Big Data: A Public Health Perspective

Big Data Lectures – Chapter 1

Lexin Li

Division of Biostatistics
University of California, Berkeley

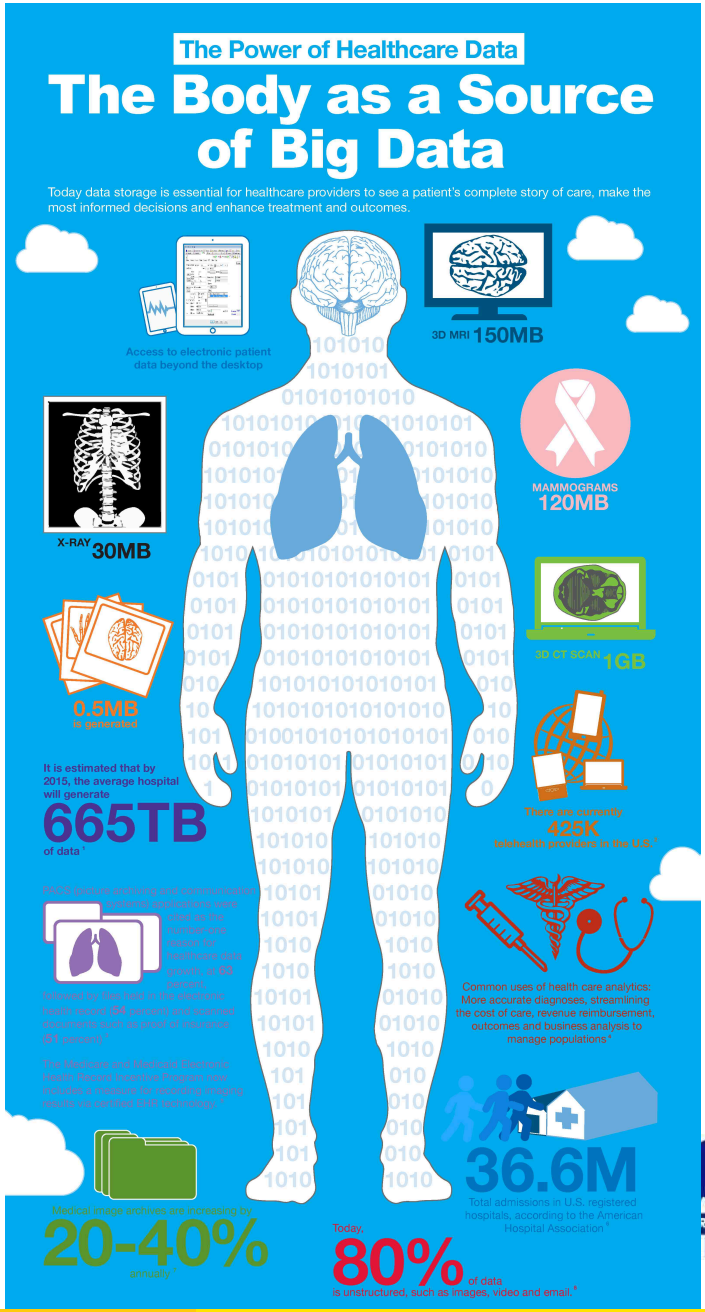
Outline

- list of topics:
  - Big Data: introduction
    - Big Data examples
    - curriculum evolution (now and then, stat vs cs)
    - what skills are demanded and what are emphasized
    - what this course offers
  - statistical concepts and principles
    - uncertainty, statistical distribution, and p-value
    - association and causation
    - supervised learning vs unsupervised learning
    - model interpretation and prediction
    - model evaluation and selection: training error, testing error, cross-validation
    - bias-variance tradeoff
  - computing: distributed storage and processing
    - Hadoop Distributed File System
    - MapReduce parallel computing
  - art of data analysis — a practical guide

