Regression and classification

Let X be a p-dimensional predictor variable and Y the target variable of interest. Assume a linear model in which

$$
\begin{aligned}
\textbf{Regression:}\quad & Y \in \mathbb{R} \\
& Y = X\beta^* + \varepsilon,
\end{aligned}
$$

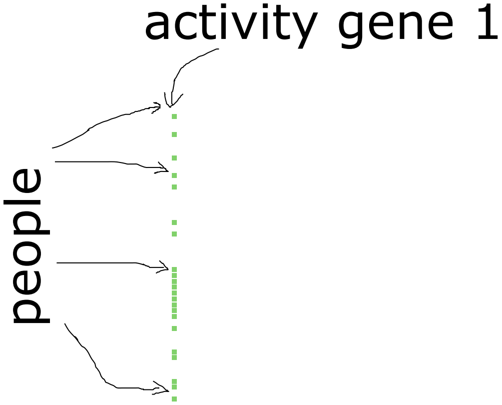
$$
\begin{aligned}
\textbf{Classification:}\quad & Y \in \{0,1\} \text{ or } \{-1,1\} \\
& P(Y=1) = f(X\beta^*), \quad \text{where } f(x) = 1/(1+\exp(-x))
\end{aligned}
$$

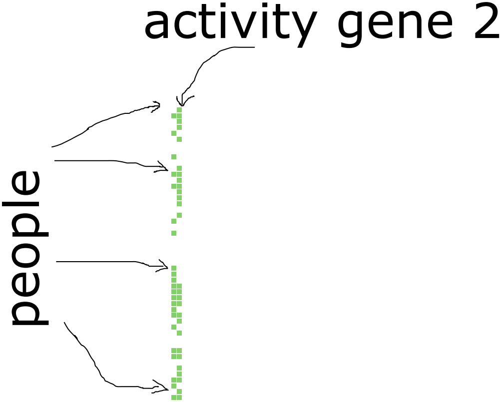
for some (sparse) vector $\beta^* \in \mathbb{R}^p$, noise $\varepsilon \in \mathbb{R}$.
Regression (or classification) is high-dimensional if $p \gg n$.

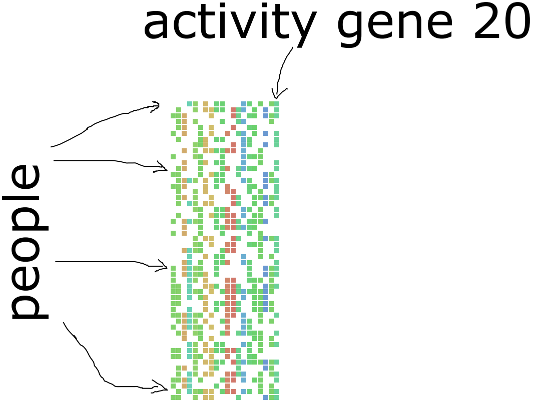
Historical start: Microarray data (Golub et al., 1999)

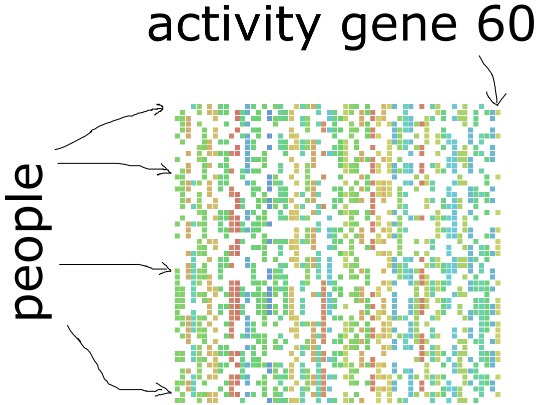
Gene expression levels of more than 3000 genes are measured for n = 72 patients, either suffering from acute lymphoblastic leukemia ("X", 47 cases) or acute myeloid leukemia ("O", 25 cases). Obtained from Affymetrix oligonucleotide microarrays.

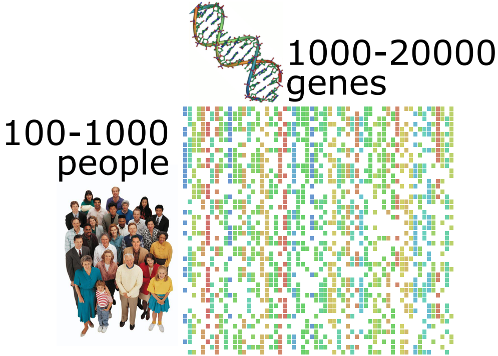
A look at (a binary version of) the data for a subset of patients and genes. Gene 1 is here either modelled as on (above average activity; filled green square) or off (below average activity; empty square)



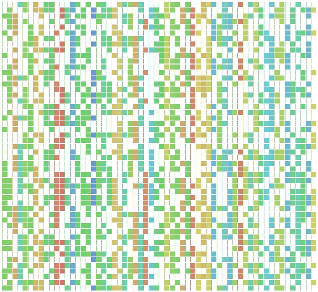

activity gene 2
people

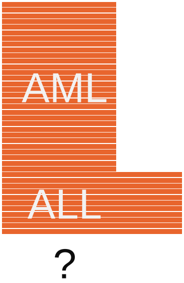

activity gene 20
people


activity gene 60
people



We have more variables (genes) than observations (patients): high-dimensional data

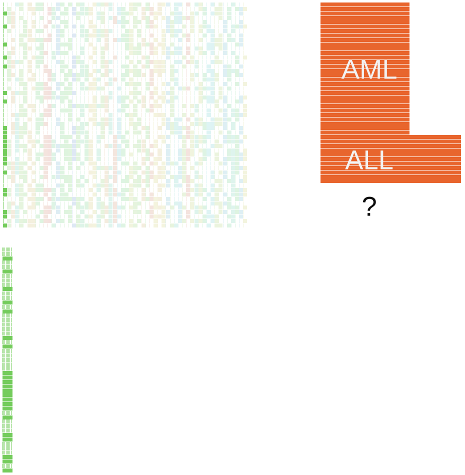




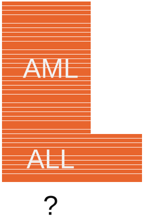
Red bars show three types of people:

- **AML**: known to have **a**cute **m**yeloid **l**eukemia
- **ALL**: known to have **a**cute **l**ymphocytic **l**eukemia
- **?**: we dont known which subtype it is

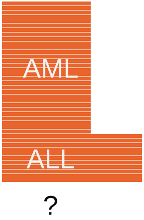
select first gene 8 times... (non-integer values are also allowed)



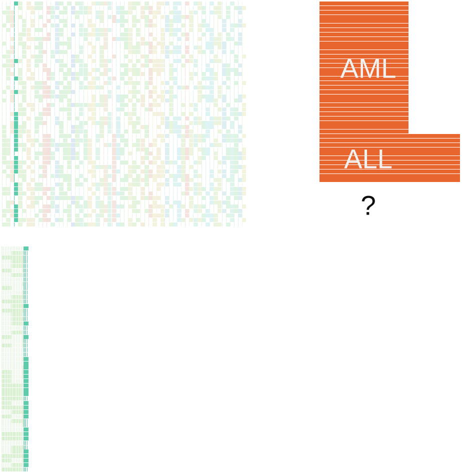
select second gene 9 times...



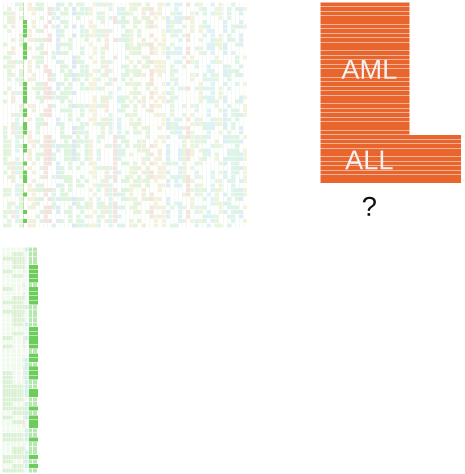
select third gene once..

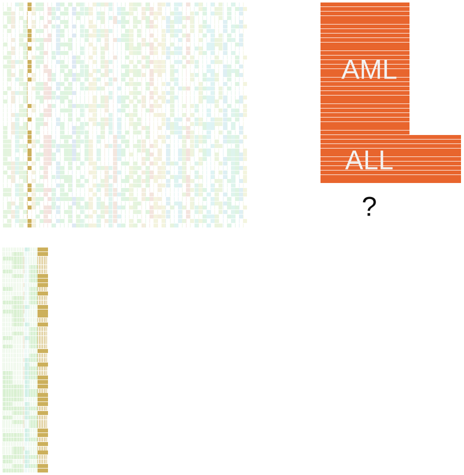


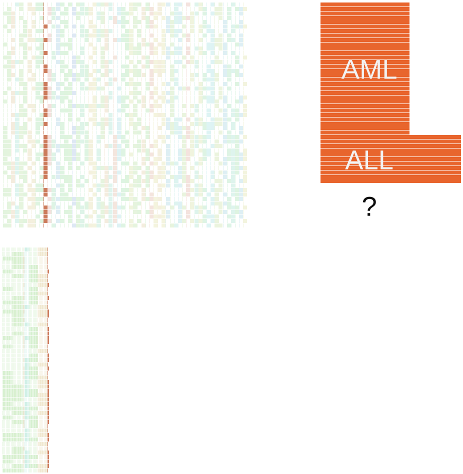
select fourth gene 4 times...

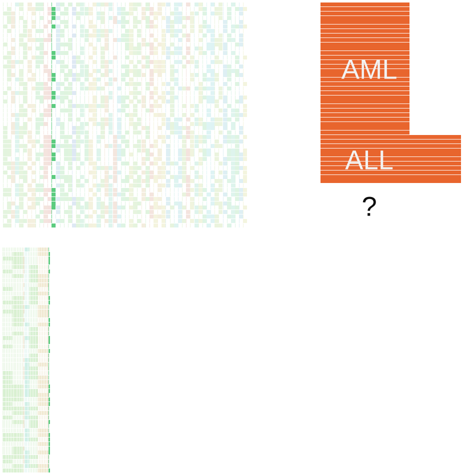


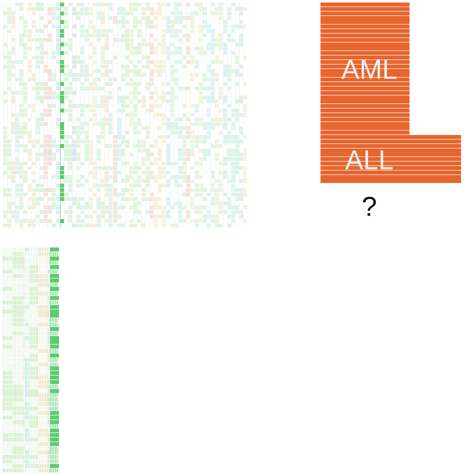
select fifth gene not at all, sixth gene 7 times...

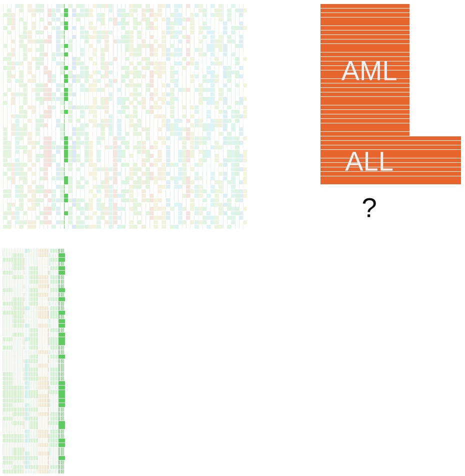


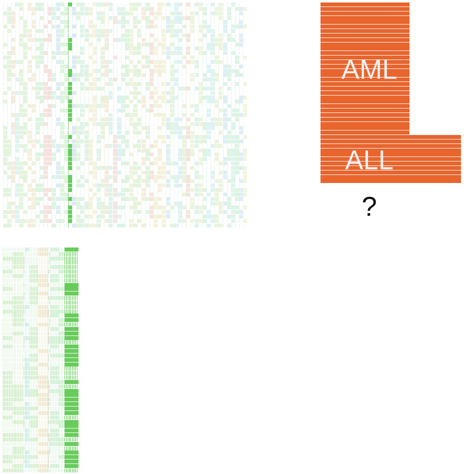

AML
ALL
?

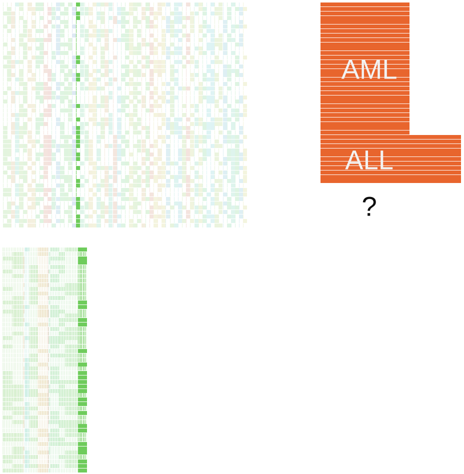

AML
ALL
?

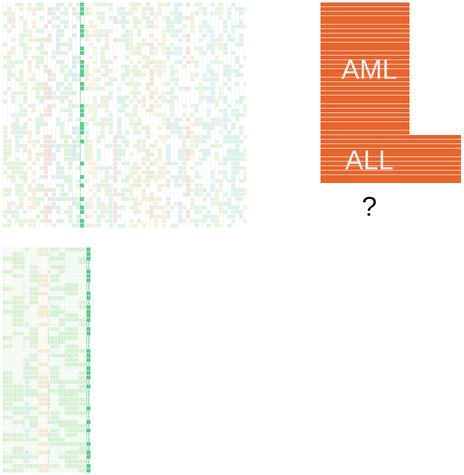

AML
ALL
?

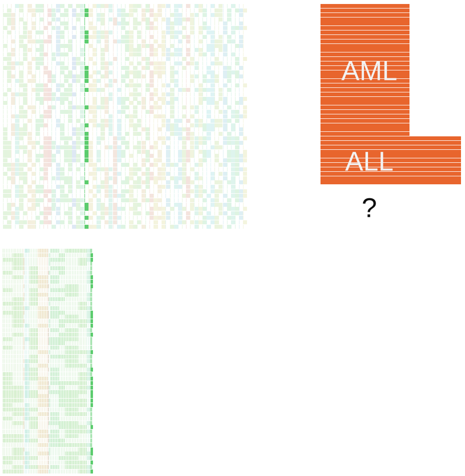

AML
ALL
?


AML
ALL
?

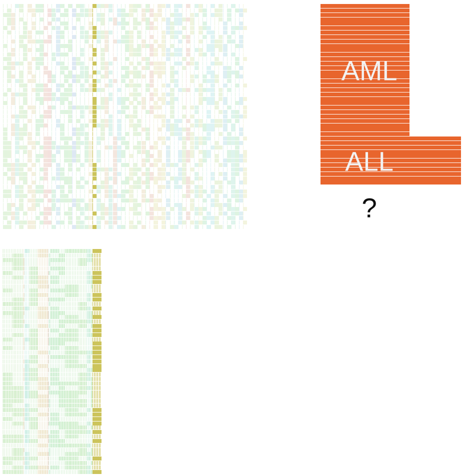

AML
ALL
?

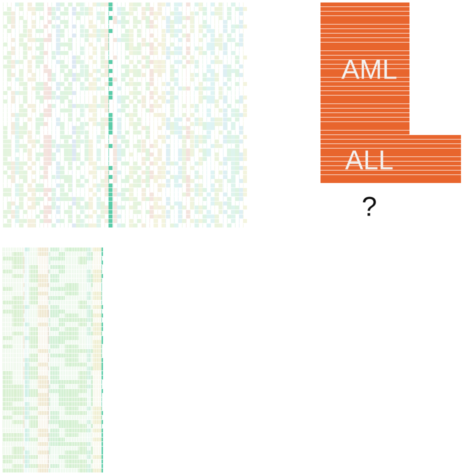

AML
ALL
?

