Université Paris-Est Marne-la-Vallée
29/09/2016

Traitement et visualisation de données ouvertes

UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Philippe Gambette

Plan

L’ère des données

Traitement et visualisation des données

Quelques outils pratiques

Source: Design vector designed by Freepik

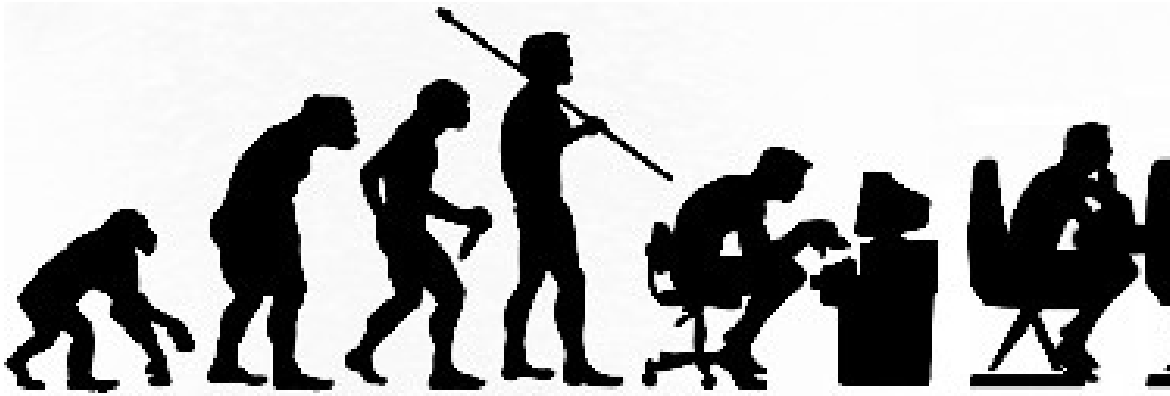
La révolution des données

3° étape de la révolution numérique ?

Révolution informatique

Révolution internet

Révolution de la donnée



Source : http://3.bp.blogspot.com/-Z8EYq8M8_AM/UGN_yXYbotI/AAAAAAAABqI/g_3rB9vg-xk/s400/airline-evolution.png

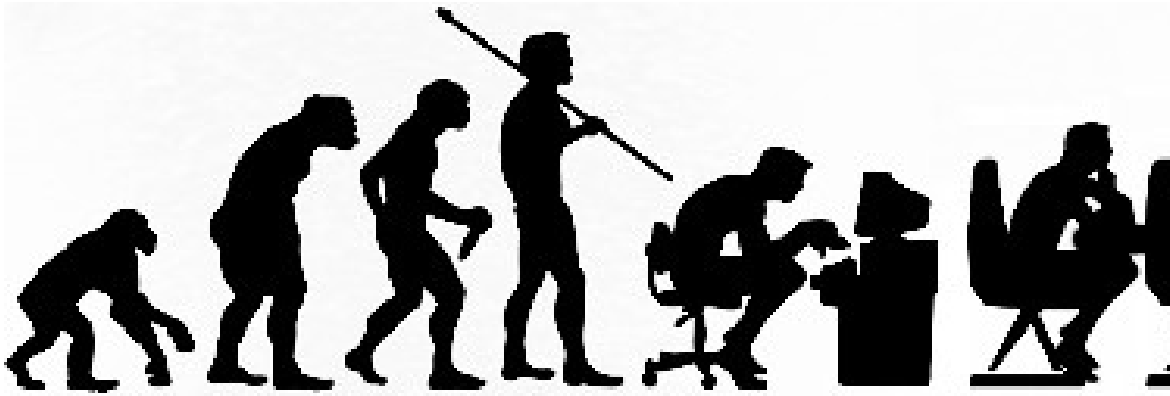
La révolution des données

« data scientist » : informatique, mathématiques, stratégie

Révolution informatique

Révolution internet

Révolution de la donnée

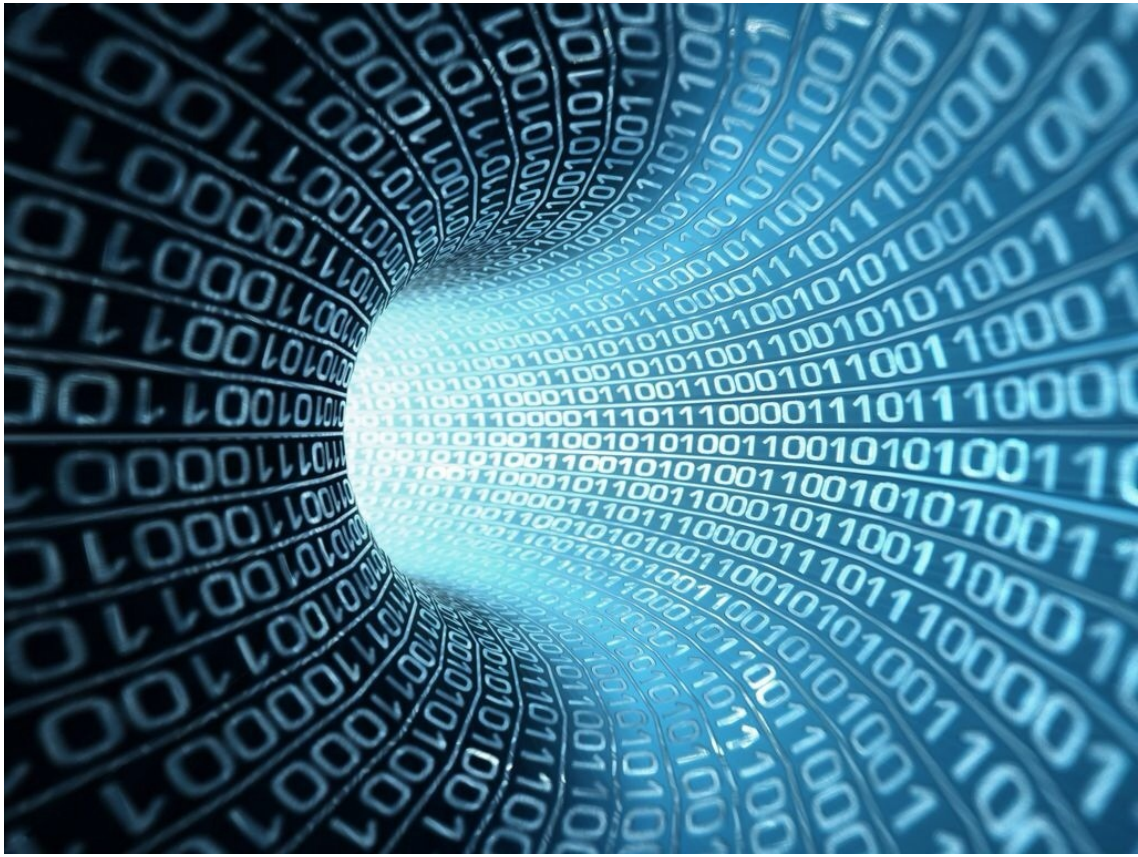


Source : http://3.bp.blogspot.com/-Z8EYq8M8_AM/UGN_yXYbotI/AAAAAAAABqI/g_3rB9vg-xk/s400/airline-evolution.png

Données ouvertes, «*open data*»

Des données de plus en plus :

- **accessibles**
- **réutilisables**
- **stockables**
- **… traitables !**

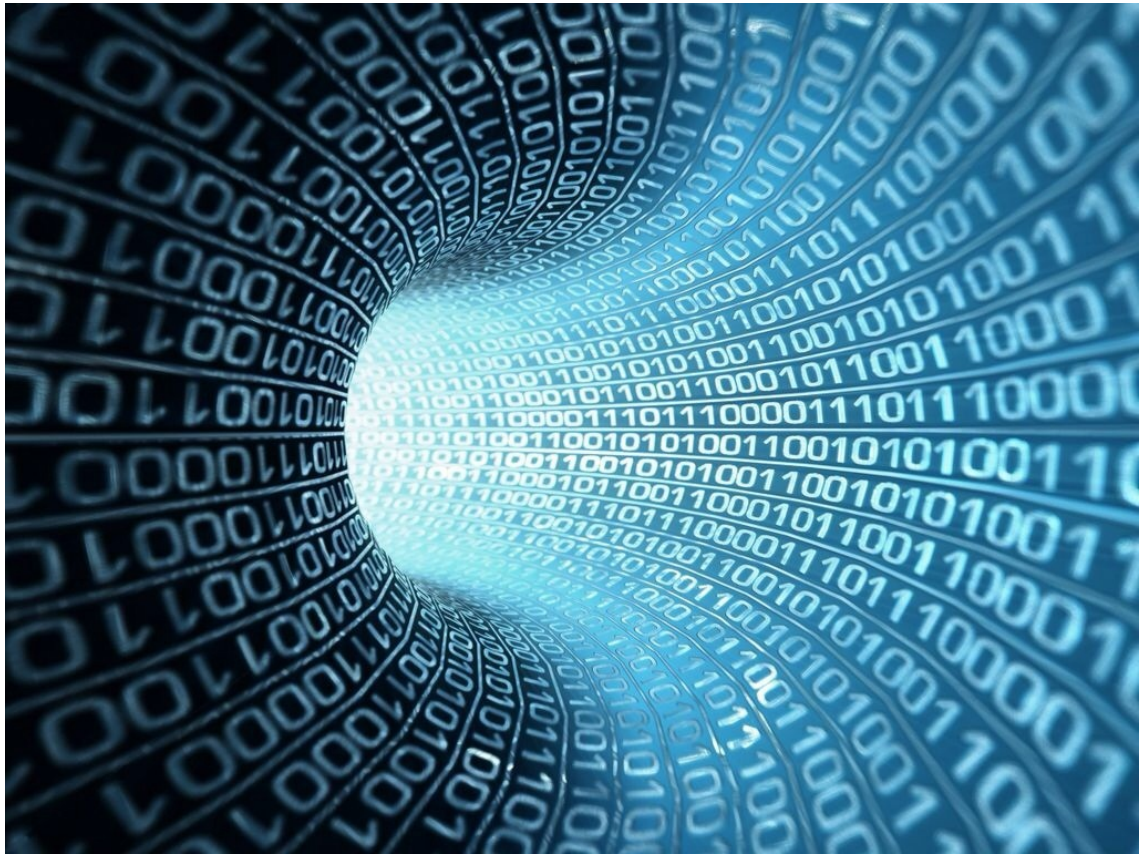


Source : http://www.economie.gouv.fr/files/eco_numerique2.png

Données ouvertes, «*open data*»

Des données de plus en plus :

- **accessibles**
- **réutilisables**
- **stockables**
- **… traitables !**



Source : http://www.economie.gouv.fr/files/eco_numerique2.png

Henri Verdier, Chief Data Officer français,
directeur d'Etalab (https://www.etalab.gouv.fr/)

Données géographiques

Des données sur :

- **Google Maps : « mashups »**
- **Open Street Map**

Données géographiques

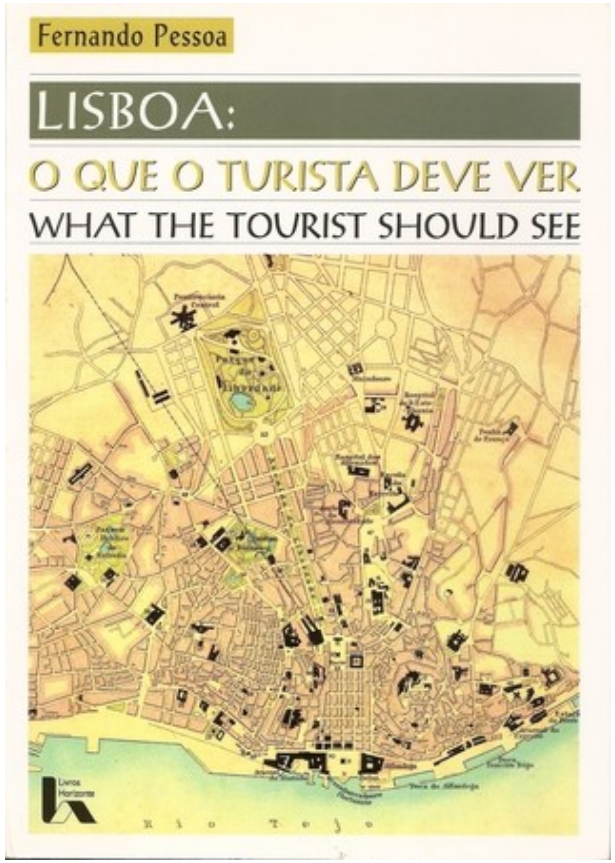
Des données sur :

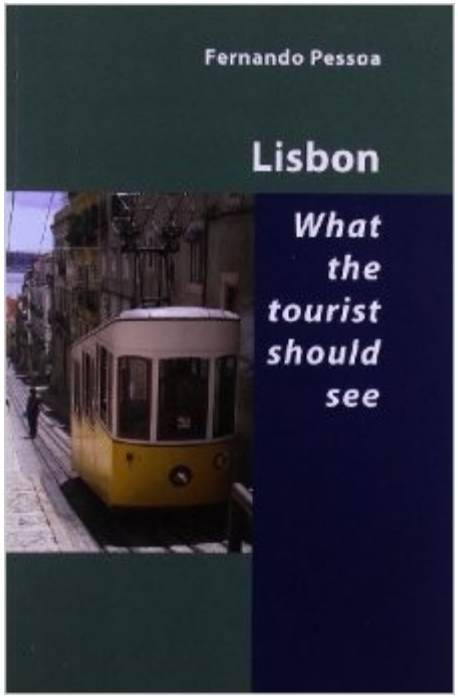
- **Google Maps : « mashups »**
- **Open Street Map**
- **Base Adresse Nationale (en open data) :**
  http://adresse.data.gouv.fr/

Données géographiques

Géolocalisation de *Lisbonne* par Pessoa

Guide touristique écrit en 1925 par Fernando Pessoa, en anglais

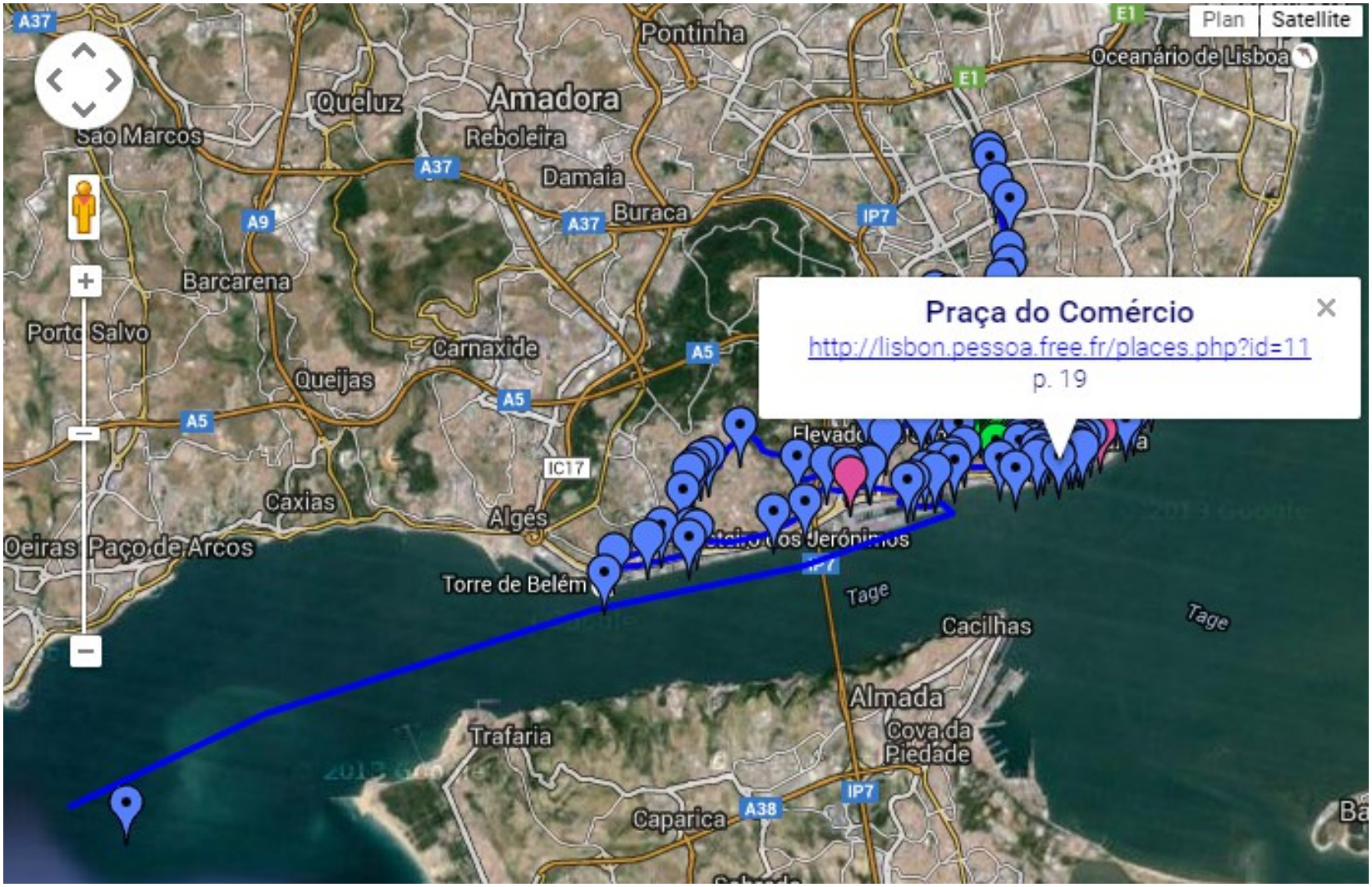




http://lisbon.pessoa.free.fr

Géolocalisation de *Lisbonne* par Pessoa

Géolocalisation manuelle Google Maps :



Géolocalisation de *Lisbonne* par Pessoa

Géolocalisation automatique Google Maps :

On the North side of the square, facing the river, there are three parallel streets; the middle one issues from a magnificent triumphal arch[12] of great dimensions, indubitably one of the largest ones in Europe. It is dated 1873, but it was designed by Veríssimo José da Costa and began to be built in 1755.

The allegoric group which crowns the arch, sculptured by Calmels, personifies Glory crowning Genius and Valour; and the recumbent figures, which represent the rivers Tagus and Douro, as well as the statues of Nun'Alvares, Viriato, Pombal and Vasco da Gama, are due to the sculptor Victor Bastos.

The Terreiro do Paço is one of the places where boats are taken to cross the river; on the right-hand side, facing the river and on it, is the provisional station of the Southern Railways. It also often happens that tourists land here, as commonly do the crews of foreign men-of-war which visit the port. There is also a carriage and motor-car stand in this square. The general aspect of the square is of a kind to give a very agreeable impression to the most exacting of tourists.

From the Praça do Commercio we can go on to the centre of city by any of the three streets which go North from there - Rua do Ouro[13] on the left, Rua Augusta (the one with the arch) in the middle, and Rua da Prata on the right. Let us choose Rua do Ouro, which, owing to its commercial importance is the main street of the city. There are several banks, restaurants, and shops of all kinds in this street; many of the shops, especially towards the upper end of the artery will be found to be as luxurious as their Parisian equivalents.

Almost at the upper end of the street, on the left-hand side as we go up, there is the *Santa Justa Elevator*[14], so called because the transversal street in which it is built is called Rua de Santa Justa. This is one of the "sights" of Lisbon and always compels great admiration from tourists from everywhere.

It is due to a French engineer, Raoul Mesnier, to whom other interesting projects are also due. The elevator is all built in iron, but it is extremely distinctive, light and safe. There are two lifts, worked by electricity. It goes up to Largo do Carmo, where there are the ruins of Carmo Church[15], now the Archeological Museum. Authority is needed to go right up to the top, above where the lifts themselves stop; from there a magnificent panorama is got of the whole city and the river. The elevator belongs to the Electric Tramway Company.

Géolocalisation de *Lisbonne* par Pessoa

Géolocalisation automatique Google Maps :

river, there are three parallel streets; the middle great dimensions, indubitably one of the largest gned by Veríssimo José da Costa and began to

ch, sculptured by Calmels, personifies Glory figures, which represent the rivers Tagus and ato, Pombal and Vasco da Gama, are due to the

ere boats are taken to cross the river; on the right-hand side, facing the river rn Railways. It also often happens that tourists land here, as commonly do ort. There is also a carriage and motor-car stand in this square. The general reeable impression to the most exacting of tourists.

o we can go on to the centre of city by any of the three streets which go North from there - Rua do Ouro[13] on the left, Rua Augusta (the one with the arch) in the middle, and Rua da Prata on the right. Let us choose Rua do Ouro, which, owing to its commercial importance is the main street of the city. There are several banks, restaurants, and shops of all kinds in this street; many of the shops, especially towards the upper end of the artery will be found to be as luxurious as their Parisian equivalents.

Almost at the upper end of the street, on the left-hand side as we go up, there is the Santa Justa Elevator[14], so called because the transversal street in which it is built is called Rua de Santa Justa. This is one of the "sights" of Lisbon and always compels great admiration from tourists from everywhere.

It is due to a French engineer, Raoul Mesnier, to whom other interesting projects are also due. The elevator is all built in iron, but it is extremely distinctive, light and safe. There are two lifts, worked by electricity. It goes up to Largo do Carmo, where there are the ruins of Carmo Church[15], now the Archeological Museum. Authority is needed to go right up to the top, above where the lifts themselves stop; from there a magnificent panorama is got of the whole city and the river. The elevator belongs to the Electric Tramway Company.

Géolocalisation de *Lisbonne* par Pessoa

Géolocalisation automatique Google Maps :

river, there are three parallel streets; the middle great dimensions, indubitably one of the largest gned by Veríssimo José da Costa and began to

ch, sculptured by Calmels, personifies Glory figures, which represent the rivers Tagus and ato, Pombal and Vasco da Gama, are due to the

Base de données MySQL ; PHP ; Javascript

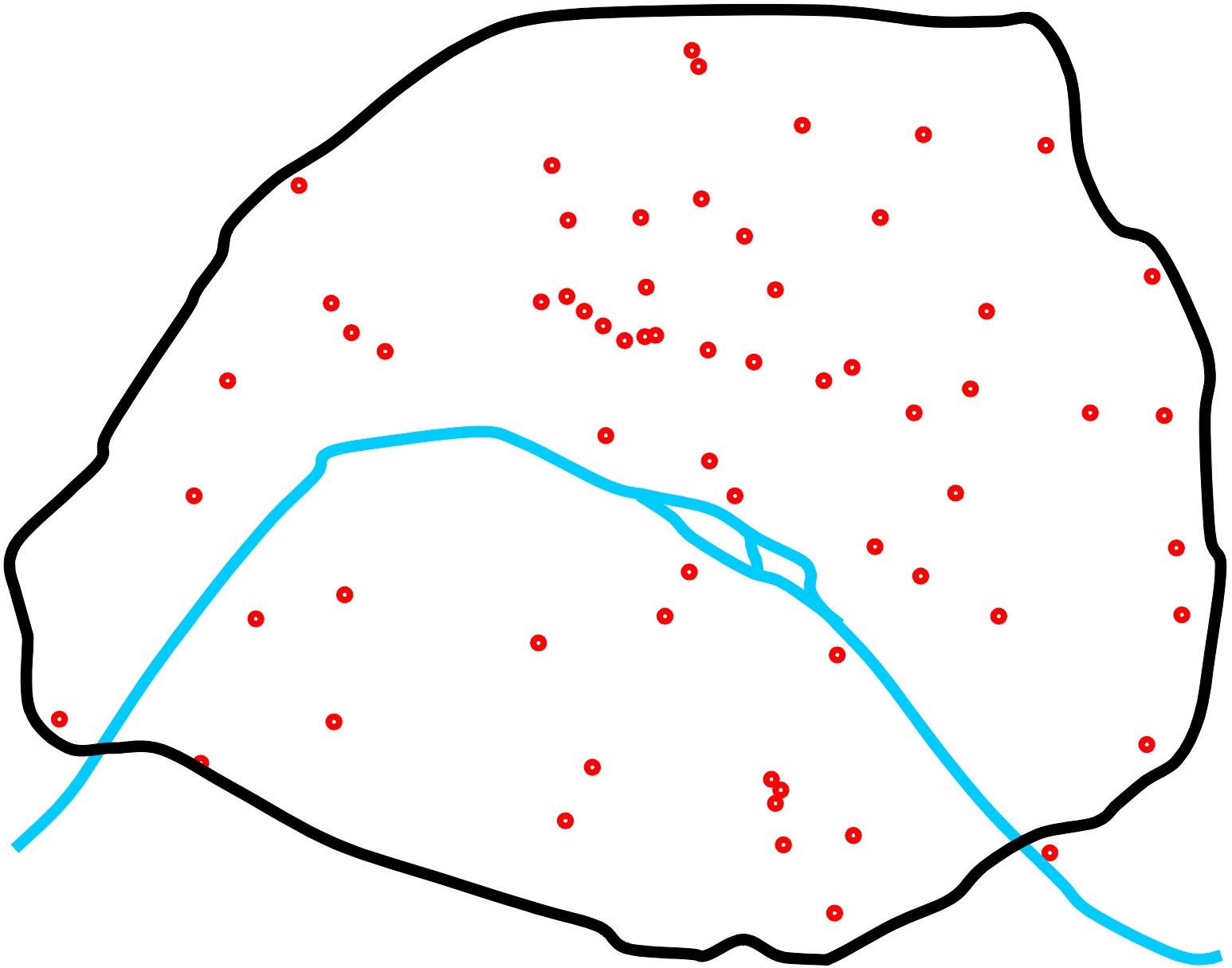
ere boats are taken to cross the river; on the right-hand side, rn Railways. It also often happens that tourists land here, as ort. There is also a carriage and motor-car stand in this square. reeable impression to the most exacting of tourists.

o we can go on to the centre of city by any of the three streets which North from there - Rua do Ouro[13] on the left, Rua Augusta (the one with the arch) in the middle, and Rua Prata on the right. Let us choose Rua do Ouro, which, owing to its commercial importance is the main street of the city. There are several banks, restaurants, and shops of all kinds in this street; many of the shops, especially towards the upper end of the artery will be found to be as luxurious as their Parisian equivalents.

Almost at the upper end of the street, on the left-hand side as we go up, there is the Santa Justa Elevator[14], so called because the transversal street in which it is built is called Rua de Santa Justa. This is one of the "sights" of Lisbon and always compels great admiration from tourists from everywhere.

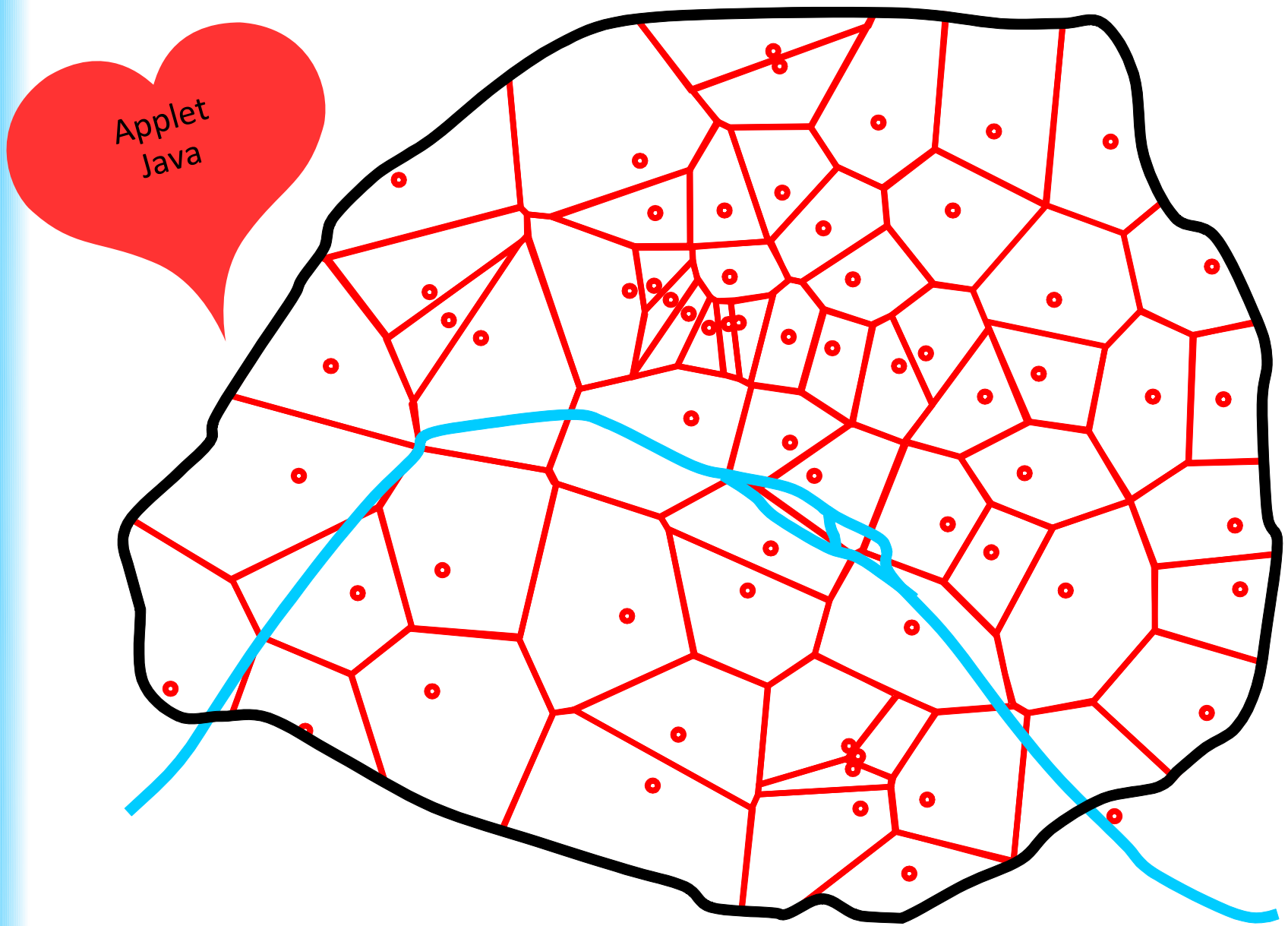
It is due to a French engineer, Raoul Mesnier, to whom other interesting projects are also due. The elevator is all built in iron, but it is extremely distinctive, light and safe. There are two lifts, worked by electricity. It goes up to Largo do Carmo, where there are the ruins of Carmo Church[15], now the Archeological Museum. Authority is needed to go right up to the top, above where the lifts themselves stop; from there a magnificent panorama is got of the whole city and the river. The elevator belongs to the Electric Tramway Company.

Diagramme de Voronoi des McDos parisiens



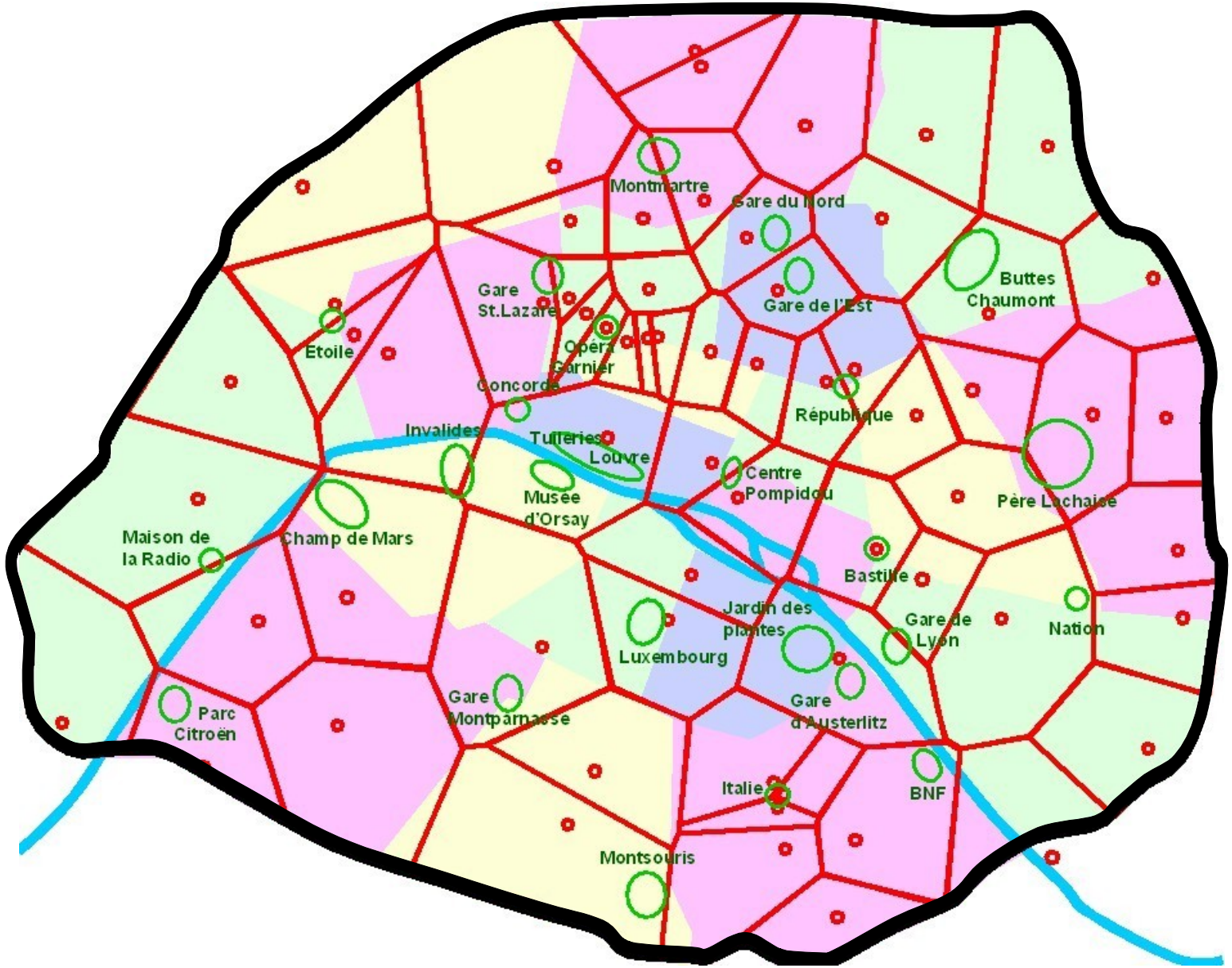
http://gambette.blogspot.fr/2006/10/mcdonalds-macdo-mac-donalds-et-vorono.html

Diagramme de Voronoi des McDos parisiens



http://gambette.blogspot.fr/2006/10/mcdonalds-macdo-mac-donalds-et-vorono.html

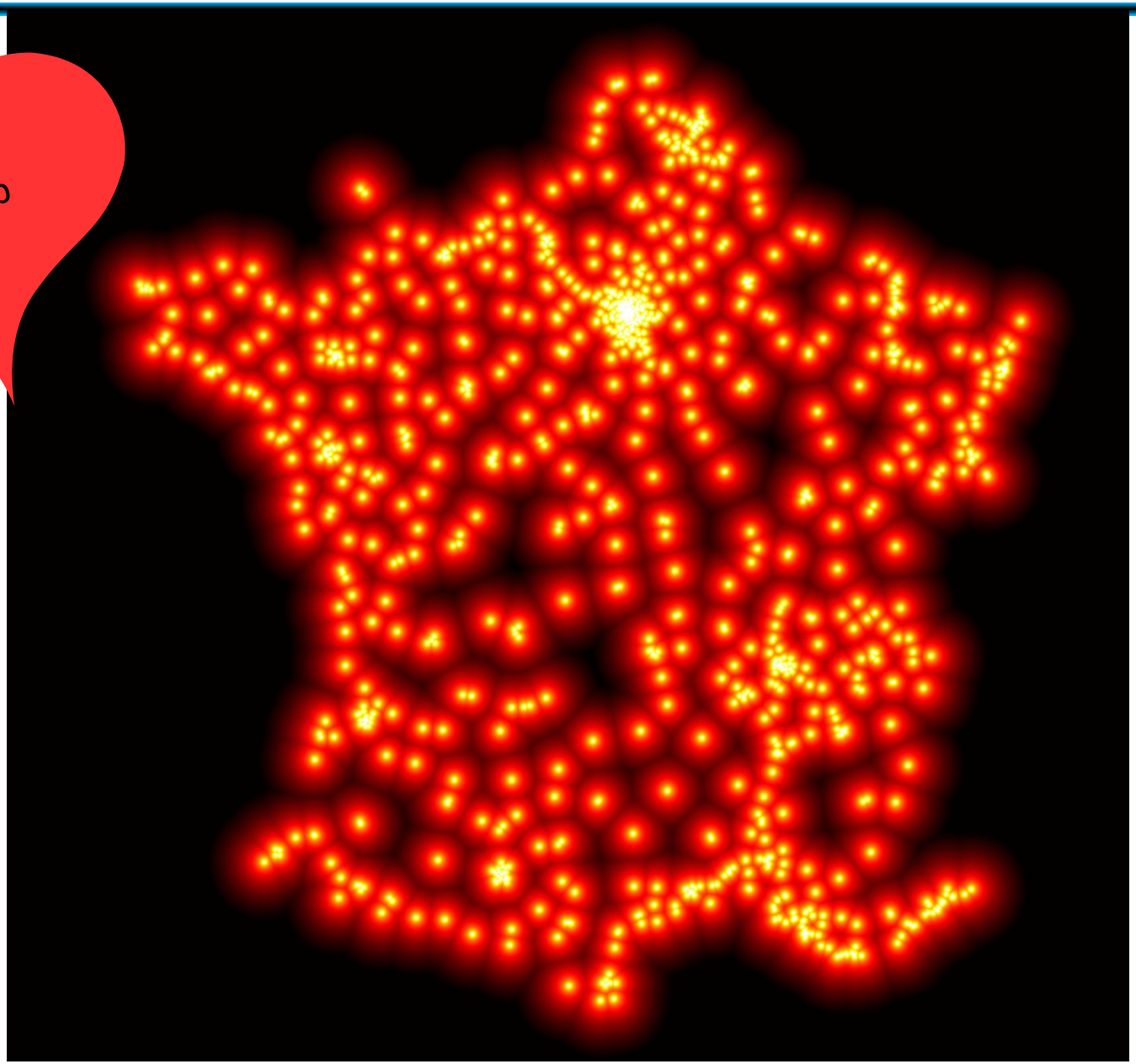
Diagramme de Voronoi des McDos parisiens



http://gambette.blogspot.fr/2006/10/mcdonalds-macdo-mac-donalds-et-vorono.html

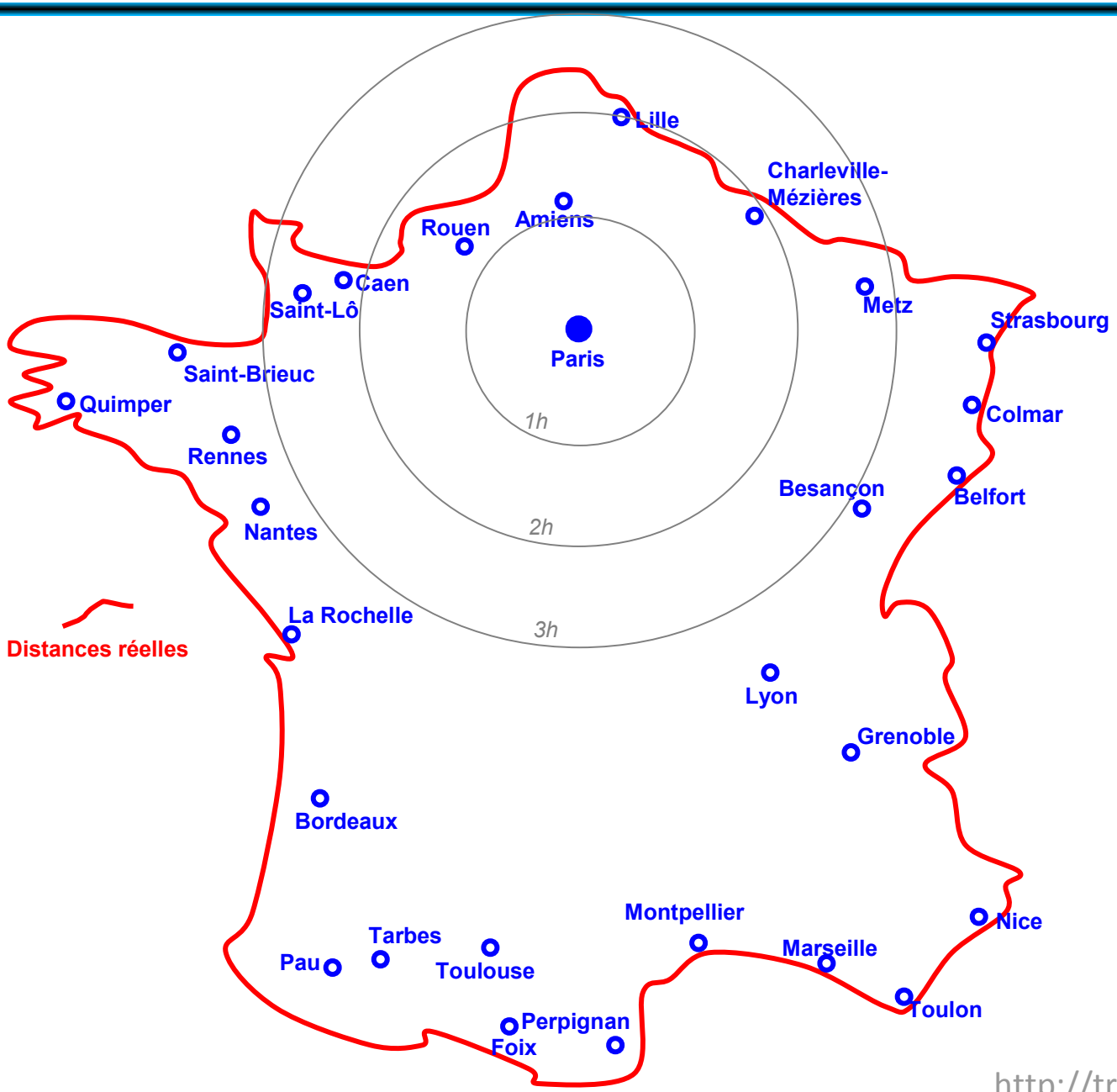
Diagramme de Voronoi des McDos français

Code Matlab

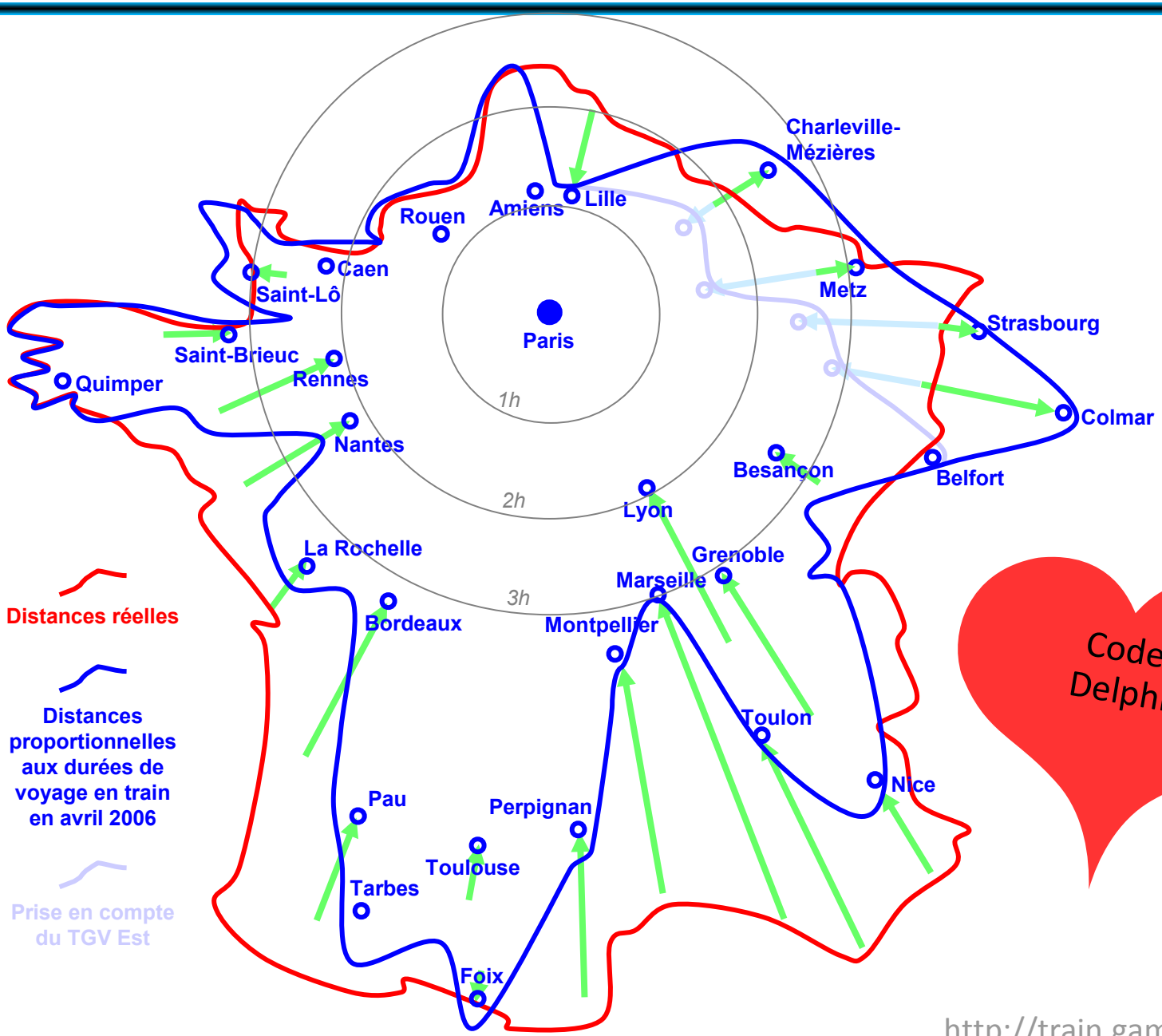


http://www.comeetie.fr/map_mcdofr.php?

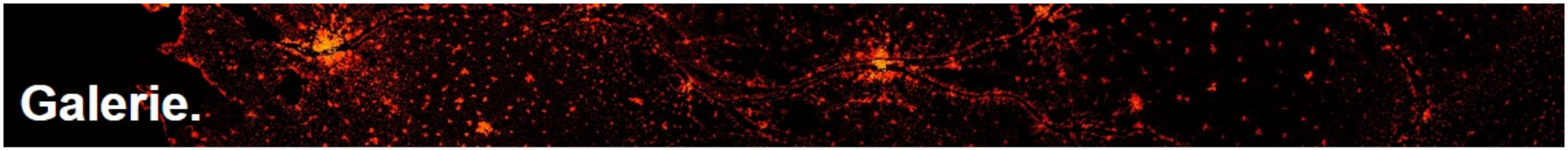
La France en train depuis Paris

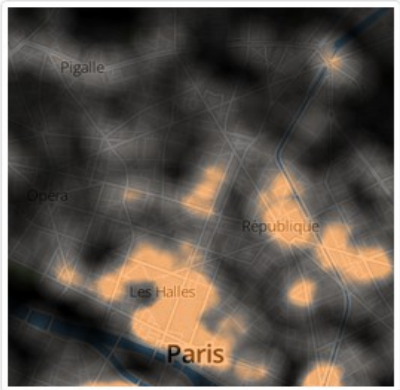


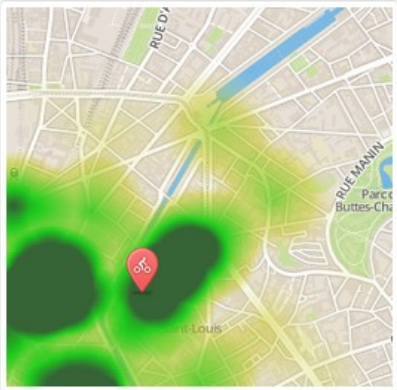
La France en train depuis Paris

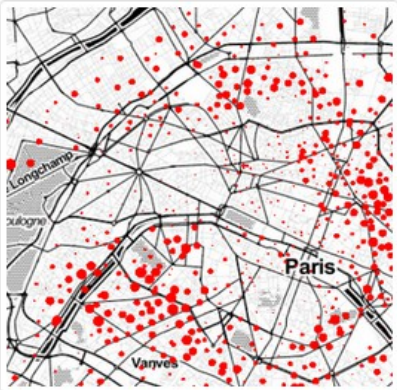


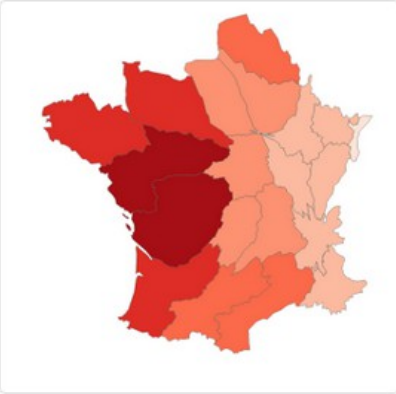
D’autres traitements de données géographiques











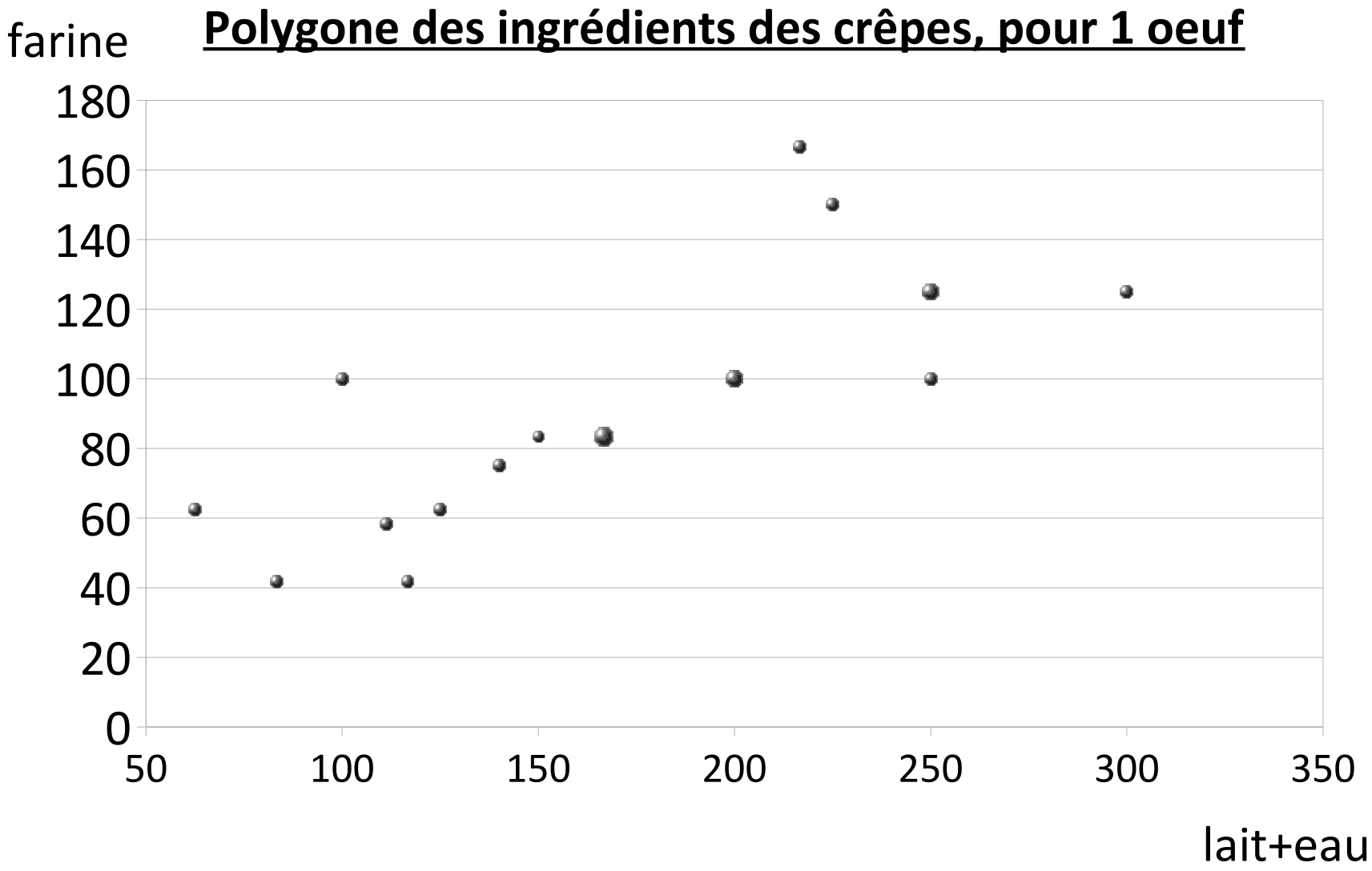
http://www.comeetie.fr/galerie.php?

Données gastronomiques

Visualisation de données de recettes de crêpes

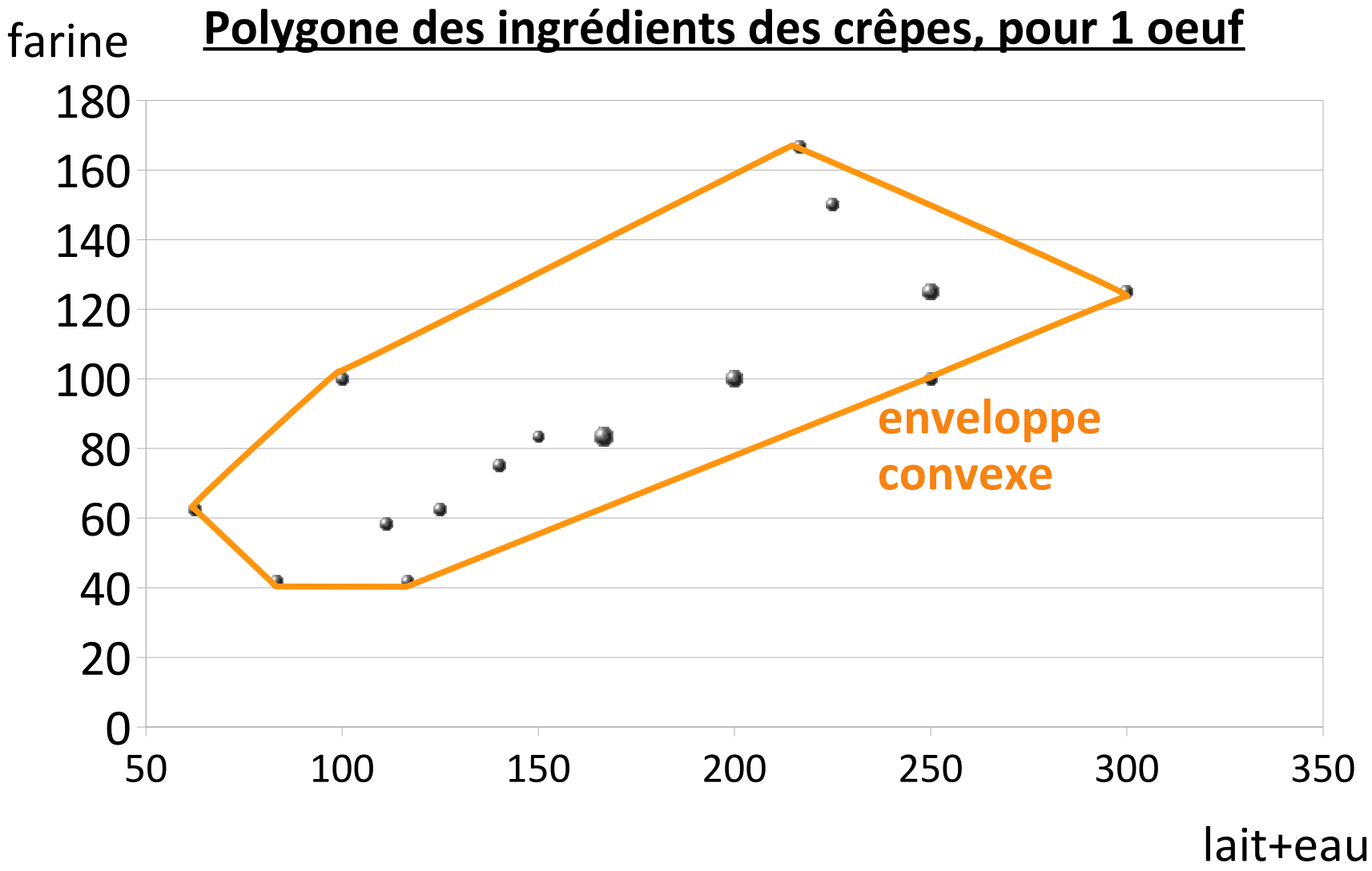
Photo Frédérique Voisin-Demery, https://flic.kr/p/dtUFN8

Visualisation de données de recettes de crêpes



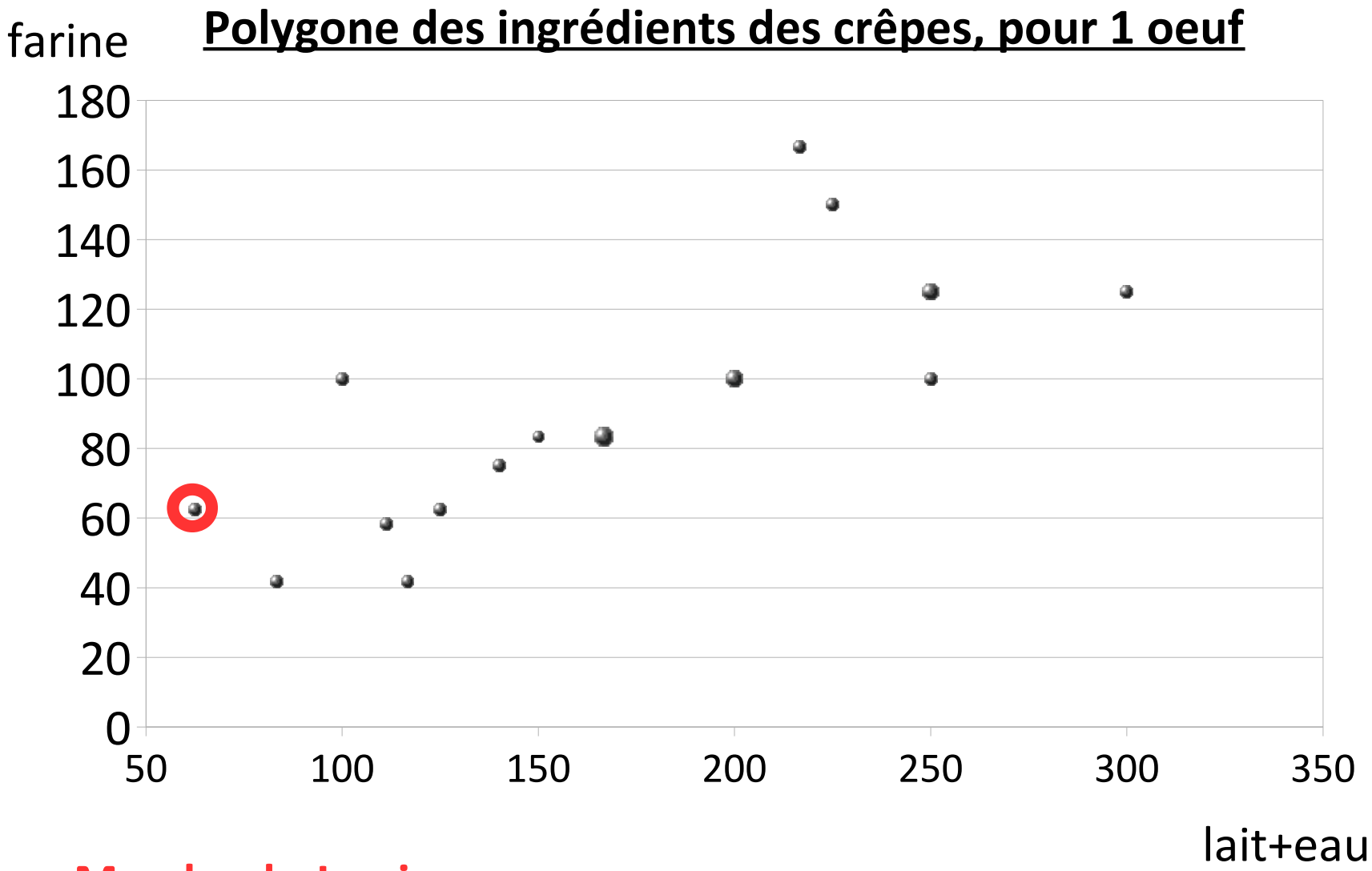
http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

Visualisation de données de recettes de crêpes



http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

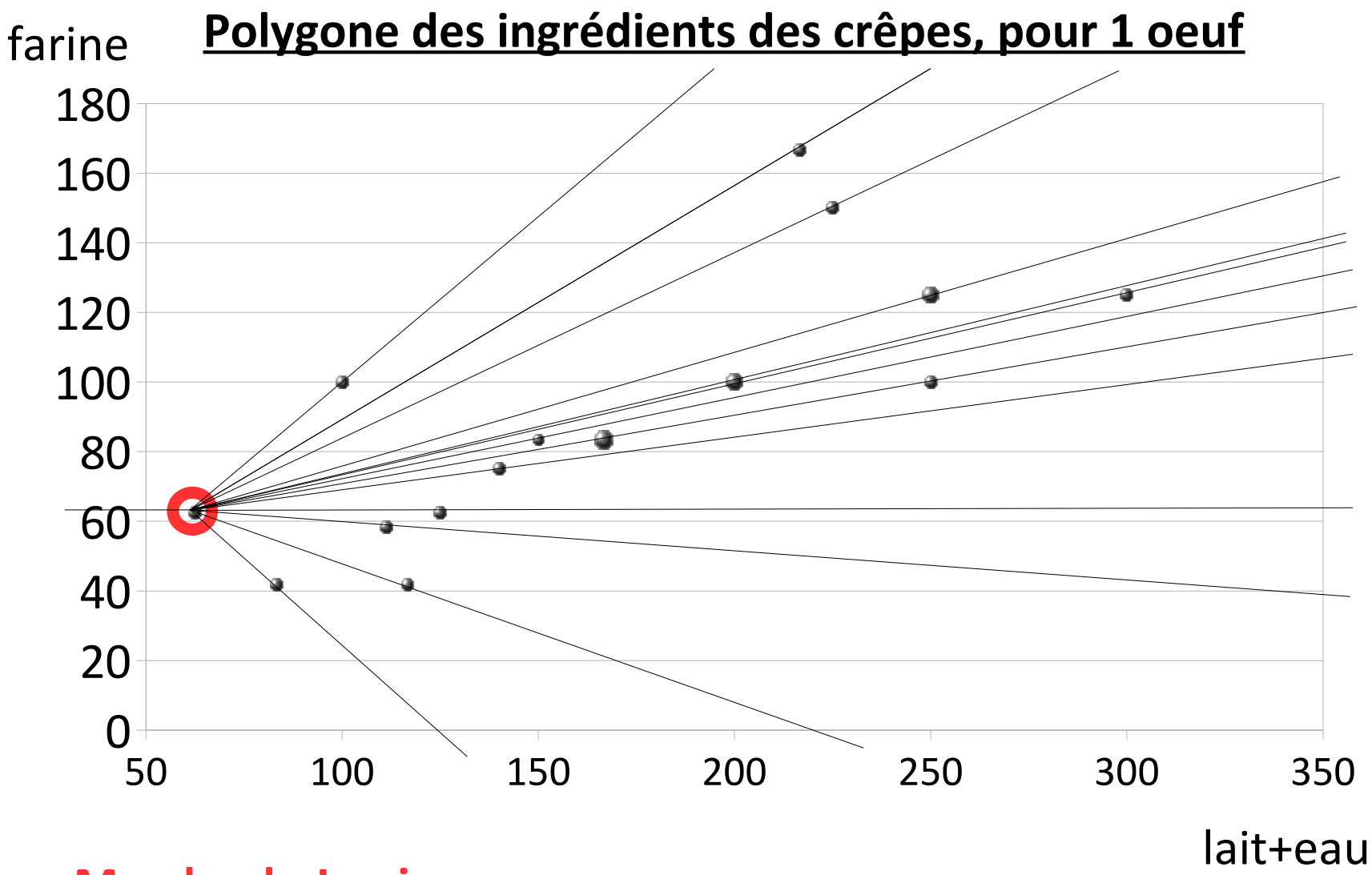
Visualisation de données de recettes de crêpes



Marche de Jarvis

http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

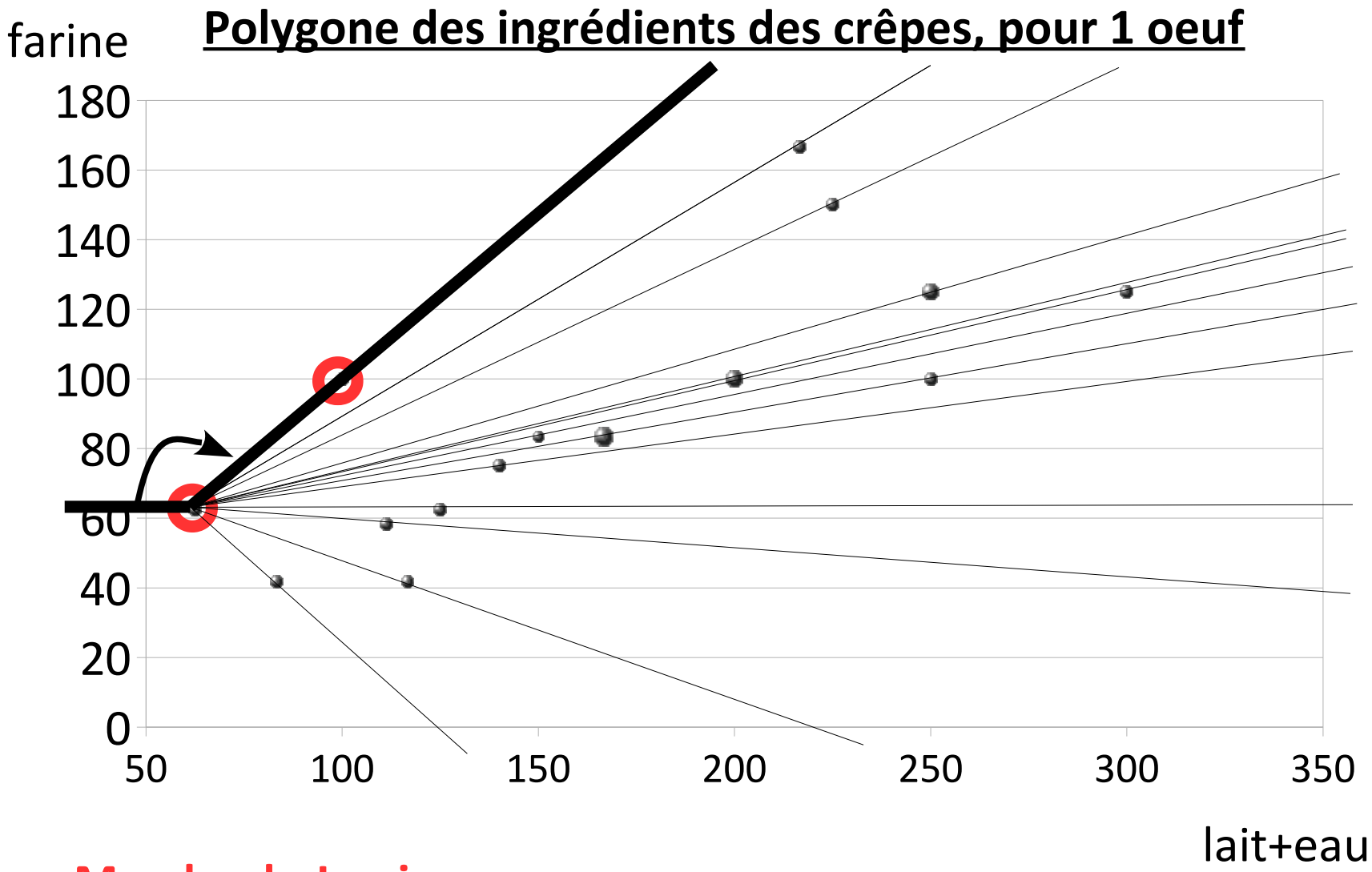
Visualisation de données de recettes de crêpes



Marche de Jarvis

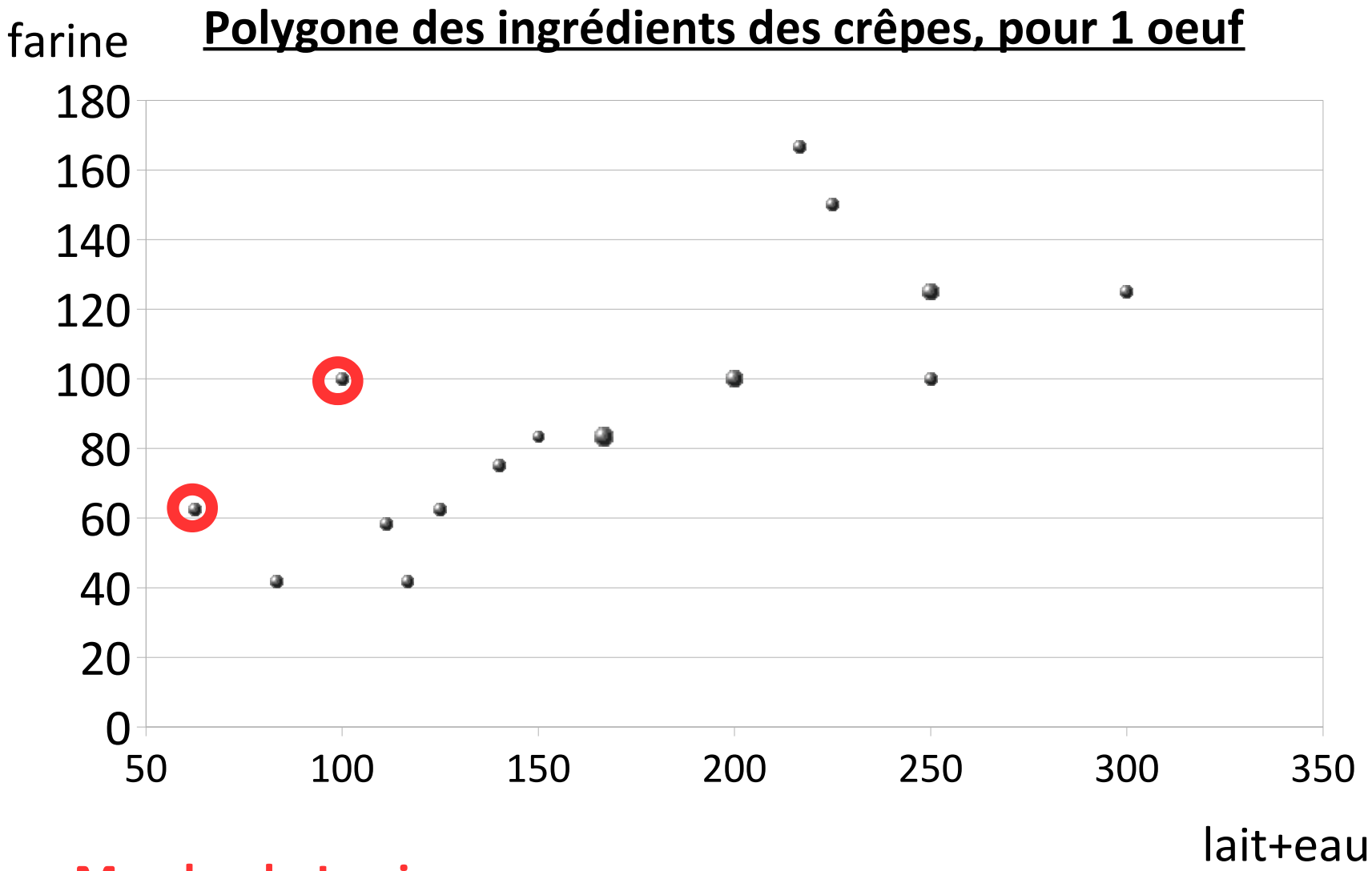
http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

Visualisation de données de recettes de crêpes



http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

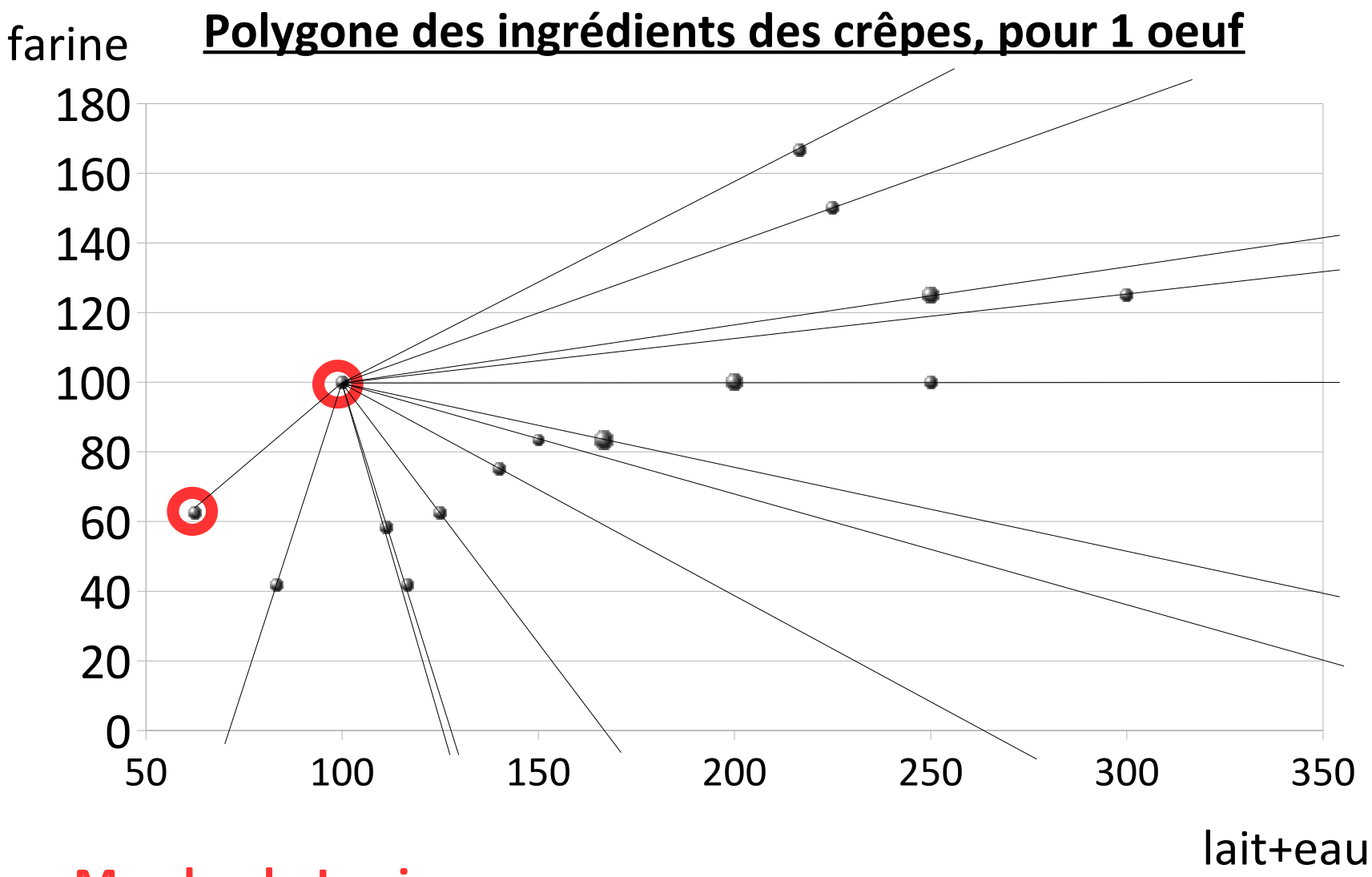
Visualisation de données de recettes de crêpes



Marche de Jarvis

http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

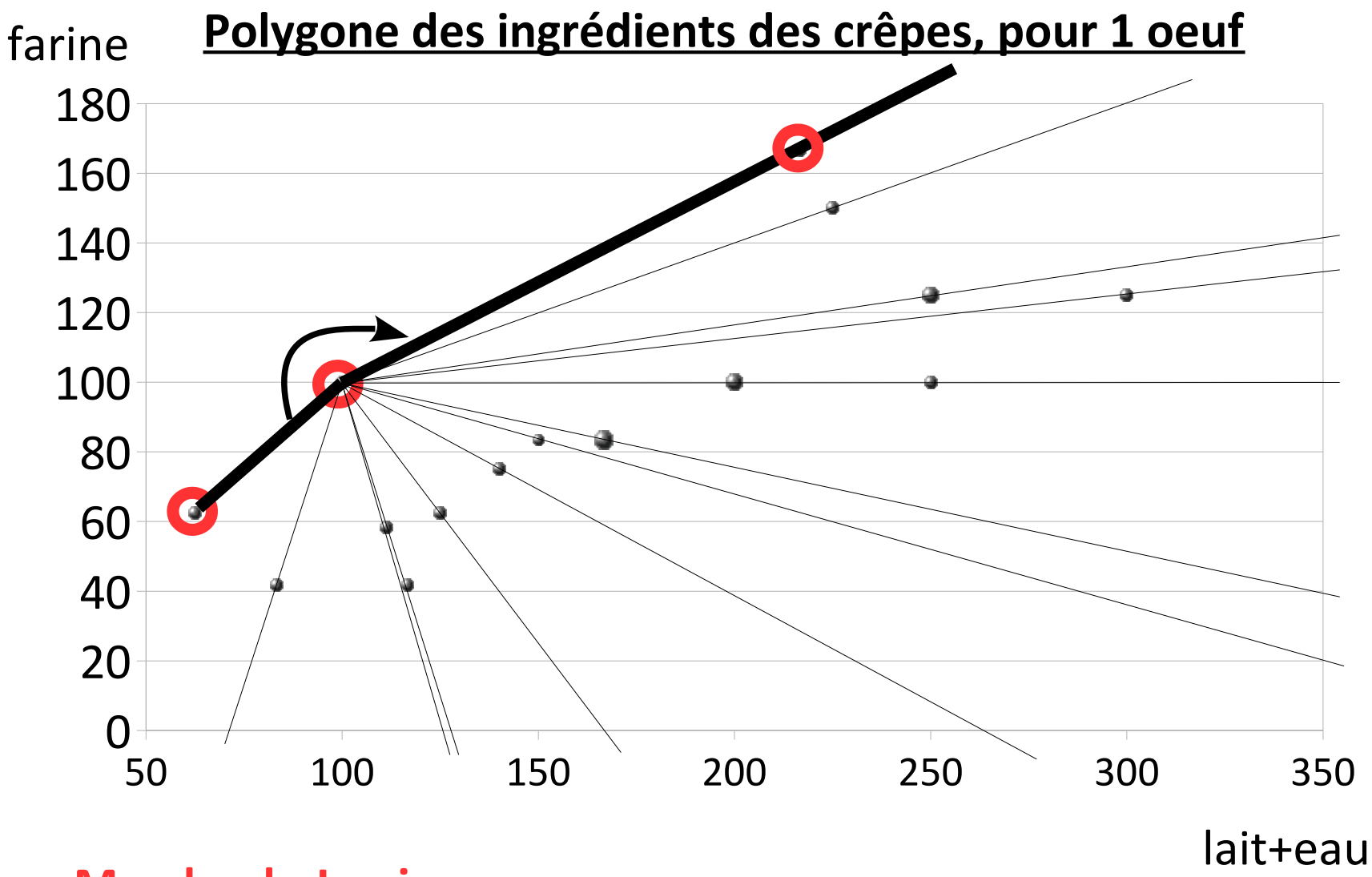
Visualisation de données de recettes de crêpes



Marche de Jarvis

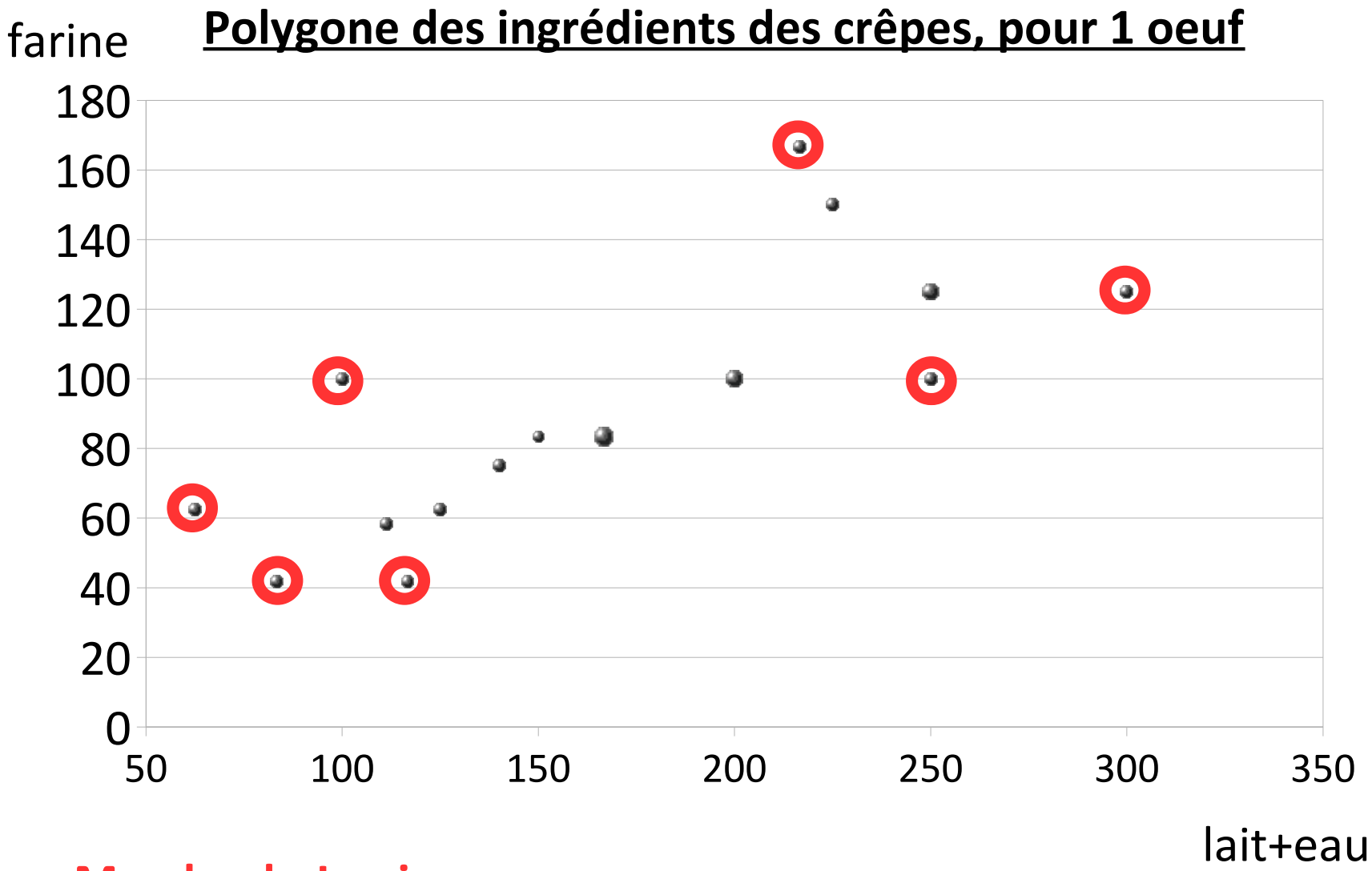
http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

