# Strategic Innovation

## Data Science

• • •

Introducing the course

Dr. Philippe J.S. De Brouwer
2022, Warsaw, Poland
last compiled: November 7, 2023

eMBA at UW

# Table of Contents

# Introducing Ourselves

# Contact Philippe

**FIGURE 1:** Scan this QR-code to obtain Dr. De Brouwer's business card and connect via LinkedIn

# Who are you

**Question**

Who are you?
What do you expect from this program?
What do you want from this program?
What should we focus on?

# Availability of Slides and other materials

Course materials:

1. materials available
   `http://www.de-brouwer.com/students/uw.html`

Materials from “the Big R-Book” (De Brouwer (2020)):

1. videos and code: `http://www.de-brouwer.com/publications/r-book/index.html`
2. code and slides: `http://www.de-brouwer.com/publications/r-book/18901229-for-teachers.html`

# Objectives

# Objectives of the program

| | |
|---|---|
| know | the basics of statistics and data manipulation |
| know | at least one analytical tool (R) |
| understand | the importance of data in decision making |
| understand | uses and limits of various methods |
| apply | understand limits of models |
| apply | make informed decisions |
| apply | write a technical paper |
| apply | write a presentation and present it |

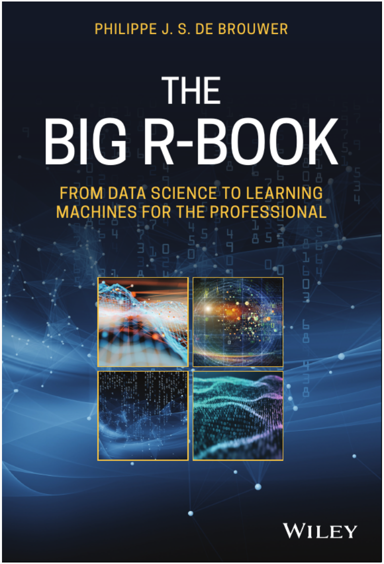
# The content of the program

# The content of the program

1. (optional) Getting started with R and its use
2. SQL Databases and importing data in R
3. Data wrangling (preparing data to build a model)
4. Building powerful models (linear regressions, generalised linear regression, non-linear regression, decision tree, random forest, SVN, neural network, etc.) and model validation
5. (optional) Introduction to companies and financial markets
6. Automating presentations, documents, etc.
7. (optional) Big Data
8. (optional) Code performance (speeding up R)
9. (extra/optional) Quantum Computers
10. (extra/optional) Crowdfunding and Fintech

# The Big R-Book

De Brouwer (2020)



# The grading of the program

# Grading of the course

- 50% presence and collaboration in classroom (including in-class assignments)
- 50% selected assignment (details see Chapter *Assignments*, page 20)

# Assignments

# Assignment

**Definition 1 (the assignment)**

Students are expected to gather data, analyse it and report the results in

1. a paper (between 5 and 50 pages)
2. a presentation (life in the classroom – during last course)

The format of the presentation (during the last lesson) is a “10 minutes elevator pitch” + 10 minutes questions

# The Grading of the assignment

- 40% for the idea, logic, coherence, and conclusions
- 30% for the written materials (paper and/or slides)
- 30% for the presentation itself (quality of slides if used + oratorical qualities)

The assignments are teamwork.

# Ideas for the content of the Assignment

1. solve a problem by using data and rely on one or more of the methods studied (eg. regression, MCDA), etc.
2. use an existing project/document/essay and improve the decisions proposed by using data

# Levels of Measurement

It is customary to refer to the theory of scales as having been developed by Stevens (1946). In that paper he argues that all measurement is done by assuming a certain scale type. He distinguished four different types of scale: nominal, ordinal, interval, and ratio scales.

# Nominal Scale

# Nominal Scale i

The nominal scale is the simplest form of classification. It simply contains labels that do not even assume an order. Examples include asset classes, first names, countries, days of the month, weekdays, etc. It is not possible to use statistics such as average or median, and the only thing that can be measured is which label occurs the most (modus of mode).

| **Scale Type** | **Nominal** |
|---|---|
| Characterization | labels (e.g. asset classes, stock exchanges) |
| Permissible Statistics | mode (not median or average), chi-square |
| Permissible Scale Transformation | equality |
| Structure | unordered set |

# Nominal Scale ii

**TABLE 1:** Characterization of the Nominal Scale of Measurement.

Note that it is possible to use numbers as labels, but that this is very misleading. When using an nominal scale, none of the traditional metrics (such as averages) can be used.

# Ordinal Scale

# Ordinal Scale i

This scale type assumes a certain order. An example is a set of labels such as very safe, moderate, risky, very risky. Bond rating such as AAA, BB+, etc. also are ordinal scales: they indicate a certain order, but there is no way to determine if the distance between, say, AAA and AA- is similar to the distance between BBB and BB-. It may make sense to talk about a median, but it does not make any sense to calculate an average (as is sometimes done in the industry and even in regulations)

# Ordinal Scale ii

| Scale Type | Ordinal Scale |
|---|---|
| Characterization | ranked labels (e.g. ratings for bonds from rating agencies) |
| Permissible Statistics | median, percentile |
| Permissible Scale Transformation | order |
| Structure | (strictly) ordered set |

**TABLE 2:** Characterization of the Ordinal Scale of Measurement.

Ordinal labels can be replaced by others if the strict order is conserved (by a strict increasing or decreasing function). For example AAA, AA-, and BBB+ can be replaced by 1, 2 and, 3 or even

# Ordinal Scale iii

by -501, -500, and 500,000. The information content is the same, the average will have no meaningful interpretation.

# section 8

# Interval Scale

This scale can be used for many quantifiable variables: temperature (in degrees Celsius). In this case, the difference between 1 and 2 degrees is the same as the difference between 100 and 101 degrees, and the average has a meaningful interpretation. Note that the zero point has only an arbitrary meaning, just like using a number for an ordinal scale: it can be used as a name, but it is only a name.

| **Scale Type** | **Interval Scale** |
|---|---|
| Characterization | difference between labels is meaningful (e.g. the Celsius scale for temperature) |
| Permissible Statistics | mean, standard deviation, correlation, regression, analysis of variance |
| Permissible Scale Transfor- | affine |

# Interval Scale ii

**TABLE 3:** Characterization of the Interval Scale of Measurement.

Rescaling is possible and remains meaningful. For example, a conversion from Celsius to Fahrenheit is possible via the following formula, $T_f = \frac{9}{5}T_c + 32$, with $T_c$ the temperature in Celsius and $T_f$ the temperature in Fahrenheit.

An affine transformation is a linear transformation of the form $\boldsymbol{y} = \boldsymbol{A}.\boldsymbol{x} + \boldsymbol{b}$. In Euclidean space an affine transformation will preserve collinearity (so that lines that lie on a line remain on a line) and ratios of distances along a line (for distinct collinear points $p_1, p_2, p_3$, the ratio $||p_2 - p_1||/||p_3 - p_2||$ is preserved).

In general, an affine transformation is composed of linear transformations (rotation, scaling and/or shear) and a translation (or “shift”). An affine transformation is an internal operation and several linear transformations can be combined into one transformation.

# Ratio Scale

# Ratio Scale i

Using the Kelvin scale for temperature allows us to use a ratio scale: here not only the distances between the degrees but also the zero point is meaningful. Among the many examples are profit, loss, value, price, etc. Also a coherent risk measure is a ratio scale, because of the property translational invariance implies the existence of a true zero point.

# Ratio Scale ii

| Scale Type | Ratio Scale |
| --- | --- |
| Characterization | a true zero point exists (e.g. VAR, VaR, ES) |
| Permissible Statistics | geometric mean, harmonic mean, coefficient of variation, logarithms, etc. |
| Permissible Scale Transformation | multiplication |
| Structure | field |

**TABLE 4:** Characterization of the Ratio Scale of Measurement.

# References

# References i


De Brouwer, P. J. (2020). *The Big R-Book: From Data Science to Learning Machines and Big Data*. New York: John Wiley & Sons, Ltd.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science 103*(2684), 677–680.

# Nomenclature i

MCDA Multi Criteria Decision Analysis, page 22