Positioning Libraries to Support Data Science

Tim Dennis, Director, UCLA Library Data Archive

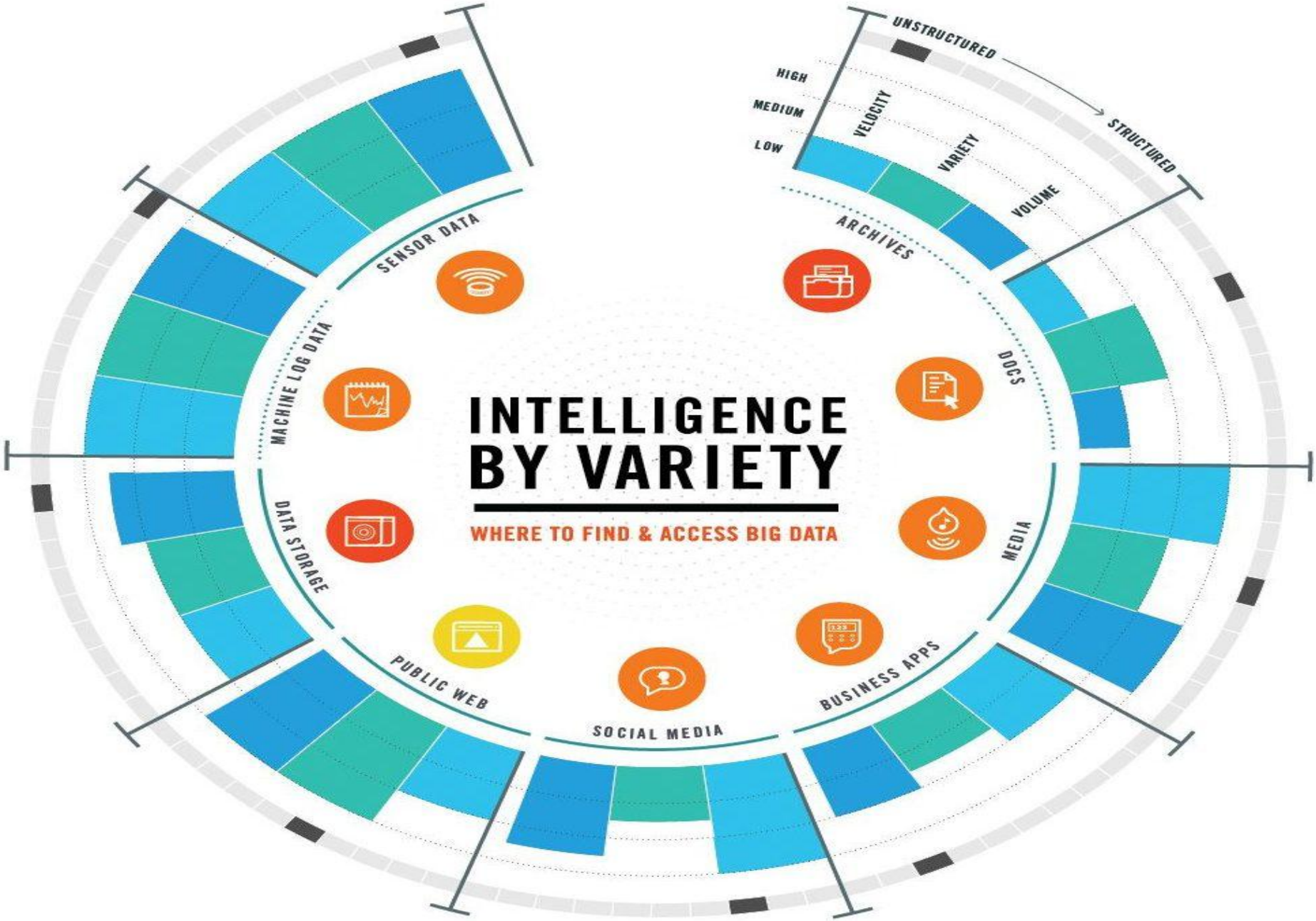
LiASA 2017 - 18th Annual Conference
October 5, 2017

