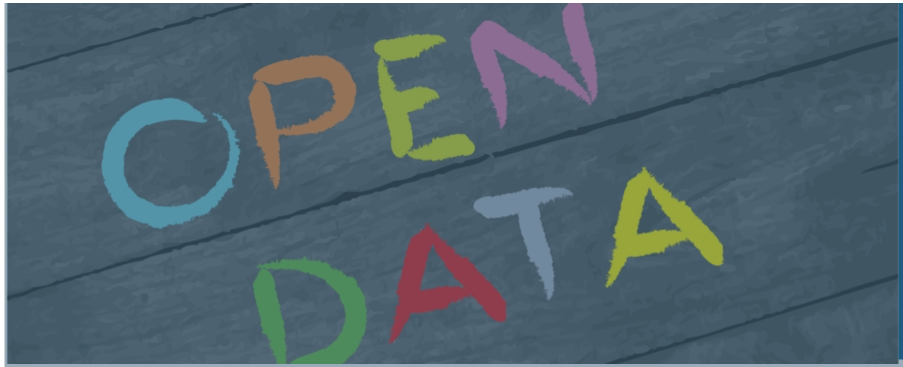
Designing Experiments

Overview of data management & exploitation solutions

Genoveva Vargas-Solar
French Council of Scientific Research, LIG
genoveva.vargas@imag.fr

http://vargas-solar.com/data-centric-smart-everything/

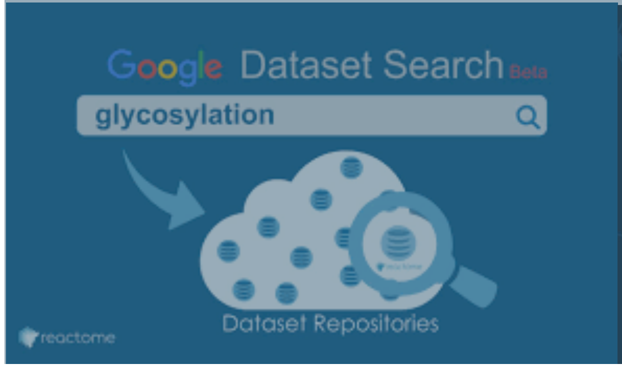


Microsoft Research Ope

Search datasets

A collection of free datasets from Microsoft Research to advance state-of-the-art research in areas such as natural language processing, computer vision, and domain specific sciences. Downloa to a cloud-based Data Science Virtu seamless development experience.

Democratized access to open data collections & algorithms released under different conditions & qualities

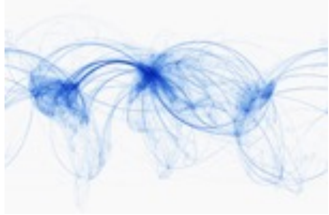


Microsoft Research Open Data BETA

Search datasets

A collection of free datasets from Microsoft Research to advance state-of-the-art research in areas such as natural language processing, computer vision, and domain specific sciences. Download or copy directly to a cloud-based Data Science Virtual Machine for a seamless development experience.

Network Science



Computational Science

Social Data Science

Digital humanities

Computation

(Algorithm: mathematical model)

Experiment setting

(Architecture: computing environment)

Volume

Velocity

Variety

Value

Veracity

1000 Yottabytes	1 Brontobyte
1000 Brontobytes	1 Geopbyte

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

	Data Science
Query Types	- Exploratory - Analytics: modelling & predicting
Execution model	Step by step
Results properties	Approximation with some error degree Data, queries, samples
Dataset content	Extension (raw)

csv, XML, JSON, BLOB, …

Data structure

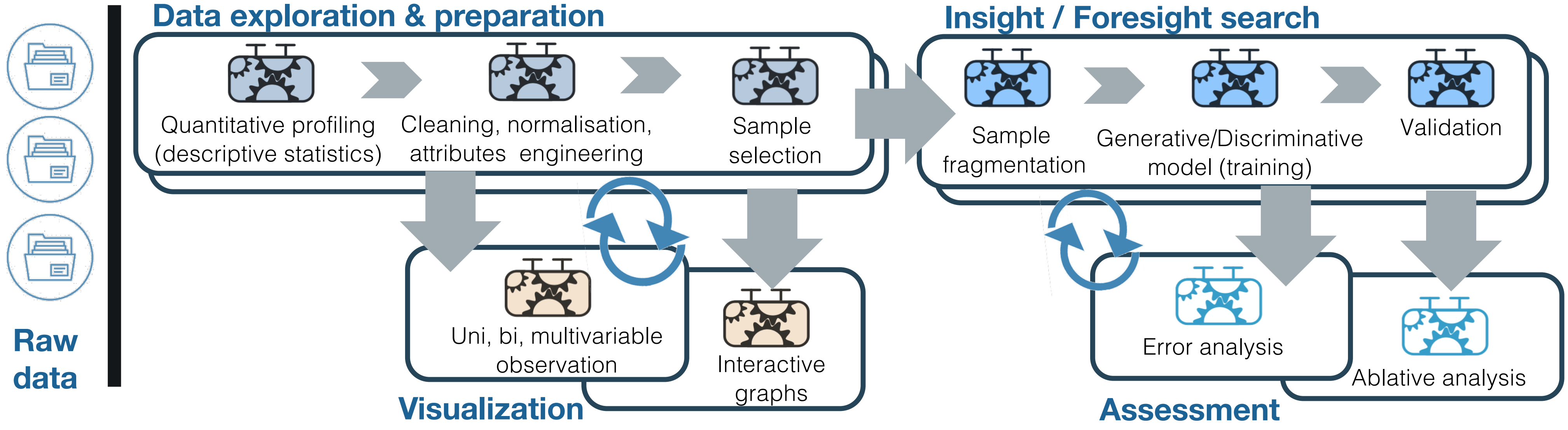
table, key-value, tuple, document, graph

- 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 CONSUMPTION PHILOSOPHIES

Diogenes approach


Data exploration & preparation
Quantitative profiling (descriptive statistics)
Cleaning, normalisation, attributes engineering
Sample selection
Insight / Foresight search
Sample fragmentation
Generative/Discriminative model (training)
Validation
Raw data
Uni, bi, multivariable observation
Interactive graphs
Visualization
Error analysis
Ablative analysis
Assessment

- **Data:** observations of phenomena often described as series of features/attributes
- **Query:** analytics objective (looks for insights or foresights) expressed as a pipeline of operations guided by the conditions and characteristics of the data
- **Result**: a model or prediction with associated assessment indexes, not definitive accepted with an associated error margin, accepted by comparison

DATA PROCESSING AND ANALYSIS

CONTEXT: URBAN DATA

"Does it make sense to invest in low-carbon technologies?"

