Big Data and Science: Myths and Reality

H.V. Jagadish

http://www.eecs.umich.edu/~jag

Six Myths about Big Data

- It’s all hype
- It’s all about size
- It’s all analysis magic
- Reuse is easy
- It’s the same as Data Science
- It’s all in the cloud

Big Data Myth 1

- Big Data is all hype.

Data Analysis Has Been Around for a While

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

W.E. Demming

1958: "A Business Intelligence System"

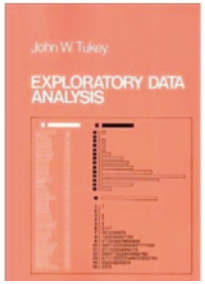
Peter Luhn

1997: "Machine Learning"

1977: "Exploratory Data Analysis

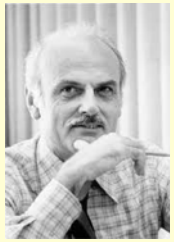
1989: "Business Intelligence"

1970: Relational Database

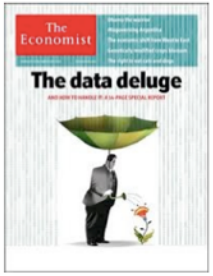


Howard Dresner

E.F. Codd

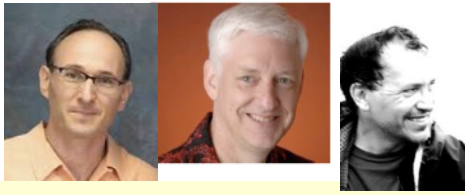


2010:"The Data Deluge"



2007:"The Fourth Paradigm"

2009:"The Unreasonable Effectiveness of Data"



1996: Google

Google™

Abridged Version of Jeff Hammerbacher's timeline for CS 194 at UCB, 2012

Breathless Journalists!!

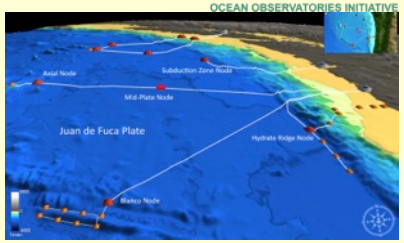
Big Data Impetus

- Can collect cheaply, due to digitization.
- Can store cheaply, due to falling media prices.
- Driven by business process automation and the web.
- But now impacting everywhere.

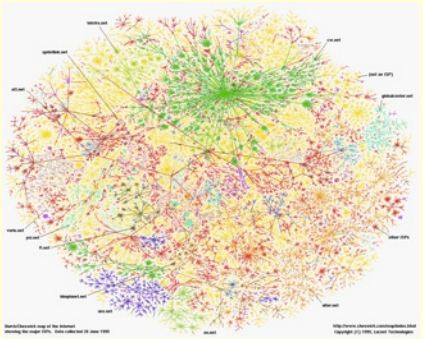
Nearly every field of endeavor is transitioning from “data poor” to “data rich”

