Big Data, An Introduction

prof. dr Arno Siebes

Algorithmic Data Analysis Group
Department of Information and Computing Sciences
Universiteit Utrecht

Outline

Today we introduce two topics

- Big Data
  - what does it mean, how did it come to be, challenges it poses, and why is it so popular.
- Data Mining
  - data becomes valuable through its analysis, my favourite term for this is data mining

Statistics and, more general, probability theory are indispensable for the analysis of data, we will revise some basic notions today as well.

What is Big Data?

Big Data

> *"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."*

by prof. Dan Ariely

- James B. Duke Professor of Psychology and Behavioral Economics at Duke University
- founding member of the Center for Advanced Hindsight

So, let us first discuss what Big Data actually means, starting with its root cause: the digital era we live in.

The Digital Era

In the roughly 70 years after the invention of the computer,

- the world has become thoroughly digitized

The work place is fast becoming totally(?) computerised

- from office automation to computer assisted diagnosis to automatic legal research
- from robot manufacturing to 3-D printing
- from sat nav to self-driving cars
- from blue collar to highly skilled

The environment is continuously monitored and controlled

- through a multitude of sensors and actuators

And everyone is always connected

- through smartphones, smart watches, tablets, laptops, wearables, ...

Digital Trails

Everything computerised

- means that everything is digital

That is,

- everything causes data to stream through computers and networks; in fact, that is often all there is.

And data that streams through a computer

- is recorded and stored.

Every process

- leaves digital trails

Ever more things that happen in the world

- are recorded in ever greater detail

Hence *Big Data*

Big Data

The non-technical term Big Data is "defined" by

- V olume: ever more massive amounts of data
- V elocity: stream in at ever greater speed
- V ariety: in an ever expanding number of types

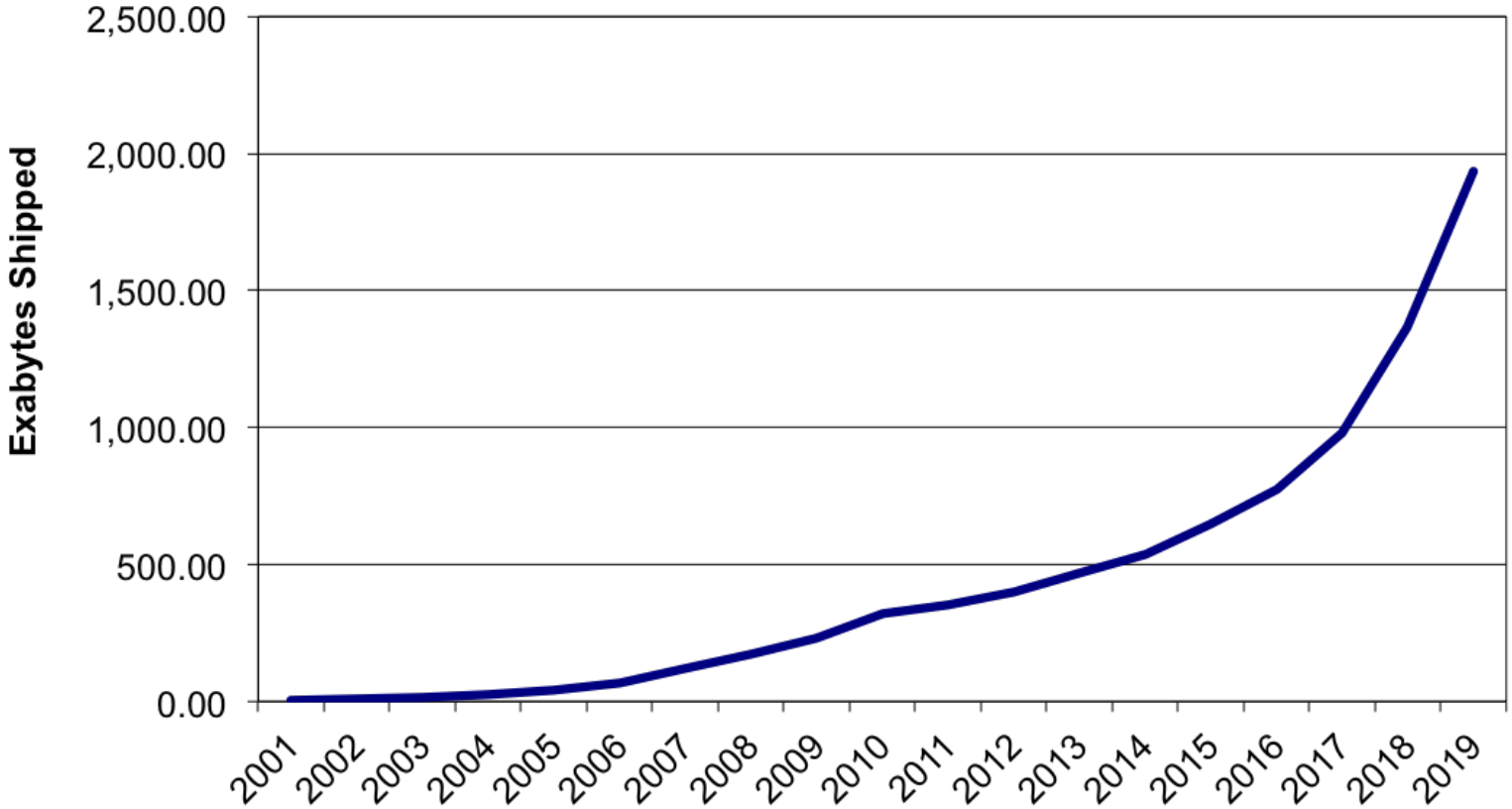
While being non-technical, the three **V**'s characterisation points out what the problem is

