FORUM STIC

Paris-Saclay 2013

Olivier MESNARD

N°6.2

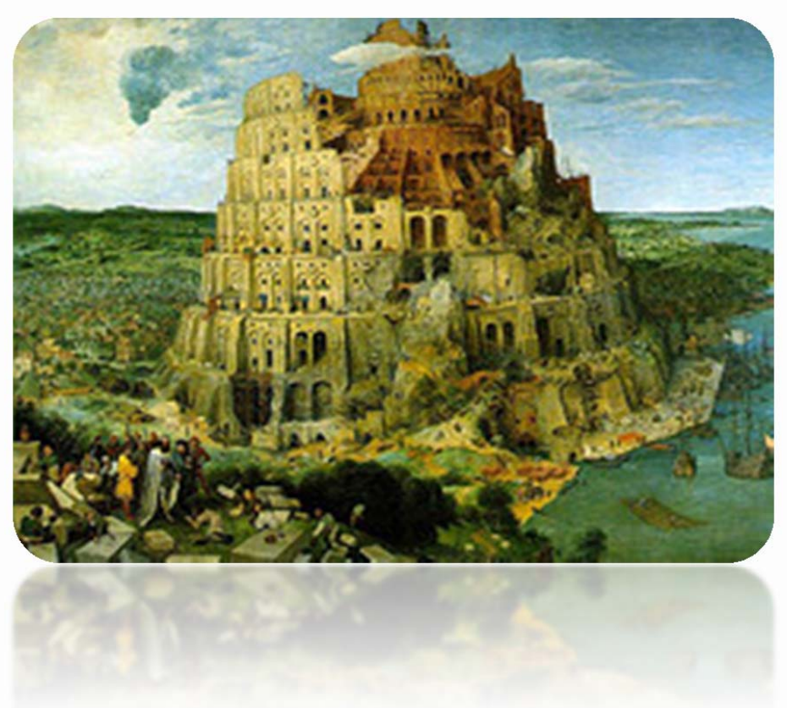
Multimedia Multilingual Integration

Context and issues

We are drowning in information and are still starving for knowledge

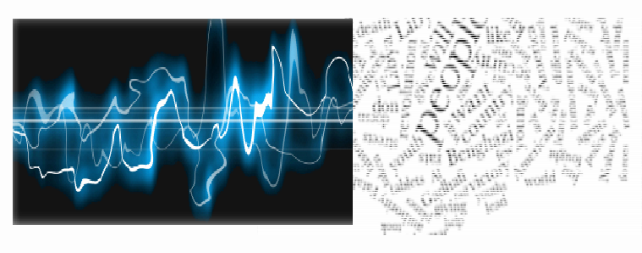
- Exploding growth of available information
  - Traffic, internet users and size of data double every two years
- Growing interest in sources of more diverse nature
  - Audio, video
  - Blogs, social networks
  - Under-resourced languages
- Poor quality of processing on user generated content
  - NLP sensitive to noise and ill-formed text, low recall
- Low productivity of practitioner of open source intelligence
  - Lack of visual analytics or human computer interaction
  - Lack of high level functionality: faceted search, clustering of documents, structured knowledge extraction, trusted sources…
- No integrated solution for content analytics
  - Cost of integrating technologies from more than one provider