Astronomy: LSST

Physics: LHC



Oceanography: OOI



Sociology: The Web

Biology: Sequencing

Economics: mobile, POS terminals

Neuroscience: EEG, fMRI

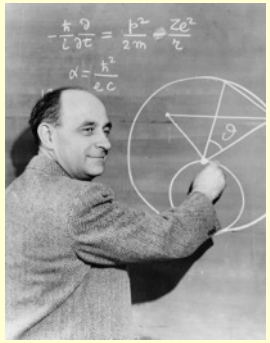
Data-Driven Medicine

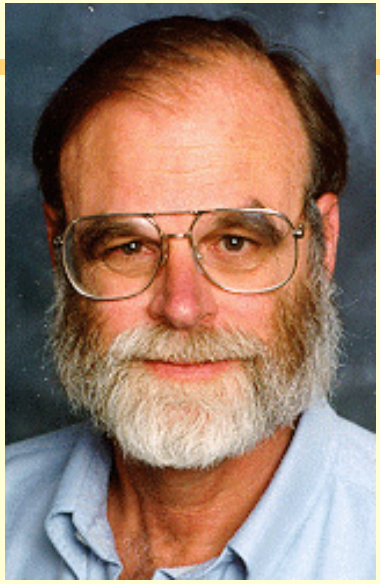
Sports

Slide courtesy of the Moore Sloan Data Science Environments Initiative

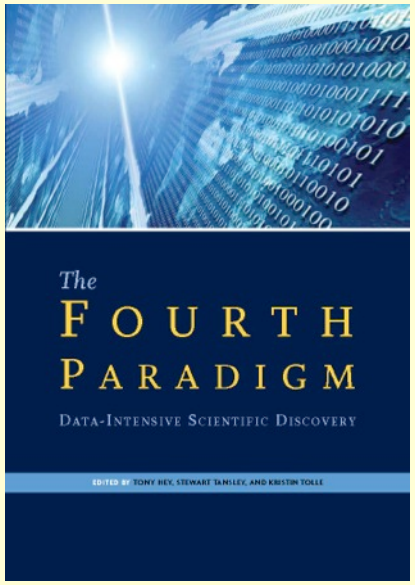
Data-Driven Science

1. Empirical + experimental
2. Theoretical
3. Computational
4. Data-Intensive



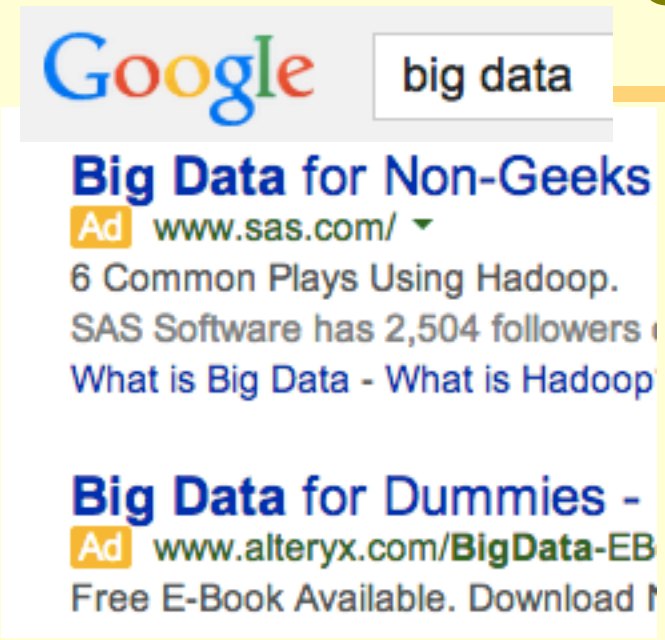


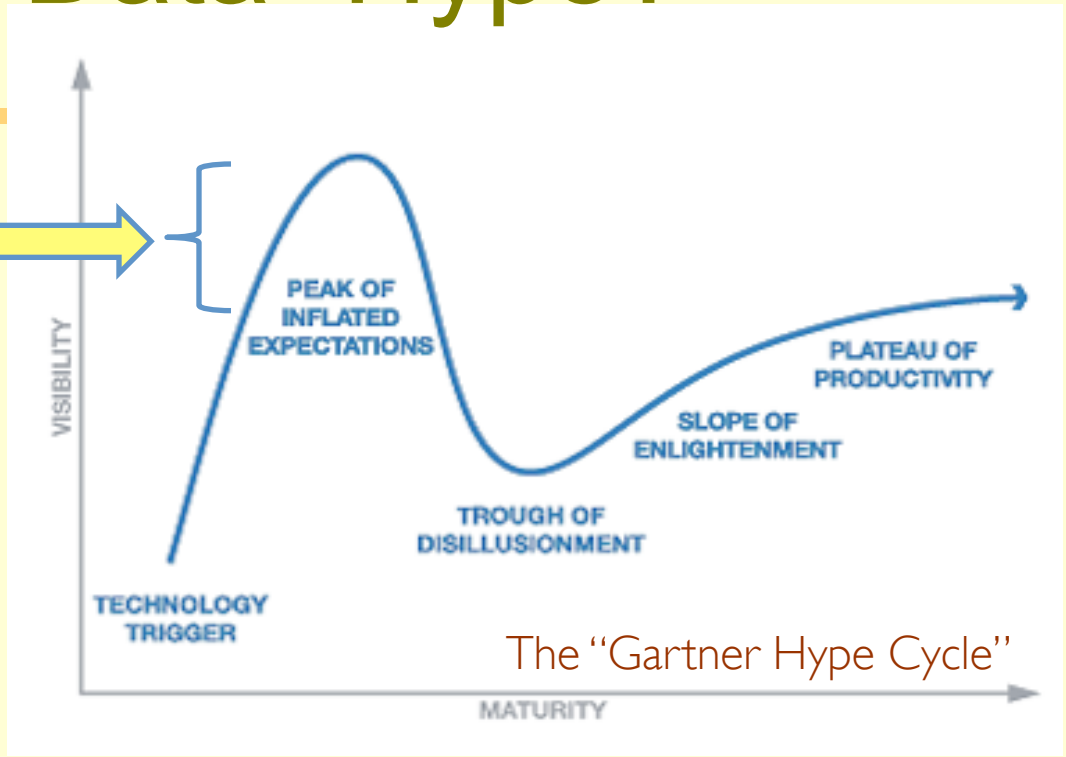
Jim Gray



Slide courtesy of the Moore Sloan Data Science Environments Initiative

“Big Data” Hype?





