

Using the IN-SPIRE clustering tool to characterize portfolios and identify potential overlap

George Santangelo

The Office of Portfolio Analysis

DPCPSI / OD

Office of Portfolio Analysis (OPA)

OPA Mission Components

- Coordinate portfolio analysis activities across NIH
- Train and consult with NIH staff to promote best practices
- Develop a science of portfolio analysis
 - Build new tools and augment pre-existing ones
 - Build a community of experts: government, academia, private sector

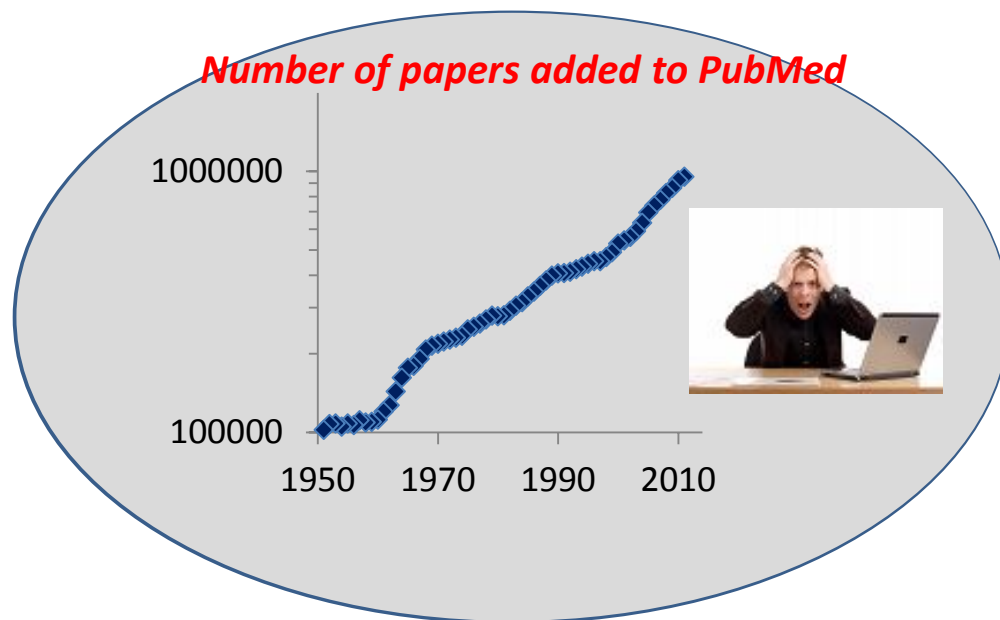
The Tao of OPA

- Beware of “garbage in, gospel out”
- You can’t manage what you can’t measure
- What you measure is what you get

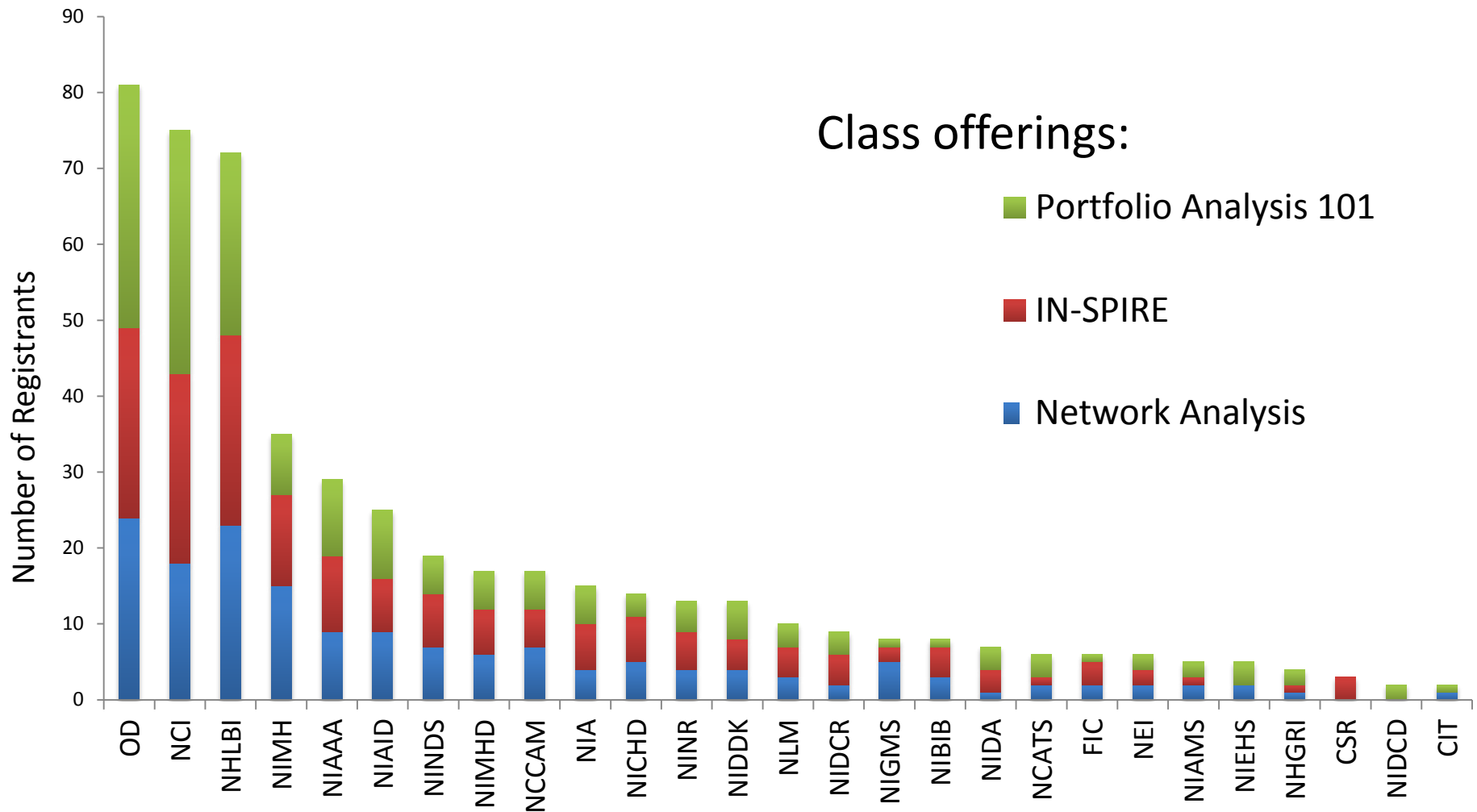


Benefits of high-throughput document clustering

- Insights that improve portfolio management and strategic planning
- Comprehensive and convincing demonstration that inappropriate overlap is minimal or non-existent
- Improved understanding of how portfolios align with the current literature as it continues to expand

