ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

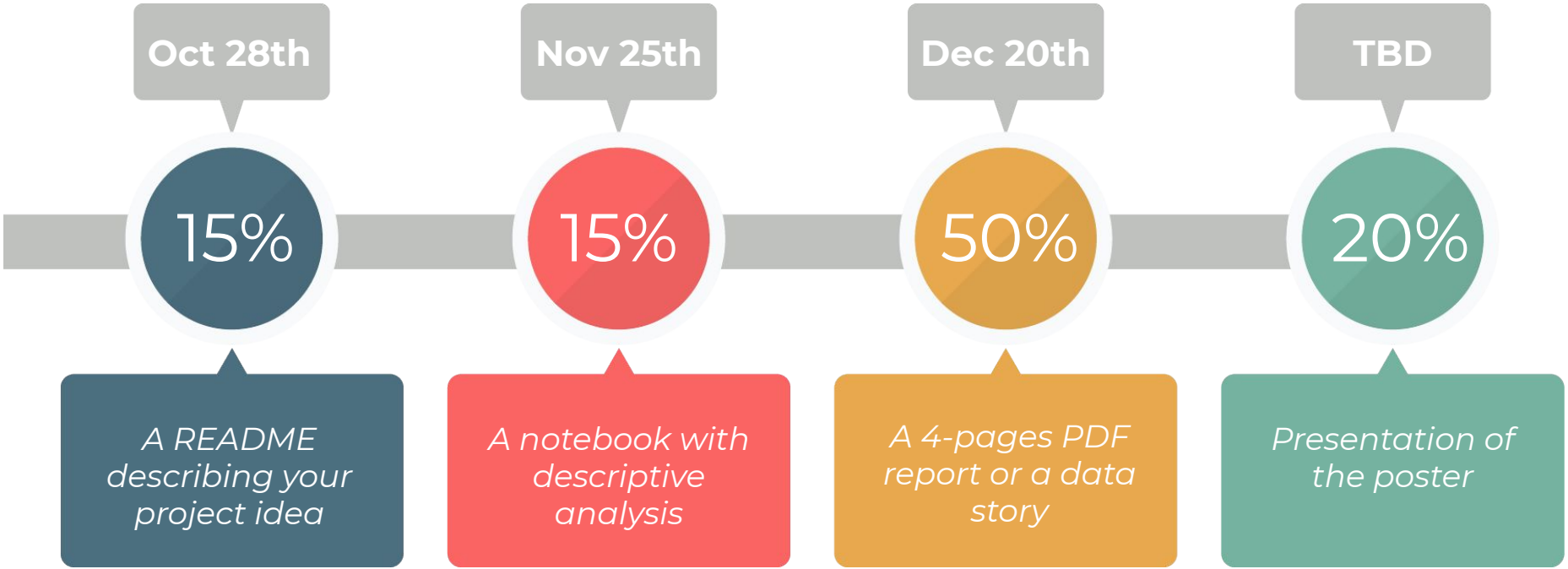
Project

Weight: 33%

“Data science for social good”

Think about how you could improve society through data analysis!

Timeline



All the dates have deadline at 23.59

Milestone 1

Backtracking Innovation

From Snapchat's Facial Recognition to the Inception of Semiconductors

Abstract

The US Patent & Trademark Office, short USPTO, offers a full-fledged directed graph of US patent citations since 1975. Many at the time obscure patents paved the way for groundbreaking new technologies in the future, as for example Google's Page Rank (#6285999) or Apple's touch screen devices (#7479949). In our project we endeavor to visualize this massive graph with more than 89 Million edges in an appealing and concise way in order to render the narration of more intricate stories possible and offer potential users a slick and interactive interface for navigating the patents graph. Inspirations on how to beautifully draw a graph network are taken from Kirell Benzi. Those stories may range from analysing node connectivities to extract influences of particular inventors to finding the shortest walks between patents. Through this we hope to gain deeper insight into the metastructure of inventions throughout the digital revolution and how they shape the future as we know it today.

Research questions

- How to apply common graph algorithms to a massive data set?
- How to efficiently travers and query a graph of that scale?

Milestone 1 - Template

Clone the repository and use the README file as template

https://github.com/epfl-ada/Projects

Milestone 1 - Repository

Create a public repository in Github and register your project:

https://go.epfl.ch/ada19_register

Milestone 1 - Examples

Backtracking Innovation
https://github.com/qantik/prayingmantissa/tree/3a8754f847a955018c5dbbaeb996d79e19200f87/project

Social topics throughout music: sentiment analysis
https://github.com/hbjornoy/DataAnalysis/tree/78a59863a500ade2db71fcc59d81d8d599efadc6/project

Evolution in music through the years
https://github.com/maxpr/ADA/tree/f399374483a21551cb39aadc5343c84cef2ca33f/project

Mental Health in Switzerland
https://github.com/othmanbck/ada/tree/987c91aa8c88d1640145586dd97421929f9b69b6/project

Milestone 2

A notebook with data collection and descriptive analysis, properly commented, and the notebook ends with a more structured and informed plan for what comes next

Report - Examples

https://go.epfl.ch/ada17_projects

https://go.epfl.ch/ada18_projects

Final presentation

Poster

+

3 minutes presentation

Datasets

Provided dataset are available here:

https://go.epfl.ch/ada19_datasets

You can propose a different dataset, but the TAs support will be limited

ADA Cluster facts

7 Nodes

266 CPUs

1.24TB of Memory

133.7 TB Disk

Master node

ssh gaspar_account@iccluster040.iccluster.epfl.ch

Useful information



Connect to the cluster	ssh	Putty
Download/upload	scp	WinScp

Connect to the master node:

ssh gaspar_account@iccluster040.iccluster.epfl.ch

Log of the jobs:
http://iccluster040.iccluster.epfl.ch:8088/cluster

Datasets

The datasets are available in /datasets

hadoop fs -ls /datasets

hadoop fs -cat /datasets/<file_name> | less