Multilinguism



Social Networks

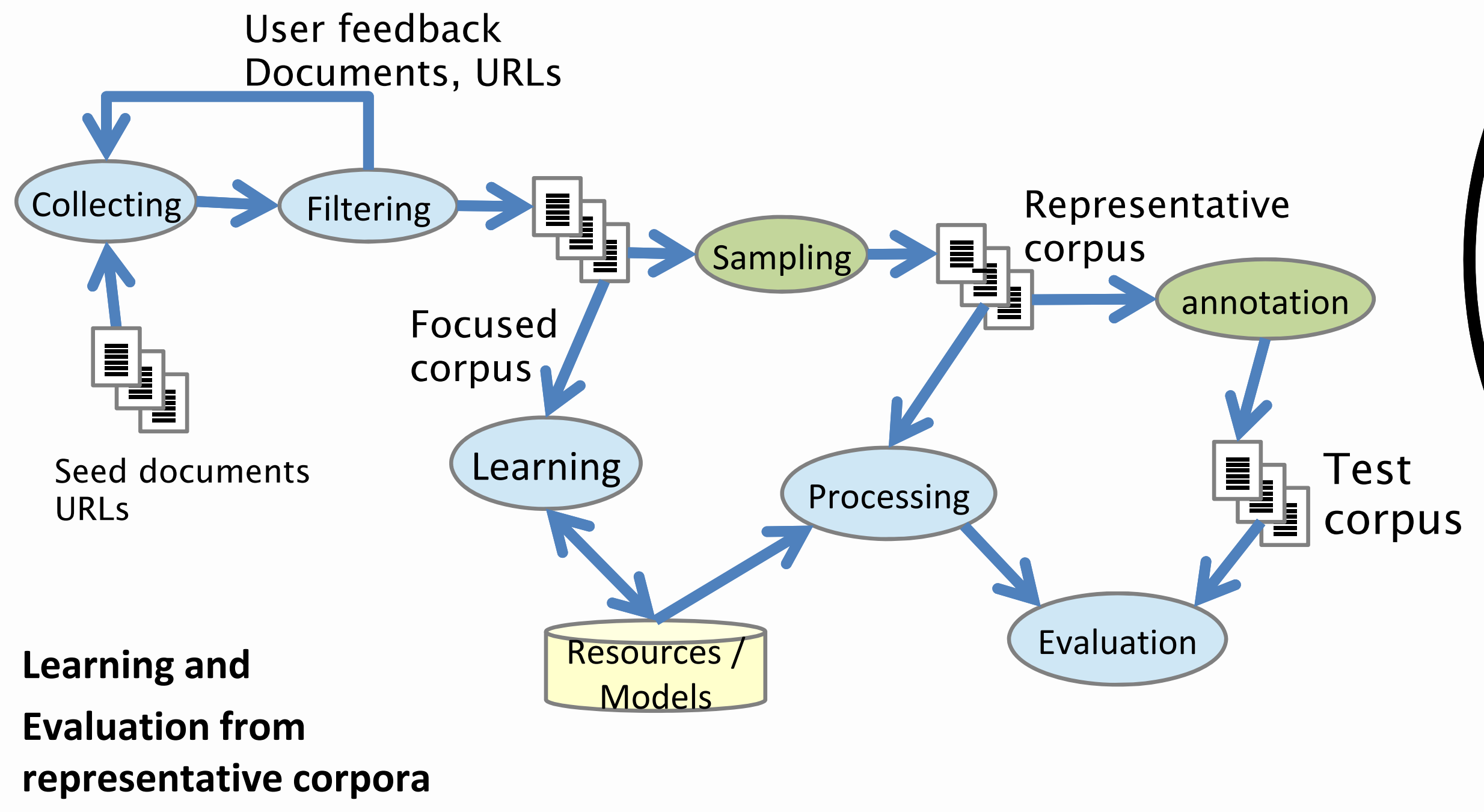
Multimedia



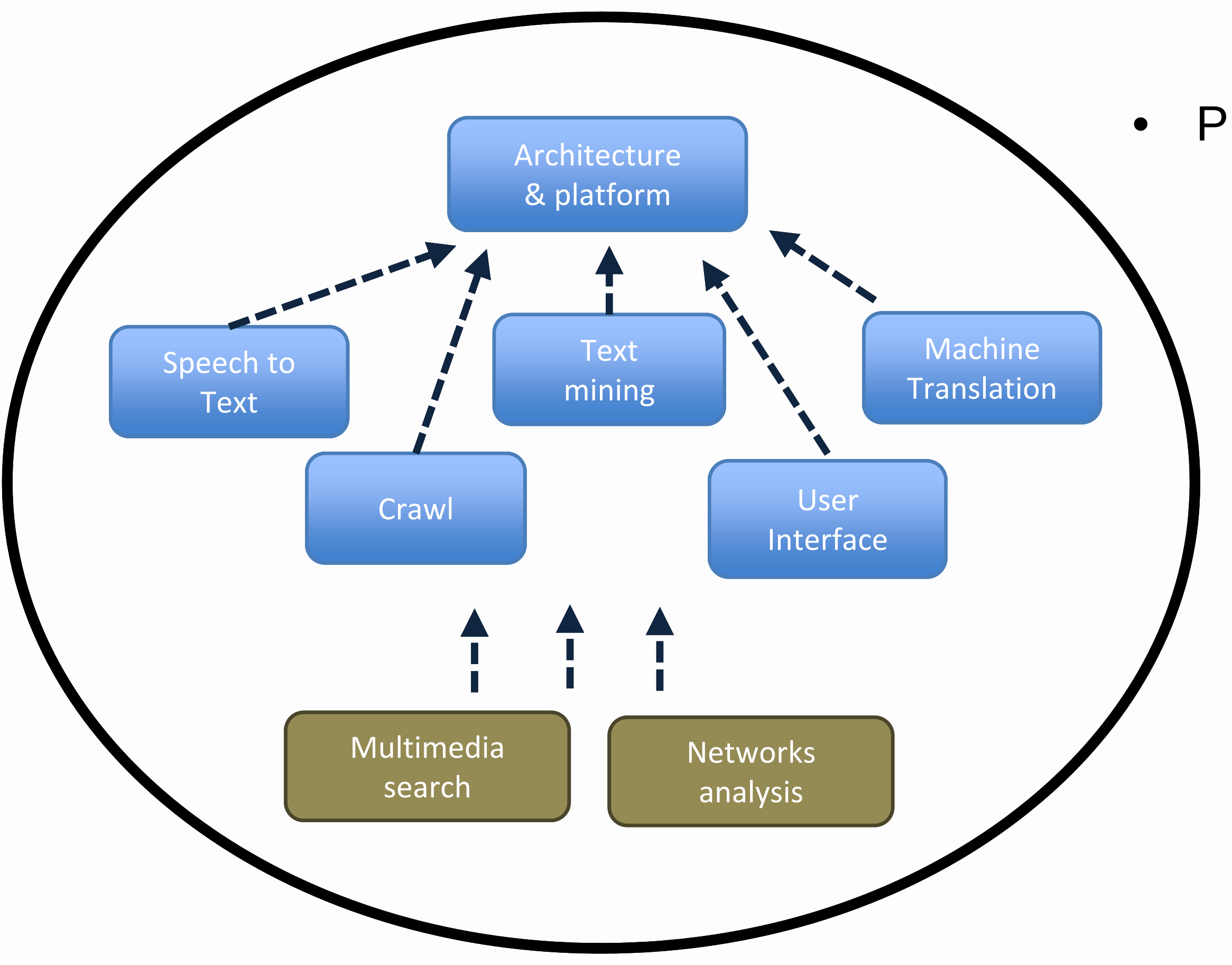
Objectives

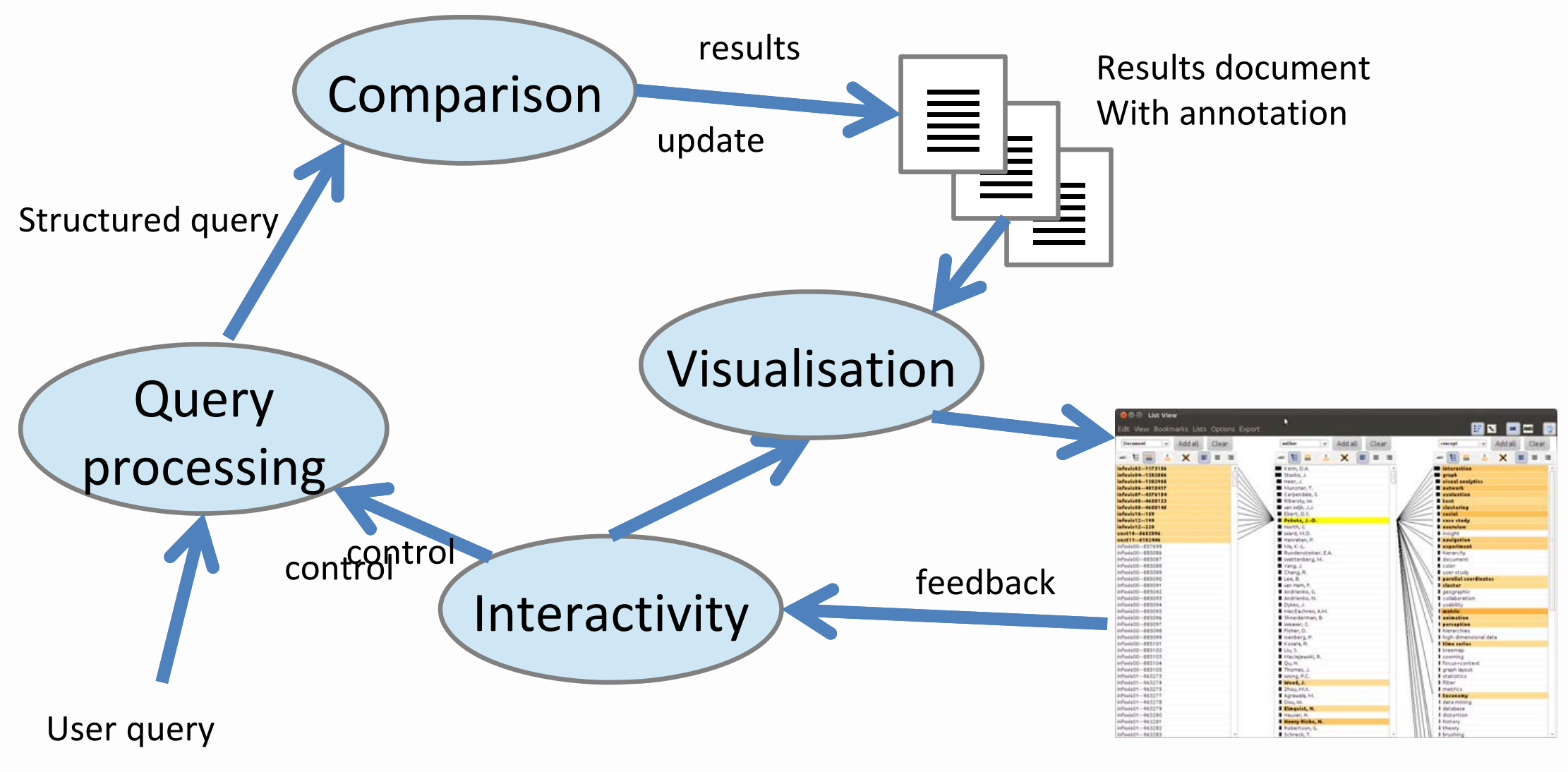
Collaboration, user focused enhancement

- Build a platform to experiment and evaluate technologies for data mining for non (or loosely) structured information (text, audio, video)
  - Integrate technologies of projects members: automatic speech recognition, word spotting, machine translation, natural language processing, crawl, information retrieval, information extraction, analysis of graphs
  - Evaluation of quality from isolated components vs. integrated processing chain using objectives metrics
  - User Evaluation with significant data
- Reduce cost and optimize the process of adaptation to some new language or domain
  - Experiment on under-resourced language
  - Experiment with short-time development constraint
- Enhance quality of search with noise and different types of text
- Evaluate scalability of solutions
- Develop high level function:
  - Cross-lingual information retrieval, knowledge base population, monitoring of crawl, ..
- Prototype of application with high capabilities
  - Enable interactivity and usability with high volume



