Data management issues in ML Studios

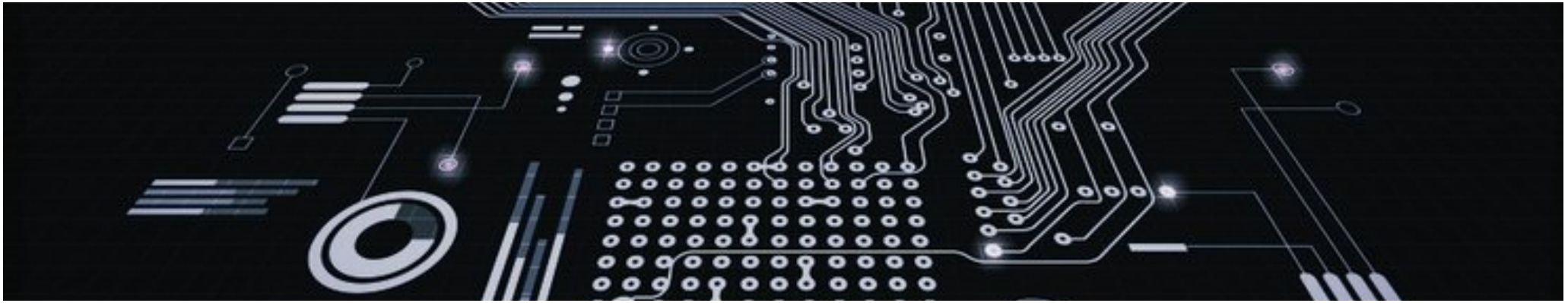
Pipelines, experiments & stacks

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

http://vargas-solar.com/data-ml-studios/

***"Data is everything and everything is data"**, Pythian*

Turning reality phenomena into data thanks to the **Big Data** trend



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**

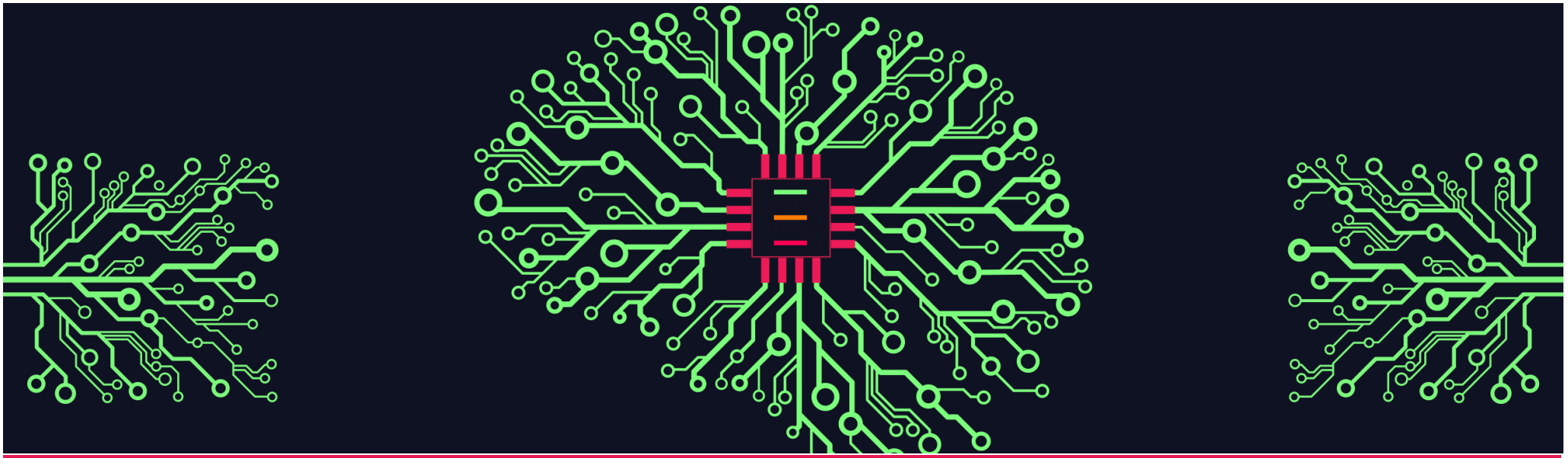
DATIFICATION

Datification is the process of rendering into data aspects of the world that have never been quantified

- Business networks, the lists of books we are reading, the films we enjoy, the food we eat, our physical activity, our purchases, our driving behaviour, and so on.
- Even our thoughts are datified when we publish them on our favourite social network;
- In a not so distant future, your gaze could be datified by wearable vision registering devices.
- At the business level, companies are datifying semi-structured data that were previously discarded: web activity logs, computer network activity, machinery signals, etc
- Reports, e-mails, or voice recordings, are now being stored not only for archive purposes but also to be analysed

Data analysis to take advantage of datafication

- Analytical knowledge is free and most of the crucial algorithms needed are accessible
- The possibility of using rich data to take evidence-based decisions is open to virtually any person or company
- Any individual can analyse huge amounts of data in short periods of time



