freenome

(What to do) when gradient descent digs too deep, and too greedily

Imran S Haque

7th DeepChem User Group Meeting
3 July 2018

@imranshaque
@freenome
freenome.com/careers

About me

- PhD CS, **Stanford** (Pande Lab) - large-scale machine learning for drug discovery
- 2011-2016, **Counsyl**: genomics technology development, large-scale medical/population genetics research
- 2016-, **Freenome**: CSO

About Freenome

- Goal: a two-pronged attack on the mortality curve of cancer through **early detection** and **early intervention**
- Founded 2015, seed funding 2016, Series A 2017 (a16z, GV, Polaris, Founders Fund, *et al.*); >$77M total
- Based three floors up from where we're sitting.

freenome

Early detection

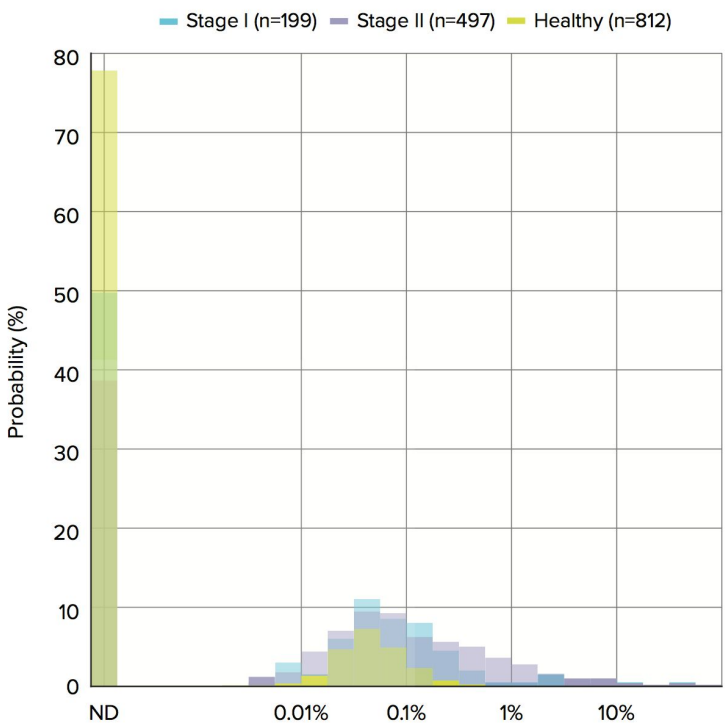
Early-stage disease detection based on multi-analyte pattern recognition

