DATA, DATA EVERYWHERE

(And not a drop to drink)

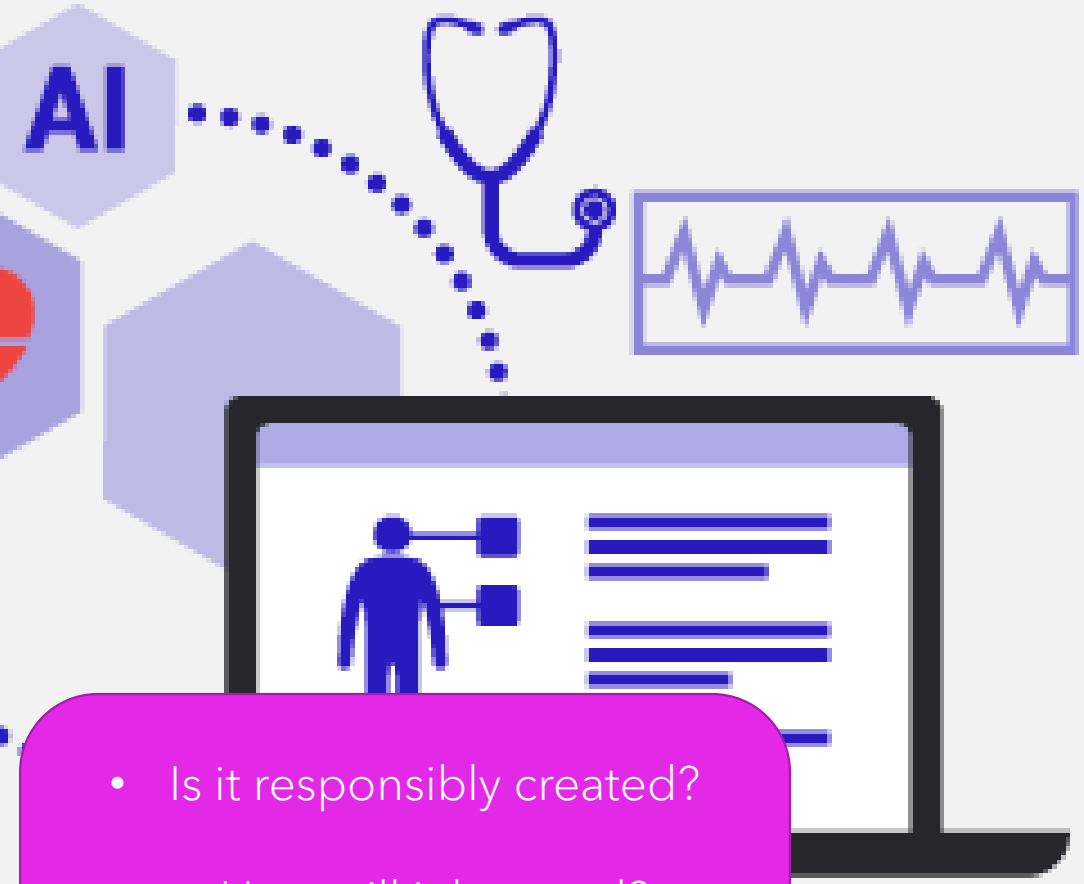
Professor Adriane Chapman
Adriane.Chapman@soton.ac.uk

SO YOU'VE BUILT A COOL WIDGET, AND IT PROVIDES YOU WITH LOTS OF DATA

WHAT SHOULD YOU DO WITH THAT DATA?

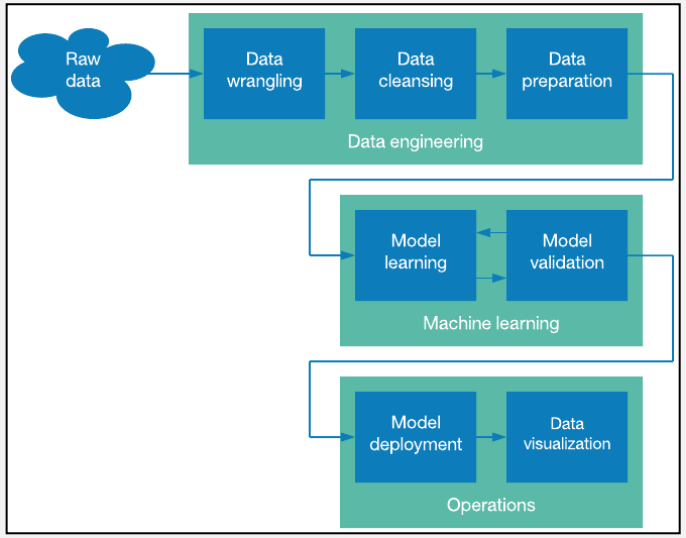
- Already overwhelmed staff being given even more information to sift through
- "What does that even mean?"

- Guidance to not consider patient-generated / app data in care
- “What does that even mean?”



Data Science in Healthcare

- Is it responsibly created?
- How will it be used?
- "What does that even mean?"

