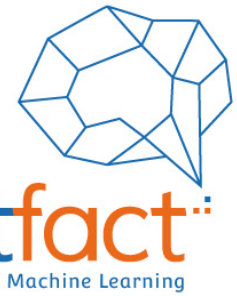
Approches prédictives dans le digital

Online and Unsupervised Learning

Sébastien Loustau, CEO

Chaire DAMI, le 28 Mars 2017



Contents

artfact

Real Time Machine Learning

Contents

Machine learning

artfact
Real Time Machine Learning

Machine learning

$$x \longrightarrow \boxed{\text{nature}} \longrightarrow y$$

- **prédire** la réponse y à partir de x,

artfact
Real Time Machine Learning

Machine learning

$$x \longrightarrow \boxed{\text{nature}} \longrightarrow y$$

- **prédire** la réponse y à partir de x,
- **comprendre** le lien entre x et y.

artfact
Real Time Machine Learning

Machine learning

$$x \longrightarrow \boxed{\text{nature}} \longrightarrow y$$

- **prédire** la réponse y à partir de x,
- **comprendre** le lien entre x et y.

$$x \longrightarrow \boxed{\text{algorithm}} \longrightarrow \widehat{y}$$

artfact
Real Time Machine Learning

Statistical Learning vs Online Learning

Statistical Learning

We observe a training set $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$.

Statistical Learning vs Online Learning

Statistical Learning

We observe a training set $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$. New x arrives. We build a model/algorithm thanks to $\mathcal{D}_n$ and predict $\hat{y}$.

artfact
Real Time Machine Learning

Statistical Learning vs Online Learning

Statistical Learning

We observe a training set $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$. New x arrives. We build a model/algorithm thanks to $\mathcal{D}_n$ and predict $\hat{y}$.

Online Learning

Data arrives sequentially.

Statistical Learning vs Online Learning

Statistical Learning

We observe a training set $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$. New x arrives. We build a model/algorithm thanks to $\mathcal{D}_n$ and predict $\hat{y}$.

Online Learning

Data arrives sequentially. At each time t, we want to make a decision based on past observations.

Statistical Learning vs Online Learning

Statistical Learning

We observe a training set $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$. New x arrives. We build a model/algorithm thanks to $\mathcal{D}_n$ and predict $\hat{y}$.

Online Learning

Data arrives sequentially. At each time t, we want to make a decision based on past observations. **No assumption over the data mechanism**.

Statistical Learning vs Online Learning

Statistical Learning

We observe a training set $\mathcal{D}_n = \{(X_i, Y_i), i = 1, \ldots, n\}$. New x arrives. We build a model/algorithm thanks to $\mathcal{D}_n$ and predict $\hat{y}$.

Online Learning

Data arrives sequentially. At each time t, we want to make a decision based on past observations. **No assumption over the data mechanism**.

Real time ML : un enjeu de taille

DATA

stockage massif
et couteux

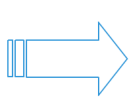
analyses
off-line

décision tardive

artfact
Real Time Machine Learning

Real time ML : un enjeu de taille

analyses
on-line



décision
temps réel

stockage
allégé

artfact
Real Time Machine Learning

Supervised learning

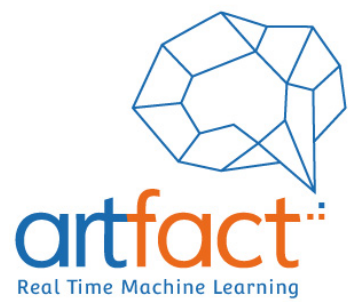
$$x \longrightarrow \boxed{\text{nature}} \longrightarrow y$$

- **prédire** la réponse y à partir de x,
- **comprendre** le lien entre x et y.

$$x \longrightarrow \boxed{\text{algorithm}} \longrightarrow \widehat{y}$$

artfact
Real Time Machine Learning

Today : unsupervised learning



Today : unsupervised learning

$$\boxed{\text{nature}} \longrightarrow x$$

- décrire les observations x,

Today : unsupervised learning

$$\boxed{\text{nature}} \longrightarrow x$$

- décrire les observations x,
- réduire la dimension de x,

artfact
Real Time Machine Learning

Today : unsupervised learning

$$\boxed{\text{nature}} \longrightarrow x$$

- décrire les observations x,
- réduire la dimension de x,
- grouper les observations x.

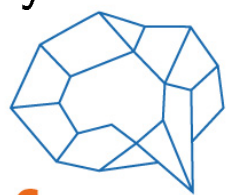
Today : unsupervised learning

$$\boxed{\text{nature}} \longrightarrow x$$

- décrire les observations x,
- réduire la dimension de x,
- grouper les observations x.

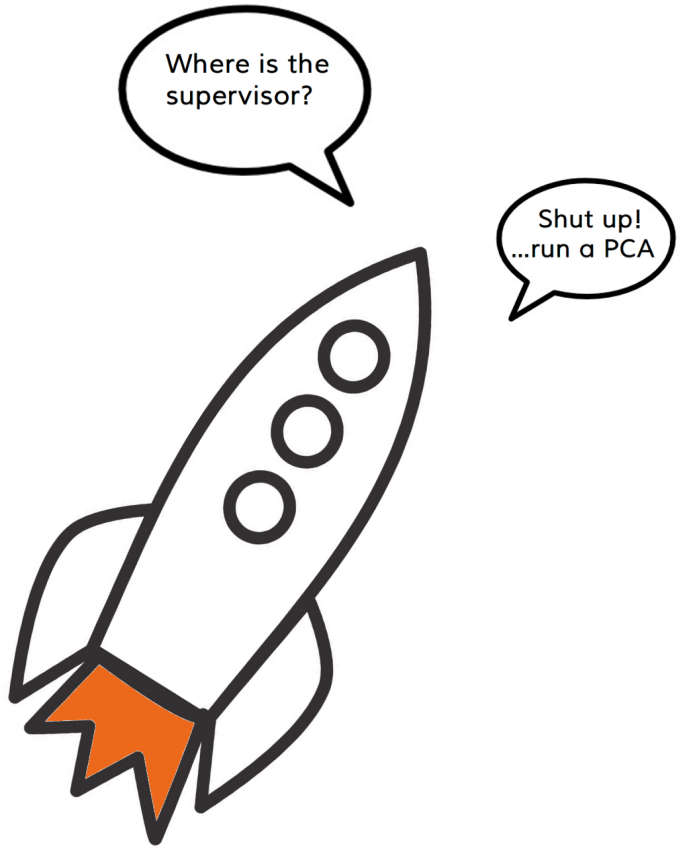
Vanilla algorithms and recent hot topics

PCA, k-means, gaussian mixtures, change points or more recently words embeddings, community detection, generative models (GAN)...



artfact
Real Time Machine Learning

Qu'est-ce que ça change ?



UNSUPERVISED LEARNING

artfact
Real Time Machine Learning

Supervised routine

Supervised routine

$$x \longrightarrow \boxed{\text{algorithm}} \longrightarrow \widehat{y}$$

Supervised routine

$$x \longrightarrow \boxed{\text{algorithm}} \longrightarrow \widehat{y}$$

- Build a model from a training sample $\{(X_i, Y_i), i = 1, \ldots, n\}$.

Supervised routine

$$x \longrightarrow \boxed{\text{algorithm}} \longrightarrow \widehat{y}$$

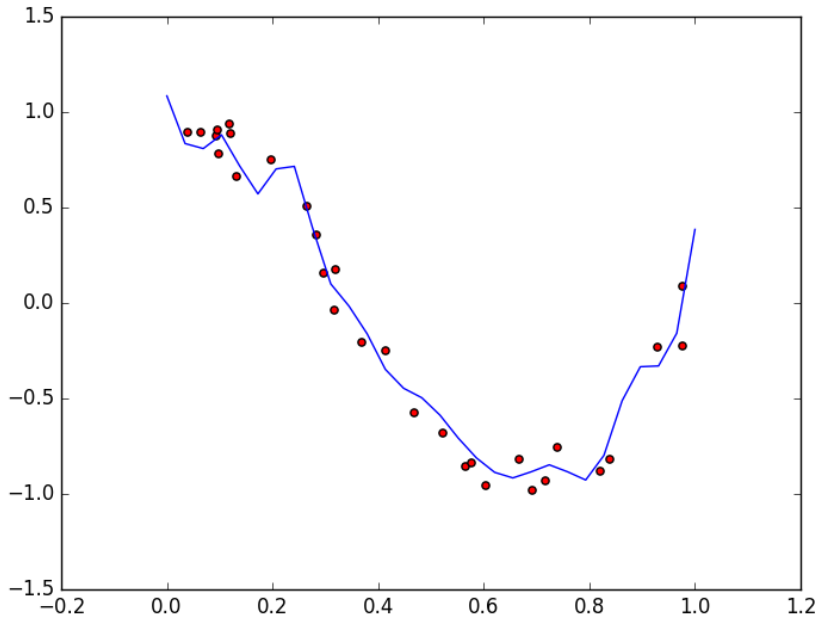
- Build a model from a training sample $\{(X_i, Y_i), i = 1, \ldots, n\}$.
- Observe new x and predict $\hat{y}$ as above.

artfact
Real Time Machine Learning

Supervised routine

$$x \longrightarrow \boxed{\text{algorithm}} \longrightarrow \widehat{y}$$

- Build a model from a training sample $\{(X_i, Y_i), i = 1, \ldots, n\}$.
- Observe new x and predict $\hat{y}$ as above.
- Problem : overfitting !

