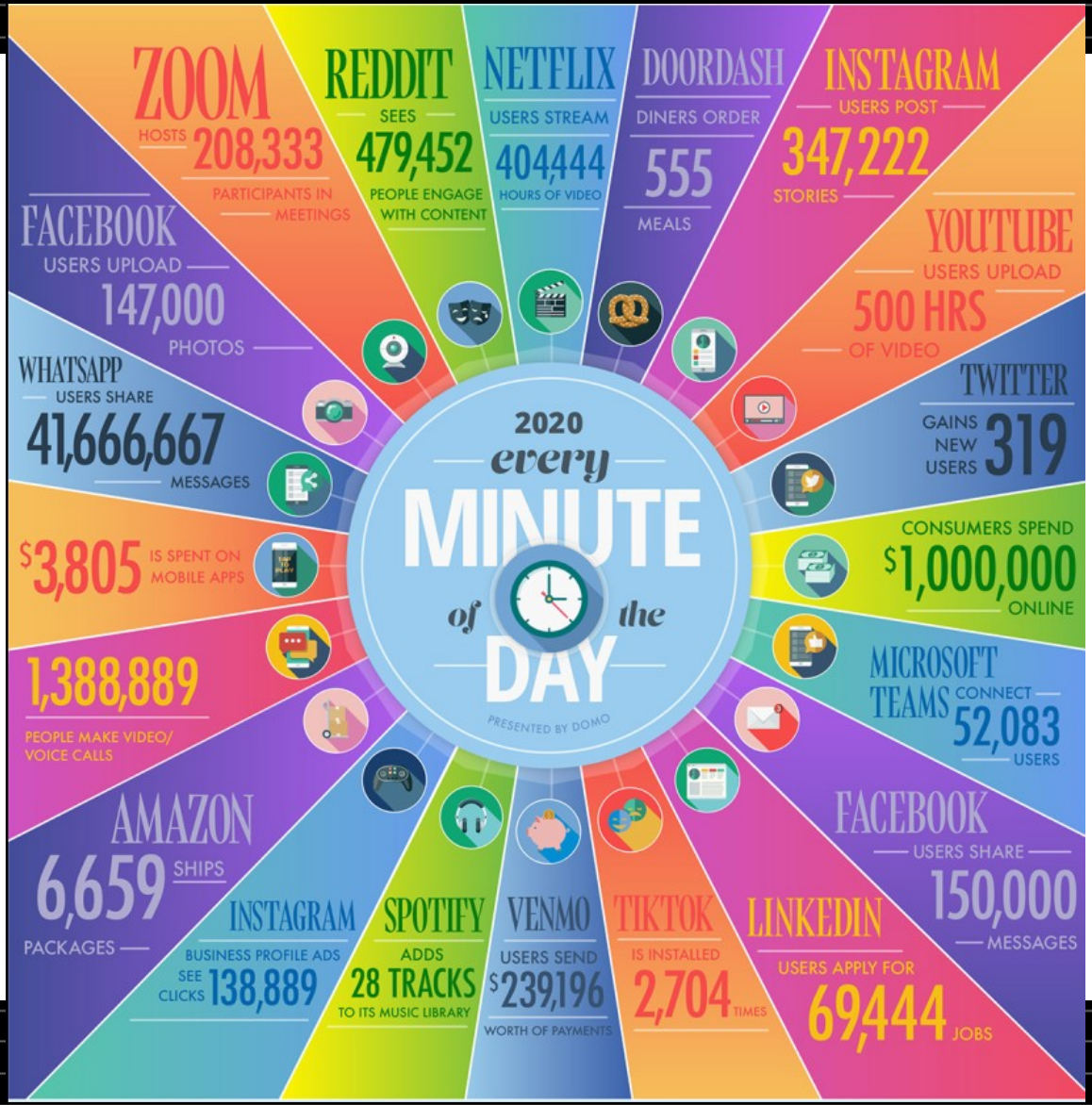
Analyse de données avec python

Une intro CP2


2020
every
MINUTE
of the
DAY
PRESENTED BY DOMO
ZOOM
HOSTS 208,333
PARTICIPANTS IN MEETINGS
REDDIT
SEES 479,452
PEOPLE ENGAGE WITH CONTENT
NETFLIX
USERS STREAM 404,444
HOURS OF VIDEO
DOORDASH
DINERS ORDER 555
MEALS
INSTAGRAM
USERS POST 347,222
STORIES
FACEBOOK
USERS UPLOAD 147,000
PHOTOS
YOUTUBE
USERS UPLOAD 500 HRS
OF VIDEO
WHATSAPP
USERS SHARE 41,666,667
MESSAGES
TWITTER
GAINS NEW USERS 319
$3,805 IS SPENT ON MOBILE APPS
CONSUMERS SPEND $1,000,000 ONLINE
1,388,889
PEOPLE MAKE VIDEO/ VOICE CALLS
MICROSOFT TEAMS CONNECT 52,083 USERS
AMAZON
6,659 SHIPS
PACKAGES
FACEBOOK
