SLAM MICHIGAN

STUDENT LEARNING AND ANALYTICS AT MICHIGAN

October 19, 2012: Interactive Large-Scale Data Analyses and Visualization for Learning

Krishna Madhavan, Assistant Professor,
School of Engineering Education, Purdue University

SLAM MICHIGAN®

STUDENT LEARNING AND ANALYTICS AT MICHIGAN

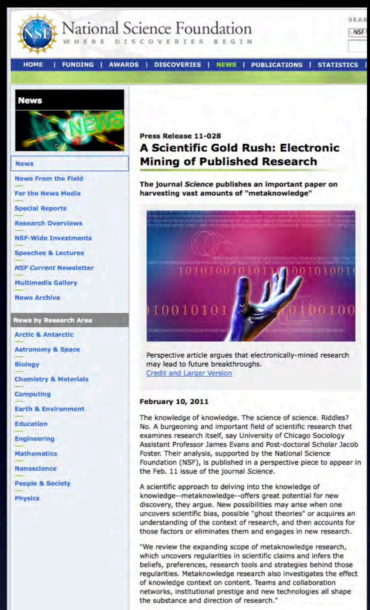
www.crlt.umich.edu/slam

Interactive Large Scale Analyses and Visualization for Learning

Krishna Madhavan
Gerhard Klimeck
Mihaela Vorvoreanu
Xin "Cindy" Chen
Purdue University, West Lafayette, IN

CYBERINFRASTRUCTURE
for engineering education

They say the next frontier is data...


National Science Foundation
WHERE DISCOVERIES BEGIN

HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS

News

News
News From the Field
For the News Media
Special Reports
Research Overviews
NSF-Wide Investments
Speeches & Lectures
NSF Current Newsletter
Multimedia Gallery
News Archive

News by Research Area
Arctic & Antarctic
Astronomy & Space
Biology
Chemistry & Materials
Computing
Earth & Environment
Education
Engineering
Mathematics
Nanoscience
People & Society
Physics

Press Release 11-028
A Scientific Gold Rush: Electronic Mining of Published Research

The journal *Science* publishes an important paper on harvesting vast amounts of "metaknowledge"

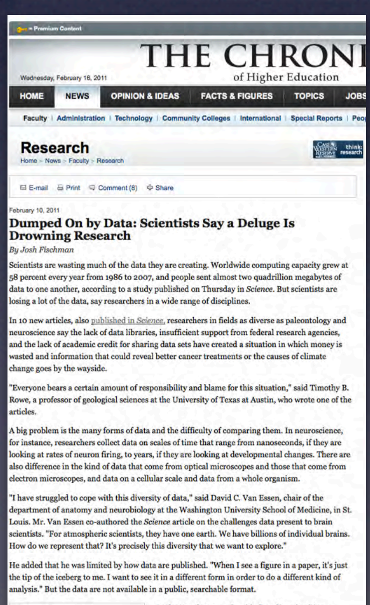
Perspective article argues that electronically-mined research may lead to future breakthroughs.
Credit and Larger Version

February 10, 2011

The knowledge of knowledge. The science of science. Riddles? No. A burgeoning and important field of scientific research that examines research itself, say University of Chicago Sociology Assistant Professor James Evans and Post-doctoral Scholar Jacob Foster. Their analysis, supported by the National Science Foundation (NSF), is published in a perspective piece to appear in the Feb. 11 issue of the journal *Science*.

A scientific approach to delving into the knowledge of knowledge--metaknowledge--offers great potential for new discovery, they argue. New possibilities may arise when one uncovers scientific bias, possible "ghost theories" or acquires an understanding of the context of research, and then accounts for those factors or eliminates them and engages in new research.

"We review the expanding scope of metaknowledge research, which uncovers regularities in scientific claims and infers the beliefs, preferences, research tools and strategies behind those regularities. Metaknowledge research also investigates the effect of knowledge context on content. Teams and collaboration networks, institutional prestige and new technologies all shape the substance and direction of research."


= Premium Content

THE CHRONI
of Higher Education
Wednesday, February 16, 2011

HOME | NEWS | OPINION & IDEAS | FACTS & FIGURES | TOPICS | JOBS

Faculty | Administration | Technology | Community Colleges | International | Special Reports | Peo

Research
Home > News > Faculty > Research

E-mail | Print | Comment (8) | Share

February 10, 2011
Dumped On by Data: Scientists Say a Deluge Is Drowning Research
By Josh Fischman

Scientists are wasting much of the data they are creating. Worldwide computing capacity grew at 58 percent every year from 1986 to 2007, and people sent almost two quadrillion megabytes of data to one another, according to a study published on Thursday in *Science*. But scientists are losing a lot of the data, say researchers in a wide range of disciplines.

In 10 new articles, also published in *Science*, researchers in fields as diverse as paleontology and neuroscience say the lack of data libraries, insufficient support from federal research agencies, and the lack of academic credit for sharing data sets have created a situation in which money is wasted and information that could reveal better cancer treatments or the causes of climate change goes by the wayside.

"Everyone bears a certain amount of responsibility and blame for this situation," said Timothy B. Rowe, a professor of geological sciences at the University of Texas at Austin, who wrote one of the articles.

A big problem is the many forms of data and the difficulty of comparing them. In neuroscience, for instance, researchers collect data on scales of time that range from nanoseconds, if they are looking at rates of neuron firing, to years, if they are looking at developmental changes. There are also difference in the kind of data that come from optical microscopes and those that come from electron microscopes, and data on a cellular scale and data from a whole organism.

"I have struggled to cope with this diversity of data," said David C. Van Essen, chair of the department of anatomy and neurobiology at the Washington University School of Medicine, in St. Louis. Mr. Van Essen co-authored the *Science* article on the challenges data present to brain scientists. "For atmospheric scientists, they have one earth. We have billions of individual brains. How do we represent that? It's precisely this diversity that we want to explore."

He added that he was limited by how data are published. "When I see a figure in a paper, it's just the tip of the iceberg to me. I want to see it in a different form in order to do a different kind of analysis." But the data are not available in a public, searchable format.

Ecologists also struggle with data diversity. "Some

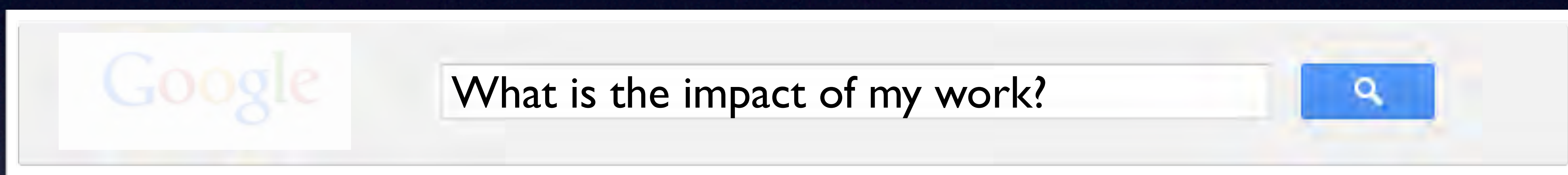
CYBERINFRASTRUCTURE
for engineering education

They say the next frontier is data...

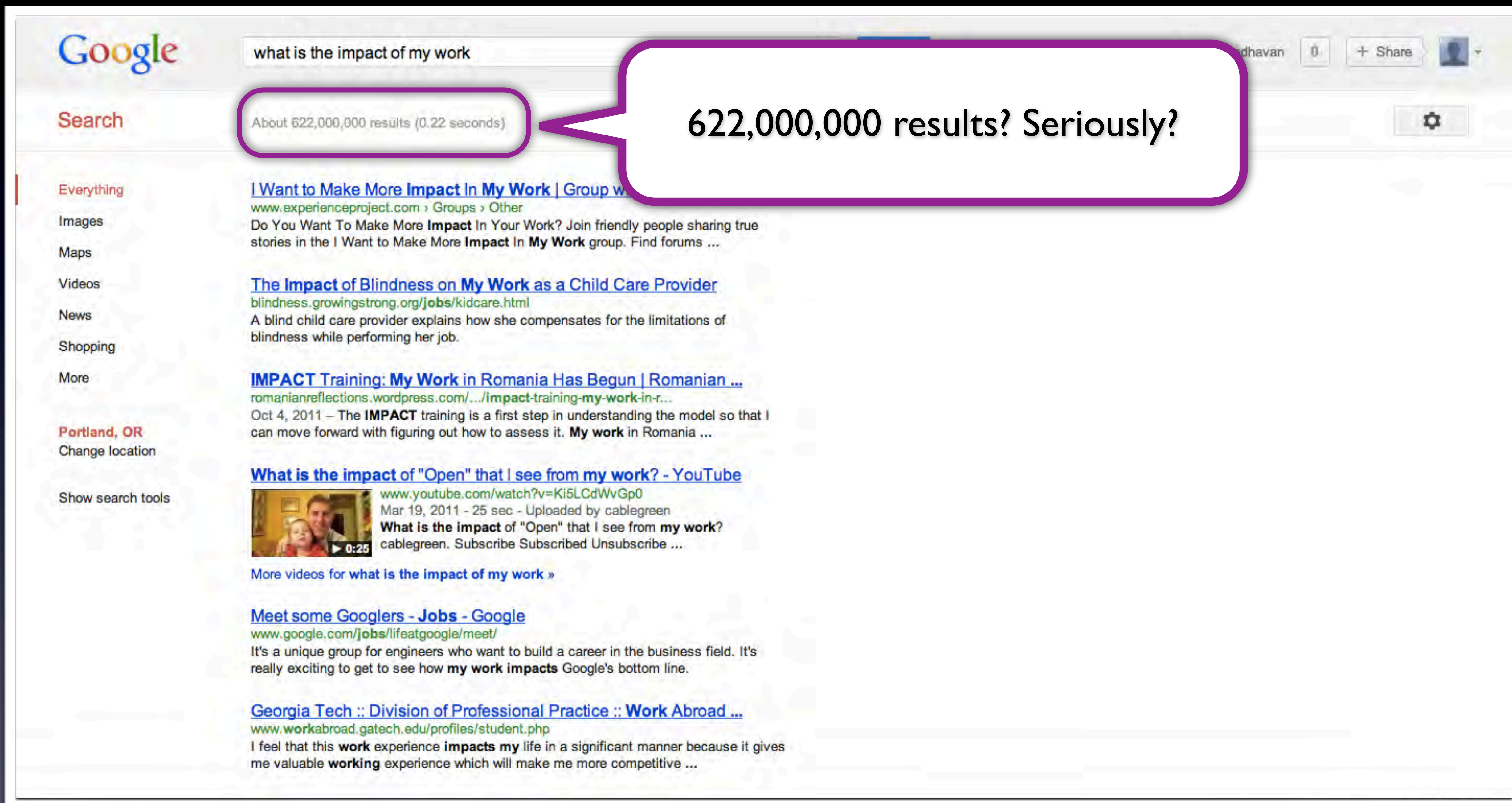
CYBERINFRASTRUCTURE
for engineering education

Data is useless...its like a genie stuck in the lamp...

No Mojo!



I can search the world's data - so what?



CYBERINFRASTRUCTURE for engineering education

What if I don't know where to start searching?

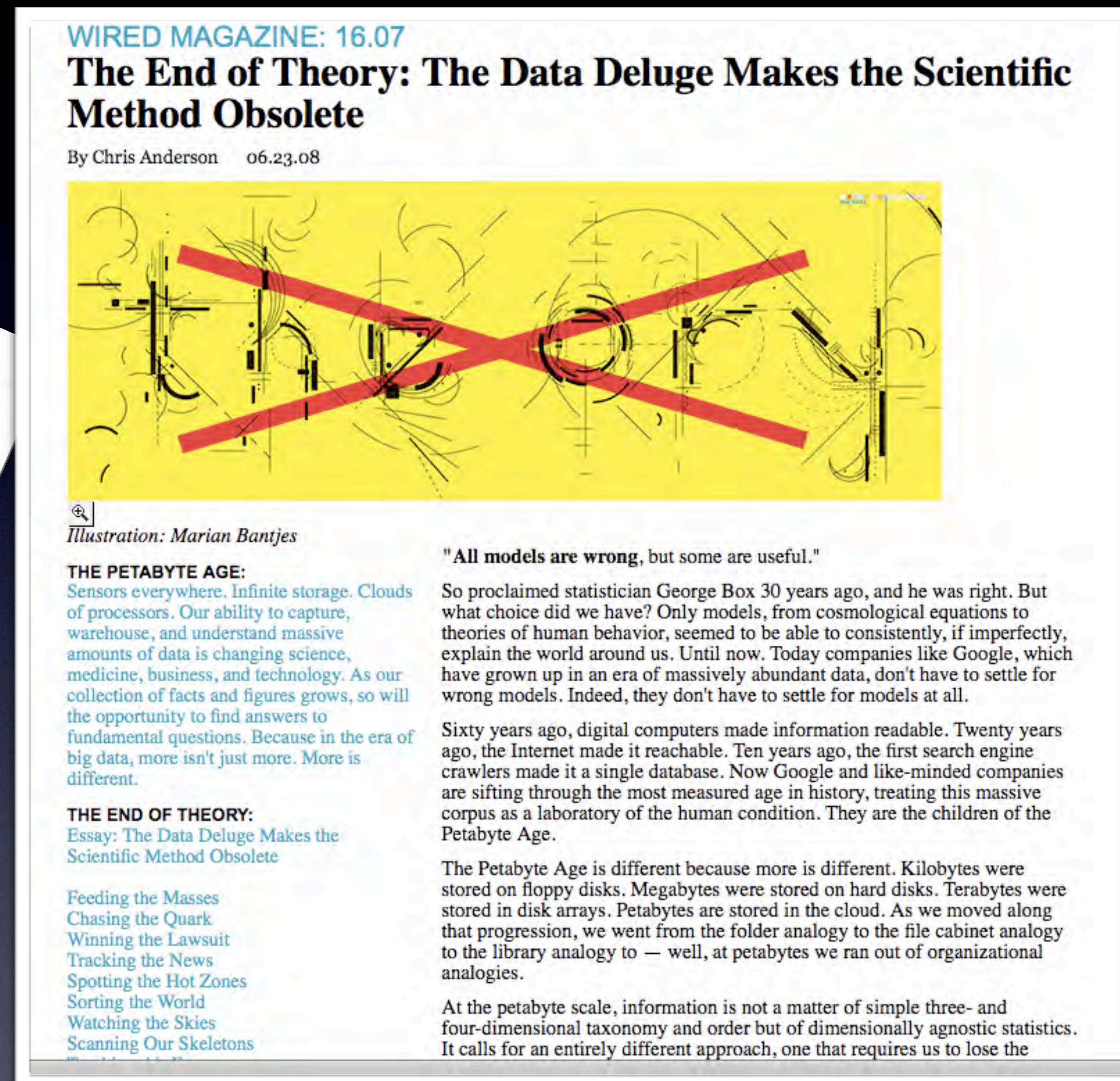
Number of results I rather get...

1

To me - nobody *really* cares about data...

People care about *sense making!*

The grand challenge is making sense of data



WIRED MAGAZINE: 16.07

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn't just more. More is different.

THE END OF THEORY:
Essay: The Data Deluge Makes the Scientific Method Obsolete

Feeding the Masses
Chasing the Quark
Winning the Lawsuit
Tracking the News
Spotting the Hot Zones
Sorting the World
Watching the Skies
Scanning Our Skeletons

"**All models are wrong**, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

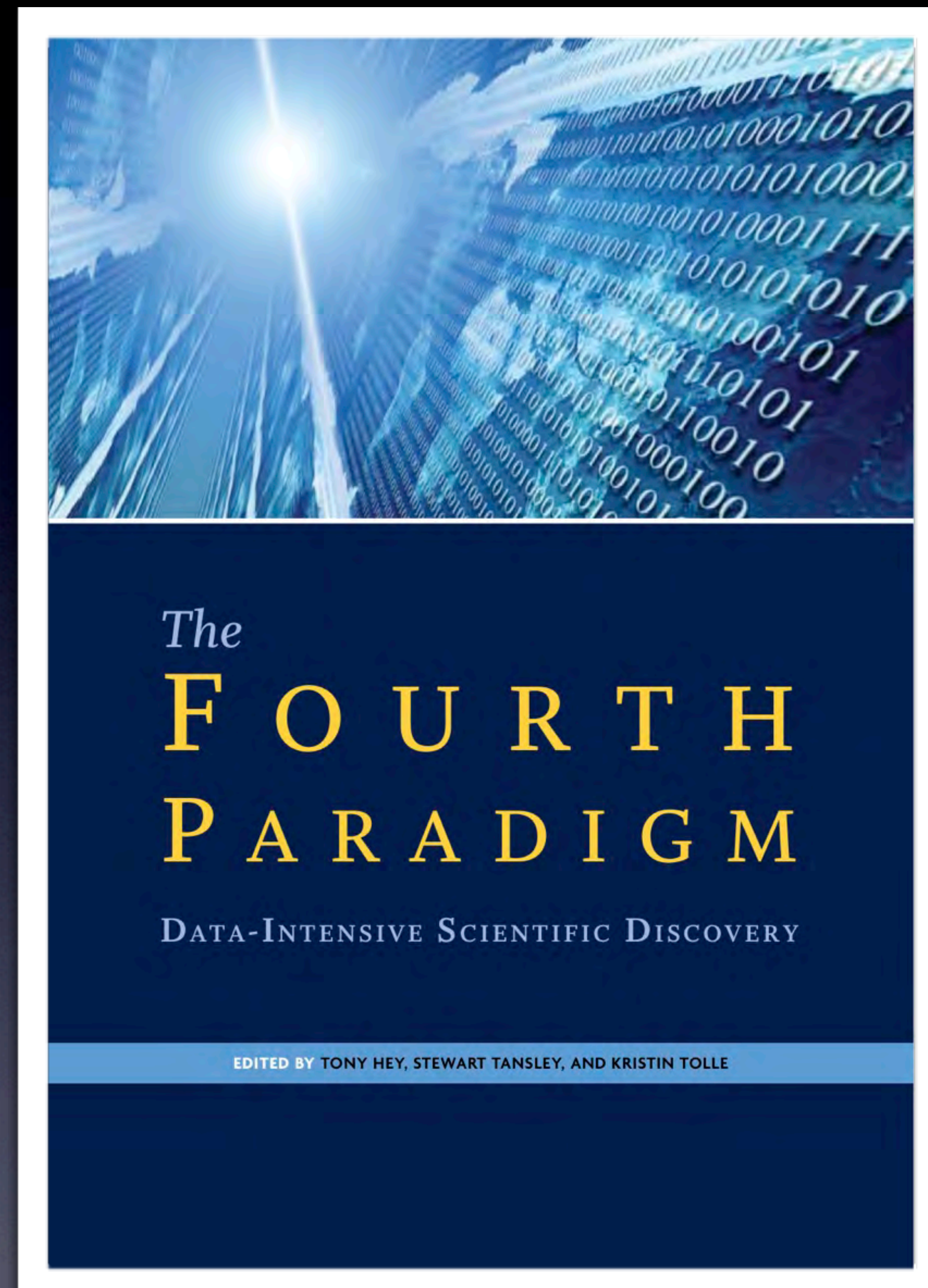
Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers made it a single database. Now Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the Petabyte Age.

The Petabyte Age is different because more is different. Kilobytes were stored on floppy disks. Megabytes were stored on hard disks. Terabytes were stored in disk arrays. Petabytes are stored in the cloud. As we moved along that progression, we went from the folder analogy to the file cabinet analogy to the library analogy to — well, at petabytes we ran out of organizational analogies.