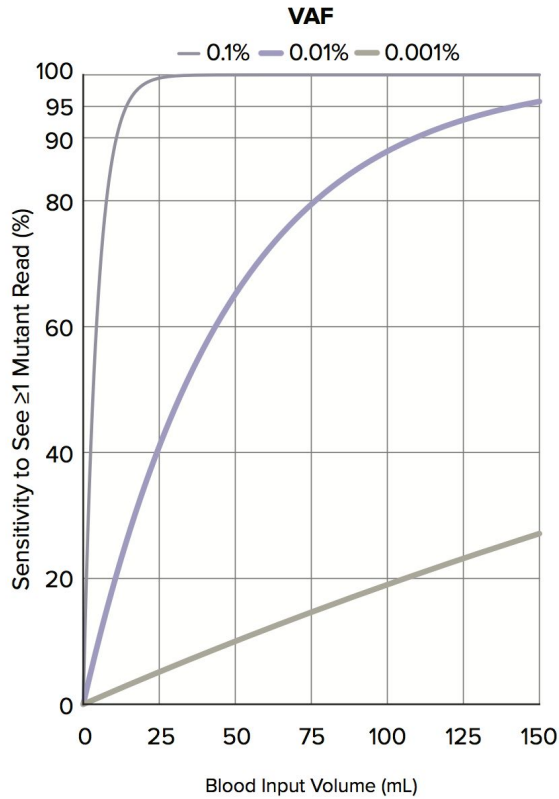
Precision oncology

PD-1, PARP, and other IO/epigenetic modulator response prediction

Tumor-derived mutations are *incredibly* rare in the blood at early stage

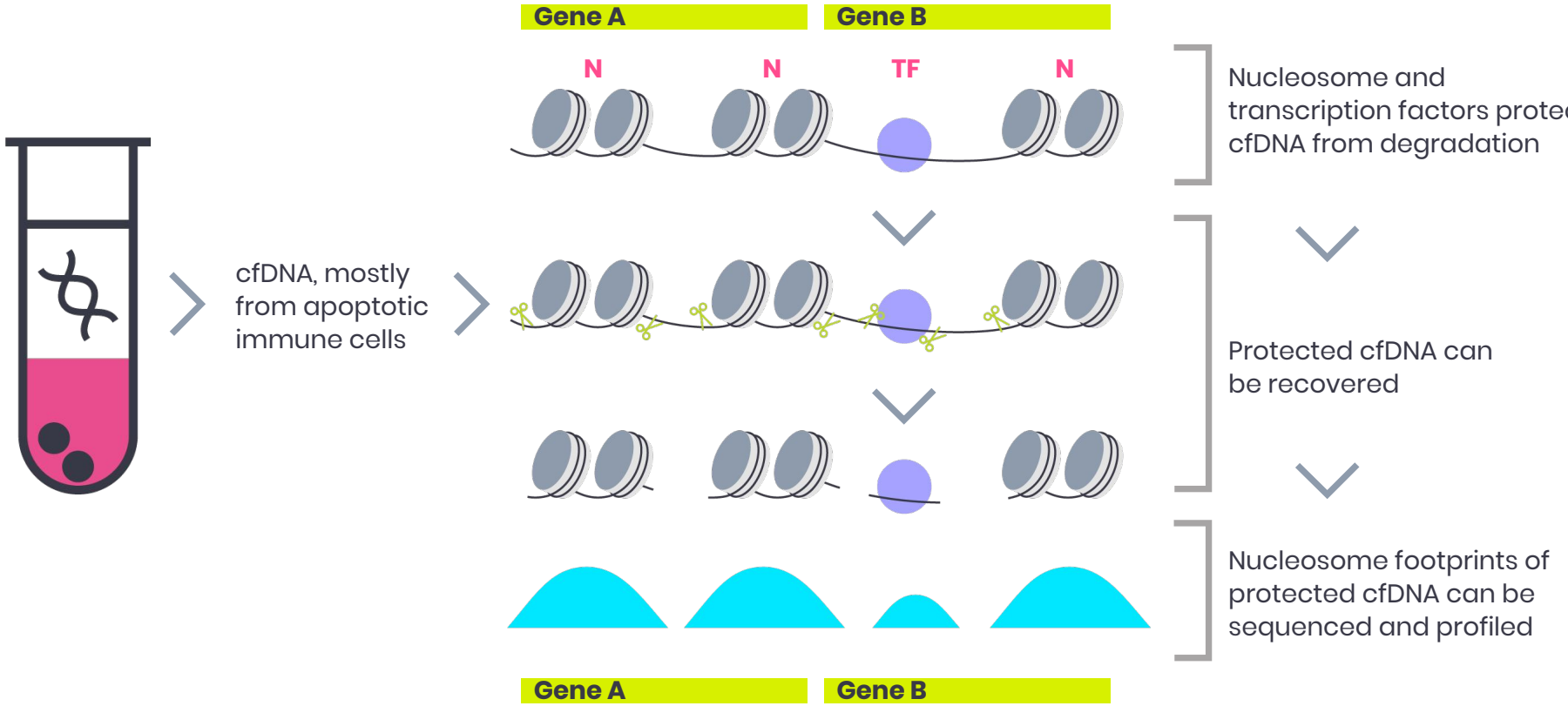
- **Over half of stage I cancer patients have ctDNA concentrations below 0.01–0.1%.**
- Binomial counting statistics imply that to be able to get **even one** mutant molecule from a site mutated at 0.01% would require **~150mL blood draw.**
- Intuition: if you can't get the 0.01%, find signal in the other 99.99% of material.

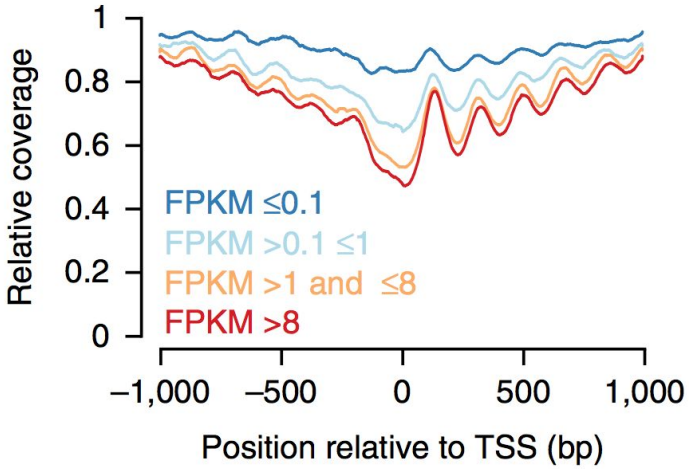




Haque IS and Elemento OE. bioRxiv 2017
Cohen JD et al. Science 2018
Phallen J et al. Sci Transl Med 2017