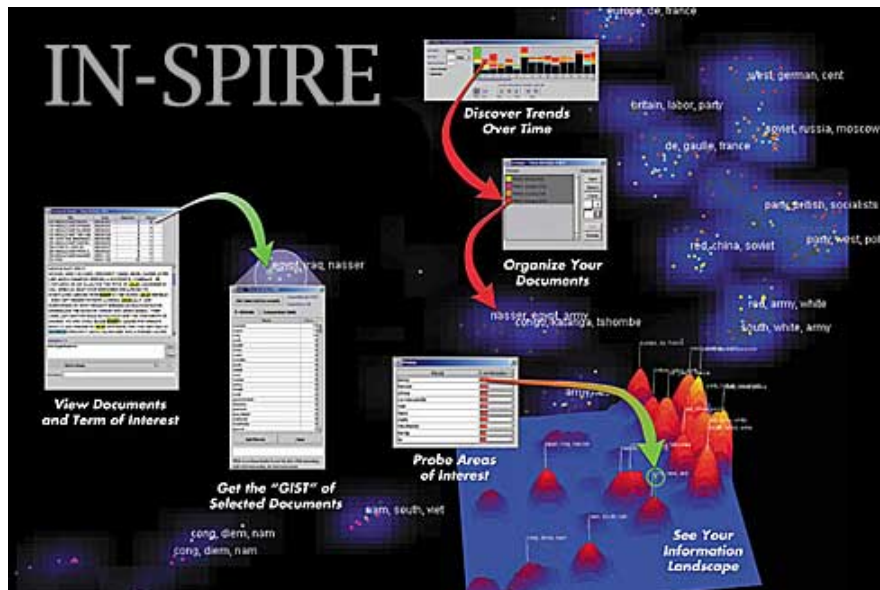


OPA Training Registrants



IN-SPIRE: Analyzing & Visualizing Information

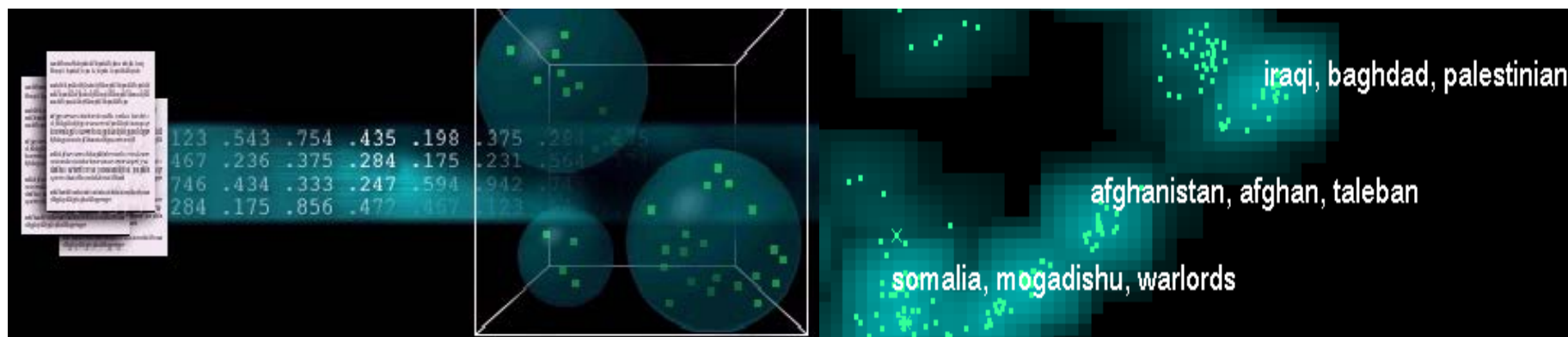
- Developed by Pacific Northwest National Labs (PNNL) in Richland, Washington



- Can import and categorize any set of documents
- IN-SPIRE 5.7.2 released in Aug 2013
- All OPA analysts have completed advanced IN-SPIRE training at PNNL

- OPA offers Introductory and Intermediate IN-SPIRE classes for NIH staff
- OPA / PNNL partnership is adding IN-SPIRE features useful to NIH

IN-SPIRE Text Processing



Extract Text from Documents

- *Create a mathematical signal (vector) for each document*

Organize According to Key Topics

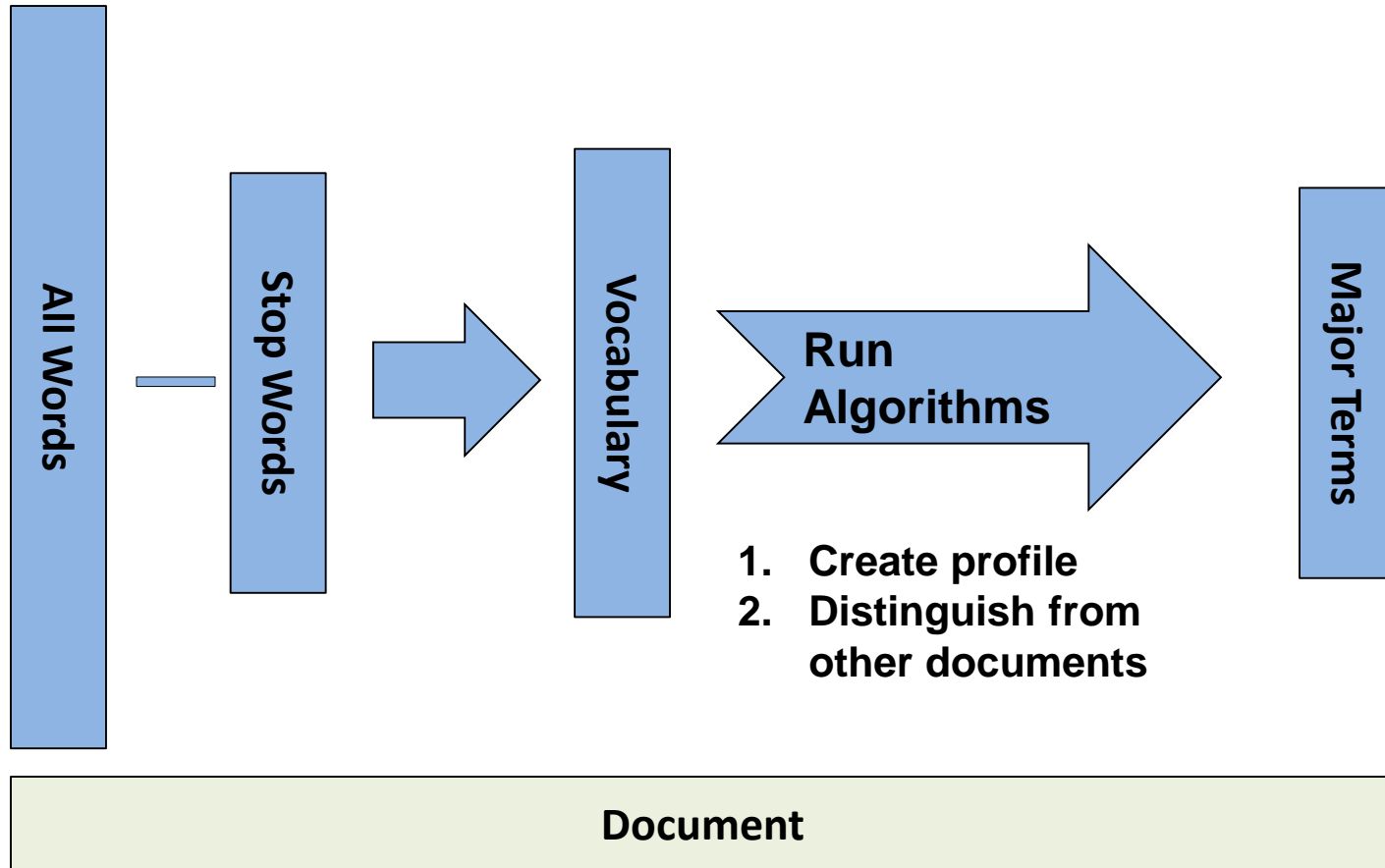
- *Cluster the document vectors in n -space*

Present each document as a “docustar” where proximity suggests similar themes

- *Project the n -space clusters into a two dimensional visualization*

IN-SPIRE Text Processing

(All Words) – (Stop Words) = Vocabulary



IN-SPIRE Text Processing: an example

In recent years, increasing numbers of people of all ages have been heeding their health professionals' advice to get active for all of the health benefits exercise has to offer. But for some people—particularly those who overdo or who don't properly train or warm up—these benefits can come at a price: sports injuries.

IN-SPIRE Text Processing: an example

In recent years, increasing numbers of people of all ages have been heeding their health professionals' advice to get active for all of the health benefits exercise has to offer. But for some people—particularly those who overdo or who don't properly train or warm up—these benefits can come at a price: sports injuries.

IN-SPIRE Analysis and Visualization

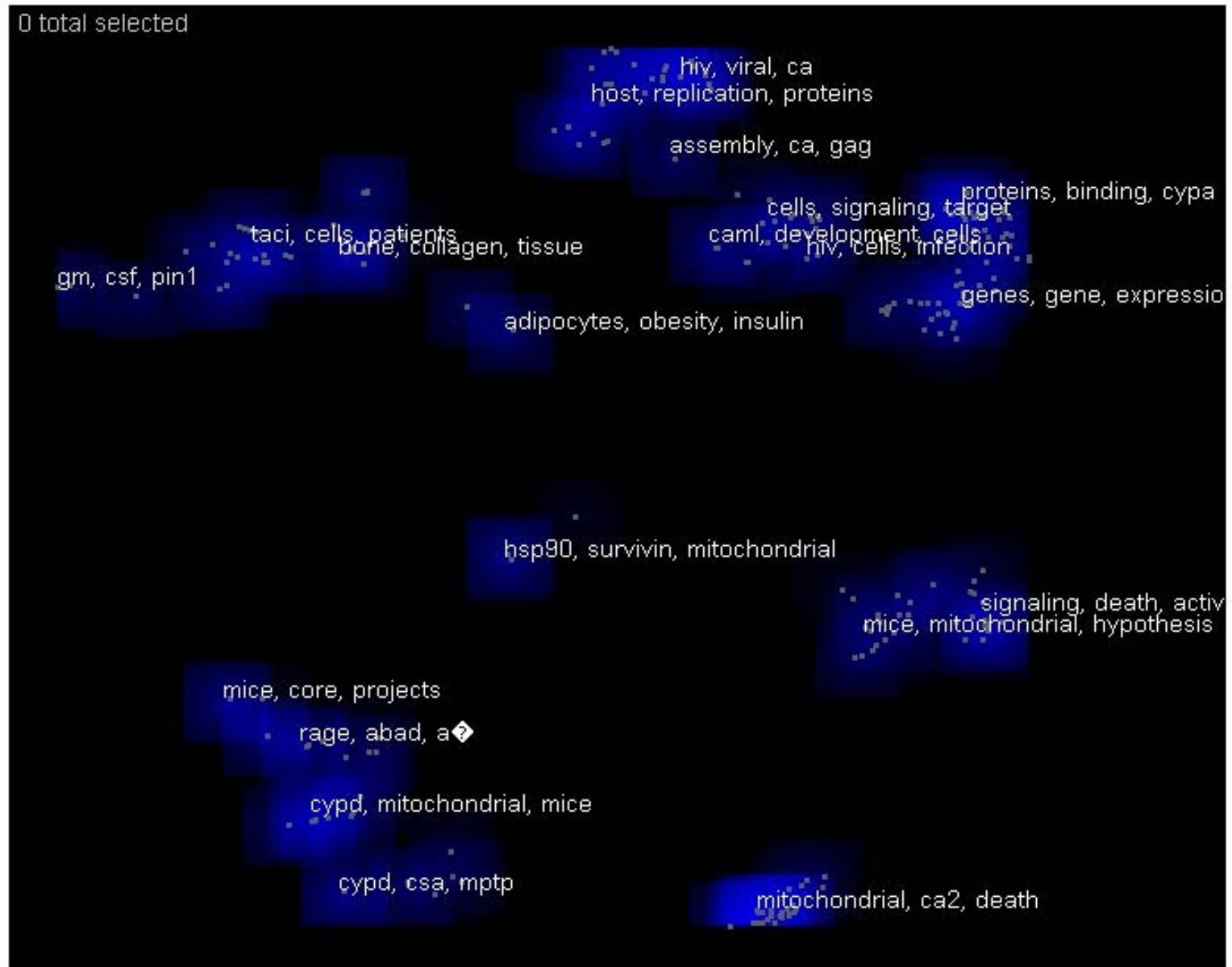
Analysis

- Thematic distribution by various metadata
- Query relationships and overlap
- Targeted search
- Time slicing
- Informed exploration and discovery

