STAT121 / AC209 / E-109

CS109 Data Science

Hanspeter Pfister
pfister@seas.harvard.edu

Joe Blitzstein
blitzstein@stat.harvard.edu

Verena Kaynig
vkaynig@seas.harvard.edu

Outline

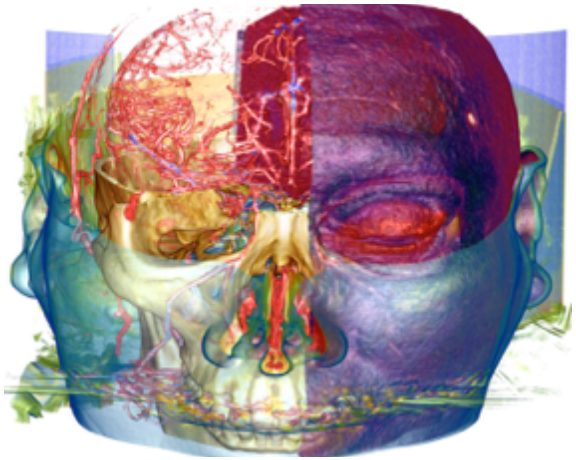
- What?
- Why?
- Who?
- How?

Outline

- What?
- Why?
- Who?
- How?

Data Science

To gain insights into data through computation, statistics, and visualization



A Data Scientist Is...

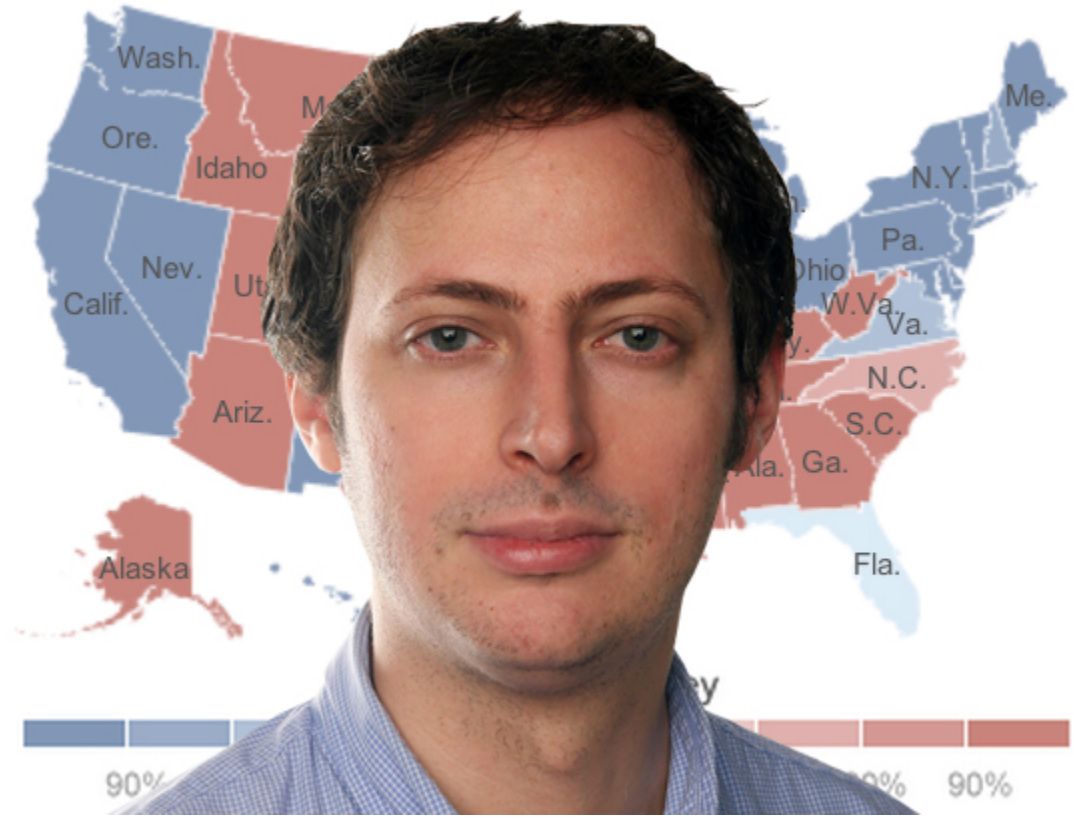
"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."

- Josh Blumenstock

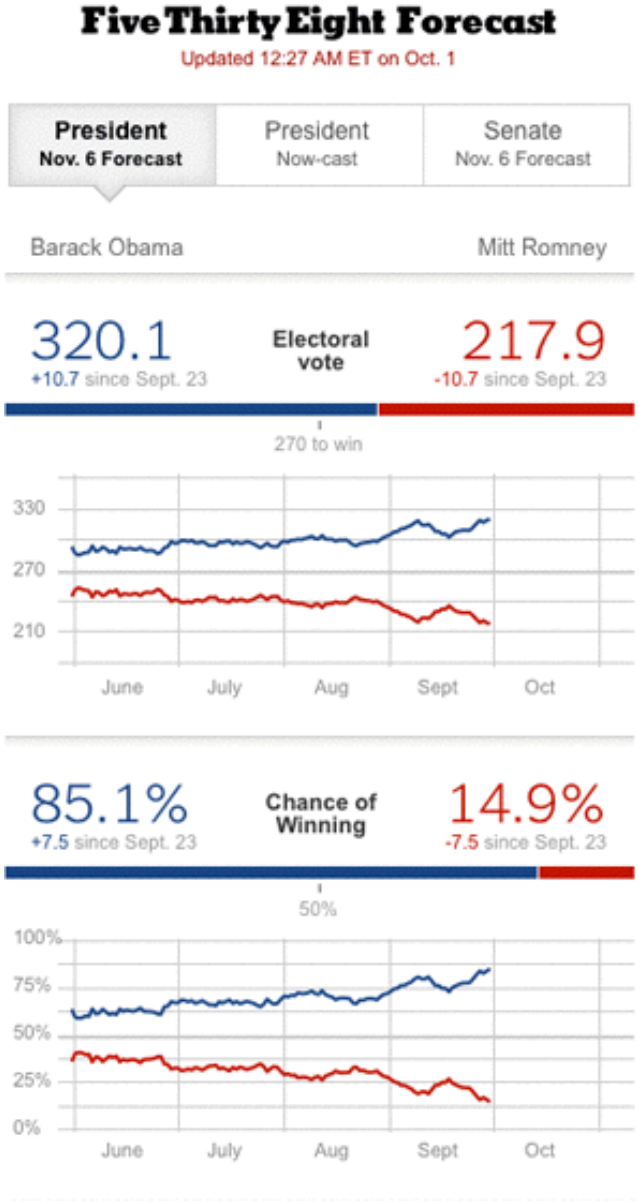
"Data Scientist = statistician + programmer + coach + storyteller + artist"

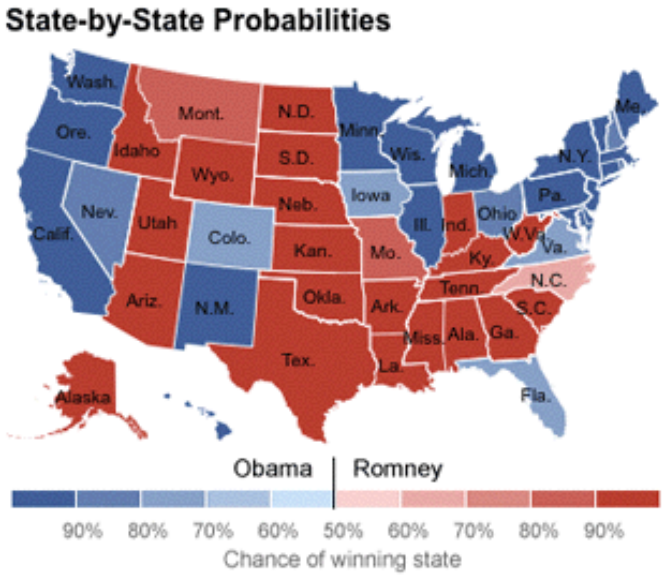
- Shlomo Aragmon

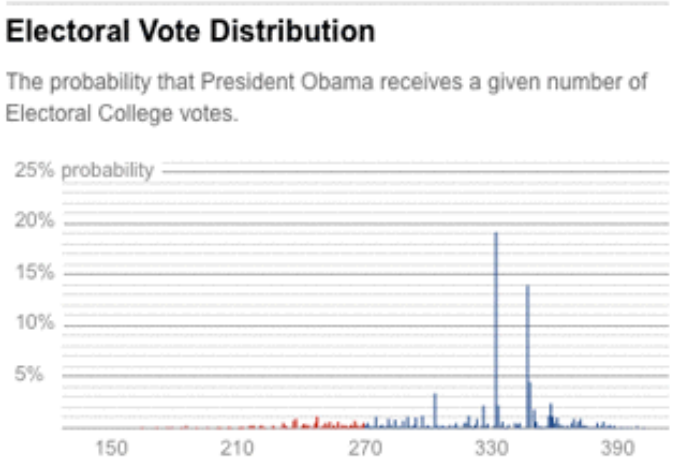
Nate Silver


Wash.
Ore.
Idaho
Nev.
Calif.
Ariz.
Alaska
Me.
N.Y.
Pa.
W.Va.
Va.
N.C.
S.C.
Ga.
Fla.
90%
90%

“Nate Silver won the election” – Harvard Business Review







#natesilverfacts

Ben Hamner @benhamner 7 Nov

#natesilverfacts: Nate Silver doesn't update according to priors, priors update according to Nate Silver @mattcutts

Expand Reply Retweet Favorite ••• More

DLDahly Epidemiology @statsepi 7 Nov

#natesilverfacts Nate Silver's models fit the test data even better than the training data.

citizenrobot @citizenrobot 7 Nov

Nate Silver knows when GRR Martin will finish the Winds of Winter #NateSilverFacts

William Chen @wzchen 11 Nov

There is no such thing as missing data, only data that Nate Silver has not chosen to reveal to you. #natesilverfacts

Retweeted by Rodrigo Aldecoa and 1 other

Expand Reply Retweeted Favorite ••• More

Is Election Predictor Nate Silver A Witch? Probably. And Quantified Self Data Will Make You One Too

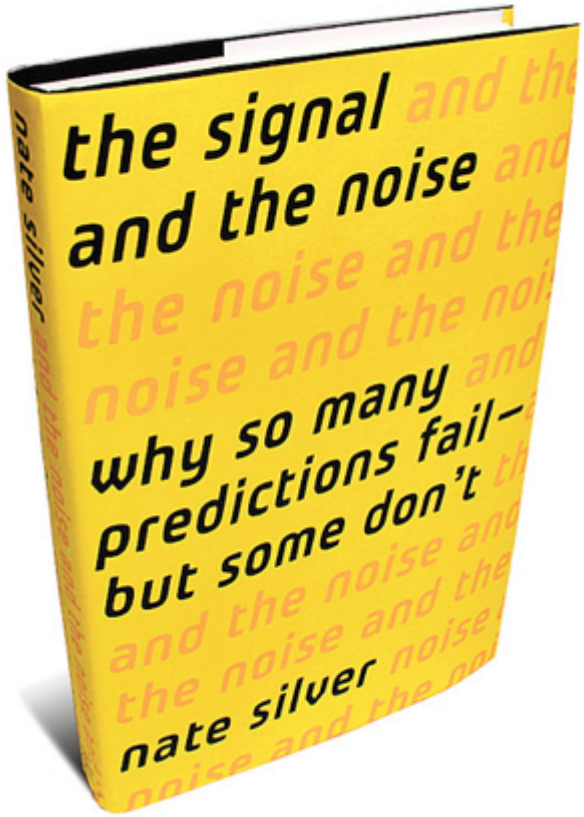
JOSH CONSTINE

Wednesday, November 7th, 2012 7 Comments

Scientists are yesterday's wizards and demigods. And Nate Silver is a scientist. One whose ability to **predict the outcome of elections** is so precise, it's nearly indistinguishable from magic. That's why **IsNateSilverAWitch.com** is so funny. But really what his flawless prediction of the presidential election signifies is the coming of age of the quantified universe.

http://techcrunch.com/2012/11/07/nate-silver-as-software/

Nate Silver on Pundits



Silver: "Pundits are no better than a coin toss."

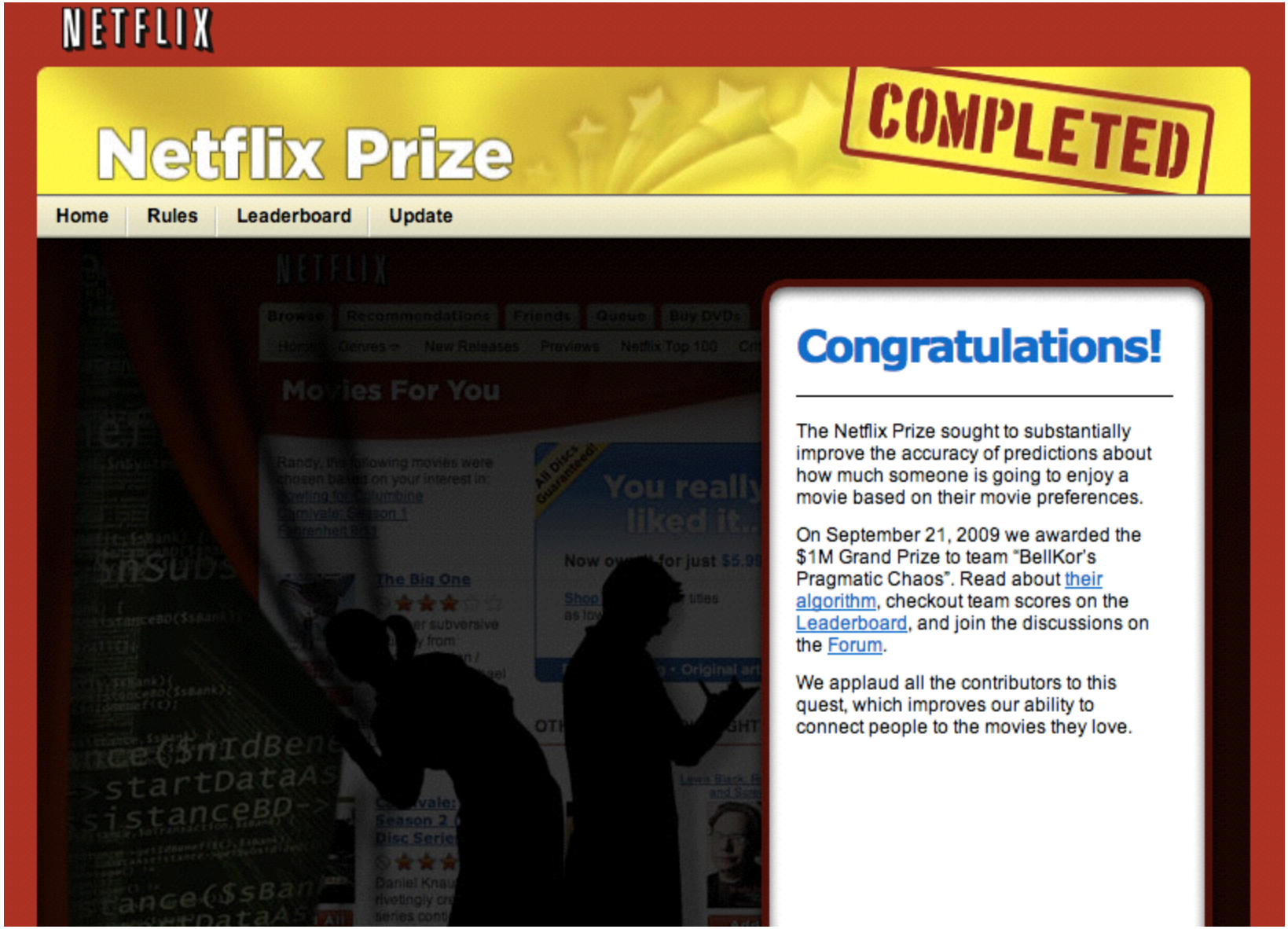
Stewart: "Do you foresee a coin getting its own show? The coin toss show?"