cfDNA beyond sequence: gene expression from chromatin structure

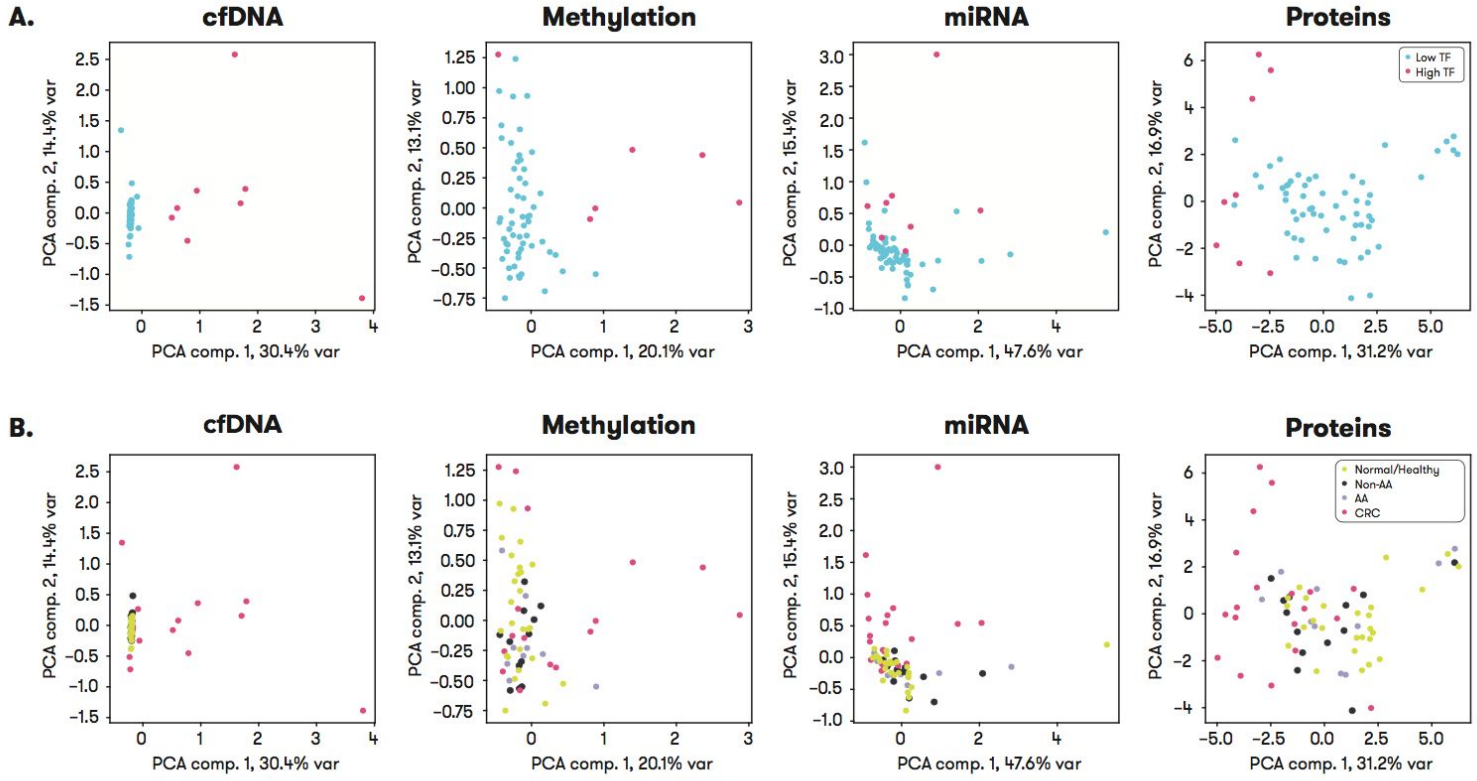




Ulz P et al. Nat Genet 2016

Multi-analyte assays drive data integration and multimodal inference

- Freenome assays multiple analytes in a single sample to recover multiple dimensions of expression and regulation
- Goal: use (multi-modal) machine learning to recover the presence/ absence of cancer from diffuse, genome-wide signals of chromatin structure and gene expression



(A) PCA as a function of TF. (B) PCA as a function of tissue class. High TF samples have consistently aberrant behavior across all 4 analytes investigated.

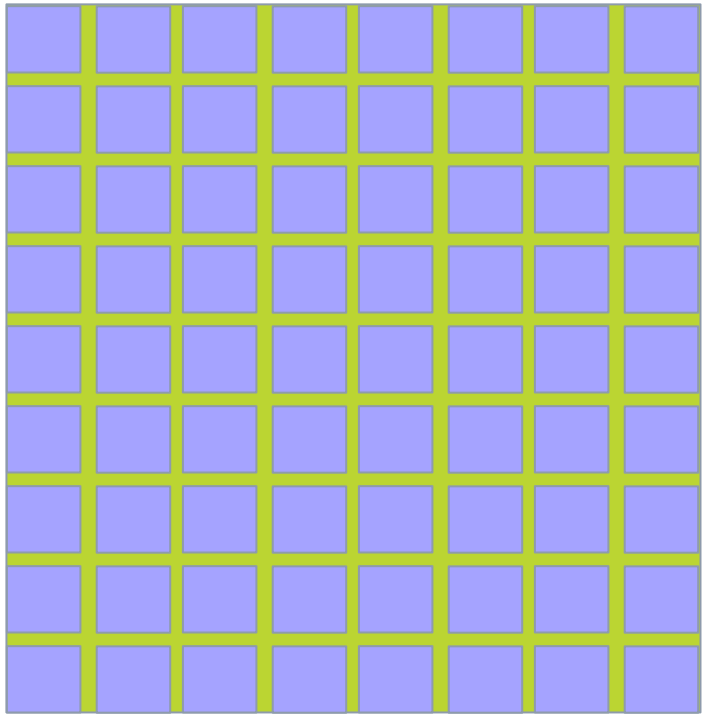
Delubac D et al, AACR 2018.

