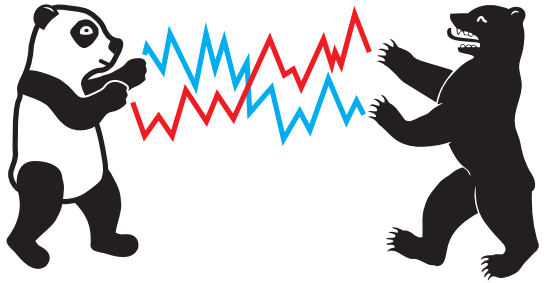
DS2

Data Science & Digital Society

C.A.S.E.

CENTER FOR APPLIED STATISTICS AND ECONOMICS



Ladislaus von Bortkiewicz Professor of Statistics
Humboldt-Universität zu Berlin
lvb.wiwi.hu-berlin.de

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SSRN

Congratulations Wolfgang K.,

You are currently in the top 10% of Authors on SSRN by total new downloads within the last 12 months.

Check out your Personalized Rankings Page to see where you rank amongst other authors on SSRN.

View your Personalized Ranking

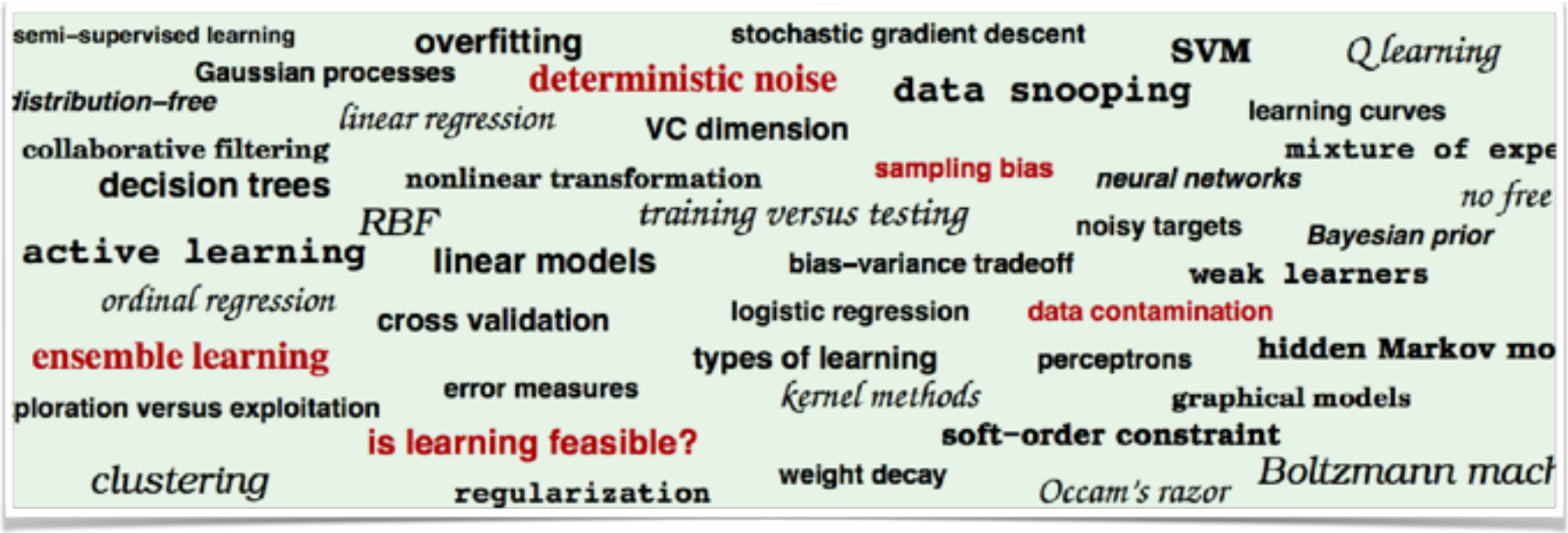
LinkedIn

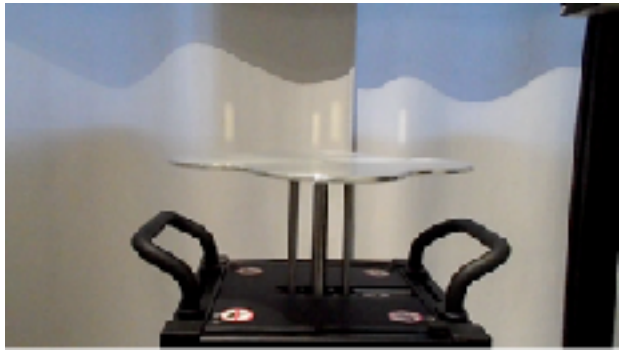
MENDELEY

SSRN

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

ML Machine Learning



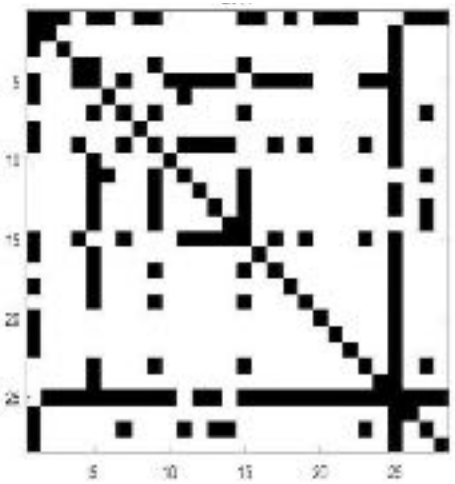


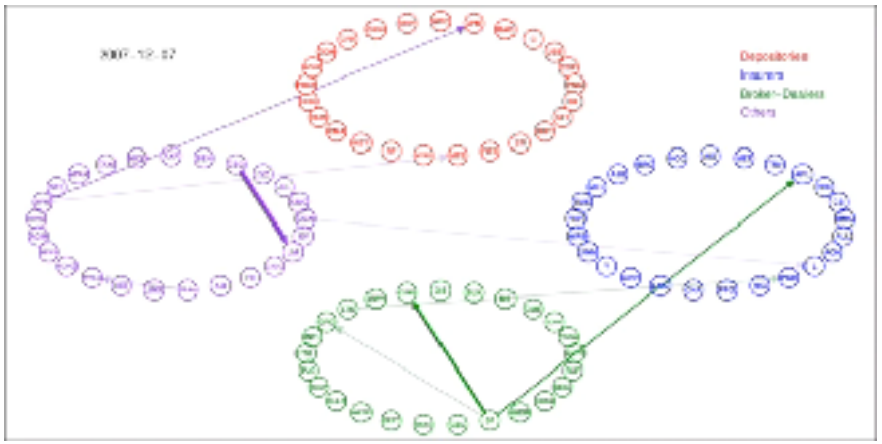
work.caltech.edu/library/181.html

W Phillips (curve), hydro engineer/economist, MONIAC Monetary National Income Analogue Computer

Data Science & Digital Society

DS2





Digital technologies connect, empower and expose individuals

Dynamic interactions ask for smart data science

Digital society creates new business relations

Questions

How do digital technologies change interactions?

How do decision processes evolve?

How does data science impact/define societies?

Society components

A Micro

Individual decision

B Meso

Dynamic networks

C Macro

Defined through data boundaries

A Micro level

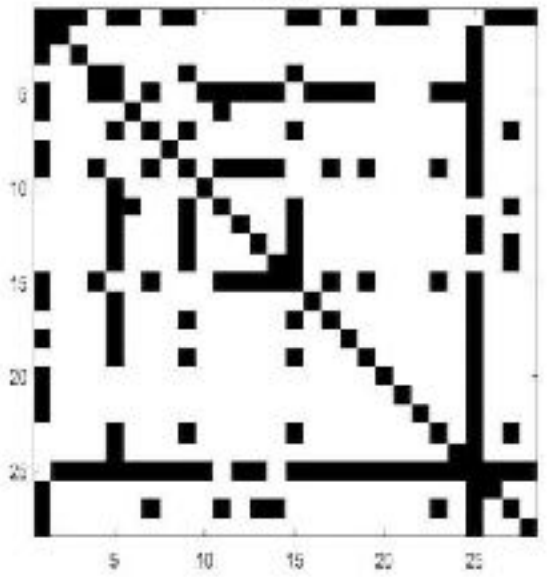
Social Media

Individual Decision Making

Digital Business & Operations

B Meso level

DS2



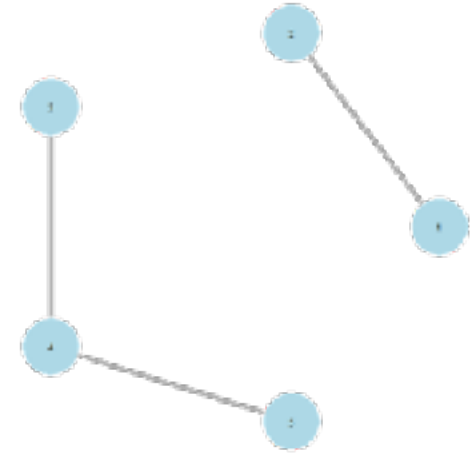
Network Emergence and Evolution

Dynamic information aggregation

Automation & Risk

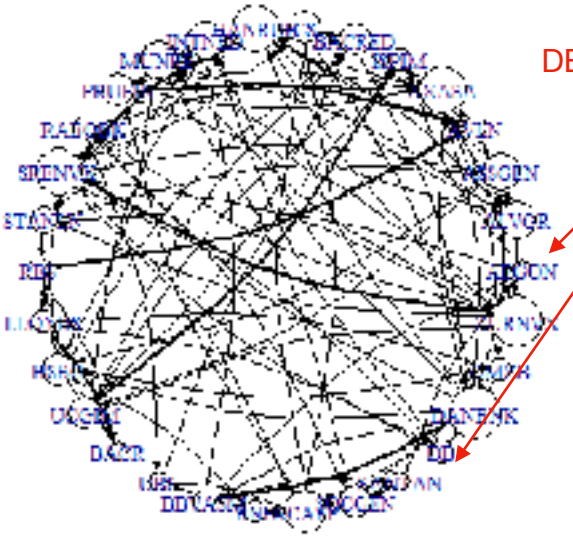
$$\begin{pmatrix} 1&1&0&0&0\\ 1&1&0&0&0\\ 0&0&1&1&0\\ 0&0&1&1&1\\ 0&0&0&1&1 \end{pmatrix} \begin{pmatrix} 1&1&0&0&0\\ 1&1&0&0&0\\ 0&0&1&1&0\\ 0&0&1&1&0\\ 0&0&0&1&1 \end{pmatrix} \begin{pmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&1&1&0\\ 0&0&0&0&1 \end{pmatrix}$$

