



Quinze ans de recherche appliquée en science des données

Sébastien Déjean

Institut de Mathématiques de Toulouse
Université Toulouse III – Paul Sabatier



Quelques chiffres

■ Production scientifique

- 44 publications internationales, 5 publications nationales, 3 *proceedings*, 1 chapitre de livre (1 en révision)

- 4 conférences invitées depuis 2014

- 4 séminaires depuis 2015

■ Encadrement

- 2 thèses en co-direction

- 2 post-doc, 1 ingénieur

- 2 à 3 stagiaires L3-M1-M2 par an

- Participation à 8 comités de thèse

Pourquoi ce titre ?

■ J. Tukey (1962)

The future of data analysis

The Annals of Mathematical Statistics

THE FUTURE OF DATA ANALYSIS¹

BY JOHN W. TUKEY

Princeton University and Bell Telephone Laboratories

Received July 1, 1961.

¹ Prepared in part in connection with research sponsored by the Army Research Office through Contract DA36-034-ORD-2297 with Princeton University. Reproduction in whole or part is permitted for any purpose of the United States Government.

■ D. Donoho (2017)

50 years of Data Science

Journal of Computational and Graphical Statistics

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS

2017, VOL. 26, NO. 4, 745–766

<https://doi.org/10.1080/10618600.2017.1384734>

50 Years of Data Science

David Donoho

Department of Statistics, Stanford University, Stanford, CA

Plan

I. Science des données

II. Collaborations interdisciplinaires

- Biologie
- Recherche d'Information

III. Le futur

- Domaine bio-médical
- Applications industrielles

Science des données

Des données à des connaissances

- *There is not yet a consensus on what precisely constitutes **Data Science** but:*
- *Data Science can be seen (defined ?) as a :*
 - *the study of the generalizable **extraction of knowledge from data**.*
 - *Data Science clearly has an **interdisciplinary** nature and requires substantial **collaborative effort**.*

F. Chamrouki, Statistical data science and some unsupervised learning problems, The Second International Symposium on Data Science and Computational Intelligence, 12/2018

Une science expérimentale

■ *Data analysis has, of necessity, to be an experimental science, and needs therefore to adopt the attitudes of **experimental science**.*

J. Tukey (1962), The future of Data Analysis

■ *Nous sommes en effet convaincus que [...] la statistique [...] doit s'honorer d'être **une science expérimentale**. [...] c'est à l'observation qu'on doit demander quel est l'ordre de la réalité : le mérite du calculateur étant de découvrir sans parti pris, sans a priori, quels courants de lois traversent l'océan des faits.*

J.-P. Benzécri (1973), L'Analyse des Données I

Une méthodologie de travail

- 1/ énoncer clairement une **question précise**
- 2/ **prévoir les méthodes** d'analyse des données
- 3/ établir un **plan d'expérience**
- 4/ **acquérir** les données
- 5/ **analyser** les données
- 6/ **interpréter** les résultats
- 7/ **répondre** à la question posée

■ Problem-Plan-Data-Analysis-Conclusion

R. J. MacKay, R. W. Oldford (2000) Scientific method, statistical method and the speed of light. *Statistical Science*.

■ Plan-Do-Check-Act(Adjust)

Shewhart, Deming 1930-1950

Our best conclusion is often a refined question.

De Veaux, College, Velleman (2008)

Le savant n'est pas l'homme qui fournit les vraies réponses; c'est celui qui pose les vraies questions.

C. Lévi-Strauss. Le Cru et le Cuit (1964)

Collaborations interdisciplinaires : Biologie

Un besoin de collaboration

S.R Searle (**1966**) Matrix Algebra for the biological sciences

*The biological sciences are **today** in the process of changing from being primarily descriptive **to being very much quantitative**. As a result, biologists find themselves **confronted more and more with large amounts of numerical data** [...]. But the mere collecting and recording of data achieve nothing; having been collected, they must be **investigated to see what information may be contained concerning the biological problem** at hand.[...]*

*Frequently, however, biologists have to subject their data to more complex calculations, requiring procedures that **involve mathematical details beyond their general experience**. In order to carry out the mathematics the biologist in this situation must either **learn the procedures himself**, or at least **learn something of the language of mathematics**, that he may **communicate satisfactorily with the mathematician** whose aid he enlists.*

► La biologie entre dans l'ère du Big Data des années 2000-2010 avec les technologies à haut-débit (génomique, transcriptomique...)

