

eleven - hackathon solidaire

22 avril 2022



Qui sommes-nous ?



Simon Georges-Kot
Manager data scientist
Ensaë 2013

simon.georges-kot
@eleven-strategy.com



Marie Guegain
Consultant data scientist
Ensaë 2020

marie.guegain
@eleven-strategy.com

AGENDA



1. Présentation d'eleven
2. Présentation du sujet du hackathon
3. Modalités pratiques

ELEVEN AT A GLANCE

2008

FOUNDED IN
PARIS

1 

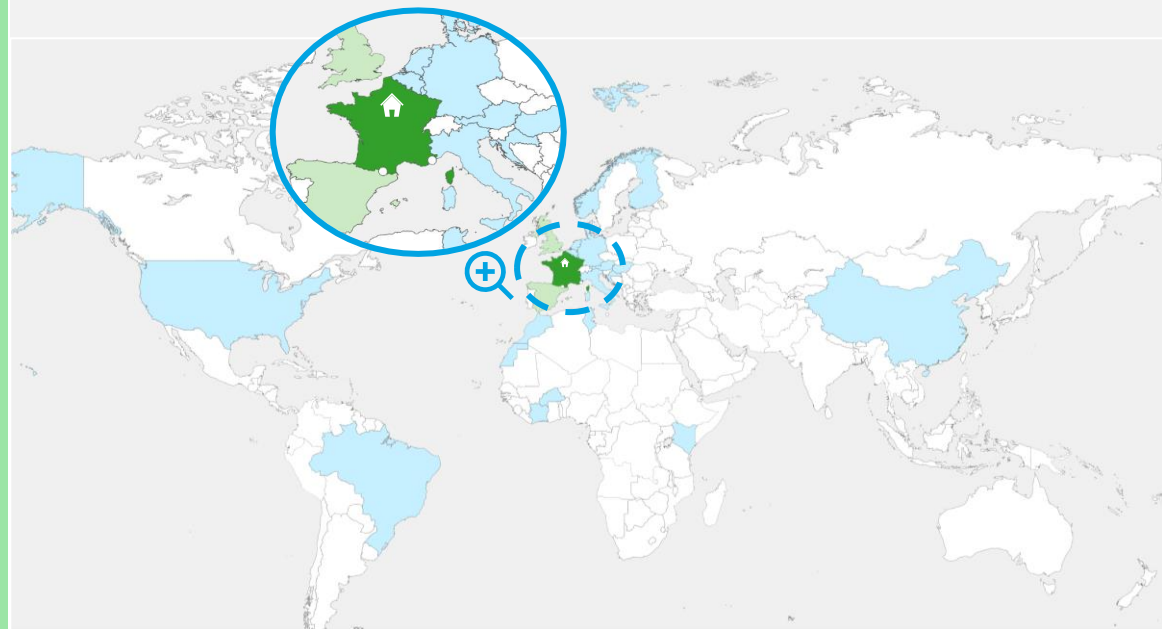
OFFICE AND
EXPANDING

 **50+**
CONSULTANTS

50/50
BUSINESS to
ENGINEERING
BACKGROUND

50/50
CORPORATE vs
PRIVATE EQUITY
CLIENTS

 **>20%**
ANNUAL GROWTH



3 COMPLEMENTARY CLIENT OFFERS



DIGITAL STRATEGY AND ACCELERATION

- How does digital disrupt my industry and business?
- What moves should I make to take advantage of this?
- How can I develop proofs of concept to gain buy-in?
- How can I scale proven concepts into a new business?



