Welcome! Take a seat anywhere

COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

Machine Learning Boot Camp

Wi-fi Network:
guest-net (no password)

Open any webpage (i.e. BBC, Amazon, etc.) on your browser, a pop-up will appear to connect to guest wi-fi. Accept terms to gain access.

Boot Camp Team

Cody Chiuzan
Course Director - Columbia

Noah Simon
Instructor – U. Washington

Yifei Sun
Instructor - Columbia

Xu Gao
Workshop Guide - Columbia

Lizzy Gibson
Workshop Guide - Columbia

Serena Zhang
Workshop Guide - Columbia

Abby Welbourn
Admin. Director - Columbia

Day 1: June 6, 2019

8:00 – 9:00 Optional training: Intro to R

9:15 – 9:30 Welcome and Introduction

9:30 – 10:00 Low vs. High Dimensional Data (Lecture)

10:00 – 10:30 Classical vs. New Methodologies (Lecture)

10:45 – 12:00 Penalized Regression Methods (Lecture)

12:00 – 1:00 Networking Lunch (11th Floor, Room 1101)

1:00 – 1:45 Penalized Regression Methods (Lab)

1:45 – 2:45 Non-Linear Regression (Lecture + Lab)

3:00 – 4:45 Classification Models (Lecture + Lab)

4:45 – 5:00 Day Overview, Q&A

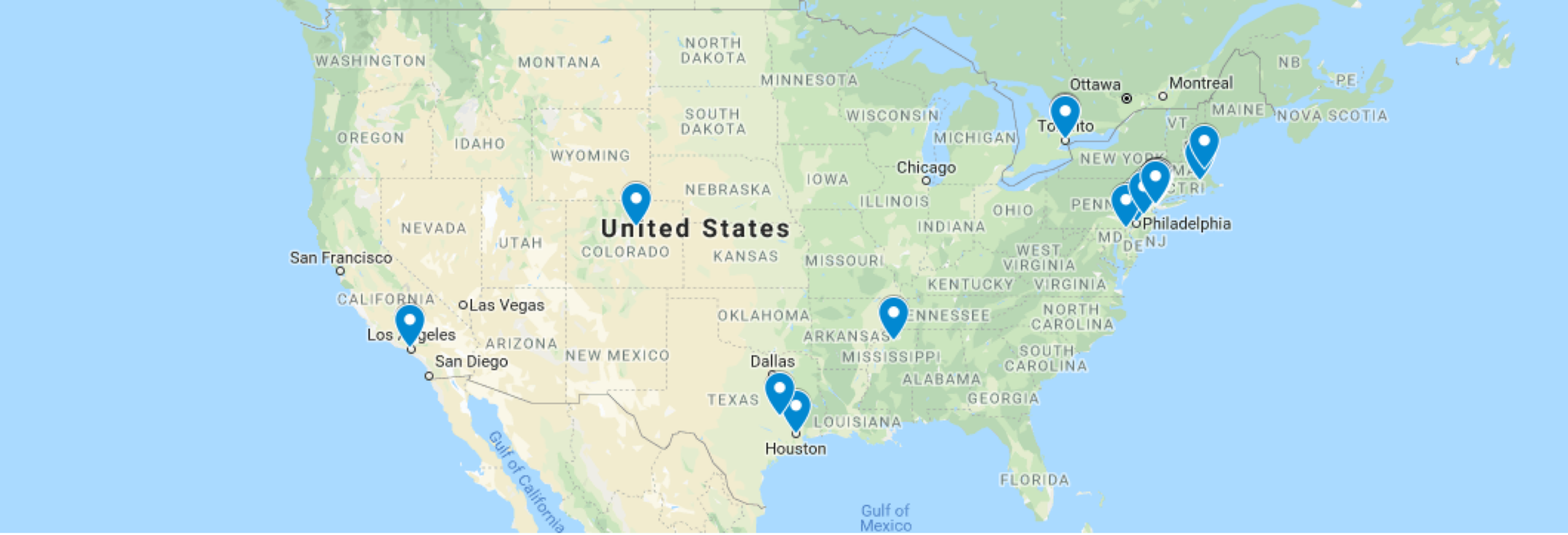
Day 2: June 7, 2019

8:30 – 8:45	Recap from Day 1: Group Q&A
8:45 – 10:00	Tree Based Methods (Lecture)
10:15 – 11:00	Tree Based Methods (Lab)
11:00 – 12:00	Model Interpretation (Lecture + Lab)
12:00 – 1:00	*Networking Lunch (11th Floor, Room 1101)*
1:00 – 2:15	Clustering Algorithms (Lecture + Lab)
2:15 – 3:30	Principal Component Analysis (Lecture + Lab)
3:45 – 4:45	Deep Learning
4:45 – 5:00	Evaluations, Questions, Wrap-Up

Quick Logistics

- Restrooms: Exit classroom- Men's on right, women's on left.
- Drinks/Snacks: Café on 10th floor (5 floors up)
- Lunch: on 11th Floor, Room 1101 (6 floors up)
- Course materials: will be sent to everyone after the Training
- Name tags: please wear during training to make connecting with others easier! Please return after the Training to help us be greener