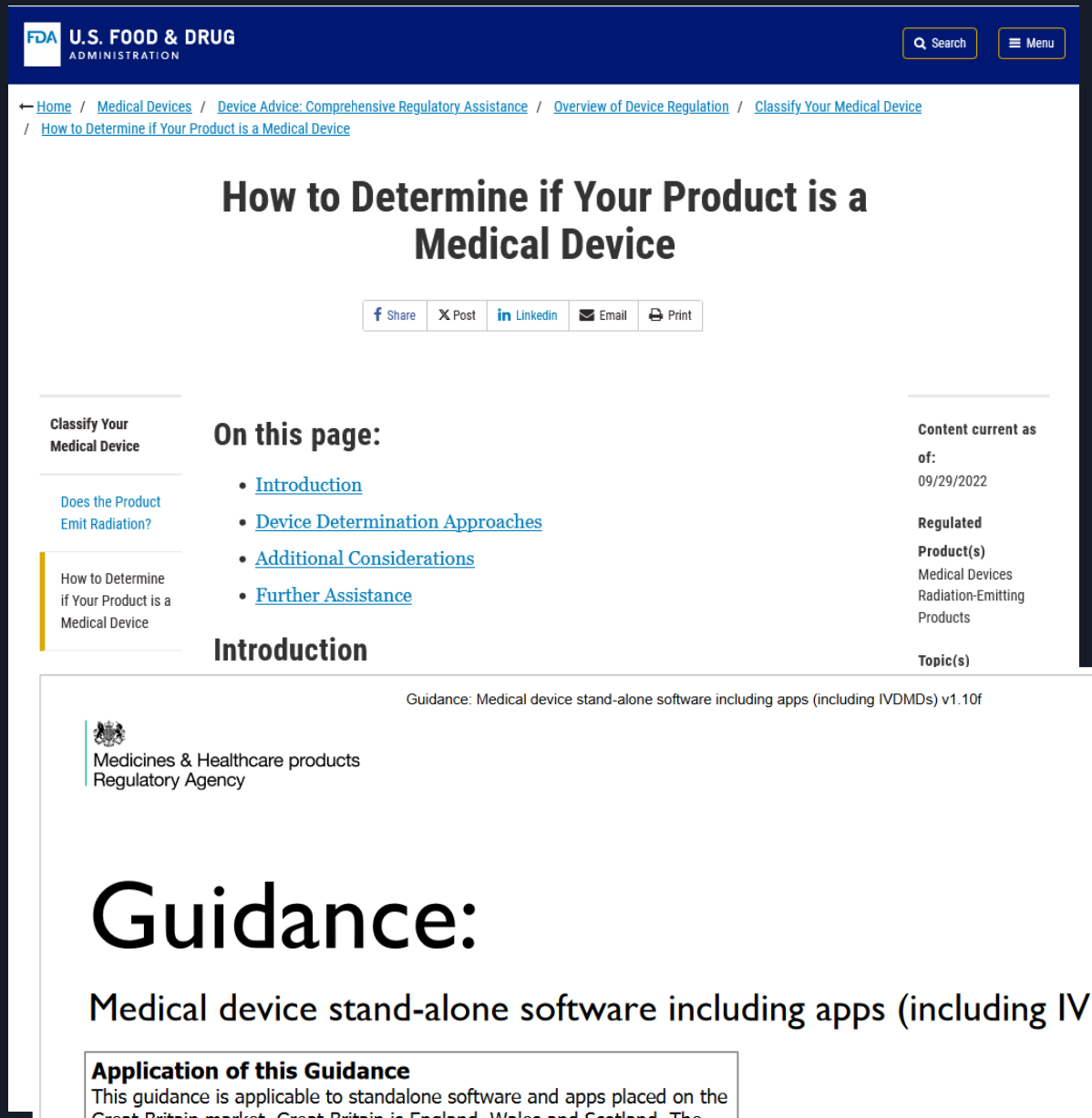


HUH. SO WHAT SHOULD YOU DO WITH IT?

"GIVE IT TO A CLINICIAN TO IMPROVE CARE"

- Don't assume that a clinician wants your data
- Don't assume that a clinician can understand your data
- What do they need?
- An App in the clinical context is a medical device!

  Develop from the start with regulatory requirements in mind!



"BUILD AN APP FOR PATIENTS AND CLINICIANS"

- Ask: For whom, and why would they want it?
- Why is knowing X interesting or helpful?
- What else would someone need to know
- Vast research in Human Computer Interaction, Usability, etc.

GOV.UK Menu

Home > Health and social care > National Health Service > Digital and data-driven health and care technology

Department of Health & Social Care

Guidance

A guide to good practice for digital and data-driven health technologies

Updated 19 January 2021

Contents

Introduction

1. How to operate ethically
2. Have a clear value proposition
3. Usability and accessibility
4. Technical assurance
5. Clinical safety
6. Data protection
7. Data transparency
8. Cybersecurity
9. Regulation
10. Interoperability and open

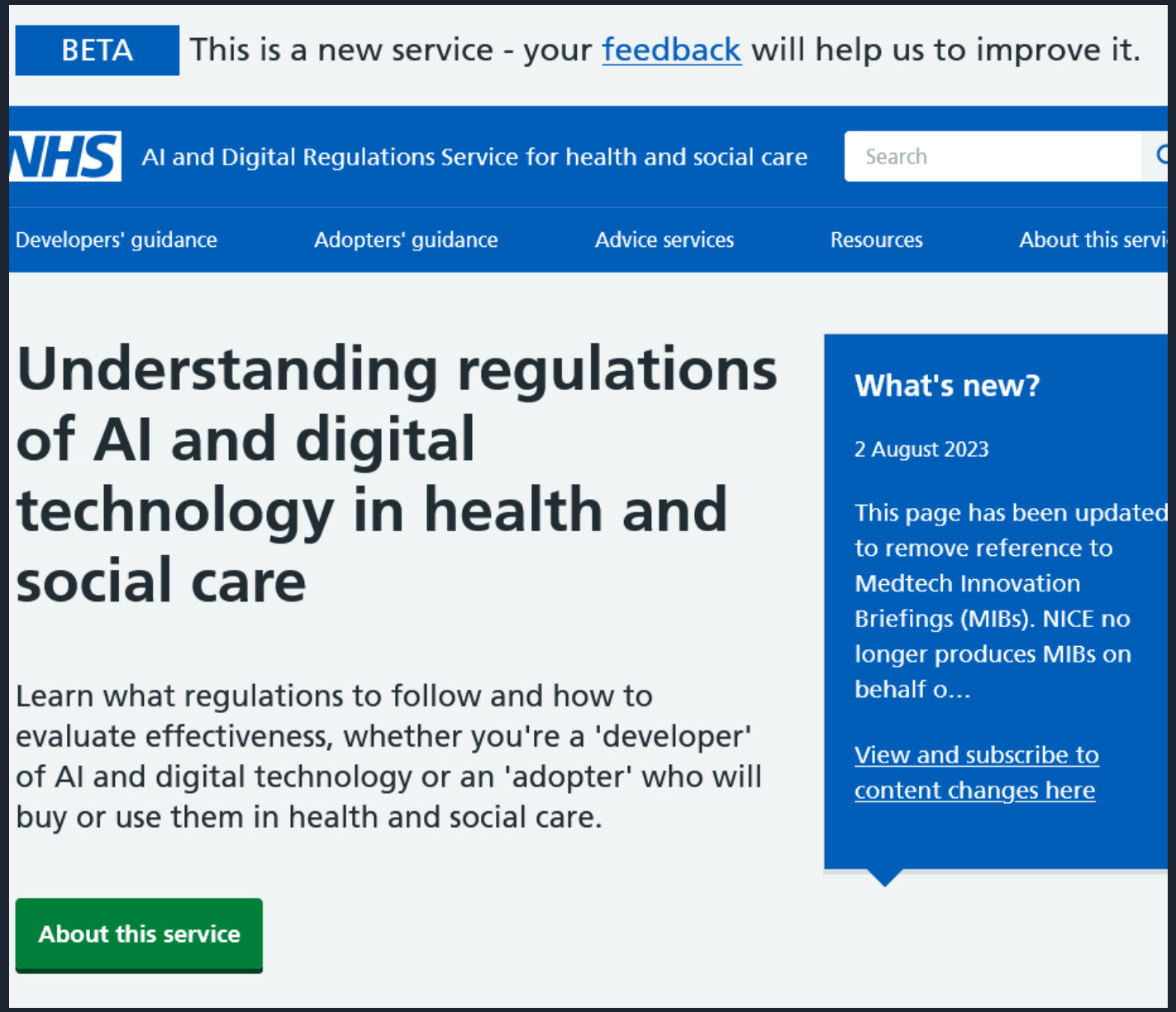
Introduction

Across the country and around the globe, digital innovators are helping us deliver our commitment to the digital transformation of health and social care, to bring benefits to patients, the workforce and the system as a whole. NHS England's Long Term Plan sets the direction towards widespread digitally-enabled care. The Secretary of State's Technology Vision goes on to articulate a clear ambition for the generation of more digital services designed around user need and adhering to key principles of privacy, security, interoperability and inclusion.

