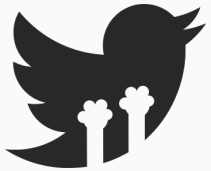


Sample.Cat Project

Comment Twitter reflète les sentiments d'une population en état de choc ?

Mathieu Gaborit

PSES-HSF – 3 juillet 2016

Contenu

1. Remember...
2. Qui est ”on” ?
3. TimeLine & Défis passés
4. La suite : Machine Learning
5. Un peu de socio
6. We have cookies

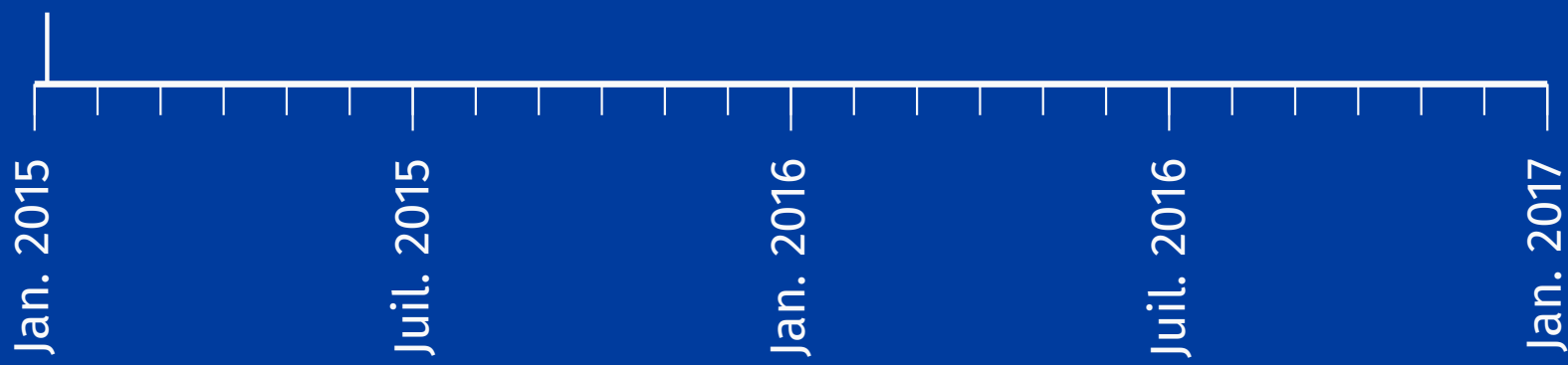
Celui qui ~~chante~~ cause

Mathieu Gaborit

- Padawan physicien
- Idéaliste pythoniste aimant les données
- Intéressé par… un peu tout

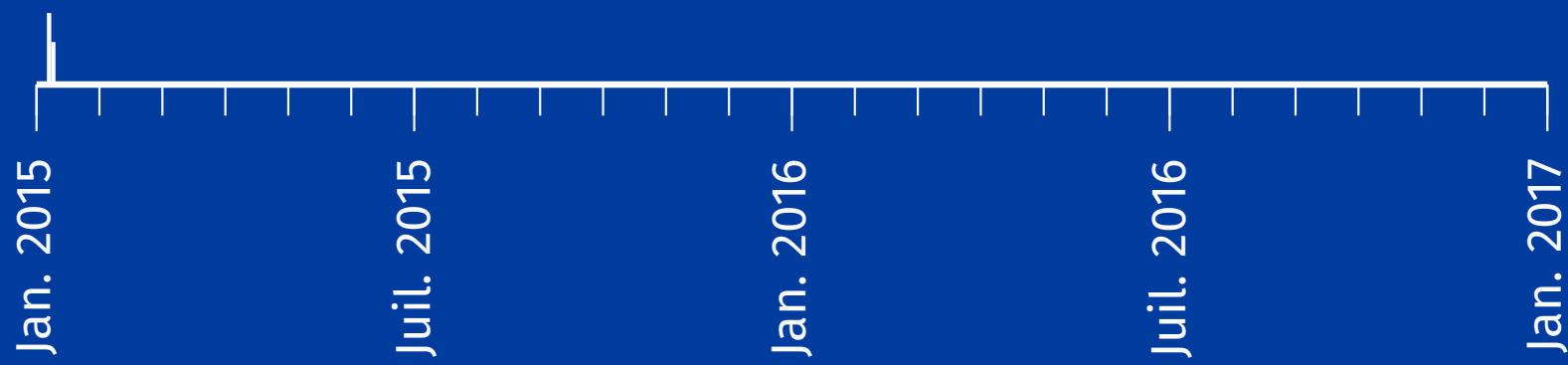
Remember...

