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 archievements*

ARE THE RESULT OF *shared* KNOWLEDGE

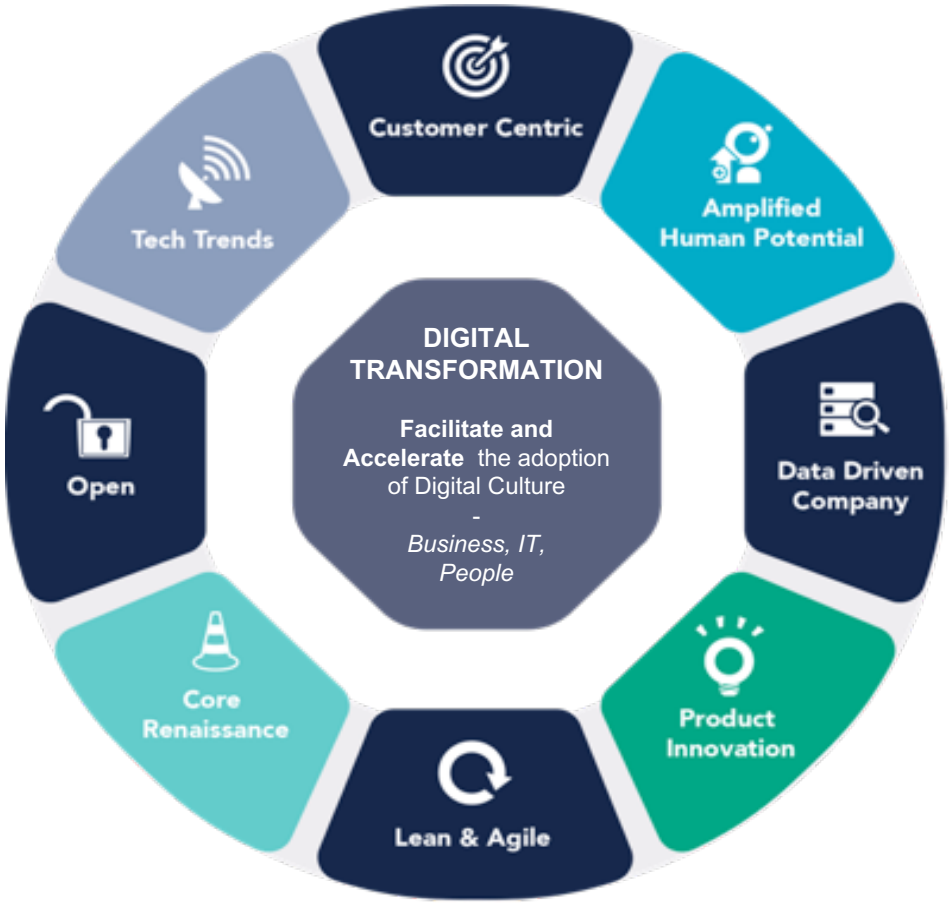
AND THE JOY OF WORKING +OGETHER

WE ARE ALWAYS *on the lookout*

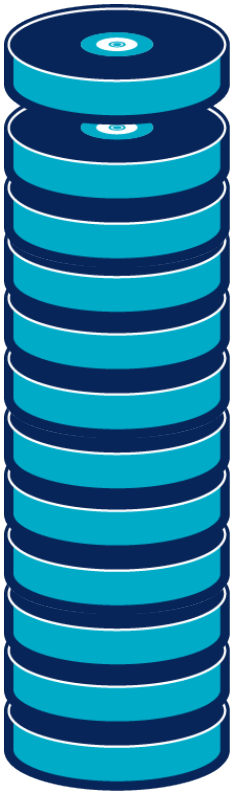
FOR BETTER *ways* OF WORKING

OCTO, DIGITAL TRANSFORMATION ACCELERATOR

Consulting & Delivery

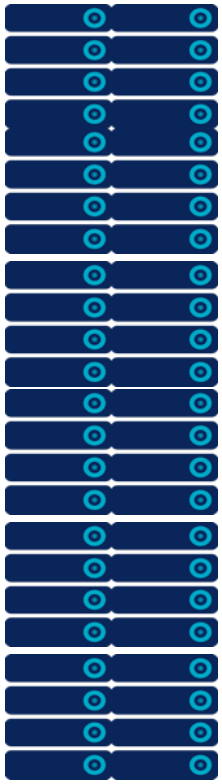


BIG DATA @ OCTO : THE NUMBERS



250

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

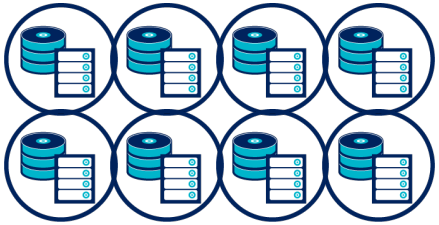


800

cores, the biggest Hadoop cluster built by OCTO

40

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

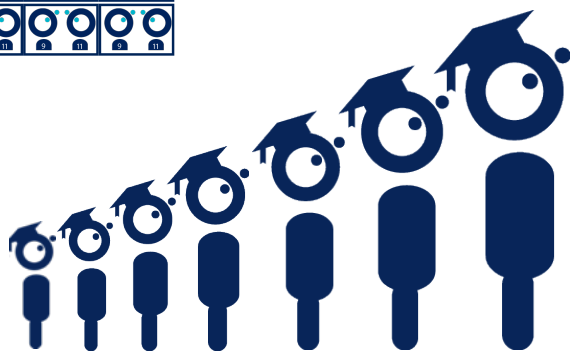


850

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

1st	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
3rd	SNCF - *prédiction de la fréquentation des gares en Ile de France*	datascience.net
6th	Imperial College London - *Loan Default Prediction*	kaggle
13th	Allstate – *purchase prediction challenge*	kaggle
2nd	Tradeshift – *Text classification*	kaggle
5th	Microsoft - *Malware classification*	kaggle

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


Afterwork BigData

FOUR PILLARS OF BIG DATA

OCTO Technology
There is a better way

FOUR PILLARS OF BIG DATA

OCTO Technology
There is a better way

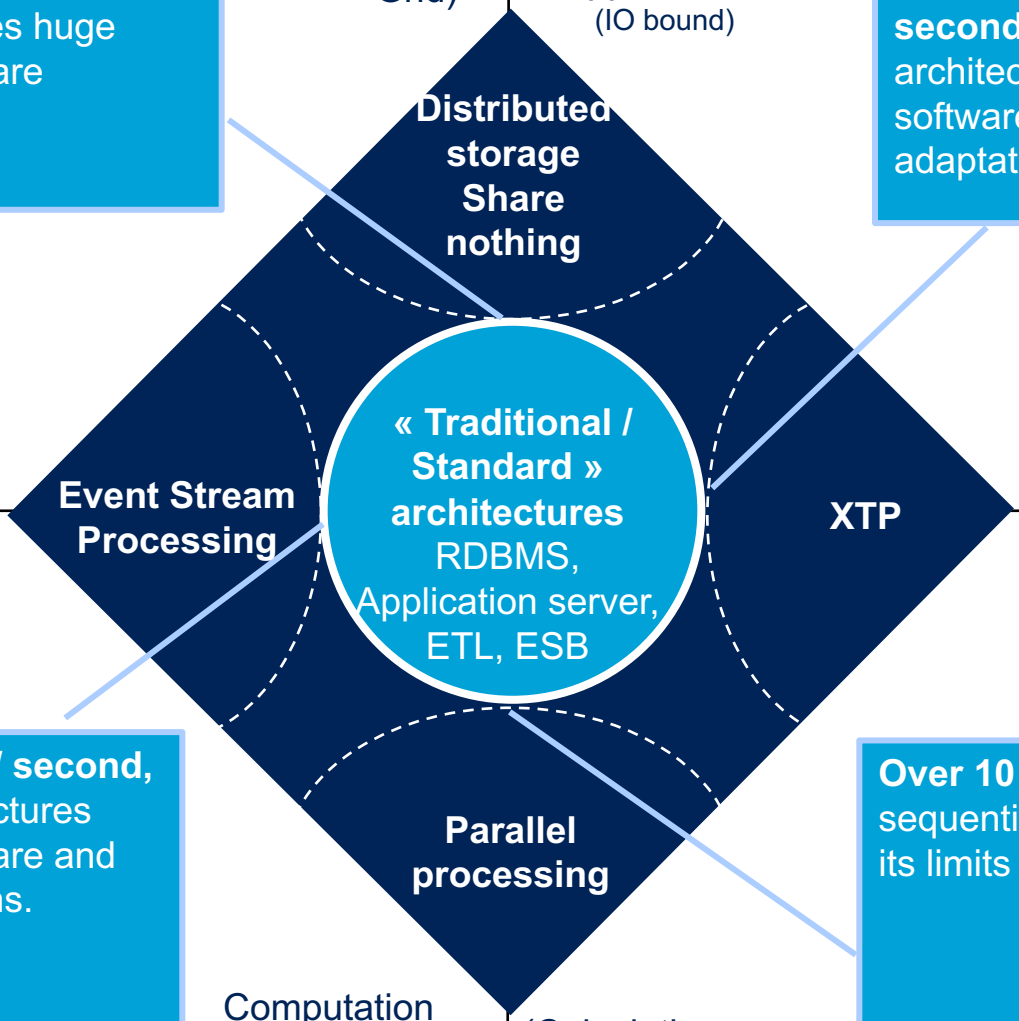
LIMITATIONS OF TRADITIONAL ARCHITECTURES

Over 10 Tb, « classical » architectures requires huge software and hardware adaptations.

(Storage Grid)

Storage oriented applications
(IO bound)

Over 1 000 transactions / second, « classical » architectures requires huge software and hardware adaptations.



(Stream Grid)

Event flow oriented application
Message Bound
(streaming)

(Transaction Grid)

Transaction oriented applications
Transaction Bound
(TPS)

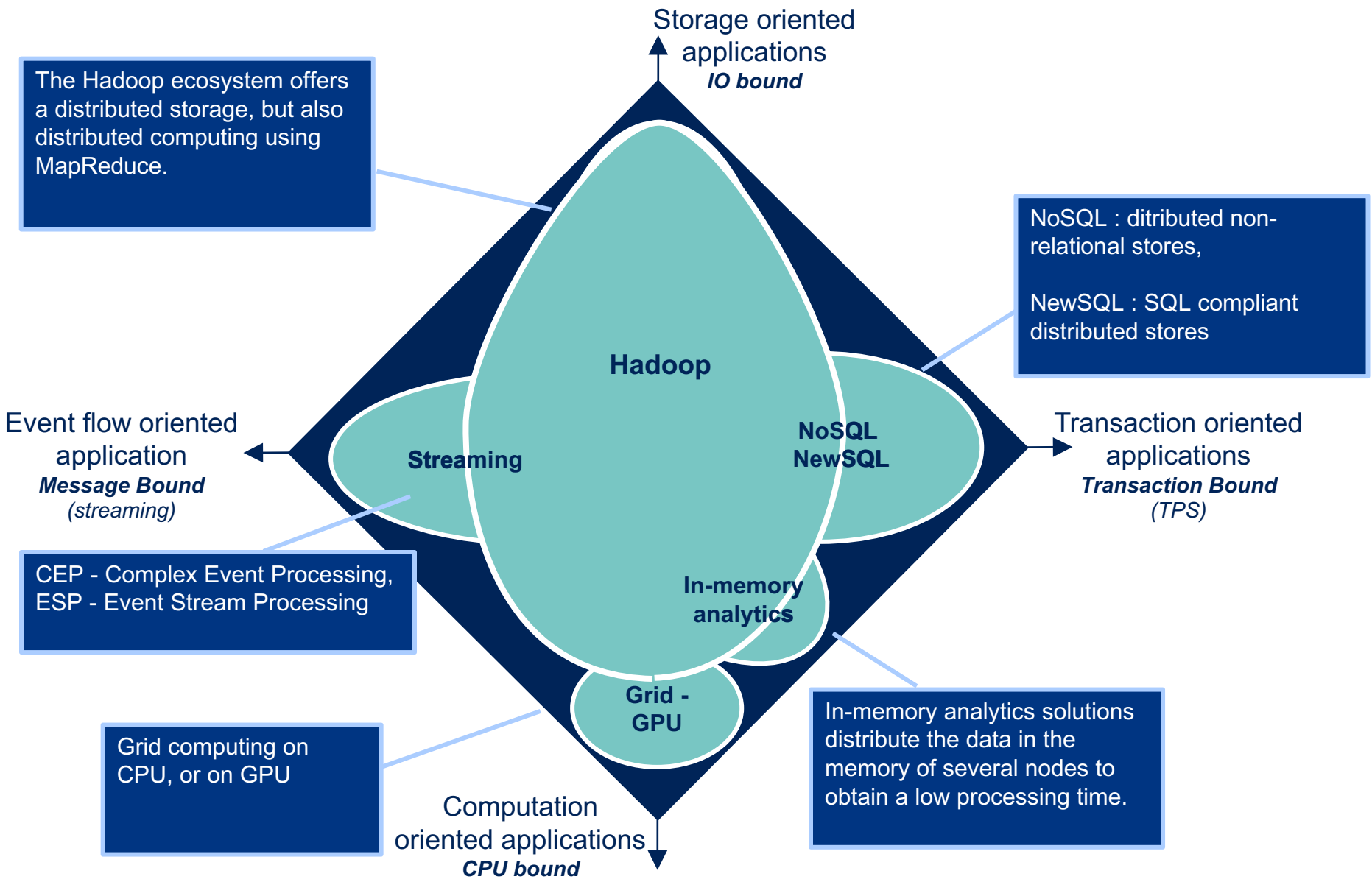
Over 1 000 events / second, « classical » architectures requires huge software and hardware adaptations.