It is our duty as NHS England and central government to capitalise on these opportunities responsibly. The healthcare system is a unique space where a variety of regulatory ecosystems overlap. Due to the privileged nature of dealing with people's health and their protected data, the system is covered by various pieces of legislation

AI! DATA SCIENCE!

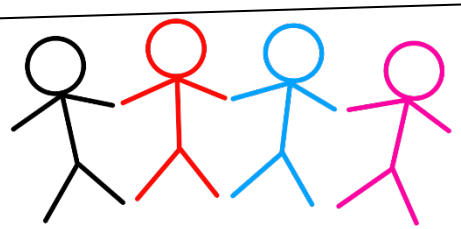
- Yes! … but let's do it responsibly
- Fast changing world, e.g. Beta service!



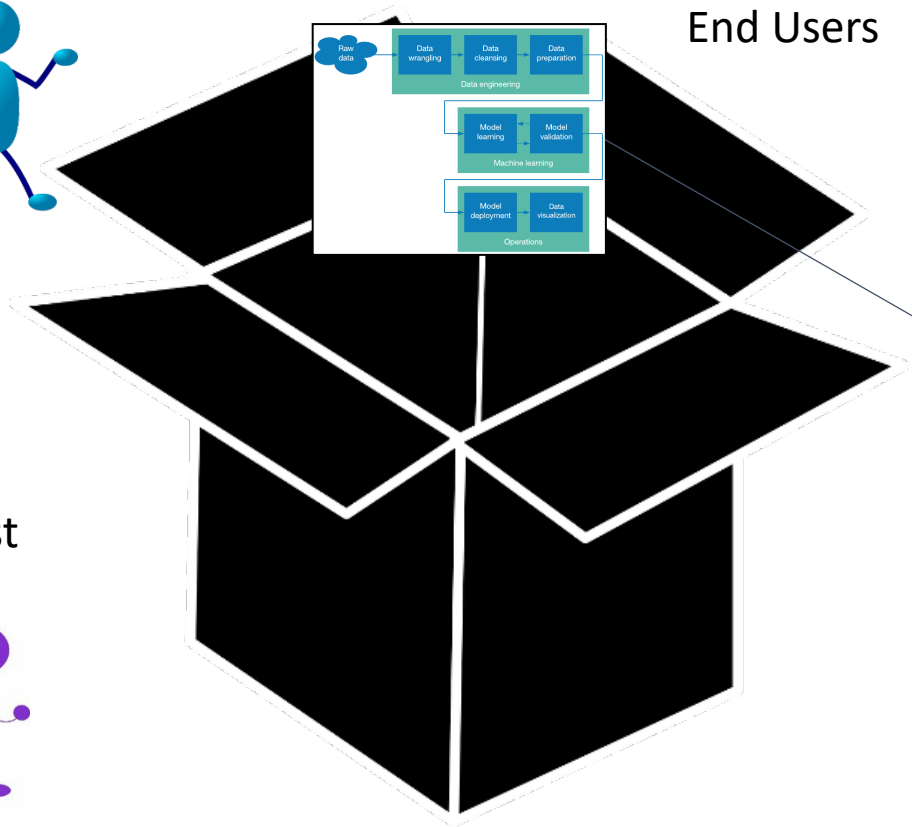
GREAT! THANKS FOR THE POINTERS.
(WHY IS SHE STILL STANDING THERE?)

I ♥ DATA

Owner
Requisition-er
Project Manager



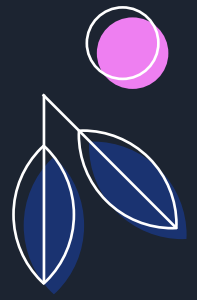
Affected Individuals
End Users



Responsible
Development,
Fairness,
Accountability,
Transparency,
Explainability

Data Scientist
Developer
Engineer

How to get data; what's going on in the data; how to fix the data; new models; etc.

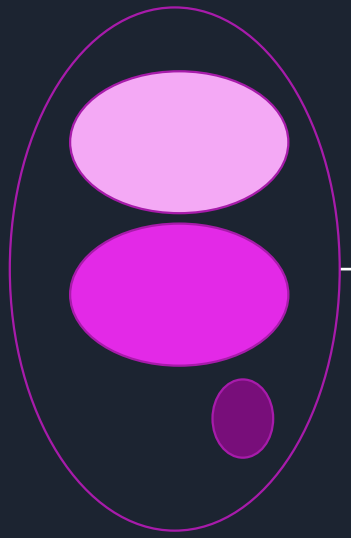


DATA- MODEL CONSEQUENCES

Fairness

ALGORITHM FAIRNESS

Data



Model

Output

Biases exist in all data

e.g. sampling bias, institutional bias, etc.

Models are an *abstraction*.

"Learn" to fit the data to a function.

Can exacerbate biases

Fairness is a measure of whether a model treats groups differently.

WHAT IS FAIRNESS IN HEALTH?

Fairness Measure	Description	PPI statements: Patient experience
Individual Fairness	Predictions for any pair of similar individuals are the same.	I want my doctor to treat me the same as all other patients, as if we were all sitting behind a screen.
Predictive Parity	Groups have equal probability of an individual with positive predictive value to belong to the positive class (TP/(TP+FP)).	I want to receive the same care as everyone else in a system that does not discriminate.
Predictive Equality	Groups have equal probability of an individual in the negative class to have a positive predictive value, a false alarm, (FP/(FP+TN)).	
Equal Opportunity	Groups have equal probability of an individual in a positive class to have a negative predictive value (FN/(TP+FN)).	
Treatment Equality	Groups have an equal ratio of false negatives and false positives (FN/FP).	

