



Introduction au Big Data

Master 2 FD, IEF - Rennes

Sylvain BARTHELEMY

Qui suis-je?

gwen**lake.**

Build next-gen apps using AI

We help organizations by operationalizing AI, machine learning and data analytics

Objectif du cours

Vous avez terminé vos études, et on vous confie un projet de big data / IA / data science que vous devez mener à bien rapidement et faire une présentation à votre équipe à l'issue du projet.

Travaux menés pendant les cours

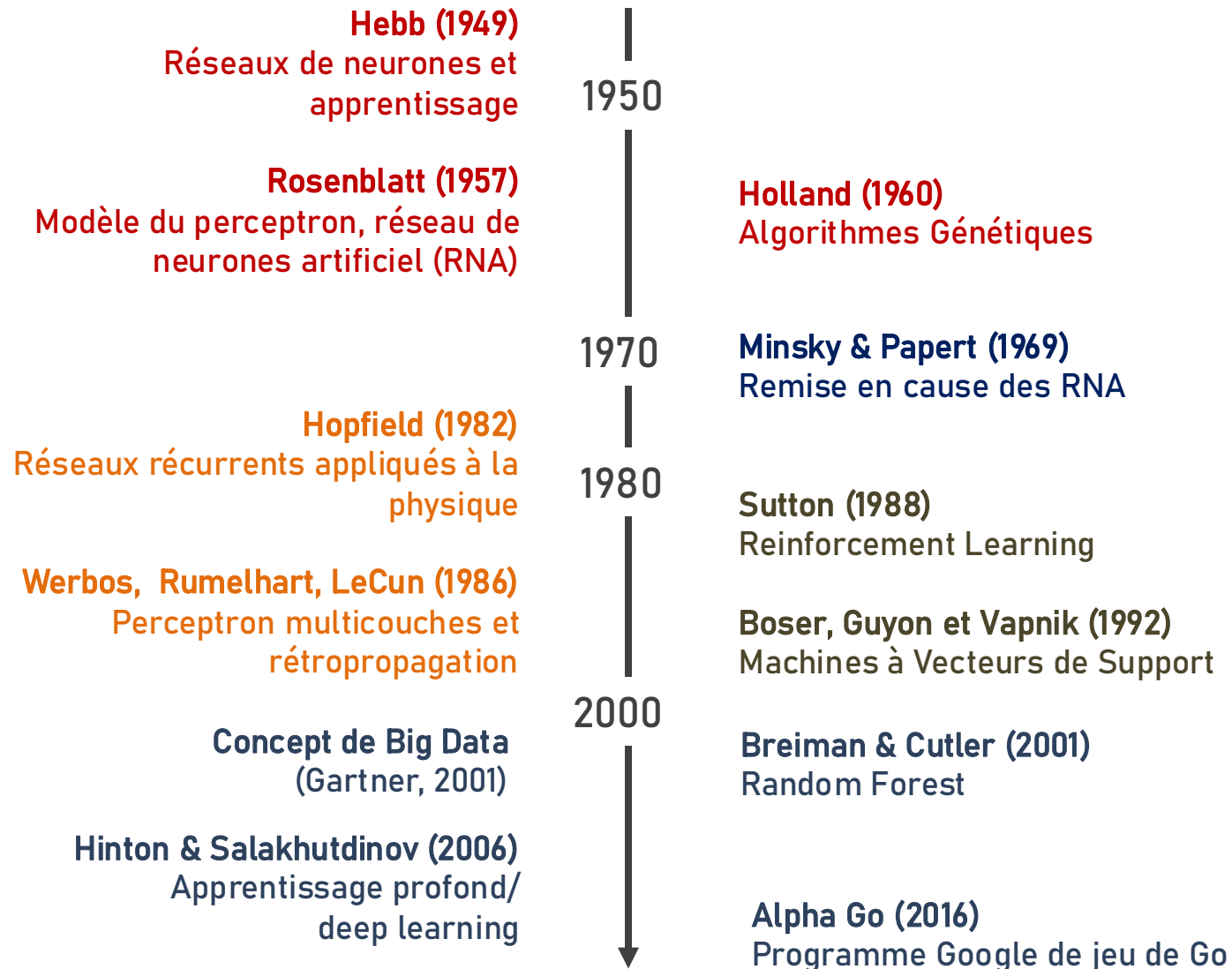
7 séances + présentation

**Evaluation sur la base de la qualité du travail mené,
comportement pendant les séances et présentation finale**

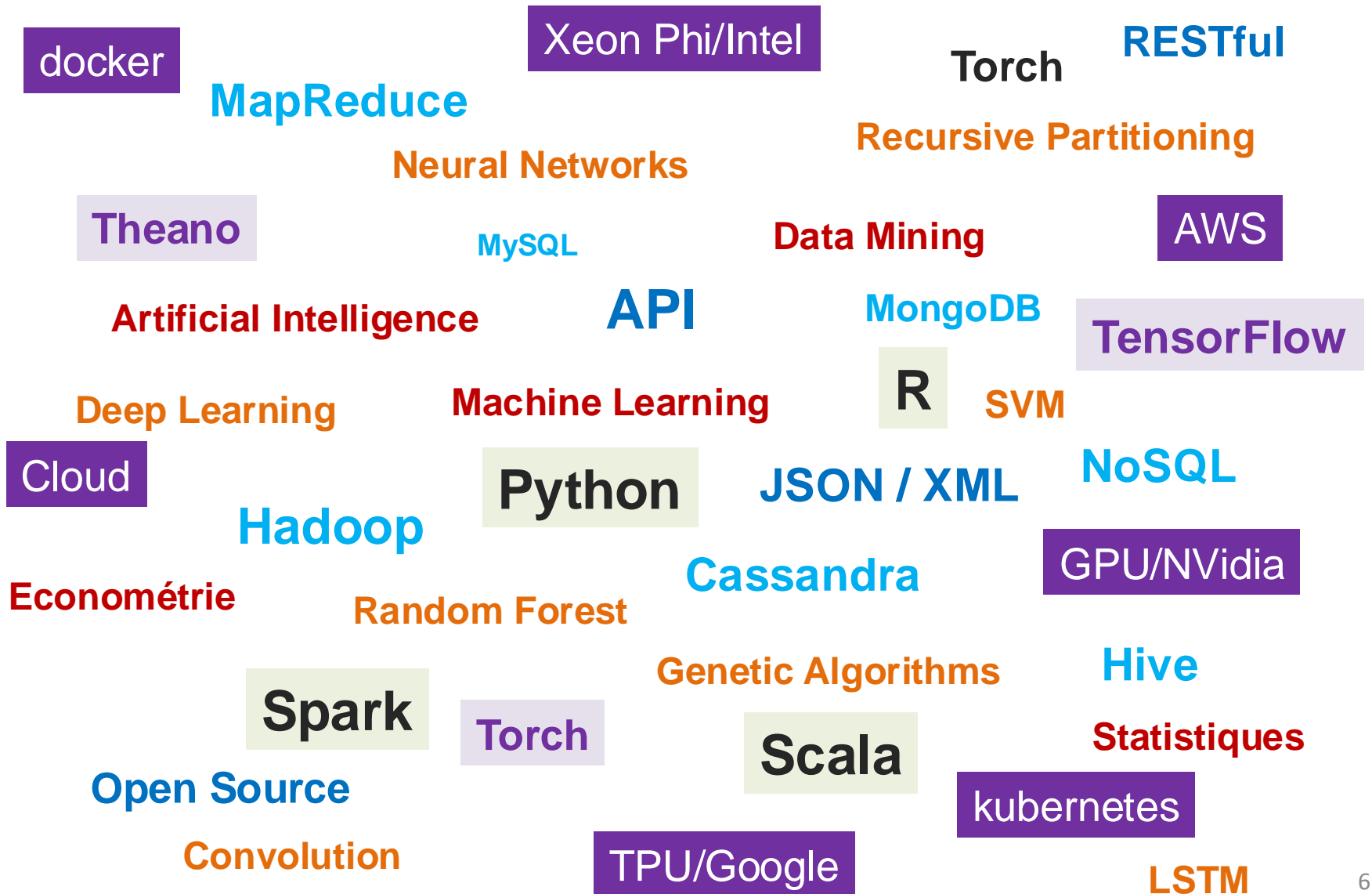


Le Big Data ?

