What is Data Science?

(a personal view: connecting data to reality)

Jordi Vitrià, PhD

UNIVERSITAT DE BARCELONA

jordi.vitria@ub.edu

Machine Learning

Fat Data

Data Science

Dirty Data

Big Data

Data Mining

Artificial Intelligence

Data Science is a **multidisciplinary methodology** to help to define what we want to do with data, how do we evaluate our algorithms, what decisions can be grounded on data, how do we combine evidences from several sources, etc.

Data Science Path

What do I want?
Does it have sense?

Question

What are my data sources? How reliable are they?

Acquire

How do I develop an understanding of the content of my data?

Describe

What are the key relationships in my data?

Discover

How do I develop an understanding of the content of my data?

Analyze

What are the likely future outcomes?

Predict

Are my expectations fulfilled?

Evaluate

Taking (big)data-based decisions is not new but now it is easier.

Big Data

Big Data

What is Big Data?

- For some people, they have big data when its size > 65536 x 256.

- In general we have big data when its size does not allow its storage and analysis in a big computer.

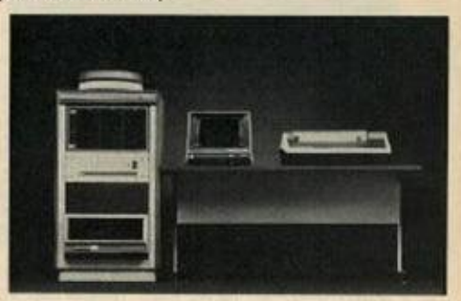
10 Megabyte Hard Disk

$3,495*

5440-12 Top Load Drive
* Factory rebuilt 10MB cartridge disk drive only
A new Cameo Data Systems controller is available for $1,495
$4,495 for a brand new Ampex 10MB drive only

We are the CP/M** and MP/M** specialist of Southern California. We can supply you with the latest CP/M ($150) or MP/M ($300) and with Standard BIOS ($150) or Custom BIOS ($300). Immediate delivery worldwide. Domestic and foreign inquiries invited...dealers too.

**CP/M and MP/M are Trademarks of Digital Research



We are a full service computer retailer. We totally integrate hardware and software into high quality, high reliability systems. Systems for use in development, process control and general business. Word processing naturally, multi tasking and multi processing too.

COMPUTER COMPONENTS

Circle 279 on inquiry card. 5848 Sepulveda Boulevard Van Nuys, California 91411 213•786-7411

July 1980.

More common

Fat Data

Big Data

Less common

Big Data

With a personal computer:

- **You can find an element in a 1 MB file in less than a second.**
- **You can find an element in a 1 GB file in less than a minute.**
- **You can find an element in a 1 TB file in less than sixteen hours.**
- **You can find an element in a 1 PB file in less than two years.**
- **You can find an element in a 1 EB file in less than two thousand years.**

Big Data

LinkedIn manages 7 trillion messages per day

Walmart generates 2.5 petabytes of data every hour.
(2,5 × 10^{16} bits = one million gigabytes).

Big Data

- On average, people send about 500 million tweets per day.
- The average U.S. customer uses 1.8 gigabytes of data per month on his or her cell phone plan.
- Amazon sells 600 items per second.
- On average, each person who uses email receives 88 emails per day and send 34. That adds up to more than 200 billion emails each day.
- MasterCard processes 74 billion transactions per year.

Big Data

Big data is more than size.
It is commonly characterized with several V:

Volume

Velocity

Variety

Big Data

The main phenomenon behind Big Data is **datification**.

The V's are a consequence of it.

Big Data

We are rendering into data many aspects of the world that have never been quantified before:

business networks

books I'm reading

location

physical activity

consumed food

purchases

physiological signals

straight thoughts

friendship

gaze

driving behavior

Artificial Intelligence and Machine Learning

Artificial intelligence is an academic discipline devoted to the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, language recognition, decision-making, planning, reasoning, etc.

Artificial intelligence is classified into two parts, General AI and Narrow AI. General AI refers to making machines intelligent in a wide array of activities that involve thinking and reasoning. Narrow AI, on the other hand, involves the use of artificial intelligence for a very specific task.

Machine learning is a subset of artifi cial intelligence that uses algorithms to learn from data (inductive behavior).

Data Science

Data Science

Technology is the collection of tools, including machinery, modifications, arrangements and procedures used by humans.