The shape of data: what does 1 TB look like?

"Traditional" ML domains (recommendations, vision, language, etc.)

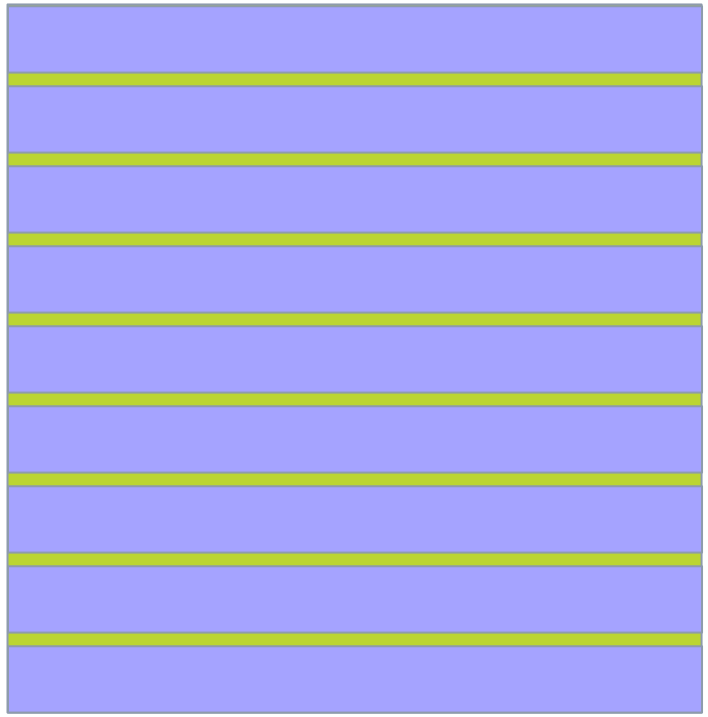
Biology and chemistry (genomics, drug discovery, etc.)

The shape of data: what does 1 TB look like?



Traditional ML

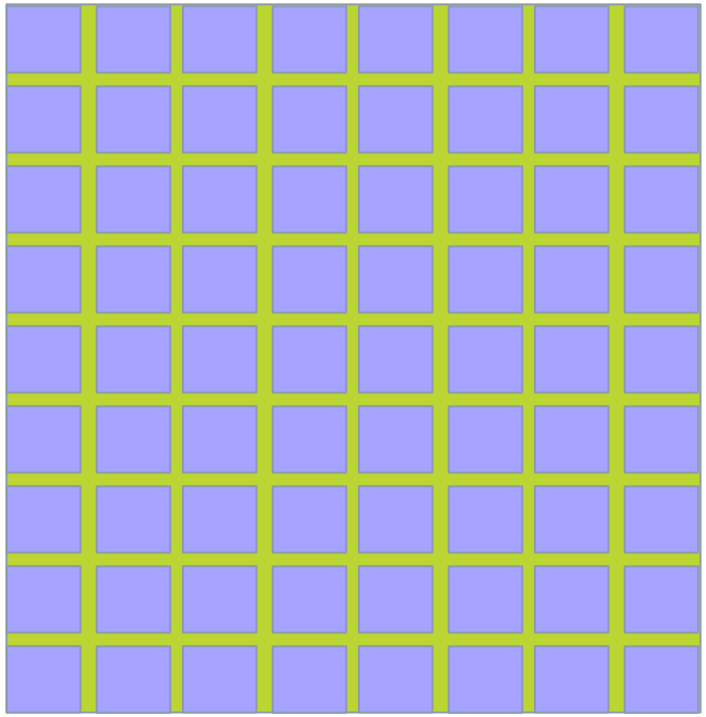
- **Many, small** samples (kB-MB)
- Cheap(ish) to acquire new instances (~$0.01)
- Easy(ish) to sample over measurement error



Biological ML

- **Few, large** samples (GB)
- Expensive to acquire new samples ($1000)
- Pervasive measurement error

The shape of data: what does 1 TB look like?