UCLA

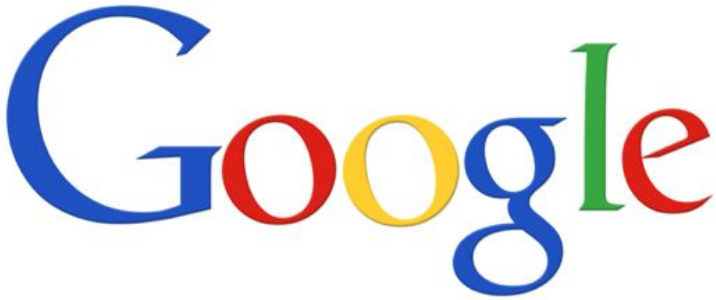
What is Data Science?

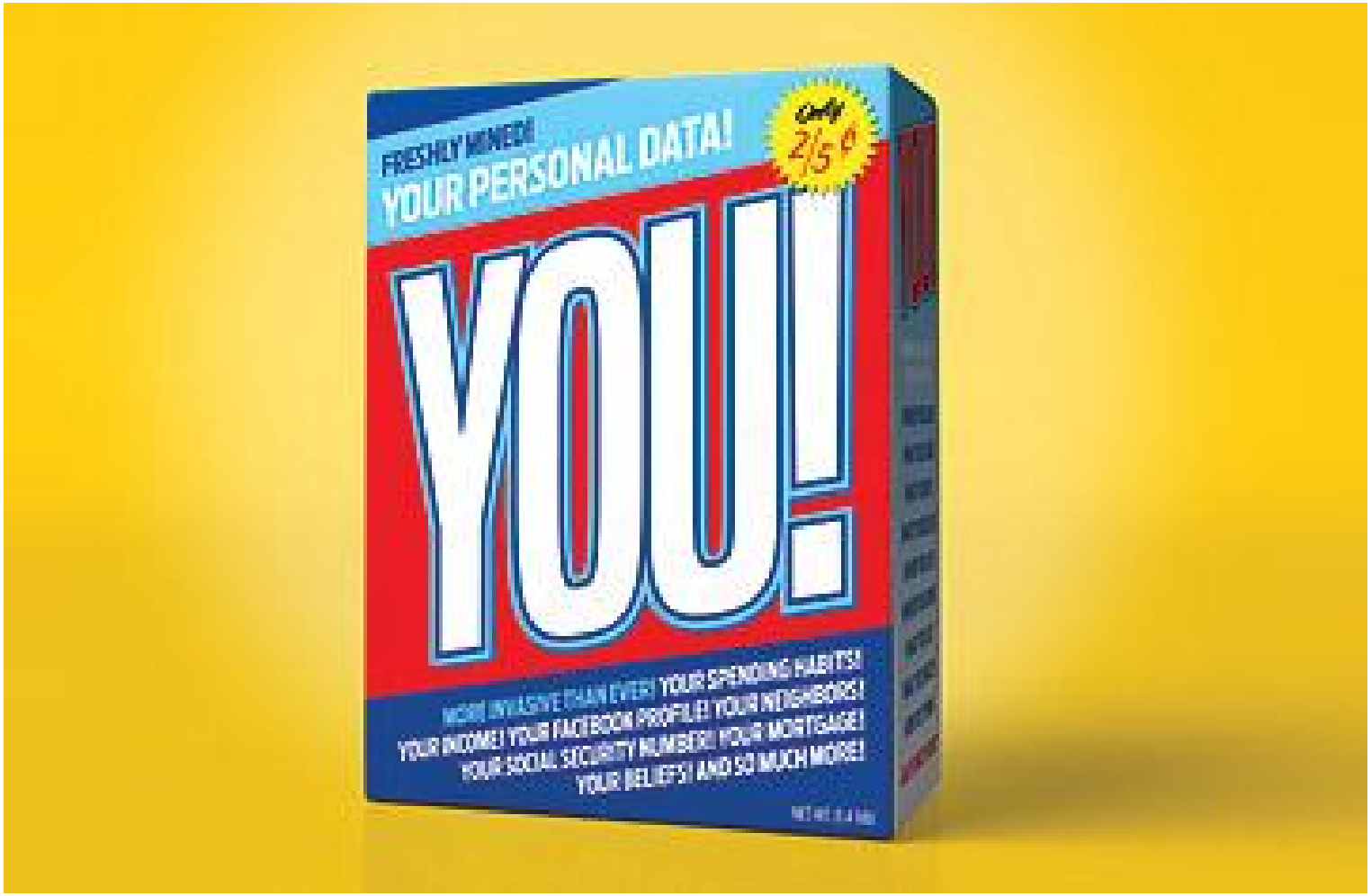
"[Data Science is] the ability to extract knowledge and insights from large and complex data sets"

-- Former US Chief Data Scientist - Dr. DJ Patil https://goo.gl/3Y6es5


UNSTRUCTURED
STRUCTURED
HIGH
MEDIUM
LOW
VELOCITY
VARIETY
VOLUME
ARCHIVES
DOCS
MEDIA
BUSINESS APPS
SOCIAL MEDIA
PUBLIC WEB
DATA STORAGE
MACHINE LOG DATA
SENSOR DATA
INTELLIGENCE
BY VARIETY
WHERE TO FIND & ACCESS BIG DATA

These companies collect and profit from data




FRESHLY MINED!
YOUR PERSONAL DATA!
Only 2.50
YOU!
MORE INVASIVE THAN EVER! YOUR SPENDING HABITS!
YOUR INCOME! YOUR FACEBOOK PROFILE! YOUR NEIGHBORS!
YOUR SOCIAL SECURITY NUMBER! YOUR MORTGAGE!
YOUR BELIEFS! AND SO MUCH MORE!

Industry Context of Use

- Need to make sense of huge swaths of increasingly dynamic and complex data in order to monetize
- Need people with skills to:
  - Munge, clean and organize data
  - Create algorithms to make data actionable
  - Store data efficiently and intelligently

“

Data is the new oil"

Clive Humby

What does that mean for academia?

UCLA

Impact on Academia

- Many academics with higher order math/CS skills are poached (or start start-ups)
- Market pressures put on Universities to produce more grads with data skills
- Research methods in many fields transformed by computationally enabled research and access to “Big Data”

82 Masters

Most releva

DISCIPLINES

<< All Disciplines

Computer Science & IT (1480)

Business Information Systems (116)

Computer Sciences (701)

Data Science & Big Data (82)

Geographical Information Systems (GIS) (78)

Health Informatics (90)

Human Computer Interaction (20)

Informatics & Information Sciences (352)

IT Security (100)

Video Games & Multimedia (30)

Web Technologies & Cloud Computing (47)

LOCATION

United Kingdom (75)

United States (82) Show All

Filters: United States × Data Science & Big Data × Clear All

Analytics, M.Sc.

American University Washington DC Kogod School of Busin... | Washington, D. C., Washington, D.C., U

44,369 EUR / year | 1 year | On campus | English (Take IELTS Test)

The Master of Science in Analytics program (MSAn) offered by the Kogod Schoo
American University Washington DC prepares students to be experts in data an
organizational decisions, and to solve dynamic business problems. The progra

Add to Comparison D

Graduate Pathway in Data Analytics Engineering, Pre-Maste

INTO George Mason University | Fairfax, Virginia, United States

29,759 EUR / year | 9 months | On campus | English (Take IELTS Test)

http://www.mastersportal.eu/

Undergrad Data Science Education

In spring 2016, UC Berkeley's first Foundations of Data Science course attracted around 300 students. This semester (2017), nearly 1,000 have enrolled...Across the UC system, campuses are quickly adding data science programs in response to soaring workplace demand.

--By Isha Salian, www.sfchronicle.com

Cathryn Carson
@CathrynCarson

Following

First day of class in Foundations of #Datascience for @BerkeleyDataSci - all materials at data8.org

10:23 AM - 23 Aug 2017

14 Retweets **47** Likes

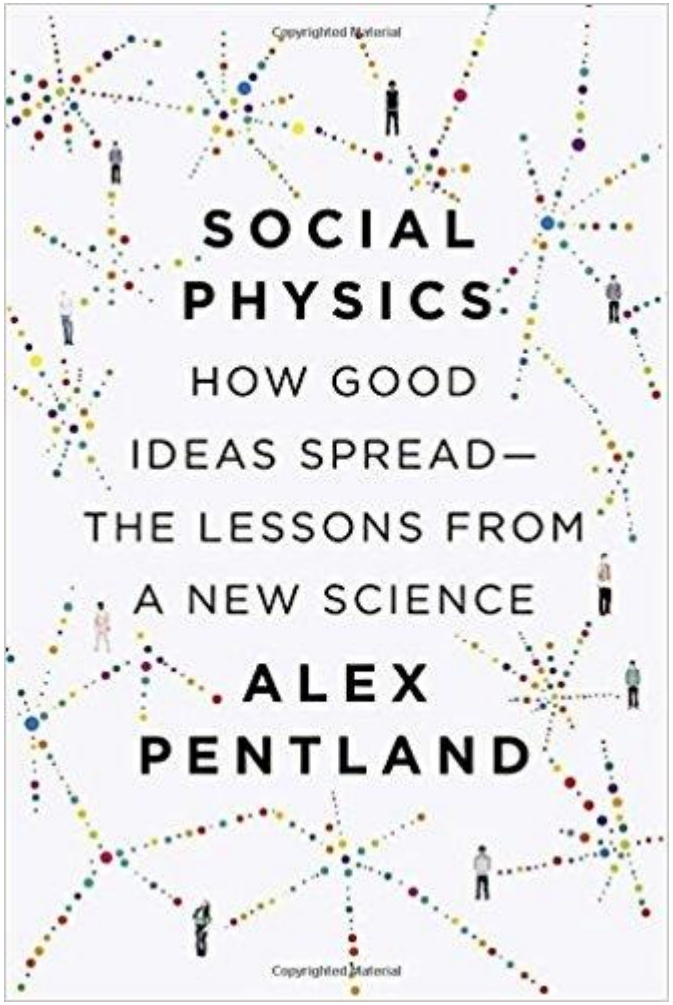
3 14 47

Research Innovation

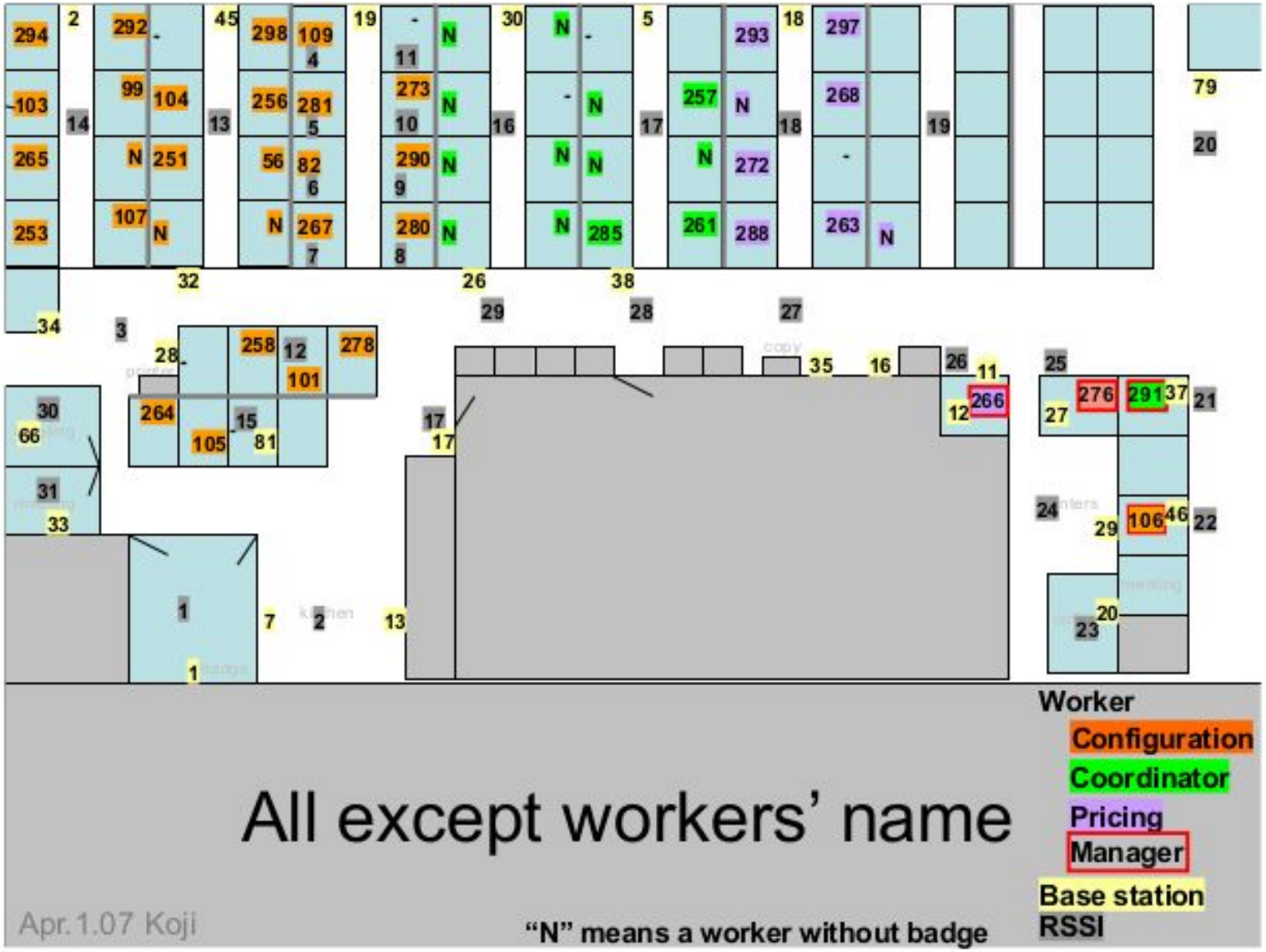
Data Science enables new research questions through new data sources:

- Cell phone data
- Social media data
- Sensor data
- Textual data

Example: Sensor Data



- Modelling humans like we do ants
- Used sensors in a business office
- Sensor badges had mic, accelerometer, and geolocation


294
2
292
45
298
109
4
19
11
N
30
N
5
293
18
297
79
103
14
99
104
13
256
281
5
273
10
N
16
N
17
257
N
18
268
19
20
265
N
251
56
82
6
290
9
N
N
N
N
272
253
107
N
N
267
7
280
8
N
N
285
261
288
263
N
32
26
38
34
3
29
28
27
28
258
12
278
101
264
15
105
81
35
16
26
11
266
12
25
276
291
37
21
27
30
66
31
33
17
17
24
29
106
46
22
1
7
2
13
20
23
1
Worker
Configuration
Coordinator
Pricing
Manager
Base station
RSSI
All except workers' name
Apr.1.07 Koji
"N" means a worker without badge

