BIGPROD

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822

Agenda for today

13.00 – 13.15 Welcome and introduction of participants

13.15 – 13.30 Introduction of the BIGPROD project

13.30 – 14.00 Presentation of the BIGPROD platform operation

14.00 – 14.30 Discussion

WHAT IS BIGPROD

BIGPROD PROJECT PARTNERS

Quantitative Science and Technology Studies team, Foresight-driven Business Strategies, VTT Technical Research Centre of Finland

Public Policy and Management Institute

Competence Center Innovation and Knowledge Economy, Fraunhofer ISI

Economics of Technology and Innovations, Faculty of Technology, Policy and Management, Delft University of Technology

Economics of Knowledge and Innovation team, Maastricht University

School of Government & Public Policy, Faculty of Humanities & Social Science, University of Strathclyde

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822

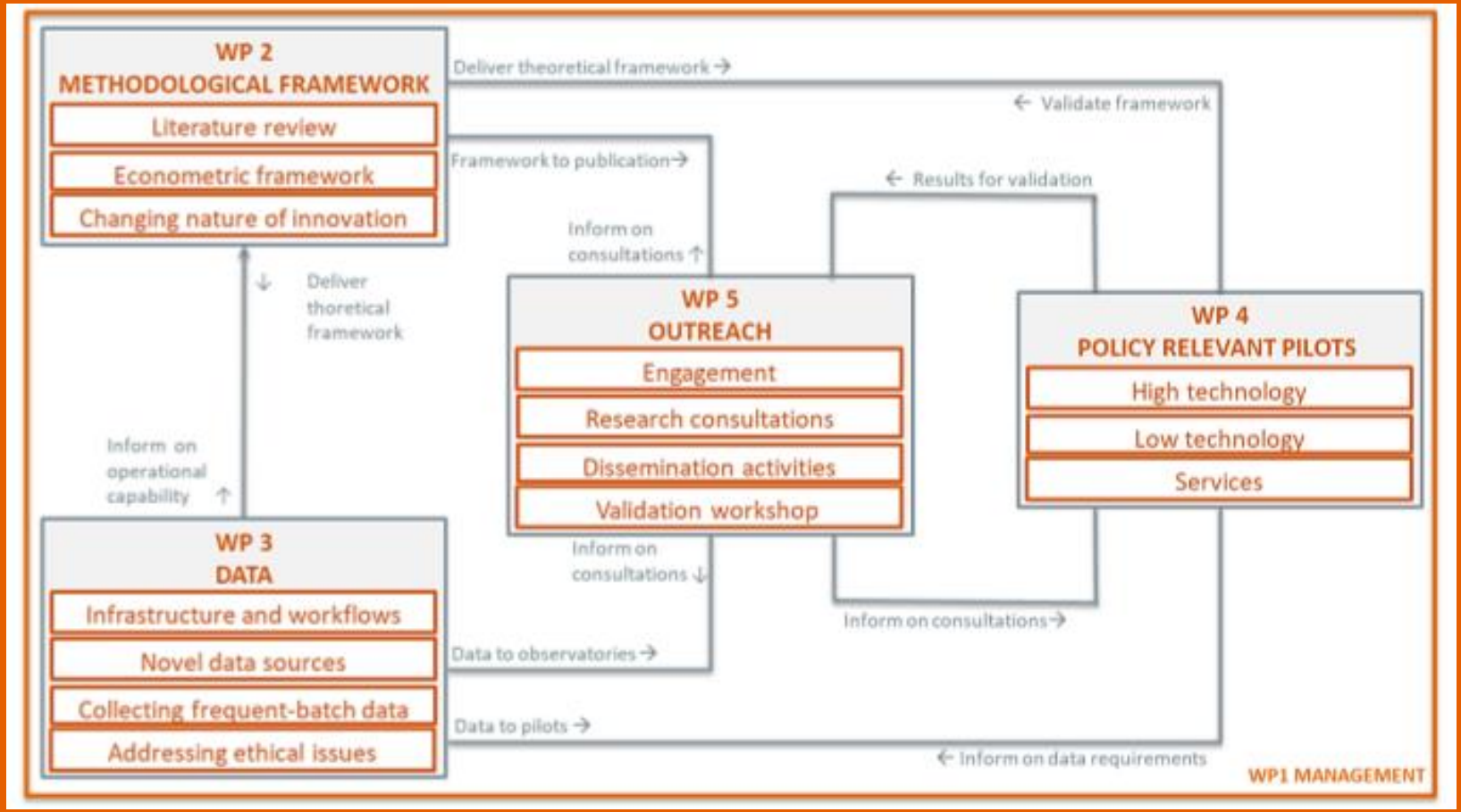
Addressing the productivity paradox

The objective of the project is

1. to extend existing econometric approaches on productivity with a theoretically sound “Big data” measures that can be operationalized and validated through pilots.
2. To have deep stakeholder consultation mitigating the skills gap, creating transparency, enabling stakeholder influence in sources and tools and enabling policy makers being informed on tools and pilots.

Objectives

1. Management and coordination for utilizing “Big data” for innovation and productivity assessment
2. Creating an extended econometric framework for the evaluation of the productivity-innovation link based on “Big data”
3. Building a large-scale data platform and framework which will yield frequent batch data on company performance and innovation activities
4. Create policy-relevant pilots that measure the impact of proposed changes, while enabling policymakers being informed on tools developed and piloted.
