# Module Master Recherche
# Apprentissage et Fouille

Michele Sebag – Balazs Kegl – Antoine Cornuéjols
http://tao.lri.fr

28 octobre 2009

# Représentation pour l'apprentissage

- Sélection d'attributs
- Changements de représentation linéaires
- Changements de représentation non linéaires
- Propositionalisation
- Une étude de cas

# Au début sont les données...

| | AGE | SEX | BMI | BP | ··· Serum Measurements ··· | | | | | | Response |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Patient | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 | y |
| 1 | 59 | 2 | 32.1 | 101 | 157 | 93.2 | 38 | 4 | 4.9 | 87 | 151 |
| 2 | 48 | 1 | 21.6 | 87 | 183 | 103.2 | 70 | 3 | 3.9 | 69 | 75 |
| 3 | 72 | 2 | 30.5 | 93 | 156 | 93.6 | 41 | 4 | 4.7 | 85 | 141 |
| 4 | 24 | 1 | 25.3 | 84 | 198 | 131.4 | 40 | 5 | 4.9 | 89 | 206 |
| 5 | 50 | 1 | 23.0 | 101 | 192 | 125.4 | 52 | 4 | 4.3 | 80 | 135 |
| 6 | 23 | 1 | 22.6 | 89 | 139 | 64.8 | 61 | 2 | 4.2 | 68 | 97 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 441 | 36 | 1 | 30.0 | 95 | 201 | 125.2 | 42 | 5 | 5.1 | 85 | 220 |
| 442 | 36 | 1 | 19.6 | 71 | 250 | 133.2 | 97 | 3 | 4.6 | 92 | 57 |

# Motivations : Trouver et élaguer des descripteurs

Avant l'apprentissage : décrire les données.

- Une description trop pauvre $\Rightarrow$ on ne peut rien faire
- Une description trop riche $\Rightarrow$ on doit élaguer les descripteurs

Pourquoi ?

- L'apprentissage n'est pas un problème bien posé
- $\Longrightarrow$ Rajouter de l'information inutile (l'age du vélo de ma grand-mère) peut dégrader les hypothèses obtenues.

# Feature Selection, Position du problème

Contexte

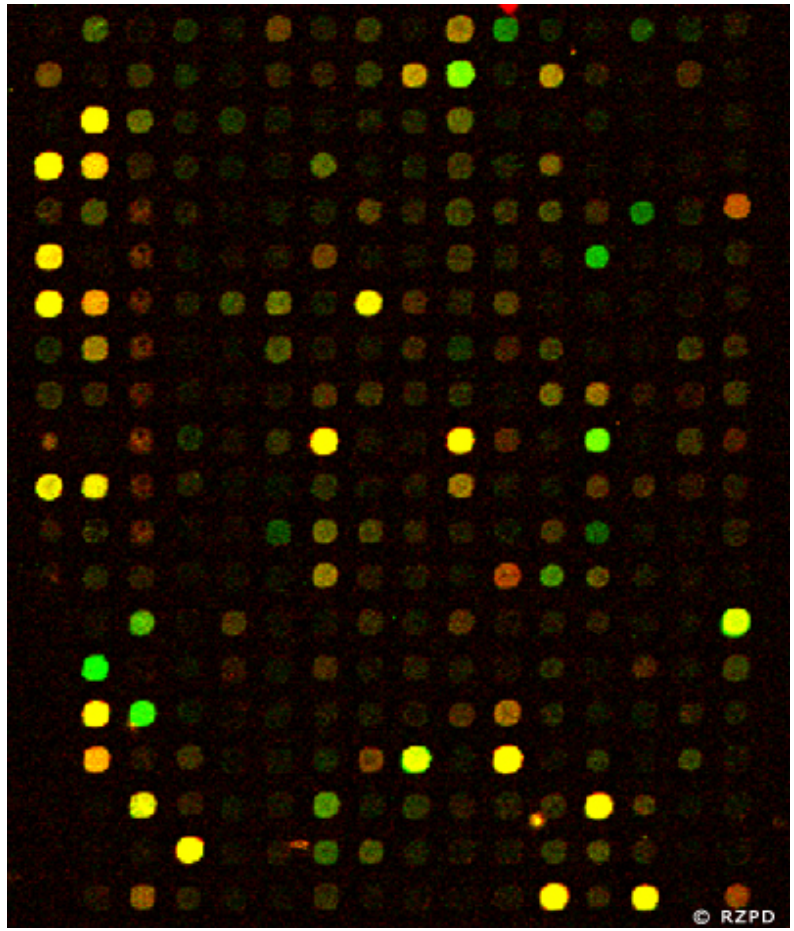
- Trop d'attributs % nombre exemples
  - En enlever — Feature Selection
  - En construire d'autres — Feature Construction
  - En construire moins — Dimensionality Reduction
- Cas logique du 1er ordre : — Propositionalisation

Le but caché : sélectionner ou construire des descripteurs ?

- Feature Construction : construire les bons descripteurs
- A partir desquels il sera facile d'apprendre
- Les meilleurs descripteurs = les bonnes hypothèses...

# Quand l'apprentissage c'est la sélection d'attributs

## Bio-informatique




- 30 000 gènes
- peu d'exemples (chers)
- but : trouver les gènes pertinents

# Il est facile de faire n'importe quoi

*Un exemple d'aventure fort désagréable...*

http://www-stat.stanford.edu/∼hastie/TALKS/barossa.pdf

# (Rappel) Définition de p-value

Contexte : observation — le rouge est sorti 14 fois sur 20

Question : est-ce le hasard ? — deux hypothèses

- $H_0$: le casino est honnête — $\Pr(\text{rouge}) = 1/2$
- ... ou non

p-value : Proba (observations | $H_0$)

*probabilité d'observer ça sous l'hypothèse $H_0$*

Nb de rouges sur N tirages $\sim \mathcal{B}(N, 1/2)$

$$\Pr(\#\text{ rouges} \geq 14) = .047$$

... On rejette l'hypothèse $H_0$ à 5% de niveau de confiance

# Position du problème

## Buts

- Sélection : trouver un sous-ensemble d'attributs
- Ordre/Ranking : ordonner les attributs

## Formulation

Soient les attributs $\mathcal{A} = \{a_1, ..a_d\}$. Soit la fonction :

$$\begin{aligned} \mathcal{F} : \mathcal{P}(\mathcal{A}) &\mapsto \mathbb{R} \\ A \subset \mathcal{A} &\mapsto Err(A) = \text{ erreur min. des hypothèses fondées sur } A \end{aligned}$$

**Trouver** $Argmin(\mathcal{F})$

## Difficultés

- Un problème d'optimisation combinatoire $(2^d)$
- D'une fonction $\mathcal{F}$ inconnue...

# Selection de features: approche filtre

Méthode univariée

Définir $score(a_i)$; ajouter itérativement les attributs maximisant *score*

ou retirer itérativement les attributs minimisant *score*

+ simple et pas cher
− optima très locaux

Backtrack possible

- Etat courant $\mathcal{A}$
- Ajouter $a_i$ à $\mathcal{A}$
- Peut être ajouter $a_i$ rend $a_j \in \mathcal{A}$ inutile ?
- Essayer d'enlever les features de $\mathcal{A}$

Backtrack = moins glouton; meilleures solutions ; beaucoup plus cher.

# Selection de features: approche wrapping

Méthode multivariée

Mesurer la qualité d'un ensemble d'attributs :

estimer $\mathcal{F}(a_{i1}, ...a_{ik})$

Contre

Beaucoup plus cher : une estimation = un pb d'apprentissage.

Pour

Optima meilleurs

# Selection de features: approche embarquée (embedded)

Principe – online

On rajoute à l'apprentissage un critère qui favorise les hypothèses à peu d'attributs.

Par exemple : trouver $w$, $h(x) = wx$, qui minimise

$$\sum_i (h(x_i) - y_i)^2 + \sum_d |w_d|$$

Premier terme : coller aux données

Deuxième terme : favoriser $w$ avec beaucoup de coordonnées nulles

Principe – offline

On a trouvé

$$h(x) = wx = \sum_d w_d x_d$$

Si $|w_d|$ petit, l'attribut $d$ n'est pas important... Les enlever et recommencer.

# Approches filtre, 1

## Notations

Base d'apprentissage : $\mathcal{E} = \{(x_i, y_i), i = 1..n, y_i \in \{-1, 1\}\}$

$a(x_i)$ = valeur attribut $a$ pour exemple $(x_i)$

## Corrélation

$$corr(a) = \frac{\sum_i a(x_i).y_i}{\sqrt{\sum_i (a(x_i))^2 \times \sum_i y_i^2}} \propto \sum_i a(x_i).y_i = \ < a, y >$$

## Limites

- Attributs corrélés entre eux
- Dépendance non linéaire

# Approches filtre, 2

## Corrélation et projection

Stoppiglia et al. 2003

Repeat

- $a^*$ = attribut le plus corrélé à la classe

$$a^* = argmax\{\sum_i a(x_i)y_i,\ a \in \mathcal{A}\}$$

- Projeter les autres attributs sur l'espace orthogonal à $a^*$

$$\forall b \in \mathcal{A} \qquad b \rightarrow \quad b - \frac{<a^*,b>}{<a^*,a^*>}\ a^*$$

$$b(x_i) \rightarrow \quad b(x_i) - \frac{\sum_j a^*(x_j)b(x_j)}{\sum_j a^*(x_j)^2} a^*(x_i)$$

# Corrélation et projection, suite

- Projeter $y$ sur l'espace orthogonal à $a^*$

$$y \rightarrow y - \frac{<a^*,y>}{<a^*,a^*>} a^*$$

$$y_i \rightarrow y_i - -\frac{\sum_j a^*(x_j) y_j}{\sum_j a^*(x_j)^2} a^*(x_i)$$

- Until Critère d'arrêt
  - Rajouter des attributs aléatoires ($r(x_i) = \pm 1$) *probe*
  - Quand le critère de corrélation sélectionne des attributs aléatoires, s'arrêter.

Limitations

quand il y a plus de 6-7 attributs pertinents, ne marche pas bien.

