Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.
Bayesian information criterion (BIC) includes a penalty for using more variables
Blind experiments or A/B testing may be used to show if relationship between various variables
Business models: A business model describes the rationale of how an organization creates, delivers, and captures value, in economic, social, cultural or other contexts.
Causation indicates that one event is the result of the occurrence of the other event
Identifying causation requires controlled experiments
Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.
Clustering tends to be associated with segmentation that allows us to recognize similar combinations of attribute values when we don't have predefined categories.
Confidentiality: We often don't own or manage corporate/internet/app data about ourselves
The source data is critical for advertisers so we cannot expect companies to be banned/excluded from using it and For many apps/websites, you must accept their privacy data sharing policies to use their services fully
The interface for selecting privacy preferences should move away from individual Internet platforms and be put into the hands of individual consumers
Data science extraction of knowledge/value from data through the complete data lifecycle process
Data scientist Skills and Job
Business: product development, business
Machine learning/Big data: unstructured data, structured data, machine learning, big and distributed data
Mathematics/Operations research: optimisation, mathematics,
graphical models, Bayesian and Monte Carlo statistics, algorithms, simulation
Programming: systems administration, back end programming, front end programming
Statistics: visualisation, temporal statistics, surveys and marketing, spatial statistics, science, data manipulation

Data developers: were relatively strong in all areas except Statistics
Data researchers: had little Programming
Data businesspersons: had little Mathematics/Operations research and even less Programming
Data creatives: were relatively strong in all areas except Mathematics/Operations research
Data value chain
Collection: getting the data
Engineering: storage and computational resources across the full lifecycle
Governance: overall management of data across the full lifecycle
Wrangling: data preprocessing, cleaning
Analysis: discovery (learning, visualisation, etc.)
Presentation: arguing the case that the results are significant and useful
Operationalisation: putting the results to work, so as to gain benefits or value
Data wrangling: Process of transforming "raw" data into data that can be analysed to generate valid actionable results and insights
Data needs to be cleaned, so it can be (re)used
Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.
Data Governance is making sure that it is done appropriately
Policies, training, providing resources Planning and understanding
Data lifecycles
Create data→processing data→analysing data→preserving data→giving access to data→re-

using data

Data management capability maturity

Data acquisition, processing and quality assurance Goal: Reliably capture and describe scientific data in a way that facilitates preservation and reuse

Data description and representation Goal: Create quality metadata for data discovery, preservation, and provenance functions

Data dissemination Goal: Design and implement interfaces for users to obtain and interact with data

Repository services/preservation Goal: Preserve collected data for long-term use

Digital container: file format whose specification describes how different elements of data and metadata coexist in a computer file

Growth laws

Moore's Law – capability and size of IT - Number of transistors per chip doubles every 2 years (starting from 1975)

Koomey's Law – capability and size of IT - Amount of battery needed will fall by a factor of 100 every decade

Bell's Law – purpose of IT - Roughly every decade a new, lower priced computer class forms based on a new programming platform, network, and interface resulting in new usage and the establishment of a new industry

Zimmerman's Law – relationship between - Zimmerman is creator of Pretty Good Privacy (PGP), an early encryption system

Surveillance is constantly increasing" and Privacy constantly decreasing

Human-in-the-loop

Need the human perspective in the design, understanding and review of the process, how it is utilised and its results

Influence diagrams: consists of nodes and arcs and is an alternative to a decision tree.

Machine-readable data: data (or metadata) which is in a format that can be understood by a computer, e.g., XML, JSON

Markup language: system for annotating a document in a way that is syntactically distinguishable from the text

Metadata

structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource

data about data, structured so that a computer can process & interpret it

- Descriptive: describes content for identification and retrieval
- Structural: documents relationships and links
- Administrative: helps to manage information

Machine Learning is concerned with the development of algorithms and techniques that allow computers to learn.

No Free Lunch Theorem

If a [learning] algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.

Use **NoSQL database** when: Storing large volume of data with little to no structure and Data changes rapidly

NoSQL databases offer a rich variety beyond traditional relational.

NIST Case studies

Data sources: where does the data comes from? data volume: how much there is?

Data velocity: how does the data change over time?

Data volume: how much there is

Data variety: what different kinds of data is there?

Data veracity: is the data correct? what problems might it have?

Software: what software needed to do the work?

Analytics: what statistical analysis & visualisation is needed?

Processing: what are the computational requirements?

Capabilities: what are key requirements of the operational system?

Security/privacy: what security/privacy requirements are there?
Lifecycle: what ongoing requirements are there?
Other: are there other notable factors?
Outliers are values outside of the expected parameters for the data
Open data: Data that is "freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control
PMML (Predictive Model Markup Language) provides a standard language for describing a (predictive) model that can be passed between analytic software (e.g. from R to SAS).
Regression trees
A regression tree is a supervised machine learning algorithm that predicts a **continuous-valued response variable** by learning decision rules from the predictors (or independent variables)
Rather than using a single function to represent the data, divide the data into similar segments, then make predictions in each segment
Stakeholders are any parties that have a relationship with a project/policy/product/data. This includes the data's source
, managers, analysts and users, IT developers, data scientists
The responsibility of looking after the data lies with all stakeholders – the data scientist is a key to making this happen
Tidy data is an approach to structuring how you store and use data
The Vs
Volume: The quantity of data to be stored
Velocity: The speed at which data enters the system and must be processed
Variety: Variations in the structure of the data to be stored
Veracity: correctness, truth, i.e., the lack of …
Variability: change in meaning over time

• What is a data scientist?

A **data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes through each stage in the data lifecycle. Moreover, **Data science** is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process.

• What is R Markdown?

To maintain a reproducible workflow, you need to record what steps you take in a process. R Markdown is an authoring format (.Rmd files) that enables us to combine embedded R code

• How has data science changed business models?

Information brokering service: buys and sells data/information for others.
Information-based differentiation: satisfies customers by providing a differentiated service built on the data/information.
Information-based delivery network: deliver data/ information for others.
Information provider: business selling the data/ information it collects.

• How can you handle missing data?

If a small fraction of cases have several missings, drop the cases.
If a variable or two, out of many, have a lot of missings, drop the variables.
If missings are small in number, but located in many cases and variables, you need to impute these values (replace with substituted values) to do most analyses.

• What is the significance of open data?

open data provides new opportunities for business, new products and services, and can raise productivity
open data supports public understanding and citizen engagement
scientists need to better publicise their data (with help from universities, etc.)
industry sectors should work with regulators and coordinate industry collaboration
collaboration across sectors in both public and private settings

• How are temporal elements included in data?

The temporal aspect of the data can be of different types. Specific and Relative.