STAT121 / AC209 / E-109

CS109 Data Science

Hanspeter Pfister
pfister@seas.harvard.edu

Joe Blitzstein
blitzstein@stat.harvard.edu

Outline

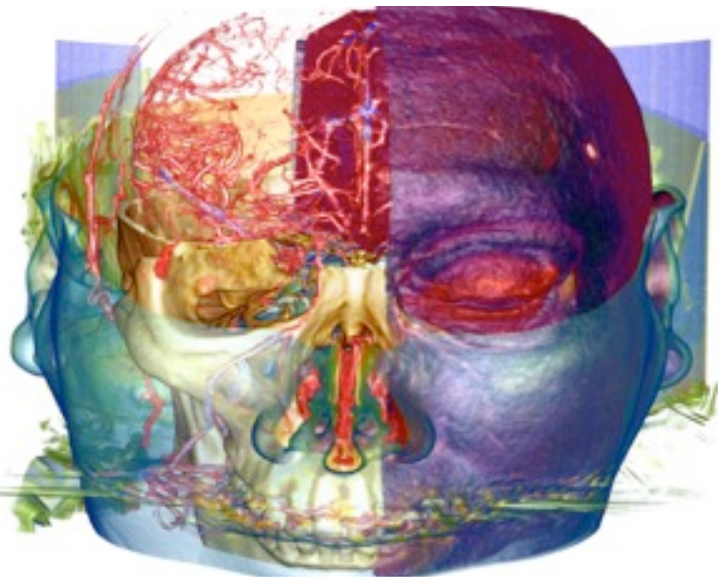
- What?
- Why?
- Who?
- How?

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

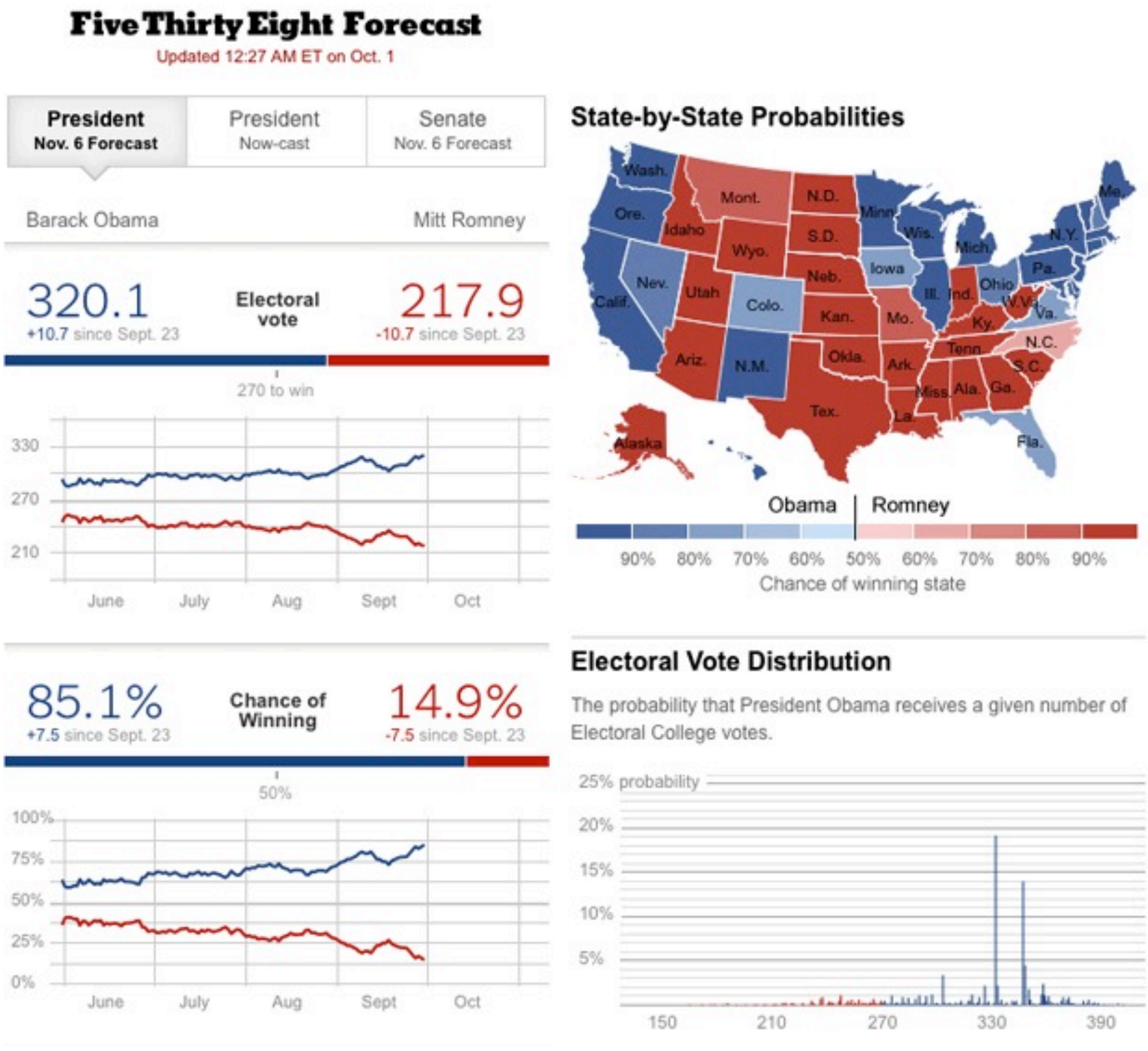
"Data Scientist = statistician + programmer + coach + storyteller + artist"

- 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

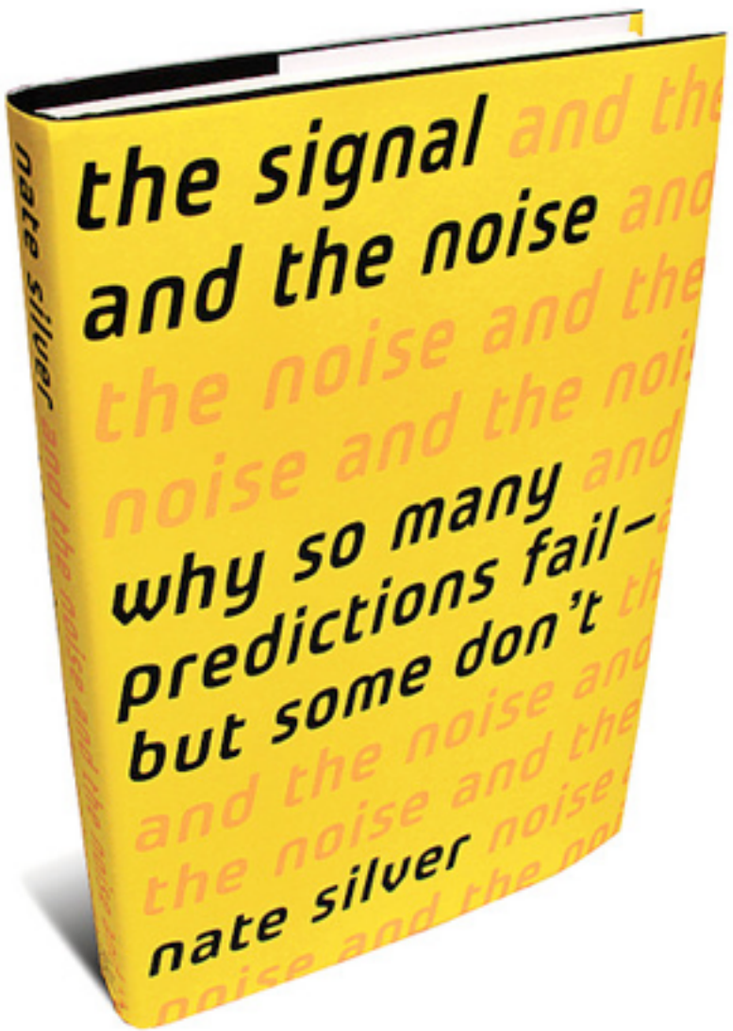
Wednesday, November 7th, 2012

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

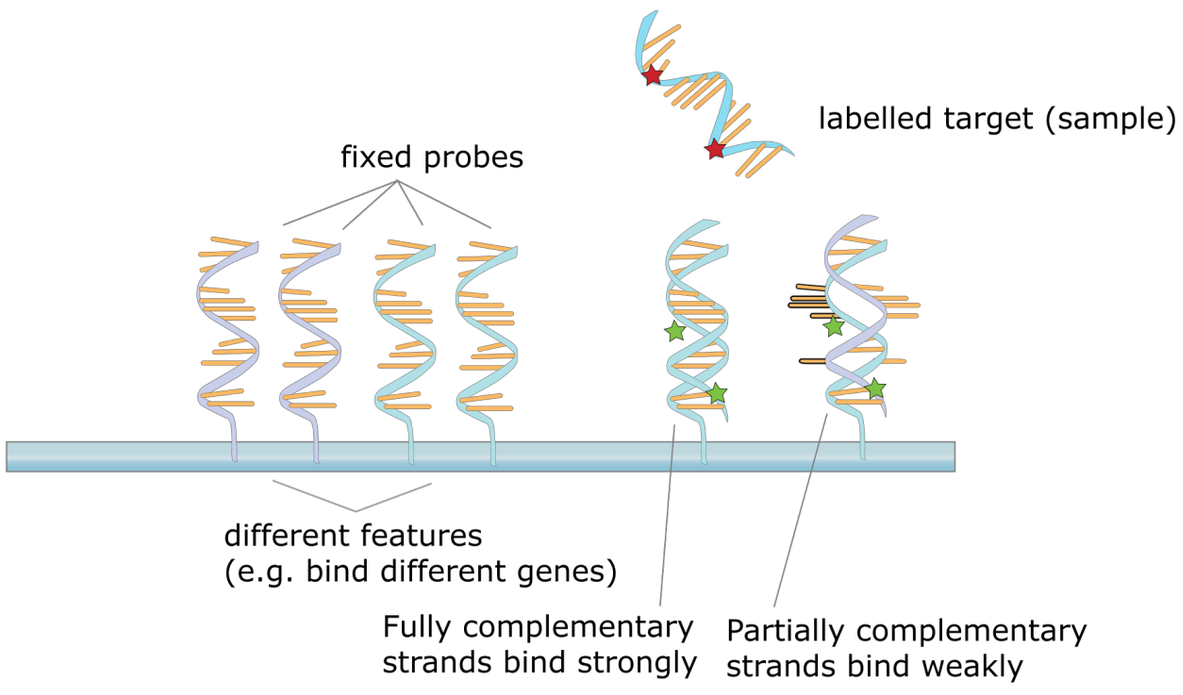
Some Key Principles

- *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?)

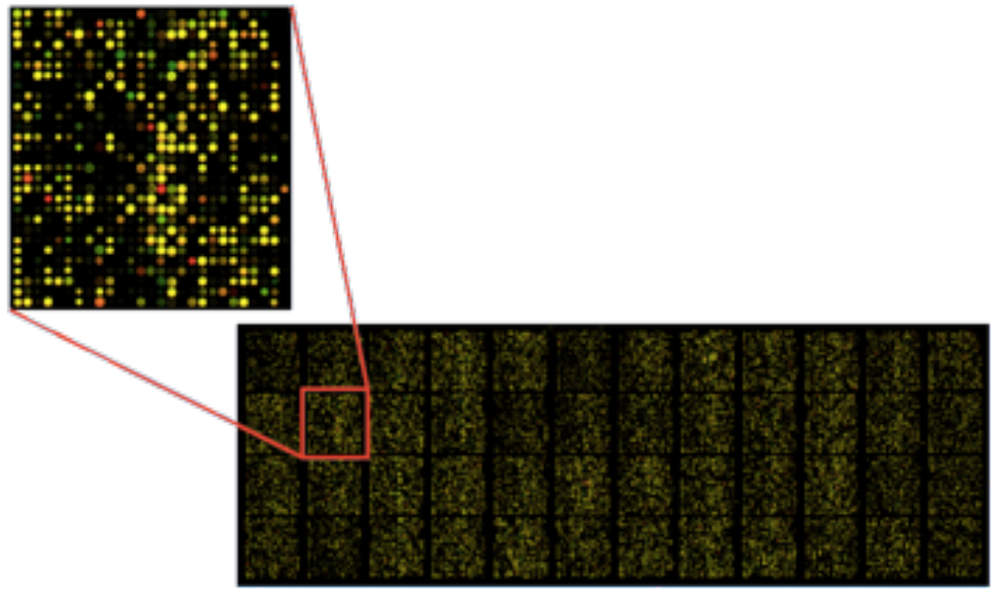
Human Genome

1953 Crick (right) and Watson (left) reveal the structure of DNA, claiming it holds the key to the identity of genes

Microarrays

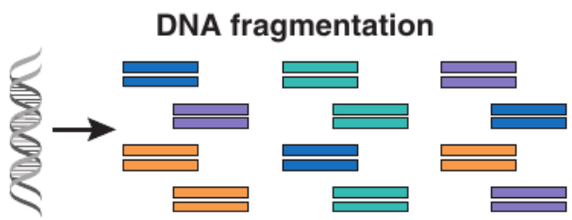


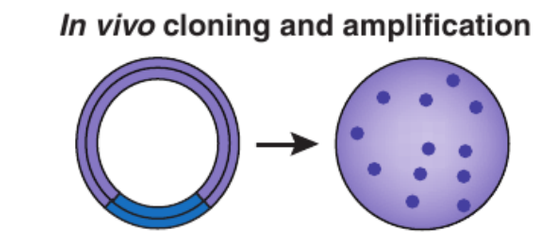
Affimetrix Chip

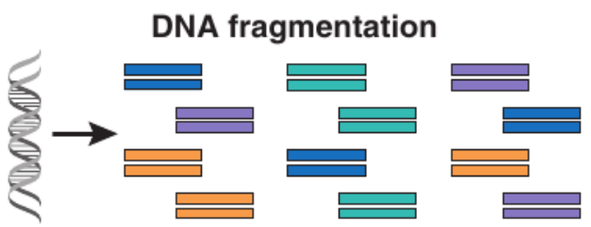


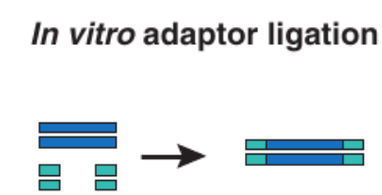
[wikipedia]

