ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI
SCIENZE STATISTICHE "PAOLO FORTUNATI"

Introduction to Unstructured Data Analysis

With Applications in Banking and Finance

Giovanni Della Lunga

Halloween Conference in Quantitative Finance

Bologna - October 26-28, 2021

Subsection 1

Why?

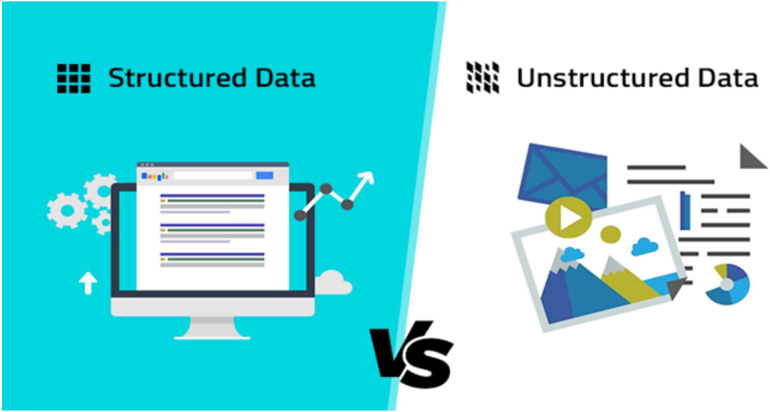
The purpose and the rationale for the subject covered in the seminar

Why?

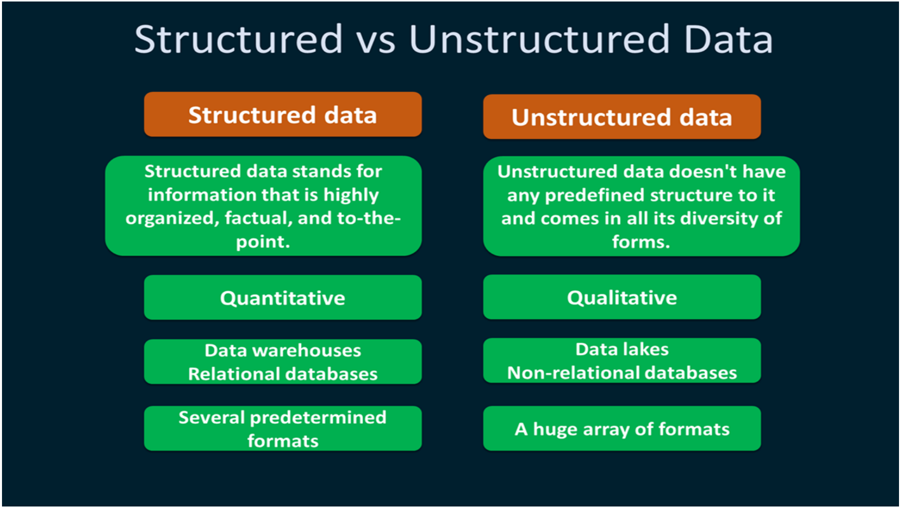
- When we think of financial data, our thinking almost always ends up displaying infinite historical series of numbers (prices, interest rates, volatility)
- Financial data usually come in as structured data.

The importance of Unstructured Data

On the other hand, unstructured data, such as call transcripts, emails text, transactional data are largely an area not yet fully exploited due to accessibility and processing challenges.



The importance of Unstructured Data



Structured vs Unstructured Data

Structured data	Unstructured data
Structured data stands for information that is highly organized, factual, and to-the-point.	Unstructured data doesn't have any predefined structure to it and comes in all its diversity of forms.
Quantitative	Qualitative
Data warehouses Relational databases	Data lakes Non-relational databases
Several predetermined formats	A huge array of formats

Subsection 2

What?

The key content, principles and topics to be learned in these lessons

Machine Learning

- Machine learning is an artificial intelligence (AI) technology which provides systems with the ability to automatically learn from experience without the need for explicit programming, and can help solve complex problems.
- It is seen as a part of artificial intelligence.
- Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.
- Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Natural Language Processing

- Although textual data is abundantly available, the entanglement of natural language makes it particularly difficult to extract useful information from them.
- Natural Language Processing (NLP), a field of Artificial Intelligence (AI), analyzes, understands, and derives meaning from unstructured data.
- NLP focuses on the interaction between data science and human language, and is scaling to lots of industries.
- Today NLP is booming thanks to the huge improvements in the access to data and the increase in computational power, which are allowing practitioners to achieve meaningful results in areas like healthcare, media, finance and human resources, among others. NLP is used for Named Entity Recognition (NER) and Sentiment Analysis as well as Parts-of-Speech tagging, and more.

Information Extraction

- In order to act on unstructured data, ML models have to perform one of the crucial processes called Information Extraction(IE).
- Information Extraction is the process of retrieving key information intertwined within the unstructured data. In other words, extracting structured data from the unstructured data.
- Unfortunately, when it comes to mining these sources for usable data, it's not quite so quick and easy. Sure, you can search documents for specific text, but what does that really tell you? Beyond word or phrase frequency, not much else.
- For these reasons you have to resort to sophisticated techniques of Text Analysis which employ a variety of methodologies to process the text, one of the most important of these being Natural Language Processing (NLP).

Subsection 3

How?

the learning tools we are going to use

Anaconda

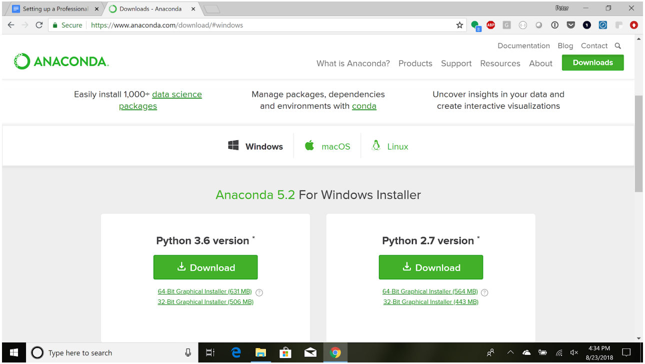
- To set up your python environment, you'll first need to have a python on your machine.
- There are various python distributions available and we have chosen one that works very well for data science.
- Anaconda comes with its own Python distribution which will be installed along with it.

Anaconda

- Data Science often requires you to work with a lot of scientific packages like scipy and numpy, data manipulation packages like pandas and IDEs and interactive Jupyter Notebook.
- Now, you don't need to worry about any python package most of them come pre-installed and if you want to install a new package, you can do that simply by using conda or pip.

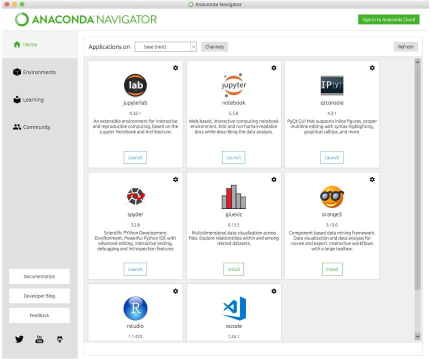
Installing Python and Anaconda

- To download an Anaconda distribution, you can use the official download page: **https://www.anaconda.com/download/**
- Here, you can select your platform and then choose the installer. For this, you can choose which version you want and whether 32-bit or 64-bit.


Setting up a Professional
Downloads - Anaconda
Secure
https://www.anaconda.com/download/#windows
Documentation
Blog
Contact
ANACONDA.
What is Anaconda?
Products
Support
Resources
About
Downloads
Easily install 1,000+ data science packages
Manage packages, dependencies and environments with conda
Uncover insights in your data and create interactive visualizations
Windows
macOS
Linux
Anaconda 5.2 For Windows Installer
Python 3.6 version *
Download
64-Bit Graphical Installer (631 MB)
32-Bit Graphical Installer (506 MB)
Python 2.7 version *
Download
64-Bit Graphical Installer (564 MB)
32-Bit Graphical Installer (443 MB)
Type here to search
4:34 PM
8/23/2018

Testing Your Installation

To test your installation, on Windows, click on Start and then Anaconda Navigator in the program list (or search for Anaconda in the search bar and select Anaconda Navigator). On a Mac, open up the finder, and in the Applications folder, double click on Anaconda-Navigator.



jupyterlab
0.32.1
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

notebook
5.5.0
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

qtconsole
4.3.1
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

spyder
3.2.8
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

glueviz
0.13.3
Multidimensional data visualization across files. Explore relationships within and among related datasets.

orange3
3.13.0
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

rstudio
1.1.423

vscode
1.26.1

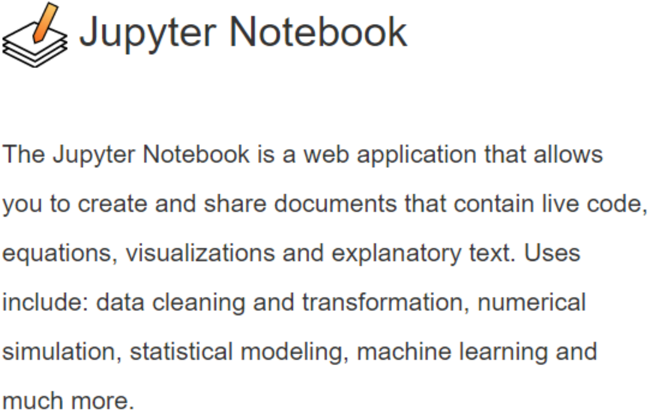
Package Managers

- Anaconda will give you two package managers- pip and conda.
- When some packages aren't available with conda, you can use pip to install them.
- Note that using pip to install packages also available to conda may cause an installation error.

Teaching tools: Jupyter Notebook

- The Python world developed the IPython notebook system.
- Notebooks allow you to write text, but you insert code blocks as ”cells” into the notebook.
- A notebook is interactive, so you can execute the code in the cell directly!
- Recently the Notebook idea took a much enhanced vision and scope, to explicitly allow languages other than Python to run inside the cells.
- Thus the Jupyter Notebook was born, a project initially aimed at Julia, Python and R (Ju-Pyt-e-R). But in reality many other languages are supported in Jupyter.

Teaching tools: Jupyter Notebook

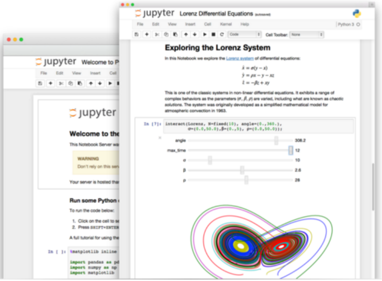


Jupyter Notebook

The Jupyter Notebook is a web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

Teaching tools: Jupyter Notebook

- Jupyter was designed to enable sharing of notebooks with other people.
- The idea is that you can write some code, mix some text with the code, and publish this as a notebook.
- In the notebook they can see the code as well as the actual results of running the code.


jupyter Welcome to P
File Edit View Insert Cell
jupyter
Welcome to the
This Notebook Server wa
WARNING
Don't rely on this serv
Your server is hosted tha
Run some Python
To run the code below:
1. Click on the cell to se
2. Press SHIFT+ENTER
A full tutorial for using the
jupyter Lorenz Differential Equations (autosaved)
File Edit View Insert Cell Kernel Help
Python 3
Code
Cell Toolbar: None

Exploring the Lorenz System

In this Notebook we explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters (σ, β, ρ) are varied, including what are known as chaotic solutions. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

```
In [7]: interact(Lorenz, N=fixed(10), angle=(0.,360.),
                 σ=(0.0,50.0),β=(0.,5), ρ=(0.0,50.0));
```

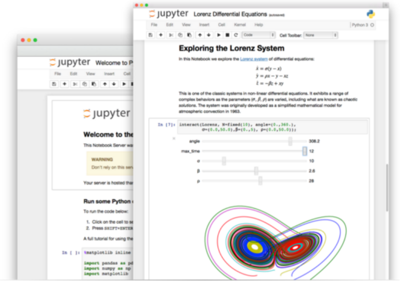
angle 308.2
max_time 12
σ 10
β 2.6
ρ 28

```
In [ ]: %matplotlib inline

        import pandas as pd
        import numpy as np
        import matplotlib
```

Teaching tools: Jupyter Notebook

- This is a nice way of sharing little experimental snippets, but also to publish more detailed reports with explanations and full code sets.
- Of course, a variety of web services allows you to post just code snippets (e.g. gist).
- What makes Jupyter different is that the service will actually render the code output.



jupyter

Welcome to the

This Notebook Server wa

WARNING

Don't rely on this serv

Your server is hosted tha

Run some Python

To run the code below:

1. Click on the cell to se
2. Press SHIFT+ENTER

A full tutorial for using the

```
In [ ]: %matplotlib inline

        import pandas as pd
        import numpy as np
        import matplotlib
```

Exploring the Lorenz System

In this Notebook we explore the Lorenz system of differential equations:

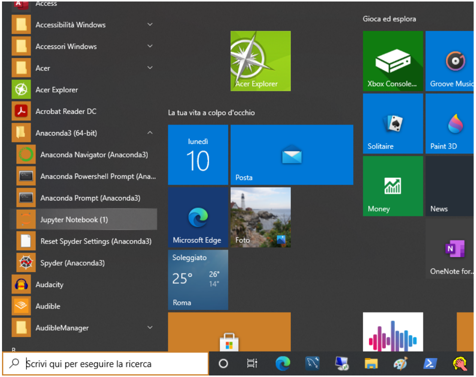
$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters (σ, β, ρ) are varied, including what are known as chaotic solutions. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

```
In [7]: interact(Lorenz, N=fixed(10), angle=(0.,360.),
                 σ=(0.0,50.0),β=(0.,5), ρ=(0.0,50.0));
```

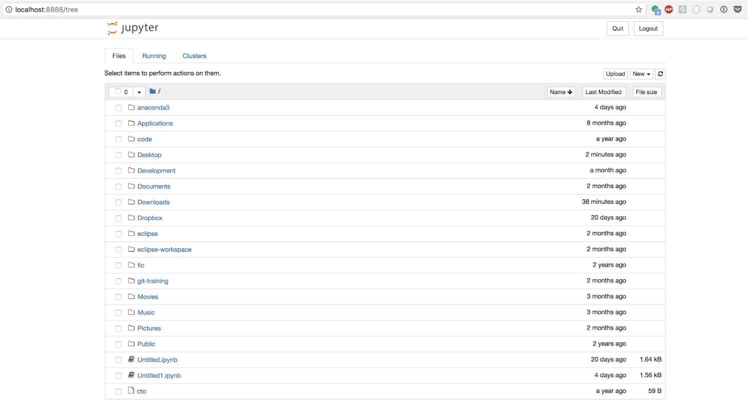
Teaching tools: Jupyter Notebook

- As we saw earlier, the Jupyter Notebook ships with Anaconda. To run it, you can get in your virtual environment and type the following command: **jupyter notebook**;
- Or directly from the Windows Menu...



Teaching tools: Jupyter Notebook

- You can find this at **http://localhost:8888/tree**
- Now to run Python here, you can create a new file.



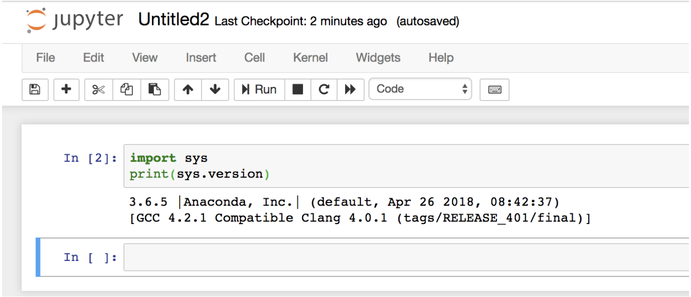
localhost:8888/tree

Select items to perform actions on them.

□ 0 ▾ /	Last Modified	File size
□ anaconda3	4 days ago	
□ Applications	8 months ago	
□ code	a year ago	
□ Desktop	2 minutes ago	
□ Development	a month ago	
□ Documents	2 months ago	
□ Downloads	38 minutes ago	
□ Dropbox	20 days ago	
□ eclipse	2 months ago	
□ eclipse-workspace	2 months ago	
□ fic	2 years ago	
□ git-training	2 months ago	
□ Movies	3 months ago	
□ Music	3 months ago	
□ Pictures	2 months ago	
□ Public	2 years ago	
□ Untitled.ipynb	20 days ago	1.64 kB
□ Untitled1.ipynb	4 days ago	1.56 kB
□ cto	a year ago	59 B

Teaching tools: Jupyter Notebook

To make sure it's working, click in the cell and type the following:


jupyter Untitled2 Last Checkpoint: 2 minutes ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code

```
In [2]: import sys
        print(sys.version)

        3.6.5 |Anaconda, Inc.| (default, Apr 26 2018, 08:42:37)
        [GCC 4.2.1 Compatible Clang 4.0.1 (tags/RELEASE_401/final)]

In [ ]:
```

Teaching tools: Google Colab

- Colaboratory, or "Colab" for short, is a product from Google Research.
- Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

https://colab.research.google.com/notebooks/intro.ipynb?hl=en

Teaching tools: Google Colab

- More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing free access to computing resources including GPUs.
- Colab notebooks are stored in *Google Drive*, or can be loaded from *GitHub*. Colab notebooks can be shared just as you would with Google Docs or Sheets.

https://colab.research.google.com/notebooks/intro.ipynb?hl=en

GitHub Repository for this Course

- GitHub is a provider of Internet hosting for software development and version control using Git. It offers the distributed version control and source code management (SCM) functionality of Git, plus its own features.
- It provides access control and several collaboration features such as bug tracking, feature requests, task management, continuous integration for every project.
- You can find all the teaching materials (notebook, slides, code, etc...) at this address **https://github.com/polyhedron-gdl** in the repository **applied-computational-finance/2021**.