- data that is too big to handle

One should compare it to the Very Large DB conference series

- "very large" was something completely different in 1975 (the first VLDB) than it is now
- but the semantics: "very large = too big to fit in memory" is still the same.

Volume: HDD Shipment Sales in Exabyte



(Forbes, Jan 29, 2015), note this is for hard disks only(!)
Recall: 1000 B = 1 kB, 1000 kB = 1 MB, 1000 MB = 1 GB, 1000 GB = 1 TB, 1000 TB = 1 P(eta)B, 1000 PB = 1 E(xa)B, 1000 EB = 1 Z(etta)B, 1000 ZB = 1 Y(otta)B

Data Production

One way to view it is to say:

- 90% of the world's data has been produced in the last two years
- we produce 2.5 quintillian (10^{18}) bytes per day

Another way is, we produced

- 100 GB/day in 1992
- 100 GB/hour in 1997
- 100 GB/sec in 2002
- 28,875 GB/sec in 2013

Velocity

In 1 second (July 13, 2016, from Internet Live Stats) there were:

- 731 Instagram photos uploaded
- 1,137 Tumblr posts made
- 2,195 Skype calls made
- 7,272 Tweets send
- 36,407 GB of Internet traffic
- 55,209 Google searches
- 126,689 YouTube videos viewed
- 2,506,086 Emails sent (including spam)

The data is not only vast but you also get it at an incredible speed.

- if you want to do something with that data, *do it now*

Variety

In a first year databases course

- you are taught about tables and tuples

Data that can be queried using SQL and is known as

- *structured* data

It is estimated that over 90% of the data we generate is

- *unstructured* data
  - text
  - tweets
  - photos
  - customer purchase history
  - click-streams

Variety means that we want, e.g., to analyse

- different kinds of data, structured and unstructured, from different data sources as one combined data source

Big Data in Society

Think about it, Facebook has

- in the order of 1.5×10^9 users
- with (on average) ≥ 50 links
  - i.e., in the order of 4×10^{10} (undirected) links in the graph
  - (compare: the brain, 10^{11} neurons, $10^{14} - 10^{15}$ connections)

Supermarkets know

- the exact content of each transaction
  - and to a large extent aggregated by loyalty cards

Banks know

- each and every (financial) transaction of their customers
  - how many people still use cash?

