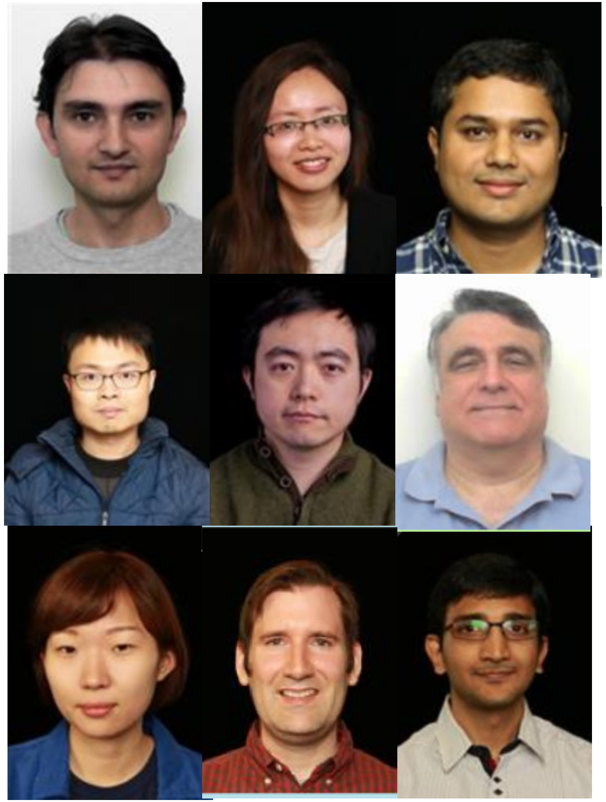
Data Integration and Machine Learning: A Natural Synergy

Xin Luna Dong @ Amazon.com
Theo Rekatsinas @ UW-Madison
Sigmod 2018

Acknowledgement



product graph

snorkel

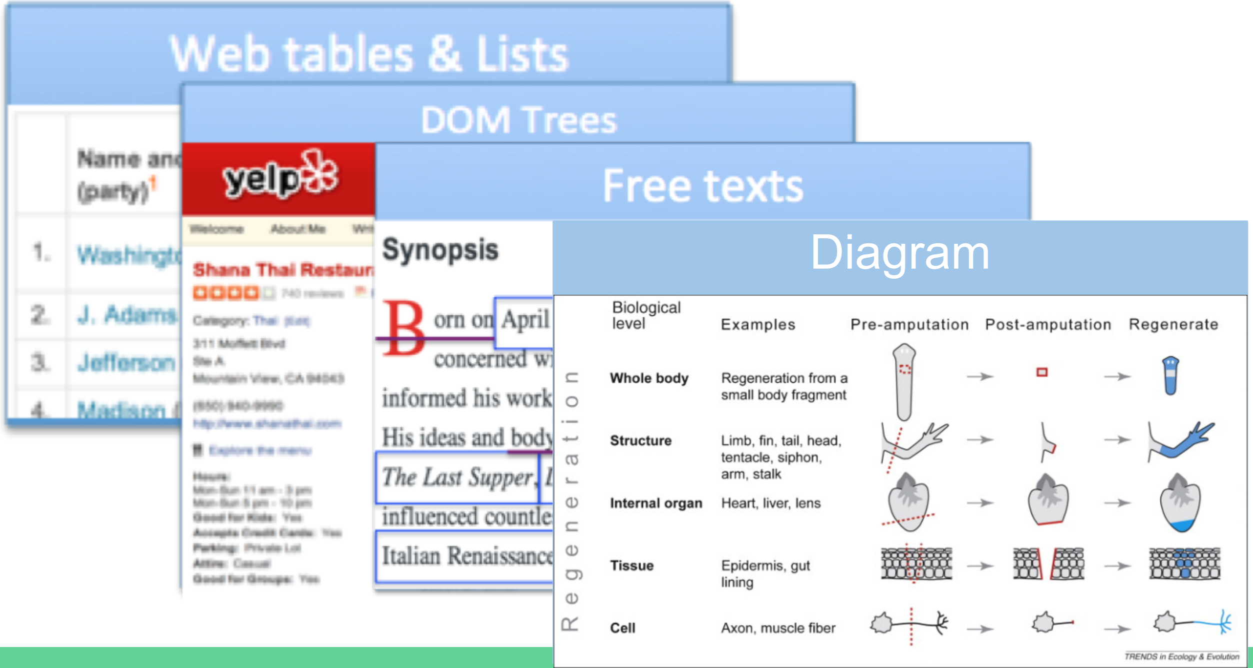
What is Data Integration?