etc

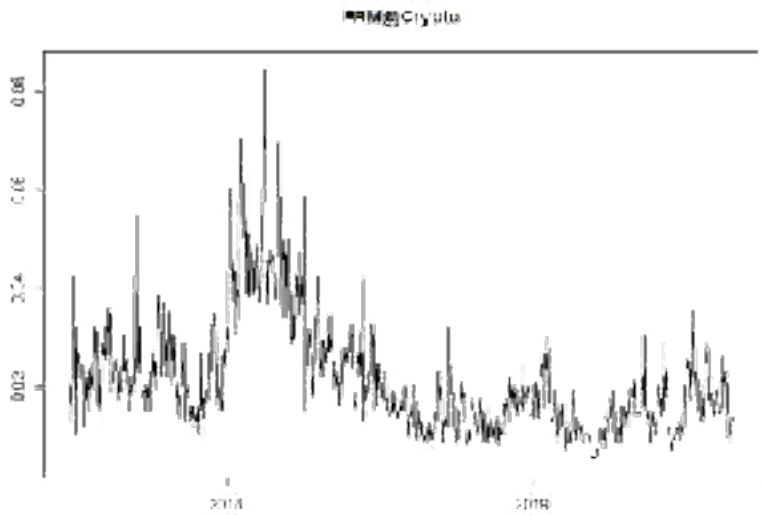


C Macro level

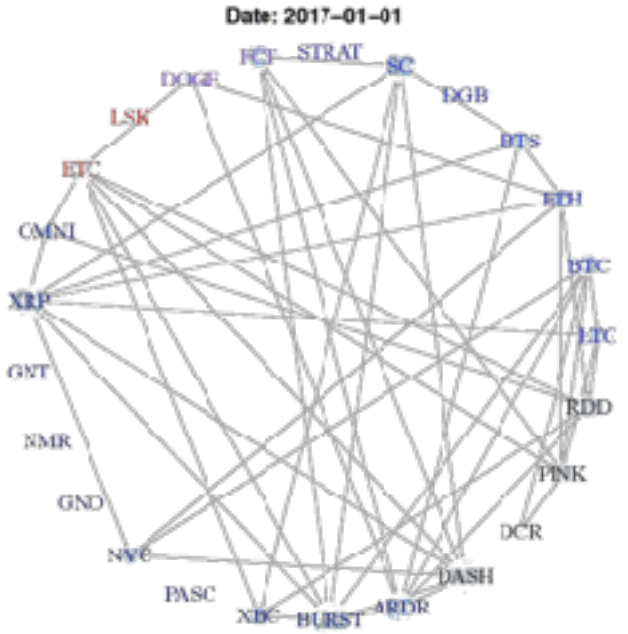
FRM@iTraxx Sen Fin Europe CDS Index



DB total centrality increases in July 2019, as does ALVGR

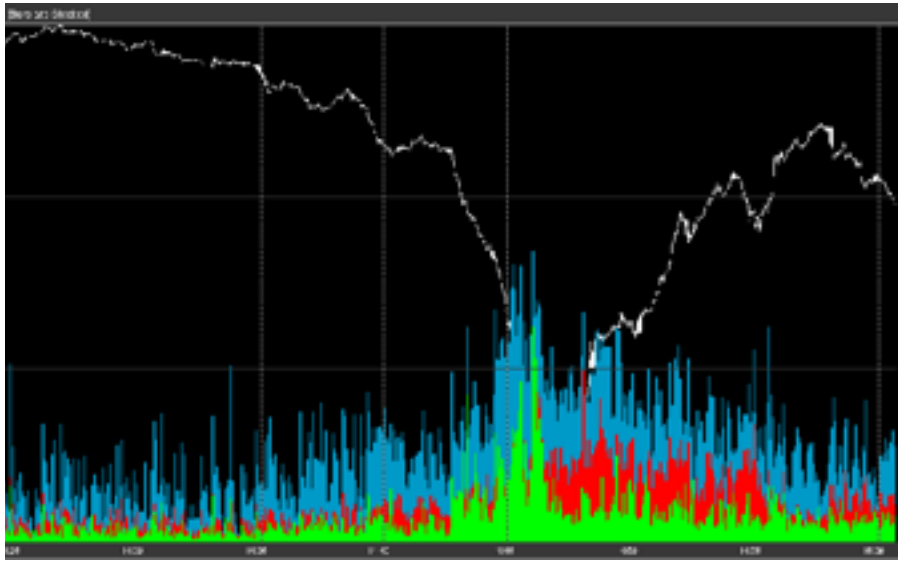


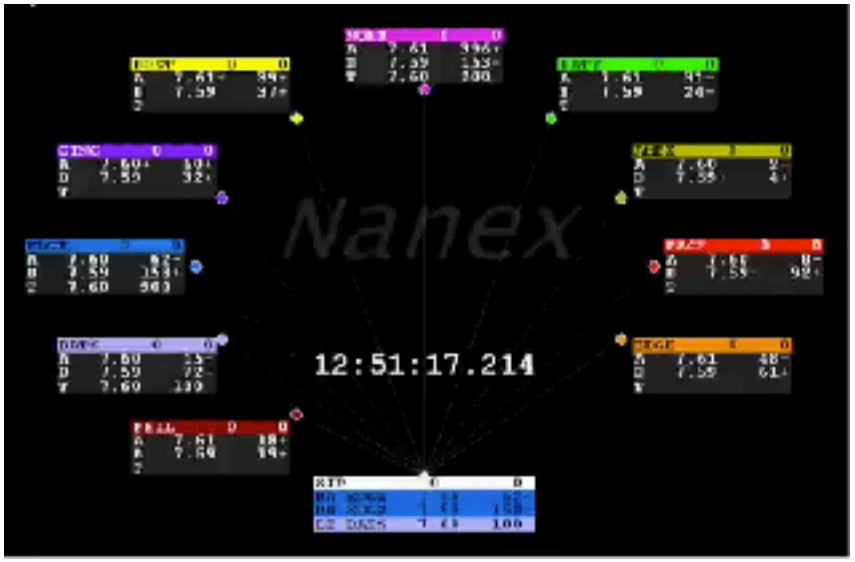
20190301
FRM: 0.032



HF Trading & Network

DS2





LOB Data Structure

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

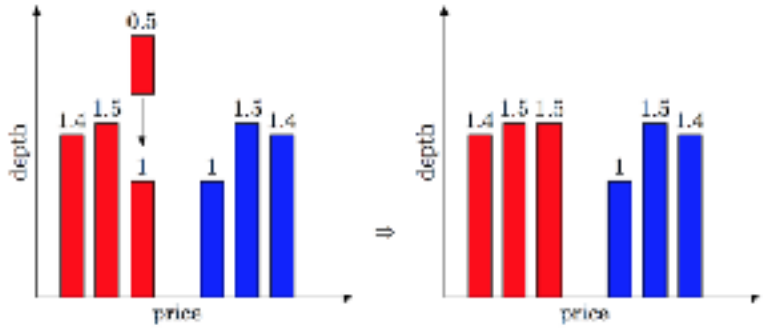
- message file: indicators for the type of event causing an update of the limit order book. e.g. 34713 *sec* = 9:38:55am

Time (sec)	Event Type	Order ID	Size	Price	Direction
⋮	⋮	⋮	⋮	⋮	⋮
34713.685155243	1	206833312	100	118600	-1
34714.133632201	3	206833312	100	118600	-1
⋮	⋮	⋮	⋮	⋮	⋮

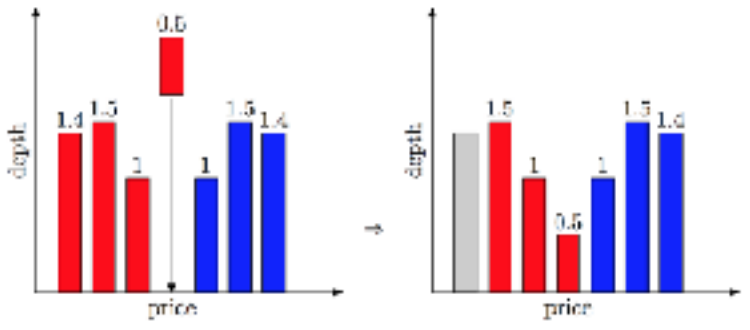
October 2, 2013 - Blackberry rallies from $7.60 to $8.00. Watch the deluge of quotes from Nasdaq (pink at 12 o'clock) overwhelm the system when the price ticks to the next level. Quote rates approach 40,000 per second on a 25 millisecond basis. (10 sec of extreme trading here)

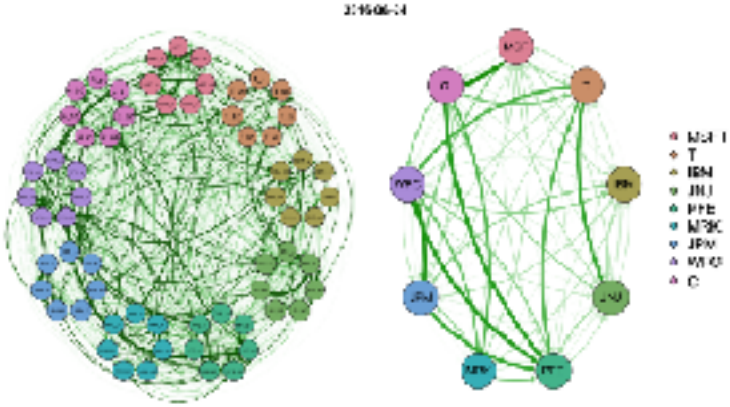
HF Trading & Network

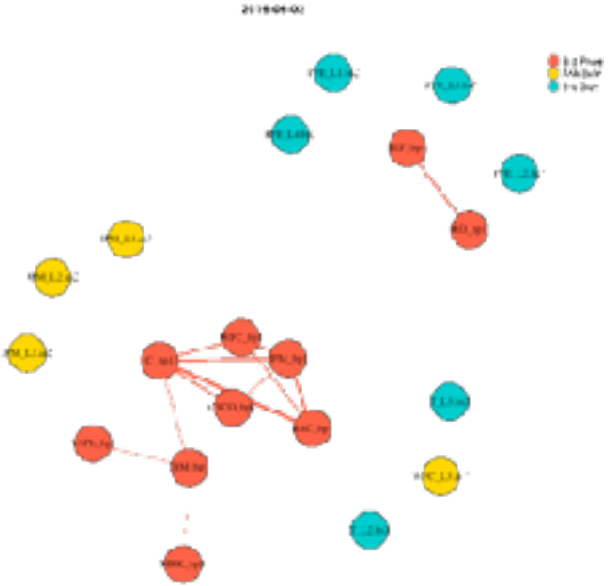
LOB - Normal Limit Order



LOB - Aggressive Limit Order

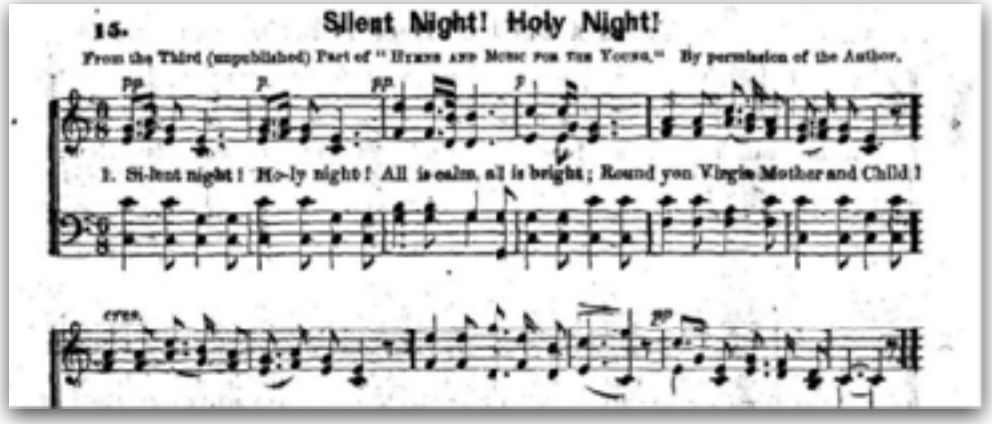




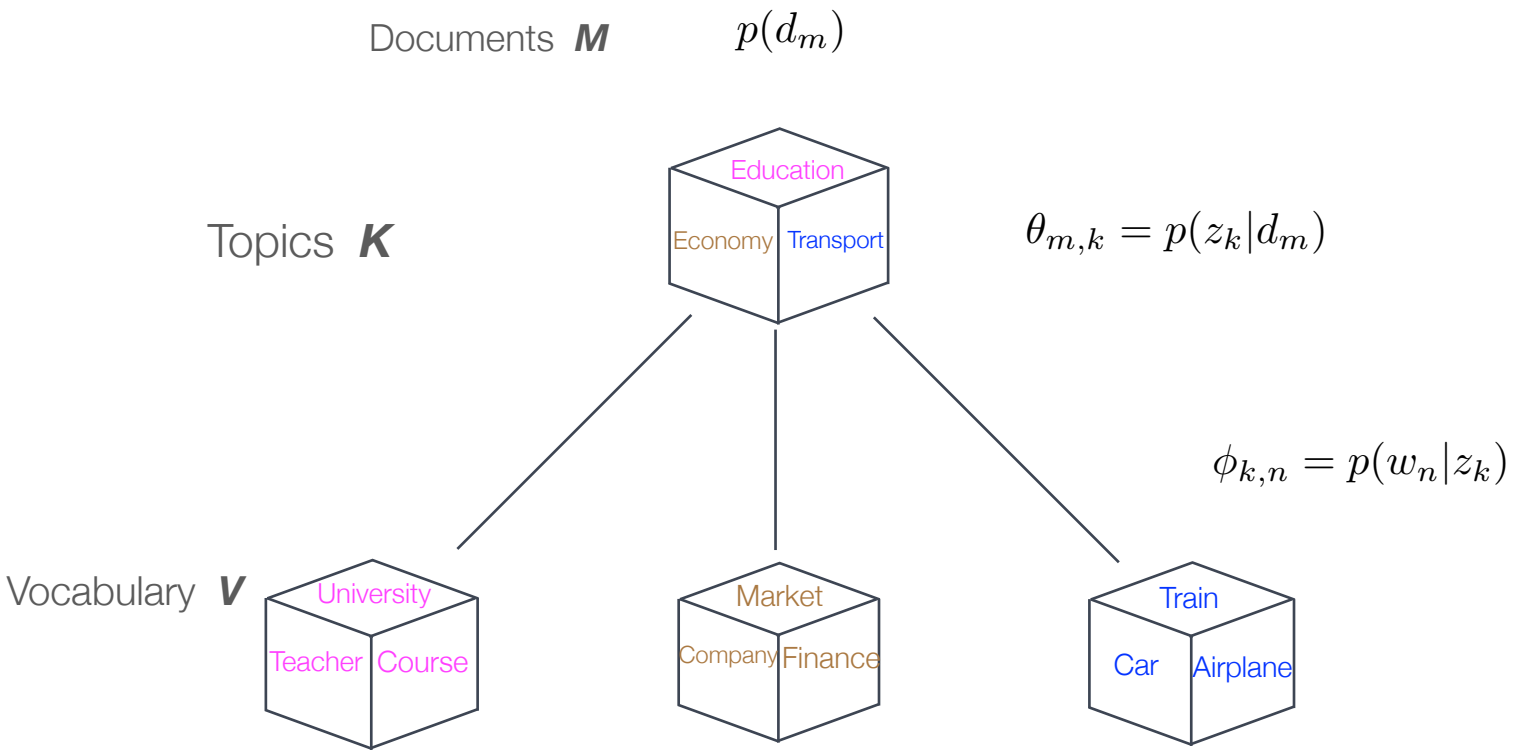


Christmas Songs

- Observed data/text
- 155 distinct words
- number of topics ?



