Biostat 202: Opportunities and Challenges of “Big Data”: Review and Case Studies

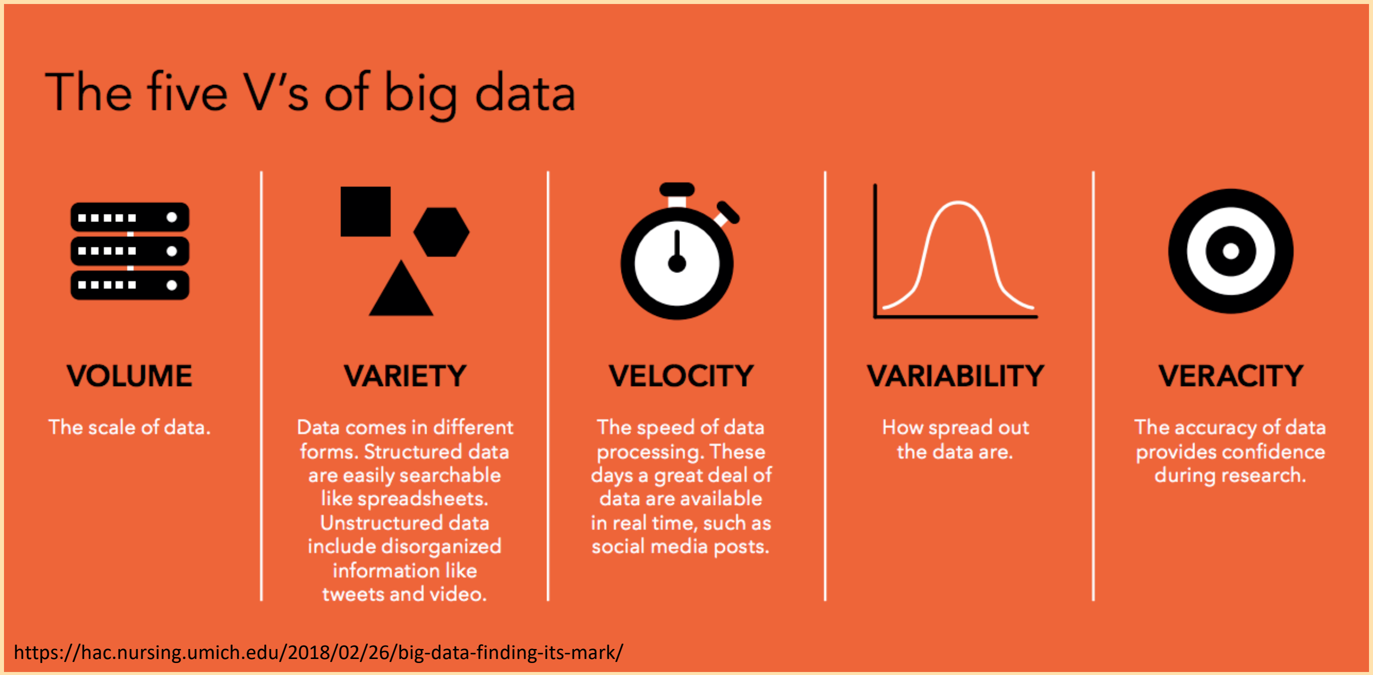
Aaron Wolfe Scheffler

Division of Biostatistics

Department of Epidemiology and Biostatistics

9/9/2019

“Big Data”