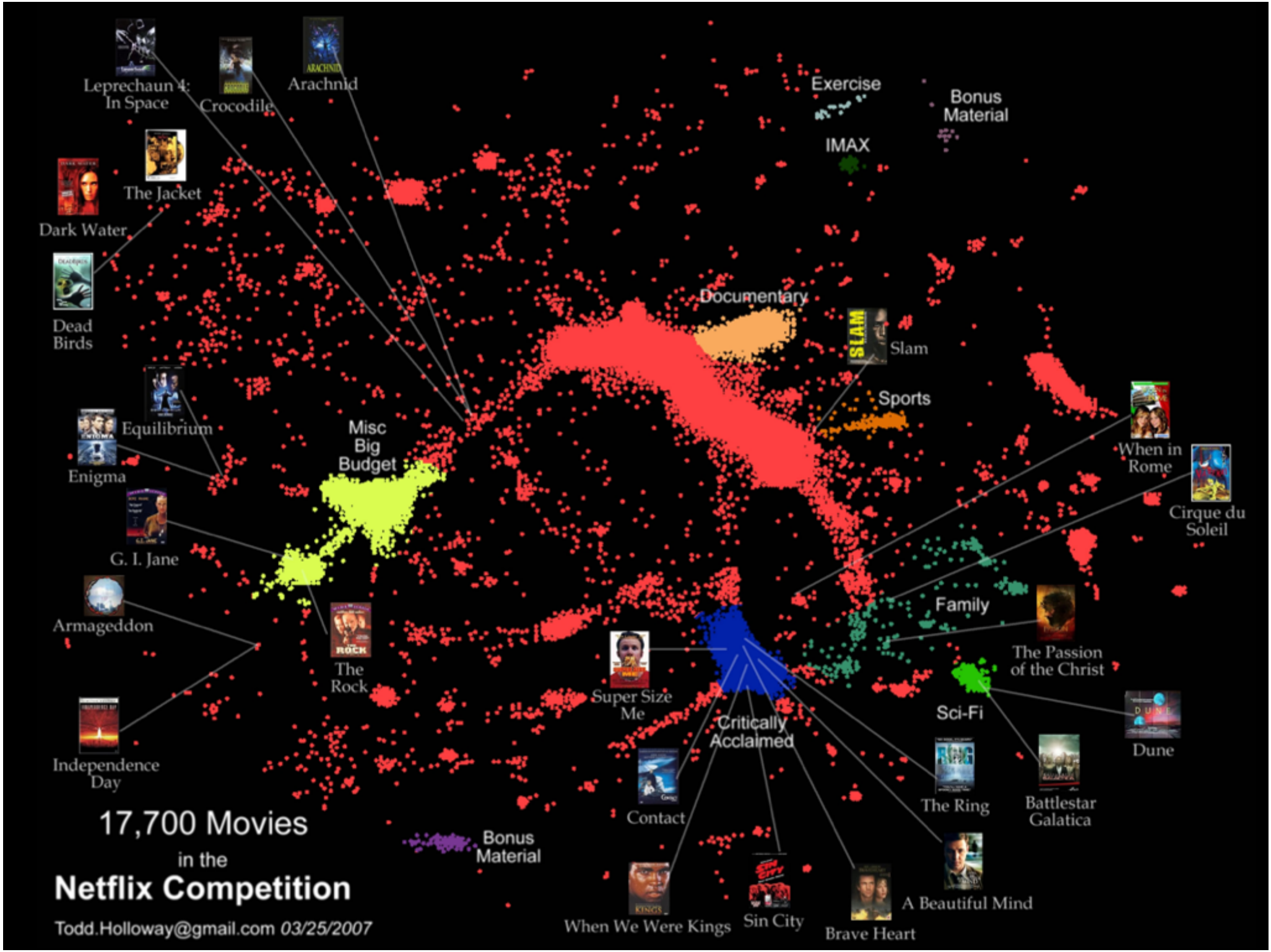
http://www.thedailyshow.com/watch/wed-october-17-2012/nate-silver

Some Key Principles

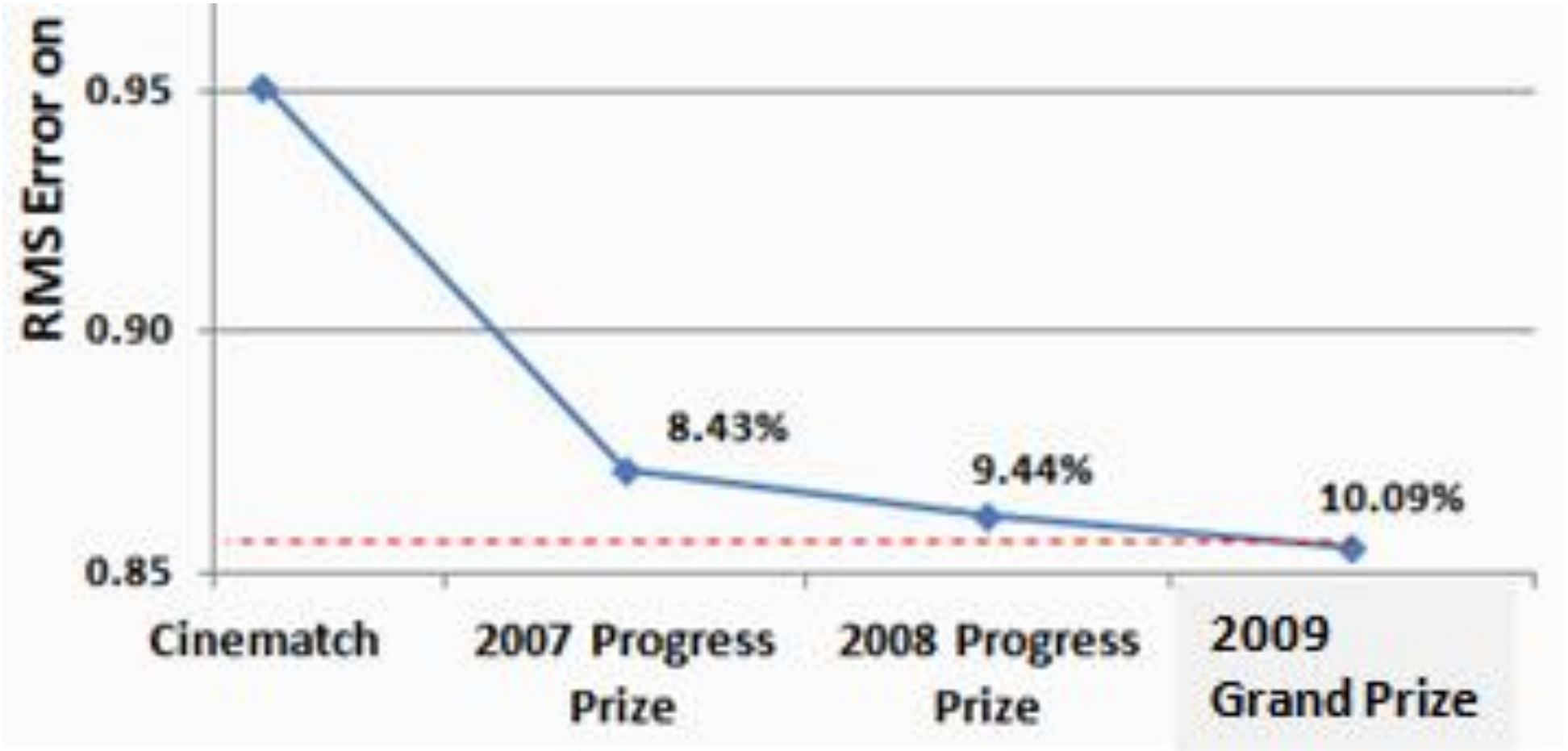
- *use many data sources* (the plural of anecdote is not data)
- *understand how the data were collected* (sampling is essential)
- *weight the data thoughtfully* (not all polls are equally good)
- *use statistical models* (not just hacking around in Excel)
- *understand correlations* (e.g., states that trend similarly)
- *think like a Bayesian, check like a frequentist* (reconciliation)
- *have good communication skills* (What does a 60% probability even mean? How can we visualize, validate, and understand the conclusions?)

Netflix Prize



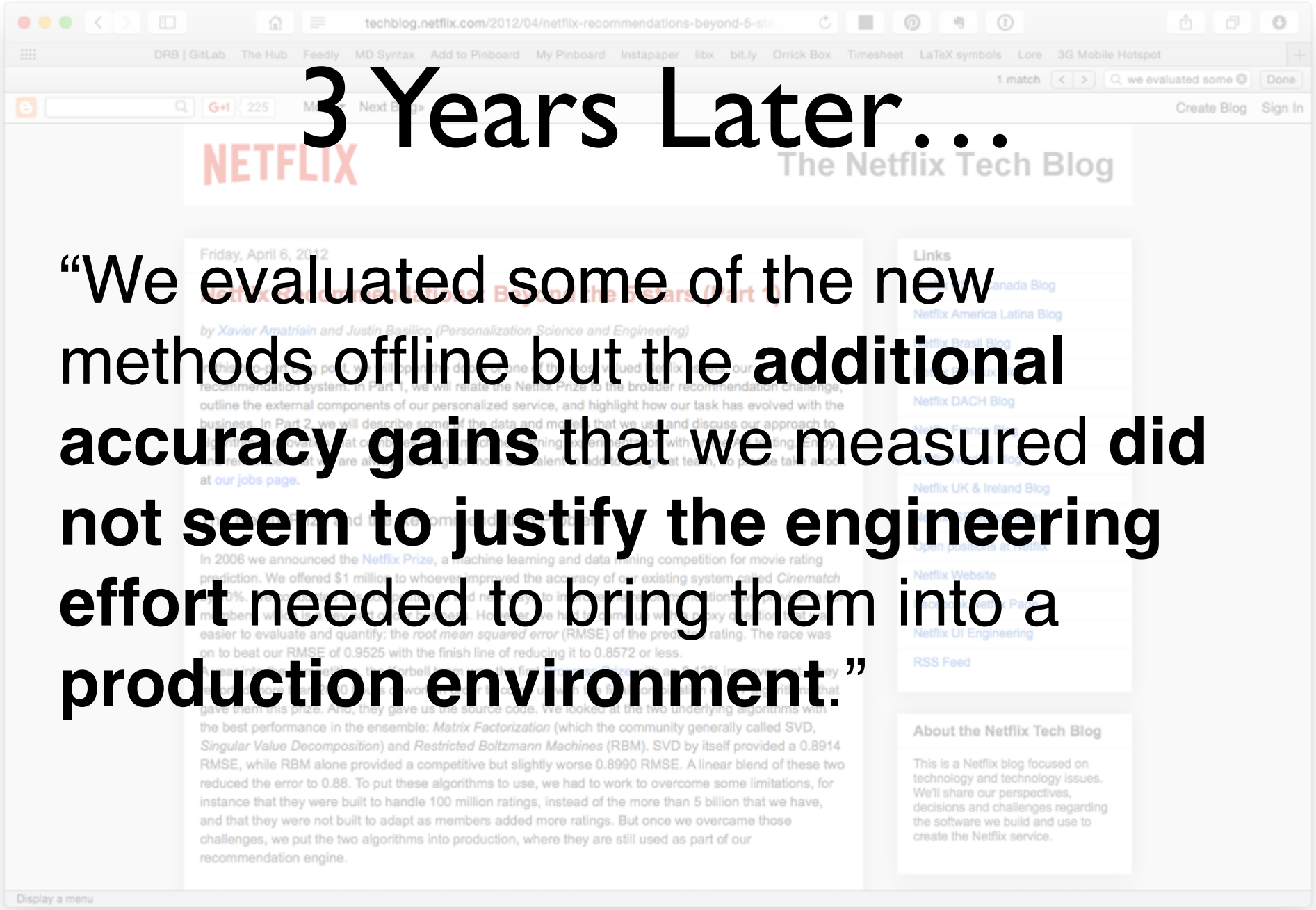

Leprechaun 4: In Space
Crocodile
Arachnid
Exercise
Bonus Material
IMAX
The Jacket
Dark Water
Dead Birds
Documentary
Slam
Sports
Equilibrium
Misc Big Budget
Enigma
When in Rome
Cirque du Soleil
G. I. Jane
Armageddon
Family
The Rock
The Passion of the Christ
Super Size Me
Critically Acclaimed
Sci-Fi
Dune
Independence Day
The Ring
Battlestar Galatica
Contact
17,700 Movies
in the
Netflix Competition
Bonus Material
A Beautiful Mind
When We Were Kings
Sin City
Brave Heart
Todd.Holloway@gmail.com 03/25/2007

Netflix Prize Progress



HBR, Oct 2012

3 Years Later…



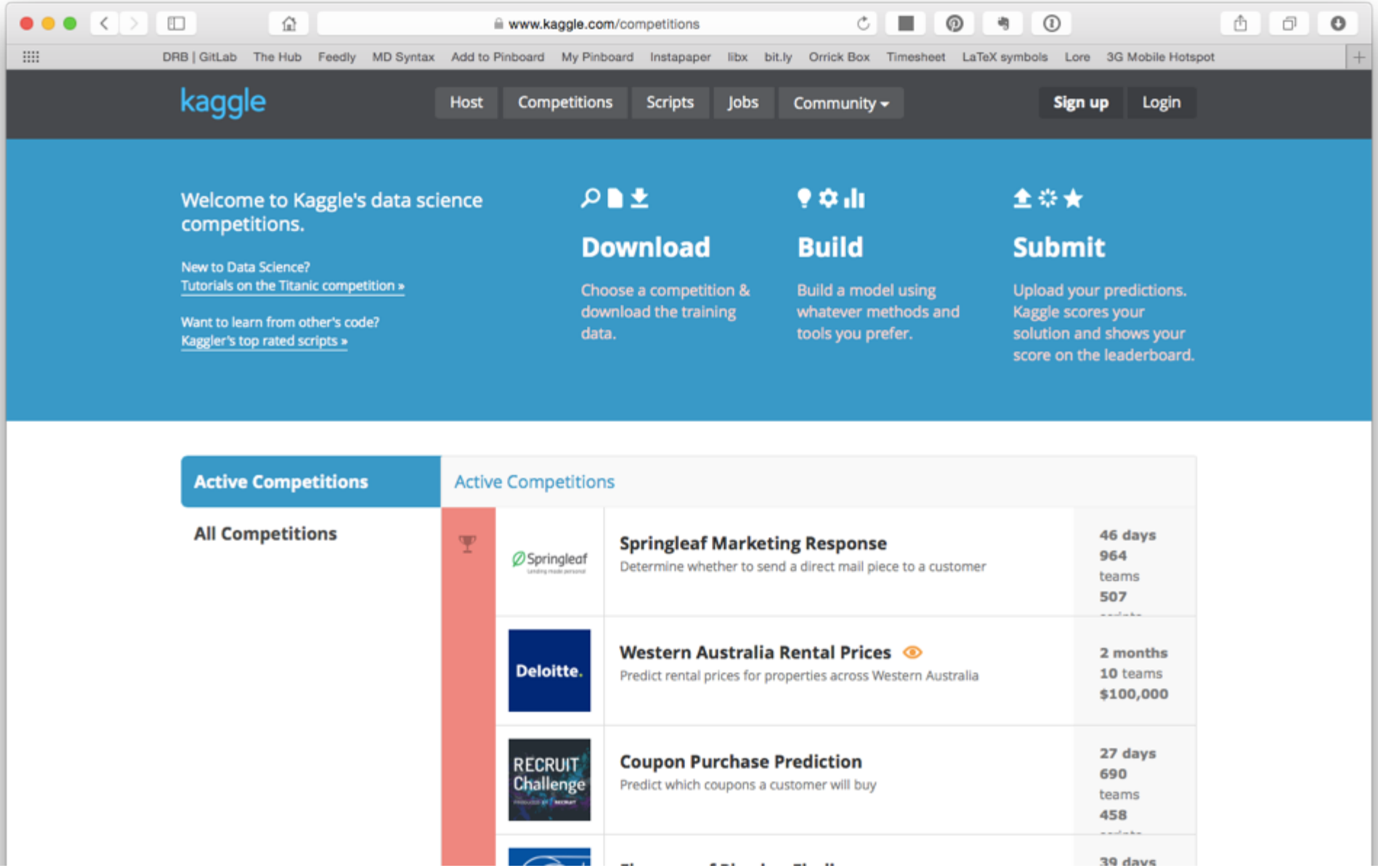
"We evaluated some of the new methods offline but the **additional accuracy gains** that we measured **did not seem to justify the engineering effort** needed to bring them into a **production environment**."