Deux contacts déterminants

Nathalie Viguerie



Ateliers de formation
N°145

Inserm

Institut national
de la santé et de la recherche médicale

**Analyse statistique
des Biopuces**

**Statistical analysis of
microarrays**

26-27 mai 2003

May 26-27, 2003

La Londe Les Maures

Pascal Martin



Deux contacts déterminants



- V. Barquissau et al. (2018). Caloric Restriction and Diet-Induced Weight Loss Do Not Induce Browning of Human Subcutaneous Abdominal White Adipose Tissue in Obese Women and Men. *Cell Reports*.
- E. Montastier et al. (2014). Adipose Tissue CIDEA Is Associated, Independently of Weight Variation, to Change in Insulin Resistance during a Longitudinal Weight Control Dietary Program in Obese Individuals. *PLoS ONE*.
- F. Capel et al. (2008). Contribution of energy restriction and macronutrient composition to changes in adipose tissue gene expression during dietary weight-loss programs in obese women. *The Journal of Clinical Endocrinology and Metabolism*.
- D. Larrouy et al. (2008). Gene expression profiling of human skeletal muscle in response to stabilized weight loss. *The American Journal of Clinical Nutrition*.
- A. Mazzucotelli et al. (2007). The Transcriptional Coactivator Peroxisome Proliferator-Activated Receptor (PPAR) γ Coactivator-1 α and the Nuclear Receptor PPAR α Control the Expression of Glycerol Kinase and Metabolism Genes Independently of PPARG Activation in Human White Adipocytes. *Diabetes*.
- I. González et al. (2013). Selecting the good level of details in undecimated wavelet transform improves the classification of samples from metabolomic data. *JP Journal of Biostatistics*.
- I. González et al. (2009). Highlighting Relationships Between Heterogeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis. *Journal of Biological Systems*.
- I. González et al. (2008). CCA: An R Package to Extend Canonical Correlation Analysis. *Journal of Statistical Software*.
- S. Déjean et al. (2007). Clustering time series gene expression data using smoothing spline derivatives. *EURASIP Journal on Bioinformatics and Systems Biology*.
- P. Martin et al. (2007). Novel aspects of PPAR α -mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. *Hepatology*.
- A. Baccini et al. (2005). Stratégies pour l'analyse statistique de données transcriptomiques. *Journal de la Société Française De Statistique*.

Interface entre statistique et biologie

La méthodologie de travail à l'épreuve du réel

| 1/ énoncer clairement une *question précise*

« J'ai plusieurs jeux de données acquis sur les mêmes échantillons biologiques et je veux les intégrer. »

| 3/ établir un plan d'expérience

Interface entre statistique et biologie

La méthodologie de travail à l'épreuve du réel

2/ *prévoir les méthodes d'analyse des données*

■ M. Tenenhaus et M. Hanafi (2010). A Bridge Between PLS Path Modeling and Multi-Block Data Analysis. in *Handbook of Partial Least Squares*.

Method	Criterion		
(1) SUMCOR (Horst 1961)	$Max \sum_{j,k} Cor(F_j, F_k)$ or $Max \sum_j Cor(F_j, \sum_k F_k)$	(11) (Hanafi and Kiers 2006)	$Max_{all} \ w_j\ =1 \sum_{j \neq k} Cov(X_j w_j, X_k w_k) $
(2) MAXVAR (Horst 1961) or GCCA (Carroll 1968)	$Max \{\lambda_{first}[Cor(F_j, F_k)]\}$ (a) or $Max \sum_j Cor^2(F_j, F_{j+1})$	(12) ACOM (Chessel and Hanafi 1996) or Split PCA (Lohmöller 1989)	$Max_{all} \ w_j\ =1 \sum_j Cov^2(X_j w_j, X_{j+1} w_{j+1})$ or $Min_{F,p,j} \sum_j \ X_j - F p_j^T\ ^2$
(3) SsqCor (Kettenring 1971)	$Max \sum_{j,k} Cor^2(F_j, F_k)$	(13) CCSWA (Hanafi et al., 2006) or HPCA (Wold et al., 1996)	$Max_{all} \ w_j\ =1, Var(F)=1 \sum_j Cov^4(X_j w_j, F)$ or $Min_{\ F\ =1} \sum_j \ X_j X_j^T - \lambda_j F F^T\ ^2$
(4) GenVar (Kettenring 1971)	$Min \{\det[Cor(F_j, F_k)]\}$	(14) Generalized PCA (Casin 2001)	$Max \sum_j R^2(F, X_j) \sum_h Cor^2(x_{jh}, \hat{F}_j)$ (c)
(5) MINVAR (Kettenring 1971)	$Min \{\lambda_{last}[Cor(F_j, F_k)]\}$ (b)		$Min_{F,p,j} \sum_j \left\ \frac{1}{\sqrt{\lambda_{first}[Cor(x_{jh}, x_{jl})]}} X_j - F p_j^T \right\ ^2$
(6) Lafosse (1989)	$Max \sum_j Cor^2(F_j, \sum_k F_k)$	(15) MFA (Escofier and Pagès 1994)	
(7) Mathes (1993) or Hanafi (2005)	$Max \sum_{j,k} Cor(F_j, F_k) $		
(8) MAXDIFF (Van de Geer, 1984 & Ten Berge, 1988)	$Max_{all} \ w_j\ =1 \sum_{j \neq k} Cov(X_j w_j, X_k w_k)$	(16) Oblique maximum variance method	$Min_{F,p,j} \sum_j \left\ X_j \left(\frac{1}{n} X_j^T X_j \right)^{-1/2} - F p_j^T \right\ ^2$
(9) MAXBET (Van de Geer, 1984 & Ten Berge, 1988)	$Max_{all} \ w_j\ =1 \sum_{j,k} Cov(X_j w_j, X_k w_k)$		
(10) MAXDIFF B (Hanafi and Kiers 2006)	$Max_{all} \ w_j\ =1 \sum_{j \neq k} Cov^2(X_j w_j, X_k w_k)$		

■ A. Tenenhaus, M. Tenenhaus (2011) Regularized generalized canonical correlation analysis. *Psychometrika*

■ A. Tenenhaus et al. (2014) Variable selection for generalized canonical correlation analysis. *Biostatistics*