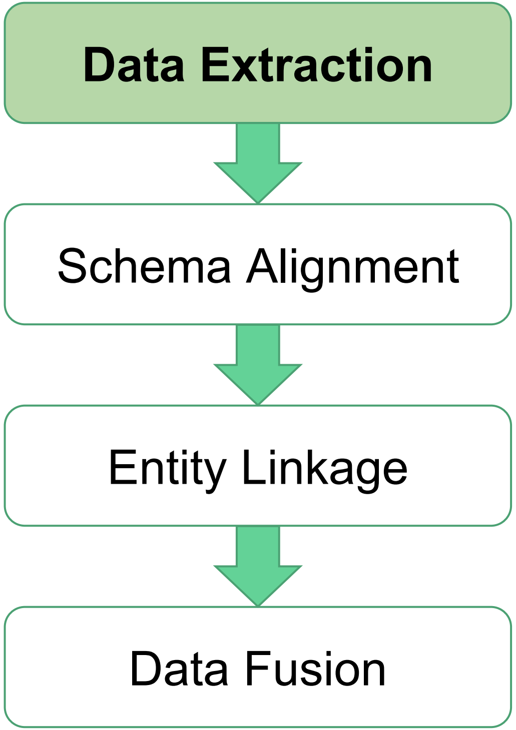
- **Data integration**: to provide unified access to data residing in multiple, autonomous data sources
    - **Data warehouse**: create a single store (materialized view) of data from different sources offline. Multi-billion dollar business.
    - **Virtual integration**: support query over a mediated schema by applying online query reformulation. E.g., Kayak.com.

- In the RDF world: different names for similar concepts
    - **Knowledge graph** is equivalent to a data warehouse. Has been widely used in Search and Voice
    - **Linked data** is equivalent to virtual integration

Why is Data Integration Hard?

- Heterogeneity everywhere
  - Different data formats



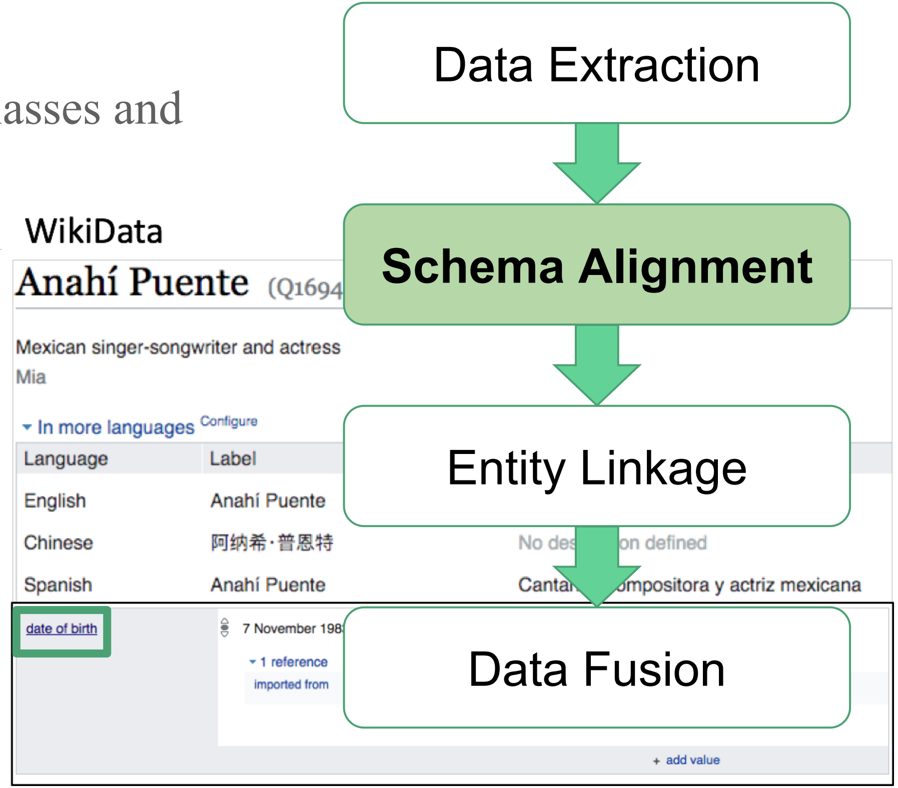


Why is Data Integration Hard?

