

FOUILLE DE DONNÉES ET AIDE A LA DECISION

Introduction au machine learning.

Anne-Claire Haury

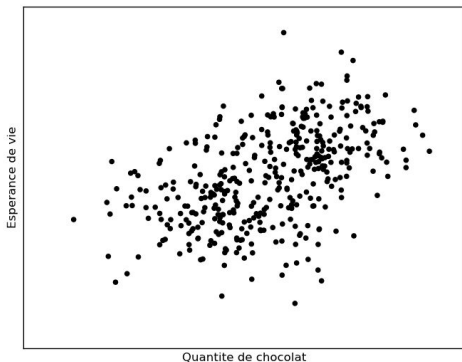
M2 Informatique
Université Denis Diderot

Premier semestre 2016-2017

INTRODUCTION

CHOCOLAT ET ESPÉRANCE DE VIE

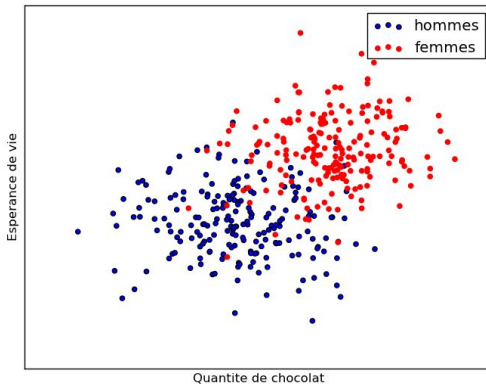
Exemple emprunté à Isabelle Guyon



Manger du chocolat **augmente** l'espérance de vie.

CHOCOLAT ET ESPÉRANCE DE VIE

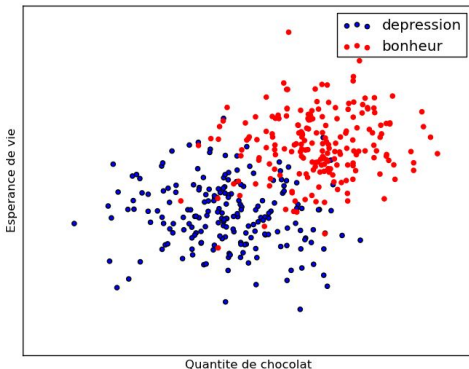
Exemple emprunté à Isabelle Guyon



Manger du chocolat **n'augmente pas** l'esperance de vie.

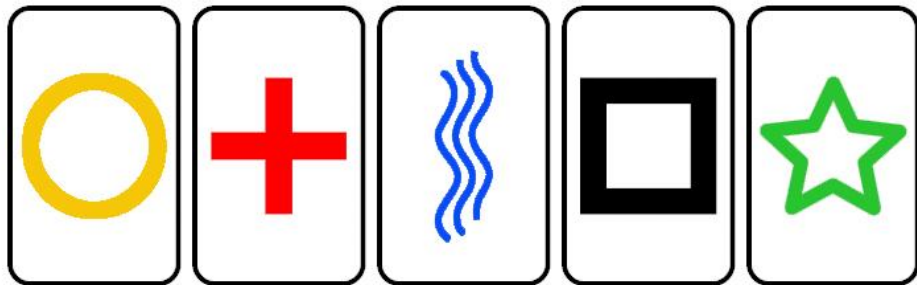
CHOCOLAT ET ESPÉRANCE DE VIE

Exemple emprunté à Isabelle Guyon



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

LES EXPÉRIENCES DE RHINE



Source: Wikipedia

PILE OU FACE?



PILE OU FACE?



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

VÊTEMENTS 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.

VÊTEMENTS 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.

Un des objectifs de ce cours : **ne plus se faire manipuler !**

VOUS AVEZ UNE MAUVAISE INTUITION STATISTIQUE (SI SI!)

Un dîner
PRESQUE
parfait



PEUT-ON FAIRE DIRE AUX CHIFFRES CE QUE L'ON VEUT ?

<i>Données fictives !</i>	Nombre de chômeurs	Nombre de travailleurs potentiels
Année 1	1.000.000	10.000.000
Année 2	1.010.000	11.000.000

“Le chômage a augmenté de 1%.”

VRAI
OU
FAUX ?

“Le taux de chômage a baissé de 0.9 points.”

“Il y a 10.000 chômeurs de plus.”

“Le taux de chômage a baissé de 10%.”

