Introduction to statistical learning

1. Introduction

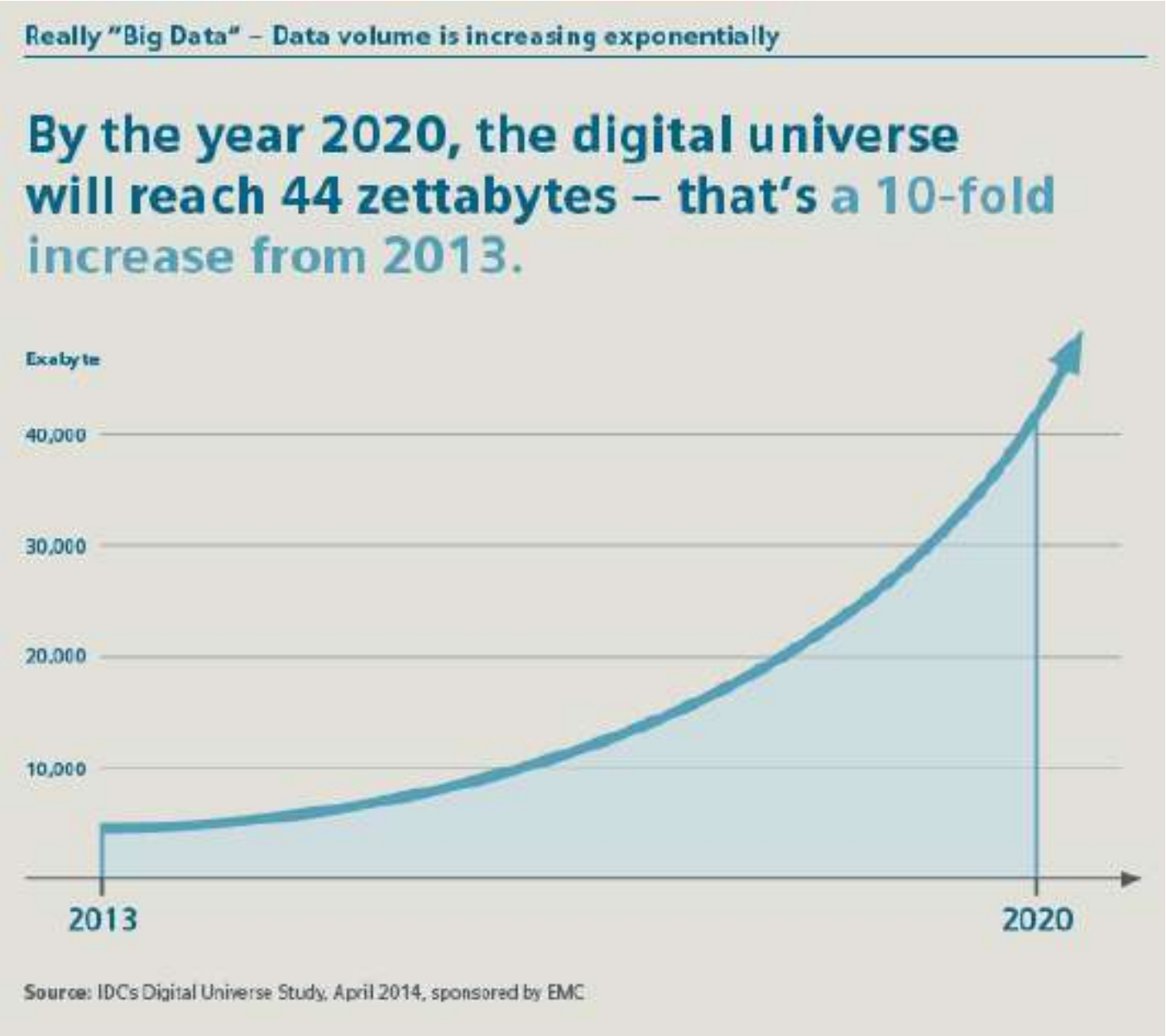
V. Lefieux

June 2018

Table of contents

Table of contents

Data everywhere



Data everywhere

- Before:
  - structured data,
  - generated by companies and organizations,
  - regular but not so frequent updates (e.g monthly).
- Now:
  - unstructured data,
  - generated by users,
  - real time data.

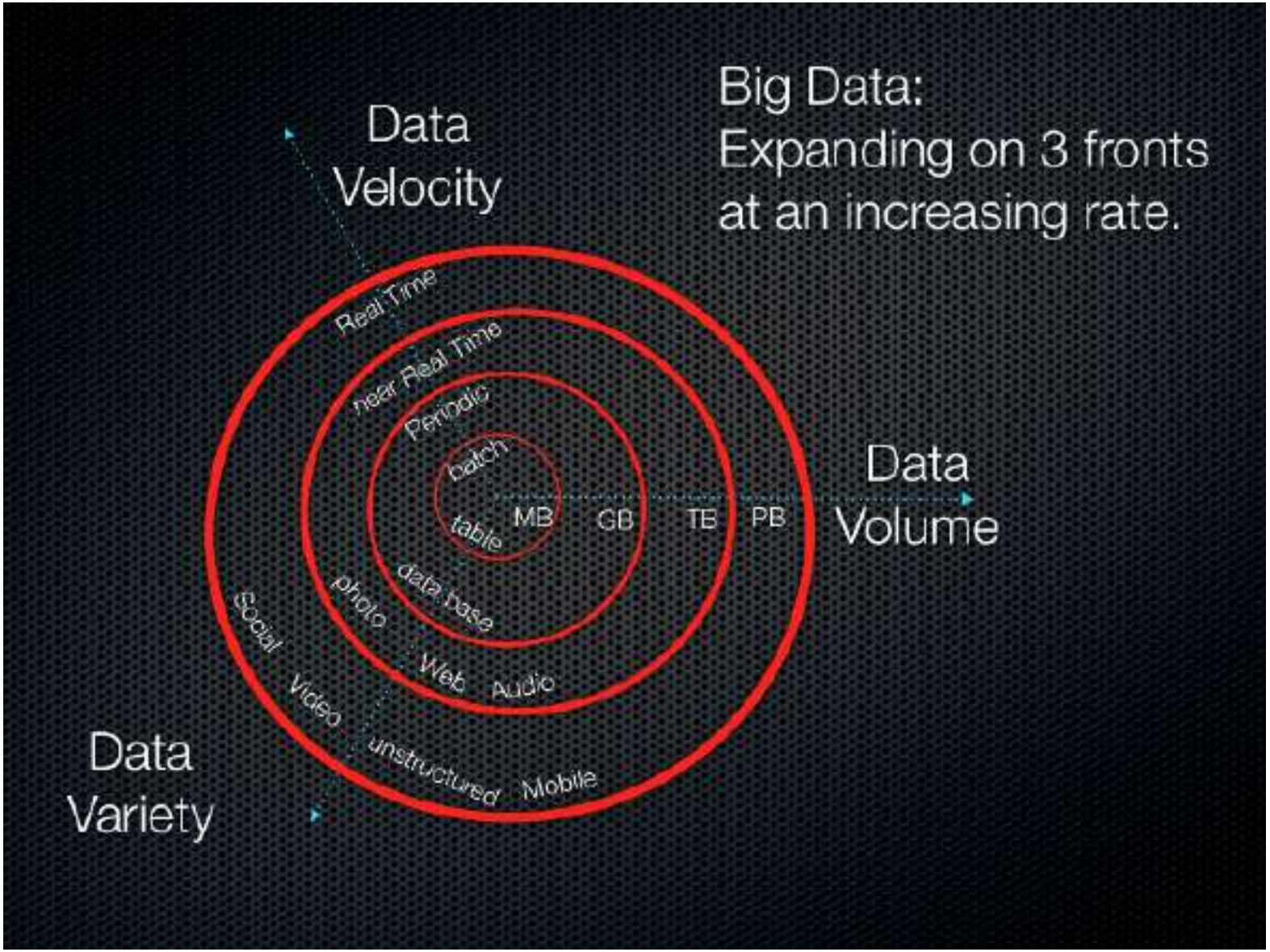
Some data generated by companies and organization

Some data generated by users

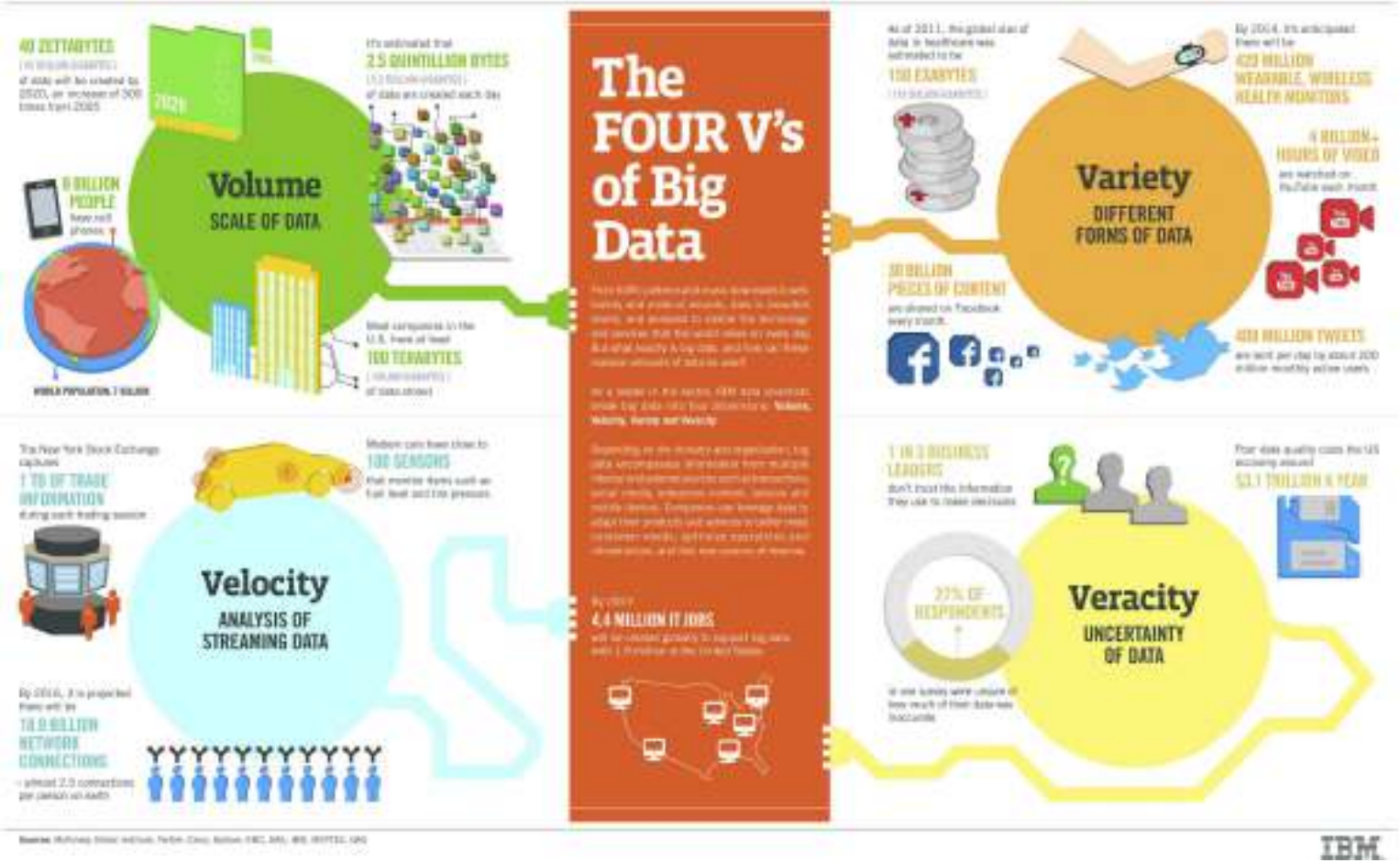
Some networks

And now health data

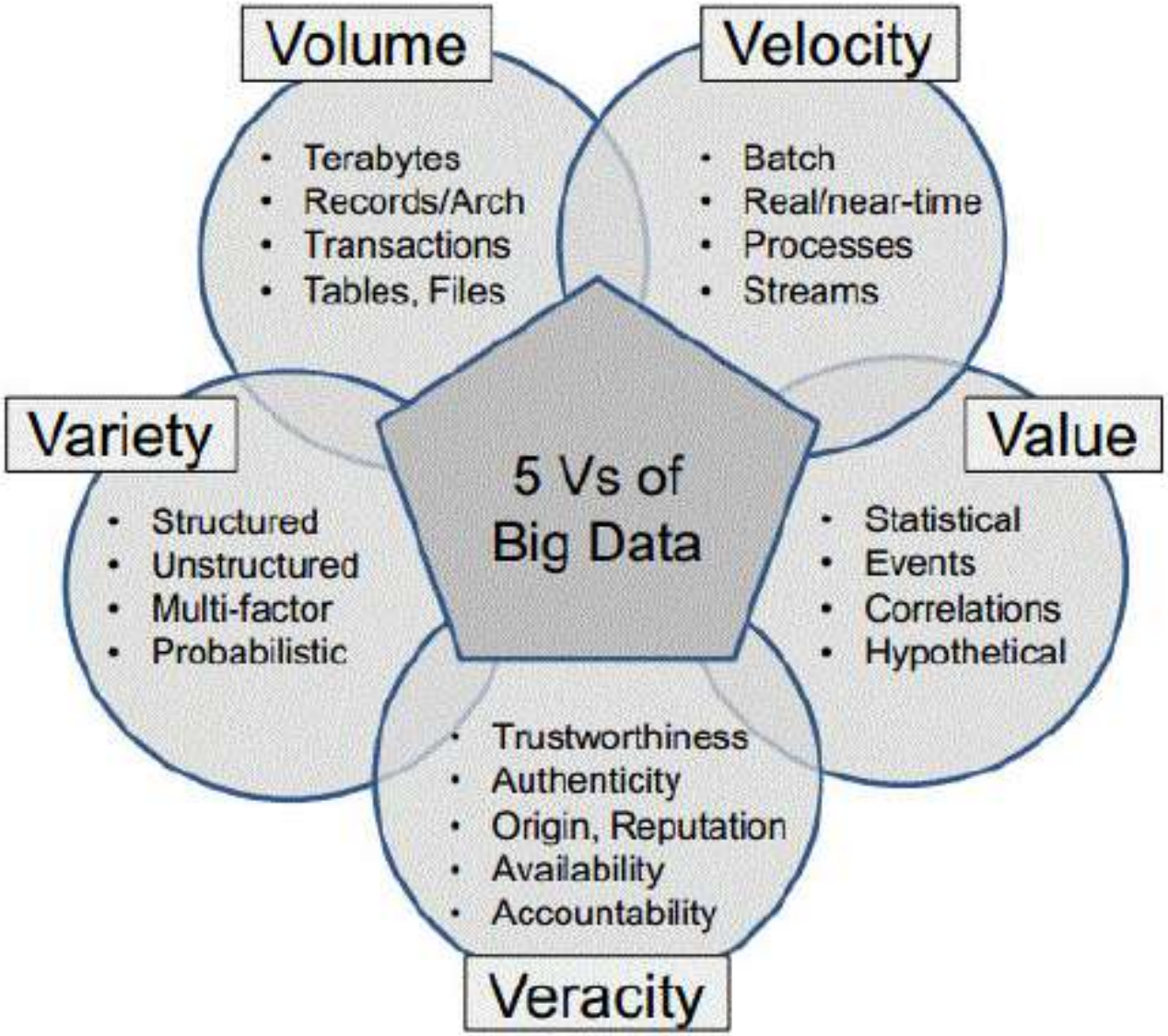
3 V ?