- Heterogeneity everywhere
  - Different ways to express the same classes and attributes

IMDB

WikiData

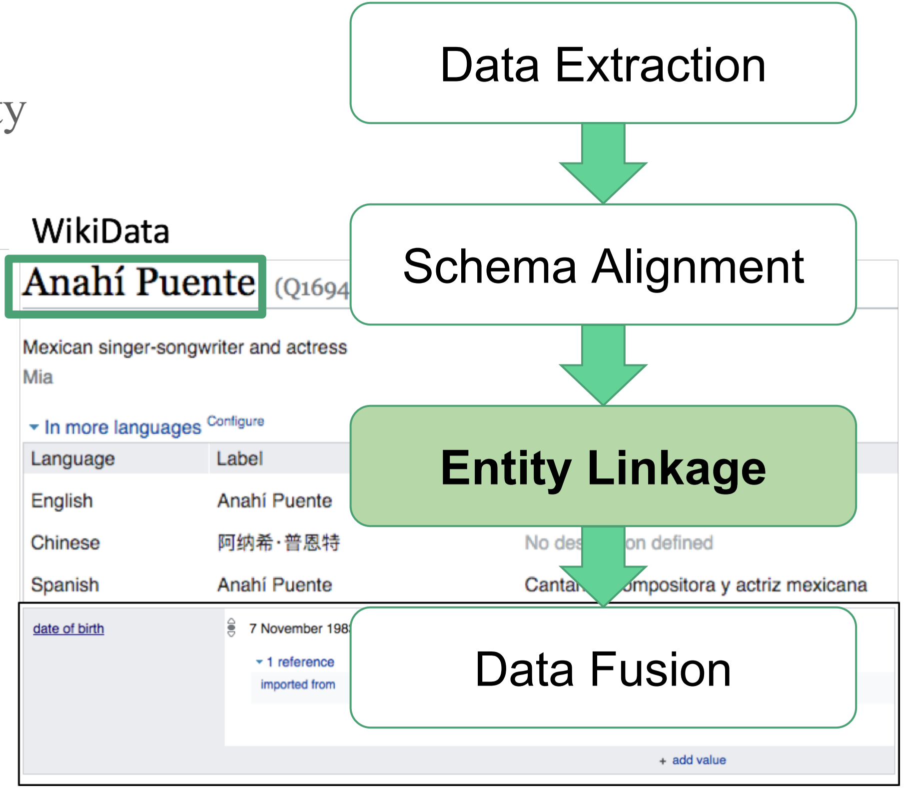


Why is Data Integration Hard?

