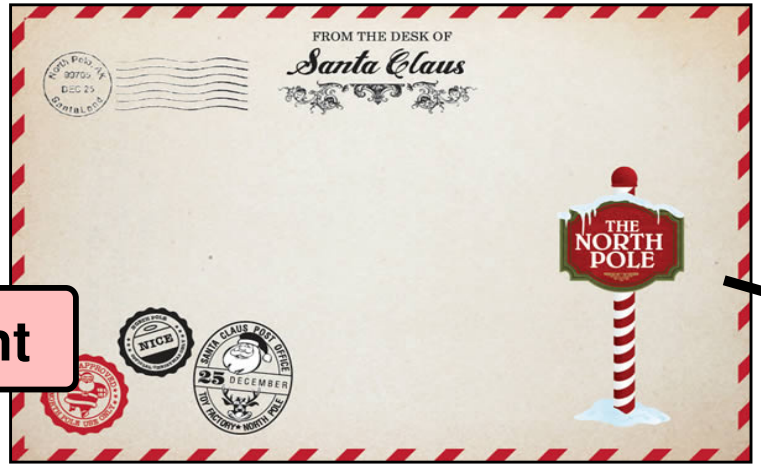
University of Basel

Archetype Analysis: a framework for selecting representative objects

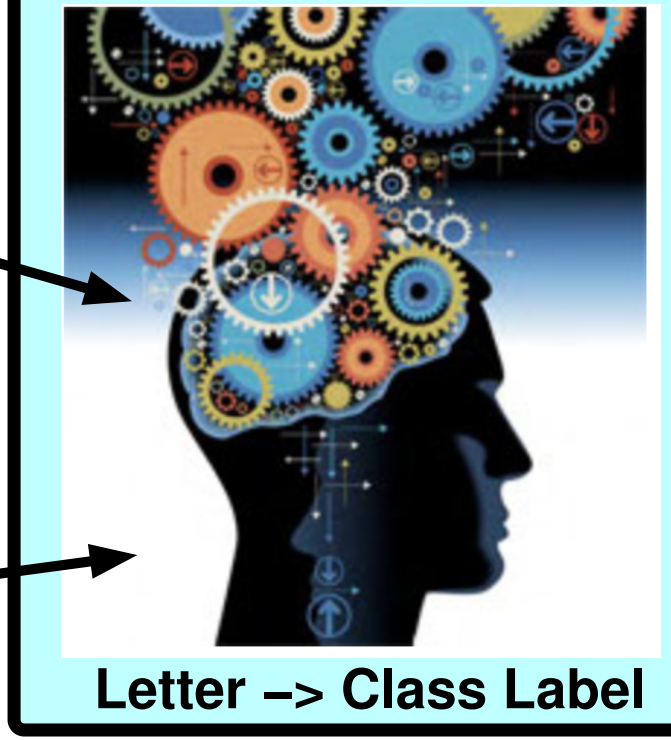
Volker Roth, Department of Mathematics and Computer Science,
University of Basel

Machine Learning: Supervised Setting



Important

spam



Statistical Learning Theory

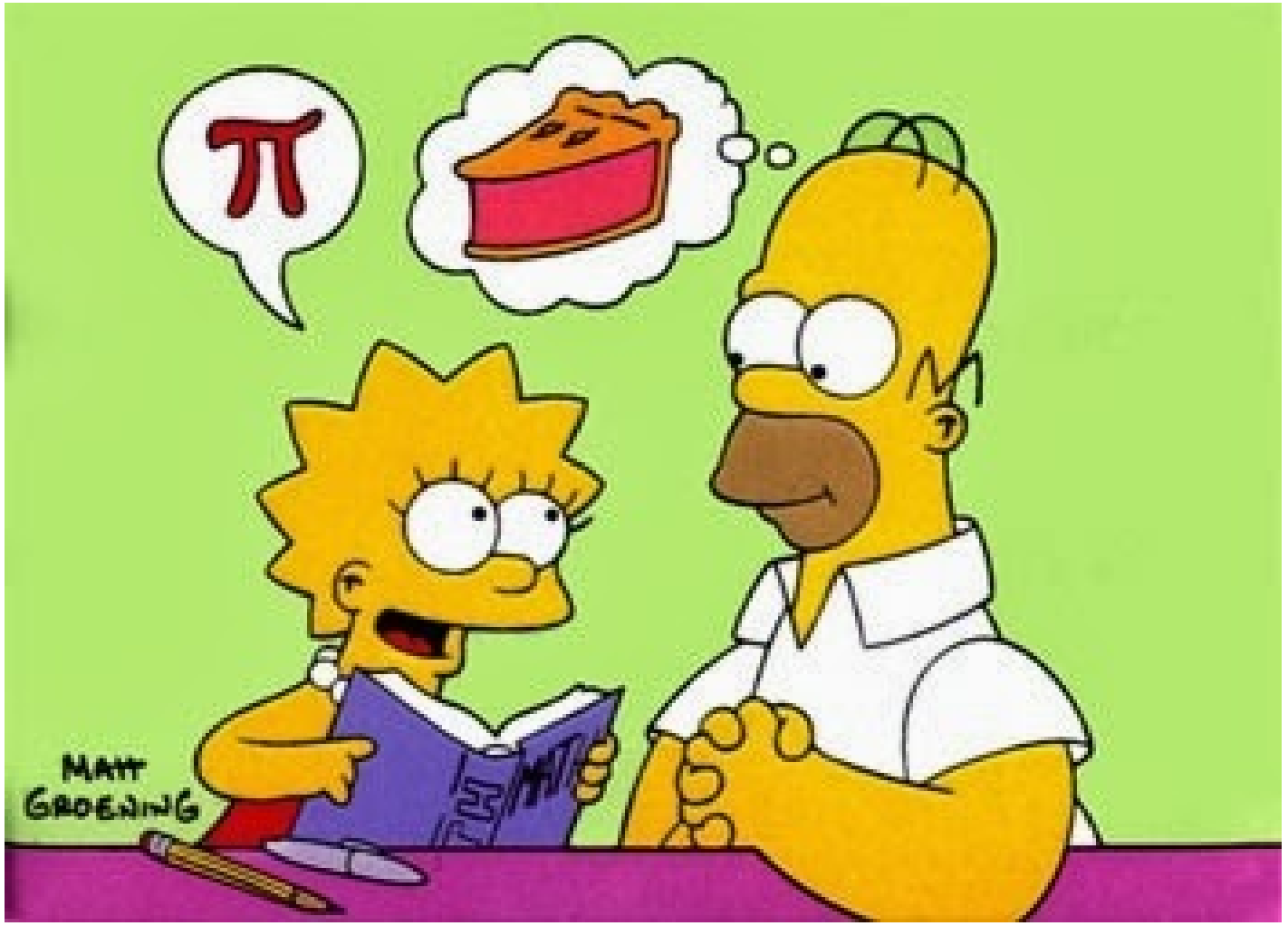
$$P\left(\sup_{f\in\mathcal{F}}\left|\hat{R}_n(f)-R(f)\right|>\varepsilon\right)\leq 8S(\mathcal{F},n)e^{-n\varepsilon^2/32}$$

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\hat{R}_n(f)-R(f)\right|\right]\leq 2\sqrt{\frac{\log S(\mathcal{F},n)+\log 2}{n}}$$

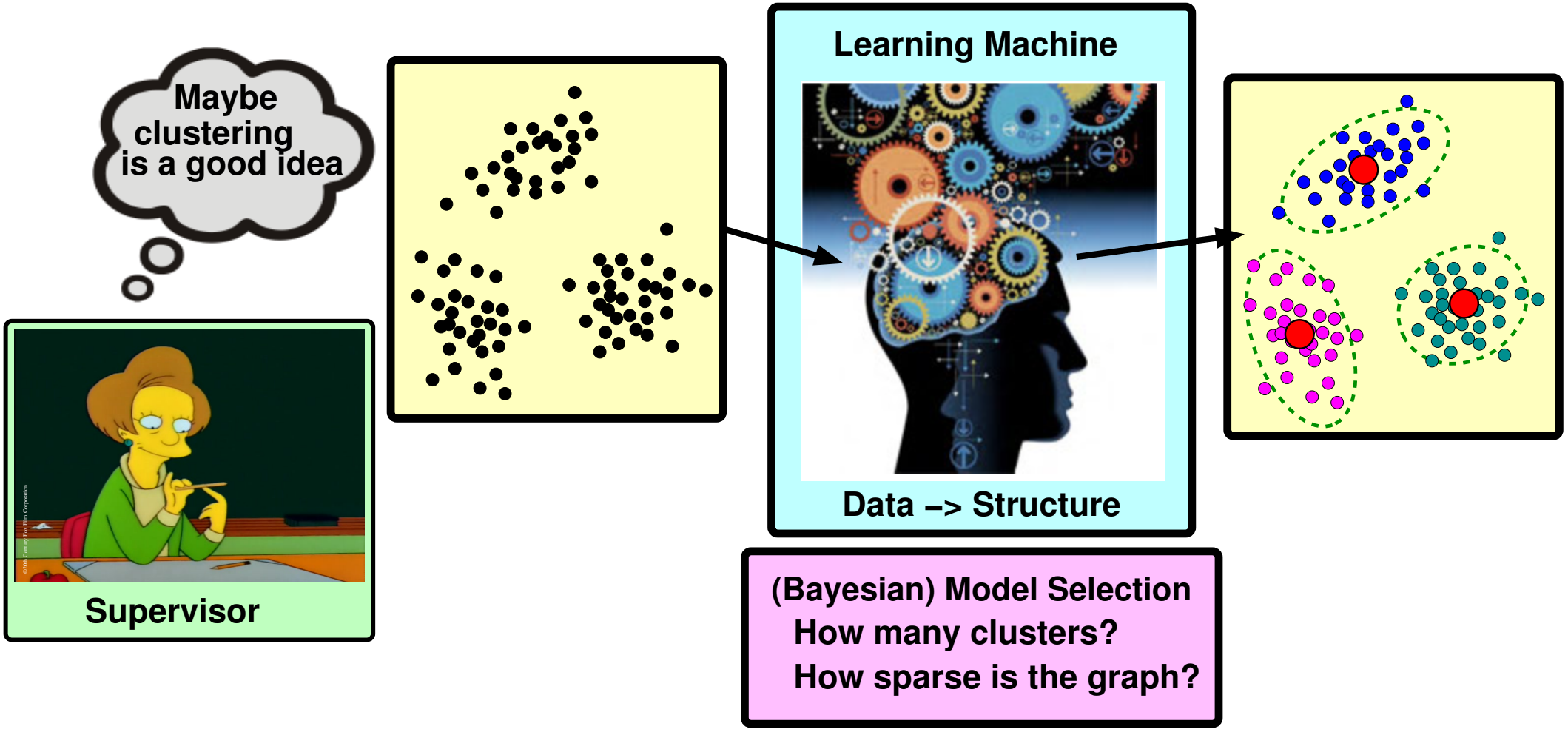
Machine Learning: Uncertain Labels

Sometimes even the best experts make errors.

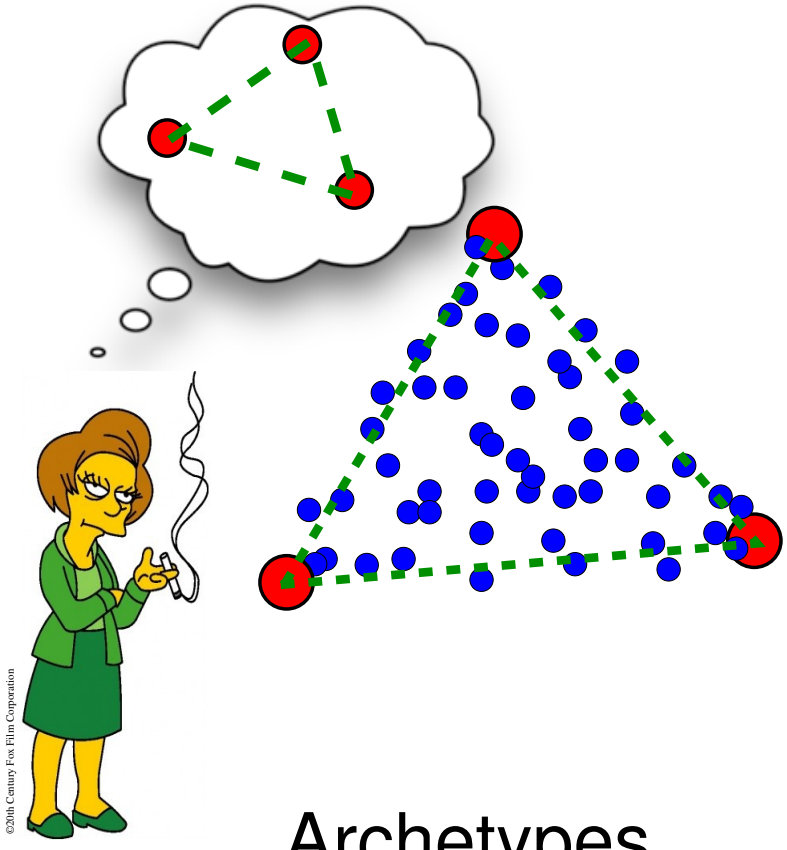
Labels might be uncertain or missing...



Machine Learning: Unsupervised Setting

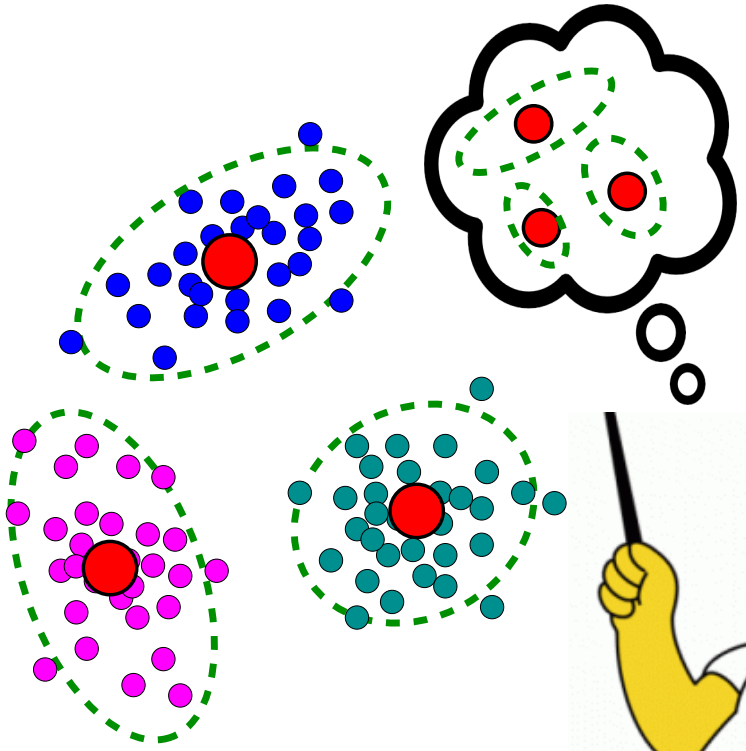


A Repeated Pattern: Search for Representative Observations



Archetypes

"most perfect possible form"



(Cluster–)Prototypes

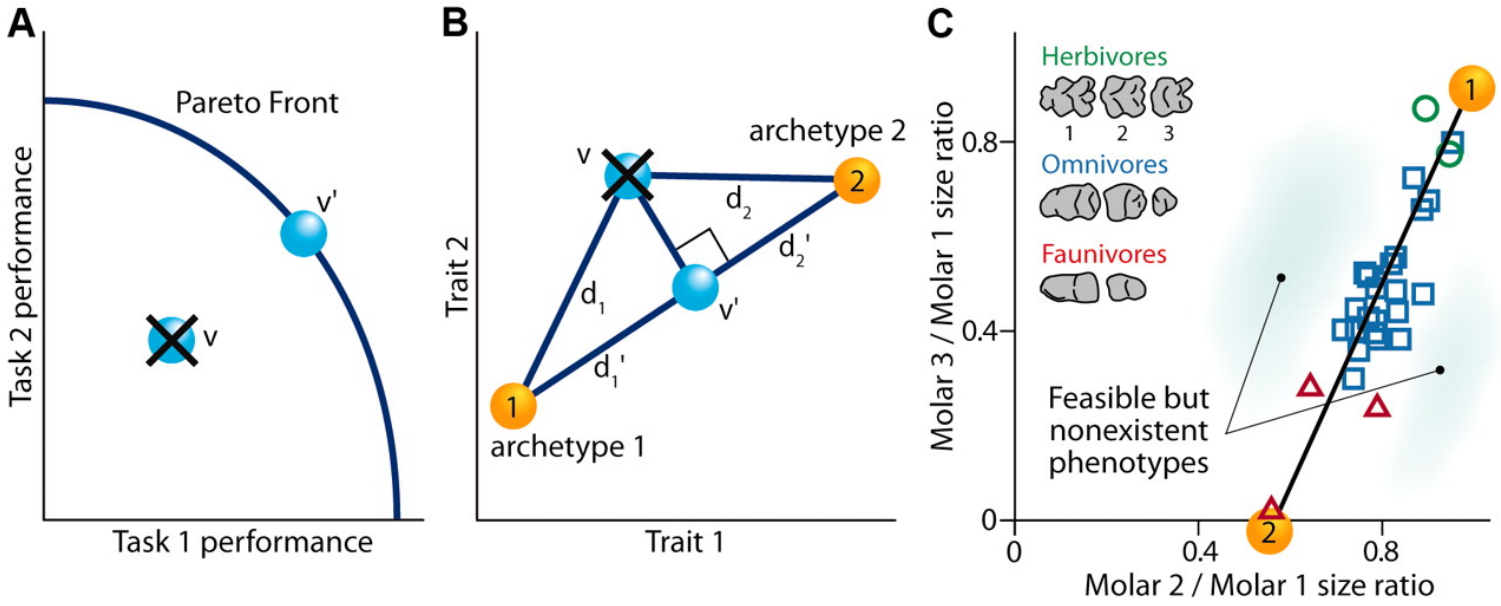
"an attempt at something"