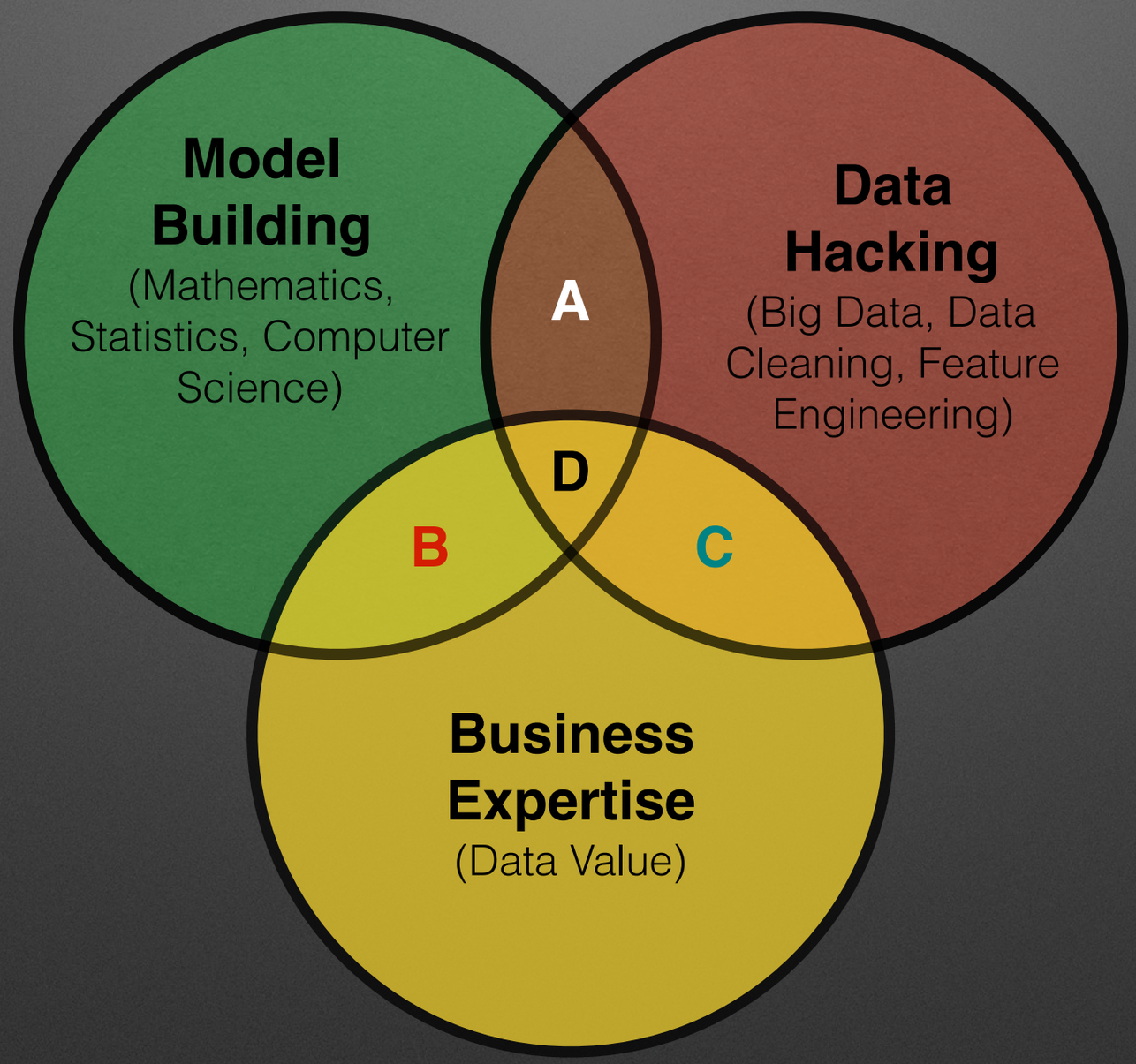
Big Data is a key **technology** to process massive amounts of data (f.e. to count items).

Methodology is the systematic, theoretical analysis of the methods applied to a field of study.

Data Science is a **methodology** to define what we want to do with data, how do we evaluate our actions, what decisions can be grounded on data, how do we combine evidences from several sources, etc.

D is an empty set!

A+B+C = D



Data Science Tasks

Background

Output

Domain Knowledge, Causality, Decision Making, Human Behavior

Data Insights: realistic models.

