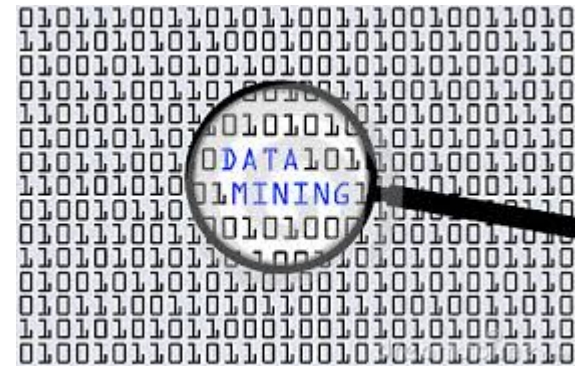
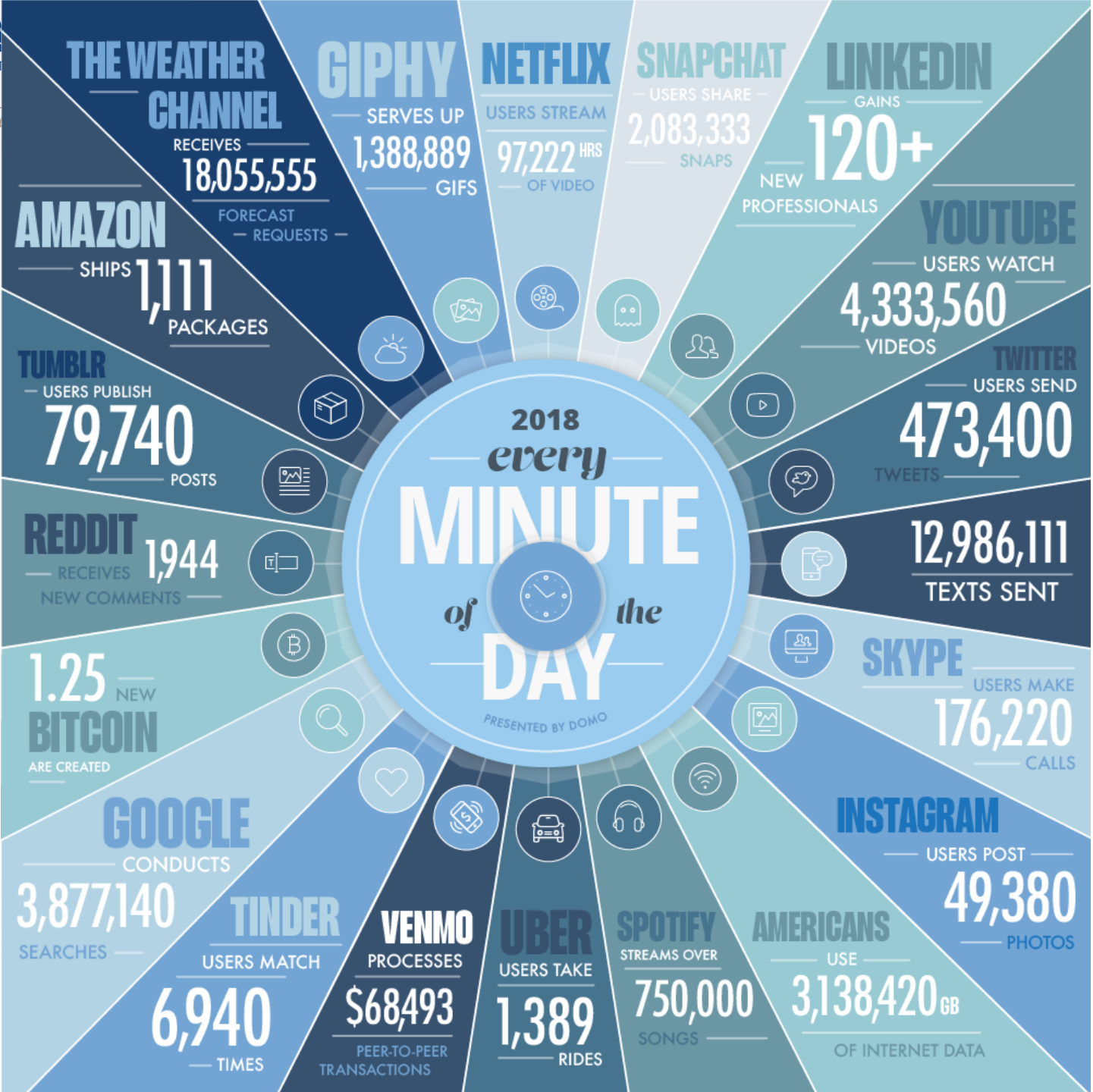
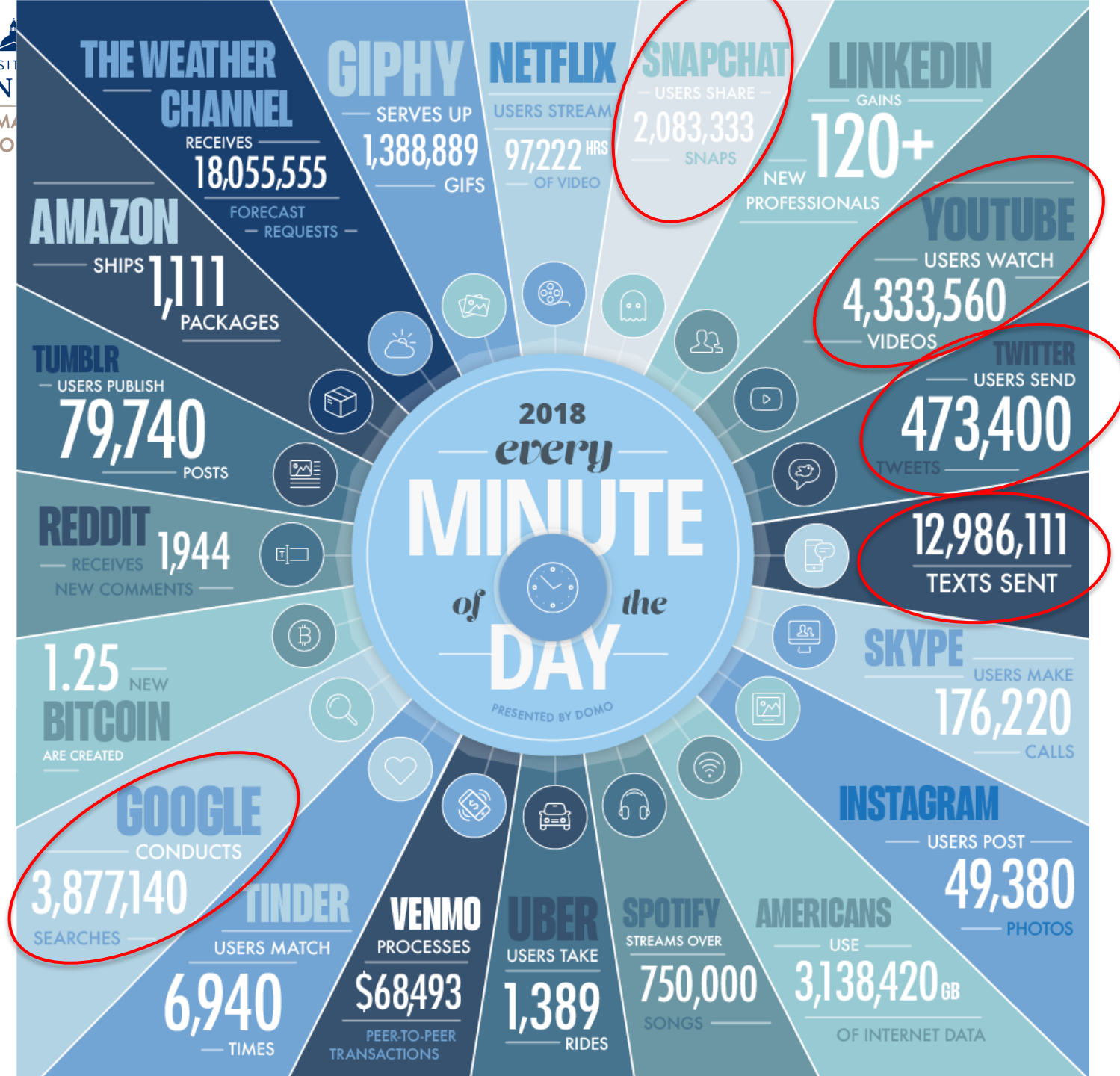