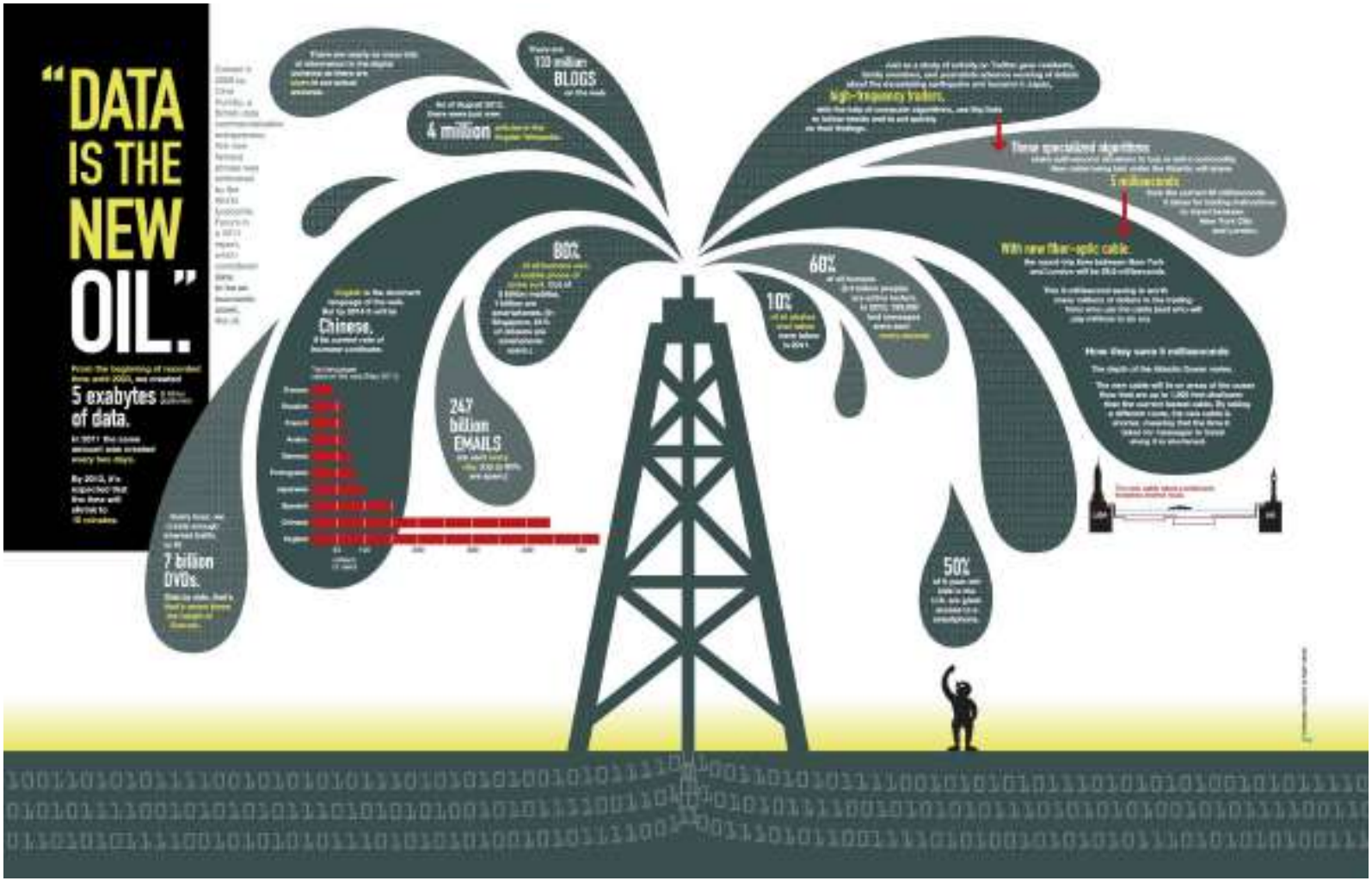
4 V ?



5 V ?

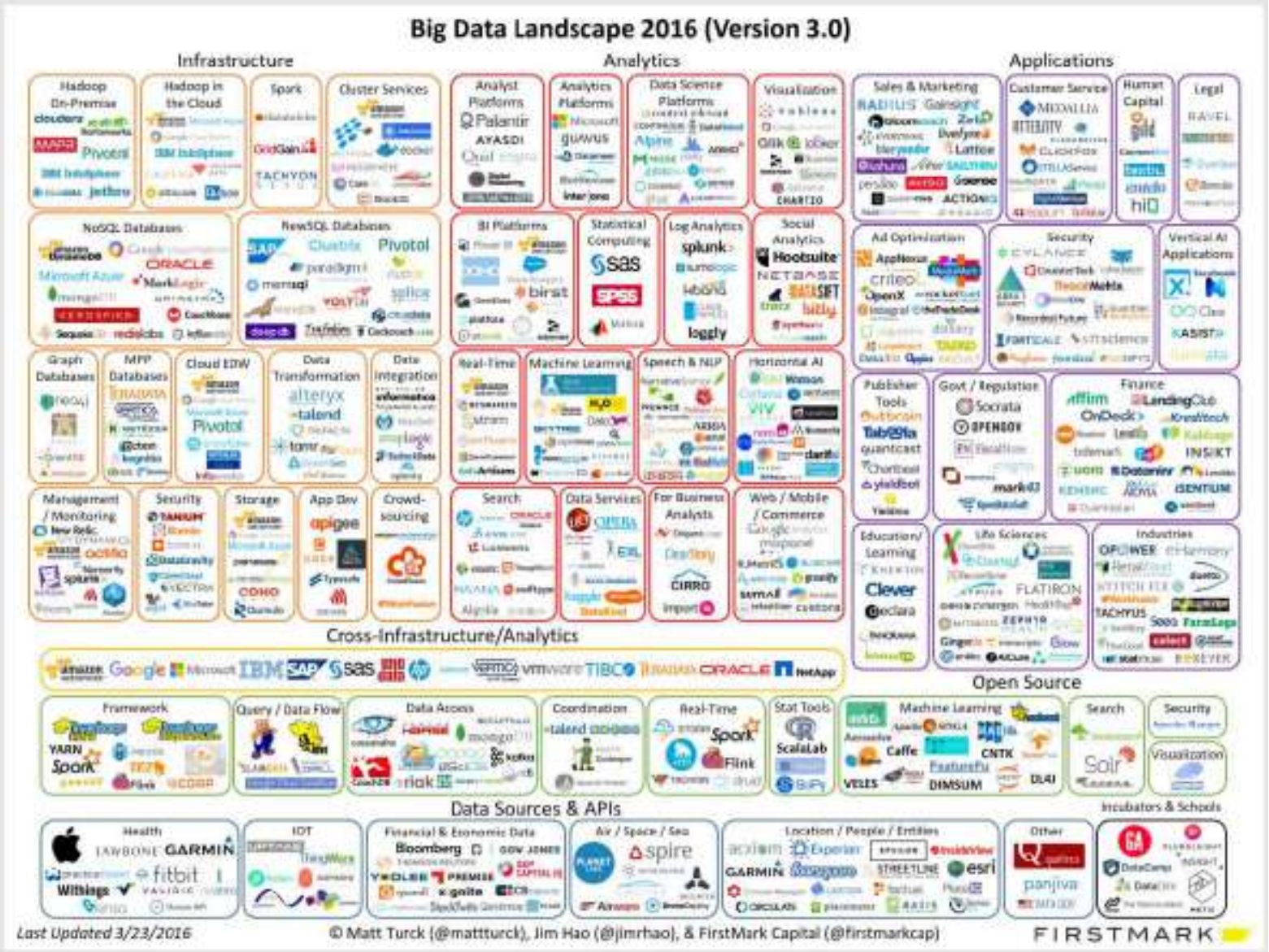

Volume
Terabytes
Records/Arch
Transactions
Tables, Files
Velocity
Batch
Real/near-time
Processes
Streams
Variety
Structured
Unstructured
Multi-factor
Probabilistic
5 Vs of
Big Data
Value
Statistical
Events
Correlations
Hypothetical
Trustworthiness
Authenticity
Origin, Reputation
Availability
Accountability
Veracity

The new oil ?



Clive Huby, 2006.

A landscape


Big Data Landscape 2016 (Version 3.0)
Infrastructure
Analytics
Applications
Cross-Infrastructure/Analytics
Open Source
Data Sources & APIs
Incubators & Schools
Last Updated 3/23/2016
© Matt Turck (@matttURCK), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)
FIRSTMARK

Gartner hype cycle 2017

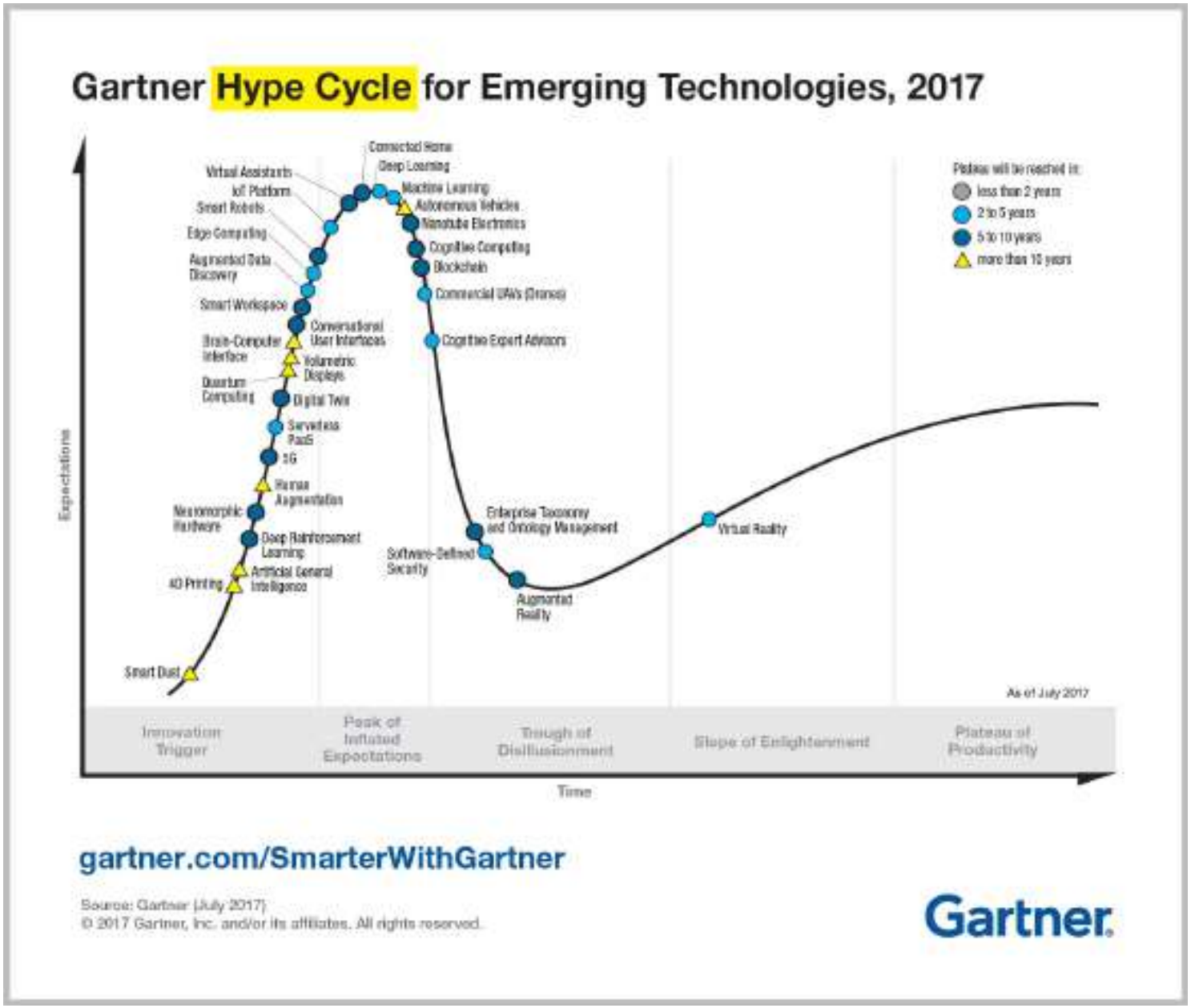


Table of contents

Data analytics

A process:

- collecting,
- organizing (cleaning and storing),
- analyzing,
- visualizing

large sets of data.

An objective: discover useful information to improve business decisions.

A new idea ?

Four major influences act on data analysis today:

- *The formal theories of statistics.*
- *Accelerating developments in computers and display devices.*
- *The challenge, in many fields, of more and ever larger bodies of data.*
- *The emphasis on quantification in an ever wider variety of disciplines.*

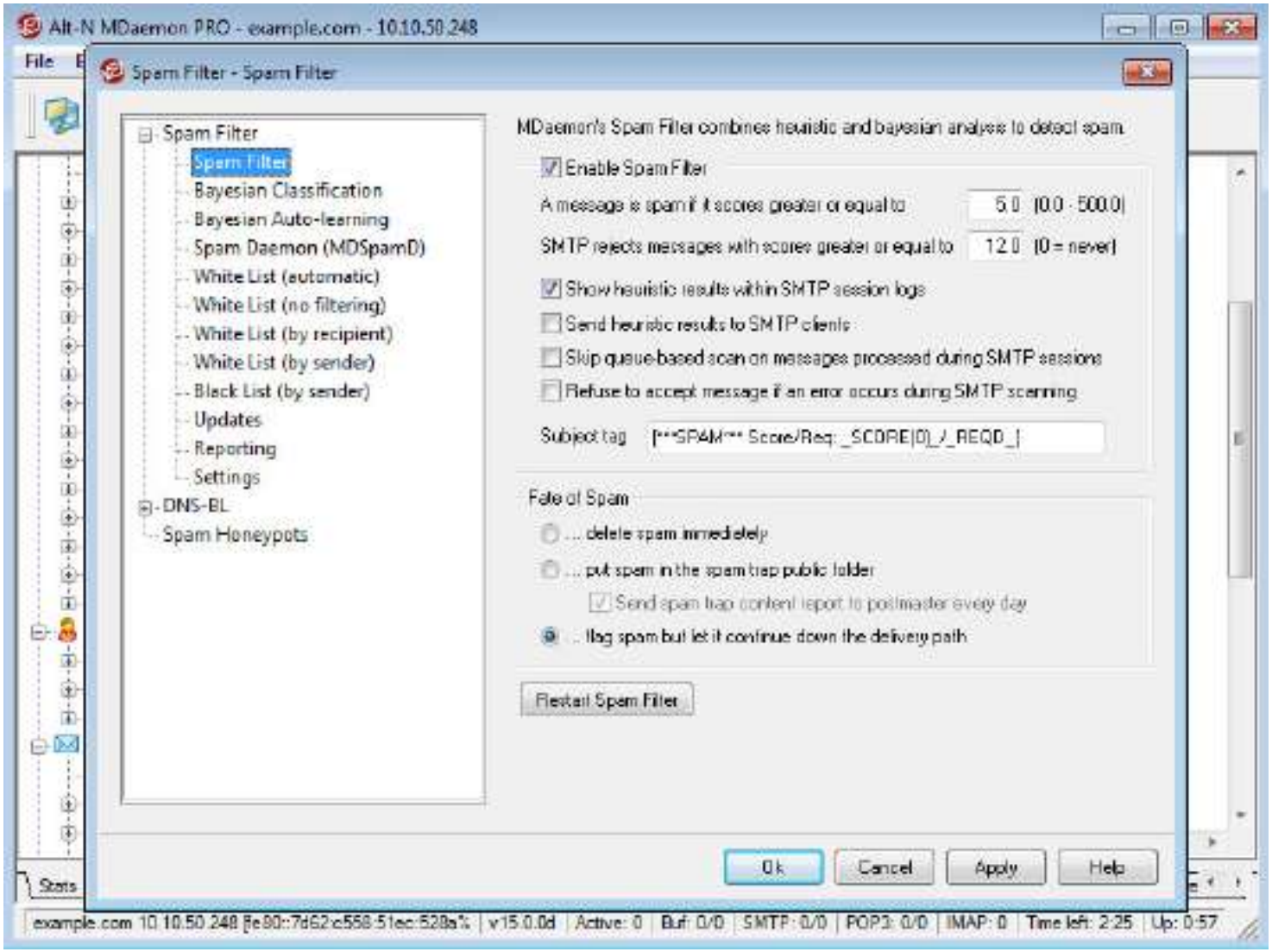
Not so new !

Data analysis and statistics: an expository overview
J. W. Tukey and M. B. Wilk
1966

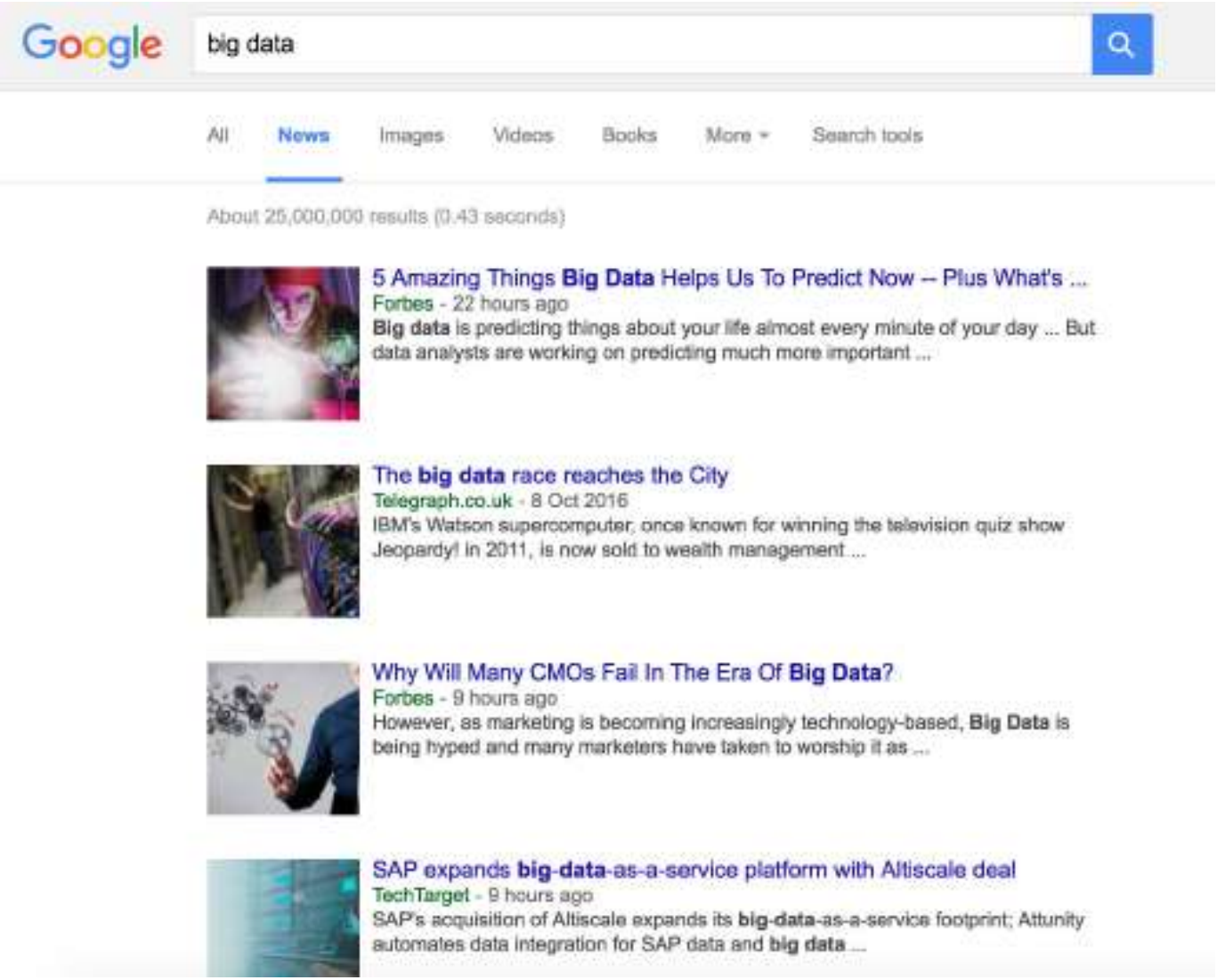
Four major influences act on data analysis today:

- *The formal theories of statistics.*
- *Accelerating developments in computers and display devices.*
- *The challenge, in many fields, of more and ever larger bodies of data.*
- *The emphasis on quantification in an ever wider variety of disciplines.*

Spam filter


Alt-N MDaemon PRO - example.com - 10.10.50.248
File
Spam Filter - Spam Filter
Spam Filter
Spam Filter
Bayesian Classification
Bayesian Auto-learning
Spam Daemon (MDSpamD)
White List (automatic)
White List (no filtering)
White List (by recipient)
White List (by sender)
Black List (by sender)
Updates
Reporting
Settings
DNS-BL
Spam Honeypots
MDaemon's Spam Filter combines heuristic and bayesian analysis to detect spam.
Enable Spam Filter
A message is spam if it scores greater or equal to
5.0
[0.0 - 500.0]
SMTP rejects messages with scores greater or equal to
12.0
[0 = never]
Show heuristic results within SMTP session logs
Send heuristic results to SMTP clients
Skip queue-based scan on messages processed during SMTP sessions
Refuse to accept message if an error occurs during SMTP scanning
Subject tag
[***SPAM*** Score/Req: _SCORE|0|_/_REQD_]
Fate of Spam
... delete spam immediately
... put spam in the spam trap public folder
Send spam trap content report to postmaster every day
... flag spam but let it continue down the delivery path
Restart Spam Filter
Ok
Cancel
Apply
Help
Stats
example.com 10.10.50.248 [fe80::7d62:c568:51ec:528a%
v15.0.0d
Active: 0
Buf: 0/0
SMTP: 0/0
POP3: 0/0
IMAP: 0
Time left: 2:25
Up: 0:57

Web search



Recommendations

Marketing


IBM Big Data & Analytics Hub
ABOUT CONTACT CONTENT BY TYPE SUBSCRIBE
USE CASES INDUSTRIES ANALYTICS TECHNOLOGY EVENTS AROUND THE WEB FOR DEVELOPERS BIG DATA & ANALYTICS HEROES
White Papers & Reports
Moving Up the Digital Marketing Maturity with Big Data Analytics
Much of the big data evolution is being driven by the vast commercial possibilities of the Internet. The scope and pace of change suggests the advent of an even more fundamental shift, focusing the broader advertising community on a series of critical questions: How can data most effectively be used to address core business needs? Where can this data be sourced? How can it be shared, optimized and enhanced for analysis and monetization?
Are you asking yourself these same questions? We hope so.
This paper will achieve the following objectives:
1. History of media and its impact to marketing
2. Critical marketing imperatives for data-driven digital marketing
3. Four critical optimizations
4. Steps to success in the digital marketing big data journey
5. Summarize concepts and steps digital marketers need to embrace if they are to harness, manage and successfully exploit big data
IBM Insight at World of Watson
Mandalay Bay, Las Vegas
October 24–27
Register now

Customer relationship management (CRM)

- *Hotel chain uses big data to increase bookings.*
- *Pizza chain earns more dough in bad weather.*
- *Music distributor applies big data for demand planning.*
- *Financial services company scores new clients.*
- *Retailer creates pregnancy detection model.*

Smart grids

And smart cities.

Genomics

nature international weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

News & Comment > News > 2015 > October > Article

NATURE | NEWS

E-alert RSS Facebook Twitter

Genome researchers raise alarm over big data

Storing and processing genome data will exceed the computing challenges of running YouTube and Twitter, biologists warn.

Erika Check Hayden

07 July 2015

Rights & Permissions

The computing resources needed to handle genome data will soon exceed those of Twitter and YouTube, says a team of biologists and computer scientists who are worried that their discipline is not geared up to cope with the coming genomics flood.

Other computing experts say that such a comparison with other 'big data' areas is not convincing and a little glib. But they agree that the computing needs of genomics will be enormous as sequencing costs drop and ever more genomes are analysed.

Refugee trauma

The mental-health crisis among migrants

The refugees and migrants surging into Europe are suffering very high levels of psychiatric disorders. Researchers are struggling to help.

Like Share 293k people like this. Be the first of your friends.