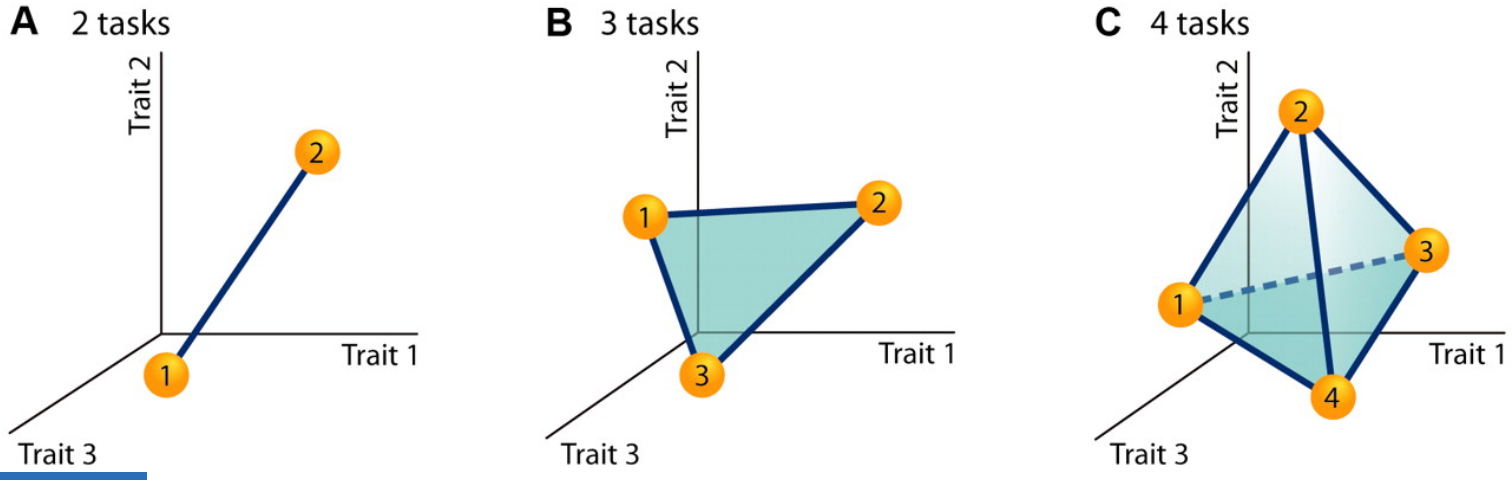
Archetype Analysis: Biological Motivation

Is there a theoretical foundation of the "archetype concept"?

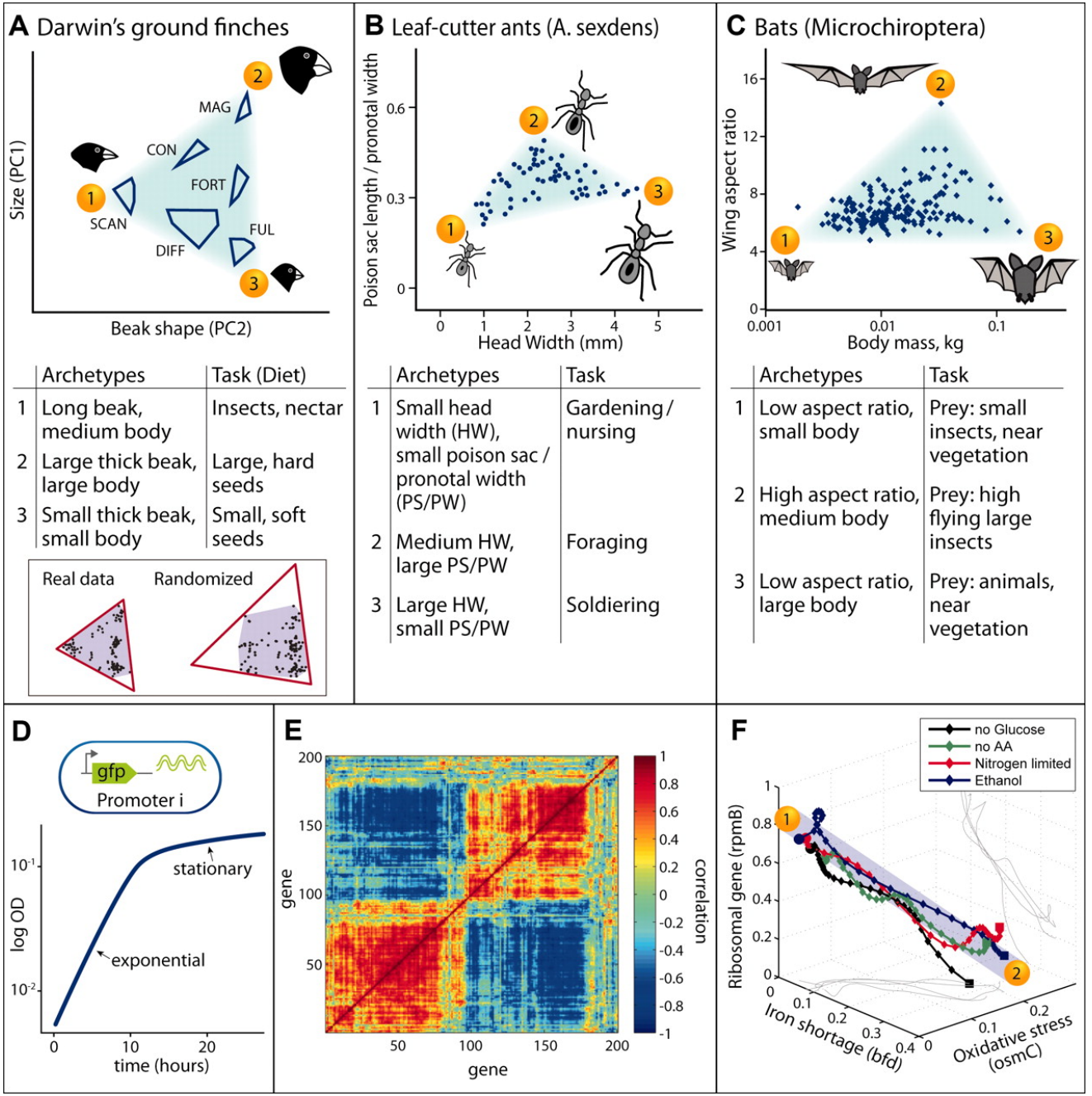
$\rightsquigarrow$ O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, U. Alon: **Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space**. Science, 2012

Archetypes and Evolutionary Trade-offs





Shoval et al.: Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space. Science, 2012



	Archetypes	Task (Diet)
1	Long beak, medium body	Insects, nectar
2	Large thick beak, large body	Large, hard seeds
3	Small thick beak, small body	Small, soft seeds

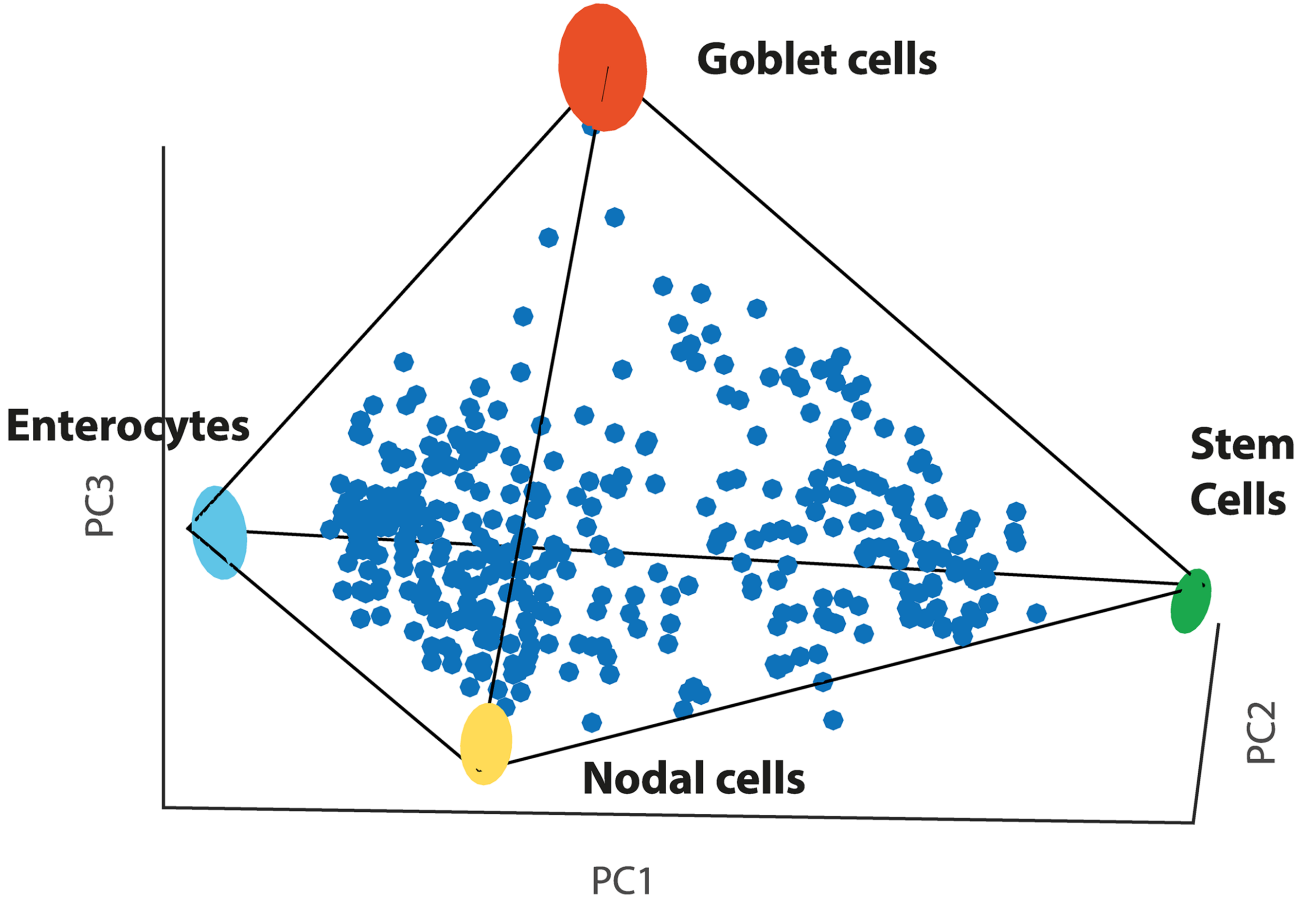
	Archetypes	Task
1	Small head width (HW), small poison sac / pronotal width (PS/PW)	Gardening / nursing
2	Medium HW, large PS/PW	Foraging
3	Large HW, small PS/PW	Soldiering

	Archetypes	Task
1	Low aspect ratio, small body	Prey: small insects, near vegetation
2	High aspect ratio, medium body	Prey: high flying large insects
3	Low aspect ratio, large body	Prey: animals, near vegetation

Shoval et al.: Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space. Science, 2012

Gene Expression Space

Human colon crypt cells fall in a tetrahedron in gene expression space.



The four vertices of this tetrahedron are each enriched with genes for a specific task related to stemness and early differentiation.

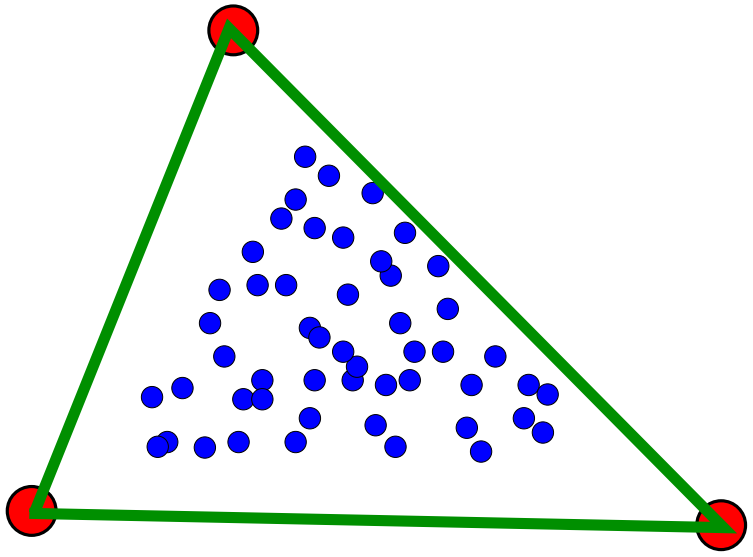
Korem et al. 2015

Computational Archetype Selection

Cutler &Breiman, *Archetypal Analysis*, Technometrics 1994.

- n observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^p$, as rows of data matrix $X \in \mathbb{R}^{n \times p}$
- Aim: find K archetypes $\Rightarrow Z \in \mathbb{R}^{K \times p}$; $K \ll n$ fixed.
- Observations are convex mixtures of archetypes

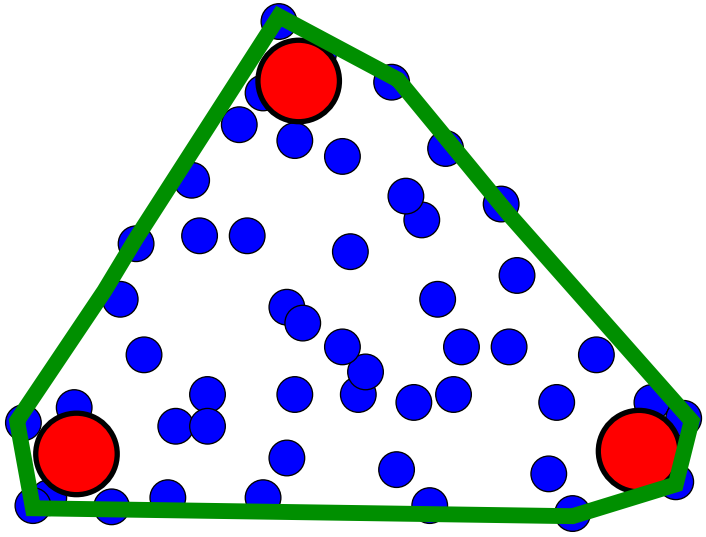
$$\boldsymbol{x}_i = Z^t \boldsymbol{a}_i + \epsilon_i, \quad a_{ij} \geq 0 \text{ and } \sum_{j=1}^{K} a_{ij} = 1.$$



Computational Archetype Selection

- Archetypes themselves are convex mixtures of observations:

$$\boldsymbol{z}_i = \sum_{j=1}^{n} b_{ij}\boldsymbol{x}_j, \quad \text{where } b_{ij} \geq 0 \quad \text{and} \quad \sum_{j=1}^{n} b_{ij} = 1$$



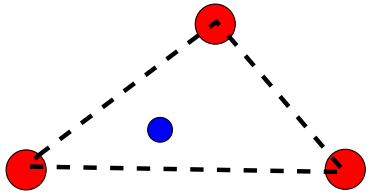
Archetypes approximate the convex hull!

- Constrained optimization problem involving two sets of coefficients $\{a_{ij}\}$ and $\{b_{ij}\}$: iteratively minimize sum of squares

$$(\hat{A}, \hat{B}) = \text{argmin}_{A,B} \|X - AZ\|^2 = \|X - ABX\|^2.$$

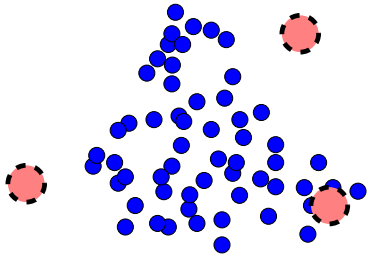
Basic Algorithm

1. Given K archetypes $Z_{K\times p}$, update the compositions $\boldsymbol{a}_i$ of the i-th object (a QP):



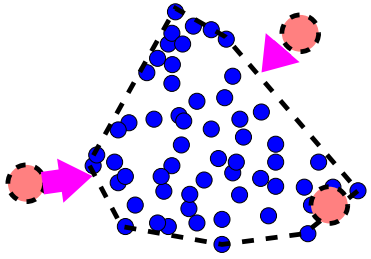
$$\boldsymbol{a}_i = \arg \min_{\boldsymbol{a}\in\mathbb{R}^p_+ : \boldsymbol{a}^t\mathbf{1}=1} \|\boldsymbol{x}_i - Z^t\boldsymbol{a}\|^2.$$

2. Given compositions $A_{n\times K}$, update archetypes by solving the least-squares problem



$$Z = \arg \min_{Z\in\mathbb{R}^{K\times p}} \|X - AZ\|^2.$$

3. Move ATs back to the convex hull (also a QP):

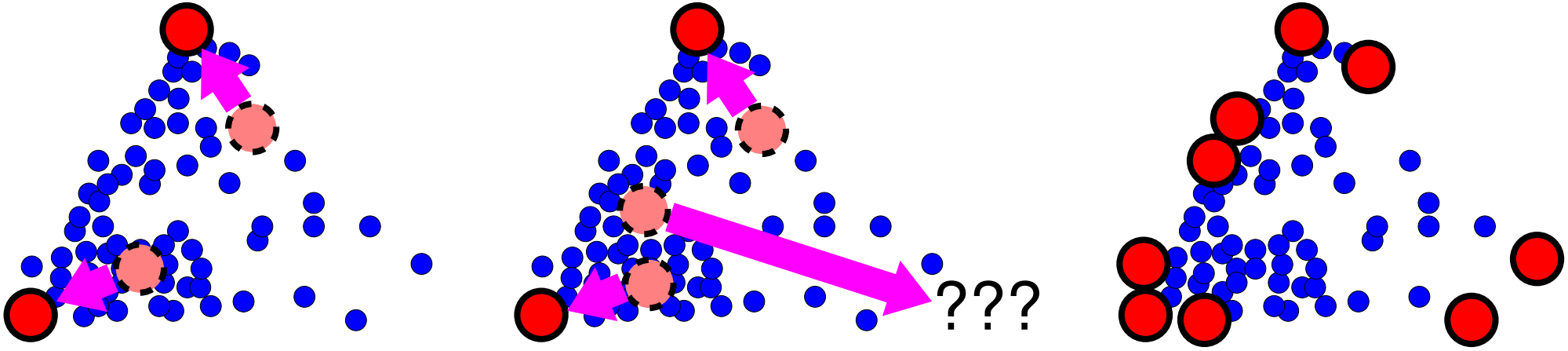


$$\boldsymbol{b}_k = \arg \min_{\boldsymbol{b}\in\mathbb{R}^n_+ : \boldsymbol{b}^t\mathbf{1}_n=1} \|\boldsymbol{z}_k - X^t\boldsymbol{b}\|^2.$$

Computational Archetype Selection

Problems:

- High computational complexity,
- Solution heavily depends on initialization of archetypes,
- Have to fix the number of archetypes K a priori.
- Assume that we have a suitable **representation** such that it makes sense to search for "triangles"...



