

# Big Data in the Entreprise:

## Lesson Learned

(in french speaking Switzerland)

Alexandre Masselot  
OCTO Technology Switzerland

@OCTOSuisse @alex\_mass  
Geneva, septembre 14, 2017



# Big Data in the Entreprise:

## Lesson Learned

(in french speaking Switzerland)

Alexandre Masselot  
OCTO Technology Switzerland

@OCTOSuisse @alex\_mass  
Geneva, septembre 14, 2017



WE BELIEVE THAT *information technology*

---

**TRANSFORMS COMPANIES**

---

WE KNOW THAT THE *greatest achievements*

---

ARE THE RESULT OF *shared* KNOWLEDGE

---

**AND THE JOY OF WORKING + TOGETHER**

---

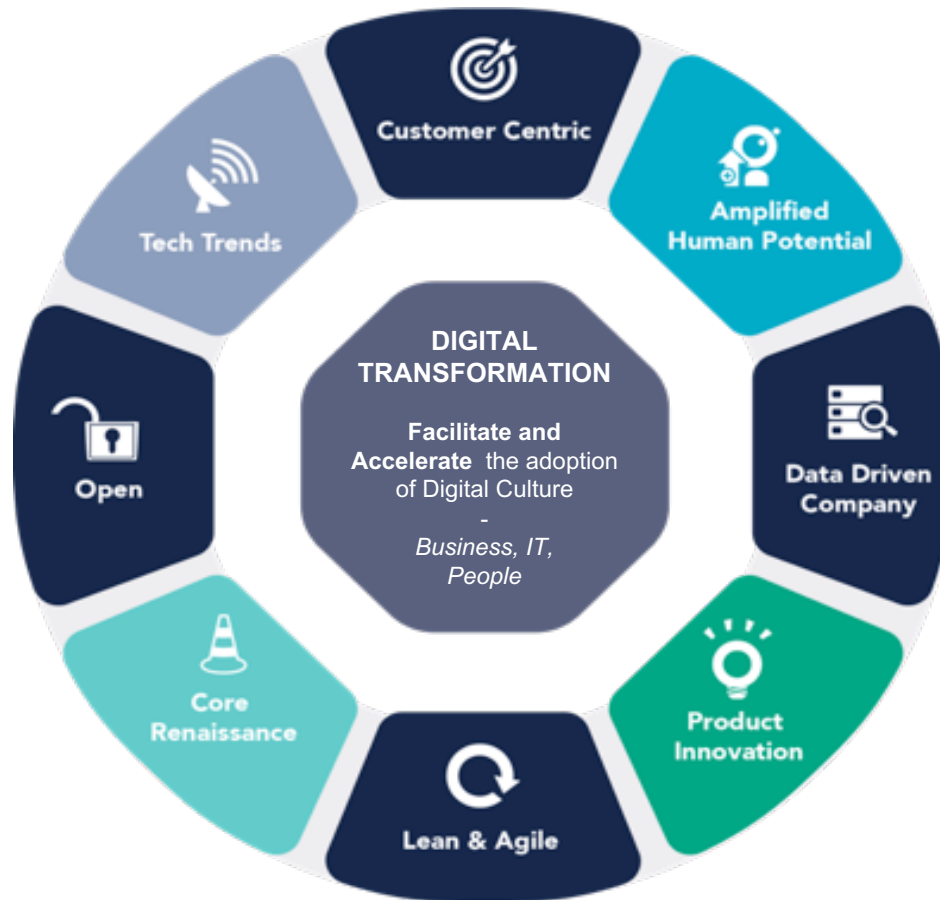
**WE ARE ALWAYS *on the lookout***

---

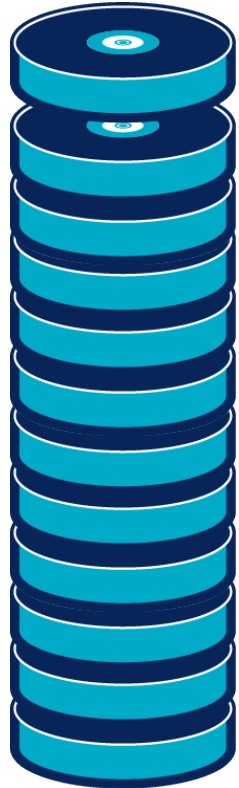
**FOR BETTER *ways* OF WORKING**

---

---

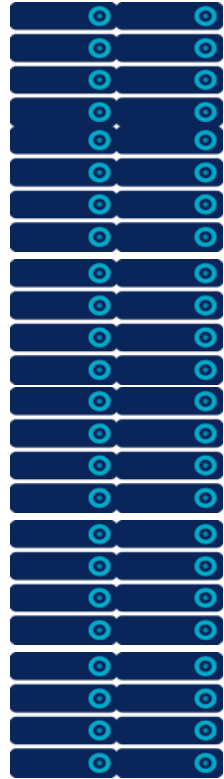


# BIG DATA @ OCTO : THE NUMBERS



**250**

TB, the biggest volume of data analyzed by OCTO's data scientists

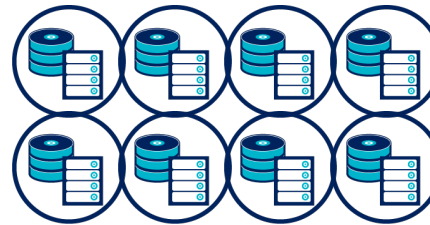


**800**

cores, the biggest Hadoop cluster built by OCTO

Is the number of Big Data projects at OCTO in the past 12 months

**40**



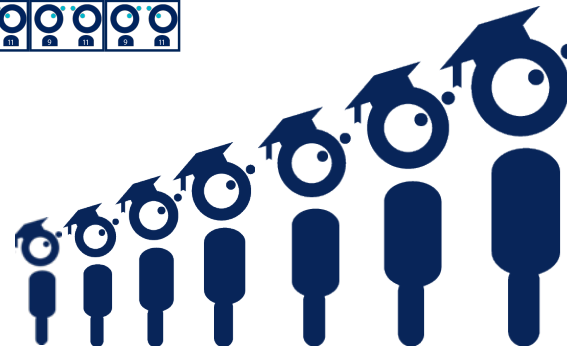
**850**

TB, the biggest volume of distributed storage on a single project



**16**

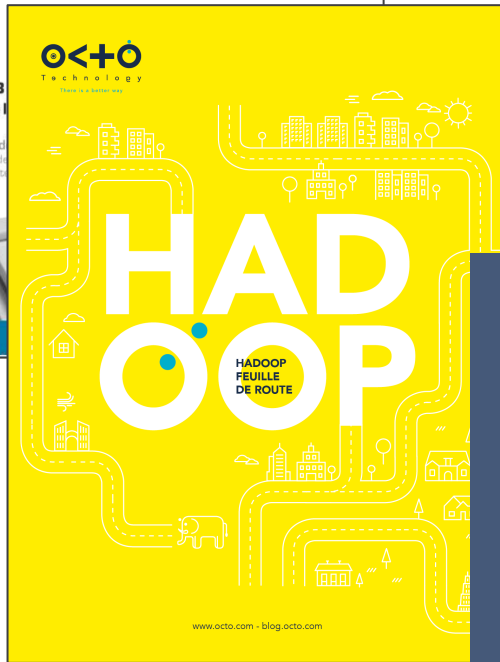
The number of active partnerships with major Big Data actors