Contact Abby Welbourn 734-972-1733 for assistance



Around-the-nation Attendees

California • Colorado • Texas • Mississippi • Canada • Delaware • Pennsylvania
New Jersey • New York • Rhode Island • Massachusetts

To show this poll

1

Install the app from
pollev.com/app

2

Start the presentation

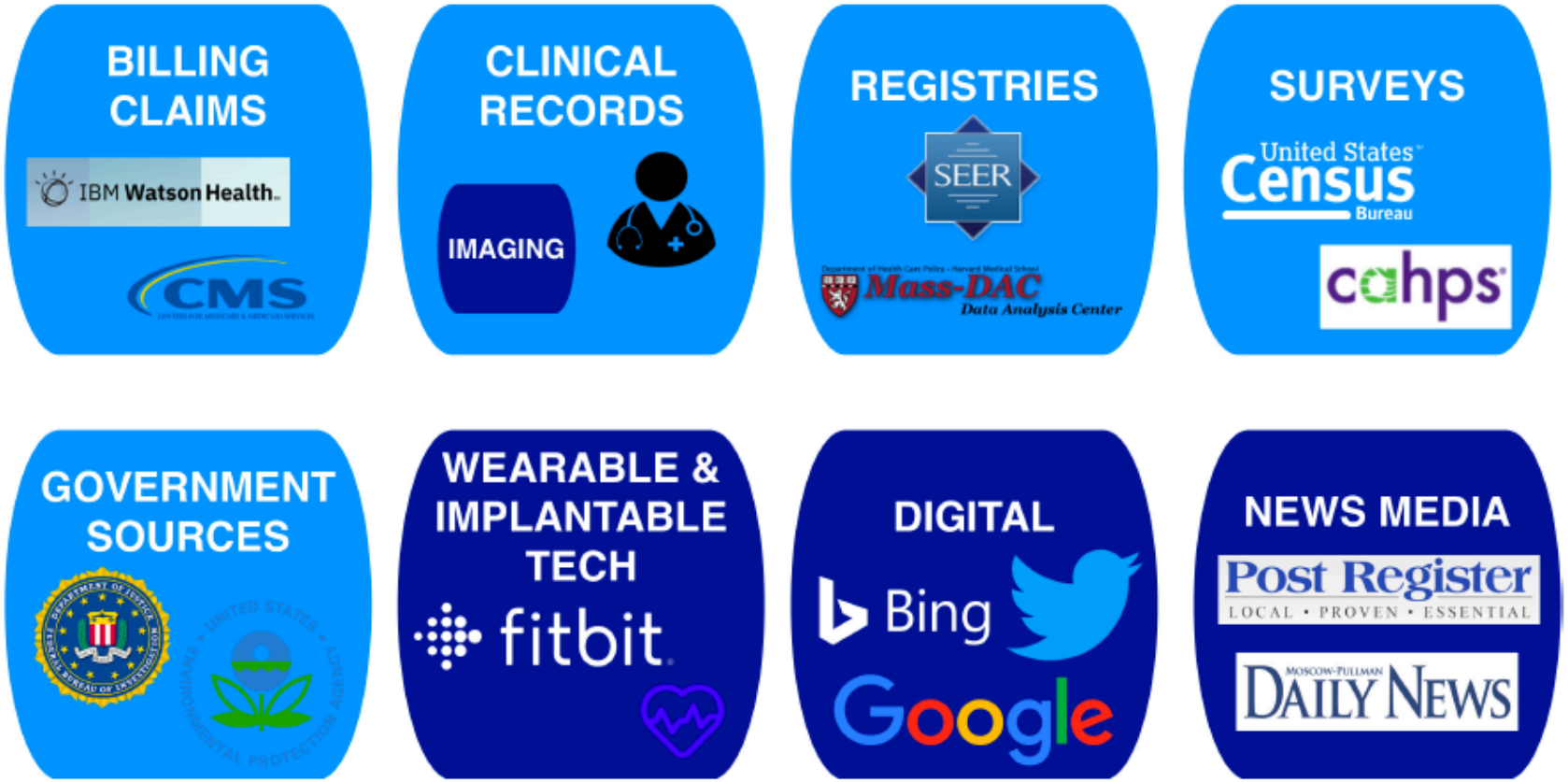
Still not working? Get help at pollev.com/app/help
or
Open poll in your web browser

BIG DATA vs SMALL DATA

Cody Chiuzan, PhD

Department of Biostatistics
Columbia University
June 6, 2019

BIG DATA



Sherri Rose, 2019

What is BIG DATA?