Learning and Evaluation from representative corpora





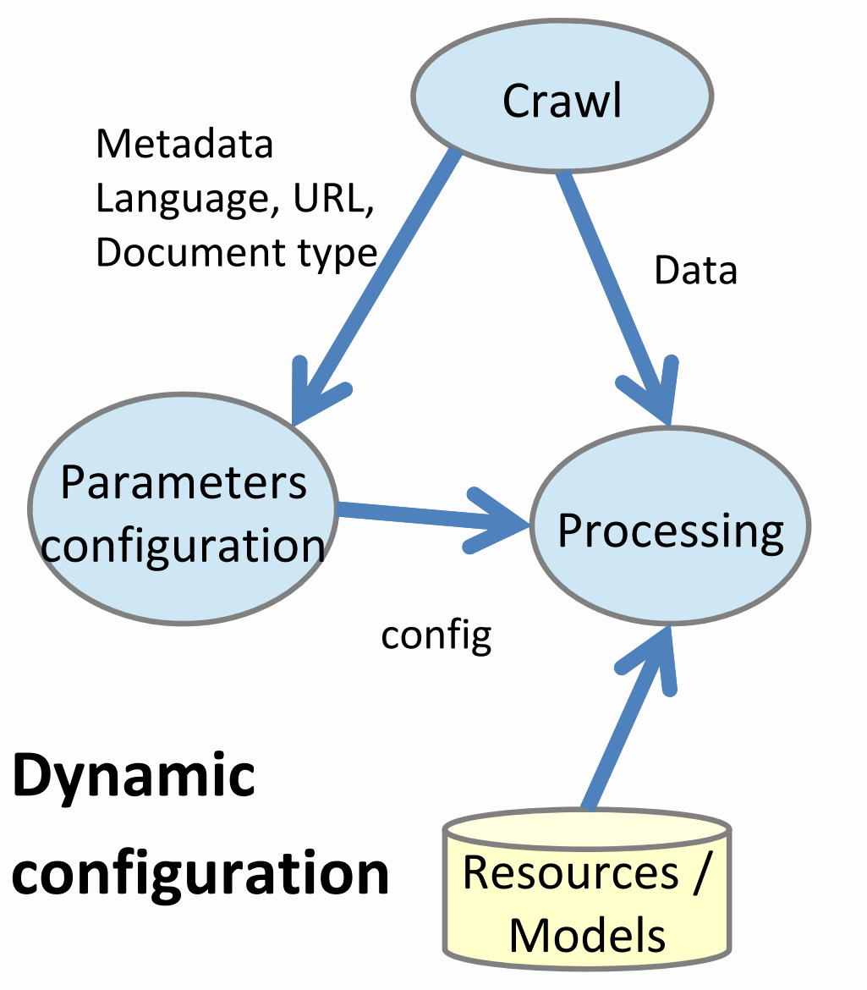
Innovations

Short term adaptation to usage context

- Domain adaptation for information extraction with limited prior knowledge
  - bootstrapping, distant supervision with database, co-training…
- Unsupervised learning of morphology model from text
  - minimal description length
- Cross-lingual projections of annotations for poor-resources languages
- …
- PhD 1: Knowledge base population from heterogeneous documents
  - Documents multi-source, multilingual, multimedia
  - Merging, aggregation, probability computation
  - Supervision by CNRS LIMSI
- PhD 2: Model and dynamic of information spread on network
  - Get rid of unlikely hypothesis: closed word, static graph and neutral message
  - Supervision by UPMC Lip6

Expected results

- Bring up an ecosystem with industrial partners, users and academics focused on unstructured data analytics
  - Ready to integrate best of breed complementary solution
  - Promote connectors relying on industrial standards
- Reduce delay and annotated resources to integrate new language and adapt system to new domain
- Optimize operating point of each component for the best quality of end-to-end output
  - Alignment of resources and metadata
  - Control and monitoring with common parameters
- User oriented evaluation
  - Tuned metrics for open source intelligence tasks
  - Integrated with high level interface
  - Evaluation by end-users and with representative data



Dynamic configuration

Contact: olivier.mesnard@irt-systemx.fr

Financial support:

Academic support: