Actuarial Data Science: Opportunities and Challenges

Junge Aktuare, Swiss Association of Actuaries

8th March 2022

Dr. Jürg Schelldorfer, Actuary SAA

Senior Actuarial Data Scientist, Swiss Re

Chair of the «Data Science» working group of the Swiss Association of Actuaries (SAA)

Disclaimer

The opinions expressed in this presentation are those of the author only. They are inspired by the work that the author is doing for both Swiss Re and the SAA, but they do not necessarily reflect any official view of either Swiss Re or the SAA.

Table of Content

Introduction

Major Topics

FINMA-regulated insurance market

IFRS 17

Actuarial Data Science

Regulation

Low interest rates

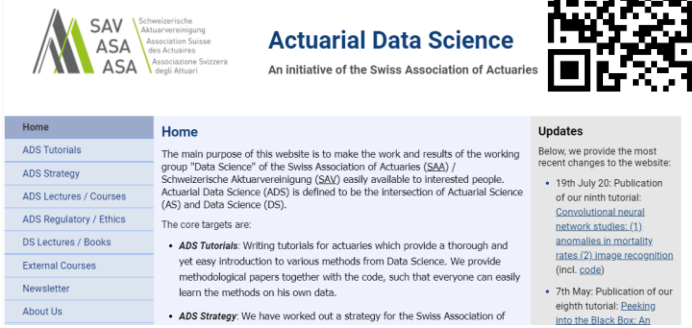
Sustainability

SAV Fachgruppe «Data Science»

– Anja Friedrich
– Frank Genheimer
– Thomas Hull
– Dr. Christian Lorentzen
– David Lüthi (Stv)
– Dr. Michael Mayer
– Dr. Daniel Meier (Stv)
– Dr. Jürg Schelldorfer (Leitung)
– Dr. Alessandro Torre
– Dr. Andreas Troxler
– Prof. Dr. Mario Wüthrich

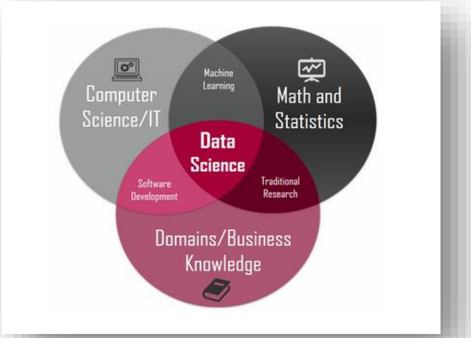
- …und viele weitere welche in den letzten 5 Jahren temporär mitgearbeitet haben.

www.actuarialdatascience.org



What is (Actuarial) Data Science?

Definition(s) und differences Data Science / Actuarial Science(1)

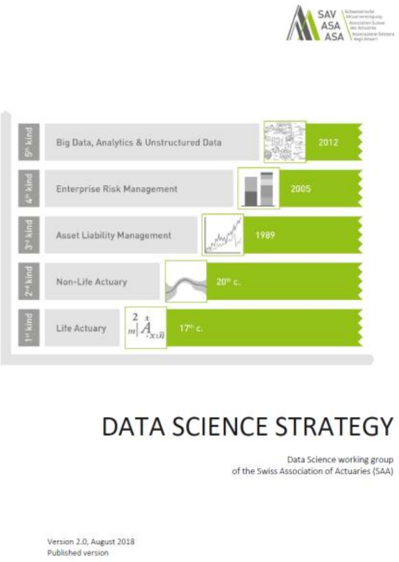


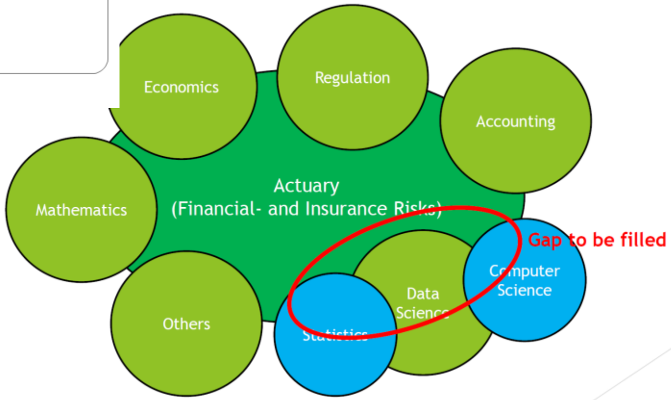
… intersection of …

- Mathematics and statistics
- Computer science / IT
- Domain / Business Knowledge