- Big data, commonly characterized by volume, variety, velocity, and veracity, goes beyond the data type and includes different aspects such as storage, data quality, and aspects of data analysis: hypothesis-generating, rather than hypothesis-testing
- The strength of big data is finding associations and finding a signal is only the first step
- Medical big data has distinct features from big data of other disciplines, but also distinct from traditional clinical epidemiology
- Lots of healthcare applications
  - Predictive modeling (disease)
  - Clinical decision support
  - Disease or safety surveillance

Medical Big Data: Sources

- Some sources of medical big data:
  - Administrative claim record
  - Clinical registries
  - Electronic health records
  - Biometric data
  - Patient-reported data
  - Medical imaging
  - Biomarker data
  - Prospective or retrospective cohort studies
  - Clinical trials

Medical Big Data: Challenges

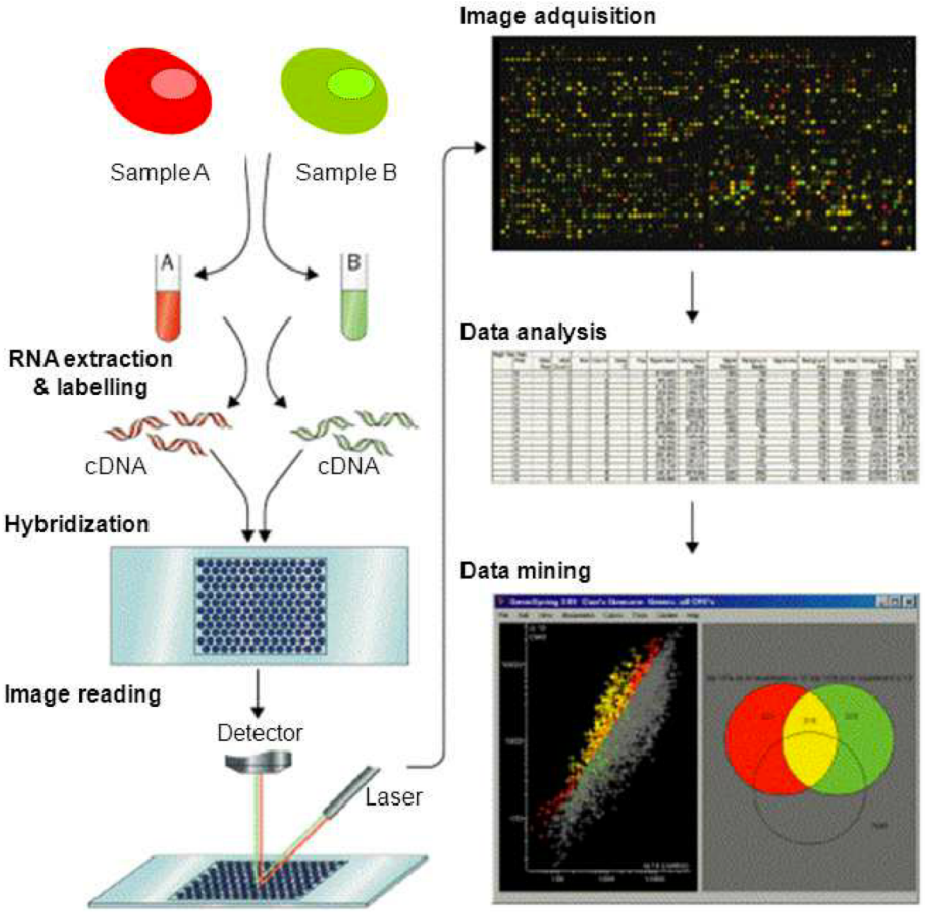
- Unique features and challenges
  - Complementary dimensions
  - Disparate sources (data inconsistency and instability)
  - Multiple scales (seconds vs years)
  - Incompleteness
  - Data quality: errors/changes in coding (e.g., ICD codes changed in 2014), error in measurement
  - Limitations of observational studies, validation, analytical issues

Medical Data: One Classification

- Large *n* and small *p* (e.g., administrative claims)
  - Data tends to be incomplete and noisy
  - Can be dealt with classical statistical approaches
  - Problems with spurious associations
- Small *n* and large *p* (e.g., microarrays)
  - Curse of dimensionality
  - Multiple testing problems
  - Classical tests do not address this type of data efficiently
- Large *n* and large *p* (combines issues from the previous two)

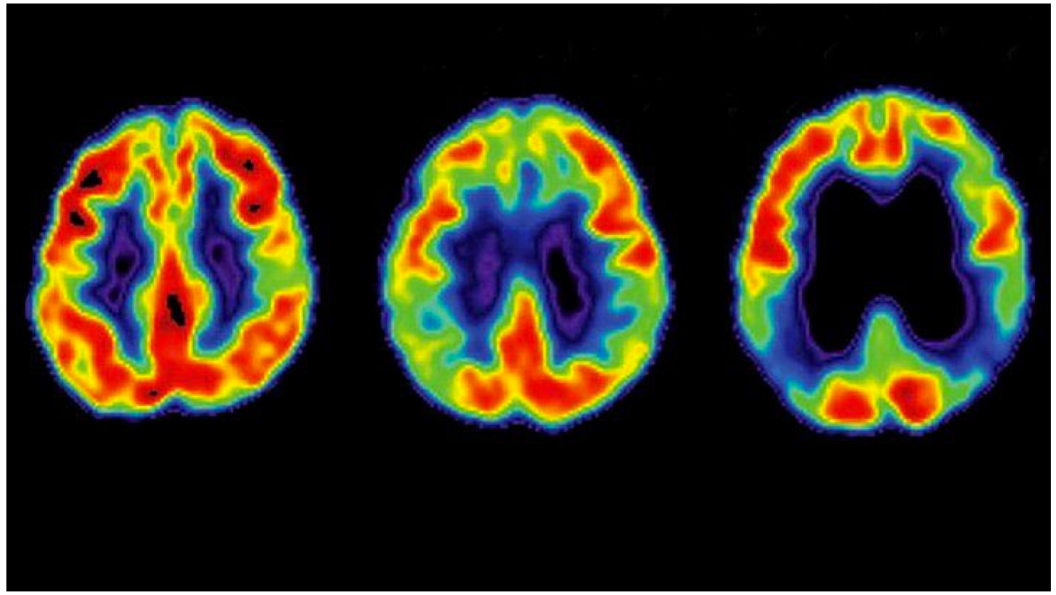
Example: Microarrays

- Measure differential gene expressions in samples (e.g., disease vs normal tissue)
- Gene expression profiles are compared via the difference in the fluorescence of the two samples
- Identify groups of genes (clustering)
- Model building: small *n*, large *p*



Example: fMRI

- Functional magnetic resonance imaging measures the brain activity by detecting changes associated with blood flow
- Which voxels are significant?
- Which groups of voxels behave similarly (clustering)?
- Networks: How are voxels or groups of voxels related to each other or through time?
- Model building: small *n*, large *p*



Alzheimer's disease: mild, moderate, and severe stage caused by neuron death and tissue loss.

Promising Results (thus far)

- Personalized medicine
- Use of clinical decision support systems (e.g., automated medical images, mining of medical records/literature – natural language processing)
- Tailored diagnosis and treatment decisions to support desired patient behaviors using mobile devices

What About 'Small' Data?

- Data scientist have gained experience with using appropriate algorithms for 'big data'

- Small datasets still exist
  - E.g., rare diseases or phenomena, aggregate modeling of states, countries (limited population)

- Small data challenges
  - Over-fitting
  - Noise becomes problematic
  - Outliers can be influential

What About 'Small' Data?

Some suggestions to deal with small data problems:

- Use simple, classical models (e.g., parametric models)
- Pool data (if reasonable)
- Use Bayesian modeling (sensible priors)
- Use confidence intervals (not just point estimates)
- Perform feature selection and use regularization
- Remember: "Statisticians are the original data scientists"!

MACHINE LEARNING

VS

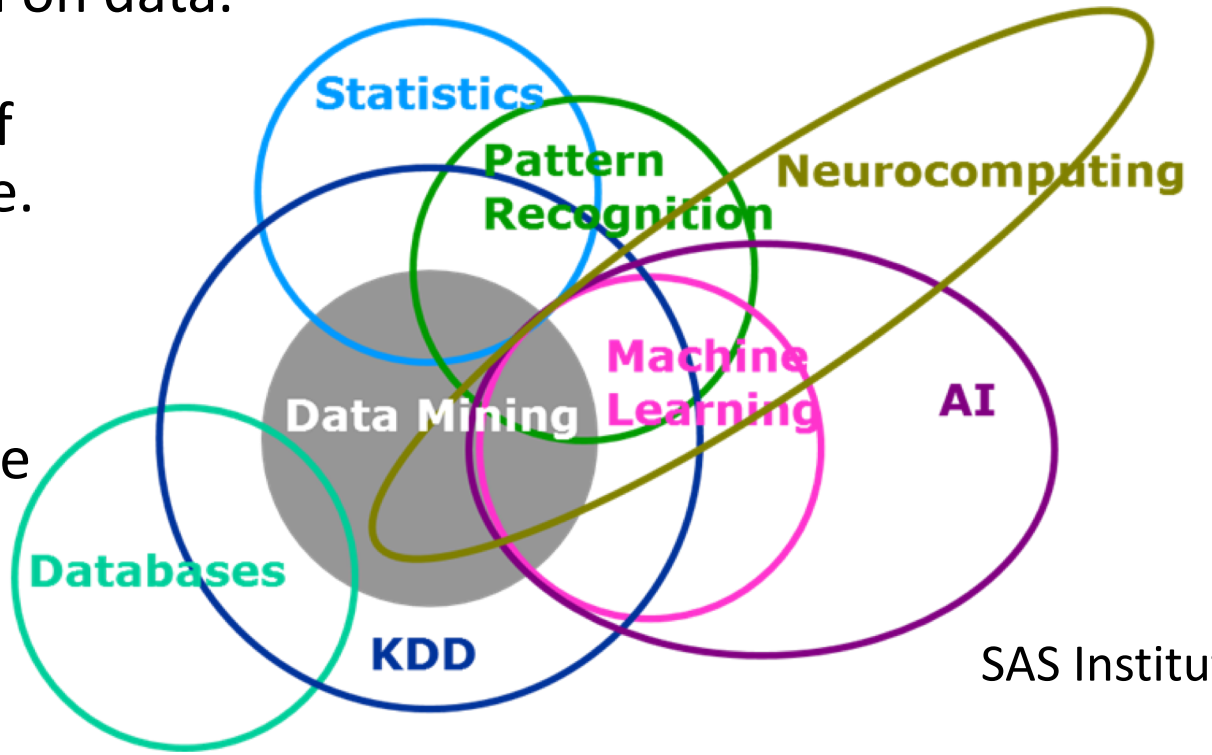
CLASSICAL METHODS

What is Machine Learning (ML)?

