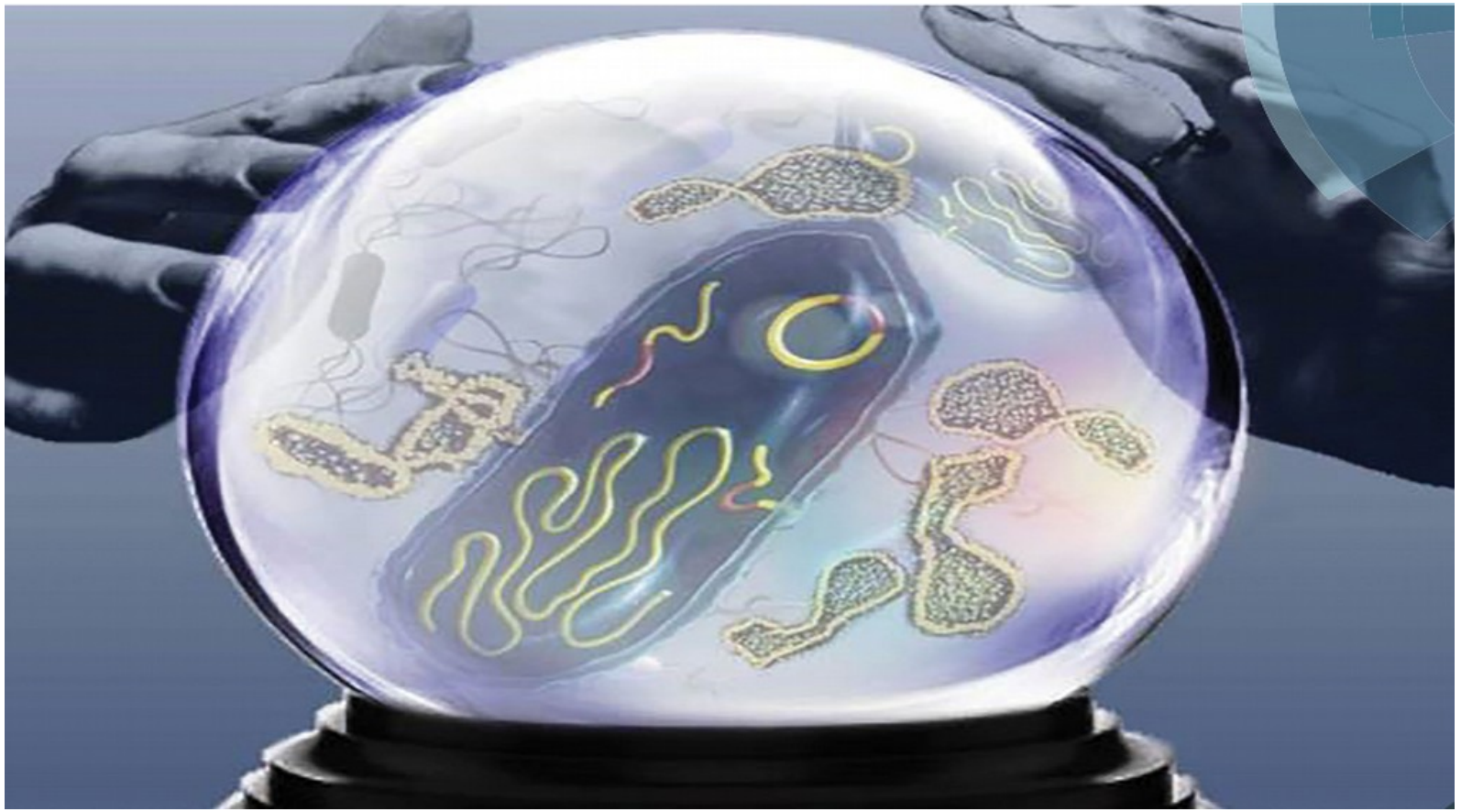
Machine Learning View of OMICs Integration

OMICs Integration and Systems Biology course

Nikolay Oskolkov, NBIS SciLifeLab

Lund, 5.10.2020



There are many ways to do same thing

Peculiarities of OMICs Integration

Peculiarity #1: Find Something in My Data

Fishing expedition

- I do not understand your biological hypothesis
- I do not have any
- Then I reject your manuscript

BMC Part of Springer Nature

Genome Biology

Home About Articles Submission Guidelines

Editorial | Open Access | Published: 03 September 2020

A hypothesis is a liability

Itai Yanai & Martin Lercher

Genome Biology **21**, Article number: 231 (2020) | Cite this article

12k Accesses | **619** Altmetric | Metrics

> " 'When someone seeks,' said Siddhartha, 'then it easily happens that his eyes see only the thing that he seeks, and he is able to find nothing, to take in nothing. [...] Seeking means: having a goal. But finding means: being free, being open, having no goal.' " Hermann Hesse

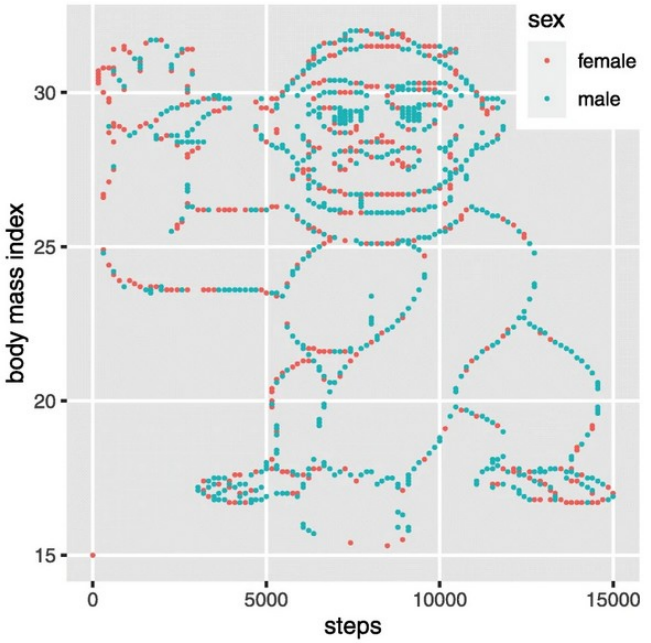
There is a hidden cost to having a hypothesis. It arises from the relationship between night science and day science, the two very distinct modes of activity in which scientific ideas are generated and tested, respectively [1, 2]. With a hypothesis in hand, the impressive strengths of day science are unleashed, guiding us in designing tests, estimating parameters, and throwing out the hypothesis if it fails the tests. But when we analyze the results of an experiment, our mental focus on a specific hypothesis can prevent us from exploring other aspects of the data, effectively blinding us to new ideas. A hypothesis then becomes a liability for any night science explorations. The corresponding limitations on our creativity, self-imposed in hypothesis-driven research, are of particular concern in the context of modern biological datasets, which are often vast and likely to contain hints at multiple distinct and potentially exciting discoveries. Night science has its own liability though, generating many spurious relationships and false hypotheses. Fortunately, these are exposed by the light of day science, emphasizing the complementarity of the two modes, where each overcomes the

a

Lister - [c:\Users\YAN...