■ A. Singh et al. (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*

► Faire preuve de pédagogie pour assurer la communication

Interface entre statistique et biologie

La simulation de données comme passerelle

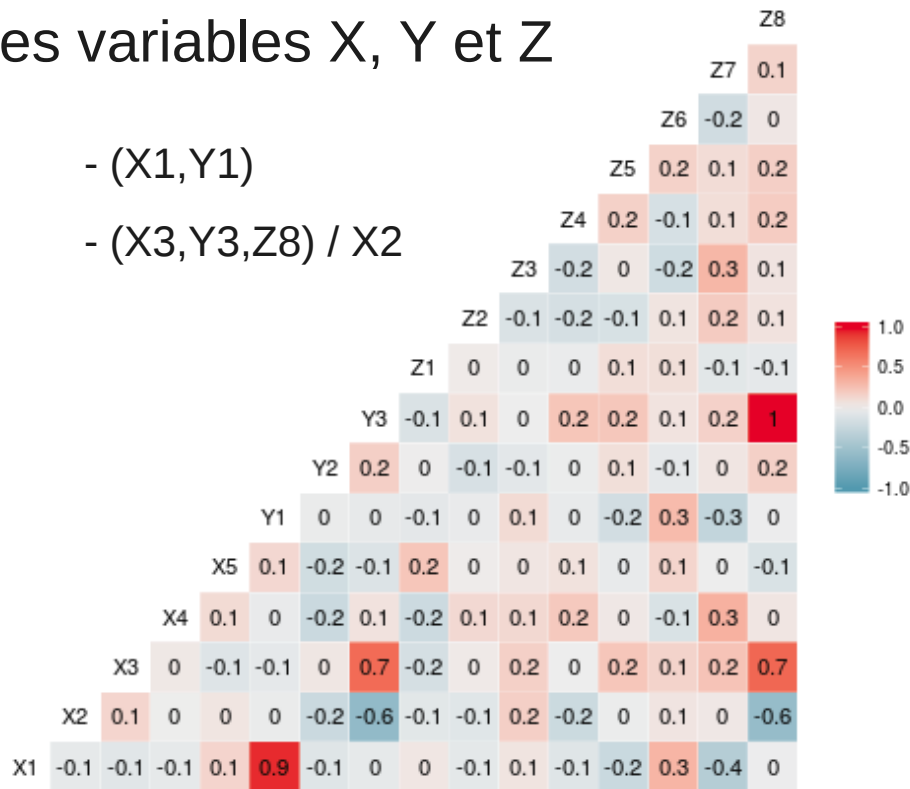
- 50 individus (e.g. échantillons biologiques)
- 3 jeux de données quantitatifs :
 - X : 5 variables (e.g. gènes)
 - Y : 3 variables (e.g. protéines)
 - Z : 8 variables (e.g. dosages, mesures...)
- 1 variable qualitative Groupe à 2 modalités A / B (e.g. contrôle/traité)

Interface entre statistique et biologie

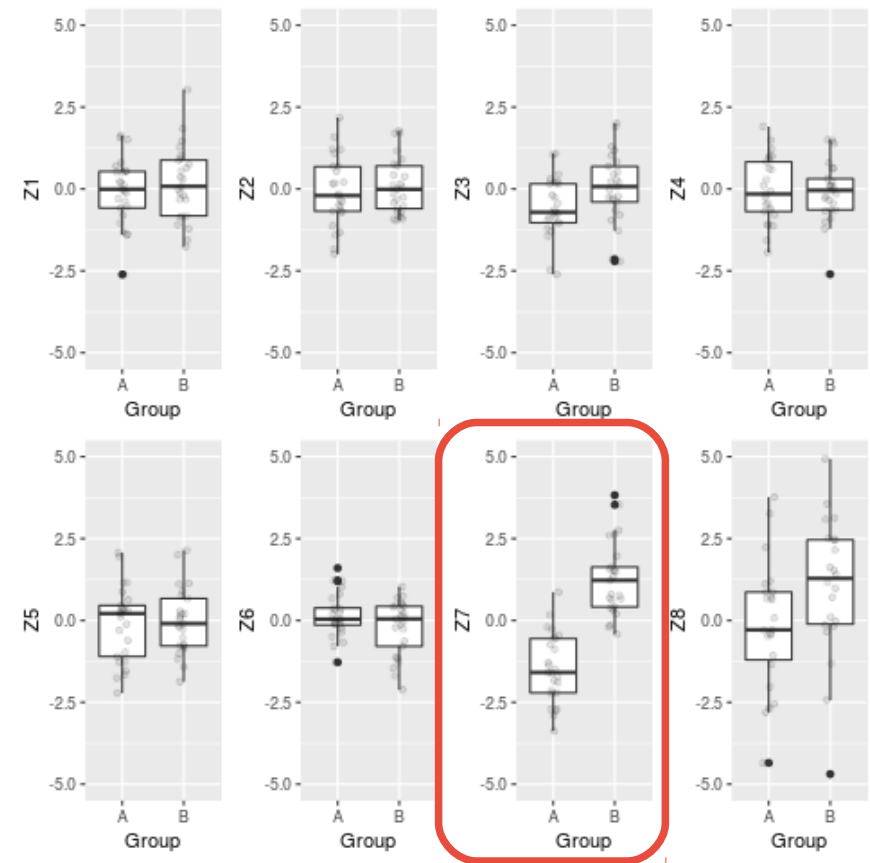
Matrice de corrélation entre les variables X, Y et Z

- (X1,Y1)

- (X3,Y3,Z8) / X2

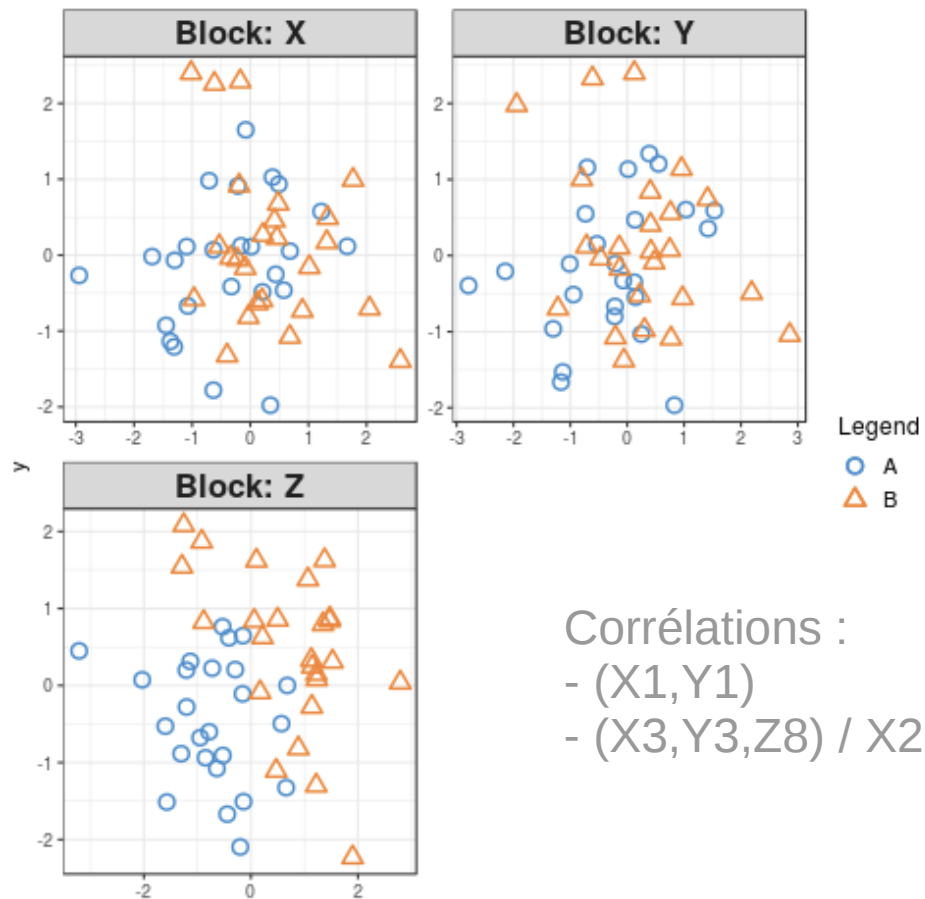


Boxplot par groupes des variables Z

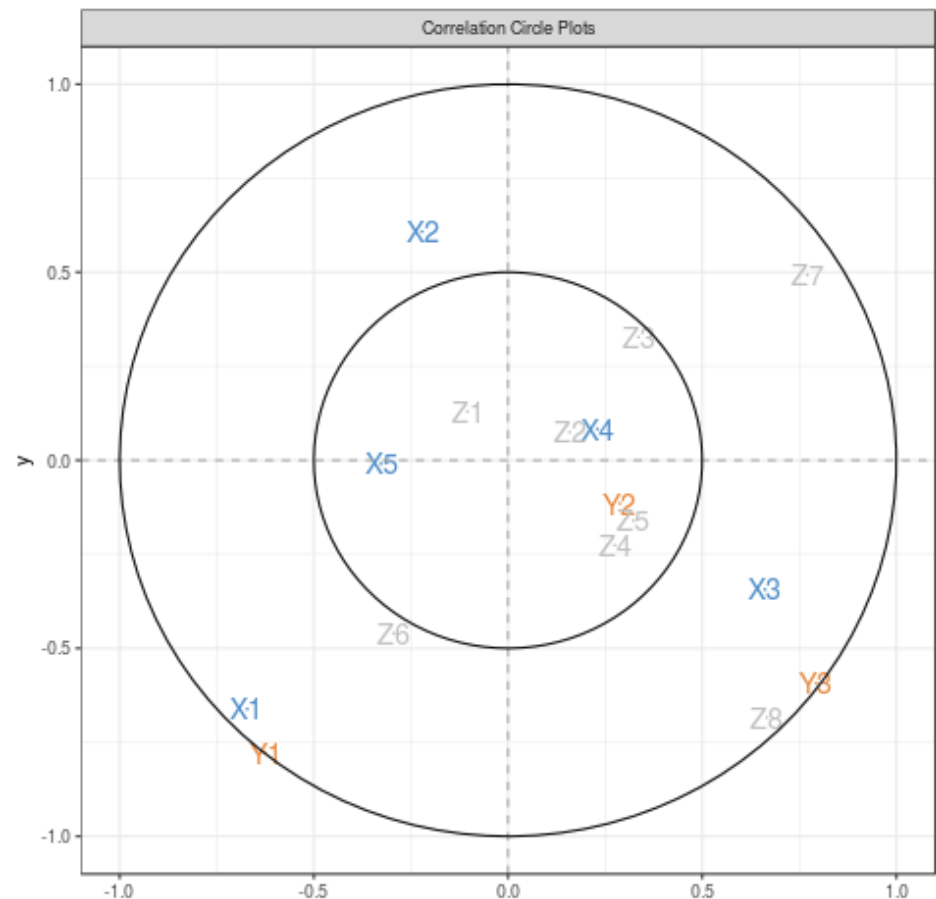


Interface entre statistique et biologie

Résultats d'une analyse PLS multi-blocs supervisée



Représentation des individus



Représentation des variables

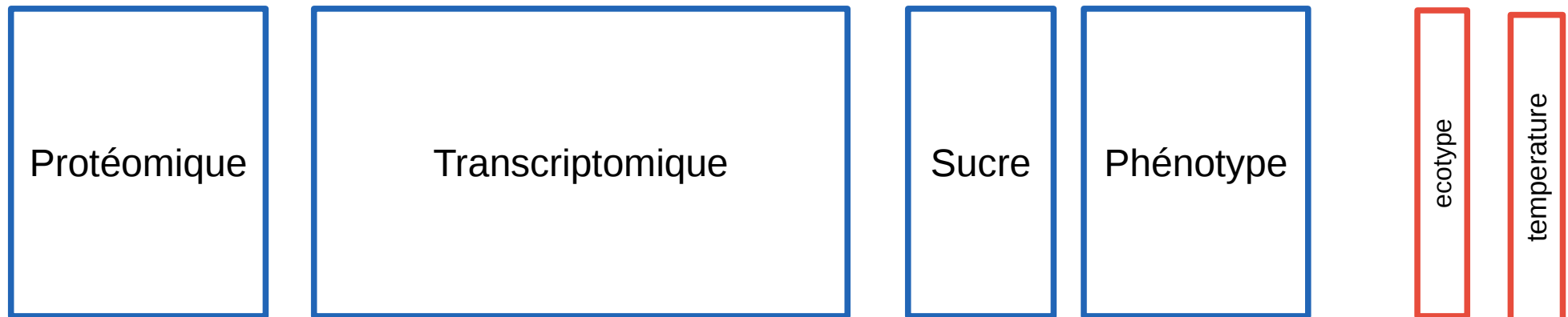
Le projet Wallomics

Harold Duruflé, Christophe Dunand...
Université Paul Sabatier

Adaptation des plantes au réchauffement climatique

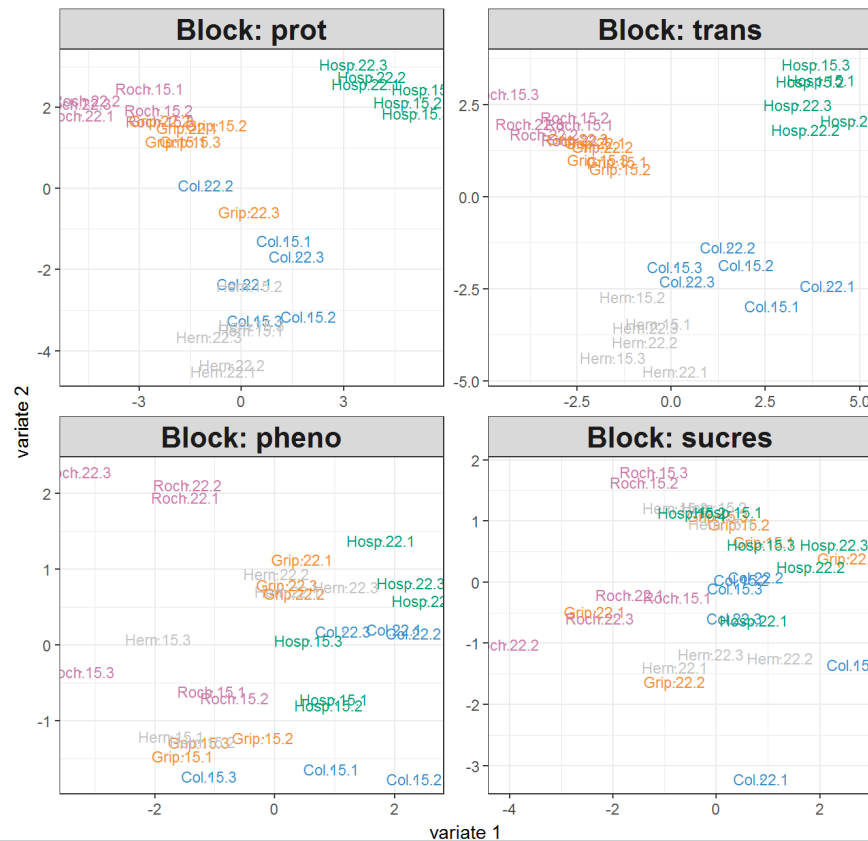
- **30** échantillons *A. thaliana*: 
 - **5** écotypes (Col, Hosp, Grip, Hern, Roch)
 - **2** niveaux de température (bas 15°C, élevé 22°C)
 - **3** réplicats
- **4** jeux de données : protéomique (400 protéines), transcriptomique (20000 gènes), métabolomique (7 sucres), phénotype (9 mesures)

| 4/ *acquérir* les données



Le projet Wallomics

Résultats d'une analyse multi-blocs supervisée par le facteur écotype avec sélection de variables



5/ *analyser* les données

- Stage de M. Selmani (INSA 4A GMM)
- H. Duruflé, M. Selmani,..., S. Déjean (2018). A powerful framework for an integrative study with heterogeneous omics data: from univariate statistics to multi-block analysis. *BioRxiv*

6/ *interpréter* les résultats
7/ *répondre* à la question posée

Travaux en cours...