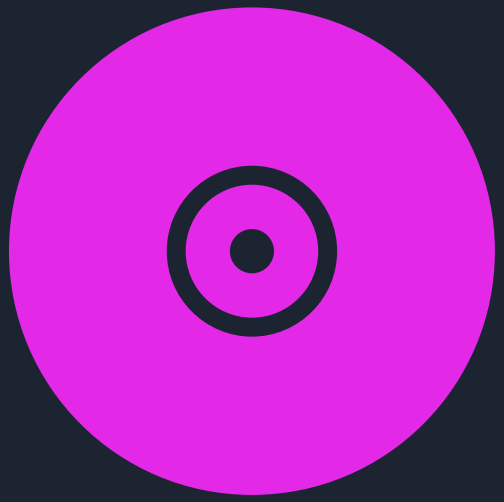
ALL DATA LIES

- Data
  Selection bias
  Institutional bias
  Societal bias

- The types of sensors and diagnostics we build change how it lies

Jung, An, Kwak, Salminen, Jansen. (2018) "Assessing the Accuracy of Four Popular Face Recognition Tools for Inferring Gender, Age, and Race." AAAI.

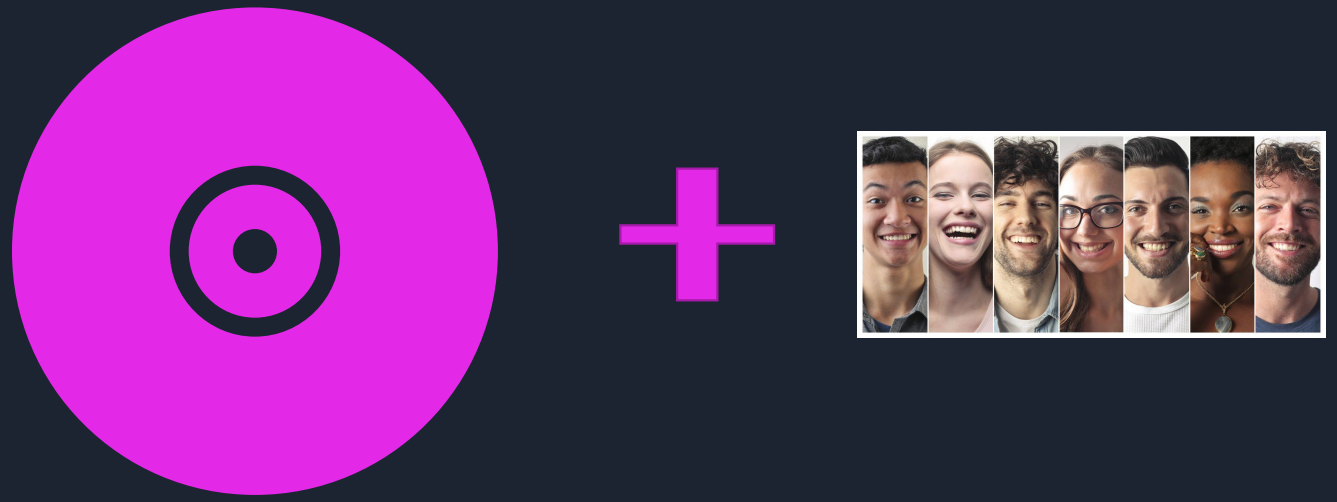
ALL MODELS ARE AN ABSTRACTION (THEREFORE, NONE ARE CORRECT)



- The frame of reference

  Definition: Given data D = {x_1, x_2, … x_n}, target labels L = {y_1, y_2, …, y_n} find a hypothesis s.t.
- Data: transformation of raw data into feature vectors determines unit of analysis and quantification. Data exists **in context.**
- Labels: Discretization matters. Number and boundaries affect results. Who defines labels? Who creates labels?
- Loss: Interact in non-obvious, domain specific ways. Penalize errors differently. Often simplified and backed into loss function

DATA – MODEL CONSEQUENCES



= Possibly Bad

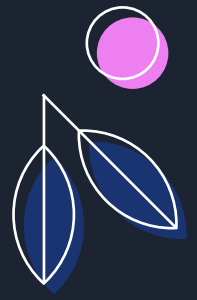
DATA – MODEL - HUMAN CONSEQUENCES

Complex systems and humans

= Not Good

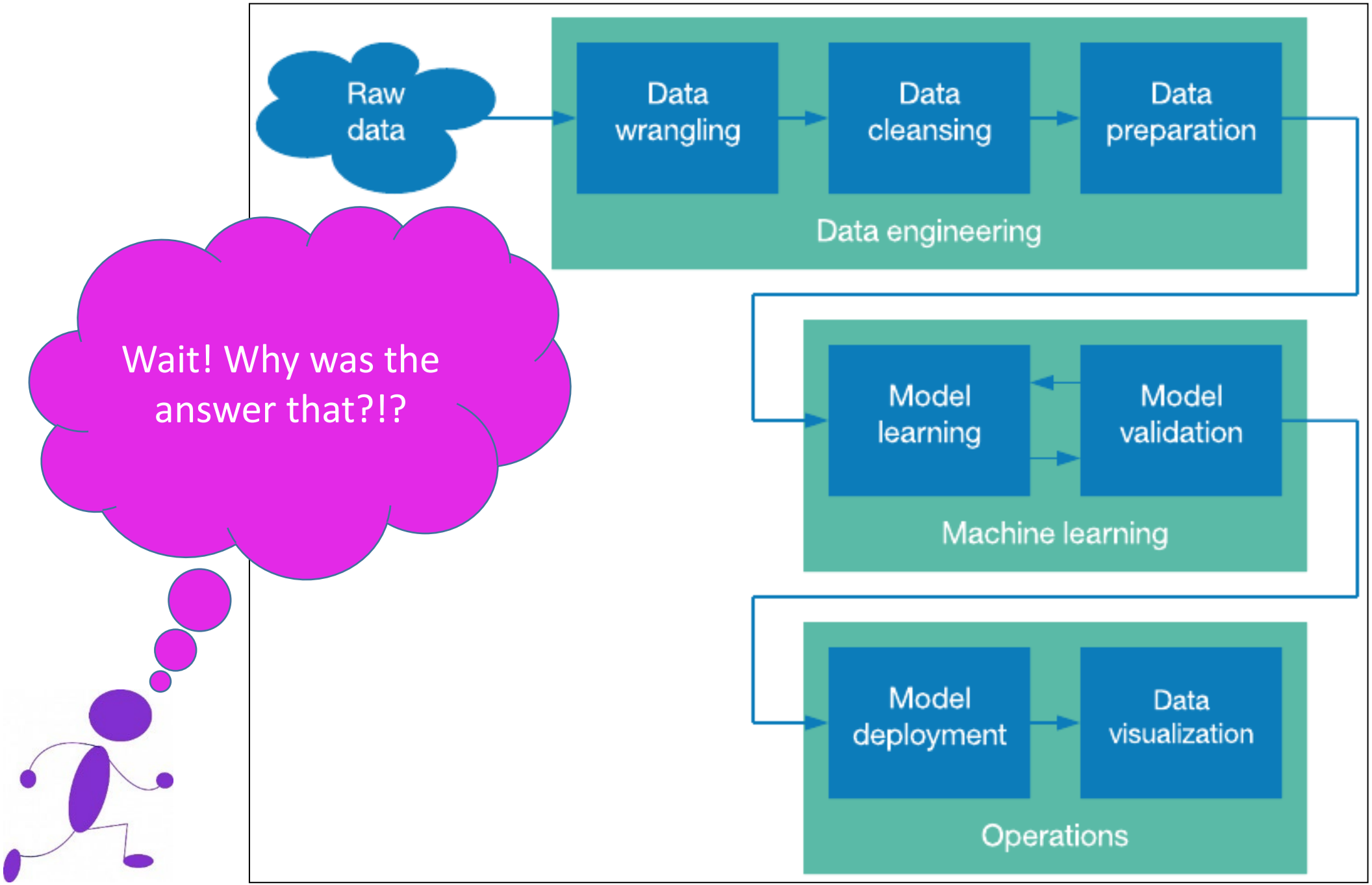
Chapman, Grylls, Ugwudike, Gammack, and Ayling. 2022. A Data-driven analysis of the interplay between Criminological theory and predictive policing algorithms. 2022 ACM FAccT. https://doi.org/10.1145/3531146.3533071

HOW DOES YOUR WIDGET CREATE DATA?
WHAT ARE THE BIASES IT WILL CONTAIN
AND HOW COULD IT SPIRAL?



PROVENANCE

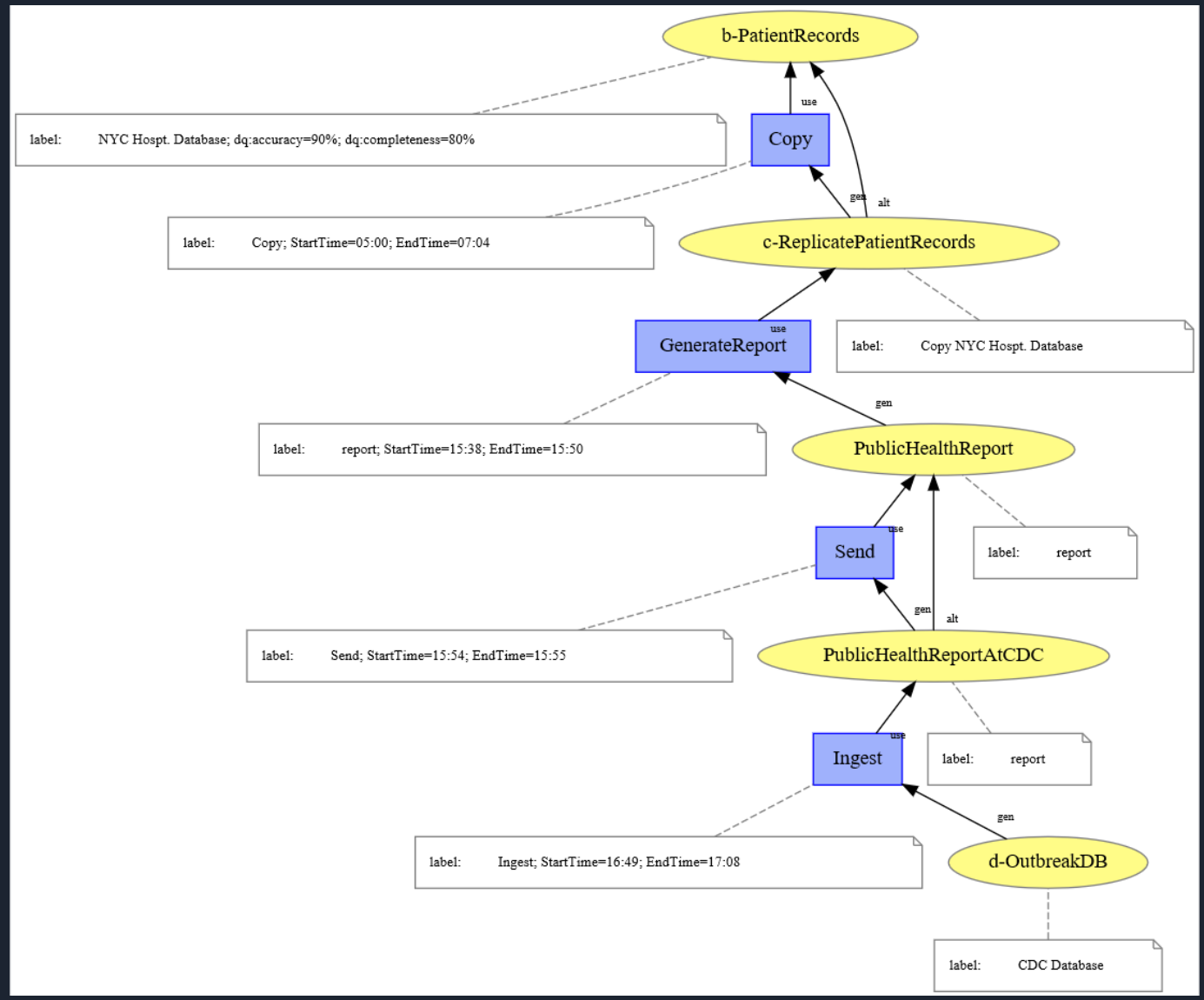
Transparency and Explanations


Raw data
Data wrangling
Data cleansing
Data preparation
Data engineering
Model learning
Model validation
Machine learning
Model deployment
Data visualization
Operations
Wait! Why was the answer that?!?

