ENACTING DATA SCIENCE PIPELINES FOR EXPLORING GRAPHS

From Libraries to Studios

Genoveva.Vargas-Solar
LIG-LAFMIA, CNRS, France
genoveva.vargas@imag.fr

José-Luis Zechinelli-Martini, *UDLAP, Mexico*
Javier Espinosa*, LAFMIA-U. Lyon 3, France*

BBIGAP Workshop, Lyon, 25th August, 2020

DATIFICATION

Rendering into **data** aspects of the world that **have never been quantified**

Any individual can analyse huge amounts of data in short periods of time

- **Analytical knowledge:** most of the crucial algorithms are accessible
- Use rich data to make **evidence-based decisions** open to virtually **any person or company**

CONNECTED DATA

1 **Observations** structured as networks with their **own interconnection rules determined** by the **variables** characterising each observation

Connected Enterprise

Employees, Costumers & Partners

Internet Connected Things

Graph

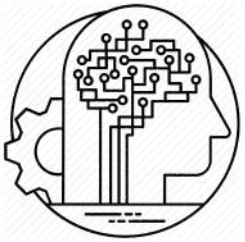
Mathematical Concept & operations
Data structures & algorithms

2

Digital Mesh

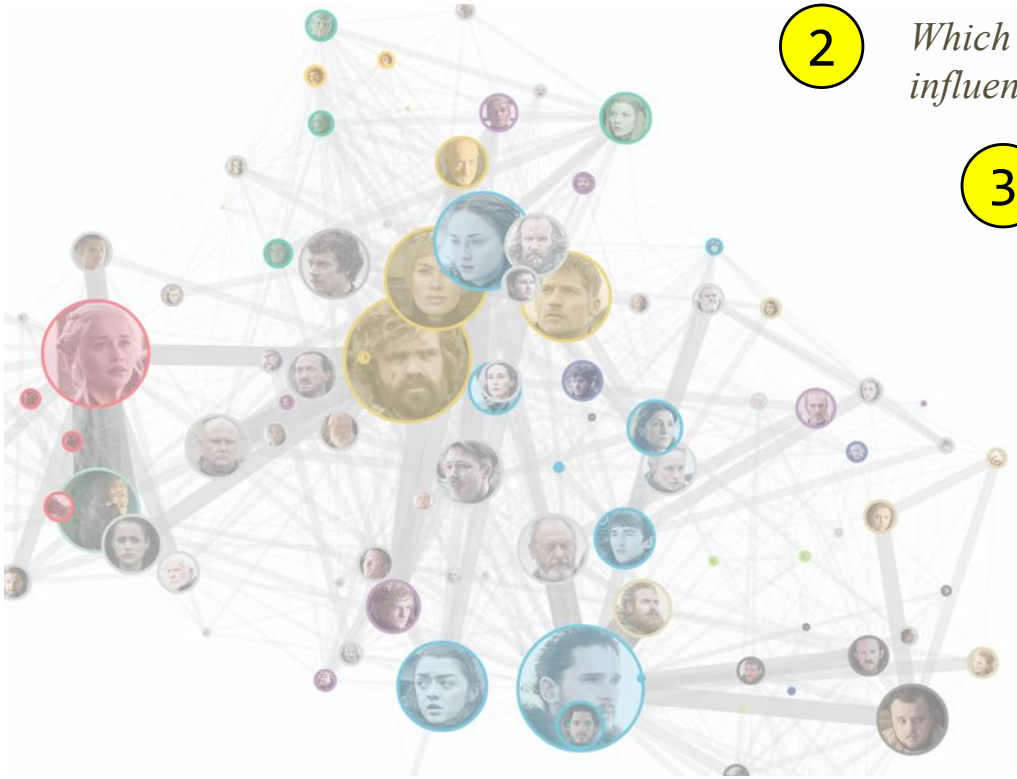
Hubs, Nodes & Links

Knowledge Network



CONNECTED DATA

1 *How many characters are there in GOT?*



2 *Which are the houses that challenge the throne and how influential are they?*

3 *How are the most popular characters?*

How does popularity evolve of main characters evolve along the story?

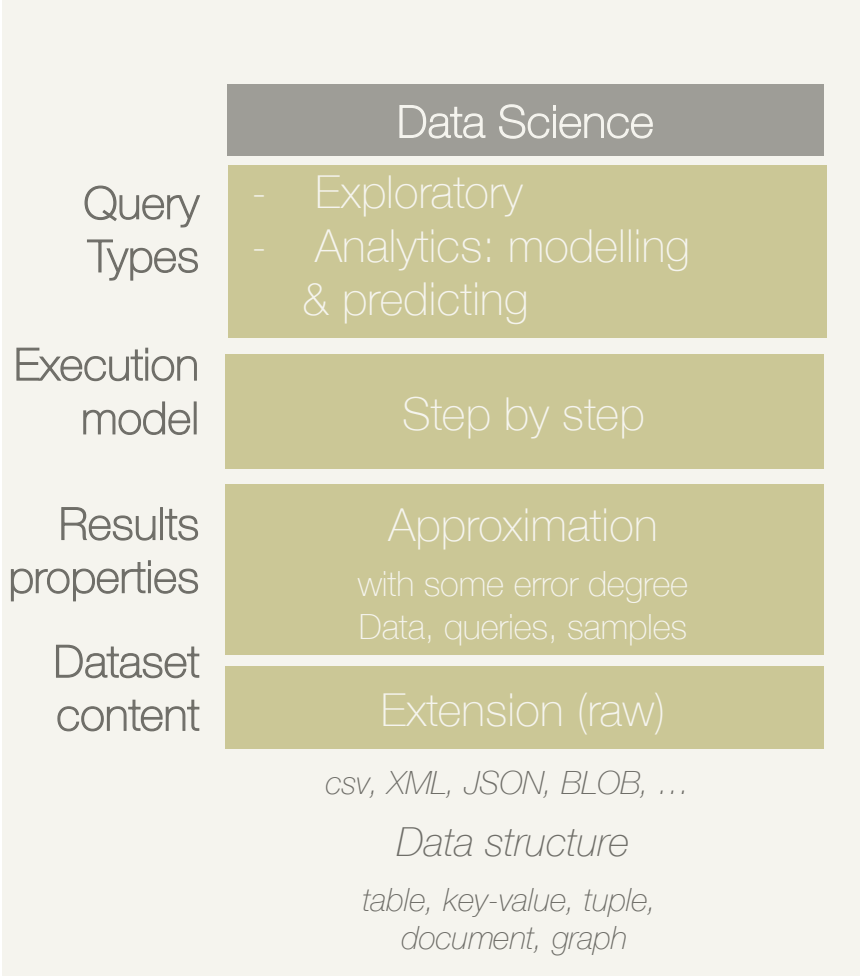
4 *Determine a geographical organisation of the different countries given the trajectories performed by armies?*

5 *Predict who can be the next King/Queen?*

QUERYING APPROACHES

	Databases ¦ information retrieval		Data Science
Query Types	- Relational, multi-dimensional, spatio-temporal, aggregation - Patterns, regular expressions	- (Dis)conjunctive - Navigational	- Exploratory - Analytics: modelling & predicting
Execution model	On demand/continuous	On demand	Step by step
Results properties	- Completeness (full/partial) - Fussiness	Approximation Precision/recall I Probabilistic	Approximation with some error degree Data, queries, samples
Dataset content	Intention model	Extension model	Extension (raw)
	Data structure *table, key-value, tuple, document, graph*	*Quantitative representation* *Frequency matrix, Statistical profiling* *Semantic representation* *Ontology, Terms graph*	*csv, XML, JSON, BLOB, …*