Prescriptive Decisions: Why? What is best?

Domain Knowledge, Statistics, Machine Learning, Complex Systems, etc.

Data Analysis: experiment design, simple models.

Predictive Decisions: How to? When to?

Knowledge

Data Processing, Visualization

Data Processing: aggregation, segmentation, labeling, metrics, etc.

Data Descriptions.

Data Processing

Data Preparation: cleaning, anomaly detection, imputation, etc.

Datasets

Data Engineering

Data Storing from reliable data flow, using good infrastructure pipelines.

Databases

Information

Data Engineering

Data Collection from instruments, logging, sensors, external data, etc.

Raw Data

Data

Data Science

Steps:

- Ask a question.
- Get the data. They can be heterogeneous and non structured.
- Data Processing (cleaning, munging, etc.).
- Data Analysis (computer science, linguistics, economy, sociology, etc.).
- Take a decision and act.

Questions

- **Description** is using data to provide a quantitative summary of certain features of the world. → What is the mean value of X?

- **Prediction** (or **association**) is using data to map some features of the world (the inputs) to other features of the world (the outputs). → How would seeing X change my belief in Y?

- **Causation**: Measuring the causal influence of a variable X in another variable Y, while excluding any influences on Y not actually due to the causal effect of X, and being able to guess what the effect will be if one performs an action. → How would expected lifespan change if more people become vegetarian?

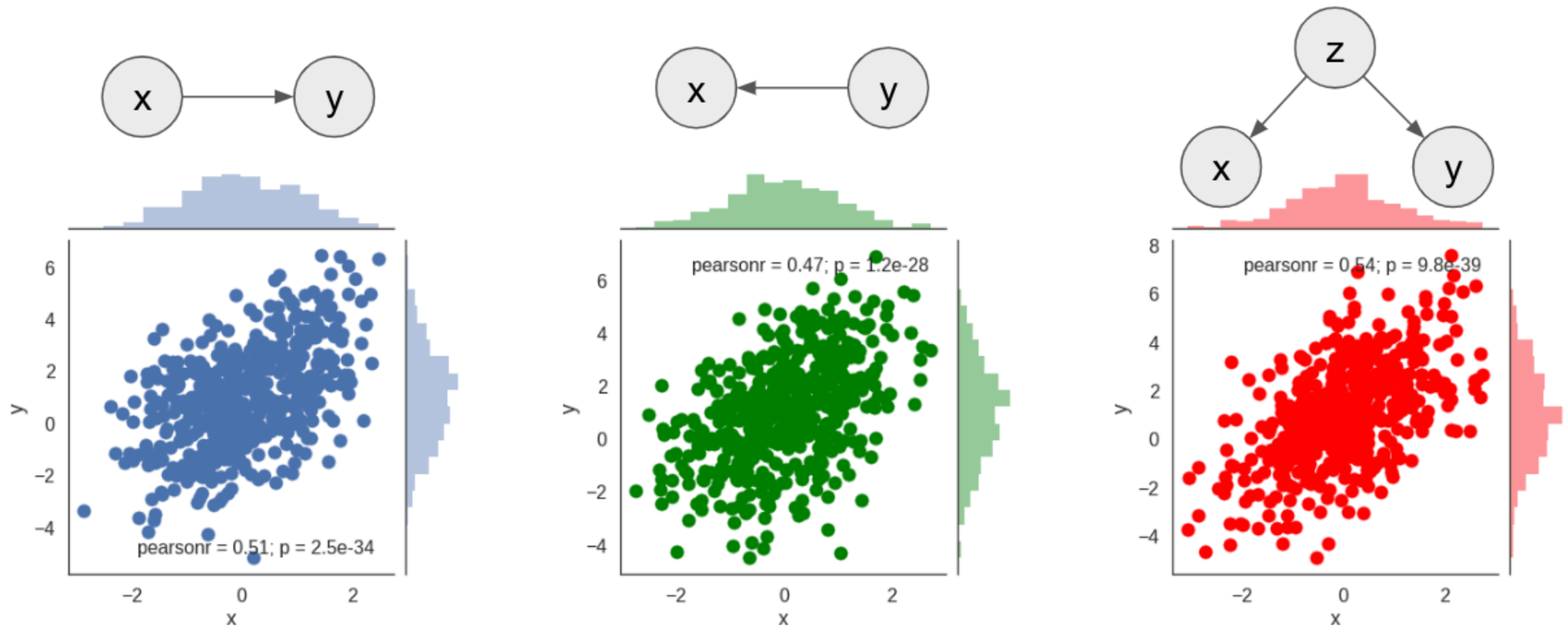
Intervention

- **Counterfactuals**: Being able to reason about hypothetical situations, things that *could* happen. → Would my grandfather still be alive if he did not smoke?

x = randn()
y = x + 1 + sqrt(3)*randn()

y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2

z = randn()
y = z + 1 + sqrt(3)*randn()
x = z



The joint distributions p(x,y) of these three causal models are indistinguishable.