5. Utilizing the most effective tools available to effect stakeholder engagement and co-creation, while simultaneously ensuring the dissemination of the knowledge gained in this process to the wider public.


WP 2
METHODOLOGICAL FRAMEWORK
Literature review
Econometric framework
Changing nature of innovation
Deliver theoretical framework →
← Validate framework
Framework to publication →
← Results for validation
Inform on consultations ↑
Deliver thoretical framework
WP 5
OUTREACH
Engagement
Research consultations
Dissemination activities
Validation workshop
WP 4
POLICY RELEVANT PILOTS
High technology
Low technology
Services
Inform on operational capability ↑
Inform on consultations ↓
Inform on consultations →
WP 3
DATA
Infrastructure and workflows
Novel data sources
Collecting frequent-batch data
Addressing ethical issues
Data to observatories →
Data to pilots →
← Inform on data requirements
WP1 MANAGEMENT

Outcomes

The consortium will

1. Provide a theoretical framework and methodology with a detailed description of the indicators, parameters and weights and how they interact together.
2. Methodologies will be made available for free as it will be used for dissemination to teach how the system works and to explain the underlying philosophy. Python programming language notebooks with the basic calculations will also be provided.
3. Implementation of a methodology to gather, store and make an accessible novel variable for public policy productivity analysis.

WHY IS BIGPROD IMPORTANT

EUROPEAN
COMMISSION

Brussels, 26.2.2020
COM(2020) 150 final

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN CENTRAL BANK AND THE EUROGROUP

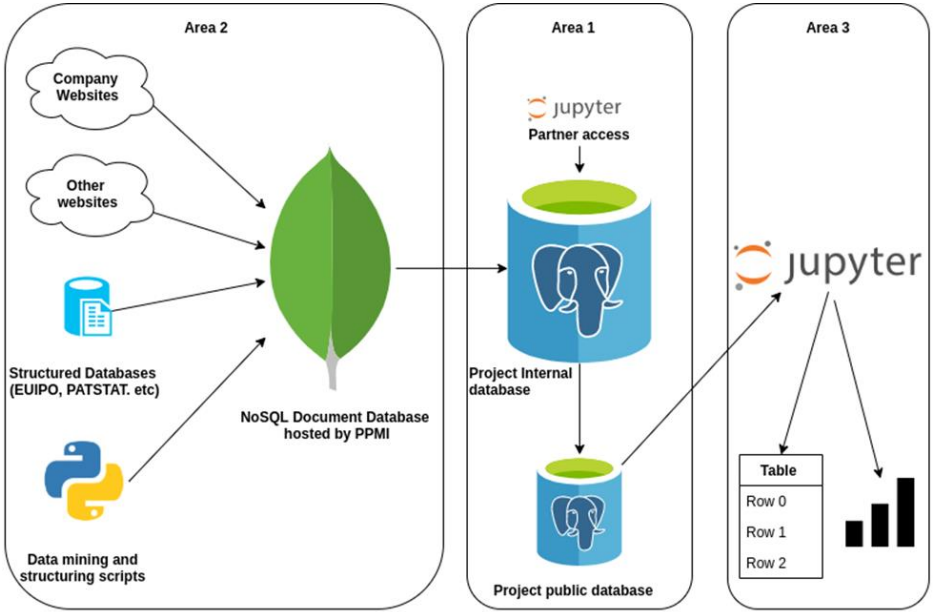
2020 European Semester: Assessment of progress on structural reforms, prevention and correction of macroeconomic imbalances, and results of in-depth reviews under Regulation (EU) No 1176(2011)