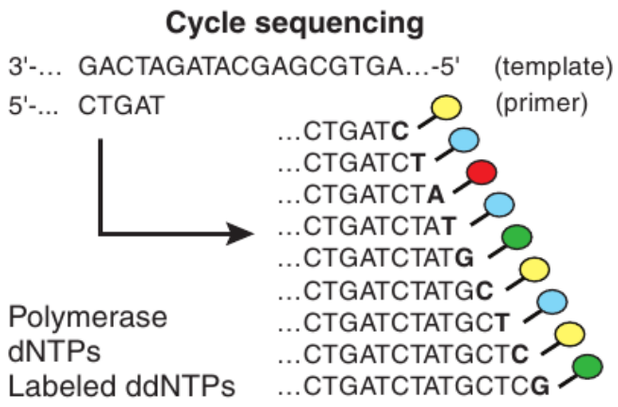
Sequencing

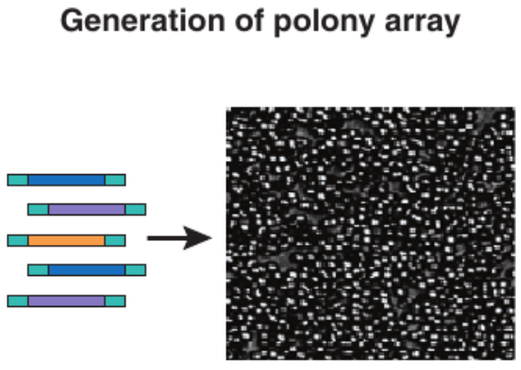

DNA fragmentation

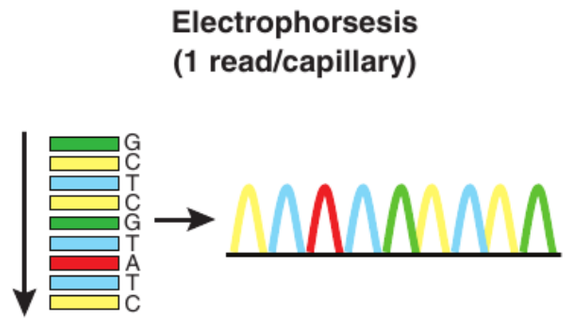

In vivo cloning and amplification

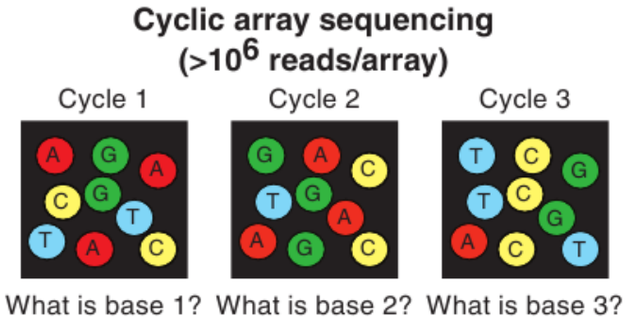

DNA fragmentation


In vitro adaptor ligation

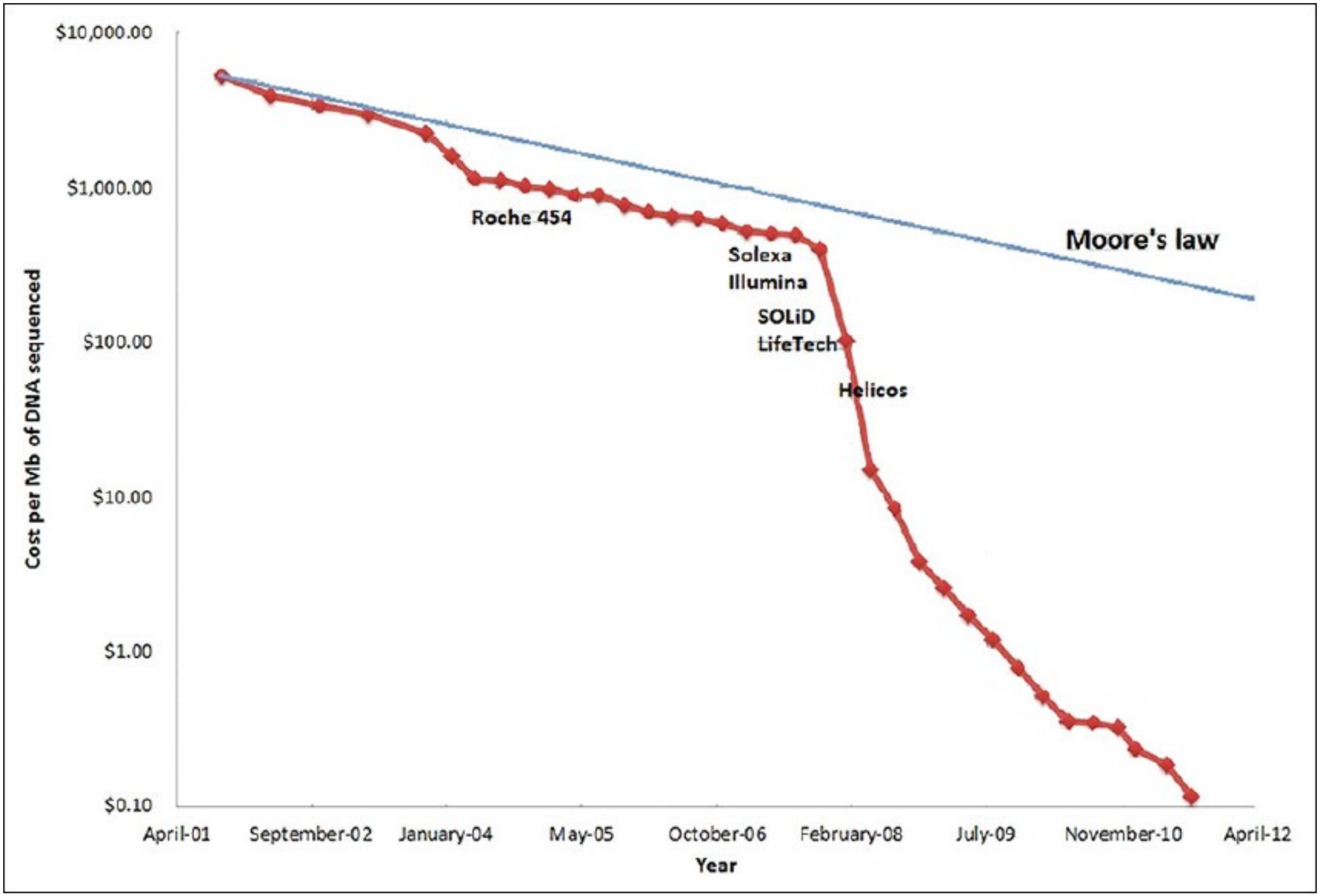

Cycle sequencing
3'-… GACTAGATACGAGCGTGA…-5' (template)
5'-... CTGAT (primer)
…CTGATC
…CTGATCT
…CTGATCTA
…CTGATCTAT
…CTGATCTATG
…CTGATCTATGC
…CTGATCTATGCT
…CTGATCTATGCTC
…CTGATCTATGCTCG
Polymerase
dNTPs
Labeled ddNTPs


Generation of polony array

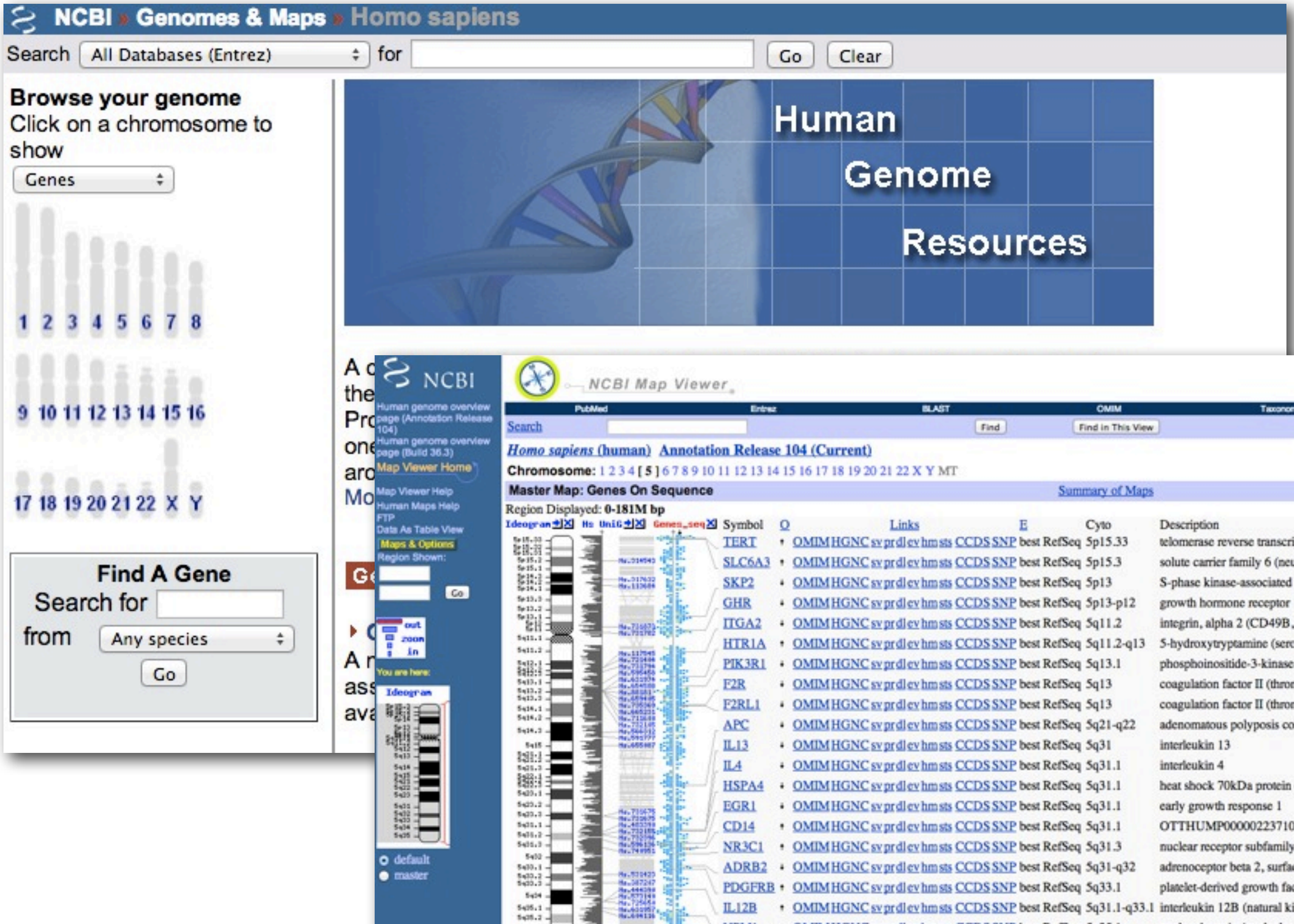

Electrophorsesis
(1 read/capillary)
G
C
T
C
G
T
A
T
C


Cyclic array sequencing
(>10^6 reads/array)
Cycle 1
Cycle 2
Cycle 3
What is base 1?
What is base 2?
What is base 3?

