F2xBD: Big Data Management: Introduction

Dr. Radu Mihailescu

Associate Professor

Programme Director MSc AI

Agenda for today

1. Introduction to Big Data Management & Admin
2. Semantic Web & Knowledge Graphs
3. Semantic Web Technologies

Course content – Part 1 (Dr. Radu Mihailescu)

Weeks 1-5:

- Big data
- Semantic web
- Linked data
- Knowledge graphs
- Knowledge representation: Ontologies, inference – RDFS, OWL
- Knowledge retrieval: SPARQL, triplestores

Course content – Part 2 (Dr. Drishty Sobnath)

Weeks 7-11:

- NoSQL data models
  - Graph model (neo4j, Cypher QL)
  - Key-value model (Redis)
  - Document model (MongoDB, JavaScript API)
  - Wide-column (Cassandra, CQL)
- Data integration

Tools needed

- Protégé: ontology editor with OWL, RDFS..., SPARQL querying
- Fuseki JENA: triplestore, SPARQL querying
- H2 or MYSQL: data integration
- A few more tools online (viewer, validator…)
- neo4j: graph database
- MongoDB: document store
- Cassandra: wide- column store
- PostgreSQL + PostGIS + QGis

Prior knowledge

- Extensive knowledge of Relational Database Systems
  - Relational model and algebra
  - SQL
  - ACID Transactions
- First order logic
  - Conjunction, disjunction, and negation
- Familiarity with distributed system concepts
  - Messaging issues
  - Client-server
  - Peer-to-peer

Expectations

- Technical course
- Hours (per week):
  - 2 hours lecture (selected topics)
  - 1 hour lab/tutorial
  - 7-10 hours private study

Assessment

- Class Test 20% - Week 7 (Thursday - TBC)
  - Covers teaching material Part 1 (weeks 1-4 - TBC)

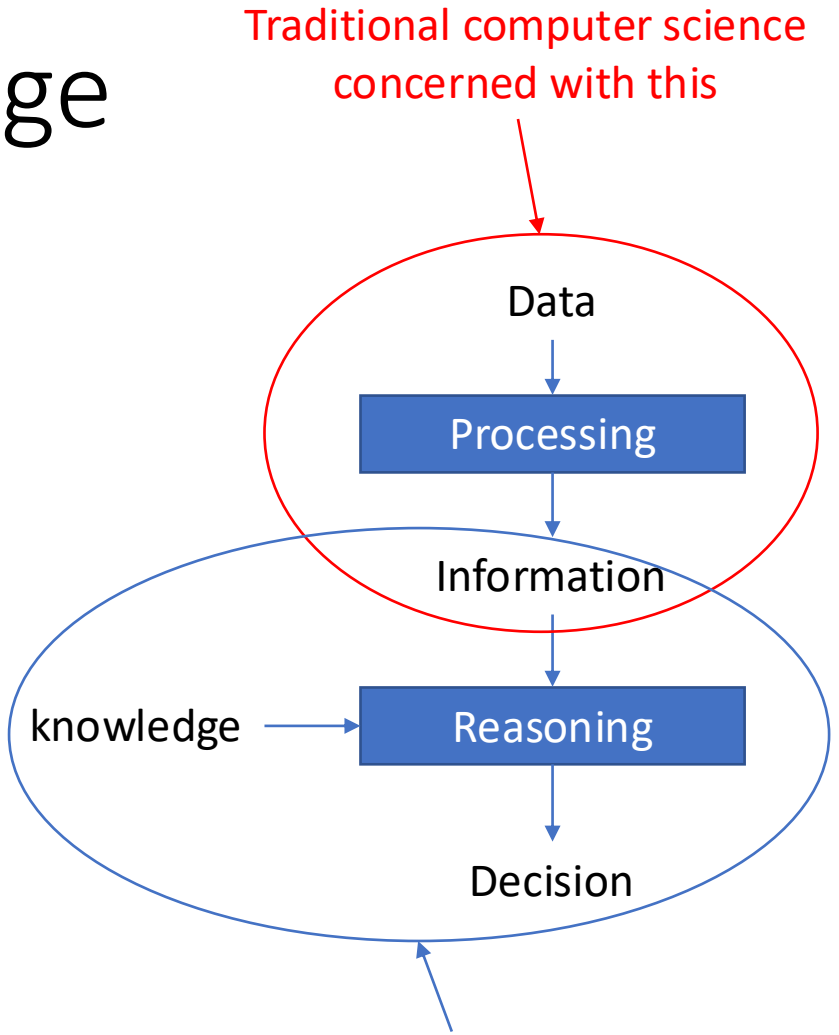
- Exam 80%
  - Part 1 (40%) – ontologies, owl, sparql queries and data integration questions
  - Part 2 (40%) – dataset to be imported in noSQL store, list of queries to create

