TRAITEMENT DE L'INFORMATION

Ophélie Fraisier
ophelie.fraisier@irit.fr

2018 — 2019

ANALYSES DE BASE

- Présentation de quelques outils statistiques pour analyser un jeu de données
  - Valeurs remarquables
  - Liens entre variables
  - Catégorisation

MÉDIANE / MOYENNE

- **Moyenne** : indicateur le plus simple pour résumer l'information fournie par un ensemble de données, elle est égale à la somme de ces données divisée par leur nombre.
- **Médiane** : si on ordonne un ensemble de données, la médiane est la valeur qui partage cet ensemble en deux parties égales.

Remarque : la médiane est un indicateur plus robuste que la moyenne car elle n'est pas affectée par les valeurs extrêmes.

Salaires nets en France

Moyenne : 2202 €

Médiane : 1772 €

(fr)

```
=MOYENNE(A1:A50)
=MEDIANE(A1:A50)
```

(en)

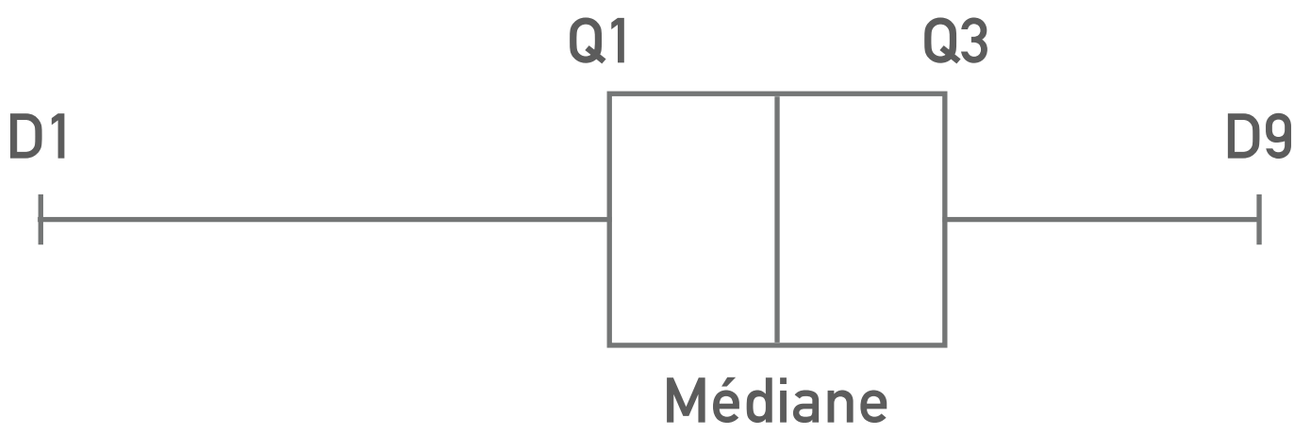
```
=AVERAGE(A1:A50)
=MEDIAN(A1:A50)
```

QUANTILES ET BOÎTE À MOUSTACHES (BOXPLOT)

- **Quantiles :** valeurs divisant un jeu de données en intervalles contenant le même nombre de données.

	Nombre d'intervalles	Notation
Médiane	2	
Quartiles	4	Q1, Q2, Q3
Déciles	10	D1 … D9
Centiles	100	C1 … C99

- **Boîte à moustache :**



(fr)

```
=CENTILE(données;alpha)
```

alpha entre 0 et 1

alpha=0,25 : Q1

alpha=0,5 : Médiane

(en)

```
=PERCENTILE(données;alpha)
```

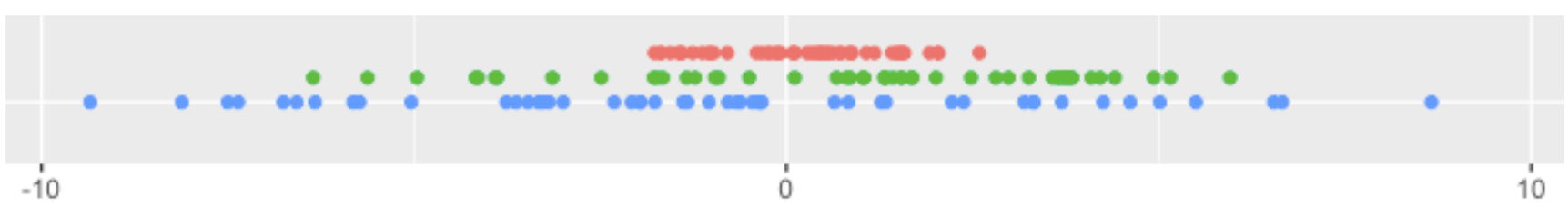
DISPERSION

- La **variance** sert à mesurer la dispersion au carré d'un ensemble de valeurs autour de leur moyenne.
- L'**écart-type** est la racine carrée de la variance.

$$s^2 = \frac{\sum(x - \bar{x})^2}{n}$$

$$s = \sqrt{s^2}$$

- Plus ces valeurs sont faibles, plus la population est homogène.



$$s = \begin{matrix} \bullet\ 1 \\ \bullet\ 3 \\ \bullet\ 5 \end{matrix}$$

(fr)

```
=ECARTYPE(A1:A50)
=VAR(A1:A50)
```

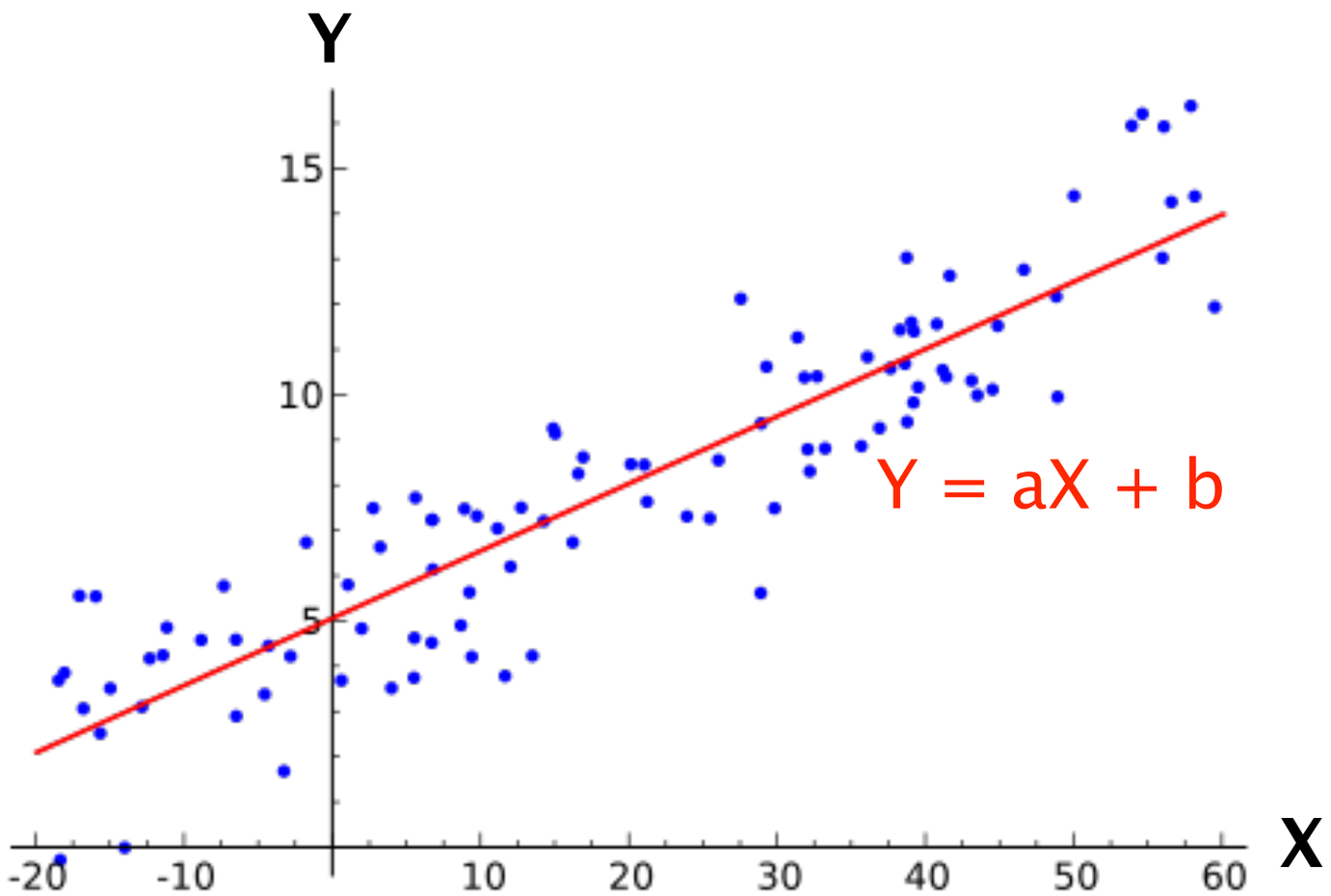
(en)

```
=STDEV(A1:A50)
=VAR(A1:A50)
```