Spatio – temporal series

Inhouse observed variables

- Electric consumption
- Indoor temperature
- Indoor humidity
- Gas consumption
- Outdoor temperature
- Outdoor humidity

Meteorological variables

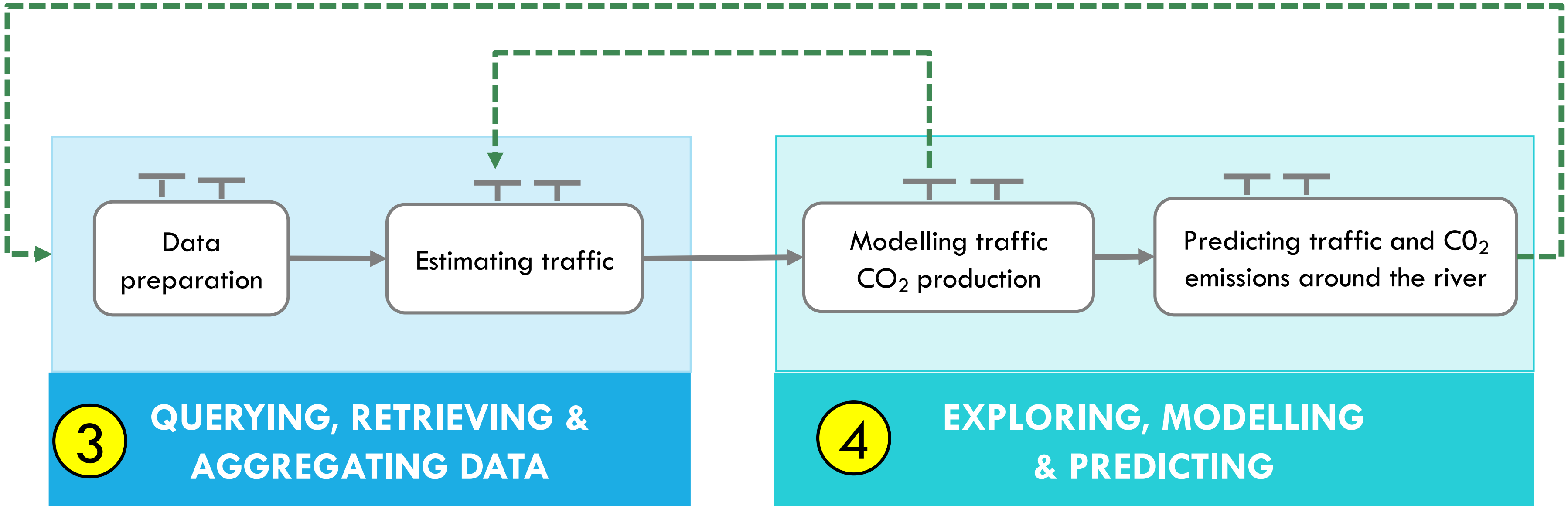
- Total precipitation
- Total cloud cover
- Shortwave radiation
- Wind direction
- Snowfall amount
- Sunshine duration
- Wind speed

- Heterogeneous masses of data , often spatio-temporal and produced as streams (*i.e.*, spatio-temporal series),
- Backbone of analytical and prediction processes

DATA SCIENCE PIPELINE

1 COMPLEX AND REPETITIVE PROCESSING & ANALYSIS TASKS

Example: 2 *Will jogging in the morning around the river next weekend reduce breathing CO_2 ?*



Data preparation

Estimating traffic

Modelling traffic CO_2 production

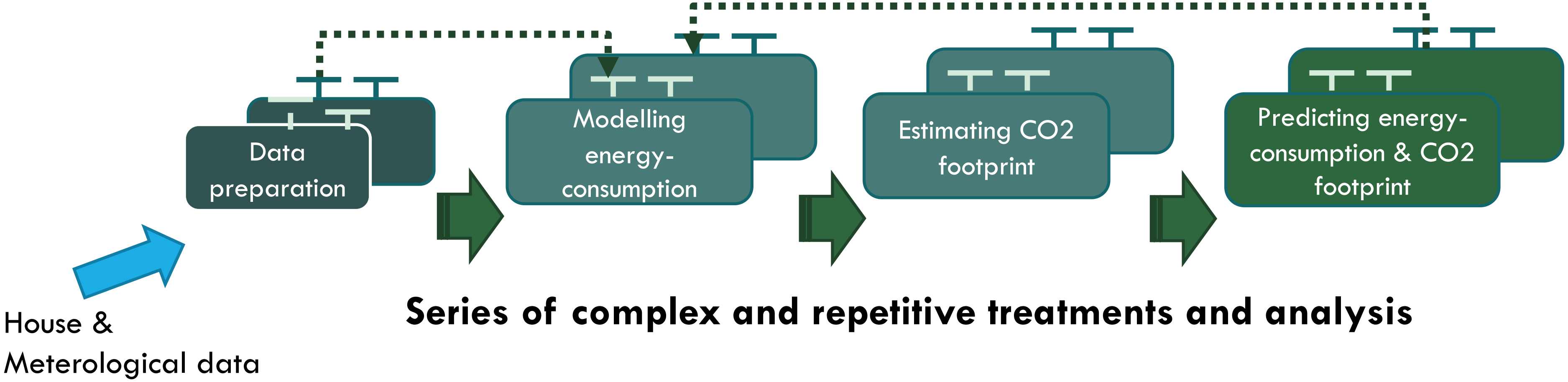
Predicting traffic and $C0_2$ emissions around the river

3 QUERYING, RETRIEVING & AGGREGATING DATA

4 EXPLORING, MODELLING & PREDICTING

DATA SCIENCE PIPELINE

"Does it make sense to invest in low-carbon technologies?"

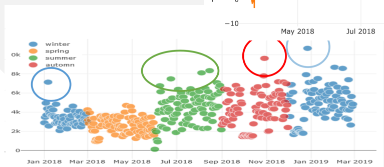

Data preparation
Modelling energy-consumption
Estimating CO2 footprint
Predicting energy-consumption & CO2 footprint
House &
Meterological data

Series of complex and repetitive treatments and analysis

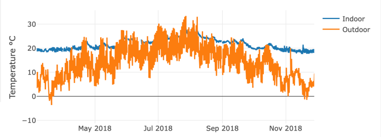
- **Exploration & preparation:** Identify which variables impact my household CO_2 footprint
- **Analysis:** Predict whether investing in low-carbon technologies will decrease my CO_2 footprint
- **Assessment:** Identify which prediction model is better for my household energy consumption pattern

DATA SCIENCE PIPELINE

Outliers



Distribution



Statistical Analysis

Count	102
Count	10152.000000
Mean	164.634161
Std	187.285460
Min	0.000000
25%	62.000000
50%	99.000000
75%	180.250000
Max	1985.000000

Understanding raw data

House &
Meterological data

Data preparation

ing energy-consumption

Estimating CO2 footprint

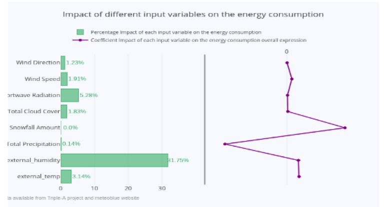
ing energy-consumption & CO2 footprint

CO_2 Cost Formula

Member state	Displaced electricity	Natural gas	Heating Oil	Biomass
UK	0.519	0.216	0.298	0.039
Netherlands	0.530	0.204	0.267	0/0.395
France	0.09	0.241	0.329	≈ 0.013
Belgium	0.258	0.202	0.279	0

$$\frac{tCO_2}{a} = \frac{[(EDPM(kWh) - EDPI(kWh)) \times REF(kgCO_2/kWh)]}{1000}$$

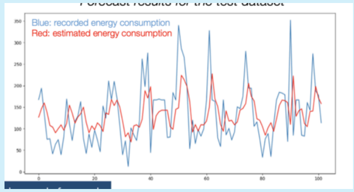
Impact of other variables on energy consumption

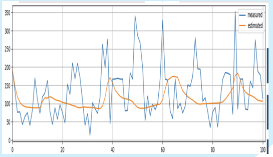

Impact of different input variables on the energy consumption

ta available from Triple-A project and meteoblue website

Naïve Forest Persistence → ARIMA → ARX

Forecast results for the test dataset

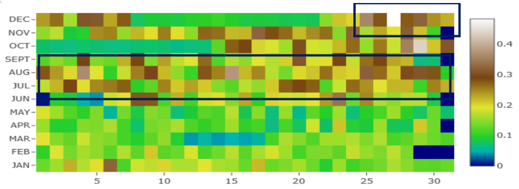




Time Series Analisis

RMSE= 65.025

RMSE= 55.84323



Investement will reduce CO_2 in Xmas and Summer !