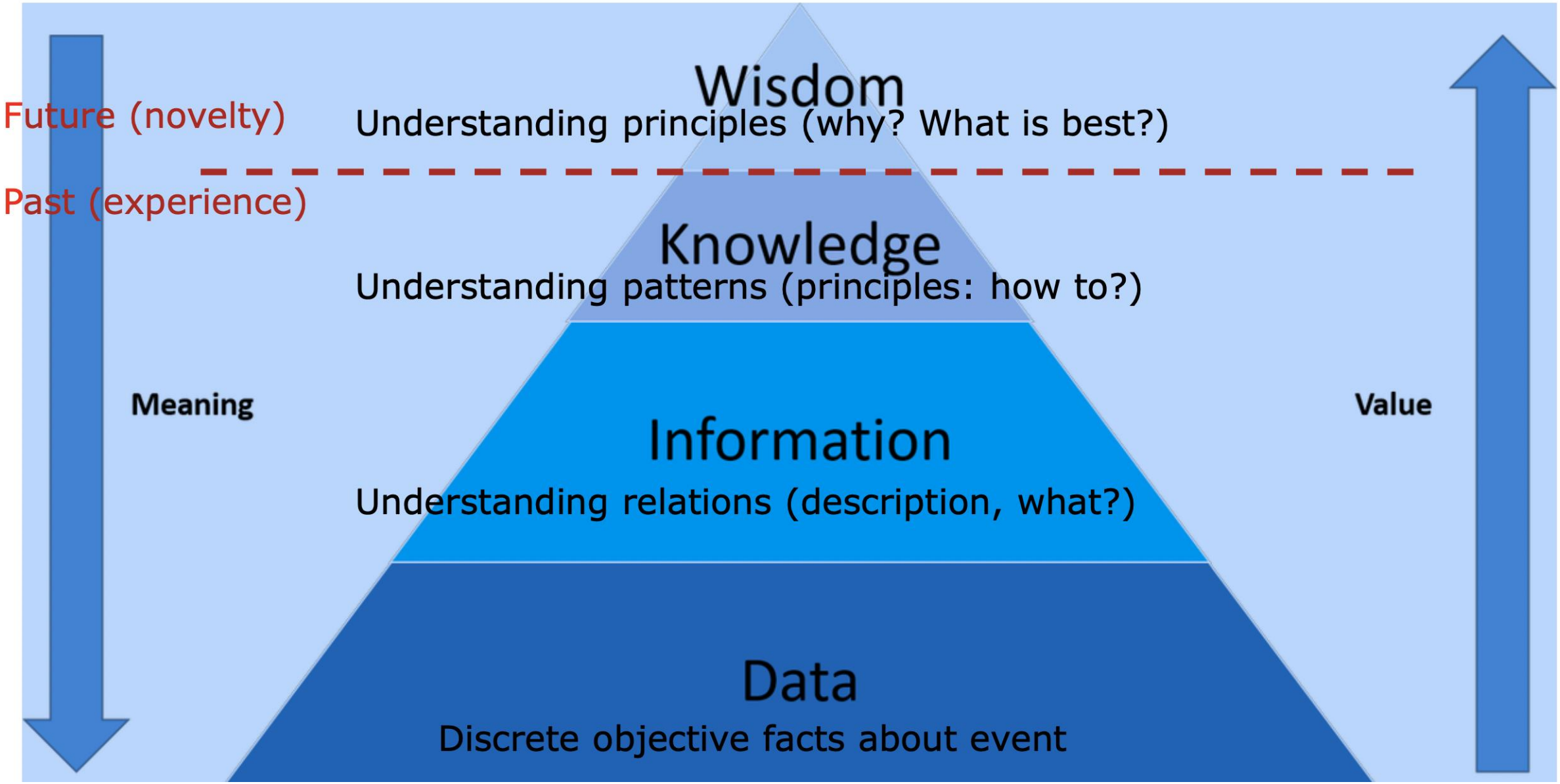
References

- Knowledge Graphs and Big Data Processing. Valentina Janev, Damien Graux, Hajira Jabeen, Emanuel Sallinger. Lecture Notes in Computer Science (LNCS, volume 12072), 2020.
- Encyclopedia of big data. Eedited by Laurie A. Schintler, Connie L. McNeely, Cham, Switzerland: Springer, 2022
- Big Data: Principles and Paradigms. Rajkumar Buyya, Rodrigo N. Calheiros, Amir Vahid Dastjerdi. Elsevier:London, 2016
- *Handbook of semantic web technologies*. DOMINGUE, John, FENSEL, Dieter, et HENDLER, James A. (ed.). Springer Science & Business Media, 2011.

Data, information, and knowledge

- **Data** consists of a collection of elementary fragmented figures and symbols usually lacking context (raw or unorganised)
  - Temperature, price, surface, distance, numbers, text…
  - Originate from observation, measurement, and collection and lacks meaning
- **Information** is data put in context (structured)
  - Integrating different data to describe a state of affair: quantity of sale, monthly income, portfolio of a given finance operator, tweets of a known person…
  - Result from data processing (grouping, calculating, sorting, dashboard…) making it interpretable: one can answer questions such as how many units did we sell? How was the monthly income of our Dubai sales unit?
- **Knowledge** a set of meanings, rules, procedures related to a domain
  - Applying knowledge to information allows making decisions and taking actions
  - Knowledge can be acquired formally or through experience
  - Knowledge is possessed by human (a cognitive state elaborated through education, training, experience…)