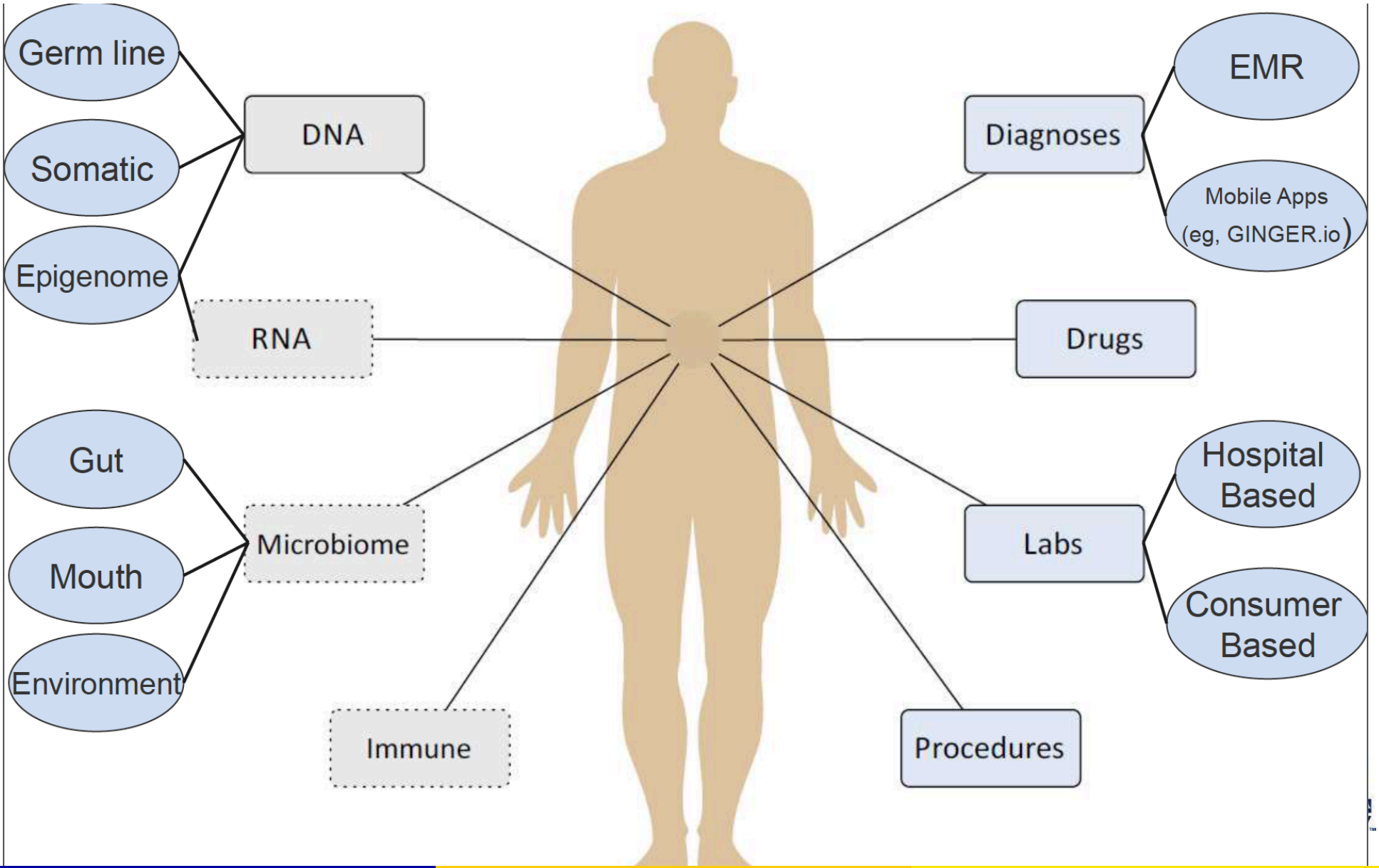
Big Data: Introduction

Big Data

- Big Data is **everywhere**...
  - McKinsey report (2011-05): *Big Data: the next frontier for innovation, competition, and productivity*
  - Wall Street Journal (2012-11-29) *Big Data is on the rise, bringing big questions*
  - a **fever** about Big Data



Big Data examples



Big Data examples

- EHR (electronic health record) / EMR (electronic medical record)
- Million Veteran Program (MVP): EHR of 8.5 million enrollees, volunteer-based, exon- and whole genome sequencing, millions of single genetic variants
- Obama's Digital Health Data in a Million-Person Precision Medicine Initiative (PMI) Cohort
- The Cancer Genome Atlas Project (TCGA)
- Mobile apps: A lot of data is getting collected by something like 50,000-100,000 mobile apps that in one way or another relate to health.Record and track your health. e.g. Asthma Health by Mount Sinai

Daytime and nighttime asthma symptoms
Daily usage of controller and rescue inhalers
Triggers such as colds, increased physical activity, strong smells, exhaust fumes, house dust, and animals
Emergency department visits, medical visits, and changes in medication

Reminders of Your medications, Local air quality

Big Data more examples

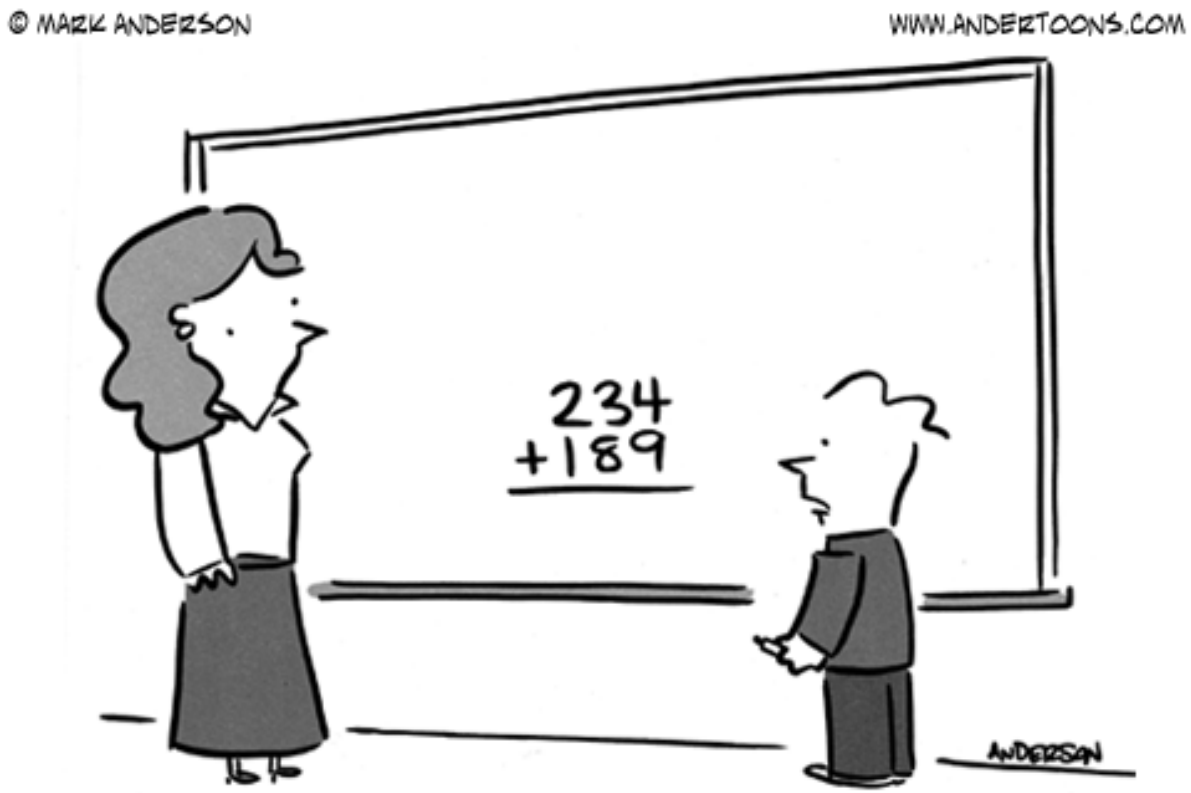
"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Big Data more examples

"Remember, the other team is counting on Big Data insights based on previous games. So, kick the ball with your other foot."