PEOPLE VS COLLINS



1964. Un vol. Les témoins affirment avoir vu un homme noir barbu et moustachu et une femme blonde avec une queue de cheval s'enfuir dans une voiture jaune. Malcolm et Janet Collins correspondent à la description...

PEOPLE VS COLLINS

Raisonnement du procureur :

- Homme noir portant une barbe : 10%
- Homme noir portant une moustache : 25%
- Femme blanche portant une queue de cheval : 10%
- Femme blanche ayant des cheveux blonds : 33%
- Voiture en partie jaune : 10%
- Couple "inter-racial" dans une voiture : 0.1%

Ils en concluent que la probabilité que les Collins soient innocents est de 1/12 millions. Ils sont donc condamnés.

La cour d'appel annule la condamnation. Quelle était l'erreur du jury lors du procès en première instance ?

EXPLICATION

Admettons que les probabilités, bien qu'estimées sans doute arbitrairement, soient justes.

L'erreur principale est d'avoir ignoré les dépendances entre les événements.

Au lieu de multiplier toutes les probabilités entre elles, il faut prendre en compte le fait que les événements ne sont pas indépendants et considérer les **probabilités conditionnelles**.

Par exemple, la probabilité d'avoir une moustache sachant que l'on a une barbe est très élevée, disons 90%. Donc la probabilité d'avoir une barbe ET une moustache devient $10\% \times 90\%$ au lieu de $10\% \times 25\%$. De même pour les autres événements.

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!).

Fouilles de données		Systèmes avancés	
Hommes	Femmes	Hommes	Femmes
90%	84.5%	70%	60%

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

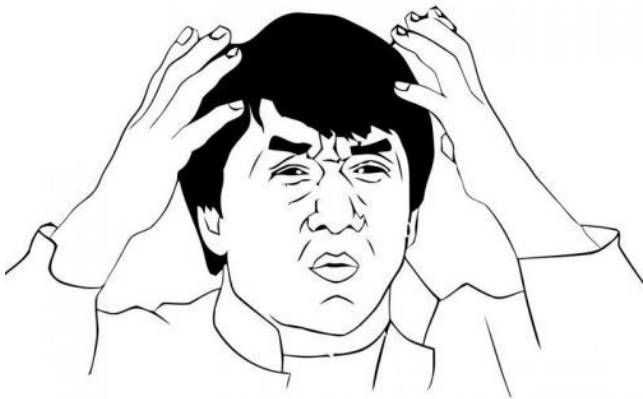
PARADOXE DE SIMPSON

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!).

Fouilles de données		Systèmes avancés	
Hommes	Femmes	Hommes	Femmes
90%	84.5%	70%	60%

Réussite globale:

Hommes	Femmes
74%	82%



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**.

Fouilles de données		Systèmes avancés	
Hommes	Femmes	Hommes	Femmes
90% (9/10)	84.5% (38/45)	70% (28/40)	60% (3/5)
Réussite globale:		Hommes	Femmes
		74% (37/50)	82% (41/50)

PARADOXE DES ANNIVERSAIRES

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

<https://goo.gl/forms/ZiD5U83M7ZPyHElq2>

ou

<https://sites.google.com/site/dataminingp7/formulaires>

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

EXPLICATION

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

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

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 $1 - p$.

EXEMPLE AVEC 50 PERSONNES

Probabilité d'avoir des anniversaires différents:

$$\begin{aligned} p &= \frac{365}{365} \times \frac{364}{365} \times \dots \times \frac{365 - 50 + 1}{365} \\ &= \frac{365 \times 364 \times \dots \times 316}{365^{50}} \\ &= 0.0296 \end{aligned}$$

Il y a donc 97% de chances qu'au moins 2 personnes aient le même anniversaire.

ET AU POKER ?

Sur un jeu de 52 cartes, quelle est la probabilité que j'aie une paire d'As ?

A - $2/52$

B - $1/52$

C - $1/221$

D - $1/1326$

D - $1/2652$

ET AU POKER ?

Sachant que j'ai As/Roi dans la main, quelle est la probabilité que mon adversaire ait une paire d'As?

A - $2/52$

B - $1/52$

C - $1/221$

D - $1/1326$

D - $1/2652$

CE QU'ON EN CONCLUT

Avoir de l'information change drastiquement la donne!

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

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 $2/3$.

