

---

# Data Science Pour Ingénieurs avec Python

Semaine 6 – Partie II

---

ENG-209 2024

## Présentations - enseignants (2<sup>e</sup> partie)



Jean-Philippe Pellet

---

*1<sup>re</sup> partie*

**Python pour habitué·e·s de C++;  
découverte de Pandas**



Éric Bouillet

---


*2<sup>e</sup> partie*

**Data Science en Python**



Olivier Verscheure

## Agenda - 2<sup>e</sup> partie

Date	Contenu
 Semaine 6	Introduction: Une journée de la Data Science Numpy: Génération de données synthétiques Matplotlib: Visualisation de données scientifiques
Semaine 7	Scikit-learn: introduction à l'apprentissage automatique
Semaine 8	- Régression linéaire et polynomiale
Semaine 9	- Classification par régression logistique
Semaine 10	- Classification par arbres de décisions
Semaine 11	- Évaluation de performances des modèles
Semaine 12	Révisions et énoncé
Semaine 13	Examen final - partie II (poids $\frac{2}{3}$ ) – due 19.10.2024 à minuit

Sous réserve d'adaptations et de modifications

---

# Introduction to Data Science

ENG-209 2024

---

# 5 main challenges

1. Getting the right data
2. Collaborating with the business
3. Building trust

# Common Misconception in Businesses



© marketoonist.com

# What is Data Science?



# What is Data Science?

Data + Algorithms → Knowledge → Benefits

*Big Data /  
Data lake*

*Machine Learning /  
“Dumb” A.I.*



# Challenge #1 - Getting the right data



# What do you see?



# A fantastic source of data

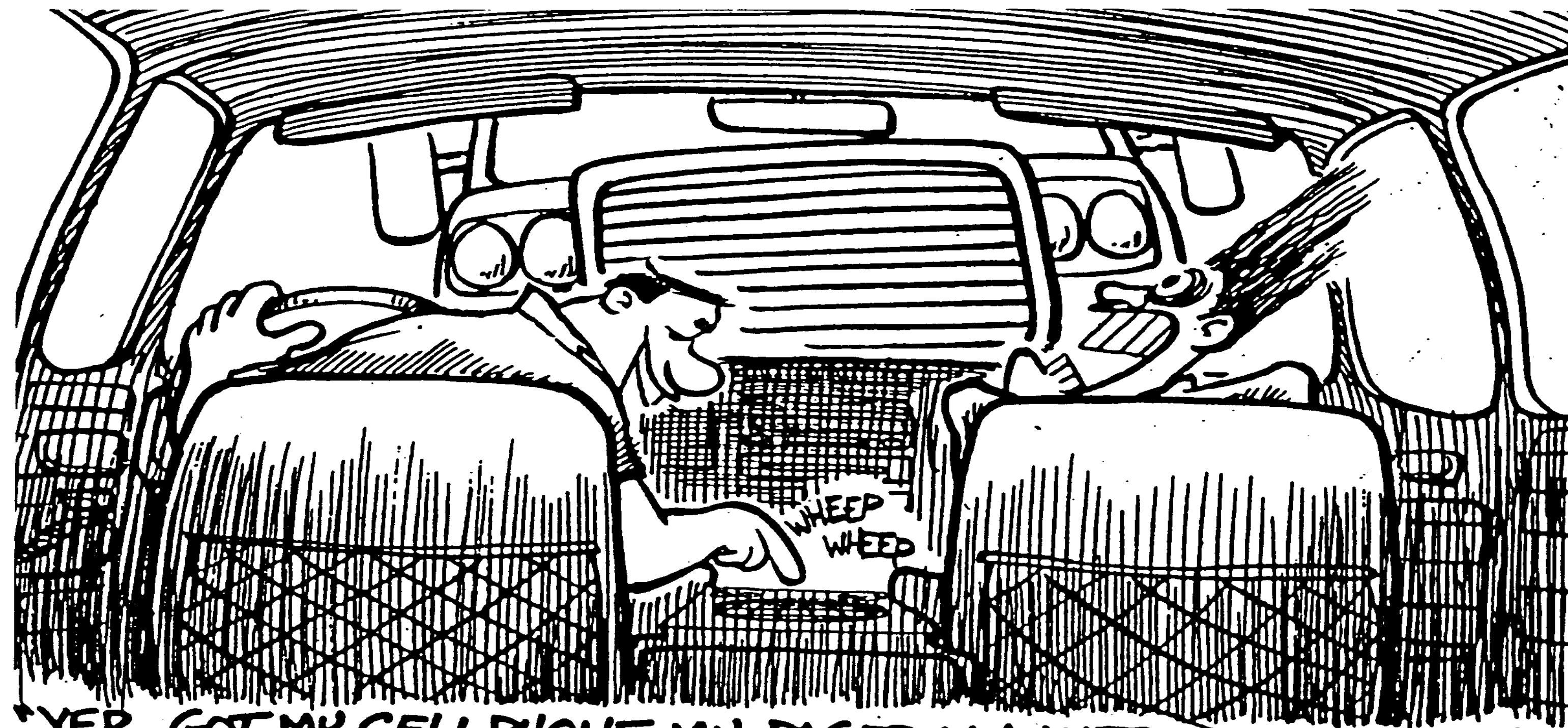


# Data is the new oil



The Economist, May 2017

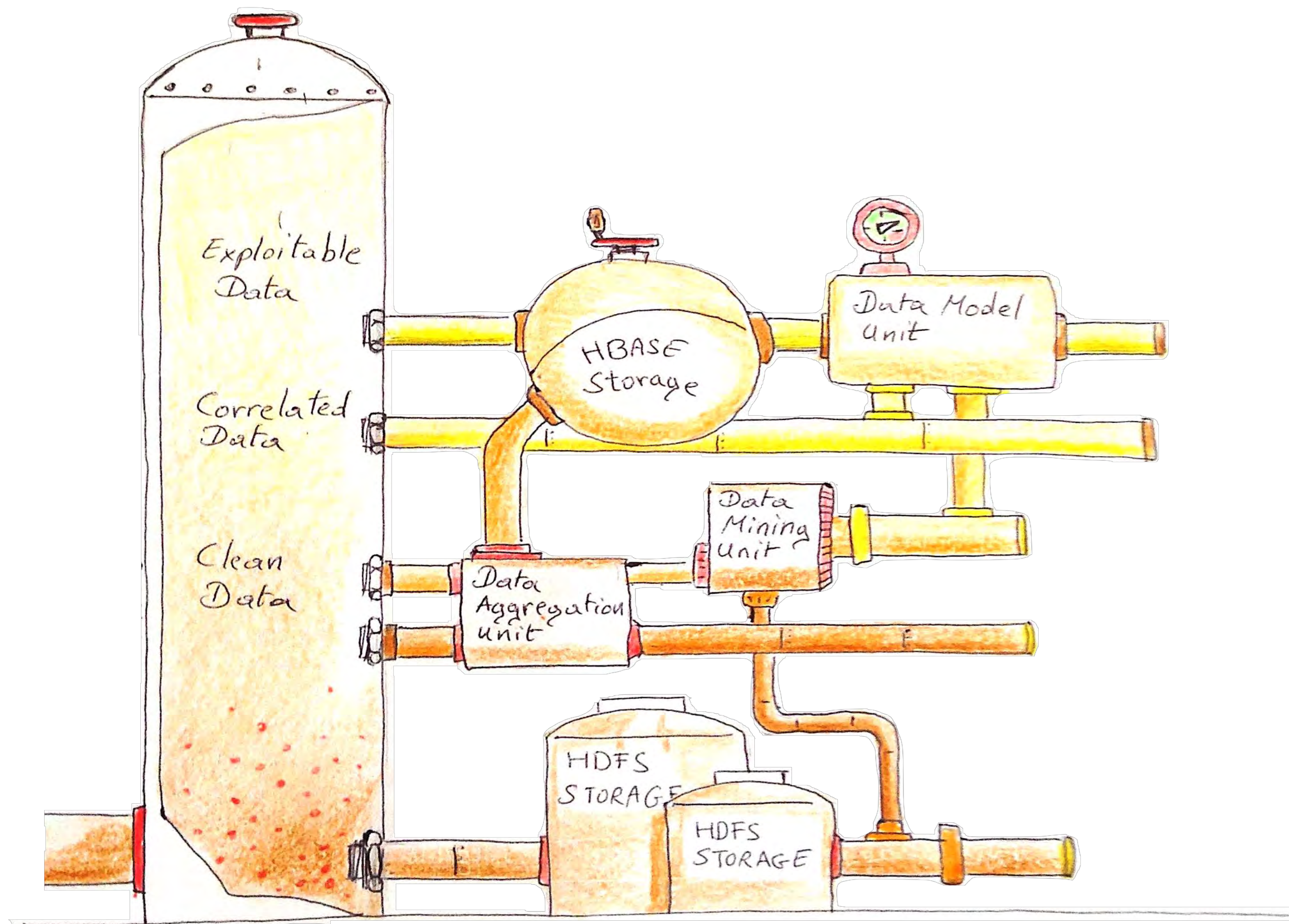
# Big data, Bad data



"YEP... GOT MY CELLPHONE, MY PAGER, MY INTERNET LINK, MY WIRELESS FAX, AND THANKS TO THIS NIFTY SATELLITE NAVIGATION SYSTEM, I KNOW PRECISELY WHERE I AM AT ALL TIMES!"

BY LOWE FOR THE SUN-SENTINEL, FLOP

# Like oil, data must be refined

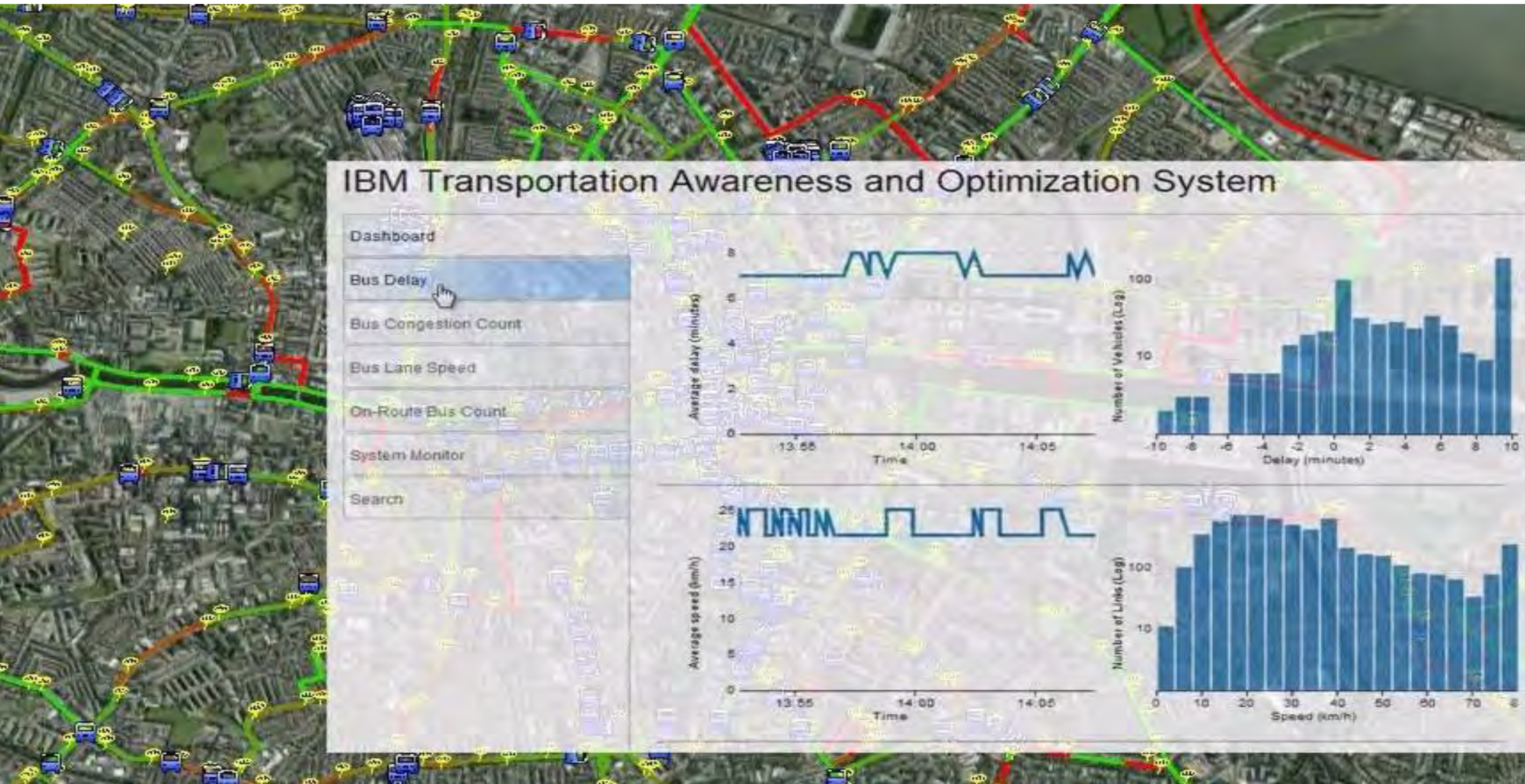




**Challenge #2 – Collaborating with domain experts**

# From raw data to information

Real-time tool for situational awareness



-  Timetables
-  Car
-  Parking capacity
-  Routes & maps
-  SCATS Induction loop
-  Bus AVL (GPS)
-  Accessibility
-  CCTV
-  Bike



# Bühler Insights Replay



1.7 milliards d'observations par jour

# What is Data Science?

Data + Algorithms → Knowledge → Benefits

*Big Data /  
Data lake*

*Machine Learning /  
“Dumb” A.I.*

# How can data science improve Urban Mobility?

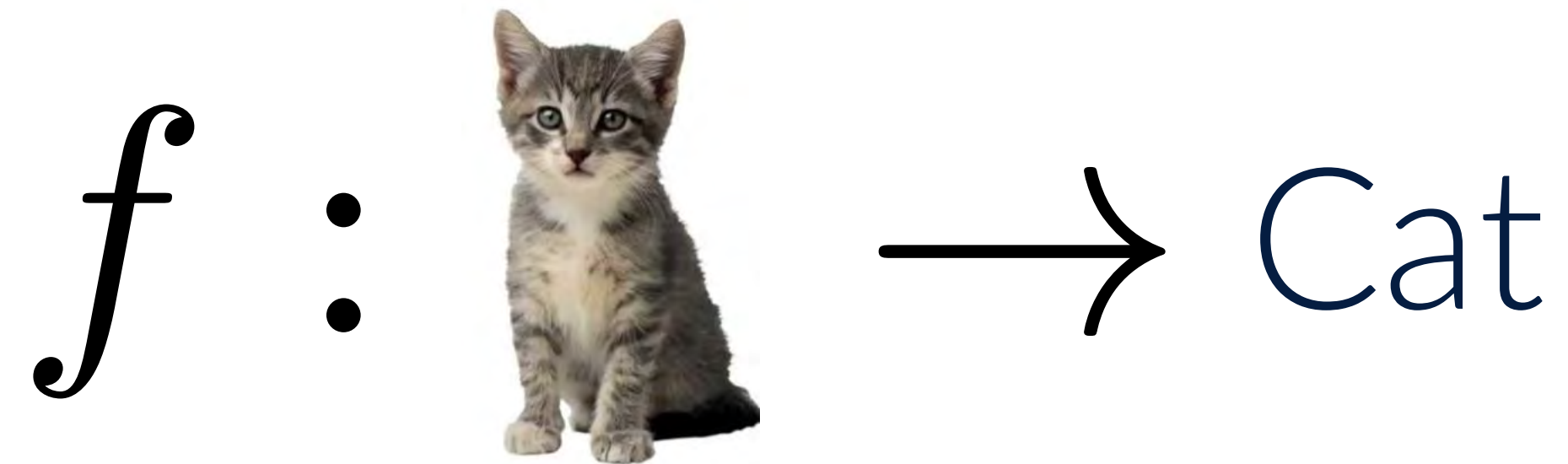
19



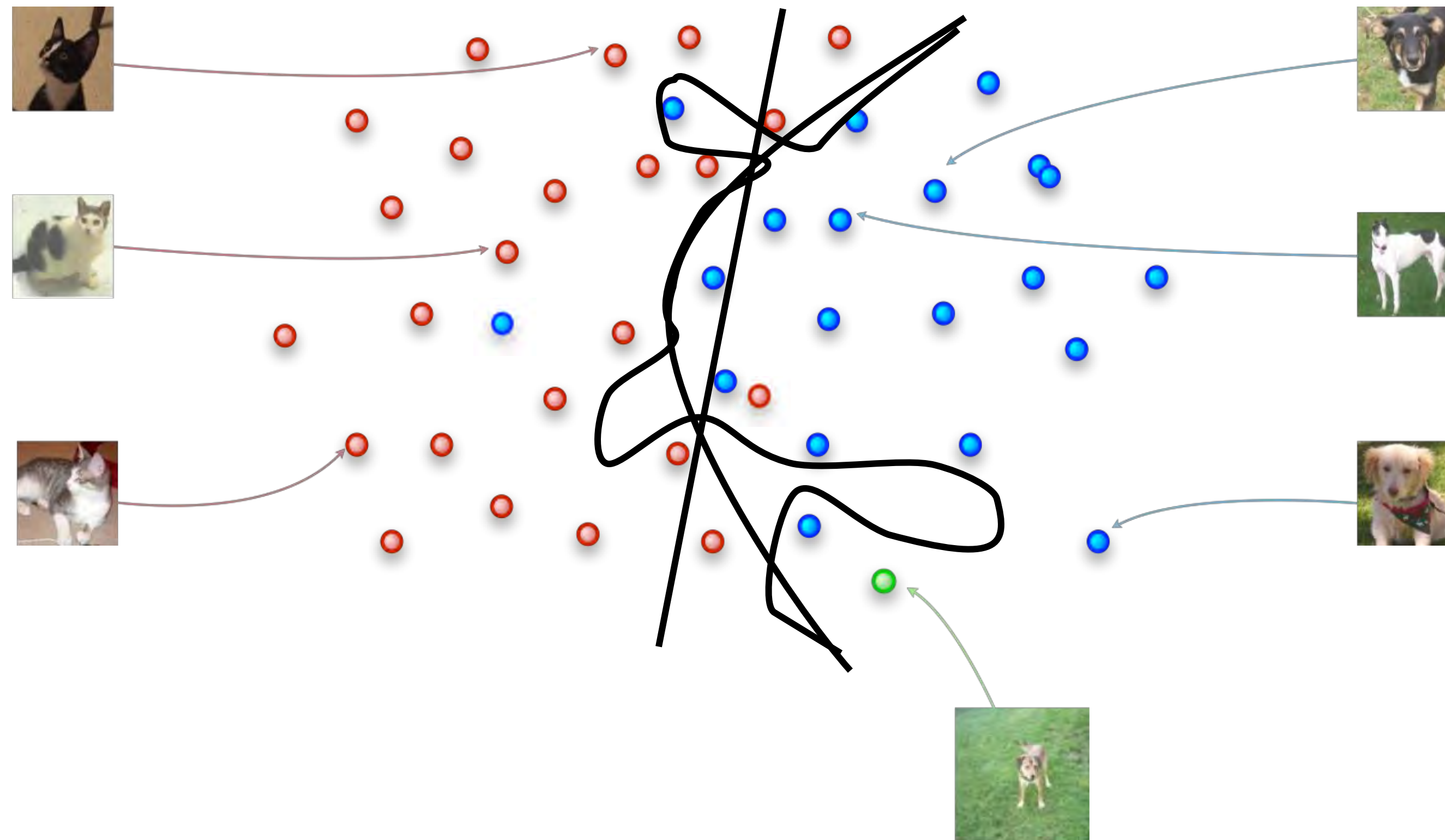
1. **From raw data to information**
  - Assimilate complex sensor data
  - Deal with volume, accuracy, diversity
2. **From information to models**
  - Understand how people move
  - Model and predict demand



# Machine Learning demystified



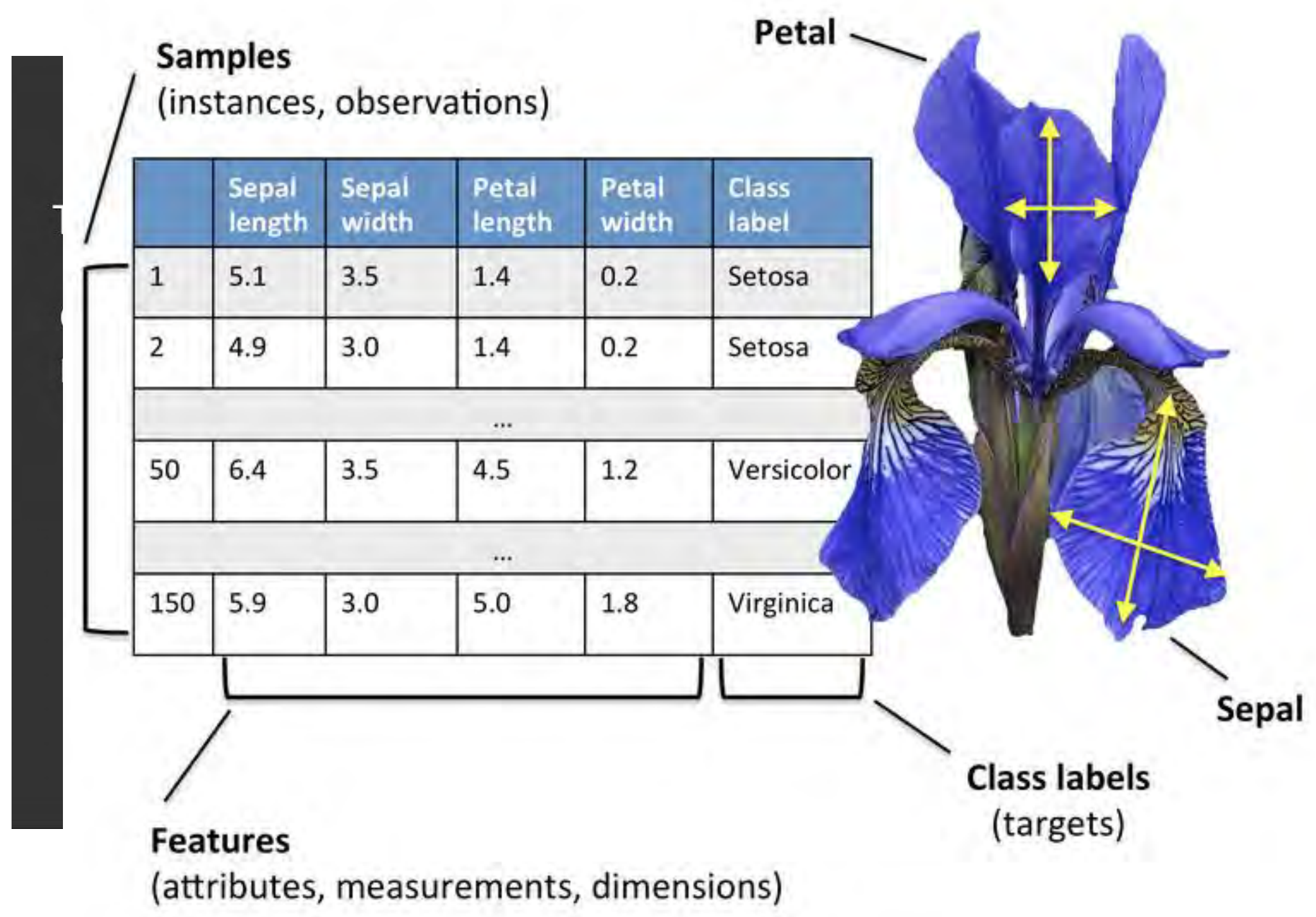
# Machine Learning demystified



# Ask, Get and Explore

## Hands-on example

UCI Machine Learning Repository



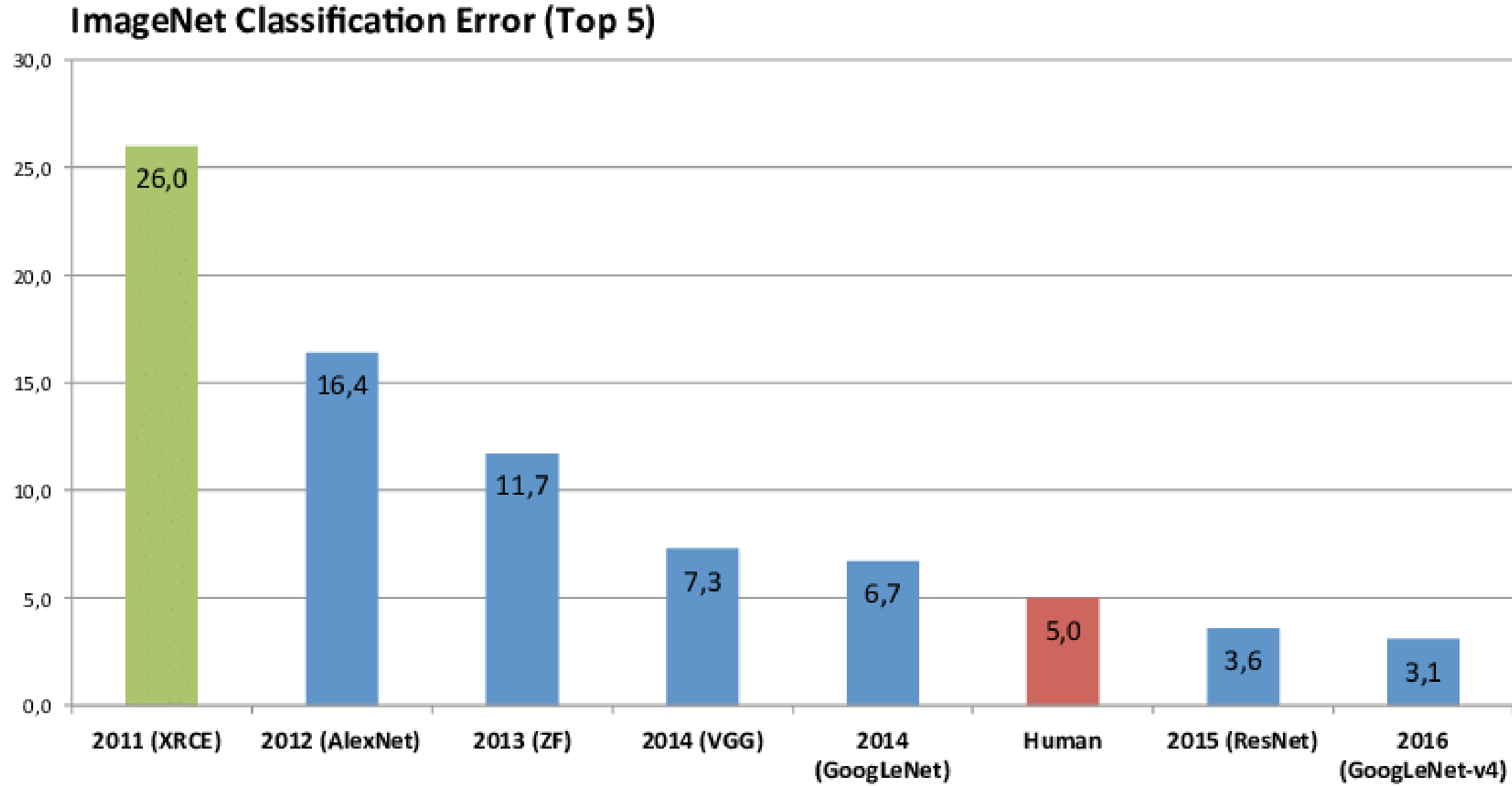
# Recent Advances in AI

## Convolutional Neural Networks (CNN)

### ImageNet Challenge

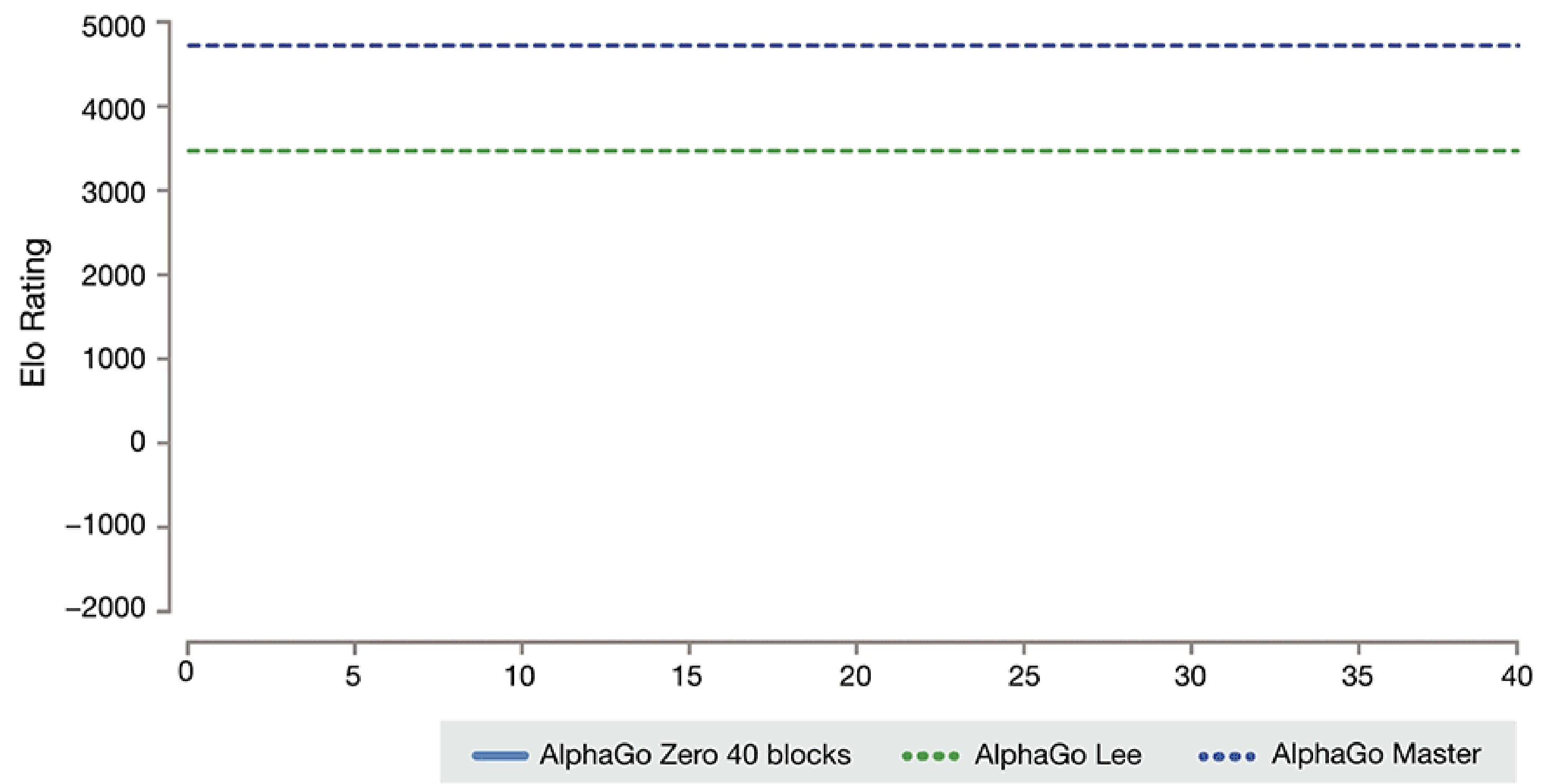
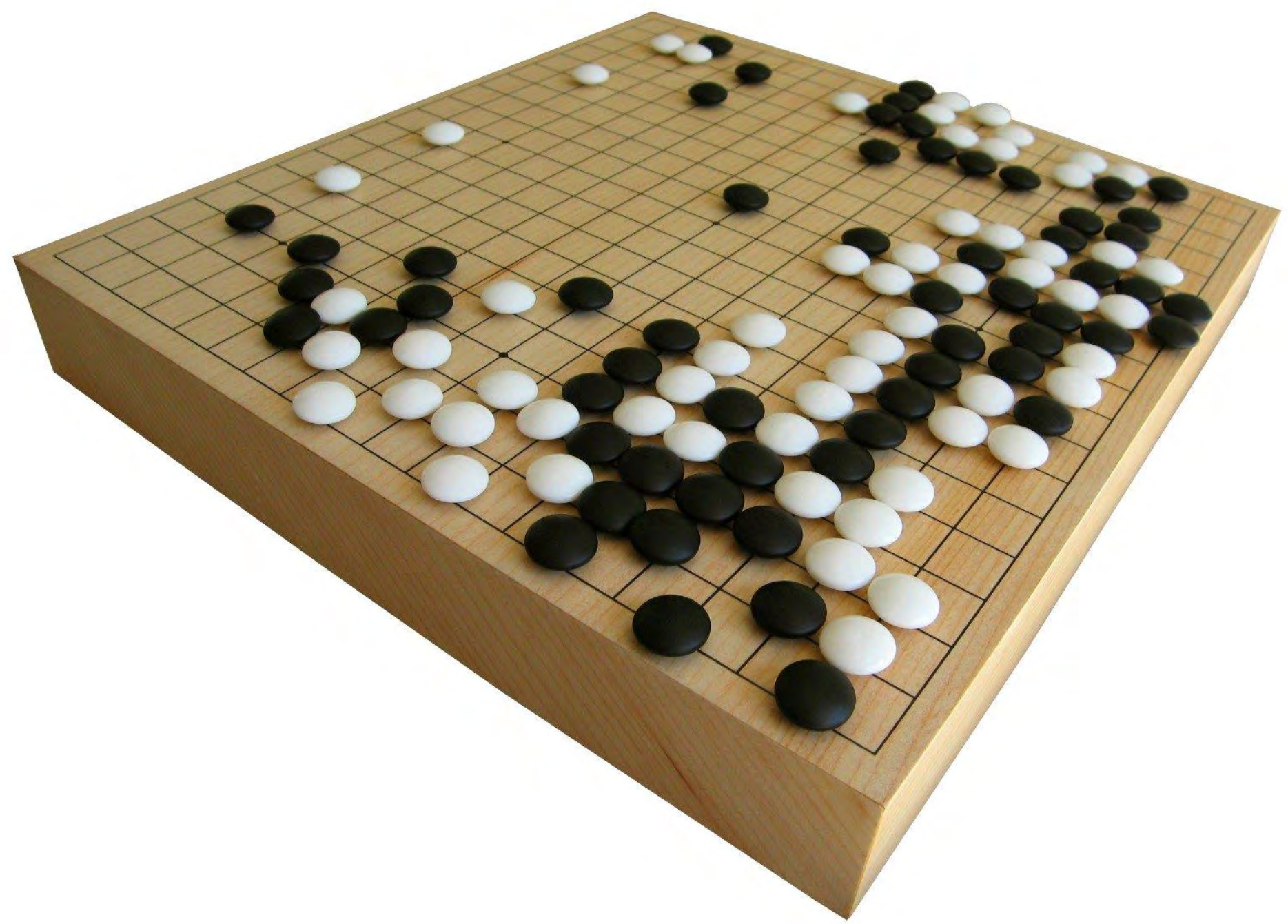


- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



# Recent Advances in AI

## Deep Reinforcement Learning in AlphaGo Zero





# This success relies on...

---

1. Large dataset of labelled data

2. Good quality data

3. Enough computing power

4. Clear and measurable objectives

# An Unexpected Outcome



It's an Indian elephant!

# Another Unexpected Outcome

Qui est Olivier Verscheure d'après ChatGPT (GPT 4.0)

Olivier Verscheure est *un scientifique de renom dans le domaine de l'analyse de données et de l'intelligence artificielle. ~~Il a reçu un Ph.D. en informatique de l'Université catholique de Louvain, en Belgique.~~*

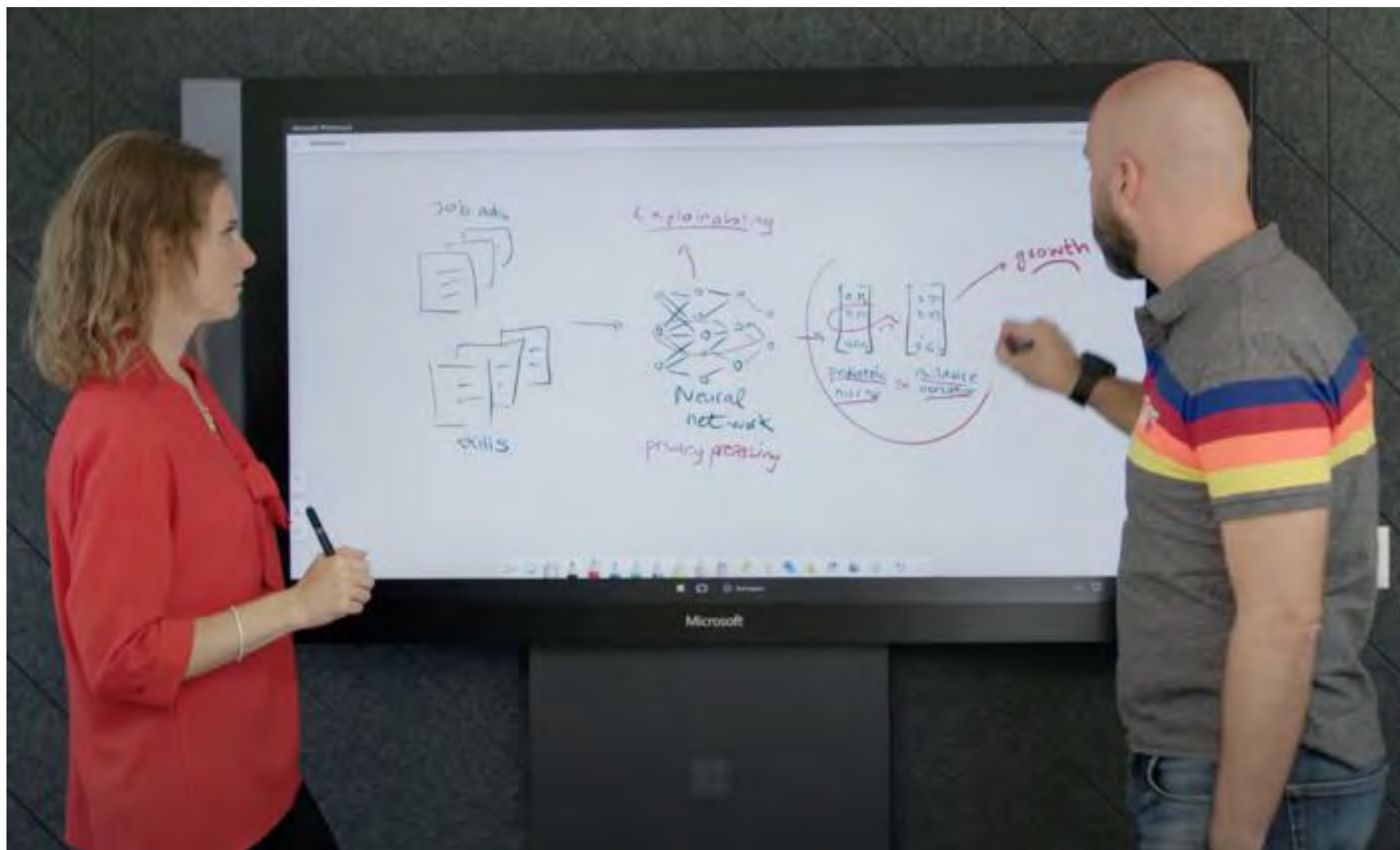
Après avoir obtenu son doctorat, Verscheure a travaillé pour IBM Research pendant plus d'une décennie. Pendant son séjour chez IBM, il a contribué à de nombreuses avancées significatives dans le domaine du traitement des données.