DATA AND AI

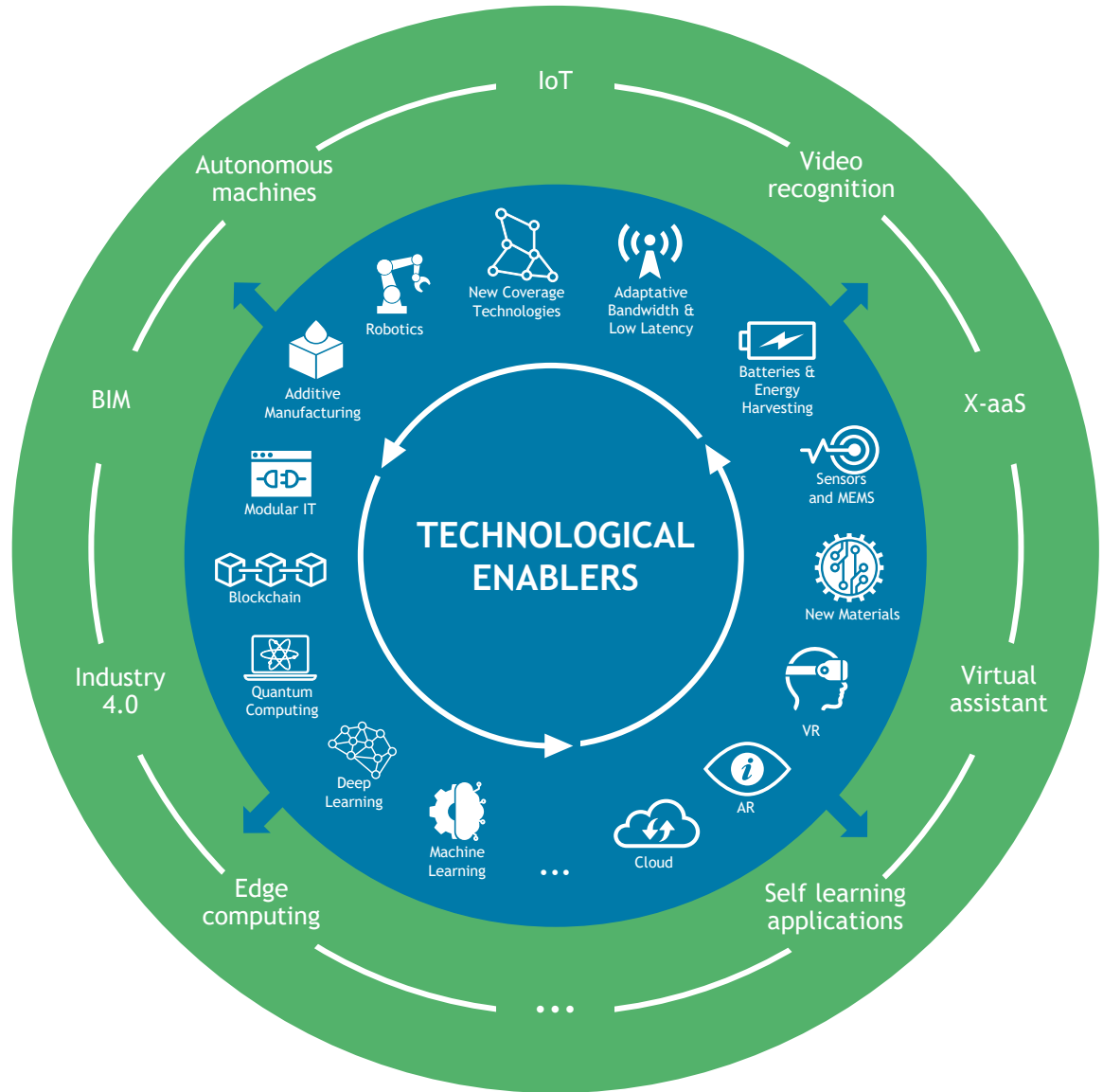
- Is data science and A.I. relevant to my business?
- What A.I. disruptions can I expect?
- How should I respond?



DUE DILIGENCES

- Is my digital-enabled target attractive?
- How can I drive digital-enabled value from my asset?
- What equity story can I tell?
- How best to position my asset for exit?

STRUCTURED AROUND TECHNOLOGICAL DISRUPTIONS



Disruptive convergences

CAC 40



Mid Cap



ADDRESSING
LARGE AND
MEDIUM CAP
CLIENTS ACCROSS
SEVERAL KEY
INDUSTRIES...

Large Cap



CARLYLE

ARDIAN



EURAZEO

...

Mid Cap



ABENEX



andera PARTNERS



naxicap PARTNERS

...

AND LEADING
EUROPEAN LARGE
CAP AND SMID
CAP PRIVATE
EQUITY FUNDS

ELEVEN'S CONSULTANTS ARE AT THE CROSSROADS OF FOUR PROFILES



STRATEGY
CONSULTANTS



ENTREPRENEURS

eleven
strategy consultants
• Paris •



TECH
ENTHUSIASTS



DATA
SCIENTISTS

eleven a accompagné un acteur de l'industrie du luxe dans la conception, l'implémentation et l'industrialisation d'un tableau de bord sur le suivi de la performance digitale de ses marques vs. concurrents



Client

Groupe leader de l'industrie du luxe et cinq de ses marques



Enjeux stratégiques

- Comment mettre en place un système de suivi de la santé digitale des marques du groupe vs. concurrents ?
- Comment anticiper les résultats des rapports des banques internationales qui influencent les décisions des investisseurs ?
- Comment collecter des données sur les concurrents (e.g., à partir de fournisseurs externes via une API) ?



Outils data & digitaux, compétences et méthodologies utilisées

- Collecte, modélisation et audit des données externes
- Industrialisation d'un tableau de bord, de la collecte à la visualisation des données
- Réflexion stratégique axée sur la sélection de KPIs digitaux
- Expertise sectorielle dans l'industrie des fournisseurs de données (search, trafic, social media listening) et du digital dans l'industrie du luxe (influenceurs, stratégie social media, etc.)



Impact

Identification des KPIs digitaux permettant de préempter une éventuelle perte de brand heat



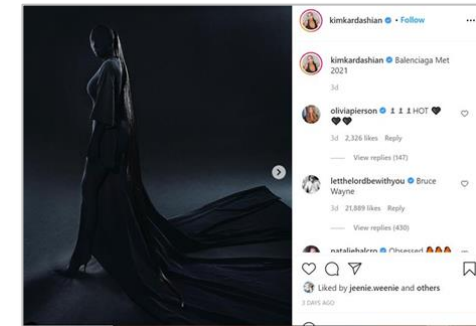
Impacts clés du projet

- ✓ Construction et industrialisation d'un tableau de bord sur la santé digitale des marques
- ✓ Sélection de KPIs digitaux pertinents et de plateformes corrélées avec les performances financières
- ✓ Contrôle de la qualité et de la véracité des données
- ✓ Plan de transition pour une intégration du suivi à 100% chez le client

