

Multi-Omics Data Integration via Machine Learning

Omics Integration and Systems Biology course

Nikolay Oskolkov, Lund University, NBIS SciLifeLab, Sweden

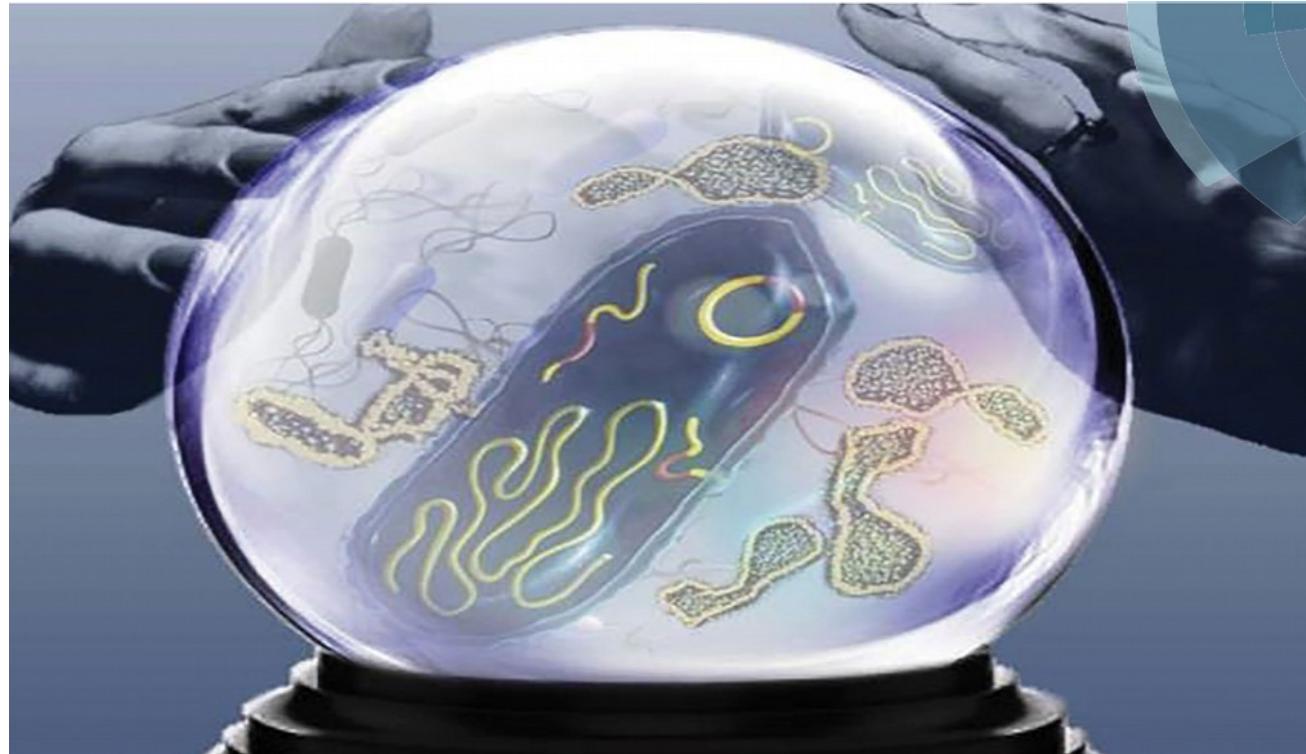
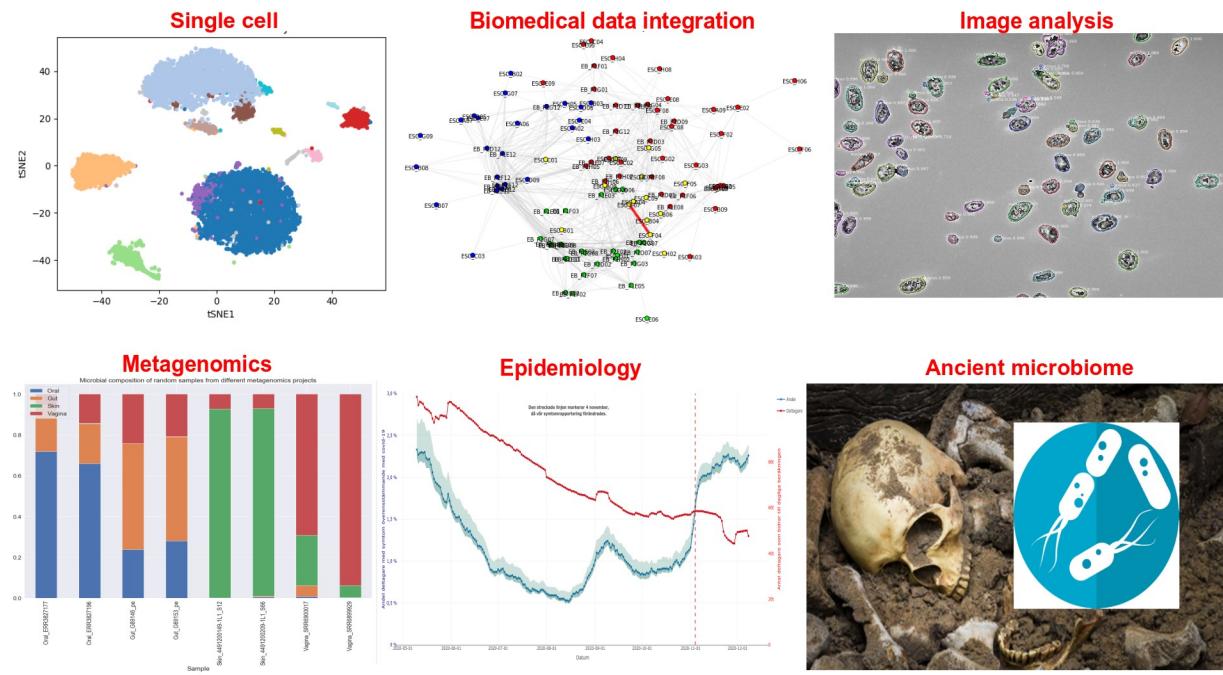
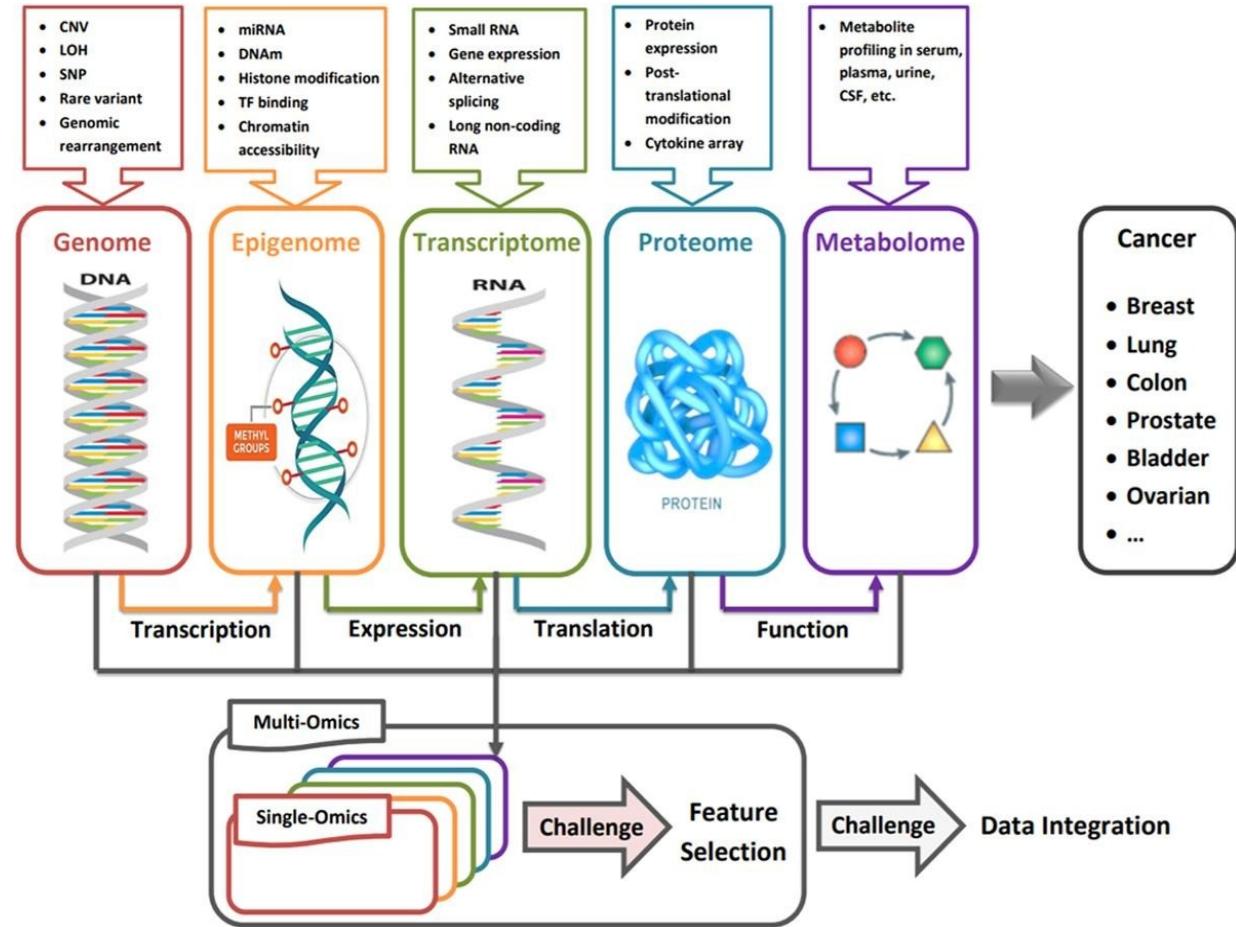


Image adapted from Molecular Omics, Issue 1, 2018

- 2007 PhD in theoretical physics
- 2011 medical genetics at Lund University
- 2016 working at NBIS SciLifeLab, Sweden





The image shows a composite screenshot of two web applications. The top half displays the 'ELIXIR Omics Integration and Systems Biology' workshop landing page, featuring a large, colorful network graph illustrating biological data integration. The bottom half shows a GitHub repository for 'NBISweden / workshop_omics_integration', specifically the 'workshop-common' branch, displaying a detailed list of commit history and repository metadata.

