MEDICAL DATA MINING

Timothy Hays, PhD

Health IT Strategy Executive
Dynamics Research Corporation (DRC)

December 13, 2012

Healthcare in America

Is a VERY Large Domain with
Enormous Opportunities for Data Mining

- US Healthcare (2009)
  - $2.5 Trillion
  - 17.3% of GDP
- Healthcare system:
  - Providers, Payers and Patients,
  - Government (Federal and State) and Private/Commercial,
  - Research to (Best) Practice,
  - Regulations, Laws, and Policies (i.e. Affordable Care Act, etc.)

Healthcare / Medical Data Mining

- Patients and Consumers
- Providers: Government, Private or Commercial, hospitals, pharmacies, clinics, doctors' offices, and other provider services
- Payers: employers, insurance carriers, other third-party payers, health plan sponsors (employers, unions, DOD, VA, HHS, etc.)

Areas where data mining can help!

Healthcare / Medical Data Mining

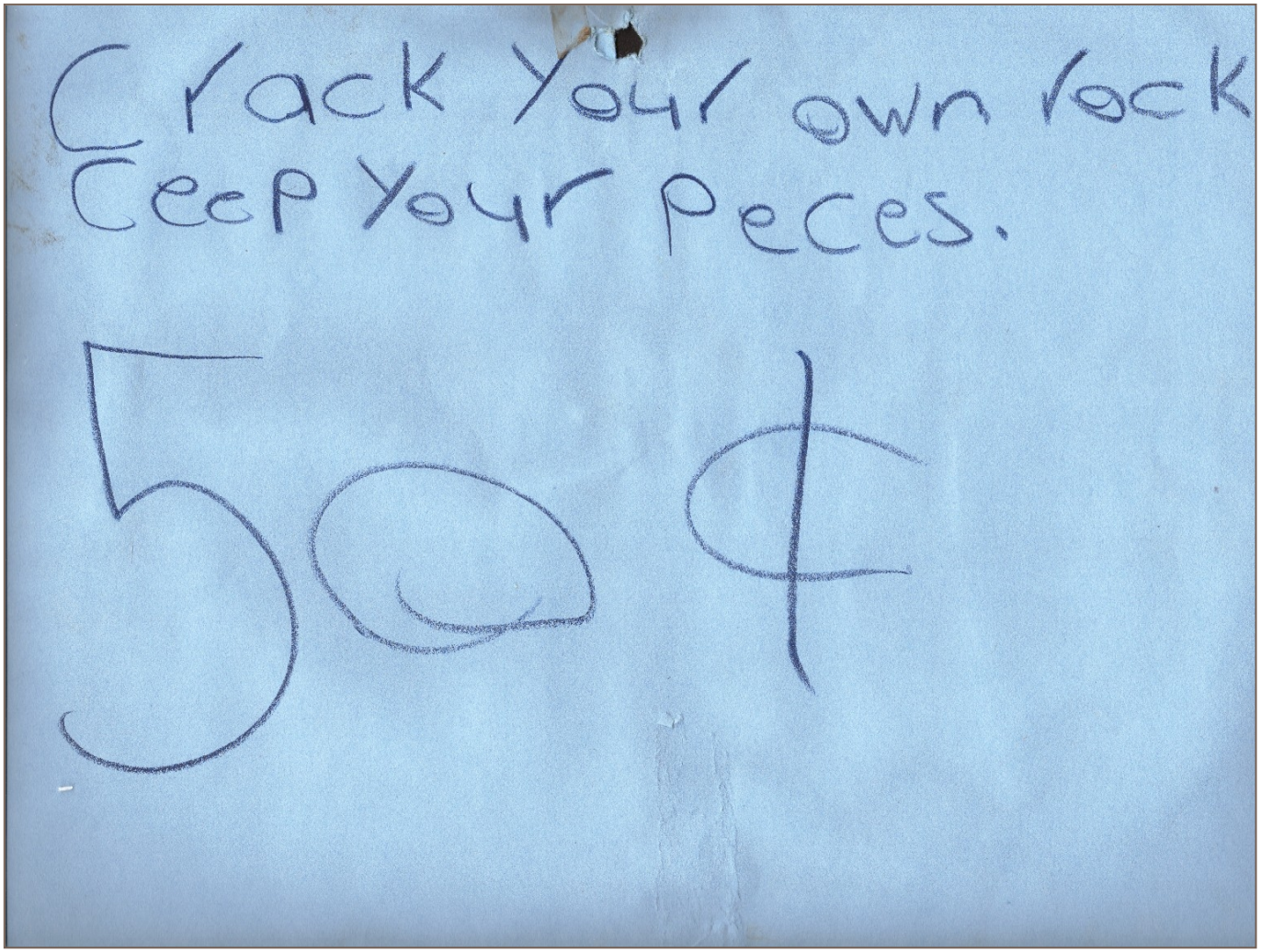
- Healthcare crosscuts:
  - Regulatory: laws, regulations, coding, guidelines, best practices, performance, costs, reporting (i.e., adverse event), etc.
  - Research: basic research, pharmaceuticals, medical devices, genetics, drug-drug interaction, diagnostic test decision support, biomedical research data mining (basic or clinical results), etc.
  - IT systems: interoperability, software development, information/data storage, security and access, reporting, “Big Data” and small data, usability, data transfer, training/educating/communicating, and so much more!

Areas where data mining can help!

So How Do We Get There?

Understanding and Then Tackling the Pieces!

Medical Data Mining



Where are the opportunities?

