Math, Data Science and Social Impact

André C. P. L. F. de Carvalho
Universidade de São Paulo
andre@icmc.usp.br

Data Science
São Paulo School of Advanced Science
on Learning from Data

Summary

- Origins of Data Science
- Mathematics of Data Science
- Responsible Data Science
- Social Data Science

The beginning...

- Babylon, 4000 BC
  - Regular censuses were conducted to decide how much food to produce
    - To feed its population
  - Census Records were written on lay plates

Eat

Data Science is not new

- 1962 John Tucker suggests that statistics should move towards data analysis
- 1974 Peter Naur
  - Turing Awards (2005)
  - Famous for the BNF notation

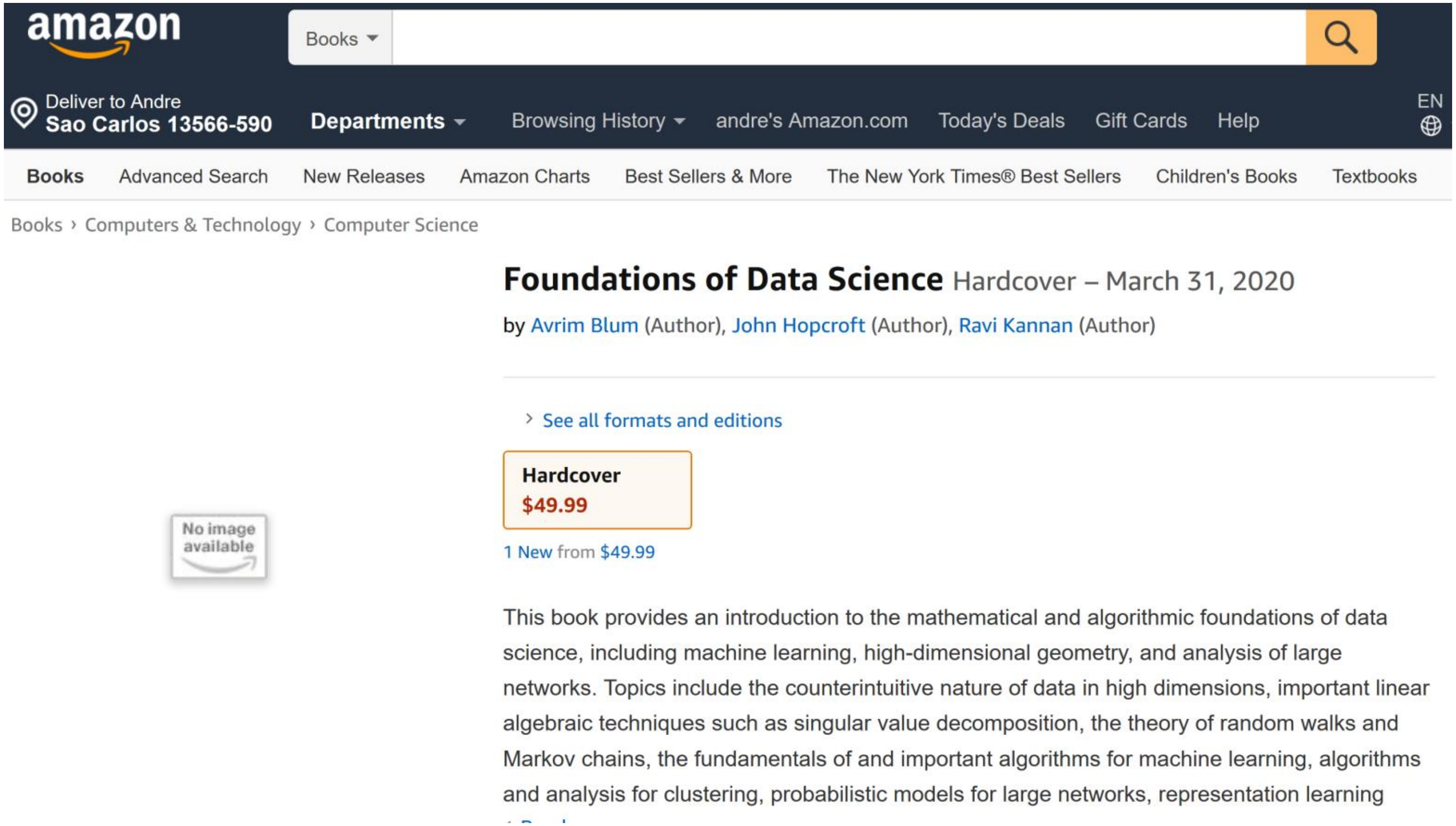
> Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences.

Concise Survey of Computer Methods

Mathematics and Data Science

- More Science than Data
- Mathematics is the main tool
  - Analysis (differential and integral calculus)
  - Discrete Mathematics
  - Linear Algebra
  - Optimization and Operational Research
  - Statistics
  - ...