(One) Definition: A set of methods that computers use to make and improve predictions or behaviors based on data.

ML arises at the intersection of statistics and computer science.

ML is driven by the unique Computational challenges of building models from very large data sets.



SAS Institute

Machine Learning: Major Types

Supervised Learning starts with the goal of predicting a known outcome or target.

Supervised learning focuses on:

- Classification - involves choosing among subgroups to best describe a new data instance
  - E.g., automated interpretation of the EKG, where pattern recognition is performed to select from a limited set of diagnoses
- Prediction - involves estimating an unknown parameter
  - E.g., estimating the risk score for coronary heart disease (CHD)

Supervised Learning Methods

Linear Regression	Standard method for modeling the relationship b/w independent variables and a continuous outcome; helps making predictions of future values of the outcome.
Logistic Regression	Similar to linear regression but operates with a binary outcome; helps with classification tasks, e.g., what are the odds of developing cancer?
Decision Tree	Classification or regression model that splits data-feature values into branches at decision nodes (e.g., if a feature is a color, each possible color becomes a new branch) until a final decision output is made.
Random Forest	Improves the accuracy of decision tree by generating multiple decision trees and taking a majority vote of them to predict the output.

Supervised Learning Methods

Support Vector Machine	Typically used for classification; it draws a division (as wide as posible) between classes; can be generalized to nonlinear problems.
Gradient Boosting Tree	Classification/regression technique that generates decision trees sequentially; each tree focuses on correcting the errors coming from the previous tree model and the final output is a combination of the results from all trees.
Simple Neural Network	Used in classification or regression problems. 'Artificial' neurons create an input layer, one or more hidden layers where calculations take place, and an output layer.

Unsupervised Learning Methods

Unsupervised Learning does not have any known outputs to predict.

- In the 'unsupervised' setting, we try to find naturally occurring patterns and groupings within the data.
- Compelling example: Precision Medicine Initiative
    - E.g., look at a collection of cytokines, lymphocytes, etc., and see if there are recurrent patterns to guide the development of future therapies

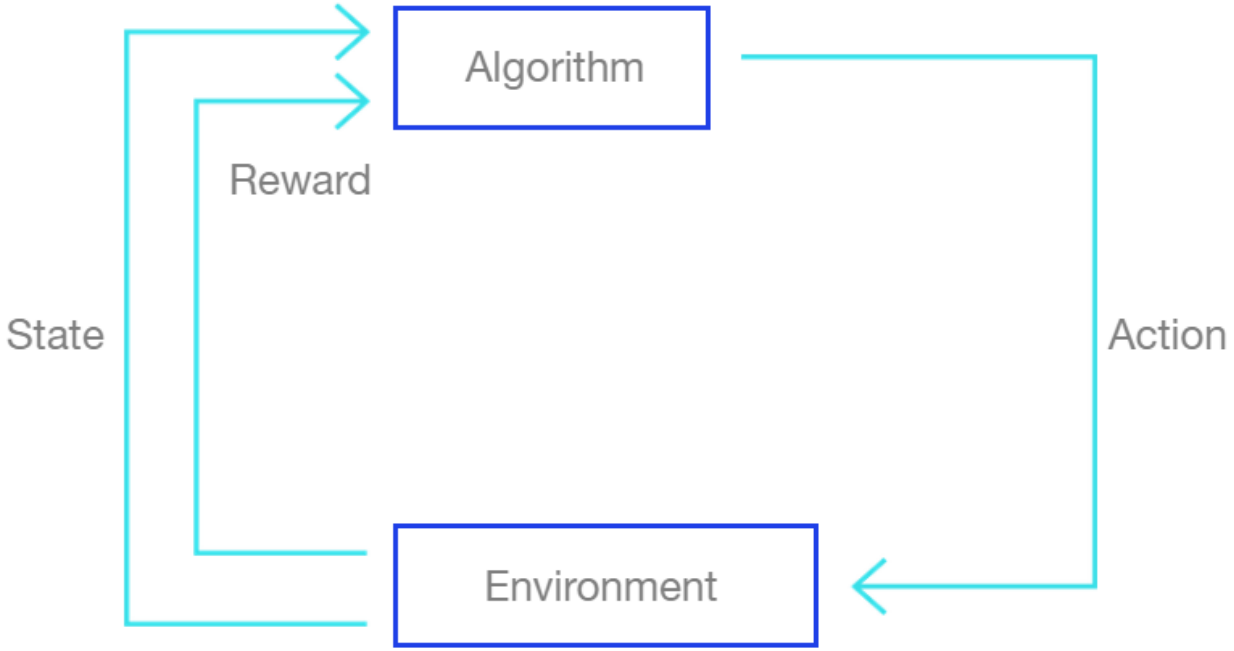
Unsupervised Learning Methods

K-means Clustering	Puts data into a number of groups (k) that each contain data with similar characteristics (as determined by the model).
Hierarchical Clustering	Splits or aggregates clusters along a hierarchical tree to form a classification system.
Principal Component	A method used for dimensionality reduction by converting a set of correlated variables into a set of uncorrelated variables, called principal components.
Gaussian Mixture Model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups (clusters).