Xavier Amatriain and Justin Basilico, 2012

Some Challenges

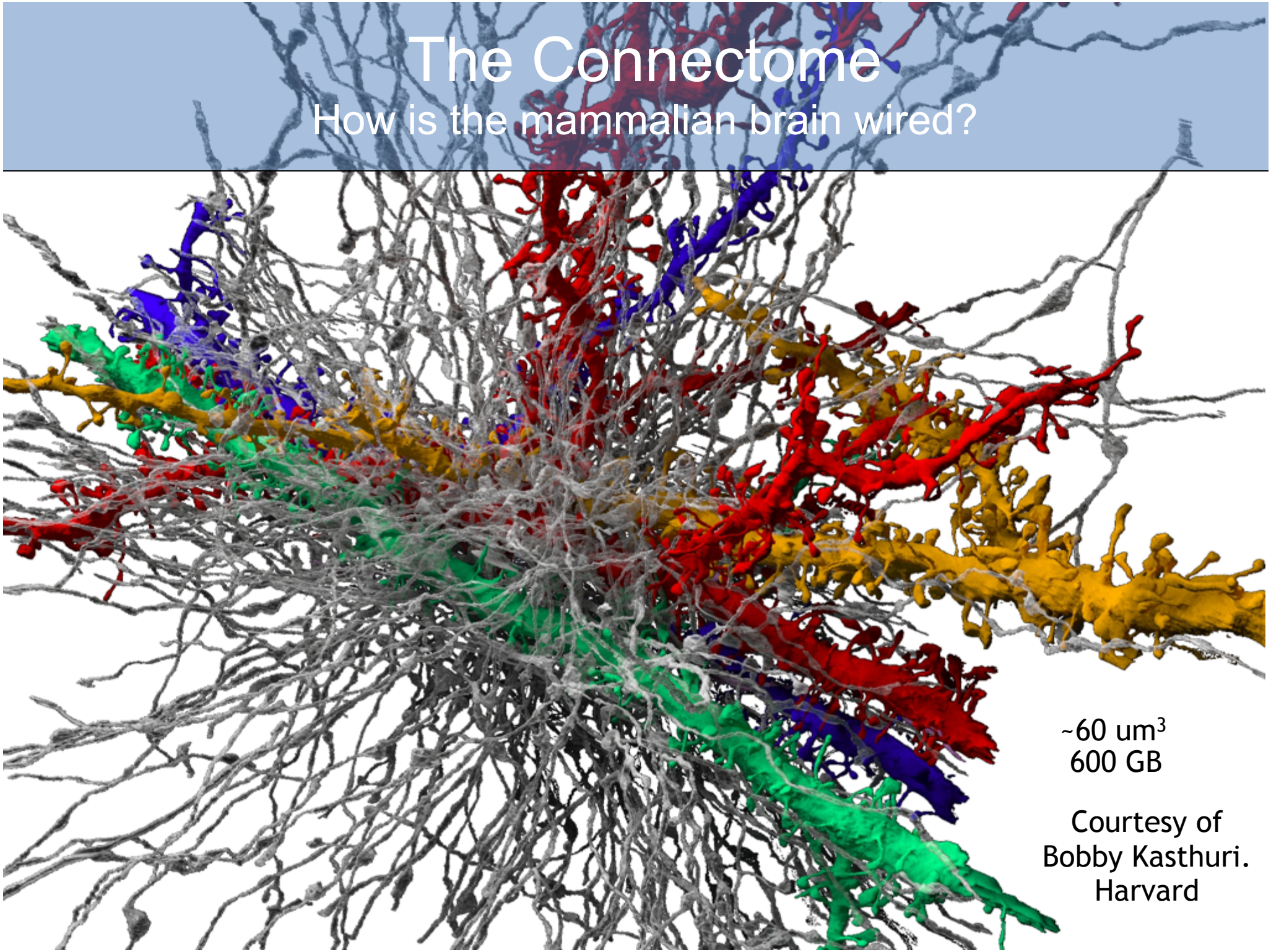
- *massive data* (500k users, 20k movies, 100m ratings)
- *curse of dimensionality* (very high-dimensional problem)
- *missing data* (99% of data missing; *not* missing at random)
- *extremely complicated set of factors that affect people's ratings of movies* (actors, directors, genre, ...)
- *need to avoid overfitting* (test data vs. training data)

Kaggle


www.kaggle.com/competitions
DRB | GitLab
The Hub
Feedly
MD Syntax
Add to Pinboard
My Pinboard
Instapaper
libx
bit.ly
Orrick Box
Timesheet
LaTeX symbols
Lore
3G Mobile Hotspot
kaggle
Host
Competitions
Scripts
Jobs
Community
Sign up
Login
Welcome to Kaggle's data science competitions.
New to Data Science?
Tutorials on the Titanic competition »
Want to learn from other's code?
Kaggler's top rated scripts »
Download
Choose a competition & download the training data.
Build
Build a model using whatever methods and tools you prefer.
Submit
Upload your predictions. Kaggle scores your solution and shows your score on the leaderboard.
Active Competitions
All Competitions
Active Competitions
Springleaf Marketing Response
Determine whether to send a direct mail piece to a customer
46 days
964
teams
507
Western Australia Rental Prices
Predict rental prices for properties across Western Australia
2 months
10 teams
$100,000
Coupon Purchase Prediction
Predict which coupons a customer will buy
27 days
690
teams
458

The Connectome

How is the mammalian brain wired?

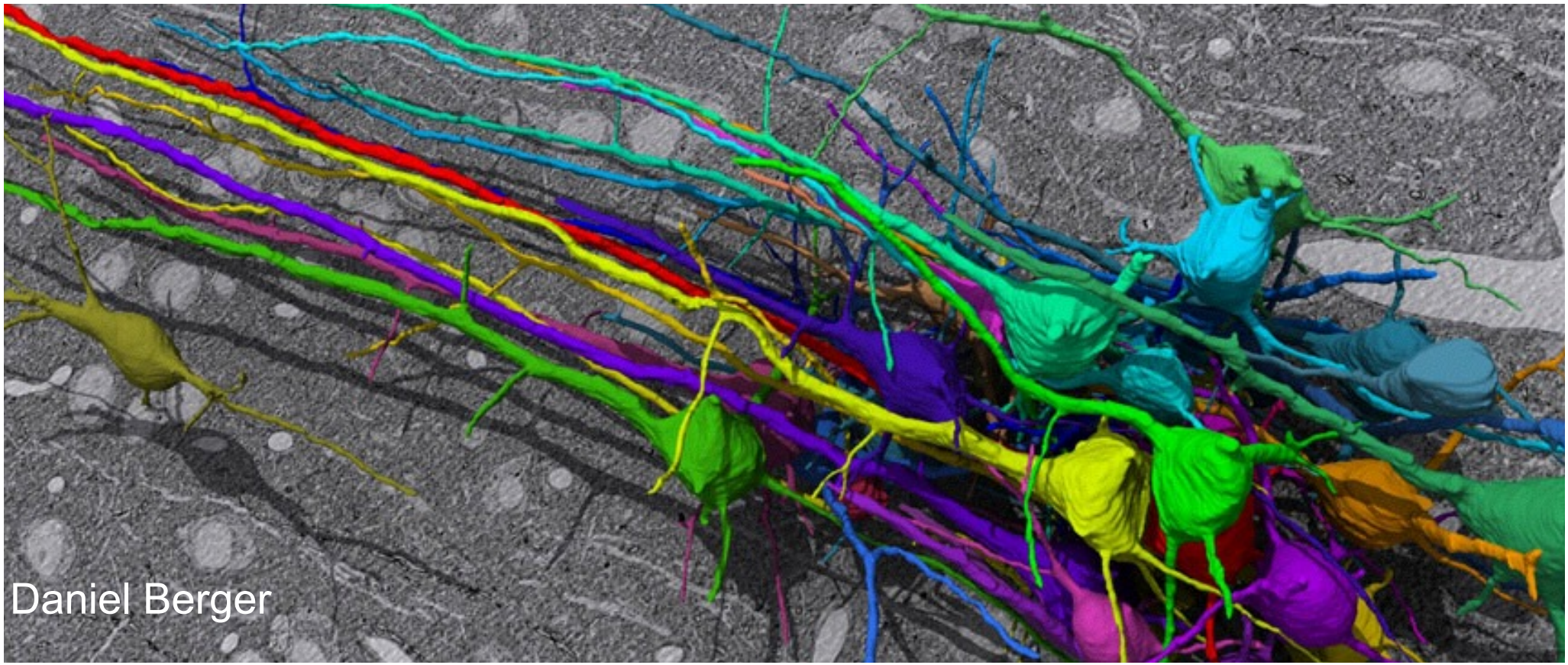


~60 um^3
600 GB

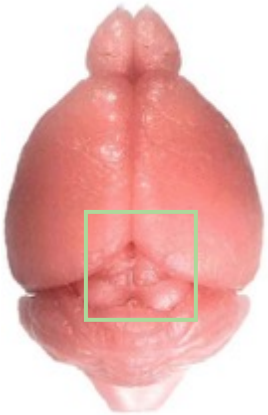
Courtesy of
Bobby Kasthuri.
Harvard

The Data Challenge

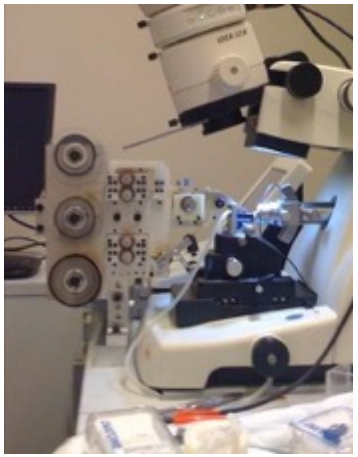
- Pixel resolution: 3-5 nm; Slice thickness: 30-50 nm
- 1 mm^3: 40 Gpixels x 25,000 slices = **~1 PByte**