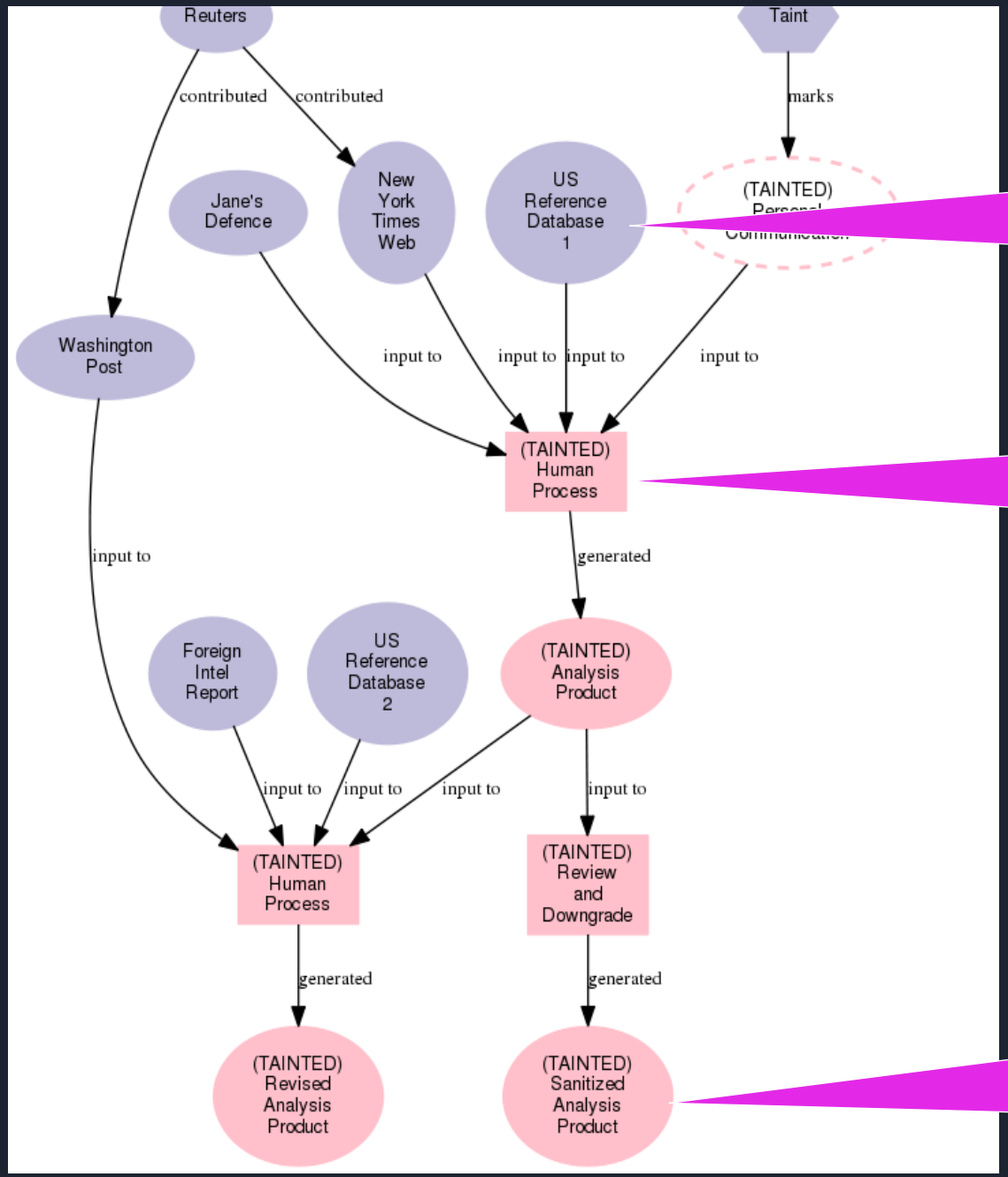
PROVENANCE*: A "FAMILY TREE"

- A record of what actually happened
- Agents, Entities, Activities
- Important attributes: timestamps

- Provenance is a type of metadata



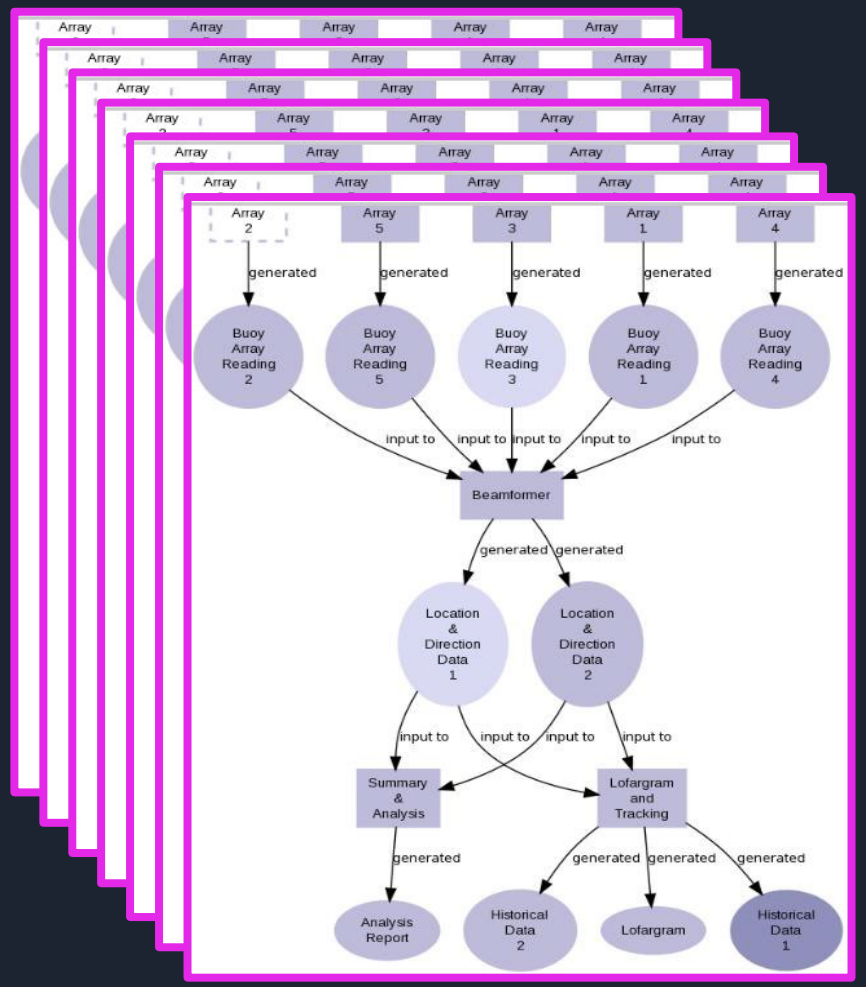
* aka Pedigree, Lineage



Organization: Who is using my data?

Developer: This data is tainted!

User: Do I “trust” this data?