RÉGRESSIONS

- 2 variables X et Y sont **corrélées** s'il existe un lien entre elles

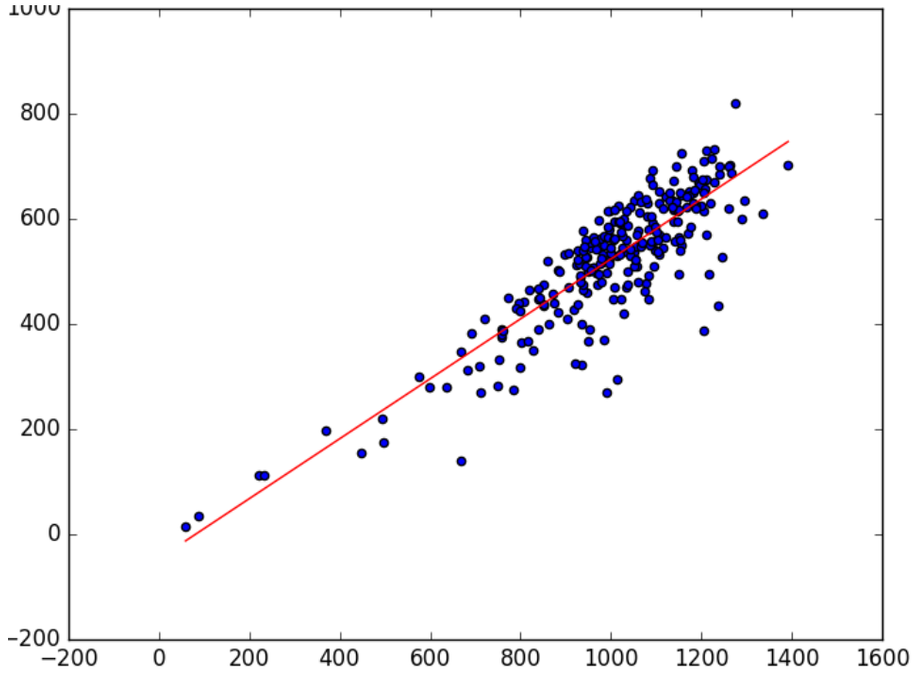
- Une **régression linéaire** permet de déterminer la **droite de régression** $Y = aX + b$



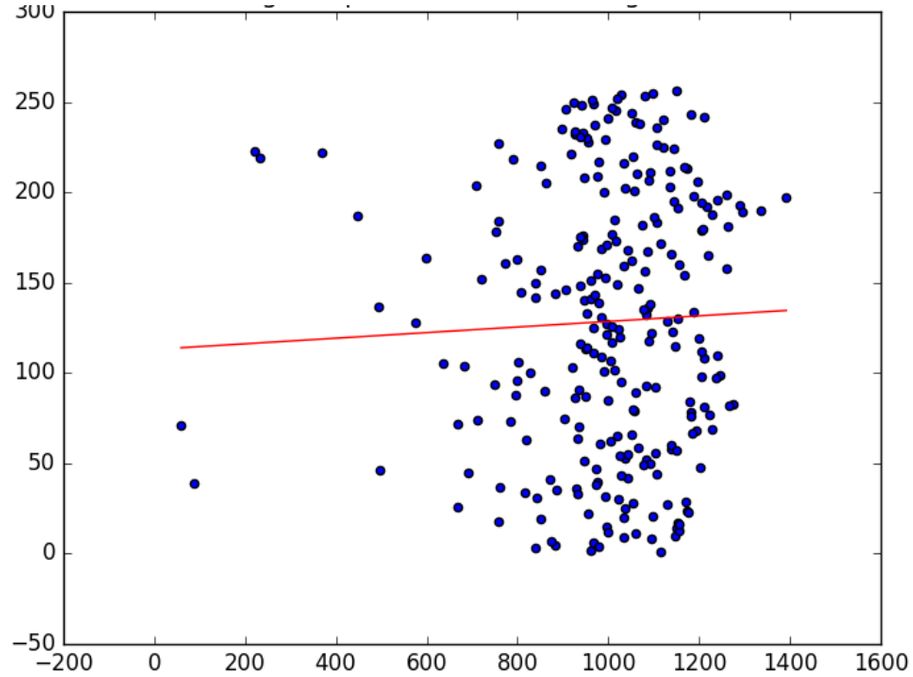
- D'autres types de régression sont possibles : régression polynomiale, régression circulaire, …

COEFFICIENT DE DÉTERMINATION R^2

- Mesure la qualité de la régression et donc le pouvoir prédictif du modèle
- Varie entre 0 et 1 :
  - = 0 : le modèle testé représente très mal les données
  - = 1 : le modèle testé passe par tous les points du nuage



$R^2 = 0{,}75$



$R^2 = 0$

REMARQUES

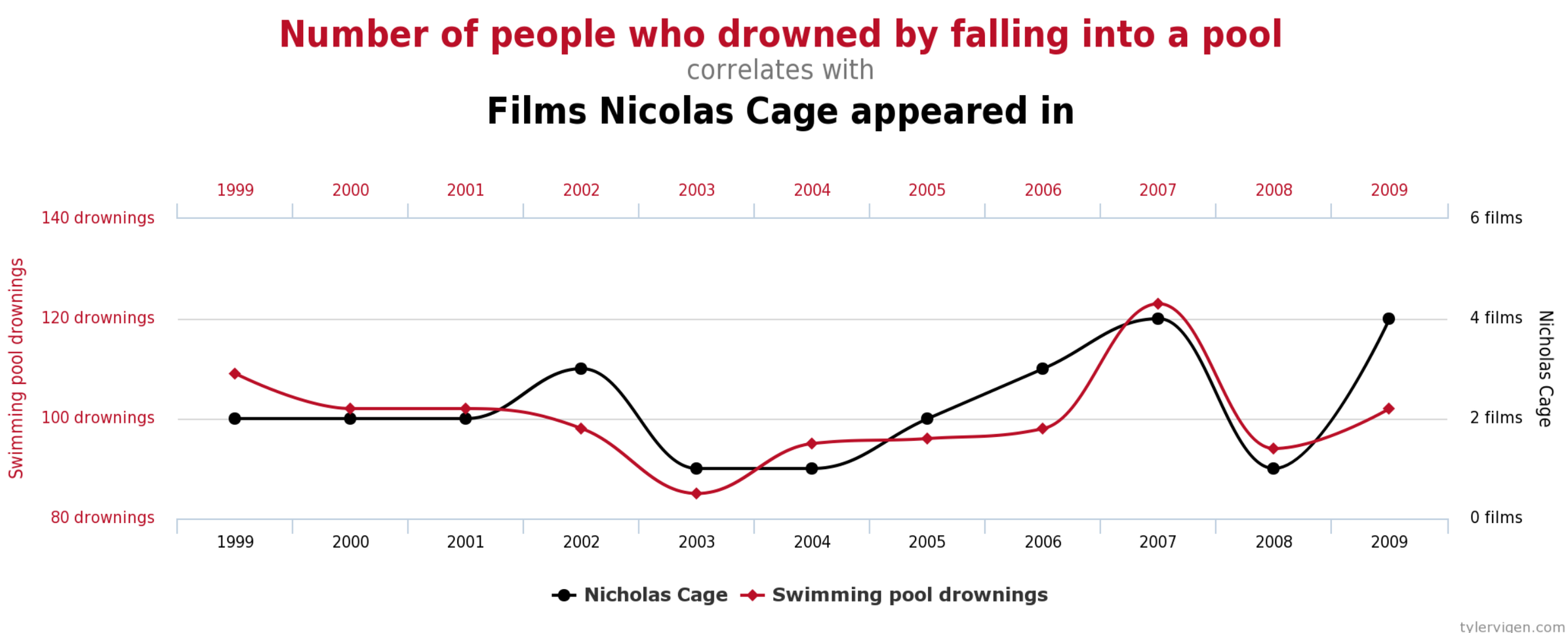
- R^2 nul ≠ Variables indépendantes

- **Corrélation ≠ Causalité**
  - Plus forte corrélation trouvée dans les années 50 :
    - Consommation de bière sur la côte Ouest des États-Unis
    - Mortalité infantile au Japon.
  - Cause des variations communes à chercher dans le contexte historique de l'après-guerre.

