The Alignment Problem: Values and Tech

Week 5

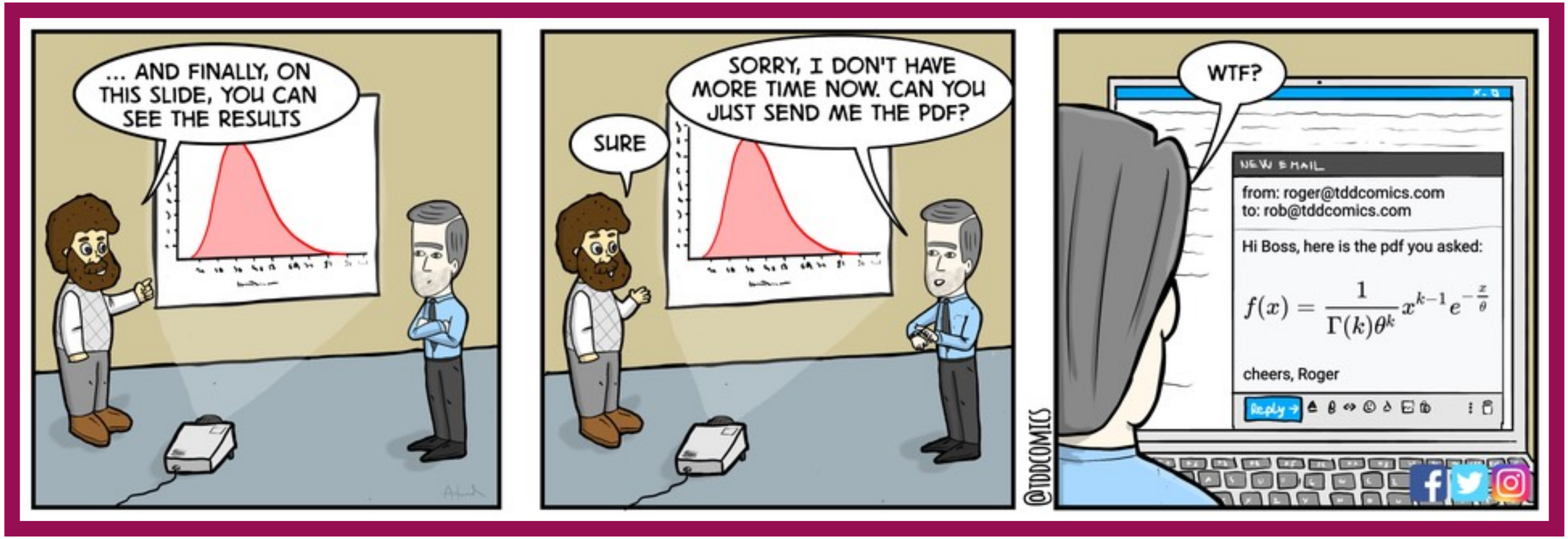
DS198-004: Data Discovery Scholars Seminar
UC Berkeley - Computation, Data Science, and Society

Fall 2021

Today

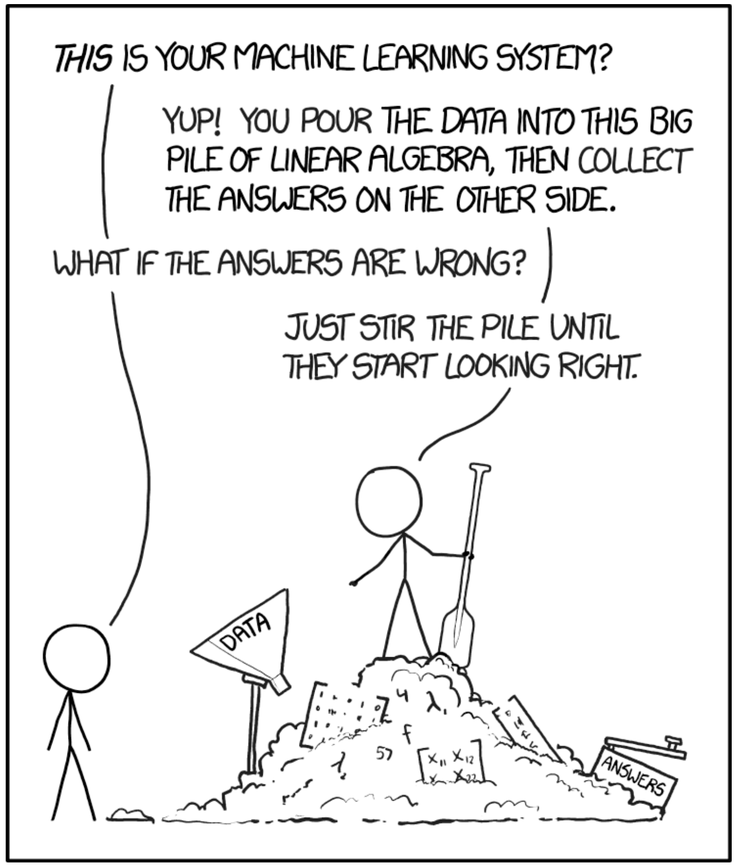
1. Meme of the Week
2. Announcements
3. Career Accelerator Program
4. Embedding Values in Technology
5. Privacy
6. Work Session

