### M1 – Apprentissage

### Michèle Sebag – Benoit Barbot LRI – LSV

27 janvier 2014

▲□▶ ▲圖▶ ▲圖▶ ▲圖▶ = ● ● ●

# Clustering

### Input

$$\mathcal{E} = {\mathbf{x}_1, \ldots, \mathbf{x}_n} \sim P(x)$$

### Output

- Models
- Clusters
- Representatives

### Assumptions, contexts

Clusters are separated by a low-density region





æ

## **Motivations**

Compression.

Ex, vector quantization in images.



▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

- Divide and conquer; preliminary for classification.
  Ex, different types of diseases.
- Check data.

### Overview

### Clustering

K-Means Generative models Expectation Maximization Selecting the number of clusters Stability

Axiomatisation

Data Streaming

Application: Autonomic Computing

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

Learning distances

# **Clustering Questions**

### Hard or soft ?

- Hard: find a partition of the data
- Soft: estimate the distribution of the data as a mixture of components.



### Parametric vs non Parametric ?

- Parametric: number K of clusters is known
- Non-Parametric: find K

(wrapping a parametric clustering algorithm)

#### Caveat:

- Complexity
- Outliers
- Validation

## **Formal Background**

#### Notations

Ε	$\{x_1,\ldots,$	<b>x</b> <sub>N</sub> }	dataset
---	-----------------	-------------------------	---------

- *N* number of data points
- K number of clusters given or optimized
- $C_k$  k-th cluster Hard clustering
- $\tau(i)$  index of cluster containing  $\mathbf{x}_i$
- Soft clustering

- $\begin{array}{ll} f_k & k\text{-th model} \\ \gamma_k(i) & \Pr(\mathbf{x}_i \sim f_k) \end{array}$
- **Solution** Hard Clustering Partition  $\Delta = (C_1, \dots, C_k)$ Soft Clustering  $\forall i \sum_k \gamma_k(i) = 1$

## Formal Background, 2

**Quality / Cost function** Measures how well the clusters characterize the data

- (log)likelihood
- dispersion

soft clustering hard clustering

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

$$\sum_{k=1}^{K} \frac{1}{|C_k|^2} \sum_{\mathbf{x}_i, \mathbf{x}_j \text{ in } C_k} d(\mathbf{x}_i, \mathbf{x}_j)^2$$

### Formal Background, 2

**Quality / Cost function** Measures how well the clusters characterize the data

- (log)likelihood
- dispersion

soft clustering hard clustering

$$\sum_{k=1}^{K} \frac{1}{|C_k|^2} \sum_{\mathbf{x}_i, \mathbf{x}_j \text{ in } C_k} d(\mathbf{x}_i, \mathbf{x}_j)^2$$

**Tradeoff** Quality increases with  $K \Rightarrow$  Regularization needed to avoid one cluster per data point

#### Exercize

$$\sum_{k=1}^{K} \frac{1}{|C_k|^2} \sum_{\mathbf{x}_i, \mathbf{x}_j \text{ in } C_k} ||\mathbf{x}_i - \mathbf{x}_j||^2 = \sum_{k=1}^{K} \frac{1}{|C_k|^2} \sum_{\mathbf{x}_i, \mathbf{x}_j \text{ in } C_k} ||\mathbf{x}_i - \bar{\mathbf{x}}_k||^2$$

with  $\bar{\mathbf{x}}_k$  = average  $\mathbf{x}_i, \mathbf{x}_i \in C_k$ .

(日)、(型)、(E)、(E)、(E)、(O)へ(C)

## **Clustering vs Classification**

Marina Maila

Marina Mella		http://videoiectures.net/	
	Classification	Clustering	
К	# classes (given)	# clusters (unknown)	
Quality	Generalization error	many cost functions	
Focus on	Test set	Training set	
Goal	Prediction	Interpretation	
Analysis	discriminant	exploratory	
Field	mature	new	

▲□▶ ▲圖▶ ▲圖▶ ▲圖▶ / 圖 / のへで

http://wideolestumes.met/

# **Non-Parametric Clustering**

#### **Hierarchical Clustering**

### Principle

- agglomerative (join nearest clusters)
- divisive (split most dispersed cluster)