**>20**

The number of OCTO certified on the Hadoop platform

# BIG DATA @ OCTO: PUBLICATIONS





## OCTO Folks Work Hard, Play Hard

1<sup>st</sup>

- Caisse de dépôts - *score de délivrance d'un brevet européen*

 **datascience.net**

2&4

- Argus - *prédiction du prix de vente de véhicules d'occasion*

 **datascience.net**

3<sup>rd</sup>

- SNCF - *prédiction de la fréquentation des gares en Ile de France*

 **datascience.net**

6<sup>th</sup>

- Imperial College London - *Loan Default Prediction*

**kaggle**

13<sup>th</sup>

- Allstate – *purchase prediction challenge*

**kaggle**

2<sup>nd</sup>

- Tradeshift – *Text classification*

**kaggle**

5<sup>th</sup>

- Microsoft - *Malware classification*

**kaggle**

**OCTO, there is a better way to learn, recruit and have fun!**

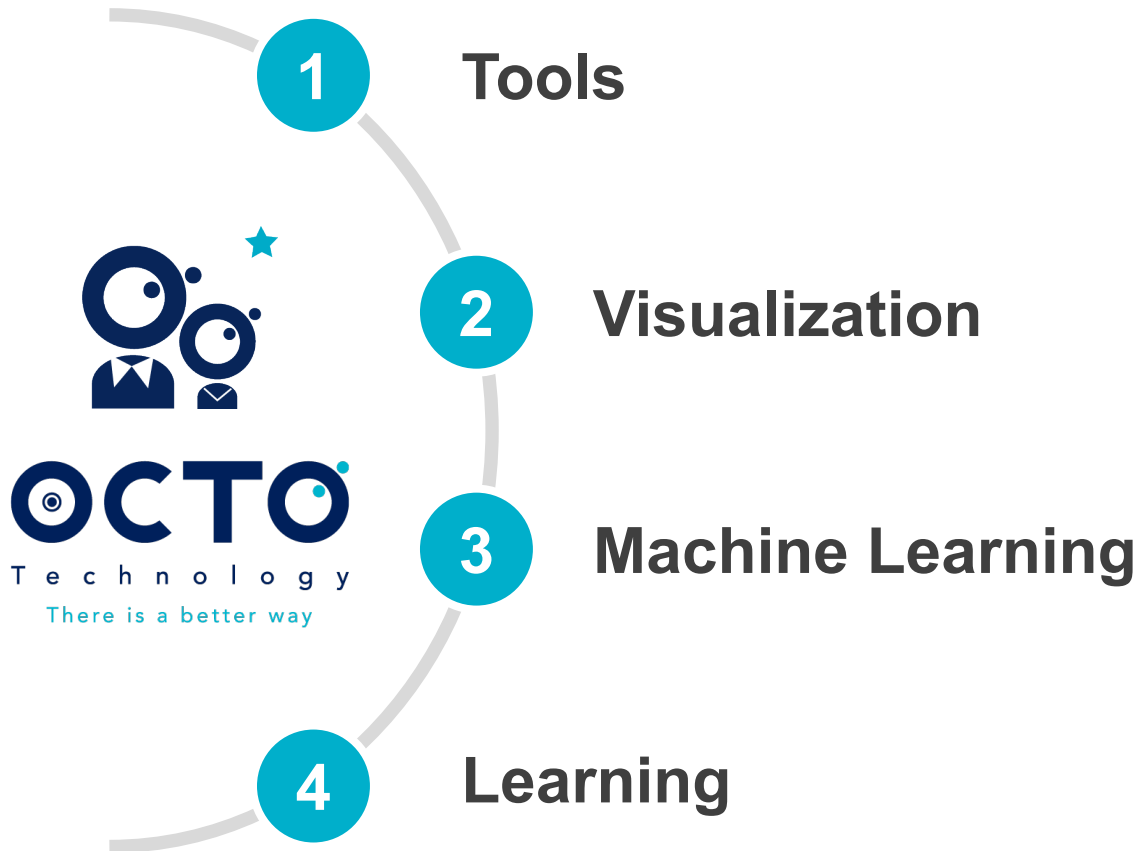




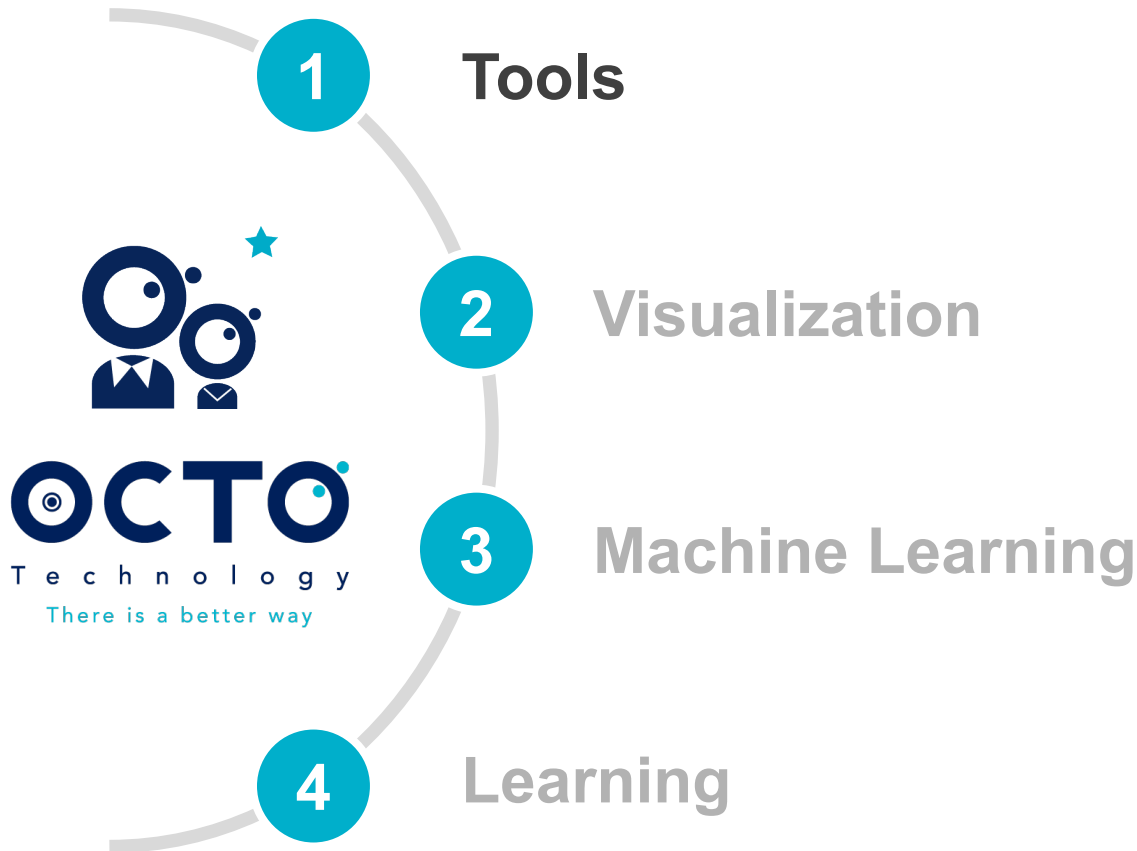
*Afterwork* **BigData**



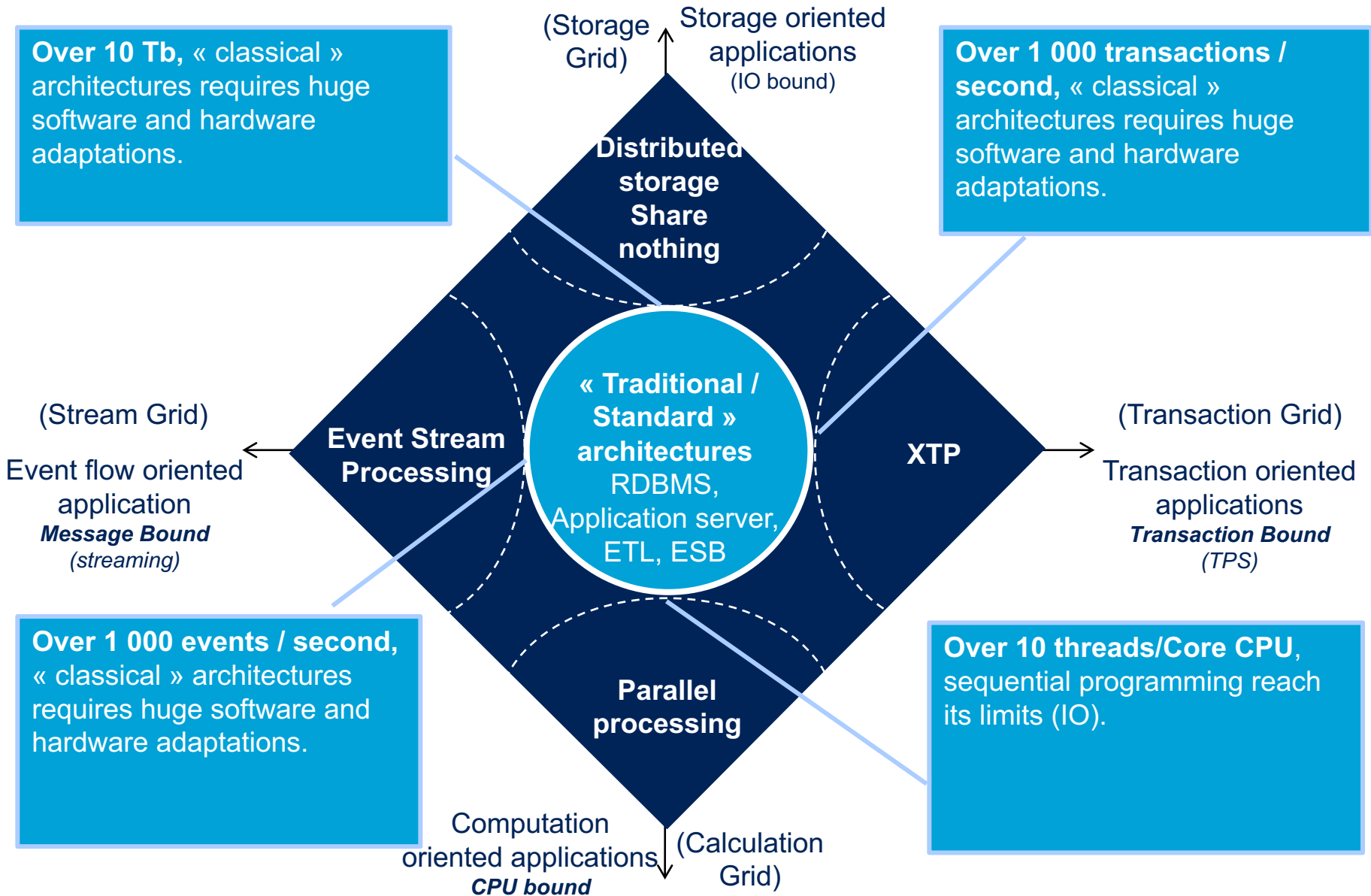
# FOUR PILLARS OF BIG DATA



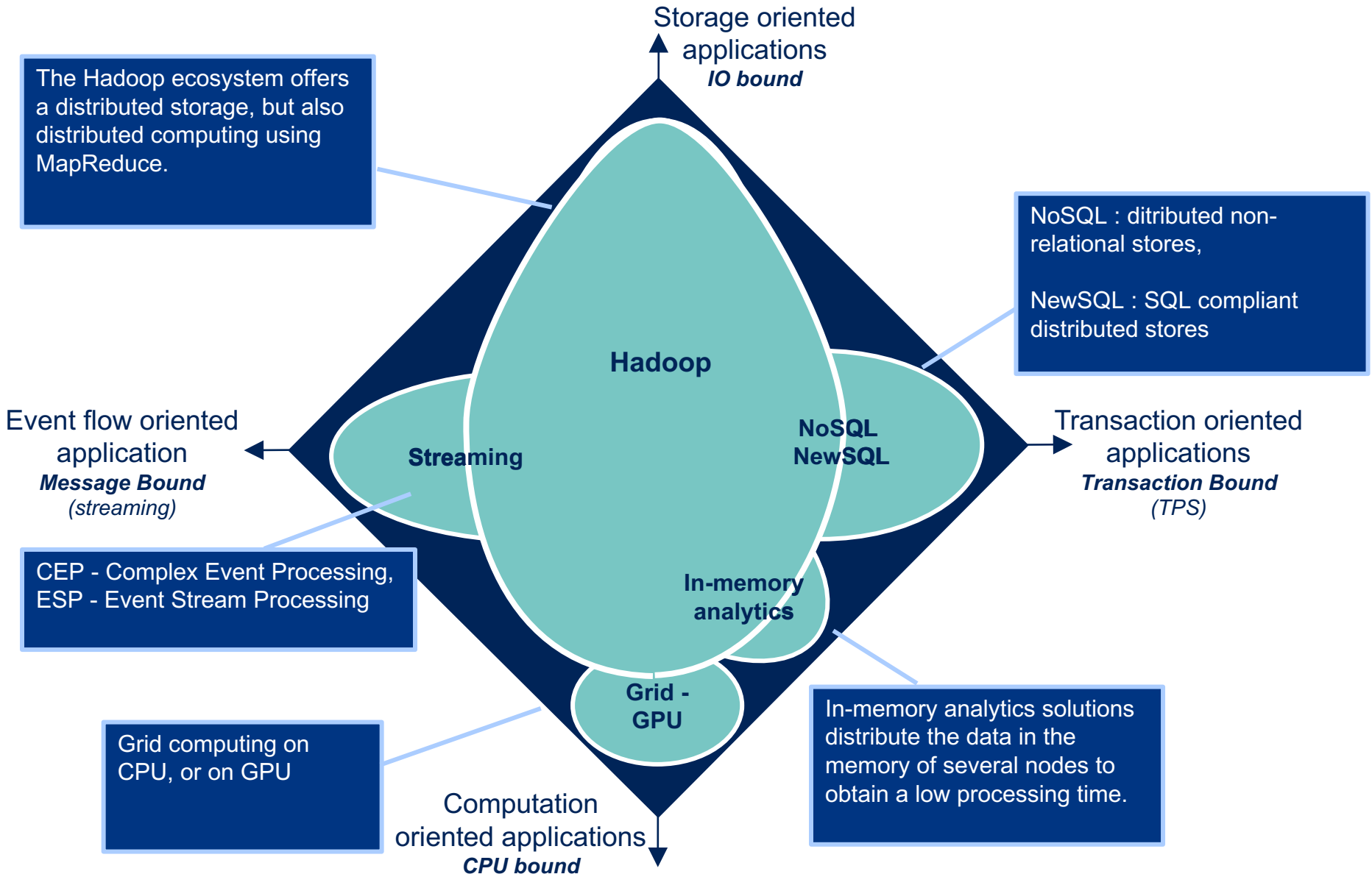
# FOUR PILLARS OF BIG DATA



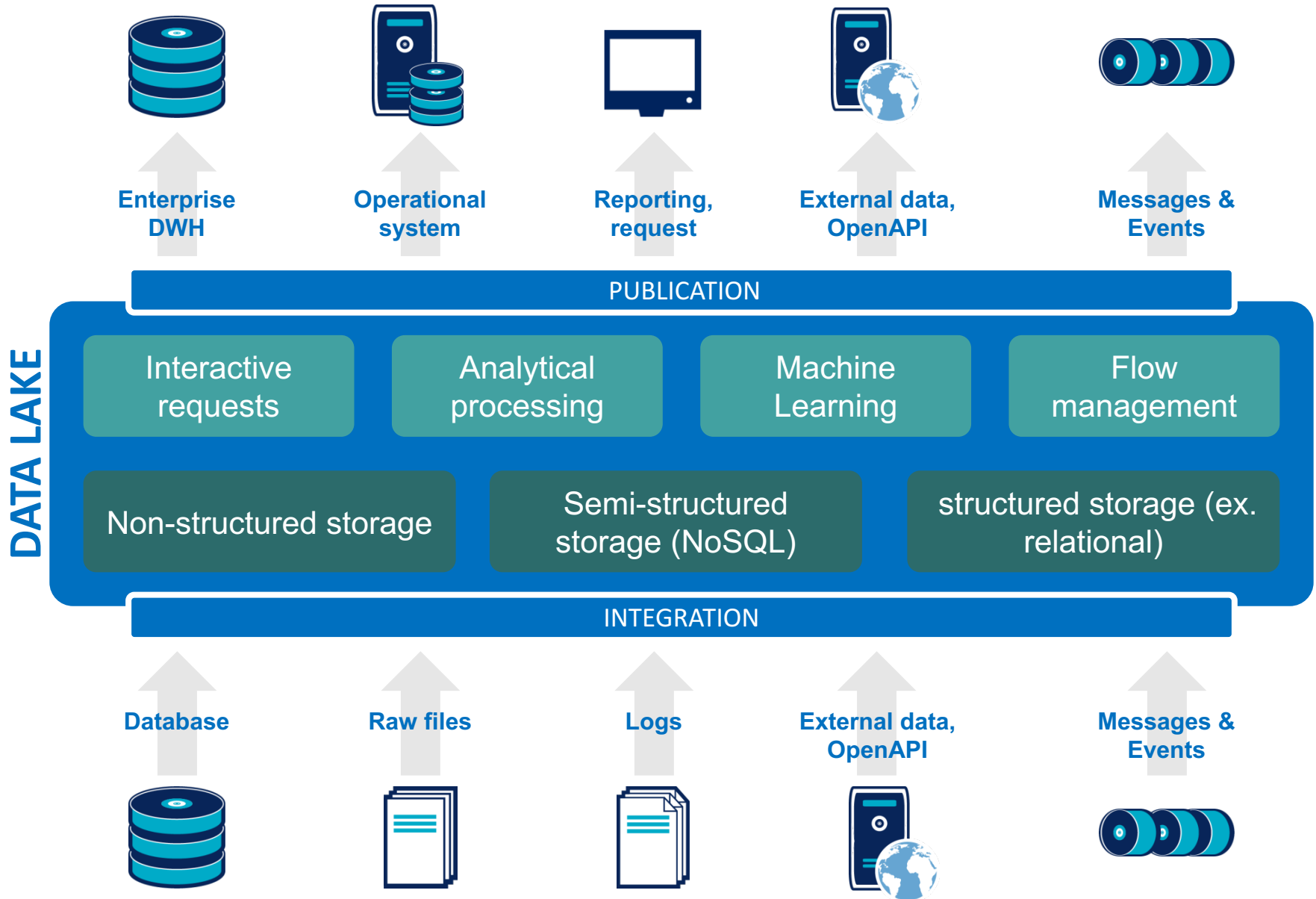
# LIMITATIONS OF TRADITIONAL ARCHITECTURES



# BIG DATA - EMERGING FAMILIES



# ANTI-PATTERN #1: "I want a Data Lake"



# ANTI-PATTERN #1: "I want a Data Lake"



You do not need to store/compute petabyte of data...



From Datalake...



... to Dataswamp

Business

JUL 1, 2016 @ 10:00 AM 921 VIEWS

## Is Your Data Lake Destined To Be Useless?



**TeradataVoice**

Big Data to Data-Driven Insights [FULL BIO](#) ▾



**Stephen Brobst**, Teradata

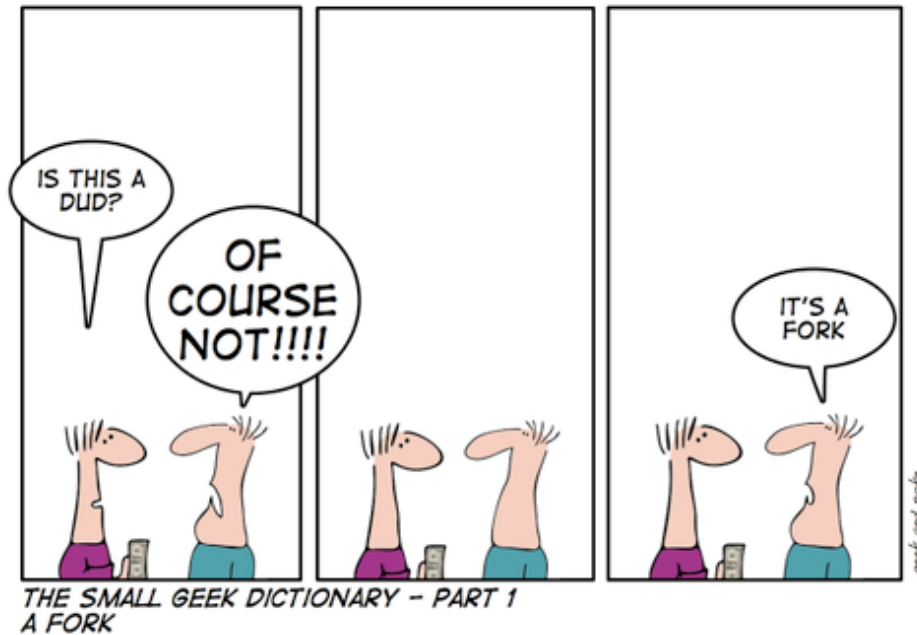
*By Stephen Brobst*

**Gartner** has predicted that through 2018, 90% of deployed data lakes will be useless. That word "useless" should grab your attention. It's worse than a "failed