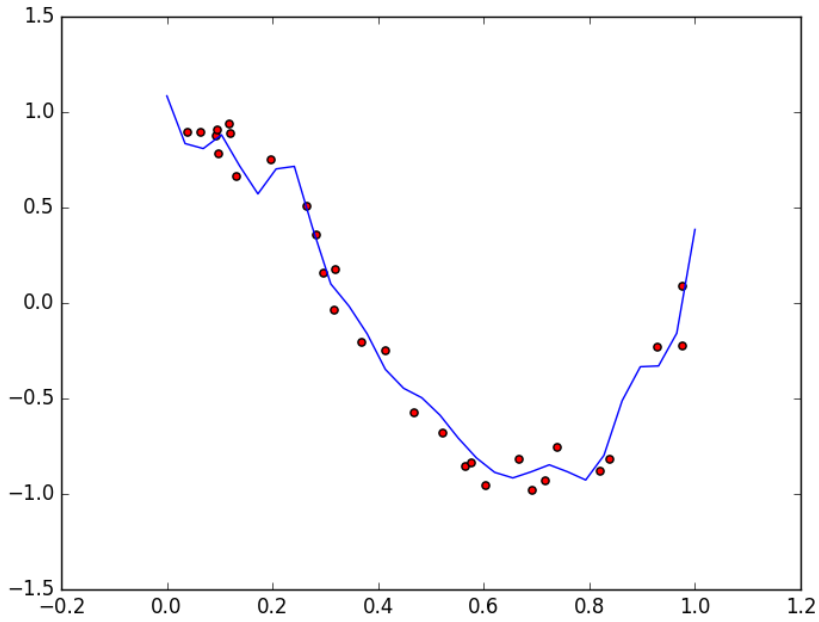


artfact
Real Time Machine Learning

Supervised routine

$$x \longrightarrow \boxed{\text{algorithm}} \longrightarrow \widehat{y}$$

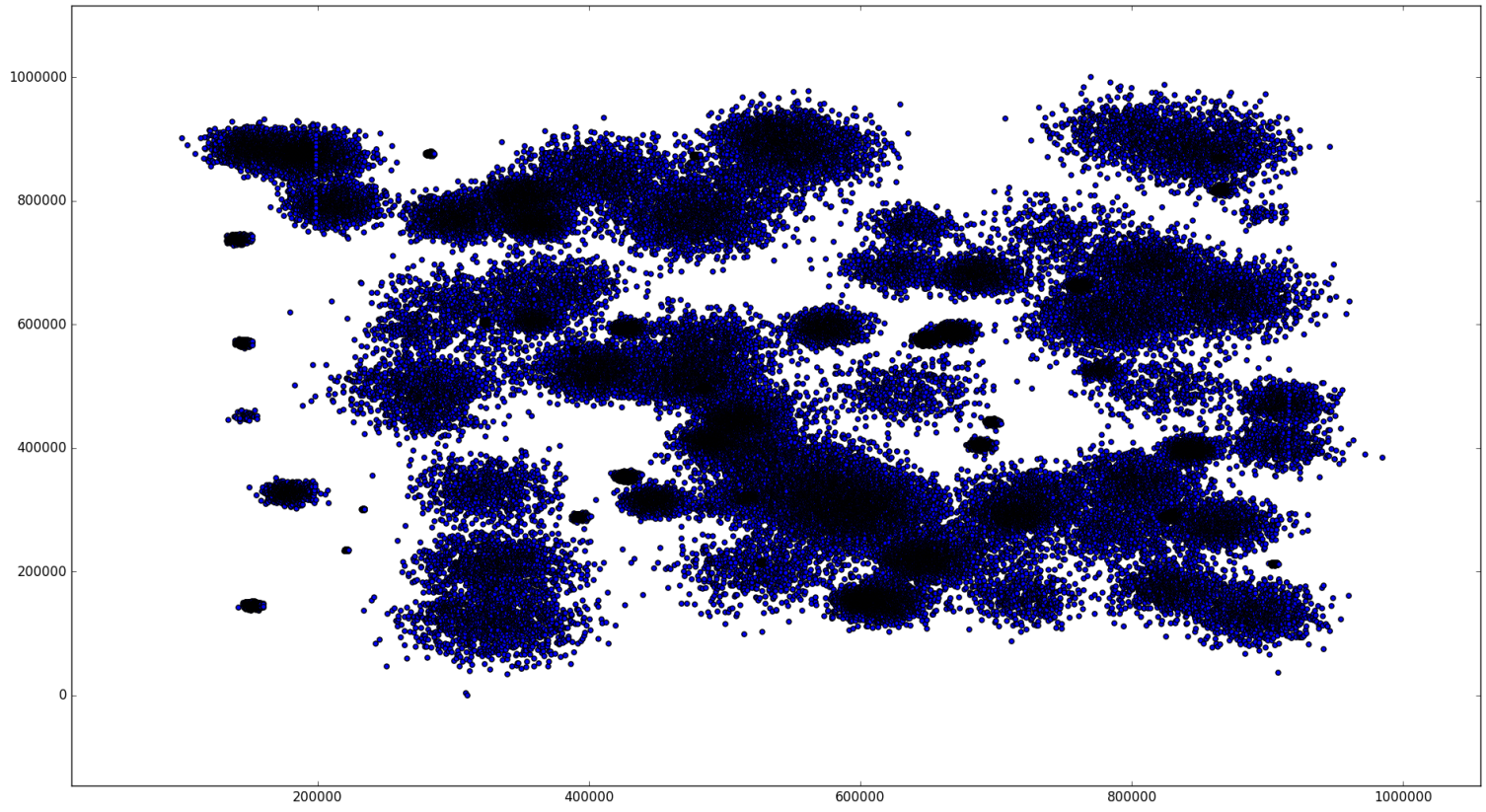
- Build a model from a training sample $\{(X_i, Y_i), i = 1, \ldots, n\}$.
- Observe new x and predict $\hat{y}$ as above.
- Problem : overfitting !



- Solution : Training set + test set, Leave-One-Out, V-fold Cross validation.

artfact
Real Time Machine Learning

Unsupervised : science or art ?