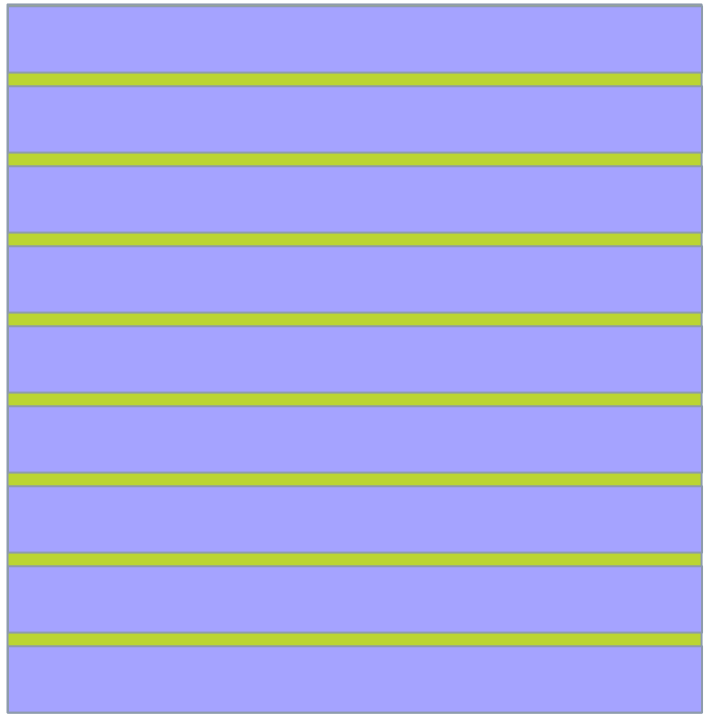
Traditional ML

- **Many, small** samples (kB-MB)
- Cheap(ish) to acquire new instances (~$0.01)
- Easy(ish) to sample over measurement error

"Bioinformatics is the study of batch effects"

The unique challenge in biological ML: samples are scarce and labels are always confounded by sample-, source-, and batch-specific errors.

Not handling these effects might be the easiest way to fail to generalize.



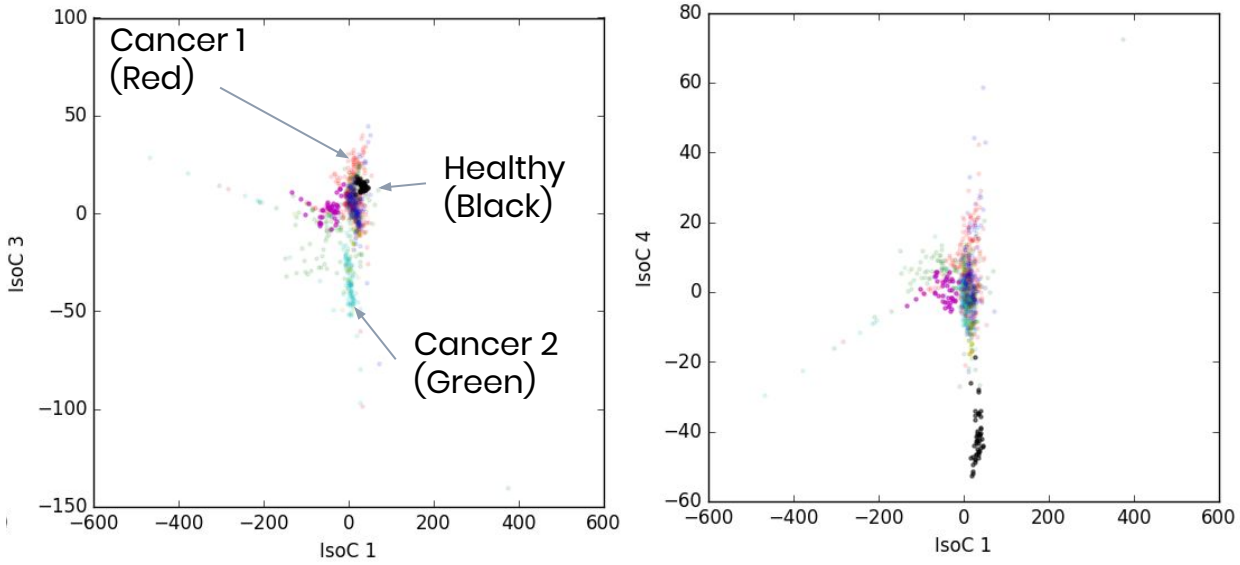
Biological ML

- **Few, large** samples (GB)
- Expensive to acquire new samples ($1000)
- Pervasive measurement error