USERS SHARE 150,000
MESSAGES
INSTAGRAM
BUSINESS PROFILE ADS SEE CLICKS 138,889
SPOTIFY
ADDS 28 TRACKS
TO ITS MUSIC LIBRARY
VENMO
USERS SEND $239,196
WORTH OF PAYMENTS
TIKTOK
IS INSTALLED 2,704 TIMES
LINKEDIN
USERS APPLY FOR 69,444 JOBS

Data Science

- Après l’accumulation de données
- Exploitation de ces données
- Compréhension de ces données
- Visualisation de ces données

Définition de la data Science

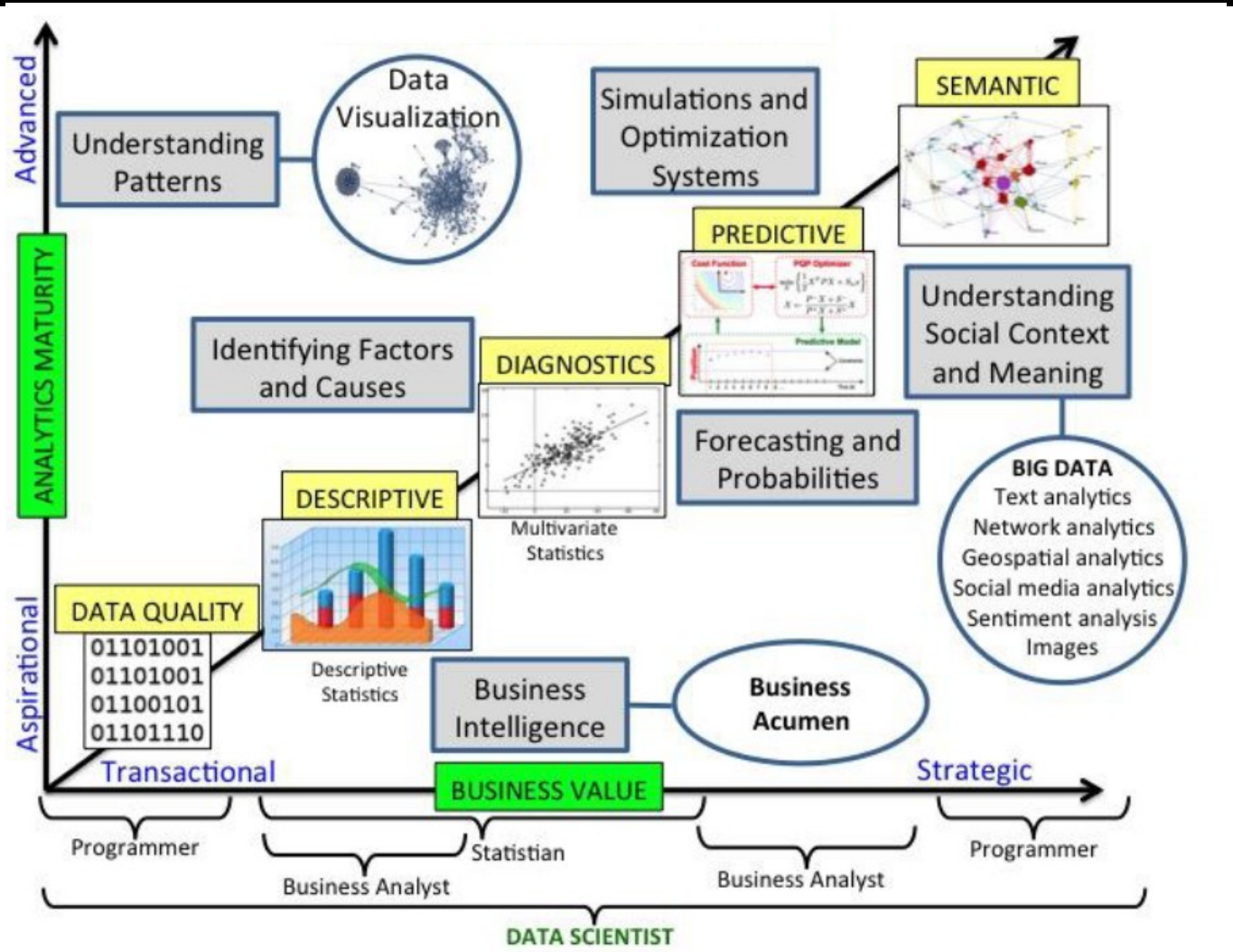
- S’appuyer sur des outils mathématiques, statistiques, informatiques et de visualisation des données pour transformer ces données brutes en informations utiles

Le data scientist

Il doit tout savoir faire !

