Apprentissage statistique pour la recherche par les données

T7 Science des données et apprentissage automatique

Sébastien Déjean et Laurent Risser

Ingénieurs de Recherche à l'Institut de Mathématiques de Toulouse

« *A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician* »

Josh Wills

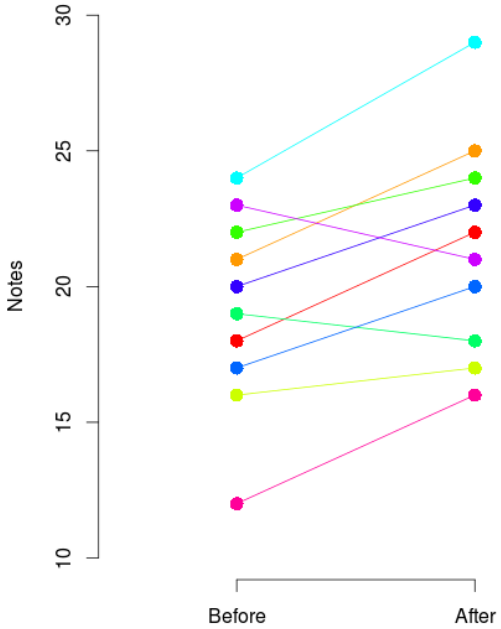
- L'analyse de données est un domaine en pleine explosion actuellement
- Domaine pourtant assez ancien mais en perpétuel renouvellement :
    1) Evolution du volume et de la structure des données quantitatives
    2) Evolution des moyens d'analyse
- Déroulement chronologique depuis les années 40 à nos jours* :
    1) Taille des données
    2) Domaines et problèmes
    3) Méthodologies
    4) Problématiques pratiques
    5) Retours d'expériences

*cf cours INSA du Pr P. Besse (wikistat.fr)

- **Données**
  - → Expérience planifiée
  - → $n \approx 30$ individus observés sur $p \approx 10$ variables
- **Types de problèmes**
  - → Effet significatif d'une molécule, comparaison de groupes, ...
- **Méthodologies**
  - → Tests statistiques
- **En pratique**
  - → Nature des données (indépendantes, appariées)
  - → Nature du test statistique (paramétrique, non paramétrique)

Données appariées

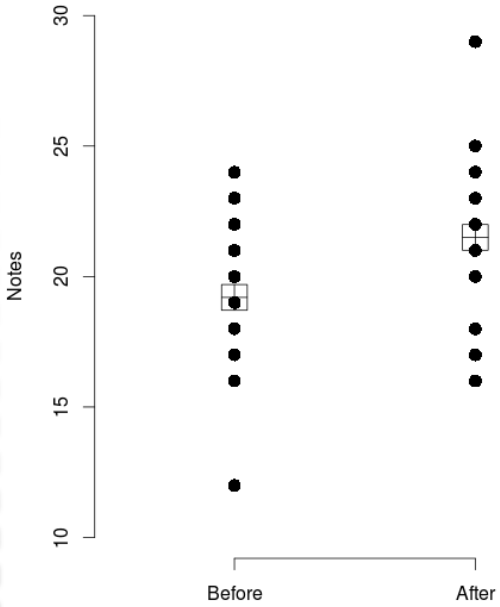
	Before	After
Louise	18	22
Léo	21	25
Emma	16	17
Gabriel	22	24
Chloé	19	18
Adam	24	29
Lola	17	20
Timéo	20	23
Inès	23	21
Raphaël	12	16



> wilcox.test(x,y, **paired=TRUE**)
Wilcoxon signed rank test with continuity correction
V = 5, p-value = **0.02428**
alternative hypothesis: true **location shift** is not equal to 0

> t.test(x,y, **paired=TRUE**)
Paired t-test
t = -3.1461, df = 9, p-value = **0.01181**
alternative hypothesis: true **difference in means** is not equal to 0

Données indépendantes



	Before	After	
Louise	18	21	Lucas
Léo	21	17	Manon
Emma	16	25	Louis
Gabriel	22	24	Jade
Chloé	19	29	Arthur
Adam	24	18	Alice
Lola	17	20	Nathan
Timéo	20	16	Lina
Inès	23	22	Hugo
Raphaël	12	23	Léa

> wilcox.test(x,y, **paired=FALSE**)
Wilcoxon signed rank test with continuity correction
W = 35, p-value = **0.2716**
alternative hypothesis: true **location shift** is not equal to 0

> t.test(x,y, **paired=TRUE**)
Paired t-test
T = -1.3529, df = 9, p-value = **0.1928**
alternative hypothesis: true **difference in means** is not equal to 0

Années 1940-1970 : Punchline

- Attention à la nature des données
- Choix du modèle primordial
- Supervision par l'expert à l'origine des données

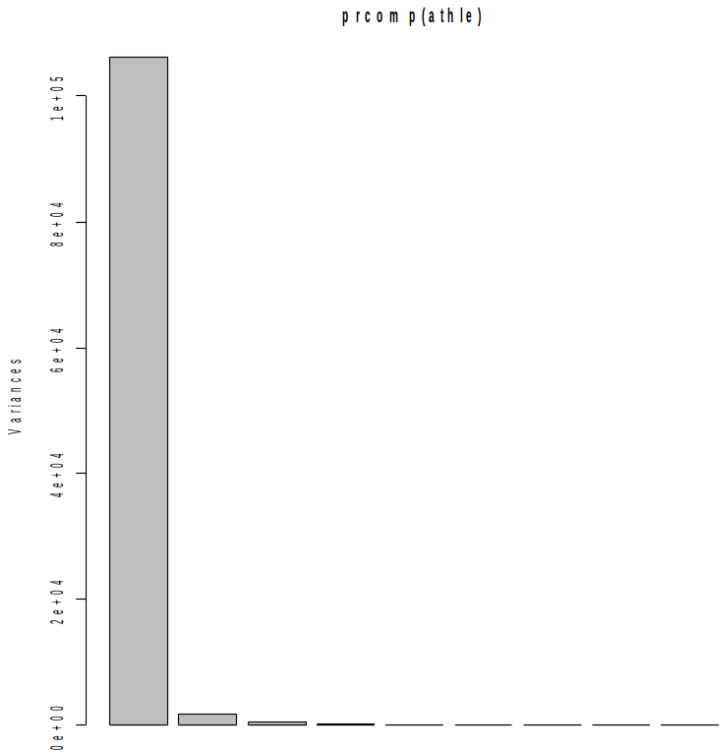
- **Données**
  → Pas forcément plus grandes mais possibilité de les traiter globalement
- **Types de problèmes**
  → Exploration
- **Méthodologies**
  → Analyse des données (M*ultivariate statistics*, Mardia et al., 1979)
- **En pratique**
  → Pré-traitement des données (centrage-réduction, conversion log, ...)

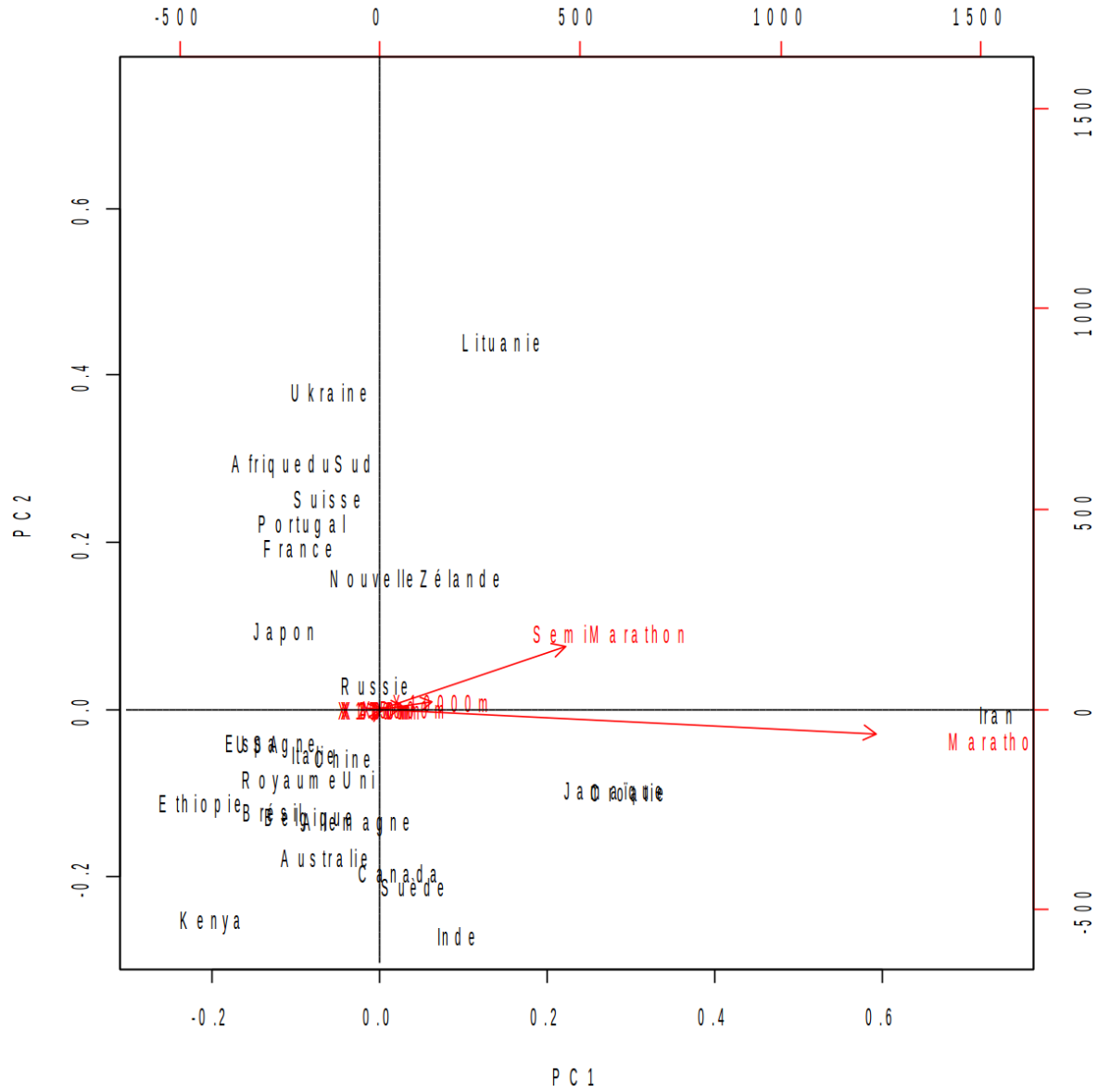
Années 1970-1980 : un exemple

Records nationaux (en secondes) de p = 9 épreuves d'athlétisme pour n = 26 pays

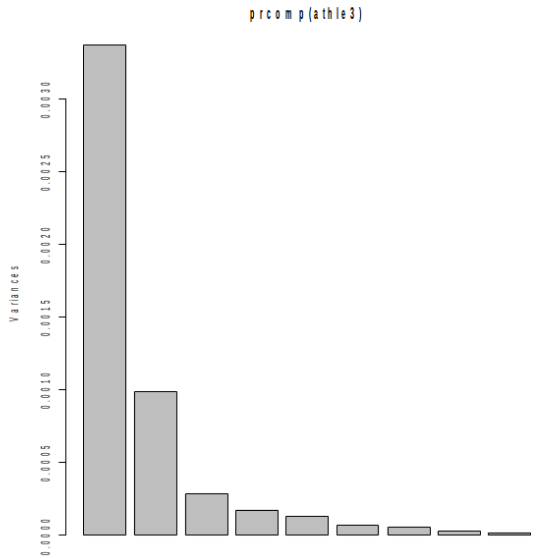
	100m	200m	400m	800m	1500m	5000m	10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueduSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

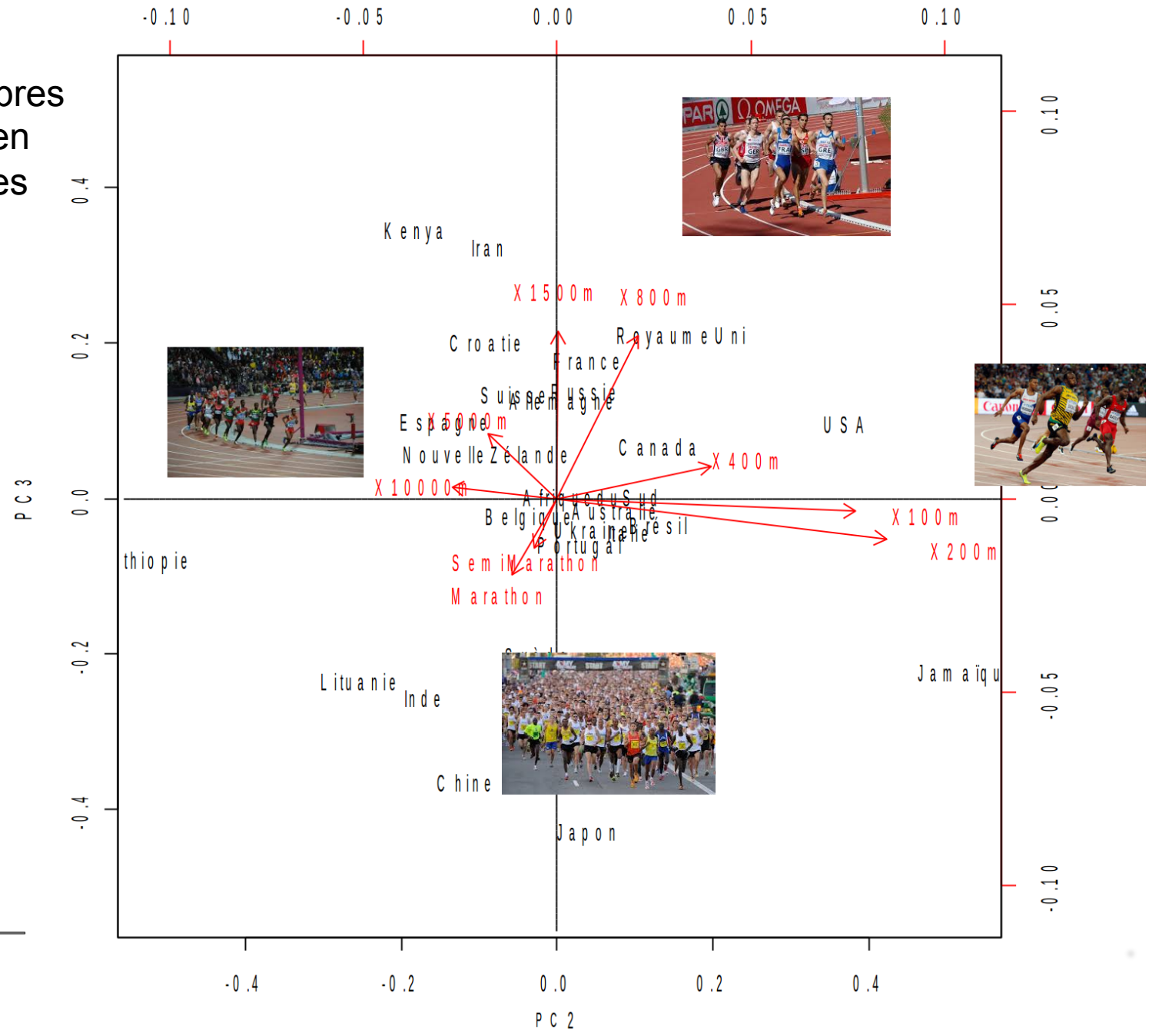
Éboulis des valeurs propres et biplot d'une Analyse en Composantes Principales sur les données **brutes**