# Approches filtre, 3

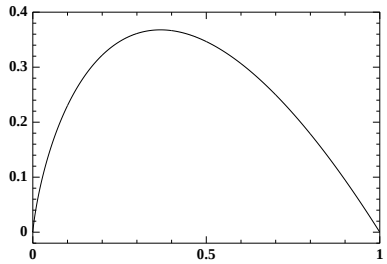
Gain d'information

arbres de décision

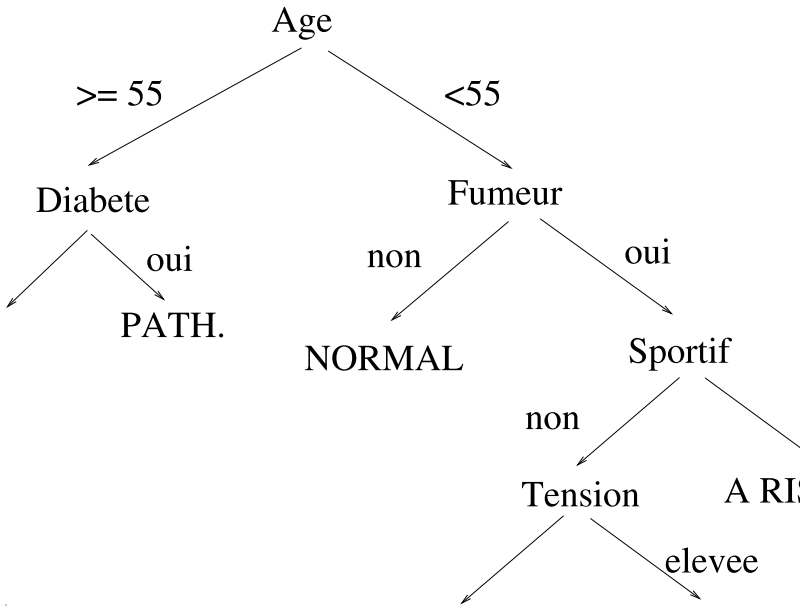
$$p([a = v]) = Pr(y = 1|a(x_i) = v)$$

$$QI([a = v]) = -p([a = v]) \log p([a = v])$$

$$QI(a) = \sum_v Pr(a(x_i) = v) QI([a = v])$$



# Gain d'information, suite




## Limitations

- Les mêmes que celles des arbres de décision
- Problème de XOR.

# Quelques scores

*en fouille de textes, contexte supervisé*

Notations : $c_i$ une classe $a_k$ un mot (ou terme)

Critères

1. Fréquence conditionnelle $P(c_i|a_k)$
2. Information mutuelle $P(c_i, a_k)Log(\frac{P(c_i,a_k)}{P(c_i)P(a_k)})$
3. Gain d'information $\sum_{c_i,\neg c_i}\sum_{a_k,\neg a_k} P(c,a)Log\frac{p(a,c)}{P(a)P(c)}$
4. Chi-2 $\frac{(P(t,c)P(\neg t,\neg c)-P(t,\neg c)P(\neg t,c))^2}{P(t)P(\neg t)P(c)P(\neg c)}$
5. Pertinence $Log\frac{P(t,c)+d}{P(\neg t,\neg c)+d}$

# Approches wrapper

## Principe générer/tester

Etant donné une liste de candidats $\mathcal{L} = \{A_1, .., A_p\}$

- Générer un candidat $A$
- Calculer $\mathcal{F}(A)$
  - apprendre $h_A$ à partir de $\mathcal{E}_{|A}$
  - tester $h_A$ sur un ensemble de test $= \hat{\mathcal{F}}(A)$
- Mettre à jour $\mathcal{L}$.

## Algorithmes

- hill-climbing / multiple restart
- algorithmes génétiques Vafaie-DeJong, IJCAI 95
- (*) programmation génétique & feature construction. Krawiec, GPEH 01

# Approches a posteriori

Principe

- Construire des hypothèses
- En déduire les attributs importants
- Eliminer les autres
- Recommencer

Algorithme : SVM Recursive Feature Elimination Guyon et al. 03

- SVM linéaire $\rightarrow h(x) = sign(\sum w_i.a_i(x) + b)$
- Si $|w_i|$ est petit, $a_i$ n'est pas important
- Eliminer les $k$ attributs ayant un poids min.
- Recommencer.

# Limites

Hypothèses linéaires

- Un poids par attribut.

Quantité des exemples

- Les poids des attributs sont liés.
- La dimension du système est liée au nombre d'exemples.

Or le pb de FS se pose souvent quand il n'y a pas assez d'exemples

# Représentation pour l'apprentissage

- Sélection d'attributs
- Changements de représentation linéaires
- Changements de représentation non linéaires
- Propositionalisation
- Une étude de cas

# Partie 2. Changements de représentation lineaires

- Réduction de dimensionalité
- Analyse en composantes principales
- Projections aléatoires

# Dimensionality Reduction — Intuition

## Degrees of freedom

- Image: 4096 pixels; but not independent
- Robotics: (# camera pixels + # infra-red) $\times$ time; but not independent

## Goal

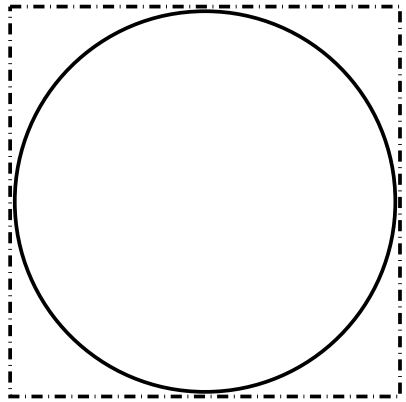
Find the (low-dimensional) structure of the data:

- Images
- Robotics
- Genes

# Dimensionality Reduction

## In high dimensions

- Everybody lives in the corners of the space
  Volume of Sphere $V_n = \frac{2\pi r^2}{n} V_{n-2}$
- All points are far from each other



## Approaches

- Linear dimensionality reduction
  - Principal Component Analysis
  - Random Projection
- Non-linear dimensionality reduction

## Criteria

- Complexity/Size
- Prior knowledge

e.g., relevant distance

# Linear Dimensionality Reduction

Training set *unsupervised*

$$\mathcal{E} = \{(\mathbf{x}_k), \mathbf{x}_k \in \mathbb{R}^D, k = 1 \dots N\}$$

## Projection from $\mathbb{R}^D$ onto $\mathbb{R}^d$

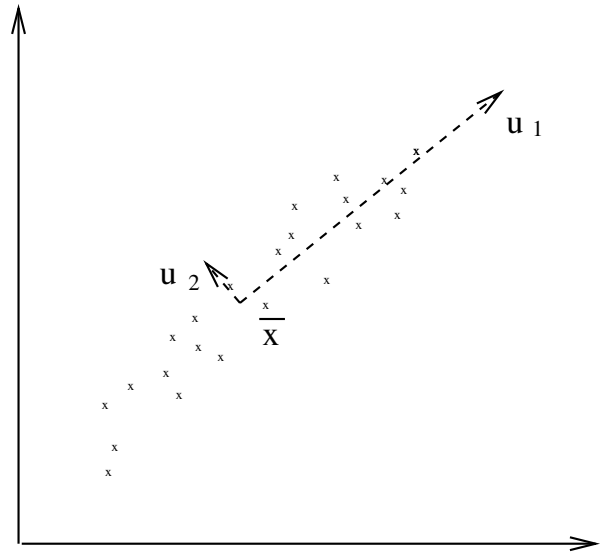
$$\begin{aligned} \mathbf{x} \in \mathbb{R}^D \rightarrow \quad & h(\mathbf{x}) \in \mathbb{R}^d, \ d << D \\ & h(\mathbf{x}) = A\mathbf{x} \\ s.t. \text{ minimize} \quad & \sum_{k=1}^{N} ||\mathbf{x}_k - h(\mathbf{x}_k)||^2 \end{aligned}$$

# Principal Component Analysis

## Covariance matrix $S$

Mean $\mu_i = \frac{1}{N}\sum_{k=1}^{N} X_i(\mathbf{x}_k)$

$$S_{ij} = \frac{1}{N}\sum_{k=1}^{N}(X_i(\mathbf{x}_k) - \mu_i)(X_j(\mathbf{x}_k) - \mu_j)$$




symmetric $\Rightarrow$ can be diagonalized

$$S = U\Delta U' \quad \Delta = Diag(\lambda_1, \ldots \lambda_D)$$

## Thm: Optimal projection in dimension $d$

projection on the first $d$ eigenvectors of $S$

Let $u_i$ the eigenvector associated to eigenvalue $\lambda_i$ $\lambda_i > \lambda_{i+1}$

$$h : \mathbb{R}^D \mapsto \mathbb{R}^d, h(\mathbf{x}) = < \mathbf{x}, u_1 > u_1 + \ldots + < \mathbf{x}, u_d > u_d$$

# Sketch of the proof

## 1. Maximize the variance of $h(\mathbf{x}) = A\mathbf{x}$

$\sum_k ||\mathbf{x}_k - h(\mathbf{x}_k)||^2 = \sum_k ||\mathbf{x}_k||^2 - \sum_k ||h(\mathbf{x}_k)||^2$

$$\text{Minimize } \sum_k ||\mathbf{x}_k - h(\mathbf{x}_k)||^2 \Rightarrow \text{ Maximize } \sum_k ||h(\mathbf{x}_k)||^2$$

$$Var(h(\mathbf{x})) = \frac{1}{N}\left(\sum_k ||h(\mathbf{x}_k)||^2 - ||\sum_k h(\mathbf{x}_k)||^2\right)$$

As

$$||\sum_k h(\mathbf{x}_k)||^2 = ||A\sum_k \mathbf{x}_k||^2 = N^2||A\mu||^2$$

where $\mu = (\mu_1, \ldots . \mu_D)$.
Assuming that $\mathbf{x}_k$ are centered ($\mu_i = 0$) gives the result.

# Sketch of the proof, 2

## 2. Projection on eigenvectors $u_i$ of $S$

Assume $h(\mathbf{x}) = A\mathbf{x} = \sum_{i=1}^{d} < \mathbf{x}, v_i > v_i$ and show $v_i = u_i$.

$$Var(AX) = (AX)(AX)' = A(XX')A' = ASA' = A(U\Delta U')A'$$

Consider $d = 1,\ v_1 = \sum w_i u_i$ $\qquad$ $\sum w_i^2 = 1$

*remind* $\lambda_i > \lambda_{i+1}$

$$Var(AX) = \sum \lambda_i w_i^2$$

maximized for $w_1 = 1, w_2 = \ldots = w_N = 0$

that is, $v_1 = u_i$.