The numbers are staggering

- many (most?) companies are to a smaller or larger extent information companies

Big Data in Science

Science has its own Big Data collections, e.g.,

Astronomy has the Australian Square Kilometre Array Pathfinder

- currently acquires 7.5 terabytes/second of sample image data
- 750 terabytes/second (25 zettabytes/year) by 2025

Biology through high speed experiments, e.g., for genomic data

- the 2015 worldwide sequencing capacity was 35 petabases/year
- expected to grow to 1 zettabase/year by 2025

The Royal Dutch Library

- has an archive containing (digitized)
  - over 300.000 books, 1.3 million newspapers, 1.5 million magazine pages, ...

Dans, Data Archiving and Network Services (KNAW and NWO)

- has over 160.000 data sets ready for re-use

Think of the potential value of Facebook's data

- for social science research

But, Why?

Big Data is a huge stream of varied data that comes in at an incredible rate.

- but why do we have Big Data?

More precisely, why

- do we generate such vast amounts of data?
- why do we want to store and/or process these amounts

The short answers are

- because we can
- because there is value

Slightly more elaborate answers on the next couple of slides

Information is Immaterial

Different from anything else, information is *not* made out of matter

- it may always be represented using matter, but that is just a representation

Moreover, we know that

- all information can be represented by a finite bit string (Shannon)
- every effective manipulation of information can be done with one machine only: a Universal Turing Machine (Turing)

Hence, we can. If

- each type of information had its own unique representation
- and each manipulation (of each type of) information would require its own machine

We would not be talking about Big Data

Immaterial Implies: No Size

And, no size means we can miniaturize. Hence

Moore's Law

The CPU has seen a 3 million fold increase of transistors

- Intel 4004 (1973): 2300 transistors
- Intel Xeon E5-2699 v4 (2016): 7.2×10^9 transistors (its L3 cache is almost three times the size of my first hard disk: 55 MB vs 20 MB)

Kryder's Law

Hard Disk capacity has seen a 1.5 million fold increase

- IBM (1956): 5 MB (for $50,000)
- Seagate (2016): 8TB (for $200)
- note: Bytes per dollar increase 4×10^8

Hence, we can. Without these there would have been no ubiquity and without ubiquity there would be no Big Data

But, Why Is It All Stored?

We can, and, clearly, storage space is cheap, but still

- that doesn't mean that every bit is sacred, does it?

The reason is (at least) twofold

- You store everything about yourself
  - Facebook, YouTube, Twitter, Whatsapp, Google+, LinkedIn, Instagram, Snapchat, Pinterest, foursquare, WeChat, ...
  - don't ask me why you do that.
- Companies love these hoards of data, because it is *valuable*

It is valuable, because the detailed data trails give insight in

- in the relation between behaviour and health
- in what you like (and can thus be recommended to you)
- and many, many more examples

Hence, Big Data

Valuable, But

In 2006, Michael Palmer blogged

> *Data is just like crude. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value*

Hence, a course on Big Data: there are deep computer science challenges to make Big Data valuable

- and not only, or even predominantly, commercially

If we solve these, society (including all the sciences one can think of) can become a data driven society

Then There Is Value

> *"Uber, the world's largest taxi company, owns no vehicles. Facebook, the world's most popular media owner, creates no content. Alibaba, the most valuable retailer, has no inventory. And Airbnb, the world's largest accommodation provider, owns no real estate. Something interesting is happening."*

Tom Goodwin, on Techcrunch.com

No tangibles and still hard to beat, why?

- they are the interface
- they *know* the customer

Too Big to Handle?

Big Data is a huge stream of varied data that comes in at an incredible rate.

- *but is it really too big to handle?*

The answer is, as always, both

Yes if you want to be able to perform any arbitrary computation on that data

No there are computations you can perform without a problem.

Too big to handle means

- that we still have to find out how to do the things we want to do efficiently (enough)

The Challenges of Big Data

Being about data, the obvious challenges are for

- databases: storing and querying massive amounts of unstructured data
- information retrieval: how to find what you are looking for
- data mining: creating knowledge out of data

But clearly, all these challenges have aspects of both

- algorithmics: efficiency is key
- software technology: massive systems that are parallel and/or distributed; moreover, highly optimised implementations are easily orders of magnitude faster (the difference between doable or not).

Big Data is a good fit for COSC

Challenges for Databases

Just a few, in random order and perhaps not even the most important ones:

- how do you actually manage exabytes of data?
  - and process them especially since they will be much bigger than main memory
- how do you query unstructured data?
  - how about search accelerators like indices?
- how do you query over multiple (independent) data sources
  - can you optimize such queries?
- how do you take into account that not every source is reliable?
  - or has the same granularity, or ... ?

For each of these questions, and many more, there are partial answers. But partial is not enough, and most importantly

- do they scale up to Big Data?

Challenges for IR

Again, just a few and not necessarily the most important ones

- Ranking results (e.g., the outcome of a query) is, as ever, important as the full answer is meaningless due to size
  - how do you rank (almost) arbitrary data?
- How do you search over multiple data sources?
  - links may make it conceptually easier, but computationally harder
- How do you rank results over multiple data sources?
  - how about sources with different degrees of reliability?
- How do you adapt to the user?
  - the sources are far more varied than the web
  - the intended usage is far more varied than Google queries

Oh, and we want our results fast.

Challenges for Data Mining

Volume and Velocity have played their role in Data Mining almost from the beginning. Variety is similarly well known

- *is it?*

It is true that there are mining algorithms for many different data types

- categorical data tables, transaction data, data streams, time series, text, graphs, networks, ...

However, all these algorithms are defined

- for one data type only

Hence, as with DB and IR, the open problem for DM is

- how do you mine over multiple (distinct) data sources?

Since this is very much an open question, it will play no role in this course.

This Course

This course is focussed on the data mining part of Big Data

- but it is not yet another data mining course

Rather than presenting a variety of techniques for various problem classes

- such as deep neural networks

We focus on very simple mining problems and techniques

- and study how Big Data influences their solution,
- predominantly by looking at sampling
- note this is not necessarily simple...

That is, foundational issues for simple techniques, showing you how "size" can be overcome.

- there are other courses that cover the advanced techniques

Data Science

You might wonder why we use the buzz-word Big Data rather than the newer and more fancy one: Data Science.

Data Science

- was first used as an alternative name for computer science (Peter Nauer, 1960)
- as was hypology, by the way (from the classical Greek $\upsilon\pi o\lambda o\gamma\iota\sigma\mu o$ (calculate))

Later it was suggested as a new name for

- *Statistics* (C.F. Jeff Wu, 1997)

or more specifically,

- the intersection of Statistics and Computing (William S. Cleveland, 2001)

Nowadays, it is related to Big Data very similar to how Knowledge discovery in Databases is related to Data Mining

- perhaps, even more tightly integrated with (specific) application areas.

Since we focus on the analysis, Big Data seems more appropriate

Data Science

Data Science is arguably best viewed as the union of larger and smaller chunks of may different areas, e.g.,

- anything related to Big Data from computer science
- statistics
- law
  - what is allowed and why not
  - but also the application in law practice
- humanities
  - digital humanities
  - ethics (not everything that is not forbidden should be done)
- and many, many more

To a large extent, data science can be seen as a new research paradigm supporting many different fields of research (and businesses, of course)

A Note on Terminology

There are many terms that denote almost the same field or parts thereof

- (computational) statistics, machine learning, data mining, statistical learning, pattern recognition, signal processing, computational learning theory, exploratory data analysis...

It is usually possible to identify a technique X that one would expect sooner in a Journal or Conference for A than one for B, e.g.

- for A = Machine Learning and B = Data Mining
- X would be "Deep Learning" for A
- and "Pattern Mining" for B

But such differences are shallow. There are more or less clear differences in the traditions and culture of these different fields

- e.g., what constitutes a good paper

But, these largely reflect what all these different names already signify:

- the field from in the researcher originally started

I started out in Databases, hence I prefer the term Data Mining

Data Mining

Learning from Data

The basic assumption common to all areas of data analysis is that the data we have is *sampled* from some *distribution*:

$$D \sim \mathcal{D}$$

This simply means that we

- aim to learn something about reality from a (small) sample

That is, we want to generalise. For example, something like:

- in all of D I see x, so if I see a *new* sample d from $\mathcal{D}$ I expect that d will also have x

The reason for the "is sampled" assumption is that if D is all there is, the analysis doesn't have much use,

- if you have data that tells you exactly when raindrops hit your garden on July 14, 2016
- what could you possibly want to learn from that?

An Interlude on Probability

Probability Theory and Statistics are part of the language of data mining.

- to make sure that we all use the same terminology in the same way, I will include some brief interludes on these topics, usually just before we are going to use them.

This, perhaps unfortunately, not the same as a refresh course on these topics

- let alone an introductory course for those of you who have never been taught these topics

The reason is simple

- we don't have time for this

If you need an introductory or refresh course, there is plenty to find on the web. Even if you know these topics very well, watch the videos of (unfortunately deceased) Hans Rosling on Youtube

Probability Distribution

- we have a set of possible outcomes of an experiment, sometimes called the sample space, denoted by Ω
- a set of events E, where each $e \in E$ is a set of outcomes, i.e., $e \subseteq \Omega$; often $E = \mathcal{P}(\Omega)$
- and a (probability) function $\mathbb{P} : E \to \mathbb{R}$

If for the triple $(\Omega, E, \mathbb{P})$ the *Kolmogorov Axioms* hold we have a probability space:

1. $\mathbb{P}(e) \geq 0$
2. $\mathbb{P}(\Omega) = 1$
3. for an (in)finite set of events $\{e_i\}_{i \in I}$, if

$$\forall i, j \in I : e_i \cap e_j = \emptyset \to \mathbb{P}\left(\bigcup_I e_i\right) = \sum_I \mathbb{P}(e_i)$$

$\mathbb{P}$ is called a probability distribution on Ω

Formalizing Probability

Different from much of Maths, Probability Theory and Statistics have a very direct connection to our experience of reality

- this makes defining intuitive notions – like probability – highly contentious

For example, the question whether probabilities are inherently subjective or objective, i.e., real properties of objects in the world

- the objective view is, of course, greatly helped by quantum physics which states that there is true randomness at sub-atomic scales
- but, does that mean that all probabilities are objective?

Such philosophical issues were further complicated by purely mathematical concerns

- handling infinitely large sets is far from easy, only the invention/discovery of measure theory made this – for probability theory – relatively easy

The Axioms for probability theory were only laid down by Kolmogorov in 1933. Note that here are, of course, various (equivalent) formulations

Some Simple Properties

From the axioms, it follows easily (homework) that

- $\mathbb{P}(\emptyset) = 0$
- $e_1 \subseteq e_2 \rightarrow \mathbb{P}(e_1) \leq \mathbb{P}(e_2)$
- $\forall e \in E : \mathbb{P}(e) \in [0, 1]$
- $\mathbb{P}(e_1 \cup e_2) = \mathbb{P}(e_1) + \mathbb{P}(e_2) - \mathbb{P}(e_1 \cap e_2)$

Note that $\mathbb{P}(e_1 \cap e_2)$ is often written as $\mathbb{P}(e_1, e_2)$ or even $\mathbb{P}(e_1 e_2)$, Using this notation we have the more general property

$$\mathbb{P}\left(\bigcup_{i \in \{1,\ldots n\}} e_i\right) = \sum_{\emptyset \neq J \subseteq \{1,\ldots n\}} (-1)^{|J|-1} \mathbb{P}(e_J)$$

where e_J is a shorthand for the intersection of all e_j with $j \in J$.

The Union Bound

During this course we will often try to provide a bound

- usually an upper bound, but lower bounds are also interesting

on probabilities and expectations (we recall the definition later).

Often using some general tools.

The first is known as the *union bound* which follows directly from the last property on the previous slide. For any set of events $\{e_1, \ldots, e_n\}$:

$$\mathbb{P}\left(\bigcup_{i \in \{1,\ldots n\}} e_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(e_i)$$

Multidimensional Distributions

For us, Ω is often the multidimensional domain of some database (table), i.e.,

$$\Omega = \prod_{I \in I} D_i$$

in which the D_i are finite (or, at most recursive enumerable). For a $t \in \Omega$, $\mathbb{P}(t)$ denotes the probability of t. Being a bit more precise, we should talk about a

- *random* variable X (taking values in Ω)
- and the probability that $X = t$, i.e., $\mathbb{P}(X = \{t\})$,
- or simply $\mathbb{P}(X = t)$, or even $\mathbb{P}(t)$

From such a multidimensional distributions, we can construct *induced* probabilities

- by constraining (or fixating) (part) of t
- or by projecting (selecting) on a subset of I.

To do this properly, we need the notion of a *conditional* probability

Conditional Probabilities

$\mathbb{P}(e_1, e_2)$ denotes the probability that

- events e_1 and e_2 both occur

With $\mathbb{P}(e_1 \mid e_2)$ we denote the probability that

- event e_1 occurs given that we already know that e_2 occurs

This conditional probability is defined as:

$$\mathbb{P}(e_1 \mid e_2) = \frac{\mathbb{P}(e_1, e_2)}{\mathbb{P}(e_2)}$$

assuming $\mathbb{P}(e_2) > 0$, of course, but who would want to condition on things that don't occur?

Conditional Probabilities, Intuitively

In our (finite) multidimensional case, $\mathbb{P}$ is simply specified by

- a *look-up table* on Ω
- each cell, indexed by $d_1 \in D_1, d_2 \in D_2, \ldots, d_n \in D_n$, contains $\mathbb{P}(X = (d_1, d_2, \ldots, d_n))$

With this table as intuition, it is easy to see that

$$\mathbb{P}(e_1, e_2) = \mathbb{P}(e_1 \mid e_2) \times \mathbb{P}(e_2)$$

(to get to "cell" (e_1, e_2), you first go to column e_2 and then look for the intersection with row e_1 in that column). Which is, of course, completely equivalent to the definition

$$\mathbb{P}(e_1 \mid e_2) = \frac{\mathbb{P}(e_1, e_2)}{\mathbb{P}(e_2)}$$

(assuming $\mathbb{P}(e_2) > 0$ again, of course).

Bayes Theorem

We have that

$$\mathbb{P}(e_2 \mid e_1) \times \mathbb{P}(e_1) = \mathbb{P}(e_1, e_2) = \mathbb{P}(e_1 \mid e_2) \times \mathbb{P}(e_2)$$

That is

$$\mathbb{P}(e_2 \mid e_1) \times \mathbb{P}(e_1) = \mathbb{P}(e_1 \mid e_2) \times \mathbb{P}(e_2)$$

Dividing by $\mathbb{P}(e_1)$ yields

$$\mathbb{P}(e_2 \mid e_1) = \frac{\mathbb{P}(e_1 \mid e_2) \times \mathbb{P}(e_2)}{\mathbb{P}(e_1)}$$

Which is the celebrated theorem of Reverend Thomas Bayes.

This may look extremely simple, but conceptually it is amazing and thus hard to understand. It is also extremely useful and we'll see it quite often during the course

Another View on Bayes

Let $\{e_1, \ldots, e_n\}$ be a set of mutually disjoint events that together span all of Ω, i.e.,

- $\forall i, j : i \neq j \rightarrow e_i \cap e_j = \emptyset$
- $\cup_{i=1}^{n} e_i = \Omega$

and let A be some event with $\mathbb{P}(A) > 0$, then

$$\mathbb{P}(e_i \mid A) = \frac{\mathbb{P}(e_i \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(e_i) \times \mathbb{P}(A \mid e_i)}{\sum_{j=1}^{n} \mathbb{P}(e_j) \times \mathbb{P}(A \mid e_j)}$$

Exercise:

- 1 in 100 people have disease X
- if you have disease X, the chance of a positive test is 95%
- if you don't have the disease, the chance of a negative test is 95%

You test positive, what is the chance that you have the disease?

What?

The answer is

- 16%

This may seem disappointing. If these kind of numbers are realistic

- which they are

medical tests don't seem that useful.

- What happens if you do a second test?

If you do the exercise again, you'll see that the probability of having the disease after two independent positive tests is far higher.

Often a cheap test is used to screen out people with a very low chance of having a condition

- followed by a more expensive test for those that test positive

Which is a very good idea

- both economical
- and from a probabilistic perspective

Marginal Distributions

In the two-dimensional, we have $\mathbb{P}(X = t) = \mathbb{P}(X = (t_1, t_2))$, which we can equivalently denote by two random variables:

$$\mathbb{P}(X_1 = t_1, X_2 = t_2)$$

If we do not care about X_2 we can *marginalize* as follows:

$$\begin{aligned}\mathbb{P}(X_1 = t_1) &= \sum_{t_2} \mathbb{P}(X_1 = t_1, Y = t_2) \\ &= \sum_{t_2} \mathbb{P}(X_1 = t_1 \mid Y = t_2) \times \mathbb{P}(Y = t_2)\end{aligned}$$

That is, the probability that we see X_1 taking the value t_1 is

- the sum of the probabilities $\mathbb{P}(X_1 = t_1, X_2 = t_2)$, where
- X_1 always takes on the same value t_1
- and X_2 takes on all its possible values

Clearly we can do this with arbitrary numbers of variables

Independence

Sometimes conditioning does nothing, this is called independence. More precisely, two random variables X_1 and X_2 are independent if, knowledge of the one doesn't give you any information about the other, i.e., if

$$\mathbb{P}(X_1|X_2) = \mathbb{P}(X_1)$$

Note that Bayes law now immediately gives us that

$$\mathbb{P}(X_2|X_1) = \mathbb{P}(X_2)$$

also holds. If X_2 gives no information about X_1, than X_1 can give you no information about X_2.

Another equivalent, again by courtesy of Bayes law, way to introduce independence is by

$$\mathbb{P}(X_1, X_2) = \mathbb{P}(X_1) \times \mathbb{P}(X_2)$$

The Continuous Case

In this course we will mostly restrict ourselves to the finite case. For proofs and examples the continuous case is sometimes easier. In that case we do not have a probability distribution, but a probability density function

$$\mathbb{P}(a \leq x \leq b) = \int_a^b p(x)dx$$

Or, in the more general multi-dimensional case:

$$\mathbb{P}(x \in O) = \int_O p(\mathbf{x})d\mathbf{x}$$

The ubiquitous example being the normal distribution

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and its multi-dimensional generalisation

Back to Learning

With the terminology refreshed, we can discuss the basic assumption

$$D \sim \mathcal{D}$$

a bit deeper. It means that there is some *random process* going on that "outputs" (basic) events from $\mathcal{D} = (\Omega, E, \mathbb{P})$ according to their probability, i.e.,

$$\mathbb{P}(\text{output} = e) = \mathbb{P}_{\mathcal{D}}(X = e)$$

More precisely, there is a process which we *model* with the distribution. And in this model we describe structure, e.g.,

- that X_1 and X_2 are independent
- that X_1 depends on X_2 and X_3, e.g. by

$$X_1 = aX_2 + bX_3 + e$$

Simulating a Process

The process is not directly accessible, we only see the outputs. But using the distribution we can simulate the underlying process.

For example, in our running example (the finite multi-dimensional case), we can simulate it as follows

- assign each $e \in \Omega$ its unique (non-overlapping) interval $[e_l, e_u] \subseteq [0, 1]$, such that $e_u - e_l = \mathbb{P}_{\mathcal{D}}(X = e)$
- use a pseudo random number generator (PRNG) which outputs uniformly on $[0, 1]$
- and output the e for which PRNG $\in [e_l, e_u]$

This model doesn't exhibit explicit structure in the process

- it is called a multinomial distribution

If there is structure, we can give a simpler description

The Holy Grail of Learning

We have D, but the real interest is in $\mathcal{D}$:

- the data set is accidental (aleatory)
  - slightly more accurate, it as aleatory but (often) it reflects the epistemic structure (mostly)
- the distribution is the true (epistemic) structure

That is, only if we know $\mathcal{D}$ we can make predictions

Slightly more precise, we do not want any description of $\mathcal{D}$, we want

a succinct description of $\mathcal{D}$

because that allows us to actually *understand* what is going on.

The Holy Grail of Learning

We have D, but the real interest is in $\mathcal{D}$:

- the data set is accidental (aleatory)
  - slightly more accurate, it as aleatory but (often) it reflects the epistemic structure (mostly)
- the distribution is the true (epistemic) structure

That is, only if we know $\mathcal{D}$ we can make predictions

Slightly more precise, we do not want any description of $\mathcal{D}$, we want

a succinct description of $\mathcal{D}$

because that allows us to actually *understand* what is going on.

Unfortunately, the holy grail is unattainable

Induction

The OED defines induction (in the sense we use it) as

> *the process of inferring a general law or principle from the observation of particular instances*

in contrast with deduction where we (may) apply general laws to specific instances.

For deduction, well, at least for, say, First Order Logic, we can prove that it is sound

- if the premisses are true, so will be the conclusion

The question is if there is a similarly good procedure for induction, i.e., (Stanford Encyclopaedia of Philosophy):

> *can we justify induction; to show that the truth of the premise supported, if it did not entail, the truth of the conclusion.?*

This is known as *The Problem of Induction*

David Hume

Mid 18th century the philosopher David Hume argued

No! There is no justification for induction

There is no procedure that will always, guaranteed,

- give you the true general rule

Hume was actually more concerned with the more general induction problem

- conformity betwixt the future and the past

how do we know that regularity we have observed in the past will also be shown in the future

- (before Newton): will the Sun also rise tomorrow?

According to Hume all justifications are circular:

- the inductive step was successful yesterday, so it will also work today

Our Limited Inductive Problem

Our problem is that with a finite number of observations, many hypotheses are consistent, which one to choose?

Given a finite number of data points

- there are infinitely many functions that go through them

If your adversary gives you a number of data points

- and you guess the general rule, and predict the next data point

your adversary has enough leeway to think of another, consistent rule

- and generate a next data point that proves you wrong

no matter how many data points you have seen, and guesses you have made, you'll always give a wrong answer

So, our limited induction problem doesn't have a solution either.

Bummer!

Philosophers thought about this since at least the ancient Greeks

- Epicurus (300 BC) had the principle of multiple explanations
  - discard no hypotheses that is consistent with the observations
- William of Occam (1287 - 1347) had the principle of simplicity
  - Numquam ponenda est pluralitas sine necessitate (Plurality must never be posited without necessity)
  - But, then? Which one is the simplest?

In the end, scientists tend to be pragmatic

- and science and technology seem to do very well

and formulate criteria (such as simplicity) to pick a hypothesis

- we'll see examples of such criteria

In fact, both Epicurus's and Occam's ideas are still very much alive and we'll meet both – not necessarily in a form they would recognize

This Course

In this course we will study

- simple cases – make simple assumptions

And study how well we can learn $\mathcal{D}$ from D.

- which involves some non-trivial math

In fact, we will often look at a marginal distribution of $\mathcal{D}$

- aiming to induce *classifiers* from D

Functions that allow us to decide the *class* of new, unseen, cases

- e.g., to determine whether or not a new patient suffers from some disease.

And, most of all,

we'll study the effect of Big Data on this task