EXPLORATORY QUERYING



- Methodologies weaving data management, greedy algorithms

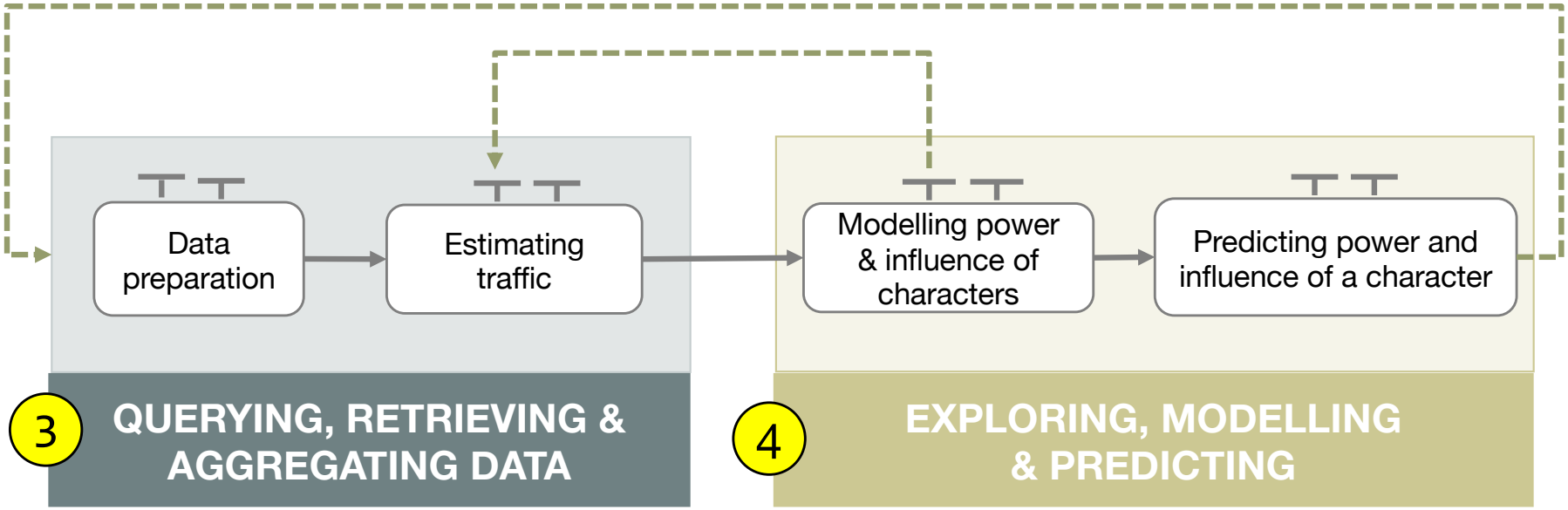
- Programming models that must be tuned to be deployed in different target architectures

Data collections as backbone for conducting **experiments**, drive hypothesis and lead to "valid" conclusions, models, simulations, understanding

DATA SCIENCE PIPELINE

1 COMPLEX AND REPETITIVE PROCESSING & ANALYSIS TASKS

2 **Example:** *Predict who will be the last king/queen of GOT?*

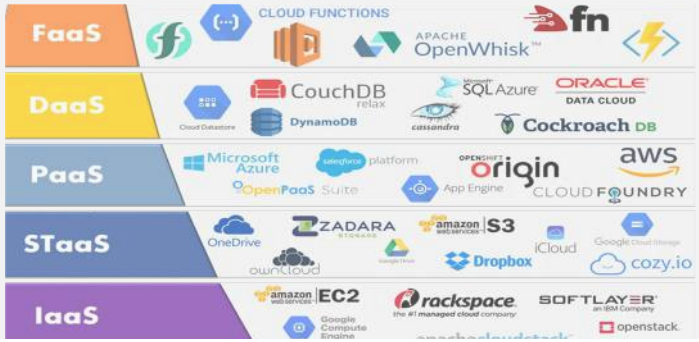


DATA SCIENCE PIPELINE

ISSUES

1. **Artisanal design** depending on data scientist/engineers "*expertise*"

2. In-house programming using many different **libraries, stacks, tools difficult to integrate**

DATA SCIENCE LABS	DATA SCIENCE STACKS	BD SERVICES PLATFORMS
kaggle.com Google Colab Azure Notebooks		

GRAPH STORES AND SYSTEMS

Graph Operations

1. **Community detection**: *label propagation, Louvain, weakly/strongly connected components, triangle count, colouring, local clustering coefficient*
2. **Centrality**: page rank, betweenness, closeness, degree, Eigenvector
3. **Similarity:** node, nearest neighbours, cosine/Euclidean/ Jaccard/ Overlap/Pearson
4. **Pathfinding & search**: adamic Adar, common/total neighbours, same community
5. **Heuristic link prediction**
   - **Pattern discovery**