artfact
Real Time Machine Learning

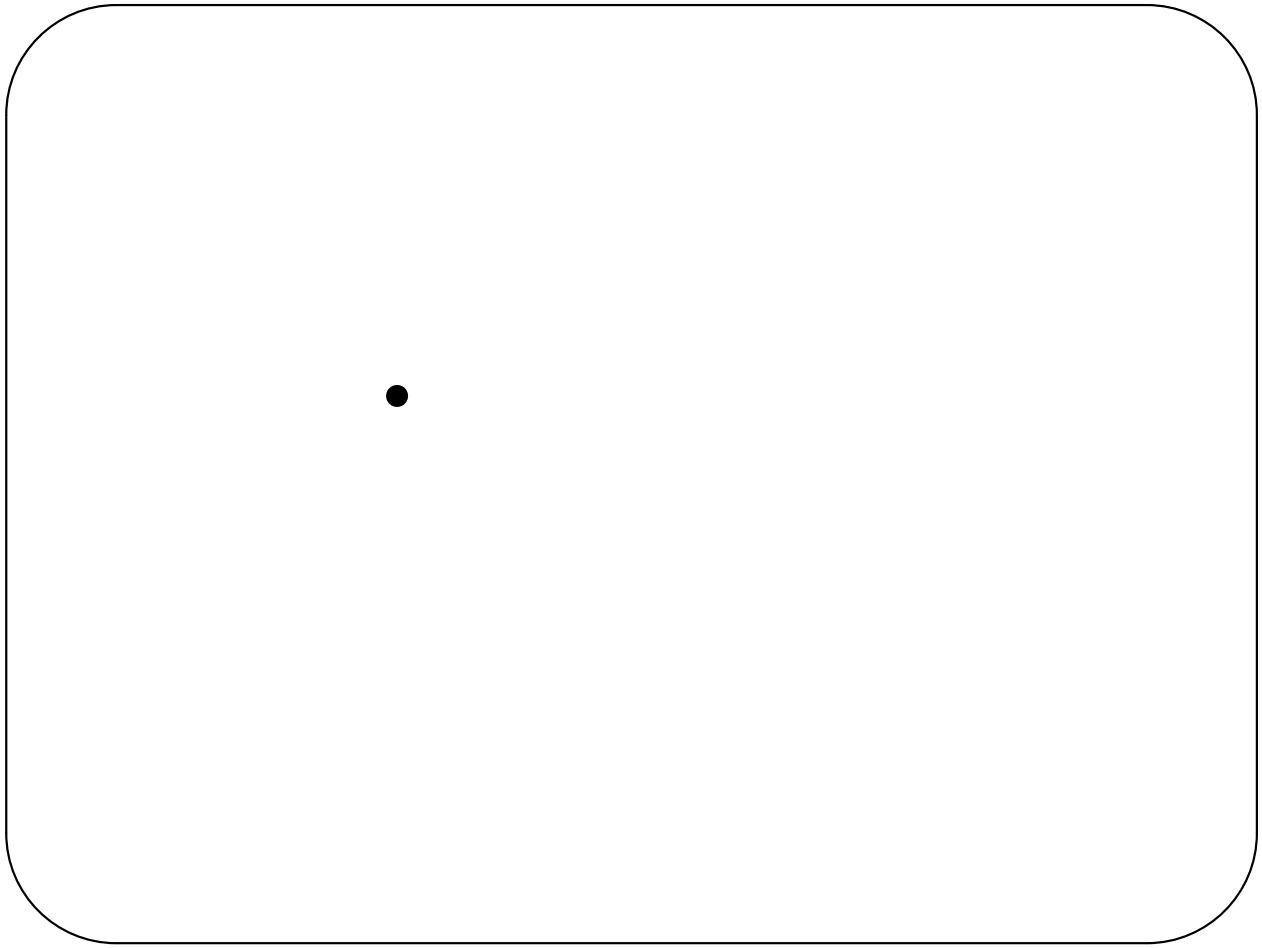
Unsupervised : the hype

artfact
Real Time Machine Learning

Contents

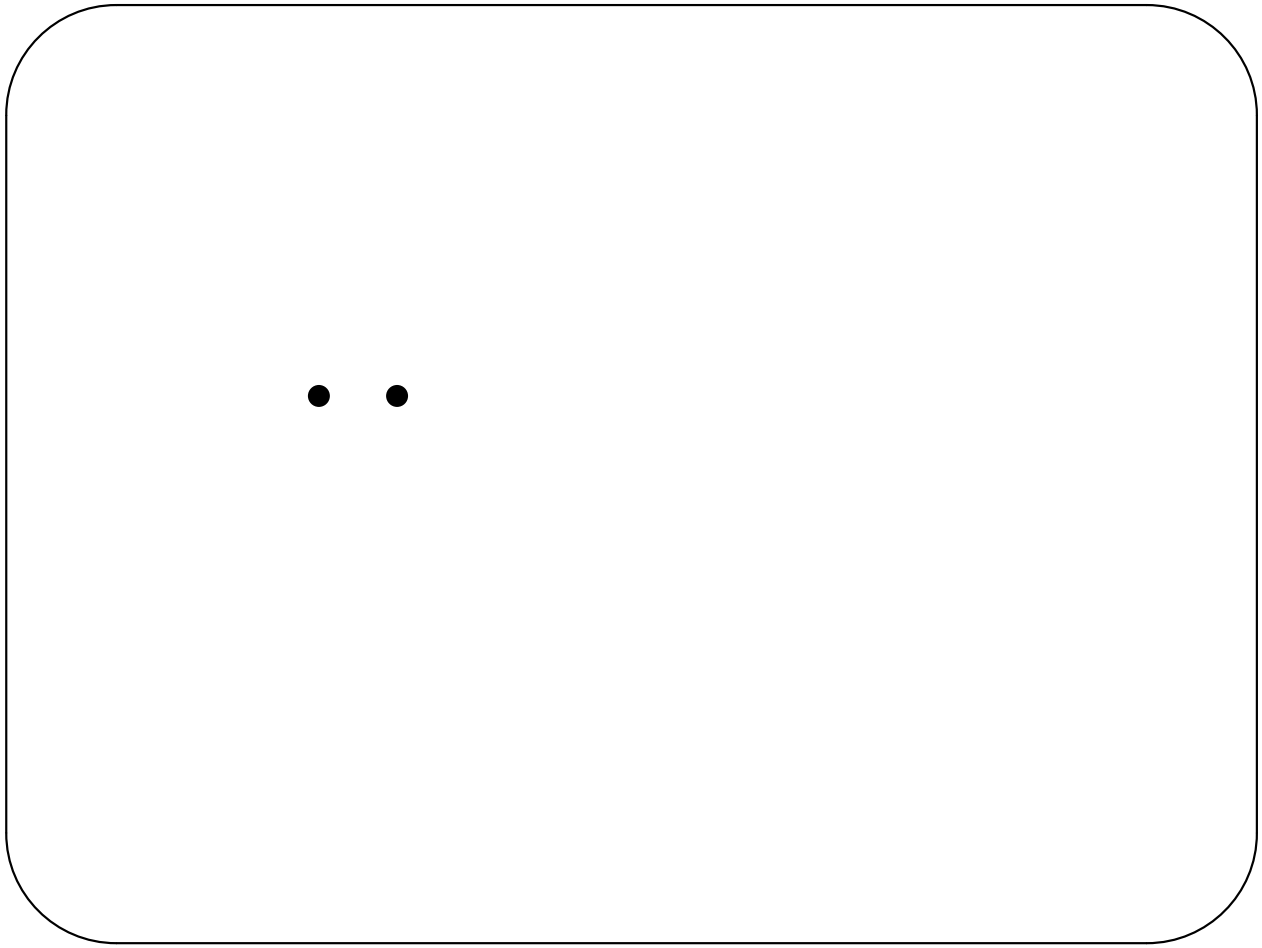
The problem of Online Clustering

The problem of Online Clustering



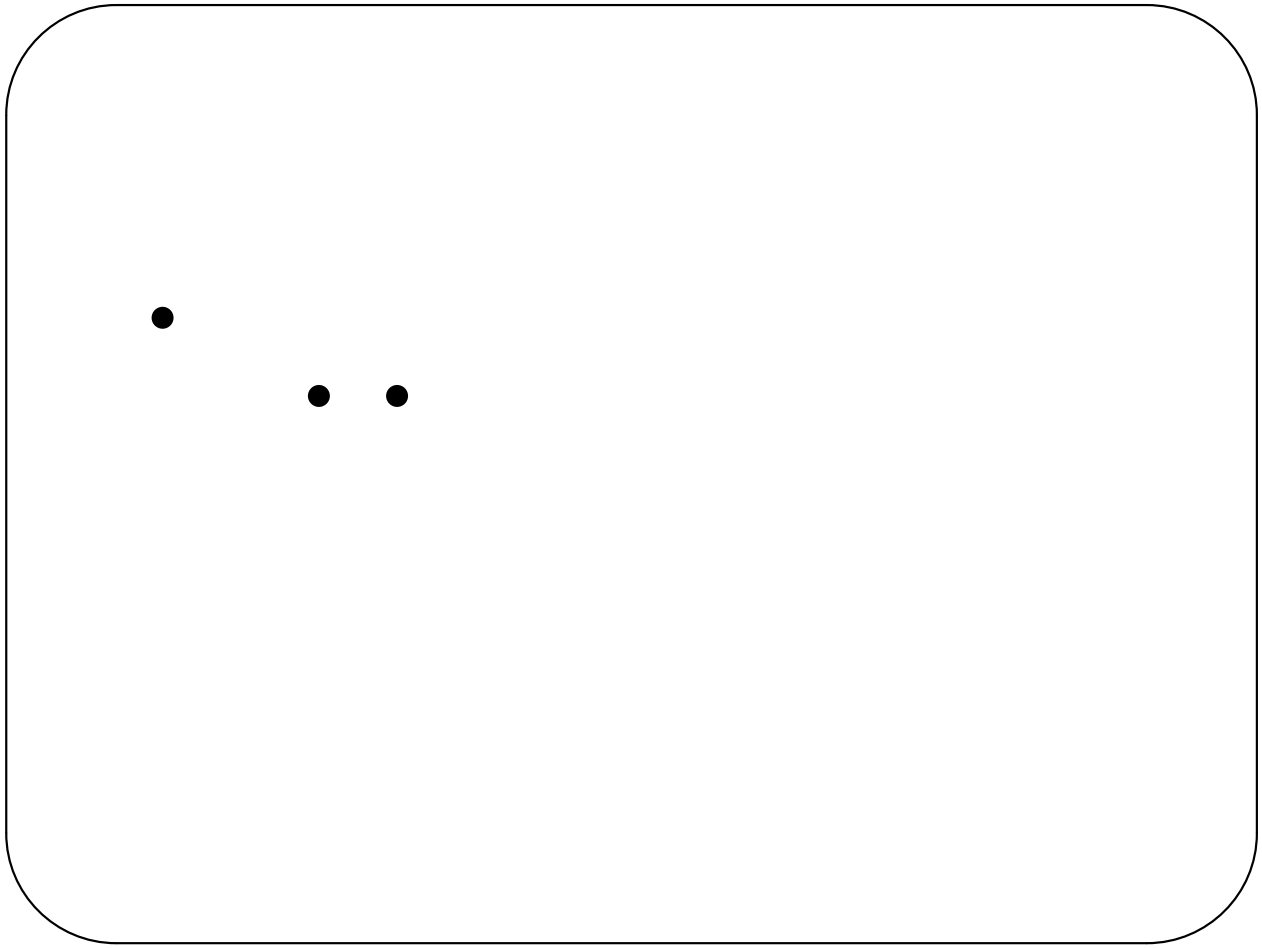
artfact
Real Time Machine Learning

The problem of Online Clustering



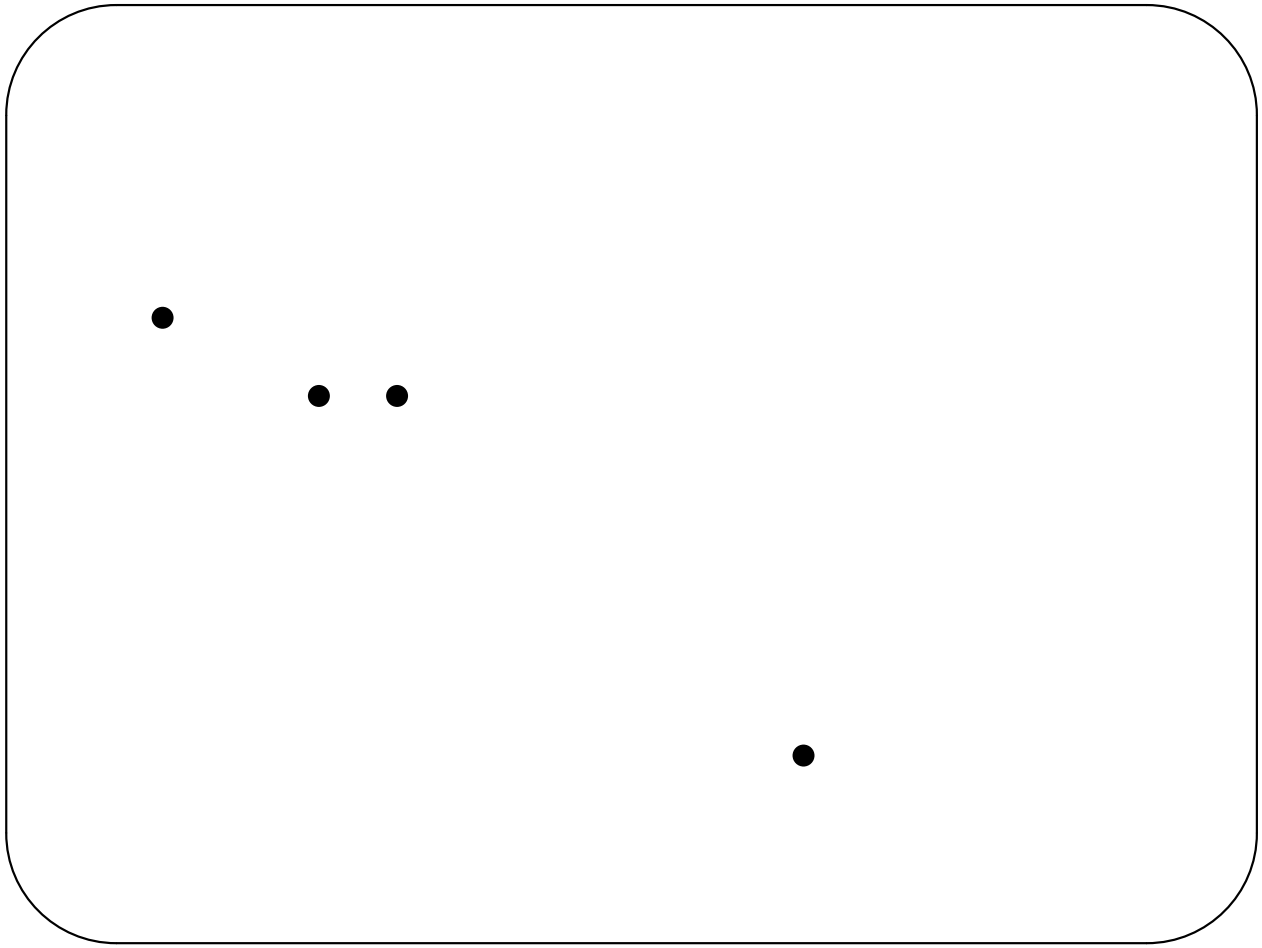
artfact
Real Time Machine Learning

The problem of Online Clustering



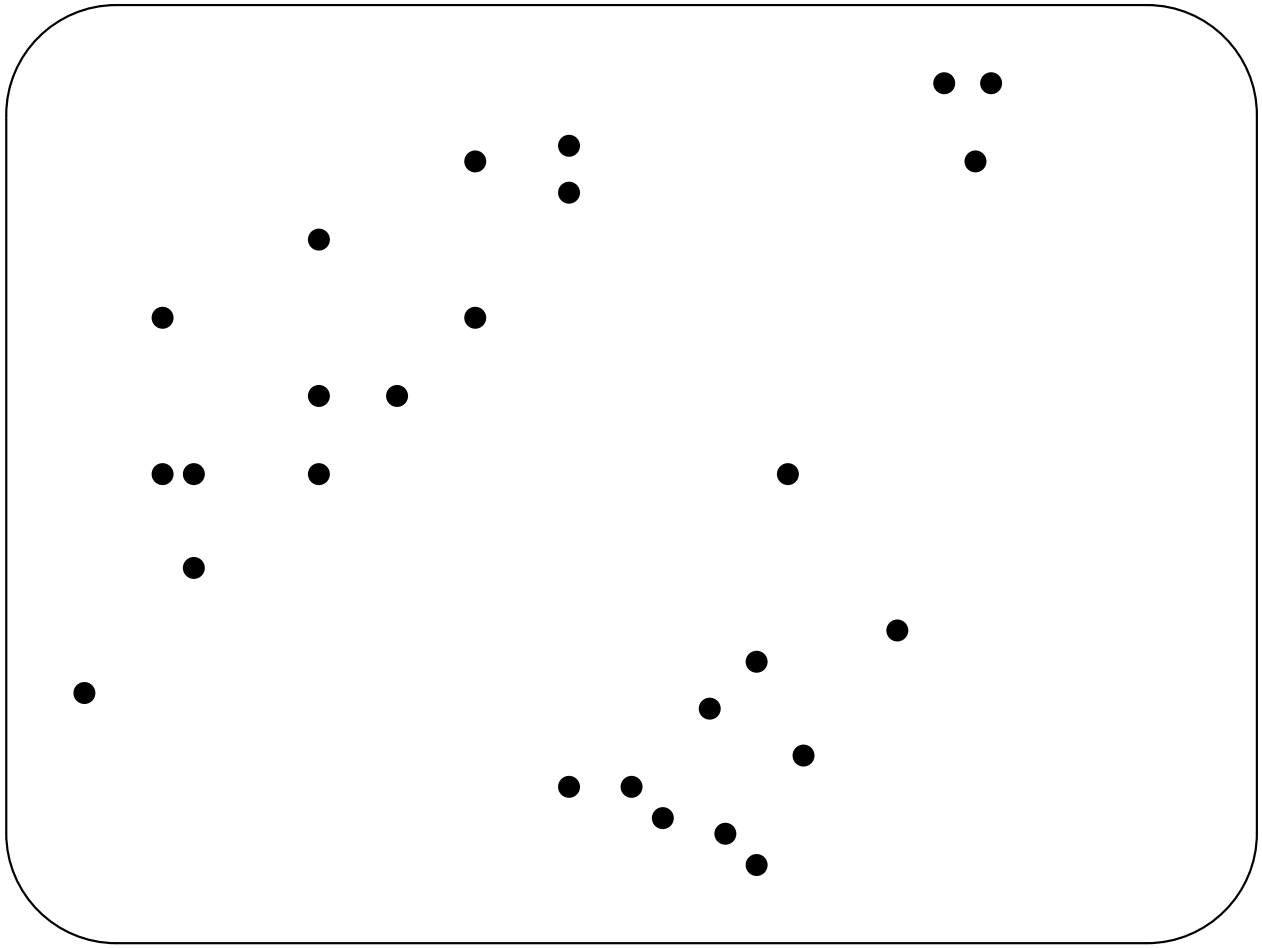
artfact
Real Time Machine Learning

The problem of Online Clustering