Introduction à la Data Science



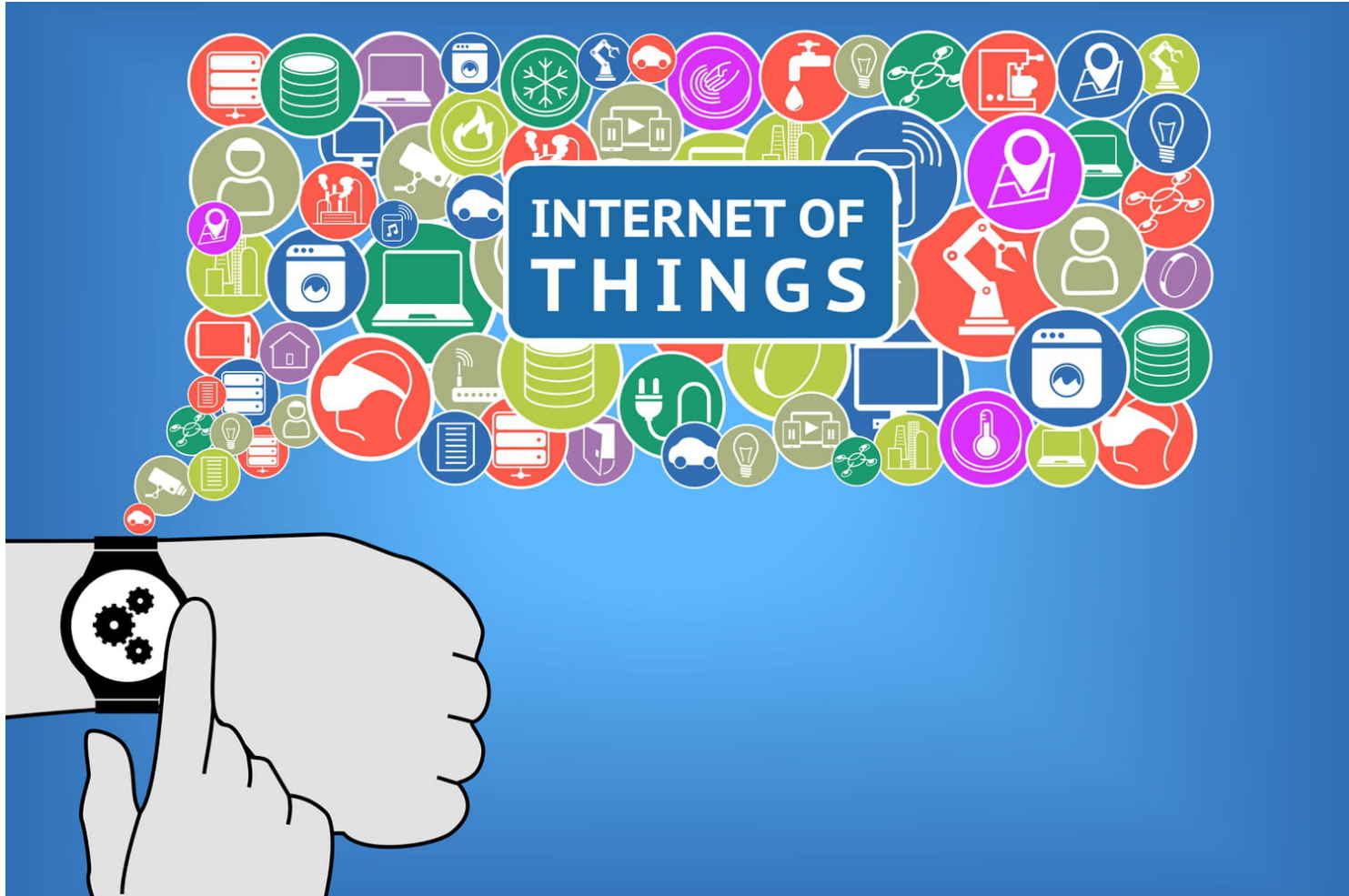
Manuele Kirsch Pinheiro
Luis Angelo Steffenel
Bénédicte Le Grand





Pourquoi ?

Des milliards d'objets connectés



Internet of Things

Objets capables d'envoyer des données automatiquement :

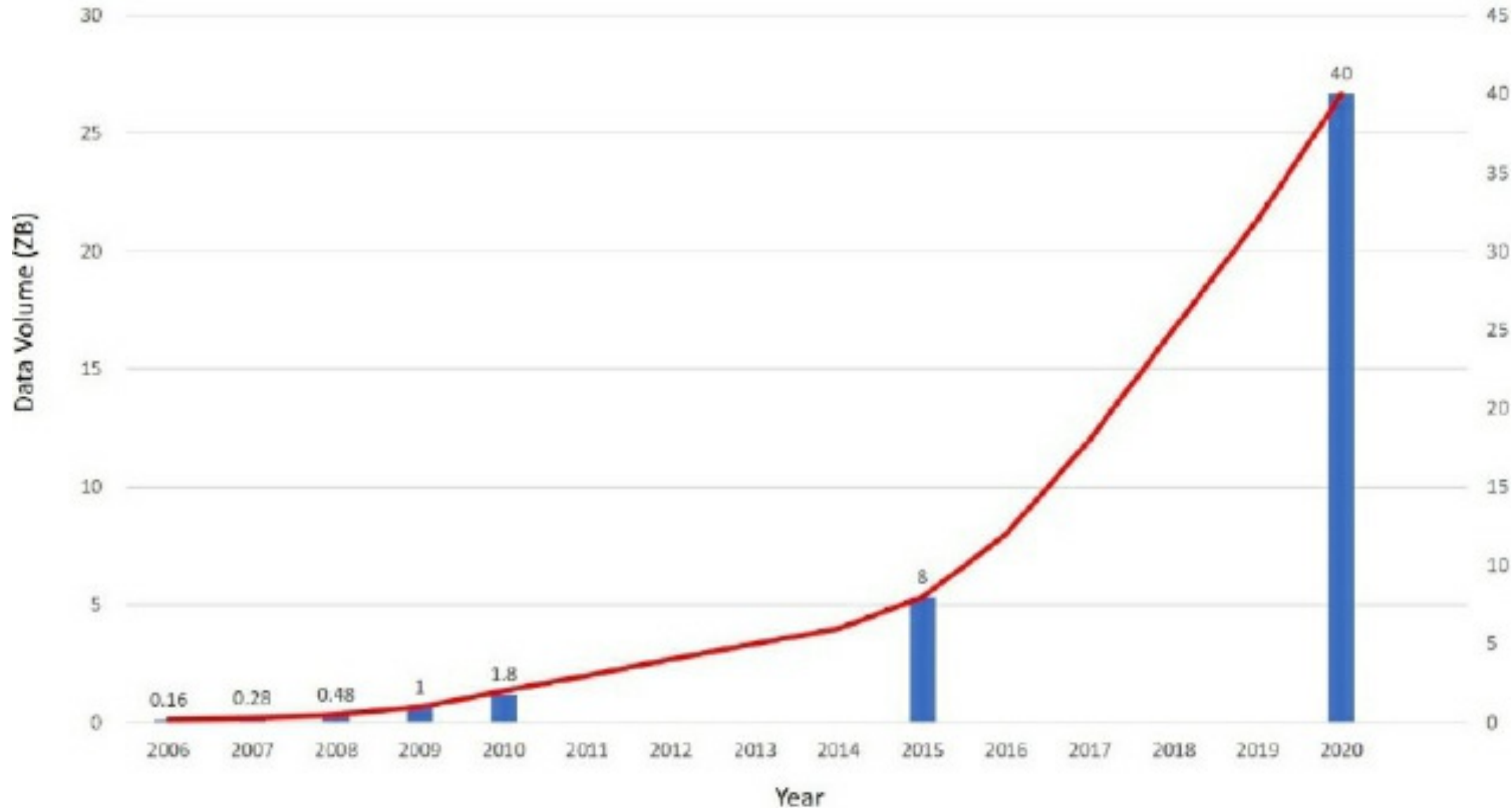
- Voiture avec équipement pour télé-péage,
- Moniteurs de places disponibles dans un parking,
- Moniteur cardiaque implanté chez un humain,
- contrôleurs de la qualité de l'eau,
- Compteur intelligent qui rapporte la consommation d'énergie,
- Détecteur de radiations,
- Traceurs d'objets dans un entrepôt,
- Applis mobiles pour tracer les mouvements et la localisation
- Thermostats intelligents qui ajustent la température des pièces en fonction des prévisions météo et de l'activité dans la maison,
- Équipements domotiques intelligents.

Selon le site [statista.com](https://www.statista.com) :

- déjà plus de 23 milliards d'équipements IoT aujourd'hui
- prévision : 75 milliards en 2025.