At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the

CYBERINFRASTRUCTURE
for engineering education

The grand challenge is making sense of data



CYBERINFRASTRUCTURE
for engineering education

Next Frontier - Sense Making

= Premium Content

THE CHRON
of Higher Education

Wednesday, February 16, 2011

HOME | NEWS | OPINION & IDEAS | FACTS & FIGURES | TOPICS | JO

Faculty | Administration | Technology | Community Colleges | International | Special Reports | P

Research

Home > News > Faculty > Research

E-mail | Print | Comment (8) | Share

ebruary 10, 2011

Dumped On by Data: Scientists Say a Deluge Is Drowning Research

By Josh Fischman

cientists are wasting much of the data they are creating. Worldwide computing capacity grew at 8 percent every year from 1986 to 2007, and people sent almost two quadrillion megabytes of ata to one another, according to a study published on Thursday in *Science*. But scientists are osing a lot of the data, say researchers in a wide range of disciplines.

n 10 new articles, also published in *Science*, researchers in fields as diverse as paleontology and euroscience say the lack of data libraries, insufficient support from federal research agencies, nd the lack of academic credit for sharing data sets have created a situation in which money is wasted and information that could reveal better cancer treatments or the causes of climate hange goes by the wayside.

'Everyone bears a certain amount of responsibility and blame for this situation," said Timothy B. Rowe, a professor of geological sciences at the University of Texas at Austin, who wrote one of the rticles.

A big problem is the many forms of data and the difficulty of comparing them. In neuroscience, or instance, researchers collect data on scales of time that range from nanoseconds, if they are ooking at rates of neuron firing, to years, if they are looking at developmental changes. There are lso difference in the kind of data that come from optical microscopes and those that come from lectron microscopes, and data on a cellular scale and data from a whole organism.

'I have struggled to cope with this diversity of data," said David C. Van Essen, chair of the lepartment of anatomy and neurobiology at the Washington University School of Medicine, in St. Louis. Mr. Van Essen co-authored the *Science* article on the challenges data present to brain cientists. "For atmospheric scientists, they have one earth. We have billions of individual brains. How do we represent that? It's precisely this diversity that we want to explore."

He added that he was limited by how data are published. "When I see a figure in a paper, it's just he tip of the iceberg to me. I want to see it in a different form in order to do a different kind of nalysis." But the data are not available in a public, searchable format.

National Science Foundation
WHERE DISCOVERIES BEGIN

HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS |

News

News
News From the Field
For the News Media
Special Reports
Research Overviews
NSF-Wide Investments
Speeches & Lectures
NSF Current Newsletter
Multimedia Gallery
News Archive

News by Research Area
Arctic & Antarctic
Astronomy & Space
Biology
Chemistry & Materials
Computing
Earth & Environment
Education
Engineering
Mathematics
Nanoscience
People & Society
Physics

Press Release 11-028

A Scientific Gold Rush: Electronic Mining of Published Research

The journal *Science* publishes an important paper on harvesting vast amounts of "metaknowledge"

Perspective article argues that electronically-mined research may lead to future breakthroughs.
Credit and Larger Version

February 10, 2011

The knowledge of knowledge. The science of science. Riddles? No. A burgeoning and important field of scientific research that examines research itself, say University of Chicago Sociology Assistant Professor James Evans and Post-doctoral Scholar Jacob Foster. Their analysis, supported by the National Science Foundation (NSF), is published in a perspective piece to appear in the Feb. 11 issue of the journal *Science*.

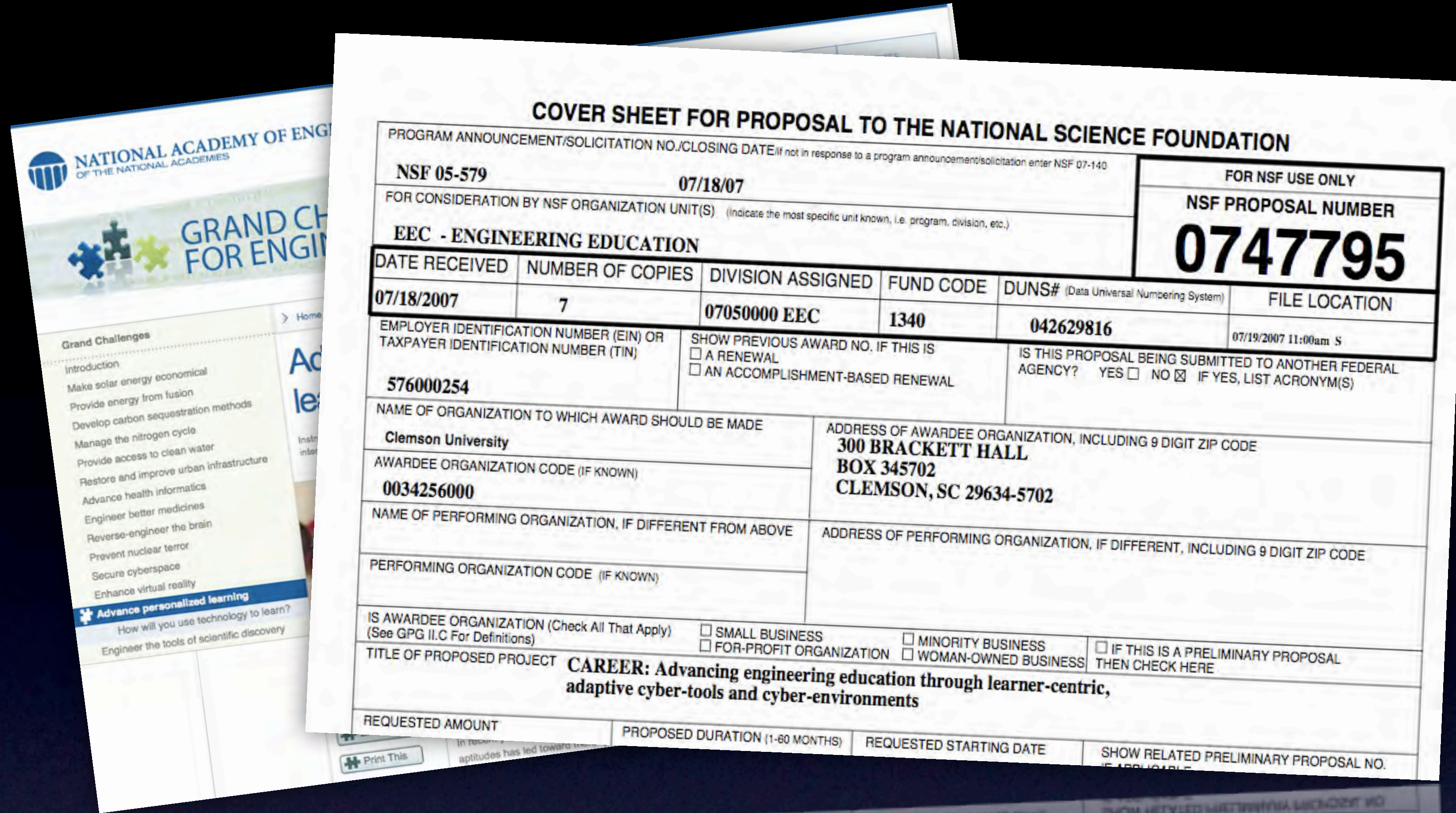
A scientific approach to delving into the knowledge of knowledge--metaknowledge--offers great potential for new discovery, they argue. New possibilities may arise when one uncovers scientific bias, possible "ghost theories" or acquires an understanding of the context of research, and then accounts for those factors or eliminates them and engages in new research.

"We review the expanding scope of metaknowledge research, which uncovers regularities in scientific claims and infers the beliefs, preferences, research tools and strategies behind those regularities. Metaknowledge research also investigates the effect of knowledge context on content. Teams and collaboration networks, institutional prestige and new technologies all shape the substance and direction of research."

CYBERINFRASTRUCTURE
for engineering education

"Advance Personalized Learning: Instruction can be individualized based on learning styles, speeds, and interests to make learning more reliable."

The National Academy of Engineering. "Grand Challenges for Engineering: Advance Personalized Learning." Available at http://www.engineeringchallenges.org/cms/8996/9127.aspx. (June 2008).



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 07-140

NSF 05-579 07/18/07

FOR NSF USE ONLY

NSF PROPOSAL NUMBER

0747795

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

EEC - ENGINEERING EDUCATION

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
07/18/2007	7	07050000 EEC	1340	042629816	07/19/2007 11:00am S

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

576000254

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

Clemson University

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE

300 BRACKETT HALL
BOX 345702
CLEMSON, SC 29634-5702

AWARDEE ORGANIZATION CODE (IF KNOWN)

0034256000

NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE

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

PERFORMING ORGANIZATION CODE (IF KNOWN)

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions) ☐ SMALL BUSINESS ☐ FOR-PROFIT ORGANIZATION ☐ MINORITY BUSINESS ☐ WOMAN-OWNED BUSINESS ☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT CAREER: Advancing engineering education through learner-centric, adaptive cyber-tools and cyber-environments

REQUESTED AMOUNT | PROPOSED DURATION (1-60 MONTHS) | REQUESTED STARTING DATE | SHOW RELATED PRELIMINARY PROPOSAL NO.

“Engineering education experiences of the future can center on students [...] with cyber-tools and cyber-environments (also known as cyberinfrastructure) acting in *well-choreographed harmony, adapting, and customizing themselves to individual learner needs and outcomes* [emphasis added].”

Madhavan, K.P.C. (2007). “CAREER: Advancing Engineering Education through Learner-centric, Adaptive Cyber-tools and Cyber-environments.” NSF CAREER Proposal. Submitted to NSF-EEC.

Defining the problem

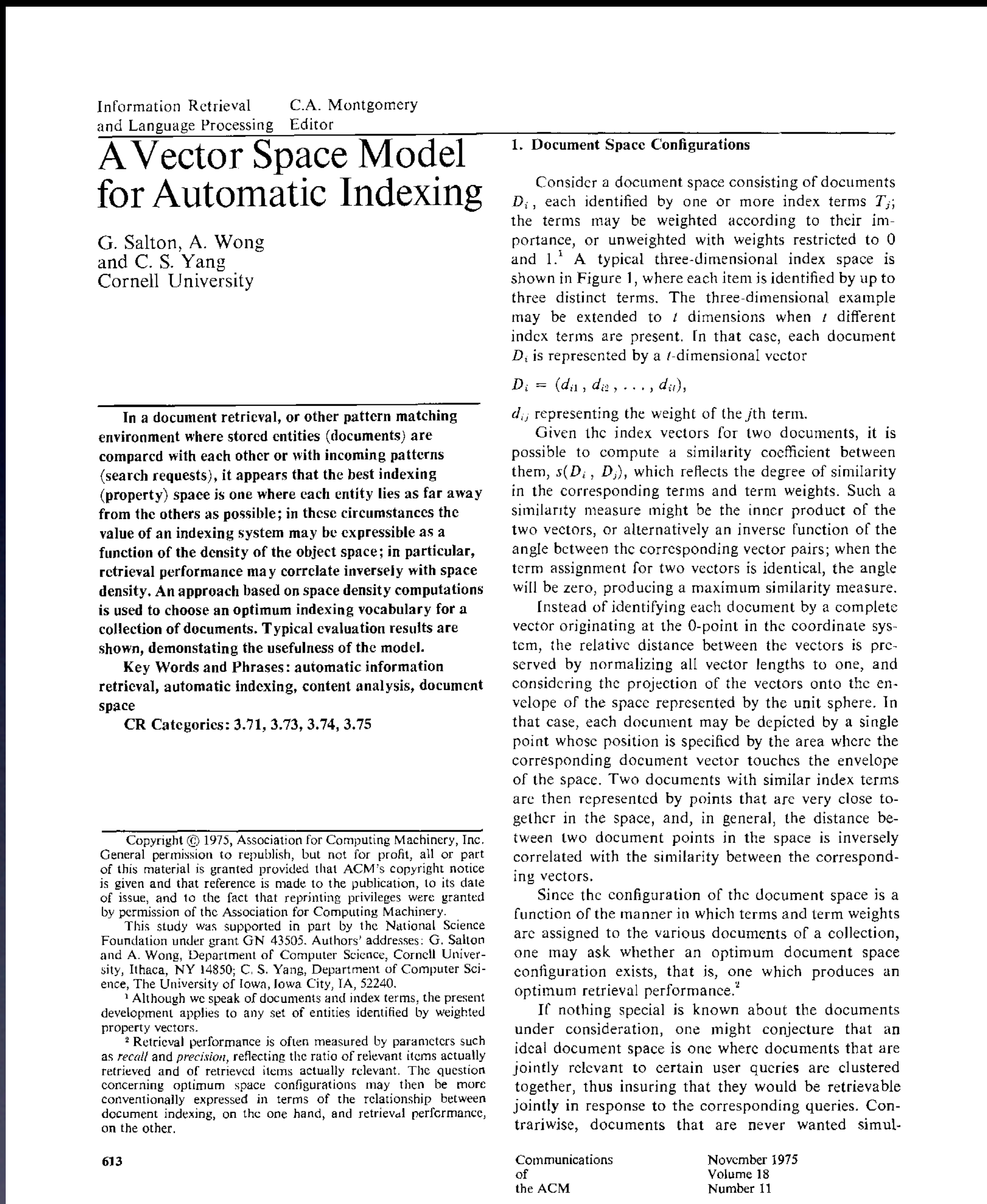
There is so much work on intelligent tutors, teaching tutor agents, recommender systems, etc.

So, why is personalized learning such a big deal? What is the role of learning analytics in tackling this grand challenge?

How can we derive *actionable intelligence* if you don't know much *data* about your users/learners?

CYBERINFRASTRUCTURE for engineering education

Personalized Learning, Basic Trigonometry, Matrices



Information Retrieval and Language Processing
C.A. Montgomery Editor

A Vector Space Model for Automatic Indexing

G. Salton, A. Wong and C. S. Yang
Cornell University

In a document retrieval, or other pattern matching environment where stored entities (documents) are compared with each other or with incoming patterns (search requests), it appears that the best indexing (property) space is one where each entity lies as far away from the others as possible; in these circumstances the value of an indexing system may be expressible as a function of the density of the object space; in particular, retrieval performance may correlate inversely with space density. An approach based on space density computations is used to choose an optimum indexing vocabulary for a collection of documents. Typical evaluation results are shown, demonstating the usefulness of the model.

Key Words and Phrases: automatic information retrieval, automatic indexing, content analysis, document space

CR Categories: 3.71, 3.73, 3.74, 3.75

This study was supported in part by the National Science Foundation under grant GN 43505. Authors' addresses: G. Salton and A. Wong, Department of Computer Science, Cornell University, Ithaca, NY 14850; C. S. Yang, Department of Computer Science, The University of Iowa, Iowa City, IA, 52240.

[1] Although we speak of documents and index terms, the present development applies to any set of entities identified by weighted property vectors.

[2] Retrieval performance is often measured by parameters such as *recall* and *precision*, reflecting the ratio of relevant items actually retrieved and of retrieved items actually relevant. The question concerning optimum space configurations may then be more conventionally expressed in terms of the relationship between document indexing, on the one hand, and retrieval performance, on the other.

1. Document Space Configurations

Consider a document space consisting of documents D_i, each identified by one or more index terms T_j; the terms may be weighted according to their importance, or unweighted with weights restricted to 0 and 1.[1] A typical three-dimensional index space is shown in Figure 1, where each item is identified by up to three distinct terms. The three-dimensional example may be extended to t dimensions when t different index terms are present. In that case, each document D_i is represented by a t-dimensional vector

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{it}),$$

d_{ij} representing the weight of the jth term.

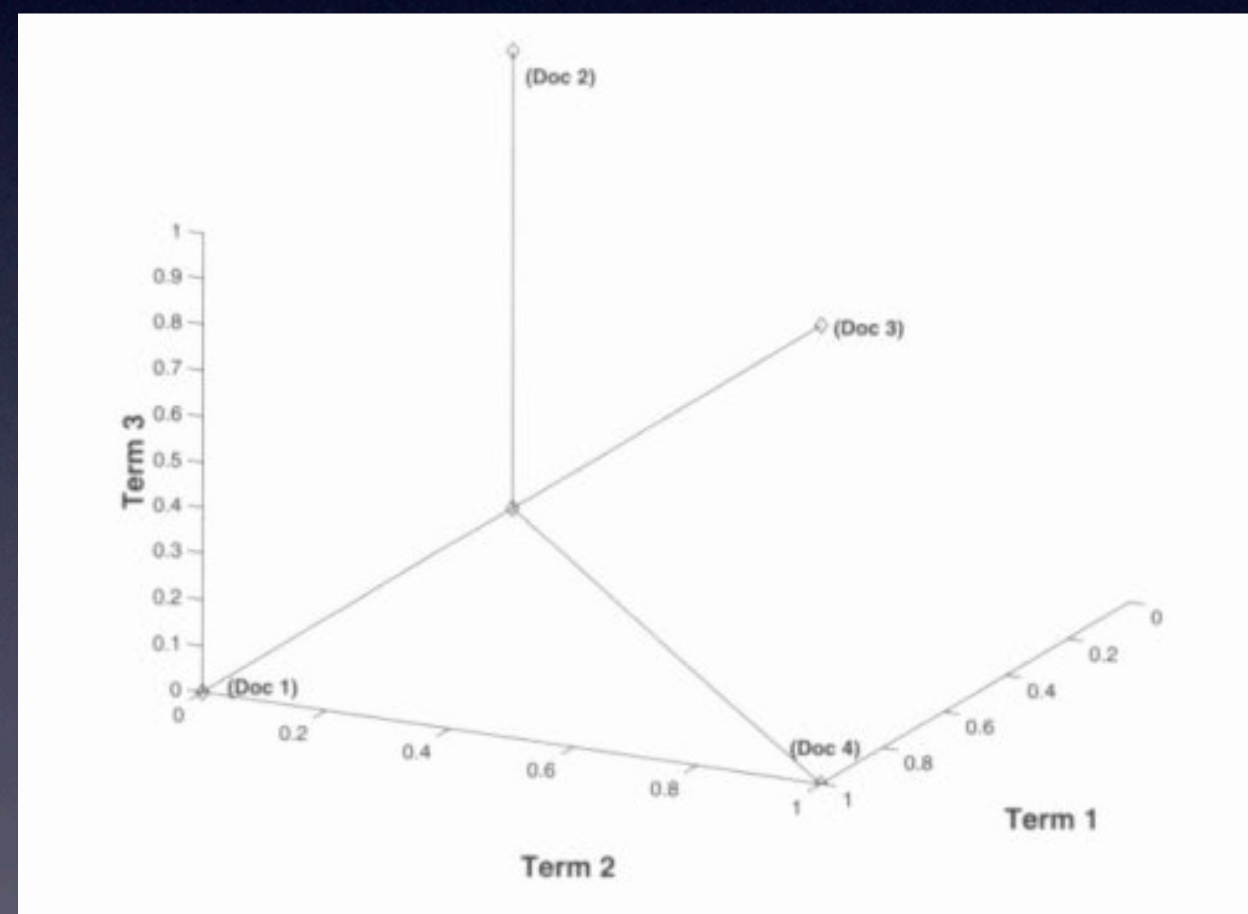
Given the index vectors for two documents, it is possible to compute a similarity coefficient between them, $s(D_i, D_j)$, which reflects the degree of similarity in the corresponding terms and term weights. Such a similarity measure might be the inner product of the two vectors, or alternatively an inverse function of the angle between the corresponding vector pairs; when the term assignment for two vectors is identical, the angle will be zero, producing a maximum similarity measure.

Instead of identifying each document by a complete vector originating at the 0-point in the coordinate system, the relative distance between the vectors is preserved by normalizing all vector lengths to one, and considering the projection of the vectors onto the envelope of the space represented by the unit sphere. In that case, each document may be depicted by a single point whose position is specified by the area where the corresponding document vector touches the envelope of the space. Two documents with similar index terms are then represented by points that are very close together in the space, and, in general, the distance between two document points in the space is inversely correlated with the similarity between the corresponding vectors.

Since the configuration of the document space is a function of the manner in which terms and term weights are assigned to the various documents of a collection, one may ask whether an optimum document space configuration exists, that is, one which produces an optimum retrieval performance.[2]

If nothing special is known about the documents under consideration, one might conjecture that an ideal document space is one where documents that are jointly relevant to certain user queries are clustered together, thus insuring that they would be retrievable jointly in response to the corresponding queries. Contrariwise, documents that are never wanted simul-

[k x n]	Learner 1	Learner 2	Learner 3	Learner n
Data 1	1	1	0	1
Data 2	0	0	0	1
Data 3	0	1	0	1
Data k	1	0	1	1

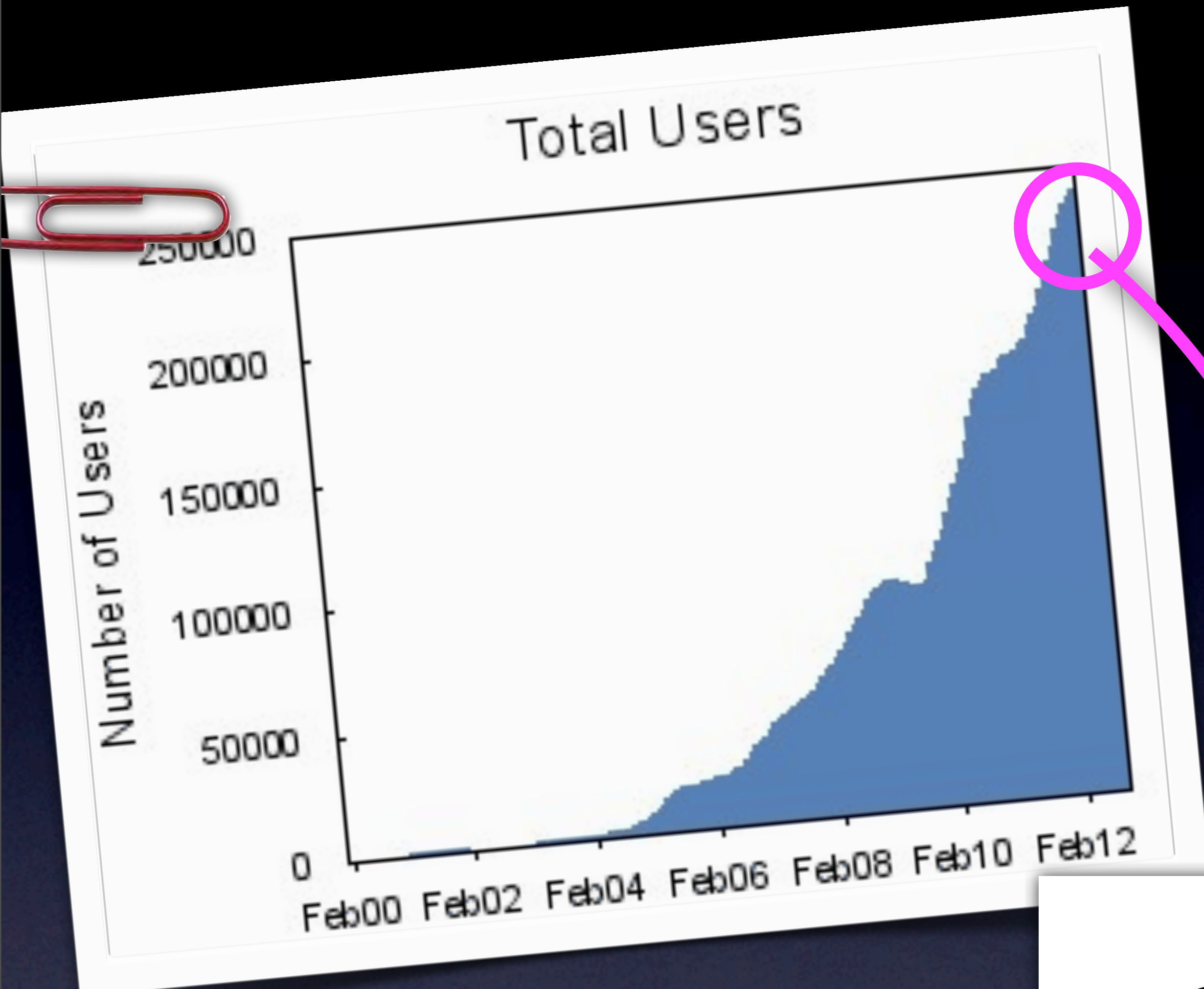


Source: Berry, M.W. and Browne, M. (2005). Understanding Search Engines: Mathematical Modeling and Software Retrieval (Software Environments, Tools, 2nd Edition)

Case Study - nanoHUB.org

Platform Perspective

Setting the context - nanoHUB.org



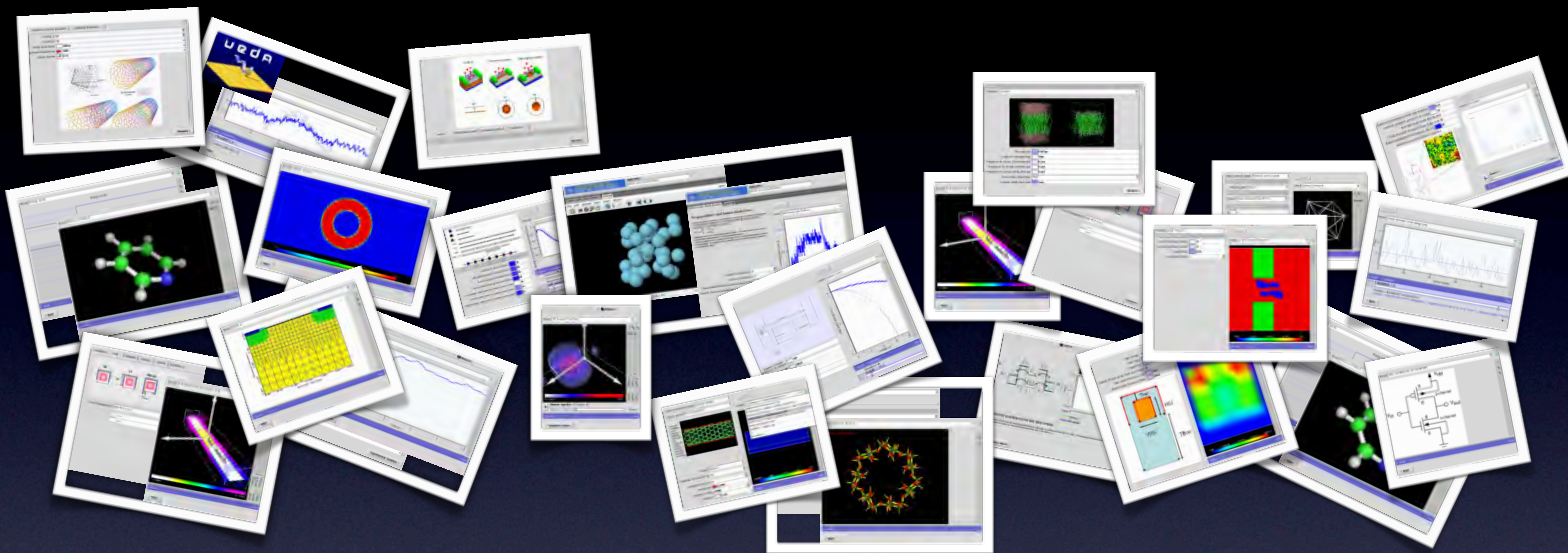
nanoHUB.org

9 years

>240,000 users *annually*

CYBERINFRASTRUCTURE for engineering education

Setting the context - nanoHUB.org



Over **260** tools

Easy to use and intuitive user-contributed tools

Setting the context - nanoHUB.org

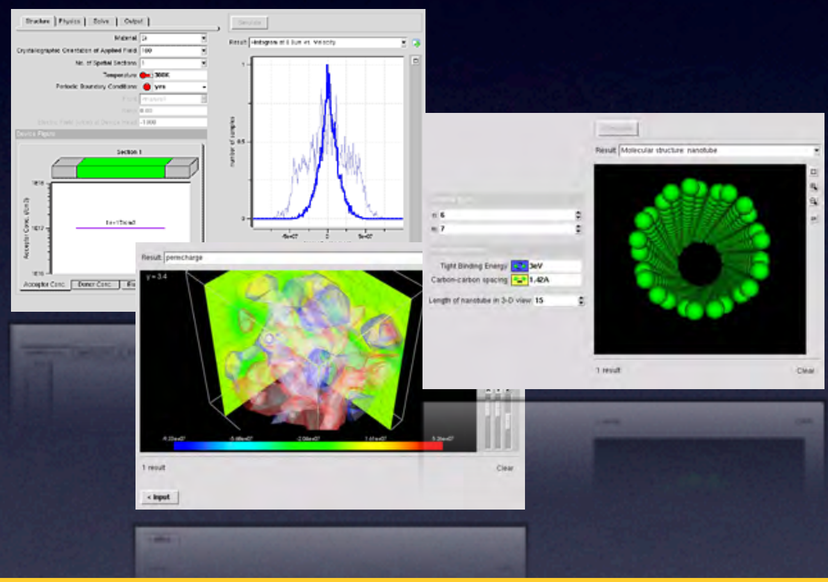
Over
3,400
resources

User Contributed

Direct Impact
RESEARCH to LEARNING

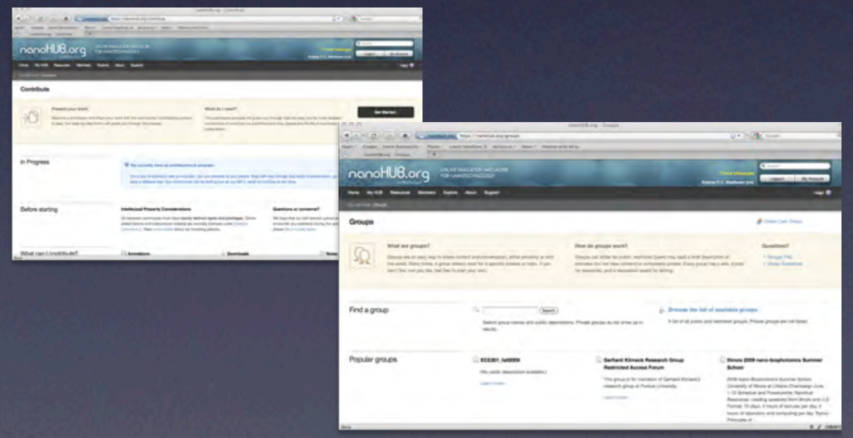
CYBERINFRASTRUCTURE for engineering education

Online Simulations



and More...

Community



Instrumenting the environment (e-infrastructure) holds the key nanoHUB.org

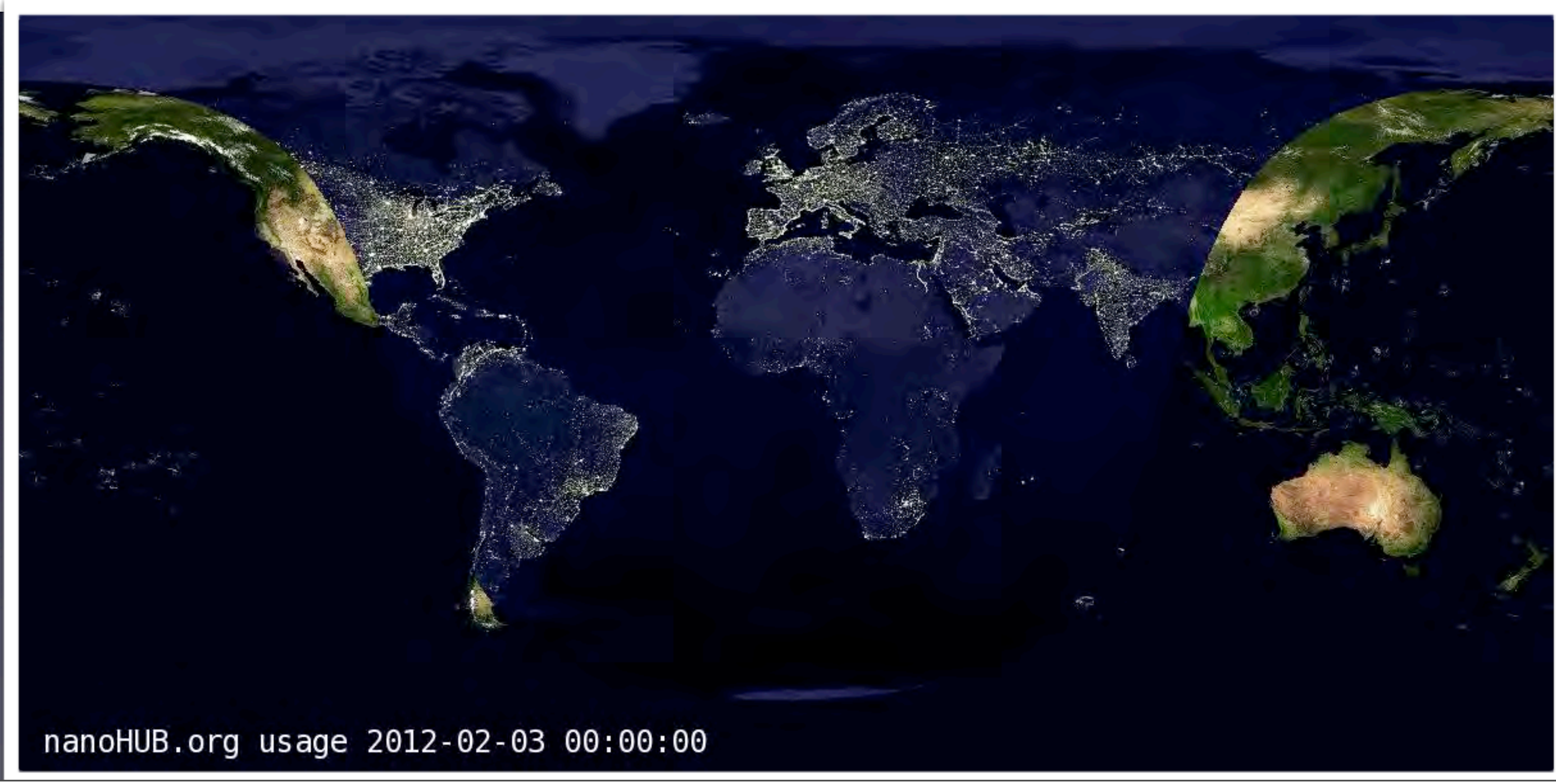
nano App Store

172 Countries worldwide

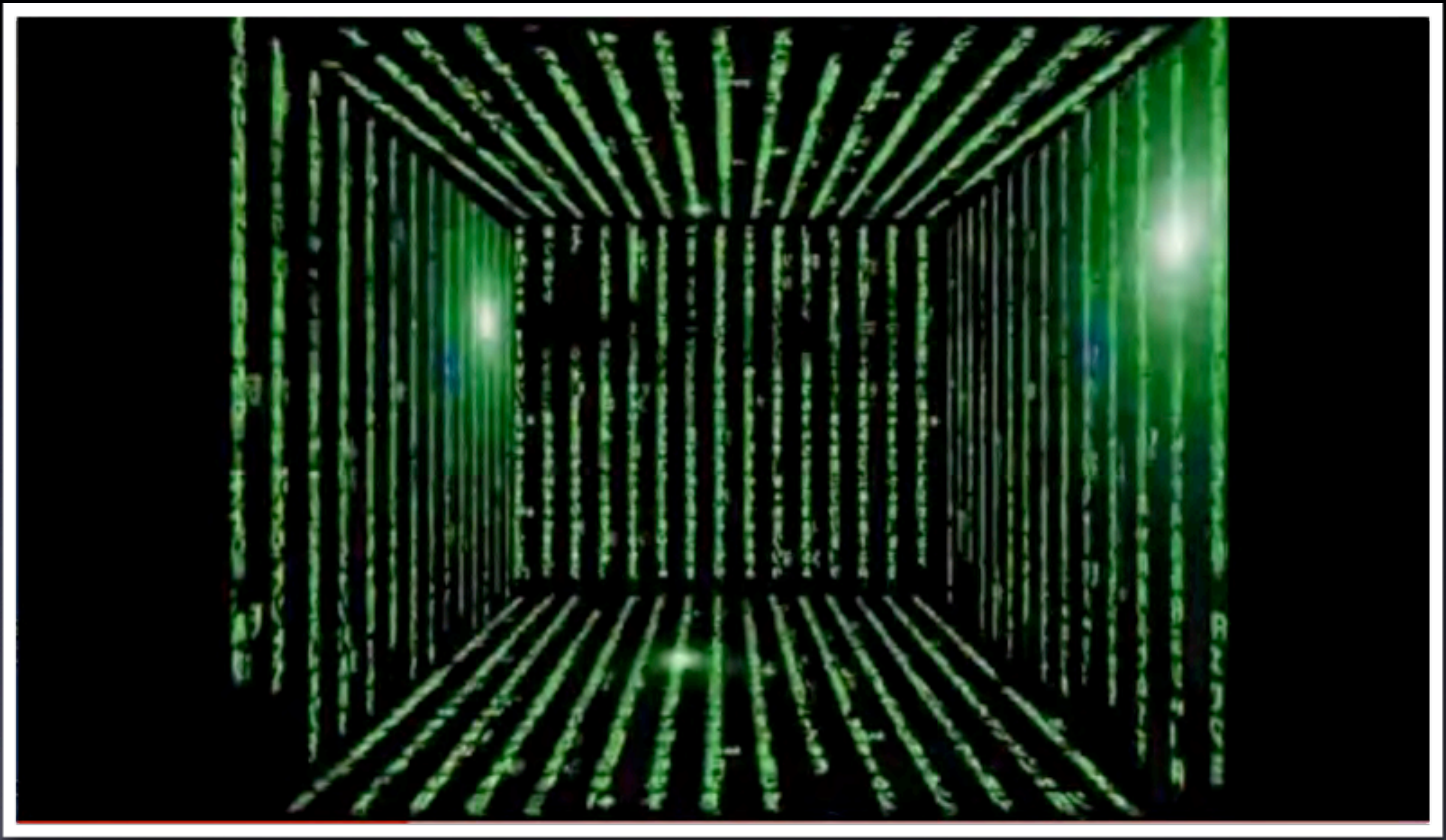
As much traffic as http://www.purdue.edu