Mathematics and Data Science

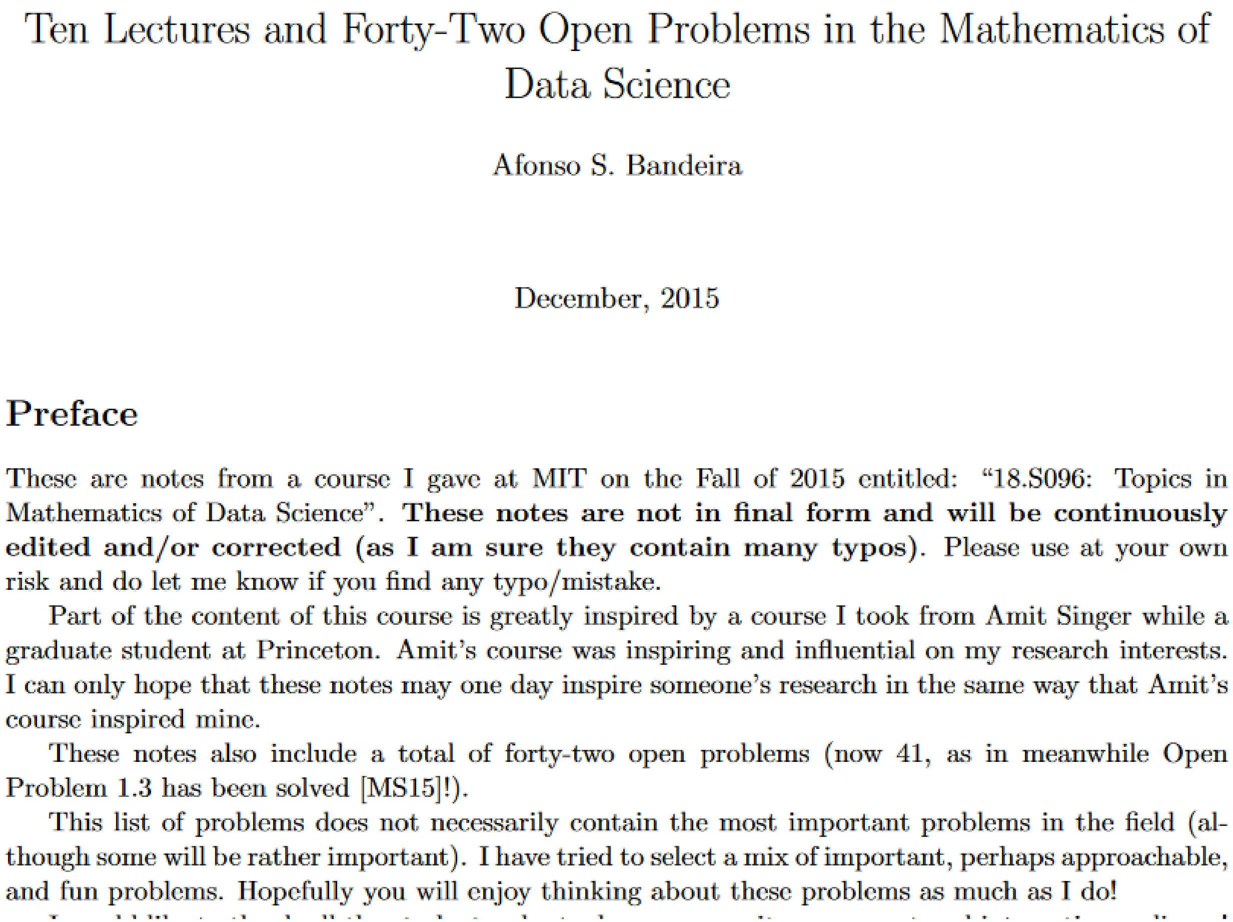


Foundations of Data Science Hardcover – March 31, 2020

by Avrim Blum (Author), John Hopcroft (Author), Ravi Kannan (Author)

This book provides an introduction to the mathematical and algorithmic foundations of data science, including machine learning, high-dimensional geometry, and analysis of large networks. Topics include the counterintuitive nature of data in high dimensions, important linear algebraic techniques such as singular value decomposition, the theory of random walks and Markov chains, the fundamentals of and important algorithms for machine learning, algorithms and analysis for clustering, probabilistic models for large networks, representation learning

