Visualisation exploratoire de données historiques

Jean-Daniel Fekete
Directeur de Recherche INRIA, Orsay
Resp. de l'équipe AVIZ (www.aviz.fr)
Jean-Daniel.Fekete@inria.fr
www.lri.fr/~fekete

Nicole Dufournaud
CESR, Tours et EHESS, Paris

Nicole.Dufournaud@laposte.net
nicole.dufournaud.net

Analyse exploratoire de données

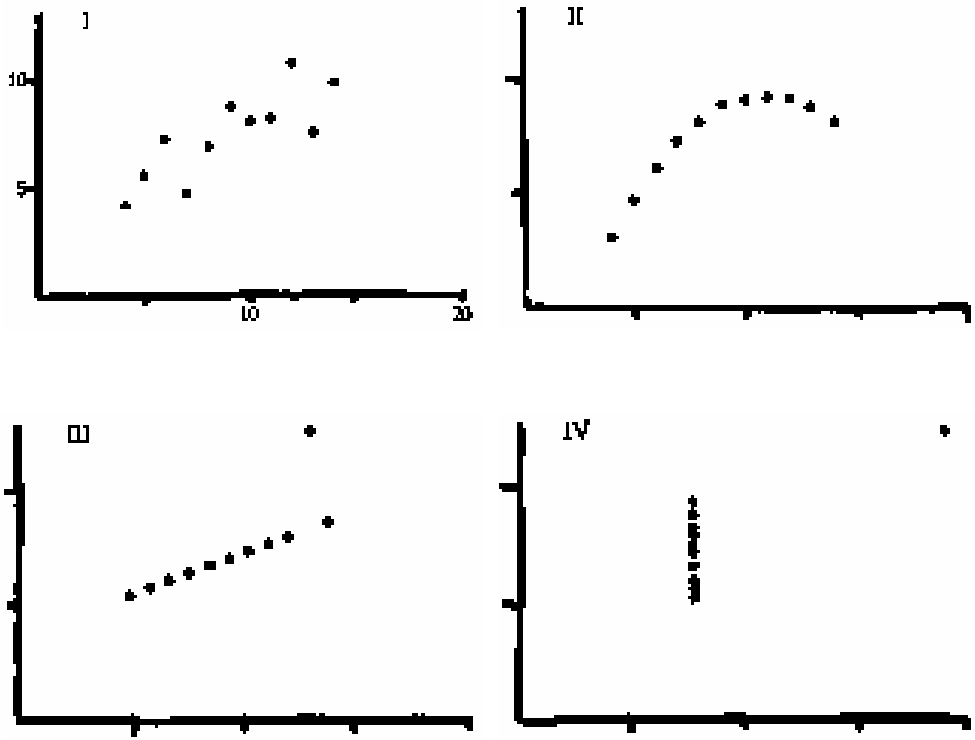
- Trois écoles de statistiques et trois paradigmes :
  - Classique (confirmatoire)
    - Problème => Données => Modèle => Analyse => Conclusions
  - Exploratoire [Tukey, John (1977), *Exploratory Data Analysis*, Addison-Wesley]
    - Problème => Données => Analyse => Modèle => Conclusions
  - Bayésienne [T. Bayes (1763), « An Essay towards solving a Problem in the Doctrine of Chances », *Philosophical Transactions of the Royal Society of London*, 53.]
    - Problème => Données => Modèles => Distribution à priori => Analyse => Conclusions

Analyse exploratoire de données et visualisation

- Les statistiques classiques s'appliquent à un problème quand :
  - On a des données en quantité suffisante
  - On a un modèle
- Sinon, les statistiques exploratoires permettent :
  - D'explorer plusieurs modèles rapidement
  - De trouver des motifs intéressant/inattendus dans les données

Statistiques et Visualisation : le Quartet d'Anscombe

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



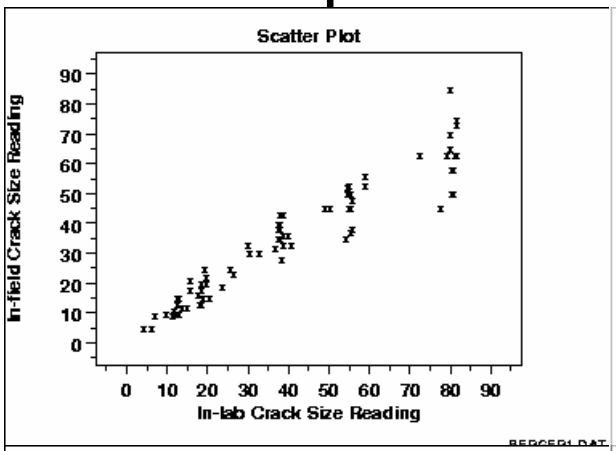
N:	11.0
mean X's :	9.0
mean Y's :	7.5
standard error of slope estimate:	0.1
sum of squares:	110.0
regression sum of squares:	27.5
residual sum of squares of Y:	13.8
correlation coefficient:	0.8
r squared:	0.7
regression line:	Y=3+0.5X