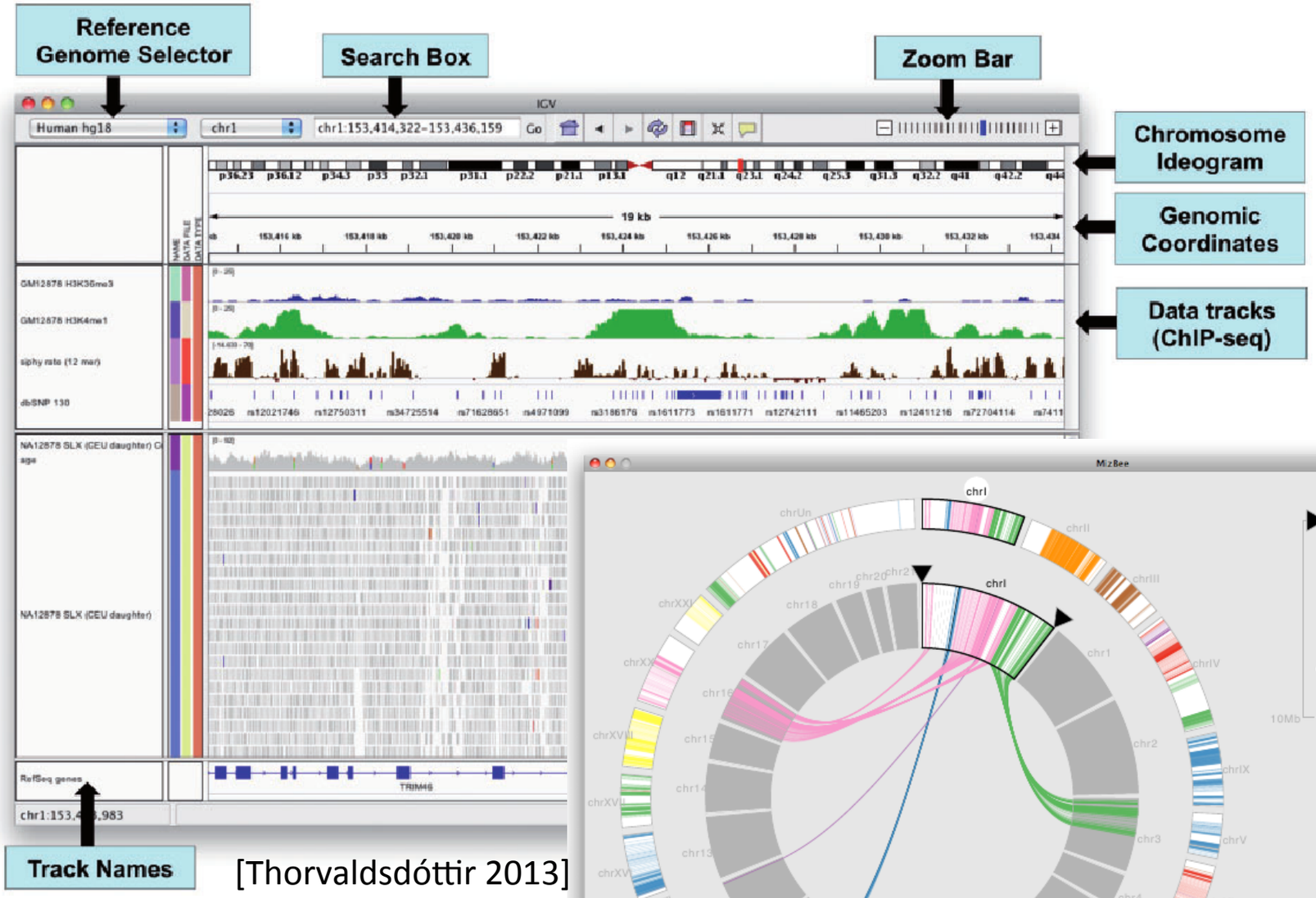
Sequencing Cost



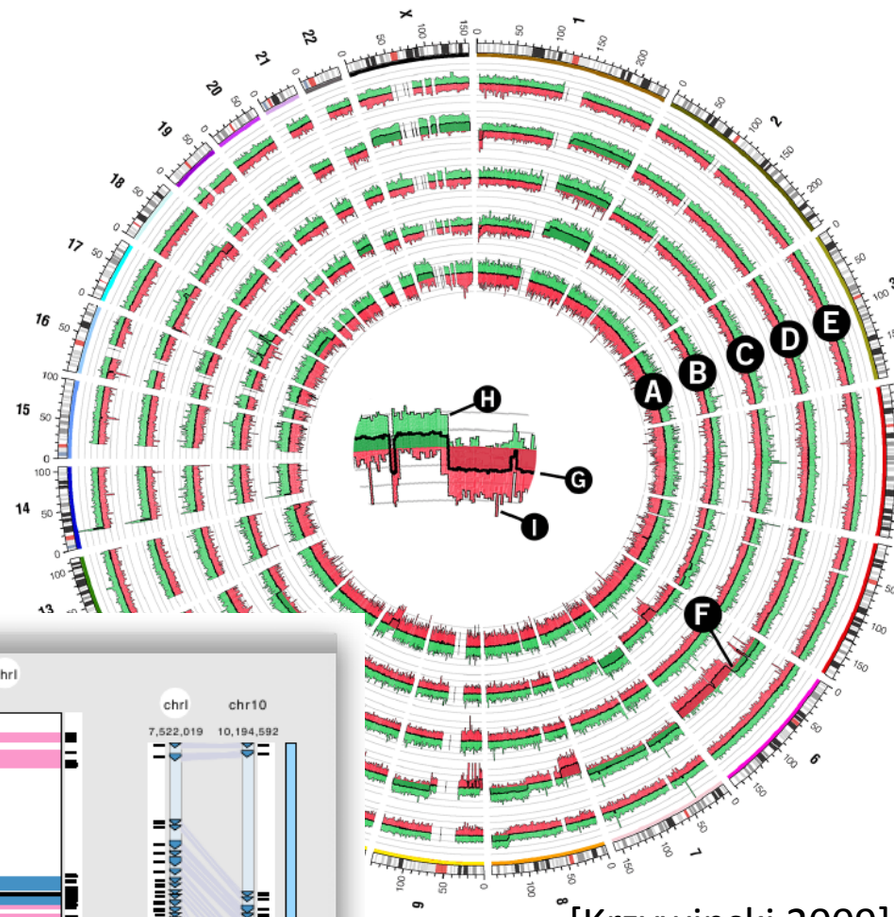
Genome Data


NCBI » Genomes & Maps » Homo sapiens
Search All Databases (Entrez) for Go Clear
Browse your genome
Click on a chromosome to show
Genes
1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16
17 18 19 20 21 22 X Y
Find A Gene
Search for
from Any species
Go
Human Genome Resources
NCBI
Human genome overview page (Annotation Release 104)
Human genome overview page (Build 36.3)
Map Viewer Home
Map Viewer Help
Human Maps Help
FTP
Data As Table View
Maps & Options
Region Shown:
Go
out
zoom
in
You are here:
Ideogram
default
master
NCBI Map Viewer
PubMed Entrez BLAST OMIM Taxonomy Structure
Search Find Find in This View Advanced Search
Homo sapiens (human) Annotation Release 104 (Current)
BLAST human sequences
Chromosome: 1 2 3 4 [5] 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT
Master Map: Genes On Sequence
Summary of Maps
Maps & Options
Region Displayed: 0-181M bp
Download/View Sequence/Evidence
Ideogram Hs UniG Genes_seq
Symbol Q Links E Cyto Description
TERT OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5p15.33 telomerase reverse transcriptase
SLC6A3 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5p15.3 solute carrier family 6 (neurotransmitter transporter), member 3
SKP2 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5p13 S-phase kinase-associated protein 2, E3 ubiquitin protein ligase
GHR OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5p13-p12 growth hormone receptor
ITGA2 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q11.2 integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)
HTR1A OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q11.2-q13 5-hydroxytryptamine (serotonin) receptor 1A, G protein-coupled
PIK3R1 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q13.1 phosphoinositide-3-kinase, regulatory subunit 1 (alpha)
F2R OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q13 coagulation factor II (thrombin) receptor
F2RL1 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q13 coagulation factor II (thrombin) receptor-like 1
APC OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q21-q22 adenomatous polyposis coli
IL13 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q31 interleukin 13
IL4 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q31.1 interleukin 4
HSPA4 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q31.1 heat shock 70kDa protein 4
EGR1 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q31.1 early growth response 1
CD14 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q31.1 OTTHUMP00000223710
NR3C1 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q31.3 nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
ADRB2 OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q31-q32 adrenoceptor beta 2, surface
PDGFRB OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q33.1 platelet-derived growth factor receptor, beta polypeptide
IL12B OMIM HGNC sv prdl ev hm sts CCDS SNP best RefSeq 5q31.1-q33.1 interleukin 12B (natural killer cell stimulatory factor 2, cytotoxic lymphocy

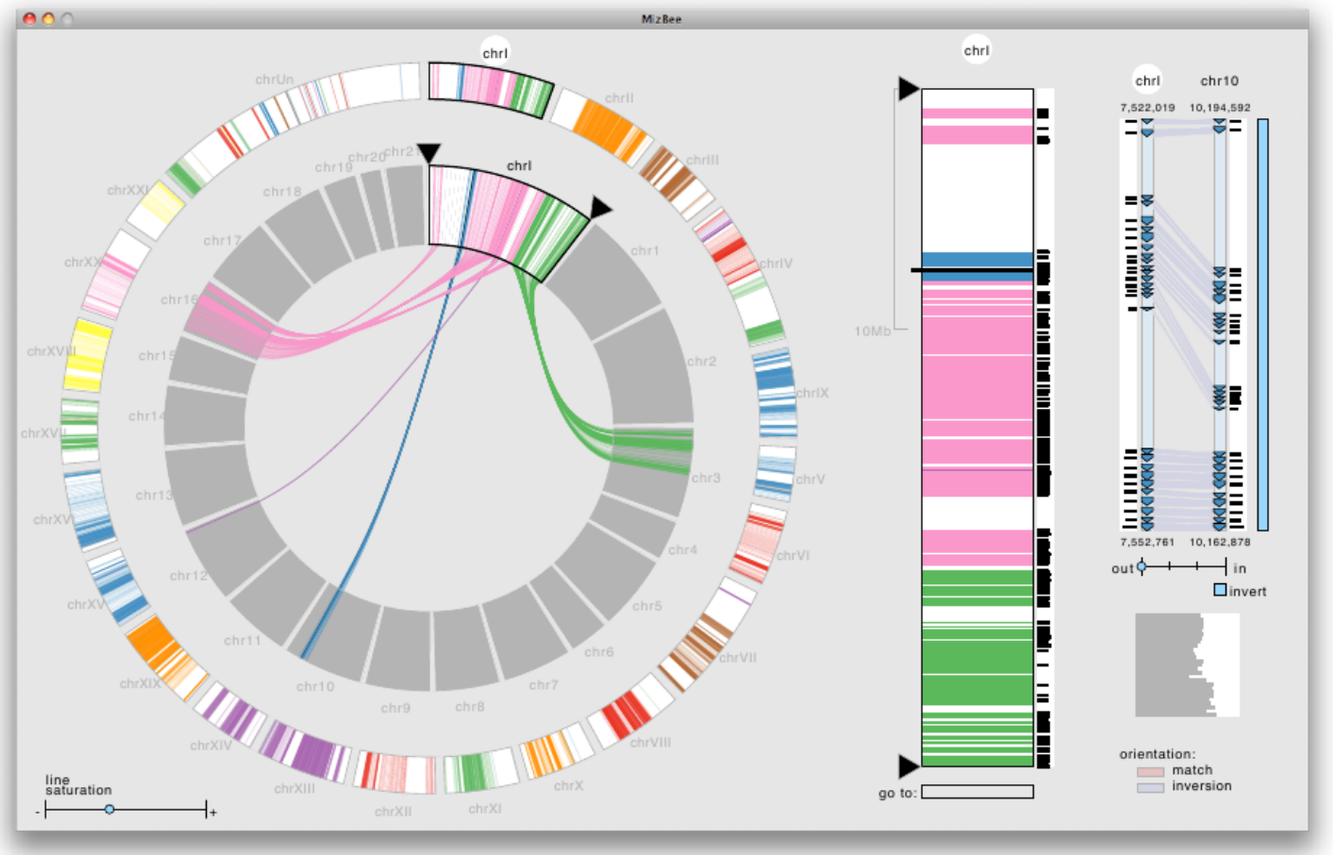
Genome Visualization



[Thorvaldsdóttir 2013]



[Krzywinski 2009]



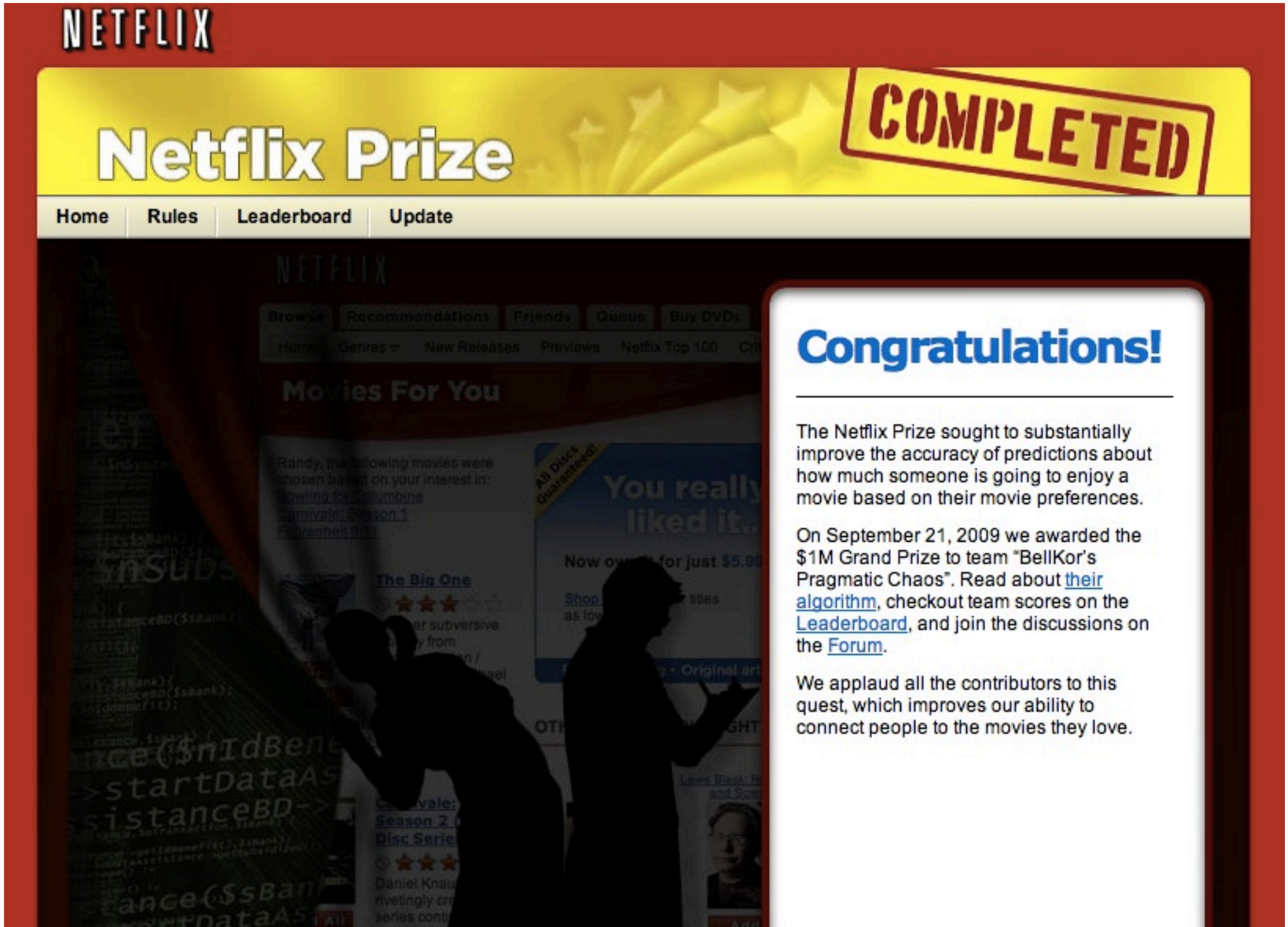
[Meyer 2009]

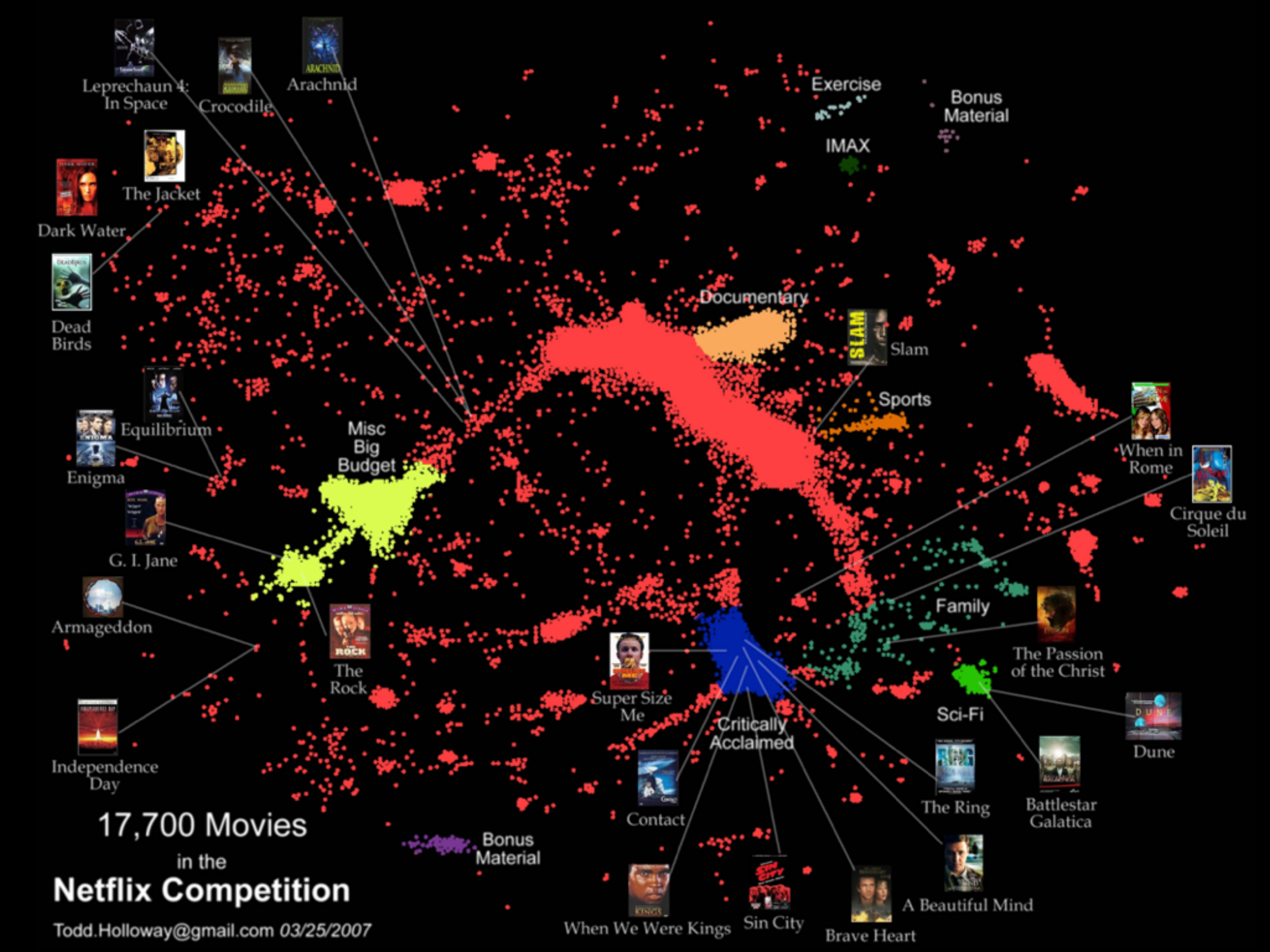
Personalized Therapy

"...10 years from now, each cancer patient is going to want to get a genomic analysis of their cancer and will expect customized therapy based on that information."