LE MACHINE LEARNING

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

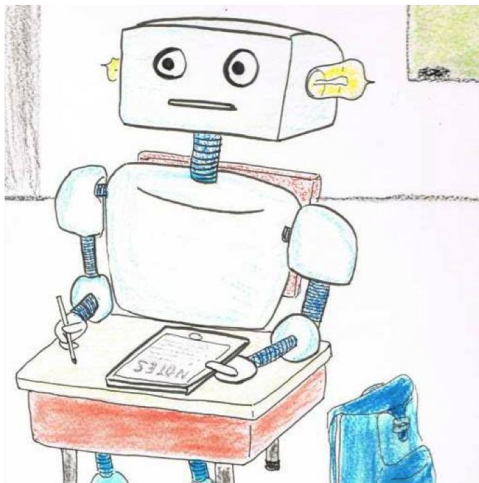
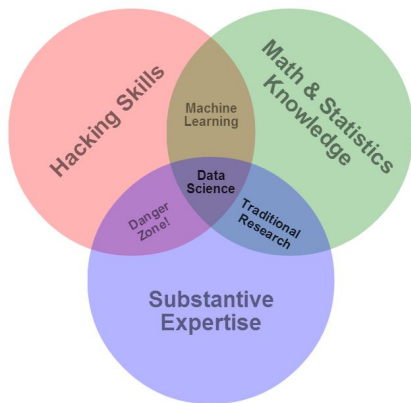


Figure: Tiré du blog du laboratoire "Computer and Cognition", NYU.

LA RENCONTRE DE PLUSIEURS DISCIPLINES



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.

Figure: Tiré de econometricsense.blogspot.fr



"Let me guess. You had to review your diagnosis."

0	1000, 4 pages	1001, 2 pages	1002, 2 pages
1	1003, 3 pages	1004, 2 pages	1005, 4 pages
2	1006, 2 pages	1007, 3 pages	1008, 2 pages
3	1009, 2 pages	1010, 3 pages	1011, 3 pages
4	1012, 4 pages	1013, 3 pages	1014, 3 pages
5	1015, 3 pages	1016, 3 pages	1017, 2 pages
6	1018, 3 pages	1019, 3 pages	1020, 2 pages
7	1021, 3 pages	1022, 3 pages	1023, 5 pages
8	1024, 3 pages	1025, 5 pages	1026, 3 pages
9	1027, 4 pages	1028, 4 pages	1029, 3 pages
10	1030, 3 pages	1031, 3 pages	1032, 3 pages
11	1033, 3 pages	1034, 3 pages	1035, 3 pages
12	1036, 3 pages	1037, 3 pages	1038, 3 pages
13	1039, 3 pages	1040, 3 pages	1041, 3 pages
14	1042, 3 pages	1043, 3 pages	1044, 3 pages
15	1045, 3 pages	1046, 3 pages	1047, 3 pages
16	1048, 3 pages	1049, 3 pages	1050, 3 pages
17	1051, 3 pages	1052, 3 pages	1053, 3 pages
18	1054, 3 pages	1055, 3 pages	1056, 3 pages
19	1057, 3 pages	1058, 3 pages	1059, 3 pages
20	1060, 3 pages	1061, 3 pages	1062, 3 pages
21	1063, 3 pages	1064, 3 pages	1065, 3 pages
22	1066, 3 pages	1067, 3 pages	1068, 3 pages
23	1069, 3 pages	1070, 3 pages	1071, 3 pages
24	1072, 3 pages	1073, 3 pages	1074, 3 pages
25	1075, 3 pages	1076, 3 pages	1077, 3 pages
26	1078, 3 pages	1079, 3 pages	1080, 3 pages
27	1081, 3 pages	1082, 3 pages	1083, 3 pages
28	1084, 3 pages	1085, 3 pages	1086, 3 pages
29	1087, 3 pages	1088, 3 pages	1089, 3 pages
30	1090, 3 pages	1091, 3 pages	1092, 3 pages
31	1093, 3 pages	1094, 3 pages	1095, 3 pages
32	1096, 3 pages	1097, 3 pages	1098, 3 pages
33	1099, 3 pages	1100, 3 pages	1101, 3 pages
34	1102, 3 pages	1103, 3 pages	1104, 3 pages
35	1105, 3 pages	1106, 3 pages	1107, 3 pages
36	1108, 3 pages	1109, 3 pages	1110, 3 pages
37	1111, 3 pages	1112, 3 pages	1113, 3 pages
38	1114, 3 pages	1115, 3 pages	1116, 3 pages
39	1117, 3 pages	1118, 3 pages	1119, 3 pages
40	1120, 3 pages	1121, 3 pages	1122, 3 pages
41	1123, 3 pages	1124, 3 pages	1125, 3 pages
42	1126, 3 pages	1127, 3 pages	1128, 3 pages
43	1129, 3 pages	1130, 3 pages	1131, 3 pages
44	1132, 3 pages	1133, 3 pages	1134, 3 pages
45	1135, 3 pages	1136, 3 pages	1137, 3 pages
46	1138, 3 pages	1139, 3 pages	1140, 3 pages
47	1141, 3 pages	1142, 3 pages	1143, 3 pages
48	1144, 3 pages	1145, 3 pages	1146, 3 pages
49	1147, 3 pages	1148, 3 pages	1149, 3 pages
50	1150, 3 pages	1151, 3 pages	1152, 3 pages
51	1153, 3 pages	1154, 3 pages	1155, 3 pages
52	1156, 3 pages	1157, 3 pages	1158, 3 pages
53	1159, 3 pages	1160, 3 pages	1161, 3 pages
54	1162, 3 pages	1163, 3 pages	1164, 3 pages
55	1165, 3 pages	1166, 3 pages	1167, 3 pages
56	1168, 3 pages	1169, 3 pages	1170, 3 pages
57	1171, 3 pages	1172, 3 pages	1173, 3 pages
58	1174, 3 pages	1175, 3 pages	1176, 3 pages
59	1177, 3 pages	1178, 3 pages	1179, 3 pages
60	1180, 3 pages	1181, 3 pages	1182, 3 pages
61	1183, 3 pages	1184, 3 pages	1185, 3 pages
62	1186, 3 pages	1187, 3 pages	1188, 3 pages
63	1189, 3 pages	1190, 3 pages	1191, 3 pages
64	1192, 3 pages	1193, 3 pages	1194, 3 pages
65	1195, 3 pages	1196, 3 pages	1197, 3 pages
66	1198, 3 pages	1199, 3 pages	1200, 3 pages