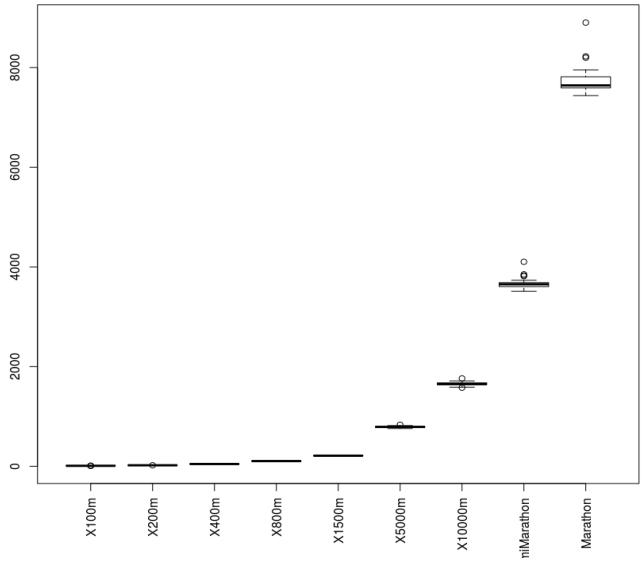


Éboulis des valeurs propres et biplot d'une Analyse en Composantes Principales sur les données **transformées en -log**

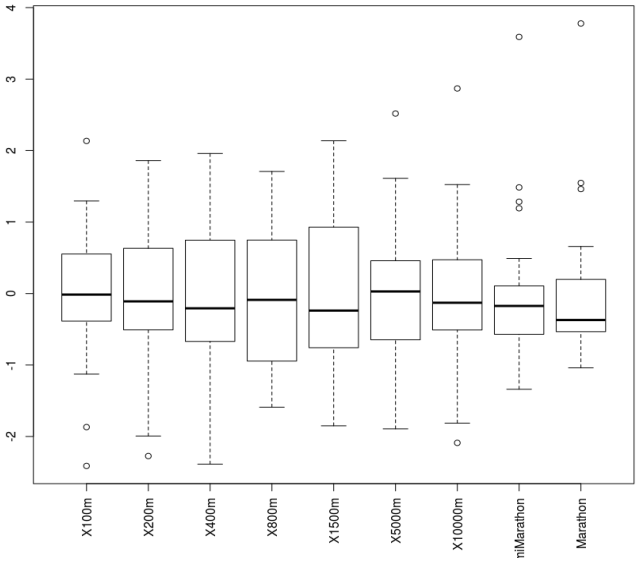




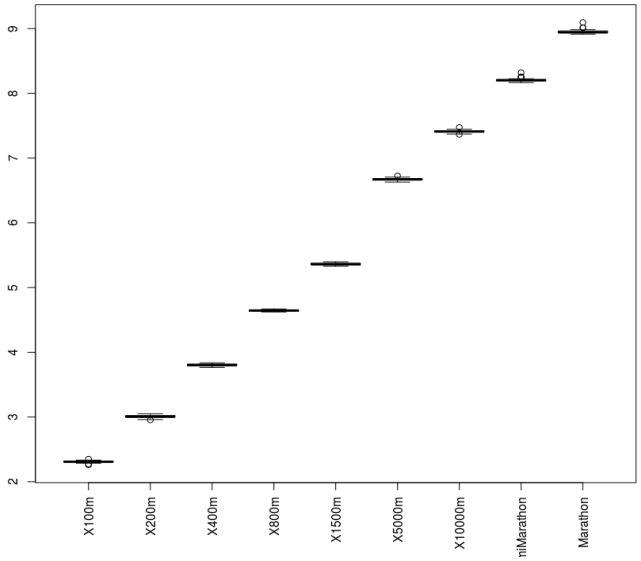
Années 1970-1980 : un exemple ⟶ Impact du pré-traitement



Données brutes



Données centrées-réduites

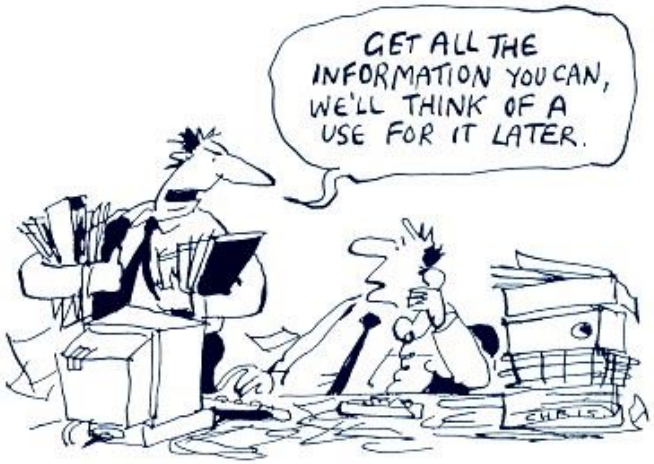


Données converties en log

Années 1970-1980 : Punchline

- Assumer un éventuel pré-traitement des données

- **Données**
  - → Taille des données (n et p) en augmentation
  - → Données qui ne sont plus planifiées
- **Types de problèmes**
  - → La fouille remplace l'exploration. *From Data Mining to Knowledge Discovery (Fayyad et al., 1996)*
- **Méthodologies**
  - → Version « V1 » des méthodes actuelles (SVM, CART, réseaux de neurones)
- **En pratique**
  - → Les logiciels de fouille regroupent dans un même environnement des outils de gestions de données, des techniques exploratoires et de modélisation statistique



- **Données**

  → p explose : dimension des observations p >> nombre d'observations n

- **Types de problèmes**

  → Exploration

- **Méthodologies**

  → Version « V2 » des méthodes actuelles : CART→ random forest, régularisation (LASSO, ridge, elastic net), parcimonie, optimisation, sélection de modèle

- **En pratique**

  → Fléau de la dimension, interprétabilité des résultats

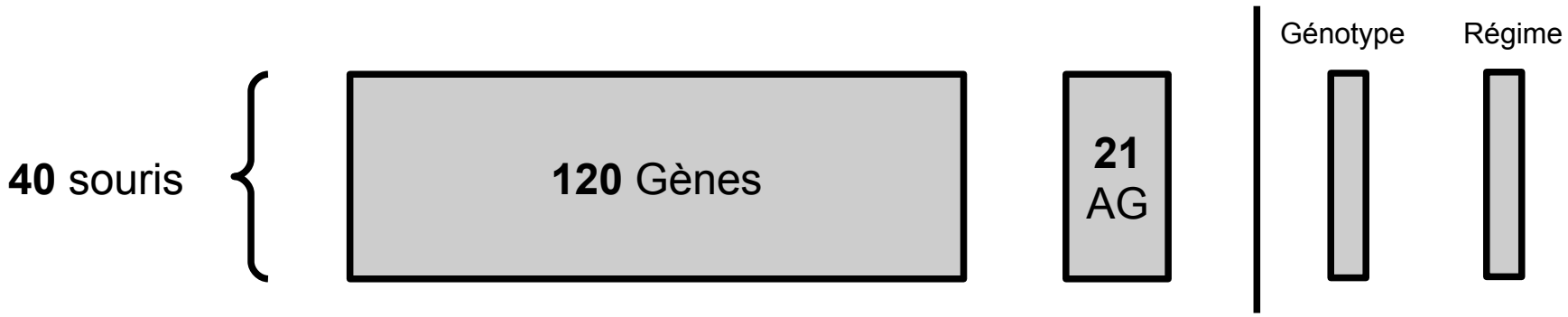
Années 2000 : Un exemple en biologie

Données acquises au laboratoire de pharmacologie et toxicologie de l'INRA de Toulouse. Elles proviennent d'une étude de nutrition chez la souris. Pour 40 souris, nous disposons :

- données d'expression de **120** gènes recueillies sur membrane nylon avec marquage radioactif,
- mesures de **21** acides gras hépatiques.

Par ailleurs, les n=**40** souris sont réparties selon deux facteurs :

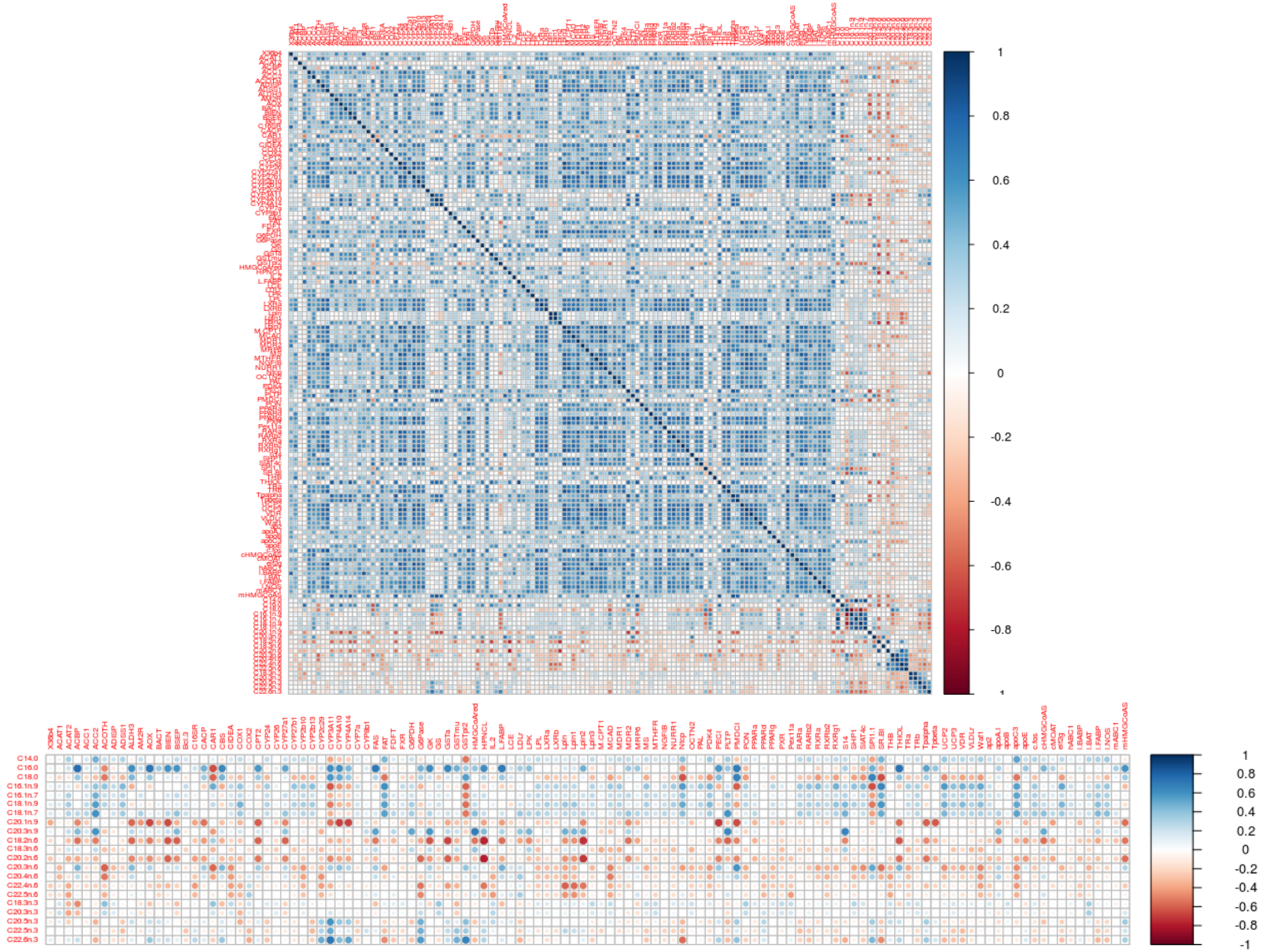
- Génotype (**2** modalités) : les souris sont soit de type sauvage (wt) soit génétiquement modifiées (PPAR) ; **20** souris dans chaque cas.
- Régime (**5** modalités) : les **5** régimes alimentaires sont notés *ref*, *coc*, *fish*, *lin*, *sun* ; **4** souris de chaque génotype sont soumises à chaque régime alimentaire.

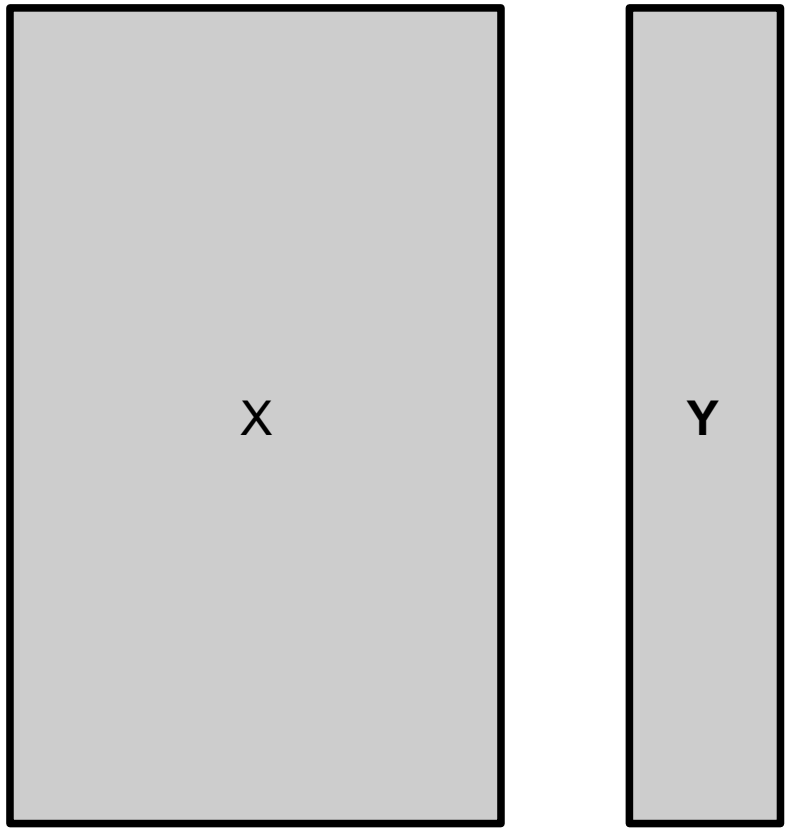


- P. Martin, H. Guillou, F. Lasserre, S. Déjean, A. Lan, J-M. Pascussi, M. San Cristobal, P. Legrand, P. Besse, T. Pineau (2007). Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. Hepatology

Années 2000 : Un exemple en biologie

Problème : proposer une méthode permettant de mettre en évidence et de représenter les relations entre les gènes et les acides gras.





Objectif de la méthode : explorer les relations linéaires entre deux ensembles de variables quantitatives observées sur les mêmes individus.

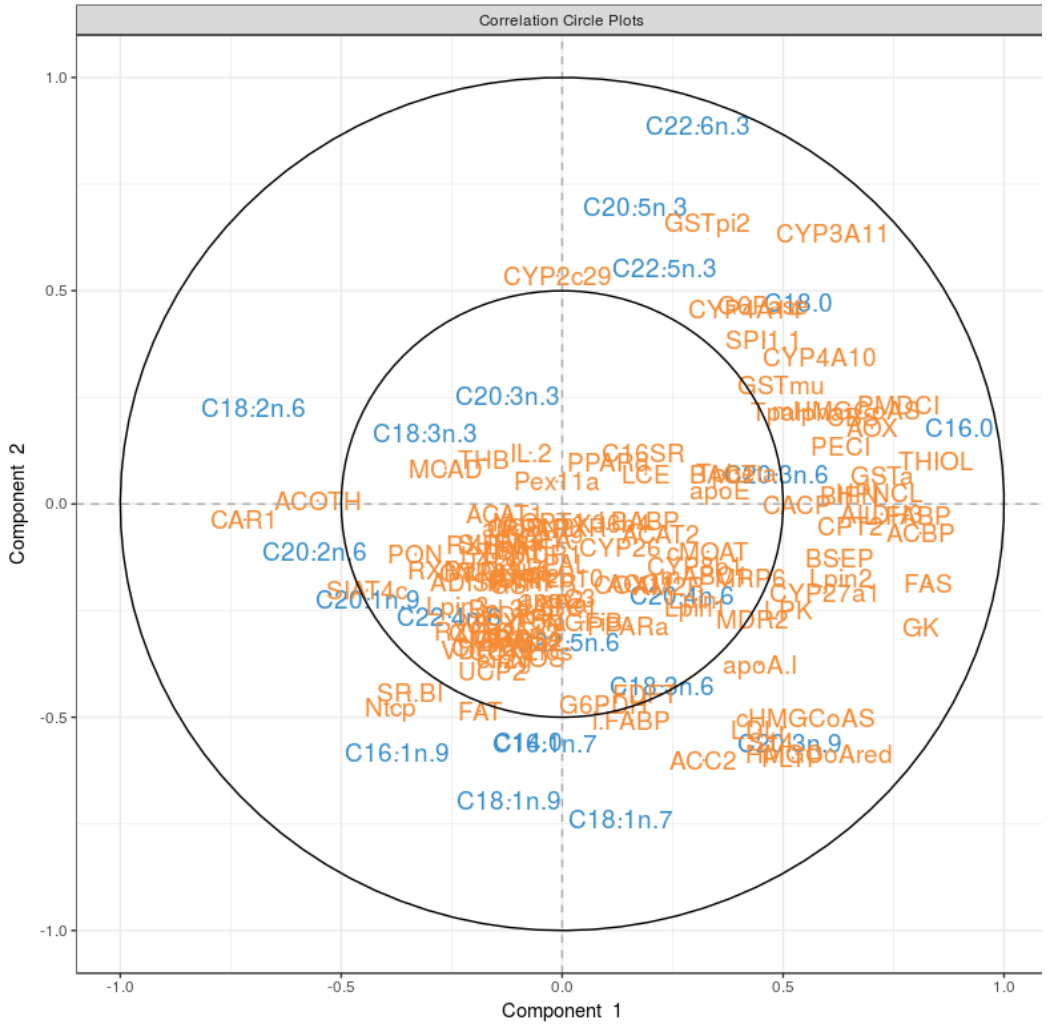
Principe : trouver la corrélation maximale entre une combinaison linéaire des variables X et une combinaison linéaire des variables Y pour avoir la première paire de variables canoniques. Itérer pour les suivantes.

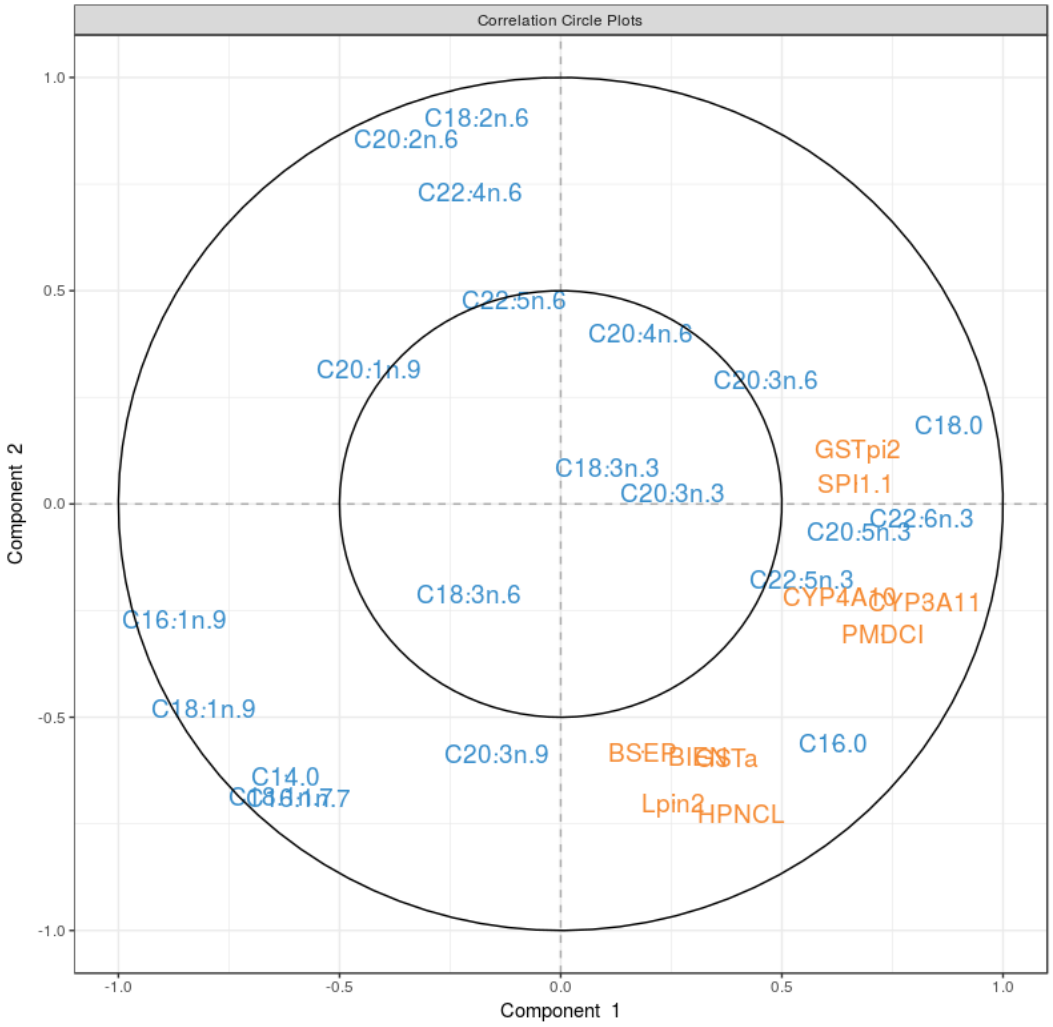
Limites : ne peut fonctionner qu'avec un nombre « suffisant » d'individus.

Alternatives : version régularisée (*ridge*) ou parcimonieuse (*sparse*).

I. González, S. Déjean, P.G.P. Martin, O. Gonçalves, P. Besse, A. Baccini (2009). Highlighting Relationships Between Heteregeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis. Journal of Biological Systems, 17(2), 173-199.

Années 2000 : Un exemple en biologie





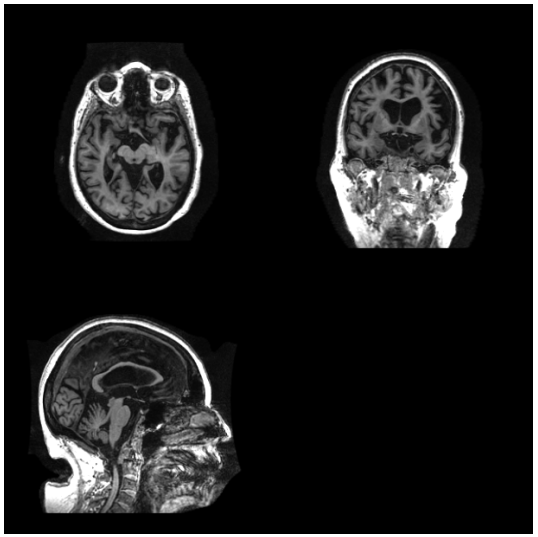
version *ridge* : l'ensemble des variables est conservé.

version *Lasso* : les variables sont sélectionnées.

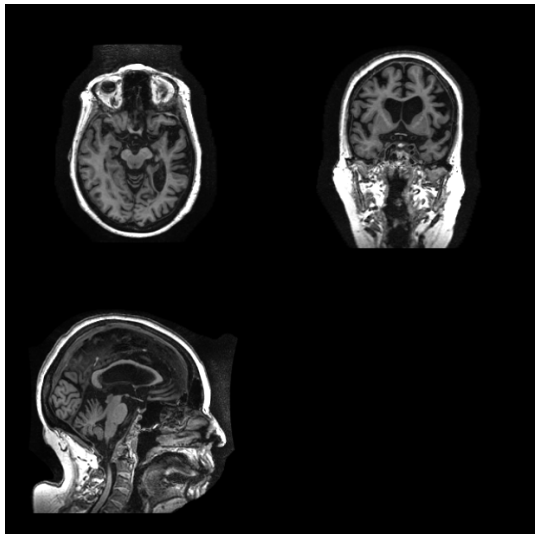
Années 2000 : Un exemple en imagerie médicale

Context du projet :

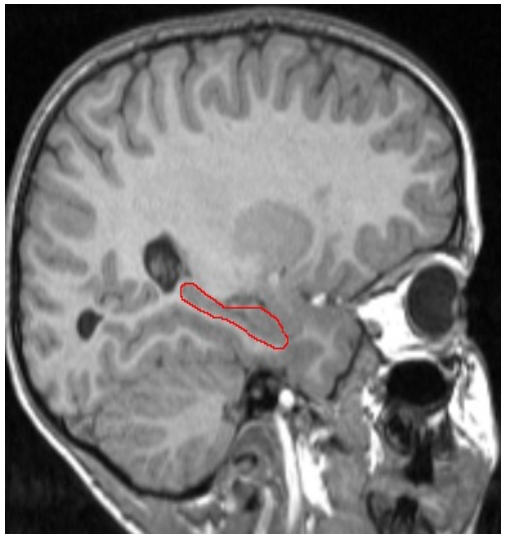
- Collaboration FX Vialard et J.B. Fiot (Univ. Dauphine)
- Base de données ADNI* de 1500 patients dont certains développent la maladie d'Alzheimer.
- IRM du cerveau des patients à différent temps d'acquisition
- Etat du patient connu a postériori
- Prédiction maladie d'Alheimer en fonction de l'évolution morphométrique du cerveau ?



