eleven – hackathon challenge

eleven x Département IMI

September 6th, 2022

École des Ponts
ParisTech

eleven
strategy • data • digital

AGENDA

École des Ponts
ParisTech

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

eleven
strategy • data • digital
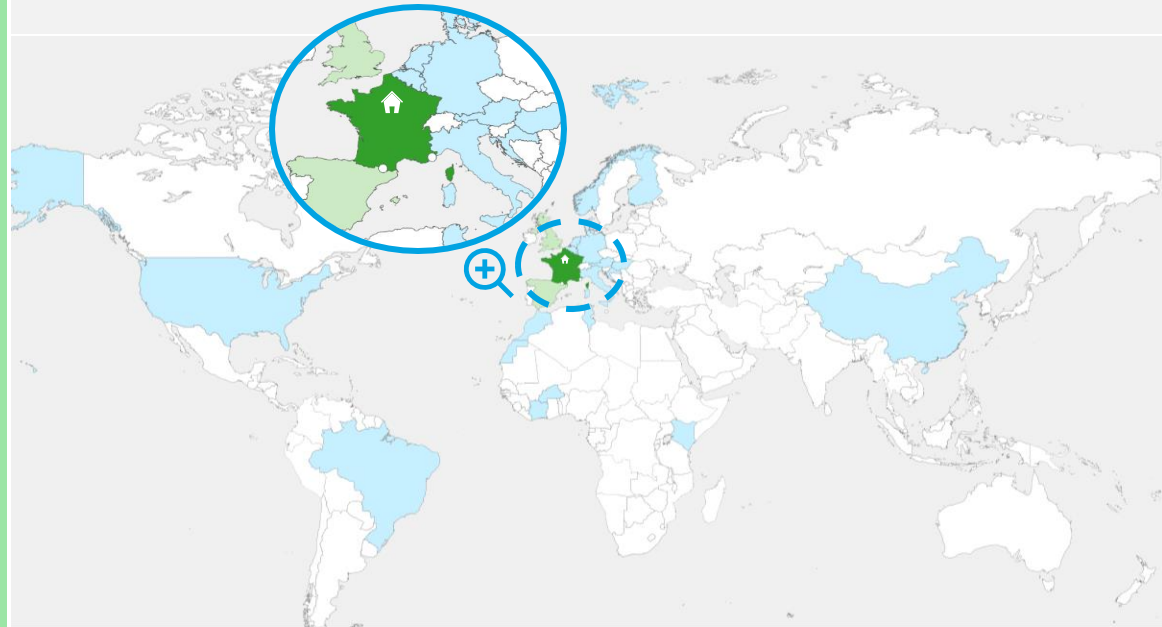
3 COMPLEMENTARY CLIENT OFFERS

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?
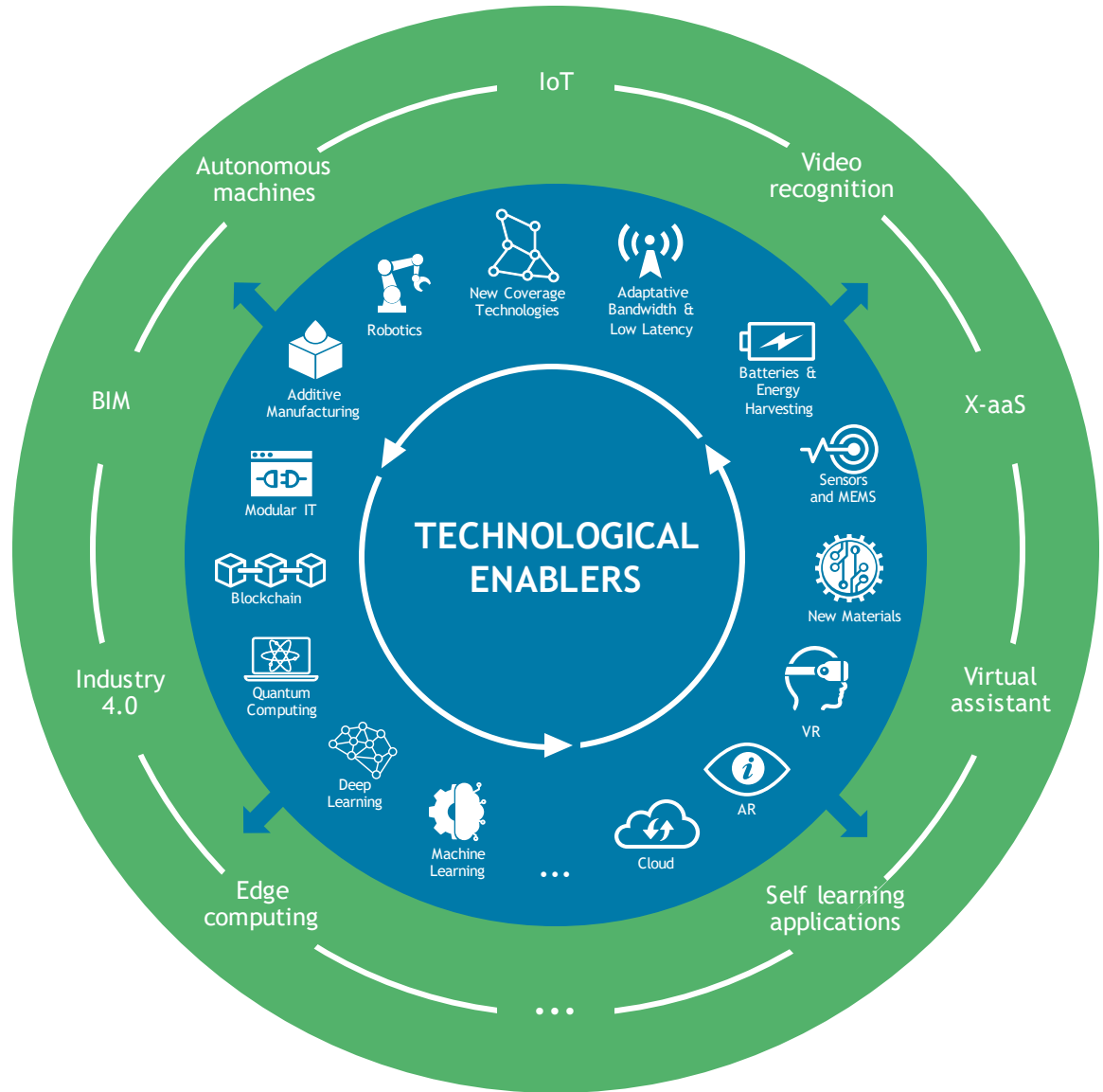
DATA AND AI