Big Data

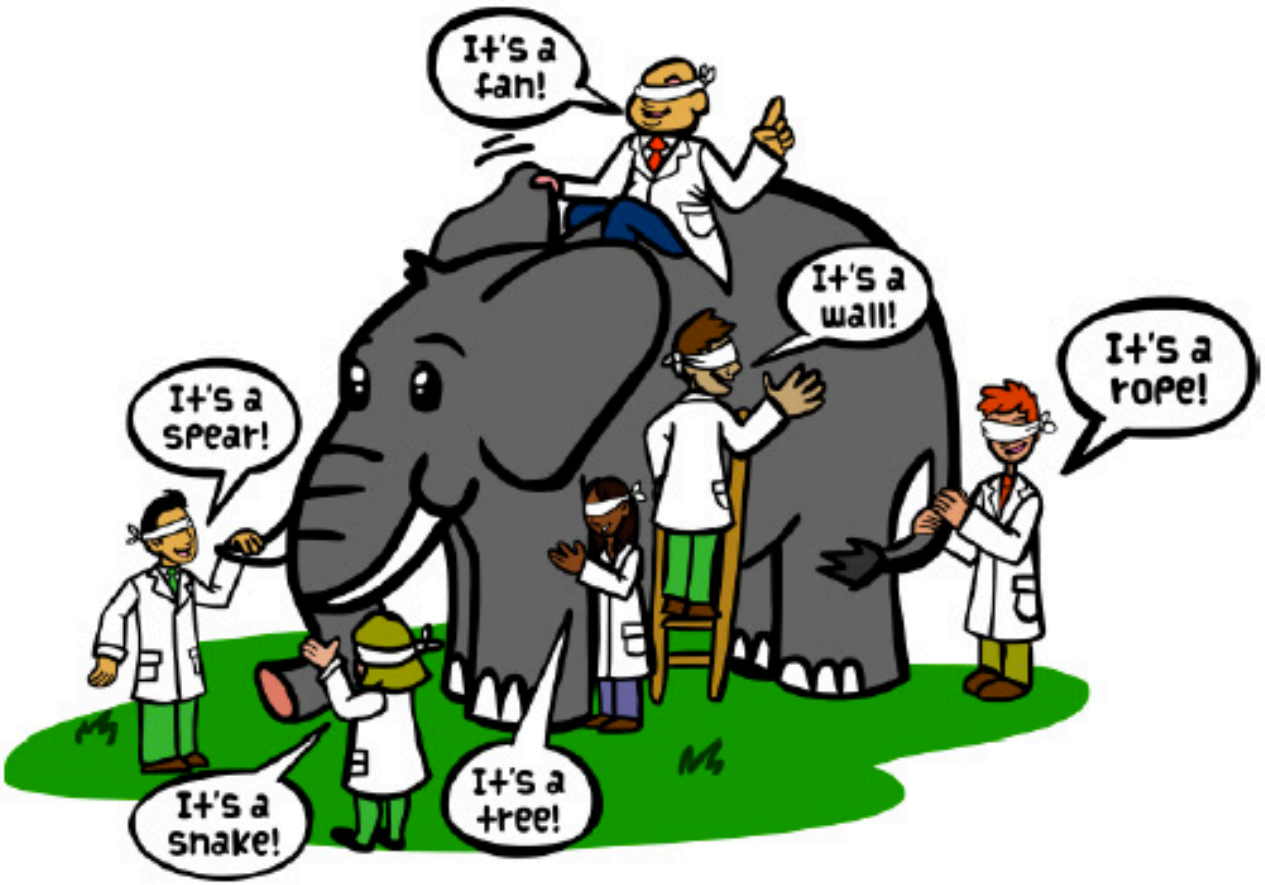
- what is Big Data?



"Does this count as big data?"

Big Data

- what is Big Data?



Big Data

- what is Big Data?
  - "There are 1,000 Hamlets in a thousand people's eyes." — William Shakespeare

Big Data

- what is Big Data?
  - "There are 1,000 Hamlets in a thousand people's eyes." — William Shakespeare
  - There are 1,000 views of Big Data in a thousand people's eyes.

Big Data

- what is Big Data?
  - "There are 1,000 Hamlets in a thousand people's eyes." — William Shakespeare
  - There are 1,000 views of Big Data in a thousand people's eyes.

- Big Data refers to the idea:
  - NSF Big Data Initiative, 2012: "that scientists **manage, analyze, visualize, and extract useful information** from **large, diverse, distributed and heterogeneous** data sets so as to accelerate the progress of scientific discovery and innovation."
  - EMC^2: "the foundation for creating new levels of **business value**, helping drive efficiency, quality, and personalized products and services, producing higher levels of customer satisfaction, with **integrated storage, analytics, and applications**"

Big Data

- a completely new thing?
  - Wal-Mart made $>$ 20 million transactions daily, and constructed an 11 terabyte database of customer transactions
  - AT&T had 100 million customers and carried on the order of 300 million calls a day on its long distance network
  - "Databases today can range in size into the terabytes – more than $1,000,000,000,000$ bytes of data. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do you draw meaningful conclusions about the forest?"

Big Data

- a completely new thing?
  - Wal-Mart made $>$ 20 million transactions daily, and constructed an 11 terabyte database of customer transactions – **year 1998**
  - AT&T had 100 million customers and carried on the order of 300 million calls a day on its long distance network – **year 1998**
  - "Databases today can range in size into the terabytes – more than $1,000,000,000,000$ bytes of data. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do you draw meaningful conclusions about the forest?" – **year 1999 Tech Report**

Big Data

- a completely new thing?
  - Wal-Mart made $>$ 20 million transactions daily, and constructed an 11 terabyte database of customer transactions – **year 1998**
  - AT&T had 100 million customers and carried on the order of 300 million calls a day on its long distance network – **year 1998**
  - "Databases today can range in size into the terabytes – more than $1,000,000,000,000$ bytes of data. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do you draw meaningful conclusions about the forest?" – **year 1999 Tech Report**
  - **"The newest answer is data mining"** – same 1999 Tech Report *Introduction to Data Mining and Knowledge Discovery*

Big Data

- curriculum evolution: then vs now; cs vs stat

then		now	
cs[1]	stat[2]	cs[3]	stat
data warehouse	regression models	distributed system	
OLAP (online analytical proc)	lasso, ridge, PCA, PLS	Map-Reduce, Hadoop	?
association rules	splines, kernel smooth	association, freq items	?
classification	CART MARS GAM	PageRank, link analysis	?
clustering	boosting	clustering	
prediction model	classification, SVM	SVD, dim reduction	
text mining	neural networks	machine learning	
multimedial mining	clustering	online advertisement	
transactional db*	$p >> n$**	recommendation sys	
social networks*	network models**	social networks	

- cs[1]: Han, J. and Kamber, M. (2000). *Data Mining: Concepts and Techniques*, 1st edition [*(2006), 2nd edition]
- stat[2]: Hastie, T., Tibshirani, R., and Friedman, J. (2001). *Elements of Statistical Learning*, 1st edition [**(2009), 2nd edition]
- cs[3]: Rajaraman, A., Leskovec, J., and Ullman, J.D. (2012+). *Mining of Massive Datasets*, manuscript

What is important

- key skill set: ability to turn **data** into **information** then into **action**
  - a deep understanding of the science or business to **ask the right questions** – usually the hardest part
  - **statistical skills** to build appropriate models given the massive, complicated and usually very messy data
    - general but useful statistical principles and statistical thinking
    - statistical and data mining methods for data with big size and high dimensionality
  - **engineering skills** to carry out all operations – consider the data size
    - computing
    - programming
  - capability to **find insights** and **tell stories** from the data – present the results super clearly and concisely

What this course offers (what's different)

- *Big Data's big problem: little talent* (WSJ, 04-29-2012)

  the courses for big data don't yet exist in universities, "though bits of it do exist in various university departments and businesses, as an integrated discipline it is only just starting to emerge"

  my take:

- course focus: how to better prepare us for academic and industrial jobs involving Big Data

- 3 major components of this course:
  - statistics component
  - computing component
  - class projects and Big Data applications in public health

Statistics component

- topics to cover:
  - hypothesis testing; A/B testing; causal inference
  - linear regression and logistic regression: basics; inference under large n; sampling; divide-and-conquer; parallel computing — handling large n
  - classification: discriminant analysis; nearest neighbor; classification and regression tree; linear support vector machine
  - regularization — handling large p
  - basis expansion: splines; wavelet; kernel methods
  - ensemble: neural networks and deep learning; boosting; random forest
  - recommendation systems
  - network and graphical models
  - dimension reduction: principal components analysis; independent components analysis; nonnegative matrix factorization; multidimensional scaling
  - unsupervised learning: clustering; association analysis
  - neuroimaging analysis

Statistics component

- topics to cover:
  - "as usual" part: **data mining and machine learning** methods, e.g., SVM, PCA, lasso, boosting, ...
  - some relatively new and important topics, e.g., recommendation system, network analysis
- emphasis:
  - emphasize on general but useful **statistical principles and thinking**
  - emphasize on statistical **models** – a parsimonious and interpretable model is favored over a black-box type learning algorithm; how to properly formulate a statistical model; how to develop intuitive **insights and interpretation**
  - emphasize on statistical inferences – how to evaluate the **uncertainty** of the outcome
  - for each method to introduce: what is it; what's the intuition behind; how is it **connected** to other methods

Computing component

- topics to cover:
  - distributed storage system: Hadoop Distributed File System
  - parallel computing paradigm: MapReduce
  - data visualization in R: `ggplot`
  - Big Data analysis and programming in R: `biglm`, `bigmemory`, `pbdR`, `doParallel`
  - Tensor Flow; Spark; Amazon Web Services*
- programming:
  - R — a must but not enough
  - Python* — commonly used in data analysis
- optimization*
  - optimization techniques, complexity analysis, scalability, ...

Class projects

- **Big Data applications in public health:**
  - project I: **electronic health record**
    - Heritage Health Prize
    - Predict HIV Progression
  - project II: **mobile health**
    - Predicting Parkinson's Disease Progression with Smartphone Data
    - Accelerometer Biometric Competition
  - project IV: **omics/imaging**
    - Second Annual Data Science Bowl: Transforming How We Diagnose Heart Disease
    - Data Science Bowl 2017: Can You Improve Lung Cancer Detection?

- data repository: `www.kaggle.com`
  — register, and search under "Competitions"

Class projects

- project due dates:

	Project	Topic	Due date
(1)	Project I	EHR	February 12
(2)	Project II	mobile health	March 12
(3)	Project III	computing	April 9
(4)	Project IV	omics/imaging	May 7

- data analysis project:
  - under each project category, **choose one data** to work on — please coordinate with me
  - ask questions: (a) the competition question; (b) a question of your own
  - data processing: handle and clean the real world big and messy data
  - data analysis: (a) summarize the winning methods; (b) propose your own solution and do a comparison; (c) carry out an analysis to address your own question

Class projects

- written report:
  - the report is **no more than 3 pages**, **including** tables and figures (and references, everything)
  - at the beginning of the report, prepare a summary of your findings (what kind of useful knowledge you have obtained via your analysis)
  - make sure to include a description on how you process the data
- oral presentation:
  - there is an oral presentation in class in the week of April 23 to 25; you can choose to present any one of the 3 data analysis projects you have done — please coordinate with me
- goals to achieve through the data analysis projects:
  - gain understanding of important Big Data applications in public health
  - gain experience of processing **big** and **messy** real world data
  - push for a super **concise and clear** presentation

Course syllabus

- instructor info: lexinli@berkeley.edu; 344B Li Ka Shing Center
- prerequisite: PB HLTH 142 (basic concepts of probability and distributions, point and interval estimation, hypothesis testing),145 (regression analysis of continuous outcome), 241 (categorical data analysis, some modern statistical learning techniques), or equivalent. If you have any concerns about the prerequisites, **please come talk to me right after the first class.**
- homework, exam and project: there are totally 3 data analysis projects + 1 computing project; there is no other homework or exam.
- grading: four projects 20% each + oral presentation 20%

Course syllabus

- recommended books:
  - *The Elements of Statistical Learning – Data Mining, Inference and Prediction*, Trevor Hastie, Robert Tibshirani, and Jerome Friedman, 2009, 2nd edition, Springer
  - *Mining of Massive Datasets*, Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman, 2014, Stanford University

- learning objectives: by the end of the semester, students will be expected
  - to have an in-depth understanding of both key statistical and analytical principles as well as a wide range of data mining and machine learning techniques
  - to gain some experience with modern optimization and computing paradigms
  - to have a good understanding of some key Big Data applications in public health, including electronic health record, omics, mobile health, among others

Statistical Principles

Statistical principles

- goals:
  - introduce a number of fundamental statistical concepts and principles that are useful for (big) data analysis
  - demonstrate how to present statistical concepts in one or two plain sentences — try yourself, wish you can do better than me

- what is statistics:
  - statistics is the study of the collection, organization, analysis, interpretation, and presentation of **data** [Wikipedia]
  - in my view, a discipline that describes **uncertainty** from the data
  - as Einstein put "God doesn't play dice with the world" — He does!
  - **distribution**: a collection of all possible outcomes and the associated chance (probability) of each outcome
  - ***p*-value**: in statistical significance testing, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true [Wikipedia] — if you believe that the conjecture (null hypothesis) is true, what is the chance that your data actually support it

Association and causation

- principle no.1:

$$\text{association} \neq \text{causation}$$

  - a correlation / association between two variables does **not** necessarily imply that one causes the other
  - most statistical analysis infer about correlation between variables; a very small number of analyses infer about causation (causal inference)