Users at all top 50 US Engr Schools
Worldwide 19% of all .edu domains

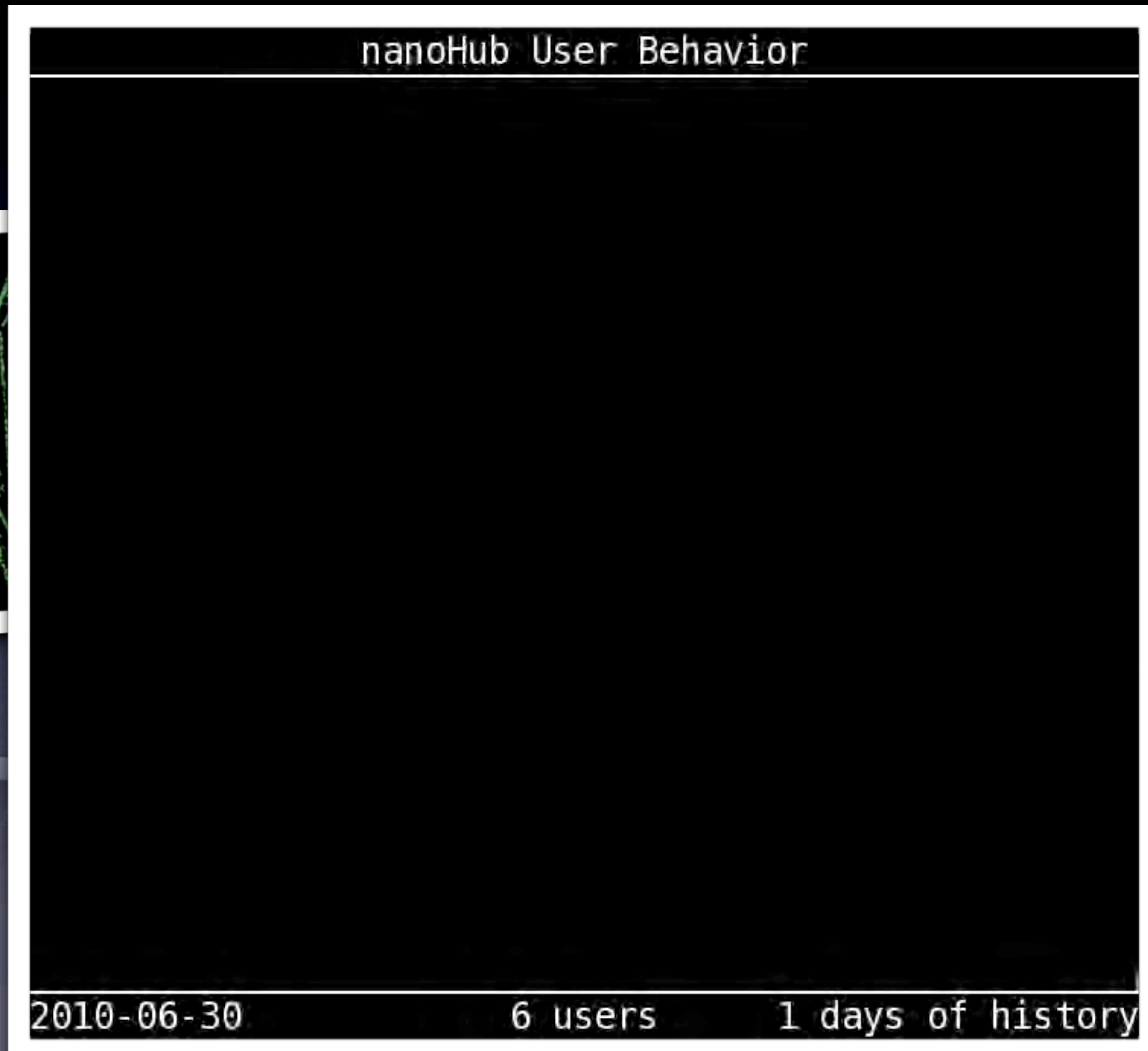
- New Registrations
- Simulation Users
- Tutorial / Lecture Users



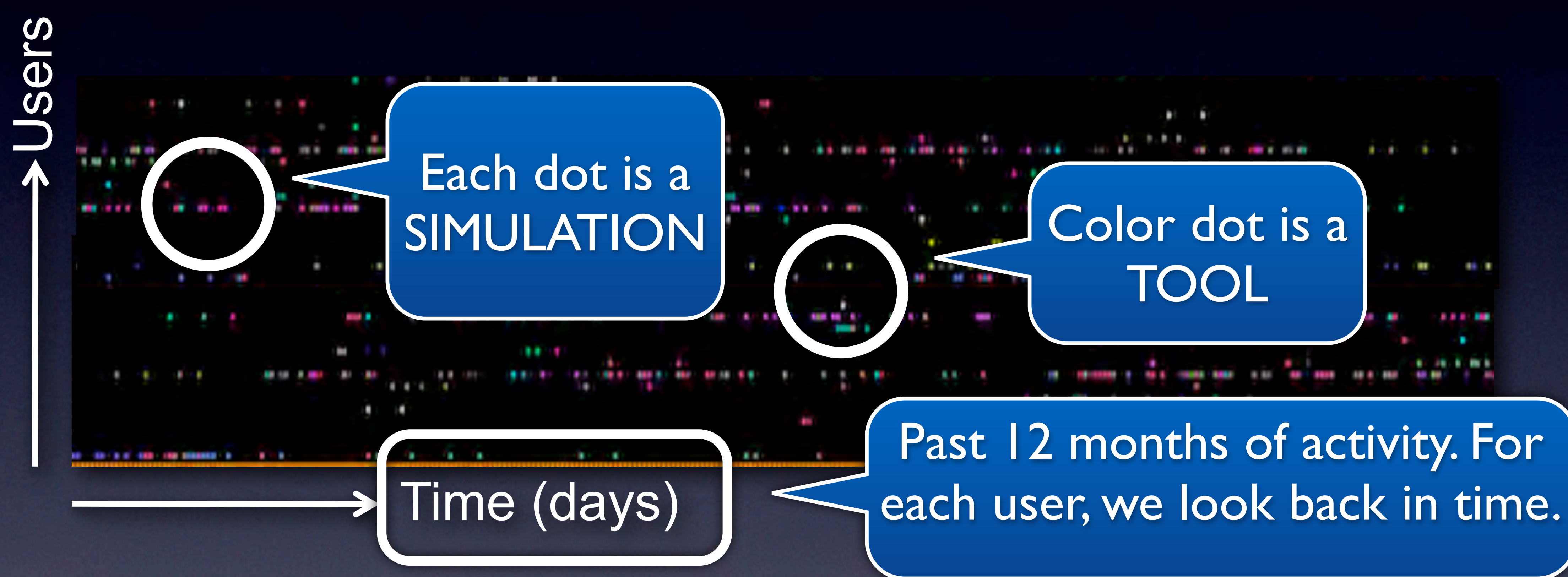
The Matrix (Movie)