Souvenez vous…



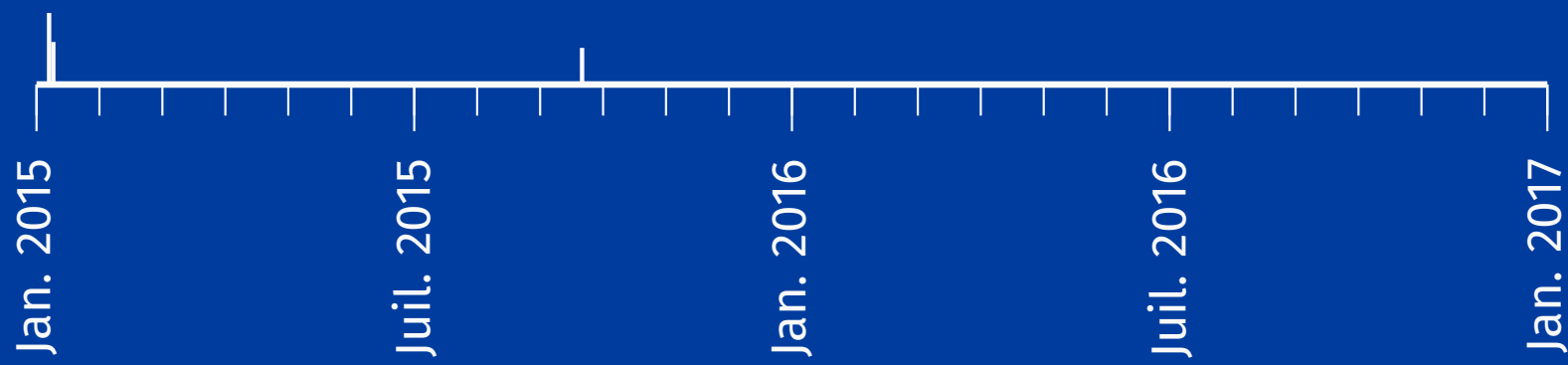
Charlie Hebdo: 15 personnes

Source Wikipédia



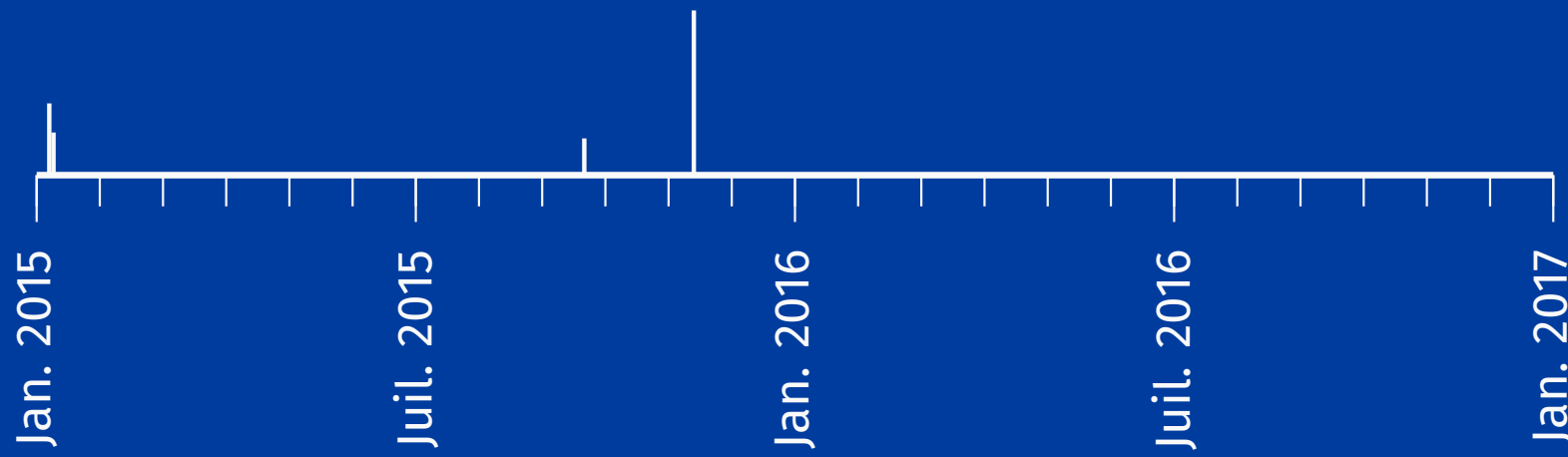
Hyper Casher: 5 personnes

Source Wikipédia



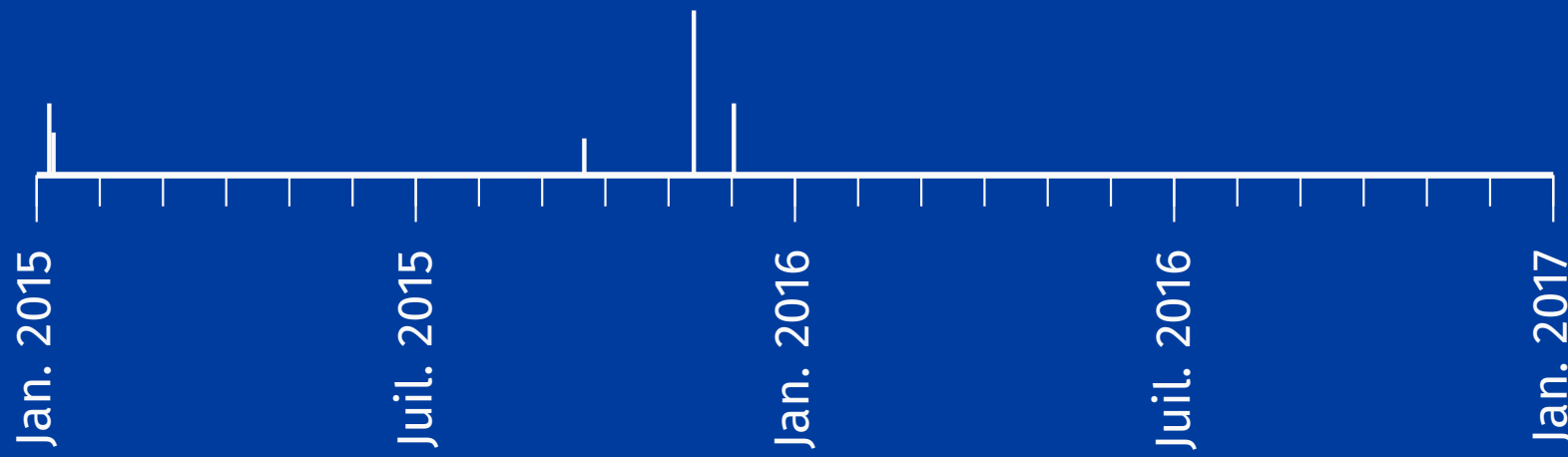
Thalys: 4 personnes

Source Wikipédia



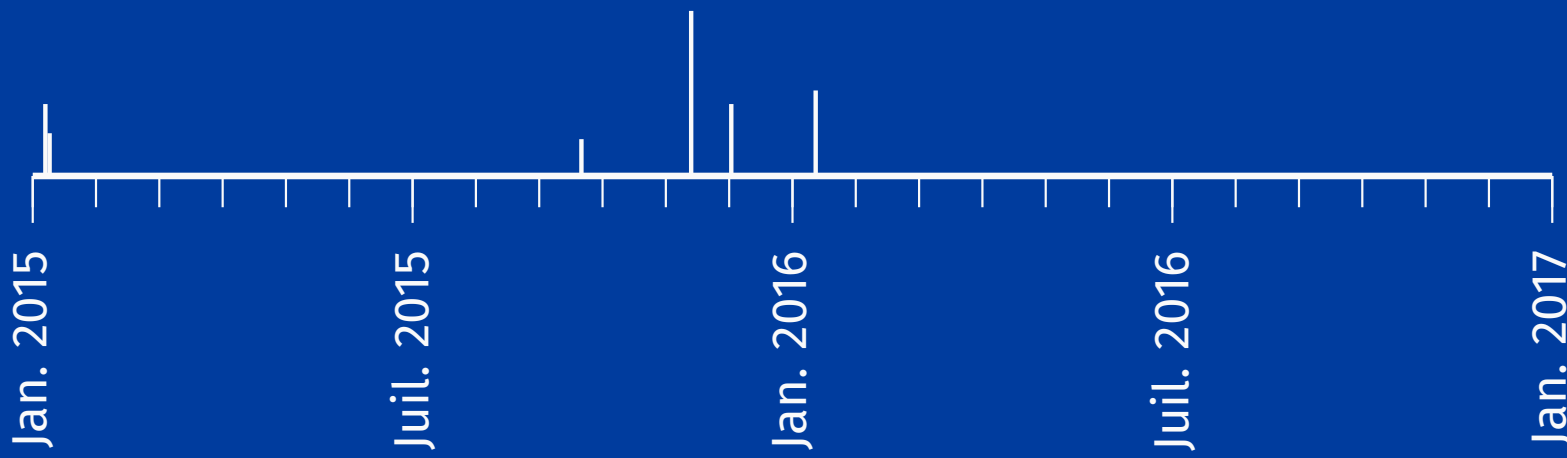
Paris Attacks: 500 personnes

Source Wikipédia



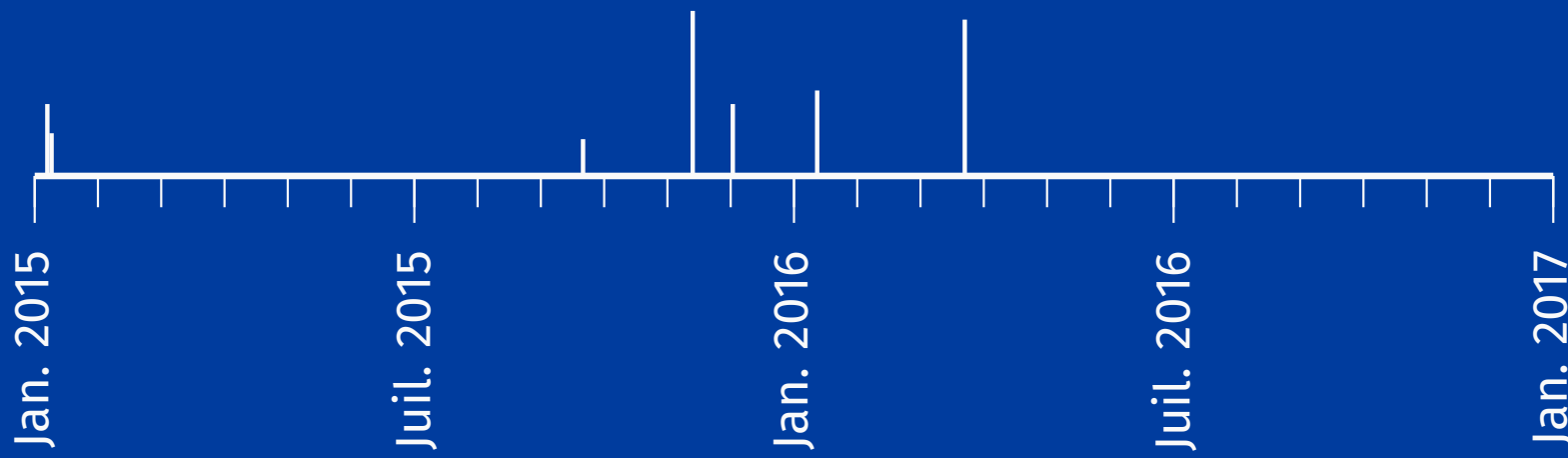
San Bernardino: 15 personnes

Source Wikipédia



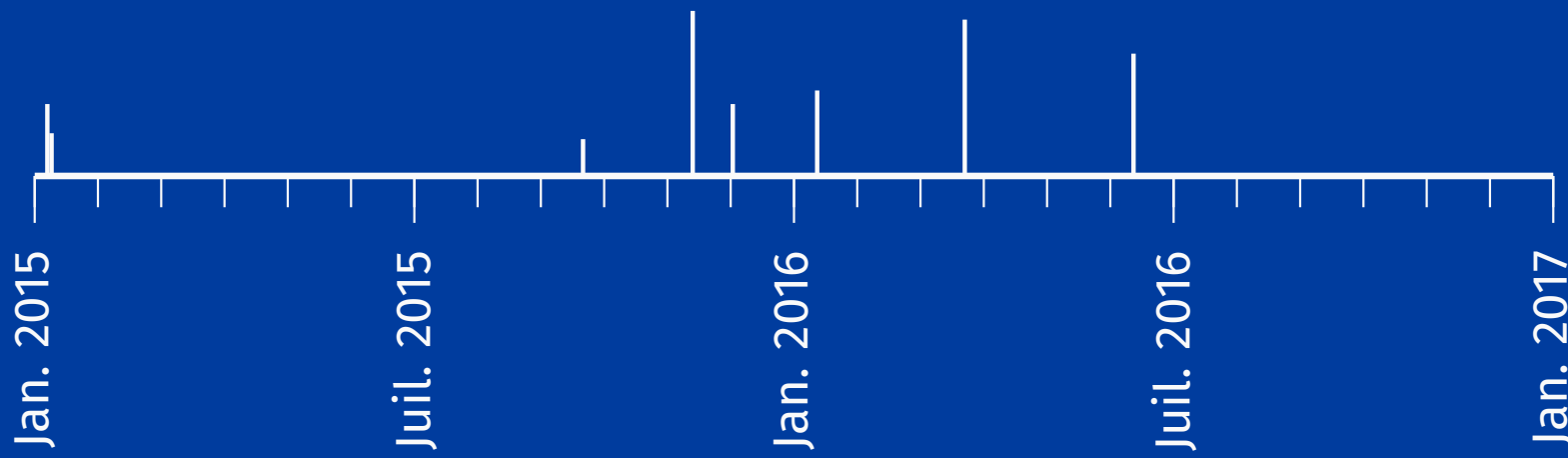
Istanbul: 25 personnes

Source Wikipédia



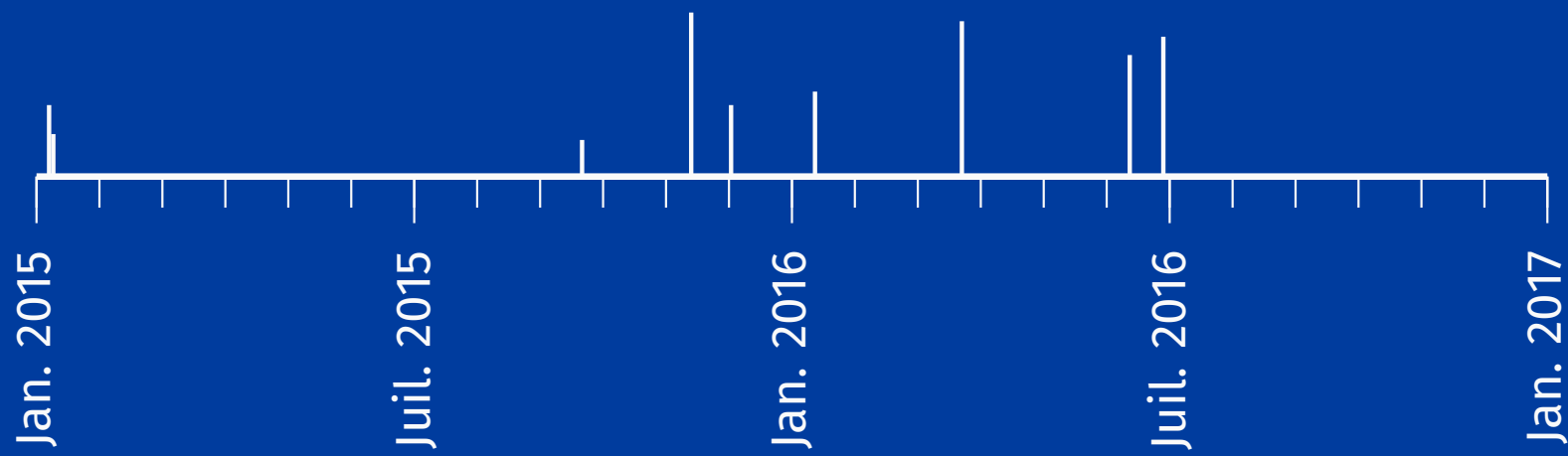
Brussels: 360 personnes

Source Wikipédia



Orlando: 100 personnes

Source Wikipédia



Istanbul (Atatürk) : 200 personnes

Source Wikipédia

Things that shape you...

Beaucoup de crises sociales depuis quelques années :

- nouvelle montée d'un "terrorisme d'état"
- retour des questions d'identité nationale
- toujours moins d'accord entre population et gouvernement
- propagation de l'info toujours plus rapide (stress partagé)

Things that shape you...