Build On History and Knowledge

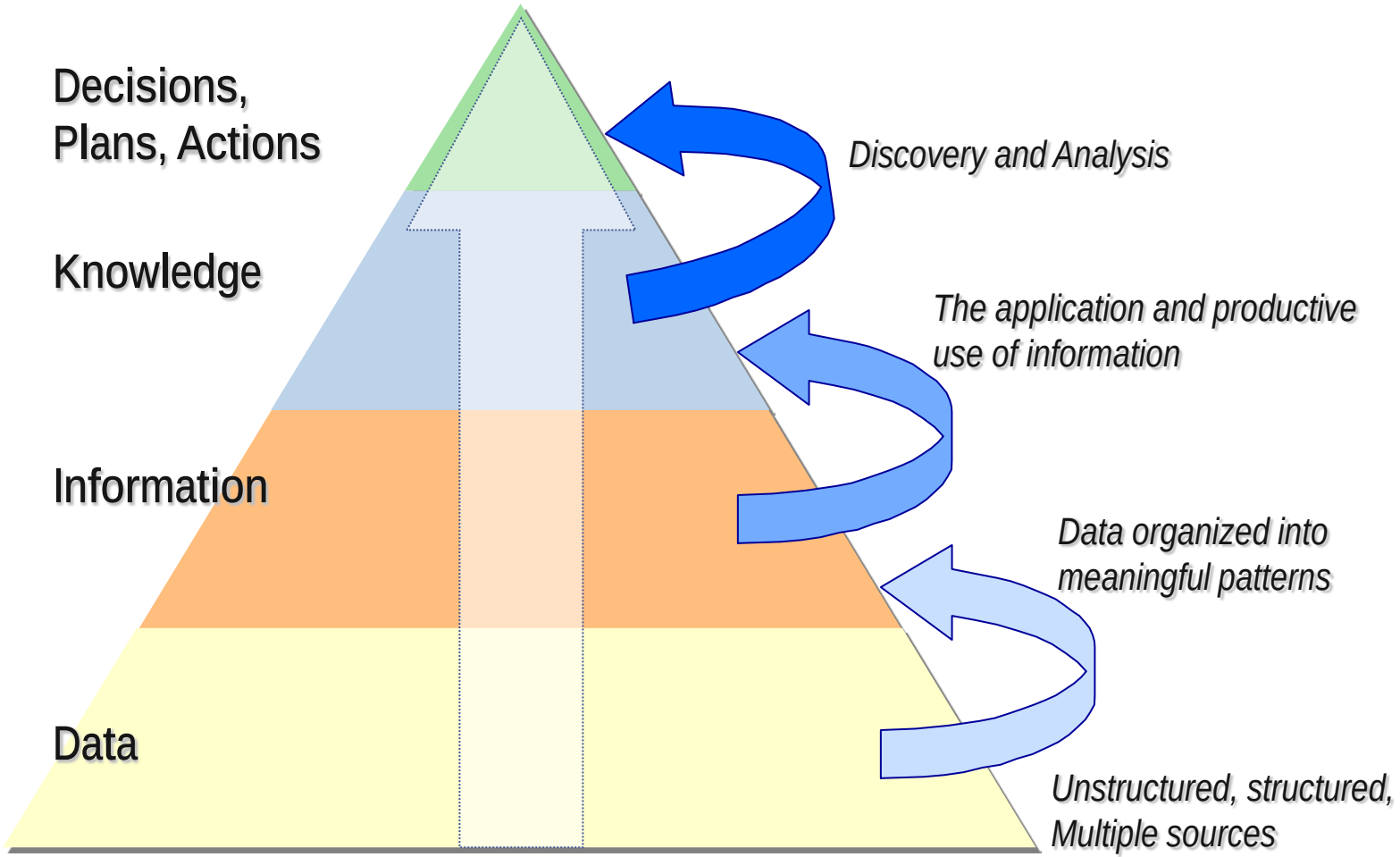
Practice domains / Fields we can learn from:

- **Knowledge management** is the theory behind knowledge capture and use
- **Informatics** is the science of information, the practice of information processing, and the engineering of information systems
- **Analytics** is the practical application of tools (i.e. algorithms) upon information to gain new insights.

- Others:
  - Business Intelligence
  - Competitive Intelligence
  - Computational Science
  - Bioinformatics
  - Health Informatics
  - Predictive Modeling
  - Decision Support
  - Artificial Intelligence, etc.

Data Mining

Effective Use of Data, Tools, and Analyses



Medical Data Mining

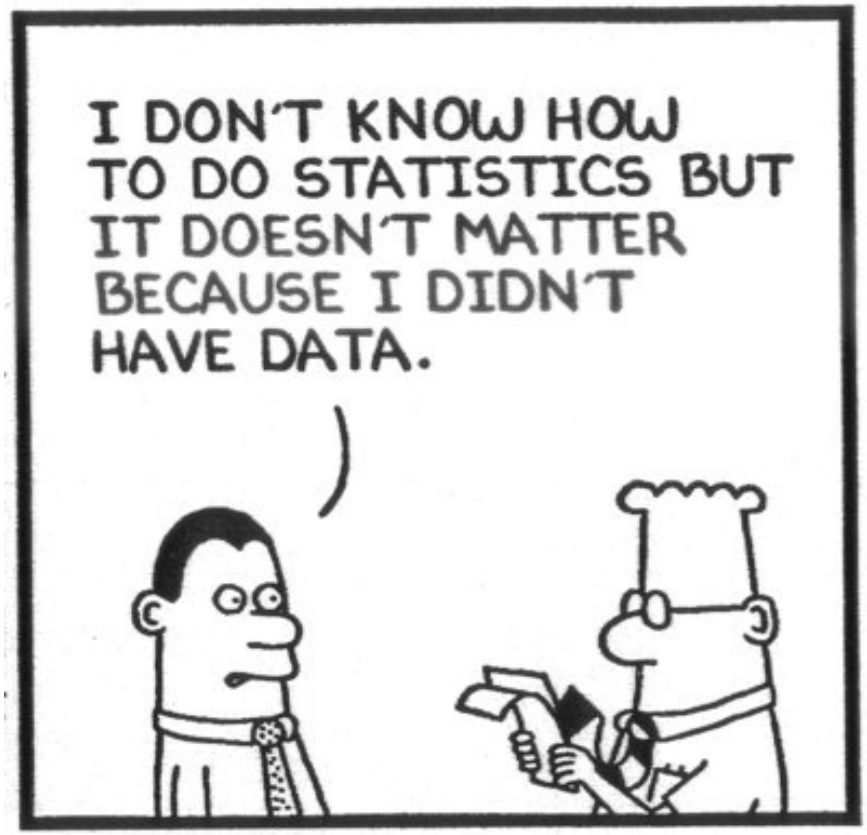
- Leads to:
  - Question based answers
  - Anomaly based discovery
  - New Knowledge discovery
  - Informed decisions
  - Probability measures
  - Predictive modeling
  - Decision support
  - Improved health
  - Personalized medicine

So, Again, How Do We Get There?

- What questions are you trying to answer?
  - Ultimately, identify answers to questions we didn't know we had
- Do you have data, tools and analyses to answer the question?
- Example areas:
  - Healthcare management (provider care practices)
  - Fraud and abuse
  - Treatment effectiveness
  - Patient involvement and relationship

Understanding and Then Tackling the Pieces!