Batch Effects

Beware the data

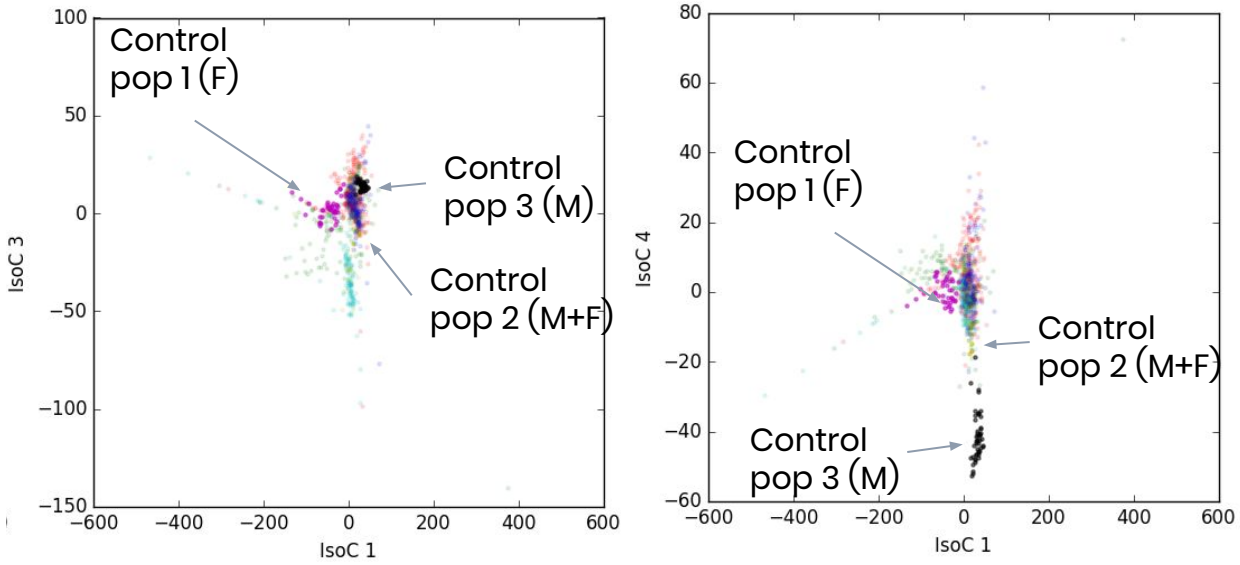
- Check out NCBI SRA: tons of public sequencing data as a great starter resource.
- But: ability to train a model (even in cross-validation) on lumped public data may be meaningless because of unknown batch effects.



Example: Isomap of public sequencing data from a single lab (name withheld to protect the innocent) on cancers + healthies. **Sweet - we can separate cases from controls!**

Beware the data

- Check out NCBI SRA: tons of public sequencing data as a great starter resource.
- But: ability to train a model (even in cross-validation) on lumped public data may be meaningless because of unknown batch effects.