### Algorithm

Init: Each point is a cluster (*n* clusters)

Loop

Select two most similar clusters

Merge them

Until there is only 1 cluster

### Hierarchical Clustering, example



790

#### Key point 1: choice of distance



< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <



Variable Index

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 - のへで

### Key point 2: choice of aggregation

Compute distance between two clusters

- Complete linkage:Largest distance between points
- Single linkage: Smallest distance between points
- Average linkage: Average distance between points
- Centroid: distance between centroids of the points

Centroid of points: point closest to their average.



## **Parametric Clustering**

Parametric: K is known

### Algorithms based on distances

- K-means
- graph / cut

#### Algorithms based on models

Mixture of models: EM algorithm

### Overview

### Clustering

K-Means Generative models Expectation Maximization Selecting the number of clusters Stability

Axiomatisation

Data Streaming

Application: Autonomic Computing

Learning distances

## K-Means

### Algorithm

1. Init:

Uniformly draw K points  $\mathbf{x}_{i_j}$  in  $\mathcal{E}$ Set  $C_j = {\mathbf{x}_{i_j}}$ 

2. Repeat

3. Draw without replacement  $\mathbf{x}_i$  from  $\mathcal{E}$ 

4. 
$$\tau(i) = \operatorname{argmin}_{k=1...K} \{ \mathbf{d}(\mathbf{x}_i, \mathbf{C}_k) \}$$
 find best cluster for  $\mathbf{x}_i$ 

5. 
$$C_{\tau(i)} = C_{\tau(i)} \bigcup \mathbf{x}_i$$
 add  $\mathbf{x}_i$  to  $C_{\tau(i)}$ 

6. Until all points have been drawn

7. If partition  $C_1 \dots C_K$  has changed **Stabilize** Define  $\mathbf{x}_{i_k} =$  **best point** in  $C_k$ ,  $C_k = \{x_{i_k}\}$ , goto 2.

### **Algorithm terminates**

## K-Means, Knobs

**Knob 1 : define**  $d(\mathbf{x}_i, C_k)$ 

- $min\{d(\mathbf{x}_i,\mathbf{x}_j),\mathbf{x}_j\in C_k\}$
- average $\{d(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_j \in C_k\}$
- $max\{d(\mathbf{x}_i,\mathbf{x}_j),\mathbf{x}_j\in C_k\}$

### Knob 2 : define "best" in $C_k$

- Medoid
- \* Average

(does not belong to  $\mathcal{E}$ )

#### favors

long clusters compact clusters spheric clusters

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

$$\begin{aligned} \underset{i \in C_k}{\operatorname{argmin}} \{\sum_{\mathbf{X}_j \in C_k} d(\mathbf{x}_i, \mathbf{x}_j)\} \\ \frac{1}{|C_k|} \sum_{\mathbf{X}_j \in C_k} \mathbf{x}_j \end{aligned}$$

### No single best choice



FIG. 1. Optimizing the diameter produces B while A is clearly more desirable.



FIG. 2. The inferior clustering B is found by optimizing the 2-median measure.

## K-Means, Discussion

### PROS

- **Complexity**  $\mathcal{O}(K \times N)$
- Can incorporate prior knowledge

initialization

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

### CONS

- Sensitive to initialization
- Sensitive to outliers
- Sensitive to irrelevant attributes

## *K*-Means, Convergence

#### For cost function

$$\mathcal{L}(\Delta) = \sum_{k} \sum_{i,j \neq \tau(i) = \tau(j) = k} d(\mathbf{x}_i, \mathbf{x}_j)$$

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへぐ

▶ for 
$$d(\mathbf{x}_i, C_k) =$$
 average  $\{d(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_j \in C_k\}$ 

▶ for "best" in 
$$C_k$$
 = average of  $\mathbf{x}_j \in C_k$ 

K-means converges toward a (local) minimum of  $\mathcal{L}$ .