{SWD(2020) 500-527 final}

"Productivity growth remains a challenge, even more so in the light of demographic change…

…There are multiple causes for this weak performance…

…Policies to foster productivity need to be tailored to national circumstances…"



PLATFORM OPERATION

BIGPROD data platform roles

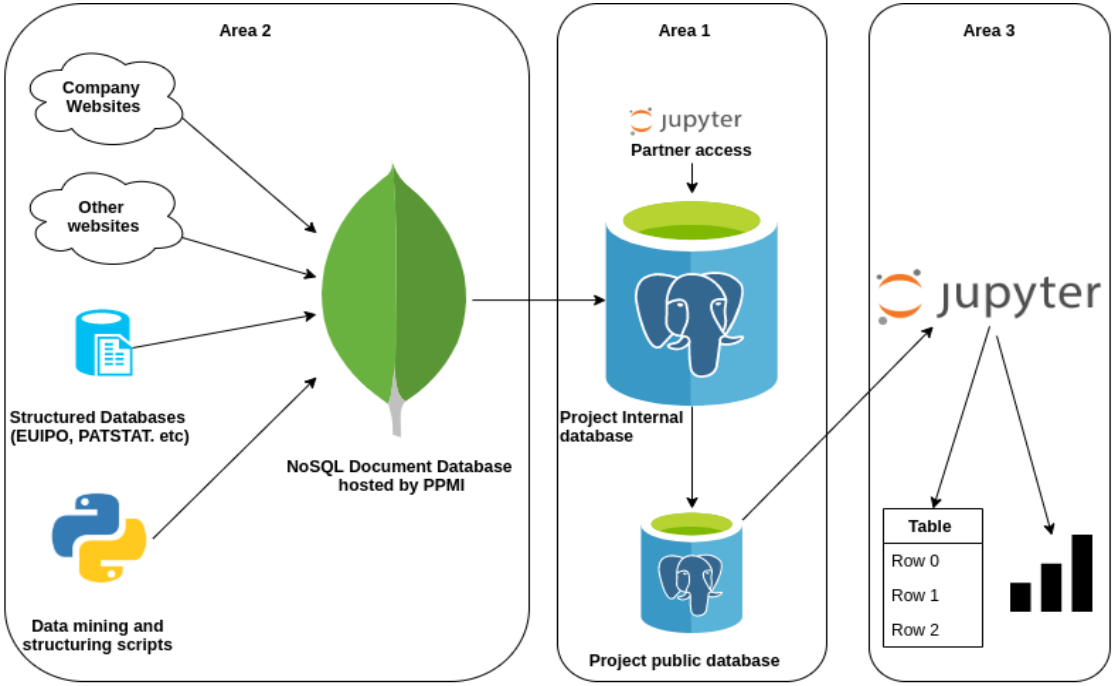
- It will store the project data assembled from various sources;
- It will facilitate the data exchange between the consortium partners;
- It will serve and expose a selected sub-section of the data to the end-users;
- It will ensure data protection, security and privacy.