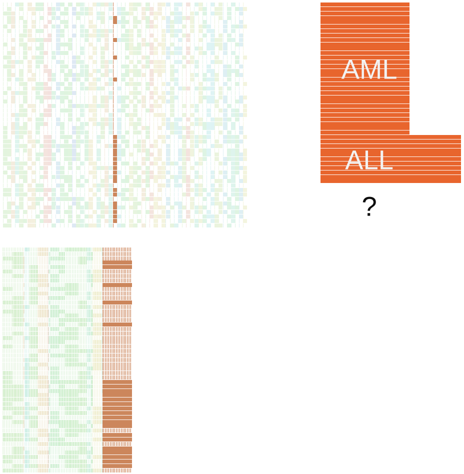

AML
ALL
?

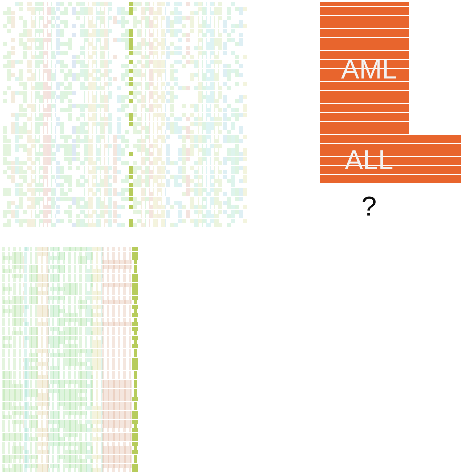


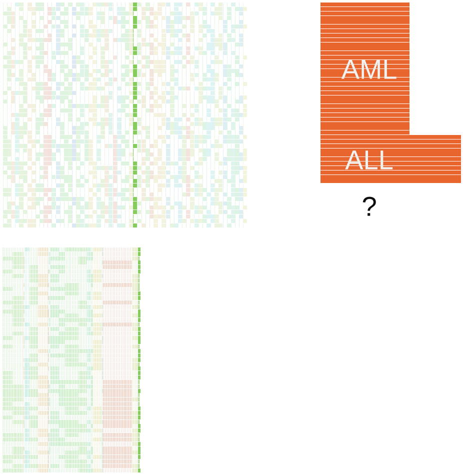
AML
ALL
?

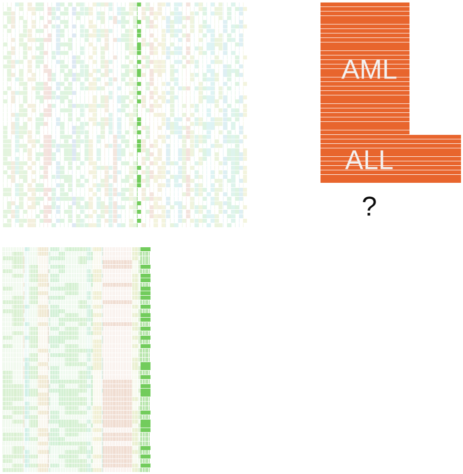

AML
ALL
?

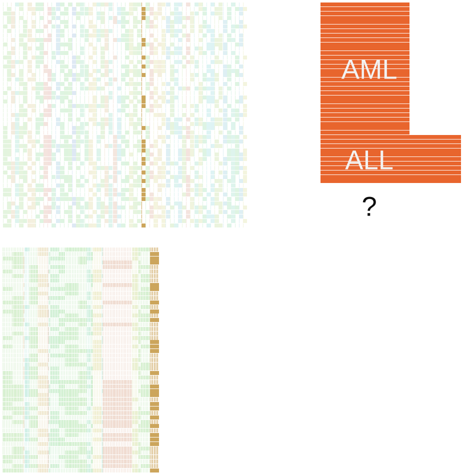

AML
ALL
?

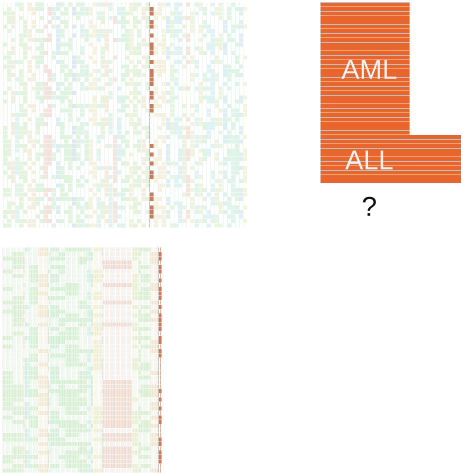

AML
ALL
?

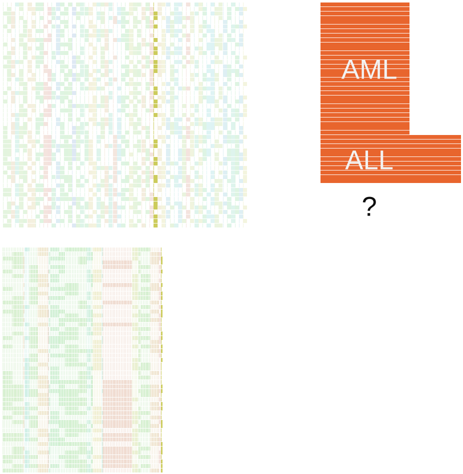

AML
ALL
?

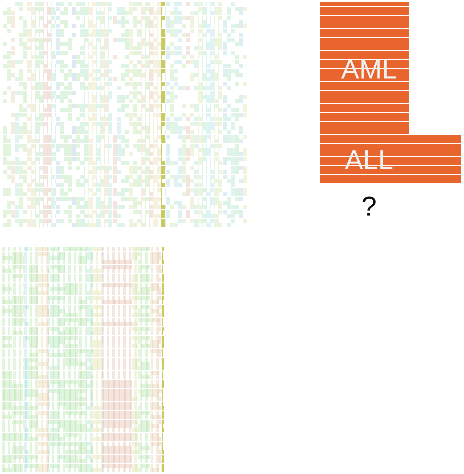

AML
ALL
?

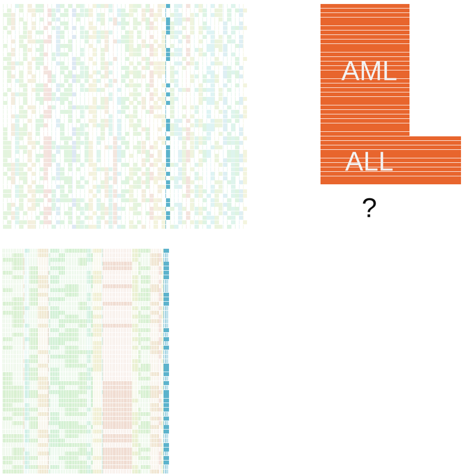

AML
ALL
?

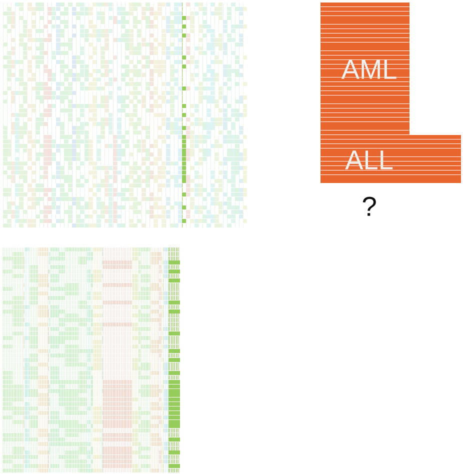

AML
ALL
?

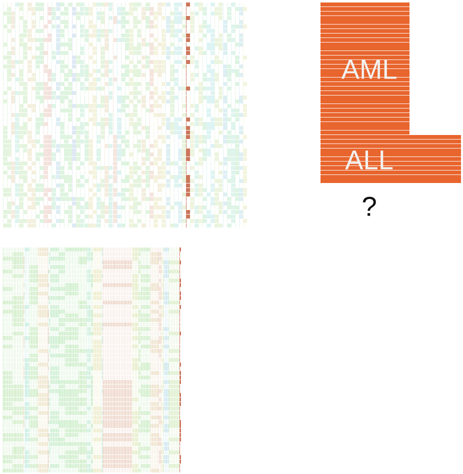

AML
ALL
?