- Heterogeneity everywhere
  - Different references to the same entity

IMDB

WikiData

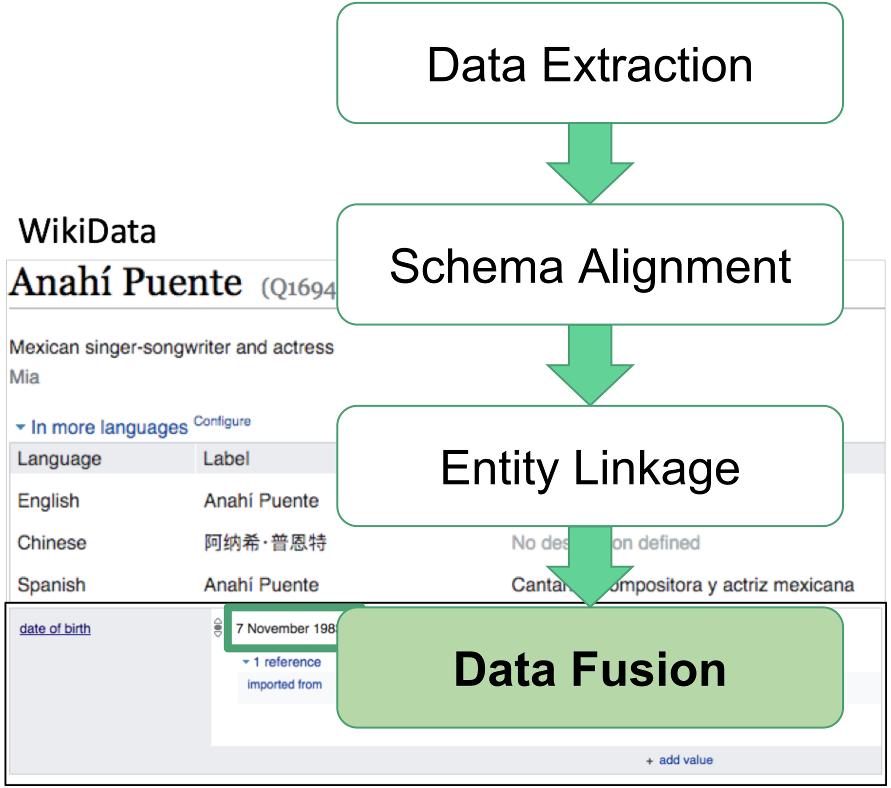


Why is Data Integration Hard?

- Heterogeneity everywhere
  - Conflicting values

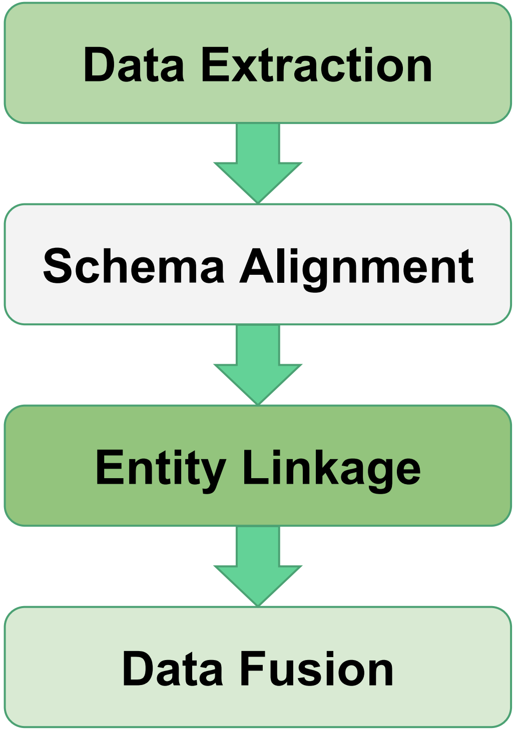
IMDB

WikiData



Importance from a Practitioner's Point of View

- Entity linkage is indispensable whenever integrating data from different sources
- Data extraction is important for integrating non-relational data
- Data fusion is necessary in presence of erroneous data
- Schema alignment is helpful when integrating relational data, but not affordable for manual work if we integrate many sources



What is Machine Learning?

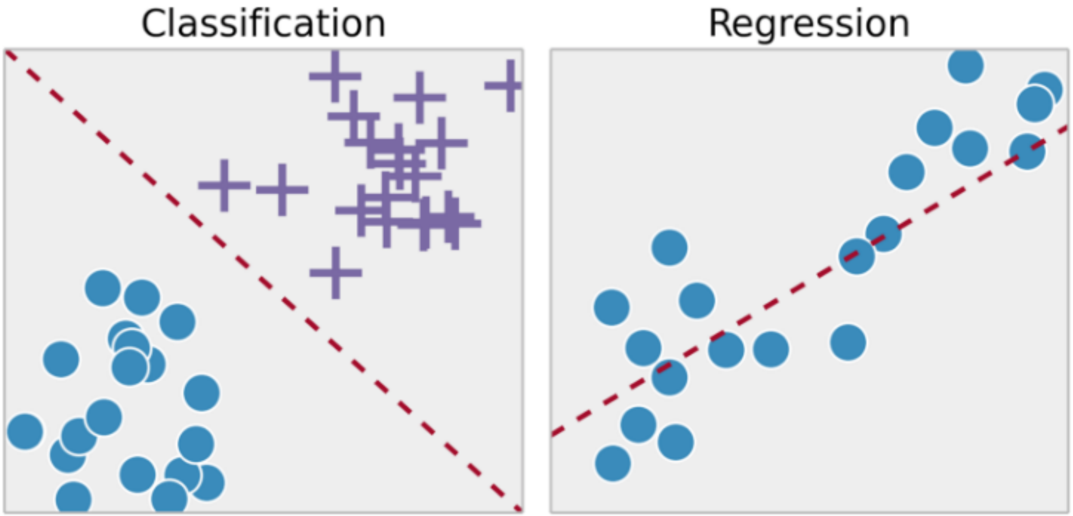
- **Machine learning:** teach computers to ***learn*** with data, not by programming

- **More Formal definition**
  A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, **improves with experience E**.