Mathematics and Data Science



Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science

Afonso S. Bandeira

December, 2015

Preface

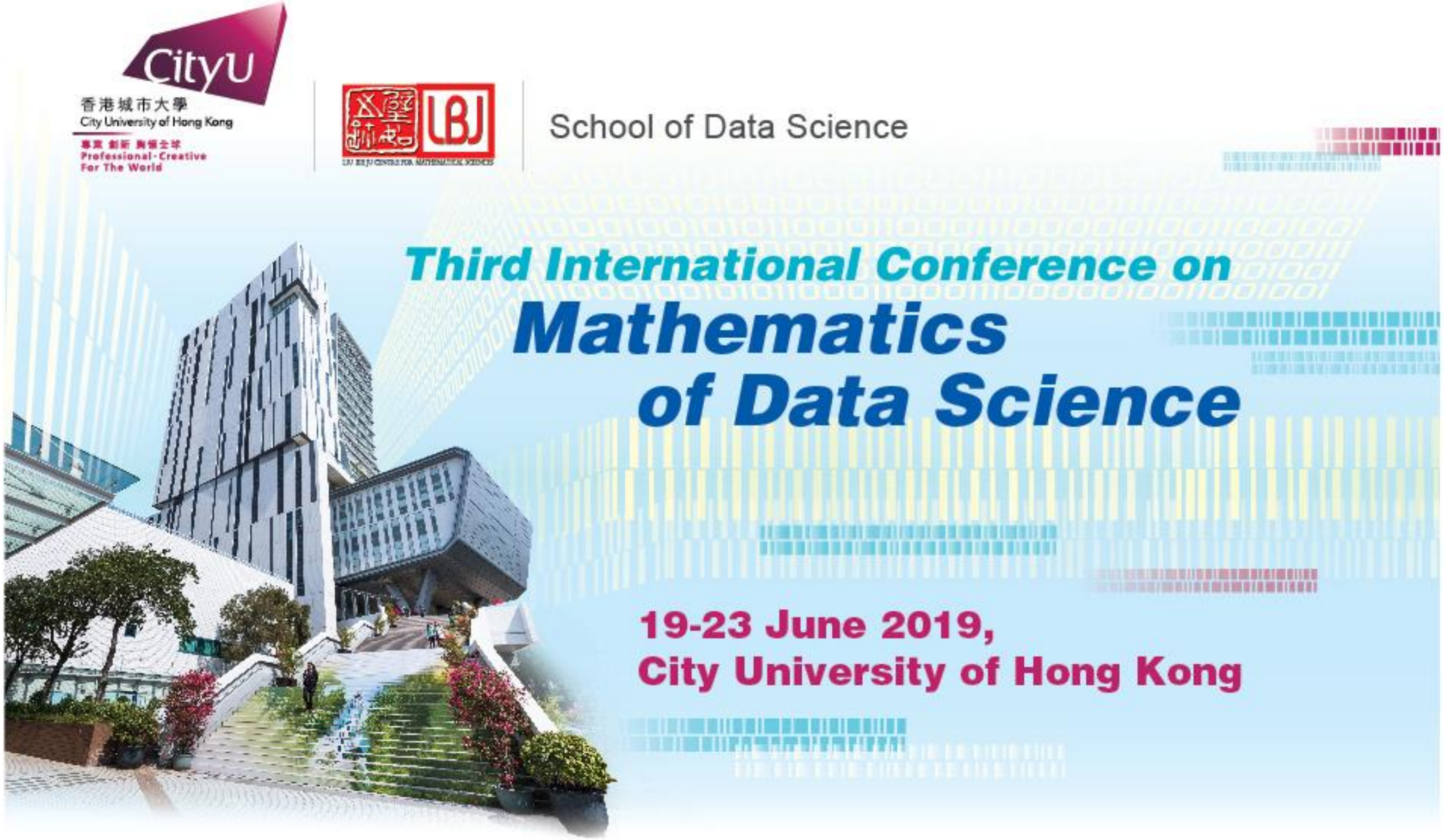
These are notes from a course I gave at MIT on the Fall of 2015 entitled: "18.S096: Topics in Mathematics of Data Science". **These notes are not in final form and will be continuously edited and/or corrected (as I am sure they contain many typos).** Please use at your own risk and do let me know if you find any typo/mistake.

Part of the content of this course is greatly inspired by a course I took from Amit Singer while a graduate student at Princeton. Amit's course was inspiring and influential on my research interests. I can only hope that these notes may one day inspire someone's research in the same way that Amit's course inspired mine.

These notes also include a total of forty-two open problems (now 41, as in meanwhile Open Problem 1.3 has been solved [MS15]!).

This list of problems does not necessarily contain the most important problems in the field (although some will be rather important). I have tried to select a mix of important, perhaps approachable, and fun problems. Hopefully you will enjoy thinking about these problems as much as I do!

Mathematics and Data Science



CityU
香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

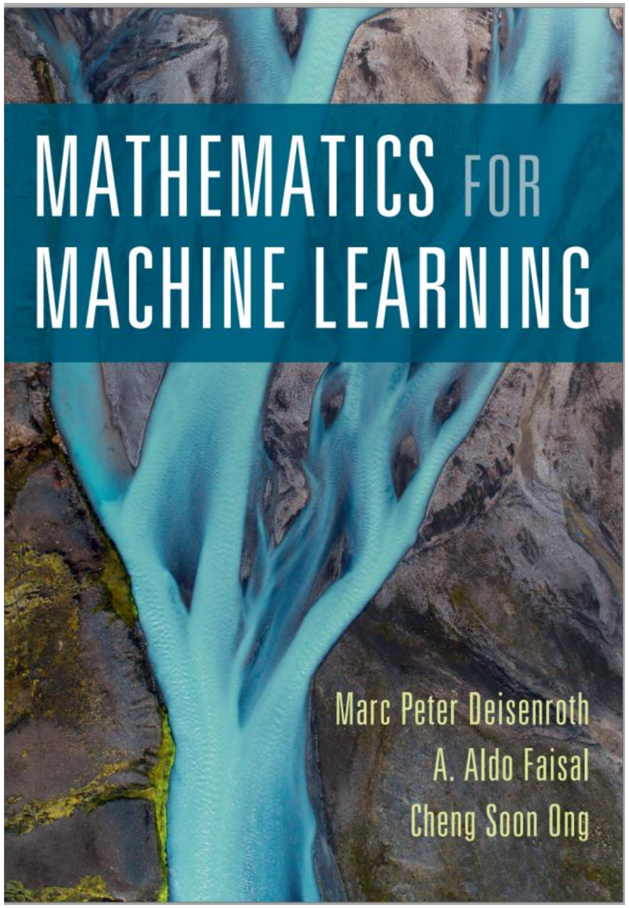
LBJ
LIU BIE JU CENTRE FOR MATHEMATICAL SCIENCES

School of Data Science

Third International Conference on
Mathematics
of Data Science

19-23 June 2019,
City University of Hong Kong

Mathematics and Data Science



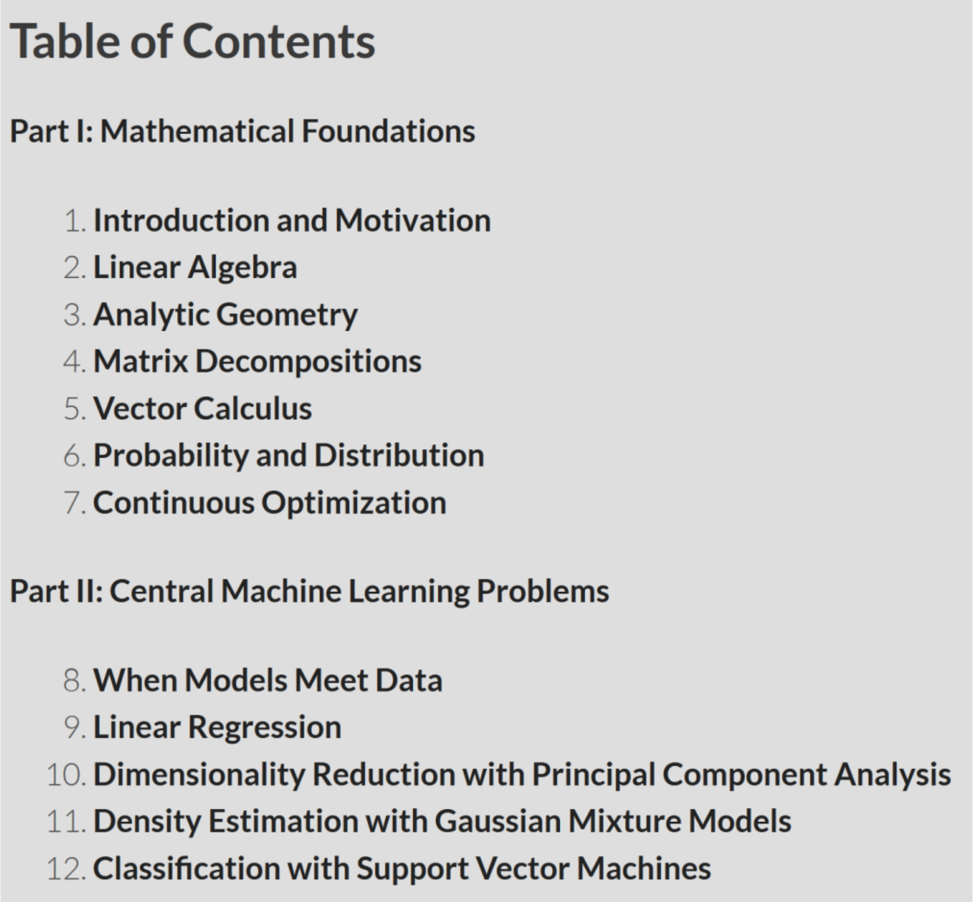


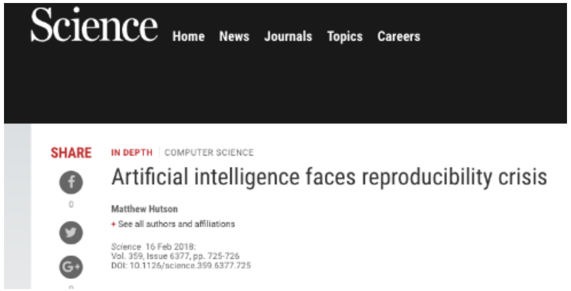
Table of Contents

Part I: Mathematical Foundations

Part II: Central Machine Learning Problems

Responsible Data Science

- Accountability
  - Who is responsible?
- Reproducibility
  - Data, code and experimental choices must be publicly available
- Privacy
  - Knowing 300 likes, ML can predict someone personality better than her/his partner



IN DEPTH | COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Matthew Hutson

+ See all authors and affiliations

Science 16 Feb 2018:
Vol. 359, Issue 6377, pp. 725-726
DOI: 10.1126/science.359.6377.725

Responsible Data Science

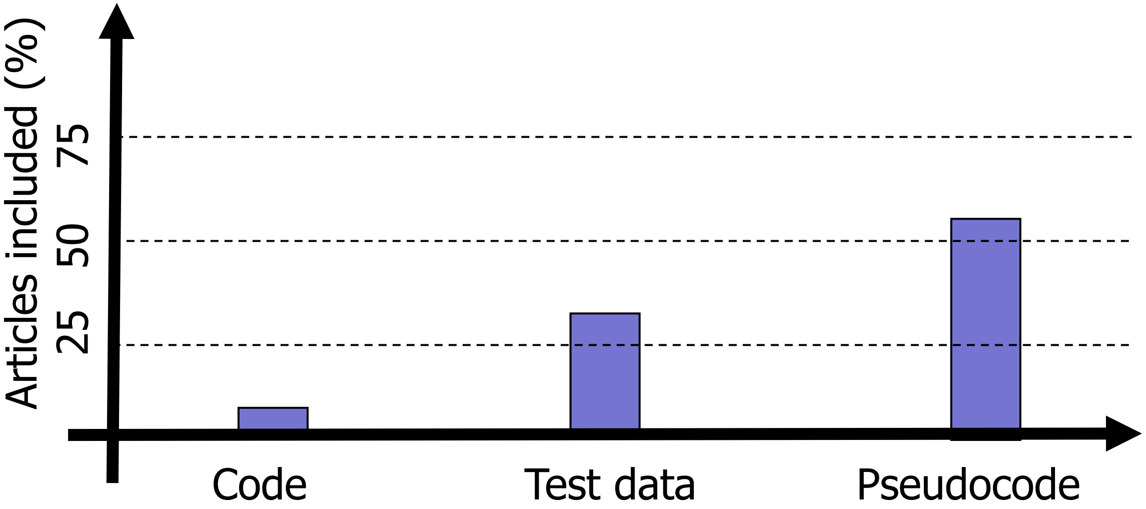
- Transparency
  - Right to explanation
    - General Data Protection Regulation (GDPR-EU)
    - Explainable AI (XAI)
- Fairness
  - Avoid decisions can be based on sensitive features
    - E.g. Citizenship, Gender, Race
  - Fair Information Practices