Director, The Cancer Genome Atlas (TCGA), Time Magazine, 6/13/11

Netflix Prize

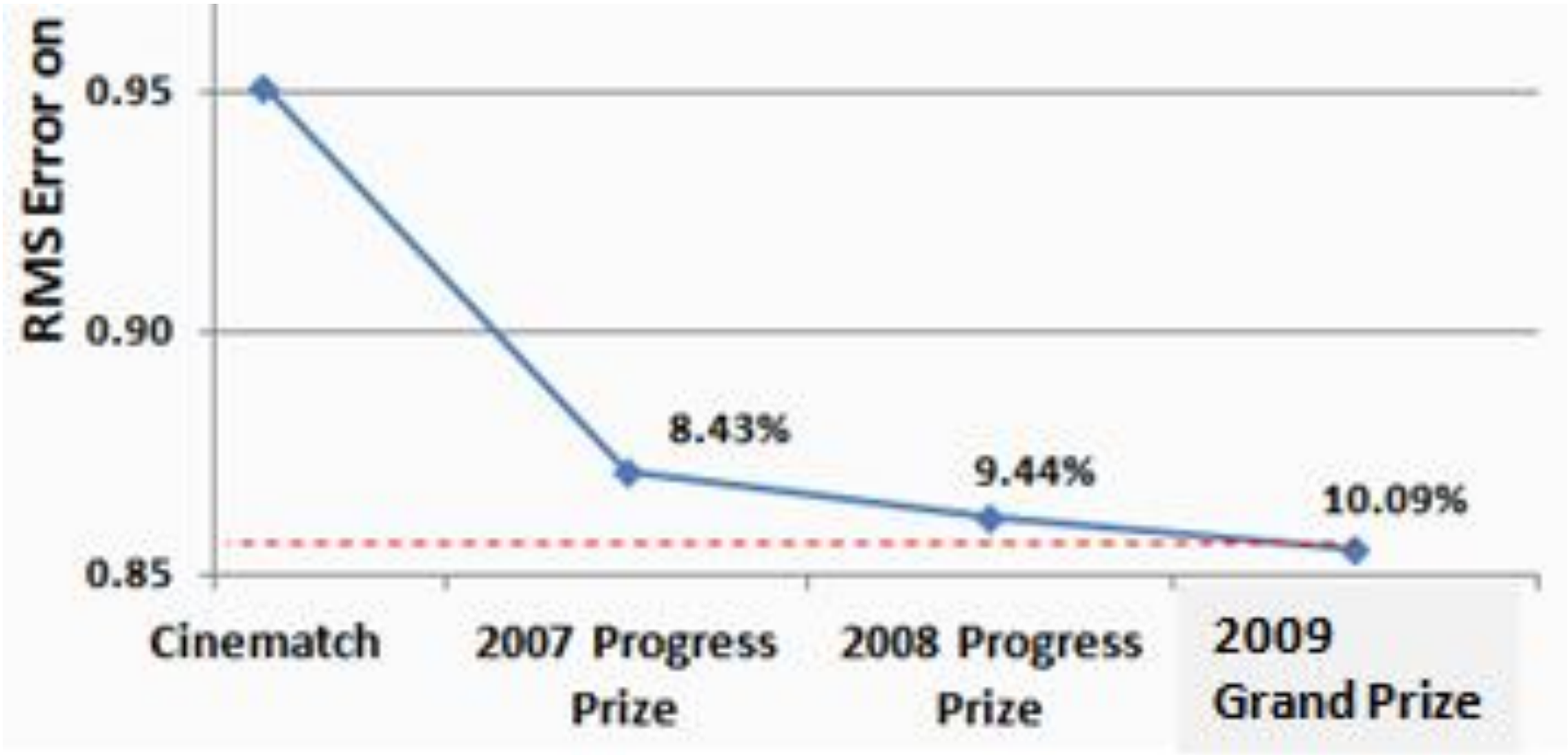



Leprechaun 4: In Space
Crocodile
Arachnid
Exercise
Bonus Material
IMAX
The Jacket
Dark Water
Dead Birds
Documentary
Slam
Sports
Equilibrium
Enigma
Misc Big Budget
When in Rome
Cirque du Soleil
G. I. Jane
Armageddon
The Rock
Family
The Passion of the Christ
Super Size Me
Critically Acclaimed
Sci-Fi
Dune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

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)

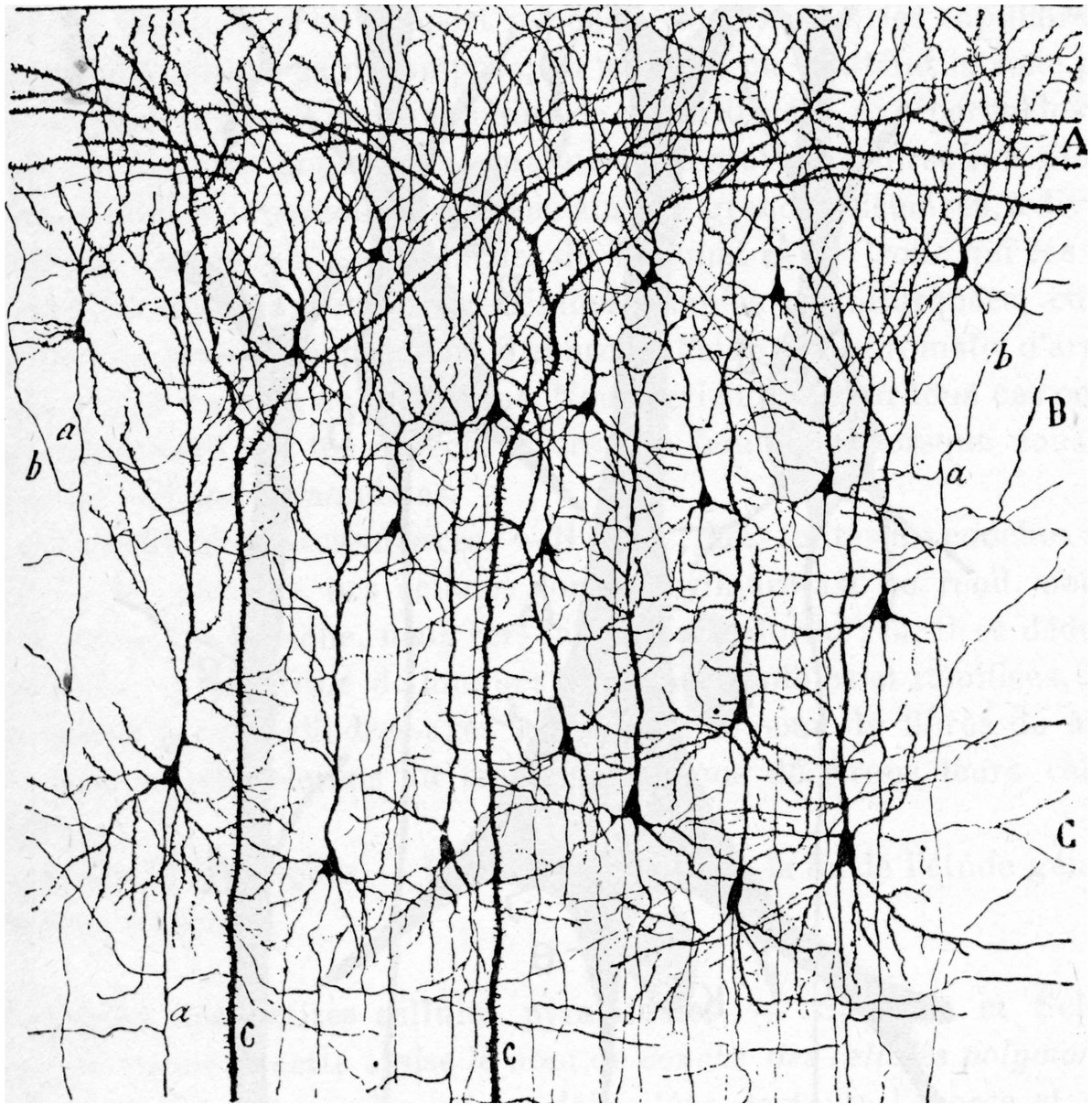
Netflix Prize Progress



http://blogs.hbr.org/cs/2012/10/big_data_hype_and_reality.html

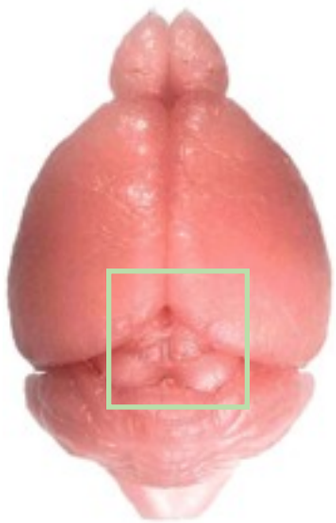
Connectome

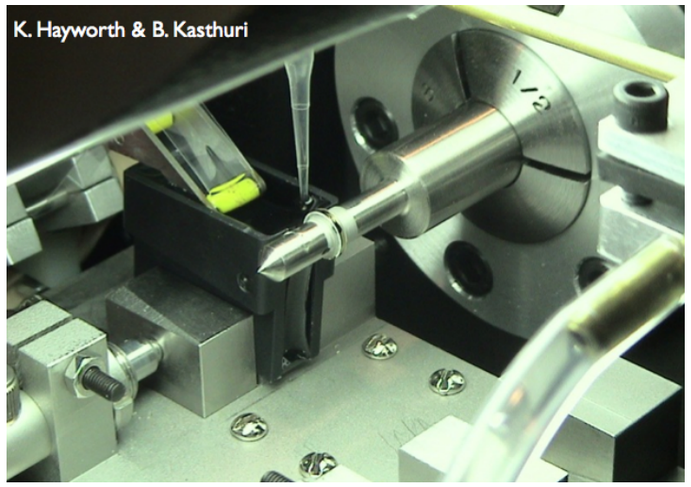
What is the connectivity of large brain circuits?

