Modeling Topic Evolution in Scientific Archives and the EPIQUE Project

Bernd Amann, David Chavalarias, Alexandre Delanoë, Ian Jeantet, Thibault Racovski

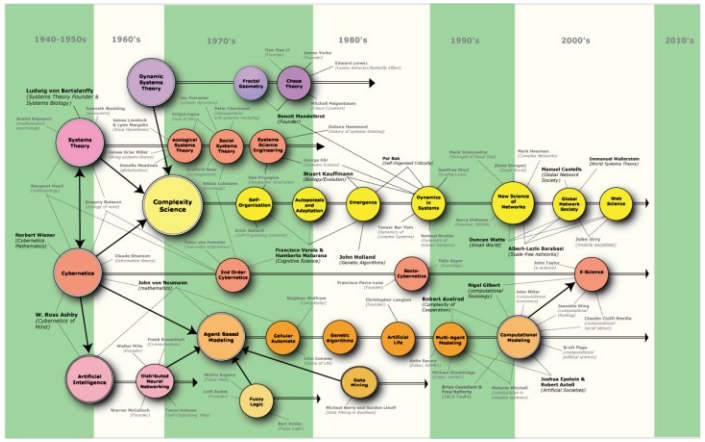
2

A Short Introduction to Topic Modeling

Problem Statement

- Given a collection of scientific **text documents**, derive a **structured representation** of scientific **concepts** and their **evolution.**
- **Questions** :
  - How transform **text into concepts** ?
  - How represent **evolution of concepts** ?
- **Text analytics** : deriving **interesting structured patterns** from text collections
- **Tools** :
  - Natural Language Processing (NLP)
  - analytical methods : linguistics, statistics, graphs, logics

text categorization, text clustering, summarization, **topic extraction**, **semantic analysis**, named entity recognition, sentiment analysis, trend detection, …



From Text to Topics

M	Cat	Milk	Cow	Grass
D1				
D2				
D3				
D4				
D5				

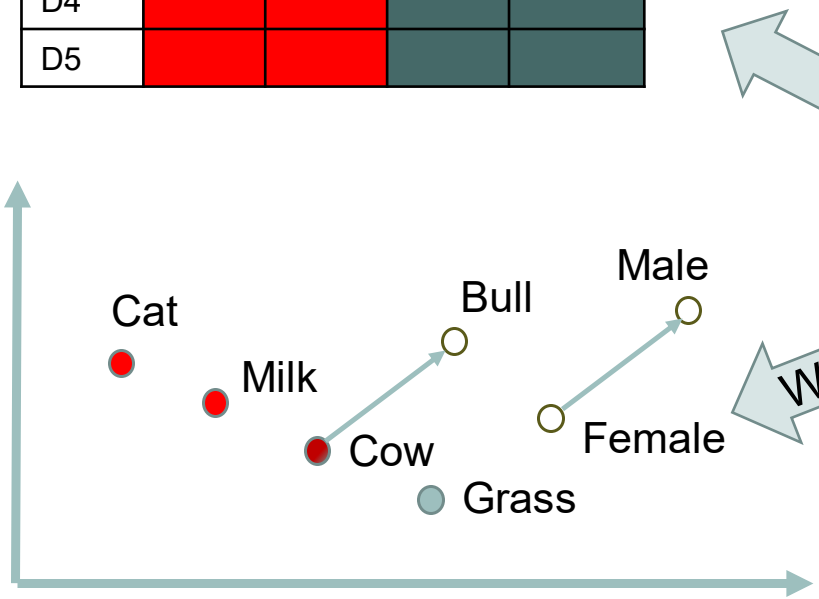
D1: **Cats** love the taste of real **milk**.

D2: Give your **cat** the **milk** they crave.

D3: Is **cows milk** bad for **cats**?

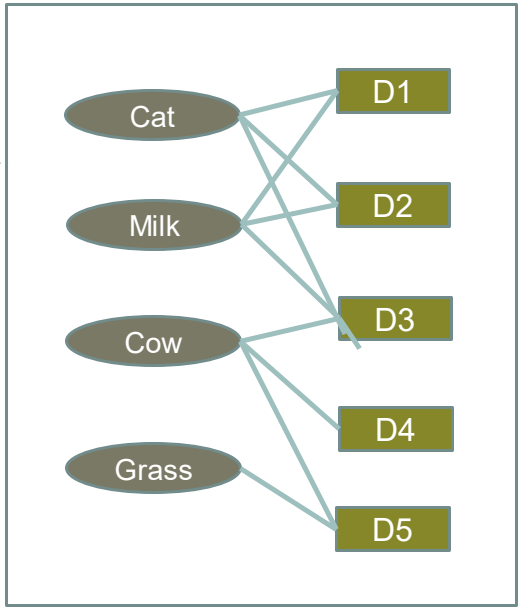
D4: Which **grass** is best for **cows**?

D5: Raising **cows** for **milk** is not an easy way to farm.

