A Geometric Perspective on Feature Selection

Pablo G. Camara

Assistant Professor of Genetics
Perelman School of Medicine, University of Pennsylvania

camara-lab.org

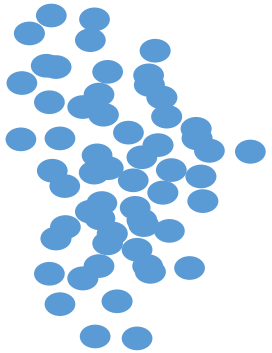
Abel Symposium, Geiranger, Norway, June 5, 2018

Feature Selection

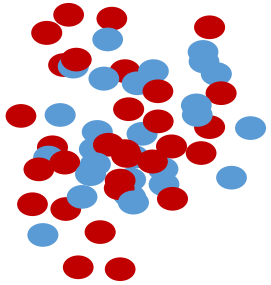
A classical problem in statistics: select features that differ among two populations

$$f_i\colon \{x_k\} \rightarrow \mathrm{F}$$

Population A



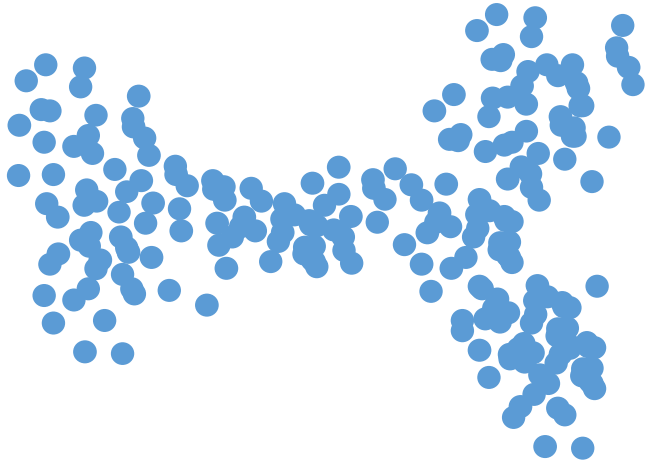
Population B



Non-parametric tests: Kolmogorov-Smirnov (1939), Wilcoxon (1945), Mann-Whitney *U* (1947)…

Unsupervised Feature Selection

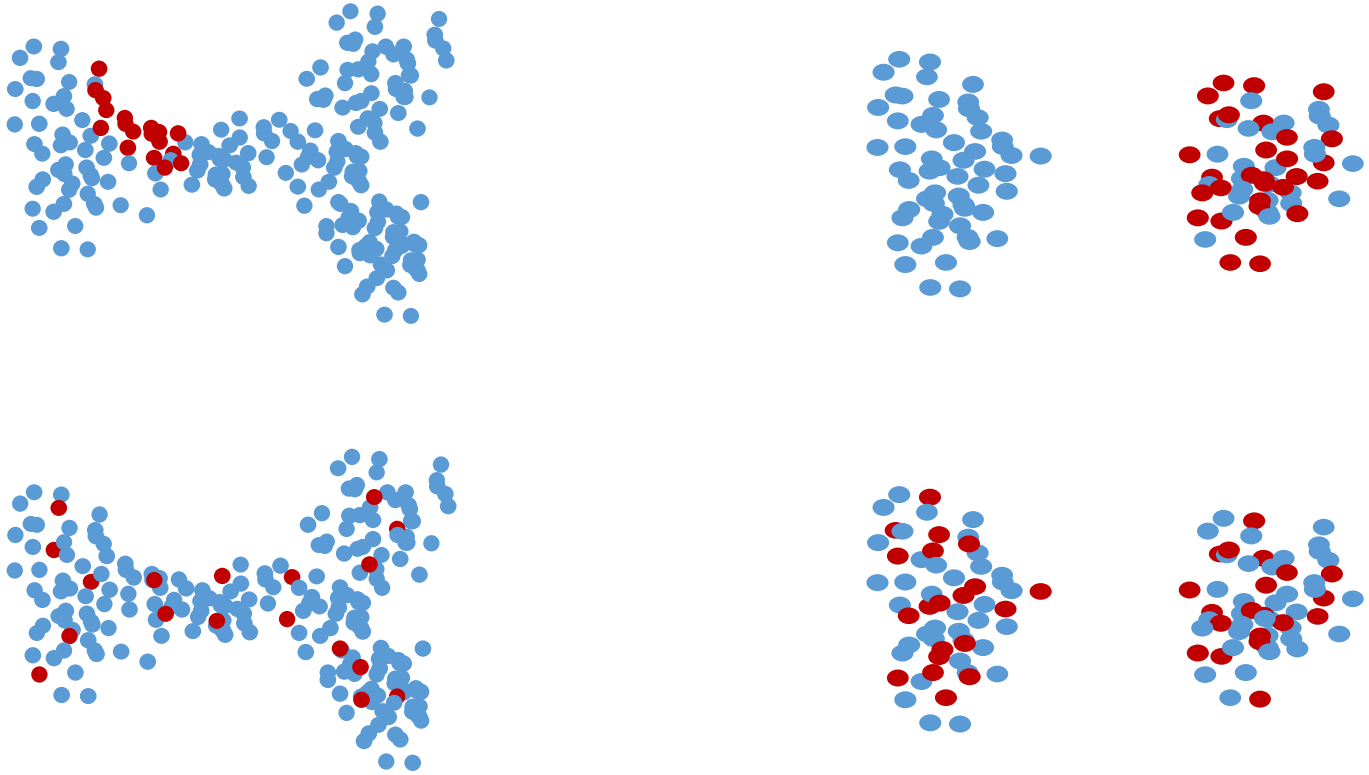
Often samples cannot be arranged into discrete populations



Can we select differential features without predefining populations?

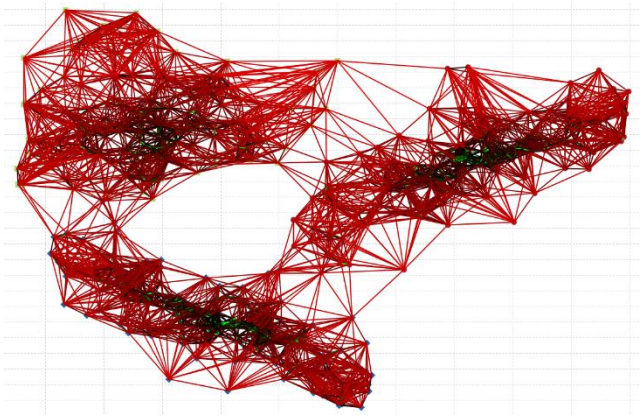
Unsupervised Feature Selection

Variance in F is often used to rank features. However, it does not make use of the metric structure of $\{x_k\}$ when available.



Laplacian Score

k-nearest neighbor graph G



Graph Laplacian: $L = D - A$

$$D = diag(A \cdot \boldsymbol{I})$$

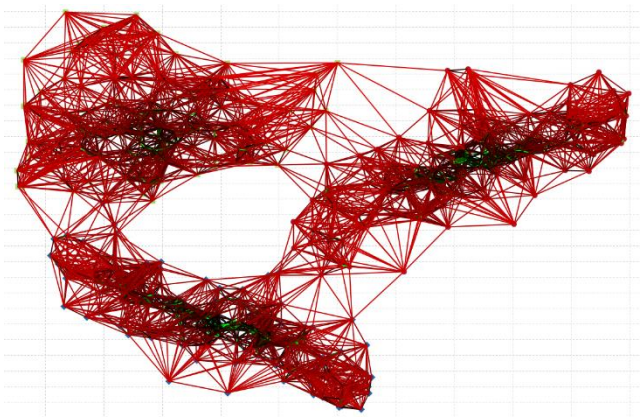
Laplacian score: $\mathrm{R}(f_i) = \frac{\langle \tilde{f}_i, L\tilde{f}_i \rangle_G}{\langle \tilde{f}_i, \tilde{f}_i \rangle_G} = \frac{\tilde{f}_i^T L \tilde{f}_i}{\tilde{f}_i^T D \tilde{f}_i}$

$$\tilde{f}_i = f_i - \mu_i$$

He, Cai, Niyogi (2006)

Laplacian Score

k-nearest neighbor graph G



Graph Laplacian: $L = D - A$ $\quad D = diag(A \cdot \boldsymbol{I})$

Laplacian score: $\mathrm{R}(f_i) = \frac{\langle \tilde{f}_i, L\tilde{f}_i \rangle_G}{\langle \tilde{f}_i, \tilde{f}_i \rangle_G} = \frac{\tilde{f}_i^T L \tilde{f}_i}{\tilde{f}_i^T D \tilde{f}_i}$ $\quad \tilde{f}_i = f_i - \mu_i$