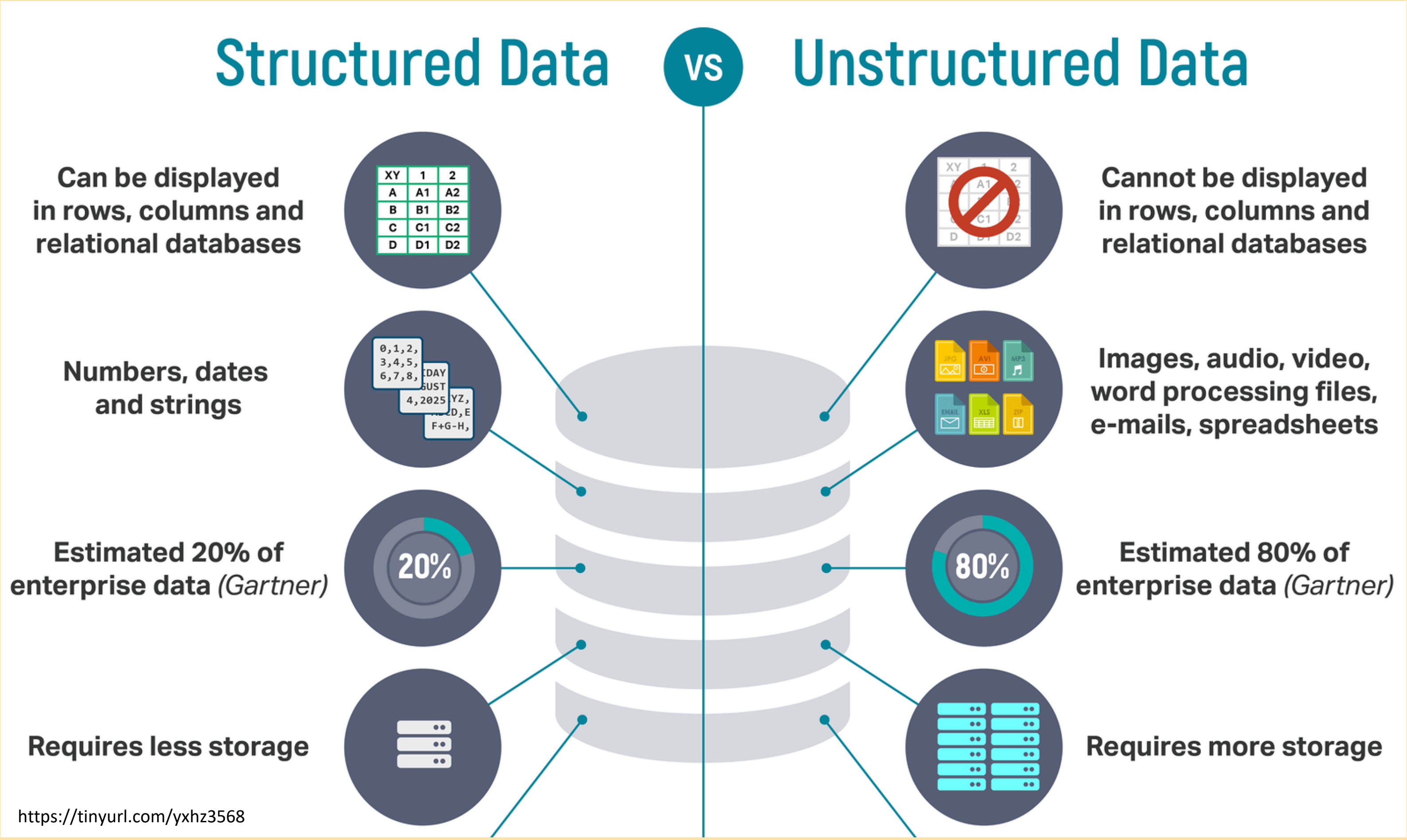


https://hac.nursing.umich.edu/2018/02/26/big-data-finding-its-mark/

Big Data

All Data

Structured vs unstructured data



Our focus

Goal: Use machine learning algorithms for the purpose of *prediction* or *identifying patterns* in data

1. Data

2. Train/fit model

Algorithm(s)