## ANTI PATTERN #3: FORKING THE SOLUTION

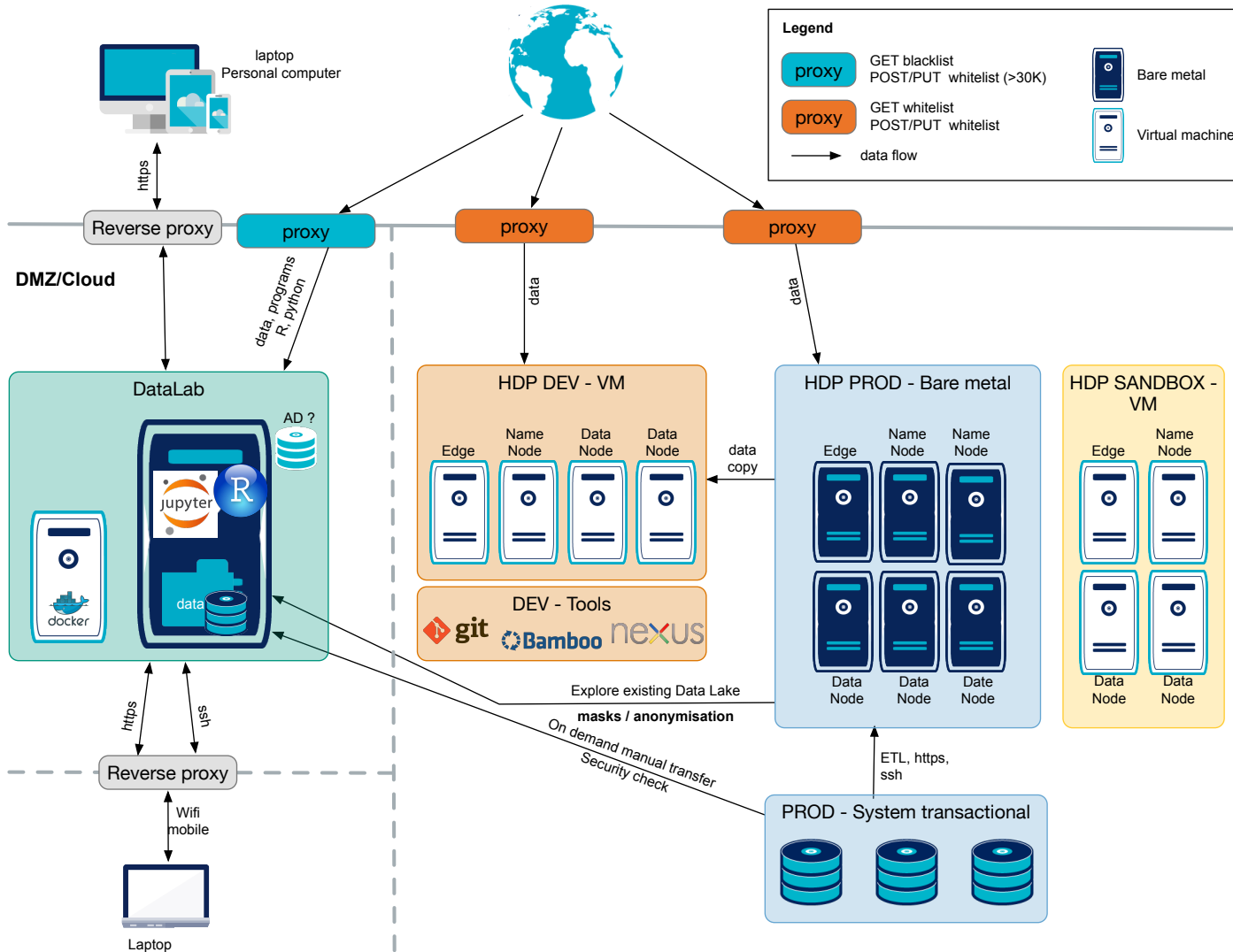


*“The NIH syndrome (Not Invented Here) is a disease.”*  
Linus Torvalds

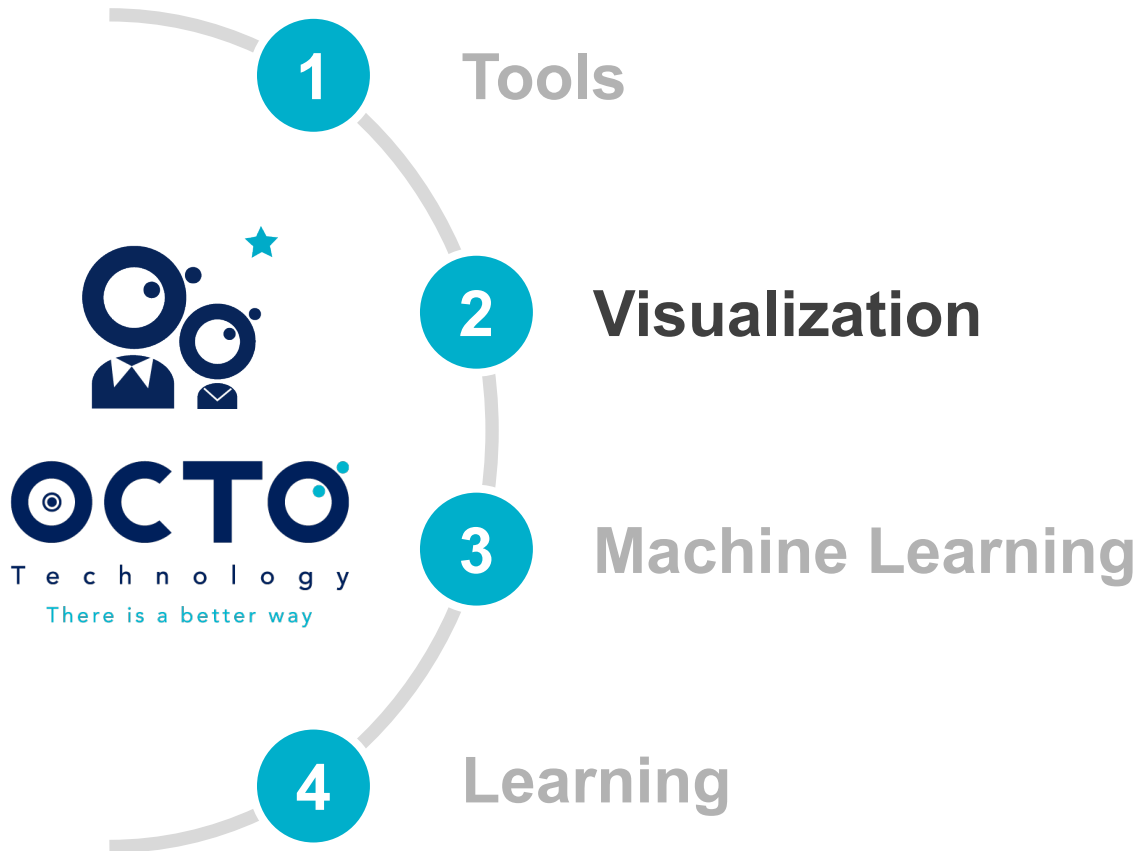




# ANTI PATTERN #4: OVER ARCHITECTING FROM DAY 1



# FOUR PILLARS OF BIG DATA



## ANOTHER PERSPECTIVE ON VISUALIZATION

Who said that? When?

*“There is danger in giving too much information to executives of small brain capacity.”*

*“As a cathedral is to its foundations, so is an effective presentation of the fact to the data.”*

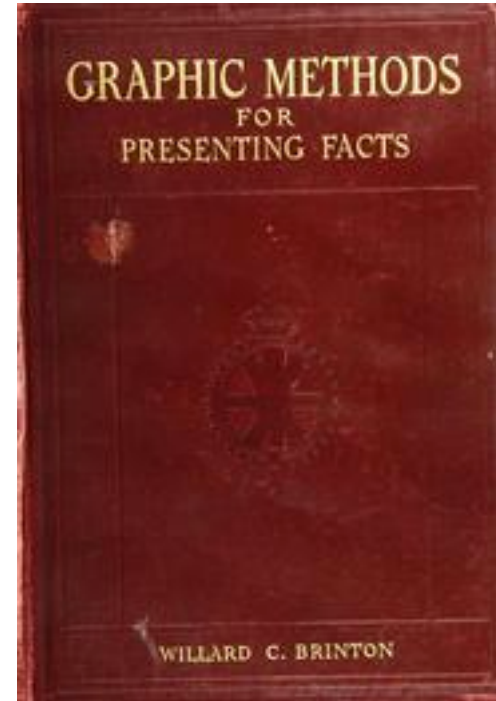
*“The answer is that the executive of the future will be forced on the analysis of facts which have been collected and arranged for his instantaneous and continuous use.”*



## ANOTHER PERSPECTIVE ON VISUALIZATION



**Willard C. Brinton**

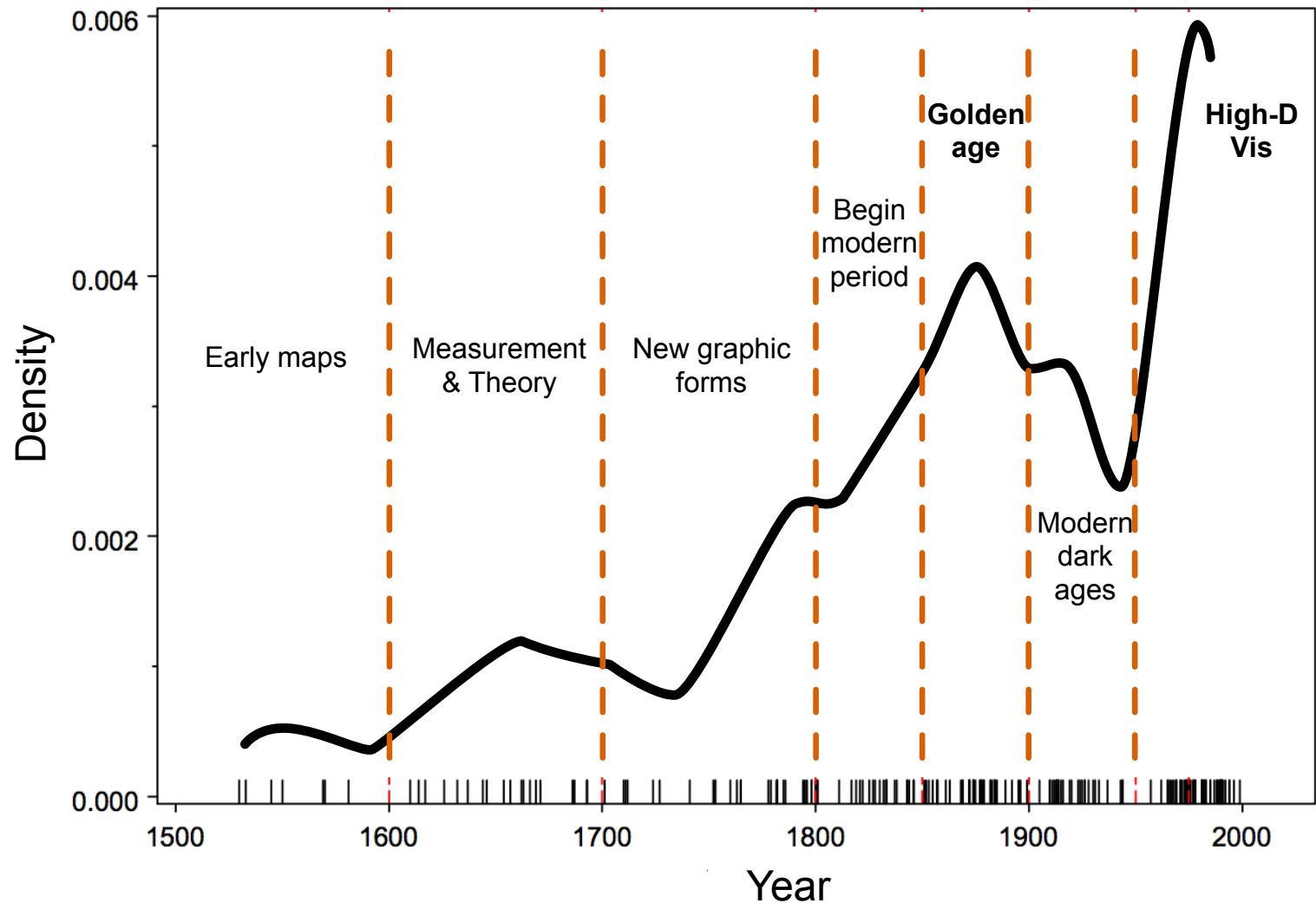


**1914**

[100yrsofbrinton.tumblr.com](http://100yrsofbrinton.tumblr.com)



# Graphics Milestones: Time course of developments



The distribution of milestone items over time, shown by a rug plot and density estimate.  
Michael Friendly et Daniel J. Denis. <https://www.researchgate.net/publication/221649568>



## ANOTHER PERSPECTIVE ON VISUALIZATION



*Photo by the International News Service*

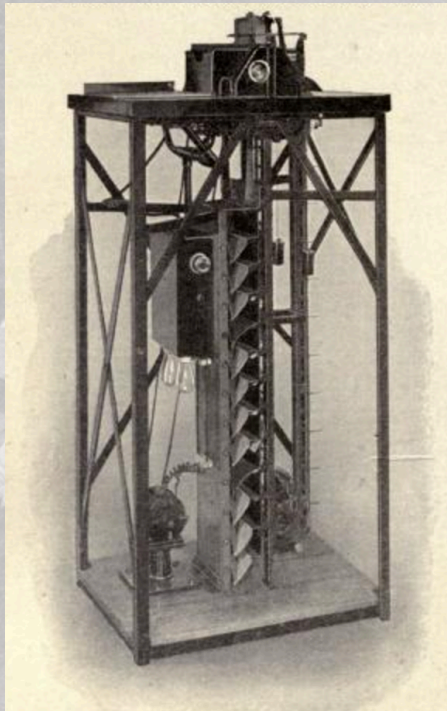
**Fig. 238. Statistical Exhibits in the Municipal Parade by the Employees of the City of New York, May 17, 1913**

Many very large charts, curves and other statistical displays were mounted on wagons in such manner that interpretation was possible from either side of the street. The Health Department, in particular, made excellent use of graphic methods, showing in most convincing manner how the death rate is being reduced by modern methods of sanitation and nursing

[100yrsofbrinton.tumblr.com](http://100yrsofbrinton.tumblr.com)

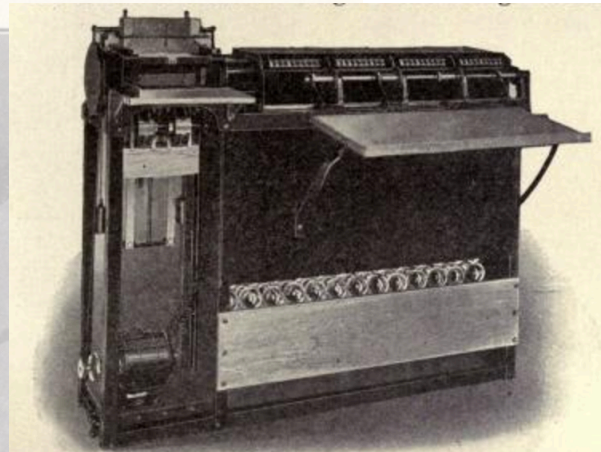


# THE PREVIOUS BIG DATA REVOLUTION (END 1800s)



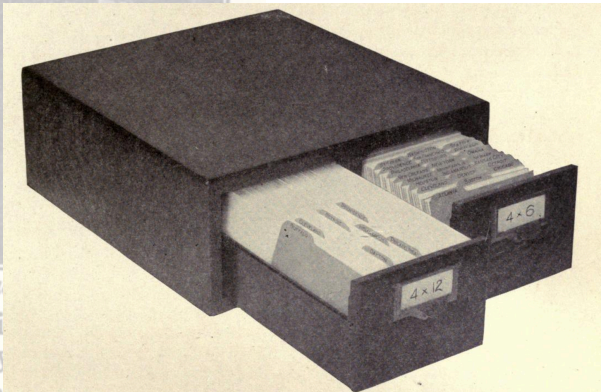
*Courtesy of the Tabulating Machine Company*  
**Fig. 229. Hollerith Card-Sorting Machine**  
 Suitable for Use by Corporations for  
 Statistical Work Relating to Sales,  
 Costs, Etc.

