NLP and AI technologies and Platform development

Jerónimo Arenas-García (Universidad Carlos III de Madrid)

Apr 27, 2021

IntelComp Opening Event

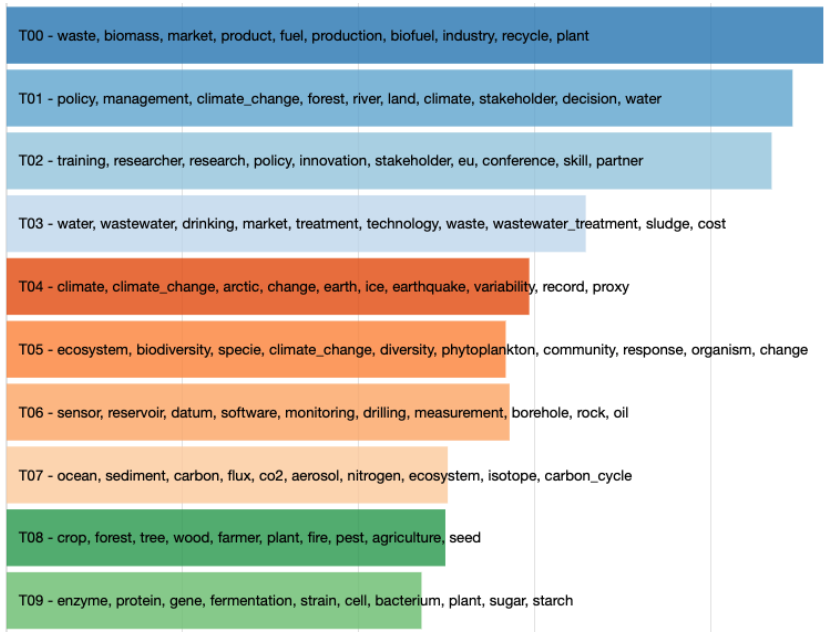
Statistical classification approach problems

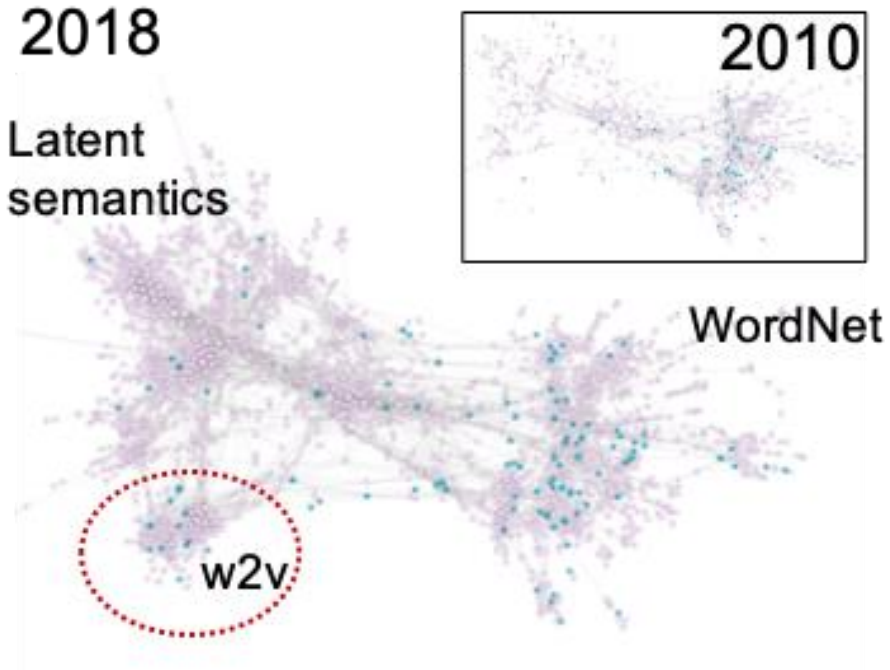
- Long delay in the availability of data from statistics and surveys (sometimes 1-2 years)
- Not suitable grain of definition on information (ex. HMI IPC codes).
- Poor document representation; binary multicomponent vector of classification labels (with few human defined components).
- Heterogeneous classifications, usually not compatible between distinct corpus (IPC-AJSC-Cordis classification-NSF/NIH class. …).
- Problems with classification time evolution: speed of adaptation, backward compatibility (ex. IPC versions).
- Disjunctive classifications for business intelligence or statistical use; hybrid projects