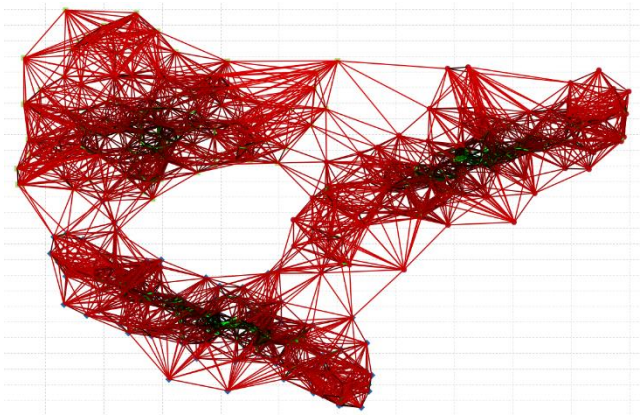
He, Cai, Niyogi (2006)

Limitations:

- Lacks statistical framework
- Curse of dimensionality
- Only uses graph structure

Laplacian Score

k-nearest neighbor graph G



Graph Laplacian: $L = D - A$

$$D = diag(A \cdot \boldsymbol{I})$$

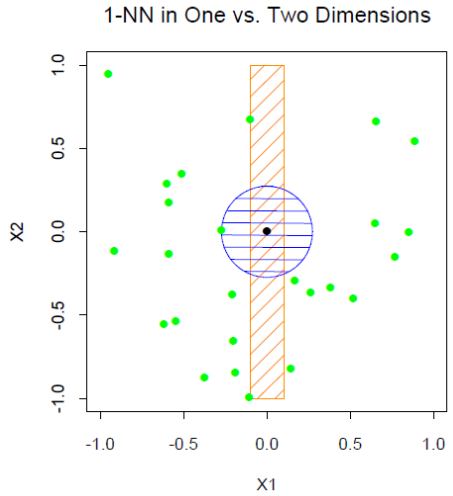
Laplacian score: $$\mathrm{R}(f_i) = \frac{\langle \tilde{f}_i, L\tilde{f}_i \rangle_G}{\langle \tilde{f}_i, \tilde{f}_i \rangle_G} = \frac{\tilde{f}_i^T L \tilde{f}_i}{\tilde{f}_i^T D \tilde{f}_i}$$

$$\tilde{f}_i = f_i - \mu_i$$

He, Cai, Niyogi (2006)

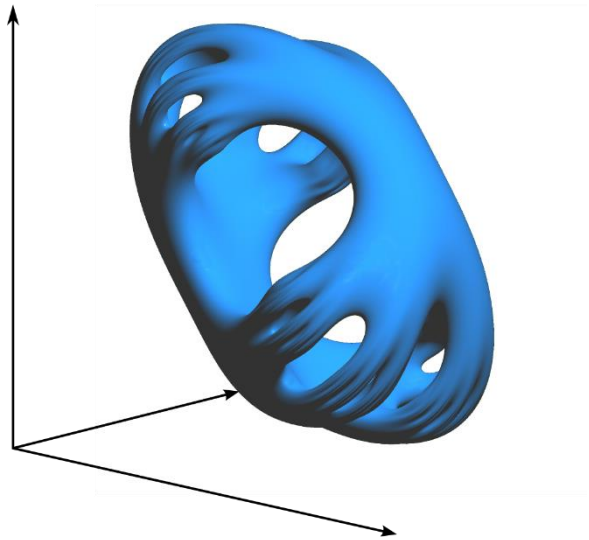
Limitations:

- Lacks statistical framework
- Curse of dimensionality
- Only uses graph structure



A Geometric Perspective on Feature Selection: Continuous Case

***D*-dimensional manifold $\mathcal{M}$**



features (differential forms in $\mathcal{M}$)

$$f_1^{(0)}, f_2^{(0)}, \dots, f_{n^{(0)}}^{(0)}$$

$$f_1^{(1)}, f_2^{(1)}, \dots, f_{n^{(1)}}^{(1)}$$

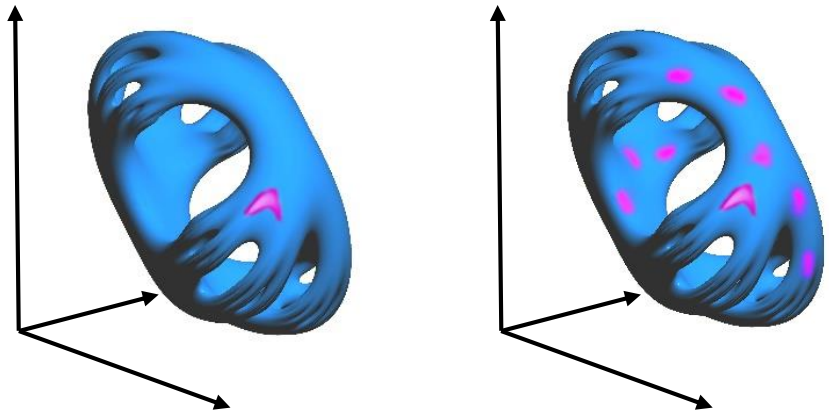
$$\vdots$$

$$f_1^{(D)}, f_2^{(D)}, \dots, f_{n^{(D)}}^{(D)}$$

A Geometric Perspective on Feature Selection: Continuous Case

Rayleigh quotient of the Laplace-Beltrami operator of $\mathcal{M}$:

$$R_{\mathcal{M}}\left(f_i^{(k)}\right) = \frac{\left\langle \tilde{f}_i^{(k)}, L^{(k)} \tilde{f}_i^{(k)} \right\rangle_{\mathcal{M}}}{\left\langle \tilde{f}_i^{(k)}, \tilde{f}_i^{(k)} \right\rangle_{\mathcal{M}}}$$

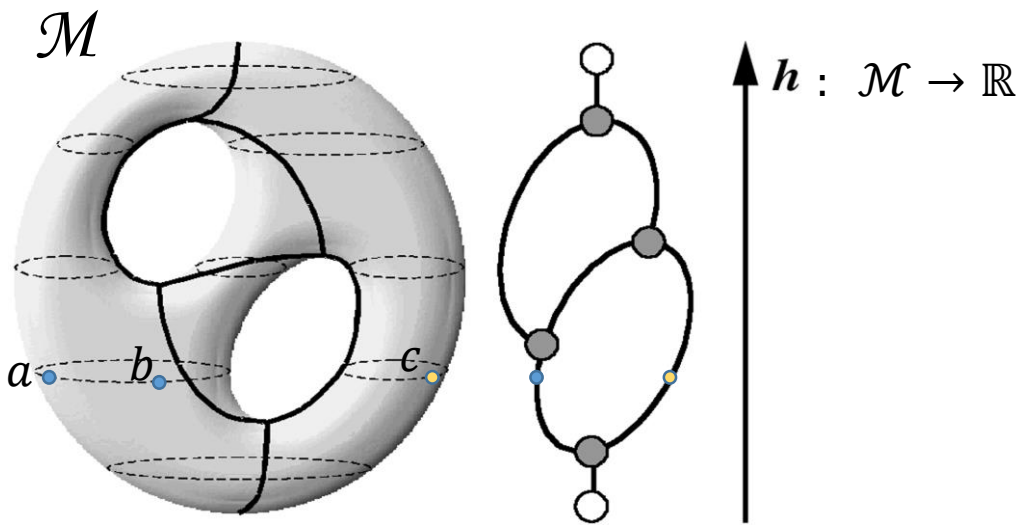


low R(f)

high R(f)

A Geometric Perspective on Feature Selection: Continuous Case

Reeb graphs reduce spaces while preserving local relationships:



