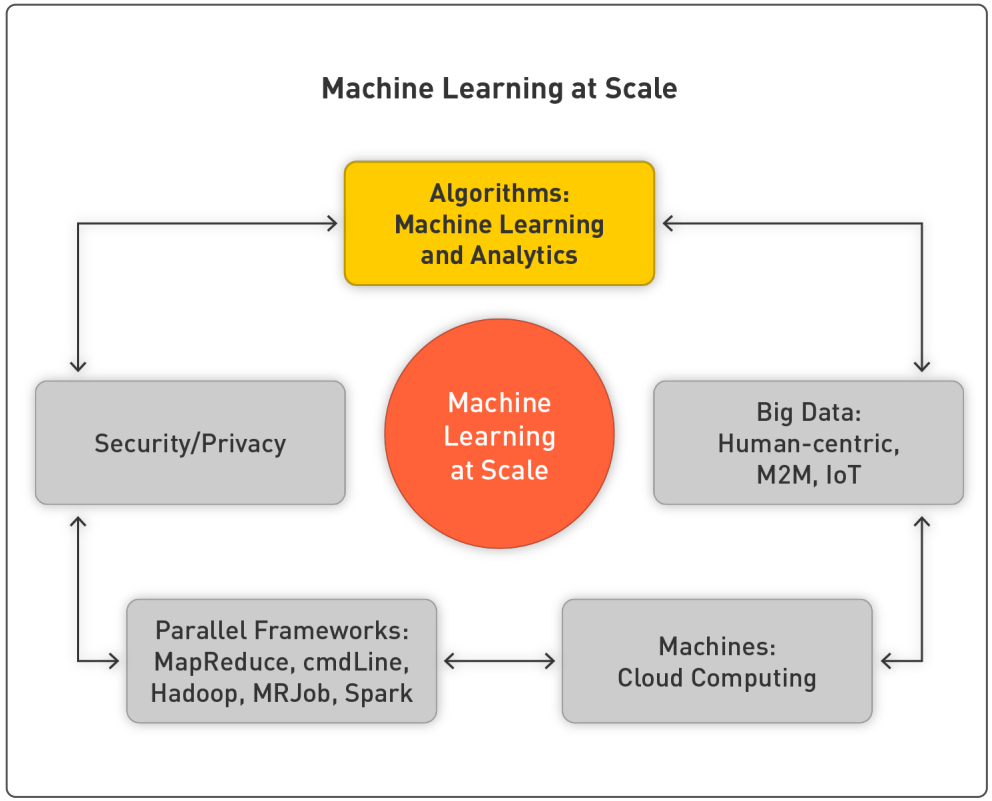

Machine Learning at Scale
Algorithms: Machine Learning and Analytics
Security/Privacy
Machine Learning at Scale
Big Data: Human-centric, M2M, IoT
Parallel Frameworks: MapReduce, cmdLine, Hadoop, MRJob, Spark
Machines: Cloud Computing

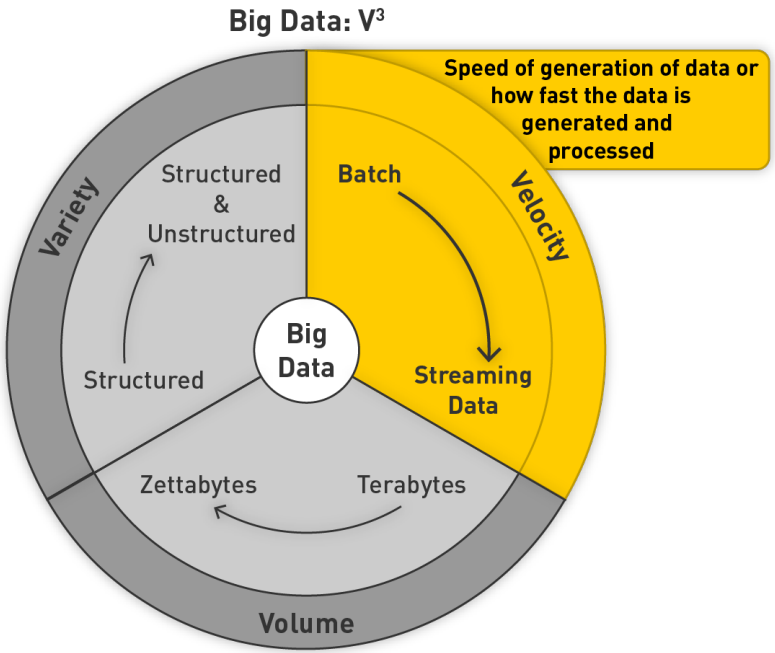
Big Data Definition

Big data is broad term for data sets so large or complex that traditional data-processing applications are **inadequate**:

- Processing
  - Laptop (8–16 GB of memory, 1 TB hard drive) overwhelmed with 4–5 GB
- Storage
  - Laptop only 1 TB
- Throughput
  - Three hours to read 1 TB on laptop

Other Challenges

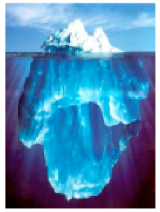
- Analysis
- Capture
- Data curation
- Search
- Sharing
- Storage
- Transfer
- Visualization
- Security
- Information privacy



IBM characterizes Big Data by its volume, velocity, and variety—or simply, V³.

Sources Driving Big Data

It's All Happening Online

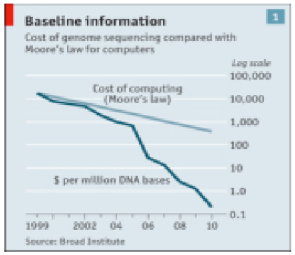


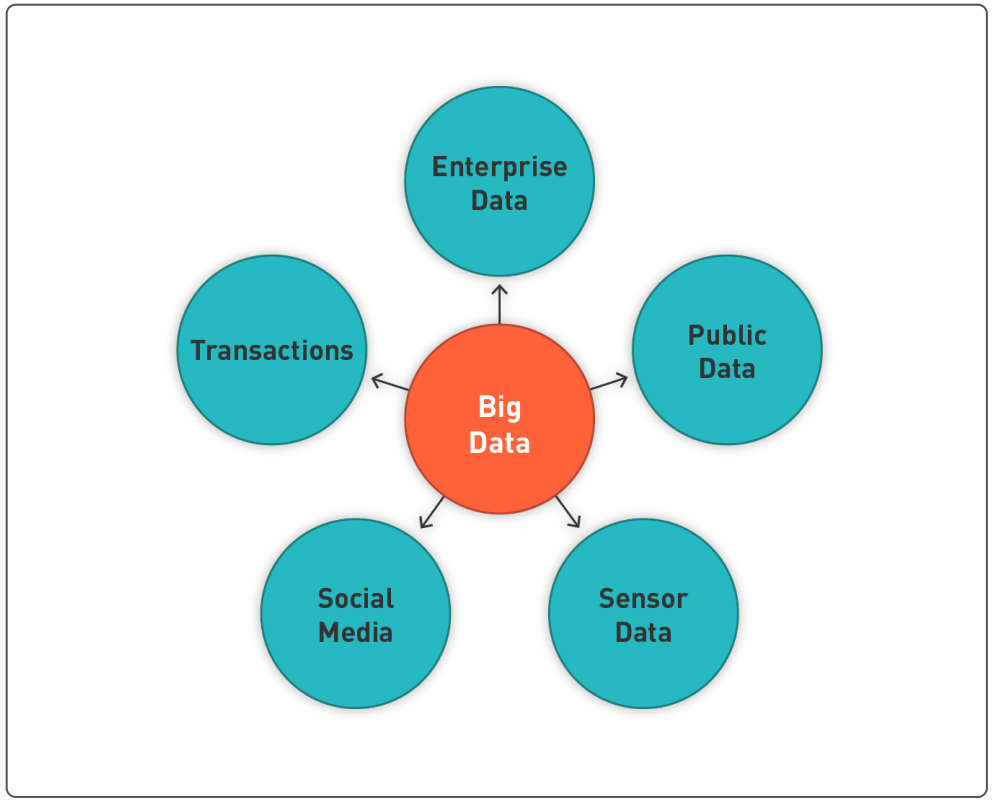
Every:
Click
Ad impression
Billing event
Fast forward, pause, . . .
Friend request
Transaction
Network message
Fault
. . .

User Generated (Web, Social & Mobile) Quantified Self

Internet of Things/M2M

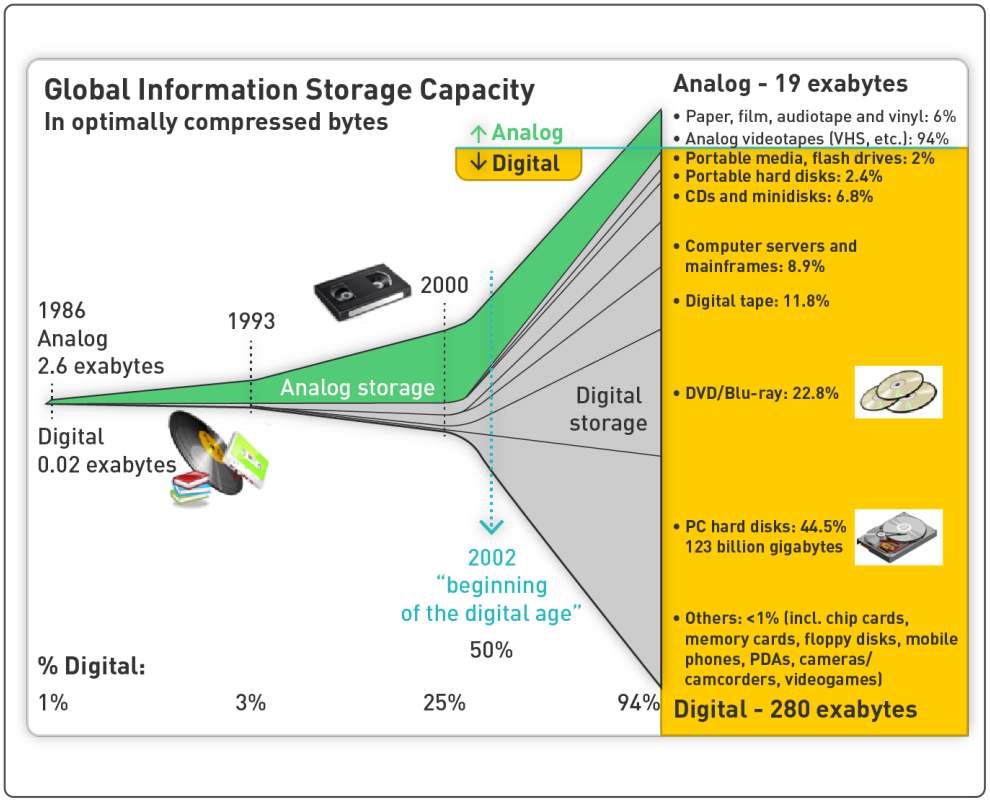
Scientific Computing




Enterprise Data
Transactions
Public Data
Big Data
Social Media
Sensor Data

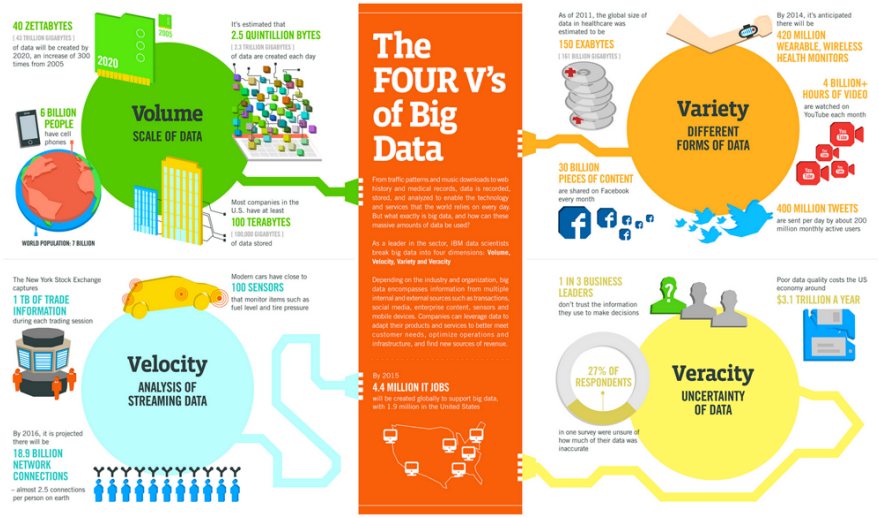
Areas of Application

New York Stock Exchange	1 TB of trade data daily
Facebook	Over 250 billion images, 250 PB of data
Ancestry.com	2.5 PB of data
Internet archives	2 PB of data
Hadron collider	15 PB of data yearly


Global Information Storage Capacity
In optimally compressed bytes
↑ Analog
↓ Digital
1986
Analog
2.6 exabytes
1993
2000
Analog storage
Digital
storage
Digital
0.02 exabytes
2002
"beginning
of the digital age"
50%
% Digital:
1%
3%
25%
94%
Analog - 19 exabytes
• Paper, film, audiotape and vinyl: 6%
• Analog videotapes (VHS, etc.): 94%
Digital - 280 exabytes
• Portable media, flash drives: 2%
• Portable hard disks: 2.4%
• CDs and minidisks: 6.8%
• Computer servers and mainframes: 8.9%
• Digital tape: 11.8%
• DVD/Blu-ray: 22.8%
• PC hard disks: 44.5%
123 billion gigabytes
• Others: <1% (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/ camcorders, videogames)

Data in Zettabytes (ZB)

- 1 ZB = 1021 = 1 billion laptops (one per person)
- In 2005: 0.12 ZB
- In 2007: 0.28 ZB
- In 2020: 40 ZB
- In context, today everyone's hard drive with 1 TB
    - Data storage today: 1–2 TB data per person (1–2 billion hard drives)
    - By 2020: 40 TB per person



Link to original image on IBM website.

Veracity: Quality of data being captured, can vary greatly

Applications of Big Data

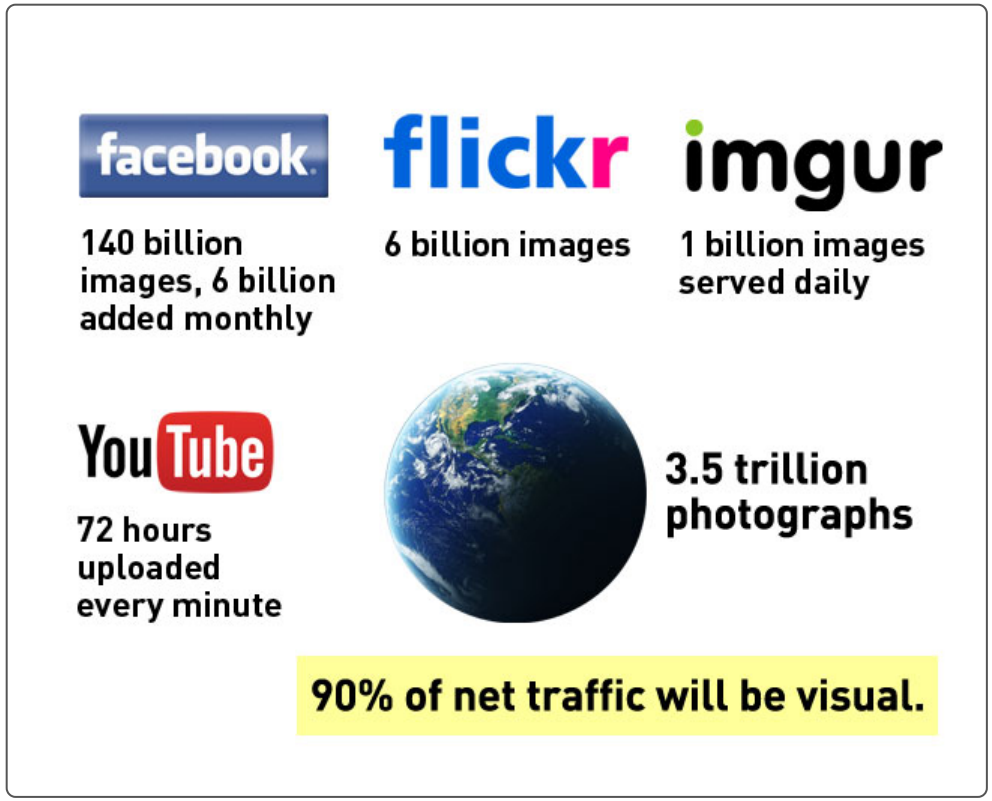
- Society based (human-centric)
- Internet of things (machine-centric)

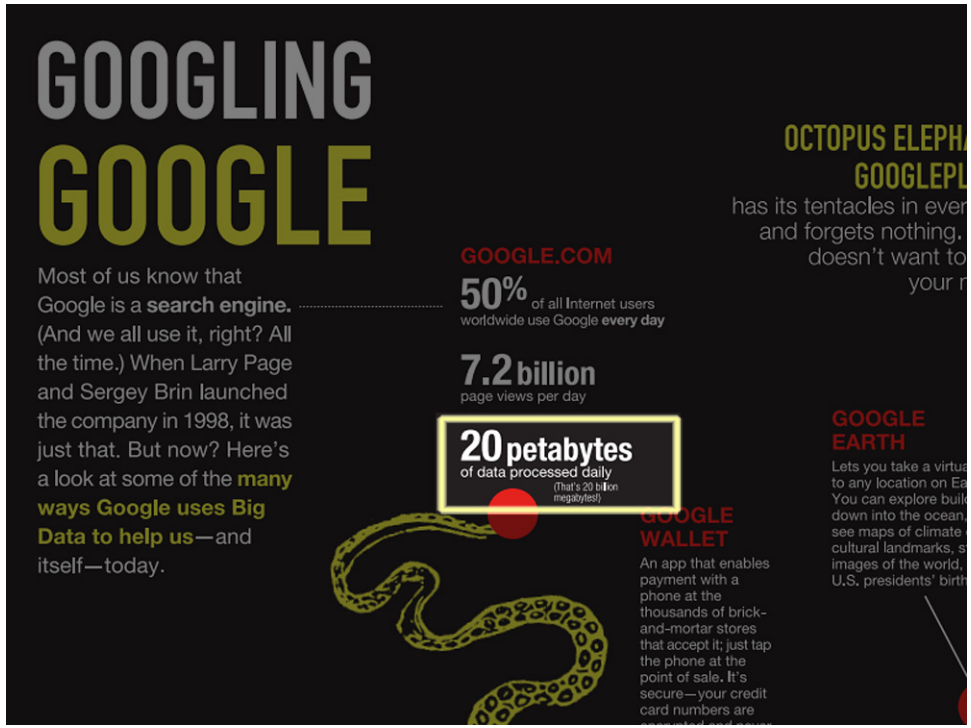
Societal and Personal Sources of Data

- Social, professional
- Quantified self (eating, sleeping, exercising)
- Voting
- Health care

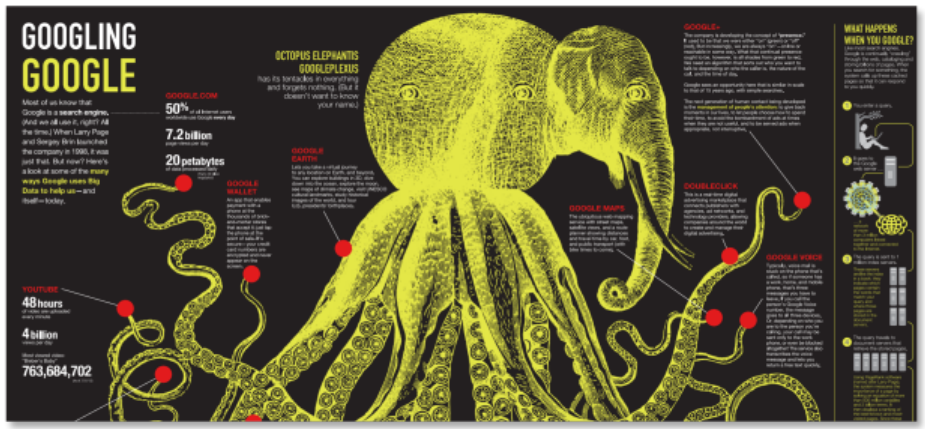
Online Society

- 3 billion people online today
- 1.5 billion on a social network
  - Facebook
  - Twitter
  - LinkedIn
  - WeChat
- Sharing huge volume of data


facebook
140 billion images, 6 billion added monthly
flickr
6 billion images
imgur
1 billion images served daily
YouTube
72 hours uploaded every minute
3.5 trillion photographs
90% of net traffic will be visual.


GOOGLING
GOOGLE
Most of us know that Google is a search engine. (And we all use it, right? All the time.) When Larry Page and Sergey Brin launched the company in 1998, it was just that. But now? Here's a look at some of the many ways Google uses Big Data to help us—and itself—today.
GOOGLE.COM
50% of all Internet users worldwide use Google every day
7.2 billion
page views per day
20 petabytes
of data processed daily
(That's 20 billion megabytes!)
OCTOPUS ELEPH
GOOGLEPL
has its tentacles in ever
and forgets nothing.
doesn't want to
your
GOOGLE EARTH
Lets you take a virtu
to any location on Ea
You can explore buil
down into the ocean,
see maps of climate
cultural landmarks, s
images of the world,
U.S. presidents' birth
GOOGLE WALLET
An app that enables payment with a phone at the thousands of brick-and-mortar stores that accept it; just tap the phone at the point of sale. It's secure—your credit card numbers are

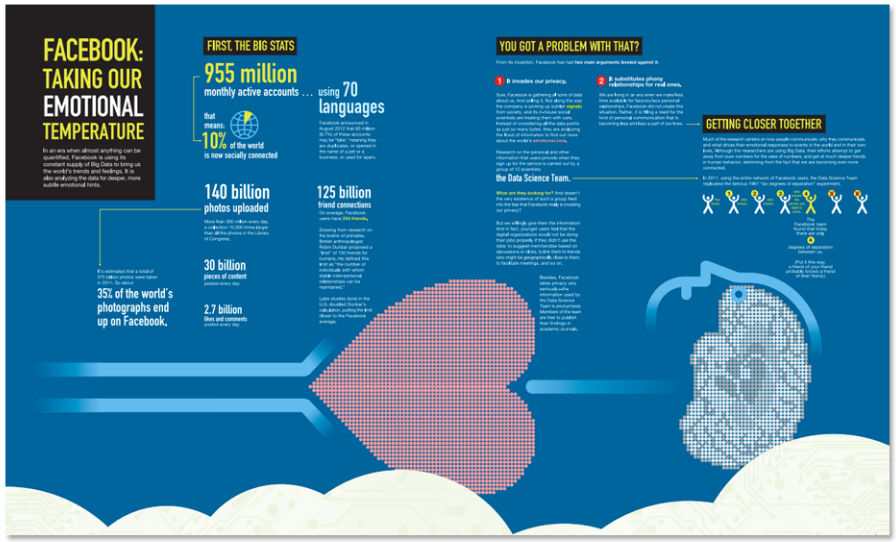

YOUTUBE
48 hours
of video are uploaded
every minute
4 billion
views per day
Most viewed video:
“Bieber’s Baby”
763,684,702
(As of 7/31/12)

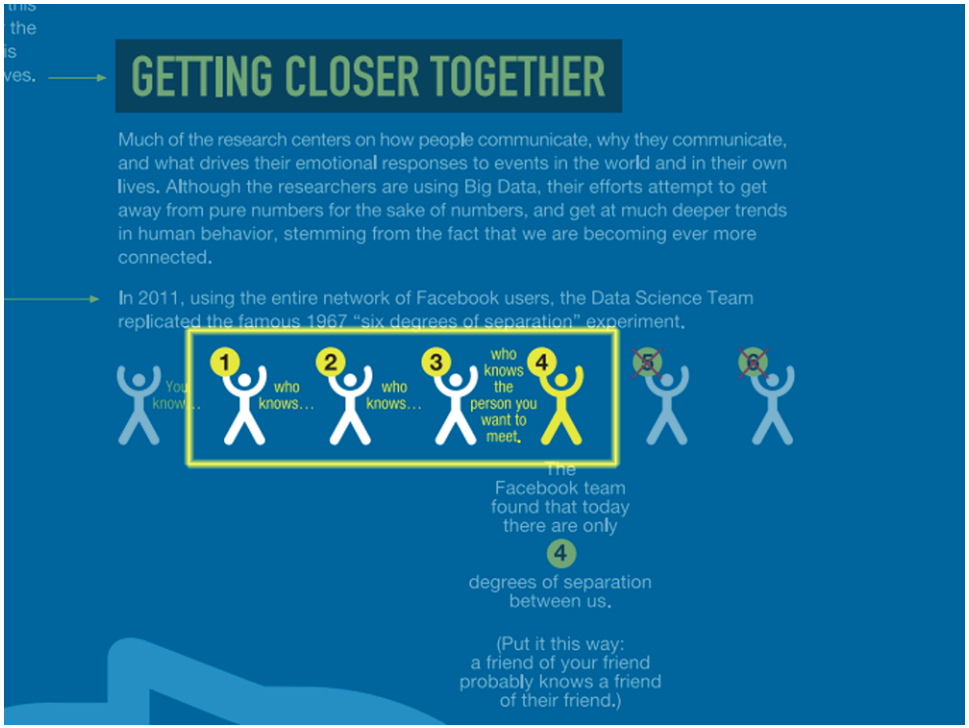


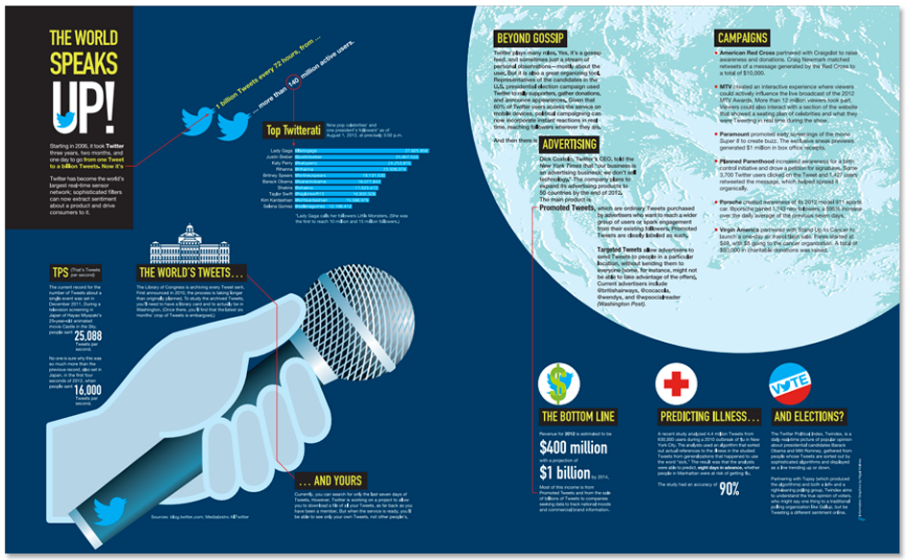
Data from 2012

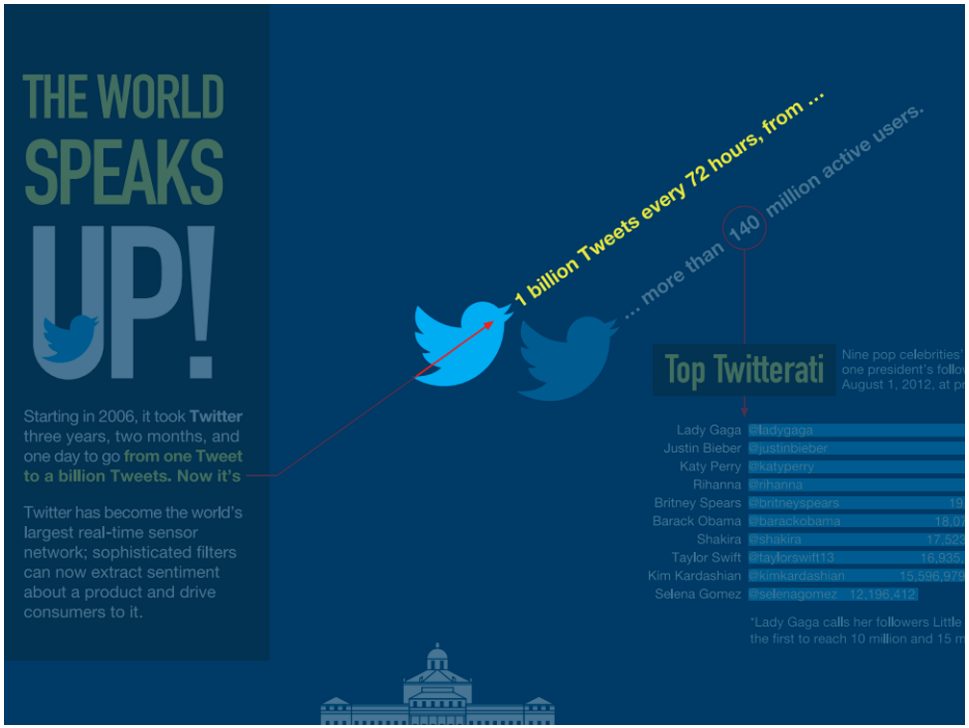
Facebook Big Data Stats

- 2.5 billion content items shared daily
- 1.5 billion people on network
- Over a trillion edges in graph
- 2.7 billion "likes" daily
- 105 TB data scanned via Hive every 30 minutes


FACEBOOK: TAKING OUR EMOTIONAL TEMPERATURE
In an era when almost anything can be quantified, Facebook is using its constant supply of Big Data to bring us the world's trends and feelings. It is also analyzing the data for deeper, more subtle emotional hints.
FIRST, THE BIG STATS
955 million
monthly active accounts . . .
using 70 languages
that means:
10% of the world is now socially connected
140 billion photos uploaded
125 billion friend connections
30 billion pieces of content
2.7 billion likes and comments
35% of the world's photographs end up on Facebook.
YOU GOT A PROBLEM WITH THAT?
1 It invades our privacy.
2 It substitutes phony relationships for real ones.
the Data Science Team,
GETTING CLOSER TOGETHER


GETTING CLOSER TOGETHER
Much of the research centers on how people communicate, why they communicate, and what drives their emotional responses to events in the world and in their own lives. Although the researchers are using Big Data, their efforts attempt to get away from pure numbers for the sake of numbers, and get at much deeper trends in human behavior, stemming from the fact that we are becoming ever more connected.
In 2011, using the entire network of Facebook users, the Data Science Team replicated the famous 1967 "six degrees of separation" experiment.
You know...
1
who knows...
2
who knows...
3
who knows the person you want to meet.
4
5
6
The Facebook team found that today there are only
4
degrees of separation between us.
(Put it this way: a friend of your friend probably knows a friend of their friend.)

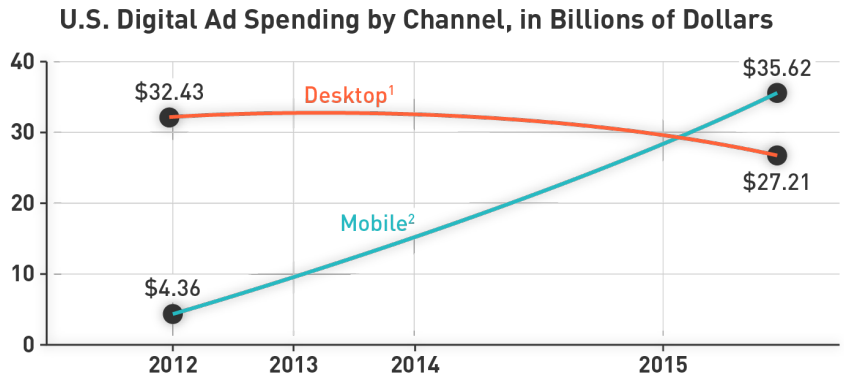

THE WORLD SPEAKS UP!
Starting in 2006, it took Twitter three years, two months, and one day to go from one Tweet to a billion Tweets. Now it's
Twitter has become the world's largest real-time sensor network, sophisticated filters can now extract sentiment about a product and drive consumers to it.
1 billion Tweets every 72 hours, from ...
... more than 140 million active users.
Top Twitterati
BEYOND GOSSIP
ADVERTISING
CAMPAIGNS
TPS (That's Tweets per second)
THE WORLD'S TWEETS...
25,088
16,000
... AND YOURS
THE BOTTOM LINE
$400 million
$1 billion
PREDICTING ILLNESS...
90%
AND ELECTIONS?


THE WORLD SPEAKS UP!
Starting in 2006, it took Twitter three years, two months, and one day to go from one Tweet to a billion Tweets. Now it's
Twitter has become the world's largest real-time sensor network; sophisticated filters can now extract sentiment about a product and drive consumers to it.
1 billion Tweets every 72 hours, from ...
... more than 140 million active users.
Top Twitterati
Lady Gaga @ladygaga
Justin Bieber @justinbieber
Katy Perry @katyperry
Rihanna @rihanna
Britney Spears @britneyspears
Barack Obama @barackobama
Shakira @shakira
Taylor Swift @taylorswift13
Kim Kardashian @kimkardashian
Selena Gomez @selenagomez 12,196,412

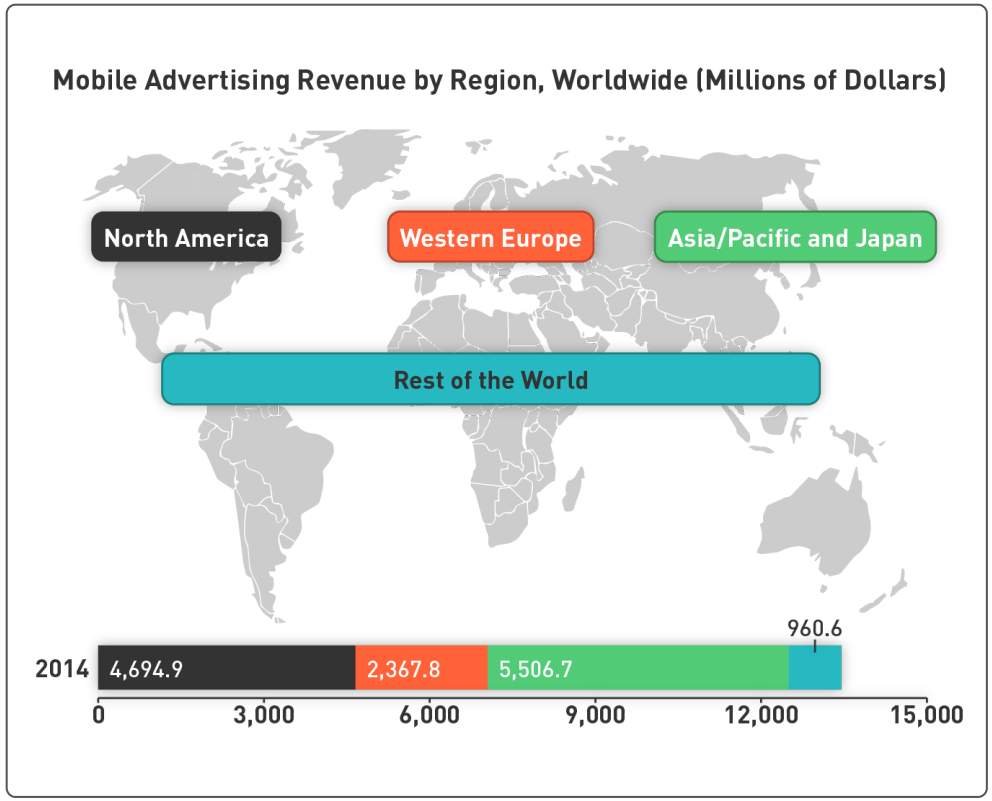

bluefin
LABS
@TheEllenShow can I just say, I love the jcpenney commercials with you, they are fabulous!
JDAngel7
RT @MoonieToons: I love the Ellen jc penny commercials lol
Stacia_Powers
These new JC Penney commercials are trying too hard. The Ellen ones are the only ones worth watching
TalGolightly
Why are even Ellen's JCPenny ads amusing? Can she do nothing without being graceful and charming?
neagle

Advertising

- Represents 2% of GDP in United States since 1900s
- Increased spending online as opposed to on traditional channels

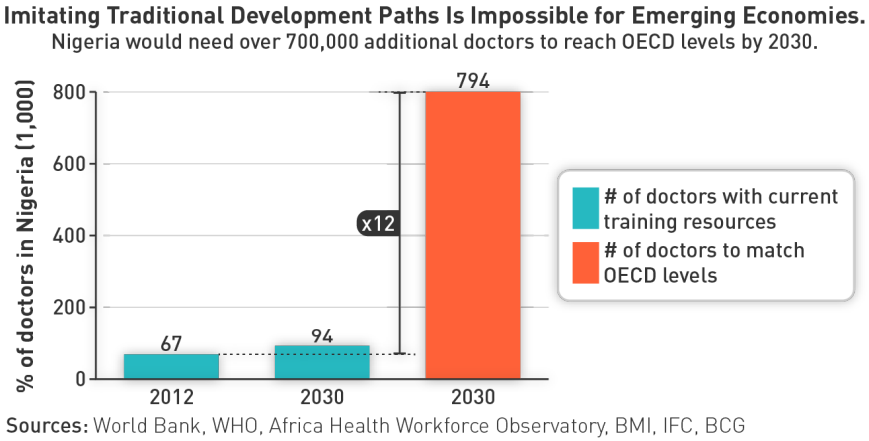


1- Includes spending primarily on desktop-based ads.
2- Includes classified, display (banners and other, rich media, and video), e-mail, lead generation, messaging-based, and search advertising; ad spending on tablets included.


Mobile Advertising Revenue by Region, Worldwide (Millions of Dollars)
North America
Western Europe
Asia/Pacific and Japan
Rest of the World
2014
4,694.9
2,367.8
5,506.7
960.6
0
3,000
6,000
9,000
12,000
15,000

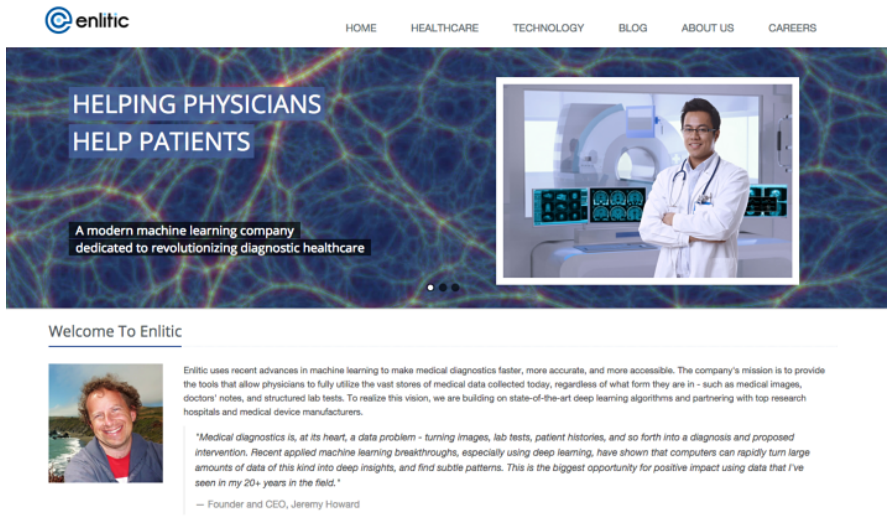
Health Care

Uneven across the world



- Nigeria: Population 170 million, 67,000 doctors (1:2600 ratio)
- United States: Population 300 million, 1 million doctors (1:300 ratio)

Technology can help address discrepancies in health care.



Jeremy Howard: Applying machine learning to medical diagnostics

Genomics

- Cost of gene sequencing: $100,000 (2005) to $4,000 today
- Near future, $40 at your local pharmacy

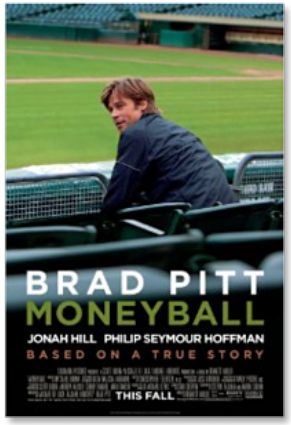
In the United States, 18% of GDP is spent on health care.

Project of Intel-GE Care Innovations Lab

Home carpet that senses movement, helpful for seniors "aging in place"

Sports

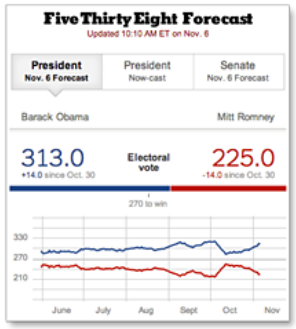
- Oakland Athletics baseball team in Oakland, CA



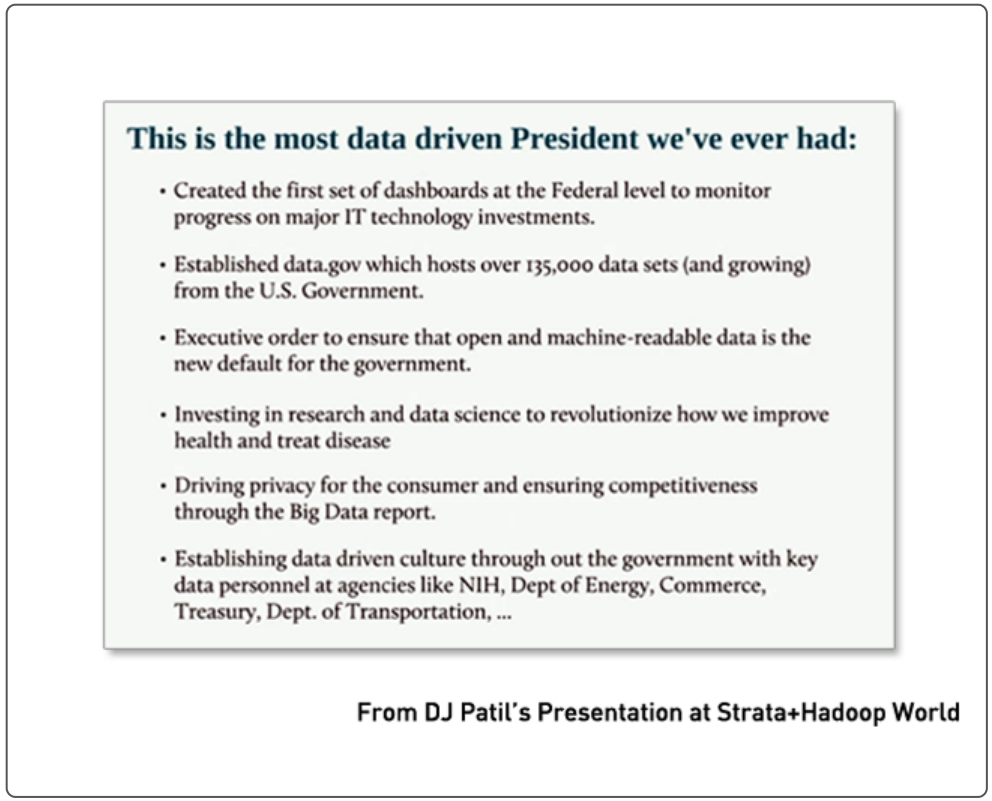
- Data science used to select best team on small budget

Government

- Barack Obama's victories in 2008 and 2012



- Attributable to effective use of data science


This is the most data driven President we've ever had:
• Created the first set of dashboards at the Federal level to monitor progress on major IT technology investments.
• Established data.gov which hosts over 135,000 data sets (and growing) from the U.S. Government.
• Executive order to ensure that open and machine-readable data is the new default for the government.
• Investing in research and data science to revolutionize how we improve health and treat disease
• Driving privacy for the consumer and ensuring competitiveness through the Big Data report.
• Establishing data driven culture through out the government with key data personnel at agencies like NIH, Dept of Energy, Commerce, Treasury, Dept. of Transportation, ...
From DJ Patil's Presentation at Strata+Hadoop World

AMP Lab (UC Berkeley)

Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

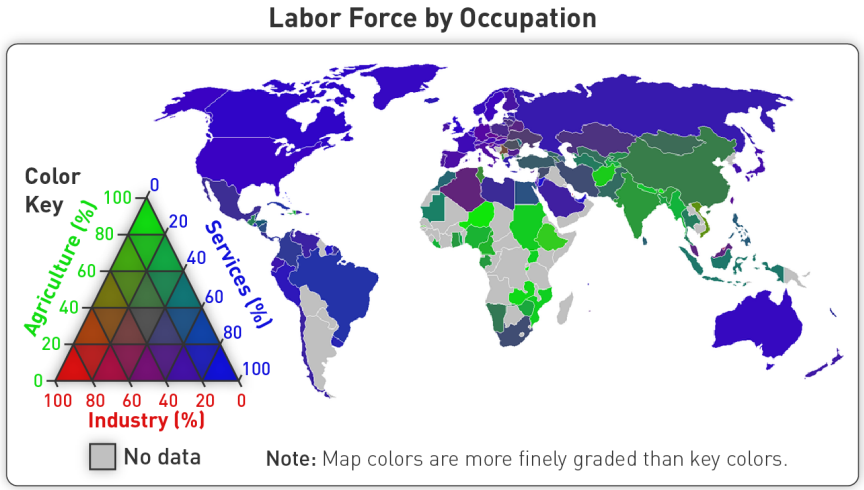
Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

National Science Foundation: In addition to funding the Big Data solicitation, and keeping with its focus on basic research, NSF is implementing a comprehensive, long-term strategy that includes new methods to derive knowledge from data; infrastructure to manage, curate, and serve data to communities; and new approaches to education and workforce development. Specifically, NSF is:

- Encouraging research universities to develop interdisciplinary graduate programs to prepare the next generation of data scientists and engineers;
- Funding a $10 million Expeditions in Computing project based at the University of California, Berkeley, that will integrate three powerful approaches for turning data into information - machine learning, cloud computing, and crowd sourcing;

Micro work-based framework connecting short tasks with workers



First world: Service oriented, can leverage machine learning

Summary

- Many big data problems where humans play a central role
- Next: Contrast with machine-centric big data

We are at a tipping point—humans are no longer at the center of the data universe.

The Internet of Things (IoT)

More data manufactured by machines—servers, cell phones, GPS-enabled cars—than by people

Advent of IoT Era

- Machine-to-machine learning (M2M)
  - Attach sensors to things animate and inanimate, and network these
  - Stream data
  - By 2020, over 26 billion things in network (majority M2M)
- Has evolved from convergence of wireless technologies, micro-electromechanical systems (MEMS), and the Internet
- Is a scenario in which objects, animals, or people are provided with:
  - Unique identifiers
  - The ability to transfer data over a network without human-to-human or human-to-computer interaction

Japan Earthquake

- Fourteen seconds before, all Shinkansen trains (bullet trains) and plants shut down
- Mesh of sensors from connected computers to form an early warning system

Tracking Nature in the Wild

- Sea lions instrumented with GPS tracking system
- Underwater hub picks up signals

Serengeti

- Animals tracked through attached GPS systems or through sensors (motion detectors and cameras)
- Work of Tanya Berger-Wolf, 2014

Challenges

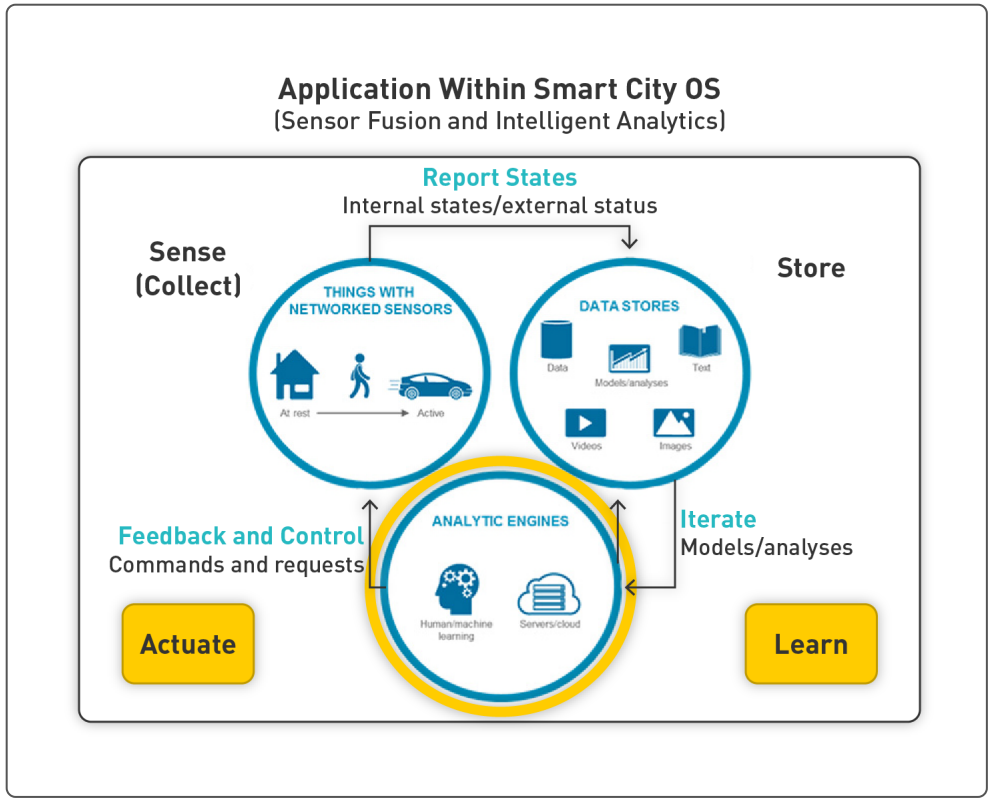
- Battery life, storage, network connectivity
- Privacy and security

Smart Homes

- By 2020, 1 billion smart meters
- One megabyte per device daily (from polling 1,000 times) → 1 petabyte of data daily

Smart Cities

- Involves capturing telemetry data from IoT and citizens (cell phones)
- Affects all sectors of society, including:
  - Government services
  - Transport and traffic management
  - Energy
  - Health care
  - Water
  - Waste


Application Within Smart City OS
(Sensor Fusion and Intelligent Analytics)
Report States
Internal states/external status
Sense
(Collect)
Store
THINGS WITH
NETWORKED SENSORS
At rest
Active
DATA STORES
Data
Models/analyses
Text
Videos
Images
ANALYTIC ENGINES
Human/machine
learning
Servers/cloud
Feedback and Control
Commands and requests
Iterate
Models/analyses
Actuate
Learn

Autonomous Vehicles

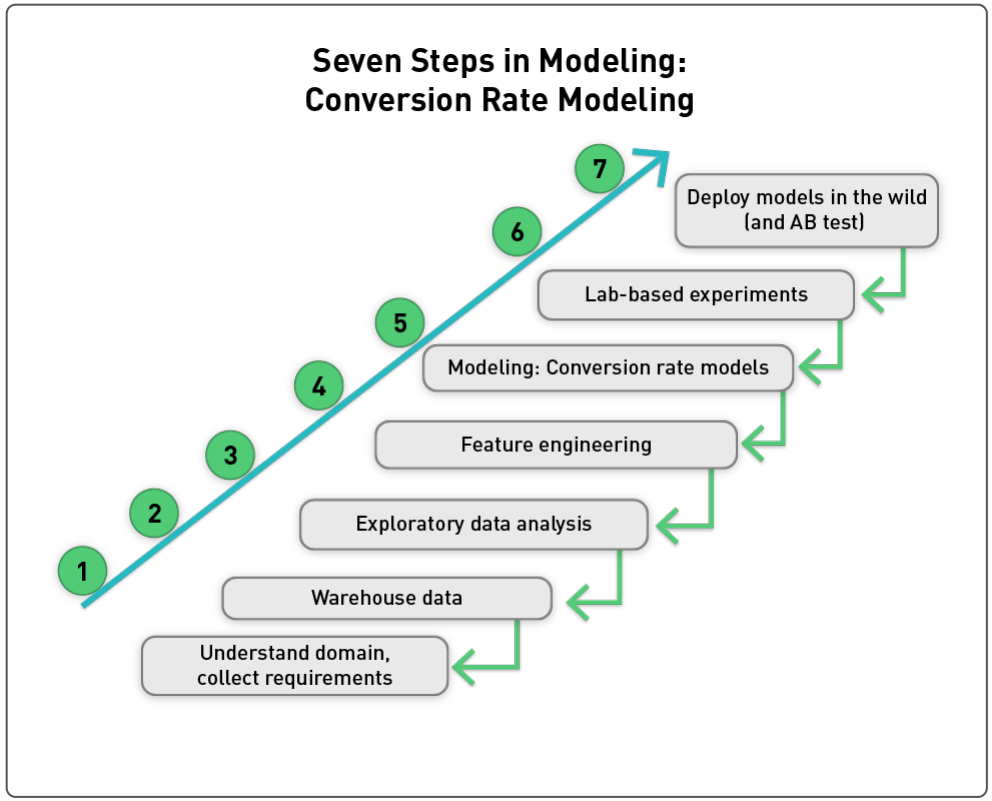
- Driverless cars (Google)
- Drones


Big data is not just about Web 2.0/Social, but IoT also.
AUTOMOTIVE
Auto sensors reporting location, problems
COMMUNICATIONS
Location-based advertising
CONSUMER PACKAGED GOODS
Sentiment analysis of what's hot, customer service
FINANCIAL SERVICES
Risk & portfolio analysis
New products
EDUCATION & RESEARCH
Experiment sensor analysis
HIGH TECHNOLOGY / INDUSTRIAL MFG.
Mfg quality
Warranty analysis
LIFE SCIENCES
Clinical trials
Genomics
MEDIA / ENTERTAINMENT
Viewers / advertising effectiveness
ON-LINE SERVICES / SOCIAL MEDIA
People & career matching
Website optimization
HEALTH CARE
Patient sensors, monitoring, EHRs
Quality of care
OIL & GAS
Drilling exploration sensor analysis
RETAIL
Consumer sentiment
Optimized marketing
TRAVEL & TRANSPORTATION
Sensor analysis for optimal traffic flows
Customer sentiment
UTILITIES
Smart Meter analysis for network capacity
LAW ENFORCEMENT & DEFENSE
Threat analysis - social media monitoring, photo analysis
THE cloudera SESSIONS
your inside track to big data

By 2020, over 26 billion devices will make up the IoT.

Introduction

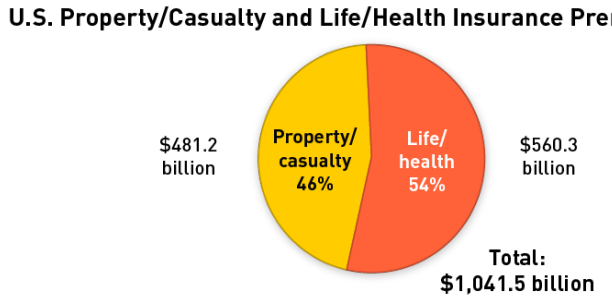
- New opportunities and business needs generated by big data applications
- Massive opportunity and challenge for machine learning
- Machine learning does not exist in a vacuum
    - Requires a sophisticated ecosystem


Seven Steps in Modeling:
Conversion Rate Modeling
1
2
3
4
5
6
7
Understand domain, collect requirements
Warehouse data
Exploratory data analysis
Feature engineering
Modeling: Conversion rate models
Lab-based experiments
Deploy models in the wild (and AB test)

Where does time go in large-scale machine learning?

Eighty percent of the work in any data project is preparing the data.

Example: U.S. Insurance Industry



- Net premiums totaled $1 trillion in 2013
  - Almost half property and casualty (auto, home, commercial)
- Highly competitive space, leading to development of new programs
  - E.g., special rates for good drivers
    - Score computed through accident history over, say, five years
    - Static, annual, small-scale, batch computation

Progressive's Snapshot Product

- "Bad driver" surveillance using telematics
- Via small box plugged into steering wheel that records, sends information
- Features:
  - How many miles driven
  - How many miles driven between midnight and 4 a.m.
  - Use of sharp or gentle braking
- Data used to analyze driving patterns, risk to insurance company
- Potential reduced premiums for user

Privacy

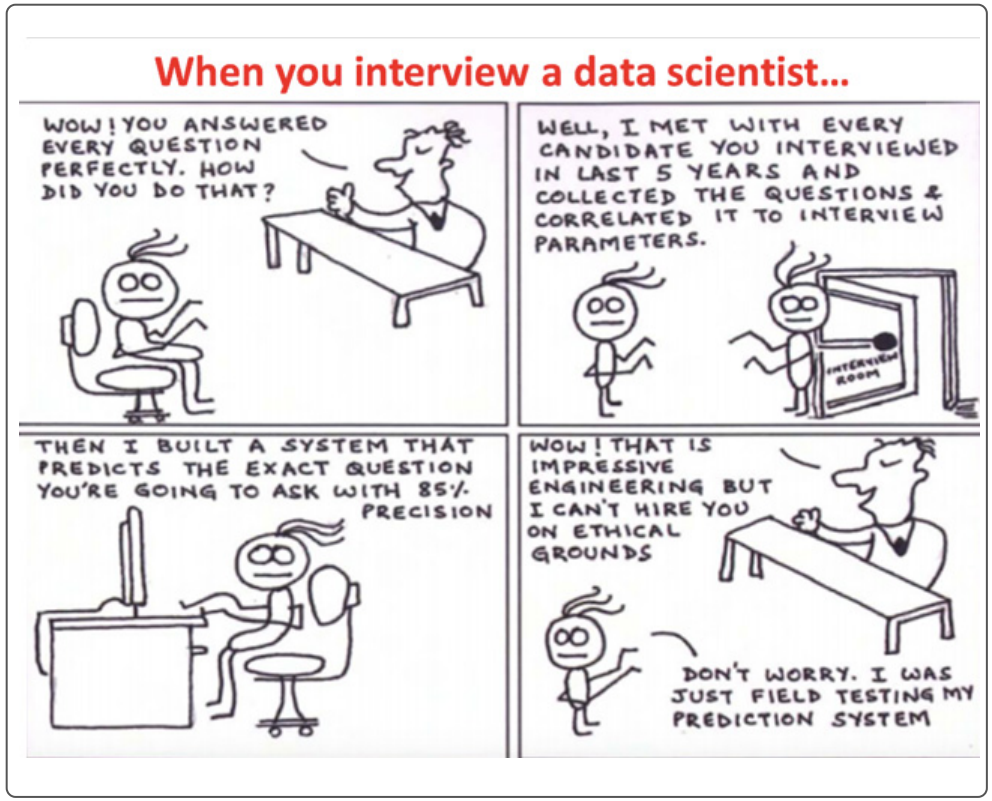
- Huge data exhaust being generated from digital and online activities
- Data not transitory but permanent
- Caution needed with storage, accessibility

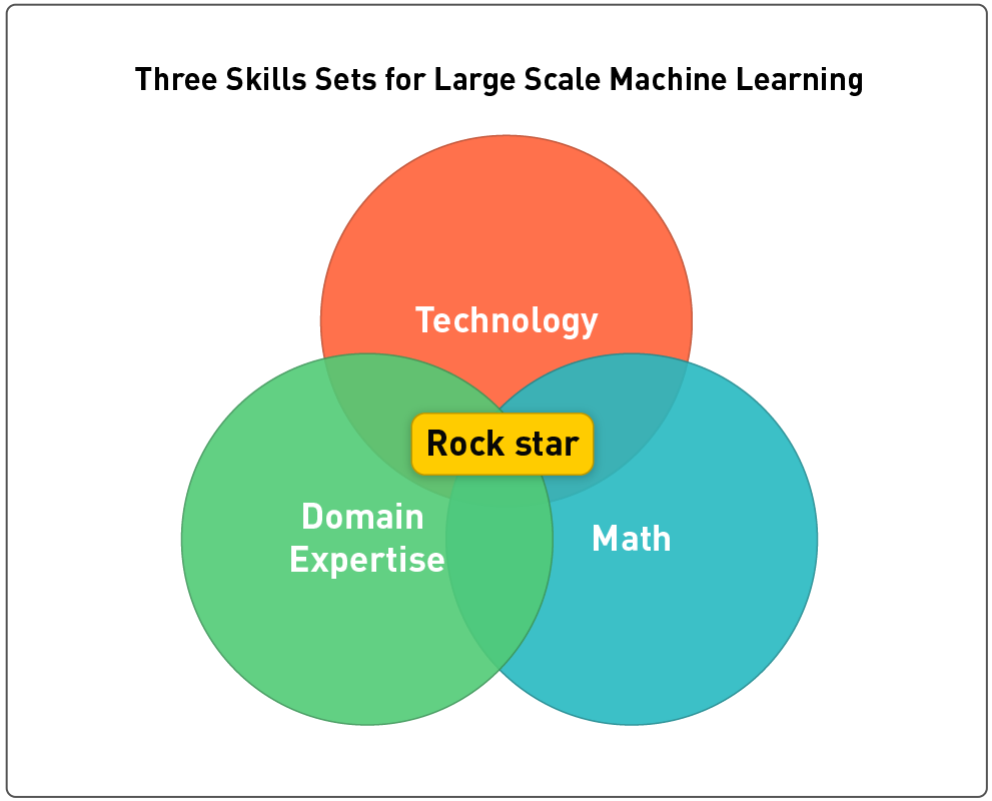
Challenges

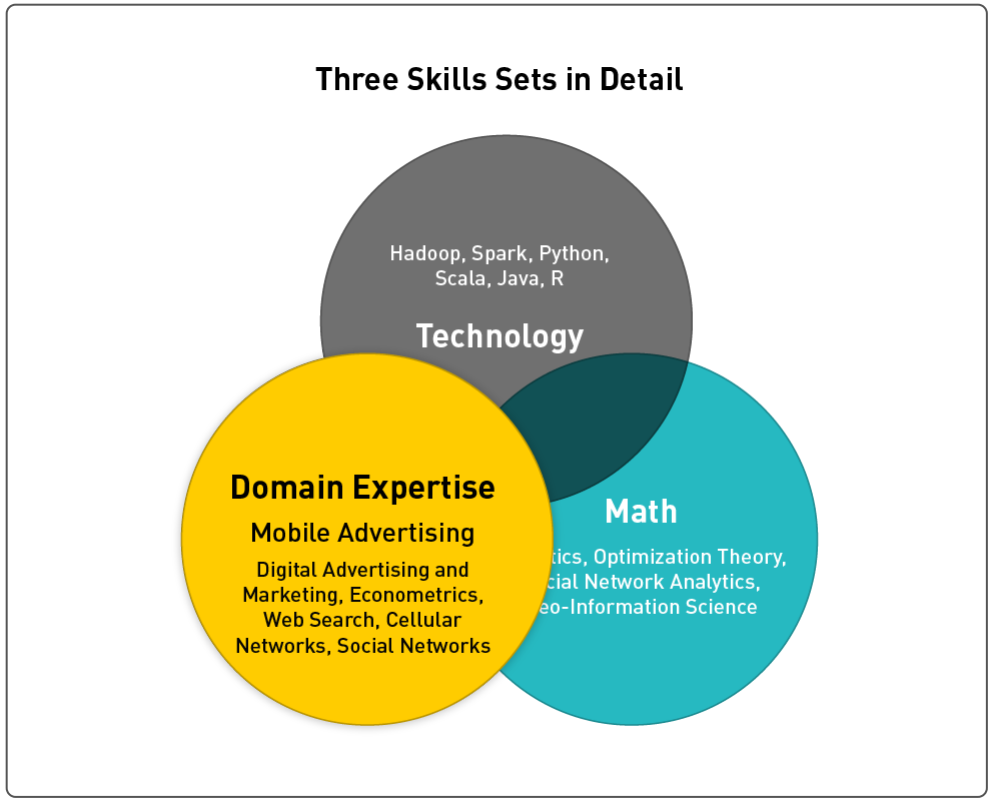
- International security
- Privacy

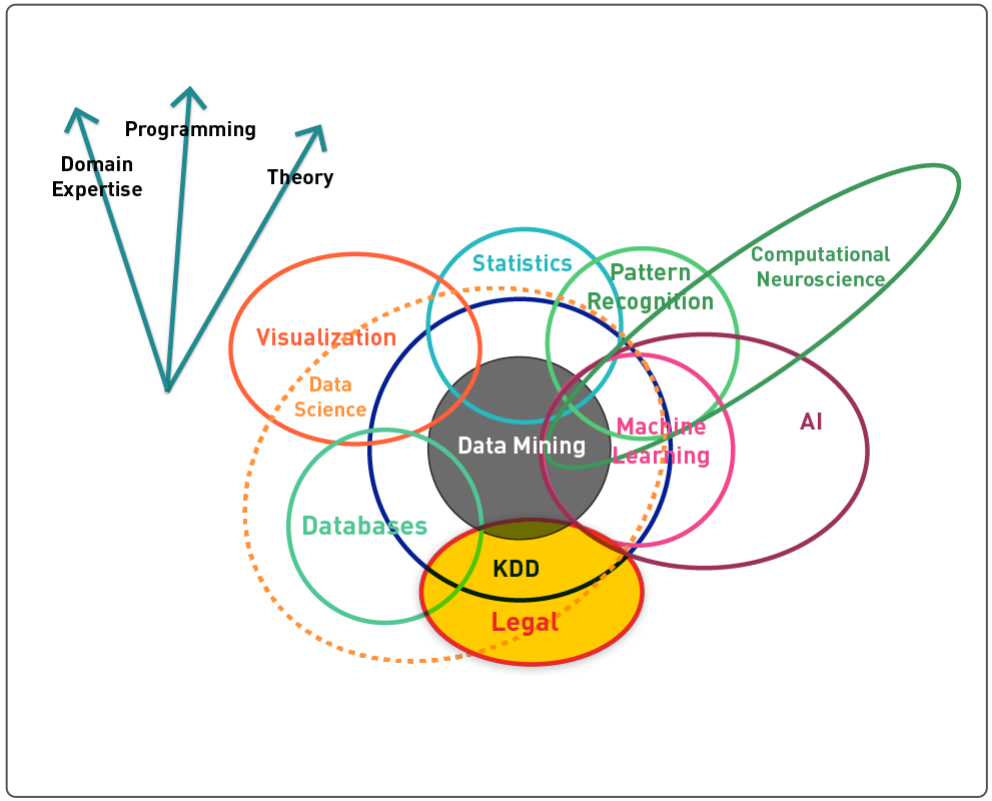
Pacemaker: Privacy vs. Ownership

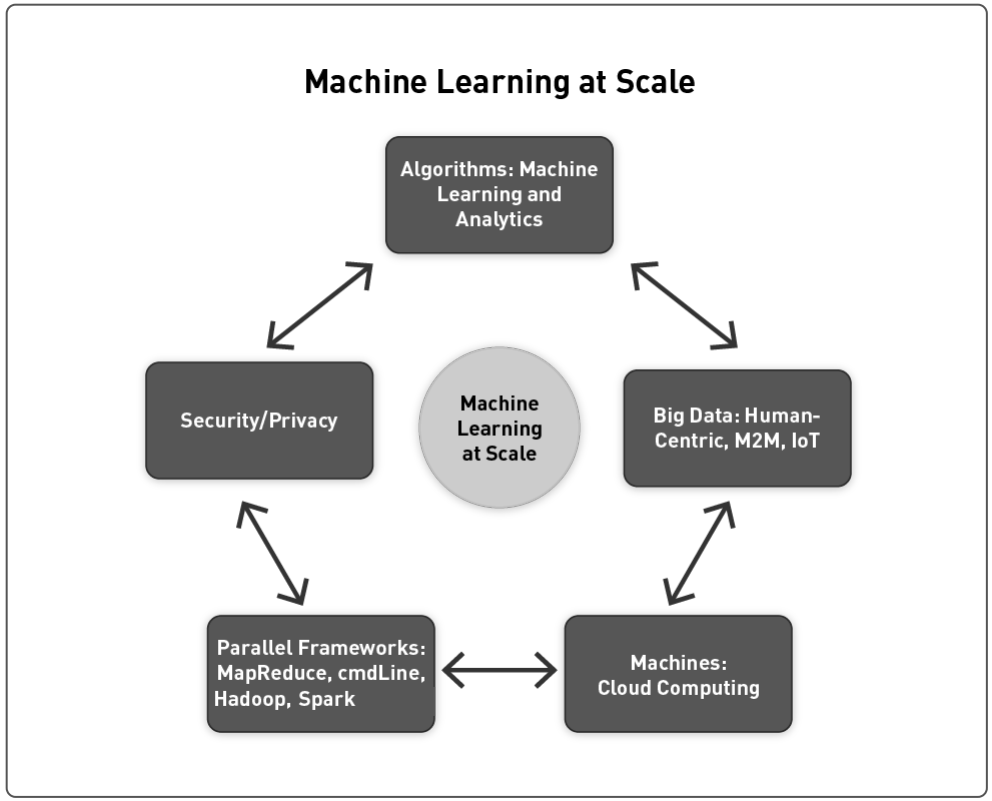
Who owns the data we generate?


When you interview a data scientist...
WOW! YOU ANSWERED EVERY QUESTION PERFECTLY. HOW DID YOU DO THAT?
WELL, I MET WITH EVERY CANDIDATE YOU INTERVIEWED IN LAST 5 YEARS AND COLLECTED THE QUESTIONS & CORRELATED IT TO INTERVIEW PARAMETERS.
INTERVIEW ROOM
THEN I BUILT A SYSTEM THAT PREDICTS THE EXACT QUESTION YOU'RE GOING TO ASK WITH 85% PRECISION
WOW! THAT IS IMPRESSIVE ENGINEERING BUT I CAN'T HIRE YOU ON ETHICAL GROUNDS
DON'T WORRY. I WAS JUST FIELD TESTING MY PREDICTION SYSTEM


Three Skills Sets for Large Scale Machine Learning
Technology
Rock star
Domain Expertise
Math


Three Skills Sets in Detail
Hadoop, Spark, Python,
Scala, Java, R
Technology
Domain Expertise
Mobile Advertising
Digital Advertising and
Marketing, Econometrics,
Web Search, Cellular
Networks, Social Networks
Math
tics, Optimization Theory,
cial Network Analytics,
eo-Information Science

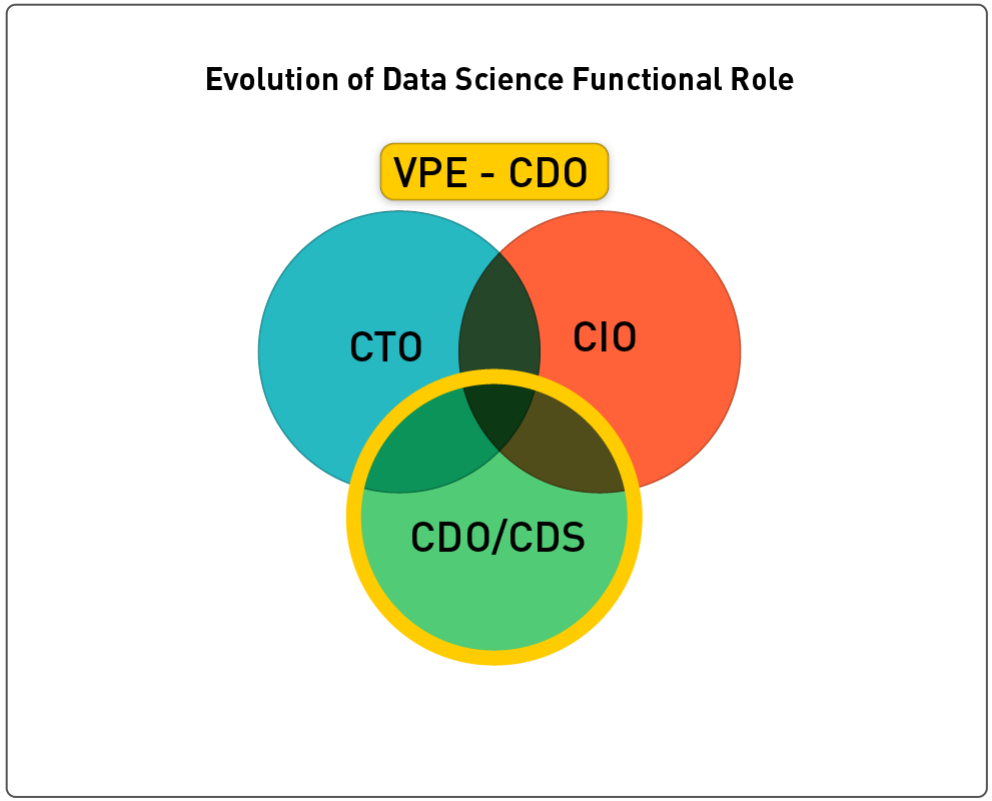

Programming
Domain Expertise
Theory
Statistics
Pattern Recognition
Computational Neuroscience
Visualization
Data Science
Data Mining
Machine Learning
AI
Databases
KDD
Legal


Machine Learning at Scale
Algorithms: Machine Learning and Analytics
Security/Privacy
Machine Learning at Scale
Big Data: Human-Centric, M2M, IoT
Parallel Frameworks: MapReduce, cmdLine, Hadoop, Spark
Machines: Cloud Computing


"Swiss army knife of the 21st century"
—MediaGuardian Innovation Awards

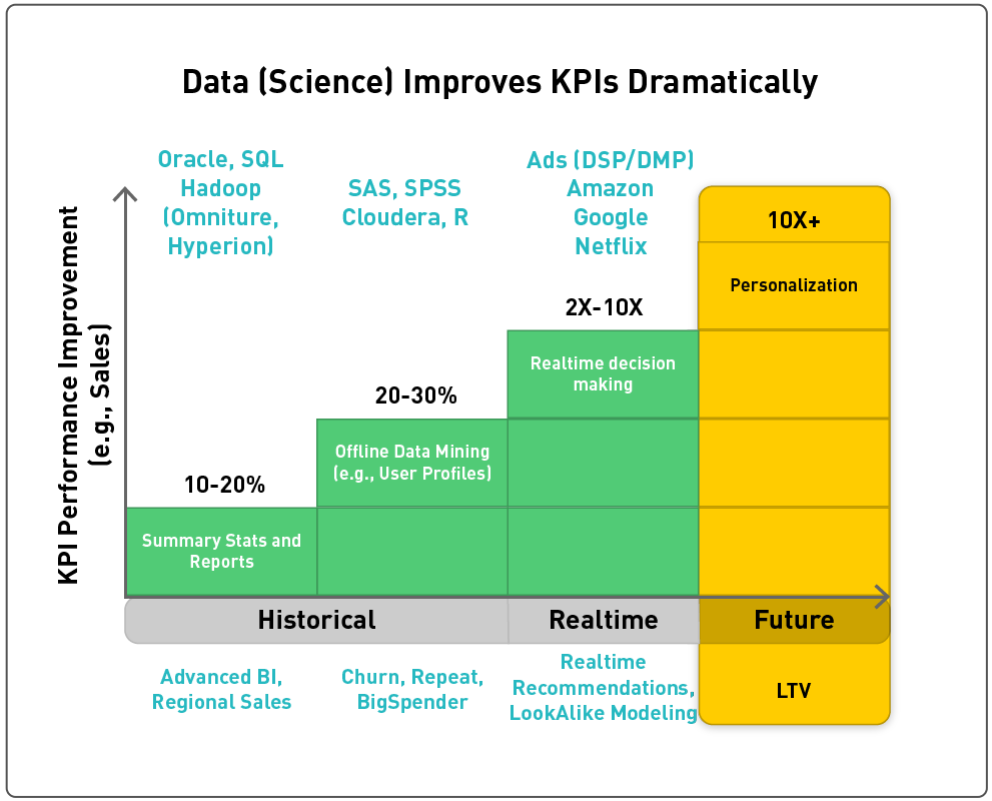
Data Science (DS) Team

- Experts in:
  - Technology
  - Math
  - Domain
  - Legal (privacy, security)
- Project managers
- Research scientists
- Communications experts


Evolution of Data Science Functional Role
VPE - CDO
CTO
CIO
CDO/CDS

The challenge is to leverage data at scale ...

... as we move from era of business intelligence to data science.


Data (Science) Improves KPIs Dramatically
KPI Performance Improvement (e.g., Sales)
Oracle, SQL
Hadoop
(Omniture, Hyperion)
SAS, SPSS
Cloudera, R
Ads (DSP/DMP)
Amazon
Google
Netflix
10X+
Personalization
2X-10X
Realtime decision making
20-30%
Offline Data Mining (e.g., User Profiles)
10-20%
Summary Stats and Reports
Historical
Realtime
Future
Advanced BI, Regional Sales
Churn, Repeat, BigSpender
Realtime Recommendations, LookAlike Modeling
LTV

Data Scientists in Huge Demand

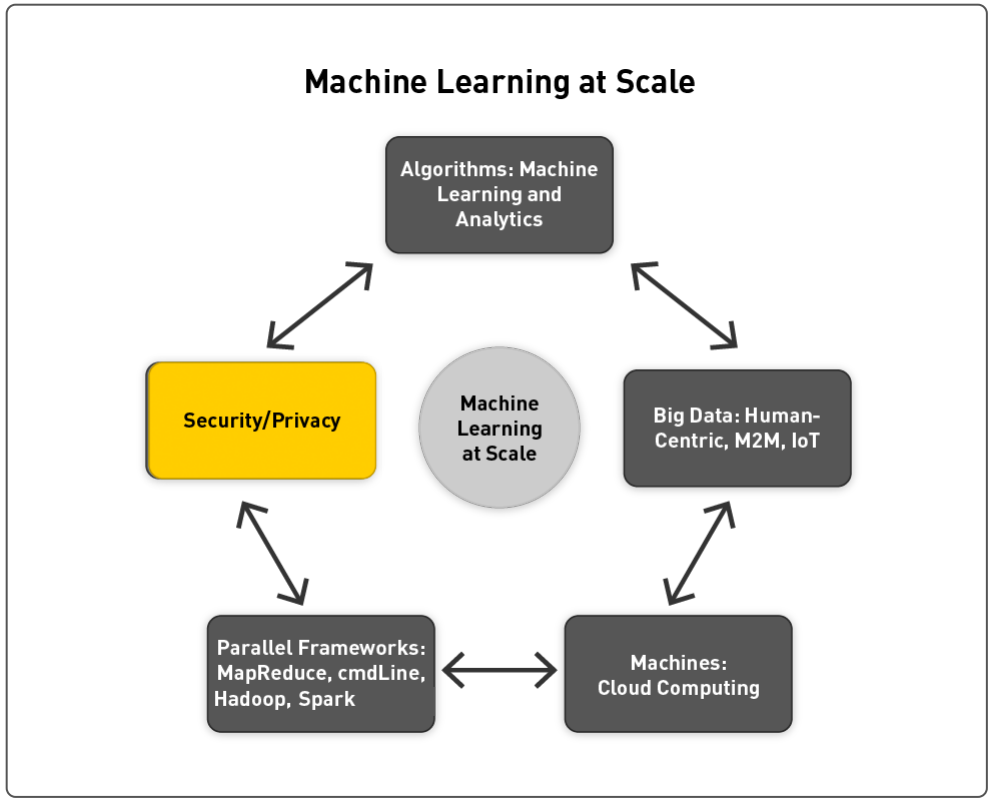
There will be over 2 million open positions in the United States.

After This Class ...

- Role
  - Individual contributors: R&D, r&D, R&d, D
  - Managers and leaders
- Focus
  - Research
    - Continue studies, get PhD
    - Teach
    - Conduct theoretical and applied research

After This Class ... (cont.)

- Focus
  - Infrastructure development: Build
  - Development
    - Architects of big data pipelines and large-scale machine learning
    - Full-stack people
    - Builders of apps (on a fully supported framework)


Machine Learning at Scale
Algorithms: Machine Learning and Analytics
Security/Privacy
Machine Learning at Scale
Big Data: Human-Centric, M2M, IoT
Parallel Frameworks: MapReduce, cmdLine, Hadoop, Spark
Machines: Cloud Computing

Large-Scale Machine Learning

- Needs supporting infrastructure and ecosystem
- Parallel computing frameworks needed

MapReduce Framework

- Allows us to divide and conquer huge data problems
- Four frameworks
    - Command-line-based approach
    - Hadoop
    - ~~MRJob~~ (This is no longer covered in the course)
    - Spark

"Embarrassingly Parallel" Problems

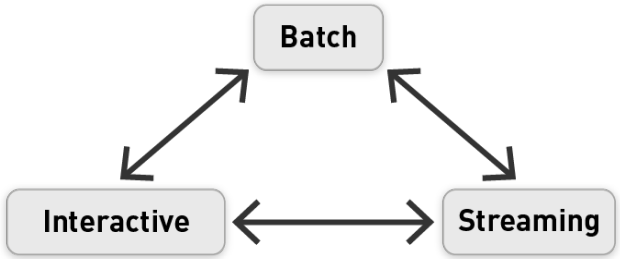
- Most machine learning problems
- Divided into subproblems, which:
  - Require little or no communication
  - Are easily distributed
  - Have linear computational flow and scale

Most machine learning algorithms are amenable to parallel computation.

Framework Requirements

- Scalable
- Fault tolerant
- Iterative
- Interactive

Few frameworks provide all these aspects



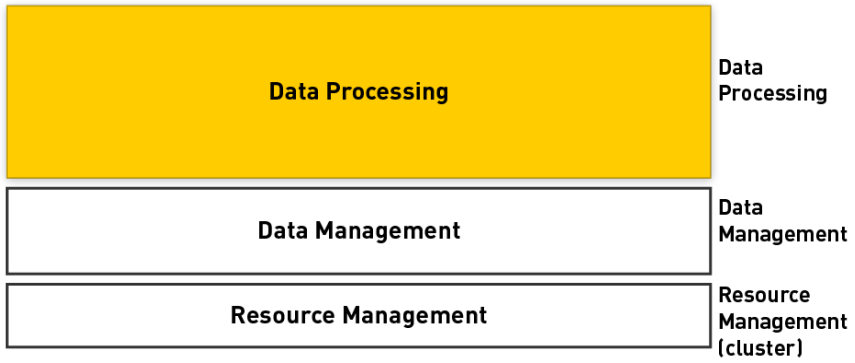
- Also compatible with existing open-source ecosystem

Four MapReduce Frameworks

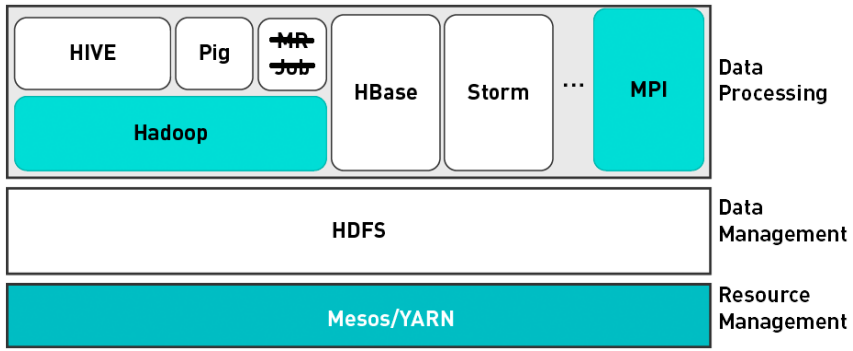
In increasing order of sophistication and ease of use

- Command-line-based approach
- Hadoop: Storing and manipulating data
- ~~MRJob~~ (This is no longer covered in the course)
- Spark:
    - Memory backed
    - Integrated framework for variety of activities
    - Specialized libraries for machine learning, graph processing, data analytics

Data Analytics Stack

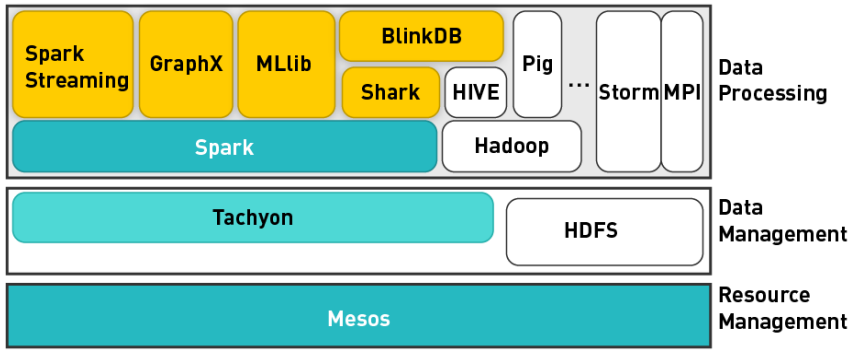


Populating the Stack

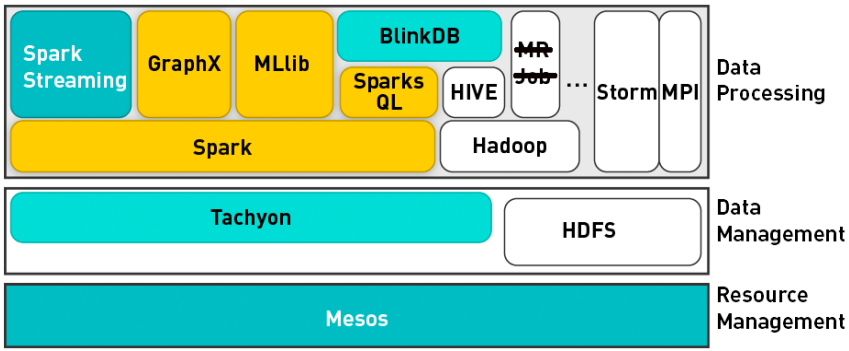


Mesos/YARN: Allow multiple frameworks on single cluster

Populating the Stack: Spark SQL



Class Phases



- Phase 0: Command line
- Phase 1: Hadoop/HDFS
- Phase 2: ~~MRJob~~ (This is no longer covered in the course)
- Phase 3: Spark

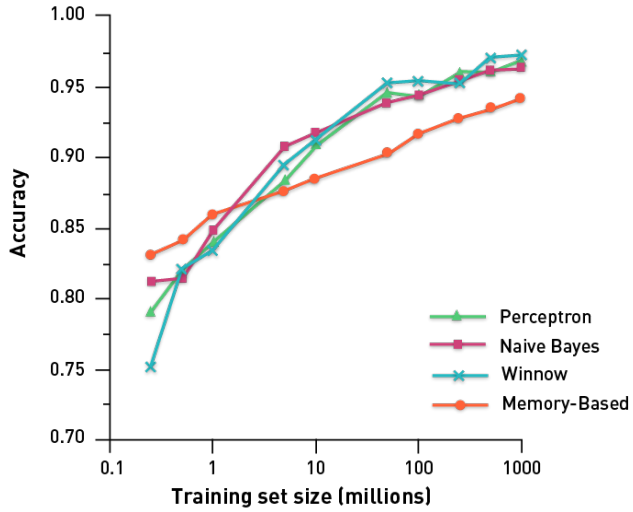
Machine Learning Algorithms

- **Supervised machine learning** (convex optimization, gradient descent, linear regression, decision trees, ensembles of models, support vector machines)
- **Unsupervised** (expectation maximization, matrix multiplication, alternating least squares)
- **Graphs** (random walks, PageRank, graph search algorithms such as breadth-first search, shortest path)
- **Hybrid algorithms** (supervised machine learning and random walks)
- **Applications** (digital advertising, social media, health care, e-commerce, entertainment, metrics, statistics)

The question: More data scientists or more data?

Machine Learning and Data Study

- Filling in confusable words in sentences
  - E.g., *For breakfast I ate ___ eggs.* Use {to, two, too}.



- Conclusion: More data leads to 10–20% boost in performance

Supervised Classification in a Nutshell

Given $D = \{(x_i, y_i)\}_i^n$ (label: y_i; (sparse) feature vector: x_i)

Induce $f : X \to Y$ s.t. loss is minimized

empirical loss = $\frac{1}{n}\sum_{i=0}^{n} \ell(f(x_i), y_i)$ (loss function: ℓ)

Consider functions of a parametric form:

$$\arg\min_{\theta} \frac{1}{n}\sum_{i=0}^{n} \ell(f(x_i;\theta), y_i)$$

(model parameters: θ)

Key insight: machine learning as an optimization problem!
(closed form solutions generally not possible)

twitter

Gradient Descent

$$w^{(t+1)} = w^{(t)} + \gamma^{(t)} \frac{1}{n} \sum_{i=0}^{n} \nabla l\left(f(\mathbf{x}_i; \theta^{(t)}), y_i\right)$$

"batch" learning: update model after considering all training instances

Stochastic Gradient Descent (SGD)

$$w^{(t+1)} = w^{(t)} + \gamma^{(t)} \nabla l\left(f(\mathbf{x}; \theta^{(t)}), y\right)$$

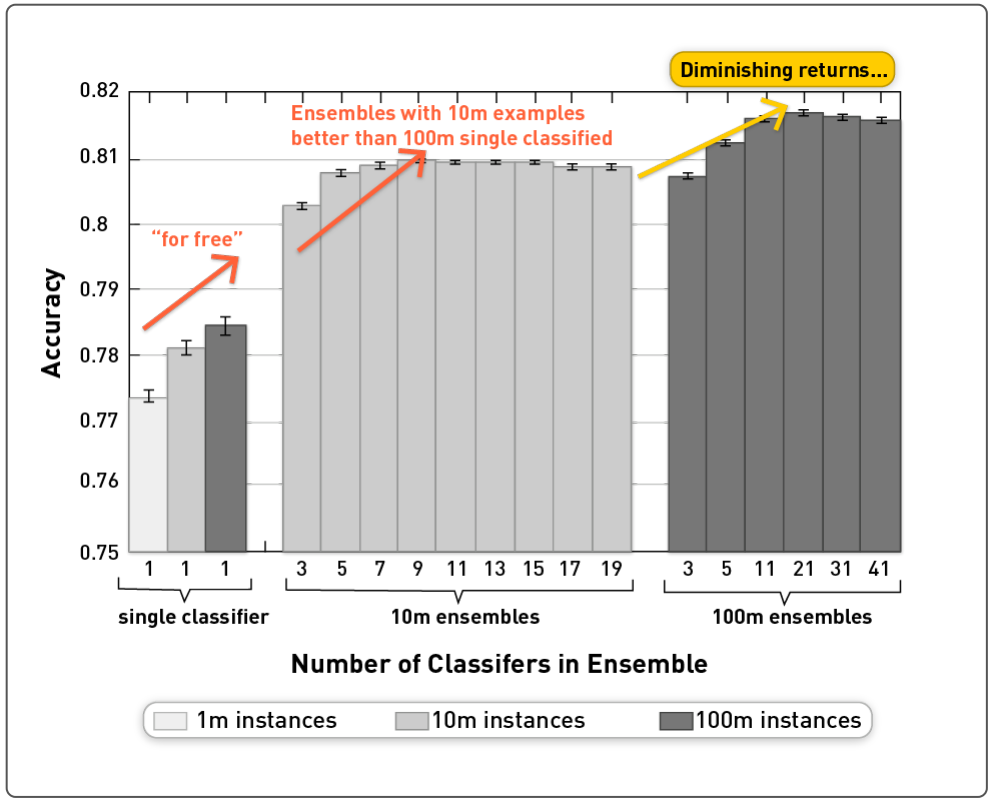
"online" learning: update model after considering each (randomly-selected) training instance

In practice... just as good!

Solves the iteration problem!
What about the single reducer problem?

twitter

Ensembles of single models can help augment our performance.


Diminishing returns...
Ensembles with 10m examples
better than 100m single classified
"for free"
Accuracy
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0.75
1 1 1
3 5 7 9 11 13 15 17 19
3 5 11 21 31 41
single classifier
10m ensembles
100m ensembles
Number of Classifers in Ensemble
1m instances
10m instances
100m instances

More Data or More Data Scientists?

- Trade-off between bias and variance
- Bias from data scientists
- Variance from more data
    - With more data can reduce variance

Bias-Variance

- Empirical studies: More data leads to big improvements
- Gigs of training data to TBs or PBs: Improved performance
- Formal perspective: Bias-variance trade-off

Machine Learning Objectives

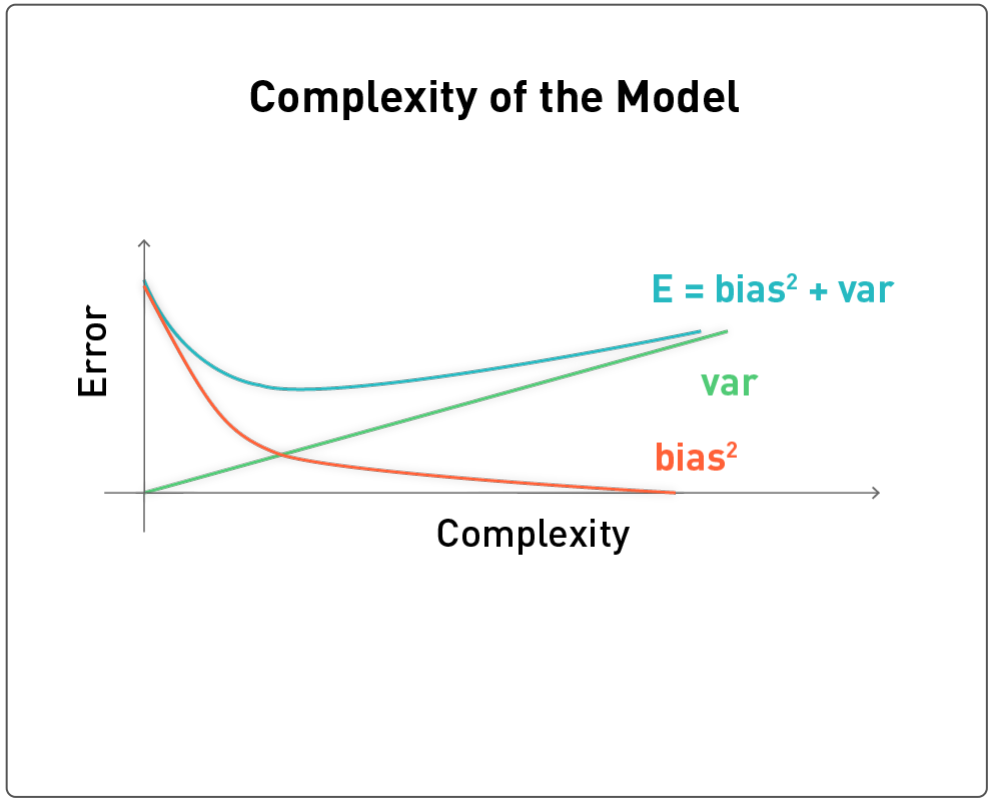
- Minimize error
- Minimize loss function
- Reduce model complexity
- Better generalization
- E.g., mean squared error for regression

Loss: Irreducible and Reducible Error

- Irreducible:
  - Inherent uncertainty
  - Associated with natural variability in system
  - E.g., noisy sensors, incorrectly documented data
- Reducible:
  - Can and should be minimized to maximize accuracy

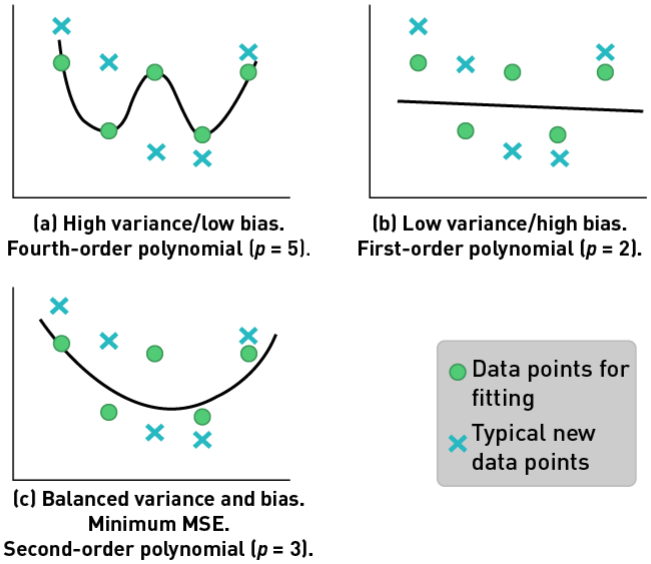
Reducible Error

- Includes both "error due to squared bias" and "error due to variance"
- Goal: Simultaneously reduce bias and variance as much as possible in order to obtain as accurate a model as is feasible.
- Trade-offs in selecting models:
  - Flexibility and complexity
  - Selecting appropriate training sets


Complexity of the Model
Error
E = bias² + var
var
bias²
Complexity

Usually, bias is a decreasing function of complexity, while variance is an increasing function of complexity.

Bias-Variance Trade-Off in Model Selection in Simple Problem



Error Due to Squared Bias

- Definition: Amount by which expected model prediction differs from the true value, over the training data
- Bias is introduced at model selection
- Repeat model-building process (through resampling) to obtain average of prediction values
- If average prediction values are substantially different than true value, bias will be high

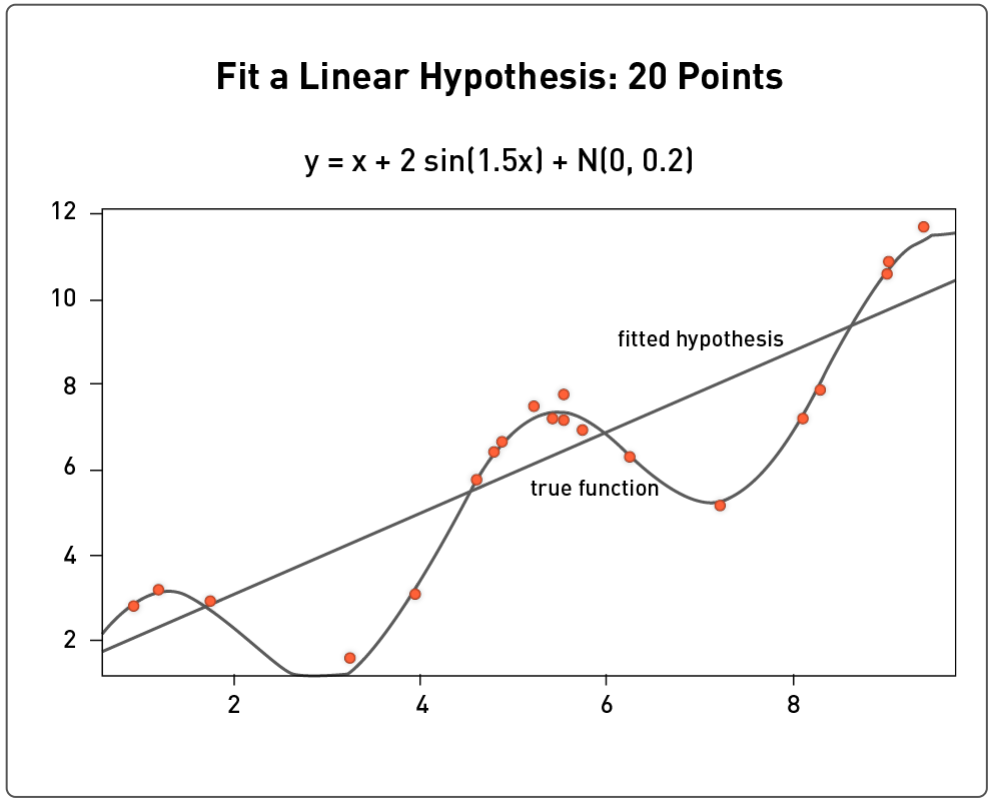
Error Due to Variance

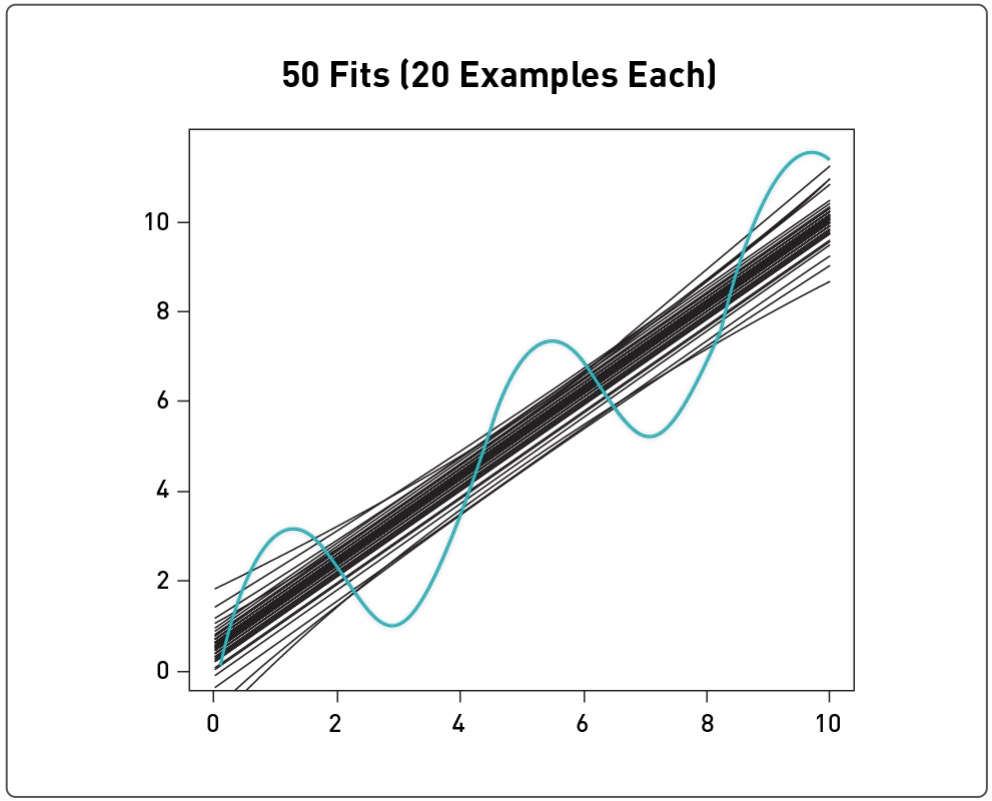
- Definition: Amount by which prediction over one training set differs from expected predicted value over all training sets
- Repeat modeling, measure inconsistency over different training data sets
- Variance measures how inconsistent predictions are from one another, over different training sets, not whether or not they are accurate (Manning et al., 2008)

Bias-Variance Analysis in Regression

- True function: $y = f(x) + \varepsilon$
  - Where ε is normally distributed with zero mean and standard deviation σ.
- Given a set of training examples, {(xi,yi)}, we fit a hypothesis $h(x) = wx + b$ to the data to minimize the squared error.

$$\sum_i [y_i - h(x_i)]^2$$


Fit a Linear Hypothesis: 20 Points
y = x + 2 sin(1.5x) + N(0, 0.2)
12
10
8
6
4
2
2
4
6
8
fitted hypothesis
true function


50 Fits (20 Examples Each)
10
8
6
4
2
0
0
2
4
6
8
10

Variance, Bias, and Noise

- Variance:

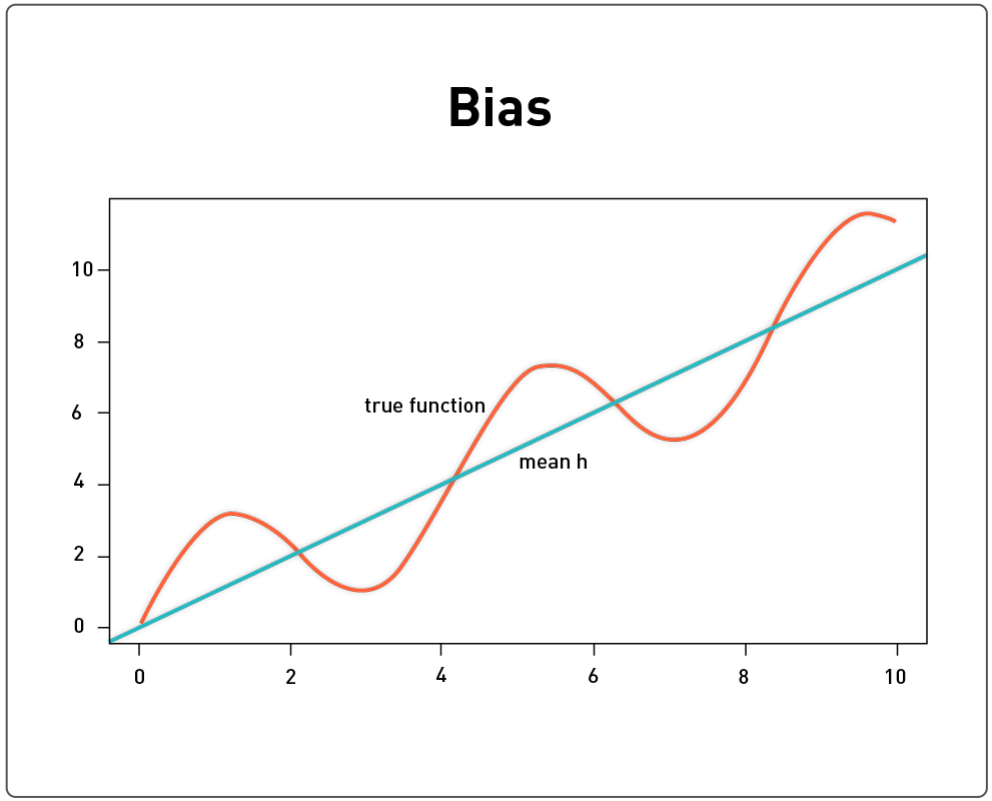
$$Eh(x^*) - E[h(x^*)]\ 2$$

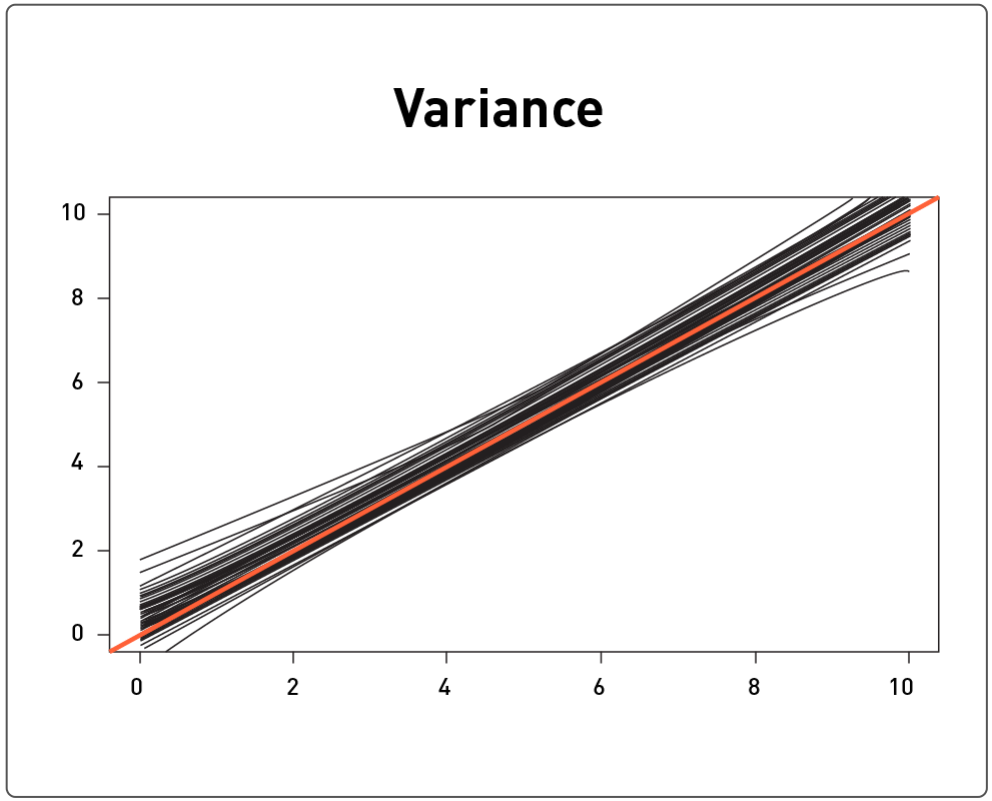
- Bias:

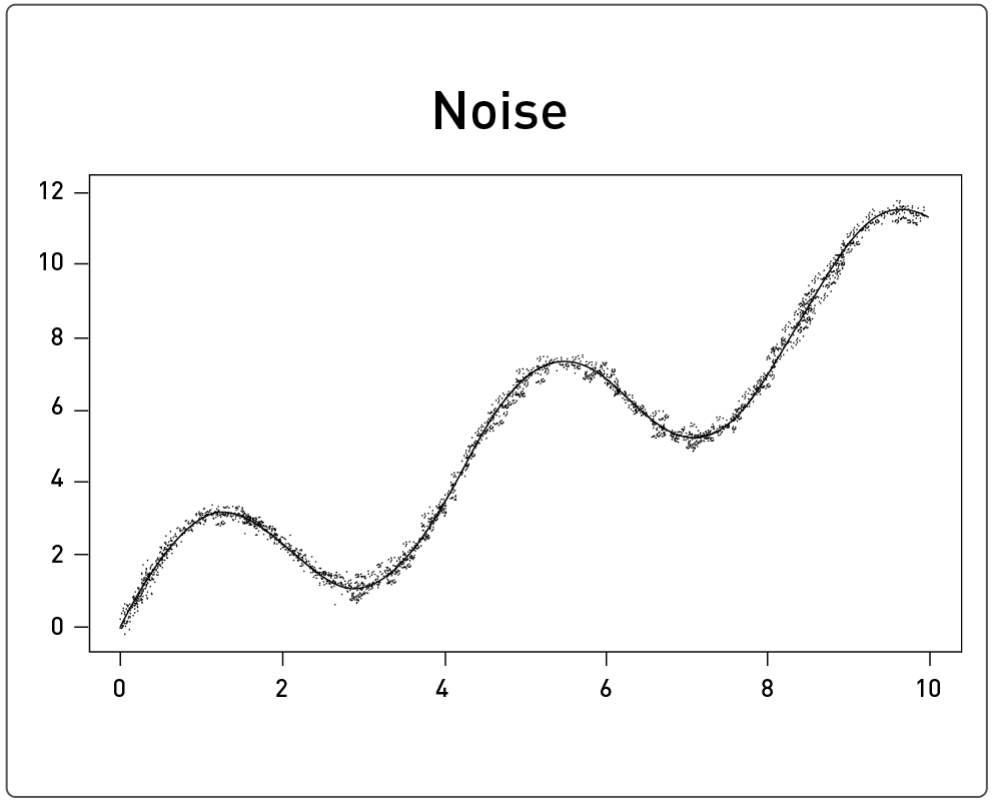
$$E[h(x^*)] - f(x^*)$$

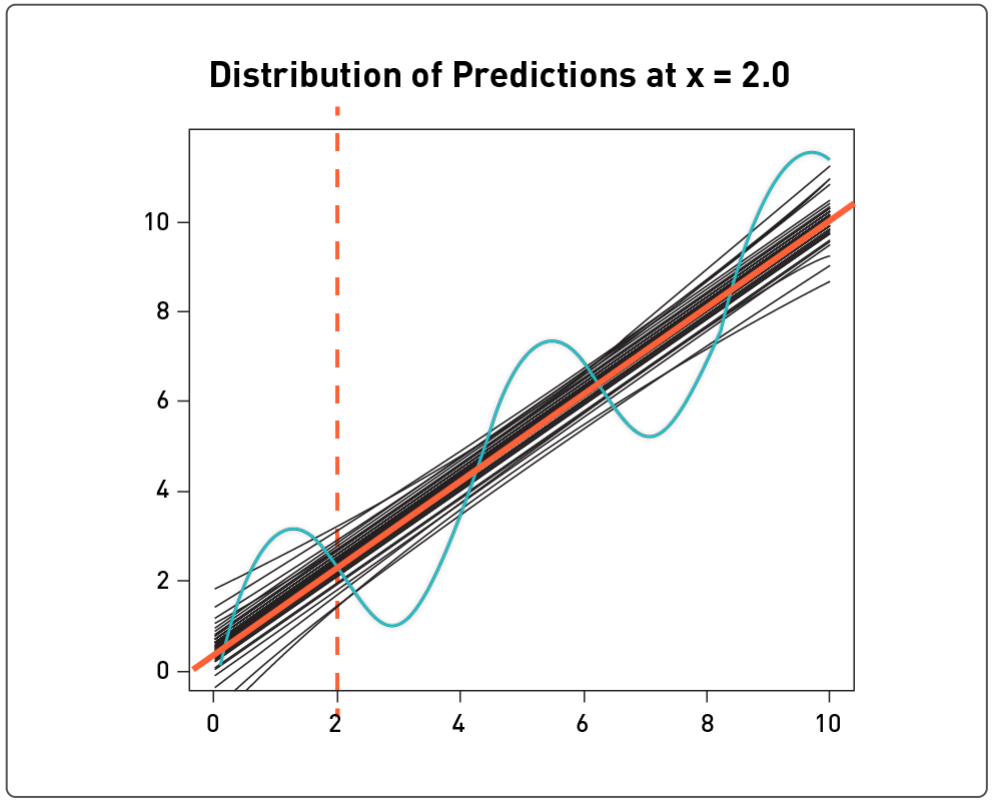
- Noise:

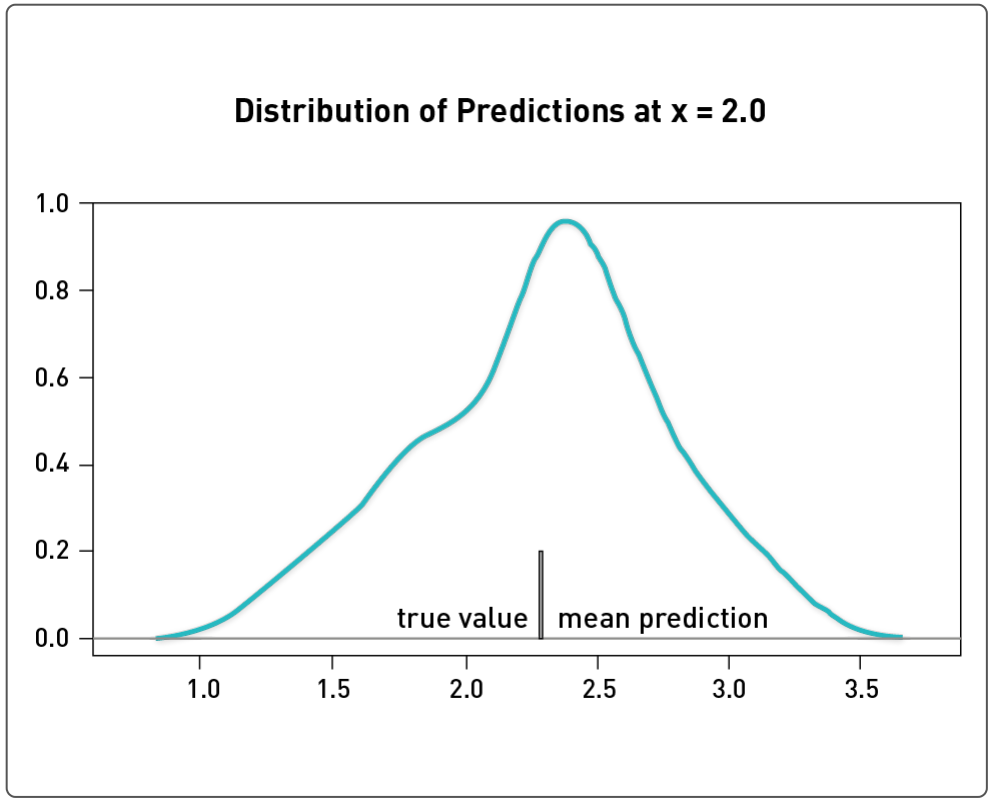
$$E[(y^* - f(x^*))2] = E[\varepsilon 2] = \sigma 2$$

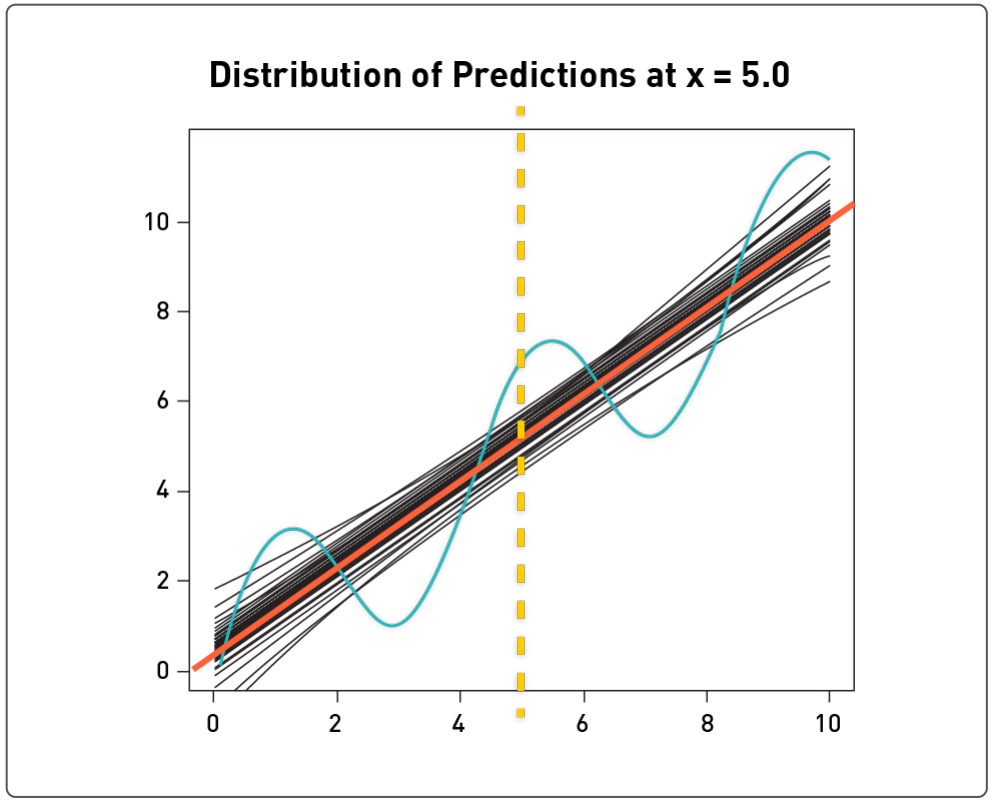

Bias
10
8
6
4
2
0
0
2
4
6
8
10
true function
mean h

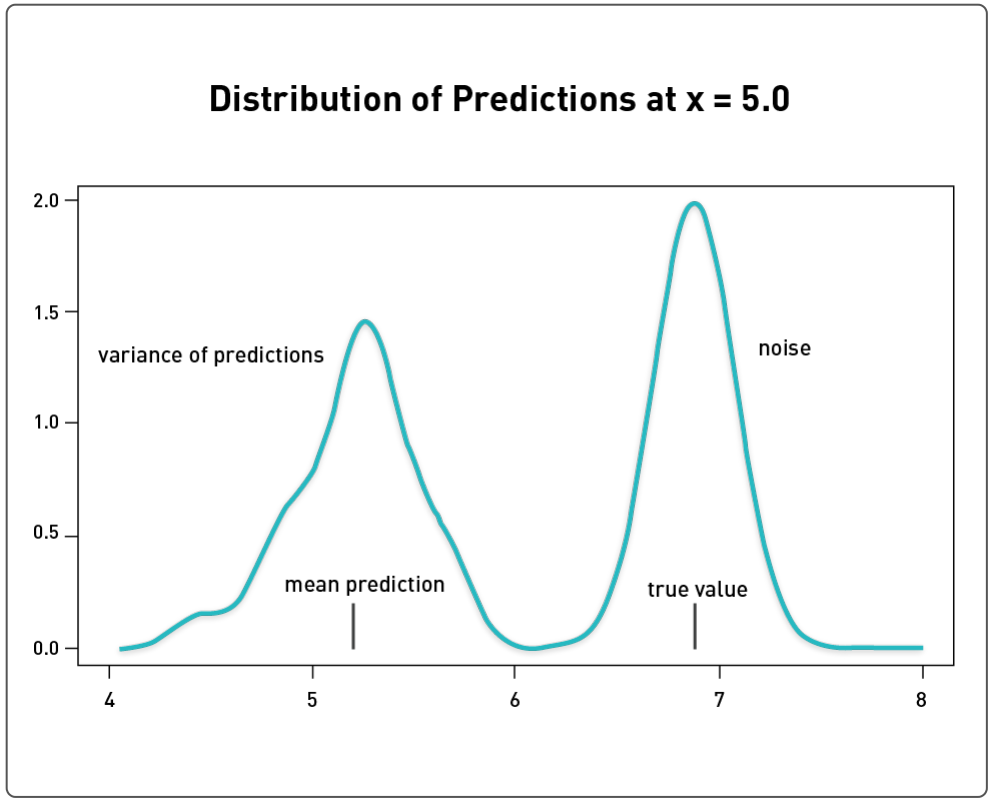

Variance
10
8
6
4
2
0
0
2
4
6
8
10


Noise
12
10
8
6
4
2
0
0
2
4
6
8
10


Distribution of Predictions at x = 2.0
10
8
6
4
2
0
0
2
4
6
8
10

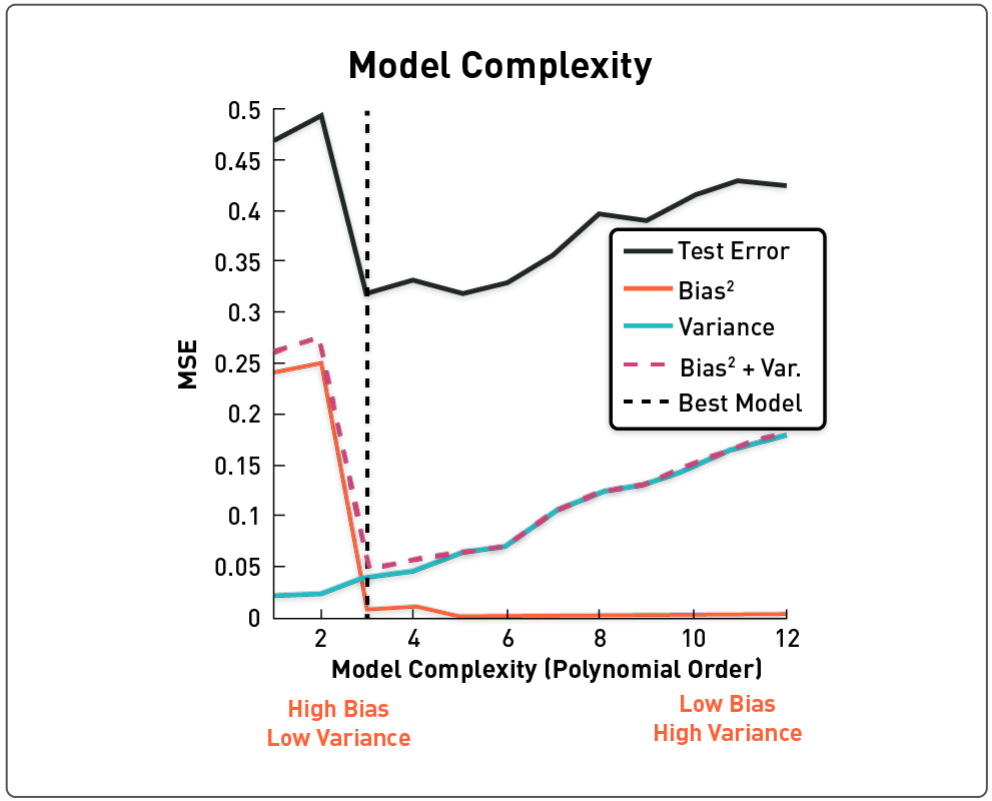

Distribution of Predictions at x = 2.0
1.0
0.8
0.6
0.4
0.2
0.0
true value
mean prediction
1.0
1.5
2.0
2.5
3.0
3.5


Distribution of Predictions at x = 5.0
10
8
6
4
2
0
0
2
4
6
8
10


Distribution of Predictions at x = 5.0
2.0
1.5
1.0
0.5
0.0
variance of predictions
noise
mean prediction
true value
4
5
6
7
8

Measuring Bias and Variance

- Simulate multiple training sets by bootstrap replicates.
- Construct *B* bootstrap replicates of *S* (e.g., *B* = 200).
  - S1, ⋯ SB
- Apply learning algorithm to each replicate Sb to obtain hypothesis hb.
- Let Tb=S\Sb be the data points that do not appear in Sb.
- Compute predicted value hb(x) for each *x* in Tb.


Model Complexity
MSE
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
2
4
6
8
10
12
Model Complexity (Polynomial Order)
Test Error
Bias²
Variance
Bias² + Var.
Best Model
High Bias
Low Variance
Low Bias
High Variance

Trade-off between bias and variance: Best to minimize both at the same time

Bias-Variance Decomposition

- Can be extended to classification problems
- Pedro Domingos (2000a; 2000b): Developed unified decomposition that covers both regression and classification

Bias-Variance Trade-Off Code: Articles

Excellent articles with code in matlab for bias-variance analysis of squared error loss:

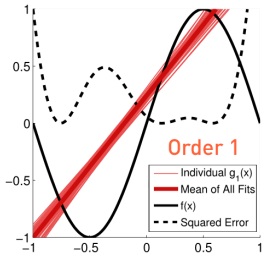
- Model Selection: Underfitting, Overfitting, and the Bias-Variance Tradeoff
- Ask a Data Scientist: The Bias vs. Variance Tradeoff
- polyfit: Polynomial curve fitting

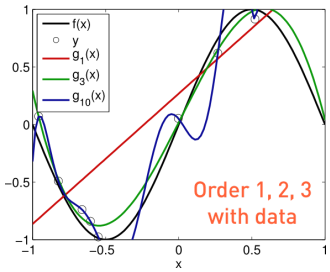
Bias-Variance Order: 1, 2, 3 Polynomials

xGrid = linspace(−1, 1, 100);

f(x) = sin(πXx)

```
% FIT MODELS TO K INDEPENDENT DATASETS
K = 50;
for iS = 1:K
  ySim = f(x) + noiseSTD*randn(size(x));
  for jD = 1:numel(degree)
    % FIT THE MODEL USING polyfit.m
    thetaTmp = polyfit(x,ySim,degree(jD));
    % EVALUATE THE MODEL FIT USING polyval.m
    simFit{jD}(iS,:) =
polyval(thetaTmp,xGrid);
  end
end

% DISPLAY ALL THE MODEL FITS
h = [];
for iD = 1:numel(degree) For polynomial = iD
  figure(iD+1)
  hold on
  % PLOT THE FUNCTION FIT TO EACH DATASET
  for iS = 1:K
    h(1) = plot(xGrid,simFit{iD}
(iS,:),'color',
  brighten(cols(iD,:),.6));
  end
  % PLOT THE AVERAGE FUNCTION ACROSS ALL FITS
  Average predication of xGrid
  h(2) = plot(xGrid,mean(simFit{iD}),
  'color',cols(iD,:),'Linewidth',5);
  % PLOT THE UNDERLYING FUNCTION f(x)
  h(3) = plot(xGrid,f(xGrid),'color','k','Linewidth',3);
  % CALCULATE THE SQUARED ERROR AT EACH POINT, AVERAGED ACROSS ALL DATASETS
  BIAS
  squaredError = (mean(simFit{iD})-f(xGrid)).^2;
  % PLOT THE SQUARED ERROR
  h(4) = plot(xGrid,squaredError,'k--','Linewidth',3);
  uistack(h(2),'top')
  hold off
  axis square
  xlim([-1 1])
  ylim([-1 1])
  legend(h,{springf('Individual g_{%d}(x)',degree(iD)),
  'Mean of All Fits','f(x)','Squared Error'},'Location','WestOuts…
  title(sprintf('Model Order=%d',degree(iD)))
end
```

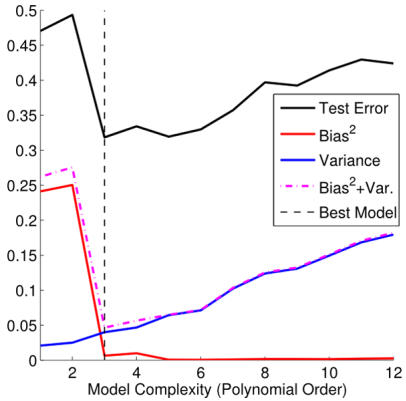




Our original goal was to approximate f(x), not the data points, per se.

Bias-Variance Order: 1 to 12 Polynomials

```
% # INITIALIZE SOME VARIABLES
xGrid = linspace(-1,1,N);
meanPrediction = zeros(K,N);
thetaHat = {};
x = linspace(-1,1,N);
x = x(randperm(N));
for iS = 1:K % LOOP OVER DATASETS
  % CREATE OBSERVED DATA, y
  y = f(x) + noiseSTD*randn(size(x));

  % CREATE TRAINING SET
  xTrain = x(1:nTrain);
  yTrain = y(1:nTrain);

  % CREATE TESTING SET
  xTest = x(nTrain+1:end);
  yTest = y(nTrain+1:end);

  % FIT MODELS
  for jD = 1:nPolyMax

    % MODEL PARAMETER ESTIMATES
    thetaHat{jD}(iS,:) = polyfit(xTrain,yTrain,jD);

    % PREDICTIONS
    yHatTrain{jD}(iS,:) = polyval([thetaHat{jD}(iS,:)],xTrain); % TRAINING
SET
    yHatTest{jD}(iS,:) = polyval([thetaHat{jD}(iS,:)],xTest); % TESTING SET

    % MEAN SQUARED ERROR
    trainErrors{jD}(iS) = mean((yHatTrain{jD}(iS,:) - yTrain).^2); %
TRAINING
    testErrors{jD}(iS) = mean((yHatTest{jD}(iS,:) - yTest).^2); % TESTING
  end
end          Test Error=Variancexi+Biasxi+irreducibleError

% CALCULATE AVERAGE PREDICTION ERROR, BIAS, AND VARIANCE
for iD = 1:nPolyMax
  trainError(iD) = mean(trainErrors{iD});
  testError(iD) = mean(testErrors{iD});
  biasSquared(iD) = mean((mean(yHatTest{iD})-f(xTest))x^2); Avg(Bias (xi))
  variance(iD) = mean(var(yHatTest{iD},1));                 Avg(Variance (xi))
end
[~,bestModel] = min(testError);
```



Model Selection

- Trade-off between bias and variance
- Practical method for selecting model:
  - Minimize error
  - Function of model complexity

Effect of Algorithm Parameters

- K-Nearest Neighbor: Increasing k typically increases bias, reduces variance
- Decision trees of depth D: Increasing D typically increases variance, reduces bias
- Radial basis function (RBF) support vector machines (SVM) with parameter σ: Increasing σ increases bias, reduces variance

Bagging tends to reduce variance without changing bias.

Wrap-Up: Bias-Variance Trade-Off

- Bias and variance of an estimator are related to squared prediction error.
- These concepts can be applied to classification problems.
- An optimal estimator will have both low variance and low bias.

Parallel Sort

```
#! /bin/ksh

MAX_LINES_PER_CHUNK=1000000
ORIGINAL_FILE=$1
SORTED_FILE=$2
CHUNK_FILE_PREFIX=$ORIGINAL_FILE.split.
SORTED_CHUNK_FILES=$CHUNK_FILE_PREFIX*.sorted

usage ( )
{
    echo Parallel sort
    echo usage: psort file1 file2
    echo Sorts text file file1 and stores the output in file2
    echo Note: file1 will be split in chunks up to
    $MAX_LINES_PER_CHUNK lines
    echo and each chunk will be sorted in parallel
}

# test if we have two arguments on the command line
if [ $# != 2 ]
then
    usage
    exit
fi

#Cleanup any leftover files
rm -f $SORTED_CHUNK_FILES > /dev/null
rm -f $CHUNK_FILE_PREFIX* > /dev/null
rm -f $SORTED_FILE

#Splitting $ORIGINAL_FILE into chunks ...
split -1 $MAX_LINES_PER_CHUNK $ORIGINAL_FILE
$CHUNK_FILE_PREFIX

for file in $CHUNK_FILE_PREFIX*
do
    sort $file > $file.sorted &
done
wait

```

```
#Merging chunks to $SORTED_FILE ...
sort -m $SORTED_CHUNK_FILES > $SORTED_FILE

#Cleanup any leftover files
rm -f $SORTED_CHUNK_FILES > /dev/null
rm -f $CHUNK_FILE_PREFIX* > /dev/null
```

General Framework for Mapping and Reducing

```
%%writefile cmdLineMapReduce.sh
ORIGINAL_FILE=$1
BLOCK_SIZE=$2
MapperParameters=$* #note $* will include the first and second parameter also
CHUNK_FILE_PREFIX=$ORIGINAL_FILE.split
SORTED_CHUNK_FILES=$CHUNK_FILE_PREFIX*.sorted
usage ( )
{
    echo Command Line Map Reduce
    echo usage: cmdLineMapReduce filename chunksize mapperParameters
    echo Note: filel will be split in chunks up to $BLOCK_SIZE chunks each
    echo NOTE: please make sure the output of the mapper is what the reducer expects
}
#Splitting $ORIGINAL_FILE INTO CHUNKS
split -b $BLOCK_SIZE $ORIGINAL_FILE $CHUNK_FILE_PREFIX
#DISTRIBUTE
for file in $CHUNK_FILE_PREFIX*
do
    #grep -i $FIND_WORD $file|wc -1 >$file.intermediateCount &
    ./mapper.py $MapperParameters $file >$file.intermediateCount &
done
wait
#MERGING INTERMEDIATE COUNT CAN TAKE THE FIRST COLUMN AND TOTAL
#numOfInstances=$(cat *.intermediateCount | cut -f 1 | paste &minussd+ &minus- |bc)
reducerOutput=$(cat *.intermediateCount | ./reducer.py)
echo "$reducerOutput"
```

Course Overview

- How to quantify data
- How to estimate time
- Powers of 10
- Parallelism
- Scale mundane tasks like grep, sort

Challenges in MapReduce Framework

- Data management
- Task management
- Fault tolerance
- Lack of interactivity

A Better Framework

- Scale out, not up
- Fault tolerance
- Minimize data movement
- Move processing to the data
- Hide file system details from application developer
- Seamless scalability

Class Phases

- Phase 0: Command line
- Phase 1: Hadoop/HDFS
- Phase 2: MRJob
- Phase 3: Spark

Summary

- Big data (what, why, where, who, how)
- Role of data scientist
- Data modeling pipeline
- Bias variance as a means of understanding more data
- MapReduce framework using command-line utility