Operačný program
Efektívna
verejná správa

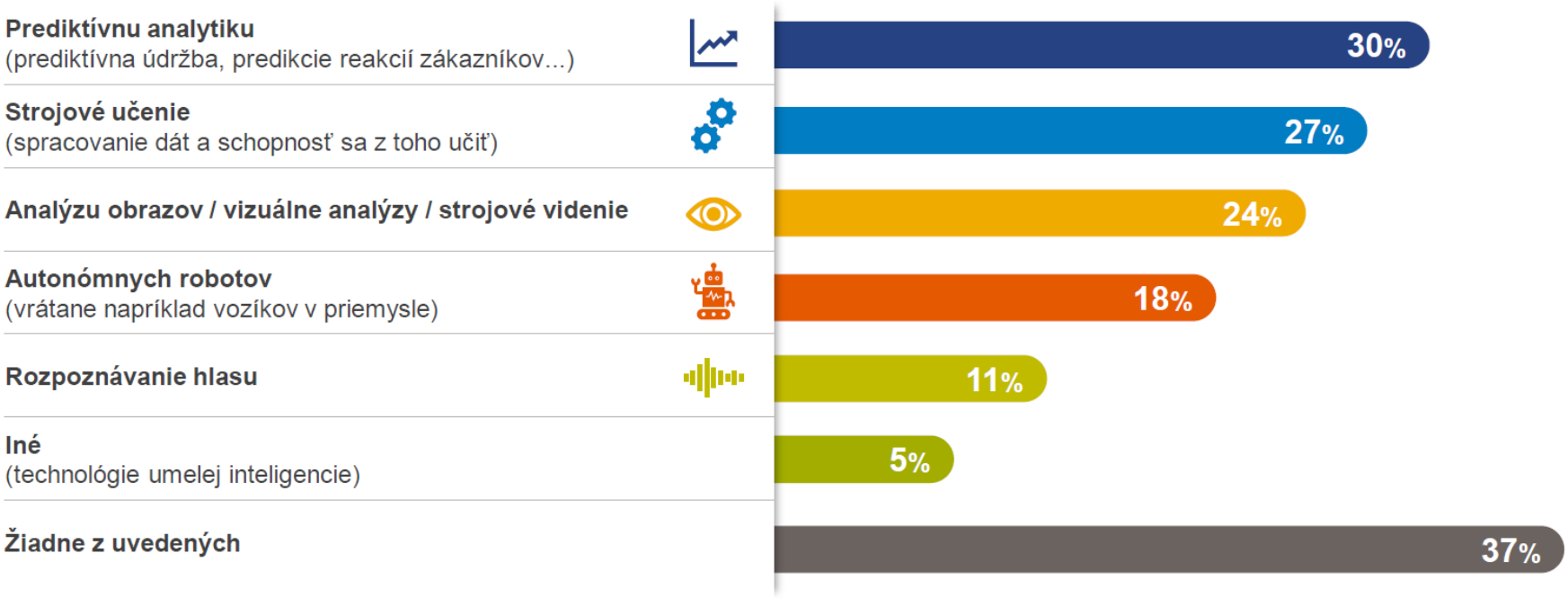
Európska únia
Európsky sociálny fond

Etika veľkých dát

doc. RNDr. Martin Takáč, PhD.
Centrum pre kognitívnu vedu FMFI UK

Online festival analytických jednotiek, 18. 3. 2021

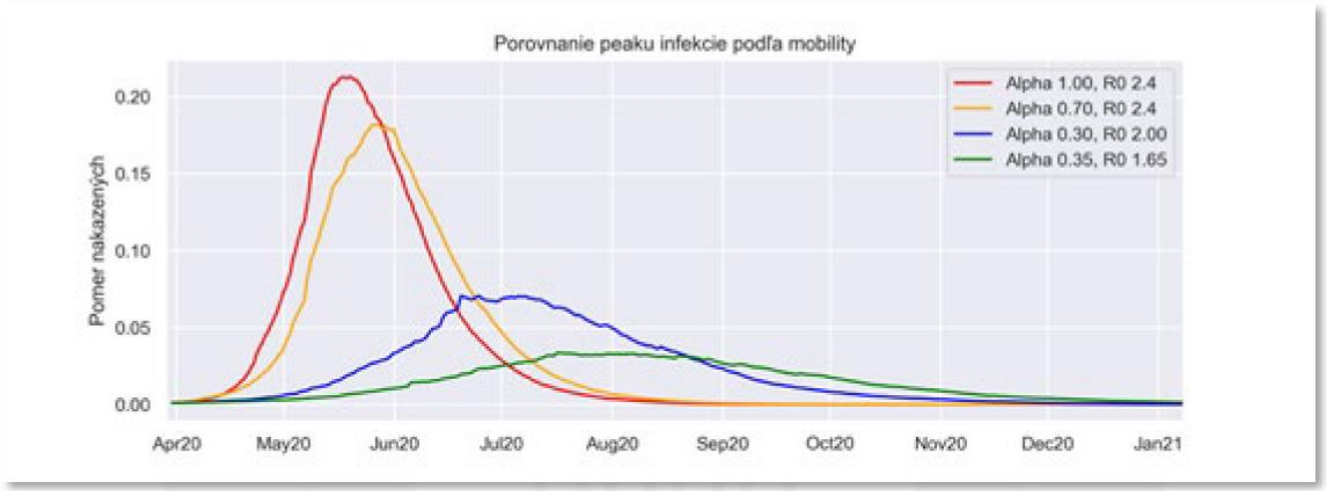
Súčasné využívanie umelej inteligencie
v spoločnostiach na Slovensku



Báza: všetci respondenti | N=206 | Q2 | Ktoré z nasledujúcich technológií umelej inteligencie dnes Vaša spoločnosť využíva?

Prediktívna analytika

Graf 1 (verzia 1): Priebeh infekcie podľa mobility obyvateľstva



zdroj: [Scenáre dopadu koronavírusu na Bratislavy, 2/4/2020]

Čím viac a presnejších dát, tým použiteľnejší model

COVID-19 SLOVENSKO: PREDPOVEDE A SKUTOČNOSŤ (JANUÁR 2021)

DÁTA BEZ PÁTOSU	DÁTUM PREDPOVEDE	PUBLIKÁCIA	PREDPOVEĎ	SKUTOČNOSŤ	ODCHÝLKA	ODCHÝLKA %
NOVÉ POZITÍVNE PRÍPADY 7 dňový priemer	3. - 13. JANUÁR 2021	FACEBOOK DENNÍKY	2,000	1,904	-96	-5%