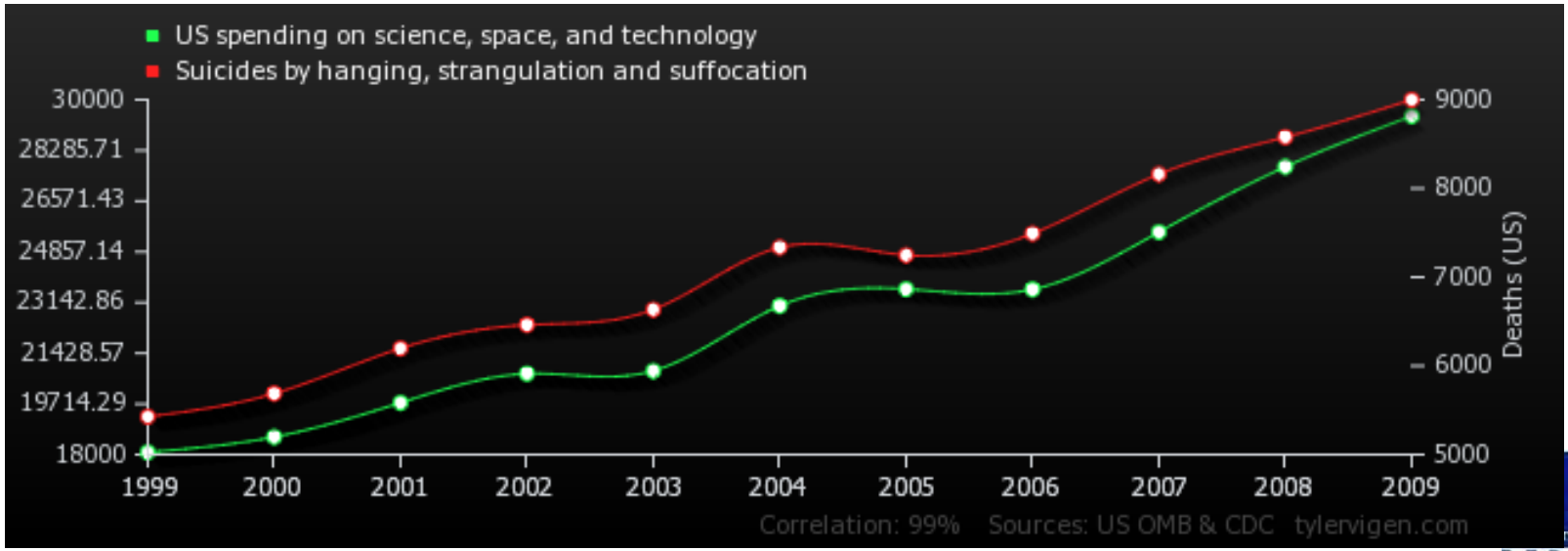
Association and causation

- principle no.1:

$$\text{association} \neq \text{causation}$$

  - a correlation / association between two variables does **not** necessarily imply that one causes the other
  - most statistical analysis infer about correlation between variables; a very small number of analyses infer about causation (causal inference)

- examples:



Association and causation

- examples (continued):
  - young children who sleep with the light on are much more likely to develop myopia in later life; therefore, sleeping with the light on causes myopia [published in *Nature*, 1999, and received much coverage in the popular press;

Association and causation

- examples (continued):
  - young children who sleep with the light on are much more likely to develop myopia in later life; therefore, sleeping with the light on causes myopia [published in *Nature*, 1999, and received much coverage in the popular press; — later study did not find that infants sleeping with the light on caused the development of myopia; it did find a strong link between parental myopia and the development of child myopia, also noting that myopic parents were more likely to leave a light on in their children's bedroom; in this case, the cause of both conditions is likely parental myopia, and the above-stated conclusion is false

Association and causation

- examples (continued):
  - young children who sleep with the light on are much more likely to develop myopia in later life; therefore, sleeping with the light on causes myopia [published in *Nature*, 1999, and received much coverage in the popular press; — later study did not find that infants sleeping with the light on caused the development of myopia; it did find a strong link between parental myopia and the development of child myopia, also noting that myopic parents were more likely to leave a light on in their children's bedroom; in this case, the cause of both conditions is likely parental myopia, and the above-stated conclusion is false
  - HDL ("good") cholesterol is negatively correlated with incidence of heart attack; therefore, taking medication to raise HDL will decrease the chance of having a heart attack

Association and causation

- examples (continued):
  - young children who sleep with the light on are much more likely to develop myopia in later life; therefore, sleeping with the light on causes myopia [published in *Nature*, 1999, and received much coverage in the popular press; — later study did not find that infants sleeping with the light on caused the development of myopia; it did find a strong link between parental myopia and the development of child myopia, also noting that myopic parents were more likely to leave a light on in their children's bedroom; in this case, the cause of both conditions is likely parental myopia, and the above-stated conclusion is false
  - HDL ("good") cholesterol is negatively correlated with incidence of heart attack; therefore, taking medication to raise HDL will decrease the chance of having a heart attack — later research called this conclusion into question; instead, it may be that other underlying factors, like genes, diet and exercise, affect both HDL levels and the likelihood of having a heart attack

Association and causation

- some lessons:
  - we laugh at obvious mistakes but often forget how easy it is to make subtle errors any time an attempt is made to use statistics to prove causality [*Little Handbook of Statistics* by G.E. Dallal]
  - with that said, we should also bear in mind that,
    - fully randomized study — A/B testing
    - with some not very complicated statistical techniques, plus some **additional assumptions** we wish to make, one can do **causal inference** with the **observational data** too

Supervised and unsupervised learning

- supervised learning vs unsupervised learning:
  - in supervised learning, the goal is to predict the value of an **outcome** measure based on a number of **input** measures
  - in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures
- examples:
  - predict whether a patient, hospitalized due to a heart attack, will have a second heart attack, based on demographic, diet and clinical measurements for that patient
  - predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data
  - identify the numbers in a handwritten ZIP code, from a digitized image
  - identify the risk factors for prostate cancer, based on clinical and demographic variables
  - identify items that appear together most frequently at a checkout counter (market basket analysis)
  - rank a set of webpages in terms of their relative importance, e.g., relevance to a search query (Google's *PageRank* algorithm)