Computation oriented applications
CPU bound

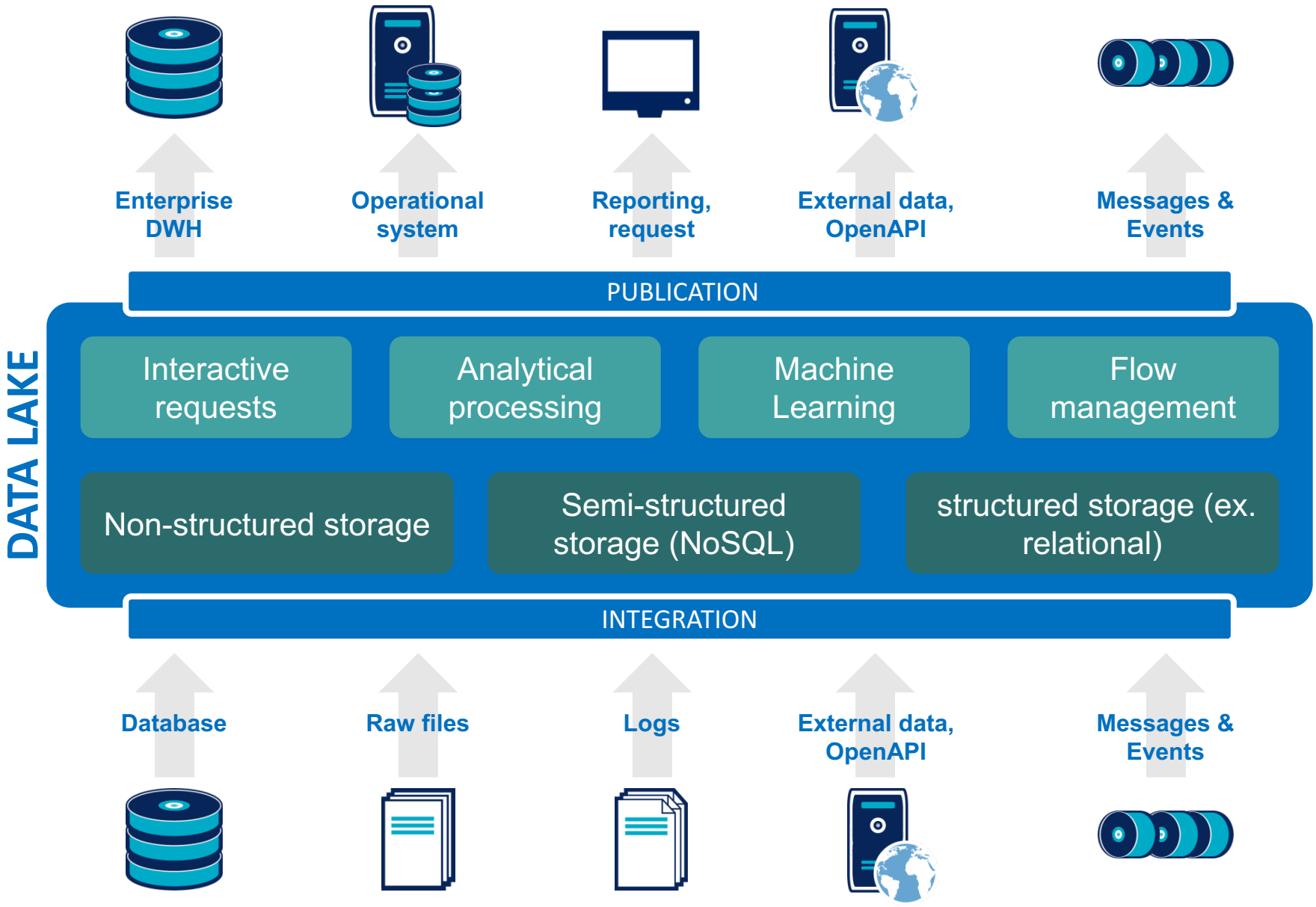
(Calculation Grid)

Over 10 threads/Core CPU, sequential programming reach its limits (IO).

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

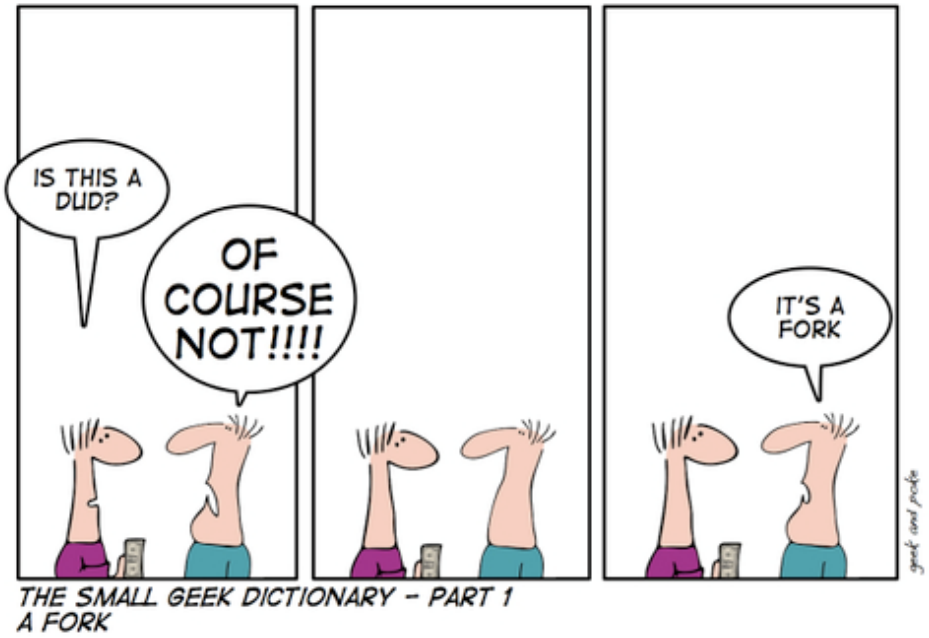
ANTI PATTERN #2: THE LOTTERY

V2 – Last updated 5/3/2017

 mattturck.com/bigdata2017

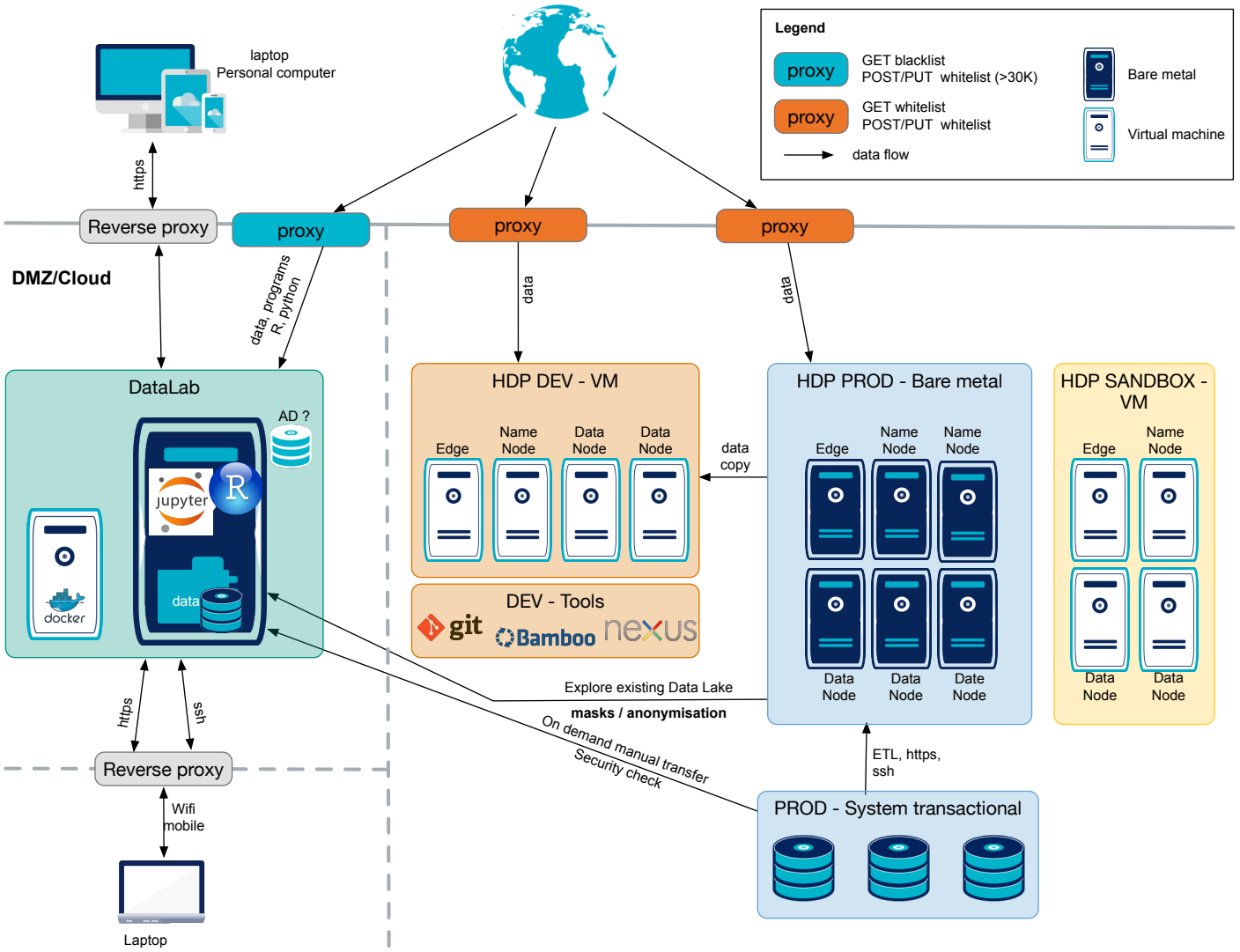
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