DATA

DATA SPECTRUM

Querying objective

Data exploration

Hindsight

Insight

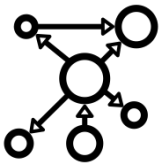
Foresight

Relational & aggregation querying

Information retrieval

Data

Data collections



Data bases

Data Rawness Degree

Raw

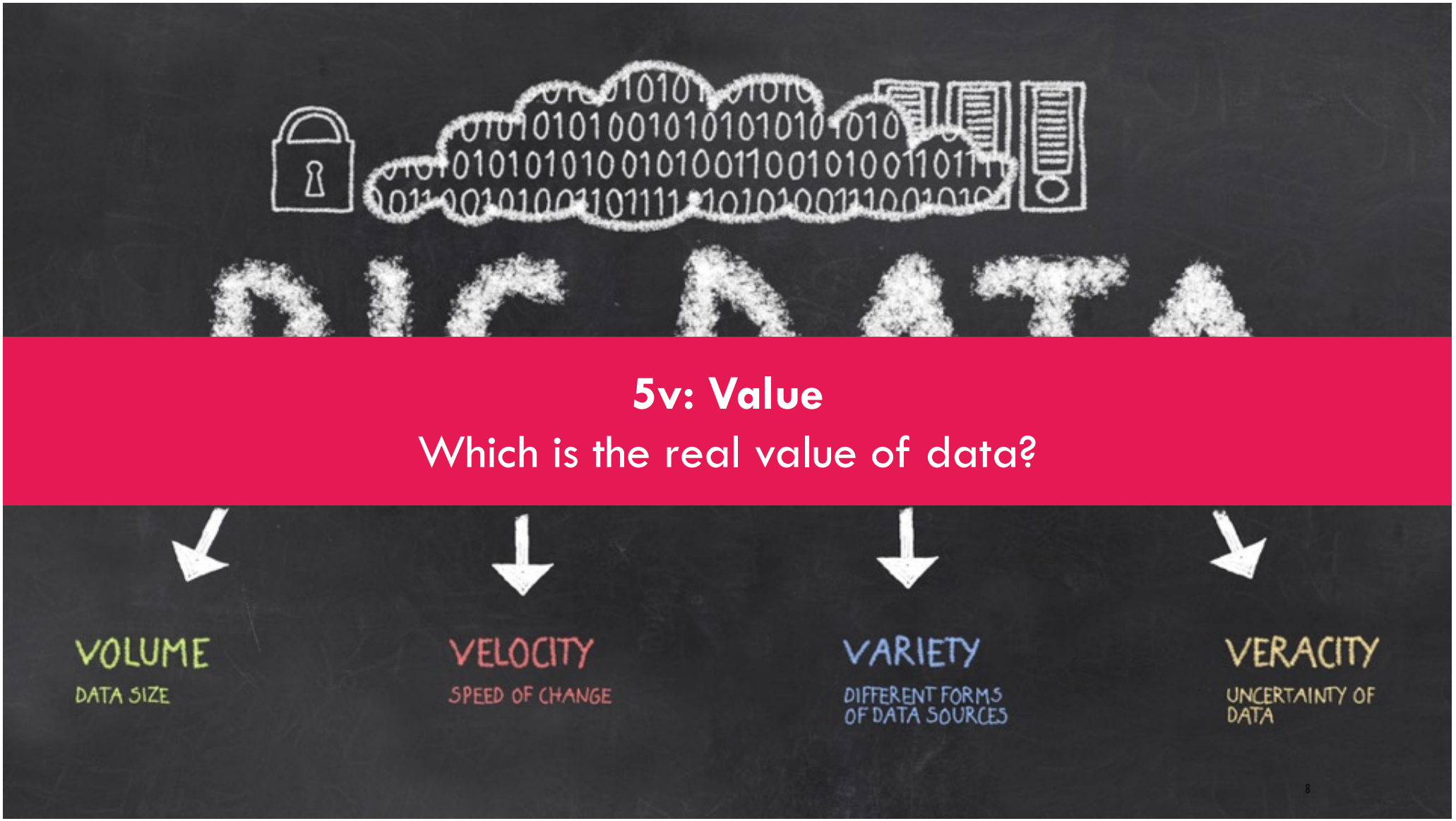
Curated

BIG DATA

- Data collections with characteristics difficult to process on single machines or traditional databases

- A new generation of tools, methods and technologies to collect, process and analyse massive data collections

→ Tools imposing the use of parallel processing and distributed storage


BIG DATA
5v: Value
Which is the real value of data?
VOLUME
DATA SIZE
VELOCITY
SPEED OF CHANGE
VARIETY
DIFFERENT FORMS OF DATA SOURCES
VERACITY
UNCERTAINTY OF DATA

BIG DATA PROPERTIES

3V

- **Volume** (size)
- **Velocity** (production rate)
- **Variety** (data types & format)

4V

- **Variability** (inconsistencies by constant meaning changes)

5V

- **Veracity** (truth and consistency)

…

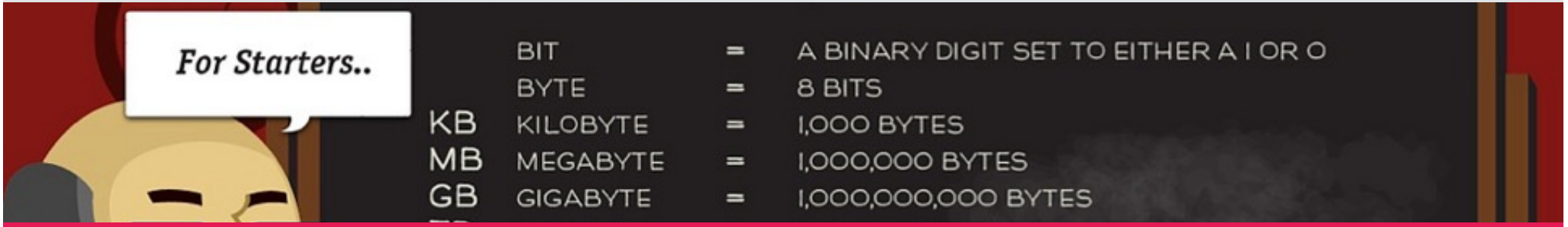
- **Value** (how much information)

10V

V's models [Jagadish 2014]

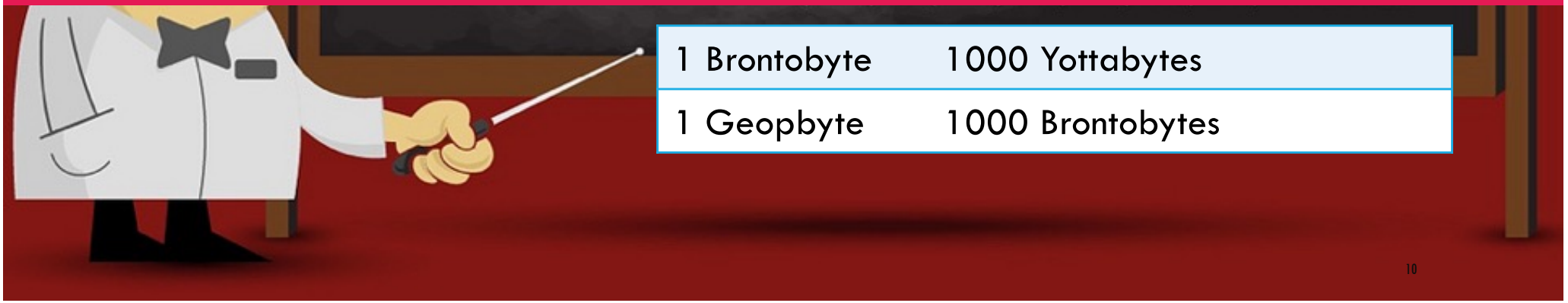
"Big Data can really be very small and not all large datasets are big!"
- Mike 2.0 [Hillard 2012]

HOW BIG IS YOUR DATA ?



Helluva lot of data !!

http://spectrum.ieee.org/computing/software/beyond-just-big-data



1 Brontobyte	1000 Yottabytes
1 Geopbyte	1000 Brontobytes

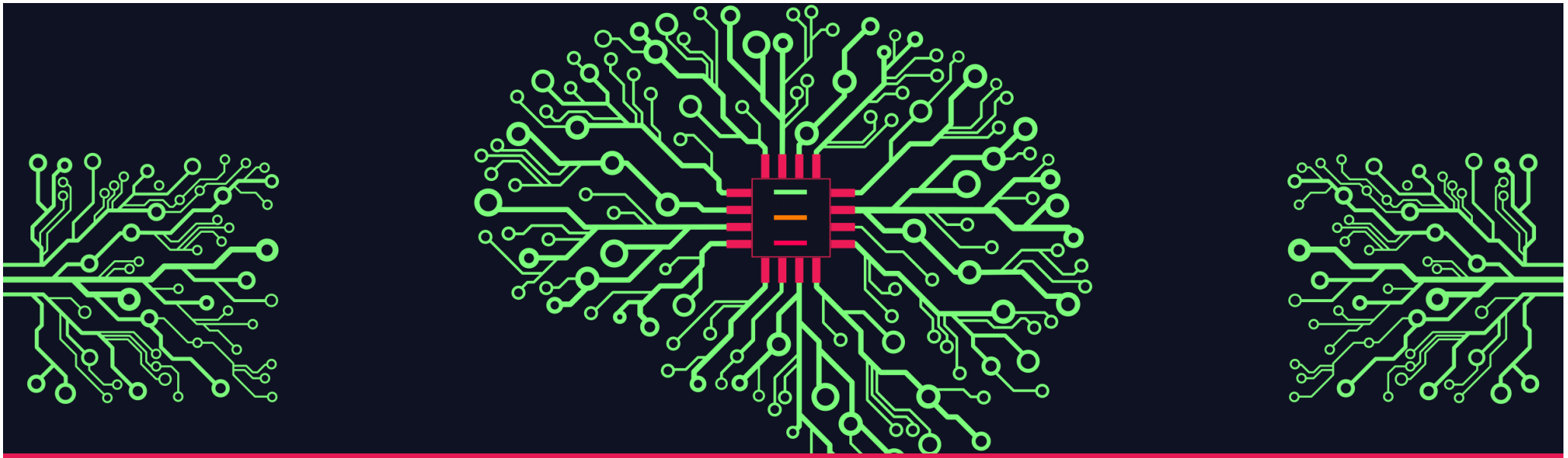
DATA COLLECTIONS/DATABASES

Consumed data:

- different sizes
- quality, uncertainty, ambiguity degree
- evolution in structure, completeness, production conditions, conditions in which data is retrieved
- content, explicit cultural, contextual, background properties
- access policies modification

Conditions of consumption:

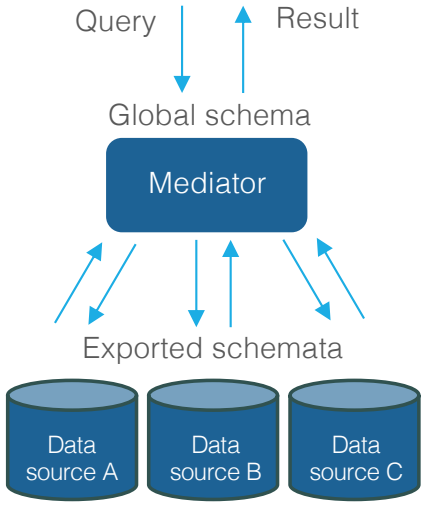
- reproducibility, transparency degree (avoid “software artefacts”)



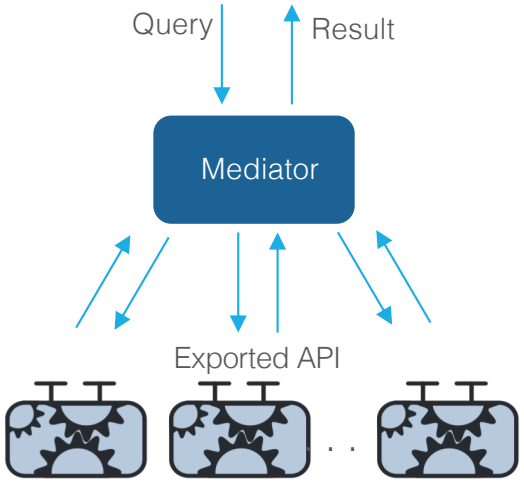
DATA CONSUMPTION

DATA CONSUMPTION PHILOSOPHIES

Oracle approach



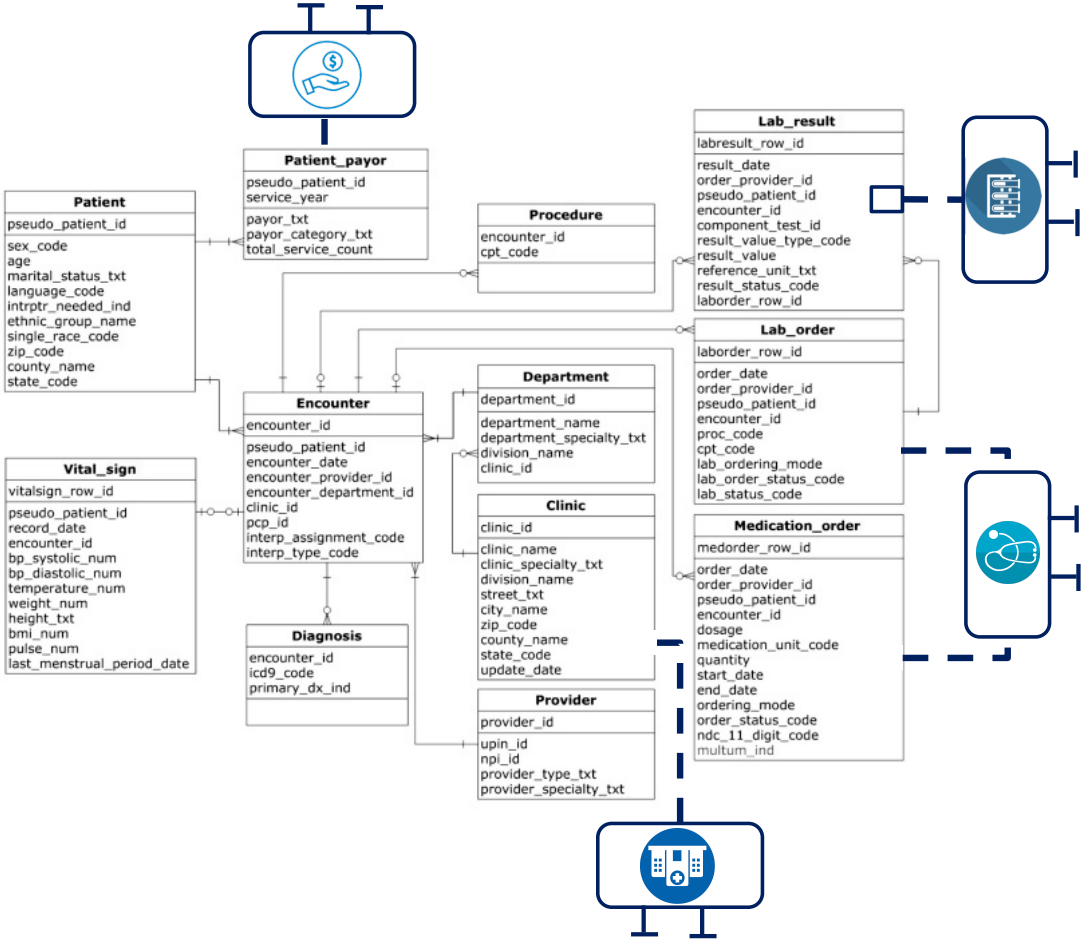
Heterogeneous data sources known in advance