	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	

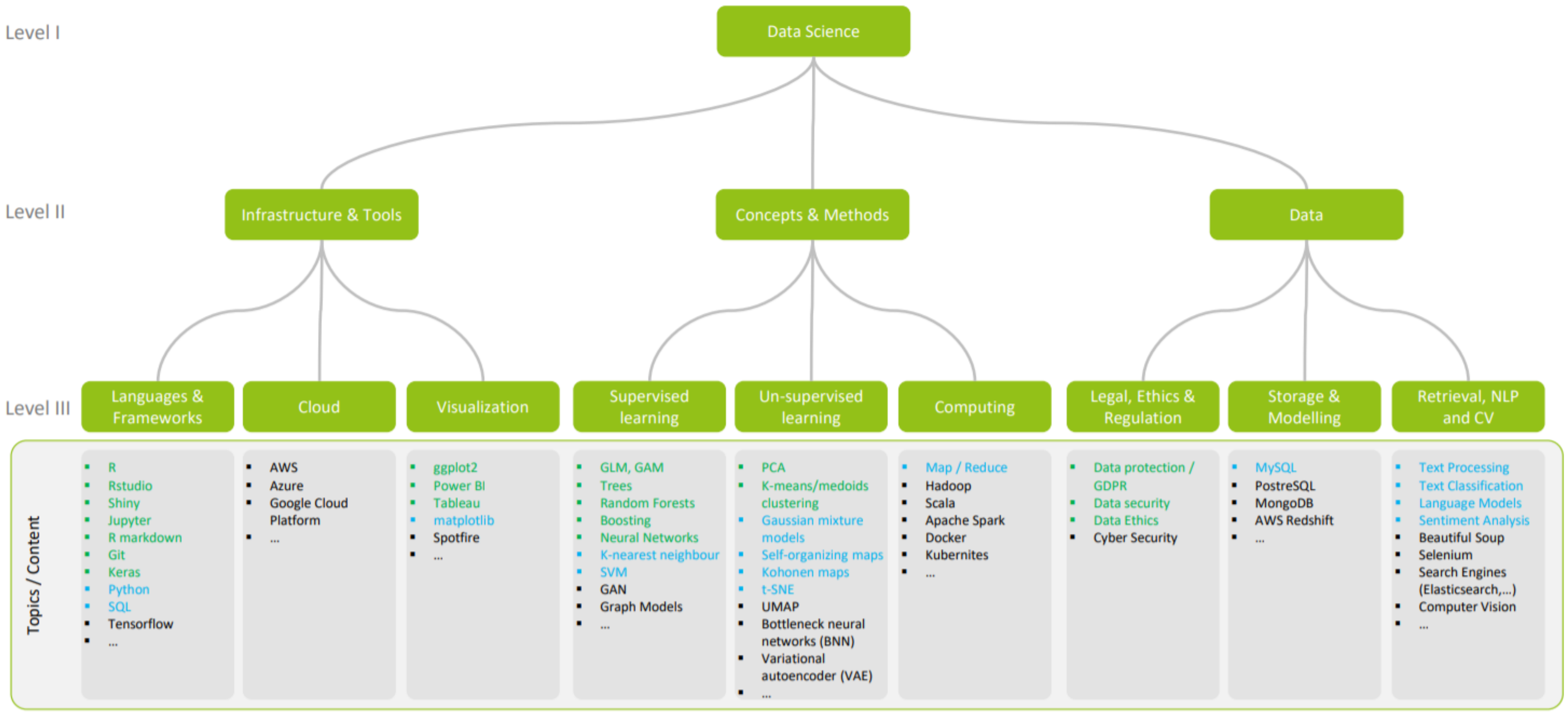
Typical course content and competences





(1) Aus: [Data Science Strategy, Data Science working group, SAA, Version 2.0, August 2018]

What is Data Science?