[Baseline]



[Baseline + 12 mois]



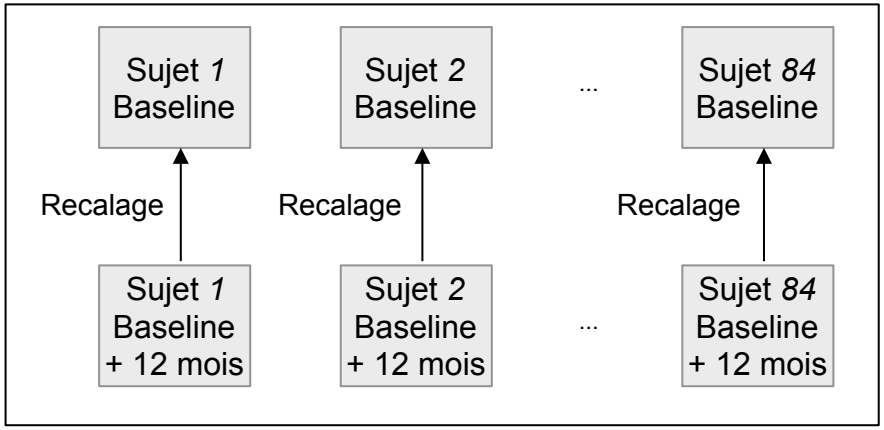
Hippocampe

Focalisation sur l'hippocampe :

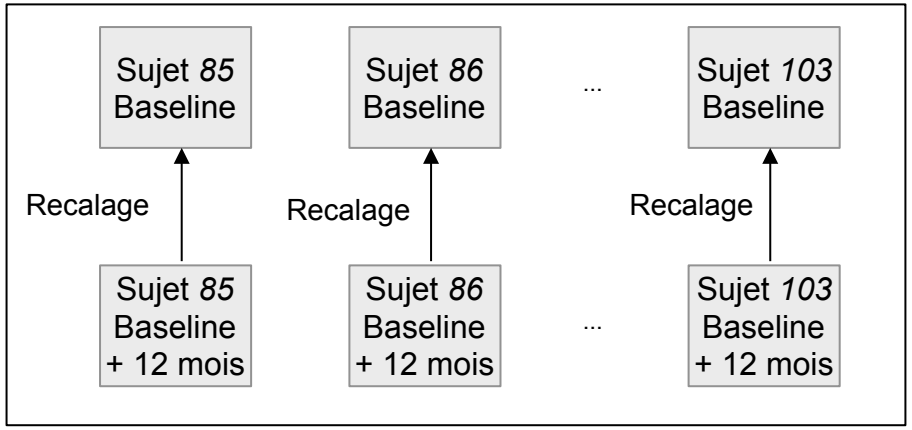
- Segmentations de [Mueller et al., Neuroimaging Clin. N. Am, 2005]
- *[Baseline]* : 103 patients sont MCI
- *[Baseline + 12 mois]* : 84 patients sont MCI / 19 patients sont AD

* http://adni.bmap.ucla.edu/

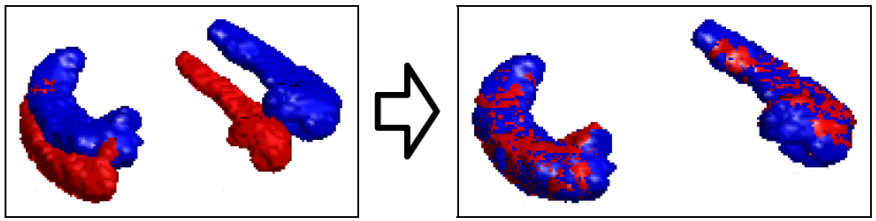
Pré-traitement 1 : Suivi des déformations entre [Baseline] et [Baseline + 12 mois]



Groupe des MCI

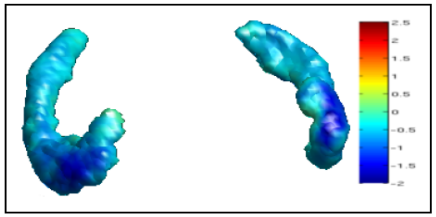


Groupe des AD



(a) Recalage rigide

[Ourselin et al. Im Vis Comp., 2001]

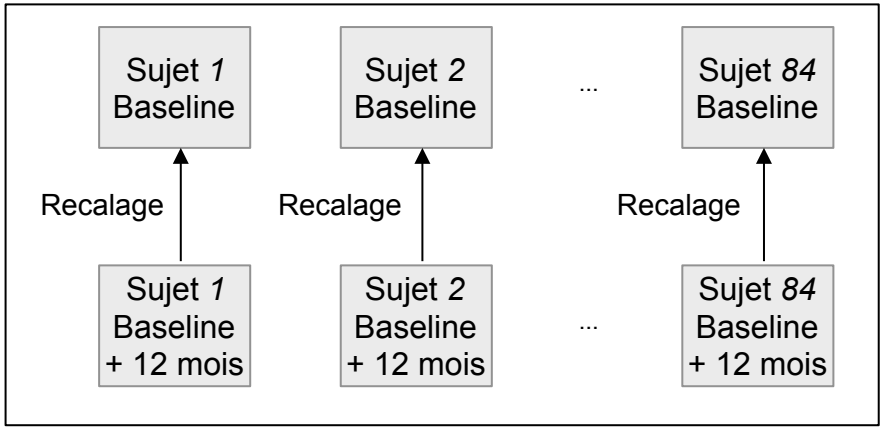


(b) Recalage élastique

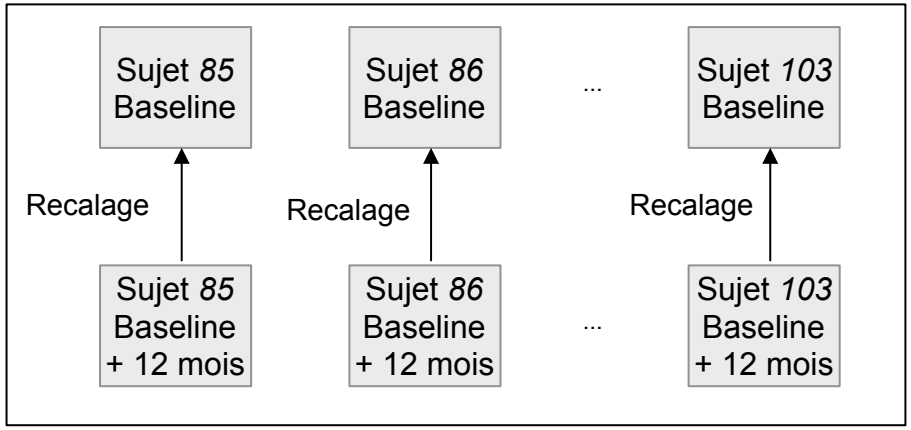
[Vialard et al. IJCV, 2012]

Pré-traitement 2 : Transport des marqueurs d'évolution sur une forme *moyenne*

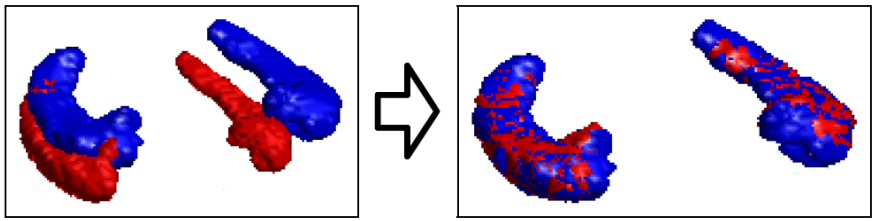
Pré-traitement 1 : Suivi des déformations entre [Baseline] et [Baseline + 12 mois]



Groupe des MCI

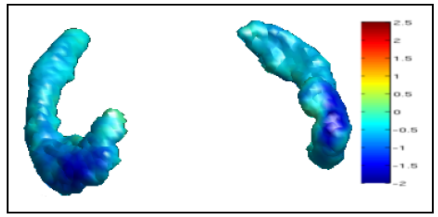


Groupe des AD



(a) Recalage rigide

[Ourselin et al. Im Vis Comp., 2001]



(b) Recalage élastique

[Vialard et al. IJCV, 2012]

Pré-traitement 2 : Transport des marqueurs d'évolution sur une forme *moyenne*

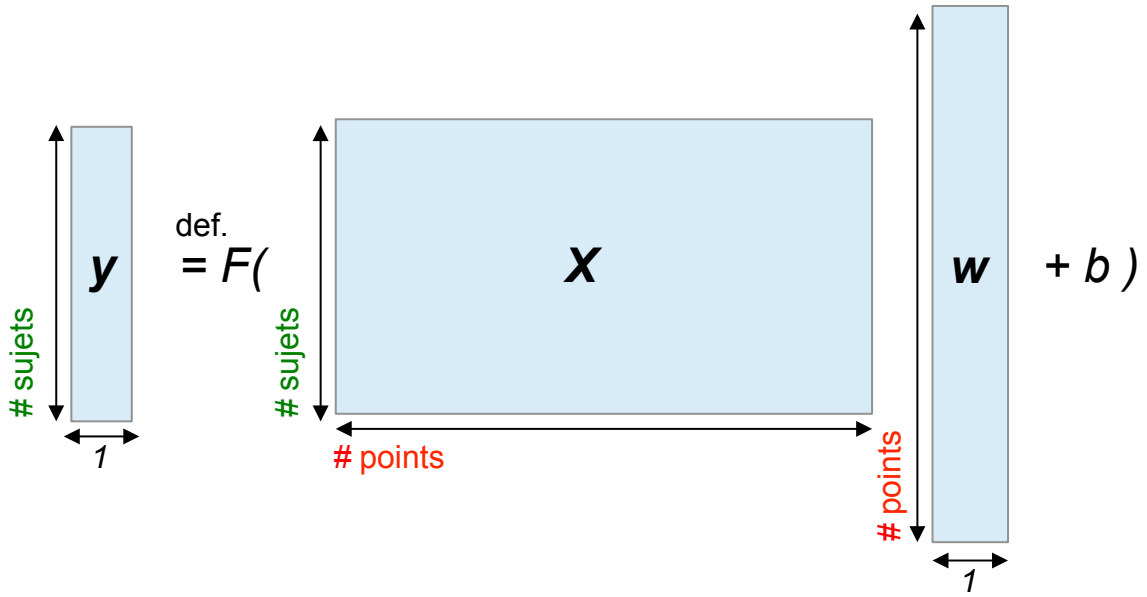
Pour chacun des n = 103 sujets :

- $\boldsymbol{x}_i$: Observation de l'évolution de la forme sur environ p = 20000 points
- y_i : Etat AD ou MCI

Apprentissage de marqueurs discriminants ???

Modèle prédictif de régression logistique qui défini la probabilité des y_i en fonctions des $\mathbf{x_i}$:

$$p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, b\right) \overset{\text{def.}}{=} \frac{1}{1+\exp\left(-y_i \ \left(\mathbf{x}_i^T \mathbf{w}+b\right)\right)}$$



Où :

$\boldsymbol{X} \in \boldsymbol{R}^{n*p}$: matrice des *n = 103* observations de dimension *p = 20000*

$\boldsymbol{y} \in \{\mp 1\}^n$: Etat (*AD = -1 / MCI = 1*)

$(\boldsymbol{w}, b) \in \boldsymbol{R}^p * \boldsymbol{R}$: parametres à estimer

Optimisation de la log-vraisemblance :

Paramètre de régularisation

$$\text{Find } (\hat{\mathbf{w}}, \hat{b}) \text{ in } \underset{\mathbf{w},b}{\operatorname{argmin}} \ \mathcal{L}(\mathbf{w}, b) + \lambda J(\mathbf{w})$$

où :

$$\mathcal{L}(\mathbf{w}, b) \overset{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^{n} \log\left(1+\exp\left(-y_i(\mathbf{x}_i^T \mathbf{w}+b)\right)\right)$$

Exploration de différents modèles de régularisation :

(1) Ridge : $J(\mathbf{w}) = ||\mathbf{w}||_2$

(2) LASSO : $J(\mathbf{w}) = ||\mathbf{w}||_1$

(3) Elastic net : $J(\mathbf{w}) = \lambda_1||\mathbf{w}||_1 + \lambda_2||\mathbf{w}||_2$

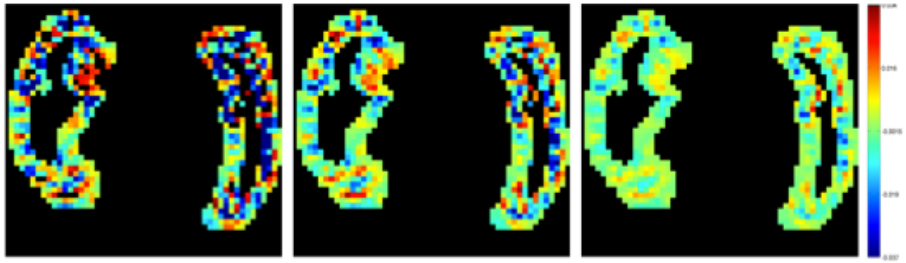
(4) Sobolev semi-norm: $J(\mathbf{w}) = \sum_{\omega\in\mathbf{\Omega}} |\nabla_{\mathbf{\Omega}}\mathbf{w}(\omega)|$

(5) Total Variation : $J(\mathbf{w}) = \left(\sum_{\omega\in\mathbf{\Omega}} |\nabla_{\mathbf{\Omega}}\mathbf{w}(\omega)|^2\right)^{1/2}$

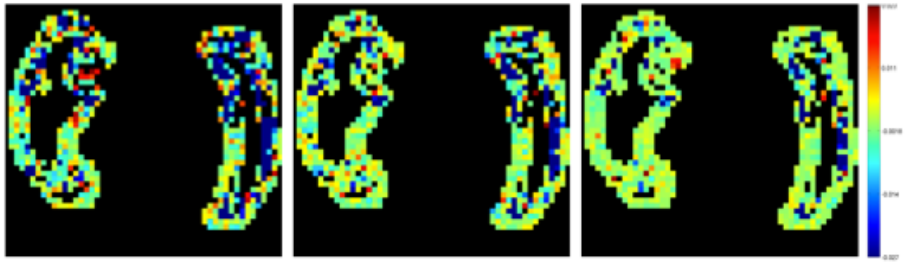
(6) Fused LASSO : $J(\mathbf{w}) = \lambda_1||\mathbf{w}||_1 + \lambda_2 \sum_{\omega\in\mathbf{\Omega}} |\nabla_{\mathbf{\Omega}}\mathbf{w}(\omega)|$

Optimisation de ***w*** avec :
Lewis & Overton, *Nonsmooth optimization via quasi-Newton methods*. Math. Programming 2012

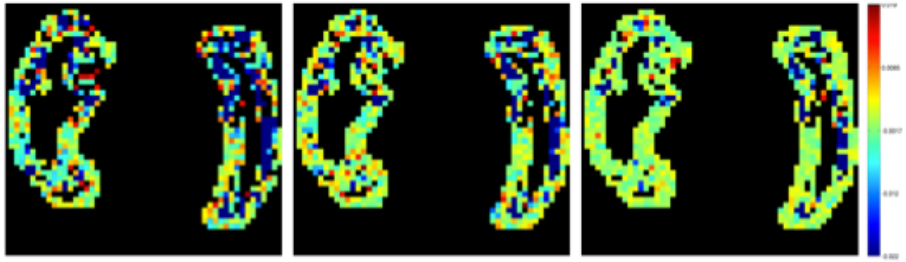
Années 2000 : Un exemple en imagerie médicale



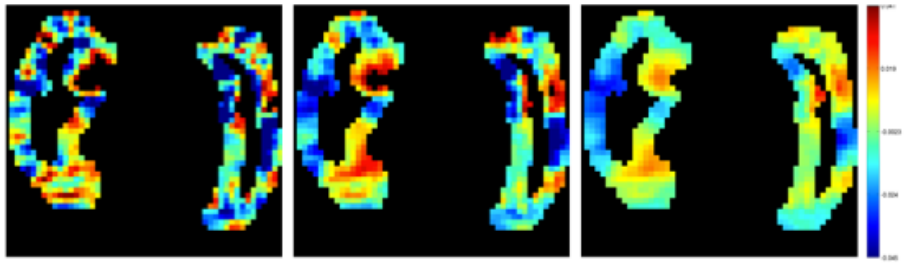
(1) Ridge



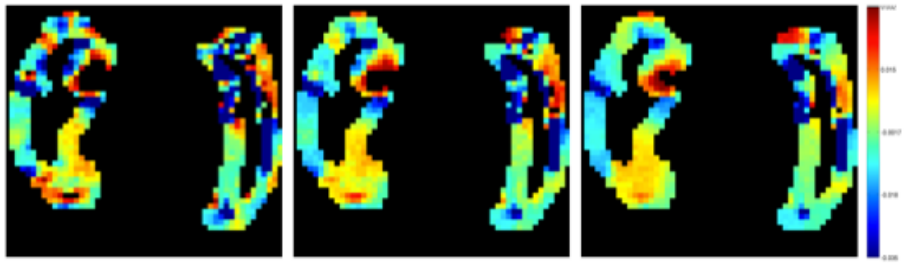
(2) LASSO



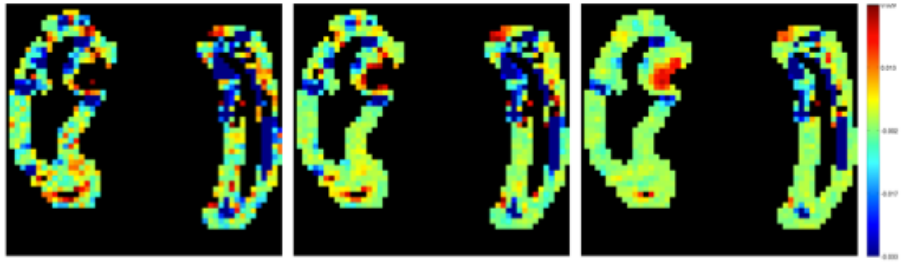
(3) Elastic net



(4) Sobolev semi-norm



(5) Total Variation



(6) Fused LASSO

Représentation de $\boldsymbol{w}$ pour trois λ sur un plan de l'hippocampe :

- Bleu et rouge : forte influence locale
- Vert : peu ou pas d'influence locale

Années 2000 : Un exemple en imagerie médicale

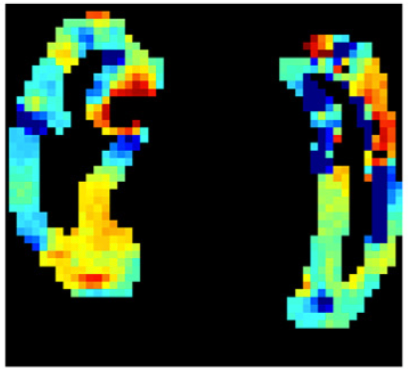
Résultats obtenus avec une méthode de cross validation (ici leave-10%-out) :

- Spec+Sens = 2 ⟶ bonne prédiction dans 100% des cas
- Spec+Sens = 1 ⟶ pile ou face aurait le même pouvoir prédictif

Regularization		λ **range**	$\hat{\lambda}$ **(optimal** λ**)**	**Spec+ Sens**
None		0	0	1.00
Standard	LASSO	$[10^{-9},10^{0}]$	0.01	1.04
	Ridge	$[10^{-9},10^{0}]$	0.001	1.06
	Elastic Net	$[10^{-9},10^{0}]^2$	$\begin{cases}\hat{\lambda}_1 = 0.01\\ \hat{\lambda}_2 = 1\end{cases}$	1.13
Spatial	Sobolev	$[10^{-9},10^{7}]$	10^4	1.17
	Total Variation	$[10^{-9},10^{0}]$	0.01	**1.31**
	Fused LASSO	$[10^{-9},10^{0}]^2$	$\begin{cases}\hat{\lambda}_1 = 0.01\\ \hat{\lambda}_2 = 10^{-4}\end{cases}$	**1.32**

Meilleurs résultats avec une régularisation en pertinente avec les données :

- Tient compte de la distribution spatial
- Permet quelques transitions franches



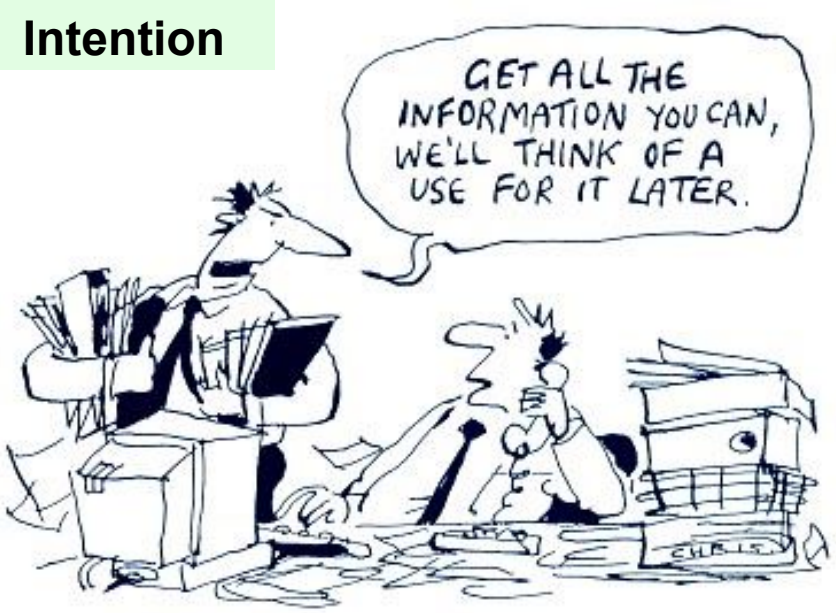
[Fiot J.B. et al., NeuroImage: Clinical, 2012]

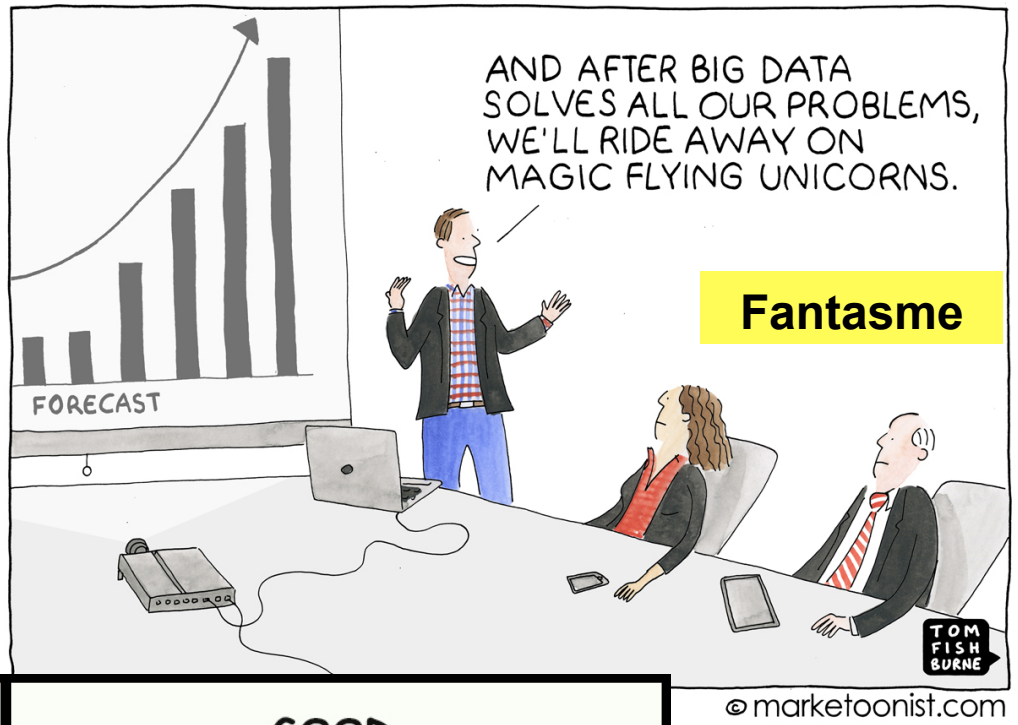
Années 2000 : Punchline

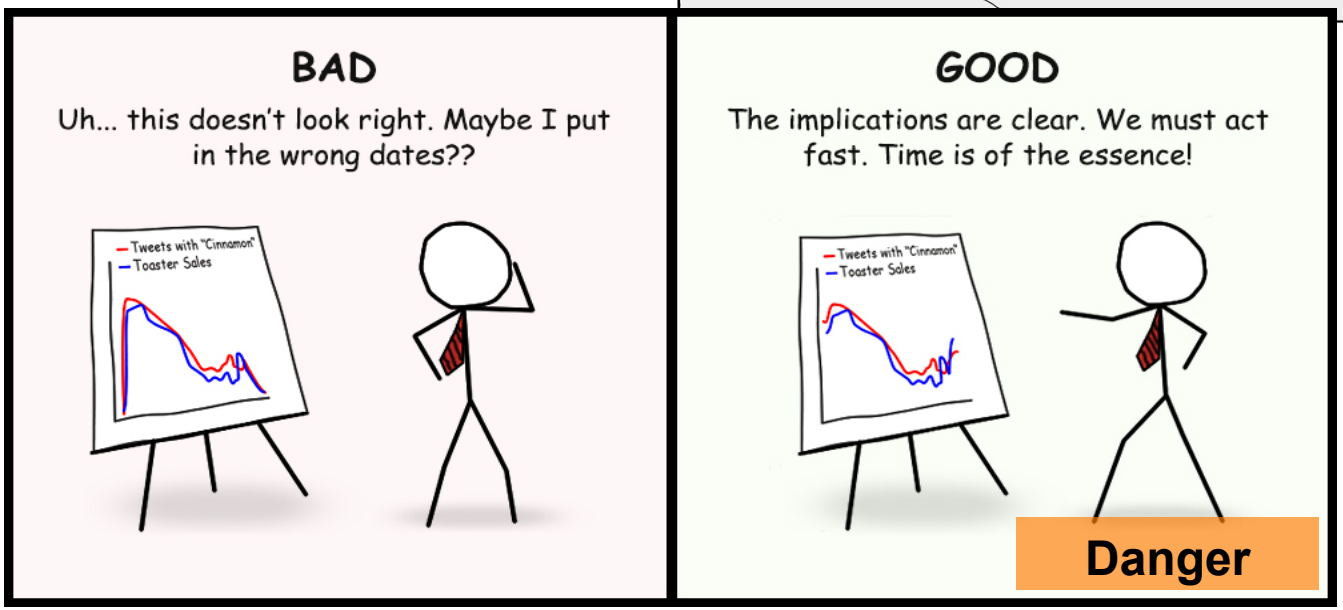
- Besoins mathématiques croissants
- Choix des modèles
- Interprétabilité des méthodes « boîte noire »

- **Données**
  - → n explose
- **Types de problèmes**
  - → Prédiction temps réel
  - → Accroissement continu des bases de données d'apprentissage
  - → Hétérogénéité des données
- **Méthodologies**
  - → Retour vers le futur (corrélation, régression logistique)
- **En pratique**
  - → Les données ne tiennent plus sur une machine
  - → Le hardware doit suivre
  - → Temps de calculs

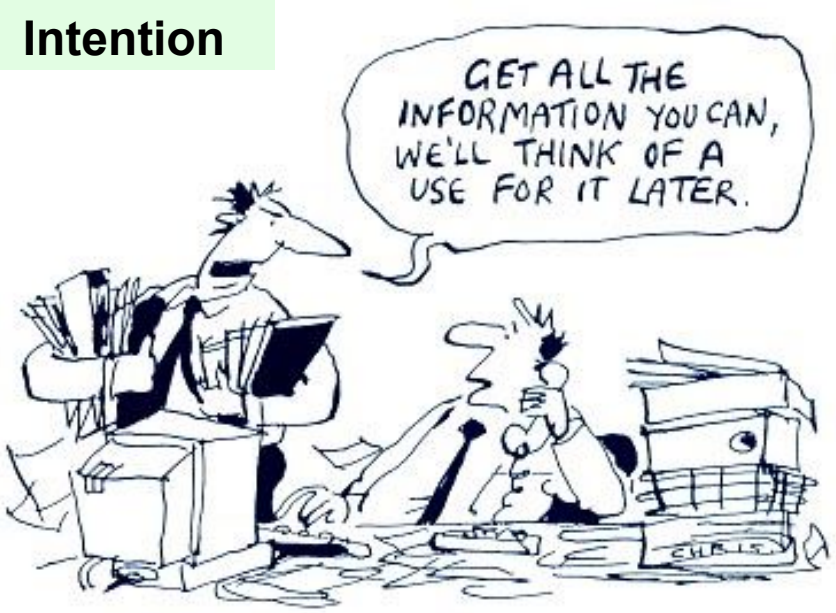
Intention

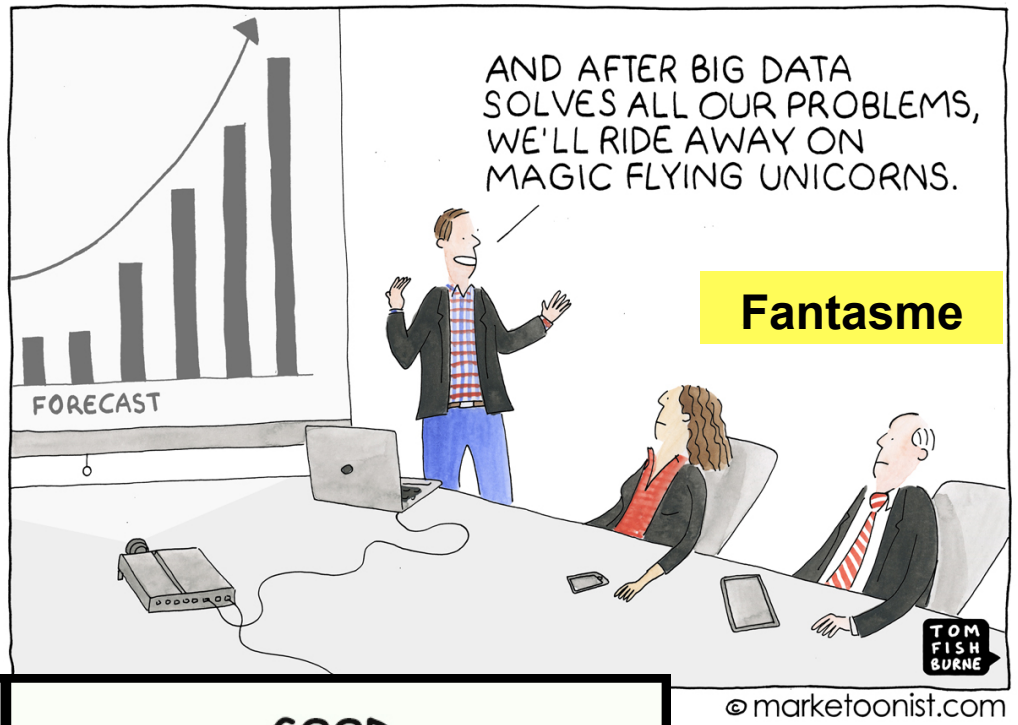

GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.


AND AFTER BIG DATA SOLVES ALL OUR PROBLEMS, WE'LL RIDE AWAY ON MAGIC FLYING UNICORNS.
FORECAST
Fantasme
TOM FISH BURNE
© marketoonist.com

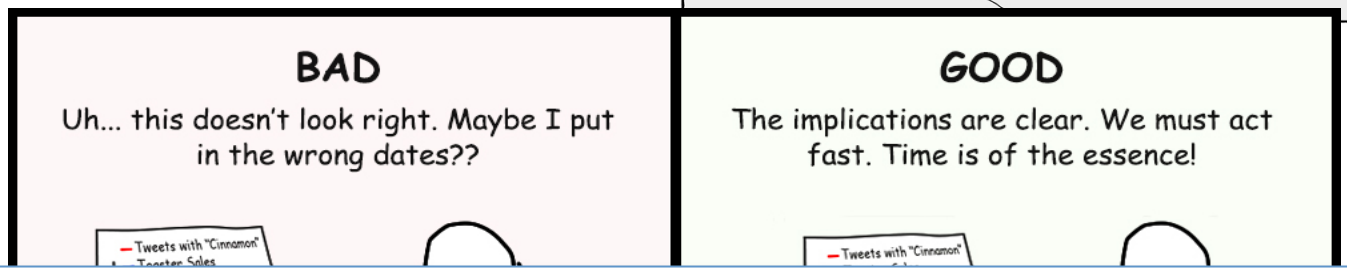

BAD
Uh... this doesn't look right. Maybe I put in the wrong dates??
Tweets with "Cinnamon"
Toaster Sales
GOOD
The implications are clear. We must act fast. Time is of the essence!
Tweets with "Cinnamon"
Toaster Sales
Danger

Intention





Fantasme



- http://www.tylervigen.com/spurious-correlations
-
- https://www.google.com/trends/correlate/

Spurious correlations

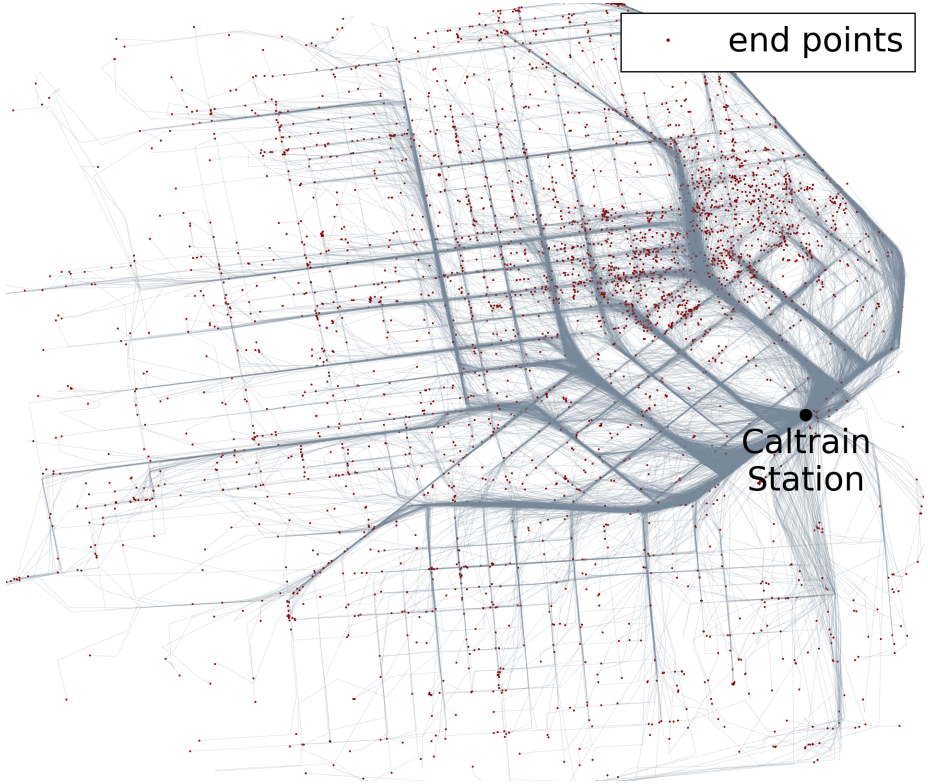
Google Correlate

Années 2010s : Un exemple d'analyse du réseau routier

Contexte :

- Thèse de B. Guillouet avec P. Besse, J.M. Loubes et F. Royer (INSA/UT3/IMT - Datasio)
- Kaggle « ECML/PKDD 15 : Taxi Trajectory Prediction»
- Prédiction de la destination d'un trajet de taxi en fonction de son trajet courant.
- (Porto) 20000 trajets connus / (San Fransisco) 4000 trajets connus
- Modèle de prédiction temps réel et automatiquement adaptable à différentes villes ?

20000 trajets sur Porto



4000 trajets sur San Francisco

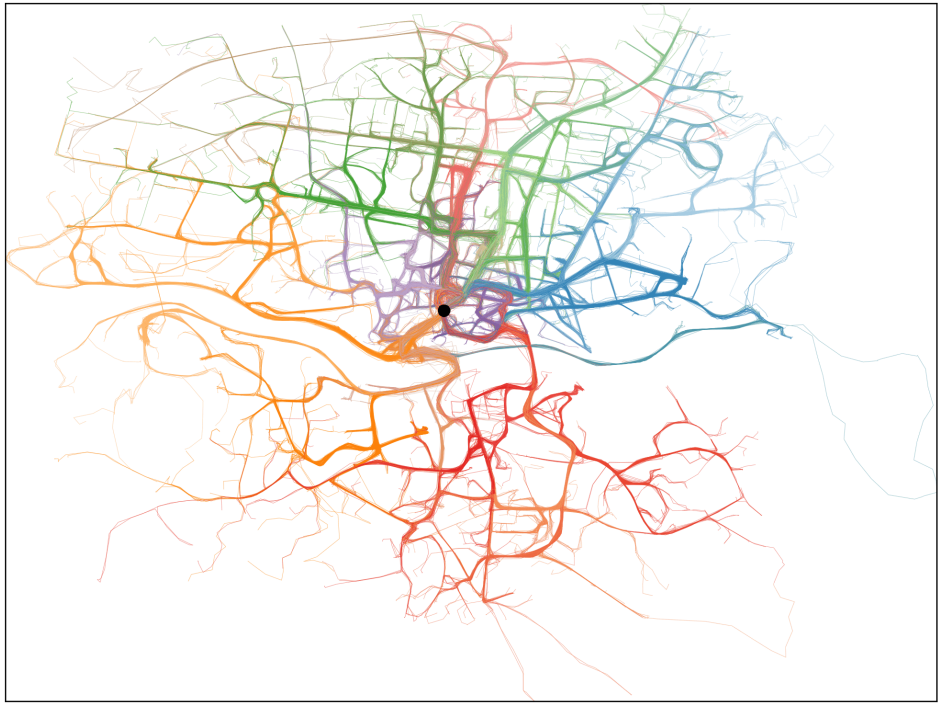
Années 2010s : Un exemple d'analyse du réseau routier

Résolution du problème :

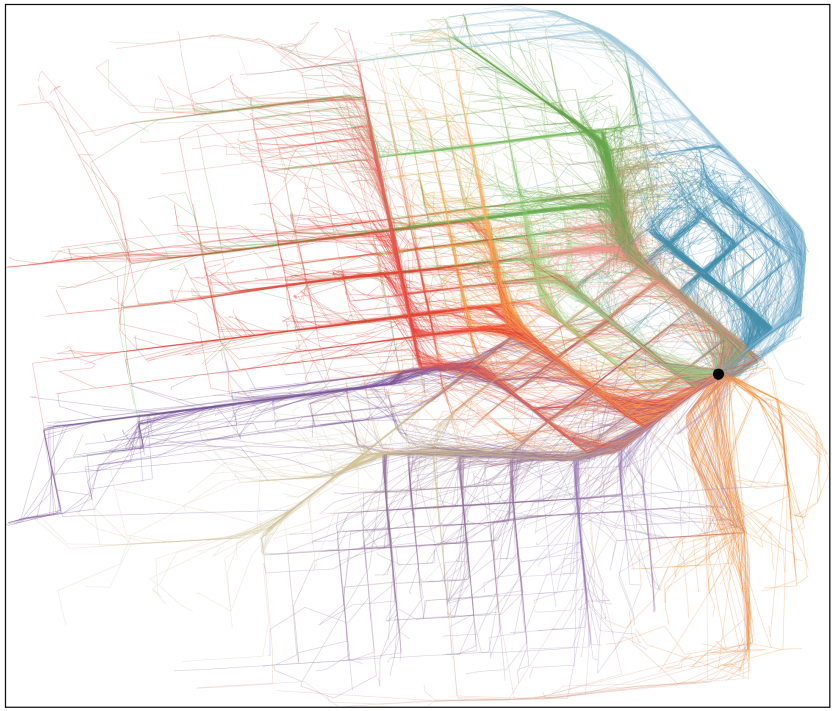
- Prétraitements des données.
- Apprentissage assez demandeur en terme de calculs mais réaliste sur un serveur standard.
- Modèle de mélange Gaussiens / Maximum de vraisemblance / Critère d'Information Bayésien

Passage à l'échelle :

- Pré-étape de clustering
- Clustering hiérarchique des parcours en fonction d'une mesure de distance adaptée