nanoHUB User Matrix

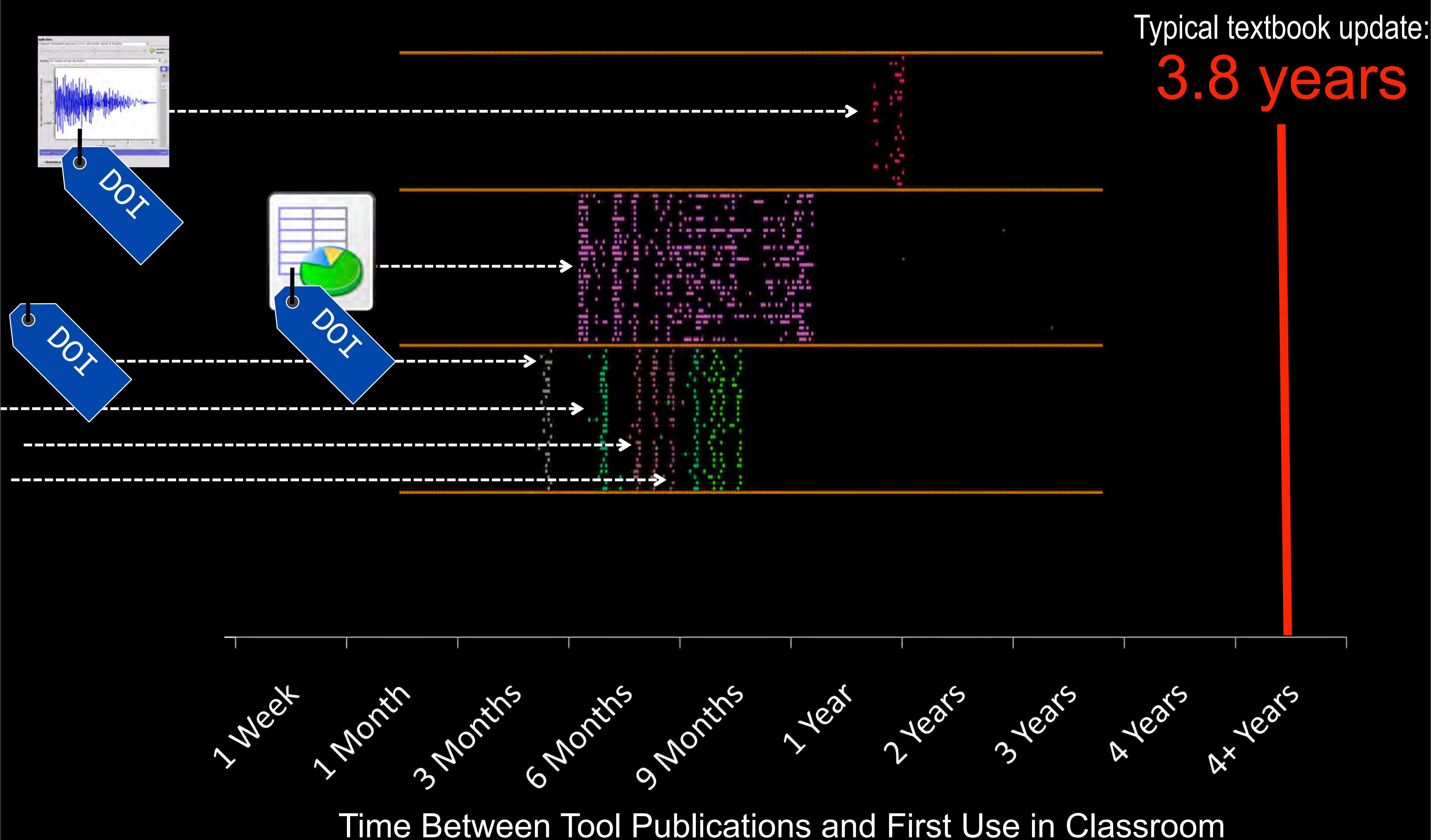


Slowing Down

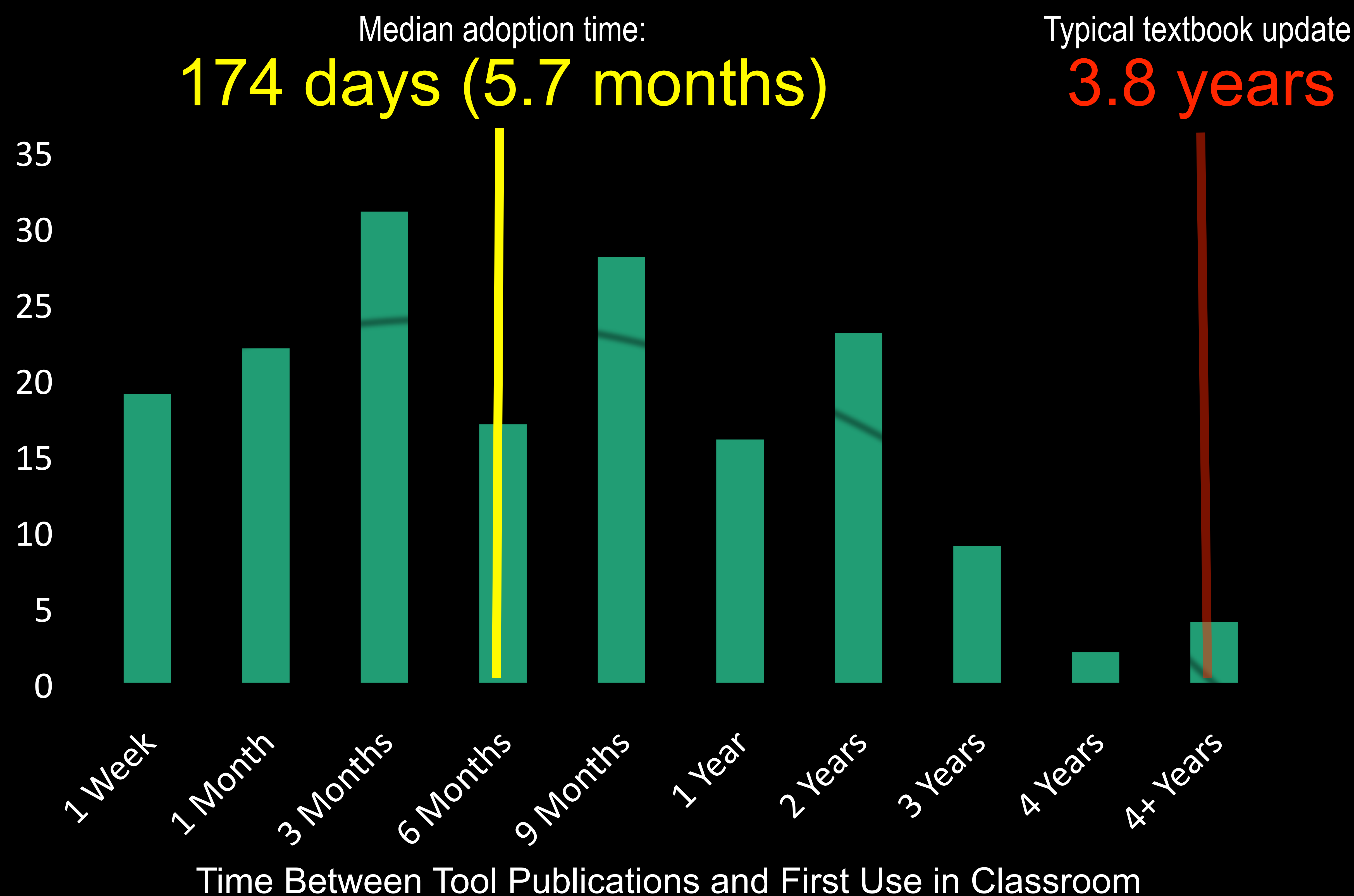


For each user we plot ALL simulation tool activities over the past 12 months

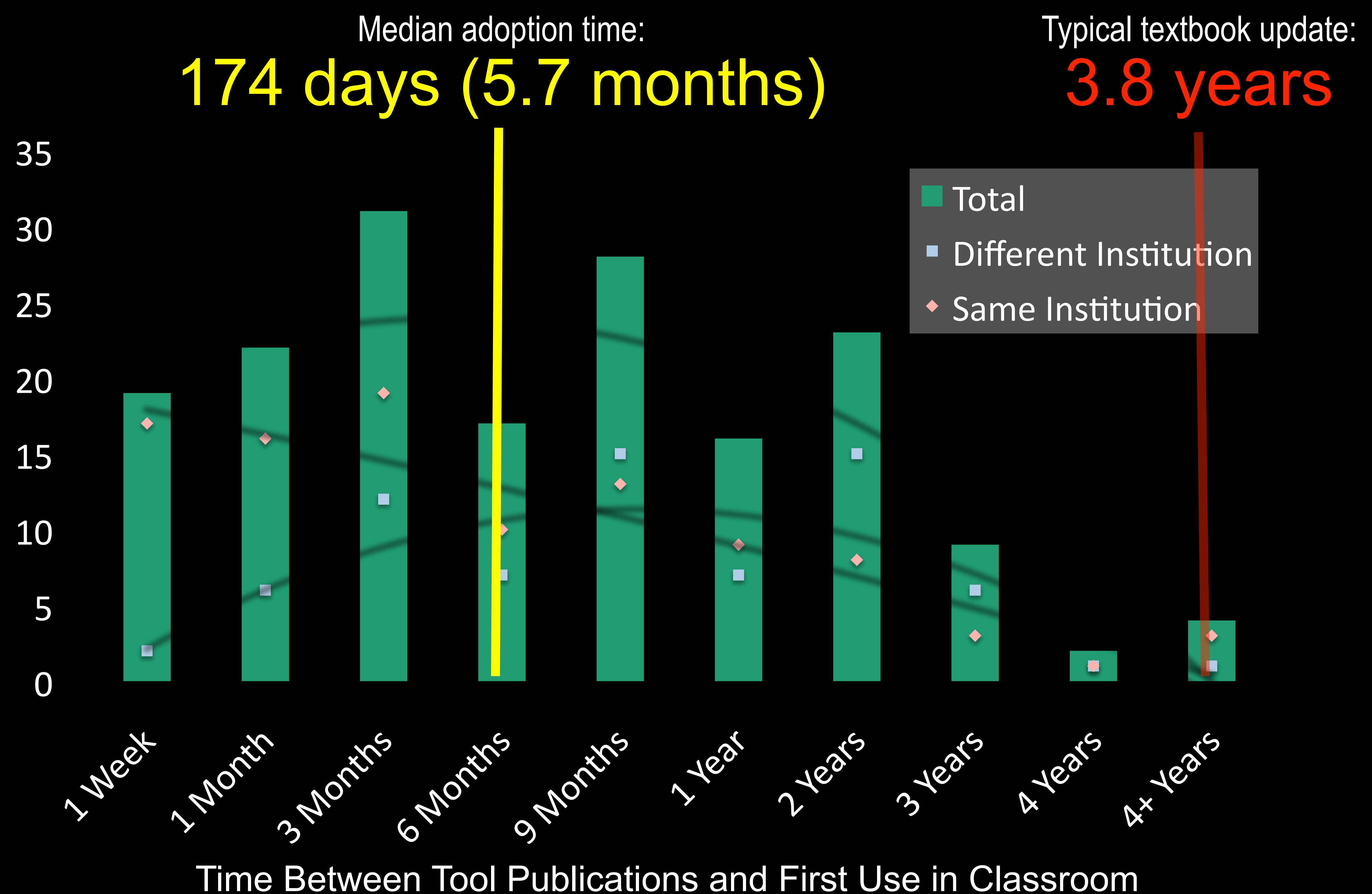
Time to First Adoption



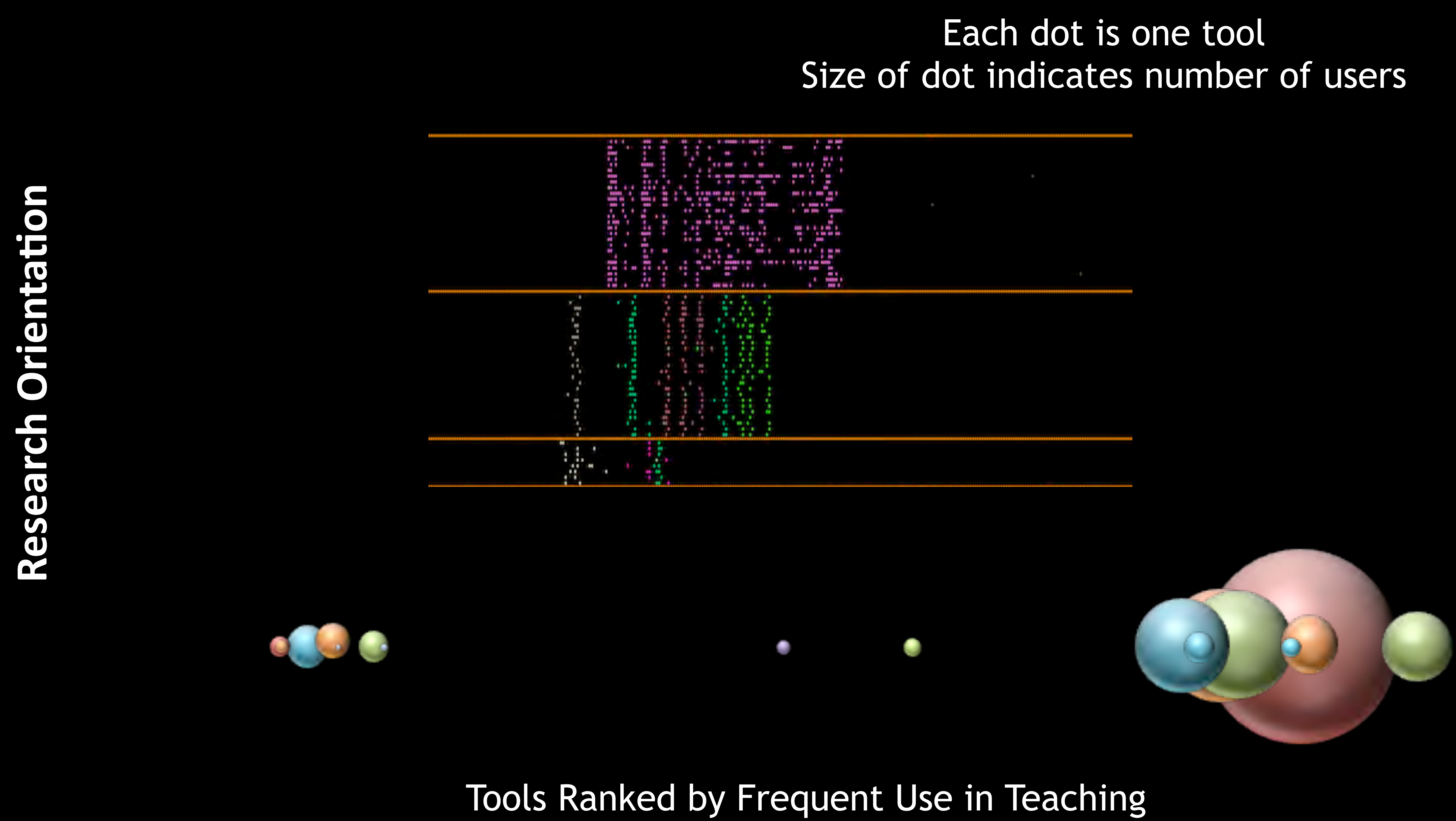
Rapid Adoption of Research



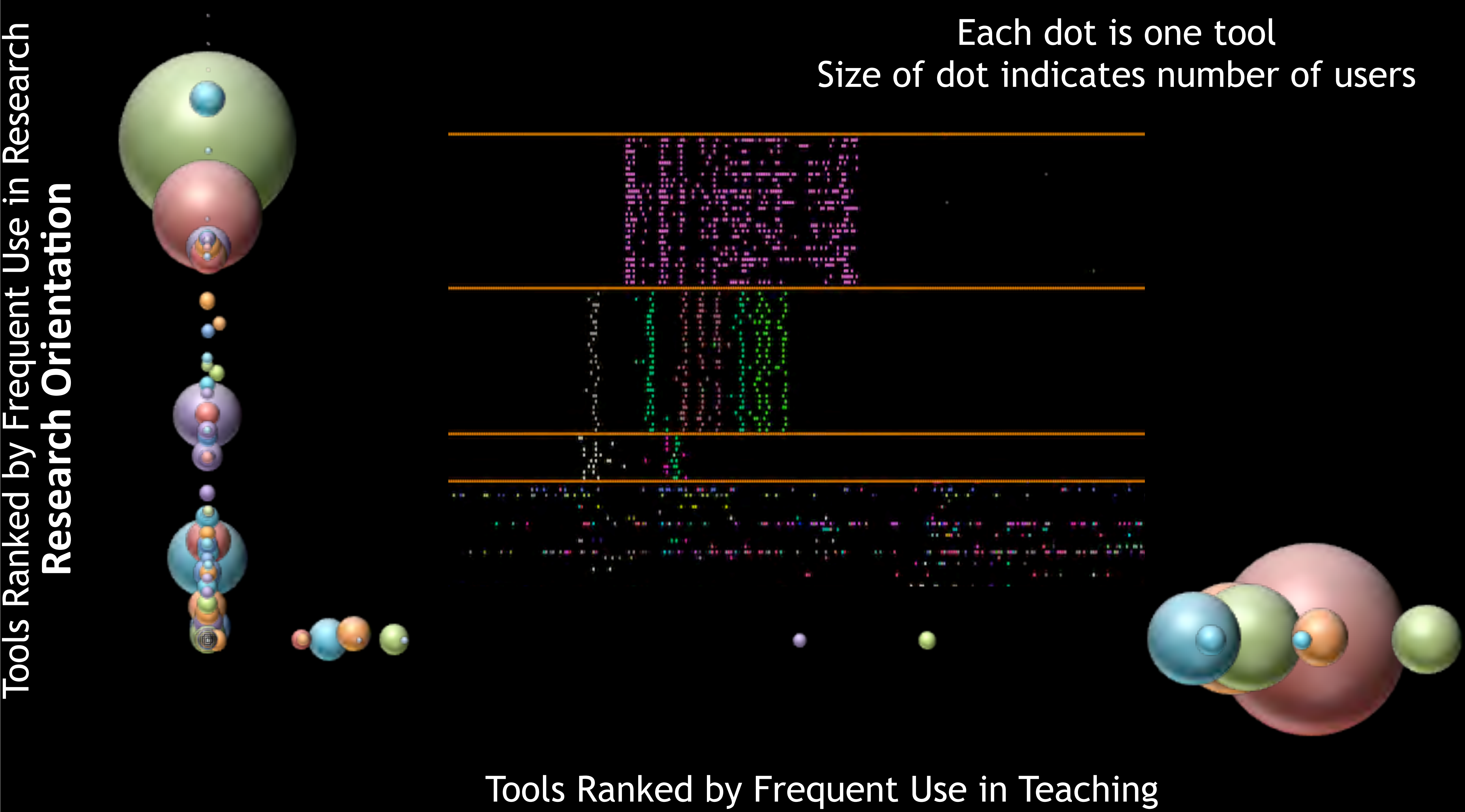
Revolutionizing Research → Classroom



Usage Patterns => Tool Qualification



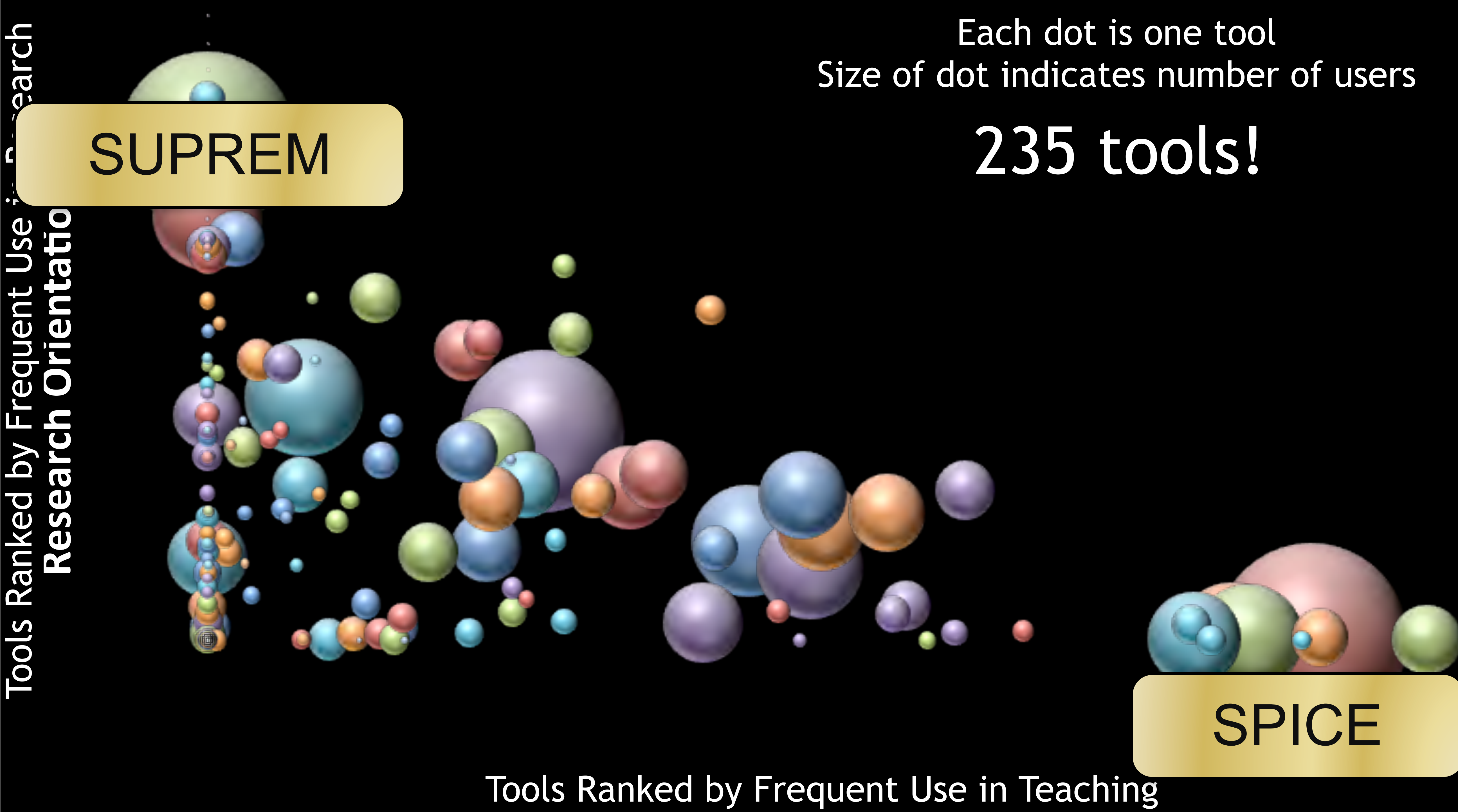
Usage Patterns => Tool Qualification



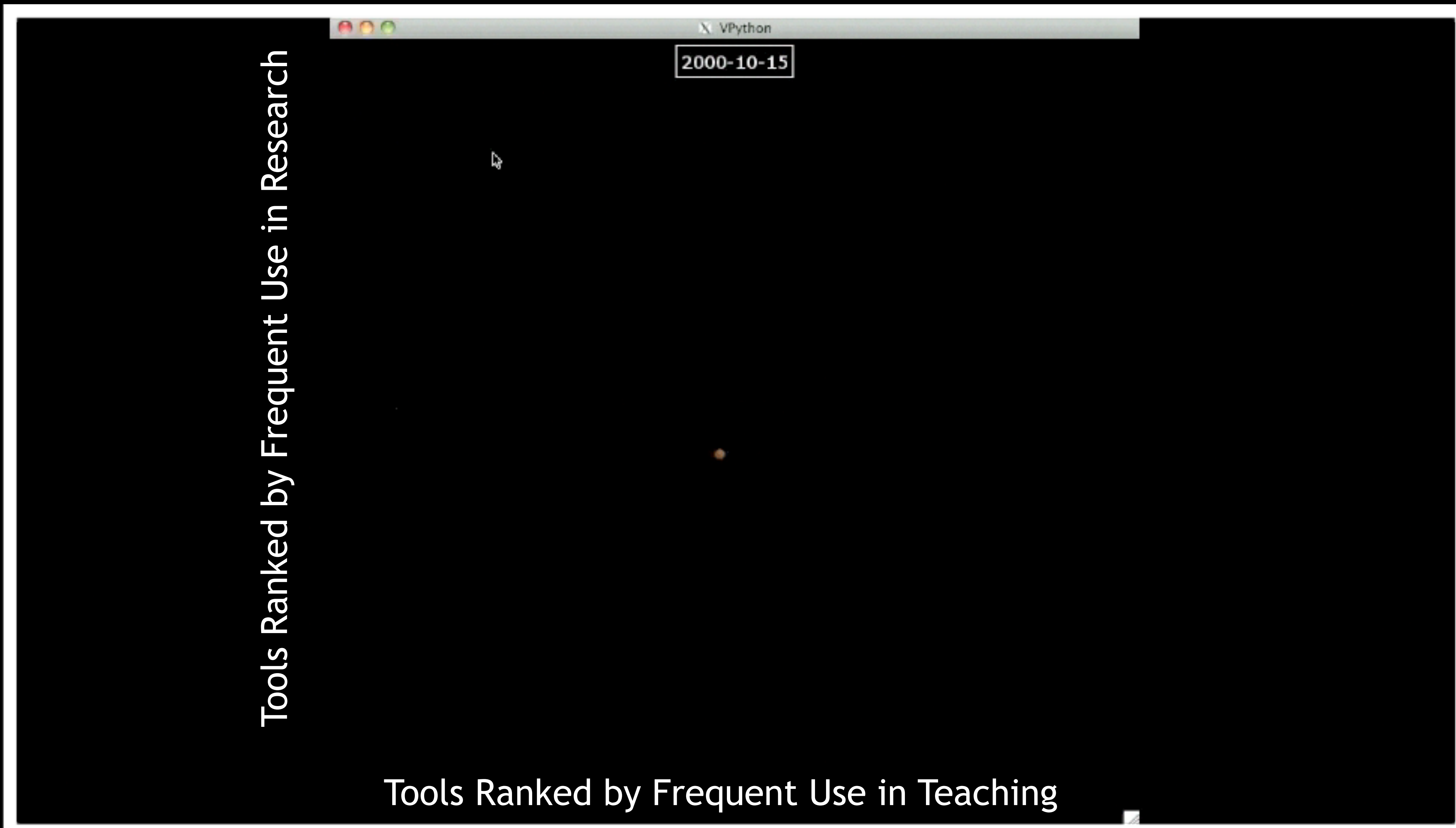
Dual Use
Education and Research are coupled!

Each dot is one tool
Size of dot indicates number of users

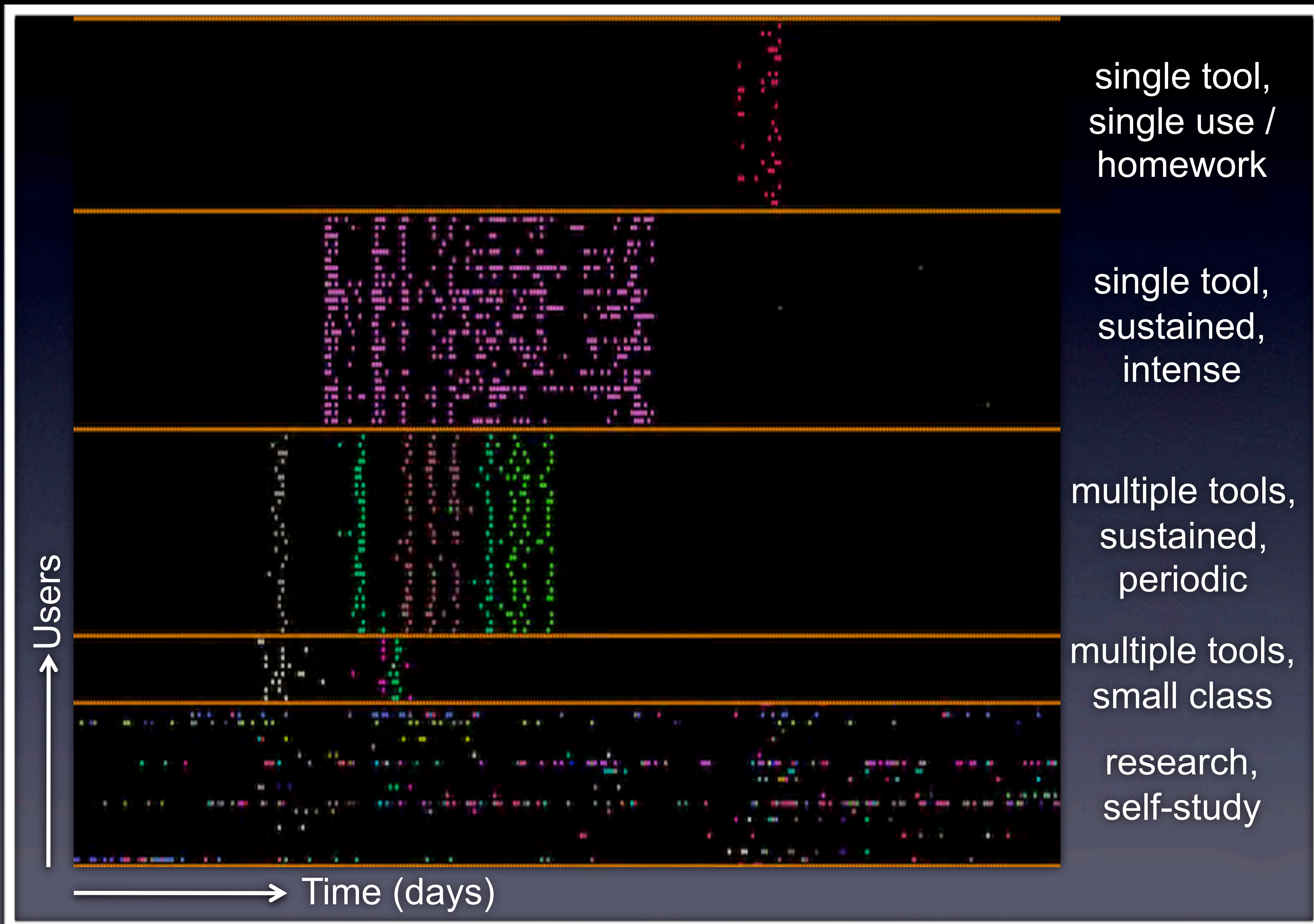
235 tools!