Introduction: High Dimensional Biological Data

Tabular

Text

Editing Wikipedia articles on **Medicine**

[View Details](#)

Editing Wikipedia can be daunting for novices, especially if you're contributing to Wikipedia for the first time as a class assignment. This guide is designed to assist students who have been assigned to contribute biomedical related content to Wikipedia. Here's what other editors will expect you to know.

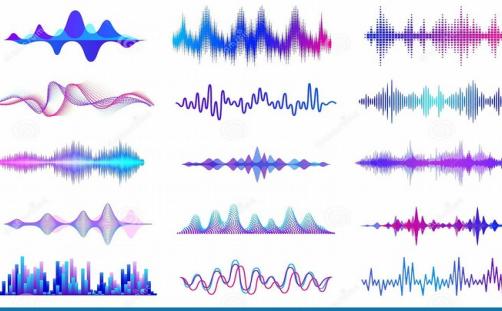
Be accurate
You're editing a resource millions of people use to make medical decisions, so it's vitally important to be accurate. Wikipedia is used more for medical information than the websites for WebMD, NIH, and the WHO. But with great power comes great responsibility!

Understand the guidelines

Wikipedia editors in the medicine area have developed additional guidelines to ensure that the content on Wikipedia is medically sound. Take extra time to read and understand these guidelines. When you edit an article, ensure your changes meet these special requirements. If not, your work is likely to be undone by other editors who do not follow these guidelines.

takes valuable volunteer time away from creating content. If you're not comfortable working under these guidelines, talk to your instructor about an alternative off-wiki assignment.

Sound

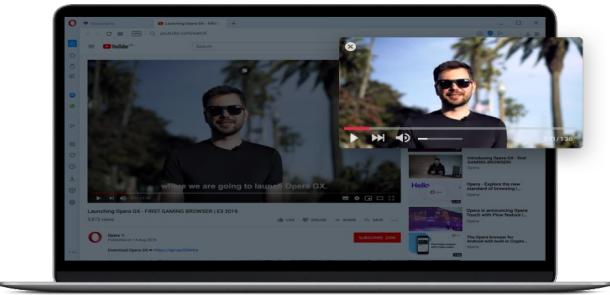


 dreamstime.com

ID 142115245 © Spicytruffel

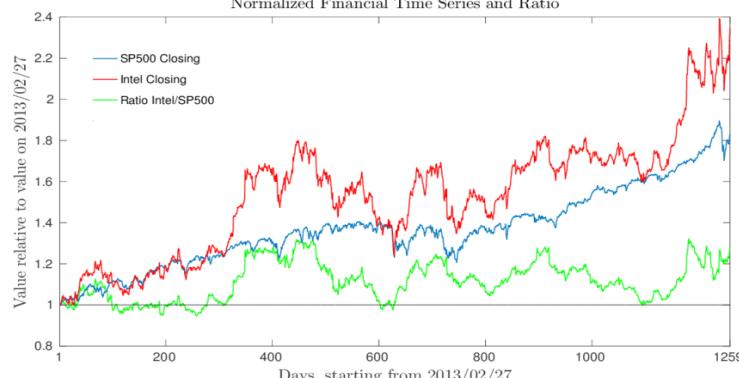
DATA

Video



Time Series

Normalized Financial Time Series and Ratio

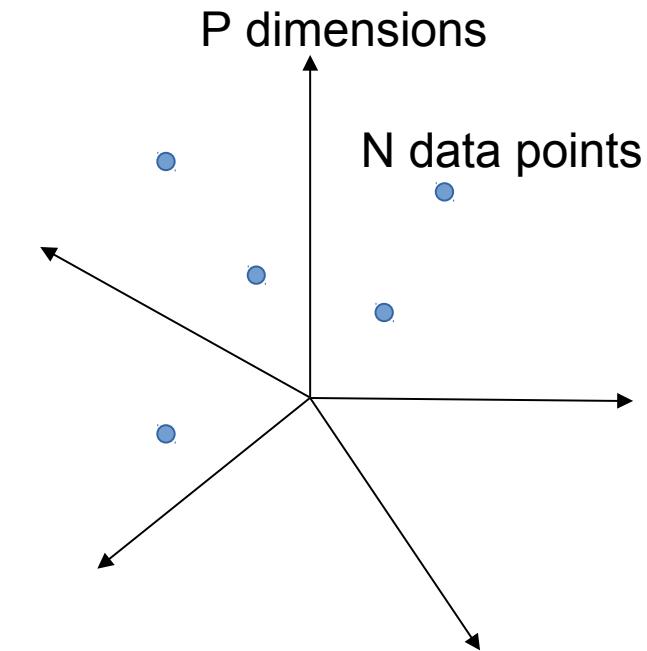


Statistical observations:
e.g. samples, cells etc.

Features: genes, proteins,
microbes, metabolites etc.

N

0	3	1	0	2	3	8	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	6	7	1	2	2
1	2	3	10	0	4	6	1	0	5
3	2	2	1	4	3	2	1	6	0
7	4	4	5	3	9	6	1	6	1
7	1	1	5	2	8	9	1	3	6
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	8	2



High Dimensional Data:
P >> N

For a robust statistical analysis, one should properly “sample” the P-dimensional space, hence large sample size is required, $N \gg P$

Types of Statistical Analysis

P is the number of features (genes, proteins, genetic variants etc.)
N is the number of observations (samples, cells, nucleotides etc.)

Biology / Biomedicine

Bayesianism



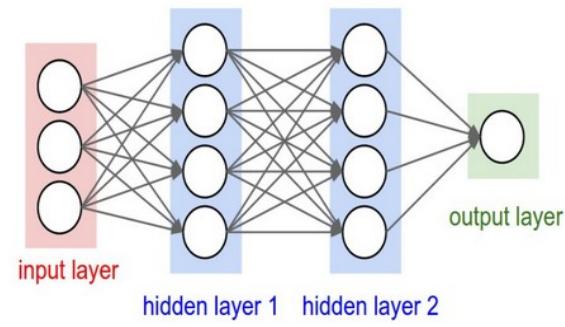
$P \gg N$

Frequentism



$P \sim N$

Deep Learning



$P \ll N$

The Curse of Dimensionality



Amount of Data

Ex.1

$$Y = \alpha + \beta X$$

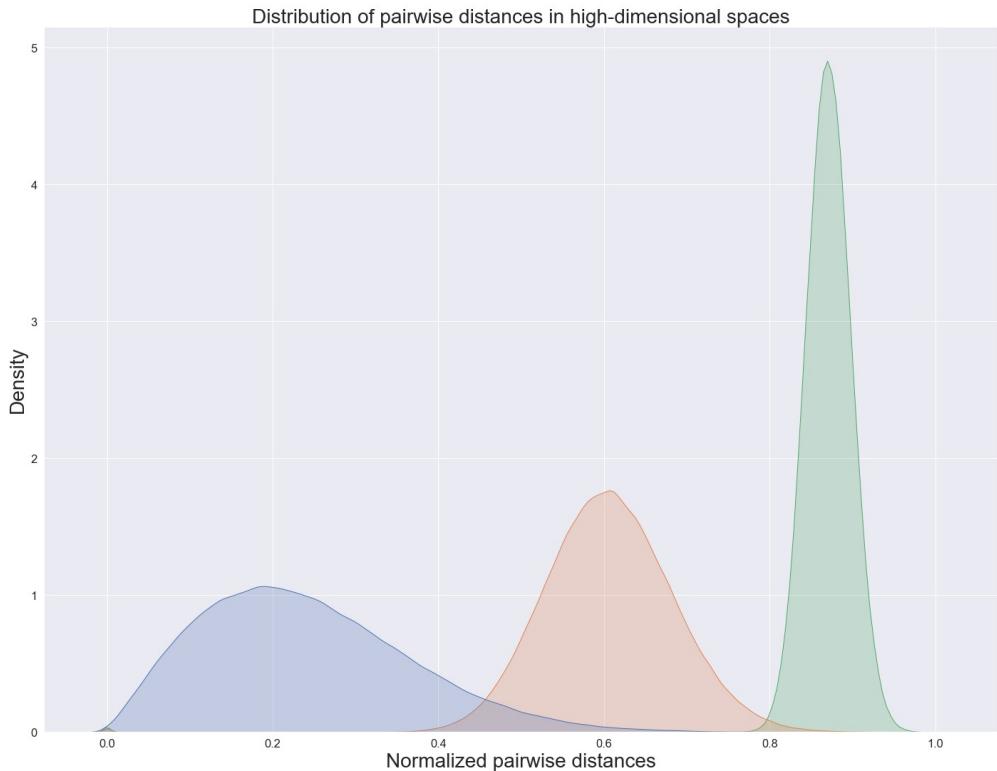
$$\beta = (X^T X)^{-1} X^T Y$$

$$(X^T X)^{-1} \sim \frac{1}{\det(X^T X)} \dots \rightarrow \infty, \quad n \ll p$$