ID	steps	bmi
1	15000	16.9
2	15000	16.9
6	14861	16.8
7	14861	16.8
8	14699	17.3
10	14560	20.5
11	14560	20.6
13	14560	20.5
17	14560	20.4
18	14560	20.4
19	14560	19.8
20	14560	19.7
22	14560	19.7
24	14560	19.6
25	14560	19.6
27	14560	19.6
29	14560	17.4
30	14560	17.4
32	14398	20.9
37	14398	17.5
40	14398	17.1
42	14259	21.1
43	14259	21.1

b



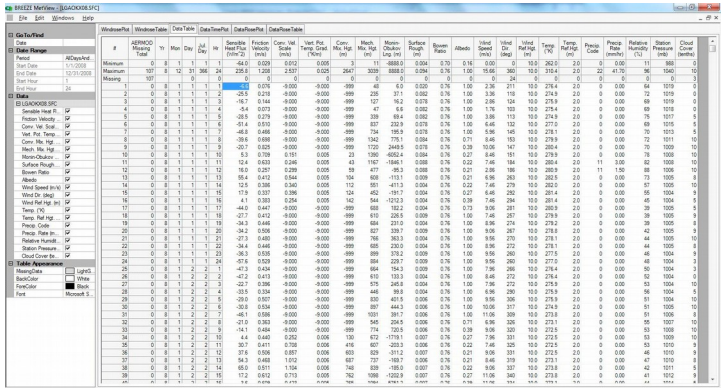
c

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

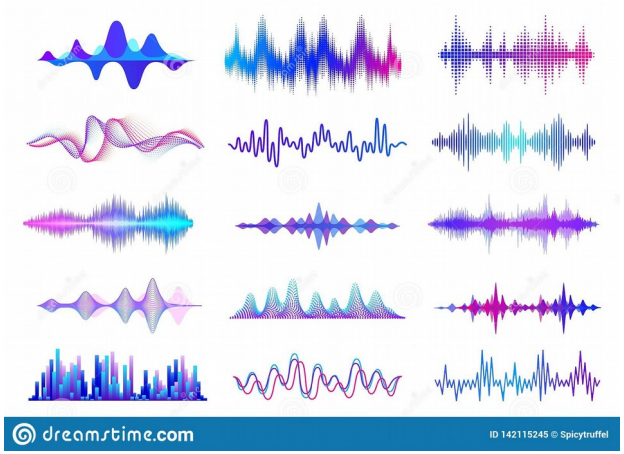
a An artificial dataset given to students with and without explicit hypotheses on the relationship between BMI and the steps taken on a particular day, for men and women. **b** A plot of the dataset. **c** The contingency table for students in the two groups ("hypothesis-focused," "hypothesis-free") that discovered the gorilla or not [6]

Peculiarity #2: Various Data Distributions

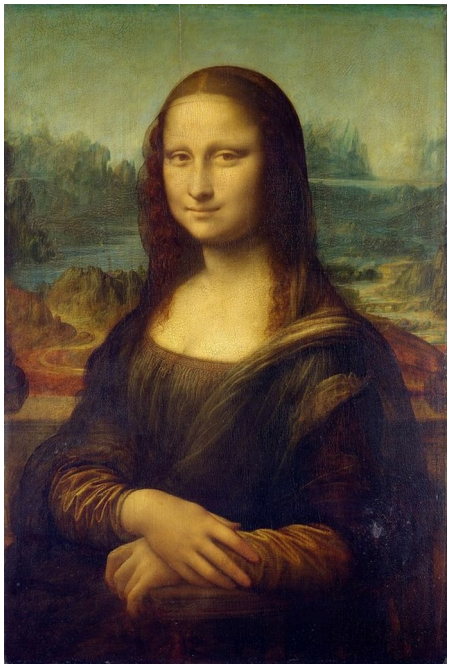
Tabular



Sound



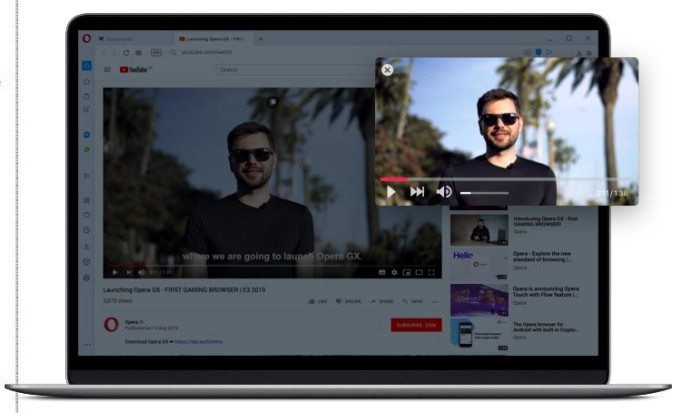
Image



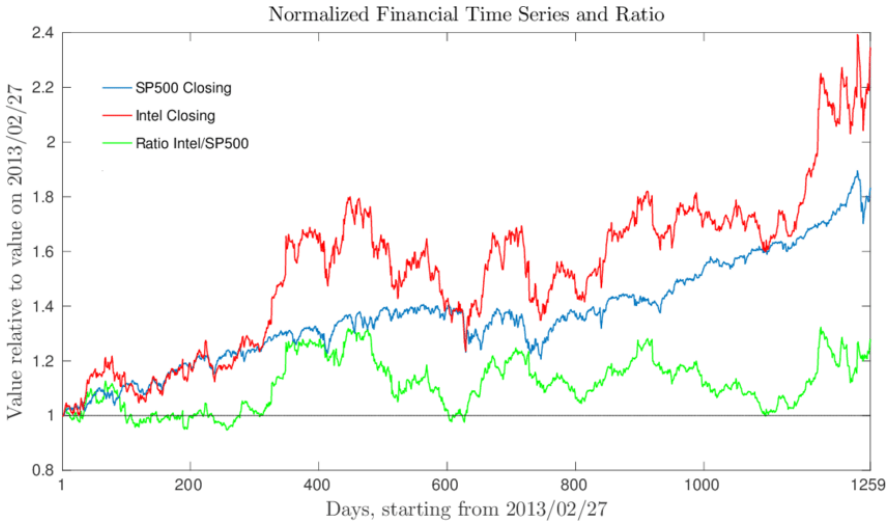
Text

DATA

Video



Time Series



Peculiarity #3: High-Dimensional Data

OMIC1 (N × P_1)

$$N \quad \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

OMIC2 (N × P_2)

$$N \quad \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

OMIC3 (N × P_3)

$$N \quad \begin{pmatrix} 0 & 3 & 1 & 0 & 2 & 3 & 8 & 1 & 1 & 3 \\ 1 & 1 & 0 & 0 & 7 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 0 & 6 & 7 & 1 & 2 & 2 \\ 1 & 2 & 3 & 10 & 0 & 4 & 6 & 1 & 0 & 5 \\ 3 & 2 & 2 & 1 & 4 & 3 & 2 & 1 & 6 & 0 \\ 7 & 4 & 4 & 5 & 3 & 9 & 6 & 1 & 6 & 1 \\ 7 & 1 & 1 & 5 & 2 & 8 & 9 & 1 & 3 & 6 \\ 5 & 0 & 1 & 6 & 2 & 0 & 0 & 0 & 1 & 5 \\ 1 & 6 & 3 & 3 & 4 & 6 & 2 & 0 & 1 & 1 \\ 1 & 2 & 2 & 4 & 1 & 1 & 3 & 0 & 8 & 2 \end{pmatrix}$$

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

Metabolomics
$N \approx P$

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

Proteomics
$N \approx P$

← **manageable**

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

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

impossible ↓

$$\setcounter{MaxMatrixCols}{64}\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Genomics
$N <<<< P$

$$\setcounter{MaxMatrixCols}{64}\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Methylomics
$N <<<< P$

The Curse of Dimensionality complicates OMICs Integration

The Curse of Dimensionality

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

Biomedicine

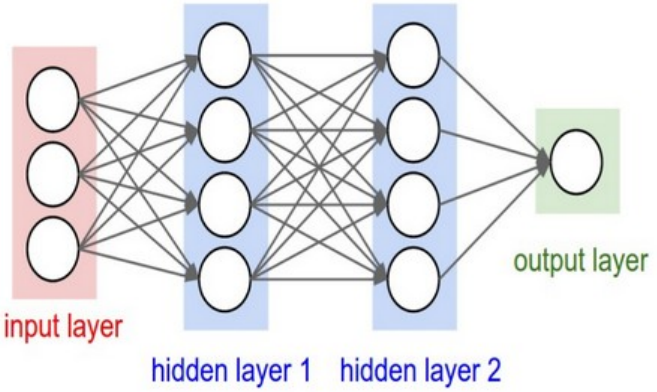
Bayesianism

P >> N

Frequentism

P ~ N

Deep Learning



P << N

Amount of Data

Ex.1

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

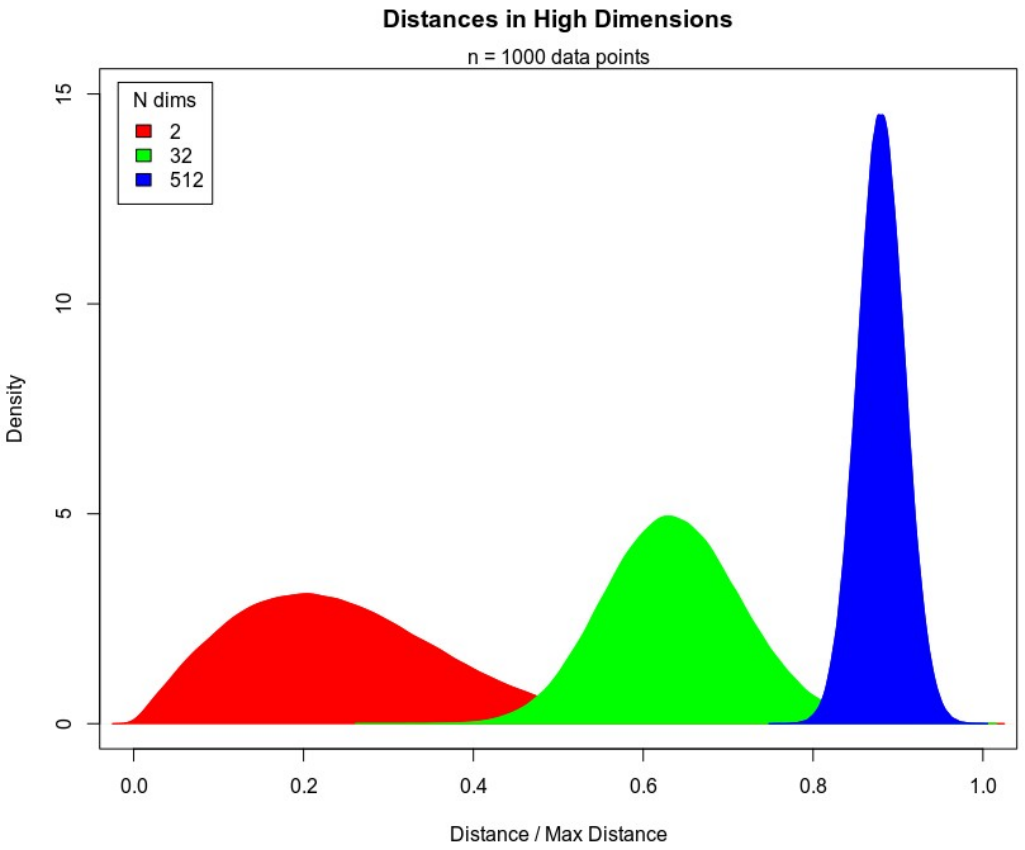
$$\beta = (X^T X)^{-1} X^T Y$$

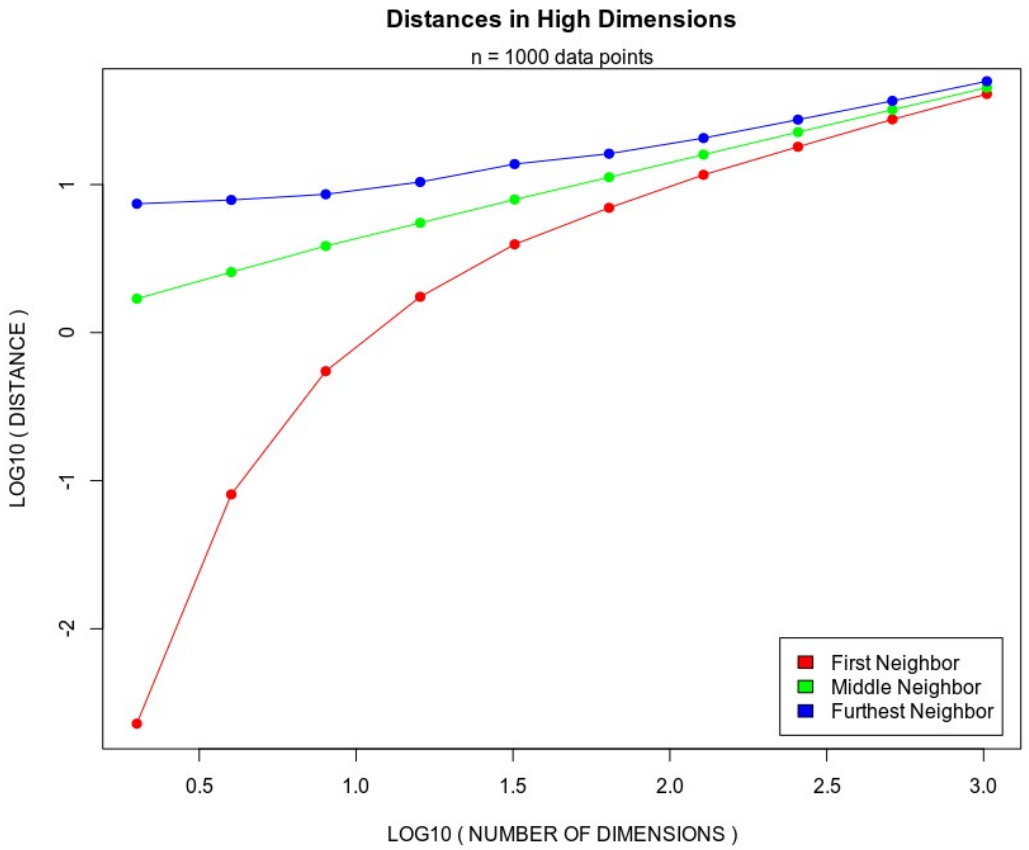
$$(X^T X)^{-1} \sim \frac{1}{\det(X^T X)} \cdots \to \infty, \quad n << p$$

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

Biased ML variance estimator in HD-space

Equidistant Points in High Dimensions





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

The differences between closest and furthest data point neighbours disappears in high-dimensional spaces – can't cluster

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

Low Dimensional Space

```r
set.seed(123); n <- 20; p <- 2
Y <- rnorm(n); X <- matrix(rnorm(n*p),n,p); summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0522 -0.6380  0.1451  0.3911  1.8829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14950    0.22949   0.651    0.523
## X1          -0.09405    0.28245  -0.333    0.743
## X2          -0.11919    0.24486  -0.487    0.633
##
## Residual standard error: 1.017 on 17 degrees of freedom
## Multiple R-squared:  0.02204,    Adjusted R-squared:  -0.09301
## F-statistic: 0.1916 on 2 and 17 DF,  p-value: 0.8274
```

Going to Higher Dimensions

```
set.seed(123456); n <- 20; p <- 10
Y <- rnorm(n); X <- matrix(rnorm(n*p),n,p); summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0255 -0.4320  0.1056  0.4493  1.0617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.54916    0.26472   2.075   0.0679 .
## X1           0.30013    0.21690   1.384   0.1998
## X2           0.68053    0.27693   2.457   0.0363 *
## X3          -0.10675    0.26010  -0.410   0.6911
## X4          -0.21367    0.33690  -0.634   0.5417
## X5          -0.19123    0.31881  -0.600   0.5634
## X6           0.81074    0.25221   3.214   0.0106 *
## X7           0.09634    0.24143   0.399   0.6992
## X8          -0.29864    0.19004  -1.571   0.1505
## X9          -0.78175    0.35408  -2.208   0.0546 .
## X10          0.83736    0.36936   2.267   0.0496 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8692 on 9 degrees of freedom
## Multiple R-squared:  0.6592,    Adjusted R-squared:  0.2805
```

Even Higher Dimensions

```
set.seed(123456); n <- 20; p <- 20
Y <- rnorm(n); X <- matrix(rnorm(n*p),n,p); summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
## ALL 20 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.34889         NA      NA       NA
## X1           0.66218         NA      NA       NA
## X2           0.76212         NA      NA       NA
## X3          -1.35033         NA      NA       NA
## X4          -0.57487         NA      NA       NA
## X5           0.02142         NA      NA       NA
## X6           0.40290         NA      NA       NA
## X7           0.03313         NA      NA       NA
## X8          -0.31983         NA      NA       NA
## X9          -0.92833         NA      NA       NA
## X10          0.18091         NA      NA       NA
## X11         -1.37618         NA      NA       NA
## X12          2.11438         NA      NA       NA
## X13         -1.75103         NA      NA       NA
## X14         -1.55073         NA      NA       NA
## X15          0.01112         NA      NA       NA
## X16         -0.50943         NA      NA       NA
```

How to define and evaluate OMICs Integration?