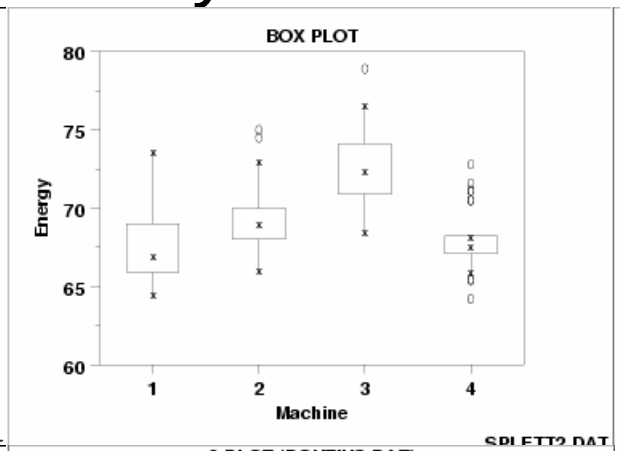
http://astro.swarthmore.edu/astro121/anscombe.html

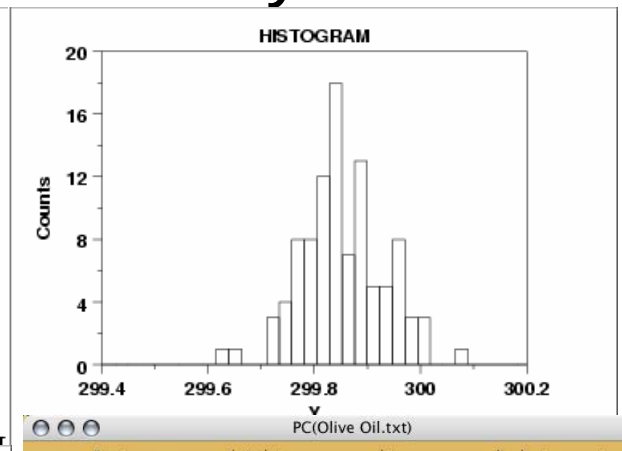
F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 [February 1973], 17-21

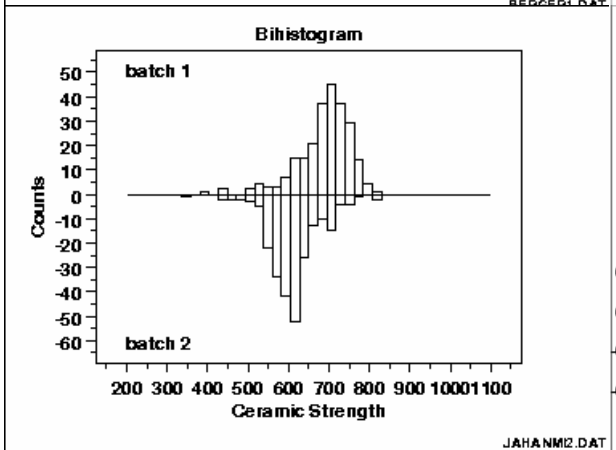
AED et Visualisation d'information

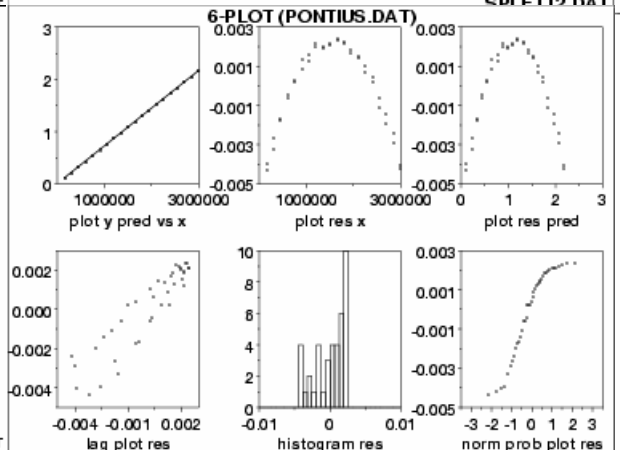
- AED: représentation simples et efficaces
  - http://www.math.yorku.ca/SCS/Gallery/