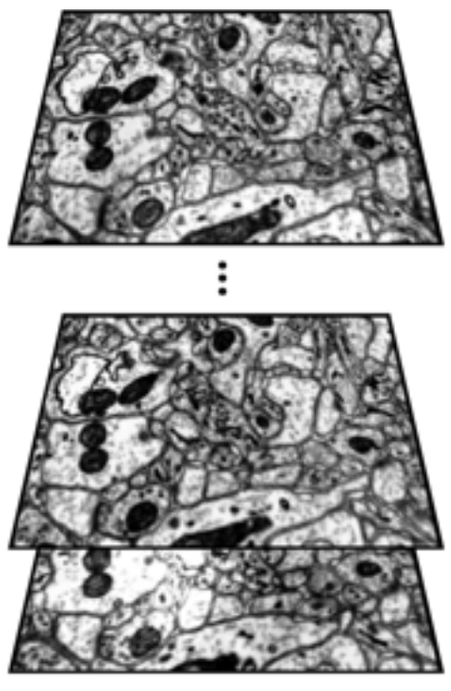


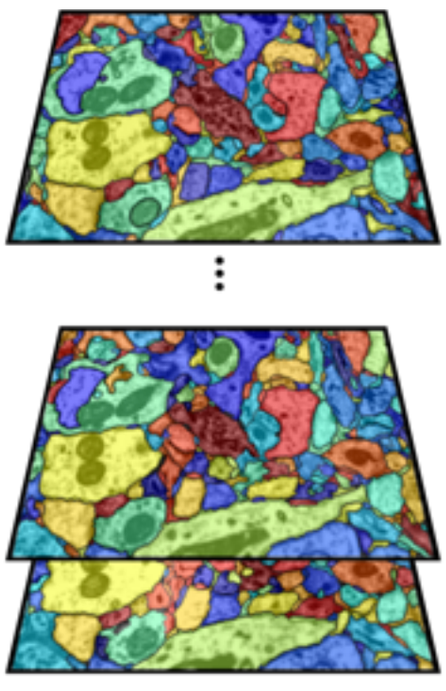
Ramón y Cajal, 1905

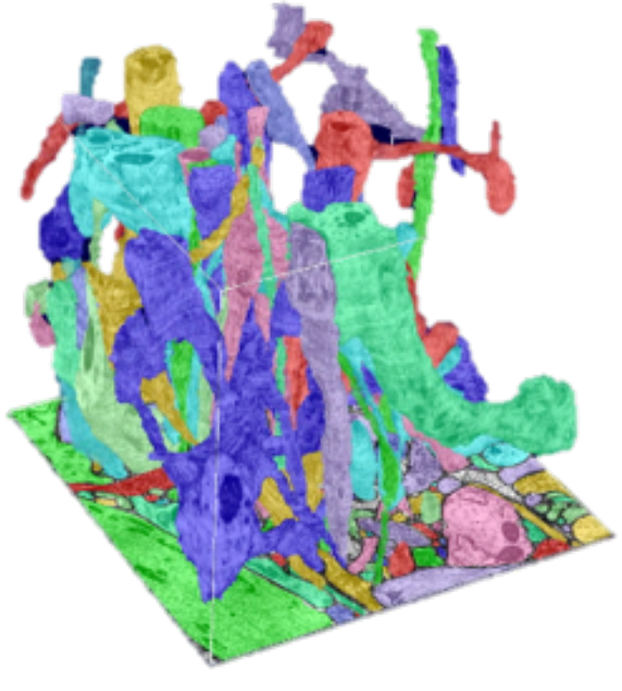
Connectome Workflow





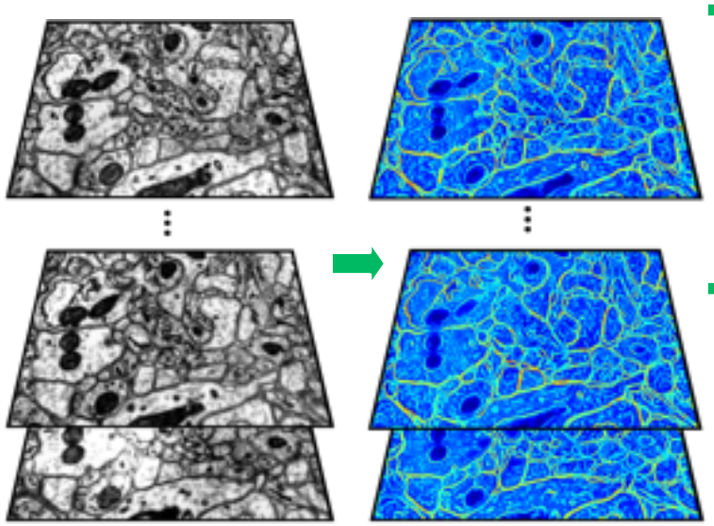




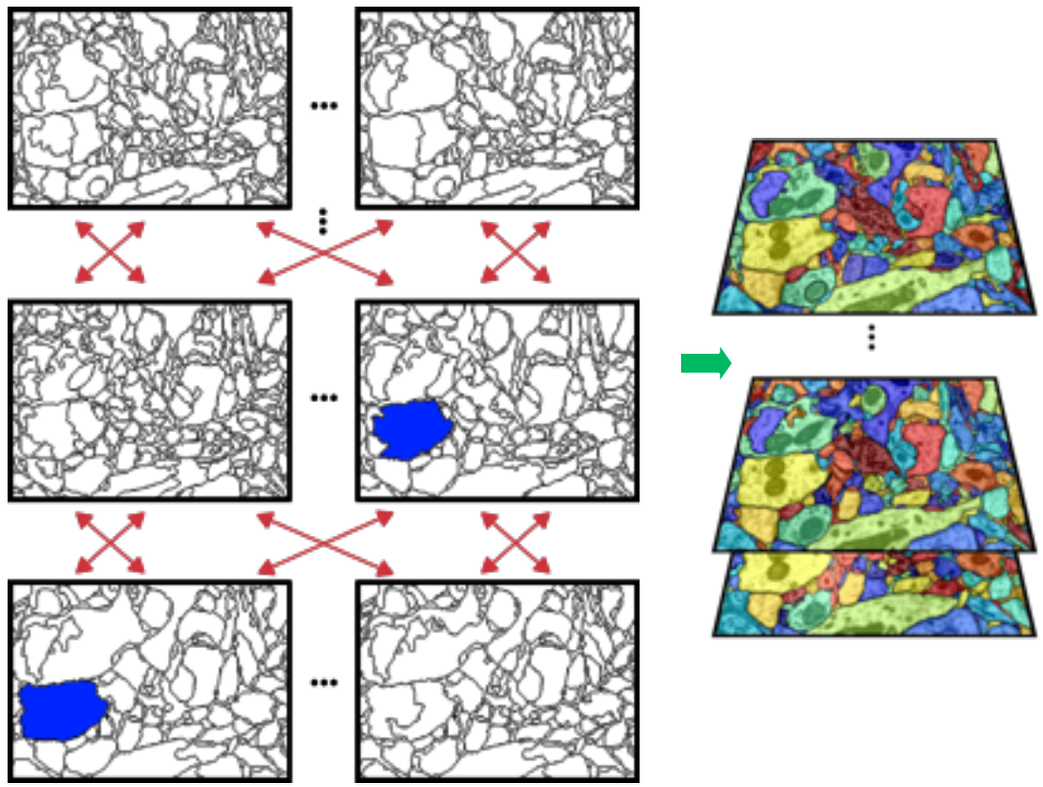


Ultra-Thin Section EM

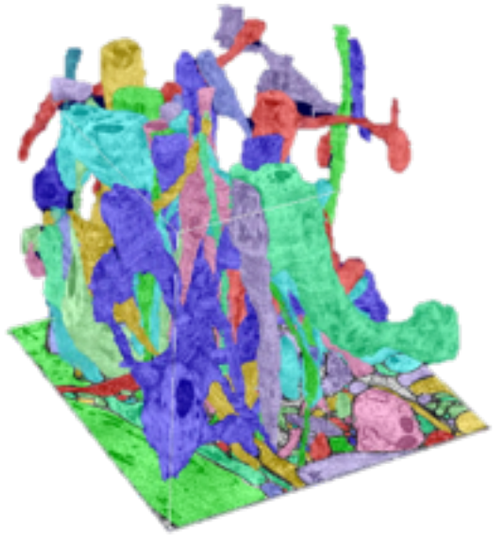
Automatic Reconstruction



2D Segmentation



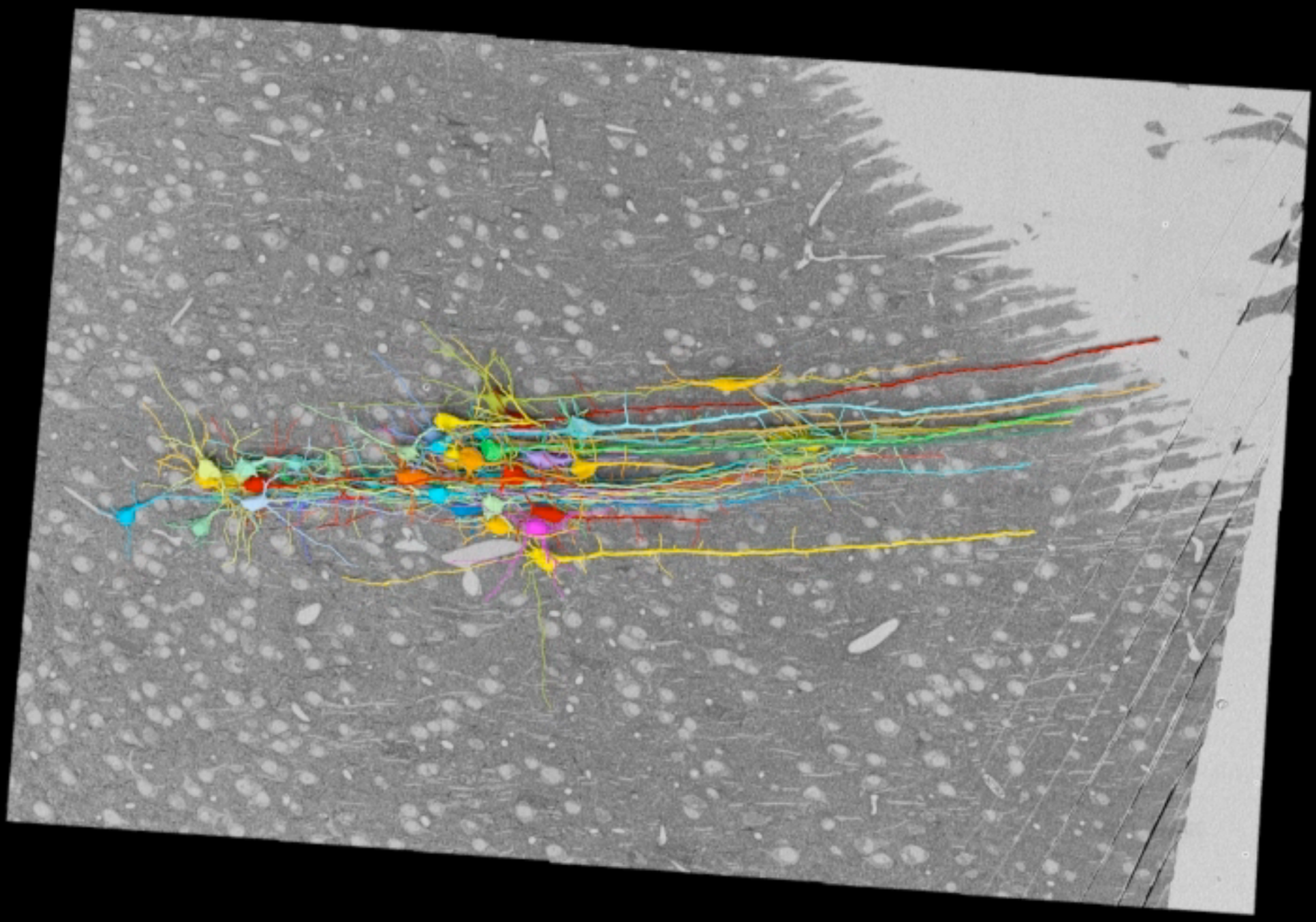
Combine Multiple 2D Segmentations with Fusion

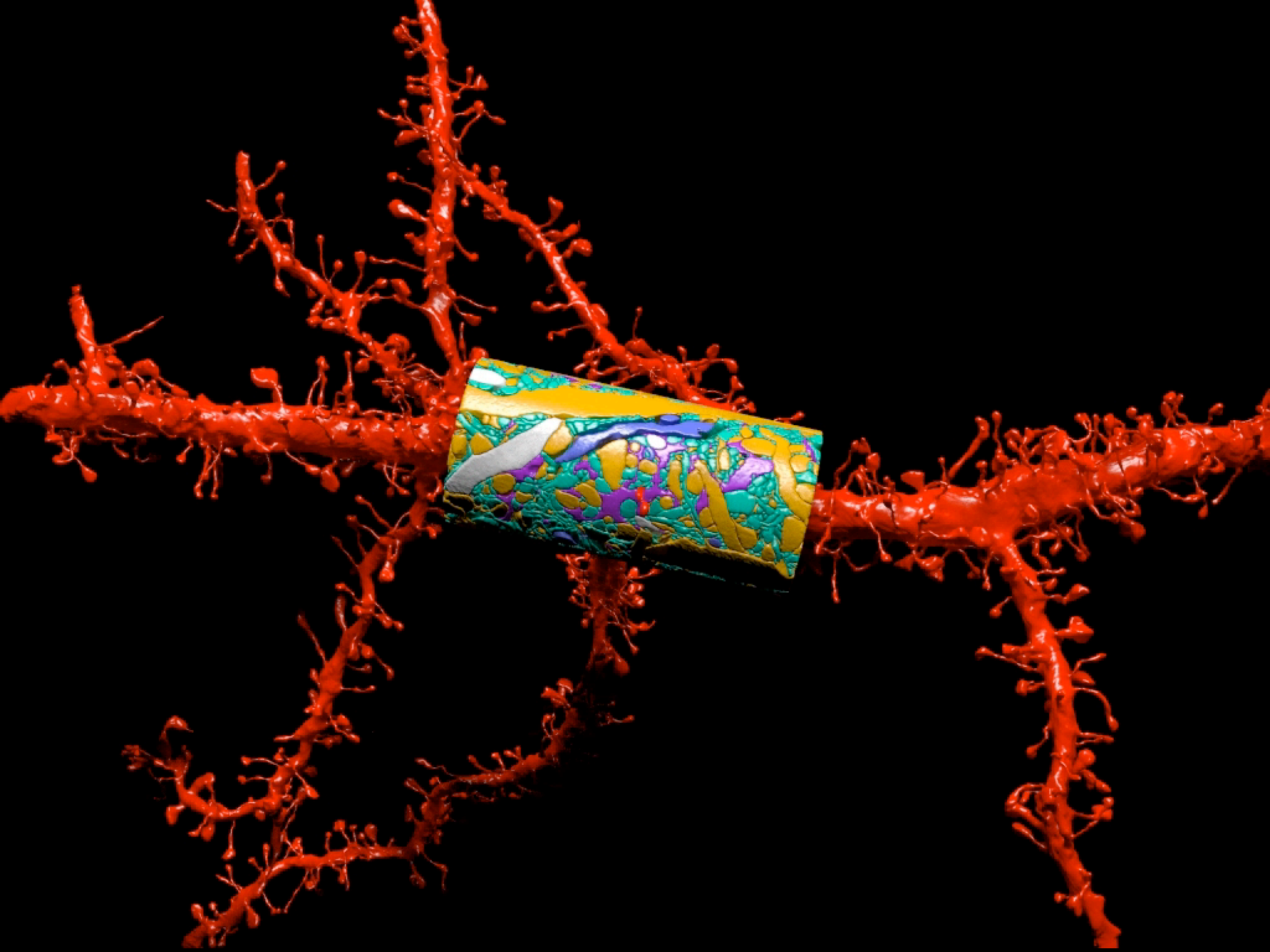


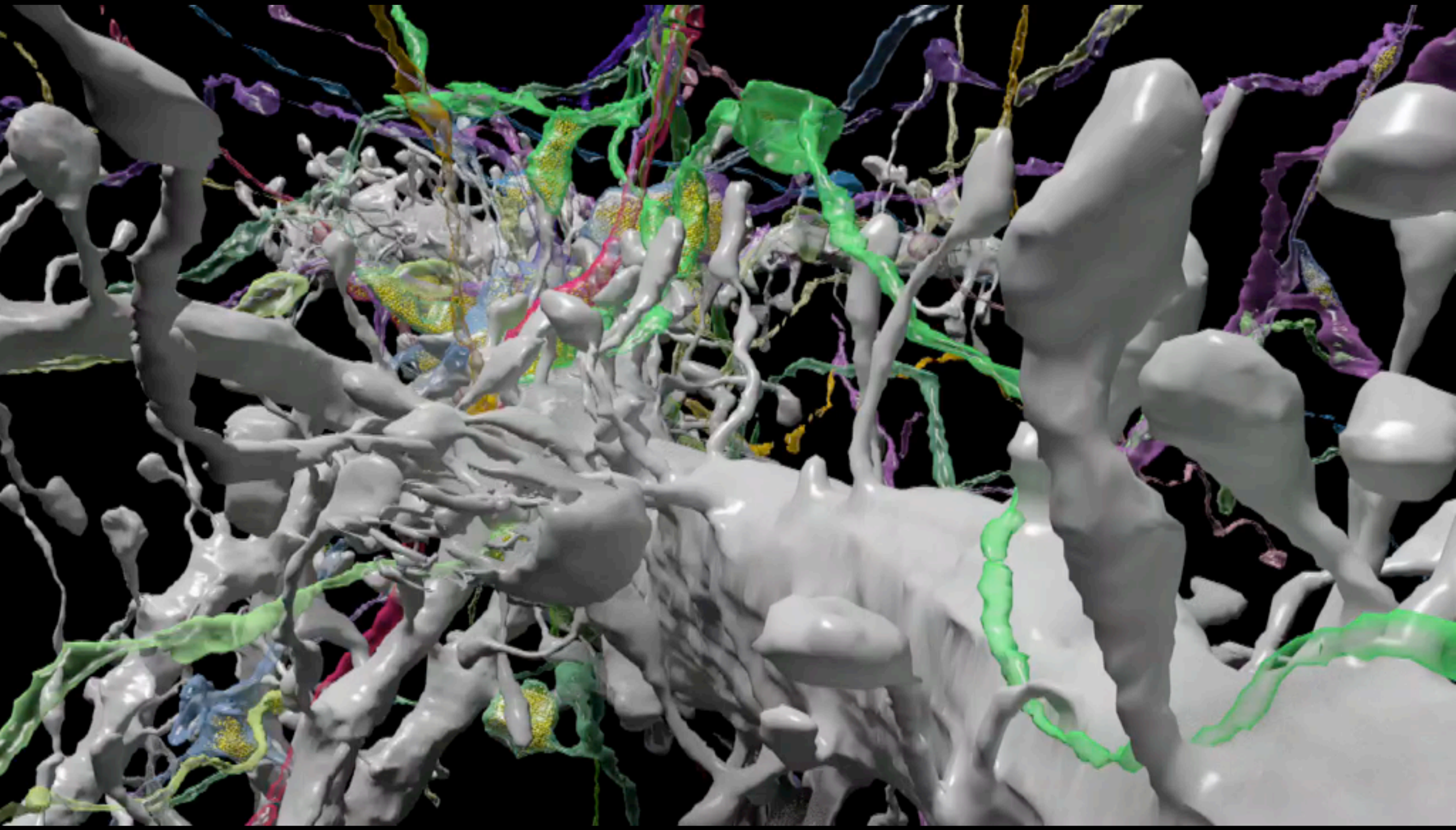
Globally Consistent 3D Segmentation

[Kaynig et al., CVPR 10]
[Vazquez et al., ICCV 2011]