This machine will sort about 12,000 cards per hour, and place in the right compartment all cards having the hole punched in the same position in the particular column for which the sorting is done



*Courtesy of the Tabulating Machine Company*  
**Fig. 230. Hollerith Tabulating Machine for Totalling**  
 the Data Contained on Punched Cards

The machine illustrated has four counters, permitting the simultaneous taking off of the data contained under four different headings on the punched card. The sorted cards are placed at the left of the tabulating machine and run through at the rate of about 3,000 per hour. Totals are read from the dials shown at the right



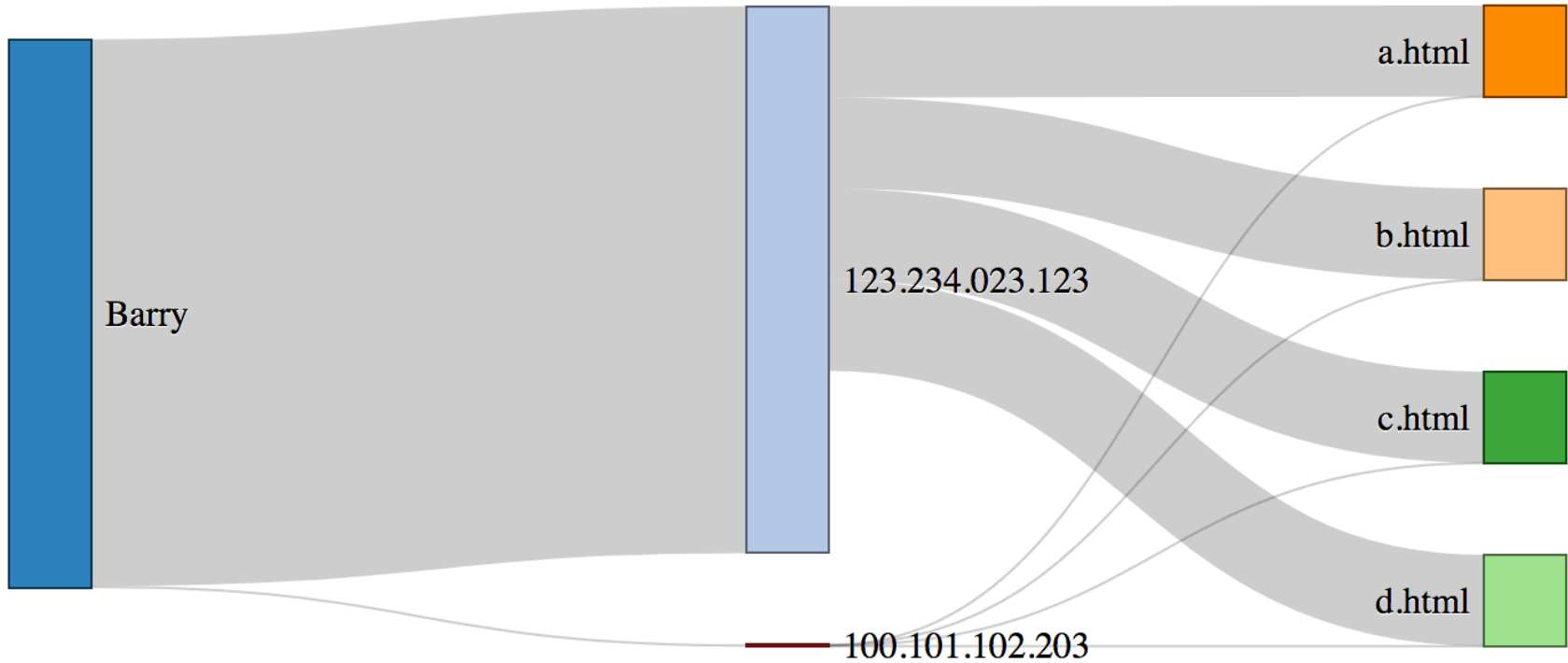
*Courtesy of the Tabulating Machine Company*  
**Fig. 217. A Standard 4-inch by 6-inch Filing Case** Is Used for the Curve Cards. Cards Twelve Inches Long Can Be Filed Lengthwise in a Drawer without Inconvenience  
 Guide cards separate the curve cards by departments. If desired, blue prints can be made from each curve card and the blue prints filed by expense-account numbers, as a cross-index of the data on the white cards which are filed by departments

**Fig. 231. A Complete** and **Fig. 230. Hollerith Tabulating Machine for Totalling**  
 The girls at the left are operating the key punches for punching the cards. A girl is at the table at the extreme right. In the corner is the card-sorting machine, and the machine is in the center. Files for punched cards are seen along the wall



# ANTI PATTERN #1: USING THE WRONG GRAPH TYPE

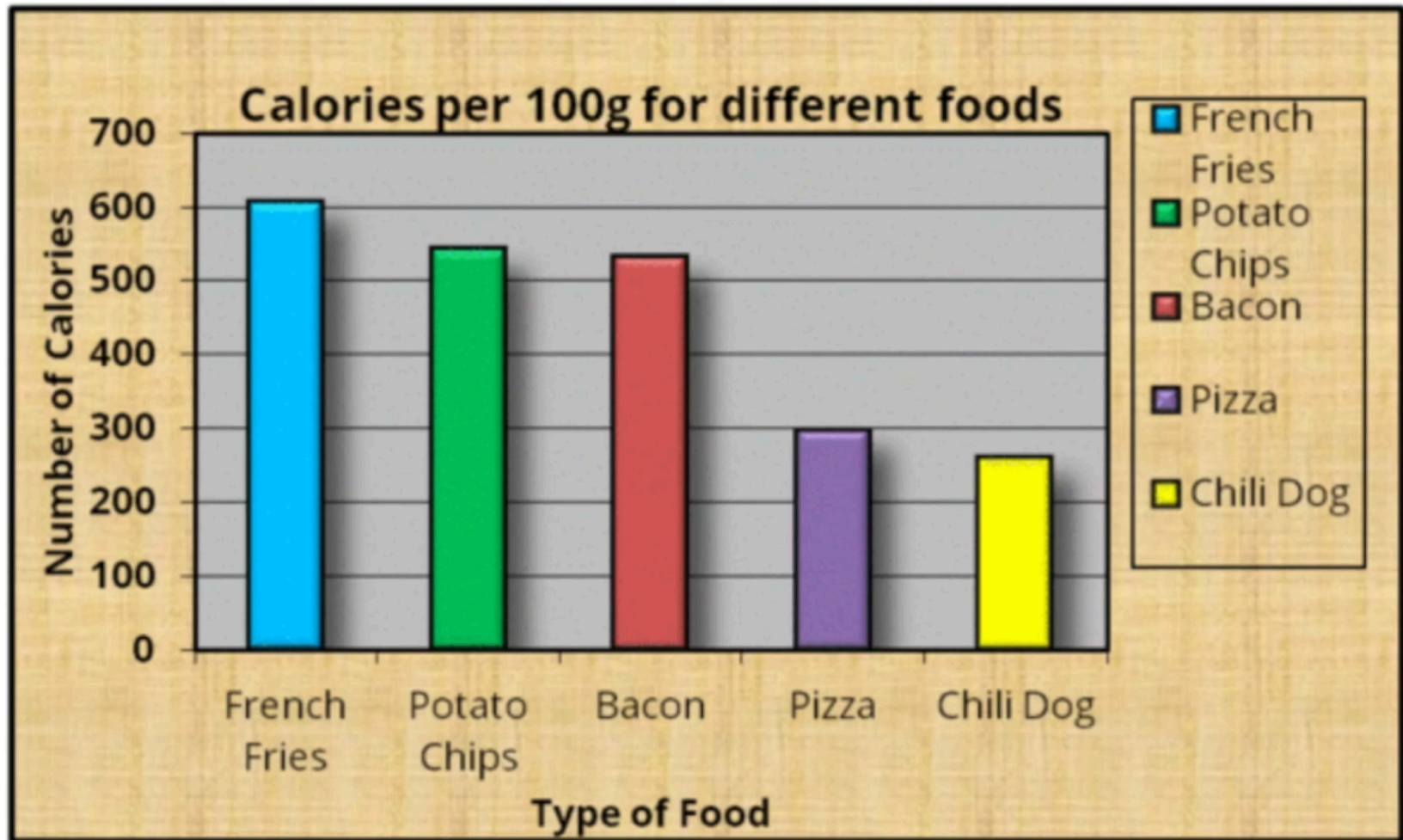
Violating all principles





## ANTI PATTERN #2: OVERLOADING

# Remove backgrounds



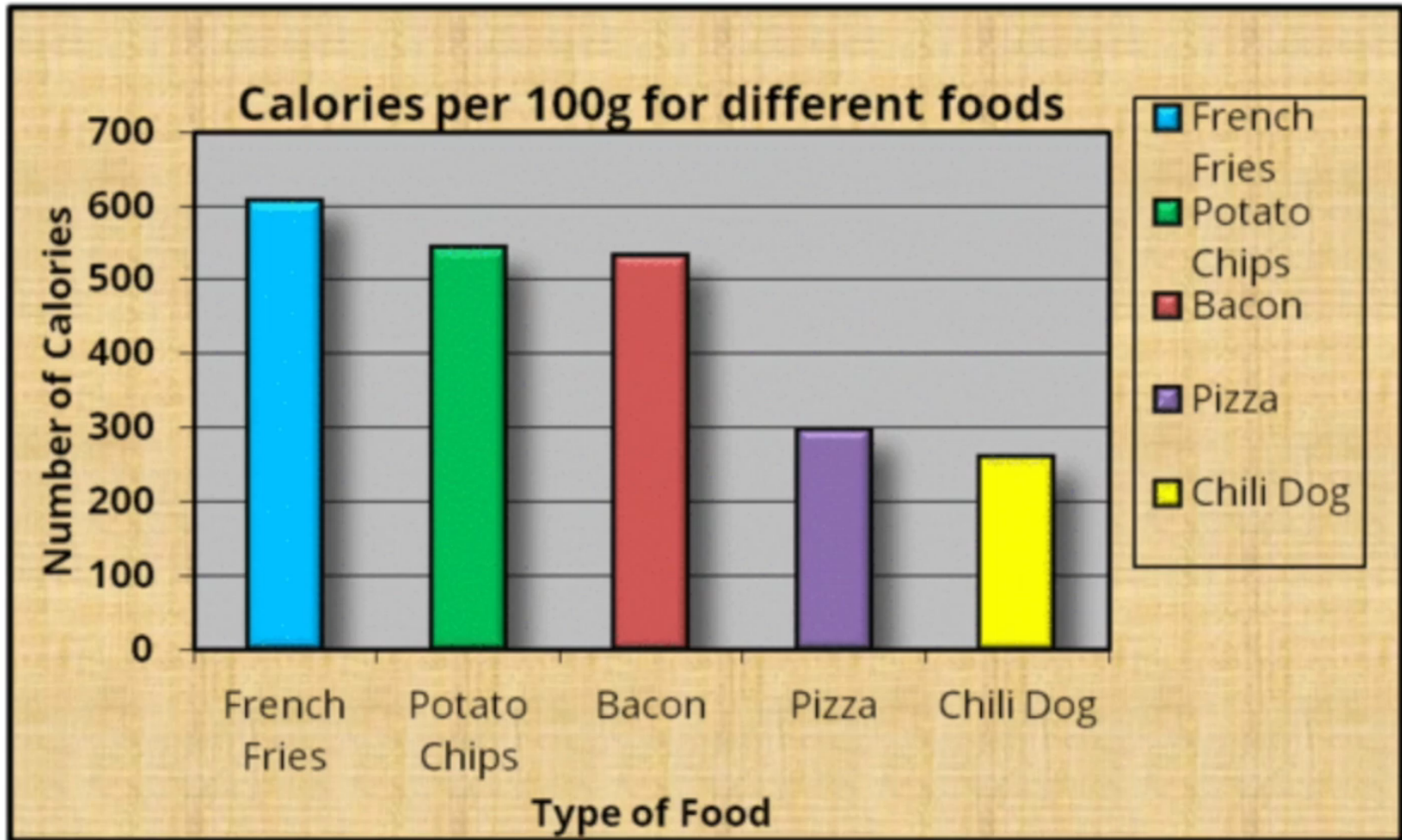
Created by Darkhorse Analytics

[www.darkhorseanalytics.com](http://www.darkhorseanalytics.com)



## ANTI PATTERN #2: OVERLOADING

# Remove backgrounds



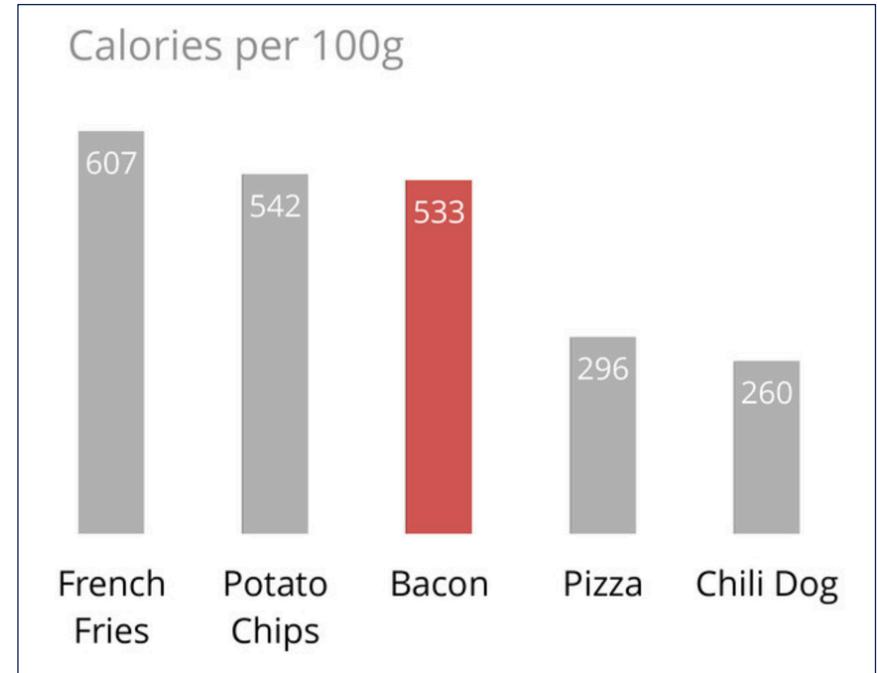
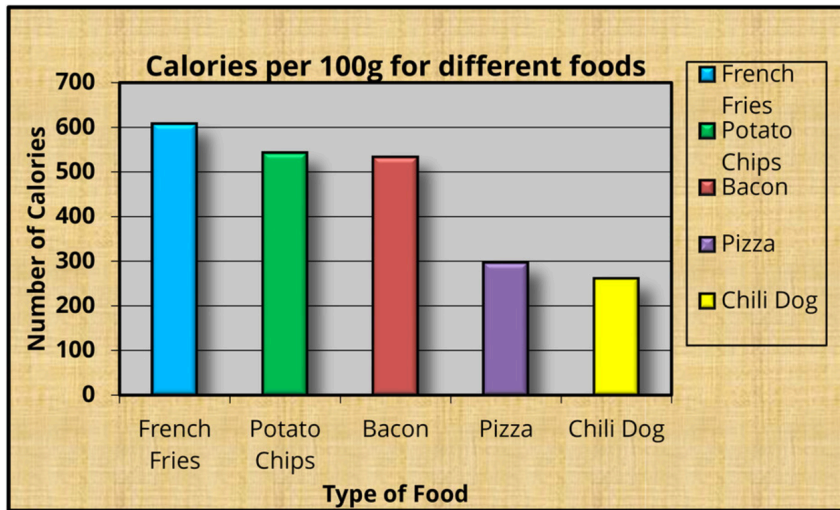
Created by Darkhorse Analytics

[www.darkhorseanalytics.com](http://www.darkhorseanalytics.com)



# ANTI PATTERN #2: OVERLOADING

## Data/Ink Ratio



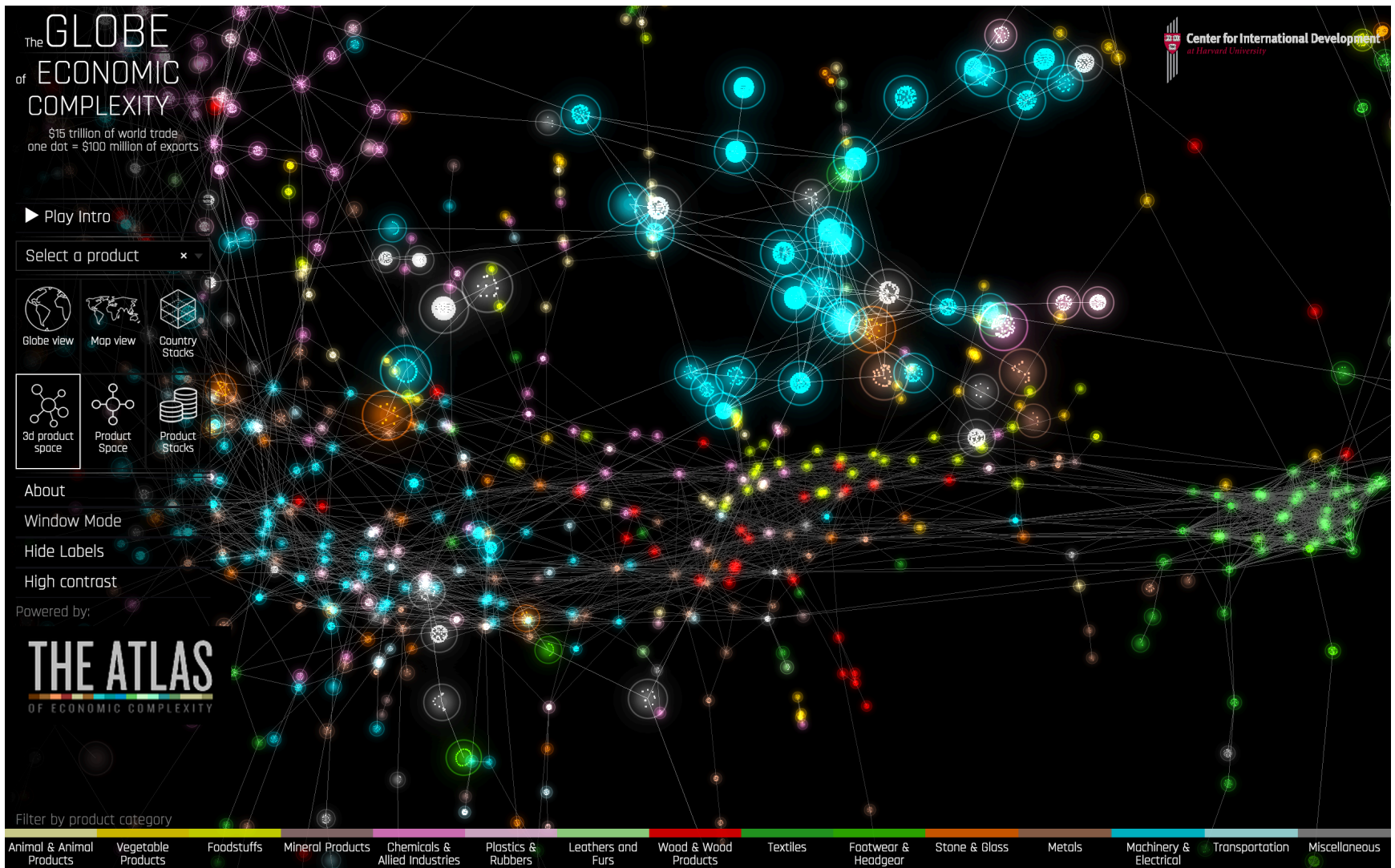
*"Perfection is achieved not when there is nothing more to add,  
but when there is nothing left to take away"*

Antoine de St Exupéry  
Terre des Hommes, 1939



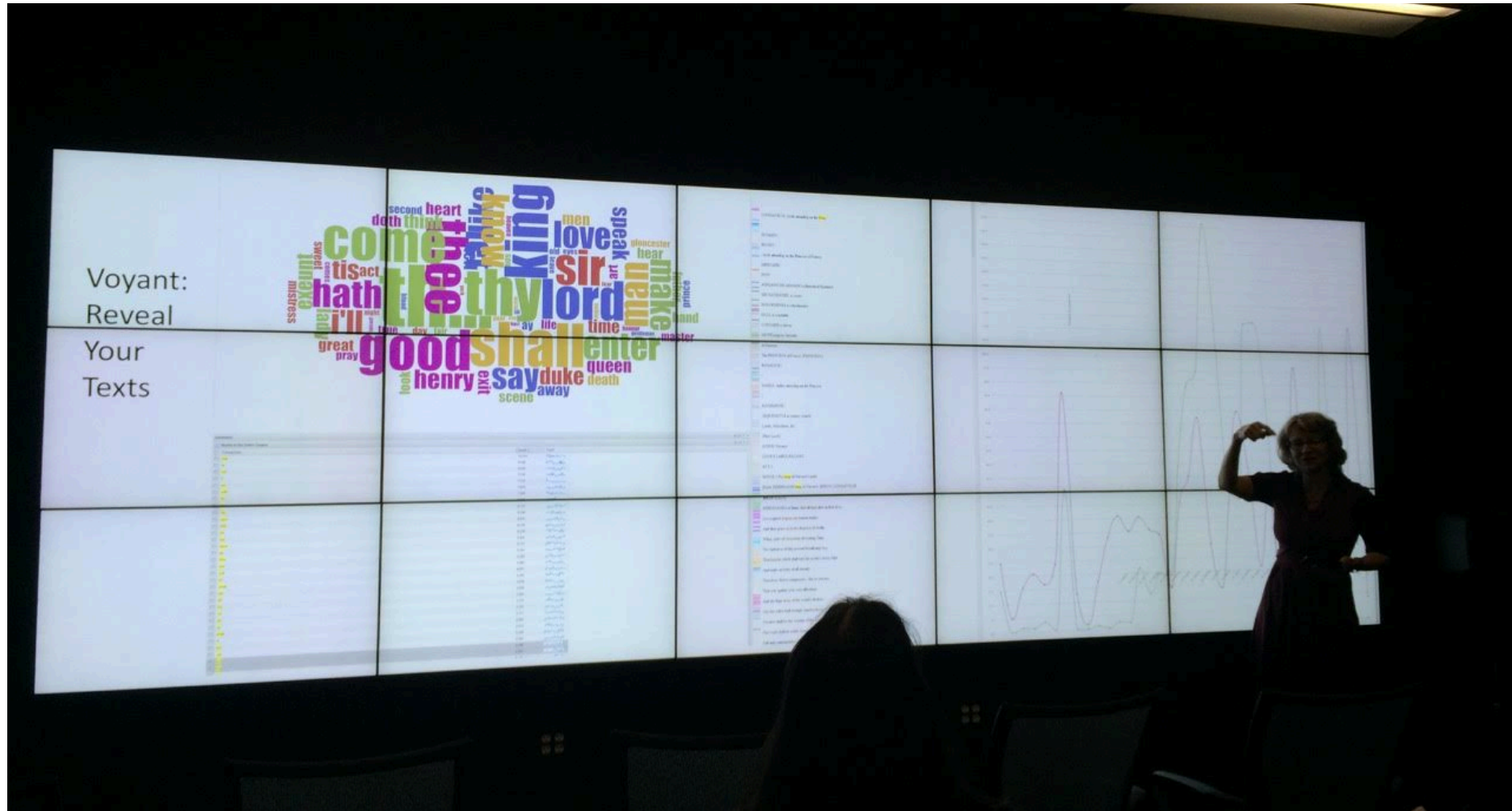
# ANTI PATTERN #3: "I REALLY WANT TO USE THIS FANCY NEW TECH"

## Three.js



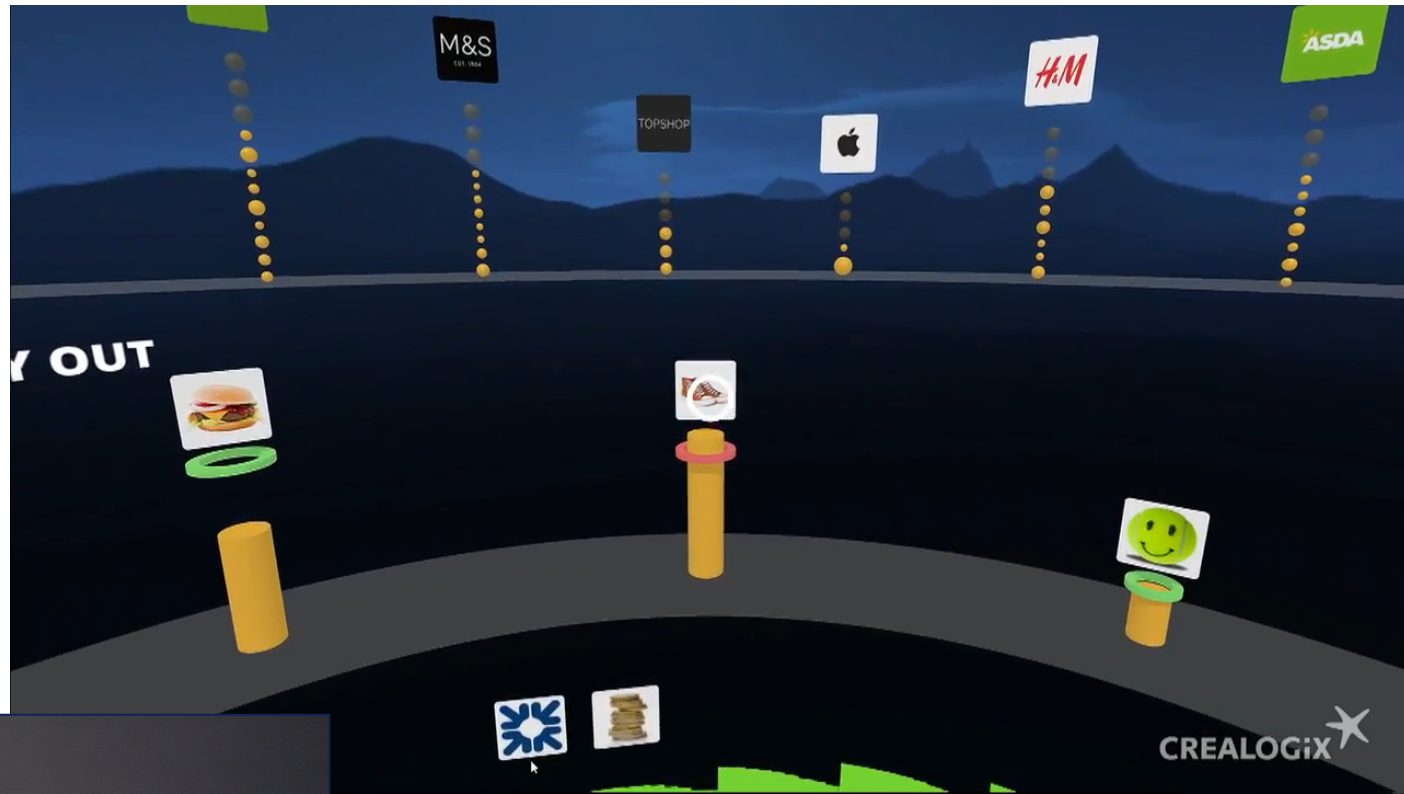
# ANTI PATTERN #3: "I REALLY WANT TO USE THIS FANCY NEW TECH"