# Principal Component Analysis, Practicalities

## Data preparation

- Mean centering the dataset

$$\begin{aligned} \mu_i &= \frac{1}{N}\sum_{k=1}^{N} X_i(\mathbf{x}_k) \\ \sigma_i &= \sqrt{\frac{1}{N}\sum_{k=1}^{N} X_i(\mathbf{x}_k)^2 - \mu_i^2} \\ z_k &= \left(\frac{1}{\sigma_i}(X_i(\mathbf{x}_k) - \mu_i)\right)_{i=1}^{D} \end{aligned}$$

## Matrix operations

- Computing the covariance matrix

$$S_{ij} = \frac{1}{N}\sum_{k=1}^{N} X_i(z_k)X_j(z_k)$$

- Diagonalizing $S = U'\Delta U$ might be not affordable... Complexity $\mathcal{O}(D^3)$

# Random projection

## Random matrix

$$A : \mathbb{R}^D \mapsto \mathbb{R}^d \qquad A[d, D] \qquad A_{i,j} \sim \mathcal{N}(0, 1)$$

define

$$h(\mathbf{x}) = \frac{1}{\sqrt{d}} A\mathbf{x}$$

## Property: $h$ preserves the norm in expectation

$$E[||h(\mathbf{x})||^2] = ||\mathbf{x}||^2$$

With high probability $1 - 2exp\{-(\varepsilon^2 - \varepsilon^3)\frac{d}{4}\}$

$$(1 - \varepsilon)||\mathbf{x}||^2 \leq ||h(\mathbf{x})||^2 \leq (1 + \varepsilon)||\mathbf{x}||^2$$

# Random projection

## Proof

$$
\begin{aligned}
h(\mathbf{x}) = & \quad \frac{1}{\sqrt{d}} A\mathbf{x} \\
E(||h(\mathbf{x})||^2) = & \quad \frac{1}{d} E\left[\sum_{i=1}^{d}\left(\sum_{j=1}^{D} A_{i,j} X_j(\mathbf{x})\right)^2\right] \\
= & \quad \frac{1}{d}\sum_{i=1}^{d} E\left[\left(\sum_{j=1}^{D} A_{i,j} X_j(\mathbf{x})\right)^2\right] \\
= & \quad \frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{D} E[A_{i,j}^2] E[X_j(\mathbf{x})^2] \\
= & \quad \frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{D} \frac{||\mathbf{x}||^2}{D} \\
= & \quad ||\mathbf{x}||^2
\end{aligned}
$$

# Random projection, 2

## Johnson Lindenstrauss Lemma

For $d > \frac{9 \ln N}{\varepsilon^2 - \varepsilon^3}$, with high probability

$$(1-\varepsilon)||\mathbf{x}_i - \mathbf{x}_j||^2 \leq ||h(\mathbf{x}_i) - h(\mathbf{x}_j)||^2 \leq (1+\varepsilon)||\mathbf{x}_i - \mathbf{x}_j||^2$$

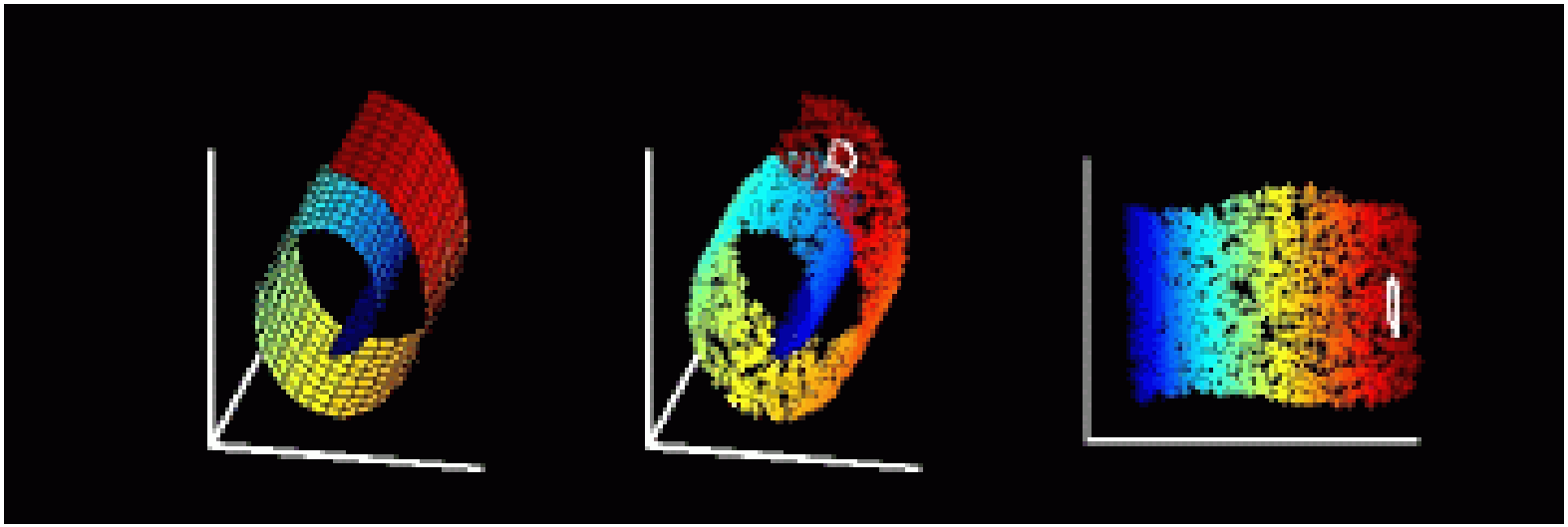
More:

`http://www.cs.yale.edu/clique/resources/RandomProjectionMethod.pdf`

# Représentation pour l'apprentissage

- Sélection d'attributs
- Changements de représentation linéaires
- Changements de représentation non linéaires
- Une étude de cas

# Non-Linear Dimensionality Reduction



## Conjecture

Examples live in a manifold of dimension $d << D$

## Goal: consistent projection of the dataset onto $\mathbb{R}^d$

Consistency:

- Preserve the structure of the data
- e.g. preserve the distances between points

# Multi-Dimensional Scaling

## Position of the problem

- Given $\{\mathbf{x}_1, \ldots, \mathbf{x}_N,\ \mathbf{x}_i \in \mathbb{R}^D\}$
- Given $sim(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}^+$
- Find projection $\Phi$ onto $\mathbb{R}^d$

$$\begin{array}{rl} x \in \mathbb{R}^D \rightarrow & \Phi(x) \in \mathbb{R}^d \\ sim(\mathbf{x}_i, \mathbf{x}_j) \sim & sim(\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)) \end{array}$$

## Optimisation

Define $X,\ X_{i,j} = sim(\mathbf{x}_i, \mathbf{x}_j)$; $X^\Phi,\ X^\Phi_{i,j} = sim(\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j))$

Find $\Phi$ minimizing $||X - X'||$

Rq : Linear $\Phi$ = Principal Component Analysis

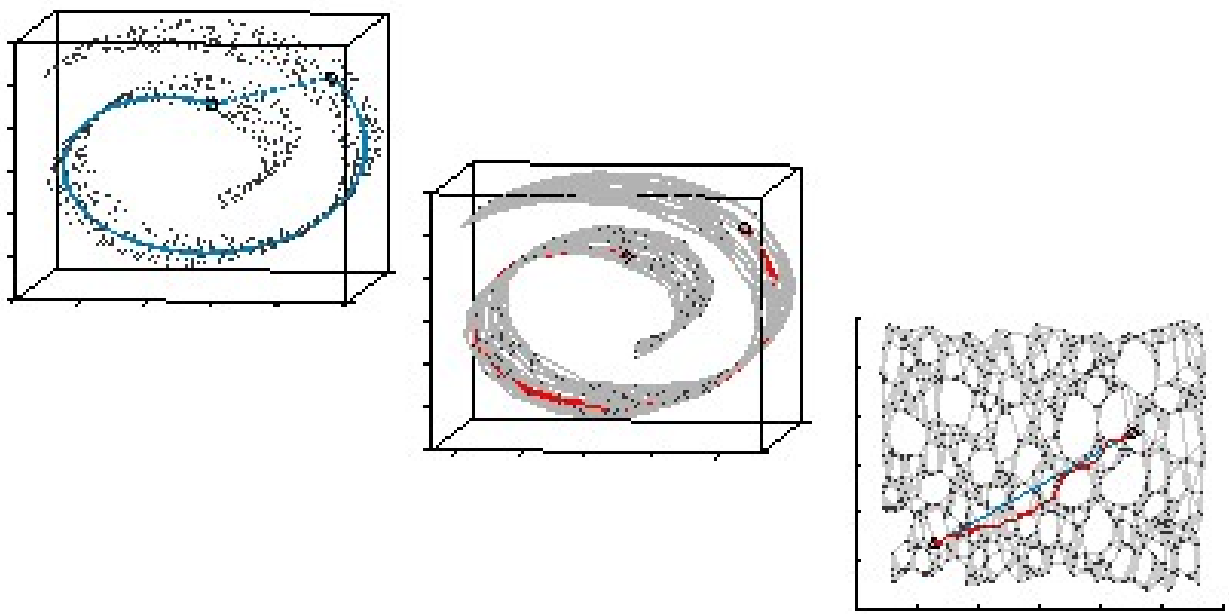
But linear MDS does not work: preserves all distances, while only *local* distances are meaningful

# Non-linear projections

## Approaches

- Reconstruct global structures from local ones and find global projection — Isomap
- Only consider local structures — LLE

Intuition: locally, points live in $\mathbb{R}^d$



# Isomap

Tenenbaum, da Silva, Langford 2000

http://isomap.stanford.edu

## Estimate $d(x_i, x_j)$

- Known if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close
- Otherwise, compute the shortest path between $\mathbf{x}_i$ and $\mathbf{x}_j$
  geodesic distance (dynamic programming)