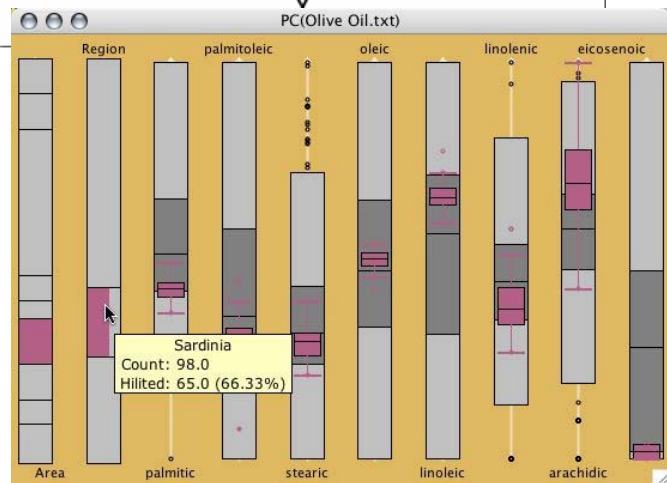








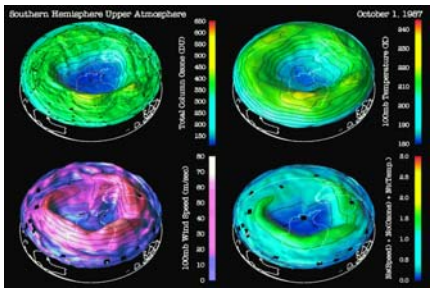




Visualisation d'information: représentations visuelles ET interaction

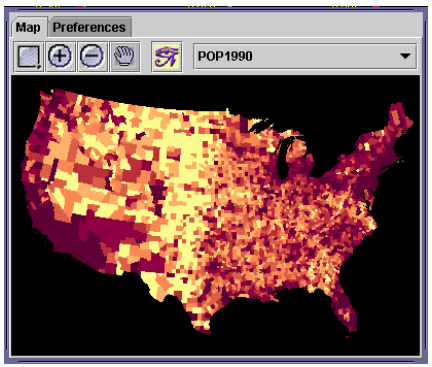
- Rétroagir sur la représentation graphique pour la changer:
  - Explorer / naviguer
  - Filtrer / Obtenir des détails
- Augmenter la vitesse change la nature de l'interaction
  - Boucle de rétroaction (Wiener 48)
  - Action/réaction dans un délais bref
  - Utilisation de la mémoire à court terme