Visualisation de données de recettes de crêpes



http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

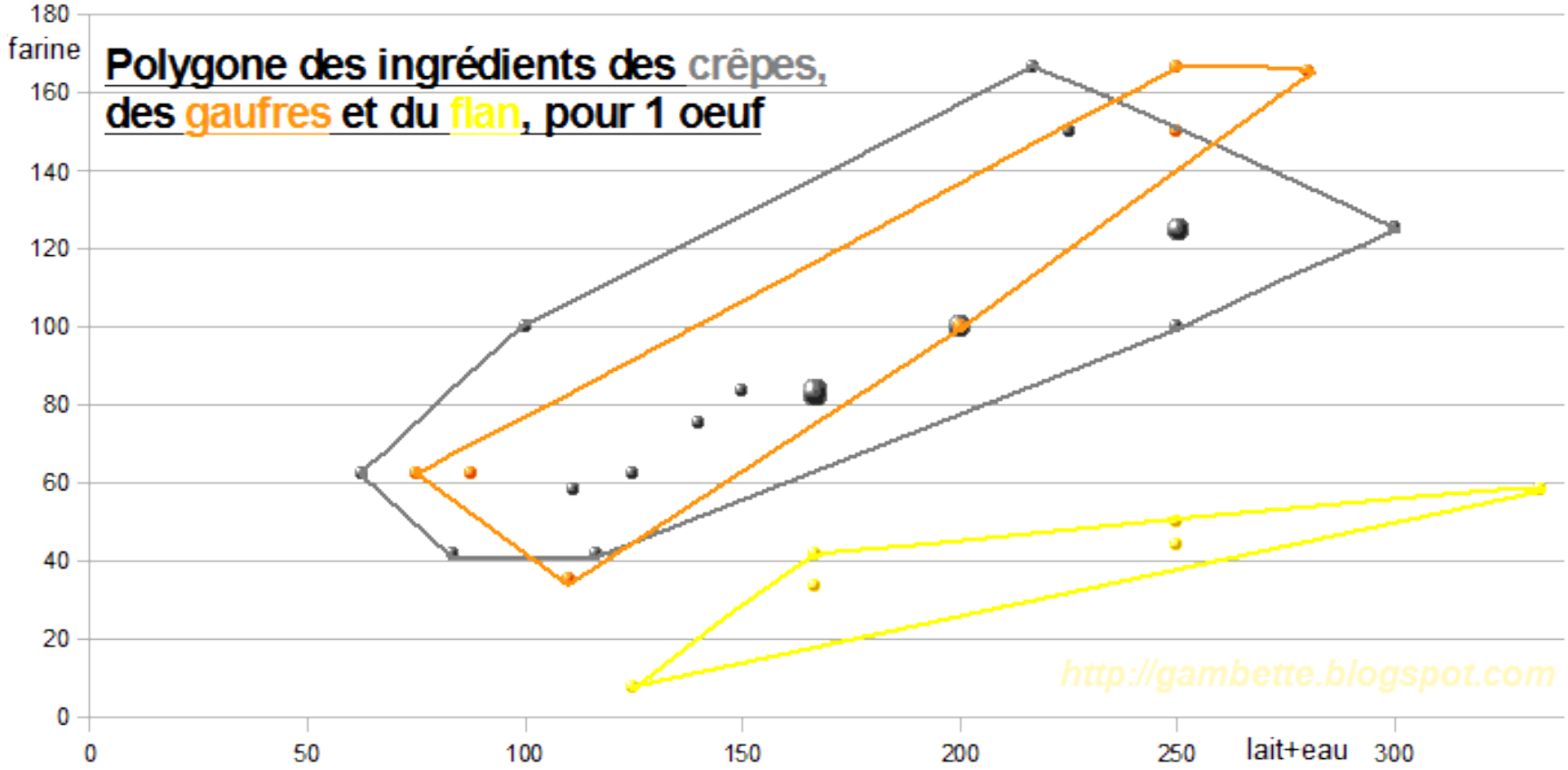
Visualisation de données de recettes de crêpes



Marche de Jarvis

http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

Visualisation de données de recettes de crêpes



http://gambette.blogspot.fr/2008/03/cuisine-polydre-des-ingrdients-et.html

Données de moteurs de recherche

Google Fight!

Suggestion de combats

Les 20 derniers combats

Select your version

Tapez 2 mots clés et cliquez sur le bouton 'Fight !'. Le gagnant est celui qui a la meilleure visibilité sur Google.

mathématiques

VS

informatique

FIGHT !

Tweet 45 Like 57 +1 28 Share 7 Pin it

COMBAT DU JOUR :

FRANCE FIGHT IRLANDE

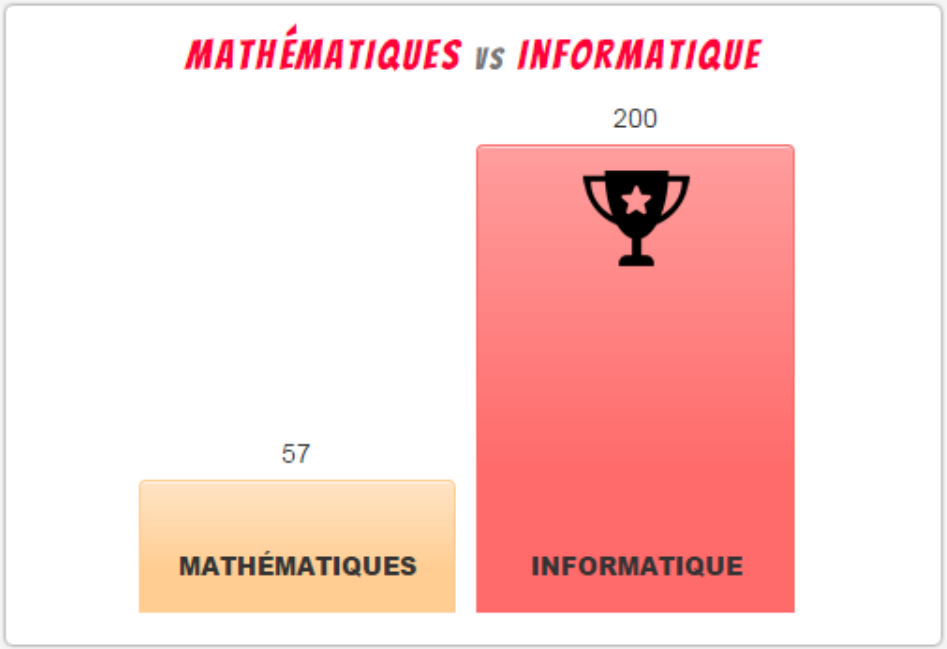
http://www.googlefight.fr/

Google Fight!

Suggestion de combats

Les 20 derniers combats

Select your version



Mode de calcul

Powered by semrush

Partagez ce combat :

Tweet Like +1 Share Pin it

ESSAYEZ AUSSI CES COMBATS

Roger Federer	FIGHT	Raphal nadal
Argent	FIGHT	Bonheur
Novak Djokovic	FIGHT	Raphael Nadal

http://www.googlefight.fr/mathematiques-vs-informatique.php

Google Fight!

Attention à la fiabilité !
http://blog.veronis.fr/2005/01/web-comptes-bidons-chez-google.html?m=0

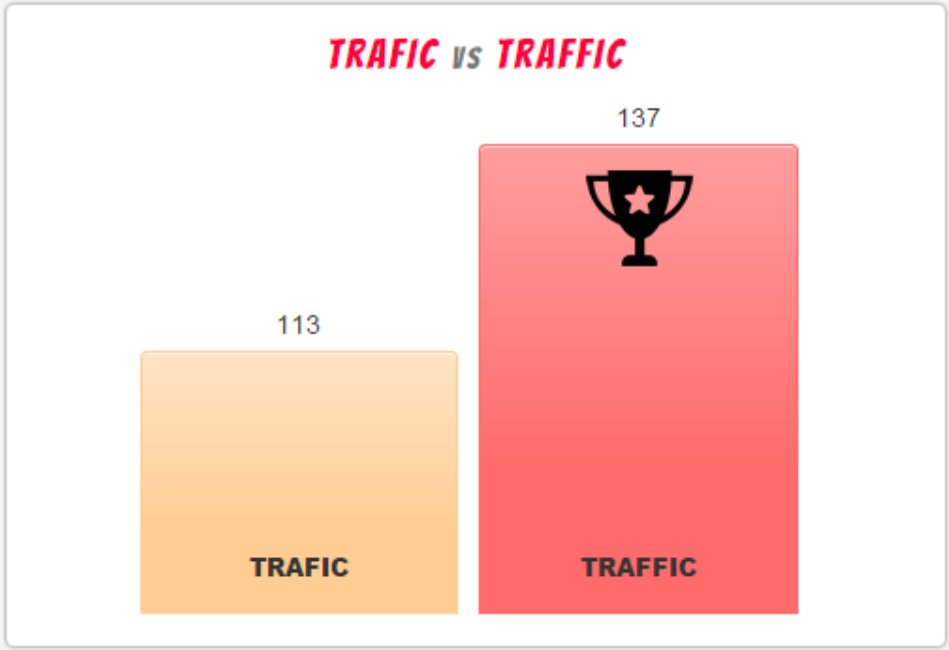
http://www.googlefight.fr/mathematiques-vs-informatique.php

Google Fight pour l'orthographe ?

Suggestion de combats

Les 20 derniers combats

Select your version



Mode de calcul

Powered by semrush

Partagez ce combat :

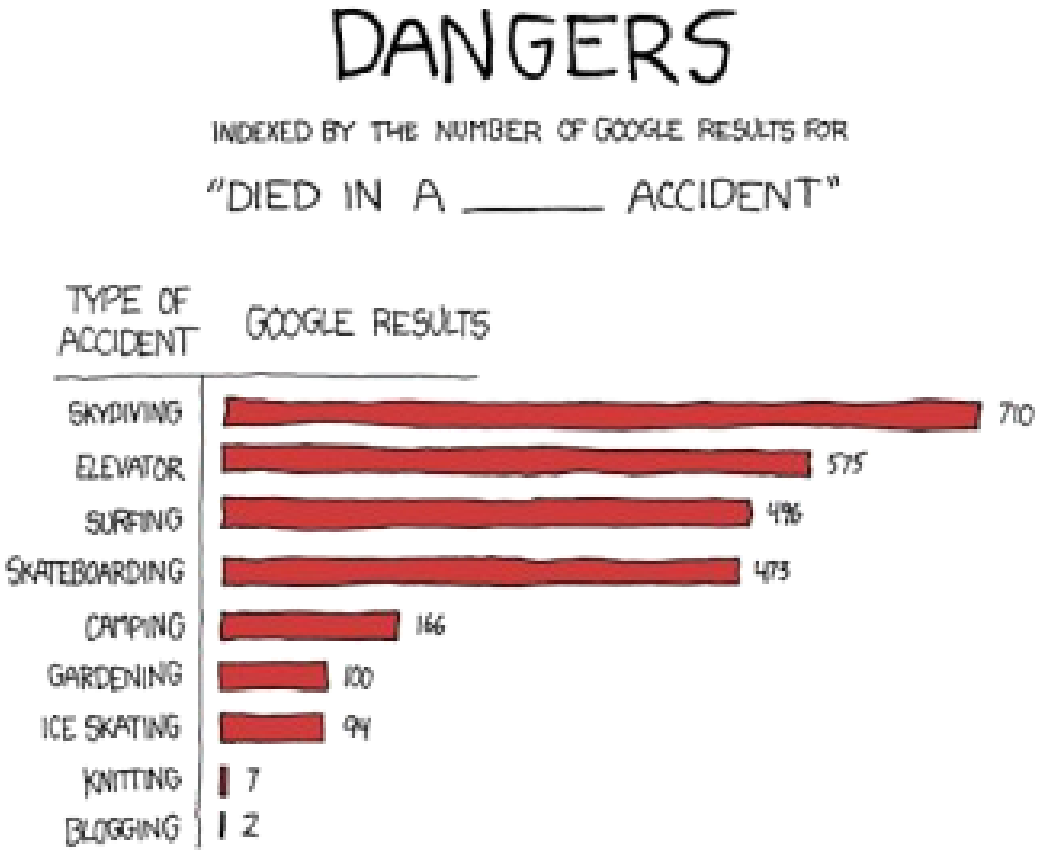
Tweet | Like | +1 | Share | Pin it

ESSAYEZ AUSSI CES COMBATS

Beyonce	FIGHT	Rihanna
Mac OS	FIGHT	Windows
Alexandrie	FIGHT	Alexandra

http://www.googlefight.fr/trafic-vs-traffic.php

Google Fights : dangers !



https://xkcd.com/369/

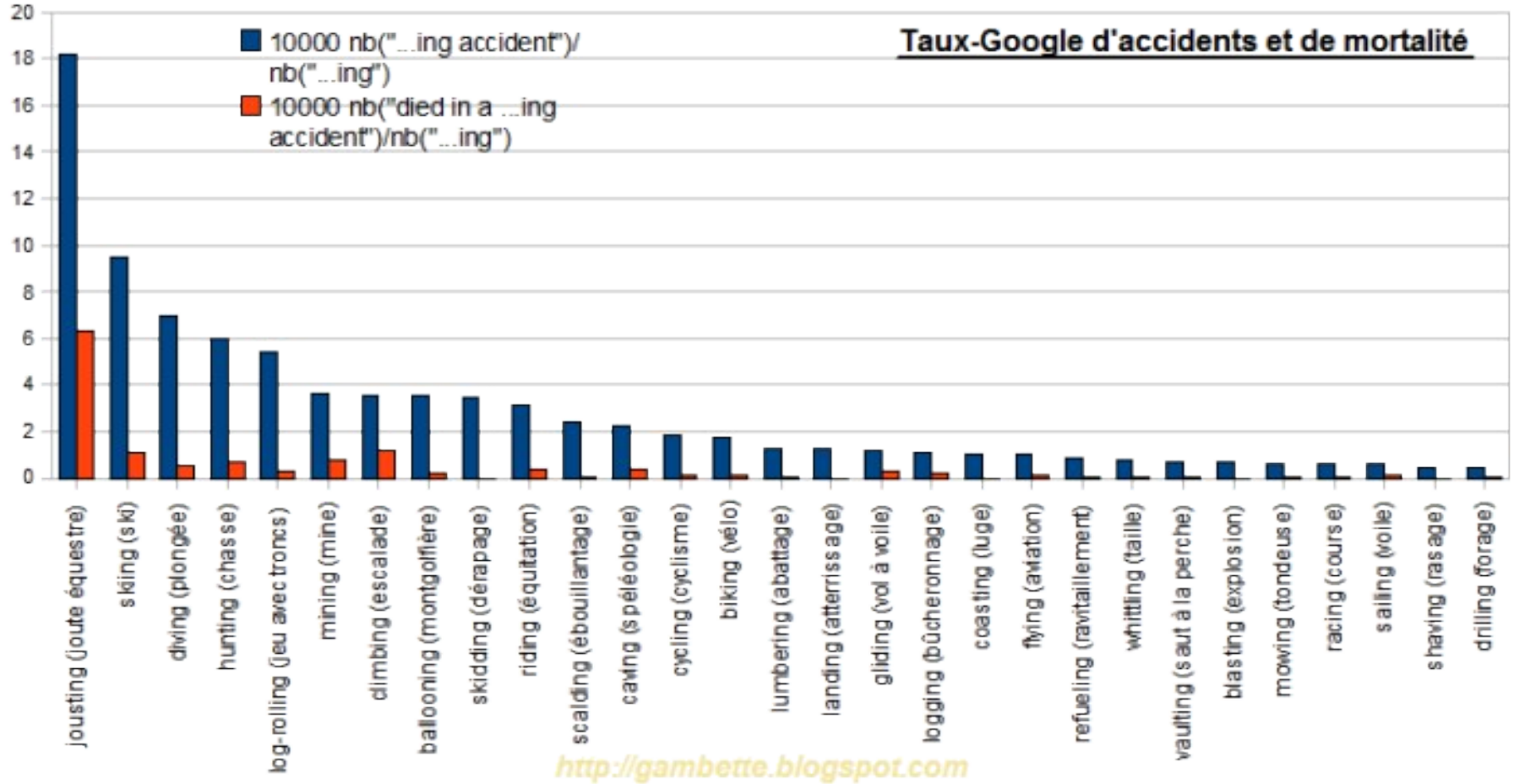
Google Fights : dangers !

http://gambette.blogspot.fr/2008/01/danger-accidents-mortels.html

Google Fights : dangers !