Archetype Analysis: Model Selection

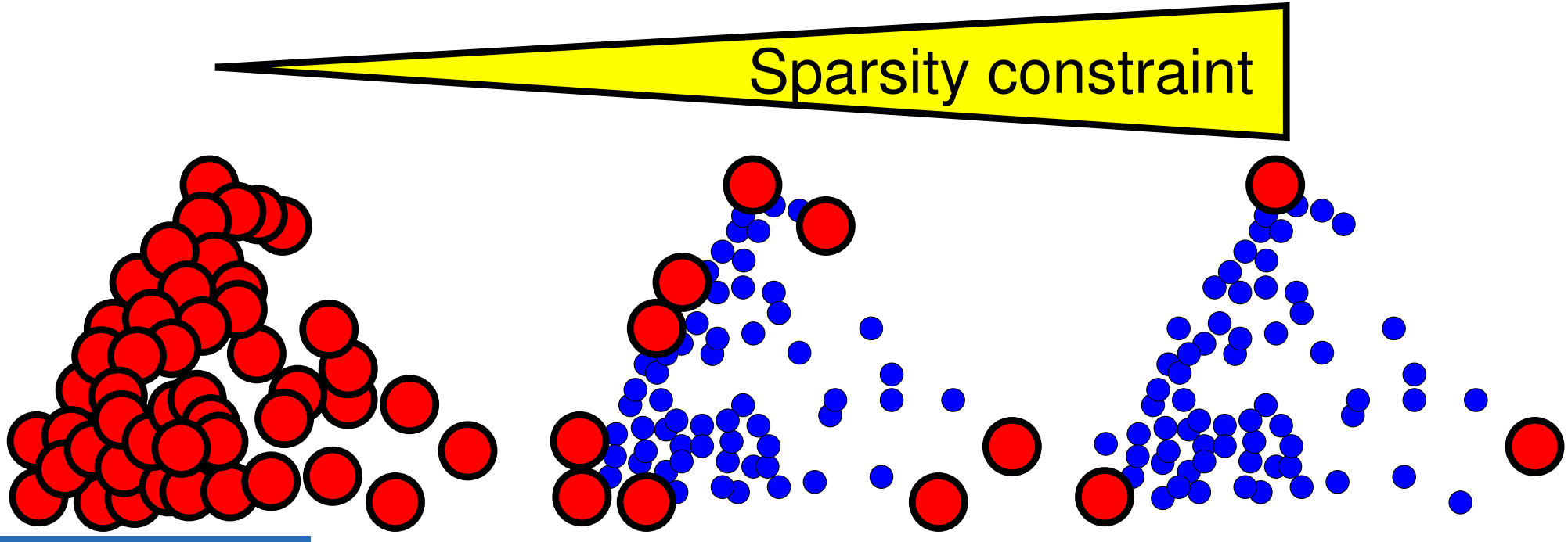
How many archetypes?

Model selection: how many archetypes?

- Solve for all numbers and choose "best"
  $\rightsquigarrow$ **Bayesian model comparison**.

- A clever way for looking at all nubers?

- Initialize all observations as archetypes, apply **sparse regression** methods to shrink most archetypes to zero.

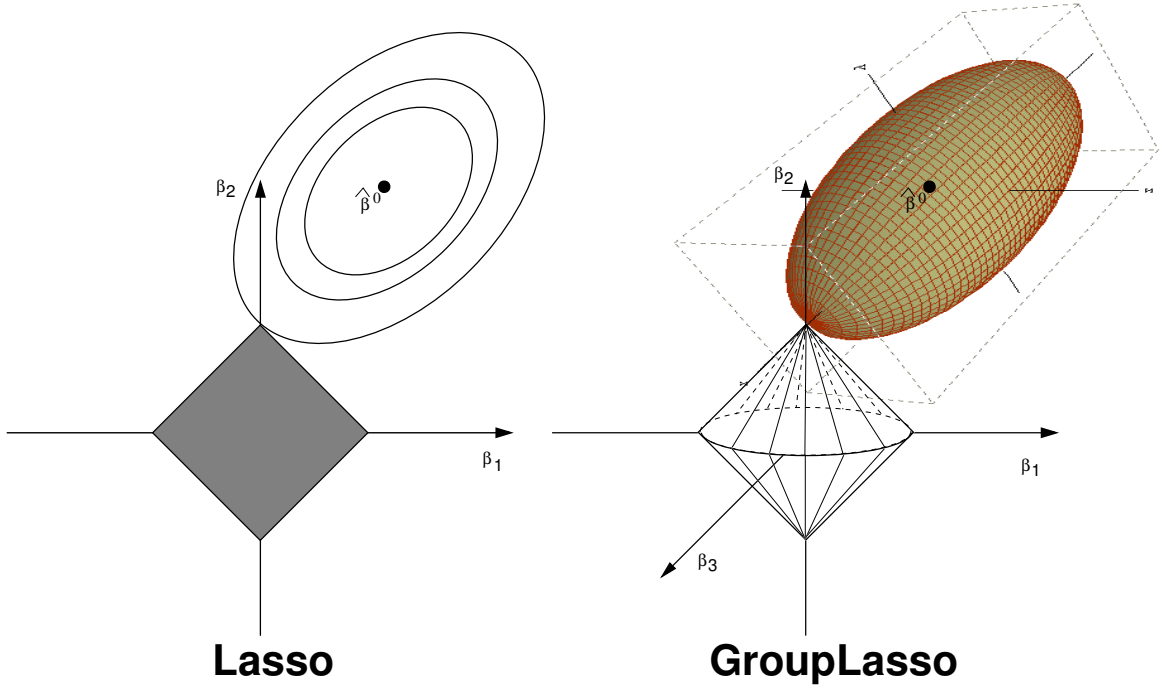


Sparse Regression via the Group-Lasso

Least-squares: $\min_{\boldsymbol{\beta}} \|X\boldsymbol{\beta} - \boldsymbol{y}\|^2$.

Treat regression coefficients β_i as **resources** needed for optimization.

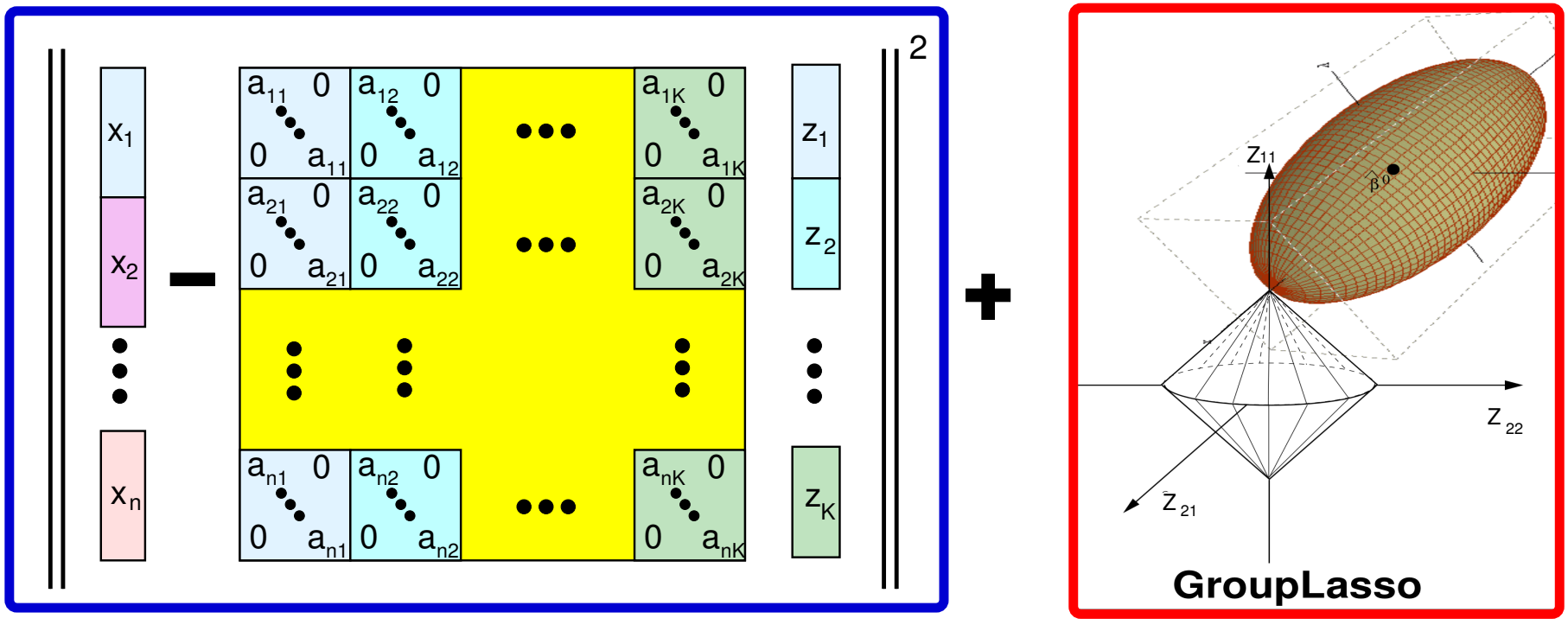
Idea: **limit resources** $\rightsquigarrow$ model must concentrate on **important features**.



$\|\boldsymbol{\beta}\|_1 \leq \kappa$ $\quad$ $\sum_{j=1}^{K} \|\boldsymbol{\beta}_j\|_2 \leq \kappa$

Automatic detection of archetypes

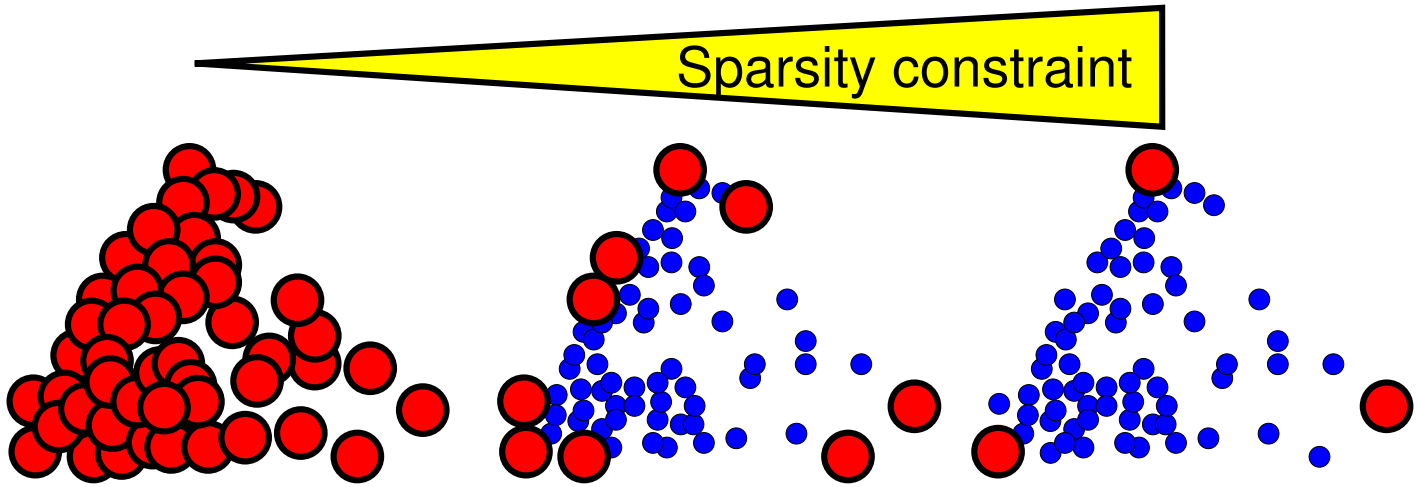
Rewrite $\|X - AZ\|$ as $\|\vec{x} - (A \otimes I_p)\vec{z}\|$, where $\vec{x}$ means column-wise vectorization of X and use $\ell_{1,2}$ **block-norm constraint**: $\sum_{j=1}^{K} \|\boldsymbol{z}_j\| \leq \kappa$.



GroupLasso constraint will shrink some archetypes $\boldsymbol{z}_j$ to zero, depending on constraint value κ.

A solution for the model selection problem

- Use GL algorithm that approximates the whole solution path on a fine grid of κ values



- For every κ-value, use BIC score

$$\mathrm{BIC}(\hat{\mu}) = \frac{\|\vec{x} - \hat{\mu}\|^2}{n\sigma^2} + \frac{\log(n)}{n} \cdot \widehat{df}(\hat{\mu}),$$

and select the best scoring model.

Archetype Analysis: Algorithms

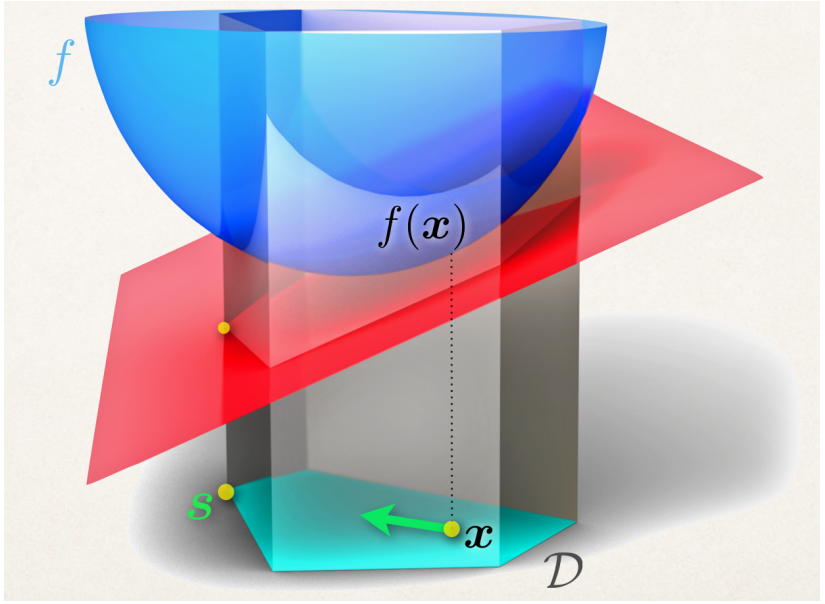
Will it work for huge datasets?

Efficient algorithms

(Bauckhage et al., 2015):
Solve step 1 and combined (2,3)-step with Frank-Wolfe algorithm
Idea: use **linear optimization oracle** over constraint set.

- Linear minimization oracle

$$\begin{aligned} \Delta(\boldsymbol{x}) \quad &= \quad \operatorname*{argmin}_{\boldsymbol{z}} \langle \boldsymbol{x}, \boldsymbol{z} \rangle, \\ &\quad \text{s.t.} \quad g(\boldsymbol{z}) \leq \kappa \end{aligned}$$



M. Jaggi, 2015

- Update $\boldsymbol{x} \leftarrow (1-\gamma)\boldsymbol{x} + \gamma\Delta(\nabla_f(\boldsymbol{x}))$

- Decrease γ

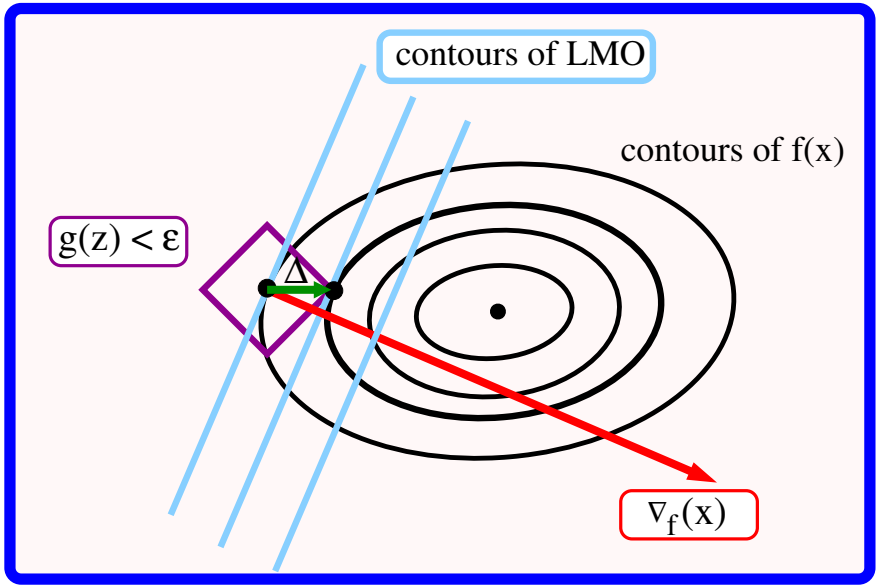
Pros: Highly efficient for one fixed value of κ
Cons: Not efficient for the whole solution path
$\rightsquigarrow$ **model selction trick does not work well**.