Reinforcement Learning

An algorithm learns to perform a task simply by trying to maximize rewards it receives for its actions (e.g, maximizes points it receives for increasing returns of an investment portfolio).

- E.g., optimize the driving behavior of self-driving cars.

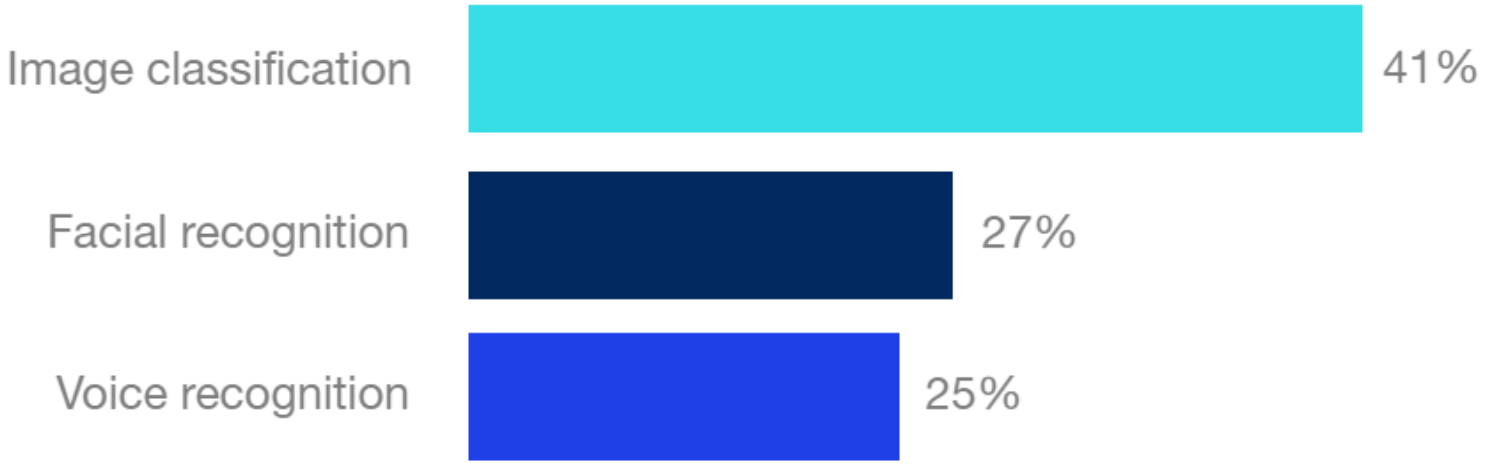


McKinsey Analytics

Deep Learning

Deep learning is a type of machine learning that can process a wider range of data resources, requires less data preprocessing by humans, and can often produce more accurate results than traditional methods.

For example, see below for % in reduction rates achieved by deep learning methods:



McKinsey Analytics

Still Confused?!