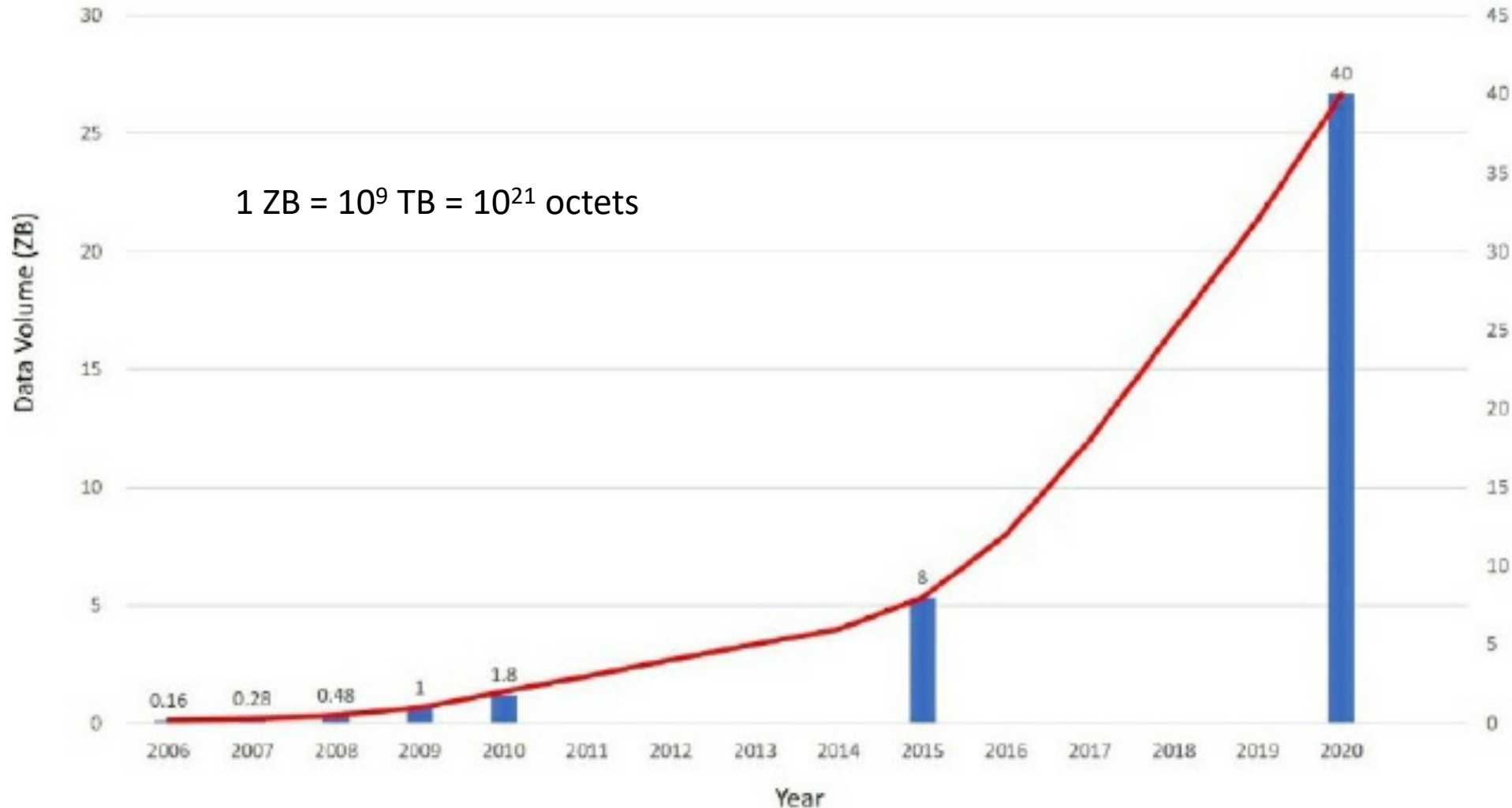
Croissance du volume de données

Global growth trend of data volume, 2006-2020



Croissance ~~du volume de données~~ de la valeur des données ???

Global growth trend of data volume, 2006-2020



- **Volume** des données générées
- **Vitesse** de production des données
- **Variété** des données
 - Structurées
 - Non structurées
- **Variabilité, Véracité, Validité, Vulnérabilité, Volatilité, Visualisation, Valeur**
- Besoin d'avoir des « insights » sur les données
 - On ne sait pas forcément précisément ce que l'on cherche
 - On ne connaît pas forcément les éléments importants dans les données



BIG DATA & AI LANDSCAPE 2018

INFRASTRUCTURE

HADOOP ON-PREMISE
cloudera Hortonworks
MAPR Pivotal
IBM InfoSphere
bluedata jethro

HADOOP IN THE CLOUD
aws Microsoft Azure
Google Cloud
TREASURE DATA
CAZENA CenturyLink

STREAMING / IN-MEMORY
aws databricks strim
confluent GridGain
dataArtisans hazelcast
TERRACOTTA Iox
WallarooLABS

NOSQL DATABASES
Google Cloud AWS
ORACLE Microsoft Azure
mongoDB MarkLogic
HEDSPARK DISTRATX
AsangoDB Couchbase
redislabs SCYLLA

NEWSQL DATABASES
SAP Clustrix
Cockroach LABS
MEMSQL
citusdata
paradigm4

GRAPH DBs
neo4j Amazon Neptune
IBM ORACLE
Cytoscape
Alation

MPP DBs
TERADATA
IBM Data Warehouse Systems
Celonis
Exasol
dremio

CLOUD EDW
aws
Google Cloud
Microsoft Azure
Pivotal
snowflake

DATA TRANSFORMATION
talend pentaho
alteryx TRIFACTA
tomr PAXATA
StreamSets UNIFI

DATA INTEGRATION
SAP Data Services
Informatica
enigma
alooma
Stitch
InfoWorks

DATA GOVERNANCE
IBM
colibra
Alation

MGMT / MONITORING
aws New Relic
rubrik
splunk
pagerduty

STORAGE
aws
Microsoft Azure
PURE STORAGE
ALACRITY
Qumulo
COHERITY

CLUSTER SVCS
aws
Microsoft Azure
docker
K8EN IO
rainforest
CASK

APP DEV
lightbowl
K8EN IO
rainforest
CASK

CROWD-SOURCING
amazon mechanicalturk
upwork
figure eight
scale
HIVE

HARDWARE
Google TPU
ARM
MYTHIC
Movidius
WAVE

GPU DBs
kinetica
IBM
INFORMatica
BLAZENDB
bryllyt PG Storm

CROSS-INFRASTRUCTURE/ANALYTICS

aws Google Cloud Microsoft IBM SAP Oracle NetApp Synapse MAAP cloudera

ANALYTICS

DATA ANALYST PLATFORMS
Microsoft pentaho
guavus AYASDI
ATTIVIO Datameer Quid incorta
interana ClearStory Origami
ENDOR MODE

DATA SCIENCE PLATFORMS
IBM KNIME dataiku
DOMINO rapidminer
CONTINUUM
ALGORITHMIA
DATAWATCH SAS

BI PLATFORMS
Microsoft AWS
looker
ATSCALE
MicroStrategy

VISUALIZATION
tableau
Google Cloud
Qlik
ZEPL
CHARTIO

MACHINE LEARNING
aws
Google Cloud
DataRobot
gamalon
ELEMENT
VIZIERE
bonsai

COMPUTER VISION
Microsoft Azure
Amazon Rekognition
clarifai
EVER AI
deepomatic

HORIZONTAL AI
IBM Watson Cortana
senient
Affective
Numenta
NOFIGURES
OSARO

SPEECH & NLP
Google Cloud
twilio
Soundhound Inc.
snips

SEARCH
ORACLE
elasticsearch
coveo
ATTIVIO
swiftype
algolia
MAANA
omnius

LOG ANALYTICS
splunk
sumologic
LOGGLY
swiftype
algolia
MAANA
omnius

SOCIAL ANALYTICS
Hootsuite
NETBASE
synthesio
simplesearch
bitly
predata
SimilarWeb

WEB / MOBILE / COMMERCE ANALYTICS
Google Analytics
mixpanel
sumall
RESCI
SIGOPT
granify

APPLICATIONS - ENTERPRISE

SALES
Salesforce
INSIDESALES.COM
conversica
clari
aviso
tact.ai
fusemachines
TROOPS

MARKETING - B2B
RADIUS
EVERSTRING
HINTIGO
sense
tubator
DataFox
ENGAGIO
mip

MARKETING - B2C
Zeta
blueyonder
ACTIONIQ
SAULTHRU
BLUECORE
QUANTIFUN
mparticle
Ampero
amperity
STREANUM
Simon
Lyfika

CUSTOMER SERVICE
MEDALLIA zendesk
CLARABridge
Gainsight
NO DATA
DigitalGenius
afiniti
AUTOMATY
Frame AI
Cognia
INTERCOM
CsDesk

HUMAN CAPITAL
HireVue
entelo
hiQ
GIGSTER
JUDICATA
PREVIEW
Stella
mya

LEGAL
Ravel
EVERSTRING
JUDICATA
PREVIEW
Stella
mya

FINANCE
Anaplan
ZUJOFO
SAHANA
TRADESHIFT

ENTERPRISE PRODUCTIVITY
slack
ORACLE
sumoto
claro
talia
butterai
Kasisto

BACK OFFICE AUTOMATION
UPath
bluewin
Aurion
Applan
Workfusion

SECURITY
CYCLANCE
StackPath
BARRACUDA
ANOMALY
DATAVISOR
CyberArk
SentinelOne
BlueTalon
SentinelOne
BlueTalon
SentinelOne
BlueTalon

