Data Science — Big Picture

Volodymyr Minin

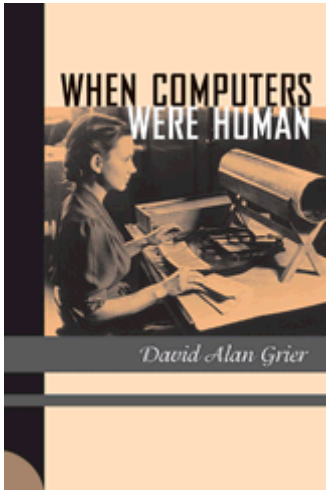
UCI Department of Statistics
Donald Bren School of Information & Computer Sciences

ISI-BUDS
June 25 2024

Computers and Data

The historical meaning of the term "computer":

"one who computes" (i.e., a person)



Since the 1700's, statisticians have been using "computers" to analyze data – so its not a new idea

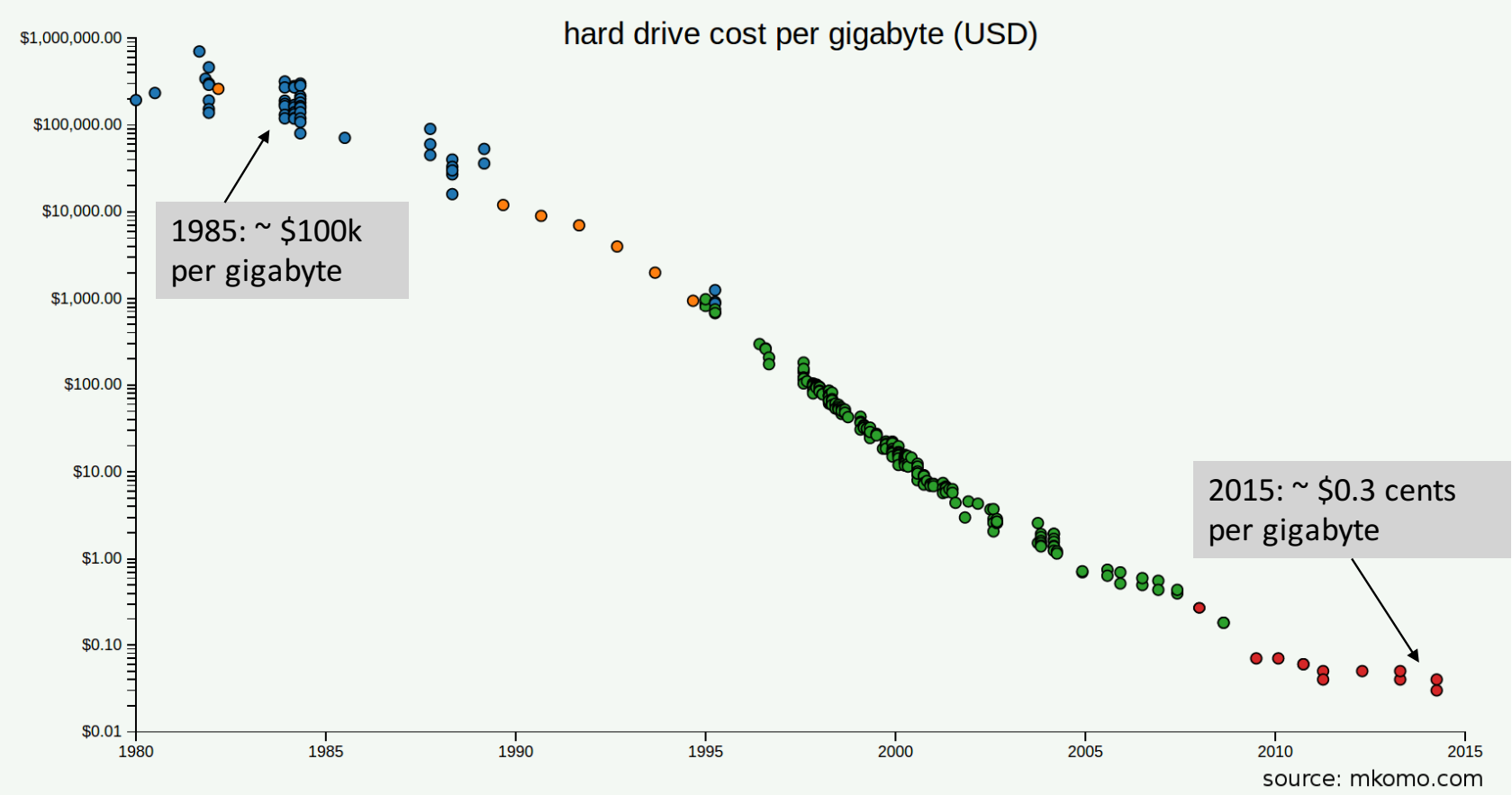
For example, Karl Pearson, one of the founders of statistics, directed a team of "computers" in his lab in London around the early 1900's