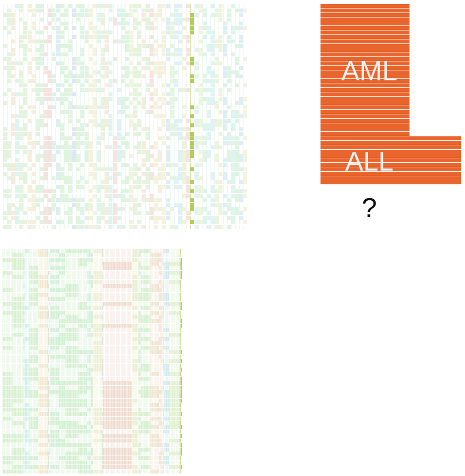

AML
ALL
?

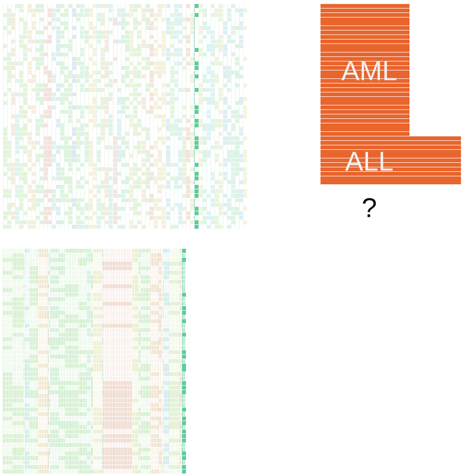

AML
ALL
?


AML
ALL
?

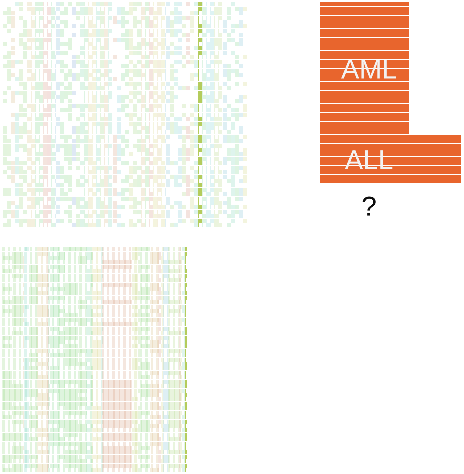

AML
ALL
?

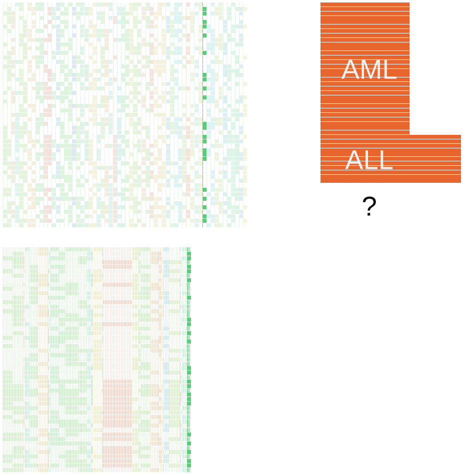

AML
ALL
?

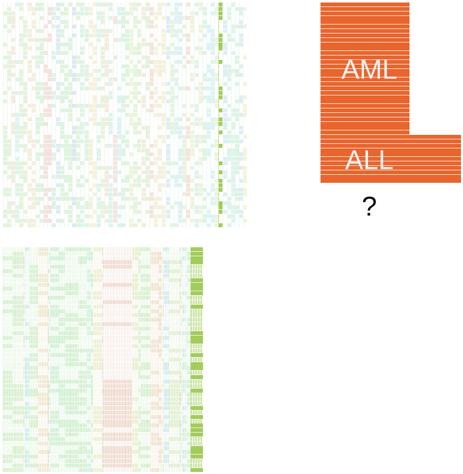

AML
ALL
?

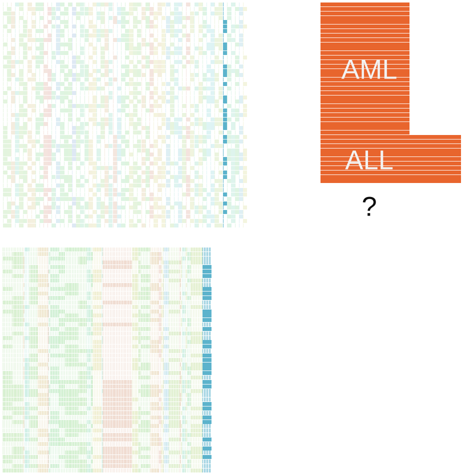


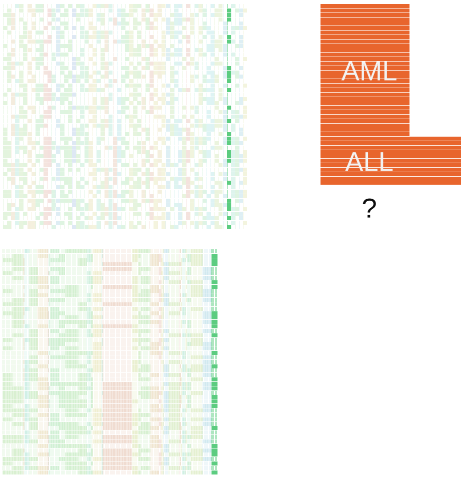
AML
ALL
?

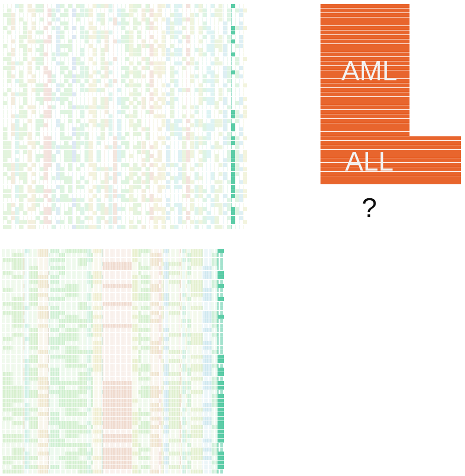

AML
ALL
?

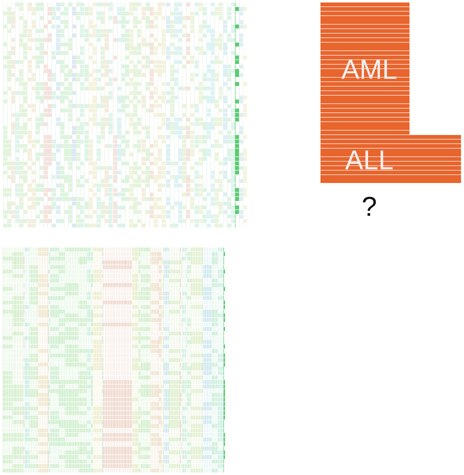

AML
ALL
?

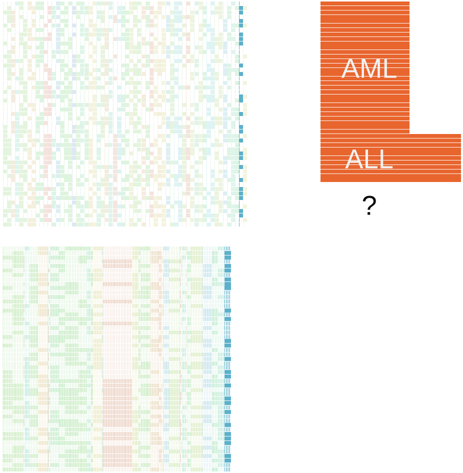

AML
ALL
?


AML
ALL
?

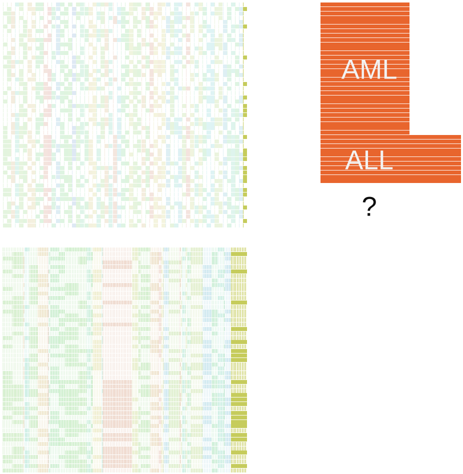

AML
ALL
?