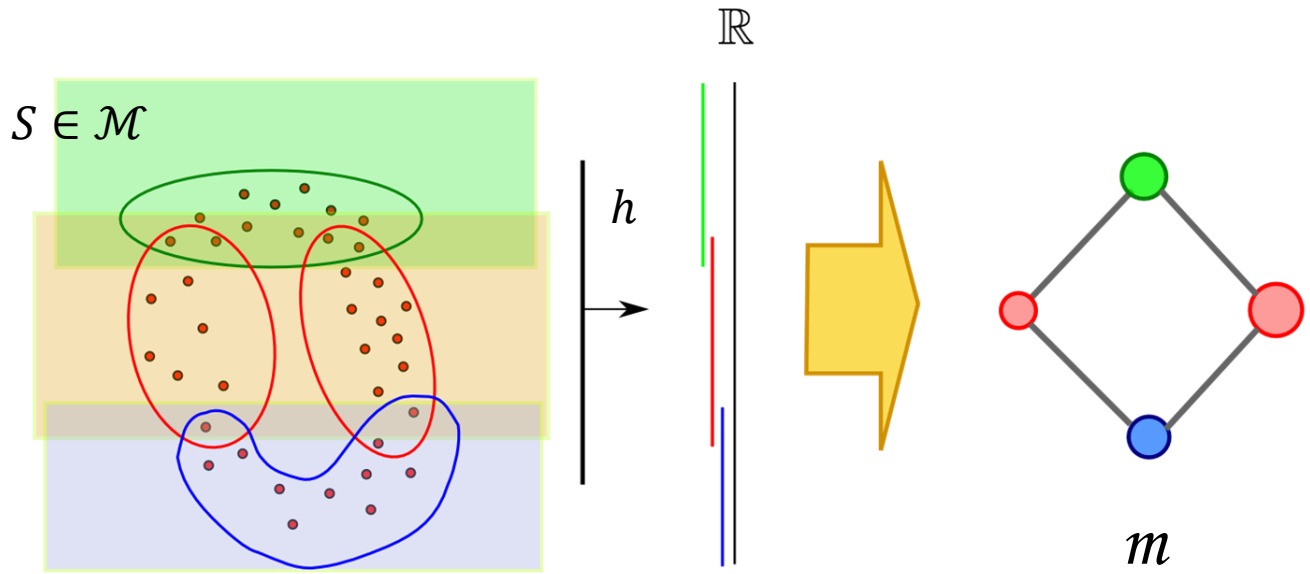
$a \sim b$ iff a and b belong to the same connected component of $h^{-1}(y)$ for some y

$$Reeb(\mathcal{M}, h) = \mathcal{M}/\sim$$

Differential forms in $\mathcal{M}$ lead to differential forms in $Reeb(\mathcal{M}, h)$ by pulling-back and integrating over connected components of $h^{-1}(y)$

A Geometric Perspective on Feature Selection: Discrete Case

Mapper provides an approximation to the topology of Reeb graphs when only given a finite set of points from $\mathcal{M}$:



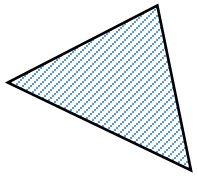
Singh, Memoli, Carlsson (2007)

Carriere, Michel, Oudot (2017)

A Geometric Perspective on Feature Selection: Discrete Case

q-forms in m: maps from q-simplices into F *Muhammad, Egerstedt (2006)*

$$f_i^{(q)}: \quad \in S_q \subset m \rightarrow \mathrm{F}$$



For each $f_i: \{x_k\} \rightarrow \mathrm{F}$ we can construct a q-form $f_i^{(q)}$ in m by averaging f_i over the elements of $\{x_k\}$ associated to each q-simplex of m.

A Geometric Perspective on Feature Selection: Discrete Case

Combinatorial Laplacian on a simplicial complex: $L^{(q)} = L^{(q),\mathrm{up}} + L^{(q),\mathrm{down}}$

$$\left(L^{(q),\mathrm{up}} f_i^{(q)}\right)([H]) = \sum_{\substack{\bar{H}\in S_{q+1}:\\ H\in\partial\bar{H}}} \frac{w(\bar{H})}{w(H)} f_i^{(q)}([H]) + \sum_{\substack{H'\in S_q: H\neq H'\\ H,H'\in\partial\bar{H}}} \frac{w(\bar{H})}{w(H)} \mathrm{sgn}([H],\partial[\bar{H}])\mathrm{sgn}([H'],\partial[\bar{H}])\, f_i^{(q)}([H'])$$

$$\left(L^{(q),\mathrm{down}} f_i^{(q)}\right)([H]) = \sum_{E\in\partial H} \frac{w(H)}{w(E)} f_i^{(q)}([H]) + \sum_{H': H\cap H' = E} \frac{w(H')}{w(E)} \mathrm{sgn}([E],\partial[H])\mathrm{sgn}([E],\partial[H'])\, f_i^{(q)}([H'])$$

Horak, Jost (2011)

A Geometric Perspective on Feature Selection: Discrete Case

Rayleigh quotient in m: $$R_m\left(f_i^{(q)}\right) = \frac{\left\langle f_i^{(q)}, L^{(q)} f_i^{(q)} \right\rangle_m}{\left\langle f_i^{(q)}, f_i^{(q)} \right\rangle_m}$$

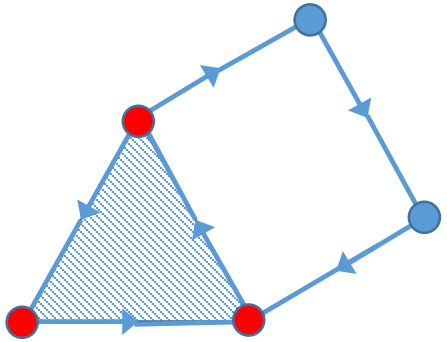
$$\left\langle f_i^{(q)}, f_j^{(q)} \right\rangle_m = \sum_{H \in S_q} w(H) f_i^{(q)}([H]) f_j^{(q)}([H])$$

For example, for 0-forms (functions on the vertices):

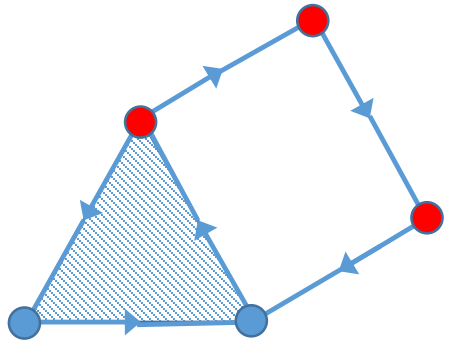
$$R\left(f^{(0)}\right) = \frac{\sum_{\alpha,\beta\in\Gamma} f_\alpha^{(0)} \left(D_{\alpha\beta} - A_{\alpha\beta}\right) f_\beta^{(0)}}{\sum_{\alpha,\beta\in\Gamma} f_\alpha^{(0)} D_{\alpha\beta} f_\beta^{(0)}}$$

Null distributions for R can be built using randomized data.

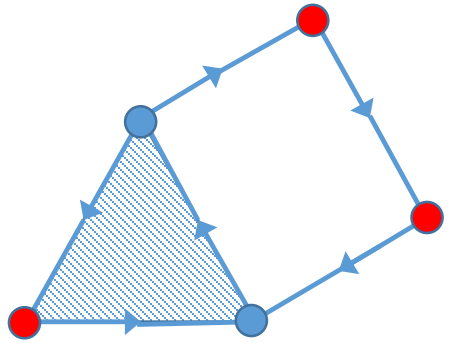
A Geometric Perspective on Feature Selection: Discrete Case



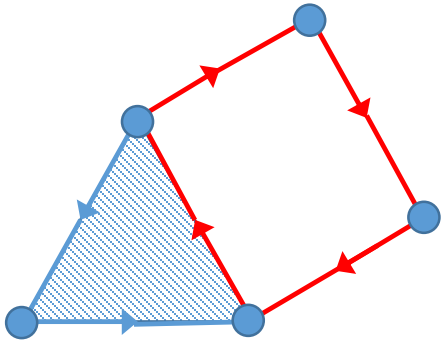
$$R_m\left(f^{(0)}\right) = \frac{1}{4}$$



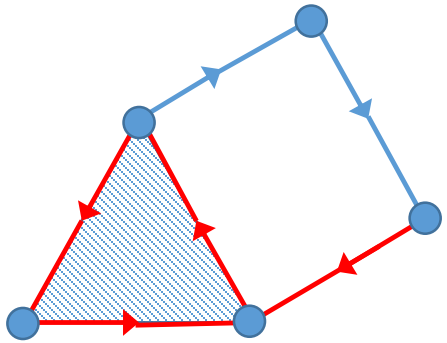
$$R_m\left(f^{(0)}\right) = \frac{3}{7}$$