2012

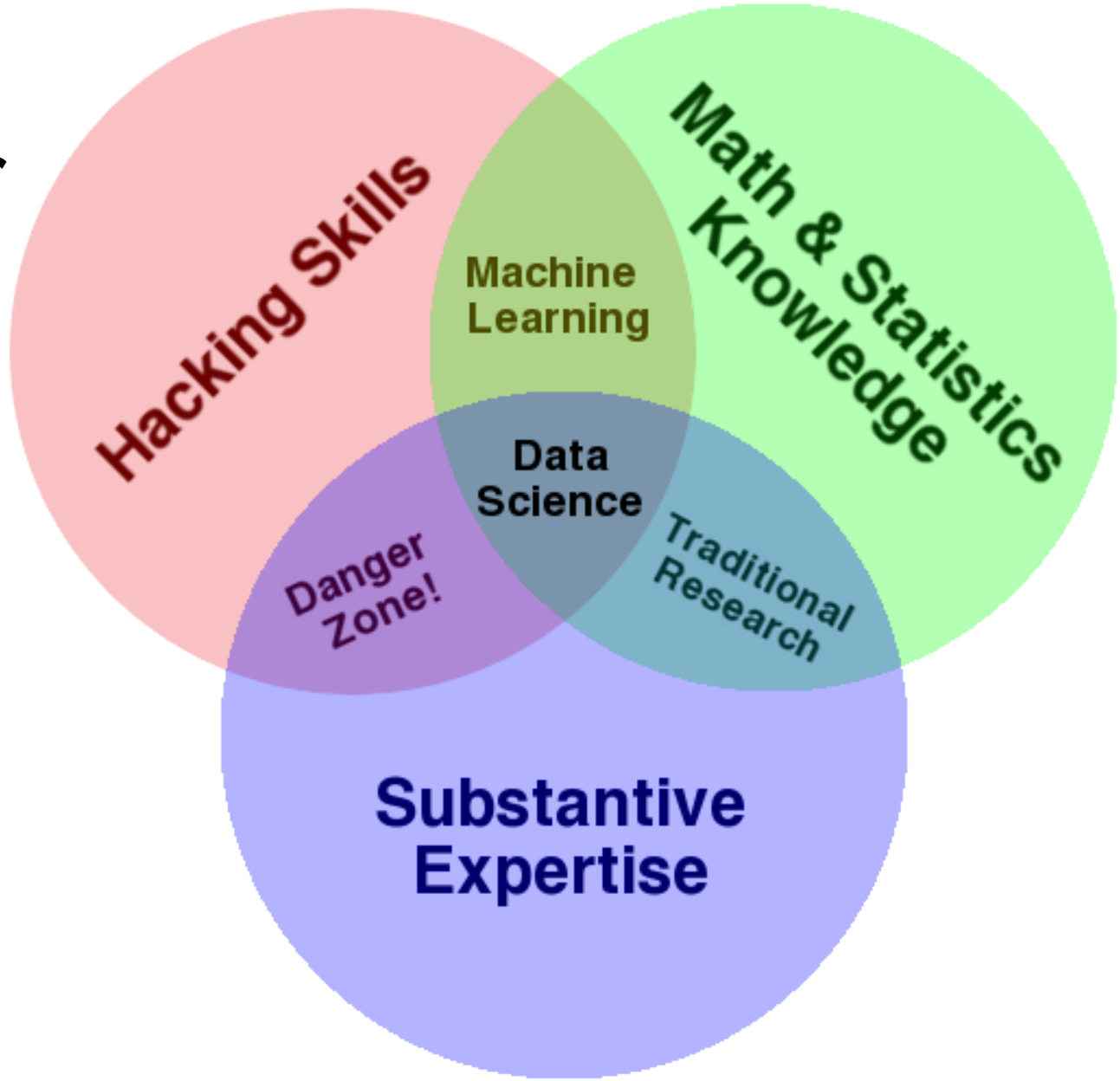






Data Science

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

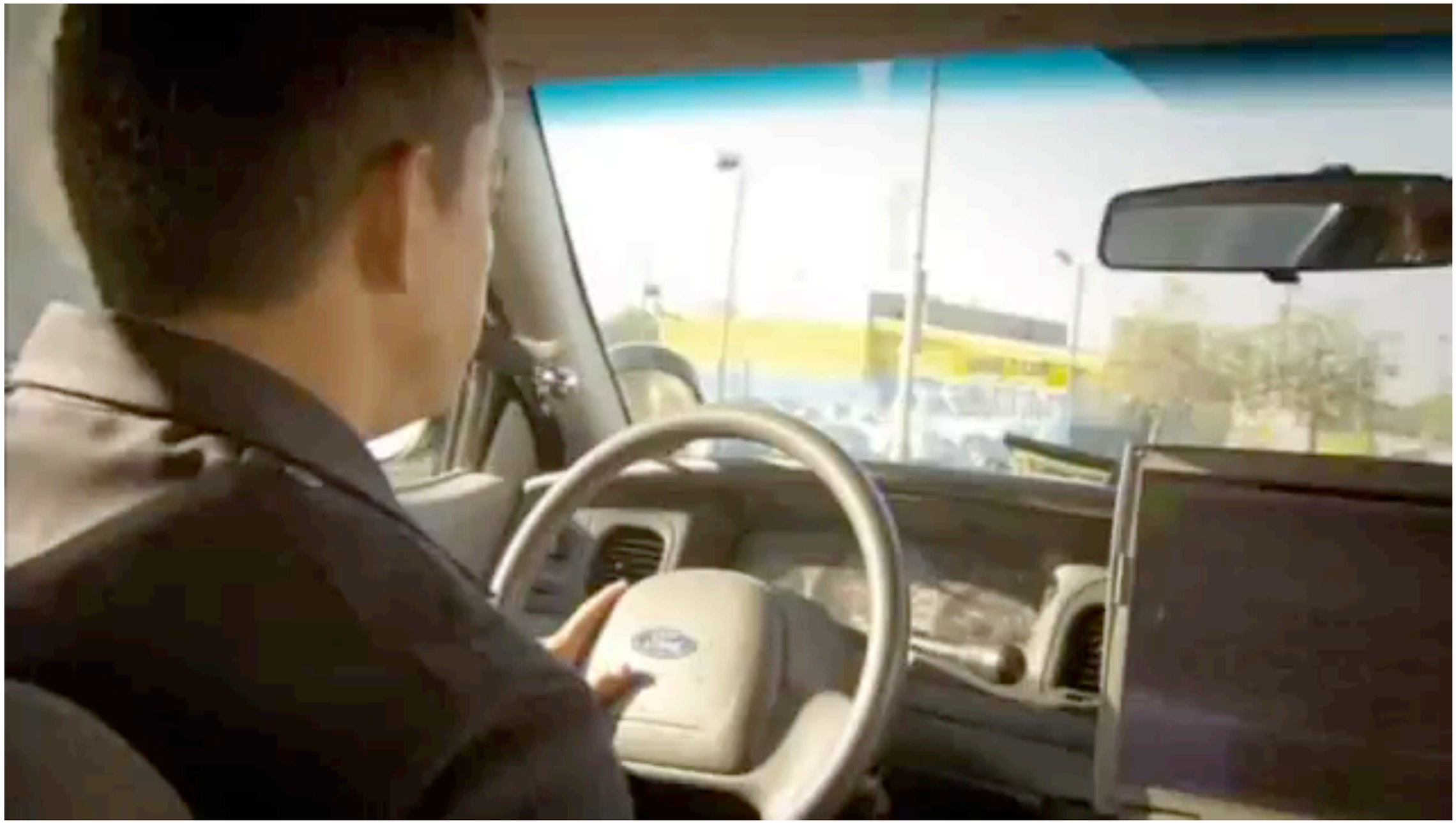
Data Science

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

Outline

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

The Age of Big Data



BBC, 2013

Crime Prevention

Cambridge police look at math to solve crimes

It is among the most notoriously difficult crimes to solve — the home break-in. There are seldom witnesses. Burglars tend to work stealthily, either under the cover of darkness or when their victims are away from home, at work or on vacation. On average, police solve no more than 13 percent of residential burglaries, according to national figures.

But two Cambridge police crime analysts and an MIT professor and doctoral student believe a computer system they developed that mathematically analyzes these crimes could be the key to solving more of them.

For the past year they have worked together to develop a calculation for quickly detecting burglary patterns, such as when and where the crime took place, how the burglar broke into the home, or whether the victim was at home sleeping or on vacation.

The algorithm, which they have named Series Finder, can analyze thousands of police incidents in minutes, looking for patterns and citing crimes that closely follow them for analysts who currently spend hours searching through computer databases trying to figure out the habits of a suspect.

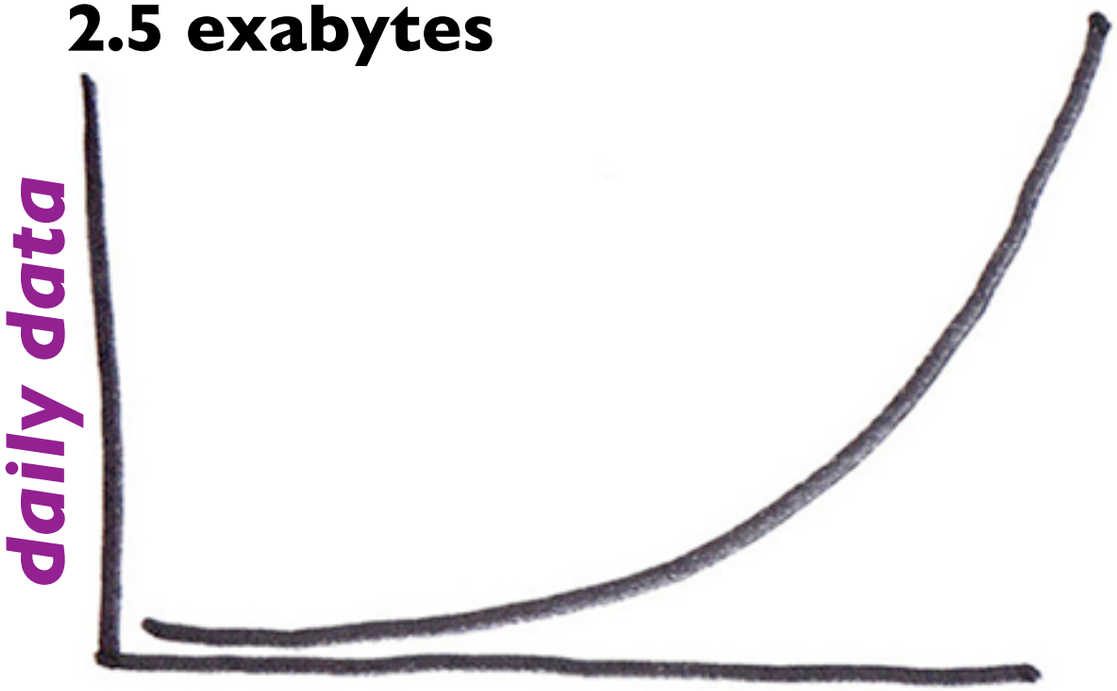
"This has the potential to be very significant," said Lieutenant Daniel Wagner, who runs the department's crime analysis unit. "This tool, if we're able to begin to use it on a daily or a regular basis, would help us identify crime series that we might not have picked up on manually being human beings."

Wagner and Rich Sevieri, the department's strategic analysis coordinator, approached officials at the MIT Sloan School of Management last year and asked

Boston Globe,
Sunday, Aug 4, 2013

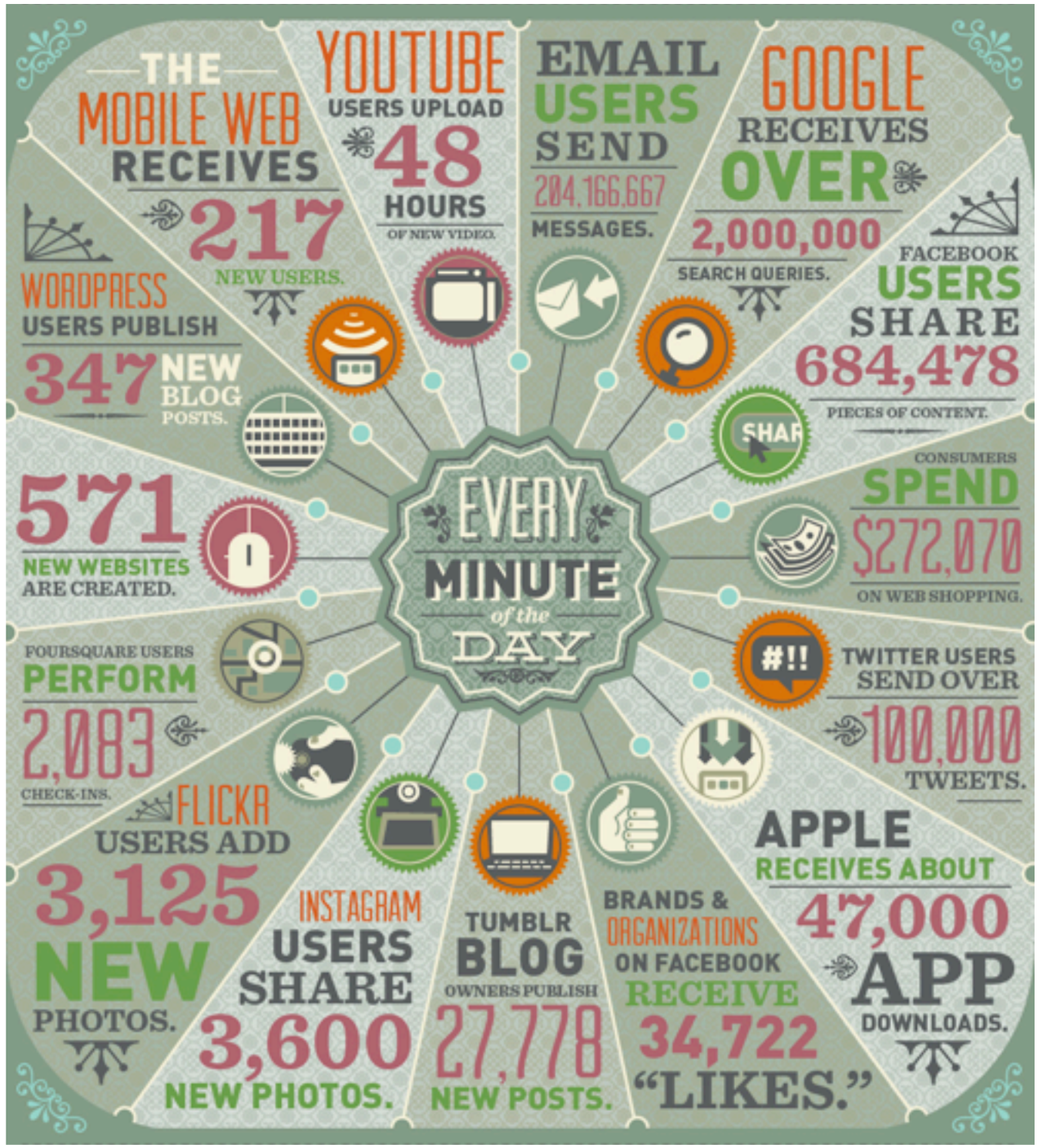
Big Data

2.5 exabytes



2012

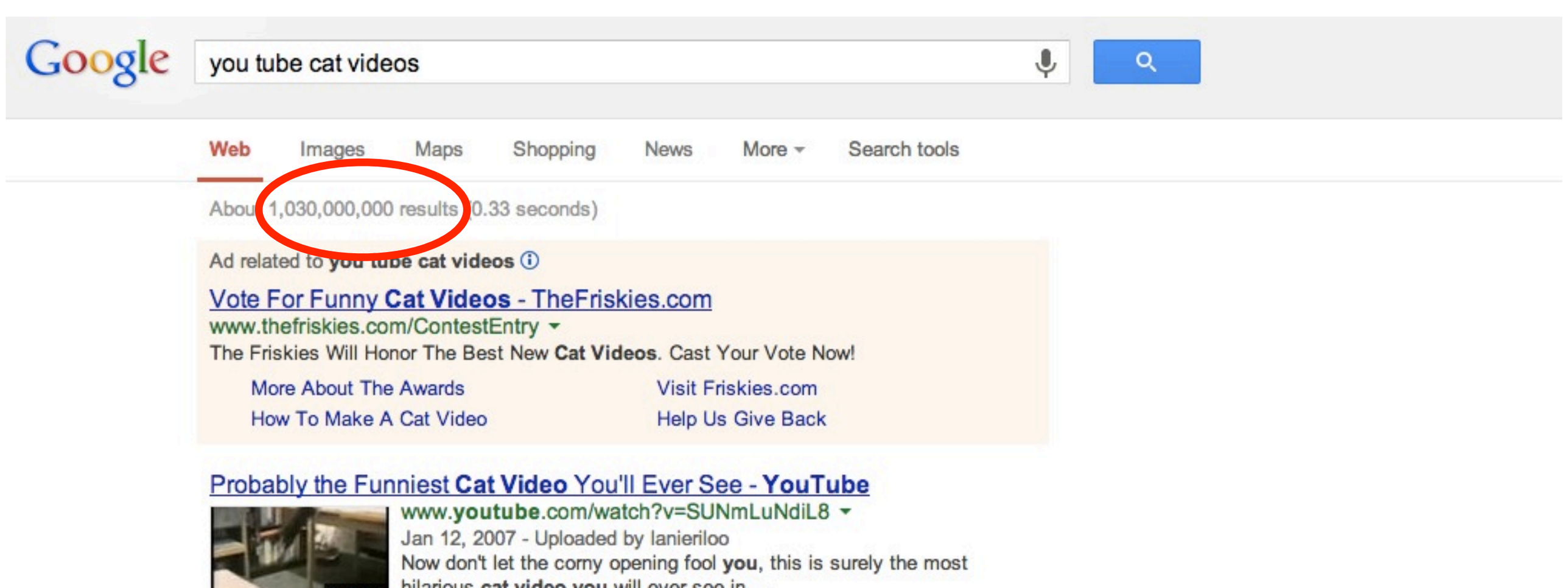
[IBMbigdata]



[Domo]

"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)

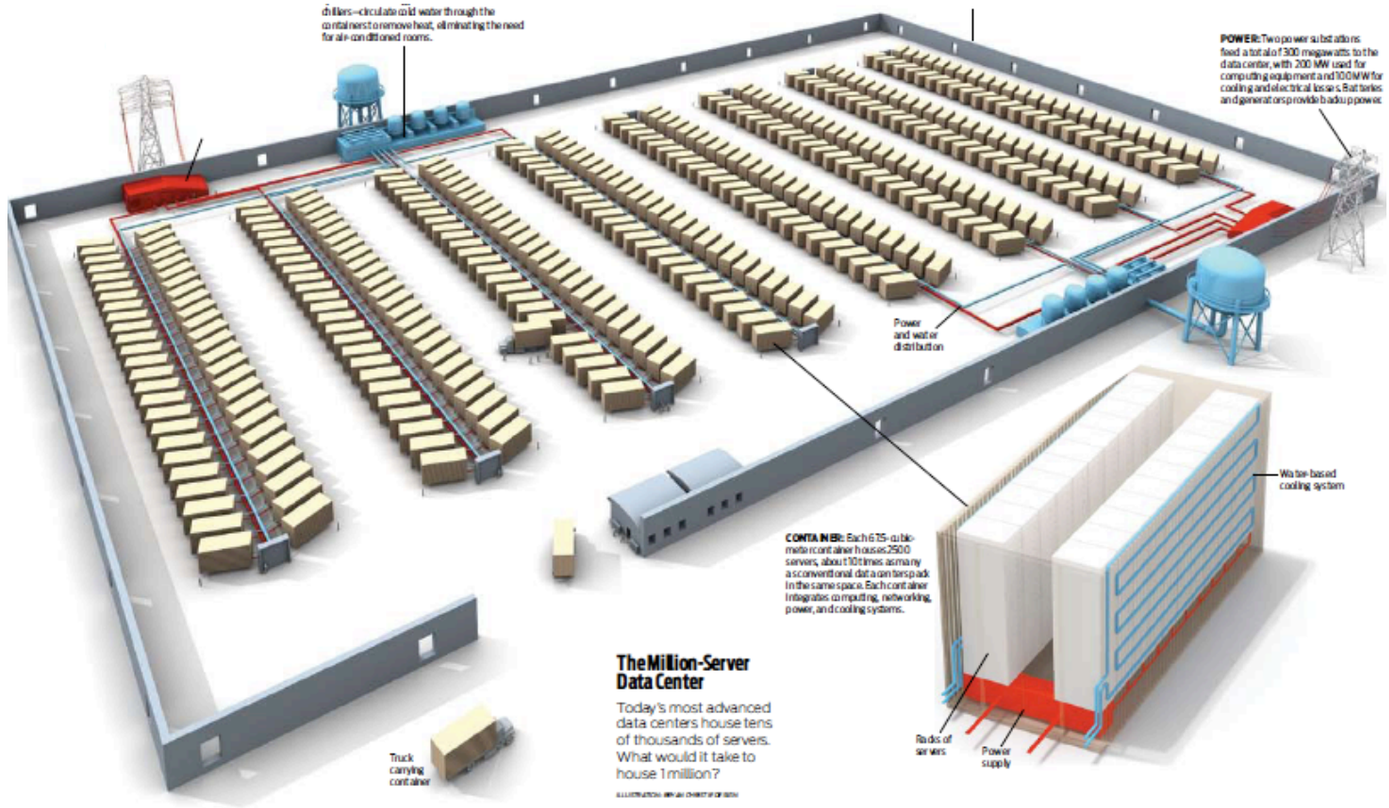


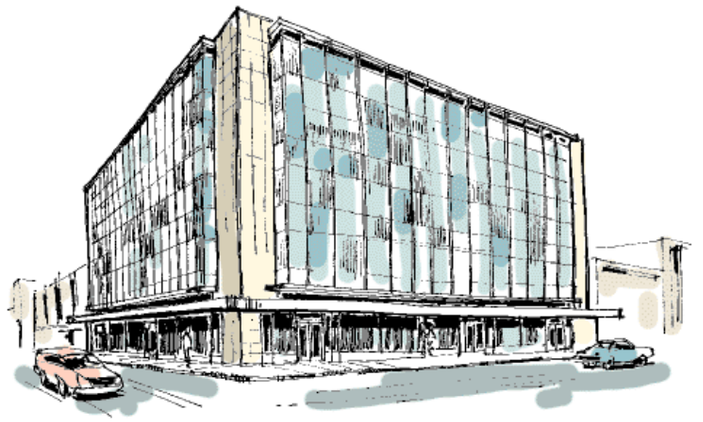
In one second on the Internet there are...

http://onesecond.designly.com/

Smarter Devices

Commodity Computing





Michael Franklin, UC Berkeley

Ubiquitous Connectivity

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

1 Zetabyte = 1 Billion Terabytes

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.

950 MILLION

USERS GENERATE 2.7 BILLION LIKES ON FACEBOOK PER DAY

90%

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

400 MILLION

NEW TWEETS ARE CREATED BY ACTIVE USERS EACH DAY

40%

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

72 HOURS

OF VIDEO IS UPLOADED TO YOUTUBE EVERY MINUTE

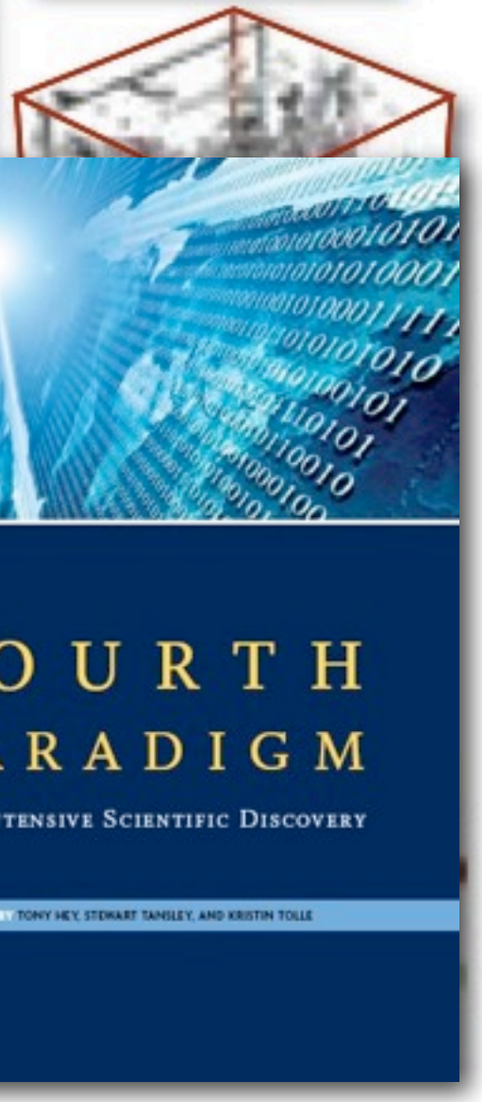
15X

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

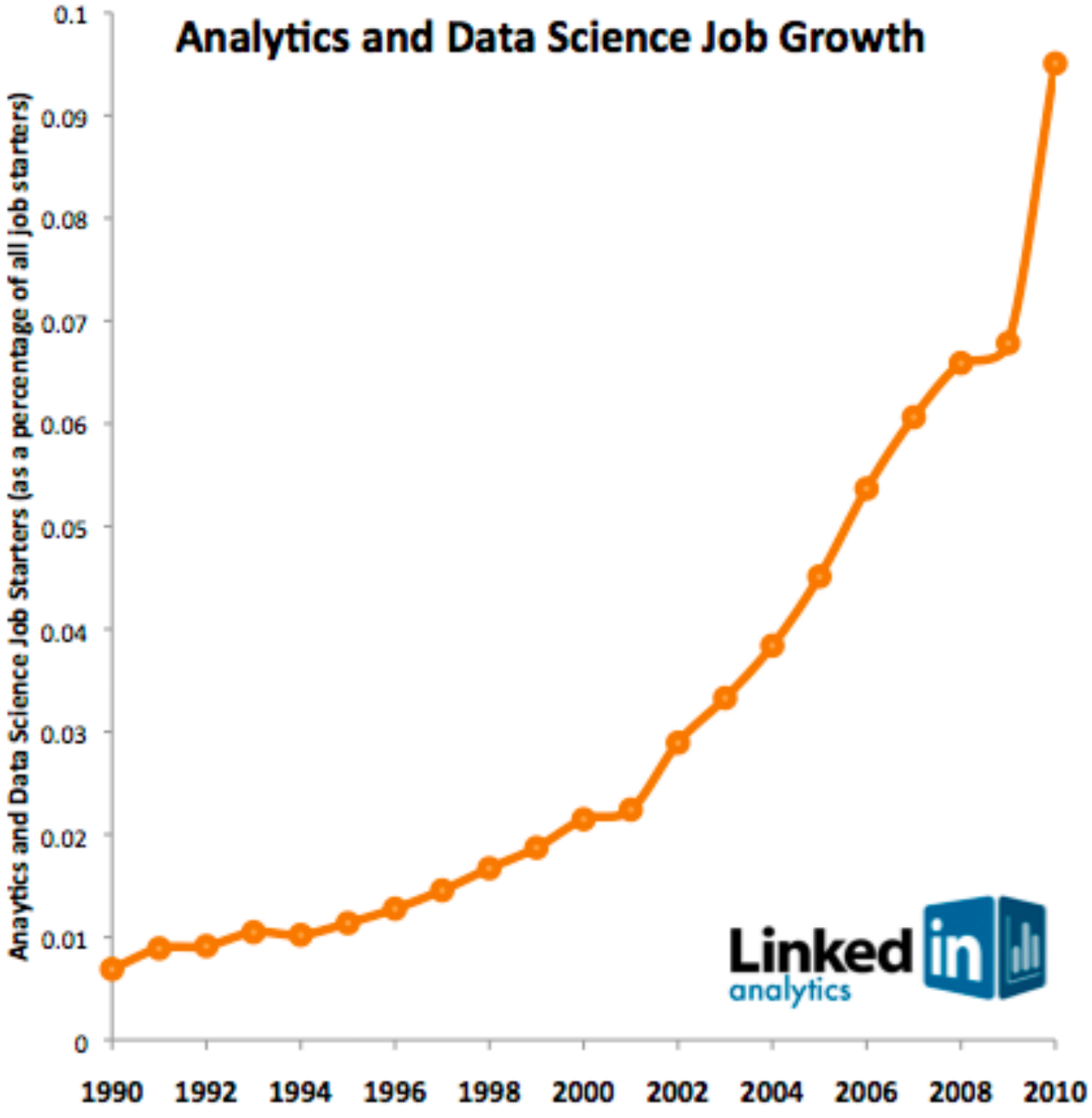
Science Paradigms

- 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

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



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"