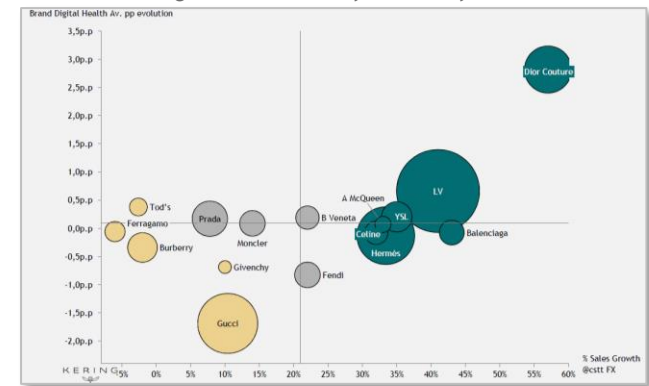


Illustrations

Exemples de plateformes suivies dans le rapport



Brand Digital Health vs Performances financières



As the **examples of two projects demonstrate**, eleven notably distinguishes by its ability to both **design digital and data strategies** and to **effectively lead and execute transformation projects** on behalf of its clients, ensuring an end-to-end continuum from strategy to **implementation**



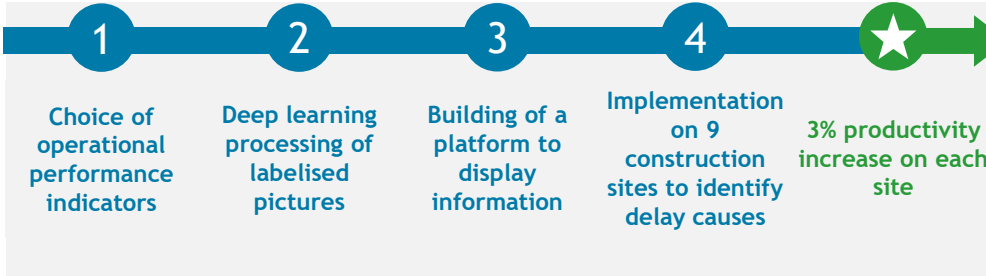
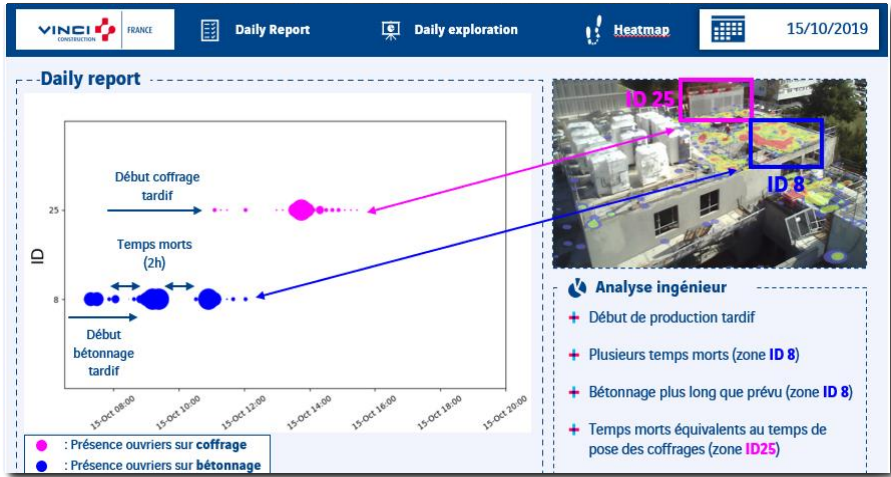
Group data strategy

Design of a **data strategy** to lay the foundations of data exploitation at the Group level and launch **high value-added use cases**





Conception and building of a **construction site monitoring tool** based on **deep learning** and **computer vision**



AGENDA



1. Présentation d'eleven
2. Présentation du sujet du hackathon
3. Modalités pratiques

Le sujet du hackathon s'inscrit dans le cadre du projet BechdelAI porté par Data for good en lien avec le collectif 50/50

Le Collectif 50/50

Présentation

Le Collectif 50/50 réunit à ce jour plus de **1500 professionnel.le.s de la création et de l'industrie du cinéma** et de l'audiovisuel français. Structuré comme un action tank, le Collectif s'engage solidairement dans une réflexion et **un combat pour l'égalité, la parité et la diversité** dans l'industrie cinématographique et audiovisuelle.

Réalisations

Le Collectif 50/50 élabore des **études**, développe des **actions**, crée des **outils** et propose des mesures incitatives aux pouvoirs publics et aux différents acteurs du secteur pour accélérer le changement.

Exemple

L'étude **Cinégalités**, fondée sur l'analyse de 115 films sortis en 2019, dresse un état des lieux des **inégalités de représentation** dans le cinéma français.

Data For Good

Présentation

Data For Good est une association loi 1901 (100% bénévole et non mercantile) créée en 2014 qui rassemble une communauté de **2500+ volontaires tech (Data, Dev, Designers)** souhaitant mettre leurs compétences à profit d'associations et d'ONG et **de s'engager pour l'intérêt général**.

Réalisations

Data for good réalise chaque année 2 **saisons d'accélération** où une dizaine de projets sont accompagnés par les bénévoles sur des thématiques environnementales, sociales et solidaires.