- 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 ACCOSS SEVERAL KEY INDUSTRIES...

8

Large Cap

Mid Cap

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 • data • digital

TECH ENTHUSIASTS

DATA SCIENTISTS

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 and end-to-end continuum from strategy to **implementation**
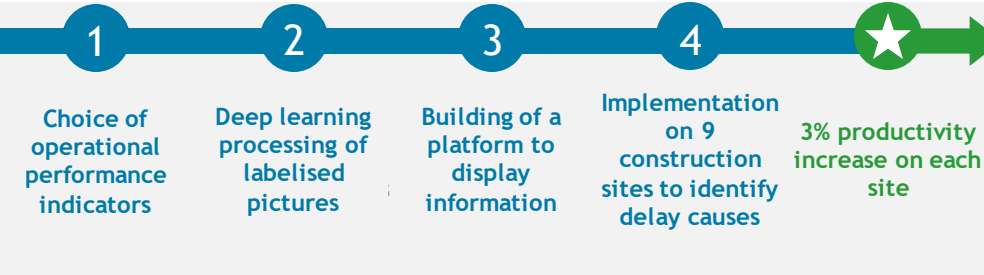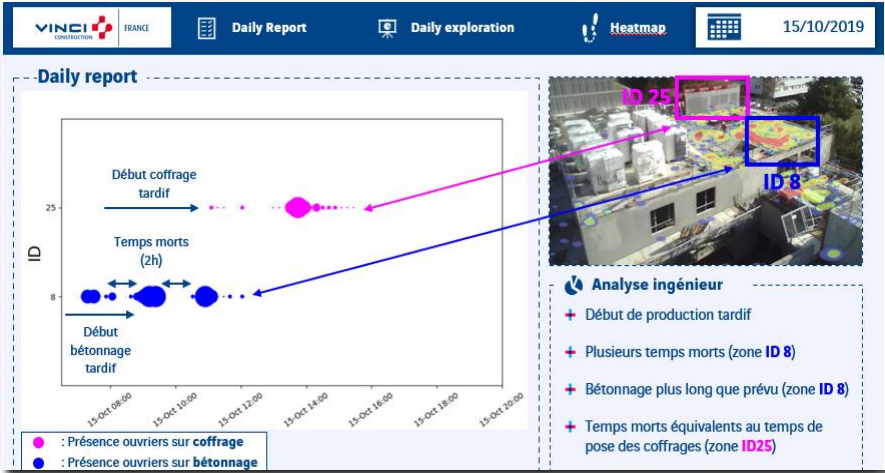
Group data strategy

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

1	2	3	4	★
Internal state of data exploitation study	**State-of-the-art comparison through market benchmark**	**Data exploitation needs collection with all business units**	**Roadmap design and target organisation definition**	**Launch of new data use cases**