ANTI PATTERN #3: FORKING THE SOLUTION

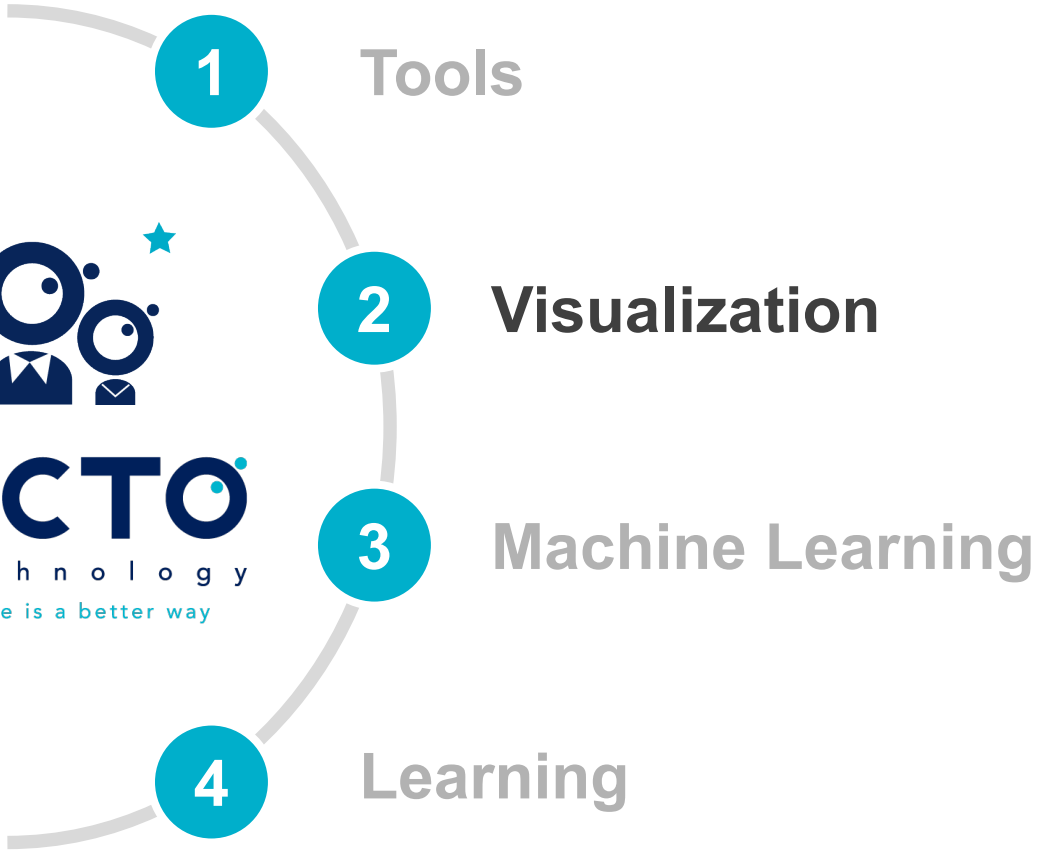


"The NIH syndrome (Not Invented Here) is a disease."
Linus Torvalds

ANTI PATTERN #4: OVER ARCHITECTING FROM DAY 1



FOUR PILLARS OF BIG DATA



OCTO Technology
There is a better way

ANOTHER PERSPECTIVE ON VISUALIZATION

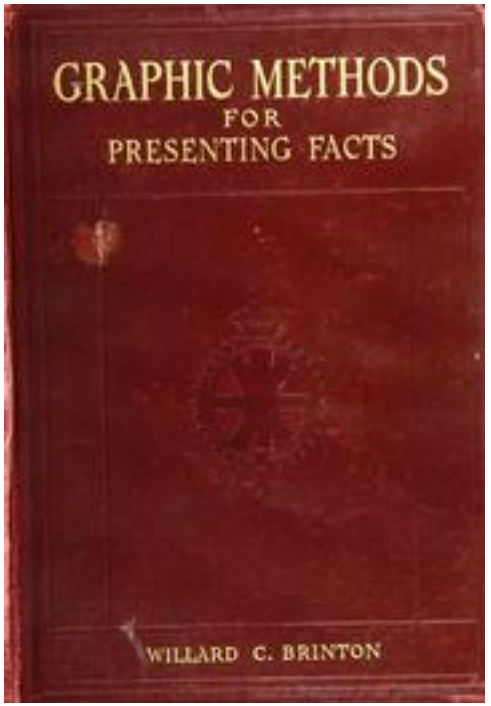
Who said that? When?

“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

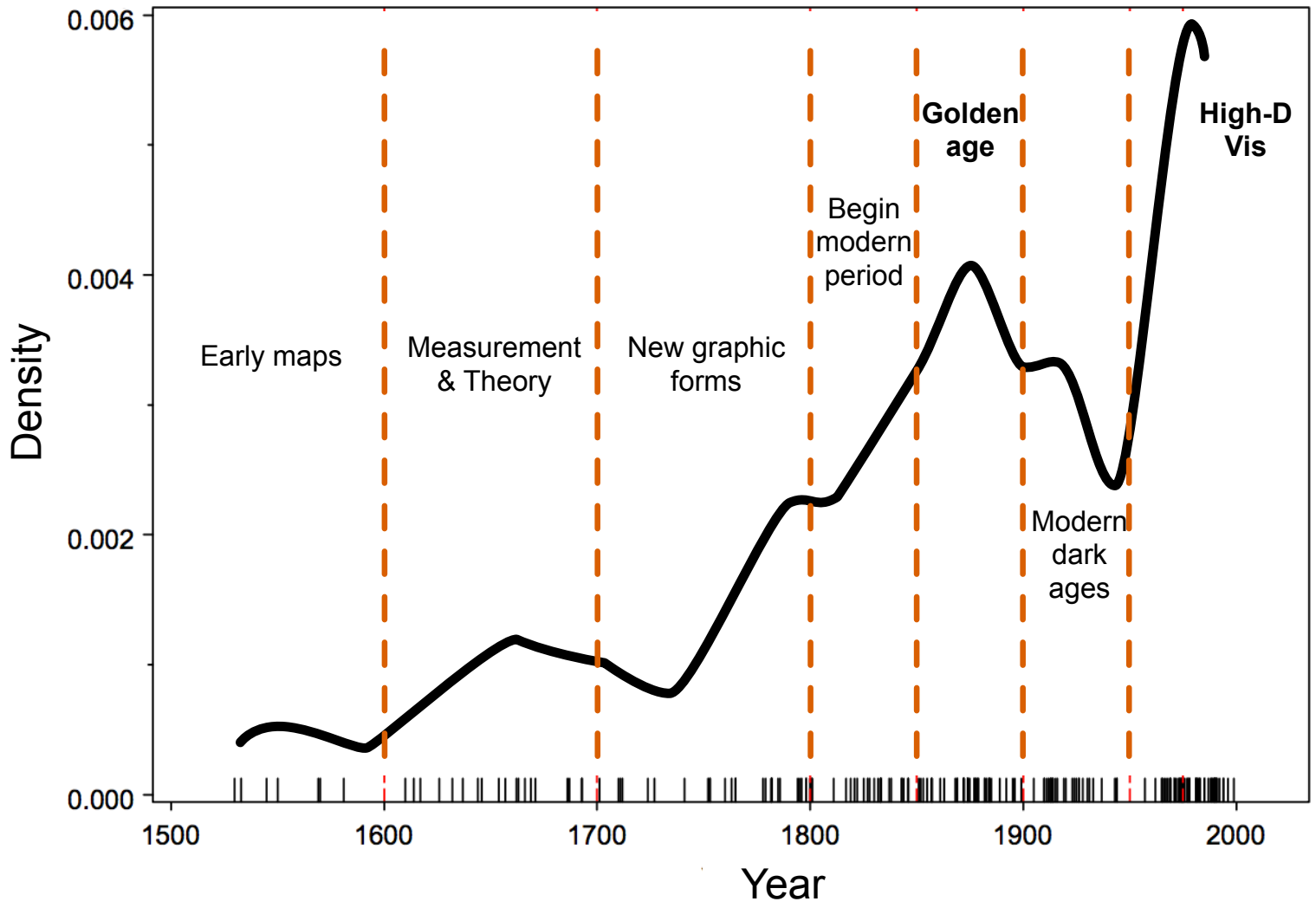


Willard C. Brinton

1914

100yrsofbrinton.tumblr.com

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

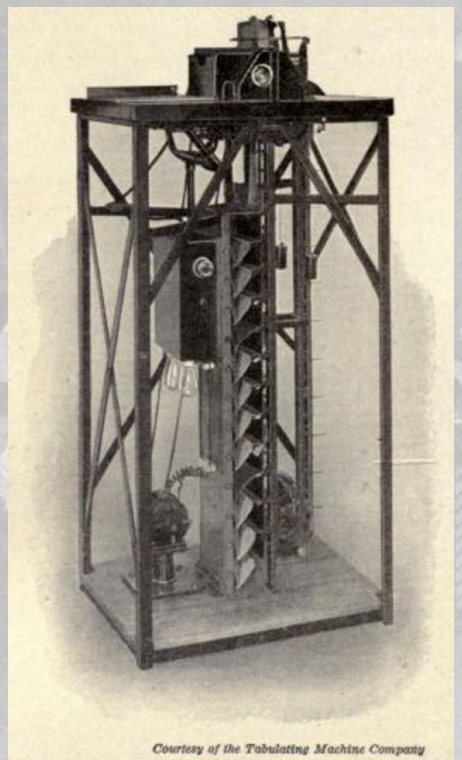
Photo by the International News Service

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

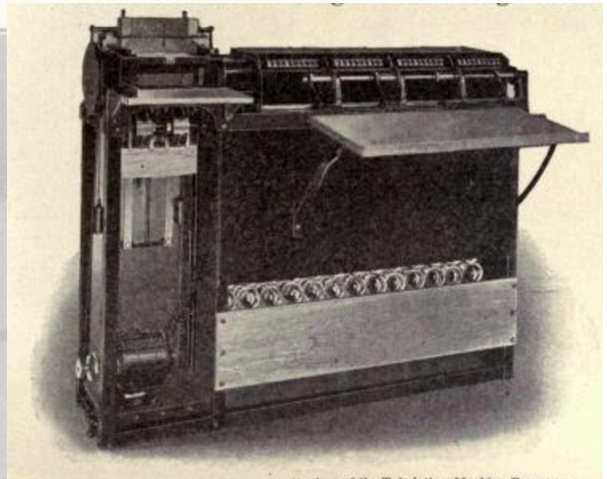
THE PREVIOUS BIG DATA REVOLUTION (END 1800s)



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

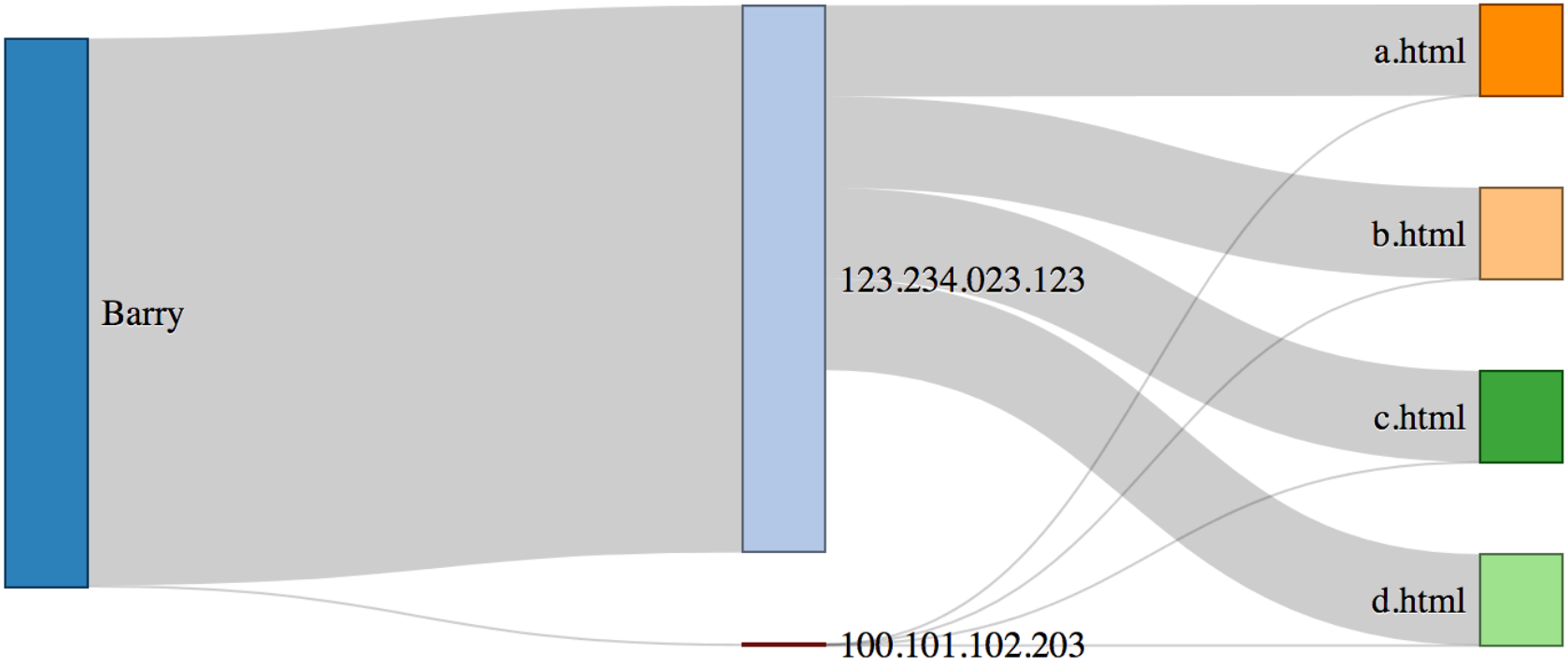
The machine illustrated has four counters, permitting the simultaneous taking off of the data contained under four different headings on the punched card. The sorted cards are placed at the left of the tabulating machine and run through at the rate of about 3,000 per hour. Totals are read from the dials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

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

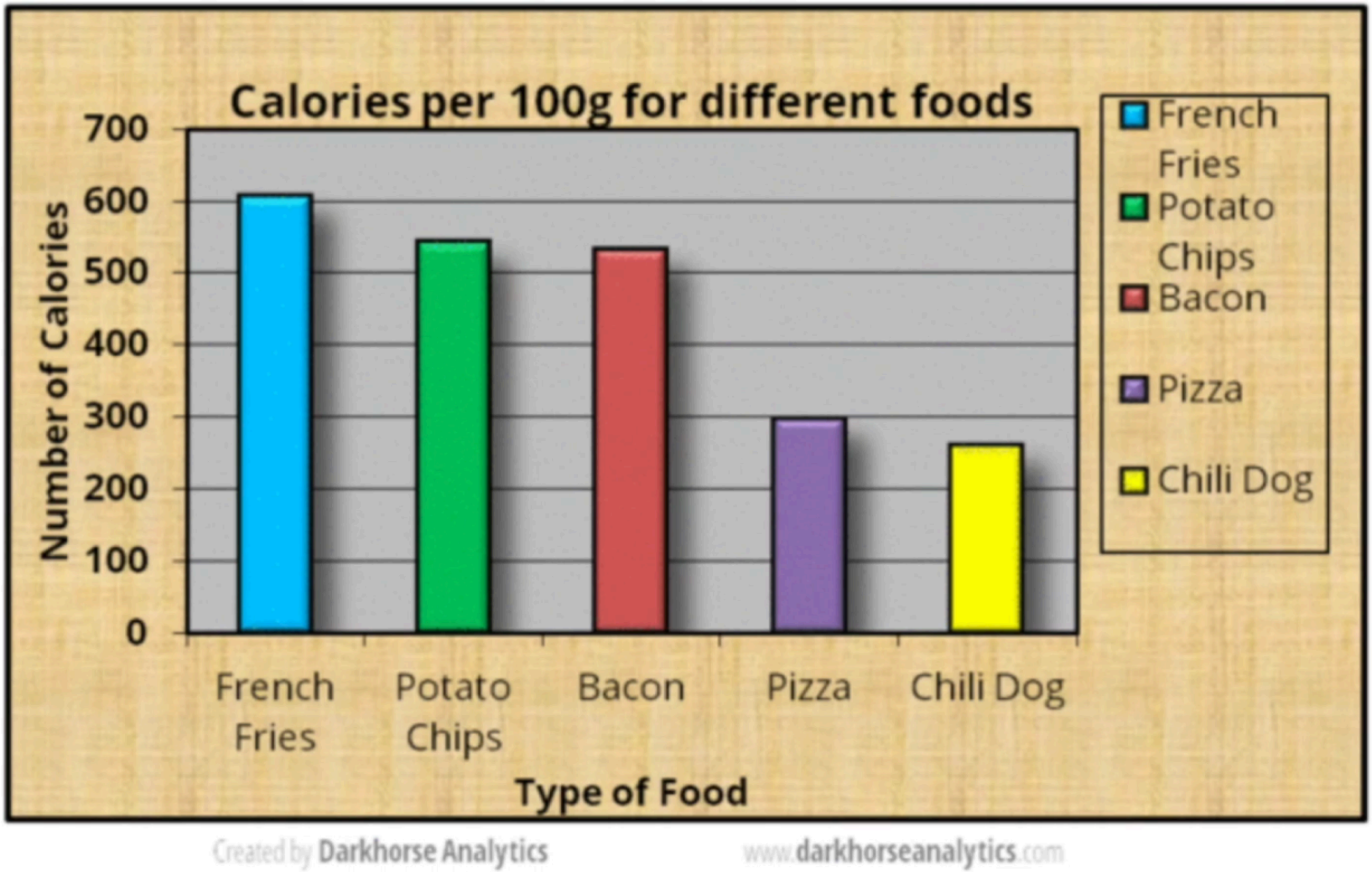
ANTI PATTERN #1: USING THE WRONG GRAPH TYPE

Violating all principles



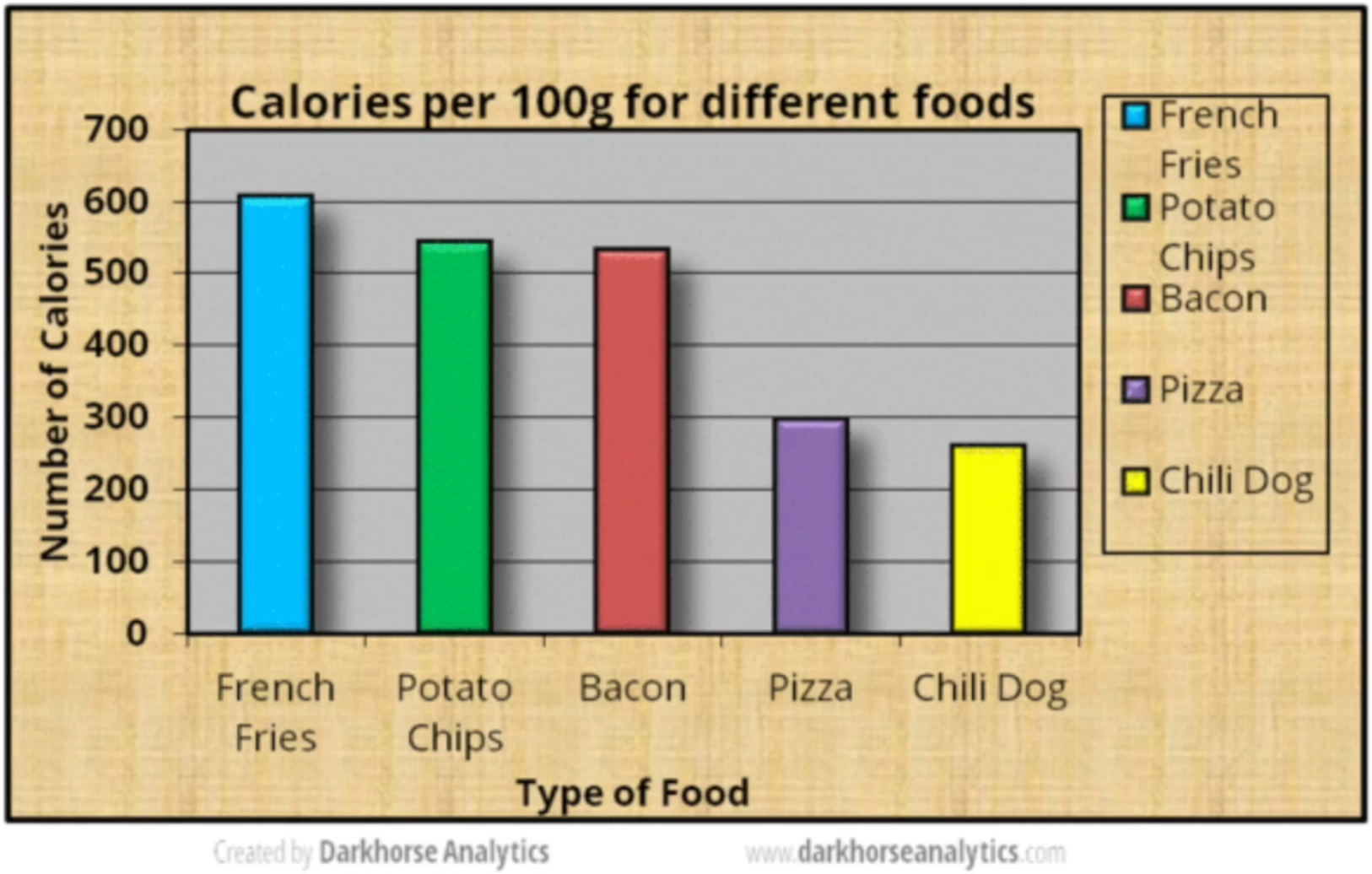
ANTI PATTERN #2: OVERLOADING

Remove backgrounds



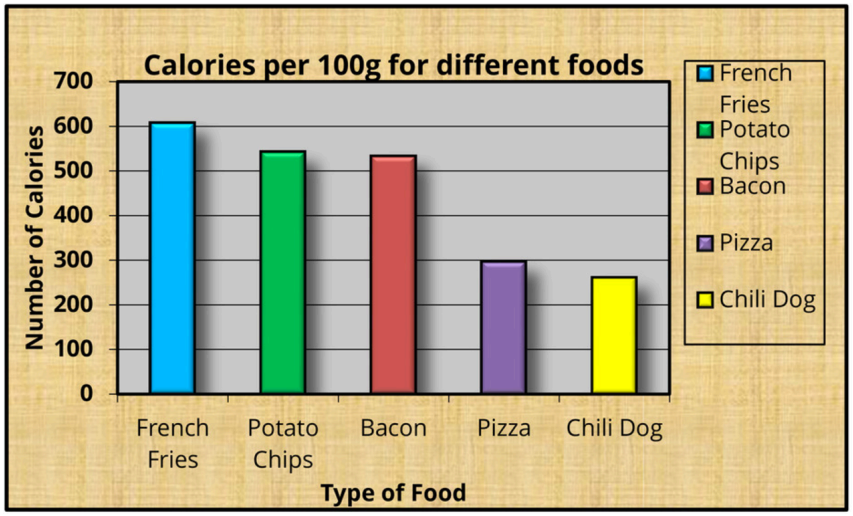
ANTI PATTERN #2: OVERLOADING

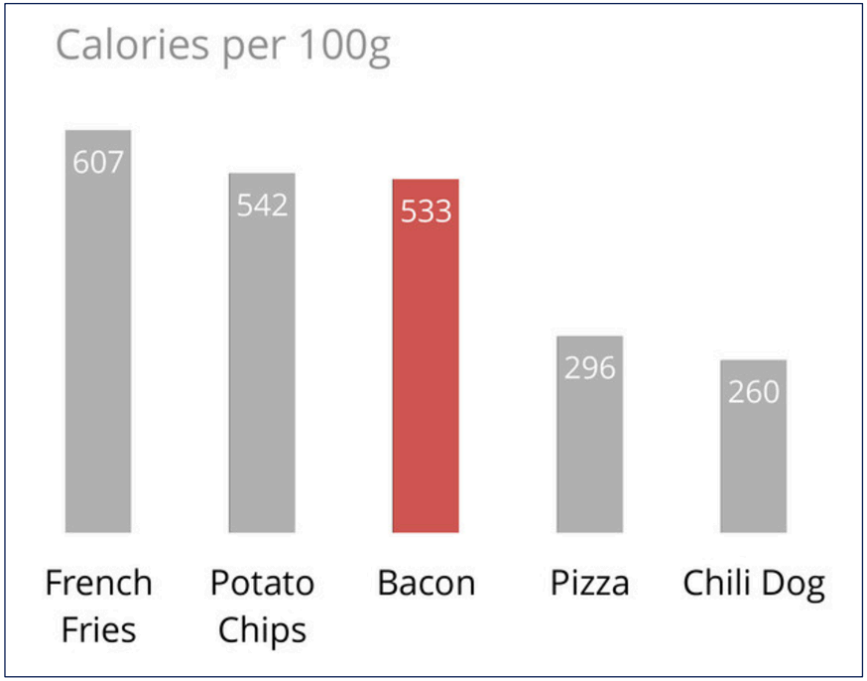
Remove backgrounds



ANTI PATTERN #2: OVERLOADING

Data/Ink Ratio

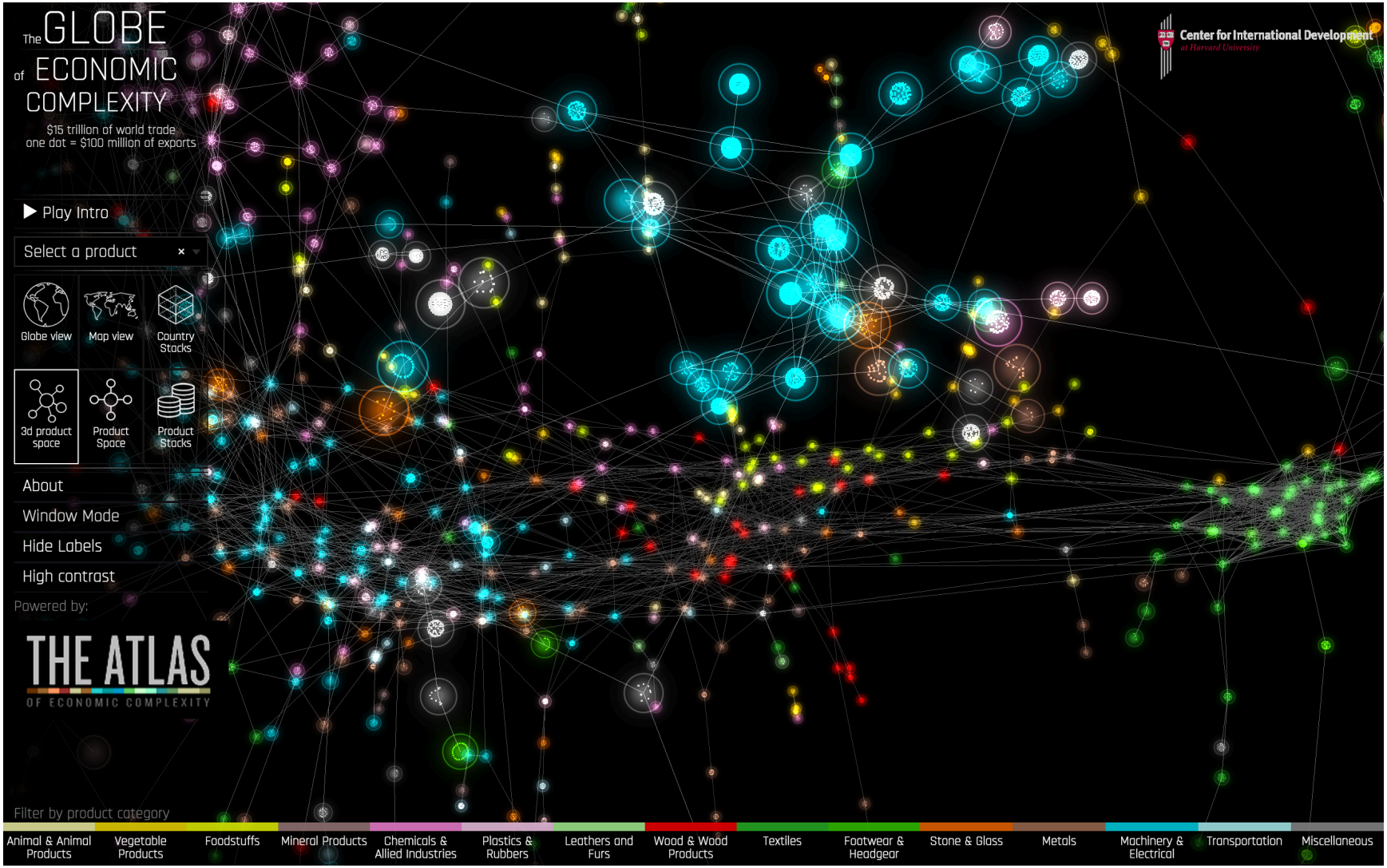




"Perfection is achieved not when there is nothing more to add, but when there is nothing left to take away"
Antoine de St Exupéry
Terre des Hommes, 1939

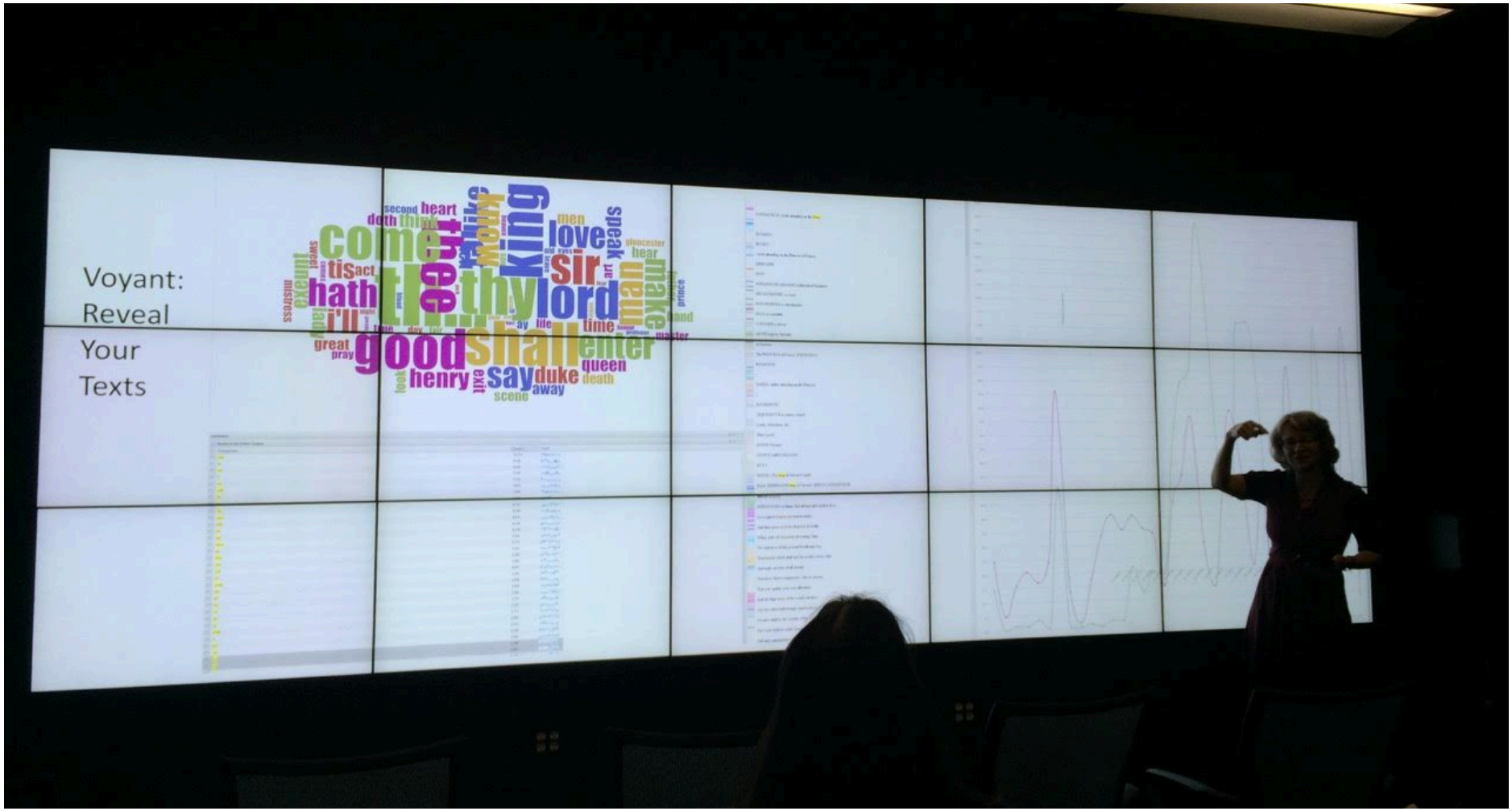
ANTI PATTERN #3: *"I REALLY WANT TO USE THIS FANCY NEW TECH"*