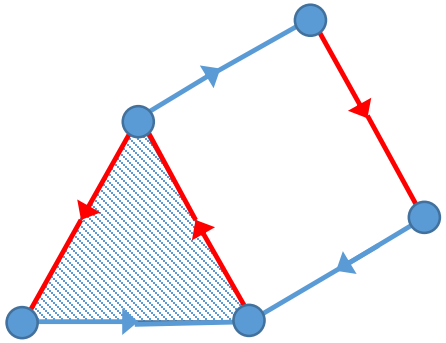
$$R_m\left(f^{(0)}\right) = \frac{2}{3}$$



$$R_m\left(f^{(1)}\right) = \frac{1}{9}$$



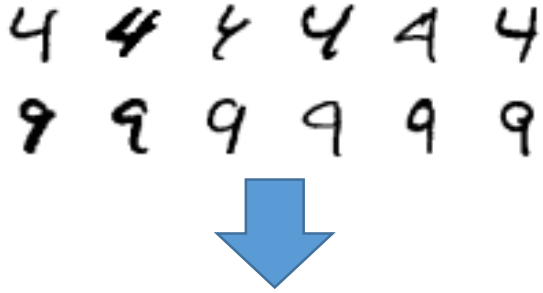
$$R_m\left(f^{(1)}\right) = 1$$



$$R_m\left(f^{(1)}\right) = 1$$

Benchmark: Gisette Dataset

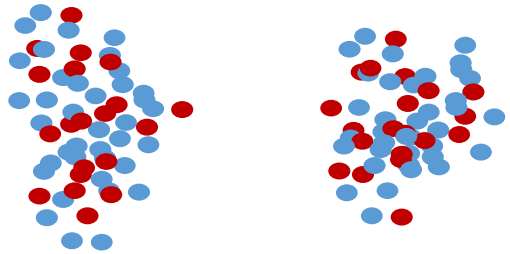
3,000 + 3,000 hand-written digits (28 x 28 pixels)



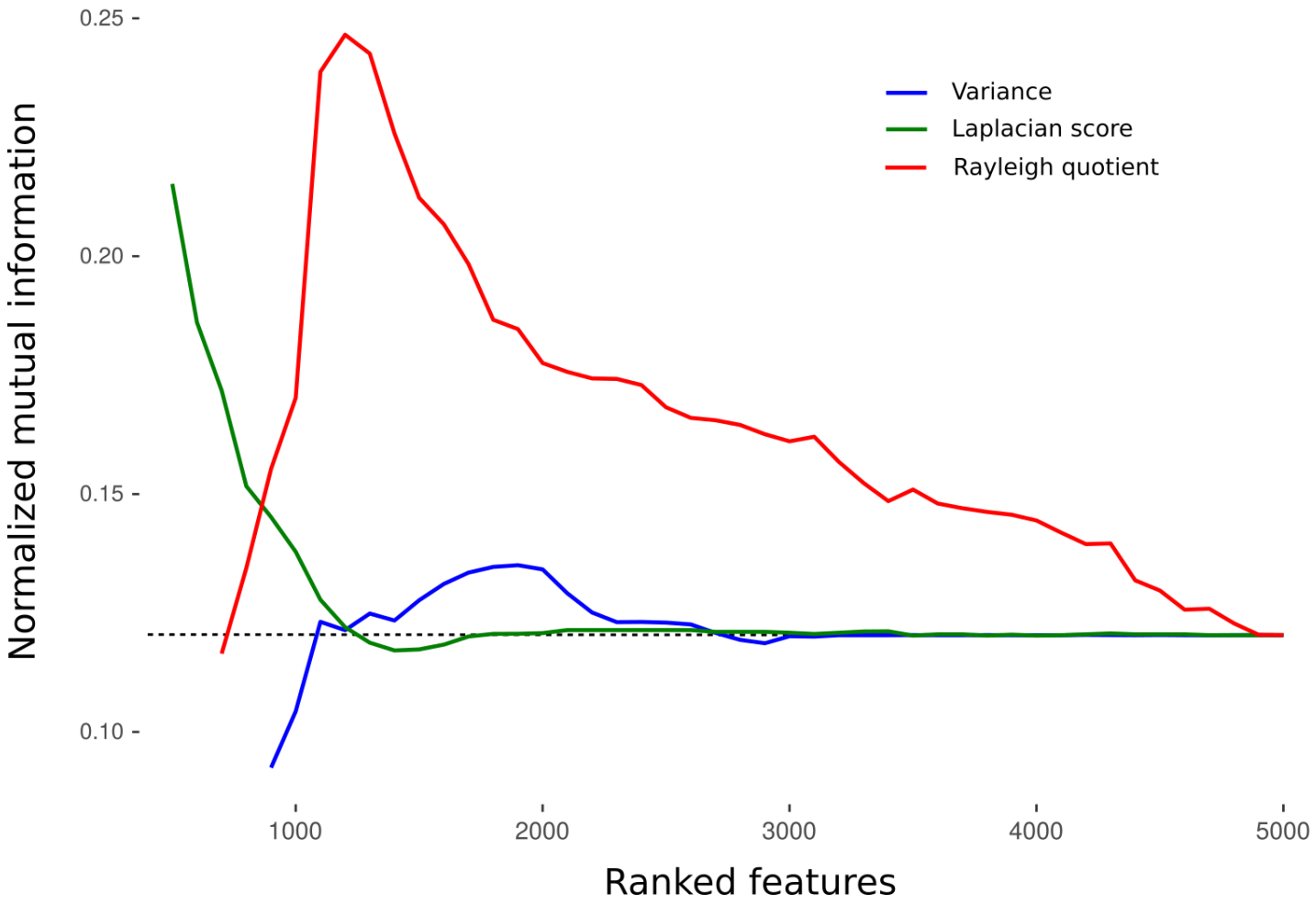
Gisette dataset
(2,500 + 2,500 features)

Guyon (NIPS 2003)

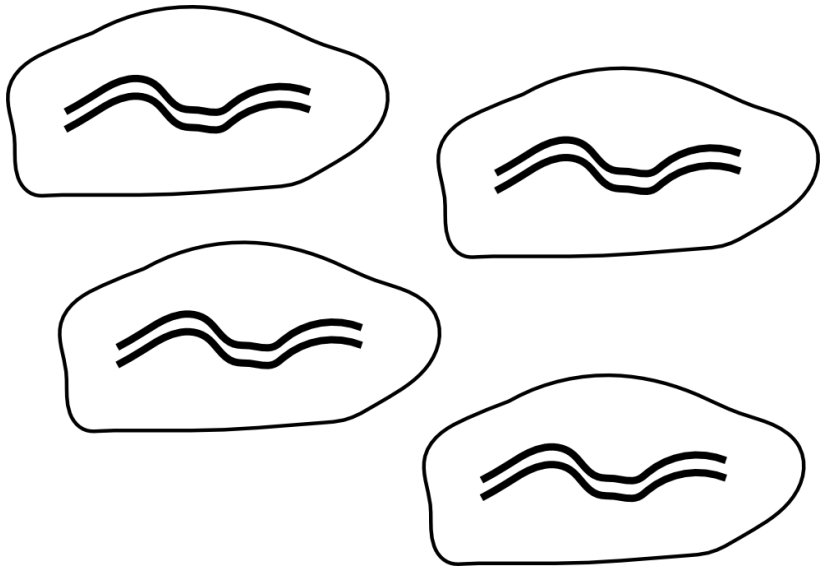
k-means clustering, with $k = 2$, using r top ranked features. Compare to actual labels using normalized mutual information.