Interactive display wall



# ANTI PATTERN #3: "I REALLY WANT TO USE THIS FANCY NEW TECH"

Virtual reality



# ANTI PATTERN #3: "I REALLY WANT TO USE THIS FANCY NEW TECH"

Data sonification

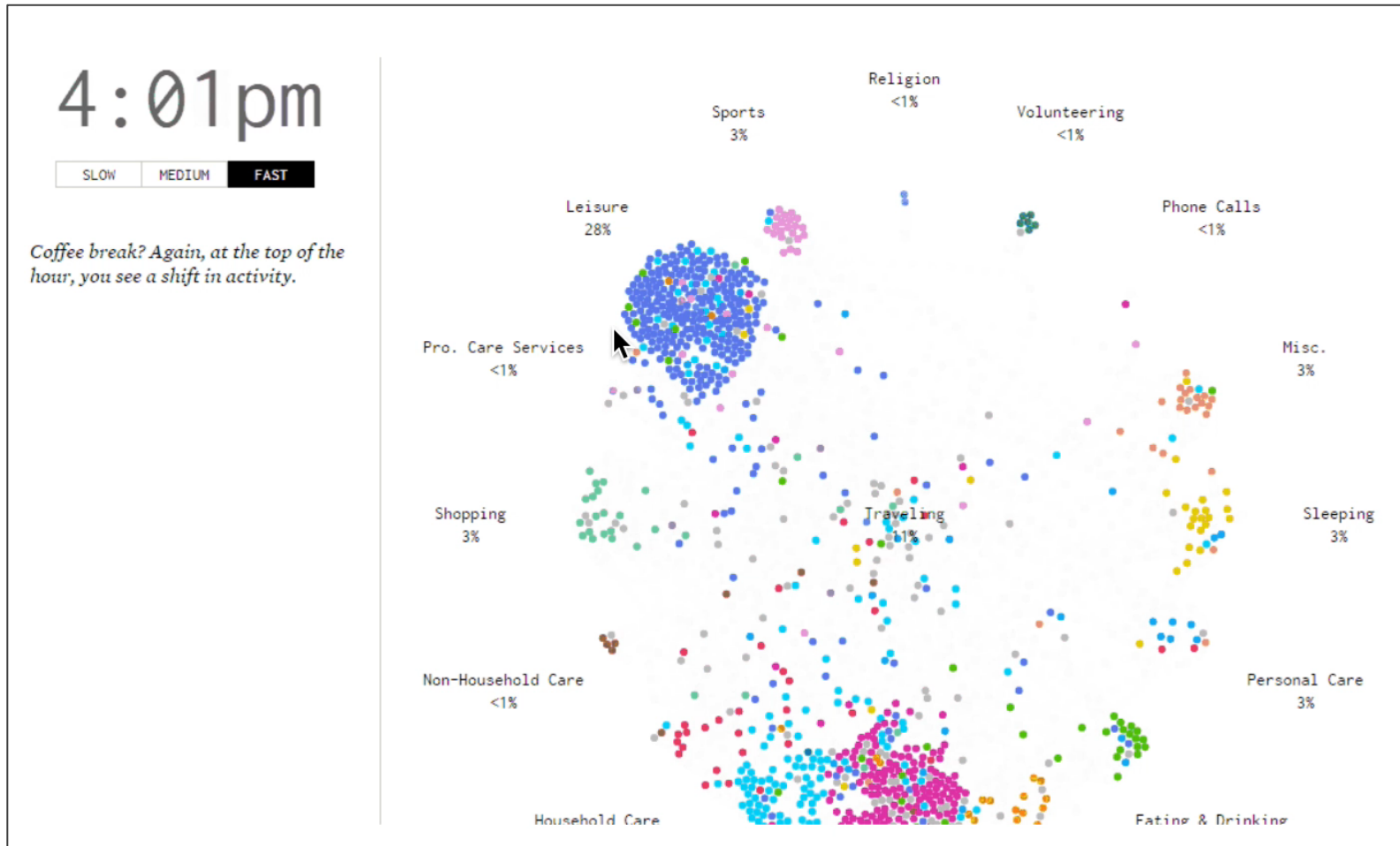


<http://earlymodernconversions.com/activity/history-visualization-lab/>



# ANTI PATTERN #3: "I REALLY WANT TO USE THIS FANCY NEW TECH"

## Animation



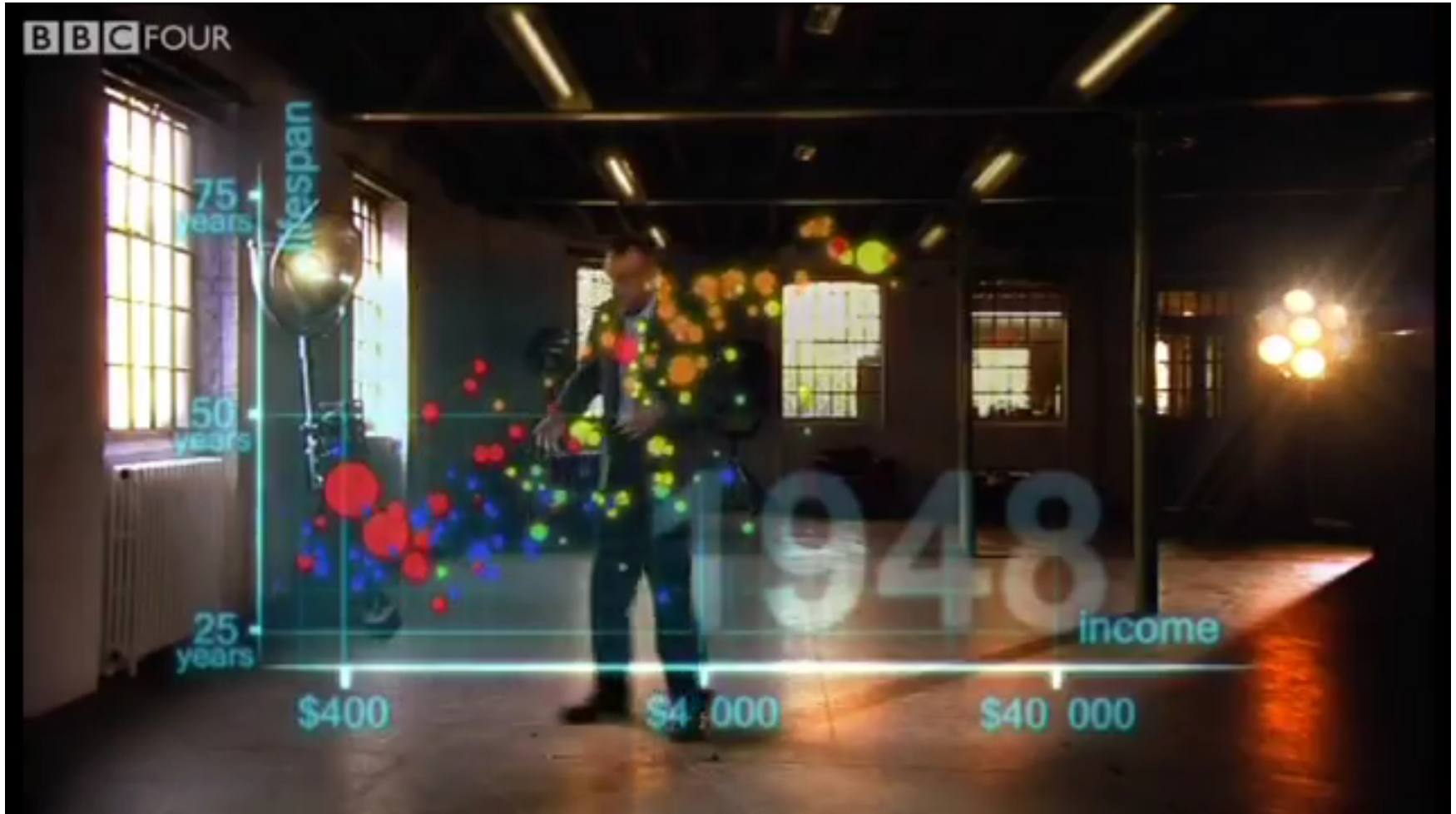
A Day in the Life of Americans – Nathan Yau





# EVEN THOUGH IT CAN BE GREAT, BY HANS ROSLING

Animation



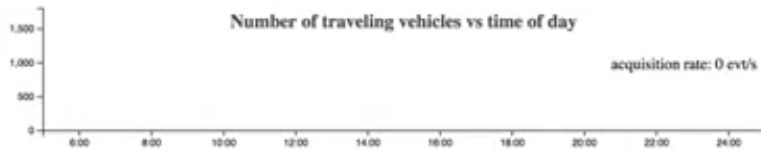
Hans Rosling... The Revolutionary



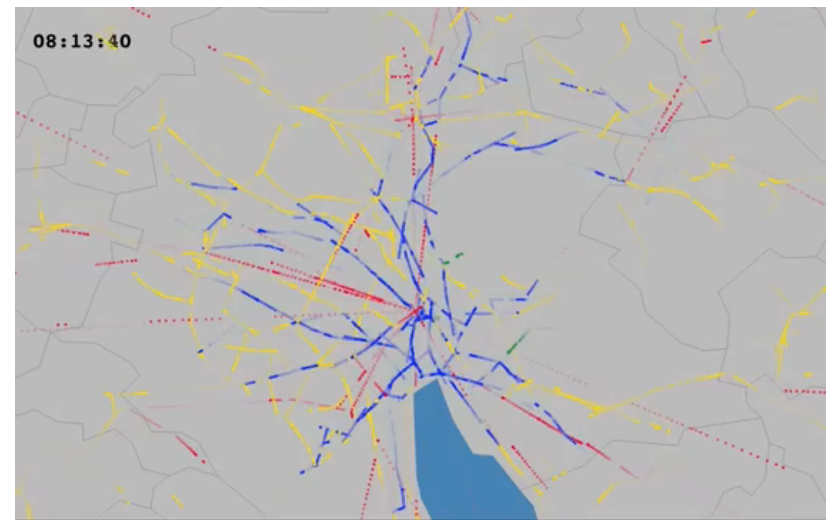
# ANTI PATTERN #4: UNDERESTIMATE YOUR BROWSER

Browser power with high throughput data

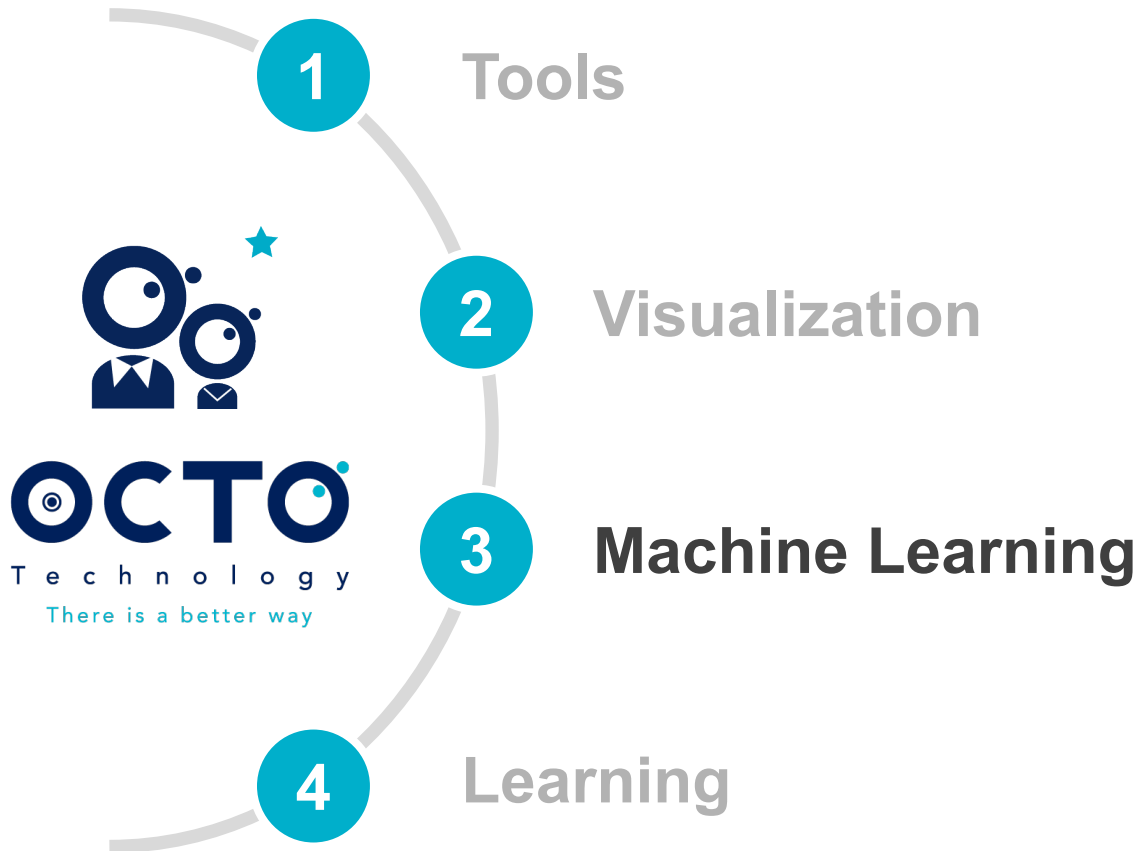
Up to 10'000 events per second



05:07:00



# FOUR PILLARS OF BIG DATA

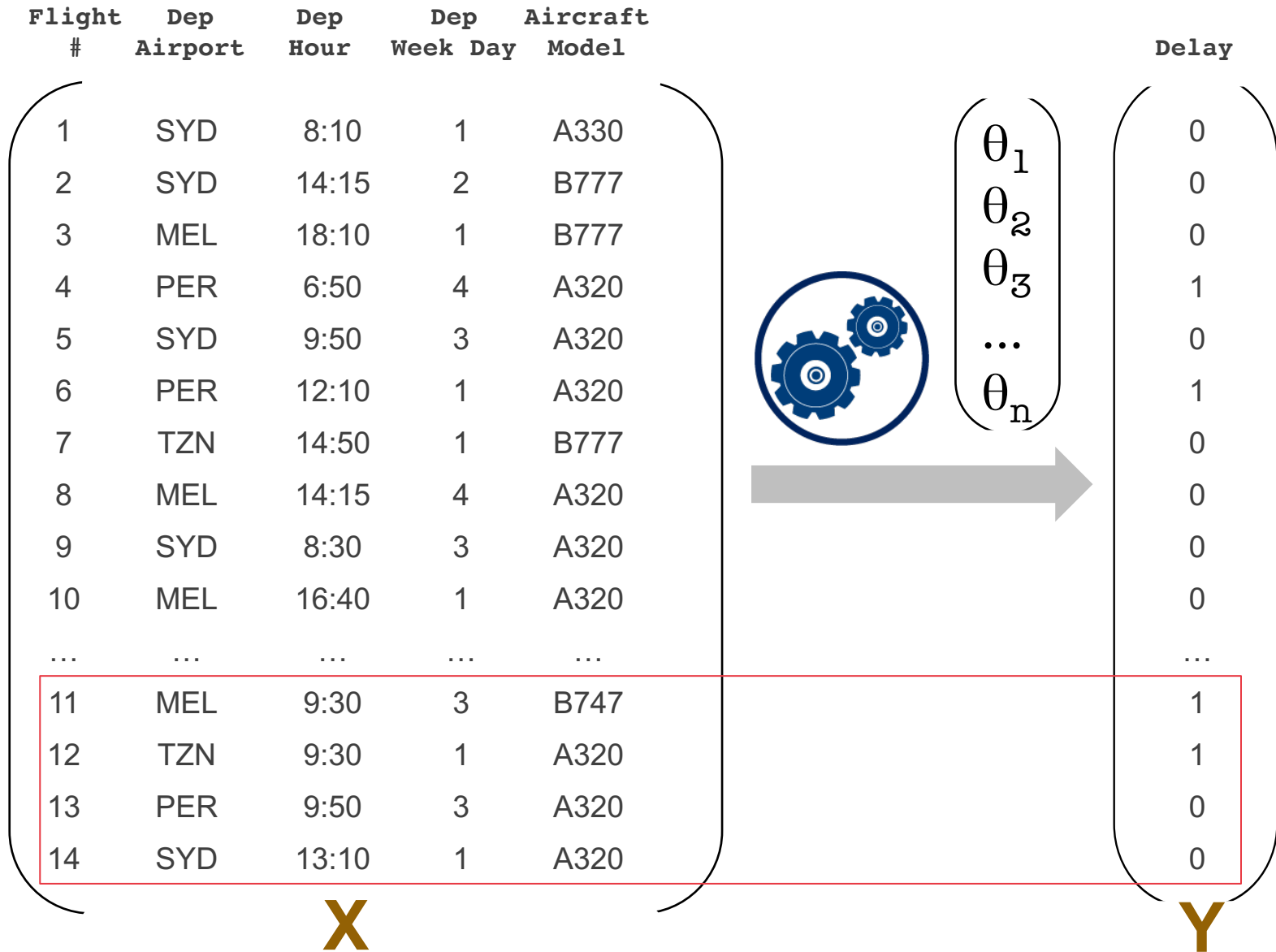


“Machine learning explores the study and construction of **algorithms** that can **learn** from and make **predictions** on **data**”

[https://en.wikipedia.org/wiki/Machine\\_learning](https://en.wikipedia.org/wiki/Machine_learning)

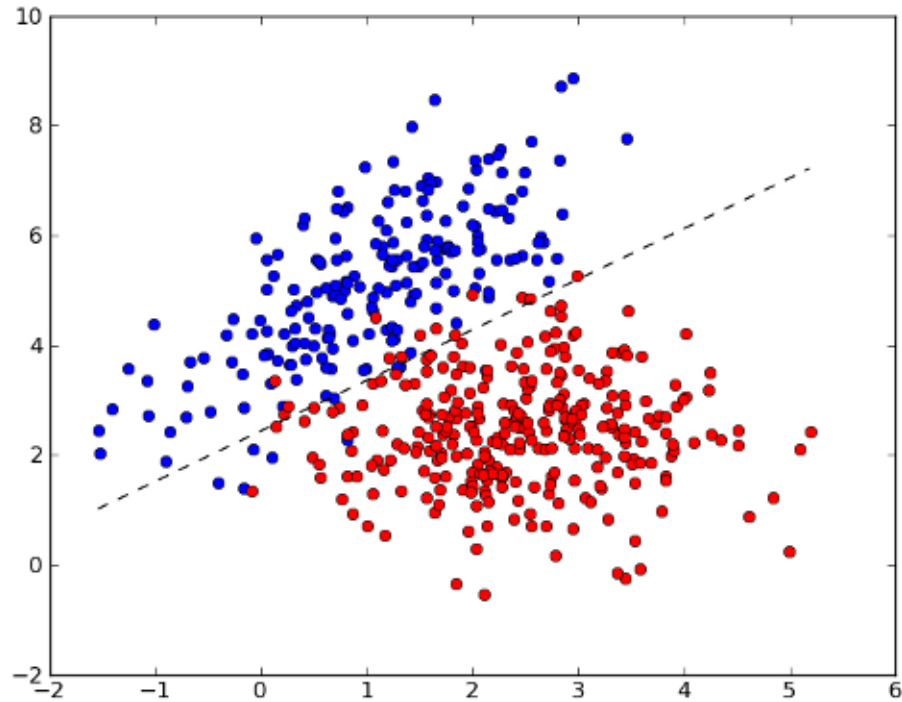


# BUILD A MODEL



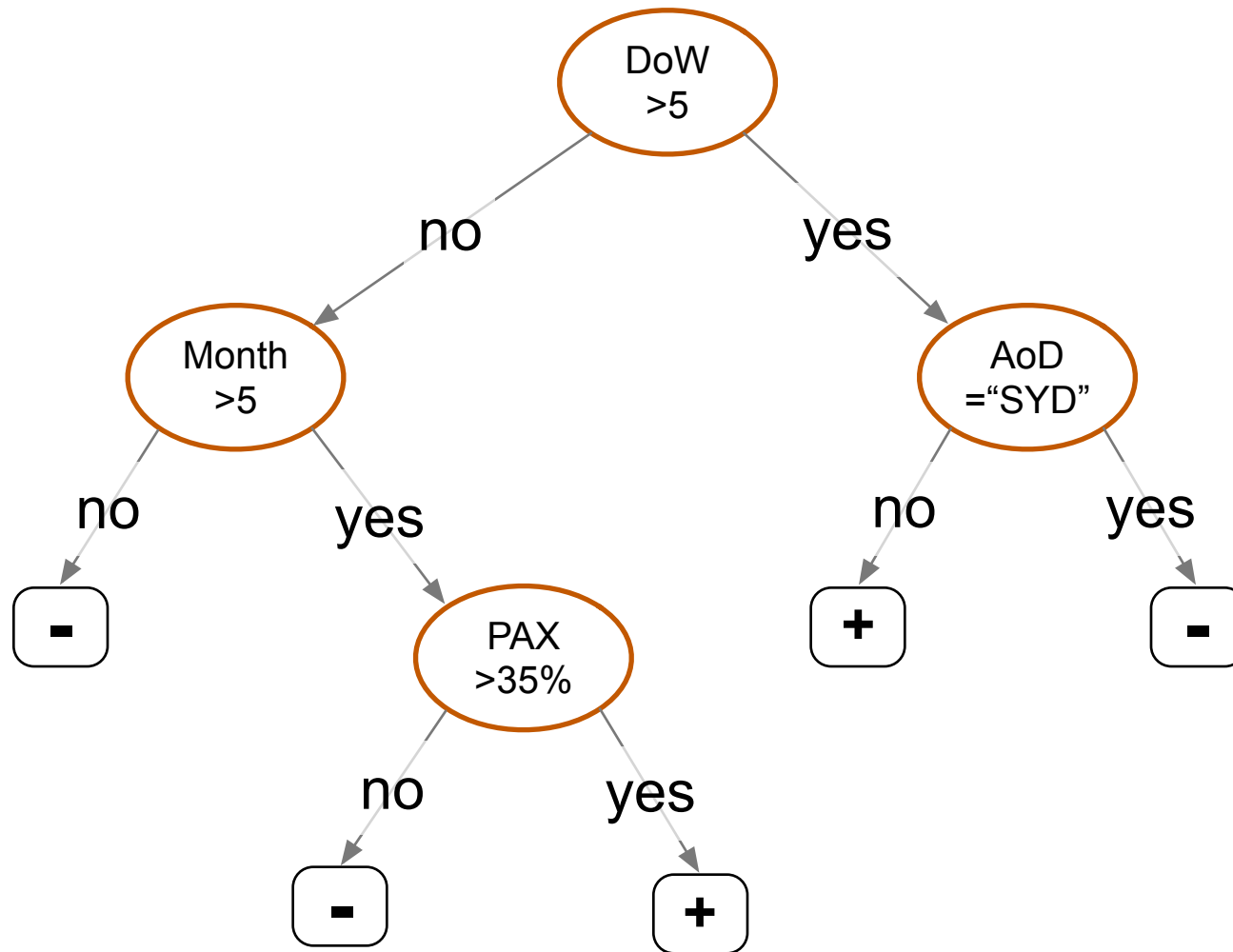
# LOGISTIC REGRESSION

## Classification algorithm



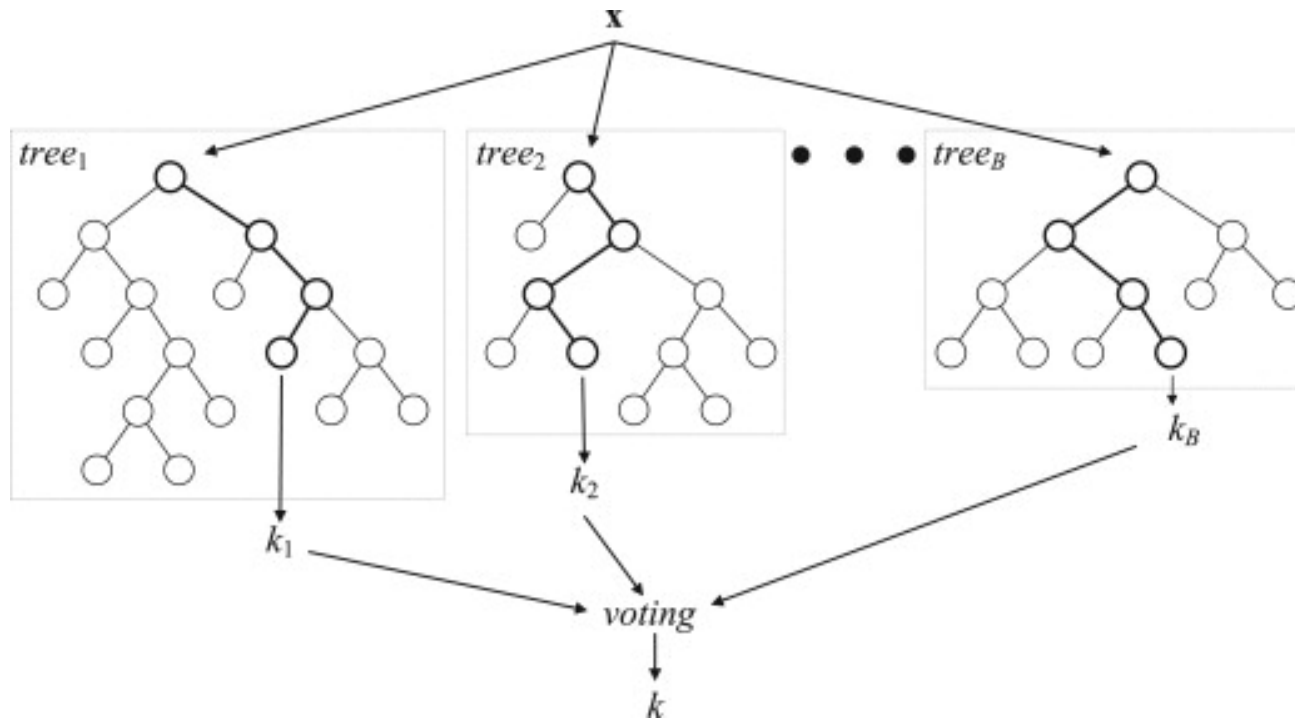
# DECISION TREE

## Classification algorithm



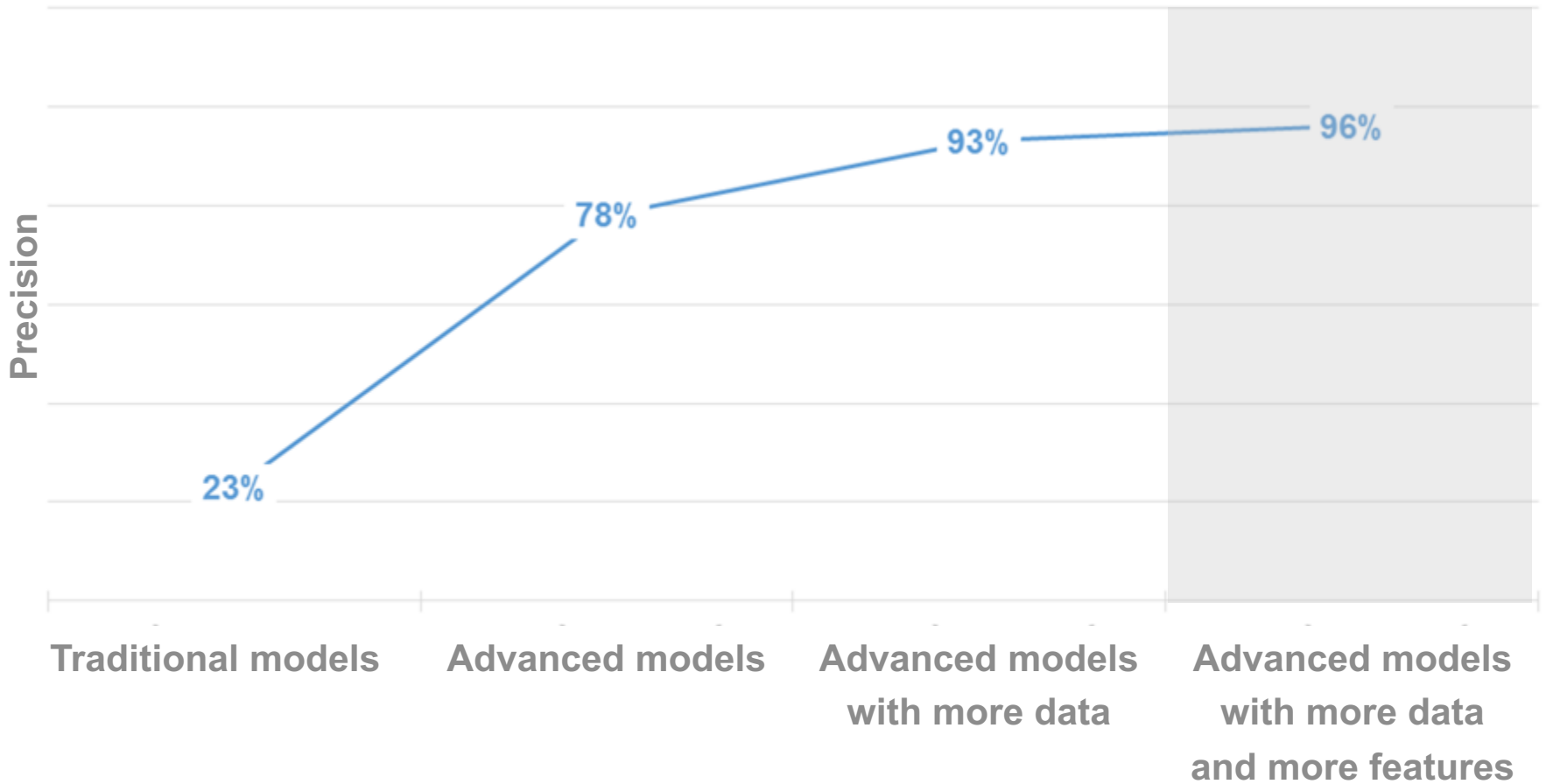
# RANDOM FOREST

## Classification algorithm

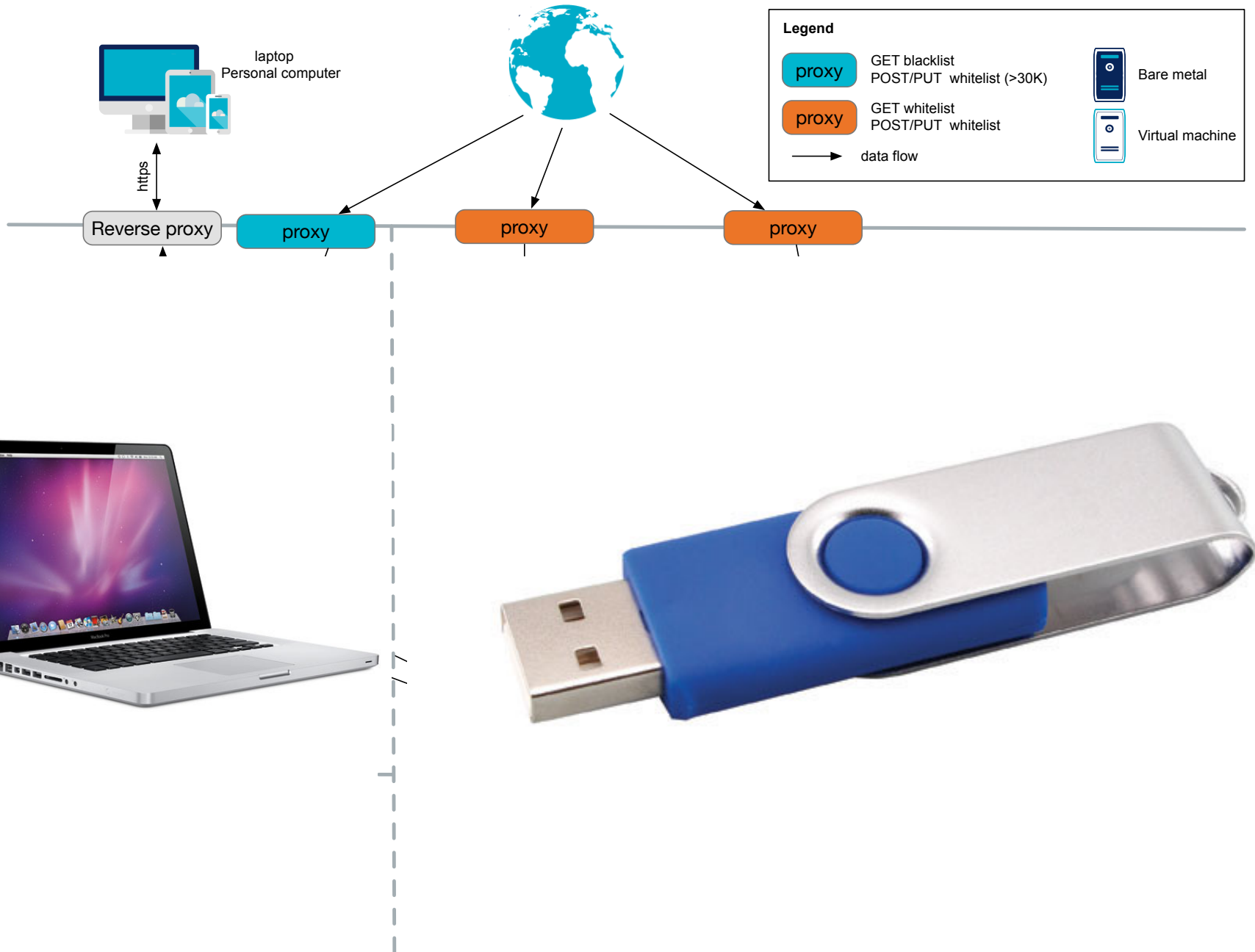




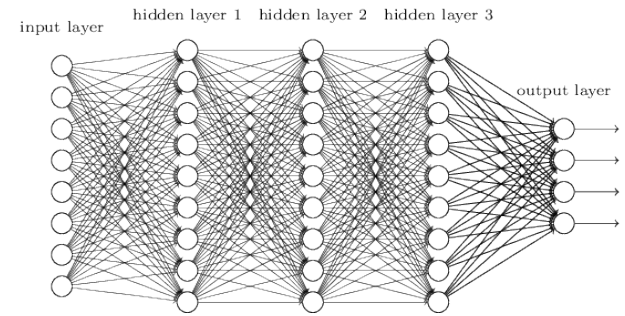
## Precision score for the TOP 20%



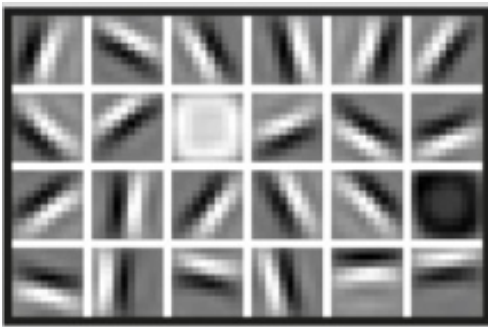
# ANTI PATTERN #1: WAIT FOR THE FULL STACK BEFORE STARTING



# ANTI PATTERN #2: START TOO COMPLEX



Identify pixels



Identify edges and simple shape



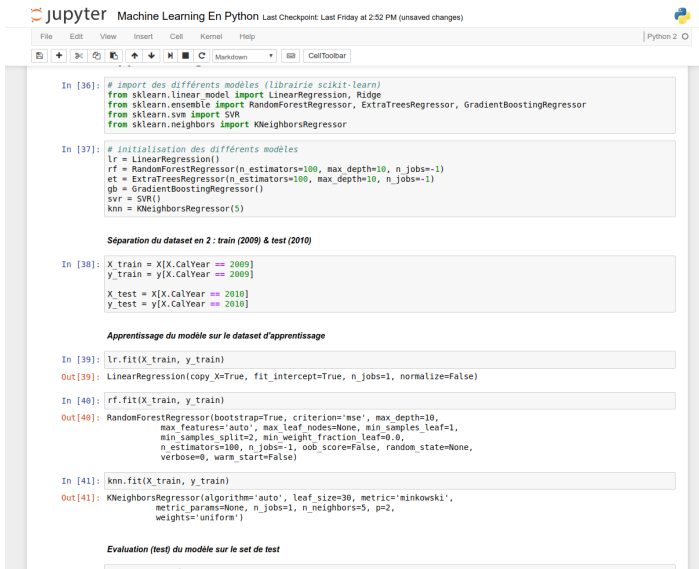
Identify complex shapes and object



Identify which shape to be used to define a human face



# ANTI PATTERN #3: FORGET THE LIFE AFTER THE NOTEBOOK



```
In [36]: # import des différents modèles (bibliothèque scikit-learn)
from sklearn.linear_model import LinearRegression, Ridge
from sklearn.ensemble import RandomForestRegressor, ExtraTreesRegressor, GradientBoostingRegressor
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor

In [37]: # initialisation des différents modèles
lr = LinearRegression()
rf = RandomForestRegressor(n_estimators=100, max_depth=10, n_jobs=-1)
et = ExtraTreesRegressor(n_estimators=100, max_depth=10, n_jobs=-1)
gb = GradientBoostingRegressor()
svr = SVR()
knn = KNeighborsRegressor(5)

Séparation du dataset en 2 : train (2009) & test (2010)

In [38]: X_train = X[X.CalYear == 2009]
y_train = y[X.CalYear == 2009]
X_test = X[X.CalYear == 2010]
y_test = y[X.CalYear == 2010]

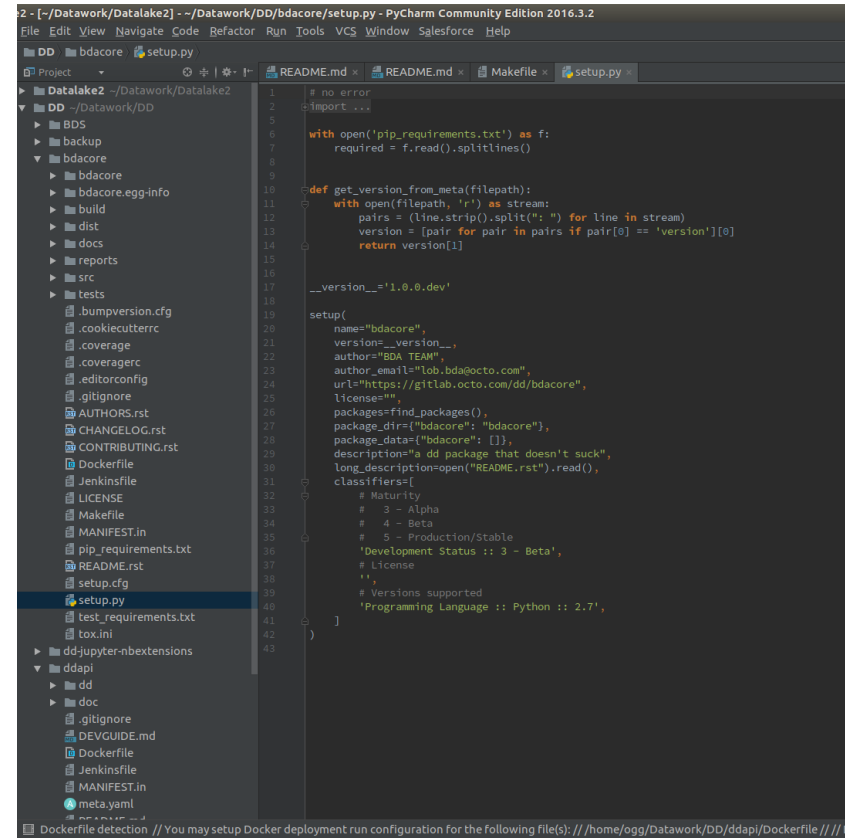
Apprentissage du modèle sur le dataset d'apprentissage

In [39]: lr.fit(X_train, y_train)
Out[39]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)

In [40]: rf.fit(X_train, y_train)
Out[40]: RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=10,
max_features='auto', max_leaf_nodes=None, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.0,
n_estimators=100, n_jobs=-1, oob_score=False, random_state=None,
verbose=0, warm_start=False)

In [41]: knn.fit(X_train, y_train)
Out[41]: KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
metric_params=None, n_jobs=1, n_neighbors=5, p=2,
weights='uniform')

Evaluation (test) du modèle sur le set de test
```



```
2 - [-/Datawork/DataLake2] --/Datawork/DD/bdacore/setup.py - PyCharm Community Edition 2016.3.2
File Edit View Navigate Code Refactor Run Tools VCS Window Salesforce Help

Project
DataLake2 --/Datawork/DataLake2
DD --/Datawork/DD
  BDS
  backup
  bdacore
    bdacore.egg-info
    build
    dist
    docs
    reports
    src
    tests
    .bumpversion.cfg
    .cookiecutter
    .coverage
    .coveragegc
    .editorconfig
    .gitignore
    AUTHORS.rst
    CHANGELOG.rst
    CONTRIBUTING.rst
    Dockerfile
    Jenkinsfile
    LICENSE
    Makefile
    MANIFEST.in
    pip_requirements.txt
    README.rst
    setup.cfg
    setup.py
    test_requirements.txt
    tox.ini
  dd-jupyter-nbextensions
  ddapi
  dd
  doc
  .gitignore
  DEVGUIDE.md
  Dockerfile
  Jenkinsfile
  MANIFEST.in
  meta.yaml
  requirements

Dockerfile detection // You may setup Docker deployment run configuration for the following file(s): //home/ogg/Datawork/DD/ddapi/Dockerfile ///

1 # no error
2 import sys
3
4
5
6 with open('pip_requirements.txt') as f:
7     required = f.read().splitlines()
8
9
10 def get_version_from_meta(filepath):
11     with open(filepath, 'r') as stream:
12         pairs = (line.strip().split(':', 1) for line in stream)
13         version = [pair for pair in pairs if pair[0] == 'version'][0]
14         return version[1]
15
16
17 __version__ = '1.0.0.dev'
18
19
20 setup(
21     name="bdacore",
22     version=__version__,
23     author="BDA TEAM",
24     author_email="lob.bda@octo.com",
25     url="https://gitlab.octo.com/dd/bdacore",
26     license="",
27     packages=find_packages(),
28     package_dir={"bdacore": "bdacore"},
29     package_data={"bdacore": []},
30     description="a dd package that doesn't suck",
31     long_description=open("README.rst").read(),
32     classifiers=[
33         # Maturity
34         # 3 - Alpha
35         # 4 - Beta
36         # 5 = Production/Stable
37         'Development Status :: 3 - Beta',
38         # License
39         # Versions supported
40         'Programming Language :: Python :: 2.7',
41     ],
42 )
43
```

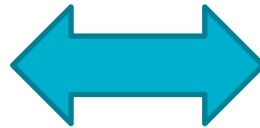


# ANTI PATTERN #3: FORGET THE LIFE AFTER THE NOTEBOOK

## Competencies



Data Scientist  
*Basically, a rockstar*



OPS

- ⦿ Advanced statistic
- ⦿ Artificial Intelligence
- ⦿ Bleeding edge algorithm

- ⦿ Run things



## Competencies



- ⦿ (Machine Learning)
- ⦿ (Statistics)

But overall...

- ⦿ Software Craftsmanship
- ⦿ A software development culture
  - > Production?
  - > Architecture
  - > Agility principles?



# ANTI PATTERN #4: MAKING A POC INSTEAD OF A PILOT PROJECT

- POC : Check the model works
- Pilot project : Check it is useful

### Customer next best action

Extract more value from your scores in three steps :

1. Choose a customer
2. Examine attributes impacts
3. Get next best action

**1 Select your customer**  
 Get insight!

**2 Select actionable features**  
 Get next action!

#### Informations

Show 10 entriesSearch:

Name	Value
score_id	12.0
source	1.0
id	1.0
prediction	0.0687926006625
crc_insan	0.0
prea	1.0
darty	0.0
code_type_comptage_insan	0.0
code_insee	44154.0
code_statut_logement	0.0

#### Insights

Show 10 entriesSearch:

Variable name	Contribution	%
nb_cont_cour_ent	<div style="width: 16%;"></div>	16%
ass_fact	<div style="width: 12%;"></div>	12%
age_tr_pred_fmt	<div style="width: 11%;"></div>	11%
lb_prof_payr	<div style="width: 7%;"></div>	7%
cel_fmt	<div style="width: 6%;"></div>	6%
crc	<div style="width: 4%;"></div>	4%
puis_sous	<div style="width: 3%;"></div>	3%
code_type_comptage	<div style="width: 3%;"></div>	3%
code_type_comptage_insan	<div style="width: 3%;"></div>	3%
code_insee	<div style="width: 3%;"></div>	3%

#### Next best action

Show 10 entriesSearch:

	Current	Best
nb_cont_cour_ent	2	6
ass_fact	0	1
age_tr_pred_fmt	3	5
Score	0.0687926007	0.2121057026

First Previous Next Last

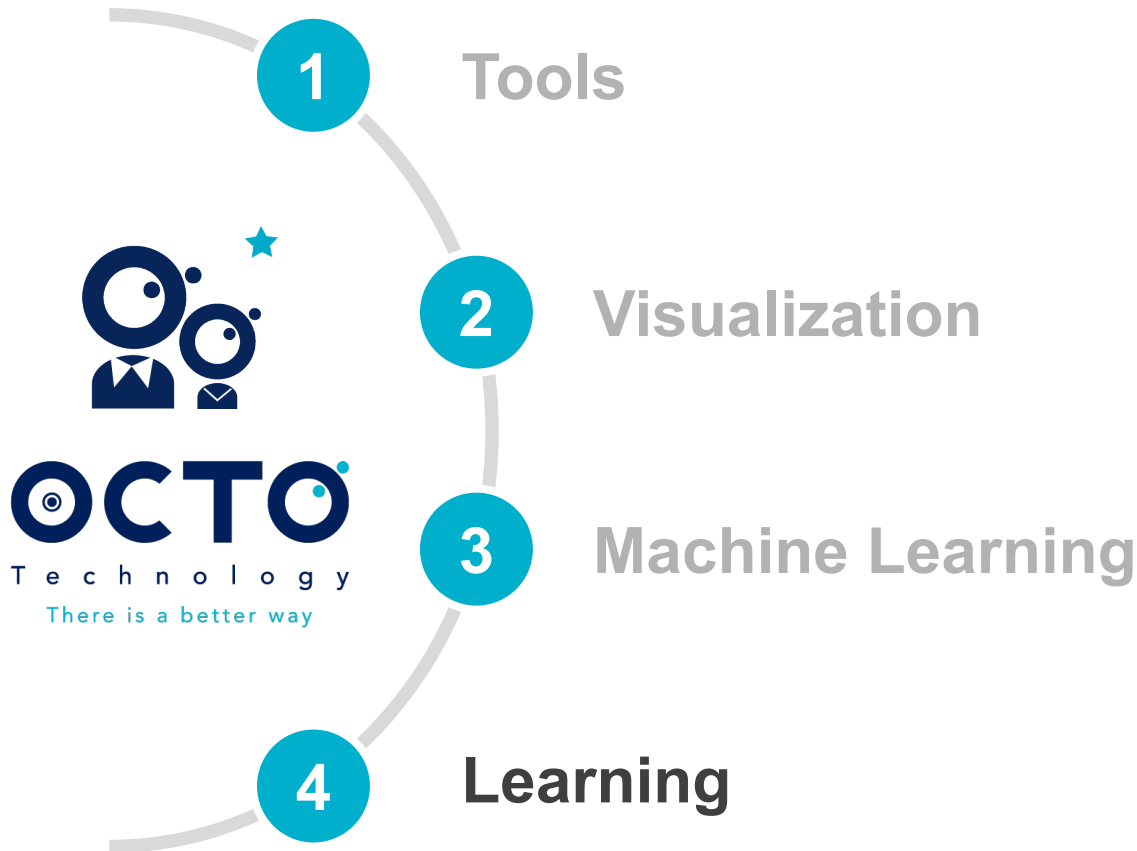
Current score 0.069.  
Best score achievable 0.212.

You can increase your score of 208.3%

First Previous Next Last



# FOUR PILLARS OF BIG DATA





## Learn To Change

### AGILE & LEAN

- BEST** Découvrir les démarches agiles et la culture agile
- NEW** Certification Scrum Master
- NEW** Adopter les bonnes pratiques de gestion de projet agile
- BEST** Facilitation graphique par la pratique
- Animer une rétrospective projet
- Le rôle de Product Owner en pratique
- BEST** Certification Leading SAFe au Scaled Framework - Séb...
- Conception logicielle : écrire des cas d'utilisation efficaces
- Kanban : mieux travailler en maîtrisant son flux de production
- Lean Startup en entreprise
- Lean IT : optimiser ses flux plutôt que ses ressources
- EXCLU** Devenir Coach Agile ou Scrum Master

### CULTURE CHANGE

#### CHANGE ET TRANSFORMATION

- EXCLU** L'atelier du changement
- EXCLU** Théorie U
- EXCLU** Explorer et partager la démarche entrepreneuriale aut...

#### FACILITATION

- EXCLU** Techniques et méthodes de facilitation de groupe : niv...
- EXCLU** Facilitation d'un plan stratégique participatif : niveau 2
- EXCLU** Enrichir et approfondir sa posture de facilitateur dans ...

#### MANAGEMENT ET LEADERSHIP

- EXCLU** Optimiser son temps et ses priorités - Ismaël Héry
- Parole performante et communication impactante
- BEST** Donner et recevoir des feedbacks efficaces
- BEST** Management collaboratif et agile : there is a better way !

### BIG DATA

#### DATA SCIENCE

- BEST** Fondamentaux de la Data Science
- BEST** Data Science : niveau avancé

#### HADOOP HORTONWORKS

Les fondamentaux d'Hadoop

- BEST** Administrer la plateforme Hadoop 2.X Hortonworks : ni...
- EXCLU** Administrer la plateforme Hadoop 2.X Hortonworks : n...
- EXCLU** Administrer la plateforme Hadoop 2.X Hortonworks : s...
- BEST** Analyse de données pour Hadoop 2.X Hortonworks av...
- EXCLU** Développer des applications pour Apache Spark avec ...

#### HADOOP CLUDERA

- BEST** Développer des applications pour Spark avec Hadoop C...
- NEW** Administrer la plateforme Hadoop Cludera
- NEW** Utiliser Pig, Hive et Impala avec Hadoop Cludera pour ...

#### SPARK DATABRICKS

- EXCLU** Programmer avec Apache Spark de Databricks

#### NOSQL

- BEST** NoSQL : découverte des solutions et architecture de la ...
- Déployer et gérer un cluster Couchbase
- Requêtes, modélisation de données, optimisation et migration...
- Gérer efficacement ses logs avec la stack ELK
- NEW** Concevoir un moteur de recherche avec Elasticsearch

*Favoriser  
l'innovation*

*Accompagner  
les transformations*

*Inspirer  
dans la durée*



# OCTO Suisse RECRUTE

[rejoins.octo.com](https://rejoins.octo.com)

*Architecte*

*Coach  
Méthodo*

*Software  
Craftsman*

**Questions ?**

*DataGeek*

*Expert  
DevOps*

*Consultant  
en Stratégie*