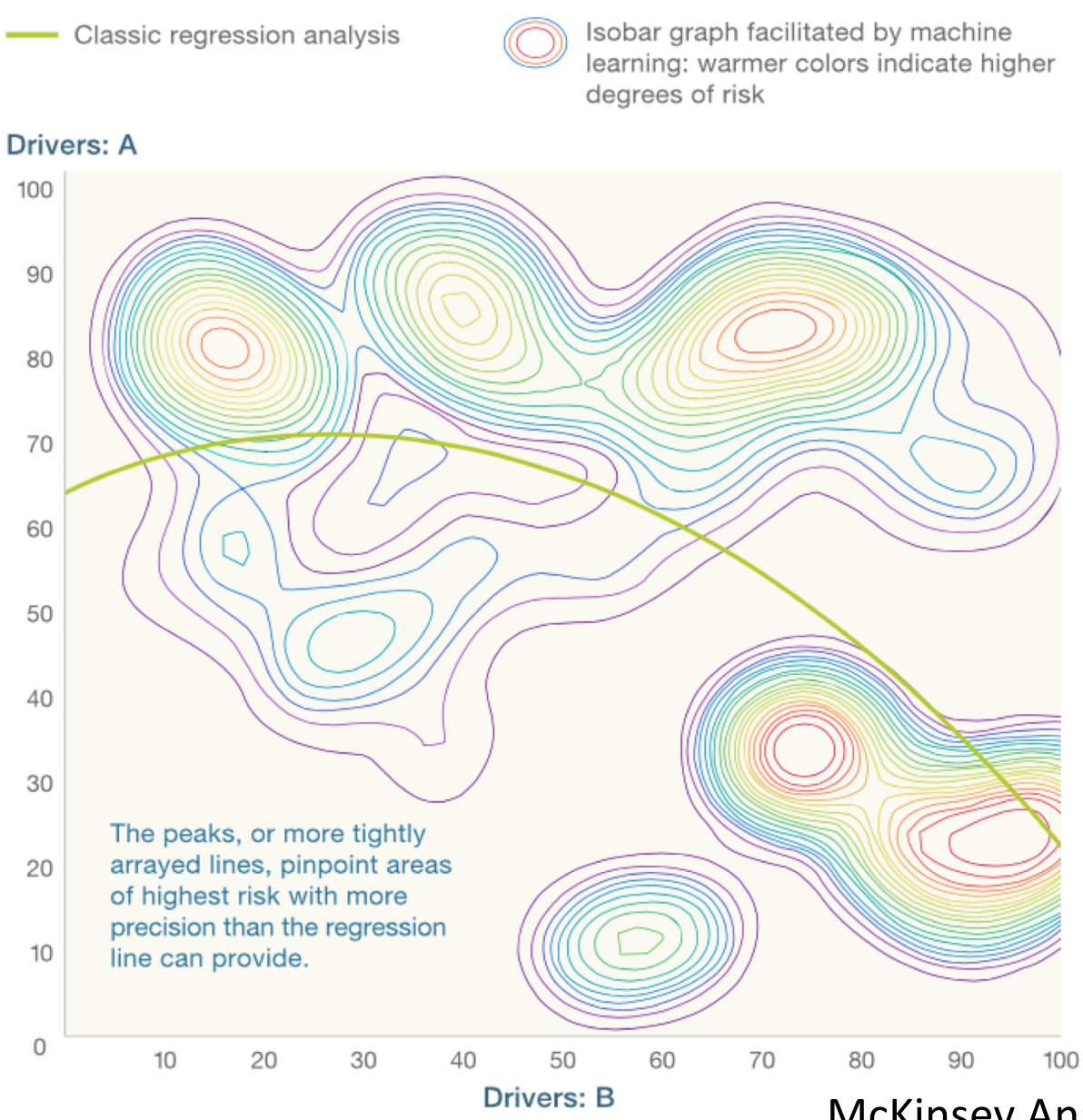
Machine Learning is …

Using algorithms that can learn from data, instead of explicitly programmed instructions.

Statistical Modelling is …

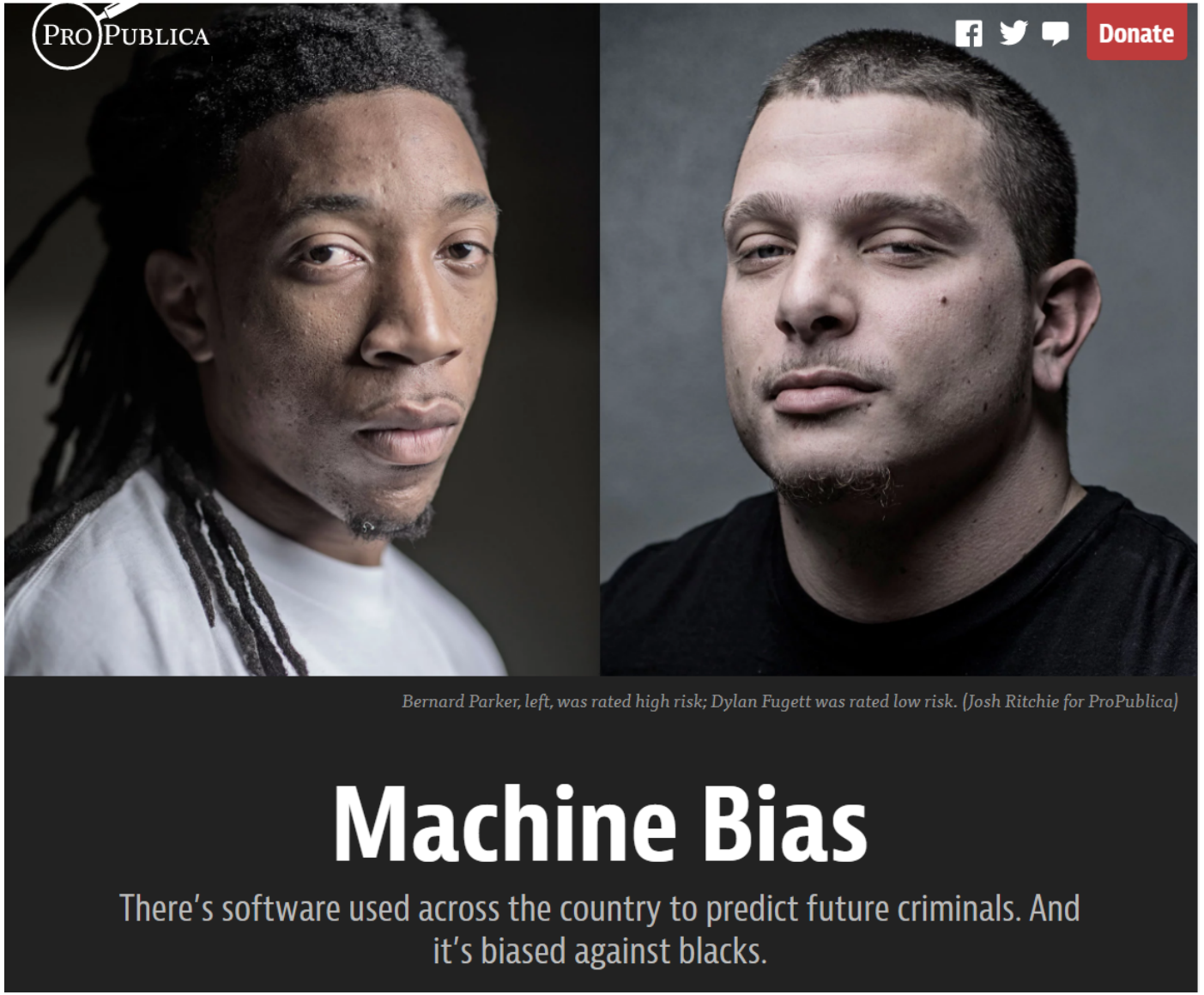
A subfield of mathematics which deals with finding relationship between variables to predict an outcome.

Machine Learning vs Statistical Modeling



McKinsey Analytics

Machine Learning vs Statistical Modeling



Machine Learning vs Statistical Modeling

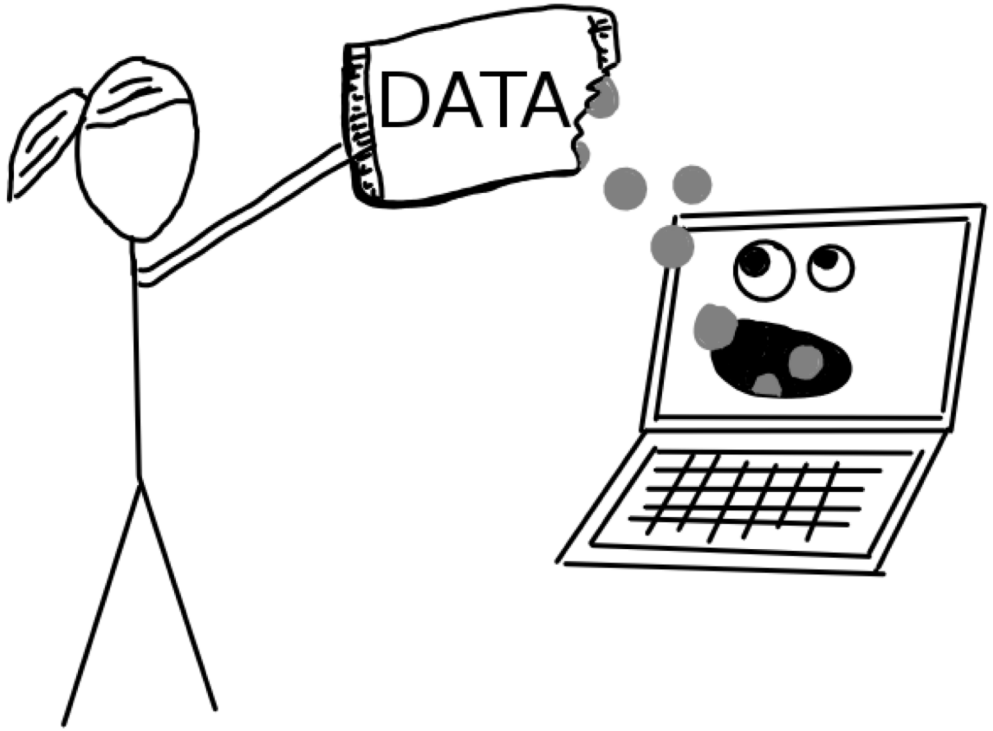
ML vs Statistical Modeling

- They come from different schools
- They belong to different ‘generations’
- Statistical modeling requires more assumptions
  - Certain relationships b/w dependent and independent variables
  - Distribution assumptions are difficult to check, inflexible
- ML algorithms are more efficient with big data
- Predictive power is generally stronger if less assumptions are made
- Focus on model-agnostic interpretability tools

Without Machine Learning

* VERY SPECIFIC INSTRUCTIONS

With Machine Learning



The Future of ML

- Machine learning (or “AI”) is associated with lots of promises and expectations
    - “AI labs”, “Data Scientists”, “Machine Learning Experts”, “AI engineers”
- Machine learning is constantly moving from science to business processes, products and real world applications
    - Decision-making
    - Drug discovery
    - Quality controls in assembly lines
    - Self-driving cars
    - Diagnosis of diseases
- Interpretability will accelerate the adoption of machine learning

THANK YOU