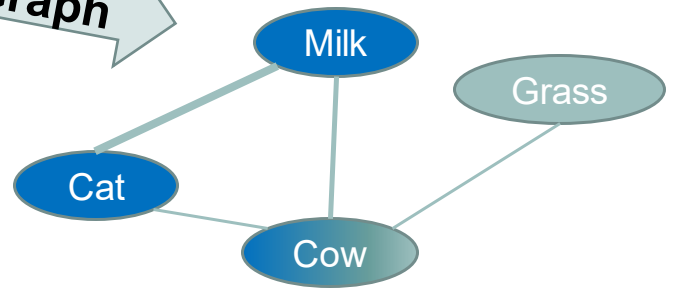


LDA

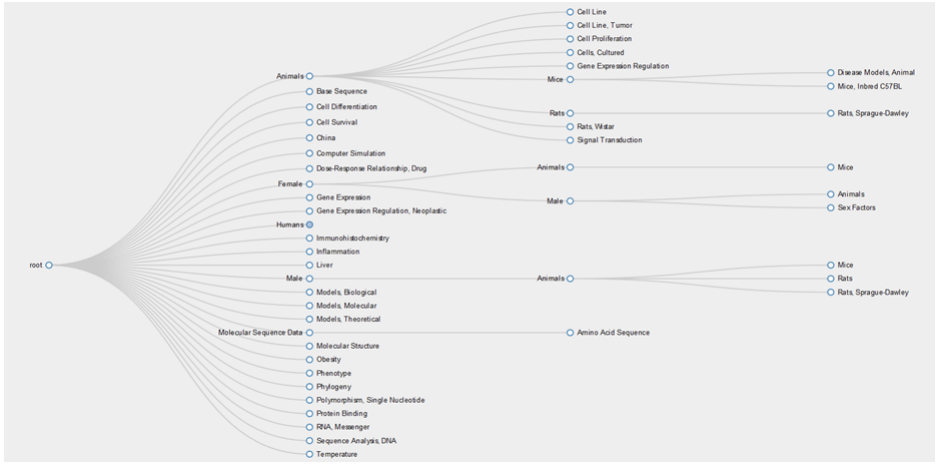
Word2Vec



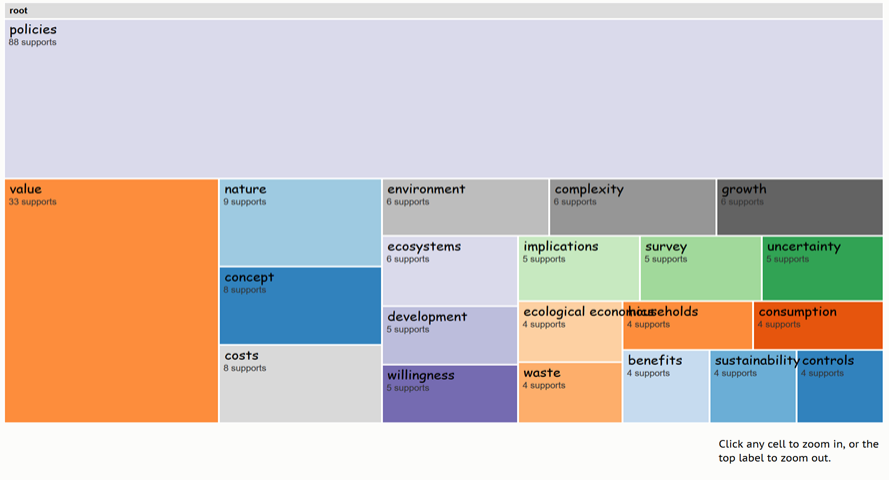
Graph

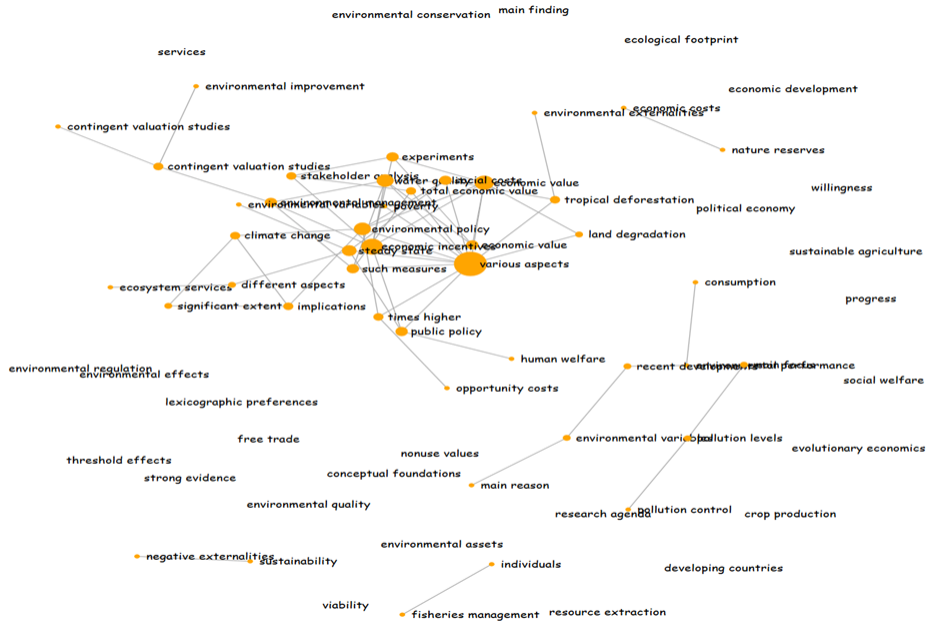


Topic Visualisation

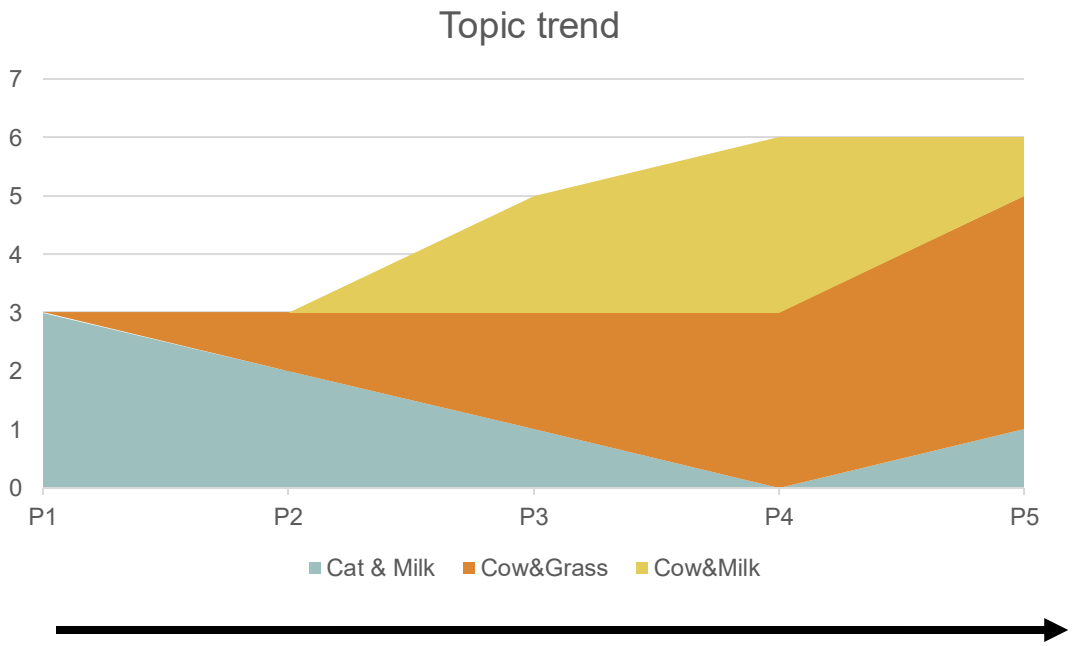