What I Mean by OMICs Integration

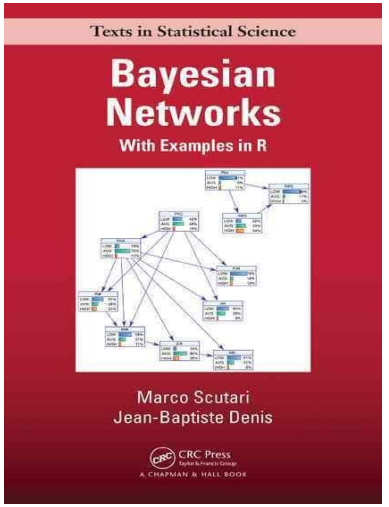
mix Omics

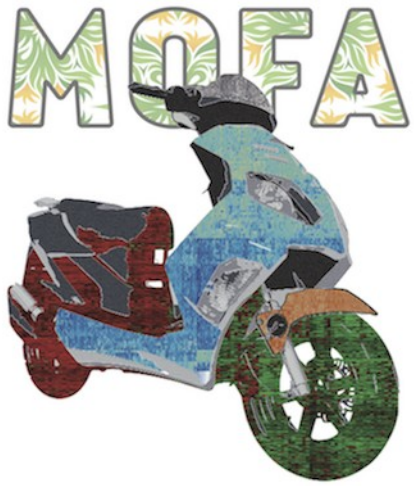
Exploration and Integration of Omics datasets

OnPLS

JIVE

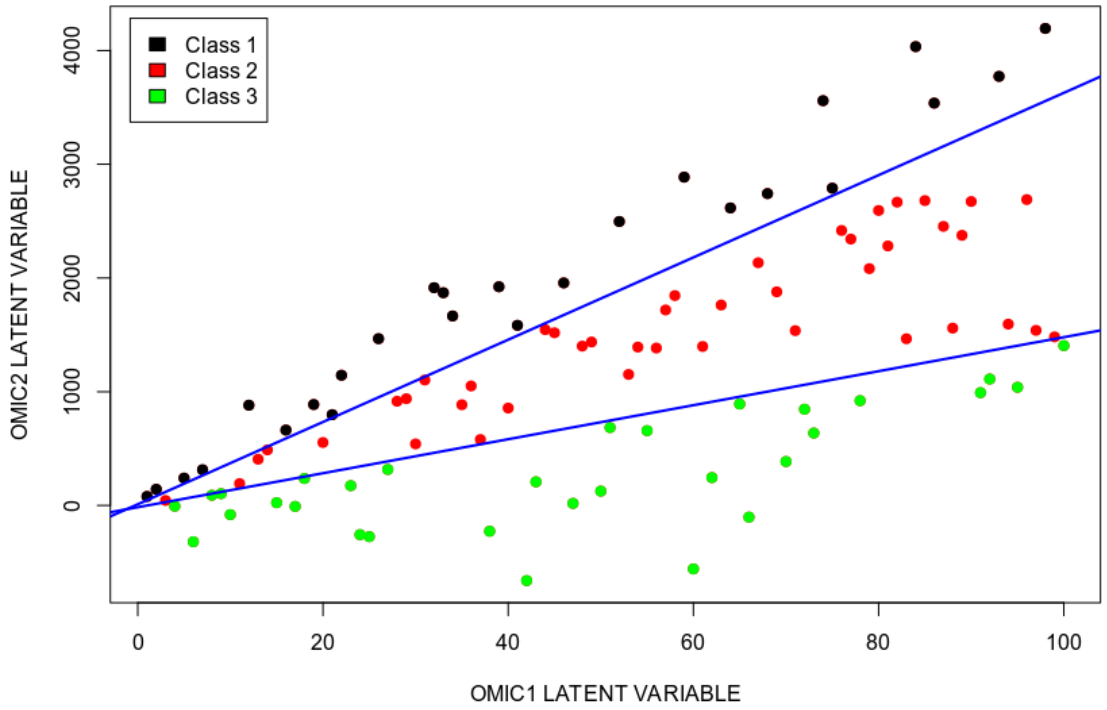
DISCO

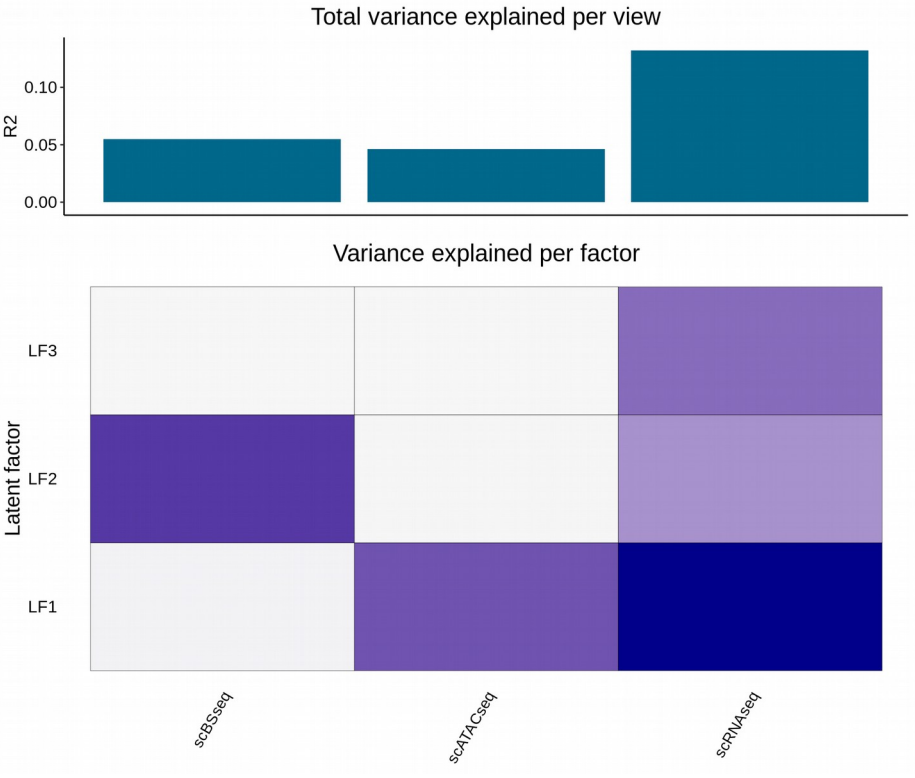




Clustering of Clusters

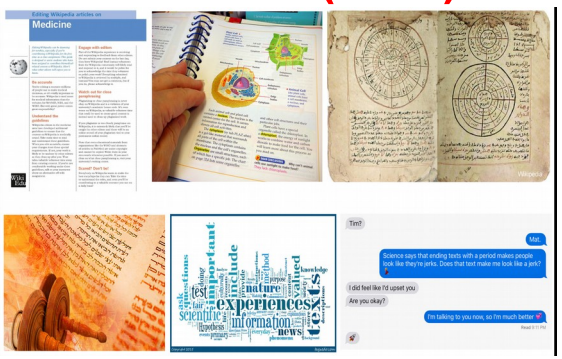
Idea Behind OMICs Integration: See Patterns Hidden in Individual OMICS