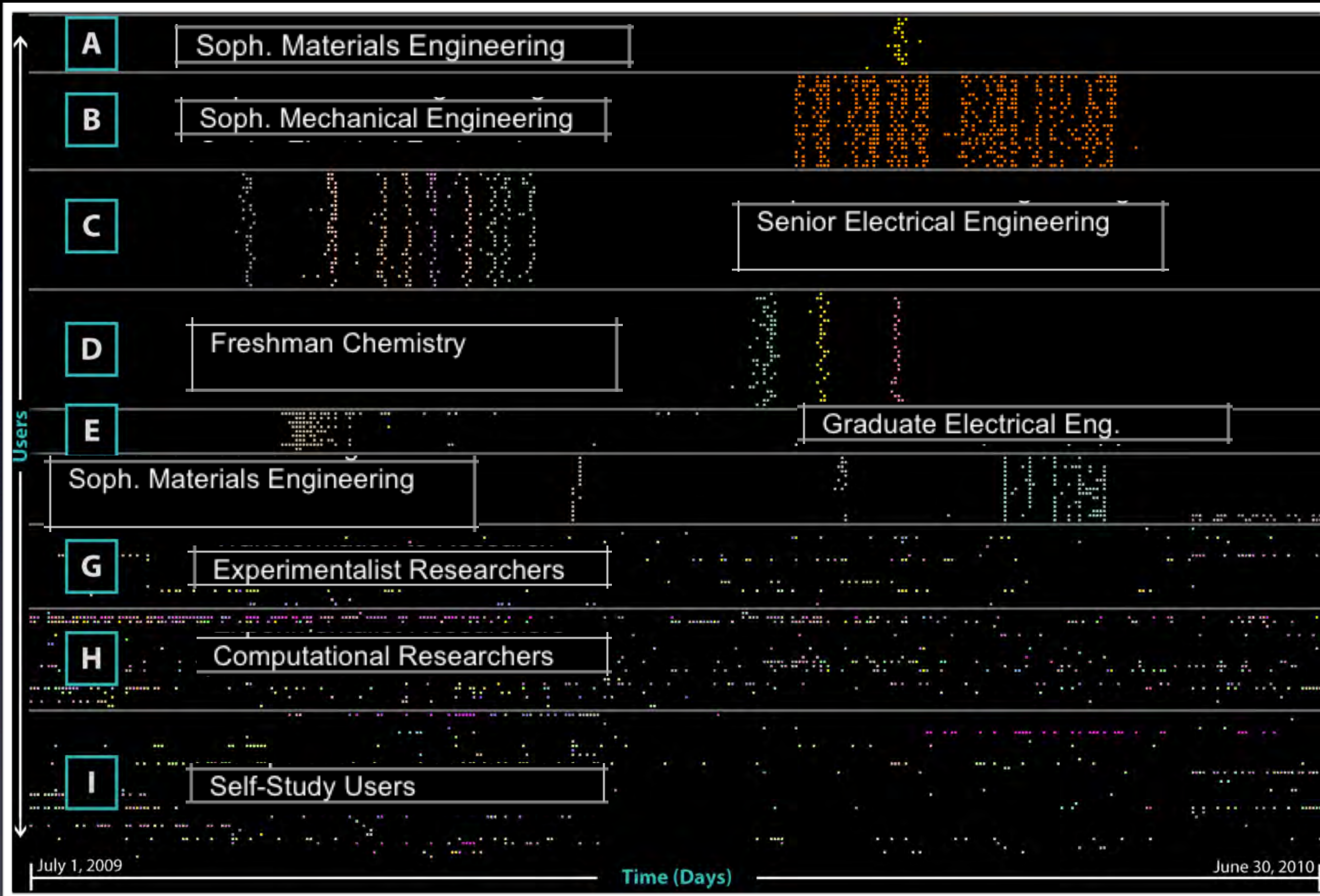
Tool Usage - Time Evolution



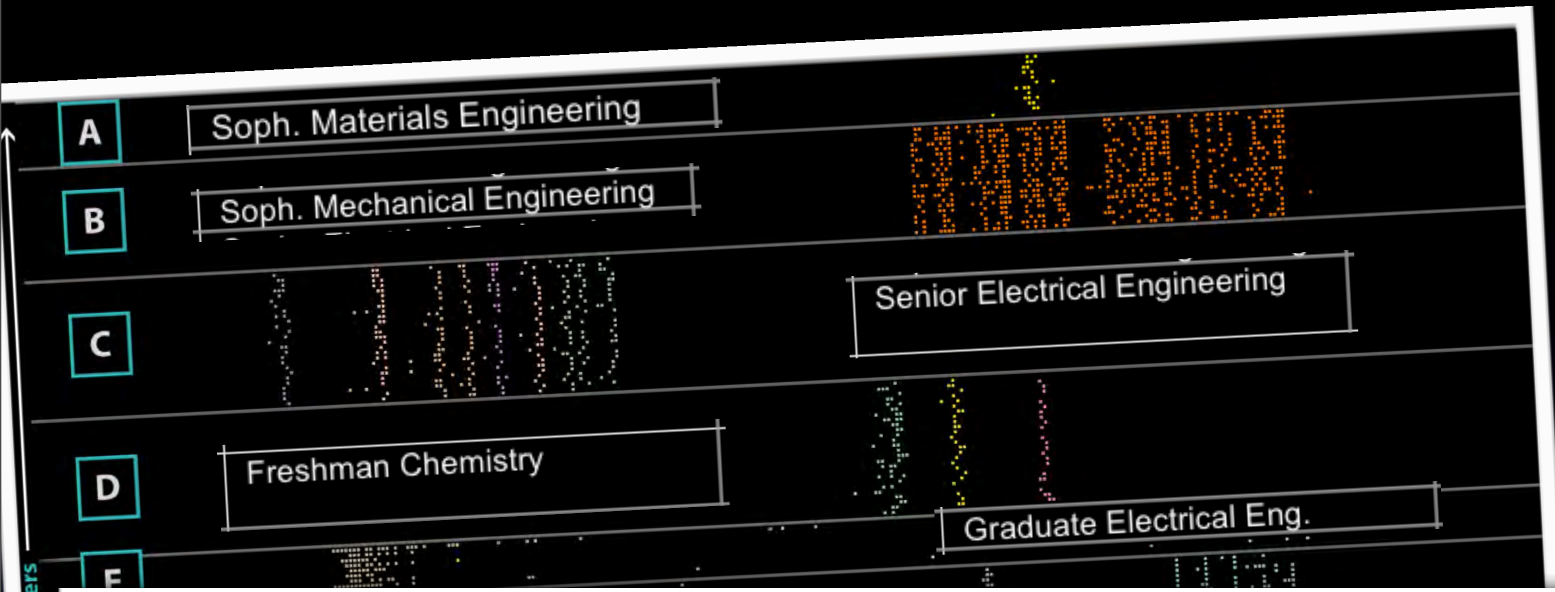
nanoHUB User Behavior



Formal Education vs. Research



Formal Education vs. Research

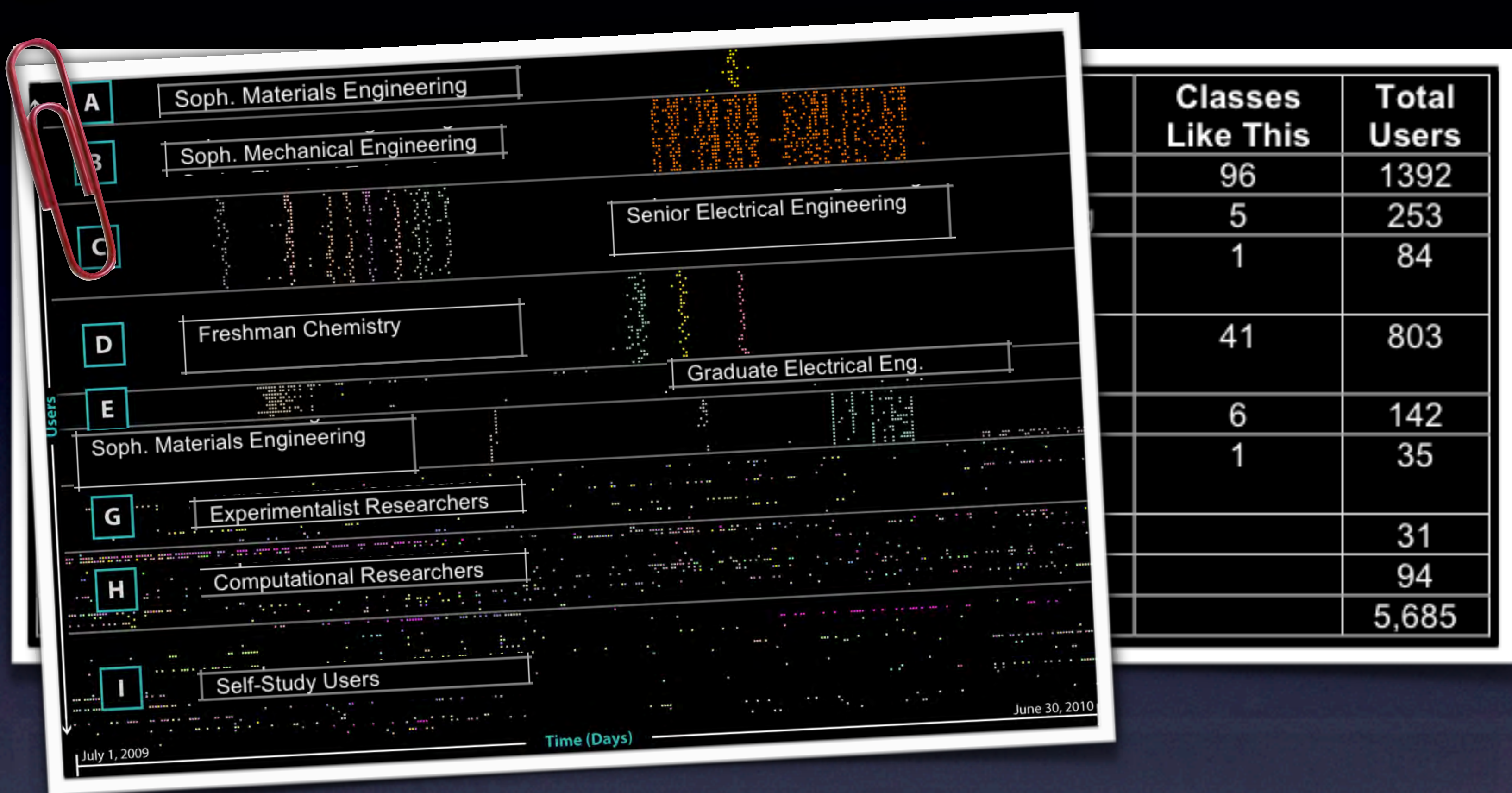


134	Courses
97	Institutions
3060	Students

95% outside NCN

	Tools and Usage Pattern	Validated Subset Shown	Classes Like This	Total Users
A	Single Tool, Single Use	Soph. Materials Engineering	96	1392
B	Single Tool, Semester Use	Soph. Mechanical Engineering	5	253
C	Multiple Tools, Periodic and Repeated Use	Senior Electrical Engineering	1	84
D	Multiple Tools, Periodic Single Use	Freshman Chemistry	41	803
E	Single Tool, Intensive Use	Graduate Electrical Eng.	6	142
F	Multiple Use in 3 Classes, Transformation to Research	Soph. Materials Engineering	1	35
G	Experimentalist Researchers	18 users		31
H	Computational Researchers	22		94
I	Self-Study Users	33 (not validated)		5,685

Formal Education vs. Research

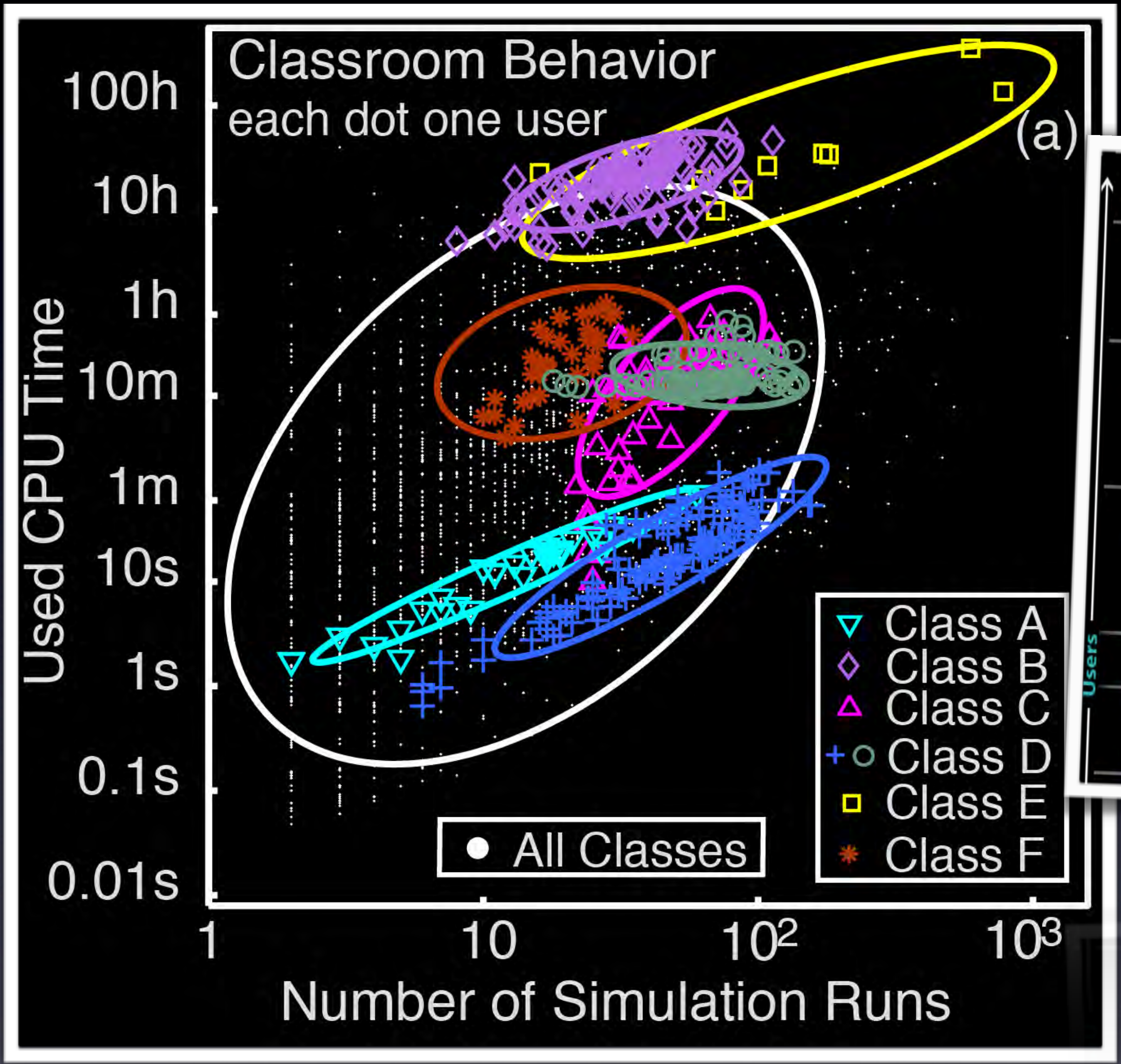


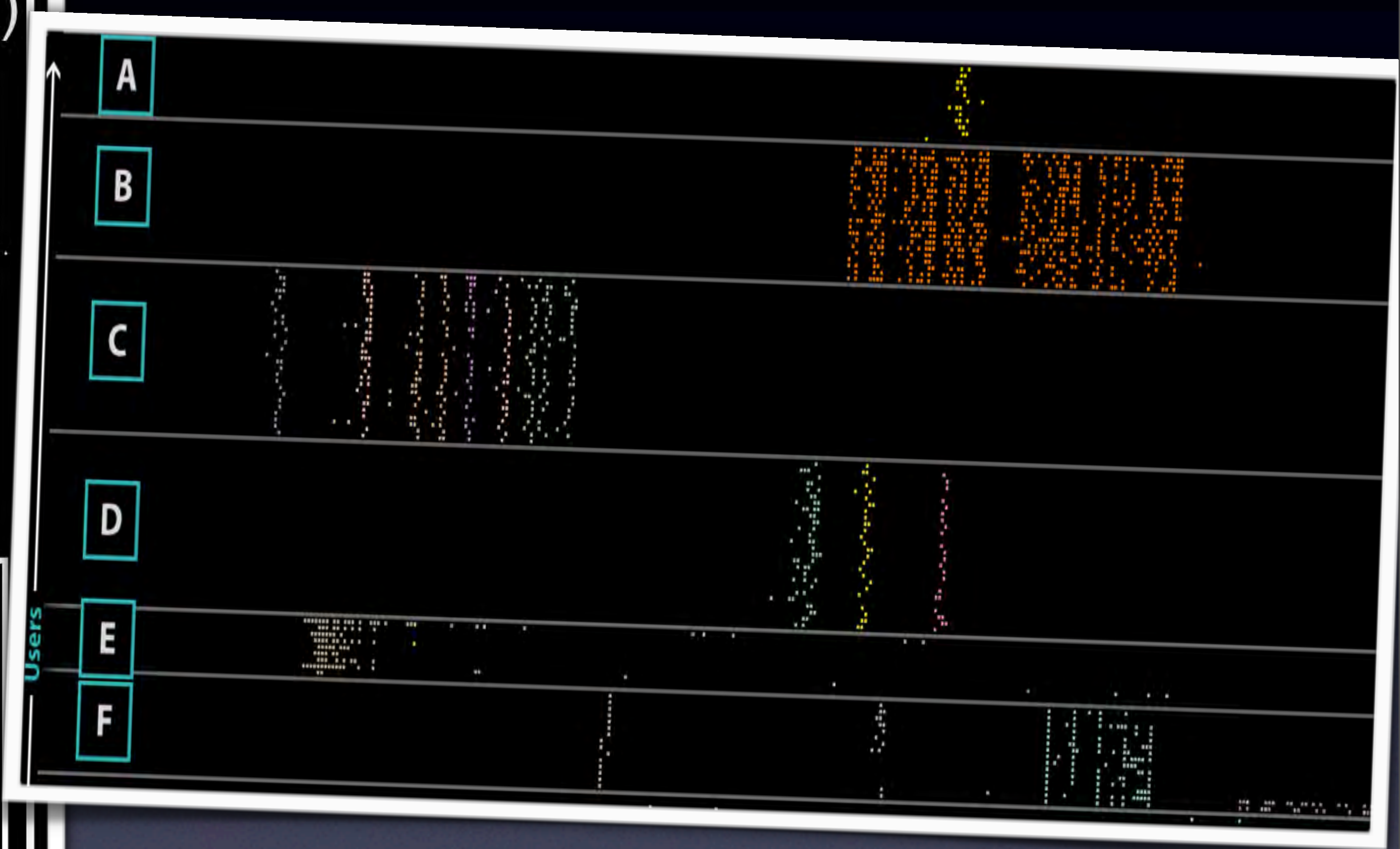
134	Courses	95% outside NCN
97	Institutions	
3060	Students	

KEY Insight	Proof of real use in education. Knowledge transfer out of research into education. Voluntary and VIRAL use!

