Big Data &AI Point Of View

Christophe Burgaud - PhD

Novembre, 2020

AGENDA

- *L’enjeu de la transformation digitate chez nos clients*
- *Retour sur les cas d’usage*
- *Concepts et definitions*
- *Marche, Technologies et Architecture*

Transformation digitale et cas d'usage

The digital age has changed the way we Live, Play, Learn and Work...

Netflix provides personalized recommendations

Waze provides a personalized driving experience

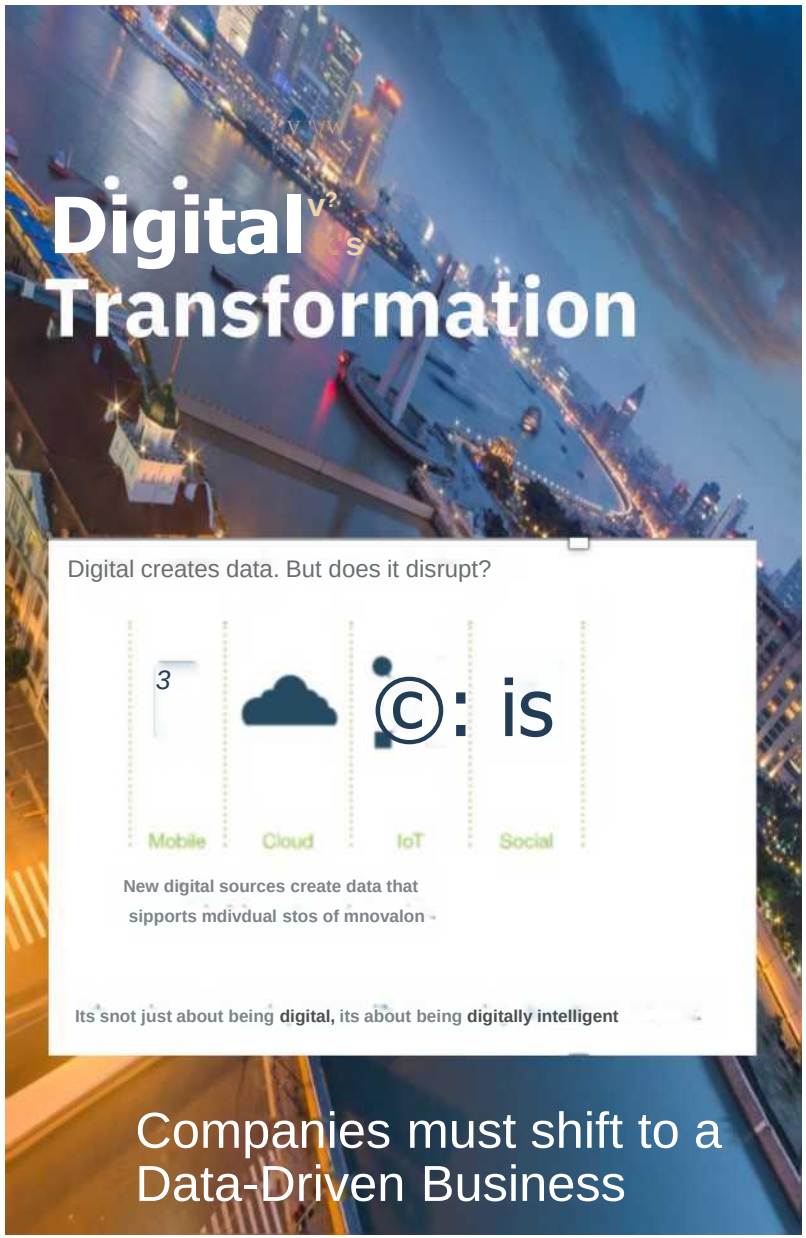
Uber delivers food that you like and is the right temperature

Self driving cars react to changing conditions

But ALL based on DATA and AI

Some Very successfully

Some not as successful



75%

of large enterprises will have **digital transformation** at the center of corporate strategy within two years

81%

of companies do not yet understand the **data required for AI**

>85%

of enterprise IT organizations will **commit to multi-cloud** architectures

La data au creur de la transformation digitale

- ***Laurent Mignon, President de la BPCE (Banque Populaire Caisse d'Epargne), 29 janvier 2019***

« *Tout depend de la maniere dont vous allez utiliser la donnee* » reprend le dirigeant.
« *La technologie vous pouvez l'appeler Machine Learning, Deep Learning, intelligence artificielle, etc.* ***le veritable probleme est comment bien utiliser la donnee*** *dont nous disposons pour ameliorer notre capacite a mieux servir nos clients. C'est le premier enjeu* » insiste-t-il.

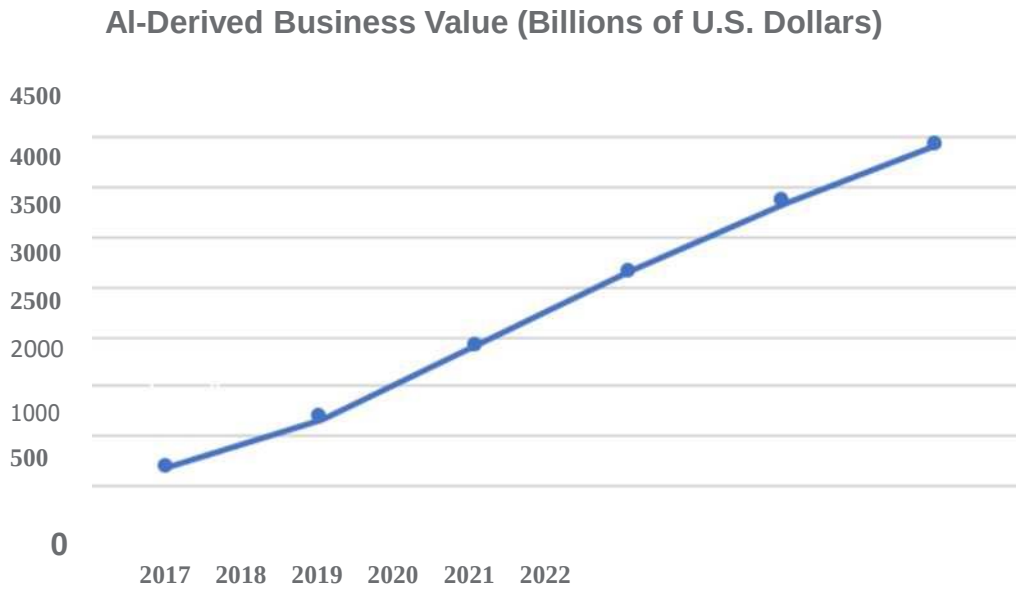
« *Si nous reussissons cela, je pense que nous transformerons en profondeur notre entreprise. De maniere implicite, l'usage de la donnee etait realise dans la tete de nos conseillers en agence. Nous devons elargir l'acces a la donnee* » dit-il.

http://www.larevuedudigital.com/la-banque-bpce-en-attente-de-son-data-lake-pour-passer-a-la-vitesse-superieure/

Business Value of Data & AI

Some main sources of AI business value:

S Customer Experience *S* Cost Reduction *S* New Revenue/New Services

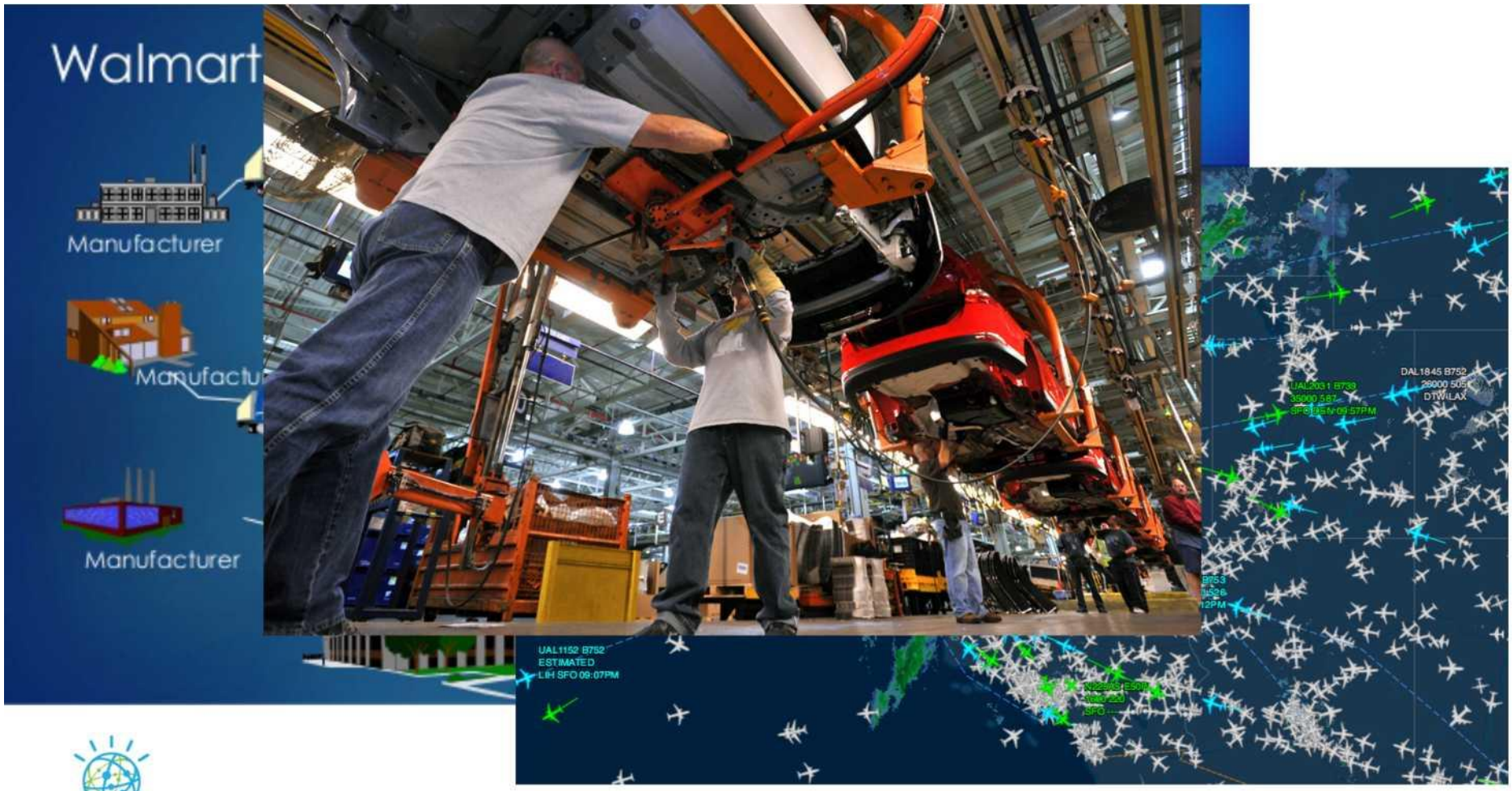


Gartner, April 2018 (https://www.gartner.com/newsroom/id/3872933)

Credit Mutuel Addresses 50% of the 350,000 emails received daily, detect automatically the intent and automate the action

Hyper or Radical Personalization

Resource Allocation & Strategic Planning



Predictions and Classifications

Industries: All

Smarter Healthcare

Multi-Channel

Homeland Security

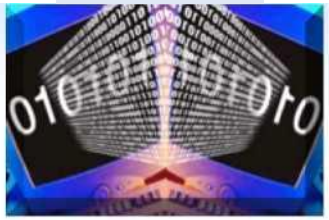
Traffic Control

Manufacturing

Trading Analytics

Finance

Log Analysis



Telecom

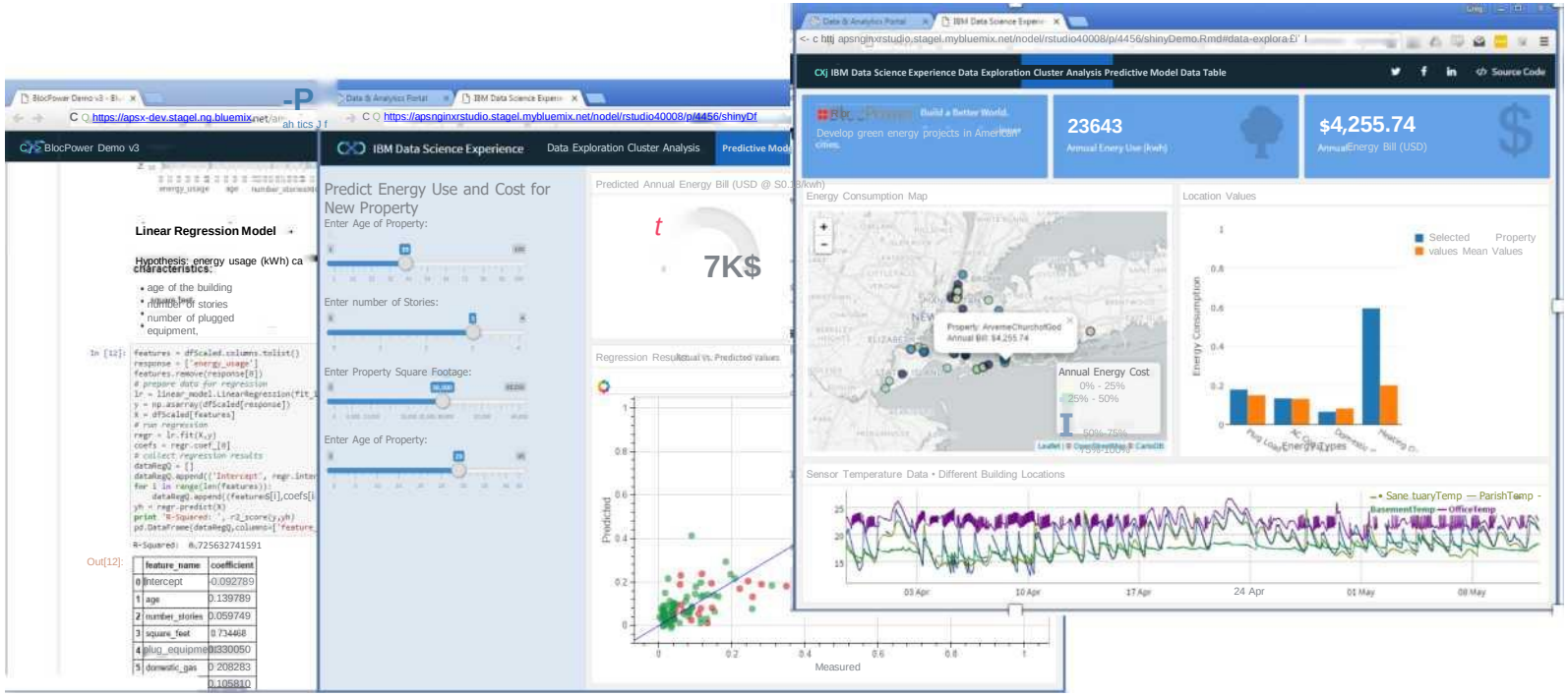
Search Quality

Fraud and Risk

Retail: Churn, NBO

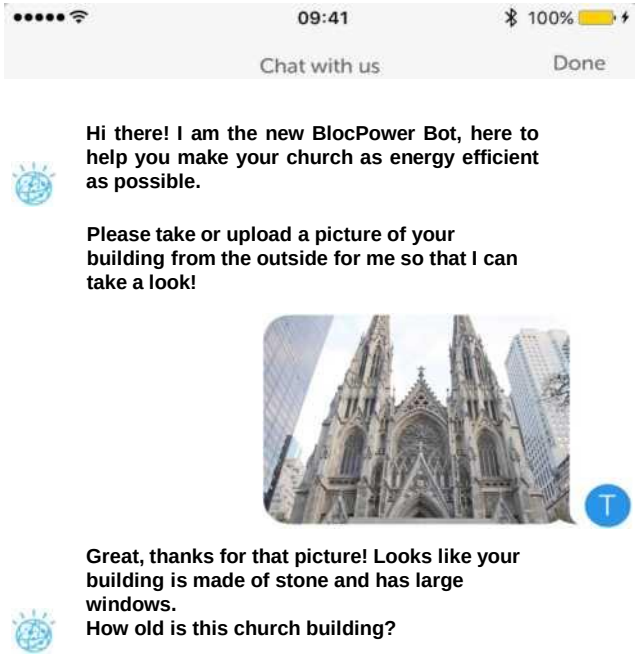
Combine

Building Model to Predict Energy Consumption of Buildings.



Combine

Chat Bot to estimate energy cost from an image of building.



