1. General Introduction

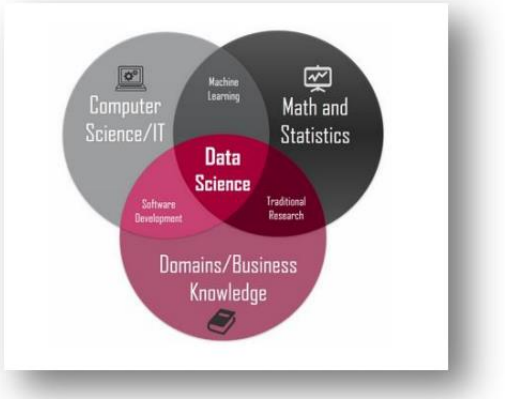
What is Data Science*?

Data science[1] is an interdisciplinary field whose ultimate purpose is to generate knowledge and extract insights from data for use in making sound business decisions and predictions. As defined in Wikipedia[2], **data science** "employs techniques and theories drawn from [...] mathematics, statistics [...] and computer science"/information technology, including signal processing, and uses probability modelling, machine learning, statistical learning, computer programming, data technology, pattern recognition, prediction, uncertainty modelling and data warehousing.

Accordingly, data science is defined as a subset of the following fields:

- Mathematics and statistics
- Computer science
- Domains/business knowledge

This is often represented as follows[1]:



By **analytics**[3], [4] we mean the process of reviewing, cleaning, transforming and modelling data to discover useful knowledge and information for decision-making purposes. In contrast to data science, analytics often does not require mathematical modelling or in-depth knowledge of computer science. For example, the analysis of standardised macroeconomic time series data of a country fall under analytics but not under data science, as no knowledge of computer science or statistical modelling is required.

"**Big data**[5] refers to data sets that are

- so big
- and complex",
- or so fast-changing or
- **so weakly** structured

"**that** [manual and] traditional [methods of data processing] are inadequate to deal with them."

Small data, according to Wikipedia[6], is data that is "small" enough for human comprehension. It is data in a volume and format that makes it accessible, informative and actionable.

By **structured data**, we mean data[7] "organised into a formatted repository, typically a [relational] database", where it is stored in rows and columns. **Unstructured data**[32] is digitised information that is not organised in a pre-defined format. Examples include image, text, video and voice data.

Statistical learning refers to a wide range of tools for modelling and understanding complex data and is a field of statistics and mathematics. **Machine learning**[8] "is a field of computer science that gives computers the ability to 'learn' with data without being explicitly programmed".

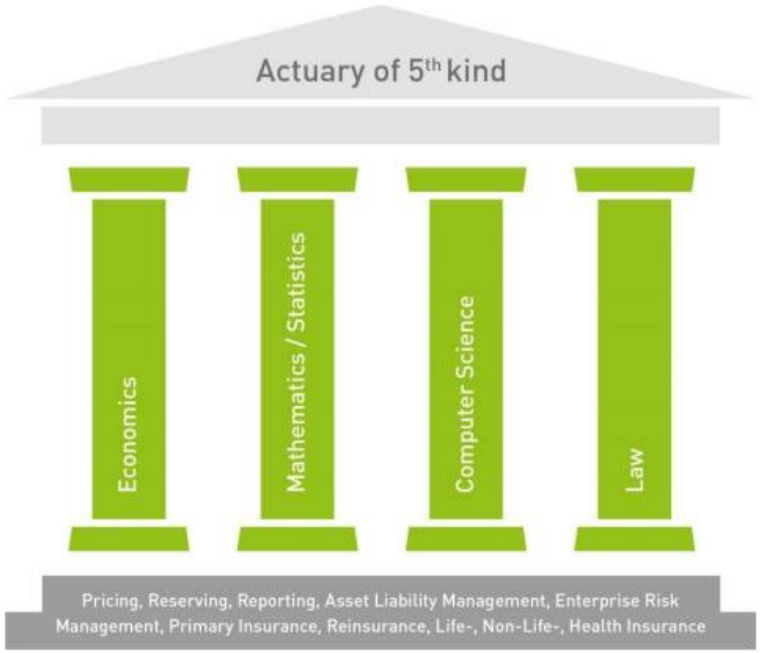
Predictive modelling[29] is a field of statistics and computer science in which statistical models are used in predicting events. The outcome to be predicted (such as a loss event) is most often in the future. Predictive modelling largely overlaps with statistical and machine learning.

*Source: «Data Science Strategy» (Link) of the Swiss Association of Actuaries (SAA)

Why does Data Science matters for actuaries*?

	Actuarial Science	Data Science
Basics	Mathematical Basics	
Data	Small Data	Small and Big Data
	Structured & static Data	Unstructured & dynamic Data
	Internal Data	External Data
Mathematics & Statistics	Probability Theory	Computational Statistics
	Life and Non-Life Insurance Mathematics	Algorithm
	Quantitative Risk Management	Information Theory
Computer science		Machine Learning & Visualisations
		Numeric Optimization
		Data Management
Programming languages	SAS, S Plus, R	Python, R
	SQL	SQL
	Excel / VBA	Julia, Spark, Scala
Domains / business knowledge	Reserving, Pricing	
	ERM, ALM, Solvency	
	Accounting, Economics, Law	

Typical course content and competences



*Source: «Data Science Strategy» (Link) of the Swiss Association of Actuaries (SAA)

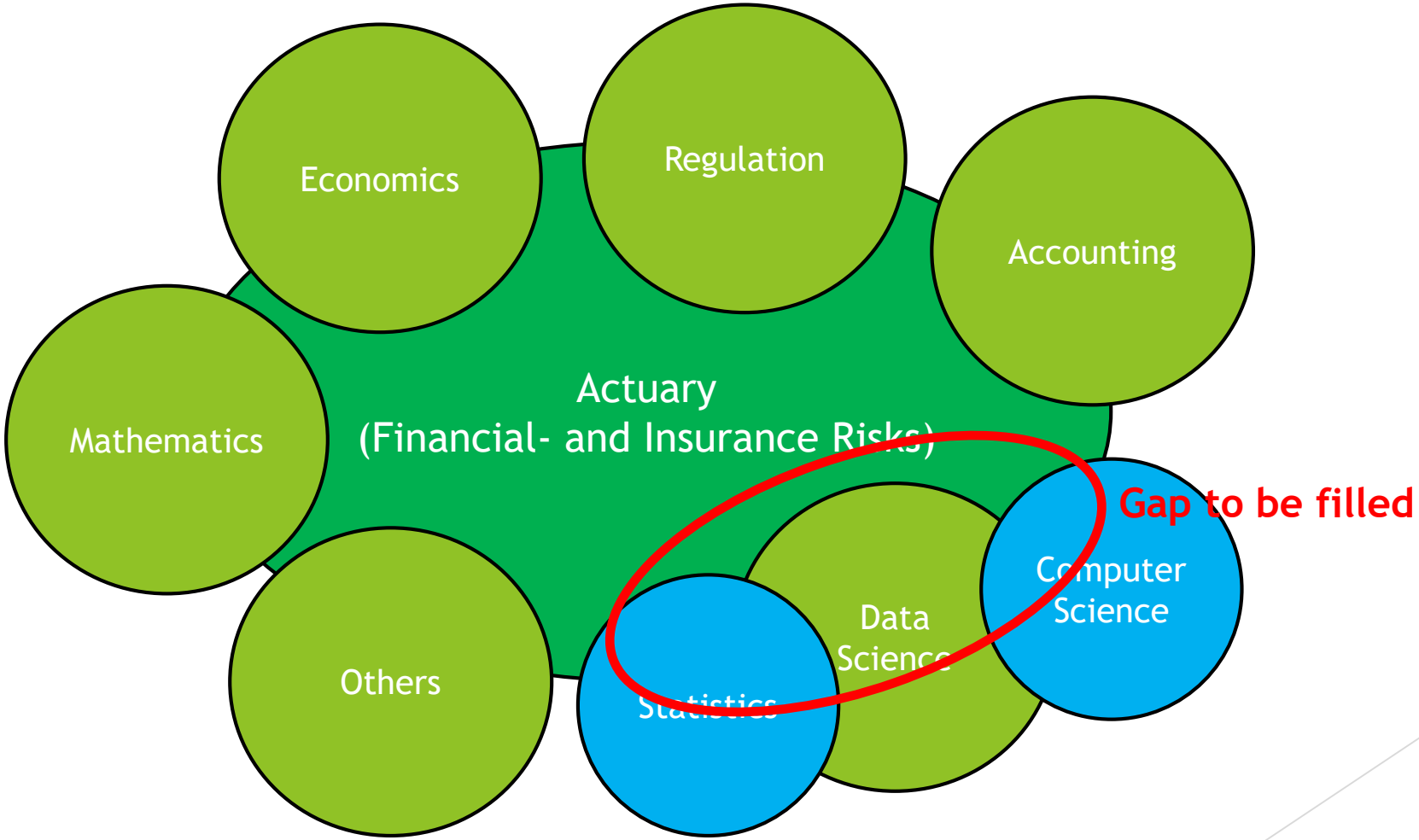
Strategy Swiss Association of Actuaries

Key action items:

- Revise basic education of actuaries, including (more) computer science and statistics.
- Upskill qualified actuaries in computer science and statistics.
- Provide courses and tutorials for upskilling.

Visit: www.actuarialdatascience.org

What does it mean for actuaries?



Why does R matters?

- R is a widely used programming software of Data Scientists
- Open source, extensively used at universities (teaching, research)
- R was built by mathematicians/statisticians
- R is very powerfull and appropriate for many data science tasks
- Ecosystem (Rstudio, Rshiny, Jupyter,…)

- Not so much about R, it is about programming and the way of working.

How to work with R?

There are various options for that:

- R console → No
- RStudio → Yes
- Jupyter Notebook (coding, documentation) → Yes
- R Markdown (coding, documentation, results) → Yes
- Rshiny (presenation, production) → Yes

Try out all of them and choose the best one (for you and the project). We will use RStudio for this training.

References

The training material is mainly based on the following literature:

- "Insurance Analytics, A Primer", 31th International Summer School of the Swiss Association of Actuaries (2018) (https://github.com/fpechon/SummerSchool)
- "Data Analytics for Non-Life Insurance Pricing" by M.V. Wüthrich and Ch. Buser (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2870308)
- "A Tutorial Introduction to R" by Aaron A. King, Stu Field, Ben Bolker, Steve Ellner (http://kingaa.github.io/R_Tutorial/)
- "Introduction to R" by Jonathan Cornelissen (https://github.com/datacamp/courses-intro-to-r)
- "An Introduction to R" by W. N. Venables, D. M. Smithand the R Core Team (https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf)
- "Introduction to Data Science with R", a video course by RStudio and O'Reilly Media (https://github.com/rstudio/Intro)
- "Case Study: French Motor Third-Party Liability Claims" from A. Noll, R. Salzmann and M.V. Wüthrich (https://www.actuarialdatascience.org/ADS-Tutorials/)
- www.actuarialdatascience.org