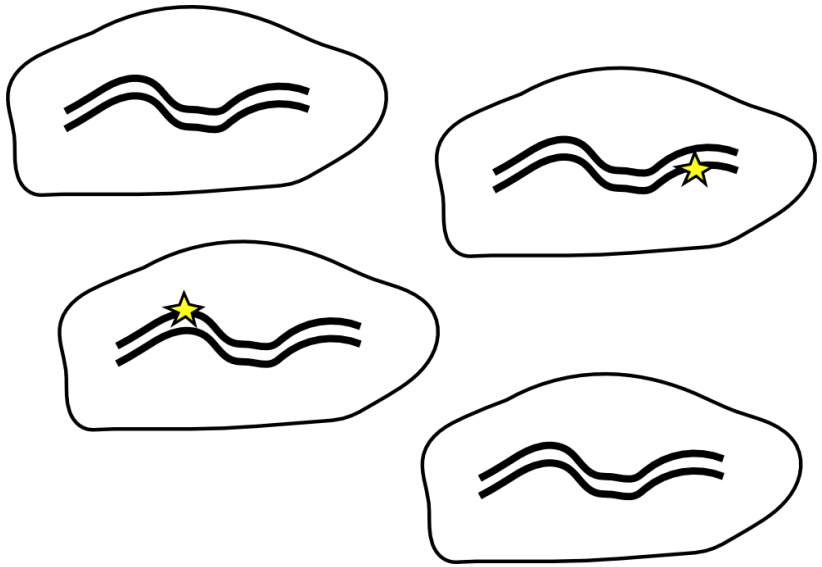
Benchmark: Gisette Dataset



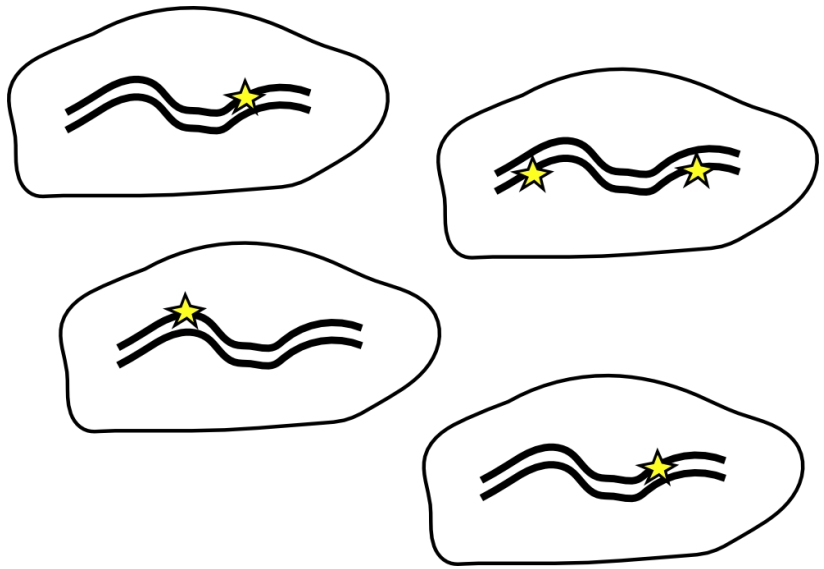
Example: Driver vs Passenger Mutations



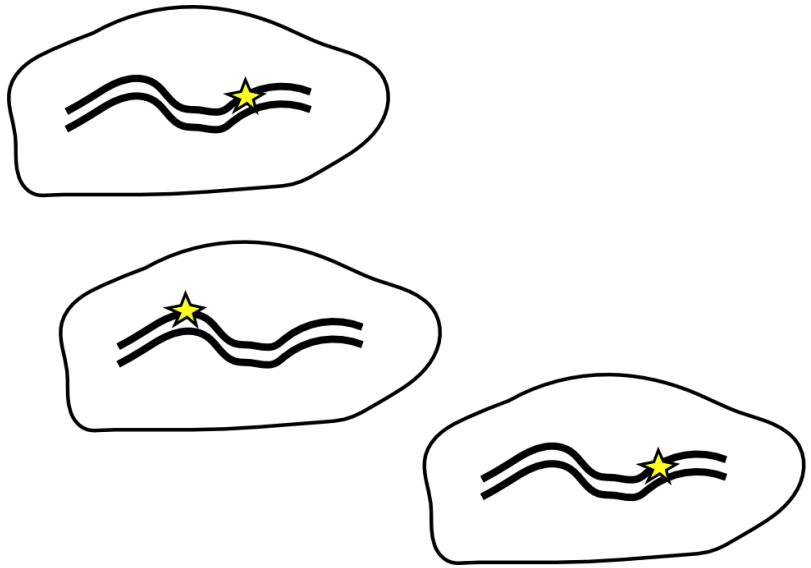
Example: Driver vs Passenger Mutations



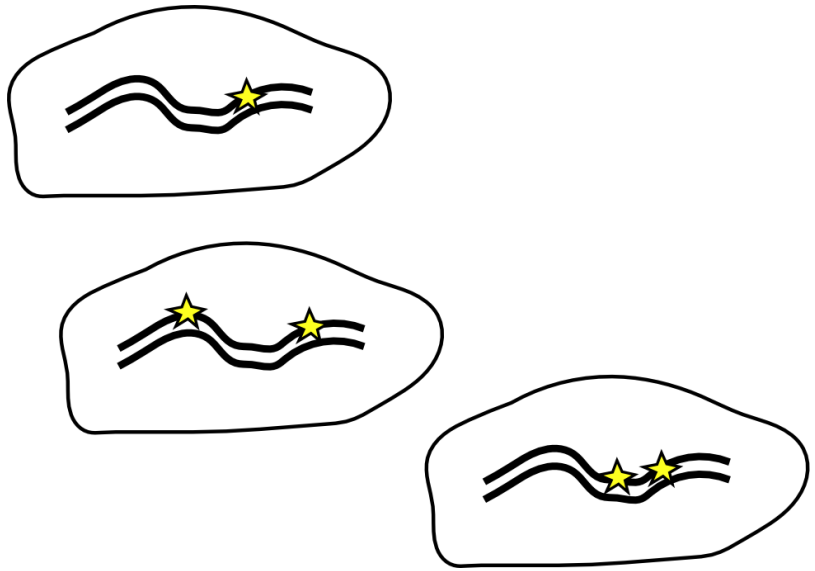
Example: Driver vs Passenger Mutations



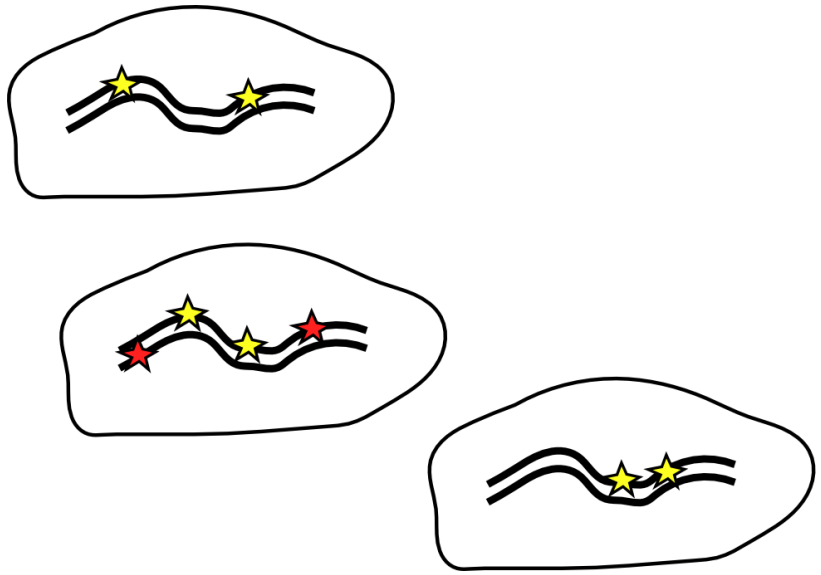
Example: Driver vs Passenger Mutations



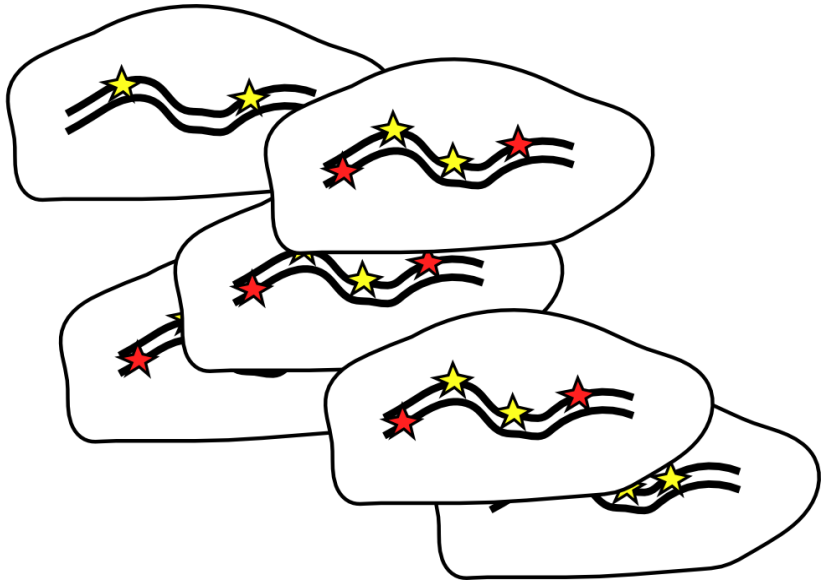
Example: Driver vs Passenger Mutations



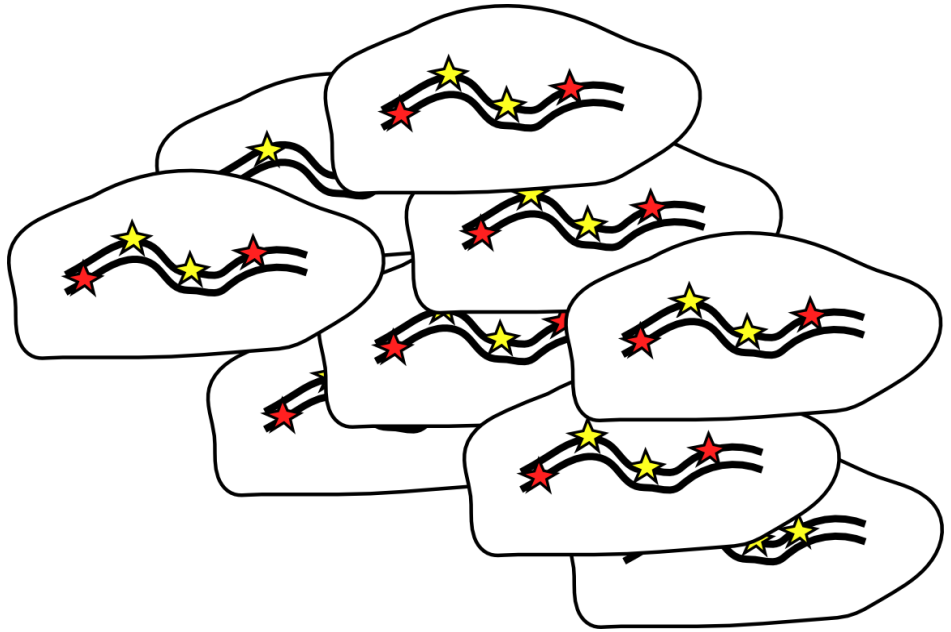
Example: Driver vs Passenger Mutations



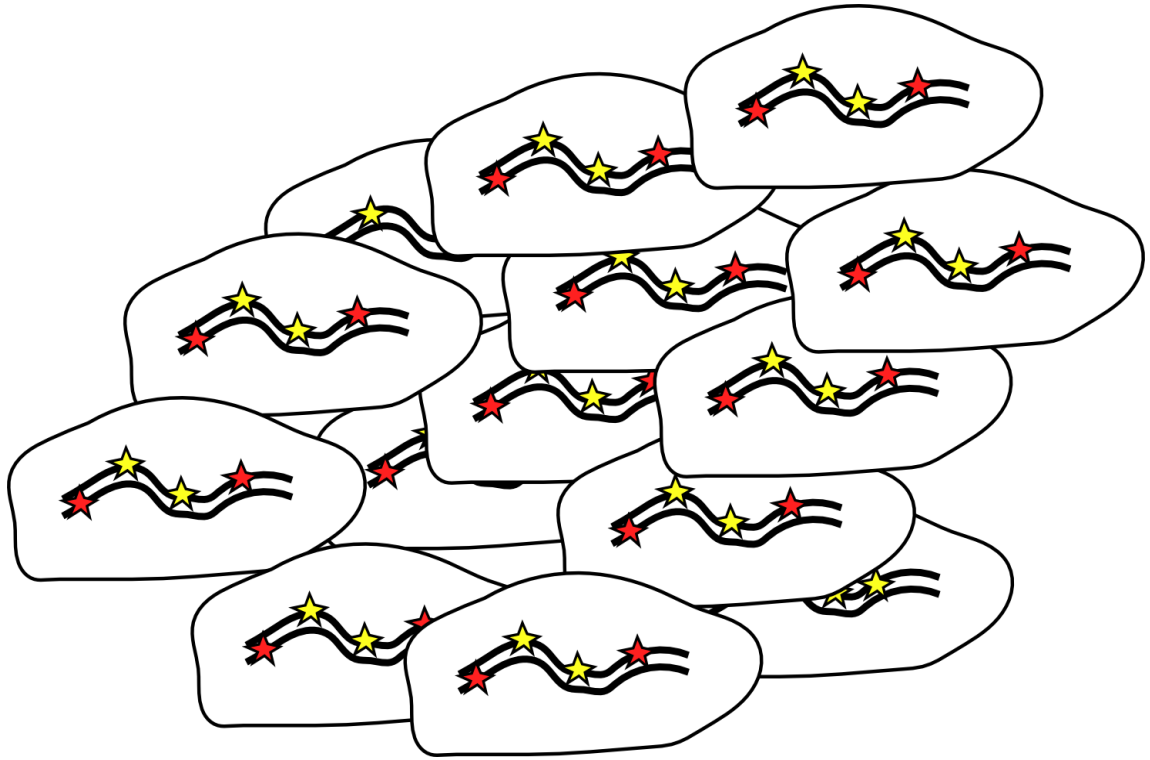
Example: Driver vs Passenger Mutations



Example: Driver vs Passenger Mutations



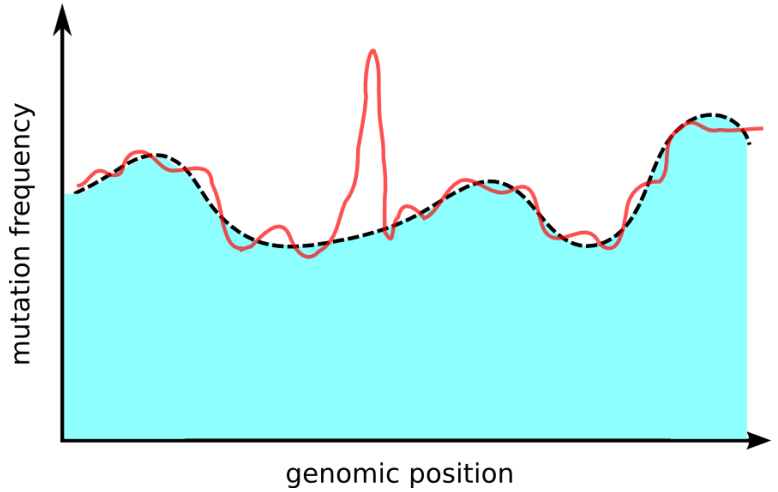
Example: Driver vs Passenger Mutations



The genome of a tumor harbors 100's to 1000's of somatic mutations, out of which only a handful are associated to the tumor progression

Example: Driver vs Passenger Mutations

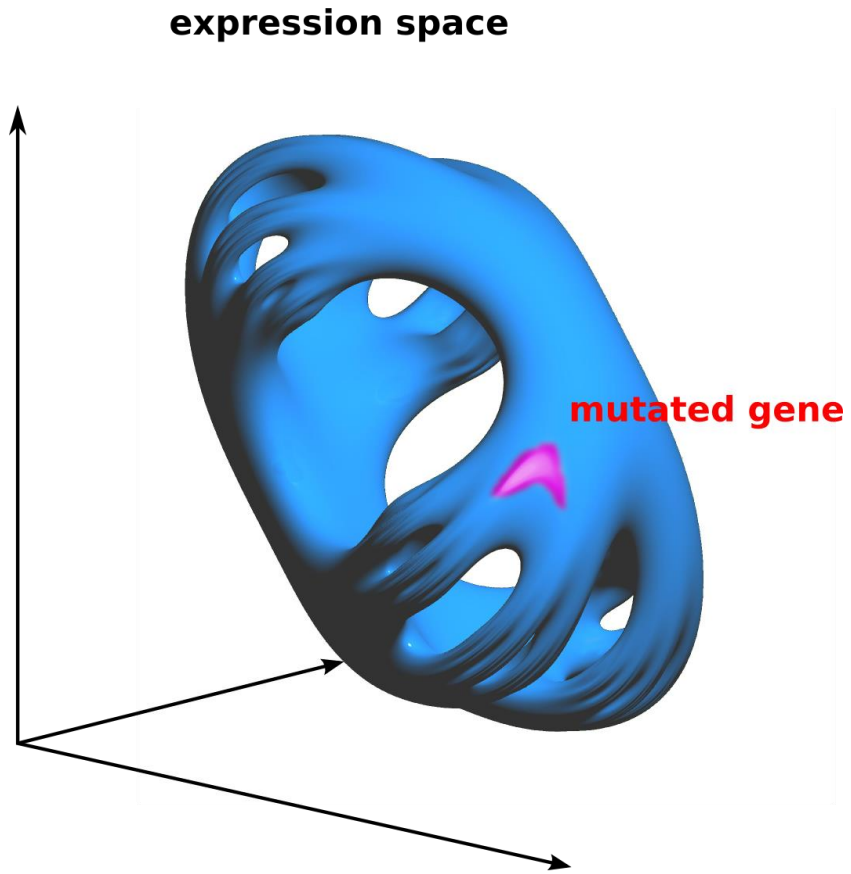
Signatures of positive selection based on recurrence across large cohorts of patients



Limited power to identify variants occurring at low frequencies (< 5%) or in hyper-mutated tumors

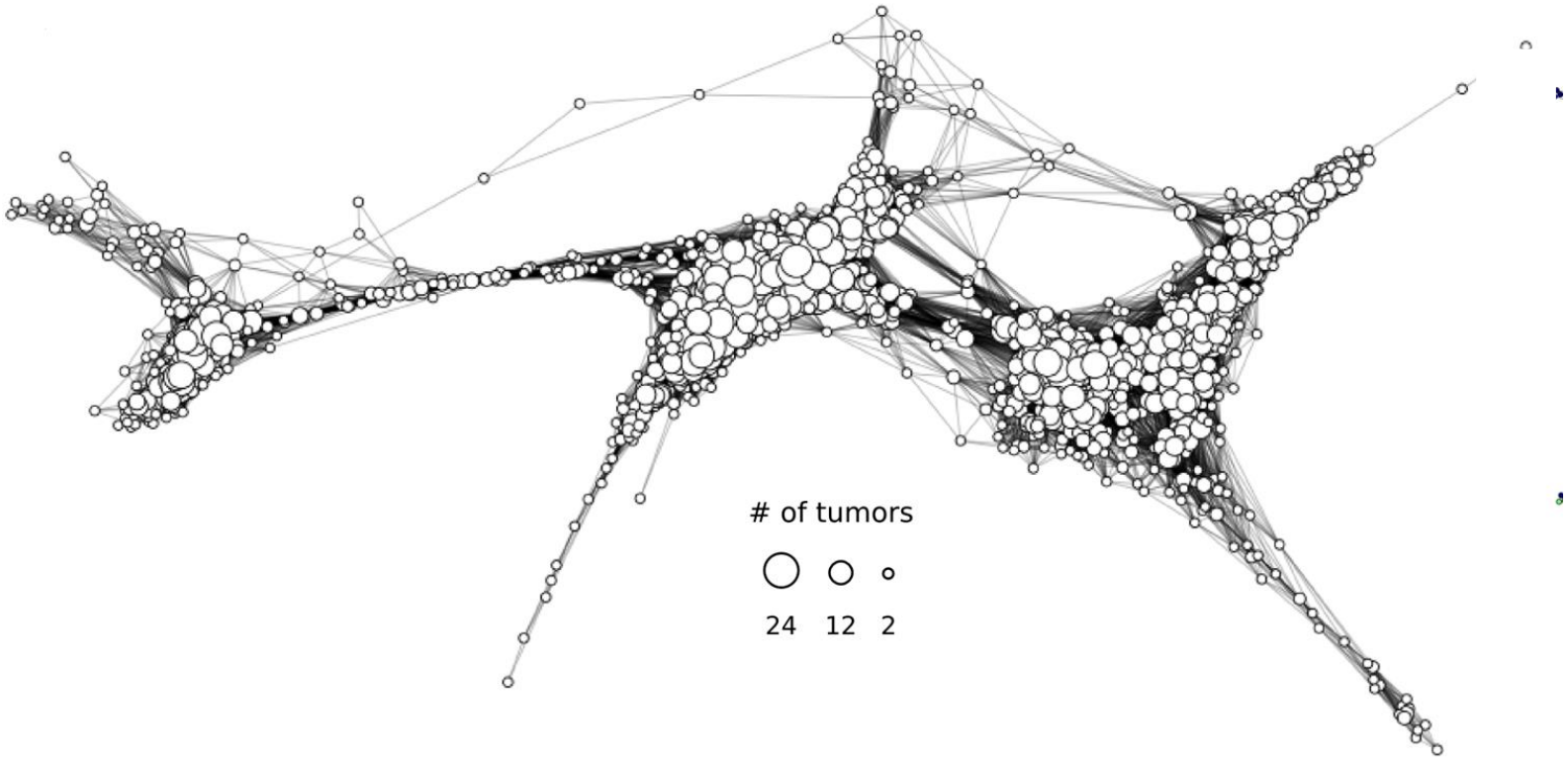
Example: Driver vs Passenger Mutations

Identification of driver mutations by localization in the gene expression space



Example: mutated cancer genes in LGG

Expression space of low-grade gliomas, RNA-seq data of 512 patients from TCGA

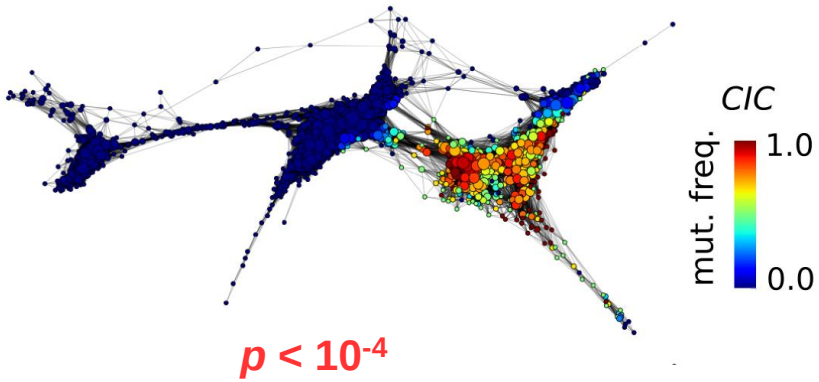


Example: mutated cancer genes in LGG

Asses significance of each mutated gene in expression space using $R_m(f_i^{(0)})$, where $f_i^{(0)}$ is the fraction of patients in each node with gene i mutated.

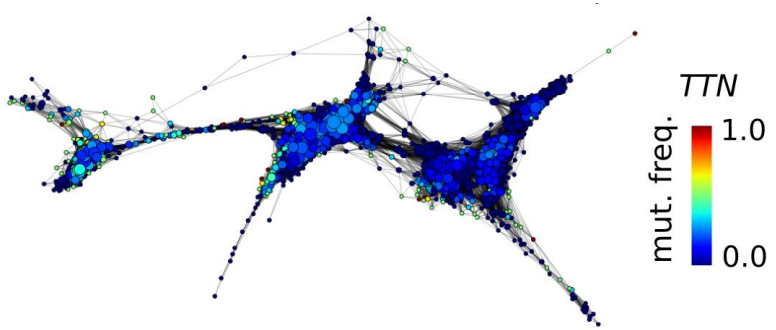
Null distribution for each gene by random permutation of patients id's.

