intelcomp

The vision, the results, the future

December 14, 2023

Information Day on a new era of STI policy making with AI

Our vision: background

- Open Research Data
- Using AI to obtain evidence for policy-making
- Appealing visualization of results
- Catalogue of web services:
  - For agenda setting
  - For proposal management

IntelComp's Goal

Use AI tools to provide policy makers with evidence based on data

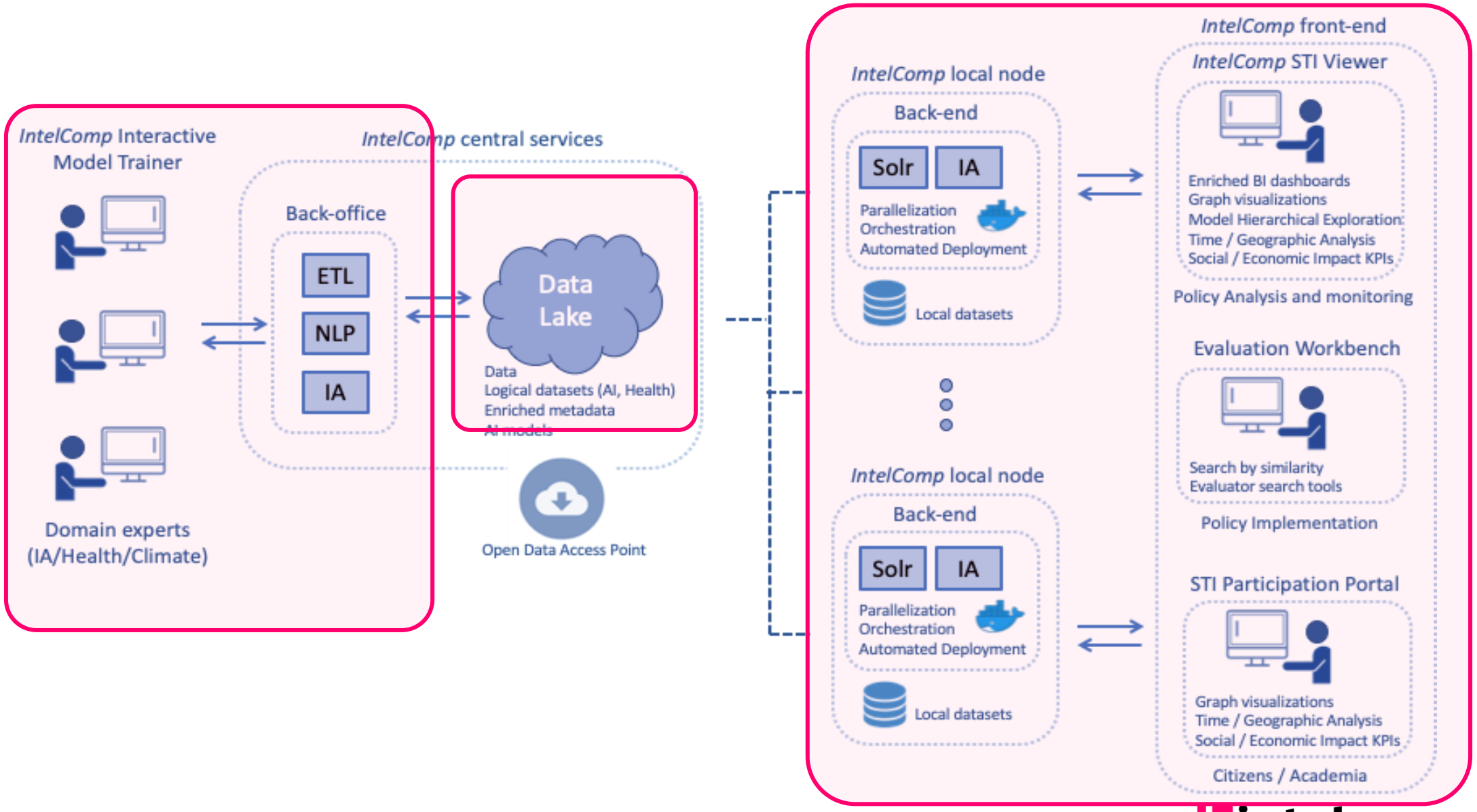
Industrialize the approach

Expert in the loop

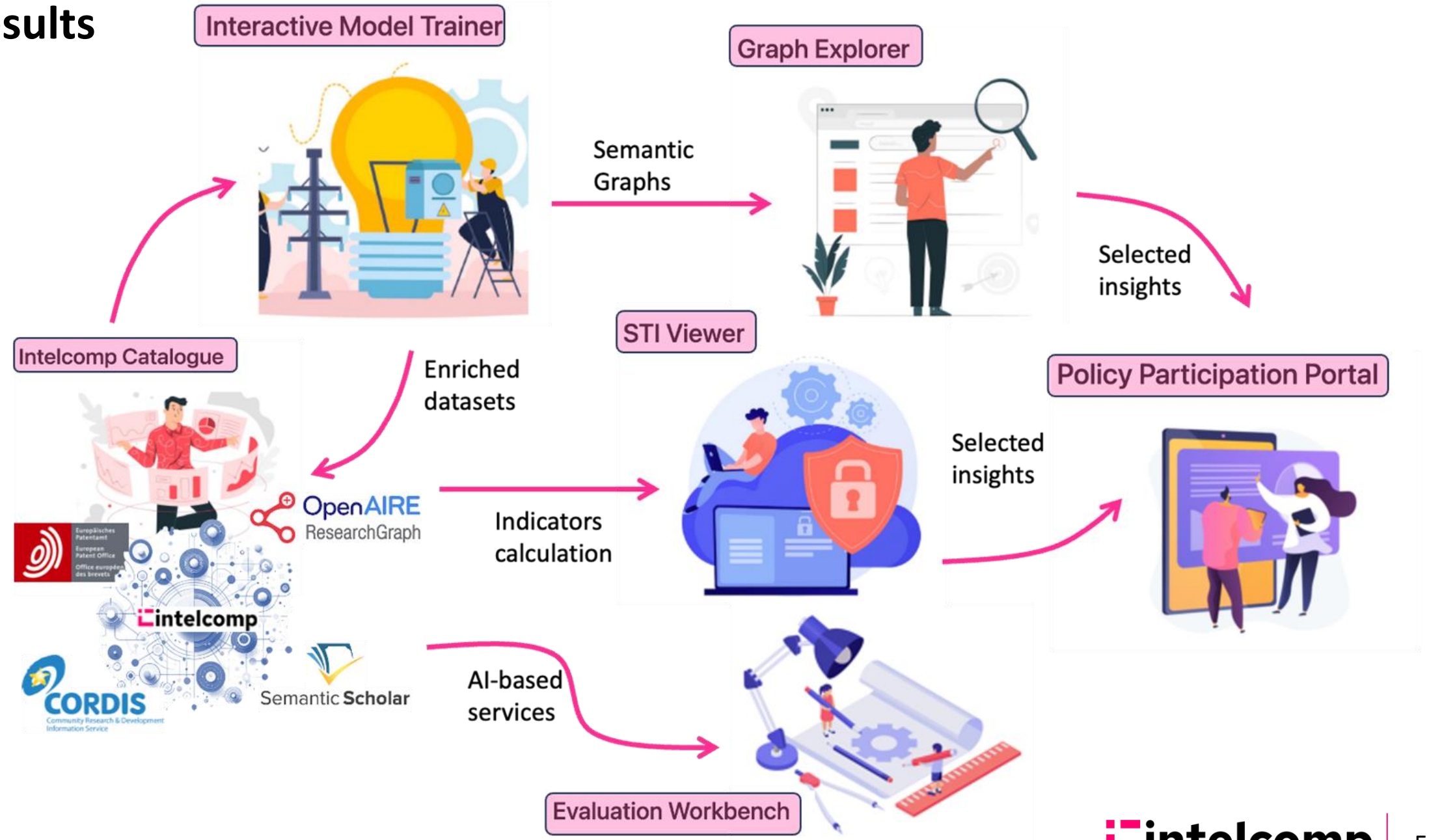
Living Labs

Integrated Platform

IntelComp's Concept & Original Design



The results

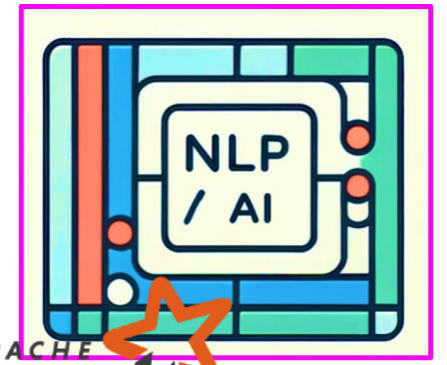

Interactive Model Trainer
Graph Explorer
Semantic Graphs
Selected insights
Intelcomp Catalogue
Enriched datasets