DETI V ŠKOLE	3. - 13. JANUÁR 2021	FACEBOOK DENNÍKY	deti stále doma	deti stále doma	✓	✓
PRIJATÍ PACIENTI DO NEMOCNÍC V JANUÁRI 2021	3. JANUÁR 2021	FACEBOOK DENNÍKY	9,000	8,395	-605	-7%
KULMINÁCIA POČTU PACIENTOV	ZAČIATOK JANUÁRA 2021	FACEBOOK DENNÍKY	koniec januára	1. februára	1 deň	+5%
POČET ÚMRTÍ "NA" COVID KU KONCU JANUÁRA 2021	3. JANUÁR 2021	FACEBOOK DENNÍKY	4,000	4,711	+711	+18%
POČET ÚMRTÍ "NA+S" COVID KU KONCU JANUÁRA 2021	3. JANUÁR 2021	FACEBOOK DENNÍKY	5,000	5,746	+746	+15%
CELKOVÝ POČET POZITÍVNYCH Z PCR TESTOV V JANUÁRI 2021	3. - 13. JANUÁR 2021	FACEBOOK DENNÍKY	110,000	65,841	-44,159	-40%

zdroj: Dáta bez pátosu, 6. 2. 2021

Podmienky pre využitie veľkých dát

- Digitalizácia a internet (bezprecedentné množstvo dát a možnosť ich prepojenia)
- Pokrok v strojovom učení (machine learning) a algoritmoch dolovania dát (data mining)
- Veľká výpočtová sila