  -- Tom Mitchell

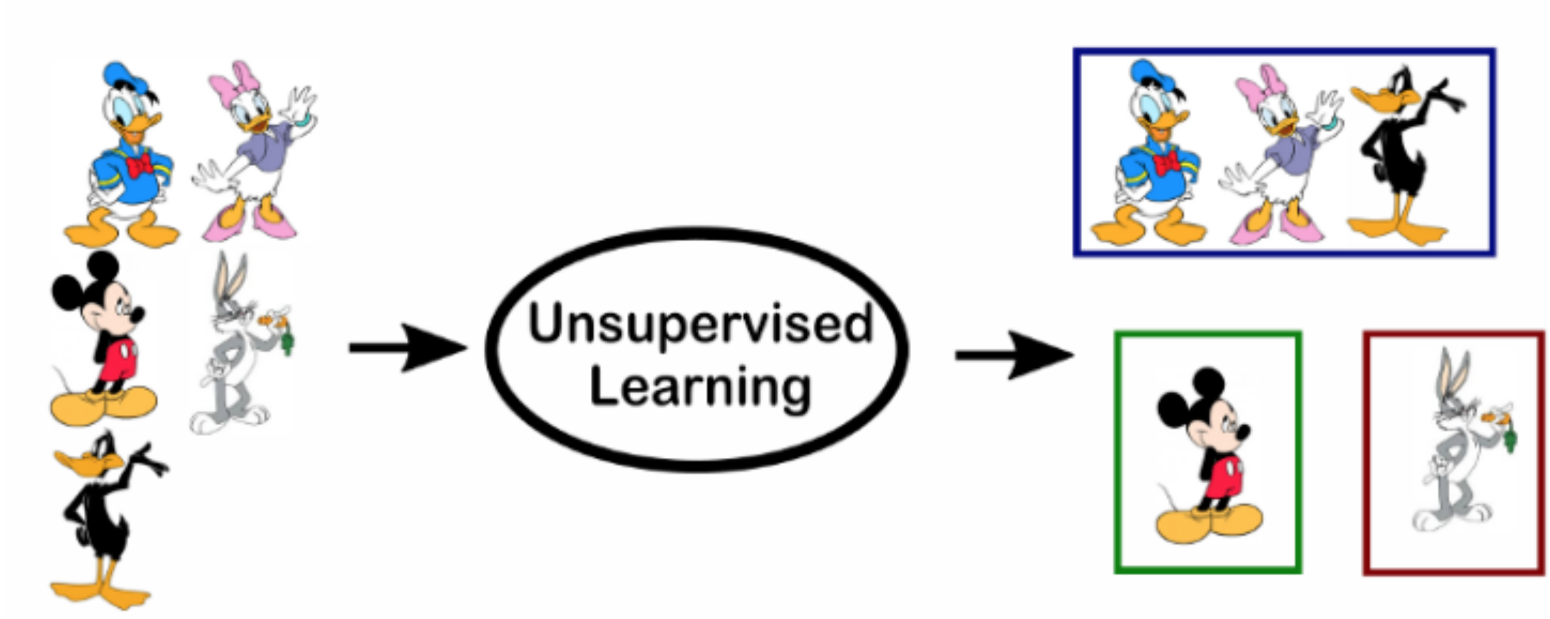
Two Main Types of Machine Learning

- Supervised learning: learn by examples



Two Main Types of Machine Learning

- Unsupervised learning: find structure w/o examples



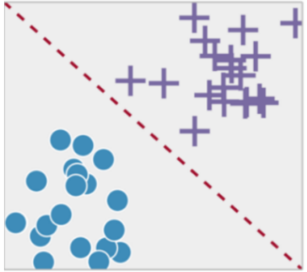
Two Main Types of Machine Learning

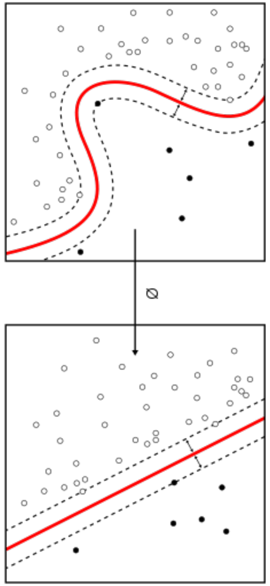
- Supervised learning: learn by examples
- Unsupervised learning: find structure w/o examples