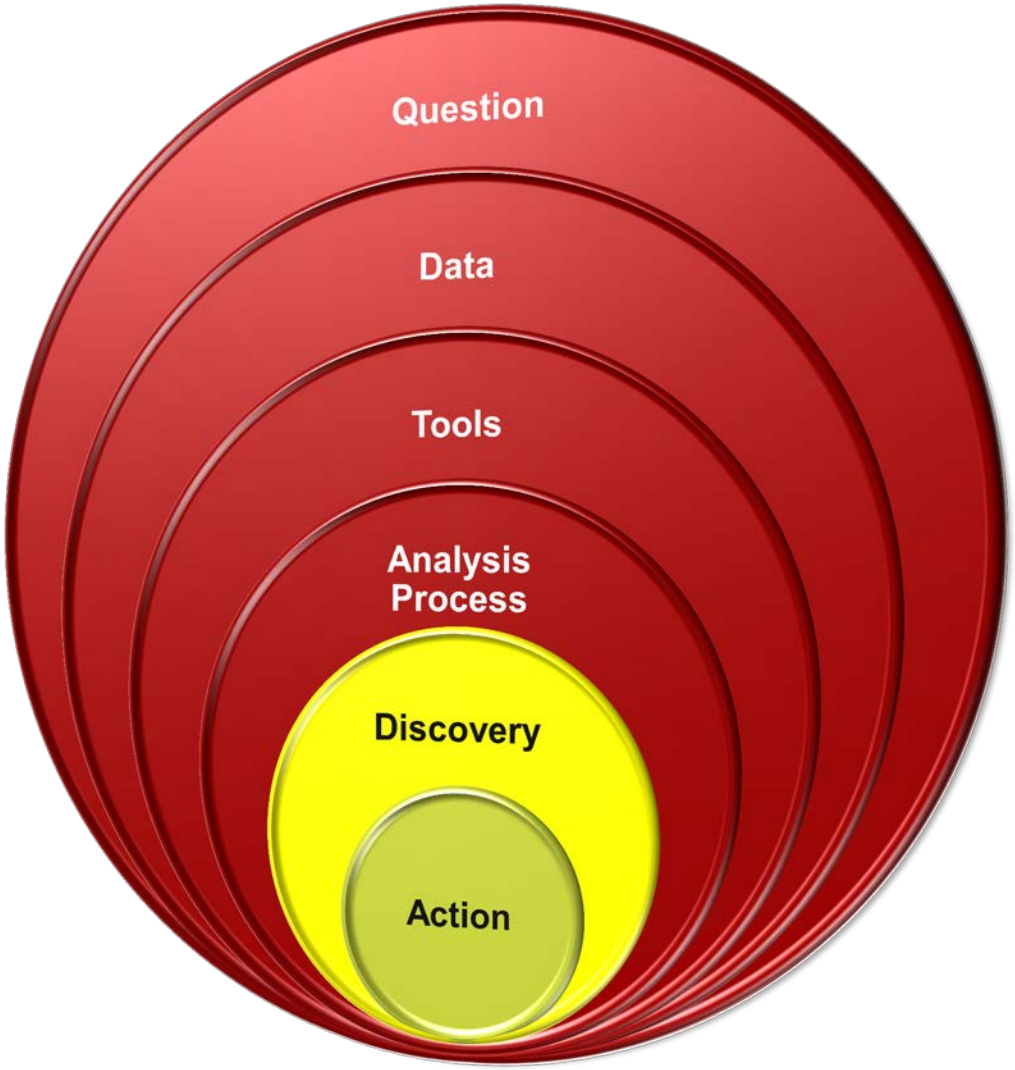
Dilbert on Data



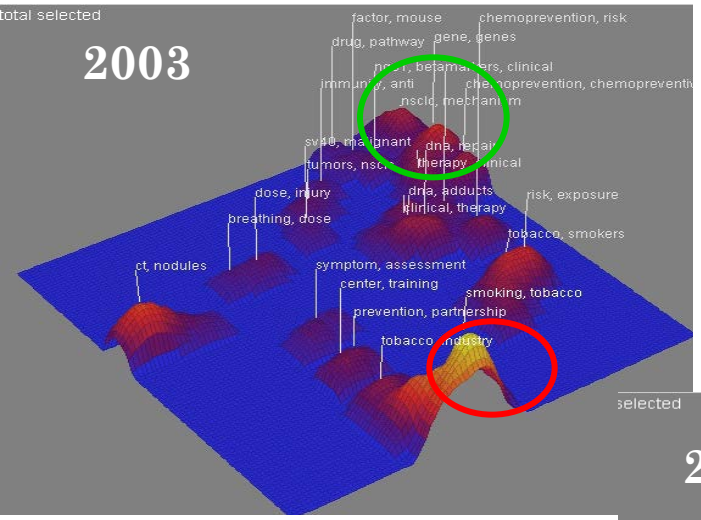
There are Exabyte's of data!

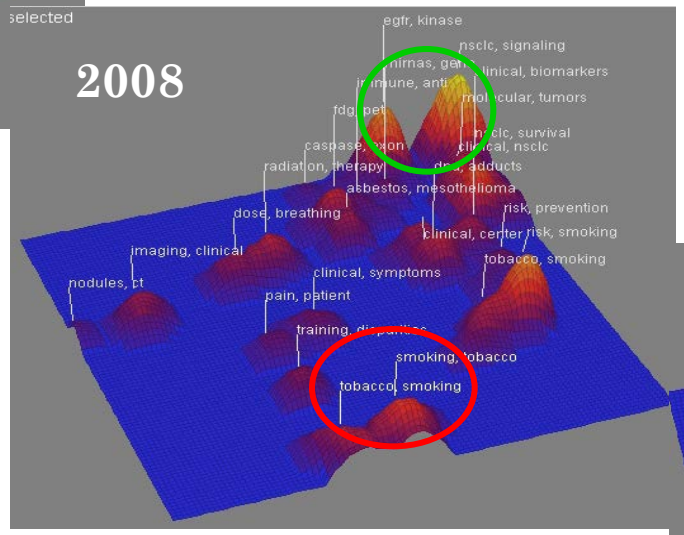
But is it the right data to answer your question?

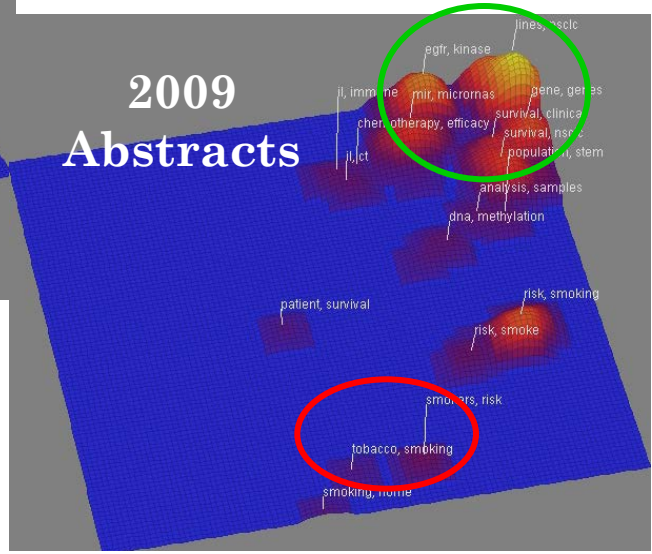
So, Again, How Do We Get There?



Evolution of the Lung Cancer Portfolio: 2003-2009







So, Again, How Do We Get There?

- Data: the V's (Volume, Velocity, Variety, Veracity, and Visibility)
- Tools/Applications: Search algorithms, text mining, natural language processing (NLP), machine learning, etc. clustering, predictive modeling, relationship and link analysis, taxonomy generation, statistical analysis, neural networks, visualization, heat maps, etc.
- Analysis: Methodologies (exploration and drill down) and subject matter/domain experts

Understanding and Then Tackling the Pieces!

Data

- The Data V's
  - Volume - large, small, combined, separate, etc.
  - Velocity – transfer: capture and retrieval includes streaming, batch processing, utilization, etc.
  - Variety - text [structured→unstructured], images, audio, video, etc.
  - Veracity - meaningfulness [use], value, variability, quality, etc.
  - Visibility - access, security,

Understanding and Then Tackling the Pieces!

Tools / Applications

- 1000's of Tools
- Integrated or add-on
- Question/Domain/Analysis specific
- Use with Repositories and Platforms
  - Helpful but not necessary for analysis
  - Relevant to data veracity/quality

Understanding and Then Tackling the Pieces!