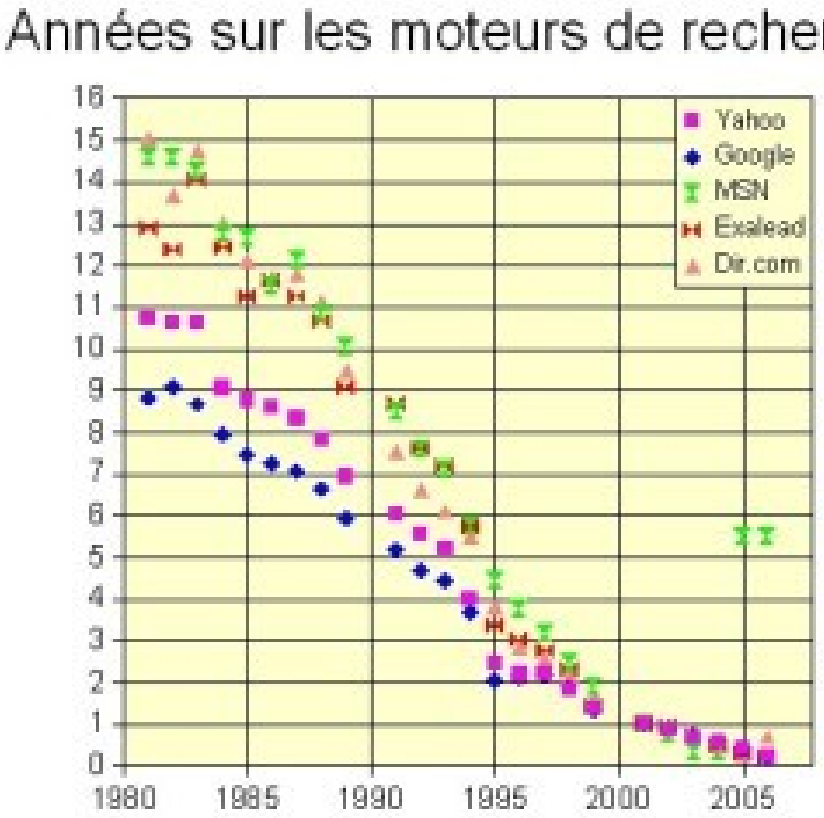
http://gambette.blogspot.fr/2008/01/danger-accidents-mortels.html

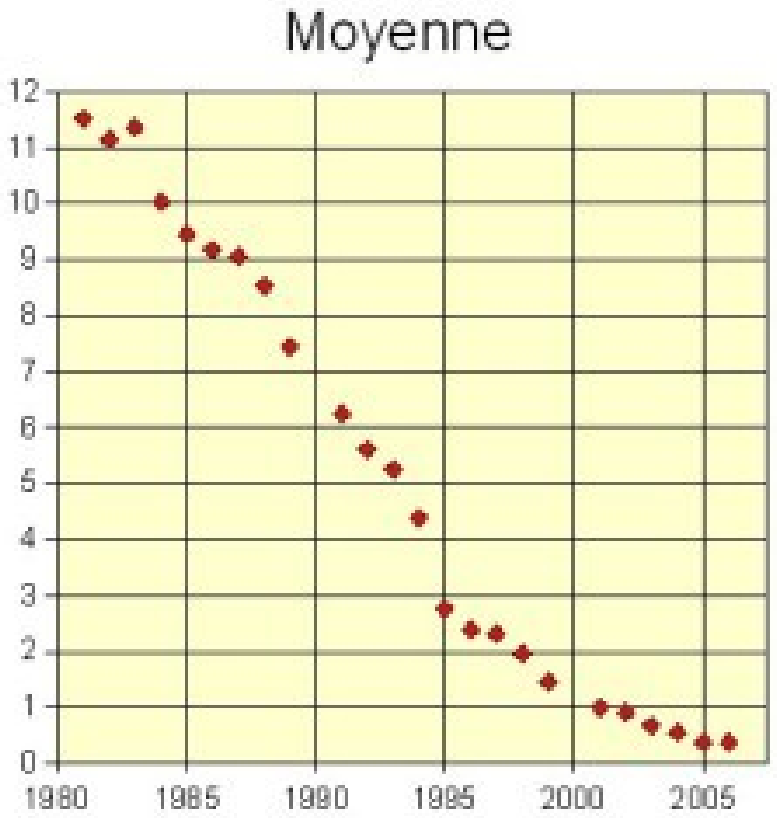
Google Fights : dangers !



http://gambette.blogspot.fr/2008/01/danger-accidents-mortels.html

Google Fights : années

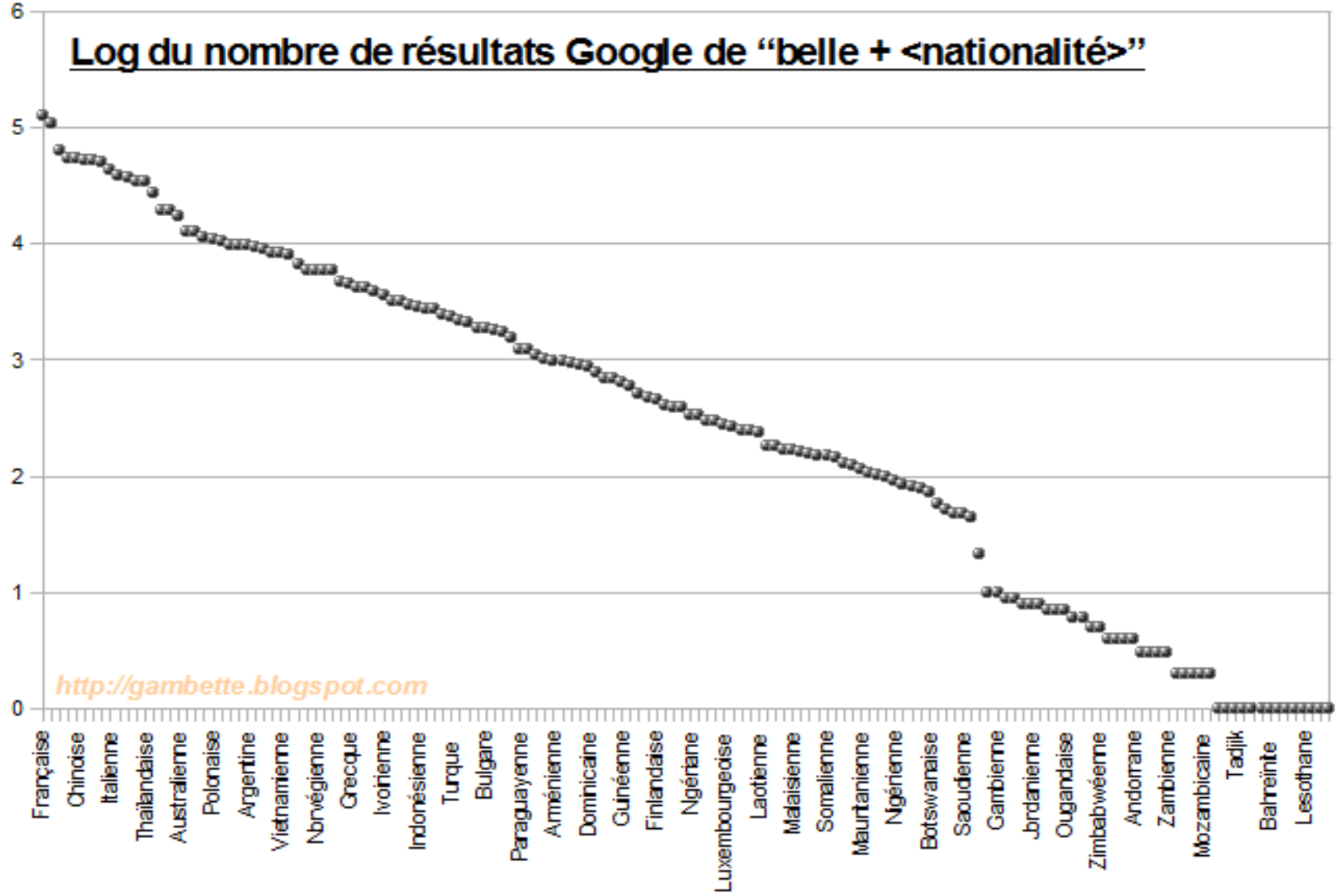




1 000 000 000/n(x), où n(x) est le nombre de résultats pour l'année x

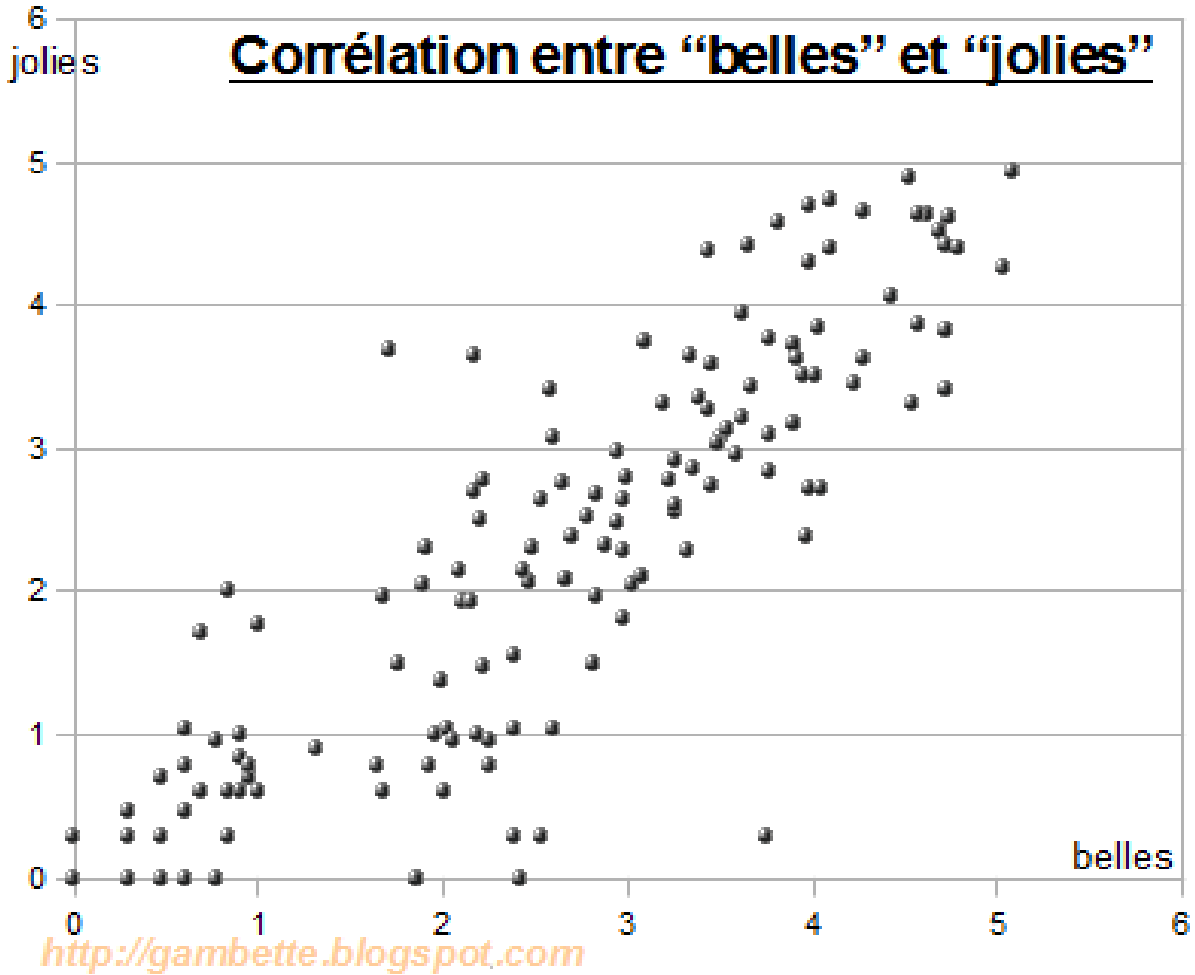
http://gambette.blogspot.fr/2006/11/la-naissance-du-web-daprs-les-moteurs.html

Google Fights : Miss Google 2010



http://gambette.blogspot.fr/2010/02/miss-google-2010.html

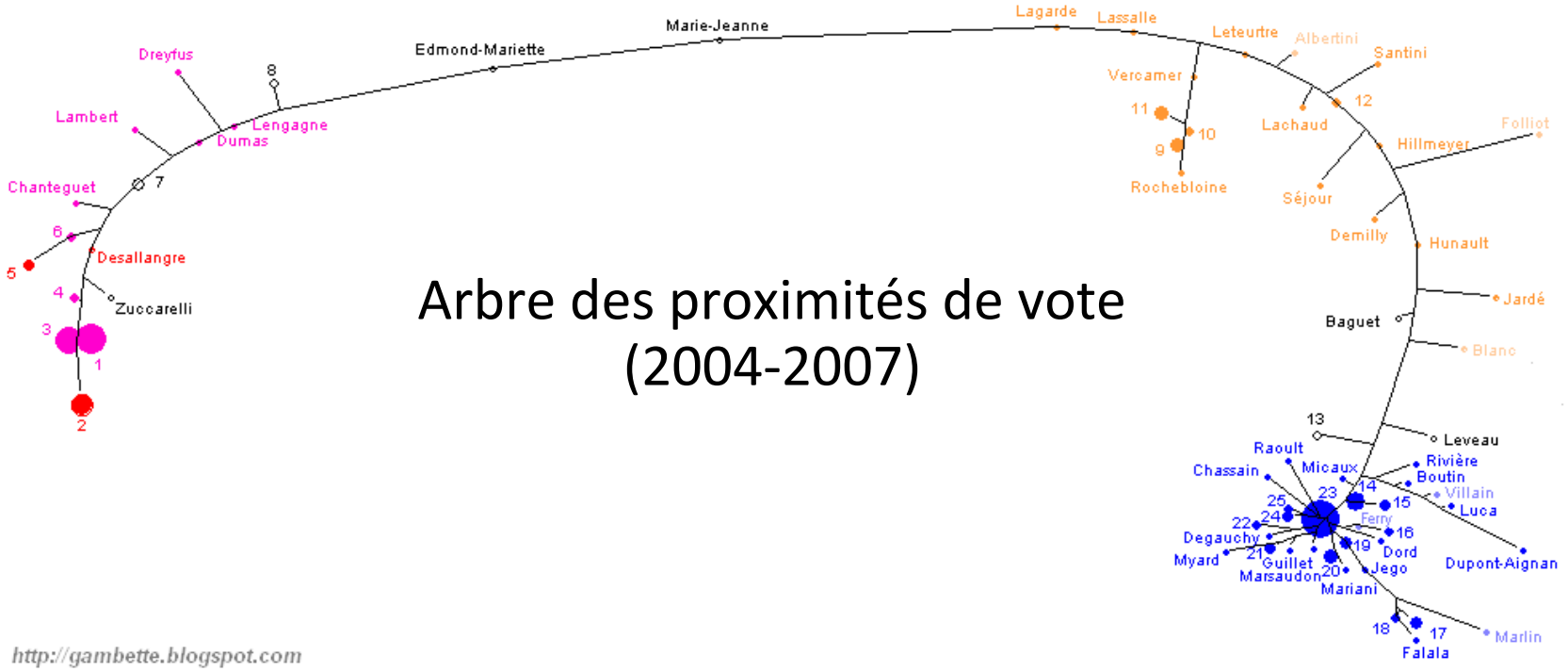
Google Fights : Miss Google 2010



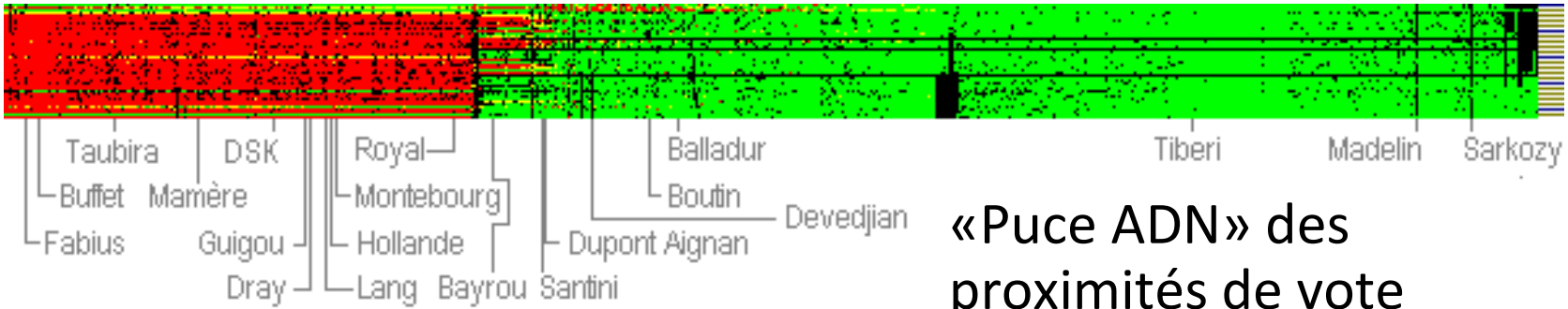
http://gambette.blogspot.fr/2010/02/miss-google-2010.html