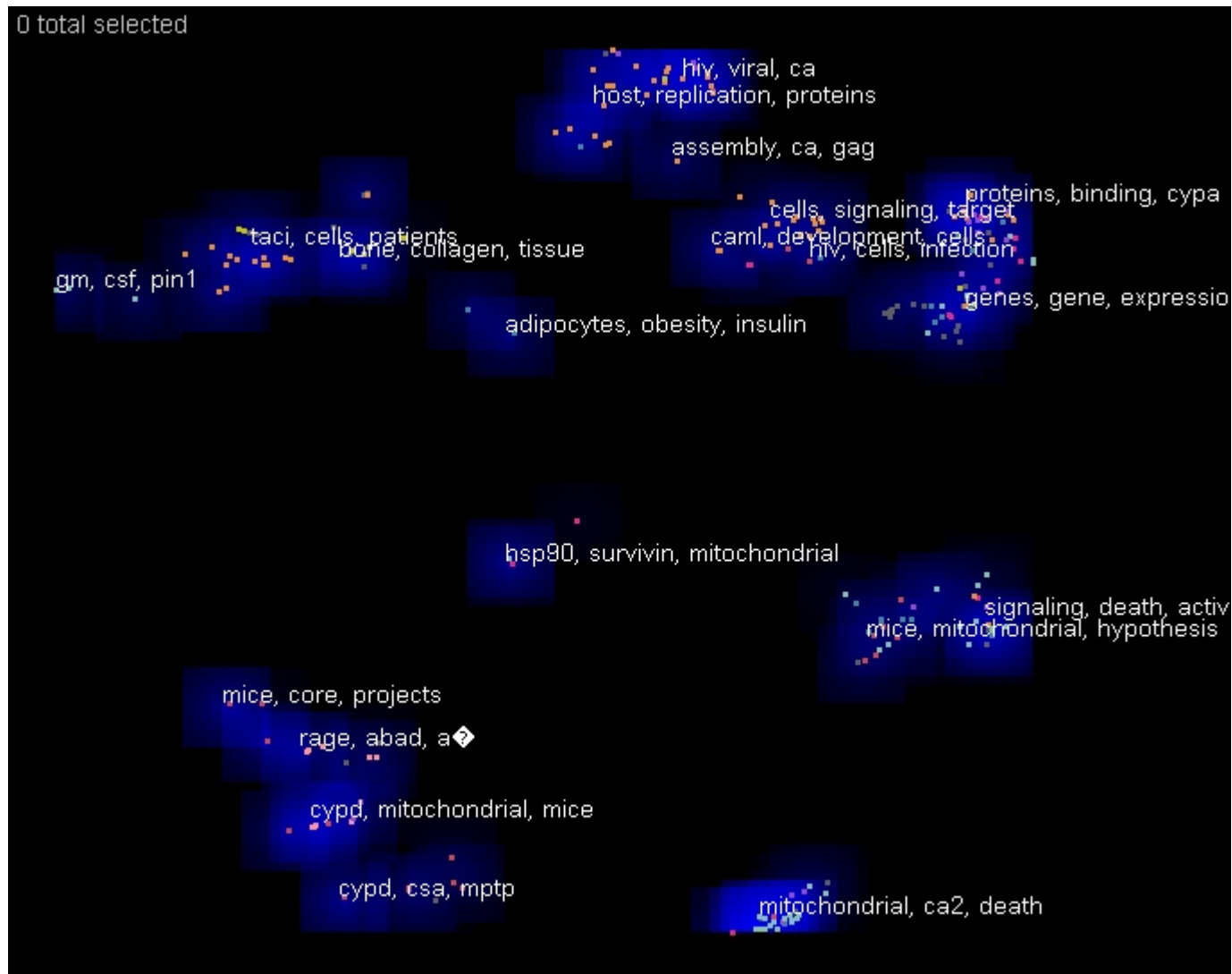
Visualization

- Galaxy View permits intuitive interaction to explore the dataset
- Theme View provides a 3-D representation of clusters