The “Gartner Hype Cycle”

Just because it’s hyped
doesn’t mean we can or should ignore it

Slide courtesy of Michael Franklin

Big Data Fact 1

- ~~Big Data is all hype.~~
- It may be hyped, but there is more than enough substance there for it to deserve our attention.

Big Data Myth 2

- Size is all that matters.
- Challenges are only at the extremes (in size).

What is Big Data

Gartner Definition:

- Volume
- Velocity
- Variety

- Veracity
- V..

Variety

- How do you even measure variety?
- No measure => hard to track progress

- “Infinite” variety on the web
  - You keep finding sites you have never seen before
- “Infinite” variety in human generated content

Veracity

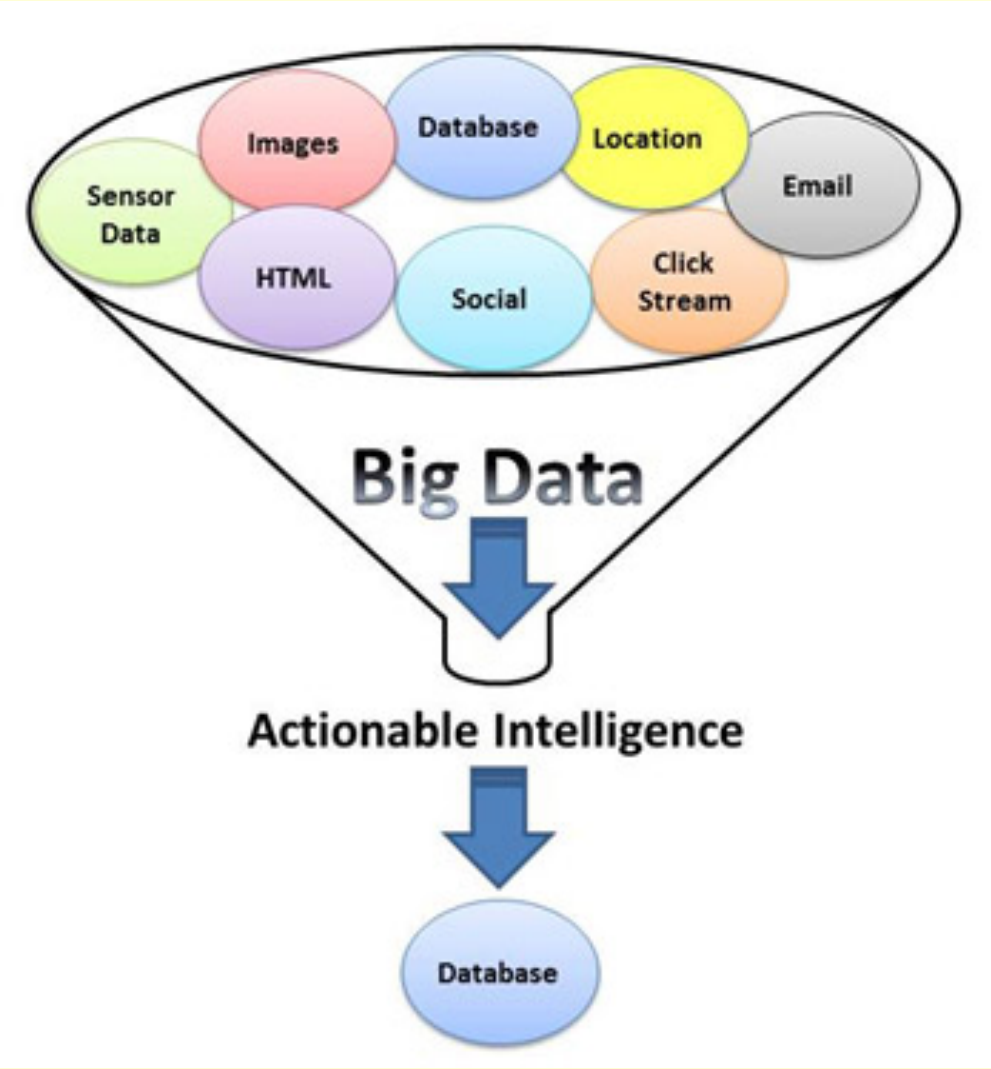
- Who do you trust?
  - Reputation on the web.
- Independence determination
  - When is it a new source and when is it a copy?

Big Data Fact 2

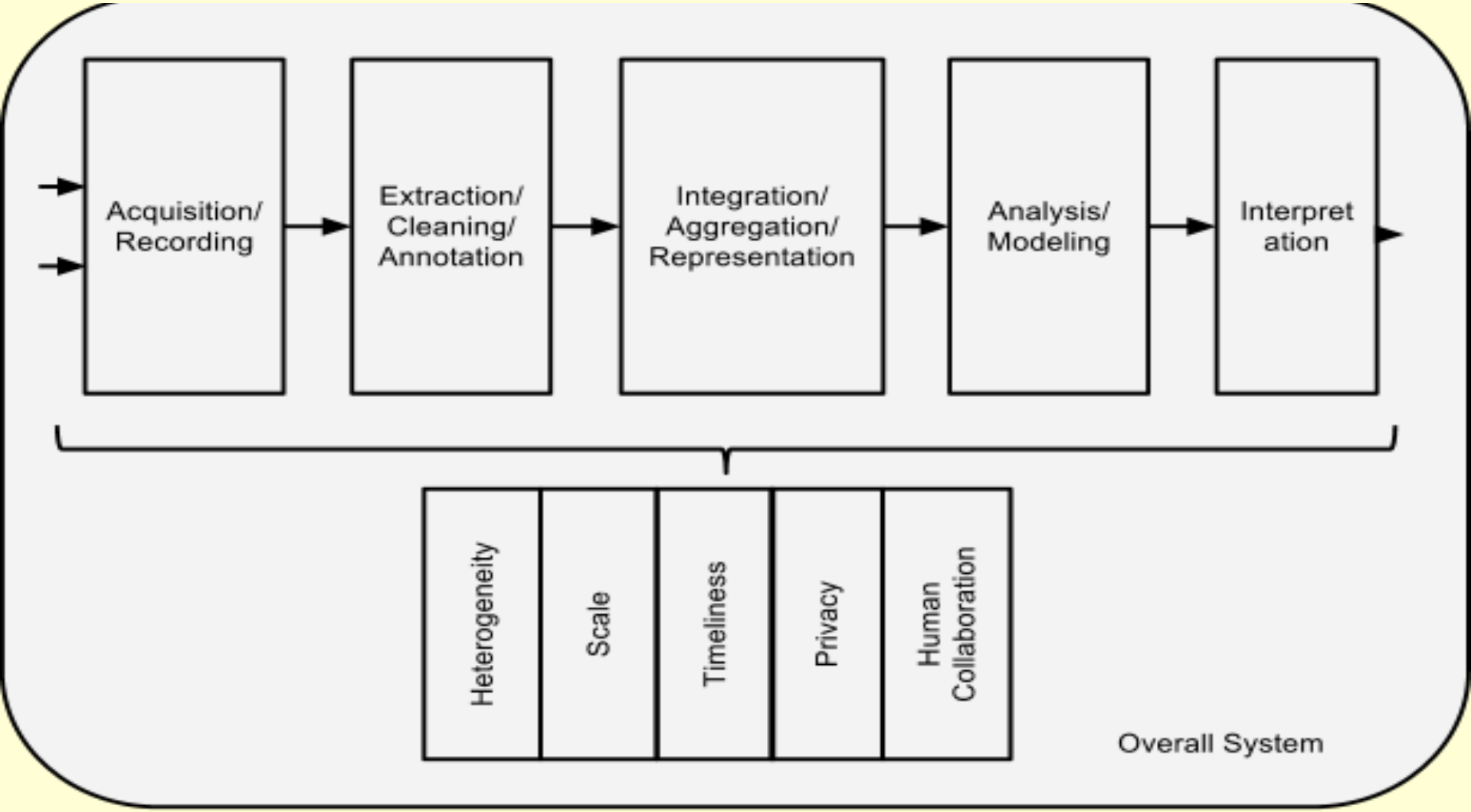
- ~~Size is all that matters.~~
- Yes, Volume and Velocity are challenging
- But Variety and Veracity are far more challenging

Big Data Myth 3

Companies Propagate This!!



From the web site of a representative silicon valley company



The Big Data Pipeline

Big Data Challenges

- In each of the steps

Read the whitepaper:

http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

Shorter version in CACM, July 2014.