Reproducibility

- ML researchers compare new algorithms to existing alternatives
  - However, code of related alternatives are not often available
    - Reproducibility crisis
  - Medicine and Psychology went through a similar crises in the last decade

Reproducibility

- AI researchers do not share their code
  - Survey with 400 algorithms propose in the 2 main AI conferences



Fonte: http://www.sciencemag.org/news/2018/02/missing-data-hinder-replication-artificial-intelligence-studiesF

Reproducibility

- Main reasons for not sharing
  - Code is not finished
  - Code belongs to the company
  - Code depends on another code, not yet published
  - To keep ahead of competitors
  - Code lost because of computer is broken or was stolen
    - ”My dog ate my program” reason
      - Nicolas Rougier, INRIA, France

Privacy protection

- Fair Information Practices (FIPs) for data
  - Collection
  - Access
  - Share
  - Use
- Also known as
  - Fair information principles
  - Fair Information Practice Principles

10 FIP principles

1. Collection: limited, lawful and by fair means; with consent or knowledge
2. Data quality: relevant, accurate, up-to-date
3. Purpose specification at time of collection
4. Notice of purpose and rights at time of collection
5. Uses limited (including disclosures) to purposes specified or compatible

10 FIP principles

6. Security through reasonable safeguards
7. Transparency of personal data practices
8. Individual right of access
9. Individual right to modify, complete and remove data
10. Data controllers accountable for implementation

Brazilian Data Protection Law

- Based on General Data Protection Regulation
- Protect, but differentiate Personal and sensitive data
  - Personal data:
    - Can identify the person (name, photo, national identification, biometry,…)
  - Sensitive personal data:
    - Can lead to biased decisions (race, gender, political preferences, religion, health, genetic, biometry, …)
- National data protection agency

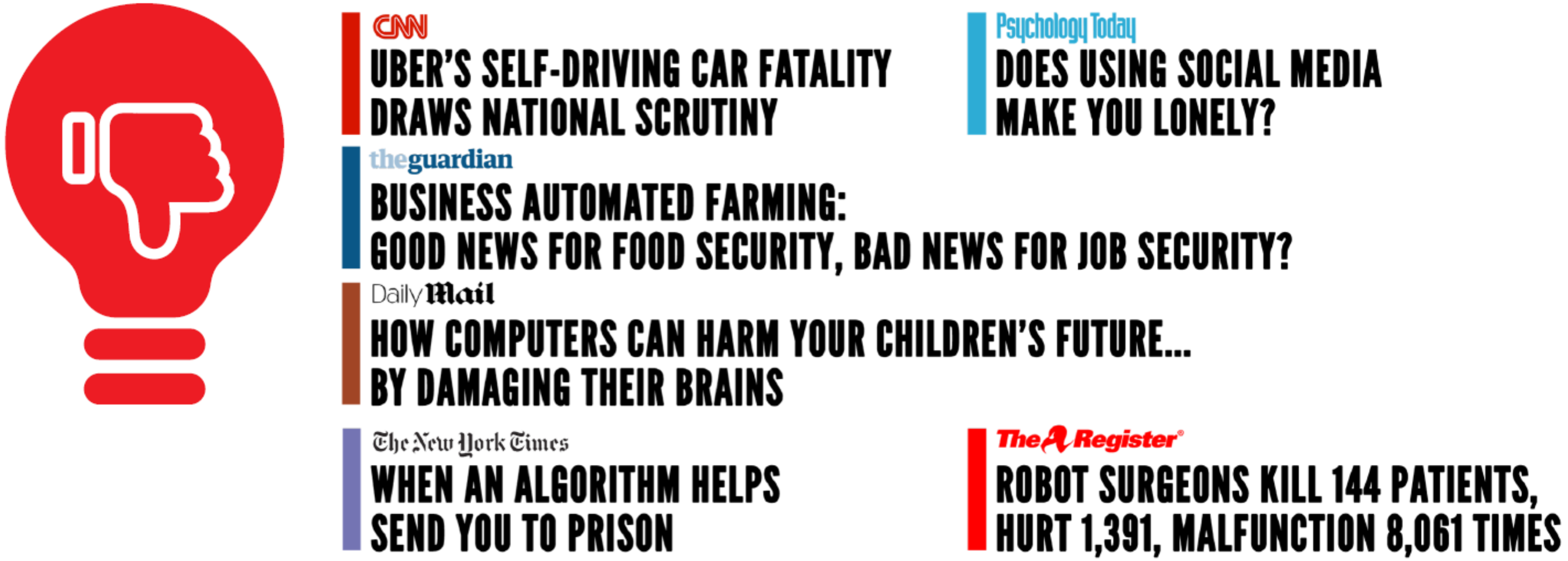
Fairness

- Decisions take by ML models can seriously affect individuals
  - Some decisions can be based on sensitive features
    - Citizenship
    - Gender
    - Race
  - Decisions based on sensitive features can lead to illegal or unfair discrimination of subgroups
  - Very active research area

Data Science for Bad

UMassAmherst | College of Information & Computer Sciences

Bad News



CNN
UBER'S SELF-DRIVING CAR FATALITY DRAWS NATIONAL SCRUTINY

Psychology Today
DOES USING SOCIAL MEDIA MAKE YOU LONELY?

theguardian
BUSINESS AUTOMATED FARMING:
GOOD NEWS FOR FOOD SECURITY, BAD NEWS FOR JOB SECURITY?

