

Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#000 Agenda

été 2020



<https://gitlab.com/inf7100>

Bio

- ▶ PhD Applied Maths (KU Leuven)
- ▶ Habilitation à Diriger des Recherches (Université de Rennes)
- ▶ Professor (Economics Dept) (Université de Rennes, École Polytechnique)
- ▶ Professor (Mathematics Dept) (Université du Québec à Montréal)
- ▶ Director (Institute of Actuaries) (Data Science for Actuaries Program)
- ▶ Blog editor
<http://freakonometrics.hypotheses.org/>



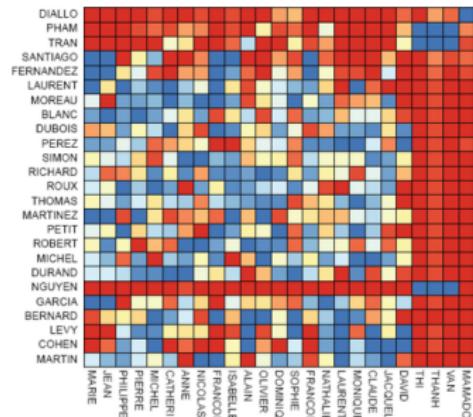
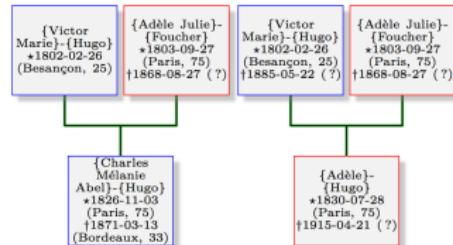
Bio

Research projects

- ▶ Machine Learning
- ▶ Reinforcement Learning
- ▶ Genealogie historique
- ▶ We are not alone!
- ▶ Tents, Tweets, and Events

Blog

- ▶ <http://freakonometrics.hypotheses.org/>
- ▶ Vulgarisation



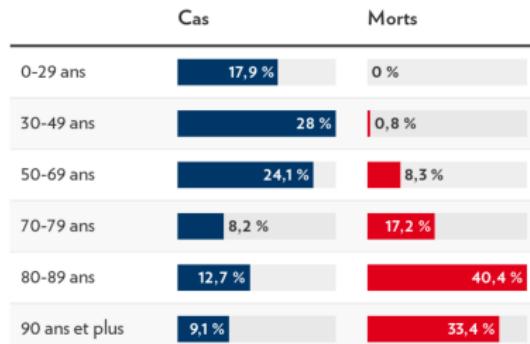
Strong positive correlation Strong negative correlation

Data Science

60 % des cas au Québec sont des femmes



La COVID-19 au Québec par groupes d'âge



Pour les cas confirmés et les décès dont l'âge est connu.

via **60% des cas au Québec sont des femmes**, La Presse, May 2020

Garbage

Two aspects : **data** and **model**

Data: “*Garbage in, garbage out*”
see [Face Recognition on Flawed Data](#),
by Clare Angelyn or [xkcd](#)

Model: machines do not understand
see [Why Should I Trust You?](#) or
[Recognition in Terra Incognita](#)



(A) Cow: **0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98



(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97



(C) No Person: 0.97, Mammal: **0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

$$\text{PRECISE NUMBER} + \text{PRECISE NUMBER} = \text{SLIGHTLY LESS PRECISE NUMBER}$$

$$\text{PRECISE NUMBER} \times \text{PRECISE NUMBER} = \text{SLIGHTLY LESS PRECISE NUMBER}$$

$$\text{PRECISE NUMBER} + \text{GARBAGE} = \text{GARBAGE}$$

$$\text{PRECISE NUMBER} \times \text{GARBAGE} = \text{GARBAGE}$$

$$\sqrt{\text{GARBAGE}} = \text{LESS BAD GARBAGE}$$

$$(\text{GARBAGE})^2 = \text{WORSE GARBAGE}$$

$$\frac{1}{N} \sum (\text{N PIECES OF STATISTICALLY INDEPENDENT GARBAGE}) = \text{BETTER GARBAGE}$$

$$\left(\text{PRECISE NUMBER} \right)^{\text{GARBAGE}} = \text{MUCH WORSE GARBAGE}$$

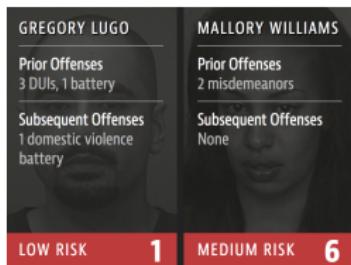
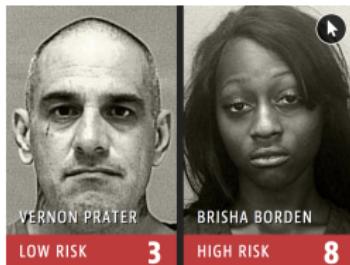
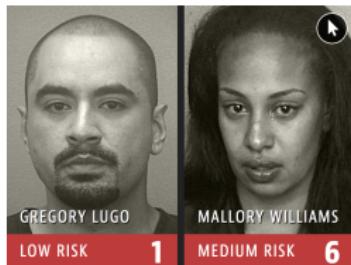
$$\text{GARBAGE} - \text{GARBAGE} = \text{MUCH WORSE GARBAGE}$$

$$\frac{\text{PRECISE NUMBER}}{\text{GARBAGE} - \text{GARBAGE}} = \text{MUCH WORSE GARBAGE, POSSIBLE DIVISION BY ZERO}$$

$$\text{GARBAGE} \times 0 = \text{PRECISE NUMBER}$$

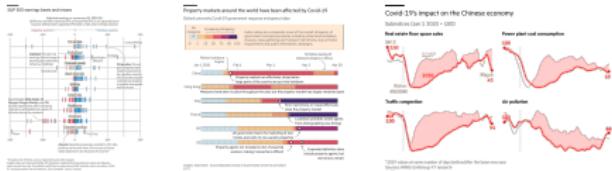
Machine Bias

see Machine Bias (Propublica, 2016), on *Correctional Offender Management Profiling for Alternative Sanctions* ([compas](#))

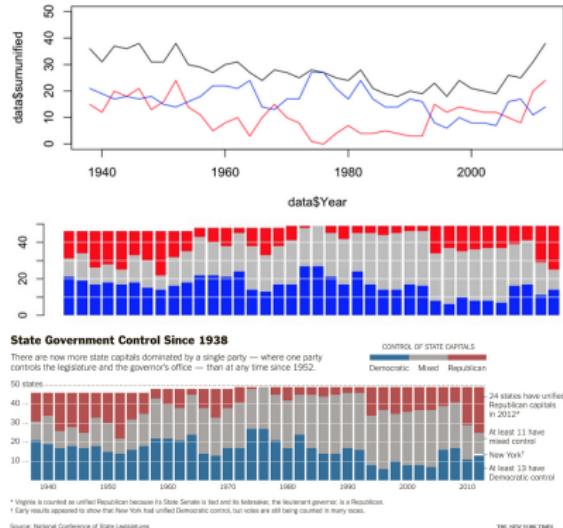
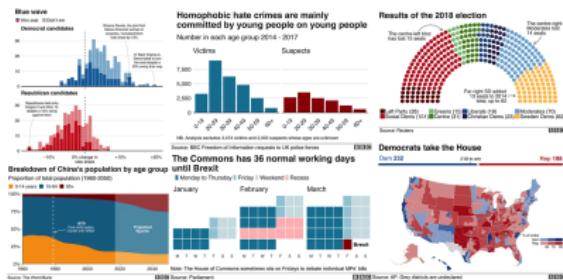


Data Science & Data Visualization

See [FT.com](#),



Going from raw data to some analysis, see [BBC Visual](#)



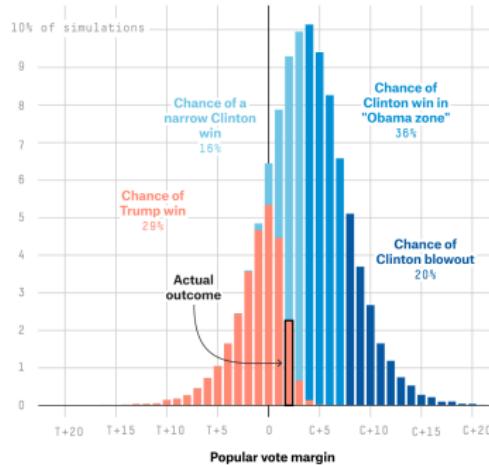
see [Amanda Cox's interview](#),
NYT Graphics Editor

Data Science

“The media’s demand for certainty – and its lack of statistical rigor – is a bad match for our complex world” The Media Has A Probability Problem, Nate Silver (2017).

FiveThirtyEight's final forecast for 2016

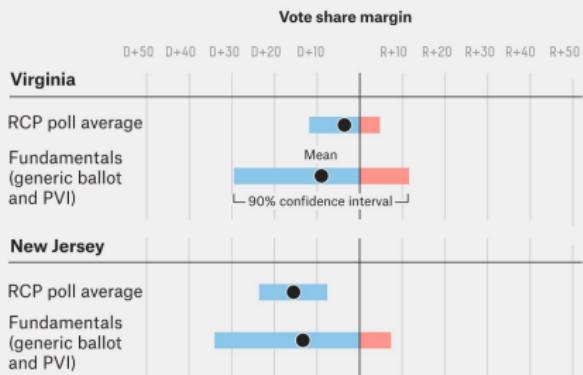
Likelihood of popular vote outcomes according to FiveThirtyEight's polls-only model at 9:35 a.m. on Election Day 2016. Based on 20,000 simulations.



FiveThirtyEight

A wide range of outcomes are possible in Virginia

Projected vote share margin in Virginia and New Jersey's upcoming gubernatorial elections



Error calculations are based on the historical accuracy of polling averages in gubernatorial races since 1998, and the historical accuracy of the “fundamentals” model in gubernatorial elections since 2001.

FiveThirtyEight

SOURCE: REALCLEARPOLITICS

Agenda

1. Observation and Experiment : on data
2. Visualization and Statistical Indicators
3. Correlation and Regression
4. Mathematics

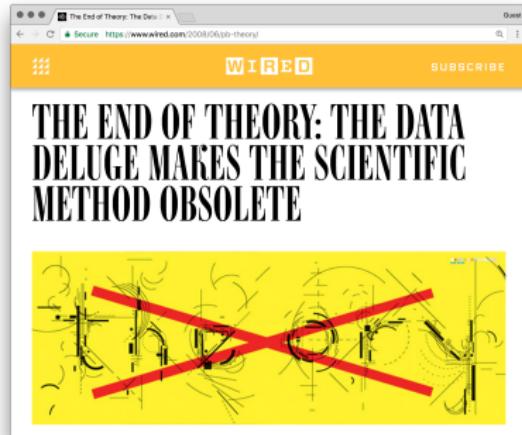
“Begin at the beginning,” the King said, gravely, ‘and go on until you come to an end, then stop” Lewis Carroll (1865, Alice’s Adventures in Wonderland)



End of Theory?

Chris Anderson's **The End of Theory: The Data Deluge Makes the Scientific Method Obsolete** (in 2008)

(see also Fulvio Mazzocchi's few remarks on the epistemology of data-driven science)



generalization is both advantageous and disadvantageous, in that while it can make a claim less concrete and less visualizable, it can also make it more powerful and more applicable as a result
(via **The Definitive Glossary of Higher Mathematical Jargon**)

Mathematics? (\neq calculus)

In order to better understand

- ▶ logarithm
- ▶ derivatives, integrals
- ▶ optimisation
- ▶ vectors, matrices
- ▶ projections
- ▶ probabilities

log : multiplicative \rightarrow additive

exp : additive \rightarrow multiplicative

E : average value

\int : sum vs. $\frac{\partial}{\partial x}$: difference

GIVEN THE PACE OF
TECHNOLOGY, I PROPOSE
WE LEAVE MATH TO THE
MACHINES AND GO PLAY
OUTSIDE.



Mathematics...

There will be a few slides on mathematical aspects of data science and machine learning...

we want the best model, but it should not be too complicated

$$\min_{m \in \mathcal{M}} \left\{ \text{criteria}(m) \right\}$$

complexity(m) \leq bound

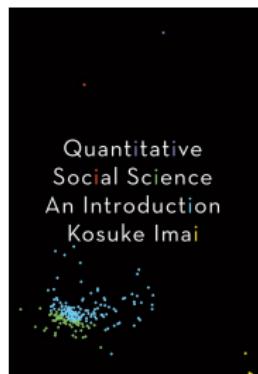
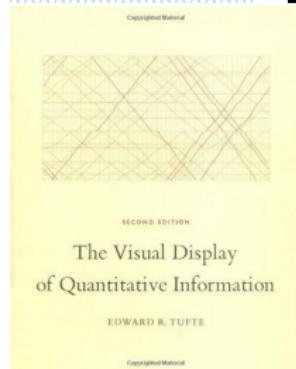
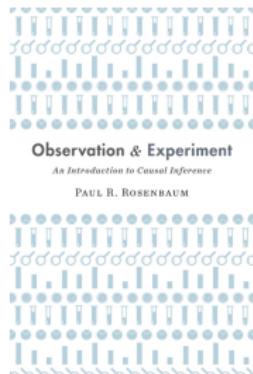
(see constrained optimization) or

find the best 3 clusters , $\begin{cases} \text{between each cluster, maximal heterogeneity} \\ \text{within each cluster, minimal heterogeneity} \end{cases}$

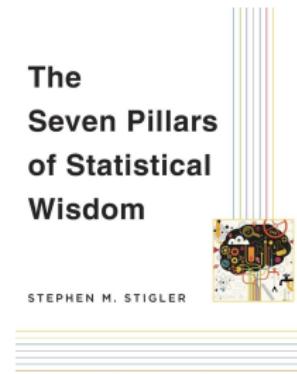
or log scales and areas in data-visualisation...

Word vectorization, datasets as matrices, pictures as tensors, etc.

References



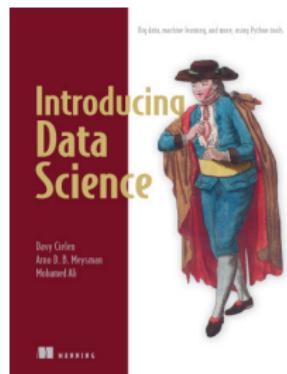
Edward R. Tufte
Envisioning Information



STEPHEN M. STIGLER



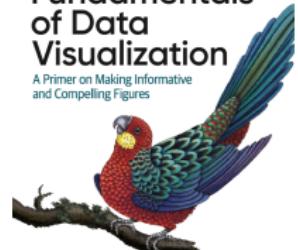
KIERAN HEALY



O'REILLY®

Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures



See **Fundamentals of Data Visualization** (on R)