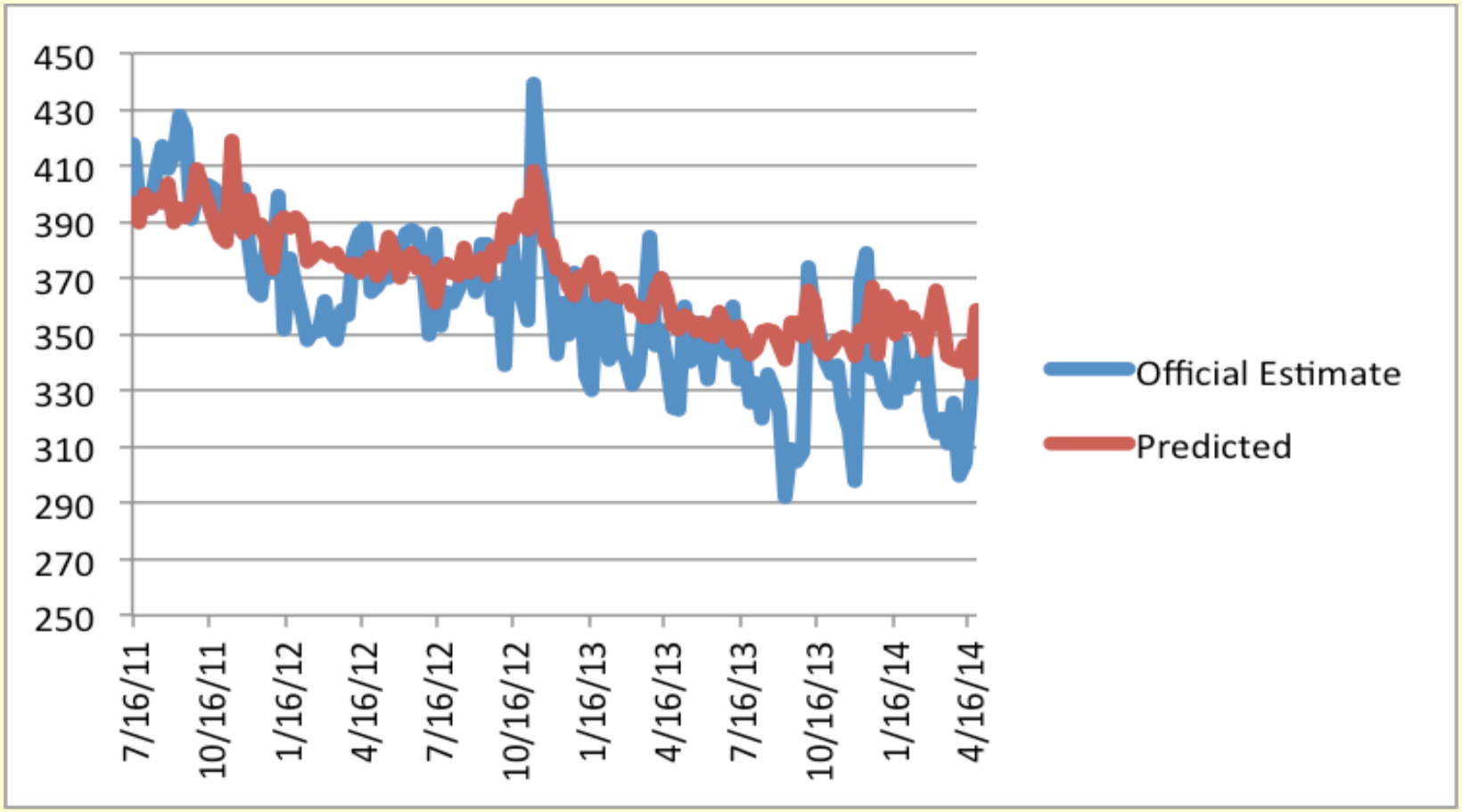
Big Data Fact 3

- Every aspect of the data ecosystem poses challenges that must be addressed.

Big Data Myth 4

- Data reuse is low hanging fruit
  - Lots of data collected for some purpose
  - Can (later) be used for a different purpose