VINCI — LEONARD

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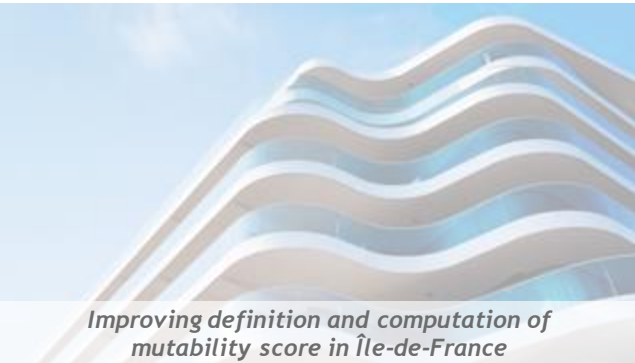

1	2	3	4	★
Choice of operational performance indicators	**Deep learning processing of labelised pictures**	**Building of a platform to display information**	**Implementation on 9 construction sites to identify delay causes**	**3% productivity increase on each site**

École des Ponts
ParisTech

eleven – use cases overview

The Right Price

Improving definition and computation of mutability score in Île-de-France

Project description

- The automatic mutability score enables to evaluate on the fly the potential profitability of a real estate project on any physical parcel given the address

Business Opportunity

- Development of new service
- Increased market share
- Improved profitability

Potential approach

Time series   Clustering

AI Bechdel Test

Automating the Bechdel Test to quantify gender inequality in movies

Project description

- The automation of the Bechdel Test aims to gauge female representation on movies by evaluating the number and quality of women-only interactions

Ethical Impact

- Evaluate bias against women in films and other media
- Identify areas for improvement for producers regarding blind spots around representation

Potential approach

NLP   Clustering

Worksite Monitoring

Ensuring a worksite safety thanks to computer vision technology

Project description

- Worksite monitoring enables to evaluate and ensure faster the safety of any worksite monitored with a stable camera

Business Opportunity

- Improvement of safety protocols
- Internal process optimization

Potential approach

Classification   Computer Vision

eleven
strategy • data • digital

eleven supercase challenge: work on real-life challenges coming from various industries leveraging your dual expertise in business and data science

Exercise:

✓ **Find your group** of 4/5 people

✓ **Choose** among the 3 available topics

✓ On your topic, **put yourself in a data consultant's shoes:**

    ✓ What is the **context** ? (ex: of the company : value chain, competition, technological trends, etc.)
    ✓ How could you generate **added-value using potential of new technological enablers** ? (time-series, NLP, computer vision, etc.)
    ✓ What should be the best way to **present your first results** ? (data visualization, 1st demo, etc.)
    ✓ How much should we **trust your first results** ? (accuracy of your model, etc.)
    ✓ What would be the **next steps if you wanted to keep improving** your results ? (fine tuning, new model, advanced feature engineering, etc.)

✓ **Synthesize your work** in a few slides (up to 15 slides)

✓ **Upload your work** on the Sharepoint (presentation, code, etc: see details at the end of this presentation)

✓ **Present to the jury** and prepare for questions

The Right Price

How to estimate a property price?

AI Bechdel Test

How to quantify women representation ?

Worksite Safety Monitoring

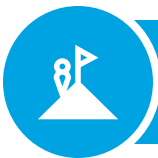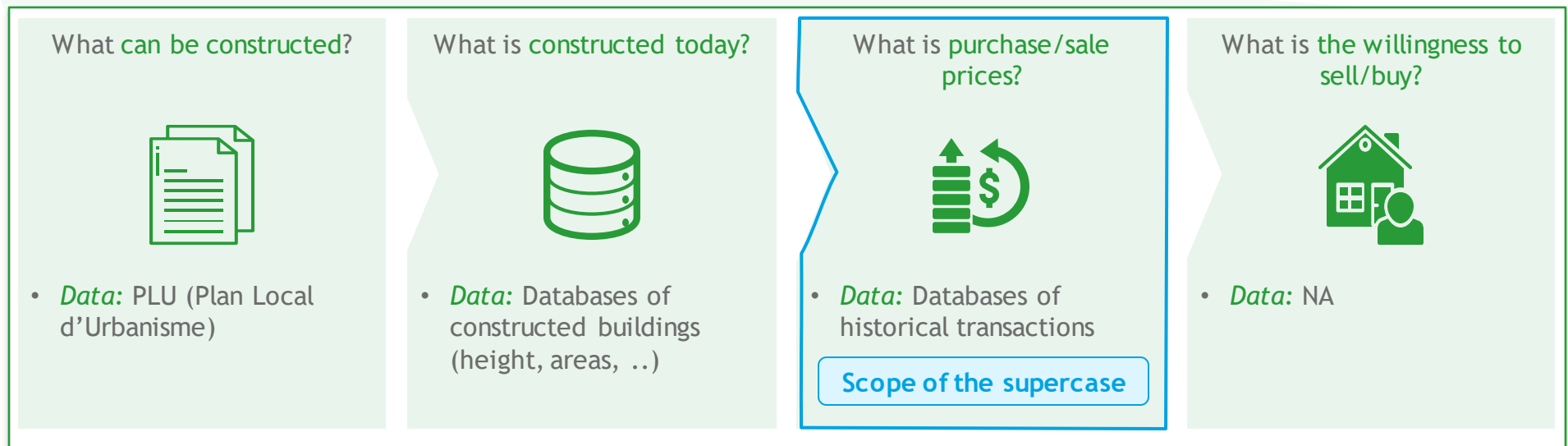How to monitor worksites' safety?

The Right Price

improving definition and computation of mutability score in Île-de-France

Context: the client, a player in real-estate industry, wants to build a robust purchase/sale price estimation model

Value chain of a real estate development project

Prospection	Construction	Operation	Renovation	Sale

What can be constructed?	What is constructed today?	What is purchase/sale prices?	What is the willingness to sell/buy?
• *Data:* PLU (Plan Local d'Urbanisme)	• *Data:* Databases of constructed buildings (height, areas, ..)	• *Data:* Databases of historical transactions	• *Data:* NA
		Scope of the supercase	
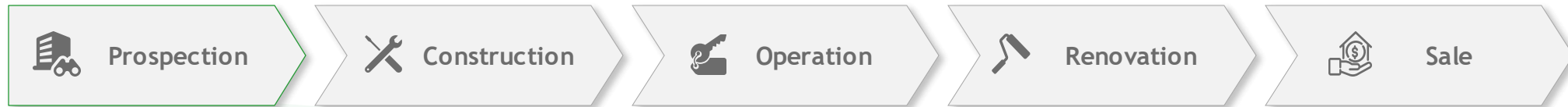
Your missions
- Process the data from historical transactions to extract meaningful features
- Build one or several price estimation model for new construction (exclusively apartments)
- Present the results to your client with the test dataset that will be provided to you

Sources: eleven strategy

eleven
strategy • data • digital

Data available: you have **access to data** from official mutation databases in Île-de-France, **non-geo-localized except but for cadastral id**, with large and various information on the properties

Note that **as in any data science project**, the data **may require further data engineering** before being fully leverageable

datemut	anneemut	moismut	coddep	libnatmut	vefa	valeur fonciere	codtypebien	libtypbien
22/10/2018	2018	10	75	Vente	FALSE	147000.0	121	UN APPARTEMENT
26/10/2015	2015	10	75	Vente	FALSE	95000.0	131	UNE DEPENDANCE
13/06/2019	2019	6	75	Vente	FALSE	1900000.0	152	BATI MIXTE - LOGEME
20/04/2017	2017	4	75	Vente	FALSE	1509000.0	121	UN APPARTEMENT
06/03/2015	2015	3	75	Vente	FALSE	28250.0	131	UNE DEPENDANCE
01/08/2017	2017	8	75	Vente	FALSE	500000.0	121	UN APPARTEMENT
03/07/2019	2019	7	75	Vente	FALSE	367000.0	121	UN APPARTEMENT
27/10/2016	2016	10	75	Vente	FALSE	90000.0	121	UN APPARTEMENT
23/11/2018	2018	11	75	Vente	FALSE	135000.0	121	UN APPARTEMENT
26/08/2016	2016	8	75	Vente	FALSE	190000.0	121	UN APPARTEMENT
28/12/2017	2017	12	75	Vente	FALSE	300000.0	14	ACTIVITE
11/06/2014	2014	6	75	Vente	FALSE	730000.0	121	UN APPARTEMENT
02/05/2018	2018	5	75	Vente	FALSE	935000.0	121	UN APPARTEMENT
11/05/2016	2016	5	75	Vente	FALSE	535000.0	121	UN APPARTEMENT
07/10/2016	2016	10	75	Vente	FALSE	339050.0	121	UN APPARTEMENT
11/07/2018	2018	7	75	Vente	FALSE	413438.0	121	UN APPARTEMENT
22/10/2018	2018	10	75	Vente	FALSE	3327000.0	14	ACTIVITE
06/02/2018	2018	2	75	Vente	FALSE	650000.0	122	DEUX APPARTEMENTS
04/09/2018	2018	9	75	Vente	FALSE	375000.0	14	ACTIVITE
09/02/2015	2015	2	75	Vente	FALSE	163000.0	121	UN APPARTEMENT
18/06/2020	2020	6	75	Vente	FALSE	768490.0	121	UN APPARTEMENT
07/06/2017	2017	6	75	Vente	FALSE	15000.0	14	ACTIVITE
13/11/2014	2014	11	75	Vente	FALSE	480000.0	121	UN APPARTEMENT
30/03/2018	2018	3	75	Vente	FALSE	823000.0	121	UN APPARTEMENT
14/09/2015	2015	9	75	Vente	FALSE	250000.0	121	UN APPARTEMENT
16/02/2018	2018	2	75	Vente	FALSE	2873000.0	122	DEUX APPARTEMENTS
11/03/2019	2019	3	75	Vente	FALSE	525000.0	121	UN APPARTEMENT

Mid-term nudge

✓ **Bonus:** You could try to retrieve the **geo-localization version** of this dataset from the other features provided

Resources : You are free to use any resources you want, here are some recommendations to help you get started

---

Languages

We highly recommend you use Python even though same kind of results could be achieved with similar tools
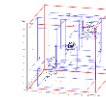

Relevant libraries

scikit learn	GeoPandas	Rtree	Streamlit
To develop a wide range of ML models	*To use geo-localized visualization tools*	*To use geo-localized advanced ML models*	*To efficiently develop a dashboard / front-end*

For information: the process of geo-location is independent from the rest of the challenge

75 Historical mutations

77 Historical mutations

→ **eleven** secret recipe – to be discovered →

75 Geo-localized mutations

77 Geo-localized mutations

Hint : Investigate cadastral open data

AI Bechdel Test

Automating the Bechdel Test to quantify gender inequality in movies

Context: The BechdelAI project aims to measure female underrepresentation and gender inequality in movies and audiovisual media

To pass the test, a movie must meet 3 conditions

1 Two women are named (first name or first name + last name)

2 They talk to each other ...

3 about anything but men

The Issue
- To compute a Bechdel score, movies must be watched entirely which prevents a scaling-up of the test that could be applied to the entire media industry

Your missions
- The objective is to automate the analysis using the audio, image and/or script of a film
- Can we detect if a woman is speaking and count the total number of distinct women speaking in an audio extract ?

École des Ponts ParisTech

eleven
strategy • data • digital

Proposed Approach : you can split the problem into two separate problems concerning (i) segmentation and clustering of speakers, and (ii) identifying the gender of a speaker
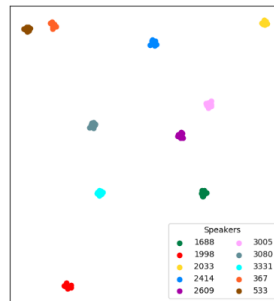
Proposed breakdown of the problem

Segmentation and clustering of speakers

Description

1. Cutting the audio into short tracks

2. For each short track, apply a feature extraction method to transform the audio into digital variables

   - What are the feature extraction methods for audio ?

3. Apply a clustering algorithm to identify the number of distinct speakers and the speaker in each audio range

   - Which clustering algorithm to choose ?

Illustration

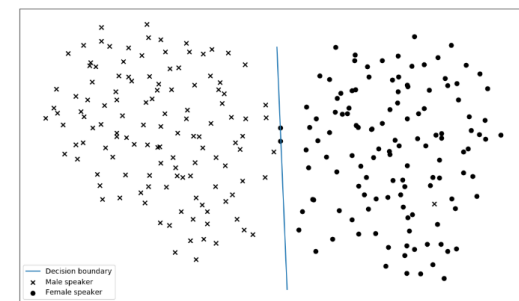A good feature extraction method must be able to separate the speakers

Gender identification of the speaker

Description

1. Cutting the audio into short distinct speaker tracks

2. For each short track, apply a feature extraction method to transform the audio into digital variables

3. Train a machine learning algorithm to learn to predict speaker gender based on numerical features of the audio

   - Which algorithm to apply?

Illustration

A good feature extraction method should separate genders

Packages

librosa   Resemblyzer   speechSegmenter   ...

Ressources

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

- https://github.com/resemble-ai/Resemblyzer

...

eleven
strategy • data • digital

Data available: you have access to 3 databases allowing you to perform (i) speakers segmentation, (ii) gender identification, and (iii) testing your algorithms on movie dialogues

Available databases

1	VoxConverse dataset
2	LibriSpeech dataset
3	moviesoundclips.net dataset

Bechdeltest.com

📇 **Description**

• Voxconverse is a dataset designed to train and evaluate speaker segmentation and clustering algorithms

• It contains 216 audio clips from YouTube video dialogues, mainly from debates or TV shows, in English

• Each audio clip is labeled in RTTM format, which indicates which speaker is speaking at which time

• The gender of the speaker is not indicated

ⓘ **Documentation and access**

• Documentation : https://www.robots.ox.ac.uk/~vgg/data/voxconverse/

• Access : here

Data available: you have access to 3 databases allowing you to perform (i) speakers segmentation, (ii) gender identification, and (iii) testing your algorithms on movie dialogues
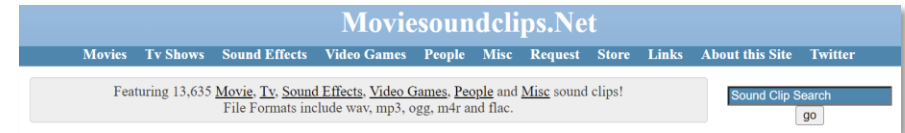
Données mises à disposition

1 — VoxConverse dataset

2 — LibriSpeech dataset

3 — moviesoundclips.net dataset

Open SLR — 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 is a dataset of more than 1,000 hours of audio built from fragments of audiobooks read by about 2,500 readers, in English

- Each reader reads one or more excerpts from one or more books, and each excerpt has only one reader

- The gender of each reader is indicated

ⓘ **Documentation and access**

- Documentation : https://www.openslr.org/12/

- Access : here

eleven
strategy • data • digital

École des Ponts
ParisTech

Data available: you have access to 3 databases allowing you to perform (i) speakers segmentation, (ii) gender identification, and (iii) testing your algorithms on movie dialogues

Données mises à disposition

1 VoxConverse dataset

2 LibriSpeech dataset

3 moviesoundclips.net dataset

Movie Sound Clips

Moviesoundclips.Net

Movies   Tv Shows   Sound Effects   Video Games   People   Misc   Request   Store   Links   About this Site   Twitter

Featuring 13,635 Movie, Tv, Sound Effects, Video Games, People and Misc sound clips!
File Formats include wav, mp3, ogg, m4r and flac.

Sound Clip Search
go

Description

- Moviesoundclips.net lists about 3,000 audio clips from English movies
- For each audio clip, the distinct number of characters of each gender who intervene is informed
- 70% of the dataset is made available to validate the performance of the algorithms during the hackathon
- The rest of the data will be used by the jury to test the performance of the algorithms proposed by the participants

ⓘ Documentation and access

- Documentation : https://www.moviesoundclips.net/
- Access : here

Worksite Safety Monitoring

Ensuring a worksite safety thanks to computer vision technology

Context: the client, a player in the construction industry is building a worksite monitoring solution to evaluate and mitigate risks on worksites

For illustration purposes

**1 Data acquisition**

- 7 camera deployed on several test construction site
- Storage in a centralized database

2 Detection model

- Detection automation based on deep learning
- Training on 1500 pictures with 11 000 annotations

3 Analysis model

- Clustering of images to group together images from the same camera
- Heatmap of workers' presence during the day
- Storage of results in a table

4 Reporting

- Synthesis of daily human activities contextualized in a picture
- User-friendly exposition of the results
- Identify other value propositions relying on the detection model

Scope of the supercase

Your missions

- Build a worker detection model based on annotated data
- Cluster the images to process each camera independently (prerequisite to draw heatmaps)
- Build a heatmap for each worksites with density of workers during the day
- Expose your result and performance on a dashboard with a test dataset that will be provided to you
- Identify additional value that could be brought by the detection model

Data available: you have access to pictures from different worksites, with the corresponding labels giving you information about on-site workers

Example of worksite pictures
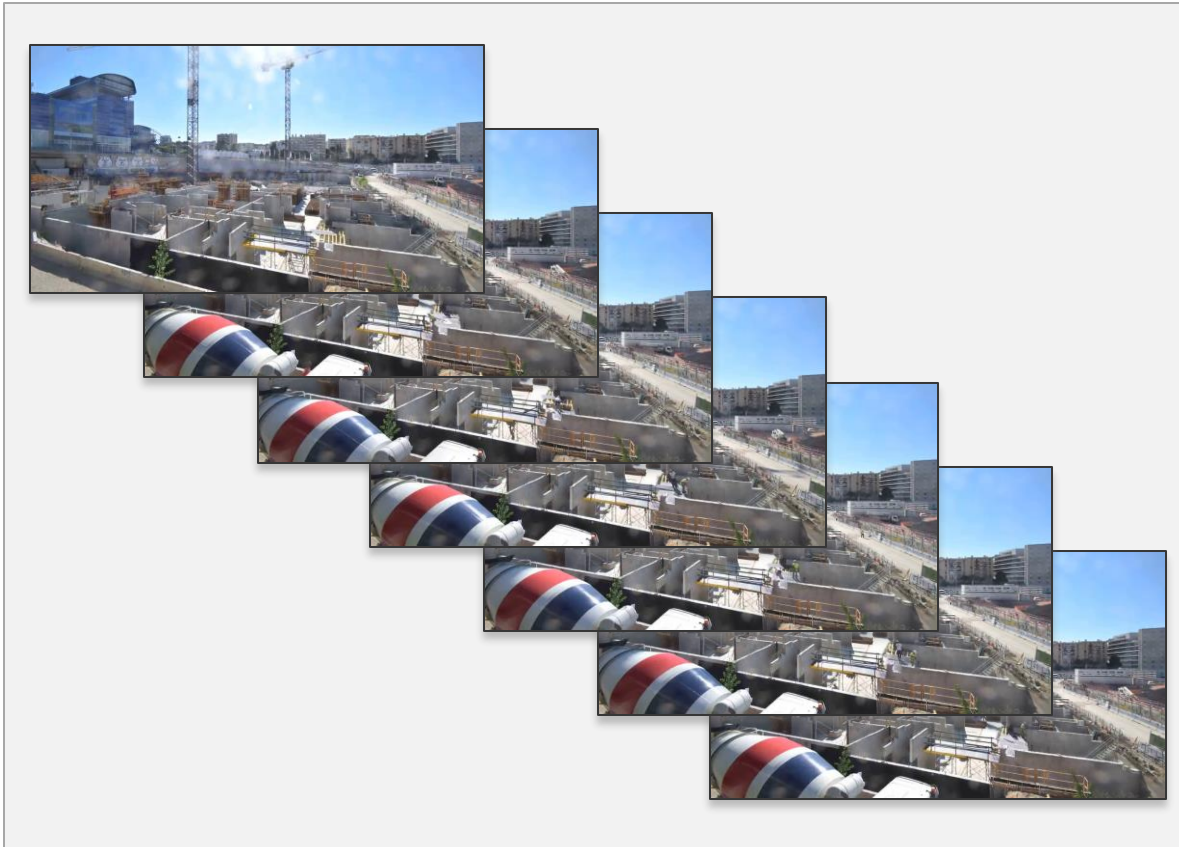
Example of labels (json format)

```
{"description": "",
 "tags": [],
 "size":
        {"height": 720, "width": 1280},
        "objects":
                [{"id": 346688445,
                 "classId": 1285799,
                 "description": "",
                 "geometryType": "rectangle",
                 "labelerLogin": "Raziajuthy",
                 "createdAt": "2020-02-04T09:08:12.615Z",
                 "updatedAt": "2020-02-28T11:38:28.963Z",
                 "tags": [],
                 "classTitle": "People",
                 "points":
                        {"exterior": [[153, 622], [177, 647]],
                        "interior": []}},
                {"id": 346688444,
                 "classId": 1285799,
                 "description": "",
                 "geometryType": "rectangle",
                 "labelerLogin": "Raziajuthy",
                 "createdAt": "2020-02-04T09:08:12.615Z",
                 "updatedAt": "2020-02-28T11:38:28.963Z",
                 "tags": [],
                 "classTitle": "People",
                 "points":
                        {"exterior": [[851, 683], [865, 720]], "interior": []}},
                {"id": 346688443,
                 "classId": 1285811,
                 "description": "",
                 "geometryType": "rectangle",
```

