Vilnius universitetas

Matematikos ir informatikos fakultetas

DIDIEJI DUOMENYS (BIG DATA). KAS TAI?

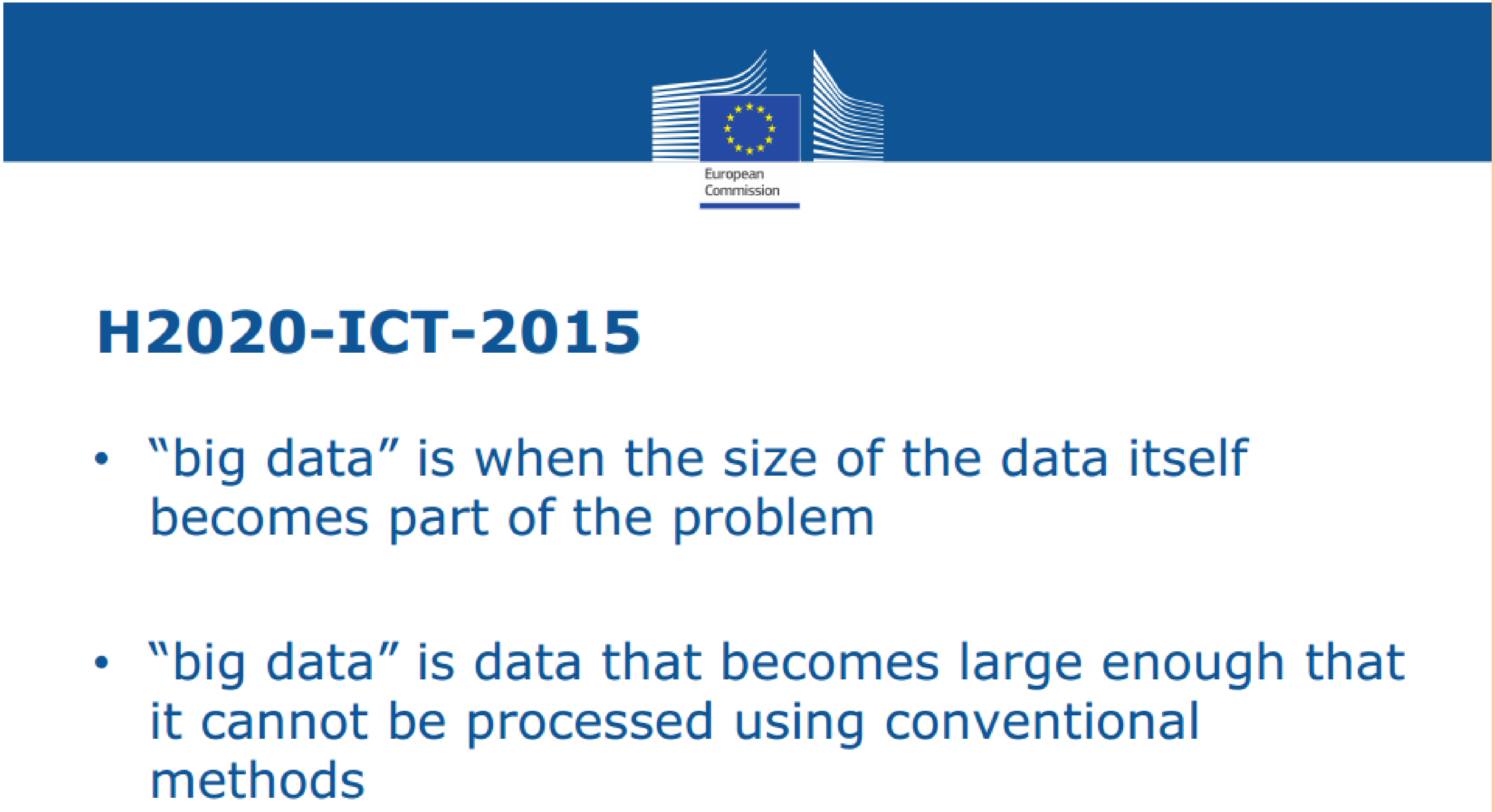
prof. dr. Olga Kurasova

Olga.Kurasova@mif.vu.lt

Didieji duomenys. Jų atsiradimo šaltiniai

- Modernios technologijos leidžia generuoti **milžiniškus duomenų kiekius**.
- Pagrindiniai **didžiųjų duomenų šaltiniai**:
  - Astronomija,
  - Meteorologija,
  - Genų inžinerija,
  - Medicina,
  - Bankinės ir finansinės sistemos,
  - Socialiniai tinklai,
  - Telekomunikacija,
  - Daiktų internetas (*Internet of Things*, IOT)
  - Saitynas (*web*).
  - Kiti

Didieji duomenys



European Commission

H2020-ICT-2015

- "big data" is when the size of the data itself becomes part of the problem

- "big data" is data that becomes large enough that it cannot be processed using conventional methods

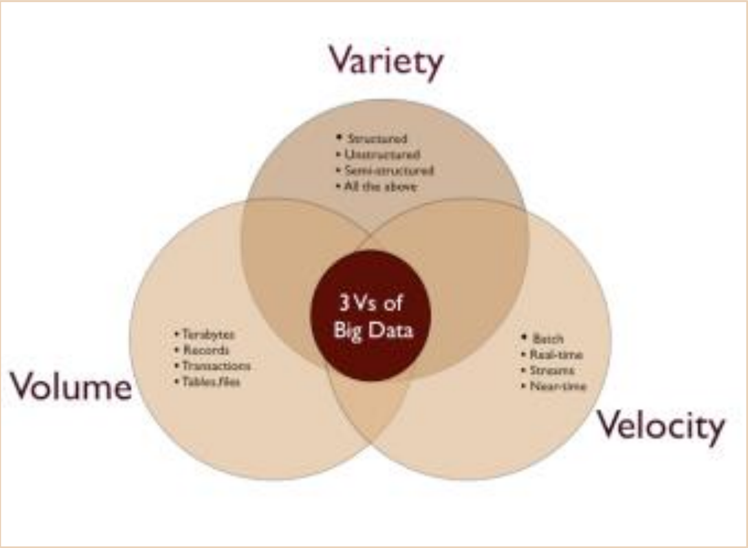
Didžiųjų duomenų ištakos

- **Visualization provides** an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of **big data**. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources (1997).
- Big data is high-**volume**, high-**velocity** and/or high-**variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

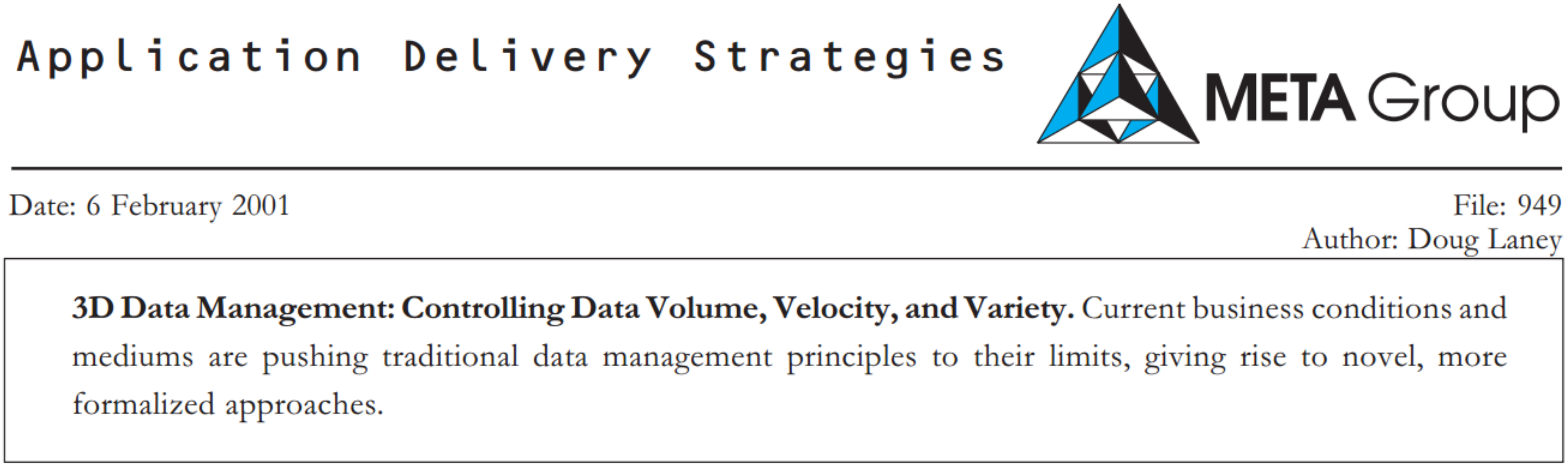
Didieji duomenys (3V)

Big data can be described by the following characteristics:

- **Volume** (didžiulis duomenų kiekis).
- **Variety** (plati duomenų įvairovė).
- **Velocity** (nuolat atsirandantys nauji duomenys).



Didieji duomenys (3V)



Application Delivery Strategies

META Group

Date: 6 February 2001

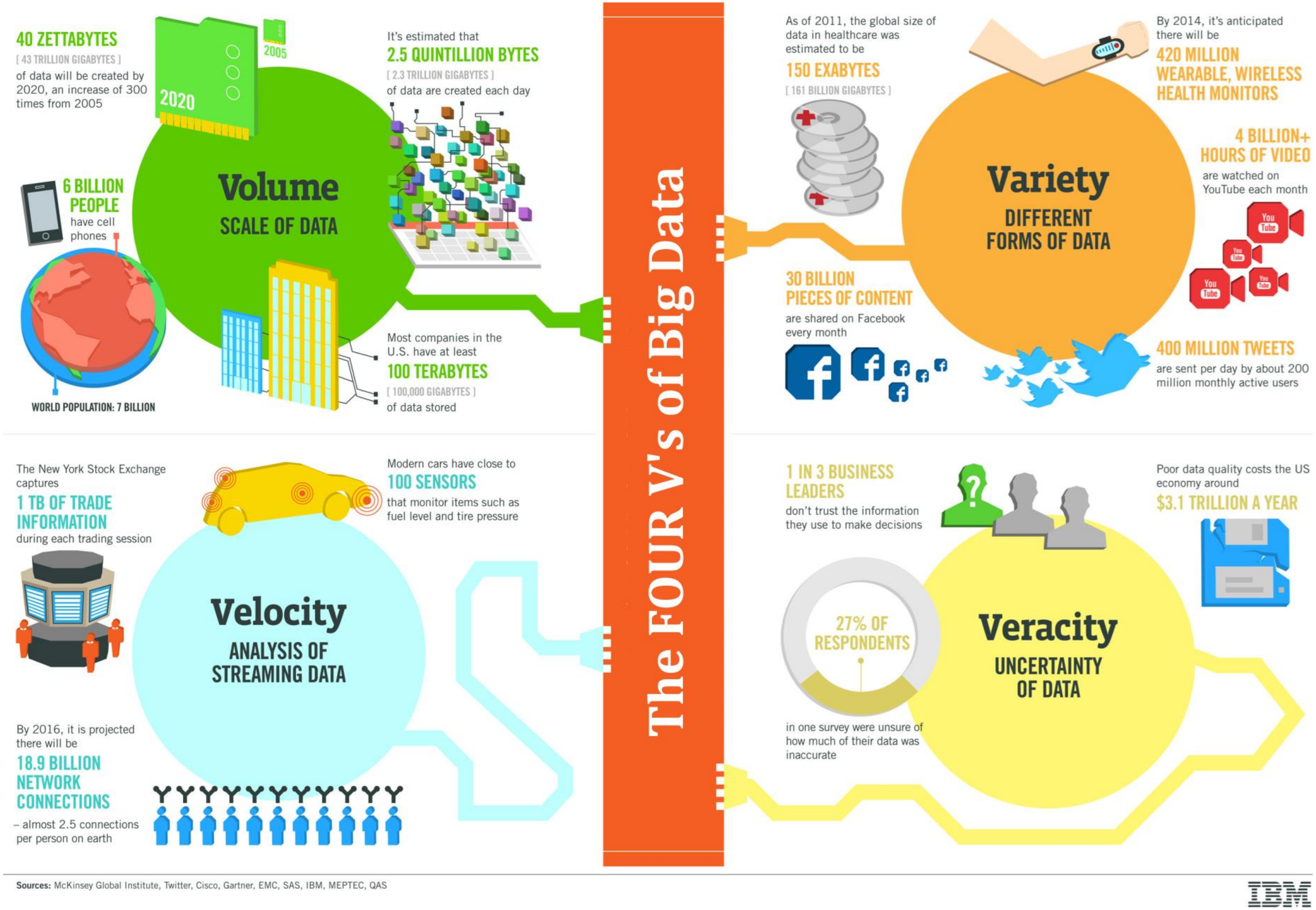
File: 949
Author: Doug Laney

3D Data Management: Controlling Data Volume, Velocity, and Variety. Current business conditions and mediums are pushing traditional data management principles to their limits, giving rise to novel, more formalized approaches.

Very difficult to define (precisely):

- data is "**big**" if it defies traditional processing & storage paradigms – bigness becomes part of the problem.
- the "3Vs": volume (size), velocity (bytes/s), variety (database, jpeg, video, numbers, text in language X...).
- ...to which we add the 4th V to denote creation of Value (by linking, aggregating, analysing, visualizing...).

Didieji duomenys (4V)



The FOUR V's of Big Data

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

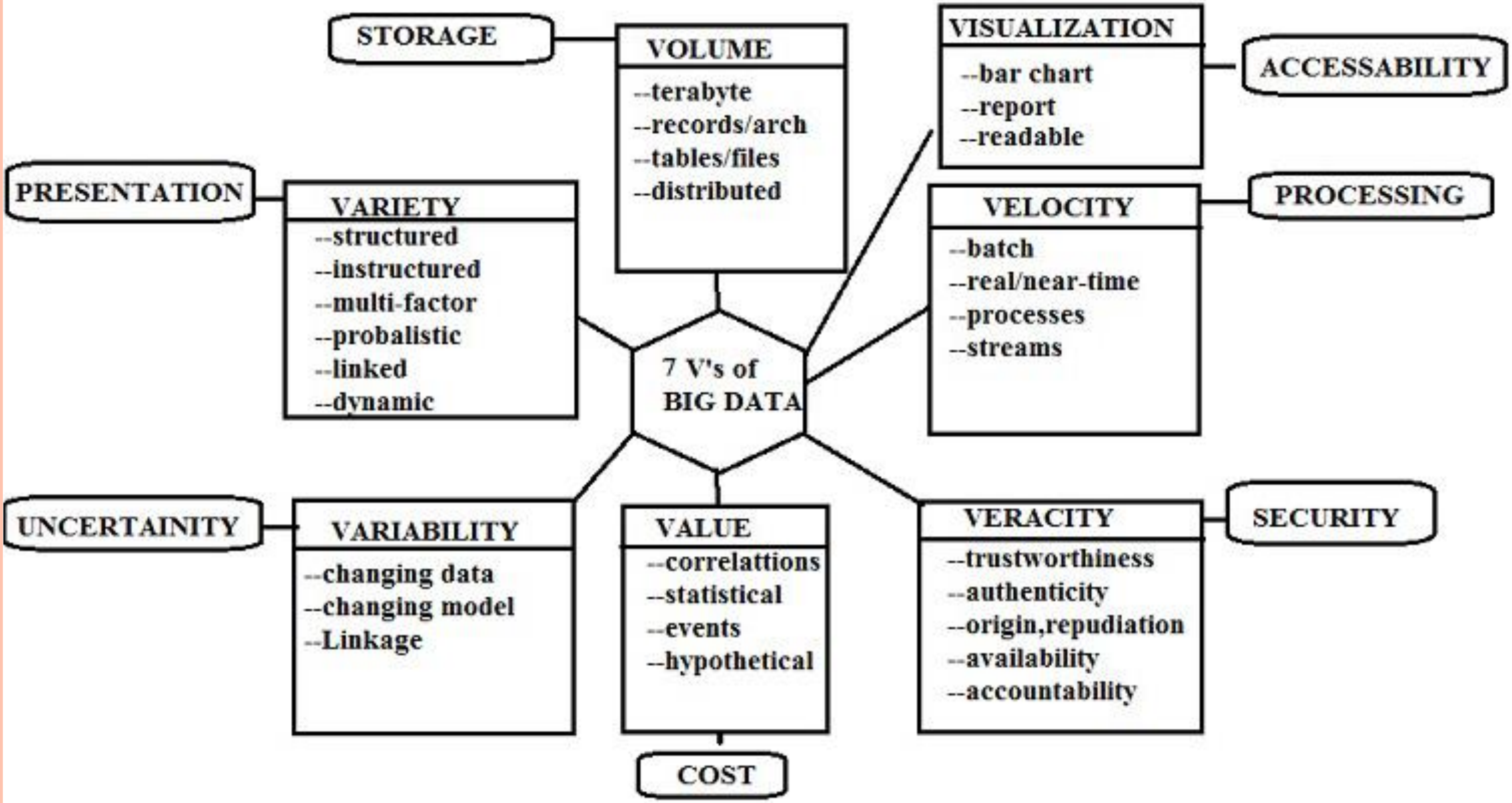
http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

Didieji duomenys (5V)

- **Volume** – from terabytes to petabytes and up.
- **Variety** – an expanding universe of data types and sources.
- **Velocity** – accelerated data flow in all directions.
- **Variability** – inconsistent data flows with periodic peaks.
- **Complexity** – the need to correlate and share data across entities.

“**Big data** is not only about analytics, it's about the whole pipeline. So when you think about big data solutions, you have to think about all the different steps: collect, store, organize, analyze, and share,” said Amazon CTO Werner Vogels.

Didieji duomenys (7V)



Didieji duomenys

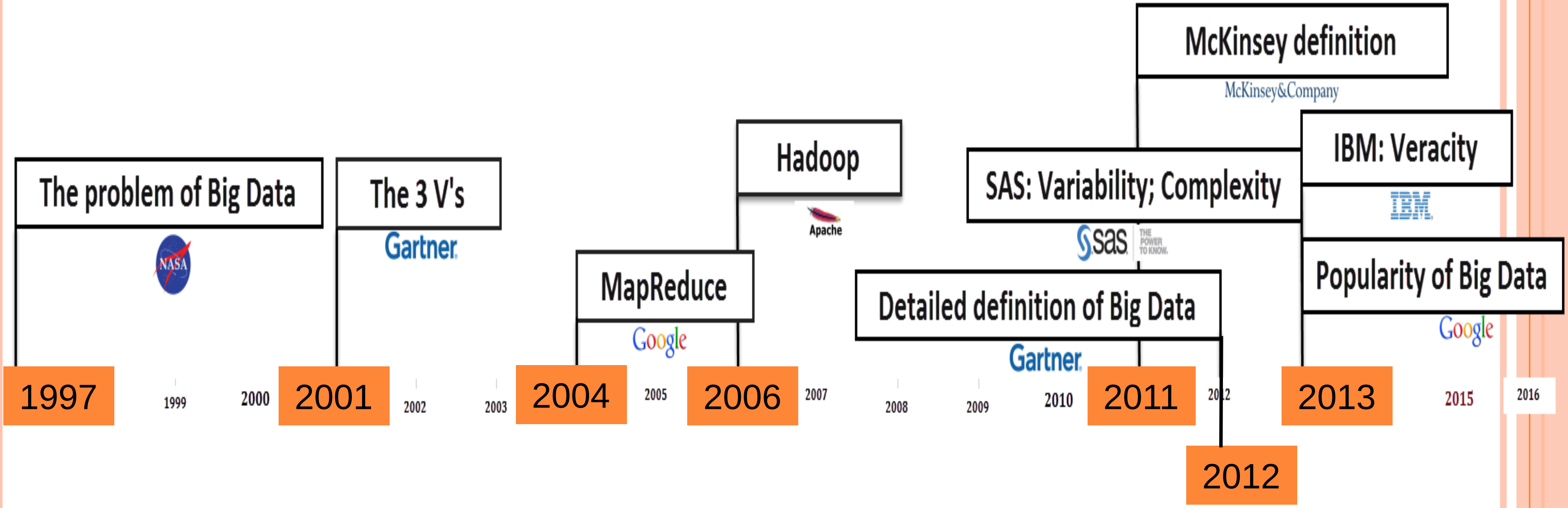
- **Big data** refers to data being collected in ever-escalating volumes, at increasingly high velocities, and for a widening variety of unstructured formats and variable semantic contexts. Big data can be historical (meaning stored data) or real-time (meaning streamed directly from the source).

Microsoft

- Every day, we create 2.5 quintillion bytes of data—so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data**.

IBM

Didieji duomenys: anksčiau ir dabar



1997 – The problem of Big Data

2001 – The 3 V's

2004 – MapReduce

2006 – Hadoop

2011 – McKinsey definition

2011 – SAS: Variability; Complexity

2012 – Detailed definition of Big Data

2013 – IBM: Veracity

2013 – Popularity of Big Data

What Types of Data are in Big Data?

- Structured Data
  - Tables, Relational Data …
- Unstructured Data
  - Raw text, Images, Video, Audio …
- Semi-structured
  - Hybrid data, such as documents with tables …
- Metadata
  - Structured data about data
- Streaming Data
  - Data that moves across networks at high speed
- Temporal Data
  - Data including trends/activities in time
- Geospatial Data
  - Data that includes information on positions in space
- Many others …

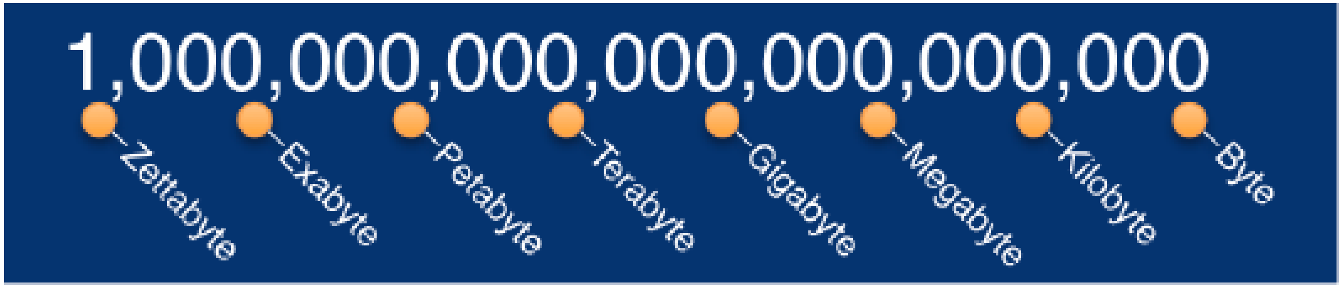
by the poster created by analysts at Altamira

Matavimo vienetai

Reikšmė	Žymėjimas	Pavadinimas
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	youttabyte

How Big is Big Data?

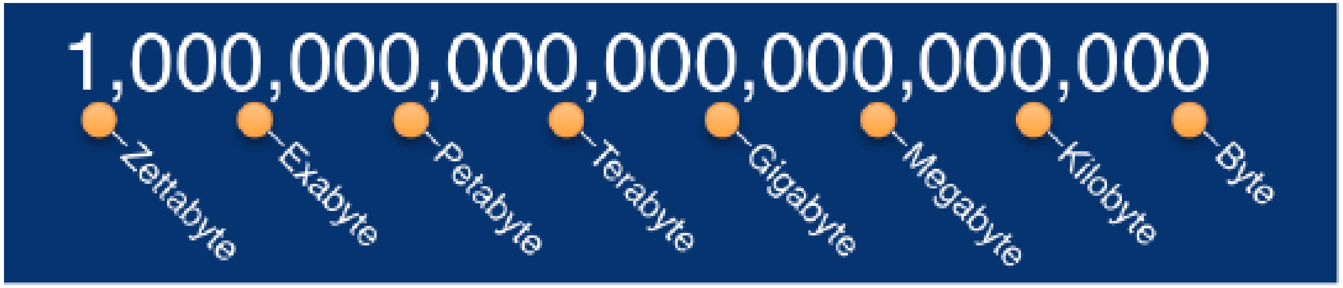
- Internet traffic is now ~**5 Zettabyte** per year (IBM)
- **Visa** processes 150 Million transactions per day (VISA)
- **Library of Congress** holds 3.2 Petabytes of data
- 207 Terabytes of video loaded daily on **YouTube** (2012)
- 50 billion **devices connected to the Internet** by 2020 (IDC)
- 50 Billion photos on **Facebook** in 2010
- 400 Million **Tweets** per day (Washington Post)



by the poster created by analysts at Altamira

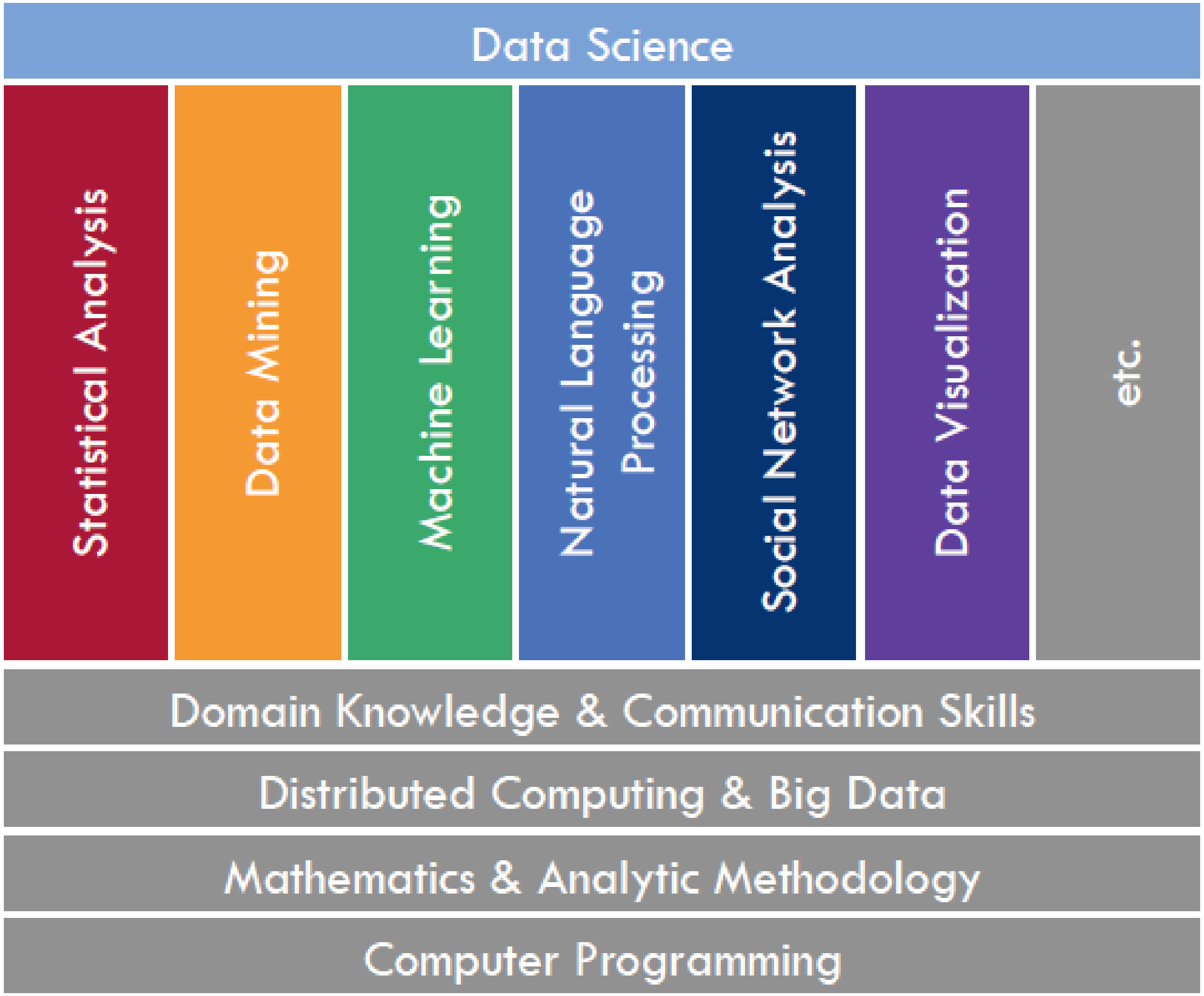
How Big is Big Data?

- Seagate sold 330 Exabytes of **hard drives** in 2011
- LHC produces 500 Exabytes of particle collision data per day **CERN**!
- **iPhone** 5s: 76 Gigaflops
- Fastest **supercomputer**: 50 Petaflops
- Interesting Comparison: **Human Brain** has 100 Billion Neurons (100 Giga-Neurons), 100 Trillion Synapses (100 Tera-Synapses), neurons “fire” 1-1000 times/second (100 Giga-fires to 100 Tera-fires per second)



by the poster created by analysts at Altamira

What is a Data Scientist?



Data Science

Statistical Analysis

Data Mining

Machine Learning

Natural Language Processing

Social Network Analysis

Data Visualization

etc.

Domain Knowledge & Communication Skills

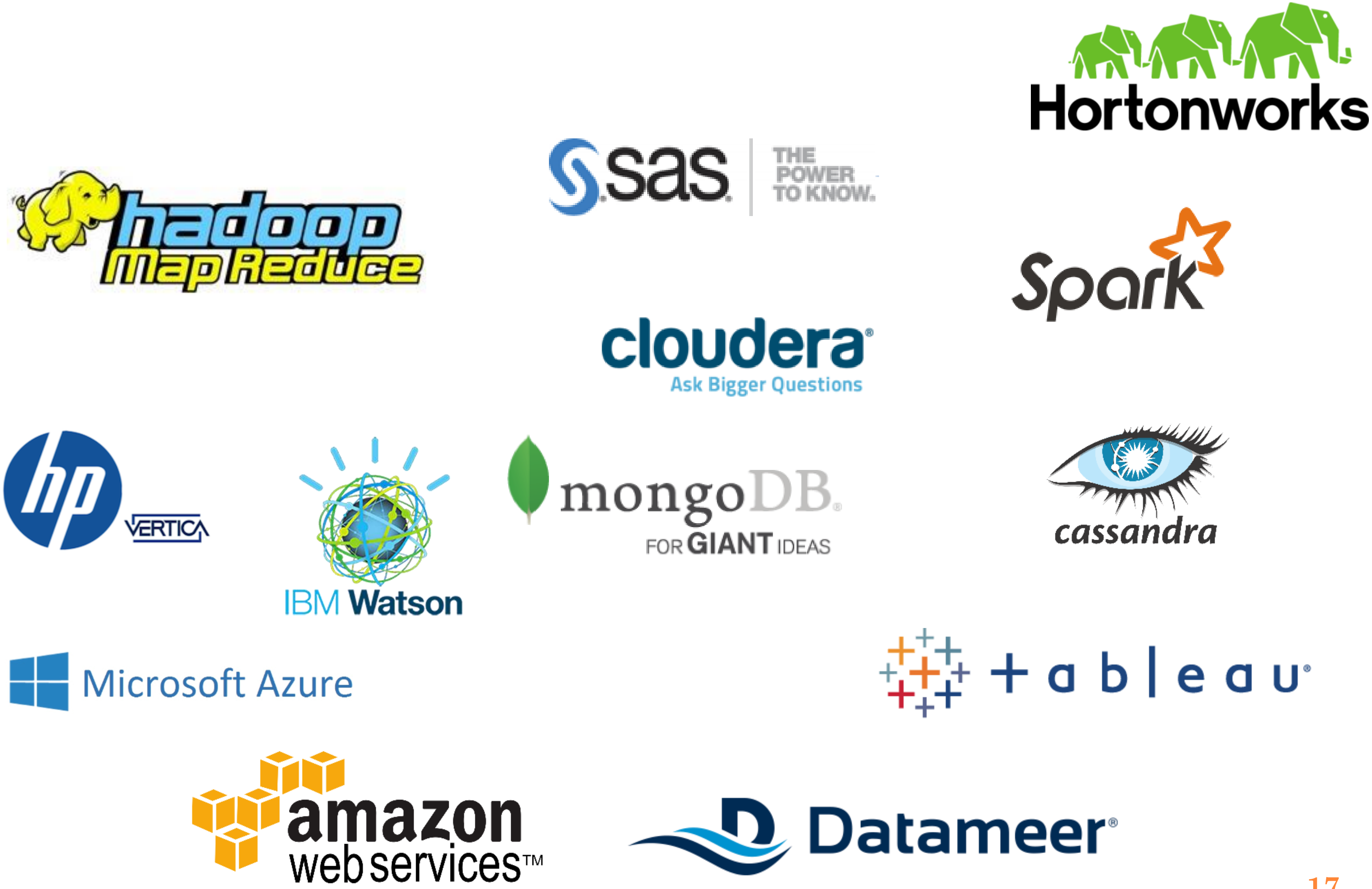
Distributed Computing & Big Data

Mathematics & Analytic Methodology

Computer Programming

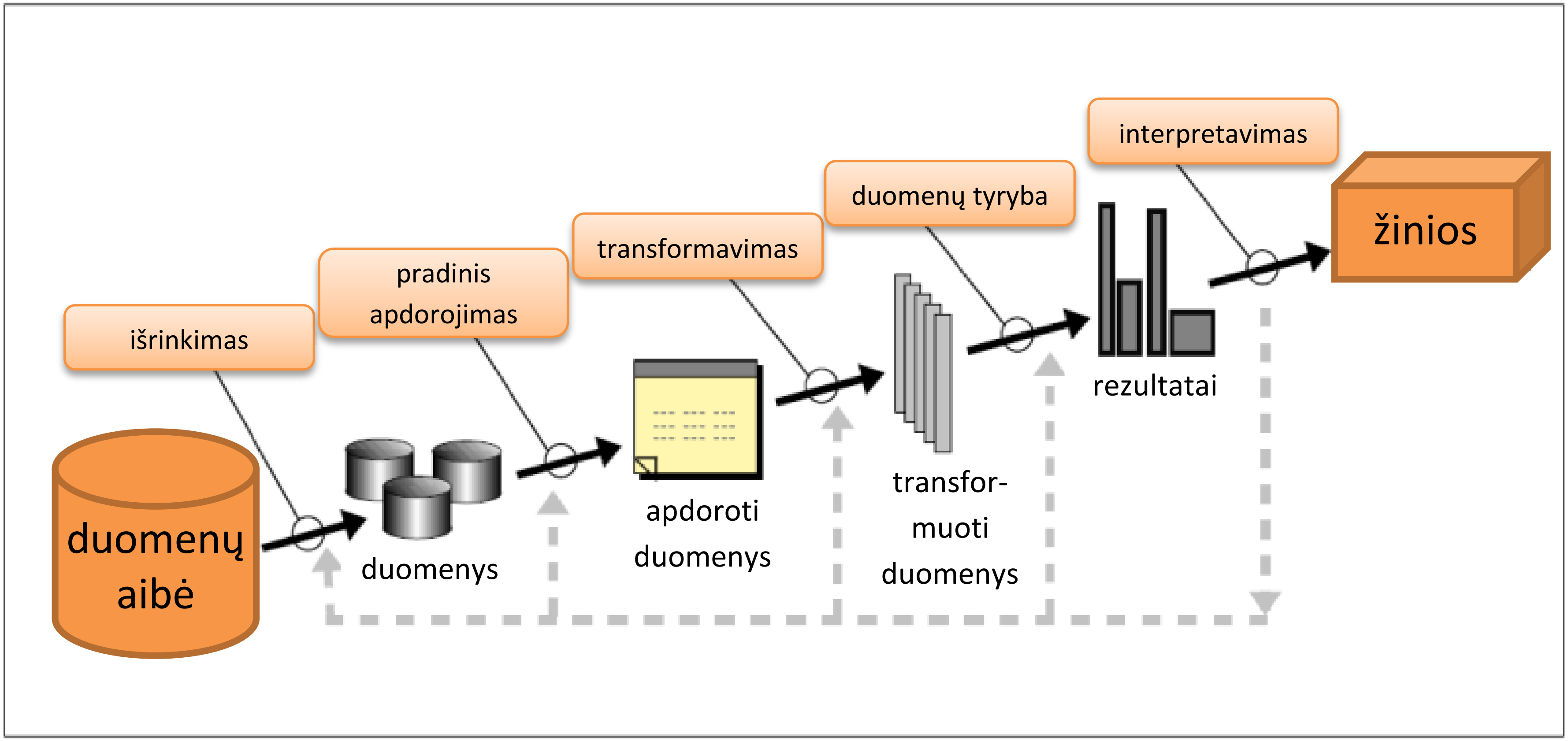
by the poster created by analysts at Altamira

Su didžiaisiais duomenimis susiję sprendimai ir technologijos



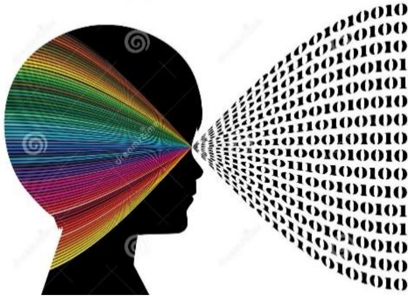
Duomenų tyryba didžiųjų duomenų eroje

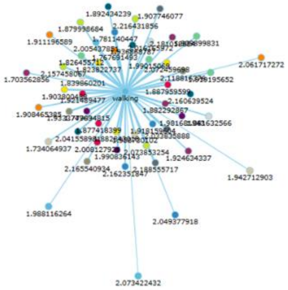
Duomenų tyryba žinių radimo procese

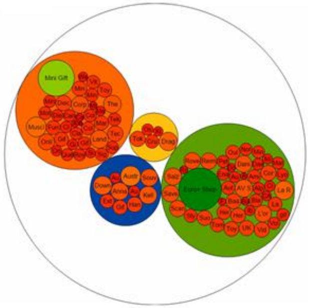


Duomenų vizuali analizė

- **Vizualizuoti** (*lot. visualis – regimas*) – nematomą atvaizdą, daiktą, reiškinį, daryti matomą (Tarptautinių žodžių žodynas, 1985).
- **Vizualizavimas** – tai informacijos kodavimas į regimuosius vaizdus.
- **Vizualizavimas** – tai grafinis informacijos pateikimas, palengvinantis informacijos suvokimą.

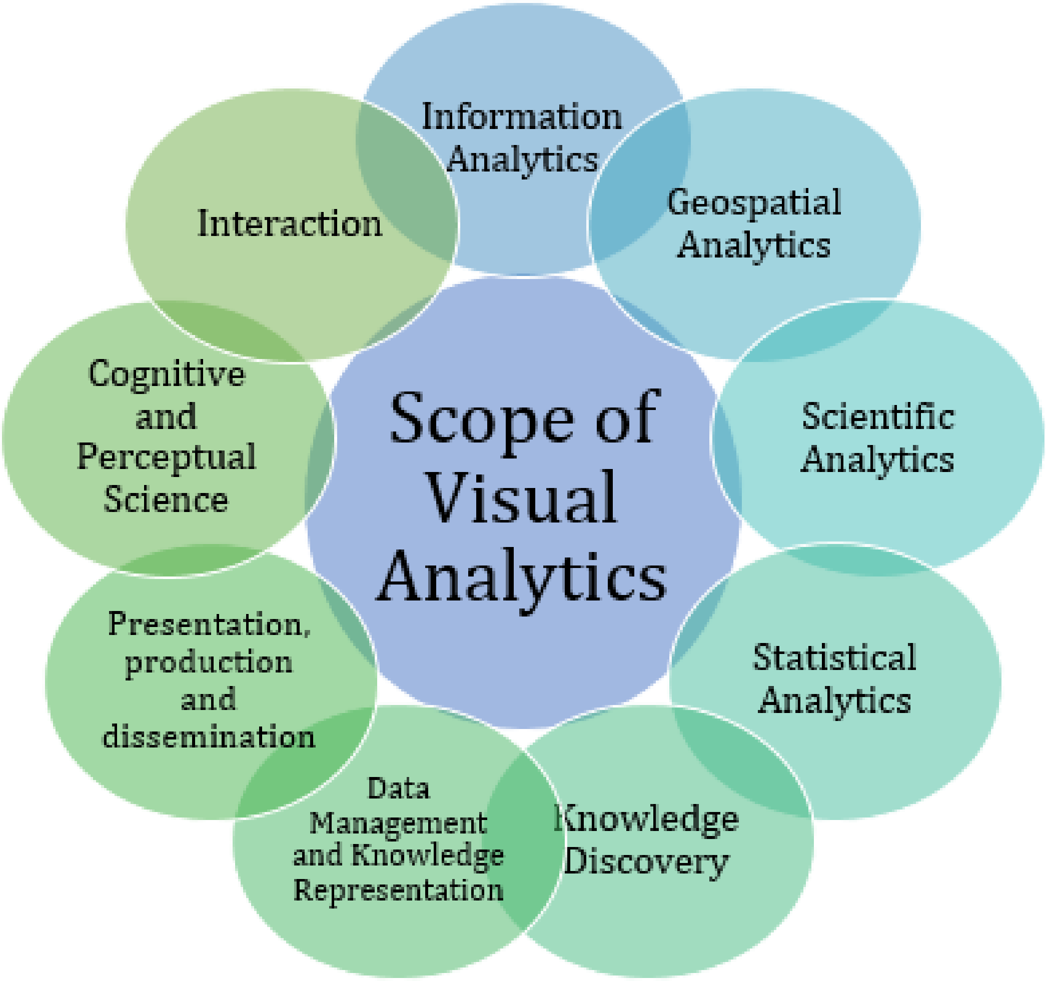




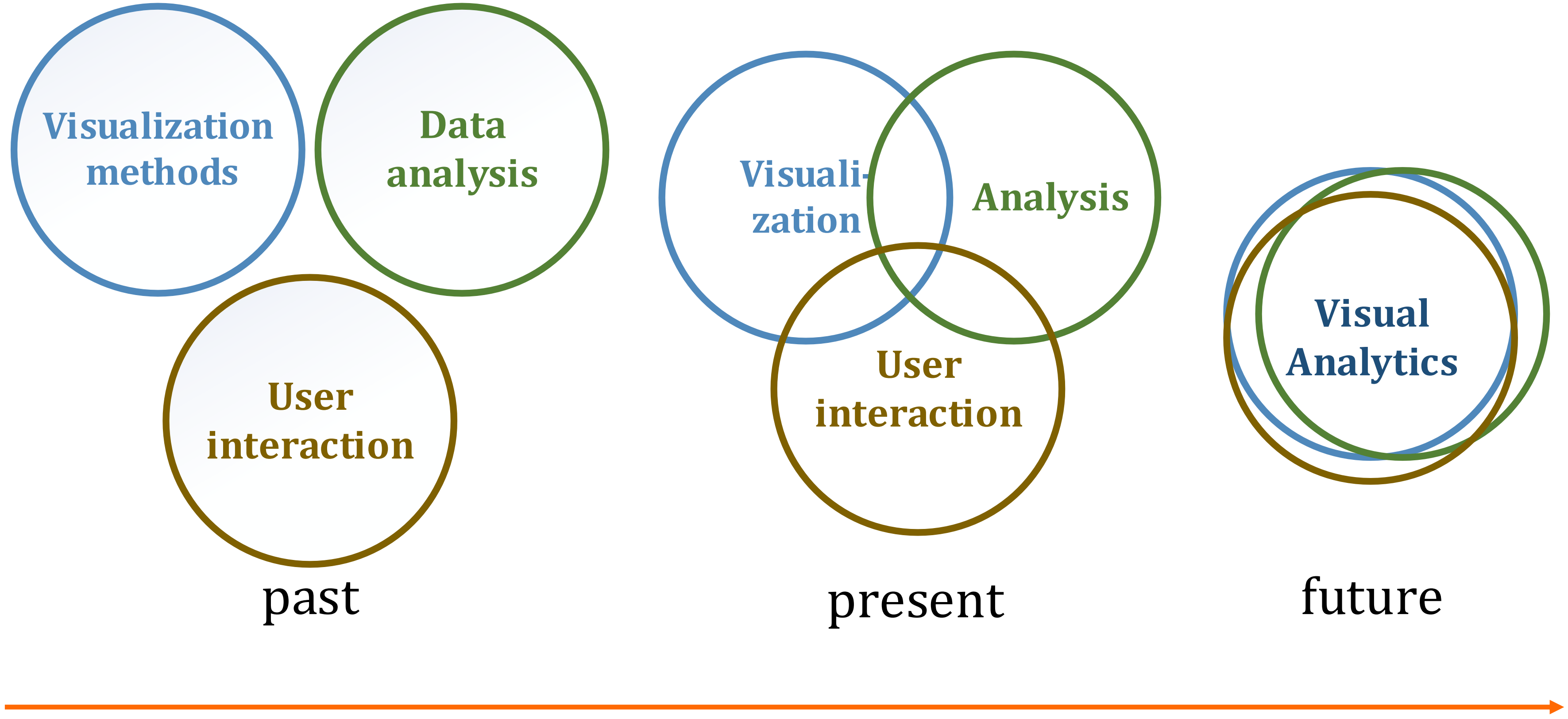


Visual Analytics

Visual Analytics combines automated analysis techniques with **interactive visualizations** for an effective **understanding**, reasoning and **decision making** on the basis of very large and complex datasets.

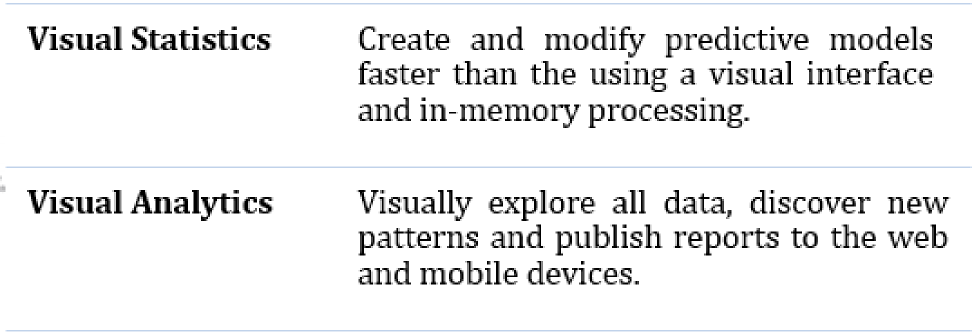

Information Analytics
Interaction
Geospatial Analytics
Cognitive and Perceptual Science
Scope of Visual Analytics
Scientific Analytics
Presentation, production and dissemination
Statistical Analytics
Data Management and Knowledge Representation
Knowledge Discovery

Vizualizavimas – anksčiau, dabar, ateityje



Visual Analytics as a Web Services

- **SAS** is an integrated system of software solutions for data management, graphics design, statistical and mathematical analysis, business forecasting and decision support.

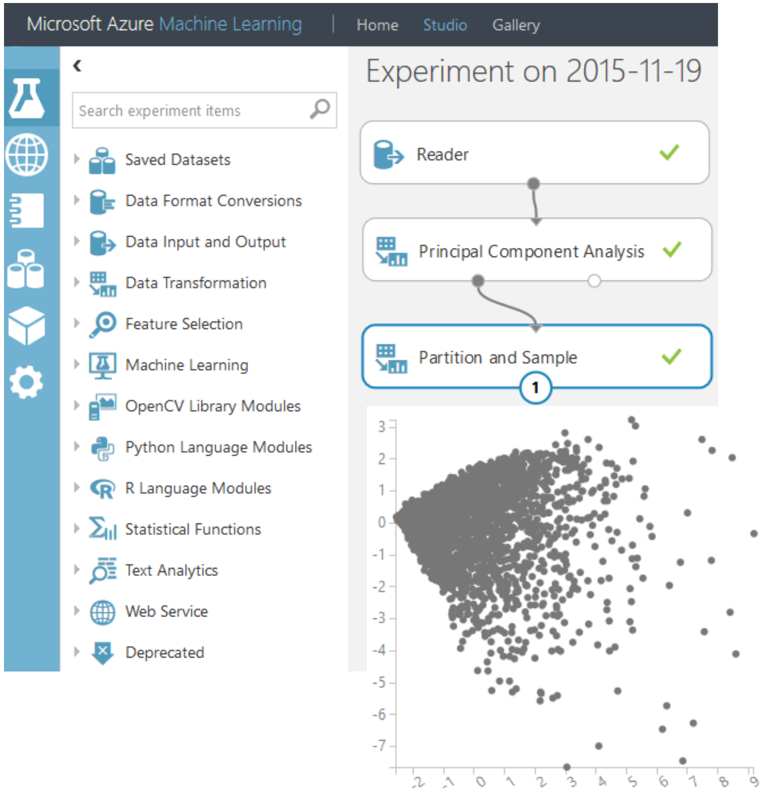


Visual Statistics	Create and modify predictive models faster than the using a visual interface and in-memory processing.
Visual Analytics	Visually explore all data, discover new patterns and publish reports to the web and mobile devices.

- **Oracle** Business Analytics is analytics for insight and innovation. **Oracle Data Visualization Cloud Service** get instant clarity with stunningly visual analysis and self-service discovery.

Visual Analytics as a Web Services

Microsoft Azure is an open, flexible, enterprise-grade cloud computing platform. **Microsoft Azure Machine Learning** is a service to build predictive analytics models and to easily deploy those models for consumption as cloud web services.

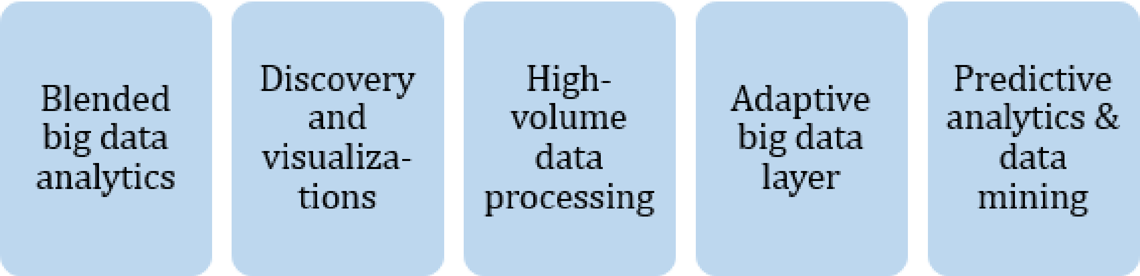


Platforms for Visual Analytics

Pentaho, a Hitachi Group Company, is a leading data integration and business analytics company with an enterprise-class, open source-based platform for diverse **big data deployments**.

pentaho

A Hitachi Group Company



Blended big data analytics

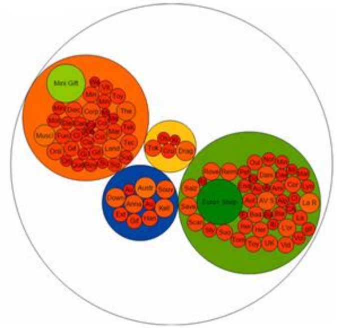
Discovery and visualiza- tions

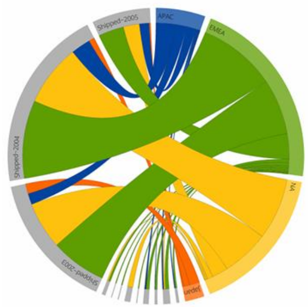
High- volume data processing

Adaptive big data layer

Predictive analytics & data mining

hadoop





Platforms for Visual Analytics

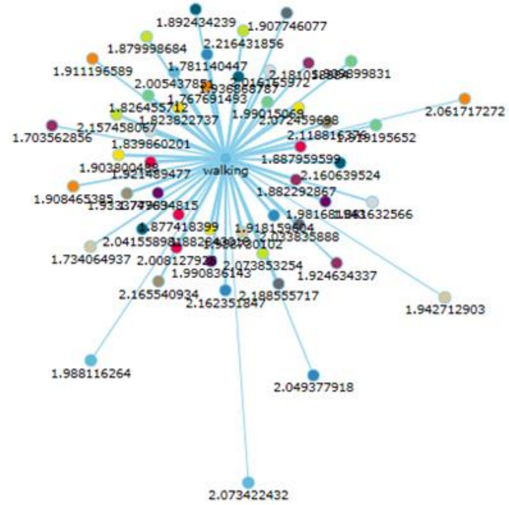
Amazon Web Services offers a broad set of global compute, storage, database, analytics, application, and deployment services.

amazon web services

Machine Learning

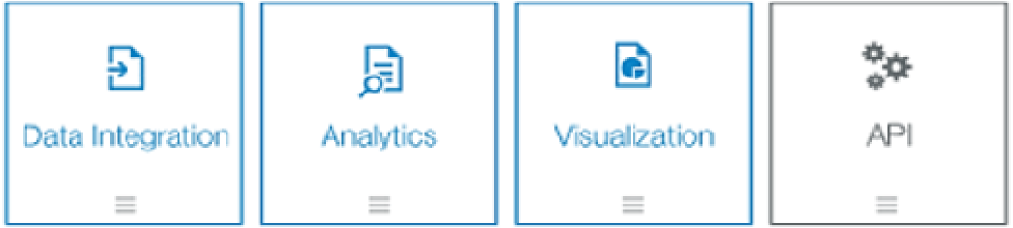
S3

EC2



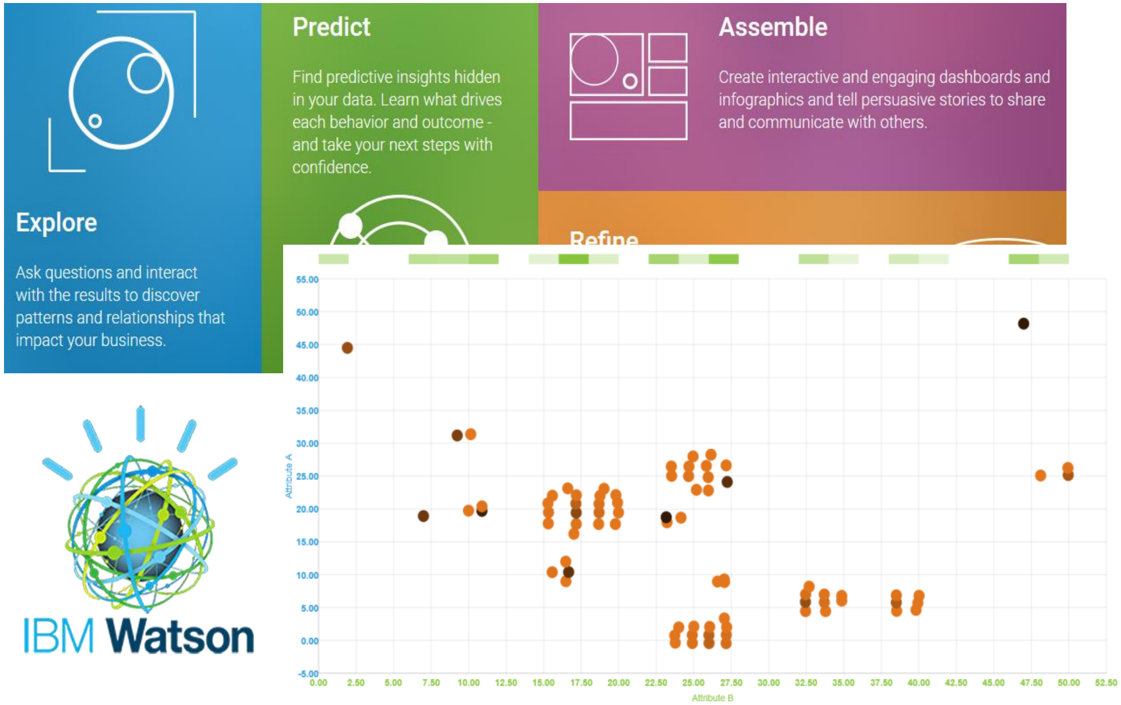
Datameer is big data analytics platform for Hadoop that empowers business users to directly integrate, analyze, and visualize any data.

Datameer
stay curious



Cloud based Data Exploration

- **IBM Watson** is a technology platform that uses **natural language processing** and machine learning to reveal insights from large amounts of unstructured data.
- **IBM Watson Analytics** offers the benefits of advanced analytics without the complexity.



Predict

Find predictive insights hidden in your data. Learn what drives each behavior and outcome - and take your next steps with confidence.

Assemble

Create interactive and engaging dashboards and infographics and tell persuasive stories to share and communicate with others.

Refine

Explore

Ask questions and interact with the results to discover patterns and relationships that impact your business.

Skaitmeninis intelektas ir didieji duomenys

- Skaitmeninio intelekto metodai **suteikia galingus įrankius** sprendžiant didžiųjų duomenų iššūkius.
- Evoliuciniai skaičiavimai, neuroniniai tinklai, neraiškiosios sistemos **geba susidoroti su dideliais kiekiais neapibrėžtumų**, susijusių su didžiųjų duomenų įvairumu ir nepastovumu.
- Tačiau didžiųjų duomenų kiekiai **kelia didelius iššūkius** ir skaitmeninio intelekto metodų efektyviam taikymui.