Connectome Workflow

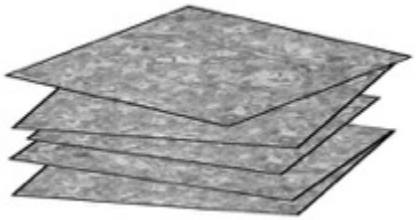


Cutting



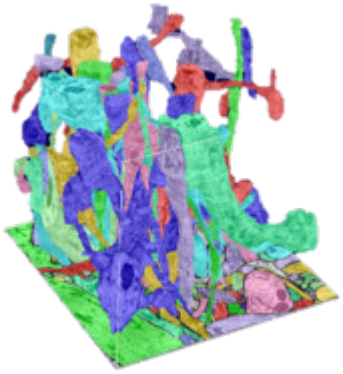
Imaging

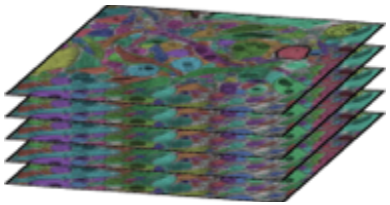
Alignment &
Registration



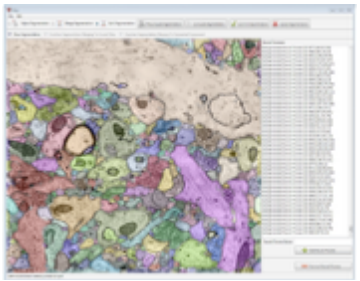
Segmentation

Visualization

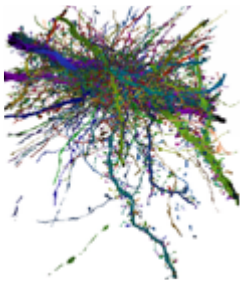


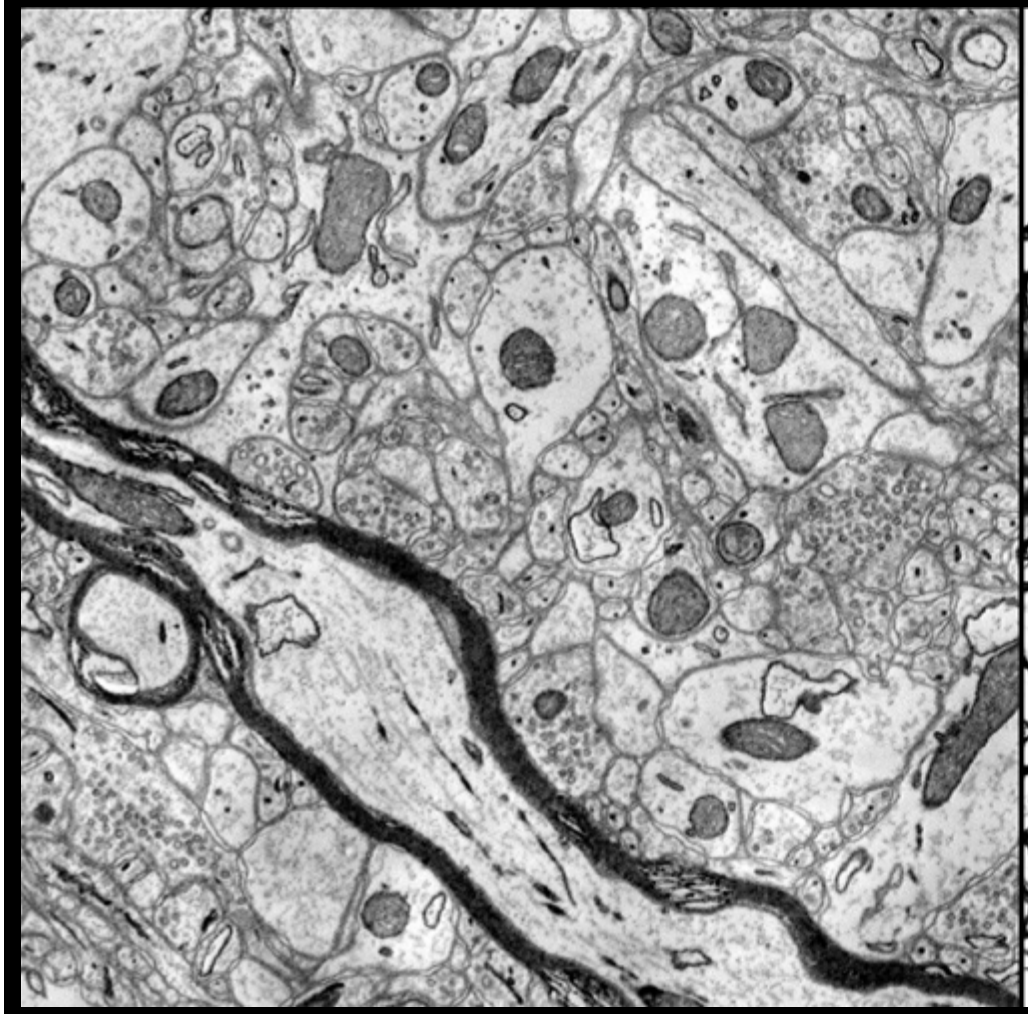


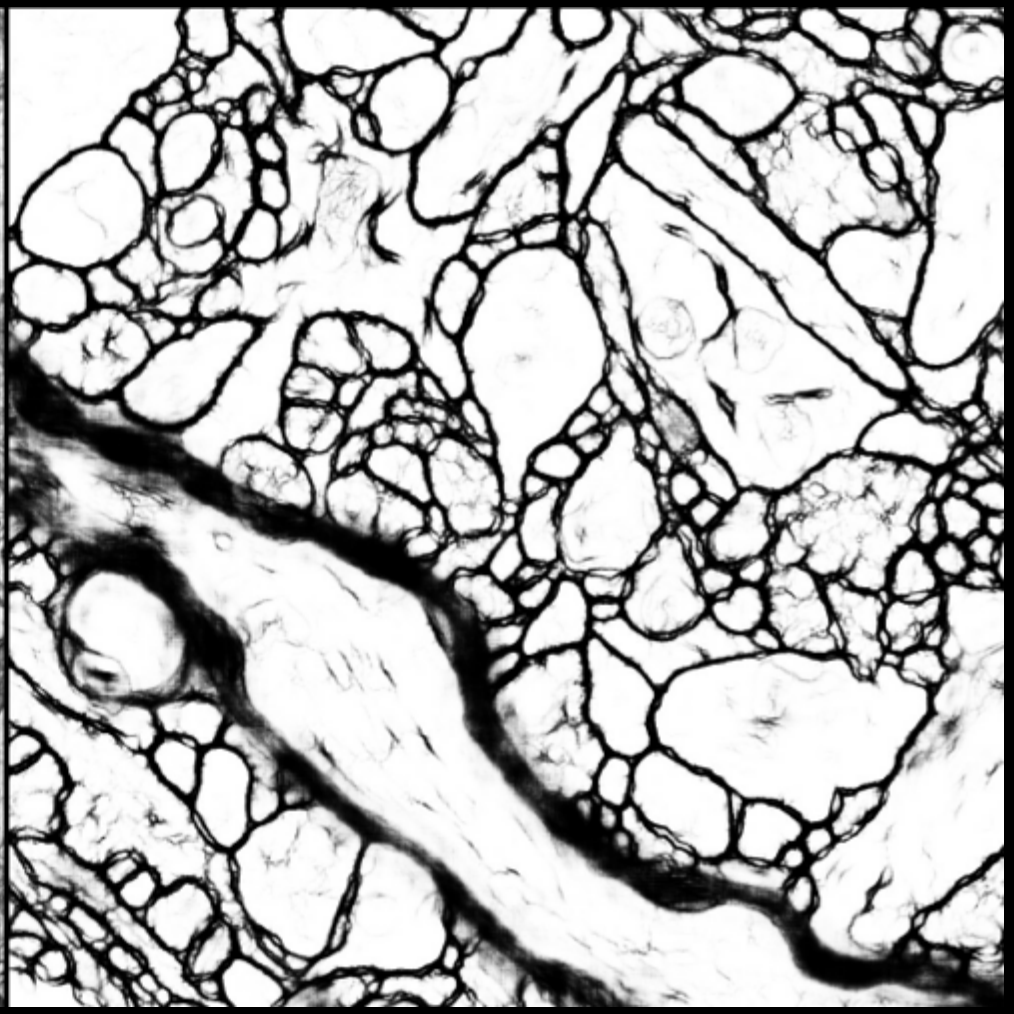
Proof Reading

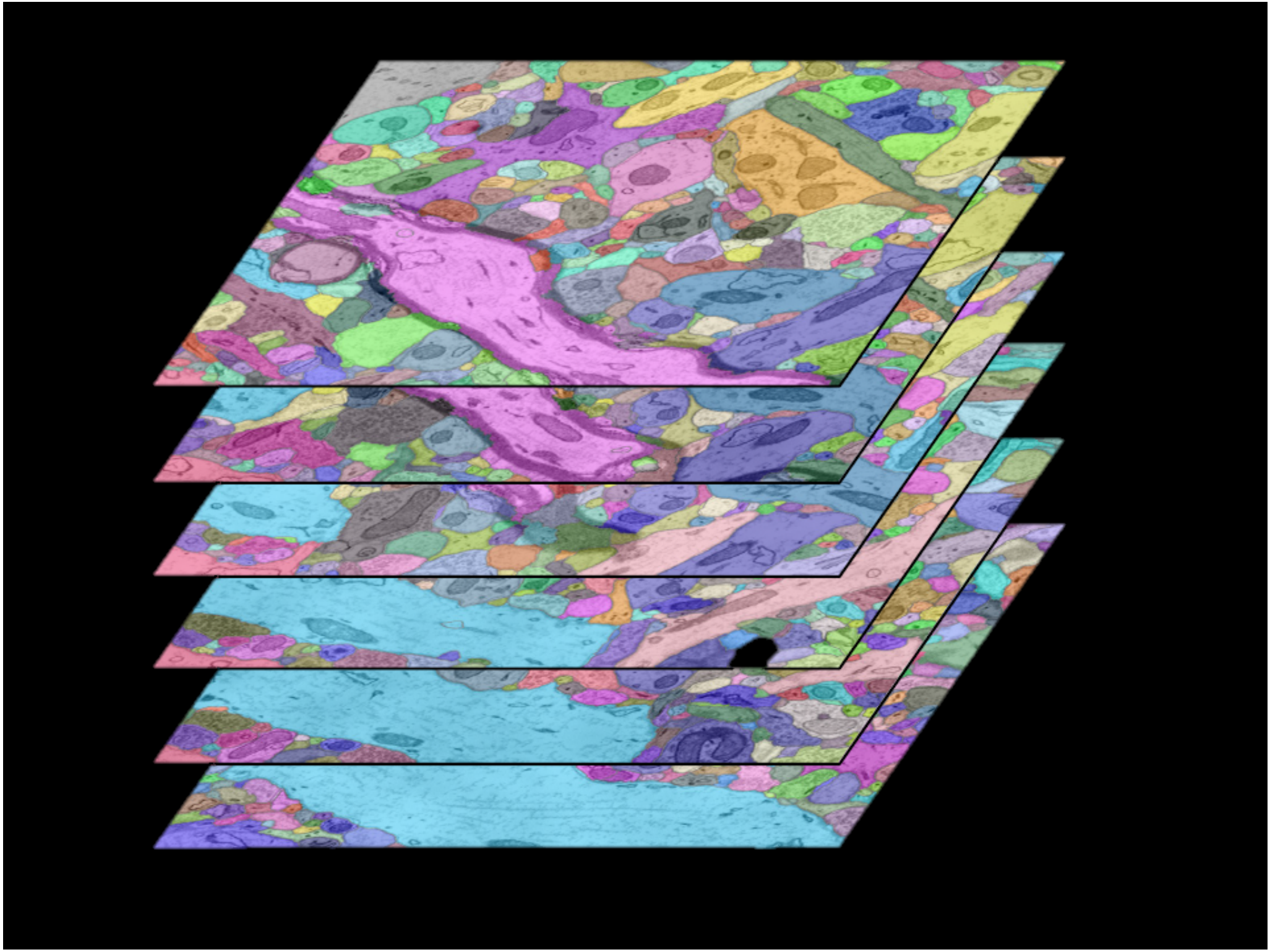


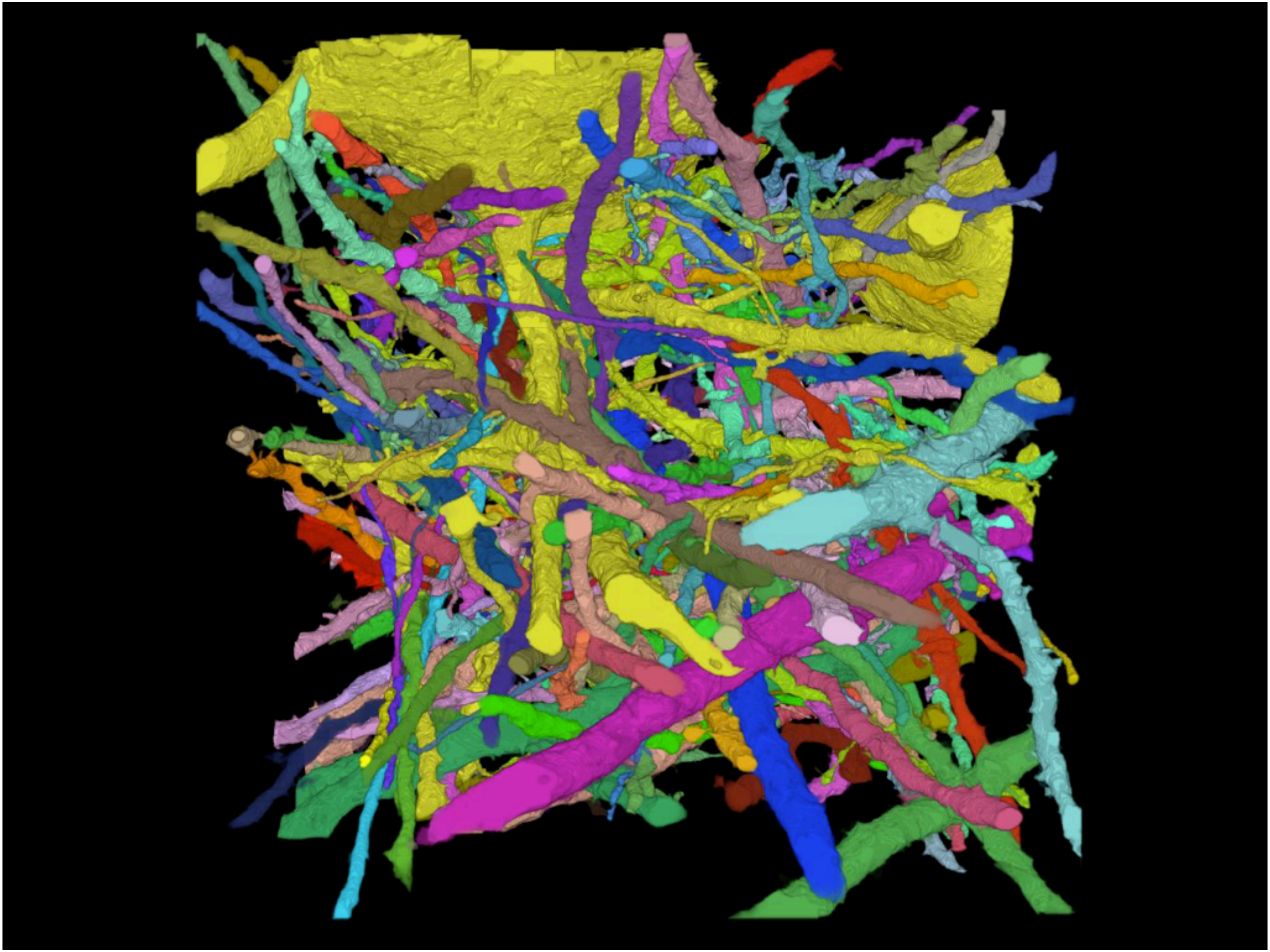
Analysis



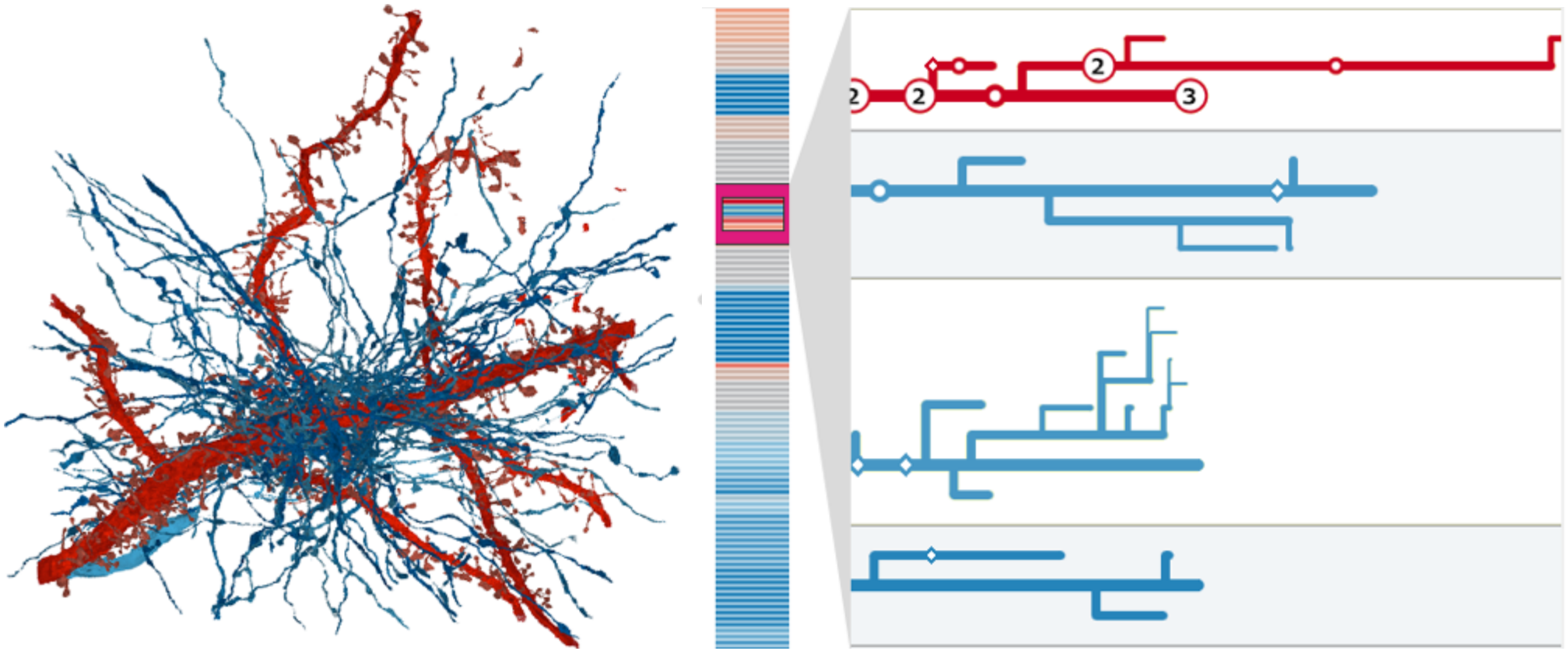








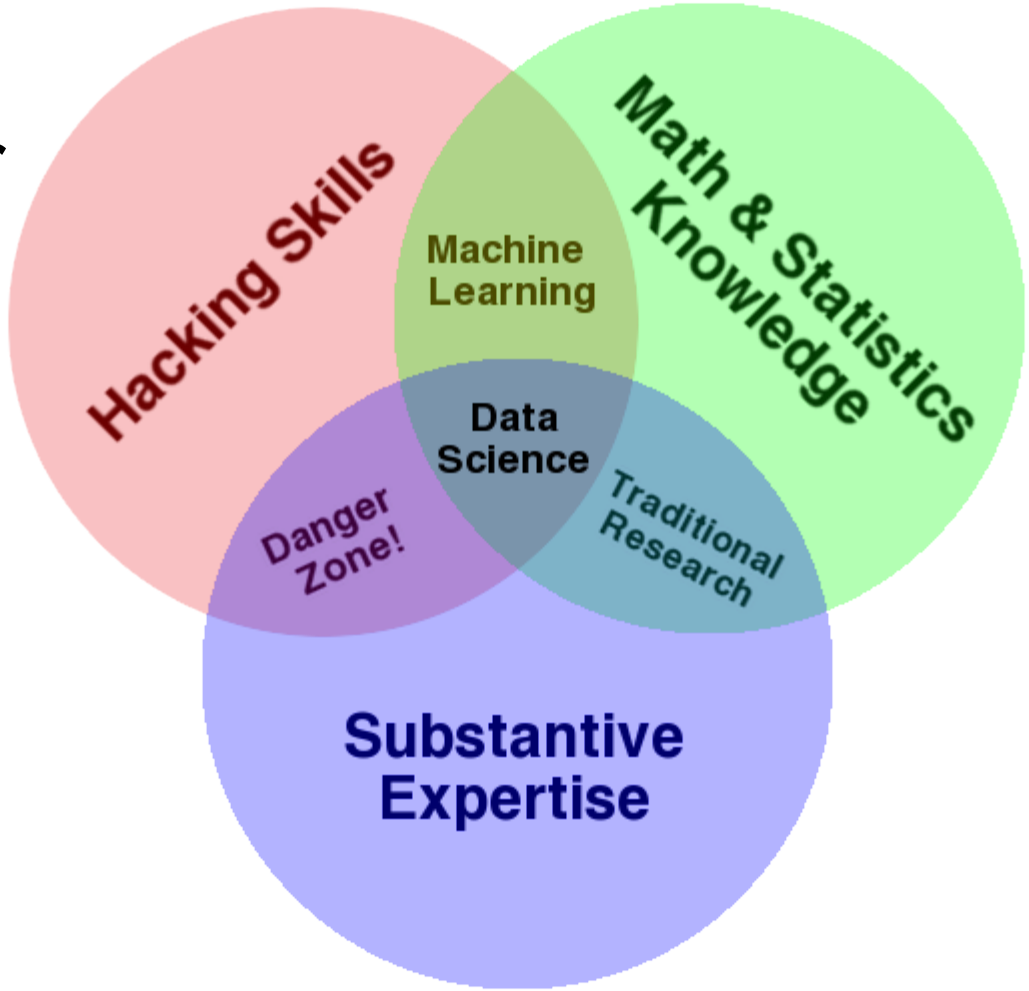
Analysis



K. Al-Awami, et al.,
"NeuroLines: A Subway Map Metaphor for Visualizing Nanoscale Neuronal Connectivity
IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 2369-2378,
2014

Data Science

Computer Science



Statistics

Domain Science

Drew Conway

Machine **Human**

Data Management

Human Cognition

Data Mining

Perception

Machine Learning

Visualization

Story Telling

Business Intelligence

Decision Making

Theory

Statistics

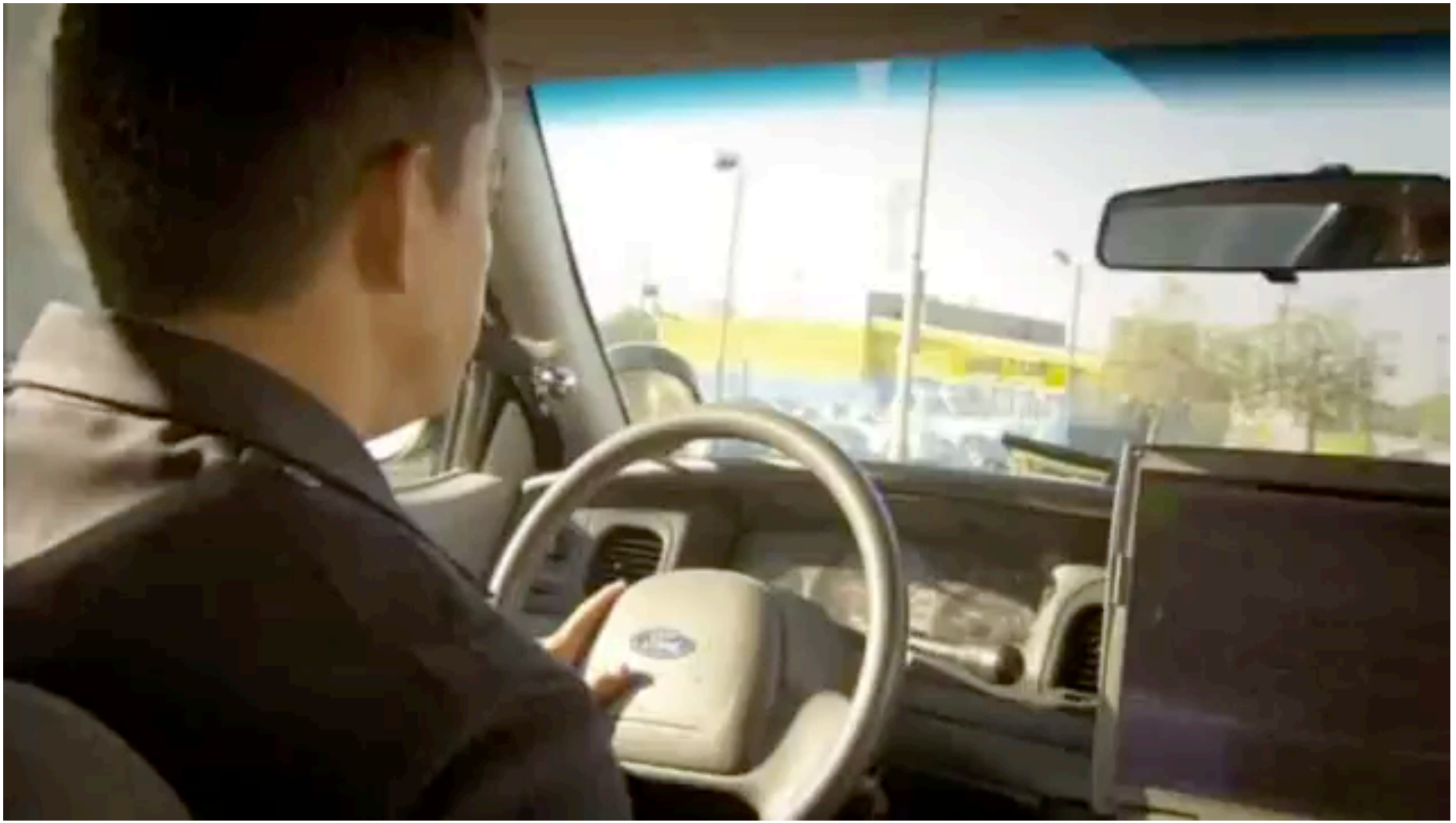
Data Science

Inspired by Daniel Keim, "Visual Analytics: Definition, Process, and Challenges"

Outline

- What?
- Why?
- Who?
- How?

The Age of Big Data

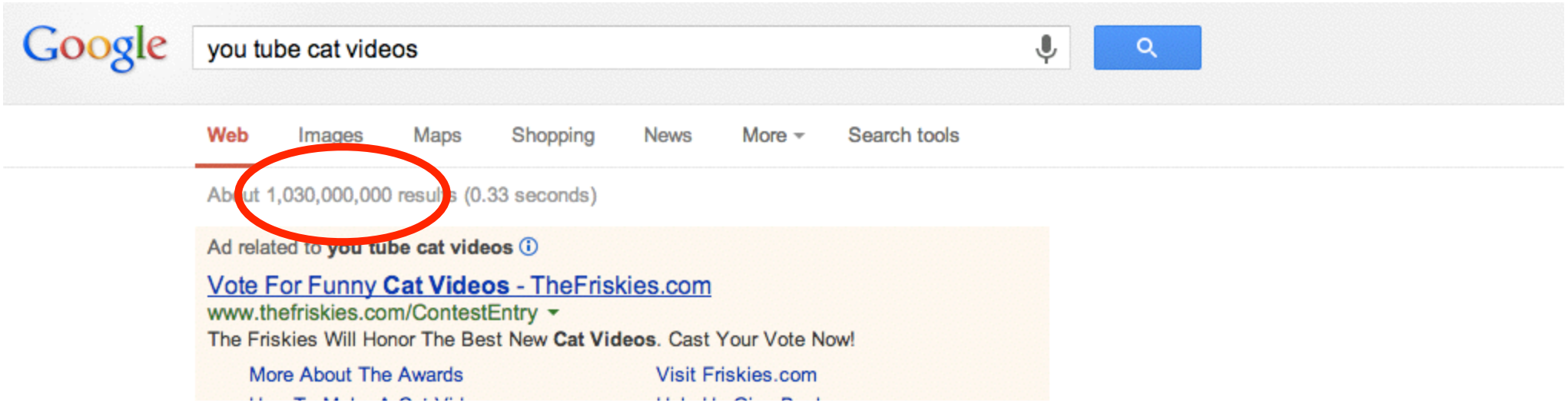


BBC, 2013

Big Data

"Between the dawn of civilization and 2003, we only created five exabytes of information; now we're creating that amount every two days."

Eric Schmidt, Google (and others)



In one second on the Internet there are...

http://onesecond.designly.com/

travers808, Visual.ly

THE BIG V'S OF BIG DATA

Turning Information Overload Into Big Sales

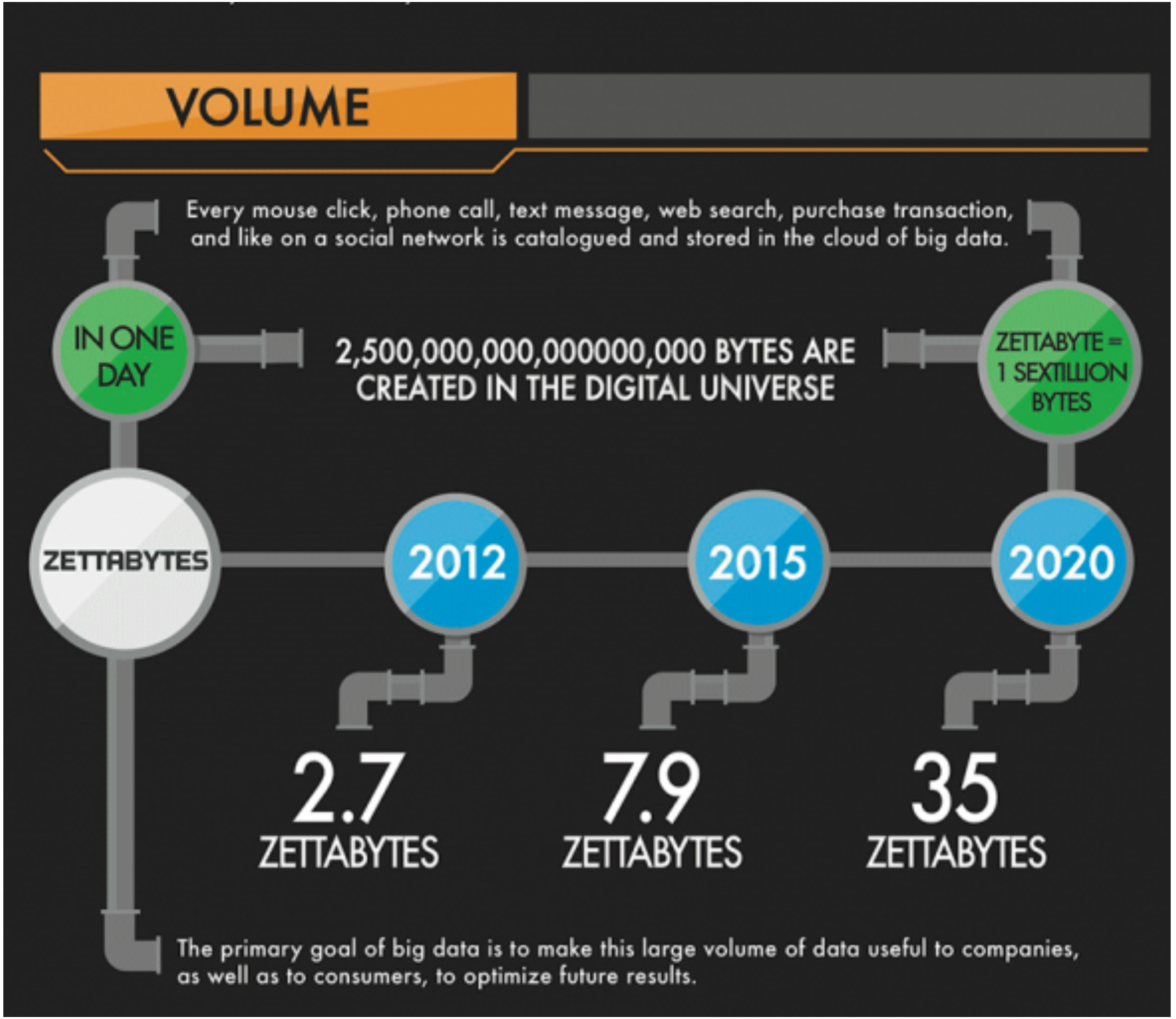
In the emerging market of Big Data, three "V" words have often been used to describe the issues at hand with information overload in our digital world.

THE EXISTING V'S

Big data has brought both great opportunity and change to the technological industry. Data scientists traditionally look at the existing V's, the ones that have classically been utilized to understand key variables of any data set.

VOLUME

Every mouse click, phone call, text message, web search, purchase transaction, and like on a social network is catalogued and stored in the cloud of big data.


VOLUME
Every mouse click, phone call, text message, web search, purchase transaction, and like on a social network is catalogued and stored in the cloud of big data.
IN ONE DAY
2,500,000,000,000000,000 BYTES ARE CREATED IN THE DIGITAL UNIVERSE
ZETTABYTE = 1 SEXTILLION BYTES
ZETTABYTES
2012
2015
2020
2.7 ZETTABYTES
7.9 ZETTABYTES
35 ZETTABYTES
The primary goal of big data is to make this large volume of data useful to companies, as well as to consumers, to optimize future results.

ZETTABYTES ZETTABYTES ZETTABYTES

The primary goal of big data is to make this large volume of data useful to companies, as well as to consumers, to optimize future results.

VARIETY

In today's multi-faceted Internet culture, the great volume of data is also extremely varied in its form. So many variables can be thrown at a company that the true value of information can often be lost in the sea of data.

PURCHASE TRANSACTIONS

WEBSITE TRAFFIC

REWARDS PROGRAMS

QUARTERLY BUSINESS REPORTS

TWITTER

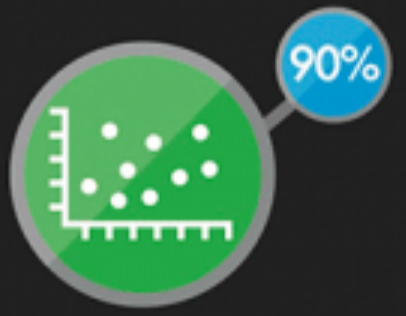
FACEBOOK

BLOG CONTENT

VELOCITY

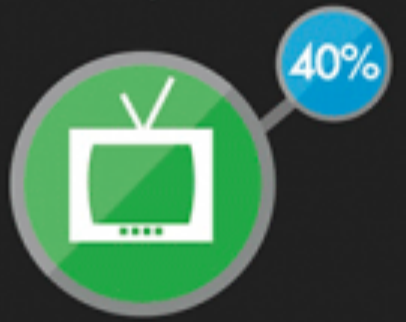
Information is being created at a faster pace than ever before. The varied channels of big data are each increasing their output of content, daily.

USERS GENERATE 2.7 BILLION LIKES ON FACEBOOK PER DAY



of the data in the world today has been created in the last two years alone

NEW TWEETS ARE CREATED BY ACTIVE USERS EACH DAY



40% of tweets are related to television and are beginning to be implemented in TV ratings

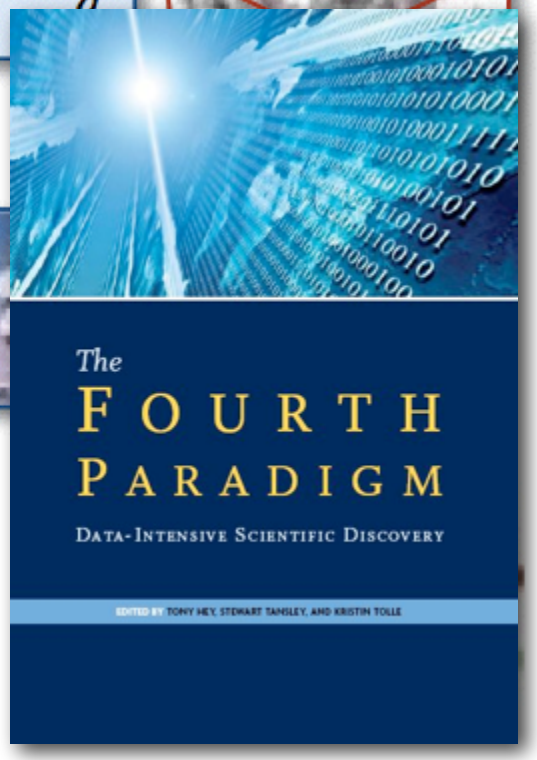
OF VIDEO IS UPLOADED TO YOUTUBE EVERY MINUTE

In 7 years, 15x the amount of data that exists today will be created every single year

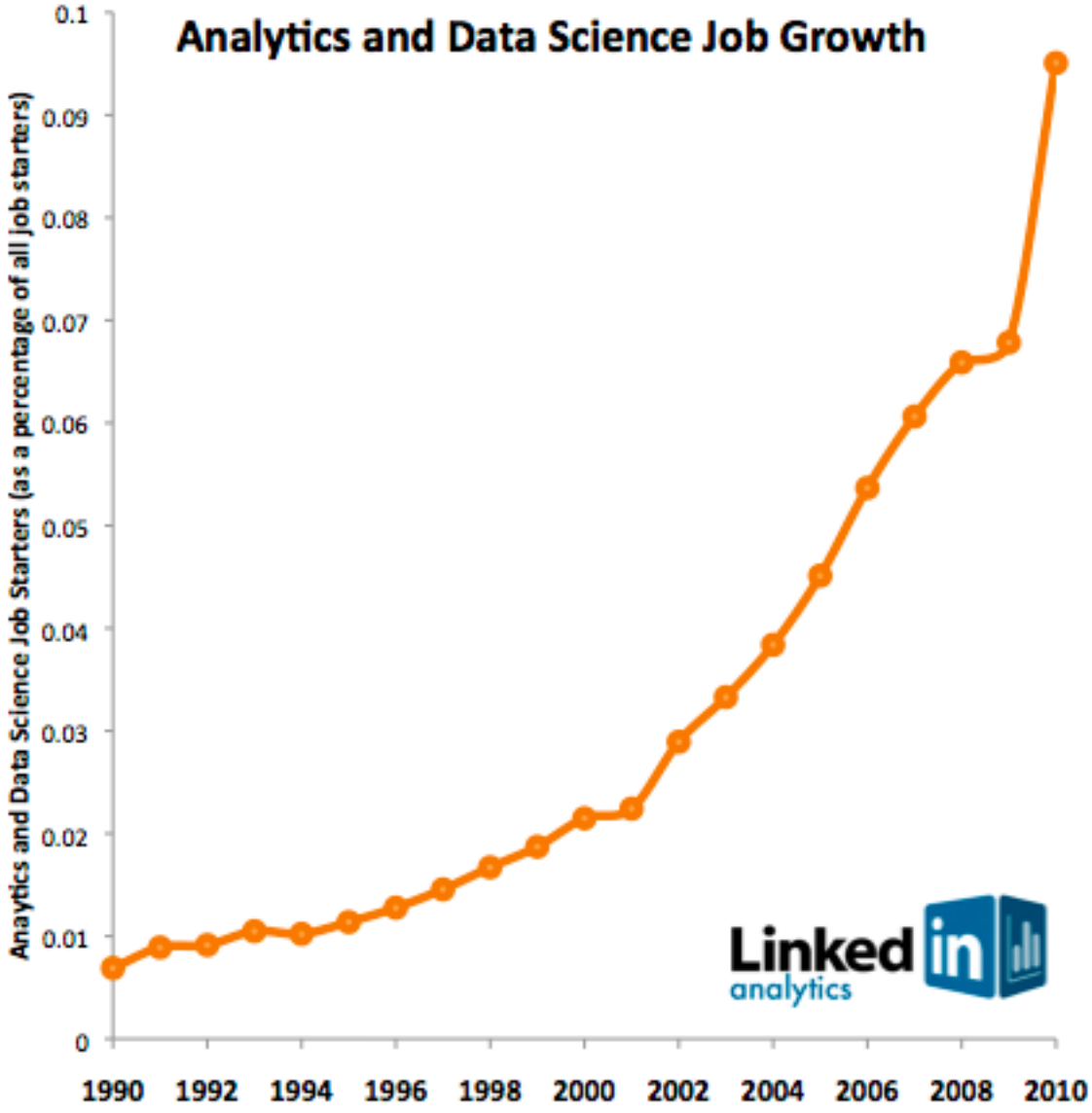
Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - K\frac{c^2}{a^2}$$



Jim Gray, Microsoft


Analytics and Data Science Job Growth
Anaytics and Data Science Job Starters (as a percentage of all job starters)
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
Linked in
analytics

"By 2018, the US could face a shortage of up to 190,000 workers with analytical skills"

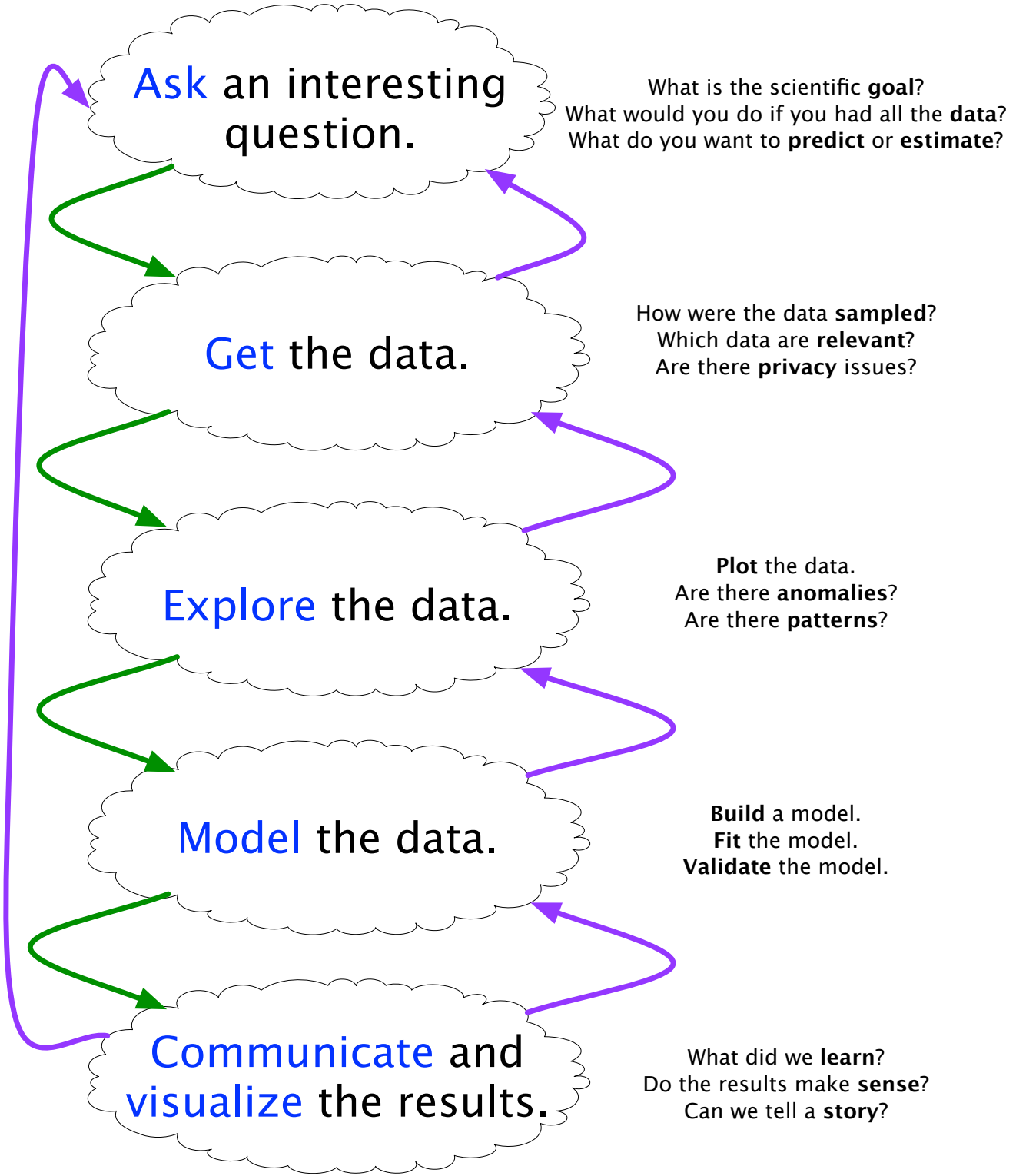
McKinsey Global Institute

"The sexy job in the next 10 years will be ~~statisticians~~." *Data Scientists?*

Hal Varian, Prof. Emeritus UC Berkeley
Chief Economist, Google

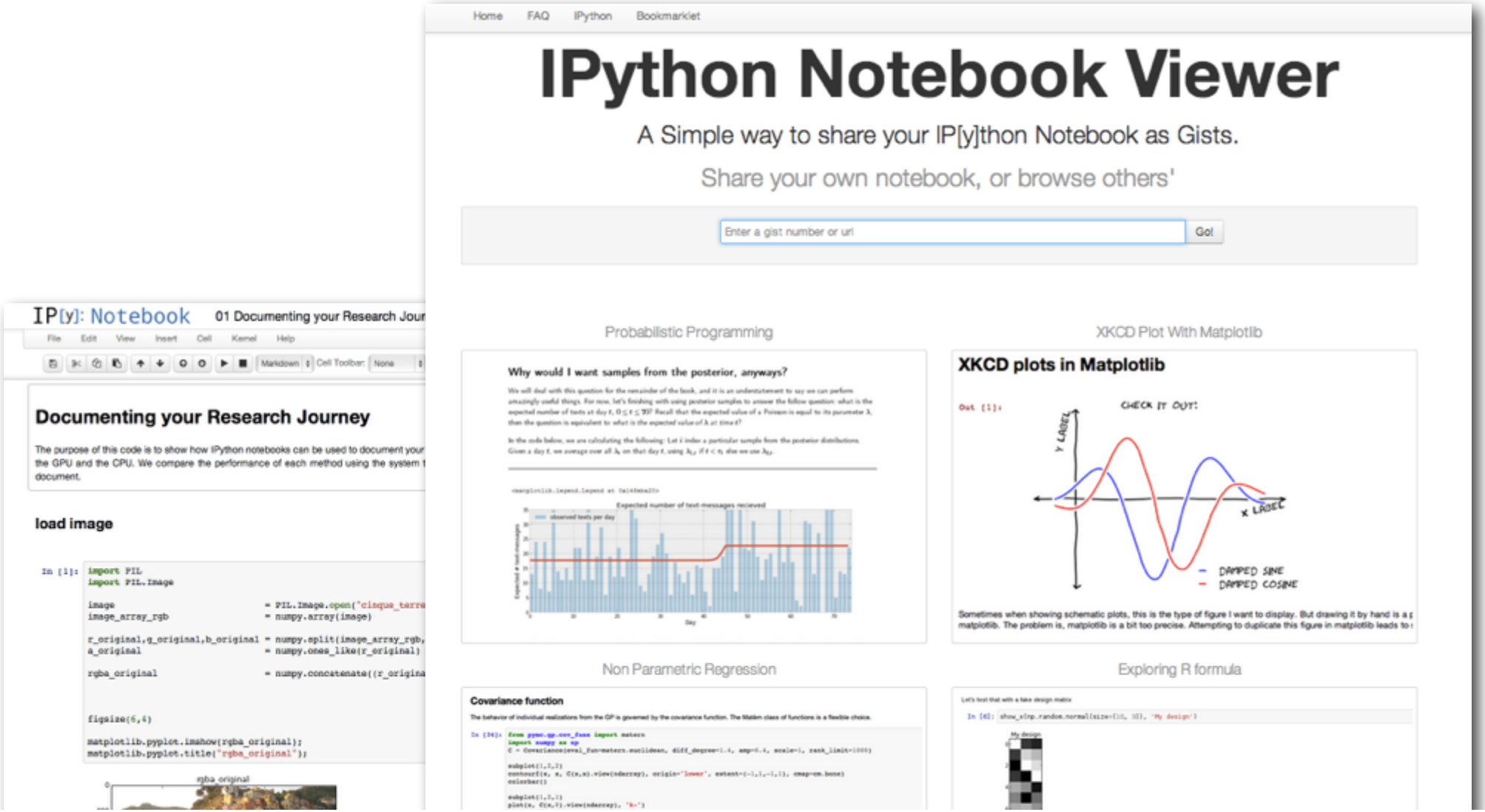
Hal Varian Explains...

The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and **ubiquitous data**." – Hal Varian


Ask an interesting question.
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?
Get the data.
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?
Explore the data.
Plot the data.
Are there **anomalies**?
Are there **patterns**?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

IPython Notebooks

http://nbviewer.ipython.org/



Outline

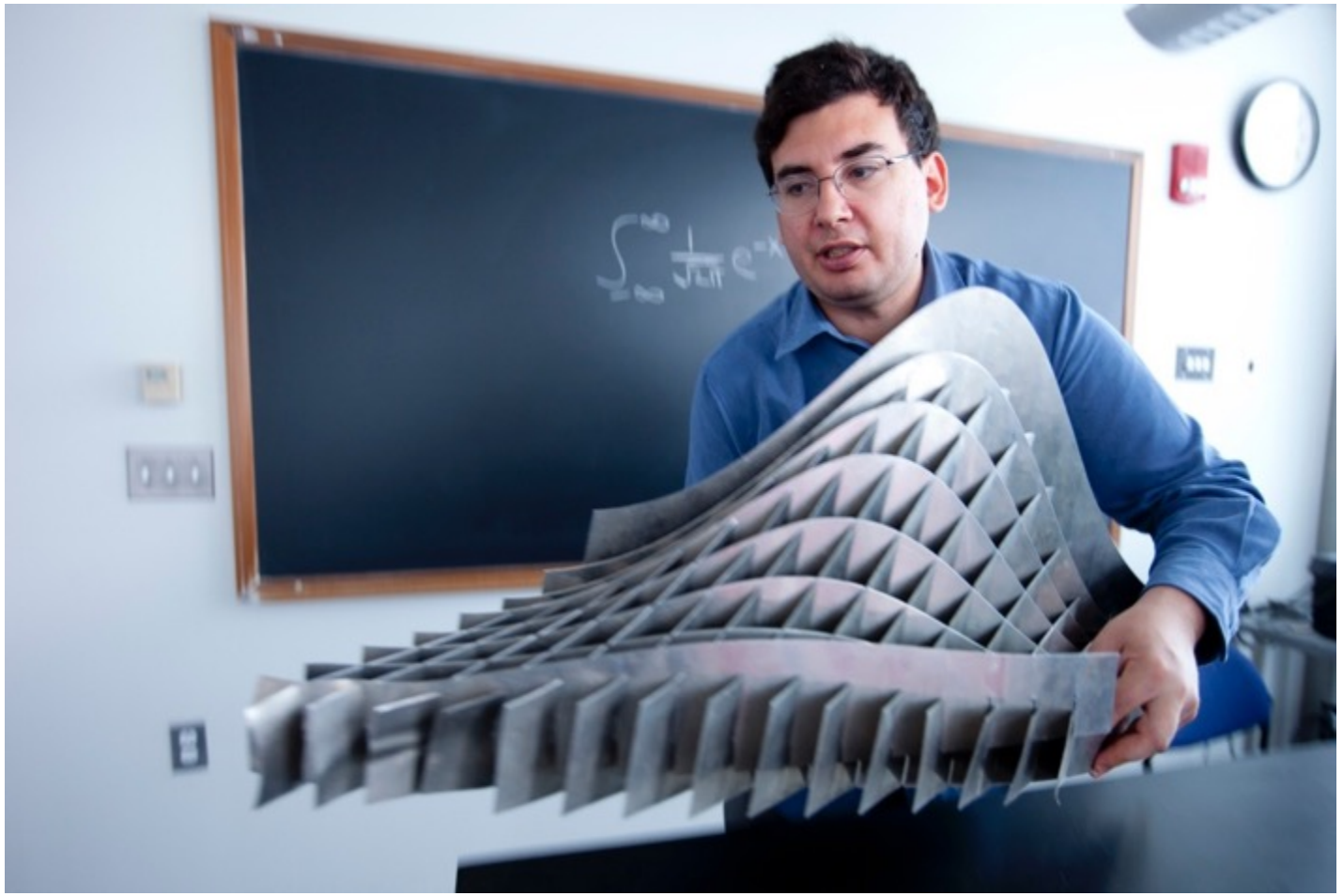
- What?
- Why?
- Who?
- How?

Hanspeter Pfister

An Wang Professor of Computer Science, SEAS
Director, Institute for Applied Computational Science
pfister@seas.harvard.edu / @hpfister

Joe Blitzstein

Professor of the Practice in Statistics,
Co-Director of Undergraduate Studies in Statistics
blitz@fas.harvard.edu, twitter @stat110, SC 714



Verena Kaynig-Fittkau

Lecturer and research scientist at IACS
vkaynig@seas.harvard.edu, NW B164

Rahul Dave

Head TF and Lecturer at IACS
rahuldave@gmail.com, NW B164

CS 109 Staff

Andrew Reece
Antonio Coppola
Austen Novis
Brian Feeny
Dana Katzenelson
Giri Gopalan
Irma Nomani
Jacob Dorabialski
Joseph Song
Kathy Li
Lawrence Kim
Leandra King
Luis Campos
Marcus Way
Michael Ma
Michael Packer
Nelson Santos
Richard Kim
Rick Wei-Jong Lee
Sail Wu
Stephen Klosterman
Xintao Qiu
Yingzhuo (Diana) Zhang
Yuhao Zhu

About You

Outline

- What?
- Why?
- Who?
- How?

CS109 Key Facets

- *data munging/scraping/sampling/cleaning* in order to get an informative, manageable data set;
- *data storage and management* in order to be able to access data quickly and reliably during subsequent analysis;
- *exploratory data analysis* to generate hypotheses and intuition about the data;
- *prediction* based on statistical tools such as regression, classification, and clustering; and
- *communication* of results through visualization, stories, and interpretable summaries.

Act I: Predictions

- Data Collection, "Munging", and Storage
- Exploratory Data Analysis (EDA)
- Classification & Regression
- Cross Validation
- Dimensionality Reduction
- Effective Communication & Writing

Act II: Recommendations

- Support Vector Machines
- Decision Trees & Random Forests
- Bagging & Boosting
- Machine Learning Best Practices
- MapReduce, Amazon's EC2, and Spark

Act III: Clustering & Text

- Bayesian Thinking & Naive Bayes
- Text Analysis: LDA & Topic Modeling
- Clustering
- Effective Presentations
- Deep Learning
- Guest Lecture: Experimental Design

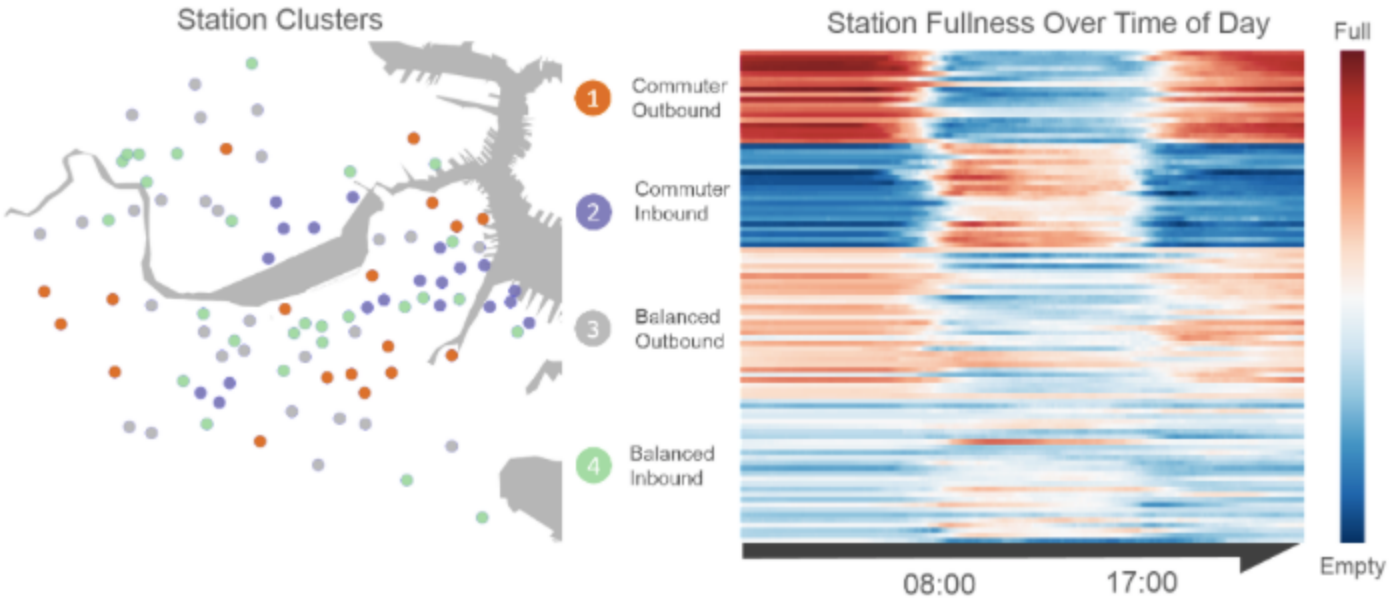
cs109.org

CS109 Data Science

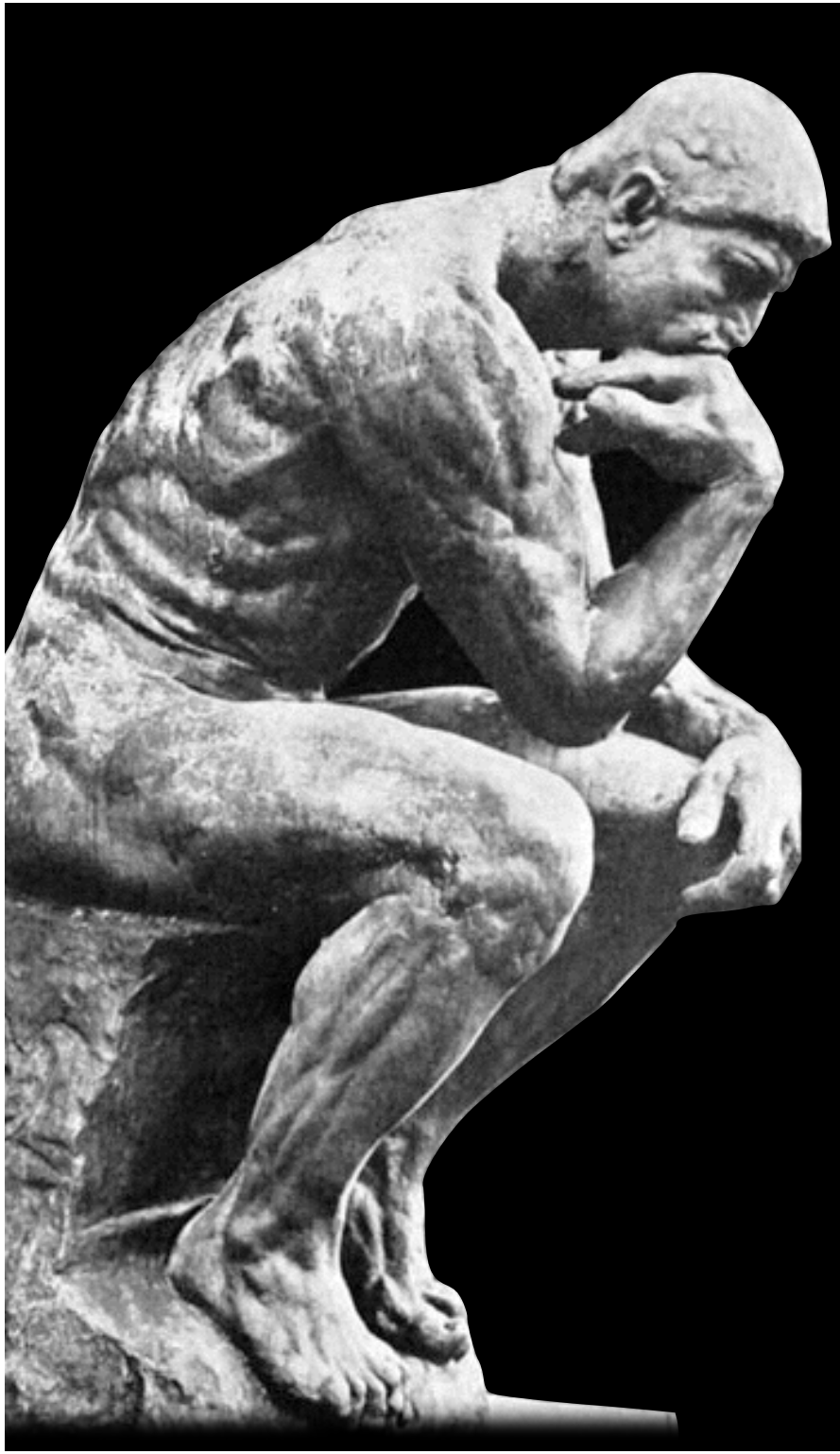
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Home Piazza Syllabus Schedule GitHub Projects Resources

CS109 Data Science



Predicting Hubway Stations Status by Lauren Alexander, Gabriel Goulet-Langlois, Joshua Wolff



Concepts...

Lectures

...and Skills

Sections

Sections

- Introduce tools & skills; available as lab notebooks and videos
- Mandatory, except for DCE students
- First (group) section this Friday!
  - 10am-12pm in MD G115
- Regular sections first week as office hours to get help with Python, Git, and HW0

Section Schedule (TBD)

	Monday	Tuesday	Wednesday	Thursday	Friday
9:00 AM			Rahul, NW-B150		
10:00 AM	Leandra	Ima	Steve		Luis
11:00 AM	NW-B150	NW-B150	NW-B150		NW-B150
12:00 PM					
1:00 PM	Diana				Michael Ma
2:00 PM	NW-B150	Lecture	Lawrence	Lecture	NW-B150
3:00 PM	Joseph	NW-B103	NW-B150	NW-B103	Michael Packer
4:00 PM	NW-B150		Antonio		NW-B150
5:00 PM		Sail, Nelson	NW-B150		
6:00 PM	Austen, Dana	NW-B150 and B166	Richard		
7:00 PM	NW-B150 and B166	Kathi	NW-B150		
8:00 PM		NW-B150			

Homework

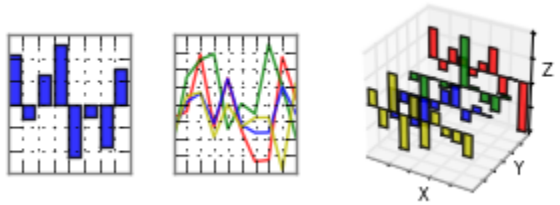
- Real-World focus
  - Scrape and wrangle messy data
  - Apply sophisticated statistical analysis
  - Visualize and communicate results
- Election data, music charts, recommendations, etc.

