A Brief Overview of Data Mining

Ashok Samal
10/14/2021

New World!

- Largest retailer does not have an inventory
  - Alibaba
- Largest hotel chain does not own a hotel
  - Airbnb
- Largest media company does not generate any original content
  - Facebook
- Largest taxi company does not own a single car
  - Uber

New World!

2008: 105B

2004: 917B

2009: 87B

1999: 446B

1892: 234 B

1903: 60 B

1902: 102 B

1955: 184 B

Year Founded: October 2021 Market Cap

POPULAR SCIENCE

THE FUTURE NOW

THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate P.52

OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins? P.38

NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics P.69

PLUS

Juan Enriquez Reprograms Life P.31

James Gleick Unsplits the Bit P.58

AND Lawrence Weschler Questions the Cloud P.76

SPECIAL ISSUE

DATA IS POWER

HOW INFORMATION IS DRIVING THE FUTURE

November 2011

Data Mining: Example (myth?)

- What products are sold together with diapers in a grocery store/supermarket?
  - Answer: Beer
- Highest volume on Friday afternoons
  - By men between the ages of 25 and 35.
- What did the supermarket do as a consequence?
  - They put the beer display next to the diapers.
- Beer sales skyrocketed.

Data Mining: Example

What is the most bought item after hurricanes?

Large-scale Data is Everywhere!

- Enormous data growth in both commercial and scientific databases
  - Advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.
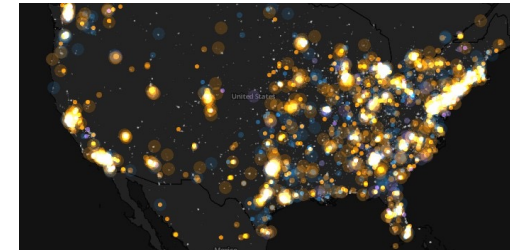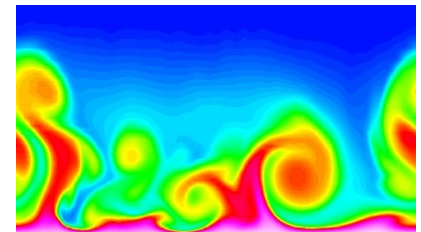
Cyber Security

E-Commerce

Traffic Patterns

Social Networking: Twitter

Sensor Networks

Computational Simulations

Spatial Data

- Geographic information is any item that is **georeferenced**
  - Atomic form

    <location, time, property>

  - Also called **geospatial** information
  - May be further augmented with images, audio or video

- Geographic information
  - Traditionally created by government authorities
    - USGS, NGA, military in many countries, state and local governments
  - Recently volunteers have become significant contributors

How much data?

- Every day, we create 2.5 quintillion (10^{18}) bytes of data (1 Exabyte)

- 90% of the data in the world today has been created in the last two years alone.

- 463 exabytes of data will be generated each day by humans as of 2025.

How much data?

SI decimal prefixes		Binary usage
Name (Symbol)	Value	
Kilobyte (KB)	10^3	2^{10}
Megabyte (MB)	10^6	2^{20}
Gigabyte (GB)	10^9	2^{30}
Terabyte (TB)	10^{12}	2^{40}
Petabyte (PB)	10^{15}	2^{50}
Exabyte (EB)	10^{18}	2^{60}
Zettabyte (ZB)	10^{21}	2^{70}
Yottabyte (YB)	10^{24}	2^{80}

How much data?

- YouTube
  - July 2011 - 48 hours of video uploads/minute
  - 1 hr of video = 80GBytes (640 x 480 x 30fps x 8bpp)
  - With 10:1 compression ratio = 8Gbytes
  - 2014: 300 hours/min
  - 2019: 500 hours/min
- More video is uploaded to YouTube in 60 days than the 3 major US networks created in 70 years.
- 250 million hours of videos watched per day on TV

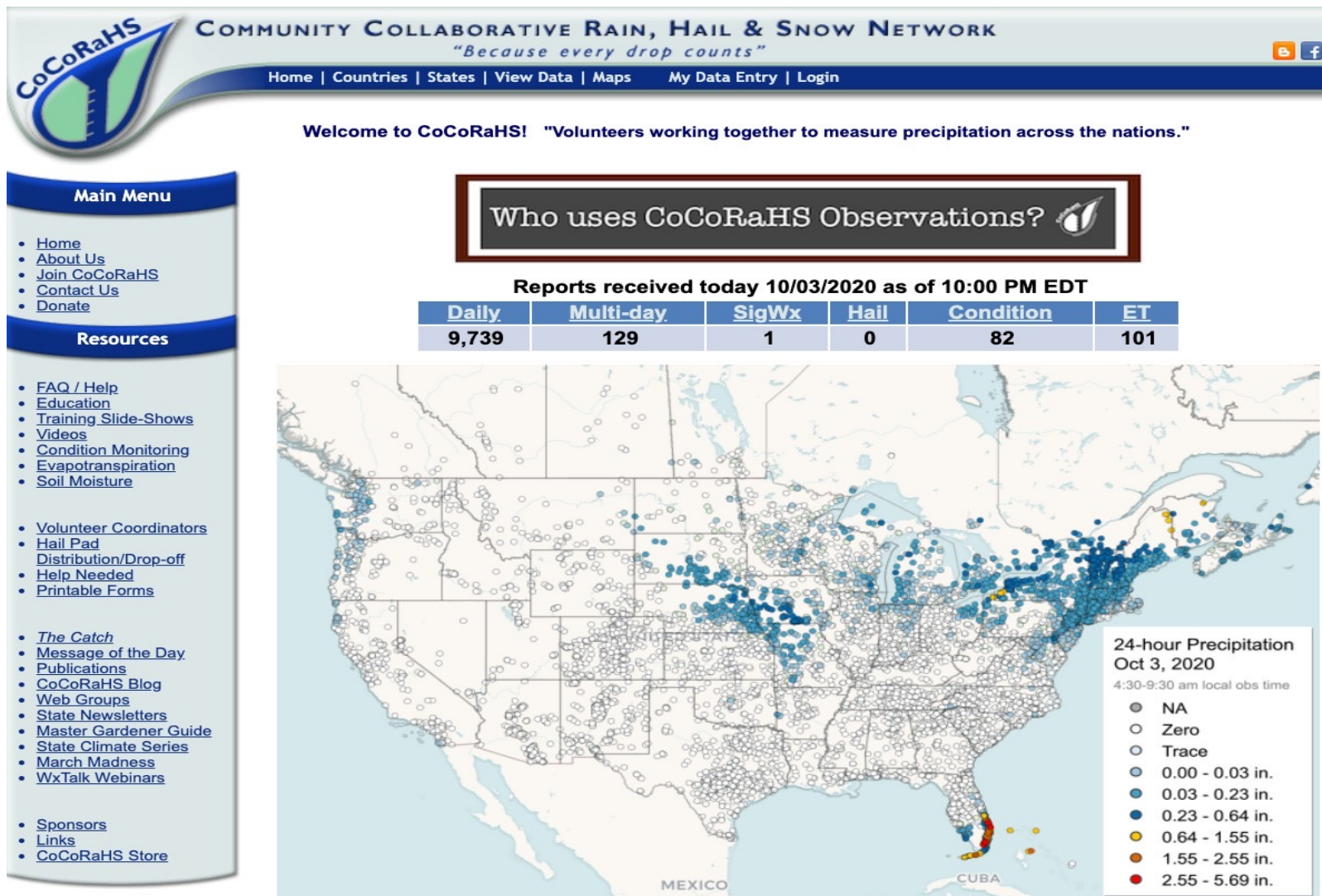YouTube Stats

How much data?

- Facebook
  - Over 2.5 billion(monthly) active users (1.6 billion daily users)
  - 350 million photos are uploaded per day (2019) - 4,000 per second.
- Twitter
  - 330 million monthly active users (145 million daily users)
  - 500 Million tweets per day (2019)
  - 6000 tweets per second (2019)
- Flickr
  - Over 10 Billion images
  - Up to 25 Million added per day (high traffic day)
  - 90 million monthly users
- Digital Images
  - 1 trillion photos taken in 2018
  - Over 6 billion smart phones by 2020 (2.6 Billion in 2015)
  - 1.4 Trillion pictures will be taken in 2020

Social Media Usage Link

Machine-to-Machine Data

- Self-Driving Cars
  - 3 PBytes per car per year
- Sensors
  - 1Trillion sensors on the Internet by 2020
  - Songdo (South Korea) Smart City
- Smart "things"
  - Windows, homes, hotels
  - Bridges
  - Tractors
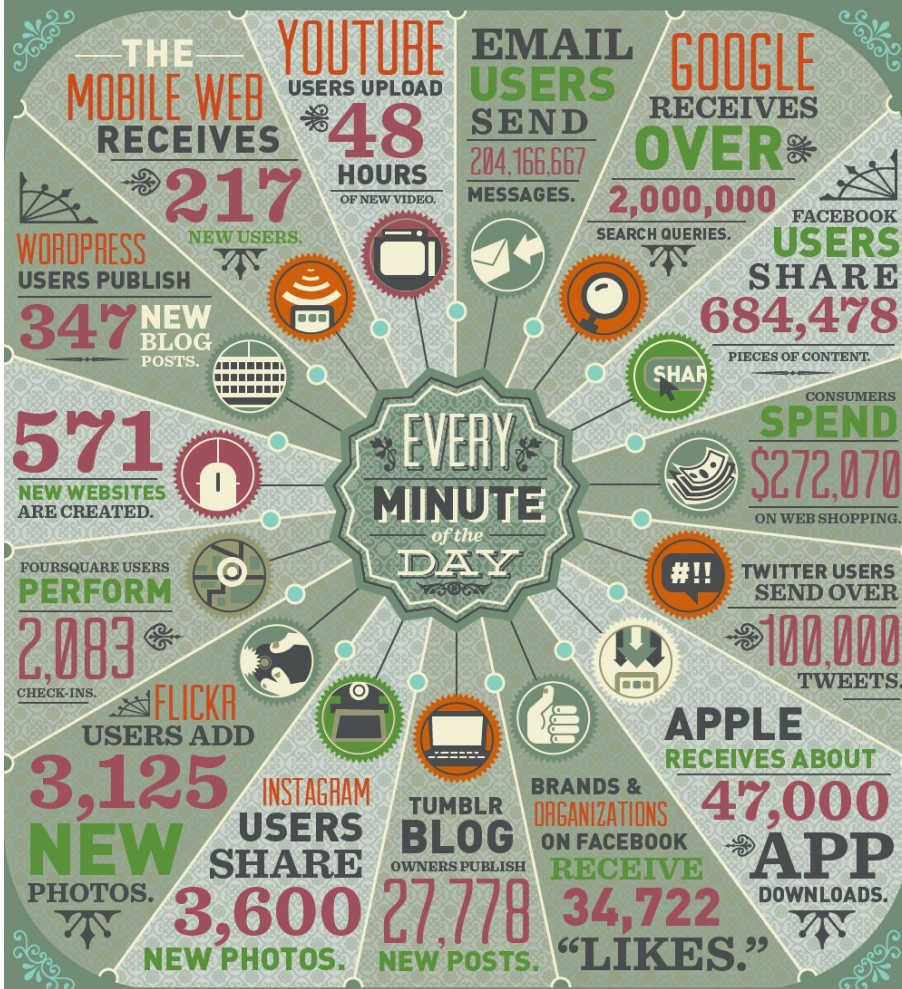 - TV

Volunteered Geoinformatics Example: www.cocorahs.org

COMMUNITY COLLABORATIVE RAIN, HAIL & SNOW NETWORK
"Because every drop counts"

Home | Countries | States | View Data | Maps | My Data Entry | Login

Welcome to CoCoRaHS! "Volunteers working together to measure precipitation across the nations."