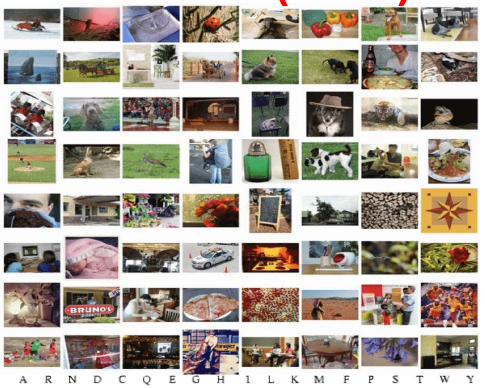
How I Evaluate OMICs Integration, Data Science: Boost in Prediction

TEXT (78%)



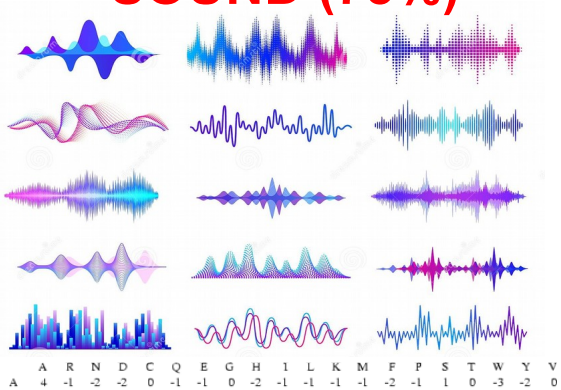
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

IMAGE (83%)



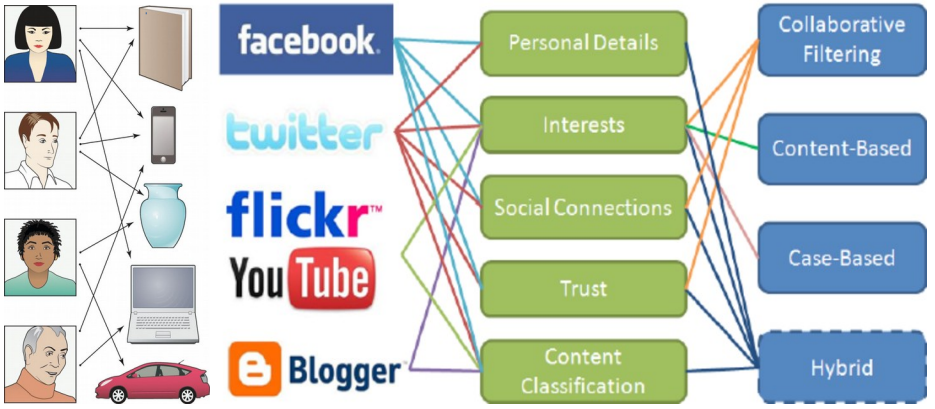
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

SOUND (75%)



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

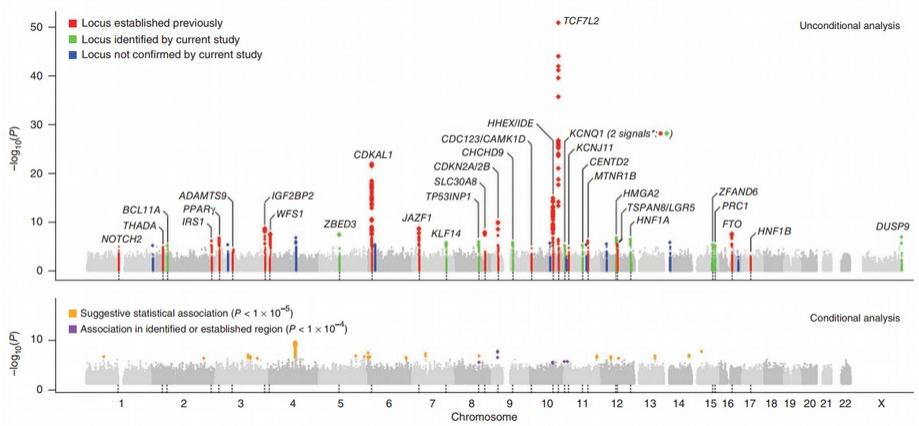
Predict Facebook user interests



Data Integration Accuracy: 96%

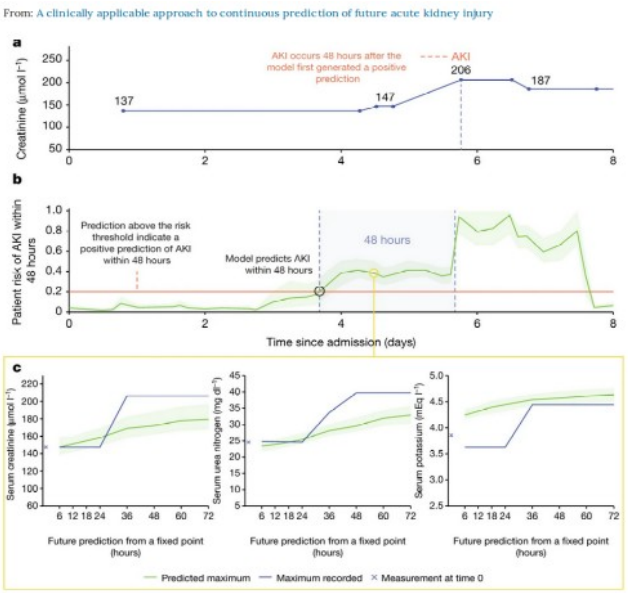
Prediction is an Ultimate Criterion of Successful OMICS Integration