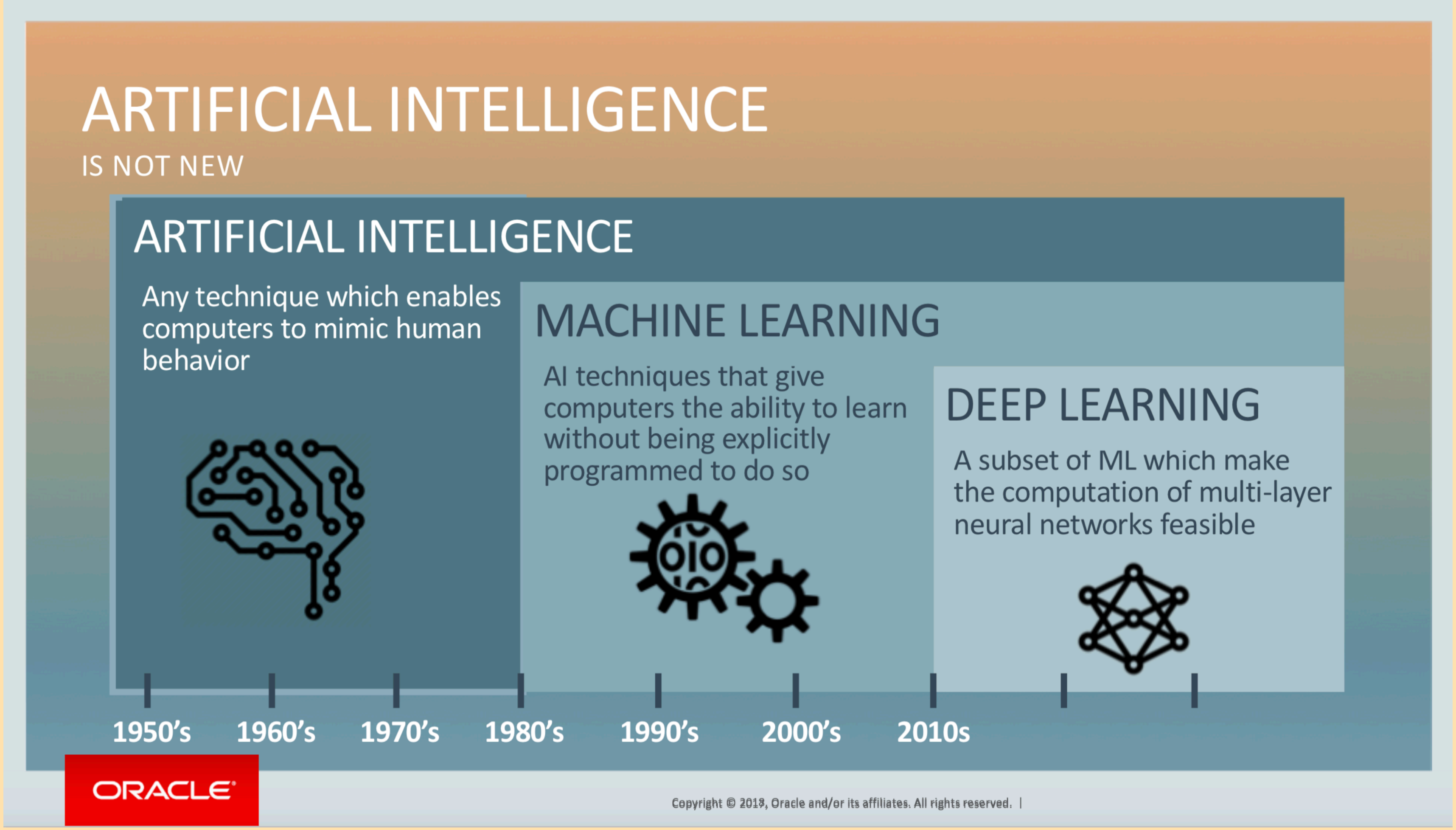
3. Output

- Prediction
- Dimension Reduction
- Clustering

4. Interpret

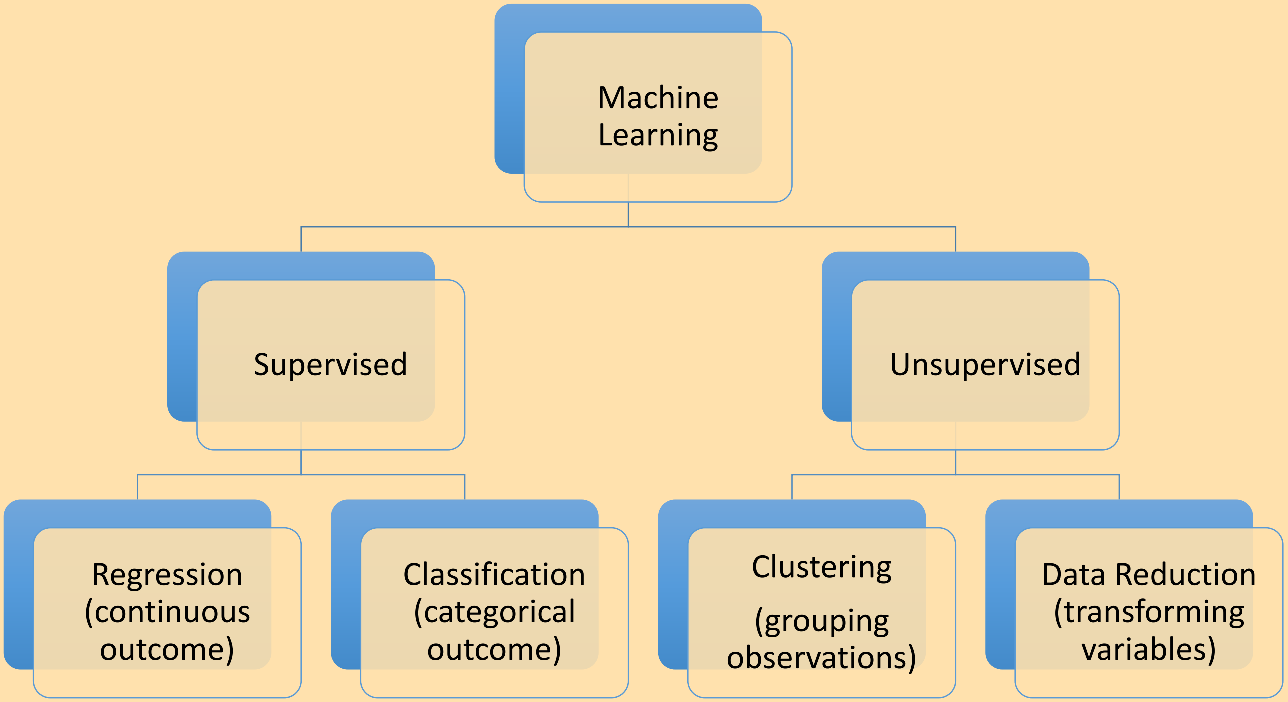
= INFORMATION?