Distributed data services with dynamic binding

- **Data**: model of a mini-world, it is a set of facts structured according to some data model
- **Query**: precisely stated it can include terms, operators (and/or/negation, relational, aggregation), and constraints
- **Result**: collection of items that completely or partially correspond to consumers requirements (precision & recall)

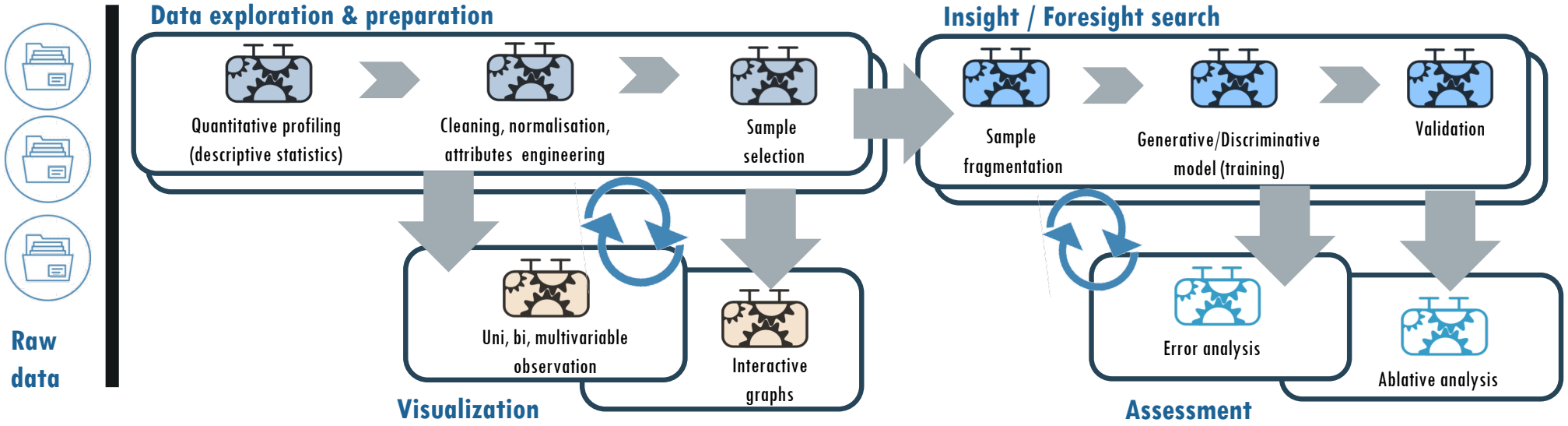
ASKING FACTS ABOUT DIABETES



- Average of lab & medication orders per patient and physician in a given clinic
- Vital signs and lab results used to emit a diagnosis for a given patient
- Number of patients with diabetes followed per clinic

DATA CONSUMPTION PHILOSOPHIES

Diogenes approach

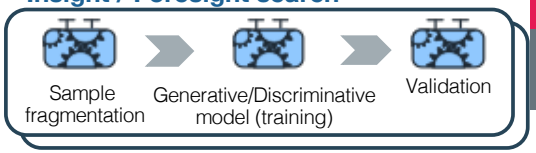


- **Data**: observations of phenomena often described as series of features/attributes
- **Query**: analytics objective (looks for insights or foresights) expressed as a pipeline of operations guided by the conditions and characteristics of the data
- **Result**: a model or prediction with associated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

UNDERSTANDING DIABETES

Predict diabetes given specific variables

Insight / Foresight search



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Data exploration & preparation

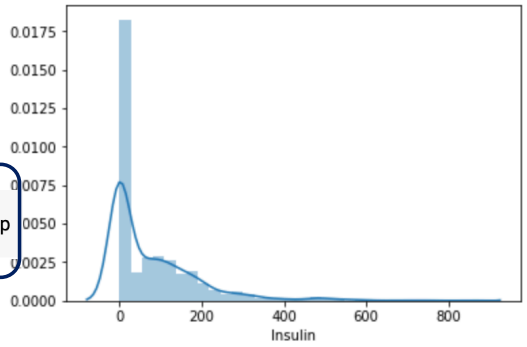
```
#Importing the requi
import numpy as np
import pandas as pd
```

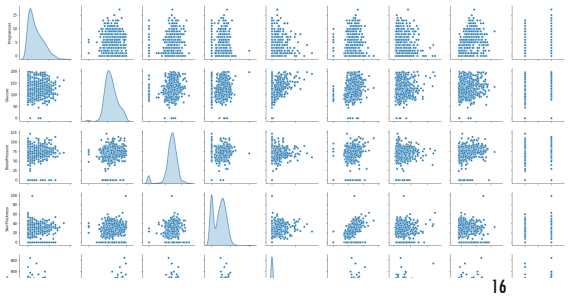
	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

```
Empty DataFrame
Columns: [Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Ou
tcome]
Index: []
Empty DataFrame
Columns: []
Index: []
```