Galaxy View

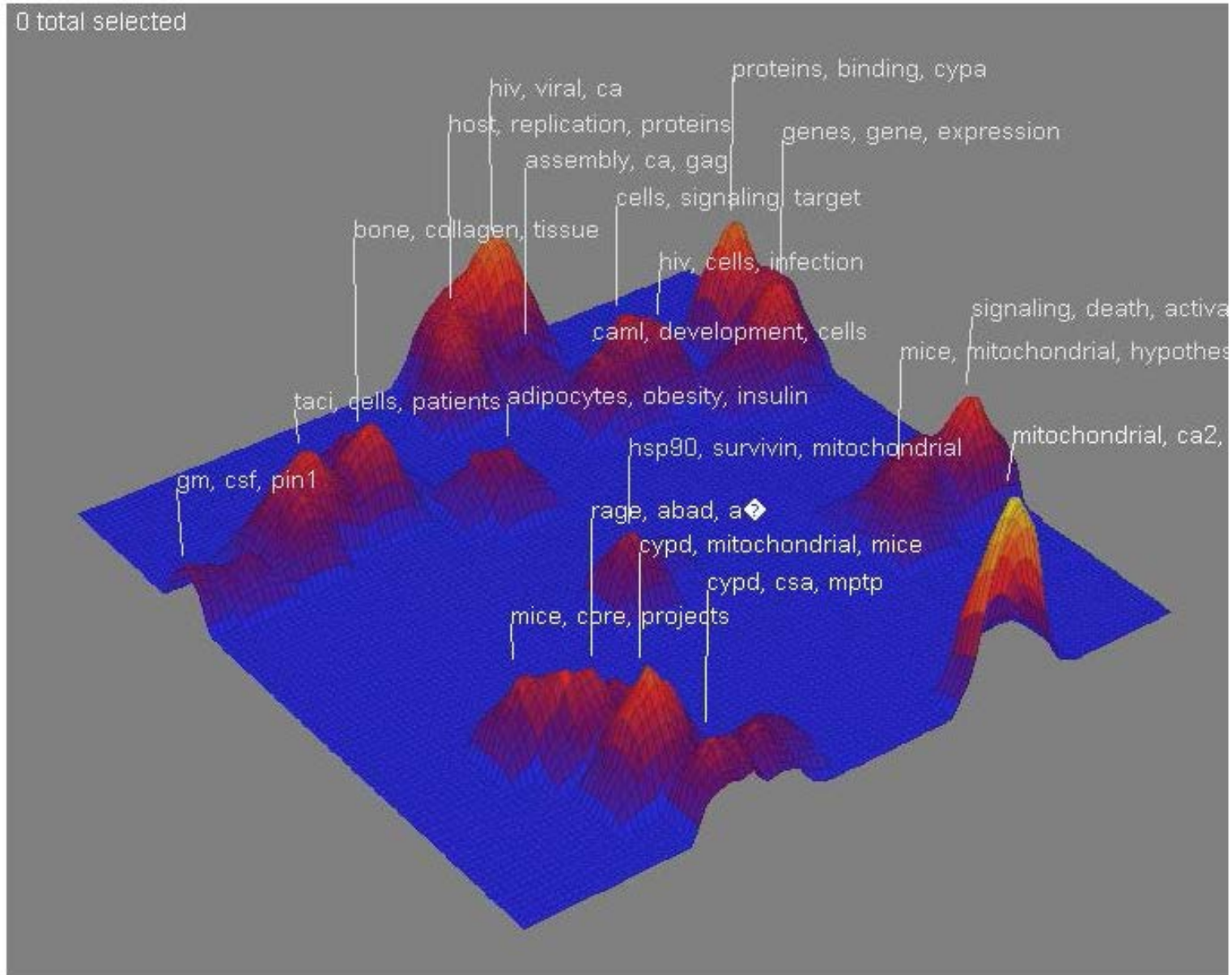


Galaxy View: Documents grouped in a variety of ways can be highlighted and exported



IC (0/476)
ai (0/139)
hl (0/85)
ns (0/64)
dk (0/44)
gm (0/35)
ca (0/26)
ag (0/21)
hd (0/12)
aa (0/10)
mh (0/8)
ar (0/6)
bx (0/5)
de (0/5)
ey (0/4)
rr (0/4)
cx (0/3)
da (0/2)