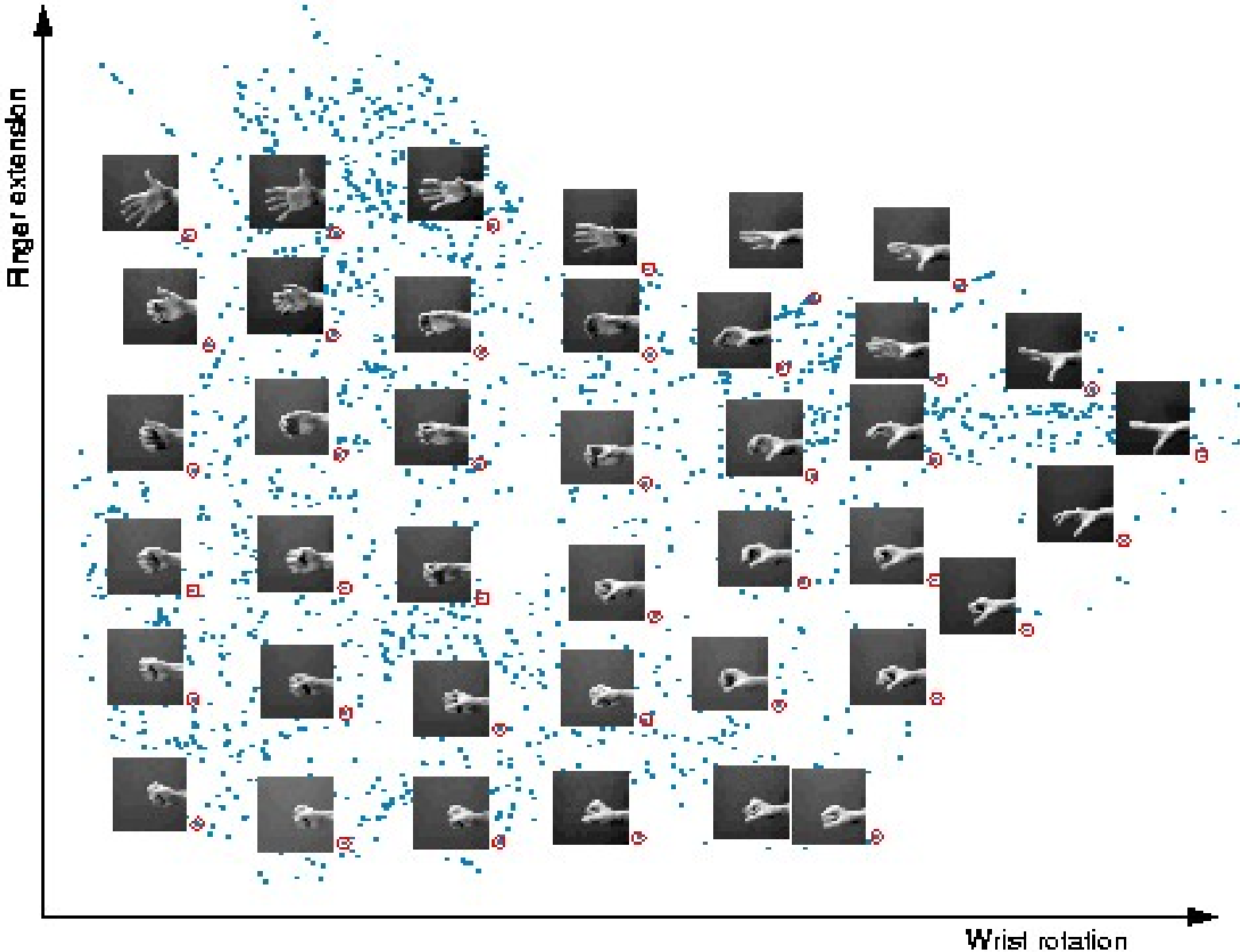
## Requisite

If data points sampled in a convex subset of $\mathbb{R}^d$,
then geodesic distance $\sim$ Euclidean distance on $\mathbb{R}^d$.

## General case

- Given $d(\mathbf{x}_i, \mathbf{x}_j)$, estimate $< \mathbf{x}_i, \mathbf{x}_j >$
- Project points in $\mathbb{R}^d$

# Isomap, 2



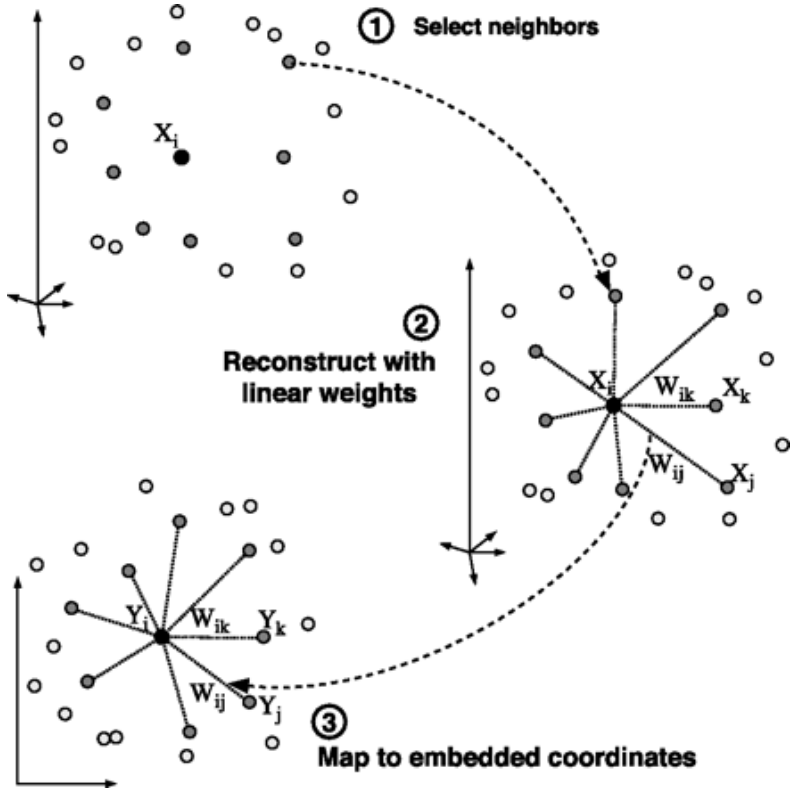
# Locally Linear Embedding

Roweiss and Saul, 2000

http://www.cs.toronto.edu/~roweis/lle/

## Principle

- Find local description for each point: depending on its neighbors



# Local Linear Embedding, 2

## Find neighbors

For each $\mathbf{x}_i$, find its nearest neighbors $\mathcal{N}(i)$

Parameter: number of neighbors

## Change of representation

Goal Characterize $\mathbf{x}_i$ wrt its neighbors:

$$\mathbf{x}_i = \sum_{j \in \mathcal{N}(i)} w_{i,j} \mathbf{x}_j \quad \text{with} \quad \sum_{j \in \mathcal{N}(i)} w_{ij} = 1$$

Property: invariance by translation, rotation, homothety

How Compute the local covariance matrix:

$$C_{j,k} = < x_j - x_i, x_k - x_i >$$

Find vector $w_i$ s.t. $Cw_i = 1$

# Local Linear Embedding, 3

## Algorithm

Local description: Matrix $W$ such that $\sum_j w_{i,j} = 1$

$$W = argmin\{\sum_{i=1}^{N} ||\mathbf{x}_i - \sum_j w_{i,j}\mathbf{x}_j||^2\}$$

Projection: Find $\{z_1, \ldots, z_n\}$ *in* $\mathbb{R}^d$ minimizing

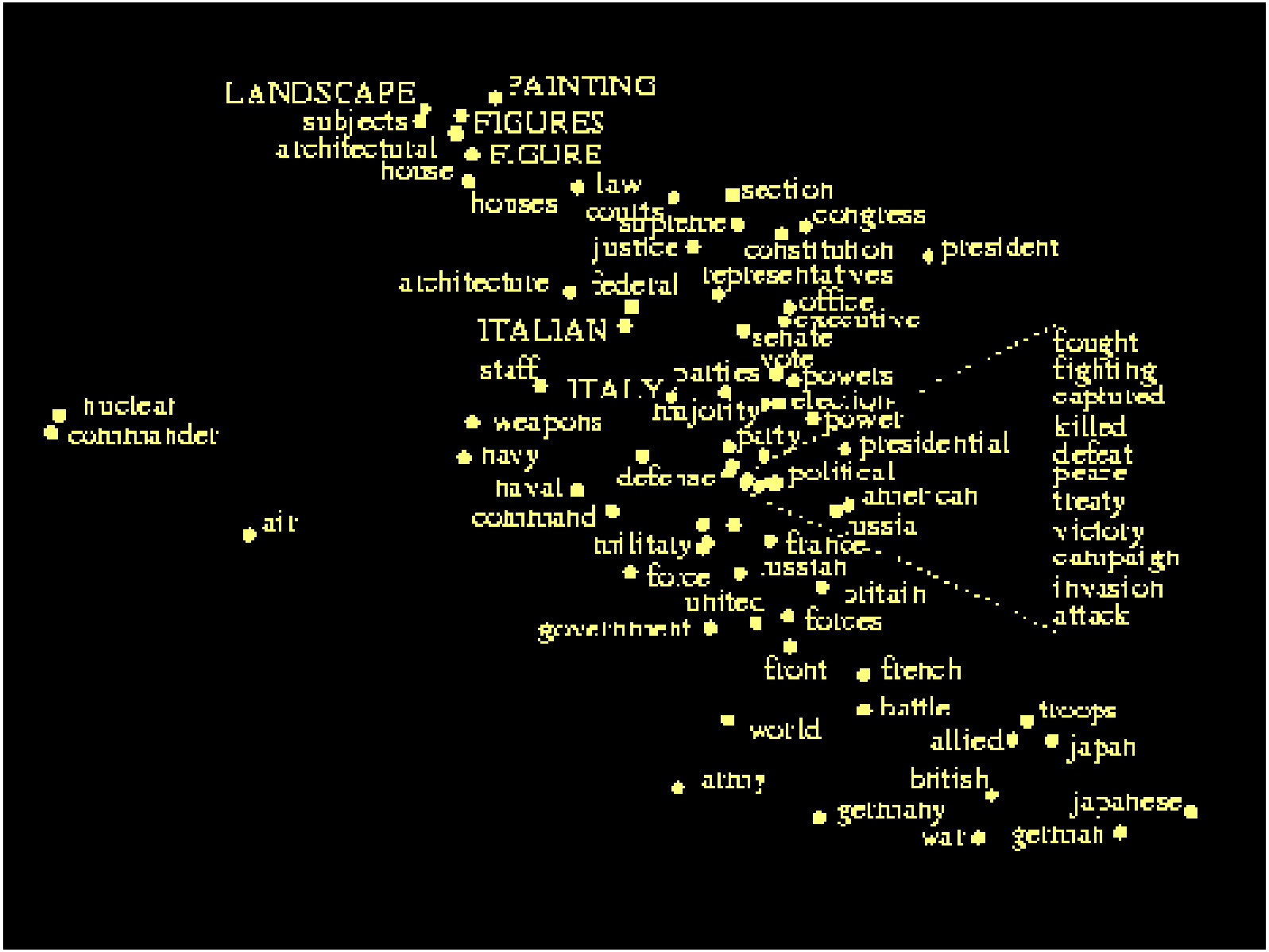
$$\sum_{i=1}^{N} ||z_i - \sum_j w_{i,j} z_j||^2$$