Alternative: Forward stagewise

Technically similar to Frank-Wolfe:

- Linear minimization oracle (LMO)

$$\begin{aligned}\Delta(\boldsymbol{x}) &= \operatorname*{argmin}_{\boldsymbol{z}} \langle \boldsymbol{x}, \boldsymbol{z} \rangle, \\ &\text{s.t.} \quad g(\boldsymbol{z}) \leq \epsilon \ll \kappa\end{aligned}$$



- Update $\boldsymbol{x} \leftarrow \boldsymbol{x} + \Delta(\nabla_f(\boldsymbol{x}))$

...but very different behaviour:

- incremental path following behaviour
  $\rightsquigarrow$ efficient for computing the whole solution path
- built-in monotonicity "regularization" $\rightsquigarrow$ very stable,
  can be extended to non-convex "norms" for increased sparsity.

Forward stagewise

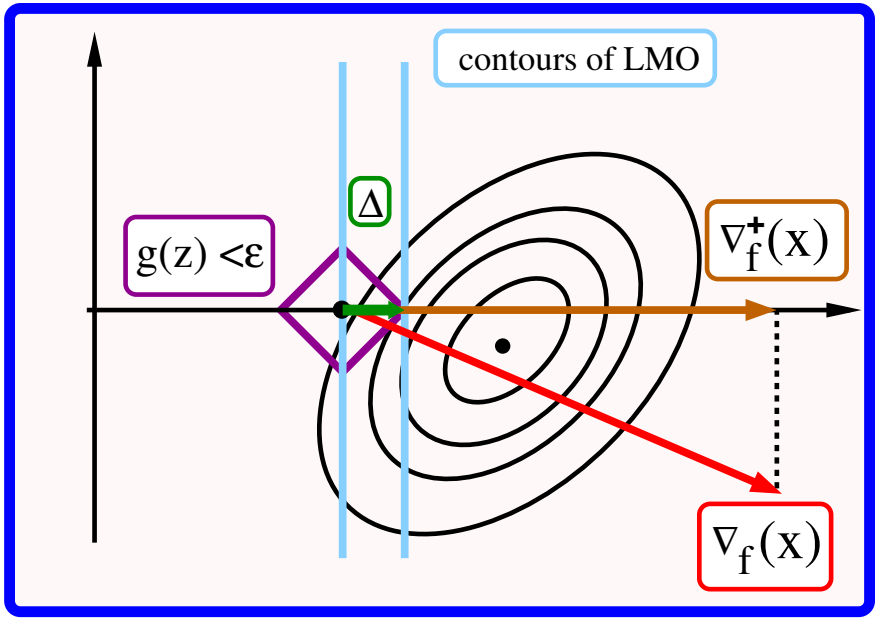
Consider step 1 in AT analysis:
Given K archetypes $Z_{K\times p}$, update the compositions $\boldsymbol{a}_i$ of the i-th object $\boldsymbol{x}_i$ under convexity constraints:

$$\boldsymbol{a}_i = \arg \min_{\boldsymbol{a}\in\mathbb{R}^p_+ : \boldsymbol{a}^t\mathbf{1}=1} \|\boldsymbol{x}_i - Z^t\boldsymbol{a}\|^2.$$

This is a non-negative lasso estimate

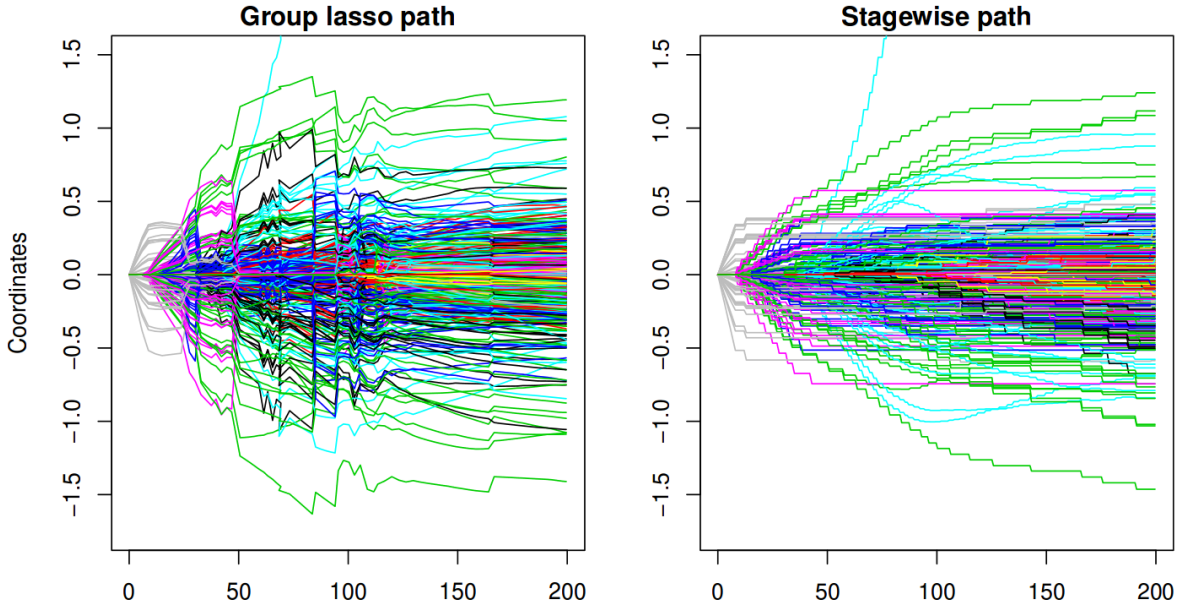
$$\begin{aligned} \min_{\boldsymbol{a}} \quad & \|\boldsymbol{x}_i - Z^t\boldsymbol{a}\|^2 \\ \text{s.t.} \quad & \|\boldsymbol{a}\|_1 = 1,\ a_j \geq 0. \end{aligned}$$

Use projected gradient:
$\boldsymbol{a} \leftarrow \boldsymbol{a} + \Delta(\nabla_f^+(\boldsymbol{a}))$



Forward stagewise

- Useful if LMO can be computed easily: all norms, block-norms etc.
- NN lasso: LMO is intersection of simplex and linear function
  $\rightsquigarrow$ find best vertex l, update l-th component of $\boldsymbol{a}$ as $a_l \leftarrow \boldsymbol{a}_l + \epsilon$
  $\rightsquigarrow$ $\boldsymbol{a}$ is monotone increasing in every component.
- Simple and efficient: iterate until $\|\boldsymbol{a}\|_1 = 1 \rightsquigarrow 1/\epsilon$ iterations needed.
- Conceptually the same behaviour for group-lasso estimate in step 2.



R. Tibshirani, 2015

Further algorithmic tricks

- Pre-select candidate points on convex hull:
  - Points on **convex hull in any linear projection** are also on the "full" convex hull.
  - Convex hull computation very efficient in 2D: $O(n \log n)$
  - Randomly project data to planes, compute points on convex hull, aggregate.

- Alternative to random projections: use **pairwise PCA projections**.
  - PCA is a good preprocessing step anyway, since convex polygon with K vertices can be embedded in a space with $< K$ dimensions.

- We can easily solve AT problems with > 100 millions of objects.

Archetype Analysis: Data representation

How to encode the data such that we can see "triangles"?

Representation issues

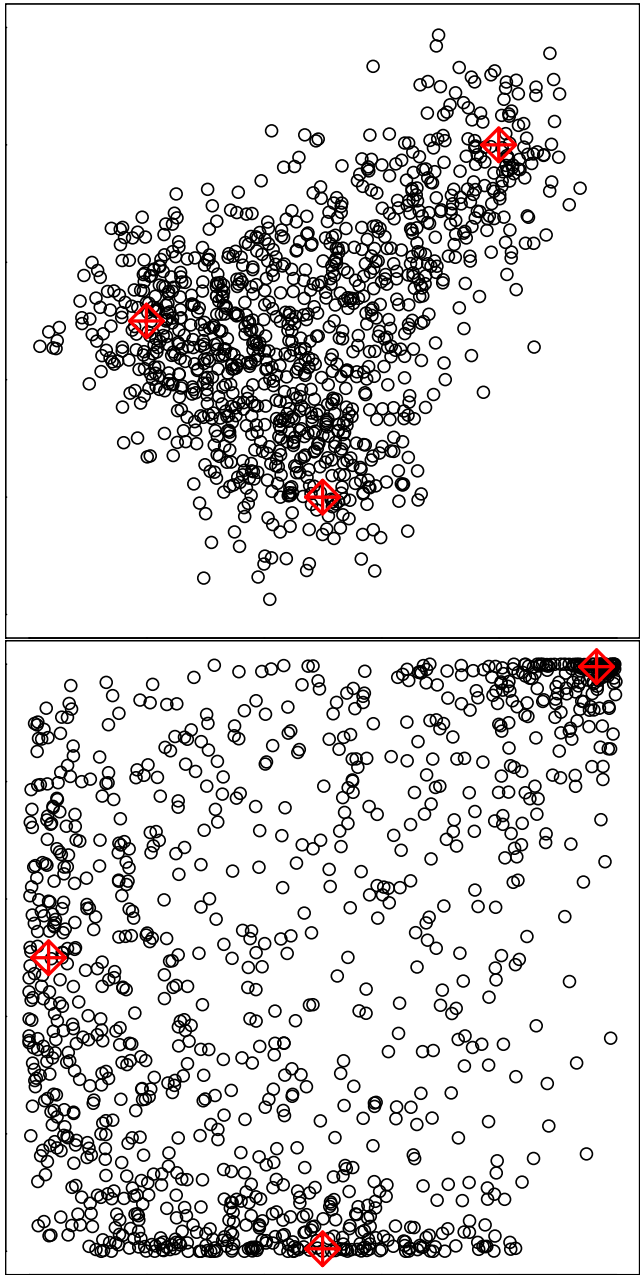
Representations

- different domains
  (p-values, body size, weight)
- monotone transformations (e.g. $\log$)

Problem: archetypal analysis is
sensitive to choice of representation

Solution: Copula extension

- representation independence
- robust against outliers
- mixed data & missing values



Copula

Copula density

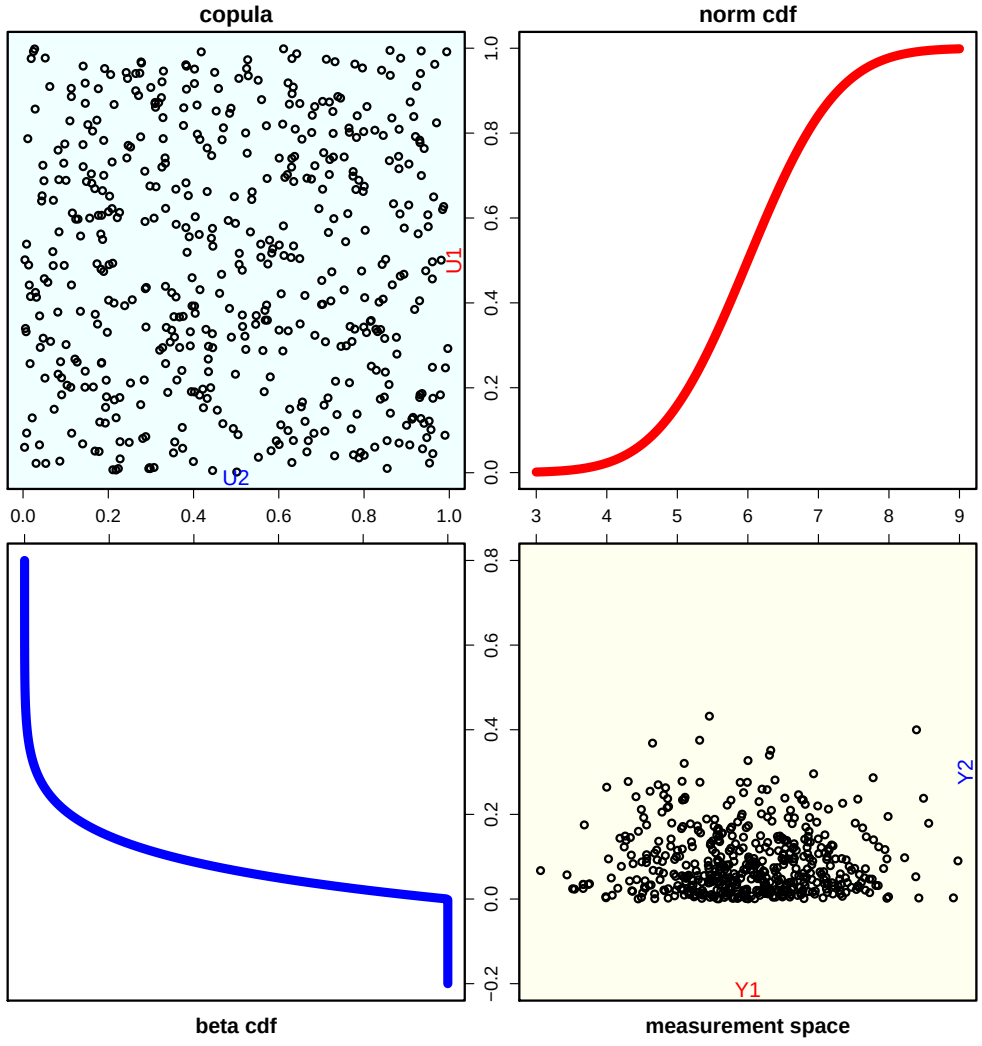
- p-dim pdf on $[0,1]^p$
- uniform marginals
- defines dependency structure

Property

- construct arbitrary multivariate distribution

$$y_j = F_j^{-1}(u_j)$$

(Sklar 1959)



Independence Copula

Copula

Copula density

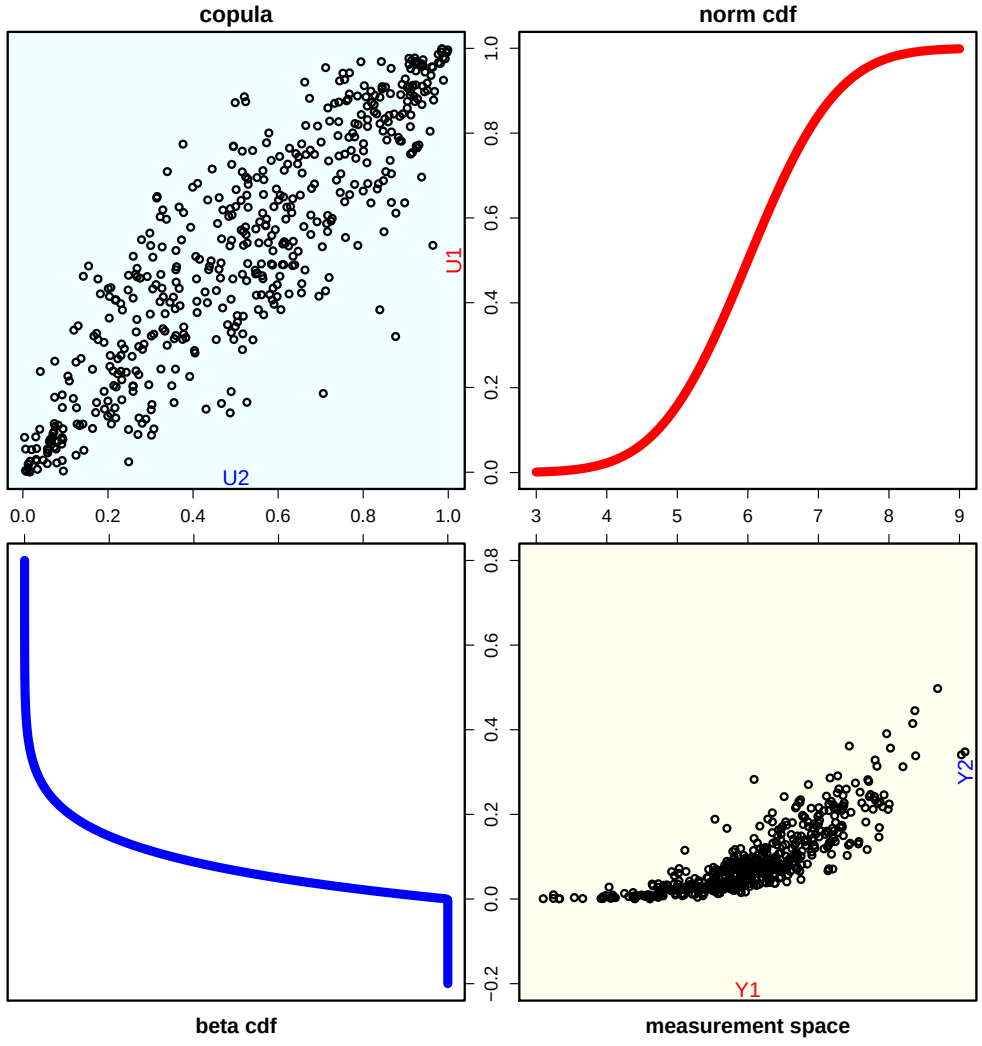
- p-dim pdf on $[0,1]^p$
- uniform marginals
- defines dependency structure

Property

- construct arbitrary multivariate distribution

$$y_j = F_j^{-1}(u_j)$$

(Sklar 1959)



Gaussian Copula

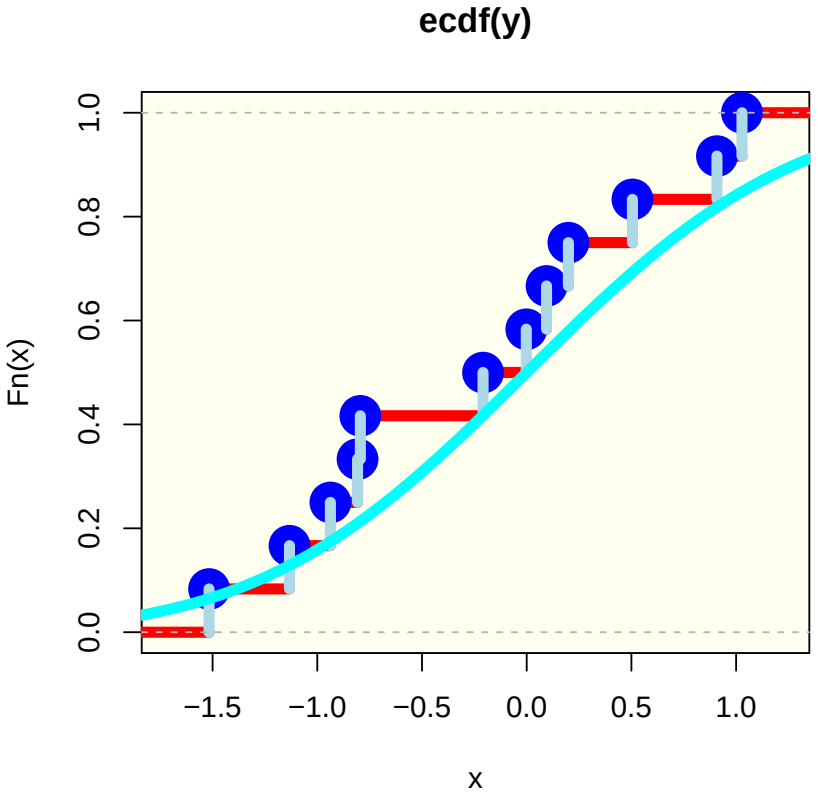
Semi-parametric Gaussian Copula

Hierarchical model:

$$
\begin{array}{l}
R \\
\vert \\
\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, R) \\
\vert \\
y_i = f_i(x_i)
\end{array}
$$

- ATs invariant against monotone marginal transformations
- ATs depend only on ranks
  $\rightsquigarrow$ insensitive to outliers.