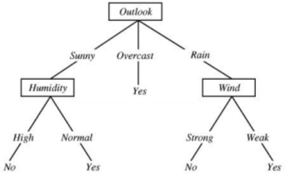
	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

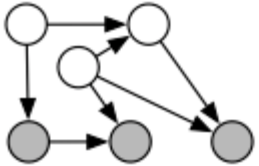
Techniques for Supervised ML

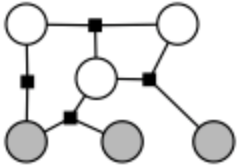
Hyperplanes	Kernel	Tree-based	Graphical Mdl	Logic Prog	Neural Netw
Linear/Logistic regression	SVM	Decision tree, Random forest	Bayes net, CRF	Pr soft logic, Markov logic net	ANN, RNN, CNN

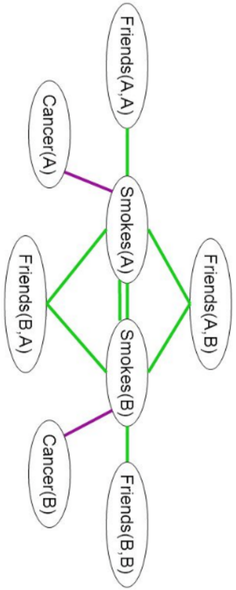


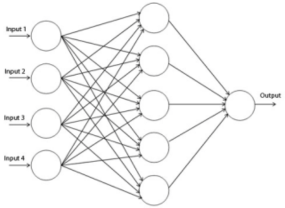


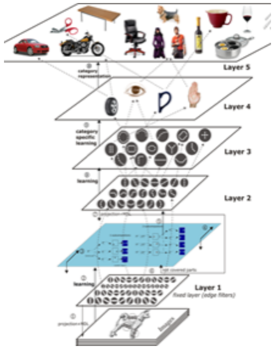












Key Lessons for ML [Domingos, 2012]

- Learning = Representation + Evaluation + Optimization
- **It's generalization that counts: generalize beyond training examples**
- Data alone is not enough: "no free lunch" theorem--No learner can beat random guessing over all possible functions to be learned
- Intuition fails in high dimensions: "curse of dimensionality"
- **More data beats a cleverer algorithm**: Google showed that after providing 300M images for DL image recognition, no flattening of the learning curve was observed.

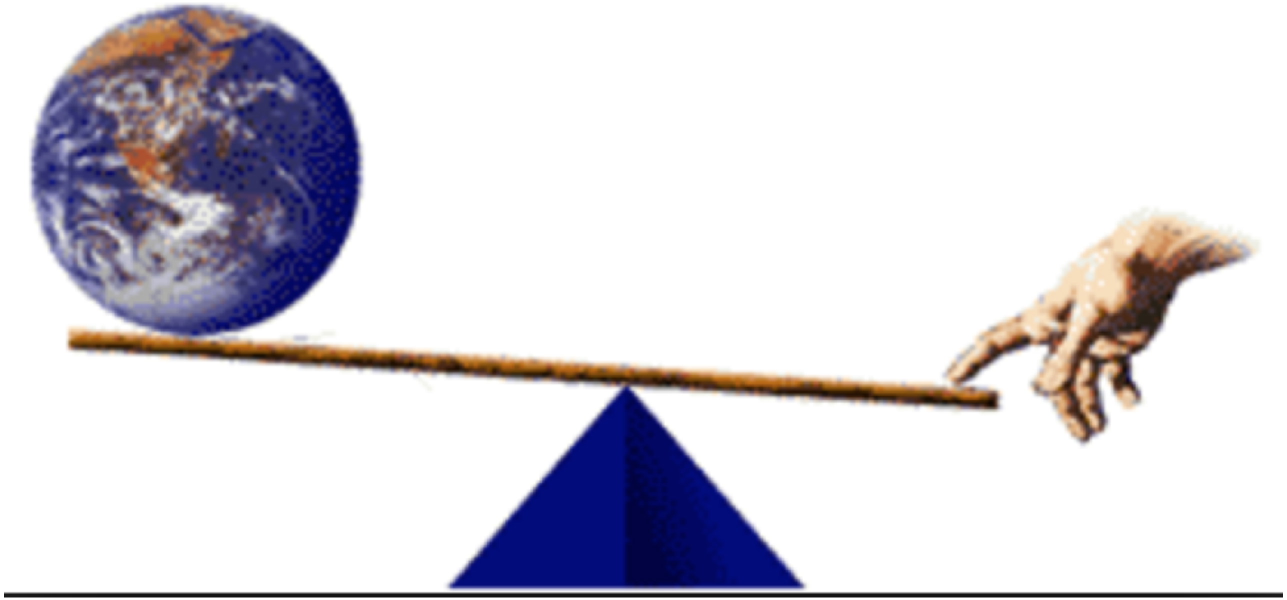
DI & ML as Synergy

- **ML for effective DI: AUTOMATION, AUTOMATION, AUTOMATION**
  - Automating DI tasks with training data
  - Better understanding of semantics by neural network

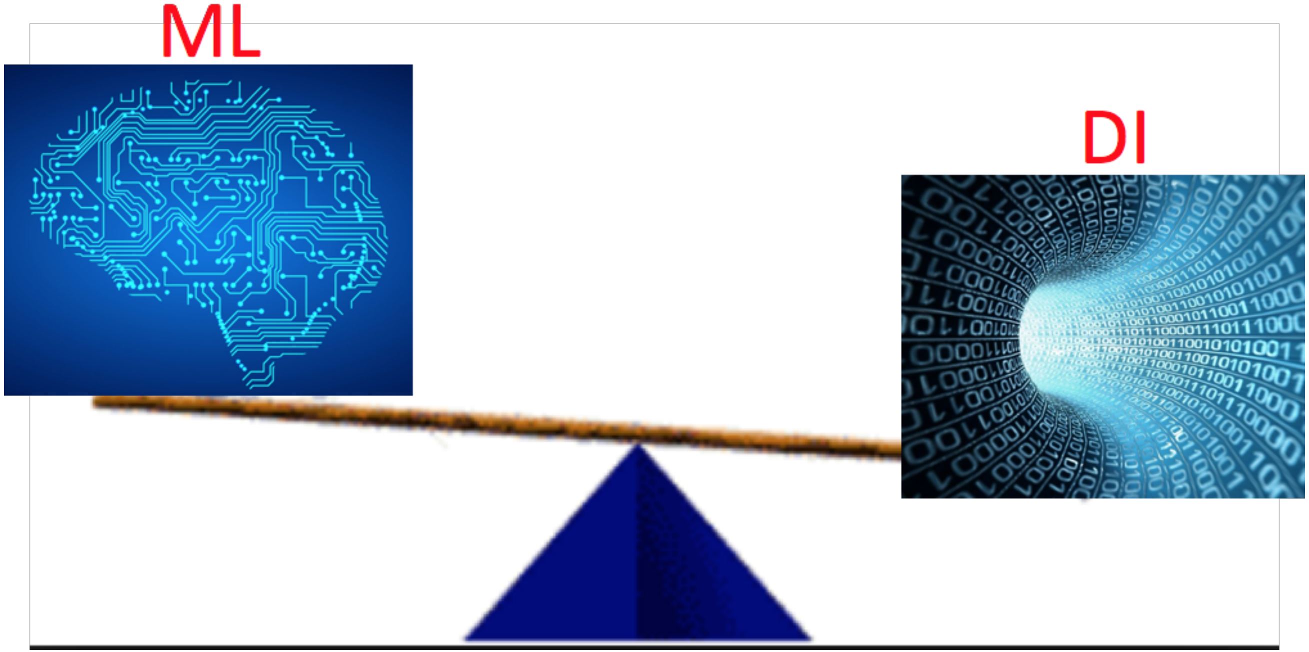
- **DI for effective ML: DATA, DATA, DATA**
  - Create large-scale training datasets from different sources
  - Cleaning of data used for training