$$Minimize\ ((I - W)Z)'((I - W)Z) = Z'(I - W)'(I - W)Z$$

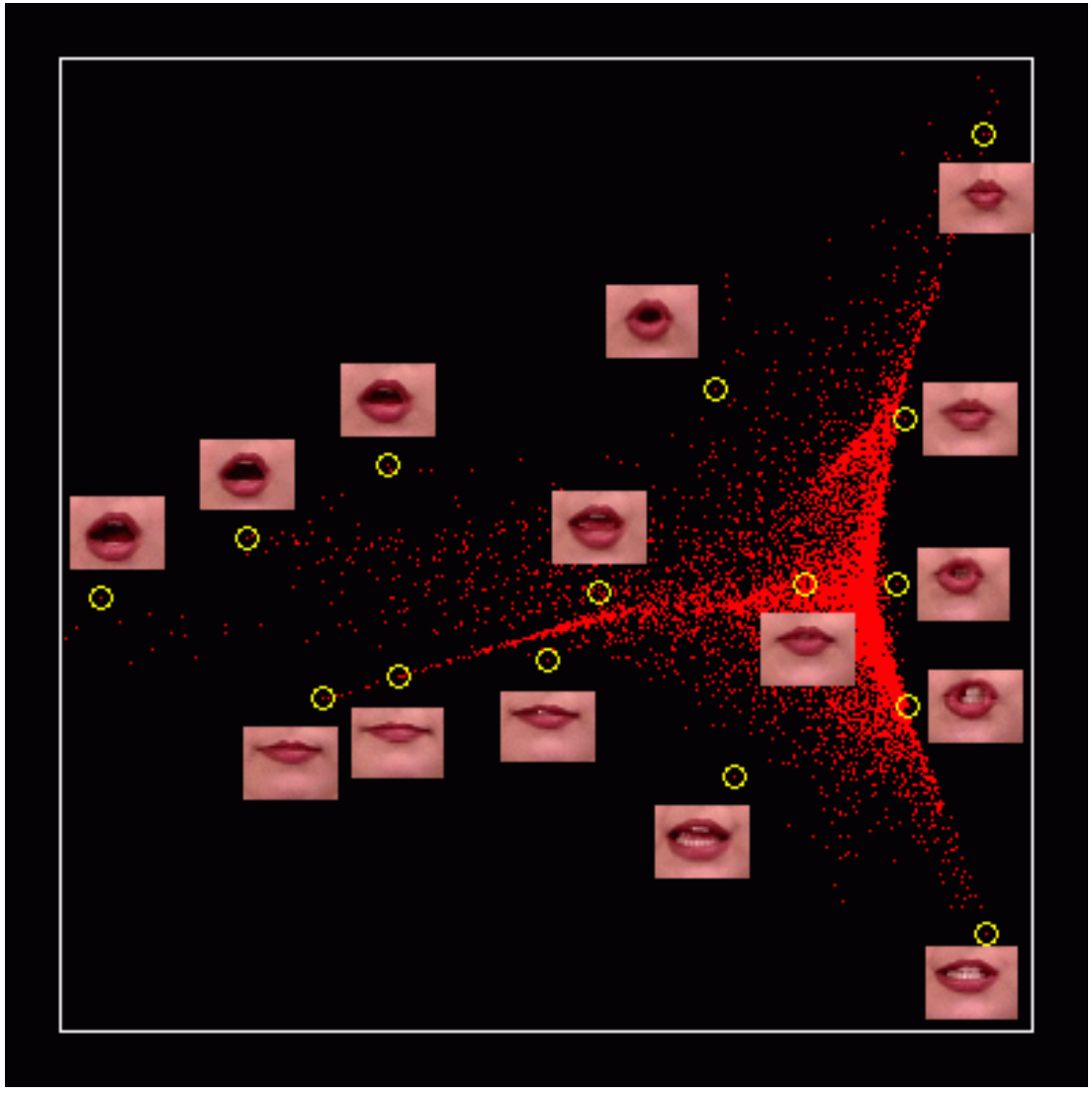
Solutions: vectors $z_i$ are eigenvectors of $(I - W)'(I - W)$

- Keeping the $d$ eigenvectors with lowest eigenvalues $> 0$

# Example, Texts



# Example, Images



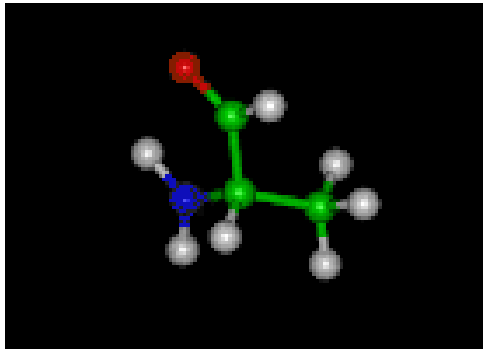
LLE

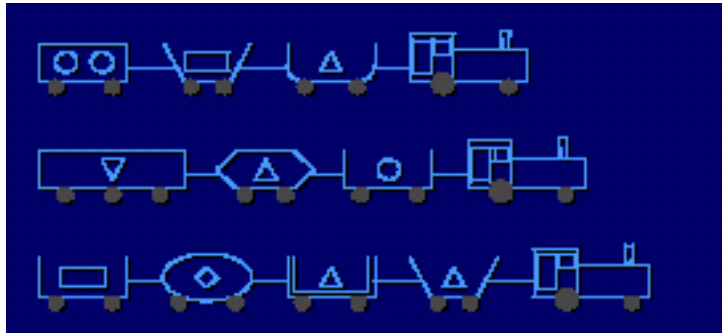
# Représentation pour l'apprentissage

- Sélection d'attributs
- Changements de représentation linéaires
- Changements de représentation non linéaires
- Propositionalisation
- Une étude de cas

# Propositionalization

## Relational domains





## Relational learning

PROS

Use domain knowledge

CONS

Covering test $\equiv$ subgraph matching

Inductive Logic Programming

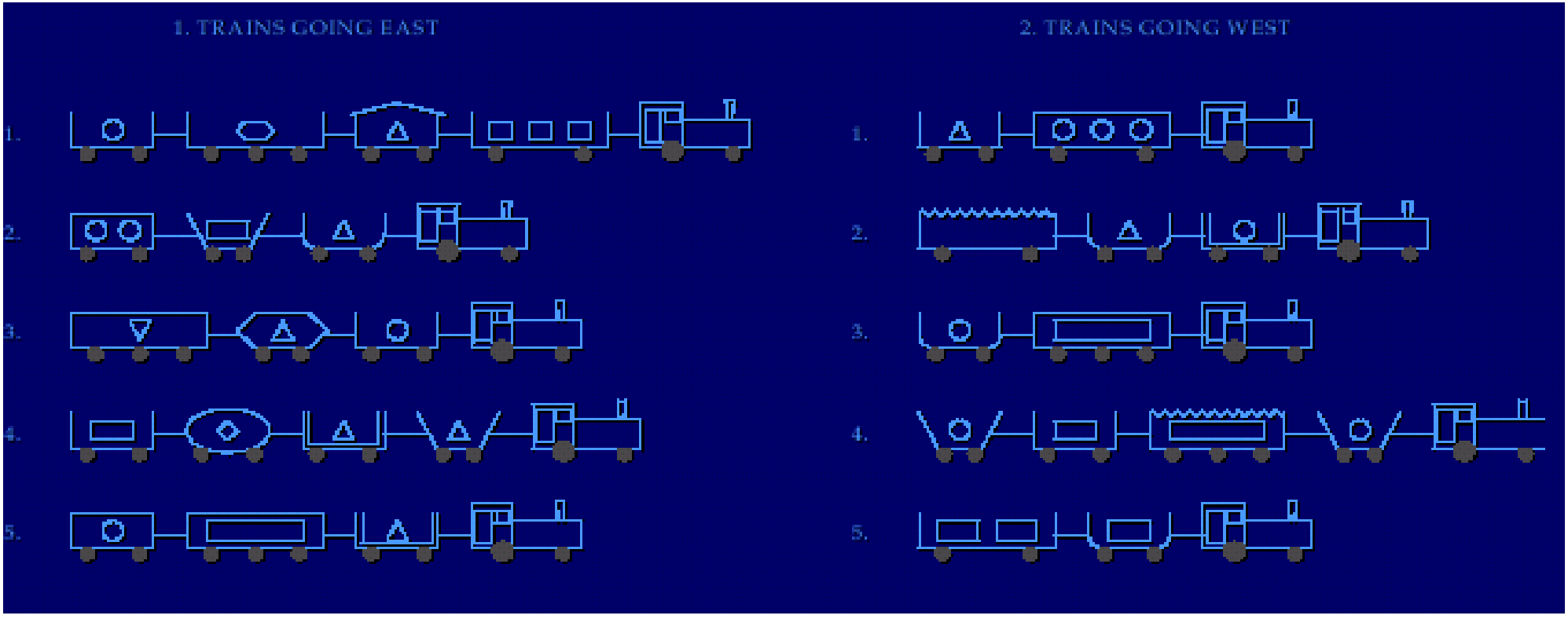
Data Mining

exponential complexity

Getting back to propositional representation: propositionalization

# West - East trains

Michalski 1983



# Propositionalization

## Linus (ancestor)

Lavrac et al, 94

$$West(a) \leftarrow \quad Engine(a,b), first\_wagon(a,c), roof(c), load(c,square,3)...$$
$$West(a') \leftarrow \quad Engine(a',b'), first\_wagon(a',c'), load(c',circle,1)...$$

| West | Engine(X) | First Wagon(X,Y) | Roof(Y) | $Load_1$ (Y) | $Load_2$ (Y) |
|---|---|---|---|---|---|
| a | b | c | yes | square | 3 |
| a' | b' | c' | no | circle | 1 |

Each column: a role predicate, where the predicate is determinate linked to former predicates (left columns) with a single instantiation in every example

# Propositionalization

## Stochastic propositionalization

Kramer, 98

Construct random formulas $\equiv$ boolean features

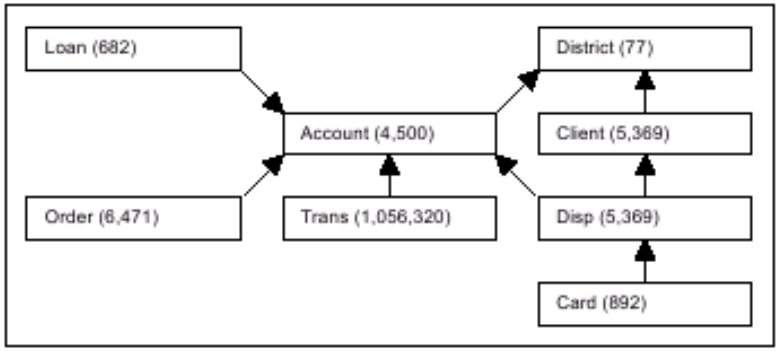
## SINUS – RDS

`http://www.cs.bris.ac.uk/home/rawles/sinus`

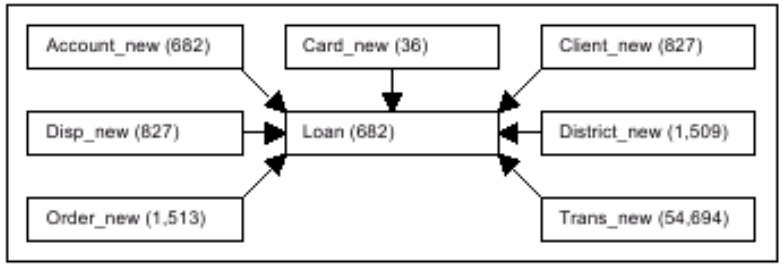
`http://labe.felk.cvut.cz/~zelezny/rsd`

- Use modes (user-declared) `modeb(2,hasCar(+train,-car))`
- Thresholds on number of variables, depth of predicates...
- Pre-processing (feature selection)

# Propositionalization




DB Schema




Propositionalization

## RELAGGS

Database aggregates

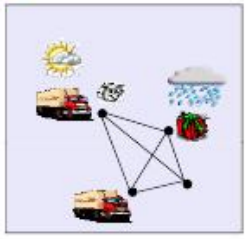
- average, min, max, of numerical attributes
- number of values of categorical attributes

# Apprentissage par Renforcement Relationnel

### Real Time Strategy Games

- Many objects of various types in complex interactions
- Good players can generalize across situations involving distinct object configurations