- Informaticien
- Mathématicien
- Financier
- Analyste
- Visualiseur !

Une autre vue



Complexe mais

- Demande croissante du métier de Data Scientist
- Apprendre à prédire le futur c’est plutôt sympa comme job
- On va apprendre à manipuler en quelques lignes des centaines ou milliers de données
- Il y a déjà plein d’outils


ANATOMY OF A
DATA SCIENTIST
SALARY
Average salary of data scientists is
$120,000/year
BENEFITS
• Harvard Business Review called data science the "Sexiest Job of the 21st Century"
• One of the fastest growing careers in the United States
• 94% of data science graduates have found jobs since 2011
RESPONSIBILITIES
• Conduct research
• Extract, clean, and analyze data from varied sources
• Solve problems
• Build automation tools
• Communicate findings to management
EDUCATION
• 88% of all data scientists have at least a Master's degree
• 46% of data scientists have a PhD
SKILLS
• Programming languages (R, Python, SQL, Hive, etc.)
• Statistics
• Multivariable calculus and linear algebra
• Machine learning
• Software engineering
• Wrangle, visualize, and communicate data to management
CAREER POSSIBILITIES
• The majority of data scientists work in the technology industry.
• Other options include marketing, consulting, healthcare and pharmaceuticals, finance, government, gaming, and many more.
RESOURCES:
https://insidebigdata.com/2017/08/05/benefits-data-scientist-career/
https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm
https://blog.udacity.com/2014/11/data-science-job-skills.html
https://online.rutgers.edu/resources/infographics/what-can-you-do-with-a-career-in-data-science/?program=mi
THE COMPUTER MERCHANT, LTD.
THE IT STAFFING COMPANY

Que faire de ces données ?

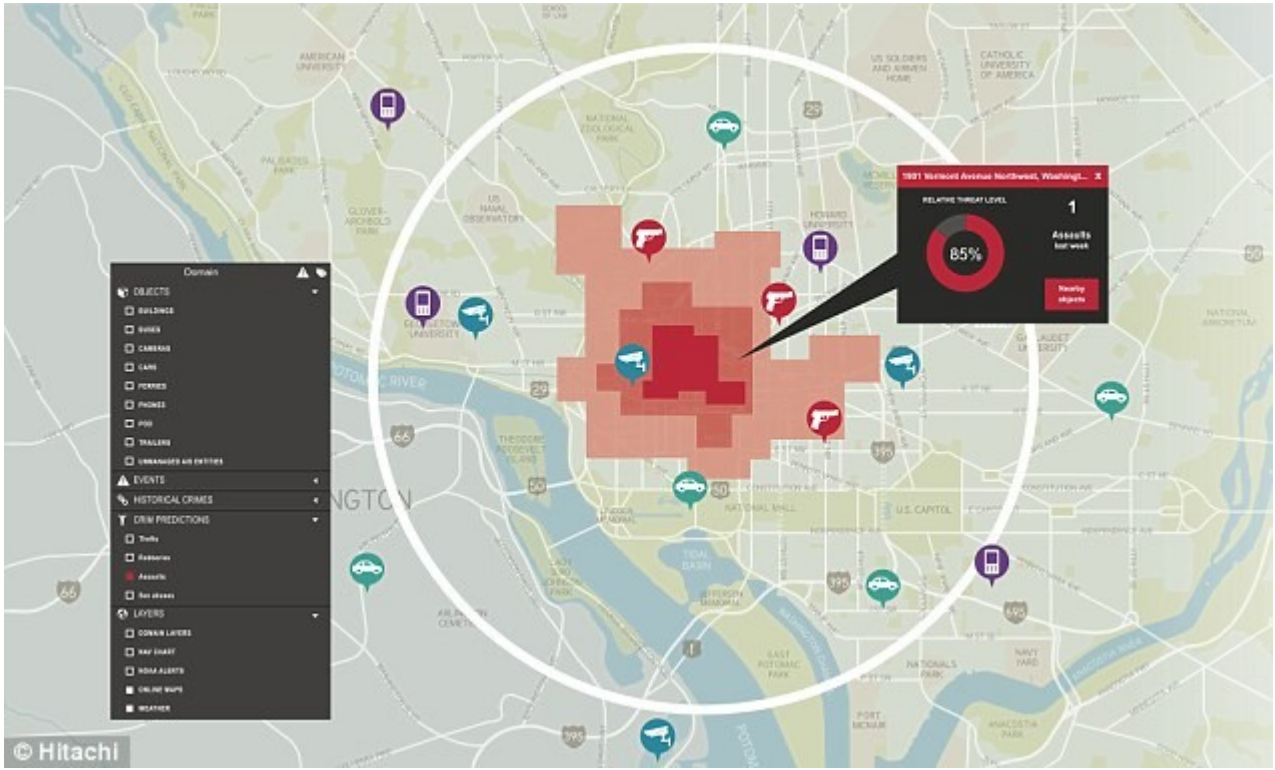
- Vendre
  - Mieux connaître son client (amazon, netflix, …)
  - Proposition d'achats supplémentaires
- Aider à la décision
  - Les banques, les assurances
  - Prévention des risques, de la délinquance,…
  - Médecine (anticiper la réponse d'un patient avant le début d'une chimiothérapie, ..)
  - Pharmacie (rapport coûts/bénéfices d'un traitement…)
  - Prévenir la pollution
- Science comportementale
  - comprendre des biais
  - analyse d'opinion sur le web