Zoomable Treemaps

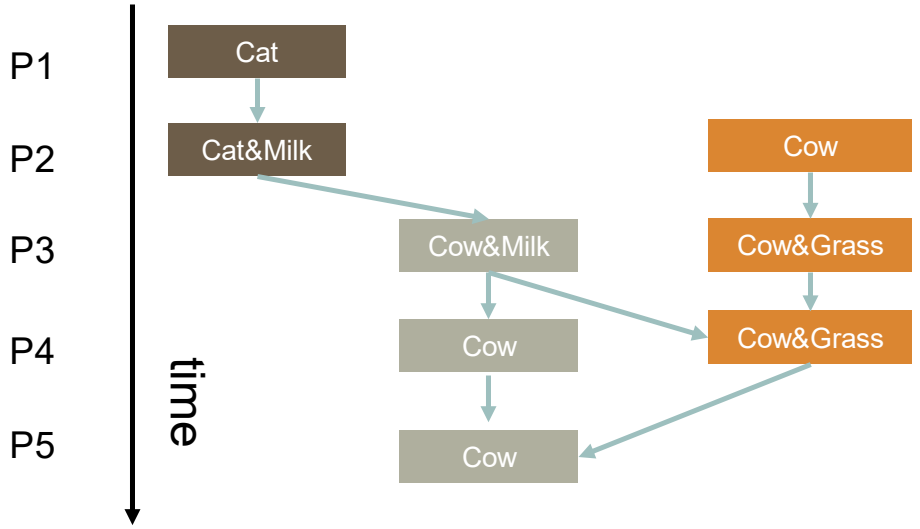




Topic Evolution



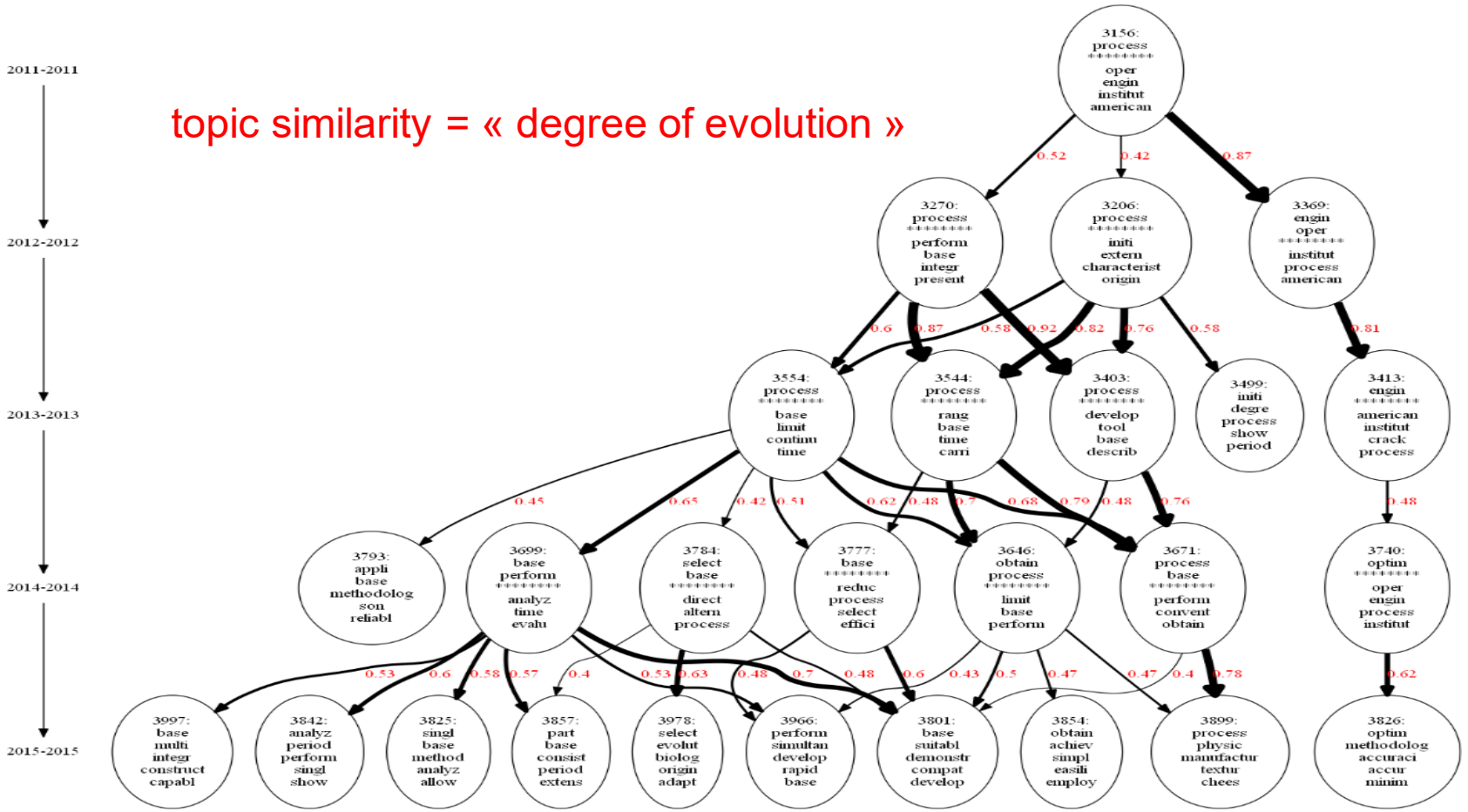
Topic Trends



Topic Networks (Phylomemies)

Phylomemy network

topic similarity = « degree of evolution »



8

So, what's the problem ?

Choices and Difficulties

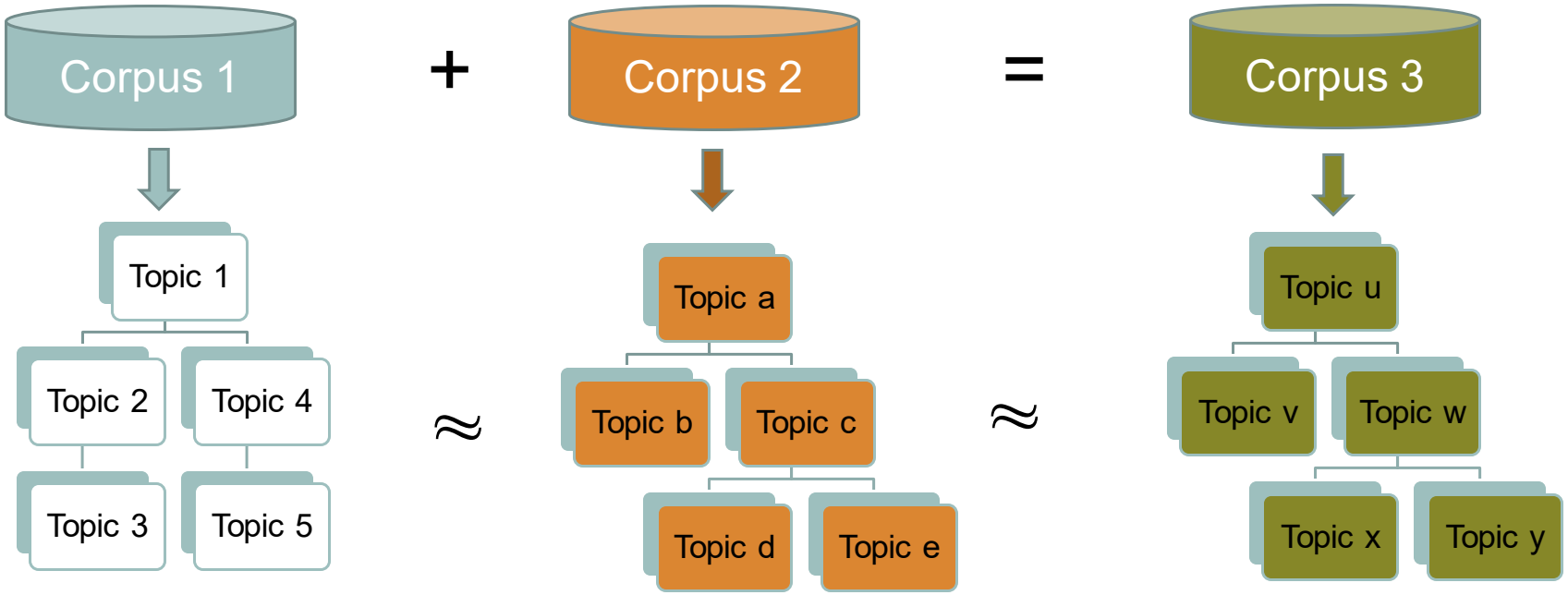
Input

- **Corpus** :
  - Bias in time and contents
  - Noise
  - Vocabulary
- **Models :**
  - LDA, graph, Word2Vec, frequent item set …
  - Cosine, Jaccard, …
- **Hyperparameters :**
  - Bounds, priors, weights, …

Output

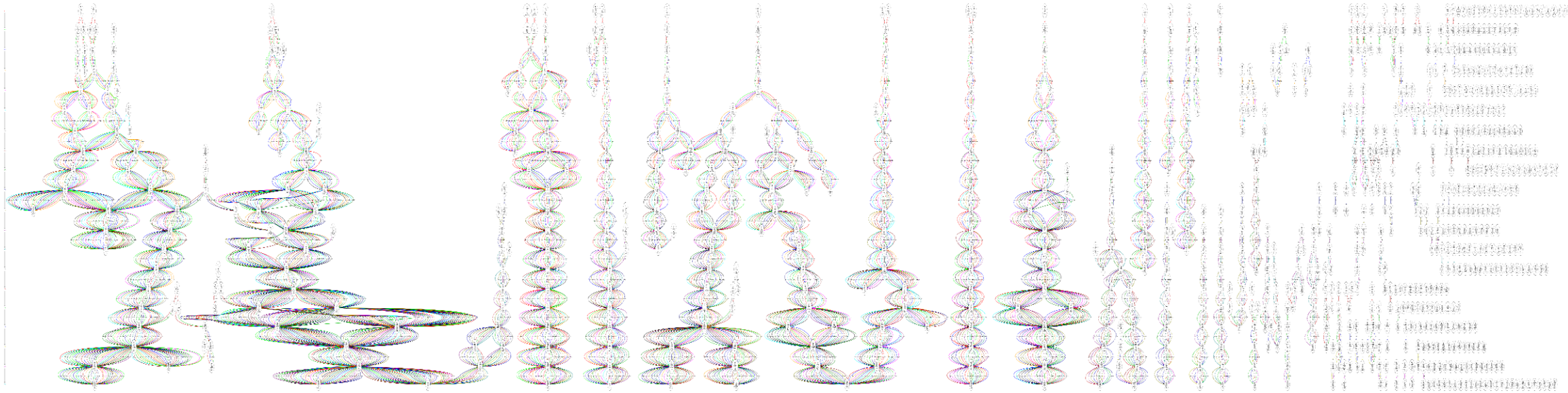
- Phylomemy **size** and **complexity**
- Interpretation & **explanation** semantics
- **Quality** measures
  - Relevance
  - Completeness
  - Diversity, …

Precision vs. Generalization

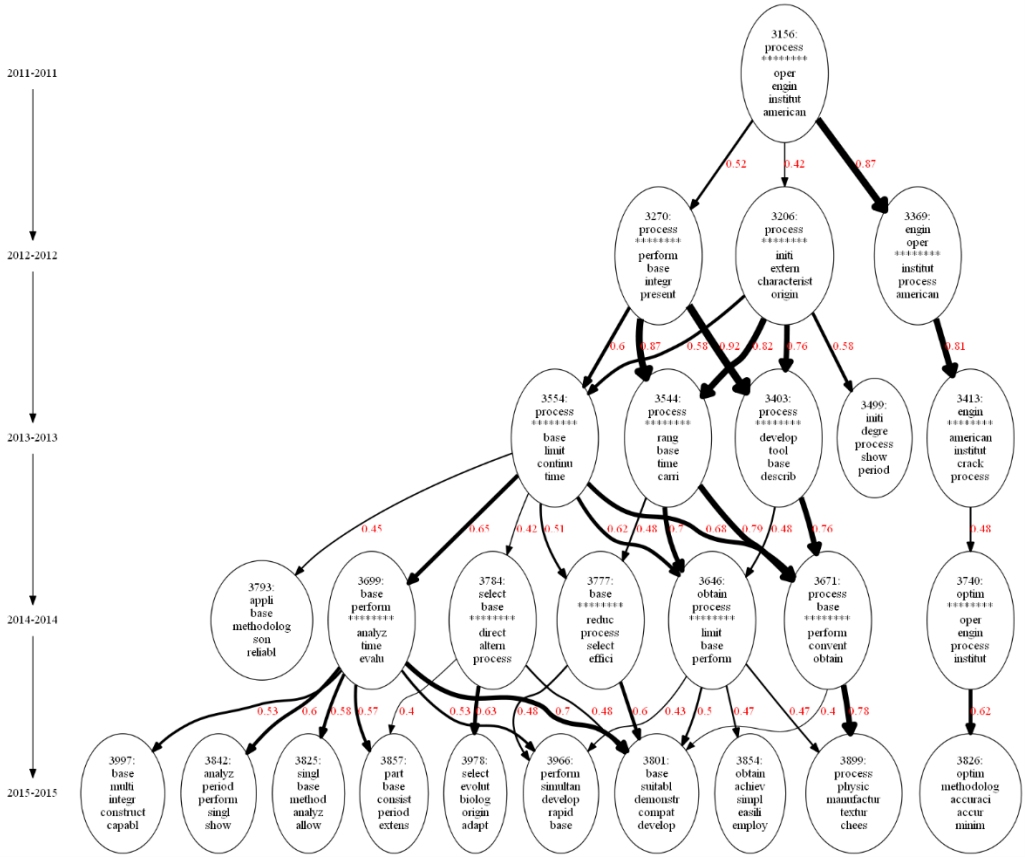


- Similar to Bias-Variance tradeoff (overfitting, underfitting)

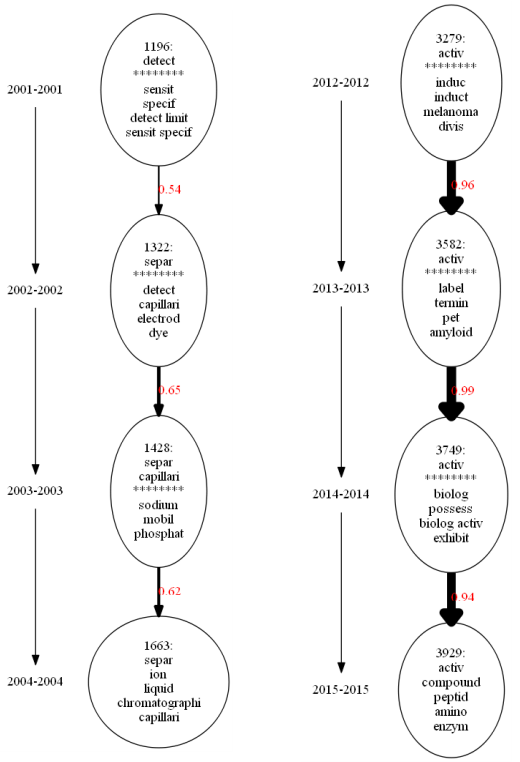
Complexity and heterogeneity



Evolution and Relevance



Which phylomemy is more « interesting » ?



High evolution

Low evolution

Observations and Solutions

Observations :

- Expert knowledge, interaction and validation is crucial
- Computation + visualisation is not sufficient

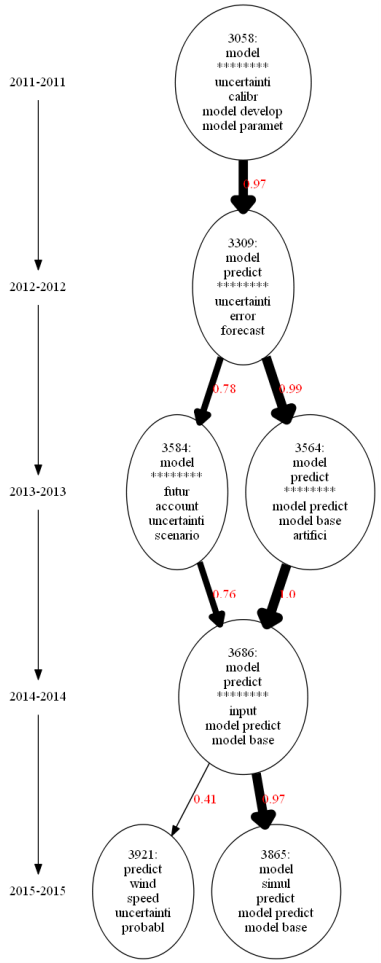
Solutions :

- Interactive exploration
- Iterative / incremental computation
- Open environment
- Assisted parametrization
- Explainable / interpretable results
- High-level filtering (query languages)

Pattern queries

get phylomemy about « model » where depth >= 4 and evolution >= 0.4 and complexity <= 0.5

id	G1Depth	G1Subnodes	G1MeanSimBeta	G1MeanSimAlpha (evolution)	G1IAO	G1IAI
345	18	47	0.709097459	0.126692764	0.642857143	0.821428571
713	16	41	0.653450044	0.049589811	0.6	0.8
942	15	40	0.651348349	0.050429224	0.591836735	0.795918367
1019	14	38	0.654102038	0.05895758	0.595744681	0.787234043
721	16	36	0.690647283	0.333010948	0.804878049	0.853658537
191	19	48	0.706519545	0.067389134	0.649122807	0.824561404
1705	11	36	0.683256752	0.251702896	0.861111111	0.972222222
1079	14	33	0.787762166	0.47119101	0.666666667	0.761904762
474	17	42	0.659919881	0.057710512	0.607843137	0.803921569
1668	11	31	0.62348373	0.119896616	0.555555556	0.833333333
1391	13	30	0.652026062	0.163769259	0.552631579	0.763157895
1402	12	29	0.653282242	0.350031431	0.540540541	0.756756757
592	17	38	0.70553714	0.378527711	0.755555556	0.822222222
1169	14	28	0.724692689	0.047291185	0.558823529	0.794117647
1515	12	28	0.684996182	0.399962117	0.787878788	0.818181818
1225	13	36	0.659921765	0.3746721	0.6	0.777777778
1354	13	26	0.729546685	0.19116916	0.628571429	0.714285714
1789	11	26	0.649534403	0.37703352	0.548387097	0.806451613
1297	13	25	0.694691714	0.050539332	0.615384615	0.923076923



EPIQUE

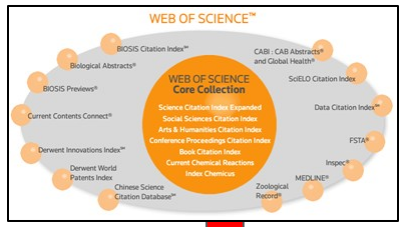
Towards quantitative epistemology – reconstruct the evolution of science at a large scale

LIP6 SU : Laboratoire d'Informatique de Paris 6, Sorbonne-Université (UPMC) (coordonnateur)
IRISA Rennes 1 : Institut de Recherche en Informatique et Systèmes Aléatoires
ISC-PIF : Institut des Systèmes Complexes de Paris - Ile de France
IHPST : Institut d'Histoire et de Philosophie des Sciences et des Techniques

Objectives

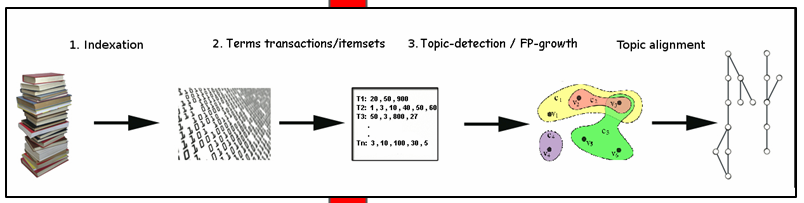
- Analyze and understand the evolution of science present in scientific archives
- *Scalable* methods for the detection and alignment of topics
- *Interactive* workflows which dynamically adapt to the distribution and evolution of data