Exemple

Accompagnement d'**Open Food Facts** (base de données derrière l'app Yuka) sur plusieurs sujets (impact carbone, éco-score, etc.) au cours de 4 saisons



Dans le cadre de la 10^e saison d'accélération de projet, le projet **BechdelAI** a pour objectif la mesure et l'automatisation du test de Bechdel, et plus généralement de la (sous)représentation féminine et des inégalités de représentation dans le cinéma et l'audiovisuel

Le projet BechdelAI a pour objectif la mesure de la sous-représentation féminine et des inégalités de représentation dans le cinéma et l'audiovisuel

BechdelAI

Dans le cadre de la 10^e saison d'accélération de projet, le projet BechdelAI a pour objectif la mesure et l'automatisation du test de Bechdel, et plus généralement de la (sous)représentation féminine et des inégalités de représentation dans le cinéma et l'audiovisuel

1

Automatisation du test de Bechdel

- Le test de Bechdel permet de quantifier de manière synthétique la représentation des femmes dans un film

Test de Bechdel

1. Y a-t-il au moins 2 femmes ?
2. Qui se parlent à un moment ? ...
3. ... d'autre chose que d'un homme ?

- Déterminer le score de Bechdel d'un film nécessite aujourd'hui de le regarder en entier, ce qui empêche un passage à l'échelle
- L'objectif est d'automatiser l'analyse en utilisant l'audio, l'image et / ou le script d'un film

2

Extension de l'étude cinégalités

- L'étude Cinégalités a permis d'analyser 115 films sortis en 2019
- Elle se fonde aujourd'hui sur une approche manuelle de l'analyse des films, qui est un frein à sa réplication et son extension
- L'objectif de développer une panoplie d'outils permettant d'automatiser une partie du travail d'analyse des films : genre et couleur de peau des personnages, corrélation avec l'âge, le contexte de représentation (familial, professionnel, etc.), ...

3

Développement d'outils d'analyse

- Le test de Bechdel constitue une mesure synthétique mais imparfaite de la représentation des femmes dans les contenus audiovisuels
- Parallèlement, de nombreuses briques technologiques sont disponibles grâce aux avancées récentes en computer vision et NLP
- L'objectif est de tirer partie de ces briques technologiques pour construire des outils d'analyse plus avancés que le test de Bechdel : analyse de posture, analyse du langage, etc.

Le hackathon portera sur **deux sujets au choix** visant à développer des **briques d'analyse** de la représentation des femmes dans les contenus audiovisuel

Sujet 1


Quels sont les types de films qui obtiennent un mauvais score au test de Bechdel ?

Description

- L'objectif est d'identifier les types de films qui **sous-représentent** systématiquement les femmes ou les représentent mal
- Par exemple, quels sont les **caractéristiques saillantes** des films qui ne passent pas le test de Bechdel ?
- Le but final est d'utiliser ces analyses afin d'effectuer des **campagnes de sensibilisation** à destination des acteurs du monde de l'audiovisuel, en **ciblant les bonnes personnes** et institutions

Technos

 Clustering

 Machine learning

 Interprétabilité

Sujet 2

Combien de personnages de chaque sexe interviennent dans une séquence audio donnée ?


Description

- L'objectif est d'utiliser **l'audio des films** pour automatiser certaines briques du test de Bechdel
- Par exemple, est-il possible de **détecter quand une femme parle**, et même de compter le **nombre distinct** de femmes qui s'expriment dans un audio donné ?
- Le but final est d'utiliser ces briques afin d'analyser un corpus très large de contenu audiovisuel par le prisme du test de Bechdel

Technos

 Analyse audio

 Classification

 Deep learning

Le hackathon portera sur **deux sujets au choix** visant à développer des **briques d'analyse** de la représentation des femmes dans les contenus audiovisuel

Sujet 1


Quels sont les types de films qui obtiennent un mauvais score au test de Bechdel ?

Description

- L'objectif est d'identifier les types de films qui **sous-représentent** systématiquement les femmes ou les représentent mal
- Par exemple, quels sont les **caractéristiques saillantes** des films qui ne passent pas le test de Bechdel ?
- Le but final est d'utiliser ces analyses afin d'effectuer des **campagnes de sensibilisation** à destination des acteurs du monde de l'audiovisuel, en **ciblant les bonnes personnes** et institutions

Technos

 Clustering

 Machine learning

 Interprétabilité

Sujet 2

Combien de personnages de chaque sexe interviennent dans une séquence audio donnée ?

Description

- L'objectif est d'utiliser **l'audio des films** pour automatiser certaines briques du test de Bechdel
- Par exemple, est-il possible de **détecter quand une femme parle**, et même de compter le **nombre distinct** de femmes qui s'expriment dans un audio donné ?
- Le but final est d'utiliser ces briques afin d'analyser un corpus très large de contenu audiovisuel par le prisme du test de Bechdel

Technos

 Analyse audio

 Classification

 Deep learning

Données : vous accédez à **deux bases de données** renseignant les **caractéristiques** d'environ **10k films** ainsi que leur score sur le **test de Bechdel**

Données mises à disposition

Bechdeltest.com

Bechdel Test Movie List

/bech-del test/ n.
1. It has to have at least two [named] women in it
2. Who talk to each other
3. About something besides a man

Movie list

2022 (1 movie)

✓ The 355

2021 (117 movies)

✗ #NoFilter

About

The **Bechdel Test**, or **Bechdel-Wallace Test**, sometimes called the *Mo Movie Measure* or *Bechdel Rule* is a simple test which names the following three criteria: (1) it has to have at least two women in it, who (2) who talk to each other, about (3) something besides a man. The test was popularized by **Alison Bechdel's** comic, the name of which Google won't let me put on this page for inciting hate, in a 1985 strip called *The Rule*. For a nice video introduction to the subject please check out *The Bechdel Test for Women in Movies* on feministfrequency.com.

If you need access to the raw data, check out the [docs for the api](#).

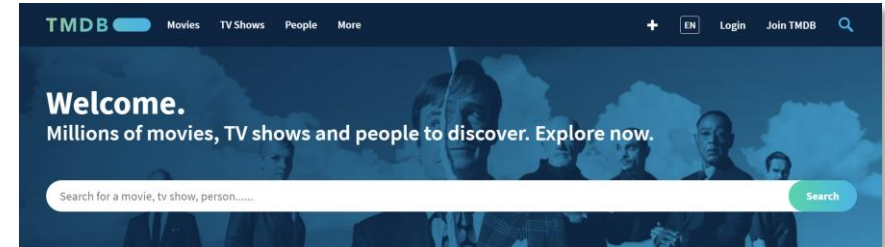
Description

- Le site bechdeltest.com recense le score de Bechdel d'environ **10k films**, notés de manière collective
- La base de donnée issue de l'API du site contient **une ligne par film**, avec un **score de 0 à 3** (pour chacun des 3 critères du test), et des informations diverses sur le film (titre, année, etc.)
- Chaque ligne contient en outre **l'identifiant IMDB** du film, qui permet d'**appairer** des informations issues d'autres sources

Documentation et accès

- Documentation : <https://bechdeltest.com/api/v1/doc>
- Accès : [ici](#)

The Movie Database (TMDB)



Description

- TMDB est une **base de données communautaire sur le cinéma** et la télévision, qui recense environ 750k films
- Les données fournies sont issues de l'API du site, uniquement pour les **films présents sur bechdeltest.com**, et contiennent **une ligne par film**
- Chaque ligne contient des informations sur le film telles que le **budget**, la **langue**, le ou les **genres**, etc.
- Certaines informations peuvent être **manquantes**, et certains films présents sur bechdeltest.com peuvent être **absents** de TMDB

Documentation et accès

- Documentation : <https://developers.themoviedb.org/3/getting-started/introduction>
- Accès : [ici](#)

Approches : vous pouvez **notamment** adopter deux approches basées pour l'une sur des méthodes **d'interprétabilité** et de l'apprentissage supervisé, et pour l'autre sur des méthodes **non supervisées**

Exemples de méthodologies possibles

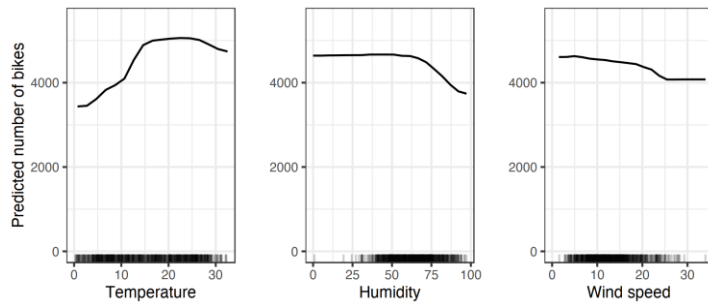
Apprentissage supervisé et interprétabilité

Description

1. Entraîner un algorithme de machine learning pour prédire le score de Bechdel d'un film à partir de ses caractéristiques
 - Quel algorithme utiliser ?
 - Comment évaluer ses performances ?
2. Utiliser des techniques d'interprétabilité pour comprendre les caractéristiques les plus prédictives du score
 - Quelle méthode mettre en œuvre ?
 - Quelles conclusions actionnables en tirer ?

Illustration

Le partial dependency plot permet d'analyser l'influence des variables



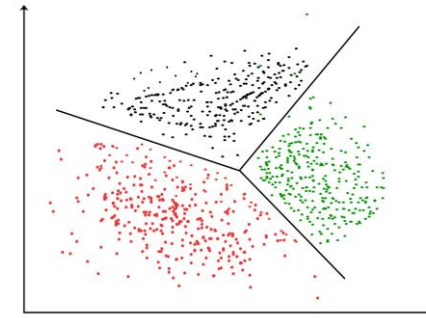
Apprentissage non-supervisé

Description

1. Utiliser un algorithme non-supervisé afin de grouper les films ayant des caractéristiques similaires
2. Pour chaque groupe de films, évaluer le score de Bechdel probable

Illustration

Le clustering permet de grouper les films similaires



Packages

pandas

scikit
learn

SHAP

...

Ressources

- <https://christophm.github.io/interpretable-ml-book/>
- <https://ai.plainenglish.io/predicting-bechdel-test-score-using-machine-learning-7253618a3f8>

...

Le hackathon portera sur **deux sujets au choix** visant à développer des **briques d'analyse** de la représentation des femmes dans les contenus audiovisuel

Sujet 1

Quels sont les types de films qui obtiennent un mauvais score au test de Bechdel ?

Description

- L'objectif est d'identifier les types de films qui **sous-représentent** systématiquement les femmes ou les représentent mal
- Par exemple, quels sont les **caractéristiques saillantes** des films qui ne passent pas le test de Bechdel ?
- Le but final est d'utiliser ces analyses afin d'effectuer des **campagnes de sensibilisation** à destination des acteurs du monde de l'audiovisuel, en **ciblant les bonnes personnes** et institutions

Technos

 Clustering

 Machine learning

 Interprétabilité

Sujet 2

Combien de personnages de chaque sexe interviennent dans une séquence audio donnée ?