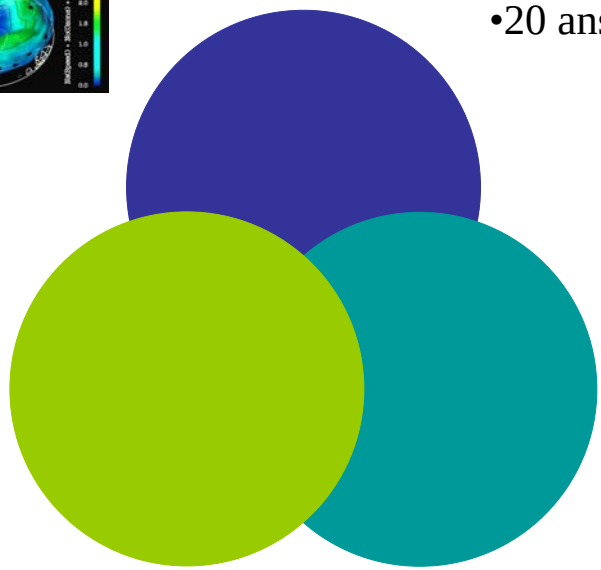
Visualisation : 3 domaines

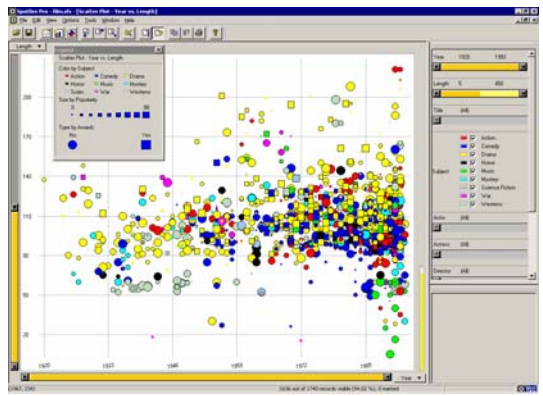


Visualisation scientifique

- Sous communauté de l'Informatique Graphique
- 20 ans d'histoire







Cartographie + Statistiques

- Communauté à part entière
- 2000/200 ans d'histoire

Visualisation d'information

- Sous communauté de l'Interaction Homme-Machine
- 10 ans d'histoire

Visualisation : 3 disciplines

Informatique

- Algorithmes et structures de données
- Géométrie
- Infographie

Design

- Nouvelles représentation
- Communication

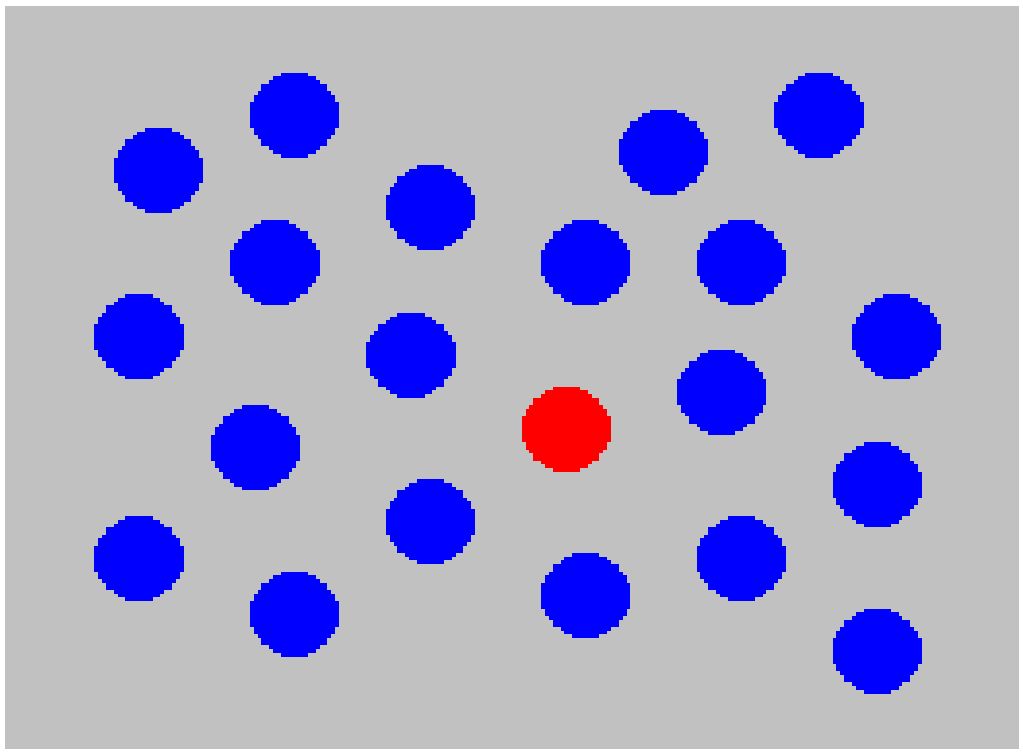
Psychologie

- Perception
- Cognition
- Interaction

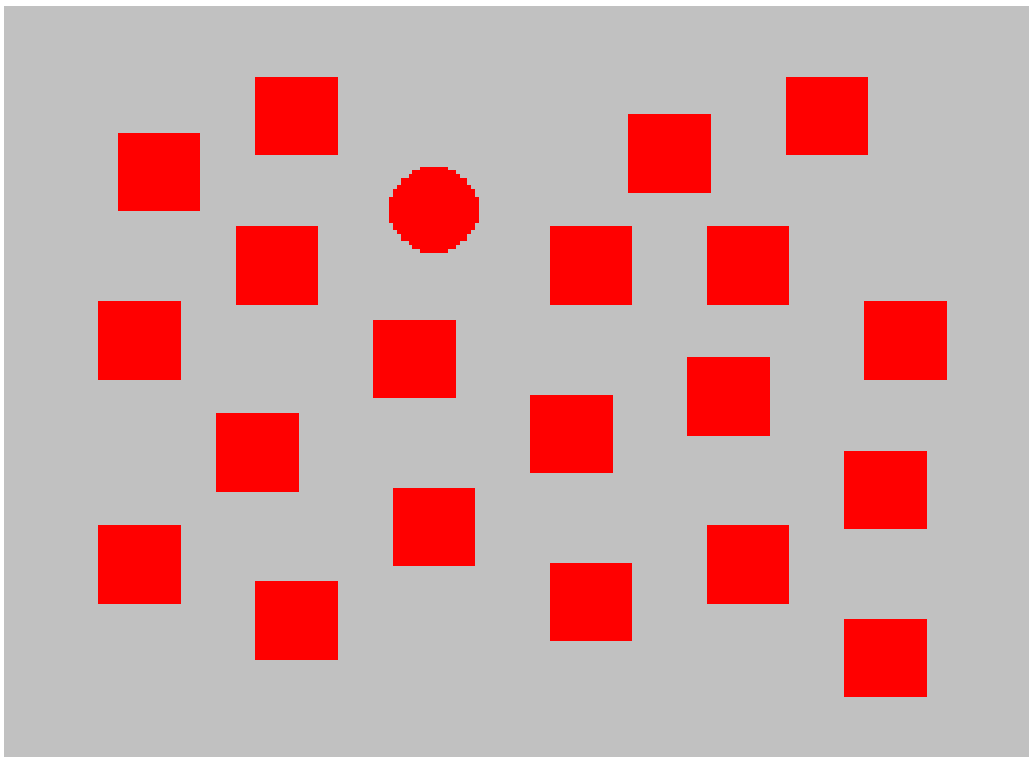
Principes de la visualisation d'information