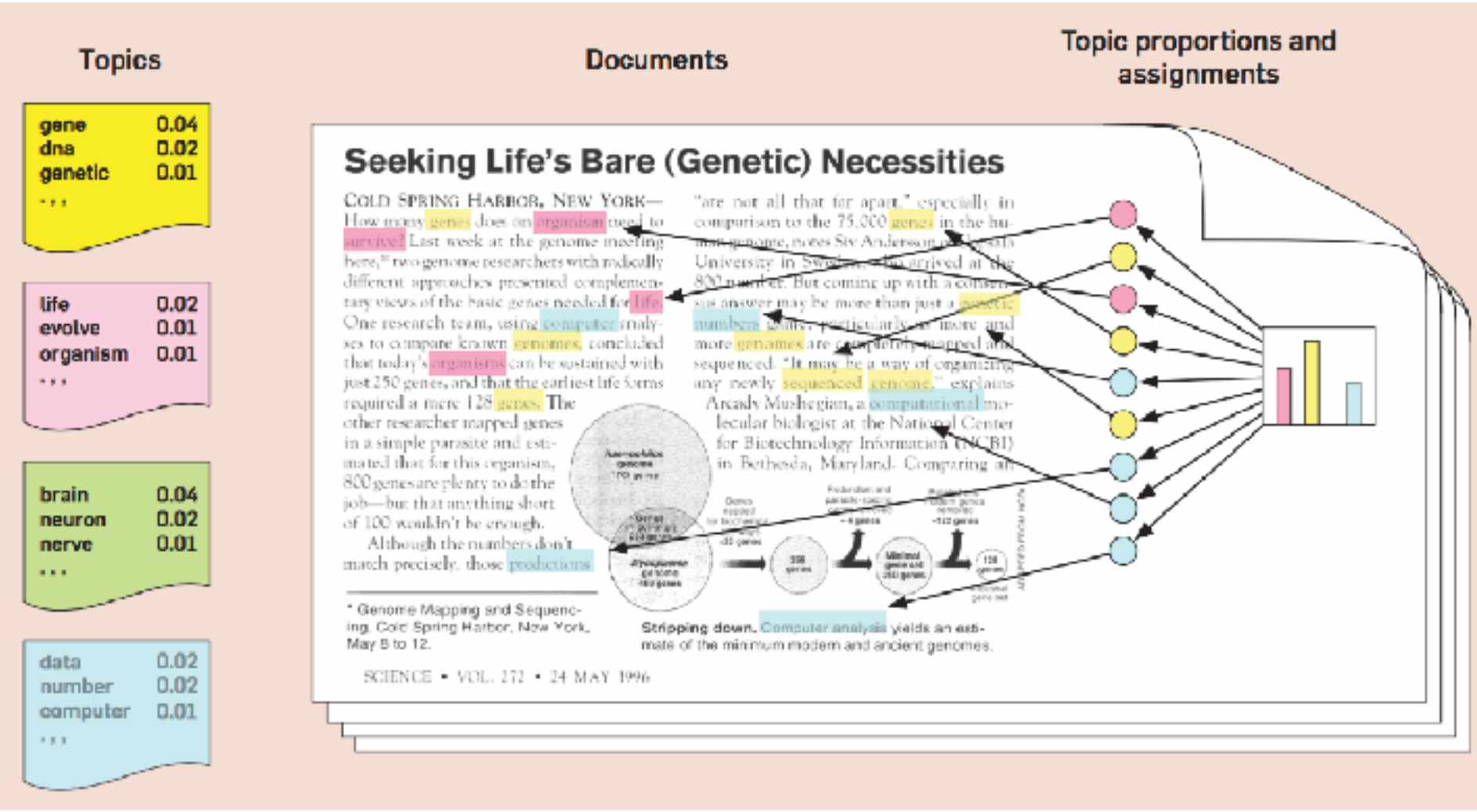
LDA Latent Dirichlet Allocation



$$p(d_m, w_n) = p(d_m)p(w_n|d_m) = p(d_m)\sum_{k=1}^{K} p(w_n|z_k)p(z_k|d_m)$$

θ ϕ

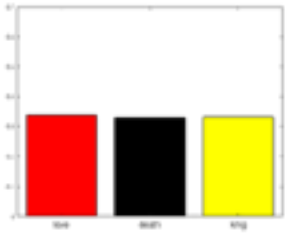
DTM Dynamic Topic Modelling



DTM Dynamic Topic Modelling

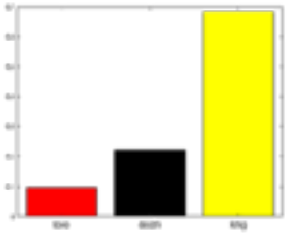
- Text I: Dirichlet parameter $\beta = (1, 1, 1)$

love love death king love love king king love love king king death king death king love king death king love love death king king king death death love love death king king love death love king king death death death king love king love love death king king love love death death death death love king king death king king love king king love king king king king love death king love king death king king death king love king death love king king death love death love death death love king death king love death love king king death



- Text II: Dirichlet parameter $\beta = (2, 5, 15)$

death king king king king king king death king king king love love death king king king king death king love king king king death king death death king king king king love king king king king king king death king death king king death death king king king king king king death king king king king death king king king king king king death king king king king king death king king king death death king king death king king love king death king death death king king king death



What to the Slave is the Fourth of July?

DS2

Frederick Douglass
July 5, 1852

Mr. President, Friends and Fellow Citizens:

He who could address this audience without a quailing sensation, has stronger nerves than I have. I do not remember ever to have appeared as a speaker before any assembly more shrinkingly, nor with greater distrust of my ability, than I do this day. A feeling has crept over me, quite unfavorable to the exercise of my limited powers of speech. The task before me is one which requires much previous thought and study for its proper performance. I know that apologies of this sort are generally considered flat and unmeaning. I trust, however, that mine will not be so considered. Should I seem at ease, my appearance would much misrepresent me. The little experience I have had in addressing public meetings, in country schoolhouses, avails me nothing on the present occasion.

http://teachingamericanhistory.org/library/document/what-to-the-slave-is-the-fourth-of-july/

Word Clouds

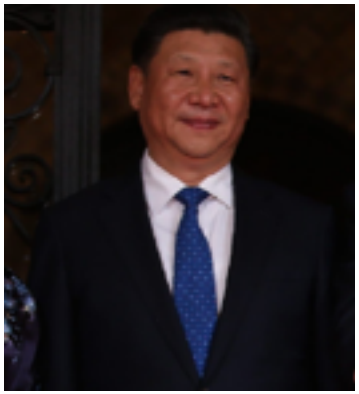
Frederick Douglass

Donald Trump

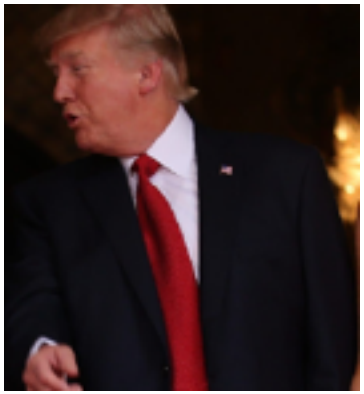
Presidential Address

DS2

Xi Jinping



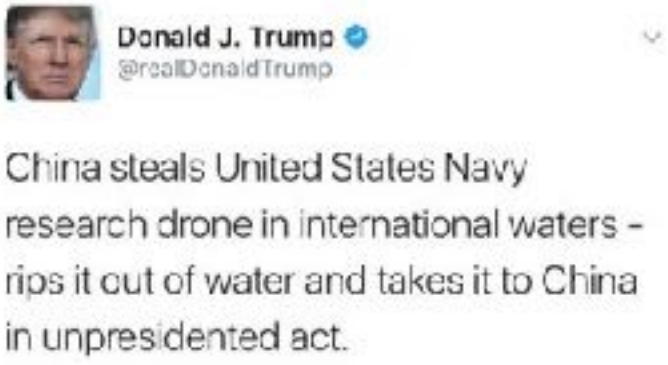
Donald Trump



Tokenization

- Neural networks take numbers as input
- Thus strings need to be „tokenized“
- Character-level vs word-level
  - Character-level uses more memory but has a smaller dictionary.
  - Word-level is faster and uses less memory.
  - We use character-level, because it is easier to tell whether an output is sensible.

The perils of character-level text generation: non-sense words.



As the network learn more and more patterns, it erroneously combines a proper word with a common prefix and a common suffix.

Trump

- ~33.000 tweets from Donald Trump

  (slightly less than 7 MB)

- Courtesy of http://www.trumptwitterarchive.com/

  which - unlike Twitter itself - has virtually all historic & deleted tweets

Markov Chains

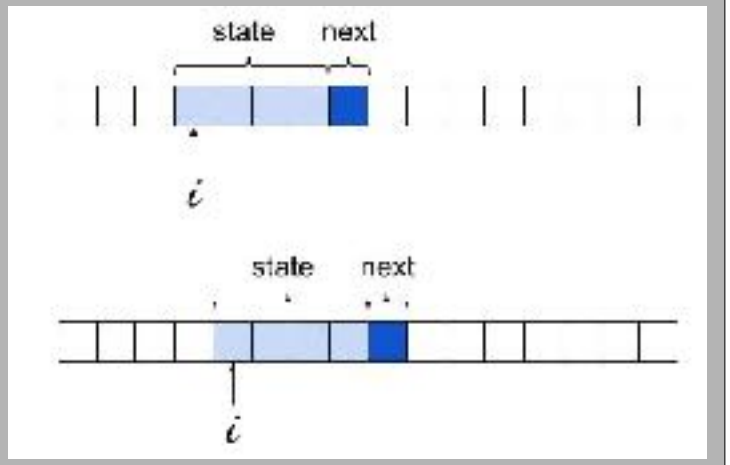

```
from collections import defaultdict, Counter
import random
import sys

# Order of Markov-Chain
STATE_LEN = 4

# Load Data and Initialise Dictionary
data = rawtxt
model = defaultdict(Counter)

# Build the dictionary
print('Learning model...')
for i in range(len(data) - STATE_LEN):
    state = data[i:i + STATE_LEN]
    next = data[i + STATE_LEN]
    model[state][next] += 1
print('Done.')

# Generation
statelist = ['Crooked Hillary', 'Fake News', 'China ', 'Robert Muller', 'We'] # Seeds to use
for state in statelist:
    if len(state)>=STATE_LEN:
        state=state[:STATE_LEN]
        print('Starting state: "{}"'.format(state))
        out = list(state) # Convert the seed text to a list of chars
        for i in range(250):
            out.extend(random.choices(list(model[state]), model[state].values())) # Look up current chain in dictionary and randomly choose next char
            state = state[1:] + out[-1]
        print(''.join(out),'\n')
```



Source: https://eli.thegreenplace.net/2018/elegant-python-code-for-a-markov-chain-text-generator/

