social computing

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

Big Data and Hadoop

Prof. Dr. Mohamed Amine Chatti
Dr. Arham Muslim

Social Computing Group, University of Duisburg-Essen
www.uni-due.de/soco/

Course Content

Learning Analytics

What?

Data, Environments, Context

Big Data

Hadoop Ecosystem

Why?

Objectives

Personalization

Recommender Systems

How?

Methods

Machine Learning / Data Mining

Information Visualization & Visual Analytics

Social Network Analysis (SNA)

Who?

Stakeholders

Learner Modeling

Privacy

Agenda

- Big Data
- Core Hadoop
- The Hadoop Ecosystem

Data is the new …

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Data is the new oil!

Data is the new gold!

Data is the new currency!

Data drives all what we do

In the last 10 minutes we generated more data than from prehistoric time until 2003!

We are all generating data!

Making a phone call

Getting a speeding ticket

Refueling your car

Sending an email

Taking the train

Making an appointment

Adjusting the temperature

Attending this lecture

Sensors are ubiquitous

GPS

Bluetooth

Touchscreen

GSM/LTE/5G

Light sensor

Finger-print scanner

Camera (back)

Camera (front)

Proximity sensor

Gyroscopic sensor

Magnetometer

Accelerometer

Wi-Fi

Microphone

Drowning in data

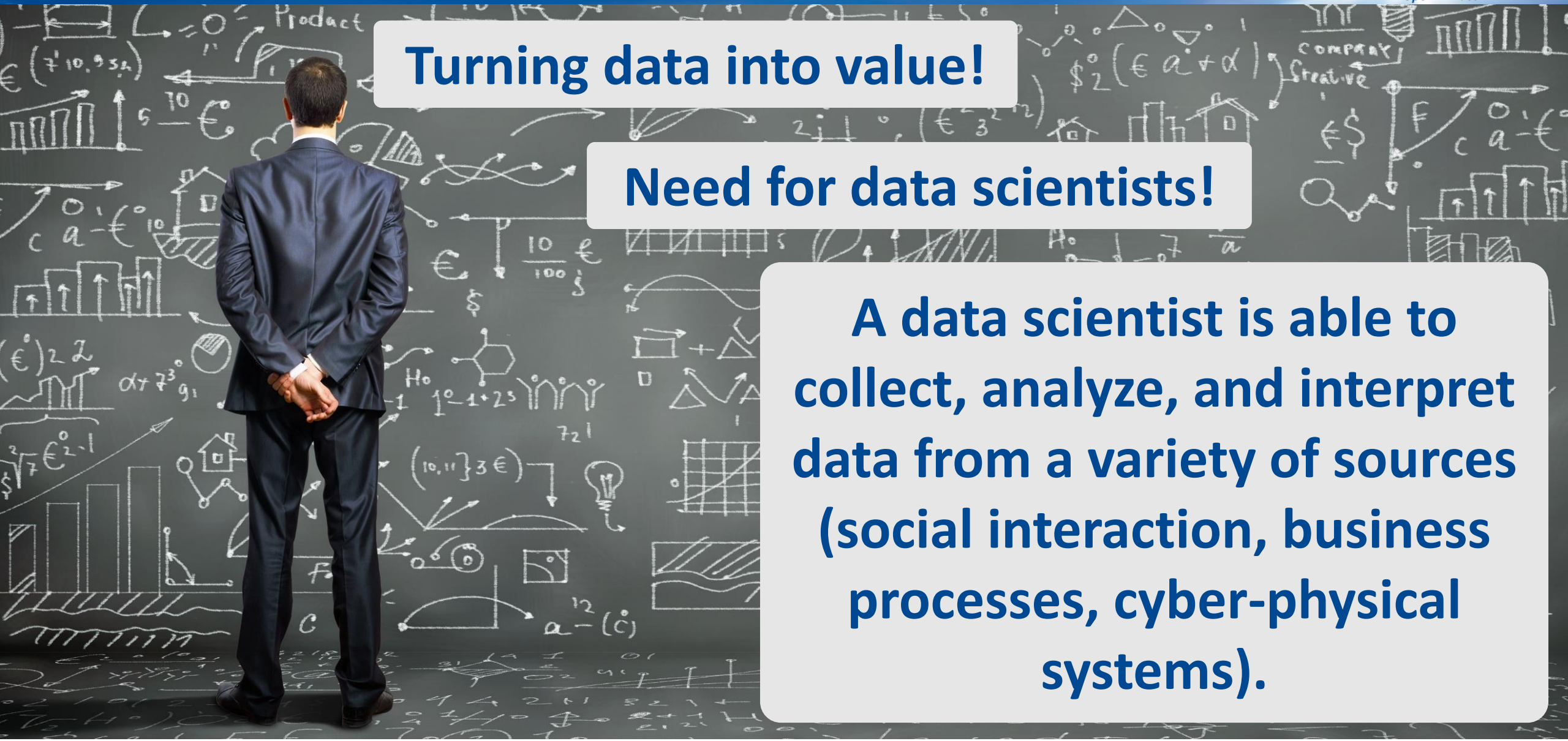
Data does not have to be “Big” to be challenging!

How to extract real value from data?

Data analytics questions are everywhere!

Data Science Questions

Data Scientists

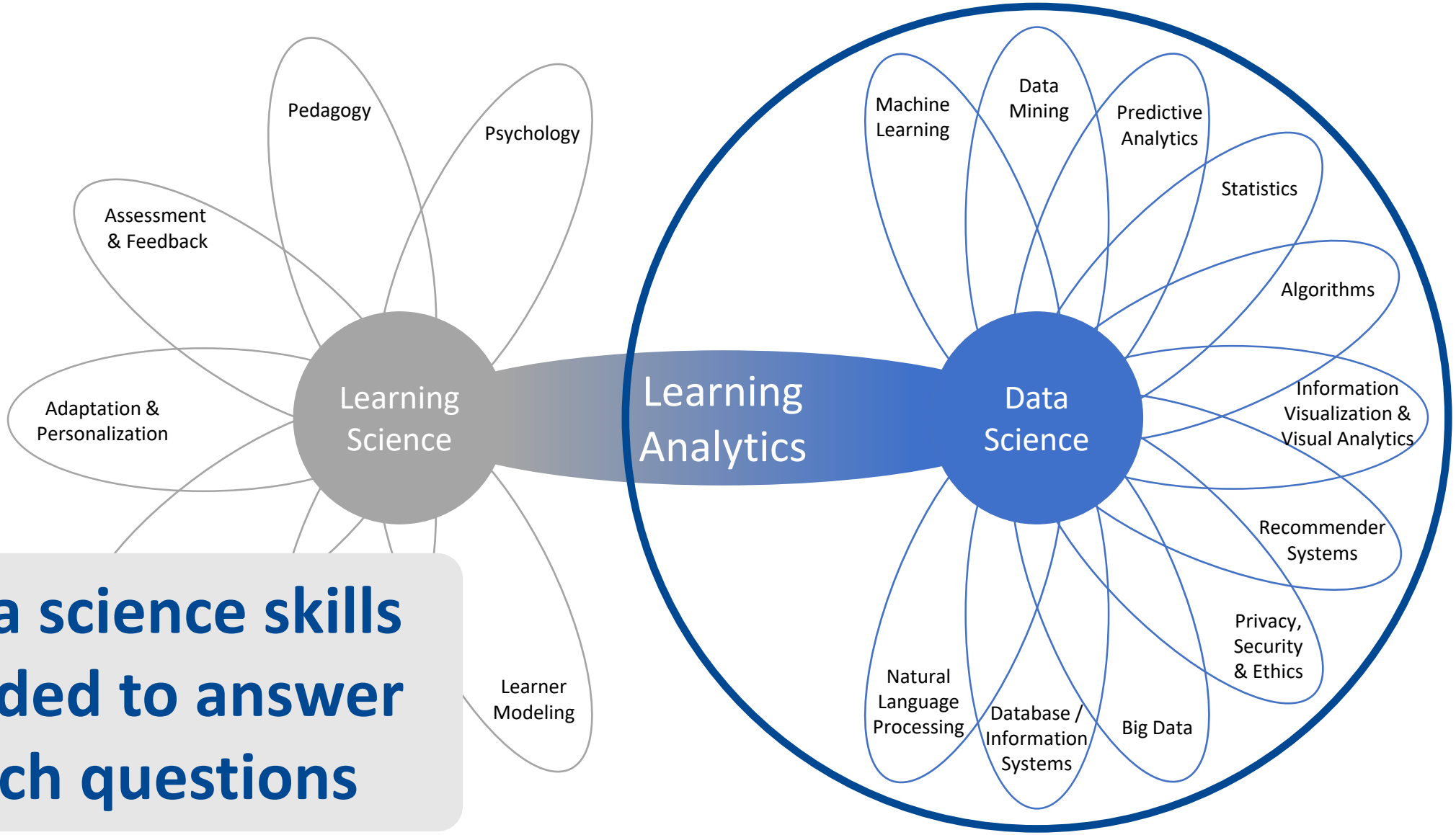


Turning data into value!

