Module IV: Knowledge Graph Fundamentals and Construction

1:00 pm - 2:05 pm

Module 4 Overview

KG Fundamentals and Construction

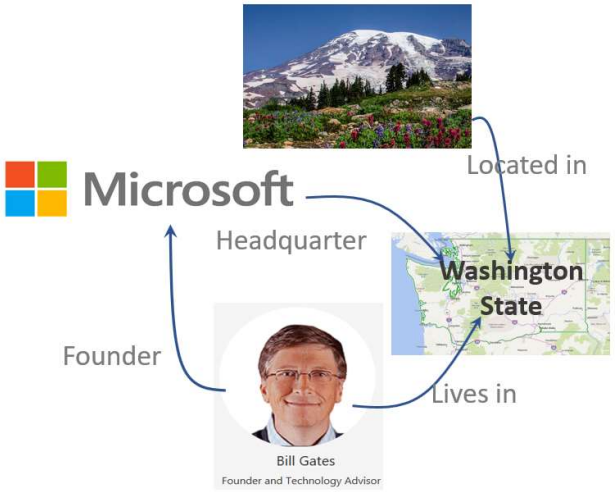
A brief history of Knowledge Graph (KG)

KG representation and examples

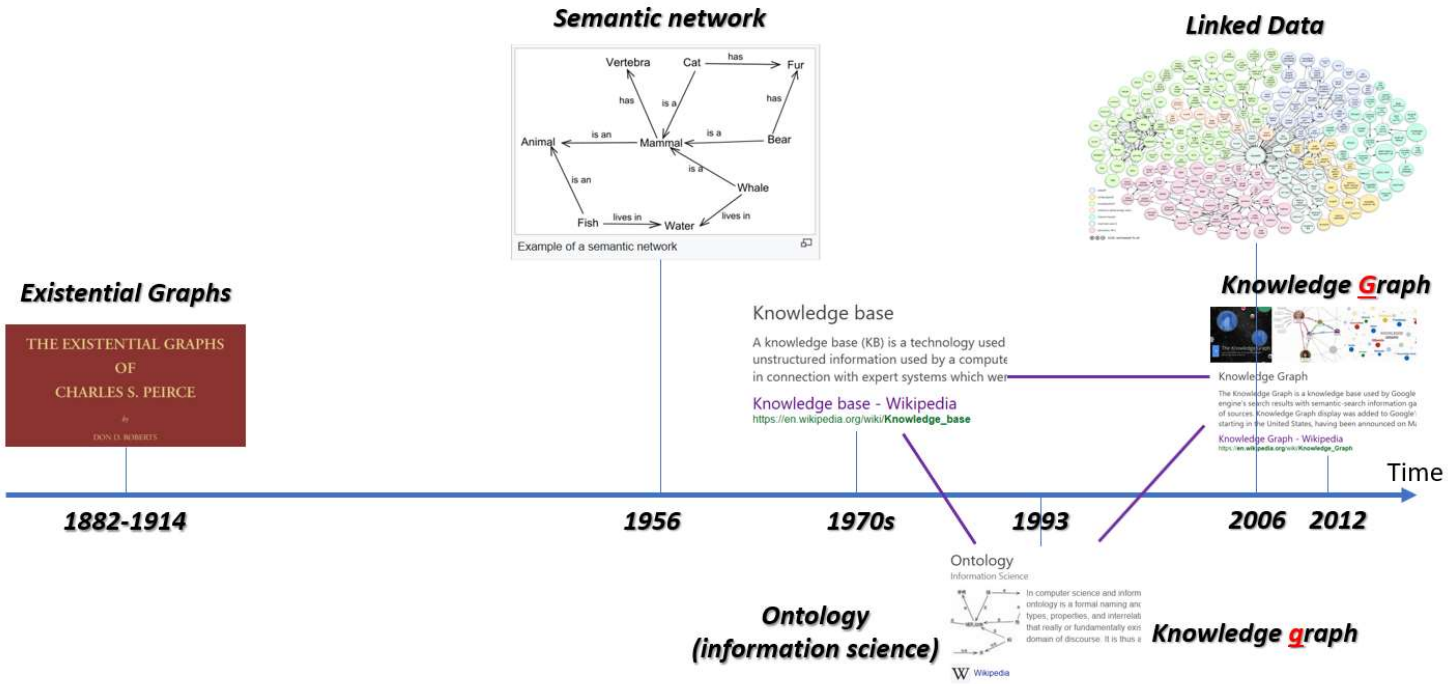
KG Construction

- General overview
- NLP techniques
- MAG case study

Lab 4 – Demo: enrich concepts



A Brief History of Knowledge Graph



Knowledge Graph Representation

- Graph
  - Node
    - Attribute1
    - Attribute2
    - Attribute3
    - …
  - Edge

    (Node1, Node2, weights)

Homogeneous	vs	Heterogeneous
Node: 1 table Edge: 1 table		Node: multiple tables Edge: multiple tables

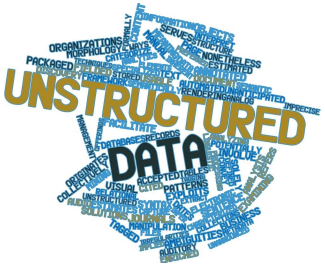
- (***S***ubject, ***P***redicate, ***O***bject)
  - Each ***node*** has a universal id (***S***)
  - Its ***attribute*** is represented as:

    (***S***, attributeName (***P***), attributeValue(***O***))
  - An ***edge*** connected two nodes (e.g. ***S1***, ***S2***):

    (***S1***, relationName (***P***), ***S2***(***O***))

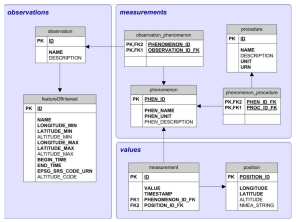
Homogeneous	vs	Heterogeneous
Node + Edge: SINGLE table		

Knowledge Graph Construction



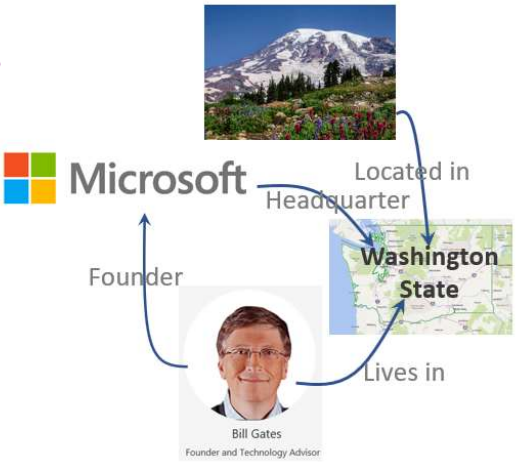
Unstructured Documents

Natural language processing (NLP)



Existing Relational Databases

Data pipeline processing

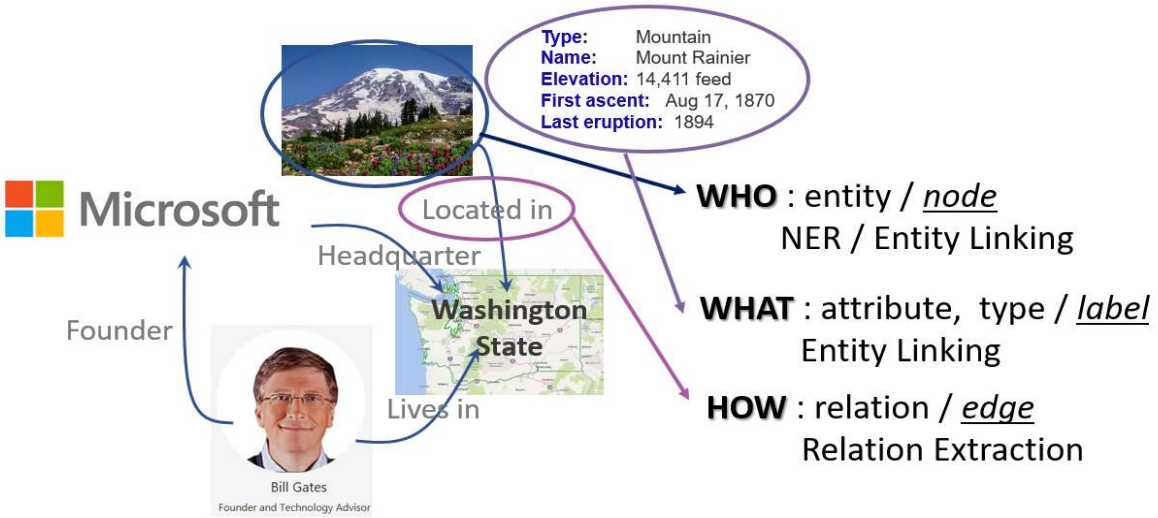


Knowledge in the ***graph*** form

Human Common Sense

Manual efforts

- Sentence Level
  - Part-of-speech (PoS) Tagging
  - Named entity recognition (NER)
  - Dependency Parsing
- Document Level
  - Coreference resolution
  - Topic model
  - Classification
- Information Extraction
  - Entity resolution
  - Entity linking
  - Relation extraction

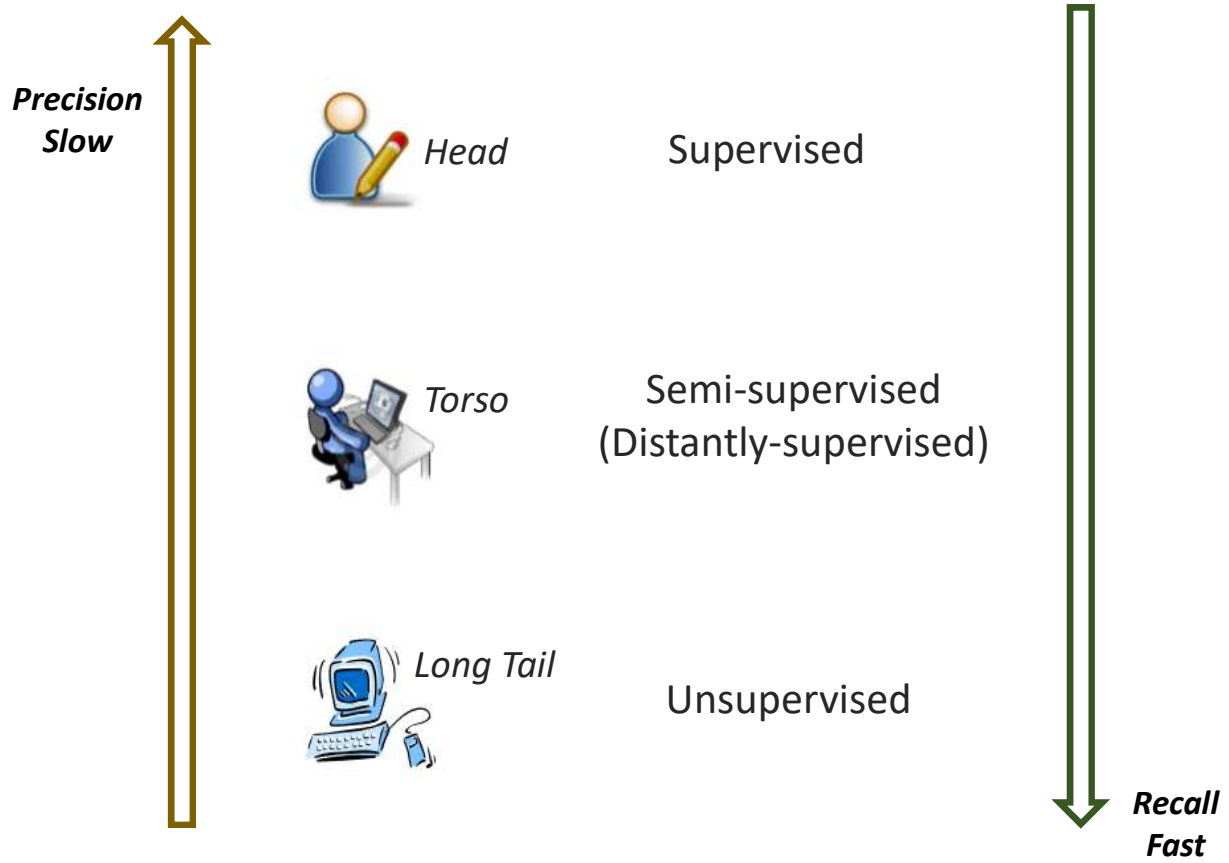


NLP Techniques

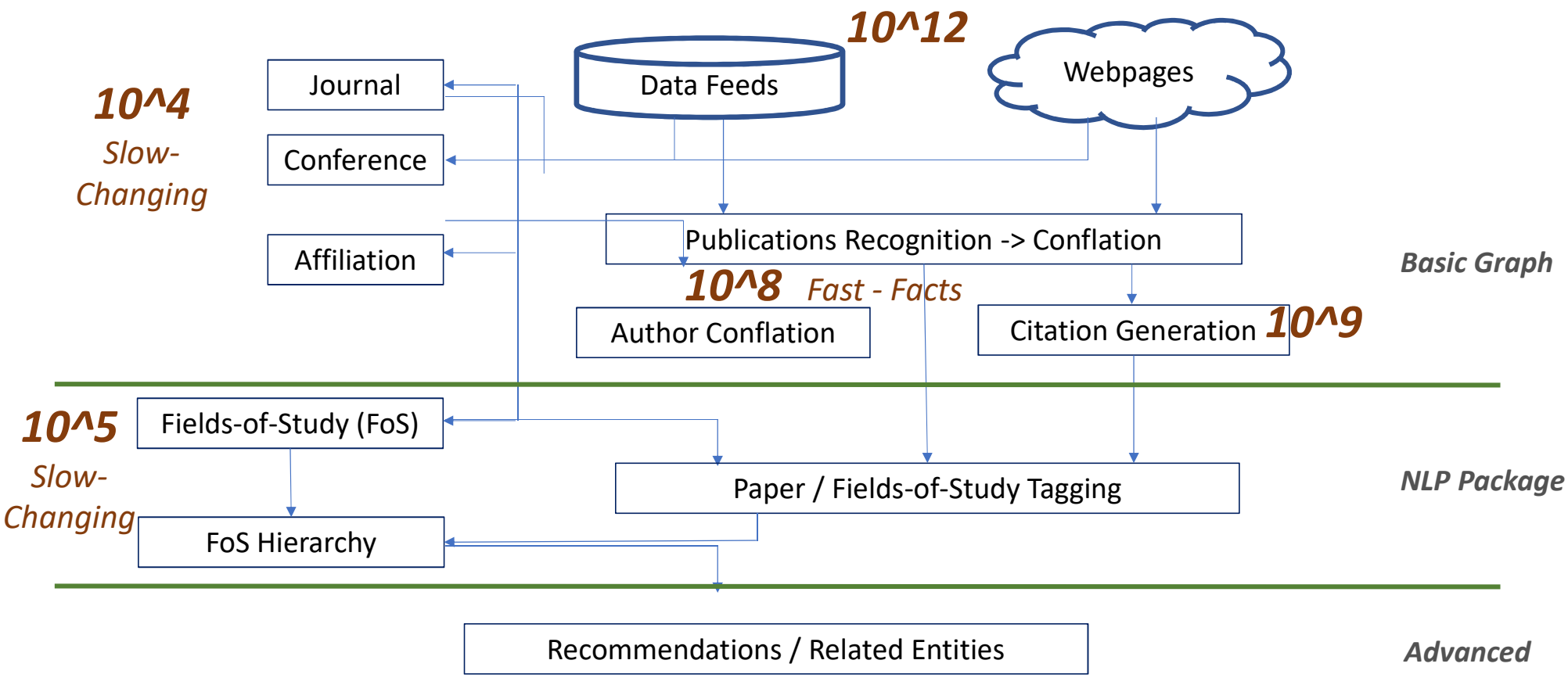
for Knowledge Graph Construction – At a glance

Knowledge Graph Construction

- Challenges:
  - Incomplete
  - Inconsistent
  - Ambiguous



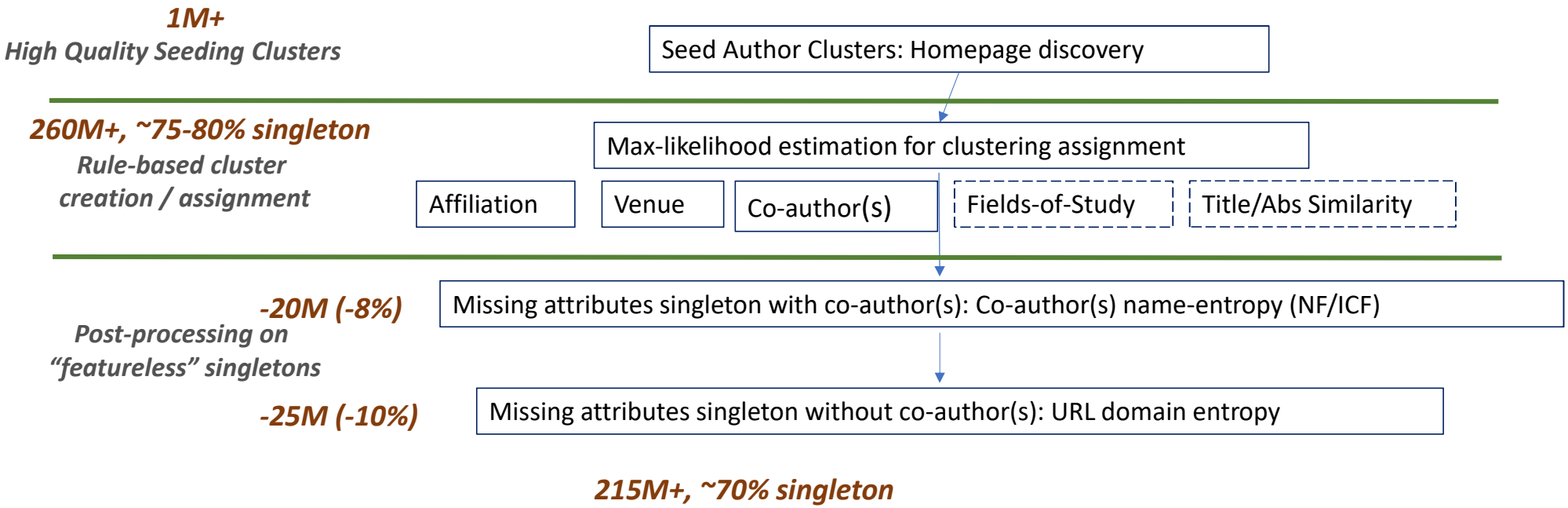
Microsoft Academic Graph (MAG) Construction



MAG Construction

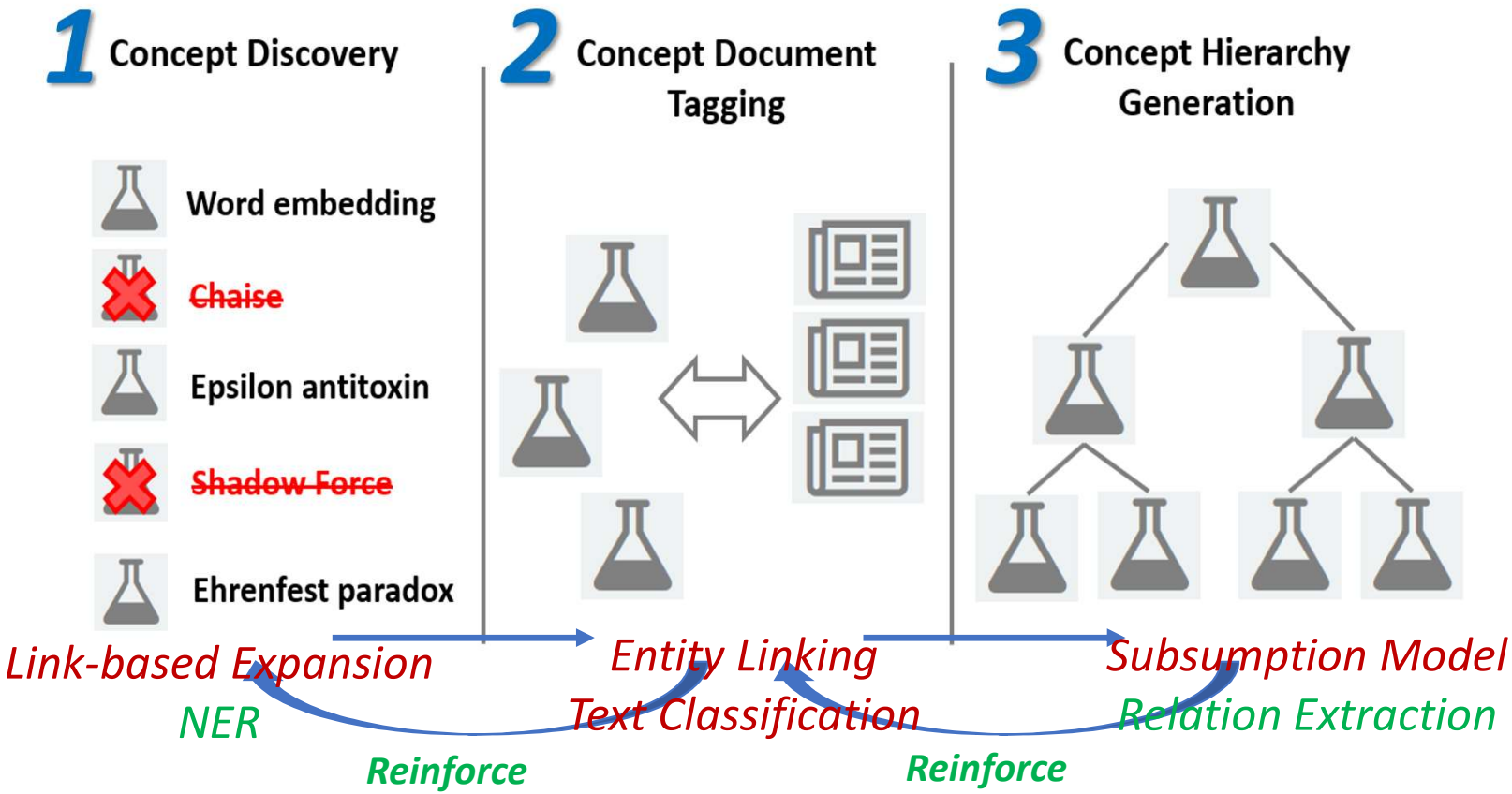
Author Recognition/conflation

Guiding principles: prefer under-conflation than over-conflation (<0.1%)



MAG Construction

Concept Recognition/Tagging/Hierarchy



MAG Construction

Concept / Publication representation

	Word	Concept	Publication
Language Features			
Discrete Space	1-hot vector	Bag-of-Words (Desc.)	Bag-of-Words (Title + Abstract)
Continuous Space	Word Vector *	Concept Embeddings **	Bag-of-Concepts (Title + Abstract)
Structural Features			
Discrete Space		Related Concepts	Citation / Venue [/Author]
Continuous Space		Heterogenous Graph Embedding	Homogenous Graph Embedding

***Mixture Model** to represent Concepts & Publications (or any documents)*

*Word vector are train on Skip-gram model with 13B tokens of 130M English publications, vocab size: ~2M
Concept embeddings - average of the word vector in concept description. (*Description*** is important!)

MAG Construction

Concept Hierarchy Results

L5	L4	L3	L2	L1	L0
Convolutional Deep Belief Networks	Deep belief network	Deep learning	Artificial neural network	Machine learning	Computer Science
(Methionine synthase) reductase	Methionine synthase	Methionine	Amino acid	Biochemistry / Molecular biology	Chemistry / Biology
(glycogen-synthase-D) phosphatase	Phosphorylase kinase	Glycogen synthase	Glycogen	Biochemistry	Chemistry
	Fréchet distribution	Generalized extreme value distribution	Extreme value theory	Statistics	Mathematics
Hermite's problem	Hermite spline	Spline interpolation	Interpolation	Mathematical analysis	Mathematics

***Completely Data-Driven** (L2-L5)*

6-Level Hierarchy with 660K+ Concepts

Lab 4: Enrich Concepts

- Demo: Named Entity Recognition (NER) for new concept discovery

Model	Description	CONLL 2003 F1
TagLM (Peters+, 2017)	LSTM BiLM in BLSTM Tagger	91.93
ELMo (Peters+, 2018)	ELMo in BLSTM	92.22
BERT-Base (Devlin+, 2019)	Transformer bidi LM + fine tune	92.4
CVT Clark	Cross-view training + multitask learn	92.61
BERT-Large (Devlin+, 2019)	Transformer bidi LM + fine tune	92.8
Flair	Character-level language model	93.09