Secteur bancaire

je collabore avec le département des ressources humaines sur un projet en lien avec le Machine Learning et les données analytiques. Il s'agit de développer un outil qui permette d'avoir des visions, à date et future, de nos besoins en compétences. Cela permet au département RH d'anticiper les recrutements sur les prochaines années et de faire ainsi en sorte que la banque ne manque pas d'experts, notamment dans les domaines porteurs d'avenir comme la Blockchain.

Source : https://group.bnpparibas/actualite/metiers-banque-data-scientist-senior

Minority report



Source dailymail.co.uk

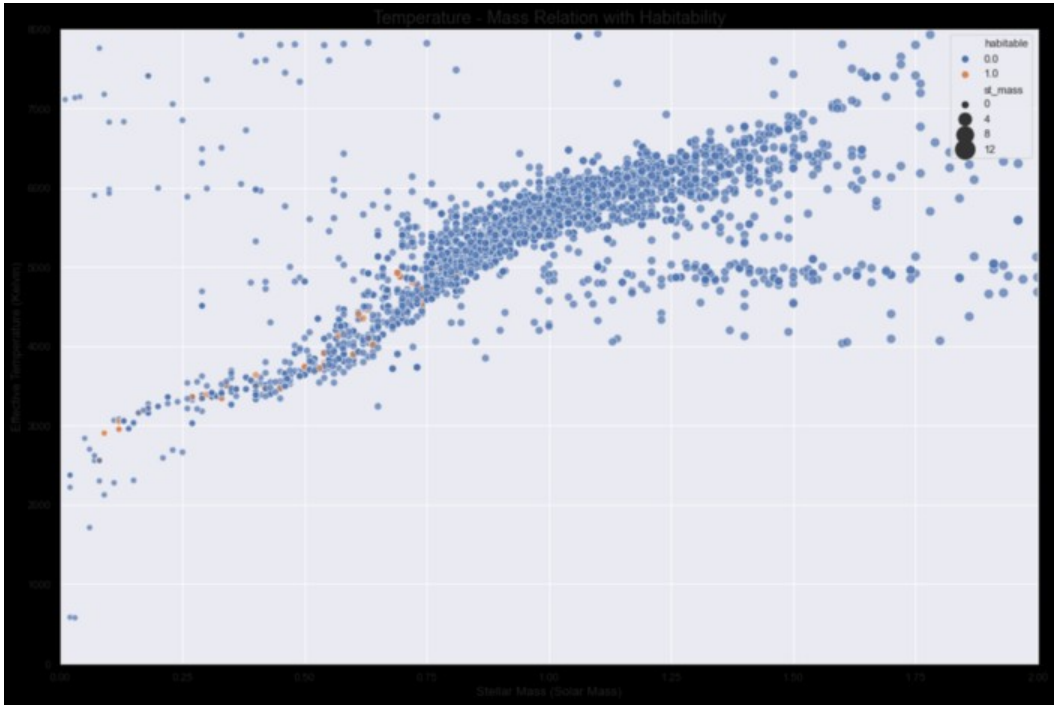
WHAT DATA DOES PCA USE?