Need for data scientists!

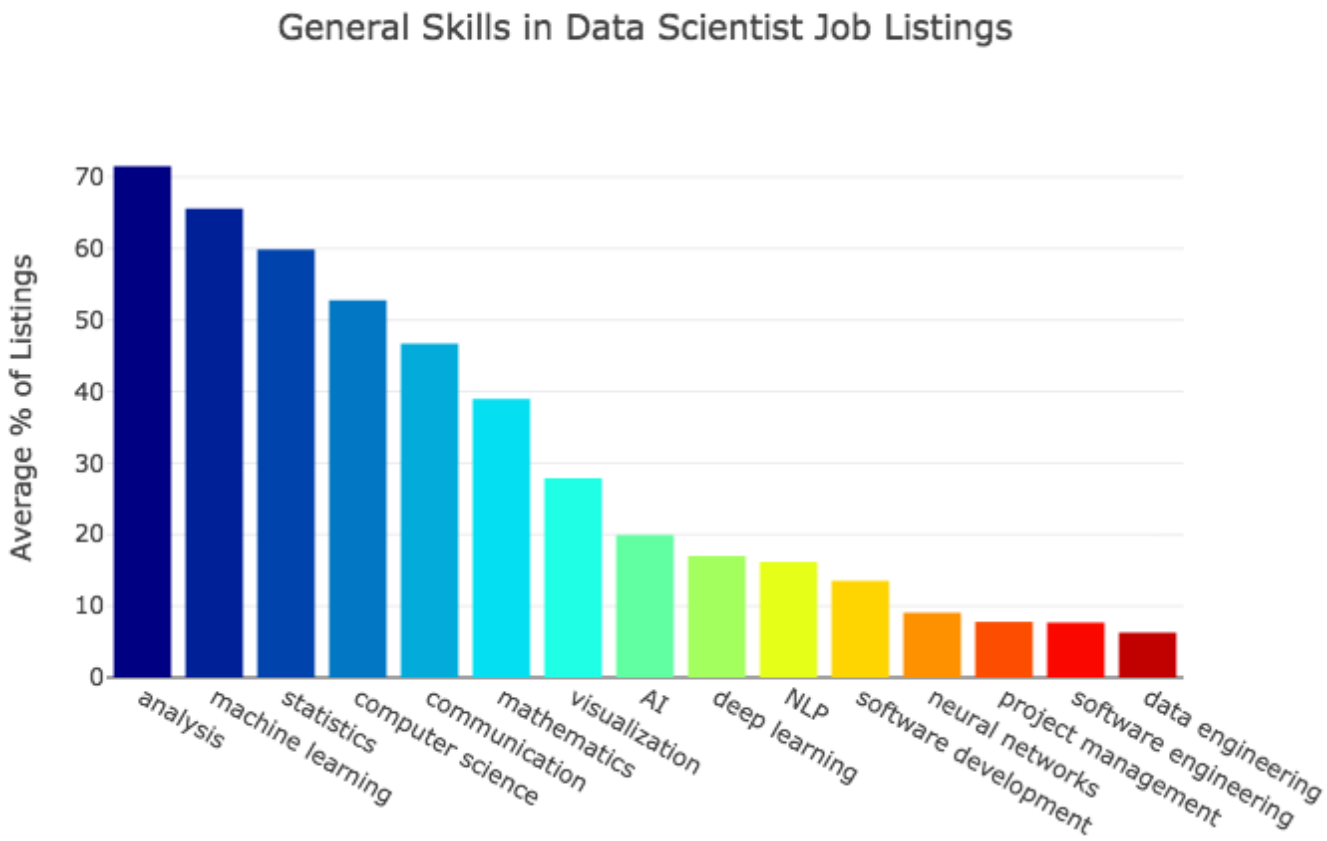
A data scientist is able to collect, analyze, and interpret data from a variety of sources (social interaction, business processes, cyber-physical systems).

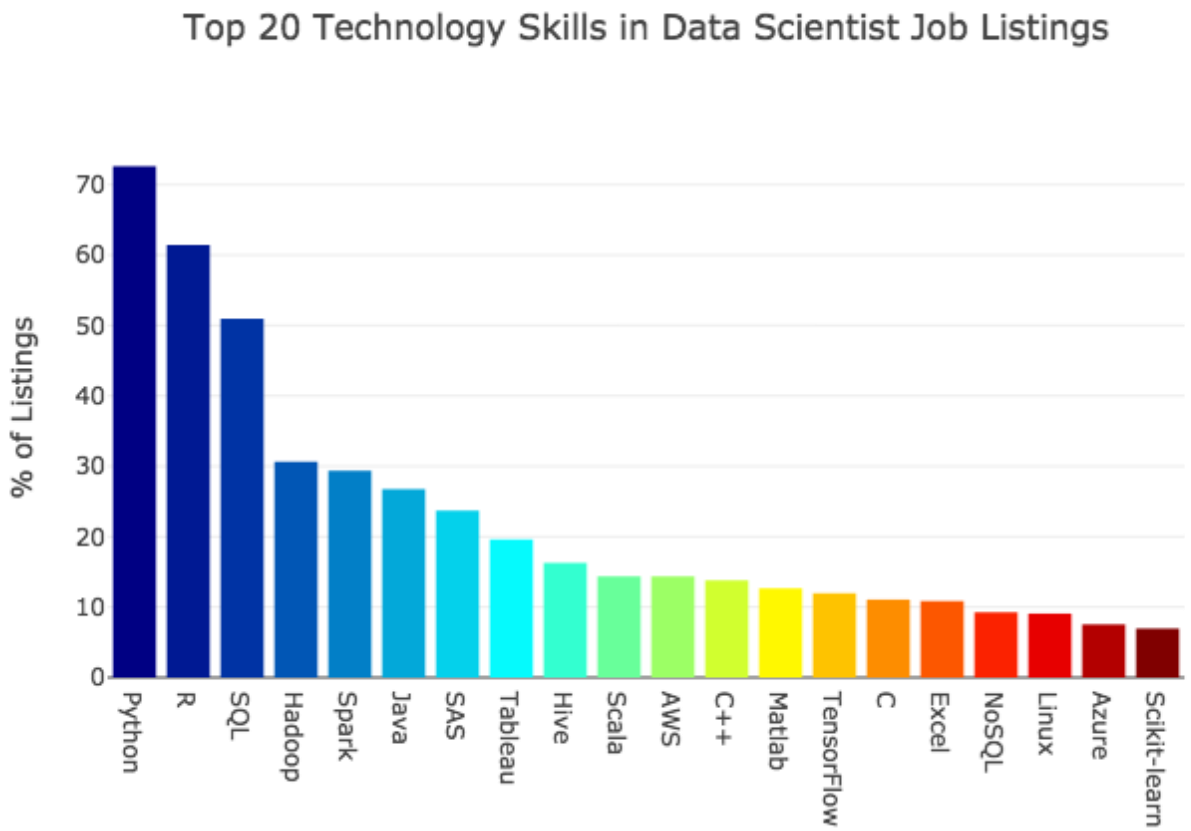
Data Science Skills



Data science skills needed to answer such questions

Data Science Skills





[https://towardsdatascience.com/the-most-in-demand-skills-for-data-scientists-4a4a8db896db]

Big Data

Who is generating Big Data?

The model of generating/consuming data has changed

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

Which sectors are generating Big Data?