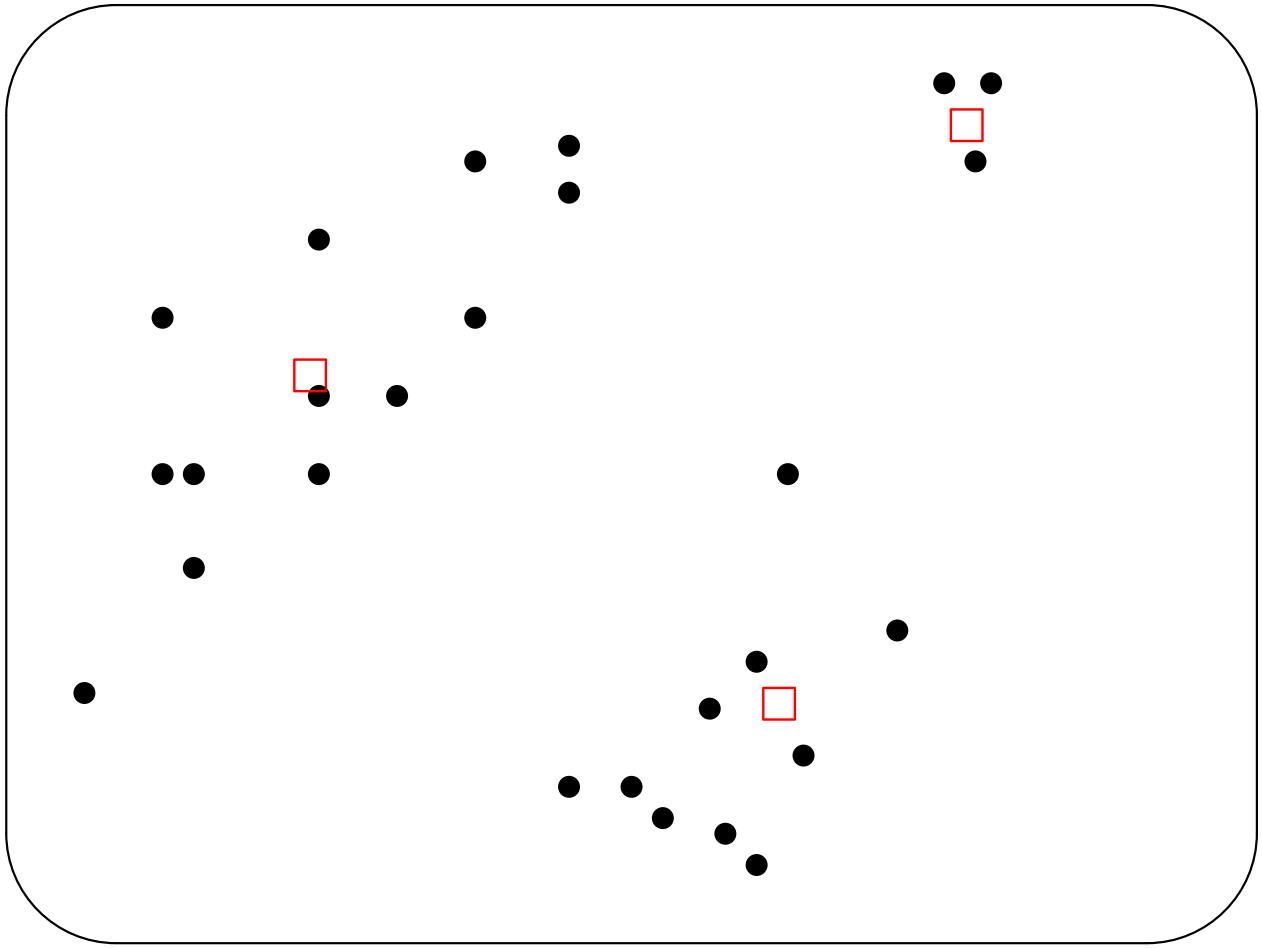
artfact
Real Time Machine Learning

The problem of Online Clustering



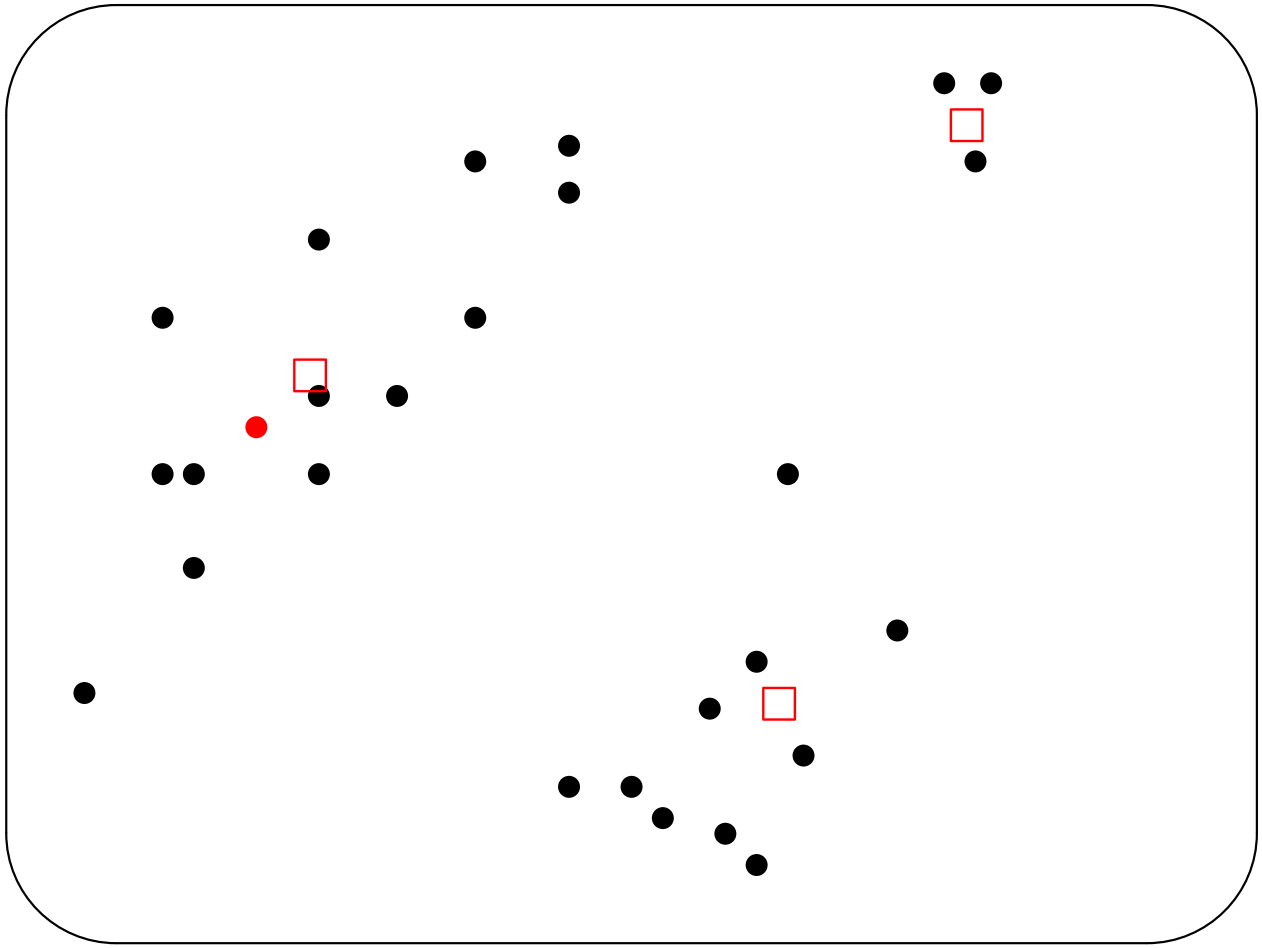
artfact
Real Time Machine Learning

The problem of Online Clustering



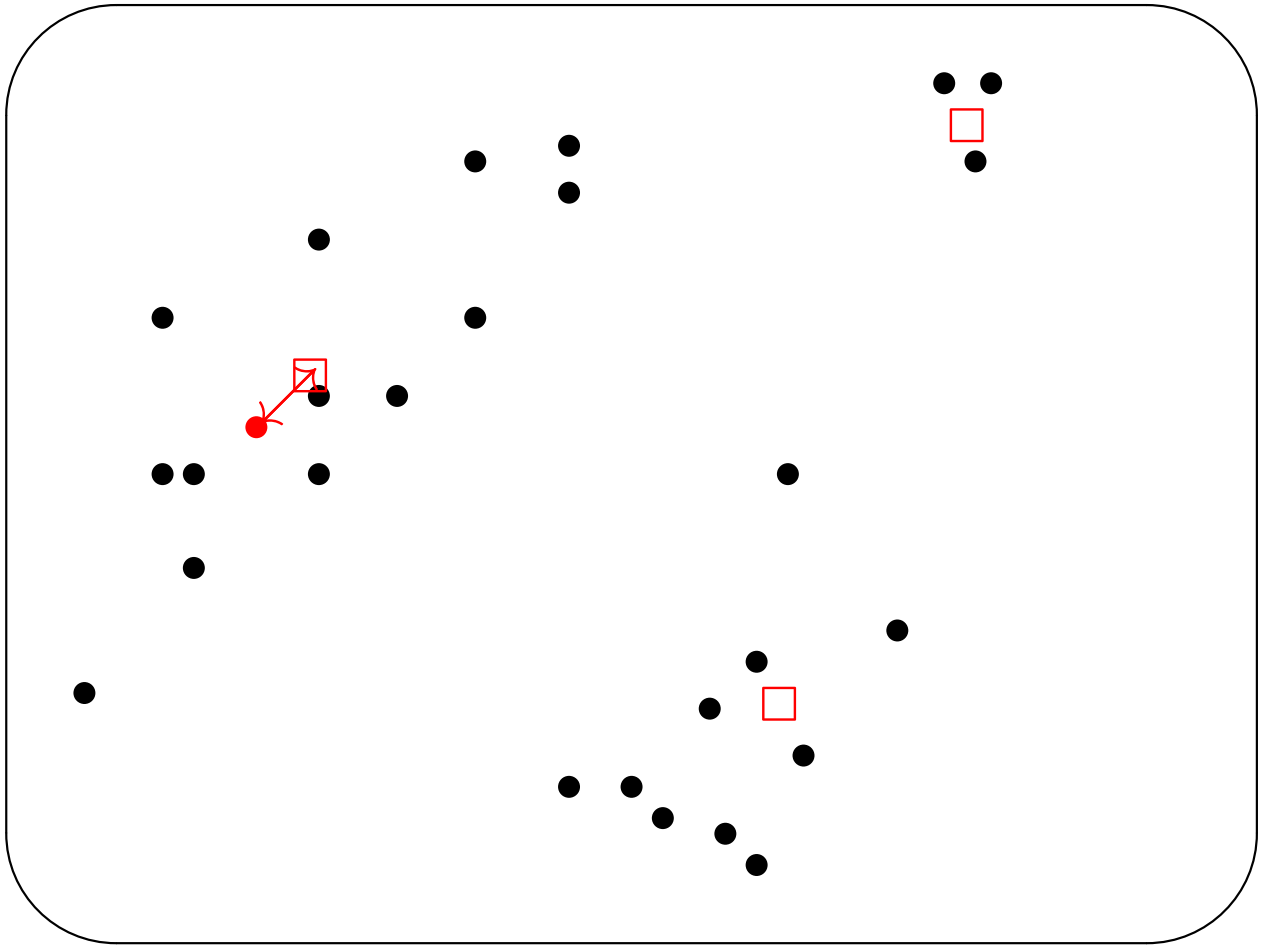
artfact
Real Time Machine Learning

The problem of Online Clustering



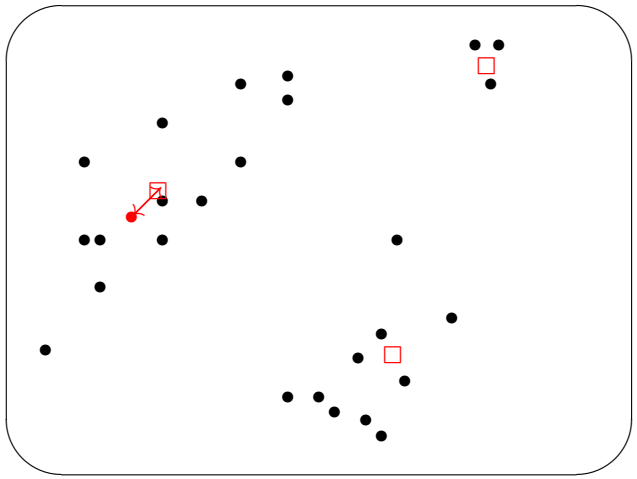
artfact
Real Time Machine Learning

The problem of Online Clustering



artfact
Real Time Machine Learning

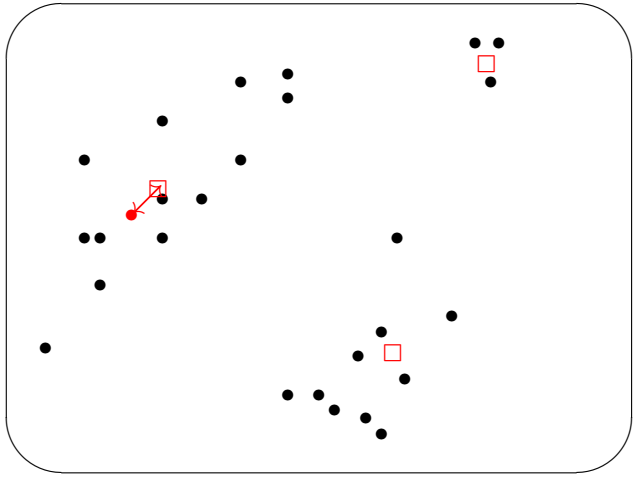
The problem of Online Clustering



- Principle : clustering as prediction !