Données en arbres

Vote des députés



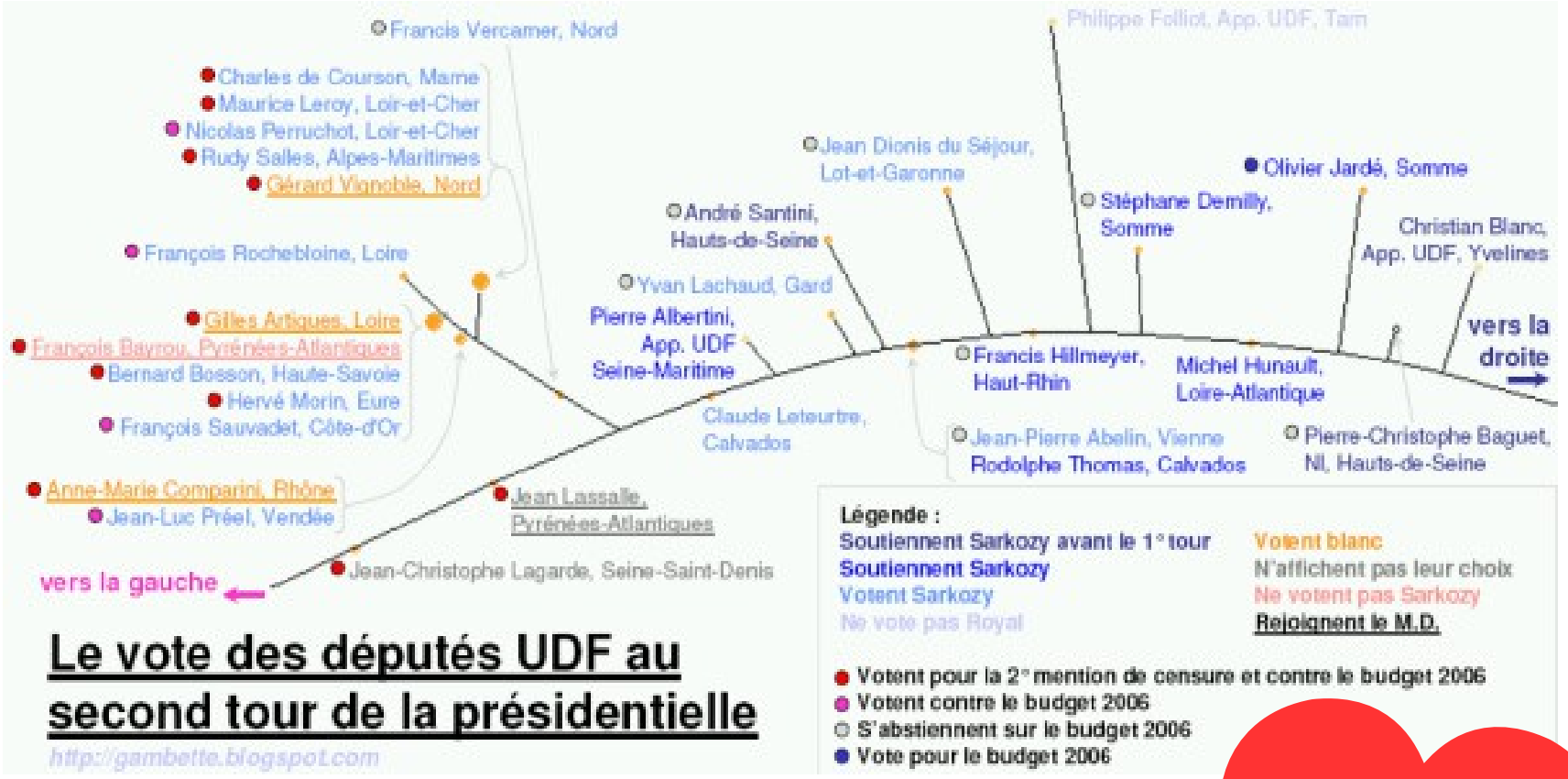
Arbre des proximités de vote (2004-2007)



«Puce ADN» des proximités de vote

http://gambette.blogspot.fr/2007/01/arbre-phylogntique-des-dputs.html
http://gambette.blogspot.fr/2007/02/la-puce-adn-des-dputs.html

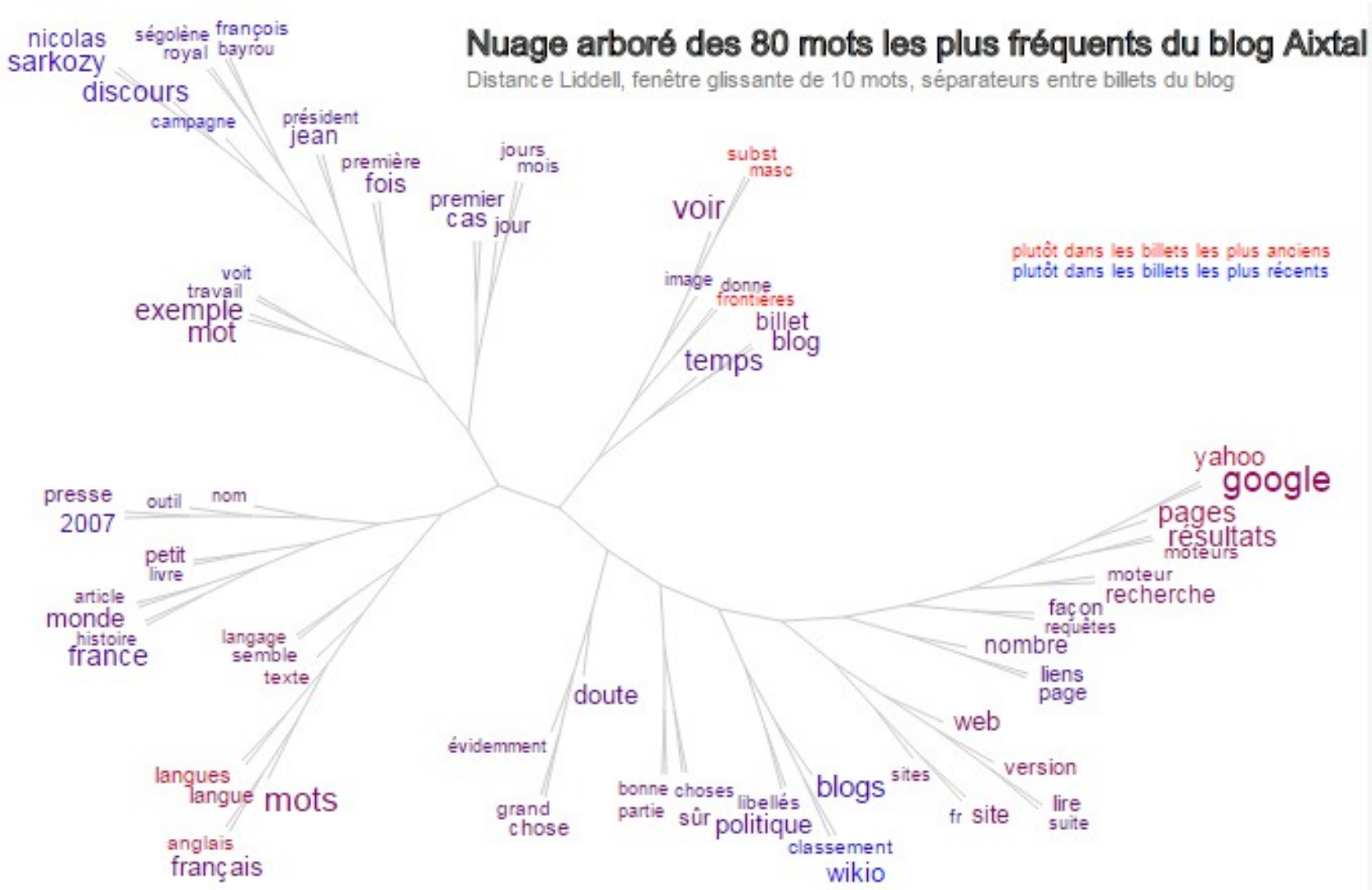
Vote des députés



Code Delphi, code Python

http://gambette.blogspot.fr/2007/05/positionnement-des-dputs-udf.html

Nuage arboré du blog de Jean Véronis



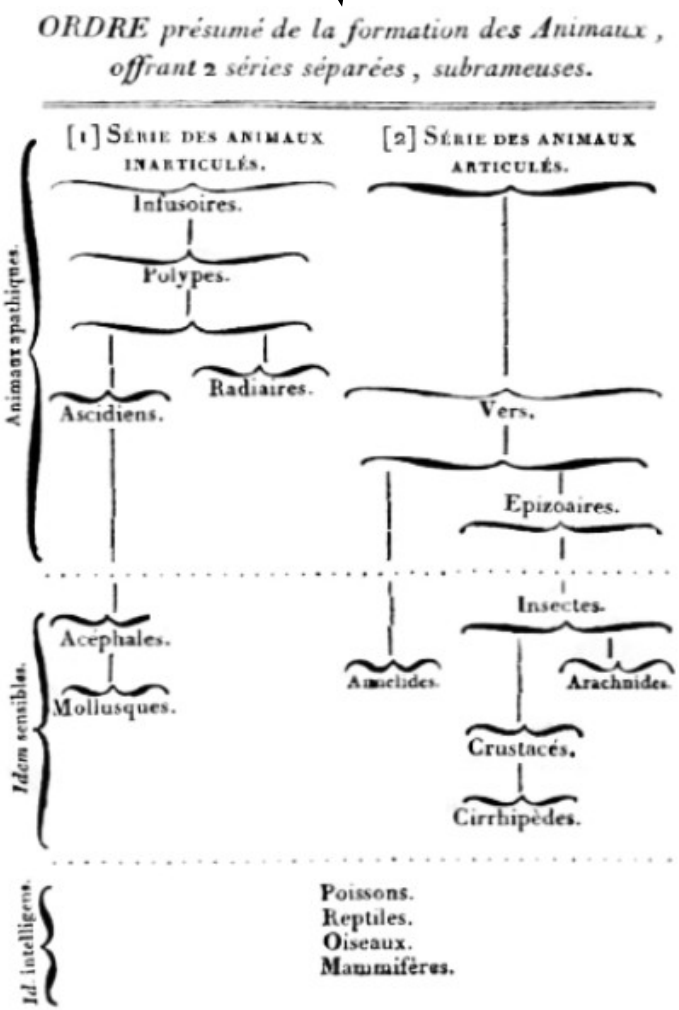
http://gambette.blogspot.fr/2013/10/hommage-en-nuage.html

Arbres phylogénétiques et arbres de mots

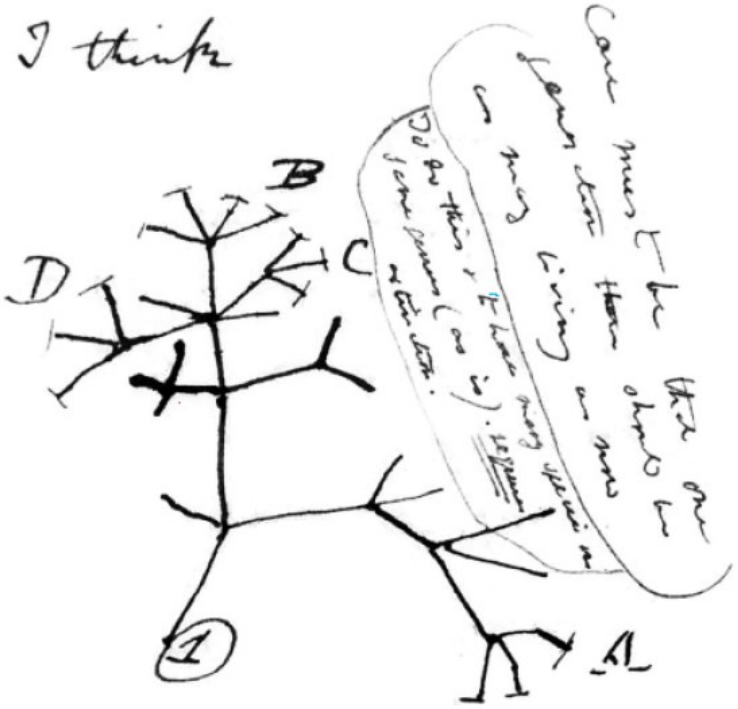
Arbre phylogénétique d'un ensemble d'espèces :

- Les **classer** en fonction de caractères communs
- Décrire leur **évolution**

D'après Lamarck (1815) *Histoire naturelle des animaux sans vertèbres*



Darwin (1837) *Carnet B*



Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Données sur les feuilles

MOTS

Position des mots

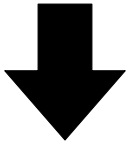
Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles



Distances entre les feuilles

	A	B	C	D
A	0	2	5	6
B	2	0	5	6
C	5	5	0	3
D	6	6	3	0

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

Distances entre les feuilles

	A	B	C	D
A	0	**2**	5	6
B	**2**	0	5	6
C	5	5	0	3
D	6	6	3	0

classification hiérarchique ascendante

Arbre

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

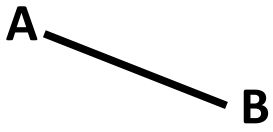
Données sur les feuilles

Distances entre les feuilles

	A+B	C	D
A+B	0	**5**	**6**
C	**5**	0	3
D	**6**	3	0

classification hiérarchique ascendante

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

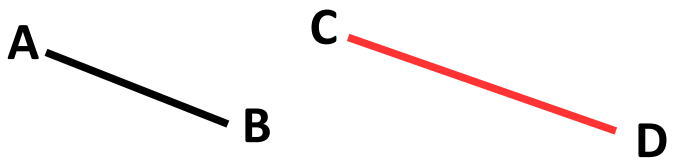
Données sur les feuilles

Distances entre les feuilles

	A+B	C	D
A+B	0	5	6
C	5	0	**3**
D	6	**3**	0

classification hiérarchique ascendante

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

Distances entre les feuilles

	A+B	C+D
A+B	0	**5,5**
C+D	**5,5**	0

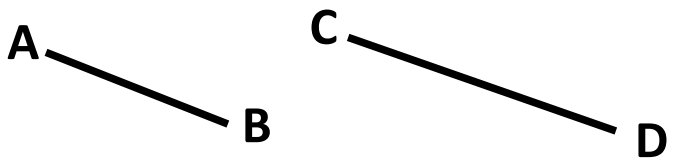
MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

classification hiérarchique ascendante

Arbre



Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles

Distances entre les feuilles

	A+B	C+D
A+B	0	**5,5**
C+D	**5,5**	0

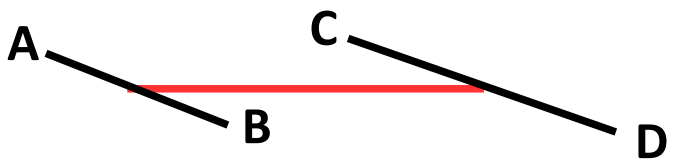
MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

classification hiérarchique ascendante

Arbre



Arbres phylogénétiques et arbres de mots

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

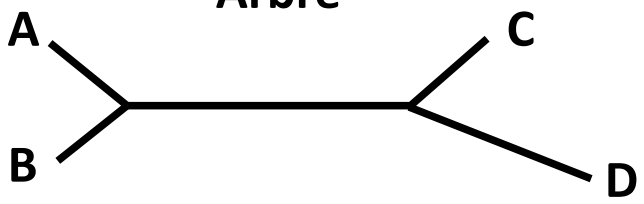
Données sur les feuilles

Distances entre les feuilles

	A	B	C	D
A	0	2	5	6
B	2	0	5	6
C	5	5	0	3
D	6	6	3	0

classification hiérarchique ascendante

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

Outils pratiques

Quelques outils pratiques

- extension **iMacros** de Firefox

Pour récupérer un ensemble de pages web

- **expressions régulières**

Pour extraire de l'information ou la changer de format

Dans la fonction rechercher/remplacer d'un éditeur de texte ou dans un script Python

- **bibliothèques Javascript** D3.js, Google Charts, Charts.js, etc.

Pour visualiser les données de manière interactive sur le web
http://www.sitepoint.com/15-best-javascript-charting-libraries/

Quelques langages utiles

- **R** : orienté statistiques

  - https://www.r-project.org/
  - http://r4ds.had.co.nz/ (R for data science)

- **Javascript** : orienté web (interactions avec l'utilisateur)

  - http://www.w3schools.com/js/

- **Python** : pour des scripts de test rapide en particulier

  - https://www.python.org/

- **Java** : pour des outils en production

  - https://www.java.com/fr/

Les expressions régulières selon xkcd

https://xkcd.com/208/

Bibliothèque Javascript d3.js

Data-Driven Documents

http://d3js.org/

Pour continuer à jouer avec les données

- Data Job 2016 – jeudi 10 novembre 2016 à Paris :
http://datajob.fr/
(entrée gratuite pour étudiants moins de 28 ans)

- Hackathons à Paris :
https://www.eventbrite.fr/d/france--paris/hackathon/

- Blog *Je véronise* :
http://gambette.blogspot.com/

- Les interventions d'Henri Verdier sur l'open data :
https://www.youtube.com/results?search_query=Henri+Verdier

- Actualités de la révolution des données :
http://radar.oreilly.com/data