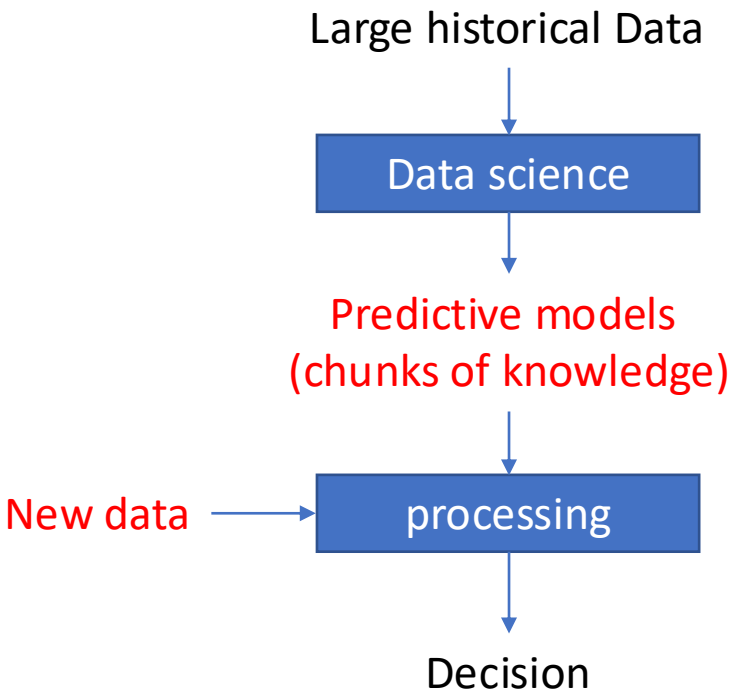




DIKW Pyramid, Ackoff 1989

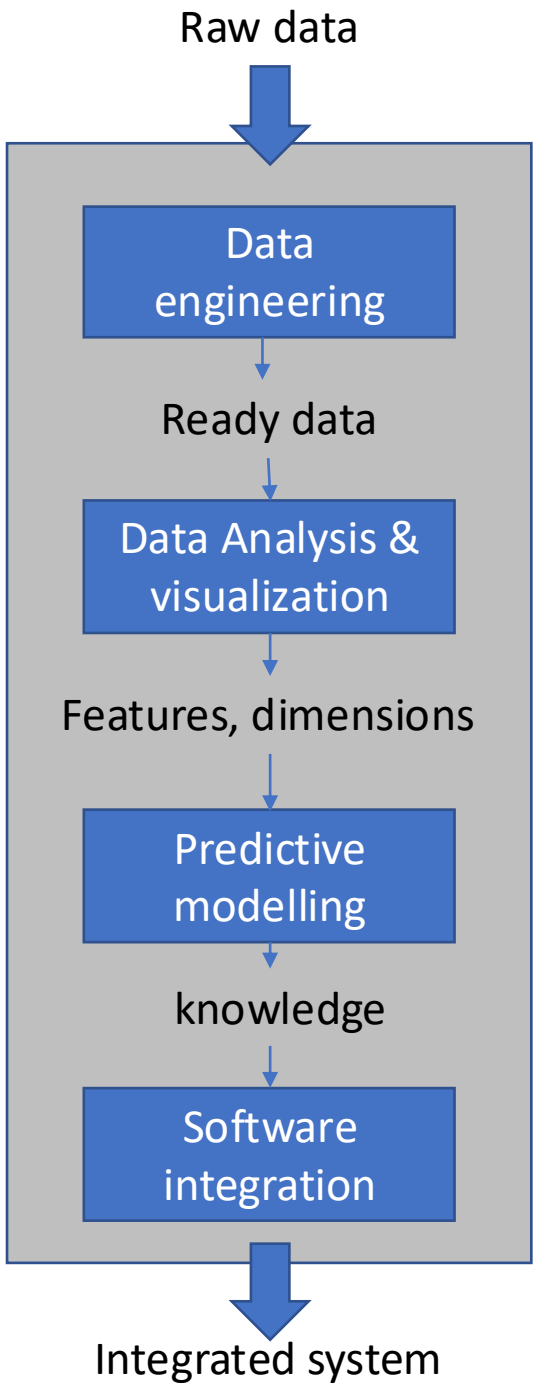
Data science

- The process of using machine learning algorithms to create predictive models (knowledge) from data to solve specific problems
- Data originates from the problems domain (health, industry, management, education…)
- Creating machine learning algorithms require
  - Maths (calculus, linear algebra, probability theory, differential geometry…)
  - Statistics (modelling, inference)
  - Computer science (coding, databases, IoT…)

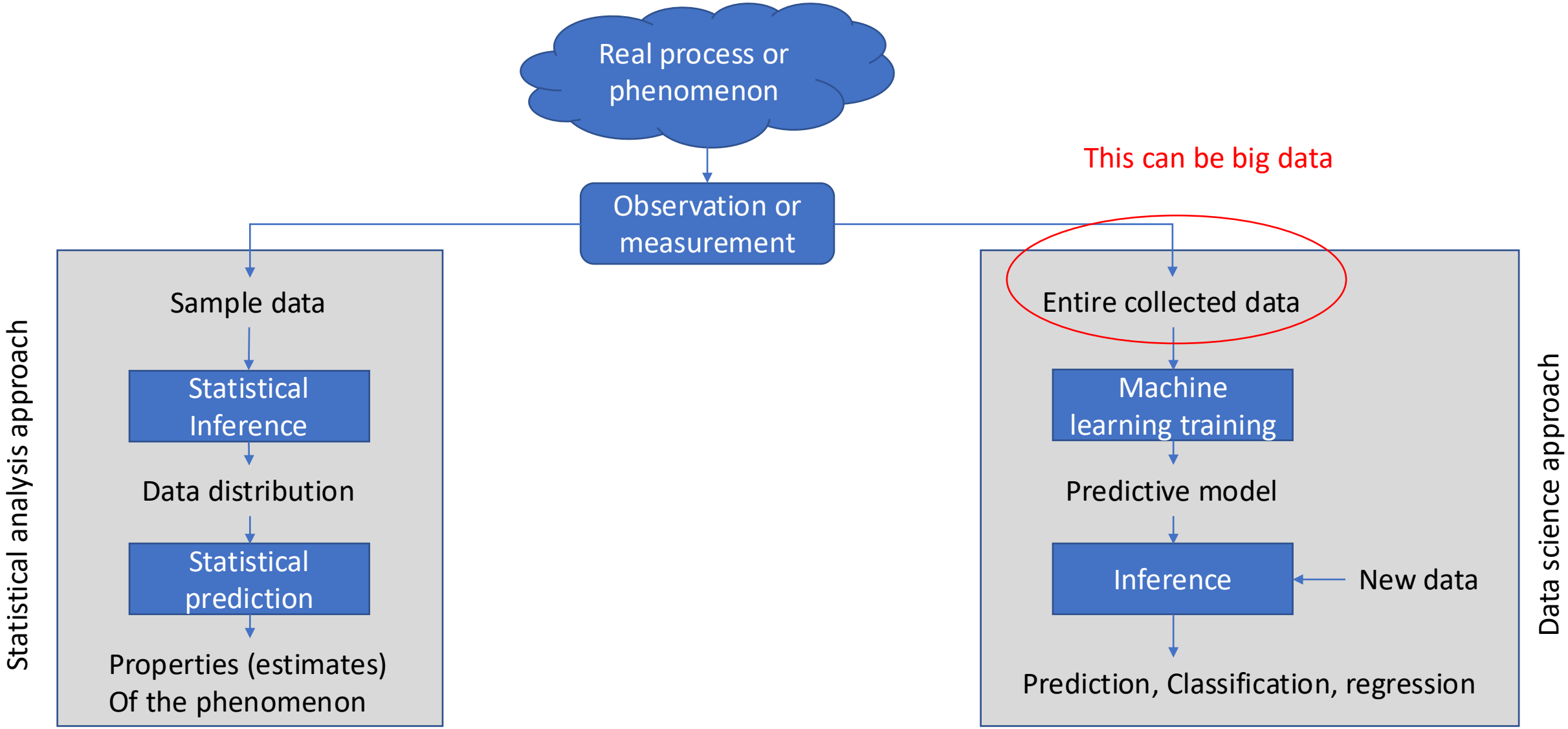


Data science process

- Data science consists of a process made of various activities to transform data into knowledge
  - **Data engineering** (make the data ready)
    - Collecting, formatting, conditioning, pre-processing, querying
  - **Data analysis** (understand the data)
    - Transforming, modelling, extracting features, visualising
  - **Predictive modelling** (extract insights)
    - Developing models (discriminative or generative)
    - Evaluate performance of models
  - **Software integration** (make use of the models)
    - Integrate models in large software systems
    - Such system would be AI systems (self-driving, automatic stocks investments…)



Statistical analysis vs. data science



Machine Learning vs. Statistics

Machine Learning

- **Algorithms**: Engineering heuristics and practical approach to find solutions
- Focus primarily on individual characteristics
- Computational algorithms to optimize an objective function
- Emphasizing uncertainty: Optional?
- Evaluation: split data (e.g. cross-validation, etc.)

Statistics

- **Models**: Mathematically principled approach to understand the system (investigation of variations)
- Focus primarily on population characteristics
- Models fit to the data and their properties
- Emphasizing uncertainty: Requirement
- Evaluation: same dataset (e.g. R2, residual analysis, etc.)

Big Data

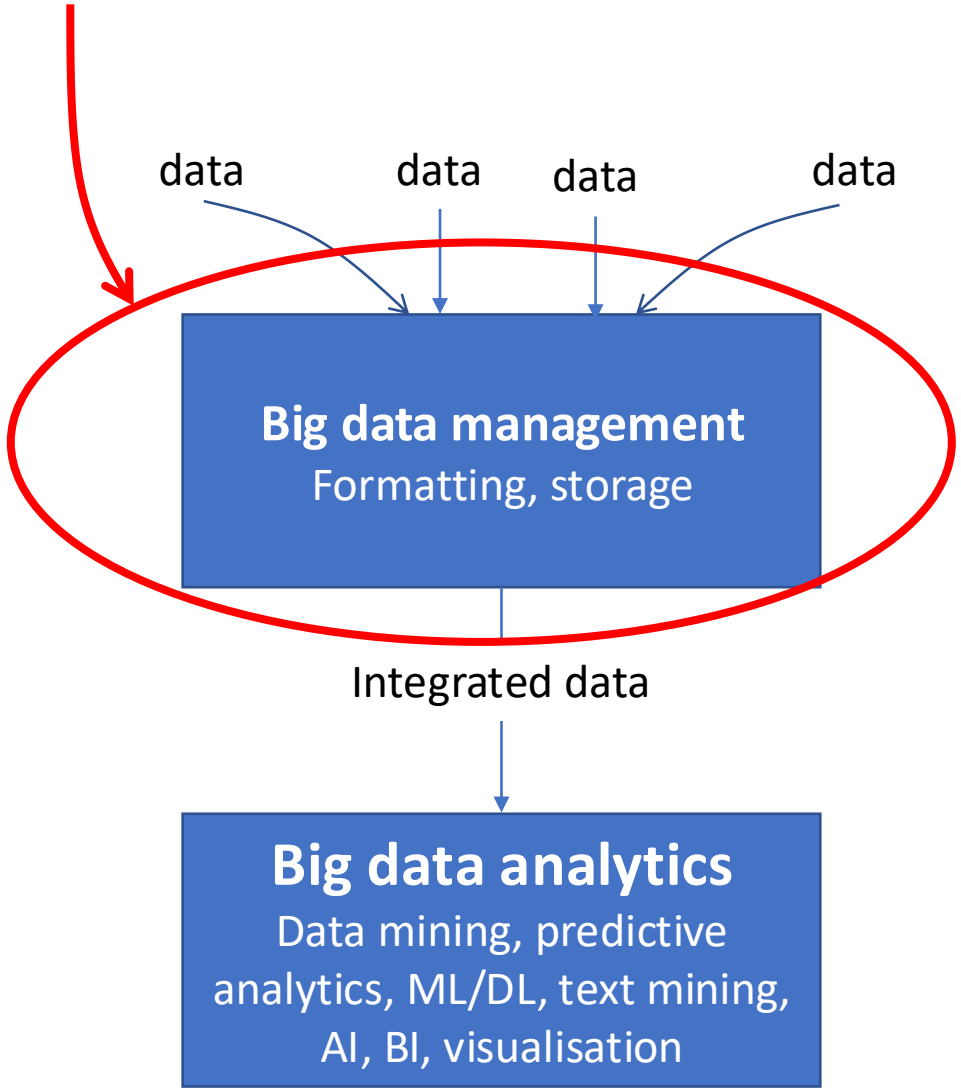
- Widespread of web, mobile, wearables, IoT
- Collection of large amount of data
  - Home, retail, health, transport…
- Fast growing data
  - 2 Zb in 2010, 18 Zb in 2016, 74 Zb in 2021 -> 2500 in 2030, 19200 in 2035!
- Questions:
  - How do we manage this data (store, retrieve…)?
  - How do we process this data?
  - How can we extract insights from this data?

Characteristics of big data

Characteristics	Description	Effect
Volume	Large amounts (size)	Which storage techniques
Velocity	Fast collection rate	What speed of processing
Variety	Structured, semi-structured, unstructured	What formatting standards
Veracity	Authenticity, provenance, accountability, uncertainty	How to authenticate data
Value	Include meanings and insights	Which capacity to derive actions

Big data analytics

- Big Data Analytics refers to the strategy of analysing large volumes of data
- Such data is gathered from a wide variety of sources
  - different kind of sensors
  - images/videos/media
  - social networks
  - transaction records
  - Mobile apps, web apps
  - Emails
- Companies use big data analytics for
  - Effective marketing
  - New revenue opportunities
  - Customer personalisation
  - Improving operational efficiency
- Big data management is the stage that makes data ready for big data analytics



Big data analytics is a sub-field of data science

Some business applications of big data analytics

- Customer acquisition and retention
- Targeted ads
- Product development
- Price optimization
- Supply chain and channel analytics
- Risk management
- Improved decision-making

Data sharing and value

- Data is usually collected by individuals, companies, organisations and governmental agencies
- Isolated analysis allow answering local questions of interest to the organisation or individual
- Sharing data and integrating it on a wider scale allows answering larger questions and deriving insights of general interest
- What are the processes, tools and standards for sharing data?

Various types of data

Type	Description
Big data	“Big data” is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” (Laney 2001) --------- “Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.” (Manyika, 2011)
Open data	“The data available for reuse free of charge can be observed as Open Data.” (Janev et al. 2018)
Linked data	The term Linked Data refers to a set of best practices for publishing structured data on the Web. These principles have been coined by Tim Berners-Lee in the design issue note Linked Data (Berners-Lee 2006).
Smart data	“Simply put, if Big Data is a massive amount of digital information, Smart Data is the part of that information that is actionable and makes sense. It is a concept that developed along with, and thanks to, the development of algorithm-based technologies, such as artificial intelligence and machine learning.” (Dallemand 2020)

Semantics

- A term invented by the French philosopher Michel Bréal to explain how terms may have various meanings for different people
- In computer science, semantics refers to the "meaning and practical use of data" (data objects used to represent a concept or entity)
- Semantic technologies are need when data from different sources and types is combined (integrated) to perform some decision making