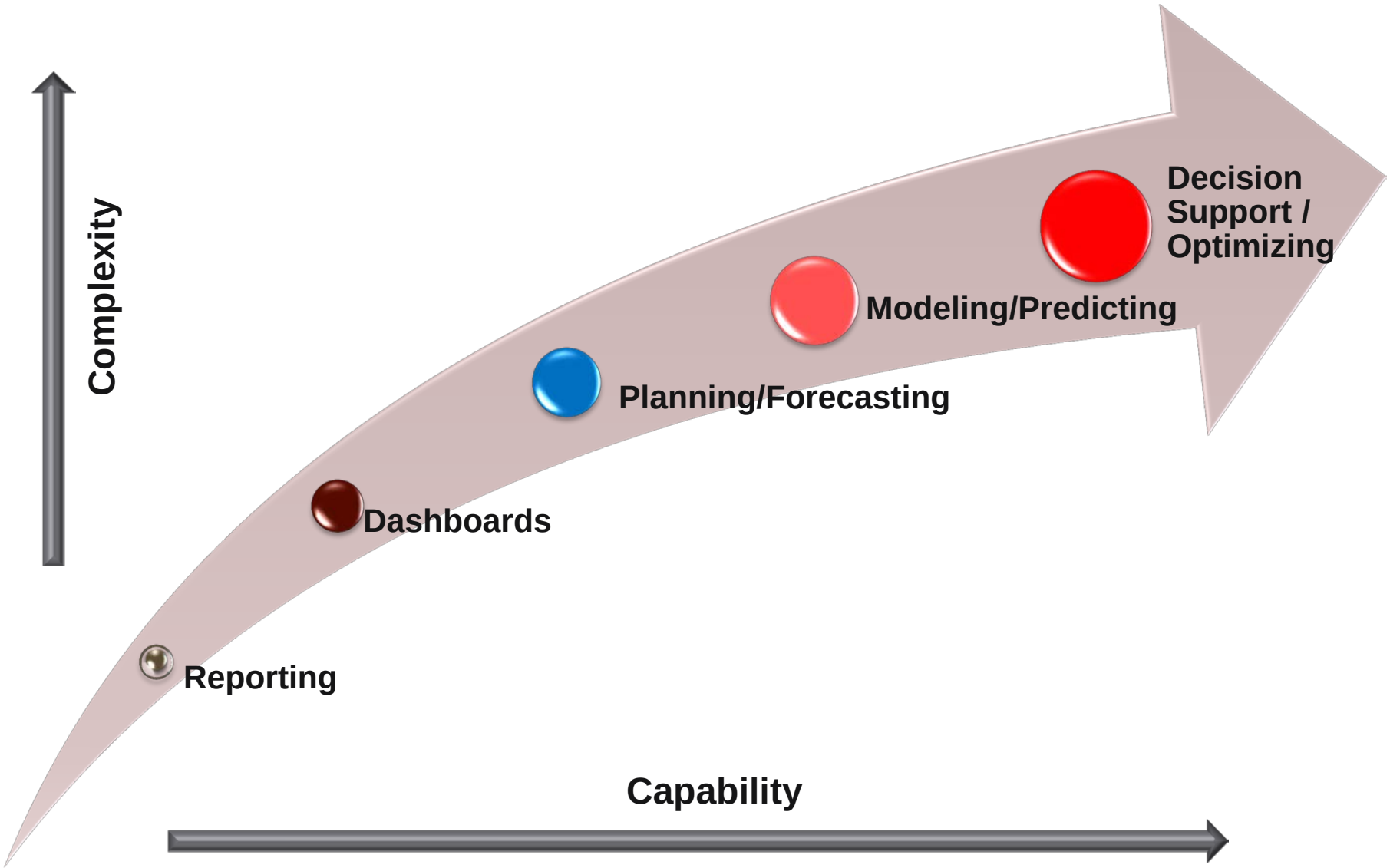
Analysis

- Methodologies
  - Research
  - Algorithms
  - Proprietary

- Subject matter experts
  - Domain (healthcare centric) expertise
  - Analysis expertise
  - Tool and application experience

Understanding and Then Tackling the Pieces!

Medical Data Mining Spectrum



Medical Data Mining

- It sounds good, but are there standards for data capture, use, definitions, sharing, etc.?
  - Are standards needed (yet)?
- Are there sufficient tools, applications, analyses and staff available to identify valuable information to improve healthcare?
- Are there sufficient benefits and incentives in core areas where data mining is essential?
  - Personalized and predictive medicine
  - Fraud and abuse
  - Research advancements
  - Improved treatments and medical devices

Dilbert on Federal & Corporate Realities

An integrated approach is key!

Biomedical Data Mining Case Study

NIH Research, Condition, and Disease Categorization Project

National Institutes of Health (NIH): Case Study

The purpose of the Research, Condition, and Disease Categorization (RCDC) project is to

1. Consistently categorize NIH-funded research projects according to research areas/categories
2. Use an automated process, and
3. Make the results available to Congress and the public

NIH: Case Study

How does it do that?

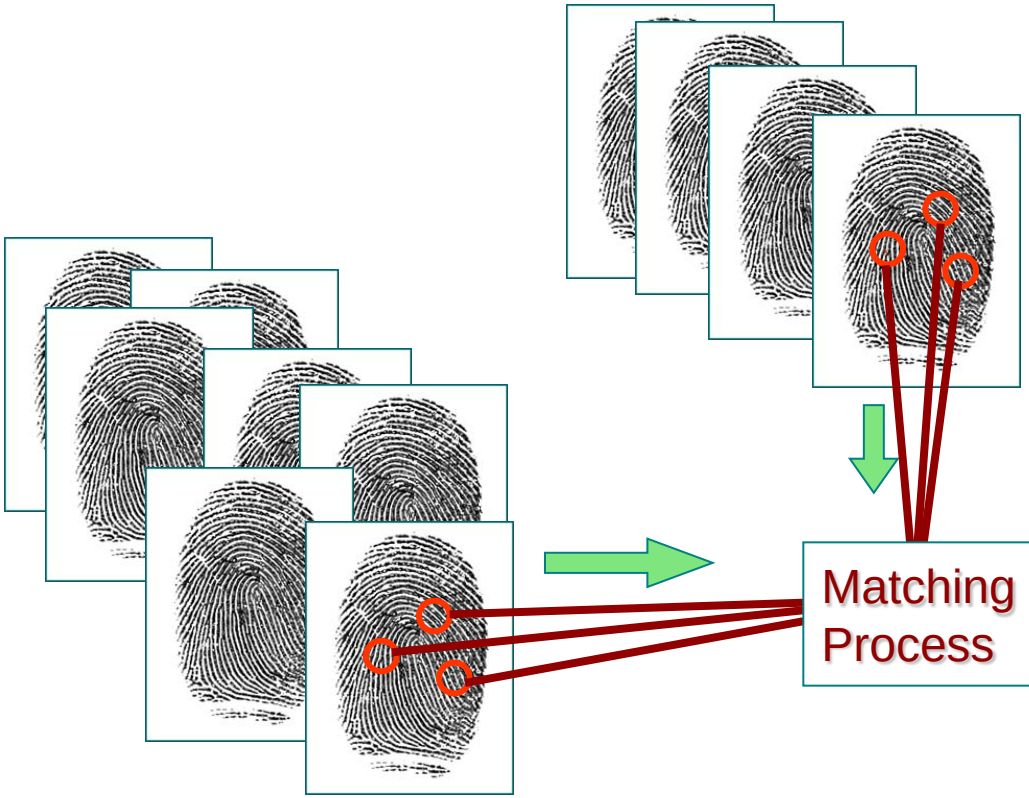
- RCDC system uses Elsevier's Collexis technology to text mine biomedical concepts from research descriptions
- NIH research experts define a weighted classification system for each of 238 categories
- All NIH research is then categorized through an automated process
- The output is reported publicly at Report.NIH.Gov

NIH: Case Study

The Research, Condition, and Disease Categorization (RCDC) project is an example of providing new, timely information out of unstructured and structured data

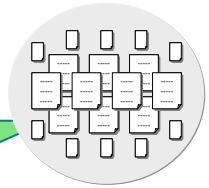
- RCDC pulls data from 7 databases (all containing well over 4 terabytes of content) that gets routed, compartmentalized, validated and ultimately used for regular reports and on-demand queries
- Allows research information to be explored proactively
- Is adaptable as
    1) Science evolves over time and
    2) Increased needs for data usage are identified

How Does RCDC Work?



Category
Definition developed by
NIH staff

Matching
Process

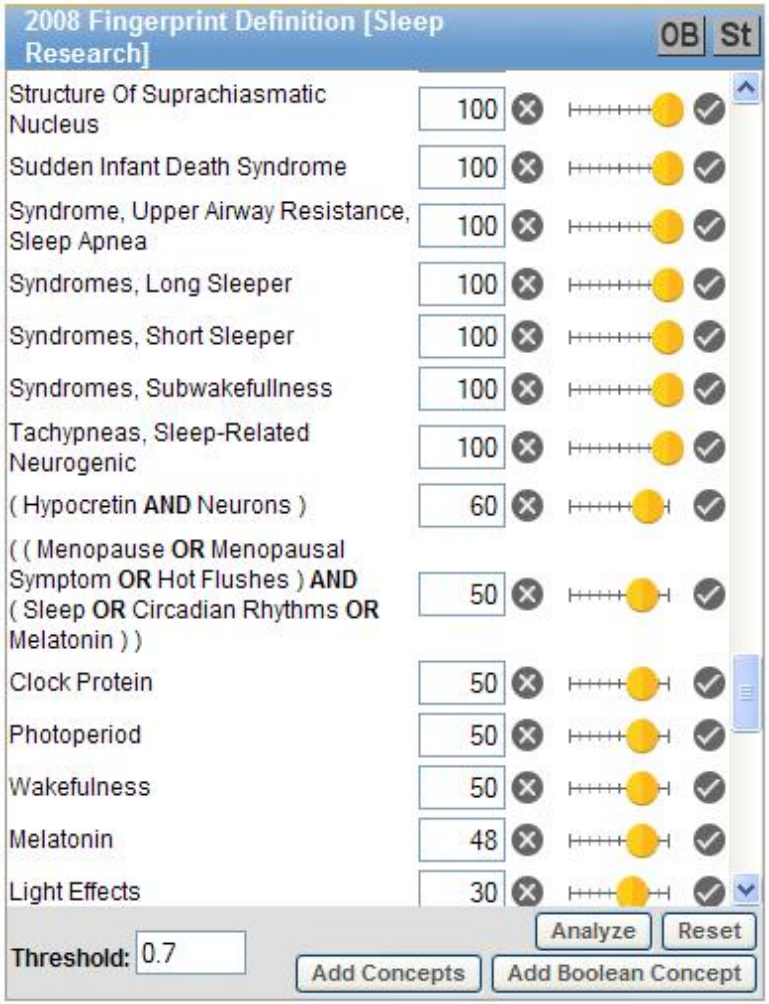


Projects with matching categories

Project Description

Matching compares individual project descriptions to the category definition – the resultant category matching score for a project is a reflection of how closely the project is related to a particular category

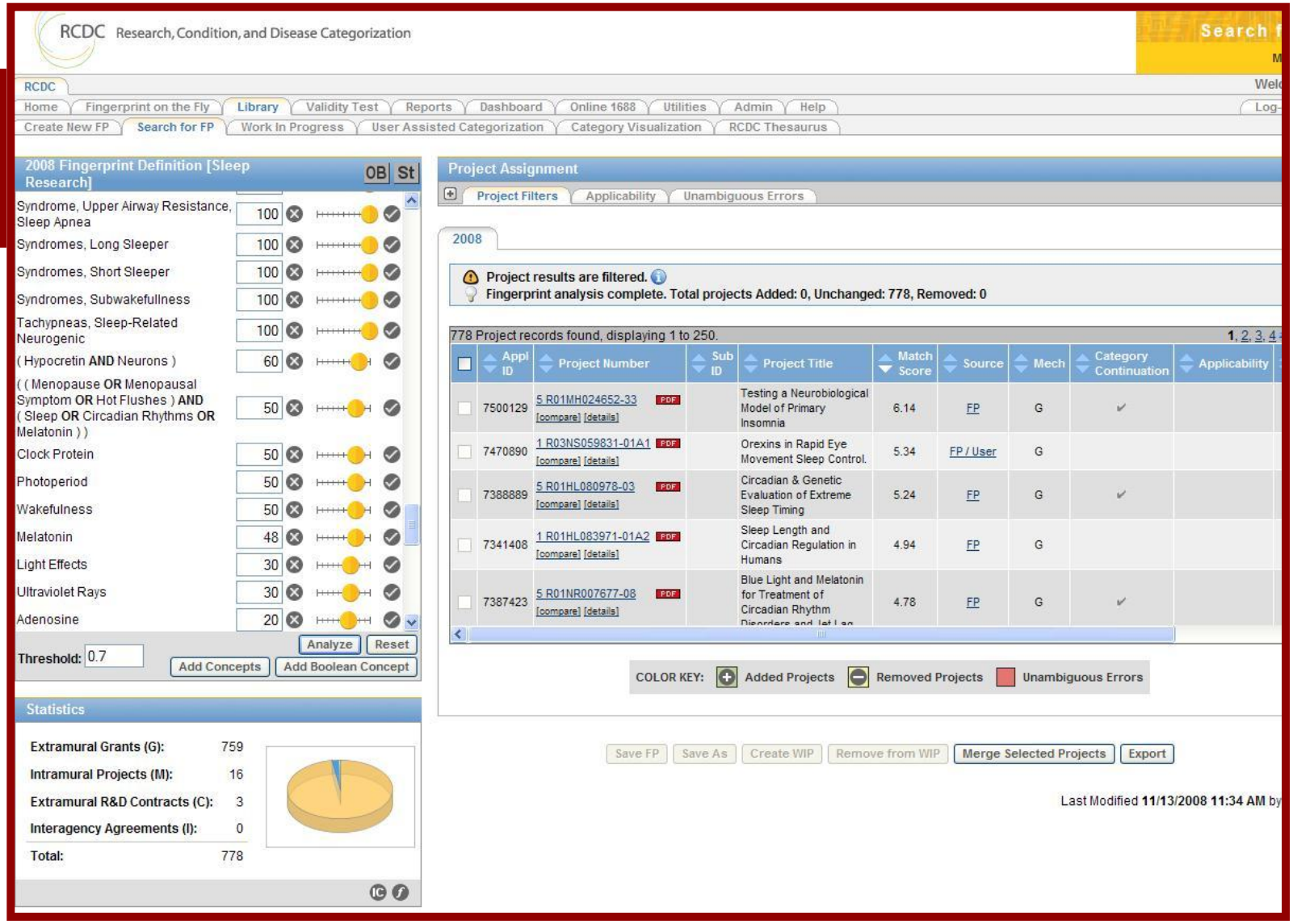
Category Definition Created in the New System