Green: Recommended for actuaries in the industry with some basic knowledge in data science

Blue: Recommended for actuaries in the industry with extended knowledge in data science (combined with green)

Black: Data science (combined with green and blue)

Version V1.1, Oktober 2019

Opportunities

1 – Unstructured data

- Actuaries are trained and used to work with tabular/structured data. The big(ger) amount of data in an insurance company are unstructured data (text, images, pdfs,…)

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.

- Unstructured data should and can be used by actuaries, as many technologies (e.g. extract text from pdf's) become a «commodity».

- Concrete Examples:
    - Free text fields to classify a claims to a specific claim type (e.g. glass, theft,…)
    - Extract information from structured pdf's
    - Extract daily COVID-19 cases from pdf's
    - Use text field to model the claims severities for worker's compensation claims data (Tutorial here)
    - Predict number of injured in car collisions from police reports <-> insurance claims (Tutorial to come)

2 – More data

- Having big data is not the standard for actuaries, it is still the exception. Insurance companies are not the big data owners, as they are in the second line of the industry (enabling business, not creating business).
- Design products and actively engage in product development. Ideas: The Geneva Assocation, Swiss Re
- Collaborations with third-party data provider

- Concrete Examples:
    - Telematics data
    - Vessel real-time data (Link)
    - Shipment goods data
    - Sensor data for property insurance (The Geneva Association: From Risk Transfer to Risk Mitigation)
    - Electronic Health Records (EHR) data
    - (Tracking data)
    - (…)

3 – Models and algorithms

- Advanced statistical and machine learning algorithms are **seamlessly** available through open source software libraries (e.g. Python, R)
- Commercial software providers have recognized the first point.
- Neural networks can be easily fitted with Python/R using a few lines of code.
- New is the seamlessly availability and the simple usage of the algorithms. The statistical models are not new.

- Concrete Examples:
  - Fitting a simple Neural Network using R/Python
  - Fitting Classification and Regression Tree, Random Forest, Boosting
  - Fitting clustering algorithms

R interface to Keras

Keras

R-CMD-check passing | CRAN 2.8.0 | license MIT

Keras is a high-level neural networks API developed with a focus on enabling fast experimentation. *Being able to go from idea to result with the least possible delay is key to doing good research.* Keras has the following key features:

- Allows the same code to run on CPU or on GPU, seamlessly.
- User-friendly API which makes it easy to quickly prototype deep learning models.
- Built-in support for convolutional networks (for computer vision), recurrent networks (for sequence processing), and any combination of both.
- Supports arbitrary network architectures: multi-input or multi-output models, layer sharing, model sharing, etc. This means that Keras is appropriate for building essentially any deep learning model, from a memory network to a neural Turing machine.

See the package website at https://tensorflow.rstudio.com for complete documentation.

https://keras.rstudio.com/

CRAN Task View: Machine Learning & Statistical Learning

Maintainer: Torsten Hothorn
Contact: Torsten.Hothorn at R-project.org
Version: 2022-03-02
URL: https://CRAN.R-project.org/view=MachineLearning

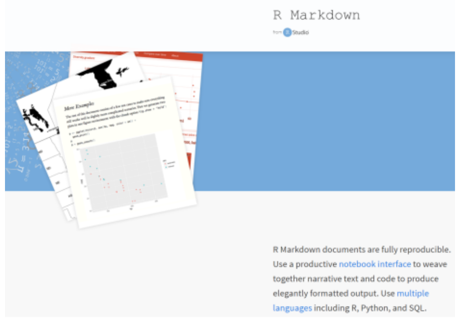
Several add-on packages implement ideas and methods developed at the borderline between computer scie

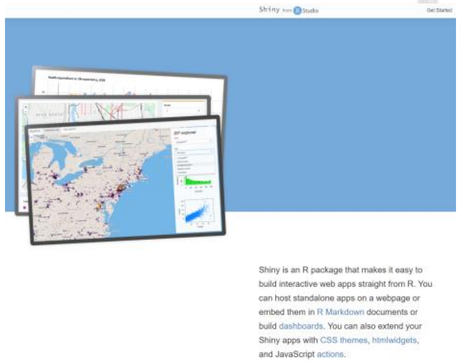
- *Neural Networks and Deep Learning* : Single-hidden-layer neural network are implemented in packa
  RcppDL (denoising autoencoder, stacked denoising autoencoder, restricted Boltzmann machine, dee
- *Recursive Partitioning* : Tree-structured models for regression, classification and survival analysis, f
  and M5. The Cubist package fits rule-based models (similar to trees) with linear regression models i
  Two recursive partitioning algorithms with unbiased variable selection and statistical stopping criteri
  distributions of the response are available in package party and partykit as well.
  Graphical tools for the visualization of trees are available in package maptree.
  Partitioning of mixture models is performed by RPMM.
  Computational infrastructure for representing trees and unified methods for prediction and visualizat
- *Random Forests* : The reference implementation of the random forest algorithm for regression and c
  inference trees is implemented in package party. randomForestSRC implements a unified treatment
  for random forest algorithms. In addition, packages ranger and Rborist offer R interfaces to fast C++
  Package RGF is an interface to a Python implementation of a procedure called regularized greedy fo
- *Regularized and Shrinkage Methods* : Regression models with some constraint on the parameter esti
  generalized linear models and Cox models can be obtained from functions available in package glm
  models). Package RXshrink can be used to identify and display TRACEs for a specified shrinkage p
  optional LASSO penalty to produce sparse solutions is implemented in package penalizedLDA. The
  an interface to the LIBLINEAR library. The ncvreg package fits linear and logistic regression model
  errors is estimated by hdm, inference on low-dimensional components of Lasso regression and of es
  are fitted by package mpath using composite optimization by conjugation operator. The islasso pack
- *Boosting and Gradient Descent* : Various forms of gradient boosting are implemented in package gb
  generalized linear, additive and nonparametric models is available in package mboost. Likelihood-ba
- *Support Vector Machines and Kernel Methods* : The function `svm()` from e1071 offers an interface to
  spaces can be estimated using rdetools which also offers procedures for model selection and predicti
- *Bayesian Methods* : Bayesian Additive Regression Trees (BART), where the final model is defined i
  by package tgp. Bayesian structure learning in undirected graphical models for multivariate continuo
- *Optimization using Genetic Algorithms* : Package rgenoud offers optimization routines based on gen
- *Association Rules* : Package arules provides both data structures for efficient handling of sparse bina
  data efficiently, in the form of self-sufficient itemsets, using either leverage or lift.
- *Fuzzy Rule-based Systems* : Package frbs implements a host of standard methods for learning fuzzy r
- *Model selection and validation* : Package e1071 has function `tune()` for hyper parameter tuning and
  other visualisation techniques for comparing candidate classifiers are available from package ROCR
- *Causal Machine Learning* : The package DoubleML is an object-oriented implementation of the dou
- *Other procedures* : Evidential classifiers quantify the uncertainty about the class of a test pattern usi
- *Meta packages* : Package caret provides miscellaneous functions for building predictive models, incl
  similar toolbox. The h2o package implements a general purpose machine learning platform that has
  machine learning algorithms, such as nearest neighbors, trees, random forests, and several feature se
- *GUI* rattle is a graphical user interface for data mining in R.
- *Visualisation (initially contributed by Brandon Greenwell)* The `stats::termplot()` function package
  renderings of the prediction function, are implemented in a few packages. gbm, randomForest and ra
  PDPs for a wide variety of machine learning models (e.g., random forests, support vector machines,
  less information. ICEbox focuses on constructing individual conditional expectation (ICE) curves, a

4 – Technology and computing power

- «2 – Big(ger) data» and «3 - Algorithms» are useful due to the available computing power to fit complex mathematical models.
- Many technologies/tools (see below) enable the development of machine learning models

- Concrete Examples:
  - Git (Link)
  - Jupyter notebooks (Link)
  - R Shiny Apps (Link): Deploy a dahsboard company-internally for management
  - R Markdown (Link)
  - Deploy and retrain models frequently
  - Deploy pricing model, re-trained every day (Parametric Flight Delay Insurance, Swiss Re)
  - Accident images
  - Satellite images

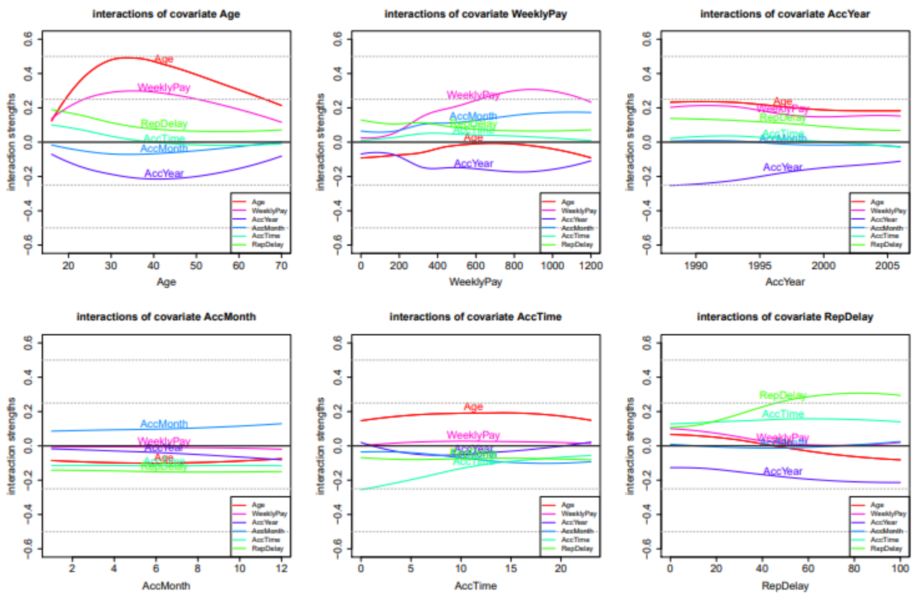




5 – Processes and efficiency

- Using new technological tools (see previous slides) enables actuaries to improve processes and increase their efficiency, and allows the actuaries to focus on the core.

- Concrete Examples:
  - R package(s) for calculating special reserves
  - Automating reports with R markdown (EAA training)
  - Find easily structure not yet caputed in a simple model (Tutorial here)
  - Speed up finding interactions in glms

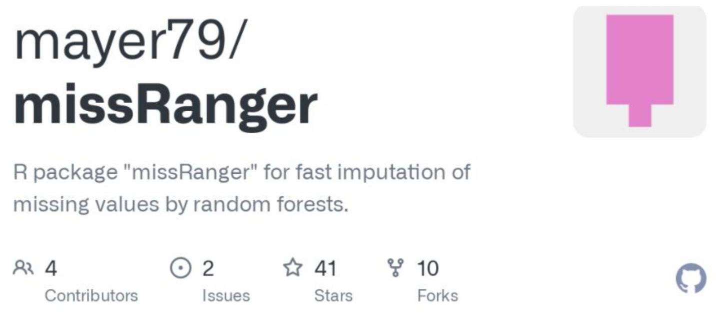


Challenges

1 – Data quality

- Data are the foundation of data science.
- The quality of the data remains as important as ever! «Garbage in – garbage out». If the data are not meeting the quality criteria (accuracy, appropriatness, completeness), there is no reason to draw any business conclusions from it.
- See Solvency II data quality definition.
- Data Governance
- Collecting / ensuring data quality can not be easily achieved → Data strategy
- Statistical / machine learning techniques can help to detect anomalies in data (anomaly, outlier detection).

- Concrete Examples:
    - Predict data points and compare with observed value (missForest, missRanger)
    - Data strategy
    - Data governance



2 – Usage and access to data

- Usage of some data is not allowed (legal, regulation)
- Having the appropriate / desired data is not the case
- Collaborations with other companies
- Buying external data

- Concrete Examples:
    - Submission data for a corporate insurer / reinsurer
    - The insured is the data owner, or has the granular data → information asymmetry

3 – XFT

- Product development means working in cross-functional teams (XFT), where actuaries / data specialists need to help design and define a product.
- Just doing the math/pricing is not enough.
- Interest beyond the actuarial core is required.
- Data quality and data collection needs time and care.

- Concrete Examples:
  - Shipment goods data
  - Swiss Re flight delay insurance
  - Swiss Re P&C Analytics (Link), see Impact+

4 – Varia

- Legacy in IT infrastructure and systems
- Legacy in mindset and lack of knowledge about opportunities/limits of Actuarial Data Science
- Sharing/Publish data is difficult than 10 years ago
- Sharing use cases is often not of interest
- Better/more complex algorithms are rarely the biggest challenge for an innovation.

- Concrete Examples:
  - No public text data on insurance claims
  - For learning, synthetic data are sufficient.
  - Synthetic data generators (Simulation Machine, …)

Take-home message

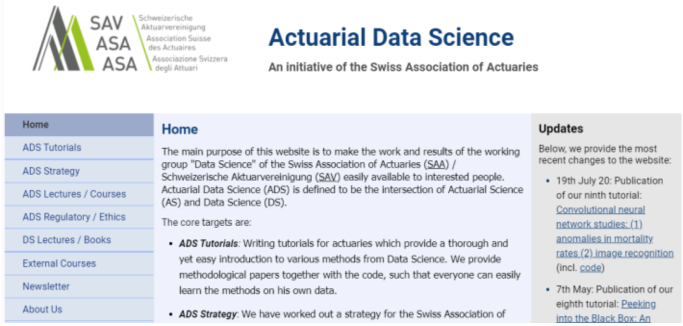
Conclusions

- Data Science != Actuarial Science
- Actuaries will collaborate with data scientists as they collaborate with IT, accounting, claims, underwriting,…
- Opportunities for additional data, better models and better toolkits
- Data quality/usage/access is a major challenge
- A very well calibrated GLM may still be as good as an advanced machine learning model in terms of accuracy.

Visit

For actuaries and data scientists in insurance

www.actuarialdatascience.org



Acknowledgements

People:

- All members of the SAA working group
- Dr. Alexander Noll
- Dr. Simon Renzmann
- Ron Richman

Insititutions:

- Swiss Association of Actuaries (SAA)
- RiskLab at ETH Zurich
- MobiLab for Analytics at ETH Zurich

Companies:

- Swiss Re

Appendix

ADS basics: Articles and repositories

The following articles/repositories are fundamental for entering the topic of Actuarial Data Science (ADS):

- Data Analytics for Non-Life Insurance Pricing, ETH Zurich, M.V. Wüthrich and C. Buser
- AI in Actuarial Science, R. Richman, SSRN, 2018
- ADS Tutorials, SAA, 2018-present
- Insurance Analytics – A Primer, International Summer School of the Swiss Association of Actuaries, 2018
- Insurance Data Science: Use and Value of Unusual Data, International Summer School of the Swiss Association of Actuaries, 2019

And do not forget the fundamentals of Statistics vs. Machine Learning:

- Statistical Modeling: The Two Cultures. L. Breimann, Statistical Science 16/3, 199-215, 2001
- To explain or to Predict?, G. Shmueli, Statistical Science 25/3, 289-310, 2010