STI Viewer
Policy Participation Portal
OpenAIRE ResearchGraph
Indicators calculation
Selected insights
intelcomp
CORDIS
Semantic Scholar
AI-based services
Evaluation Workbench

Data Catalogue

- IntelComp's data lake collects several STI-related open data sets
  - Metadata heterogeneity
  - Variable curation quality
  - Very large datasets
- IntelComp's approach:
  - Unstructured datalake
  - Distributed storage
    (HDFS, parquet)
  - English as an anchor language
  - Text-based representations

Technologies

- Scalable NLP and NMT
  - GPU processing
  - Spark NLP for parallel processing
- Text-based AI
  - Supervised classification
  - Zero-shot classification
  - Document domain selection
  - Topic Modeling
- LLMs
  - Transformer-based document representations
  - GPT models to improve interpretability



TILDE

APACHE Spark

Spark NLP

dask

Hugging Face

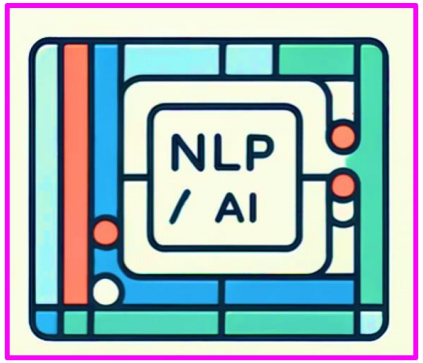
PyTorch

MALLET
machine learning for language toolkit

CTM

Solr

AI and NLP service catalogue



System for subcorpus generation

- Transformers
- Expert in the loop approach
- Relevance feedback

Topic Modeling Service

- Mallet
- Neural solutions (prodLDA, CTM)
- Expert assisted Curation tools
- Graphical tools for topic model exploration

Semantic Similarity & Graph Analysis

- Find documents with similar semantics
- Graph-based impact analysis

Classification service

- Transformers
- Zero-shot classification

Interactive Model Trainer

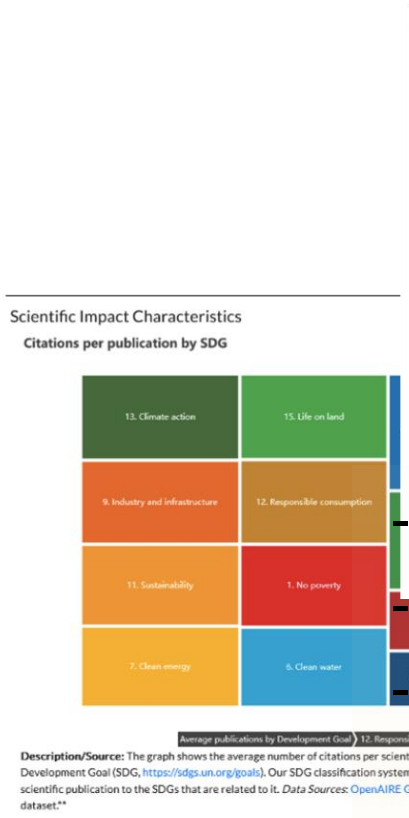
The goal is better experience with the STI Viewer and the EWB