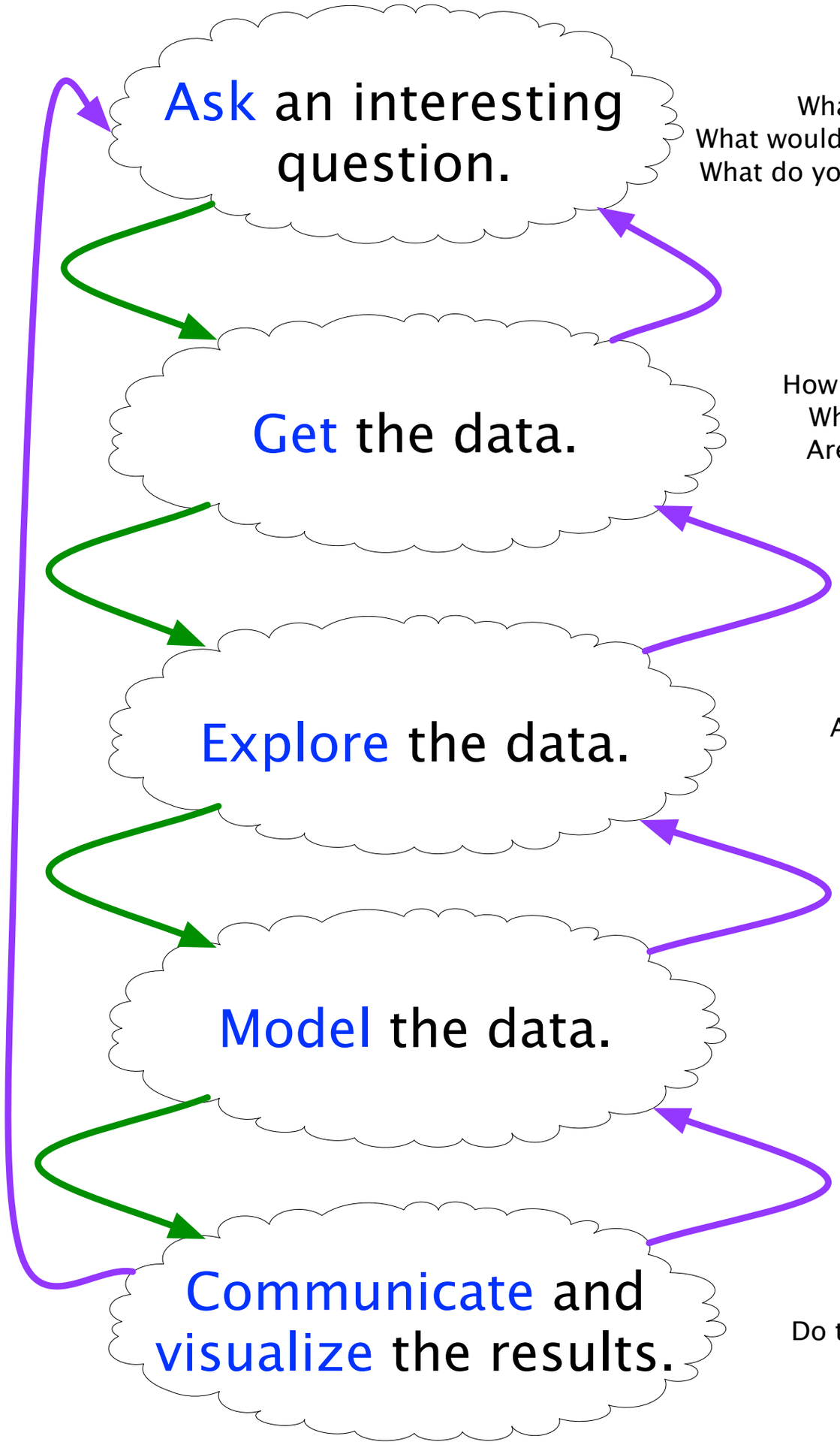
McKinsey Global Institute

"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



What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Build a model.
Fit the model.
Validate the model.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Outline

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

Hanspeter Pfister

Harvard
School of Engineering
and Applied Sciences

Hanspeter Pfister
Professor of Computer Science

Maxwell Dworkin 227
33 Oxford Street
Cambridge, MA 02138

Tel: 617.496.8269
Fax: 617.496.3012
pfister@seas.harvard.edu

gvi.seas.harvard.edu/pfister

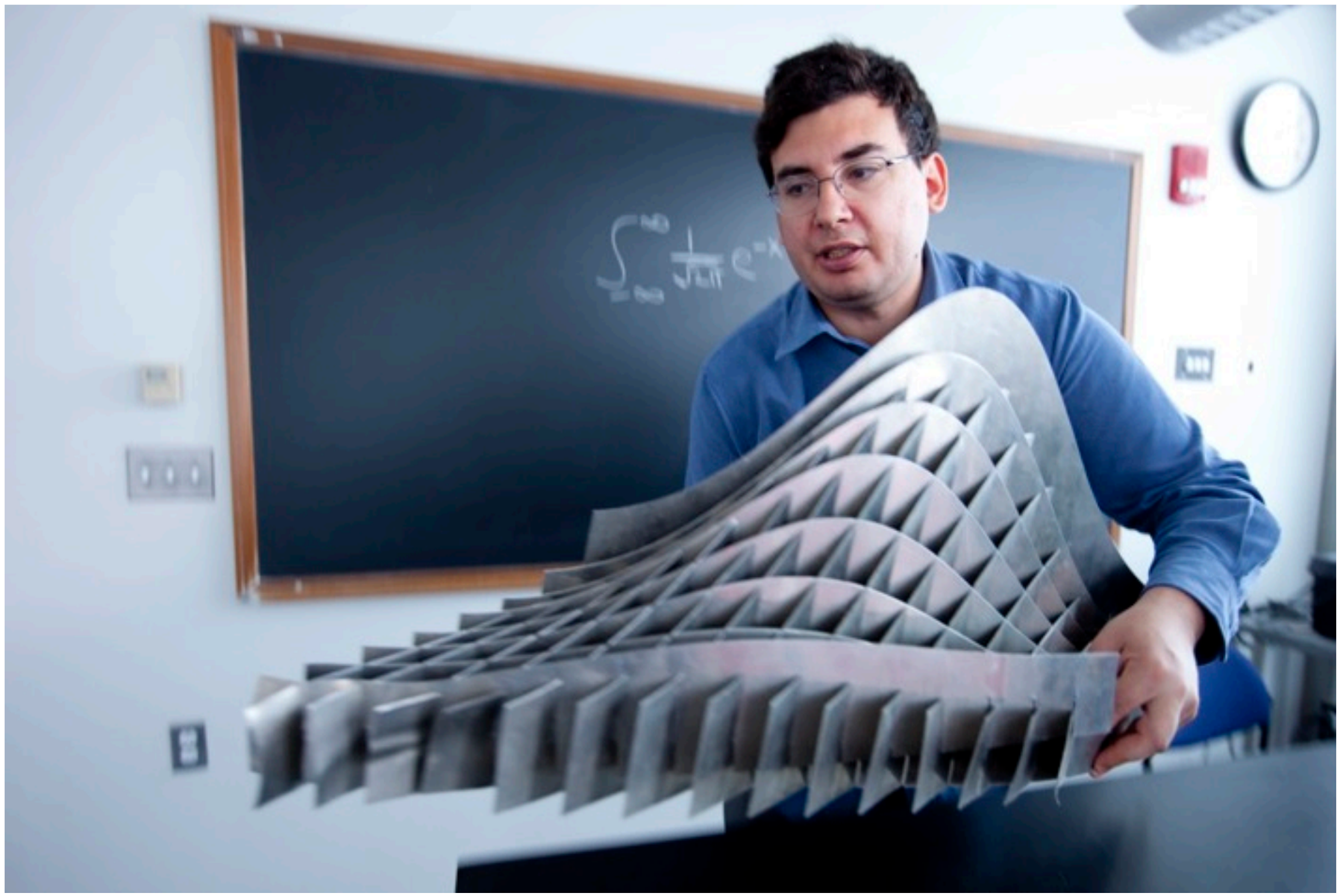
An Wang

My Background

- Grew up in Switzerland
- M.Sc. in EE from ETH Zurich
- Ph.D. in CS from SUNY Stony Brook
- 11 years in industry (MERL)
- At Harvard since 2007, Visual Computing Group (4 Ph.D., 7 PD)
- Teach CS109 / CS171, taught CS175 / CS264 / CS205
- Director of the Institute of Applied Computational Science (IACS)
- Two daughters, Lilly (10) and Audrey (7)

Joe Blitzstein

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



CS109 Staff

- Chris Beaumont, Head TF
- Johanna Beyer
- Nicolas Bonneel
- Alex D'Amour
- Rahul Dave
- Brandon Haynes
- Ray Jones
- Steffen Kirchhoff
- Seymour Knowles-Barley
- Alexander Lex
- Deqing Sun
- Tim Brenner, A/V

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 - especially big 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 1: Predictions

- Data Science Process
- Data Types and Data "Munging"
- Probability Review
- Classification & Regression
- Cross Validation, Clustering
- Visualization & Story Telling

Act II: Recommendations

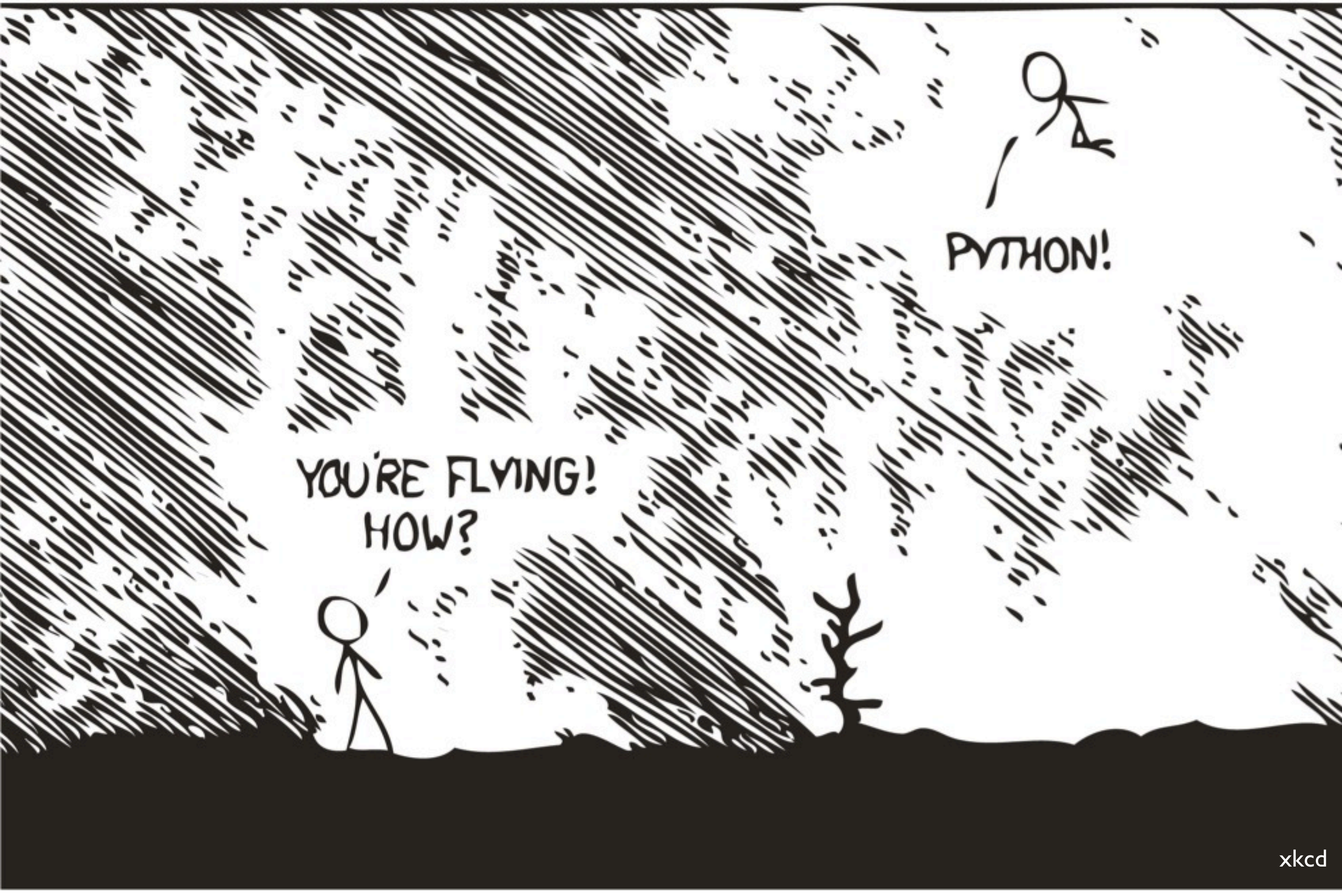
- Bayesian Thinking & Computation
- Monte Carlo Methods
- Machine Learning Methods
- MapReduce and Amazon's EC2
- Databases (Margo Seltzer)

Act III: Network Analysis

- Network Visualization
- Network Sampling
- Community Detection
- Guest Lecture

Abstractions...

...and Tools

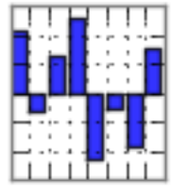

PYTHON!
YOU'RE FLYING!
HOW?
xkcd

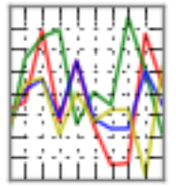
IP[y]: IPython

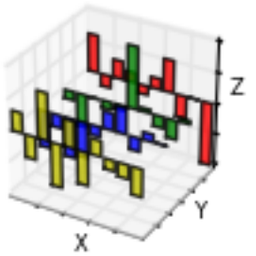
Interactive Computing

pandas

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







scikits learn

machine learning in Python

NumPy

SciPy.org Sponsored By ENTHOUGHT

matplotlib

mrjob

Homework

- Real-World focus
  - Scrape and wrangle messy data
  - Apply sophisticated statistical analysis
  - Visualize and communicate results
- Election data, movie reviews, Yelp! data, etc.

Final Project

- Pick a project of your choosing
- Teams of up to 2 students
- Process books, web sites, screencasts
- IPython (exceptions possible)
- Best project prizes!

cs109.org

CS109 Data Science

Harvard
School of Engineering
and Applied Sciences

Home Piazza Syllabus Schedule Homework Readings Resources

Learning from data in order to gain useful predictions and insights. This course introduces methods for five key facets of an investigation: data wrangling, cleaning, and sampling to get a suitable data set; data management to be able to access big data quickly and reliably; exploratory data analysis to generate hypotheses and intuition; prediction based on statistical methods such as regression and classification; and communication of results through

Instructors:
Hanspeter Pfister, Computer Science
Joe Blitzstein, Statistics

Staff:
Chris Beaumont, Head TF
Johanna Beyer
Nicolas Bonneel
Alex D'Amour
Rahul Dave
Brandon Haynes
Ray Jones
Steffen Kirchhoff
Seymour Knowles-Barley
Alexander Lex

Is this course for me ???

Prerequisites

Programming experience

- 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
  - Good test of your basic skills
  - Installation of several Python frameworks
    - Not graded, do it as soon as possible
- Read syllabus carefully
- Do readings
  - Post comments to Piazza using #readings