Daily Mail
HOW COMPUTERS CAN HARM YOUR CHILDREN'S FUTURE...
BY DAMAGING THEIR BRAINS

The New York Times
WHEN AN ALGORITHM HELPS SEND YOU TO PRISON

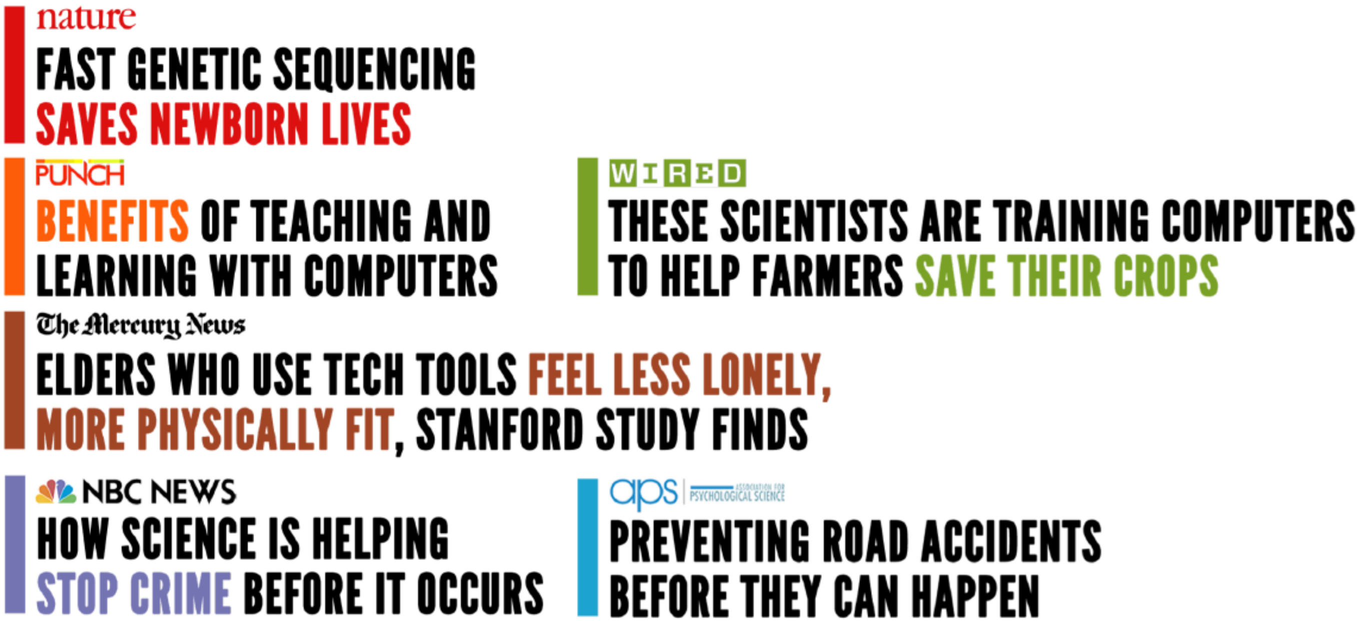
The Register
ROBOT SURGEONS KILL 144 PATIENTS, HURT 1,391, MALFUNCTION 8,061 TIMES

Laura Haas, Umass Amherst

Data Science for Good

UMass Amherst | College of Information & Computer Sciences

Good News



nature

FAST GENETIC SEQUENCING SAVES NEWBORN LIVES

PUNCH

BENEFITS OF TEACHING AND LEARNING WITH COMPUTERS

WIRED

THESE SCIENTISTS ARE TRAINING COMPUTERS TO HELP FARMERS SAVE THEIR CROPS

The Mercury News

ELDERS WHO USE TECH TOOLS FEEL LESS LONELY, MORE PHYSICALLY FIT, STANFORD STUDY FINDS

NBC NEWS

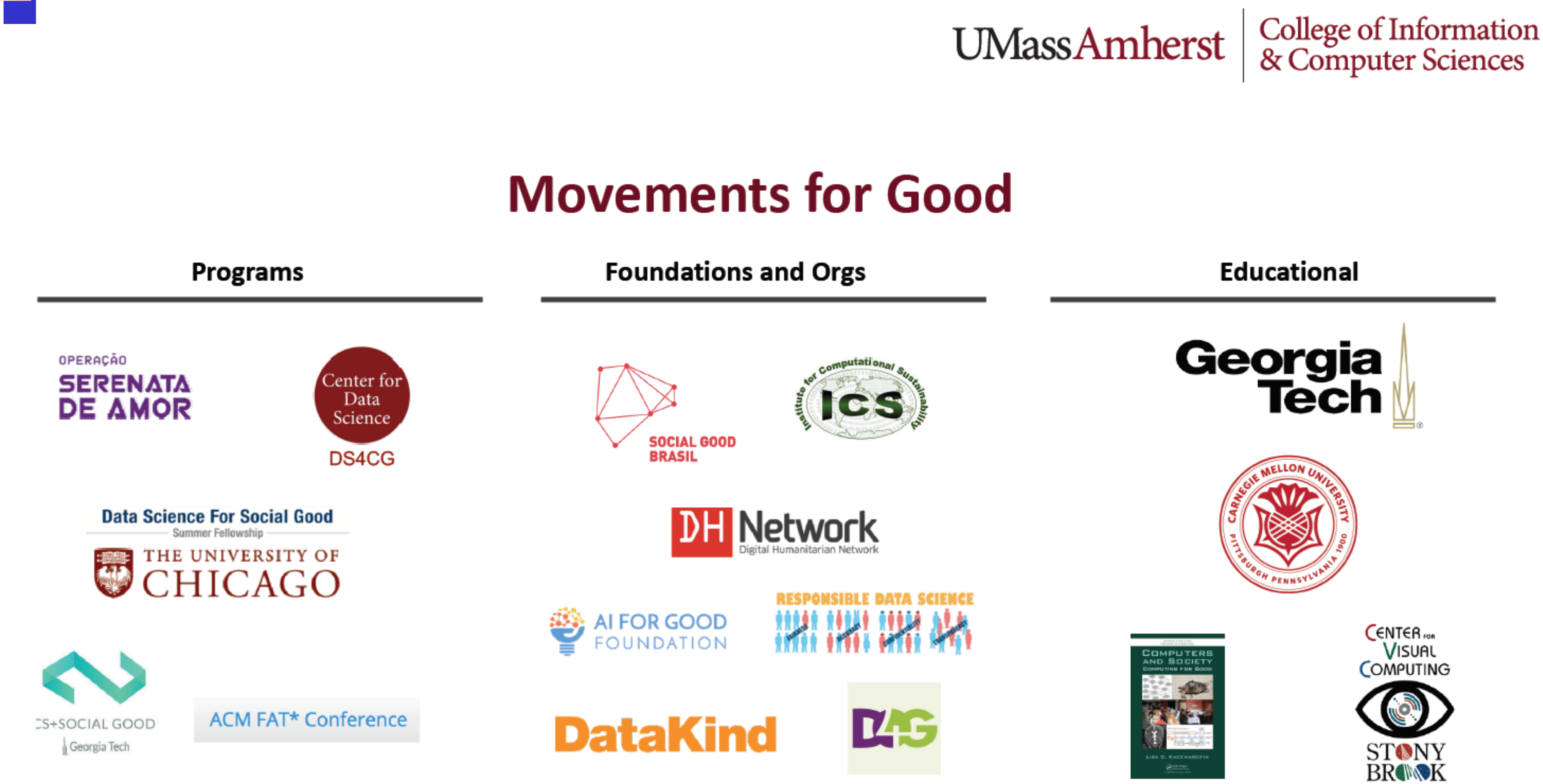
HOW SCIENCE IS HELPING STOP CRIME BEFORE IT OCCURS

aps | ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

PREVENTING ROAD ACCIDENTS BEFORE THEY CAN HAPPEN

Laura Haas, Umass Amherst

Data Science for Good



UMassAmherst | College of Information & Computer Sciences

Movements for Good

Programs

OPERAÇÃO SERENATA DE AMOR

Center for Data Science DS4CG

Data Science For Social Good Summer Fellowship THE UNIVERSITY OF CHICAGO

CS+SOCIAL GOOD Georgia Tech

ACM FAT* Conference

Foundations and Orgs

SOCIAL GOOD BRASIL

Institute for Computational Sustainability ICS

DH Network Digital Humanitarian Network

AI FOR GOOD FOUNDATION

RESPONSIBLE DATA SCIENCE

DataKind

D4G

Educational

Georgia Tech

CARNEGIE MELLON UNIVERSITY PITTSBURGH PENNSYLVANIA 1900

COMPUTERS AND SOCIETY COMPUTING FOR GOOD

CENTER FOR VISUAL COMPUTING STONY BROOK

Agradecimento a Laura Haas, Umass Amherst

Data Science for good

- University of Chicago summer program
- Non-profit movement
  - Bring social and economical benefits to people and communities
  - Some programs are funded by companies
- Adopted by other institutions
- Contribution for a fair society

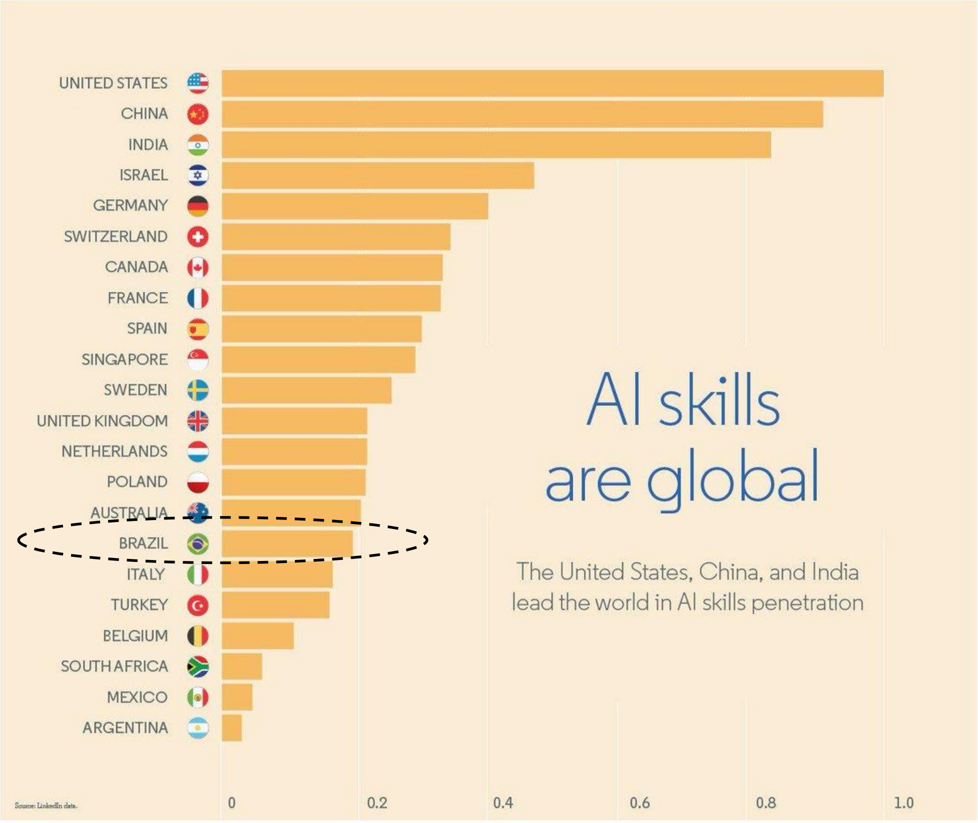
DATA SCIENCE FOR SOCIAL GOOD

рахмат
danke
謝謝
ngiyabonga
Баярлалаа
spas
teşekkür ederim
спасибо
mersi
kia ora
barka
welalin
tack
mahalo
tapadh leat
faafetai lava
vinaka
спасібі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
nanni
nandri
kiitos
dankie
dhanyavad
mauruuru
thank you
gracias
хвала
asante
manana
bedankt
bayarlalaa
gràcie
hvala
köszönöm
akun
dankon
aciū
obrigada
tenki
murakoze
enkosi
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
děkuji
sukriya
kop khun krap
obrigado
sagolun
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
ありがとう
tanemirt
grazie
najis tuke
kam sah hamnida
terima kasih
rahmet
rahmat
ευχαριστώ
diolch
dhanyavadagalu
shukriya
merce
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci

Questions?

Human resources

16º



AI skills are global

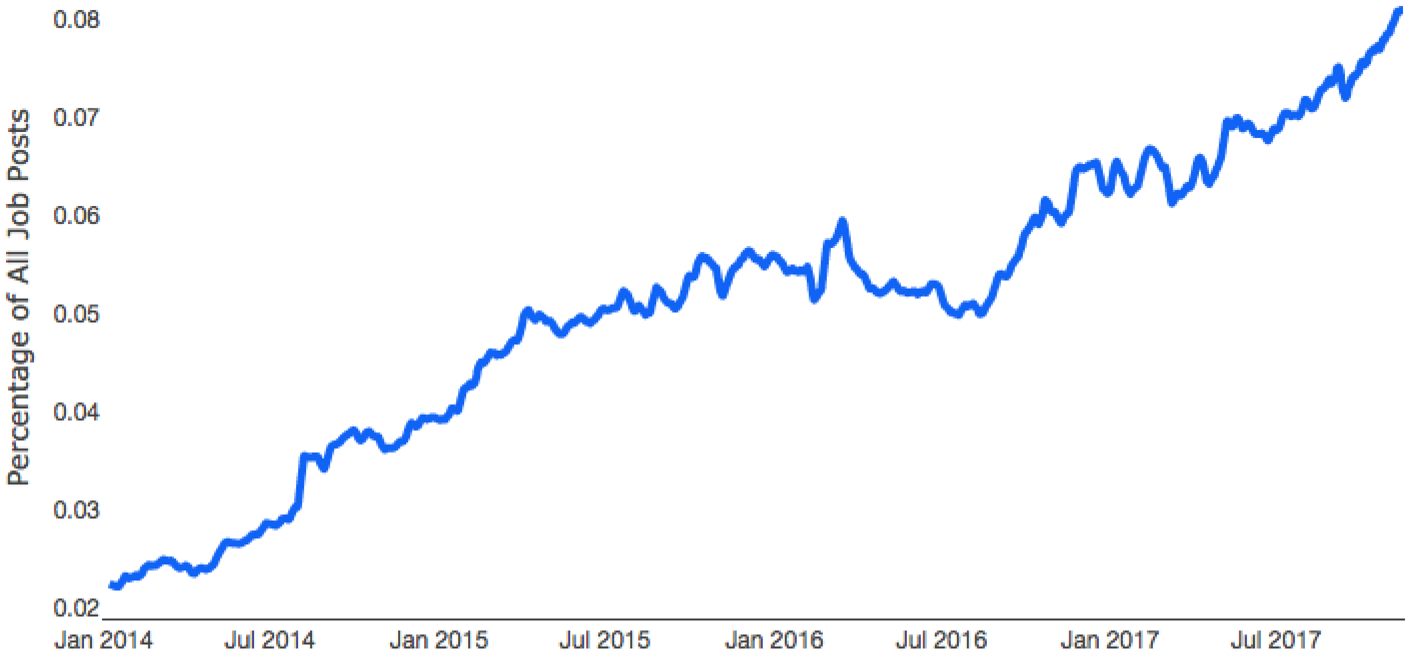
The United States, China, and India lead the world in AI skills penetration

Source: LinkedIn data.

Note: Country comparison is done by benchmarking every country against the United States. Specifically, we divide the AI skill penetration index of each country by that of the United States using the common occupations between the United States and each country. Therefore, the country index reflects how much AI skills penetrate a country's occupations relative to the United States over the past 3 years.

Jobs in Data Science

Data Science Jobs on Indeed Have Quadrupled over the last 4 years



https://medium.com/indeed-data-science/transitioning-from-academia-to-industry-perspectives-from-indeeds-data-scientists-de890acd1bf