Visualization

```
import seaborn as sns, numpy as np
```

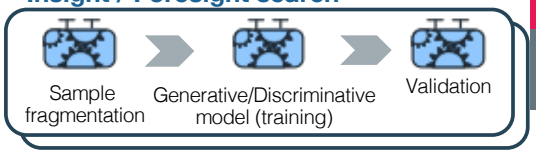




UNDERSTANDING DIABETES

Predict diabetes given specific variables

Insight / Foresight search



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1

Data exploration & preparation

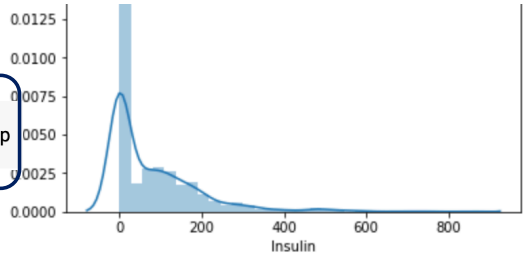
```
#Importing the requi
import numpy as np
```

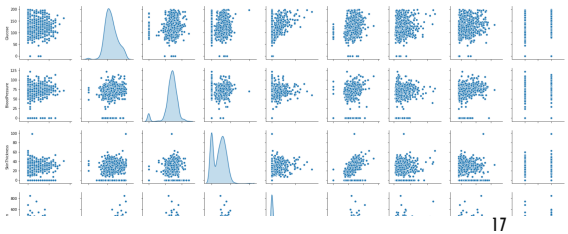
	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00

```
prima test:          Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin  BMI  \
Outcome
0                    500          500      500            500            500      500
1                    268          268      268            268            268      268

         DiabetesPedigreeFunction  Age
Outcome
0                             500  500
1                             268  268
```

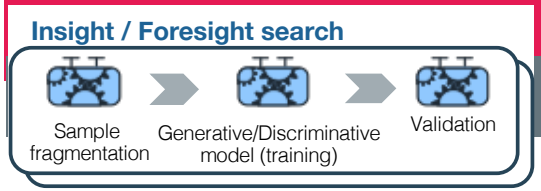
```
import seaborn as sns, numpy as np
```





UNDERSTANDING DIABETES

Predict diabetes given specific variables



Sample fragmentation

```
X = pima_df.iloc[:,0:8]
Y = pima_df.iloc[:,8]
```

```
test_size = 0.30 # taking 70:30 training and test set
seed =1 # Random number seeding for reapeatability of the code

#Splitting Data into train-test where 70% will be used for training #and rest 30% for Testing our m
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=test_size, random_state=seed)
```

```
from sklearn.linear_model import LogisticRegression
import matplotlib.pyplot as plt
#To Split our Data set into training and test data
from sklearn.model_selection import train_test_split
# To calculate accuracy measures and confusion matrix
from sklearn import metrics
```

Generative/Discriminative model (training)

```
# Use method LogisticRegression() import
Logistic_model = LogisticRegression()
#Let's pass our training data sets which
#Our fit method below actually do all th
# So here we get our optimal surface
# which will be our model
Logistic_model.fit(X_train, y_train)
```

```
column_label = list(X_train.columns) # To label all the coefficient
model_Coeff = pd.DataFrame(Logistic_model.coef_, columns = column_label)
model_Coeff['intercept'] = Logistic_model.intercept_
print("Coefficient Values Of The Surface Are: ", model_Coeff)
```

Validation

```
y_predict = Logistic_model.predict(X_test)
print("Y predict/hat ", y_predict)
```

```
Y predict/hat  [0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0
 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 1 0 1 0 1 0 1 0 0 0 0 0 1 1 1 1 1 0
 1 0 1 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0
 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0
 0 0 0 1 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 1 1 1 1 1 1 0 0 1 0 1 0 0 0 0 1 0 1 0
 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0
 0 1 0 0 0 0 0 0 1 0]
```

UNDERSTANDING DIABETES

Predict diabetes given specific variables

Error analysis

Ablative analysis

Assessment

```
#coefficient can be calculated as shown below making use of #model.coef_
column_label = list(X_train.columns) # To label all the coefficient
model_Coeff = pd.DataFrame(Logistic_model.coef_, columns = column_label)
model_Coeff['intercept'] = Logistic_model.intercept_
print("Coefficient Values Of The Surface Are: ", model_Coeff)
```

```
Coefficient Values Of The Surface Are:     Pregnancies   Glucose  BloodPressure  SkinThickness  Insulin
BMI  \
0     0.094379  0.025543      -0.019857      -0.001549 -0.00007  0.056307

   DiabetesPedigreeFunction       Age  intercept
0                  0.389514  0.008663  -5.058877
```

- These values are nothing but the z= 0.094Preg+0.0255Plas+.....+(-5.05)
- Which, get's fed into our Sigmoid function sigmoid, g(z) = 1/(1 + e ^−z).

```
logmodel_score = Logistic_model.score(X_test, y_test)
print("This is how our Model Scored:\n\n", logmodel_score)
```

```
This is how our Model Scored:

 0.7748917748917749
```

```
# Note That In Confusion Matrix
# First argument is true values,
# Second argument is predicted values
# this produces a 2x2 numpy array (matrix)
print(metrics.confusion_matrix(y_test, y_predict))
# Lets run this and see the outcome below:
```

```
[[132  14]
 [ 38  47]]
```

Recall: Recall(For Non_diabetic) = TP/(TP+FN) Here TP = 132, FN= 14

- Recall = 132/(132+14)= 132/146 = 0.90 = 90 %
- Recall(For Diabetic) = TP/(TP+FN) TP = 47, FN = 38
- Recall(For Diabetic)= 47/85 = 0.55 = 55 % ,

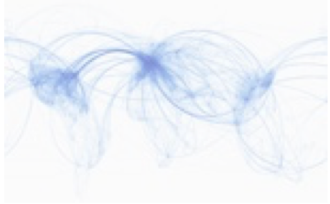
Precision:

- Precision (For Non- Diabetic) = TP/(TP+FP)= 132/ 170 = 0.77 = 77%
- Precision (For Diabetic) = TP/(TP+FP)= 47/ 61 = 0.77 = 77%

- predicted 47 patient to be diabetic(True Positive) and 132 to be non-diabetic(True Negative)
- predicted 14 patient to be diabetic(False Positive) and 38 to be non-diabetic(False Negative)

DATA CENTRIC SCIENCES

Network Science



Computational Science

Social Data Science

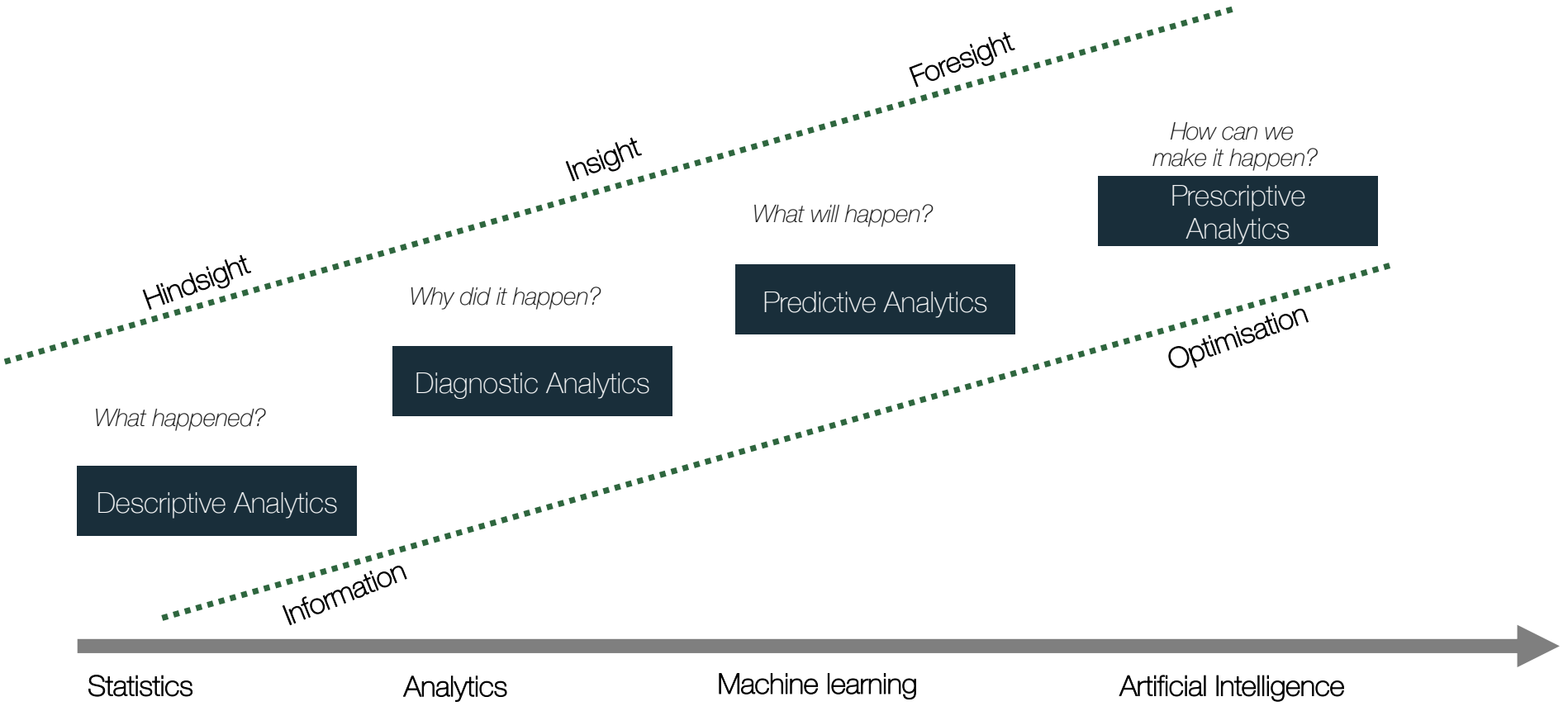
Digital humanities

Develop methodologies weaving data management, greedy algorithms, and programming models that must be tuned to be deployed in different target computer architectures

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


DATA SCIENCE
MACHINE LEARNING
ANALYTICS
PREDICTIVE
PROJECTS
SOCIAL MEDIA
SERVICES
WWW
BIG DATA
CONTENT
COMPUTING
TECHNOLOGY
PROMOTION
PROCESSING
CONSUMER
ORGANIZATION
PLANNING
EVENTS
PROGRAMMING
CODING
SOFTWARE
MODELS
E-MARKETING
COMMUNICATION
BRANDING
CONSUMER DEMAND
WEB MARKETING
DATA MINI
VISION
ENGINEERING
RESEARCH
KDD
WEB DEV
STRATEGY
WORLDWIDE
VISUALIZATION
PRO
PRICING
SEGMENTATION
SOCIAL NETWORKS
BIG DATA
WEB SERVICES
MATHS
PATTERN
ENGINEERING
PLANNING
MEDIA
STATISTICS
BIG DATA
MOBILE
INFORMATION
SERVICE
PROJECTS
INFORMATION
SOLUTIONS

EXPERIMENTS OBJECTIVE



DATA SCIENCE

The representation of complex environments by rich data opens up the possibility of applying all the scientific knowledge regarding how to infer knowledge from data

Definition:

- Methodology by which actionable insights can be inferred from data
- Complex, multifaceted field that can be approached from several points of view: ethics, methodology, business models, how to deal with big data, data engineering, data governance, etc.

Objective:

- Production of beliefs informed by data and to be used as the basis of decision making
- N.B. In the absence of data, beliefs are uninformed and decisions are based on best practices or intuition

DATA SCIENCE STRATEGIES

Probing reality: data can be gathered by passive or by active methods.

- In the latter case, data represents the response of the world to our actions.
- *What is the best button size and colour? The best answer can only be found by probing the world.*

Pattern discovery: datified problems can be analysed automatically to discover useful patterns & natural clusters that can greatly simplify their solutions.

- *Profile users is a critical ingredient in important fields as programmatic advertising or digital marketing.*