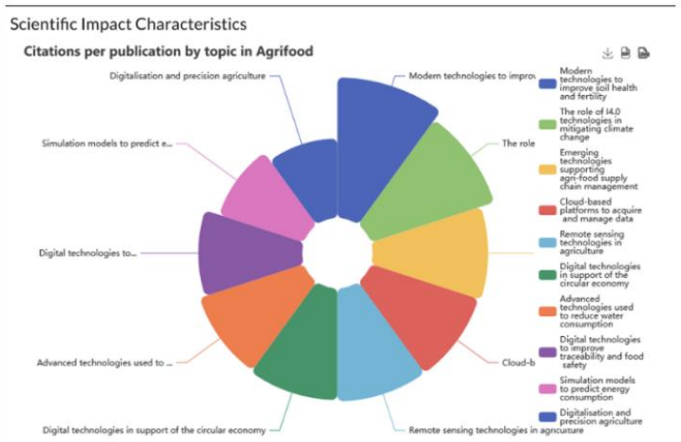
- Maximize alignment with domain experts' intuition
- Improve recommendations of the EWB

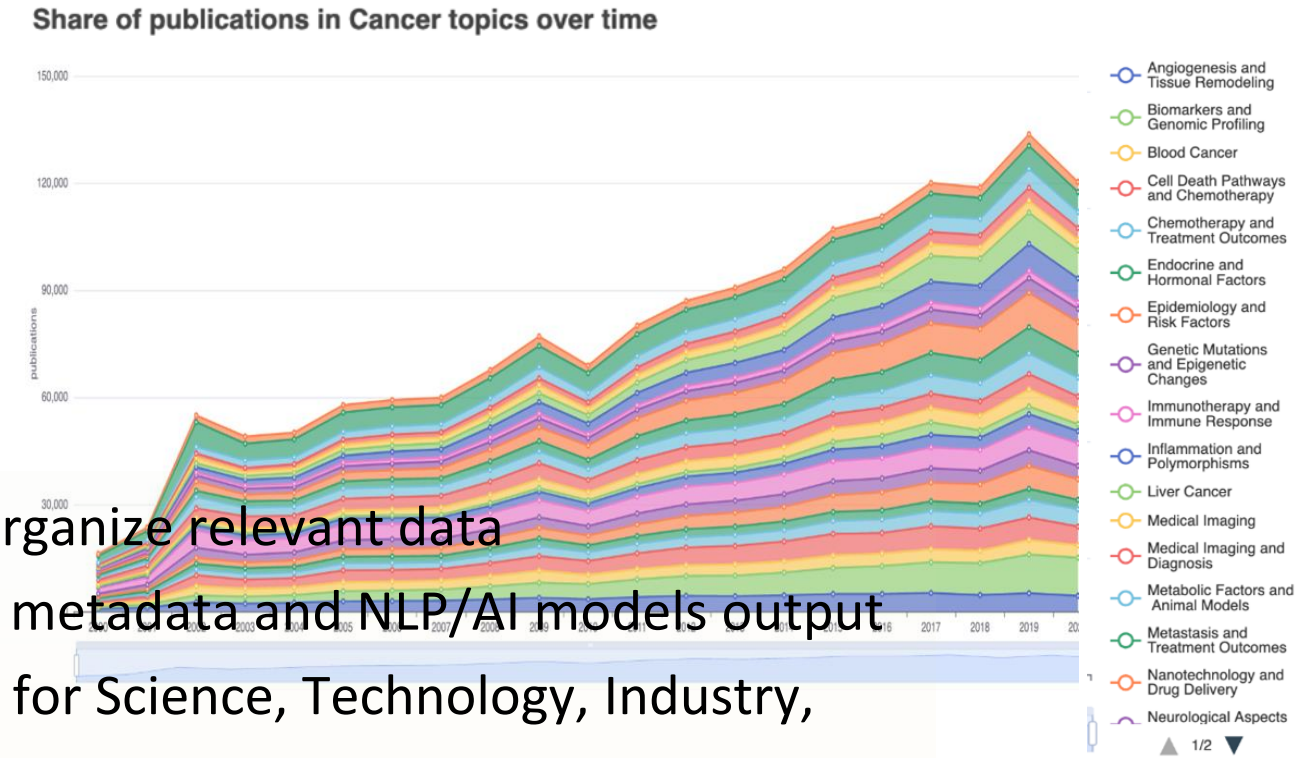
Indicators calculation and visualization: STI Viewer