APPLICATIONS - INDUSTRY

ADVERTISING
AppNexus
criteo
ORACLE
MOAT
thetradedesk
distillery
TAPAB
Oppler

EDUCATION
edX
OpenX
KIDaptive
KIDaptive
KIDaptive

GOVERNMENT
OPENGOV
GRIDSMART
Passport
SmartProcure
STREETLIGHT DATA

REAL ESTATE
REDFIN
Opendoor
CREDIC
VT S
CREDIC
VT S

FINANCE - INVESTING
KENSIC
Quantopian
KENSIC
Quantopian

FINANCE - LENDING
ondeck
affirm
KREDIT
AVANT
INSIGHT
100Cred
MoneyLion
aire
agrib

INSURANCE
Lemonade
CYNCE
TECHNOLAB

HEALTHCARE
Flatiron
Clever
Ginger
Glow
3D Med
zebra
PIMA
OVI
TEMPUS
patientCloud
AI Cure
Qventus
prognos
enatic
Imag

LIFE SCIENCES
Benevolent
Verily
WuXi
ZEPHYRUS
ZEPHYRUS
ZEPHYRUS

TRANSPORTATION
UBER
TESLA
CLEARPATH
drive.ai
nautix
nautix
nautix

AGRICULTURE
FARMERS
Granular
John Deere
BLUE RIVER
Blue River
Blue River

COMMERCE
STITCH FIX
D&G
F&G
F&G

INDUSTRIAL
Siemens
PREDIX
UPTAKE
TACHYUS
KODAK

OPEN SOURCE

FRAMEWORK
TensorFlow
PyTorch
Keras
Caffe
MXNet
Theano
PySpark
Flink
Kafka
Druid
Storm

QUERY / DATA FLOW
Spark SQL
Presto
SLAM DATA
Apache Drill
Apache Airflow

DATA ACCESS
nifi
mongoDB
couchDB
CouchDB
CouchDB

COORDINATION
talend
Apache Zookeeper
Apache Airflow

STREAMING
Spark
Flink
Kafka
Druid
Storm

STAT TOOLS
R
Python
Julia

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow
theano
Caffe
MXNet
Theano
PySpark
Flink
Kafka
Druid
Storm

SEARCH
elasticsearch
Solr

LOGGING & MONITORING
kibana
logstash
Prometheus

VISUALIZATION
Tableau
Rodeo

COLLABORATION
Slack
Jupyter

SECURITY
Apache Ranger
KNOX
Sentry

DATA SOURCES & APIs

HEALTH
Apple
VALIDIC
practicefusion
fitbit
GARMIN
kinsa

IOT
GE Digital
UPTAKE
helium
samsara

FINANCIAL & ECONOMIC DATA
Bloomberg
THOMSON REUTERS
DOW JONES
BSP CAPITAL IQ
CBRIGHT
xignite
Quandl
PREMIERE
Estimize
eSSENTIAL
StockTwits
FLARE
earnest

AIR / SPACE / SEA
Orbital Insights
planet
Airware
AIRBOTICS
INDUSTRIAL
WINDWARD
telemetry
DroneDeploy

PEOPLE / ENTITIES
axiom
experian
Epsilon
InsideView
Crimson Hexagon
BASIS
SAFEGRAPH

LOCATION INTELLIGENCE
FOUR SQUARE
sense360
placeIQ
esri
Mapillary
cubiq
A Badar

OTHER
qualtrics
DATA GOV
enigma
CRUX

DATA RESOURCES

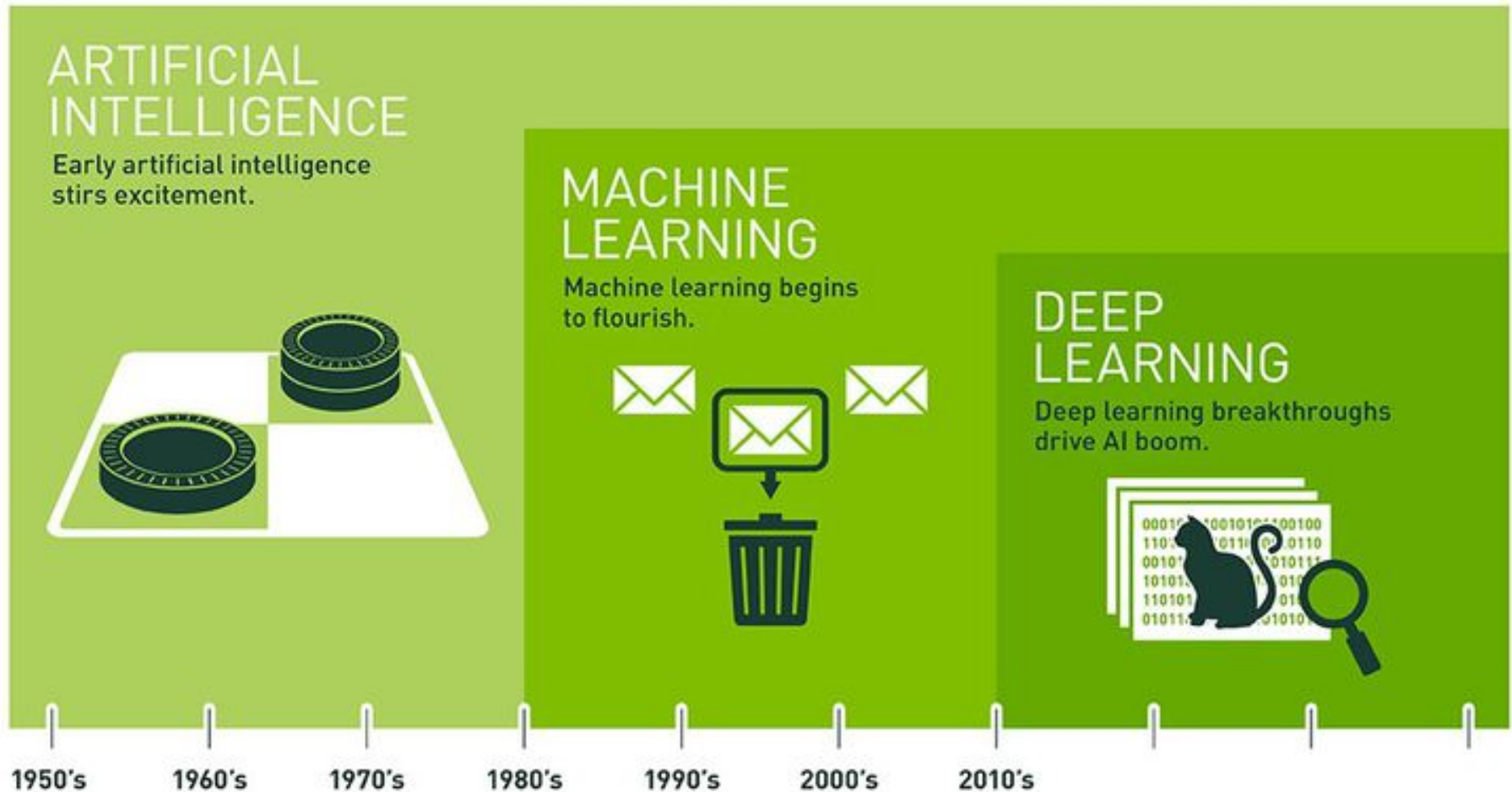
DATA SERVICES
Palantir
QIQA
fractal
kaggle
DataKind
mapbox

INCUBATORS & SCHOOLS
PLURALSIGHT
DataCamp
DataElite
The Data Incubator

RESEARCH
facebook research
MIRI
VECTOR INSTITUTE
ALLEN INSTITUTE
AIZ

- Applique :
 - des principes scientifiques, des méthodes, des algorithmes et des processus
 - pour extraire des connaissances, de l'information
 - en collectant, traitant et analysant des données structurées et non structurées
 - Sources de données structurées : par exemple SGBD (Oracle, MySQL,...)
 - Sources de données non structurées : texte, audio, vidéo, documents.