Resource Requirements

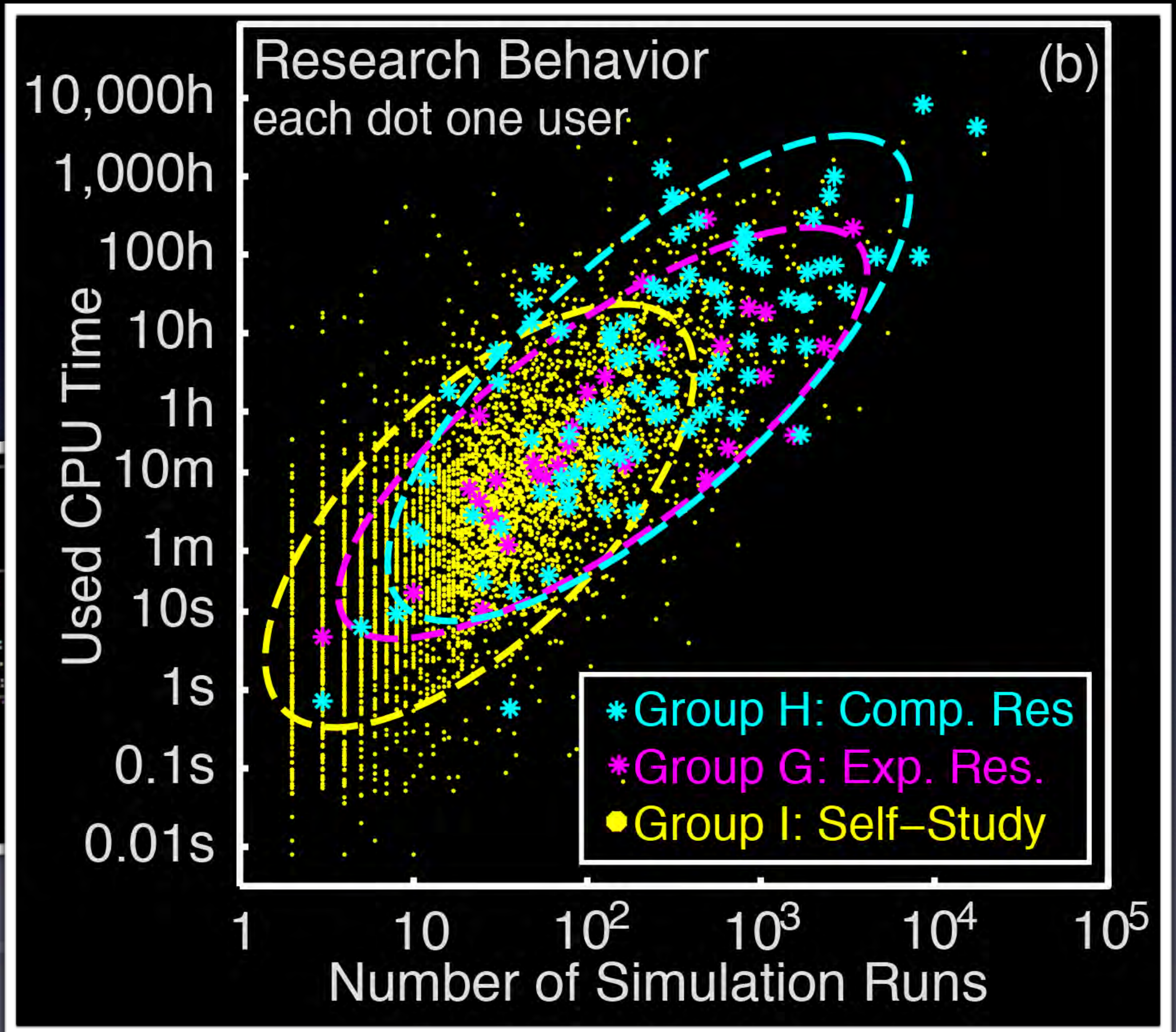
Simulations vs. CPU Consumption





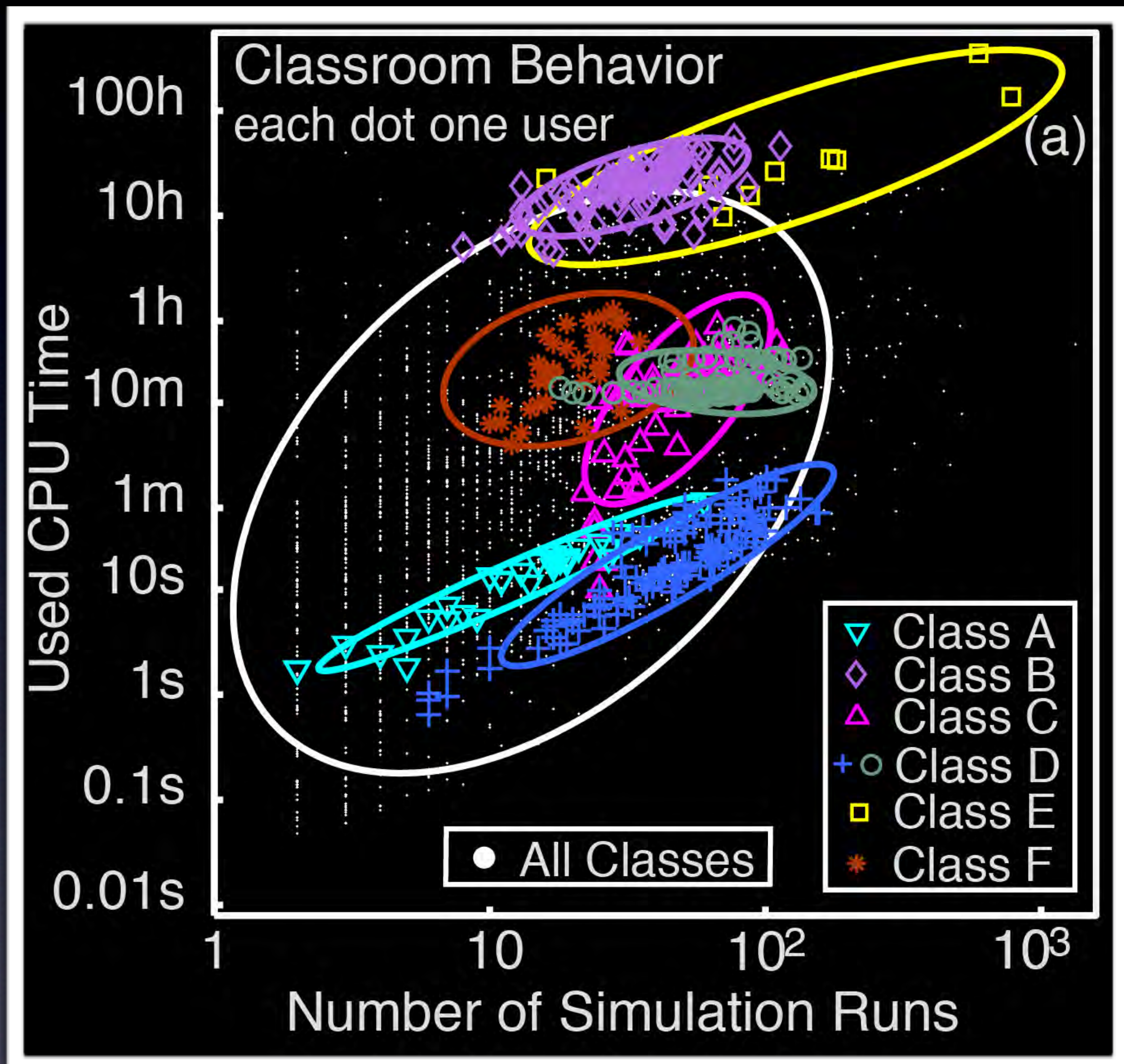
Resource Requirements

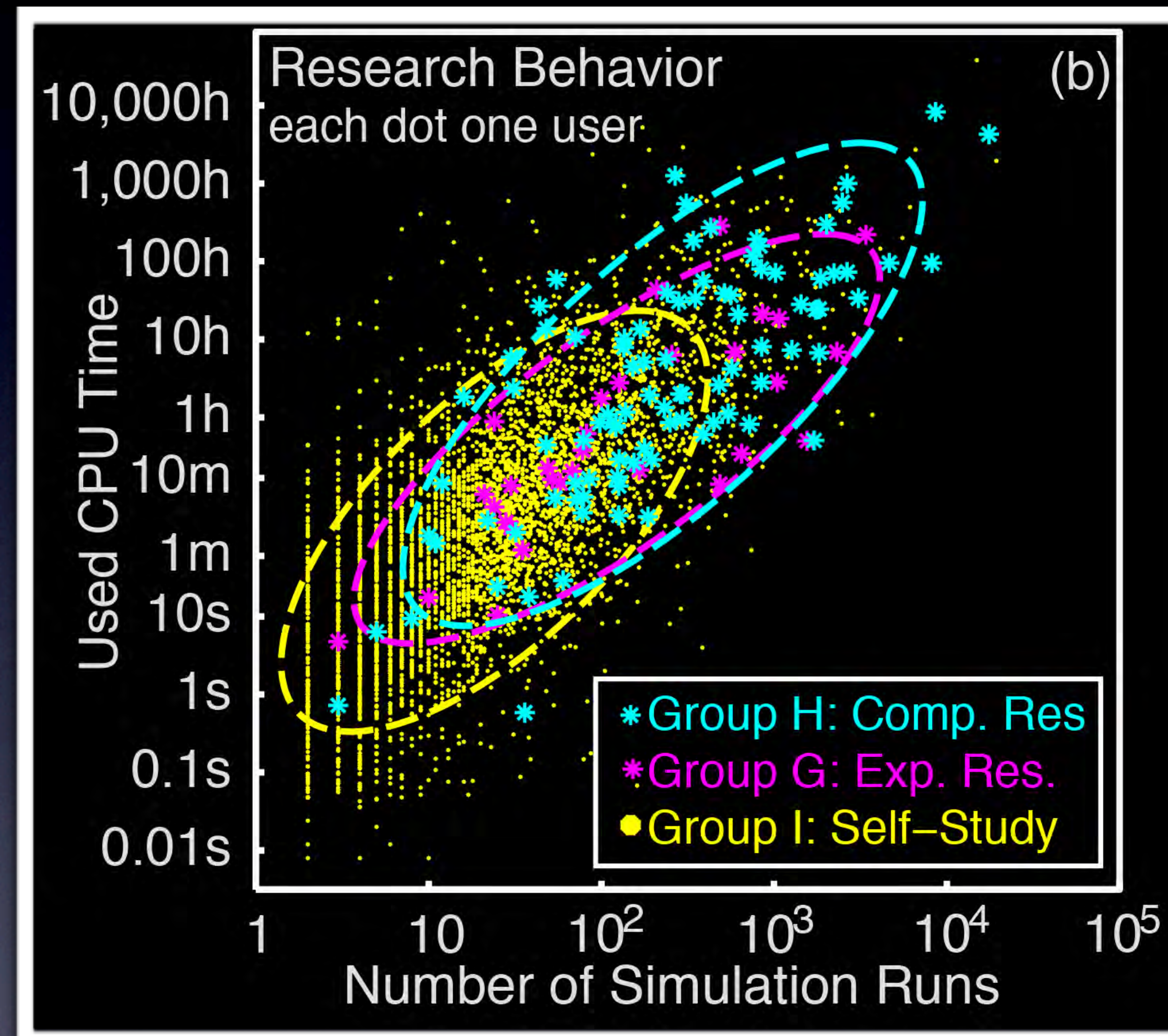
Simulations vs. CPU Consumption



Resource Requirements

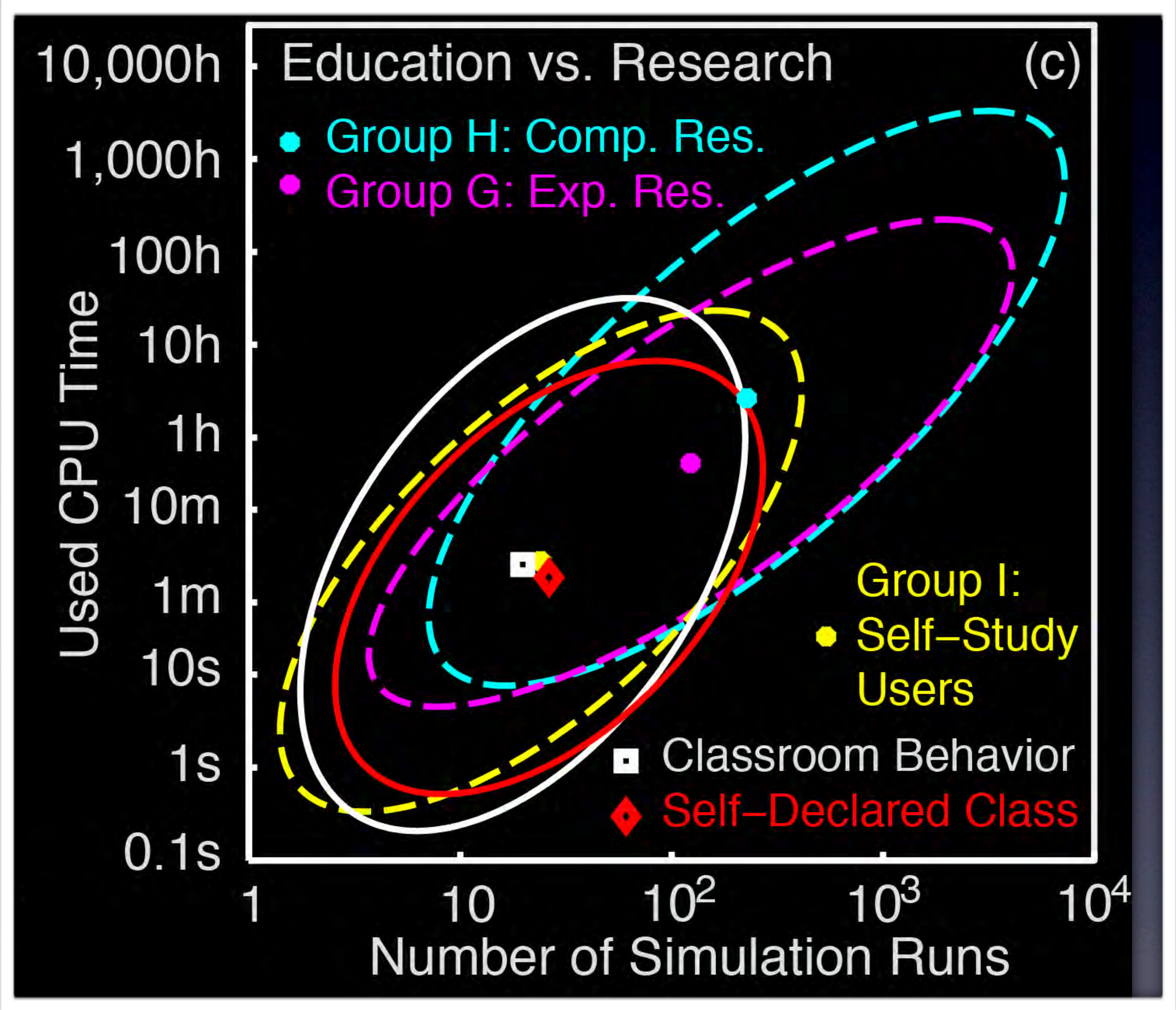
Simulations vs. CPU Consumption





Resource Requirements

Simulations vs. CPU Consumption



200	Runs	Average Research
8	Hrs CPU	
10000	Runs	Top Research
10000	Hrs CPU	
20	Runs	Average Edu
5	Min CPU	
400	Runs	Top Edu
20	Hrs CPU	

Case Study - Informal Spaces

Systems Perspective

Worked in collaboration Xin "Cindy" Chen and Dr. Mihaela Vorvoreanu (CGT, Purdue)

Instrumenting Informal Spaces

Social Media Proliferation

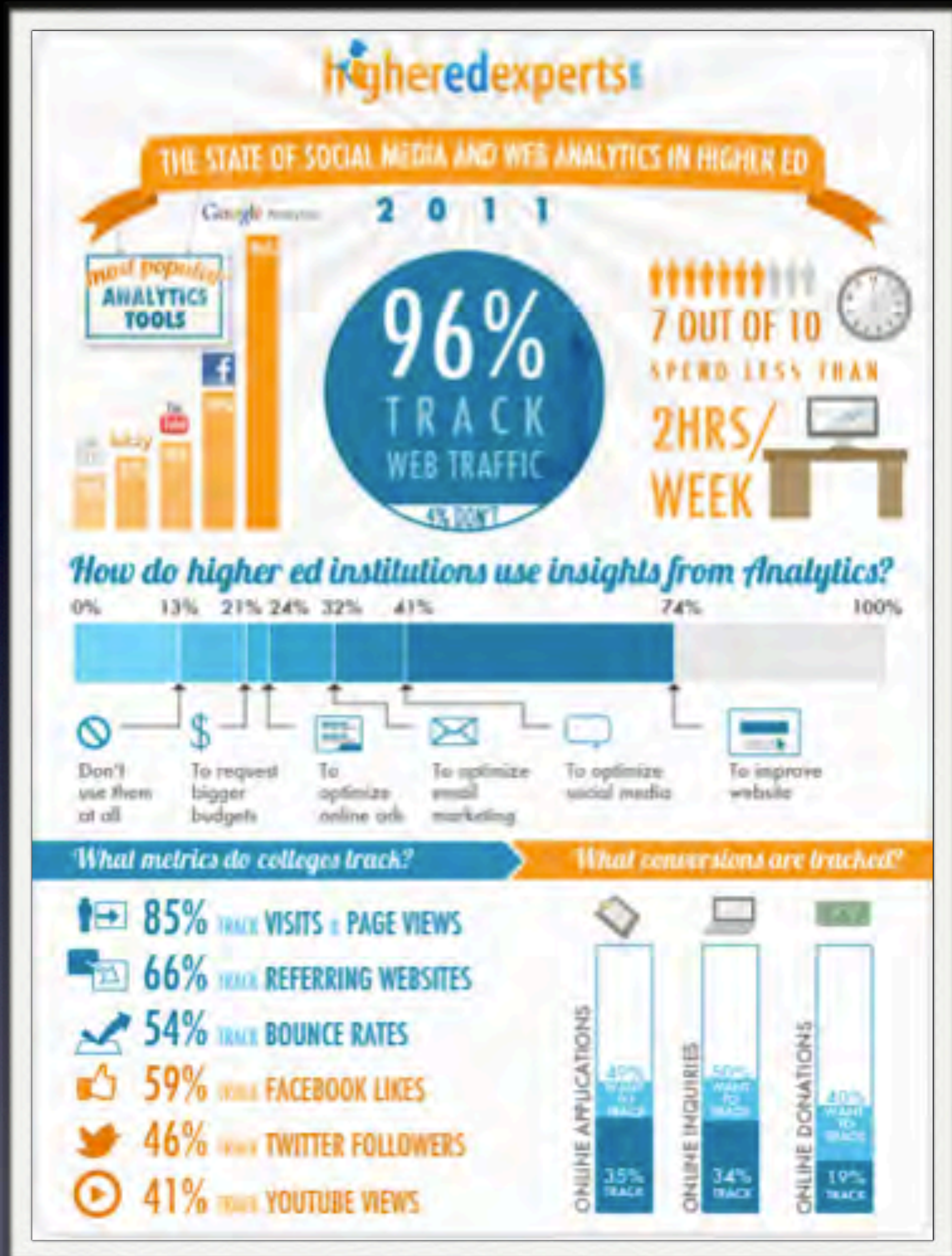
CYBERINFRASTRUCTURE
for engineering education

What insights do we gain from user generated data?

Dashboard(s-ing)

CYBERINFRASTRUCTURE
for engineering education

How does higher education use social media data?

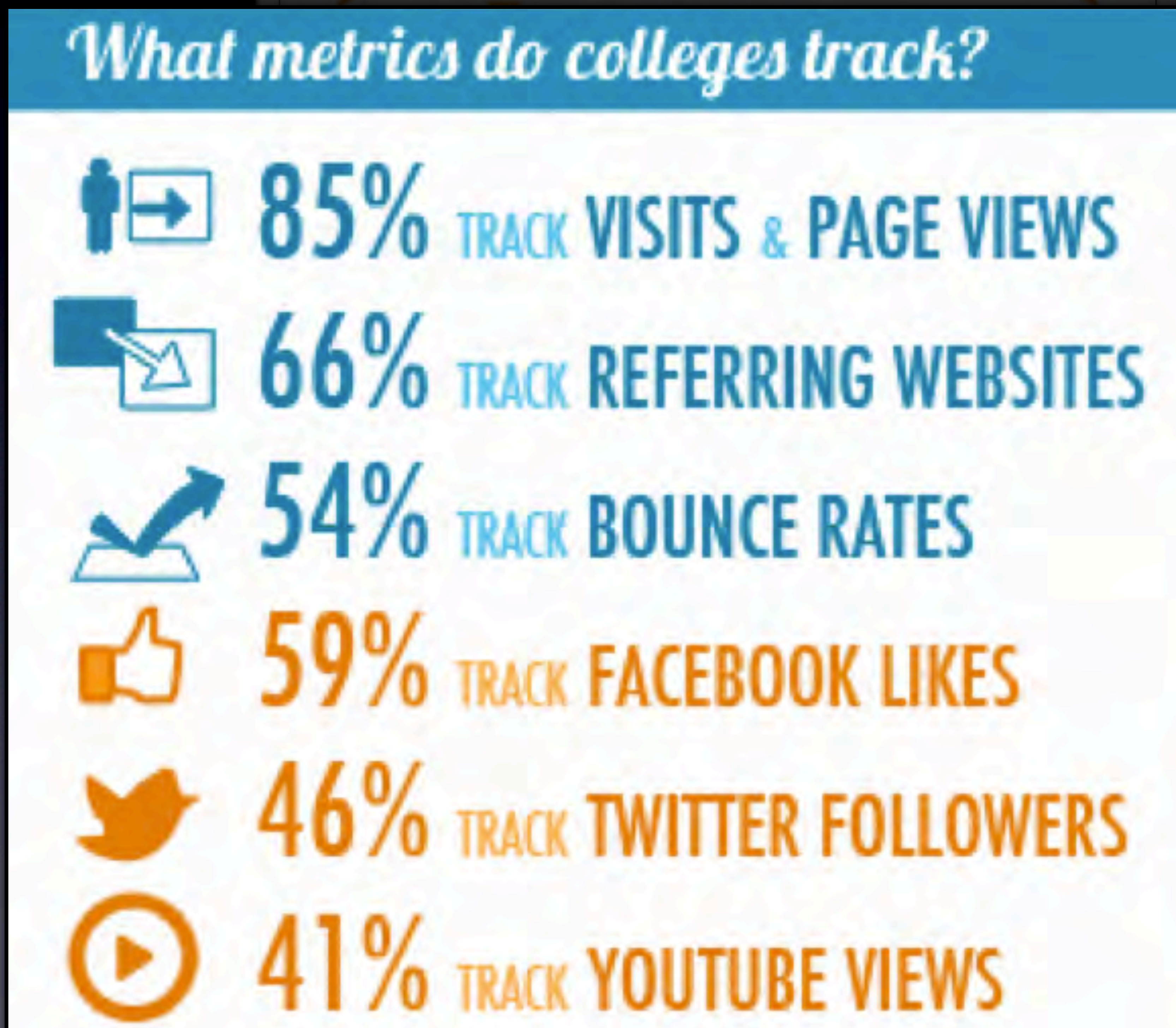


The State of Web and Social Media Analytics in Higher Ed, Survey by Higher Ed Experts, July 2011

CYBERINFRASTRUCTURE for engineering education

How does higher education use social media data?