Framework for STI Data Policy → Analysis in the context of the LLs → Technical feasibility analysis → Ingestion of Data Sources → Data Analysis & Model Training → "User Stories" definition

Identification of relevant indicators and data sources → Prioritization of the indicators → Tickets for the Developers → New data listed in Data Catalogue → Metadata Enriched Datasets → Visualization Dashboards





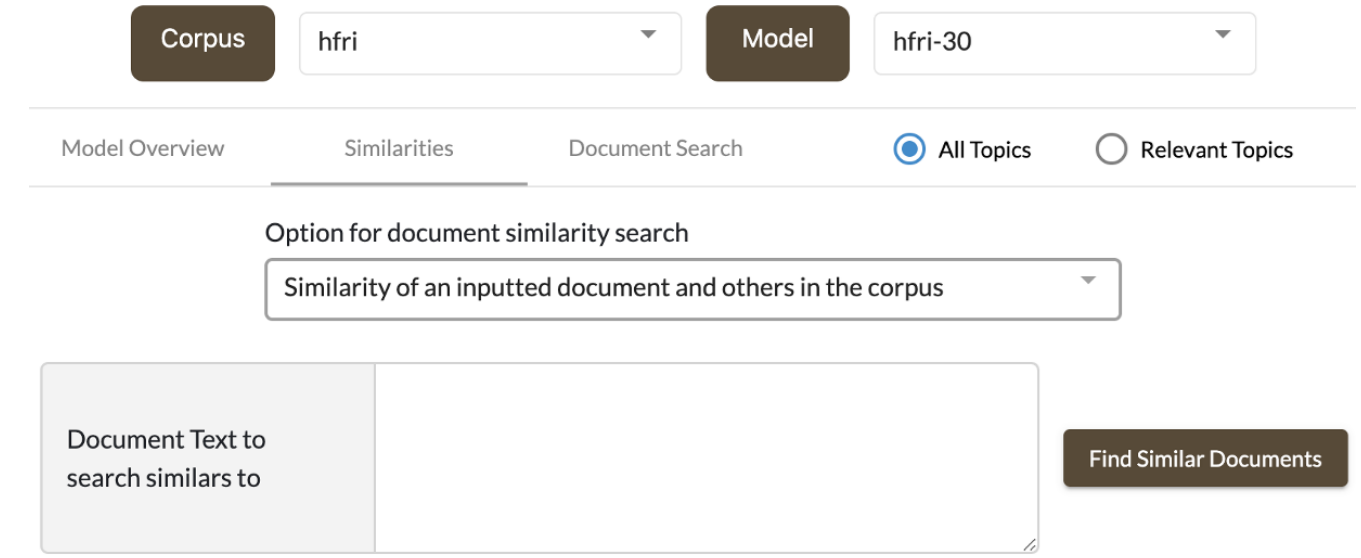


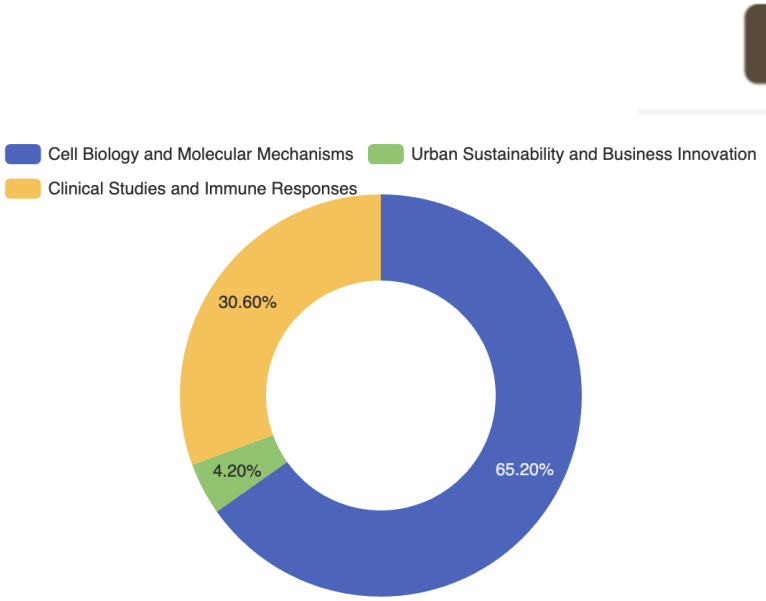