Markov Chain Results

- Length=1
  - "Cayooererdauer htre @Jumimangrut crith for tpand ton wile p fof Trses ald n w ote baler7://e: Miathtum! f ndaye: e won touses dl Hoiz: tra.Trere.."Tutheco is yot wt dil Jis? hitlefo WallDaG d ithig cks Gringraciou inathe wd! @reaureis JJKit ick Ave ba"
- Length=2
  - "My htter of to cagesio dieve wing ell Con buttps://example.com onameople.conicat wis eve watch pronalserso arearent:ht @OH2B: ht--and to. Sperins ands we wat a nof thecom/"
- Length=3
  - "Chicago. Pleasting dontry Very rivery patem ally heal have been not betting Farned in it donaldTrump https://example.com/"
- Length=4
  - Roberts on the greatAgain & Republic and that? Will not wanted by the Democrats) is been threat fasterf101 That's a talentwarrive to has last nighting anyone in the can people awaii. https://example.com/
- Length=5
  - "Crooked agreement Make America and change coming you answer when he debt if that last night become Counting trying the deal argumental_boss: @realDonaldTrump leaders can't be on to ends in Florida announced his beyond win! https://example.com/"
  - "China is only way ahead in thing and inefficientis is sent the attended I wish he begun!,"
  - Robert & Order todayshow is the election and unfair and neighbors are going to Obama had to a paid term that Obama will leader just levels on really imagine is brings of New Jersey Happy Birthdahdah: Whats good donaldTrump @DonaldTrump for our countr

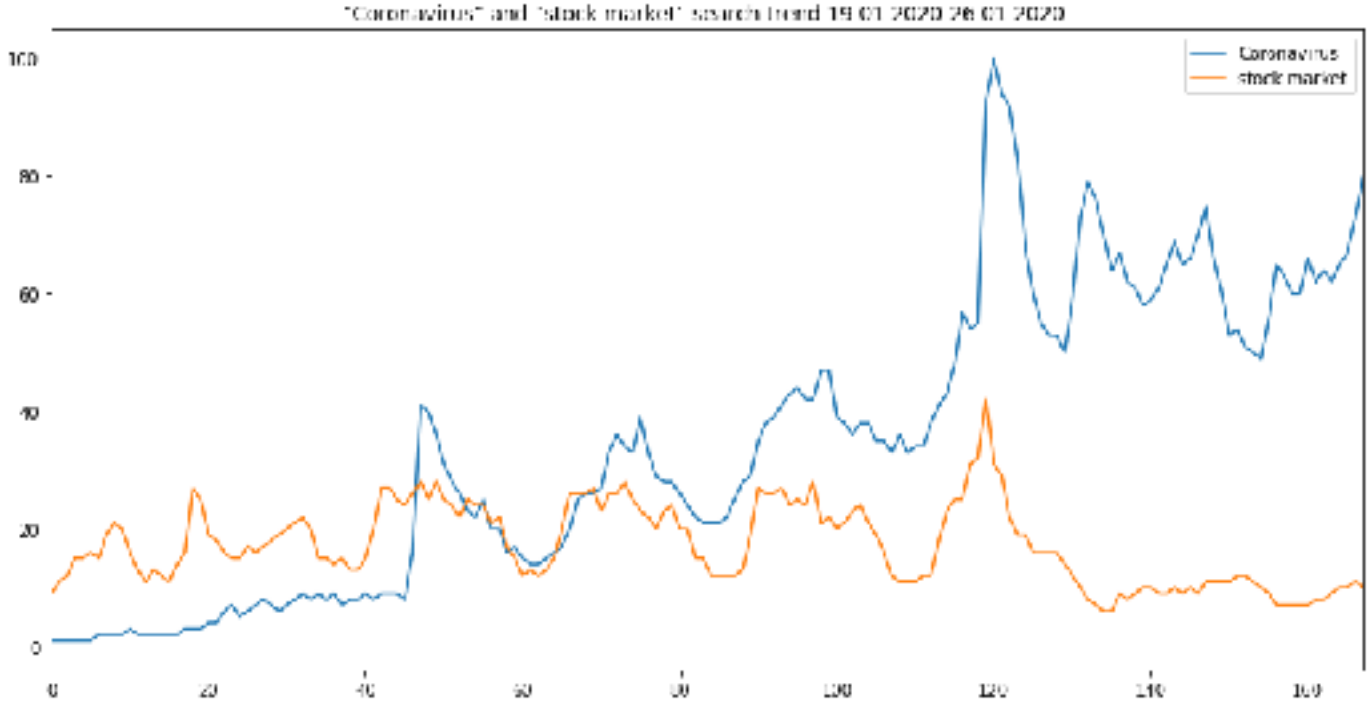
Fake News

DS2

- Umbrella term for an epidemic
- Disinformation and distrust in experts
- Economics, Finance not immune

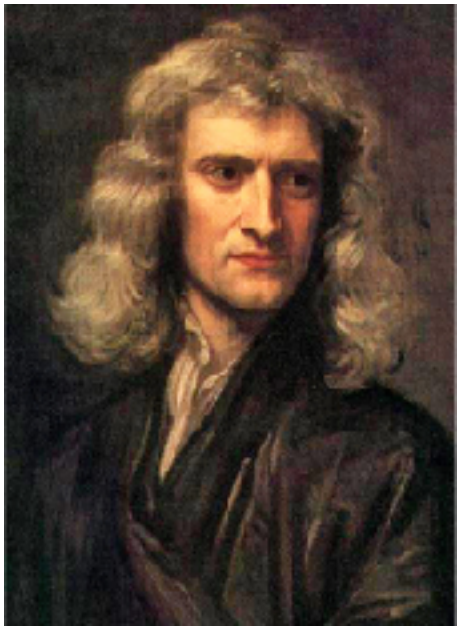
News

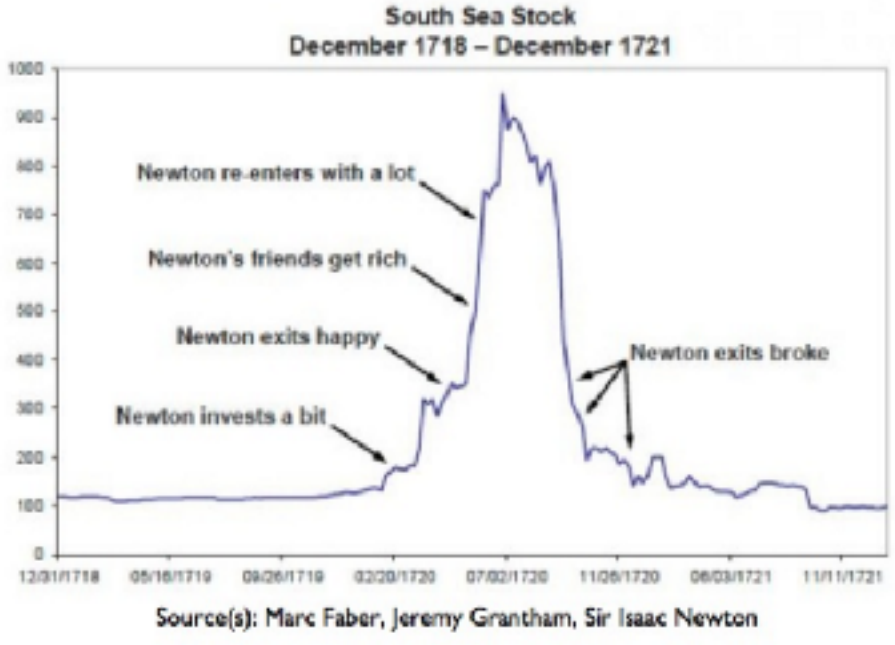
- Corona V
- Pharma
- Stocks



Fake News

Back in the spring of 1720, Sir Isaac Newton owned shares in the South Sea Company, the hottest stock in England. As a rumor among the people appeared that ore of silver/gold was found, this was "fake news " as we call them today. The great physicist muttered that he 'could calculate the motions of the heavenly bodies, but not the madness of the people.' Newton dumped his South Sea shares, pocketing a 100% profit totaling £7,000. But just months later, swept up in the wild enthusiasm of the market, Newton jumped back in at a much higher price — and lost £20,000 (or more than $3 million in [2002-2003's] money. For the rest of his life, he forbade anyone to speak the words 'South Sea' in his presence.





https://doi.org/10.1098/rsnr.2018.0018

ML Trumpbot

DS2

TrumpBot: Seq2Seq with Pointer Sentinel Model

Trumpbot

Filip Zivkovic
Department of Computer Science
Stanford University
zivkovic@stanford.edu

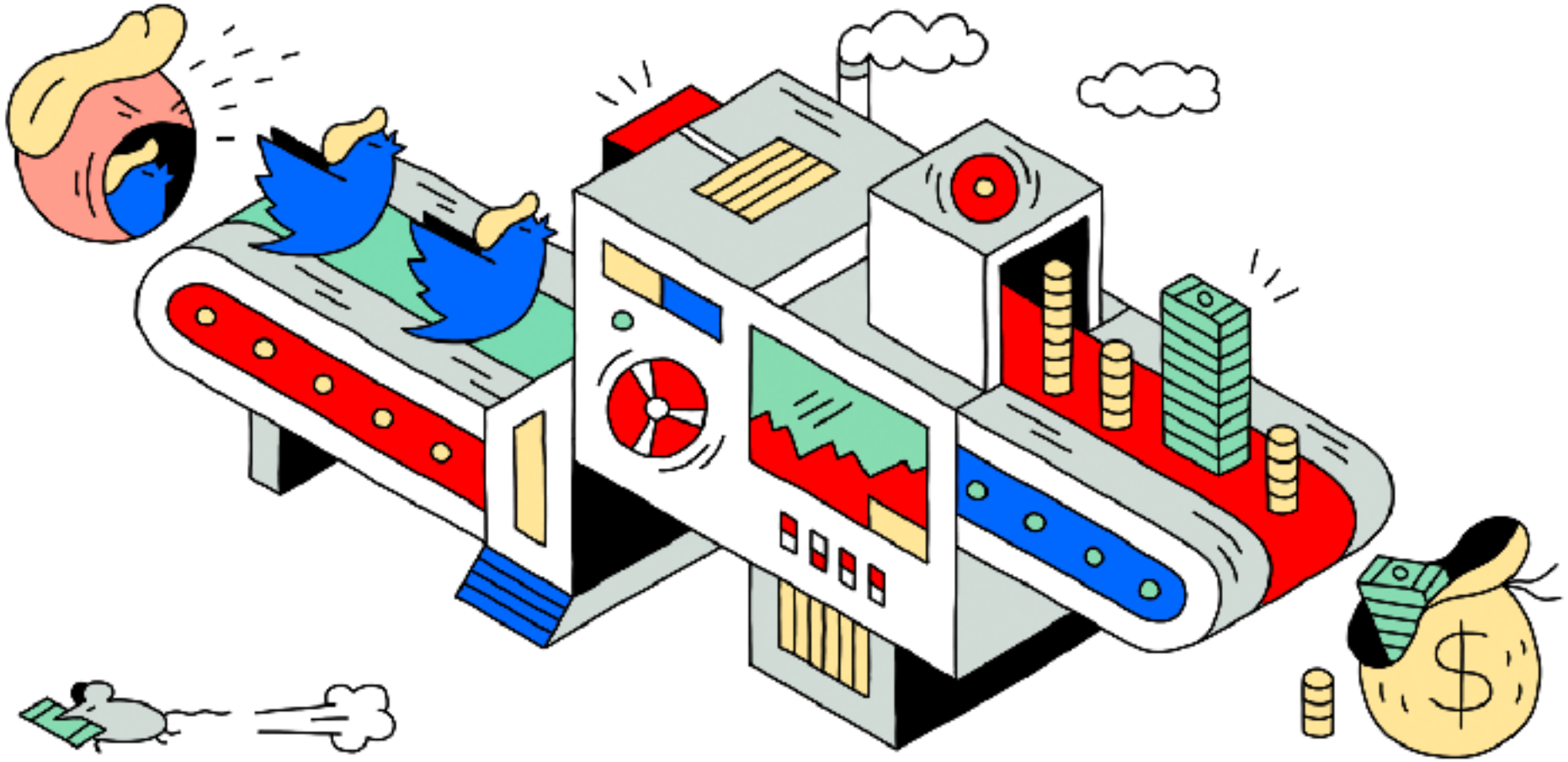
Derek Chen
Department of Computer Science
Stanford University
derekchen14@stanford.edu

Abstract

Recurrent neural networks have become prominent across many natural language processing tasks, from named entity recognition to machine translation. We apply a Seq2Seq model to the goal of creating a chatbot able to return self-generated text responses based on arbitrary inputs. In doing so, we explore the benefits of encoder-decoder networks, trade-offs between lacking data versus cross-training, various attention scoring functions, and finally, an attempt at adapting Pointer Sentinel Mixture Model to a dynamic length decoder. In order to create a chatbot of specific persona, we also scraped Twitter and the general web for public data of speech in the style of Donald Trump.

Query	Vinyals - No Cross Training	Luong - No cross training	Luong - w/ Cross Training
Do we have a trade deficit with China?	we have a trade deficit .	we owe japan .	no .
Nuclear weapons, Iran, and war have what in common?	it 's a mess .	iran is taking our economic dollars dollars .	the army .
The president is terrible	it is time to be president .	i beat hillary clinton	elitism
Crooked hillary is running for president, will she win?	crookedhillary is unfit to serve . bigleaguetruth debate	hillary will change the voter debate . bigleaguetruth debate	hillary clinton wants to makeamericagreatagain

Trump2cash



24.06.2018 Turkish Presidential Elections

DS2

- 45 Days
- 31 Cities
- 43 Rallies

Cemre Ünal DEDA class

24.06.2018 Turkish Presidential Elections

DS2

Istanbul

Gaziantep

Kahramanmaras

Istanbul - Kartal

Mardin

Denizli

Ankara

Ordu

Adana

Antalya

Yalova

Rize

24.06.2018 Turkish Presidential Elections

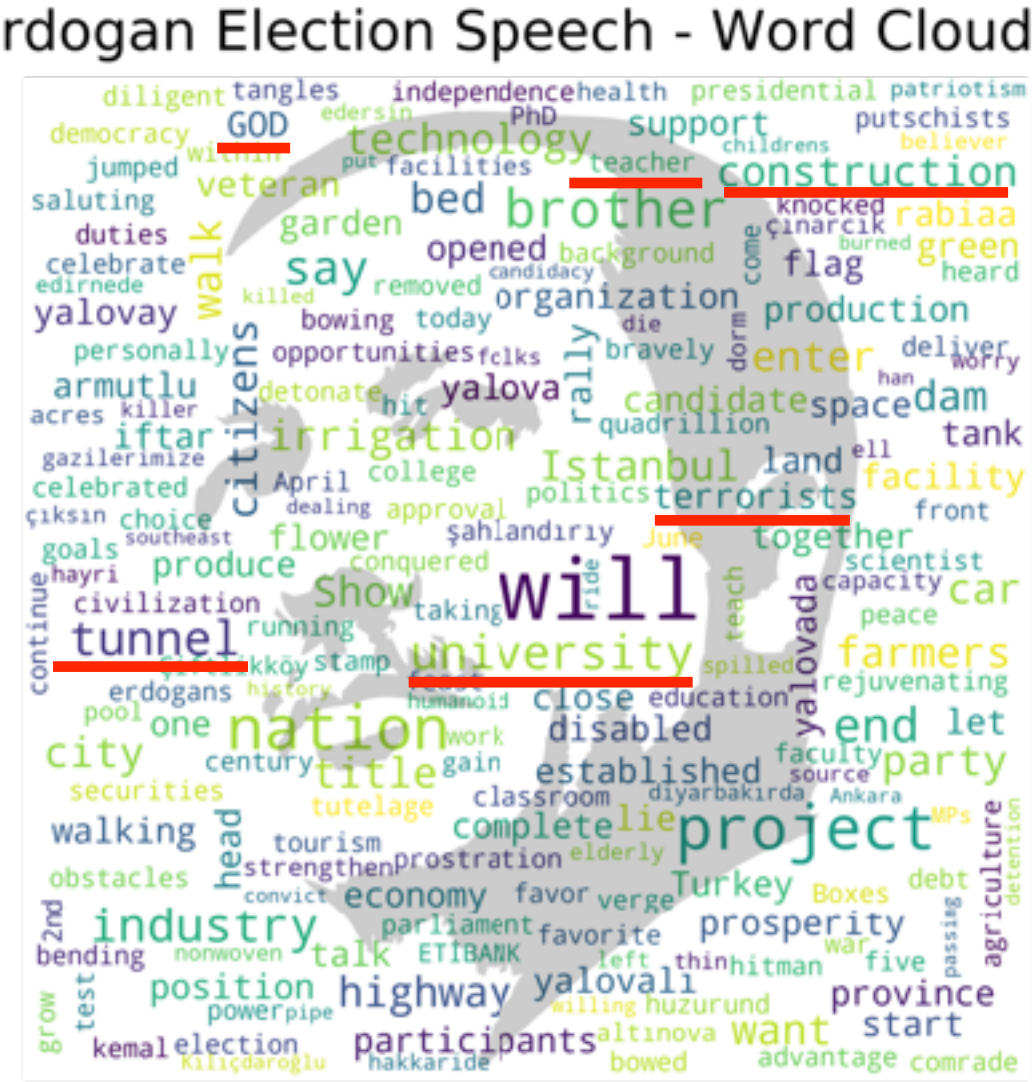
DS2

24.06.2018 Turkish Presidential Elections

Rize, Erdogan's Hometown

24.06.2018 Turkish Presidential Elections

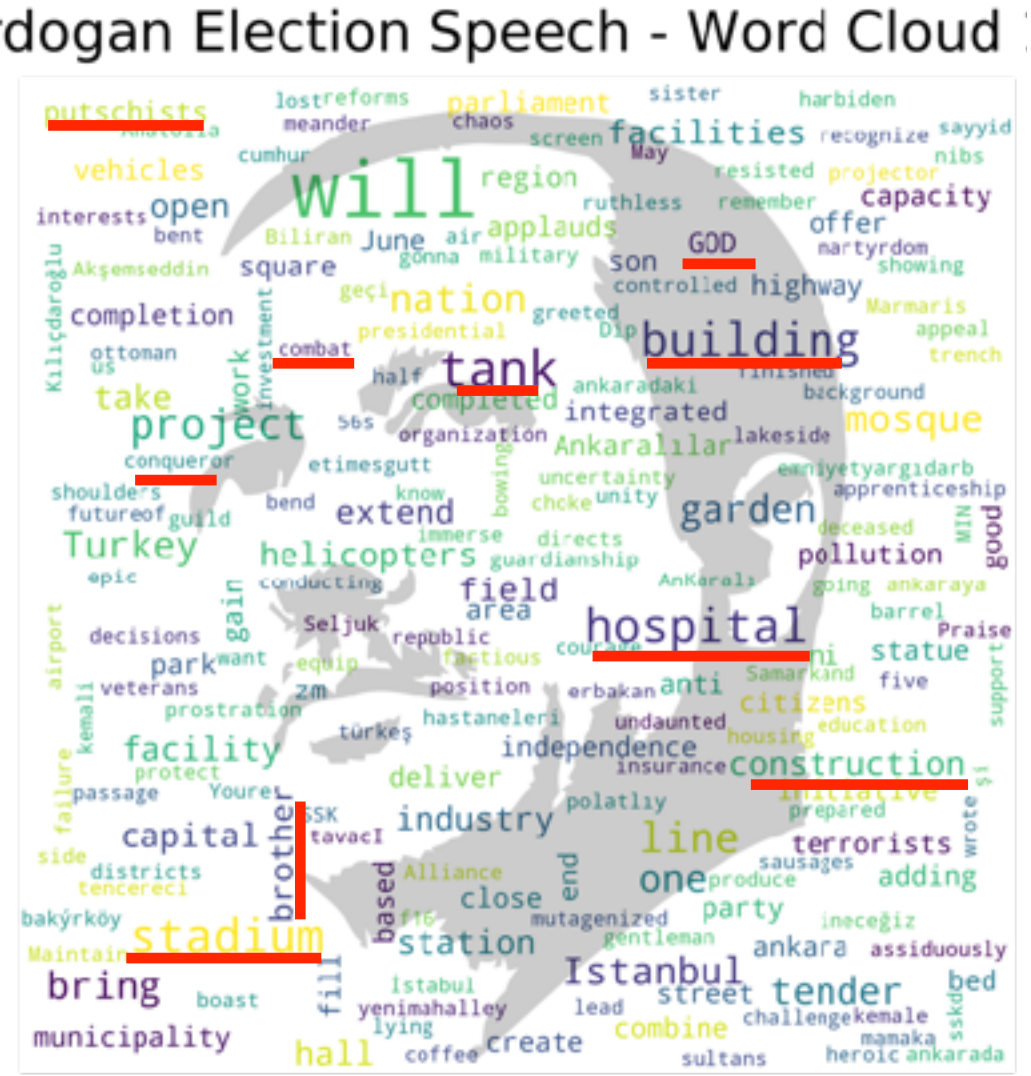
DS2



Yalova, Erdogan's Main Opponents Hometown

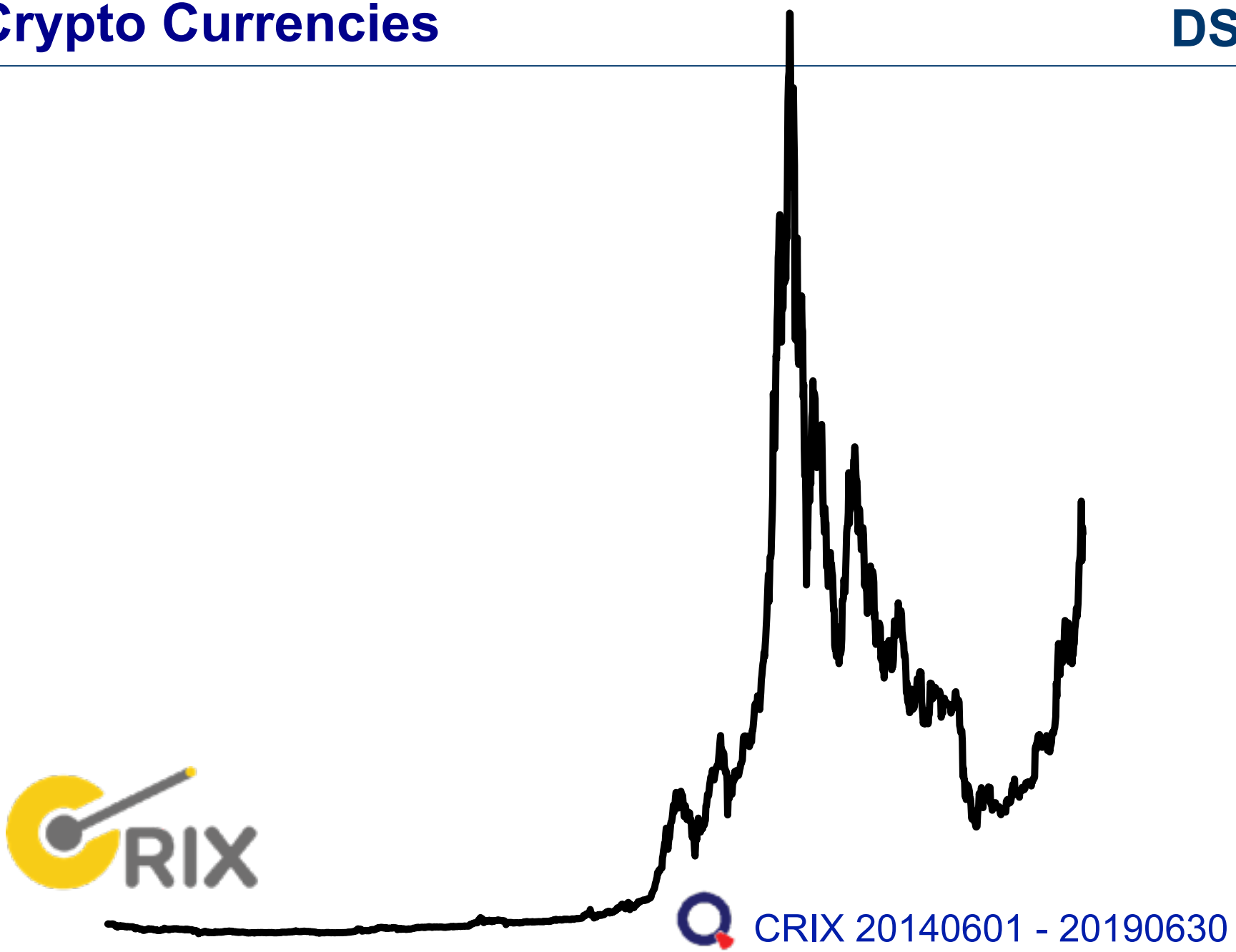
24.06.2018 Turkish Presidential Elections