$$\text{Ex.2} \quad E[\hat{\sigma}^2] = \frac{n-p}{n} \sigma^2$$

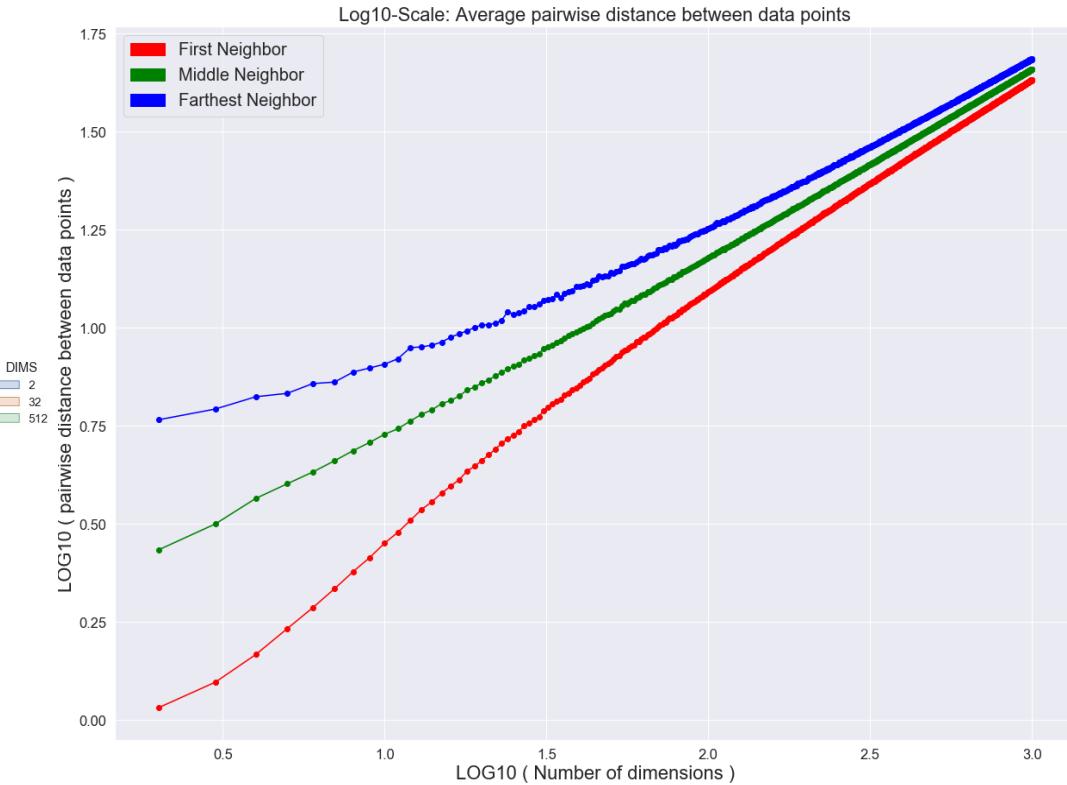
Biased ML variance estimator in HD-space

More on the Curse of Dimensionality

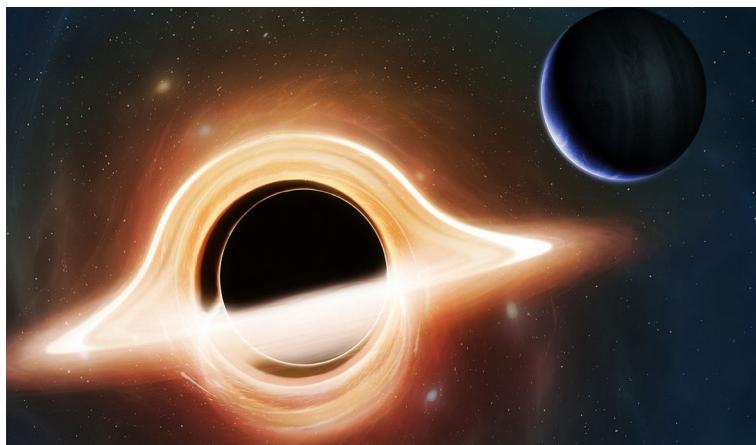
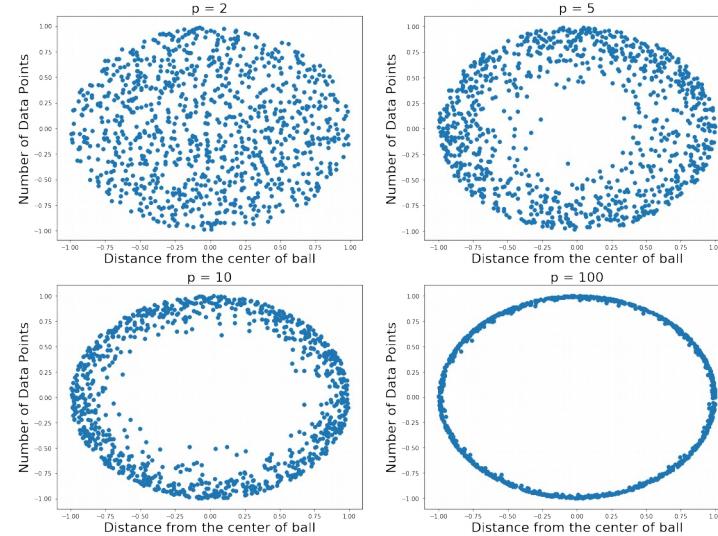
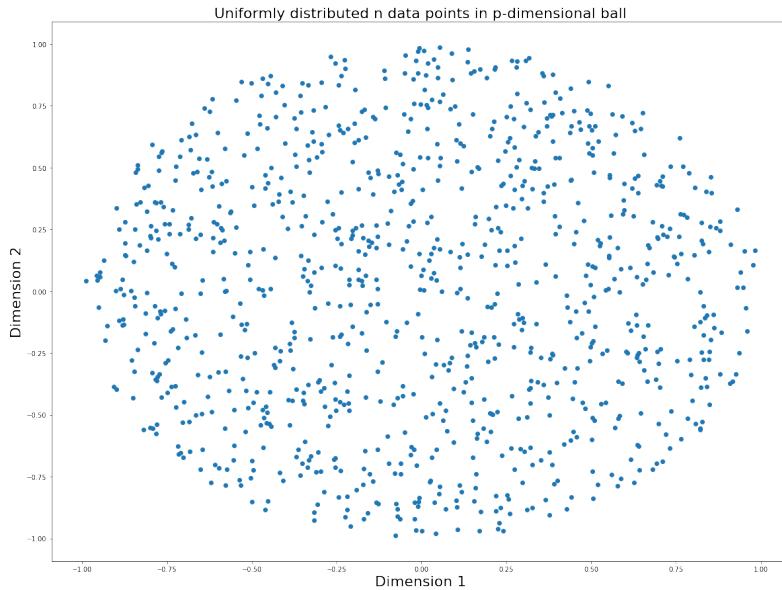


Data points become far from each other and equidistant in high dimensions

In high-dimensional space we can not separate cases and controls any more



The differences between closest and farthest data point neighbours disappears in high-dimensional spaces: can't run cluster analysis



High-dimensional data can be viewed as having a “**hole in the middle**”, hence the concept of mean / centroid loses its validity, hence we can’t use Gaussian distribution

The curse(s) of dimensionality

There is such a thing as too much of a good thing.

Naomi Altman and Martin Krzywinski

We generally think that more information is better than less. However, in the 'big data' era, the sheer number of variables that can be collected from a single sample can be problematic. This embarrassment of riches is called the 'curse of dimensionality'¹ (CoD) and manifests itself in a variety of ways. This month, we discuss four important problems of dimensionality as it applies to data sparsity^{1,2}, multicollinearity³, multiple testing⁴ and overfitting⁵. These effects are amplified by poor data quality, which may increase with the number of variables.

Throughout, we use n to indicate the sample size from the population of interest and p to indicate the number of observed variables, some of which may have missing values for some samples. For example, we may have $n = 1,000$ subjects and $p = 200,000$ single-nucleotide polymorphisms (SNPs).

First, as the dimensionality p increases, the 'volume' that the samples may occupy grows rapidly. We can think of each of the n

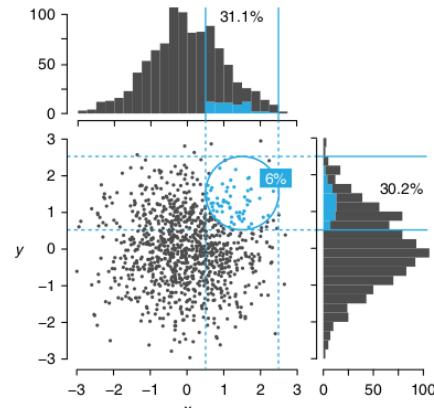


Fig. 1 | Data tend to be sparse in higher dimensions. Among 1,000 (x, y) points in which both x and y are normally distributed with a mean of 0 and s.d. $= 1$, only 6% fall within σ of $(x, y) = (1.5, 1.5)$ (blue circle). However, when the data are projected into a lower dimension—shown by histograms—about 30% of the points (all bins within blue solid lines) fall within ± 1.5 plus

A and 100 to have the minor allele a. If we tabulate on two SNPs, A and B, we will expect only ten samples to exhibit both minor alleles with genotype ab. With SNPs A, B and C, we expect only one sample to have genotype abc, and with four or more SNPs, we expect empty cells in our table. We need a much larger sample size to observe samples with all the possible genotypes. As p increases, we may quickly find that there are no samples with similar values of a predictor.

Even with just five SNPs, our ability to predict and classify the samples is impeded because of the small number of subjects that have similar genotypes. In situations where there are many gene variants, this effect is exacerbated, and it may be very difficult to find affected subjects with similar genotypes and hence to predict or classify on the basis of genetic similarity.