## **K-Means**, **Practicalities**

### Initialization

- Uniform sampling
- Average of  $\mathcal{E}$  + random perturbations
- Average of  $\mathcal{E}$  + orthogonal perturbations
- Extreme points: select  $\mathbf{x}_{i_1}$  uniformly in  $\mathcal{E}$ , then

Select 
$$x_{i_j} = argmax\{\sum_{k=1}^{j} d(\mathbf{x}_i, x_{i_k})\}$$

### **Pre-processing**

Mean-centering the dataset

### Overview

### Clustering

K-Means Generative models Expectation Maximization Selecting the number of clusters Stability

Axiomatisation

Data Streaming

Application: Autonomic Computing

Learning distances

## **Model-based clustering**

### **Mixture of components**

- Density  $f = \sum_{k=1}^{K} \pi_k f_k$
- *f<sub>k</sub>*: the *k*-th component of the mixture

• 
$$\gamma_k(i) = \frac{\pi_k f_k(x)}{f(x)}$$

• induces  $C_k = \{\mathbf{x}_j \mid k = argmax\{\gamma_k(j)\}\}$ 

### Nature of components: prior knowledge

- Most often Gaussian:  $f_k = (\mu_k, \Sigma_k)$
- Beware: clusters are not always Gaussian...



### Model-based clustering, 2

#### Search space

• Solution : 
$$(\pi_k, \mu_k, \Sigma_k)_{k=1}^K = \theta$$

### Criterion: log-likelihood of dataset

$$\ell(\theta) = \log(\Pr(\mathcal{E})) = \sum_{i=1}^{N} \log \Pr(\mathbf{x}_i) \propto \sum_{i=1}^{N} \sum_{k=1}^{K} \log(\pi_k f_k(\mathbf{x}_i))$$

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

to be maximized.

## Model-based clustering with EM

### Formalization

- Define  $z_{i,k} = 1$  iff  $\mathbf{x}_i$  belongs to  $C_k$ .
- $E[z_{i,k}] = \gamma_k(i)$  prob. **x**<sub>i</sub> generated by  $\pi_k f_k$
- Expectation of log likelihood

$$E[\ell(\theta)] \propto \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_i(k) \log(\pi_k f_k(\mathbf{x}_i))$$
$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_i(k) \log \pi_k + \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_i(k) \log f_k(\mathbf{x}_i)$$

#### **EM optimization**

E step Given  $\theta$ , compute

$$\gamma_k(i) = \frac{\pi_k f_k(\mathbf{x}_i)}{f(x)}$$

M step Given  $\gamma_k(i)$ , compute

$$\theta^* = (\pi_k, \mu_k, \Sigma_k)^* = \operatorname{argminE}[\ell(\theta)]$$

### **Maximization step**

 $\pi_k$ : Fraction of points in  $C_k$ 

$$\pi_k = \frac{1}{N} \sum_{i=1}^N \gamma_k(i)$$

 $\mu_k$ : Mean of  $C_k$ 

$$\mu_k = \frac{\sum_{i=1}^N \gamma_k(i) \mathbf{x}_i}{\sum_{i=1}^N \gamma_k(i)}$$

 $\Sigma_k$ : Covariance

$$\Sigma_k = \frac{\sum_{i=1}^N \gamma_k(i)(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)'}{\sum_{i=1}^N \gamma_k(i)}$$

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

### Overview

### Clustering

K-Means Generative models Expectation Maximization Selecting the number of clusters Stability

Axiomatisation

Data Streaming

Application: Autonomic Computing

Learning distances

## Choosing the number of clusters

K-means constructs a partition whatever the K value is. Selection of K

Bayesian approaches

Tradeoff between accuracy / richness of the model

Stability

Varying the data should not change the result

Gap statistics

Compare with null hypothesis: all data in same cluster.

### **Bayesian approaches**

#### **Bayesian Information Criterion**

$${\it BIC}( heta) = \ell( heta) - rac{\# heta}{2} \log N$$

Select  $K = \operatorname{argmax} BIC(\theta)$ where  $\#\theta = \operatorname{number}$  of free parameters in  $\theta$ :

 $\blacktriangleright$  if all components have same scalar variance  $\sigma$ 

$$\#\theta = K - 1 + 1 + Kd$$

• if each component has a scalar variance  $\sigma_k$ 

$$\#\theta = K - 1 + K(d+1)$$

• if each component has a full covariance matrix  $\Sigma_k$ 

$$\#\theta=K-1+K(d+d(d-1)/2)$$

### **Gap statistics**

### Principle: hypothesis testing

- 1. Consider hypothesis  $H_0$ : there is no cluster in the data.  $\mathcal{E}$  is generated from a no-cluster distribution  $\pi$ .
- Estimate the distribution f<sub>0,K</sub> of L(C<sub>1</sub>,...C<sub>K</sub>) for data generated after π. Analytically if π is simple Use Monte-Carlo methods otherwise
- 3. Reject  $H_0$  with confidence  $\alpha$  if the probability of generating the true value  $\mathcal{L}(C_1, \ldots, C_K)$  under  $f_{0,K}$  is less than  $\alpha$ .

Beware: the test is done for all K values...

### Gap statistics, 2

**Algorithm** Assume  $\mathcal{E}$  extracted from a no-cluster distribution, e.g. a single Gaussian.

- 1. Sample  ${\mathcal E}$  according to this distribution
- 2. Apply K-means on this sample
- 3. Measure the associated loss function

Repeat : compute the average  $\overline{\mathcal{L}}_0(K)$  and variance  $\sigma_0(K)$ Define the gap:

$$Gap(K) = \overline{\mathcal{L}}_0(K) - \mathcal{L}(C_1, \dots C_K)$$

Rule Select min K s.t.

$$Gap(K) \ge Gap(K+1) - \sigma_0(K+1)$$

What is nice: also tells if there are no clusters in the data...

### Overview

### Clustering

K-Means Generative models Expectation Maximization Selecting the number of clusters Stability

Axiomatisation

Data Streaming

Application: Autonomic Computing

Learning distances

# Stability

### Principle

- Consider  $\mathcal{E}'$  perturbed from  $\mathcal{E}$
- Construct  $C'_1, \ldots C'_K$  from  $\mathcal{E}'$
- ▶ Evaluate the "distance" between  $(C_1, \ldots, C_K)$  and  $(C'_1, \ldots, C'_K)$
- ▶ If small distance (stability), K is OK

### **Distortion** $D(\Delta)$

$$D(\Delta) = \sum_{i} ||\mathbf{x}_i - \mu_{\tau(i)}||^2 = tr(S) - tr(X'SX)$$

Minimal distortion  $D^* = tr(S) - \sum_{k=1}^{K-1} \lambda_k$ 

# Stability, 2

### Results

- $\Delta$  has low distortion  $\Rightarrow (\mu_1, \dots \mu_K)$  close to space  $(v_1, \dots v_K)$ .
- $\Delta_1$ , and  $\Delta_2$  have low distortion  $\Rightarrow$  "close"
- (and close to "optimal" clustering)

Meila ICML 06

### **Counter-example**




## Overview

#### Clustering

K-Means Generative models Expectation Maximization Selecting the number of clusters Stability

#### Axiomatisation

Data Streaming

Application: Autonomic Computing

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

Learning distances

## Kleinberg's axiomatic framework for clustering

Kleinberg 2002 Given  $\mathcal{E} = {\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{x}_i \in X}$ , a clustering builds a partition  $\Gamma$ depending on distance d. Let denote  $\Gamma = f(d)$ .

$$\left(\begin{array}{rrrr}1 & 10 & 10\\ 10 & 0 & 1\\ 10 & 1 & 0\end{array}\right)$$

 $\Gamma = (\{1\}, \{2, 3\}).$ 

# Kleinberg's axiomatic framework for clustering Properties

Scale invariance

$$\forall \alpha > 0, f(\alpha d) = f(d)$$

#### Richness

$$Range(f) =$$
 Power set of  $\mathcal{E}$ 

#### Consistency

If  $f(d) = \Gamma$  and d' is a  $\Gamma$ -enhancing transformation of d, then

$$f(d') = \Gamma$$

くしゃ ( 雪 ) ( 雪 ) ( 雪 ) ( 雪 ) ( 雪 ) ( 雪 ) (

where d' is  $\Gamma$ -enhancing if

•  $d'(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_j)$  if  $\mathbf{x}_i$  and  $\mathbf{x}_j$  in same cluster of  $\Gamma$ 