Three autonomous areas

- Platform facilitates different levels of access and performs diverse functions, such as
  - data manipulation,
  - data visualisation and
  - story-telling.
- Due to this the platform has three autonomous areas:
  - Area 1: Cloud hosted SQL database;
  - Area 2: On premises hosted NoSQL database operating in PPMI;
  - Area 3: Jupyter Notebook Server

Main data collected

- Company descriptive data;
- Indicators calculated from the data scraped from company websites;
- Indicators calculated from review and other websites;
- Indicators calculated by matching company records with other databases (PATSTAT and EUIPO);
- Results from the econometric modelling;
- Other indicators.



Raw data from the company websites, structured company data databases (EUIPO, PATSTAT), other websites (review websites, etc.) is collected into a NoSQL document database.

Data is cleaned and structured. Several indicators are calculated from raw data. Data is aggregated to company level.

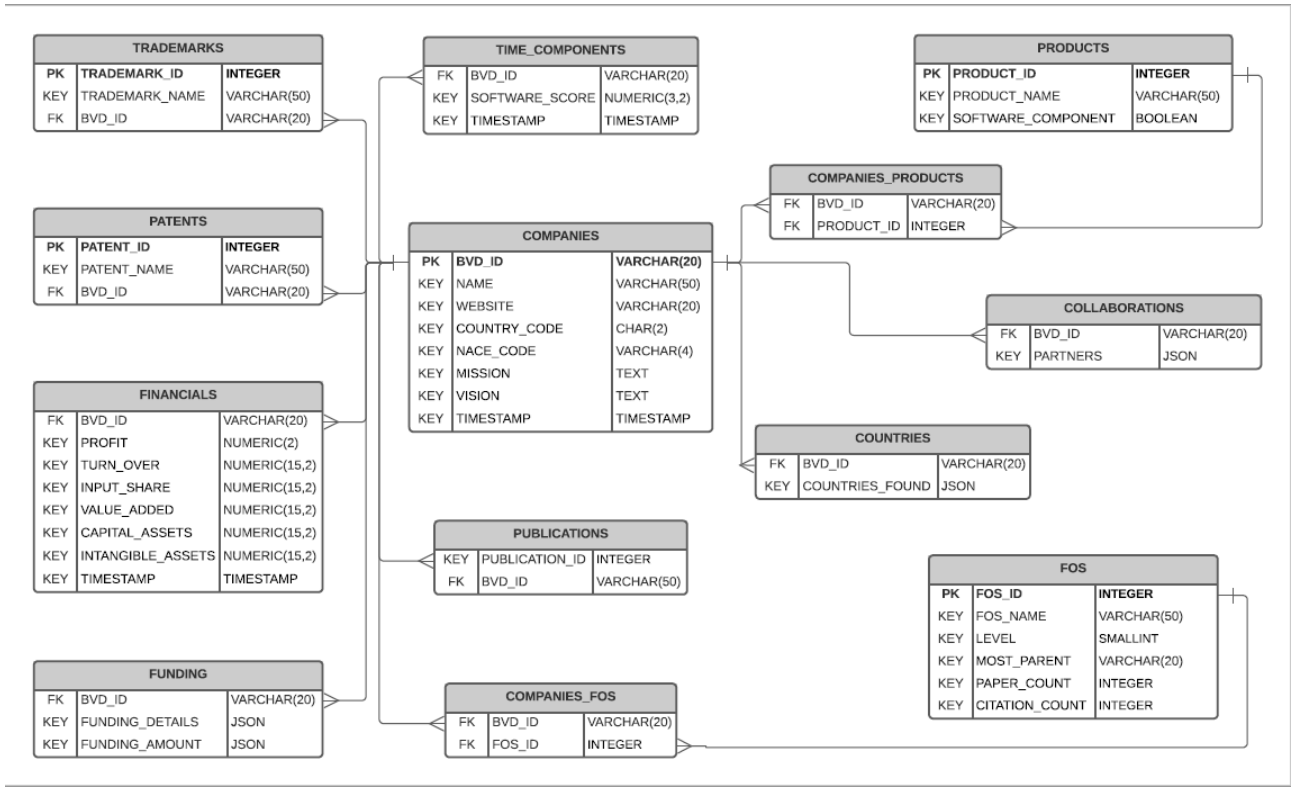
Company level data is published on the main project database. Each partner has access to the database to pull required data (for analysis or indicator construction).