Example: Sensor Data

Findings:

- Individual's style of speaking (rate, emotional tone, mirroring, turn-taking predicts 30% of the variance in performance (raises)
- Can determine influence networks and nodes
- Exposure to peer behavior influences individual behavior

Example: Sensor Data

Data Collection Sites:

- Dorm network
- Grad school network
- Thin data on cell phone data from Ivory Coast

Pentland asserts “we are scratching the surface” on this type of research

Example: Social Media Data

American Political Science Review Page 1 of 18 May 20

doi:10.1017/S00030554130000

How Censorship in China Allows Government Criticism but Silences Collective Expression

GARY KING *Harvard University*
JENNIFER PAN *Harvard University*
MARGARET E. ROBERTS *Harvard University*

We offer the first large scale, multiple source analysis of the outcome of what may be the most extensive effort to selectively censor human expression ever implemented. To do this, we have devised a system to locate, download, and analyze the content of millions of social media posts originating from nearly 1,400 different social media services all over China before the Chinese government is able to find, evaluate, and censor (i.e., remove from the Internet) the subset they deem objectionable. Using modern computer-assisted text analytic methods that we adapt to and validate in the Chinese language, we compare the substantive content of posts censored to those not censored over time in each of 85 topic areas. Contrary to previous understandings, posts with negative, even vitriolic, criticism of the state, its leaders, and its policies are not more likely to be censored. Instead, we show that the censorship program is aimed at curtailing collective action by silencing comments that represent, reinforce, or spur social mobilization, regardless of content. Censorship is oriented toward attempting to forestall collective activities that are occurring now or may occur in the future—and, as such, seem to clearly expose government intent.

Example: Big Text Revolution

- Digital text encodes much of human activity, including the past
- Often relatively cheap to purchase
- Def. “Big Data”
- How to parse it and make sense from it is a growing research field

How can Libraries engage?

UCLA

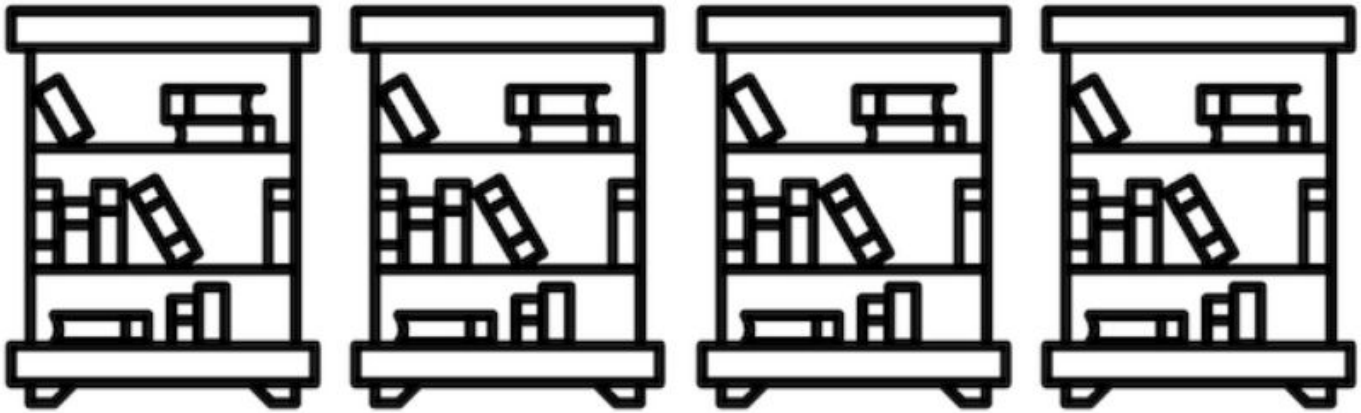
Two areas where libraries can jump in

1. Collections
2. Data services

Collections

1. Purchase licensed data sources (Corpora of text, social media data, etc.)
2. Treat purchased data as a first class citizen of your collection
3. Collections as data

Collections as Data



Collections as Data is an Institute of Museum and Library Services supported effort that aims to foster a strategic approach to developing, describing, providing access to, and encouraging reuse of collections that support computationally-driven research and teaching in areas including but not limited to Digital Humanities, Public History, Digital History, data driven Journalism, Digital Social Science, and Digital Art History.

Example: Collections as Data

- Free Speech Movement collection - artifacts from FSM digitized
- The Library & Research IT created API
- Held competition (Hackathon) for students to create novel interfaces for it

FREE SPEECH MOVEMENT DIGITAL ARCHIVE

Enter your keyword...

Search the Archive!

Text Image Audio

September 14, 1964

Free Speech Movement

A short timeline of the events of the FSM

Data Services

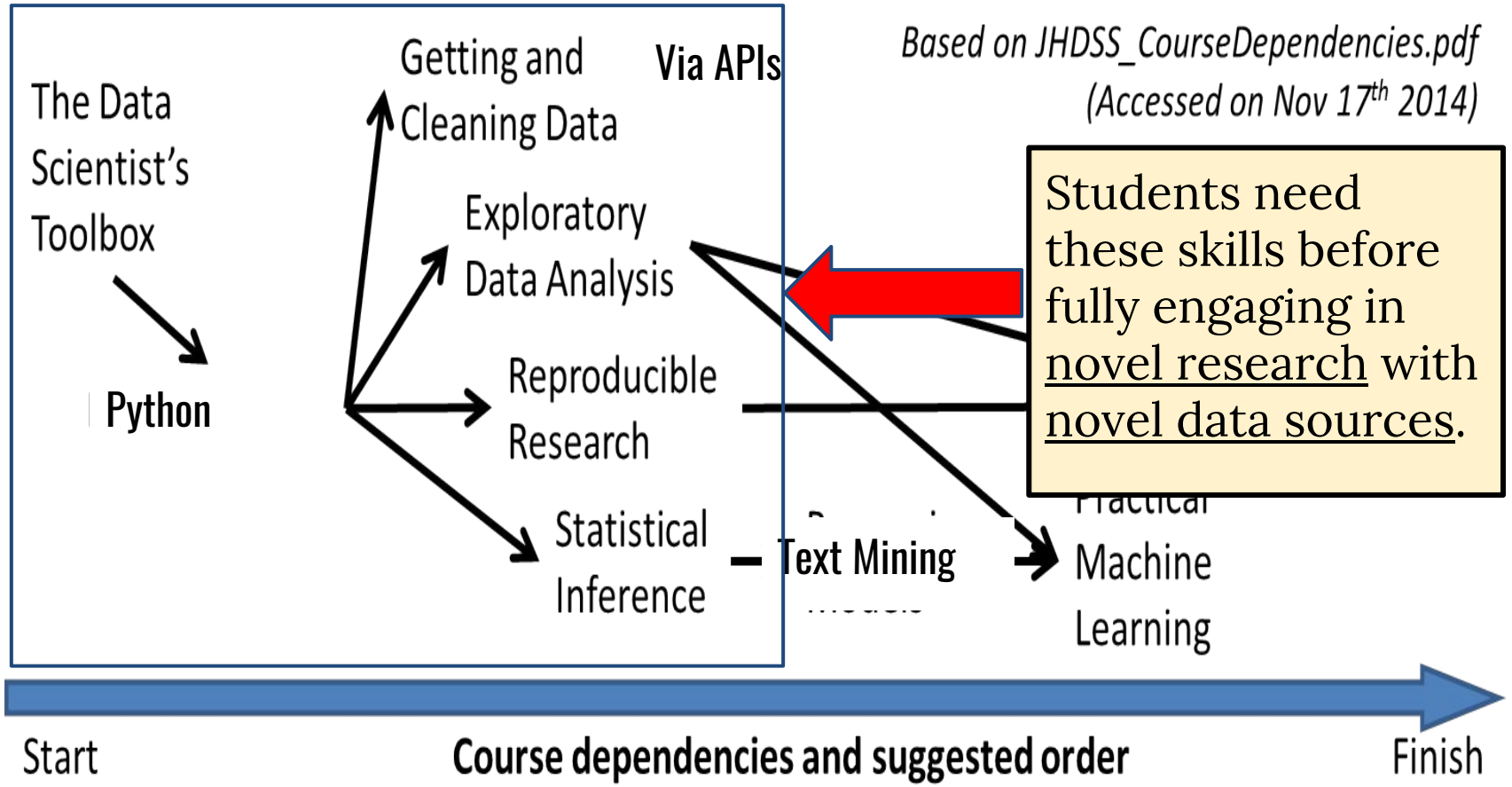
1. Finding data
2. Help with research data needs:
    - management, backup and storage
    - identify data repositories
    - educating and preparing data for sharing and publishing
    - metadata and standards support
3. Provide consultation service on using data
4. Data Instruction