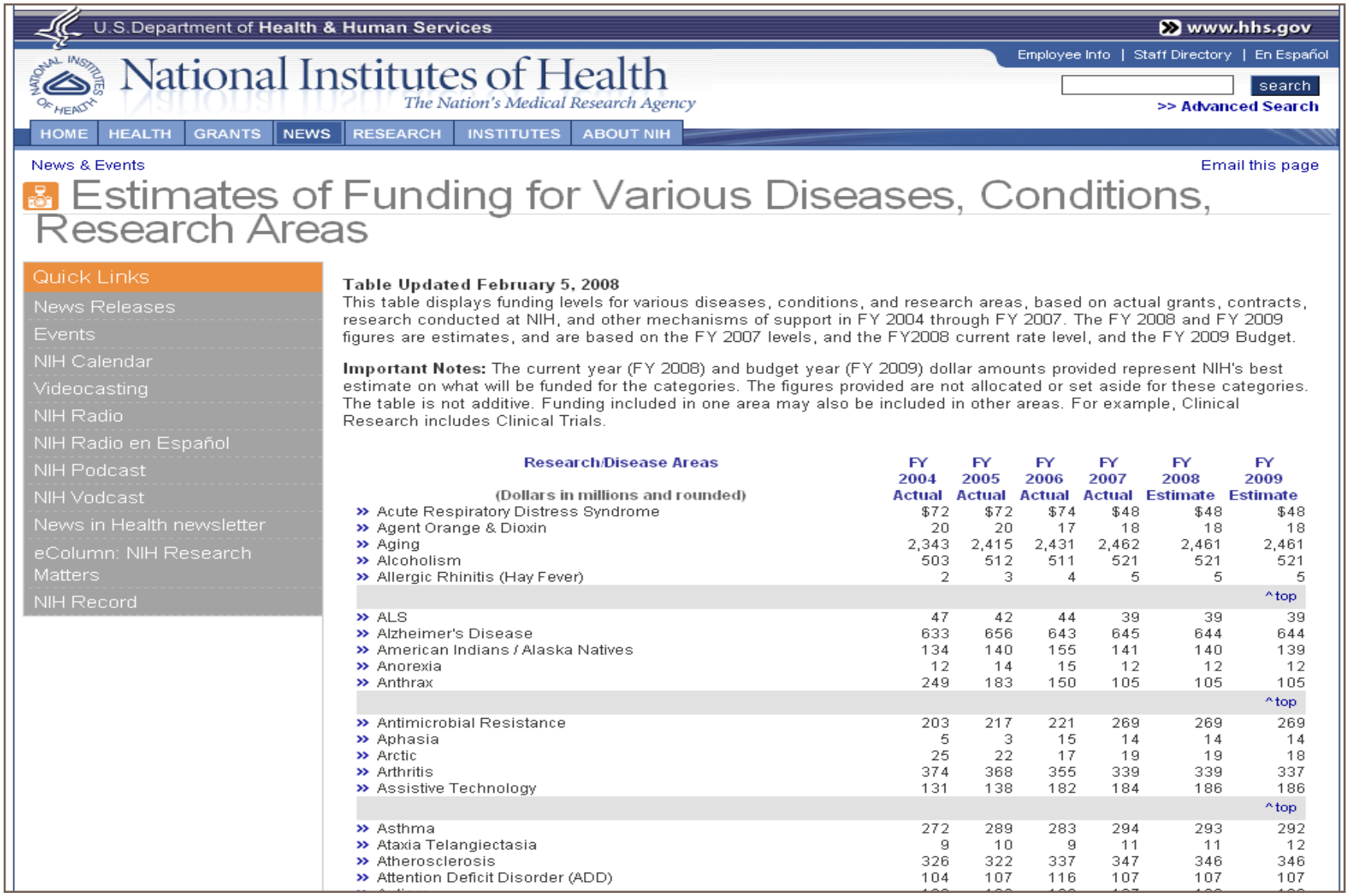
- A definition is a list of scientific terms from a thesaurus (300,000 terms and synonyms).
- Terms are selected by NIH Scientific Experts to define that research category.
- Terms are weighted to fine-tune the matching process.
- Terms from grants/projects are matched against definitions to produce category project lists.

Sample: Sleep Research Draft Fingerprint

Analytical Tool



Prior View of Data (FY 2009)



U.S. Department of Health & Human Services — www.hhs.gov

Employee Info | Staff Directory | En Español

National Institutes of Health
The Nation's Medical Research Agency

search
>> Advanced Search

HOME | HEALTH | GRANTS | NEWS | RESEARCH | INSTITUTES | ABOUT NIH

News & Events — Email this page

Estimates of Funding for Various Diseases, Conditions, Research Areas

Quick Links
- News Releases
- Events
- NIH Calendar
- Videocasting
- NIH Radio
- NIH Radio en Español
- NIH Podcast
- NIH Vodcast
- News in Health newsletter
- eColumn: NIH Research Matters
- NIH Record