Beaucoup de crises sociales depuis quelques années :

- nouvelle montée d'un "terrorisme d'état"
- retour des questions d'identité nationale
- toujours moins d'accord entre population et gouvernement
- propagation de l'info toujours plus rapide (stress partagé)

Hypothèse : ces crises sociales modifient
la perception du monde et les réactions futures

Problème à deux échelles

Problème à deux échelles

A l'échelle d'une personne

- analyse psychologique
- réflexion sur le penchant affectif
- lien avec des peurs personnelles
- prise en compte des expériences passées

Problème à deux échelles

A l'échelle d'une personne

- analyse psychologique
- réflexion sur le penchant affectif
- lien avec des peurs personnelles
- prise en compte des expériences passées

A l'échelle d'une population

- analyses sociologique & politique
- effets à moyen et long terme
- réflexion sur l'appartenance à des groupes, des communautés
- problème statistique

Comment mesurer ?

Plusieurs approches et en particulier :

Comment mesurer ?

Plusieurs approches et en particulier :

Sondage classique/internet

Sympa pour...

- la mise en oeuvre
- le contrôle du volume

Mais pas pour...

- la représentativité
- l'impartialité des questions

Comment mesurer ?

Plusieurs approches et en particulier :

Sondage classique/internet

Sympa pour...

- la mise en oeuvre
- le contrôle du volume

Mais pas pour...

- la représentativité
- l'impartialité des questions

Data-mining sur des réseaux sociaux

Pas toujours idéal à cause...

- du défi technique
- de la définition des marqueurs

Mais...

- impartial par essence
- représentativité analysée

Quelles informations peut-on tirer de
l'activité d'une population sur Twitter ?

Qui est ”on” ?

Team – Part 1

Ahmet Aker, PhD

- PhD en NLP
- *Research Fellow* à USFD (Sheffield, UK)
- Intéressé par l'extraction d'informations et la classification

Baekkwan Park, PhD

- PhD en Sciences Politiques
- Chercheur post-doc à MSU (East Lansing, USA)
- Intéressé par les problématiques sociales et NLP

Team – Part 2

Ben Michalski

- Ingénieur logiciel sur des projets web
- *Problem-solver on steroids*
- Intéressé par le côté BigData
  (récup., stockage, manip.)

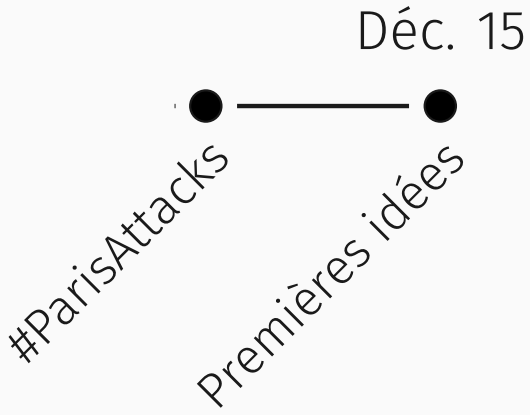
Fred Blain, PhD

- Co-fondateur du HAUM
- Chercheur post-doc à USFD (Sheffield, UK)
- Intéressé par les problématiques NLP/ML

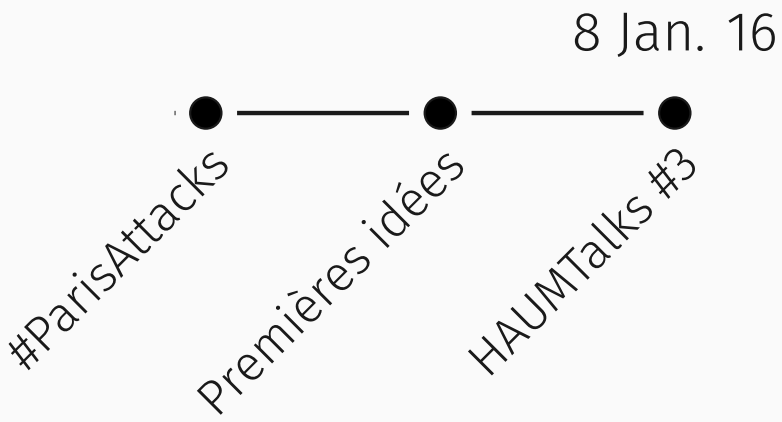
TimeLine & Défis passés

Timeline

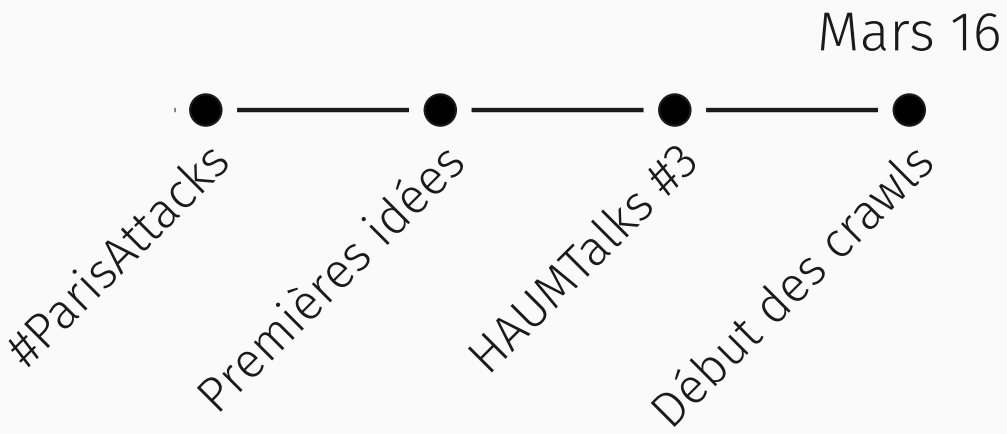
Timeline



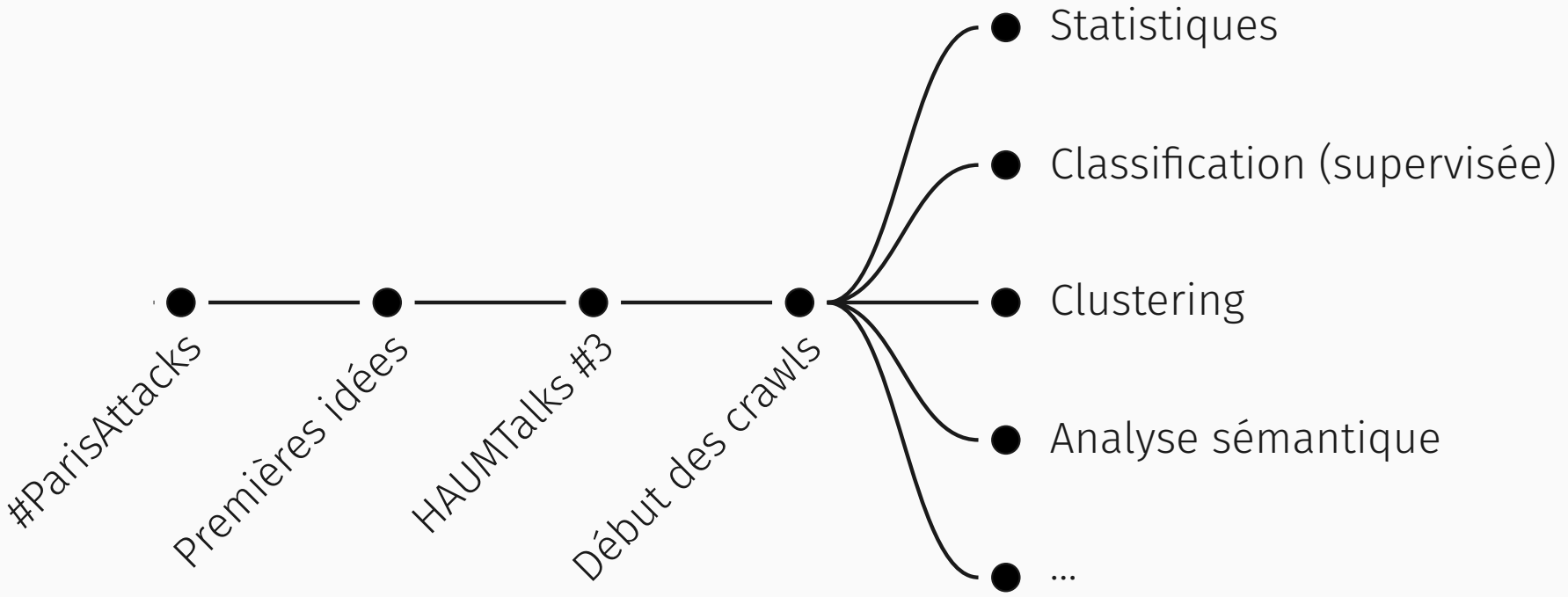
Timeline



Timeline


Mars 16
#ParisAttacks
Premières idées
HAUMTalks #3
Début des crawls

Timeline


#ParisAttacks
Premières idées
HAUMTalks #3
Début des crawls
Statistiques
Classification (supervisée)
Clustering
Analyse sémantique
...

Réduction du champ

2 questions pour l'instant :

1. Peut-on définir un temps caractéristique social en observant Twitter ?
2. Quelle est l'influence des médias sur le sentiment global ?

Réduction du champ

2 questions pour l'instant :

1. Peut-on définir un temps caractéristique social en observant Twitter ?
2. Quelle est l'influence des médias sur le sentiment global ?

Outils pour y répondre :

1. Classification supervisée, Markov Chains
2. Analyse statistique, traitement de signal, NLP
3. Quelques idées farfelues à tester

Prérequis

Statistiques, Machine Learning non supervisé

Besoin de grands jeux de données

Analyse sémantique, Machine Learning supervisé

Besoin de jeux qualifiés et annotés

TimeLine & Défis passés

Récupérer des tweets : défi #1

Des tweets dans le passé

Twitter est l'instantané même...

...la récupération “facile” dans le passé est limitée à 10 jours.

Solution : mettre en place un crawler sauce Michalski !

Des tweets dans le passé

Twitter est l'instantané même...

...la récupération "facile" dans le passé est limitée à 10 jours.

Solution : mettre en place un crawler sauce Michalski !

1. Scrapping web sur la recherche avancée (géoloc) : ∼86kT
2. Enrichissement *via* `/statuses/lookup`
3. Extraction des hashtags : ∼26k#
4. Filtrage (violent, par nombre d'occurences)
5. Scrapping et enrichissement
   jour par jour, hashtag par hashtag

Des tweets dans le passé

Twitter est l'instantané même...

...la récupération "facile" dans le passé est limitée à 10 jours.

Solution : mettre en place un crawler sauce Michalski !