A note about metadata

Metadata is crucial to your Big Data project's overall success and to your enterprise data architecture organization.

--Data Science Central
https://goo.gl/u2ocJA

New Research Depends on Basic CS Skills



Source: https://goo.gl/t9DiBW

The Library UC SAN DIEGO

Training

- If data training doesn't exist on your campus, consider starting a training program for graduate students.
  - Use Data and Software Carpentry's model to start with high quality workshops and build a local network
- If a data training program exists, partner with whoever is running it and offer to help or support
- Start meetup style groups (HackyHour, Python and Pizza, Data and Donuts)

Software Carpentry

Software development best practices

Domain agnostic

- Command line
- Version control with git
- Programming in Python or R

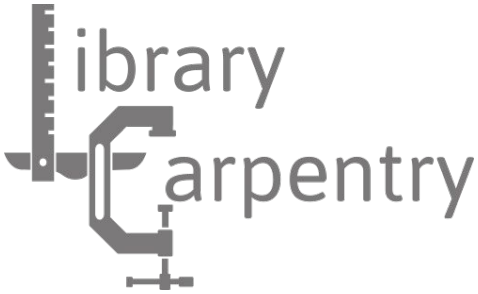
Data Carpentry

Working effectively with data
includes domain-specific content

- Data organization
- Data cleaning
- Data analysis and visualization in R or Python

Ok, why else should I learn this stuff?

- So you aren't replaced by a robot
- Learning data and software skills is very marketable (ask the students at your school)
- It's fun!
- It will help your researchers



Library Carpentry

Working effectively with data using best practices

- Basic computation skills
- Versioning and collaboration through Github
- Data analysis and cleanup through OpenRefine
- If interested, join the community call Oct. 10, 4pm. Details: https://goo.gl/PxnQCG

Data Science

Give[s] us new ways of grasping patterns, collectivities, and systematic effects that remain invisible to us without statistical and computational tools; of understanding the linkages from data to knowledge to decision-making under conditions of uncertainty; of exploiting domain-specific computational possibilities fluidly and reliably and seeking cross-fertilization across them; and of **critically engaging** the constructive and creative possibilities opened up by data collection and computation, as well as their **challenging ethical and social entanglements**.

-- Data Sciences @ Berkeley, The Undergraduate Experience

Data Literacy for Society

- Libraries (Public and Academic) should be offering advice and education on cyber-security so people can protect their personal data
- As an equity issue they should be introducing users to free and open source replacements for expensive proprietary systems
- They should also be fostering the teaching of code because this is where the jobs of the future will exist
- Providing data literacy training so people can navigate every day society (news, advertisements, etc.)

Questions and Discussion

Thanks so much!

Questions?

Tim Dennis <timdennis@ucla.edu>
Twitter: @jt14den