…..but for many years, "computers" could only work on relatively small problems

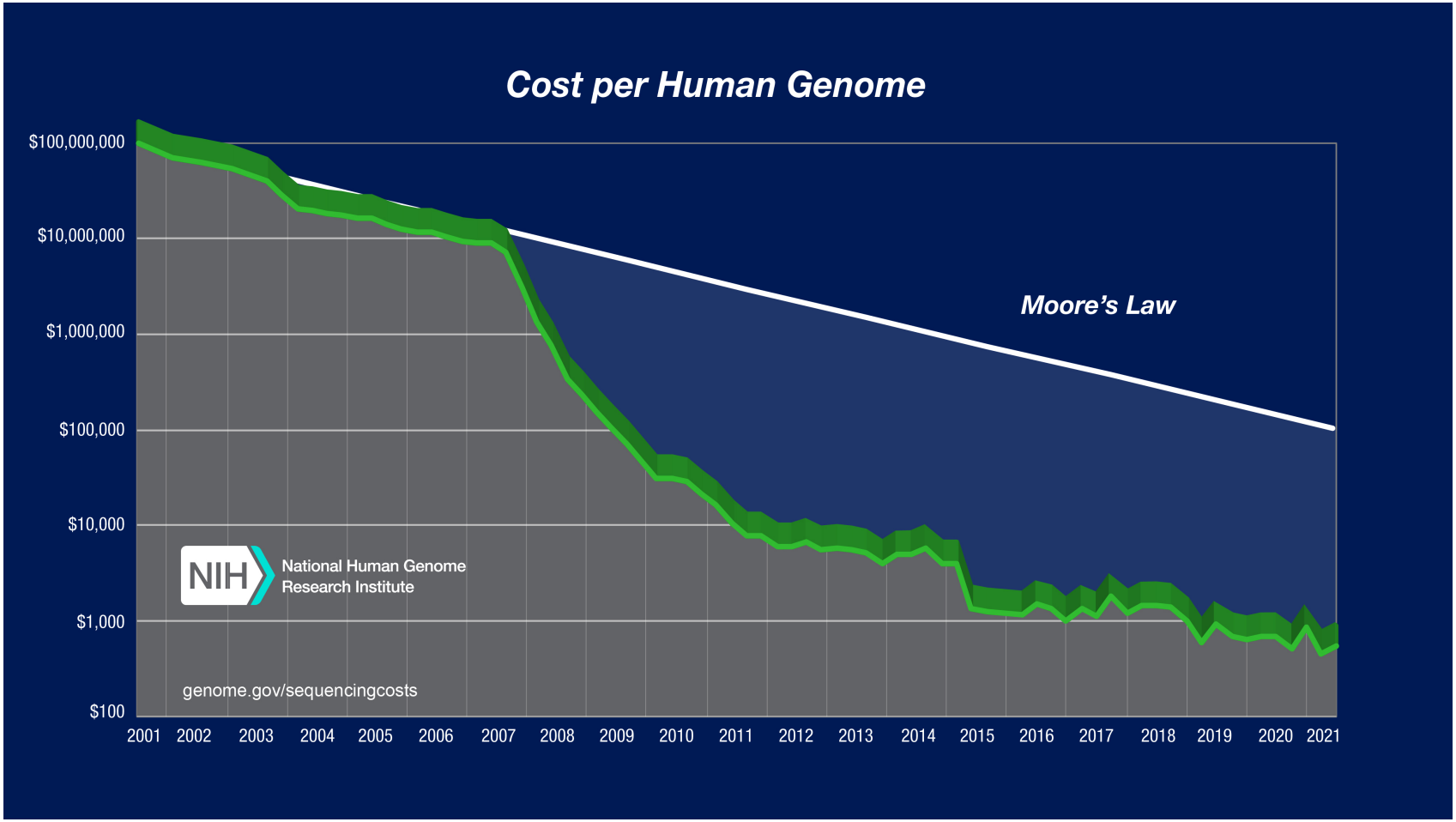
Statistics and Modern Computing

- Post World War II
  - Increasing use of computing to solve algorithmic aspects of statistical analyses
- 1960's
  - Development of statistical computing and exploratory data analysis
- 1980's
  - Computing allowed statisticians to explore more flexible models
  - Increase in use of "non-parametric" techniques and simulation methods
- 1990's
  - Development of "machine learning" — very flexible predictive modeling techniques developed in computer science
- Today
  - Data science = computing + statistics + applications