FiveThirtyEight

Nate Silver's Political Calculus

APPLICATIONS WEB

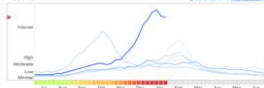


Explore flu trends - United States

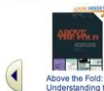
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more](#)

National

2012-2013 Fast years



Customers Who Bought This Item Also Bought



Above the Fold:
Understanding the ...
Brian Miller
★★★★☆ (15)
Paperback
\$17.49



Learning PHP, MySQL,
JavaScript, and CSS: A ...
Robin Nixon
★★★★☆ (21)
Paperback
\$23.99



Learning Web Design: A
Beginner's Guide to ...
Jennifer Niederst Robbi...
★★★★☆ (19)
Paperback
\$28.53



LES PROJETS

EXEMPLES DE PROJETS

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.

Créer un moteur de recommandation d'amis.

Prédire quel film va gagner les oscars.

Prédire le temps d'attente à l'aéroport.

Prédire des résultats sportifs.

Prédire la disponibilité des vélib.

Un sujet de votre choix sous réserve de validation.

ORGANISATION

Jusqu'à 3 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)

CODE

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 DE PROJET

Rapport à rendre avec le programme.

Une dizaine de pages (plus si nécessaire) comprenant:

- Présentation du projet/motivation.

- Description (visuelle et/ou tableau) des données.

- Méthodo utilisée.

- Résultats.

- 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

Collecte

Encodage

Description

(Visualisation)

Prédiction ou Compréhension

Evaluation

SITE WEB

<https://sites.google.com/site/dataminingp7>

INSCRIPTION AU COURS

<https://goo.gl/forms/tQlzW2fKZY7jNFi12>

<https://sites.google.com/site/dataminingp7/formulaires>

GOOGLE INTERN OPEN HOUSE

Mardi 11 octobre, ouverture des portes à 17h30.

Il est encore temps de s'inscrire.

Pour visiter les bureaux, rencontrer des Googlers et tout savoir sur les stages. Faire vite car il y a un nombre de places limitées.

<https://services.google.com/fb/forms/paris-internshipopenhouse/>