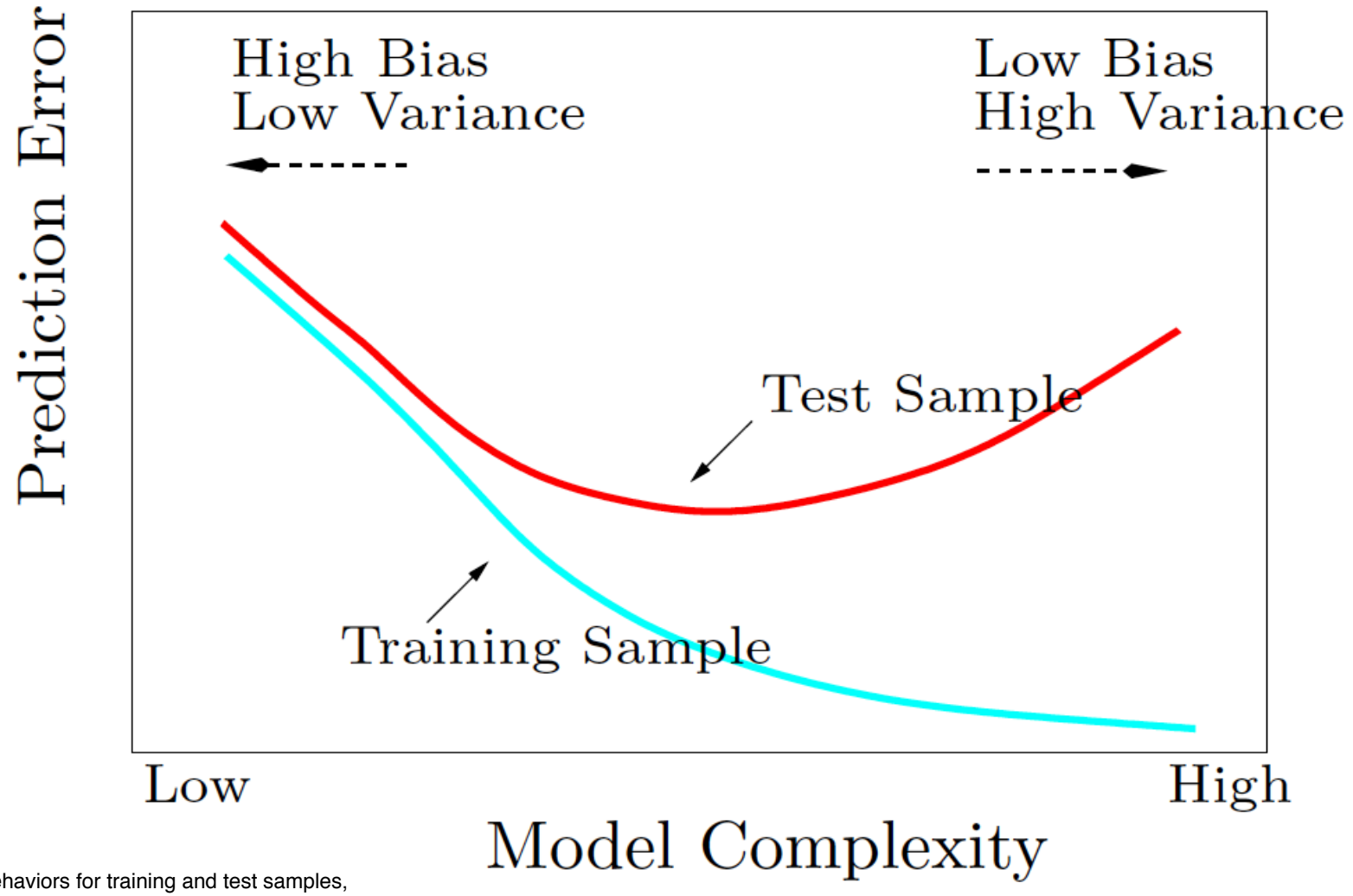
Model interpretation and prediction

- model interpretation vs prediction:
  - the two goals do **not** always align with each other
  - black-box type prediction engines can be very effective, and are often among the best performers in real data problems
  - however, data mining applications generally require interpretable models, so it is often not enough to simply produce predictions

- examples of black-box type learning methods:
  - neural networks
  - random forests
  - ensemble methods

may not always agree; I'm not against black-box solutions!

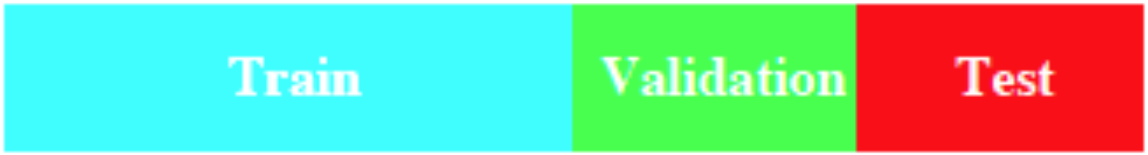
Model evaluation and selection



different behaviors for training and test samples, we want our model to explain the data as well as possible.

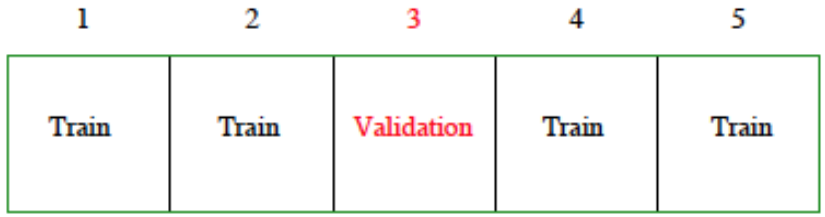
Model evaluation and selection

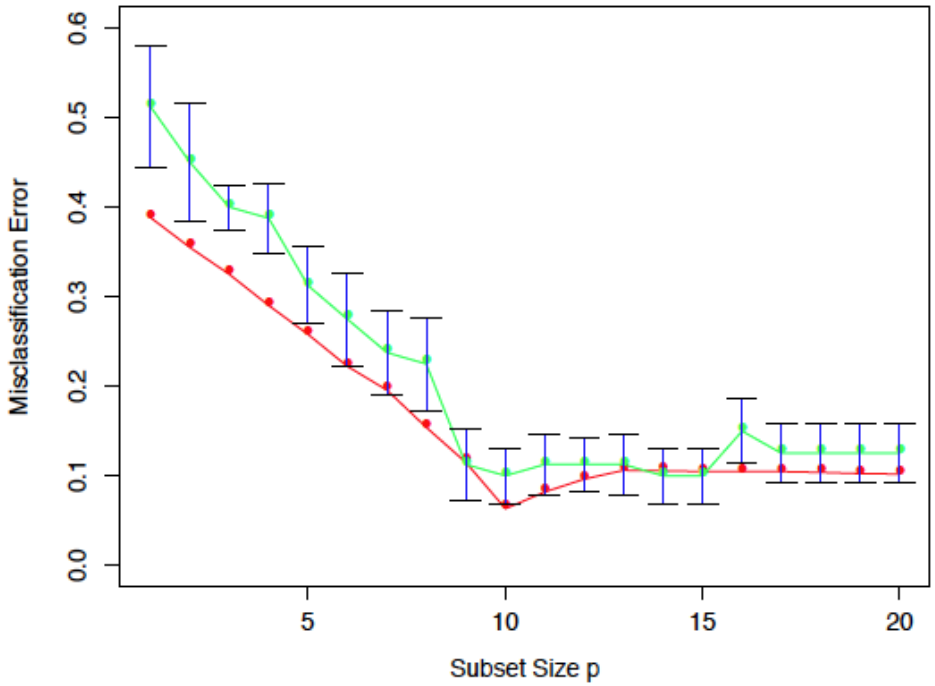
- model evaluation and selection:
  - the **generalization** performance of a learning method relates to its prediction capability on **independent test data**
  - assessment of this performance is extremely important in practice, since it guides the choice of learning method or model, and gives a measure of the model quality
- **randomly split** the data into a **training set** for model fit, a **validation set** for parameter tuning, and a **testing set** for independent evaluation
- keep in mind: **validation** is as important as model fitting



Model evaluation and selection

- cross-validation:
  - if we do not have a large enough data size, use cross-validation
  - leave-one-out; k-fold cross-validation





Bias-variance tradeoff

- **bias-variance tradeoff**:
  - assume $Y = f(\boldsymbol{X}) + \varepsilon$ where $E(\varepsilon) = 0, \text{var}(\varepsilon) = \sigma_\varepsilon^2$, then the expected prediction error of a regression fit $\hat{f}(\boldsymbol{X})$ at an input point $\boldsymbol{X} = \boldsymbol{x}_0$, using squared-error loss, is:

$$\begin{aligned} \text{Err}(\boldsymbol{x}_0) &= E[\{Y - \hat{f}(\boldsymbol{x}_0)\}^2 | \boldsymbol{X} = \boldsymbol{x}_0] \\ &= \sigma_\varepsilon^2 + [E\{\hat{f}(\boldsymbol{x}_0) - f(\boldsymbol{x}_0)\}]^2 + E[\{\hat{f}(\boldsymbol{x}_0) - E\hat{f}(\boldsymbol{x}_0)\}^2] \\ &= \sigma_\varepsilon^2 + \text{Bias}^2\{\hat{f}(\boldsymbol{x}_0)\} + \text{Var}\{\hat{f}(\boldsymbol{x}_0)\} \\ &= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance} \end{aligned}$$

  - typically, **the more complex the model $\hat{f}$, the lower the (squared) bias but the higher the variance**

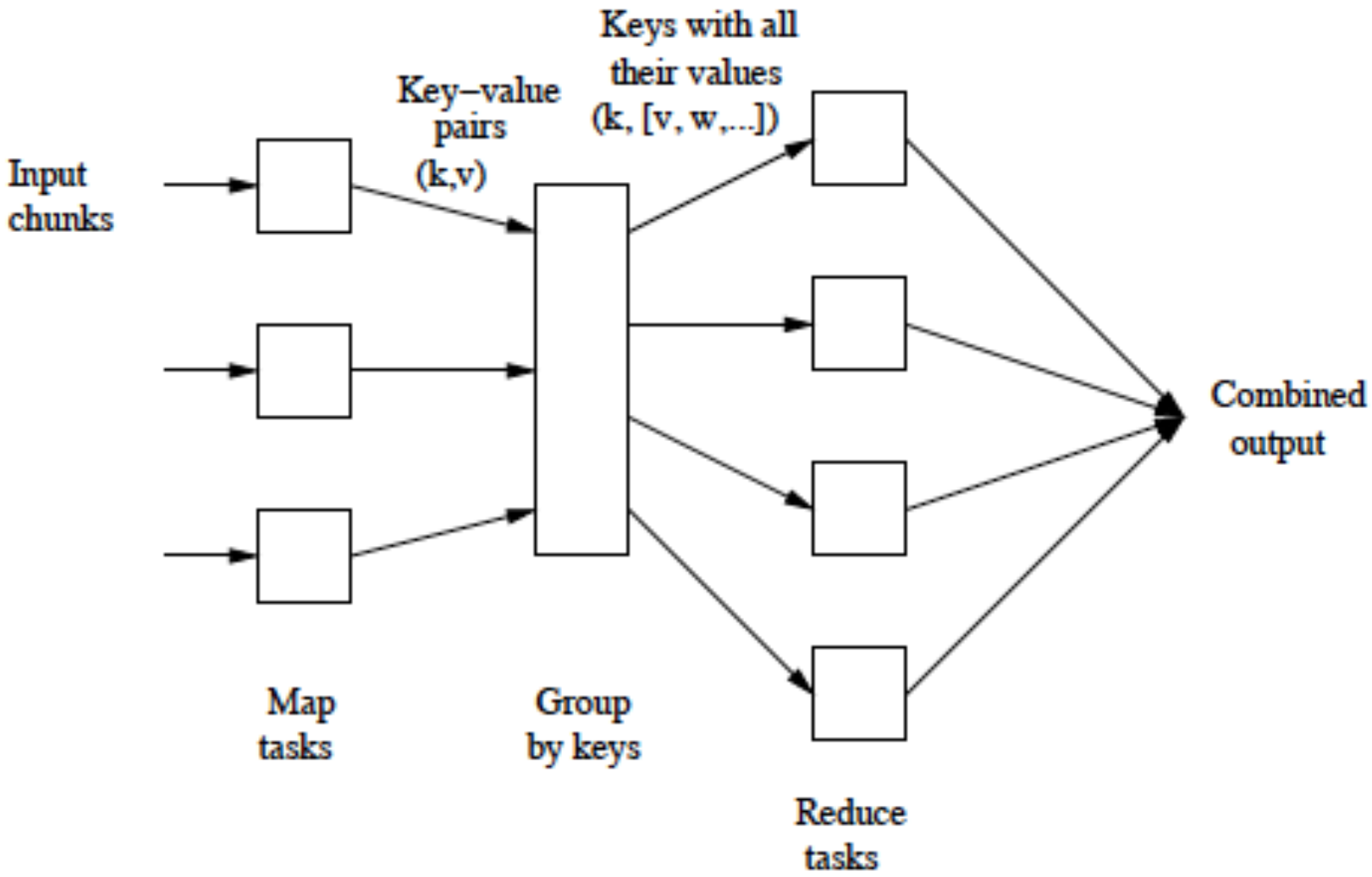
Computing: Distributed Storage and Processing

Apache Hadoop

- Apache Hadoop:
  - storage part: Hadoop Distributed File System (HDFS)
  - processing part: Hadoop Map Reduce (parallel computing)
- storage:
  - the data is extremely **large**, and extremely **regular**
  - the hardware consists of thousands of (or much more) independent components, **nodes**
  - stored on **racks** and connected by another level of network, **switches**
  - files are divided into **chunks** and stored **redundantly**; chunks are **replicated**, e.g., three times at three computing nodes on different racks
- processing:
  - computations are divided into **tasks**, such that if any one task fails to execute to completion, it can be restarted without affecting other tasks
  - **parallel** and **reliable**

