical selen sculli bepc kracow su
gigaoperations cusb dibaryons berkelman hydrogedeuterium matsci
hoodbhoy qau negele jaffee kohli composit garduate bunny decows
torroidal sciulli elektronen nationali mandelstam gallard exoti
scintillilng positronphoton vagradov akulinichev dectection fag
kapusta eivind osnes stronglyinteracting gladney persis kozi no
hadroproduction bcd herculus chinitz kimio electronweak microve
antinuclei takeutchi sangyo budker ifug laborated confronta usp
bialas kwiecinski crid usha svx mccormack antielectron goeler p
zhiwan dingzhong instititute emilian benetruy cleo niels clas b
multiparticle bosons shivpuri muon higgs cerenkov charm muons v
```

```
Column 12 CIP: Physics(-39.30), Chemistry(-30.68), Allied Healt
Intervention, and Treatment Professions(-28.53), Criminal Justi
Corrections(-28.27), Astronomy and Astrophysics(-28.27)
AppFields: Physics(303), Energy_Research_Resources(6), Physical
Chemistry(4), Electron_and_Energy_Sources(3)
8822553
8822844
9514793
9602567
9406402
9016679
8821008
9214998
9115027
335392
9602872
9122027
8921320
8922269
8903053
9220308
9008221
8800716
8906760
9510439
9602108
9220581
8902475
9215295
9309296
8920466
8906413
```

- For each cluster (column), brightest square represents topic (row) with the best “fit”
- Topics are associated with ***one or more*** disciplines based on statistical similarity
- Can look at source project abstract to validate
  - Abstract is labeled as physics (even though the term “physics” appears only once in the abstract)
- Ability to process thousands of documents simultaneously

Test Set: Abstracts of Awarded Grants from NSF

- Public abstract database as of January 2014
  - Total of over 500,000 abstracts

- Extract awards for which we can establish some form of "ground truth"
  - External validation that machine classification seems "accurate"

- Run AGM routine to measure pointwise mutual information between terms in external taxonomy & terms in abstracts
  - Clusters of abstracts are derived from calculating a probability distribution over term clusters (topics)
  - Use Hellinger divergence metric to identify CIP term most closely related to a given abstract cluster

Machine Classification for Two Facets

	Classification by Scientific Discipline	**Classification by Socio-Economic Objective (SEO)**
External taxonomy	Classification of Instructional Programs (NCES)	Nomenclature for Analysis & Comparison of Scientific Programmes & Budgets (OECD) + Australia-New Zealand Standard Research Classification SEO facet
Validation term set	NSF funding organization	Field of application (subset)
Data set with validation	278,000 awards	143,536 awards
Key caveats	• Combine awarding division & program to derive discipline • CIP is an instructional classification, not a research classification	• Field of application terms are NOT standardized, and usage is inconsistent across awards • SEO termsets were very sparse

Validation Metrics

- Recall
  - Machine classification matches what is found in the validation data
  - Measures ability of method to produce true positives

- Precision
  - Machine classification matches ONLY what was found in the validation data
  - Measures ability of method to avoid producing false positives

Sample Set: Classification by Discipline

Program	Number of awards
Algebra, Number Theory	2897
Archaeology	1917
Marine Geology and Geophysics	2363
Plant Genome Research Project	451
Political Science	1309
Social Psychology	558
Elementary Particle Accel User	517
Synthesis	373
Total	**10385**

Program	Most relevant term from CIP
Algebra, Number Theory	Statistics; Mathematics
Archaeology	Archeology
Marine Geology And Geophysics	Marine Sciences; Geological and Earth Sciences/Geosciences
Plant Genome Research Project	Plant Sciences; Genetics
Political Science	Political Science and Government
Social Psychology	Research and Experimental Psychology
Elementary Particle Accelerator User	Physics
Synthesis	Chemistry

Results: Classification by Discipline

Program	Precision	Recall
Algebra, Number Theory	100%	99%
Archaeology	100%	97%
Marine Geology and Geophysics	99%	95%
Plant Genome Research Project	98%	88%
Political Science	99%	73%
Social Psychology	94%	72%
Elementary Particle Accelerator User	98%	89%
Synthesis	93%	85%

Sample Set: Classification by SEO

Application field	Number of awards
Agriculture	1284
Climate related activities	709
Law	248
Health	3346
Total	**5587**

Field of Application	SEO-based category mapped to

Agriculture	Forestry; Horticultural Crops;
	Summer Grains and Oilseeds;
	Winter Grains and Oilseeds;
	Harvesting and Packing of Plant Products;
	Environmentally Sustainable Plant Production

Climate related activities	Climate and Climate Change;
	Renewable Energy;
	Air Quality;
	Energy Conservation and Efficiency;
	Preparation and Production of Energy Sources

Law	Government and Politics; Justice and the Law

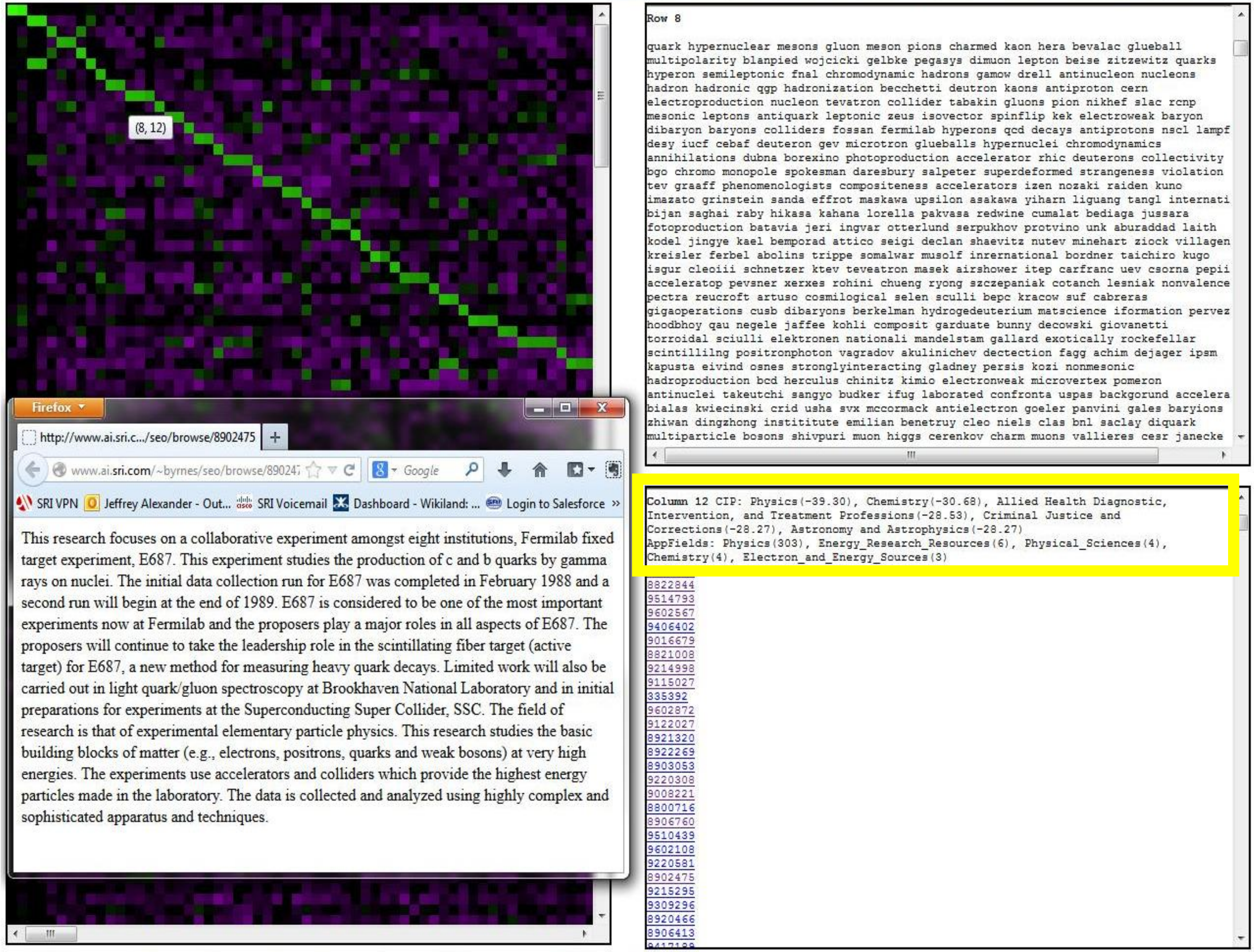
Health	Clinical Health; Health and Support Services;
	Public Health;

Results: Classification by SEO

Field of Application	Precision	Recall
Agriculture	37%	90%
Climate related activities	77%	93%
Law	96%	90%
Health	79%	52%

Summary of Findings—Classification by Discipline

- Machine learning approach performed well in classifying abstracts by discipline
  - CIP provides a rich language model for disciplines
  - Note that we had to be selective in use CIP terms, as it includes terms for training and not research (e.g., computer technician)
  - Interest in ability to show multiple disciplines associated with a given set of abstracts
    - Can be used as a measure of interdisciplinarity
    - Can also highlight unusual combinations of disciplines, which MAY be indicative of potentially transformative research


(8, 12)
Firefox
http://www.ai.sri.c.../seo/browse/8902475
www.ai.sri.com/~byrnes/seo/browse/890247
Google
SRI VPN
Jeffrey Alexander - Out...
SRI Voicemail
Dashboard - Wikiland: ...
Login to Salesforce

This research focuses on a collaborative experiment amongst eight institutions, Fermilab fixed target experiment, E687. This experiment studies the production of c and b quarks by gamma rays on nuclei. The initial data collection run for E687 was completed in February 1988 and a second run will begin at the end of 1989. E687 is considered to be one of the most important experiments now at Fermilab and the proposers play a major roles in all aspects of E687. The proposers will continue to take the leadership role in the scintillating fiber target (active target) for E687, a new method for measuring heavy quark decays. Limited work will also be carried out in light quark/gluon spectroscopy at Brookhaven National Laboratory and in initial preparations for experiments at the Superconducting Super Collider, SSC. The field of research is that of experimental elementary particle physics. This research studies the basic building blocks of matter (e.g., electrons, positrons, quarks and weak bosons) at very high energies. The experiments use accelerators and colliders which provide the highest energy particles made in the laboratory. The data is collected and analyzed using highly complex and sophisticated apparatus and techniques.

```
Row 8

quark hypernuclear mesons gluon meson pions charmed kaon hera bevalac glueball
multipolarity blanpied wojcicki gelbke pegasys dimuon lepton beise zitzewitz quarks
hyperon semileptonic fnal chromodynamic hadrons gamow drell antinucleon nucleons
hadron hadronic qgp hadronization becchetti deutron kaons antiproton cern
electroproduction nucleon tevatron collider tabakin gluons pion nikhef slac rcnp
mesonic leptons antiquark leptonic zeus isovector spinflip kek electroweak baryon
dibaryon baryons colliders fossan fermilab hyperons qcd decays antiprotons nscl lampf
desy iucf cebaf deuteron gev microtron glueballs hypernuclei chromodynamics
annihilations dubna borexino photoproduction accelerator rhic deuterons collectivity
bgo chromo monopole spokesman daresbury salpeter superdeformed strangeness violation
tev graaff phenomenologists compositeness accelerators izen nozaki raiden kuno
imazato grinstein sanda effrot maskawa upsilon asakawa yiharn liguang tangl internati
bijan saghai raby hikasa kahana lorella pakvasa redwine cumalat bediaga jussara
fotoproduction batavia jeri ingvar otterlund serpukhov protvino unk aburaddad laith
kodel jingye kael bemporad attico seigi declan shaevitz nutev minehart ziock villagen
kreisler ferbel abolins trippe somalwar musolf inrernational bordner taichiro kugo
isgur cleoiii schnetzer ktev teveatron masek airshower itep carfranc uev csorna pepii
acceleratop pevsner xerxes rohini chueng ryong szczepaniak cotanch lesniak nonvalence
pectra reucroft artuso cosmilogical selen sculli bepc kracow suf cabreras
gigaoperations cusb dibaryons berkelman hydrogedeuterium matscience iformation pervez
hoodbhoy qau negele jaffee kohli composit garduate bunny decowski giovanetti
torroidal sciulli elektronen nationali mandelstam gallard exotically rockefellar
scintillilng positronphoton vagradov akulinichev dectection fagg achim dejager ipsm
kapusta eivind osnes stronglyinteracting gladney persis kozi nonmesonic
hadroproduction bcd herculus chinitz kimio electronweak microvertex pomeron
antinuclei takeutchi sangyo budker ifug laborated confronta uspas backgorund accelera
bialas kwiecinski crid usha svx mccormack antielectron goeler panvini gales baryions
zhiwan dingzhong instititute emilian benetruy cleo niels clas bnl saclay diquark
multiparticle bosons shivpuri muon higgs cerenkov charm muons vallieres cesr janecke
```

```
Column 12 CIP: Physics(-39.30), Chemistry(-30.68), Allied Health Diagnostic,
Intervention, and Treatment Professions(-28.53), Criminal Justice and
Corrections(-28.27), Astronomy and Astrophysics(-28.27)
AppFields: Physics(303), Energy_Research_Resources(6), Physical_Sciences(4),
Chemistry(4), Electron_and_Energy_Sources(3)
8822844
9514793
9602567
9406402
9016679
8821008
9214998
9115027
335392
9602872
9122027
8921320
8922269
8903053
9220308
9008221
8800716
8906760
9510439
9602108
9220581
8902475
9215295
9309296
8920466
8906413
```

Summary of Findings: Classification by SEO

- Machine learning performance was fairly poor in classifying abstracts by SEO
  - Poor quality of language models—too sparse and non-specific
  - 'Field of application' terms (validation termset) aligns poorly with SEO terms
  - 'Field of application' labeled by NSF program officer, perhaps arbitrarily
  - May reflect difficulty in associating SEO (broader impact) with topics in fundamental research

- Results may be much better if we used expert judgment for validation, rather than metadata

Caveats and Future Research

- Classification by discipline may be especially effective due to the nature of the NSF research portfolio
  - Primarily funds academic research, which is organized by discipline
  - Predominantly funds more fundamental science, which is rooted strongly in specific disciplines relative to more applied research

Division	Machine-classified disciplinary term	Percentage of org's awards
Division of Chemistry		
	Chemistry	49%
	Physics	19%
	Biochemistry, Biophysics and Molecular Biology	18%
Division of Electrical, Communications and Cyber Systems		
	Electrical, Electronics and Communications Engineering	18%
	Materials Sciences	16%
	Physics	11%
Division of Environmental Biology		
	Ecology, Evolution, Systematics, and Population Biology	56%
	Plant Sciences	14%
	Genetics	9%
Division of Experimental & Integrative Activities		
	Computer Software and Media Applications	22%
	Computer Engineering	13%
	Computer Systems Analysis	11%
	Rehabilitation and Therapeutic Professions	6%
Division of Polar Programs		
	Atmospheric Sciences and Meteorology	31%
	Geological and Earth Sciences/Geosciences	24%
	Ecology, Evolution, Systematics, and Population Biology	12%
Division of Information & Intelligent Systems		
	Computer Software and Media Applications	23%
	Rehabilitation and Therapeutic Professions	12%
	Health and Medical Administrative Services	9%
Division of Ocean Sciences		
	Geological and Earth Sciences/Geosciences	37%
	Atmospheric Sciences and Meteorology	16%
	Ecology, Evolution, Systematics, and Population Biology	14%
Division of Integrative Organismal Systems		
	Genetics	32%
	Neurobiology and Neurosciences	29%
	Zoology/Animal Biology	14%
Emerging Frontiers		
	Ecology, Evolution, Systematics, and Population Biology	32%
	Genetics	21%
	Museology/Museum Studies	15%

Caveats and Future Research

- Planning to run similar experiment on NASA abstracts
  - Projects are much more applied, interdisciplinary
  - Access to appropriate admin data records will be crucial

- Describing of a “Classification of R&D Activities” system and toolkit
  - How to use machine-generated tags as part of an integrated classification system with multiple facets
  - E.g., disciplines, related technologies, SEOs, character of work, application areas

SRI International

Thank You

Headquarters
333 Ravenswood Avenue
Menlo Park, CA 94025
+1.650.859.2000

Additional U.S. and
international locations

www.sri.com