DATA SCIENCE STRATEGIES

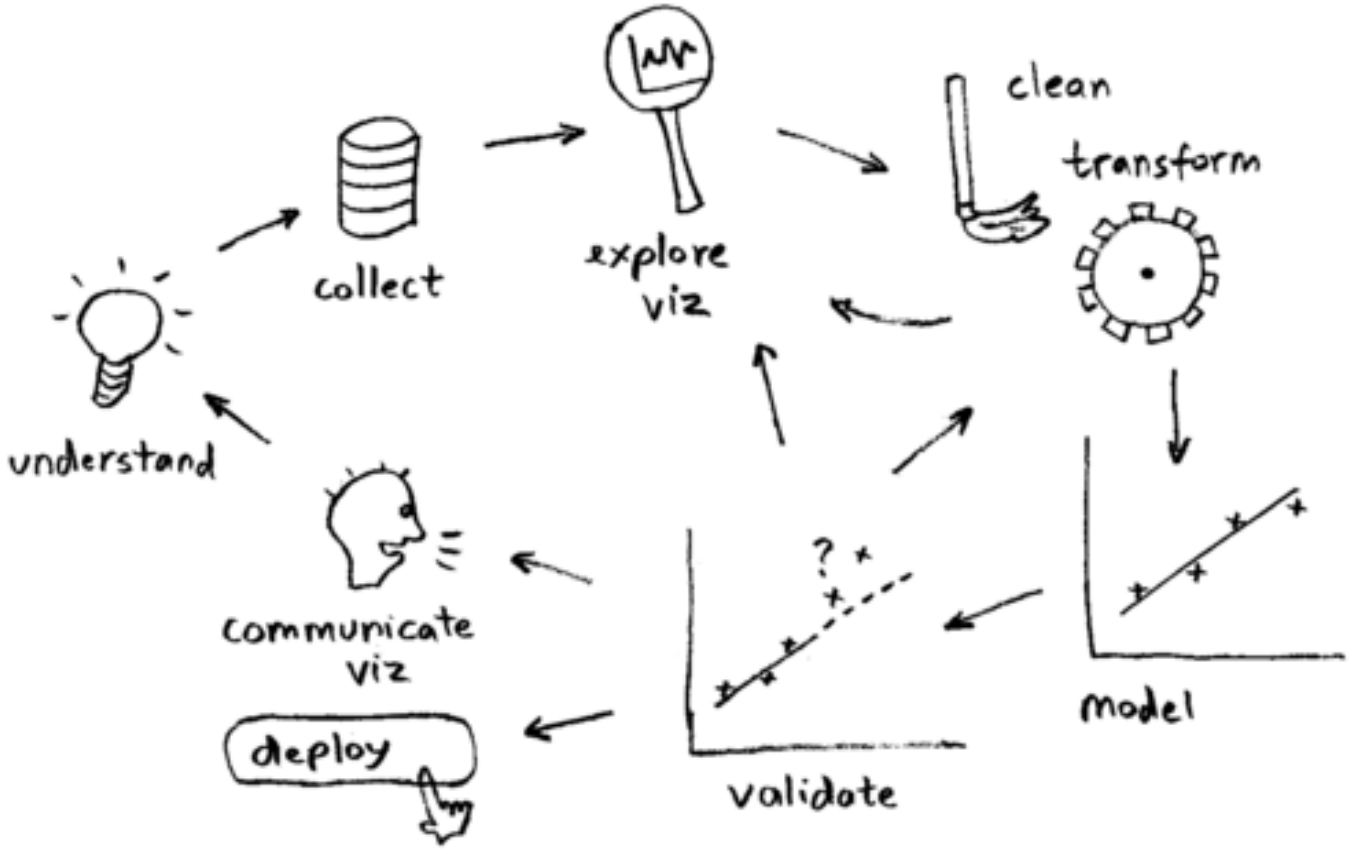
Predicting events: predictive analytics allows decisions to be taken in response to future events.

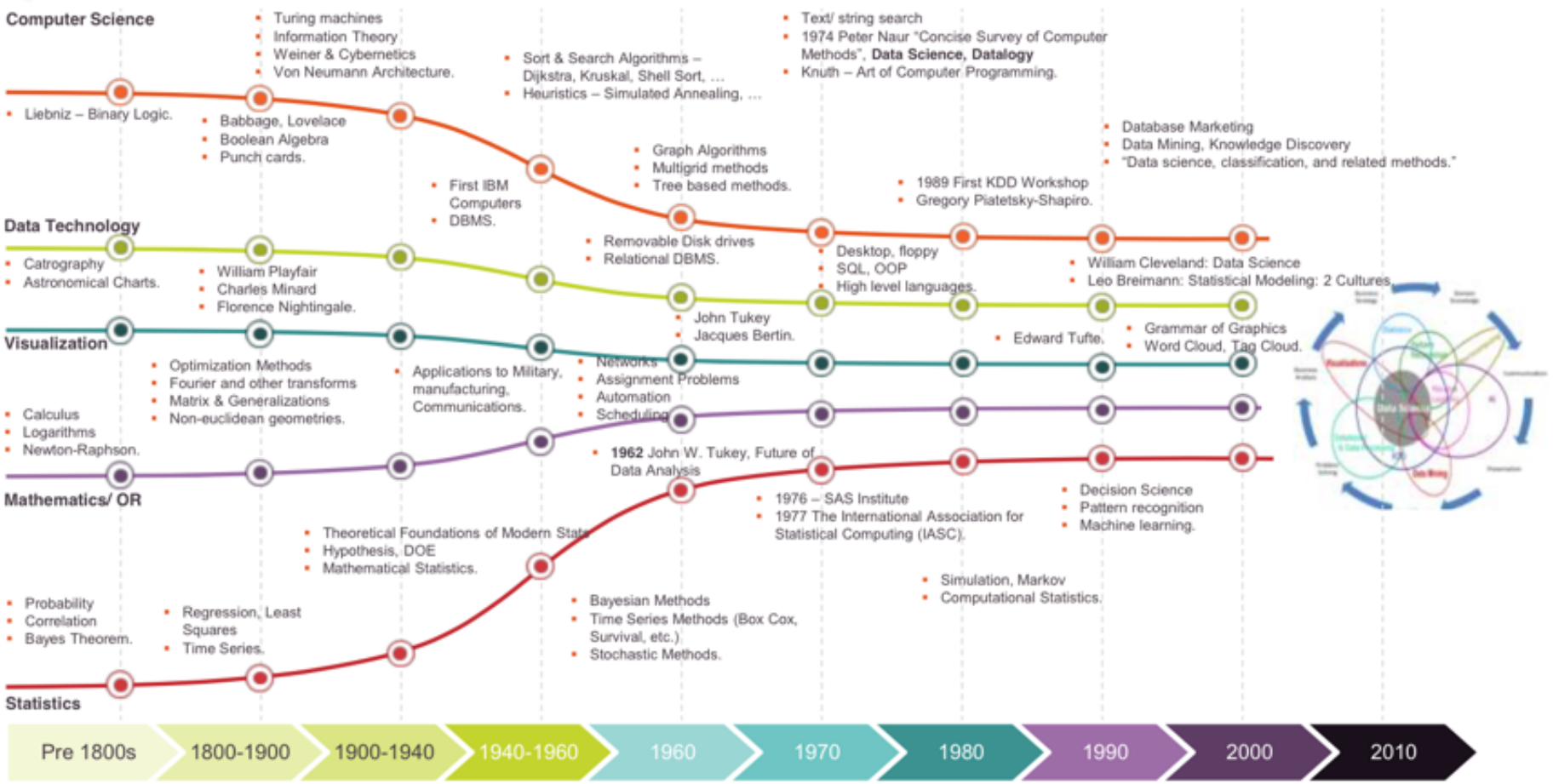
- *For example, optimize the tasks planned for retail store staff during the following week, by analysing data such as weather, historic sales, traffic conditions, etc.*

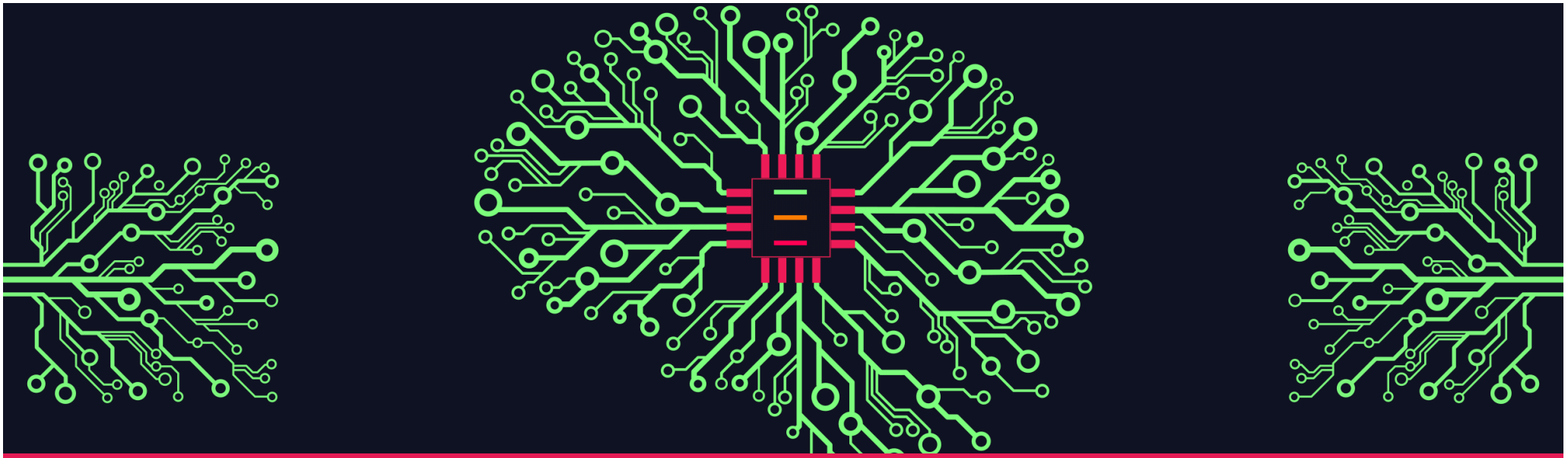
Understanding people in the world: understanding natural language, computer vision, psychology and neuroscience.

- In order to make optimal decisions, it is necessary to know the real processes that drive people's decisions and behaviour.

DATA SCIENCE WORKFLOW




Computer Science
Liebniz – Binary Logic.
Turing machines
Information Theory
Weiner & Cybernetics
Von Neumann Architecture.
Babbage, Lovelace
Boolean Algebra
Punch cards.
Sort & Search Algorithms – Dijkstra, Kruskal, Shell Sort, …
Heuristics – Simulated Annealing, …
First IBM Computers
DBMS.
Graph Algorithms
Multigrid methods
Tree based methods.
Text/ string search
1974 Peter Naur "Concise Survey of Computer Methods", Data Science, Datalogy
Knuth – Art of Computer Programming.
1989 First KDD Workshop
Gregory Piatetsky-Shapiro.
Database Marketing
Data Mining, Knowledge Discovery
"Data science, classification, and related methods."
Data Technology
Catrography
Astronomical Charts.
William Playfair
Charles Minard
Florence Nightingale.
Removable Disk drives
Relational DBMS.
Desktop, floppy
SQL, OOP
High level languages.
William Cleveland: Data Science
Leo Breimann: Statistical Modeling: 2 Cultures
Visualization
John Tukey
Jacques Bertin.
Edward Tufte.
Grammar of Graphics
Word Cloud, Tag Cloud.
Optimization Methods
Fourier and other transforms
Matrix & Generalizations
Non-euclidean geometries.
Applications to Military, manufacturing, Communications.
Networks
Assignment Problems
Automation
Scheduling
Calculus
Logarithms
Newton-Raphson.
Mathematics/ OR
1962 John W. Tukey, Future of Data Analysis
1976 – SAS Institute
1977 The International Association for Statistical Computing (IASC).
Decision Science
Pattern recognition
Machine learning.
Theoretical Foundations of Modern Stats
Hypothesis, DOE
Mathematical Statistics.
Simulation, Markov
Computational Statistics.
Probability
Correlation
Bayes Theorem.
Regression, Least Squares
Time Series.
Bayesian Methods
Time Series Methods (Box Cox, Survival, etc.)
Stochastic Methods.
Statistics
Pre 1800s
1800-1900
1900-1940
1940-1960
1960
1970
1980
1990
2000
2010



THIS LECTURE

OBJECTIVES

Introduce & guide to develop data science pipelines for studying efficient enactment strategies to explore problems that can go beyond known analytics scales & that can contribute to perform continuous on-line data centric sciences experiments

SPECIFIC OBJECTIVES

- Study the architectures and environments that ease the deployment of data science solutions.
- Teach how to apply data science method & techniques for concrete experiments running on top of target architectures.

LEARNING OUTCOMES

Deploy Data Science experiments on target environments, exhibiting the pipelines behind and execution strategies to be considered for running experiments at scale:

- Understand theoretically and technically the steps of a general data science process.
- Apply tools for executing data science pipelines.
- Learn how to make decisions on the data analytics techniques to apply according to the data properties and the analytics objective.
- Know how to define strategies to scale analytics solutions for dealing with Big Data settings using different computing resources.

CONTENT

Introduction

- Data centric sciences: Principles and common aspects
- Digital data collections: Characteristics and properties
- Data science: Big data, data analytics algorithms & tools

From centralized to high scale WIDES, data science laboratories to artificial intelligences studios

- In house data analytics environments: Jupyter
- Targeting large scale: Zeppelin
- Data science virtual machines: cloud solutions
- Data science labs: CoLab, Kaggle, Azure Notebooks
- Artificial intelligence execution environments & studios: tensor, café, Azure ML studio

Designing experiments environments

- Data labs: data collections, quality, and profiling
- Architectural settings: from in house to large scale experiments

Data engineering

- Data formats, transformations, distribution
- Studying data quality
- Statistical properties
- Techniques for adjusting data and building data samples
- An overview of applied mathematics to machine learning

Designing data science pipelines

- Linear regression
- Logistic regression
- Clustering
- Graph processing: network science

CONTACT & ADVISING

Dr. Genoveva Vargas-Solar
French Council of Scientific Research, LIG lab

ADVISING & QUESTIONS: genoveva.vargas@imag.fr