Three.js



ANTI PATTERN #3: *“I REALLY WANT TO USE THIS FANCY NEW TECH”*

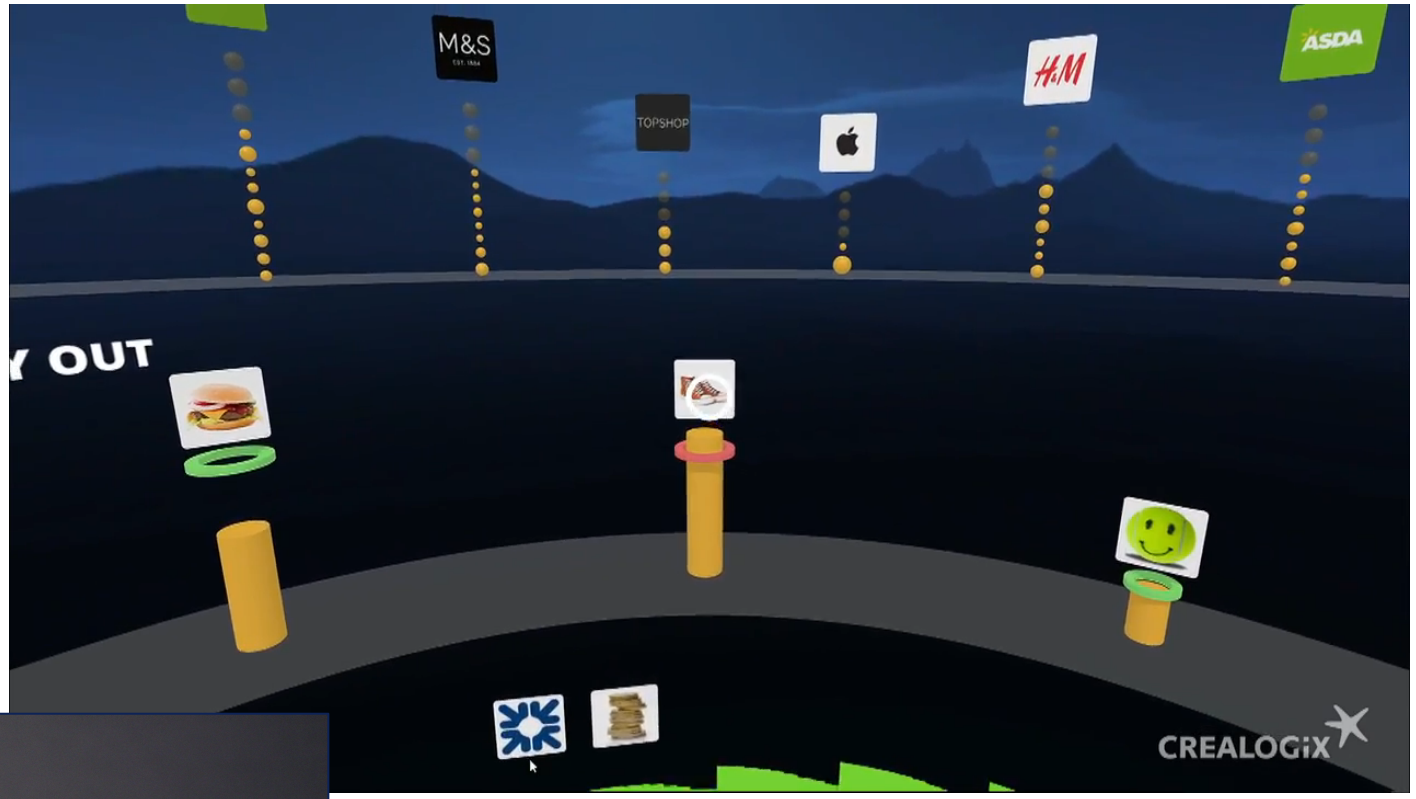
Interactive display wall



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

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

Virtual reality



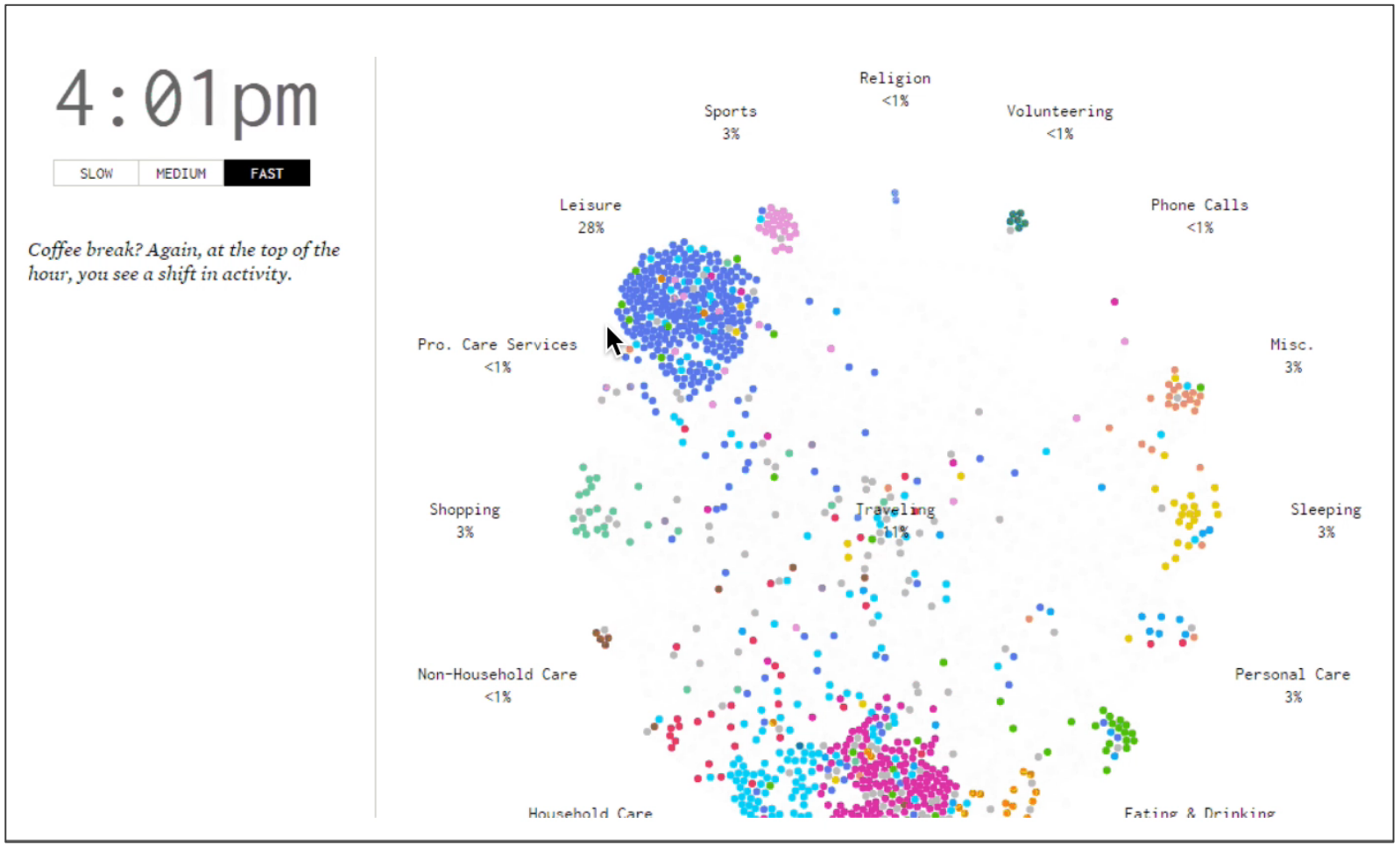
ANTI PATTERN #3: *"I REALLY WANT TO USE THIS FANCY NEW TECH"*

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

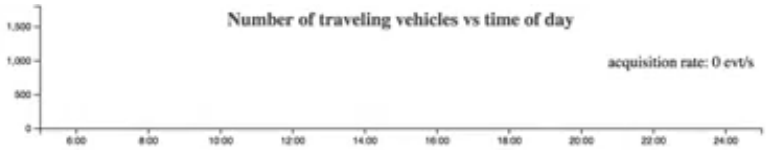
Animation

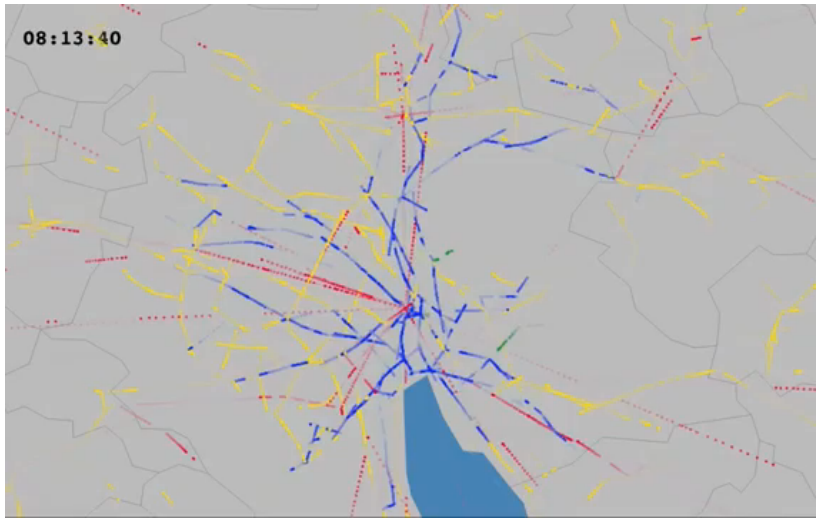
Hans Rosling… The Revolutionary

ANTI PATTERN #4: UNDERESTIMATE YOUR BROWSER

Browser power with high throughput data

Up to 10’000 events per second



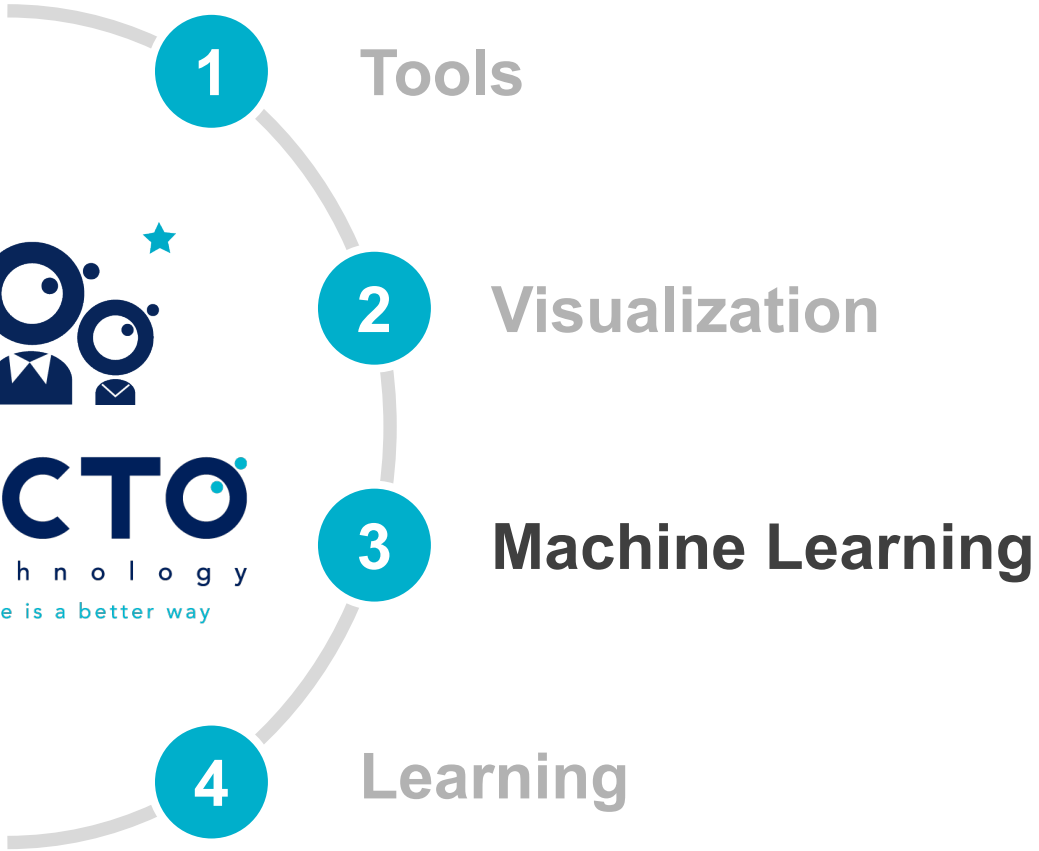


ANGULARJS by Google

d3.js

pixi.jsv3

FOUR PILLARS OF BIG DATA



OCTO Technology
There is a better way

“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

BUILD A MODEL

Flight #	Dep Airport	Dep Hour	Dep Week Day	Aircraft Model
1	SYD	8:10	1	A330
2	SYD	14:15	2	B777
3	MEL	18:10	1	B777
4	PER	6:50	4	A320
5	SYD	9:50	3	A320
6	PER	12:10	1	A320
7	TZN	14:50	1	B777
8	MEL	14:15	4	A320
9	SYD	8:30	3	A320
10	MEL	16:40	1	A320
…	…	…	…	…
11	MEL	9:30	3	B747
12	TZN	9:30	1	A320
13	PER	9:50	3	A320
14	SYD	13:10	1	A320