- User-defined storylines to organize relevant data
- Visualizations based on raw metadata and NLP/AI models output
- Used in the three living labs for Science, Technology, Industry, regulations indicators

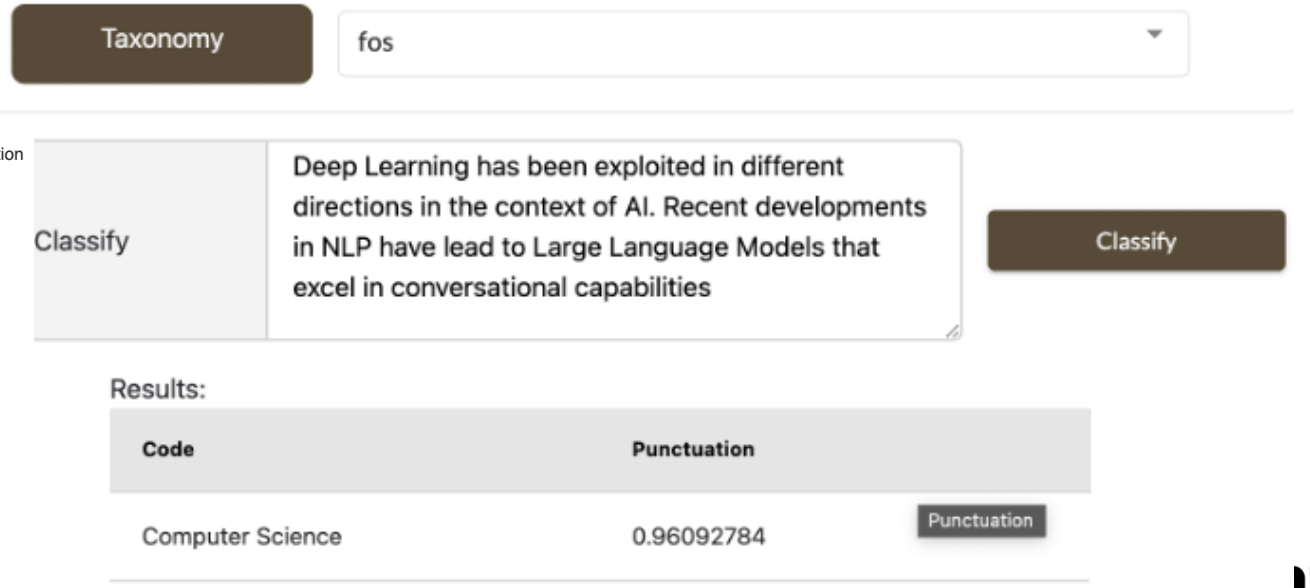
Graph explorer

- Nodes represent projects / publications, etc …
- Location is associated with semantic closeness
- Hierarchical exploration
- Use metadata to:
  - Filter nodes
  - Coloring criteria
- Export to Participation Portal

