From Official Statistics to Official Data Science

Mark van der Loo, Statistics Netherlands

CBS, Department of Methodology

Complutense University of Madrid, Spring 2019

UNIVERSIDAD COMPLUTENSE MADRID

Agenda

1. Why are computing skills important?
   - Some personal observations.
   - Experiences as a research methodologist
2. Official Statistics as a (Data) Science

Observations

Example one

Methodologist specifies

$$\text{mean}(x) = \frac{x_1}{\pi_1} + \frac{x_2}{\pi_2} + \cdots + \frac{x_n}{\pi_n}$$

Software developer implements

```
sum(x)/3.14
```

Example two

Methodologist specifies

$$\text{geometric\_mean}(x) = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$$

Software developer implements

```
geom_mean = function(x) prod(x)^(1/length(x))
```

Example two (continued)

Software developer tests implementation

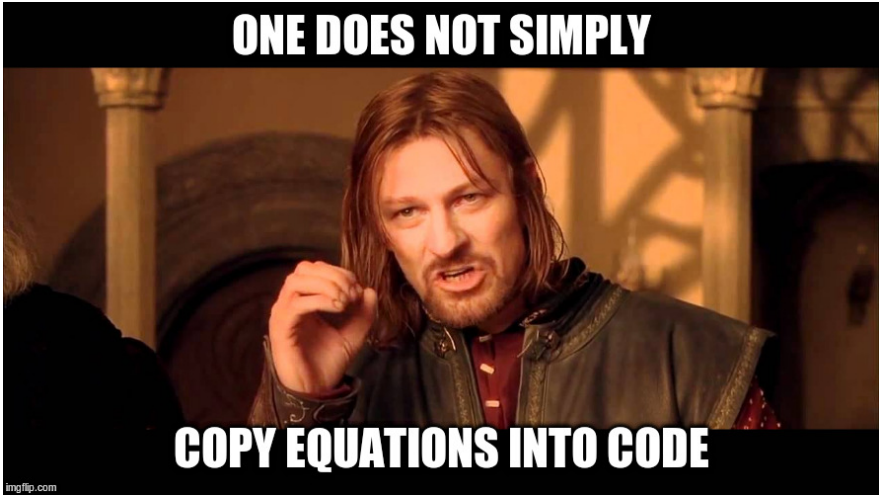
```
geom_mean(c(4,4)) == sqrt(16)
```

```
## [1] TRUE
```

User puts some actual data in: $1, 2, \ldots, 200$

```
geom_mean(1:200)
```

```
## [1] Inf
```


ONE DOES NOT SIMPLY
COPY EQUATIONS INTO CODE
imgflip.com

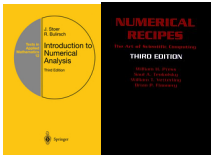
Lessons learned

Implementing methods is *not trivial*

It is called *scientific computing* or *numerical mathematics*, and it is a scientific field.

For (project) management in particular

You need to be able to recognize these situations to put the right person on the job.



A question to statistics managers

Your ‘computer person’ retires or leaves. You need to hire someone that will modernize the systems developed by this person.

a. What do you put in the job advertisement?
b. How do you interview this person to asses maturity in (statistical) programming?

A question for strategic management

Core question

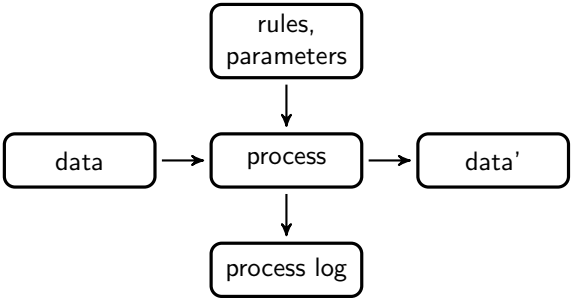
Do you think that statistical computing is a core competence for the statistical office?

and if so,

How much of it is needed (FTE)? Should there be associated career paths? ...

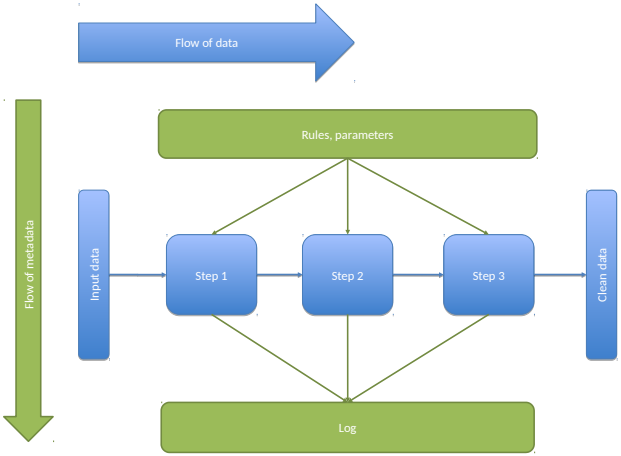
Experiences as a research methodologist

High-level process view (CSPA, GSIM)



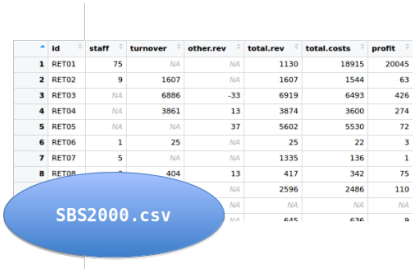
Separation of concerns + Modular approach

Slightly more realistic process view



Data cleaning using R-based packages (1)

```
library(validate)
SBS2000 <- read.csv("SBS2000.csv")
rules   <- validator(.file = "rules.R")
```



	id	staff	turnover	other.rev	total.rev	total.costs	profit
1	RET01	75	NA	NA	1130	18915	20045
2	RET02	9	1607	NA	1607	1544	63
3	RET03	NA	6886	-33	6919	6493	426
4	RET04	NA	3861	13	3874	3600	274
5	RET05	NA	NA	37	5602	5530	72
6	RET06	1	25	NA	25	22	3
7	RET07	5	NA	NA	1335	136	1
8	RET08	7	404	13	417	342	75
				NA	2596	2486	110
				NA	NA	NA	NA
				NA	645	636	9

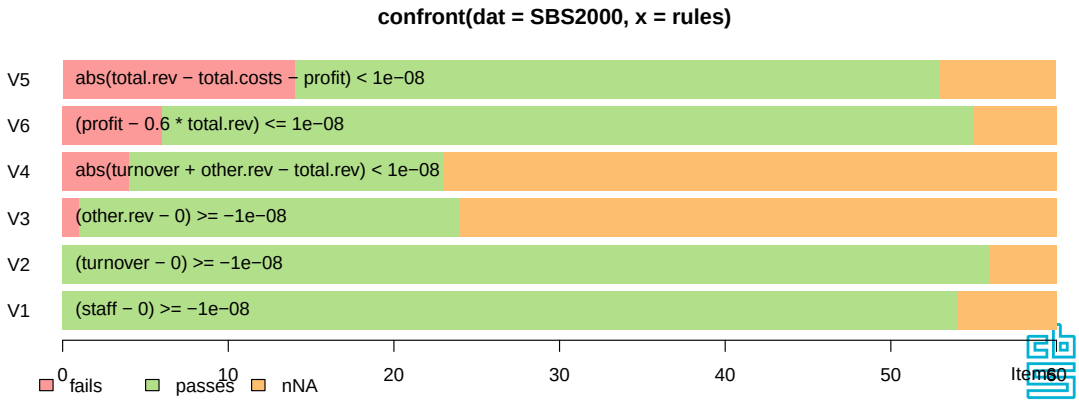
SBS2000.csv

```
1
2 # range resitrictions
3 staff >= 0
4 turnover >= 0
5 other.rev >= 0
6
7
8 # balance restrictions
9 turnover + other.rev == total.rev
0 total.rev - total.costs == profit
1 profit <= 0.6 * total.rev
2
```

rules.R

Data cleaning using R-based packages (2)

```
out <- confront(SBS2000, rules)
plot(out)
```



Data cleaning using R-based packages (3)

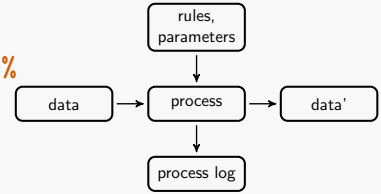
```
library(lumberjack); library(rspa);
library(simputation); library(errorlocate)

SBS2000 %L>%
    start_log( cellwise$new(key="id") ) %L>%
    replace_errors( rules ) %L>%
    tag_missing() %L>%
    impute_mf( . - id ~ . - id ) %L>%
    match_restrictions( rules, eps=1E-8 ) %L>%
    dump_log() -> clean_data
```

Data cleaning using R-based packages (3)

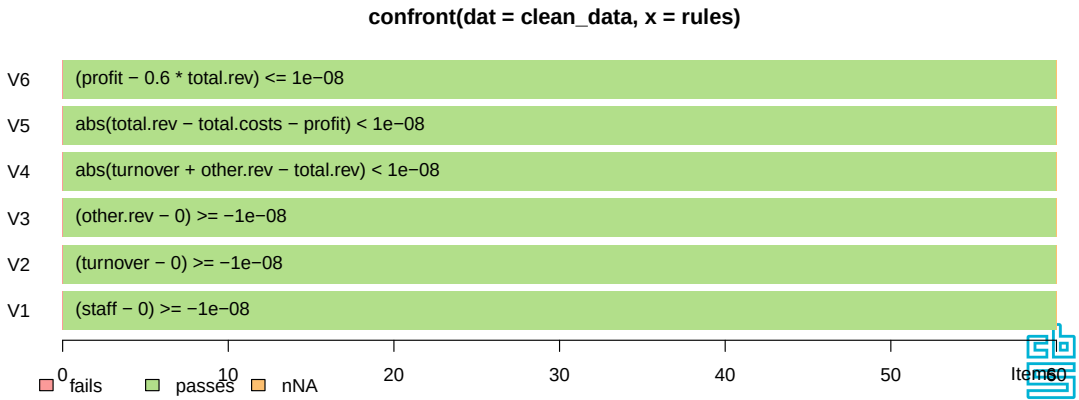
```
library(lumberjack); library(rspa);
library(simputation); library(errorlocate)

SBS2000 %L>%
    start_log( cellwise$new(key="id") ) %L>%
    replace_errors( rules ) %L>%
    tag_missing() %L>%
    impute_mf( . - id ~ . - id ) %L>%
    match_restrictions( rules, eps=1E-8 ) %L>%
    dump_log() -> clean_data
```



Data cleaning using R-based packages (4)

```
out <- confront(clean_data, rules)
plot(out)
```



Data cleaning using R-based packages (5)

```
read.csv("cellwise.csv") %L>% head(3)
```

```
##   step                    time            expression   key  variable  old
## 1    1 2019-05-10 11:05:31 CEST replace_errors(rules) RET01 total.rev 1130
## 2    1 2019-05-10 11:05:31 CEST replace_errors(rules) RET03 other.rev  -33
## 3    1 2019-05-10 11:05:31 CEST replace_errors(rules) RET07 total.rev 1335
##   new
## 1  NA
## 2  NA
## 3  NA
```

What went into this?

Methodology

Calculus, linear algebra, algorithm design, (convex) optimization, linear programming, formal logic, mathematical modeling.

Implementation

Parsing and language theory, functional programming, object orientation, numerical methods, algebraic data types. LOTS of programming experience, compiled languages, APIs and technical standards. Also: version control, documenting and testing, CI tools, UX design.

The Dolly Parton Principle

Dolly

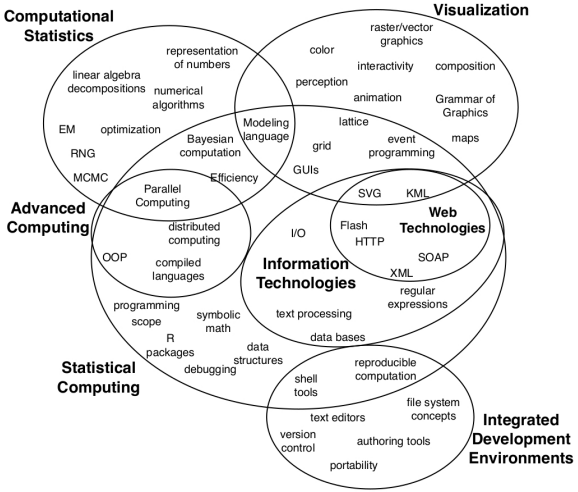
It takes a lot of money to look so cheap.

Me, writing software

It takes a lot of thinking to look so simple.

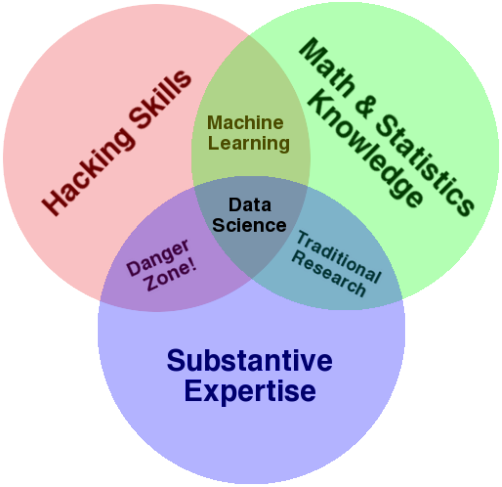
Official Statistics as a (Data) Science

Data science skill set



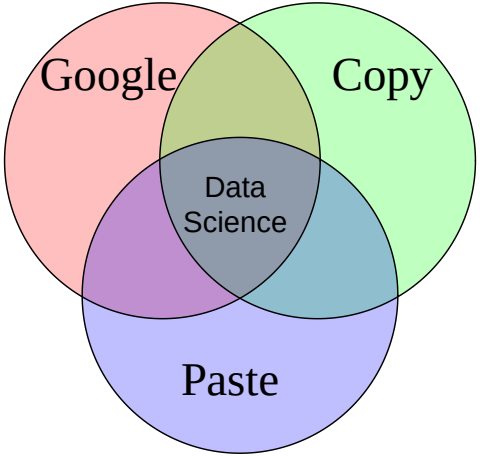
Nolan and Temple Lang (2010) *The American Statistician* **64**(2) 97–107

Data science skill set



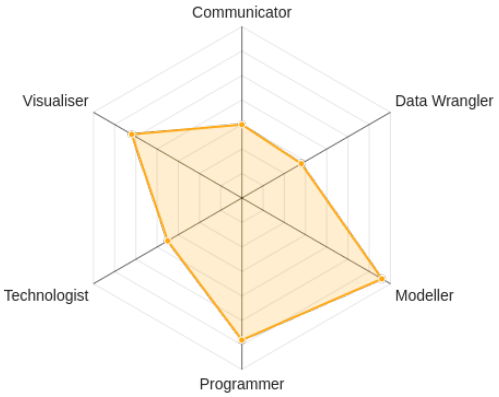
Drew Conway (2013) blog post

Data science skill set?



Me, reproduced from memory as seen at *The Internets*

Types of data scientists



Mango Solutions Data Science Radar

UNIVERSIDAD COMPLUTENSE MADRID

Data Science

Science of planning for, acquisition, management, analysis of, and inference from data.

StatNSF (2014); De Veaux *et al* 2017 *Annu. Rev. Stat.* **4** 15–31

Is data science a science?

[...] there is a solid case for some entity called 'Data Science' to be created, which would be a true science: facing essential questions of a lasting nature and using scientifically rigorous techniques to attack those questions

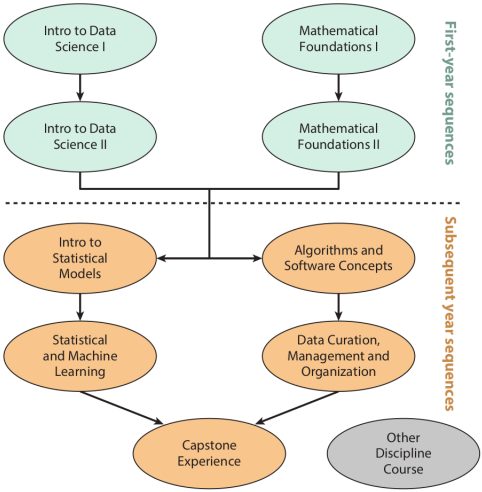
Donoho (2015) *50 years of data science*.

Key competencies of a data science major

1. Computational and statistical thinking
2. Mathematical foundations
3. Model building and assessment
4. Algorithms and software foundation
5. Data curation
6. Knowledge transference—communication and responsibility

De Veaux *et al* 2017 *Annu. Rev. Stat.* **4** 15–31

Curriculum



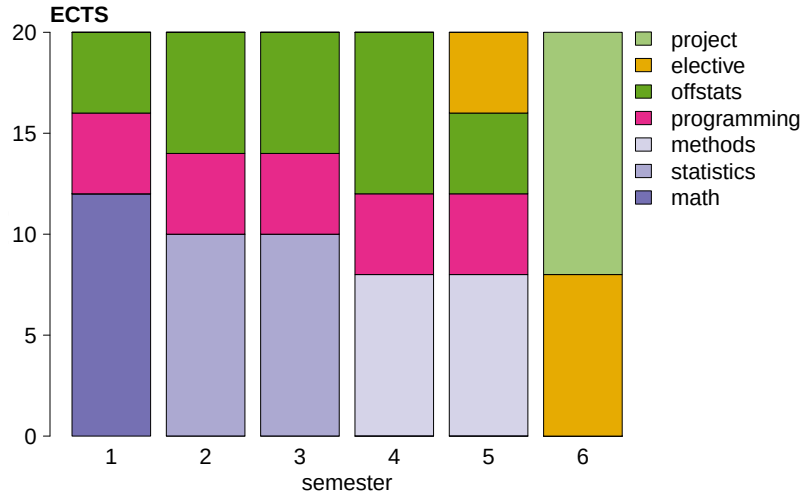
De Veaux *et al* 2017 *Annu. Rev. Stat.* **4** 15–31

UNIVERSIDAD COMPLUTENSE MADRID

Extra subject areas of an official statistics major

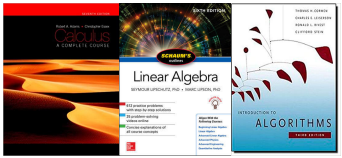
1. Macroeconomics
2. Demography
3. Ontologies and metadata
4. Policy, governance, international context
5. Privacy and data safety

Mark's Official Data Science Bachelors Curriculum



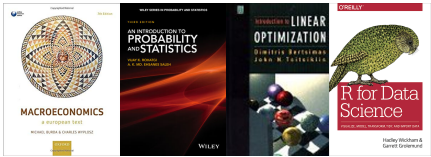
Semester I

- Calculus (6 ECTS)
  - Set theory, calculus on the real line, investigating functions (min, max, asymptotes), multivariate calculus, Lagrange multiplier method
- Linear algebra (6 ECTS)
  - Vectors and vector spaces, linear systems of equations and matrices, matrix inverse, eigenvalues, inner product spaces.
- Introduction to programming (4 ECTS)
  - Imperative programming, algorithm design, recursion, complexity, practical assignments.
- Public policy and administration (4 ECTS)
  - Government structure and institutions, policy-making and implementation, role of official statistics, international context, privacy



Semester II

- Probability and statistics I (6 ECTS)
  - Probability, discrete and continuous distributions, measures of location and variation, Bayes' rule, sampling distributions, estimation of mean and variance, CLT, ANOVA, linear models.
- Linear programming and optimization (4 ECTS)
  - Recognizing and modeling LP problems, simplex method, duality, sensitivity analysis, intro nonlinear optimization. Practical assignments using software tools.
- Programming with data I (4 ECTS)
  - Statistical analysis, data visualisation and reporting, programming skills and reproducibility, version control, testing, project.
- Macroeconomics (6 ECTS)
  - National Accounts, economic growth, labour market, consumption and investments, inflation, macro-economic equilibrium, budget policy and government debt. The main surveys.



UNIVERSIDAD COMPLUTENSE MADRID

Semester III

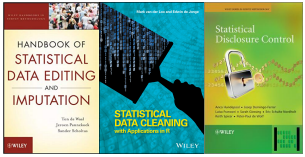
- Models in computational statistics (6 ECTS)
  - GLM, regularization, Tree models, Random Forest, SVM, unsupervised learning, model selection, lab with practical assignments.
- Probability and statistics II (4 ECTS)
  - Bayesian inference, Gibbs sampling and MCMC, maximum likelihood and Fisher information, latent models
- Programming with data II (4 ECTS)
  - Relational algebra and data bases, data representation, regular expressions, and technical standards, ontologies and metadata, practical assignments.
- Demography (6 ECTS)
  - Fertility, mortality, life table and decrement processes, age-specific rates and probabilities, stable and nonstable population models, cohorts, data and data quality. The main surveys.

Semester IV

- Methods for official statistics I (4 ECTS)
  - Advanced survey methods, weighting and estimation, calibration, SAE, handling non-response
- Methods for official statistics II (4 ECTS)
  - Time series, seasonal adjustment, benchmarking and reconciliation, time series models
- Programming with data III (4 ECTS)
  - Infrastructure for computing with big data, map-reduce, key-value stores, project.
- Communication (4 ECTS)
  - Scientific and technical writing, principles of visualization, dissemination systems.
- Ethics and philosophy of science (2 ECTS)

Semester V

- Methods for official statistics III (4 ECTS)
  - Principles of data editing, Fellegi-Holt error localization, methods for imputation.
- Methods for official statistics IV (4 ECTS)
  - Information Security and Statistical Disclosure Control
- Research methods in social science (4 ECTS)
  - Questionnaire design and field research, measurement models and latent variables
- Elective course (4 ECTS)
  - In the area of social science, economics, econometrics, computer science, or math&statistics
- Large programming project (4 ECTS)
  - E.g. a small production system, a dashboard, data cleaning system



Semester VI

- Elective courses (8 ECTS)
  - Preparing for thesis research
- Bachelor's thesis (12 ECTS)
  - Research in Macroeconomy, Demography, or Methodology. Preferably at an NSI or international organization.

Some interesting research areas

Methodology

- Complexity theory, econophysics, agent-based modeling
- Network theory
- Streaming data

Content / output

- Beyond GDP
- Globalization, regionalization
- SDG, energy transition

Take-home message

Official statistics is a (data) science, applied to society.