Programming

IP[y]: IPython
Interactive Computing

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

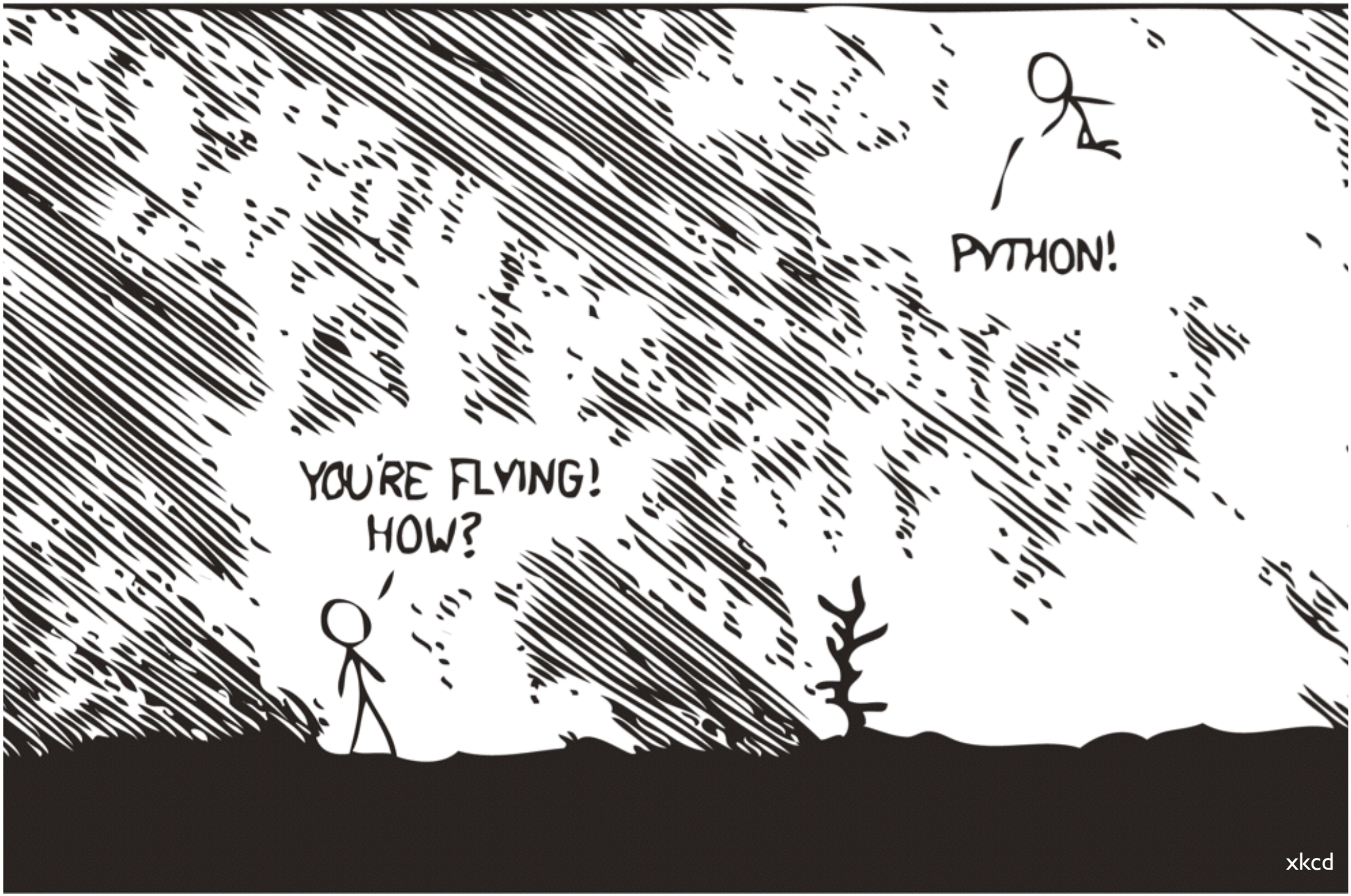


scikits learn
machine learning in Python

NumPy

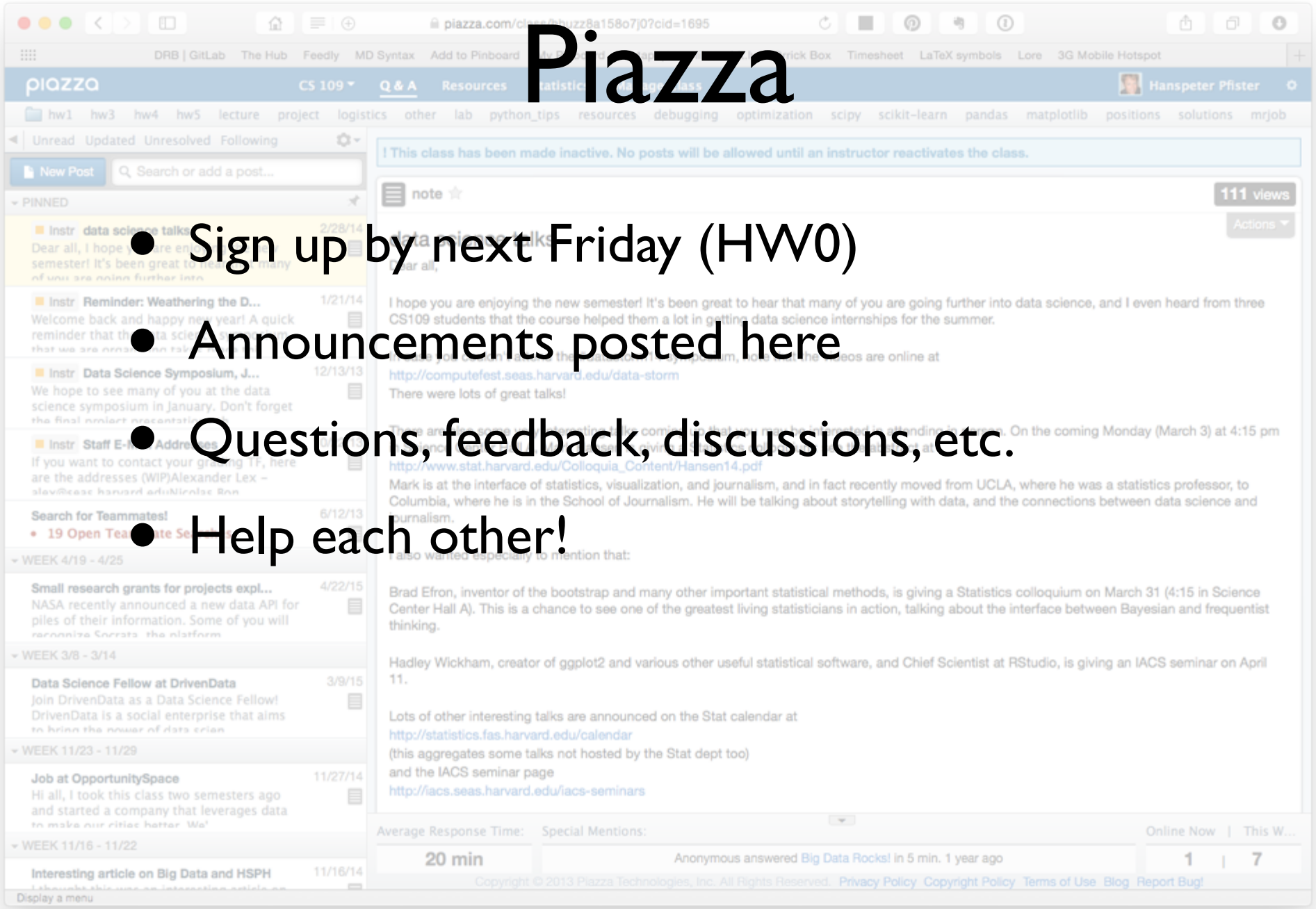
SciPy.org Sponsored By ENTHOUGHT

matplotlib


PYTHON!
YOU'RE FLYING!
HOW?
xkcd

Piazza

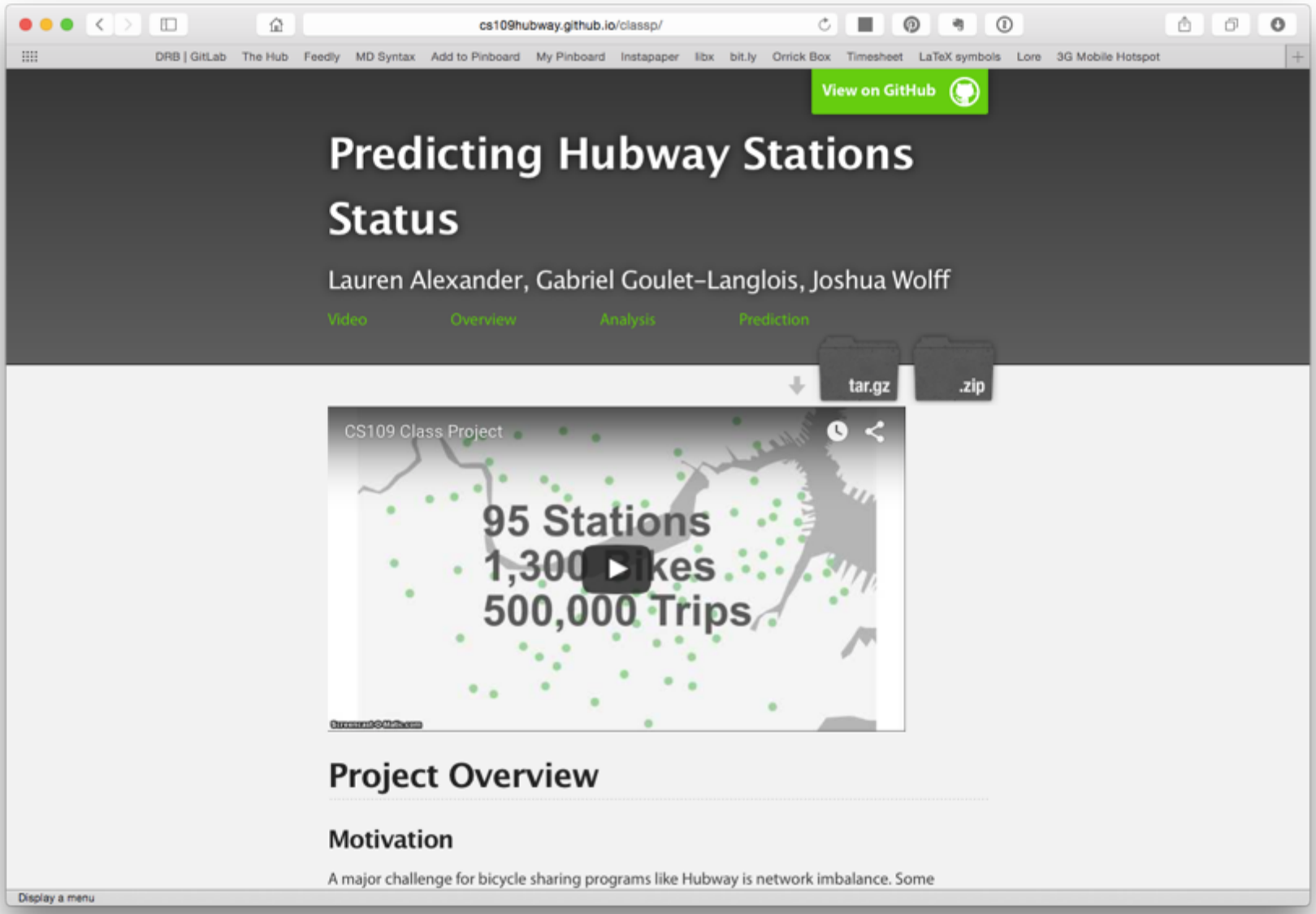
- Sign up by next Friday (HW0)
- Announcements posted here
- Questions, feedback, discussions, etc.
- Help each other!



Grades

- No exams!
- 50% Homework
- 40% Projects (3-4 person teams)
- 10% Participation (Piazza & Sections)
- 10 point scale, holistic grading

Projects


cs109hubway.github.io/classp/
DRB | GitLab
The Hub
Feedly
MD Syntax
Add to Pinboard
My Pinboard
Instapaper
libx
bit.ly
Orrick Box
Timesheet
LaTeX symbols
Lore
3G Mobile Hotspot
View on GitHub
Predicting Hubway Stations Status
Lauren Alexander, Gabriel Goulet-Langlois, Joshua Wolff
Video
Overview
Analysis
Prediction
tar.gz
.zip
CS109 Class Project
95 Stations
1,300 Bikes
500,000 Trips
Project Overview
Motivation
A major challenge for bicycle sharing programs like Hubway is network imbalance. Some
Display a menu

Policies

- HWs due on Thursdays, 11:59 pm EST
- 6 late days for HW (no questions asked)
- Cannot submit HW later than 2 days
- Regrading requests within 7 days in writing
  - Grade may improve or go down

Collaboration Policy

- Work you turn in must be your own
- Projects are a 3-4 person team effort
  - With project group peer assessment
- Acknowledge all help and code you used
- Harvard Honor Code

Is this course for me ???

Prerequisites

Programming experience

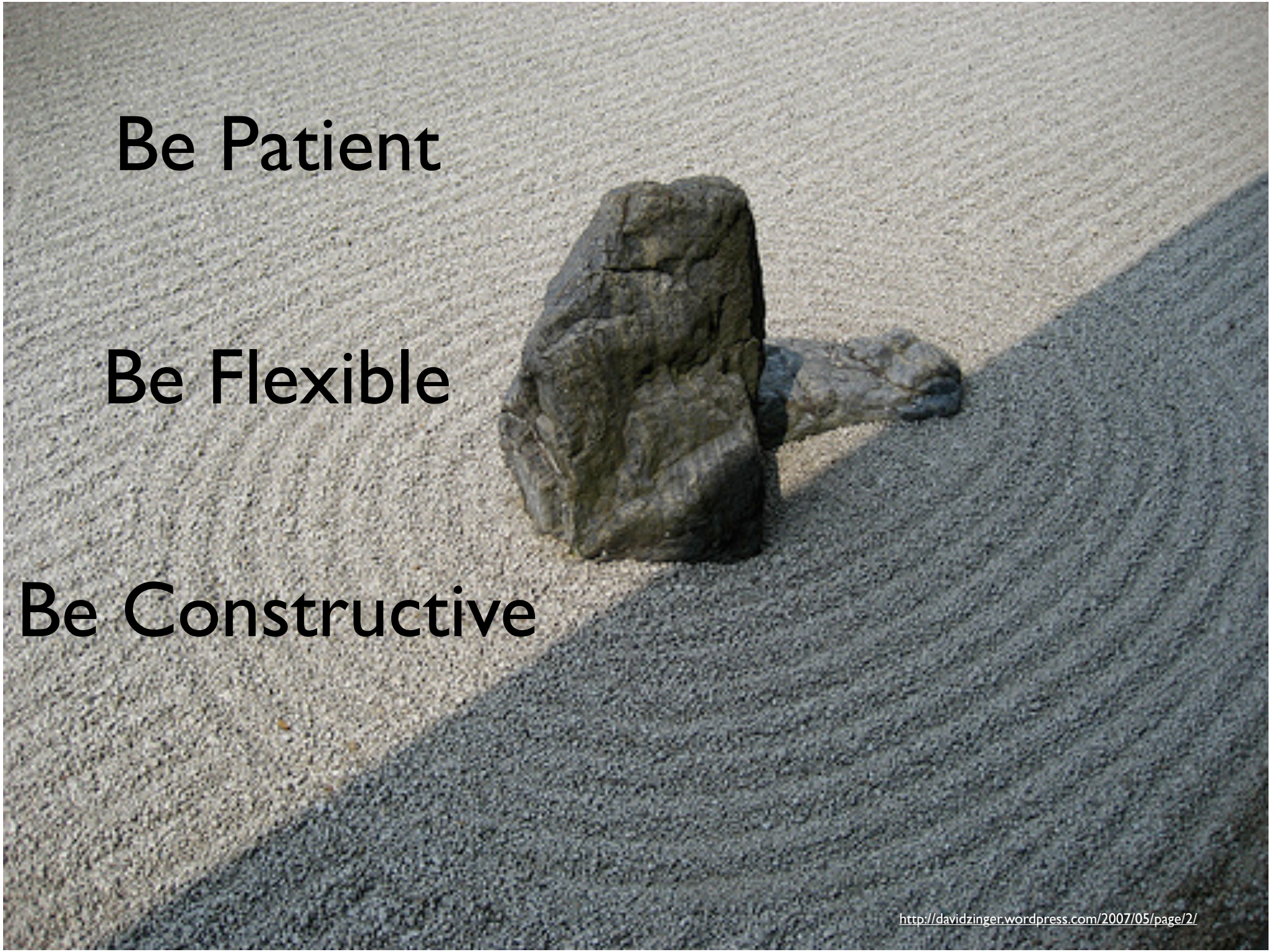
- CS50 and/or C, C++, Java, Python, etc.

Basic statistical knowledge

- STAT100, ideally STAT110

Willingness to learn new software & tools

- This can be time consuming
- You will need to read online documentation


Be Patient
Be Flexible
Be Constructive
http://davidzinger.wordpress.com/2007/05/page/2/

Next Steps

- HW 0, mandatory, needs to be submitted!
  - Good test of your basic skills
  - Complete the survey by tomorrow! Needed to be able to submit HW 0
  - Installation of several Python frameworks
  - Not graded, do it as soon as possible
- Read syllabus carefully

Important Links

- Create a github account at http://github.com
- Then fill in our survey at http://goo.gl/forms/bJwajS8zO8
- HW 0 document at https://github.com/cs109/2015lab1/blob/master/hw0.ipynb
- Week 1 notebooks at https://github.com/cs109/2015lab1
- HW repositories will be created for you on github. See HW 0 for details.