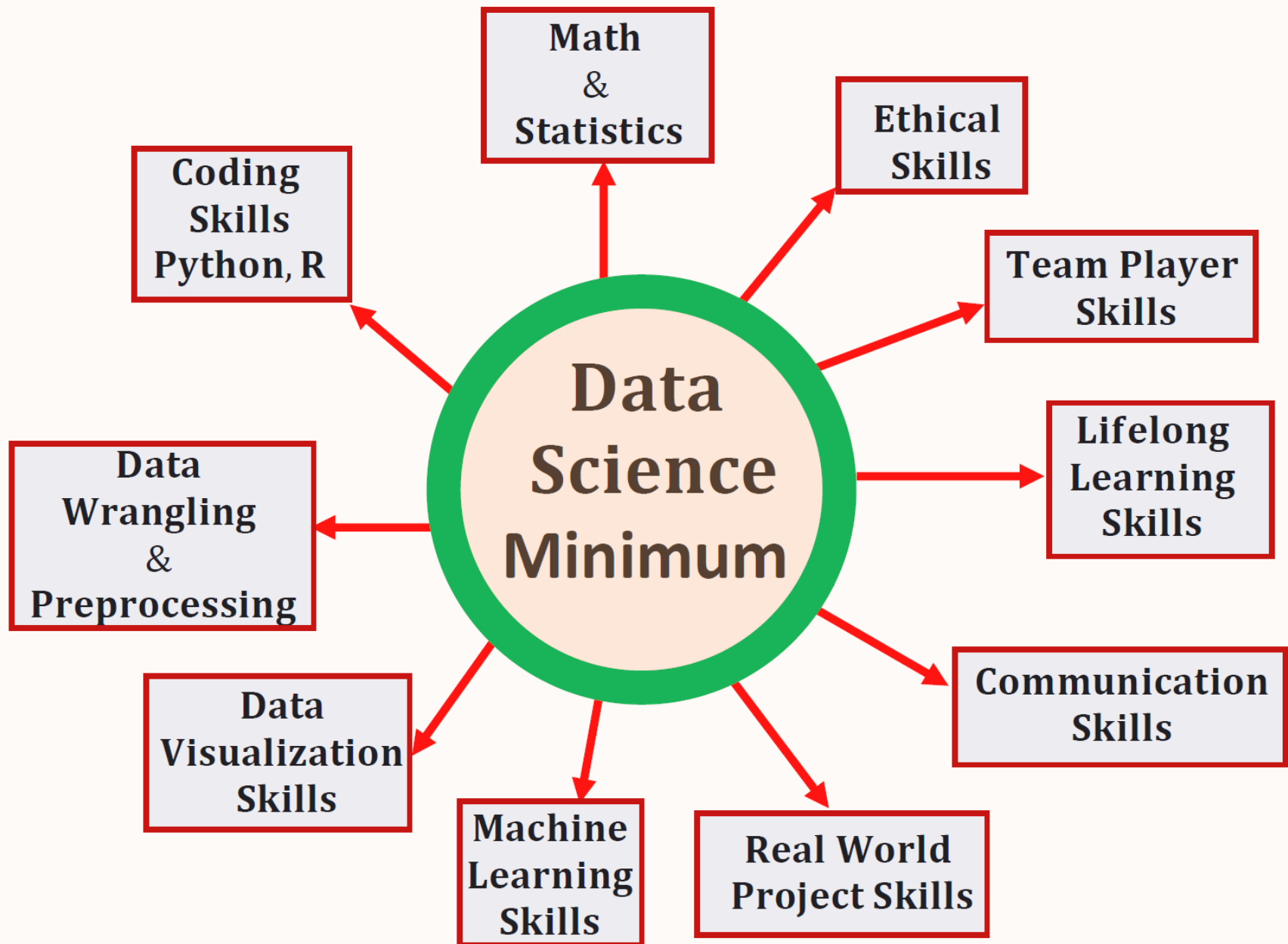
IA, Machine Learning, Deep Learning ?



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Domaines d'application

- Marketing, relation client, systèmes de recommandation
- Santé / médecine
- Secteur des banques et assurances, détection de fraudes
- Ressources humaines
- Cybersécurité
- Reconnaissance vocale/faciale, assistants personnels
- Prévisions de trafic, météo
- Analyse de réseaux sociaux, détection de tendances
- Maintenance prédictive
- ...



THE DATA SCIENCE PROCESS



Data Engineers

Data Analysts

Machine Learning Engineers

Data Scientists

THE DATA SCIENCE PROCESS



Data Engineers

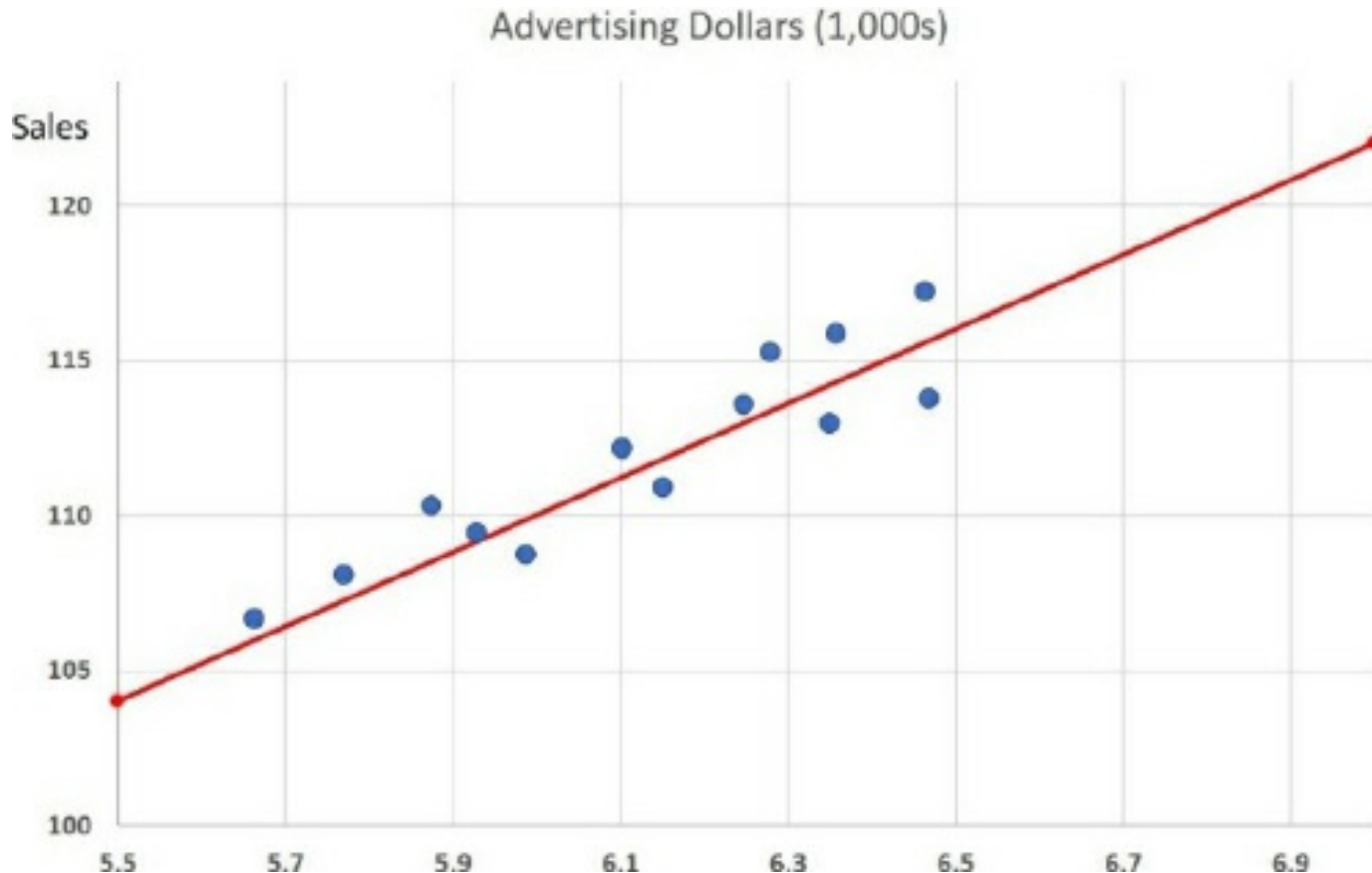
Data Analysts

Machine Learning Engineers

Data Scientists

Construction d'un modèle

- Modèle = ensemble d'hypothèses à propos des données
- Exemple de modèle : relation quasi linéaire entre le volume des ventes et la somme dépensée en publicité