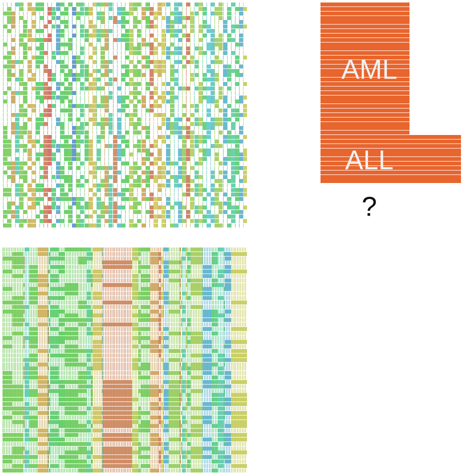

AML
ALL
?

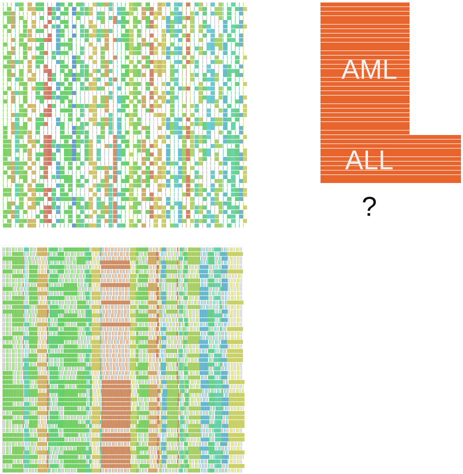

AML
ALL
?



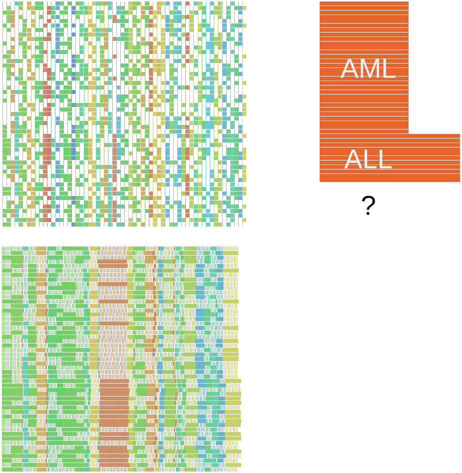
AML
ALL
?

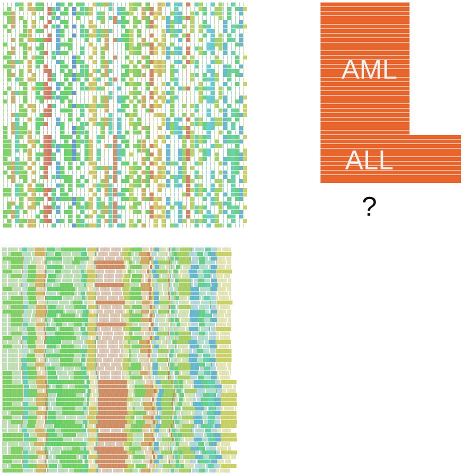

AML
ALL
?

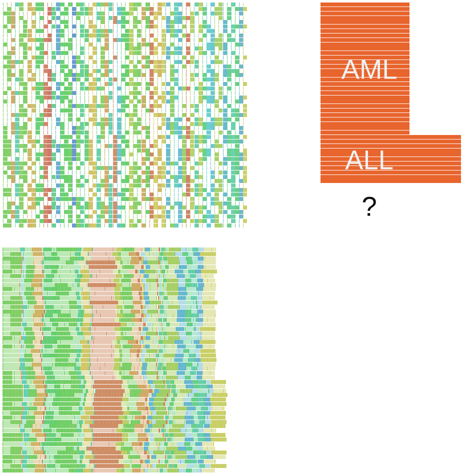

AML
ALL
?

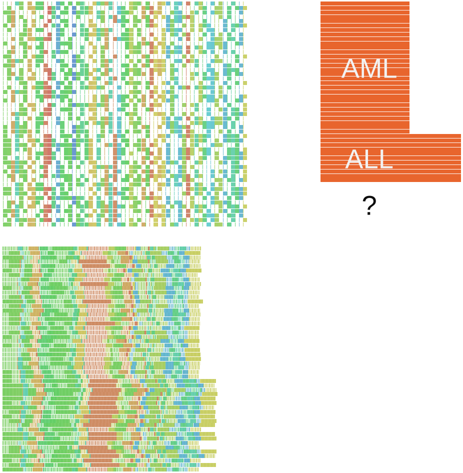


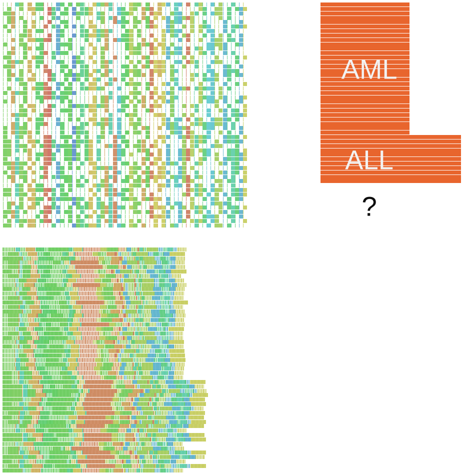
AML
ALL
?

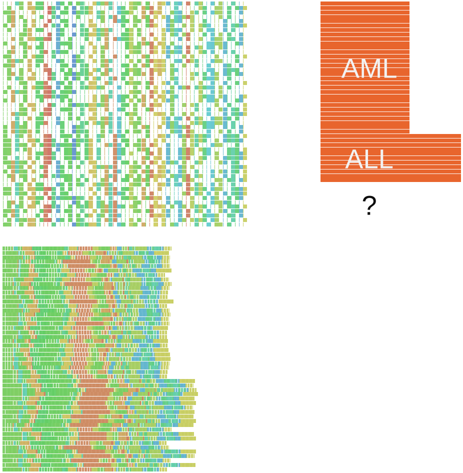

AML
ALL
?

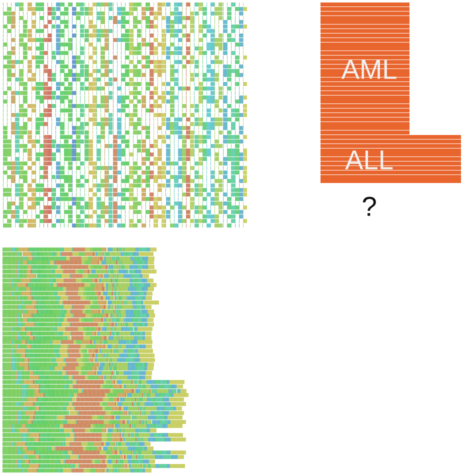

AML
ALL
?

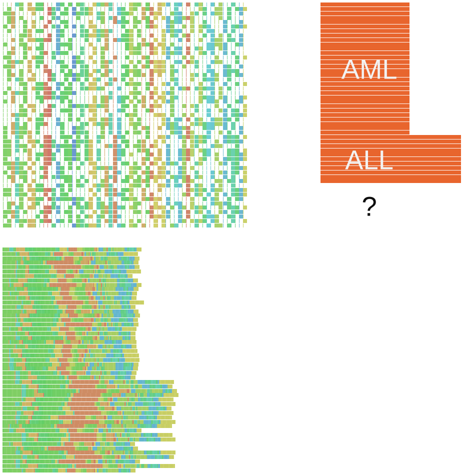

AML
ALL
?


AML
ALL
?

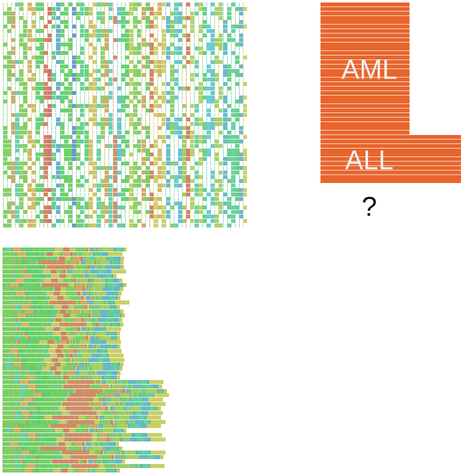

AML
ALL
?

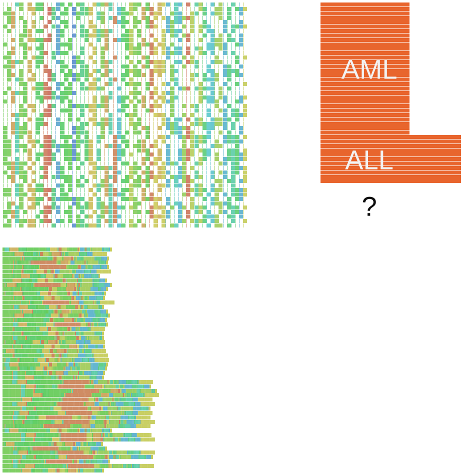

AML
ALL
?

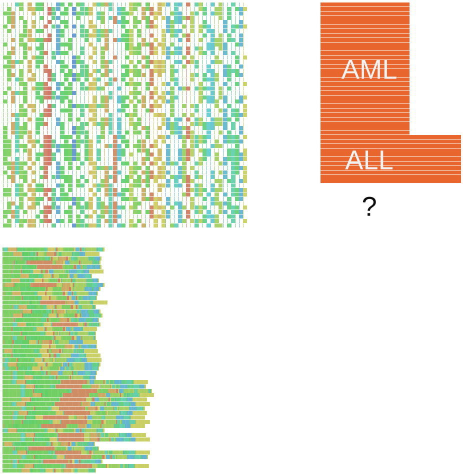

AML
ALL
?

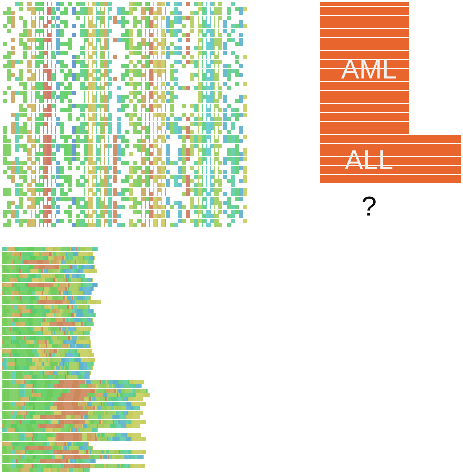


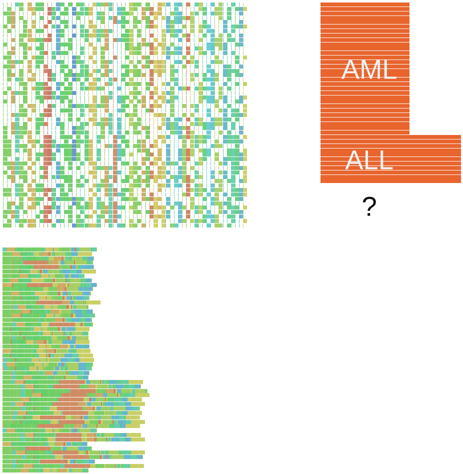
AML
ALL
?

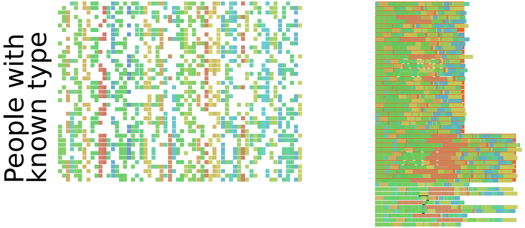

AML
ALL
?

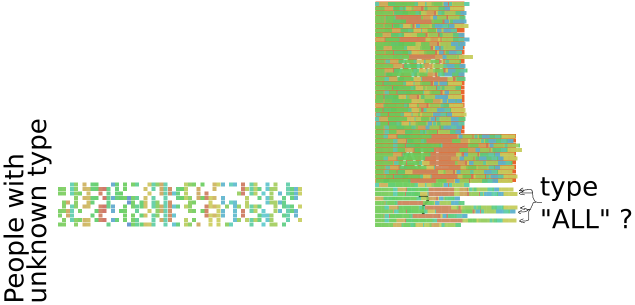

AML
ALL
?


AML
ALL
?


AML
ALL
?


AML
ALL
?


People with
known type


People with
unknown type
type
"ALL" ?

Selecting a small subset of variables

How do we get the best set of 10 genes out of all available variables?

- If we check all possible combinations of best set of 10 genes out of 60 genes in total, and a computer that checks a million sets per second, it takes about

$$20.9 \text{ hours } \approx 1 \text{ day.}$$

Selecting a small subset of variables

How do we get the best set of 10 genes out of all available variables?

- If we check all possible combinations of best set of 10 genes out of 60 genes in total, and a computer that checks a million sets per second, it takes about

$$20.9 \text{ hours} \approx 1 \text{ day.}$$

- If we have to select the best set of 10 genes out of 3000 genes, and have thousand such machines, it takes about

  500 x estimated time since big bang

Basis Pursuit (Chen et al. 99) and Lasso (Tibshirani 96)

Let Y be the n-dimensional response vector and X the $n \times p$-dimensional design.

Basis Pursuit:

$$\hat{\beta} = \text{argmin } \|\beta\|_1 \text{ such that } Y = X\beta.$$

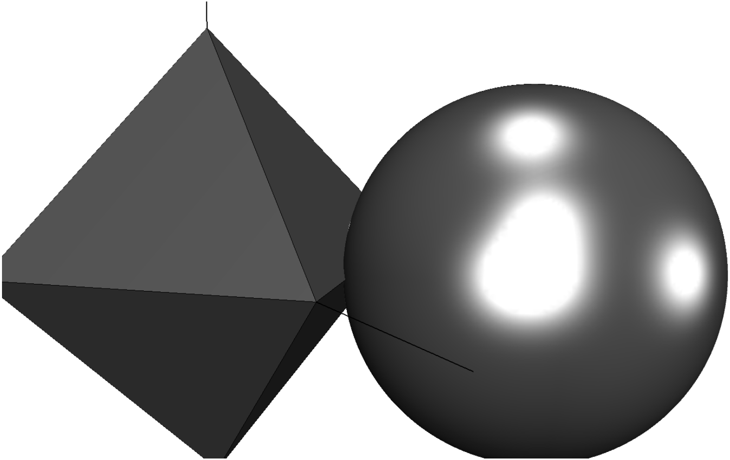
Lasso:

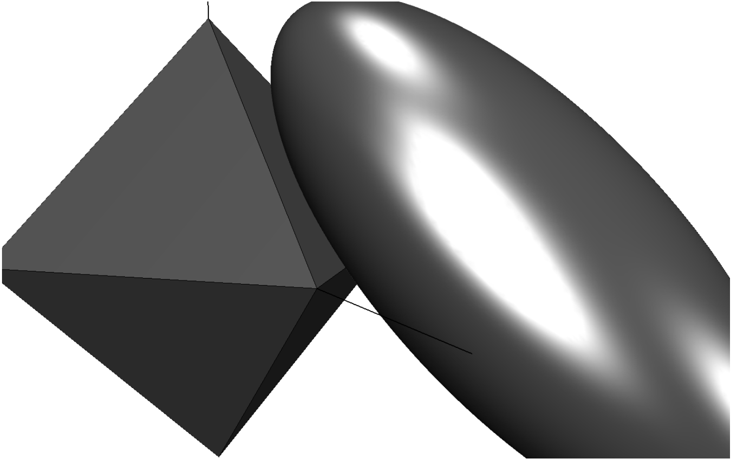
$$\hat{\beta}^\tau = \text{argmin } \|\beta\|_1 \text{ such that } \|Y - X\beta\|_2 \le \tau.$$

Equivalent to

$$\hat{\beta}^\lambda = \text{argmin } \|Y - X\beta\|_2 + \lambda\|\beta\|_1.$$

Combines sparsity (some $\hat{\beta}$-components are 0) and convexity.





When does it work?

- For *prediction* oracle inequalities in the sense that

$$\|X(\hat{\beta} - \beta^*)\|_2^2/n \;\leq\; c\sigma^2 \frac{\log(p)s}{n}$$

  for some constant $c > 0$ and noise variance $\sigma^2 > 0$, need *Restricted Isometry Property* (Candes, 2006) or weaker *compatibility condition* (Geer, 2008). Slower convergence rates possible with weaker assumptions (Greenstein and Ritov, 2004).

When does it work?

- For *prediction* oracle inequalities in the sense that

$$\|X(\hat{\beta} - \beta^*)\|_2^2/n \;\le\; c\sigma^2 \frac{\log(p)s}{n}$$

  for some constant $c > 0$ and noise variance $\sigma^2 > 0$, need *Restricted Isometry Property* (Candes, 2006) or weaker *compatibility condition* (Geer, 2008). Slower convergence rates possible with weaker assumptions (Greenstein and Ritov, 2004).

- For correct variable selection in the sense that

$$P\Big(\exists \lambda : \{k : \hat{\beta}_k^\lambda \neq 0\} = \{k : \beta_k^* \neq 0\}\Big) \approx 1,$$

  need strong *irrepresentable* (Zhao and Yu, 2006) or *neighbourhood stability* condition (NM and Bühlmann, 2006).

Compatibility condition

The usual minimal eigenvalue of the design

$$\min\{\|X\beta\|_2^2 : \|\beta\|_2 = 1\}$$

always vanishes for high-dimensional data with $p > n$.

Compatibility condition

The usual minimal eigenvalue of the design

$$\min\{\|X\beta\|_2^2 : \|\beta\|_2 = 1\}$$

always vanishes for high-dimensional data with $p > n$.

The ϕ be the (L, S)-restricted eigenvalue (Geer, 2007):

$$\phi^2(L,S) \; = \min\{s\|X\beta\|_2^2 : \|\beta_S\|_1 = 1 \text{ and } \|\beta_{S^c}\|_1 \leq L\},$$

where

- $S = \{k : \beta_k^* \neq 0\}$,
- $s = |S|$, and
- $(\beta_S)_k = \beta_k 1\{k \in S\}$

.

- If $\phi(L, S) > c > 0$ for some $L > 1$, then we get oracle rates for prediction and convergence of $\|\beta^* - \hat{\beta}^\lambda\|_1$.
- If $\phi(1, S) > 0$, then the following two are identical

$$\text{argmin}\|\beta\|_0 \text{ such that } X\beta = X\beta^*$$
$$\text{argmin}\|\beta\|_1 \text{ such that } X\beta = X\beta^*.$$

- If $\phi(L, S) > c > 0$ for some $L > 1$, then we get oracle rates for prediction and convergence of $\|\beta^* - \hat{\beta}^\lambda\|_1$.
- If $\phi(1, S) > 0$, then the following two are identical

$$\text{argmin}\|\beta\|_0 \text{ such that } X\beta = X\beta^*$$
$$\text{argmin}\|\beta\|_1 \text{ such that } X\beta = X\beta^*.$$

The latter equivalence requires otherwise the stronger *Restricted Isometry Property* which implies that $\exists\delta < 1$ such that

$$\forall b \text{ with } \|b\|_0 \leq s : \qquad (1-\delta)\|b\|_2^2 \leq \|Xb\|_2^2 \leq (1+\delta)\|b\|_2^2,$$

which can be a useful assumption for random designs X, as in compressed sensing.

Applications of linear models

GxP-Validierung
Validierung von GxP relevanten
Systemen einfach und verständlich
www.q-finity.de

Go Go Kurse Züri Oberland
Kursstart: Samstag 26. April 2014
14:30 -15:30 Uhr / 5 x 60 min
www.dance4fun.ch

CAS Unternehmensführung
Bilden Sie sich auf hohem universi-
tären Niveau in Wirtschaft weiter!
www.cas-guf.uzh.ch

Abplanalp Ramsauer AG
Archivreorganisation, Archivplan
Behördenberatung, Stellvertretung
www.abplanalp.ch

Kommunikations-Coaching
Spezialisiert auf Konfliktlösungen.
Verbesserung Ihrer Kommunikation.
www.ladera.ch

SAL Sprachausbildungen
Diplom-Ausbildung in Journalismus,
Sprachunterricht, Übersetzen.
www.sal.ch

Seminar- & Schulungsräume
mit moderner Infrastruktur
in der Stadt Zürich
www.cf-studies.ch

Applications of linear models

Applications of linear models

GxP-Validierung
Validierung von GxP relevanten Systemen einfach und verständlich
www.q-finity.de

Go Go Kurse Züri Oberland
Kursstart: Samstag 26. April 2014
14:30 -15:30 Uhr / 5 x 60 min
www.dance4fun.ch

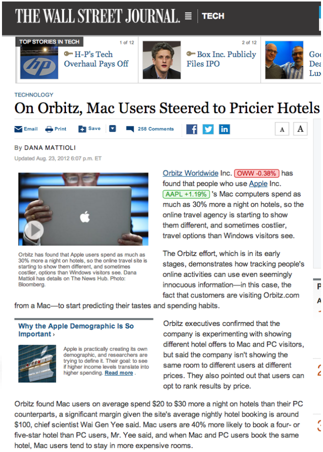
CAS Unternehmensführung
Bilden Sie sich auf hohem universitären Niveau in Wirtschaft weiter!
www.cas-guf.uzh.ch

Abplanalp Ramsauer AG
Archivreorganisation, Archivplan
Behördenberatung, Stellvertretung
www.abplanalp.ch

Kommunikations-Coaching
Spezialisiert auf Konfliktlösungen. Verbesserung Ihrer Kommunikation.
www.ladera.ch

SAL Sprachausbildungen
Diplom-Ausbildung in Journalismus, Sprachunterricht, Übersetzen.
www.sal.ch

Seminar- & Schulungsräume
mit moderner Infrastruktur
in der Stadt Zürich
www.cf-studies.ch


THE WALL STREET JOURNAL. | TECH

TOP STORIES IN TECH
1 of 12 H-P's Tech Overhaul Pays Off
2 of 12 Box Inc. Publicly Files IPO

TECHNOLOGY

On Orbitz, Mac Users Steered to Pricier Hotels

Email | Print | Save | 258 Comments

By DANA MATTIOLI
Updated Aug. 23, 2012 6:07 p.m. ET

Orbitz has found that Apple users spend as much as 30% more a night on hotels, so the online travel site is starting to show them different, and sometimes costlier, options than Windows visitors see. Dana Mattioli has details on The News Hub. Photo: Bloomberg.

Orbitz Worldwide Inc. OWW -0.38% has found that people who use Apple Inc. AAPL +1.19% 's Mac computers spend as much as 30% more a night on hotels, so the online travel agency is starting to show them different, and sometimes costlier, travel options than Windows visitors see.

The Orbitz effort, which is in its early stages, demonstrates how tracking people's online activities can use even seemingly innocuous information—in this case, the fact that customers are visiting Orbitz.com from a Mac—to start predicting their tastes and spending habits.

Why the Apple Demographic Is So Important ›

Apple is practically creating its own demographic, and researchers are trying to define it. Their goal: to see if higher income levels translate into higher spending. Read more.

Orbitz executives confirmed that the company is experimenting with showing different hotel offers to Mac and PC visitors, but said the company isn't showing the same room to different users at different prices. They also pointed out that users can opt to rank results by price.

Orbitz found Mac users on average spend \$20 to \$30 more a night on hotels than their PC counterparts, a significant margin given the site's average nightly hotel booking is around \$100, chief scientist Wai Gen Yee said. Mac users are 40% more likely to book a four- or five-star hotel than PC users, Mr. Yee said, and when Mac and PC users book the same hotel, Mac users tend to stay in more expensive rooms.

Medical data

OMOP: Observational Medical Outcomes Project (omop.org)

1. Collect medical information (drugs taken, symptoms diagnosed) for 100.000 patients
2. In total, about 15.000 drugs and 15.000 distinct symptoms encoded.

Try to detect drug-drug interactions or make risk assesements based on medical data:
Is drug A changing the risk of a heart attack if taken together with drug B for patients with a symptom S ?

Try to detect drug-drug interactions or make risk assesments based on medical data:
Is drug A changing the risk of a heart attack if taken together with drug B for patients with a symptom S ?

Can generate very high-dimensional data quickly if expanding interactions as new dummy variables (more than $> 10^{12}$ interactions of third order).

Compressed sensing: one-pixel camera

High quality JPEG
File Size: 77.9 kb

Medium quality JPEG
File Size: 19.11 kb

Images are often sparse after taking a wavelet transformation X:

$$u = Xw, \qquad \text{where}$$

- $w \in \mathbb{R}^n$: original image as n-dimensional vector
- $X \in \mathbb{R}^{n \times n}$: wavelet transformation
- $u \in \mathbb{R}^n$: vector with wavelet coefficients

High quality JPEG
File Size: 77.9 kb

Medium quality JPEG
File Size: 19.11 kb

Original wavelet transformation:

$$u = Xw, \qquad \text{where}$$

The wavelet coefficients u are often sparse in the sense that it has only a few large entries. Keeping just a few of them allows a very good reconstruction of the original image w.
Let $\tilde{u} = u1\{|U| \geq \tau\}$ be the hard-thresholded coefficients (easy to store). Then re-construct image as $\tilde{w} = X^{-1}\tilde{u}$.

Conventional way:

- measure image w with 16 million pixels
- convert to wavelet coefficients $u = Xw$
- throw away most of u by keeping just the largest coefficients

Is efficient as long as pixels are cheap.

For situations where pixels are expensive (different wavelengths, MRI) can do compressed sensing: observe only

$$y = \Phi u = \Phi(Xw),$$

where for $q \ll n$, matrix $\Phi \in \mathbb{R}^{q \times n}$ has iid entries drawn from $\mathcal{N}(0, 1)$. One entry of q-dimensional vector y is thus observed by a random transformation of the original image.



Each random mask corresponds to one row of Φ.

Reconstruct u by Basis Pursuit:

$$\hat{u} = \operatorname{argmin} \|\tilde{u}\|_1 \text{ such that } \Phi\tilde{u} = y.$$

Observe

$$y = \Phi u = \Phi(Xw),$$

where for $q \ll n$, matrix $\Phi \in \mathbb{R}^{q \times n}$ has iid entries drawn from $\mathcal{N}(0,1)$. Reconstruct wavelet coefficients u by Basis Pursuit:

$$\hat{u} = \mathrm{argmin} \|\tilde{u}\|_1 \text{ such that } \Phi\tilde{u} = y.$$

Observe

$$y = \Phi u = \Phi(Xw),$$

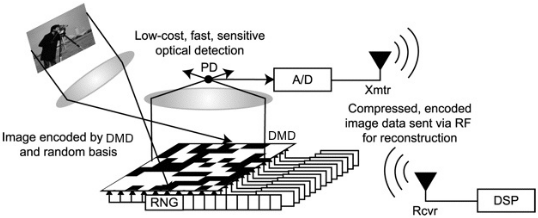
where for $q \ll n$, matrix $\Phi \in \mathbb{R}^{q \times n}$ has iid entries drawn from $\mathcal{N}(0, 1)$. Reconstruct wavelet coefficients u by Basis Pursuit:

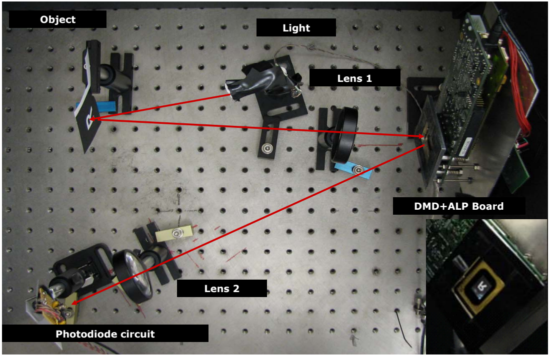
$$\hat{u} = \text{argmin} \|\tilde{u}\|_1 \text{ such that } \Phi\tilde{u} = y.$$

Matrix Φ satisfies for $q \geq s \log(p/s)$ with high probability the *Random Isometry Property*, including the existence of a $\delta < 1$ such that (Candes, 2006) for all s-sparse vectors

$$(1 - \delta)\|b\|_2^2 \leq \|\Phi b\|_2^2 \leq (1 + \delta)\|b\|_2^2.$$

Hence, if original wavelet coeffcients are s-sparse, we only need to make of order $s \log(n/s)$ measurements to recover u exactly (with high probability)!



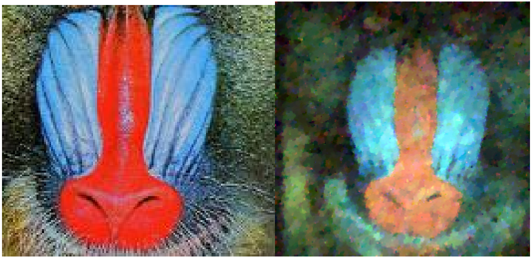


dsp.rice.edu/cs/camera

Original

16384 Pixels
1600 Measurements
(10%)

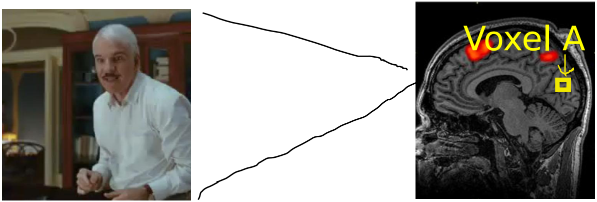
16384 Pixels
3300 Measurements
(20%)



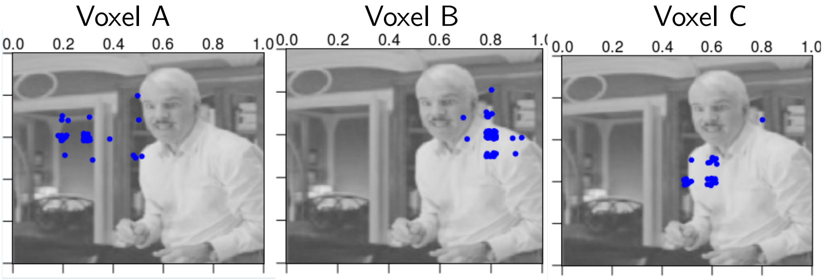
dsp.rice.edu/cs/camera

Mind reading

Can use Lasso-type inference to infer for a single voxel in the early visual cortex which stimuli lead to neuronal activity using fmri-measurements (Nishimoto et al., 2011 at Gallant Lab, UC Berkeley).



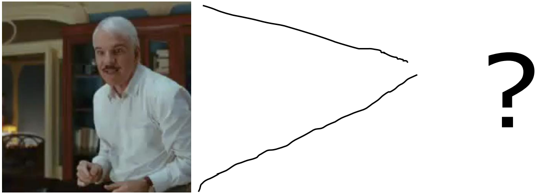
Show movies and detect which parts of the image a particular voxel of 100k neurons is sensitive to.



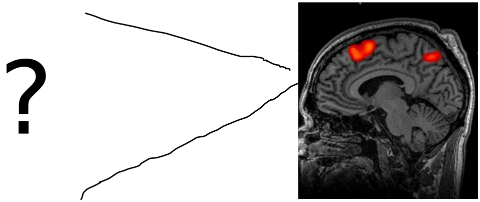
Dots indicate large regression coefficients and thus important regions for a region/voxel in the brain:

- Voxel A is stimulated by activity in the centre-left of the visual field
- Voxel B is stimulated by activity in the top right of the visual field
- Voxel C is stimulated by activity in the very centre of the visual field

Allows to forecast brain activity at all voxels, given an image.



Given only brain activity, can reverse the process and ask which image best explains the neuronal activity (given the learned regressions).



Top: seen image/movie
Bottom: image reconstructed from brain activity