MapReduce

- introduction:
  - a style of computing to manage many large-scale computations in a way that is parallel and tolerant of hardware faults
  - you write two functions: *Map* and *Reduce*
  - the system manages the parallel execution, coordination of tasks that execute *Map* or *Reduce*, and also deals with the possibility that one of these tasks will fail to execute
- key steps:
  - Map tasks are given one or more chunks from a distributed file system; turn the chunk into a sequence of key-value pairs
  - key-value pairs from each Map task are collected by a master controller and sorted by key; keys are divided among all the Reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task
  - Reduce tasks work on one key at a time, and combine all the values associated with that key in some way


Keys with all
their values
(k, [v, w,...])
Key–value
pairs
(k,v)
Input
chunks
Combined
output
Map
tasks
Group
by keys
Reduce
tasks

MapReduce

- an example: matrix-vector multiplication
  - $\boldsymbol{x} = (x_i) = \boldsymbol{M}\boldsymbol{v} \in \mathbb{R}^n$, where $\boldsymbol{M} = (m_{ij}) \in \mathbb{R}^{n \times n}$, $\boldsymbol{v} = (v_j) \in \mathbb{R}^n$
  - this calculation is at the heart of the ranking of Web pages that goes on at search engines, and there, n is in the tens of billions
  - $x_i = \sum_{j=1}^{n} m_{ij} v_j$
  - Map function: each Map task will take the entire vector $\boldsymbol{v}$ and a chunk of the matrix $\boldsymbol{M}$; from each matrix element m_{ij} it produces the key-value pair $(i, m_{ij} v_j)$; all terms of the sum that make up the component x_i will get the same key
  - Reduce function: a Reduce task has simply to sum all the values associated with a given key i; the result is a pair (i, x_i)
- more tasks by MapReduce:
  - Rajaraman et al. (2012): relational algebra, selection, projection, union, intersection, difference, aggregation, matrix multiplication, ...
  - Chu et al. (2006): locally weighted linear regression, k-means, logistic regression, naive Bayes, SVM, ICA, PCA, LDA, EM, and back propagation in neural networks, ...

Art of Data Analysis

— based on Cassie Kozyrkov's
Practical Guide to Data Science

Real world data analysis

- real world data:
  - vast amount
  - all sorts of data, and much information can be irrelevant
  - noisy; many many missing values, data errors
- what to do?

Real world data analysis

- real world data:
  - vast amount
  - all sorts of data, and much information can be irrelevant
  - noisy; many many missing values, data errors
- what to do?
  - cry a little in dread and/or excitement
  - plot it!
  - apply your favorite model
  - search the academic literature
  - use a data-mining package
  - none of the above

Real world data analysis

- real world data:
  - vast amount
  - all sorts of data, and much information can be irrelevant
  - noisy; many many missing values, data errors
- what to do?
  - cry a little in dread and/or excitement
  - plot it!
  - apply your favorite model
  - search the academic literature
  - use a data-mining package
  - **none of the above**

Real world data analysis

- think first!
  - what is the question of interest?
  - how will my results be used?
  - what shape do interesting results take?
  - does anybody care? why?
  - what are the resulting action items?

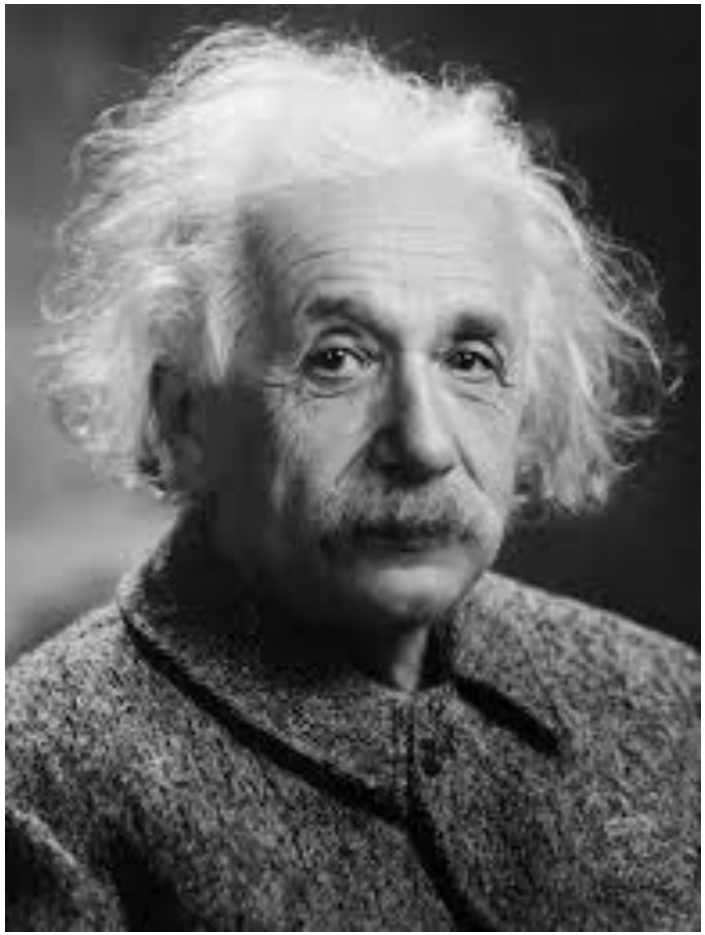
Real world data analysis

- think first!
  - what is the question of interest?
  - how will my results be used?
  - what shape do interesting results take?
  - does anybody care? why?
  - what are the resulting action items?

- important steps:
  - "one of the most important things you'll do on the job is come up with the right hypotheses." — ask the right question, and properly formulate the problem
  - extract the right, relevant data for your analysis — may require engineering skills
  - explore and understand the data; plot it, if possible!
  - choose the right tools and carry out an appropriate analysis
  - present your findings

Big Data

"If you can't explain it simply, you don't understand it well enough." — Albert Einstein



Additional readings

- Hastie, H., Tibshirani, R., and Friedman, J. (2001). *Elements of Statistical Learning*. Springer. Chapter 7
- Rajaraman, A., Leskovec, J., and Ullman, J.D. (2012). *Mining of Massive Datasets*. Chapter 2
- Chu, C.T. Kim, S.K., Lin, Y.A., Yu, Y., Bradski, G., Ng, A.Y., and Olukotun, K. (2006). Map-Reduce for Machine Learning on Multicore, In *NIPS*, edited by Bernhard Schölkopf, John C. Platt, Thomas Hoffman, 281-288
- *The Wall Street Journal*, Article 11-29-2012, "Big Data is on the rise, bringing big questions"
- *The Wall Street Journal*, Article 04-29-2012, "Big Data's big problem: little talent"
- *McKinsey&Company*, Report 05-2011, "Big Data: The next frontier for innovation, competition, and productivity"