A historical view



https://tinyurl.com/y37tgvt6

Supervised vs unsupervised



Supervised vs unsupervised

- **Supervised Learning**
- Have data where you know outcome in order to *train* the model
- Train the model to predict future outcomes from sets of predictors

Prediction: Regression (continuous outcomes) and classification (categorical outcomes)

- **Unsupervised Learning**
- No desired outcomes
- Separate data points into groups with “similar” properties (clustering)
- Or reduce data to fewer variables in some way that still captures important structure of the data

Clustering and data reduction

Model Validation

- Train-test
  - Assess performance of one model
- Train-validate-test
  - Assess performance of multiple models (e.g. parameter tuning)
- Cross-validation
  - Use training data as a validation set iteratively when you have small data

Evaluation Metrics

- Continuous targets
  - Correlation
  - Mean squared error
- Categorical targets
  - Accuracy
  - Sensitivity
  - Specificity
  - Area under the ROC curve (AUROC)
- Clustering
  - Silhouette measure

Ensemble methods

- Train a bunch of predictive models and pool them to form a final prediction.
- **Advantage:**
  - improvement in predictive accuracy
- **Disadvantages:**
  - more computation
  - it is difficult to understand an ensemble of classifiers
- Examples: bagging, boosting, and stacking

Causal Inference

- Prediction vs causation
  - estimate, predict, associate = prediction
  - effect, intervention = causal
- Threats to causal inference in observational data
  - Selection bias
- Some uses of big data and machine learning in causal inference
  - Propensity scores (induce balance, attempt to replicate randomization)
  - Potential outcomes (estimate average causal effect [ACE])

Thoughts on Big Data

1. Data mining can be used for good and evil.
2. Big data isn't the same as big information.
3. Modern machine learning methods can search for associations quickly and flexibly.
4. Having big data doesn't solve selection bias.
5. Distinguish prediction problems from causal investigation.
6. Data mining methods are good at finding spurious associations and hard to use reproducibly. Need to wary of over-fitting.
7. Randomization can be used with big data.

Case Studies

Example

Psychoactive Medications and Adverse Outcomes among Older Adults Receiving Hemodialysis

Julie H. Ishida, MD, MAS,[*†] *Charles E. McCulloch, PhD,*[‡] *Michael A. Steinman, MD,*[§] *Barbara A. Grimes, PhD,*[‡] *and Kirsten L. Johansen, MD*[*†‡]

BACKGROUND: Guidelines recommend avoidance of several psychoactive medications such as hypnotics in older adults due to their adverse effects. Older patients on hemodialysis may be particularly vulnerable to complications related to use of these agents, but only limited data are available about the risks in this population.

OBJECTIVES: To evaluate the association between the use of psychoactive medications and time to first emergency department visit or hospitalization for altered mental status, fall, and fracture among older patients receiving hemodialysis.

PARTICIPANTS: A total of 60 007 adults 65 years or older receiving hemodialysis with Medicare Part D coverage in 2011.