Give me a Fulscrum, I will Move the Earth

-- Archimedes



Give me a DI funnel, I will Move ML



Many Systems Where DI & ML Leverage Each Other

NELL

MacroBase

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Magellan

HoloClean

snorkel

Dedupe.io

KNOWLEDGE VAULT

BigGorilla

tamr

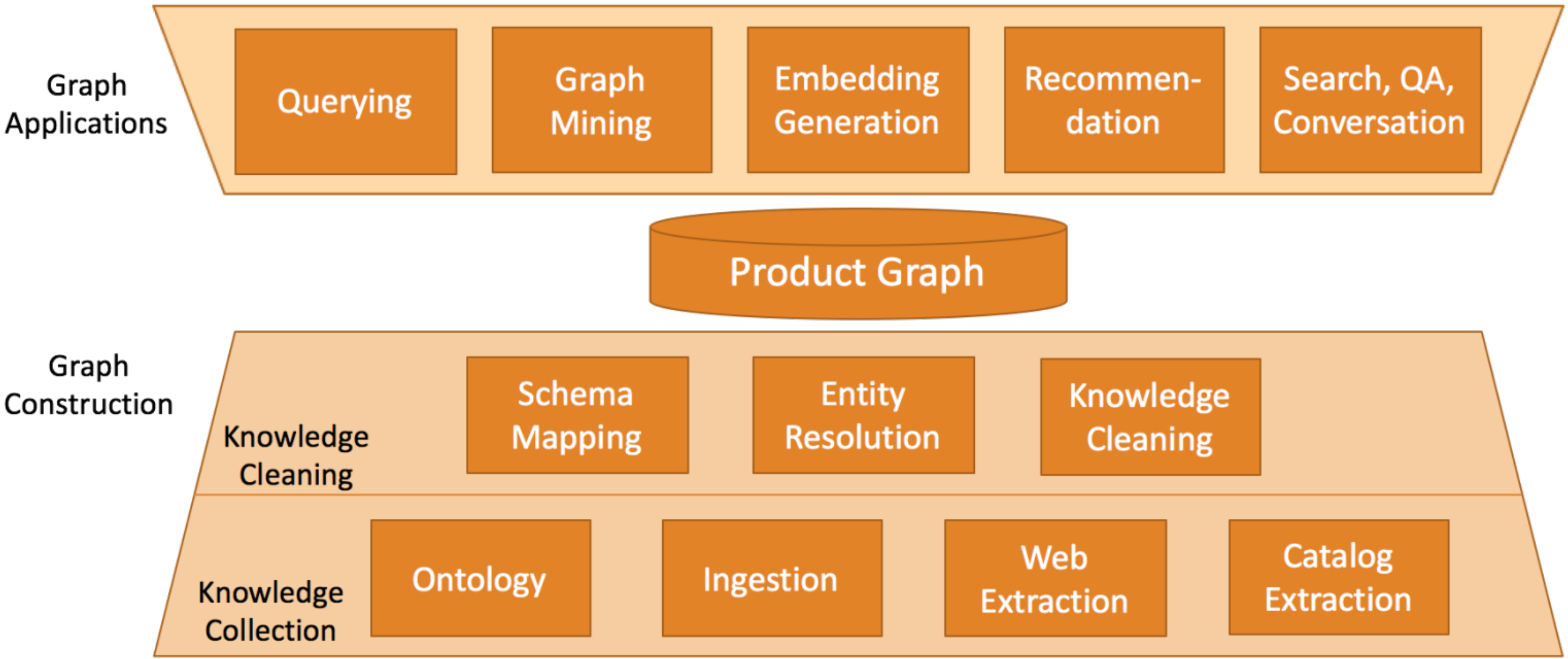
amperity

product graph

TRIFACTA

Increasing number of systems both in industry and academia.

Example System: Product Graph [Dong, KDD'18]



Goal of This Tutorial

- **NO-GOALS**
  - Present a comprehensive literature review for all topics we are covering

- **GOALS**
  - Present state-of-the-art for DI & ML synergy
  - Show how ML has been transforming DI and vice versa
  - Give some taste on which tool is working best for which tasks
  - Discuss what remains challenging