- L'œil et la perception humaine sont remarquablement adaptés à la reconnaissance de motifs visuels
- La transformation de données abstraites en information visuelle permet d'utiliser cette aptitude
- Parmi toutes les représentations possibles, seules quelques-unes « fonctionnent » :
  - il faut les trouver et les répertorier
- La psychologie nous donne une base d'explication : la perception préattentive (Triesman, 85)
  - Sans effort
  - D'un coup d'œil
  - En temps constant
- Êtes-vous préattentifs ?

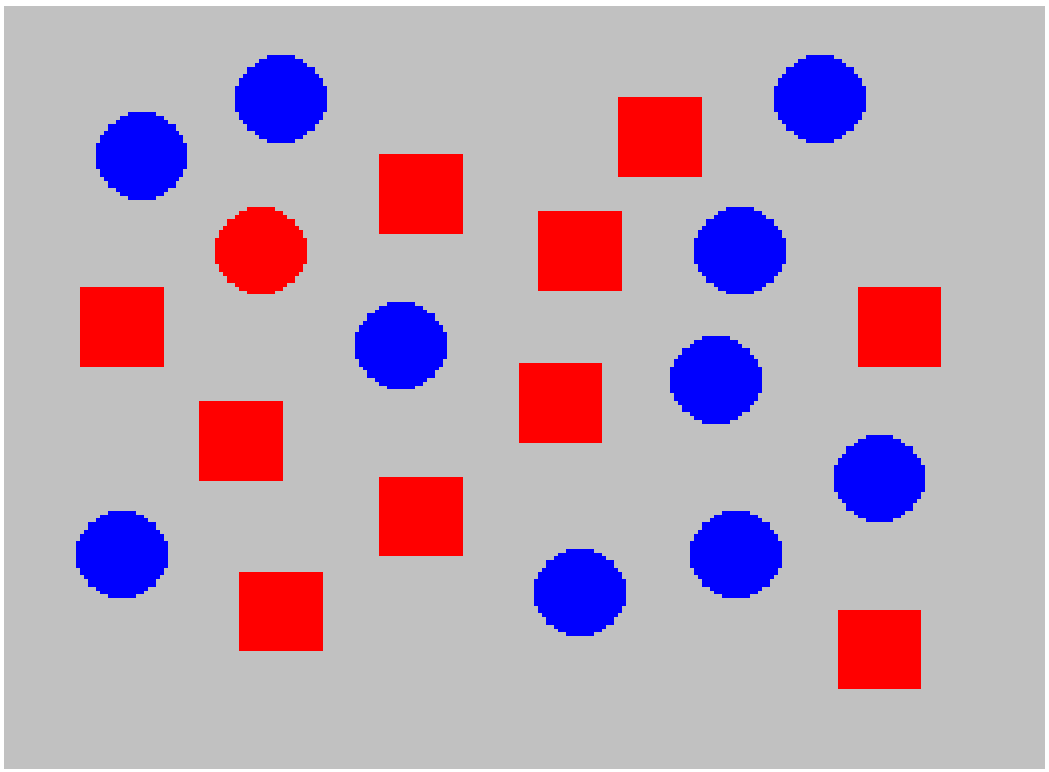
Perception préattentive (1)



Perception préattentive (2)



Perception préattentive (3)



Conclusion sur la visualisation

- Utilisez les outils disponibles pour explorer vos données :
  - Mondrian pour les données tabulées http://rosuda.org/Mondrian/
  - Guess pour les réseaux sociaux http://graphexploration.cond.org/
  - Many Eyes pour travailler collectivement
    - http://www.many-eyes.com
  - Improvise pour lier les représentations visuelles (cartes et données tabulées par exemple)
    - http://www.personal.psu.edu/cew15/improvise/

Bibliographie

- André Jacques et Chabin Marie-Anne (coord.), « Les documents anciens », *Document numérique*, Paris, Hermes, volume 3, n° 1-2, juin 1999.
- Dufournaud Nicole et Fekete Jean-Daniel, « XML/TEI pour le dépouillement des sources historiques. Leçons tirées d'une thèse en histoire », *Document numérique « Visualisation pour les bibliothèques numériques »*, Hermès-Lavoisier, pp. 37–56, Volume 9/2, 2006.
- Dufournaud Nicole et Fekete Jean-Daniel, « Compus Visualization and Analysis of Structured Documents for Understandig Social Life in the 16th Century », *Actes du colloque international Digital Libraries*, San Antonio, ACM, 2000, pp. 47-55.