Description

- L'objectif est d'utiliser **l'audio des films** pour automatiser certaines briques du test de Bechdel
- Par exemple, est-il possible de **détecter quand une femme parle**, et même de compter le **nombre distinct** de femmes qui s'expriment dans un audio donné ?
- Le but final est d'utiliser ces briques afin d'analyser un corpus très large de contenu audiovisuel par le prisme du test de Bechdel

Technos

 Analyse audio

 Classification

 Deep learning

Approches : vous pouvez découper le sujet en **deux problèmes distincts** concernant (i) la **segmentation** et le regroupement de locuteurs, et (ii) **l'identification du sexe** d'un locuteur

Proposition de découpage du problème

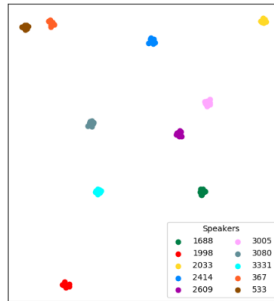
Segmentation et regroupement en locuteurs

Description

1. Découper l'audio en plages courtes
2. Pour chaque plage courte, appliquer une méthode de feature extraction pour transformer l'audio en variables numériques
 - Quelles sont les méthodes d'extraction de feature pour l'audio ?
3. Appliquer un algorithme de clustering pour identifier le nombre de locuteurs distincts et le locuteur de chaque plage
 - Quel algorithme de clustering choisir ?

Illustration

Une bonne méthode de feature extraction doit permettre de séparer les locuteurs



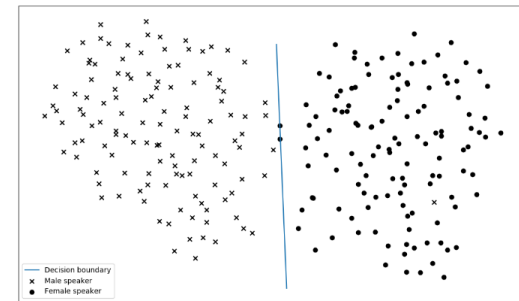
Identification du sexe du locuteur

Description

1. Découper l'audio en plages de locuteurs distincts
2. Pour chaque plage, appliquer une méthode de feature extraction pour transformer l'audio en variables numériques
3. Entraîner un algorithme de machine learning afin d'apprendre à prédire le sexe du locuteur en fonction des caractéristiques numériques de l'audio
 - Quel algorithme appliquer ?

Illustration

Une bonne méthode de feature extraction doit permettre de séparer les sexes



Packages



...

Ressources

- <https://medium.com/saarthi-ai/who-spoke-when-build-your-own-speaker-diarization-module-from-scratch-e7d725ee279>
- <https://github.com/resemble-ai/Resemblyzer>

...

Données : vous accédez à **trois bases de données** vous permettant d'effectuer (i) la **segmentation** de locuteur, (ii) **l'identification du sexe**, et (iii) le **test** de vos algorithmes sur des dialogues de film

Données mises à disposition

1

VoxConverse dataset

2

LibriSpeech dataset

3

moviesoundclips.net dataset

Bechdeltest.com



Description

- Voxconverse est un jeu de donnée conçu pour entraîner et évaluer les algorithmes de **segmentation et regroupement de locuteurs**
- Il contient **216 clips** audios issus de dialogues de vidéos YouTube, principalement de débats ou **émissions télévisées**, en **anglais**
- Chaque clip audio est **labélisé** au format RTTM, qui indique quel locuteur prend la parole à quel moment
- Le **sexe** du locuteur n'est pas renseigné

Documentation et accès

- Documentation : <https://www.robots.ox.ac.uk/~vgg/data/voxconverse/>
- Accès : [ici](#)

Données : vous accédez à **trois bases de données** vous permettant d'effectuer (i) la **segmentation** de locuteur, (ii) **l'identification du sexe**, et (iii) le **test** de vos algorithmes sur des dialogues de film

Données mises à disposition

1

VoxConverse dataset

2

LibriSpeech dataset

3

moviesoundclips.net dataset

The Movie Database (TMDb)

OpenSLR

Open Speech and Language Resources

[Home](#) [Resources](#)

LibriSpeech ASR corpus

Identifier: SLR12

Summary: Large-scale (1000 hours) corpus of read English speech

Category: Speech

License: CC BY 4.0



Description

- LibriSpeech est un jeu de données de plus de 1 000h d'audio construit à partir de fragments d'**audiobooks** lus par environ 2 500 lecteurs, en **anglais**
- Chaque lecteur lit un ou plusieurs extraits d'un ou plusieurs livres, et chaque extrait ne comporte qu'**un seul lecteur**
- Le **sexe** de chaque lecteur est renseigné