If we treat the distance between points (e.g., Euclidian distance) as a measure of similarity, then we interpret greater distance as greater dissimilarity. As n increases, this

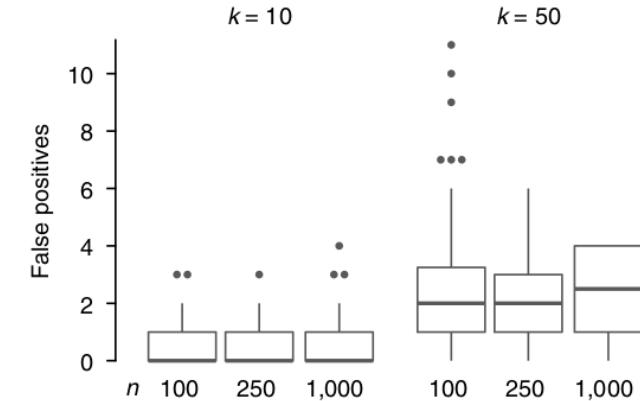


Fig. 3 | The number of false positives increases with each additional predictor. The box plots show the number of false positive regression-fit P values (tested at $\alpha = 0.05$) of 100 simulated multiple regression fits on various numbers of samples ($n = 100, 250$ and $1,000$) in the presence of one true predictor and $k = 10$ and 50 extraneous uncorrelated predictors. Box plots show means (black center lines), 25th and 75th percentiles (box edges), and minimum and maximum values (whiskers). Outliers (dots) are jittered.

Correcting for multiple testing does not solve the problem of too many false-positive hits

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \text{Metabolomics} \\ N \approx P$$

Metabolomics

N ≈ P

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \quad \text{Proteomics} \quad N \approx P$$

Proteomics N ≈ P

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \text{Metagenomics} \\ N \approx P$$

Metagenomics N ≈ P

— manageable

$$P = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

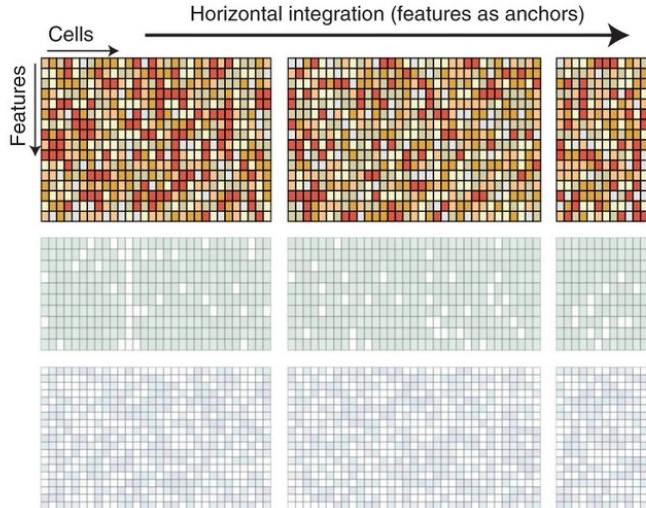
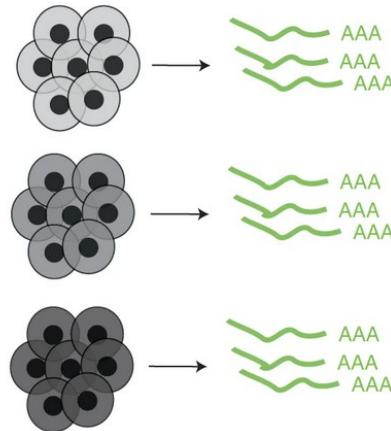
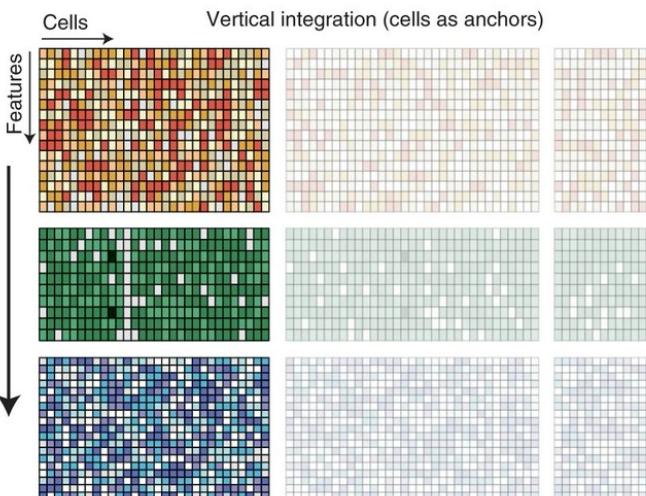
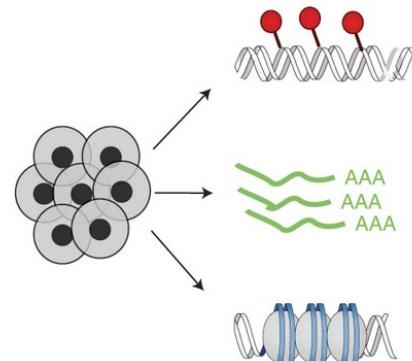
Transcriptomics N << P (Single cell: N <= P)

challenging

Genomics
N <<< P

Methylomics N <<< P

Multi-Omics Data Integration

a**b**

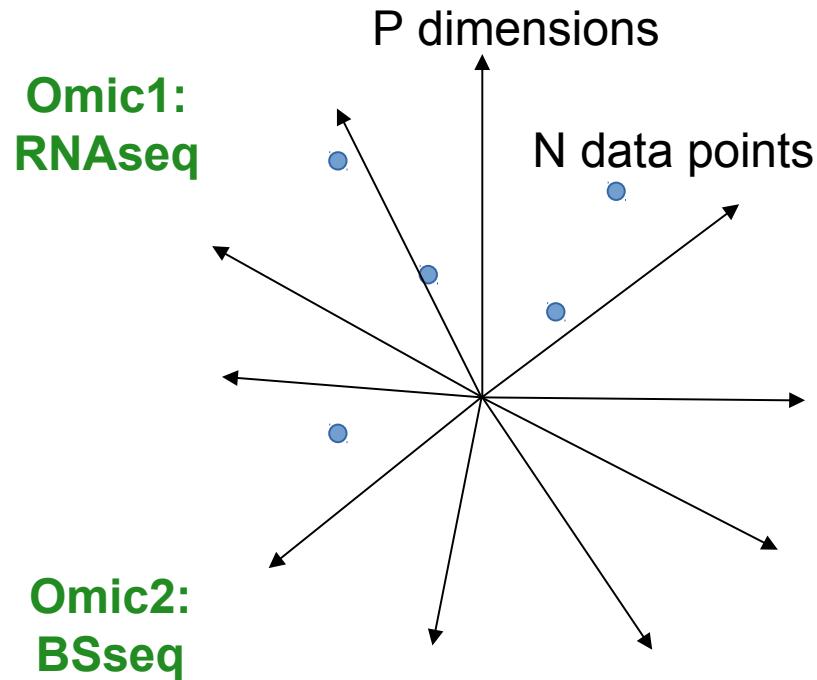
Statistical observations:
e.g. samples, cells etc.

	N									
0	3	1	0	2	3	8	1	1	3	
1	1	0	0	7	1	2	2	3	3	
1	2	2	0	0	6	7	1	2	2	
1	2	3	10	0	4	6	1	0	5	
3	2	2	1	4	3	2	1	6	0	
7	4	4	5	3	9	6	1	6	1	
7	1	1	5	2	8	9	1	3	6	
5	0	1	6	2	0	0	0	1	5	
1	6	3	3	4	6	2	0	1	1	
1	2	2	4	1	1	3	0	8	2	

Features: genes, proteins,
microbes, metabolites etc.

	N									
0	3	1	0	2	3	8	1	1	3	
1	1	0	0	7	1	2	2	3	3	
1	2	2	0	0	6	7	1	2	2	
1	2	3	10	0	4	6	1	0	5	
3	2	2	1	4	3	2	1	6	0	
7	4	4	5	3	9	6	1	6	1	
7	1	1	5	2	8	9	1	3	6	
5	0	1	6	2	0	0	0	1	5	
1	6	3	3	4	6	2	0	1	1	
1	2	2	4	1	1	3	0	8	2	

P₂



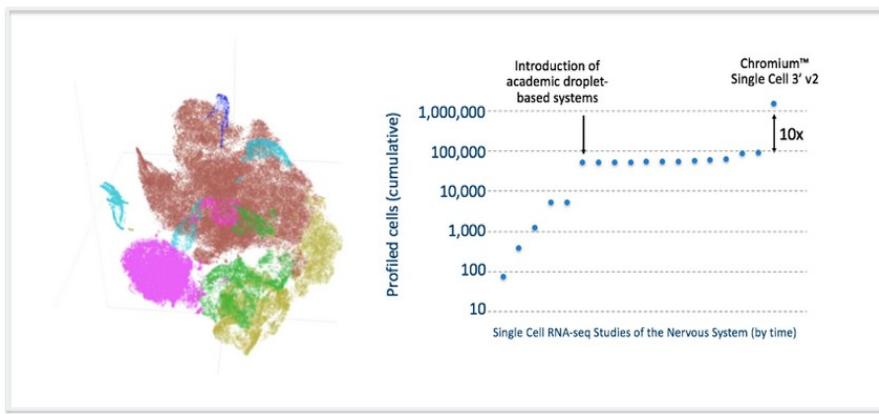
P₁ + P₂ >> N integration across features leads to even more high-dimensional data

CAREERS BLOG 10X UNIVERSITY

10X GENOMICS SOLUTIONS & PRODUCTS RESEARCH & APPLICATIONS EDUCATION & RESOURCES

< Back to Blog

< Newer Article Older Article >



Our 1.3 million single cell dataset is ready 0 KUDOS



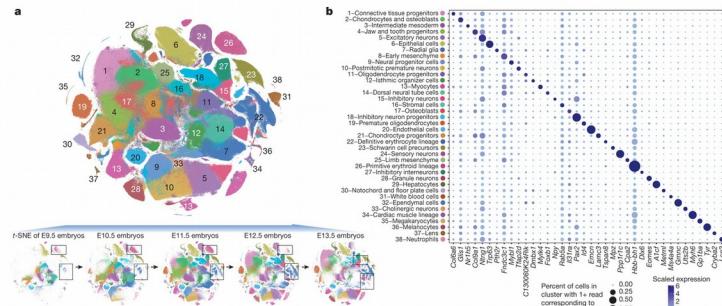
POSTED BY: grace-10x, on Feb 21, 2017 at 2:28 PM

At ASHG last year, we announced our 1.3 Million Brain Cell Dataset, which is, to date, the largest dataset published in the single cell RNA-sequencing (scRNA-seq) field. Using the Chromium™ Single Cell 3' Solution (v2 Chemistry), we were able to sequence and profile 1,308,421 individual cells from embryonic mice brains. Read more in our application note [Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium™ Single Cell 3' Solution](#).