- Readings in Information Visualization, Card, Mackinlay, Shneiderman, Morgan Kaufmann, 1999
- Information Visualization: Perception for Design (Interactive Technologies), Colin Ware, Morgan Kaufmann; 2e édition (2004)
- Information Visualization: Design for Interaction, Robert Spence, Prentice Hall; 2e édition (2007)
- Sémiologie Graphique, Bertin, 1967, Réimpression EHESS 2000
- The Visual Display of Quantitative Data, Tufte, 1983, Cheshire, CT: Graphics Press

« Rôles et pouvoirs des femmes au 16e siècle dans la France de l'Ouest »

Application de la visualisation à une problématique historique

- Problème
- Méthode
- Visualisations
- Conclusions

Problèmes

- Sujet à controverse en histoire des femmes
  - Idées reçues
  - Pas de données sérielles
  - Sources difficiles d’accès
- Quelle méthode utiliser ?
  - Vérifiable, réfutable, convaincante
  - Incrémentale, réutilisable

Méthode

- S’appuyer sur les sources primaires
- Transcrire exhaustivement
- Utiliser les technologies XML+TEI
  - Rendre la transcription exploitable, analysable
  - Garder la diplomatique (pas atomiser comme les BDD)
- Lier les résultats des analyses aux données pour vérification
- 1000 actes manuscrits transcrits, codés et annotés

Visualisations

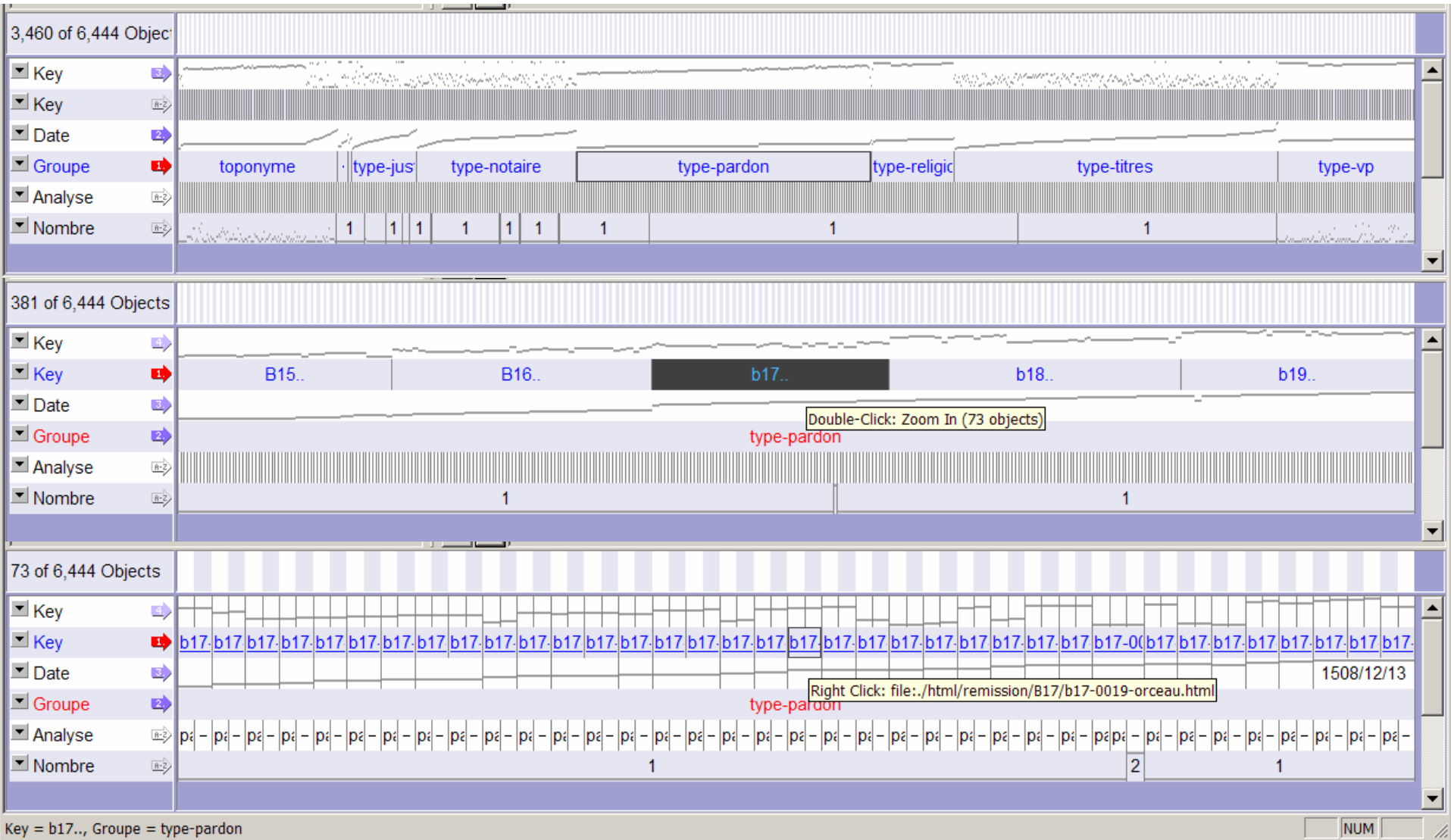
Quatre types de données et représentations:

- Visualisation / exploration de grille de dépouillement
- Arbres généalogiques
- Réseau social
- Cartes

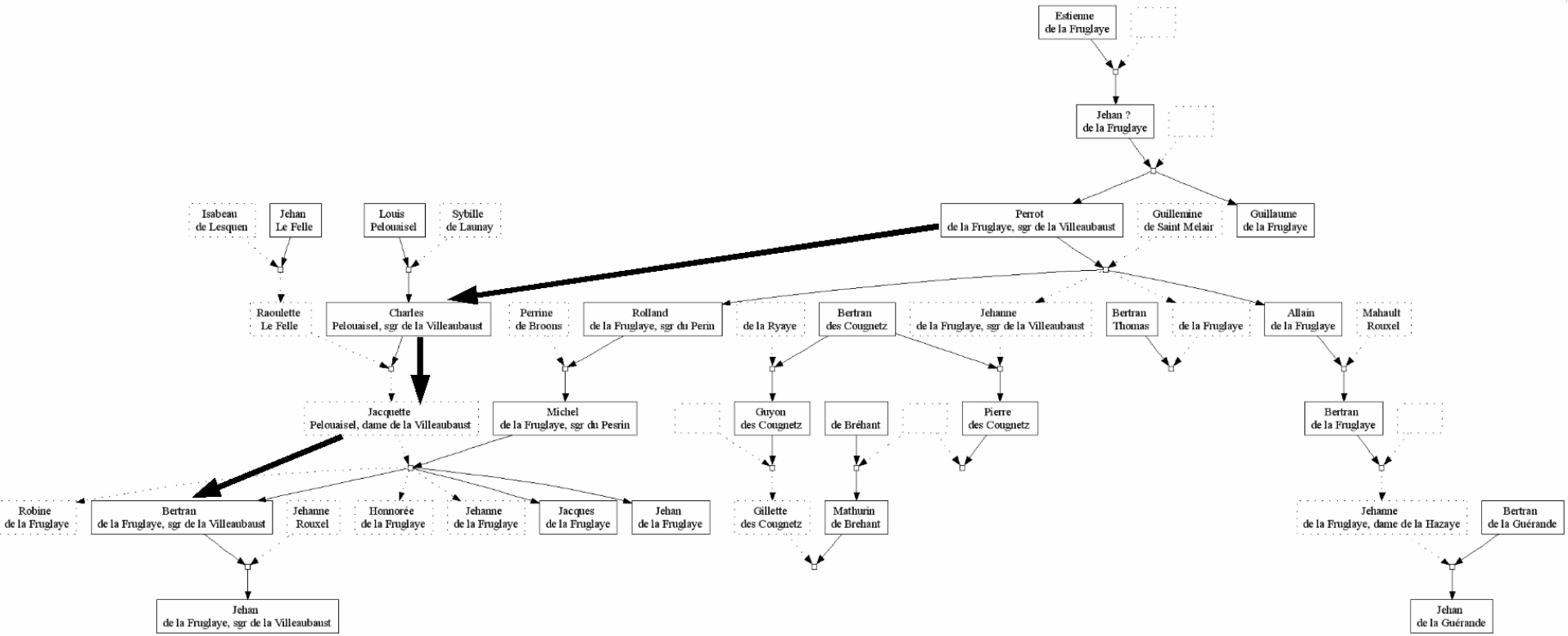
Visualisation / exploration de grille de dépouillement

- Grille de dépouillement
  - Catégorie / sous-catégorie
    - Sexe : masc / fem
    - Type de document : titre de famille / remission / …
    - Economie : sel / commerce / …
  - Région de texte associée à ses catégories
    - <rs ana=« economie-sel »>œillet de marais</rs>
- Extraction automatique de table
  - Nom Document / Date / Cat / Sous-Cat / Nbre
- Programme InfoZoom pour visualiser / naviguer