Statistics searches for candidates

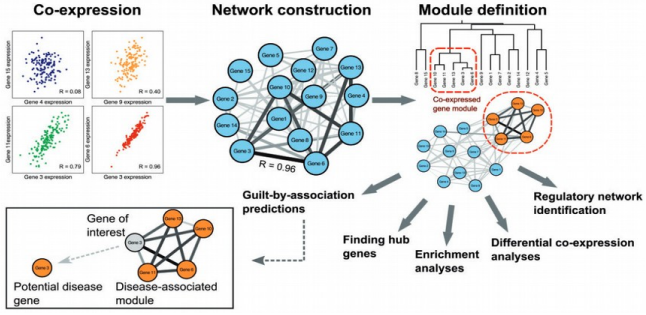


Consequence

Machine Learning optimizes prediction

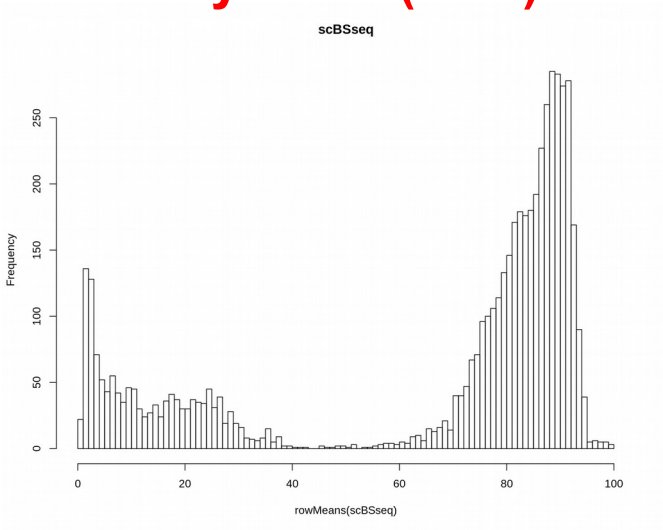


Consequence

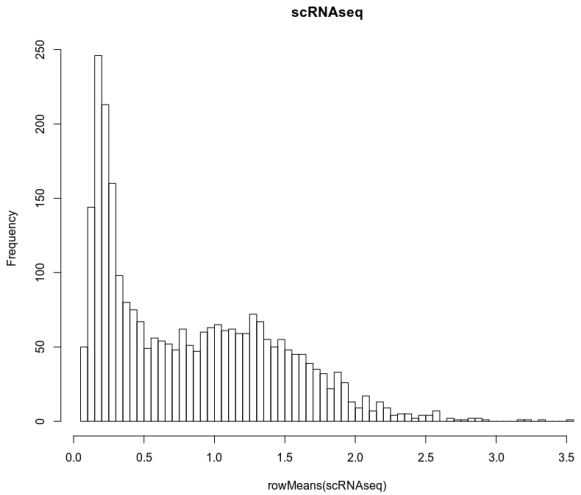


How I Evaluate OMICs Integration, Life Sciences: Boost in Prediction

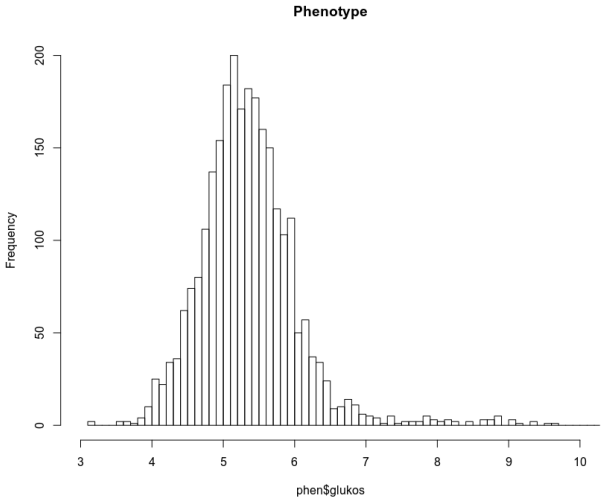
Methylation (78%)



Gene Expression (83%)



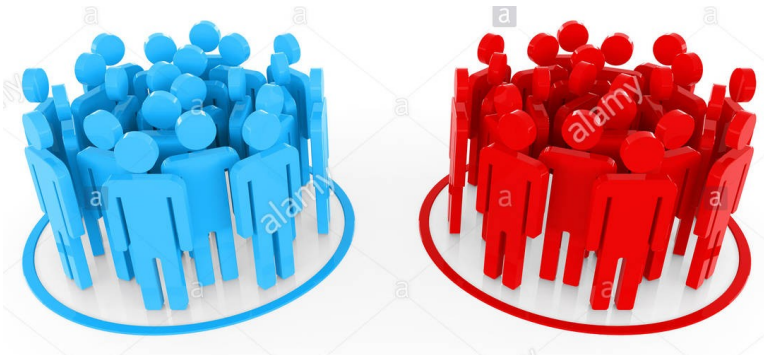
Phenotype (75%)



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

2) Explicitly model distributions:
MOFA, Bayesian Networks

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



HEALTHY SICK

Data Integration Accuracy: 96%

Protocol of OMICs Integration

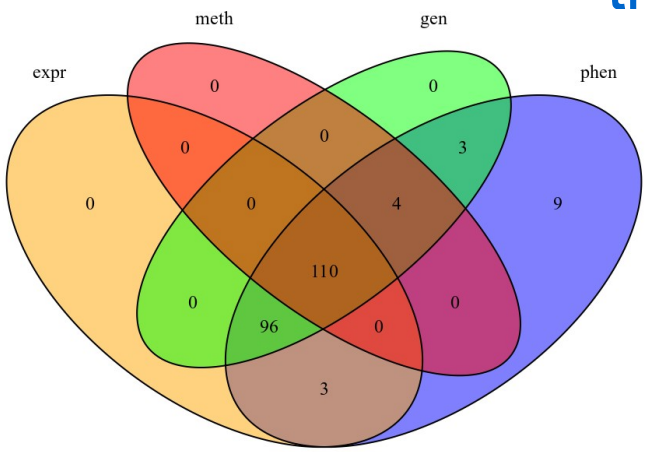
What Method of Integration to Select?