- Idea: reconstruct latent space
  $x_j = (\Phi^{-1} \circ F_j)(y_j)$
- Use empirical cdf's
  $\hat{F}_j = \frac{\text{ranks}(Y[\cdot, j])}{n+1}$



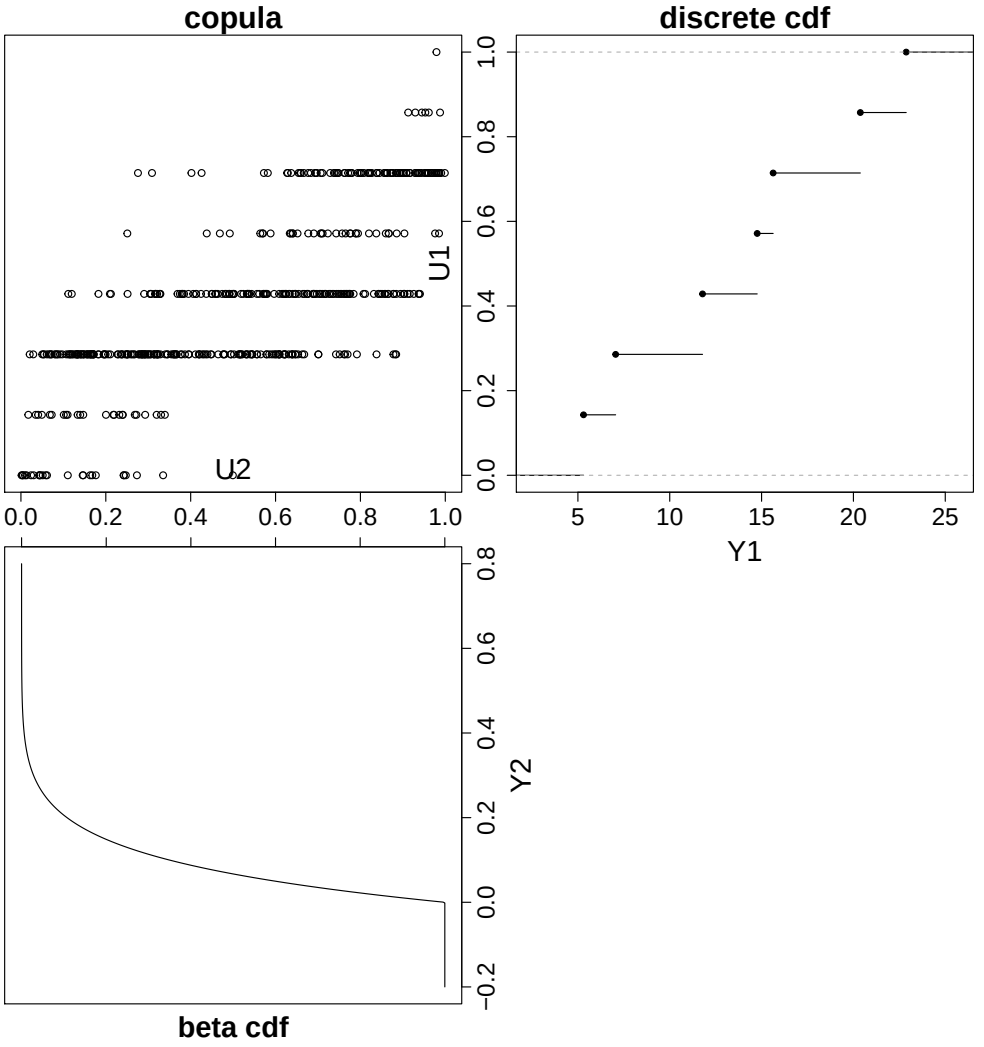
Mixed Continous / Discrete Data

Problem with discrete data: ties, empirical copula no longer uniform

Extended rank likelihood (Hoff, 2007):
stochastic association-preserving mapping
Algorithm: Gibbs sampler

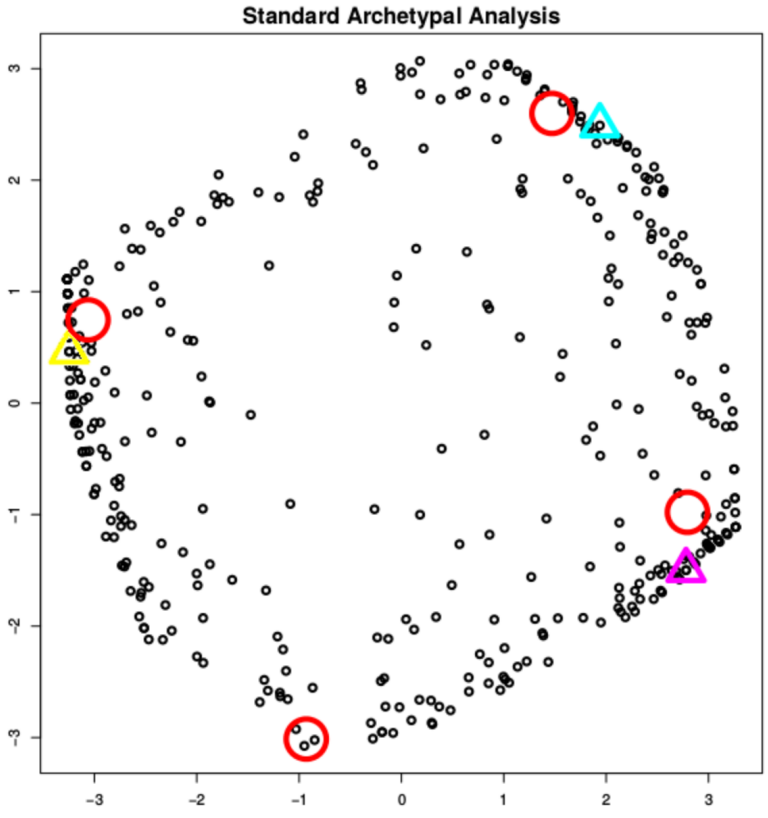
1. sample latent variables $\boldsymbol{x}$, $R|\boldsymbol{x}$

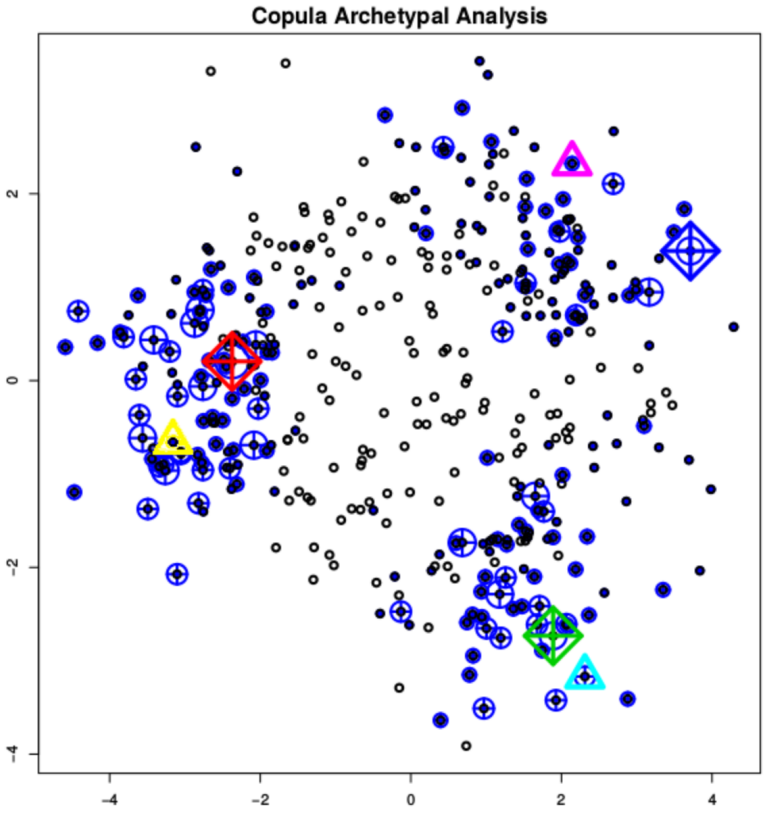
2. compute archetypes $\boldsymbol{z}$



Artificial Data

- Monotone transformation with beta marginals
- Variables quantised to 5 levels





Why use the Gaussian copula?

Generative model

$$\boldsymbol{a}_i \sim \mathrm{Dir}_K(\boldsymbol{\alpha})$$

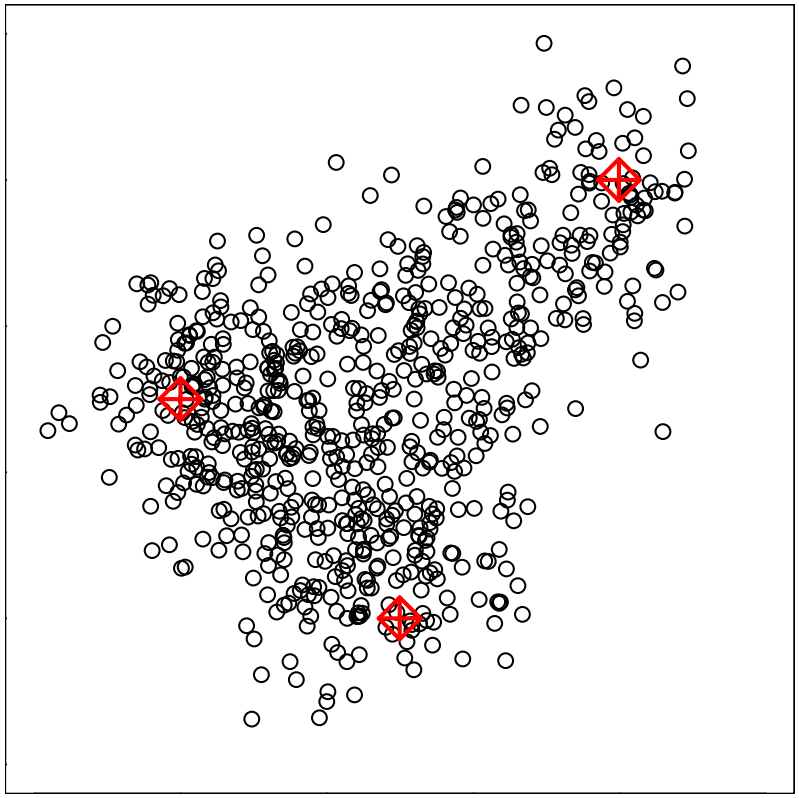
$$\boldsymbol{x}_i | Z, \boldsymbol{a}_i \sim \mathcal{N}(Z^t \boldsymbol{a}_i, \eta I_p)$$

$$X | Z, A \sim \mathcal{MN}\left(\overbrace{AZ}^{M}, I, \eta I\right)$$

$$X^t X \sim \mathcal{W}_{\mathsf{nc}}\left(n, nI, M^t M\right)$$

$$\approx \mathcal{W}_{\mathsf{c}}\left(n, \frac{1}{n} M^t M + \eta I\right)$$

(Steyn and Roux, 1972)

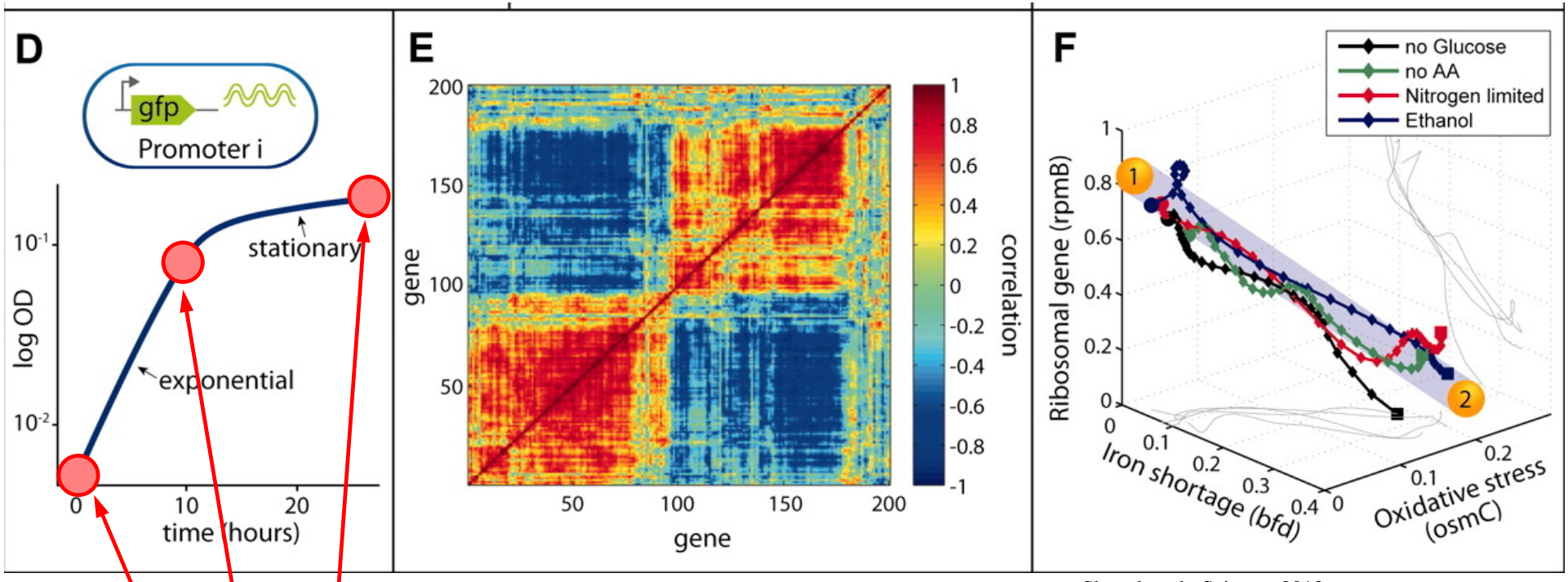


Under the assumed generative model,
a Gaussian covariance structure is plausible.

Archetype Analysis: Applications

What is it good for?

Analysis of E.coli Data

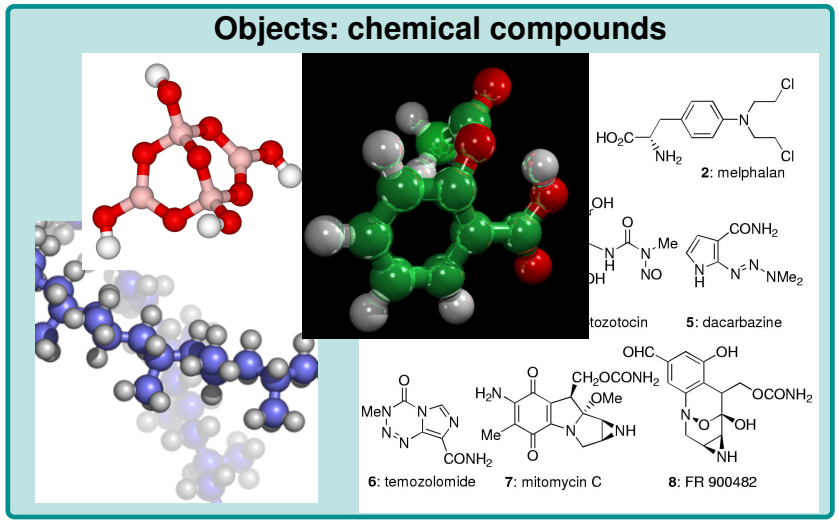


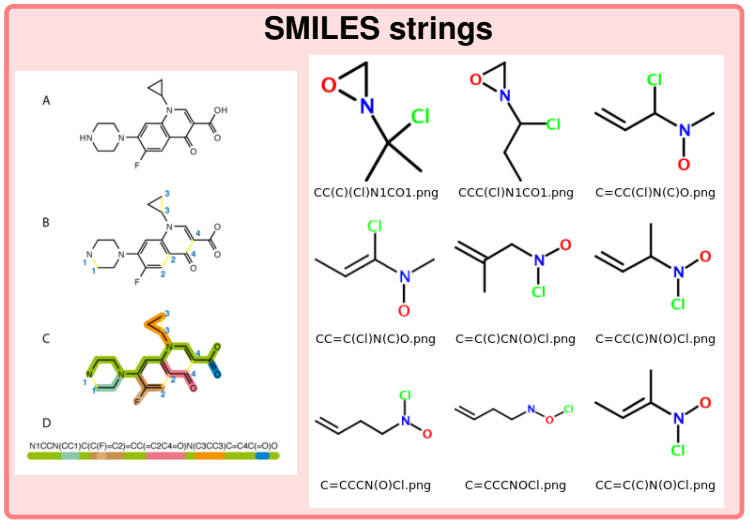
Shoval et al., Science, 2012

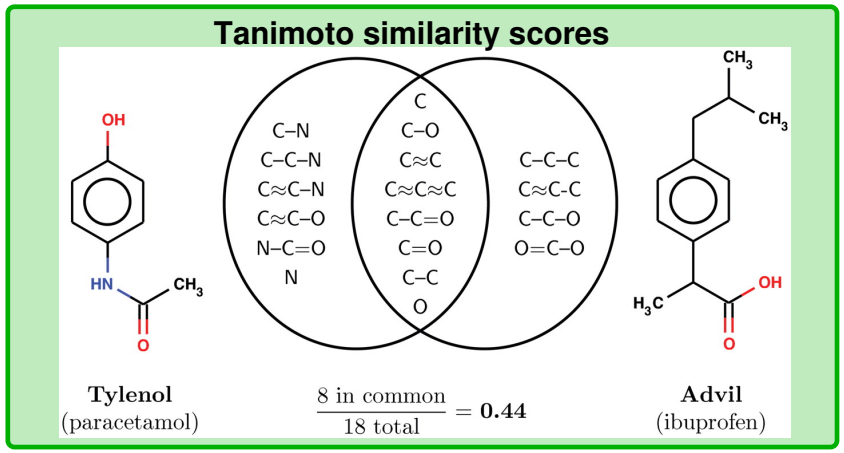
Identified Archetypes

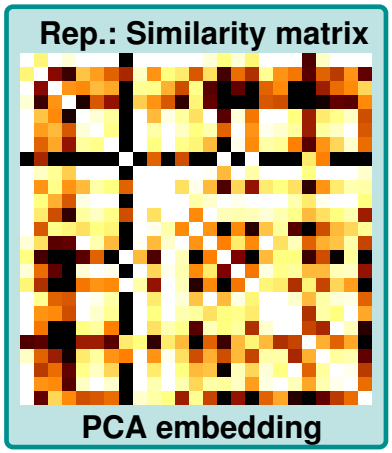
Archetypical Chemical Compounds

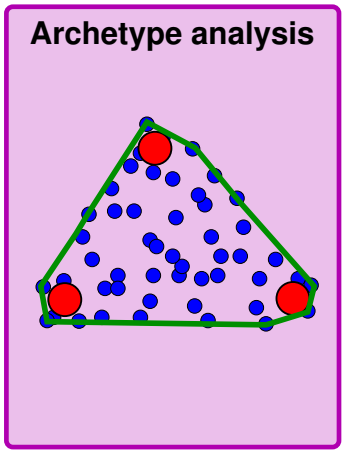
Idea: find ATs in list of compounds identified in an AIDS antiviral screen performed by the Developmental Therapeutics Program of the NCI/NIH, enriched with all available anti-HIV drugs



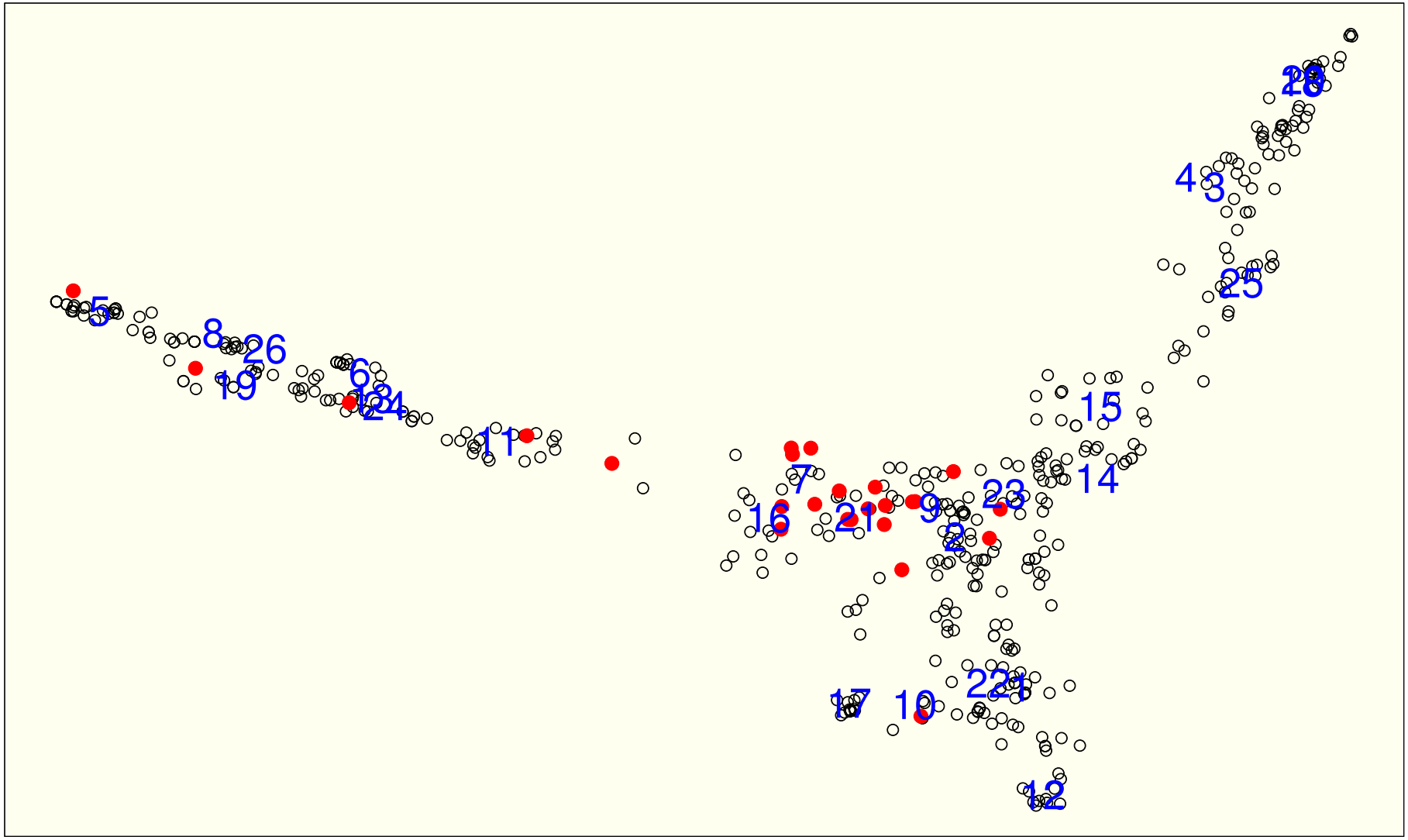




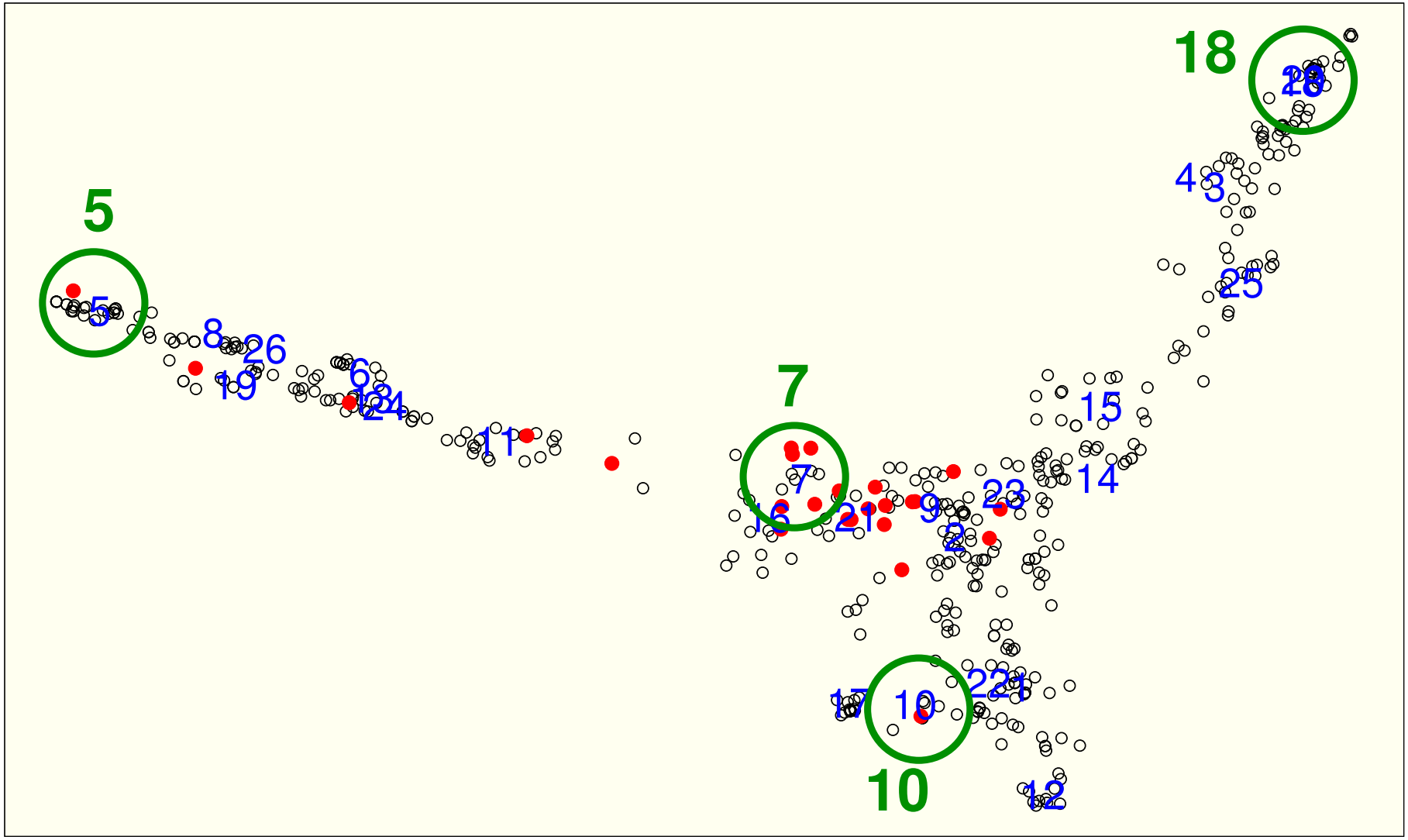




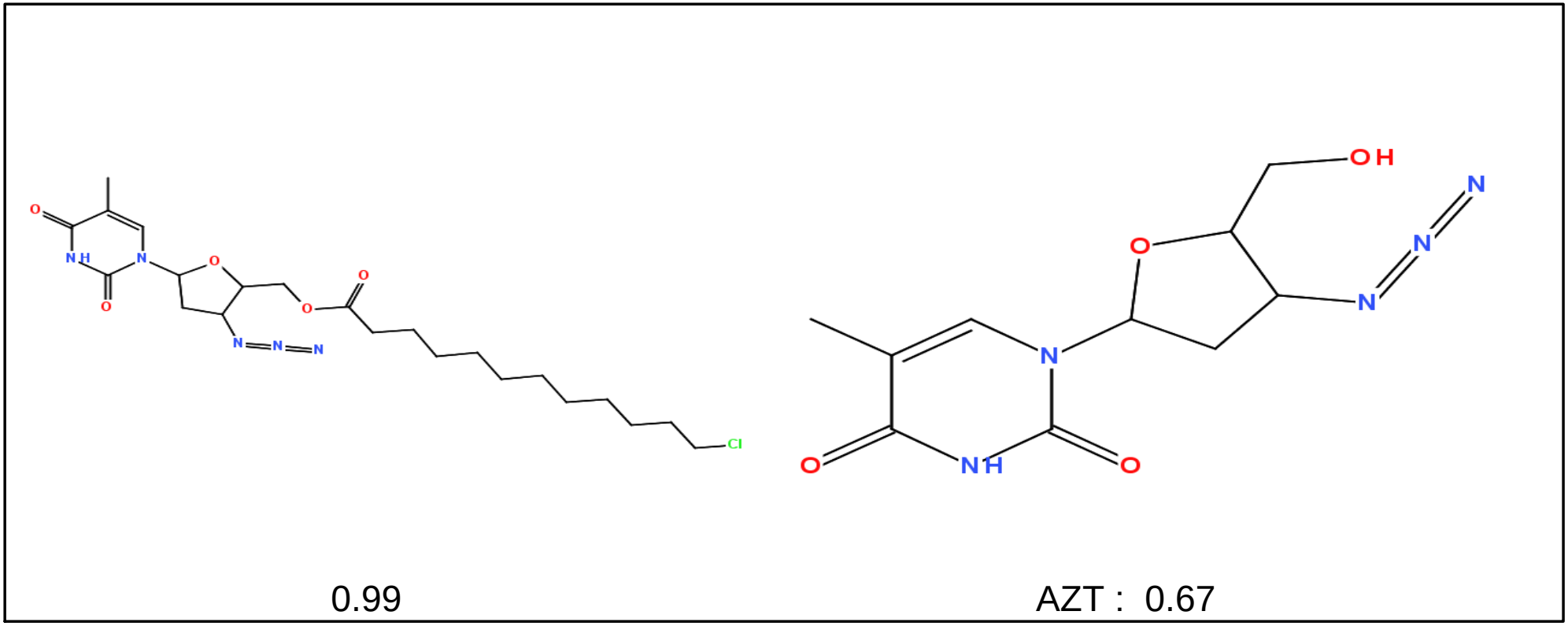
Archetypical Chemical Compounds



Archetypical Chemical Compounds

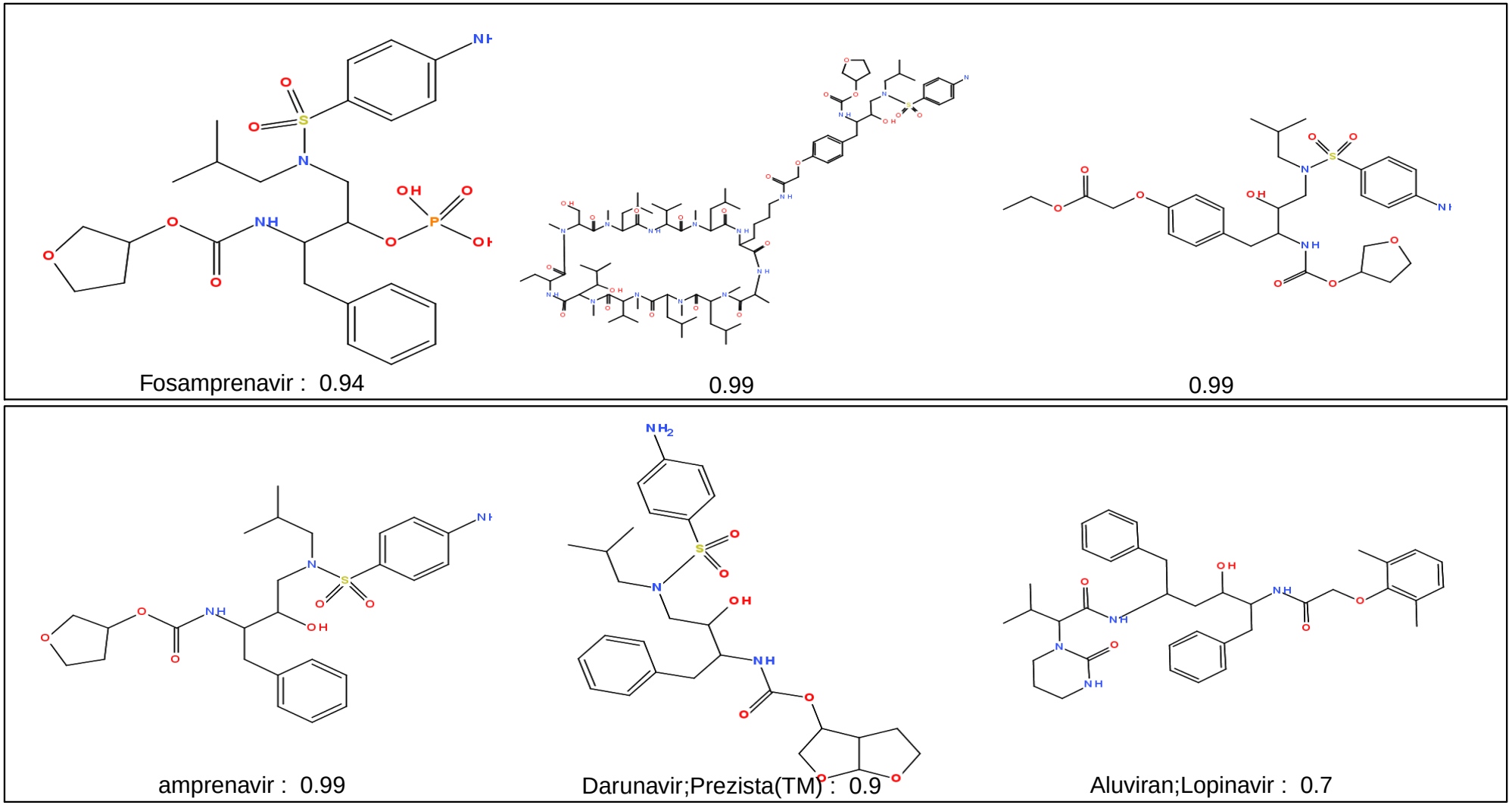


Compounds explained by AT 5



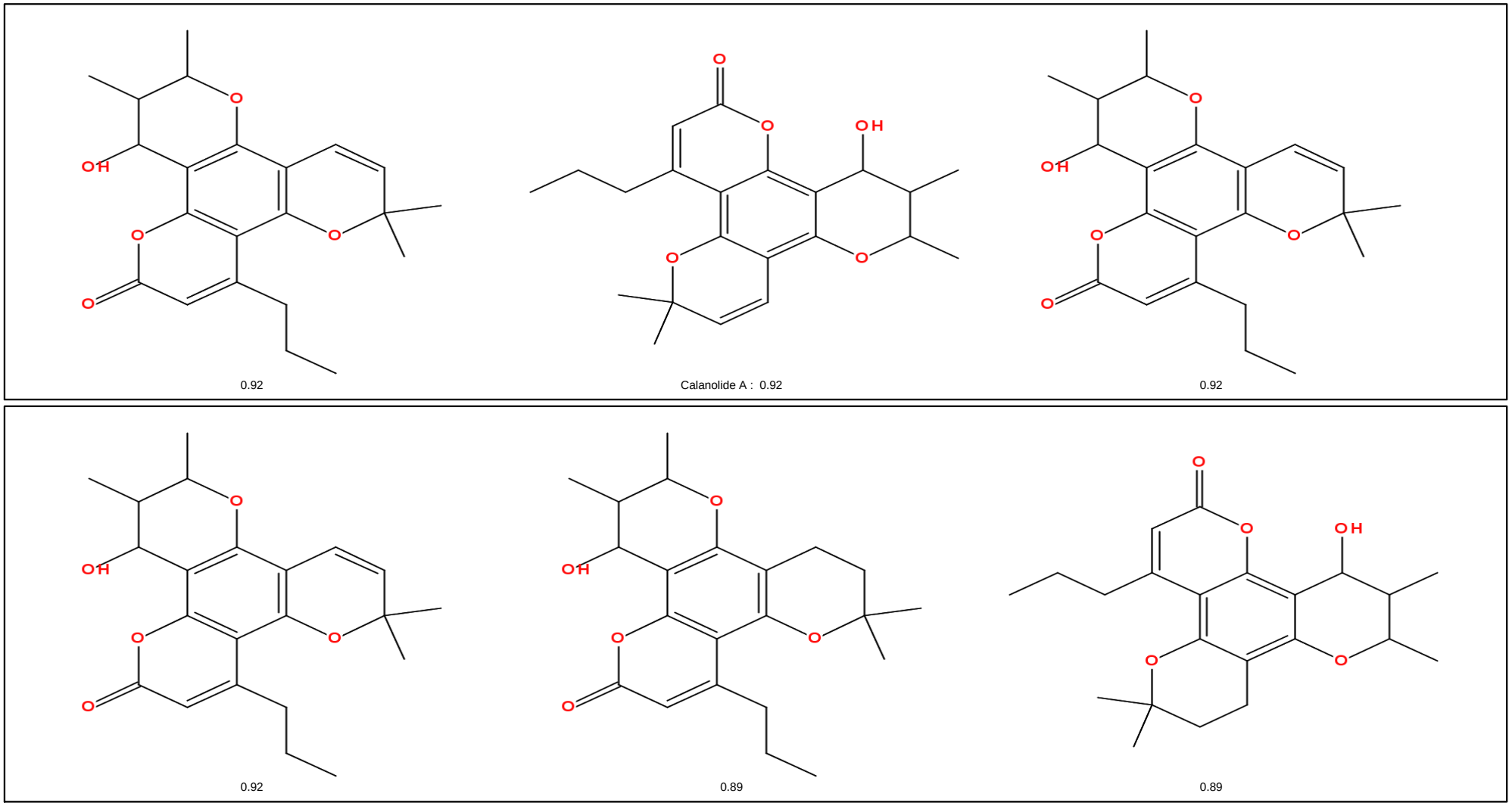
AZT It is of the nucleoside reverse-transcriptase inhibitor (NRTI) class. It inhibits the enzyme (reverse transcriptase) that HIV uses to synthesize DNA.

Compounds explained by AT 7



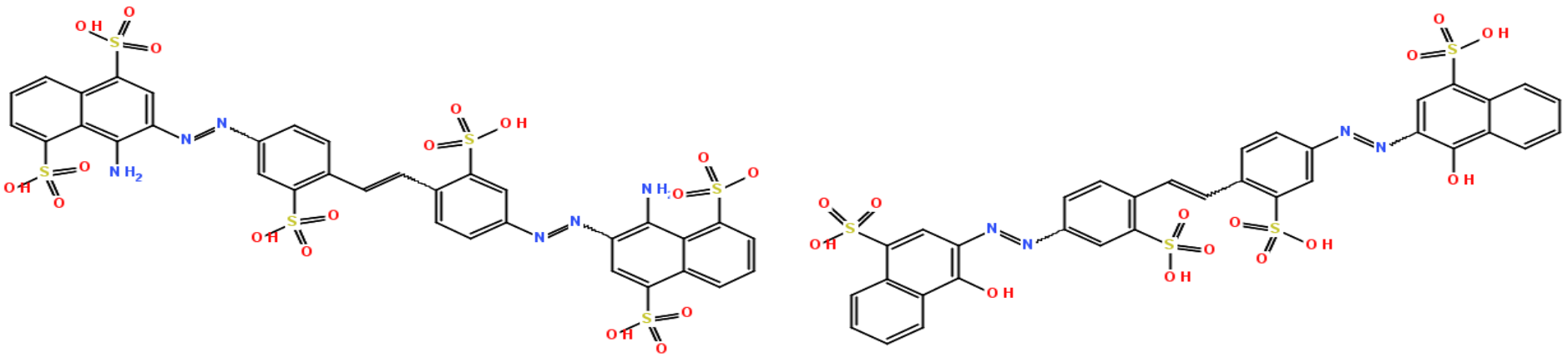
HIV protease inhibitors (they block a peptide cleaving enzyme).

Compounds explained by AT 10



Calanolide A is an experimental non-nucleoside reverse transcriptase inhibitor (NNRTI).

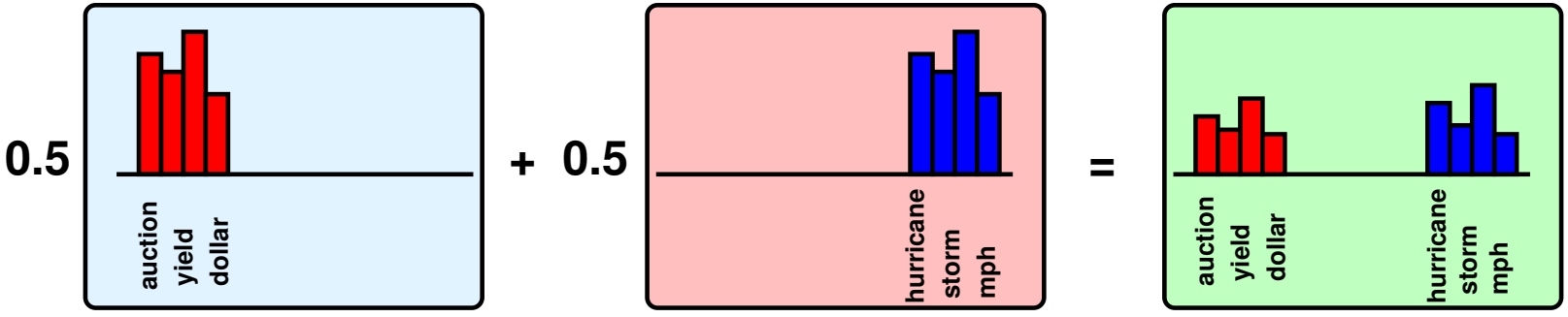
Compounds explained by AT 18



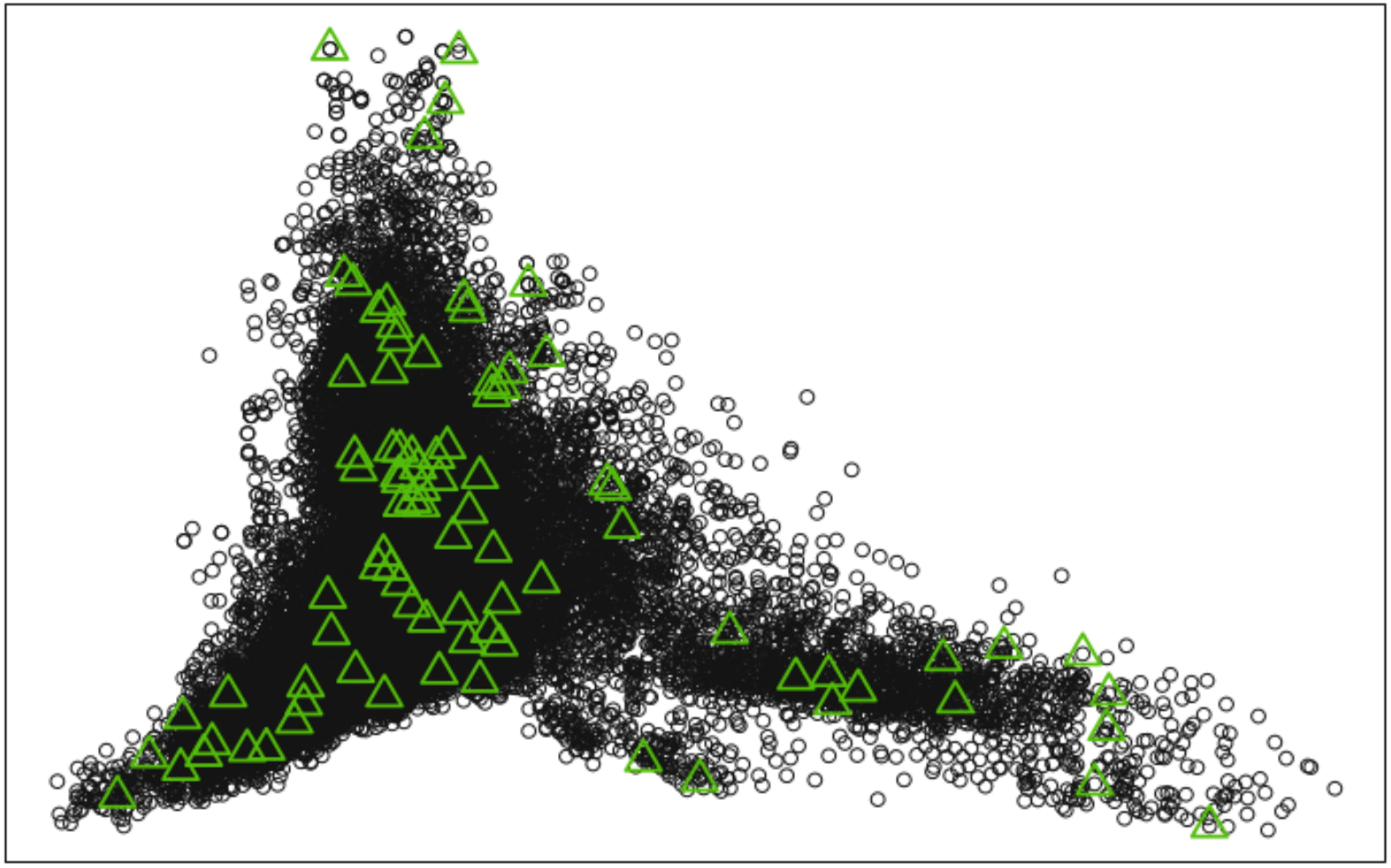
0.81 1−Naphthalenesulfonic acid, 3,3'−[1,2−ethenediylbis [(3−sulfo−4,1−phenylene)azo]]bis[4−hydroxy−, tetrasodium sal

Text categorization

- Reuters Corpus Volume 1 (RCV1): an archive of news documents
- 4 categories: Economics, Government, Corporate, Markets
- 23149 documents, vocabulary of 57180 words.
- Documents represented by word frequencies:
  Term frequency (TF) times **Inverse Document Frequency** (IDF).
- Automatically detect "pure" or archetypical documents
  (which might represent "pure" topics)