Data storage became cheaper



Data revolution in Biology



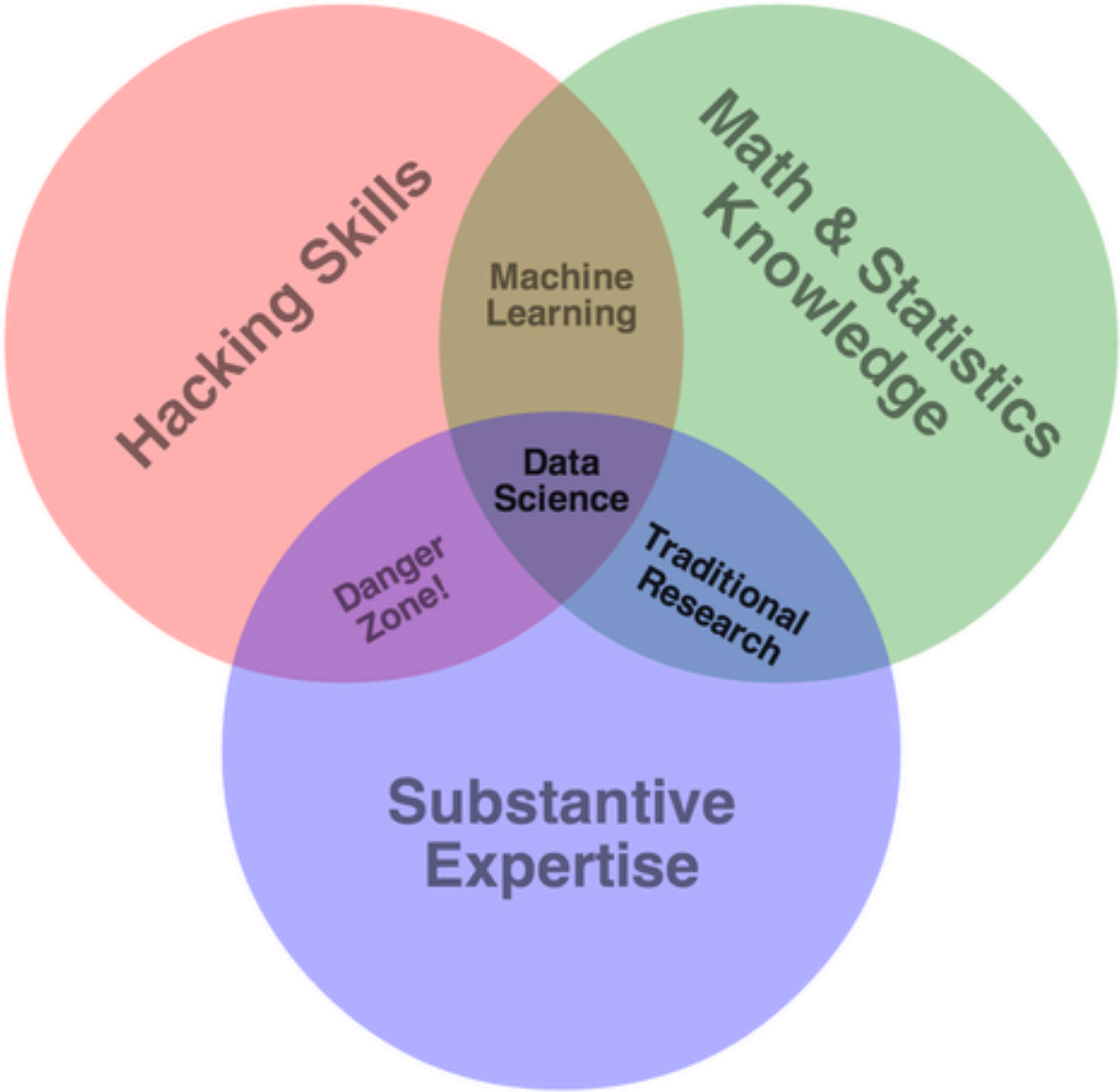
A Paradigm shift in data analysis

- Technological drivers
  - Sensors (cheap and ubiquitous, e.g., GPS on your phone)
  - Data storage (we are all “data owners”)
  - Computational power
  - Data analysis methods (statistics and machine learning)
  - Internet and wireless communication (can collect and share data)
- Convergence — tremendous demand for data analysis
  - In business, in sciences, in medicine, in engineering, and more.....
- In the past, this demand was met by statistics
  - Does not scale up — there are not nearly enough statisticians
  - Need more tools than just statistics: need databases, algorithms, machine learning,...

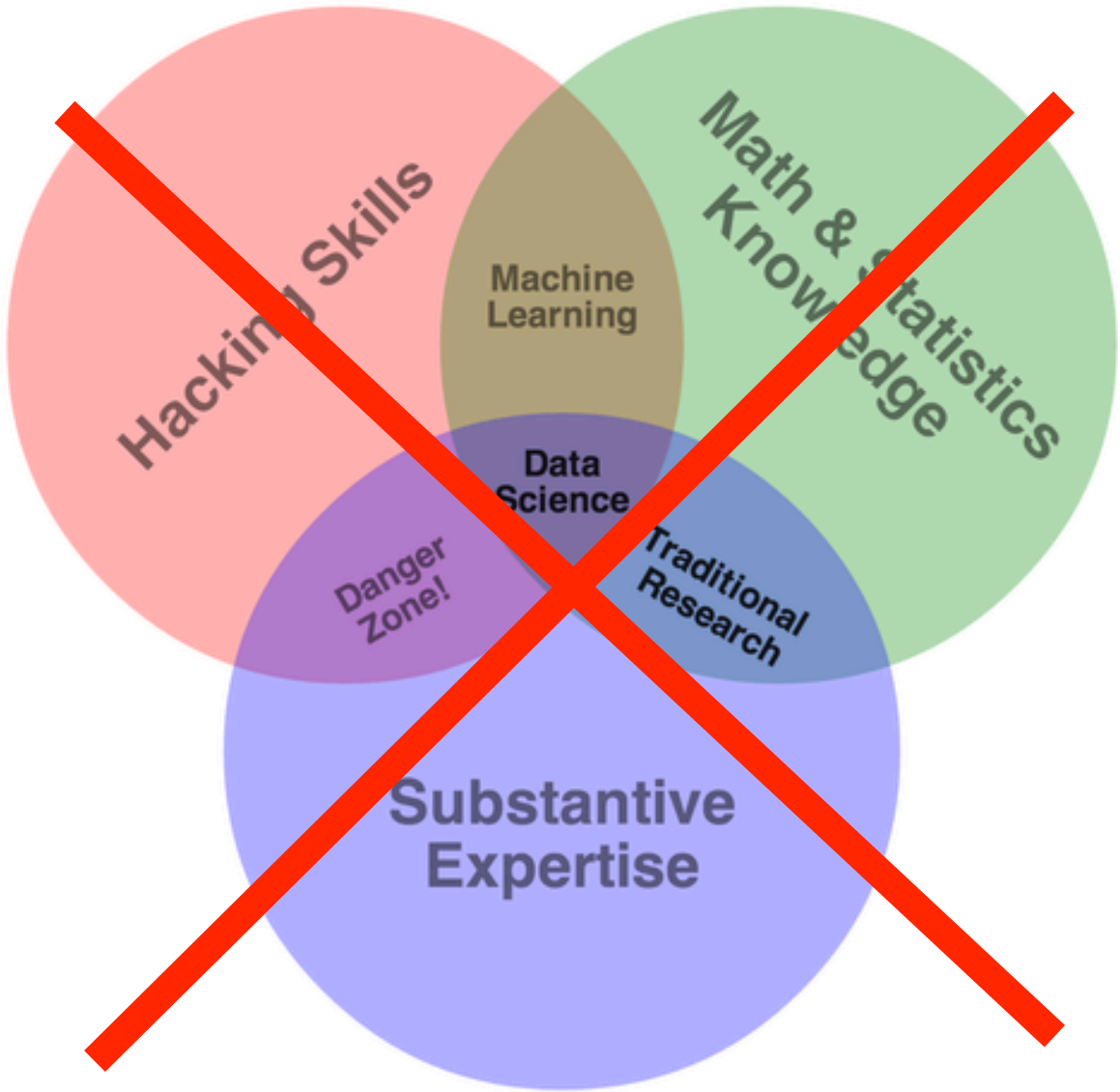
What is Data Science?