artfact
Real Time Machine Learning

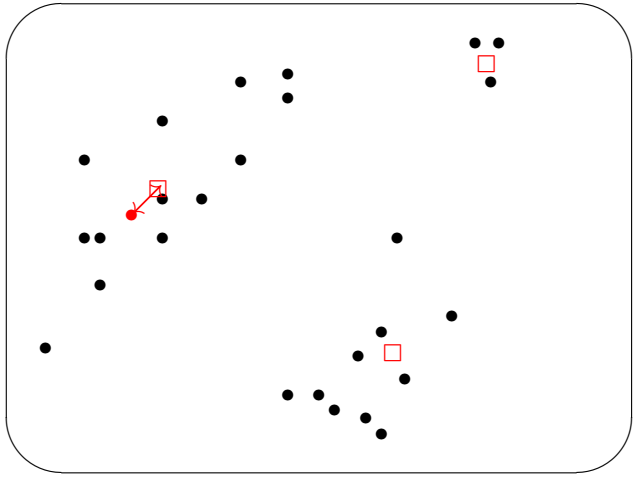
The problem of Online Clustering



- Principle : clustering as prediction !
- Sparsity assumption : points are grouped into s clusters.

artfact
Real Time Machine Learning

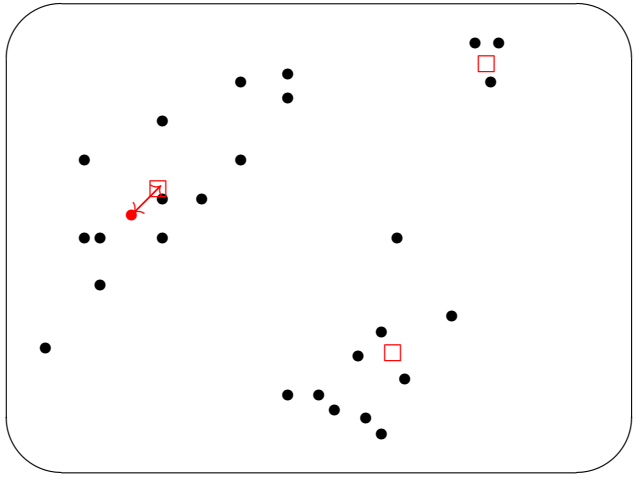
The problem of Online Clustering



- Principle : clustering as prediction !
- Sparsity assumption : points are grouped into s clusters.
- Use PAC-Bayesian regularization to choose the number of clusters.

artfact
Real Time Machine Learning

The problem of Online Clustering



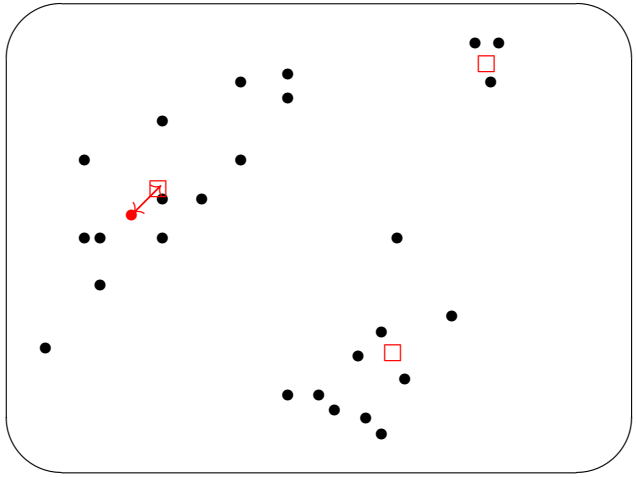
- Principle : clustering as prediction !
- Sparsity assumption : points are grouped into s clusters.
- Use PAC-Bayesian regularization to choose the number of clusters.

We prove new kind of sparsity regret bounds:

$$\sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c} \in \mathbb{R}^{dp}} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \lambda |\mathbf{c}|_0 \right\},$$

artfact
Real Time Machine Learning

The problem of Online Clustering



- Principle : clustering as prediction !
- Sparsity assumption : points are grouped into s clusters.
- Use PAC-Bayesian regularization to choose the number of clusters.

We prove new kind of sparsity regret bounds:

$$\sum_{t=1}^{T} \ell(\hat{\mathbf{c}}_t, x_t) - \inf_{\mathbf{c}\in\mathbb{R}^{dp}} \left\{ \sum_{t=1}^{T} \ell(\mathbf{c}, x_t) + \lambda |\mathbf{c}|_0 \right\},$$

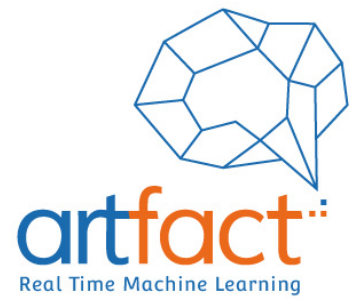
where $|\mathbf{c}|_0 = \text{card}\{j = 1, \ldots, p : c_j \neq 0_{\mathbb{R}^d}\}$ and

$$\ell(\mathbf{c}, x) = \min_{j=1,\ldots,p} \|c_j - x\|_2^2.$$

artfact
Real Time Machine Learning

Business case

- **Client** : éditeur de logiciel dans le commerce conversationnel



Business case

- **Client** : éditeur de logiciel dans le commerce conversationnel
- **Inputs** (X): comportement de navigations

Business case

- **Client** : éditeur de logiciel dans le commerce conversationnel
- **Inputs** (X): comportement de navigations
- **Objectifs** : satisfaction clients, instantanéité, conversion sans canibalisme

artfact
Real Time Machine Learning

Business case

- **Client** : éditeur de logiciel dans le commerce conversationnel
- **Inputs** (X): comportement de navigations
- **Objectifs** : satisfaction clients, instantanéité, conversion sans canibalisme

artfact
Real Time Machine Learning

L'existant

- Absence de stockage

L'existant

- Absence de stockage
- Absence d'architecture Big Data

L'existant

- Absence de stockage
- Absence d'architecture Big Data
- Réglage manuel des règles de ciblage

artfact
Real Time Machine Learning

L'existant

- Absence de stockage
- Absence d'architecture Big Data
- Réglage manuel des règles de ciblage

artfact
Real Time Machine Learning

Proof of concept (POC) : périmètre

- 11 sites
- 20M pvues/mois
- 160k events/heure

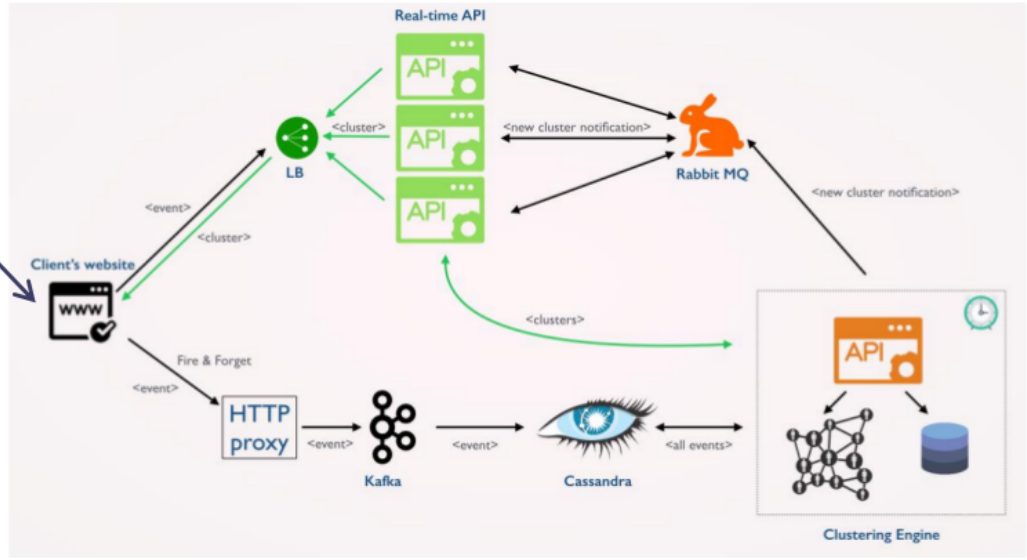
Proof of concept (POC) : objectifs

- Automatiser le ciblage
- Simplifier l'administration et la mise à jour
- Améliorer les performances