Enterprise Big Data

Table of contents

The data scientist

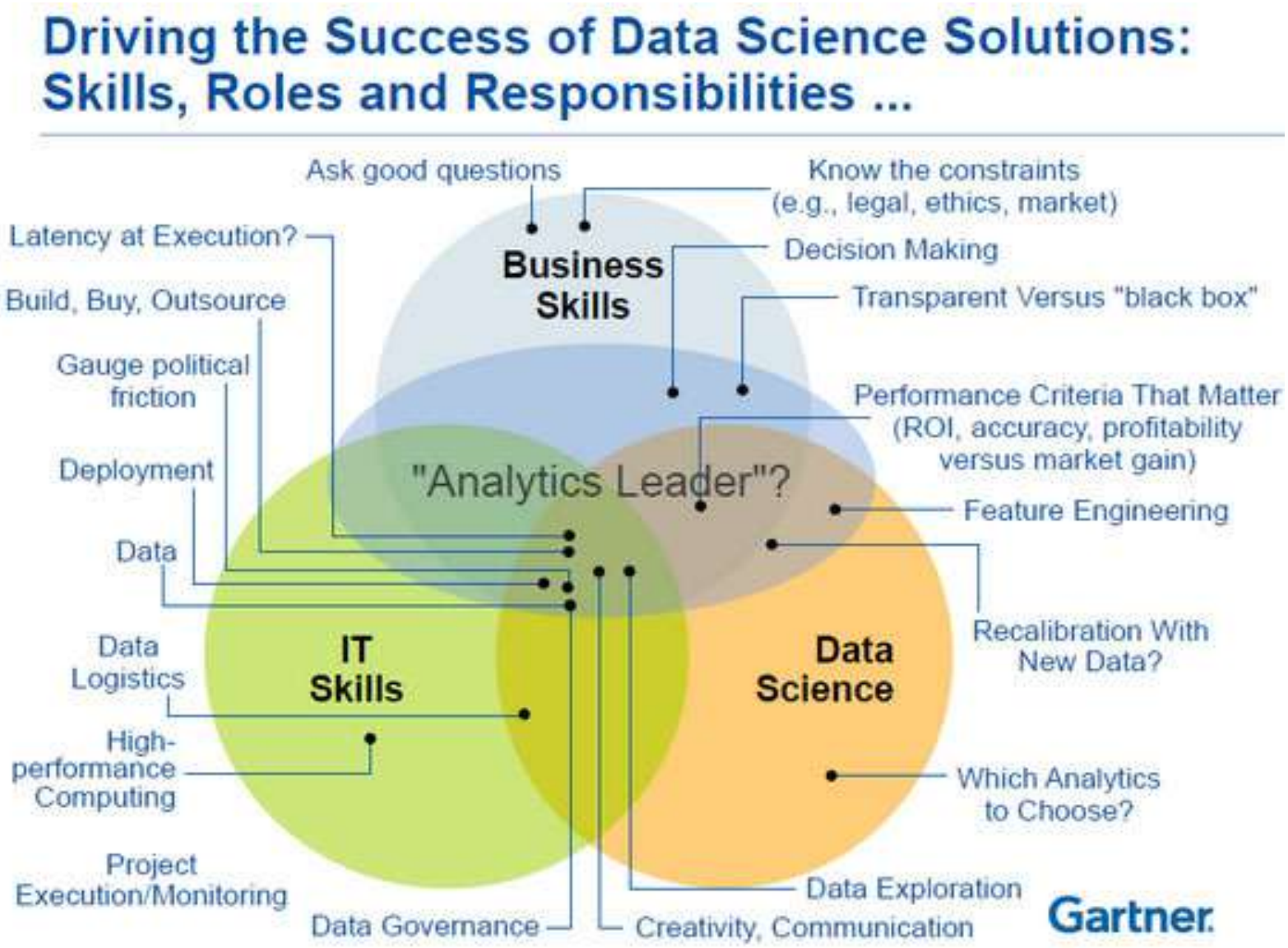
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Data scientist skills



Superhero skills ?

Some definitions: Data science

WIKIPEDIA
The Free Encyclopedia

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).

Table of contents

Some definitions: Machine learning

WIKIPEDIA
The Free Encyclopedia

Machine learning is a field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.

Some definitions: Statistical learning

Statistical learning theory is a framework for machine learning drawing from the fields of statistics and functional analysis. Statistical learning theory deals with the problem of finding a predictive function based on data. Statistical learning theory has led to successful applications in fields such as computer vision, speech recognition, bioinformatics and baseball.

Statistical learning vs Machine learning

- Machine learning, from Artificial Intelligence: large scale applications, prediction accuracy.
- Statistical learning, from Statistics: interpretability, precision, uncertainty, inference.
- For some statisticians: statistical learning is a mathematical formalisation of the machine learning.

Some concepts: online/offline learning

- Online learning (real-time): under time constraints.

  Some examples:

  - Personalized advertising.
  - Personalized healthcare.
  - Navigation & transit tools.
  - Autonomous cars.
  - Load curve forecasts.
  - Weather forecasts.

- Offline learning (batch).

Some concepts: supervised/unsupervised learning

- Supervised learning:
  Infer (predict) a function/relationship from labeled training data (e.g. classification, regression).

- Unsupervised learning:
  Find "structure" in unlabeled data (e.g. clustering).
  Even if it is more subjective than supervised learning, it can be useful as a pre-processing step for supervised learning.

Supervised learning

There are many different paradigms, including:

- Parametric statistics (linear or non-linear).
- Non-parametric statistics (local estimation methods, e.g smoothing kernel methods, k-nearest neighbors).
- Tree based methods.
- Support Vector Machines.
- Deep learning.

Some key points

- Trade-off between prediction accuracy and interpretability.
- Avoid over-fitting.
- Parsimonious model vs (full) black box: "less is more".

Table of contents

Outline

- Introduction.
- Unsupervised learning: PCA & clustering.
- Supervised learning:
  - Cross validation & bootstrap.
  - Reminders on linear regression & logistic regression.
  - Tree based methods.
  - Support Vector Machines.

Software tools