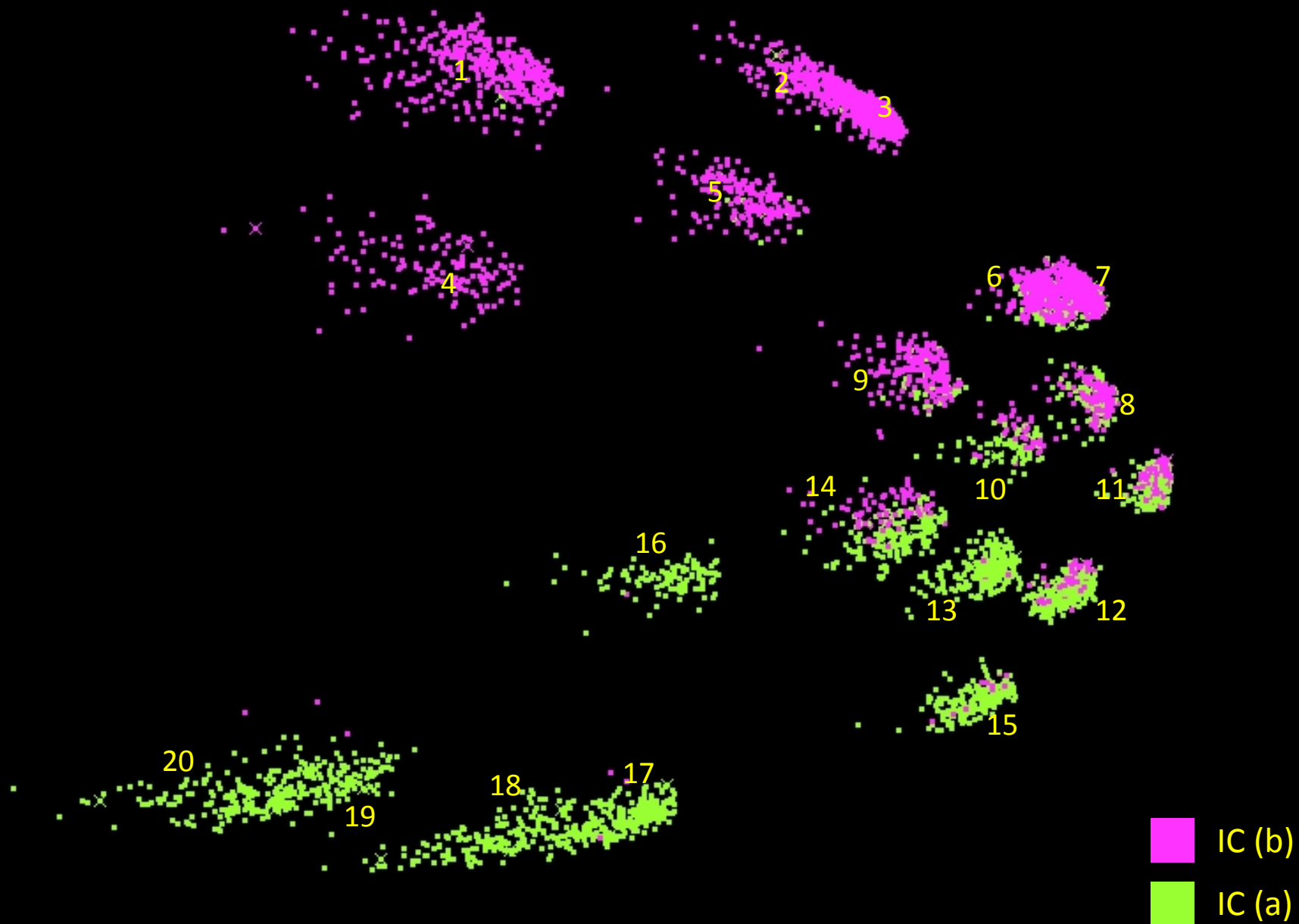
Theme View



Analyzing Overlap at NIH

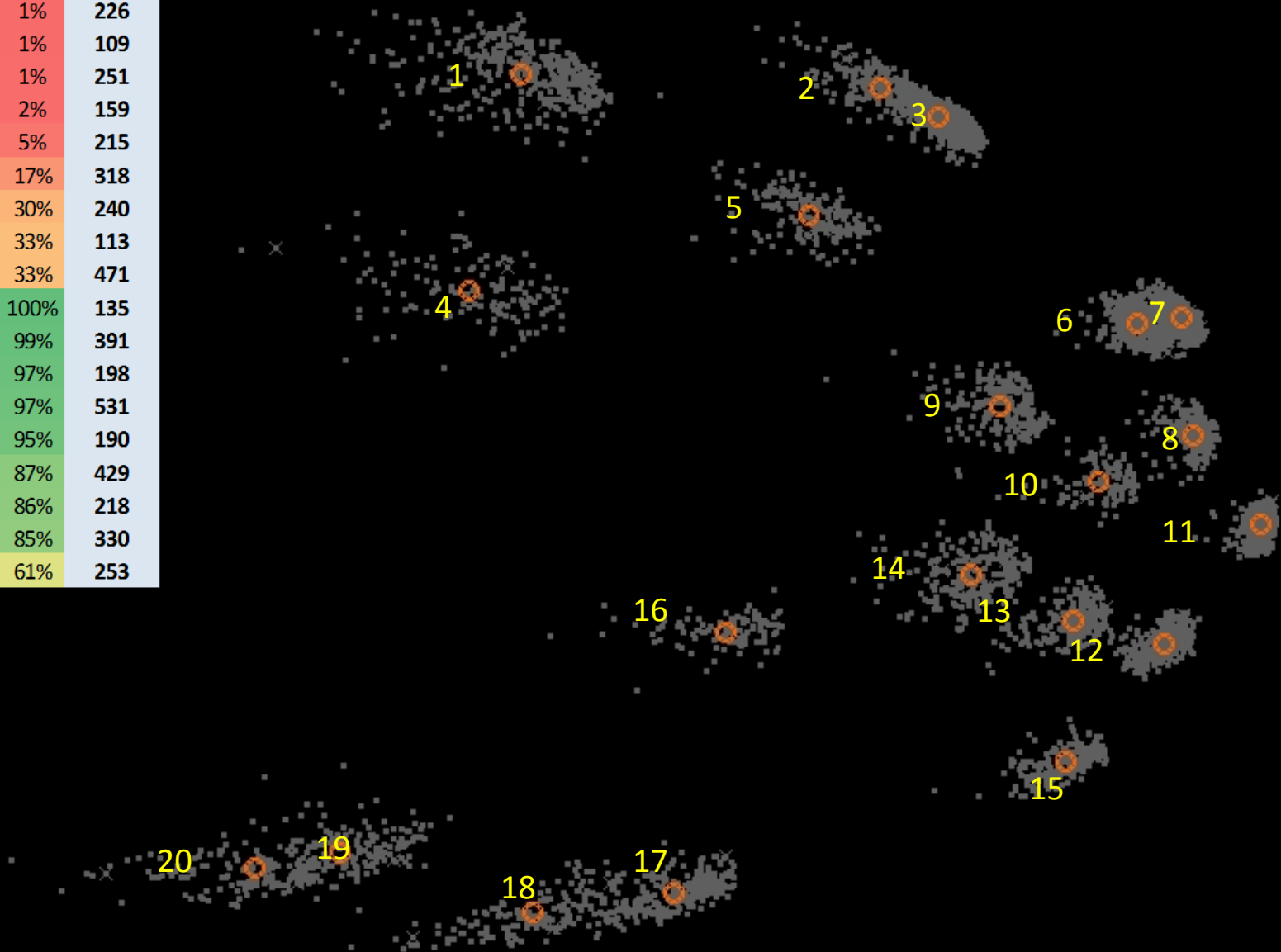
- IN-SPIRE analysis of seven of the largest NIH Institutes (ICs)
 - R01
 - Funded and Unfunded Applications
 - Active awards in FY12
- Head-to-head comparisons of these seven different ICs

IC (a) vs IC (b)