	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 ~110 samples it is a good idea to do **linear** OMICs integration
2) T2D is a phenotype of interest, therefore **supervised** integration

Feature Pre-Selection to Overcome the Curse of Dimensionality



29% of Diabetics

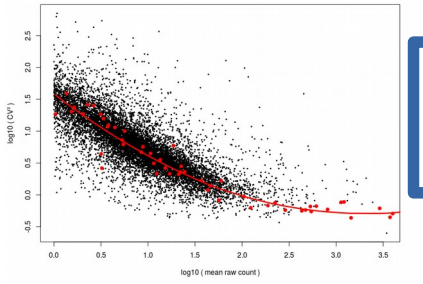
Data Set (4 OMICs)

~~WGS (~30 mln dims)~~
~~BSseq (~30 mln dims)~~

Train Set (n = 88)

Test Set (n = 22)

unsupervised



supervised

Feature Pre-Selection

Evaluation

OMICs Integration

Trained Model

Protocol of Integrative OMICS Analysis

1) Check that there is a relation between the OMICs (MOFA)

2) Choose integrative model based on amount of data and goal (linear, supervised)

3) Do feature pre-selection (supervised or unsupervised) on train data set

4) Integrate the OMICs using your favorite model chosen in 2) on train data set

5) Compare prediction of integrative model with predictions from individual OMICs

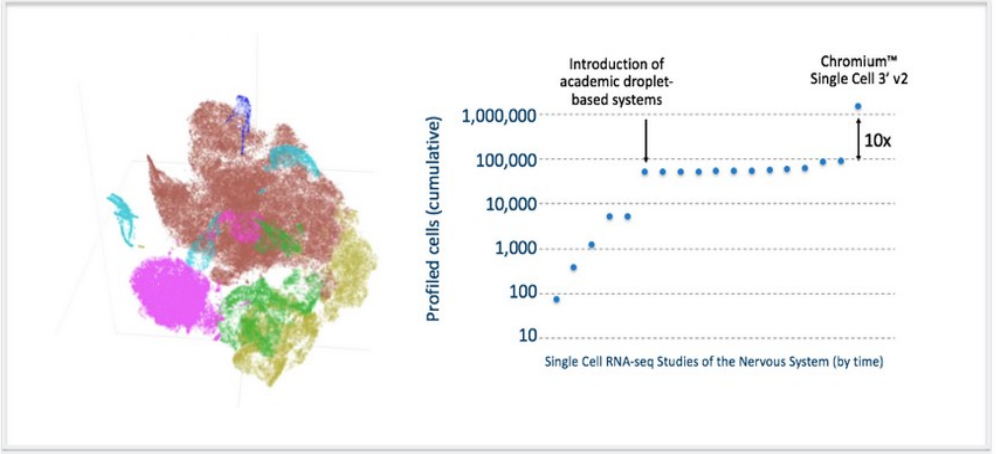
OMICs Integration in Single Cell

Big Data in Single Cell

CAREERS BLOG 10X UNIVERSITY

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

« Back to Blog ‹ Newer Article Older Article ›



Our 1.3 million single cell dataset is ready to download

0 KUDOS

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

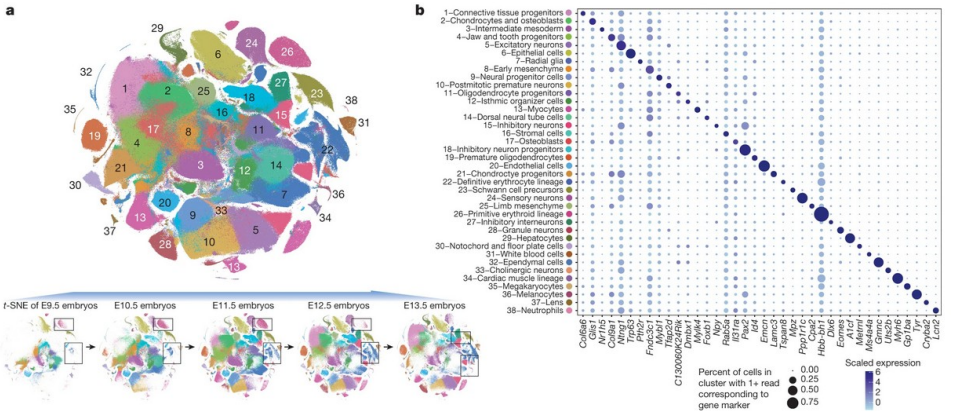
At ASHG last year, we announced our 1.3 Million Brain Cell Dataset, which is, to date, the largest dataset published in the single cell RNA-sequencing (scRNA-seq) field. Using the Chromium™ Single Cell 3' Solution (v2 Chemistry), we were able to sequence and profile 1,308,421 individual cells from embryonic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

Watch out Underfitting!
Paradise for Deep Learning!

MENU nature

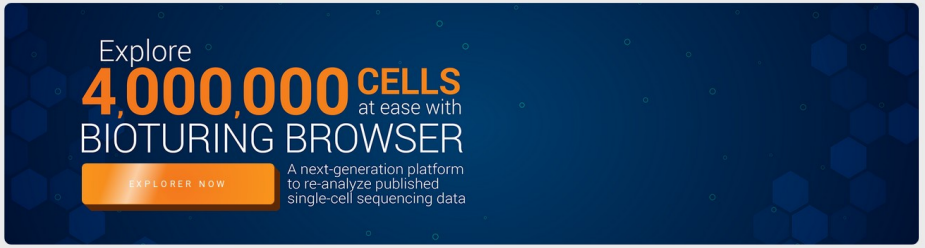
Fig. 2: Identifying the major cell types of mouse organogenesis.

From: The single-cell transcriptional landscape of mammalian organogenesis



a, t-SNE visualization of 2,026,641 mouse embryo cells (after removing a putative doublet cluster), coloured by cluster identity (ID) from Louvain clustering (in **b**), and annotated on the basis of marker genes. The same t-SNE is plotted below, showing only cells from each stage (cell numbers from left to right: n = 151,000 for E9.5; 370,279 for E10.5; 602,784 for E11.5; 468,088 for E12.5; 434,490 for E13.5). Primitive erythroid (transient) and definitive erythroid (expanding) clusters are boxed. **b**, Dot plot showing expression of one selected marker gene per cell type. The size of the dot encodes the percentage of cells within a cell type in

BioTuring Solutions Resources



Single Cell Analysis

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

by biomembers · August 30, 2019

Human Cell Atlas, single-cell data

We are glad to announce that we will upsize 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.

Search

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

National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

LUNDS UNIVERSITET