DS2



Ankara, Capital of Turkey

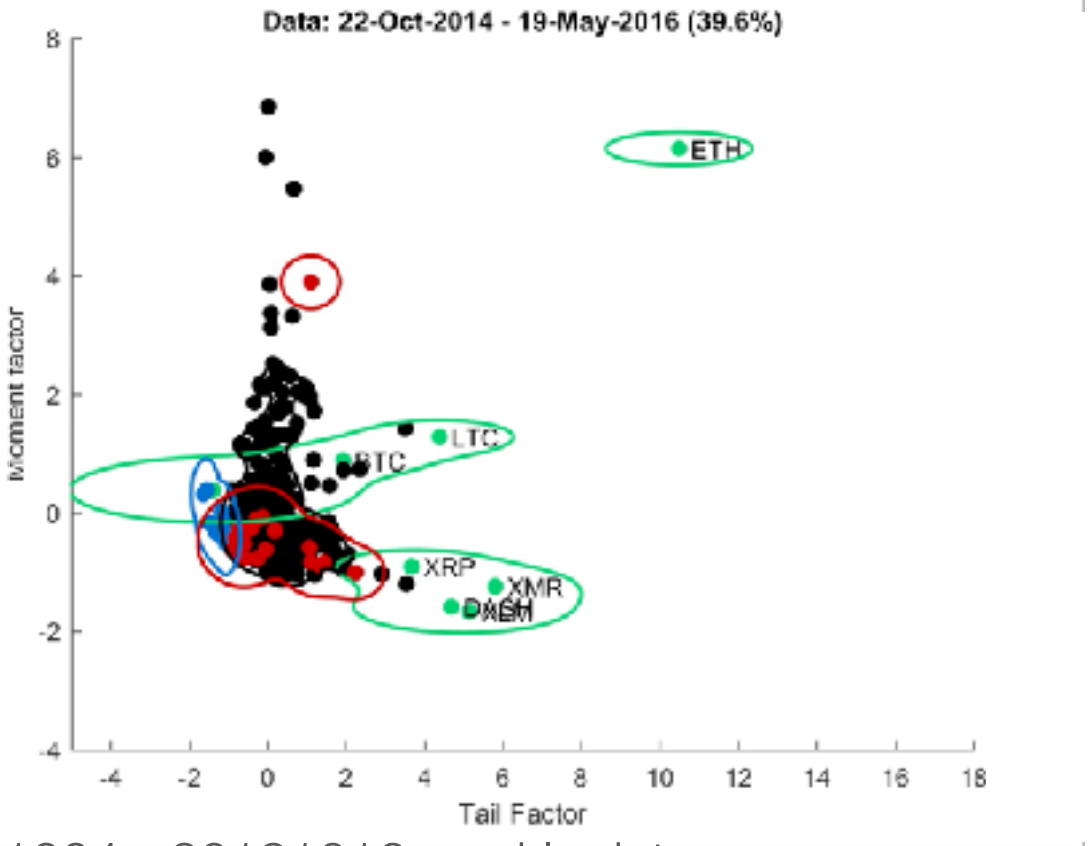
Crypto Currencies



Crypto Currencies

DS2

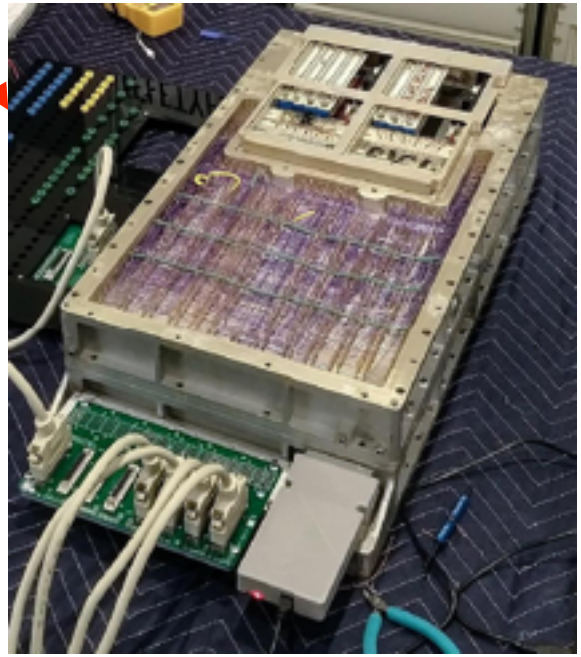
- CCs, Commodities, FX rate, Stocks (23 dimensions)
- Quantile (0.5, 1, 2.5, 5, 95, 97.5, 99, 99.5) %
- Expected Shortfall ES (0.5, 1, 2.5, 5, 95, 97.5, 99, 99.5) %
- Skew, Kurt, Var, stable, AR(1), Hurst Exponent
- Tail + Moment Factor

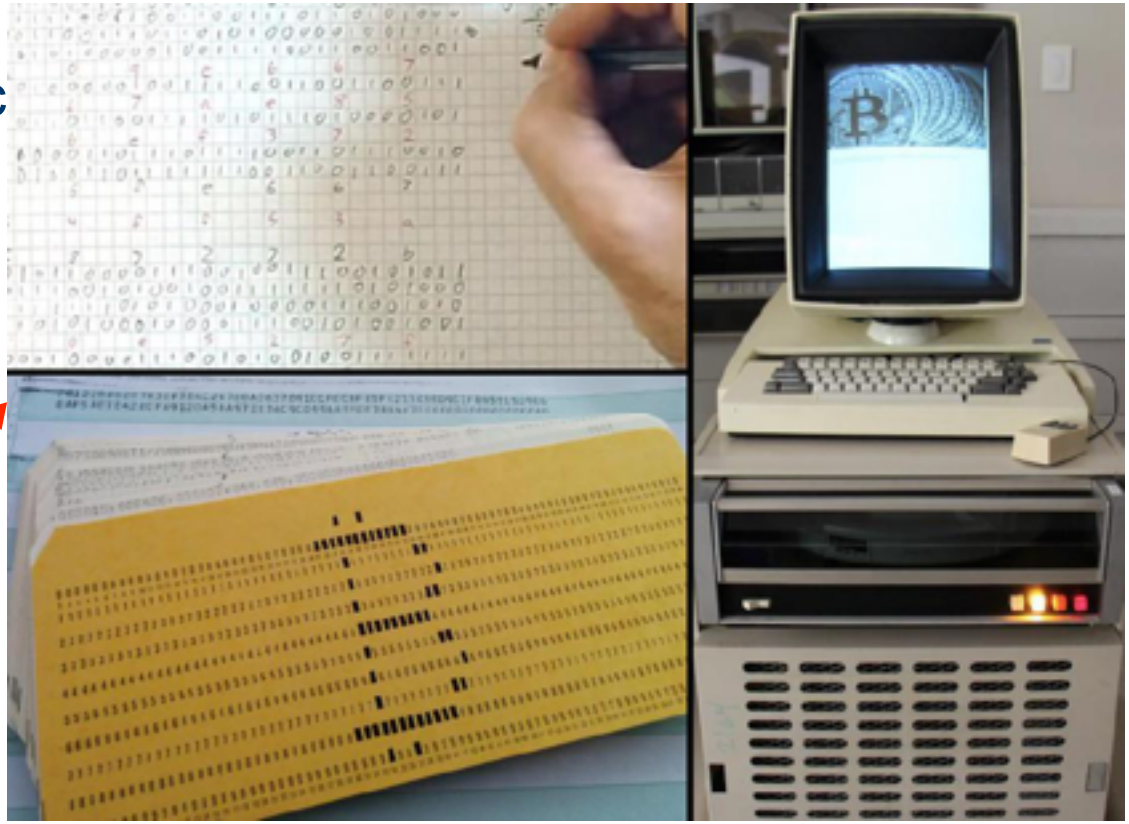


data range: 20141004 - 20181016 weekly data

Rise of the Machines

- Hand calc 0.67 hash 1D 7.8E6 sec
- 1960 Apollo Guidance Computer
  10.3 sec per hash
- 1960 Punch card computer IBM 1410
  80 sec per hash
- 1973 Xerox Alto 1.5 sec





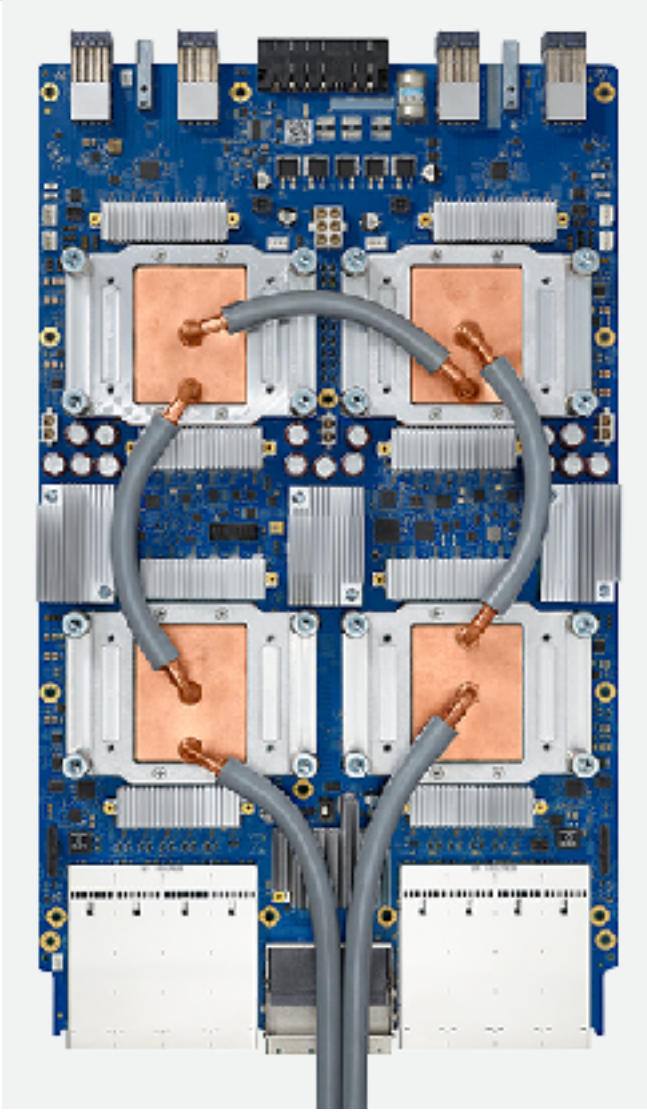
Rise of the Machines

- Cray 2 is the world's fastest supercomputer: 1985-1990
- 1.9 GFLOPs*
- 5,500 pounds
- $32 million (current $)
- 舞 锡市 太湖之光, (wu3xi1shi4tai4hu2zhi1guang1) 95 PFLOPS**

*10^9 FLoating point OPerations per Second **10^15 FLOPS

Rise of the Machines

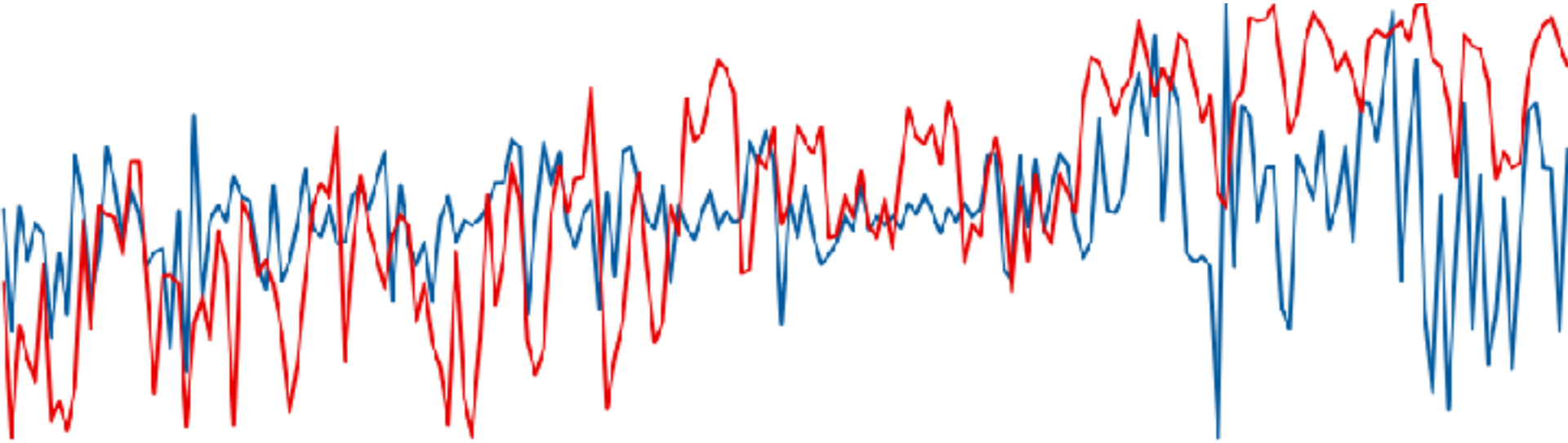
- 2017 Nvidia Titan Xp
- 125 TFLOPs, 16 GB
- 1.1 K USD
- ASICS, Antminer
- Google's Tensor Processing Unit (TPUv2)
- 180 TFLOPs, 64GB per TPU
- TPUs important for MLE*