Resources: all classical Machine Learning might come handy and a good understanding of Computer Vision libraries will be helpful

PyTorch

Standard library for deep learning

OpenCV

Standard library classical computer vision

TensorFlow

An alternative to Pytorch

scikit learn

To develop a wide range of ML models

pandas

To exploite model's output and aggregate them

Streamlit

To efficiently develop a dashboard / front-end

AGENDA

1. About eleven

2. Case presentations

   a)   The Right Price

   b)   AI Bechdel Test

   c)   Worksite Monitoring

3. General information

   〉

   a)   Expected output

   b)   Practical information

eleven
strategy • data • digital

École des Ponts
ParisTech

Expected output

You are expected to deliver the following:

- The file with your code (for this assignment we highly recommend using Python)
- A PowerPoint presentation of your work (including your experiment process, your thoughts, the hardships you had to overcome…)

PRESENTATION

CODE

⟩

École des Ponts
ParisTech

Schedule for the four days: the timing may be short, do not hesitate to split the work between the members of the team

	Tuesday 6th	Wednesday 7th	Thursday 8th	Friday 9th
AM session	**9.00 am Kick-off session** *Amphi Caquot 1* *Topic choice – 15mn* **Training (Sound/CV/TRP)** *P012, P302, P303*	**Free working session** The Right Price : *P302* AI Bechdel Test : *V404* Worksite Monitoring : *P303*	**Free working session** The Right Price : *P302* AI Bechdel Test : *P202* Worksite Monitoring : *P303*	**9.00 am > 12.00 pm Last working session** The Right Price : *P302* AI Bechdel Test : *P102* Worksite Monitoring : *P303*
PM session	**Q&A - tech** *P302, P303, P102* **5.00pm - The Right Price** *Salle: P302* **4.00pm - AI Bechdel Test** *Salle: P102* **5.00pm - Worksite Safety Monitoring** *Salle: P303*	**Q&A - tech** *P302, P303, V404* **The Right Price** *Salle: P302* **5.00pm-6.00pm - AI Bechdel Test** *Salle: V404* **Worksite Safety Monitoring** *Salle: P303*	**Q&A - business** *P302, P303, P202* **The Right Price** *Salle: P302* **AI Bechdel Test** *Salle: P202* **Worksite Safety Monitoring** *Salle: P303*	**1.00 pm > 3.30 pm Pitches session** *Amphi Caquot 1* Pitches will take place in front of a 3 people jury: 1 technical profiles, 1 business profile and 1 PhD **3.45 pm > 5.15 pm Closing session** *Amphi Caquot 1* The best team of each topic will pitch in front of the whole class **5.15 pm Cocktail** *La Grande Ruche*

Final presentation details and best practices:

On Friday 9th, you will have to present your work in front of a jury during a closed-door session

The modalities of the presentation will be as follow:

- 10min group pitch based on a PowerPoint presentation
- ~10min Q&A session with the jury
- ~5min debrief from the jury

For each supercase, a winner will be announced. The three winners will then present their work to the other students (same modalities with questions from the students)

The presentation must be as professional as possible. Here are some advices and best practices that may be useful:

- Structure your presentation: start by stating the problem that you want to solve, then present the way you tackled it, and finally describe your solution. The "story" of the presentation should be natural and easy to follow

- Be concise and precise: focus on the most important messages, as you only have 20 minutes to present the work achieved for the entire week. You should limit the number of slides you present (you can still add appendices if needed)

- Be organized as a team: split up the speaking time between the team members beforehand to make it smoother

- Be honest: tell where you encountered issues or challenges

- C-suite level: you should convince both the CEO and the CTO/CDO of the company

Evaluation criteria:

Although different in their essence, the cases will be graded based on similar criteria.
NB: any provided code will be tested in order to ensure its good functioning.

The contribution of each criteria may not reflect the actual value in a final mission restitution

	Topic	Description	Contribution to the final grade*
	Engagement	The engagement of the team during the exercise (how far you've gone, how autonomous you have been, how much you have asked questions when stuck, etc.)	2 points
Business aspect	Presentation quality	The quality of your final presentation: how professional it looks (slide quality), how clear and complete it is (storytelling), how pertinent your answers are, etc.	3 points
	Business methodology	The creativity and relevance of the methodology (i.e. scientific approach) you choose regarding the problem you try to solve, and the data provided, the business sense behind your methodology and the pragmatism of your presentation	6 points
Technical aspect	Technical choices	The explanation of your technical choices and your ability to present them in non-technical terms	3 points
	Model efficiency or Analysis relevance	The performance of your model (specific to each case), the relevance for the problem in question, the quality of the analysis led	5 points
	Code good practices	Your code must be well structured, easy to run and easy to understand with clear readme and requirements.	1 point

Please note that all groups will be graded at the end of the week

The Team workplace

For this challenge, we opened Teams for you to ask your questions when eleven consultants are not on campus

On this workplace, you will find three channels:

1) **général**: for all questions and information related to the organization of the challenge

2) **TheRightPrice**: for all questions specifically related to *TheRightPrice* case

3) **AI Bechdel Test**: for all questions specifically related to *AI Bechdel Test* case

4) **WorksiteSafetyMonitoring**: for all questions specifically related to WorksiteMonitoring case

Additional information may also be pinned in these channels (schedules, classroom numbers, etc.)

Please use the right channel to ensure fluidity of the interactions

Before asking something, also make sure that the requested information has not been given already ;)
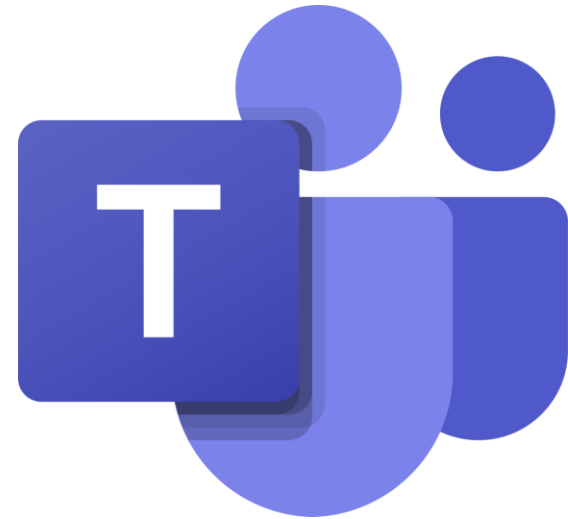
You may download the Microsoft Teams application on your device or access it via your usual browser

Download instructions & submission process

How to download datasets ?

You can download datasets and potential additional information at the following links:

- **Case #1**: TRP – get your files here

Nom ∨	Modifié par ∨
📁 Data localisée	Marco CAMPAGNA
📁 Data non localisée	Marco CAMPAGNA
📄 lexique_variables.csv	Marco CAMPAGNA

- **Case #2**: ABT – get your files here

Nom ∨	Modifié par ∨
📁 librispeech	Marco CAMPAGNA
📁 moviesoundclips.net	Marco CAMPAGNA
📁 voxconverse	Marco CAMPAGNA

- **Case #3**: WSM – get your files here

Nom ∨	Modifié par ∨
📁 Detection_Test_Set	Marco CAMPAGNA
📁 Detection_Train_Set	Marco CAMPAGNA

How to submit your works ?

Each team will receive a link by email with a Sharepoint folder to submit their assignment (both Presentation + Code)

Nom ∨	Modifié ∨	Modifié par ∨	Taille du fichier ∨	Partage

Faites glisser les fichiers ici

Notes:

- You can organize your folder as you wish.
- You can keep old files that should not be submitted to the jury in a *0_Archives* folder

Enjoy the challenge!