A subset is data which can be exposed to the public (anonymised, aggregated data for specified indicators) is sent to a smaller DB, which can be accessed from outside.

Project Public data can be accessed through Jupyter notebook server. Jupyter server has access to the public data DB and contains sample tables and visualizations which can be tweaked by the user

Area 1: Cloud-hosted SQL Database

- Merges the following
  - Orbis” database;
  - Indicators derived from unstructured company website/ review website data;
  - Indicators derived from semi-structured data sources (PATSTAT, EUIPO);
  - Indicators derived from the econometric modelling;


TRADEMARKS
PK TRADEMARK_ID INTEGER
KEY TRADEMARK_NAME VARCHAR(50)
FK BVD_ID VARCHAR(20)
TIME_COMPONENTS
FK BVD_ID VARCHAR(20)
KEY SOFTWARE_SCORE NUMERIC(3,2)
KEY TIMESTAMP TIMESTAMP
PRODUCTS
PK PRODUCT_ID INTEGER
KEY PRODUCT_NAME VARCHAR(50)
KEY SOFTWARE_COMPONENT BOOLEAN
COMPANIES_PRODUCTS
FK BVD_ID VARCHAR(20)
FK PRODUCT_ID INTEGER
PATENTS
PK PATENT_ID INTEGER
KEY PATENT_NAME VARCHAR(50)
FK BVD_ID VARCHAR(20)
COMPANIES
PK BVD_ID VARCHAR(20)
KEY NAME VARCHAR(50)
KEY WEBSITE VARCHAR(20)
KEY COUNTRY_CODE CHAR(2)
KEY NACE_CODE VARCHAR(4)
KEY MISSION TEXT
KEY VISION TEXT
KEY TIMESTAMP TIMESTAMP
COLLABORATIONS
FK BVD_ID VARCHAR(20)
KEY PARTNERS JSON
FINANCIALS
FK BVD_ID VARCHAR(20)
KEY PROFIT NUMERIC(2)
KEY TURN_OVER NUMERIC(15,2)
KEY INPUT_SHARE NUMERIC(15,2)
KEY VALUE_ADDED NUMERIC(15,2)
KEY CAPITAL_ASSETS NUMERIC(15,2)
KEY INTANGIBLE_ASSETS NUMERIC(15,2)
KEY TIMESTAMP TIMESTAMP
COUNTRIES
FK BVD_ID VARCHAR(20)
KEY COUNTRIES_FOUND JSON
PUBLICATIONS
KEY PUBLICATION_ID INTEGER
FK BVD_ID VARCHAR(50)
FOS
PK FOS_ID INTEGER
KEY FOS_NAME VARCHAR(50)
KEY LEVEL SMALLINT
KEY MOST_PARENT VARCHAR(20)
KEY PAPER_COUNT INTEGER
KEY CITATION_COUNT INTEGER
FUNDING
FK BVD_ID VARCHAR(20)
KEY FUNDING_DETAILS JSON
KEY FUNDING_AMOUNT JSON
COMPANIES_FOS
FK BVD_ID VARCHAR(20)
FK FOS_ID INTEGER