Stores

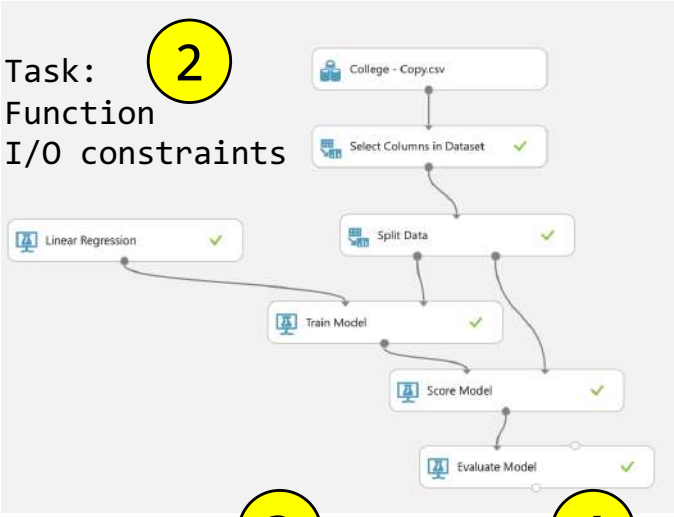
1. - **GraphDB lite**
   - **Neo4J**
   - **OrientDB**
   - **GraphEngine**
   - **HyperGraphDB**
   - **MapGraph**
   - **ArangoDB**
   - **Titan**
   - **BrightStarDB**
   - **CayLayGraph**
   - **WhiteDB**
   - **Orly**
   - **CosmosDB**
2. - **Proprietary data structures** (RDF, matrices/tables, etc)
   - **Query languages with built in functions**

Analytics tools

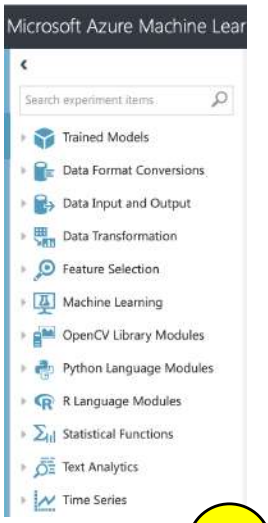
1. **Python networkx:** combined with numpy & pandas
   shortest_path(),
   shortest_path_length(),
   all_pairs_dijkstra_path()
2. **Spark Graphix**
3. **Deep Learning Models:** node classification, link prediction and graph clustering
4. **Graph processing systems (Pregel, Giraph):** global pattern matching, shortest paths, max-flow or min-cut, minimum spanning trees, diameter, eccentricity, connected components, PageRank, traversals

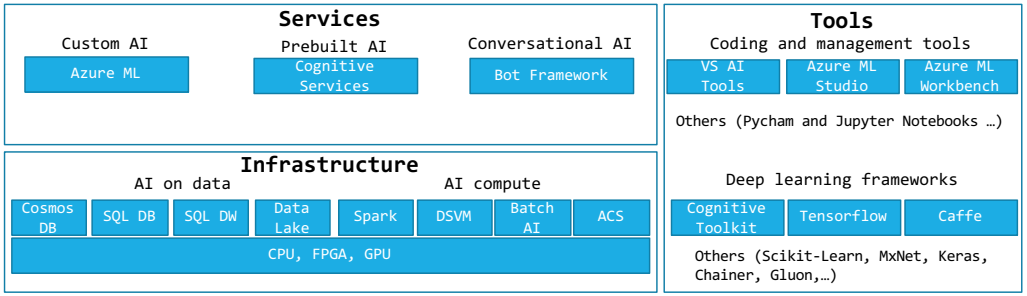
MACHINE LEARNING STUDIOS

Pipeline 1



DS Tools





Tracking 3

Record & query experiments: code, configs, results .. Etc

Projects 4

Packaging format for reproducible runs on any platform

Models 5

General model format that supports diverse deployment tools

Enactment (AI Platform, e.g. Microsoft, Databricks MLFlow, Google AI)