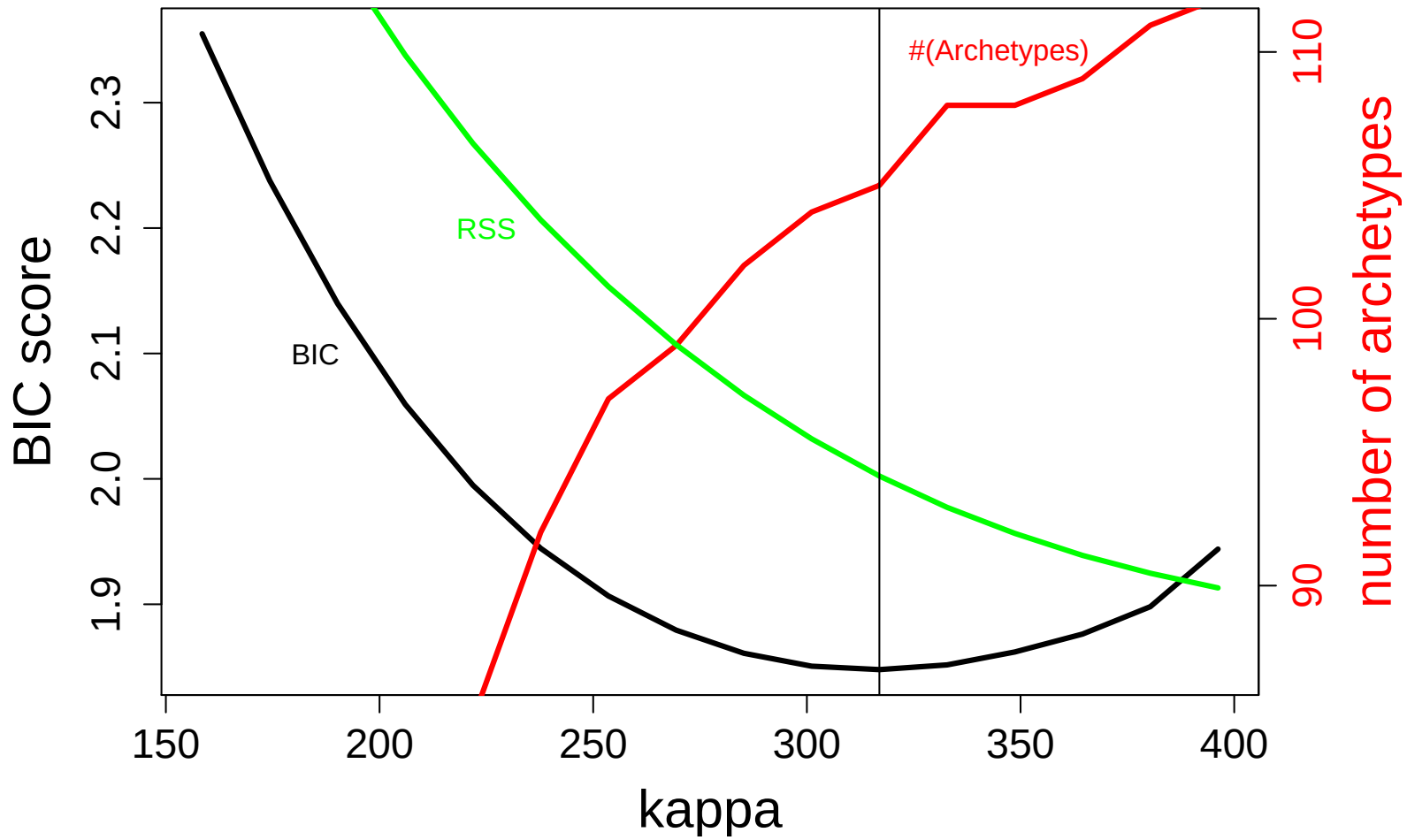


Text categorization

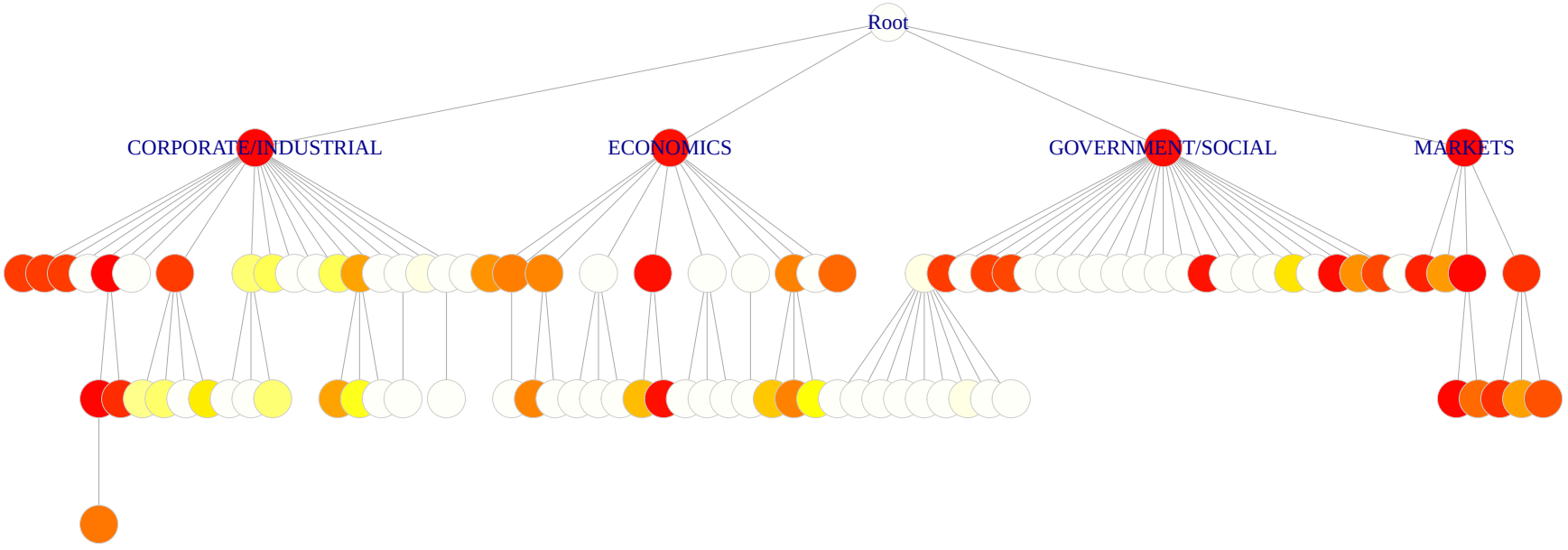


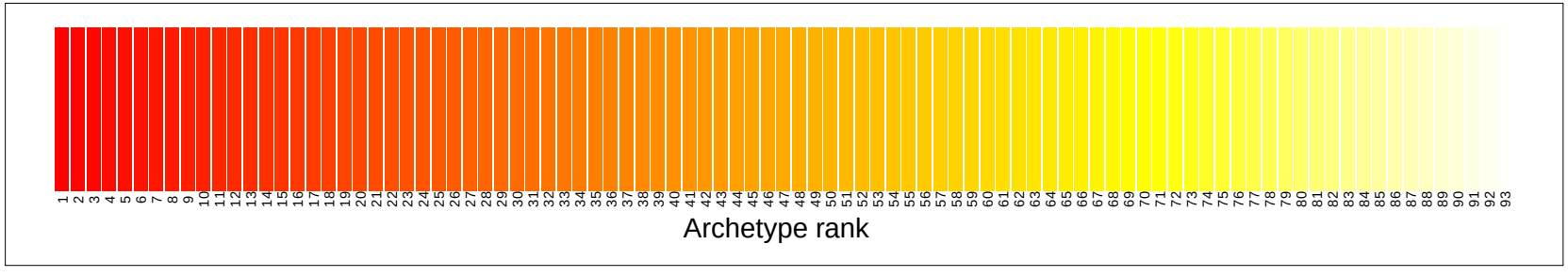
Prabhakaran et al., Springer LNCS 7476, 2012

Text categorization



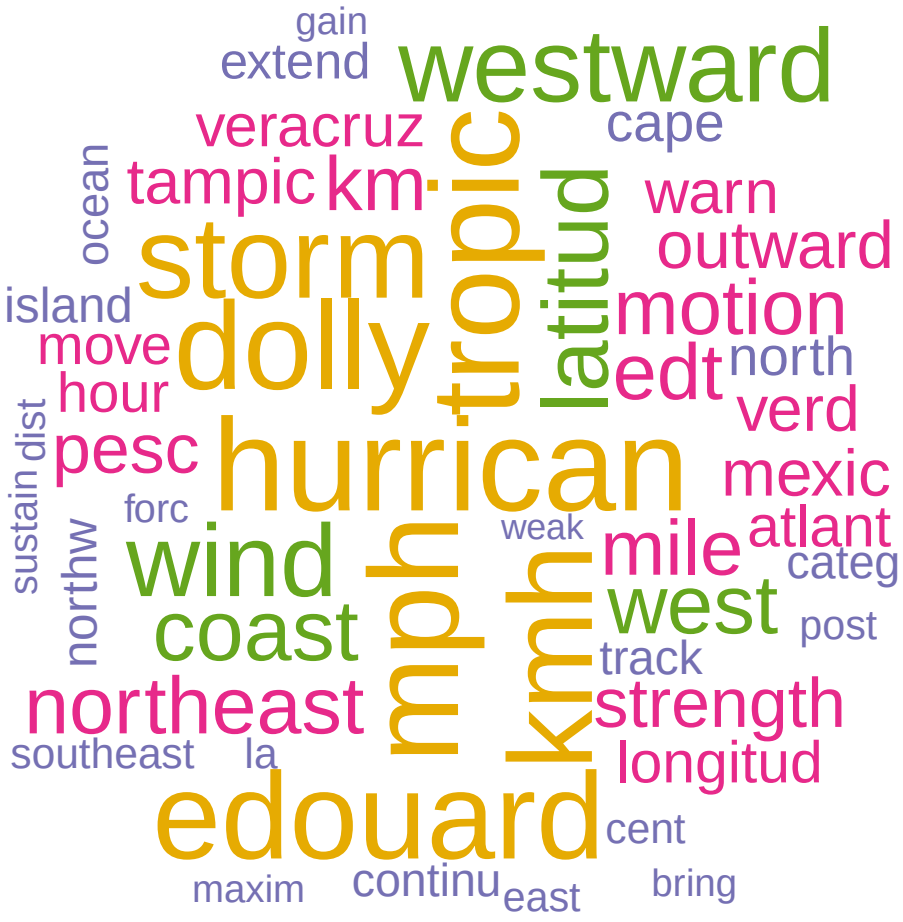
Text categorization



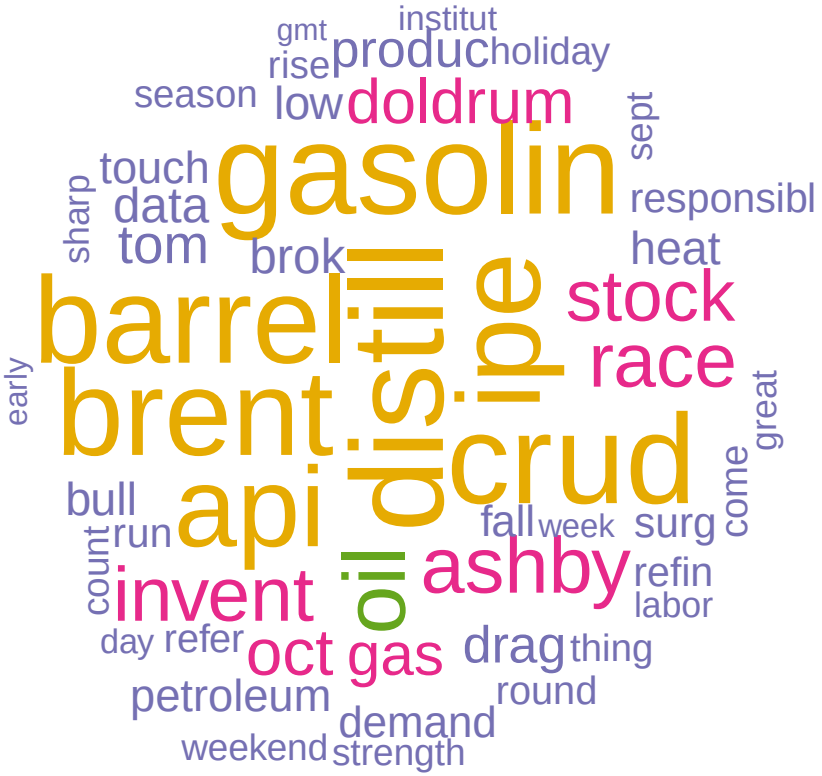


Text categorization

Text categorization



Text categorization



Pose analysis

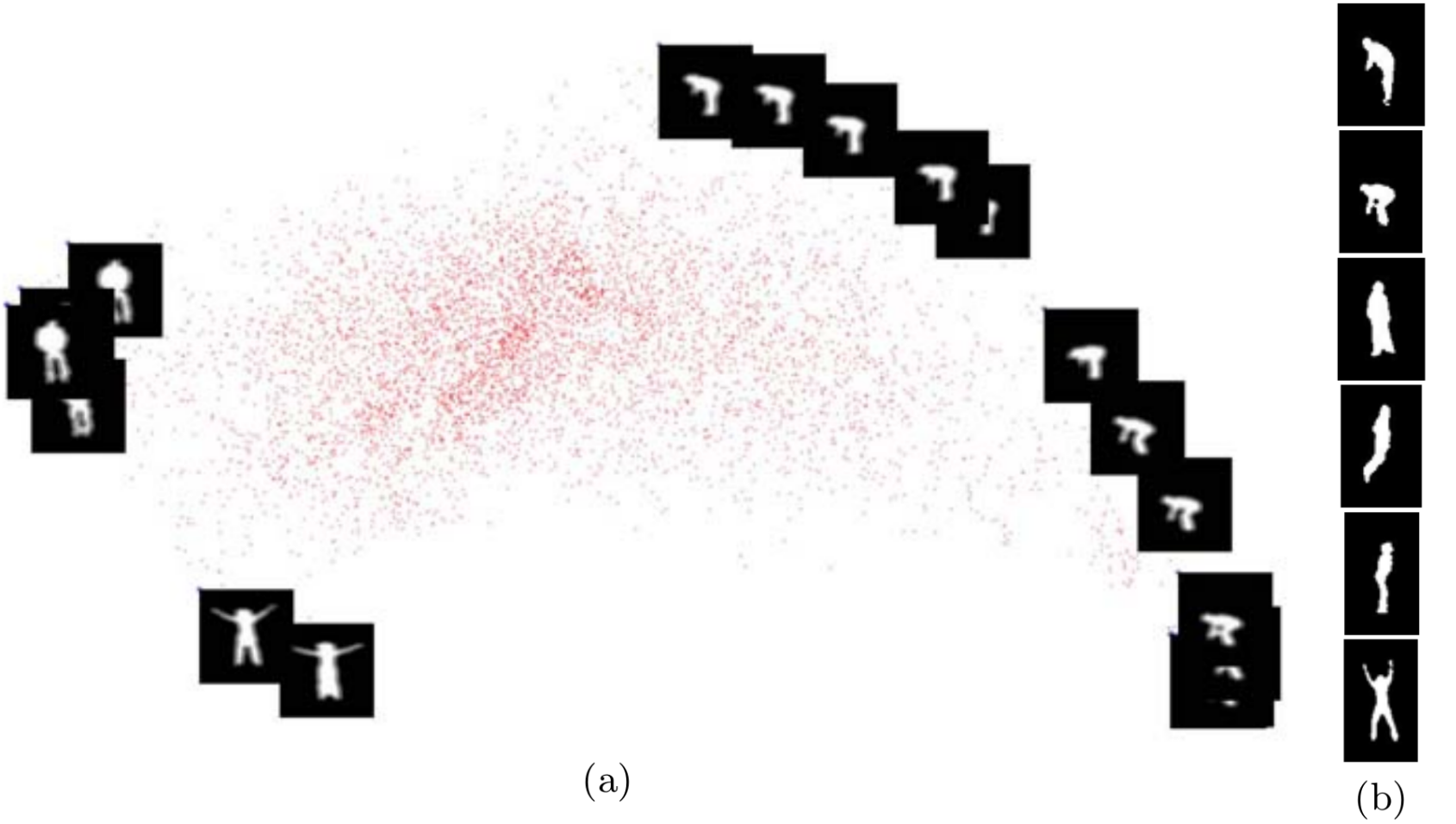
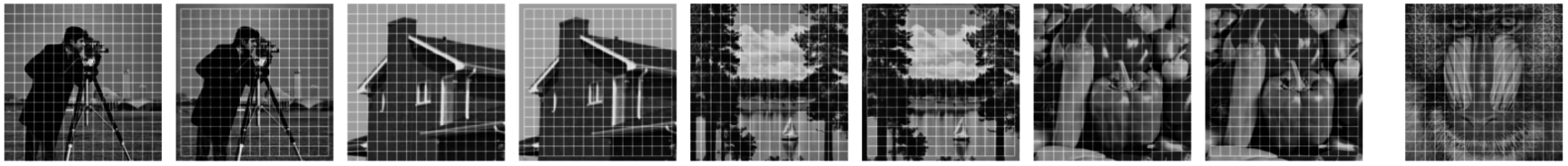


Fig. 5. (a) 2D projection of the Weizman set containing 5.000 body poses; points on the convex hull are shown as pictures. (b) 6 archetypal poses extracted from the data.

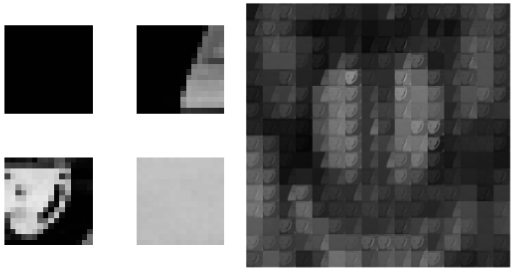
Bauckhage et al., 2009

Image Encoding

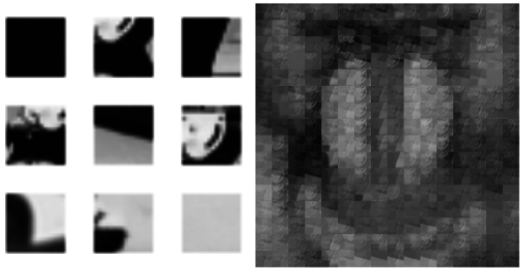


(a) training (b) test

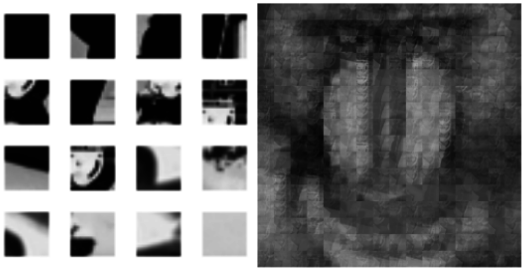
Fig. 3: Image patches (16×16 pixels) for archetypal autoencoding experiments.



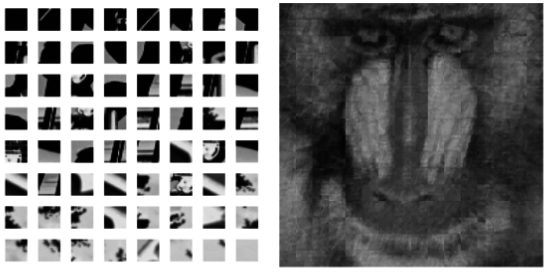
(a) $k = 4$ archetypes



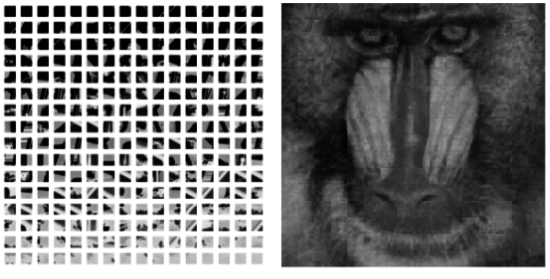
(b) $k = 9$ archetypes



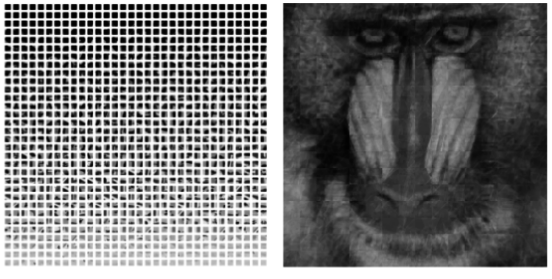
(c) $k = 16$ archetypes



(d) $k = 64$ archetypes



(e) $k = 256$ archetypes



(f) $k = 1024$ archetypes

Conclusion

- Archetype analysis is a powerful tool identifying **representative objects** in large data collections.
- Many **technical challenges:**
  - data types, representation
    $\rightsquigarrow$ invariances due to semi-parametric copula construction
  - computational/memory complexity
    $\rightsquigarrow$ stagewise forward algorithms, convex hull approximations, etc.
- Many **application areas:** objects can be bacteria, genes, documents, chemical compounds, images, etc. etc.
- **Open questions:** AT analysis essentially is an auto-encoding technique
  $\rightsquigarrow$ neural implementations,
  $\rightsquigarrow$ use as building-blocks in deep belief networks, etc.

Acknowledgments

Department of Mathematics and Computer Science, University of Basel: **Dinu Kaufmann, Sebastian Keller, Damian Murezzan, Sonali Parbhoo, Sandhya Prabhakaran, Mélanie Rey, Aleksander Wieczorek, Mario Wieser**

University Hospital Zurich: **Francesca Di Giallonardo, Yannick Duport, Christine Leemann, Stefan Schmutz, Nottania K. Campbell, Beda Joos, Osvaldo Zagordi, Huldrych F. Günthard, Karin J. Metzner**

Department of Biosystems Science and Engineering, ETH Zurich: **Armin Töpfer, Christian Beisel, Niko Beerenwinkel**

Functional Genomics Center Zurich: **Maria R.Lecca, Andrea Patrignani**

Inst. Medical Virology, U Zurich: **Peter Rusert, Alexandra Trkola**