Scientific corus

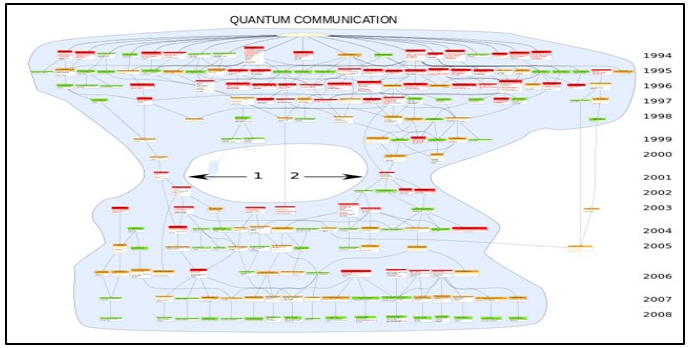


Web of Science
arXiv
DBLP
Medline …

Workflow

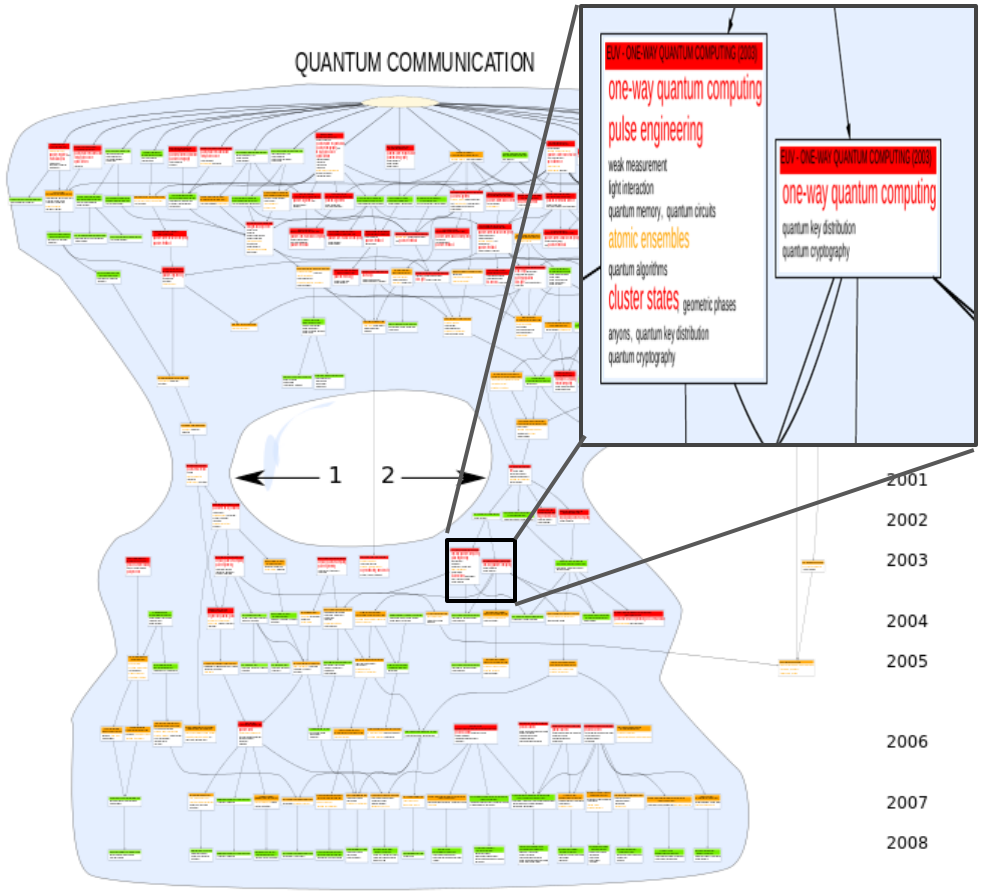


Phylomemies



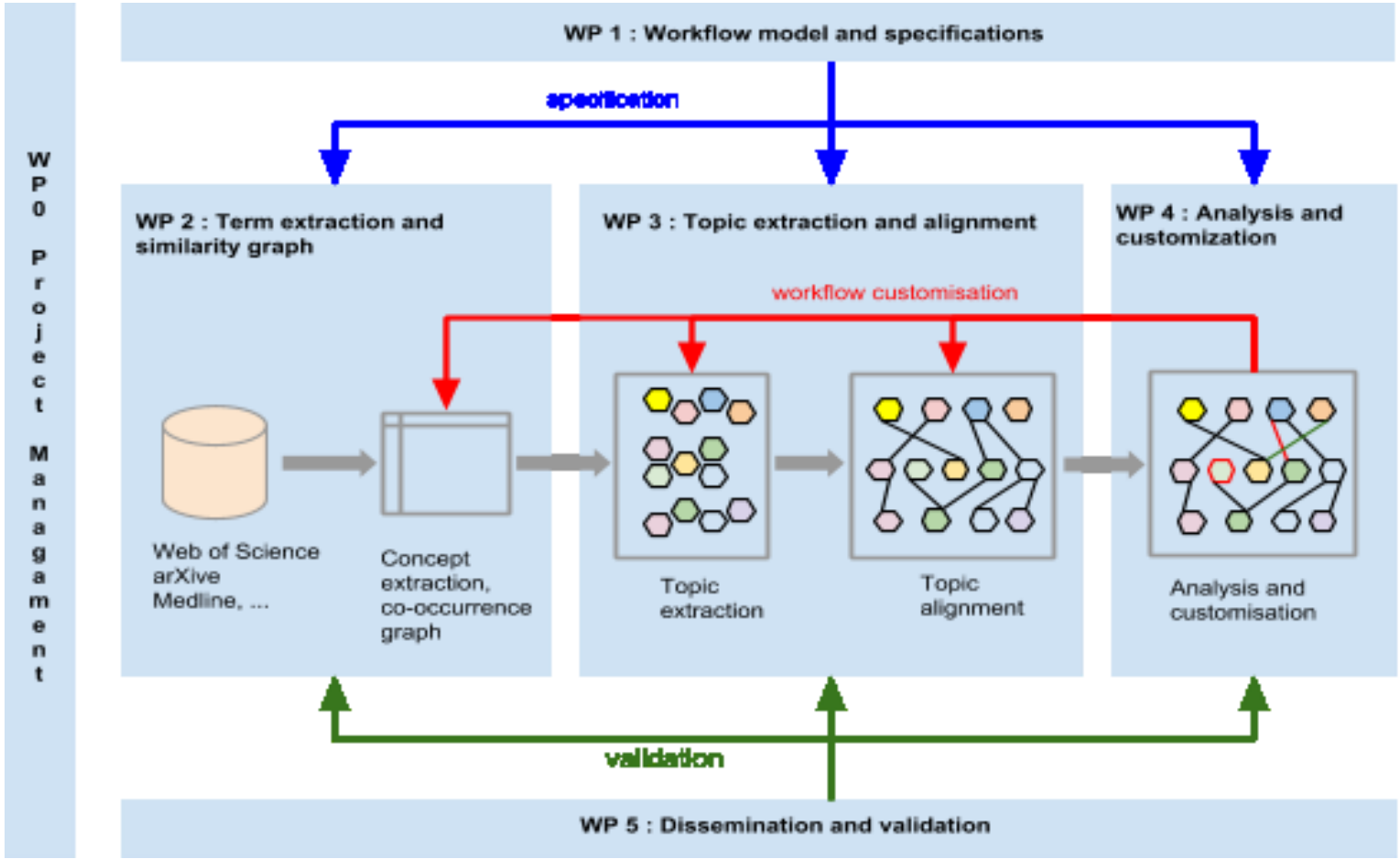
Applications

- **Validation of theories about the evolution of science**
- Cross-fertilisation of science and techniques
- Systematic scientific reviews and state of the art
- Detection of emergent scientific domains
- Innovation index and potential of technological transfert
- Classification systems for document archives
- Teaching and education (sceince maps)
- Vulgarization of science



Thank you for your attention

Project organisation

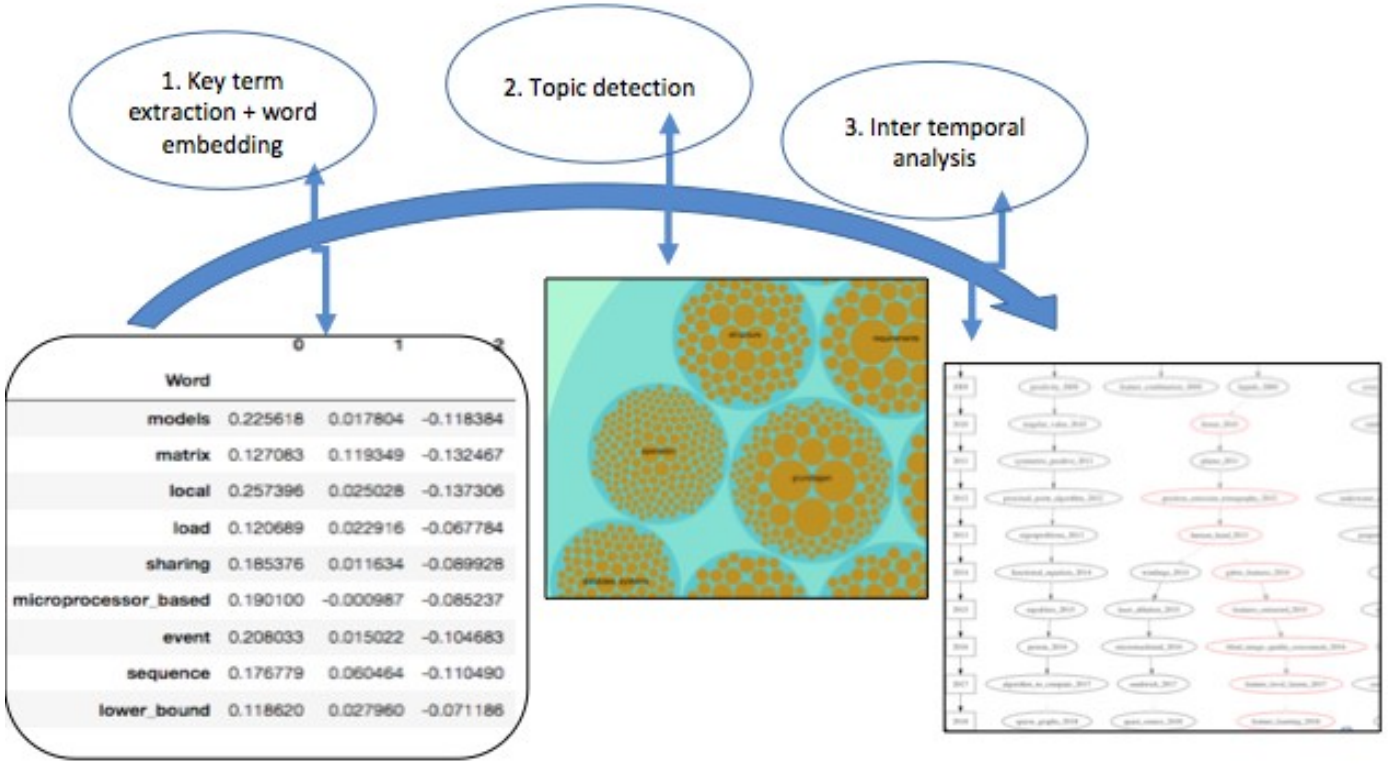


phase 1

phase 2

phase 3

Word2Vec + clustering + temporal analysis



Expected results

- Models of the dynamics of science in text corpora
  - Comparison, integration and validation of methods for unsupervised learning (text, graph) and databases for the representation and extraction of the evolution of scientific corpus
- Maps of the evolution of science
  - Visualization, validation and improvement of hypotheses on the evolution of scientific fields
  - several case study
- EPIQUE platform
  - Tools for reconstruction and the exploration of the dynamics of multi-scale in major scientific corpora
  - Integration in the open platform Gargantext of the ISC - PIF

Bias-Variance Tradeoff

Bias error :

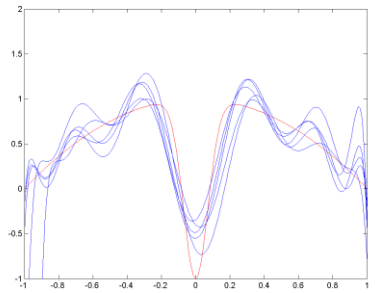
- Difference between **prediction** and **correct value**
- Low bias : complex model, *overfitting* (noise)

Variance error:

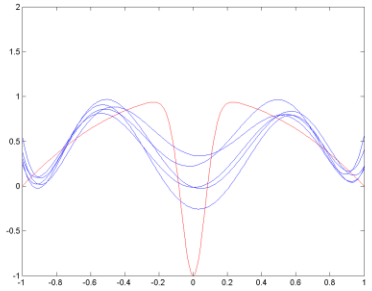
- **Variablity of a model prediction** for a given data point (low variance = generalization)
- Low variance : simple model, *underfitting*

Global error :

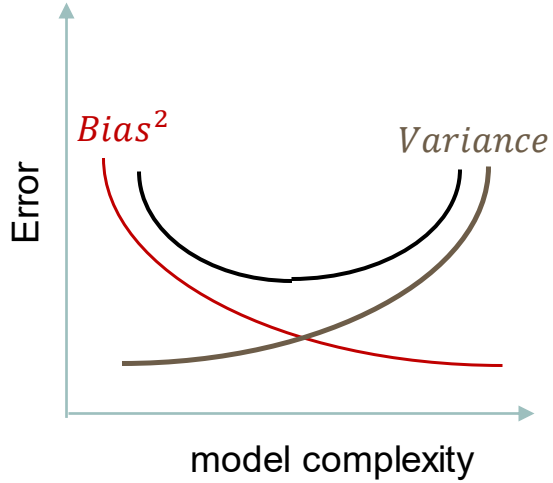
- $Bias^2 + Variance + Irreducible\ Error$



low bias / high variance



high bias / low variance



Validation of theories

Scientific Activity

- Definition of concepts, theories, hypotheses, data (observations, experiences)
- Formulate theories on the basis of the data
- Do experiments that can change theories

Rival theories of the evolution of science

- Bayesian theories change conjectures and refutations (Popper)
- Darwinian epistemology (Hull): social epistem. + selection of theories cumulative induction
- Research program (Lakatos)
- Paradigm shift (Kuhn)

Objective: Understand the processes underlying the dynamics of science from scientific corpus?

Science Evolution Signatures

- Identify the signatures of the evolutionary process of science:
  - the signature of a Darwinian selection of theories
  - the signature of a paradigm shift
  - the signature of selection in molecular evolution
- Discover and validate these signatures in the scientific corpus :
  - phylomemetic networks : with reference to McShea (1994, 1999) on the signatures of natural selection in the evolution of life.

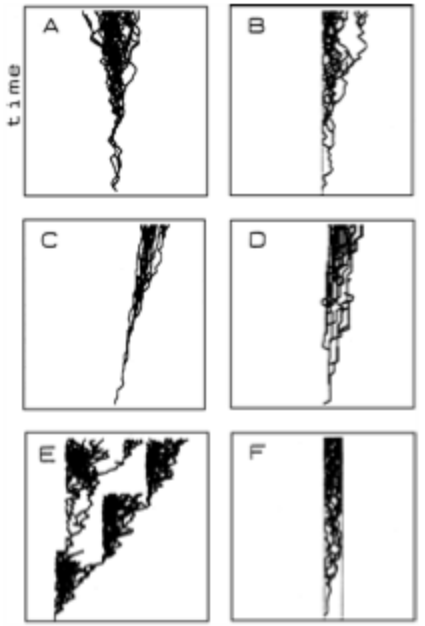


Figure 2 Output of a computer model for simulating the diversification of a group. The vertical axis is time and the horizontal axis is the state variable. In each figure, a group begins as a single species. At every time step, each lineage may increase (move right) or decrease (move left) in state space, speciate, and/or become extinct, each according to fixed probabilities. Boundaries (vertical lines in *B* and *F*) are "reflecting," meaning changes that would cross lineages to cross them are nullified. Biases are introduced (*C* and *D*) by setting the probability of moving right higher than that for moving left. (See 63 for further details.) (*A*) No trend—no boundary, no bias. (*B*) Passive trend—lower boundary, no bias. (*C*) Driven trend—no boundary, strong bias. (*D*) Weakly driven—no boundary, weak bias. (*E*) Driven, at the large scale (in that an increasing bias is present in the origin of groups, although change within groups is passive)—no boundary, strong bias. (*F*) No trend—upper and lower boundary, no bias.