# ADA511 Data science and data-driven engineering

Steffen Mæland, PierGianLuca Porta Mana

2023-05-27

# Table of contents

# Preface

To be written.

# 1 Introduction

To be written: motivation and structure of this course.

# 2 Data: use and communication

## 2.1 Statements

What is “data”?

“Data” is often presented as numbers; but obviously it’s more than that. I give you this number: “5”. OK it’s a number, but is it “data”? What is it about? what should you do with it? Instead, if I tell you: “*The number of lectures in this course is 5*”, then we can say that I’ve given you data (even if it’s actually false data). So “data” is not just numbers. A number is not “data” unless there’s some verbal, non-numeric context associated with it – even if this context is only implicitly understood.

Data can also be completely non-numeric. A clinician saying “*The patient has fully recovered from the disease*” (we imagine to know who’s the patient and what was the disease) is giving us a piece of information that we could further use, for instance, to make prognoses about other, similar patients. The clinician’s statement surely is “data”. It is essentially non-numeric data, even if in some situations we can represent it as “1”, say, while “0” would represent “not recovered”.

From these two examples, and with some further thought, we realize that “data” – and in general any piece of information or hypothesis – can universally be represented and communicated by *statements* , also called *sentences* or *propositions*[1]. In some cases we can summarize or represent such statements as numbers. But the numbers alone, by themselves, are not data.

In these notes we use the words “data” and “information” as synonyms. So our conclusion is that *information* is represented by *statements.*

## 2.2 Well-posed and ill-posed statements

We face problems when the statements that should convey information and data are not clear. Suppose that an electric-car model consumes 150 Wh/km and has a range of 200 km; a second car model consumes 250 Wh/km and has a range of 600 km. Someone says “I think the second model is better; what do you think?”. It isn’t clear how we should answer; what does “better” mean? If it refers to consumption, then the first car model is “better”. If it refers to range,

[1] These terms are not equivalent in Logic, but we’ll use them as synonyms here.

then the second model is “better”. If it refers to a combination of these two characteristics, or to something else, then we simply can’t answer. Here we have a problem with querying and giving data, because the statement underlying such query is not clear.

We say that such statement are **not well-posed**, or that they are **ill-posed**.

This may seem an obvious discussion to you. Yet you’d be surprised by how often unclear statements appear in scientific papers about data engineering! Not seldom we find discussions and disagreements that actually come from unclear underlying statements, that two parties interpret in different ways.

As a data engineer, you’ll often have the upper hand if you are on the lookout for ill-posed statements. Whenever you face an important question, or you’re given an important piece of information, or you must provide an important piece of information, *always take a little time to examine whether the question or information is actually well-posed.*

- *[TODO] Exercise: give actual paper to analyse*

# 3 Truth inference and probability inference

## 3.1 Truth, falsity, and their consistency

## 3.2 Inferences without uncertainty: the truth calculus

## 3.3 Making room for uncertainty:Plausibility, credibility, degree of belief, probability

## 3.4 Inferences with uncertainty: the probability calculus

### 3.4.1 The Three Fundamental Laws of inference

- *Exercise: Monty-Hall problem & variations*
- *Exercise: clinical test & diagnosis*

### 3.4.2 Bayes's theorem

## 3.5 Common points of certain and uncertain inference

*No premises? No conclusions!*

# 4 Data and information

## 4.1 Kinds of data

### 4.1.1 Binary

### 4.1.2 Nominal

### 4.1.3 Ordinal

### 4.1.4 Continuous

- unbounded
- bounded
- censored

### 4.1.5 Complex data

2D, 3D, images, graphs, etc.

### 4.1.6 "Soft" data

- orders of magnitude
- physical bounds

## 4.2 Data transformations

- log
- probit
- logit

# 5 Allocation of uncertainty among possible data values: probability distributions

## 5.1 The difference between Statistics and Probability Theory

*Statistics* is the study of collective properties of collections of data. It does not imply that there is any uncertainty.

*Probability theory* is the quantification and propagation of uncertainty. It does not imply that we have collections of data.

## 5.2 What's "distributed"?

Difference between distribution of probability and distribution of (a collection of) data.

## 5.3 Distributions of probability

### 5.3.1 Representations

- Density function
- Histogram
- Scatter plot

Behaviour of representations under transformations of data.

## 5.4 Summaries of distributions of probability

### 5.4.1 Location

Median, mean

### 5.4.2 Dispersion or range

Quantiles & quartiles, interquartile range, median absolute deviation, standard deviation, half-range

### 5.4.3 Resolution

Differential entropy

### 5.4.4 Behaviour of summaries under transformations of data and errors in data

## 5.5 Outliers and out-of-population data

(Warnings against tail-cutting and similar nonsense-practices)

## 5.6 Marginal and conditional distributions of probability

## 5.7 Collecting and sampling data

### 5.7.1 "Representative" samples

Size of minimal representative sample = (2^entropy)/precision

- *Exercise: data with 14 binary variates, 10000 samples*

### 5.7.2 Unavoidable sampling biases

In high dimensions, all datasets are outliers.

Data splits and cross-validation cannot correct sampling biases

## 5.8 Quirks and warnings about high-dimensional data

# 6 Making decisions

## 6.1 Decisions, possible situations, and consequences

## 6.2 Gains and losses: utilities

### 6.2.1 Factors that enter utility quantification

Utilities can rarely be assigned a priori.

## 6.3 Making decisions under uncertainty: maximization of expected utility

# 7 The most general inference problem