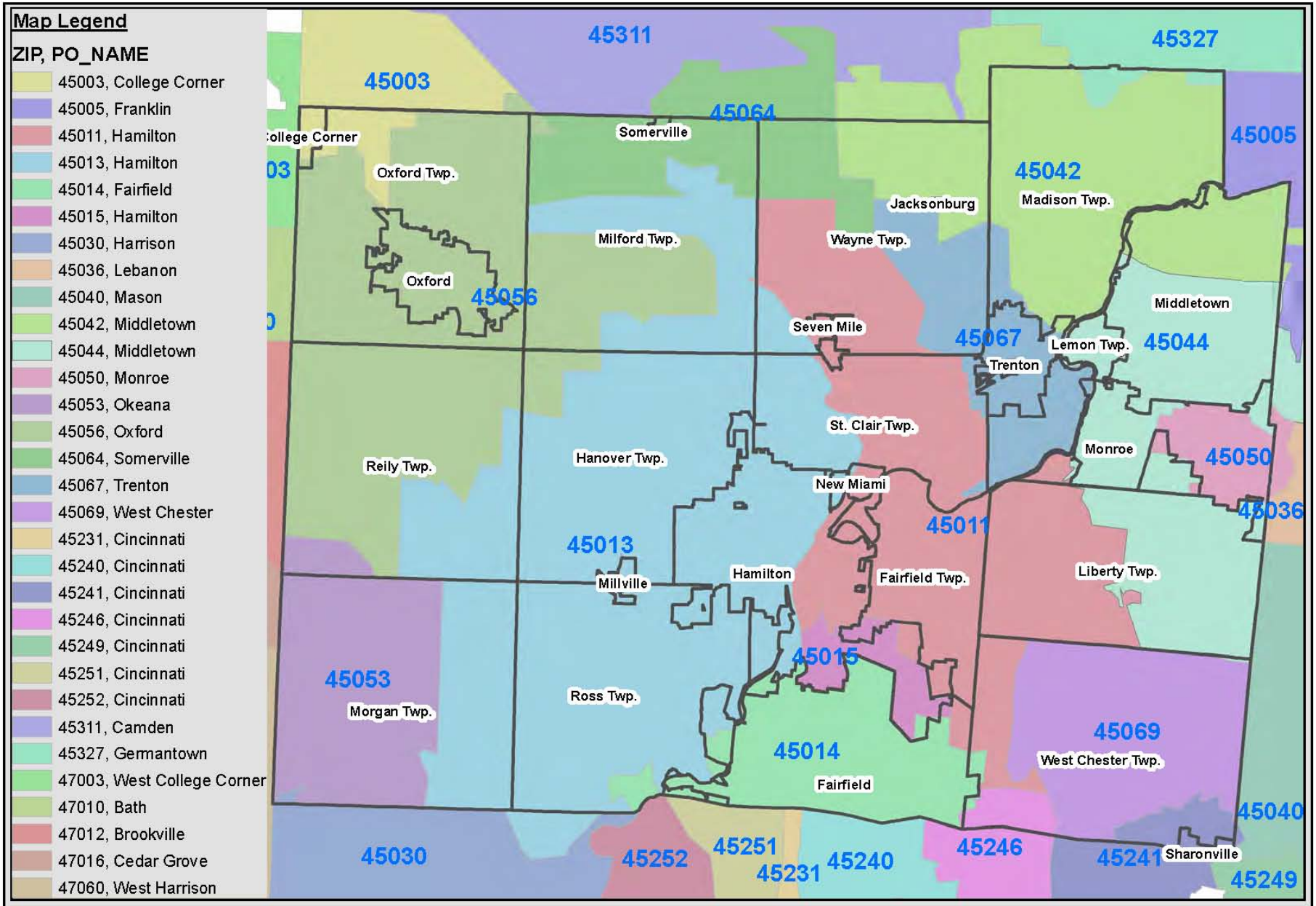
Unemployment Rate Prediction based on Tweets



Cafarella, Levenstein, Shapiro
http://econprediction.eecs.umich.edu/

Data is Organized “Wrong”

- E.g. administrative data is often rolled up by administrative jurisdiction.
- Consider Butler County, Ohio.



