Breiman L.:

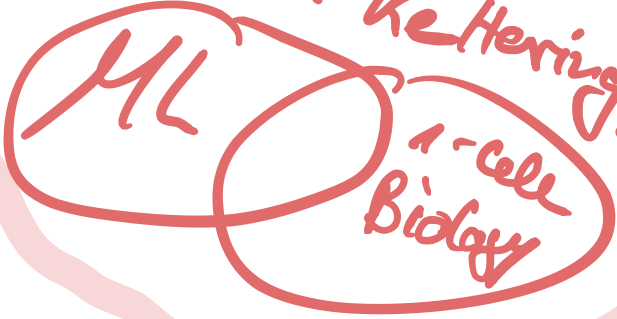
Bin Yu (UC Berkeley)

- Veridical DS -

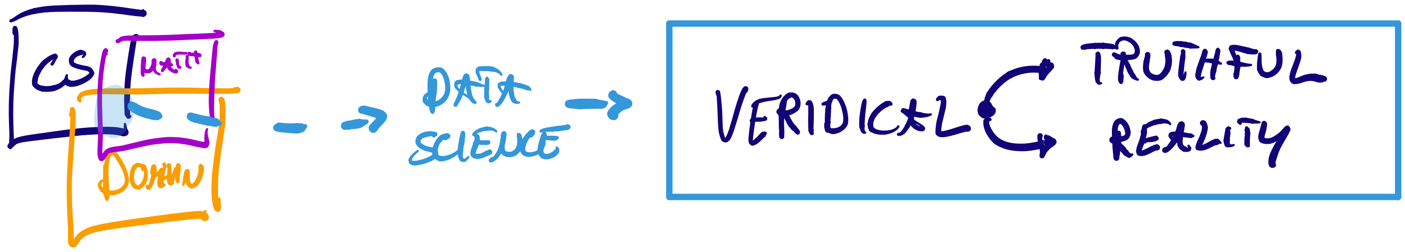
NeuRIPS 2019
- Day 2 -

Keynote:

Dana Pe'er (Sloan Kettering)



Bin Yu (UC Berkeley): Veridical Data Science



⇒ Better Communication ⊕ Rigorous Evaluation Needed!

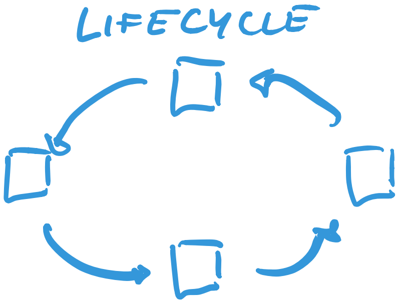
PCS FRAMEWORK

PREDICTABILITY (CS)

COMPUTABILITY (CS)

STABILITY (STATS)

STABILITY → Robustness → "Shaking" all parts of lifecycle



⇓

PERTUBATIONS

* Data (Clean)
* Model (R2R)

PROB. INFERENCE

* Data as realisation = assumption!

↳ Stability: E.g. from same RV / distribution

* What does p-value mean?
* p-value as measure of model bias

⇓

1. Problem Formulation
2. Prediction "Screening"
3. Target Value Perturbation Distribution
4. Summarization

ITERATIVE RANDOM FORESTS (iRF)

⊙ Drosophila → 4 interacting genes

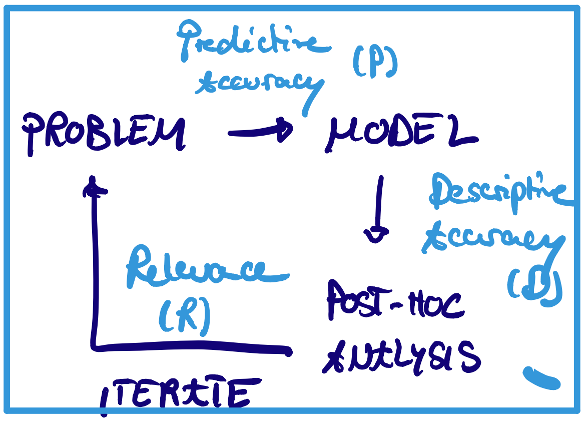
⊙ RF ⇒ perturbating features + data

⊕ Soft-dim. reduction

⊕ Random Intersection Trees

⇒ INTERPRETATION ⇒ HYPOTHESIS GEN.

≈ EXTRACTION OF KNOWLEDGE FROM MODEL



AGGLOMERATIVE CONTEXTUAL DECOMPOSITION (ACD)

Dana Pe'er (Sloan Kettering): ML meets Single-Cell Biology

TINY COMPUTER

⊙ Key Challenge Single-Cell: Data Analysis of RNA Sequencing ⇒ GENE X CELL MATRIX

⊙ CELL PHENOTYPES → Low-Dim. Manifold → NHOOD GRAPH TRAVERSAL
⇒ Shaped by regulatory nets & feedback

↙ DENSE REGIONS

↘ CELL STATE TRANSITIONS

↳ spectral Clustering of paths through graph

⊙ BEAUTY OF TISSUE → Single sample contains abundance of cells at different maturity levels → ASYNCHRONOUS

⊙ SIMPLE MODELLING → e.g. Markov Chain

↳ KEY ASSUMPTION: Development moves forward ⇒ Not given in disorders

↳ Spatio-temporal map of mammalian endoderm

⊙ HCA – Human Cell Atlas

Current(-ish) views from the Atlases

BIOLOGY GOAL: NOT PREDICT BUT UNDERSTAND

↳ Importance of Outliers!

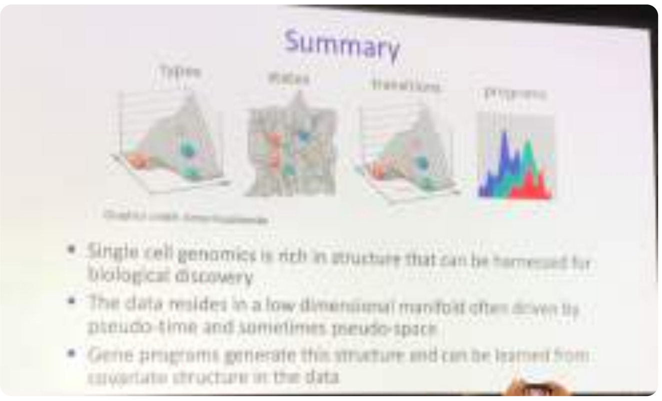
CELL TYPES ↔ CLUSTERS ↔ STATES

⊙ Use raw image directly instead of intermediate segmentation output

COVARIATION-DRIVEN MANIFOLD LEARNING

↓

Covariation Between Components



Summary

- Single cell genomics is rich in structure that can be harnessed for biological discovery
- The data resides in a low dimensional manifold often driven by pseudo-time and sometimes pseudo-space
- Gene programs generate this structure and can be learned from covariate structure in the data