*MLE = „Maximum Likelihood Estimation" (age>=45)
MLE = „Machine Learning in Economics" (age<=45)

CRIX and sentiments

DS2



Daily CRIX log-return, Daily stock twits sentiment
201407 - 201807

Després R, Chen CYH, Renault T (2019)

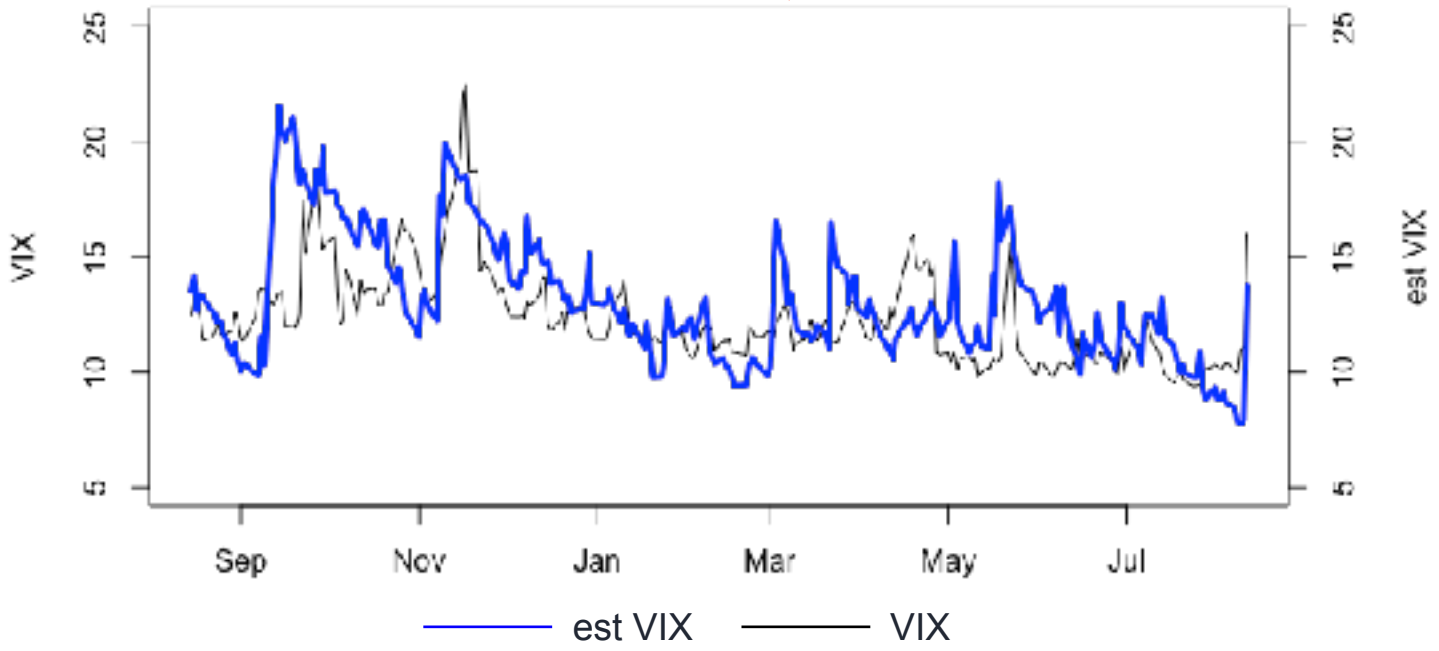
CRIX and sentiments

Term	Sentiment weight
🚀	0.90
hodl	0.54
hodl !	0.85
hackers	−0.74
miner	0.62
bitcoin 😂	−0.73
scam	−0.77
fucking scam	−0.86

Crypto specific terms

CRypto currency IndeX

adaptive λ (time-varying window estimation) to the rescue



VIX simulation

CRypto currency IndeX

- Option Pricing on CRIX and CCs (Chen CYH et al, 2018)
- Stochastic Vola Corr Jump (SVCJ)model
- VCRIX as a natural component

$$d\log Y_t = \mu dt + \sqrt{V_t} dW_{y,t} + Z_{y,t} dN_t$$
$$dV_t = \kappa(\theta - V_t)dt + \sigma_V \sqrt{V_t} dW_{v,t} + Z_{v,t} dN_t$$

Price options!!

- Portfolio choice
- Portfolio management
- Trading strategy
- Sentiment analysis
- DS2

PWC

Micro Credits

- Classics: Scoring based on financial ratios
- Modern: Score based on social network data
- Micro Credits: **webank.com**

Banking Fintech

DS2

09190701 St Wiperti, Quedlinburg, Ottone Heinrich I

20190810 Dürnstein a.d. Donau, 11921201Richard Löwenherz

Springer Handbooks of Computational Statistics

Wolfgang Karl Härdle · Henry Horng-Shing Lu · Xiaotong Shen *Editors*

Handbook of Big Data Analytics

Addressing a broad range of big data analytics in cross-disciplinary applications, this essential handbook focuses on the statistical prospects offered by recent developments in this field. To do so, it covers statistical methods for high-dimensional problems, algorithmic designs, computation tools, analysis flows and the software-hardware co-designs that are needed to support insightful discoveries from big data. The book is primarily intended for statisticians, computer experts, engineers and application developers interested in using big data analytics with statistics. Readers should have a solid background in statistics and computer science.

Visit the Quantlet platform. The Quantlet platform quantlet.de, quantlet.com, quantlet.org is an integrated QuantNet environment consisting of different types of statistics-related documents and program codes. Its goal is to promote reproducibility and offer a platform for sharing validated knowledge native to the social web. QuantNet and the corresponding Data-Driven Documents-based visualization allows readers to reproduce the tables, pictures and calculations inside this Springer book.

Statistics

ISBN 978-3-319-18283-4

9 783319 182834

springer.com

Härdle · Lu · Shen *Eds.*

Handbook of Big Data Analytics

Springer Handbooks of Computational Statistics

Wolfgang Karl Härdle
Henry Horng-Shing Lu
Xiaotong Shen *Editors*

Handbook of Big Data Analytics

QUANTLETS

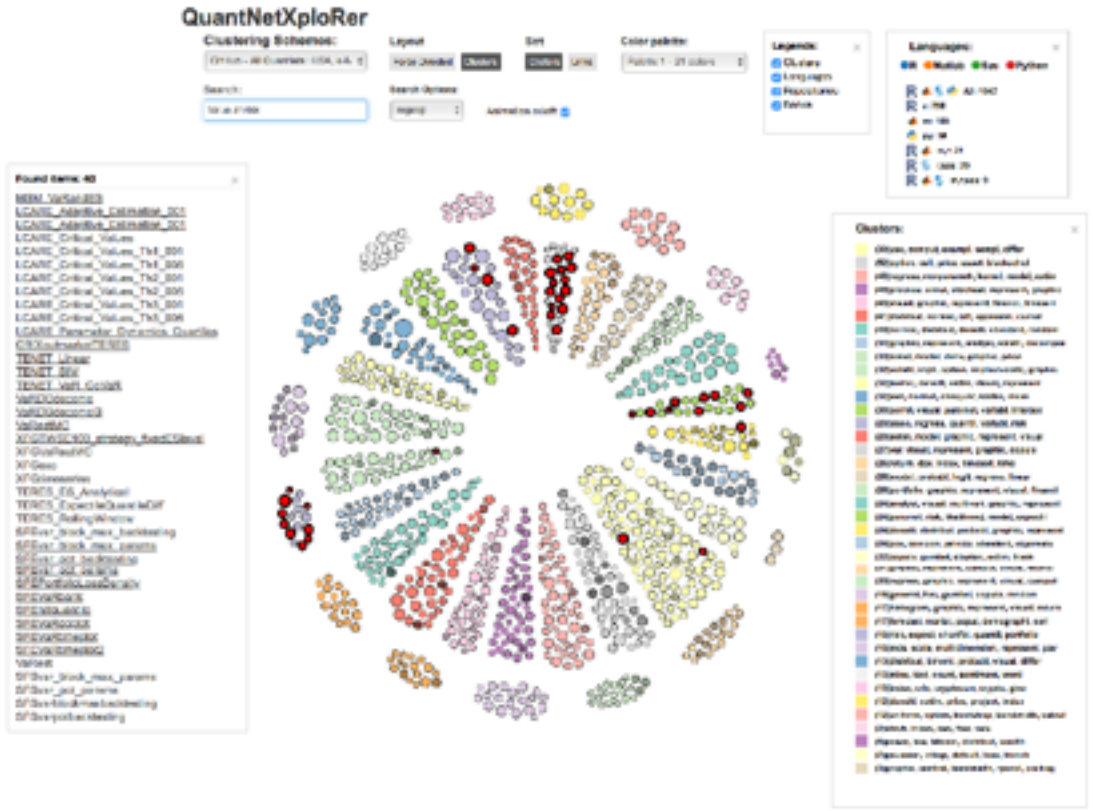
Springer

FinTech sources

FinTec data science

https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html

Dynamic Topic Modelling

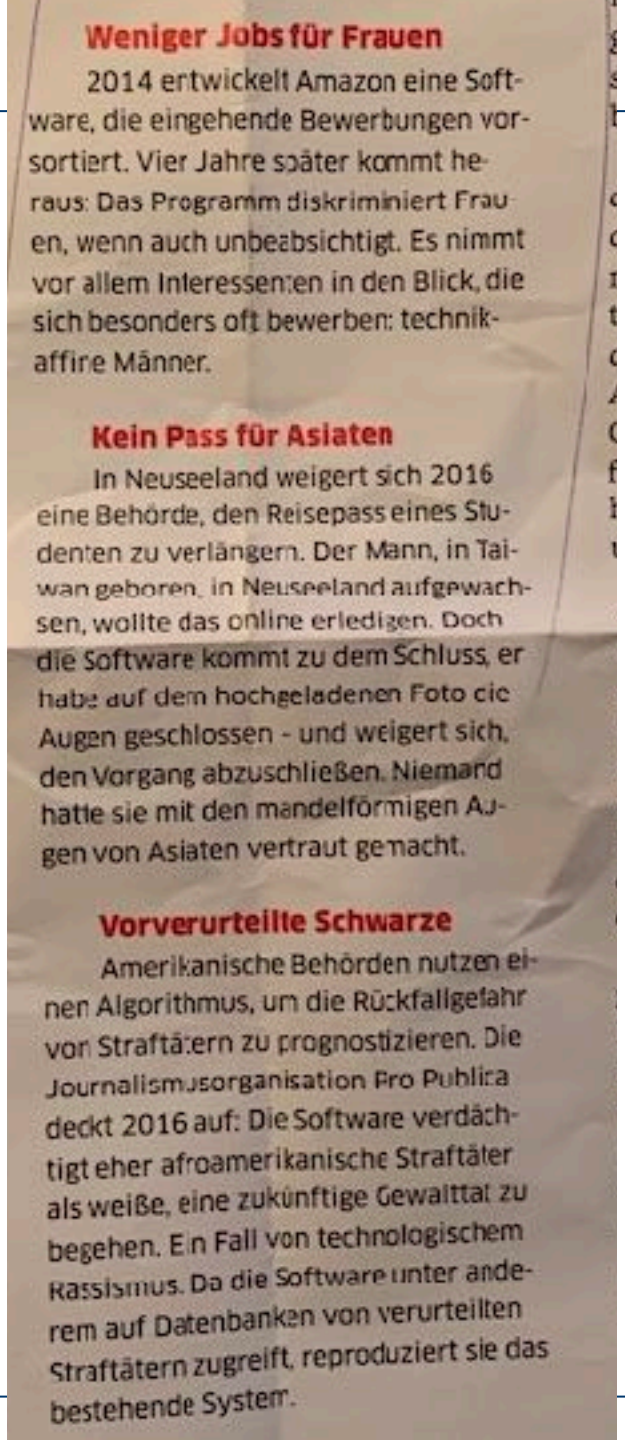


Crypto option pricer

https://github.com/QuantLet/SVCJOptionApp/tree/master/

危险 Gefahr

- Gender bias
- no 护照
- Afro Americans



Weniger Jobs für Frauen

2014 entwickelt Amazon eine Software, die eingehende Bewerbungen vorsortiert. Vier Jahre später kommt heraus: Das Programm diskriminiert Frauen, wenn auch unbeabsichtigt. Es nimmt vor allem Interessenten in den Blick, die sich besonders oft bewerben: technikaffine Männer.