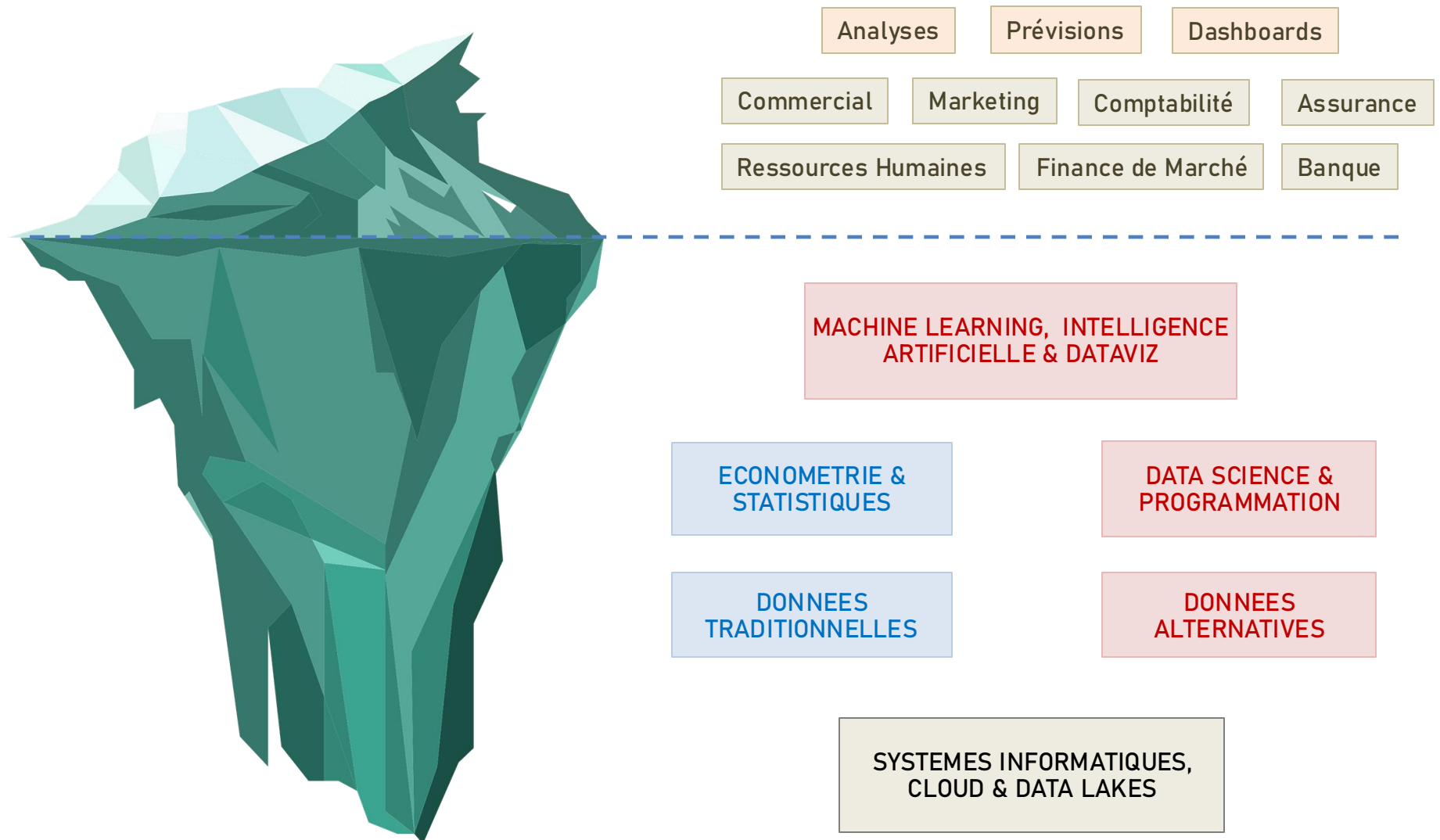
Une brève histoire du machine learning et de l'IA