MEASUREMENTS: The predictors were use of sedative-hypnotics and anticholinergic antidepressants (modeled as separate time-varying exposures). The outcomes were time to first emergency department visit or hospitalization for altered mental status, fall, and fracture (modeled separately).

Example

Psychoactive Medications and Adverse Outcomes among Older Adults Receiving Hemodialysis

Julie H. Ishida, MD, MAS,[*†] *Charles E. McCulloch, PhD,*[‡] *Michael A. Steinman, MD,*[§] *Barbara A. Grimes, PhD,*[‡] *and Kirsten L. Johansen, MD*[*†‡]

- **Goal:** predict time to first emergency department (ED) visit or hospitalization due to altered mental status etc...
  - Supervised learning, prediction

- **Target:** time to first ED visit or hospitalization (continuous)

- **Predictors:** psychoactive medication, demographics, etc…

- **Techniques:** regression

- **Validation**: form validation data set (separate EHR system?), assess correlation and mean squared error

Case Study # 1

Predicting posttraumatic stress disorder following a natural disaster

Anthony J. Rosellini [a, b, *], Francisca Dussaillant [c], José R. Zubizarreta [a, d], Ronald C. Kessler [a], Sherri Rose [a]

"Earthquakes are a common and deadly natural disaster, with roughly one-quarter of survivors subsequently developing posttraumatic stress disorder (PTSD). Despite progress identifying risk factors, limited research has examined how to combine variables into an optimized post-earthquake PTSD prediction tool that could be used to triage survivors to mental health services. The current study developed a post-earthquake PTSD risk score using machine learning methods designed to optimize prediction. The data were from a two-wave survey of Chileans exposed to the 8.8 magnitude earthquake that occurred in February 2010. Respondents (n=23,907) were interviewed roughly three months prior to and again three months after the earthquake. Probable post-earthquake PTSD was assessed using the Davidson Trauma Scale (DTS; Davidson et al.,1997)…[where a] DTS total score > 40 indicates a probable PTSD diagnosis… Pre-earthquake [independent] variables included (i) dichotomous indicators of sex, marital status, and living situation, (ii) several dichotomous and categorical employment and education-specific variables (e.g., employment status; years of school), (iii) categorical variables representing the condition of the walls and roof of the home (rated by the interviewer), and (iv) health variables, including a categorical overall health rating and several dichotomous variables representing 30-day health and long-term health problems."

Case Study # 2

Effectiveness of Ceftriaxone Plus Doxycycline in the Treatment of Patients Hospitalized with Community-Acquired Pneumonia

Scott A. Flanders, MD[1]
Vicky Dudas, PharmD[2]
Kathleen Kerr, BA[3]
Charles E. McCulloch, PhD[4]
Ralph Gonzales, MD, MSPH[3,4,5]

"Limited data exist on the effectiveness of ceftriaxone plus doxycycline in the treatment of patients hospitalized with community-acquired pneumonia (CAP). We performed a retrospective cohort study of all adults hospitalized for pneumonia between January 1999 and July 2001 at an academic medical center. Outcomes were compared for patients with CAP treated with ceftriaxone plus doxycycline versus other appropriate initial empiric antibiotic therapies. A total of 216 patients were treated with ceftriaxone plus doxycycline and 125 received other appropriate initial empiric antibiotic therapies. Medical record review by trained research assistants blinded to the research question was used to gather demographic data, comorbid illnesses, physical examination findings on initial presentation, and laboratory or radiographic results on initial presentation. Data from the hospital administrative database were used to identify the initial empiric antibiotic regimen. All antibiotics prescribed within the first 48 hours of hospitalization were considered initial empiric therapy with few exceptions. [Medical records were also] used to identify length of stay, death during the index hospitalization, and return to the emergency department or readmission within 30 days of discharge. The National Death Index was used to identify all deaths that occurred after hospital discharge."

Case Study # 3

Predicting Renal Recovery After Dialysis-Requiring Acute Kidney Injury

Check for updates

Benjamin J. Lee[1,2,3], Chi-yuan Hsu[1,4], Rishi Parikh[4], Charles E. McCulloch[5], Thida C. Tan[4], Kathleen D. Liu[1,6], Raymond K. Hsu[1], Leonid Pravoverov[7], Sijie Zheng[4,7] and Alan S. Go[1,4,5]

"After dialysis-requiring acute kidney injury (AKI-D), recovery of sufficient kidney function to discontinue dialysis is an important clinical and patient-oriented outcome. Predicting the probability of recovery in individual patients is a common dilemma. This cohort study examined all adult members of Kaiser Permanente Northern California (KPNC) who experienced AKI-D between January 2009 and September 2015 and had predicted inpatient mortality of <20%. The primary outcome was recovery of native kidney function after AKI-D, defined as renal replacement therapy (RRT) independence within 90 days after RRT initiation and survival for > 4 weeks after RRT discontinuation. Candidate predictors included demographic characteristics, comorbidities, laboratory values, and medication use. For this analysis, we classified patients as having AKI-D if they underwent RRT (acute intermittent hemodialysis and/or continuous RRT) during hospitalization in the absence of any preadmission chronic RRT and had peak in patient serum creatinine concentration > 50% of preadmission baseline…We excluded patients who had baseline eGFR values <15ml/min per 1.73 m2 (because it is difficult in this eGFR range to distinguish true AKI-D from progression of severe CKD) or predicted probability of in patient mortality > 20% using a KPNC-validated risk score (because the issue of renal recovery is clinically relevant only among those patients with AKI-D who are likely to survive the acute hospitalization)."

Case Study # 4

A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm birth

Nadav Rappoport[1,2], Jonathan Toung[3], Dexter Hadley[1,2], Ronald J. Wong[3
Kazumichi Fujioka[3], Jason Reuter[4], Charles W. Abbott[4], Sam Oh[5], Dong
Celeste Eng[5], Scott Huntsman[5], Dale L. Bodian[6], John E. Niederhuber[6,7], Xiu
Ge Zhang[8], Weronika Sikora-Wohfeld[3], Christopher R. Gignoux[4], Hui Wang[3
Laura L. Jelliffe-Pawlowski[9], Jeffrey B. Gould[3], Gary L. Darmstadt[3], Xiaobin
Carlos D. Bustamante[4], Michael P. Snyder[4], Elad Ziv[5], Nikolaos A. Patsopoul
Louis J. Muglia[8], Esteban Burchard[5], Gary M. Shaw[3], Hugh M. O'Brodovich
David K. Stevenson[3], Atul J. Butte[1,2,5] & Marina Sirota[1,2,5]

"Preterm birth (PTB), or the delivery prior to 37 weeks of gestation, is a significant cause of infant morbidity and mortality. Although twin studies estimate that maternal genetic contributions account for approximately 30% of the incidence of PTB, and other studies reported fetal gene polymorphism association, to date no consistent associations have been identified. In this study, we performed the largest reported genome-wide association study analysis on 1,349 cases of PTB and 12,595 ancestry-matched controls from the focusing on genomic fetal signals. We tested 2,015,750 SNPs for their association to the spontaneous PTB phenotype…with sex and the first ten principal components of genetic ancestry as covariates in the model. The ancestry-matched control population was obtained from the Health and Retirement Study, the vast majority of whom are older than 55 years of age at the time of recruitment, which started in 1992. In the early 1990s, major advances in neonatal care, led by the introduction of surfactant therapy, greatly increased the likelihood of survival for infants born before 30 weeks. Since all of the controls were retirees born well before 1990, we assumed that they were very unlikely to have been born extremely preterm (before 30 weeks), and, therefore, can serve as an appropriate control population for our genome analyses."

Consider the following (suggestions):

1. What type of data? Is there an outcome, if so continuous or categorical?
2. What is the research question (predictive, causal)?
3. What model should you use to investigate the research question?
4. How would you assess model validity/performance?
5. How would you interpret model output to address the research question?

Biostatistics 202: Machine Learning Guide

Supervised learning - prediction of a target

	Model	Interpretability?
Regression - continuous target		
	Linear regression	High
Classification - categorical target		
	Logistic regression	High
	Classification trees (decision trees)	High
	K-nearest neighbors	Low
	Support vector machines	Low
	Artificial neural networks	Low
Ensemble methods	Random forests (bagging)	Medium
	Adaptive boosted trees (boosting)	Medium
	Stacking	Low

*Note: all classification methods may be adjusted to allow for prediction of continuous targets.

Unsupervised learning - clustering and/or data reduction

Clustering	
	k-Means
	Heirarchical clustering
Data Reduction	
	Principal component analysis
	t-distributed stochastic neighbor embedding

FILL IN WITH YOUR GROUP

- **Goal:**
- **Target:**
- **Predictors:**
- **Techniques:**
- **Validation**:

Next: visualization lab