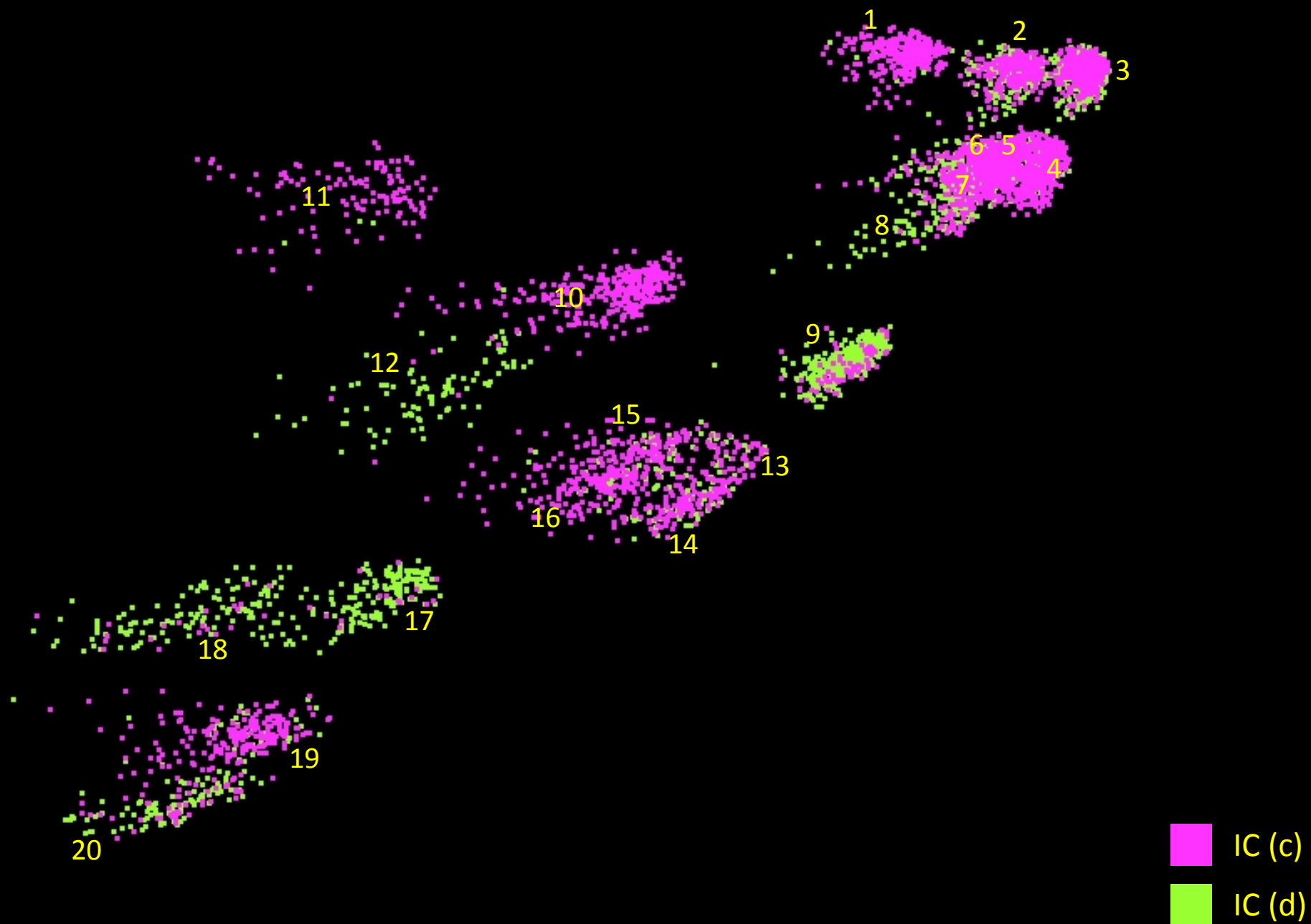


Cluster	IC (a)	IC (b)	No. of Projects
18	100%	0%	148
20	100%	0%	142
13	99%	1%	226
16	99%	1%	109
17	99%	1%	251
19	98%	2%	159
15	95%	5%	215
12	83%	17%	318
14	70%	30%	240
10	67%	33%	113
11	67%	33%	471
4	0%	100%	135
1	1%	99%	391
2	3%	97%	198
3	3%	97%	531
5	5%	95%	190
7	13%	87%	429
9	14%	86%	218
6	15%	85%	330
8	39%	61%	253

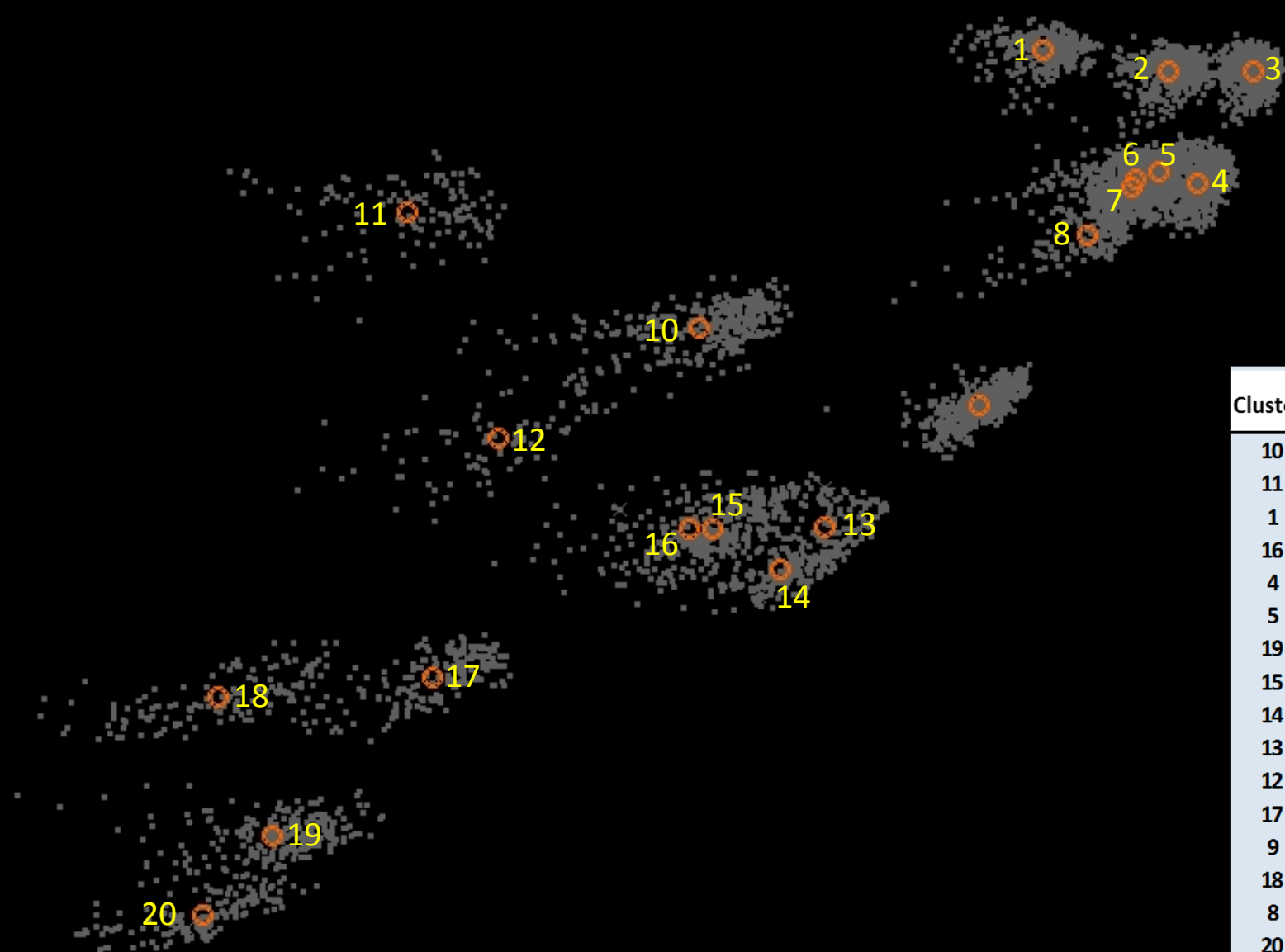
IC (a) vs IC (b)



IC (c) vs IC (d)



IC (c) vs IC (d)



Cluster	IC (c)	IC (d)	No. of Projects
10	100%	0%	317
11	98%	2%	126
1	97%	3%	309
16	94%	6%	114
4	89%	11%	509
5	89%	11%	441
19	85%	15%	230
15	76%	24%	246
14	75%	25%	172
13	75%	25%	126
12	3%	97%	98
17	8%	92%	177
9	15%	85%	431
18	16%	84%	129
8	21%	79%	165
20	36%	64%	149
7	41%	59%	449
3	50%	50%	907
6	51%	49%	316
2	53%	47%	419

Insights gained from IC/IC IN-SPIRE comparisons

- Most clusters are dominated by projects from a single Institute
- A few clusters contained a significant mixture of projects from both Institutes
- We chose to analyze further a cluster of 316 projects (153 funded, 163 unfunded) that contained approximately equal numbers of projects from IC (c) and IC (d)

Post-hoc Receipt & Referral

- Those 316 projects underwent a blind, independent evaluation by a subject matter expert (SME) from each of the two ICs
- 262 of the 316 projects were correctly coded by both IC SMEs

54 of 316 projects were difficult to resolve

29,515 R01 applications in FY12
(5,340 funded, 24,175 unfunded)



21 IC/IC comparisons of all R01s funded in FY12
(a total of 18,896 R01s)



IC (c) vs IC (d)



IN-SPIRE



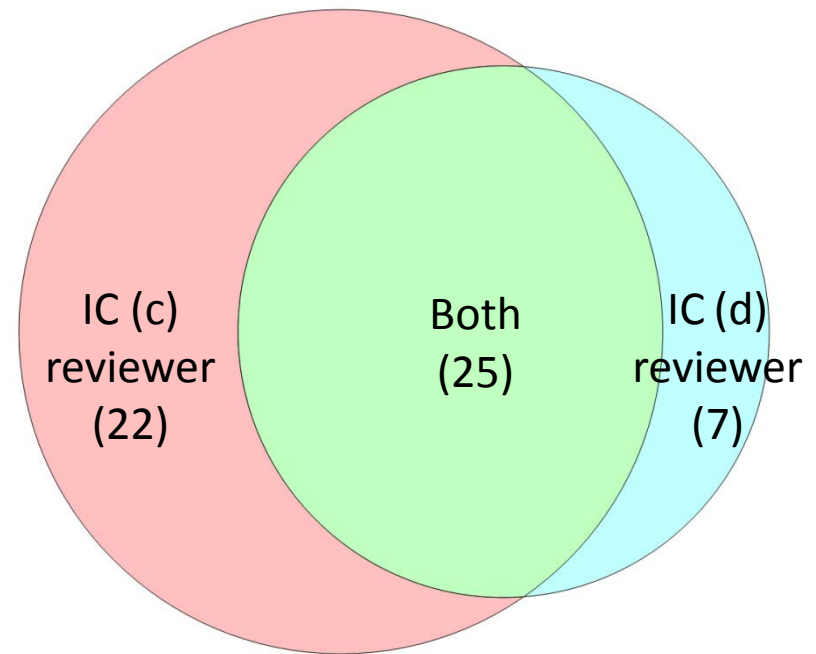
Mostly Distinct Clusters
(4840 R01s)



“Fuzzy” Cluster
(316 R01)



262 R01s resolved by
two SMEs



Summary

- The IN-SPIRE tool: multi-dimensional clustering of any set of documents
 - Categorize and dynamically analyze projects, publications, patents, et al.
 - Cluster within or between categories of documents
 - Cluster funded and/or unfunded applications, flag potential overlap
- By using IN-SPIRE, NIH staff can demonstrate that inappropriate overlap is minimal or non-existent
- OPA will continue to share databases, methods, and best practices in analyzing portfolios with IN-SPIRE and many other effective tools
 - Training / Consultation
 - Portfolio Analysis Interest Group (PAIG)
 - Blog (*The Analyst*)
- Tools in development: Track translational research (*iTrans*), effective bibliometrics (*iCite*), content analysis (*Portfolio Learning Tool*, *Semantic Analyst*)

Acknowledgments

OPA Analysts

Paula Fearon
Ian Hutchins
Jean Yuan
Carole Christian
Rob Harriman
Terry Bishop
Kristina McLinden
Geetha Senthil

OPA Software Developers

Kevin Small
Ehsan Haque
Fai Chan
Kirk Baker

OPA IT Specialist

Chuck Lynch

NIH Center for Information Technology

Calvin Johnson
Krishna Collie


NIH National Library of Medicine


Tom Rindflesch


Pacific Northwest National Labs

Dennis McQuerry

Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI)

 National Institutes of Health

 U.S. Department of Health and Human Services

 National Institutes of Health
Office of Portfolio Analysis

[Printer Friendly](#) | Text Size [A](#) [A](#) [A](#)

SEARCH

■ OPA HOME

■ TRAINING

■ TOOLS LAB

■ TOOLS DATABASE


■ MEETINGS

■ OPA STAFF

The [OPA Tools Lab](#) is located in B301 in building 1. For access please contact us.

For updates on training and other OPA activities, please sign up for our [listserv](#).

► [OPA analysis of Nicholson and Ioannidis dataset](#)



DPCPSI Home > OPA

The Office of Portfolio Analysis (OPA) was established in 2011 to provide NIH staff with multiple services, including consultation and training in the use of tools that will allow interested individuals to do their own analyses.

OPA Mission Statement:

Our purpose is to enhance the impact of NIH-supported research by enabling NIH research administrators and decision makers to evaluate and prioritize current, as well as emerging, areas of research that will advance knowledge and improve human health.

Our business is:

- To innovate by identifying and developing new, sophisticated tools that expand and improve NIH-wide efforts in portfolio analysis (PA);
- To apply and disseminate both current and newly developed tools, including computational approaches, which are capable of analyzing a wide range of parameters of biomedical research funding and the resulting impact;
- To promote trans-NIH coordination of PA activities and enhance collaboration among all PA stakeholders at NIH.

Our values are to be approachable, goal-oriented, and to provide services with the highest level of integrity and transparency.

Contact Us

Please let us know if you have comments or suggestions about portfolio analysis at NIH, or if you are interested in asking questions about current or emerging areas of research, especially if you aren't sure how to get reliable answers...
[click here](#) to contact us.