REMARQUES

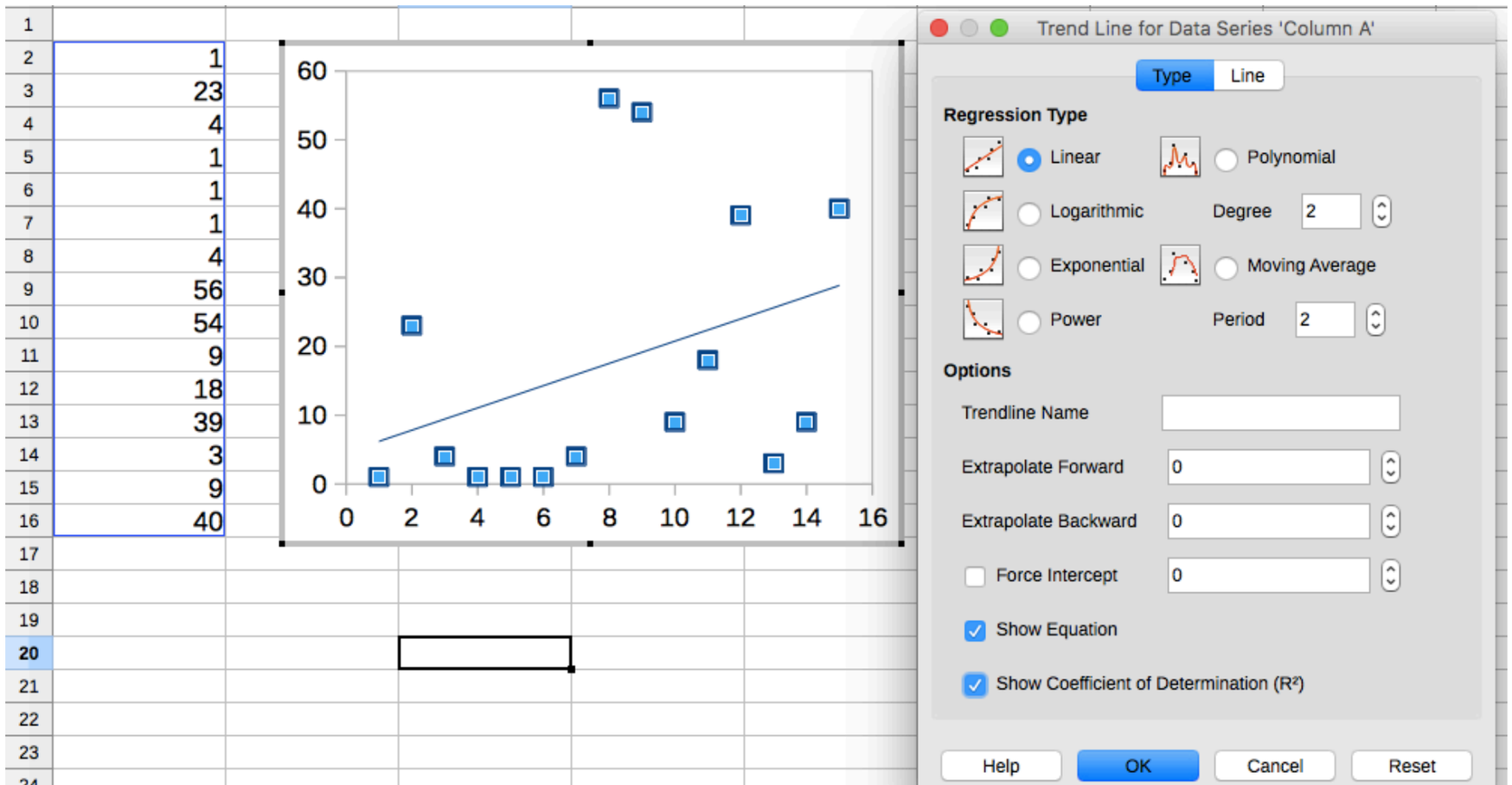
- R^2 nul ≠ Variables indépendantes

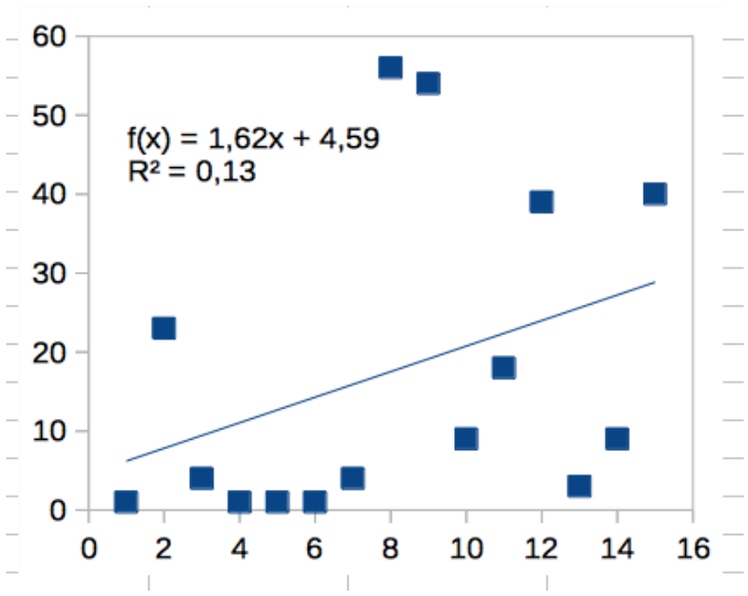
- **Corrélation ≠ Causalité**



RÉGRESSIONS DANS LIBREOFFICE

- Créer un diagramme avec les données souhaitées : **Insertion > Graphique**
- **Sélectionner la série de données sur le graphe**, puis faire **Insertion > Courbes de tendance**
- Sélectionner le type de régression souhaitée et cocher « **Afficher l'équation** » et « **Afficher le coefficient de détermination** »



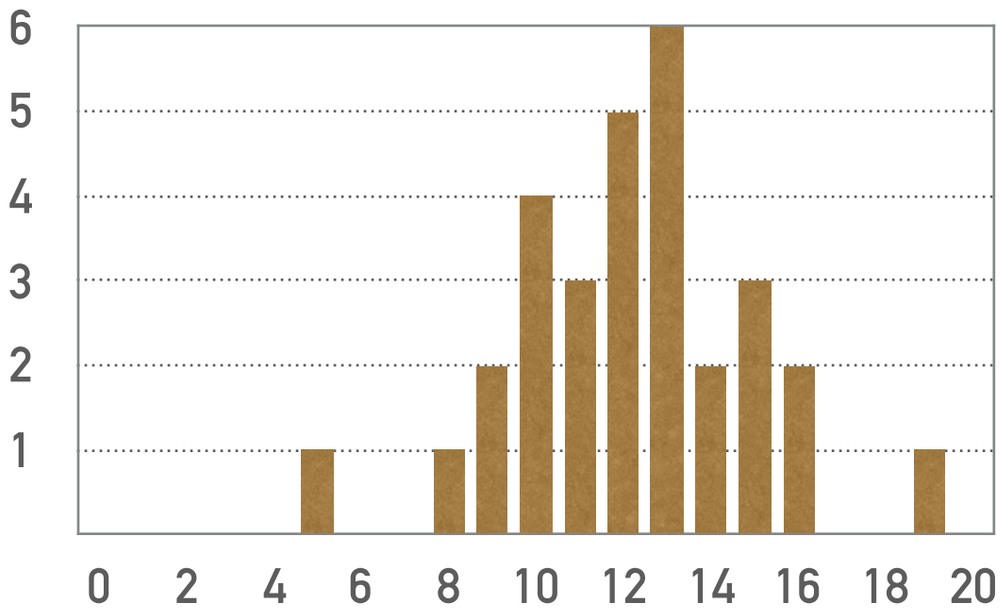


REPRÉSENTATIONS GRAPHIQUES

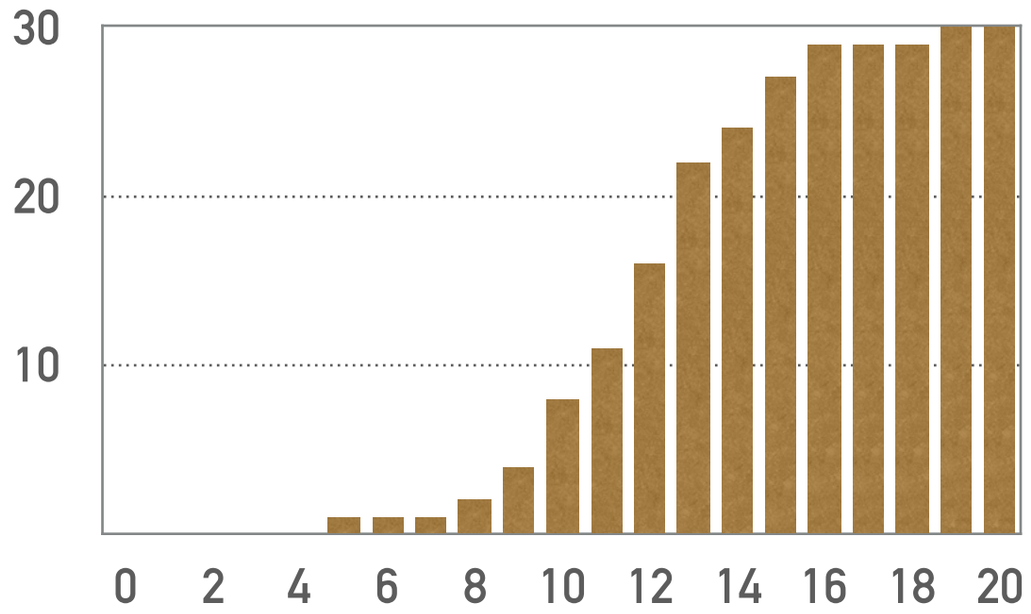
- Représentations de données quantitatives
- Nuages de mots
- Réseaux

DONNÉES QUANTITATIVES — DISTRIBUTION

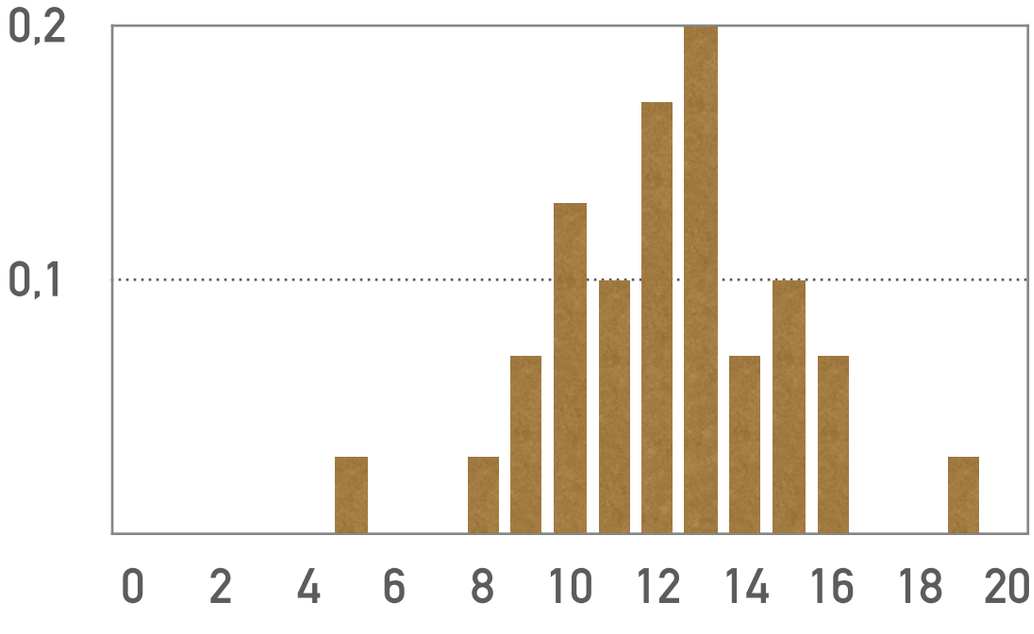
Effectifs



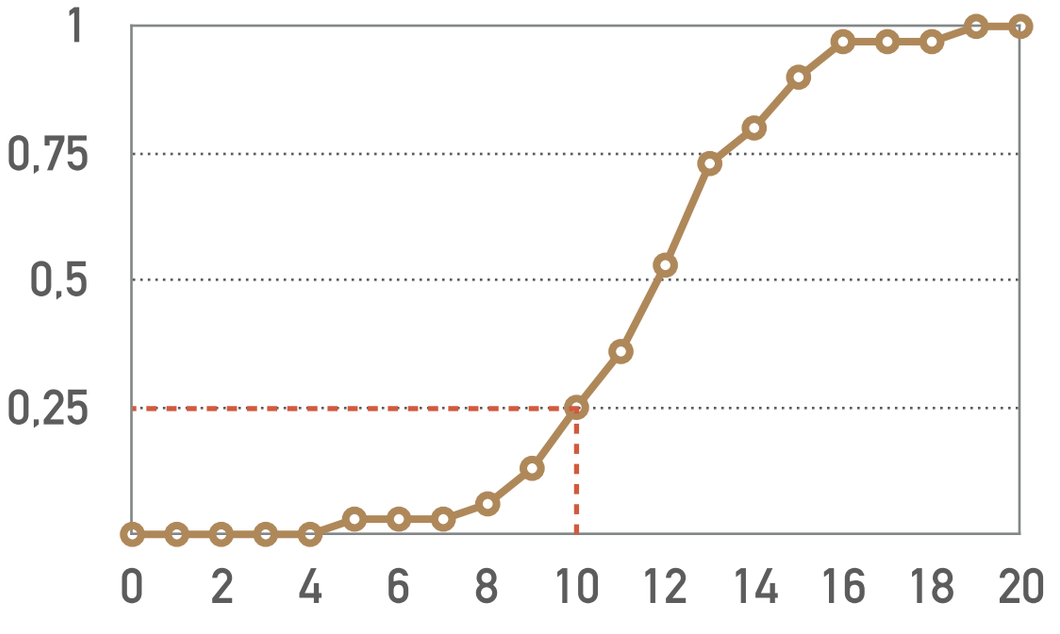
Effectifs cumulés



Fréquences

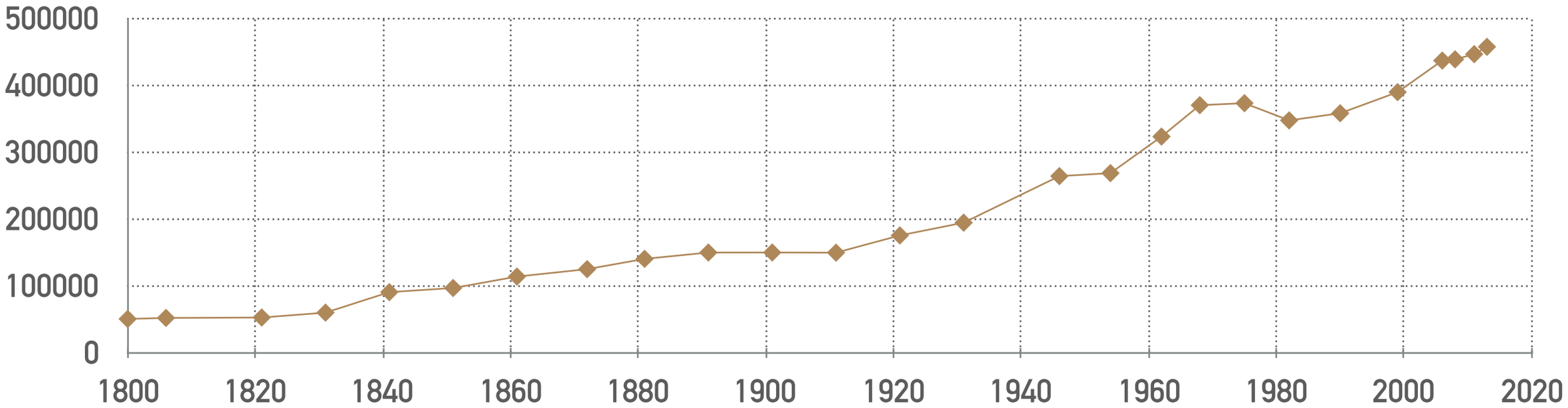


Fréquences cumulées

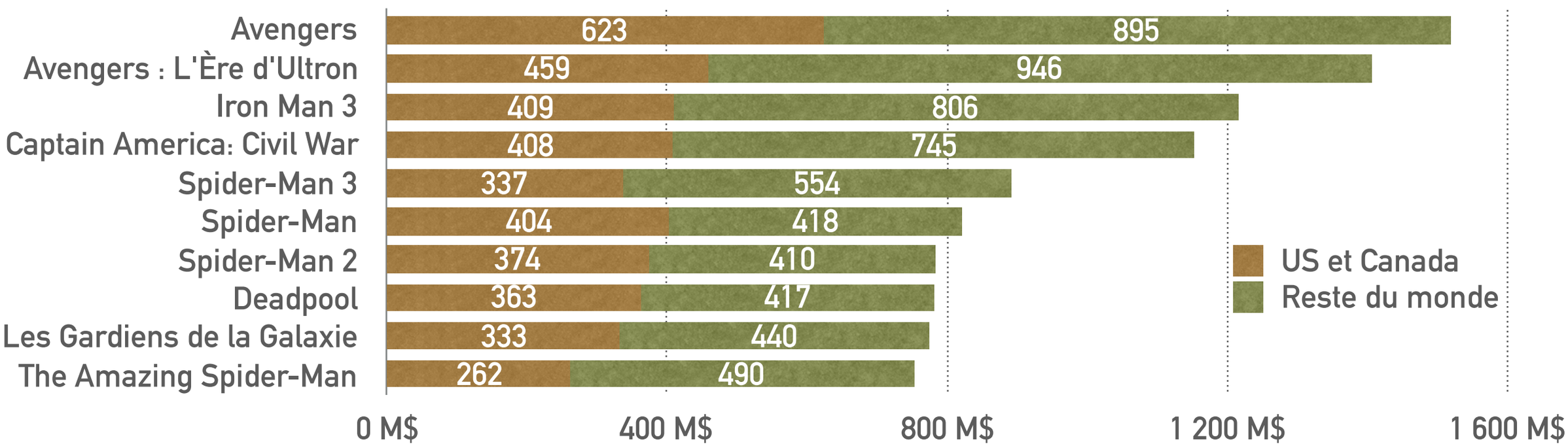


DONNÉES QUANTITATIVES — GRAPHIQUES

Évolution de la population toulousaine

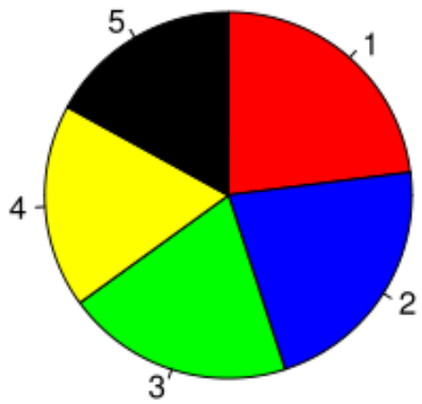


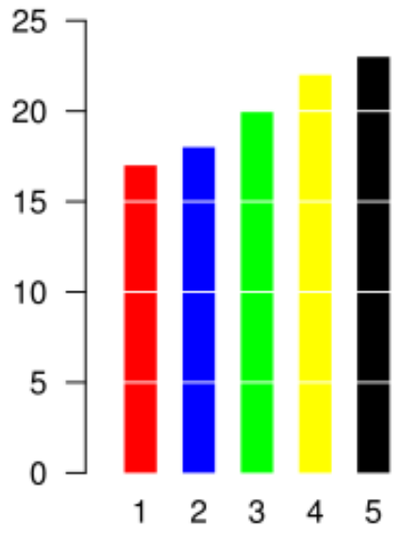
Recettes des films Marvel

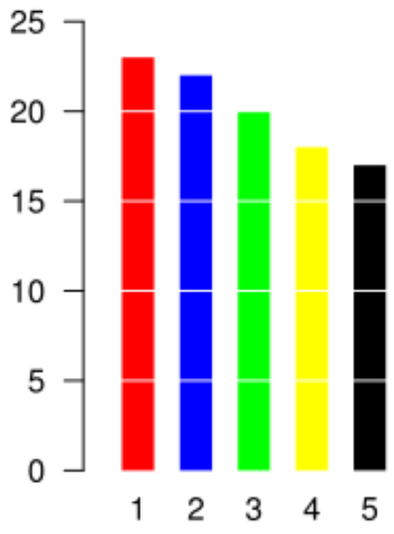


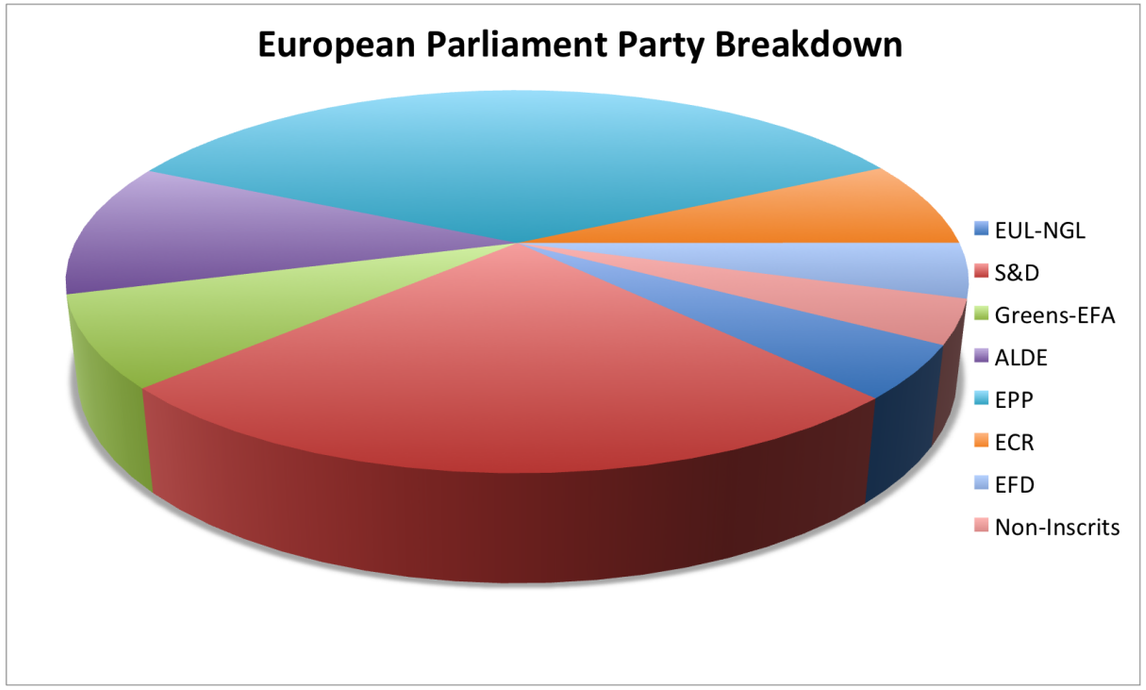
DONNÉES QUANTITATIVES — PIE CHARTS

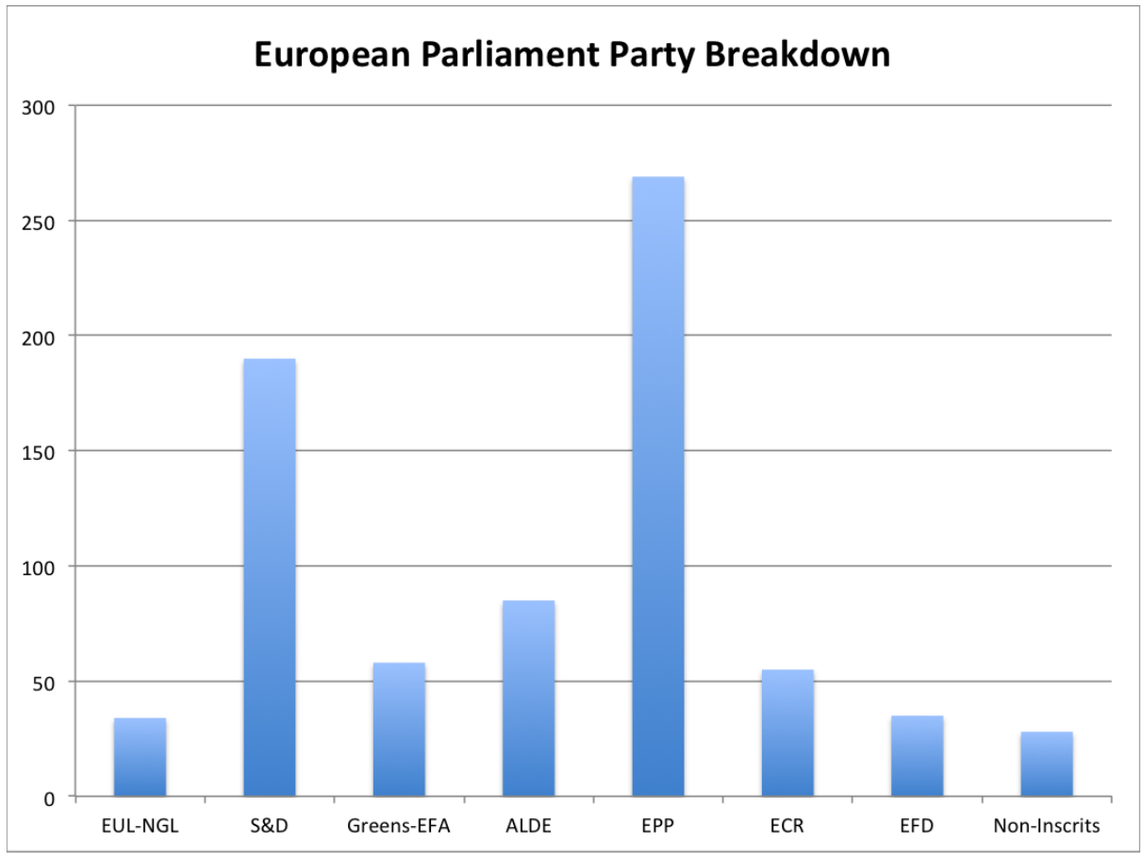
- Peu efficace pour comparer des valeurs