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Social Media

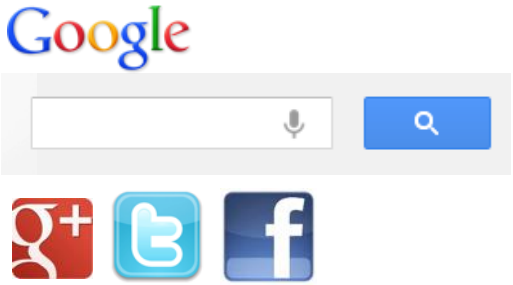
User Tracking & Engagement

Learning / Education

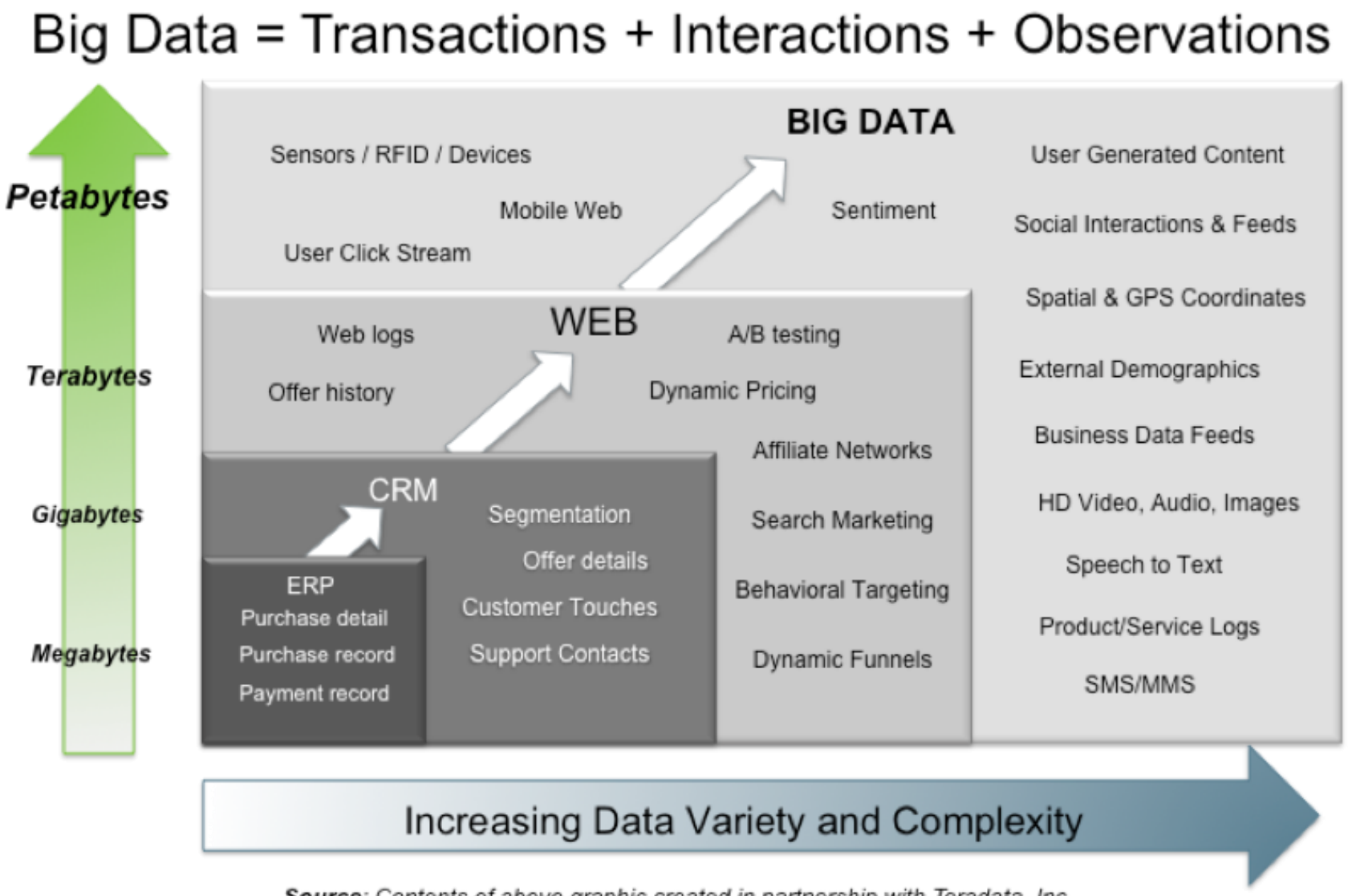
eCommerce

Financial Services

Web / Real Time Search



What kind of Big Data is generated?



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

How big is big?

- 2008: Google processes 20 PB a day
- 2009: Facebook has 2.5 PB user data + 15 TB /day
- 2011: Yahoo! has 180-200 PB of data
- 2012: Facebook has 500 TB/day
- 2013: YouTube has 1000 PB video storage; 4 billion views/day
- ...
- 2018: Facebook stores 300 PB of data; 600 TB/day
- Accumulated digital universe of data today is 4.4 ZB

Big Data – Definitions

- Big date is data that is too large or too complex to be managed using traditional data processing, analysis, and storage techniques (Wikipedia)

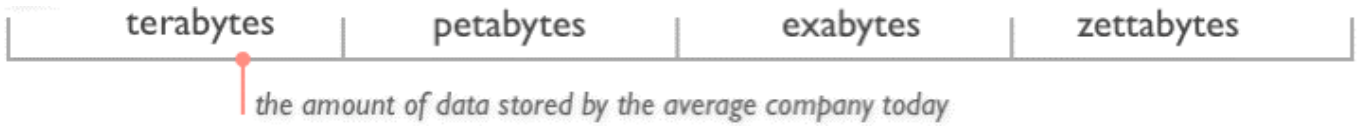
- Big data is **high-volume**, **high-velocity** and **high-variety** information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (Gartner)

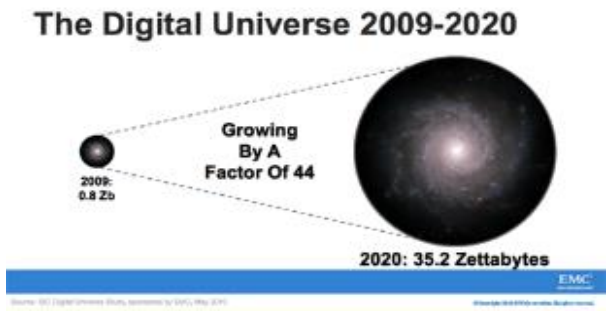
- Big Data consists of extensive datasets, primarily in the characteristics of **volume**, **velocity**, and/or **variety** that require a scalable architecture for efficient storage, manipulation, and analysis (National Institute of Standards and Technology)

Big Data: 3V's

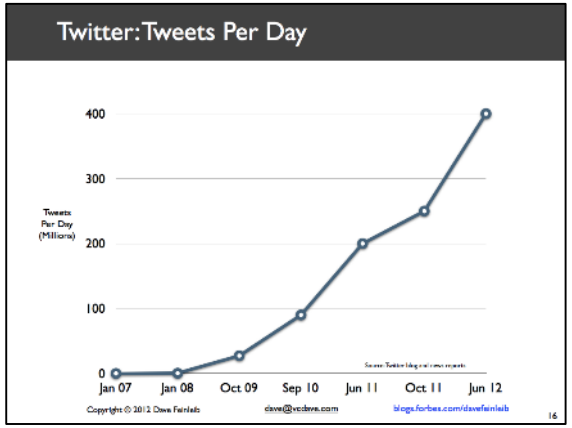
Volume (Scale)

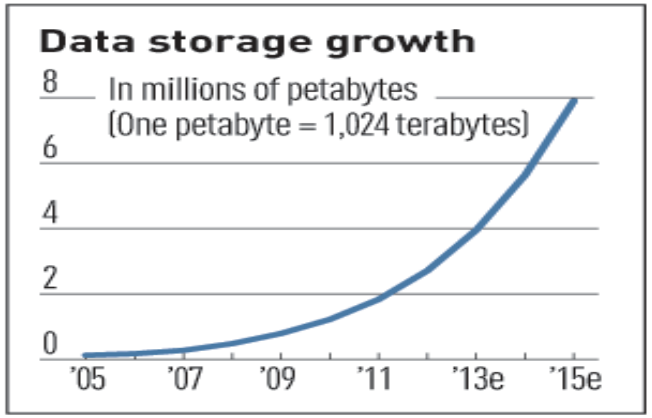
- Data volume
  - 90% of all data has been generated in just the last 2 years
  - 44x increase from 2009 to 2020
  - From 0.8 zettabytes to 35zb





- Data volume is increasing exponentially

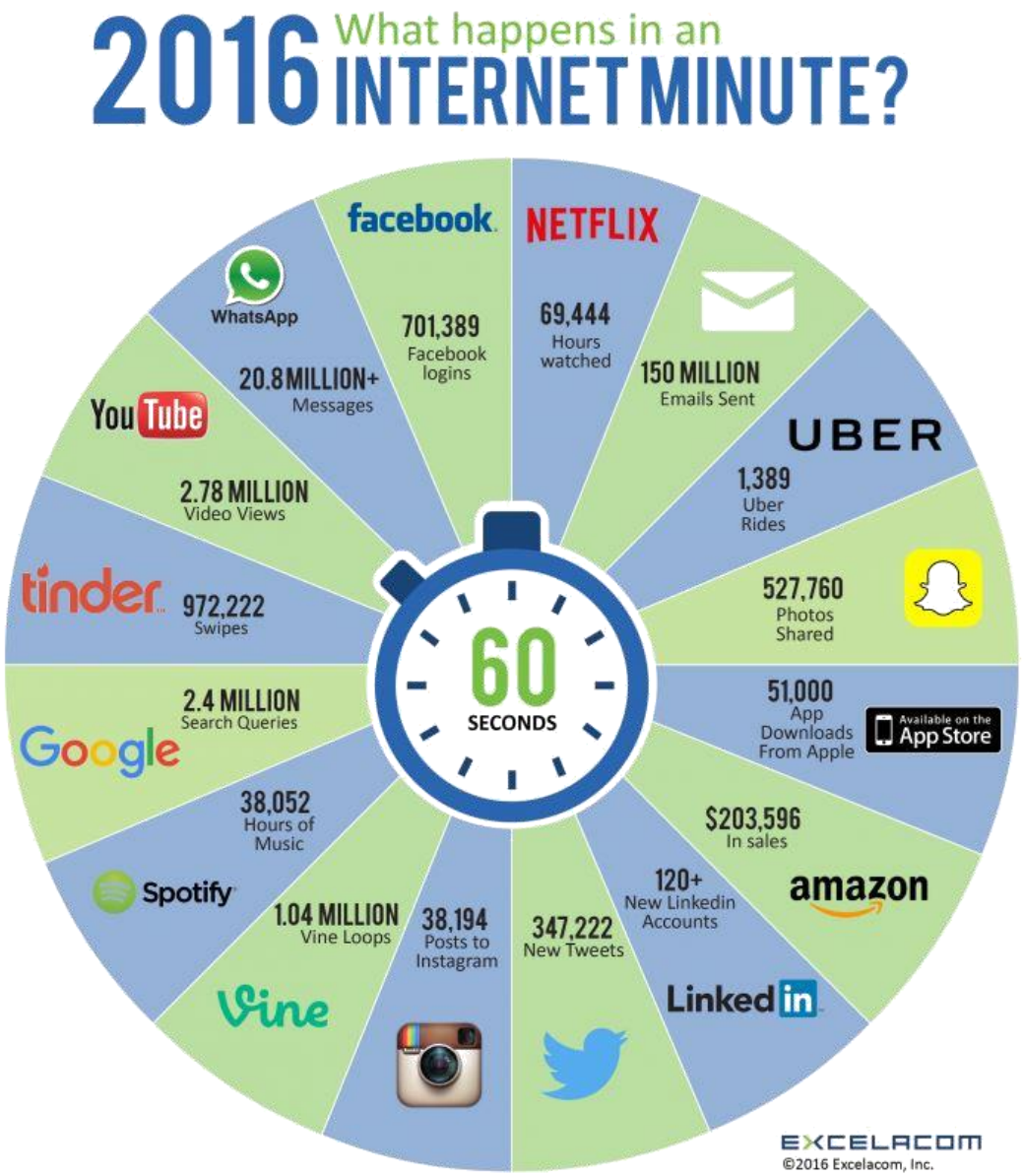




Velocity (Speed)

- The speed at which the data is generated and used

- "From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days … and the pace is accelerating" (Eric Schmidt, Google)



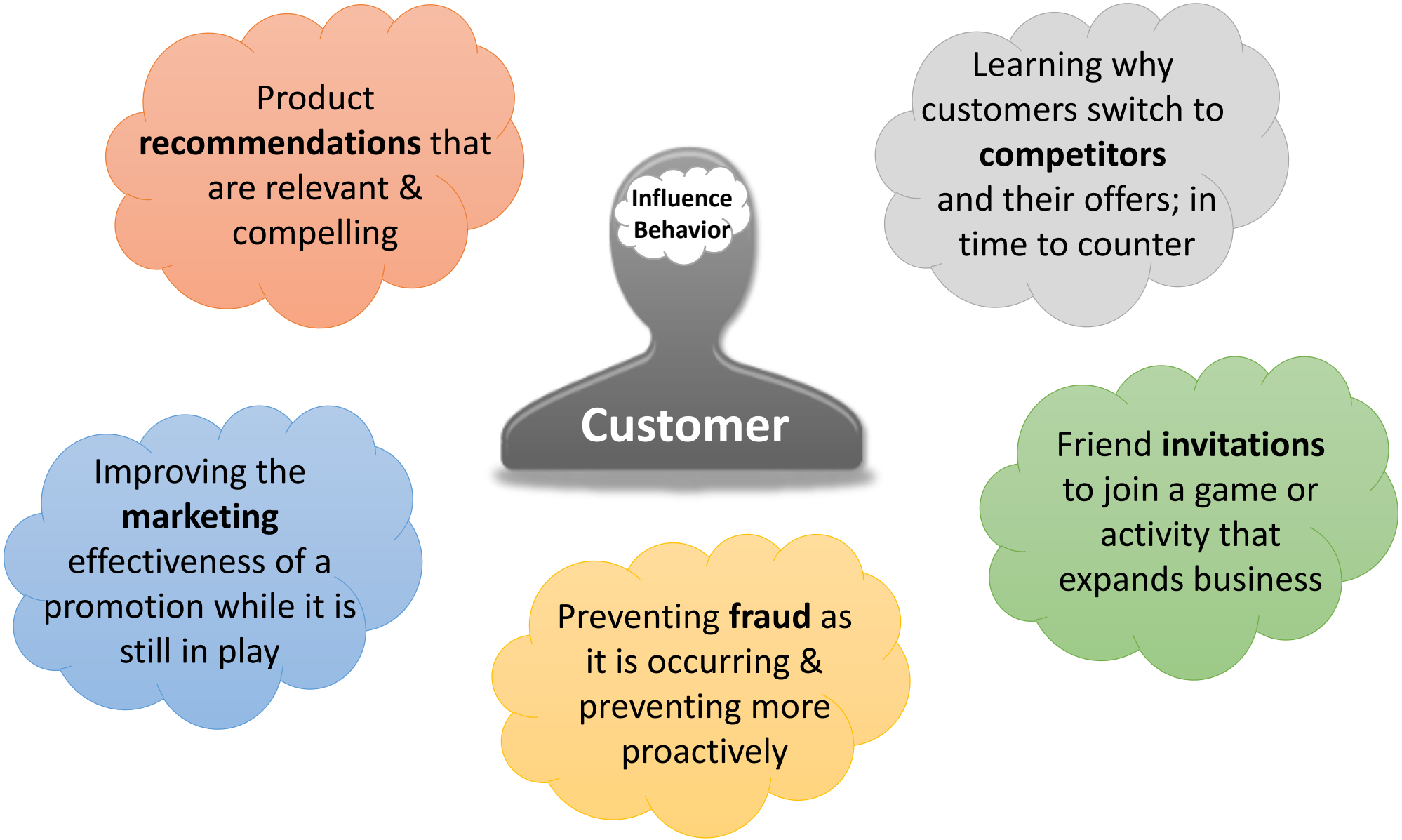
Velocity (Speed)

- Data is being generated fast and needs to be **processed fast**
- Late decisions ➡ missing opportunities

- Examples
  - E-Promotions: Based on your current location, your purchase history, what you like
  ➡ send promotions **right now** for store next to you

  - Healthcare monitoring: Sensors monitoring your activities and body
  ➡ any abnormal measurements require **immediate reaction**

Real-Time Analytics / Decision Requirement



Real-time / Fast Data

Social media and networks
(all of us are generating data)

Mobile devices
(tracking all objects all the time)

Scientific instruments
(collecting all sorts of data)

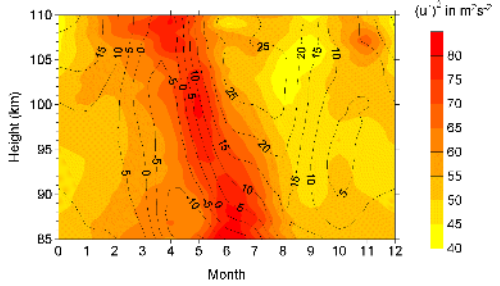
Sensor technology and networks
(measuring all kinds of data)

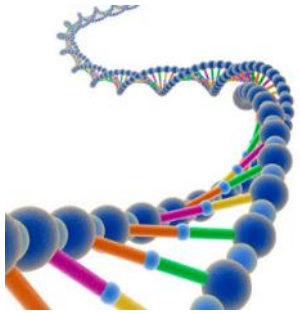
- The **progress** and **innovation** is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and **discover knowledge** from the collected data in a **timely** manner and in a scalable fashion

Variety (Different applications generating data)

Variety (Complexity)