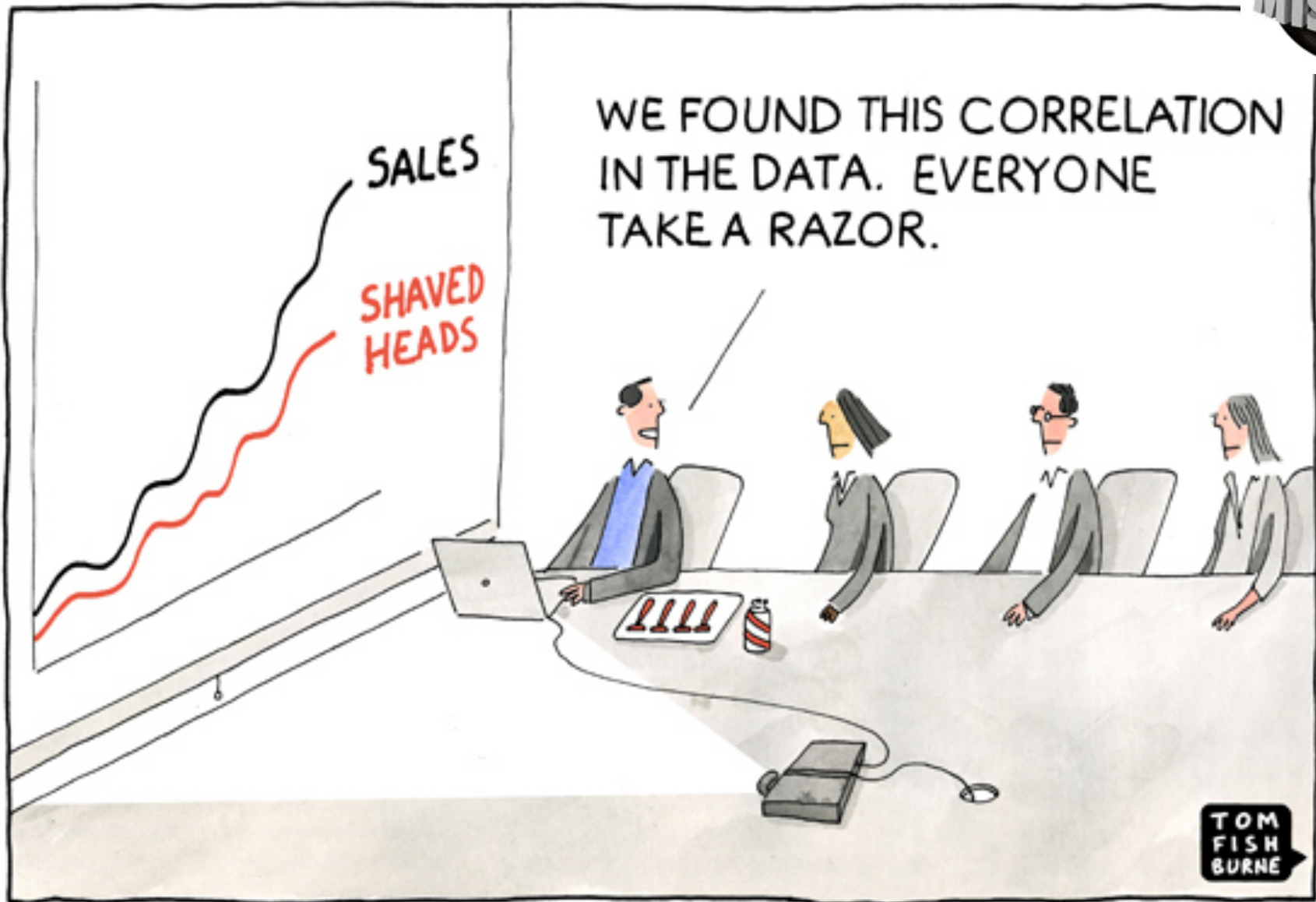


Importance d'avoir un « bon » modèle

- Pour faire des hypothèses / prédictions correctes

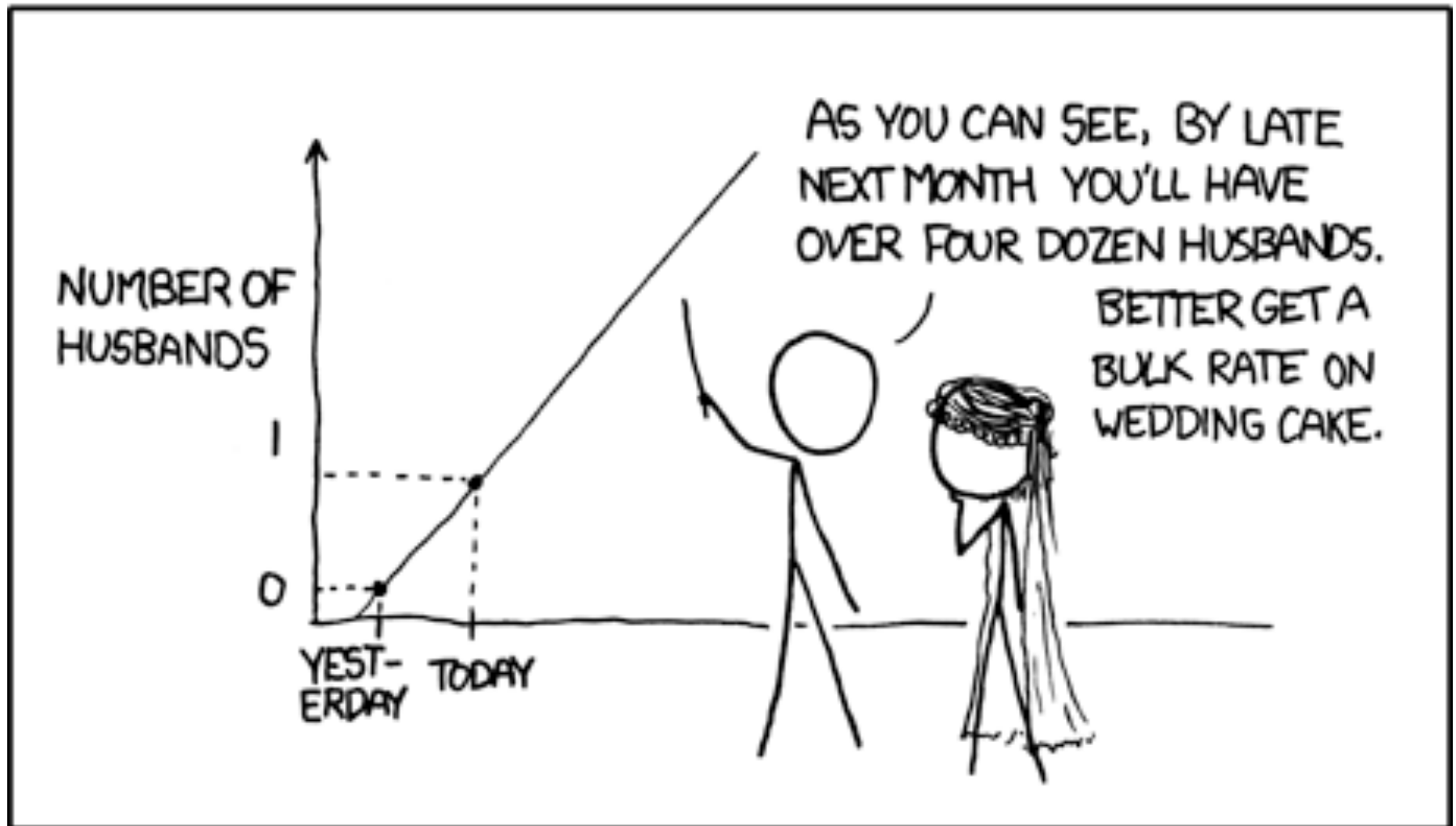
- Quel algorithme choisir ?
- Comment évaluer la qualité du modèle construit ?



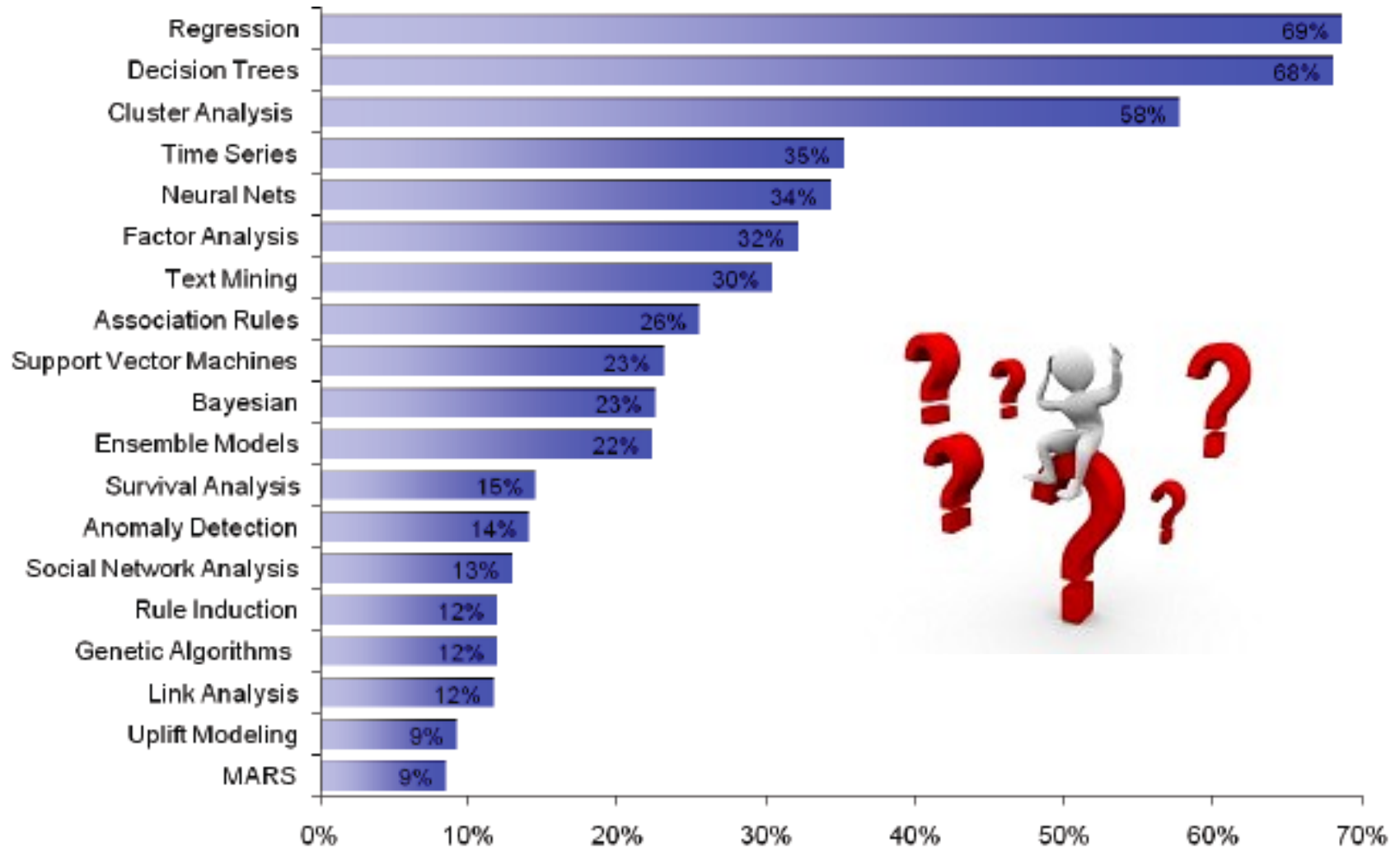




MY HOBBY: EXTRAPOLATING



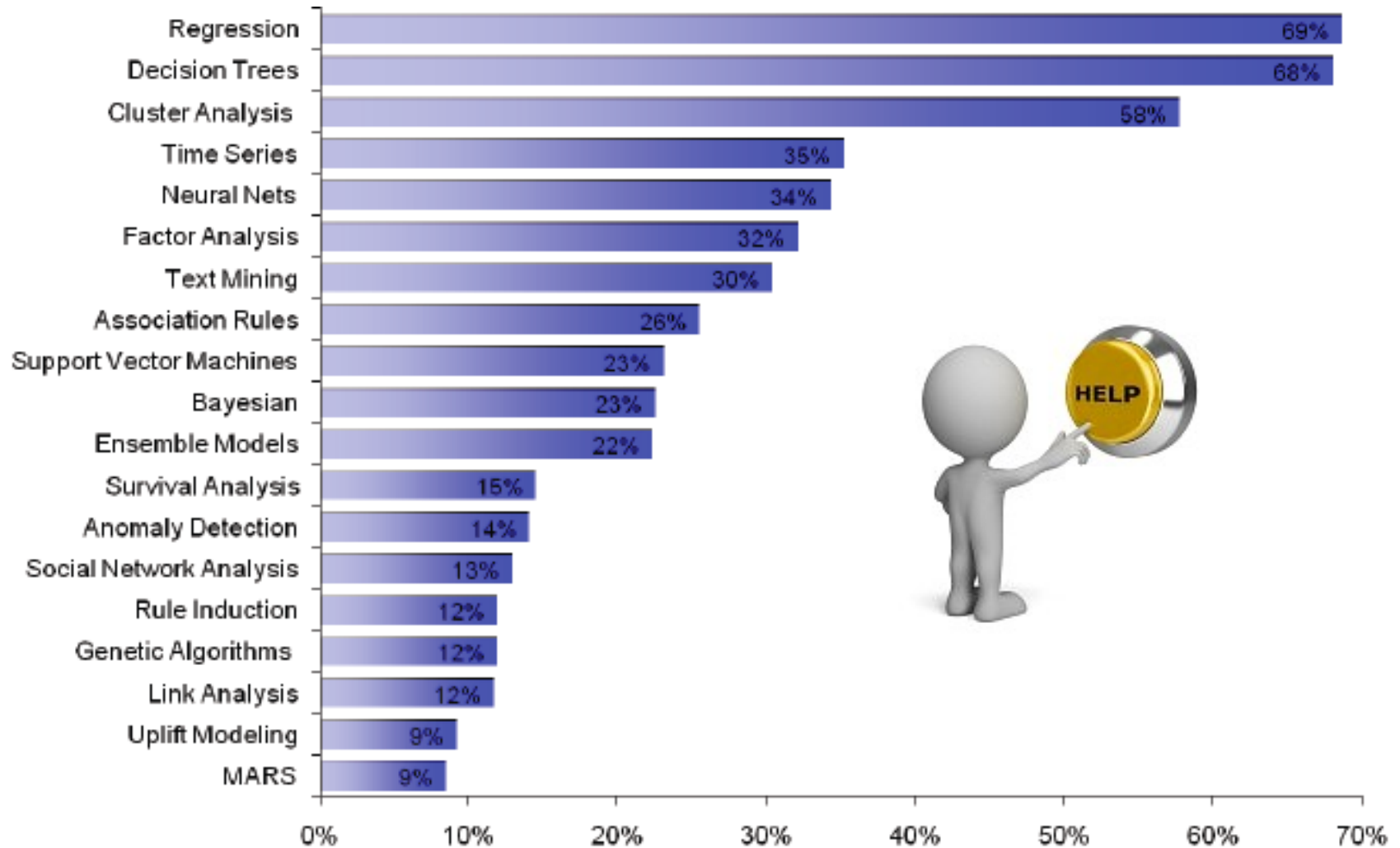
Quel algorithme utiliser ?



Question: What algorithms/analytic methods do you TYPICALLY use? (Select all that apply)

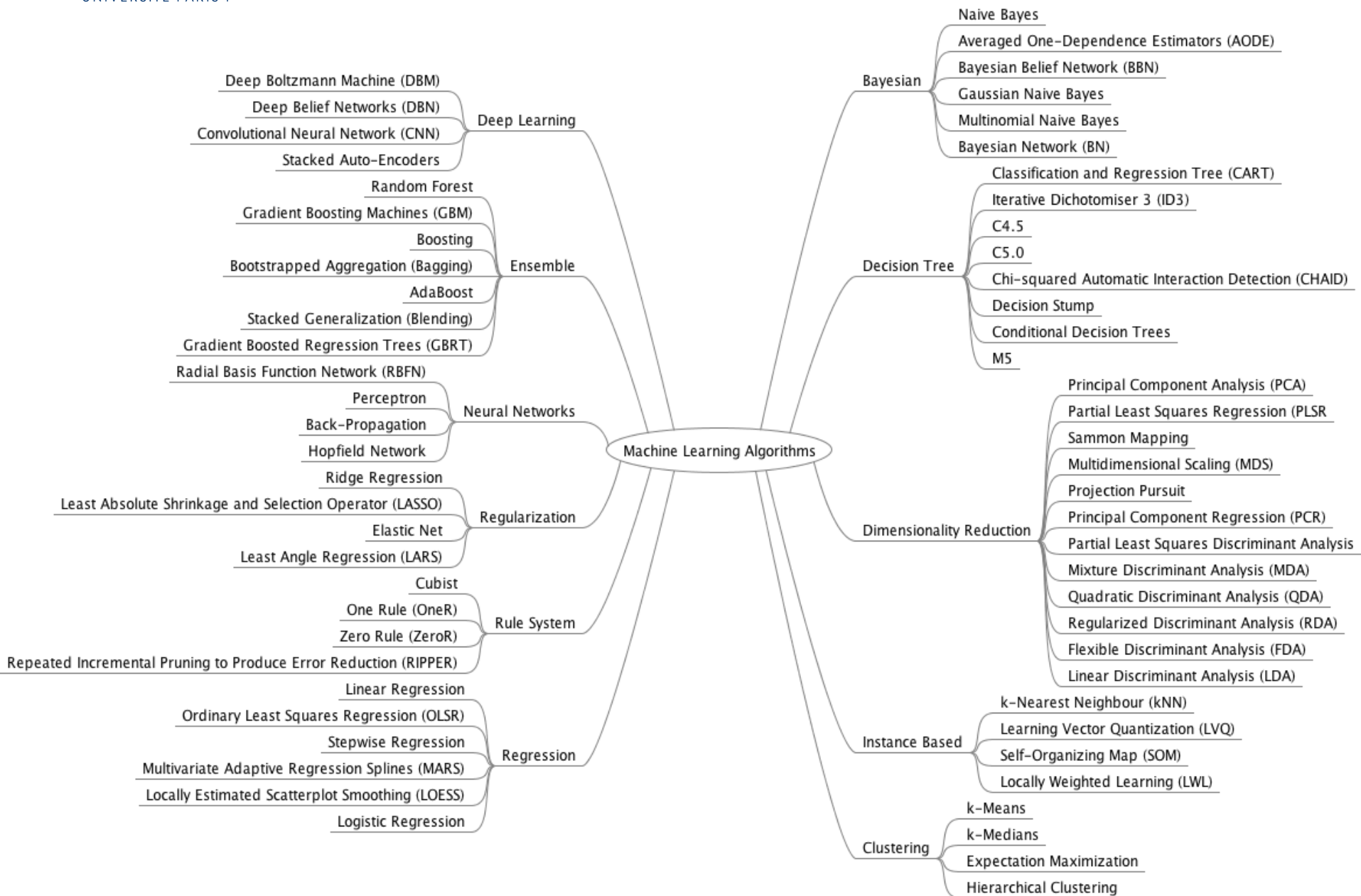
©2012 Rexer Analytics

Quel algorithme utiliser ?



Question: What algorithms/analytic methods do you TYPICALLY use? (Select all that apply)

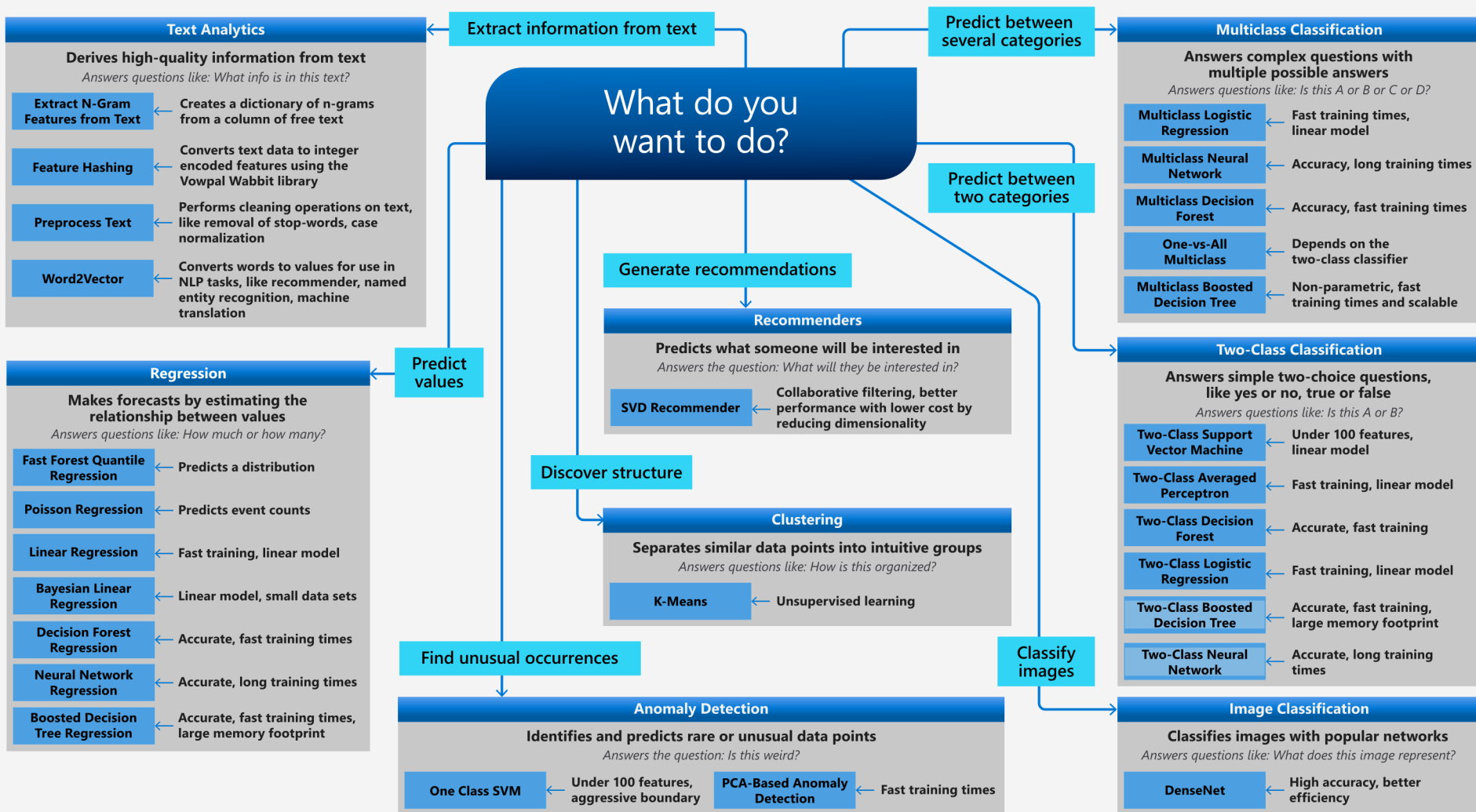
©2012 Rexer Analytics

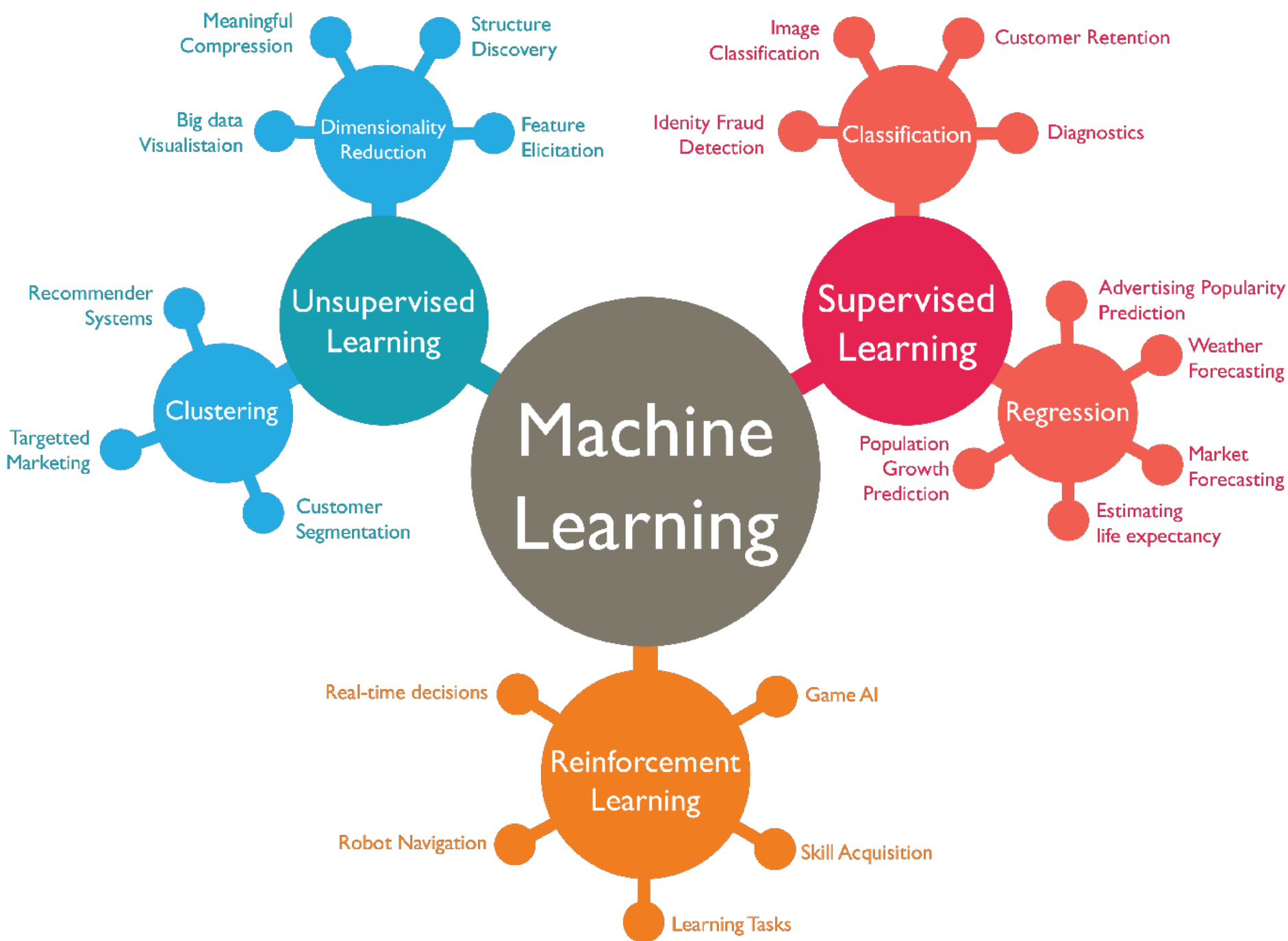




Microsoft Azure Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.







THE DATA SCIENCE PROCESS



Bienvenue dans l'aventure !!!