With lots and lots of provenance, have the ability to see what resources are truly being used

Manager: Analyst Joe's prior confidences are always correct

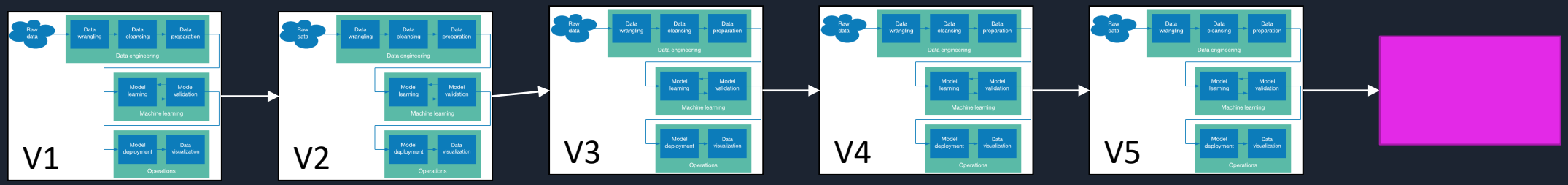
Manager: How often did someone use this expensive subscription source?

Manager: What parts of my team need retraining? (i.e. less productive)

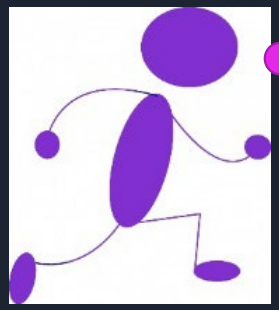
Collector: Is the data I have collected fit for my use?

Analyst: If you are interested in X, others were interested in Y and Z.

TRANSPARENCY



Wait! Why was the answer that?!?



Pina, Chapman, de Oliveira, Mattosa. Deep Learning Provenance Data Integration: a Practical Approach. WWW Companion 2023

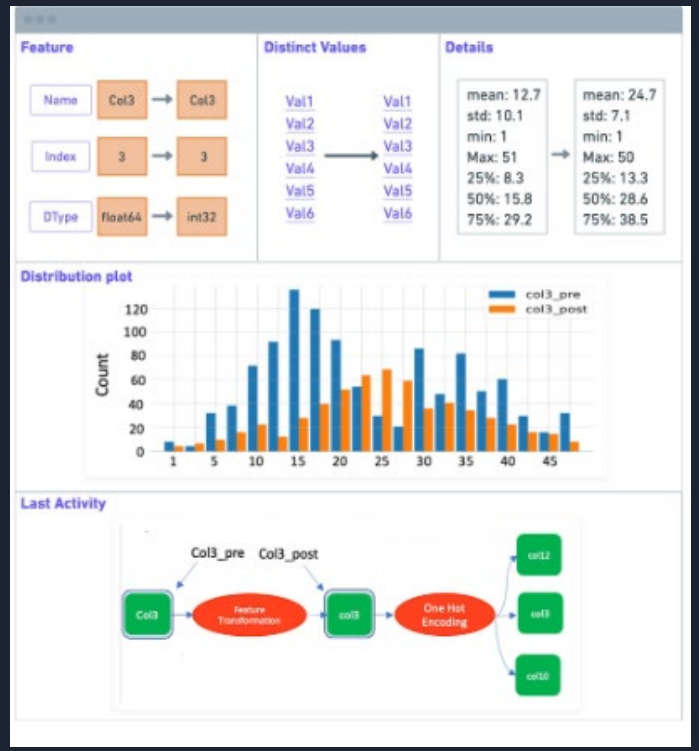
Chapman, Missier, Simonelli, and Torlone. 2020. Capturing and querying fine-grained provenance of preprocessing pipelines in data science. Proc. VLDB.

EXPLANATIONS

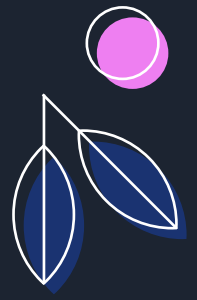
- Currently target mainly developers
- E.g. LIME, InterpretML, ELI5, SHAP

- What needs to go in an explanation for a clinician (or patient) to understand or trust a decision?

- Context?
- Additional information?
- Provenance



Chapman, Lauro, Missier, and Torlone. 2022. DPDS: assisting data science with data provenance. Proc. VLDB.



DATA EXCHANGE REASONING

Accountability

HEALTH AND WELLNESS DATA NEEDS TO BE HANDLED CAREFULLY

Latest: Strava suggests military users 'opt out' of heatmap as row deepens

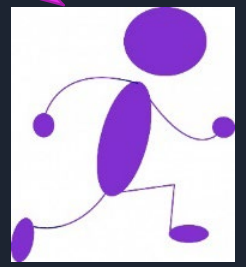
What Happens When You Track Your Boyfriend on Strava

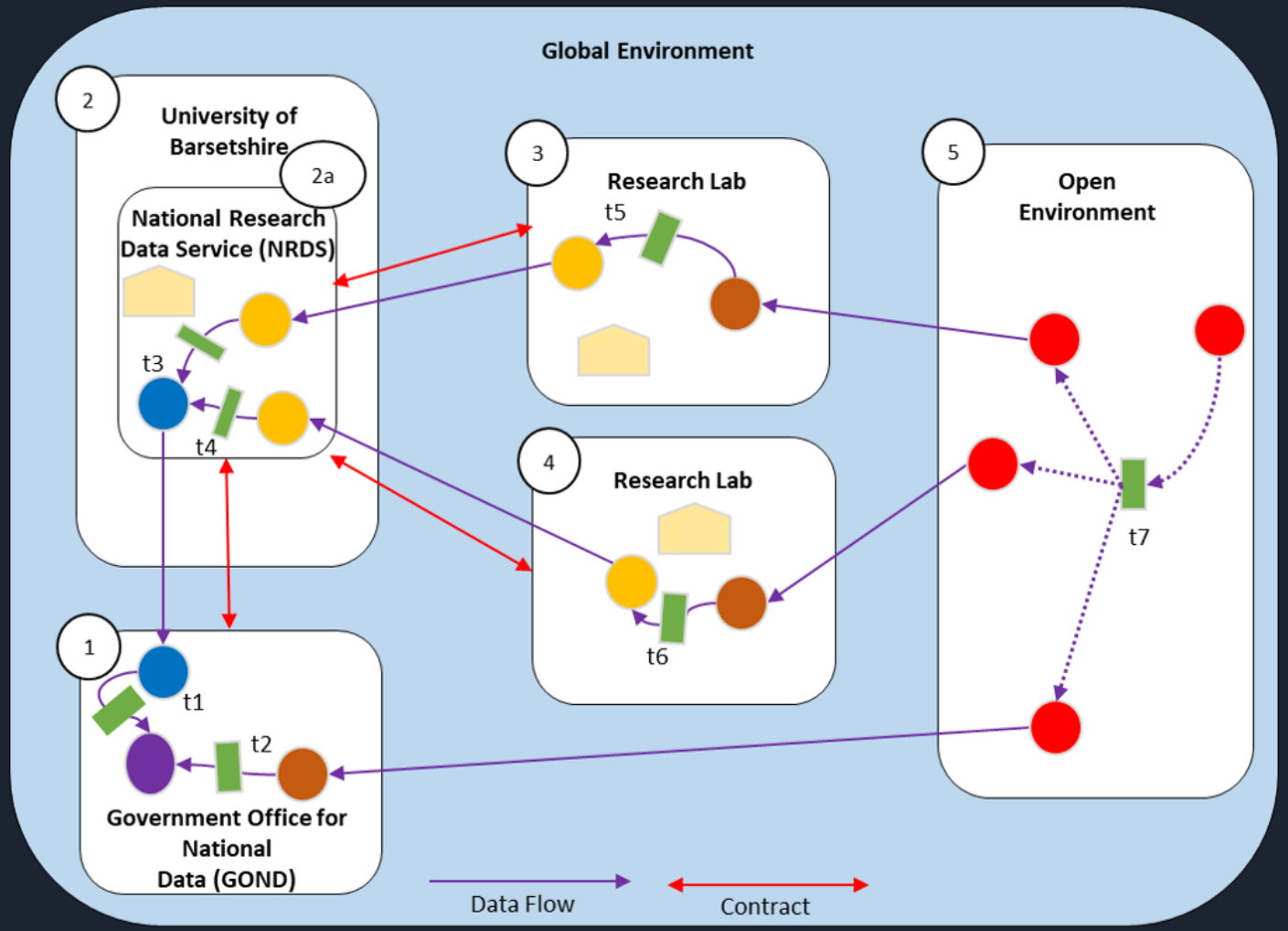
This is one woman's tale of romance, performance, surveillance, and loss.

WIRED What Happens When You Track Your Boyfriend on Strava

SIGN IN | SUBSCRIBE

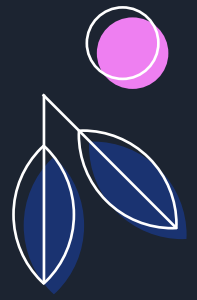
I need that data. How do I protect it so it cannot be de-anonymized based on other data out there?





Jarwar, Chapman, Elliot, Blount, and Raji, Modelling Data Environments within Prov to Assist Decision Making for Anonymisation. Available at SSRN

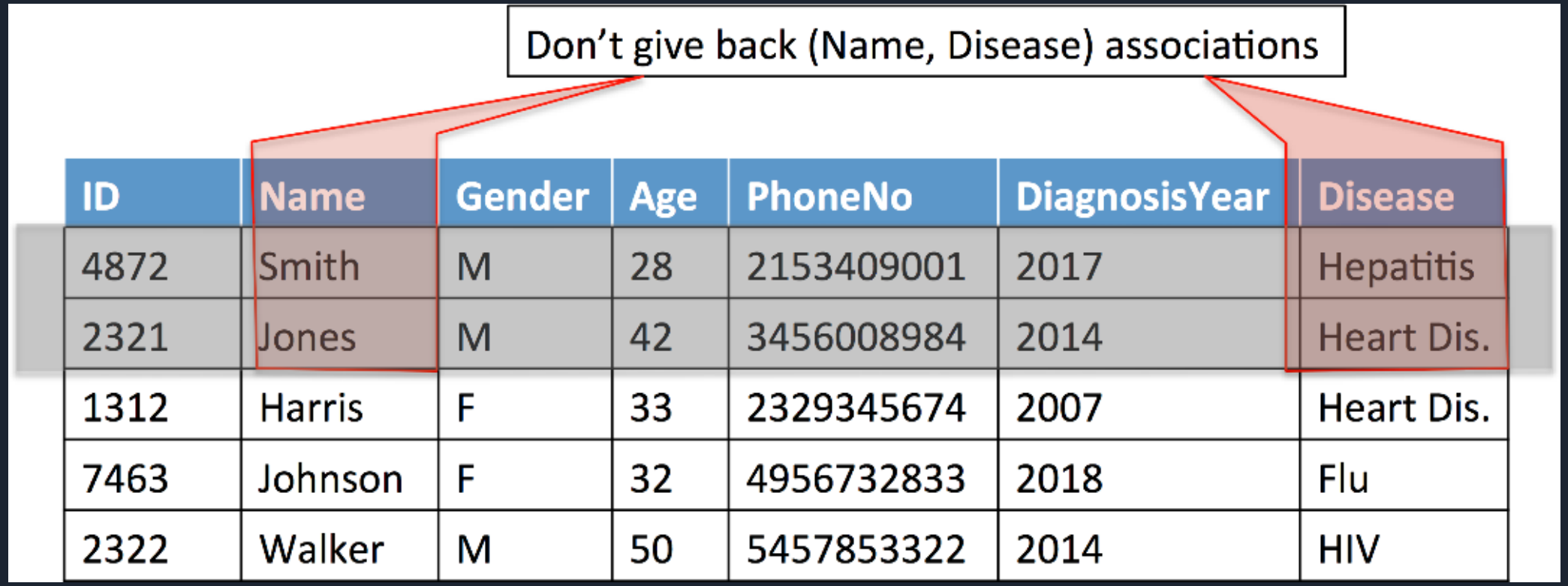
Jarwar, Chapman, Elliot, & Raji. (2021, Jul 21). Provenance, Anonymisation and Data Environments: A Unifying Construction.



MY DATA; MY WAY

Responsible Development

Data shall be shared according to this process, for only these reasons

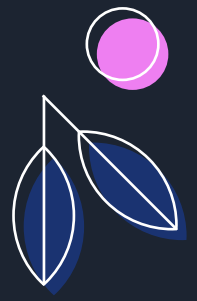


ID	Name	Gender	Age	PhoneNo	DiagnosisYear	Disease
4872	Smith	M	28	2153409001	2017	Hepatitis
2321	Jones	M	42	3456008984	2014	Heart Dis.
1312	Harris	F	33	2329345674	2007	Heart Dis.
7463	Johnson	F	32	4956732833	2018	Flu
2322	Walker	M	50	5457853322	2014	HIV

Konstantinidis et al. Personal Consent in Databases. *In revision* Sigmod 2021.; Konstantinidis et al. The Need for Machine-Processable Agreements in Health Data Management. J of Algorithmics 2020.

CONCLUSIONS

- Understand what is appropriate for you to do with your data
- Who really needs it and what do they need to understand it?
- Be responsible in your development, it will save you time and money in the future
- Be aware of the major concerns and research in this area



QUESTIONS?

p.s. Those of us who build algorithms, play with data, study HCI, etc. – we need health data. Find a buddy!