• 
$$d'(\mathbf{x}_i,\mathbf{x}_j) \geq d(\mathbf{x}_i,\mathbf{x}_j)$$
 otherwise

## **Examples**

#### Run single linkage till you get k clusters

Scale invariance Yes, consistency Yes, richness No

**Run single linkage while distances**  $\leq c \cdot max_{i,j}d(\mathbf{x}_i, \mathbf{x}_j)$ , c > 0

## **Examples**

#### Run single linkage till you get k clusters

Scale invariance Yes, consistency Yes, richness No

**Run single linkage while distances**  $\leq c \cdot max_{i,j}d(\mathbf{x}_i, \mathbf{x}_j)$ , c > 0

Scale invariance Yes, consistency No, richness Yes

Run single linkage until distances  $\leq$  some threshold r

## **Examples**

#### Run single linkage till you get k clusters

Scale invariance Yes, consistency Yes, richness No

**Run single linkage while distances**  $\leq c \cdot max_{i,j}d(\mathbf{x}_i, \mathbf{x}_j)$ , c > 0

Scale invariance Yes, consistency No, richness Yes

Run single linkage until distances  $\leq$  some threshold r

Scale invariance No, consistency Yes, richness Yes

# Impossibility result

#### Thm

- There is no consistent way of choosing a level of granularity
- There exists no f satisfying all three axioms



▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

## Overview

#### Clustering

K-Means Generative models Expectation Maximization Selecting the number of clusters Stability

#### Axiomatisation

#### Data Streaming

Application: Autonomic Computing

▲□▶ ▲圖▶ ★ 国▶ ★ 国▶ - 国 - のへで

#### Learning distances

# Part 2. Data Streaming

- When: data, specificities
- What: goals
- How: algorithms

More: see Joao Gama's tutorial,

http://wiki.kdubiq.org/summerschool2008/index.php/Main/Materials

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

## **Motivations**



Electric Power Network

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへぐ

## Data

#### Input

- Continuous flow of (possibly corrupted) data, high speed
- Huge number of sensors, variable along time (failures)
- Spatio-temporal data

### Output

- Cluster: profiles of consumers
- Prediction: peaks of demand
- Monitor Evolution: Change detection, anomaly detection

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

## Where is the problem ?

Standard Data Analysis

- Select a sample
- Generate a model (clustering, neural nets, ...)

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

## Where is the problem ?

Standard Data Analysis

- Select a sample
- Generate a model (clustering, neural nets, ...)

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

Does not work ...

- World is not static
- Options, Users, Climate, ... change

# **Specificities of data**

#### Domain

- Radar: meteorological observations
- Satellite: images, radiation
- Astronomical surveys: radio
- Internet: traffic logs, user queries, ...
- Sensor networks
- Telecommunications

#### Features

- Most data never seen by humans
- Need for REAL-TIME monitoring, (intrusion, outliers, anomalies,,,)

NB: Beyond ML scope: data are not iid (independent identically distributed)

## **Data streaming Challenges**

#### Maintain Decision Models in real-time

incorporate new information

comply with speed

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

- forget old/outdated information
- detect changes and adapt models accordingly

Unbounded training sets Prefer fast approximate answers...

- Approximation: Find answer with factor  $1 \pm \epsilon$
- Probably correct:  $Pr(answer correct) = 1 \delta$
- PAC:  $\epsilon, \delta$  (Probably Approximately Correct)
- Space  $\approx \mathcal{O}(1/\epsilon^2 \log(1/\delta))$

# Data Mining vs Data Streaming

	Traditional	Stream
Nr. of Passes	Multiple	Single
Processing Time	Unlimited	Restricted
Memory Usage	Unlimited	Restricted
Type of Result	Accurate	Approximate
Distributed	No	Yes

## What: queries on a data stream

#### Sample

- Count number of distinct values / attribute
- Estimate sliding average (number of 1's in a sliding window)
- Get top-k elements