MENU nature

Fig. 2: Identifying the major cell types of mouse organogenesis.

From: The single-cell transcriptional landscape of mammalian organogenesis



BioTuring™

Solutions Resources

Explore 4,000,000 CELLS at ease with BioTuring BROWSER

EXPLORER NOW

A next-generation platform to re-analyze published single-cell sequencing data

Single Cell Analysis

5,500,000 cells will be indexed into BioTuring Single-cell Data Repository this September

by @biometers • August 30, 2019

f | t | in

Human Cell Atlas, single-cell data

We are glad to announce that we will upscale the current single-cell database in BioTuring Single-cell Browser to 5,500,000 cells this September. With this release, we will double the current number of publications indexed in BioTuring Single-cell Browser, and cross the number of cells hosted on available public single-cell data repositories like Human Cell Atlas (HCA) and Broad Institute's Single-cell Portal.

RECENT POSTS

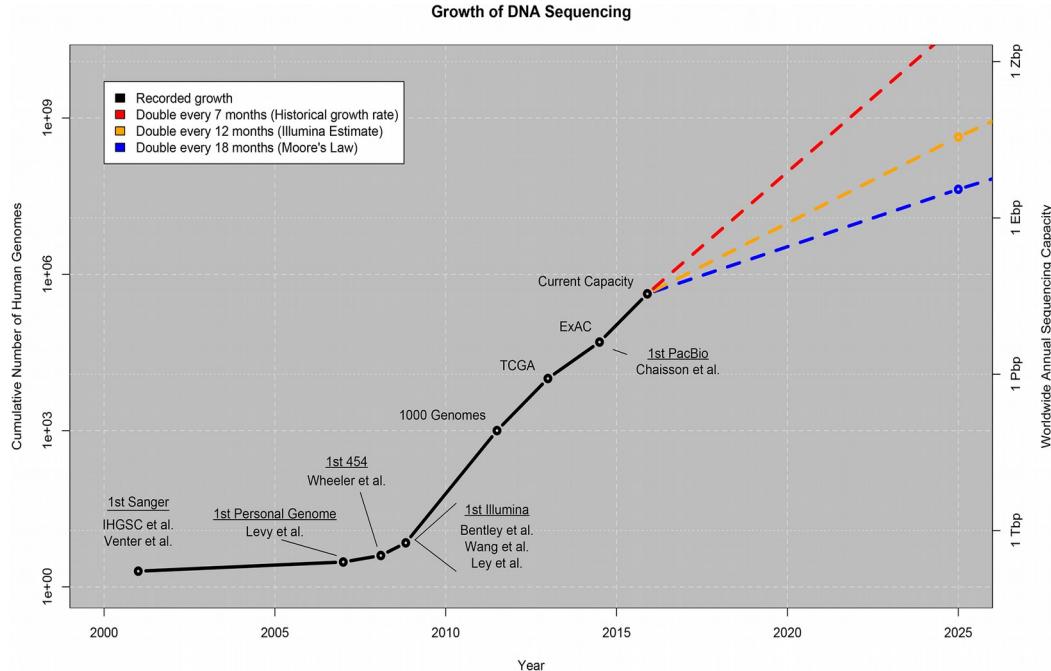
A new tool to interactively visualize single-cell objects (Seurat, Scanpy, SingleCellExperiments,...)

September 26, 2019

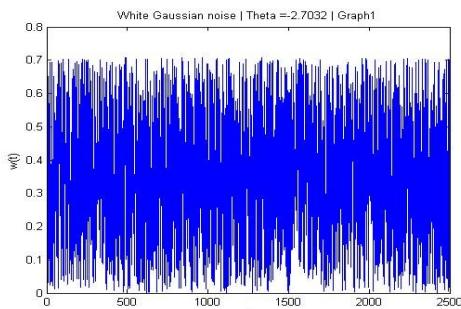
5,500,000 cells will be indexed into BioTuring Single-cell Data Repository this September

August 30, 2019

Big in Size or Sample Size?



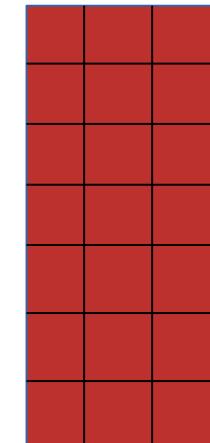
Stephens et al., (2015). Big Data: Astronomical or Genomical? PLoS Biology 13(7)



A file with **White Noise** can also take a lot of disk space

Genomics / WGS: Little Data

$$N_1 \sim 10^3$$

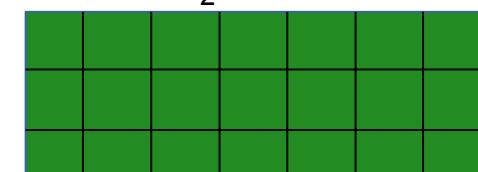


$$P_1 \sim 10^6$$

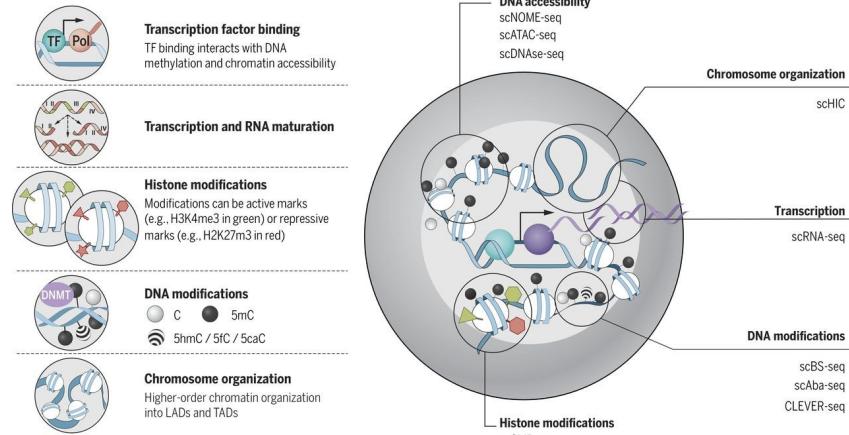
$$N_1 * P_1 = N_2 * P_2 = 10^9$$

scRNAseq: Big Data

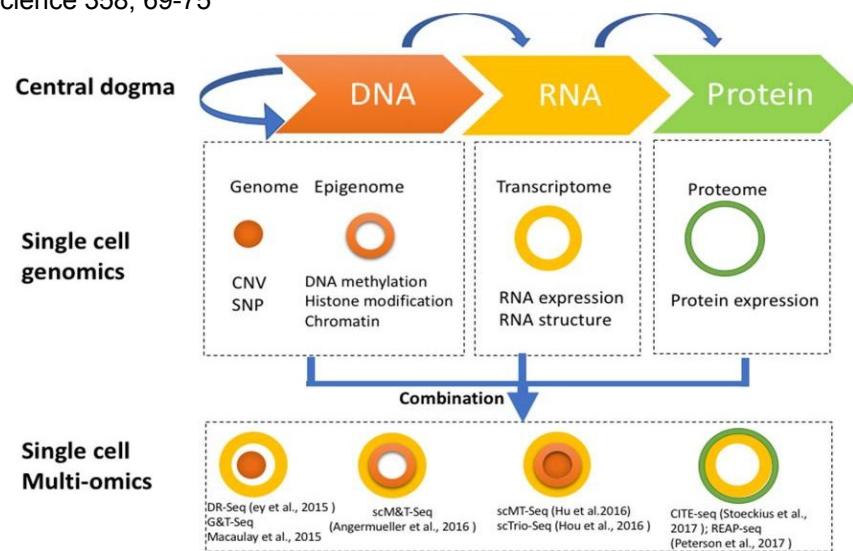
$$N_2 \sim 10^6$$



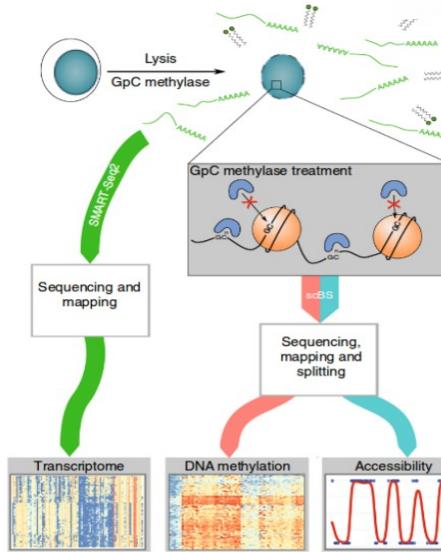
$$P_2 \sim 10^3$$



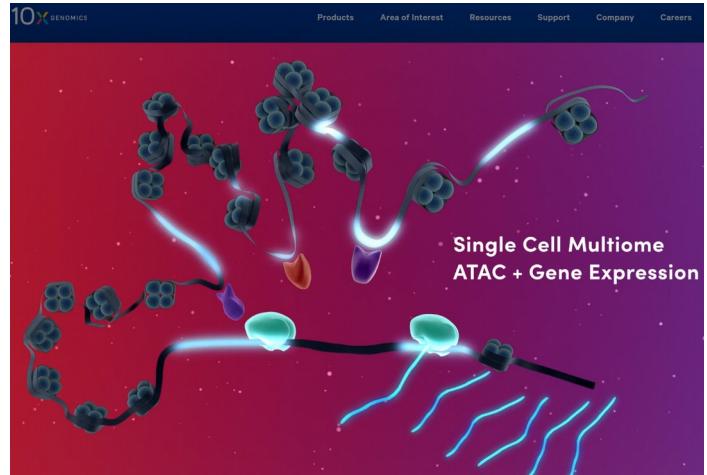
Kelsey et al., 2017, Science 358, 69-75



Hu et al., 2018, Frontier in Cell and Developmental Biology 6, 1-13



Clark et al., 2018, Nature Communications 9, 781

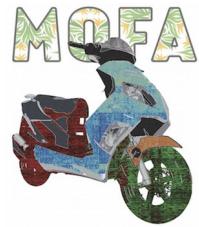
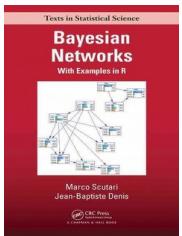


How to define and evaluate multi-Omics data integration?