Area 2: No-SQL database hosted by PPMI

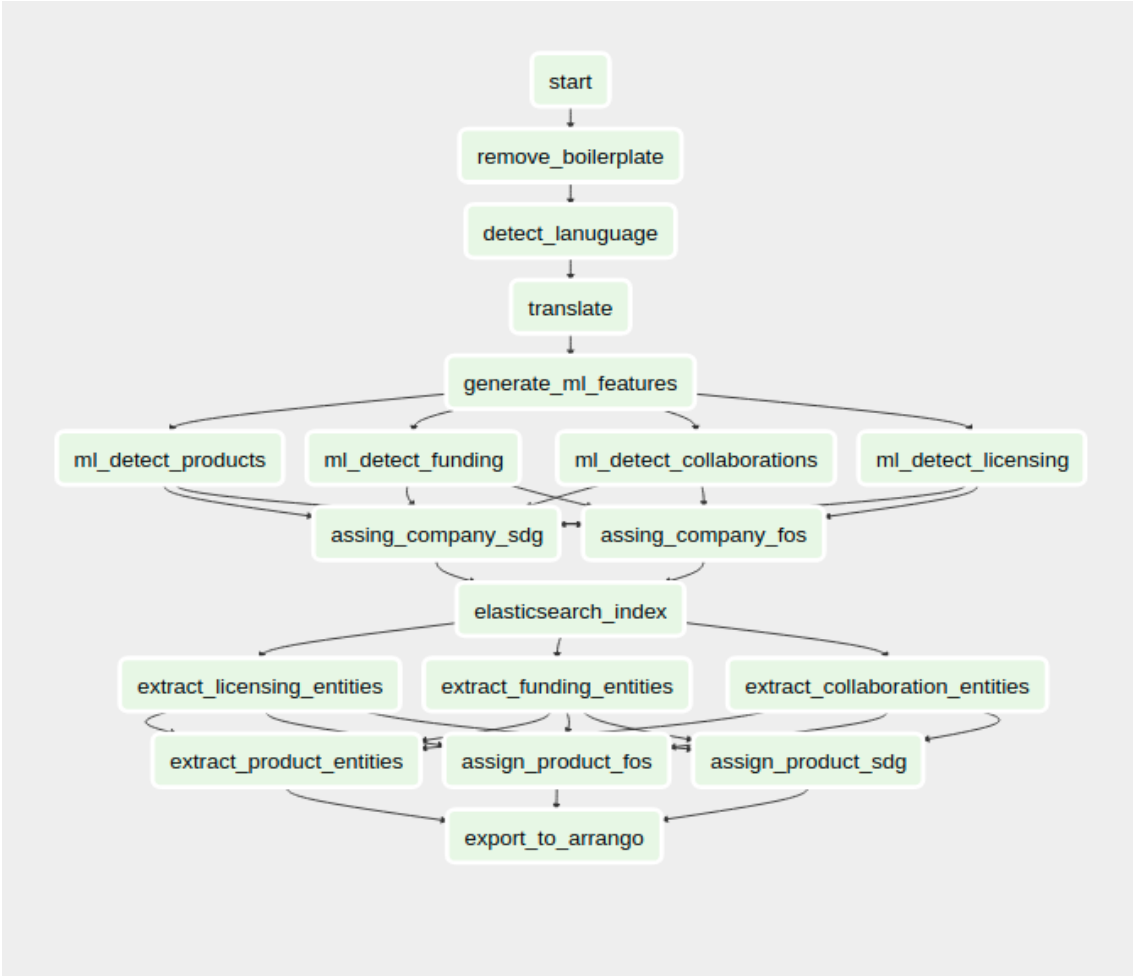
- In Area 2 company and other websites data as well as semi-structured database data is pooled together. Then various data mining, information extraction, text classification and text fragment matching algorithms are run in order to:
  - Identify and extract valuable pieces of information from the collected raw data;
  - Identify texts with relevant content for further detailed analysis;
  - Match fragments of text, e.g. product names to other records to link and enrich the data;
  - Construct indicators from the collected data.

Operations in Area 2

- ***Web scraping***
- PPMI already had developed a powerful web-scraper which can fully traverse company domain and extract text from various elements, including dynamic Java Scrip sections of the page.