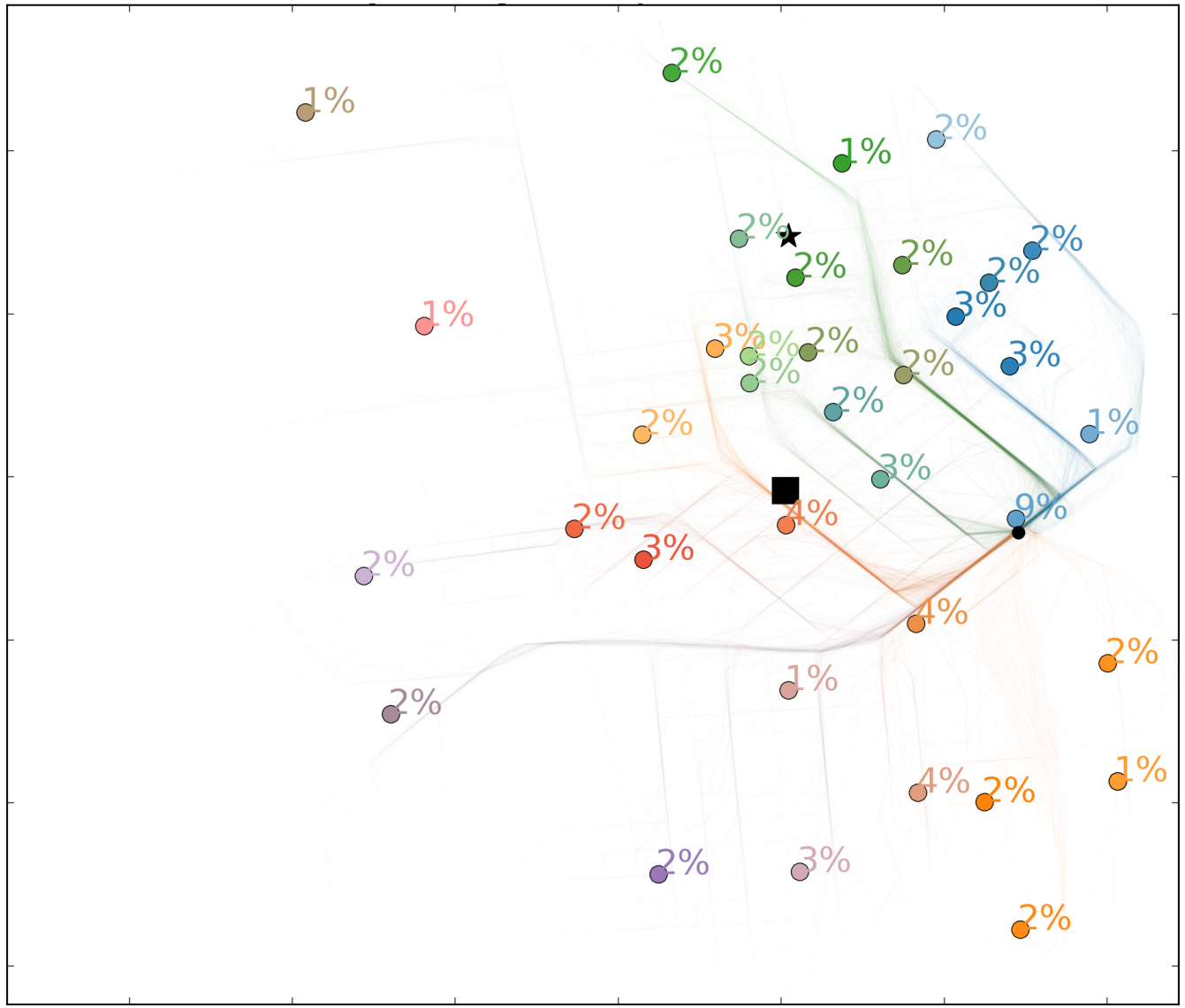


Clusters de trajets sur Porto

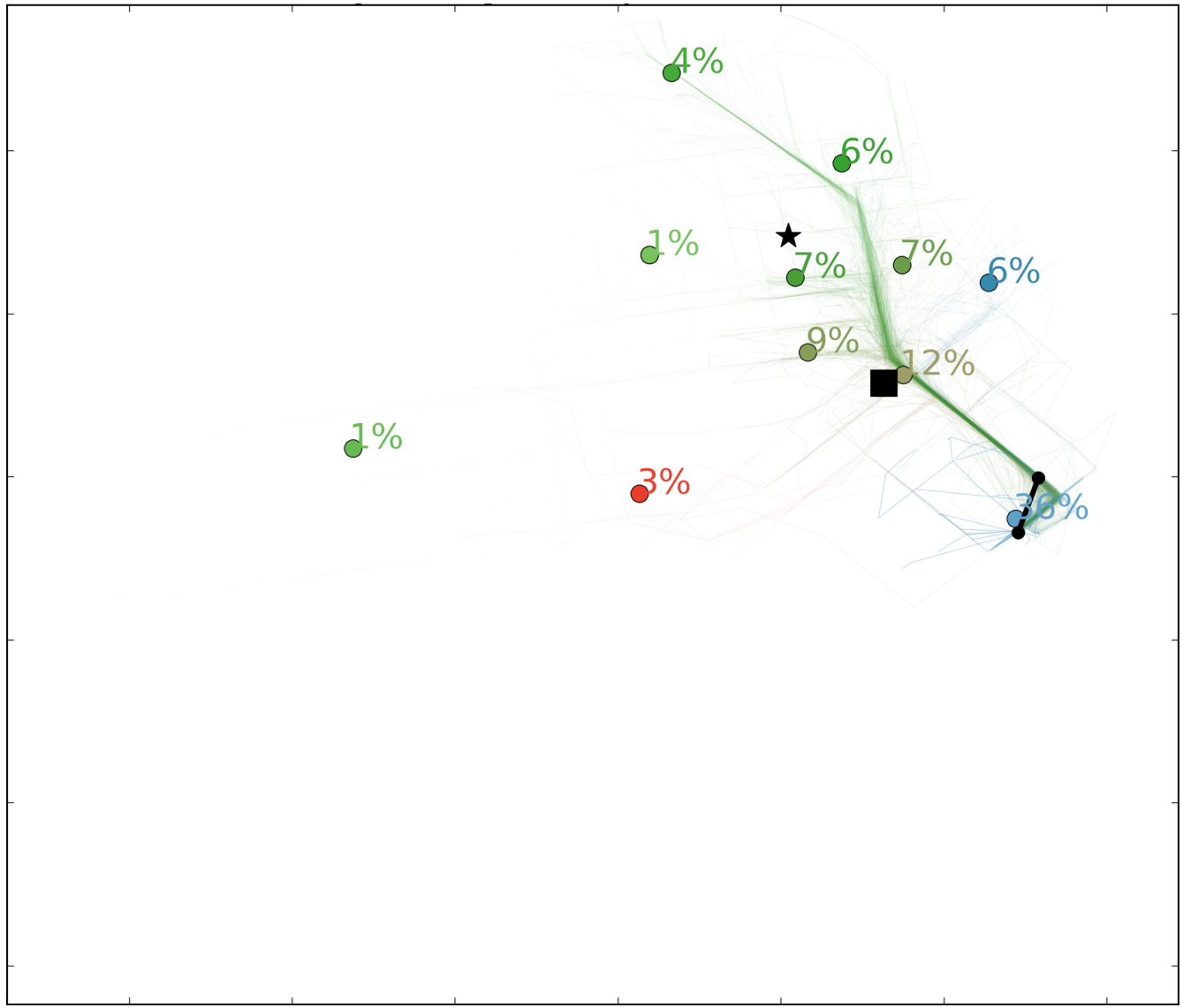


Clusters de trajets sur San Francisco

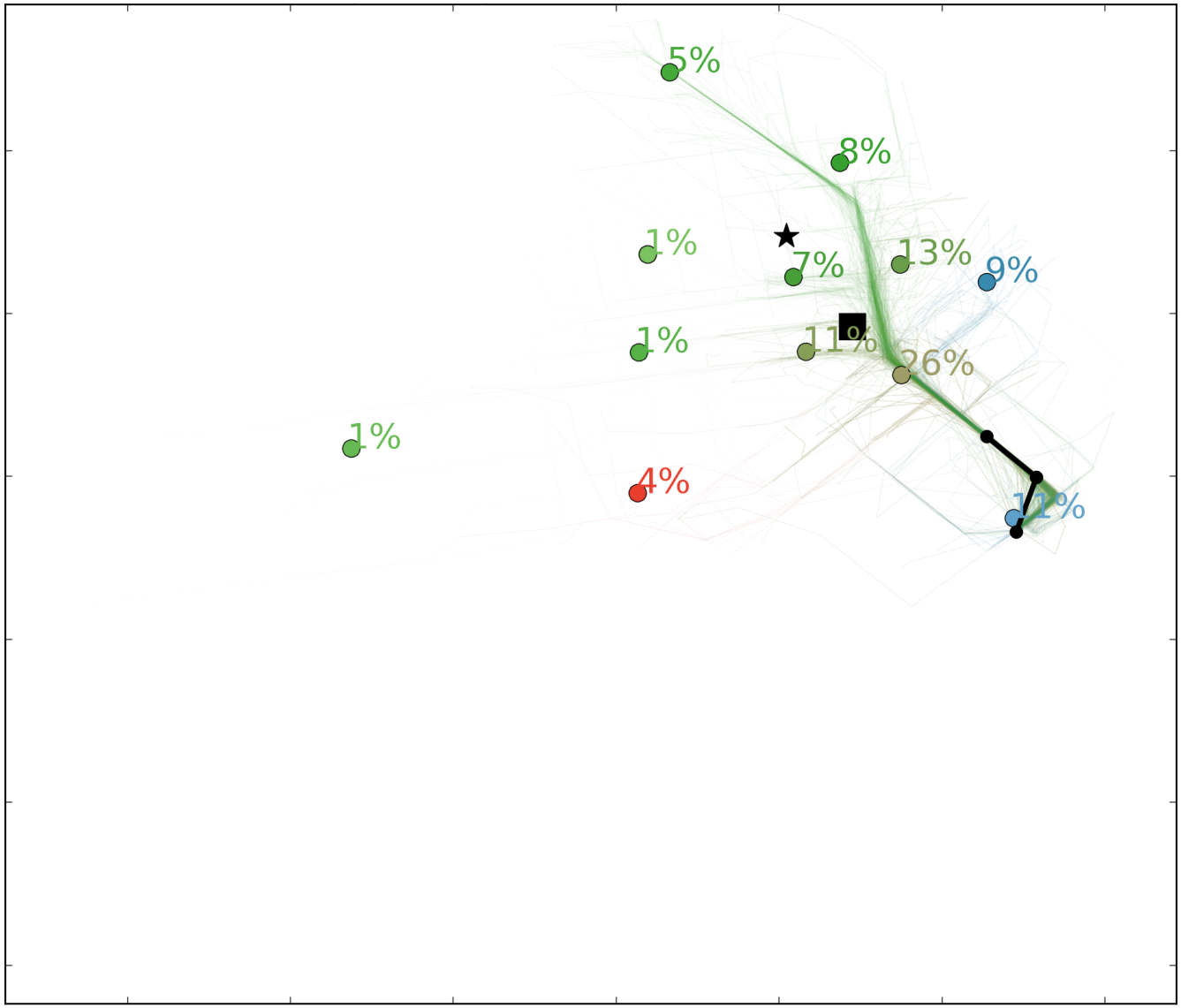
Années 2010s : Un exemple d'analyse du réseau routier



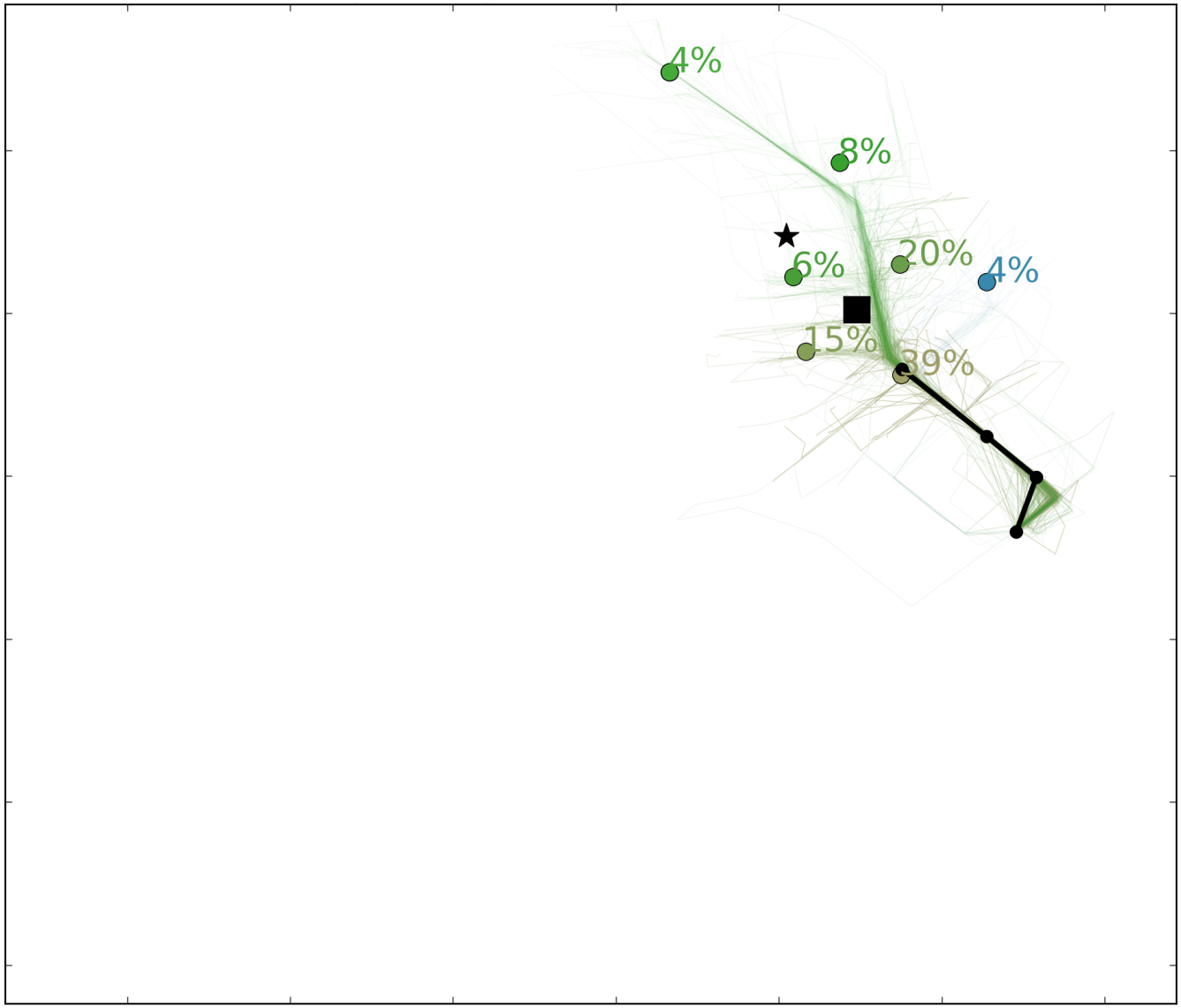
Années 2010s : Un exemple d'analyse du réseau routier



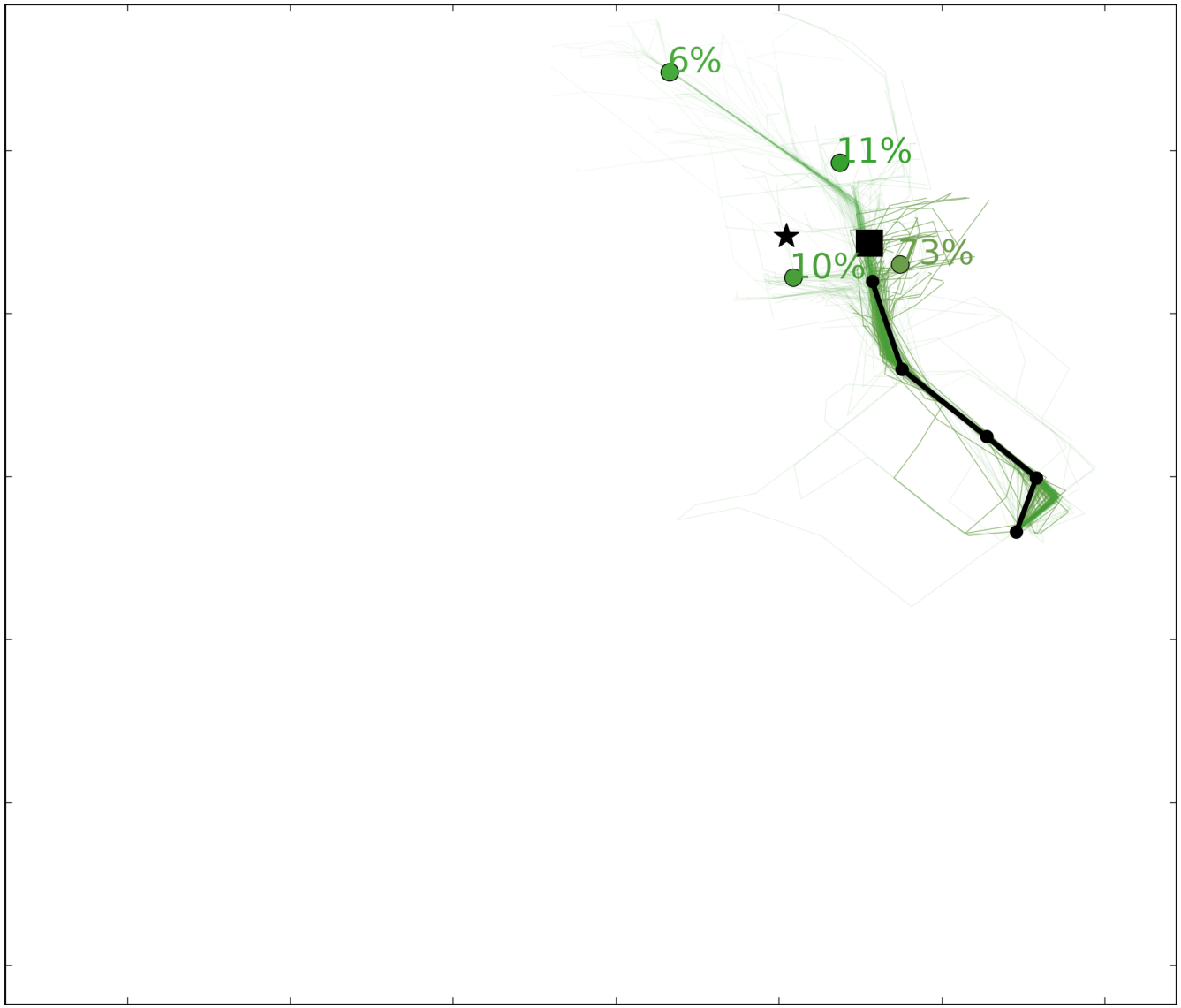
Années 2010s : Un exemple d'analyse du réseau routier



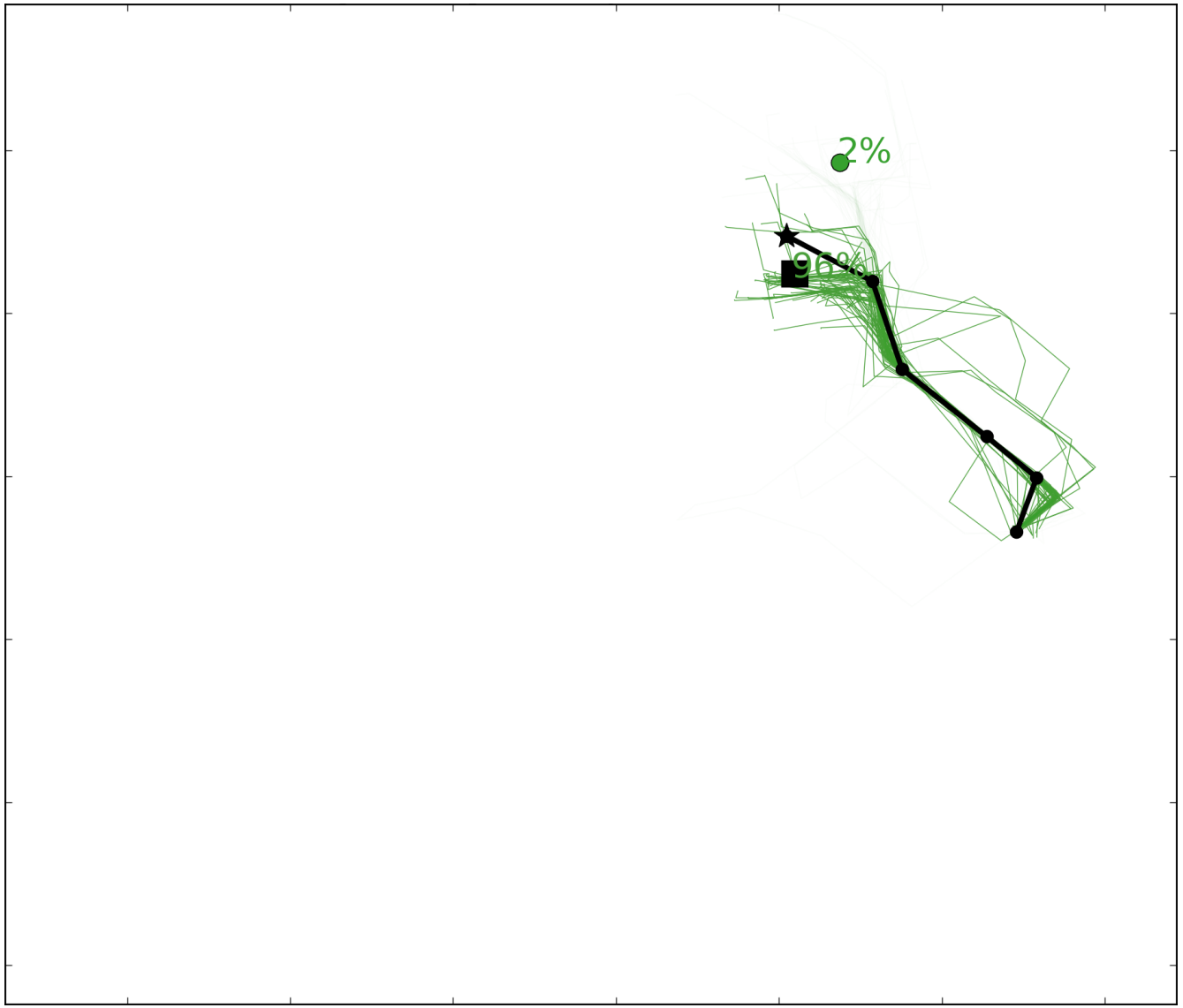
Années 2010s : Un exemple d'analyse du réseau routier



Années 2010s : Un exemple d'analyse du réseau routier



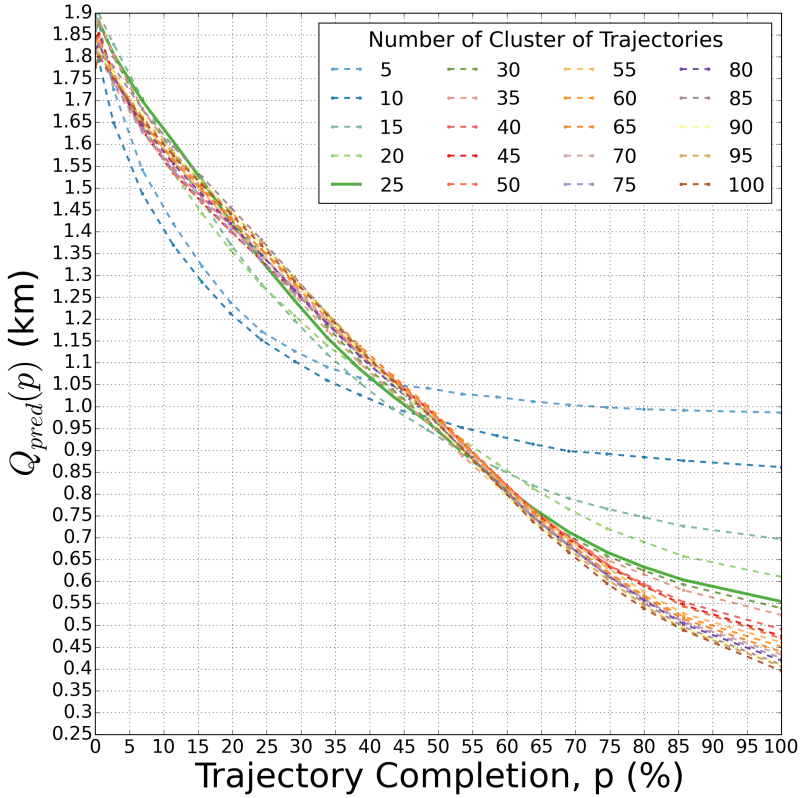
Années 2010s : Un exemple d'analyse du réseau routier



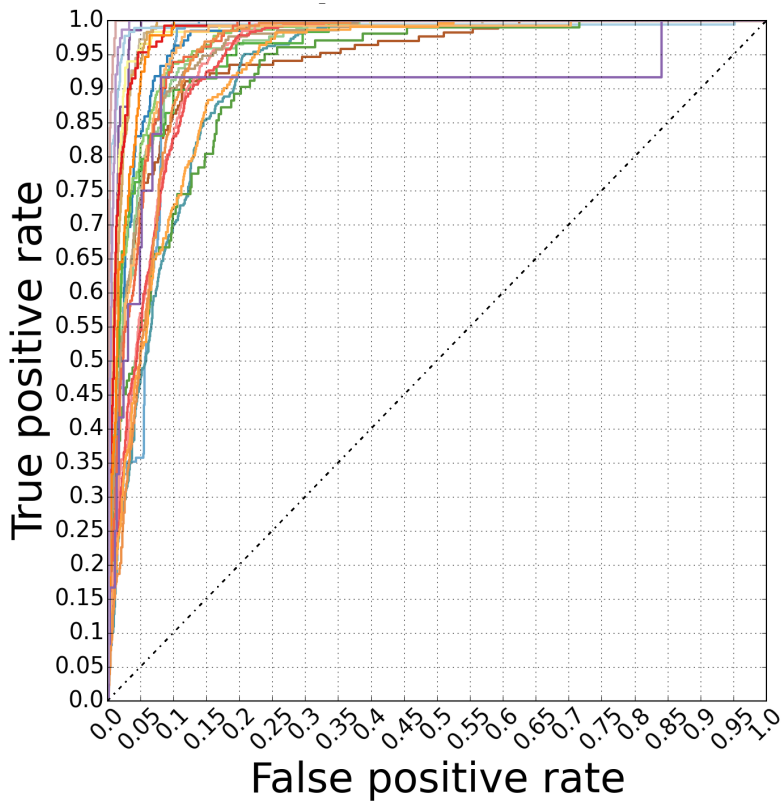
Trajet effectué à 100%

Evaluation de la méthode :

- Précision de la prédiction
- Cross validation (leave 10% out)
- Robustesse en fonction des paramètres
- Robustesse en fonction du type de réseau routier
- Respect de la contrainte de temps de prédiction



Courbes de précision



Courbes ROC

Années 2010s : Punchline

- Nouvelles contraintes temporelles
- Parallélisation (données réparties sur plusieurs machines)
- … ou bien méthodes de passage à l'échelle
- Retour vers des méthodes simples

Conclusion

- **Emergence du métier de *data scientist*.**
- **Ce sont les données qui décident !!!**

hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century