Fouille de données et aide à la décision.

Introduction au datamining.

Anne-Claire Haury

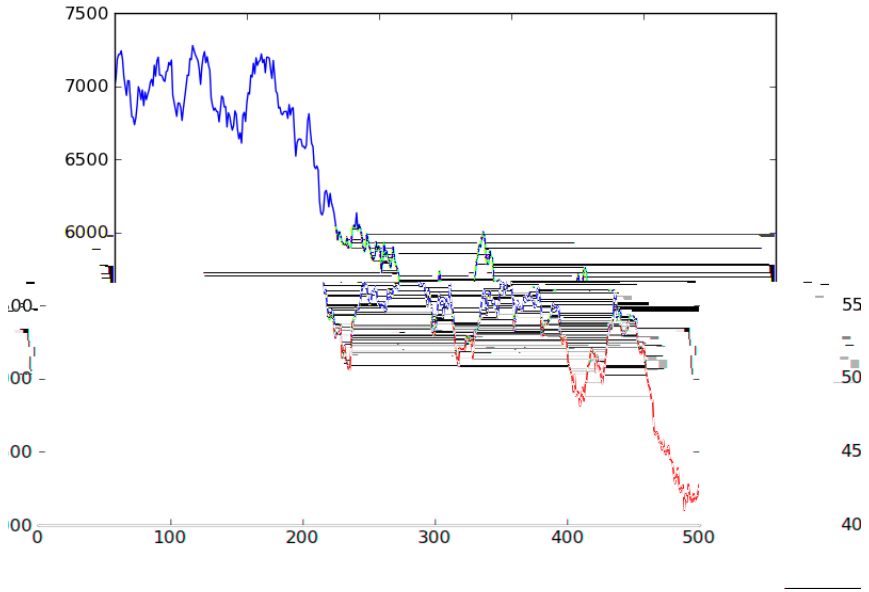
M2 Informatique
Université Denis Diderot