OnPLS

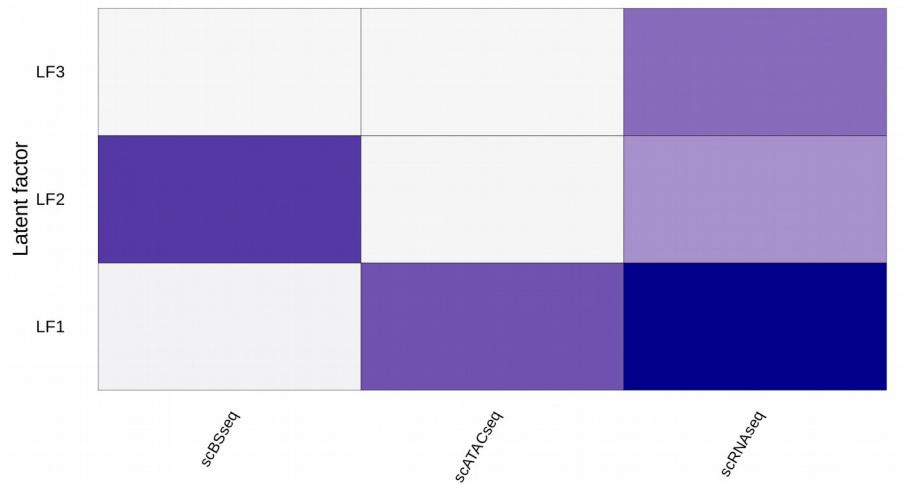
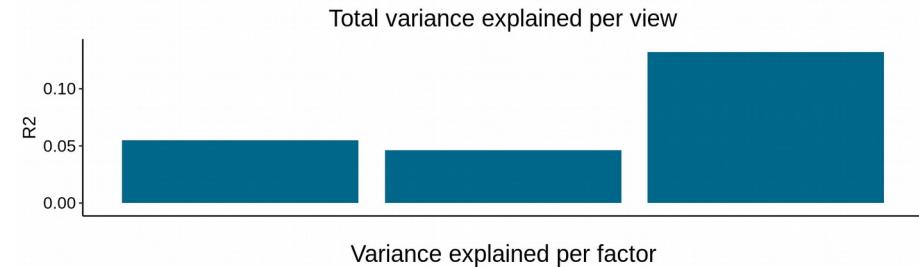
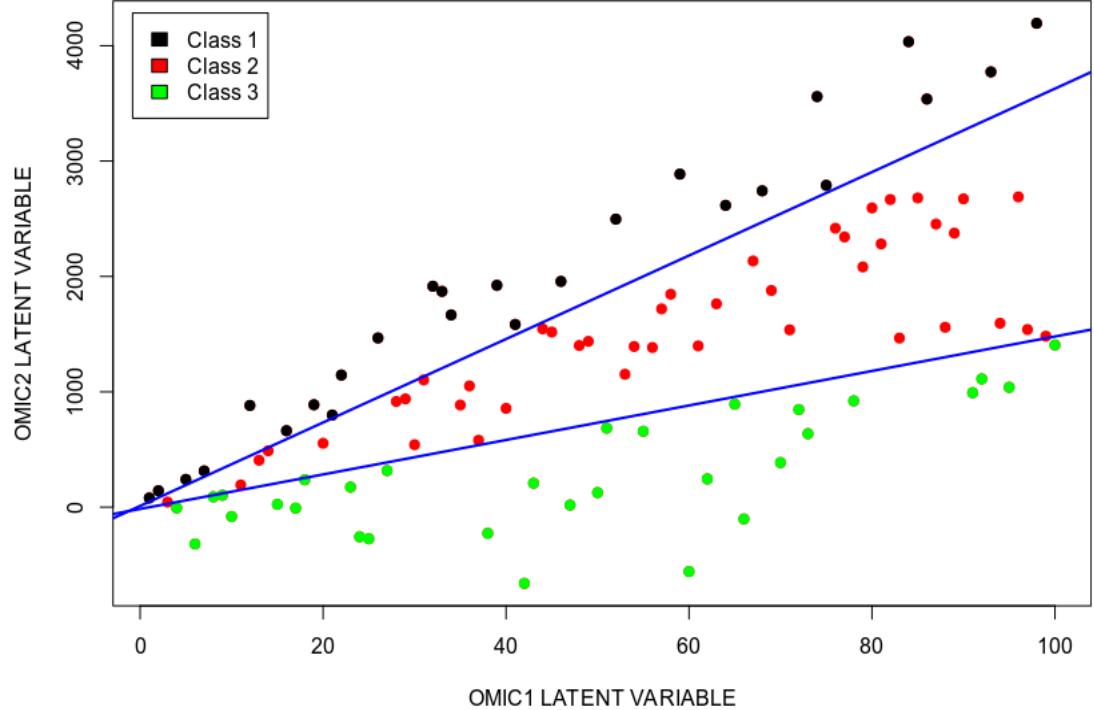
JIVE

DISCO



Clustering of Clusters

Idea Behind Omics Integration:
See Patterns Hidden in Individual Omics



How I Evaluate Omics Integration

TEXT (78%)

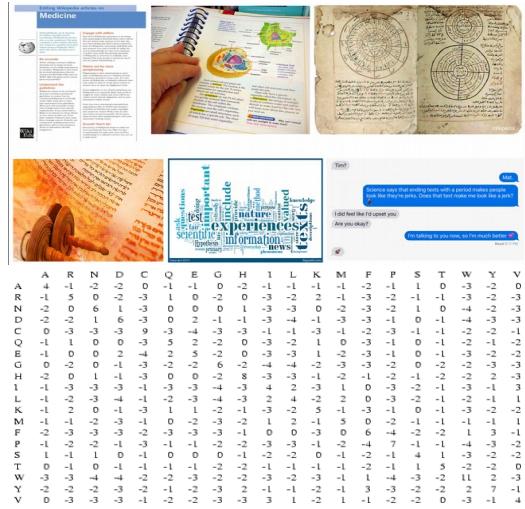
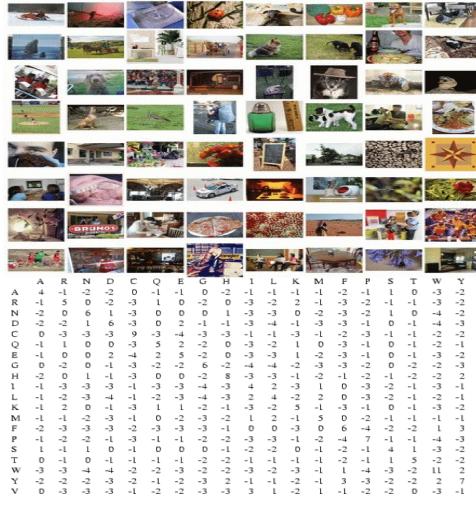
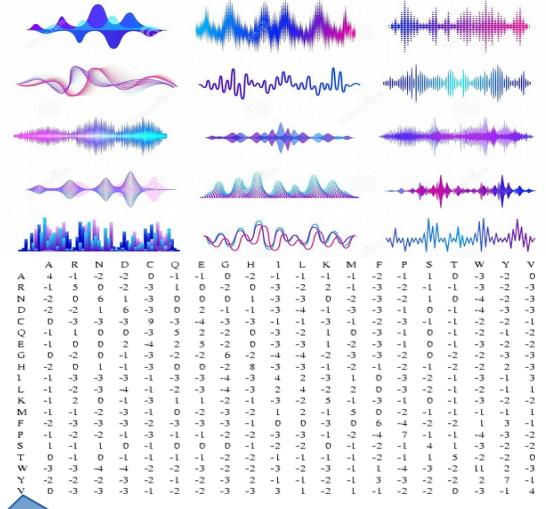


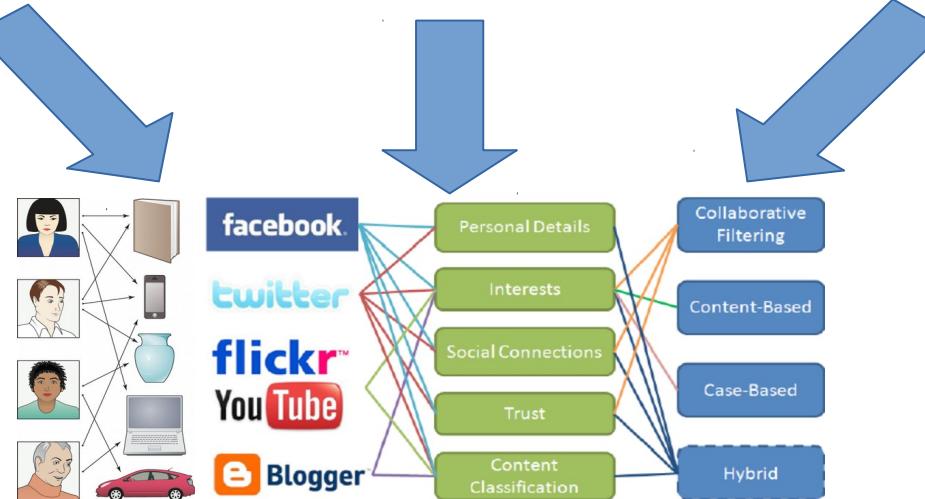
IMAGE (83%)



SOUND (75%)



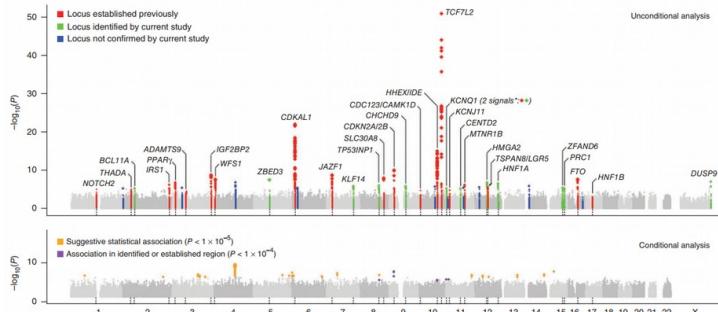
Predict Facebook user interests



Data Integration Accuracy: 96%

Prediction as a Criterion of Success

Statistics searches for candidates



Consequence

NEWS FEATURE PERSONAL GENOMES

NATURE/Vol 456/6 November 2008



The case of the missing heritability

B. Maher, Nature 456, 18-21 (2008)

Machine Learning optimizes prediction



Letter | Published: 31 July 2019

A clinically applicable approach to continuous prediction of future acute kidney injury

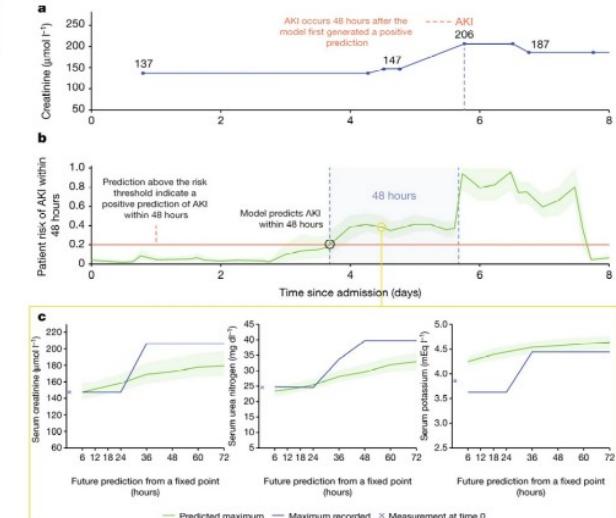
Nenad Tomalec, Kavir Glorot, [...] Shahril Mohamed

Nature 572, 116–119 (2019) | Download Citation |

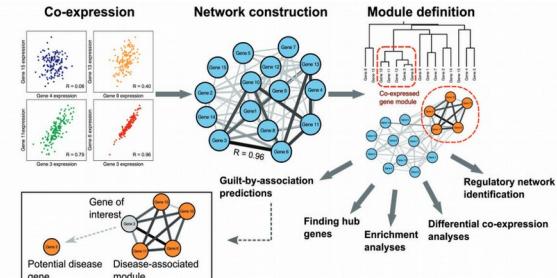
Abstract

The early prediction of deterioration could have an important role in supporting healthcare professionals, as an estimated 11% of deaths in hospital follow a failure to promptly recognize and treat deteriorating patients¹. To achieve this goal requires predictions of patient risk that are continuously updated and accurate, and delivered at an individual level with sufficient context and enough time to act. Here we develop a deep learning approach for the continuous risk prediction of future deterioration in patients, building on recent work that models adverse events from electronic health records^{2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17} and using acute kidney injury – a common and potentially life-threatening condition¹⁸ – as an exemplar. Our model was developed on a large, longitudinal dataset of electronic health records that cover diverse

From: A clinically applicable approach to continuous prediction of future acute kidney injury

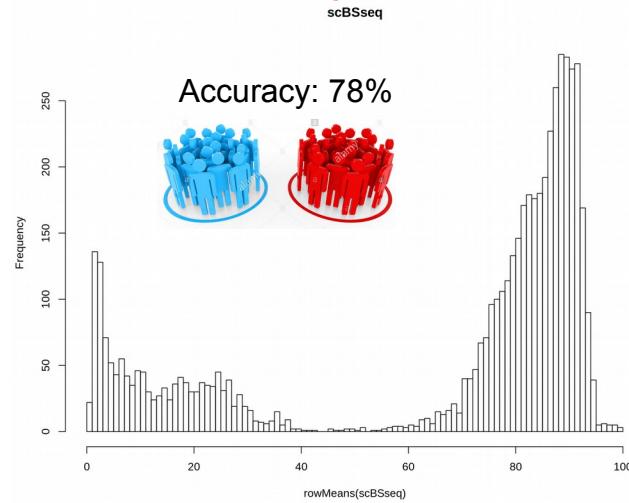


Consequence

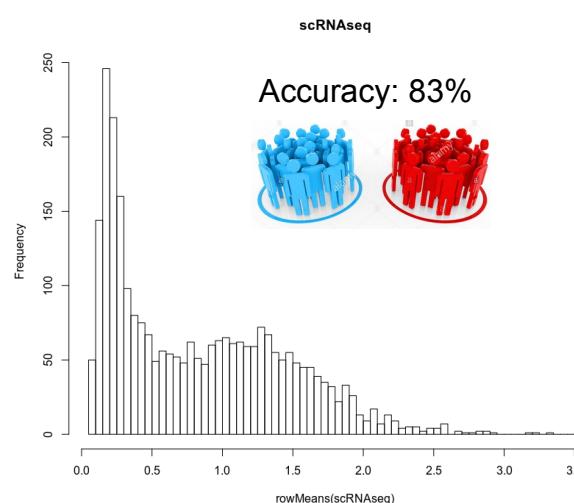


How I Evaluate Omics Integration

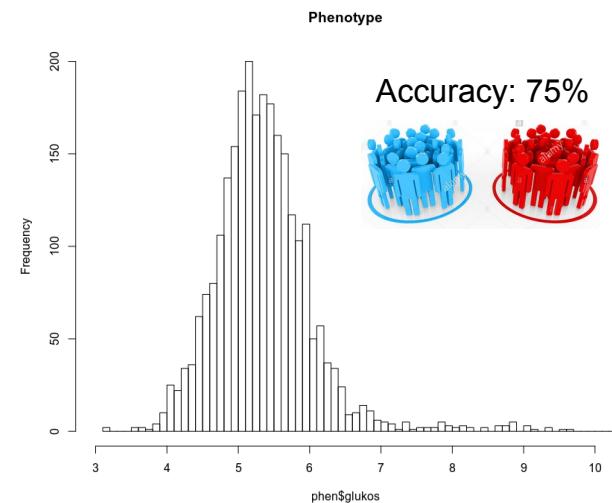
Methylation



Gene Expression



Clinical variable



1) Convert to common space:
Neural Networks, SNF, UMAP

2) Explicitly model distributions:
MOFA, Bayesian Networks

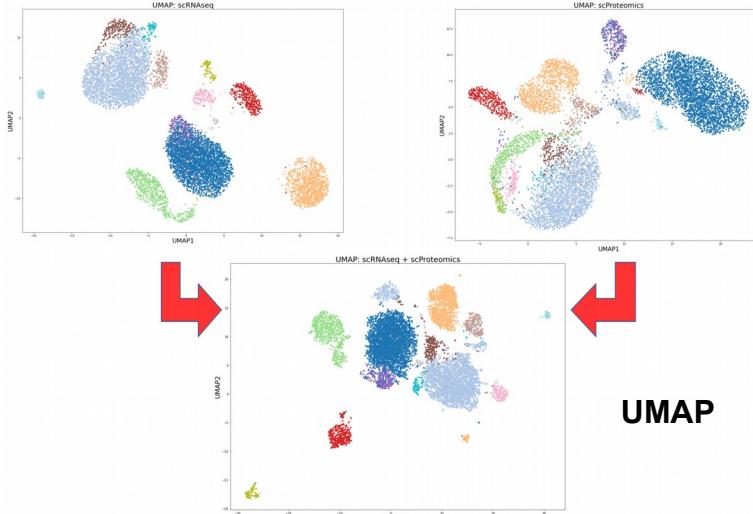
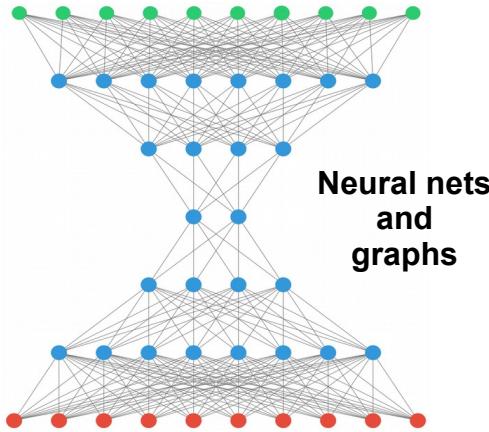
3) Extract common variation:
PLS, CCA, Factor Analysis



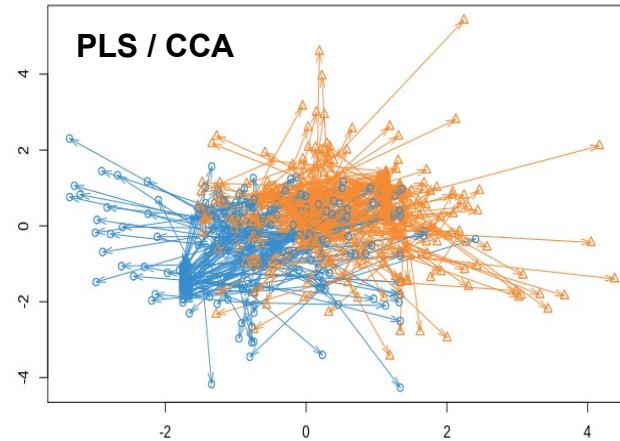
**Data Integration
Accuracy: 96%**

Overview of Integrative Methods

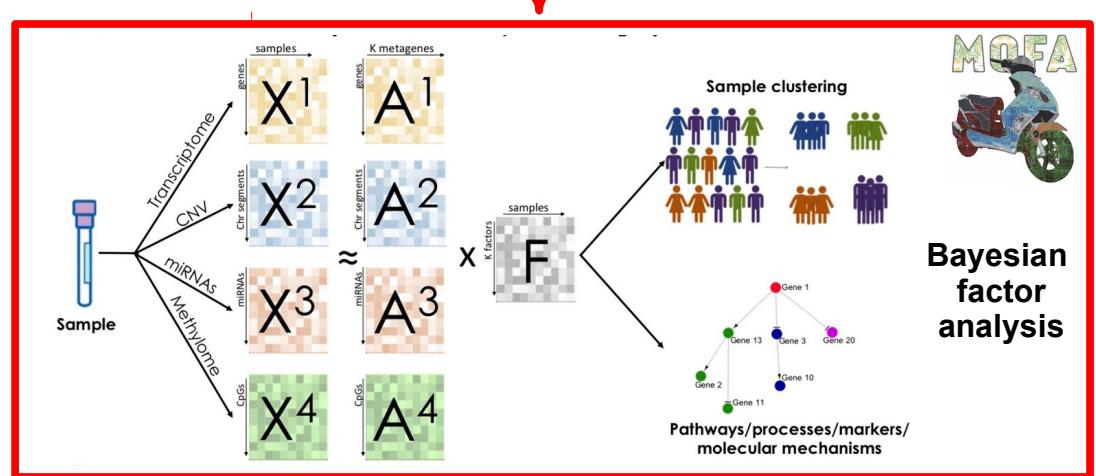
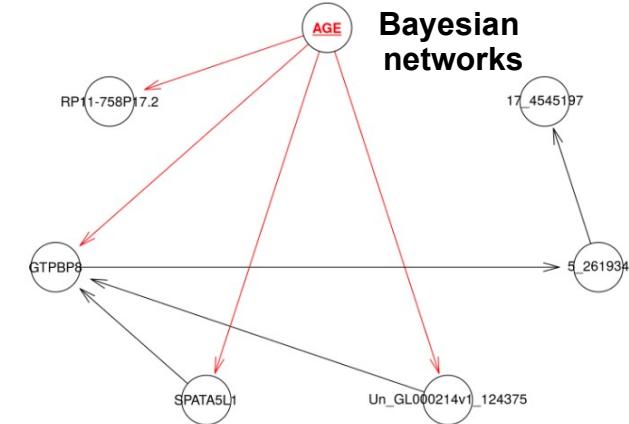
Convert to common space



Extract common variation



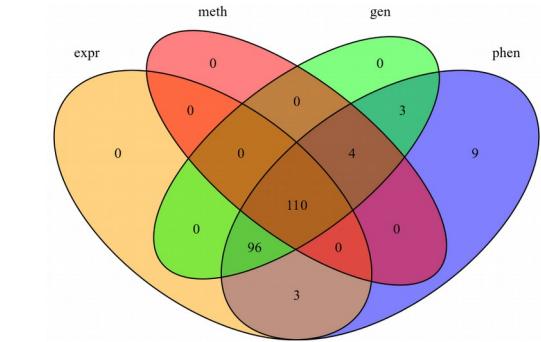
Combine via Bayes rule



	Linear	Non-Linear
Supervised	PLS / OPLS / mixOmics, LASSO / Ridge / Elastic Net	Neural Networks, Random Forest, Bayesian Networks
Unsupervised	Factor Analysis / MOFA	Autoencoder, SNF, UMAP, Clustering of Clusters

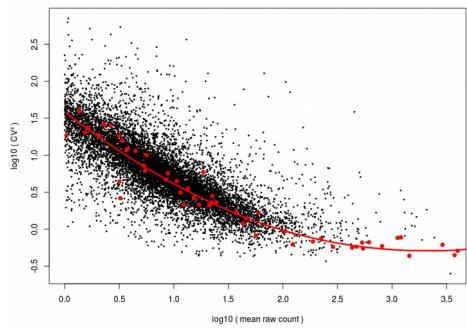
For Example:

- 1) With ~100 samples it is a good idea to do **linear** Omics integration
- 2) T2D is a phenotype of interest, therefore **supervised** integration



Data Set (4 Omics)
110 overlapping individuals

Unsupervised:
remove low-variance



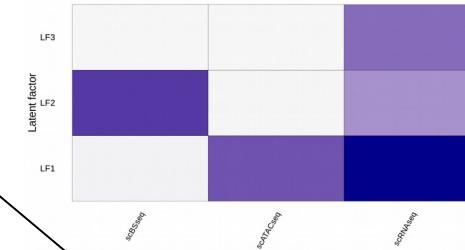
Train Set (n = 80)

Supervised:
LASSO



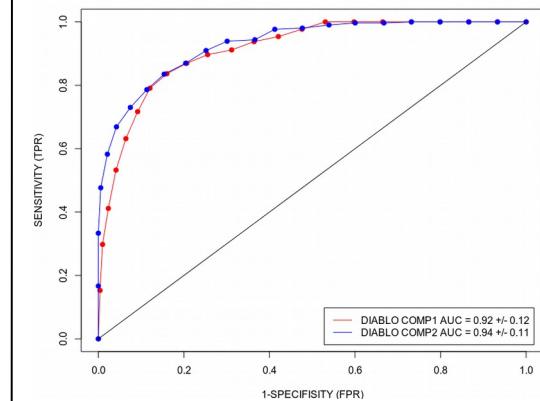
Feature Selection

Check covariance



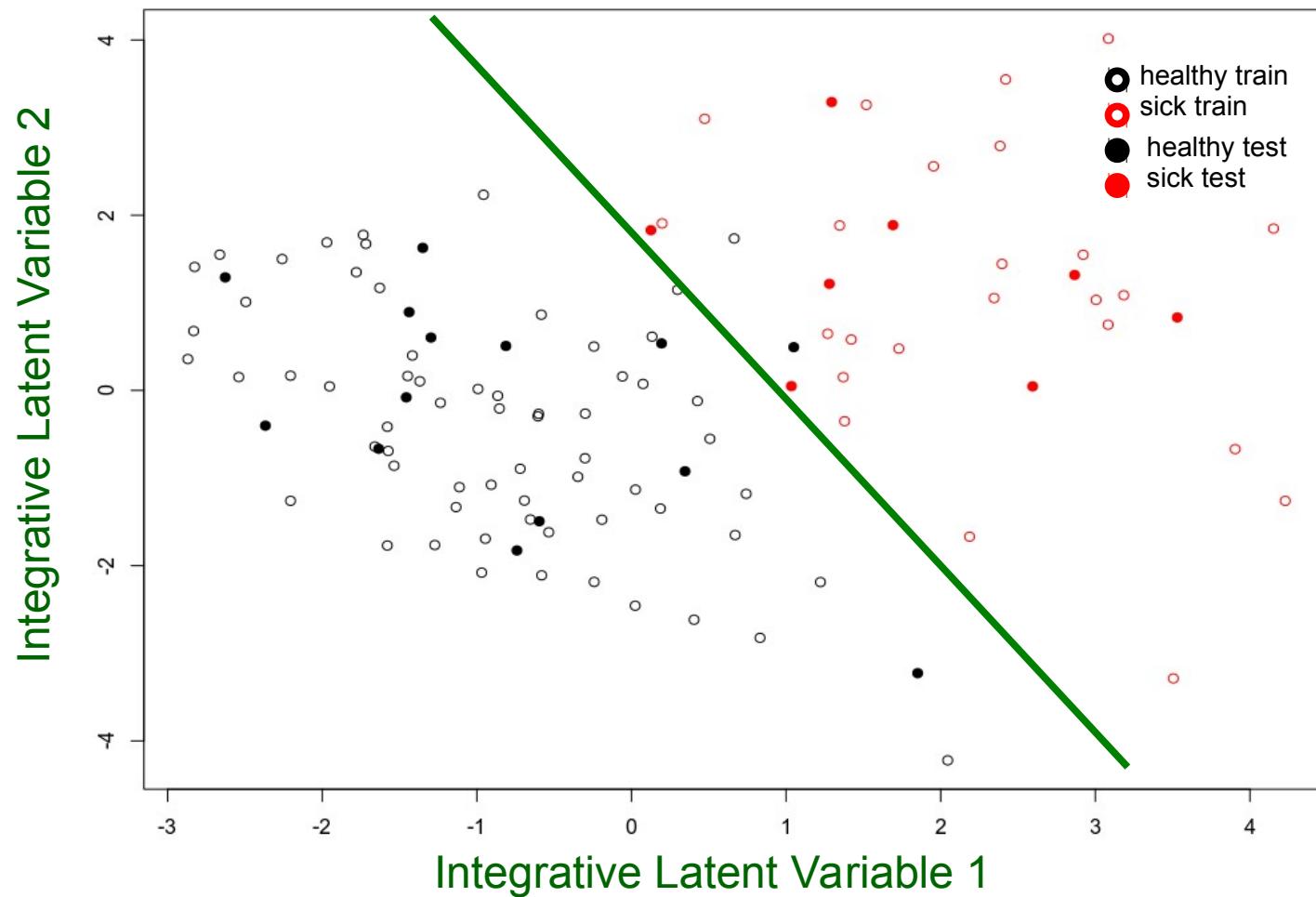
Test Set (n = 30)

Evaluation



Omics Integration

Trained Model



- 1) Biological data are **high-dimensional** and are notoriously difficult to analyze
- 2) Integration across Omics is very sensitive to the **Curse of Dimensionality**
- 3) Integrating across Omics we expect to discover **novel patterns** in the data
- 4) Increased **prediction accuracy** is an indication of successful data integration
- 5) **Single cell** Omics have enough statistical power for integrative analysis



*Knut och Alice
Wallenbergs
Stiftelse*



**LUNDS
UNIVERSITET**