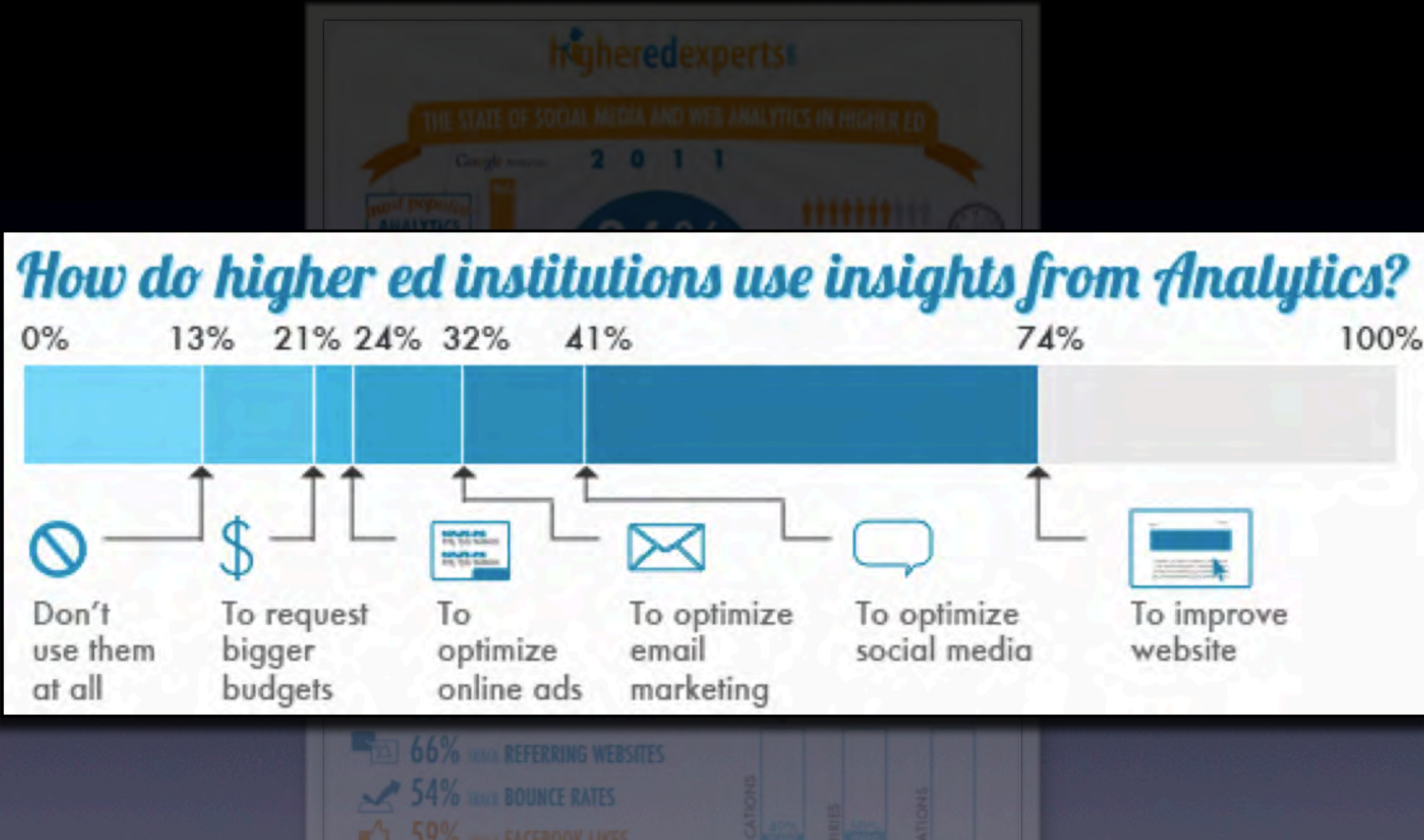
Mostly Number Counting, No Content Analysis



The State of Web and Social Media Analytics in Higher Ed, Survey by Higher Ed Experts, July 2011

CYBERINFRASTRUCTURE for engineering education

How does higher education use social media data?



Mostly for Marketing, Not Directly Related to Current Students

The State of Web and Social Media Analytics in Higher Ed, Survey by Higher Ed Experts, July 2011

CYBERINFRASTRUCTURE for engineering education

Methods

Strategy

Collect web content relevant to engineering students to understand their college experiences

Challenge

Relevant vocabulary is undefined; time span is undefined.

CYBERINFRASTRUCTURE
for engineering education

Methods

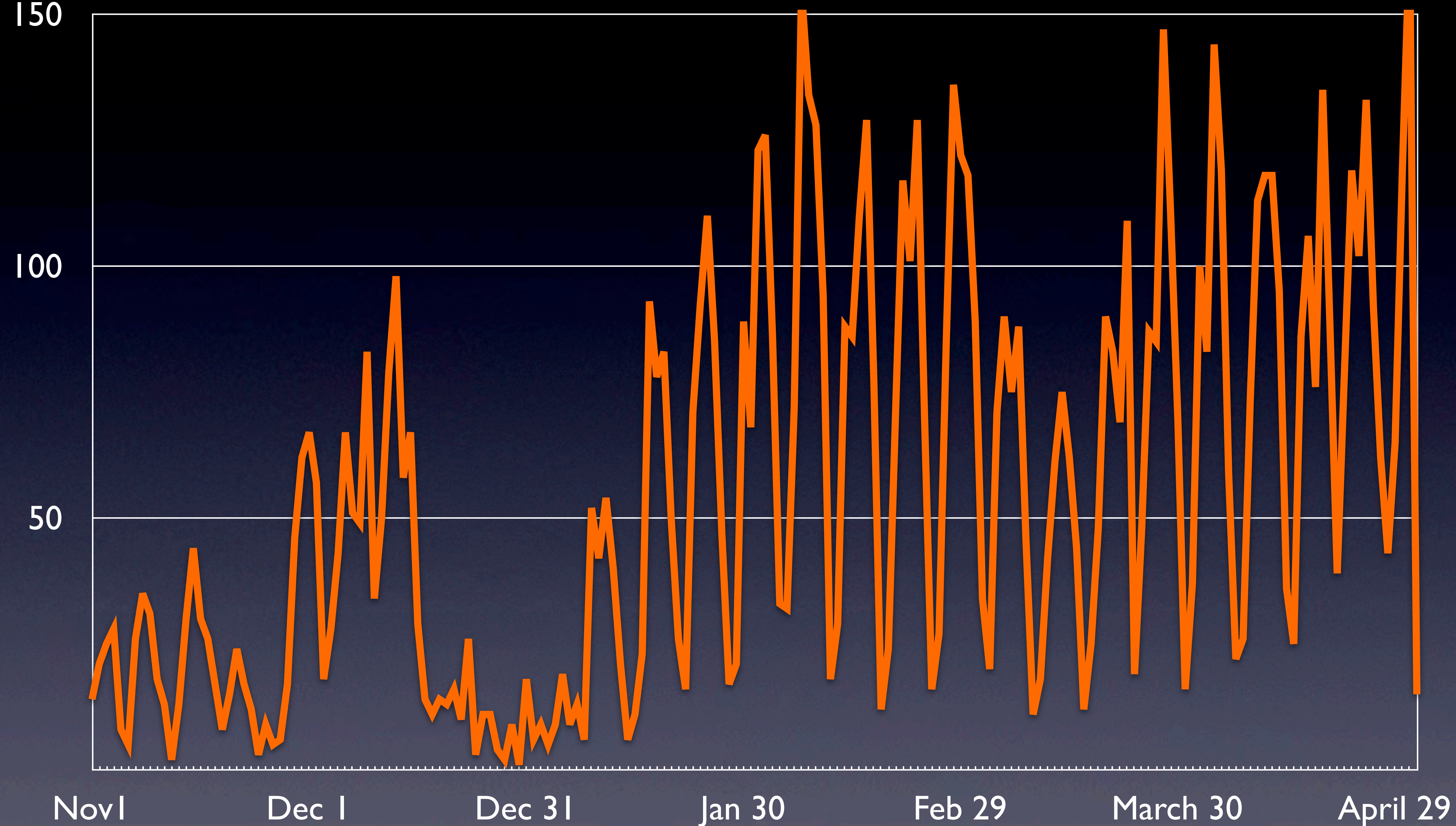
Iterative process of retrieving relevant data using *Radian6*

Nov. 1st, 2011 -- May. 2nd, 2012
#engineeringProblems: 10,006 tweets

1. Qualitative Content Analysis
2. Keyphrase Extraction and Topic Modeling

CYBERINFRASTRUCTURE
for engineering education

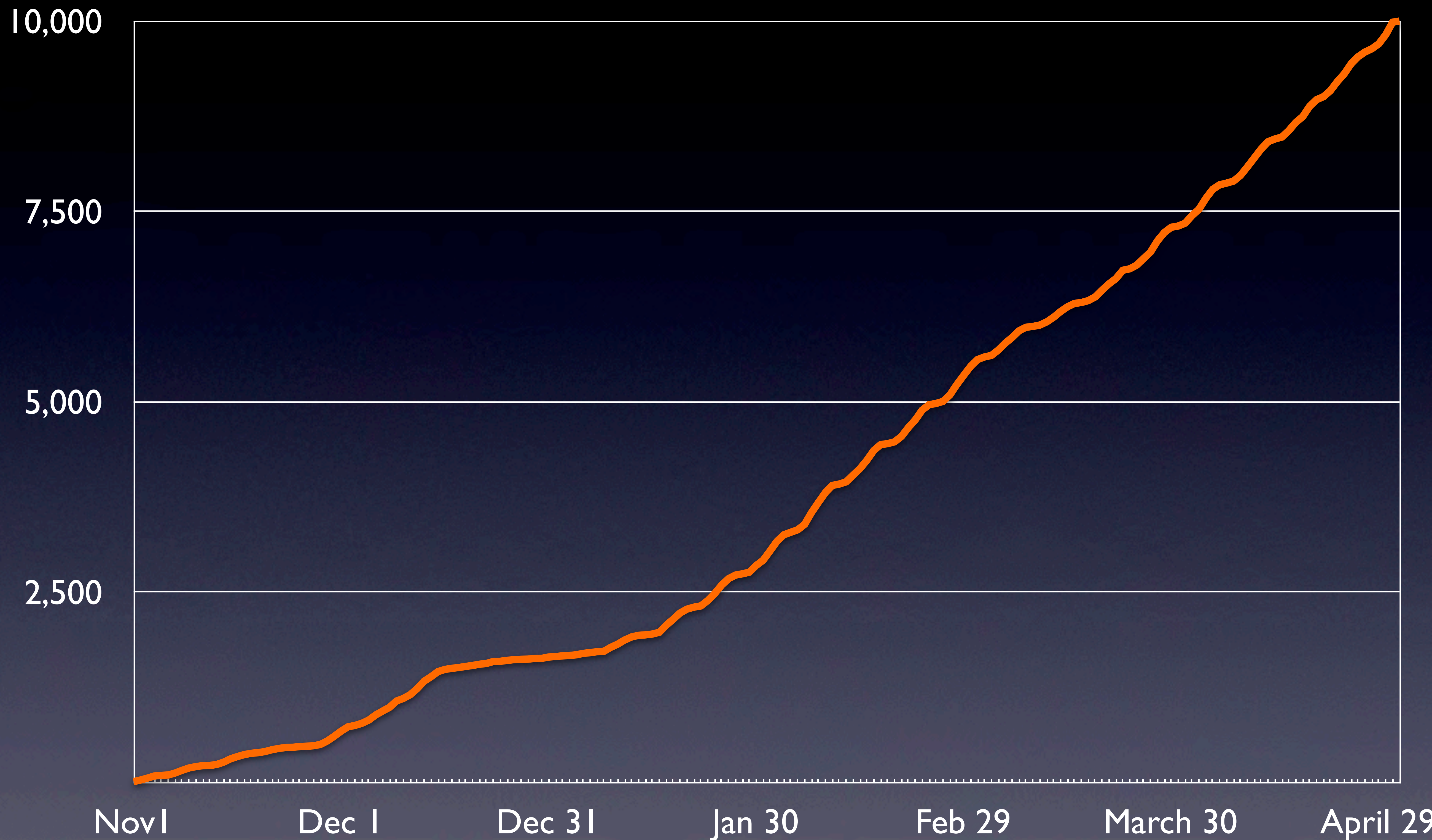
The trend: number of tweets per day using #engineeringProblems



Total No. of Tweets: 10, 006

CYBERINFRASTRUCTURE
for engineering education

The trend: accumulative number of tweets using #engineeringProblems



Total No. of Tweets: 10, 006

CYBERINFRASTRUCTURE
for engineering education

Qualitative Results

1. Sacrifice and Negative Feelings
2. Issues with Classes, Professors, Homework, and Exams
3. Gender and Other Minority Issues
4. Engineer Stereotypes and Identity Formation

CYBERINFRASTRUCTURE
for engineering education

Text Analysis

From the machine perspective, text is ***unstructured, nominal, qualitative*** data. It needs to be transformed in order to be visualized.

CYBERINFRASTRUCTURE
for engineering education

Unstructured Text

↓

Transformation

↓

Structured Data

↓

Visual Representation

CYBERINFRASTRUCTURE
for engineering education

Unstructured Text

↓

Transformation ▶ Keyphrase Extraction and Topic Modeling

↓

Structured Data

↓

Visual Representation

CYBERINFRASTRUCTURE
for engineering education

Extract prominent key terms and identify main themes from large text corpora.

Topic Modeling Results

Topic 0: problems, this week, calculator, forget, calc (calculus), happy, feeling, really, learn, hopefully, finish, numbers, year, right now, too much work, it's bad, solutions manual, guess, everyday, scores, multiple test, find out, exams, differential equations, pretty, glad, can't follow, coffee, easy, angle

Topic 1: ever, professor, words with friends, math with friends, trying, I'm awful, calculate, favor, pretty sure, engineering building, URL, hard, sometimes, the only girl, stop, more time, stay, pressure, GPA, back pack weighs more, sleep, determine, calculate how far, complicated, bitch, business major, starting, girls bathroom, don't understand, finally

Topic 2: awkward moment, Friday night, actually doing any, amount, yeah, don't know, curve, actually, free time, days, weekends, still, book, even, last night, drunk, same week, purpose, sitting, next week, don't even know how, senior design, feeling not tired, buy beer, napping, for hours, don't know, pull, force

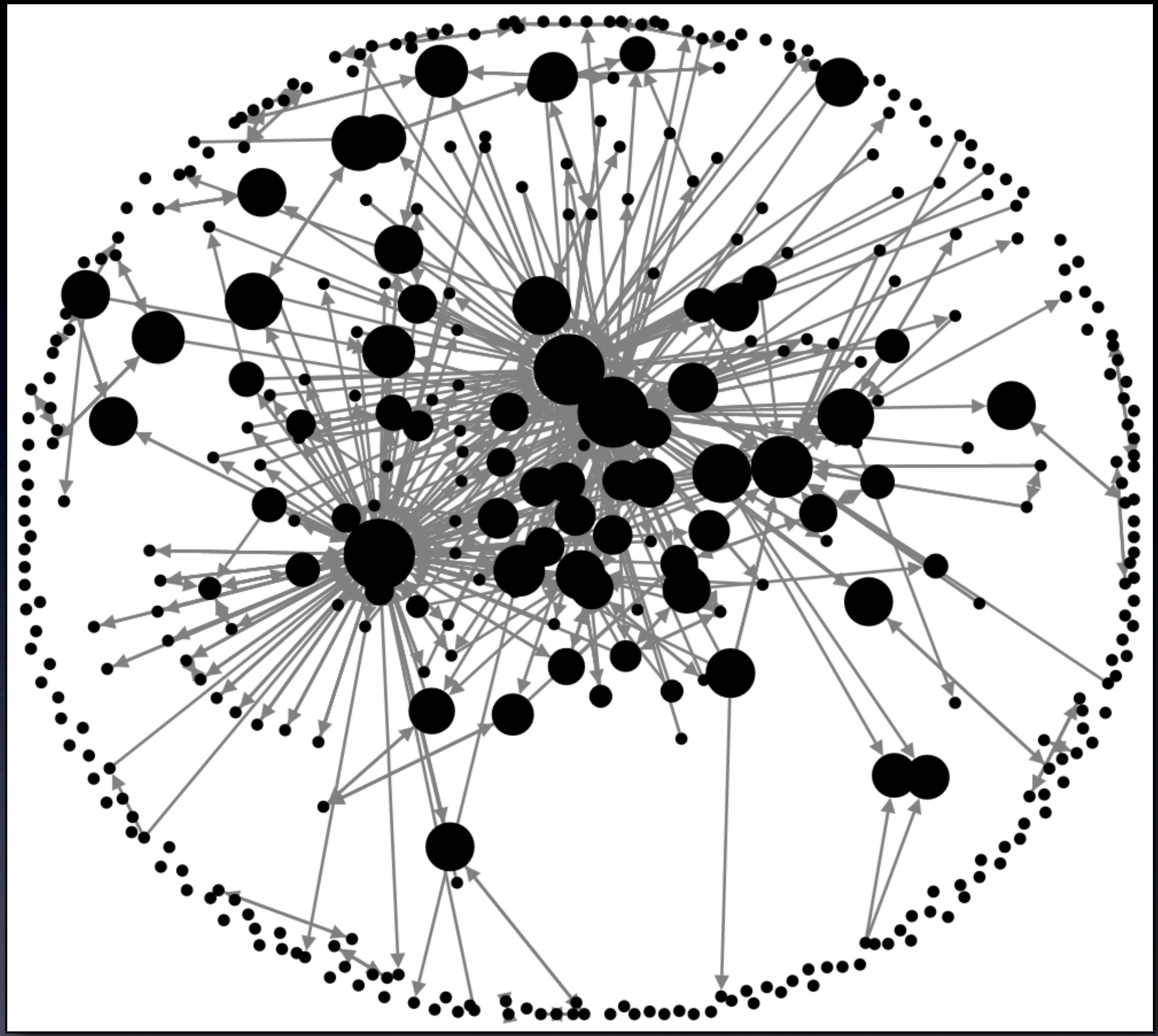
Not Converge to Distinct Topics, Need Manual Curation

Implications and Future Direction

1. Research and Policy Implications
2. Social Support, Community Building
3. Social Media Analytics Tool for Education

CYBERINFRASTRUCTURE
for engineering education

Are these Twitter users building a community?



Partial #engineeringProblems Network based on Mentions, Replies, and Follows

Questions?

Krishna Madhavan
School of Engineering Education
Network for Computational Nanotechnology
cm@purdue.edu