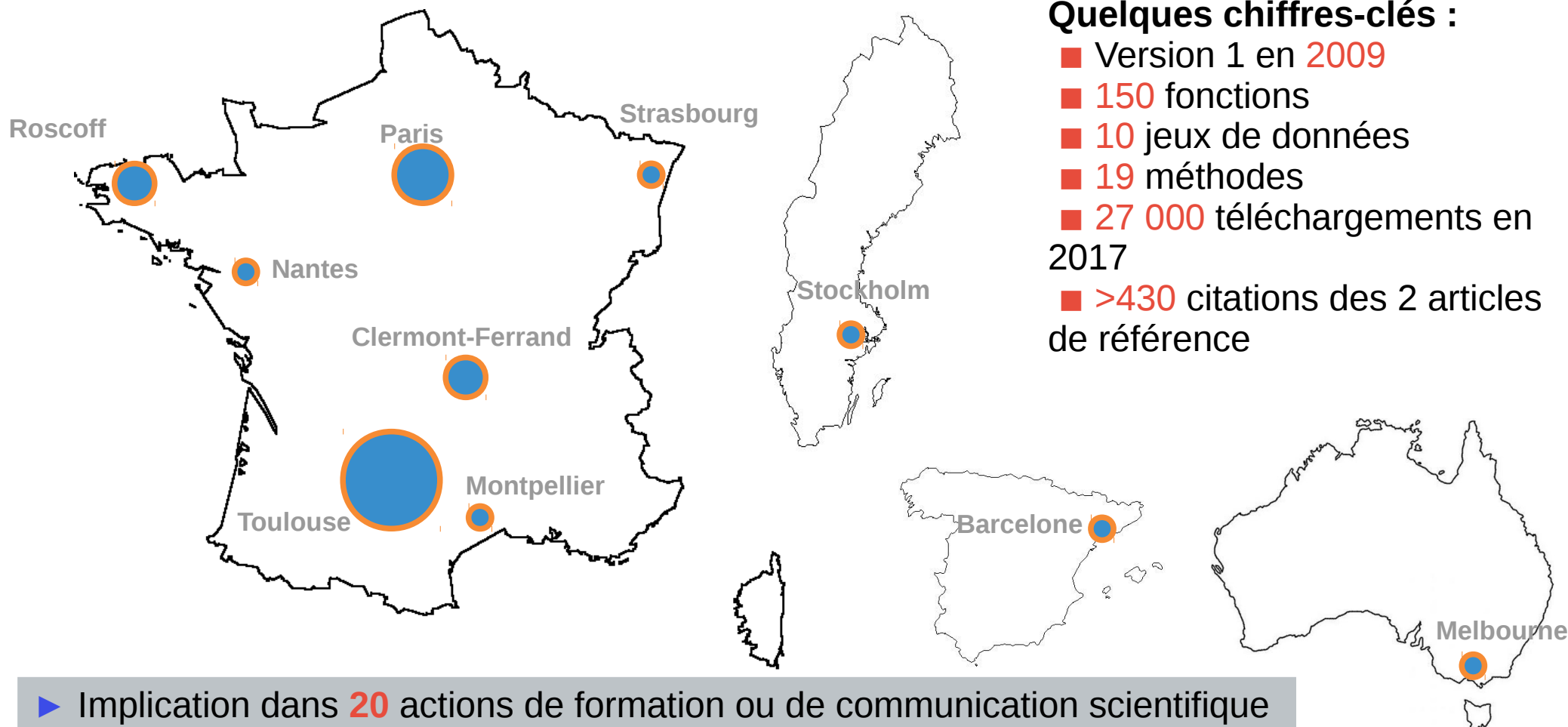
► Ouverture de nombreuses pistes de recherche en biologie

Diffusion d'un outil pour la communauté

Co-développeur du package mixOmics

Quelques chiffres-clés :

- Version 1 en 2009
- 150 fonctions
- 10 jeux de données
- 19 méthodes
- 27 000 téléchargements en 2017
- >430 citations des 2 articles de référence

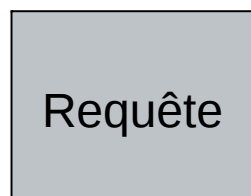


- ▶ Implication dans 20 actions de formation ou de communication scientifique
- ▶ Contribution à la formation d'environ 400 personnes

Collaborations interdisciplinaires : Recherche d'Information

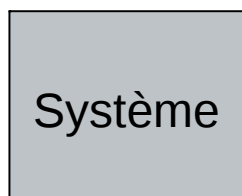
Contexte

Quoi ?



<title> Corporate Pension Plans/Funds
<desc> Document will report on problems associated with pension plans/funds such as fraud, skimming, tapping or raiding.
<narr> A relevant document will report on problems associated with pension plans/funds and also U.S. Government regulatory controls on pension plans. Examples of problems that are considered relevant are fraud, skimming, tapping or raiding.

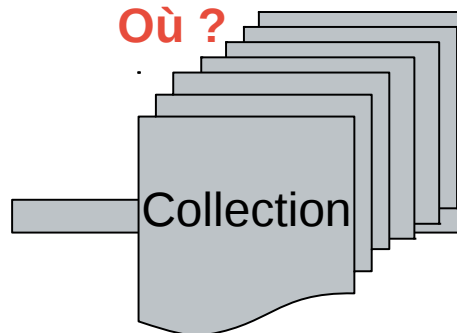
Comment ?



Quelle méthode pour :

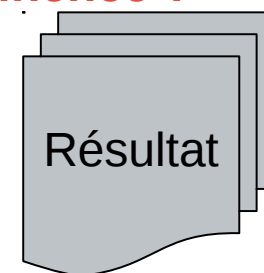
- l'indexation
- la recherche d'information
- l'extension, la reformulation de requêtes
- ...

Où ?



- Des millions de documents
- Des gigaOctets de fichiers informatiques

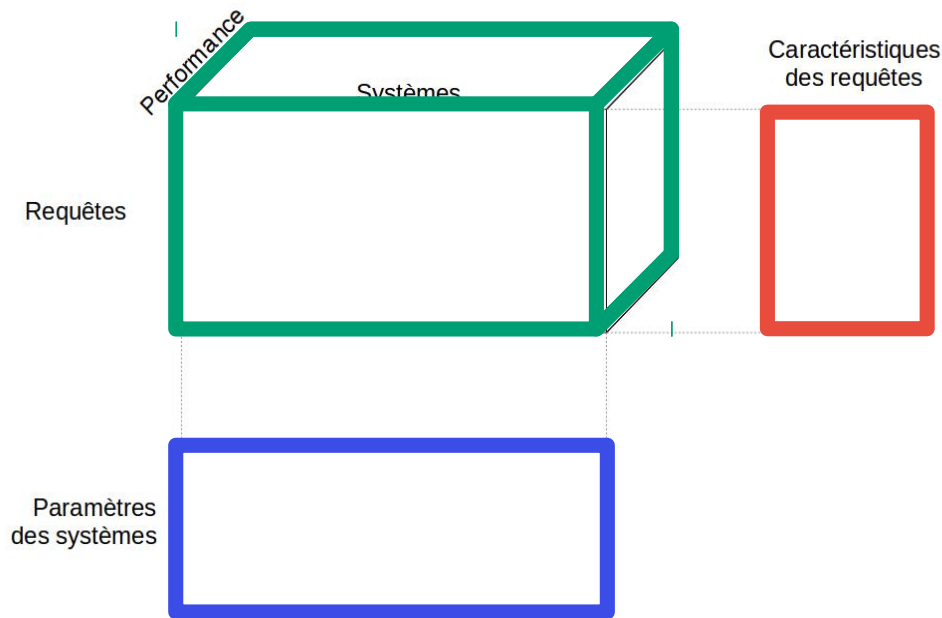
Pertinence ?



- Liste de documents pertinents connue
- nombreux (>100) indicateurs de performances basés sur Rappel et Précision

Quel système utiliser pour répondre au mieux à une requête ?

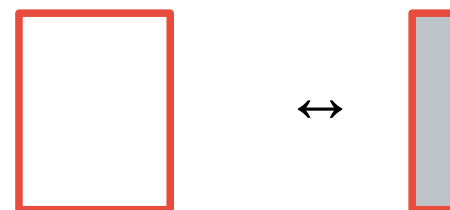
Un problème d'intégration de données



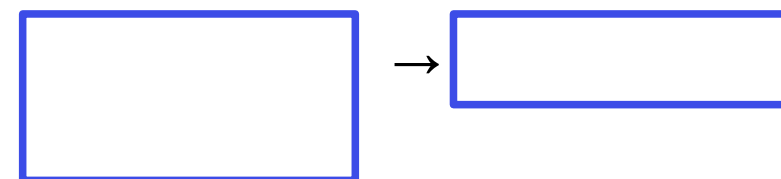
Exemple :

- Requête R : **20** verbes, **3** noms propres, indice de polysémie moyenne **2.5**...
- Système S : Reformulation de requête **Klbfree**, modèle RI **BM25**, Champs de la requête **Title + Descr + Narr**, Inverse Document Frequency **TRUE**...
- Performances pour une requête R et un système S : précision à 5 documents **0.75**, Précision à 10 documents **0.82**, Rappel à 5 documents **0.48**, Rappel à 10 documents **0.52**, *Mean Average Precision* **0.35**...

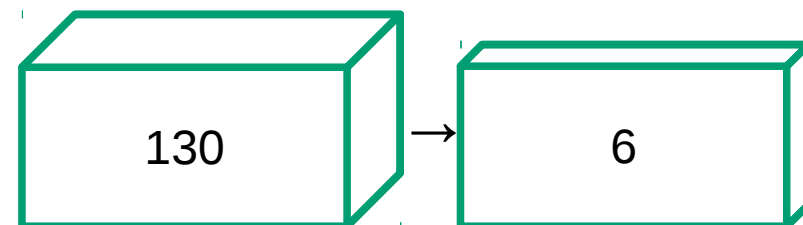
- A-G Chifu, S. Déjean, S. Mizzaro and J. Mothe (2017). Human-Based Query Difficulty Prediction. In: Jose J. et al. (eds) *Advances in Information Retrieval*. ECIR 2017. **Lecture Notes in Computer Science**, vol 10193. Springer, Cham
- J. Ayter, A. Chifu, S. Déjean, C. Desclaux, J. Mothe (2016). Statistical Analysis to Establish the Importance of Information Retrieval Parameters. **Journal of Universal Computer Science**
- J. Compaoré, A.M. Gueye, S. Déjean, J. Mothe, J. Randriamparany (2011). Mining Information Retrieval Results; Significant IR parameters. **International Conference on Advances in Information Mining and Management**
- A. Baccini, S. Déjean, L. Lafage, J. Mothe (2011). How many performance measures to evaluate Information Retrieval Systems? **Knowledge and Information System**
- A. Baccini, S. Déjean, D. Kompaoré, J. Mothe (2010). Analyse des critères de performance de systèmes de recherche d'information. **Technique et Science Informatiques**



Analyses de corrélation, questionnaire



Classification and Regression Trees, Random forests



Analyse exploratoire, classification non supervisée

Collaborations interdisciplinaires

Synthèse

- Une thématique commune : l'intégration de données
 - Passerelles entre biologie et RI
 - Quels sont les **gènes** impliqués dans la discrimination des **différentes conditions** de l'expérience ?
 - Quelles sont les **caractéristiques d'une requête** qui traduisent au mieux **la difficulté ressentie** par un humain ?
- Un positionnement dans le domaine mathématique permet de lier ce types de questions et de les traiter par des méthodes similaires

Domaine bio-médical

Une évolution déjà initiée

Big Data en santé : mythe ou réalité ?

Séminaire

Comité d'éthique CHU de Toulouse

Toulouse, le 26 juin 2017



Sébastien DEJEAN - Philippe SAINT-PIERRE - Nicolas SAVY

Institut de Mathématiques de Toulouse
Toulouse, France



Research funded by :

Project "Big Data" Grant RBIO-2015-14054319 - Région Occitanie

Project BIDASA Grant - Mastodons CNRS



■ Co-organisation de la conférence Statistique et Santé (janvier 2018)

■ Conférencier invité au Nordic Precision Medicine Forum (Stockholm, mars 2019)

Une évolution déjà initiée

■ Voice4PD-MSA : Diagnostic différentiel entre la maladie de Parkinson et l'atrophie multisystématisée par analyse numérique de la parole - *Differential diagnosis between Parkinson's disease and Multiple System Atrophy using digital speech analysis*

- Projet financé par l'Agence Nationale de la Recherche, Appel à projets Technologies pour la santé
- Partenaires : INRIA (GeoStat), Centres Hospitaliers Universitaires de Bordeaux et Toulouse, IRIT, IMT

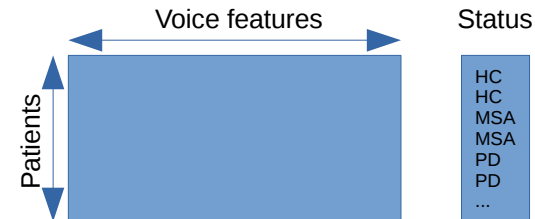
1/ Enregistrement



2/ Traitement du signal



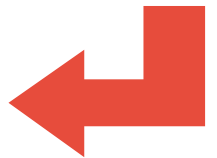
3/ Extraction de paramètres



4/ Classification



- Méthode multivariées supervisées (PLS-DA, SVM, Random Forest...)
- Sélection des paramètres les plus pertinents (méthodes sparse...)



Une évolution déjà initiée

- Contacts avec la *Fédération Régionale de Recherche en Psychiatrie et Santé Mentale*
 - Quelles sont les **facteurs** qui conduisent à la **contention** d'un patient en hôpital psychiatrique ?
 - Projet OptimIPS-TC : Optimiser le Parcours de Soins des patients Traumatisés Crâniens
 - Neurologie, psychiatrie, médecine physique et de réadaptation
 - Extraction, intégration, organisation, fouille de données
 - Analyses multivariées et visualisation de données complexes
- Difficulté accrue par la forte variabilité entre les individus et par l'hétérogénéité de la qualité des données

Le futur

Applications industrielles

Pistes pour y parvenir

- Membre de la cellule valorisation de l'IMT depuis sa création en 2017
- Actions en direction des entreprises
- Organisation d'événements scientifiques
 - Journée Math et industrie, 26 janvier 2018
 - Séminaire industriel (2 séances 11/2018, 03/2019)
 - Visites d'entreprises

▶ Échelles de temps radicalement différentes

Pistes pour y parvenir

- Le groupe « ingé-stat » : ingénieur.es statisticien.nes toulousain.es de tous horizons (recherche et développement ; public et privé)
- Environ 4 rencontres annuelles depuis 2012 : exposés techniques, étude de cas, échange de bonnes pratiques, retours d'expérience...
- Zoom sur la rencontre du 28 février 2019

Rencontre « ingé-stat » 28/02/2019

- Organisation par des collègues d'Airbus
- Présentations autour de la science des données
- La plateforme Skywise

www.airbus.com/aircraft/support-services/skywise.html

- **Skywise** - Unleashing the potential of data

■ *In aviation, data used to be locked in silos for maintenance, flight management, aircraft monitoring and safety, and other critical workflows. Airbus has discovered the **value of integrating disparate data sources**. [...] The aim is to make the data sharable, open and transparent to create value for our customers.*

Skywise / OptimiPS-TC

usinenouvelle.com/article/avec-skywise-airbus-lance-l-internet-des-avions.N618963

Imaginez une flotte d'avions qui transmettraient l'ensemble de leurs données de vol **sur une seule et même plate-forme numérique**, de la moindre vanne défectueuse à un décollage poussif, en passant par des vibrations anormales de la voilure. **Collectées, agrégées, retraitées, ces données formeraient un trésor.** Il deviendrait possible pour l'utilisation de chaque avion **de s'enrichir de l'exploitation des autres.** À la clé, **des millions d'euros d'économie.** Changer cette pièce juste avant qu'elle ne tombe en panne, définir la meilleure trajectoire de vol pour réduire la consommation de carburant... **Une révolution pour** l'aéronautique, **un changement de modèle économique pour** les industriels, **un scénario de rêve** pour les compagnies aériennes. **Et qui prend forme.**

Imaginez un ensemble de patients qui transmettraient l'ensemble de leurs données de santé **sur une seule et même plate-forme numérique**, de la moindre prise de tension artérielle jusqu'aux examens génétiques ou d'imagerie lors d'un diagnostic de cancer, en passant par les résultats du bilan lors d'une lombalgie commune. **Collectées, agrégées, retraitées, ces données formeraient un trésor.** Il deviendrait possible pour le soin de chaque patient **de s'enrichir de l'exploitation des autres.** À la clé, **des vies sauvées et des millions d'euros d'économie.** Mettre en place des mesures de prévention et de dépistages personnalisés, définir le meilleur parcours de soin pour réduire l'impact des maladies et le handicap qui en découle... **Une révolution pour** la santé, **un changement de modèle économique pour** les hôpitaux, **un scénario de rêve** pour les patients et les cliniciens. [...]

Adaptation réalisée avec l'aide de Pr X. de Boissezon (CHU Toulouse)

Conclusion

Conclusion

*The future of data analysis can involve great process, the overcoming of real difficulties, and the provision of **a great service to all fields of science and technology**. Will it? That remains to us, to our willingness to take up the **rocky road of real problems** [...]. Who is for the challenge?*

J. Tukey