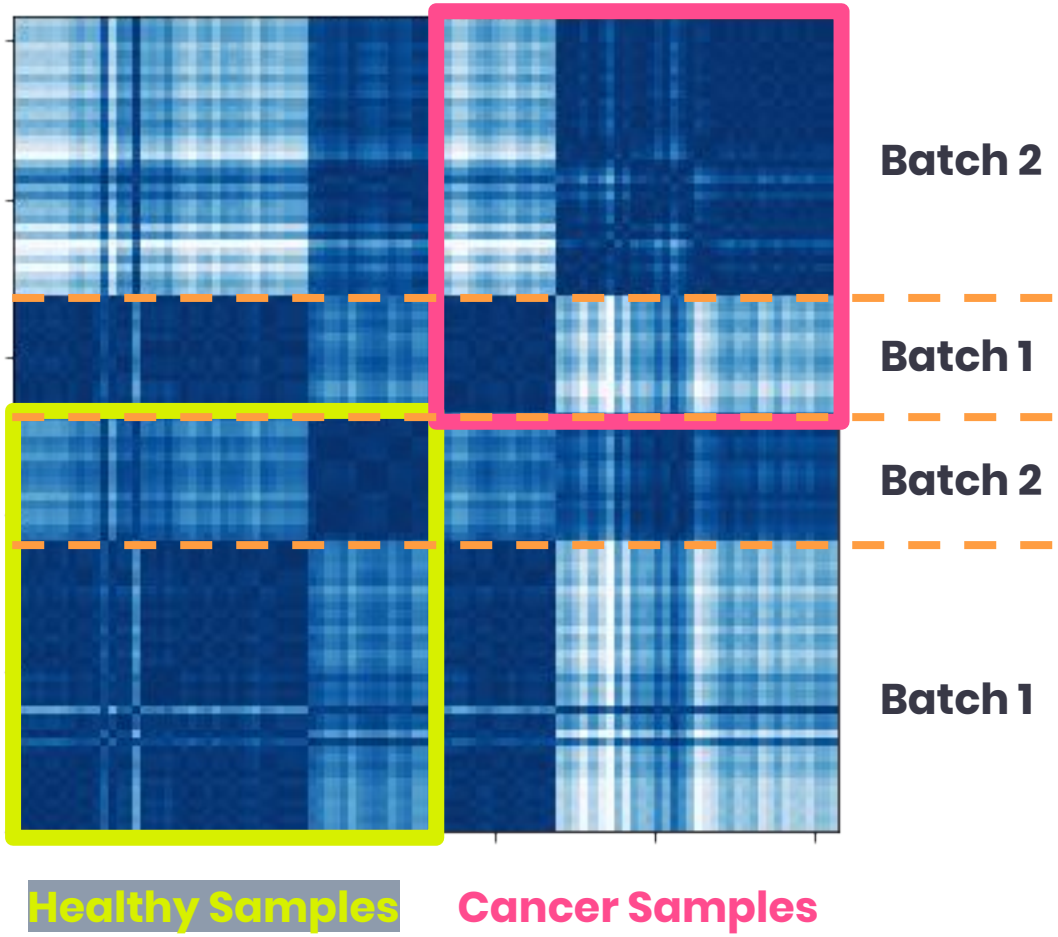


Example: Isomap of public sequencing data from a single lab (name withheld to protect the innocent) on cancers + healthies. **Non-cancer samples do not cluster together: separated by sequencing batch/protocol.**

Protip 0: Look at your data

- A simple examination of sample correlation (in the learned feature space) shows that samples' feature vectors correlate more strongly with those in the same batch than with those in the same class but different batches.

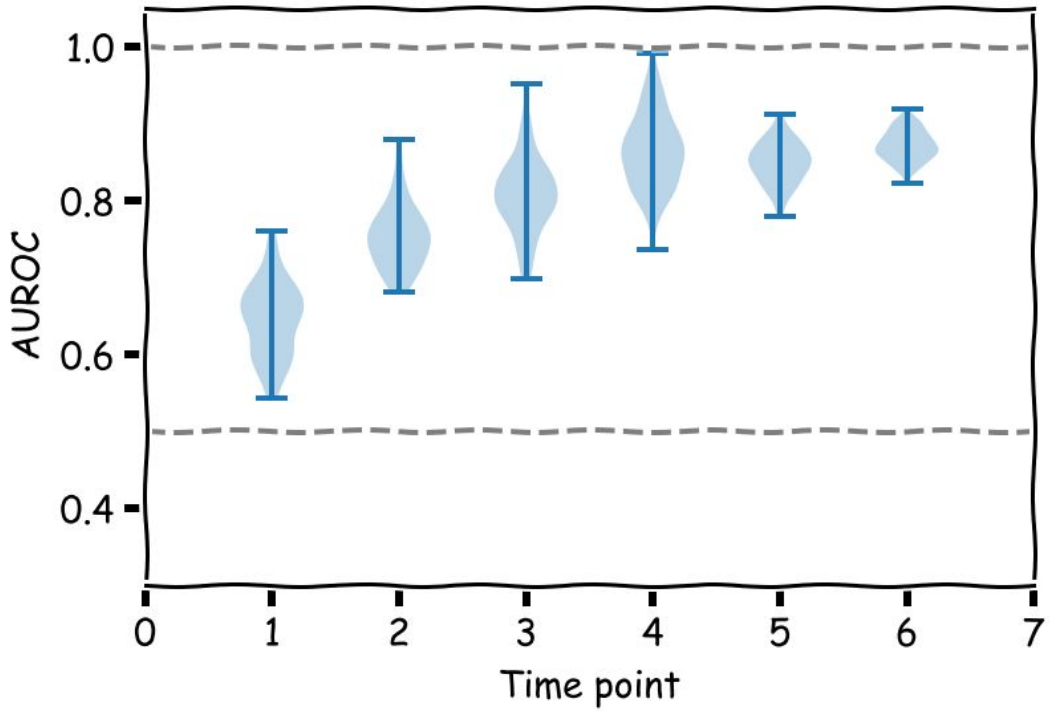
- **Simple plots** (pairwise correlation, PCA/Isomap/t-SNE) **are essential; don't leave home without them.**



Protip 1: Discrete learning curves

- Assuming you're acquiring more data as time goes on, you expect your performance to improve.
- **Good:** Plotting cross-validation performance with CIs at each time point is a super simple sanity check.