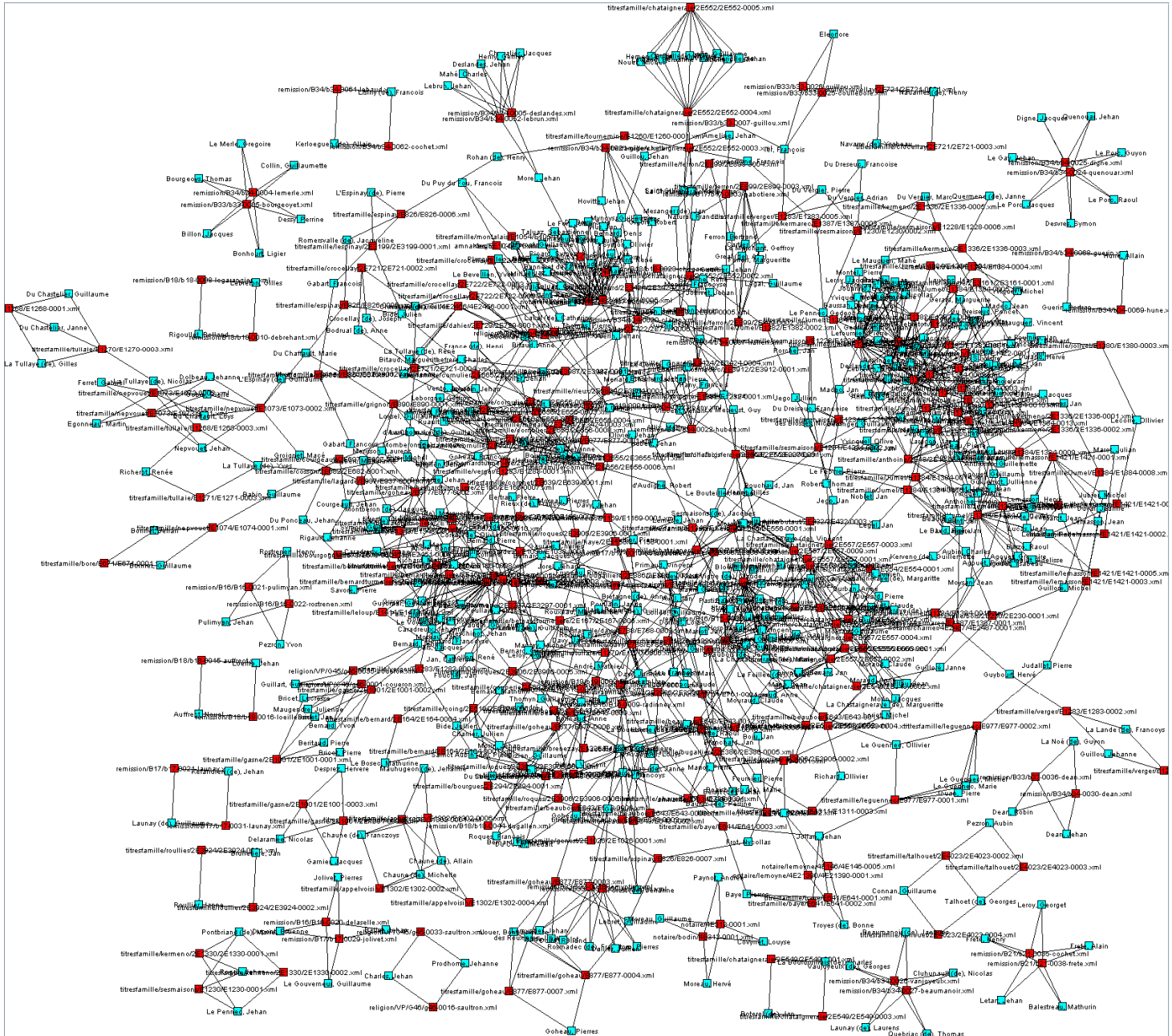
Demo


3,460 of 6,444 Objec
Key
Key
Date
Groupe
Analyse
Nombre
toponyme
type-jus
type-notaire
type-pardon
type-religic
type-titres
type-vp
1
381 of 6,444 Objects
B15..
B16..
b17..
b18..
b19..
Double-Click: Zoom In (73 objects)
type-pardon
73 of 6,444 Objects
1508/12/13
Right Click: file:./html/remission/B17/b17-0019-orceau.html
Key = b17.., Groupe = type-pardon
NUM

Transfert de terres « La Fruglaye »



Réseau social Personne/Document



Cartes de migrations

Conclusions historiques

- La méthode a permis d'obtenir des résultats à partir de données lacunaires
- Liens entre analyses et sources pour vérifier
- Travail de transcription / codage long et fastidieux mais
  - Analyse et interprétations beaucoup plus faciles
- Publication des sources en ligne:
  - nicole.dufournaud.net/these