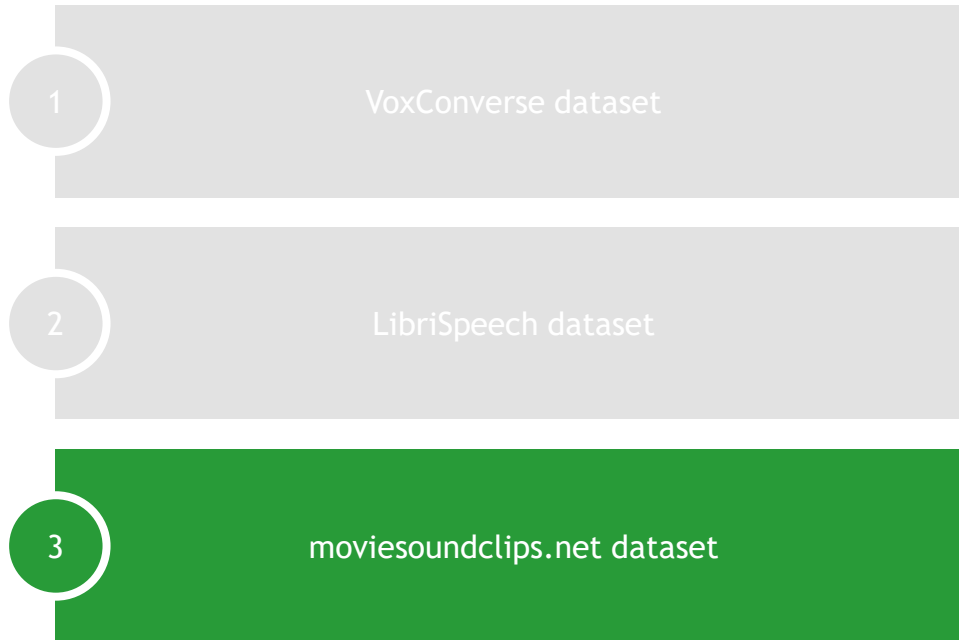


Documentation et accès

- Documentation : <https://www.openslr.org/12/>
- Accès : [ici](#)

Données : vous accédez à **trois bases de données** vous permettant d'effectuer (i) la **segmentation** de locuteur, (ii) **l'identification du sexe**, et (iii) le **test** de vos algorithmes sur des dialogues de film

Données mises à disposition



Movie Sound Clips



Description

- Moviesoundclips.net recense environ 3 000 **clips audio issus de films en anglais**
- Pour chaque clip audio, le **nombre distinct de personnages de chaque sexe** qui intervient est renseigné
- 70 % du jeu de données est mis à disposition pour **valider la performance** des algorithmes pendant le hackathon
- Le reste des données sera utilisé **par le jury** pour tester la performance des algorithmes proposés par les participants

Documentation et accès

- Documentation : <https://www.moviesoundclips.net/>
- Accès : [ici](#)

AGENDA



1. Présentation d'eleven
2. Présentation du sujet du hackathon
3. Modalités pratiques

Le **rendu final** pour chaque équipe sera composé du **code** produit et d'une **présentation** (support et oral), qui compteront chacun dans le **classement final** des équipes

Code

Consignes

- Soigner la **présentation du code** : écrire des fonctions, les commenter, suivre le standard PEP8, ...
- Rendre les **résultats répliquables** et auditable : utiliser git, inclure un fichier README.md pour expliquer comment installer le projet et ce qu'il contient, un fichier requirements.txt, ...

Sujet 1

- Le code devra inclure une fonction permettant de prédire le score de Bechdel d'un film à partir de ses caractéristiques

```
def predict_bechdel(df):  
    """  
    Parameters  
    -----  
    df: pandas.DataFrame  
        DataFrame of movies for which to  
    predict the Bechdel test result.  
    Returns  
    -----  
    bechdel_score: int  
        The Bechdel test result predicted  
    for the given movies.  
    """  
    return bechdel_score
```

Sujet 2

- Le code devra inclure une fonction permettant de compter le nombre de locuteur distinct de chaque sexe dans un audio

```
def count_speakers(audio_path):  
    """  
    Parameters  
    -----  
    audio_path: string  
        Path to .wav file for which to  
    count the number of distinct speakers.  
    Returns  
    -----  
    result: dict  
        Dictionary of the form {'M': nb.  
    male speakers, 'F': nb. female speakers}  
    """  
    return result
```

Présentation

Consignes

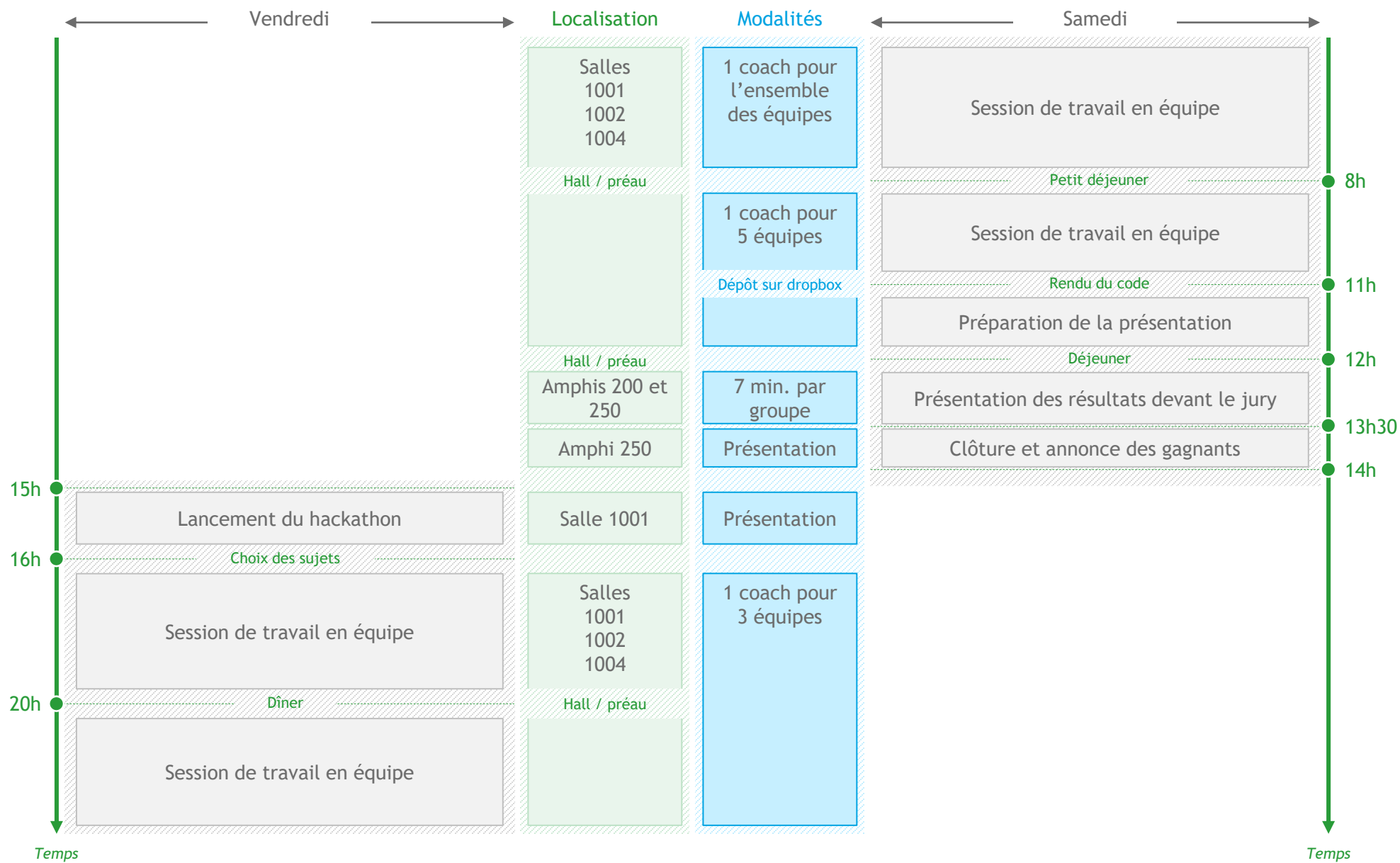
- La présentation ne doit pas dépasser **7 minutes** et sera suivie d'une ou deux questions
- Produire un **support clair** et visuellement attractif. N'hésitez pas à être **créatifs** sur le rendu !
- **Synthétiser l'approche** et les briques techniques choisies de manière simple pour permettre de toucher un auditoire non expert, en restant transparent sur les choix limites de l'approche choisie
- **Prendre de la hauteur** sur les résultats en expliquant comment les utiliser dans le cadre des objectifs du Collectif 50/50
- **Travailler l'oral** pour être impactant