We decide to use also TEXT contents (title, abstract, contents, …)

NLP tools for Topic Modeling and Document Representation

- Topics offer insight at different levels of granularity
- Topics allow to align heterogeneous corpora and track topics over them
- Topics can be used to measure similarity among documents
- Topics can be used to calculate “semantic graphs”
- Topics can be exploited together with other metadata

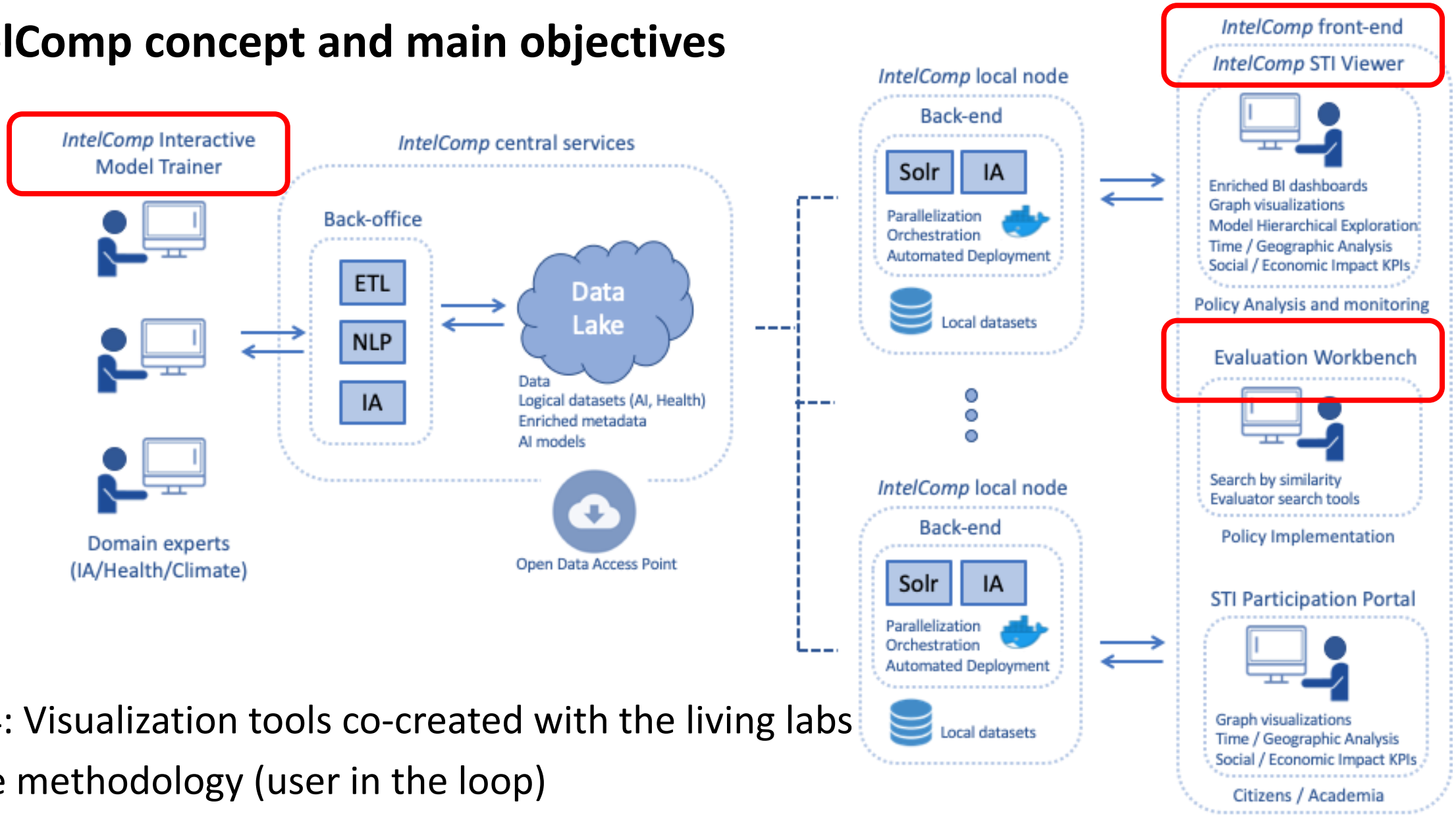




IntelComp WP3: NLP and AI services

- Scalable multilingual NLP pipelines
- Neural Automatic classification
- Topic modeling (static, temporal, hierarchical)
- Information retrieval based on topic similarity and keyword search
- Scalable document graph generation and analysis (GPU implementations)
- Lead-lag detection (thematic lead/lag between corpora)
- Agent disambiguation and characterisation (authors, organisations)
- Agent profiling with embedding techniques
- Short- and Long-term Impact analysis

IntelComp concept and main objectives



WP4: Visualization tools co-created with the living labs

Agile methodology (user in the loop)

Interactive visualizations (D3.js, dc.js, sigma.js)

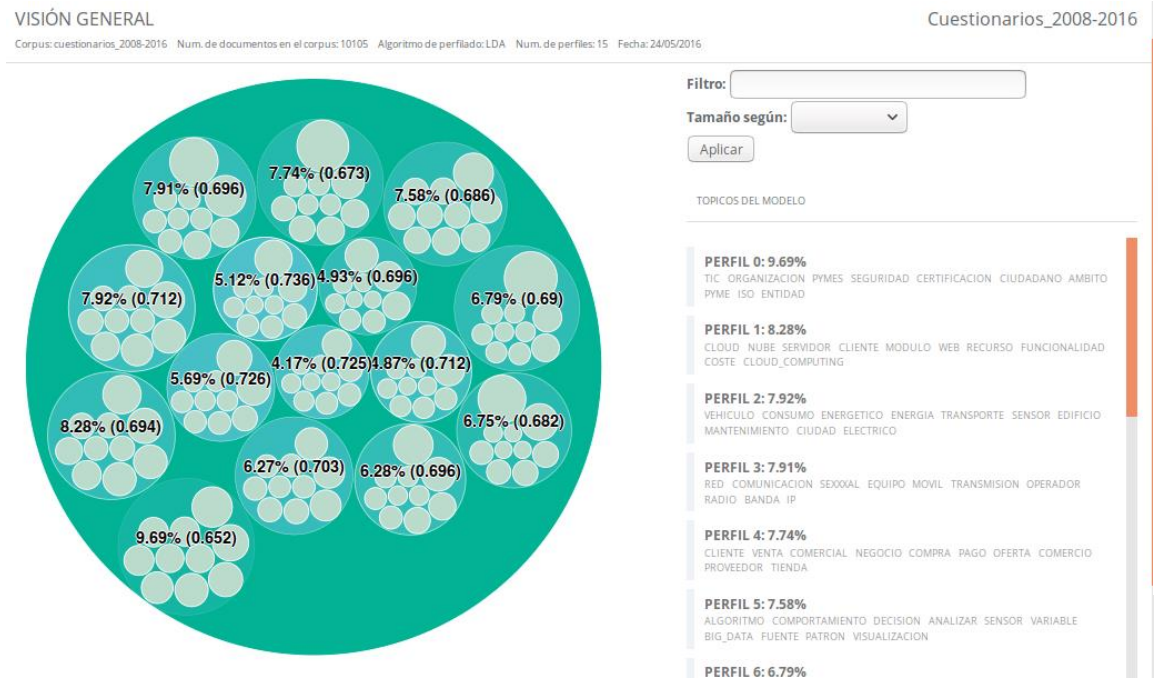
Tentative Services for Policy Makers

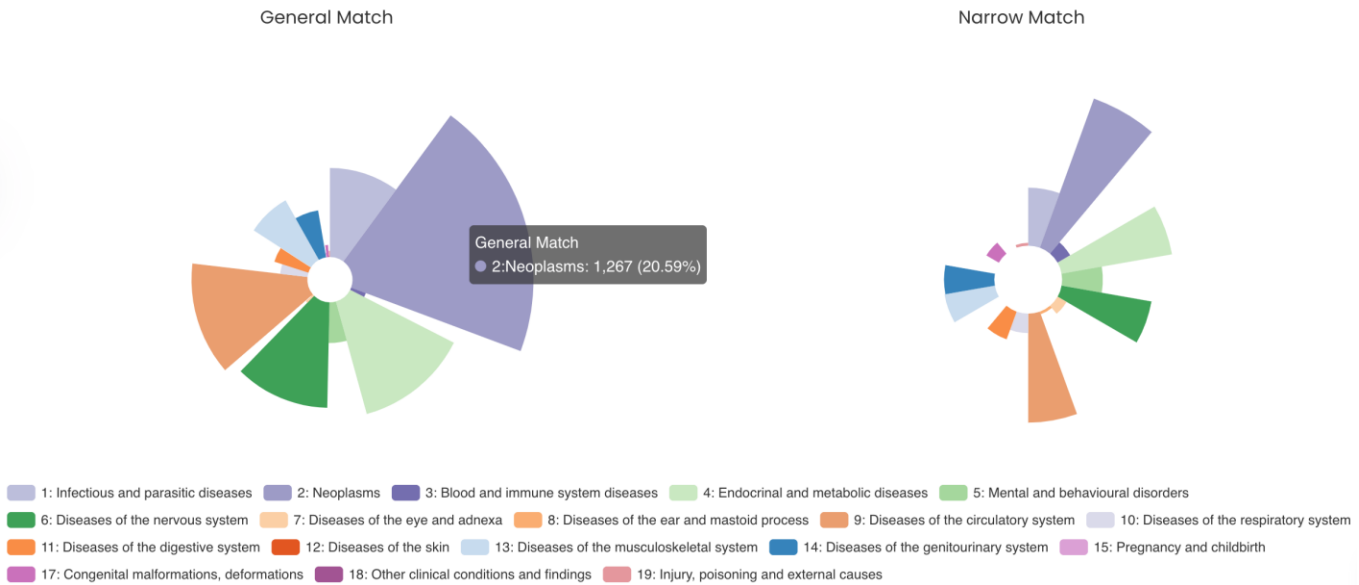
Policy Making Phase	R&D Policy Task	*IntelComp* services that address the task	Init TRL	End TRL
Policy / Programme design	Describe context to detect key emerging scientific areas and technologies	Dynamic thematic analysis of patents, publications and granted projects. Detection of lead-lags across different geographic areas.	7	9
	Describe context to identify target population	Thematic profiling of researchers, organizations and regions.	4	9
	Diagnose strengths in key scientific areas and technologies	Comparison of thematic distribution of patents and scientific production and granted projects.	7	9
	Define coherent funding instruments	Thematic distribution of funding by instrument; by agent profile; and by geographical level	7	8
	Co-creation design funding instruments	Thematic analysis and comparison of open consultations and "revealed" societal demands.	6	8
Policy / Programme implementation	Assign evaluators to proposals	Thematic characterization of available evaluators of a given proposal. Identification of key authors using graph analysis tools.	6	9
	Proposal positioning within STI information space	Locate patents, papers and proposals semantically similar to the given proposal.	7	9
	Avoid bias in selecting proposals	Comparison of thematic distribution of funded and not funded proposals.	7	9
	Assign proposals to areas or topics	Automatic classifiers for taxonomies, to help in the assignment of proposals to areas or topics.	6	9
	Duplicate research funding	Detection of similar documents by semantic content.	7	9