artfact
Real Time Machine Learning

Industrialisation : architecture

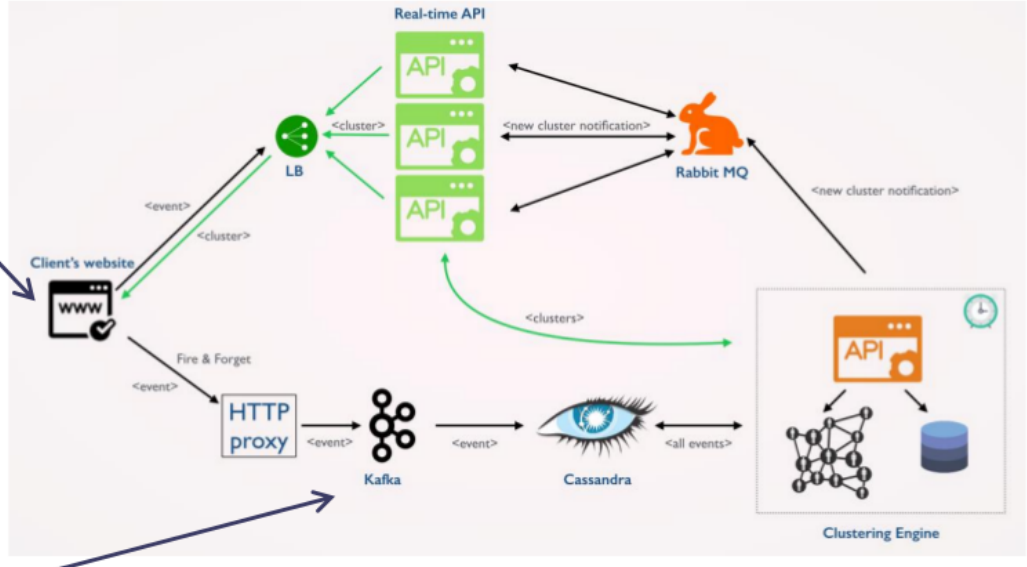
Site e-commerce:
Source des données



artfact
Real Time Machine Learning

Industrialisation : architecture

Site e-commerce:
Source des données

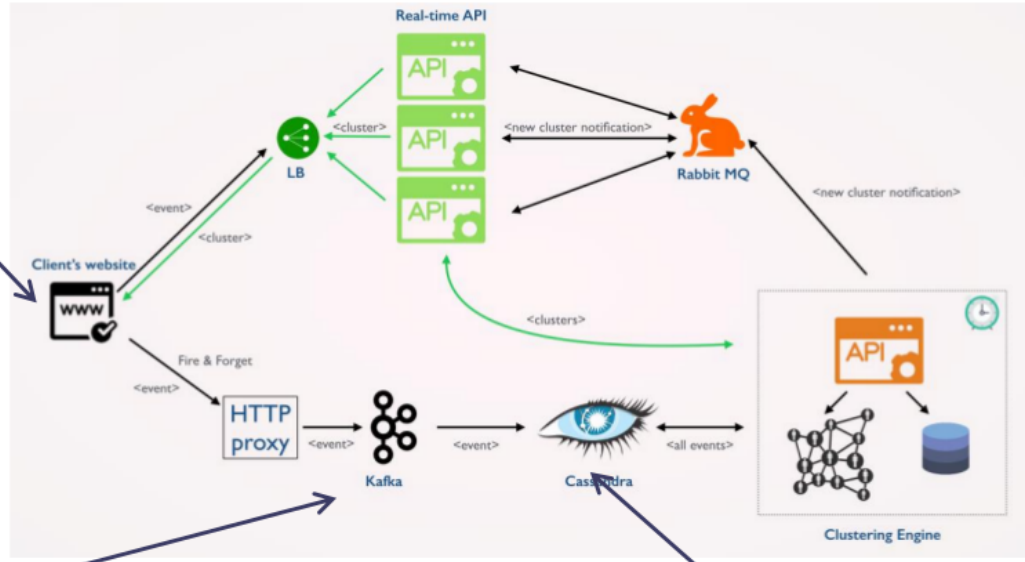


Collecte des events

artfact
Real Time Machine Learning

Industrialisation : architecture

Site e-commerce:
Source des données



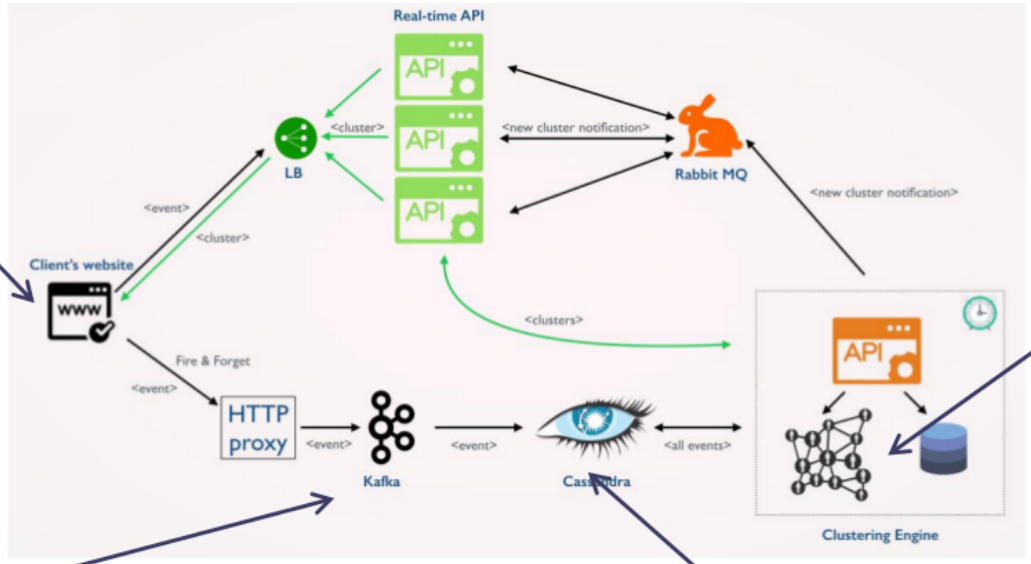
Collecte des events

Stockage des events

artfact
Real Time Machine Learning

Industrialisation : architecture

-commerce:
des données



Algorithme

Collecte des events

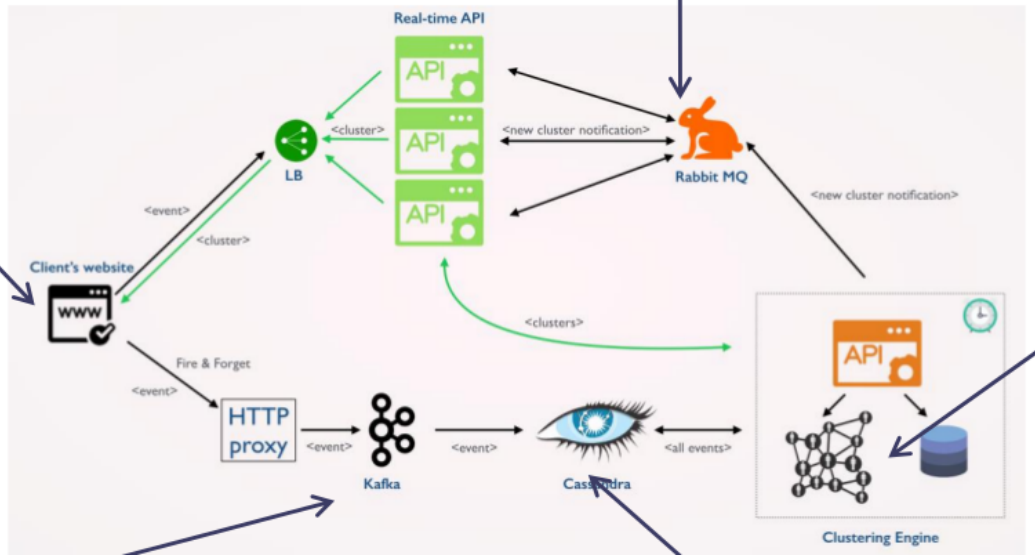
Stockage des events

artfact
Real Time Machine Learning

Industrialisation : architecture

Stockage des résultats du clustering
> Probabilité d'achat, d'ajout au panier etc.

rce:
nées



Algorithme

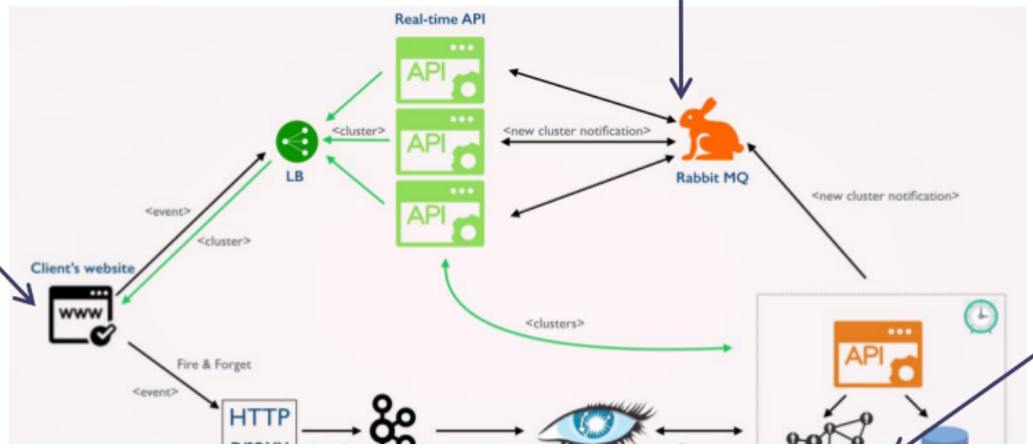
artfact
Real Time Machine Learning

Industrialisation : architecture

Affectation des nouveaux arrivants aux clusters

Stockage des résultats du clustering
> Probabilité d'achat, d'ajout au panier etc

erce:
nnées



Algorithme

artfact
Real Time Machine Learning

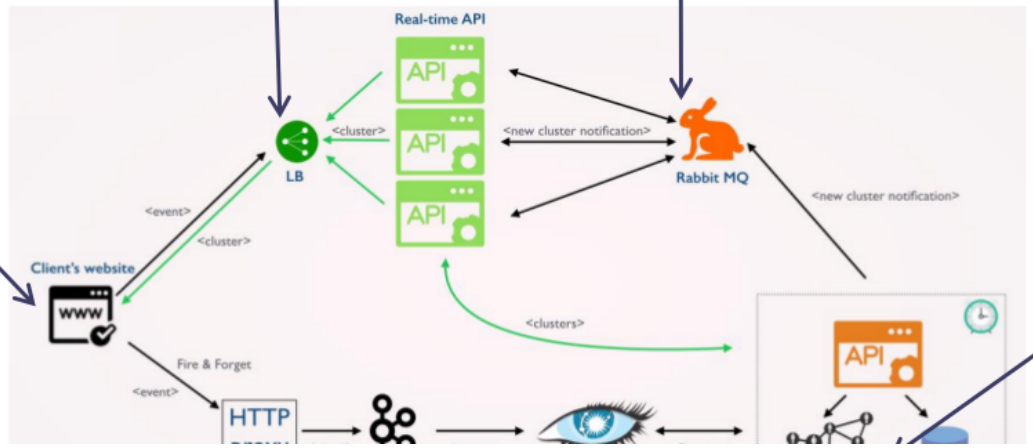
Industrialisation : architecture

Affectation des nouveaux arrivants aux clusters

Stockage des résultats du clustering
> Probabilité d'achat, d'ajout au panier etc

Ciblage

erce:
nnées



Algorithme

artfact
Real Time Machine Learning

Contents

Graph clustering

- Proposer l'analyse précédente sur un graphe

Graph clustering

- Proposer l'analyse précédente sur un graphe
- Nouveaux challenges

Graph clustering

- Proposer l'analyse précédente sur un graphe
- Nouveaux challenges
- Have a look !

Ingrédients

- Théorie des graphes
- Optimisation par Monte Carlo Markov Chain (MCMC)
- Calculs parallèles

Qu'est-ce qu'un graphe ?

Définition (Wikipédia) :

> *Un graphe est un ensemble de points nommés nœuds (parfois sommets ou cellules) reliés par des traits (segments) ou flèches nommées arêtes (ou liens ou arcs). L'ensemble des arêtes entre nœuds forme une figure similaire à un réseau.*

artfact
Real Time Machine Learning

Qu'est-ce qu'un graphe ?

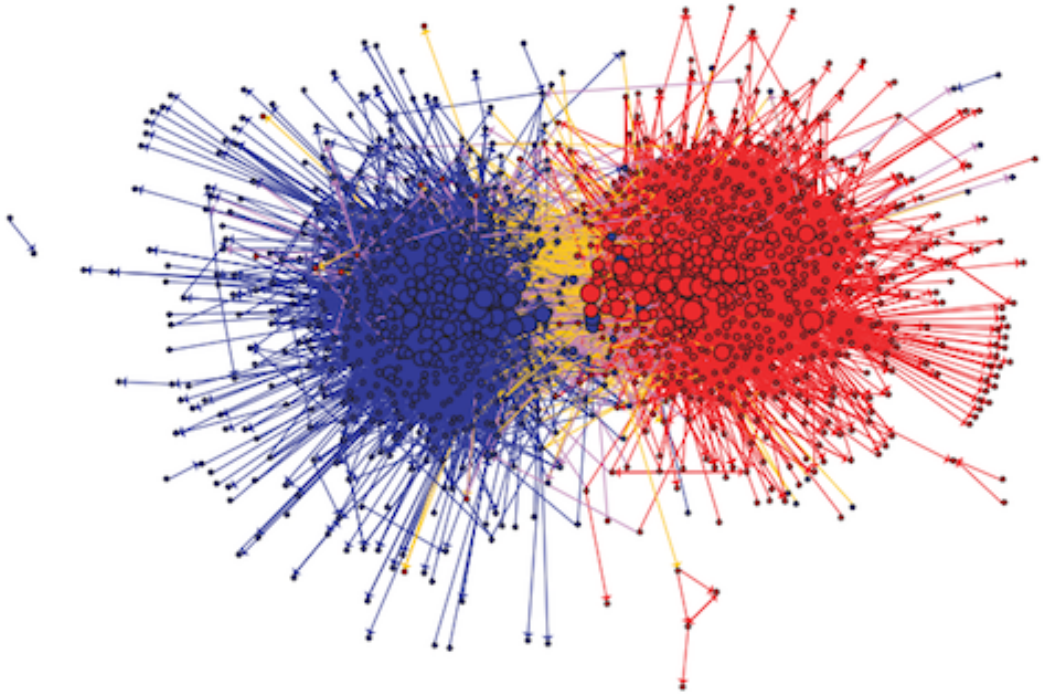
Définition (Wikipédia) :

> *Un graphe est un ensemble de points nommés nœuds (parfois sommets ou cellules) reliés par des traits (segments) ou flèches nommées arêtes (ou liens ou arcs). L'ensemble des arêtes entre nœuds forme une figure similaire à un réseau.*

Keyp point : **relational structure**.