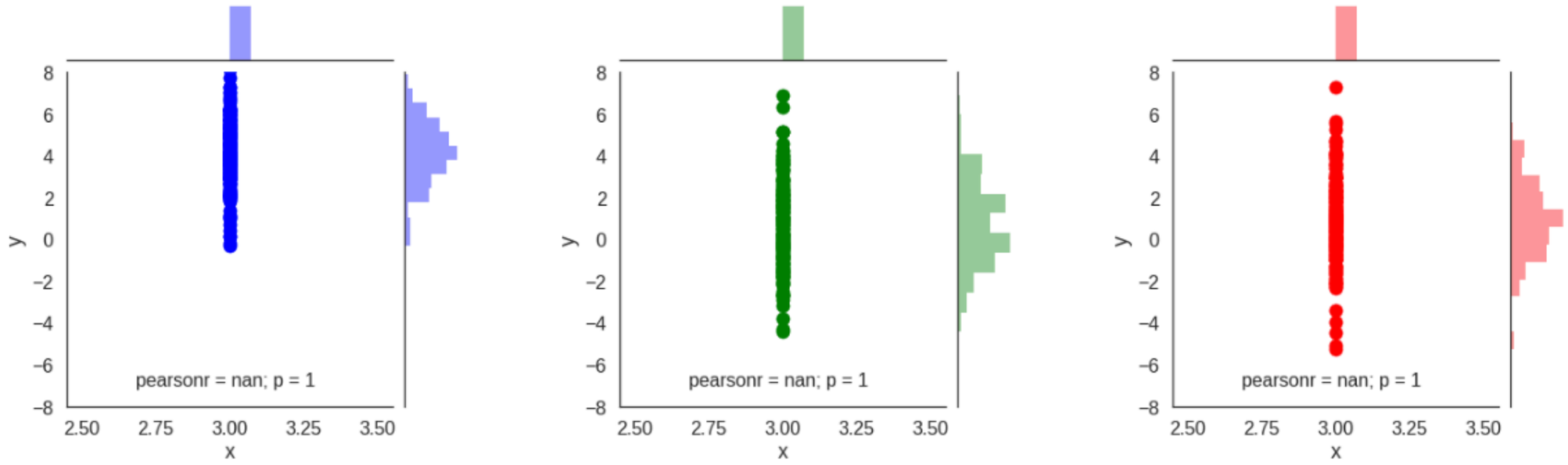
Source: https://www.inference.vc/

Intervention x=3

x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3

y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3

z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3



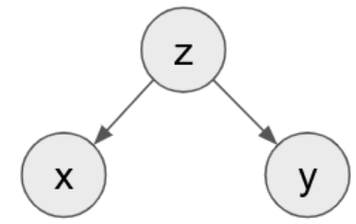
But their marginals p(y/x) are different if there is an intervention!

Source: https://www.inference.vc/

Do calculus and observational data

$$P(y|do(X)) = p(y|x)$$

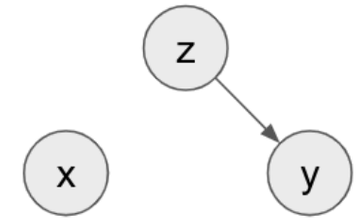
$$P(y|do(X)) = p(y)$$



$$P(y|do(X)) = p(y)$$

$$P(y|do(X)) = p(y|x)$$

$$P(y|do(X)) = p(y)$$



$$P(y|do(X)) = p(y)$$

Counterfactuals: David Blei's election example

Given that Hilary Clinton did not win the 2016 presidential election, and given that she did not visit Michigan 3 days before the election, and given everything else we know about the circumstances of the election, what can we say about the probability of Hilary Clinton winning the election, had she visited Michigan 3 days before the election?