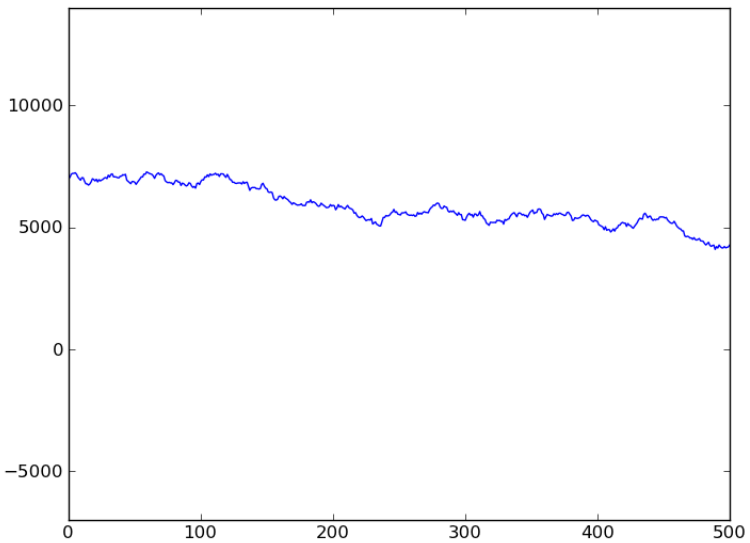
Second semestre 2014-2015

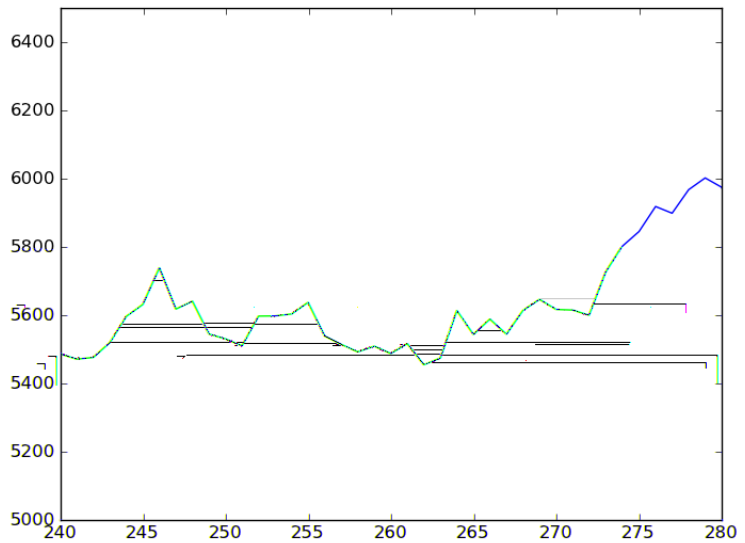
Introduction

Outline

Illusion d'optique

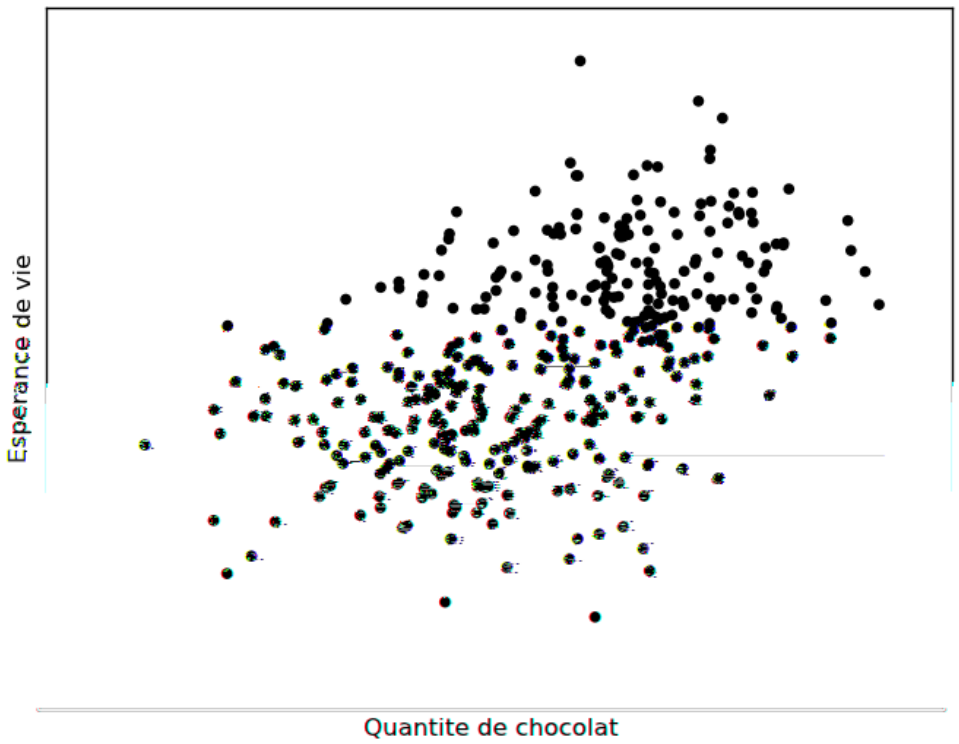






Chocolat et espérance de vie

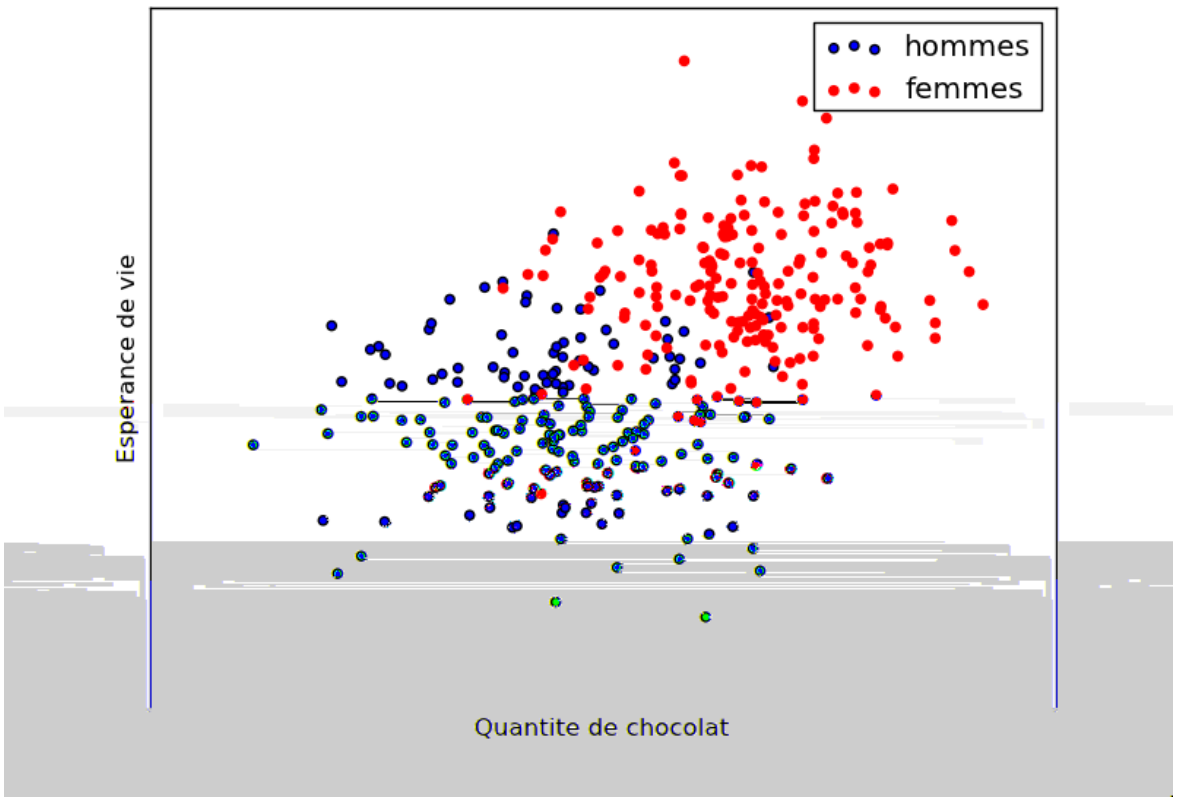
Exemple emprunté à Isabelle Guyon.



Manger du chocolat augmente l'espérance de vie.