### The Logistics Domain

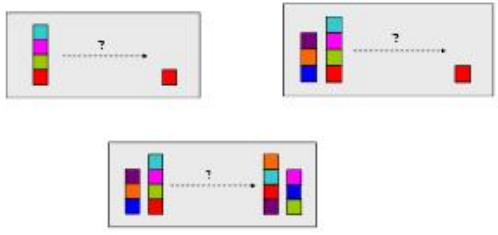


- Move many objects around with many other objects
- Identities and numbers of objects always changing

### Robot Soccer

- Reasoning about relationship between objects (players and ball) key to good play

### and of course ..... Blocksworld



- Would like a policy that is independent of number of objects/blocks

# Propositionalisation

## Contexte variable

- Nombre de robots, position des robots
- Nombre de camions, lieu des secours

## Besoin: Abstraire et Generaliser

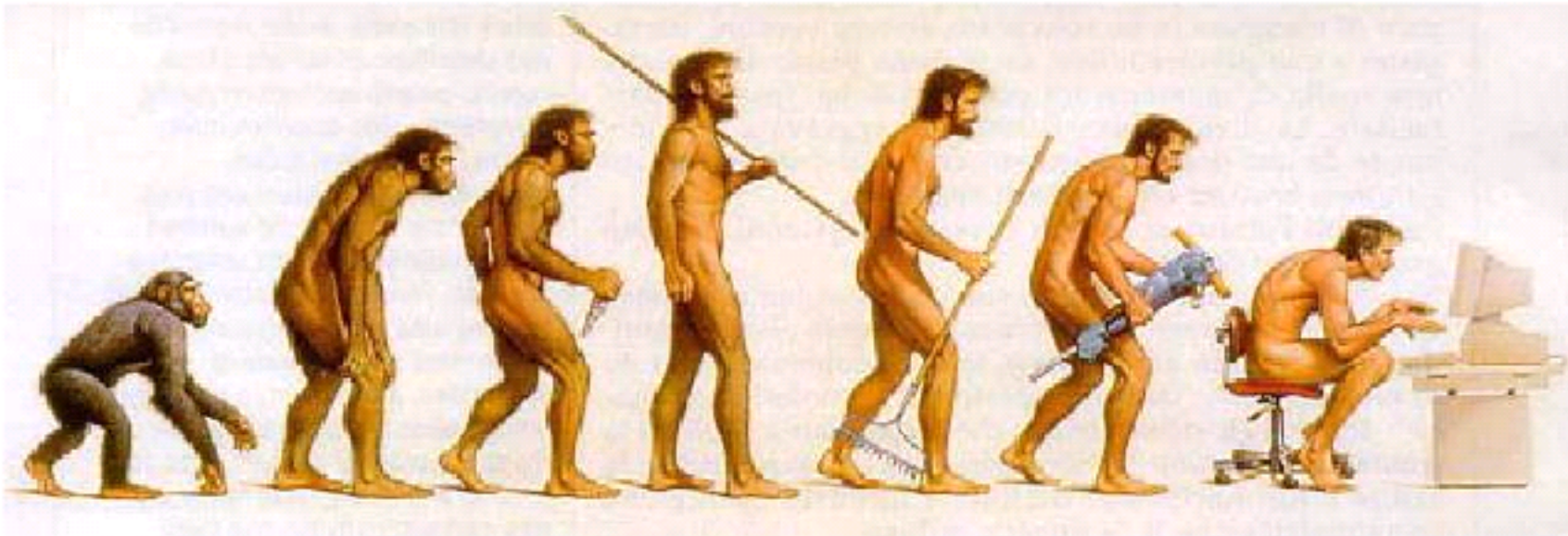
## Attributs

- Nombre d'amis/d'ennemis
- Distance du plus proche robot ami
- Distance du plus proche ennemi

# Représentation pour l'apprentissage

- Sélection d'attributs
- Changements de représentation linéaires
- Changements de représentation non linéaires
- Propositionalisation
- Une étude de cas

# Case study: Autonomic Computing



Considering current technologies, we expect that the total number of device administrators will exceed 220 millions by 2010.

Gartner 6/2001

in Autonomic Computing Wshop, ECML / PKDD 2006

Irina Rish & Gerry Tesauro.

# Autonomic Computing

## The need

- Main bottleneck of the deployment of complex systems: shortage of skilled administrators

## Vision

- Computing systems take care of the mundane elements of management by themselves.
- Inspiration: central nervous system (regulating temperature, breathing, and heart rate without conscious thought)

## Goal

Computing systems that manage themselves in accordance with high-level objectives from humans

Kephart & Chess, IEEE Computer 2003

# Autonomic Computing

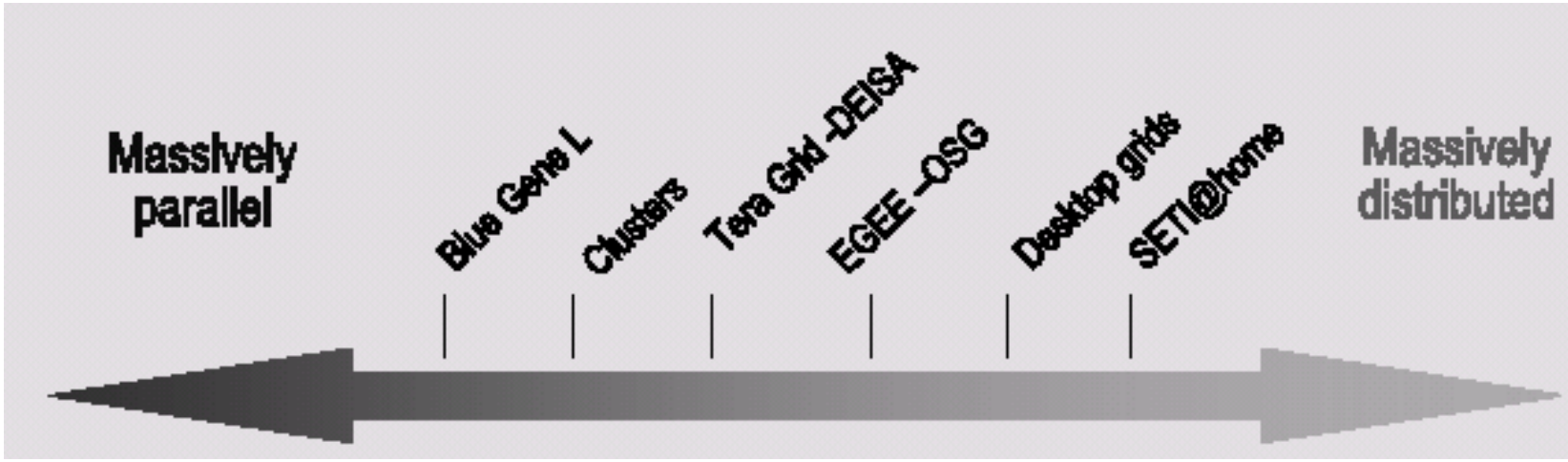
## Activity: A growing field

- IBM Manifesto for Autonomic Computing 2001
  http://www.research.ibm.com/autonomic
- ECML/PKDD Wshop on Autonomic Computing 2006
  http://www.ecmlpkdd2006.org/workshops.html
- JIC. on Measurement and Performance of Systems 2006
  http://www.cs.wm.edu/sigm06/
- NIPS Wshop on Machine Learning for Systems 2007
  http://radlab.cs.berkeley.edu/MLSys/
- Networked System Design and Implementation 2008
  http://www.usenix.org/events/nsdi08/

# Autonomic Grid System

- Grid Systems
  Presentation of EGEE, Enabling Grids for e-Science in Europe
- Acquiring the data
  The grid observatory
- Preparation of the data
  - Functional dependencies
  - Dimensionality reduction
  - Propositionalization

# Computing Systems: The landscape




## parallel

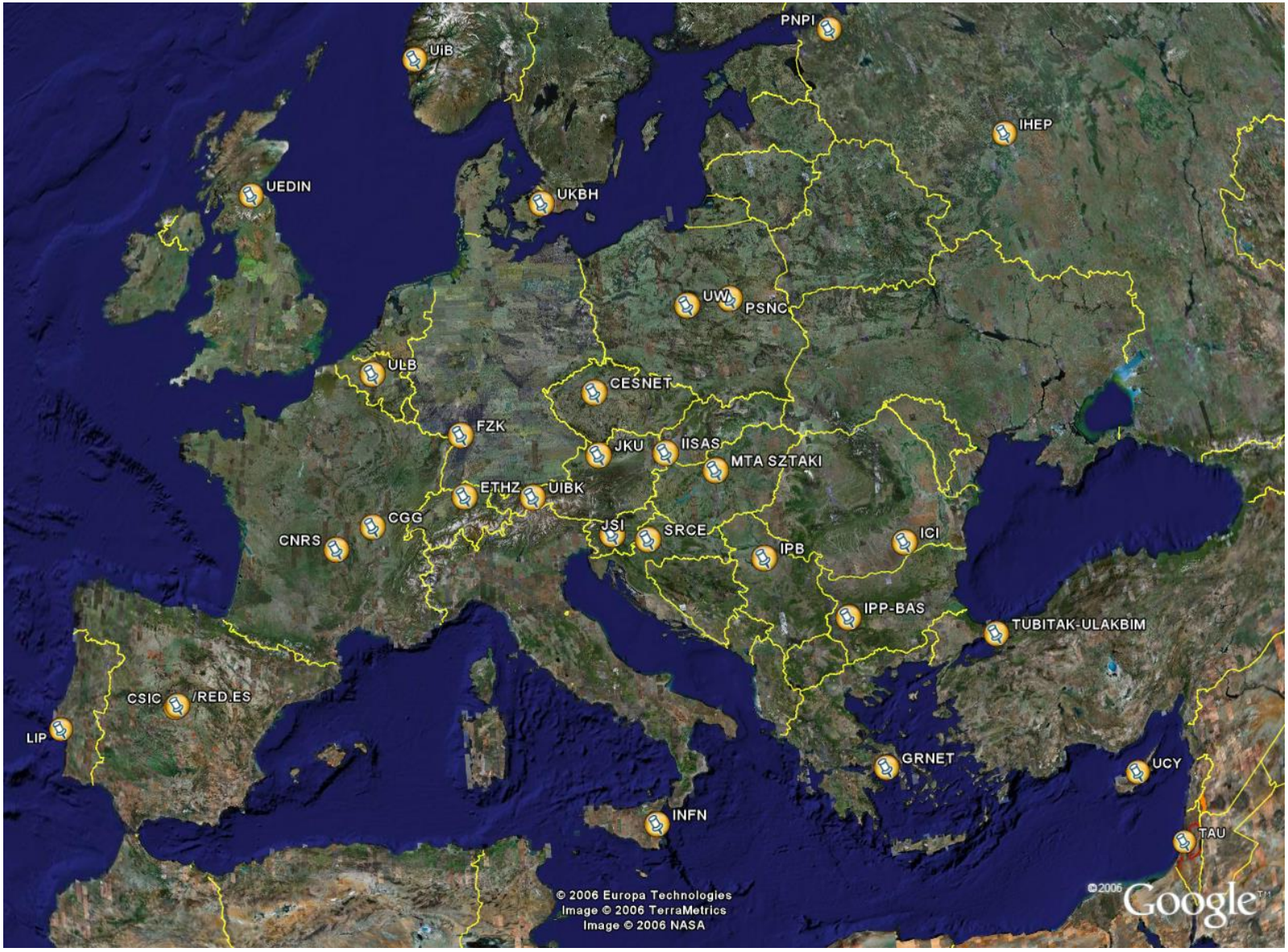
- homogeneous soft and hard
- resources
  - dedicated
  - static
  - controlled
- reduced software stack
- no built-in fault tolerance

## distributed

- heterogeneous soft and hard
- resources
  - shared
  - dynamic
  - aggregated
- middleware
- faults: the norm

# Storage and Computation have to be distributed

# EGEE: Enabling Grids for E-Science in Europe



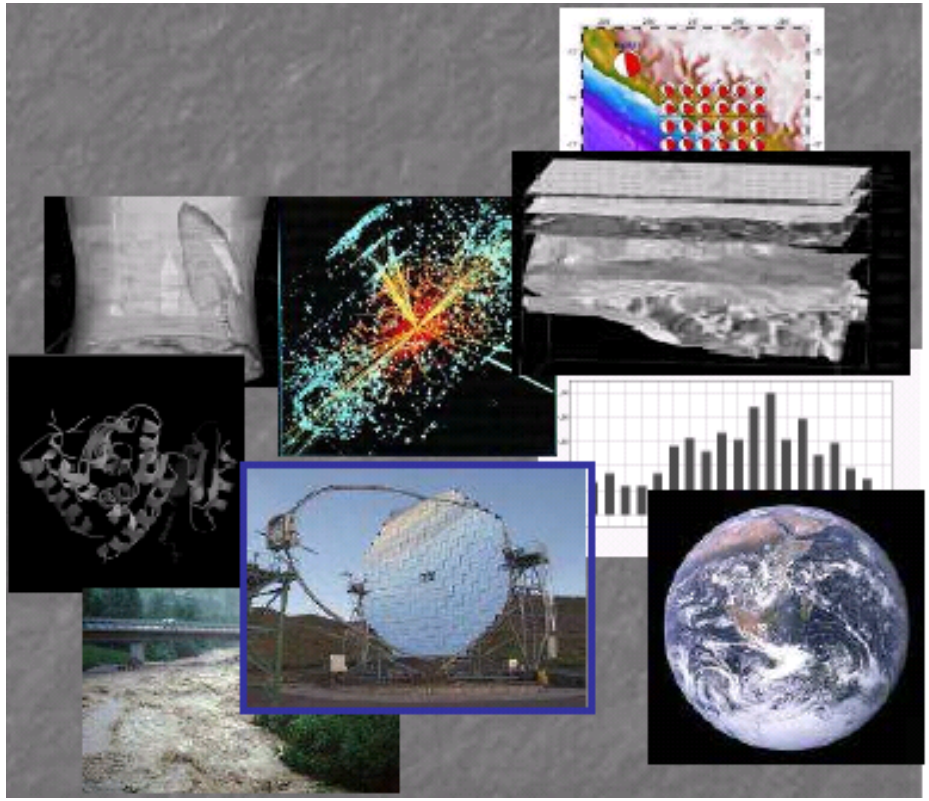
# EGEE, 2

- Infrastructure project started in 2001 $\rightarrow$ FP6 and FP7
- Large scale, production quality grid
- Core node: Lab. Accelerateur Linéaire, Université Paris-Sud
- 240 partners, 41,000 CPUs, all over the world
- 5 Peta bytes storage
- 24 $\times$ 7, 20 K concurrent jobs
- Web: www.eu-egee.org

Storage as important as CPU

# Applications

- High energy physics
- Life sciences
- Astrophysics
- Computational chemistry
- Earth sciences
- Financial simulation
- Fusion
- Multimedia
- Geophysics



# Autonomic Grid

## Requisite: The Grid Observatory

- Cluster in the EGEE-III proposal 2008-2010
- Data collection and publication: filtering, clustering

## Workload management

- Models of the grid dynamics
- Models of requirements and middleware reaction: time series and beyond
- Utility based-scheduling, local and global: MAB problem
- Policy evaluations: very large scale optimization

## Fault detection and diagnosis

- Categorization of failure modes from the Logging and Bookkeeping: feature construction, clustering,
- Abrupt changepoint detection

# Autonomic Grid: The Grid Observatory

## Data acquisition

- Data have not been stored with DM in mind — never
- Data [partially] automatically generated — here
  for EGEE services
  - redundant
  - little expert help

*It's no longer: the expert feeds the machine with data. Rather, machines feed machines...* *J. Gama*

## Data preprocessing

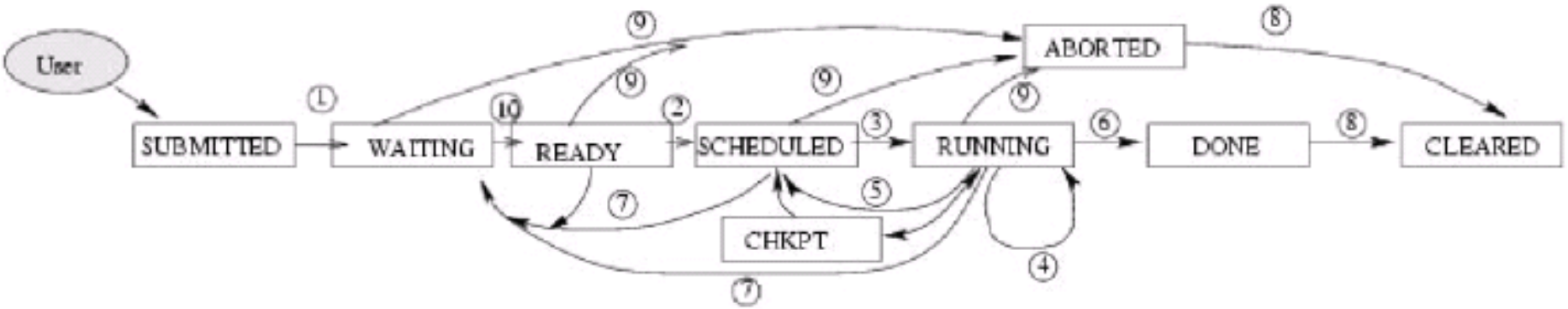
- 80% of the human cost
- Governs the quality of the output

# The grid system and the data

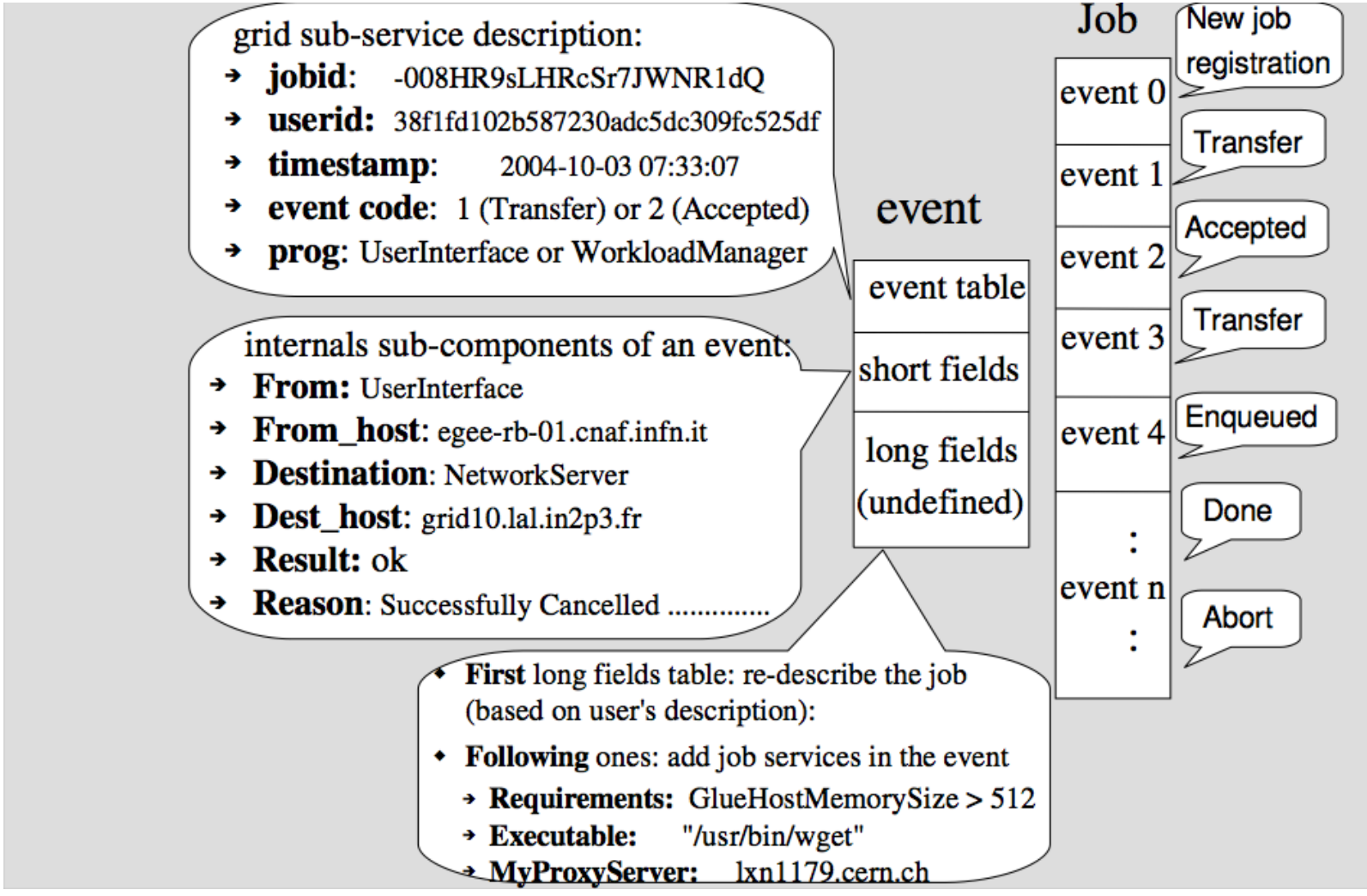
## The Workload Management System

- User Interface — User submits job description and requirements, and gets the results
- Resource Broker — Decides Computing Element
- Job Submission Service — Submits to CE and Checks
- Logging and Bookkeeping Service — Archive the data