DATA SCIENCE PIPELINE

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

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

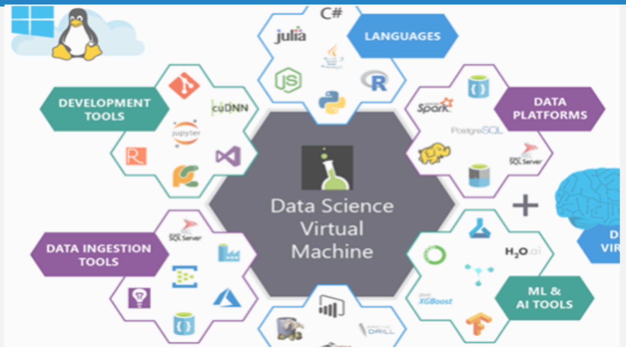
DATA SCIENCE LABS

kaggle.com

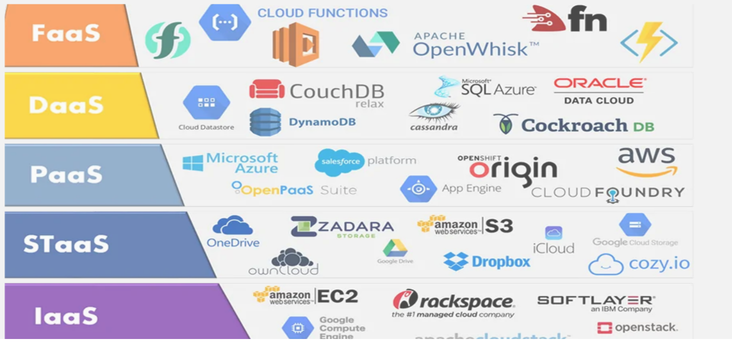
Google Colab

Azure Notebooks

DATA SCIENCE STACKS



BD SERVICES PLATFORMS



DATA SCIENCE PIPELINE

CHALLENGES

1. Efficient execution on distributed architectures requiring **important engineering effort**

2. **Tuning** for improving **data management** across nodes hosting software components and libraries

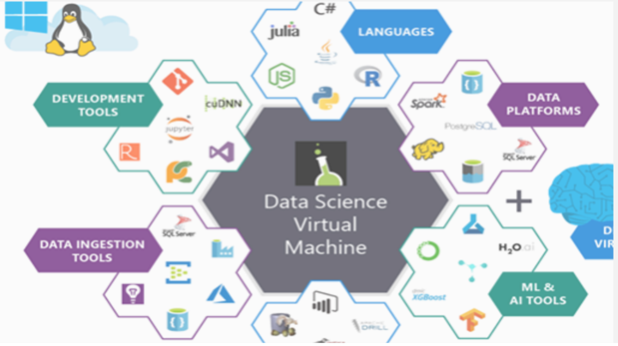
DATA SCIENCE LABS

kaggle.com

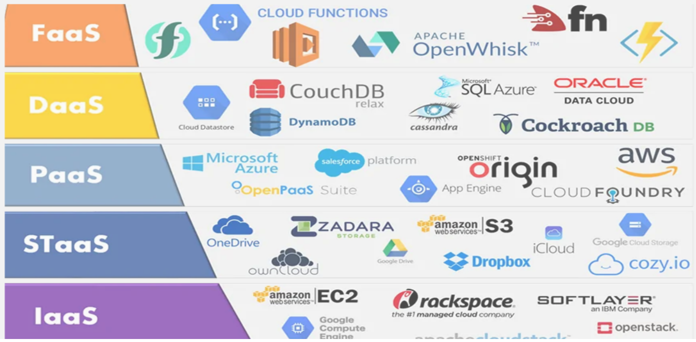
Google Colab

Azure Notebooks

DATA SCIENCE STACKS

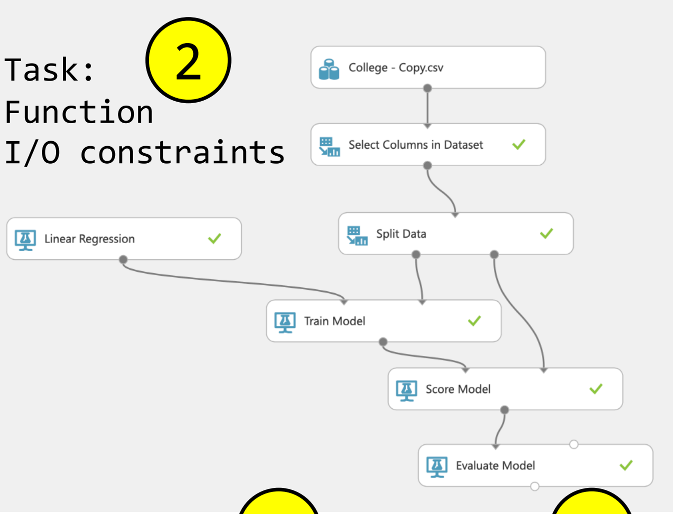


BD SERVICES PLATFORMS



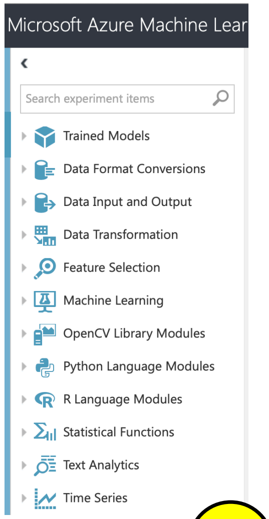
Machine Learning Studios

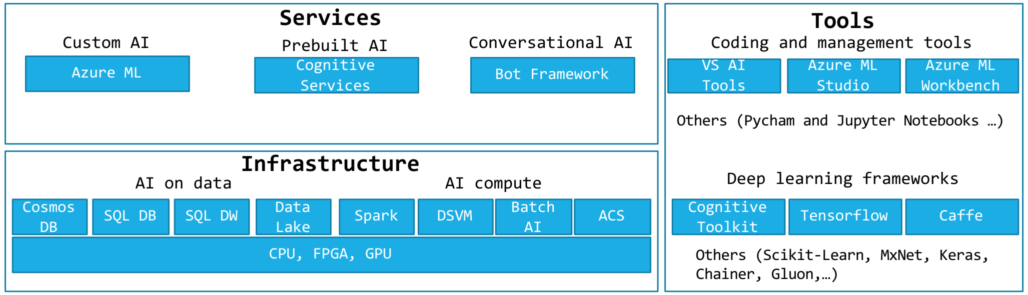
Pipeline 1



Task: 2
Function
I/O constraints

DS Tools





Services

Custom AI: Azure ML

Prebuilt AI: Cognitive Services

Conversational AI: Bot Framework

Infrastructure

AI on data: Cosmos DB, SQL DB, SQL DW, Data Lake

AI compute: Spark, DSVM, Batch AI, ACS

CPU, FPGA, GPU

Tools

Coding and management tools: VS AI Tools, Azure ML Studio, Azure ML Workbench

Others (Pycham and Jupyter Notebooks …)

Deep learning frameworks: Cognitive Toolkit, Tensorflow, Caffe

Others (Scikit-Learn, MxNet, Keras, Chainer, Gluon,…)

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

PREPARING DATA

DATA MODELS

Tuple

- Row in a relational table, where attributes are pre-defined in a schema, and the values are scalar

Document

- Allows values to be nested documents or lists, as well as scalar values.
- Attributes are not defined in a global schema

Extensible record

- Hybrid between tuple and document, where families of attributes are defined in a schema, but new attributes can be added on a per-record basis

DATA STORES

Key-value

- Systems that store values and an index to find them, based on a key

Document

- Systems that store documents, providing index and simple query mechanisms

Extensible record

- Systems that store extensible records that can be partitioned vertically and horizontally across nodes

Graph

- Systems that store model data as graphs where nodes can represent content modelled as document or key-value structures and arcs represent a relation between the data modelled by the node

Relational

- Systems that store, index and query tuples

KEY STORE VALUES

“Simplest data stores” use a data model similar to the memcached distributed in-memory cache

Single key-value index for all data

Provide a persistence mechanism

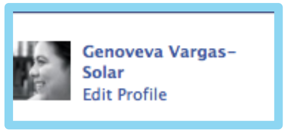
Replication, versioning, locking, transactions, sorting

API: inserts, deletes, index lookups

No secondary indices or keys

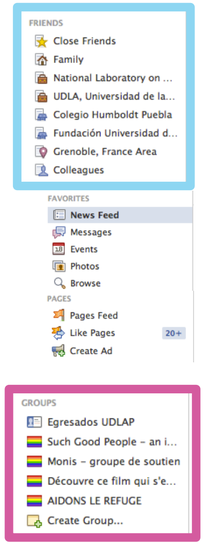
System	Address
Redis	code.google.com/p/redis
Scalaris	code.google.com/p/scalaris
Tokyo	tokyocabinet.sourceforge.net
Voldemort	project-voldemort.com
Riak	riak.basho.com
Membrain	schoonerinfotech.com/products
Membase	membase.com

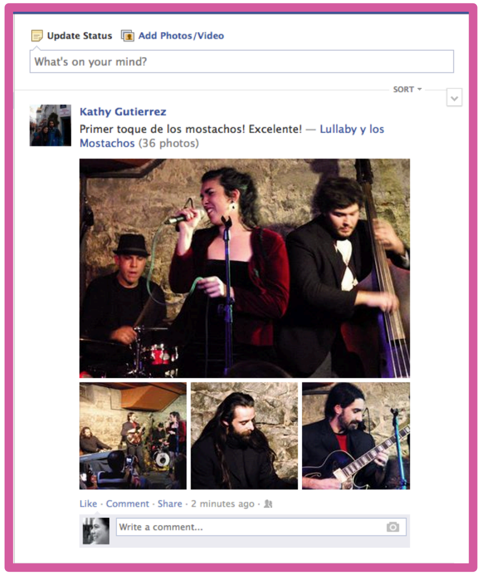
```
SELECT   name, pic, profile_url
FROM     user
WHERE    uid = me()
```


Genoveva Vargas-Solar

```
SELECT   message, attachment
FROM     stream
WHERE    source_id = me() AND type = 80
```

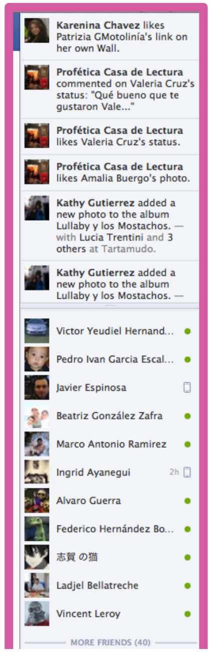
```
SELECT   name
FROM     friendlist
WHERE    owner = me()
```





Kathy Gutierrez

Primer toque de los mostachos! Excelente! — Lullaby y los Mostachos (36 photos)


Karenina Chavez likes Patrizia GMotolinía's link on her own Wall.

Profética Casa de Lectura commented on Valeria Cruz's status: "Qué bueno que te gustaron Vale..."

Profética Casa de Lectura likes Valeria Cruz's status.

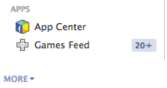
Profética Casa de Lectura likes Amalia Buergo's photo.

Kathy Gutierrez added a new photo to the album Lullaby y los Mostachos. — with Lucia Trentini and 3 others at Tartamudo.

Kathy Gutierrez added a new photo to the album Lullaby y los Mostachos. —

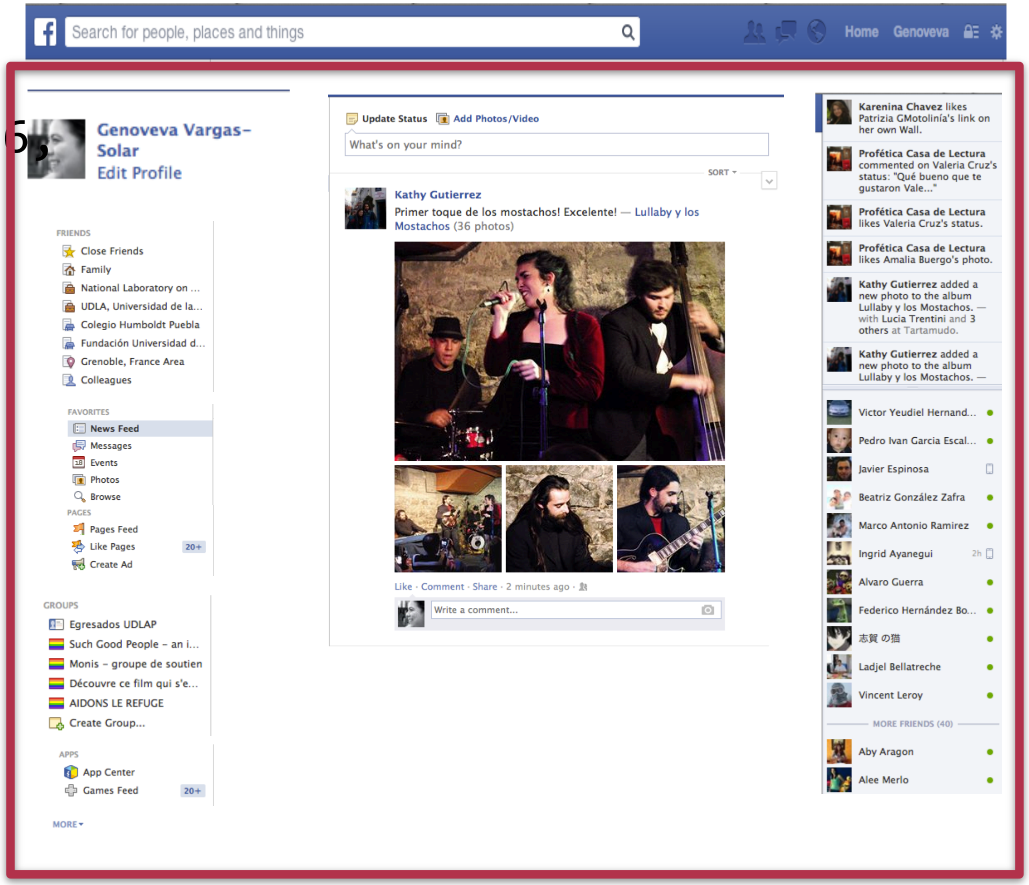
```
SELECT   name
FROM     group
WHERE    gid IN ( SELECT   gid
                  FROM     group_member
                  WHERE    uid = me() )
```

```
SELECT   name, pic
FROM     user
WHERE    online_presence = "active"
         AND
         uid IN ( SELECT   uid2
                  FROM     friend
                  WHERE    uid1 = me() )
```



https://developers.facebook.com/docs/reference/fql/

<805114856,



Search for people, places and things

Home Genoveva

Genoveva Vargas-Solar
Edit Profile

FRIENDS
- Close Friends
- Family
- National Laboratory on ...
- UDLA, Universidad de la...
- Colegio Humboldt Puebla
- Fundación Universidad d...
- Grenoble, France Area
- Colleagues

FAVORITES
- News Feed
- Messages
- Events
- Photos
- Browse

PAGES
- Pages Feed
- Like Pages 20+
- Create Ad

GROUPS
- Egresados UDLAP
- Such Good People - an i...
- Monis - groupe de soutien
- Découvre ce film qui s'e...
- AIDONS LE REFUGE
- Create Group...

APPS
- App Center
- Games Feed 20+

MORE ▾

Update Status Add Photos/Video

What's on your mind?

SORT ▾

Kathy Gutierrez
Primer toque de los mostachos! Excelente! — Lullaby y los Mostachos (36 photos)

Like · Comment · Share · 2 minutes ago ·

Write a comment...

Karenina Chavez likes Patrizia GMotolinía's link on her own Wall.

Profética Casa de Lectura commented on Valeria Cruz's status: "Qué bueno que te gustaron Vale..."

Profética Casa de Lectura likes Valeria Cruz's status.

Profética Casa de Lectura likes Amalia Buergo's photo.

Kathy Gutierrez added a new photo to the album Lullaby y los Mostachos. — with Lucia Trentini and 3 others at Tartamudo.

Kathy Gutierrez added a new photo to the album Lullaby y los Mostachos. —

- Victor Yeudiel Hernand... ●
- Pedro Ivan Garcia Escal... ●
- Javier Espinosa
- Beatriz González Zafra ●
- Marco Antonio Ramirez ●
- Ingrid Ayanegui 2h
- Alvaro Guerra ●
- Federico Hernández Bo... ●
- 志賀 の猫 ●
- Ladjel Bellatreche ●
- Vincent Leroy ●

MORE FRIENDS (40)

- Aby Aragon ●
- Alee Merlo ●

>

DOCUMENT STORES

Support more complex data: pointerless objects, i.e., documents

Secondary indexes, multiple types of documents (objects) per database, nested documents and lists, e.g. B-trees

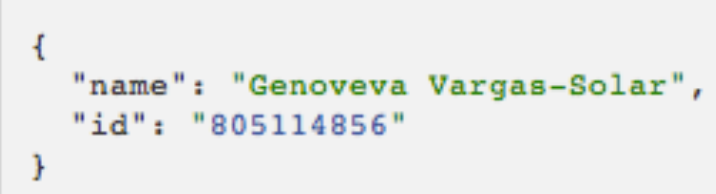
Automatic sharding (scale writes), no explicit locks, weaker concurrency (eventual for scaling reads) and atomicity properties

API: `select, delete, getAttributes, putAttributes` on documents

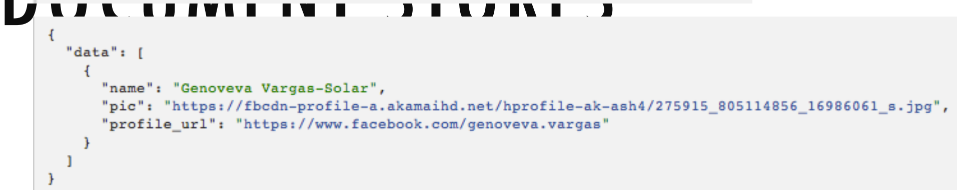
Queries can be distributed in parallel over multiple nodes using a map-reduce mechanism

SYSTEM	ADDRESS
SimpleDB	`amazon.com/simpledb`
Couch DB	`couchdb.apache.org`
Mongo DB	`mongodb.org`
Terrastore	`code.google.com/terrastore`

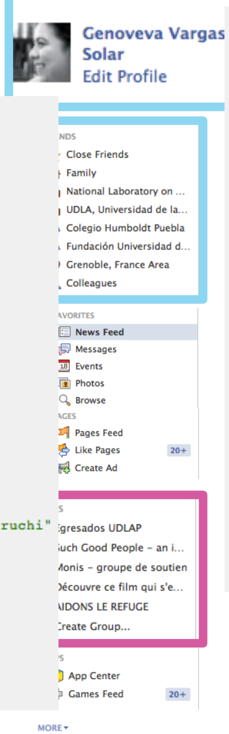
DOCUMENT STORES

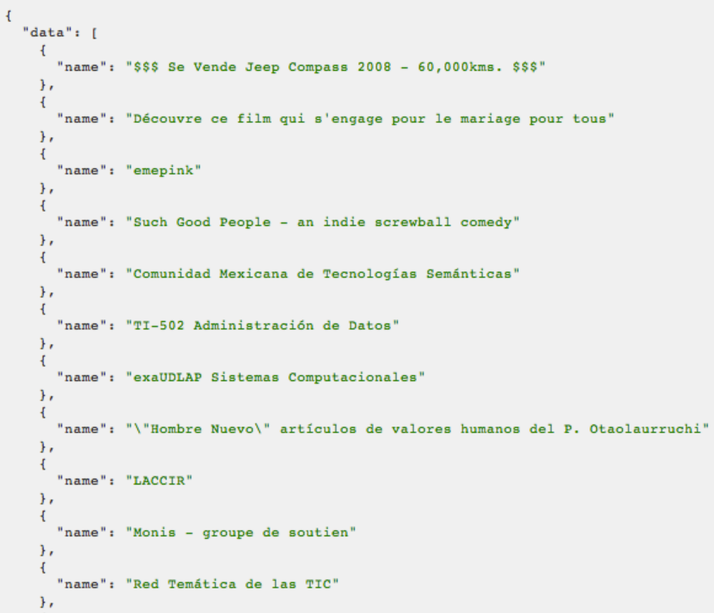


```
{
  "name": "Genoveva Vargas-Solar",
  "id": "805114856"
}
```

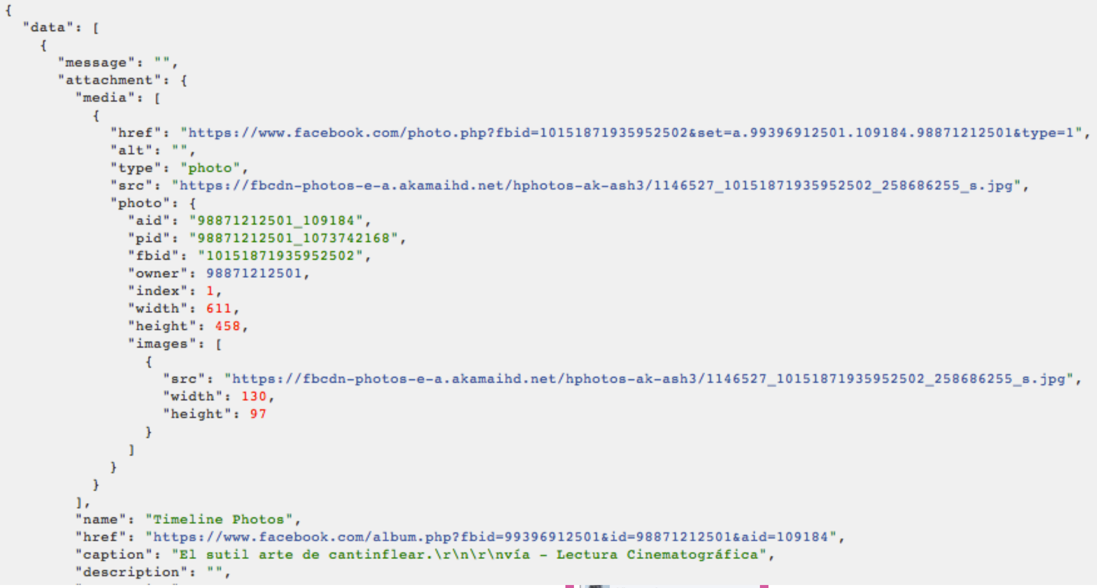


```
{
  "data": [
    {
      "name": "Genoveva Vargas-Solar",
      "pic": "https://fbcdn-profile-a.akamaihd.net/hprofile-ak-ash4/275915_805114856_16986061_s.jpg",
      "profile_url": "https://www.facebook.com/genoveva.vargas"
    }
  ]
}
```

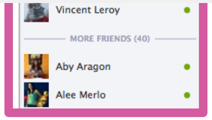




```
{
  "data": [
    {
      "name": "$$$ Se Vende Jeep Compass 2008 - 60,000kms. $$$"
    },
    {
      "name": "Découvre ce film qui s'engage pour le mariage pour tous"
    },
    {
      "name": "emepink"
    },
    {
      "name": "Such Good People - an indie screwball comedy"
    },
    {
      "name": "Comunidad Mexicana de Tecnologías Semánticas"
    },
    {
      "name": "TI-502 Administración de Datos"
    },
    {
      "name": "exaUDLAP Sistemas Computacionales"
    },
    {
      "name": "\"Hombre Nuevo\" artículos de valores humanos del P. Otaolaurruchi"
    },
    {
      "name": "LACCIR"
    },
    {
      "name": "Monis - groupe de soutien"
    },
    {
      "name": "Red Temática de las TIC"
    },
```



```
{
  "data": [
    {
      "message": "",
      "attachment": {
        "media": [
          {
            "href": "https://www.facebook.com/photo.php?fbid=10151871935952502&set=a.99396912501.109184.98871212501&type=1",
            "alt": "",
            "type": "photo",
            "src": "https://fbcdn-photos-e-a.akamaihd.net/hphotos-ak-ash3/1146527_10151871935952502_258686255_s.jpg",
            "photo": {
              "aid": "98871212501_109184",
              "pid": "98871212501_1073742168",
              "fbid": "10151871935952502",
              "owner": 98871212501,
              "index": 1,
              "width": 611,
              "height": 458,
              "images": [
                {
                  "src": "https://fbcdn-photos-e-a.akamaihd.net/hphotos-ak-ash3/1146527_10151871935952502_258686255_s.jpg",
                  "width": 130,
                  "height": 97
                }
              ]
            }
          }
        ],
        "name": "Timeline Photos",
        "href": "https://www.facebook.com/album.php?fbid=99396912501&id=98871212501&aid=109184",
        "caption": "El sutil arte de cantinflear.\r\n\r\nvía - Lectura Cinematográfica",
        "description": "",
```



EXTENSIBLE RECORD STORES

Basic data model is rows and columns

Basic scalability model is splitting rows and columns over multiple nodes

- Rows split across nodes through sharding on the primary key
  - Split by range rather than hash function
  - Rows analogous to documents: variable number of attributes, attribute names must be unique
  - Grouped into collections (tables)
  - Queries on ranges of values do not go to every node

Columns are distributed over multiple nodes using “column groups”

- Which columns are best stored together
- Column groups must be pre-defined with the extensible record stores

SYSTEM	ADDRESS
HBase	hbase.apache.com
HyperTable	hypertable.org
Cassandra	incubator.apache.org/cassandra

SCALABLE RELATIONAL SYSTEMS

SQL: rich declarative query language

Databases reinforce referential integrity

ACID semantics

Well understood operations:

- Configuration, Care and feeding, Backups, Tuning, Failure and recovery, Performance characteristics

Use small-scope operations

- Challenge: joins that do not scale with sharding

Use small-scope transactions

- ACID transactions inefficient with communication and 2PC overhead

Shared nothing architecture for scalability

Avoid cross-node operations

System	Address
MySQL C	mysql.com/cluster
Volt DB	voltdb.com
Clustrix	clustrix.com
ScaleDB	scaledb.com
Scale Base	scalebase.com
Nimbus DB	nimbusdb.com

1970 - 2000 RELATIONAL DB

More than 30 years: maturity!

Theoretical & Practical aspects (DBMS)

Domains & R $\subseteq$ D_1 x D_2 x . . . . D_n, Algebra →

1st Order Predicate Logic

Languages: SQL (wins), QUEL, QBE

DBMS Prototypes (1975), Products (1980)

A major improvement in DB: provide **data independence** & a simple, **tabular view** of data

Normal Forms & Dependencies (DB design, **consistency**)

Controversial: missing values, duplicates

R x S

R ∪ S

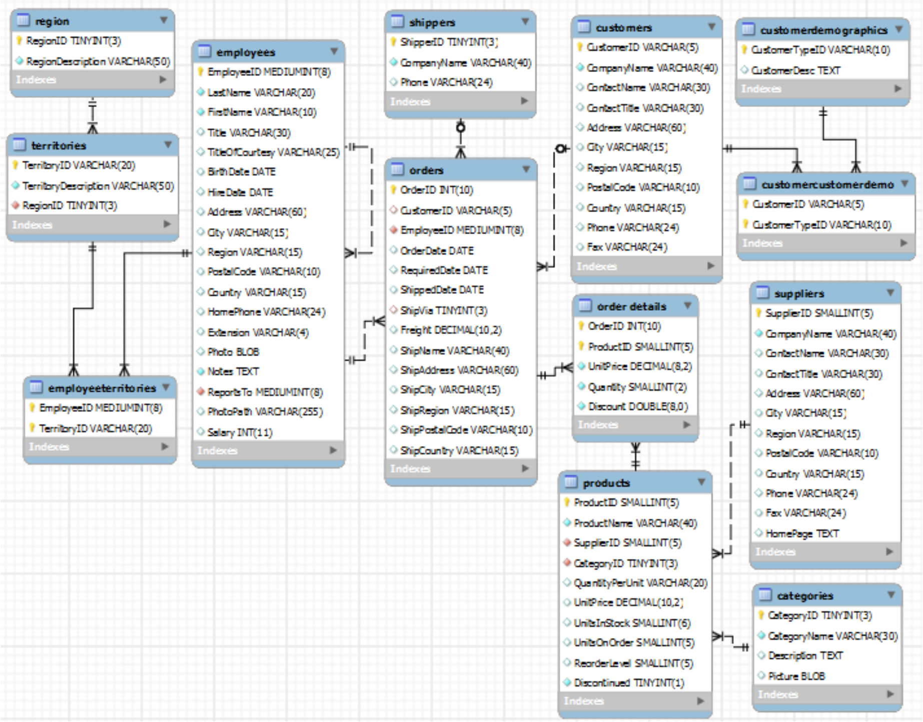
R - S

R[α]

R : φ

R * S

EXAMPLE



region
- RegionID TINYINT(3)
- RegionDescription VARCHAR(50)
- Indexes

territories
- TerritoryID VARCHAR(20)
- TerritoryDescription VARCHAR(50)
- RegionID TINYINT(3)
- Indexes

employeeterritories
- EmployeeID MEDIUMINT(8)
- TerritoryID VARCHAR(20)
- Indexes

employees
- EmployeeID MEDIUMINT(8)
- LastName VARCHAR(20)
- FirstName VARCHAR(10)
- Title VARCHAR(30)
- TitleOfCourtesy VARCHAR(25)
- BirthDate DATE
- HireDate DATE
- Address VARCHAR(60)
- City VARCHAR(15)
- Region VARCHAR(15)
- PostalCode VARCHAR(10)
- Country VARCHAR(15)
- HomePhone VARCHAR(24)
- Extension VARCHAR(4)
- Photo BLOB
- Notes TEXT
- ReportsTo MEDIUMINT(8)
- PhotoPath VARCHAR(255)
- Salary INT(11)
- Indexes

shippers
- ShipperID TINYINT(3)
- CompanyName VARCHAR(40)
- Phone VARCHAR(24)
- Indexes

orders
- OrderID INT(10)
- CustomerID VARCHAR(5)
- EmployeeID MEDIUMINT(8)
- OrderDate DATE
- RequiredDate DATE
- ShippedDate DATE
- ShipVia TINYINT(3)
- Freight DECIMAL(10,2)
- ShipName VARCHAR(40)
- ShipAddress VARCHAR(60)
- ShipCity VARCHAR(15)
- ShipRegion VARCHAR(15)
- ShipPostalCode VARCHAR(10)
- ShipCountry VARCHAR(15)
- Indexes

customers
- CustomerID VARCHAR(5)
- CompanyName VARCHAR(40)
- ContactName VARCHAR(30)
- ContactTitle VARCHAR(30)
- Address VARCHAR(60)
- City VARCHAR(15)
- Region VARCHAR(15)
- PostalCode VARCHAR(10)
- Country VARCHAR(15)
- Phone VARCHAR(24)
- Fax VARCHAR(24)
- Indexes

customerdemographics
- CustomerTypeID VARCHAR(10)
- CustomerDesc TEXT
- Indexes

customercustomerdemo
- CustomerID VARCHAR(5)
- CustomerTypeID VARCHAR(10)
- Indexes

order details
- OrderID INT(10)
- ProductID SMALLINT(5)
- UnitPrice DECIMAL(8,2)
- Quantity SMALLINT(2)
- Discount DOUBLE(8,0)
- Indexes

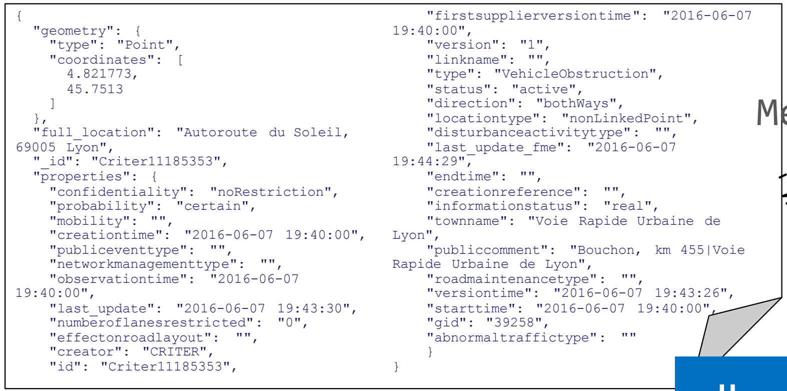
products
- ProductID SMALLINT(5)
- ProductName VARCHAR(40)
- SupplierID SMALLINT(5)
- CategoryID TINYINT(3)
- QuantityPerUnit VARCHAR(20)
- UnitPrice DECIMAL(10,2)
- UnitsInStock SMALLINT(6)
- UnitsOnOrder SMALLINT(5)
- ReorderLevel SMALLINT(5)
- Discontinued TINYINT(1)
- Indexes

suppliers
- SupplierID SMALLINT(5)
- CompanyName VARCHAR(40)
- ContactName VARCHAR(30)
- ContactTitle VARCHAR(30)
- Address VARCHAR(60)
- City VARCHAR(15)
- Region VARCHAR(15)
- PostalCode VARCHAR(10)
- Country VARCHAR(15)
- Phone VARCHAR(24)
- Fax VARCHAR(24)
- HomePage TEXT
- Indexes

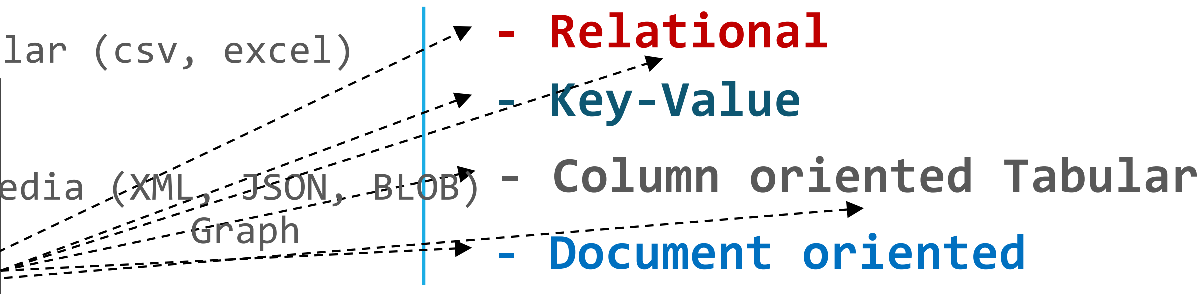
categories
- CategoryID TINYINT(3)
- CategoryName VARCHAR(30)
- Description TEXT
- Picture BLOB
- Indexes

MODELING DATA COLLECTIONS

Raw data collections



```
{
  "geometry": {
    "type": "Point",
    "coordinates": [
      4.821773,
      45.7513
    ]
  },
  "full_location": "Autoroute du Soleil,
69005 Lyon",
  "_id": "Criter11185353",
  "properties": {
    "confidentiality": "noRestriction",
    "probability": "certain",
    "mobility": "",
    "creationtime": "2016-06-07 19:40:00",
    "publiceventtype": "",
    "networkmanagementtype": "",
    "observationtime": "2016-06-07
19:40:00",
    "last_update": "2016-06-07 19:43:30",
    "numberoflanesrestricted": "0",
    "effectonroadlayout": "",
    "creator": "CRITER",
    "id": "Criter11185353",
    "firstsupplierversiontime": "2016-06-07
19:40:00",
    "version": "1",
    "linkname": "",
    "type": "VehicleObstruction",
    "status": "active",
    "direction": "bothWays",
    "locationtype": "nonLinkedPoint",
    "disturbanceactivitytype": "",
    "last_update_fme": "2016-06-07
19:44:29",
    "endtime": "",
    "creationreference": "",
    "informationstatus": "real",
    "townname": "Voie Rapide Urbaine de
Lyon",
    "publiccomment": "Bouchon, km 455|Voie
Rapide Urbaine de Lyon",
    "roadmaintenancetype": "",
    "versiontime": "2016-06-07 19:43:26",
    "starttime": "2016-06-07 19:40:00",
    "gid": "39258",
    "abnormaltraffictype": ""
    }
}
```



tabular (csv, excel)

Media (XML, JSON, BLOB)

Graph

- Relational
- Key-Value
- Column oriented Tabular
- Document oriented

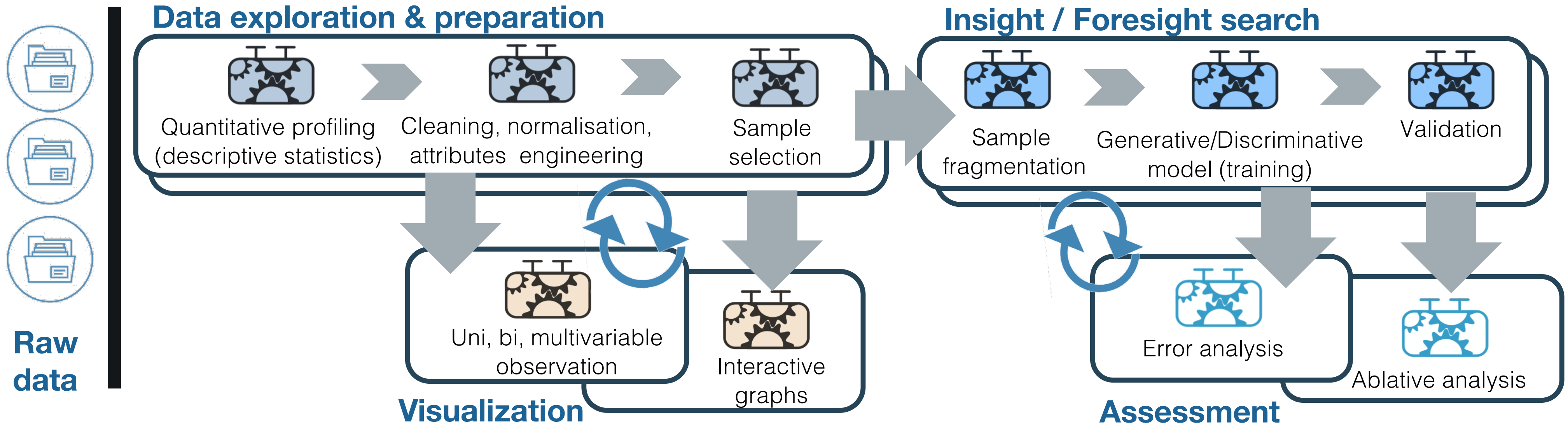
- How to transform data collections ?
- Which is the best adapted model?

→ **Polyglot persistence**

Approaches dealing with transformation rules inspired in the relational case

EXPERIMENT DESIGN

Diogenes approach


Raw data
Data exploration & preparation
Quantitative profiling (descriptive statistics)
Cleaning, normalisation, attributes engineering
Sample selection
Visualization
Uni, bi, multivariable observation
Interactive graphs
Insight / Foresight search
Sample fragmentation
Generative/Discriminative model (training)
Validation
Assessment
Error analysis
Ablative analysis

PREPARING DATA

Obtaining the data: Read from a file or obtained by scraping the web

Parsing the data: Format the data which can be in plain text, fixed columns, CSV, XML, HTML, etc.

Cleaning the data: A simple strategy is to remove or ignore incomplete records

Building data structures: A data structure that lends itself to the analysis we are interested in.

Databases provide a mapping from keys to values, so they serve as dictionaries

ANALYSING INCOME ACCORDING TO GENDER

Financial parameters related to the US population*

- **Features**: Age, sex, marital, country, income, education, occupation, capital gain, etc.
- **Question**: Are men more likely to become high-income professionals than women, i.e., to receive an income of over $50,000 per year?

- Preparing data collections
  - Read and check the data
  - Represent the data, for instance using a tabular data structure with features (columns) and records (rows)
  - Group the data

* UCI's Machine Learning Repository : https://archive.ics.uci.edu/ml/datasets/Adult

EXPLORATORY DATA ANALYSIS

Measurements and categories represent a sample distribution of a variable:

- which approximately represents the population distribution of the variable
- to make tentative assumptions about the population distribution

Different *techniques*:

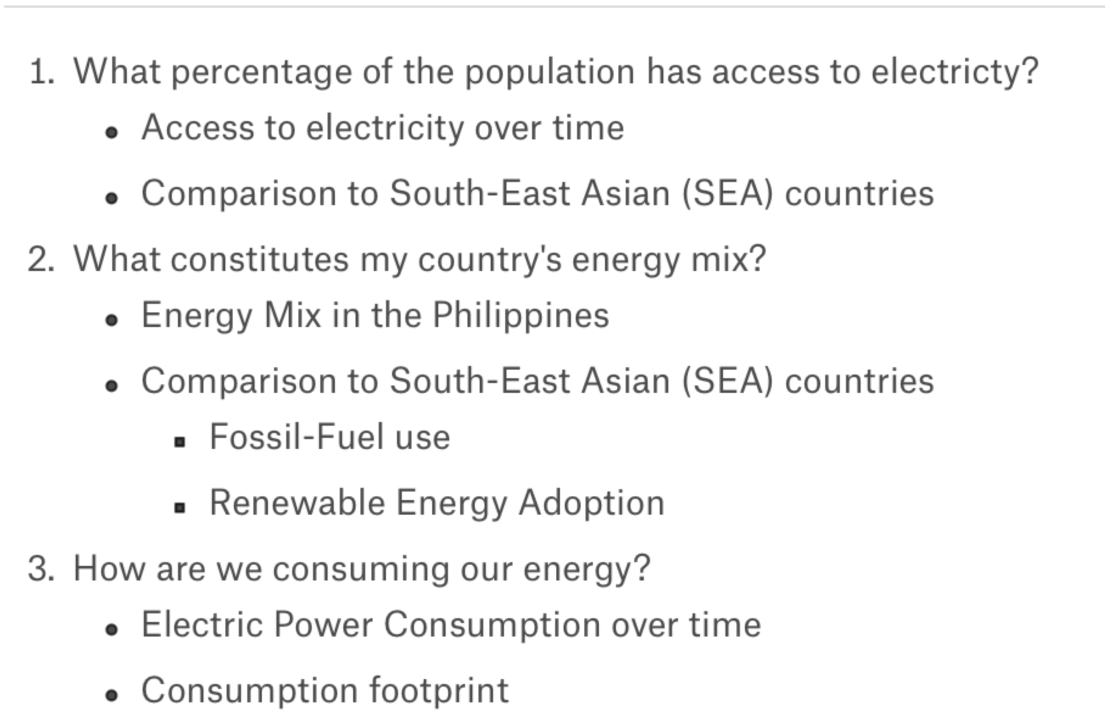
- **Summarizing the data**
- **Data distributions**
- **Outlier treatment**
- **Kernel density**

https://www.kaggle.com/robikscube/hourly-energy-consumption

SMART ENERGY

UNDERSTANDING ENERGY CONSUMPTION IN THE PHILIPPINES

https://www.kaggle.com/ljvmiranda/philippines-energy-use



1. What percentage of the population has access to electricty?
    - Access to electricity over time
    - Comparison to South-East Asian (SEA) countries
2. What constitutes my country's energy mix?
    - Energy Mix in the Philippines
    - Comparison to South-East Asian (SEA) countries
        - Fossil-Fuel use
        - Renewable Energy Adoption
3. How are we consuming our energy?
    - Electric Power Consumption over time
    - Consumption footprint

HANDS ON