artfact
Real Time Machine Learning

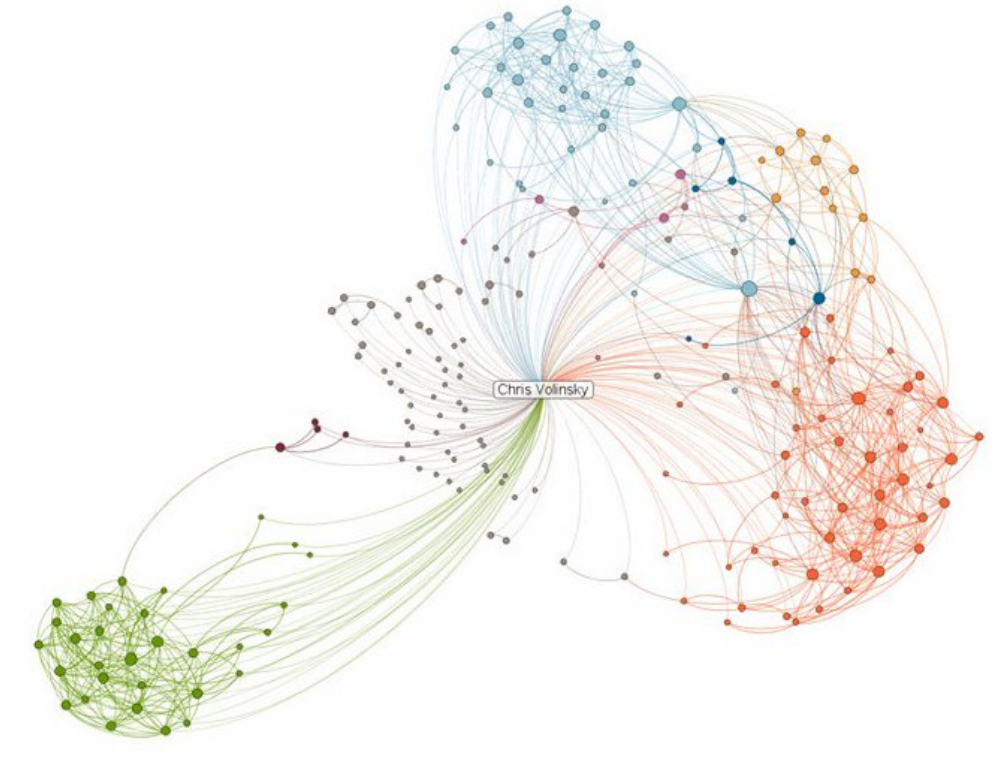
Illustrations : political blog



Network of U.S. political blogs by Adamic and Glance (2004)

artfact
Real Time Machine Learning

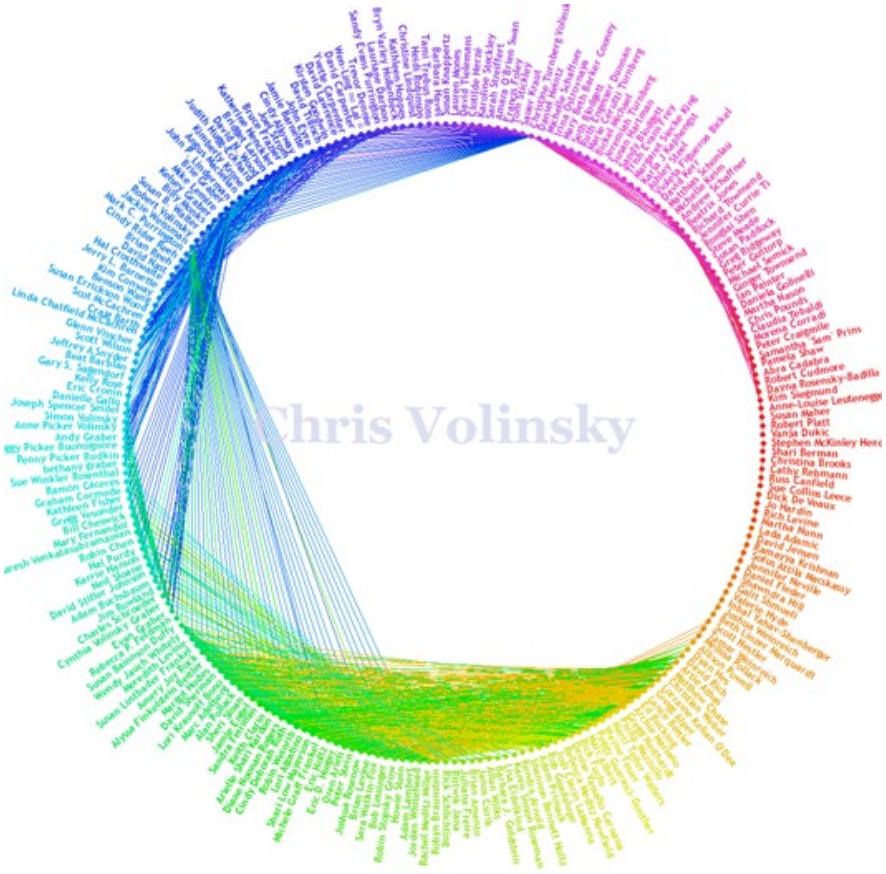
Illustrations : LinkedIn community



LinkedIn community of Chris Volinski (2011, Columbia University), see www.socilab.com

artfact
Real Time Machine Learning

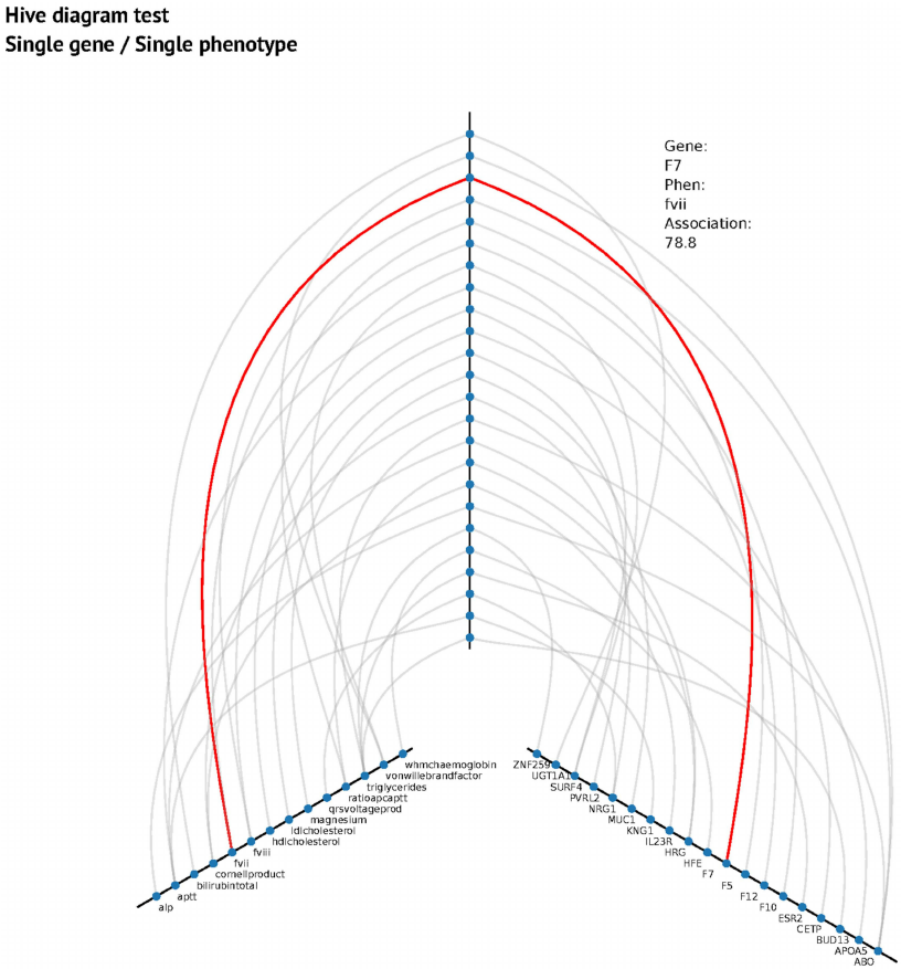
Illustrations : Facebook friends



Facebook of Chris Volinski (2011, Columbia University)

artfact
Real Time Machine Learning

Illustrations : Gene/phenotype interactions



Canonical Correlation Analysis for Gene-Based Pleiotropy Discovery (Plos Computational Biology, 2016)

artfact
Real Time Machine Learning

Formalisme mathématique

Definition

Un graphe G est une structure constituée d'un couple (E, V) où :

Formalisme mathématique

Definition

Un graphe G est une structure constituée d'un couple (E, V) où :

- $V = \{v_1, \ldots, v_N\}$ est un ensemble de N nœuds,

Formalisme mathématique

Definition

Un graphe G est une structure constituée d'un couple (E, V) où :

- $V = \{v_1, \ldots, v_N\}$ est un ensemble de N nœuds,
- E est un ensemble d'arêtes ou liens.

Formalisme mathématique

Definition

Un graphe G est une structure constituée d'un couple (E, V) où :

- $V = \{v_1, \ldots, v_N\}$ est un ensemble de N nœuds,
- E est un ensemble d'arêtes ou liens.

$|V| = N$ est appelé l'ordre de G et $|E|$la taille de G.

artfact
Real Time Machine Learning

Formalisme mathématique

Definition

Un graphe G est une structure constituée d'un couple (E, V) où :

- $V = \{v_1, \ldots, v_N\}$ est un ensemble de N nœuds,
- E est un ensemble d'arêtes ou liens.

$|V| = N$ est appelé l'ordre de G et $|E|$la taille de G.

Definition

Pour un graphe $G = (E, V)$, le degré d'un nœud $v \in V$ est le nombre d'arêtes dans E partant de v.

artfact
Real Time Machine Learning

Propriétés élémentaires

Definition

A **sparse graph** is a graph (E, V) such that $|E|$ is of order $|V|$.

Propriétés élémentaires

Definition

A **sparse graph** is a graph (E, V) such that $|E|$ is of order $|V|$.

Propriétés élémentaires

Definition

A **sparse graph** is a graph (E, V) such that $|E|$ is of order $|V|$.

Definition

A **scale-free network** is a network whose degree distribution follows a power law, that is to say:

$$\mathbb{P}(d_i = k) \sim k^{-\gamma}.$$

Propriétés élémentaires

Definition

A **sparse graph** is a graph (E, V) such that $|E|$ is of order $|V|$.

Definition

A **scale-free network** is a network whose degree distribution follows a power law, that is to say:

$$\mathbb{P}(d_i = k) \sim k^{-\gamma}.$$

Definition

A graph G with N vertices satisfies the **small-world property** if in expectation, 2 randomly chosen vertices i and j have distances proportional to $\log N$.

artfact
Real Time Machine Learning

Détection de communautés (ou graph clustering)

- Partition des nœuds qui vérifie : beaucoup d'arêtes à l'intérieur des groupes, peu d'arêtes entre les groupes.

artfact
Real Time Machine Learning

Détection de communautés (ou graph clustering)

- Partition des nœuds qui vérifie : beaucoup d'arêtes à l'intérieur des groupes, peu d'arêtes entre les groupes.
- Tâche non-supervisée (!)

artfact
Real Time Machine Learning

Détection de communautés (ou graph clustering)

- Partition des nœuds qui vérifie : beaucoup d'arêtes à l'intérieur des groupes, peu d'arêtes entre les groupes.
- Tâche non-supervisée (!)
- Méthodes spectrales consistant à diagonaliser :

$$L = D - A,$$

où A est la matrice d'adjacence et D la matrice des degrés.

artfact
Real Time Machine Learning

Détection de communautés (ou graph clustering)

- Partition des nœuds qui vérifie : beaucoup d'arêtes à l'intérieur des groupes, peu d'arêtes entre les groupes.
- Tâche non-supervisée (!)
- Méthodes spectrales consistant à diagonaliser :

$$L = D - A,$$

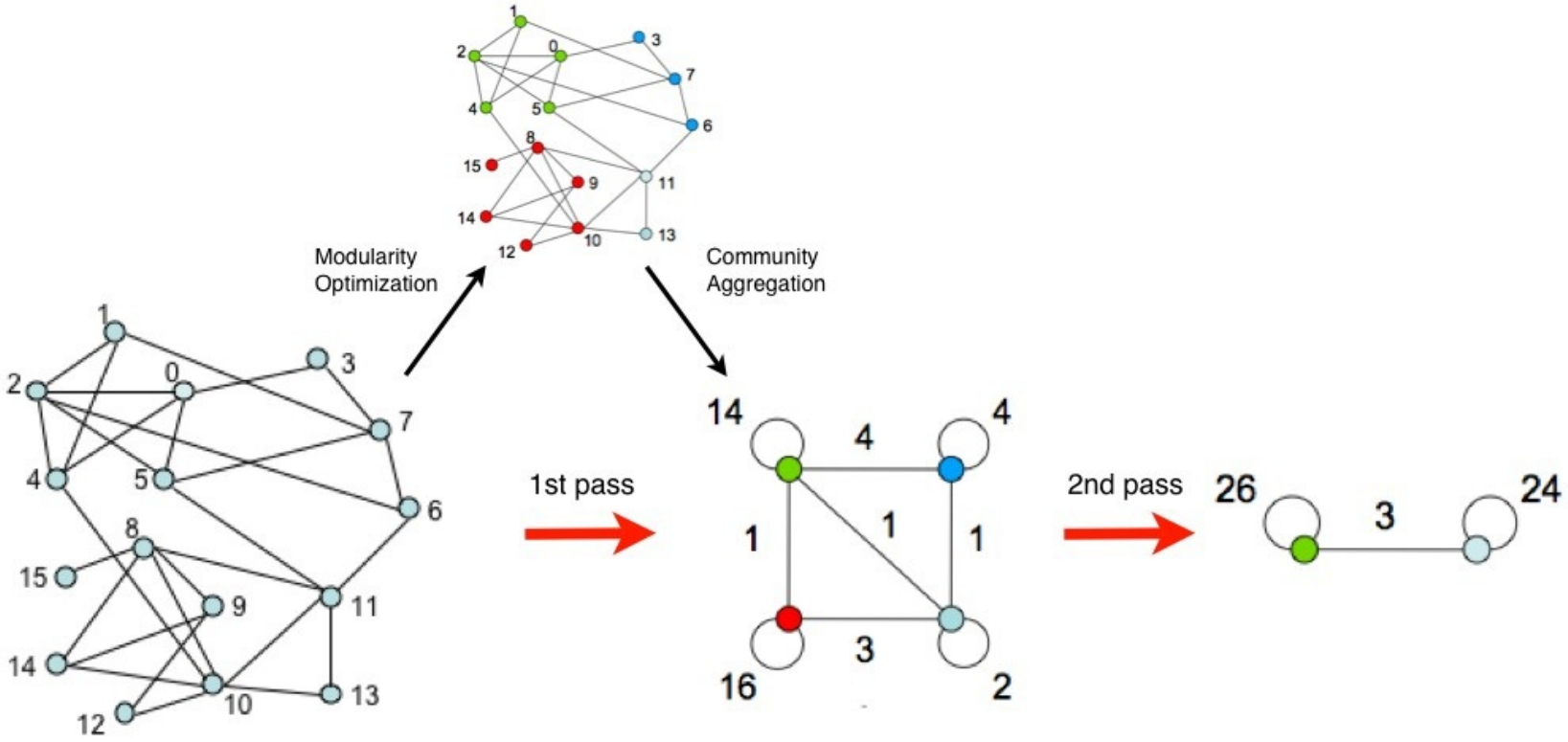
où A est la matrice d'adjacence et D la matrice des degrés.

- Optimisation d'un critère appelé la modularité :

$$\mathcal{M}_G(\mathbf{c}) = \frac{1}{2m} \sum_{(i,j)} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\mathbf{c}}(i,j).$$

artfact
Real Time Machine Learning

Algorithme glouton



artfact
Real Time Machine Learning

Algorithme MCMC

- Select at random a neighborhood C' of the current coloration C
- Accept the proposal with standard Metropolis accept/reject ratio

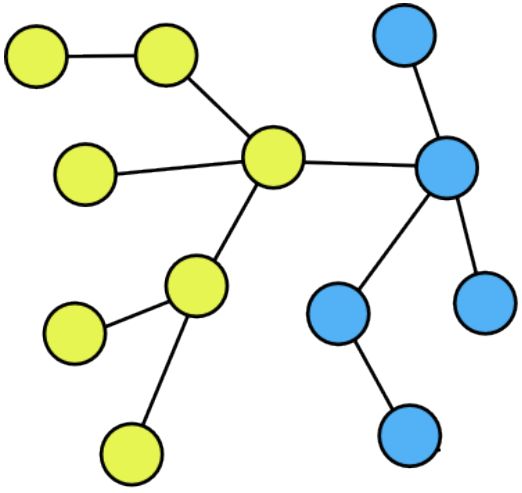
Algorithm 1 MH for Community Detection

1: Initialization $\lambda > 0$, $C^{(0)}$.
2: For $k = 1, \ldots, N$:
3: Draw $C' \sim p(\cdot|C^{(k-1)})$ where $p(\cdot|C^{(k-1)}) \in \mathcal{P}(\mathcal{N}^{C^{(k-1)}})$ is the proposal distribution over $\mathcal{N}^{C^{(k-1)}}$, a neighborhood of $C^{(k-1)}$.
4: Update $C^{(k)} = C'$ with acceptance ratio :

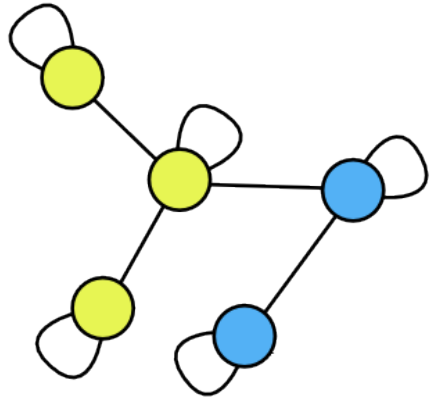
$$\rho = 1 \wedge \left(r_{C^{(k-1)} \to C'} \frac{\exp\left(\lambda Q^{C'}\right)}{\exp\left(\lambda Q^{C^{(k-1)}}\right)} \right), \text{ where } r_{C \to C'} := p(C^{(k-1)}|C')/p(C'|C^{(k-1)}).$$

artfact
Real Time Machine Learning

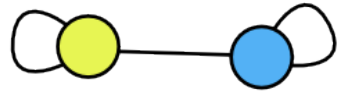
Algorithme final



L = 0



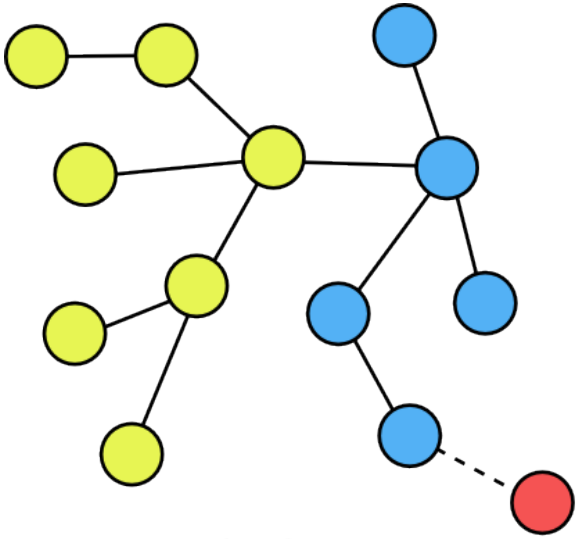
L = 1



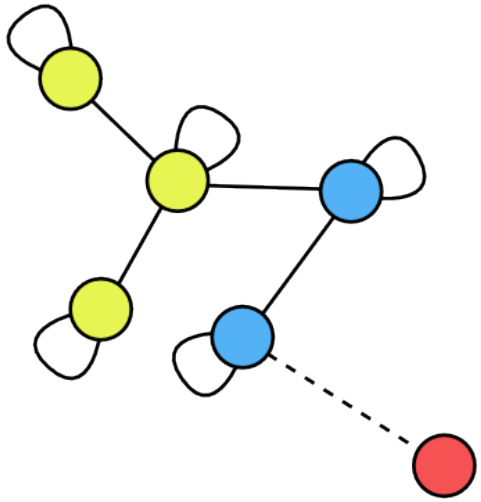
L = 2

artfact
Real Time Machine Learning

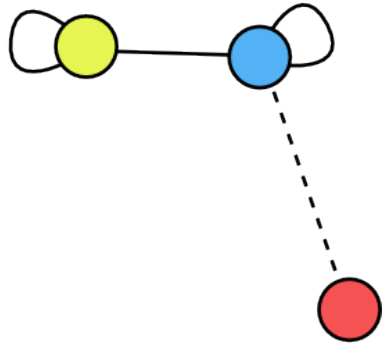
Algorithme final



L = 0



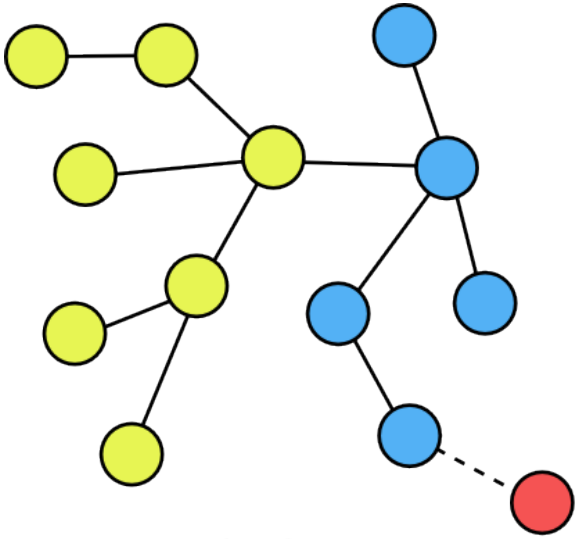
L = 1



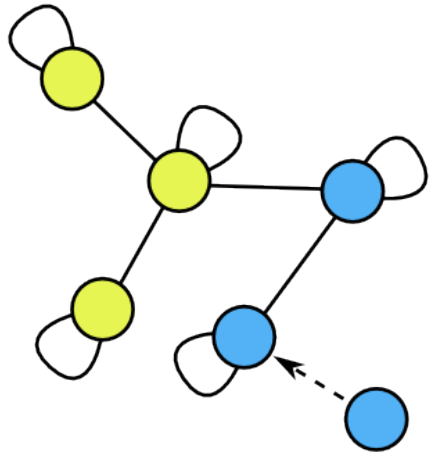
L = 2

artfact
Real Time Machine Learning

Algorithme final



L = 0



L = 1

L = 2

artfact
Real Time Machine Learning

Business case : appel à volontaire

artfact
Real Time Machine Learning

Business case : appel à volontaire

artfact
Real Time Machine Learning

Business case : appel à volontaire

artfact
Real Time Machine Learning

Propaganda :-)

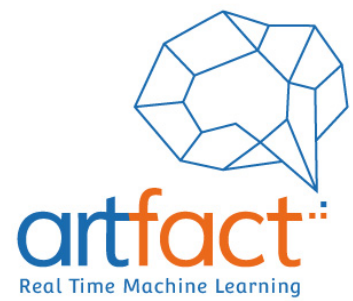
Website : http://www.artfact-online.fr

Propaganda :-)

Website : http://www.artfact-online.fr
Twitter : @artfact-lab

Propaganda :-)

Website : http://www.artfact-online.fr
Twitter : @artfact-lab
Youtube : Le Machine Learning par l'exemple



Propaganda :-)

Website : http://www.artfact-online.fr
Twitter : @artfact-lab
Youtube : Le Machine Learning par l'exemple
Meetup Machine Learning : Nantes Rennes Pau

Propaganda :-)

Website : http://www.artfact-online.fr
Twitter : @artfact-lab
Youtube : Le Machine Learning par l'exemple
Meetup Machine Learning : Nantes Rennes Pau
Plateforme d'innovation ouverte : Learnation Square

Propaganda :-)

- Merci de votre attention ! -

Website : http://www.artfact-online.fr
Twitter : @artfact-lab
Youtube : Le Machine Learning par l'exemple
Meetup Machine Learning : Nantes Rennes Pau
Plateforme d'innovation ouverte : Learnation Square

artfact
Real Time Machine Learning