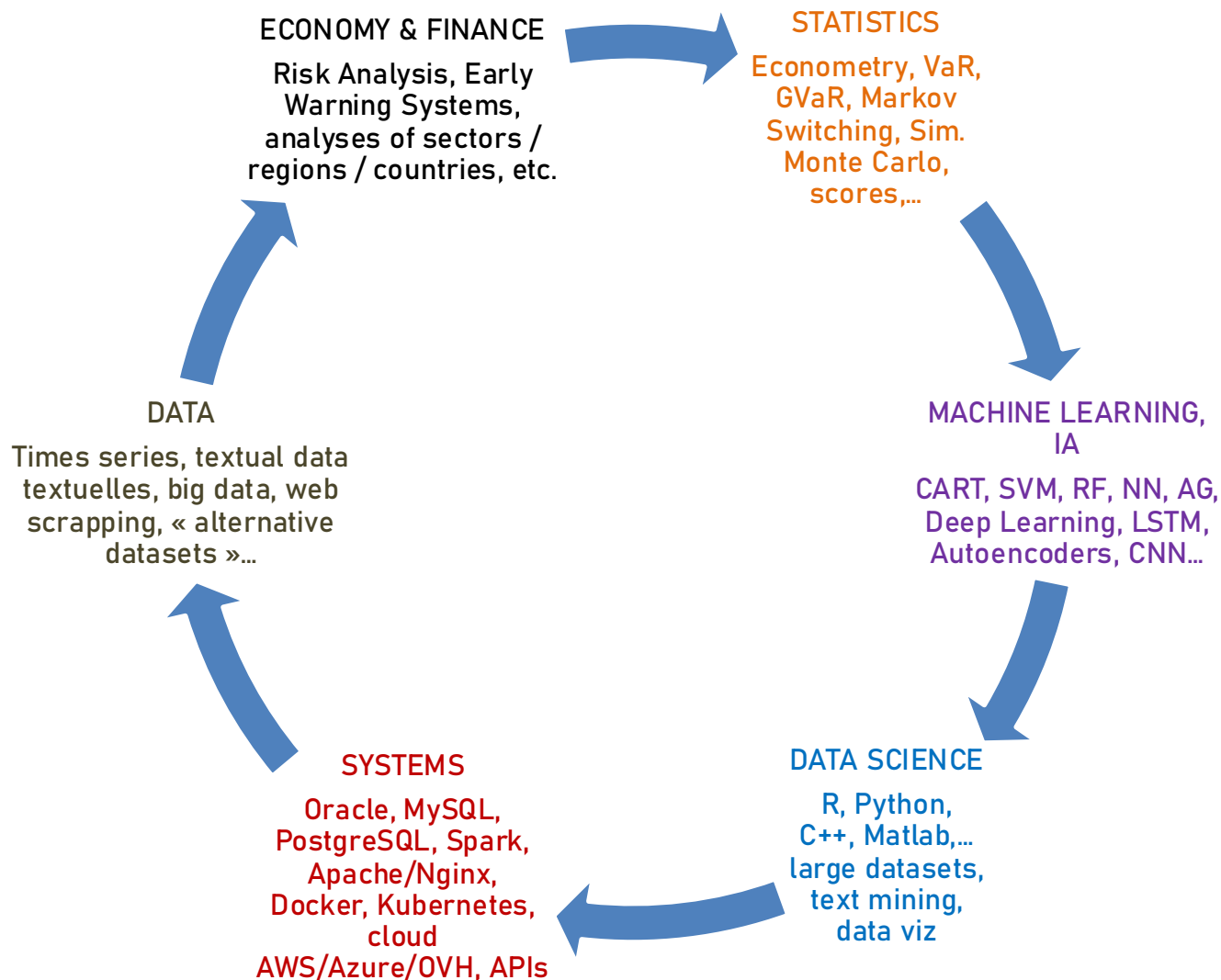
Big data and AI



Le syndrome de l'iceberg



These new techniques deeply transform economic analysis



Définition « historique » du big data

Ce n'est pas simplement le concept de données massives, mais :

- **Volumes**

Capacité à stocker des volumes de données gigantesques.

- **Vitesse**

Capacité à traiter des volumes de données importants rapidement.

- **Variées**

Capacité à traiter des données de différents types: textes, vidéos, images, données numériques, séries temporelles, etc.

Bases des données

- Les bases de données traditionnelles SGDB (MySQL, Postgresql) et la notion ACID (Atomicité, Cohérence, Isolation, Durabilité)
- Les bases de données objets ou orientées colonnes NoSQL
- Les bases en RAM type Redis
- Stockage objet (AWS S3)
- Les bases textuelles (type ElasticSearch/OpenSearch)
- Les bases de séries temporelles (InfluxDB, M3)
- Les clusters Hadoop, Hive et Kafka

The background of the slide is a dark blue, abstract network of interconnected white dots and lines, resembling a molecular structure or a data network. This pattern covers the top and bottom portions of the slide, framing a central white horizontal band.

GIT

Les bases de git

- **Installer une version de git:**
<https://git-scm.com/> ou Github Desktop
- **Créer un compte sur GitHub:**
<https://github.com/>
- **Créer un token (classic):**
<https://github.com/settings/tokens>
Generate new token (classic), on donne un nom, puis on coche « repo » et expiration (on évite les tokens éternels !)

Les bases de git

- **Je récupère un repo existant sur ma machine:**
git clone <https://github.com/langchain-ai/langchain>
- **J'initialise un répertoire existant:**
git init
git add .
- **Je valide mes modifications:**
git commit -m 'initial commit'
- **J'envoie mes premières modifications sur github:**
git remote add origin <https://github.com/sylbarth/test.git>
git branch -M main
git push -u origin main
- **Autres modifications:**
git commit -m 'new modification'
git push



Docker

Installation de Docker sous Windows

The screenshot shows the Docker Docs website. The top navigation bar is blue with the Docker logo and 'docker docs' text, followed by a search bar and links for Home, Guides, Manuals, Reference, Samples, and Contribute. Below this is a breadcrumb trail: Home / Manuals / Docker Desktop / Install Docker Desktop / Install on Windows. A left sidebar contains a list of navigation links, with 'Install on Windows' highlighted. The main content area has the title 'Install Docker Desktop on Windows' and an estimated reading time of 9 minutes. It includes a 'Docker Desktop terms' section, a 'Download Docker Desktop for Windows' section with a button labeled 'Docker Desktop for Windows', and a 'System requirements' section. The 'System requirements' section has two tabs: 'WSL 2 backend' (selected) and 'Hyper-V backend and Windows containers'. Under the 'WSL 2 backend' tab, there is a list of requirements for Windows 11 and Windows 10.

docker docs Search the docs Home Guides Manuals Reference Samples Contribute

Home / Manuals / Docker Desktop / Install Docker Desktop / Install on Windows

Understand permission requirements for Mac

Install on Windows

Understand permission requirements for Windows

Install on Apple silicon

Install on Linux

Installation per Linux distro

Quick Start Guide and sign in

Explore Docker Desktop

Change settings

Troubleshoot and diagnose

Additional resources

Dev Environments (Beta)

Extensions (Beta)

Extensions SDK (Beta)

Containerd Image Store (Beta)

FAQs

Give feedback

Release notes

Previous versions

Install Docker Desktop on Windows

Estimated reading time: 9 minutes

Docker Desktop terms

Commercial use of Docker Desktop in larger enterprises (more than 250 employees OR more than \$10 million USD in annual revenue) requires a paid subscription.

Welcome to Docker Desktop for Windows. This page contains information about Docker Desktop for Windows system requirements, download URL, instructions to install and update Docker Desktop for Windows.

Download Docker Desktop for Windows

[Docker Desktop for Windows](#)

For checksums, see [Release notes](#)

System requirements

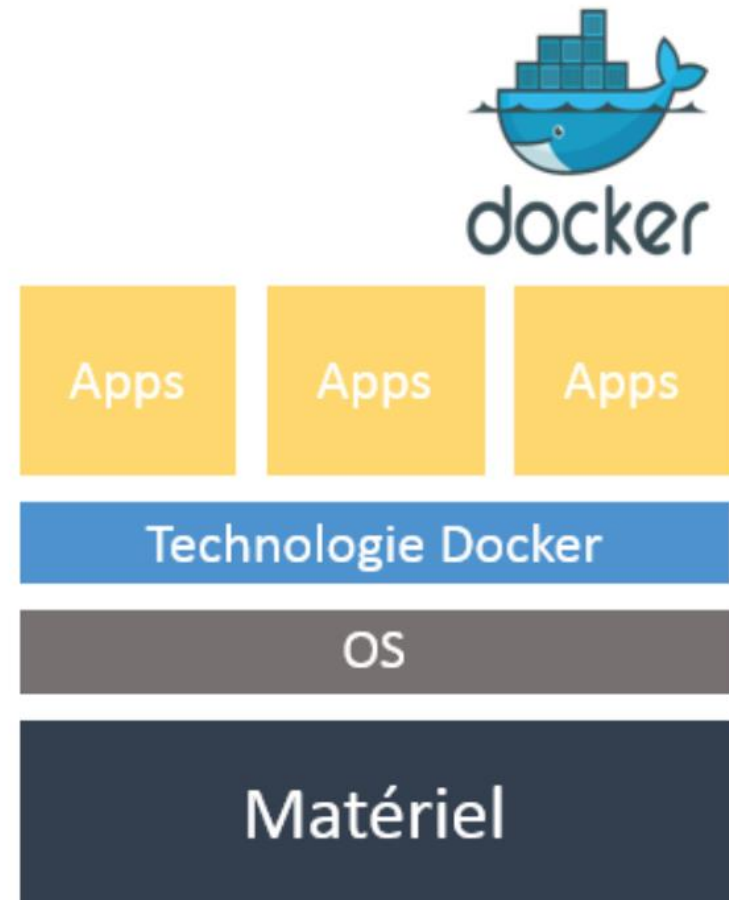
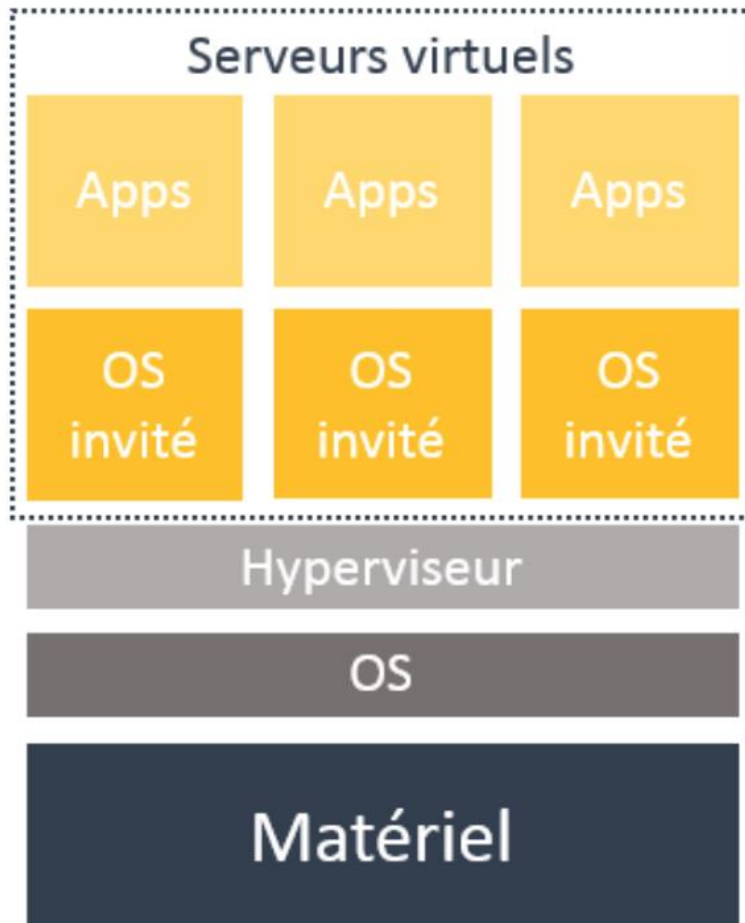
Your Windows machine must meet the following requirements to successfully install Docker Desktop.