Naša digitálna stopa

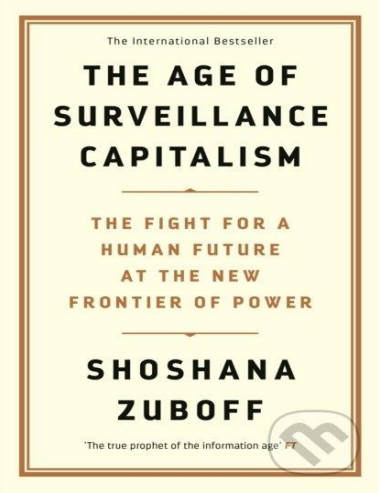
- Údaje z internetových prehliadačov (navštívené stránky, kliky, strávený čas, …)
- Vyhľadávacie frázy
- Sociálne siete (príspevky, lajky, štruktúra sietí priateľov)
- Internetové A/B testovanie
- Obsah emailov
- Prehliadaný obsah služieb (youtube, Spotify, Netflix atď.)
- Nákupy online, platby kreditnou kartou
- Údaje v databázach (daňové priznania, sociálna poisťovňa, digitálny zdravotný záznam, záverečné práce, školský prospech)
- Smartfóny/GPS – údaje o polohe a pohybe
- Hlasové dáta – Siri, Alexa, Cortana, …
- Obrazové dáta/ CCTV kamery – rozpoznávanie tvárí a emócií
- Internet of things – telesné senzory, senzory v domácnostiach, používanie zariadení (napr. tlačiarne, vykurovanie)

Naša digitálna stopa

- Čo sa dá z týchto údajov vyčítať?
- Kto ich vlastní?
- Kto má k nim prístup?
- Prečo sú (komerčne) zaujímavé?

Naša digitálna stopa

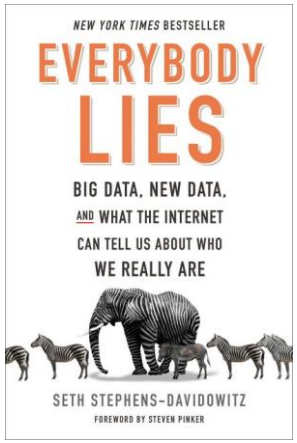
- Čo sa dá z týchto údajov vyčítať?
- Kto ich vlastní?
- Kto má k nim prístup?
- **Prečo sú (komerčne) zaujímavé?**
  - vedieť predpovedať a ovplyvňovať preferencie a správanie ľudí dáva ekonomickú a politickú moc



Naša digitálna stopa

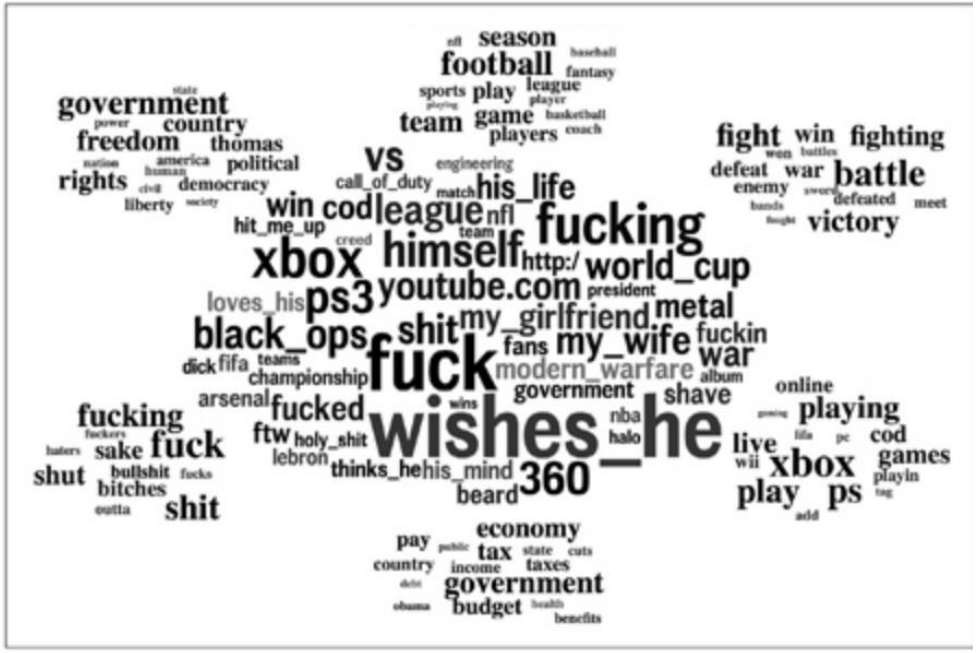
- **Čo sa dá z týchto údajov vyčítať?**
- Kto ich vlastní?
- Kto má k nim prístup?
- Prečo sú (komerčne) zaujímavé?