1. Scrapping web sur la recherche avancée (géoloc) : ∼86kT
2. Enrichissement *via* `/statuses/lookup`
3. Extraction des hashtags : ∼26k#
4. Filtrage (violent, par nombre d'occurences)
5. Scrapping et enrichissement
   jour par jour, hashtag par hashtag

Autre solution (payante) : gnip.com

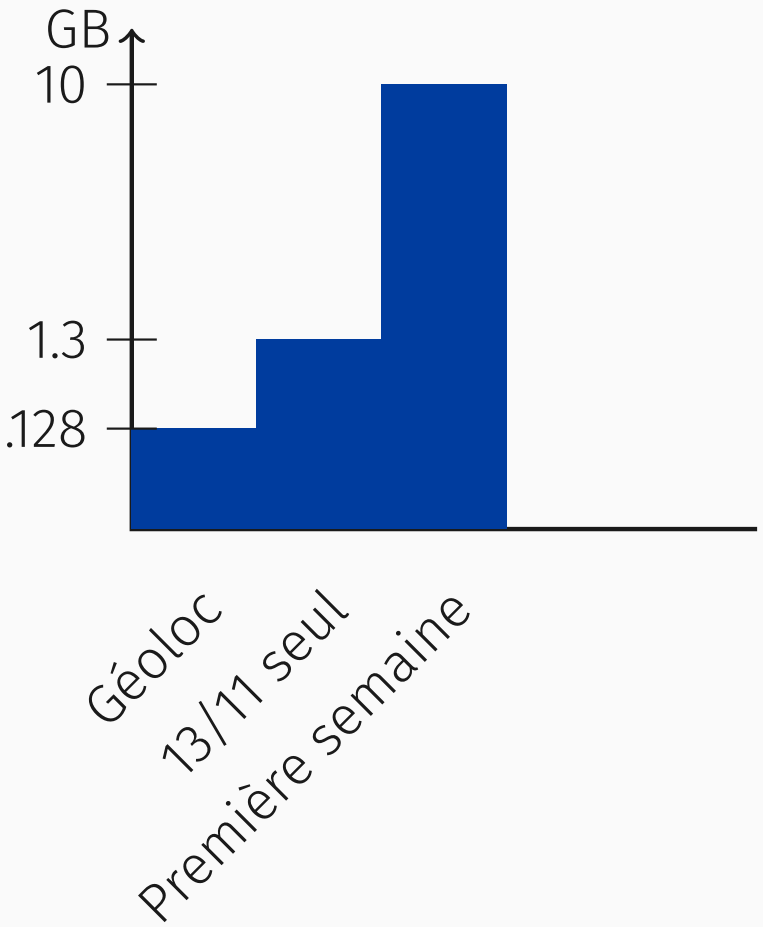
Résultat des courses

Bonus : Travailler *off-line* assure la complétude des conversations.

Résultat des courses

Bonus : Travailler *off-line* assure la complétude des conversations.

```
$ du -h --classe-ça-correctement
```



TimeLine & Défis passés

Étude préliminaire : dynamique

Qu'apporte une analyse quantitative ?

Dynamique $\Leftrightarrow$ Évolution d'un système

Il y a plein de modèles pour étudier les systèmes dynamiques.

En particulier lorsqu'ils sont linéaires et *smooth*.

Qu'apporte une analyse quantitative ?

Dynamique $\Leftrightarrow$ Évolution d'un système

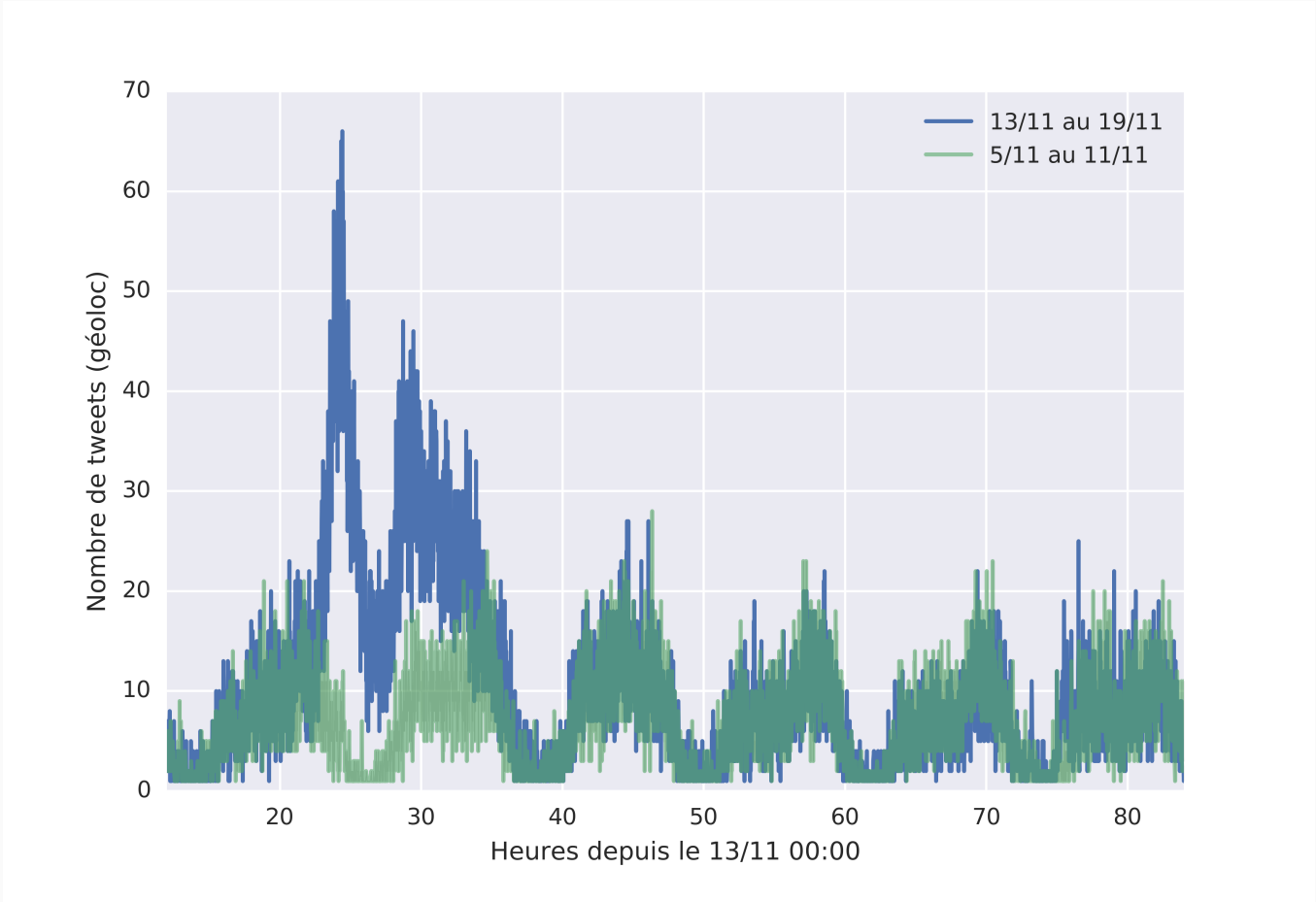
Il y a plein de modèles pour étudier les systèmes dynamiques.

En particulier lorsqu'ils sont linéaires et *smooth*.

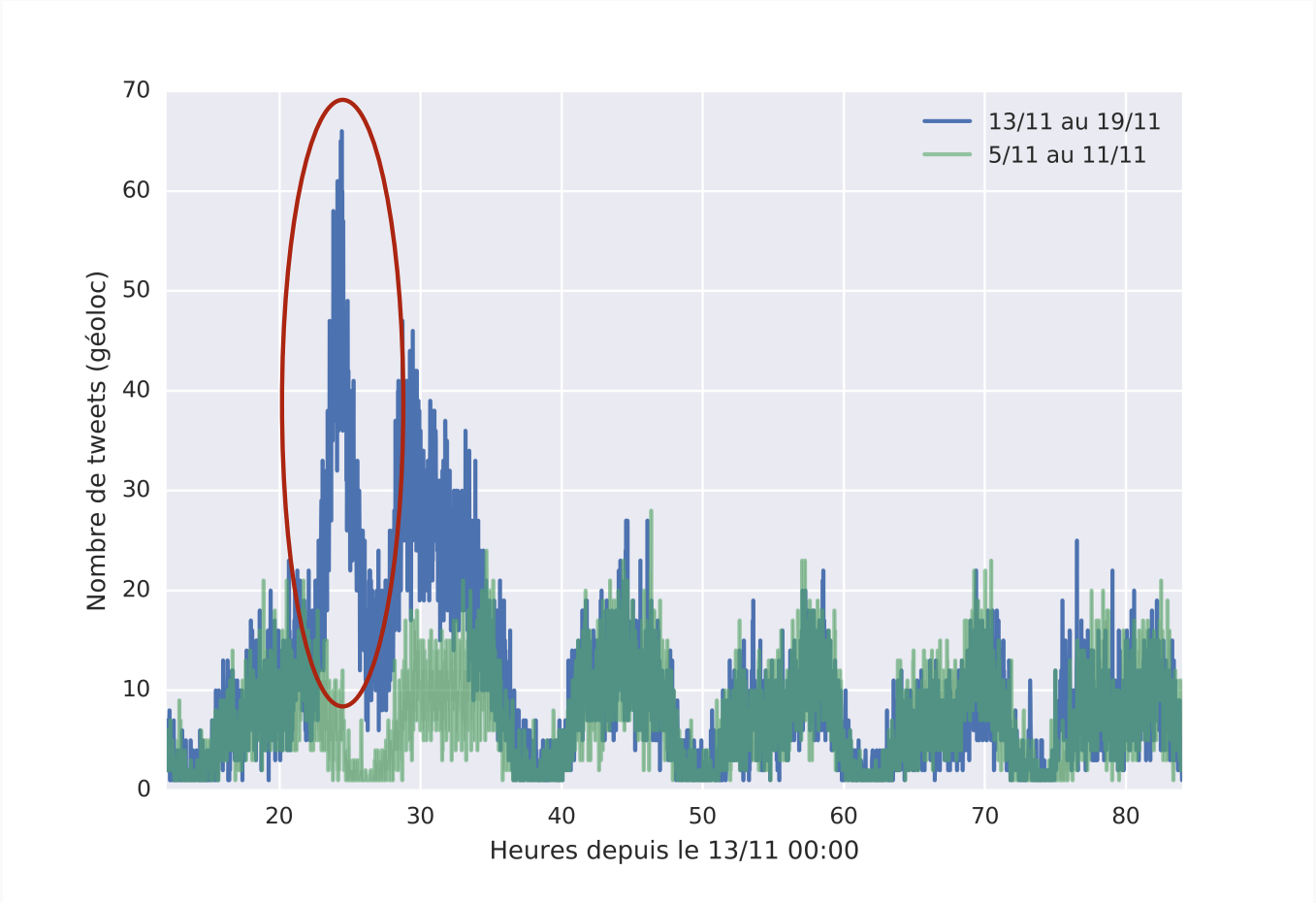
Idée

Représenter le nombres de tweets en fonction du temps et analyser le tout comme une série temporelle.