~~Verscheure a ensuite rejoint Swisscom, le principal fournisseur de télécommunications en Suisse, où il a dirigé le département Swisscom Data, Science & Analytics Lab.~~ Verscheure est également co-fondateur du Swiss Data Science Center, une collaboration entre l'École polytechnique fédérale de Lausanne (EPFL) et l'École polytechnique fédérale de Zurich (ETH Zurich). Le centre vise à promouvoir l'innovation dans le domaine de l'analyse des données en Suisse.

Verscheure a publié de nombreux articles dans des revues scientifiques et a également été conférencier invité à diverses conférences internationales. *Son travail est reconnu pour son approche innovante de l'analyse des données et de l'intelligence artificielle.*

Cependant, je vous recommande de rechercher des informations plus récentes en ligne pour obtenir les détails les plus précis et à jour sur Olivier Verscheure, car ma dernière mise à jour des données est de septembre 2021.

# Data bias



Midjourney prompt: "A photo of two people standing in front of a white drawing board in the middle. On the left is a female with blond curly hair, wearing a red top and black pants. On the right is a bald male with a beard, wearing a grey polo. **The female holds a pen in her hand and she is showing schematics on the white board.** --ar 16:9 --quality 2"

# A Disturbing Outcome



Turning a *STOP* sign into a 45 MPH speed limit

# Things are improving but...

ChatGPT 4

**You**  
En quelle(s) année(s) était-il chez Swisscom?

**ChatGPT**  
Olivier Verscheure était chez Swisscom de 2015 à 2016. Durant cette période, il a occupé le poste de Directeur du Swisscom Digital Lab à l'EPFL, où il a travaillé sur divers projets de recherche et de développement en collaboration avec Swisscom [1].

**You**  
Pour quel projet chez Swisscom Olivier est-il particulièrement reconnu?

**ChatGPT**  
Olivier Verscheure est particulièrement reconnu pour son travail au Swisscom Digital Lab à l'EPFL, où il a dirigé des projets de recherche et de développement en collaboration entre Swisscom et l'EPFL. Cependant, les sources consultées ne fournissent pas de détails spécifiques sur un projet particulier chez Swisscom pour lequel Olivier Verscheure serait spécifiquement reconnu. En tant que Directeur du Swisscom Digital Lab, son rôle impliquait de superviser plusieurs projets de recherche et d'innovation dans le domaine du numérique et des télécommunications, visant à explorer et à implémenter des solutions technologiques avancées. Son travail a contribué à renforcer la collaboration entre Swisscom et l'EPFL, en favorisant le développement de technologies innovantes [1].



En résumé, Olivier Verscheure est un chercheur et un leader dans le domaine de l'informatique et de la science des données, dont le travail a eu un impact significatif tant dans le milieu académique qu'industriel.

# Things are improving but...

ChatGPT 4



You

A-t-il travaillé chez Swisscom?



ChatGPT

Oui, Olivier Verscheure a travaillé chez Swisscom avant de rejoindre le Swiss Data Science Center (SDSC). Avant de devenir directeur exécutif du SDSC, il était le Directeur du Swisscom Digital Lab à l'EPFL (École Polytechnique Fédérale de Lausanne), où il a supervisé des projets de recherche et de développement en collaboration entre Swisscom et l'EPFL. Dans ce rôle, il a contribué à l'exploration et à l'implémentation de nouvelles technologies et solutions numériques, en mettant l'accent sur l'innovation dans le domaine des télécommunications et des services numériques.

Son passage chez Swisscom a permis de renforcer les liens entre le secteur industriel et le milieu



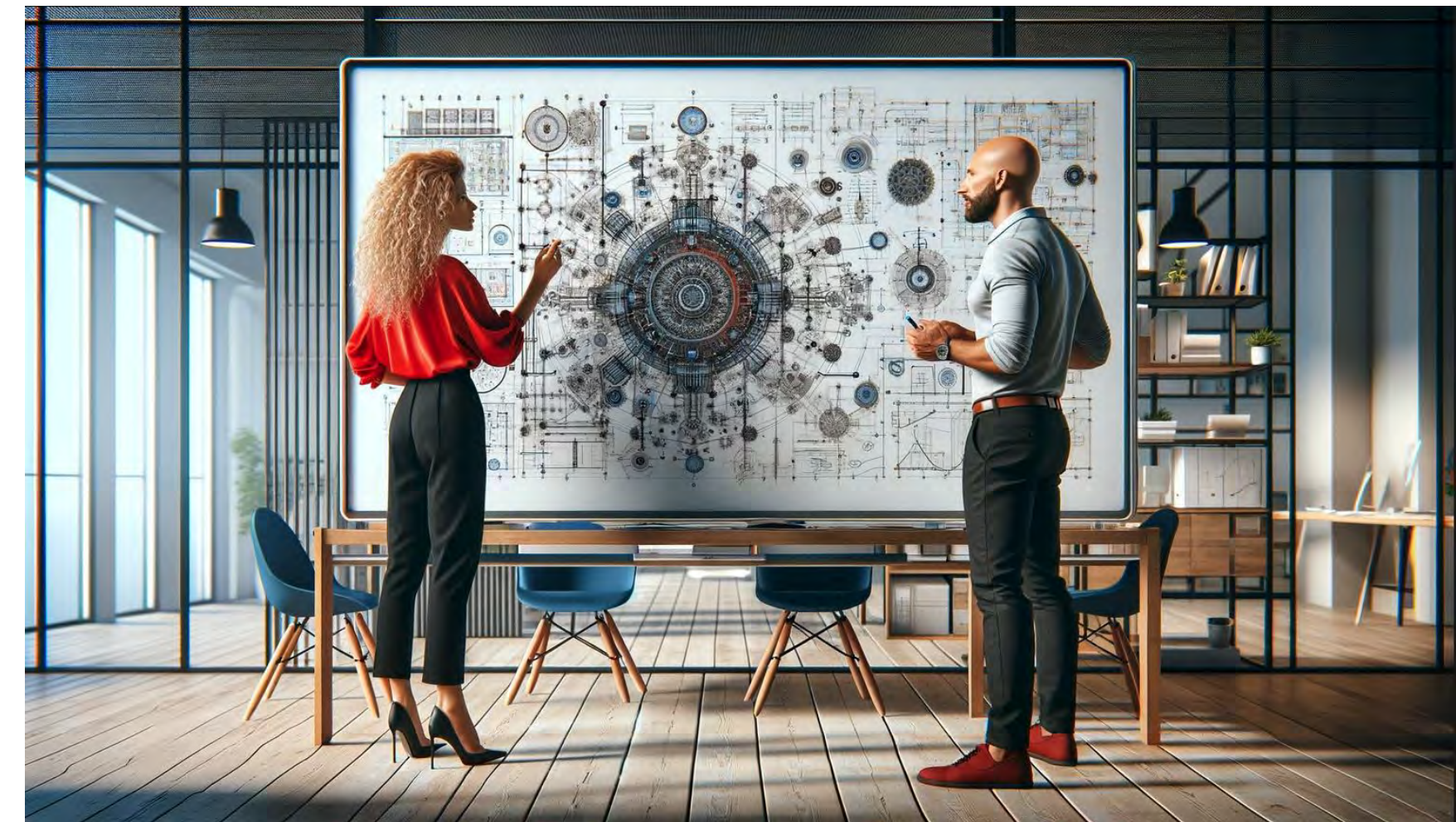
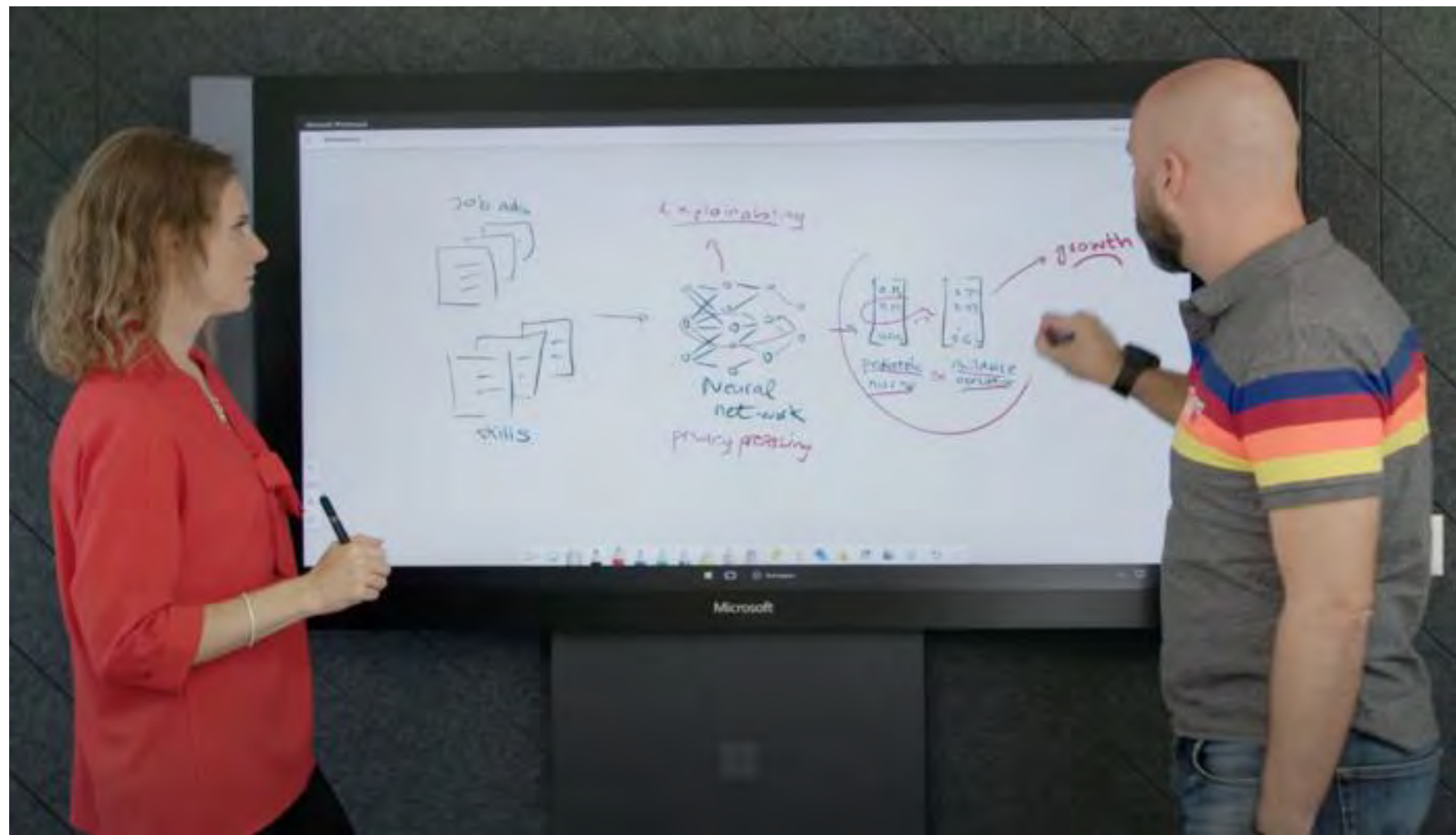
Olivier Verscheure - Executive Education HEC Lausanne



ent et l'application de technologies avancées dans des contextes réels [1].

👍 🗨️ 📄

# Things are improving but...



ChatGPT 4

**You**  
The woman should be holding the pen, not the man.

**ChatGPT**  
The image was designed with the woman holding the pen while explaining the schematics on the whiteboard, and the man listening attentively. If it seems there was a misunderstanding or if the image doesn't align with your expectations, my apologies for any confusion. I can't modify the image directly, but I can create a new one if you have further adjustments in mind. Would you like me to try again with any specific changes?





# Obstacles to a wider adoption

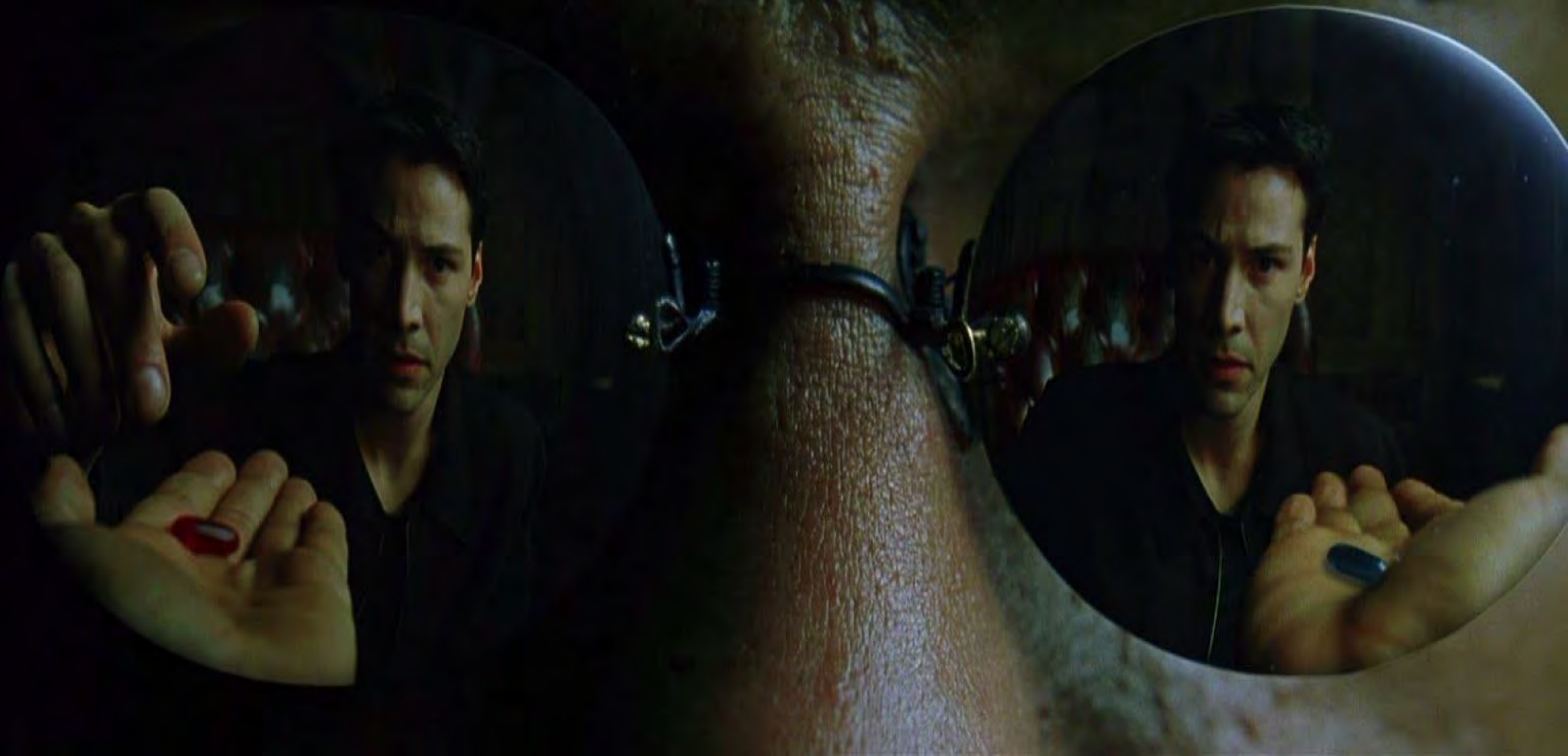
1. ~~Large dataset of labelled data~~ -> Labelling is expensive

2. ~~Good quality data~~ -> Data is usually missing/Increased uncertainty

3. ~~Clear and measurable objectives~~ -> Knowledge discovery/causality

4. Lack of interpretability/lack of trust

# Challenge #3 – Building trust



# Involvement of the Stakeholders

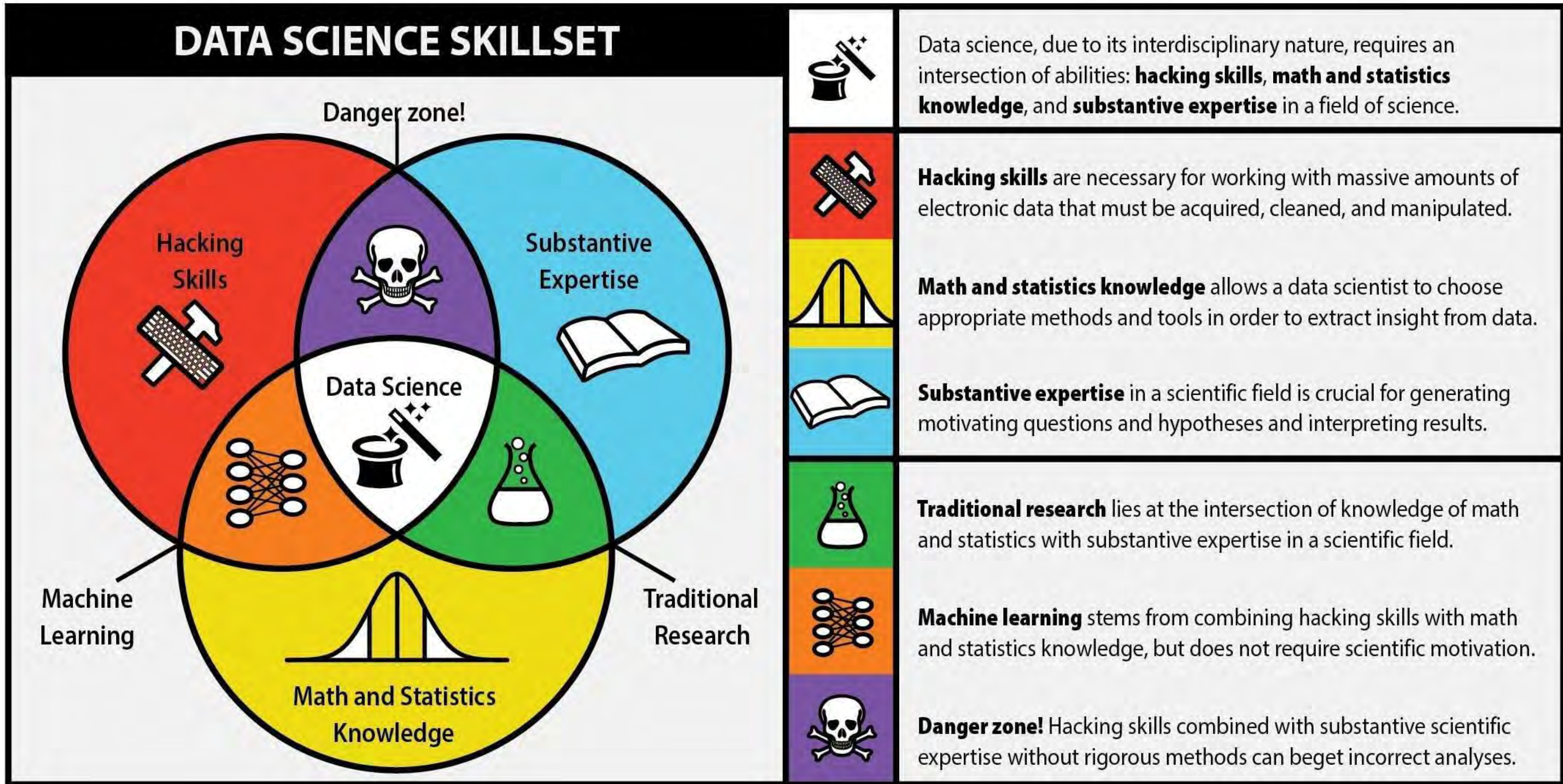
Domain expertise

Data + Algorithms → Knowledge → Benefits

*Big Data /  
Data lake*

*Machine Learning /  
“Narrow” A.I.*

# From AI/Machine Learning to Data Science



# Use Cases from Swiss Industry



Bühler: Optimize milling plant component settings to save energy



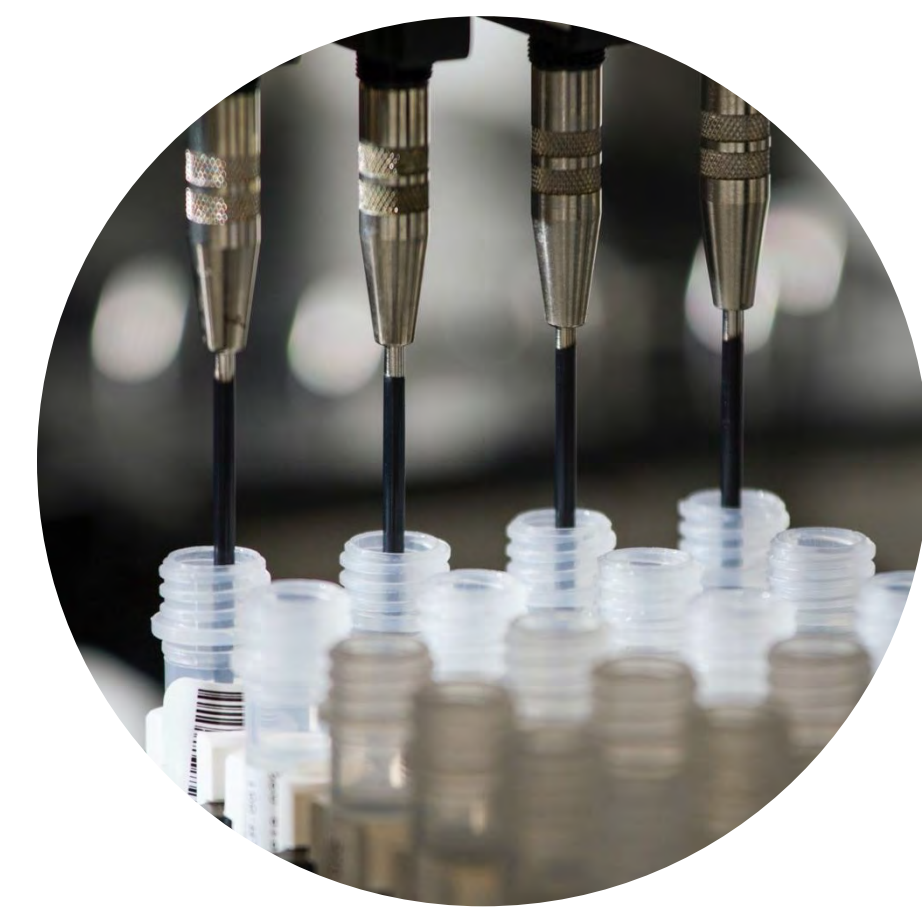
Adecco: Match job offers to candidates based on skills



Private bank: Detect money laundering activities



Richemont: Allocate luxury products to market based on demand



Pharma: Predict adverse effects of tumor drugs

IN GOD WE TRUST.



ALL OTHERS MUST BRING DATA.

– W. EDWARDS DEMING, STATISTICIAN, PROFESSOR, AUTHOR

---

# Cours pratique semaine 6

Pandas et Numpy en revue  
Visualisation

---

# Pandas – l’outil d’analyse des données

- Data Series et Data Frames
- chargement, sauvegardes, par exemple à partir de fichiers (csv)
- sélection, classement, agrégations, et traitement (transformation)



# Pandas – l’outil d’analyse des données

- Data Series et Data Frames
- chargement, sauvegardes, par exemple à partir de fichiers (csv)
- sélection, classement, agrégations, et traitement (transformation)
- jointures – [join](#) (ou [merge](#))
- groupes et agrégations groupées – [groupby](#) (objets [GroupBy](#))
- fenêtres temporelle – [rolling](#) ([Windowing](#))
- interpolation temporelle – [resampling](#)
- pivoter, réarranger des DataFrames – [pivot](#) et son opération inverse [melt](#) ([Reshaping](#))
- visualiser (un des objectifs de ce cours)

# Les données numériques avec Numpy

- Bibliothèque numérique au cœur de presque tout l'écosystème Data Science en Python (dont pandas)
- Traitement et stockage efficace de larges tableaux numériques (ndarray)
  - **Multidimensionnel** - dimension 0: scalaire, 1: vecteur, 2: matrice, 3+: tenseur
  - **Homogène** - tous les éléments sont du même type
- Un nombre important de fonctions mathématiques

# Concepts numpy vus cette semaine

- Attributs de tableaux numpy
  - `shape`: le nombre de dimensions et nombre d'éléments par dimension
- Créations de tableaux numpy
  - `arange`, `linspace`, `zeros`, `ones`, `full`, `random`.\*
- Changer le format d'un tableau numpy
  - `reshape`
- Opérations sur les tableaux numpy
  - `dot`, `sum`, `@`, `multiply`, `.T` (transpose)

# La visualisation scientifique avec python

- **matplotlib**

- Le paquet de visualisation Python, sur lequel sont fondés beaucoup d'autres paquets de visualisations
- Adapté pour une visualisation rapide et minimaliste d'exploration
- Programmation de visualisation bas niveau (plus de code)

- **pandas**

- Offre des raccourcis d'aide à l'exploration visuelle de la donnée directement à partir de dataframes
- Basé sur matplotlib (par défaut)

- Et autres méthodes de visualisation avancées que nous verrons par la suite

# La visualisation scientifique avec pandas

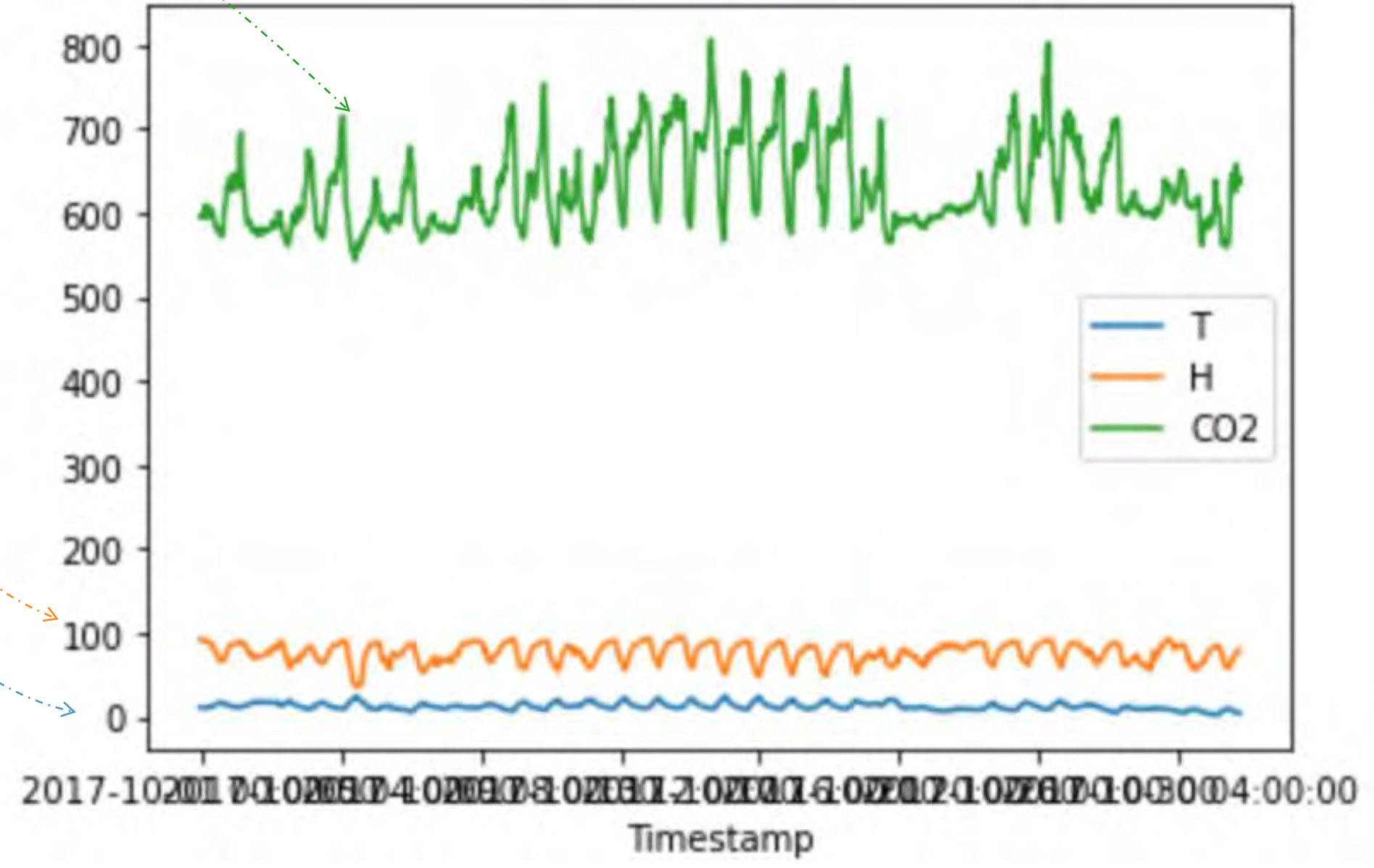
- Méthodes de visualisation rapide des objets DataFrame - `pandas.DataFrame.plot.*`
  - `.plot`: méthode générique (line plot) pour tracer des ensembles de données
  - `.plot.hist`: histogrammes
  - `area`, `bar`, `barh`, `density`, `hexbin`, `kde`, `line`, `pie`, `scatter` ... ([documentation](#))

# Visualisation avec pandas - exemples

```
df = pandas.read_csv('sensor1044.csv', index_col='Timestamp')
```

```
df.plot()
```

	T	H	CO2
Timestamp			
2017-10-01 00:00:00	11.396667	92.440806	595.004391
2017-10-01 00:30:00	11.325000	91.770691	597.017879
2017-10-01 01:00:00	11.240000	91.096761	596.487010
2017-10-01 01:30:00	11.260000	91.275734	598.787639
2017-10-01 02:00:00	11.290000	91.061157	595.366394
...	...	...	...
2017-10-31 21:30:00	5.275000	76.427979	647.675754
2017-10-31 22:00:00	5.060000	77.560944	633.452211
2017-10-31 22:30:00	4.826667	78.838867	634.303983
2017-10-31 23:00:00	4.650000	79.272471	643.250212
2017-10-31 23:30:00	4.456667	79.830688	636.954641

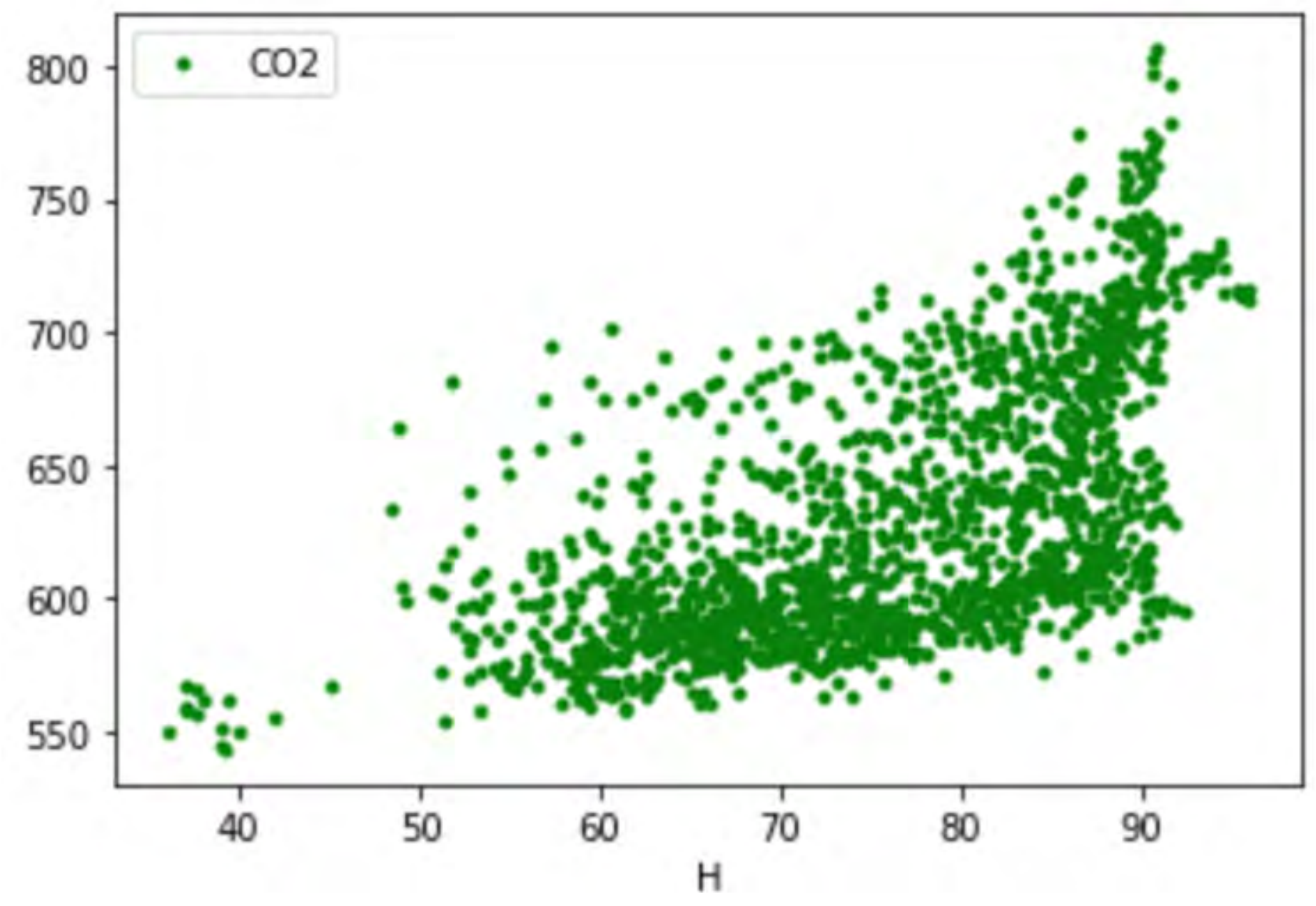


colonne index

# Visualisation avec pandas – exemples

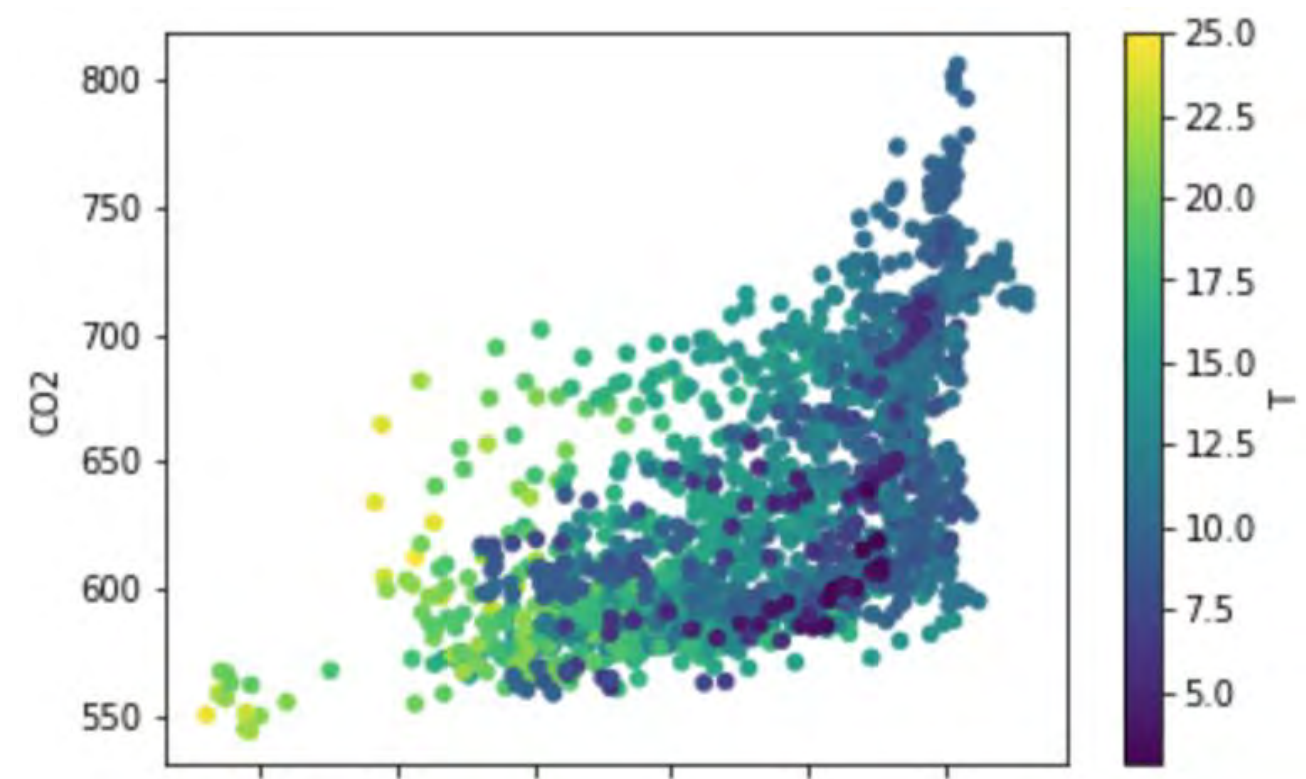
```
df = pandas.read_csv('sensor1044.csv', index_col='Timestamp')  
df.plot(x='H', y='CO2', style='.')
```

	T	H	CO2
Timestamp			
2017-10-01 00:00:00	11.396667	92.440806	595.004391
2017-10-01 00:30:00	11.325000	91.770691	597.017879
2017-10-01 01:00:00	11.240000	91.096761	596.487010
2017-10-01 01:30:00	11.260000	91.275734	598.787639
2017-10-01 02:00:00	11.290000	91.061157	595.366394
...	...	...	...
2017-10-31 21:30:00	5.275000	76.427979	647.675754
2017-10-31 22:00:00	5.060000	77.560944	633.452211
2017-10-31 22:30:00	4.826667	78.838867	634.303983
2017-10-31 23:00:00	4.650000	79.272471	643.250212
2017-10-31 23:30:00	4.456667	79.830688	636.954641

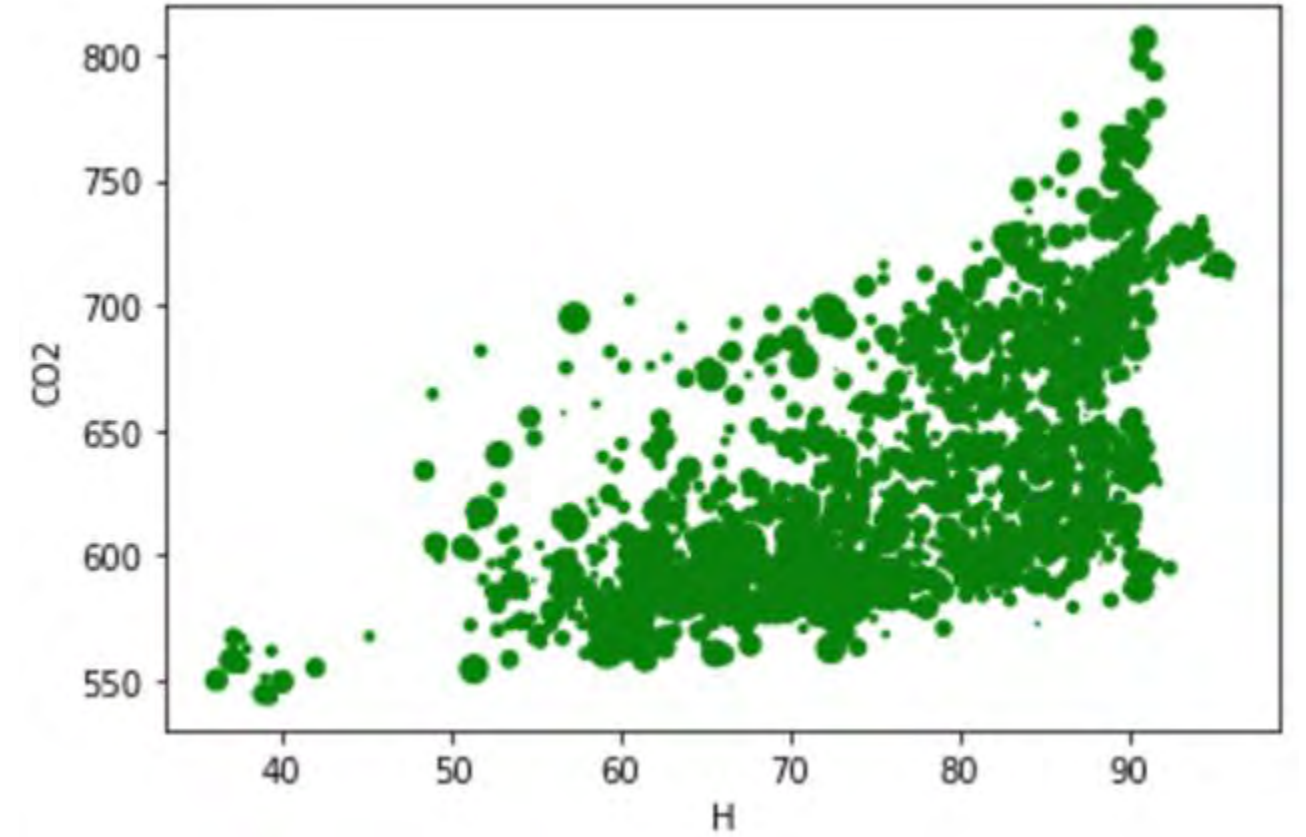


# Visualisation avec pandas - exemples

```
df.plot.scatter(x='H', y='CO2', c='T')
```



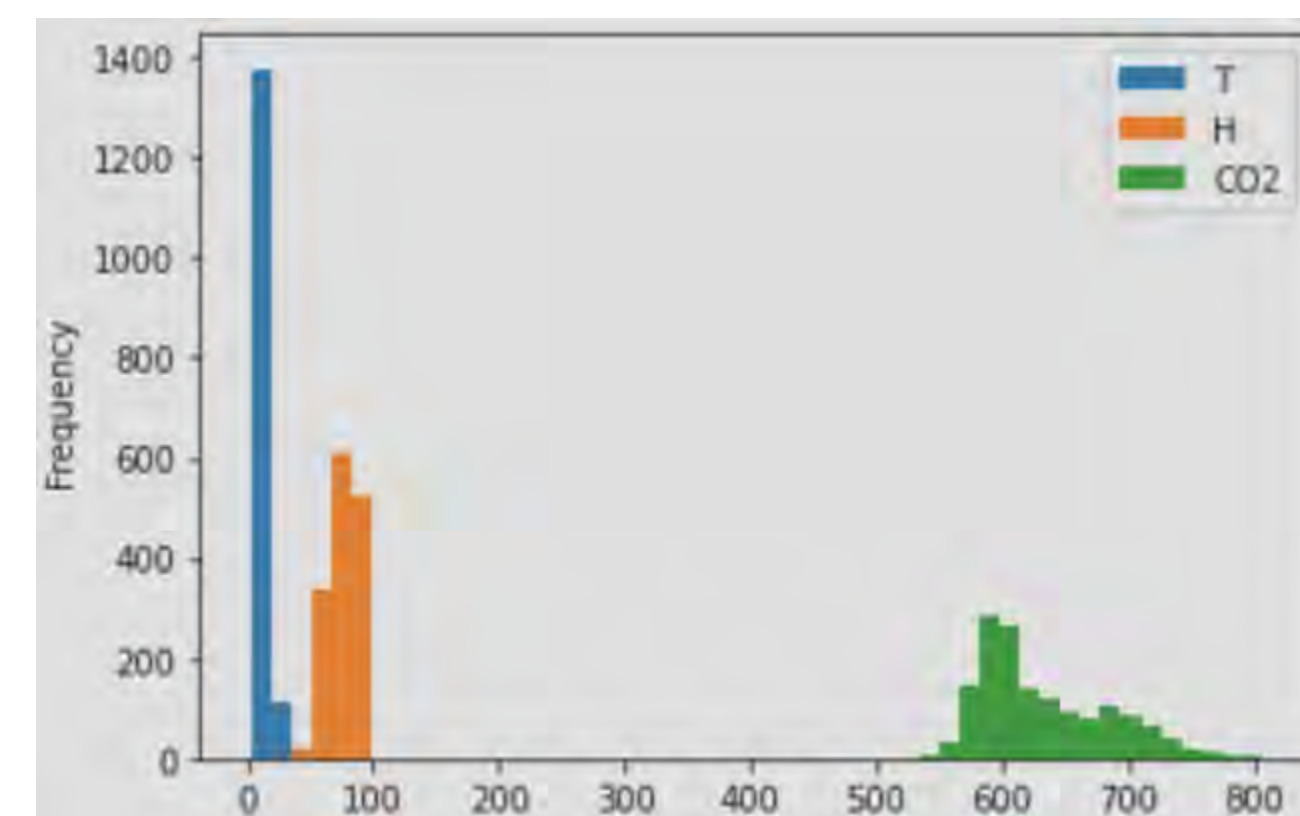
```
df.plot.scatter(x='H', y='CO2', s='T')
```



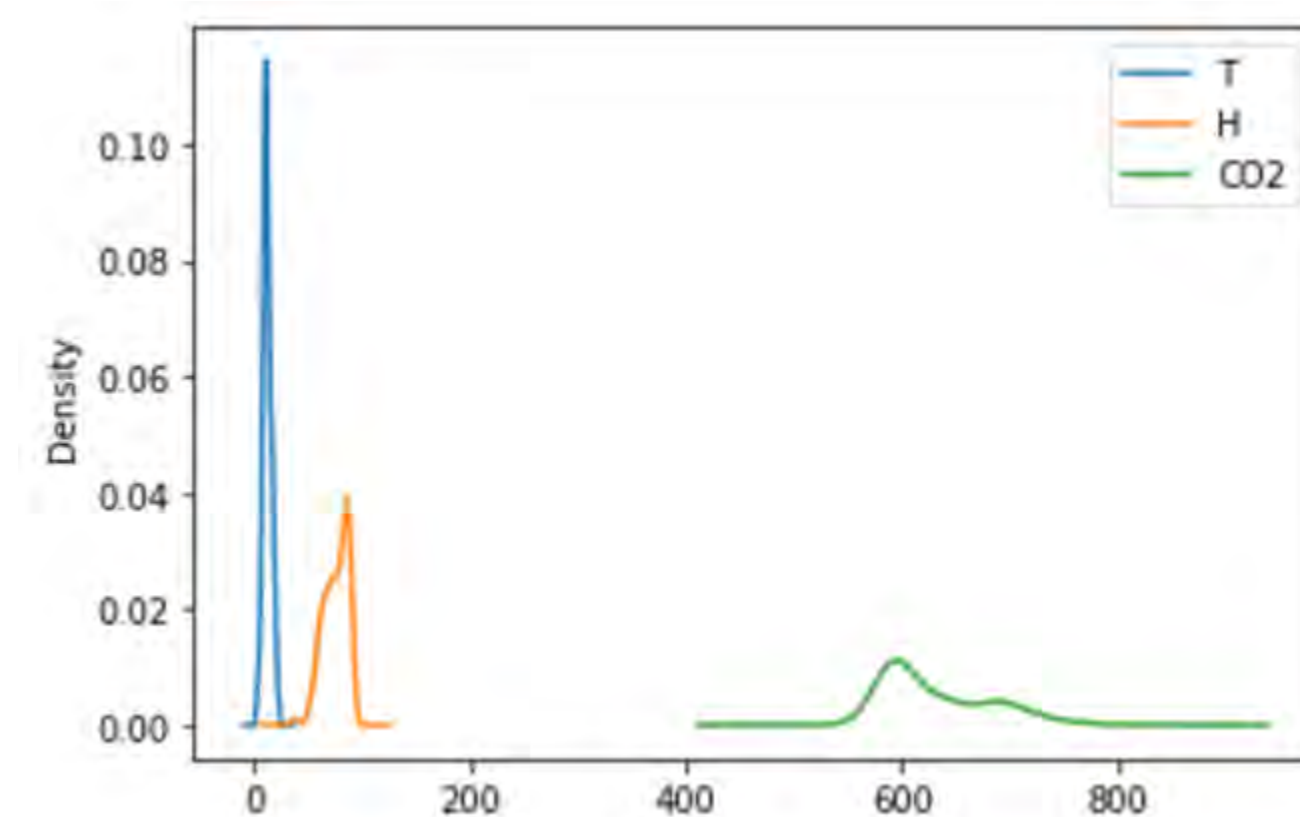


# Visualisation avec pandas - exemples

```
df.plot.hist(bins=50)
```



```
df.plot.density()  
df.plot.kde()
```



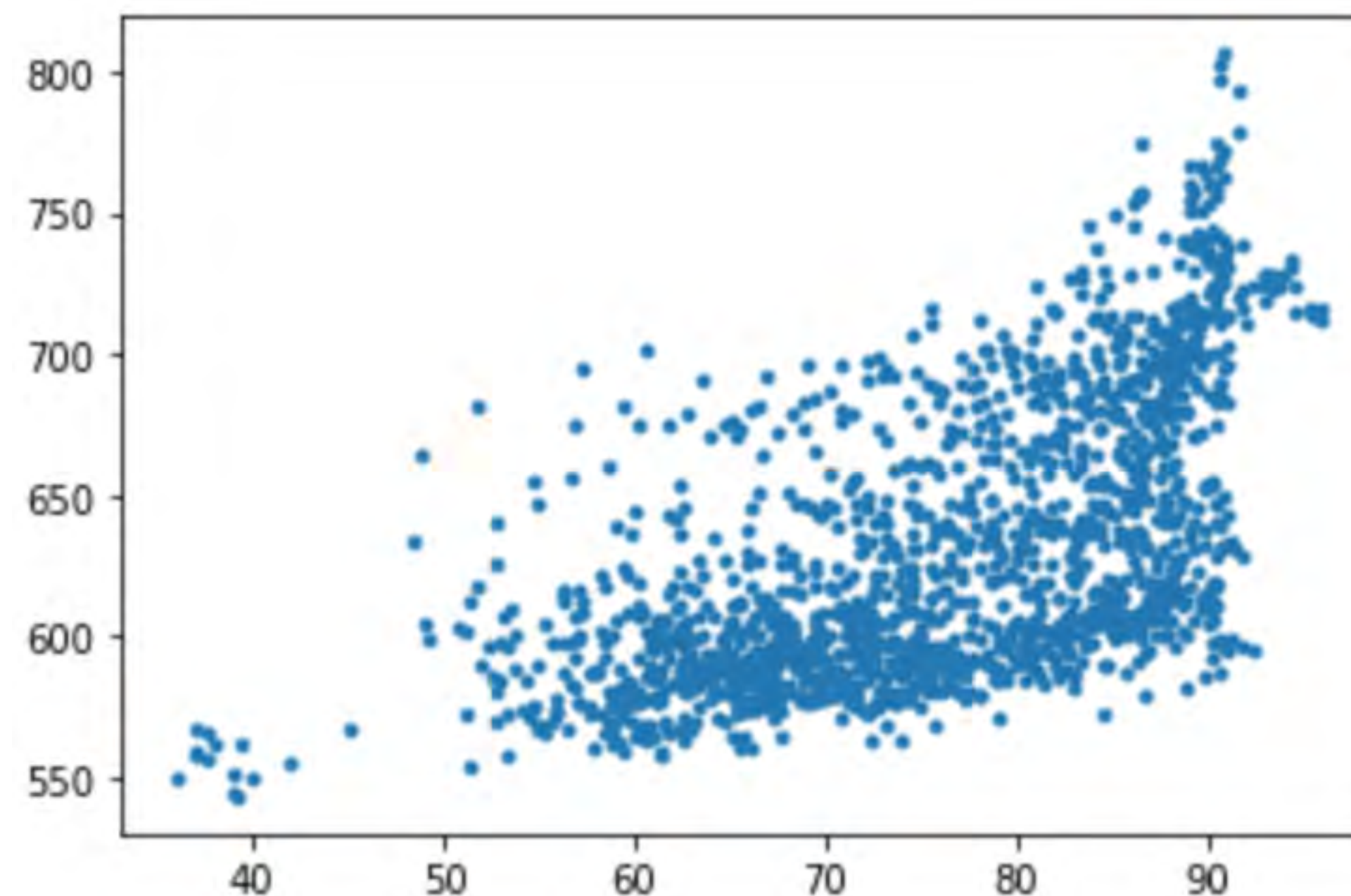
# Visualisation scientifique avec matplotlib

- Pour des fonctionnalités de visualisation plus avancées (que pandas)
  - Combiner des graphiques de différents types
  - Contrôle des axes, labels, (sous)titres, légendes, unités, placement et style
  - Affichage de données dynamiques
  - Méthodes de visualisation interactives
  - ...