#### Application: Compute entropy of the stream

$$H(x) = \sum p_i \log_2(p_i)$$

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

useful to detect anomalies

# Sampling

Uniform sampling: each one out of n examples is sampled with probability 1/n. What if we don't know the size ? Standard

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

- Sample instances at periodic time intervals
- Loss of information

#### **Reservoir Sampling**

- Create buffer size k
- Insert first k elements
- Insert *i*-th element with probability k/i
- Delete a buffer element at random

#### Limitations

- Unlikely to detect changes/anomalies
- Hard to parallelize

## **Count number of values**

#### Problem

Domain of the attribute is  $\{1, \ldots, M\}$ Piece of cake if memory available... What if the memory available is log(M) ? Flajolet-Martin 1983

Based on hashing:  $\{1, \ldots, M\} \mapsto \{0, \ldots, 2^L\}$  with L = log(M).

 $x \rightarrow hash(x) = y \rightarrow position least significant bit, lsb(x)$ 

## Count number of values, followed



Result

R = position of rightmost 0 in H $M \approx 2^R / .7735$ 

## **Decision Trees for Data Streaming**

#### Principle

Grow the tree if evidence best attribute > second best

#### Algorithm

parameter: confidence  $\delta$  (user-defined)

While true

Read example, propagate until a leaf

If enough examples in leaf

Compute IG for all attributes;

 $\begin{aligned} \epsilon &= \sqrt{\frac{R^2 \ln(1/\delta)}{2n}}\\ \text{Keep best if IG(best) - IG(second best }) > \epsilon \end{aligned}$ 

Mining High Speed Data Streams, Pedro Domingos, Geoffrey Hulten, KDD-00

# **Open issues**

#### What's new

Forget about iid;

Forget about more than linear complexity (and log space)

▲ロト ▲帰ト ▲ヨト ▲ヨト 三日 - の々ぐ

#### Challenges

Online, Anytime algs Distributed alg. Criteria of performance Integration of change detection

## Overview

#### Clustering

K-Means Generative models Expectation Maximization Selecting the number of clusters Stability

Axiomatisation

Data Streaming

Application: Autonomic Computing

◆□▶ ◆□▶ ◆三▶ ◆三▶ 三三 のへぐ

Learning distances

# **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.

・ロト ・聞ト ・ヨト ・ヨト

# **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

# **Toward Autonomic Grid**

#### EGEE, Enabling Grids for E-sciencE 2001-2011

- 50 countries
- 300 sites
- 180,000 CPUs
- 5Petabytes storage
- 10,000 users
- 300,000 jobs/ day



#### http://public.eu-egee.org/

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <

#### EGEE-III : WP Grid Observatory

- Job scheduling
- Job profiling

# **Data Streaming for Job Profiling**

X. Zhang, C. Furtlehner, M.S., ECML 08; KDD 09

#### Position of the problem

- Jobs arrive and are processed
- Want to detect outliers and anomalies
- Want to predict the traffic / dimension the system
- The job distribution is non-stationary





Preliminary step: Clustering the jobs

# Clustering with Message Passing Algorithm: Affinity Propagation

Frey and Dueck, Science 2007 Affinity Propagation w.r.t. State of art

	K-means	K-centers	AP
exemplar	artefact	actual point	actual point
parameter	K	K	s <sup>*</sup> (penalty)
algorithm	greedy search	greedy search	message passing
performance	not stable	not stable	stable
complexity	N  imes K	N  imes K	$N^2 log(N)$

WHEN ? WHY ? CONS When averages don't make sense Stable, minimal distortion Computational complexity

# **Affinity Propagation**

Given  $\mathcal{E} = \{e_1, e_2, ..., e_N\}$ elements  $d(e_i, e_i)$ their dissimilarity Find  $\sigma : \mathcal{E} \mapsto \mathcal{E}$  $\sigma(e_i)$ , exemplar representing  $e_i$ such that:  $\sigma = \operatorname{argmax} \sum_{i=1}^{N} S(e_i, \sigma(e_i))$ where  $\begin{cases} S(e_i, e_j) = -d^2(e_i, e_j) & \text{if } i \neq j \\ S(e_i, e_i) = -s^* \end{cases}$ **s**\*: penalty parameter

#### Particular cases

 $s^* = \infty$ , only one exemplar
 $s^* = 0$ , every point is an exemplar
N clusters

## Affinity Propagation, 2

#### Two types of messages

- a(i, k): Availability of i as examplar for k
- r(i,k) : Responsibility of *i* to *k*

#### **Rules of propagation**

$$\begin{array}{rcl} r(i,k) &=& S(e_i,e_k) - \max_{k',k' \neq k} \{ a(i,k') + S(e_i,e_k') \} \\ r(k,k) &=& S(e_k,e_k) - \max_{k',k' \neq k} \{ S(e_k,e_k') \} \end{array}$$

$$\begin{array}{lll} a(i,k) &=& \min\{0,r(k,k) + \sum_{i',i' \neq i,k} \max\{0,r(i',k)\}\}\\ a(k,k) &=& \sum_{i',i' \neq k} \max\{0,r(i',k)\} \end{array}$$

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > <



◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?



・ロト ・ 日下 ・ 日下 ・ 日下 ・ 今日・



◆□▶ ◆□▶ ◆三▶ ◆三▶ ◆□▶ ◆□



◆□> ◆□> ◆豆> ◆豆> ・豆 ・ のへで



◆□> ◆□> ◆豆> ◆豆> ・豆 ・ のへで



◆□> ◆□> ◆豆> ◆豆> ・豆 ・ のへで
# Iterations of Message passing



◆□> ◆□> ◆豆> ◆豆> ・豆 ・ のへで

# Iterations of Message passing



◆□▶ ◆□▶ ◆臣▶ ◆臣▶ ─臣 ─のへで

# **Hierarchical Affinity Propagation**



#### Thm

Let *h* be the height of the tree, *b* the branching factor,  $N_0$  the size of each subproblem, *K* the average number of examplars for each sub problem. Then

$$C(h) \propto N^{rac{h+2}{h+1}}$$

# **Extending AP to Data Streaming**

#### StrAP : sketch

- 1. Job  $j_t$  arrives
- 2. Does it fit the current model  $\mathcal{M}_t$ ?
  - YES: update  $M_t$
  - NO:

 $j_t \rightarrow \mathsf{Reservoir}$ 

- 3. Has the distribution changed ?
  - YES: build  $\mathcal{M}_{t+1}$  from  $\mathcal{M}_t$  and the reservoir

### **Stream Model**: $\mathcal{M}_t = \{(j_i, n_i, \Sigma_i, t_i)\}$

- ▶ *j<sub>i</sub>* examplar job
- $n_i$  number of jobs represented by  $j_i$
- $\Sigma_i$  sum of distortions incurred by  $j_i$
- t<sub>i</sub> last time step when a job was affected to j<sub>i</sub>

### Has the distribution changed ?



#### Page-Hinkley statistical change detection test

D. Hinkley. Inference about the change-point in a sequence of random variables. Biometrika, 1970

・ロト ・四ト ・ヨト ・ヨト ・ヨ

E. Page. Continuous inspection schemes. Biometrika, 1954

# **EGEE Job Streaming**

Dynamics of the distribution: schedule of restarts

Accuracy (succ/failed jobs)



#### Snapshots





(日)、

æ

# The EGEE traffic: months at a glance

#### A posteriori

build super-examplars from examplars aggregate the traffic

each s.e. a row along time



# EGEE Job Streaming, end

#### **Further work**

- List / Interpret outliers. Build a catalogue of situations
- 2. From job clustering to day clustering A day is a histogram of job clusters
- 3. Sequence modelling Caveat: nature of random variables
- 4. Fueling Job scheduling with realistic distribution models.

## Overview

#### Clustering

K-Means Generative models Expectation Maximization Selecting the number of clusters Stability

Axiomatisation

Data Streaming

Application: Autonomic Computing

▲□▶ ▲圖▶ ▲臣▶ ▲臣▶ ―臣 … のへで

Learning distances