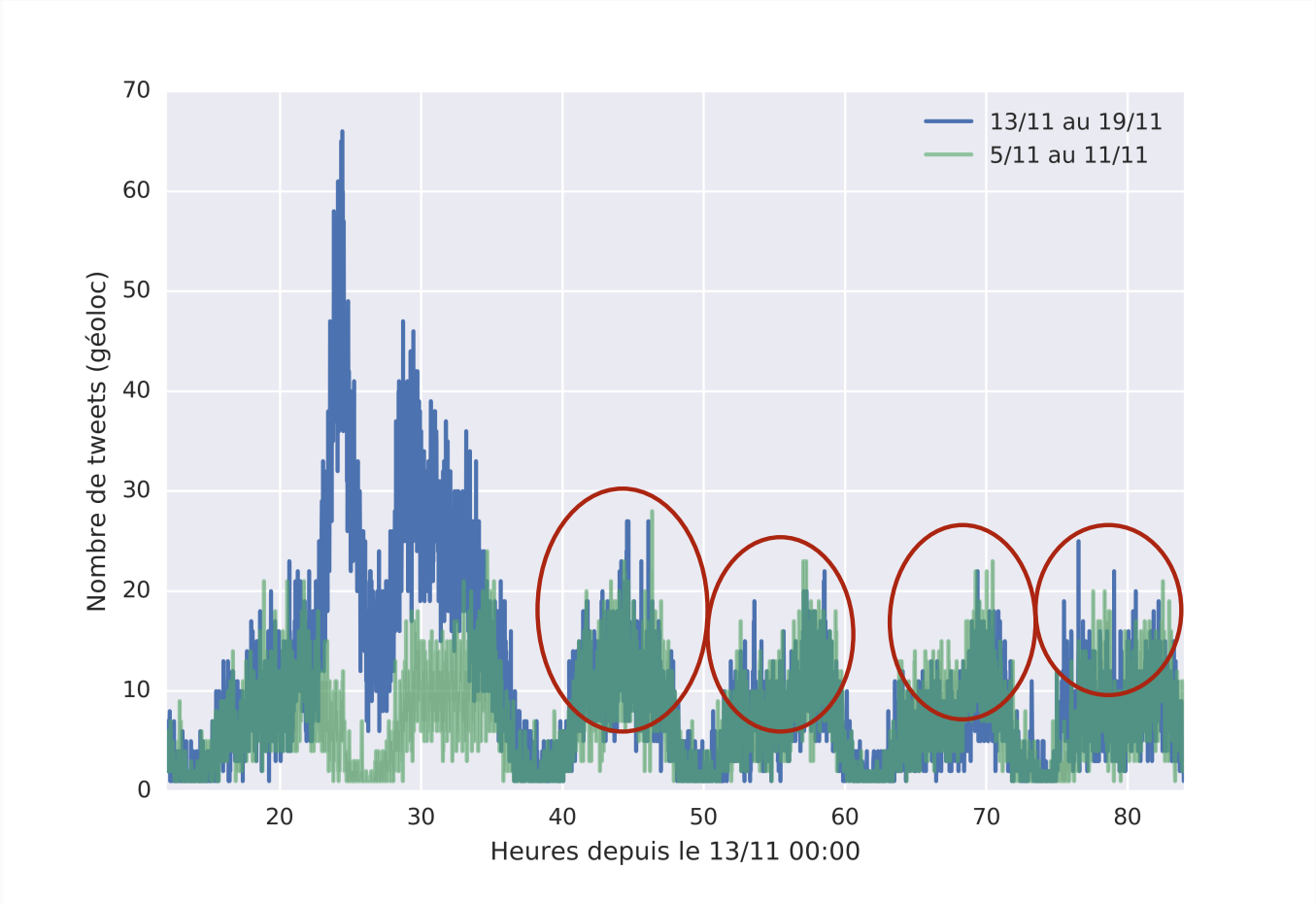
Y a-t-il une modification du motif Jour/Nuit ?



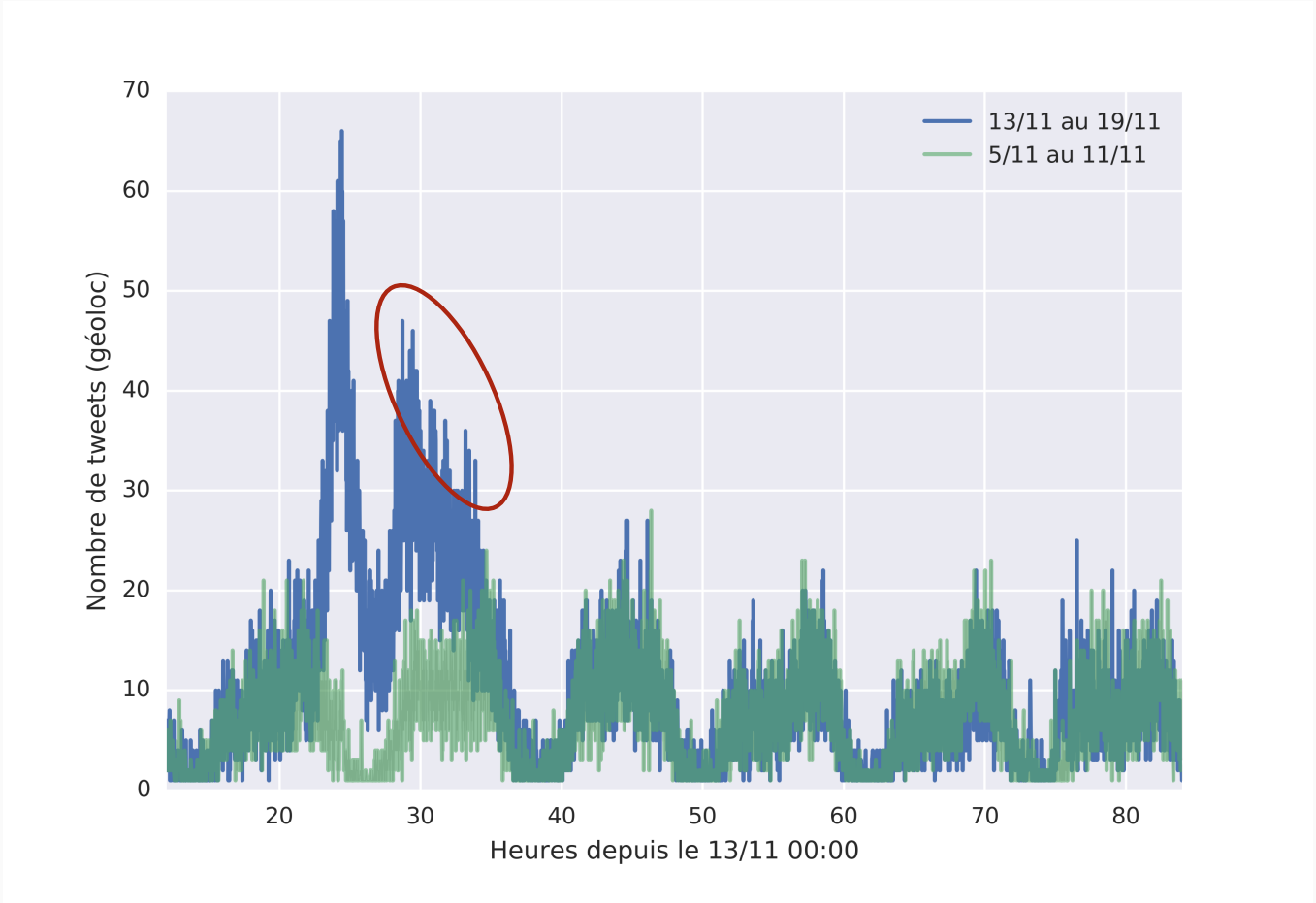
Y a-t-il une modification du motif Jour/Nuit ?



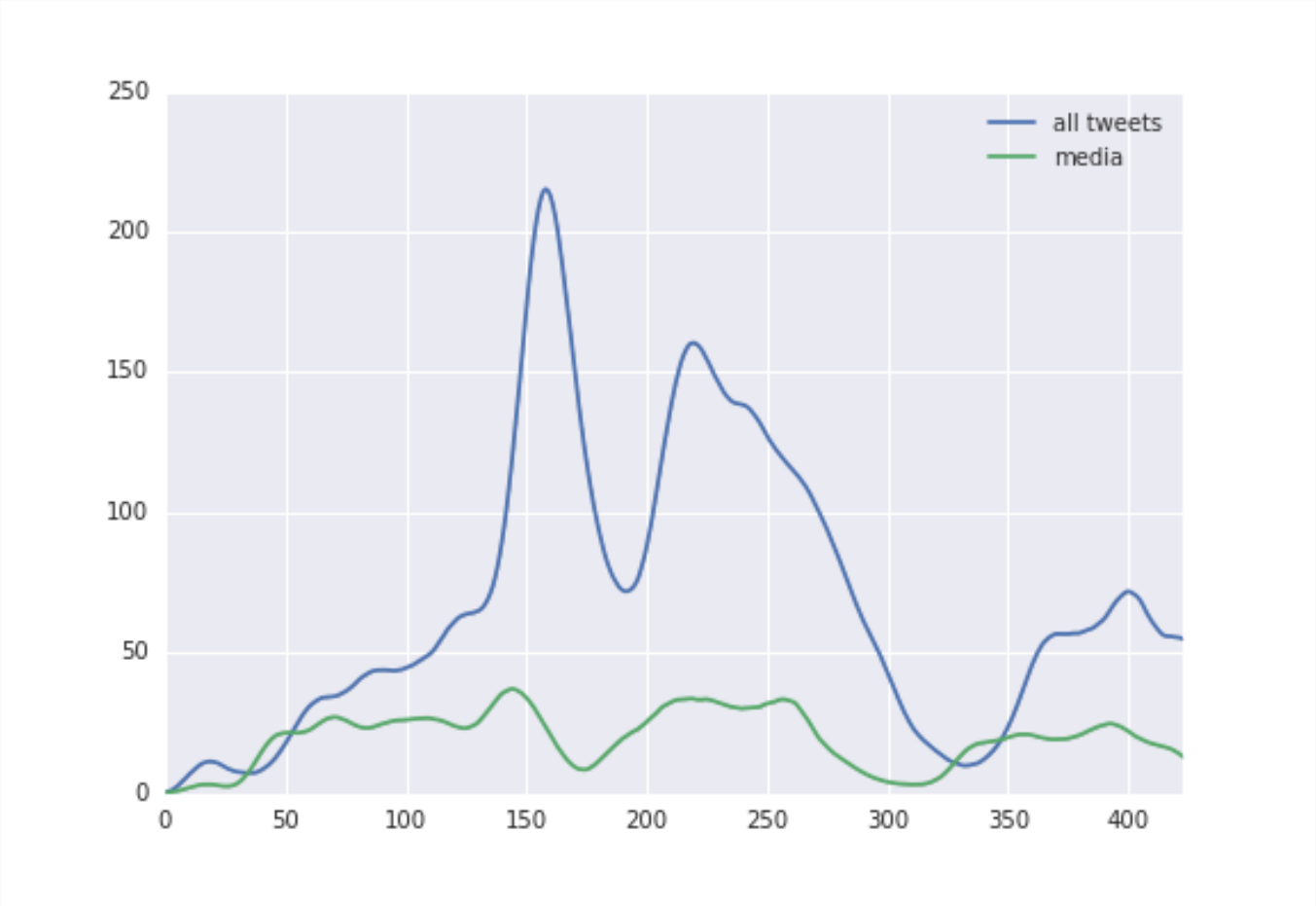
Y a-t-il une modification du motif Jour/Nuit ?



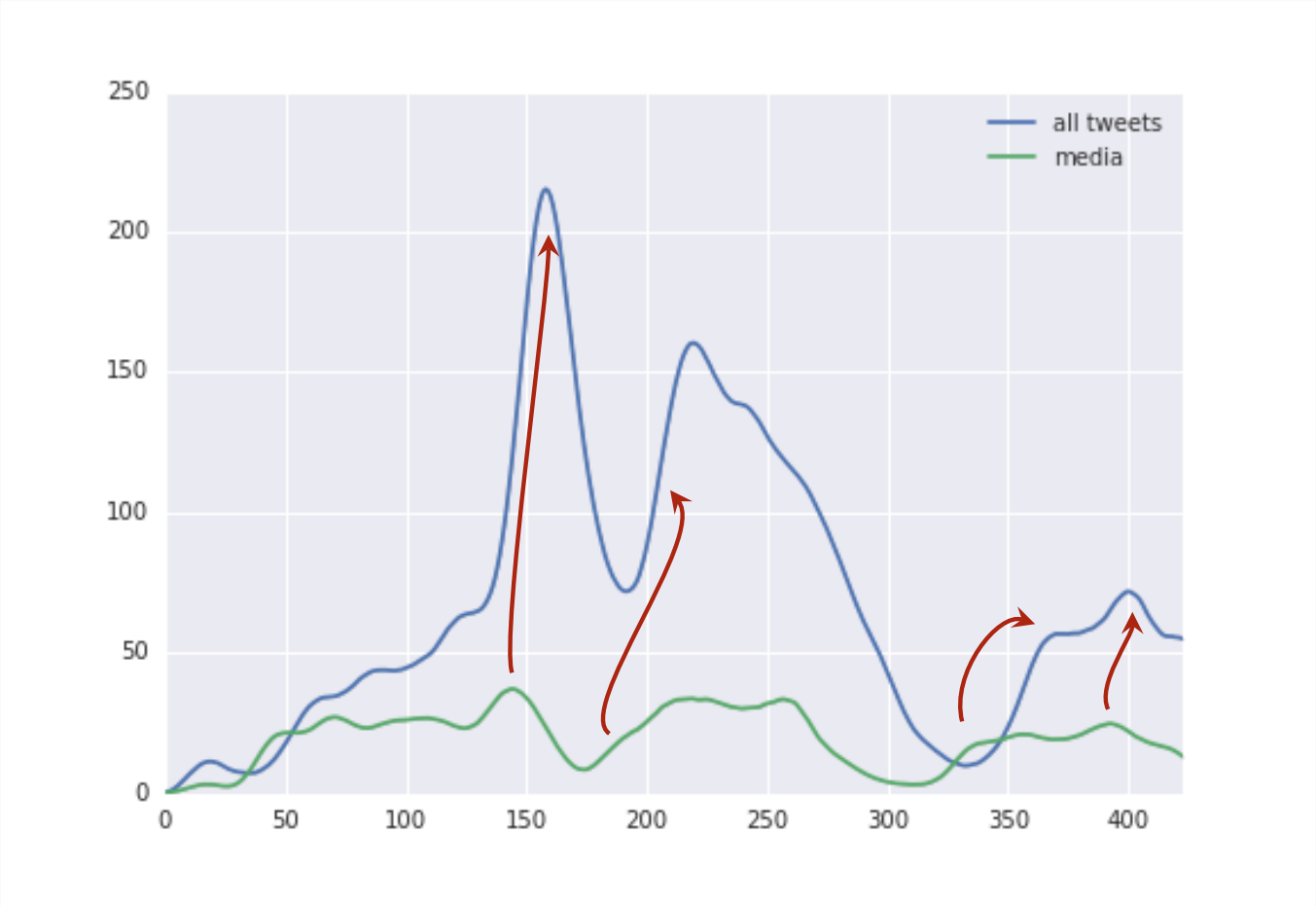
Y a-t-il une modification du motif Jour/Nuit ?



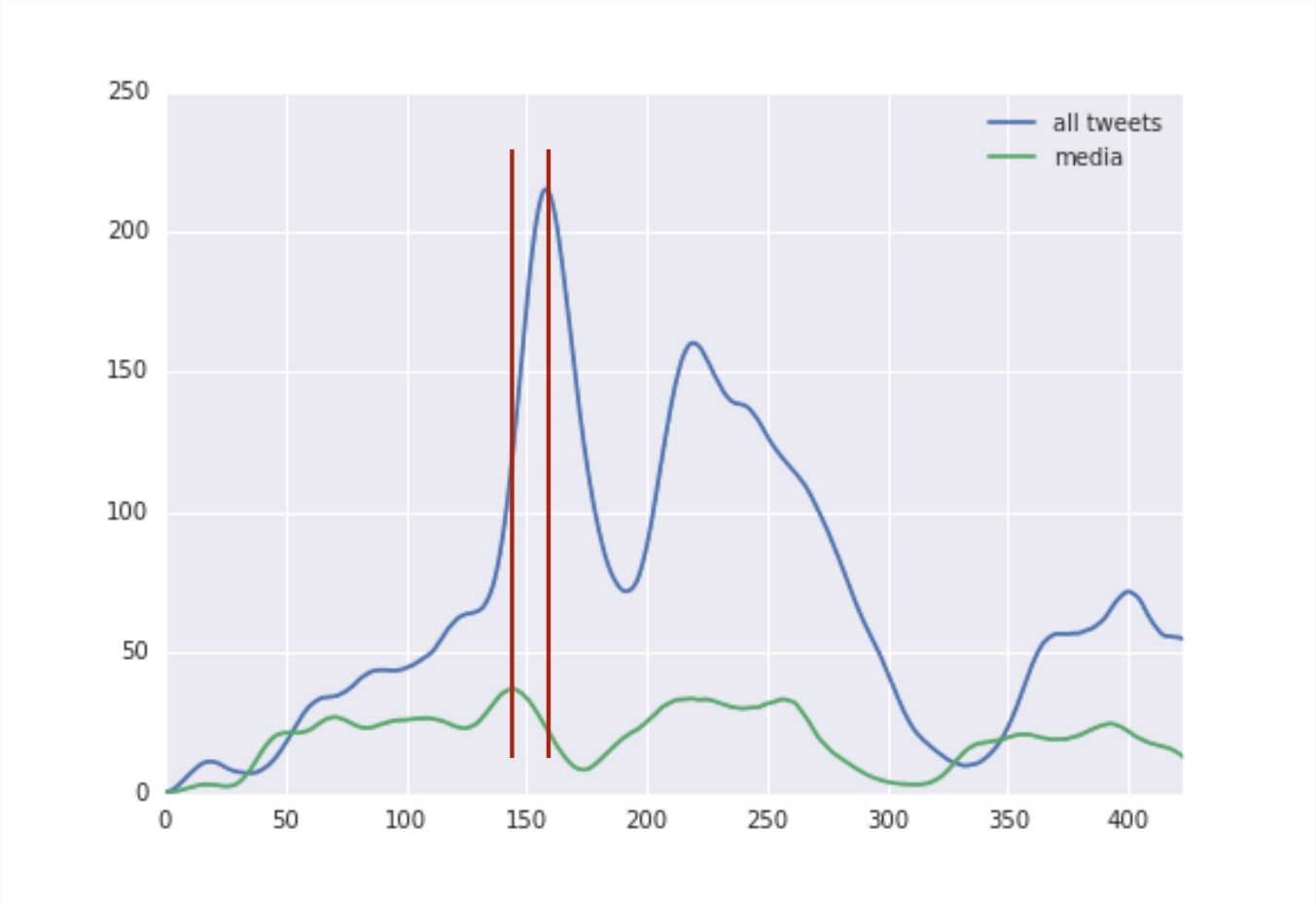
Voit-on une corrélation média/public sur le nombre de tweets ?



Voit-on une corrélation média/public sur le nombre de tweets ?



Voit-on une corrélation média/public sur le nombre de tweets ?



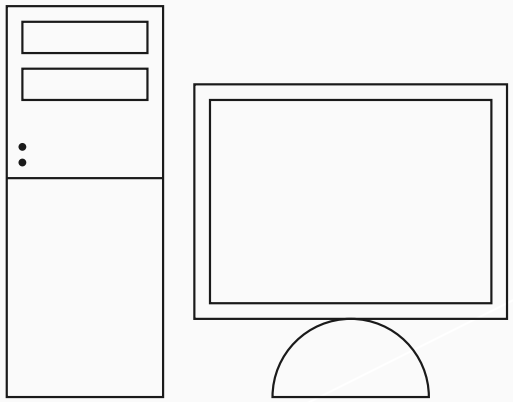
La suite : Machine Learning

La suite : Machine Learning

5min ML crash-course

Deux approches du Machine Learning

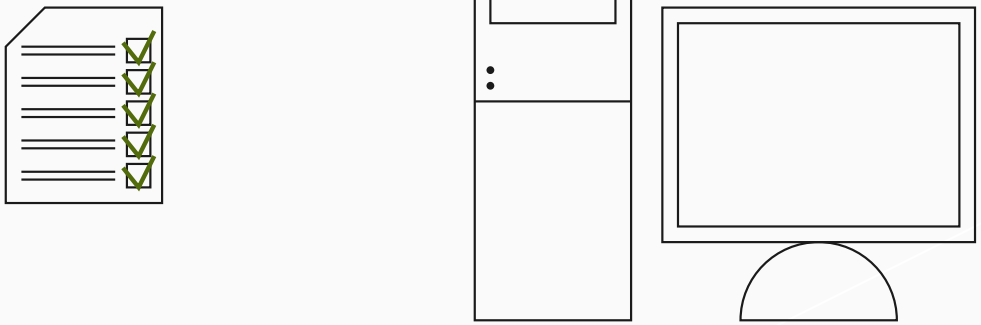
Apprentissage supervisé



Deux approches du Machine Learning

Apprentissage supervisé

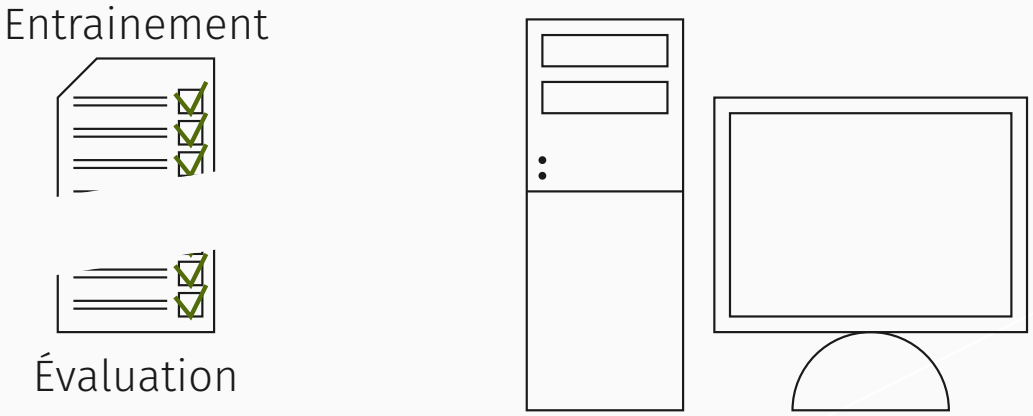
Entraînement sur un jeu entièrement résolu



Deux approches du Machine Learning

Apprentissage supervisé

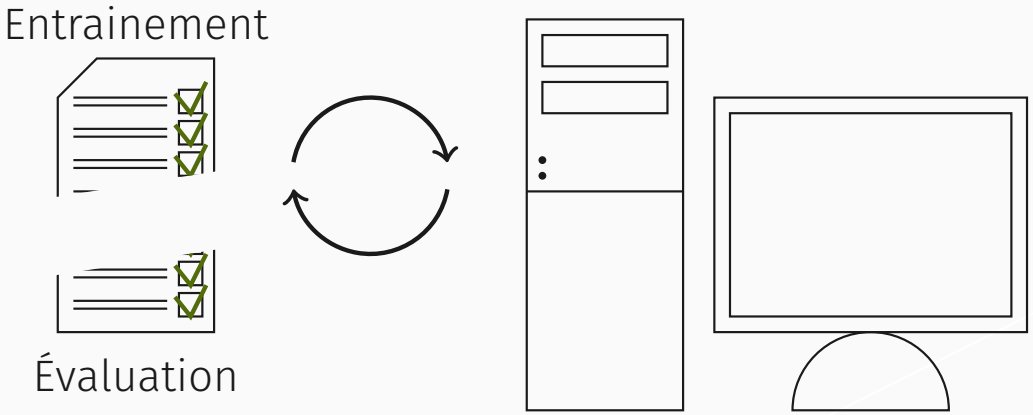
Entraînement sur un jeu entièrement résolu



Deux approches du Machine Learning

Apprentissage supervisé

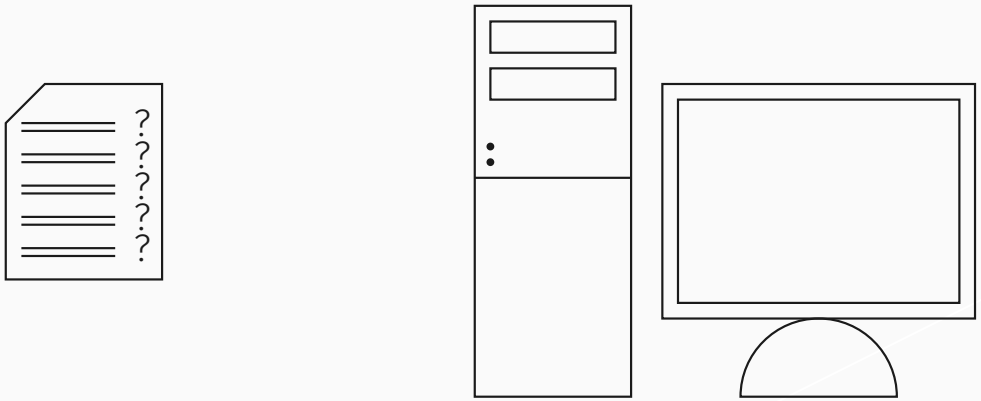
Entraînement sur un jeu entièrement résolu



Deux approches du Machine Learning

Apprentissage supervisé

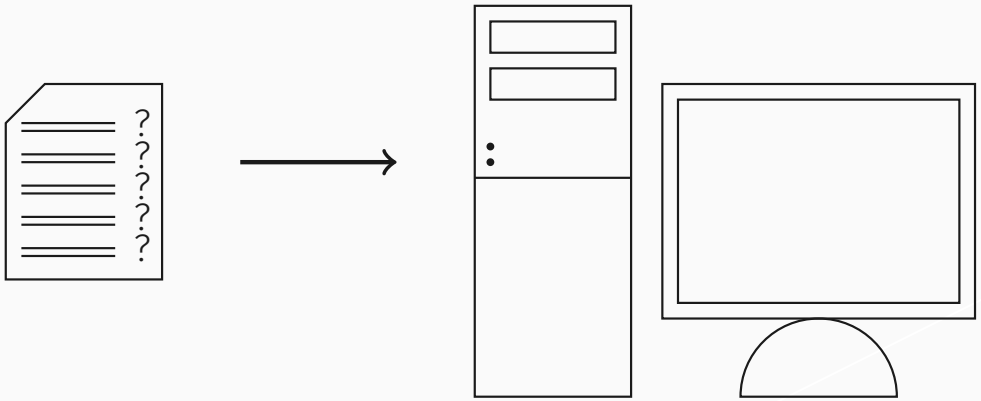
Exécution de la tâche elle-même



Deux approches du Machine Learning

Apprentissage supervisé

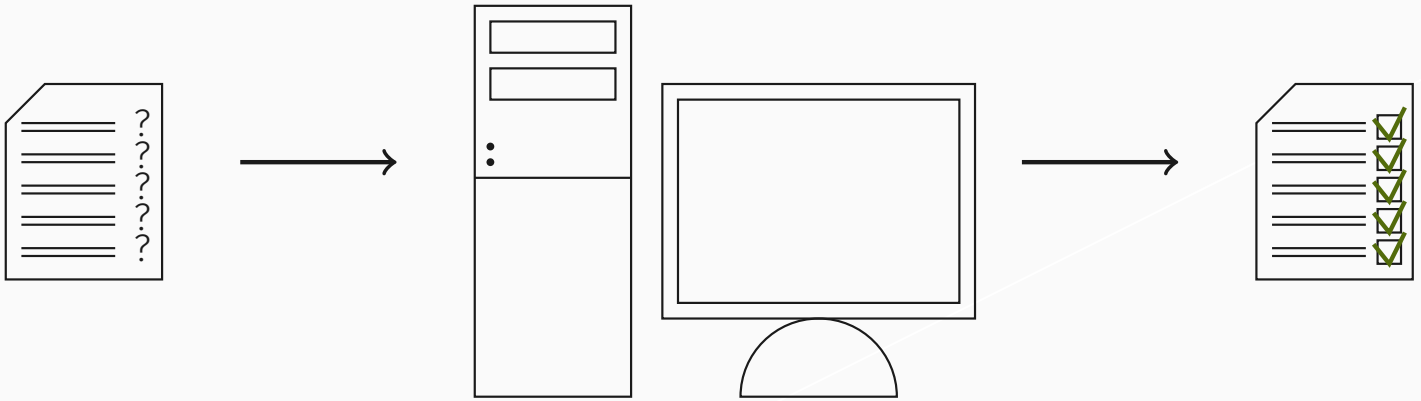
Exécution de la tâche elle-même



Deux approches du Machine Learning

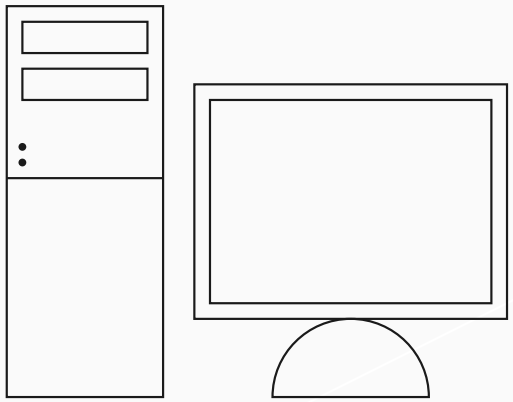
Apprentissage supervisé

Exécution de la tâche elle-même



Deux approches du Machine Learning

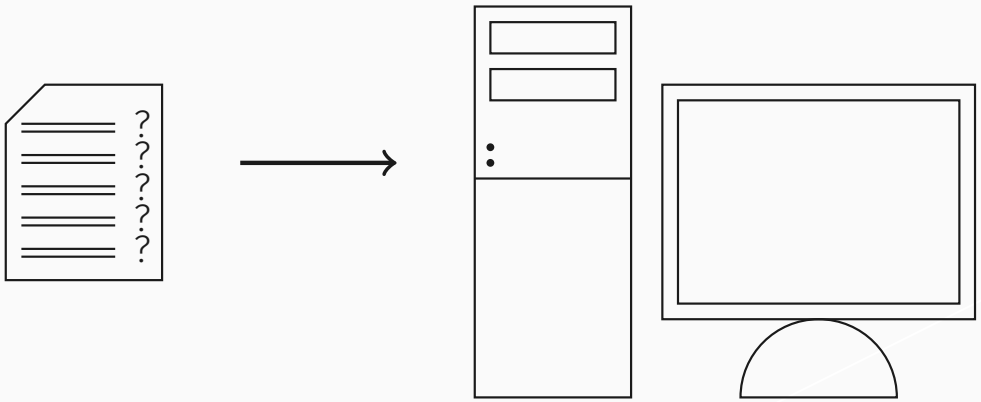
Apprentissage non supervisé



Deux approches du Machine Learning

Apprentissage non supervisé

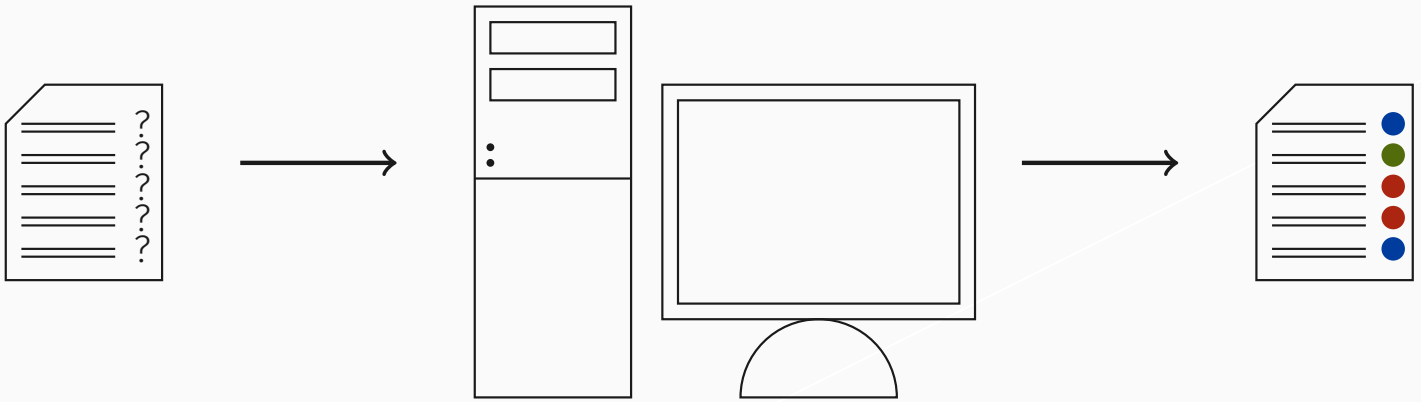
Régression libre



Deux approches du Machine Learning

Apprentissage non supervisé

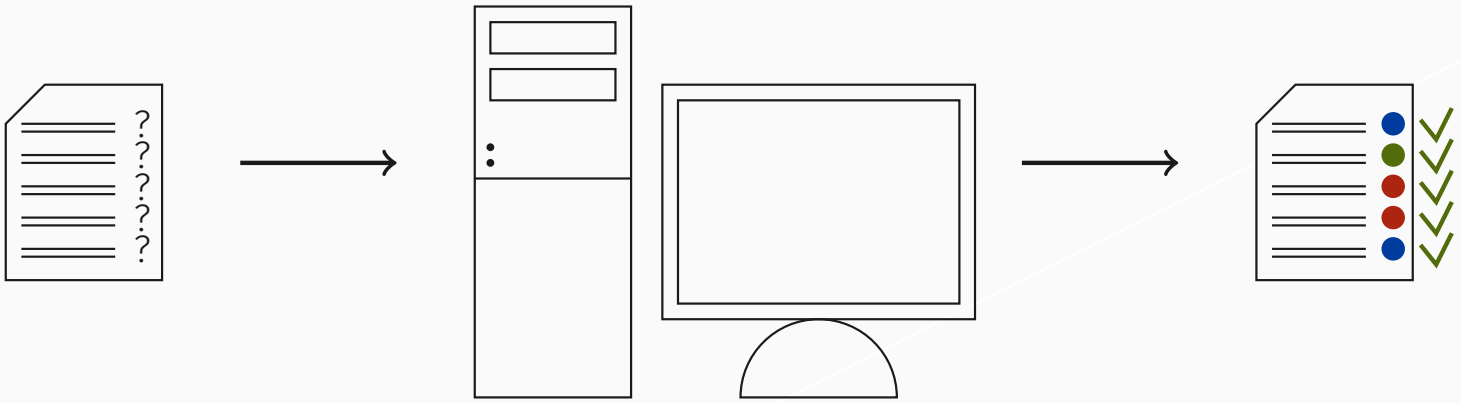
Régression libre



Deux approches du Machine Learning

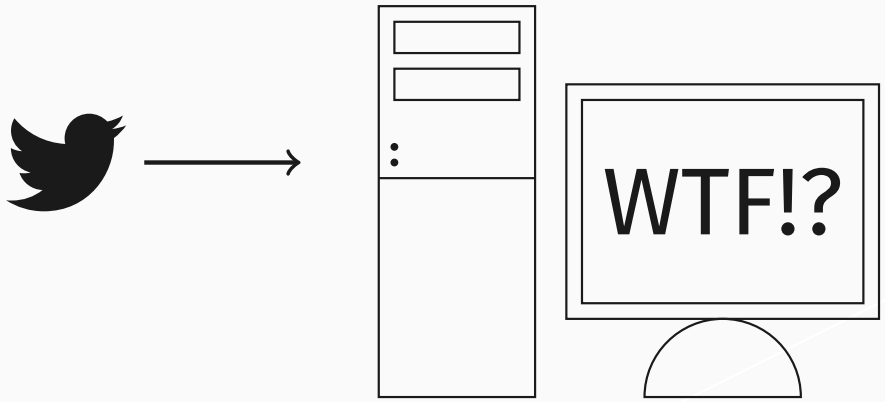
Apprentissage non supervisé

Nommage des classes



Deux approches du Machine Learning

Et pour du texte ?



Non-supervisé sur du texte ? Srly ?

Un ordinateur, ça fait des calculs. Sur des nombres.

Travailler sur du texte implique de le rendre *machine-friendly*.

Non-supervisé sur du texte ? Srly ?

Un ordinateur, ça fait des calculs. Sur des nombres.

Travailler sur du texte implique de le rendre *machine-friendly*.

Many problems, one solution...

Blow up the moon !

Non-supervisé sur du texte ? Srly ?

Un ordinateur, ça fait des calculs. Sur des nombres.

Travailler sur du texte implique de le rendre *machine-friendly*.

Many problems, one solution...

Vectoriser ? Comment ?

Principe :

- identifier certains aspects (*features*)
- transformer chaque élément sémantique en une suite de nombres (*vectors*)
- représenter ces vecteurs dans l'espace et constater comment ils se regroupent

Quelques techniques : *Brown Clusters*, *Bag of Words*, *words embeddings*

La suite : Machine Learning

Applications aux gazouillis

Tweeter n’est pas écrire

Un tweet c’est :

- 140 caractères (très très court)
- informel/condensé
- pas uniquement du texte (emojis, images)
- rarement clair (ironie, sarcasme)
- extrêmement dépendant du contexte

Un tweet est une entitée psychologiquement intéressante :

- courte
- tranchée
- immédiate

Alors comment faire ?

Emojis/Emoticones/Abbréviations

Inclus au lexique et dans les vecteurs. [BRS16]

Dépendance vis-à-vis du contexte

- Réflexion à l'échelle des conversations
- Lien entre les métadonnées du tweet et des données extérieures (activité médiatique, *etc.*)

Ironie/Sarcasme

- Utilisation de jeux spécialisés [Bos+16]
- Recours à des ~~esclaves~~ humains

La suite : Machine Learning

Classification

Classification : objectifs

Sample.Cat $\Leftrightarrow$ *sentiment analysis*

En particulier :

- Identifier le type d'opinion (*sentiment*) pour chaque tweet
- Observer les changements dans l'opinion au cours du temps
- Mettre au jour l'(in)existence d'une dynamique dans le lien média/opinion

Dynamic Topic Modelling

Classification : 2 étapes

Besoin de faire des classes pour trier les données...

...oui, mais lesquelles ?

Classification : 2 étapes

Besoin de faire des classes pour trier les données...

...oui, mais lesquelles ?

Double approche :

1. utiliser une approche non-supervisée pour identifier des ensembles
2. nommer les ensembles
3. mettre en oeuvre un outil d'annotation
4. *se procurer* des annotateurs humains
5. utiliser une approche supervisée pour classifier le jeu

Classification : 2 étapes

Besoin de faire des classes pour trier les données...

...oui, mais lesquelles ?

Double approche :

1. utiliser une approche non-supervisée pour identifier des ensembles
2. nommer les ensembles
3. mettre en oeuvre un outil d'annotation
4. ~~*se procurer* des annotateurs humains~~
5. utiliser une approche supervisée pour classifier le jeu

Prochain défi technique

Implementer un outil de crowdtagging

pas trop biaisé

From Scratch ? Pourquoi ?

- Libre (donc vérifiable)
- Paramètrable (nombre d'annotateurs/item, *etc.*)
- Contrôle des rejets
- Choix des canaux de diffusion (Twitter, *etc.*)
- Pré/post traitement *on-line*

Un peu de socio

Un peu de socio

Intérêt sociologique

Points d'intérêts

Intérêts socios

1. Réaction et type de langage
2. Rapport à l'information

Points d'intérêts

Intérêts socios

1. Réaction et type de langage
2. Rapport à l'information

Sample.Cat $\Leftrightarrow$ plusieurs *millions* de tweets

Intérêts scientifiques

1. Approche *on-line*
2. Libération des ressources

Réactions, langages, information

Réaction et type de langage

- L'étude de la dynamique et de la forme des réactions permet de comprendre l'évolution au cours d'une crise.
- L'immense jeu de données permet de construire des modèles statistiques ayant un sens.

Rapport à l'information

- L'information en direct innonde nos sociétés et particulièrement en temps de crise.
- L'analyse proposée questionne les conséquences d'un accès immmédiat et ininterrompu à l'information.

Temps réel, libération du jeu

Approche *on-line*

- Les métadonnées récupérées permettent de *rejouer l'activité Twitter* de la soirée du 13/11 et de voir si des informations auraient pu être tirées du flux.
- L'approche *online* utilisée parfois comme argument aux écoutes est ainsi mise à l'épreuve.

Libération des ressources

Le projet devrait permettre d'aboutir à un jeu de données annoté immense, en français (habituellement sous-dotté).

Un peu de socio

Un écosystème de recherche

13·11 – Étude d’une attaque

Programme de recherche transdisciplinaire

12	Années de recherche
35	Partenaires
1000	Témoignages vidéos
8	Champs de recherche (de la santé aux mathématiques)
20	Millions d’euros de financement

Comportements sociaux et réseaux

- Des thèses en socio sur l'axe réseaux sociaux
- Beaucoup de recherches sur le *data-mining* appliqué aux médias sociaux
- Réflexion autour du suivi d'opinion(s) et de rumeurs
- Terrain de jeu pour réseaux de neurones

Application à la politique

- Bing Political Index : suivi des opinions pour la présidentielle américaine
- Brexit Watch par W2O Group (basé sur Twitter)

We have cookies

Join the purring side !

On cherche en particulier :

- d'autres amateurs de *machine learning*/*bigdata*
- au moins un autre physicien fou (je me sens seul)
- des gens pour le côté psycho/socio (Baekkwan va se sentir seul)

Interessé(e)s ? Welcome!

meow@sample.cat — #sample.cat @ irc.freenode.net

Des questions ?

mat@sample.cat