X

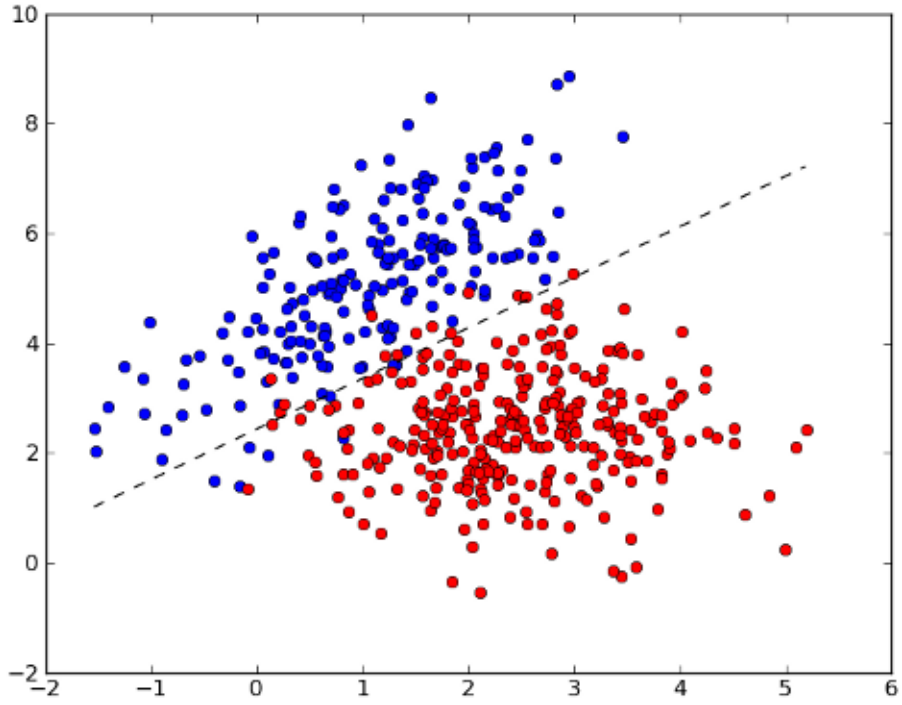
$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \dots \\ \theta_n \end{pmatrix}$$