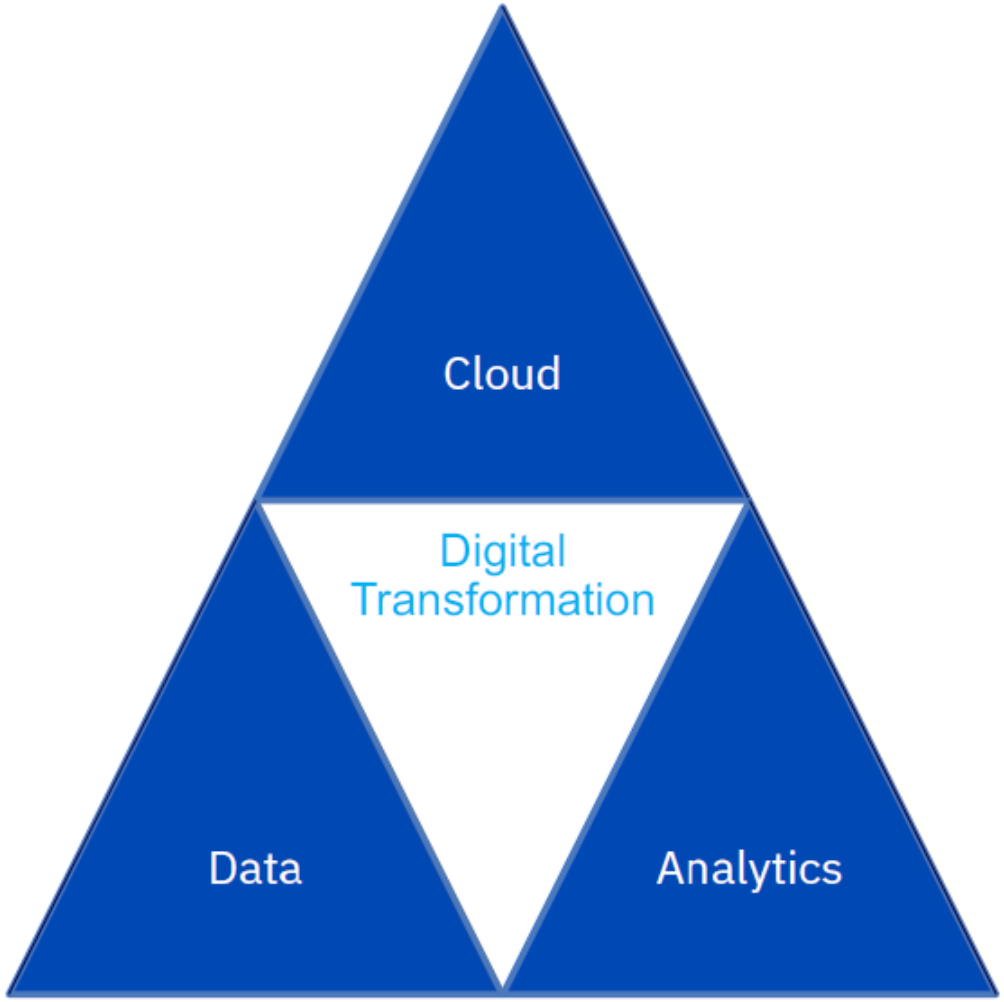
Great, thanks for that picture! Looks like your building is made of stone and has large windows

I estimate your building has a high energy usage intensity (EUI), with a 97.01% probability

Buzzwords and challenges

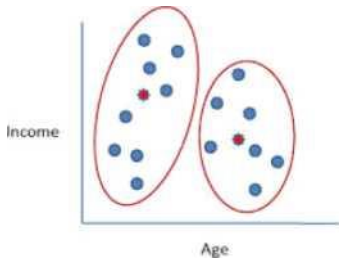
- AI, machine learning, cloud, self-service, data governance, etc...there is no shortage of buzzwords in data today.
- Every organization is seeking to outpace their competition by leveraging data to drive differentiation for their business.
- To win this race, companies are building up data science teams, investing in faster/more scalable cloud data platforms and utilizing the growing variety of publicly available datasets and algorithms.
- How do you stay ahead of what's next and help drive the successful adoption of new technology and processes within your organization?

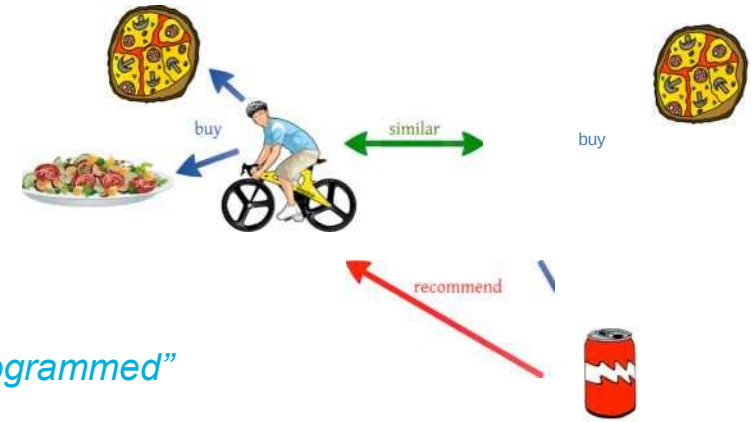
Transformation digitale: points clés



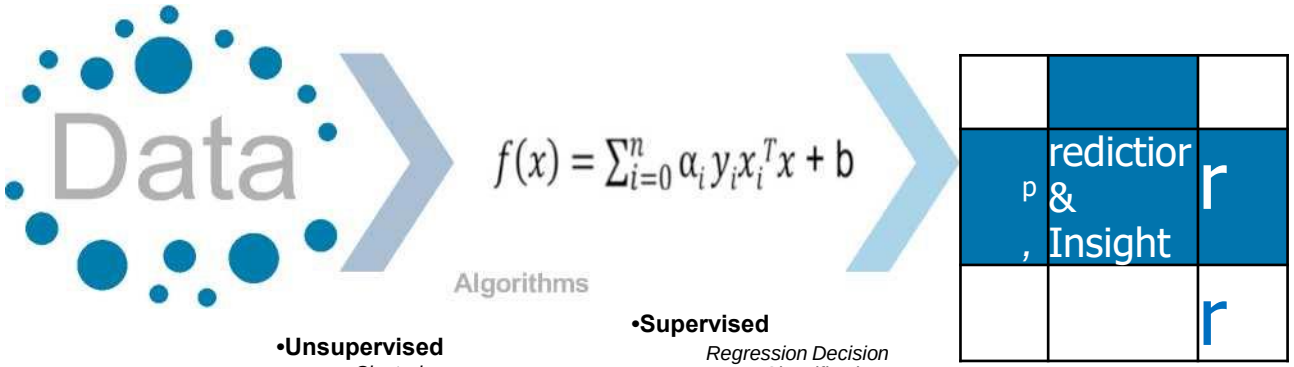
Concepts et definitions

Machine Learning... What is it?





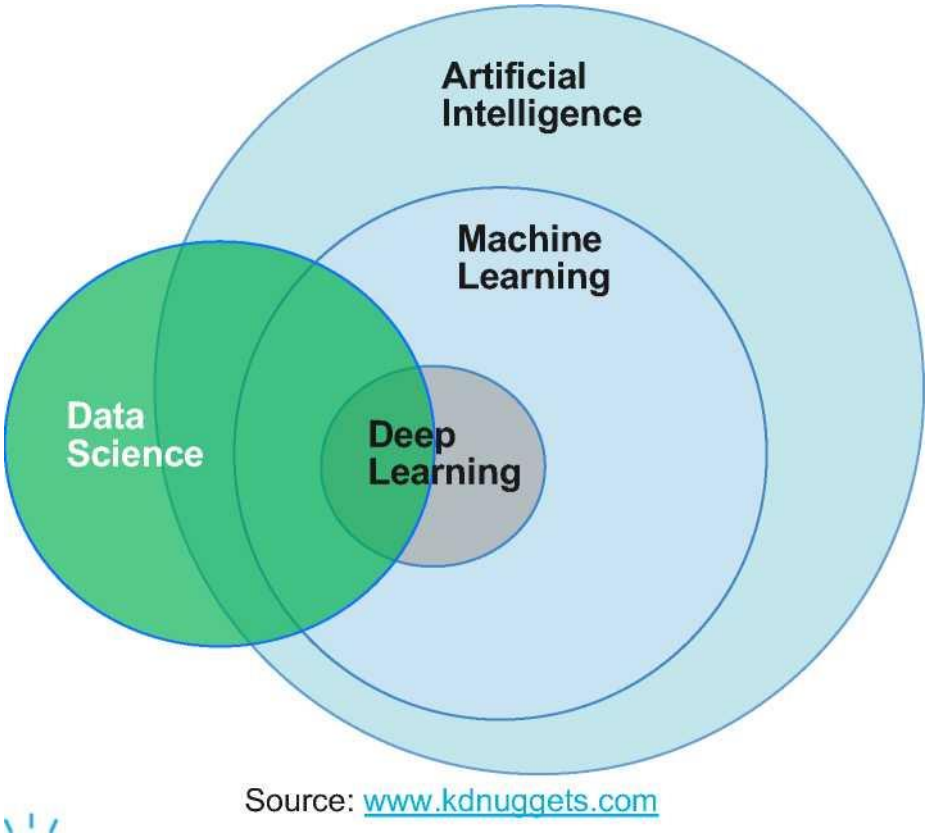
"Computers that learn without being explicitly programmed"
"Using algorithms to understand patterns in data"



•Unsupervised
Clustering
Association

•Supervised
Regression Decision
Trees Classification
Recommender systems

Artificial Intelligence



Source: www.kdnuggets.com

All Machine Learning is AI but not all AI is Machine Learning.

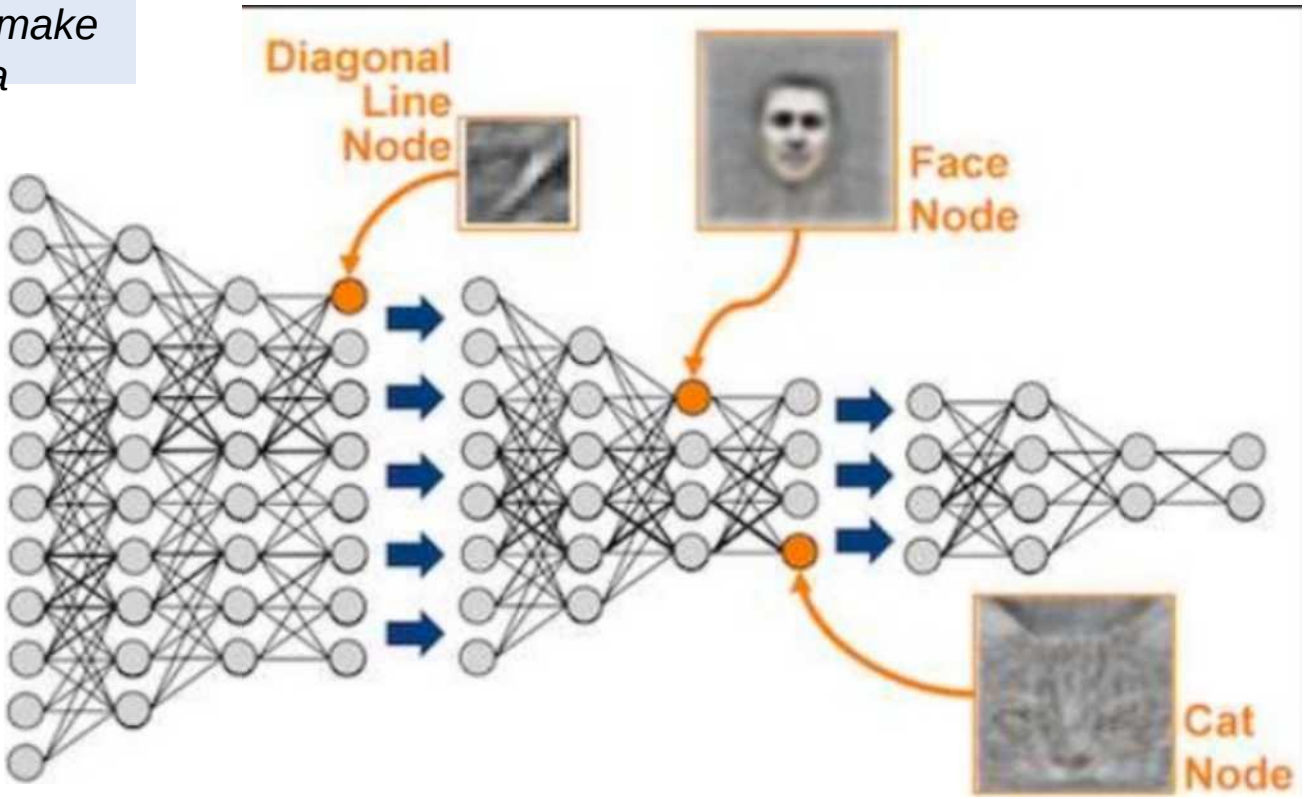
It's the fastest growing part of AI and shows the most promise.

Deep Learning is a subset of Machine Learning and focuses even more narrowly on a subset of Machine Learning techniques and requires "thought".

Deep Learning Example

Technical Definition:

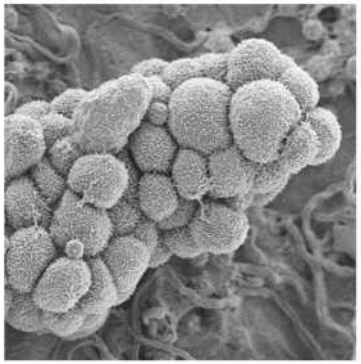
Deep learning is a class of machine learning algorithms in the form of a neural network that uses a cascade of layers (tiers) of processing units to extract features from data and make predictive guesses about new data



Deep Learning Use Cases

INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation



MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading Drug Discovery

MEDIA & ENTERTAINMENT

Video Captioning Video Search Real Time Translation

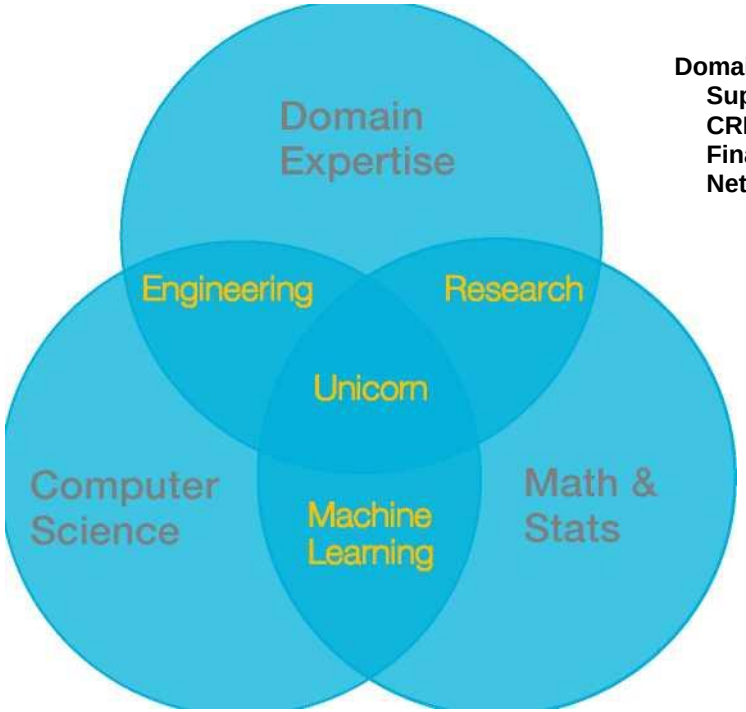
SECURITY & DEFENSE

Face Detection Video Surveillance Satellite Imagery

AUTONOMOUS MACHINES

Pedestrian Detection Lane Tracking Recognize Traffic Sign

What makes a "Data Scientist" ?



Domain Knowledge
Supply Chain
CRM
Financials
Networking

Scripting, SQL Python, R, Scala

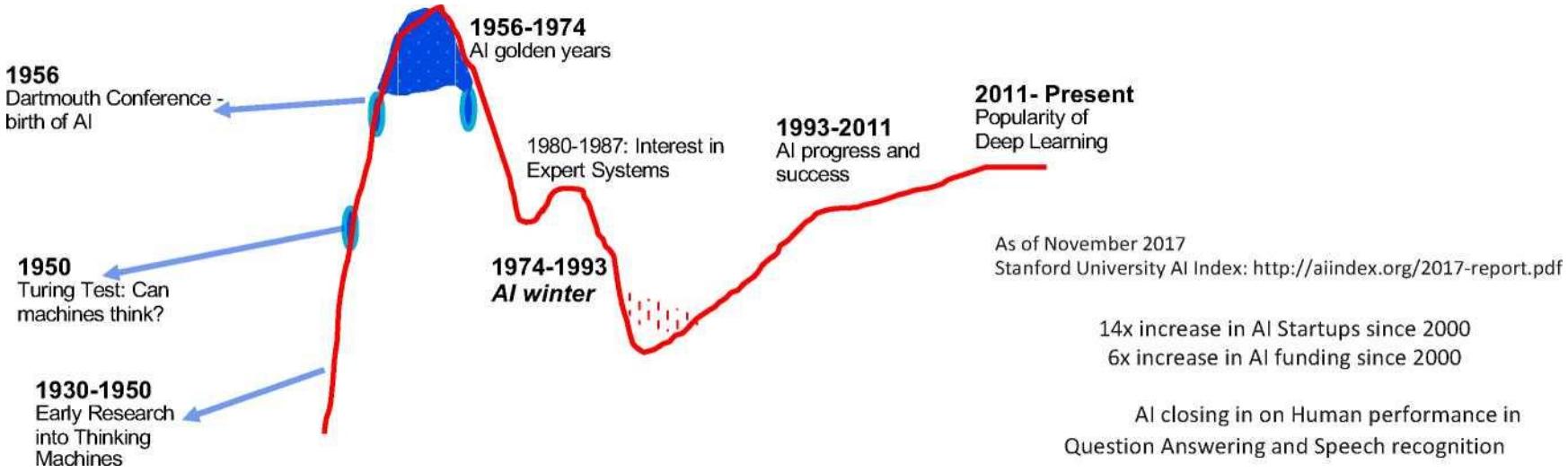
Data Pipelines
Big Data/ Apache Spark

Mathematics
Computational

Data Science Projects Require multiple Skills

. .Data Scientists have different UI and tool preferences. Some data scientists prefer to code their models in Python using scikit-learn; others favor Scala or Apache Spark; some prefer running data models in spreadsheets; and yet others are more comfortable building data science models by creating visual pipelines via point-and-click UI.
Diversity of tools is an important characteristic of this market. *" - Gartner MQ 2017*

AI and Machine Learning have been around for some time, so why the hype now?

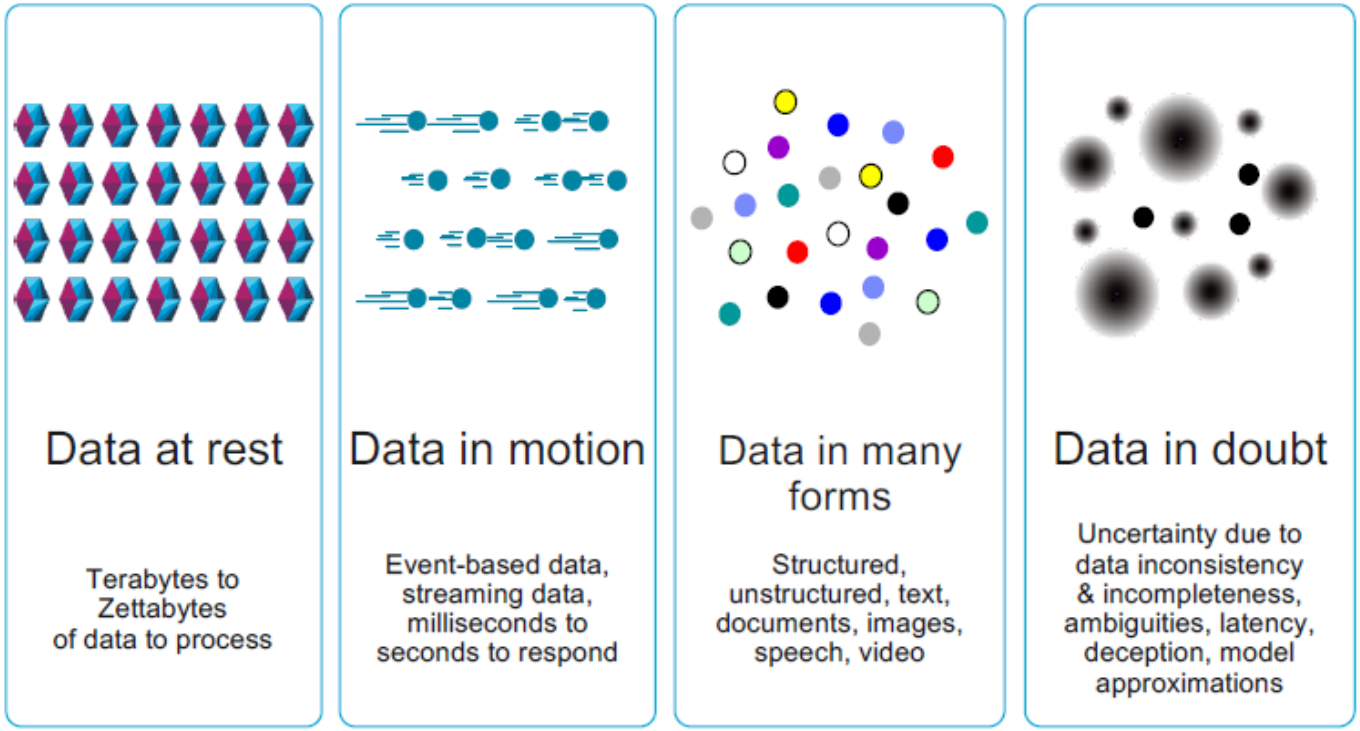


Big Data: Structured and unstructured data, as well as its storage, has grown exponentially decade. Data is also more accessible.

Distributed processing : The Machine Learning algorithms and models take advantage of new technologies : Spark, GPUs and Big Data to train models with large data sets. The new challenge is to industrialize and democratize AI in applications

Big Data Disruption: Align Data and AI to the desired speed of the business

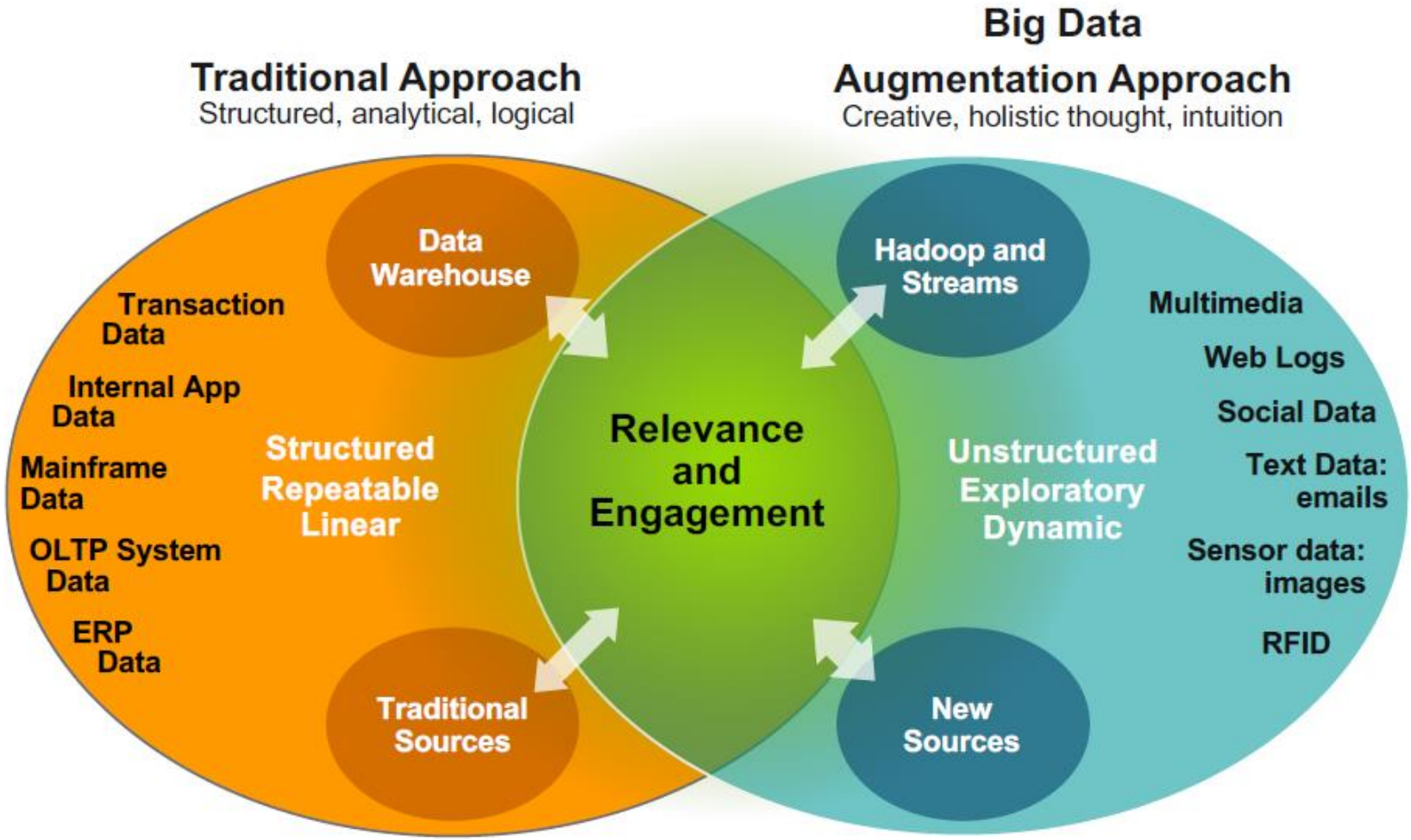
- Fit-for-purpose data architectures are necessary to accommodate the specialized data needs of the business, and individuals within its organizations.
- Gone are the days of a single, structured, data-at-rest architecture.



Une Velle Véfinition du Vig Vata... les 11 V

"*Vast, Volumes of Vigorously, Verified, Vexingly Variable Verbose yet Valuable Visualized high Velocity Data*"... I just would add "... creating **Value"** :-) **Décideo**

Big Data Approaches Help Accumulate Context



Market and Technology

Marketplace Observations and hot topics

- "Big Data" is not any more Hadoop technology only: architectures rely more and more on Object Storage

  "Big Data 2018: Cloud storage becomes the de facto data lake"

- Spark is the layer for massive distributed processing

- Data Science and Machine Leaning industrialization
  - Gain insight from more data exploration on more data (volumes, internal, external,...)
  - Challenge is to deploy and monitor predictive and prescriptive modeling

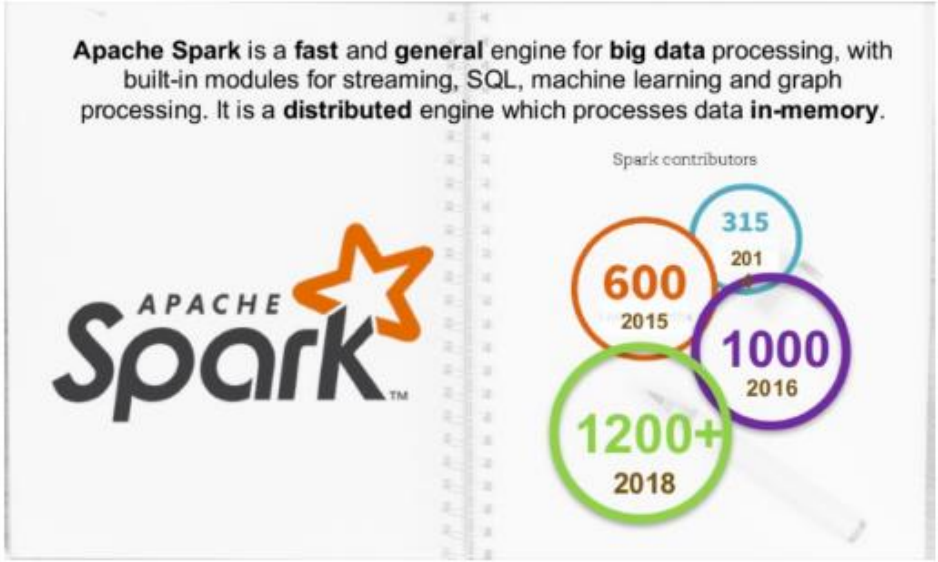
- Requirement to interoperate with Enterprise tools/application that understand SQL :
  - **BI tools like Cognos, Tableau , QlikView, ...**

- Data Governance is a hot topic
  - Sensitive Data (RGPD)
  - Security

- Understand data and apply rules/policies

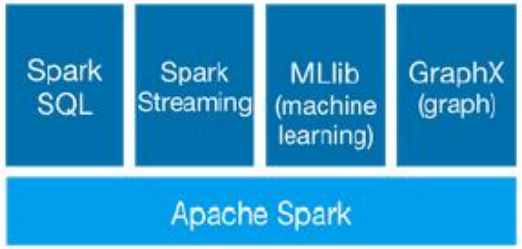
http://www.zdnet.com/article/big-data-2018-cloud-st[illegible]de-facto-data-lake/

What is Spark ?



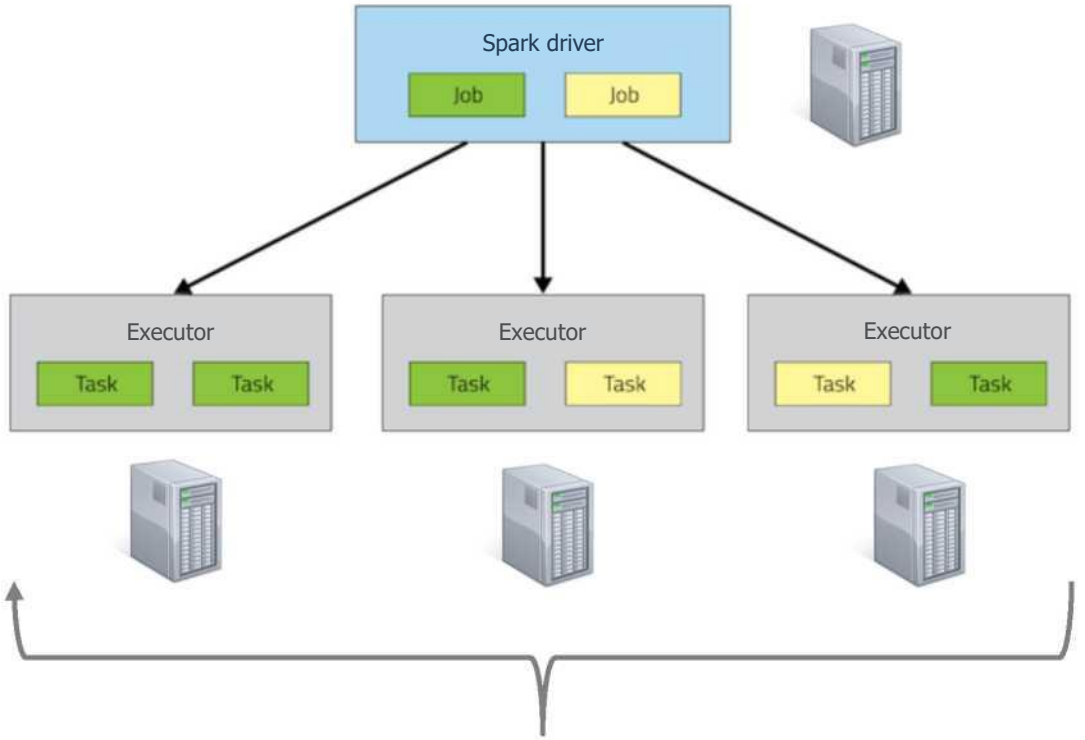
Spark is an in-memory framework

Supports general workloads as well as streaming, interactive queries and machine learning providing performance gains



Why Spark?

Spark DataFrames
In-memory distribution



HDFS
On-disk distribution

RDBMS

Files

Key Reasons for the Interest in Spark

Performant → In-memory architecture greatly reduces disk I/O

Productive →

- **Concise and expressive syntax,** especially compared to prior approaches
- **Single programming model** across a range of use cases and steps in data lifecycle
- **Integrated with common programming languages** - Java, Python, Scala, R
- **New tools** continually reduce skill barrier for access (e.g. SQL for analysts)

Leverages existing investments * Works well within **existing Hadoop ecosystem**

Improves with age → **Large and growing community** of contributors continuously improve full analytics stack and extend capabilities

What Spark Is Not!

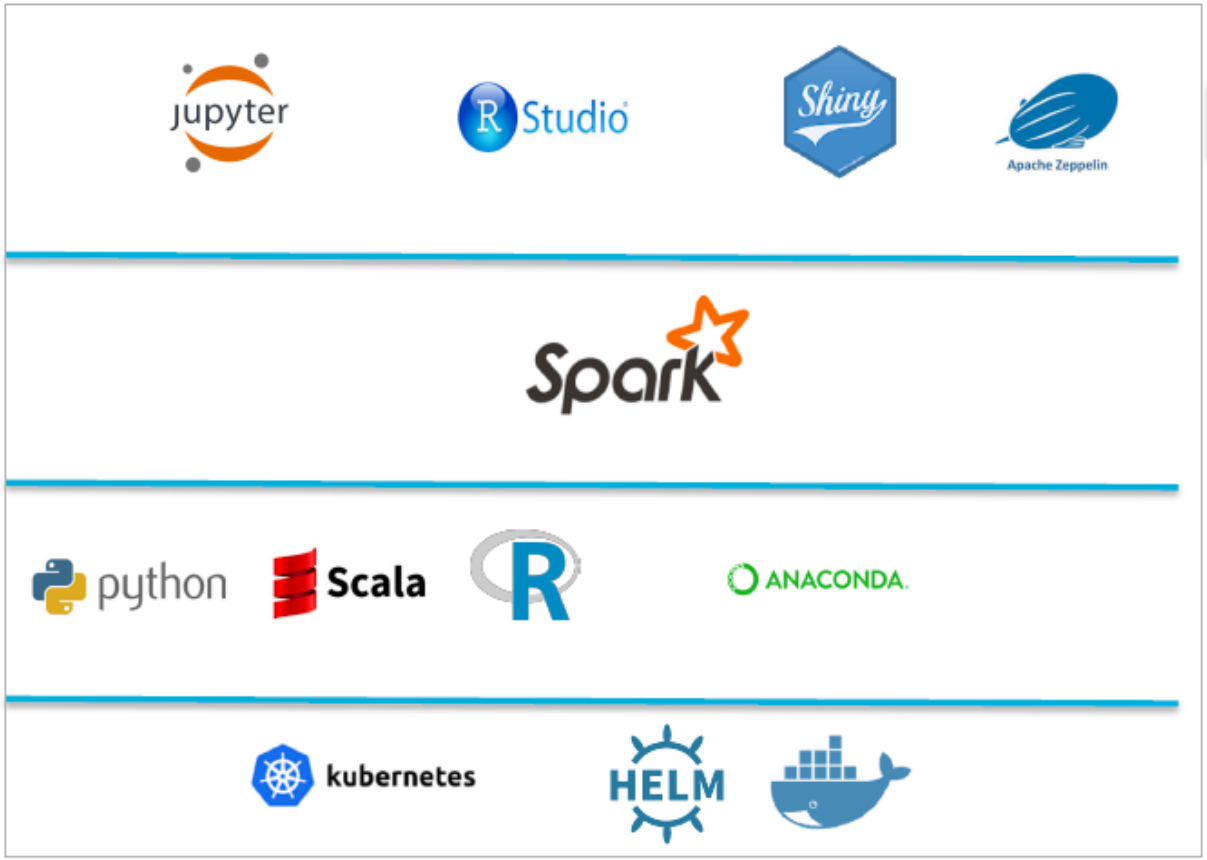
- **Not *only* for Hadoop - Spark can work within Hadoop (especially with HDFS), but Spark is a standalone system**

- **Not a data store - Spark attaches to other data stores but does not provide its own**

- **Not *only* for machine learning - Spark includes machine learning and does it very well, but it can handle much broader tasks equally well**
  - Distributed File System
  - Data preparation
  - Streaming processing
  - SQL engine
  - Graph Engine
  - Distributed R

- **Not a language!!!**

Data Science and Big Data technologies convergence



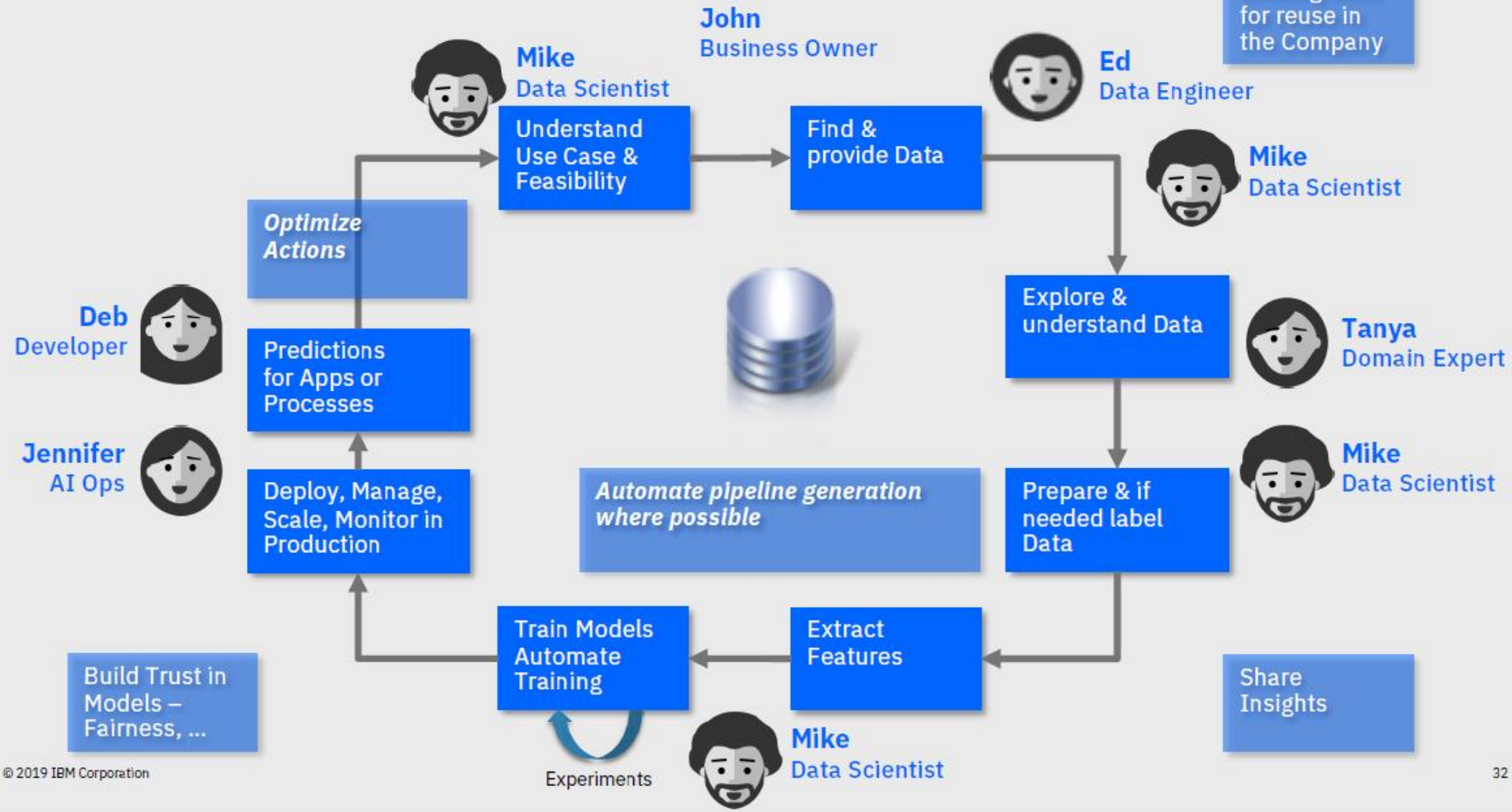
IDEs & Visual tooling
Projects: collaboration & versioning

Big Data processing

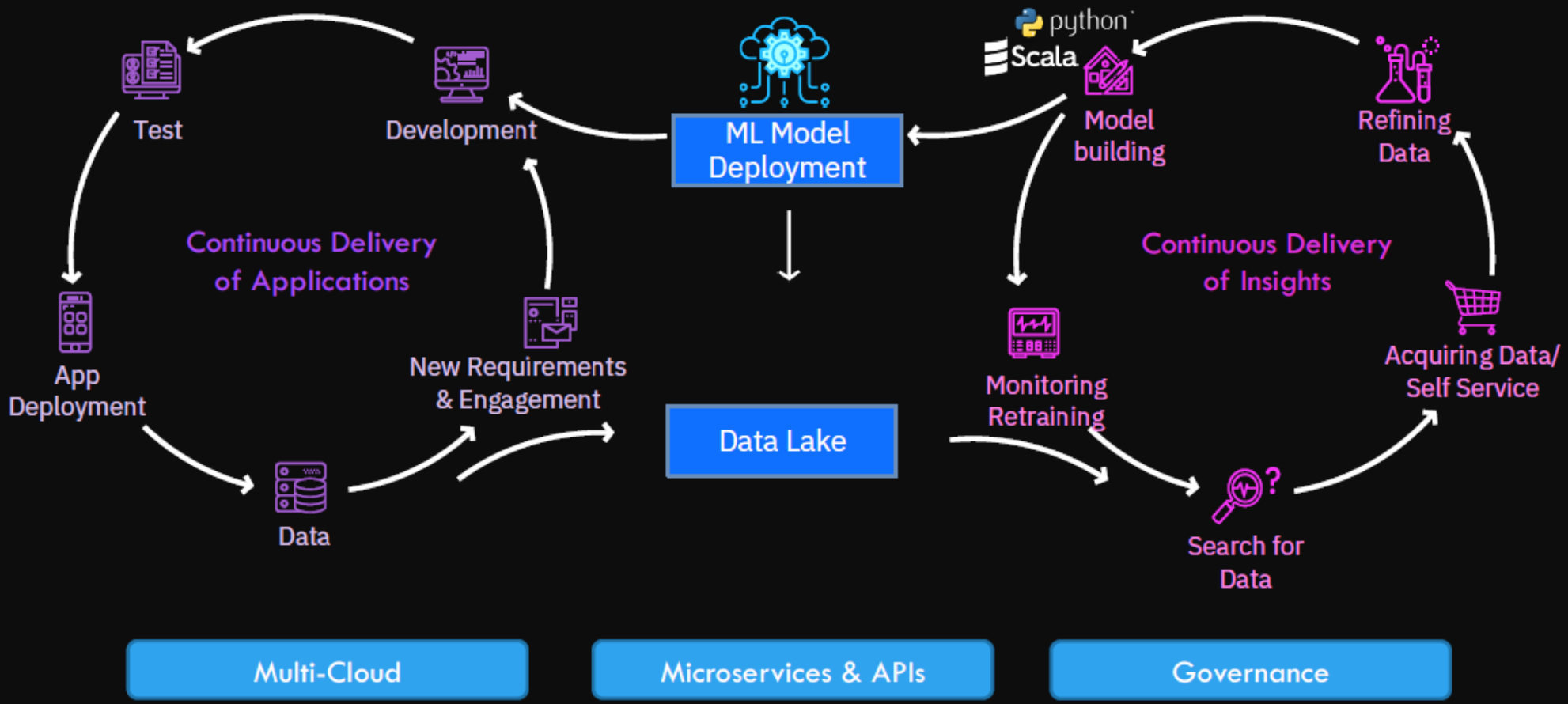
Languages

Cloud-Native Platform & Eco-System
Scale-out & load-balancing

Data Science "A Day in the Life"



Delivering Insight Applications



Organizations lack the resources to manage and monitor AI in production.

60%

of companies see **regulatory constraints** as a barrier to implementing AI.

– *IBM IBV AI 2018*

63%

cite availability of **technical skills** as a challenge to implementation.

- IBM IBV AI 2018

Without expensive Data Science resources handholding multiple AI models in a production application:

1. No way to **validate** if AI models are **compliant with regulations** and will achieve expected business outcomes before deploying
2. Difficult to **track and measure** indicators of business success in production
3. Resource intensive and unreliable processes for **ongoing business monitoring and compliance**
4. Impossible for business users to **feedback** subtle domain knowledge into model lifecycle

Data & AI Work Together

Data makes AI feasible & necessary.

AI makes Data valuable.

The Data Challenge

- *Which data should we use to train ML models?*
- *Do we have the right data to train an ML model? The data that are easy to get may not be the most informative.*
- *Are they available and easy to access ? What about the volumes ?*
- *Are they structured? Unstructured ?*
- *Do we have labeled data and are those labels valid?*
- *How do we make sure that data can be used without violating any compliance and regulation requirements?*

The AI challenge

- *Understand the business use case*
- *Choosing the right Machine Learning Algorithm*
  - *Supervised Learning*
  - *Unsupervised Learning*
- *Industrialize it in Applications*

35

How Smart Is AI ? Can I Use It For My Problem Domain ?

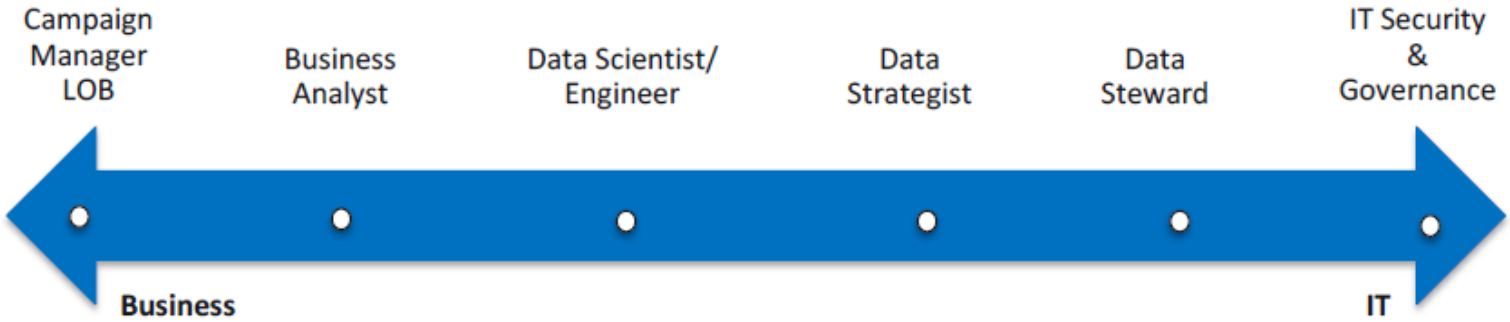
IT'S ONLY AS GOOD AS THE DATA

Data Lake Architecture

Hybrid, Public, Private Cloud

85% of enterprises will commit to multi-Cloud architectures

Who benefits from a Data Lake?



- LOB users, business analysts and data scientists can easily find the information they need without extensive IT involvement
- Data strategists and data stewards can make information available to users in an organized and well-governed manner
- IT security and governance teams can be assured that information is governed according to well-defined organizational and regulatory policies

Quelques principes d'architecture

- **Governance :**
  - ♦♦♦**Share the same understanding on data : Data Catalog**

- **User Experience :**
  - ♦**Collaborative : Share the assets (data, modeles, dashboards, flows, ...)**
  - ♦ **Managed Services**

- **Deploy Anywhere:**
  - ♦ **Cloud native architecture**
  - ♦ **Cloud agnostic & multi-cloud - any vendor cloud or data center**
  - ♦ **Coherent, efficient, and scalable data & analytics services**

Skills & Tips

5 tips for machine learning success

- **Contribute to Open Source**
  - Use Github and learn from some tutorials.
  - Scan Apache's open source projects and choose one to explore. Play around with Jupyter notebooks, and reach out to other contributors. Then, when contributing to open source becomes more important in your own dev community, you'll be ready.

- **Participate to Challenges and stay active in the community**
  - Kaggle, Hackathon

- **Keep learning Python**
  - If you're already using Python for machine learning, you might be frustrated by the lack of support, especially for large data sets.
  - But many data scientists love Python, so more open source projects are improving their support. You'll be able to do more as time goes on. Hang in there.

Pay attention to data gravity
- It almost always makes sense to move computation to the data rather than moving data to the computation. Same for tools. Bring tools to the data instead of bringing data to the tools.
- Simple advice, but it's hard to follow when you need to set up a data management architecture quickly or cheaply. If the people managing your data have security or data integrity concerns, spend time making your case. In the long term, it makes a huge difference, especially with machine learning, where computation is getting more and more intense.

Know your data:
- Maybe this seems obvious, but check the data quality; You'll probably find and correct the obvious ones, but don't get overconfident.
- Data quality issues are everywhere.

Get started with Cloud
- Sooner or later, you'll probably need to work with the cloud, so plan in advance.
- Watson Studio is a good example to do it

Do not Stop Educating

Data Science Central THE ONLINE RESOURCE FOR BIG DATA PRACTITIONERS

Meetup

KDnuggets™ https://www.kdnuggets.com/

https://www.reddit.eom/r/MachineLearning/

IDS© Elm UBI

became V:.|| COGNITIVE CLASS*

https://cognitiveclass.ai/

Data Science and Cognitive Computing Courses

Build Data Science and Cognitive Computing skills for free today.

Free Courses Sign Up

Free courses, free materials: ***MOOCs on Big Data, Data Science; Spark, Hadoop, R, SQL, NoSQL***

Study

Cloud-based sandbox for exercises

IBM Careers France Etudiants Experiments

Rejoignez notre reseau de talents

Suivre IBM: jn # O ^ @ D

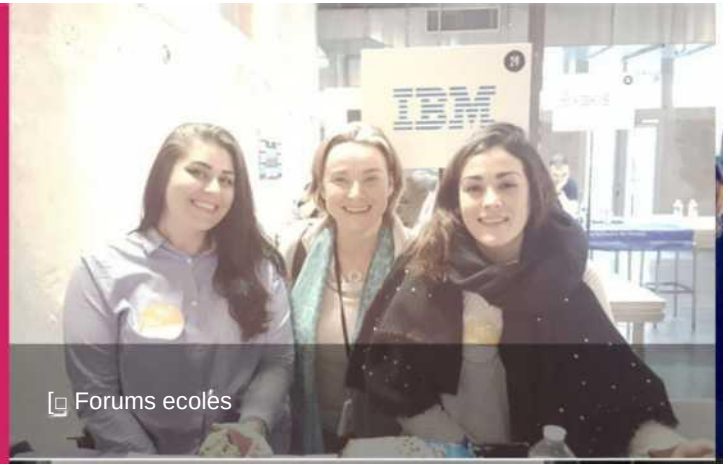
Rechercher parmi les offres juniors

Rechercher parmi les offres experiments

OldgCb

[□ En savoir plus

\U Recherchez les offres de stages



Passionne(e)s par la #TransfoDigitale ? Visionnez le replay du chat video JobTeaser anime par des

https://www.ibm.com/fr-fr/emplovment/entry_level.html