N W E S

1 inch = 16,889 feet

FOR INFORMATIONAL PURPOSES ONLY NOT INTENDED FOR USE AS A SURVEY

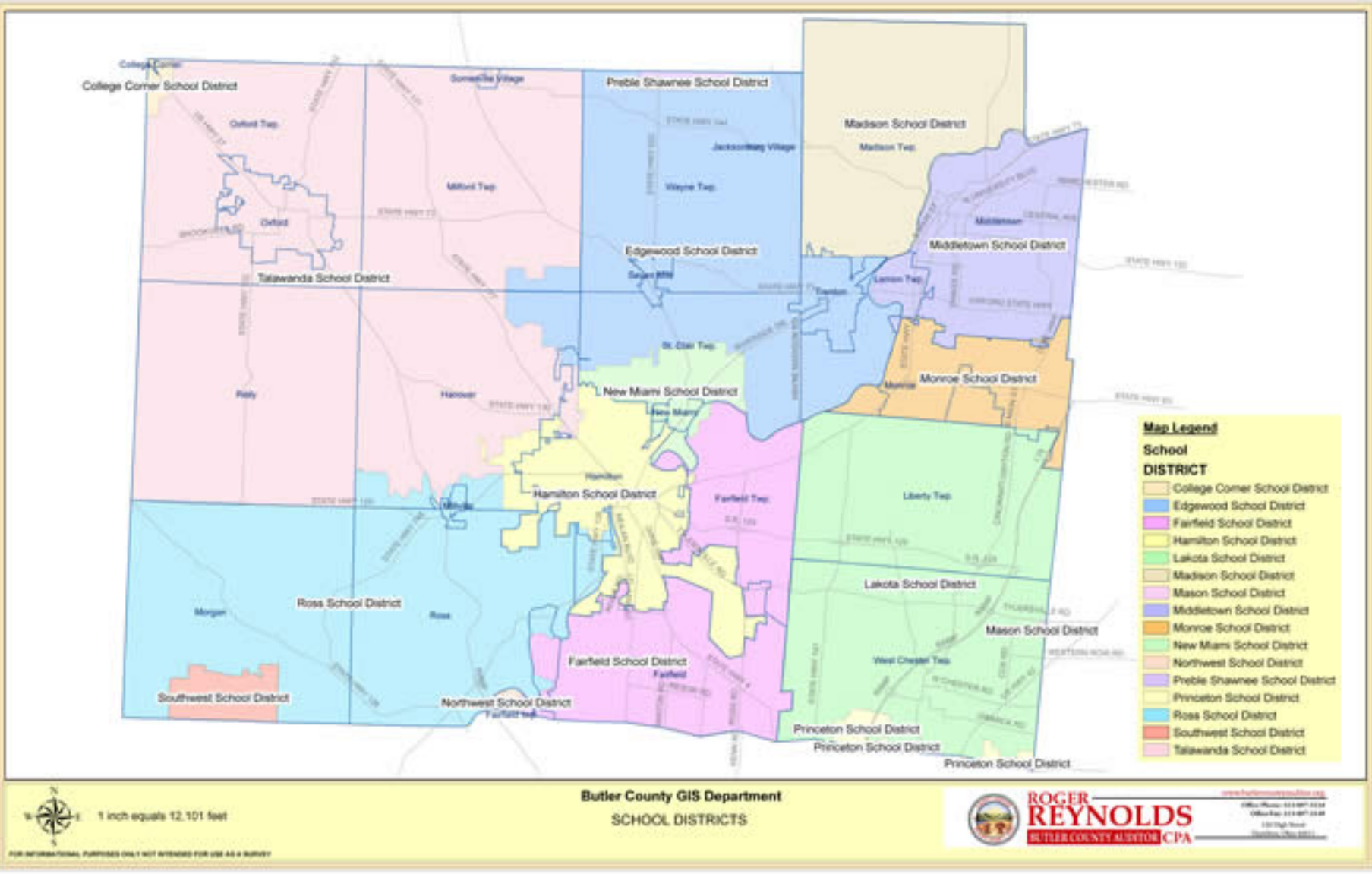
Butler County GIS Department

Zip Codes

Created: 2012

www.butlercountyauditor.org
130 High Street
Hamilton, Ohio, 45011
(513) 887-3154

ROGER REYNOLDS
BUTLER COUNTY AUDITOR CPA



1 inch equals 12,101 feet

FOR INFORMATIONAL PURPOSES ONLY NOT INTENDED FOR USE AS A SURVEY

Butler County GIS Department

SCHOOL DISTRICTS

ROGER REYNOLDS
BUTLER COUNTY AUDITOR CPA

Data is Organized “Wrong”

- E.g. administrative data is often rolled up by administrative jurisdiction.
- How to compare data rolled by school district with data rolled up by zip code?

- Working with Gates Foundation
- Create *estimated* data rolled up by desired jurisdiction.

Research Data Reuse

- Much data is now available
  - Strong push from federal agencies
  - Parallel push from reproducibility advocates
- But obstacles remain
  - Incentives to record metadata.
    - Very hard for third party to use otherwise
  - Data citation methodology and convention

