Intro to Data Science for Crime Scientists

Advanced Crime Analysis UCL

Bennett Kleinberg

7 Jan 2019

Advanced Crime Analysis

Data Science for Crime Science

What is this? and Why do we need it?

Why not just “Crime Analysis”?

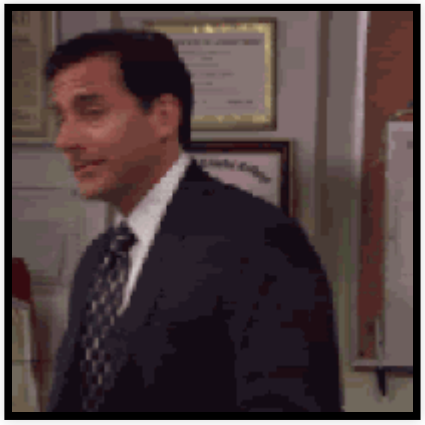
Crime Science in the 21st century

New research questions!

New ways to solve problems!

Also: uncomfortable!!

Aaah: so we're talking Big Data!



Problems with “Big Data”

- what is “big”?
- data = data?
- complexity of data?
- sexiness of small data

Game-changers in Crime Science

1. existence of data
2. availability of data
3. availability of computing resources
4. obsession with prediction

The time is now

POLICY & LAW

PALANTIR HAS SECRETLY BEEN USING NEW ORLEANS TO TEST ITS PREDICTIVE POLICING TECHNOLOGY

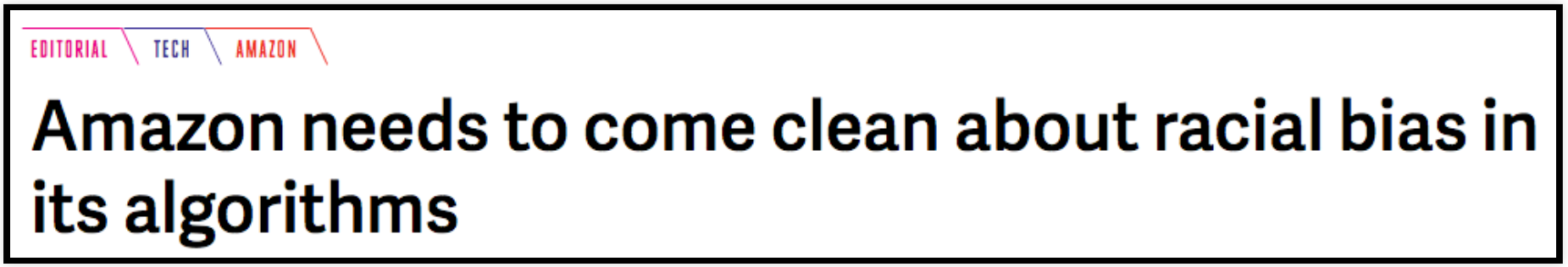
Palantir deployed a predictive policing system in New Orleans that even city council members don't know about

–> Lecture 7 + 8

https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd

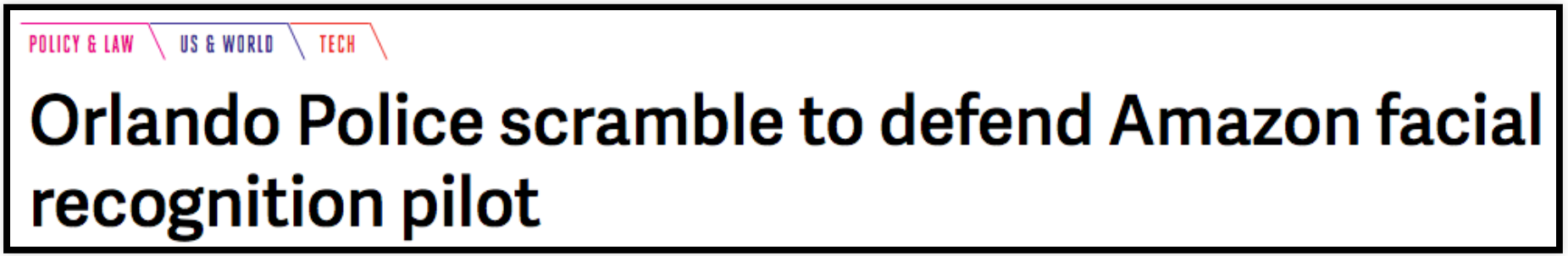
The time is now

https://www.theverge.com/2018/5/22/17379968/amazon-rekognition-facial-recognition-surveillance-aclu



https://www.theverge.com/2018/5/23/17384632/amazon-rekognition-facial-recognition-racial-bias-audit-data

The time is now



https://www.theverge.com/2018/5/24/17391632/amazon-facial-recognition-orlando-police-rekognition

–> Lecture 9

The time is now

https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

–> Lecture 9

The time is now

100,000 false positives for every real terrorist:
Why anti-terror algorithms don't work
by Timme Bisgaard Munk

http://firstmonday.org/ojs/index.php/fm/article/view/7126/6522

–> Lecture 7, 8, 9

The time is now

Police use a computer to expose false testimony

A lie-detection system being used by Spanish police highlights concerns about algorithms.

https://www.nature.com/articles/d41586-018-05285-9

–> Lecture 4 + 5

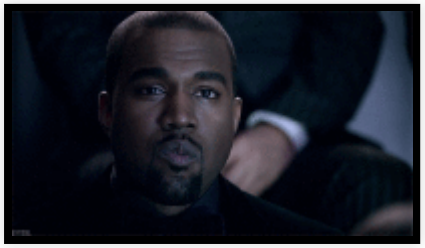
The time is now

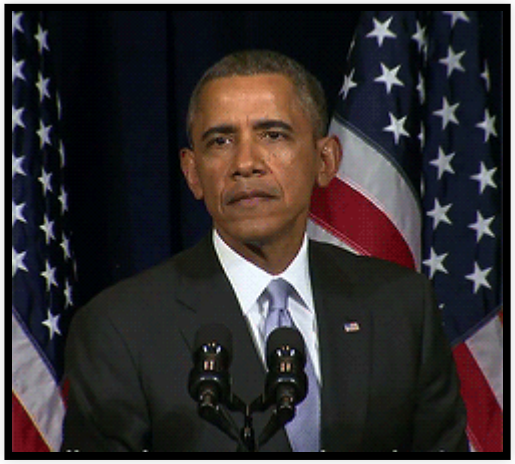
This fake news detection algorithm outperforms humans

https://thenextweb.com/artificial-intelligence/2018/08/22/this-fake-news-detection-algorithm-outperforms-humans/

–> Lecture 4 + 5

What is going on there?





Current situation

Data Science Wild West

- Anything goes
- A lot of great things
- A lot of sh**

You'll learn to tell hype from promise!

Becoming a real problem-solver

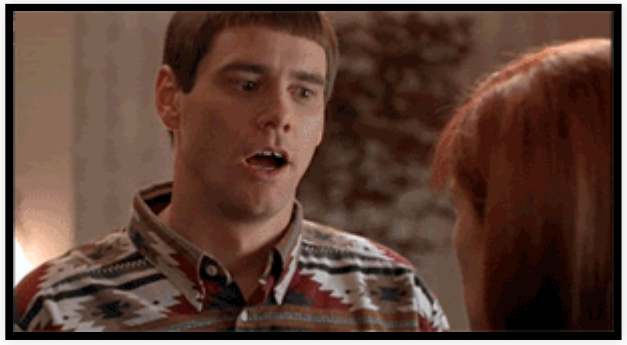
Principle 1: There’s no magic in Data Science

Principle 2: Golden data never comes in a spreadsheet

Principle 3: Data treasures are hidden in front of you

Principle 1:

There's no magic in Data Science



Principle 1:

There's no magic in Data Science

All names of current FBI most wanted terrorists?

Let's start here: https://www.fbi.gov/wanted/terrorism

Using data science techniques...

```
## Loading required package: xml2
```

```
##  [1] "SHAYKH AMINULLAH"
##  [2] "FAKER BEN ABDELAZZIZ BOUSSORA"
##  [3] "ABDULLAH AL-RIMI"
##  [4] "IBRAHIM SALIH MOHAMMED AL-YACOUB"
##  [5] "RAMADAN ABDULLAH MOHAMMAD SHALLAH"
##  [6] "ABDELKARIM HUSSEIN MOHAMED AL-NASSER"
##  [7] "ALI ATWA"
##  [8] "ABDUL RAHMAN YASIN"
##  [9] "HUSAYN MUHAMMAD AL-UMARI"
## [10] "ALI SAED BIN ALI EL-HOORIE"
## [11] "ABD AL AZIZ AWDA"
## [12] "AHMAD IBRAHIM AL-MUGHASSIL"
## [13] "JABER A. ELBANEH"
## [14] "JAMEL AHMED MOHAMMED ALI AL-BADAWI"
## [15] "MOHAMMED ALI HAMADEI"
## [16] "AYMAN AL-ZAWAHIRI"
## [17] "AHMAD ABOUSAMRA"
## [18] "ADNAN G. EL SHUKRIJUMAH"
```

The magic? A few lines of code

```
library(rvest)
target_page = read_html('https://www.fbi.gov/wanted/terrorism')
target_page %>%
  html_nodes('p.name') %>%
  html_text()
```

```
##  [1] "SHAYKH AMINULLAH"
##  [2] "FAKER BEN ABDELAZZIZ BOUSSORA"
##  [3] "ABDULLAH AL-RIMI"
##  [4] "IBRAHIM SALIH MOHAMMED AL-YACOUB"
##  [5] "RAMADAN ABDULLAH MOHAMMAD SHALLAH"
##  [6] "ABDELKARIM HUSSEIN MOHAMED AL-NASSER"
##  [7] "ALI ATWA"
##  [8] "ABDUL RAHMAN YASIN"
##  [9] "HUSAYN MUHAMMAD AL-UMARI"
## [10] "ALI SAED BIN ALI EL-HOORIE"
## [11] "ABD AL AZIZ AWDA"
## [12] "AHMAD IBRAHIM AL-MUGHASSIL"
## [13] "JABER A. ELBANEH"
## [14] "JAMEL AHMED MOHAMMED ALI AL-BADAWI"
## [15] "MOHAMMED ALI HAMADEI"
## [16] "AYMAN AL-ZAWAHIRI"
## [17] "AHMAD ABOUSAMRA"
## [18] "ADNAN G. EL SHUKRIJUMAH"
```

Principle 2:

Golden data never come in a spreadsheet

Golden data never come in a spreadsheet

```
1
00:00:00,000 --> 00:00:01,829
although there's no hard evidence<font color="#E5E5E5"> to</font>

2
00:00:01,829 --> 00:00:03,419
support Warren's claim of Native

3
00:00:03,419 --> 00:00:06,060
American<font color="#E5E5E5"> ancestry she has cited family</font>

4
00:00:06,060 --> 00:00:09,990
<font color="#E5E5E5">lore and not just a stray remarks about</font>

5
00:00:09,990 --> 00:00:12,750
```

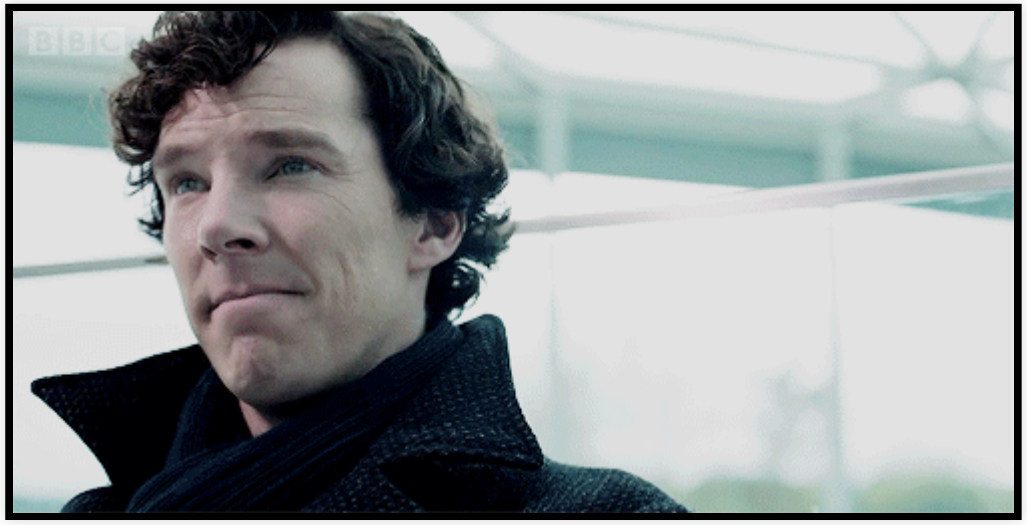
Principle 3:

Data treasures are hidden in front of you

Data treasures are hidden in front of you

Yeah, really cool.

But: I don’t need this “programming” for this!



“Programming” is only the vehicle.

Matter of volume

10 min. break

This module

Aim

- introduction to data science techniques
- being able to use state-of-the-art tools for crime analysis
- learning how to solve difficult problems
- “your quantitative masterpiece”

More on learning outcomes in the module handbook

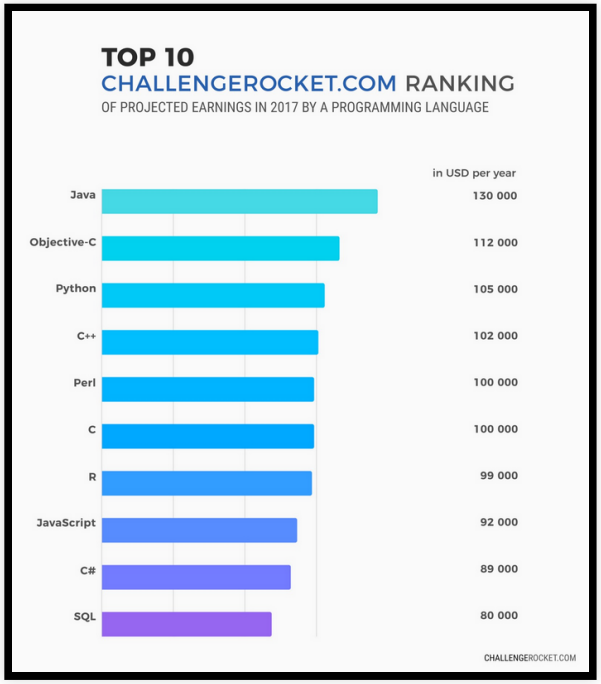
Things you'll learn

- Access Twitter data
- Build a web-scraper to the FBI's most wanted terrorists
- Write code that crawls through details of all missing persons in the UK
- Analyse the language of toxic YouTubers
- Build a linguistic model of Tweets in London
- Build your own machine learning models to predict whether a news article is fake or not
- your own capstone project

Tools we'll use

- open-source + free
- wide support community (e.g., on Stackoverflow)
- made for statistics
- state-of-the-art libraries

But still…



- R grows fast
- Highly desirable/required in industry (Google, Facebook, Microsoft, Amazon, …)

Structure of the module

- 9 Lectures (Mondays, 13-15h)
- 5 Tutorials (alternating Tuesdays, 11-13h)

Teaching assistant: Felix Soldner

Assessment

- Class test
- Applied Data Science Project

Class test

- 30% of final grade
- 1-hour closed-book exam
- open questions & MC questions
- Date: 18 Mar 2019, 13-15h

Applied Crime Analysis Project

- 70% of final grade
- your capstone project
- apply all skills learned in this module
- solve a problem you like
- we’re here to help: peer feedback + 1-on-1 feedback sessions

Feedback sessions

- Peer-feedback
  - shared evaluation among all students
  - at end of the “Text data II” lecture on 4 Feb. 2019
- 1-on-1 feedback
  - 10 min individual feedback from me and Felix
  - final advice to fine-tune your project
  - 4 March 2019

Outlook

- Webscraping
- Text mining
- Machine learning

What's next?

Homework for today:

1. Getting ready for R (on Moodle)
2. R for Crime Scientists in 12 Steps

Tomorrow's tutorial: "WTF!? session"

Next week: Web scraping