Príspevky na sociálnych sieťach



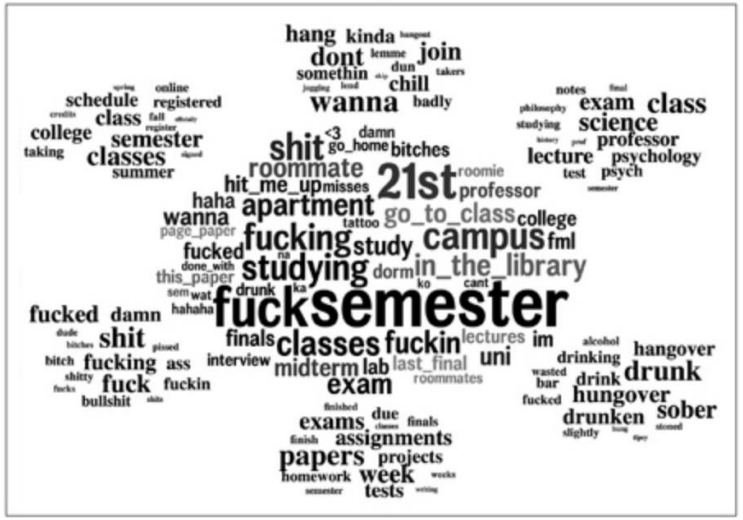
- Štúdia jazykových výrazov na Facebooku (Stephens-Davidowitz, 2017)

- Muži



Ženy

Jazyk na Facebooku podľa veku



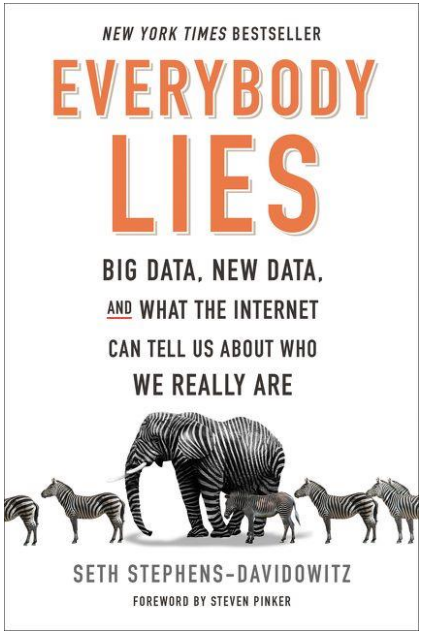
19-22 r.

23-29 r.

30-65 r.

Vyhľadávacie frázy

- Presnejšie ako štatistiky z prieskumov (self-reportov) a sociálnych sietí (nie je motivácia klamať)
- V prieskumoch a na sociálnych sieťach ľudia klamú (Stephens-Davidowitz, 2017)