Chocolat et espérance de vie

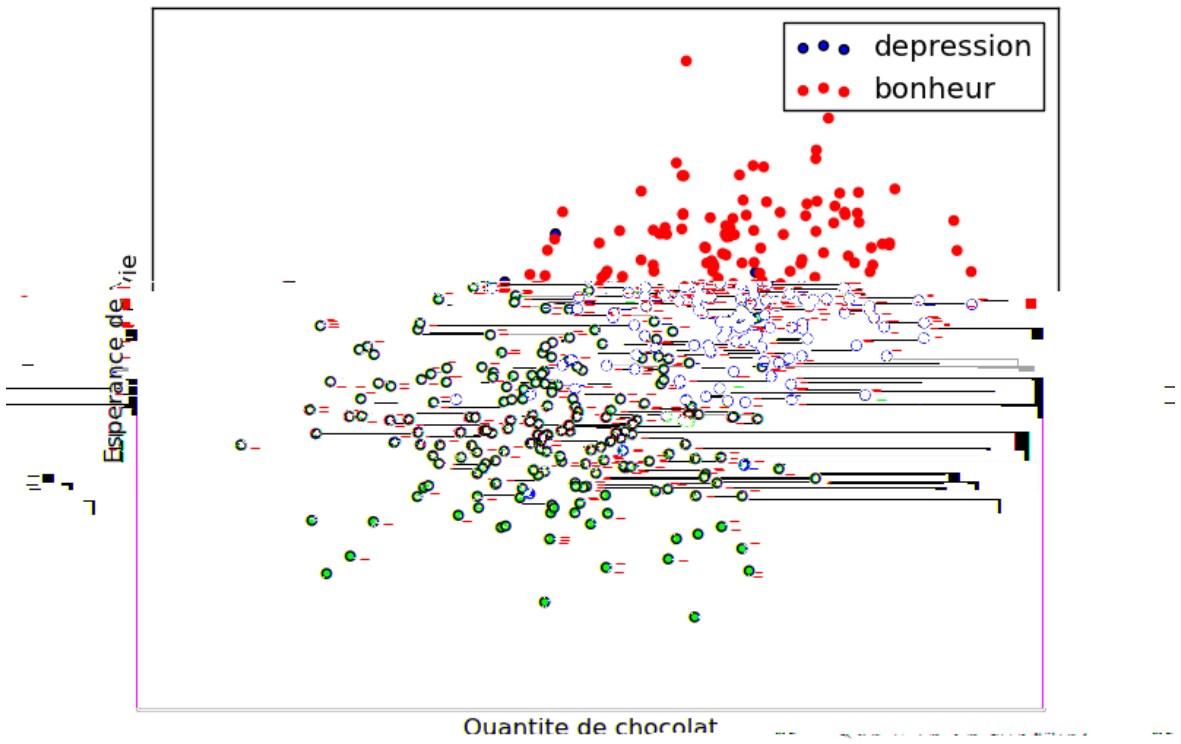
Exemple emprunté à Isabelle Guyon.



Manger du chocolat n'augmente pas l'espérance de vie.

Chocolat et espérance de vie

Exemple emprunté à Isabelle Guyon.



Manger du chocolat augmente peut-être l'espérance de vie.

Les expériences de Rhine

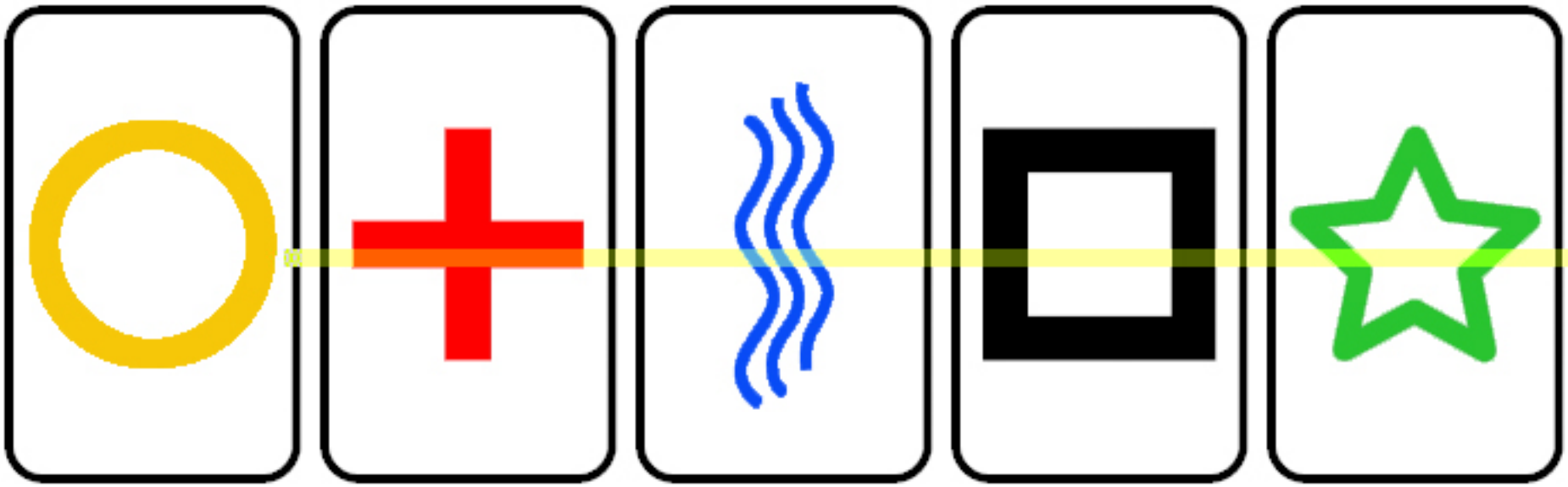


Figure de Wikipédia.

Les expériences de Rhine

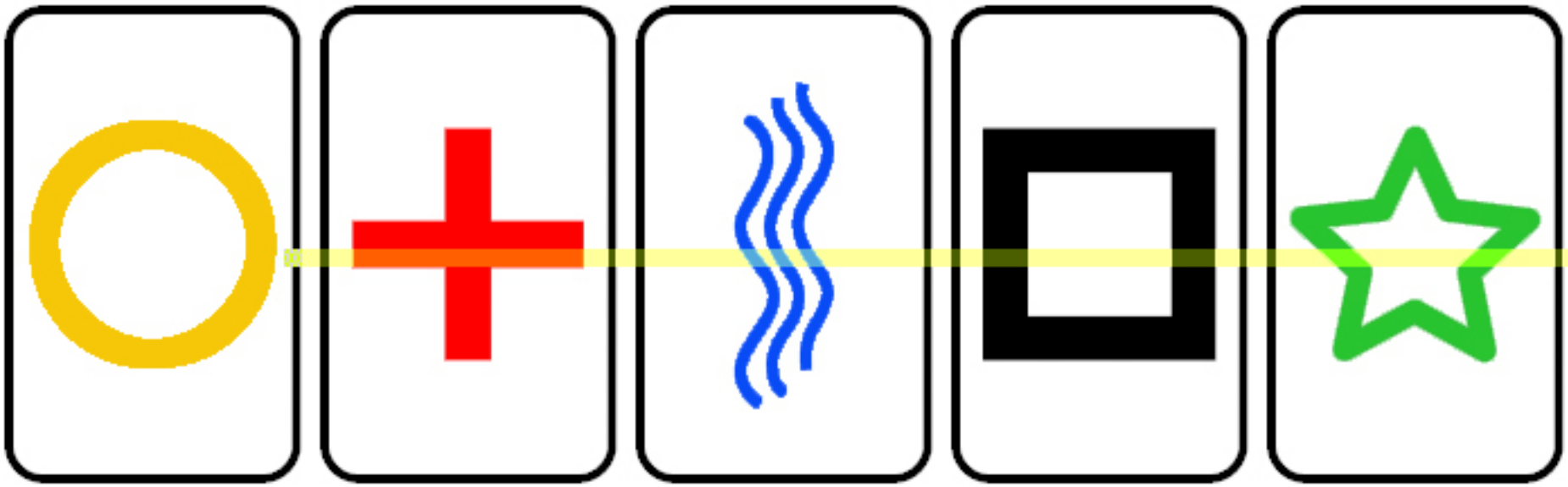


Figure de Wikipédia.

Conclusion de Rhine: lorsque les gens savent qu'ils ont des dons extra-sensoriels, les dons disparaissent...

Pile ou face ?

Pile ou face ?

Conclusion: porter un t-shirt rouge augmente les chances de tirer des faces...

Vetements et fécondité

Women Are More Likely to Wear Red or Pink at Peak Fertility

Alec T. Beall

Jessica L. Tracy

University of British Columbia

Alec T. Beall, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia V6T 1Z4, Canada E-mail: alec@psych.ubc.ca

Author Contributions Both authors contributed to the study design. Data collection, analyses, and interpretations were performed by A. T. Beall under the supervision of J. L. Tracy. Both authors contributed to the composition of the manuscript, with A. T. Beall composing initial drafts. Both authors approved the final version of the manuscript for submission.

Abstract

Although females of many species closely related to humans signal their fertile window in an observable manner, often involving red or pink coloration, no such display has been found for humans. Building on evidence that men are sexually attracted to women wearing or surrounded by red, we tested whether women show a behavioral tendency toward wearing reddish clothing when at peak fertility. Across two samples (N = 124), women at high conception risk were more than 3 times more likely to wear a red or pink shirt than were women at low conception risk, and 77% of women who wore red or pink were found to be at high, rather than low, risk. Conception risk had no effect on the prevalence of any other shirt color. Our results thus suggest that red and pink adornment in women is reliably associated with fertility and that female ovulation, long assumed to be hidden, is associated with a salient visual cue.

Vetements et fécondité

Women Are More Likely to Wear Red or Pink at Peak Fertility

Alec T. Beall

Jessica L. Tracy

University of British Columbia

Alec T. Beall, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia V6T 1Z4, Canada E-mail: alec@psych.ubc.ca

Author Contributions Both authors contributed to the study design. Data collection, analyses, and interpretations were performed by A. T. Beall under the supervision of J. L. Tracy. Both authors contributed to the composition of the manuscript, with A. T. Beall composing initial drafts. Both authors approved the final version of the manuscript for submission.

Abstract

Although females of many species closely related to humans signal their fertile window in an observable manner, often involving red or pink coloration, no such display has been found for humans. Building on evidence that men are sexually attracted to women wearing or surrounded by red, we tested whether women show a behavioral tendency toward wearing reddish clothing when at peak fertility. Across two samples (N = 124), women at high conception risk were more than 3 times more likely to wear a red or pink shirt than were women at low conception risk, and 77% of women who wore red or pink were found to be at high, rather than low, risk. Conception risk had no effect on the prevalence of any other shirt color. Our results thus suggest that red and pink adornment in women is reliably associated with fertility and that female ovulation, long assumed to be hidden, is associated with a salient visual cue.

Conclusion: les femmes atteignant leur pic de fécondité portent 3 fois plus de vêtements rouges que les autres...

Esprit statistiquement critique

- Les absurdités et manipulations à base de chiffres sont partout : politique, presse, et même recherche.
- Les chiffres ont, pour la plupart des gens, une autorité intrinsèque ("c'est scientifique").
- Les conclusions ne sont que le fruit de **l'interprétation**. Il faut dissocier résultats et conclusion.
- On ne fait rien dire du tout aux chiffres, mais on peut les utiliser pour faire passer ses opinions.
- Objectif premier de ce cours : ne plus se faire manipuler !

Il est naturel d'avoir une mauvaise intuition

13 à table?

Combinaison parfaite

Bien traiter les chiffres, c'est être impartial : ne pas chercher à obtenir un résultat en particulier car ils sont parfois très contre-intuitifs.

Paradoxe de Simpson

100 étudiants (50 hommes et 50 femmes) sont répartis sur 2 cours : fouille de données et systèmes avancés. Voici leurs pourcentages de validation des cours (exemple fictif!).

Fouille de données		**Systèmes avancés**	
Hommes	Femmes	Hommes	Femmes
90%	84.5%	70%	60%

Les hommes réussissent mieux **chacun** des cours.

Réussite globale	
Hommes	Femmes

Paradoxe de Simpson

100 étudiants (50 hommes et 50 femmes) sont répartis sur 2 cours : fouille de données et systèmes avancés. Voici leurs pourcentages de validation des cours (exemple fictif!).

Fouille de données		**Systèmes avancés**	
Hommes	Femmes	Hommes	Femmes
90%	84.5%	70%	60%

Les hommes réussissent mieux **chacun** des cours. Et pourtant :

Réussite globale	
Hommes	Femmes
74%	82%

Comment est-ce possible ?

Paradoxe de Simpson : explication

Les femmes sont plus nombreuses dans le cours où elles réussissent le mieux. Dans le cours où elles réussissent mieux, elles font un meilleur score que les hommes dans le cours où **ils** réussissent mieux. C'est donc une question de **répartition** des hommes et des femmes dans les cours.

Fouille de données		**Systèmes avancés**	
Hommes	Femmes	Hommes	Femmes
90%	84.5%	70%	60%
9/10	38/45	28/40	3/5

Réussite globale	
Hommes	Femmes
74%	82%
37/50	41/50

Paradoxe des anniversaires

Quelle est la probabilité que deux personnes parmi vous aient la même date d'anniversaire ?

Paradoxe des anniversaires

Quelle est la probabilité que deux personnes parmi vous aient la même date d'anniversaire ?

- $>$ 50% si vous êtes plus de 23
- $>$ 80% si vous êtes plus de 35
- $>$ 90% si vous êtes plus de 41
- $>$ 95% si vous êtes plus de 47
- $>$ 99% si vous êtes plus de 58

Paradoxe des anniversaires

Quelle est la probabilité que deux personnes parmi vous aient la même date d'anniversaire ?

- $>$ 50% si vous êtes plus de 23
- $>$ 80% si vous êtes plus de 35
- $>$ 90% si vous êtes plus de 41
- $>$ 95% si vous êtes plus de 47
- $>$ 99% si vous êtes plus de 58

Vérifions la théorie !

Paradoxe des anniversaires : explication

Il serait **très** improbable que vous ayez tous une date d'anniversaire différente.
Itérons:

- La première personne *choisit* sa date parmi 365 dates. Il reste 364 choix pour la seconde.
- La seconde *choisit* sa date. Il reste 363 choix.
- ...
- La n-ème personne a $(365 - n + 1)$ choix.

Si on transforme cela en **probabilités**, on obtient :

$$p = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365 - n + 1}{365}$$

p est la probabilité que les n personnes aient des anniversaires différents. Très rapidement, cette probabilité devient **infime** (on ne multiplie que des nombres < 1).
La probabilité que deux personnes **au moins** partage la même date est donc $1 - p$.

Paradoxe des Trois Portes (Monty Hall)

Un candidat à un jeu télévisé se trouve devant 3 portes. Derrière 2 portes, il n'y a rien. Derrière 1 des portes, une voiture.

- il choisit une porte
- l'animateur ouvre l'une des deux autres **qui ne cache pas la voiture**
- il reste donc 1 porte choisie au départ et une autre porte fermée
- l'animateur propose au candidat de changer de porte

Le candidat a-t-il intérêt à changer de porte ?

Paradoxe des Trois Portes (Monty Hall)

Un candidat à un jeu télévisé se trouve devant 3 portes. Derrière 2 portes, il n'y a rien. Derrière 1 des portes, une voiture.

- il choisit une porte
- l'animateur ouvre l'une des deux autres **qui ne cache pas la voiture**
- il reste donc 1 porte choisie au départ et une autre porte fermée
- l'animateur propose au candidat de changer de porte

Le candidat a-t-il intérêt à changer de porte ?

OUI

Paradoxe des Trois Portes : explication

Regardons les probabilités :

- Au départ, le candidat a 1 chance sur 3 de choisir la bonne porte
- Lorsque le présentateur en ouvre une autre qui ne contient pas la voiture, il apporte une information supplémentaire : la porte restante a donc 2 chances sur 3 de contenir la voiture.
- Le candidat **doit donc changer de porte**, passant sa probabilité de gagner de 1/3 à 2/3.

Outline

Une science à la mode

Pourquoi ?

- Stockage et traitement des données : de moins en moins cher.
- Impossible de les comprendre "à la main". Exemples : SNCF, génétique, finance, réseaux sociaux, publicité...
- Dépendance d'un grand nombre de facteurs.
- Big Data: le mot magique (qui n'a pas toujours de sens)
- Compétences recherchées par les entreprises (mots-clés) : datamining, analyse de données, big data, traitement automatique de texte, d'images, machine learning...
- $$$$$

Rendre les ordinateurs intelligents

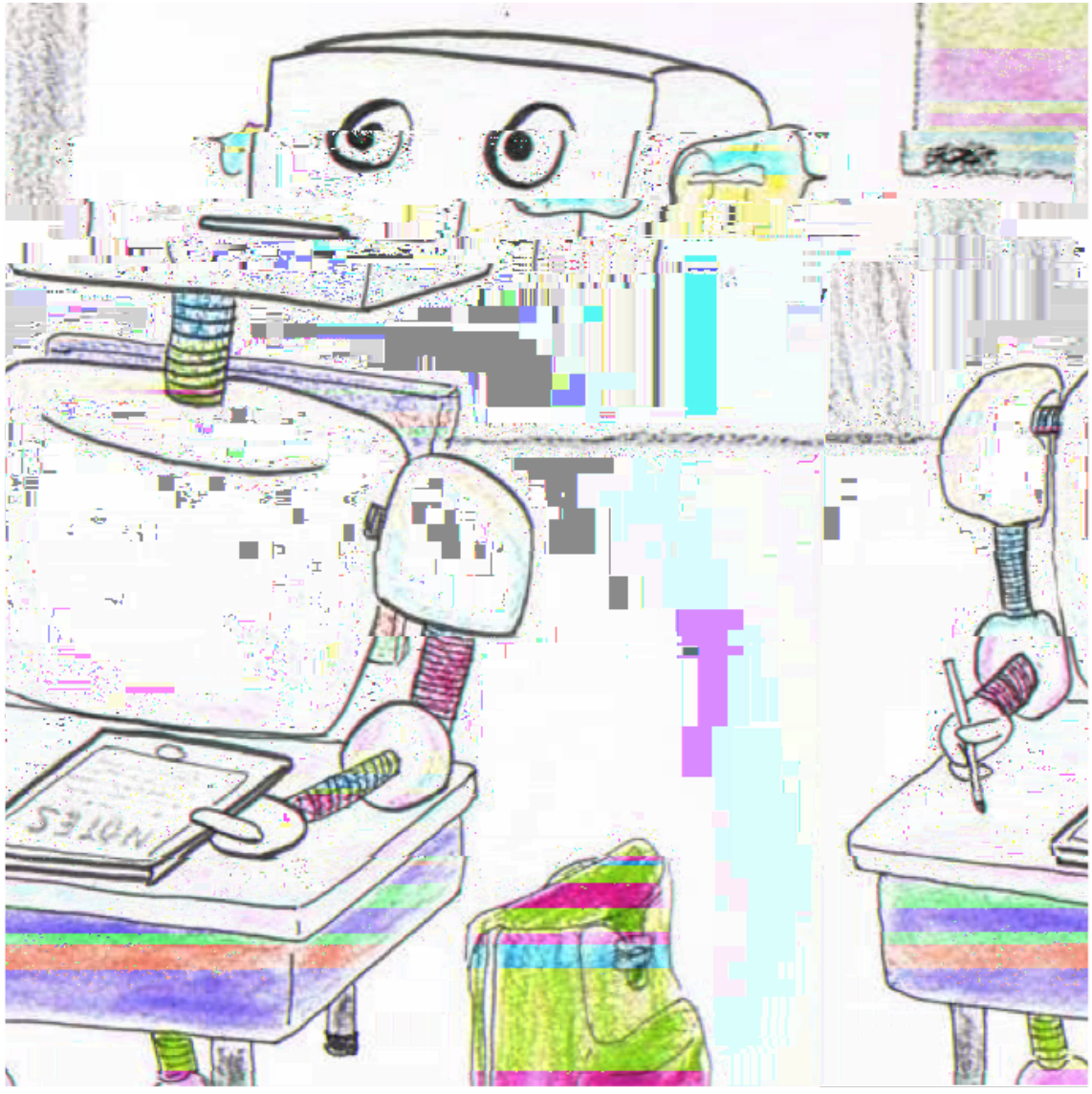


Figure: *Tiré du blog du laboratoire "Computer and Cognition", NYU.*

La rencontre de plusieurs disciplines

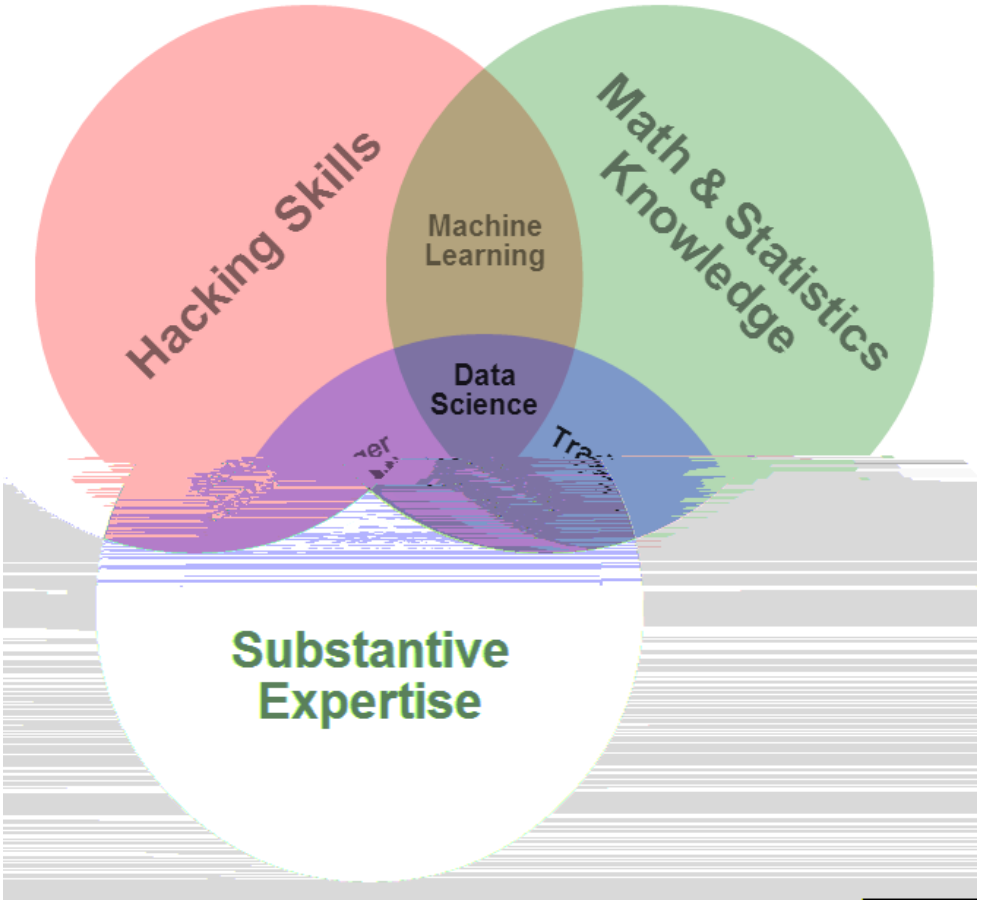
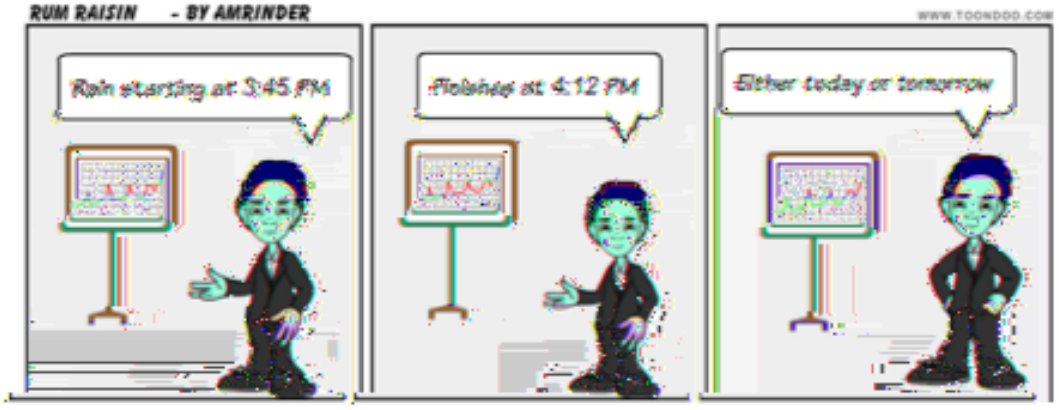


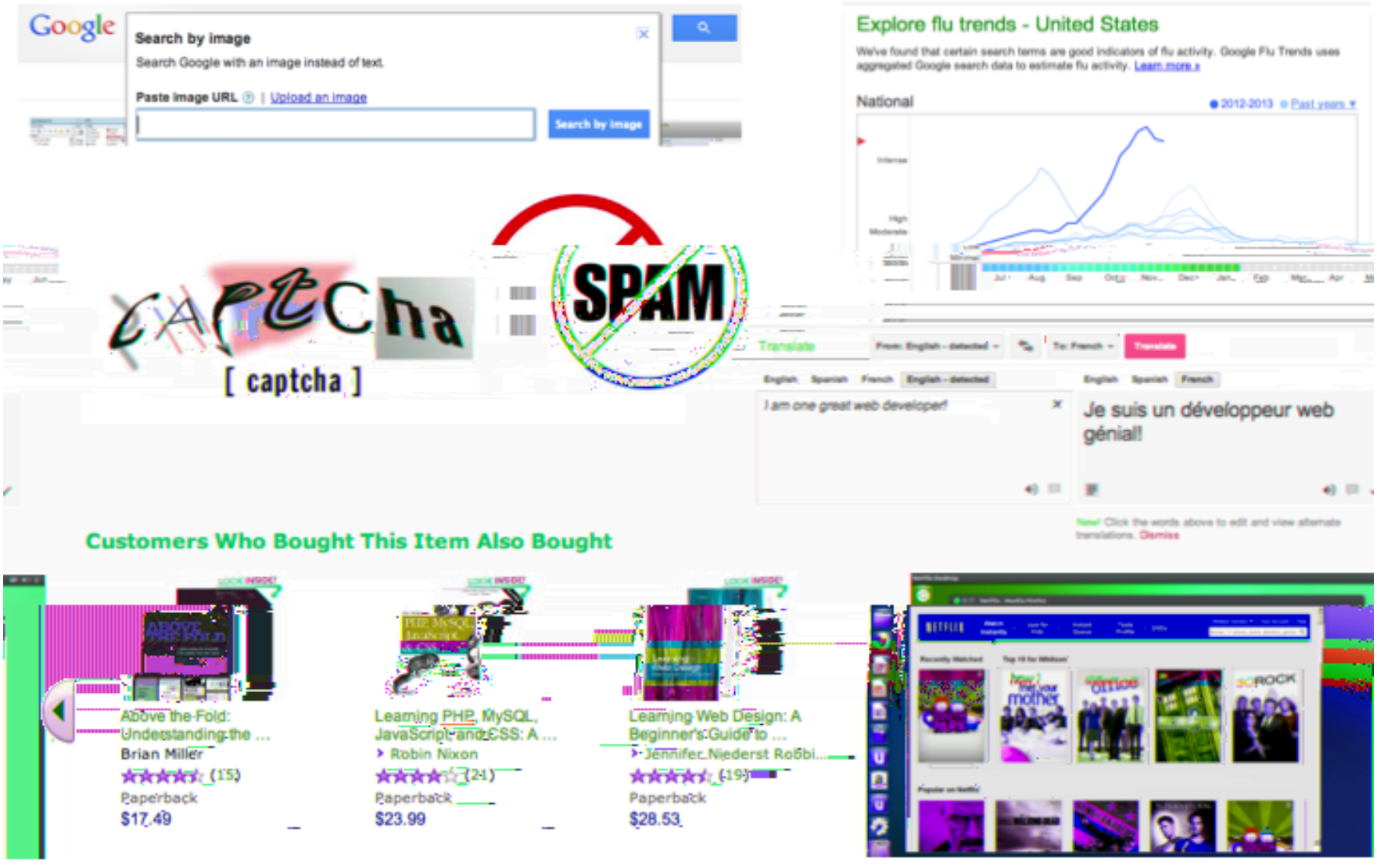
Figure: *Tiré de econometricsense.blogspot.fr*

- Changement de cap de plus en plus observé : des statistiques traditionnelles aux modèles algorithmiques.
- Besoin de modélisation mais aussi de méthodes rapides et généralisables à la grande dimension.

Quelques applications



Applications web



Outline

Exemples

- Développer un anti-spam.
- Classer automatiquement des articles.
- Créer un moteur de recommandation d'articles ("vous avez aimé... vous aimerez")
- Créer un moteur de recommandation d'images.
- Crér un moteur de recommandation d'amis.
- Prédire l'utilisation d'un système (en commun avec le cours de F. Armand)
- Prédire des résultats sportifs.
- Un sujet de votre choix sous réserve de validation.

Organisation

- Jusqu'à 4 personnes par projet (à condition de travail équivalent).
- Rapport écrit, programme et oral.
- Certains projets plus difficiles/longs que d'autres. En fonction de l'importance du cours dans votre cursus.
- Les projets sont valorisables sur un CV.
- Projet de A à Z: collecte des données + encodage + visualisation + méthodo + résultats.

Collecte des données

En fonction du projet:

- Sources officielles (INSEE, data.gouv.fr, opendata.paris.fr).
- Crawler le web, parser le code html.
- Récupérer les données via des API (Facebook, Twitter, Amazon...)
- Questionnaire web.
- Sondage dans la rue.
- Prise de mesures (ex: programme qui stocke CPU, mémoire d'une machine toutes les minutes)
- ...

Programme

- Un programme (un minimum documenté, au moins commenté) doit accompagner le projet.
- Langage de votre choix. (Python plus simple ?)
- En fonction de votre projet: appli web, page html, exécutable, script... (Pas forcément d'interface.)

Rapport

- Rapport à rendre avec le programme.
- Une dizaine de pages (plus si nécessaire) comprenant:
  1. Présentation du projet/motivation.
  2. Description (visuelle et/ou tableau) des données.
  3. Méthodo utilisée.
  4. Résultats.
  5. Conclusion.

Planning

- Semaine 2: Choix du projet et formation des équipes.
- Semaines 2 à 4: Collecte des données.
- Semaines 4 à 9: Analyse et rédaction. 1 RDV de suivi par groupe et suivi par mail en permanence.
- Semaine 9: rendu du rapport.
- Semaine 10: oral/démo (pendant le dernier cours). Vote de tout le monde et prix du meilleur projet.
- Avant le stage: obtention des notes (pas le plus important!)

L’esprit du cours

- Interactif
- Travail d’équipe
- Appliqué
- Toute proposition de thèmes à aborder est toujours la bienvenue.

Etapes d'un projet de datamining

1. Collecte
2. Encodage
3. Description
4. (Visualisation)
5. Prédiction ou Compréhension
6. Evaluation

Outline

Exemple 1 : Anti-spam textos

Données:

- messages sms (en anglais)
- 747 spams. *Exemple: "WINNER!! As a valued network customer you have been selected to receive a prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only."*
- 4827 non-spams. *Exemple: "Nah I don't think he goes to usf, he lives around here though"*
- 9663 mots distincts.

Problème difficile! Exemple filé sur le cours, à la fin nous l'aurons implémenté.

Objectif:

- comprendre ce qui différencie les deux types de messages;
- établir une règle de séparation;
- appliquer la règle à un nouveau message entrant.

Exemple 2 : reconnaissance de chiffres manuscrits

Données:

- chiffres manuscrits de 0 à 4
- 901 exemples
- Environ 180 par classe.
- Chaque image : 8 pixels x 8 pixels

Objectif:

- être capable de classifier automatiquement ces images.
- établir une règle de séparation;
- appliquer la règle à un nouveau chiffre entrant. (application code postaux)

Plan du cours (sujet à modifications)

- Séance 1: Choix et encodage des données
- Séance 2: Statistiques descriptives
- Séance 3: Tests statistiques, décision rapide
- Séance 4: Visualisation, réduction de dimension
- Séance 5: Similarités et distances
- Séance 6: Clustering
- Séance 7: Apprentissage supervisé (1)
- Séance 8: Apprentissage supervisé (2)
- Séance 9: Introduction aux graphes
- Séance 10: Présentation des projets