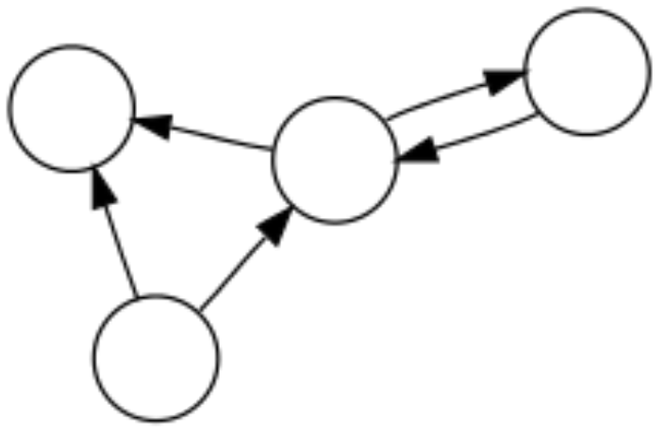


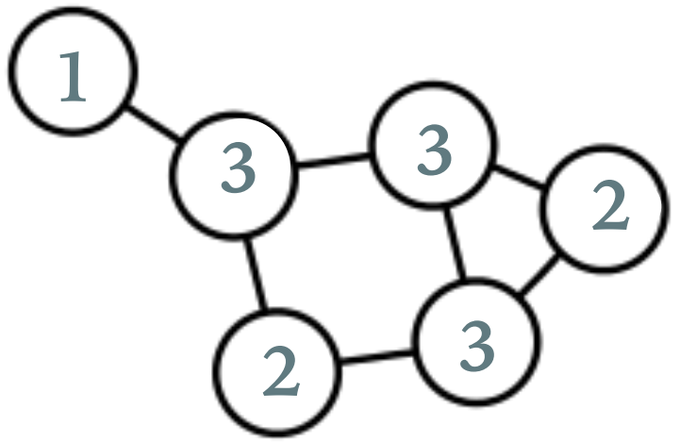
NUAGES DE MOTS

Nuage de mots représentants les titres des films sortis aux États-Unis de 2000 à 2014

GRAPHES

- Graphe = ensemble de **nœuds** (ou sommets) reliés par des **liens** (ou arêtes)
- Un graphe peut être **dirigé** ou **non dirigé**



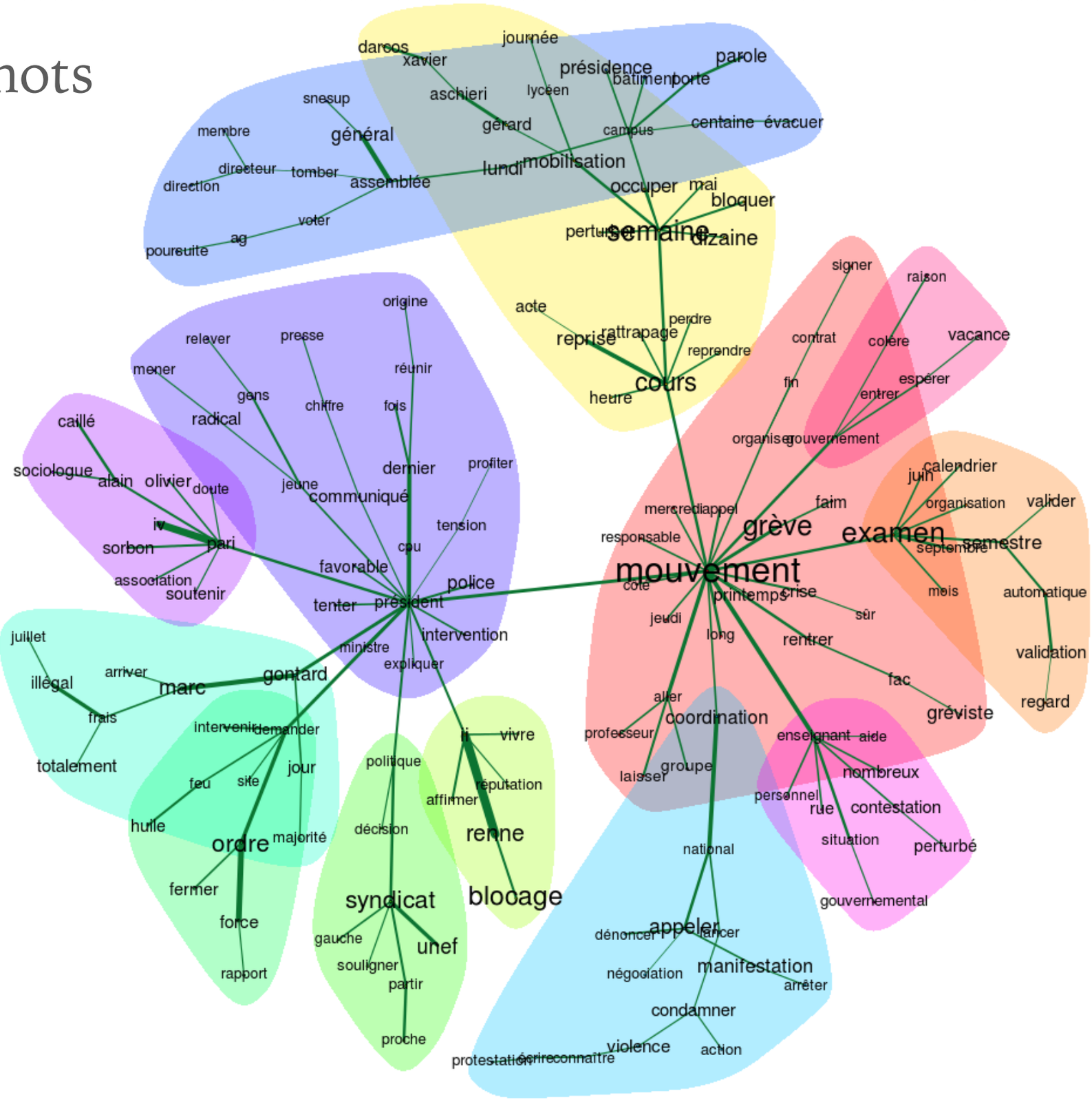


- **Degré** d'un nœud = nombre de liens le connectant aux autres
- Il est possible de donner un **poids** à un lien

GRAPHES

- Liens entre des mots

Discours sur la réforme de l'université

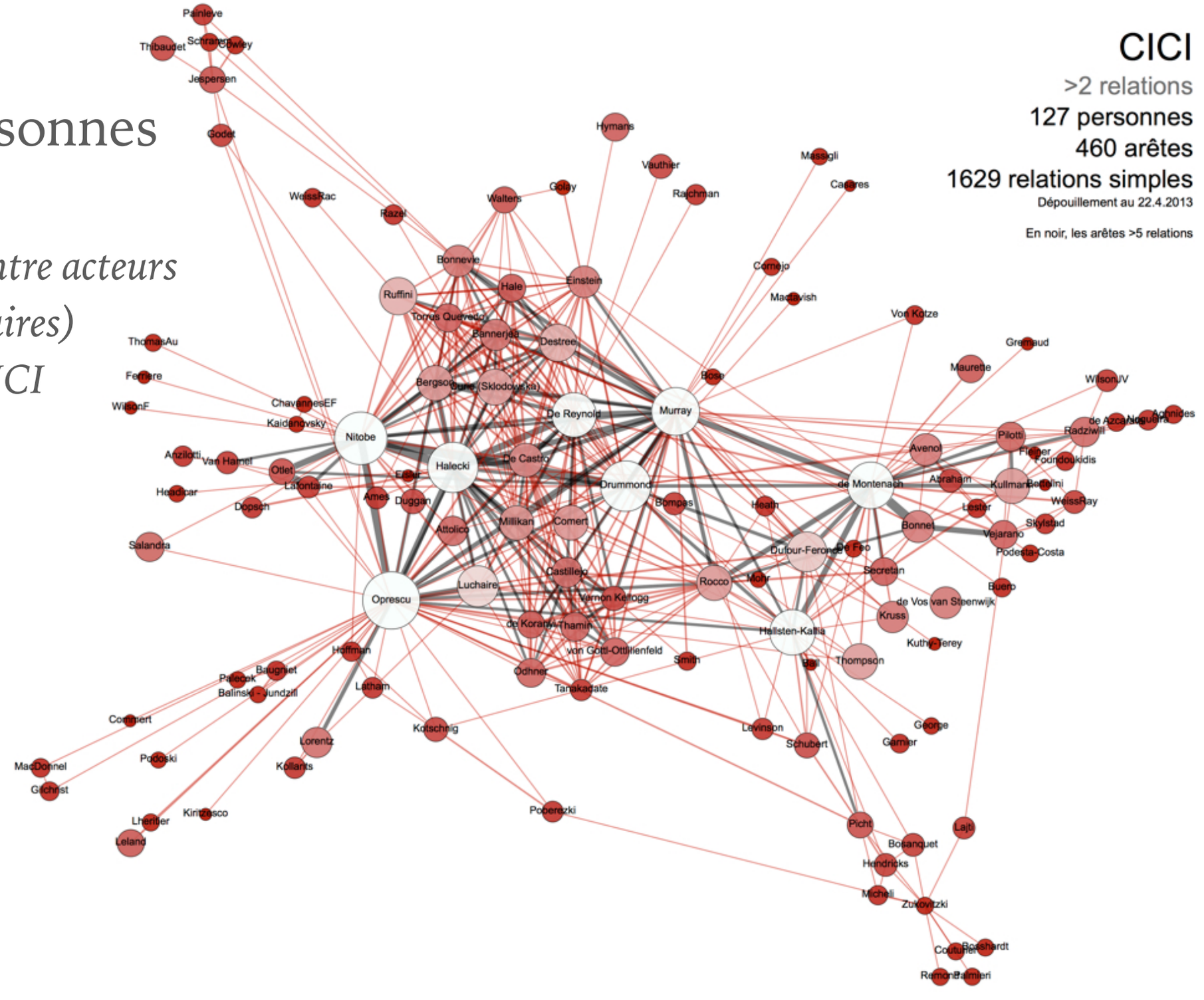


Source : IRaMuTeQ

GRAPHES

- Liens entre personnes

Réseau des relations entre acteurs (expéditeurs/destinataires) des documents de la CICI

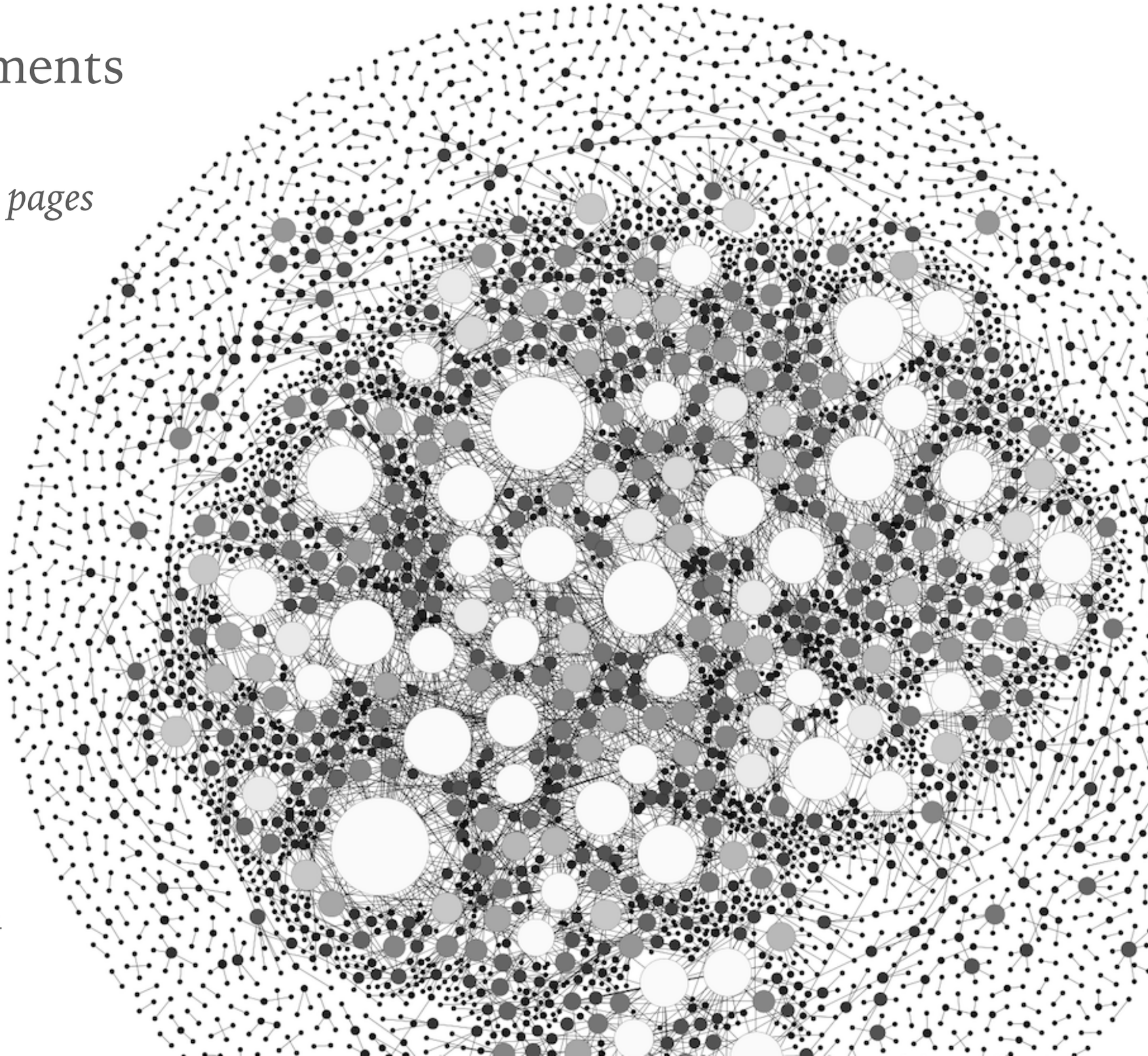


Source : www.martingrandjean.ch/analyse-de-reseau-nouvel-outil-exploration-fonds-archives/

GRAPHES

- Liens entre documents

Réseau d'hyperliens entre pages Wikipedia



Source : www.martingrandjean.ch/la-connaissance-est-un-reseau-perspective-sur-lorganisation-archivistique-et-encyclopedique/