Kein Pass für Asiaten

In Neuseeland weigert sich 2016 eine Behörde, den Reisepass eines Studenten zu verlängern. Der Mann, in Taiwan geboren, in Neuseeland aufgewachsen, wollte das online erledigen. Doch die Software kommt zu dem Schluss, er habe auf dem hochgeladenen Foto die Augen geschlossen - und weigert sich, den Vorgang abzuschließen. Niemand hatte sie mit den mandelförmigen Augen von Asiaten vertraut gemacht.

Vorverurteilte Schwarze

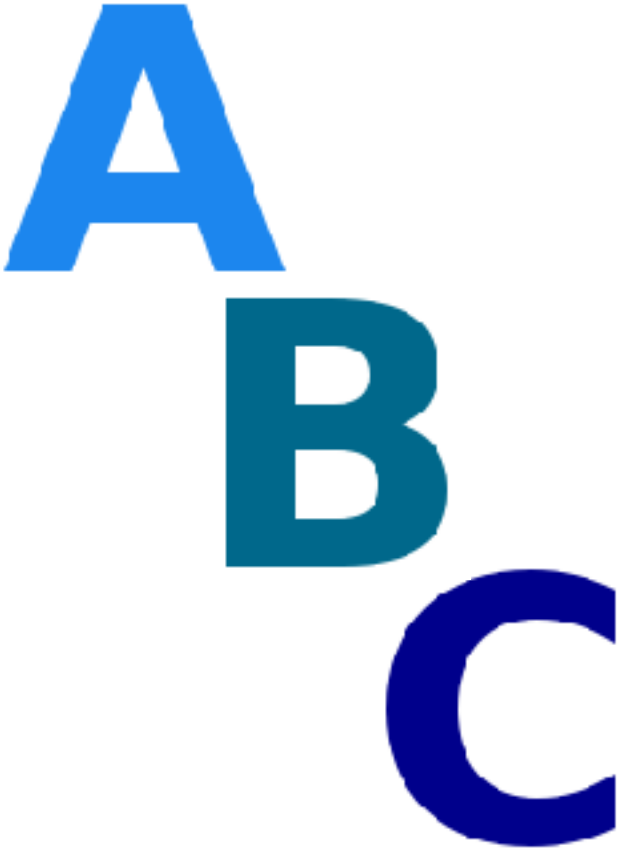
Amerikanische Behörden nutzen einen Algorithmus, um die Rückfallgefahr von Straftätern zu prognostizieren. Die Journalismusorganisation Pro Publica deckt 2016 auf: Die Software verdächtigt eher afroamerikanische Straftäter als weiße, eine zukünftige Gewalttat zu begehen. Ein Fall von technologischem Rassismus. Da die Software unter anderem auf Datenbanken von verurteilten Straftätern zugreift, reproduziert sie das bestehende System.

State, nation? data control center!

DS2

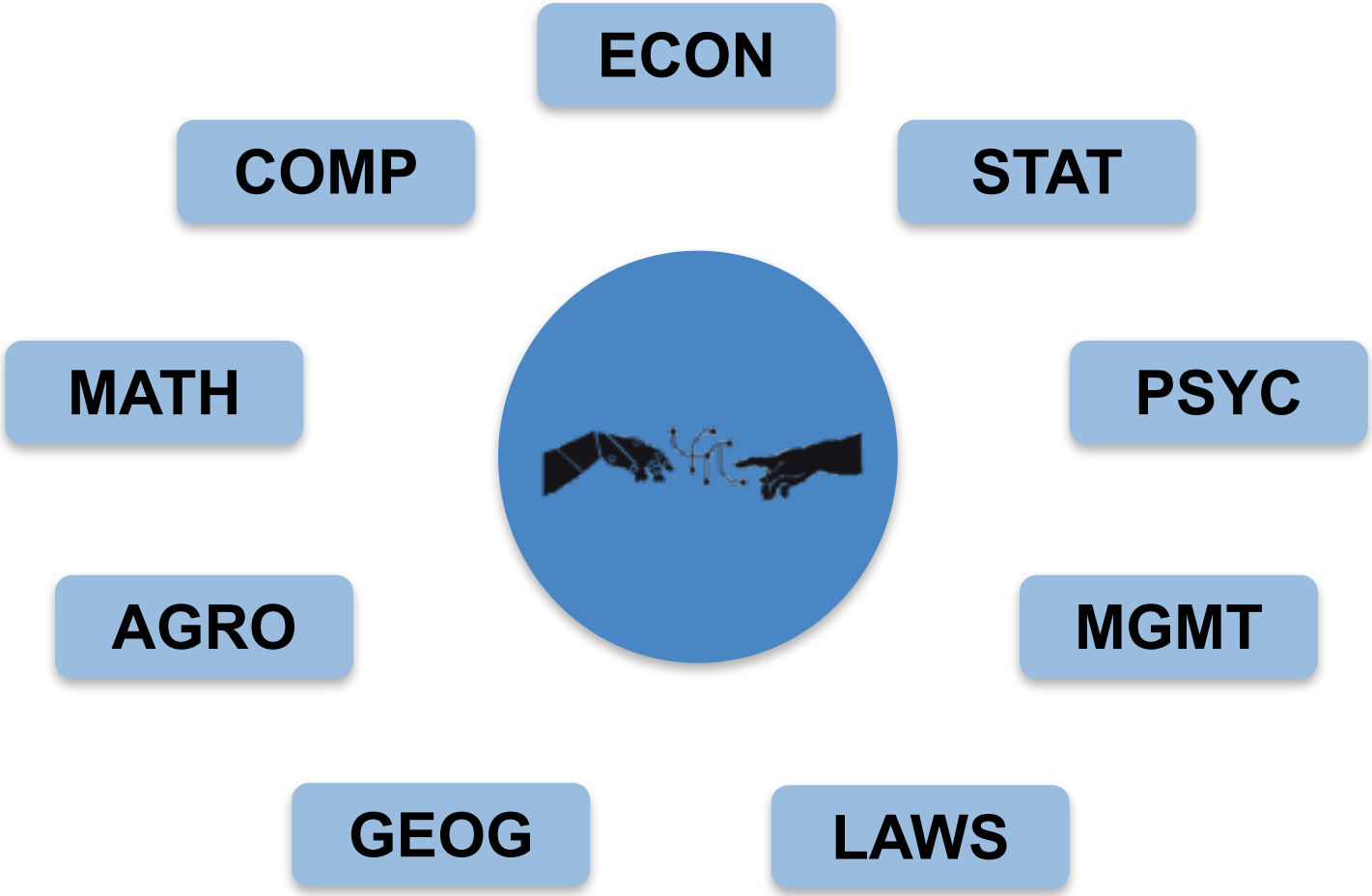
- Data & service provision
- Transparency & replicability
- Massive Open Online Research

High Frequency Markets
Dynamic Topic Modelling
Forecasting Volatility
Investors Preferences
Dynamic Risk Structuring
Cyber security insurance
Herding behavior
FinTec data science
System Network Risks
Transparency of Quant Data
Crypto currencies acceptance
State = Data Control Center



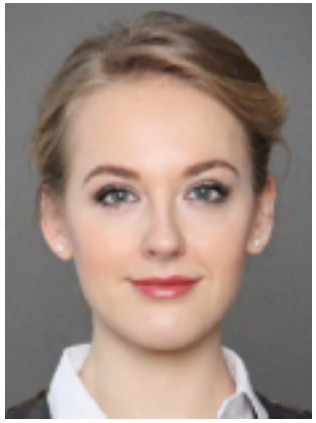
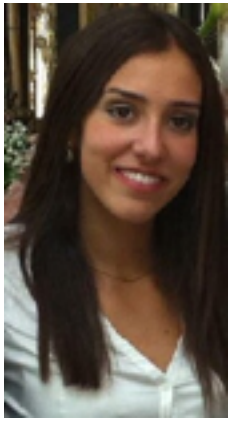
Data Science & Digital Society

DS2



DS2

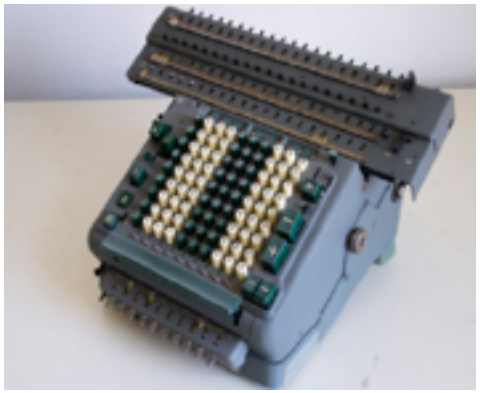


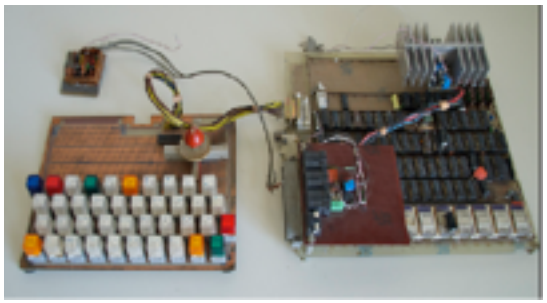



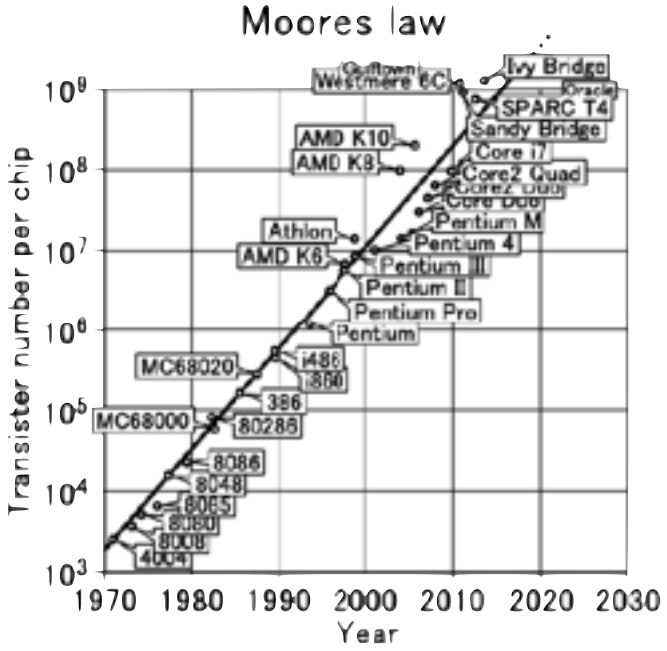



Chen Härdle Kim Matic Petukhina Reule Spilak Engelmann Trimborn Ünal

Rise of the Machines







https://phys.org/news/2015-08-silicon-limits-power-electronics-revolution.html

- http://computermuseum.wiwi.hu-berlin.de

Rise of the Machines

- 2016 iPhone 7* , 178 GFLOPs
- 2019 iPhone XS Max* , 1300 GFLOPs
- A12 Bionic Chip w/ Neural Engine, 256 GB storage
- 138 g
- 1.3 K EUR

*A10 Fusion. The Apollo guidance system had only 4K of RAM.

Rise of the Machines

DS2






- 1981 - $300,000
- 1987 - $50,000
- 1990 - $10,000
- 1994 - $1,000
- 1997 - $100
- 2000 - $10
- 2004 - $1
- 2010 - $0.10
- 2017 - $0.01
- 2018 - $0.004*

Cost per GB storage p/m

Event Stream

20170327 ASE, Bukarest, RO
20181022 厦门大学， 厦门, 中国
20190903 中山大学， 广州，中国
20190917 厦门大学， 厦门，中国
20190924 蒲江，北京， 成都, 中国
20191028 St Gallen U, CH, SDA class
20200206 Copenhagen, DK, FINTECH
20200309 Klagenfurt U, SDA class
20200420 HU Berlin, DEDA Class