## Job Lifecycle



# The data


grid sub-service description:
jobid: -008HR9sLHRcSr7JWNR1dQ
userid: 38f1fd102b587230adc5dc309fc525df
timestamp: 2004-10-03 07:33:07
event code: 1 (Transfer) or 2 (Accepted)
prog: UserInterface or WorkloadManager
internals sub-components of an event:
From: UserInterface
From_host: egee-rb-01.cnaf.infn.it
Destination: NetworkServer
Dest_host: grid10.lal.in2p3.fr
Result: ok
Reason: Successfully Cancelled .....
event
event table
short fields
long fields (undefined)
First long fields table: re-describe the job (based on user's description):
Following ones: add job services in the event
Requirements: GlueHostMemorySize > 512
Executable: "/usr/bin/wget"
MyProxyServer: lxn1179.cern.ch
Job
event 0
event 1
event 2
event 3
event 4
event n
New job registration
Transfer
Accepted
Transfer
Enqueued
Done
Abort

# Data Tables

## Events

| jobid | event | code | host | time_stamp | arrived | level |
|---|---|---|---|---|---|---|
| ---BrI1BgbIqkwtszqGfmA | 0 | 17 | atlfarm008.mi.infn.it | 2004-09-17 16:17:48 | 2004-09-17 16:17:49 | 8 |
| ---BrI1BgbIqkwtszqGfmA | 1 | 1 | atlfarm008.mi.infn.it | 2004-09-17 16:17:48 | 2004-09-17 16:17:49 | 8 |
| ---BrI1BgbIqkwtszqGfmA | 2 | 2 | lxb0728.cern.ch | 2004-09-17 16:17:53 | 2004-09-17 16:17:53 | 8 |
| ---BrI1BgbIqkwtszqGfmA | 3 | 4 | lxb0728.cern.ch | 2004-09-17 16:18:00 | 2004-09-17 16:18:01 | 8 |
| ---BrI1BgbIqkwtszqGfmA | 4 | 1 | atlfarm008.mi.infn.it | 2004-09-17 16:18:00 | 2004-09-17 16:18:01 | 8 |
| ---BrI1BgbIqkwtszqGfmA | 5 | 5 | lxb0728.cern.ch | 2004-09-17 16:18:01 | 2004-09-17 16:18:01 | 8 |

## Short Fields

| | | |
|---|---|---|
| 0 | JOBTYPE | SIMPLE |
| 0 | NS | lxb0728.cern.ch:7772 |
| 0 | NSUBJOBS | 0 |
| 0 | SEED | uLU0BArrdV98O41PLThJ5Q |
| 0 | SEQCODE | UI=000001:NS=0000000000:WM=000000:BH=0000000000:JSS=000000:LM=000000:LRMS=000000:APP=000000 |
| 0 | SRC_INSTANCE | |
| 1 | DESTINATION | NetworkServer |
| 1 | DEST_HOST | lxb0728.cern.ch |
| 1 | DEST_INSTANCE | lxb0728.cern.ch:7772 |
| 1 | DEST_JOBID | |
| 1 | REASON | |
| 1 | RESULT | START |
| 1 | SEQCODE | UI=000002:NS=0000000000:WM=000000:BH=0000000000:JSS=000000:LM=000000:LRMS=000000:APP=000000 |
| 1 | SRC_INSTANCE | |
| 2 | FROM | UserInterface |
| 2 | FROM_HOST | lxb0728.cern.ch |
| 2 | FROM_INSTANCE | |
| 2 | LOCAL_JOBID | |
| 2 | SEQCODE | UI=000003:NS=0000000001:WM=000000:BH=0000000000:JSS=000000:LM=000000:LRMS=000000:APP=000000 |
| 2 | SRC_INSTANCE | 7772 |
| 3 | QUEUE | /var/edgwl/workload_manager/input.fl |
| 3 | REASON | |
| 3 | RESULT | OK |
| 3 | SEQCODE | UI=000003:NS=0000000003:WM=000000:BH=0000000000:JSS=000000:LM=000000:LRMS=000000:APP=000000 |
| 3 | SRC_INSTANCE | |

# Data Tables

## Long Fields (4Gb)

```
+-----------------------+-------+------
| jobid                 | event | name | value
+-----------------------+-------+------
| ---BrI1BgbIqkwtszqGfmA |     0 | JDL  |[ requirements = ( ( ( ( Member("VO-atlas-lcg-release
-0.0.2",other.GlueHostApplicationSoftwareRunTimeEnvironment) ) && Member("VO-atlas-release
-8.0.5",other.GlueHostApplicationSoftwareRunTimeEnvironment) ) && ( other.GlueCEPolicyMaxCPUTime >= ( Member("LCG
-2\_1_0",other.GlueHostApplicationSoftwareRunTimeEnvironment) ? ( 36000000 / 60 ) : 36000000 ) / other.GlueHostBenchmarkSI00 ) ) &&
(other.GlueHostNetworkAdapterOutboundIP == true ) ) && ( other.GlueHostMainMemoryRAMSize >= 512 ); RetryCount = 0; edg_jobid =
 "https://lxb0728.cern.ch:9000/---BrI1BgbIqkwtszqGfmA"; Arguments = "dc2.003048.evgen.H4_170_WW._00002.pool.root
dc2.003048.simul.H4_170_WW._00208.pool.root.2 -6 6 50 350 208"; Environment = {
"LEXOR_WRAPPER_LOG=lexor_wrapper.log","LEXOR_STAGEOUT_MAXATTEMPT=5","LEXOR_STAGEOUT_INTERVAL=60",
"LEXOR_LCG_GFAL_INFOSYS=lxb2011.cern.ch:2170","LEXOR_T_RELEASE=8.0.5",
"LEXOR_T_PACKAGE=8.0.5.6/JobTransforms","LEXOR_T_BASEDIR=JobTransforms-08-00-05-06",
"LEXOR_TRANSFORMATION=share/
dc2.g4sim.trf","LEXOR_STAGEIN_LOG=dq_233387_stagein.log","LEXOR_STAGEIN_SCRIPT=dq_233387_stagein.sh",
"LEXOR_STAGEOUT_LOG=dq_233387_stageout.log","LEXOR_STAGEOUT_SCRIPT=dq_233387_stageout.sh" };
 MyProxyServer = "lxb0727.cern.ch"; JobType = "normal"; Executable =
"lexor_wrap.sh"; StdOutput = "dc2.003048.simul.H4_170_WW._00208.job.log.2"; OutputSandbox = {
"metadata.xml","lexor_wrapper.log","dq_233387_stagein.log","dq_233387_stageout.log",
"dc2.003048.simul.H4_170_WW._00208.job.log.2" }; VirtualOrganisation = "atlas";
 rank = ( other.GlueCEStateEstimatedResponseTime > 999 ) ?  -( other.GlueCEStateEstimatedResponseTime ) :  -(
other.GlueCEStateRunningJobs ); Type = "job"; StdError = "dc2.003048.simul.H4_170_WW._00208.job.log.2";
DefaultRank =  -other.GlueCEStateEstimatedResponseTime;
InputSandbox = {
"/home/negri/windmill-0.9.15/lexor/inputsandbox/lexor_wrap.sh",
"/home/negri/windmill-0.9.15/lexor/inputsandbox/dqlcg.py",
"/home/negri/windmill-0.9.15/lexor/inputsandbox/edgrmpi.sh",
"/home/negri/windmill-0.9.15/lexor/inputsandbox/dqrep.pl",
"/home/negri/windmill-0.9.15/lexor/inputsandbox/run_dqlcg.sh","/tmp/lexor/negri/dq_233387_stagein.sh",
"/tmp/lexor/negri/dq_233387_stageout.sh" } ]
+-----------------------+-------+------
```

# Preparation of the data

1. Functional dependencies
2. Dimensionality reduction — curse of dimensionality
   - Principal Component Analysis
   - Random Projection
   - Non linear Dimensionality Reduction
3. Propositionalization

# Functional dependency

## Definition

Given attributes $X$ and $X'$, $X'$ depends on $X$ on $\mathcal{E}$ ($X' \prec X$) iff

$$\exists f : dom(X') \mapsto dom(X) \; s.t. \; \forall i = 1 \ldots N, X(\mathbf{x}_i) = f(X'(\mathbf{x}_i))$$

## Examples

- $X'$ = City code, $X$ = City name
- $X'$ = Machine name, $X$ = IP
- $X'$ = Job ID, $X$ = User ID

## Why removing FD ?

- Curse of dimensionality
- Biased distance

# Functional dependency, 2

## Trivial cases

$$\#dom(X) = \#dom(X') = N \text{ number of examples}$$

## Algorithm

- Size:
$$(X' \prec X) \Rightarrow \#dom(X) \leq \#dom(X')$$
- Sample

  Repeat

  Select $v \in dom(X')$

  $\mathcal{E}_v =$ select $\mathbf{x}_i$ where $X'(\mathbf{x}_i) = v$

  Define $X(\mathcal{E}_v) = \{w \in dom(X), \exists x \in \mathcal{E}_v \ / \ X(x) = w\}$

  If $(\#X(\mathcal{E}_v) > 1)$ return false

  Until stop

  return true

# Going ubiquitous in Data Preparation

## Principles: same as usual

- Act locally
- Think globally

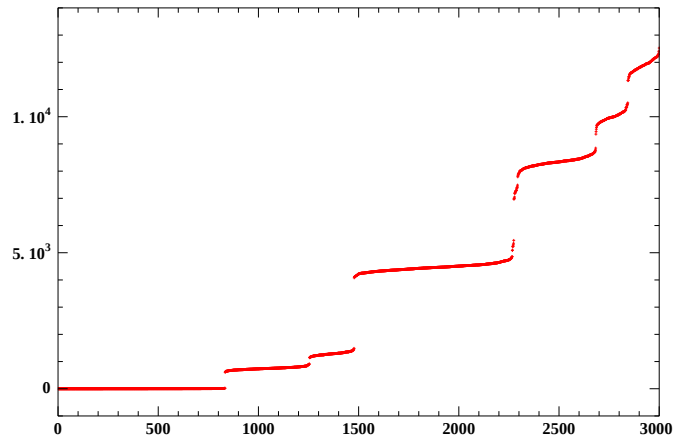
## The local level

- An ideal feature $\equiv$ a good hypothesis
- What is a promising hypothesis ?
  - Behaves well on (part of) the data
  - Is not trivial

# Going ubiquitous in Data Preparation, 2

## What is a good behaviour?

- Showing regularities
- Locally constant




## How to test triviality?

- Syntactical analysis:
  $xy - yx = 0$
- Statistical triviality:
  - Test on random data
  - Test on permutations of the data

# Going ubiquitous in Data Preparation, 3

## Internally: an optimization problem

- Define bins
- Compute histogram, associated quantity of information
- Compare histograms on real data / on random data

## Externally: an optimization problem

- Upon receiving a new feature
- Check whether this is relevant to your data
- Check whether this brings new information