# Visualisation scientifique avec matplotlib - base

```
import matplotlib.pyplot as plt

plt.plot(df.H, df.CO2, '.')
plt.show()
```



# Visualisation scientifique avec matplotlib - avancé

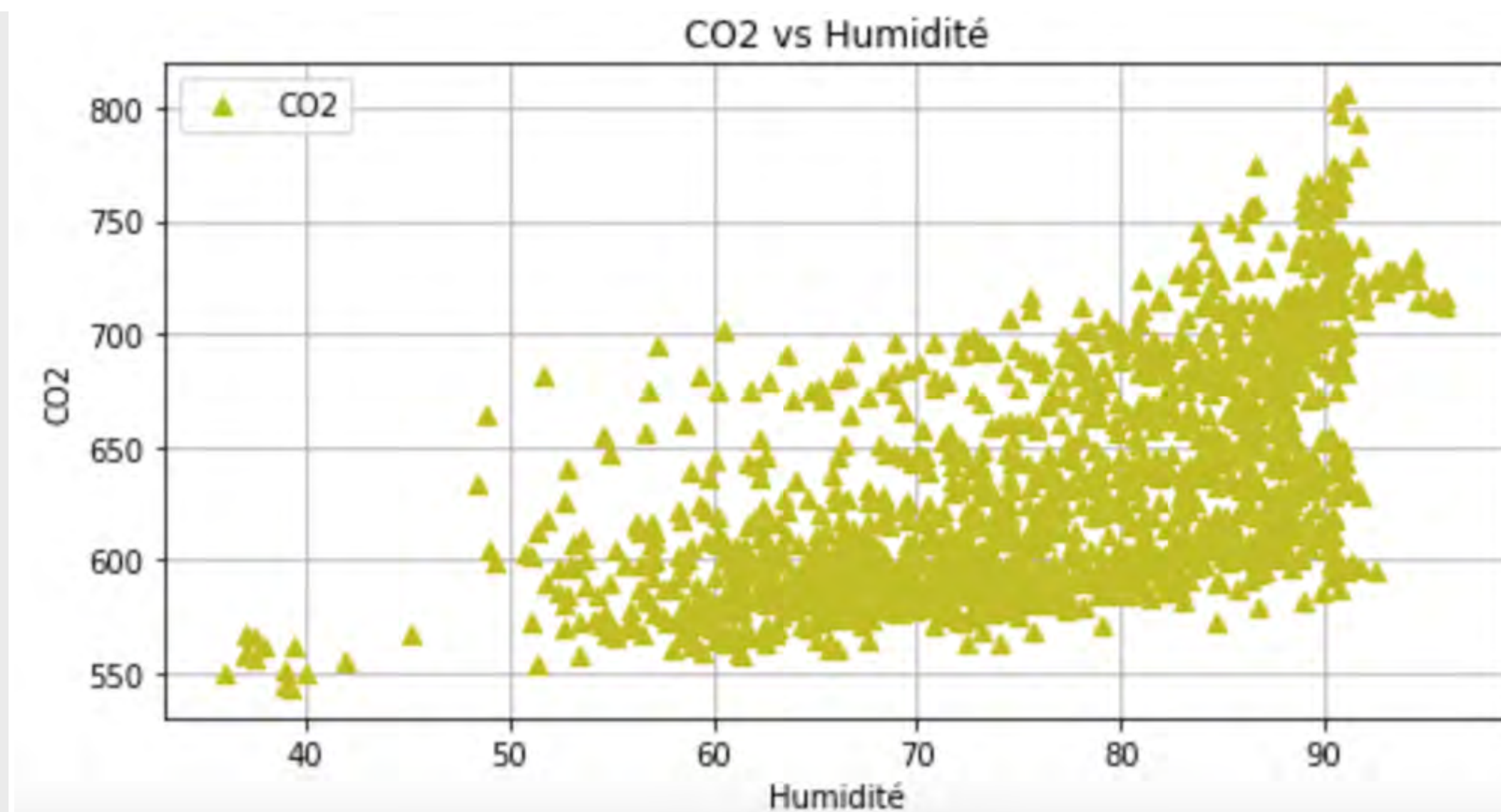
```
import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(8, 4))

ax.plot(df.H, df.CO2, '^--y', label='CO2')

ax.grid(True)
ax.legend(loc='best')
ax.set_title('CO2 vs Humidité')
ax.set_xlabel('Humidité')
ax.set_ylabel('CO2')

plt.show()
```



# Visualisation scientifique avec matplotlib - avancé

```
import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(8, 4))

ax.plot(df.H, df.CO2, '^-y', label='CO2')

ax.grid(True)
ax.legend(loc='best')
ax.set_title('CO2 vs Humidité')
ax.set_xlabel('Humidité')
ax.set_ylabel('CO2')

plt.show()
```

*Prépare l'emplacement du graphe et retourne les axes de la figure. La figure peut être composée de plusieurs sous-figures.*

*Construit le(s) graphe(s) – il est possible d'invoquer plusieurs méthodes de graphe utilisant des données différentes (plot, bar, hist, ...)*

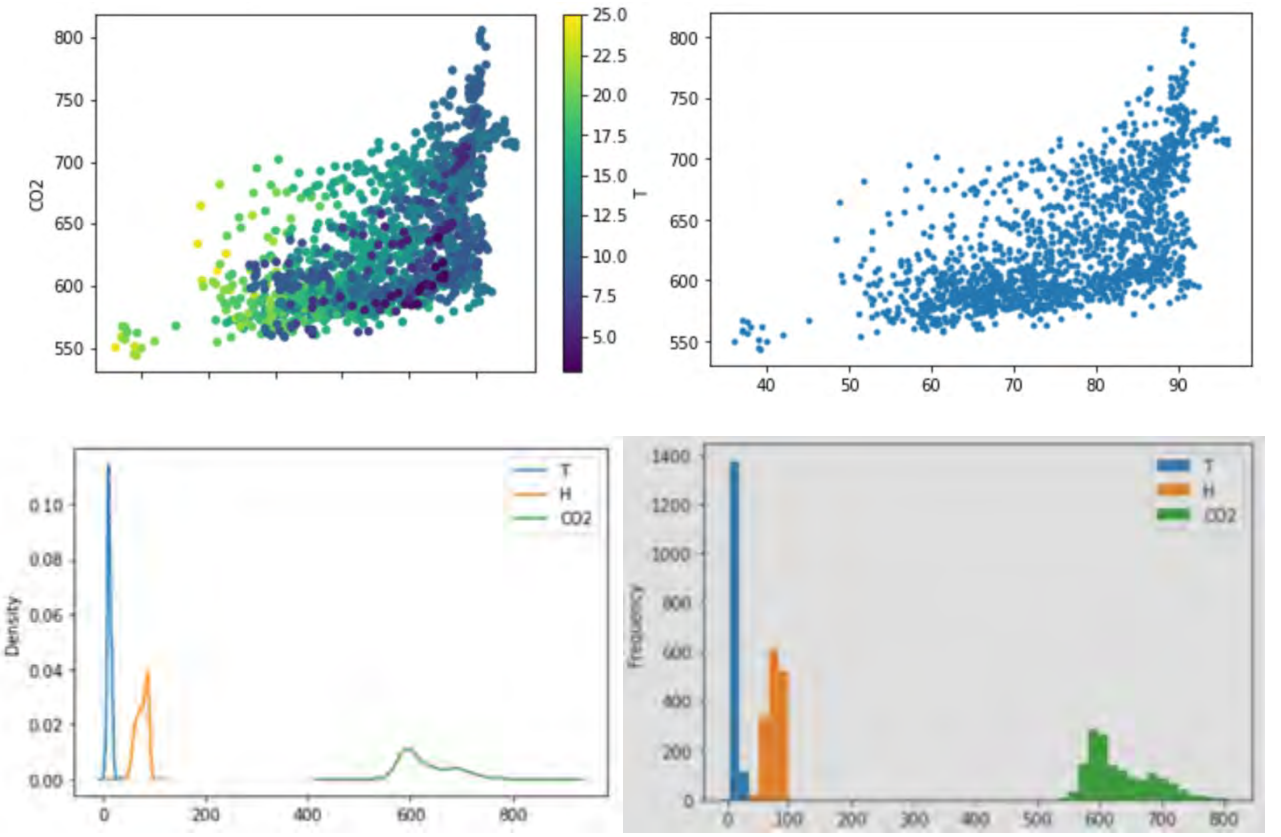
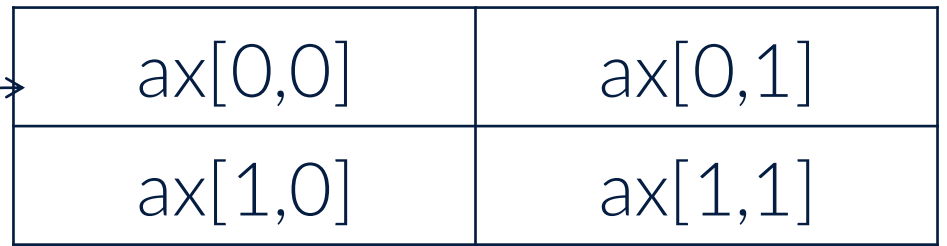
*Modifie les paramètres de la figure - le titre de la figure et des axes, l'affichage de la grille etc.*

*Affiche la figure*

# Visualisation scientifique avec matplotlib - avancé

```
import matplotlib.pyplot as plt  
  
fig, ax = plt.subplots(2, 2, figsize=(8, 4))  
  
ax[0, 0].plot(...)
```

Il est possible de créer une figure composée de plusieurs sous-figures arrangées en grille.



# Visualisation scientifique avec matplotlib - conclusions

- L'interface de programmation matplotlib est bas-niveau
  - très flexible mais nécessite "beaucoup" de code pour obtenir une visualisation présentable
  - la bibliothèque matplotlib est très avancée, n'hésitez pas à visiter le site [officiel](#)
- Autres solutions de visualisations
  - [plotly](#), seaborn, folium, cufflinks, bokeh, osmnx, pygal, vispy, ...

# Autoévaluation - objectifs

- Je suis capable de/d'...
  - Créer des tableaux de données aléatoires à partir de lois de distribution connues
  - Changer la forme (dimensions) d'un tableau de données
  - Effectuer des opérations mathématiques et opérations d'algèbre linéaire sur les tableaux numpy
  - Charger, sélectionner et visualiser des données avec pandas et matplotlib



## Exercices semaine 6

- Méthode 1 – VM de la VDI
  - Télécharger semaine06.zip dans ~/Desktop/myfiles/ENG209\_2024Fall
  - Dans un terminal

```
cd ~/Desktop/myfiles/ENG209_2024Fall
unzip semaine06.zip
./setup_part2.sh
```

- Les notebooks sont dans  
~/Desktop/myfiles/ENG209\_2024Fall/eng209-2024-partie-2/semaine06

## Exercices semaine 6

- Méthode 2 (avancé)

- Créer un compte <https://renkulab.io> (edu-ID)
- <https://renkulab.io/projects/eng209/eng209-2024/eng209-2024-partie-2>
- **Fork** Project
- **Start** Project
- Une fois le projet ouvert - dans un terminal: `./post-init.sh`
- Les notebooks sont dans `./semaine06`
- Note
  - Il est aussi possible d'ouvrir (Start) le projet anonymement sans créer de compte mais dans ce cas vous serez dans l'incapacité de sauver votre travail.