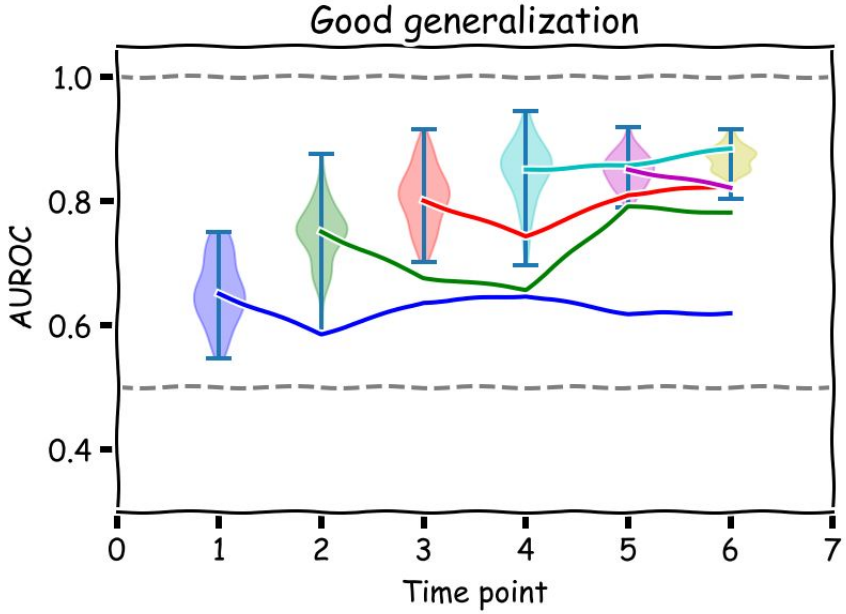
  *(Note: this is **not** your training curve; it's a discretization of your learning curve)*



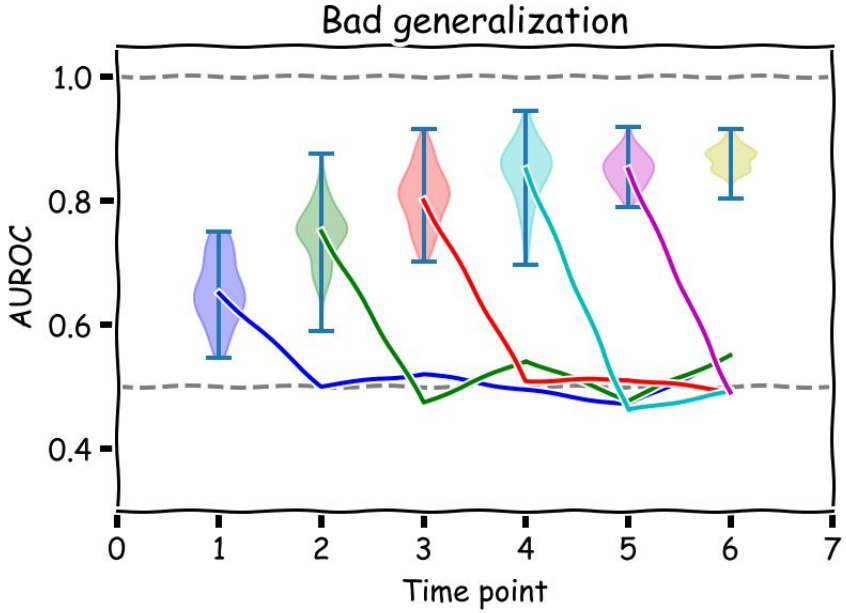
Protip 2: Performance through time

- Assuming you're acquiring more data as time goes on, you expect your performance to improve.
- **Better:** Freeze each timepoint's model, and evaluate on new data as it comes in as a new test set each time.

  A model that is generalizing well should stay within its predicted confidence intervals.



Good models resilient to batch effect will stay within their confidence bounds with each new timepoint (batch).



Bad models confounded by batch effect may fall to random performance on new timepoints (batches).

Protip 3: Down with **KFold**

- Key insight from tips #1/2: in the presence of batch effects, taking random samples of existing data is **not** the same as testing new samples, because new samples are from a new batch.
- There's nothing like more data. (mmmm, data....)
- In the absence of new data, the first thing to do is delete `KFold` from your vocabulary. **Stratify your cross-validation by batch (or other relevant covariates), not at random**, to simulate the new-data experiment.

Batch effect correction and domain adaptation

- There's a rich literature on addressing batch effects in genomics data (e.g., ComBat, HCP, and more recently loads of methods for scRNA-seq), but none is ideal.
- Similar problems have been addressed in the machine learning literature under the name **domain adaptation**, with interesting work on using adversarial training to perform the adaptation/transfer.
- Deep networks on genomic data are unstable and hard to train. Adversarial networks are unstable and hard to train. This sounds like a great area for further research…

Johnson WE, Rabinovic A, Li C. Biostatistics 2007
Mostafavi S et al. PLoS One 2013
Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Nat Biotech 2018
Ganin Y et al. JMLR 2016

Conclusions

Conclusions

- Genomics is a great application area for machine learning, but has unique data attributes (small numbers of large instances with high batch-to-batch variability) that make it challenging compared to other domains.

- Active steps during model development to control batch effect are essential: visualization, monitoring through time, batch-stratified holdouts.

- Development of improved batch-effect mitigation algorithms is a fundamental problem of great importance for the ability to do good genomic DL.

</talk>

@imranshaque

ihaque@freenome.com

Questions?

Thanks to the Freenome R&D team for the data, methods, and hard work that made this talk possible!