Cloud ML Studios

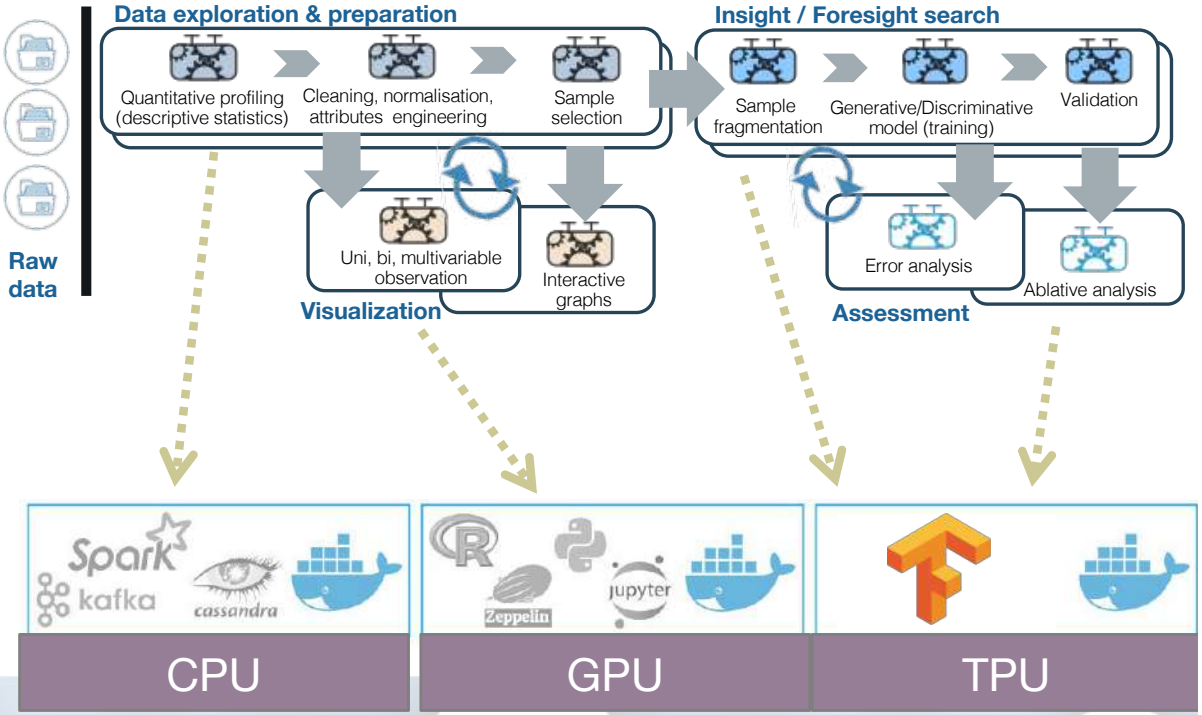
	Amazon	Microsoft	Google	IBM
Automated and semi-automated ML services				
	AmazonML	MS AzureML Studio	Cloud AutoML	IBM Watson Model Builder
Classification	✓	✓	✓	✓
Regression	✓	✓	✓	✓
Clustering	✓	✓	✖	✖
Anomaly detection	✖	✓	✖	✖
Recommendation	✖	✓	✓	✖
Ranking	✖	✓	✖	✖
Platforms for custom modelling				
	Amazon SageMaker	Azure ML Services	Google ML Engine	IBM WatsonM Studio
Built-in algorithms	✓	✖	✓	✓
Supported Frameworks	Tensorflow, MXNet, Keras, Gluon, Pytorch, Caffe2, Chainer, Torch	Tensorflow, Scikit-Learn, MS Cognitive Toolkit, SparkML	Tensorflow, Scikit-Learn, XGBoost, Keras	Tensorfoow, SparkMLib, Scikit-Learn, XGBoost, PyTorch, IBM SPSS, PMML

CONCLUSION & OUTLOOK

Next data science querying environments providing services at scale addressing data and processing divide are yet to come

1. Data science queries as **services coordinations** for processing graphs

2. **Deployment** on target architectures (cloud, GPU) for large-scale processing



Genoveva VARGAS-SOLAR.COM

Senior Scientist, CNRS, France

José-Luis Zechinelli - Martini

Universidad de las Américas Puebla, Mexico

joseluis.zechinelli@udlap.mx

Javier Espinosa

University Jean Moulin Lyon 3, France

javier.espinosa@imag.fr