Hitachi's Predictive Crime Analytics blends 'real-time event data from public safety systems and sensors with historical and contextual crime data, social media and other sources.'

These include:

- CCTV and video management systems such as Genetec and Pelco.
- Emergency call data
- Gunshot detection systems including Shotspotter
- Live weather radar
- Twitter feeds
- Traffic systems
- Crime and incident data

Détection des exoplanètes en astronomie



A vous de jouer ?

- La base de données des joueurs de la FIFA
- https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

```
Entrée [1]: import pandas as pd

Entrée [2]: https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

e:\langages\python3\lib\site-packages\IPython\core\interactiveshell.py:3146: DtypeWarning: Columns (23,35) have mixed type
s.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,

Entrée [3]: données = pd.read_csv('c:/Temp/CompleteDataset.csv',sep=',',low_memory=False)

Entrée [4]: données

Out[4]:
```

	Unnamed: 0	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club	Club Logo	...	RB	RCB	RCM
0	0	Cristiano Ronaldo	32	https://cdn.sofifa.org/48/18/players/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Real Madrid CF	https://cdn.sofifa.org/24/18/teams/243.png	...	61.0	53.0	82.0
1	1	L. Messi	30	https://cdn.sofifa.org/48/18/players/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	93	93	FC Barcelona	https://cdn.sofifa.org/24/18/teams/241.png	...	57.0	45.0	84.0
2	2	Neymar	25	https://cdn.sofifa.org/48/18/players/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	94	Paris Saint-Germain	https://cdn.sofifa.org/24/18/teams/73.png	...	59.0	46.0	79.0

Les données

- structurées / non structurées
- complètes / incomplètes / incorrectes
- textuelles / non textuelles

Données structurées

- format normalisé permettant de fournir des informations sur une page et de classer le contenu de cette page

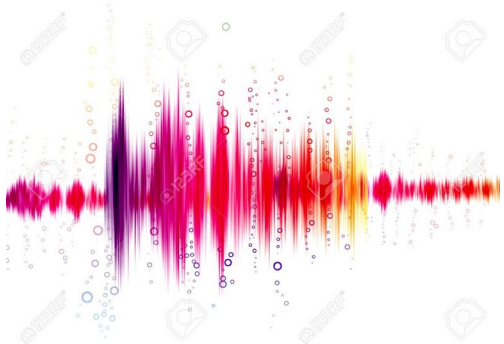
```
<html>
  <head>
    <title>Party Coffee Cake</title>
    <script type="application/ld+json">
    {
      "@context": "https://schema.org/",
      "@type": "Recipe",
      "name": "Party Coffee Cake",
      "author": {
        "@type": "Person",
        "name": "Mary Stone"
      },
      "datePublished": "2018-03-10",
      "description": "This coffee cake is awesome and perfect for parties.",
      "prepTime": "PT20M"
    }
    </script>
  </head>
  <body>
  <h2>Party coffee cake recipe</h2>
  <p>
    This coffee cake is awesome and perfect for parties.
  </p>
  </body>
</html>
```

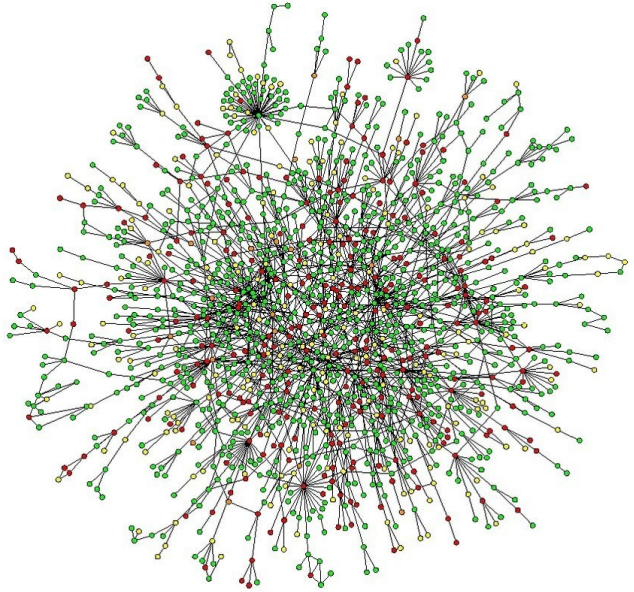
Données semi-structurées

- Contiennent des éléments d'informations
- Par exemple fichier json, XML, texte brut

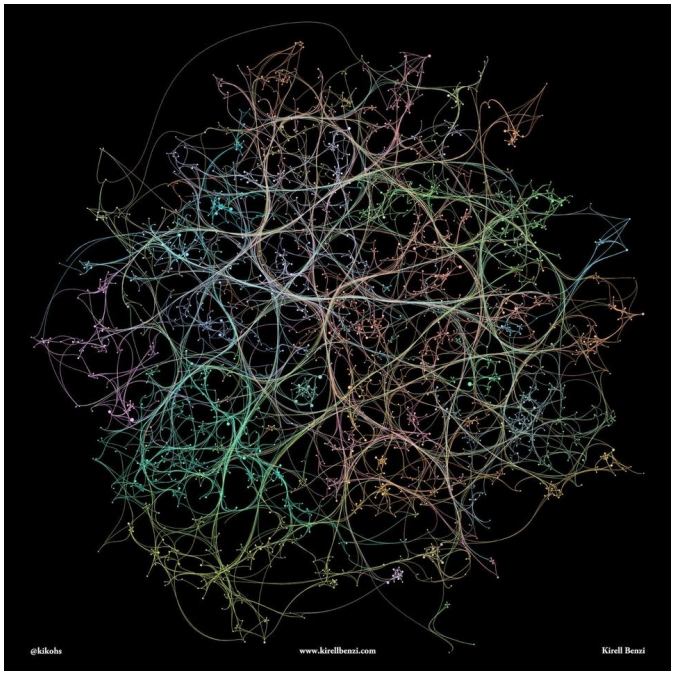
Données non structurées

- Tous les autres !





Data Science et Arts



startups de viva tech

https://actu.epfl.ch/news/
que-peut-nous-apprendre-wikipedia-sur-les-intera-4/

Dans le cours de CP2 l'éco-système python

- python
- numpy
- pandas
- sci-kit (peut-être)
- matplotlib, seaborn

Python

- *cf* résumé du cours
- un module pratique pour notre cours
- les expressions régulières (module re)
- utile pour analyser un fichier texte

Les données pour ce sours

- données structurées
  - listes
  - tableaux
  - réseaux

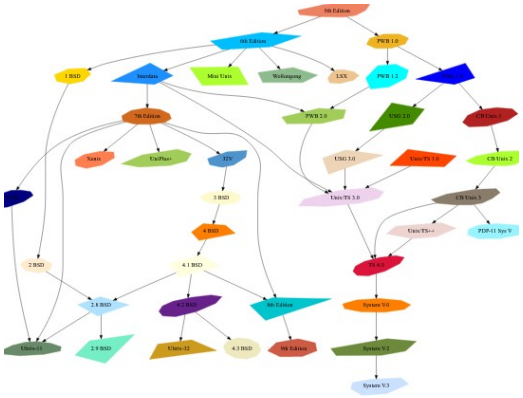
Les données structurées

- CSV (comma separated values)

```
,Name,Age,Photo,Nationality,Flag,Overall,Potential,Club,Club
Logo,Value,Wage,Special,Acceleration,Aggression,Agility,Balance,Ball
control,Composure,Crossing,Curve,Dribbling,Finishing,Free kick accuracy,GK
diving,GK handling,GK kicking,GK positioning,GK reflexes,Heading
accuracy,Interceptions,Jumping,Long passing,Long
shots,Marking,Penalties,Positioning,Reactions,Short passing,Shot power,Sliding
tackle,Sprint speed,Stamina,Standing
tackle,Strength,Vision,Volleys,CAM,CB,CDM,CF,CM,ID,LAM,LB,LCB,LCM,LDM,LF,LM,LS,LW
,LWB,Preferred Positions,RAM,RB,RCB,RCM,RDM,RF,RM,RS,RW,RWB,ST
0,Cristiano
Ronaldo,32,https://cdn.sofifa.org/48/18/players/20801.png,Portugal,https://cdn.so
fifa.org/flags/38.png,94,94,Real Madrid
CF,https://cdn.sofifa.org/24/18/teams/243.png,€95.5M,€565K,2228,89,63,89,63,93,
95,85,81,91,94,76,7,11,15,14,11,88,29,95,77,92,22,85,95,96,83,94,23,91,92,31,80,8
5,88,89.0,53.0,62.0,91.0,82.0,20801,89.0,61.0,53.0,82.0,62.0,91.0,89.0,92.0,91.0,
66.0,ST LW ,89.0,61.0,53.0,82.0,62.0,91.0,89.0,92.0,91.0,66.0,92.0
1,L.
```

Les données structurées

- Excel
- réseaux GraphViz (dot)



```
digraph "unix" {
        graph [ fontname = "Helvetica-Oblique",
                fontsize = 36,
                label = "\n\n\n\nObject Oriented Graphs\nStephen North, 3/19/93",
                size = "6,6" ];
        node [  shape = polygon,
                sides = 4,
                distortion = "0.0",
                orientation = "0.0",
                skew = "0.0",
                color = white,
                style = filled,
                fontname = "Helvetica-Outline" ];
        "5th Edition" [sides=9, distortion="0.936354", orientation=28, skew="-0.126818", color=salmon2];
```

de graphviz.org

Les données semi-structurées

- XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<breakfast_menu>
<food>
    <name>Belgian Waffles</name>
    <price>$5.95</price>
    <description>
   Two of our famous Belgian Waffles with plenty of real maple syrup
   </description>
    <calories>650</calories>
</food>
<food>
    <name>Strawberry Belgian Waffles</name>
    <price>$7.95</price>
    <description>
    Light Belgian waffles covered with strawberries and whipped cream
    </description>
    <calories>900</calories>
</food>
```

de w3schools.com

Les données semi-structurées

- JSON semblable à XML mais plus simple

```
{
    "glossary": {
        "title": "example glossary",
                "GlossDiv": {
            "title": "S",
                        "GlossList": {
                "GlossEntry": {
                    "ID": "SGML",
                                        "SortAs": "SGML",
                                        "GlossTerm": "Standard Generalized Markup Language",
                                        "Acronym": "SGML",
                                        "Abbrev": "ISO 8879:1986",
                                        "GlossDef": {
                        "para": "A meta-markup language, used to create markup languages such as DocBook.",
                                                "GlossSeeAlso": ["GML", "XML"]
                    },
                                        "GlossSee": "markup"
                }
            }
        }
    }
}
```

de json.org

Ce ne sont que quelques exemples

- On peut cependant remarquer la multiplicité des données
- Leur hétérogénéité
- Parfois des données incomplètes

Ce que nous ne ferons pas

- Analyse en composantes principales
- Algorithmes simples d'apprentissage
  - la classification supervisée
  - la régression
  - le regroupement (clustering)

Exemple prédire si un passager du Titanic va survivre ou non

- TP dessus mais pas sur la prédiction

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...

de w3schools.com

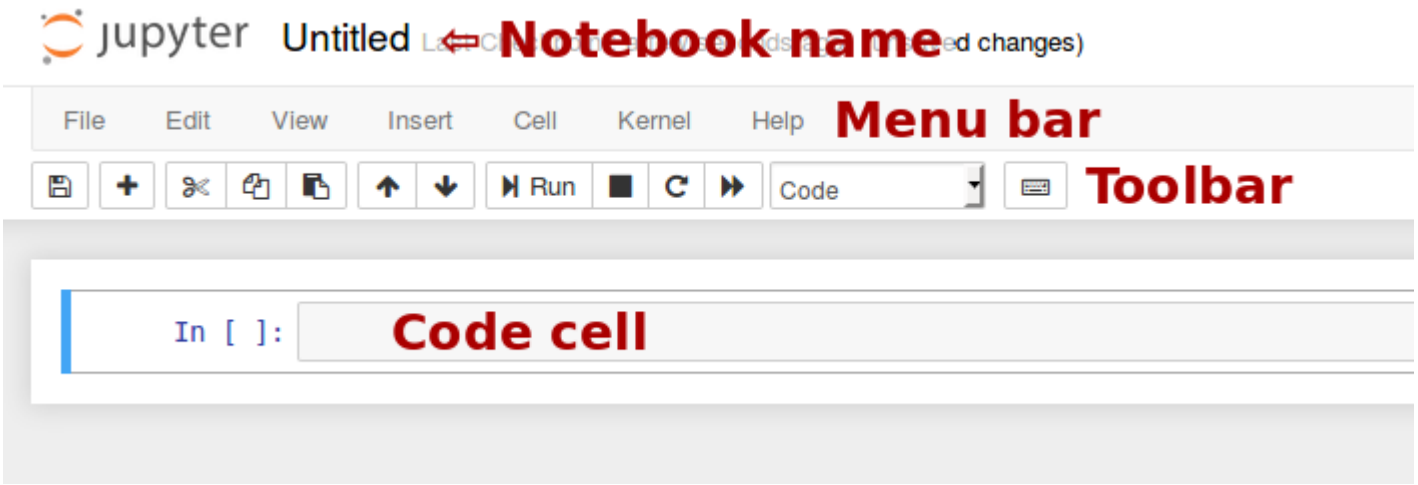
Notre environnement de travail

- Jupyter notebooks
- Julia Python R
- https://jupyter.org/

Étend la console vers le web

- Permet d'éditer du code dans le navigateur
- Exécution du code et résultat dans le navigateur
- Affiche les résultats de plusieurs sources HTML, LaTeX, PNG, SVG, matplotlib…
- Formatage du etxte en utilisant le langage markdown
- Formules mathématiques manipulées nativement grâce à MathJax

L'interface des notebook



Description

- Nom du notebook → sauvegardé automatiquement au format .ipynb
- Les cellules :
  - du code (pour nous en python) qui dépend du Kernel → shift + Enter pour l'exécution
  - du code Markdown pour la présentation
  - des cellules de texte non évaluées

Du code

- voir le notebook Démonstration notebook

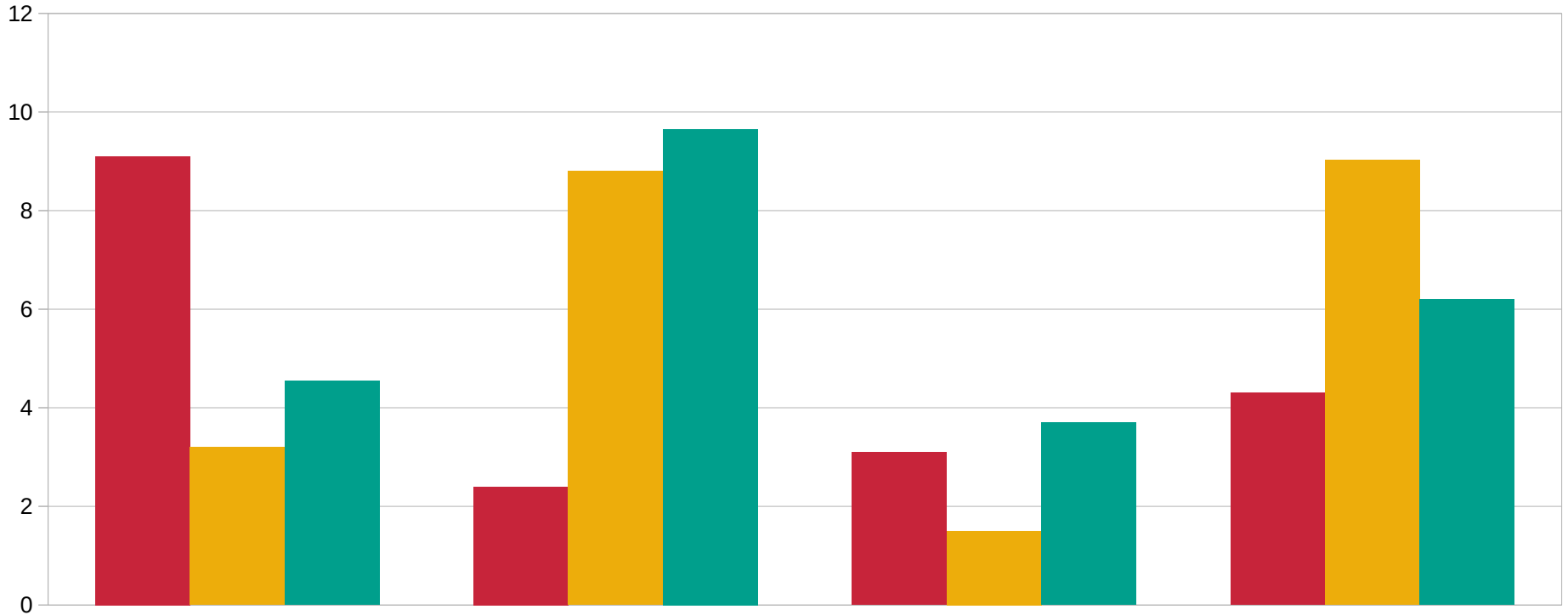
et H

Markdown

- Sur-ensemble de HTML
- Permet de faire des présentations, d'écrire du texte, d'inclure des images…
- https://daringfireball.net/projects/markdown/

Markdown (suite)

- Un lien s' écrit comme ça [Google](http://google.fr)
- On peut mettre également des images ![alt text](/path/to/img.jpg "Title")
- *cf* notebook Démonstration notebook


12
10
8
6
4
2
0