Evaluation of Proposals: Evaluation Workbench



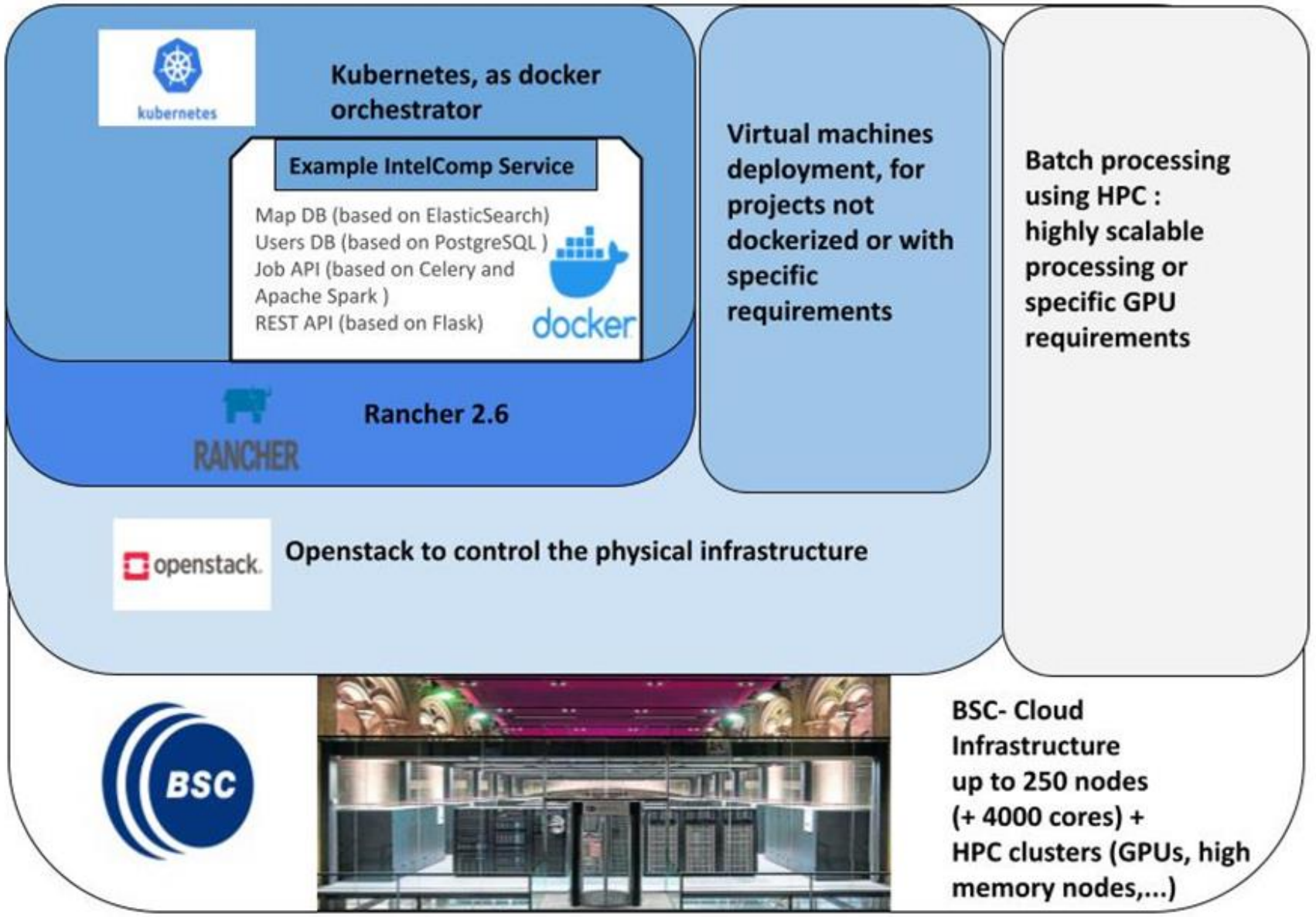




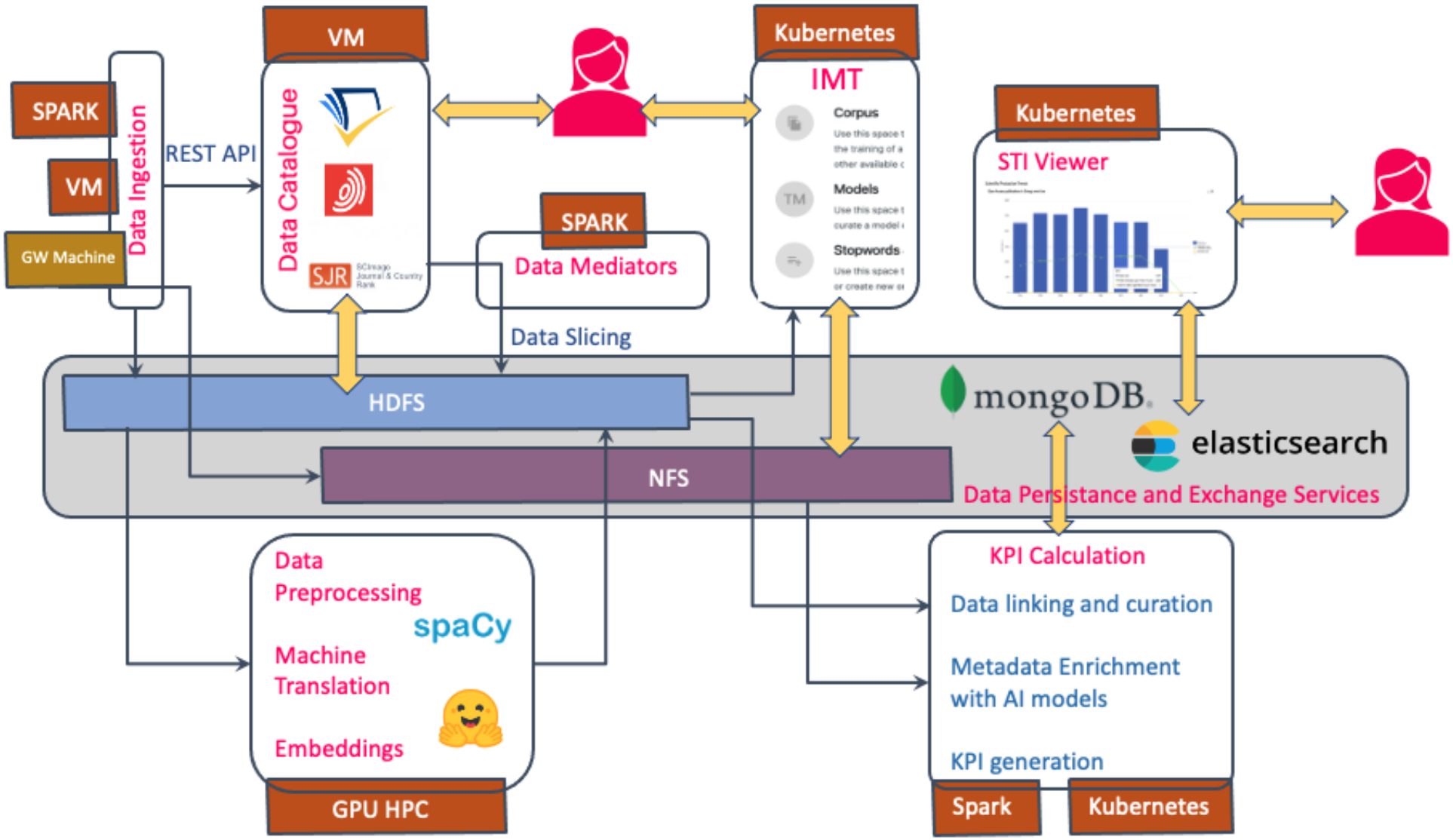
Powered by Topic Models and efficient services provided by Python and Solr backend:

1. Topic analysis of projects / experts
2. Search by similarity of projects / experts
3. Taxonomical classification of project abstracts

Infrastructure



Service Deployment



The future

- Platform infrastructure active in BSC for 2 years
- Ensuring and simplifying high-value workflows
- Data maintenance and update
- Continue fine-tuning of web services
- More intuitive interaction with the platform, LLM-like

intelcomp

The vision, the results, the future

https://apps.intelcomp.bsc.es