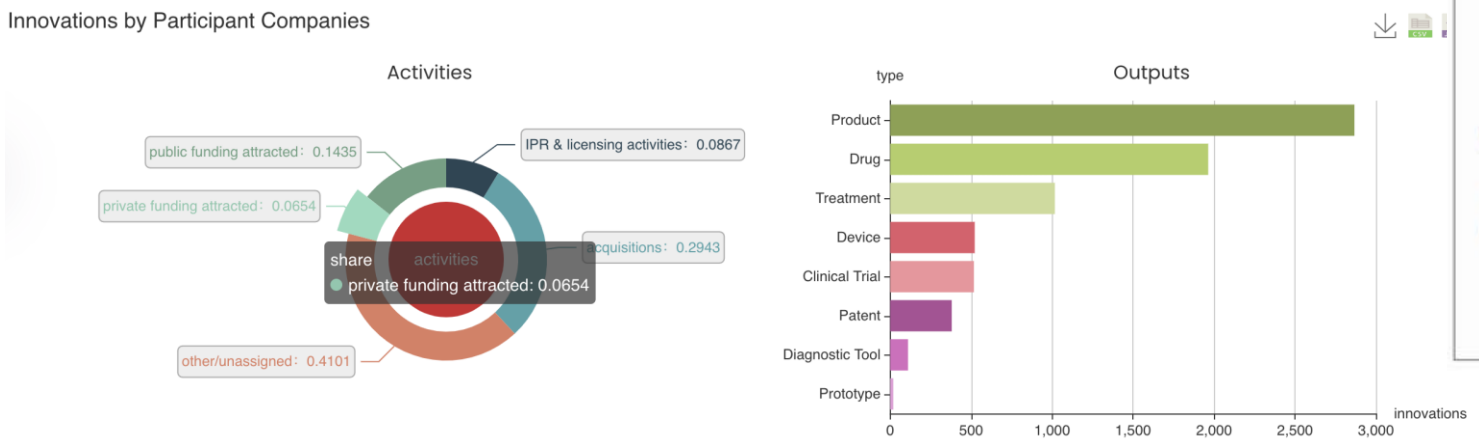
Tentative Services for Policy Makers

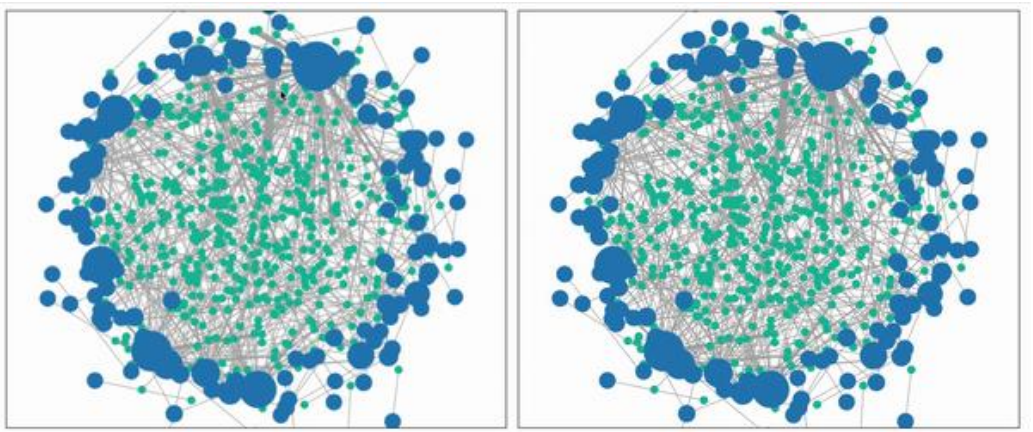
Policy Making Phase	R&D Policy Task	*IntelComp* services that address the task	Init TRL	End TRL
Policy/ Programme assessment	Link grants with outputs through citations and agents	Author and organization disambiguation.	6	9
	Link grants with outputs through semantic analysis	Semantic distance calculation for documents from heterogeneous corpora.	7	9
	Assessment of project and program direct results	Quantitative and qualitative monitoring information on direct results distributed by topic	6	9
Policy/ Programme impact assessment	Link grants with outputs through citations and agents	Author and organization disambiguation	6	8
	Link grants with outcomes through semantic analysis	Node influence analysis in citation and co-citation graphsNode influence analysis in semantic graphs	6	9
	Assessment of project and program impact	Quantitative and qualitative monitoring information on impact distributed by topic	5	6
	Influence in Media / Social Media	Identification of Trending topics Lag analysis between research field and media impact	N/A	9