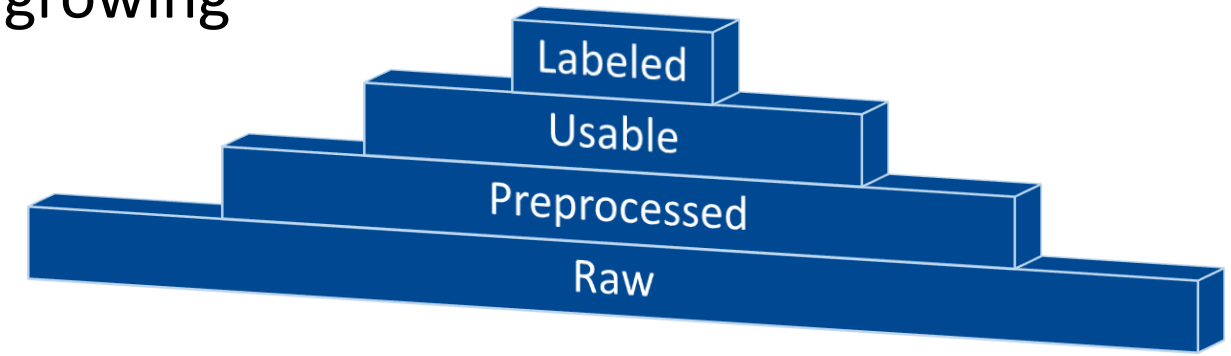
- A single application can generate/collect many types of data
  - **Structured data**: Relational data (Tables/Transaction)
  - **Unstructured data**: Text files, audio, video, web pages, computer programs, social media etc.
  - **Semi-structured data**: XML, Semantic Web RDF





- To extract knowledge, all these types of data need to linked together

- ~ 90% of all data is unstructured and it is growing faster than structured data



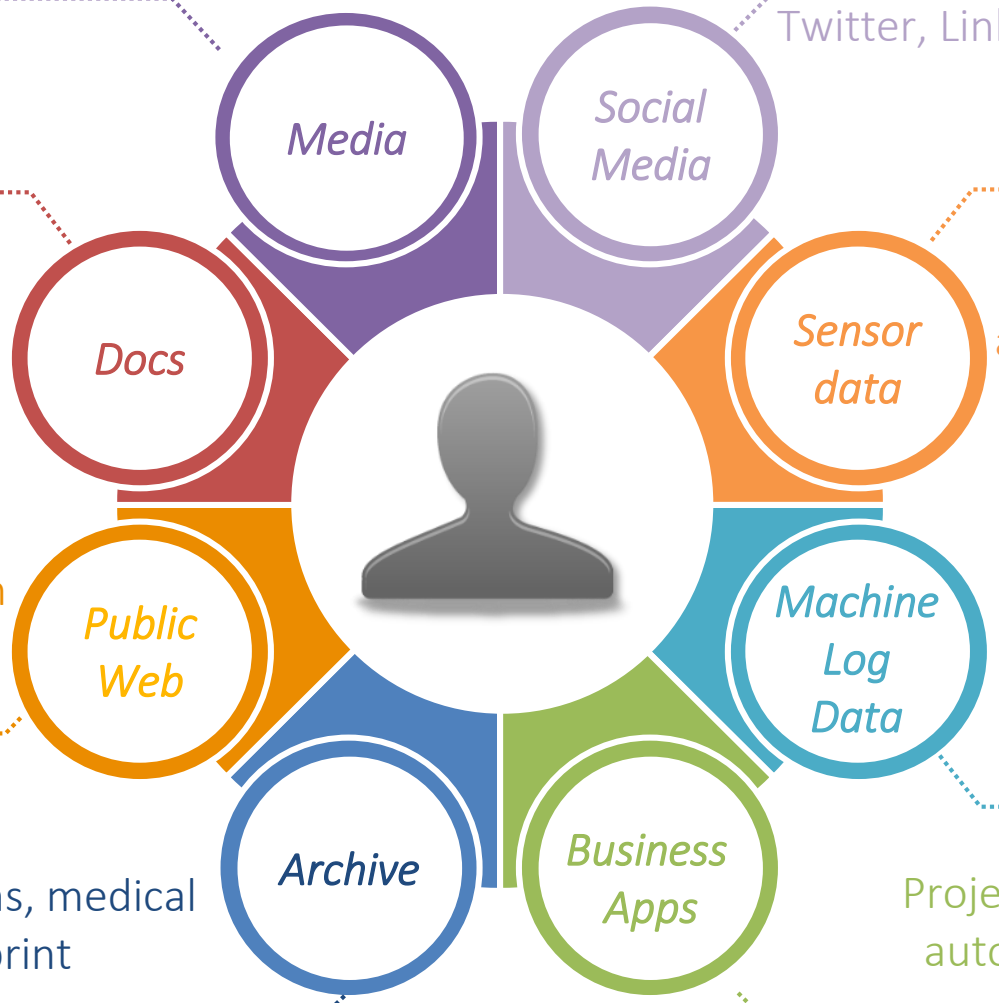
Variety (Different types of data)

Images, videos, audio, live streams, podcasts, etc.

Twitter, Linkedin, Facebook, SlideShare, YouTube, Instagram, Flickr, Pinterest, RSS, etc.

PDF, CSV, DOCX, PPTX, HTML, TXT, XML, XLS, JSON, etc.

medical devices, car sensors, satellites, video games, cell towers, assembly lines, smart farm machinery, etc.



Government, weather, traffic, health care services, economic, Wikipedia, etc.

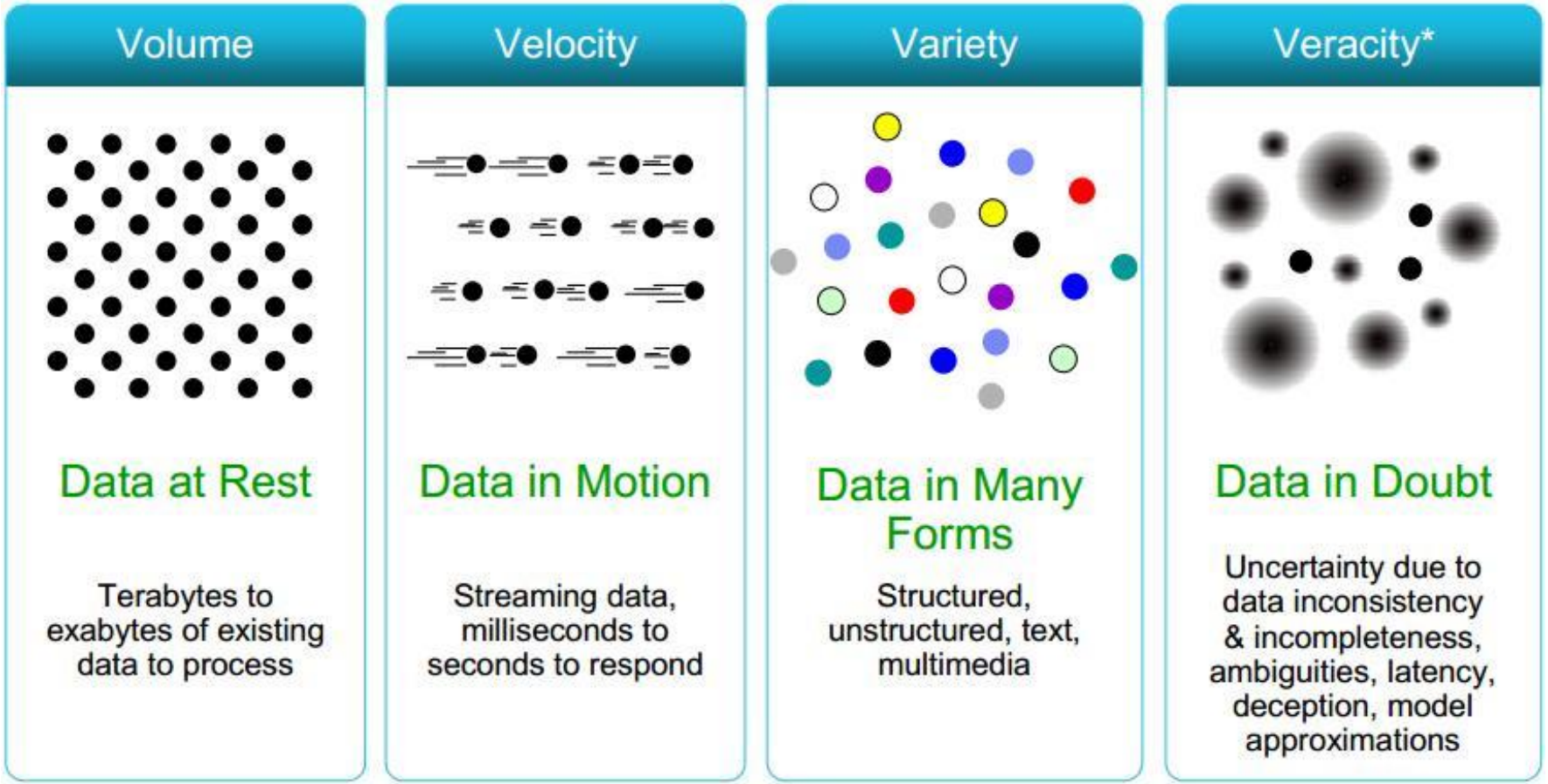
Clickstream data, server logs, business process logs, audit logs, mobile app usage, etc.

Scanned documents, insurance forms, medical record, customer correspondence, print stream files, etc.

Project management, CRM, ERP, marketing automation, talent management, expense management, intranets, portals, etc.

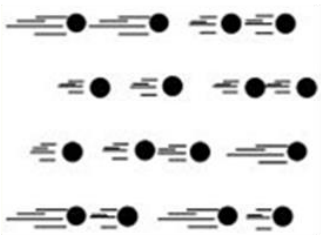
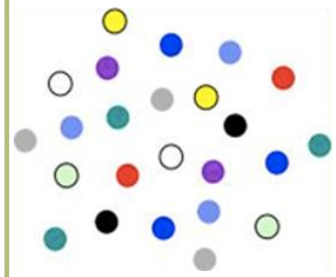
[adapted from Big Data Analytics, PwC]

Some make it 4V's



[https://www.datasciencecentral.com/profiles/blogs/data-veracity]

Or even 5V's also 7V's

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	**Streaming data, requiring milliseconds to seconds to respond**	**Structured, unstructured, text, multimedia,…**	**Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations**	**Business models can be associated to the data**

[http://informationcatalyst.com/vision-experience/big-data-value/]

Learning Data as Big Data

big data
petabytes
large
internet
search
sets
new
storage
analysis
research
software
science
san
future
million
information
technologies
management
nas
government
based
social
development
world
economic
exabytes
critiques
mapreduce
national
billion
parallel
real
every
times
years
past
used
volume
size
time
process
learning
amount
initiative
year
business
databases
institute
processing
set
now
massive
scientific
terabytes
handle
sensor
rate
well
across
state
queries
analytics
approaches
many
processes
systems
results
announced
applications
work
database
factor
influence
accuracy
replication
trends
test
per
annual
binding
crime
per
cost
data
use
challenges
between
shared
day
factor
time
networks
terms
tools
framework
even
decision
worldwide
project
attractive
technology
high
less
now
work
text
handle
applications
announced
scientific
terabytes
sensor
rate
well
across
state
queries
mapreduce
billion
national
parallel
real
every
times
years
past
social
development
world
economic
exabytes
critiques

Big Data Storage

Handling Data Volume

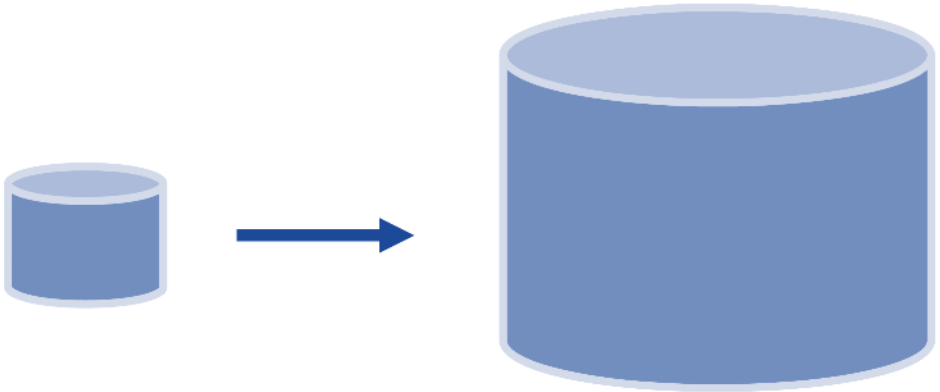
- **Small data**: (data fits into the main memory)
  - File system: csv-files, excel files
  - Read everything from file into memory
  - Manipulate data in memory (e.g. excel, python)

- **Medium data**: (data fits on machine but not into memory)
  - Database systems, files
  - Read only necessary part of the data (replace data in memory)
  - Manipulate data on disk (e.g. SQL queries, temporary views)

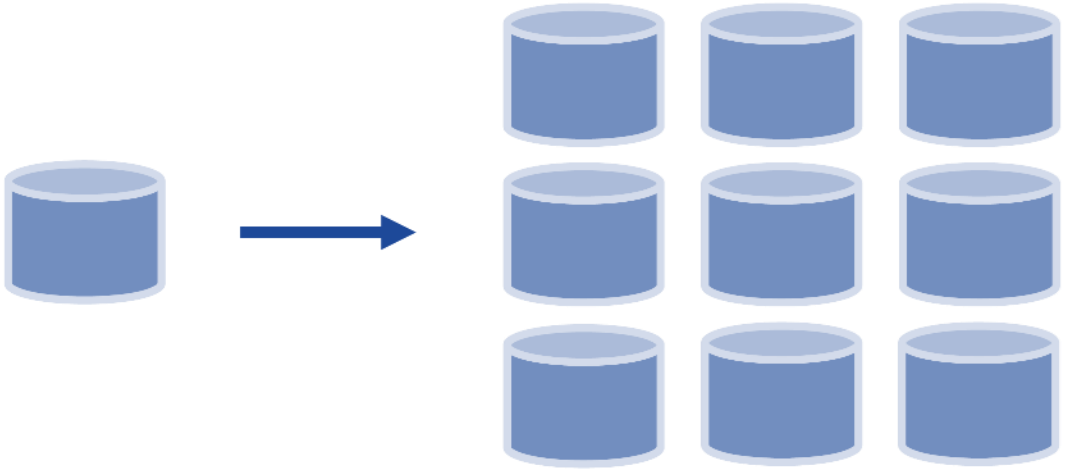
- **Big data**: (data does not fit on one machine)
  - Distributed file systems (e.g. HBase, HDFS), NoSQL databases
  - Manipulate data using cloud framework (e.g. Map reduce, Spark)

Approaches to scale up storage systems

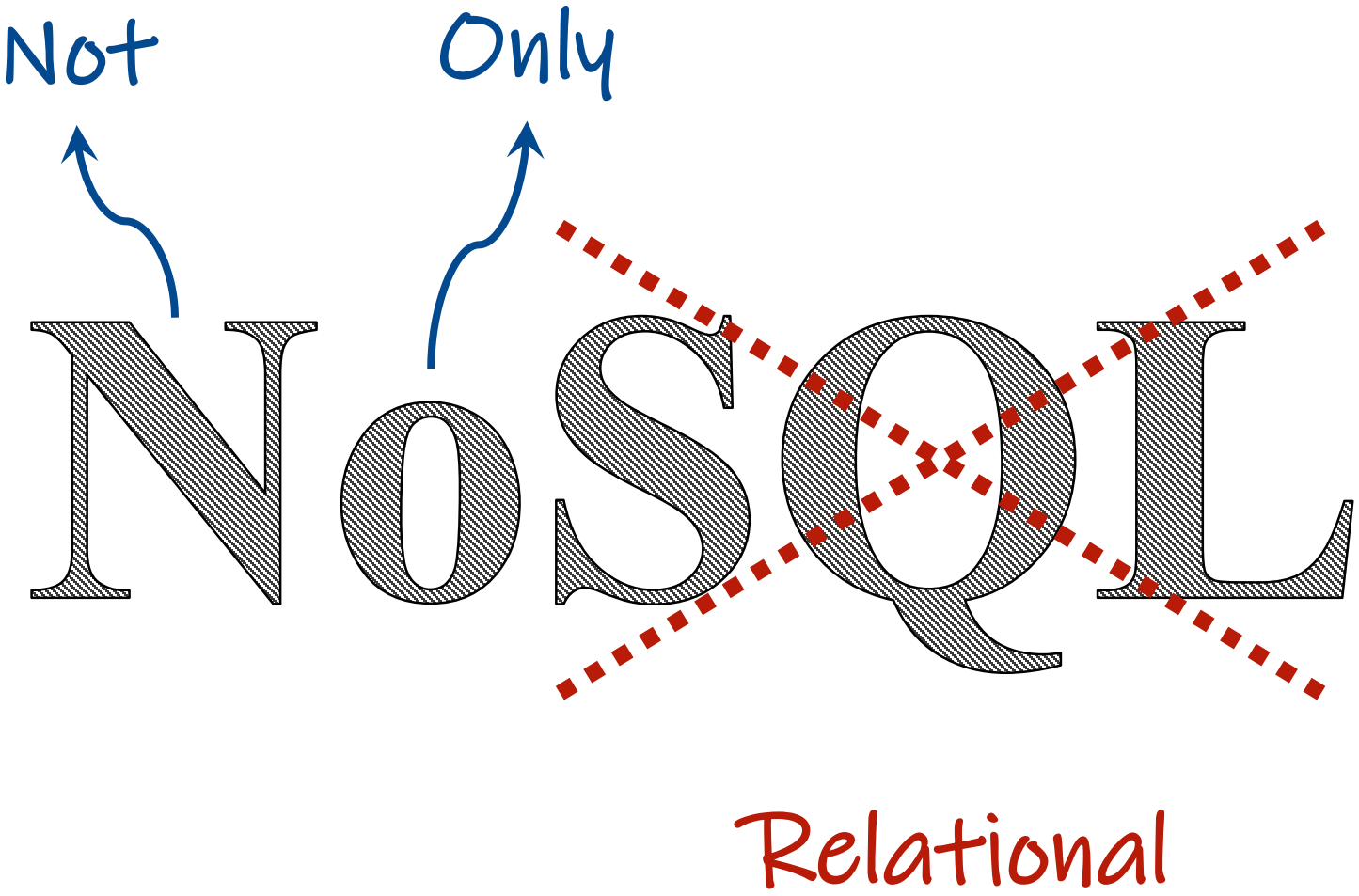
- Vertical scaling
  - Enlarge a single machine
  - – Limited in space
  - – Expensive



- Horizontal scaling
  - Use many commodity machines and form *computer clusters* or *grids*
  - – Cluster maintenance



NoSQL Databases



NoSQL vs. Relational Databases (1/2)

	NoSQL	Relational
Scaling	Horizontal scaling	Vertical scaling
Storage Mechanism	Distributed storage across multiple nodes/locations	Single seamless store of data mostly single machine/location
Data Size	Unlimited	Limited depending on selected RDBMS
Data Structure	Supports both structured and unstructured data	Supports only structured data
Data Formats	Supports various data formats (e.g. text, JSON, XML, etc)	Supports a single tabular data format
Query Language	Per-product definition and manipulation language	Standard definition and manipulation language

NoSQL vs. Relational Databases (2/2)

	NoSQL	Relational
Schema	Schema on Read: Data is simply stored, no transformation	Schema on Write: Schema must be created before data can be loaded
Transformation	Applied during read time to extract the required data	Explicit load operation has to take place which transforms data to DB internal structure
Store Process	1. In a high velocity use case, the data is prepared and analyzed for alerting, and only then is it stored 2. In a volume use case, the data is often stored in a raw state in which it was produced	Data is stored after preparation (for example after the extract-transform-load and cleansing processes)
Characteristics	Flexibility	Maturity, stability, efficiency

NoSQL – History

- Mid 2000's: Birth of the NoSQL Movement
  - Problem of computer clusters:
    Relational databases do not scale well horizontally
  - Big Players like Google or Amazon developed their own storage systems: NoSQL databases were born
- Today: Age of NoSQL
  - Several different NoSQL systems available (>225)

NoSQL – Characteristics

- There is no unique definition but some characteristics for NoSQL Databases:
  - Horizontal scalability (cluster-friendliness)
  - Non-relational
  - Distributed
  - Schema-less
  - Open-source (at least most of the systems)

NoSQL – Concepts

- **ACID** – The holy grail of RDBMSs:
  - ***A**tomicity*: All or nothing - Either all of the actions in a transaction are completed (i.e. transaction is committed) or none of them are completed (i.e. transaction is rolled back)
  - ***C**onsistency*: Any transaction brings the database from one valid state to another (Valid = does not violate any predefined rules e.g. constraints)
  - ***I**solation*: Concurrent execution of transactions results in a system state that would be obtained if transactions were executed serially
  - ***D**urability*: Once a transaction has been committed to the database, it stays there forever

NoSQL – Concepts

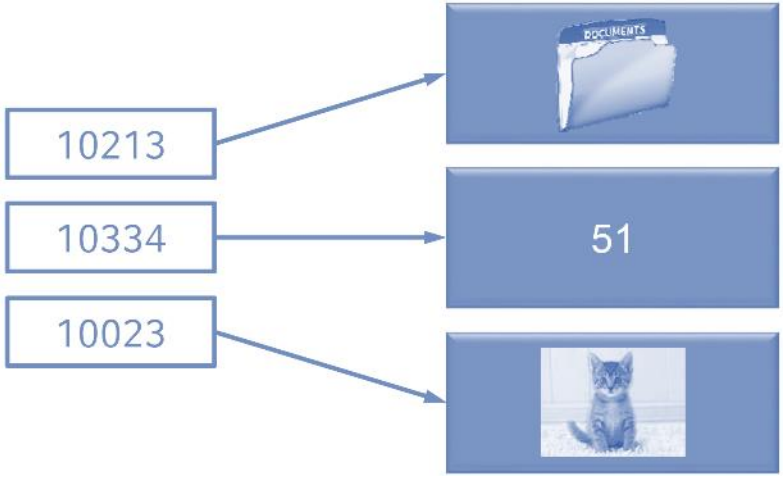
- **BASE** – An artificial concept for NoSQL databases:
  - ***B**asically **A**vailable*: The system is generally available, but some data might not at any time (e.g. due to node failures)
  - ***S**oft State*: The system's state changes over time. Old data may expire if not refreshed
  - ***E**ventual consistency*: The system is consistent from time to time, but not always. Updates are propagated through the system if there is enough time
- BASE is settled on the opposite site to ACID when considering a "consistency-availability spectrum"
  - BASE transactions allow a database to be in a temporarily inconsistent state that will eventually be resolved

NoSQL – Storage Types

- There are many NoSQL solutions, these are widely classified as:
  - Key-Value Store
  - Column Store
  - Document Store
  - Graph Store

Key-Value Store

- Most simple form of database systems
- Store key/value pairs and retrieve values by keys
- Values can be of arbitrary format



Column Store

- A database consists of multiple tables
- Each table consists of multiple rows, sorted by row key
- Each row contains a row key and one or more column families
- Each column family is defined when the table is created
- Column families can contain multiple columns (family:column)
- A cell is uniquely identified by (table,row,family:column)
- Data in each column is stored together (vs. row storage in RDBS)

	Name		Home		Office	
Key	First	Last	Phone	Email	Phone	Email
101	Florian	Krepsbach	555-1212	florian@ wobegon.org	666-1212	fk@phc.com
102	Marilyn	Tollerud	555-1213		666-1213	
103	Pastor	Inqvist			555-1214	inqvist@ wels.org

HBase – Data Model

	Name		Home		Office	
Key	First	Last	Phone	Email	Phone	Email
101	Florian	Krepsbach	555-1212	florian@ wobegon.org	666-1212	fk@phc.com
102	Marilyn	Tollerud	555-1213		666-1213	
103	Pastor	Inqvist			555-1214	inqvist@ wels.org

	Name			Home		Office		Social
Key	First		Last	Phone	Email	Phone	Email	FacebookID
101	Florian	Garfield	Krepsbach	555-1212	florian@ wobegon.org	666-1212	fk@phc.com	
102	Marilyn		Tollerud	555-1213		666-1213		
103	Pastor		Inqvist			555-1214	inqvist@ wels.org	

New columns can be added at runtime
(Good for data with high variety -> simply add columns)

Column families cannot be added at runtime

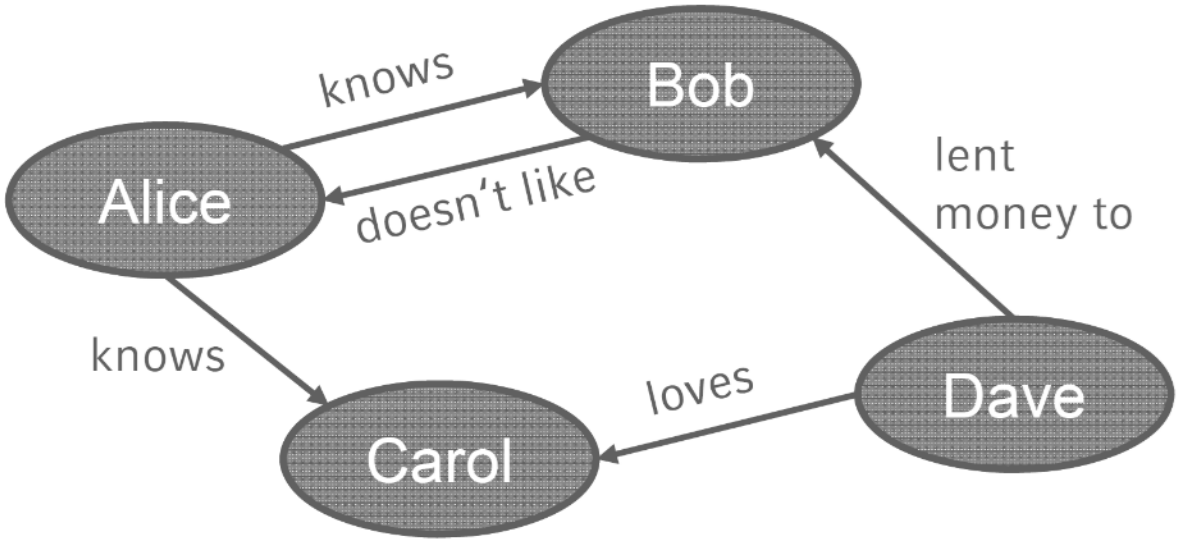
Document Store

- Store documents in form of XML or JSON
- Each document is assigned a unique key (used to retrieve the document)
- Semi-structured data records that do not have a homogeneous structure
- Data records can have more than one value (arrays) !
- Documents include internal structure, or metadata
- Data structure enables efficient use of indexes

```
<contact>
  <first_name>Max</first_name>
  <last_name>Mustermann</last_name>
  <street>Musterstraße 12</street>
  <city>Musterstadt</city>
  <zip>12345</zip>
  <country>D</country>
</contact>
```

Graph Store

- Use graphs to store and represent relationships between entities
- Composed of nodes and edges
- Each node and each edge can contain properties (Property-Graphs)



NoSQL – Storage Types (Summary)

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Increasing Data Complexity

Key – Value Store	Column Store	Document Store	Graph Store
Store key/value pairs and retrieve values by keys ***Pros***: Simplicity & Scalability ***Cons***: Lack of advanced features/queries	Store data tables as sections of columns of data, rather than as rows of data ***Pros***: Scalability & Flexibility ***Cons***: Complexity	Store document in form of XML or JSON ***Pros***: Simplicity & High data nesting ***Cons***: Scalability	Store data whose relations are well represented as a graph ***Pros***: Graph Joins ***Cons***: Flexibility
amazon web services Dynamo DB, riakKV, redis, Memcached	cassandra, druid, Apache HBASE, VERTICA, HYPERTABLE, Cloud Bigtable	terrastore, CouchDB relax, mongoDB, RAVENDB Safe by Default, Optimized for Efficiency	OrientDB, AllegroGraph Franz Inc., HyperGraphDB, sones, neo4j, OPENLINK Virtuoso, INFINITEGRAPH

Big Data Analytics

Big Data Analytics – Definitions

- Big data analytics refers to the process of collecting, organizing and analysing large sets of data to **discover patterns** and other useful information (Webopedia) [https://www.webopedia.com/TERM/B/big_data_analytics.html]

- Big data analytics examines large amounts of data to **uncover hidden patterns**, **correlations** and **other insights** (SAS) [https://www.sas.com/en_us/insights/analytics/big-data-analytics.html]

- Big data analytics is the process of examining large and varied data sets to **uncover hidden patterns**, **unknown correlations**, **market trends**, **customer preferences** and other useful information that can help organizations make more-informed business decisions (TechTarget) [https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics]

It's not just about the data…

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

Big ***Data*** + *Big* ***Data Analytics***

Refers to the DATA only

Methods of using Big Data to generate insight

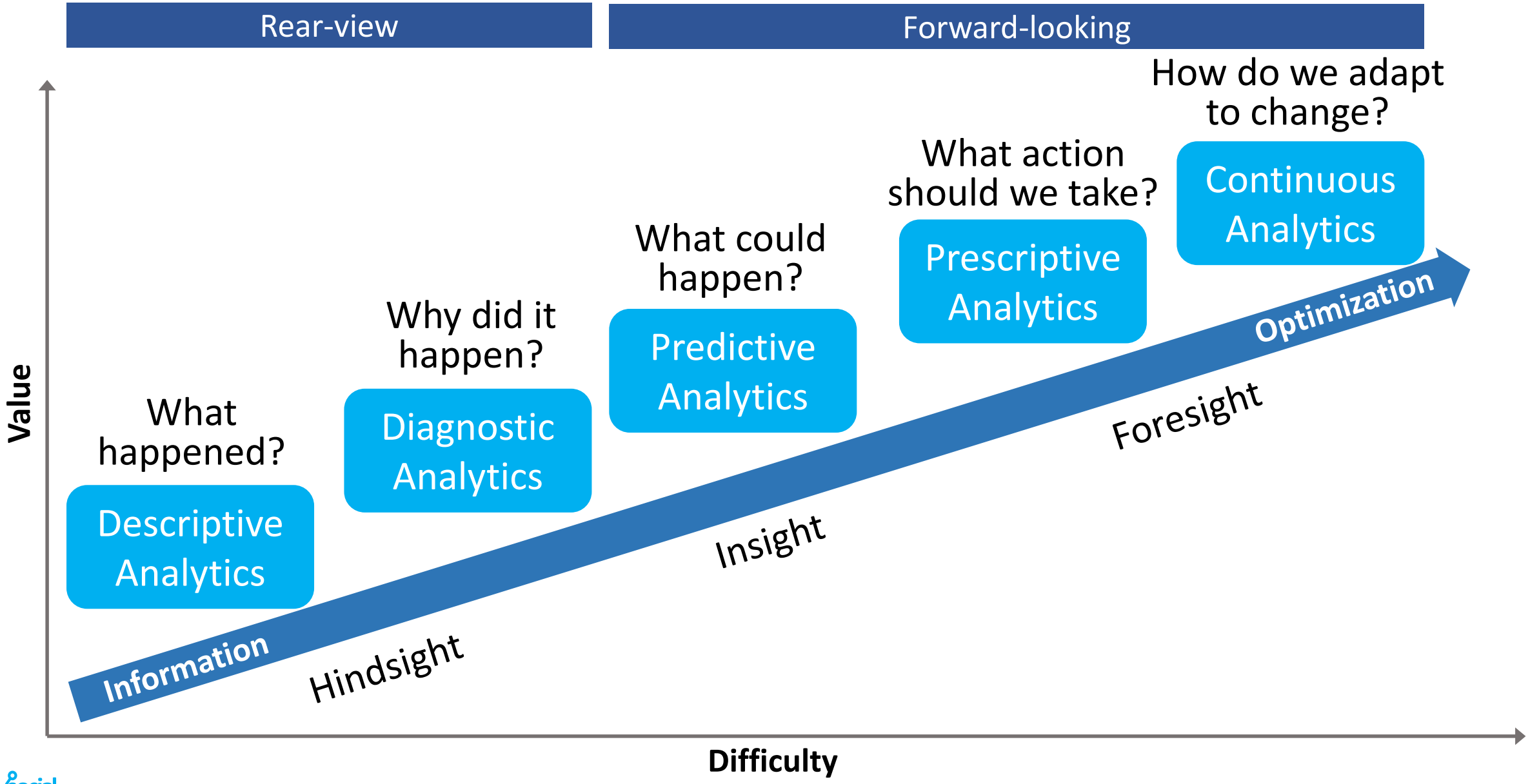
Structured → Unstructured

Traditional → Emerging

Regression *Discover relationships between variables*	***Time Series Analysis*** *Discover relationships over time*	***Signal Analysis*** *Distinguish between noise and meaningful information*	***Cluster Analysis*** *Discover meaningful groupings of data points*
A/B/N Testing *Experiment to find the most effective variation of a website, product, etc*	***Classification*** *Organize data points into known categories*	***Simulation Modeling*** *Experiment with a system virtually*	***Spatial Analysis*** *Extract geographic or topological information*
Visualization *Use visual representations of data to find and communicate info*	***Predictive Modeling*** *Use data to forecast or infer behavior*	***Complex Event Processing*** *Combine data sources to recognize events*	***Sentiment Analysis*** *Extract consumer reactions based on social media behavior*
Network Analysis *Discover meaningful nodes and relationships on networks*	***Optimization*** *Improve a process or function based on criteria*	***Deep QA*** *Find answers to human questions using artificial intelligence*	***Natural Language Processing*** *Extract meaning from human speech or writing*

[Slide Credit: Big Data Analytics, PwC]

Types of Big Data Analytics



BIG DATA & AI LANDSCAPE 2018

Big Data – Summary

- Big Data
  - **3 V's** (Volume, Velocity, Variety)
  - **5 V's** (Volume, Velocity, Variety, Veracity, Value)
  - Big Data + Big Data Analytics

- Big Data Storage
  - ACID vs. BASE (Consistency-Availability spectrum)
  - **NoSQL** (Key-Value, Column, Document, Graph Store)

- Big Data Analytics
  - Big data methods to generate insights (Turning data into **value**!)
  - Types (Descriptive, Diagnostic, predictive, Prescriptive, Continuous Analytics)