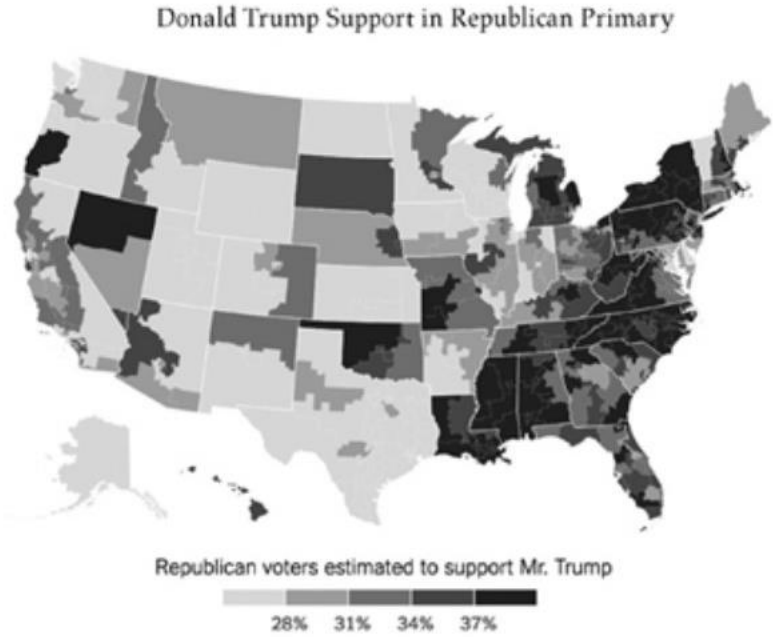


Vyhľadávacie frázy

- Politické názory
- Predsudky, rasizmus
- Osobné, zdravotné a rodinné problémy
- Sexuálne fantázie
- Zneužívanie detí, nechcené tehotenstvo
- …

Vyhľadávacie frázy

- Prediktory volebného víťazstva



Vyhľadávacie frázy

Schopnosť včas predpovedať:

- Epidémie (rýchlejšie ako z údajov zdravotného systému)
- Sociálne nepokoje a revolúcie

Prediktívne modely

- Nezávislé premenné -> závislé premenné
- Trénovanie modelu:
  - Existujúce údaje -> známy výsledok
- Použitie:
  - Nové údaje -> predikcia výsledku
- Predpoveď sa môže týkať populácie, ale aj jednotlivcov

Individuálne predikcie

- Nezávislé premenné -> závislé premenné
- Trénovanie modelu:
  - Existujúce údaje o populácii -> známe výsledky
- Použitie:
  - Nový jedinec -> zovšeobecnenie
- Predikcie obvykle slúžia na minimalizáciu (korporátnych) rizík alebo nákladov
- Pravdepodobnosť, že tento jedinec:
  - (Ne)bude schopný splácať pôžičku/hypotéku
  - je vhodným zamestnancom/študentom alebo odíde/prepadne
  - …

Ktoré slová sú najlepším prediktorom (ne)schopnosti splácať dlh?

- God
- promise
- debt-free
- minimum payment
- lower interest rate
- will pay
- graduate
- thank you
- after-tax
- hospital

Ktoré slová sú najlepším prediktorom (ne)schopnosti splácať dlh?

- Splatí

debt-free
lower interest rate
after-tax
minimum payment
graduate

- Nesplatí

God
promise
will pay
thank you
hospital

zdroj: C. O'Neil: Weapons of math destruction, 2016

Prediktívne modely v práve

- Odhad rizík, „predictive policing“, súdne rozhodnutia
- Optimistický pohľad
  - Znížia náklady, zvýšia efektívnosť, zrýchlia rozhodnutia a vymožiteľnosť práva
  - Eliminujú subjektívnosť a zaujatosť
- Naozaj?

Prediktívne modely v práve

- Strojová zaujatosť (machine bias), stereotypy
- Softvér na predikciu recidívy v USA zaujatí proti Američanom čiernej pleti:
  - Štatistický test izoloval efekt rasy od efektov predošlej kriminálnej histórie, veku a pohlavia. Výsledok: ľudia čiernej pleti označení o 77% častejšie ako vysoko rizikoví pre spáchanie násilného trestného čina a o 45% častejšie pre spáchanie akéhokoľvek trestného činu (*Pro Publica,* Angwin et al., 2016).

Kedy je algoritmus zaujatý?

- Predpokladajme predošlú históriu recidívy 30% v skupine A a 70% v skupine B.
- Algoritmus je nezaujatý, ak predpovedá recidívu v skupine B s pravdepodobnosťou:
  - 70% ?
  - 50%?
  - iné možnosti?

Úspešnosť predikcií sa dá vyhodnotiť

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Nezaujatý model = presný model?

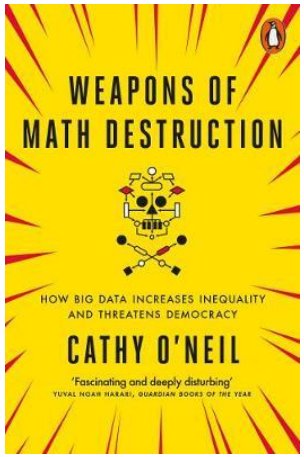
- Rozhodnutie Európskeho súdneho dvoru z r. 2011: Aby sa eliminovala zaujatosť a diskriminácia na základe pohlavia pri stanovovaní cien poistného, **poisťovne nesmú dávať ženám - vodičkám lacnejšie poistenie** (ani dávať mužom vyššiu penziu kvôli tomu, že v priemere žijú kratšie) (Kuschke 2012).
- Ale štatistiky ukazujú, že ženy majú v priemere nižšiu pravdepodobnosť spôsobenia dopravnej nehody.
- Poistenie je založené na **presnom odhade rizika**.
- Akýkoľvek algoritmus strojového učenia na stanovenie ceny poistného objaví zástupné ukazovatele (proxies) pre pohlavie/rod.
- Nepriama diskriminácia skupiny občanov na základe zástupných ukazovateľov je tiež nezákonná.
- Možné riešenie: výpočet poistného na mieru, na základe individuálnych údajov, napr. ponuka nižšieho poistného pre vodičov, ktorí si nainštalujú trasovacie zariadenie vyhodnocujúce bezpečnosť ich štýlu jazdy.

Faktory nezaujatosti predikčného modelu

- Na akých *kritériách* sú založené predikcie? (napr. rod, rasa, zdravotné postihnutie, ekonomický status alebo ich zástupné ukazovatele)
  - Potenciálne porušenie ústavných práv
- Sú *vstupné dáta* pre model nezaujaté?
  - Inak model replikuje historicky existujúce predsudky
- Je algoritmus *používaný* rovnako a férovo?
  - Obvykle viac zasahuje ľudí s nižším socioekonomickým statusom
- Sú rozhodnutia transparentné a zdôvodniteľné?
  - Algoritmy sú často komplikované a proprietárne
- Za akých podmienok je dovolené obmedziť *moje* práva na základe údajov o *iných ľuďoch*?
  - Prezumpcia viny: *Ľudia ako vy*… neplatia pôžičky, sú teroristi, atď.

Problémy s predikčnými modelmi

- Čo z nich robí „zbrane matematického ničenia“? (O’Neil, 2016)
  - Netransparentnosť (rozhodnutia nie sú zdôvodnené)
  - Škála (rovnaký algoritmus nasadený masovo)
  - Škoda (závažný dopad na kvalitu života)



Otázka

- V ktorých oblastiach je akceptovateľné obmedziť slobodu/práva jedinca na základe údajov o iných ľuďoch?

Teória detekcie signálu	**Udalosť bola predpovedaná**	**Udalosť nebola predpovedaná**
Udalosť nastala	Zásah (true positive)	**Neodhalenie (false negative)**
Udalosť nenastala	**Falošný poplach (false positive)**	- (true negative)

Otázka

- V ktorých oblastiach je akceptovateľné obmedziť slobodu/práva jedinca na základe údajov o iných ľuďoch?

Teória detekcie signálu	**Udalosť bola predpovedaná**	**Udalosť nebola predpovedaná**
Udalosť nastala	Zásah (true positive)	**Neodhalenie (false negative)**
Udalosť nenastala	**Falošný poplach (false positive)**	- (true negative)

- Tam, kde cena za neodhalenie výrazne presahuje cenu za falošný poplach
  - epidémie, terorizmus, ...

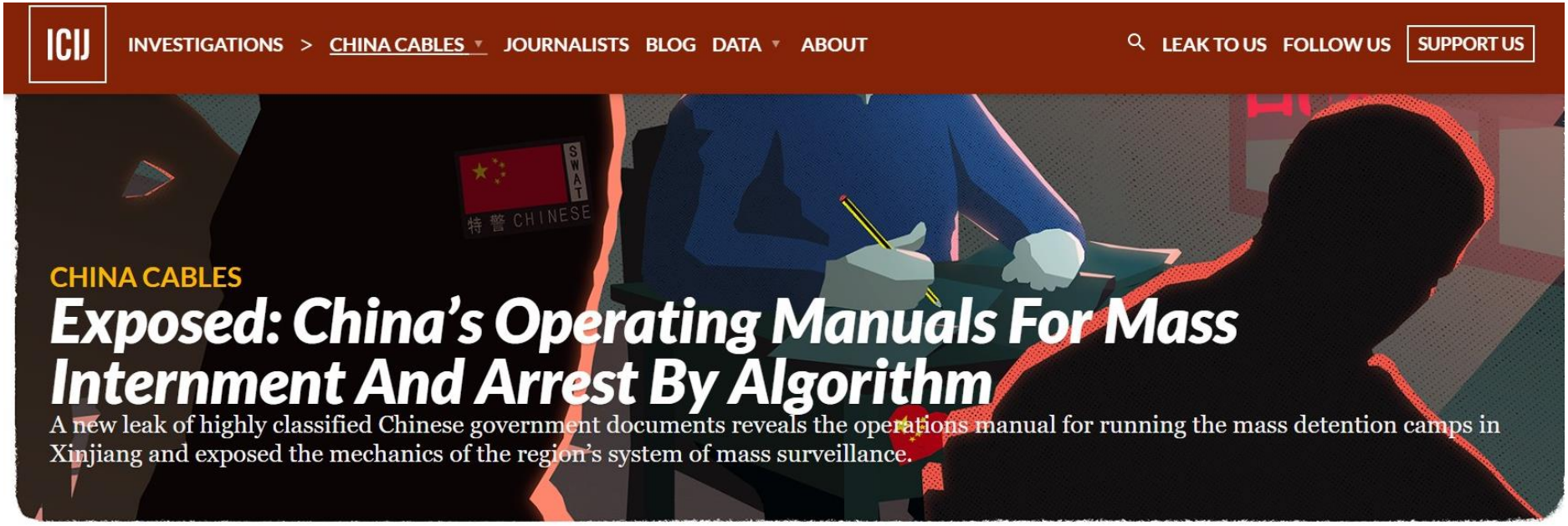
Stanovenie cenovej matice

- Problém: ak sú algoritmy používané spoločnosťami aby minimalizovali *ich* riziko, cena za neodhalenie (ktorú znáša spoločnosť) má väčšiu váhu ako cena za falošný poplach (ktorú znáša jedinec).
  - Riziko zneužitia, diskriminácie a posilňovania socioekonomickej nerovnosti

Systém sociálnych kreditov v Číne

- Jednotný parameter „dôveryhodnosti“ na základe sledovaného správania
- Systém odmien a trestov
- Obsahuje samoposilňujúcu spätnú väzbu – kredit jedinca je ovplyvnený kreditom ľudí, s ktorými má blízke vzťahy
- V Číne nie je spoločnosťou vnímaný tak zle ako na Západe
  - Menší dôraz na individuálnu slobodu a ochranu súkromia, úžitok z vymožiteľnosti morálky, bezpečia a stability

Zneužiteľnosť



NOVEMBER 24, 2019

READING TIME

20 MINUTES

A new leak of highly classified Chinese government documents has uncovered the operations manual for running the mass detention camps in Xinjiang and exposed the mechanics of the region's Orwellian system of mass surveillance and "predictive policing."

NEWSLETTER

Rozpoznávanie tvárí

Automated Inference on Criminality using Face Images

Xiaolin Wu
McMaster University
Shanghai Jiao Tong University
xwu510@gmail.com

Xi Zhang
Shanghai Jiao Tong University
zhangxi_19930818@sjtu.edu.cn

Abstract

We study, for the first time, automated inference on criminality based solely on still face images, which is free of

management science, criminology, etc.

In all cultures and all periods of recorded human history, people share the belief that the face alone suffices to reveal innate traits of a person. Aristotle in his famous work

(a) Three samples in criminal ID photo set S_c.

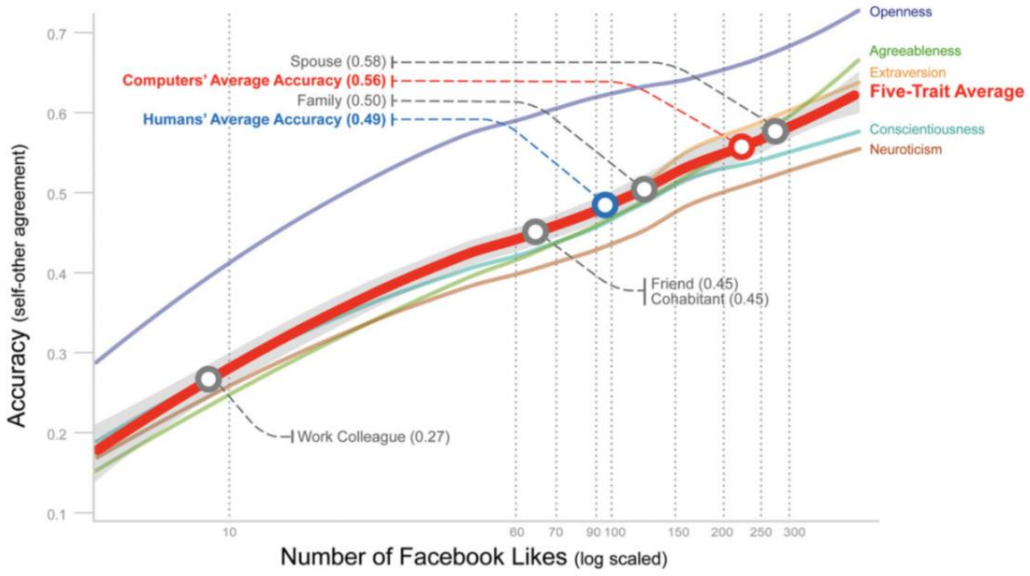
(b) Three samples in non-criminal ID photo set S_n

Etnická zaujatosť

- Algoritmy majú nedostatok vstupných dát o minoritných skupinách (napr. trénované prevažne na bielej / Kaukazskej rase)
- Algoritmy rozpoznávania tvárí častejšie zlyhávajú pre nebelochov.
- Rozpoznávanie reči a analýza textu (napr. detekcia nenávistných komentárov) funguje menej spoľahlivo na ľudí s prízvukmi a nie rodených anglicky hovoriacich
- Timnit Gebru
- Joy Buolamwini
- Dokumentárny film "Coded Bias"

Veľké dáta a politická moc

- Troll boti - http://politicalbots.org/ [Howard and Kollanyi, 2016]
- (Kossinski et al, 2015) Cambridge University's Psychometric Centre
  - 86,00 používateľov Facebooku, apka 'myPersonality'
  - Psychologický osobnostný profil Big-5 (OCEAN)
  - Predikcia OCEAN z lajkov
  - Vysoká presnosť



- https://applymagicsauce.com/

Veľké dáta a politická moc

- V r. 2015 Alexandr Kogan (Global Science Research, GSR) reimplementoval Kossinského model a pomocou Mechanical Turk získal demografické údaje a lajky od užívateľov Facebooku a ich priateľov (~350) [Davies, 2015].
- **Cambridge Analytica** = SCL (Strategic Communication Laboratories, UK) + Renaissance (hedge fond, USA) kúpila údaje od GSR a zlúčila ich s údajmi o americkom elektoráte (>50 mil. voličov)
- Apky na „Canvassing“ [Graessegger and Krogerus, 2017]
- Personalizovaná pro-Brexitová a pro-Trumpová (2016) kampaň
- Cambridge Analytica a Facebook čelili vyšetrovaniu v USA a UK.

Zdroj: https://ai-and-society.wiki.otago.ac.nz/images/6/69/Ai-elections-update.pdf

Záver

- Veľké osobné dáta sú lukratívnym artiklom
- Kto má údaje a modely, má moc
- Sú potenciálnou hrozbou pre demokraciu
- Regulácia je potrebná, na národnej aj medzinárodnej úrovni
- Dobrý príklad – GDPR