Barème indicatif

- Performance des algorithmes / 6
- Interprétation des résultats et recommandations / 6
- Clarté de la présentation (écrit et oral) / 6
- Qualité du code / 2

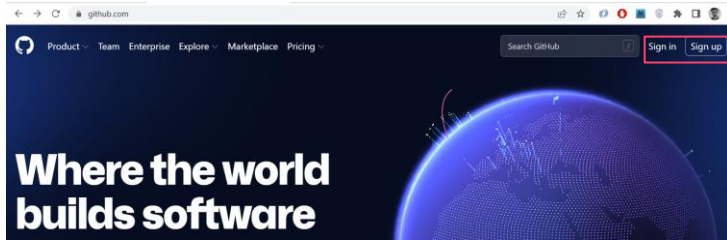
Le **hackathon** se déroulera sur **23h** pendant lesquelles les équipes bénéficieront du **soutien des coaches** et de **trois repas**



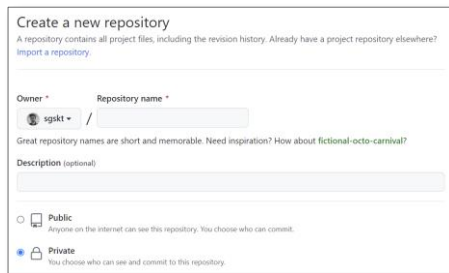


Soumission du code

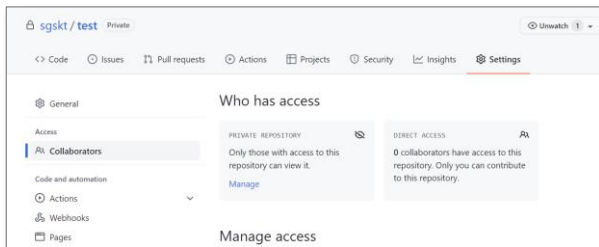
- Créer un compte sur github.com si vous n'en avez pas déjà un



- Une personne dans l'équipe crée un **dépôt privé** et invite le reste des membres comme collaborateurs



- Inviter les **coachs** comme collaborateurs additionnel



- Pousser le code à intervalle régulier. **A 12h, le commit le plus récent sera considéré comme votre soumission finale**



Soumission de la présentation

- La présentation devra être déposée **avant 12h** dans [ce dossier](#)
- Merci de nommer votre fichier de la façon suivante :

`20220423_XX_hackathon_ensae.pptx`

où XX est votre **numéro d'équipe**



Equipes gagnantes

- Le podium sera constitué de **deux équipes gagnantes**, sans classement entre elles, qui se répartiront un prix de **1 500 €**
- Les gagnants seront annoncés lors de la **clôture de l'évènement**
- Le prix sera remis dans les **deux mois** qui suivront




Points divers


- N'hésitez pas à **open-sourcer le code** que vous avez produit pendant le hackathon pour qu'il bénéficie à la communauté 😊

Coachs du hackathon




Emma
 *emmarriau*



Hélène
 *LNBAud*



Simon
 *sgskt*



Théo




Marie
 *marieg-eleven*



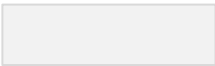
Grégoire
 *greg-lep*



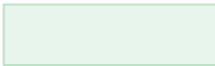
Charles
 *CarloRfg*



Chloé



Coachs eleven



Coachs Data For Good

Enjoy the challenge!