Delay
0
0
0
1
0
1
0
0
0
0
…
1
1
0
0

Y

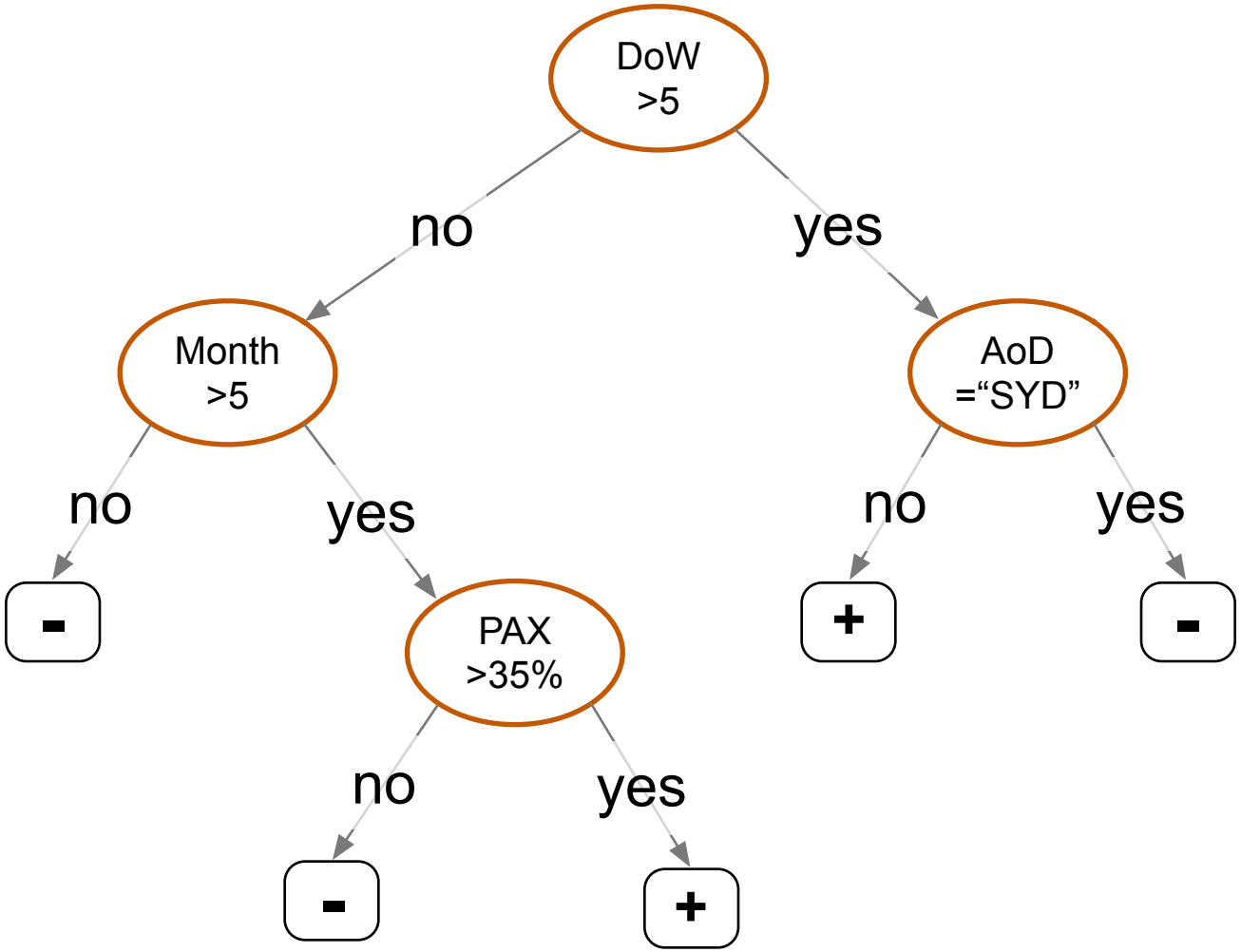
LOGISTIC REGRESSION

Classification algorithm



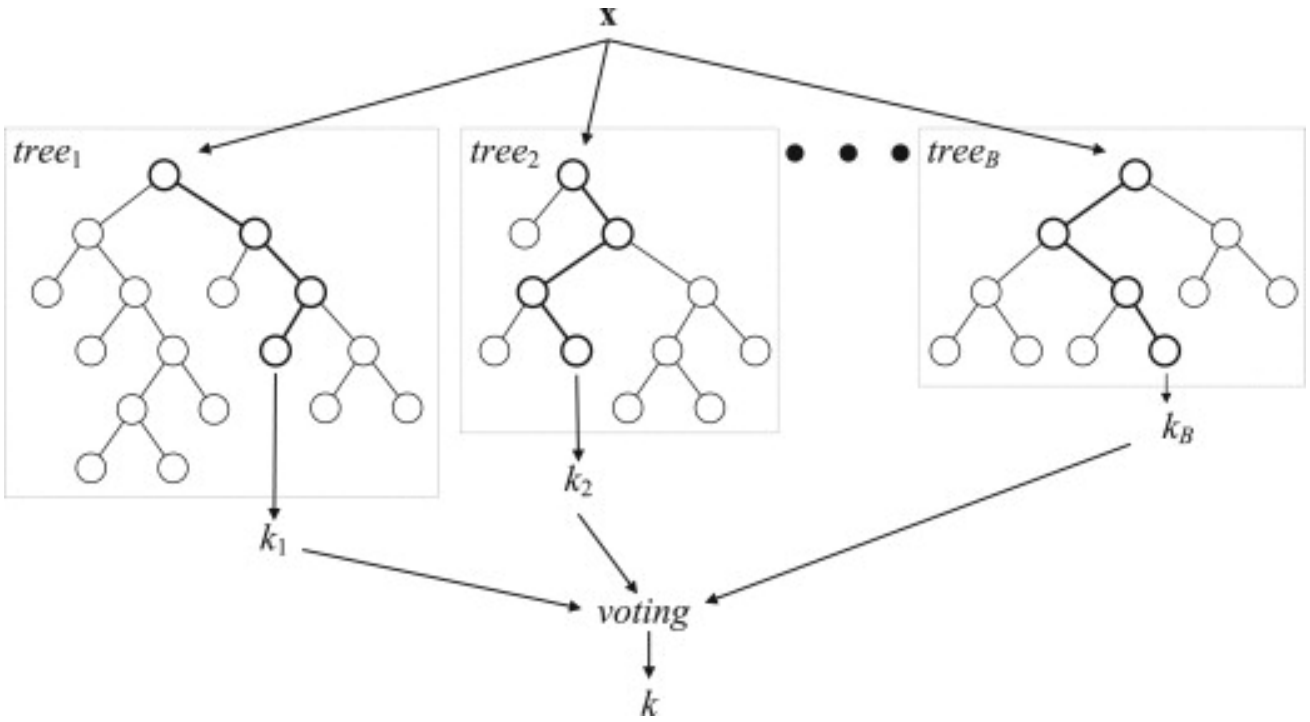
DECISION TREE

Classification algorithm



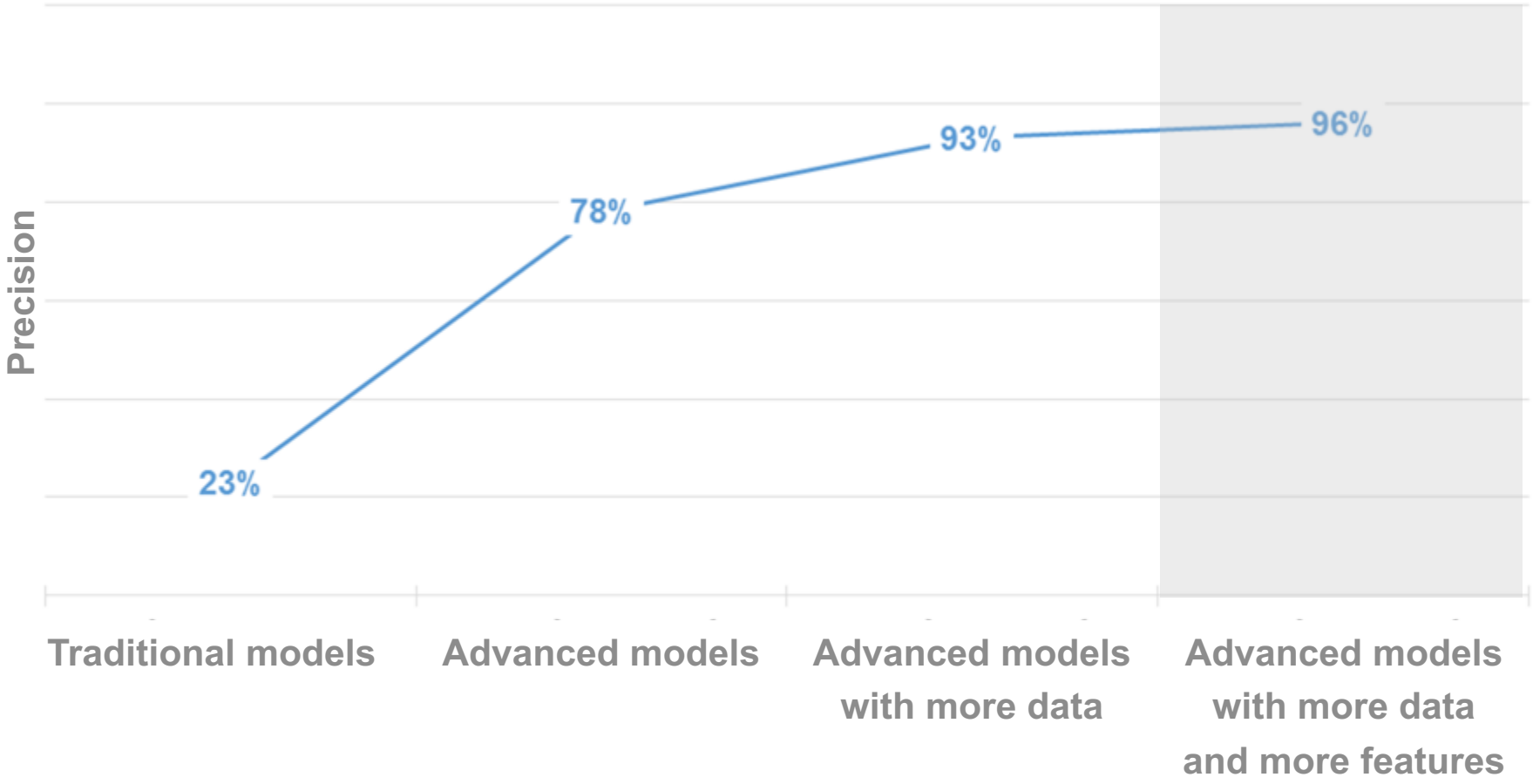
RANDOM FOREST

Classification algorithm

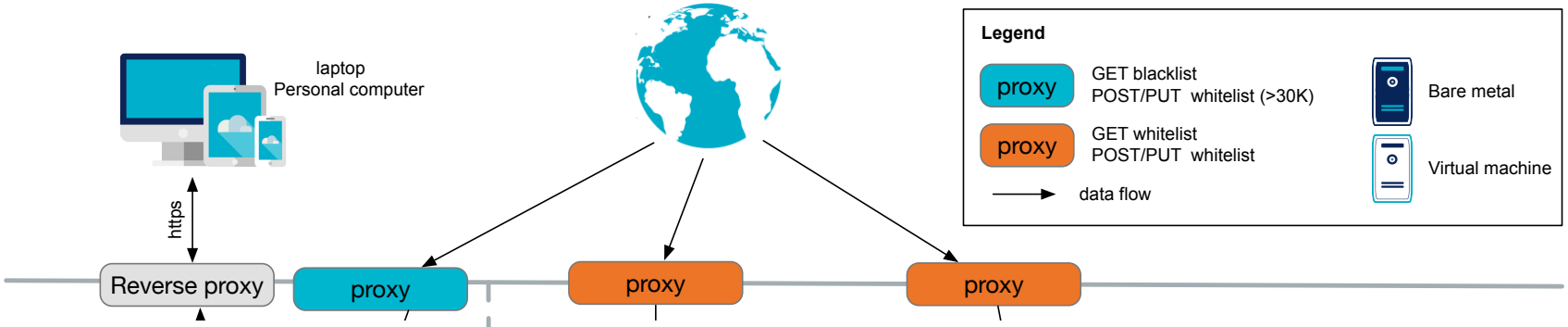


MODELS & DATA

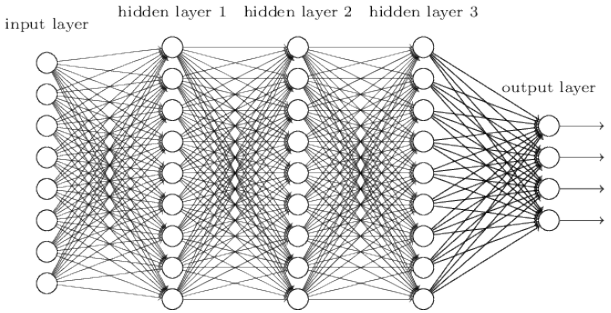
Precision score for the TOP 20%



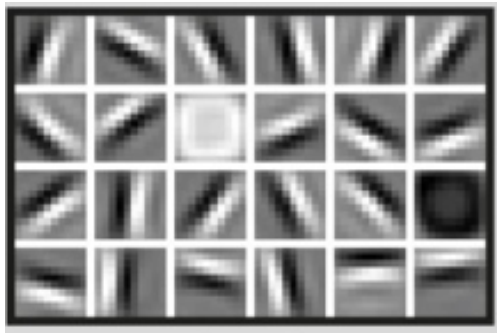
ANTI PATTERN #1: WAIT FOR THE FULL STACK BEFORE STARTING



ANTI PATTERN #2: START TOO COMPLEX



Identify pixels

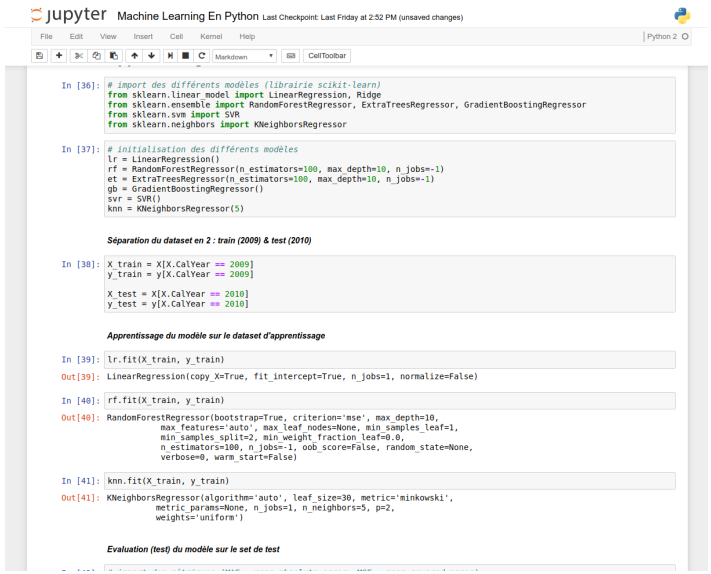


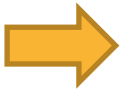
Identify edges and simple shape

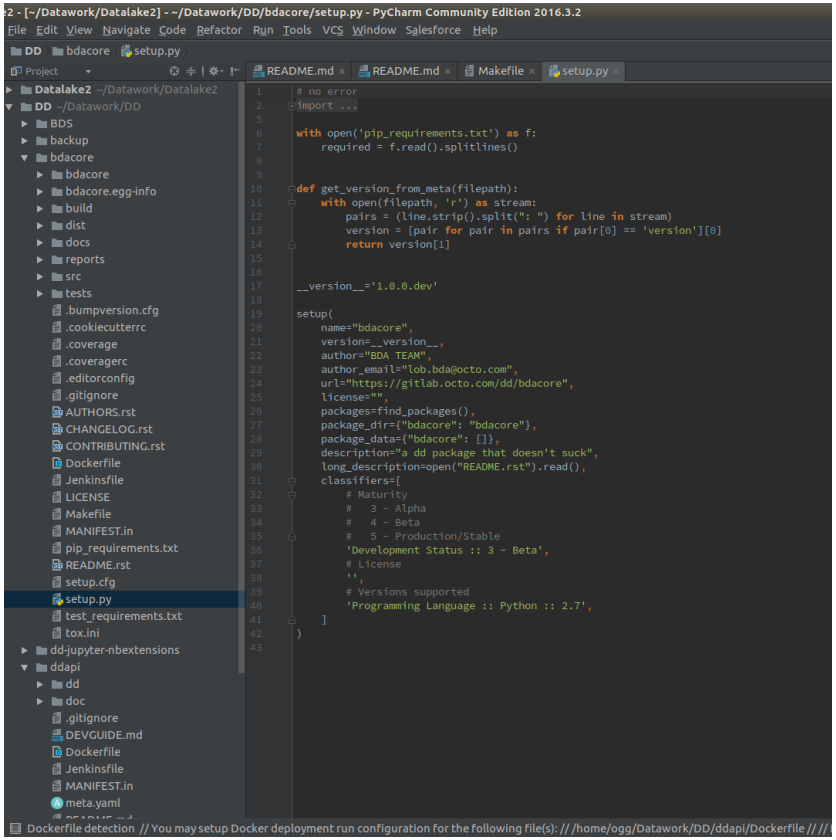
Identify complex shapes and object

Identify which shape to be used to define a human face

ANTI PATTERN #3: FORGET THE LIFE AFTER THE NOTEBOOK







ANTI PATTERN #3: FORGET THE LIFE AFTER THE NOTEBOOK

Competencies

Data Scientist

Basically, a rockstar

OPS

- Advanced statististic
- Artificial Intelligence
- Bleeding edge algorithm

- Run things

ANTI PATTERN #3: FORGET THE LIFE AFTER THE NOTEBOOK

Competencies

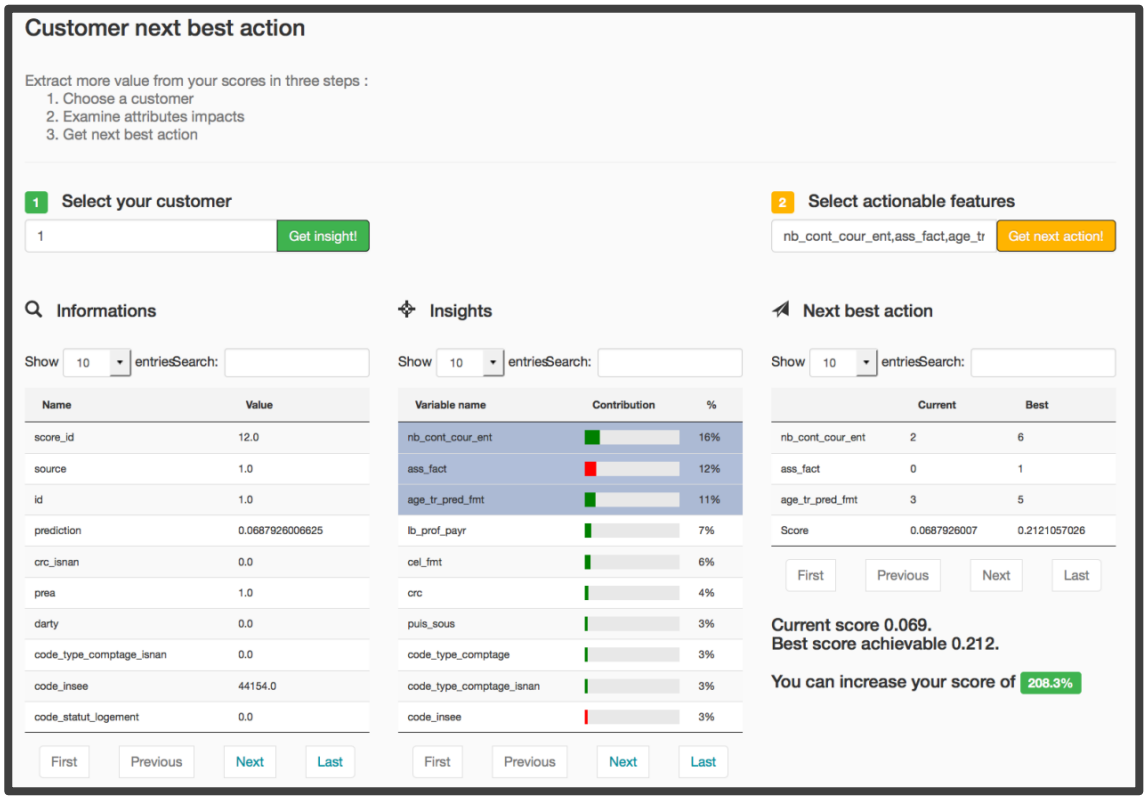
- (Machine Learning)
- (Statistics)

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



FOUR PILLARS OF BIG DATA

OCTO Technology
There is a better way

OCTO academy

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 maitrisant 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 CLOUDERA

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

academy.octo.ch


SOCIETE
GENERALE


AIRBUS


bouygues
TELECOM


ENGIE


RTE
meetic


EDF


TOTAL


Technology


Boursorama


GROUPE
GENERALI


Cdiscount.com
N°1 du e-commerce en France !


CANAL+


CNP
ASSURANCES


AIR LIQUIDE


Pentair


PUBLIGroupe

OCTO Suisse RECRUTE

rejoins.octo.com

Architecte

Coach Méthodo

Software Craftsman

Questions ?

DataGeek

Expert DevOps

Consultant en Stratégie

OCTO Technology

There is a better way