WSL 2 backend Hyper-V backend and Windows containers

WSL 2 backend

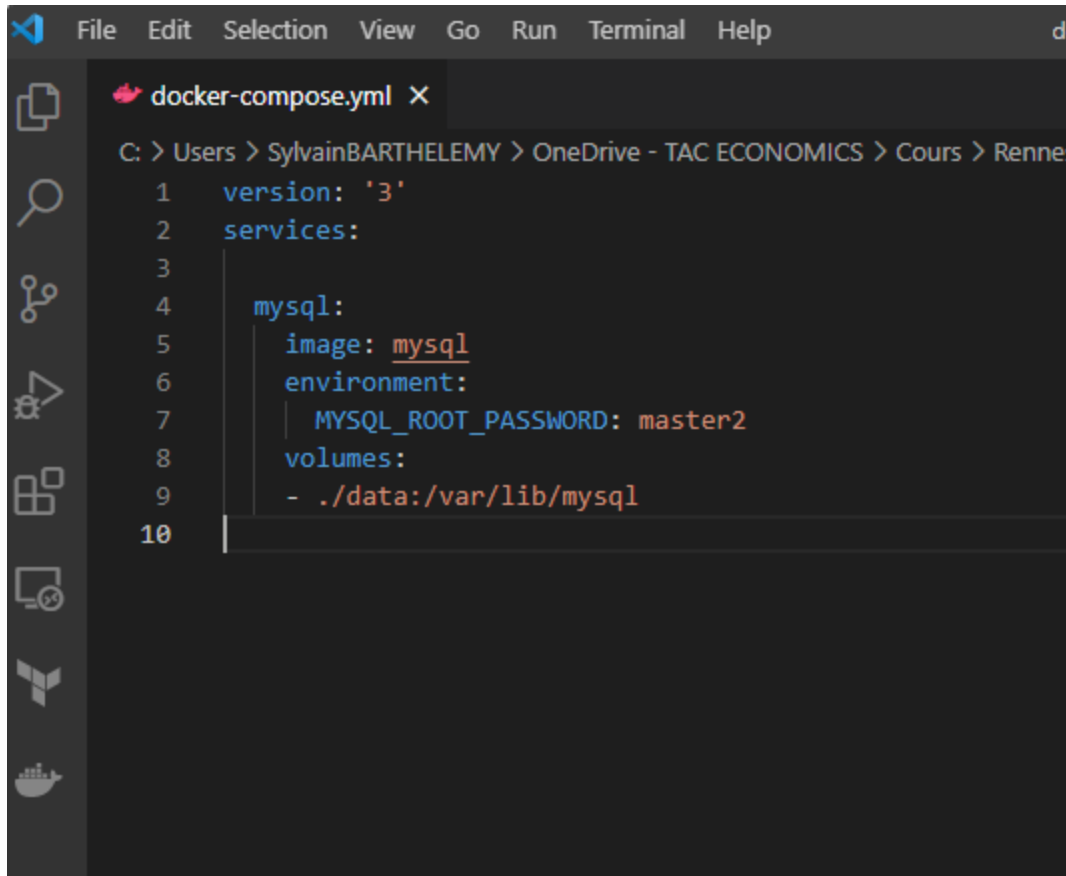
- Windows 11 64-bit: Home or Pro version 21H2 or higher, or Enterprise or Education version 21H2 or higher.
- Windows 10 64-bit: Home or Pro 21H1 (build 19043) or higher, or Enterprise or Education 20H2 (build 19042) or higher.
- Enable the WSL 2 feature on Windows. For detailed instructions, refer to the [Microsoft documentation](#).
- The following hardware prerequisites are required to successfully run WSL 2 on Windows 10 or Windows 11:

Docker ?



source: <https://www2.itroom.fr/docker-en-un-coup-d-oeil/>

Base de données MySQL



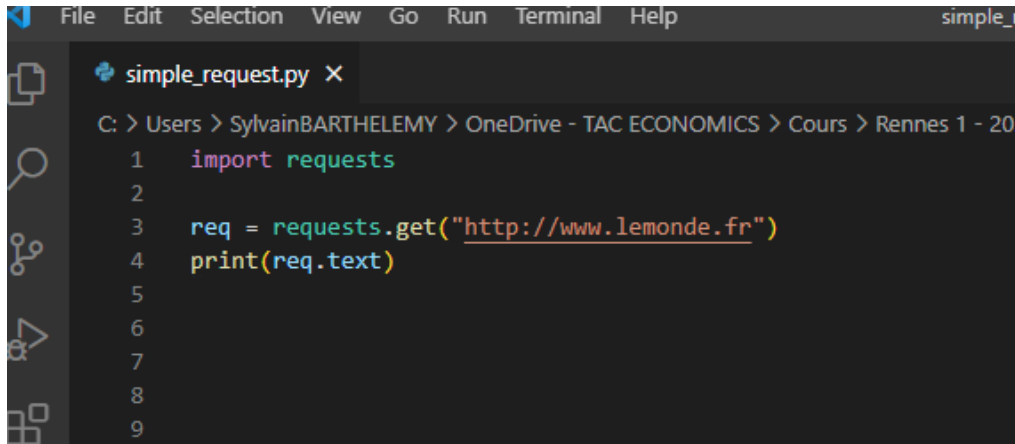
The image shows a code editor window with a dark theme. The title bar at the top reads 'File Edit Selection View Go Run Terminal Help'. The active file is 'docker-compose.yml'. The file path is 'C: > Users > SylvainBARTHELEMY > OneDrive - TAC ECONOMICS > Cours > Rennes'. The code is as follows:

```
1  version: '3'
2  services:
3
4      mysql:
5          image: mysql
6          environment:
7              MYSQL_ROOT_PASSWORD: master2
8          volumes:
9              - ./data:/var/lib/mysql
10
```

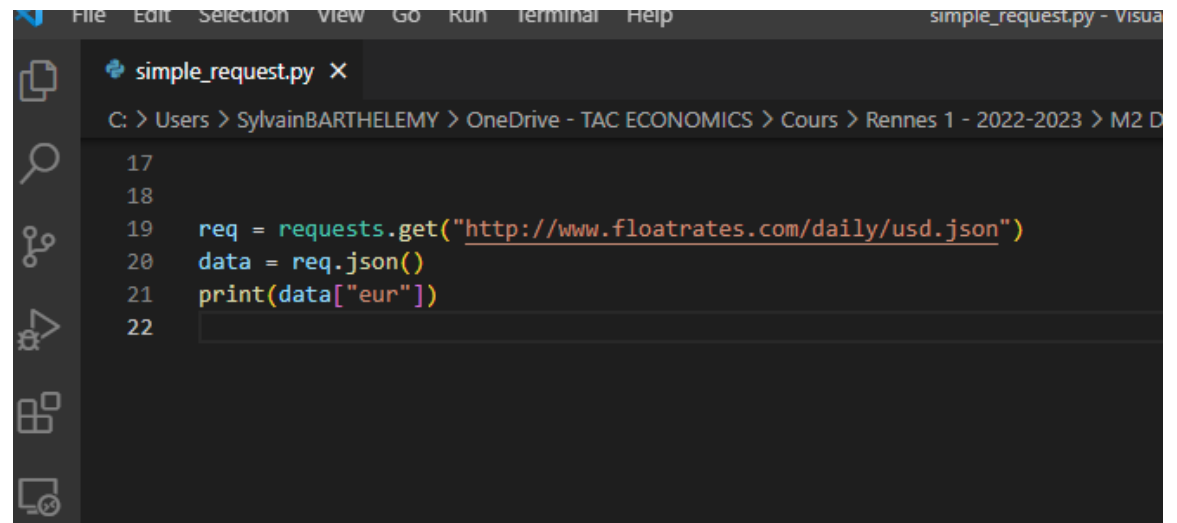



Scrapping, ML, IA

Introduction au Web Scraping RSS en Python



```
File Edit Selection View Go Run Terminal Help simple_r
simple_request.py X
C: > Users > SylvainBARTHELEMY > OneDrive - TAC ECONOMICS > Cours > Rennes 1 - 20
1 import requests
2
3 req = requests.get("http://www.lemonde.fr")
4 print(req.text)
5
6
7
8
9
```



```
File Edit Selection View Go Run Terminal Help simple_request.py - Visual
simple_request.py X
C: > Users > SylvainBARTHELEMY > OneDrive - TAC ECONOMICS > Cours > Rennes 1 - 2022-2023 > M2 D
17
18
19 req = requests.get("http://www.floatrates.com/daily/usd.json")
20 data = req.json()
21 print(data["eur"])
22
```

Introduction au Web Scraping

Lecture d'articles sur le New York Times

{T} Developers

[Home](#) [APIs](#) [Covid-19 Data](#) [Get Started](#) [Sign In](#)

The New York Times Developer Network

All the [APIs](#) Fit to Post

GET STARTED

✓ Get Started

Learn how to sign up for an API key.

📖 APIs

Learn about and try out NYT's APIs.

📖 FAQ

Get answers to frequently asked questions.

🖋 Branding

Read about NYT's branding & attribution.

📊 Covid-19 Data

Get info about NYT Covid-19 data.

Python, les essentiels

- **Pandas**
Manipulation en dataframes, comme sous R
- **Sklearn**
Algorithmes de machine learning
- **Nltk**
Travaux de traitement du langage naturel (NLP)
- **PyTorch & Tensorflow**
Deep learning
- **Lanchain**
Utilisation des grands modèles de langage



Déroulement du cours

Règles

- Groupes de 2 personnes max
- Préparation du travail pendant les cours, avec l'aide de l'enseignant
- Présentation des résultats devant toute la classe

Déroulement du travail

1. Revue de la littérature, recherches sur Internet
2. Cadrage du sujet
3. Définir un plan d'action et un partage des tâches
4. Création d'un premier dataset « test »
5. Mise en place d'un ou plusieurs outils pour récupérer les données et les traiter
6. Analyses quantitatives
7. Identification et présentation des résultats clés
8. Préparation d'une présentation PPT
9. Présentation des résultats devant la classe (30m avec questions)

Planning

Jour	Etudiants	Activité	Enseignants	Salles	Début	Fin	Durée (h)
Lundi 02/09/2024	M2 FD, M2 IEF FA, M2 IEF FI	Big Data	BARTHELEMY Sylvain	B01 - Salle info 01-07 (001)	13h30	17h00	03h30
Mercredi 04/09/2024	M2 FD, M2 IEF FA, M2 IEF FI	Big Data	BARTHELEMY Sylvain	B01 - Salle info 01-07 (001)	13h30	17h00	03h30
Mardi 24/09/2024	M2 FD, M2 IEF FA, M2 IEF FI	Big Data	BARTHELEMY Sylvain	B01 - Salle info 01-07 (001)	09h00	12h30	03h30
Mardi 24/09/2024	M2 FD, M2 IEF FA, M2 IEF FI	Big Data	BARTHELEMY Sylvain	B01 - Salle info 017 (017)	13h30	17h00	03h30
Mercredi 25/09/2024	M2 FD, M2 IEF FA, M2 IEF FI	Big Data	BARTHELEMY Sylvain	B01 - Salle info 01-07 (001)	09h00	12h30	03h30
Mardi 08/10/2024	M2 FD, M2 IEF FA, M2 IEF FI	Big Data	BARTHELEMY Sylvain	B01 - Salle info 01-07 (001)	09h00	12h30	03h30
Mercredi 09/10/2024	M2 FD, M2 IEF FA, M2 IEF FI	Big Data	BARTHELEMY Sylvain	B01 - Salle info 01-07 (001)	09h00	12h30	03h30
Jeudi 10/10/2024	M2 FD, M2 IEF FA, M2 IEF FI	Big Data	BARTHELEMY Sylvain	B01 - Salle info 01-07 (001)	13h30	17h00	03h30

Sujets

1. Usage de la presse et de la conjoncture pour analyser, comprendre et prévoir les évolutions des marchés financiers
2. Twitter et les marchés financiers
3. Mieux comprendre le rôle des influenceurs twitter sur les cours boursiers
4. Analyse des données INSEE des communes et visualisation des disparités et spécificités géographiques
5. Machine learning, IA et stratégies de trading automatisée
6. Google trends et indicateurs économiques et financiers (conjoncture, emploi, immobilier, ...)
7. Analyse long terme et intra-day du marché des cryptos
8. Stratégie de référencement (SEO) et text mining/big data
9. Lecture de la presse et analyse des entités nommées en temps réel
10. Autres sujets au choix, à la discrétion des élèves

Me contacter

Sylvain BARTHELEMY

sylvain.barthelemy@gwenlake.com

www.gwenlake.com