Interactive visualizations

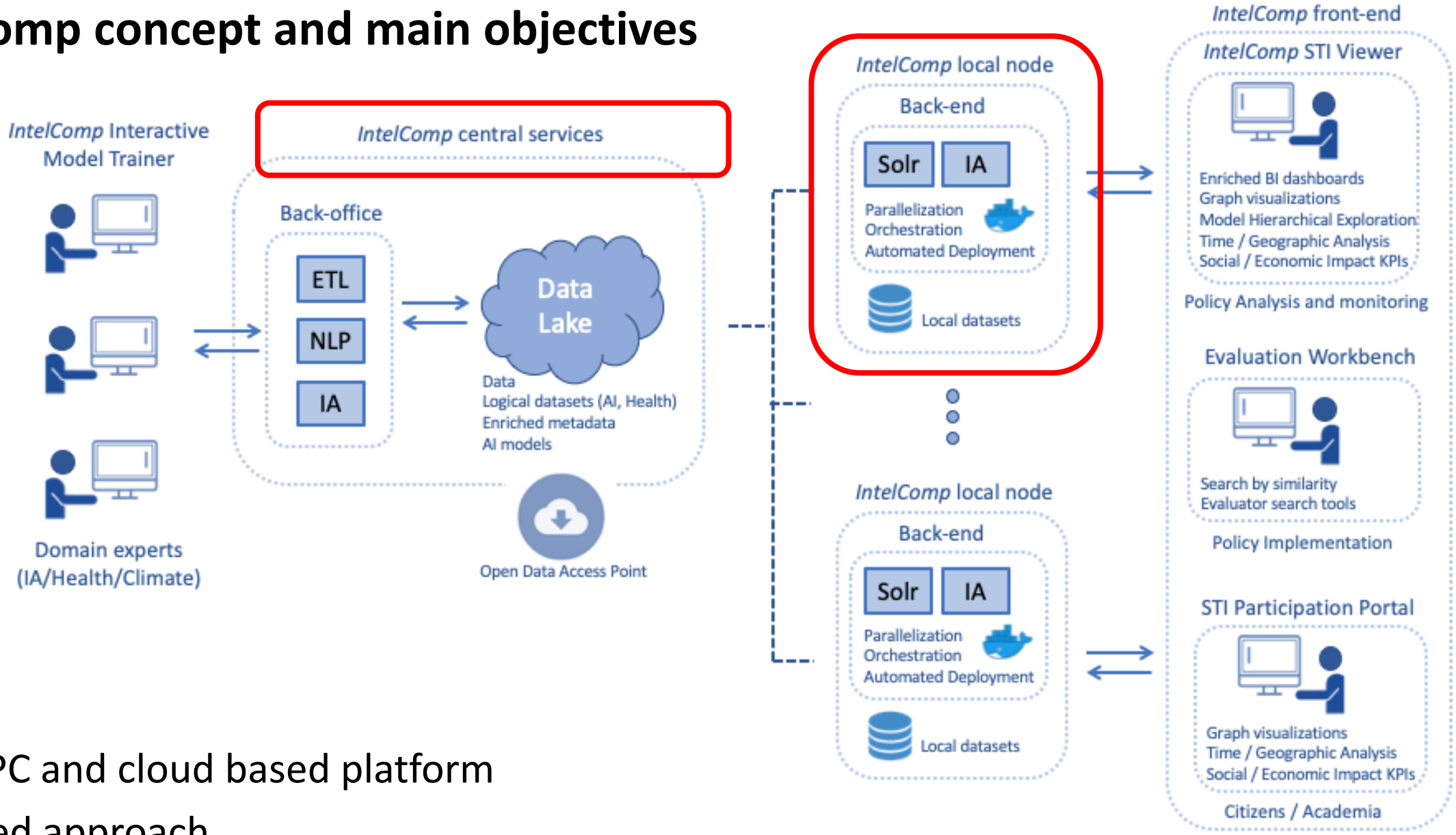








IntelComp concept and main objectives



WP5: HPC and cloud based platform

Federated approach

High Scalability for model computation and service deployment

Implementation and Deployment

- Java 1.9 & Python 3 code
- Frontend: Bootstrap JS + D3.js + Banana Lucidworks (AngularJS)
- NLP Pipeline: IXA Pipes (OpenNLP + StanfordCore NLP)
- Topic library: Mallet, Gensim, Pytorch NN implementation (beta)
- Search Engine SolR 7.X + Banana Lucidworks
- Apache + Tomcat, Postgres
- Prometheus + Graphana
- CD/CI environment GOCD
- Deployment: Ansible + Kubernetes (K8s) + Dockers containers

Implementation and Deployment

- Java 1.9 & Python 3 code
- Frontend: Bootstrap JS + D3.js + Banana Lucidworks (AngularJS)
- NLP Pipeline: IXA Pipes (OpenNLP + StanfordCore NLP)
- Topic library: Mallet, Gensim, Pytorch NN implementation (beta)
- Search Engine SolR 7.X + Banana Lucidworks
- Apache + Tomcat, Postgres
- Prometheus + Graphana
- CD/CI environment GOCD
- Deployment: Ansible + Kubernetes + Docker

- Efficient Search of Documents Based on Text
- Filters and Facets allow restricting the analysis to a subcorpus selection, e.g.,
  - For a particular funding institution, call
  - For a particular time span
  - For a particular (group of) contry(ies)
- d3.js visualizations can be developed and integrated in banana dashboards

intelcomp

https://intelcomp.eu

Jerónimo Arenas García (jeronimo.arenas@uc3m.es)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004870. H2020-SC6-GOVERNANCE-2018-2019-2020 / H2020-SC6-GOVERNANCE-2020

IntelComp WorkPackages