mutated cancer gene



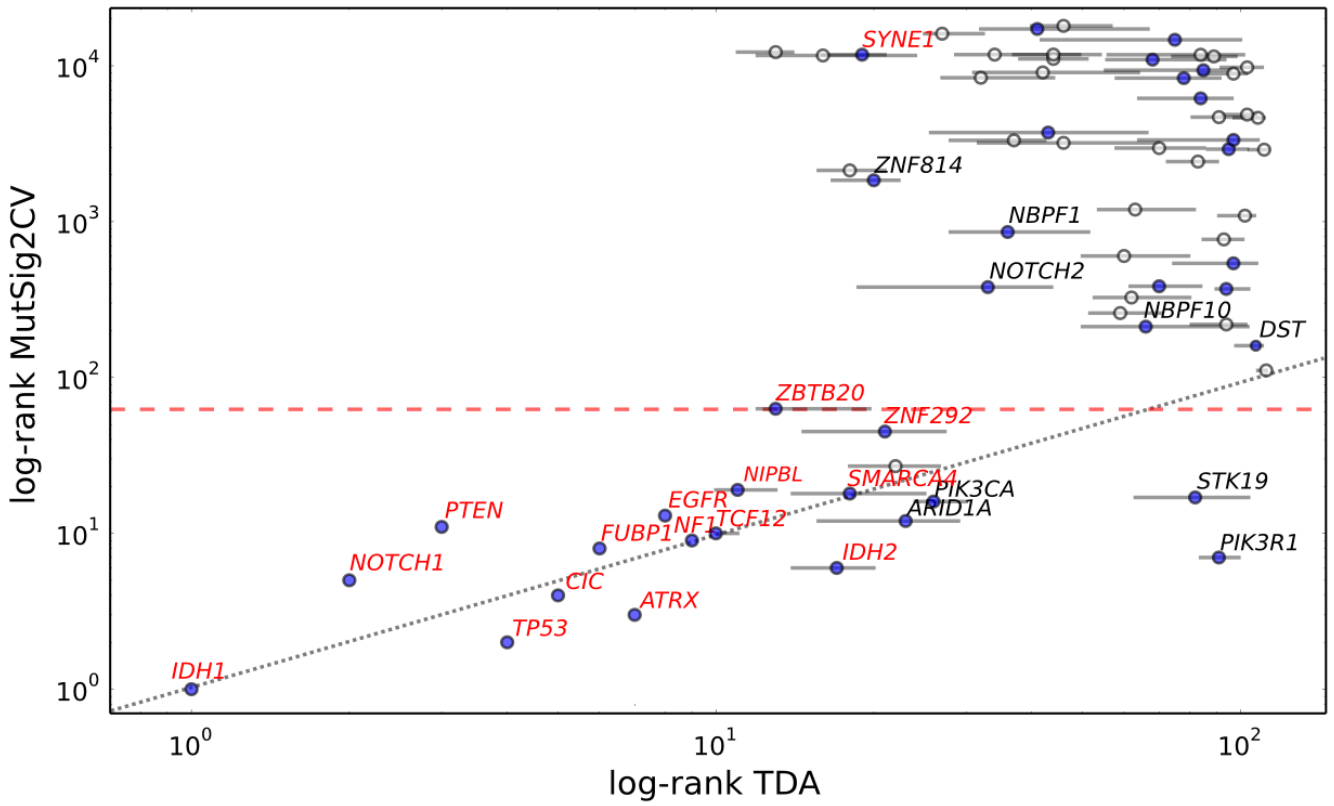
$p < 10^{-4}$

mutated passenger gene



$p = 0.4$

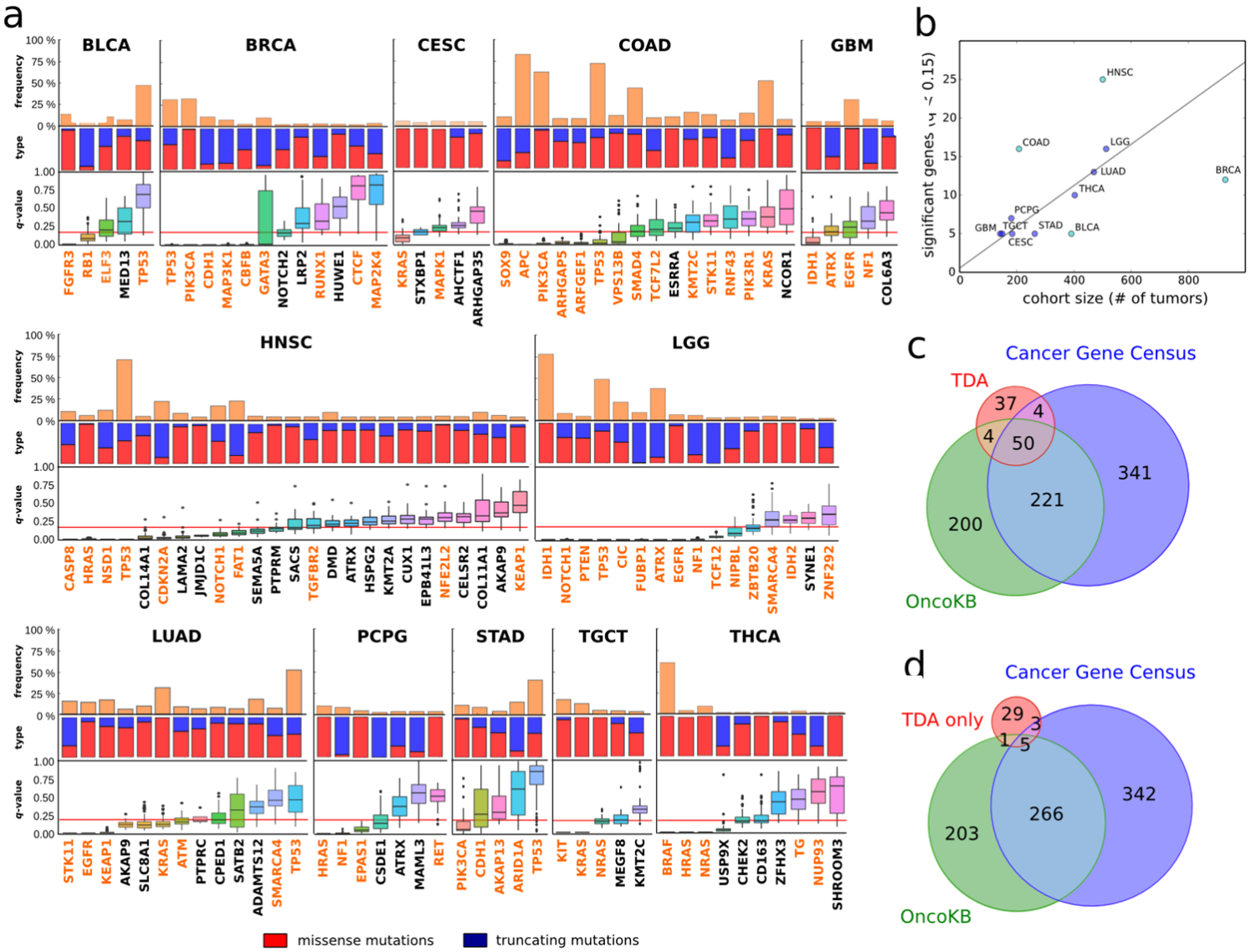
Example: mutated cancer genes in LGG



Good agreement with recurrence-based methods

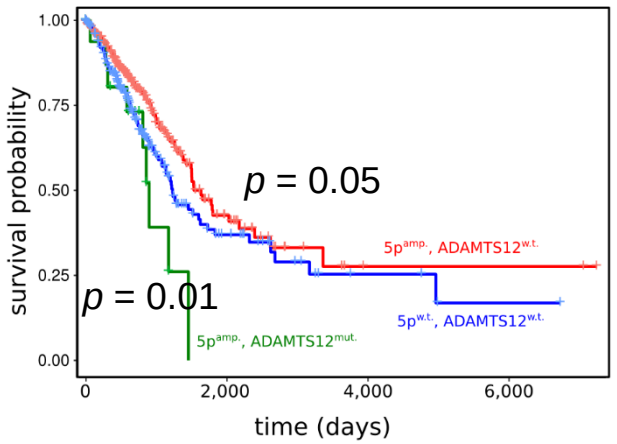
Mutated cancer genes in 12 tumor types

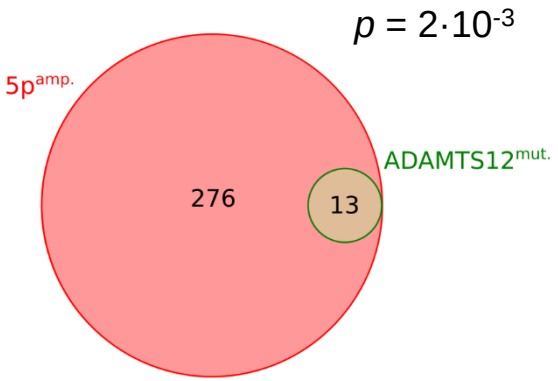
4,334 patients from 12 tumor types



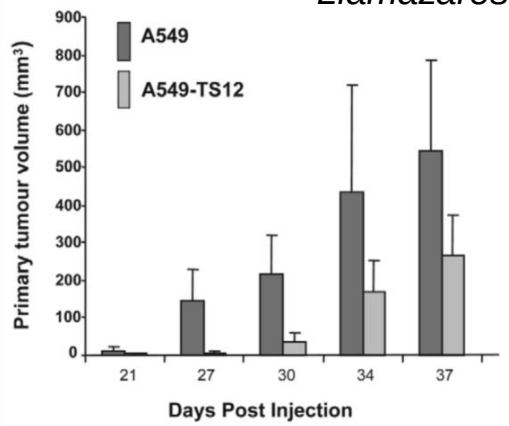
ADAMTS12 in LUAD

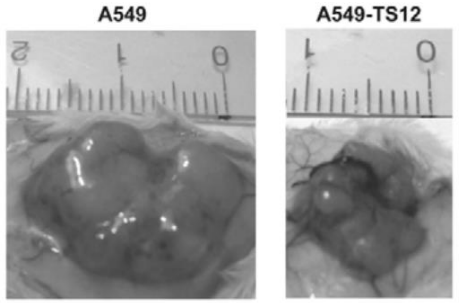
Metalloprotease that inhibits the Ras-MAPK pathway





Llamazares et al., J Cell Sci '07:





Summary

- Unsupervised feature selection can be formulated in geometric terms (if metric is available)

- General framework for co-homological feature selection generalizing ideas of network spectral analysis

- Application to genomic cancer data allows identification of new cancer genes

Acknowledgments

Camara Lab

- Rachael Aubin
- Kiya Govek
- Emma Troisi
- **Venkata Yamajala**

Funding:

Perelman School of Medicine University of Pennsylvania

Institute for Biomedical Informatics

STAND UP TO CANCER™

- Columbia University: **Raul Rabadan, Udi Rubin**
- IAS: **Arnold Levine**

Benchmark: Gisette Dataset