Big Data Fact 4

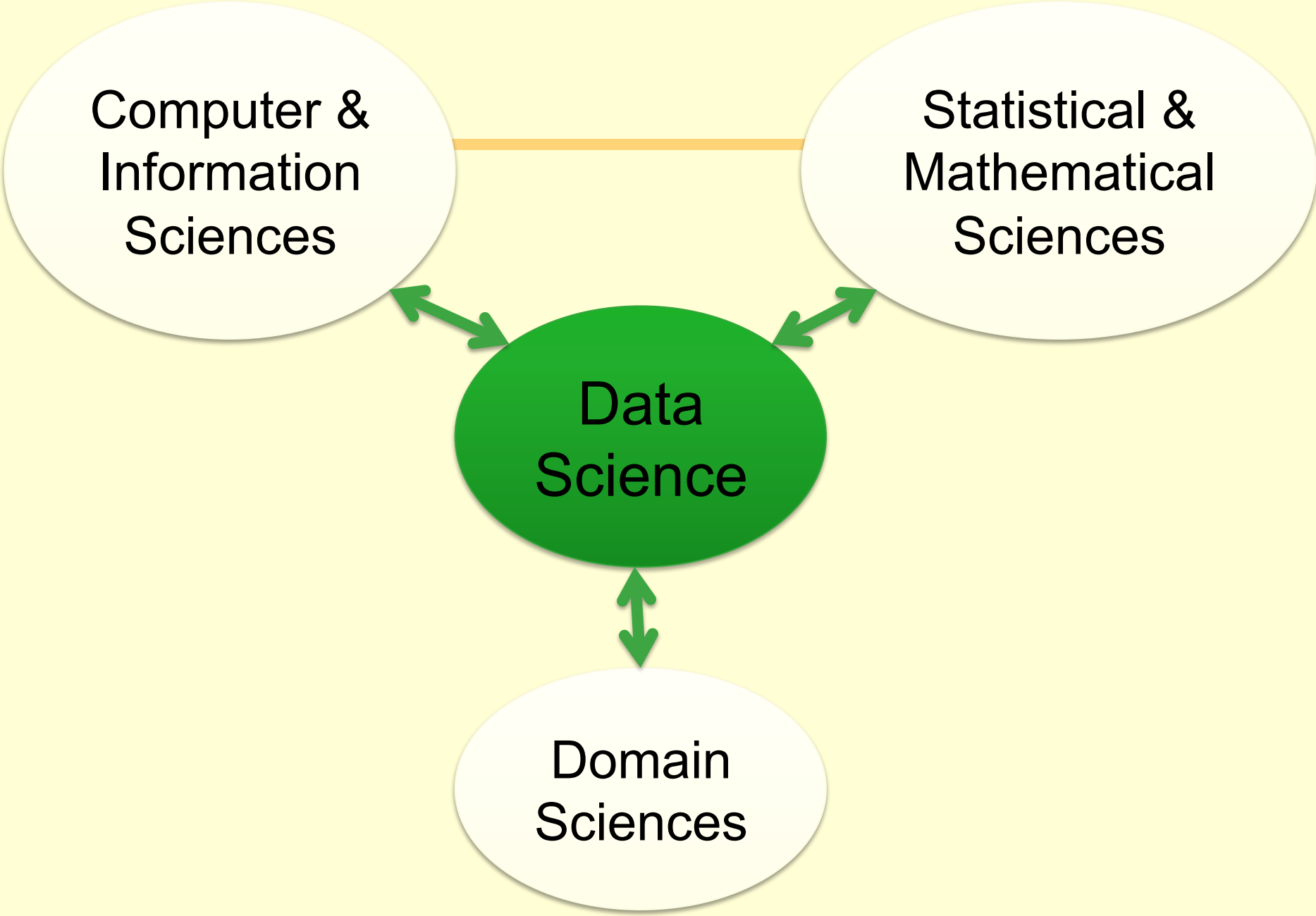
- ~~Data reuse is low hanging fruit~~
- Data reuse is critical to address
  - Holds out great promise
  - But also poses many challenging questions

Big Data Myth 5

- Data Science is the same as Big Data

Data Science

- The use of data to address problems in a domain of interest.
- Requires data management, data analysis, and domain knowledge.
- Often involves “Big Data”
- But may not …


Computer & Information Sciences
Statistical & Mathematical Sciences
Data Science
Domain Sciences

Data Science Status

- Importance widely recognized in academia.
  - Partly driven by employer demand
- Multi-disciplinary nature recognized.
- Common solution is to have some sort of structure that overlays and crosses traditional departments
  - E.g. http://minds.umich.edu

Big Data Fact 5

- ~~Data Science is the same as Big Data~~
- Data Science is related to, but different from, Big Data

Big Data Myth 6

- The central challenge with Big Data is that of devising new computing architectures and algorithms.

Big Data Myth 6 (reprise)

- Big Data is all in the cloud
- Big Data = Map Reduce style computation

What is Big Data

- Volume
- Velocity
- Variety
- Veracity

More than you know how to handle.

Humans and Big Data

- We can buy bigger systems, more machines, faster CPU, larger disks.
- But human ability does not scale!
- Big Data poses huge challenges for human interaction.

Usability for Data Science

- Data Science tasks usually involve data analysis by a domain expert with limited database expertise.
- If domain expert is to succeed, data must be usable.
- Usability matters most when the data are “big”.

Database Usability

- Improve user's ability to complete a task with a (big) database through better:
  - Query formulation
  - Result presentation
- HCI principles are very useful
- But, usability is not interface design.
- See http://www.eecs.umich.edu/db/usable

Big Data Fact 6

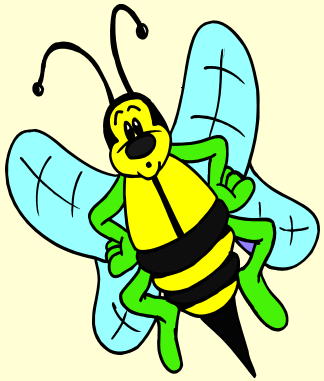
- ~~Big Data is all about the cloud.~~
- The cloud has its place in the constellation of relevant technologies, but is not a required piece of every solution.
- In fact, there are many other challenges that are at least as important
  - *cf.* National Academies report on "Frontiers of Massive Data Analysis"

Acknowledgments

- NSF Grants 1017296 and 1250880

Big Data and Data Science

- Lots of Buzz



- With good reason
  - Great potential
  - Many challenges