PROMISES AND PROBLEMS OF DATA SCIENCE FOR CRIME SCIENCE

ADVANCED CRIME ANALYSIS UCL

BENNETT KLEINBERG

4 MARCH 2019

UCL

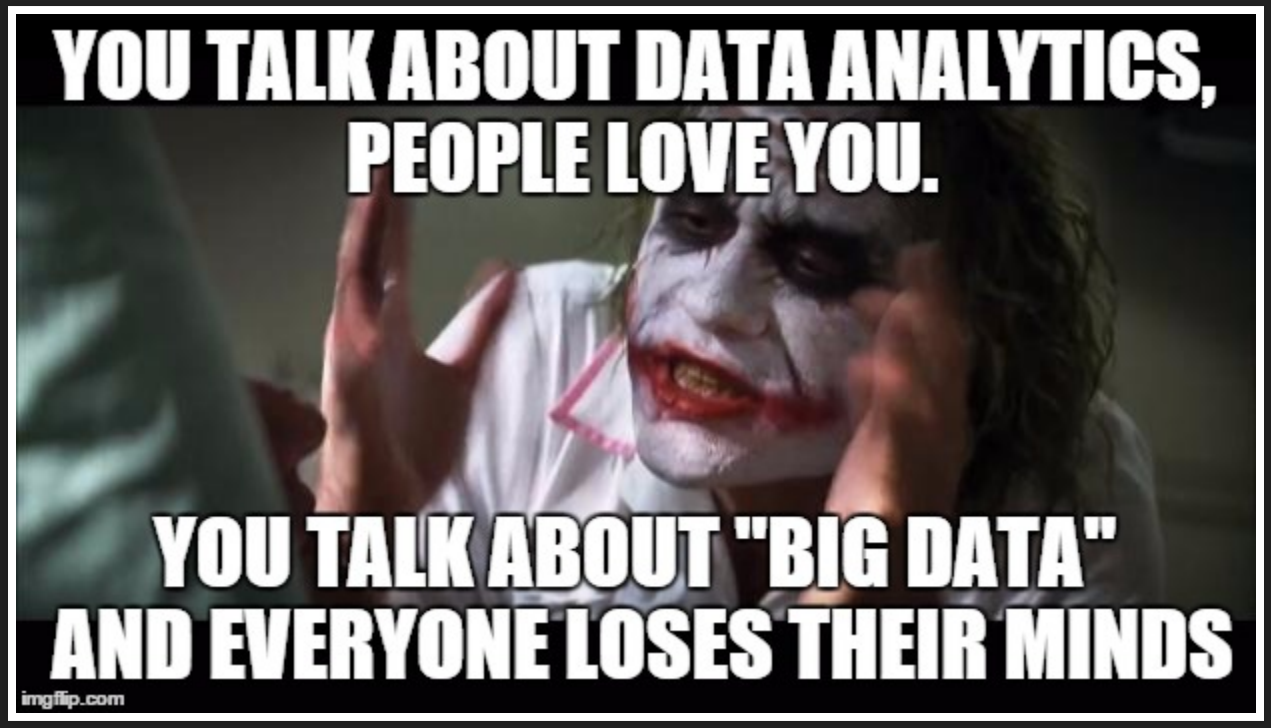
Advances, Promises and Problems

TODAY

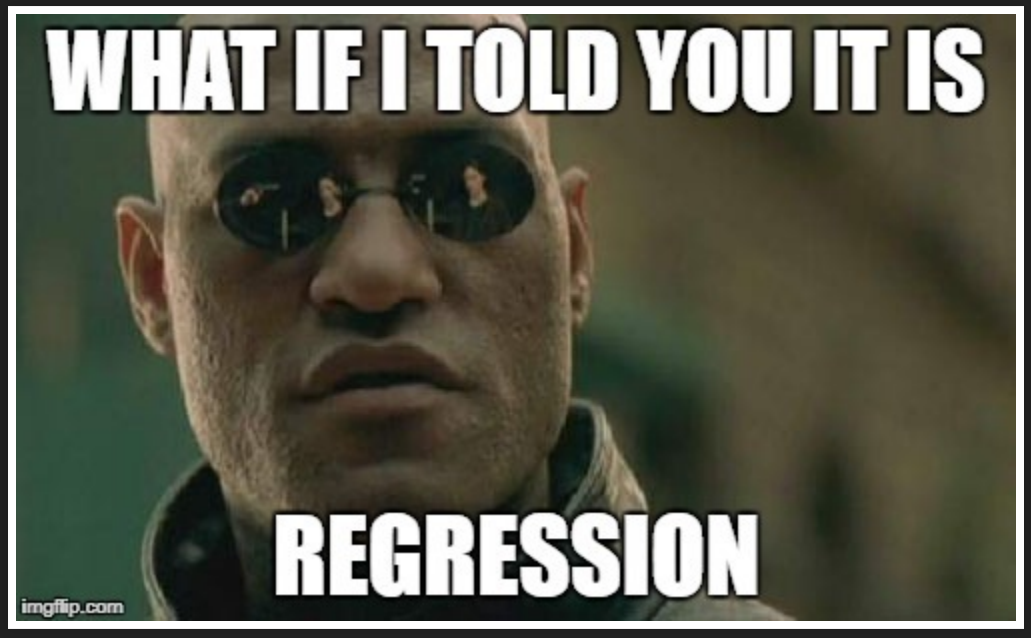
- problematic trends in data science
- fallacies in data science
- ethical considerations of data science for crime scientists
- an outlook
- “R Markdown” talk (Isabelle)

WHAT DO YOU THINK? COULD THERE BE PROBLEMS?

PROBLEMATIC TRENDS



PROBLEMATIC TRENDS



PROBLEMATIC TRENDS

PROBLEMATIC TRENDS

Extreme view: current academic data science is catering hype to compensate the Google envy.

PROBLEMATIC TRENDS

Assumptions, assumptions. Everywhere assumptions.

Regular article | Open Access

Tampering with Twitter's Sample API

Jürgen Pfeffer , Katja Mayer and Fred Morstatter

EPJ Data Science 2018 **7**:50

https://doi.org/10.1140/epjds/s13688-018-0178-0

Received: 12 December 2017 | **Accepted:** 11 December 2018 | **Published:** 19 December 2018

Pfeffer et al. (2018)

Abstract

Social media data is widely analyzed in computational social science. Twitter, one of the largest social media platforms, is used for research, journalism, business, and government to analyze human behavior at scale. Twitter offers data via three different Application Programming Interfaces (APIs). One of which, Twitter's Sample API, provides a freely available 1% and a costly 10% sample of all Tweets. These data are supposedly random samples of all platform activity. However, we demonstrate that, due to the nature of Twitter's sampling mechanism, it is possible to deliberately influence these samples, the extent and content of any topic, and consequently to manipulate the analyses of researchers, journalists, as well as market and political analysts trusting these data sources. Our analysis also reveals that technical artifacts can accidentally skew Twitter's samples. Samples should therefore not be regarded as random. Our findings illustrate the critical limitations and general issues of big data sampling, especially in the context of proprietary data and

undisclosed details about data handling.

Pfeffer et al. (2018)

- cognition -> language assumption
- online behaviour -> real behaviour assumption
- methodological flaws: random sampling
- even if: bias population remains!

WHY IS THIS A PROBLEM?

INTERMEZZO: REPRODUCIBILITY CRISIS

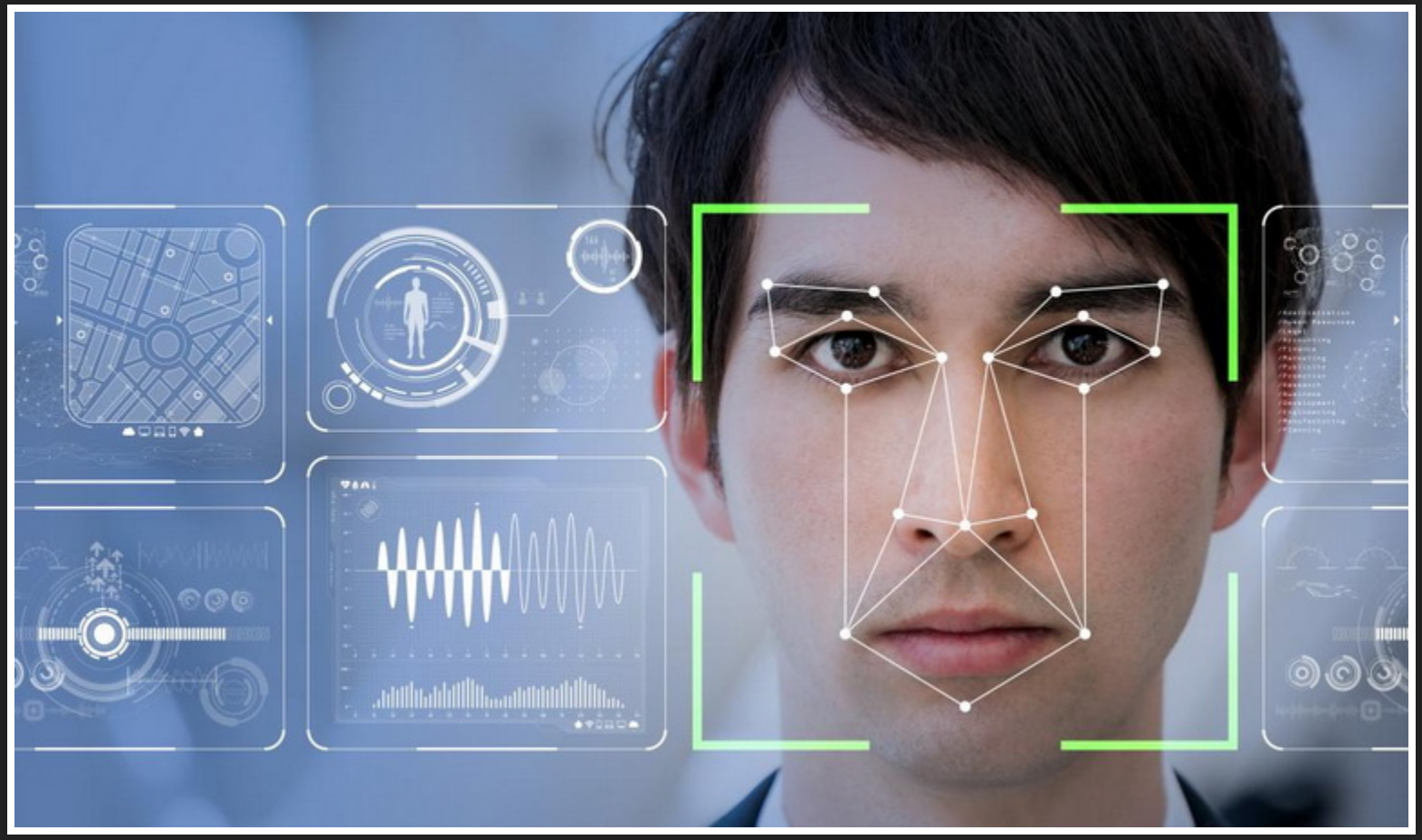
If we care about data science, we need to do a better job.

THE TECHNOLOGY FALLACY

THE TECHNOLOGY FALLACY

Img source

THE TECHNOLOGY FALLACY



Img source

THE TECHNOLOGY FALLACY

Popular belief: technology will solve all problems.

- esp. true for data
- “so we just need more data”
- so why not use it for all the difficult problems?

THE TECHNOLOGY FALLACY

Recent case:

TECH YOUTUBE

YouTube still can't stop child predators in its comments

A new video reopens discussion on an ongoing problem

By Julia Alexander | Feb 19, 2019, 12:50pm EST

Full article, Exposing YouTube video

THE TECHNOLOGY FALLACY

Mmh, that’s strange…?

- apparently not a solved problem
- and there’s more
  - Facebook
  - Twitter, etc. and content removal
- still: very much relying on humans

THE TECHNOLOGY FALLACY

Problem:

- this creates unrealistic expectations
- biggest challenge for data science: expectation management

THE NAIVITÉ FALLACY

THE NAIVITÉ FALLACY

UK government reveals new AI tool for flagging extremist content

THE NAIVITÉ FALLACY

The UK Home Office on Monday unveiled a £600,000 artificial intelligence (AI) tool to automatically detect terrorist content.

The Home Office cited tests that show the new tool can automatically detect 94% of Daesh propaganda with 99.995% accuracy. That accuracy rate translates into only 50 out of one million randomly selected videos that would require human review. The tool can run on any platform and can integrate into the video upload process to stop most extremist content before it ever reaches the internet.

source

THE NAIVITÉ FALLACY

THE NAIVITÉ FALLACY

Problem 1: A secret government agency has developed a scanner which determines whether a person is a terrorist. The scanner is fairly reliable; 95% of all scanned terrorists are identified as terrorists, and 95% of all upstanding citizens are identified as such. An informant tells the agency that exactly one passenger of 100 aboard an aeroplane in which you are seated is a terrorist. The agency decide to scan each passenger and the shifty looking man sitting next to you is tested as "TERRORIST". What are the chances that this man *is* a terrorist? Show your work!

THE NAIVITÉ FALLACY

	Terrorist	Passenger	
Terrorist	950	50	1,000
Passenger	4,950	94,050	99,000
	5,900	94,100	100,000

```
P(terrorist|alarm) = 950/5900 = 16.10%
```

THE NAIVITÉ FALLACY

Put simply: you can sell anything.

HERE'S AN IDEA

```
ai_terrorism_detection = function(person){
  person_classification = 'no terrorist'
  return(person_classification)
}
```

"UCL RESEARCHERS USE AI TO FIGHT TERRORISM!"

"AI 99.9999% ACCURATE IN SPOTTING TERRORISTS!"

DATA SCIENCE HEADLINES

UK government reveals new AI tool for flagging extremist content

GUIDE TO DATA SCIENCE HEADLINES

“UK government reveals new AI tool for flagging extremist content”

=

“UK government ~~reveals new AI tool for flagging extremist content~~ buys snake oil”

THE NAIVITÉ FALLACY

What to do about it:

- avoid the hype
- there is no rocket science here
- 95% is just (a type of) regression
- if it sounds too good to be true, it is

Beware of the hype!

THE CATEGORY MISTAKE OF DATA SCIENCE

CATEGORY MISTAKE

https://www.youtube.com/watch?v=fCLI6kxFFTE

CATEGORY MISTAKE

- So we are getting there with self-driving cars.
- Hence: we can also address the other challenges.

!!!!

CATEGORY MISTAKE

Geller, 1999, 538 article

“I would not be at all surprised if earthquakes are just practically, inherently unpredictable.”

(Ned Field)

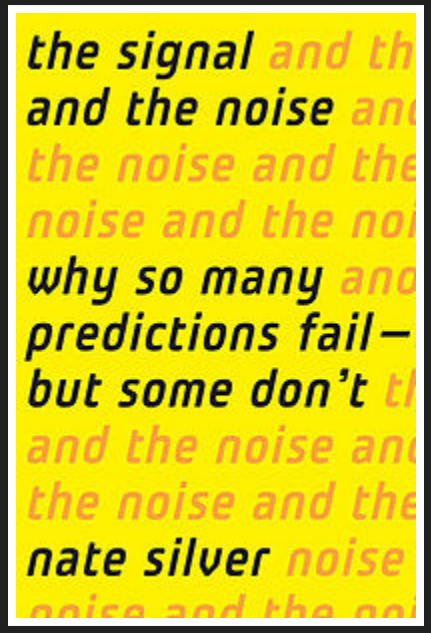
CATEGORY MISTAKE

- Building a sophisticated visual recogntion system != predicting everything
- Static phenomena vs. complex systems

Human behaviour might be the ultimate frontier in prediction.

If you only read one book in 2019…

Read: “The Signal and the noise”, Nate Silver



ETHICAL ISSUES

ETHICS & DATA SCIENCE

- data sources
- (machine) learning systems
- reinforcing systems
- responsible practices

ETHICS & DATA SCIENCE

Your turn: do you see problems for these aspects?

- data sources
- (machine) learning systems

ETHICS & DATA SCIENCE

What about “reinforcing systems”?

ETHICS & DATA SCIENCE

Choose 1:

1. FP/FN issue in the hand of practitioners
2. academics' responsibility

AN OUTLOOK

What would an ideal Data Science look like?

BE SPECIFIC…

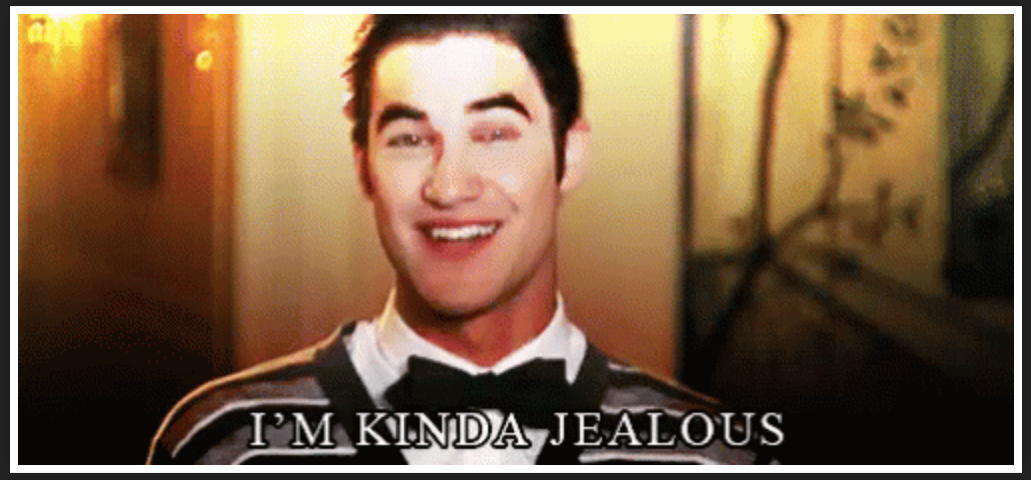
Academic data science

vs

“Industry” data science

Extreme view:

current academic data science is catering hype to compensate the Google envy.


I'M KINDA JEALOUS

ACADEMIC DATA SCIENCE

What it is doing	What it should be doing
creating “cool” studies	testing assumptions
pumping out non-reproducible papers	investing in fundamental data science research
hiring people to do cool things with our data	starting with the problem
getting on the data science train	focus on methods of data science

OUTLOOK

- we need *boring* studies!
  - longitudinal studies
  - assumption checks
  - replications
- we need to accept that Google & Co. are a different league in applying things
- we need to focus on the “ACADEMIC” part
- we need unis as control mechanism, not as a player

FOR THE FUTURE

Assumptions, assumptions, , assumptions. Everywhere assumptions.

Test them!

THIS WEEK

FEEDBACK submission + revision + your project

NEXT WEEK

- Lecture : The Applied Data Science pipeline
- Tutorial: full pipeline + your project

END