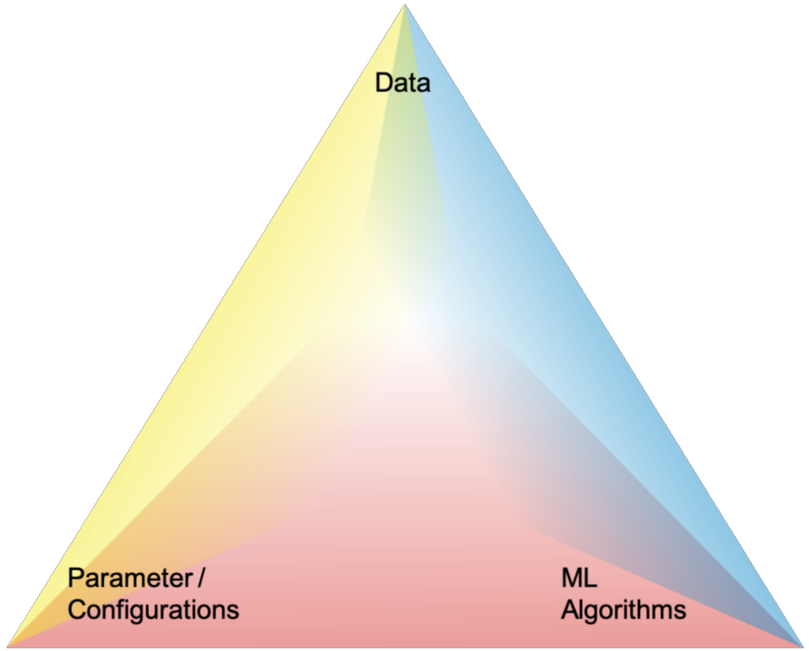
Meme of the Week


... AND FINALLY, ON THIS SLIDE, YOU CAN SEE THE RESULTS
SURE
SORRY, I DON'T HAVE MORE TIME NOW. CAN YOU JUST SEND ME THE PDF?
WTF?
NEW E MAIL
from: roger@tddcomics.com
to: rob@tddcomics.com
Hi Boss, here is the pdf you asked:
$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$
cheers, Roger
reply
@TDDCOMICS

The Alignment Problem

Repeatability, Reproducibility and Replicability





Adapted from https://suneeta-mall.github.io/slides/3Rs.html#/what-is-ds

Robustness, Reproducibility and Replicability

- Reproducibility: ability to recreate results by following the same steps using the same data
- Replicability: ability to recreate results by following slightly different processes and data
- Why do we care?
  - Imagine if we wouldn’t be able to reproduce findings/results.
    - Data = Knowledge

Adapted from https://www.kdnuggets.com/2019/11/reproducibility-replicability-data-science.html

Robustness, Reproducibility and Replicability

- Reproducibility: ability to recreate results by following the same steps using the same data
- Replicability: ability to recreate results by following slightly different processes and data
- Why do we care?
  - Imagine if we wouldn’t be able to reproduce findings/results.
    - Data = Knowledge
- Robustness: the ability to withhold adversarial attacks or different environments

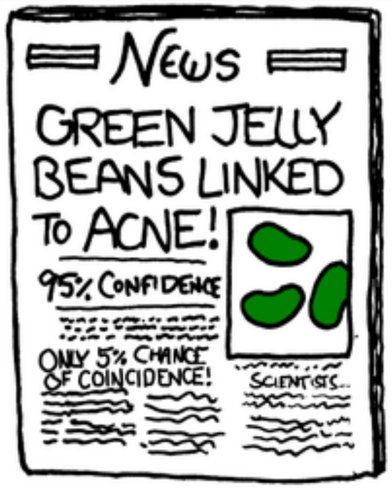
Adapted from https://www.kdnuggets.com/2019/11/reproducibility-replicability-data-science.html

Robustness, Reproducibility and Replicability

- Reproducibility
  - Ensure you have **documentation**, create **requirements.txt** files, update your **packages** and maybe use a **docker** container. Write **functions** and use **OOP**.
- Replicability
  - Harder to guarantee than reproducibility

Robustness, Reproducibility and Replicability

- Replicability
  - P-Hacking
    - Repeated process of running a random process until you reach statistical significant level
    - What is the issue here?
  - **Related Topics**
    - Random seed
    - Researcher's bias



Robustness, Reproducibility and Replicability

- **Robustness**

ANNALS OF TECHNOLOGY

THE PASTRY A.I. THAT LEARNED TO FIGHT CANCER

In Japan, a system designed to distinguish croissants from bear claws has turned out to be capable of a whole lot more.

By James Somers

March 18, 2021

Robustness, Reproducibility and Replicability

- Robustness
  - Machine Learning instances can break if you use them on things that they are not designed for.
    - Why? High-dimensionality and unexplainable black boxes. (Research ongoing)
    - Defining and understanding your problem and data
  - How do we build robustness in?
    - Adversarial Training
    - Explainability methods
    - Understanding data and distribution shift (More in tracking lecture)

Sustainability in Data Science

- Sustainability
  - Data Science for Environmental Sustainability
    - Image recognition for trash
    - Using data to estimate crop yields to fight famine
    - Image recognition for sorting waste
    - Climate change and forecasting

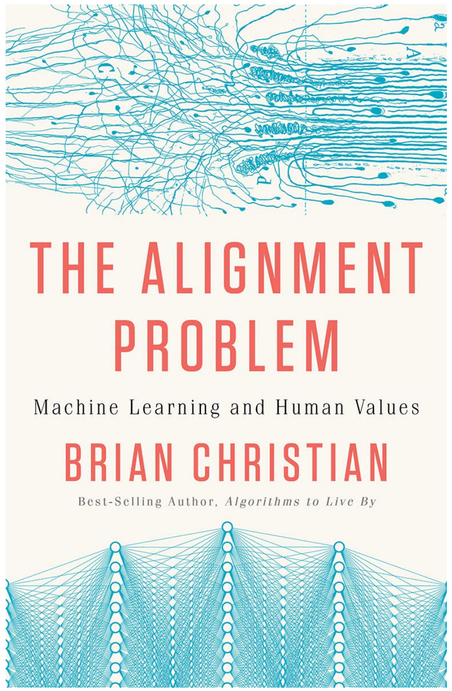
Sustainability in Data Science

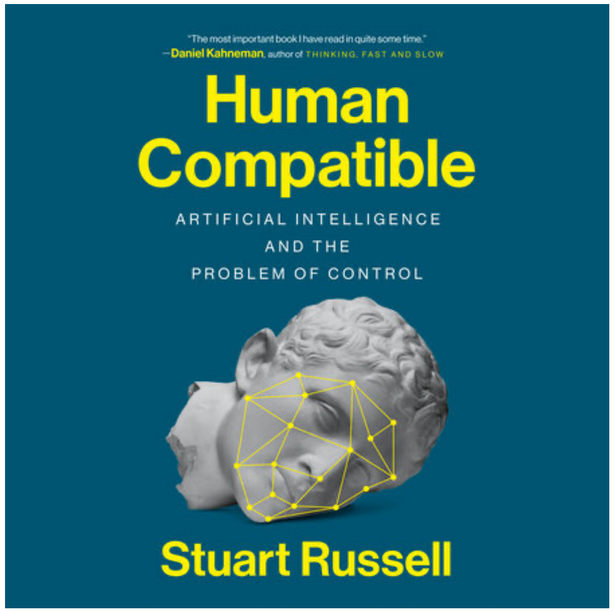
- Sustainability
  - Nice and well, but … let's take a look at the impact of technology
  - GPUs/Data Servers
    - Take a lot of electricity
    - Tradeoff b/w social impact and sustainability
    - Use green GPUs + off hours (spot instances)
  - Blockchain Mining
    - Bitcoin mining consumes 110 TeraWatt/Hour (~ Sweden)

Aligning values with tech is hard

Must Reads

- Must reads
  - The Alignment Problem
  - Human Compatible





Resources

- **Privacy**
  - [TECH] https://blog.openmined.org/private-machine-learning-explained/
  - [TECH] https://towardsdatascience.com/security-and-privacy-in-artificial-intelligence-and-machine-learning-part-1-c6f607feb94b
  - [LAW] https://iapp.org/news/a/embedding-data-ethics-into-your-culture-of-privacy/

Resources

- **Sustainability**
  - https://towardsdatascience.com/data-science-for-sustainability-b912d5fb5d24
  - https://www.discoverdatascience.org/resources/data-science-and-sustainability/
  - https://towardsdatascience.com/deep-learning-and-carbon-emissions-79723d5bc86e

Resources

- **3 Rs**
  - https://online-learning.harvard.edu/course/principles-statistical-and-computational-tools-reproducible-data-science?delta=1
  - https://www.datarobot.com/trusted-ai-101/performance/robustness-and-stability/
  - https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

Privacy

Go here!

https://tinyurl.com/ds1983privacy