Main Menu

- Home
- About Us
- Join CoCoRaHS
- Contact Us
- Donate

Resources

- FAQ / Help
- Education
- Training Slide-Shows
- Videos
- Condition Monitoring
- Evapotranspiration
- Soil Moisture

- Volunteer Coordinators
- Hail Pad Distribution/Drop-off
- Help Needed
- Printable Forms

- *The Catch*
- Message of the Day
- Publications
- CoCoRaHS Blog
- Web Groups
- State Newsletters
- Master Gardener Guide
- State Climate Series
- March Madness
- WxTalk Webinars

- Sponsors
- Links
- CoCoRaHS Store

Who uses CoCoRaHS Observations?

Reports received today 10/03/2020 as of 10:00 PM EDT

Daily	Multi-day	SigWx	Hail	Condition	ET
9,739	129	1	0	82	101

24-hour Precipitation
Oct 3, 2020
4:30-9:30 am local obs time

- NA
- Zero
- Trace
- 0.00 - 0.03 in.
- 0.03 - 0.23 in.
- 0.23 - 0.64 in.
- 0.64 - 1.55 in.
- 1.55 - 2.55 in.
- 2.55 - 5.69 in.

MEXICO   CUBA

DATA NEVER SLEEPS

DOMO

How Much Data Is Generated Every Minute?

EVERY MINUTE of the DAY

THE MOBILE WEB RECEIVES 217 NEW USERS.

YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.

EMAIL USERS SEND 204,166,667 MESSAGES.

GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.

WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.

FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.

571 NEW WEBSITES ARE CREATED.

CONSUMERS SPEND $272,070 ON WEB SHOPPING.

FOURSQUARE USERS PERFORM 2,083 CHECK-INS.

TWITTER USERS SEND OVER 100,000 TWEETS.

FLICKR USERS ADD 3,125 NEW PHOTOS.

APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.

INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.

TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.

BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."

WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING

These are just some of the more common ways that Internet users add to the big data pool. In truth, depending on the niche of business you're in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:

The global Internet population grew 6.59 percent from 2010 to 2011 and now represents **2.1 BILLION PEOPLE.**

These users are real, and they are out there leaving data trails everywhere they go. The team at Domo can help you make sense of this seemingly insurmountable heap of data, with solutions that help executives and managers bring all of their critical information together in one intuitive interface, and then use that insight to transform the way they run their business. To learn more, visit www.domo.com.
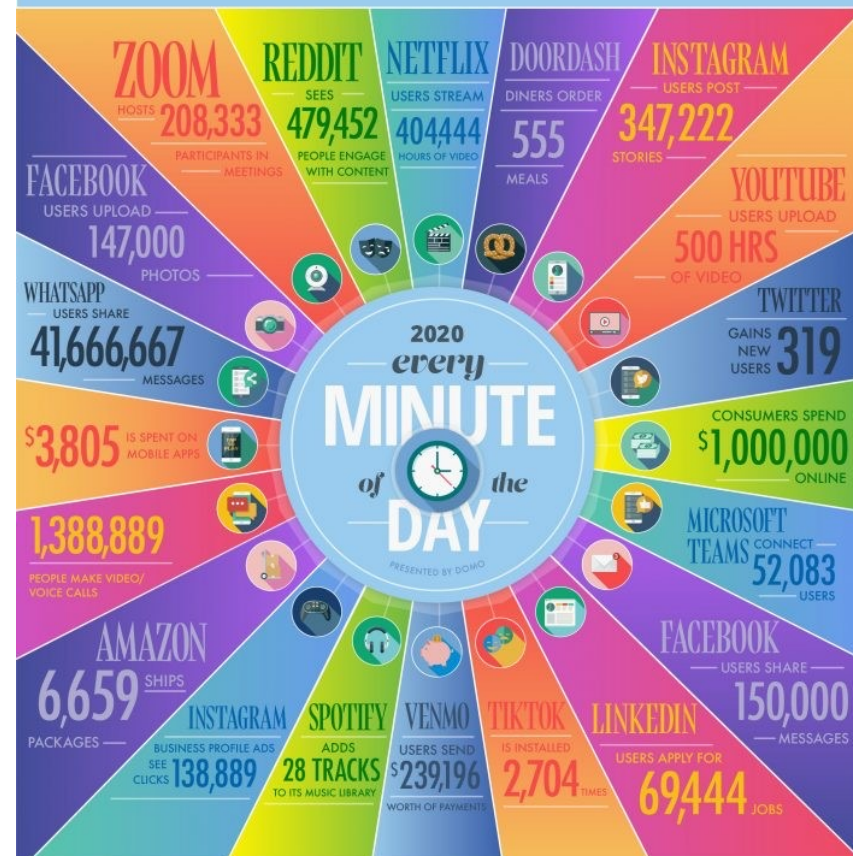
DOMO

DATA NEVER SLEEPS 8.0

DOMO

How much data is generated every minute?

In 2020, the world changed fundamentally—and so did the data that makes the world go round. As COVID-19 swept the globe, nearly every aspect of life—from work to working out—moved online, and people depended more and more on apps and the internet to socialize, educate and entertain ourselves. Before quarantine, just 15% of Americans worked from home. Now over half do. And that's not the only big shift. In our 8th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute—a trend that shows no sign of stopping.

2020 every MINUTE of the DAY
PRESENTED BY DOMO

ZOOM HOSTS 208,333 PARTICIPANTS IN MEETINGS

REDDIT SEES 479,452 PEOPLE ENGAGE WITH CONTENT

NETFLIX USERS STREAM 404,444 HOURS OF VIDEO

DOORDASH DINERS ORDER 555 MEALS

INSTAGRAM USERS POST 347,222 STORIES

FACEBOOK USERS UPLOAD 147,000 PHOTOS

YOUTUBE USERS UPLOAD 500 HRS OF VIDEO

WHATSAPP USERS SHARE 41,666,667 MESSAGES

TWITTER GAINS NEW USERS 319

$3,805 IS SPENT ON MOBILE APPS

1,388,889 PEOPLE MAKE VIDEO/ VOICE CALLS

CONSUMERS SPEND $1,000,000 ONLINE

MICROSOFT TEAMS CONNECT 52,083 USERS

AMAZON SHIPS 6,659 PACKAGES

INSTAGRAM BUSINESS PROFILE ADS SEE CLICKS 138,889

SPOTIFY ADDS 28 TRACKS TO ITS MUSIC LIBRARY

VENMO USERS SEND $239,196 WORTH OF PAYMENTS

TIKTOK IS INSTALLED 2,704 TIMES

LINKEDIN USERS APPLY FOR 69,444 JOBS

FACEBOOK USERS SHARE 150,000 MESSAGES

The world's internet population is growing significantly year over year: As of April 2020, the internet reaches 59% of the world's population and now represents 4.57 billion people — a 6% increase from January 2019.

As the world changes, businesses need to change with the times—and that requires data. Every click, swipe, share or like tells you something about your customers and what they want, and Domo is here to help your business make sense of all of it. Domo gives you the power to make data-driven decisions at any moment, on any device, so you can make smart choices in a rapidly changing world.

GLOBAL INTERNET POPULATION GROWTH 2014–2020 (IN BILLIONS)

2014	2016	2018	2020
3.0	3.4	4.3	4.5

Learn more at domo.com

DOMO
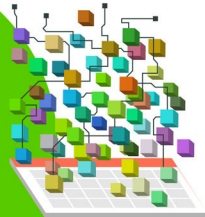
The FOUR V's of Big Data

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

2005
2020

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
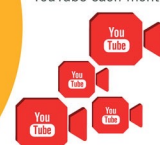
400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

CHARACTERISTICS (V'S) OF BIG DATA

GLOBAL INTERNET TRAFFIC IN 2013 WAS APPROXIMATELY 5,000,000,000,000,000,000,000 BYTES.
(5) BILLION GIGABYTES.
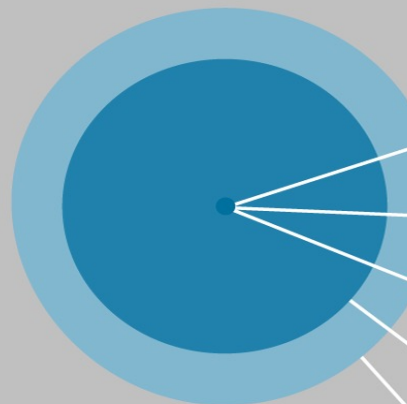
1992
100 GB/DAY

1997
100 GB/HOUR

2002
100 GB/SECOND

2013
28,875 GB/SECOND

2018
50,000 GB/SECOND

Global internet population
GREW 14.3% BETWEEN 2011 & 2013

3 BILLION
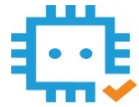The number of people who have access to the internet today equals that of the world's population in 1960

= 1960

Looking Ahead

- 163 Zettabytes of data generated per year by 2025 (IDC)

- Revenues for big data and business analytics (BDA) will grow from $130B billion in 2016 to $203B in 2020 (IDC)

The Data Scientist Shortage in 2020

Demand for Data Scientists

Job Listings

37%

Year on Year Growth in 2019

Job Ranking

#3

Ranking For Top Jobs in 2020

Salaries

14%

Average Salary Increase

Hiring

67%

Companies Expanding the Data Science Team

Why Is There Still a Shortage in 2020?

Artificial Intelligence

74%

Annual AI-related Hiring Growth 2015-2019

Big Data Gets Bigger

83%

Companies Investing in Big Data Projects

Tech Talent Shortage

85 MILLION

Global Tech Talent Shortage by 2030

Turnover

2 YEARS

Average Data Scientist Turnover

The Data Scientist Shortage
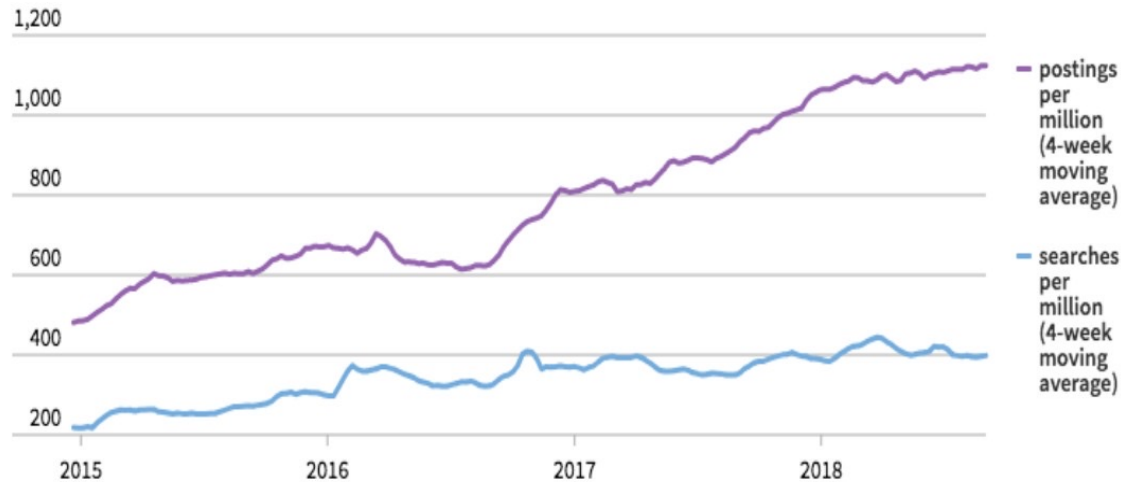
3X job postings versus job searches

- postings per million (4-week moving average)
- searches per million (4-week moving average)

250,000 / **2020 Shortage**

Difficulty Finding Security & Data Science Skillsets

Skillset	Percentage
Cybersecurity	39%
Data science/analytics	35%
AI/ML/RPA	31%
Cloud services/integration	18%
Legacy technologies	18%
DevOps/DevSecOps/Agile processes	17%
Internet of Things	17%
Cloud architecture	16%
Design thinking/UX	16%
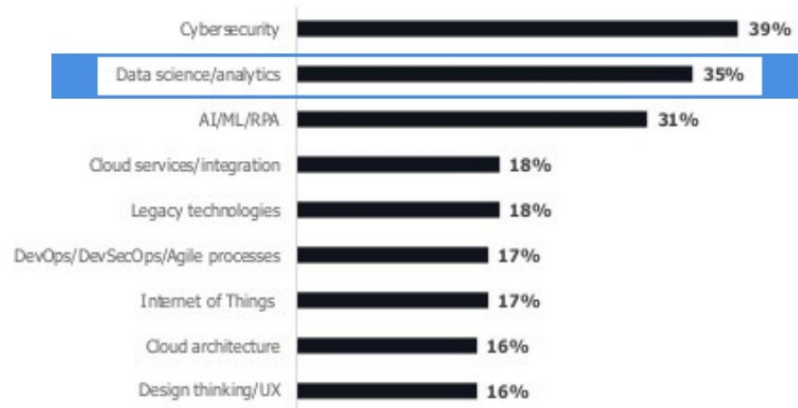
Q: In which technology-related areas do you anticipate your organization will have the most difficulty in finding appropriate skillsets?

Evolution of Sciences

- Before 1600: Empirical science
  - Gaining knowledge by observation
  - They are sometimes experimental

- 1600-1950s: Theoretical science
  - Each discipline grew a *theoretical* component.
  - Theoretical models often motivate experiments and generalize our understanding.

- 1950s-1990s: Computational science
  - In this period, most disciplines grew a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - It traditionally meant simulation.
  - It grew out of our inability to find closed-form solutions for complex mathematical models.
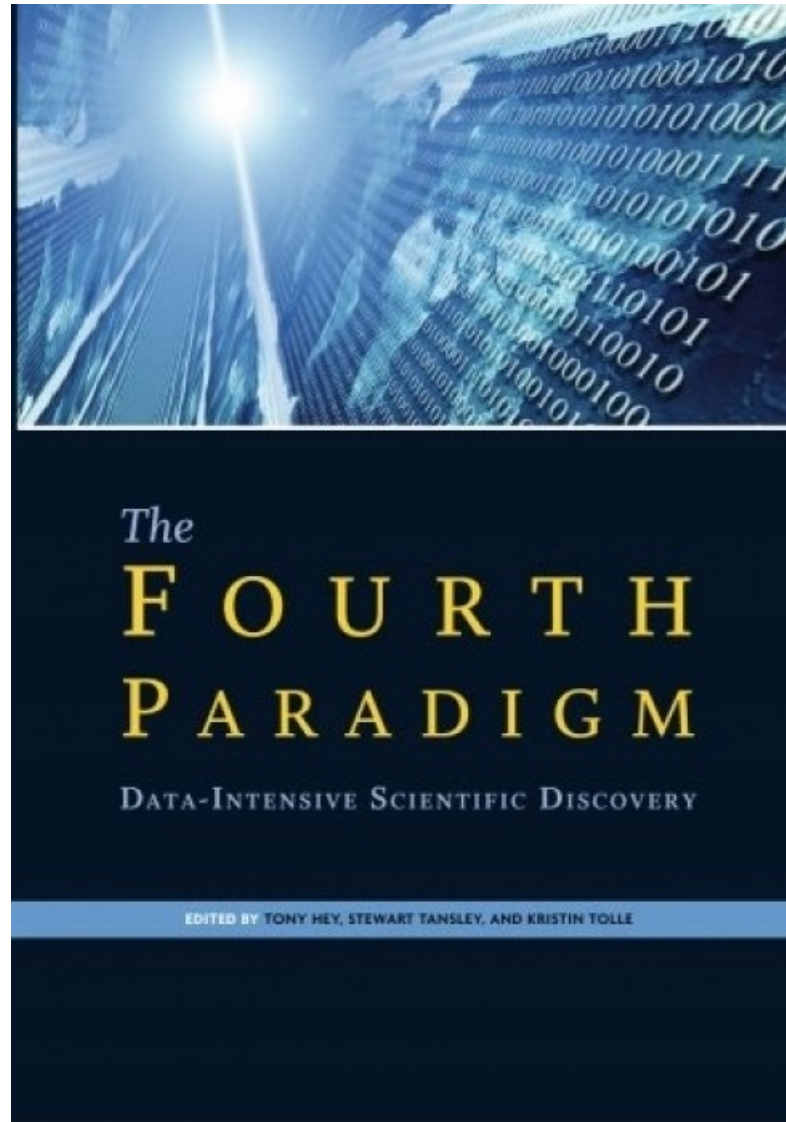
Unify experimental, theoretical and simulation approaches!

Evolution of Sciences

- 1990-now: Data science
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes.
  - X-info and Comp-X (e.g. bioinformatics, computational ecology)
  - Data exploration is the major new challenge.

What is Data Mining?

Why Not Traditional Data Analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

Input Data → **Data Processing** → **Data Mining** → **Post Processing** → Knowledge

Feature Selection
Dimensionality Reduction
Normalization
Data Subsetting

Filtering Patterns
Visualization
Pattern Interpretation
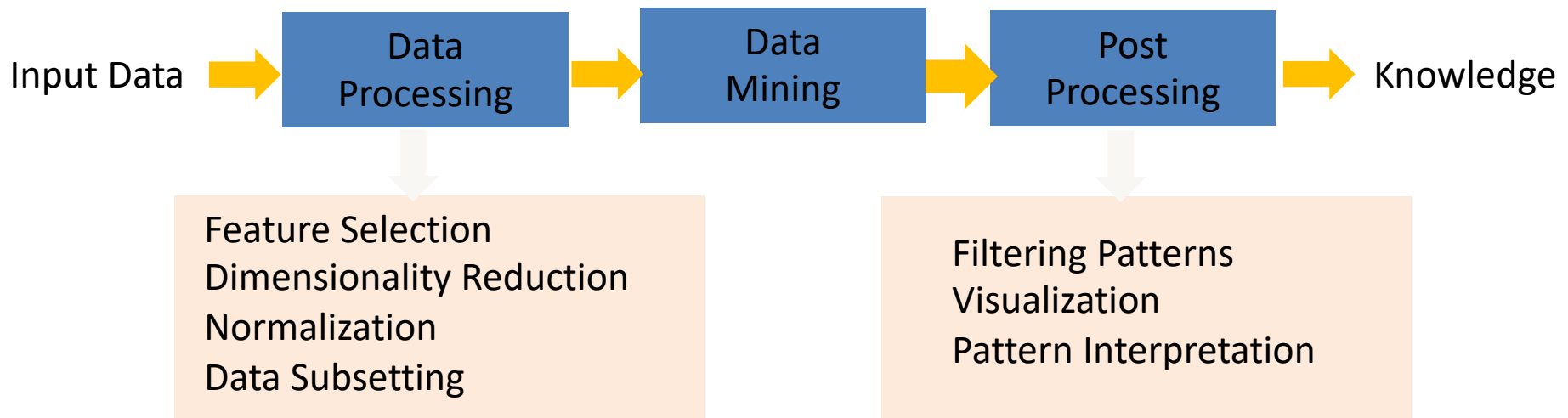
Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

What is data mining?

- **Novel**: previously unknown, not obvious
- **Valid**: broadly applicable (on new data) with some certainty

- **Meaningful**: humans should be able to understand
- **Useful**: should be possible to act on the result (actionable)

Meaningful Patterns

- A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless

- Statisticians call it Bonferroni's principle:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find meaningless patterns

What is (not) mining?

- ## What is NOT data mining?

    - Look up phone number in a phone directory

    - Query a web search engine for information about "Amazon"

- ## What is data mining?

    - Find certain names that are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)

    - Predict if a customer will consume over $100 in a store

Data Mining: A KDD Process

– Data mining: the core of knowledge discovery process.

Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Selection

Data Warehouse

Data Cleaning

Data Integration

Databases
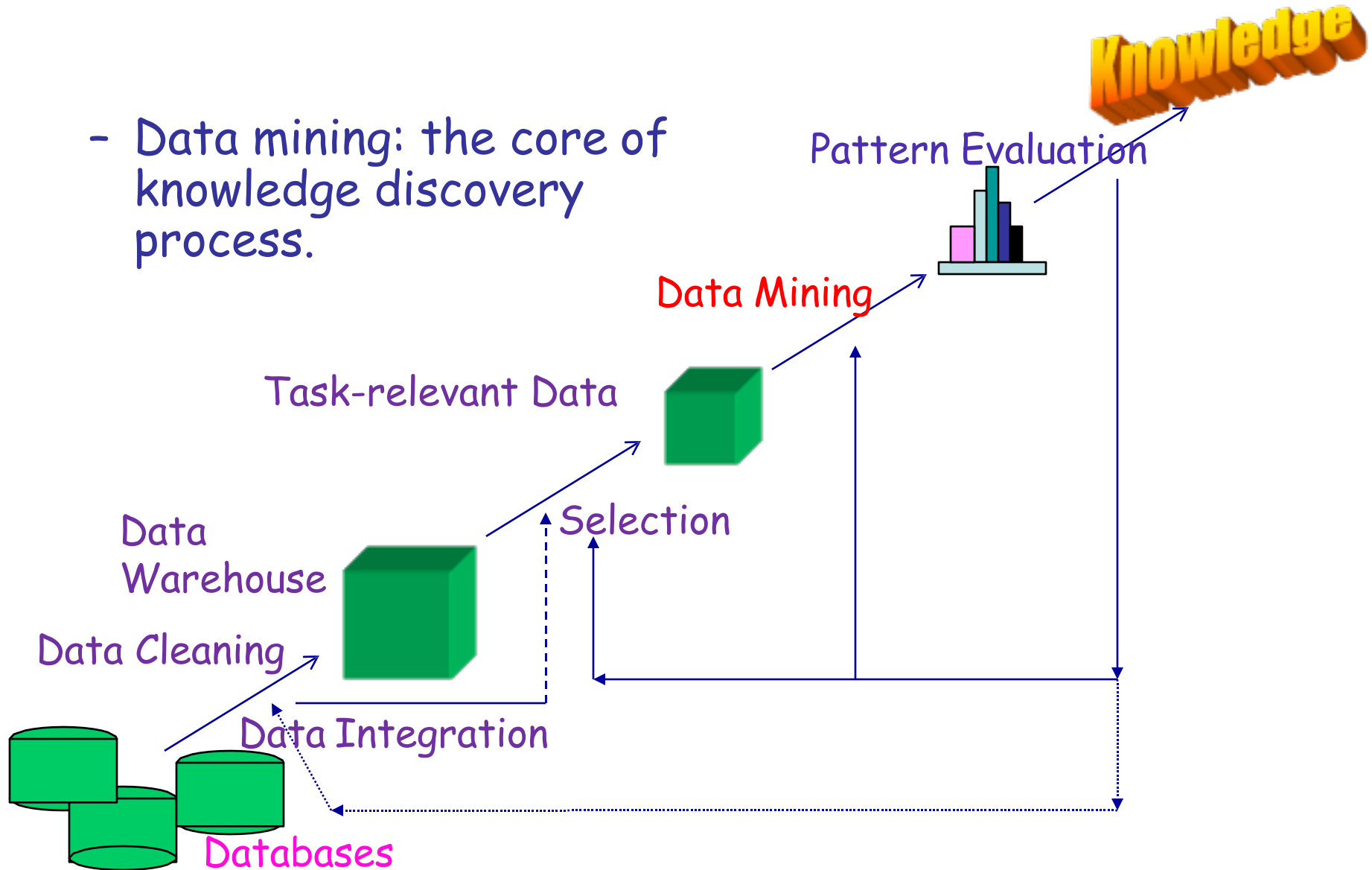
Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - Google has Petabytes of web data
    - Facebook has billions of active users
  - Purchases at department/grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit card transactions
- Computers have become cheaper and more powerful
- Competitive pressure is strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Data Mining? Scientific Viewpoint

- Data collected and stored at enormous speeds
  - Remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - Telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - Scientific simulations
    - terabytes of data generated in a few hours

- Data mining helps scientists
  - In automated analysis of massive datasets
  - In hypothesis formation
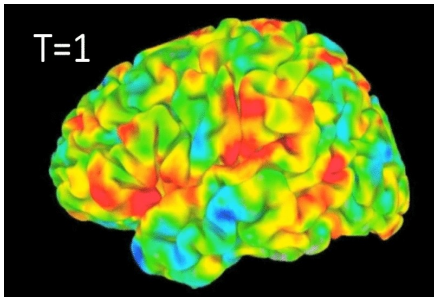
fMRI Data from Brain

Sky Survey Data
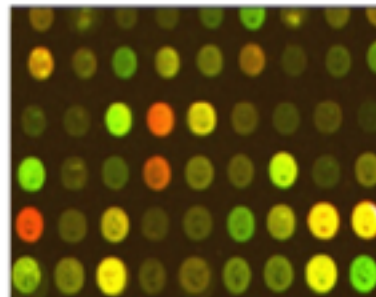
Gene Expression Data
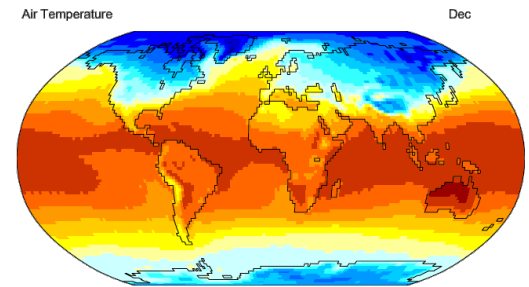
Surface Temperature of Earth

Great opportunities to improve productivity in all walks of life

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

Big data—a growing torrent

$600 to buy a disk drive that can store all of the world's music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. **5%** growth in global IT spending

235 terabytes data collected by the US Library of Congress in April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

Big data—capturing its value

$300 billion potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion potential annual value to Europe's public sector administration—more than GDP of Greece

$600 billion potential annual consumer surplus from using personal location data globally

60% potential increase in retailers' operating margins possible with big data

140,000–190,000 more deep analytical talent positions, and **1.5 million** more data-savvy managers needed to take full advantage of big data in the United States

Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs
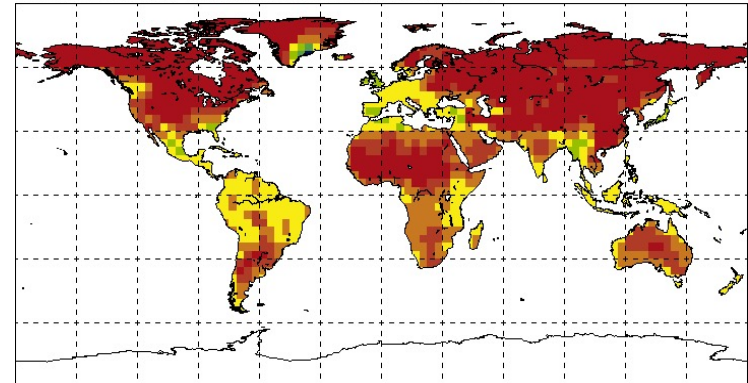
Predicting the impact of climate change

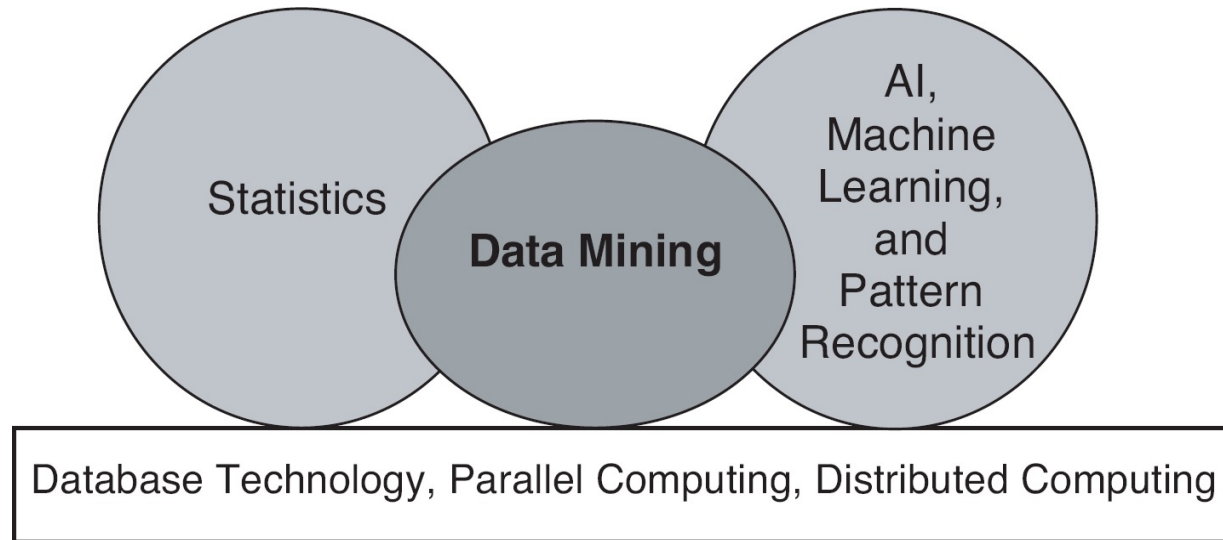Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed

- A key component of the emerging field of data science and data-driven discovery

Data Mining Tasks

- ## Prediction Methods

  - Use some variables to predict unknown or future values of other variables.

- ## Description Methods

  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks ...
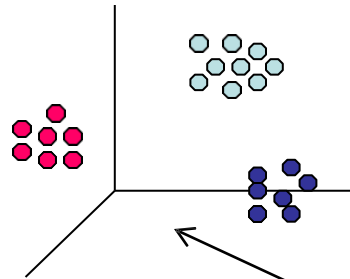
Clustering

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Predictive Modeling

Association Rules

Anomaly Detection

Milk →
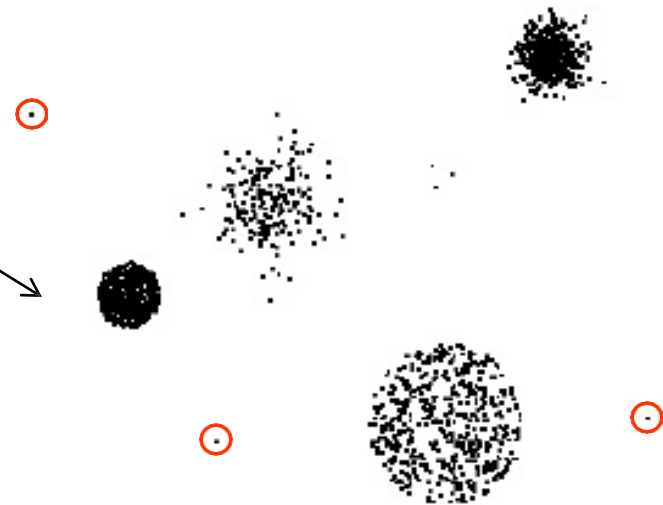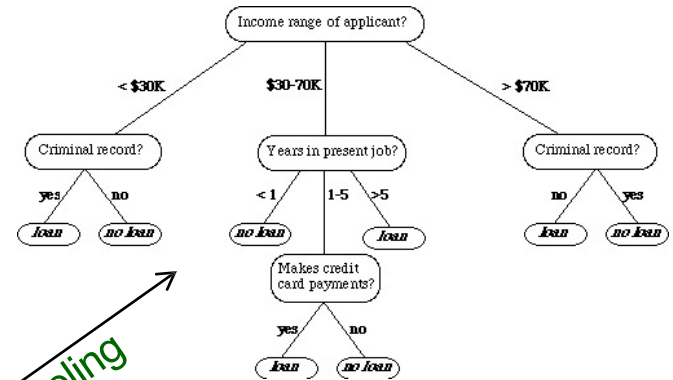
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness
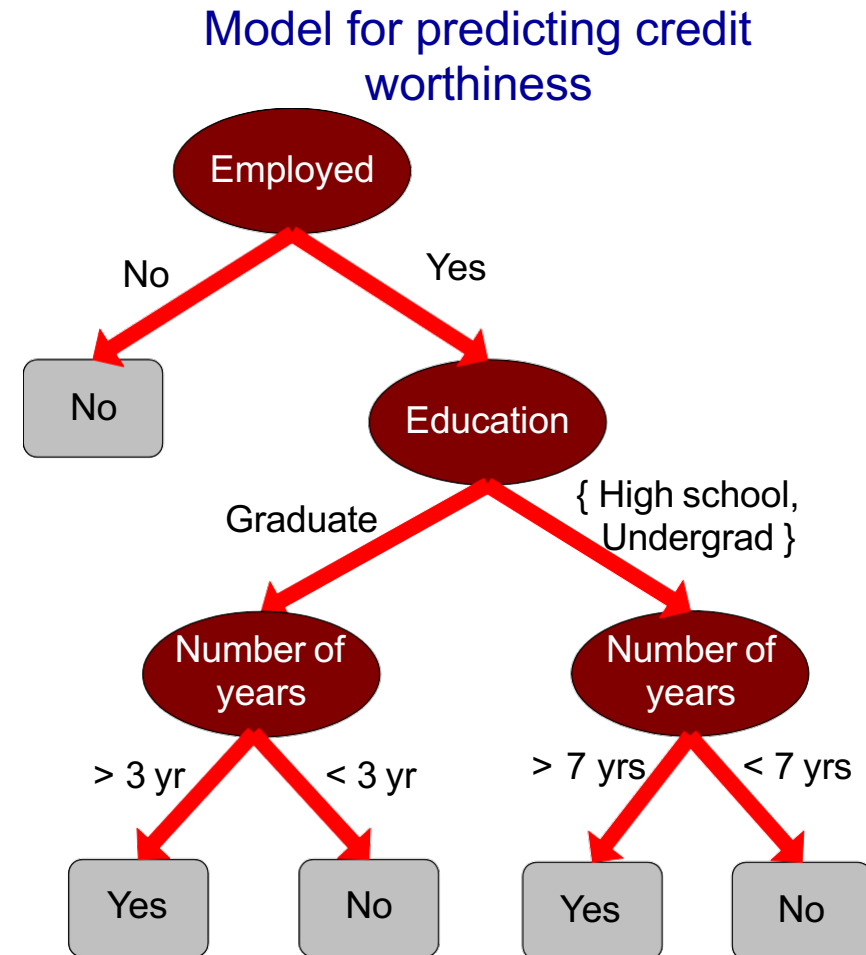
Classification Example

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

Test Set

Training Set → Learn Classifier → Model
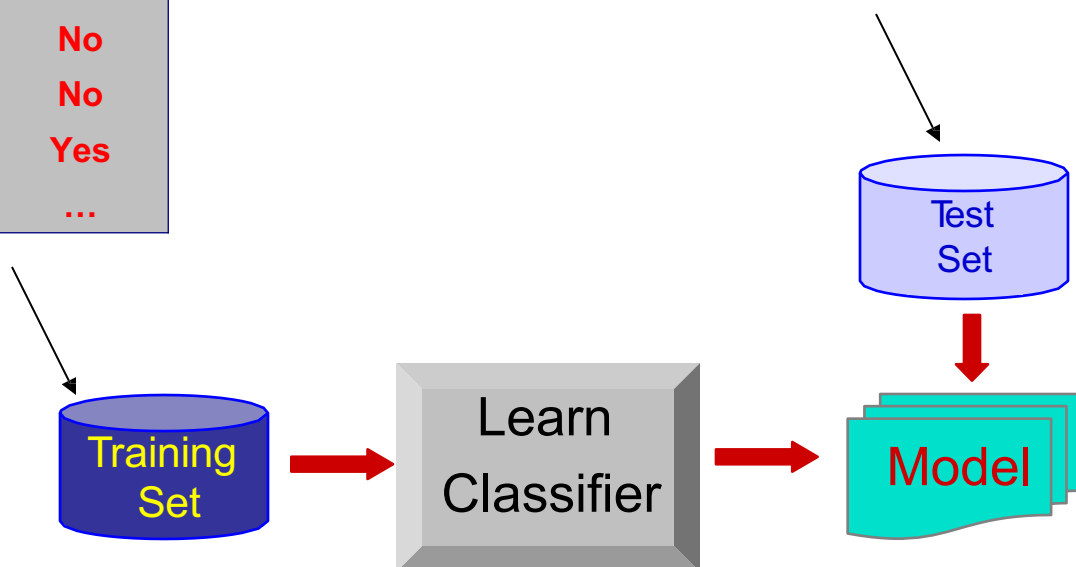
Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent

- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data

- Categorizing news stories as finance, weather, entertainment, sports, etc.

- Identifying intruders in the cyberspace

- Predicting tumor cells as benign or malignant

- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
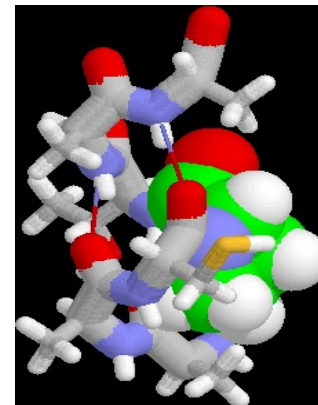
Classification: Application 1

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc.
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 2

- Churn prediction for telephone customers
  - **Goal:** To predict whether a customer is likely to be lost to a competitor.
  - **Approach:**
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Classification: Application 3

- Sky Survey Cataloging
  - **Goal**: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - **Approach:**
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classifying Galaxies

Early
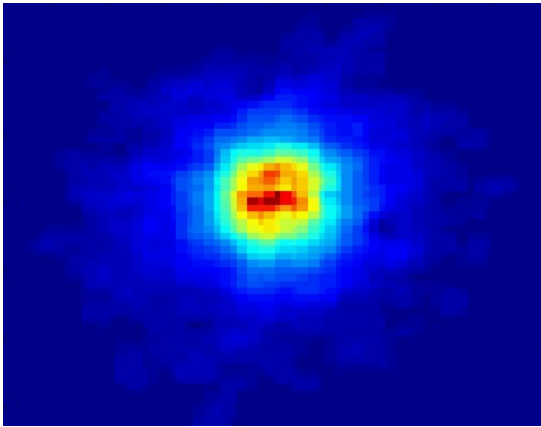
Intermediate
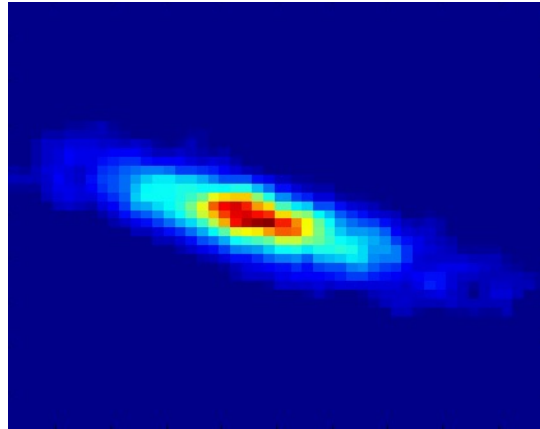
Late
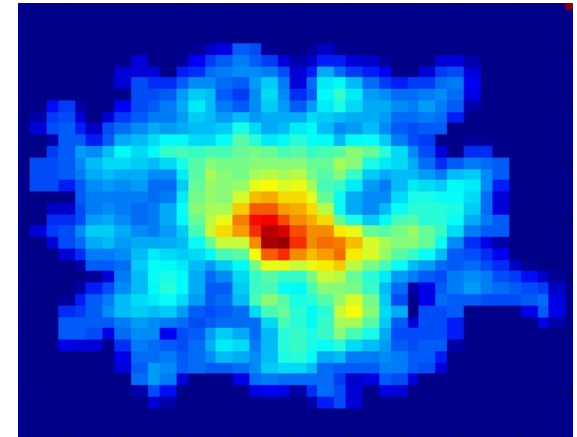
Class:
- Stages of Formation

Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.

- Extensively studied in statistics, neural network fields.

- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Spatial Predictive Models

- Location Prediction: Bird Habitat Prediction
  - Given training data
  - Predictive model building
  - Predict new data

Nest sites for 1995 Darr location
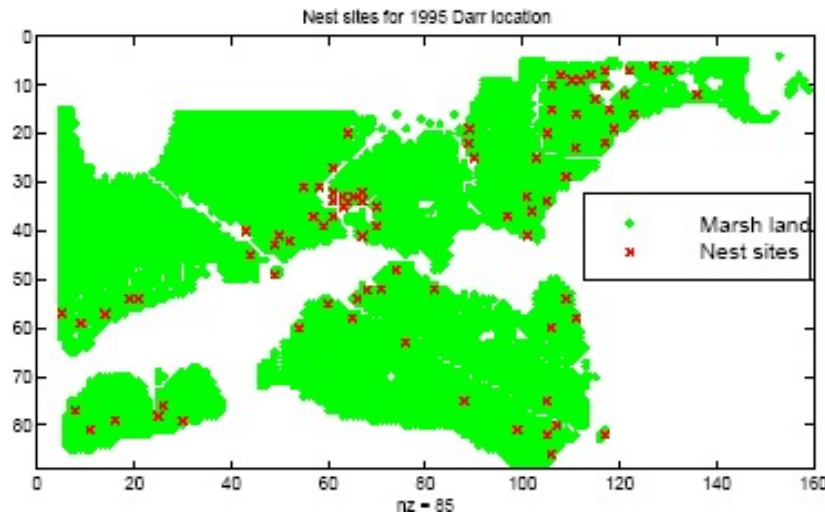
Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

Intra-cluster distances are minimized

Inter-cluster distances are maximized
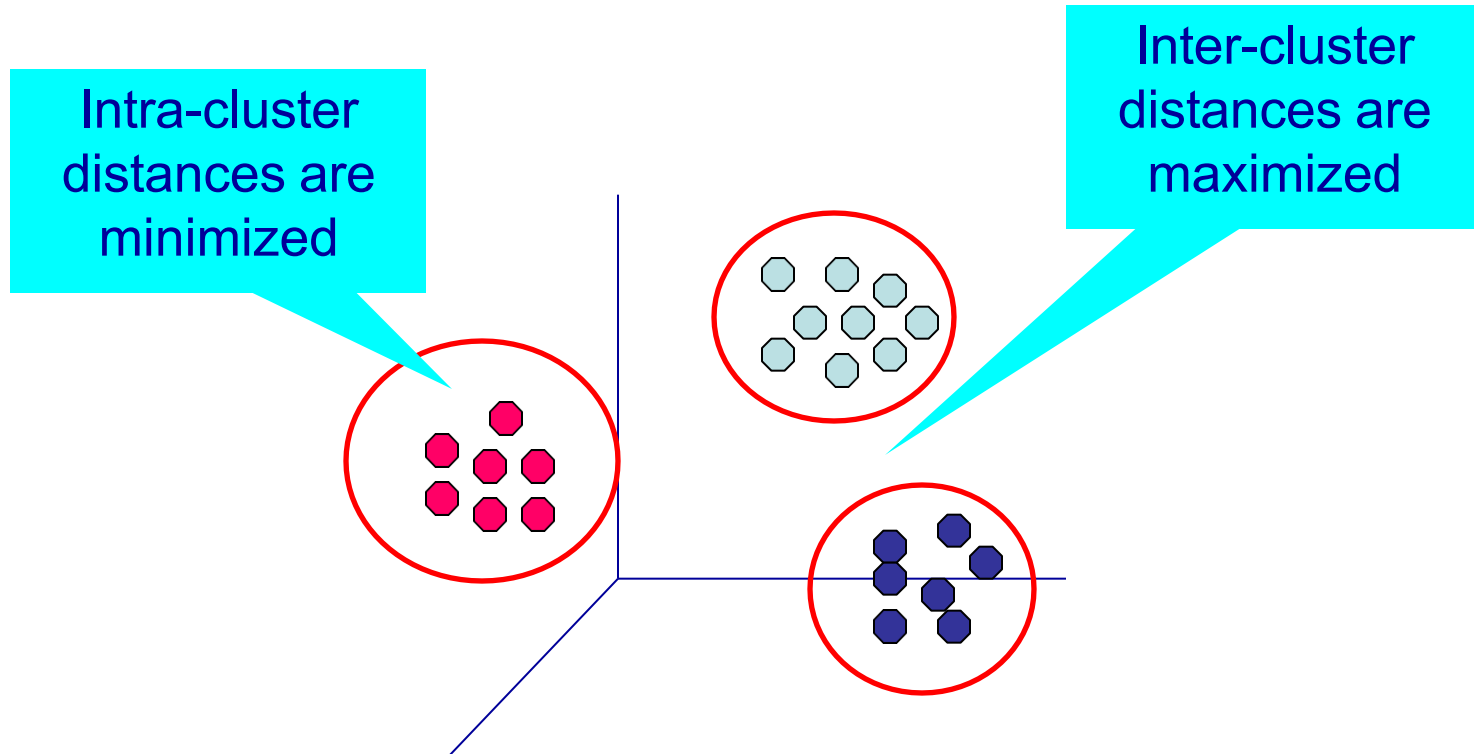
Applications of Cluster Analysis

- ## Understanding
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations

- ## Summarization
  - Reduce the size of large data sets
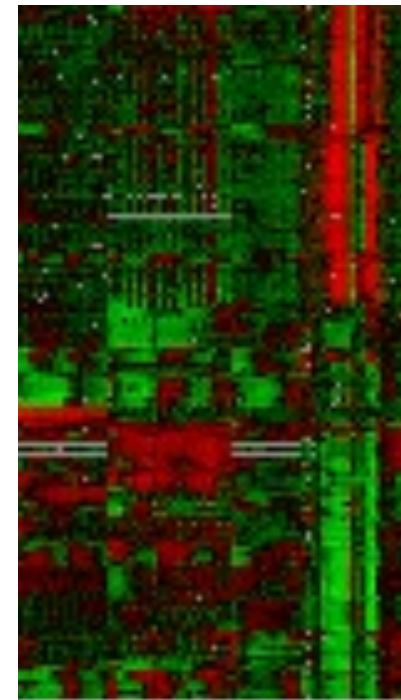
Courtesy: Michael Eisen

Clusters for Raw SST and Raw NPP
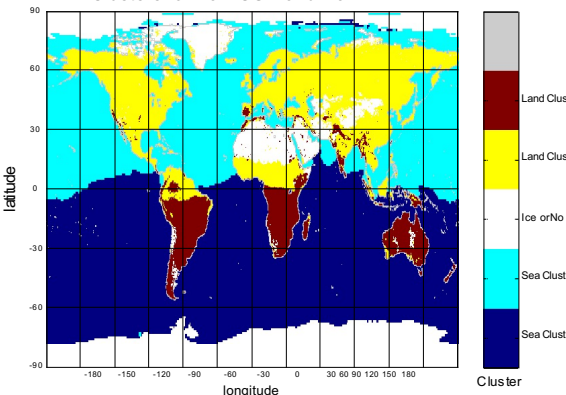
Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.

57
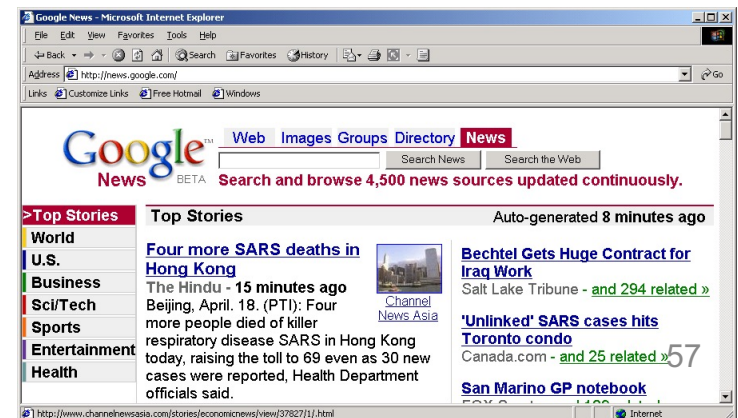
Clustering: Application 1

- Market Segmentation
  - **Goal:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - **Approach:**
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering

  - **Goal:** To find groups of documents that are similar to each other based on the important terms appearing in them.

  - **Approach:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
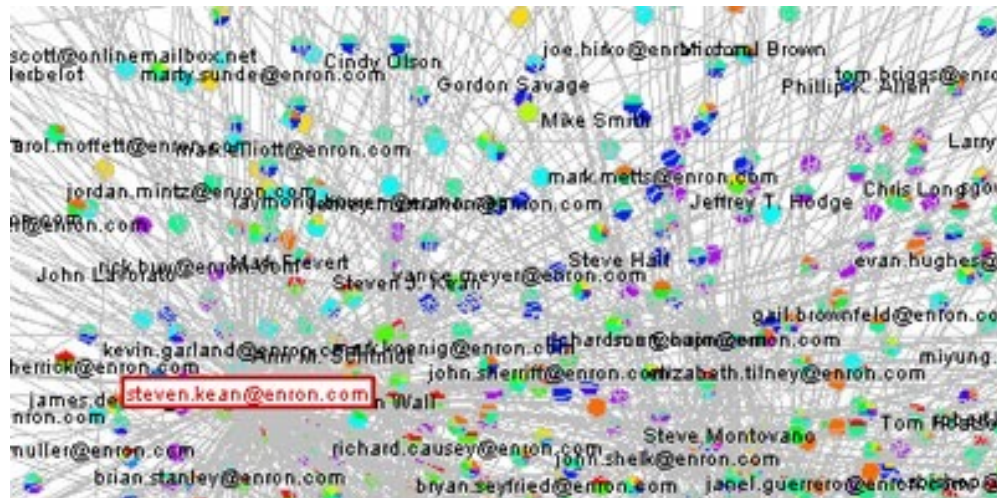
Enron email dataset

Spatial Clustering

- The 1854 Asiatic Cholera in London

Spatial Clustering

Flint Drinking Water Tests for Lead - January 2016

LEAD RANGE - PPB
- 0
- 1-4
- 5-14
- 15-49
- 50-149
- 150 AND ABOVE

Map generated by Michigan Radio's Mark Brush with help from Elias Brush, Cass Adair, and SmartyStreets

© OpenStreetMap contributors © CartoDB, CartoDB attribution
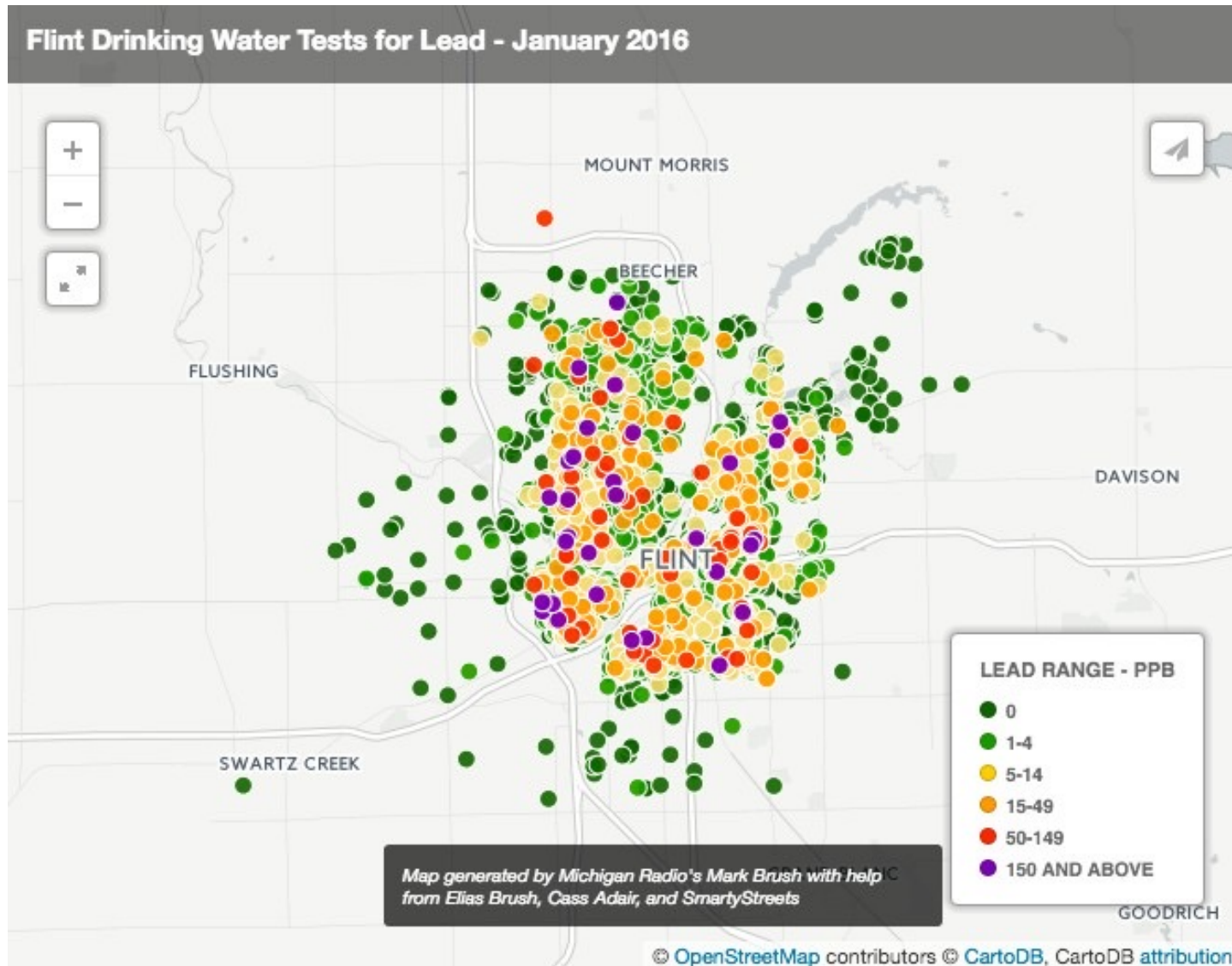
Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
$\{Milk\} \rightarrow \{Coke\}$
$\{Diaper, Milk\} \rightarrow \{Beer\}$

Association Analysis: Applications

- ## Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management

- ## Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period

- ## Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]
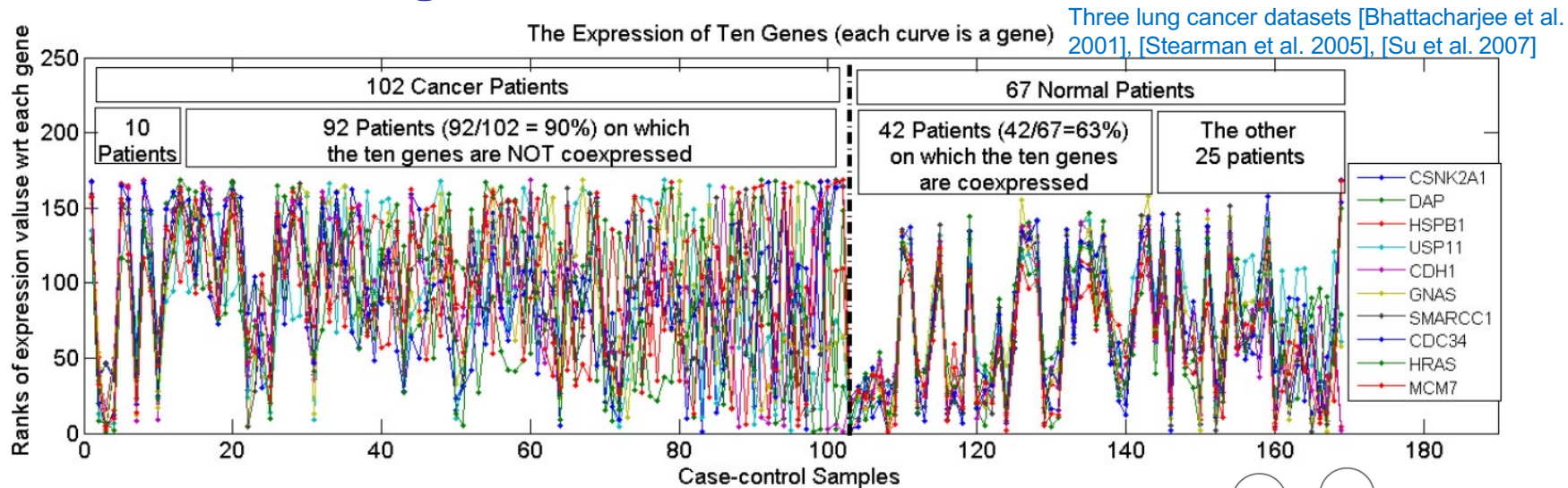
Enriched with the TNF/NFB signaling pathway
which is well-known to be related to lung cancer
P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

[Fang et al PSB 2010]
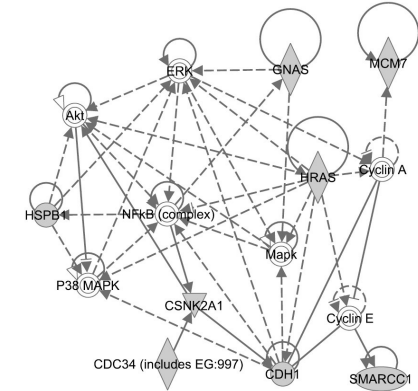
Spatial Co-location Patterns

- Given:
  - A collection of different types of spatial events
- Find: Co-located subsets of event types

Co-location Patterns – Sample Data
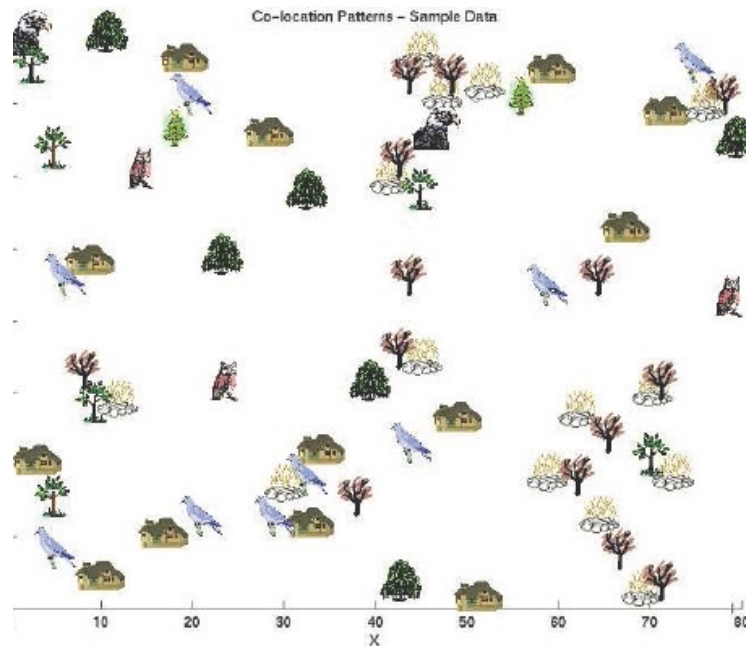
Answers: 🌳 🔥 and 🐦 🏠
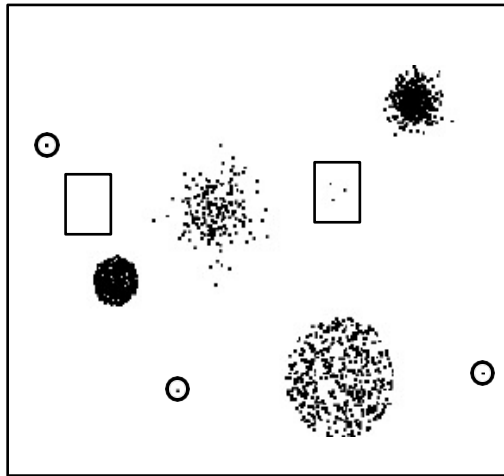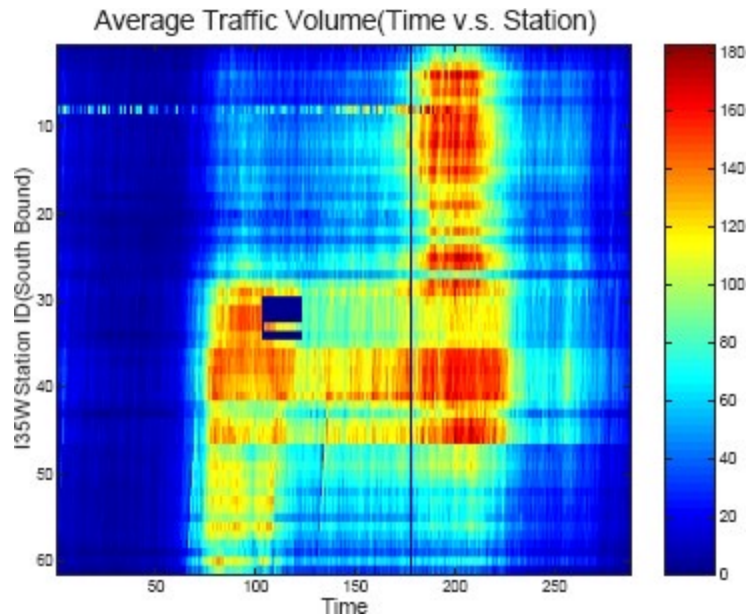
Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.
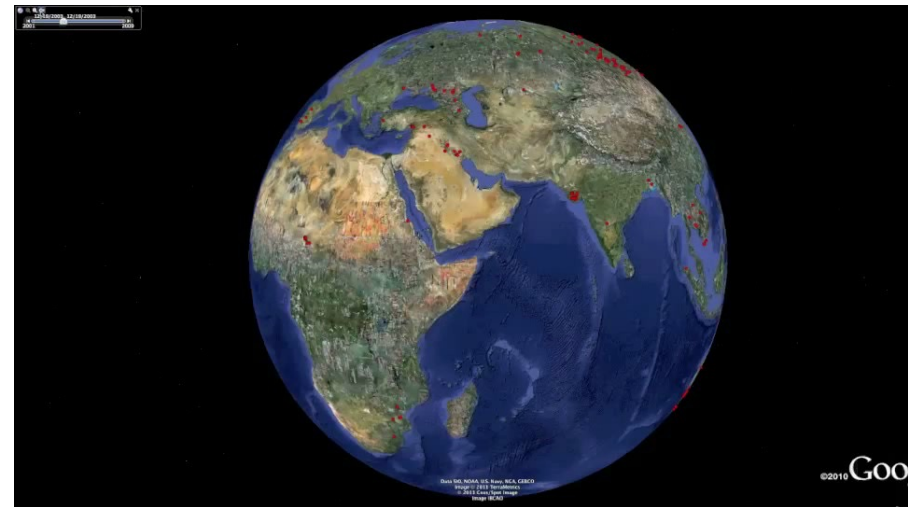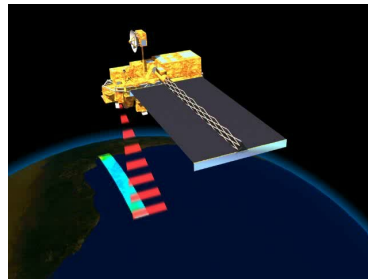
Spatial Outliers

- Spatial Outliers
  - Traffic Data in Twin Cities
  - Abnormal Sensor Detections
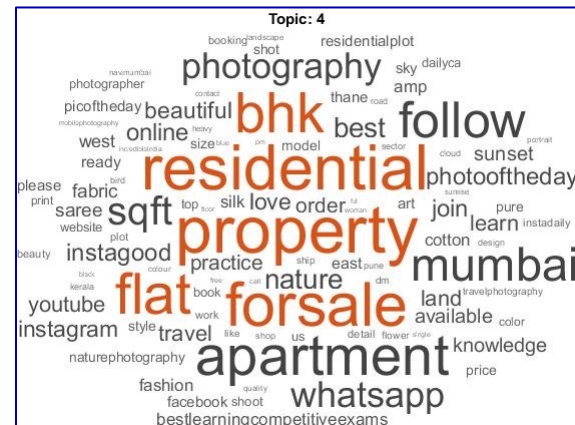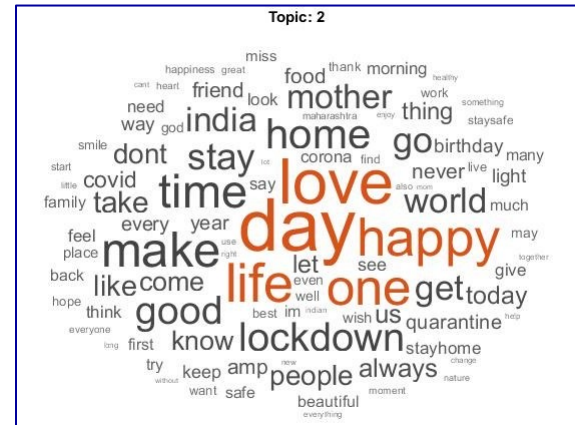 - Spatial and Temporal Outliers

Text Mining

The process of extracting interesting information and knowledge from unstructured text

Text Summarization

https://www.archives.gov/founding-docs/bill-of-rights-transcript

Topic Detection

What do people Tweet about?

Courtesy: Yunhao Fan

Sentiment Analysis

- Sentiment analysis is the detection of attitudes "enduring, affectively colored beliefs, dispositions towards objects or persons"
  1. Holder (source) of attitude
  2. Target (aspect) of attitude
  3. Type of attitude
     - From a set of types
       - Like, love, hate, value, desire, etc.
     - Or (more commonly) simple weighted polarity:
       - positive, negative, neutral, together with strength
  4. Text containing the attitude
     - Sentence or entire document

Credit: Dan Jurafsky

Sentiment Analysis

- Sentiment analysis is the detection of attitudes "enduring, affectively colored beliefs, dispositions towards objects or persons"

  1. Holder (source) of attitude
  2. Target (aspect) of attitude
  3. Type of attitude
     - From a set of types
       – Like, love, hate, value, desire, etc.
     - Or (more commonly) simple weighted polarity:
       – positive, negative, neutral, together with strength
  4. Text containing the attitude
     - Sentence or entire document

Credit: Dan Jurafsky

Why sentiment analysis?

- Movie: is this review positive or negative?
- Products: what do people think about the new iPhone?
- Public sentiment: how is consumer confidence? Is despair increasing?
- Politics: what do people think about this candidate or issue?
- Prediction: predict election outcomes or market trends from sentiment

Acceptance Speeches of Democratic Nomination for President

- Biden   2020: Negative (Confidence 83.7%)
- Clinton 2016: Positive (93.65%)
- Obama 2012: Positive (93.65%)
- Obama 2008: Positive (74.50%)
- JFK   1960: Positive (99.50%)

Data Mining and Privacy

Privacy Properties of Telephone Metadata

"You have my telephone number, connecting with your telephone number.

There are no names... in that database."

-President Obama

Mayer, Mutchler, and Mitchell (Stanford)

Data Mining and Privacy

Re-Identification

Lookup Source	% Matched
Google Places	16.6
Yelp	10.5
Facebook	13.7
All Automated Sources	31.9

Lookup Source	% Matched
Intelius	65
Google Search	58
All Automated Sources	26
All Sources	82

Automated approaches

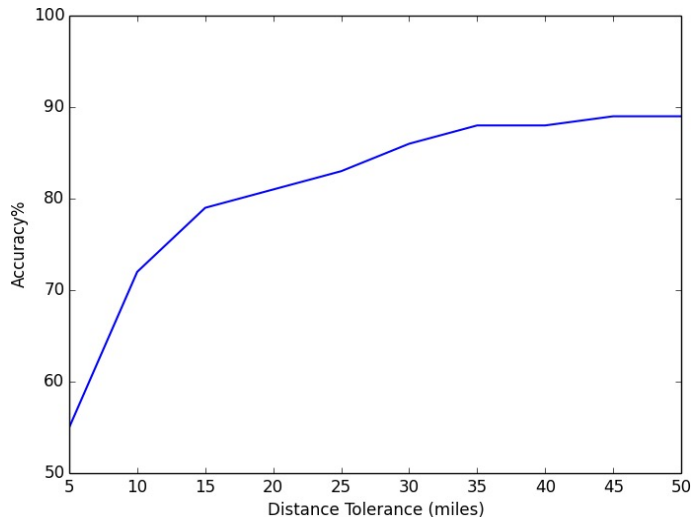Manual and combined approaches.

Data Mining and Privacy

"All it is, is the number pairs, when those calls took place, how long they took place.

So that database is sitting there."

-President Obama

Data Mining and Privacy

Home Location
Inference

Religion
Inference

≈ ¾ accuracy

(naïve heuristic on a small sample)

Methodology: re-identify businesses, cluster their locations

Methodology: comparison to Facebook data

Data Mining and Privacy

Sensitive Trait Inference

- Relapsing-Remitting Multiple Sclerosis (?)
- Cardiac Arrhythmia (✓)
- Owning an Assault Rifle (✓)
- Building a Grow House(?)
- Seeking an Abortion (?)

Methodology: automated and manual number re-identification

Idea: Intelligence law and policy should be informed by science, not lawyerly intuition

Privacy Issues

- Cambridge Analytica
  - Political consulting firm
  - What people"like" on Facebook can be used to predict personality traits (Kosinski, Stillwell and Graepel 2013)
  - Micro-targeted ads in 2016 US Elections
- Developed user profile for over 80 million user profiles
  - Demographics: age, education, sex, …
  - Psychographics: interests, opinions, values, …
  - Five personality traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism

"users who liked the 'Hello Kitty' brand tended to be high on 'Openness' and low on 'Conscientiousness,' 'Agreeableness,' and 'Emotional Stability [i.e., neuroticism].'"

Send a terrorism related ad to an older man who owns gun and is neurotic.

Privacy Issues

- How did they gather data for so many users?
  - Via a quiz app called thisismydigitallife
  - 270K users took the quiz
- How did they get 80 Million
  - Used loophole in Facebook API that allowed access to the friends of the users
- No clear evidence that ads could change voting preferences or behavior
- Personal data was collected and used without users' consent

The United States has no legal definition of personal data.

Ethics of Face Recognition

The Many Uses of Facial Recognition
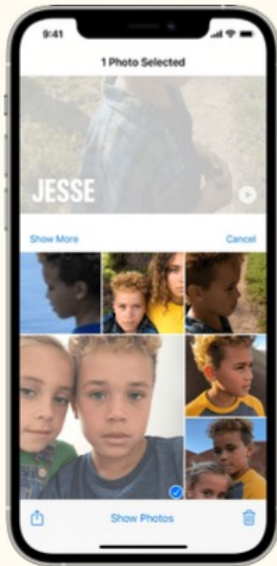
- Two ethical challenges of facial recognition
  - The risks of error
  - The threat to privacy

Erroneous Identification

- Nijeer Parks spent 10 days in jail and spent $5000 in legal fees for shoplifting and attempted assault of an officer (he was 30 mi away from the incident)

- Being the subject of an investigation, being arrested—they matter!
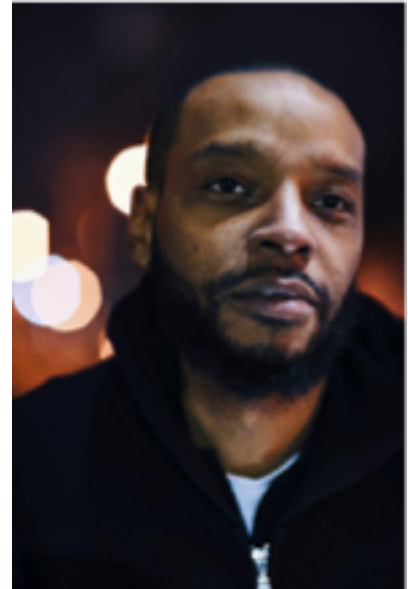
Nijeer Parks
(M. Sadek, NYT)

A Disparate Impact

- In an early academic study, Buolamwini found dark-skinned women were especially apt to be mis-matched, sometimes with false positive rates of over 30%.

- "To fail on one in three, in a commercial system, on something that's been reduced to a binary classification task, you have to ask, would that have been permitted if those failure rates were in a different subgroup?" —Buolamwini

Joy Buolamwini
(Stanford HAI)

National Academy of Sciences Recommendations

- Academic institutions should encourage the development of a basic understanding of data science in all undergraduates.

- Academic institutions should embrace data science as a vital new field that requires specifically tailored instruction delivered through majors and minors in data science.

- As data science programs develop, they should focus on attracting students with varied backgrounds and degrees of preparation and preparing them for success in a variety of careers.
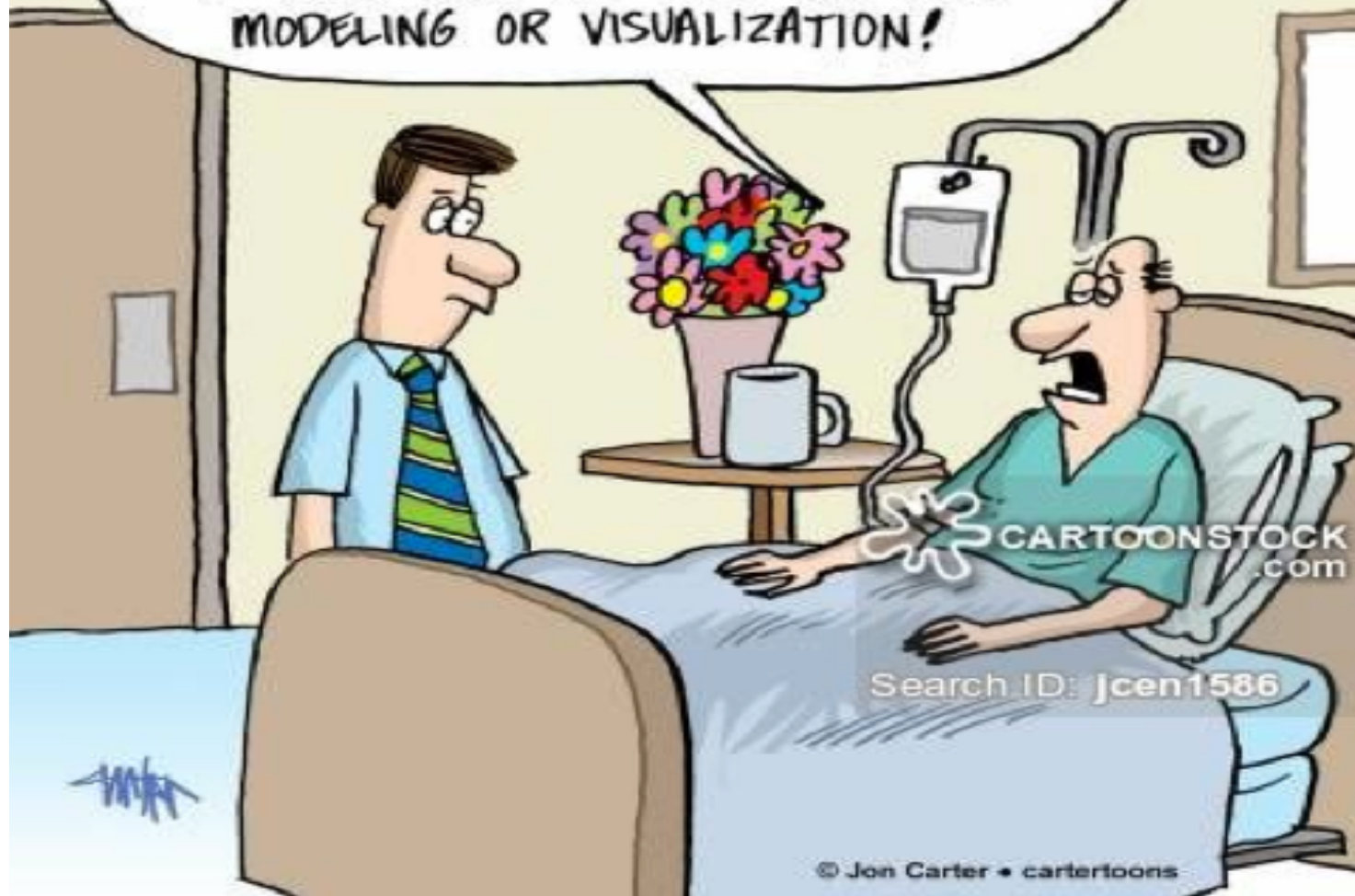
Data Science for Undergraduates: Opportunities and Options

National Academy of Sciences Recommendations

- Academic institutions should ensure that ethics is woven into the data science curriculum from the beginning and throughout.

- The data science community should adopt a code of ethics; such a code should be affirmed by members of professional societies, included in professional development programs and curricula, and conveyed through educational programs.

Some final thoughts!

- Work with real datasets
- Think about scalability from the outside
- Worry about ethics