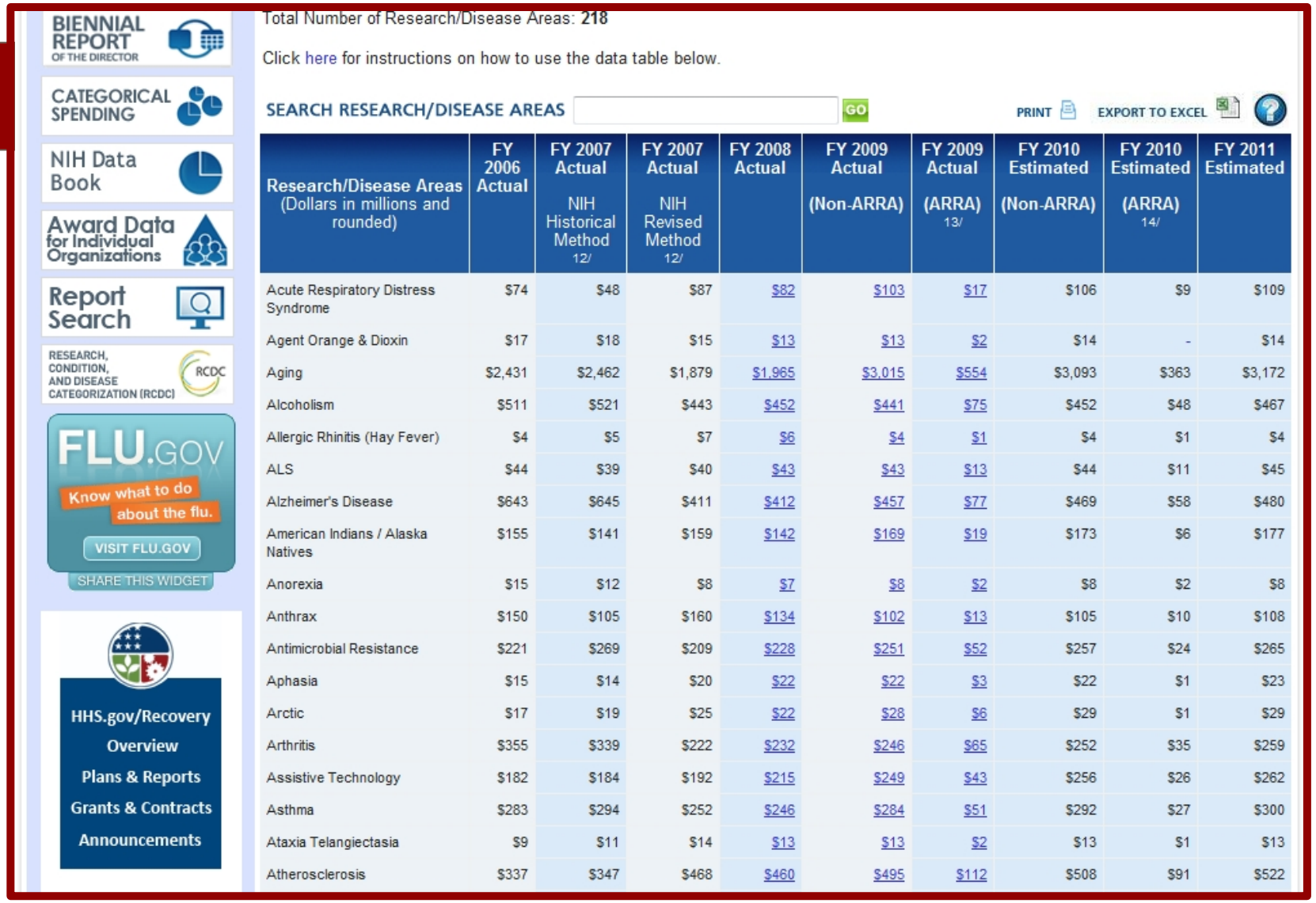
Table Updated February 5, 2008
This table displays funding levels for various diseases, conditions, and research areas, based on actual grants, contracts, research conducted at NIH, and other mechanisms of support in FY 2004 through FY 2007. The FY 2008 and FY 2009 figures are estimates, and are based on the FY 2007 levels, and the FY2008 current rate level, and the FY 2009 Budget.

Important Notes: The current year (FY 2008) and budget year (FY 2009) dollar amounts provided represent NIH's best estimate on what will be funded for the categories. The figures provided are not allocated or set aside for these categories. The table is not additive. Funding included in one area may also be included in other areas. For example, Clinical Research includes Clinical Trials.

Research/Disease Areas (Dollars in millions and rounded)	FY 2004 Actual	FY 2005 Actual	FY 2006 Actual	FY 2007 Actual	FY 2008 Estimate	FY 2009 Estimate
» Acute Respiratory Distress Syndrome	$72	$72	$74	$48	$48	$48
» Agent Orange & Dioxin	20	20	17	18	18	18
» Aging	2,343	2,415	2,431	2,462	2,461	2,461
» Alcoholism	503	512	511	521	521	521
» Allergic Rhinitis (Hay Fever)	2	3	4	5	5	5
						^top
» ALS	47	42	44	39	39	39
» Alzheimer's Disease	633	656	643	645	644	644
» American Indians / Alaska Natives	134	140	155	141	140	139
» Anorexia	12	14	15	12	12	12
» Anthrax	249	183	150	105	105	105
						^top
» Antimicrobial Resistance	203	217	221	269	269	269
» Aphasia	5	3	15	14	14	14
» Arctic	25	22	17	19	19	18
» Arthritis	374	368	355	339	339	337
» Assistive Technology	131	138	182	184	186	186
						^top
» Asthma	272	289	283	294	293	292
» Ataxia Telangiectasia	9	10	9	11	11	12
» Atherosclerosis	326	322	337	347	346	346
» Attention Deficit Disorder (ADD)	104	107	116	107	107	107

http://www.nih.gov/news/fundingresearchareas.htm

Summary level data



BIENNIAL REPORT OF THE DIRECTOR

CATEGORICAL SPENDING

NIH Data Book

Award Data for Individual Organizations

Report Search

RESEARCH, CONDITION, AND DISEASE CATEGORIZATION (RCDC)

FLU.GOV
Know what to do about the flu.
VISIT FLU.GOV
SHARE THIS WIDGET

HHS.gov/Recovery
Overview
Plans & Reports
Grants & Contracts
Announcements

Total Number of Research/Disease Areas: **218**

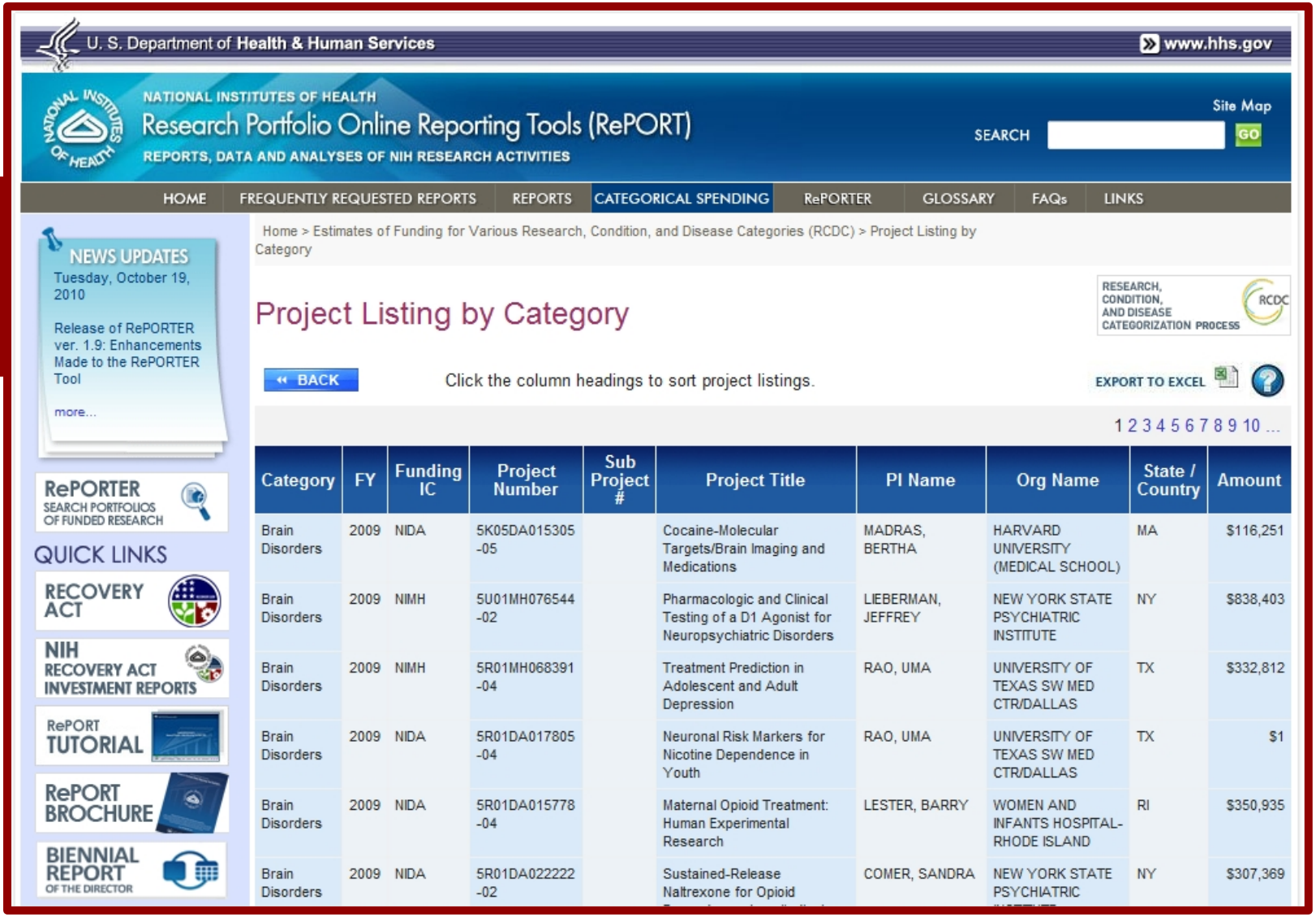
Click here for instructions on how to use the data table below.