- ***Text mining***
- Company Products
- Collaborations
- Funding

- ***Text Classification***
- Fields of Study
- SDGs


start
remove_boilerplate
detect_lanuguage
translate
generate_ml_features
ml_detect_products
ml_detect_funding
ml_detect_collaborations
ml_detect_licensing
assing_company_sdg
assing_company_fos
elasticsearch_index
extract_licensing_entities
extract_funding_entities
extract_collaboration_entities
extract_product_entities
assign_product_fos
assign_product_sdg
export_to_arrango

Fields of Study (FOS) tagging

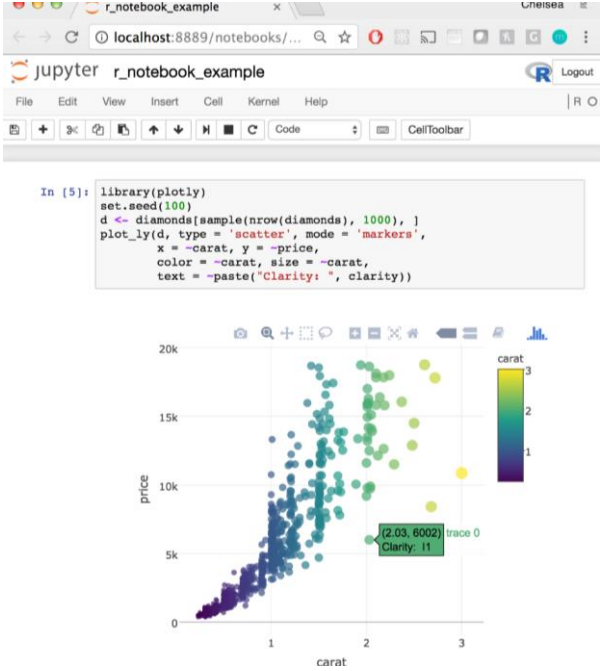
- We assign topic Field of Study categories based on those from MAG (https://academic.microsoft.com/topics) to companies and products based on their texts;
- By doing this, we enable to link texts based on the topics they cover, which allows for new ways to cluster and slice the data;

c. Area 3: Jupyter Notebook Server

They allow to facilitate the story telling – Jupyter notebooks work great in combining text narrative, code commands and data visualisations.

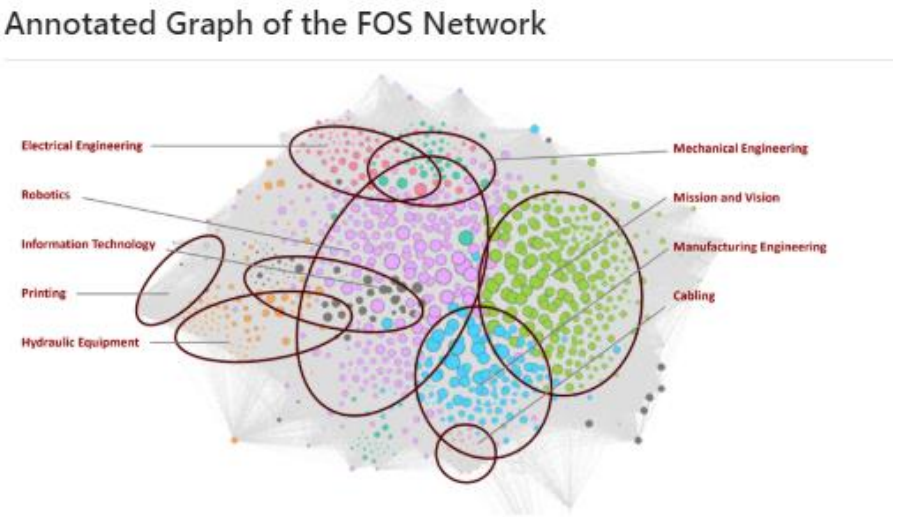
They allow to facilitate the interactive exploration of the data – Jupyter notebooks also allow the users to perform the data exploration and analysis on their own.

Jupyter notebooks facilitate several programming languages - all the main languages used form data analysis, such as R, python, Scala, and Julia.



Current operational status

- Objective is to get to ~200 000 companies
  - Samples for high, low and services have been created.
  - Webscraping is in progress.
- Early results are coming in and analysis has begun.
  - Meeting on te 27th will focus on these.



Discussion

Our next session will show our early descriptive findings