- Data science involves the full lifecycle of data: from messy unstructured data to predictions and decisions
- Data science is broader than just databases, statistics, ML, algorithms, but these are all critical components
- Key aspects of data science include
  - Domain knowledge and problem definition
  - Data preparation/organization/management
  - Understanding of uncertainty (statistics)
  - Computing, algorithms, fitting models, machine learning
  - Iterative exploration and experimentation
  - Human judgement and interpretation

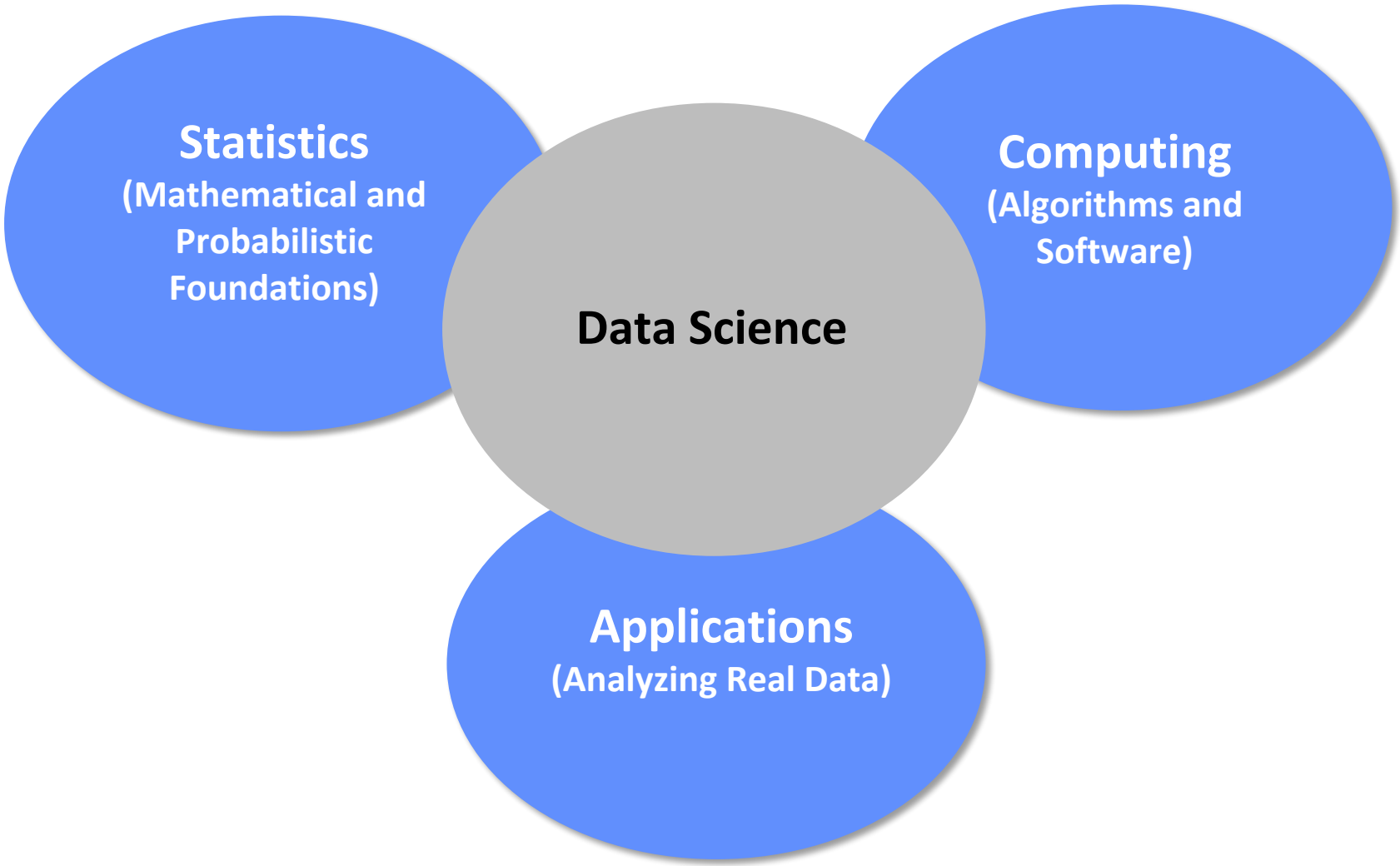
Components of Data Science



Components of Data Science



Components of Data Science

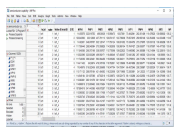


Data pipeline

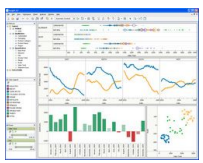
Unstructured Data

Extracted Data

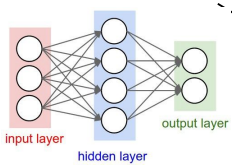
Transformed Data



Data for Modeling

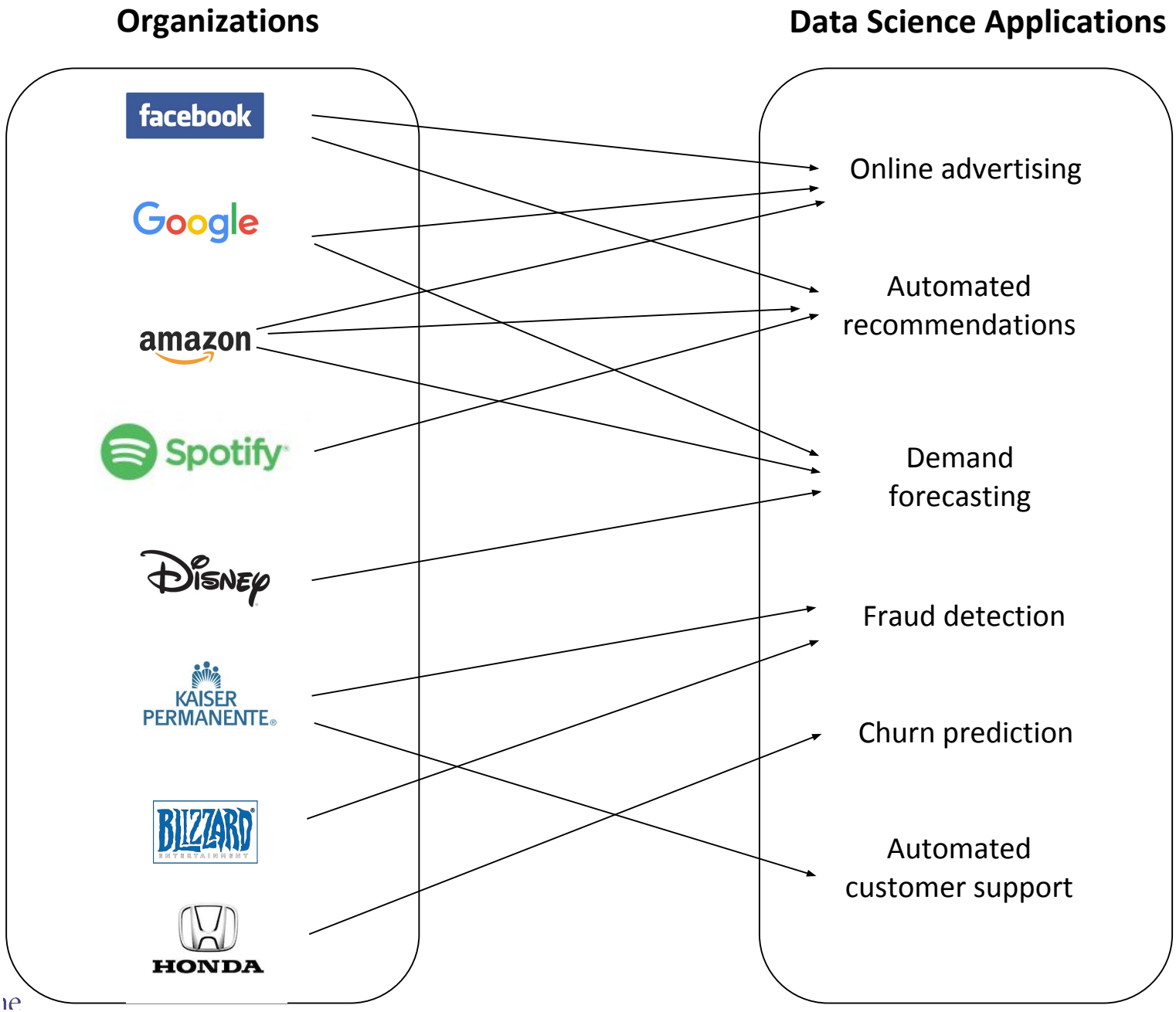


Predictive Model



Predictions/ Decisions

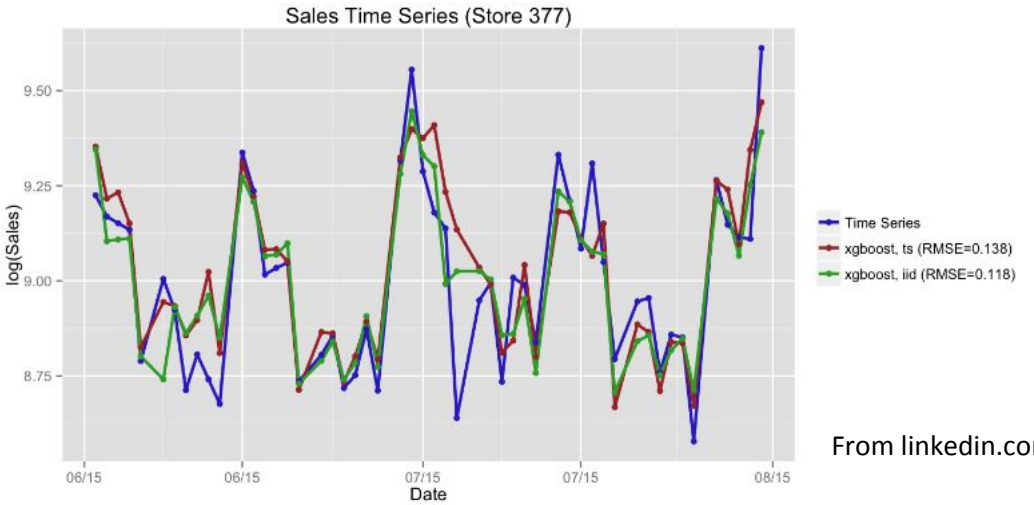
How is Data Science used?



How does Amazon forecast how many items for its warehouses?

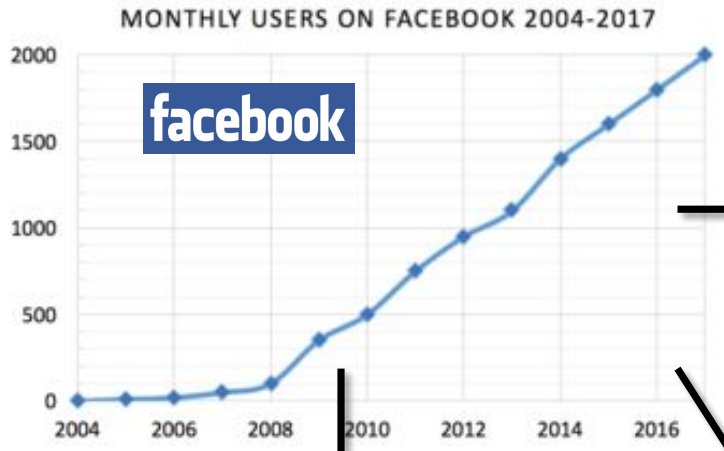
From dailymail.co.uk

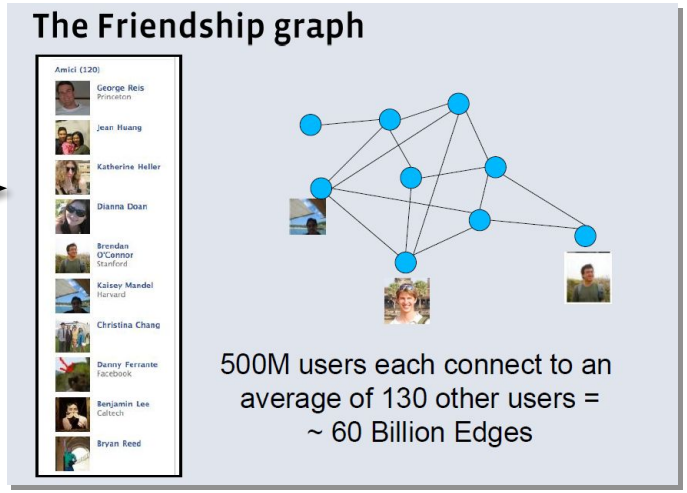
From www.formaspace.com

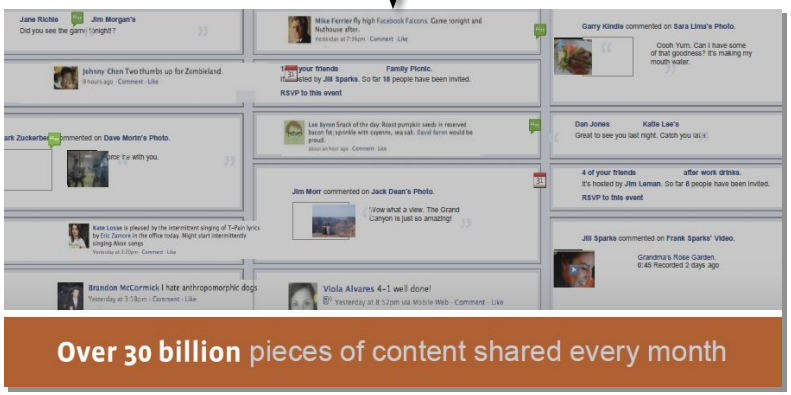


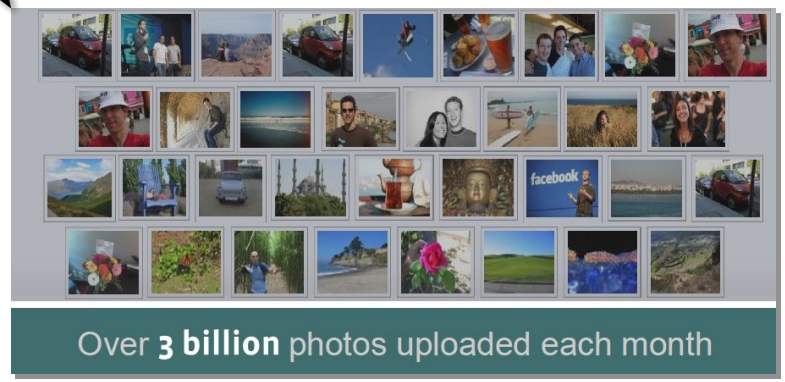
From linkedin.com

How does Facebook predict what content to show you?









Graphics from Lars Backstrom, ESWC 2011

How do companies decide what ads to show you?

?

?

?

U.S. INTERNATIONAL 中文网

The New York Times

Tuesday, March 4, 2014 | Today's Paper | Personalize Your Weather

WORLD U.S. NEW YORK BUSINESS OPINION SPORTS SCIENCE ARTS FASHION & STYLE VIDEO All Sections

TURMOIL IN UKRAINE

Putin, Flashing Disdain, Defends Action in Crimea

By STEVEN LEE MYERS
59 minutes ago

President Vladimir V. Putin's first public remarks on the political upheaval in Ukraine were aimed at both international and domestic audiences, defending Russia from the fury of global criticism and rallying support at home.

NEWS ANALYSIS

No Easy Way Out of Ukraine Crisis

By PETER BAKER 54 minutes ago

White House officials are weighing their options, knowing that reversing the occupation of Crimea would be difficult, if not impossible, in the short run.

Uriel Sinai for The New York Times

Ukrainian riot police officers stood guard at an anti-Russian rally in Donetsk on Tuesday.

Crimea's Pro-Russian Leader Says Region Is Secure

By DAVID M. HERSZENHORN 8:21 PM ET

The prime minister of the autonomous region offered the assurance on Tuesday even as armed standoffs continued.

RELATED COVERAGE

- **Kerry Takes Offer of Aid to Ukraine** 33 minutes ago
- **Cyberattacks Rise as Crisis Spills to Internet** 6:47 PM ET
- VIDEO: **Confrontation in Crimea**

The Opinion Pages

OP-ED CONTRIBUTOR

Has Privacy Become a Luxury Good?

By JULIA ANGWIN

It takes a lot of money and time to avoid hackers and data miners.

- **Editorial: Frustration With Afghanistan**
- **Brooks: Putin Can't Stop**
- **Cohen: Russia's Crimean Crime**

DRAFT

My Character to Kill

By ALEX BERENSON

I'm not sure I can say goodbye to a man who has defined my creative life for so long — and who will pay the mortgage for at least one more contract.

- Op-Docs: **'Chinese, on the Inside'**

MARKETS » At 10:03 PM ET

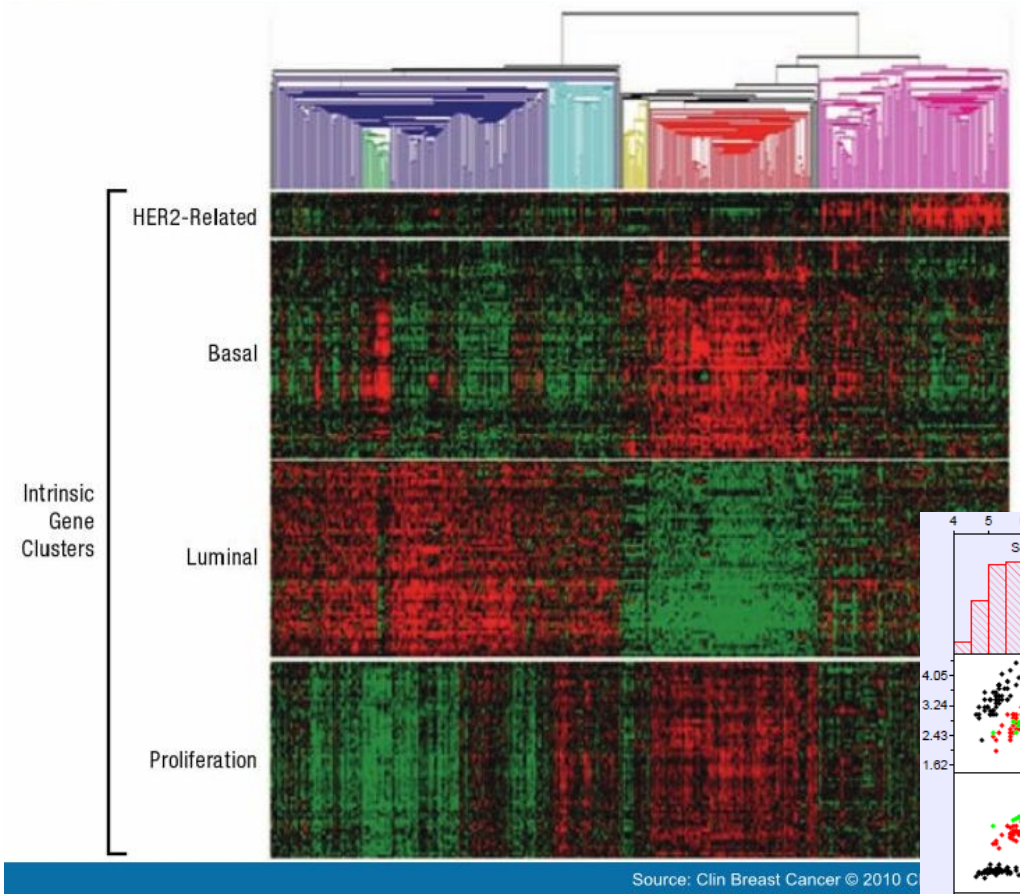
JAPAN Nikkei		HangSeng		CHINA Shanghai	
14,942.78		22,690.46		2,059.39	
+221.30		+32.83		−12.09	
+1.50%		+0.14%		−0.58%	

Data delayed at least 15 minutes

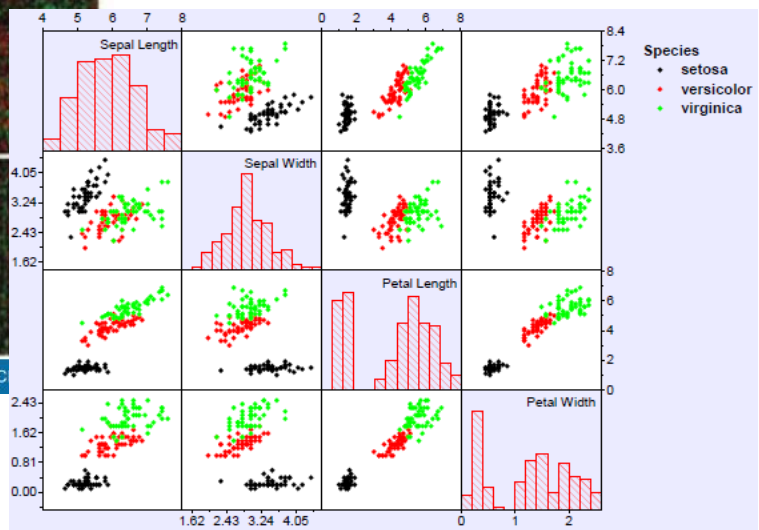
Get Quotes | My Portfolios »

?

How can we make personalized recommendations in medicine?



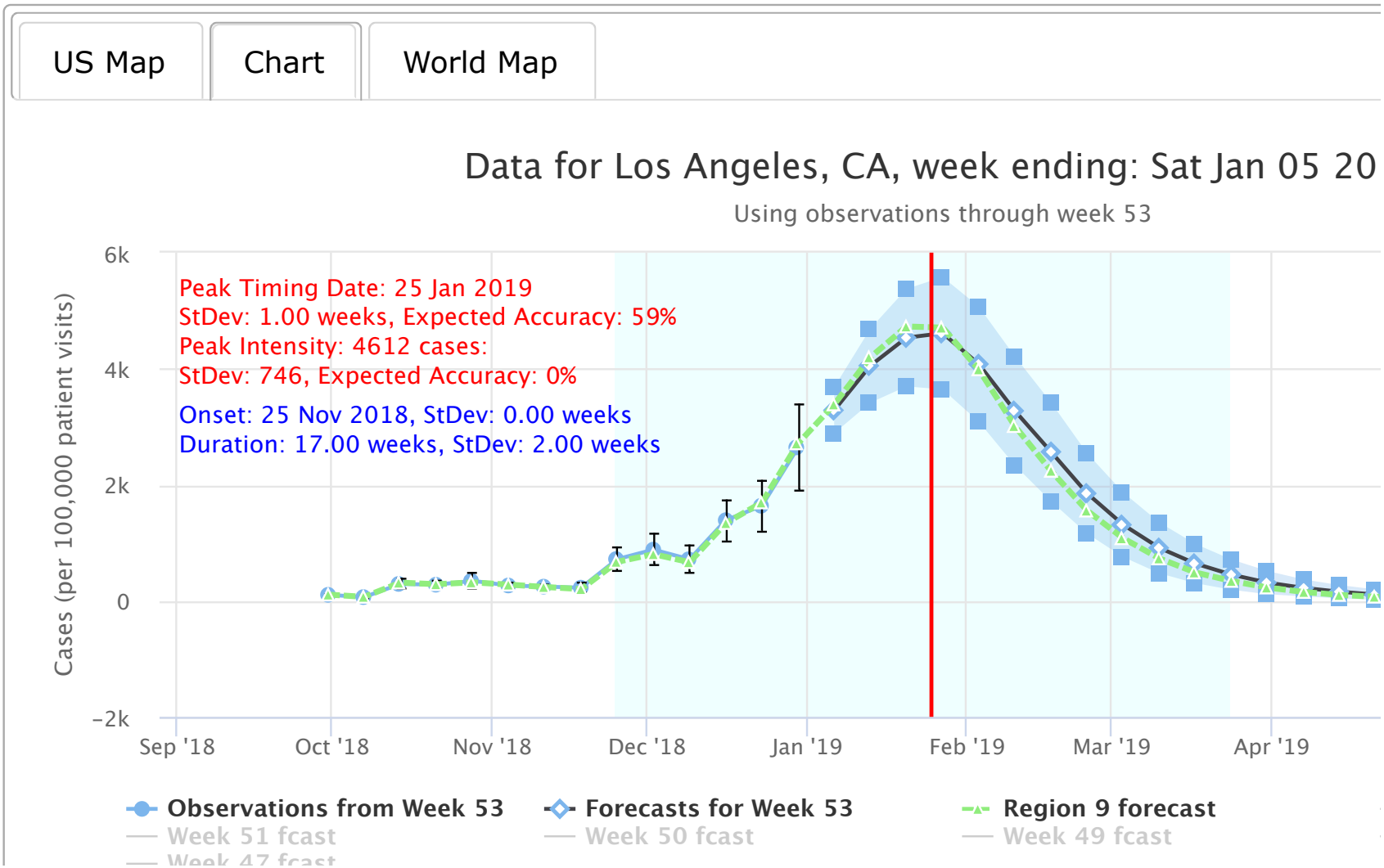
Data Matrix:
Rows = genes
Columns = patients



From www.originlab.com

How do public health workers predict infectious disease outbreaks?

Influenza Observations and Forecast

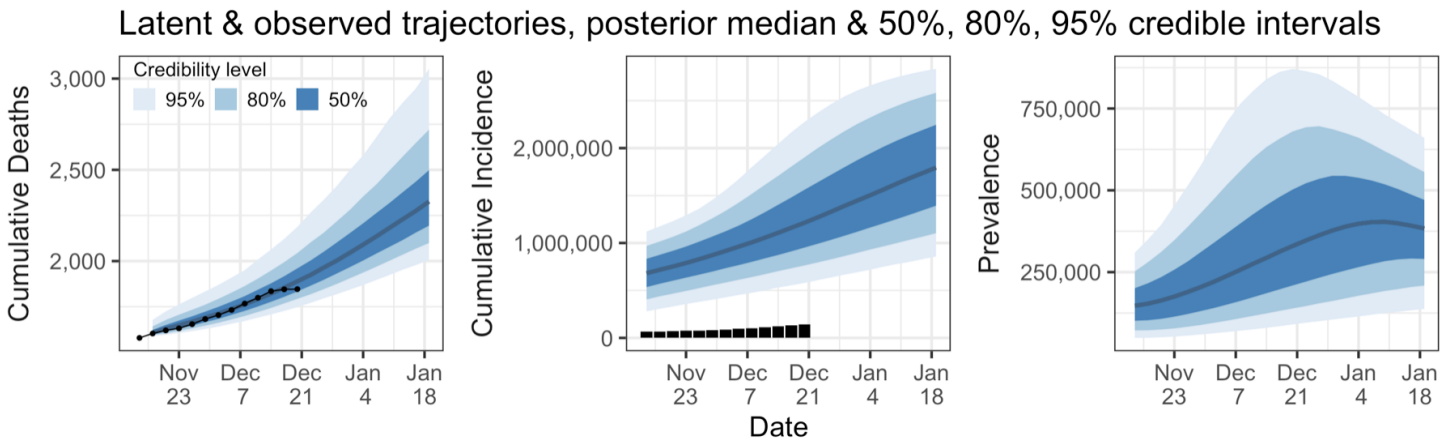


https://cpid.iri.columbia.edu

Orange County, CA COVID-19 Situation Report, December 28, 2020

Report period: Nov 15 - Dec 20 (we don't use the most recent data due to reporting delays)

The goal of this report is to inform interested parties about dynamics of SARS-CoV-2 spread in Orange County, CA and to predict epidemic trajectories. Methodological details are provided below and in the accompanying manuscript. We are also contributing to COVID Trends by UC Irvine project that provides data visualizations of California County trends across time and space.



https://www.stat.uci.edu/oc_covid_model/

Data visualization: why visualize and explore?

- People are good at pattern recognition
  - At spotting clusters, trends, outliers, structure, etc. that computers many miss
- Usually two types of users
  1. The data scientist who wants to explore/analyze/understand
     - For the data scientist, visualization and exploration are part of an iterative process
  2. The person who needs a quick summary to make a decision
     - For the consumer we want to communicate information quickly and clearly
     - e.g., for a medical doctor, for a policy-maker, for a company executive
- For data scientists...its always a good idea to look at your data
  - Helps to understand where the semantics of the data...what the measurements actually mean

What is exploratory data analysis?

- EDA is broader than just visualization
- EDA = {visualization, clustering, dimension reduction,…}
- For small numbers of variables, EDA = visualization
- For large numbers of variables, we need to be cleverer
  - Clustering, dimension reduction, embedding algorithms
  - These are techniques that essentially reduce high-dimensional data to something we can look at
- Pioneered by John Tukey (statistician at Bell Labs, Princeton) in the 1960's
  - "let the data speak"

Questions?