SEARCH RESEARCH/DISEASE AREAS GO PRINT EXPORT TO EXCEL

Research/Disease Areas (Dollars in millions and rounded)	FY 2006 Actual	FY 2007 Actual NIH Historical Method 12/	FY 2007 Actual NIH Revised Method 12/	FY 2008 Actual	FY 2009 Actual (Non-ARRA)	FY 2009 Actual (ARRA) 13/	FY 2010 Estimated (Non-ARRA)	FY 2010 Estimated (ARRA) 14/	FY 2011 Estimated
Acute Respiratory Distress Syndrome	$74	$48	$87	$82	$103	$17	$106	$9	$109
Agent Orange & Dioxin	$17	$18	$15	$13	$13	$2	$14	-	$14
Aging	$2,431	$2,462	$1,879	$1,965	$3,015	$554	$3,093	$363	$3,172
Alcoholism	$511	$521	$443	$452	$441	$75	$452	$48	$467
Allergic Rhinitis (Hay Fever)	$4	$5	$7	$6	$4	$1	$4	$1	$4
ALS	$44	$39	$40	$43	$43	$13	$44	$11	$45
Alzheimer's Disease	$643	$645	$411	$412	$457	$77	$469	$58	$480
American Indians / Alaska Natives	$155	$141	$159	$142	$169	$19	$173	$6	$177
Anorexia	$15	$12	$8	$7	$8	$2	$8	$2	$8
Anthrax	$150	$105	$160	$134	$102	$13	$105	$10	$108
Antimicrobial Resistance	$221	$269	$209	$228	$251	$52	$257	$24	$265
Aphasia	$15	$14	$20	$22	$22	$3	$22	$1	$23
Arctic	$17	$19	$25	$22	$28	$6	$29	$1	$29
Arthritis	$355	$339	$222	$232	$246	$65	$252	$35	$259
Assistive Technology	$182	$184	$192	$215	$249	$43	$256	$26	$262
Asthma	$283	$294	$252	$246	$284	$51	$292	$27	$300
Ataxia Telangiectasia	$9	$11	$14	$13	$13	$2	$13	$1	$13
Atherosclerosis	$337	$347	$468	$460	$495	$112	$508	$91	$522

http://report.nih.gov/rcdc/categories/

Drill down level data, not available on web in the past



U. S. Department of Health & Human Services — www.hhs.gov

NATIONAL INSTITUTES OF HEALTH
Research Portfolio Online Reporting Tools (RePORT)
REPORTS, DATA AND ANALYSES OF NIH RESEARCH ACTIVITIES

Site Map | SEARCH | GO

HOME | FREQUENTLY REQUESTED REPORTS | REPORTS | CATEGORICAL SPENDING | RePORTER | GLOSSARY | FAQs | LINKS

NEWS UPDATES
Tuesday, October 19, 2010
Release of RePORTER ver. 1.9: Enhancements Made to the RePORTER Tool
more...

RePORTER SEARCH PORTFOLIOS OF FUNDED RESEARCH

QUICK LINKS
RECOVERY ACT
NIH RECOVERY ACT INVESTMENT REPORTS
RePORT TUTORIAL
RePORT BROCHURE
BIENNIAL REPORT OF THE DIRECTOR

Home > Estimates of Funding for Various Research, Condition, and Disease Categories (RCDC) > Project Listing by Category

Project Listing by Category

RESEARCH, CONDITION, AND DISEASE CATEGORIZATION PROCESS (RCDC)

« BACK | Click the column headings to sort project listings. | EXPORT TO EXCEL

1 2 3 4 5 6 7 8 9 10 ...

Category	FY	Funding IC	Project Number	Sub Project #	Project Title	PI Name	Org Name	State / Country	Amount
Brain Disorders	2009	NIDA	5K05DA015305-05		Cocaine-Molecular Targets/Brain Imaging and Medications	MADRAS, BERTHA	HARVARD UNIVERSITY (MEDICAL SCHOOL)	MA	$116,251
Brain Disorders	2009	NIMH	5U01MH076544-02		Pharmacologic and Clinical Testing of a D1 Agonist for Neuropsychiatric Disorders	LIEBERMAN, JEFFREY	NEW YORK STATE PSYCHIATRIC INSTITUTE	NY	$838,403
Brain Disorders	2009	NIMH	5R01MH068391-04		Treatment Prediction in Adolescent and Adult Depression	RAO, UMA	UNIVERSITY OF TEXAS SW MED CTR/DALLAS	TX	$332,812
Brain Disorders	2009	NIDA	5R01DA017805-04		Neuronal Risk Markers for Nicotine Dependence in Youth	RAO, UMA	UNIVERSITY OF TEXAS SW MED CTR/DALLAS	TX	$1
Brain Disorders	2009	NIDA	5R01DA015778-04		Maternal Opioid Treatment: Human Experimental Research	LESTER, BARRY	WOMEN AND INFANTS HOSPITAL-RHODE ISLAND	RI	$350,935
Brain Disorders	2009	NIDA	5R01DA022222-02		Sustained-Release Naltrexone for Opioid	COMER, SANDRA	NEW YORK STATE PSYCHIATRIC	NY	$307,369

http://report.nih.gov/rcdc/categories/

NIH: Case Study

Ahead of its time

- RCDC opened the door for analysis and review of research that was not previously possible (including decision intelligence practices)
- Additional uses of new analytical, visualization, and exploration technologies are now taking place because the platform exists!

Benefits

- Enhanced NIH's ability to:
  - Leverage existing information and processes
  - Conduct text mining and perform scientific portfolio analysis
  - Provide transparency into government spending on research
- Greatly improved process
  - Consistent methodology (one definition per category)
  - Reproducible numbers
  - New open platform to support decision intelligence
- Improved public understanding of NIH spending
  - Access to project listings not available previously
  - Searchable, accessible query tools and reports

Summary

- The opportunity and future for Medical Data Mining is HUGE!
- Practice areas cover the landscape: Patient, Provider, Payer, Research, Regulatory and IT
- Tackle it in chucks!
  - Question based data mining
  - Don’t try to build the be-all end-all data source – use what’s available to begin to answer critical questions sooner rather than later
- Aspects of Data are critical
- The right Tool for the right job
- Analysis requires well trained analysts