Let's try to unpack this. We are are interested in the probability that:

- she hypothetically wins the election

conditioned on four sets of things:

- she lost the election
- she did not visit Michigan
- any other relevant an observable facts
- she hypothetically visits Michigan

Why would quantifying this probability be useful? Mainly for credit assignment.

What are the limits of Data Science

- Data science must be bounded by ethical limits.
- Data science cannot substitute intuition or creativity.

> If I had asked people what they wanted, they would have said faster horses.
>
> Henry Ford.

What are the limits of Data Science

- Data science models reproduce what we do and how we do it (including bad things and wrong strategies). Prediction is a dangerous game!

Rich Caruana gives the example of a **pneumonia risk prediction** model on which he had worked. The purpose of the model was to evaluate whether a patient with pneumonia was at high or low risk, to help decide whether or not the patient should be admitted to the hospital. "On the basis of the patient data," says Caruana, "the model had found that patients with a history of asthma have a lower risk of dying from pneumonia. In reality, everybody knows that asthma is a very high risk factor for pneumonia. What the model found is the result of the fact that asthma patients get healthcare faster, which lowers their chance of dying compared to the general population."

https://cacm.acm.org/news/214618-in-black-box-algorithms-we-trust-or-do-we/fulltext

Ethical Data Science

If a DS system is making automatic decisions, someone has the **responsibility** of those decisions.

Problems:

- Choosing a wrong model.
- Building a model with inadvertently discriminatory rules.
- Not providing explanations about decisions.
- Not respecting privacy.
- Etc.

Ethical Data Science

Responsible data science challenges:

- Data science **without prejudice** – How to avoid unfair conclusions even if they are true?
- Data science **without guesswork** – How to answer questions with a guaranteed level of accuracy?
- Data science that **ensures confidentiality** – How to answer questions without revealing secrets?
- Data science that **provides transparency** – How to clarify answers such that they become indisputable?

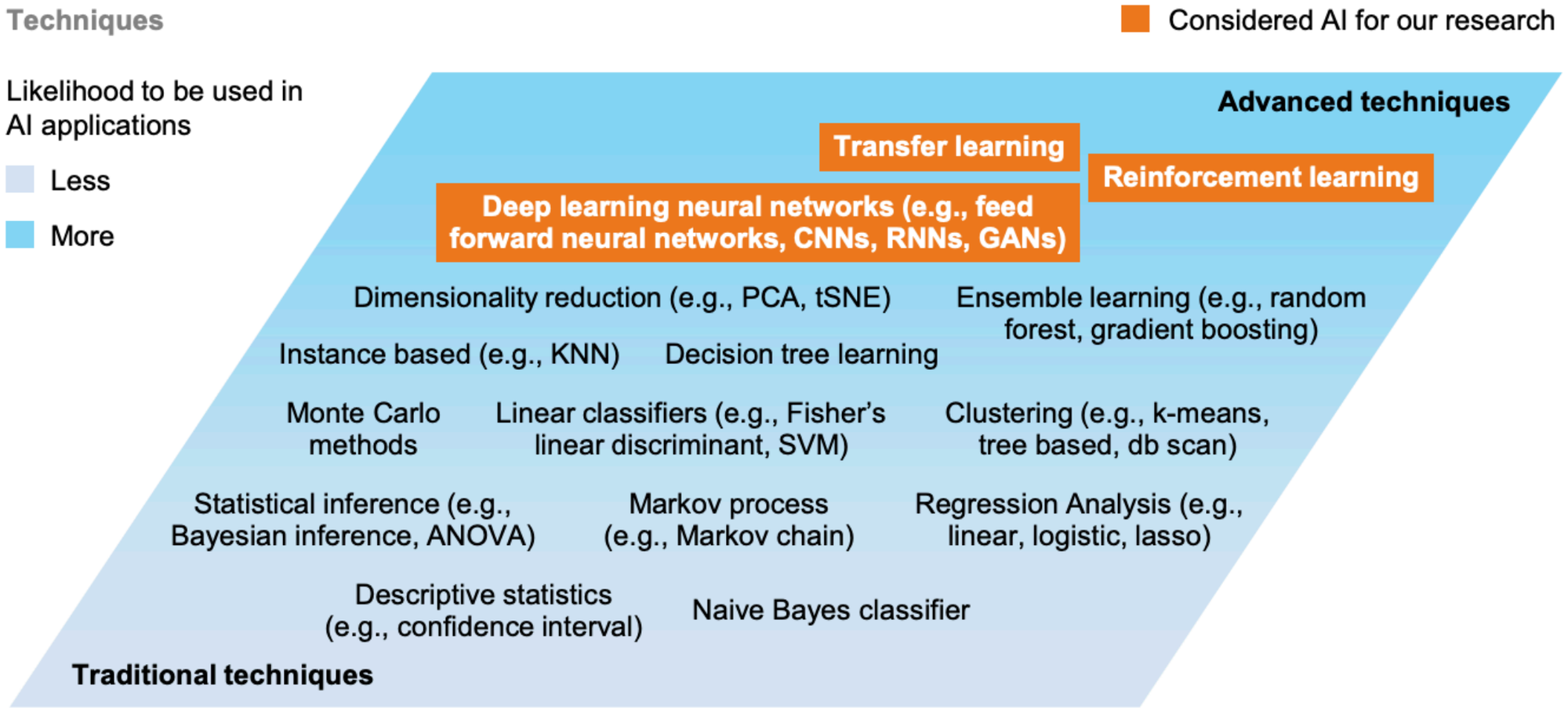
Canonical Problems and Tools

Classification	To which category does this data point belong?	Medical diagnosis: does this tissue show signs of disease? Banking: is this transaction fraudulent? Computer vision: what type of object is in this picture? Is it a person? Is it a building?
Regression	Given this input from a dataset, what is the likely value of a particular quantity?	Finance: what is the value of this stock going to be tomorrow? Housing: what would the price of this house be if it were sold today? Food quality: how many days before this strawberry is ripe? Image processing: how old is the person in this photo?
Clustering	Which data points are similar to each other?	E-commerce: which customers are exhibiting similar behaviour to each other, how do they group together? Video Streaming: what are the different types of video genres in our catalogue, and which videos are in the same genre?
Dimensionality reduction	What are the most significant features of this data and how can these be summarised?	E-commerce: what combinations of features allow us to summarise the behaviour of our customers? Molecular biology: how can scientists summarise the behaviour of all 20,000 human genes in a particular diseased tissue?
Semi-supervised learning	How can labelled and unlabelled data be combined?	Computer vision: how can object detection be developed, with only a small training data set? Drug discovery: which of the millions of possible drugs could be effective against a disease, given we have so far only tested a few?
Reinforcement learning	What actions will most effectively achieve a desired endpoint?	Robots: how can a robot move through its environment? Games: which moves were important in helping the computer win a particular game?

Canonical Problems and Tools

Classification	To which category does this data point belong?	Medical diagnosis: does this tissue show signs of disease? Banking: is this transaction fraudulent? Computer vision: what type of object is in this picture? Is it a person? Is it a building?
Regression	Given this input from a dataset, what is the likely value of a particular quantity?	Finance: what is the value of this stock going to be tomorrow? Housing: what would the price of this house be if it were sold today? Food quality: how many days before this strawberry is ripe? Image processing: how old is the person in this photo?
Clustering	Which data points are similar to each other?	E-commerce: which customers are exhibiting similar behaviour to each other, how do they group together? Video Streaming: what are the different types of video genres in our catalogue, and which videos are in the same genre?
Dimensionality reduction	What are the most significant features of this data and how can these be summarised?	E-commerce: what combinations of features allow us to summarise the behaviour of our customers? Molecular biology: how can scientists summarise the behaviour of all 20,000 human genes in a particular diseased tissue?
Semi-supervised learning	How can labelled and unlabelled data be combined?	Computer vision: how can object detection be developed, with only a small training data set? Drug discovery: which of the millions of possible drugs could be effective against a disease, given we have so far only tested a few?
Reinforcement learning	What actions will most effectively achieve a desired endpoint?	Robots: how can a robot move through its environment? Games: which moves were important in helping the computer win a particular game?

Canonical Problems and Tools



SOURCE: McKinsey Global Institute analysis

---

4 See Jacques Bughin, Brian McCarthy, and Michael Chui, "A survey of 3,000 executives reveals how businesses succeed with AI," *Harvard Business Review*, August 28, 2017.

5 Michael Chui, James Manyika, and Mehdi Miremadi, "What AI can and can't do (yet) for your business," *McKinsey Quarterly*, January 2018.

6 For a detailed look at AI techniques, see *An executive's guide to AI*, McKinsey Analytics, January 2018. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai