



# 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?

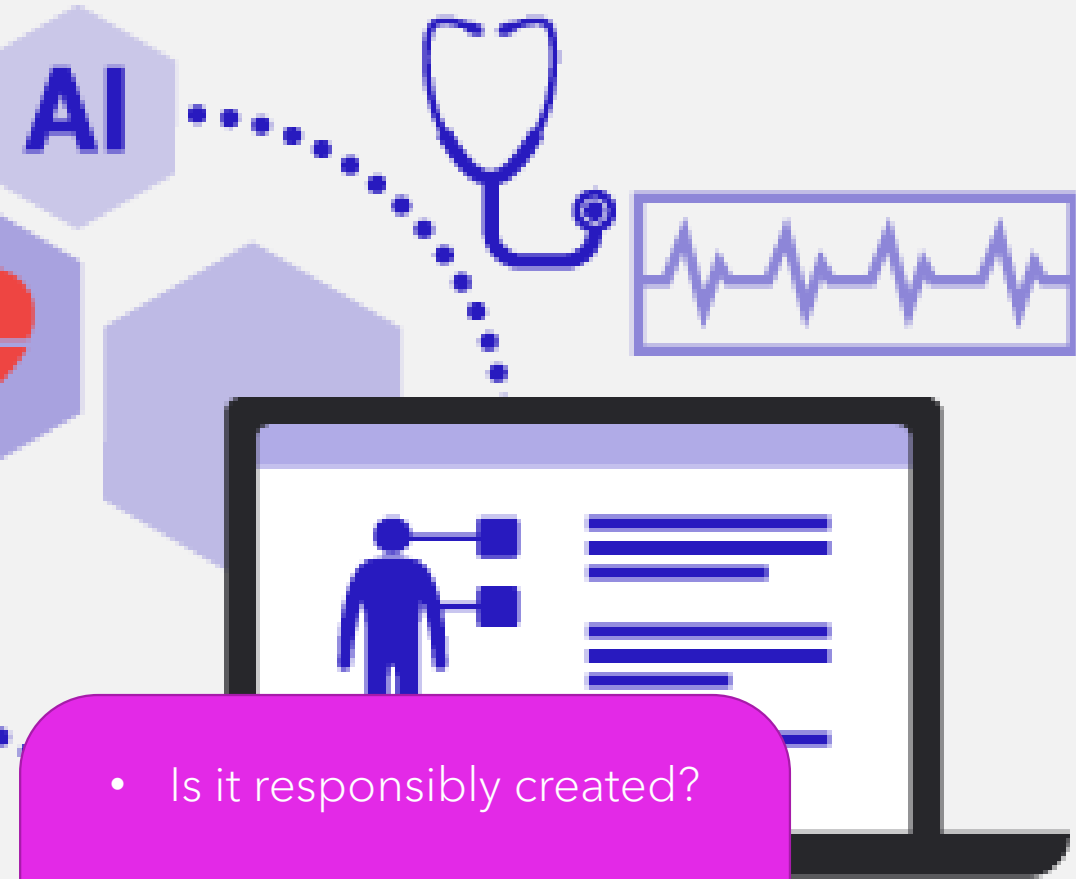


- Already overwhelmed staff being given even more information to sift through
- "What does that even mean?"



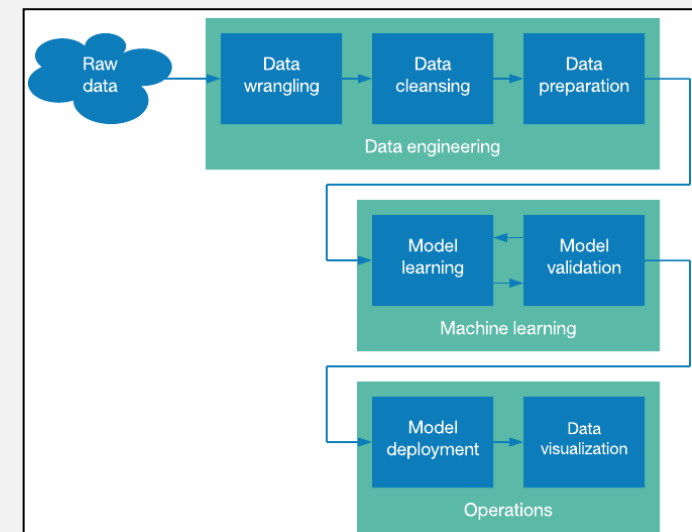
- Guidance to not consider patient-generated / app data in care
- “What does that even mean?”





# 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!*

The screenshot shows the FDA website page for "How to Determine if Your Product is a Medical Device". The page includes a navigation bar with the FDA logo and search/menu options. The main heading is "How to Determine if Your Product is a Medical Device". Below the heading are social sharing buttons for Facebook, X, LinkedIn, Email, and Print. A sidebar on the left contains a "Classify Your Medical Device" section with links for "Does the Product Emit Radiation?" and "How to Determine if Your Product is a Medical Device". A central section titled "On this page:" lists links for "Introduction", "Device Determination Approaches", "Additional Considerations", and "Further Assistance". A right sidebar shows "Content current as of: 09/29/2022", "Regulated Product(s): Medical Devices, Radiation-Emitting Products", and "Topic(s)". The main content area features a sub-heading "Introduction" and a box containing the text: "Guidance: Medical device stand-alone software including apps (including IVDMDs) v1.10f" and "Medicines & Healthcare products Regulatory Agency". Below this is a large heading "Guidance:" followed by the sub-heading "Medical device stand-alone software including apps (including IVDMDs)". A final box titled "Application of this Guidance" states: "This guidance is applicable to standalone software and apps placed on the Great Britain market. Great Britain is England, Wales and Scotland. The UKCA (UK Conformity Assessed) mark is used for certain goods, including medical devices, being placed on the Great Britain market. This guidance gives examples of software and apps which meet the definition of a..."



# “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.

The screenshot shows the GOV.UK website interface. At the top, there is a navigation bar with the GOV.UK logo, a 'Menu' dropdown, and a search icon. Below the navigation bar, a breadcrumb trail reads: Home > Health and social care > National Health Service > Digital and data-driven health and care technology. The main content area features a blue header with the text 'Guidance' and the title 'A guide to good practice for digital and data-driven health technologies', updated on 19 January 2021. On the left side, there is a 'Contents' section with a list of 10 topics: 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, and 10. Interoperability and open. The main text area is titled 'Introduction' and contains two paragraphs of text discussing digital innovation in healthcare and the NHS's commitment to digital transformation.

GOV.UK

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!

The screenshot shows the NHS AI and Digital Regulations Service website. At the top, a blue banner contains the text "BETA" in white, followed by "This is a new service - your [feedback](#) will help us to improve it." Below this is the NHS logo and the text "AI and Digital Regulations Service for health and social care". A search bar is visible on the right. A navigation menu includes "Developers' guidance", "Adopters' guidance", "Advice services", "Resources", and "About this service". The main content area features a large heading "Understanding regulations of AI and digital technology in health and social care" and a sub-heading "Learn what regulations to follow and how to evaluate effectiveness, whether you're a 'developer' of AI and digital technology or an 'adopter' who will buy or use them in health and social care." A green button labeled "About this service" is at the bottom left. On the right, a blue callout box titled "What's new?" dated "2 August 2023" contains the text: "This page has been updated to remove reference to Medtech Innovation Briefings (MIBs). NICE no longer produces MIBs on behalf o..." and a link "View and subscribe to content changes here".

**BETA** This is a new service - your [feedback](#) will help us to improve it.

**NHS** AI and Digital Regulations Service for health and social care

Search

Developers' guidance Adopters' guidance Advice services Resources About this service

## Understanding regulations of AI and digital technology in health and social care

Learn what regulations to follow and how to evaluate effectiveness, whether you're a 'developer' of AI and digital technology or an 'adopter' who will buy or use them in health and social care.

[About this service](#)

**What's new?**

2 August 2023

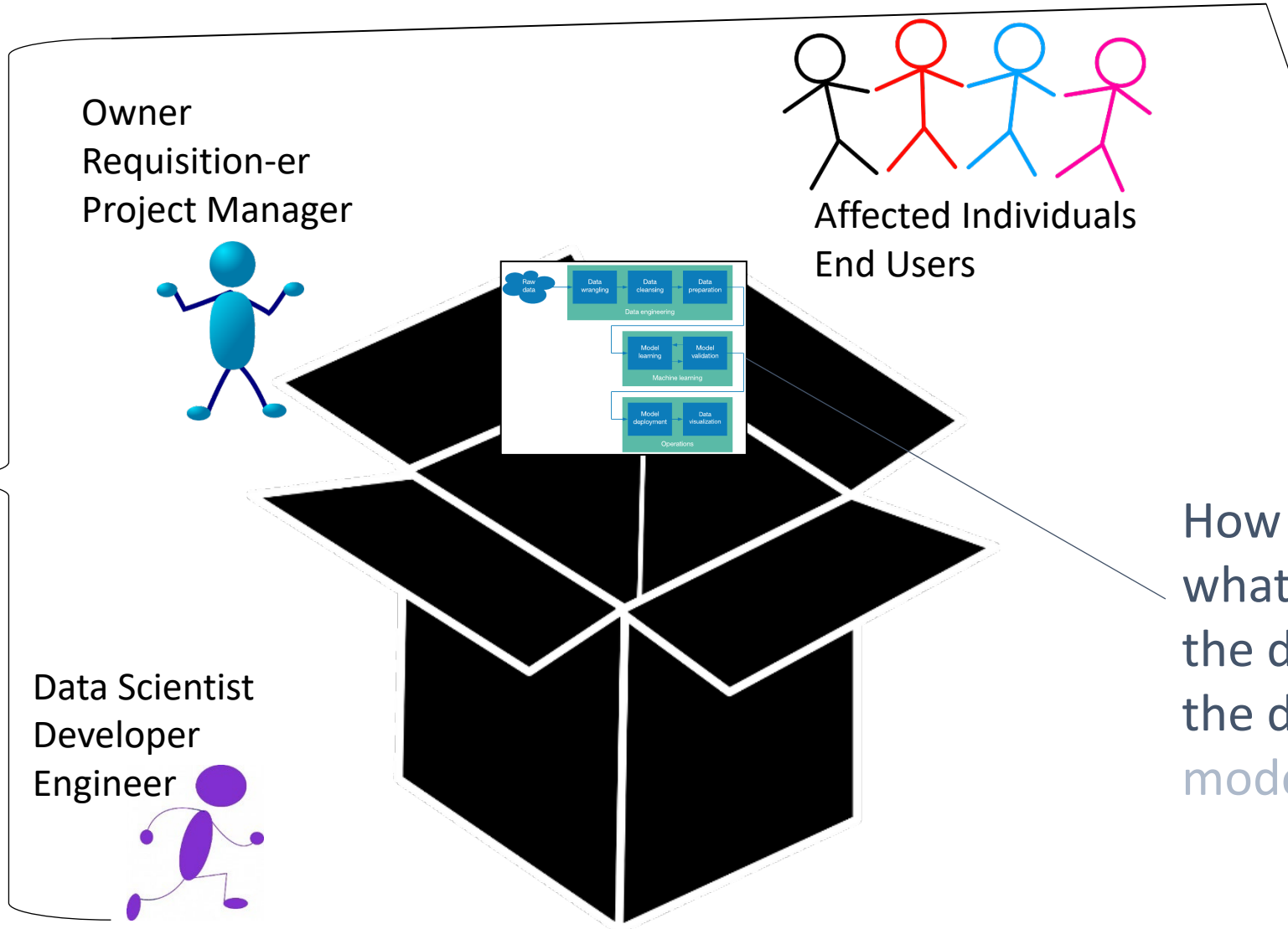
This page has been updated to remove reference to Medtech Innovation Briefings (MIBs). NICE no longer produces MIBs on behalf o...

[View and subscribe to content changes here](#)

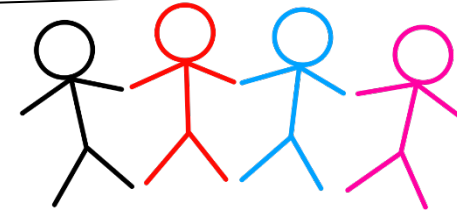
GREAT! THANKS FOR THE POINTERS.  
(WHY IS SHE STILL STANDING THERE?)

# I ❤️ DATA

Responsible  
Development,  
Fairness,  
Accountability,  
Transparency,  
Explainability



Owner  
Requisition-er  
Project Manager



Affected Individuals  
End Users

Data Scientist  
Developer  
Engineer



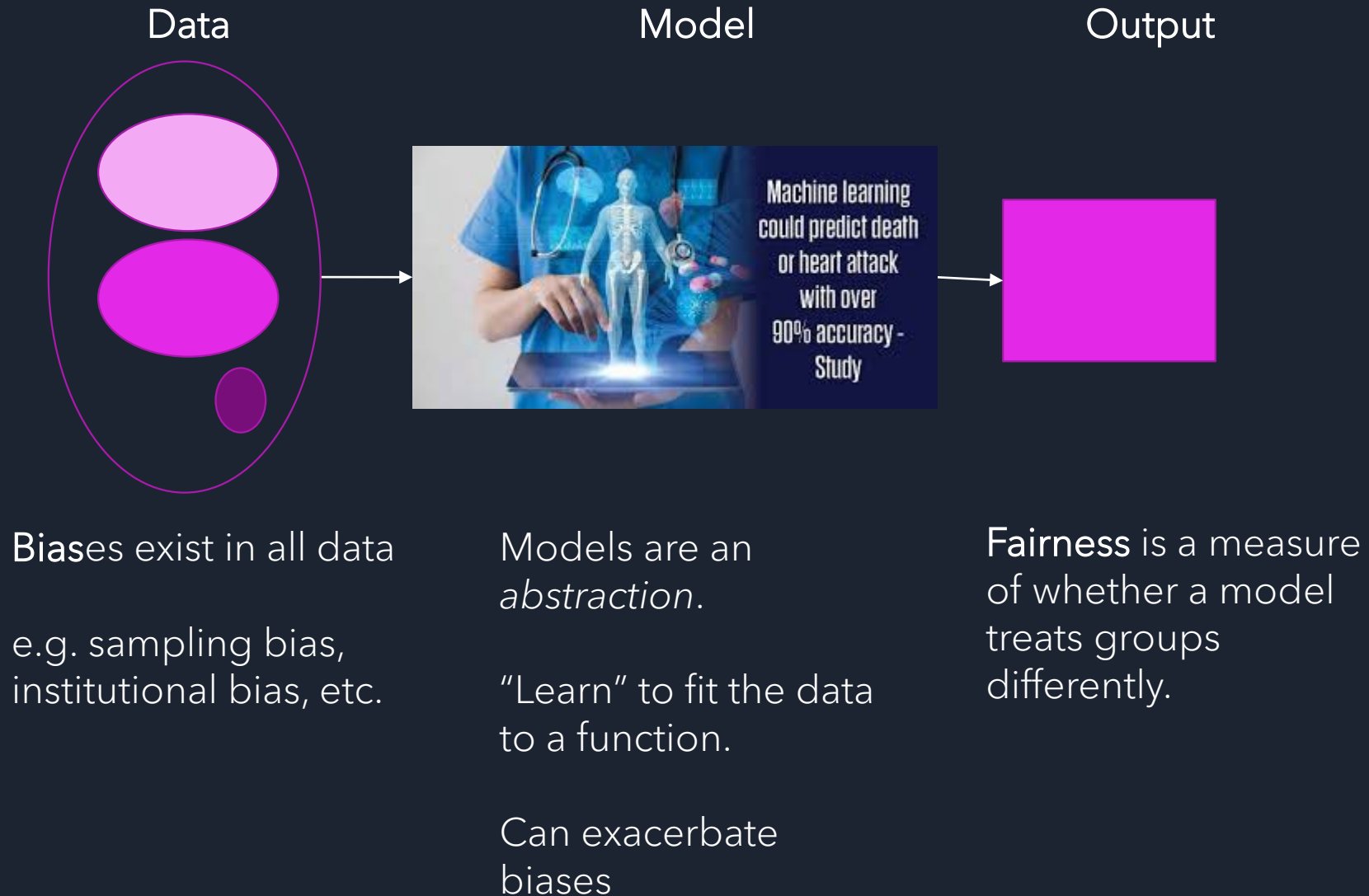
How to get data;  
what's going on in  
the data; how to fix  
the data; new  
models; etc.



DATA- MODEL  
CONSEQUENCES

| *Fairness*

# ALGORITHM FAIRNESS



# 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

## Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

Isobel Asher Hamilton Oct. 10, 2018, 5:47 AM



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)



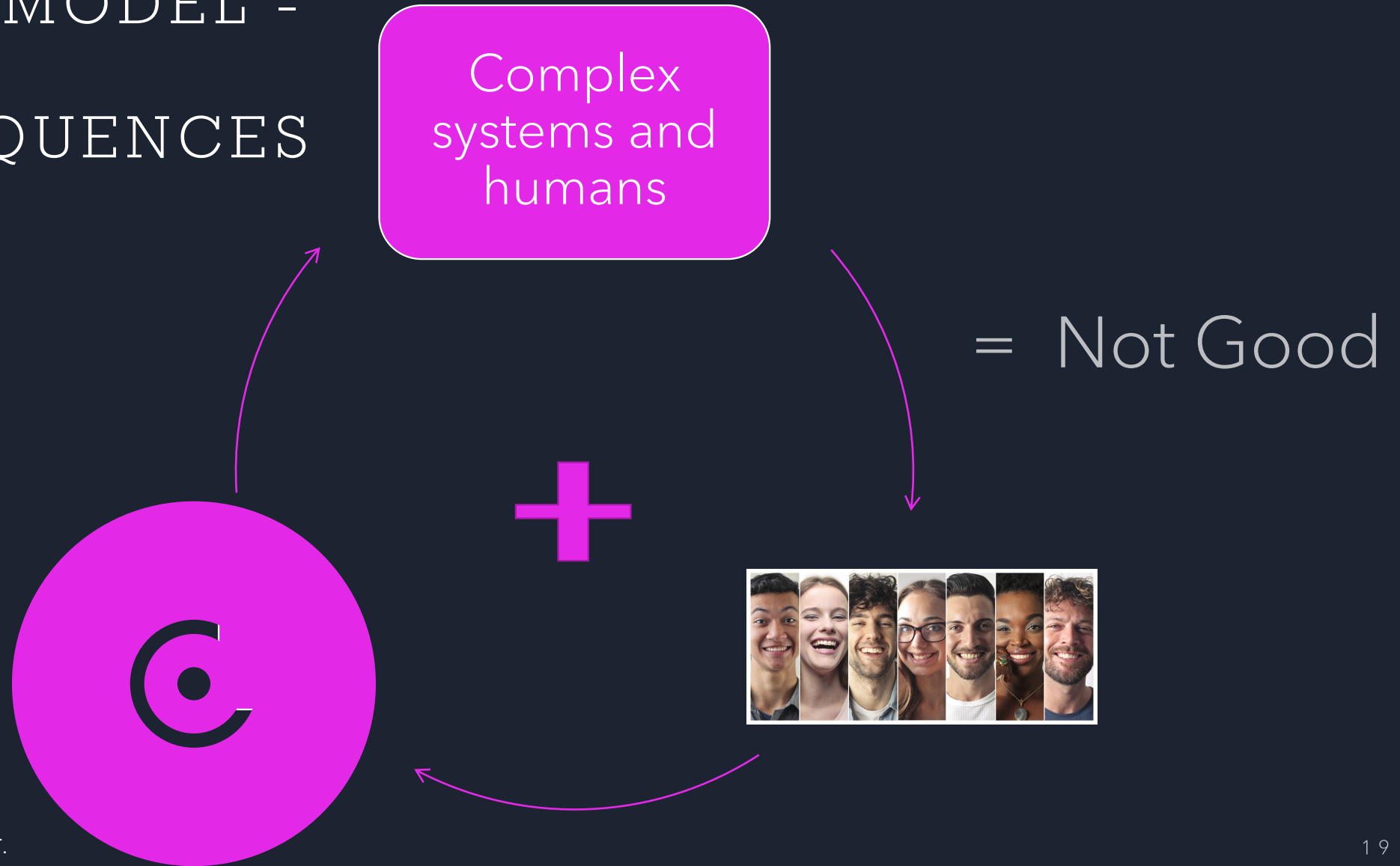
- The frame of reference  
*Definition: Given data  $D = \{x_1, x_2, \dots, x_n\}$ , target labels  $L = \{y_1, y_2, \dots, y_n\}$  find a hypothesis s.t.*
- Data: transformation of raw data into feature vectors determines unit of analysis and quantification. Data exists in context.
- Labels: Discretization 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

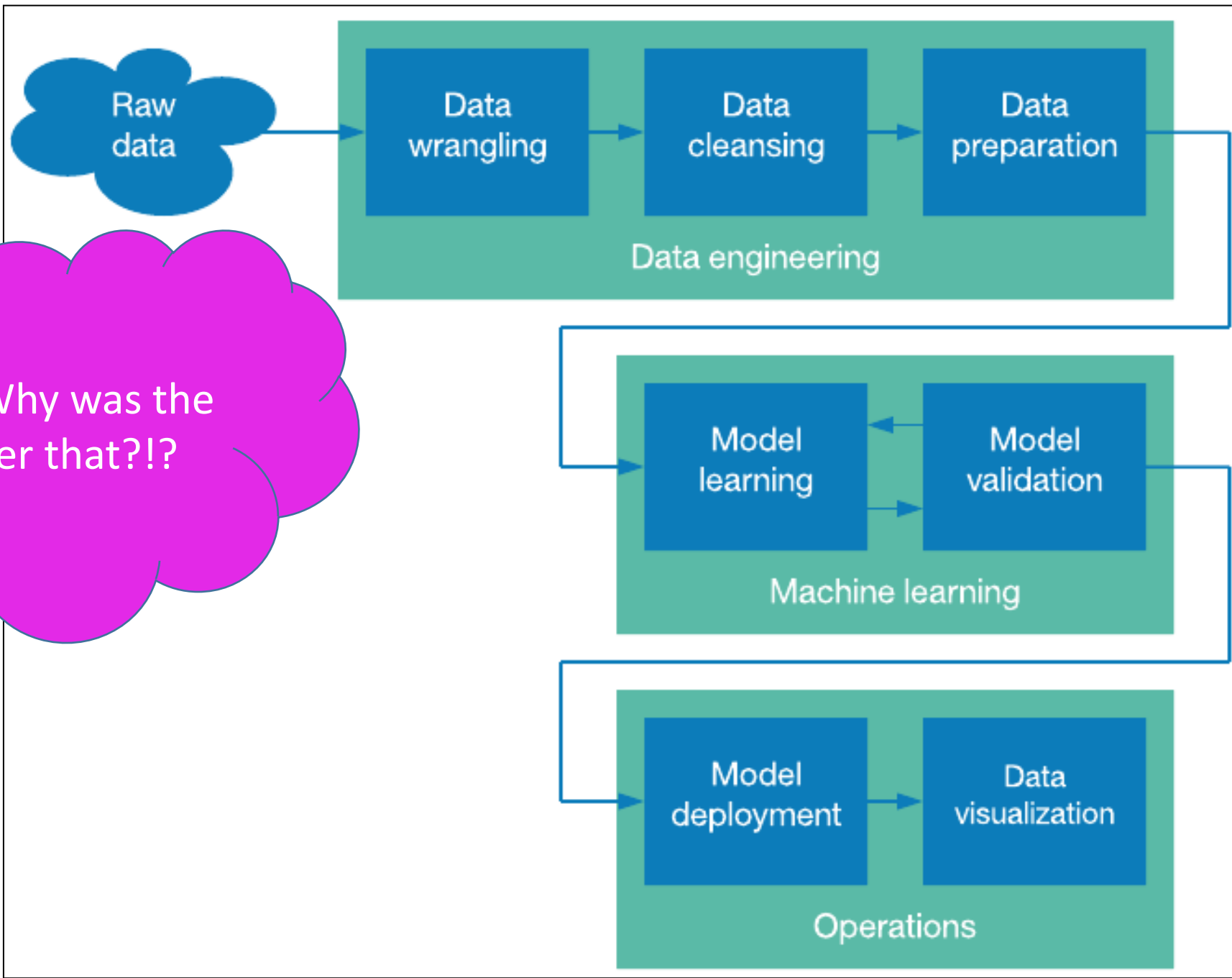
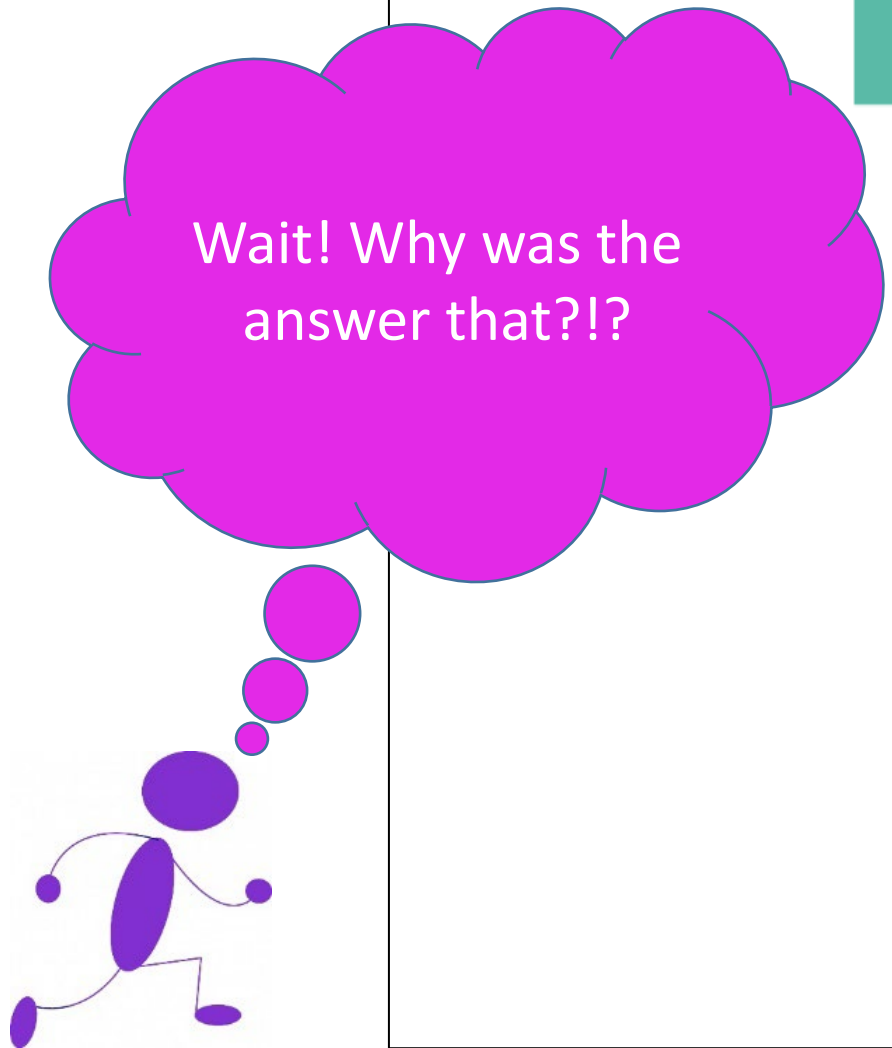


HOW DOES YOUR WIDGET CREATE DATA?  
WHAT ARE THE BIASES IT WILL CONTAIN  
AND HOW COULD IT SPIRAL?



PROVENANCE

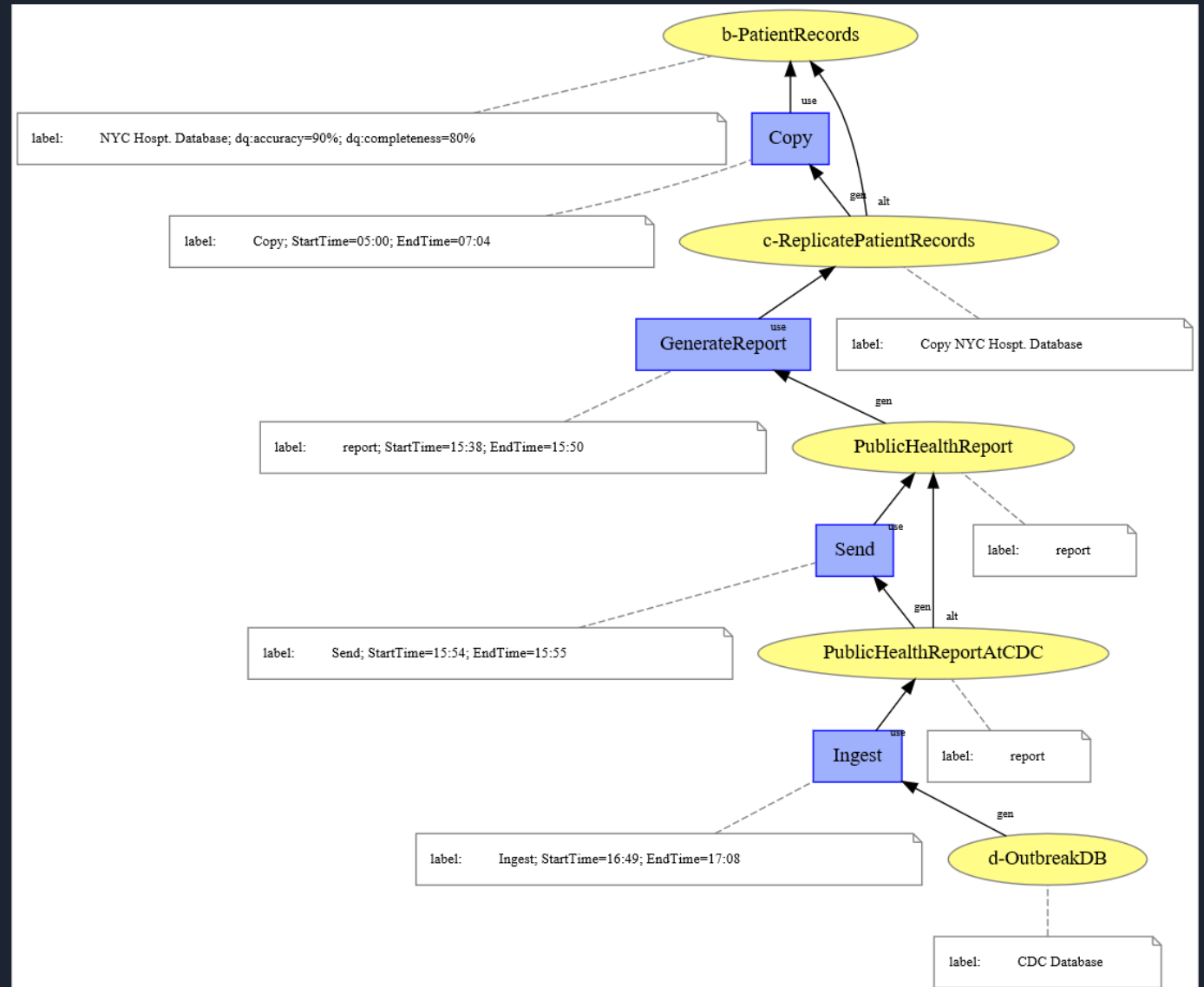
*Transparency and Explanations*



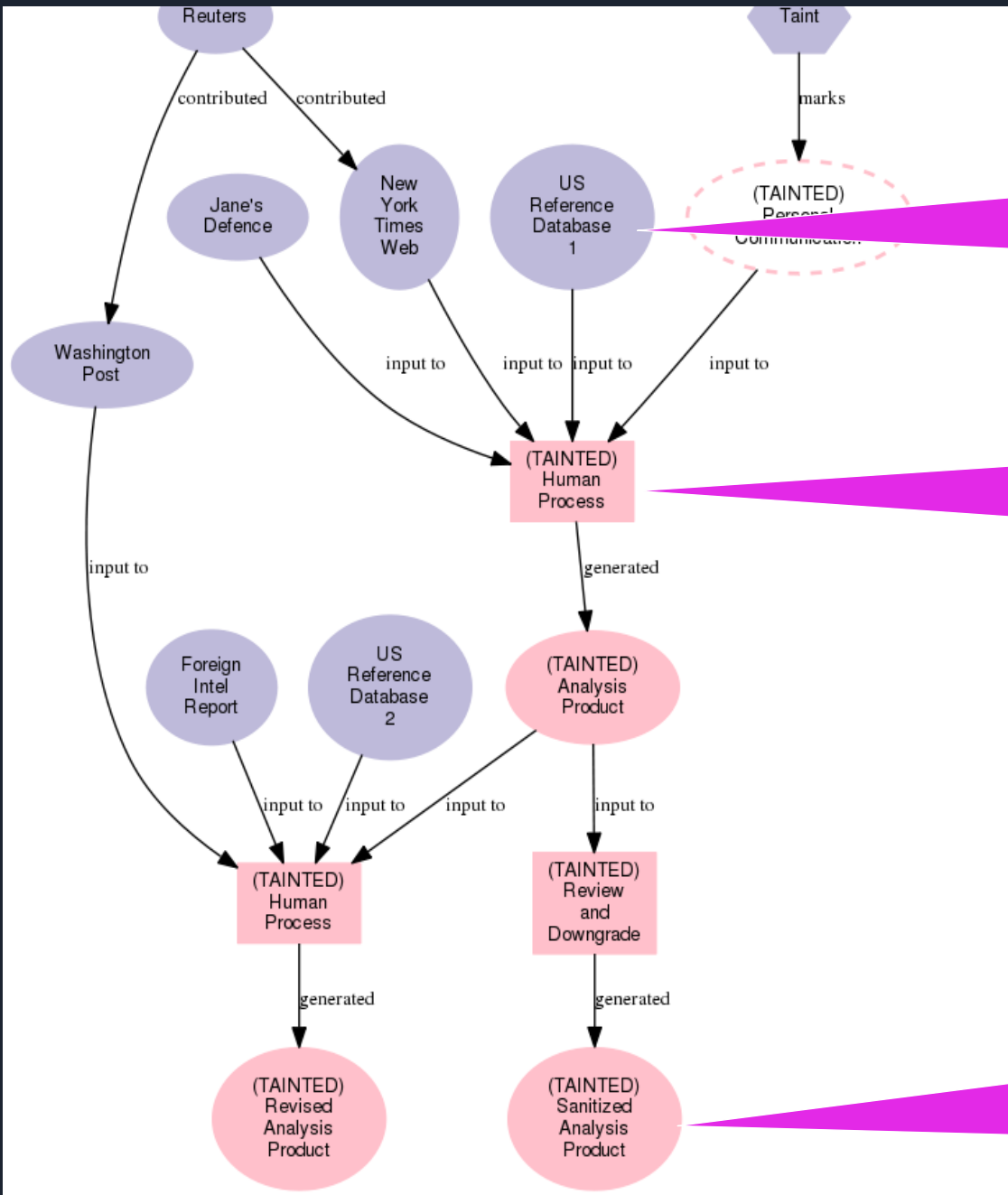


# PROVENANCE\*: A "FAMILY TREE"

- A record of what actually happened
- Agents, Entities, Activities
- Important attributes: timestamps
- Provenance is a type of metadata



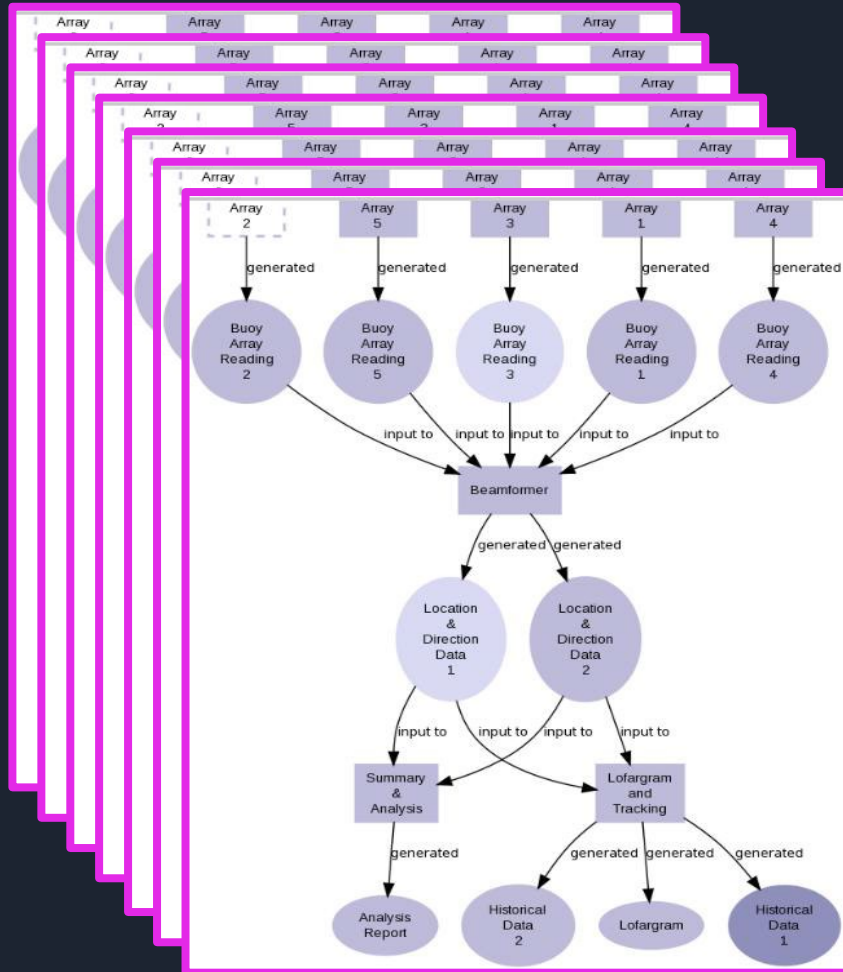
\* aka Pedigree, Lineage



Organization: Who is using my data?

Developer: This data is tainted!

User: Do I "trust" this data?



Manager: Analyst Joe's prior confidences are always correct

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

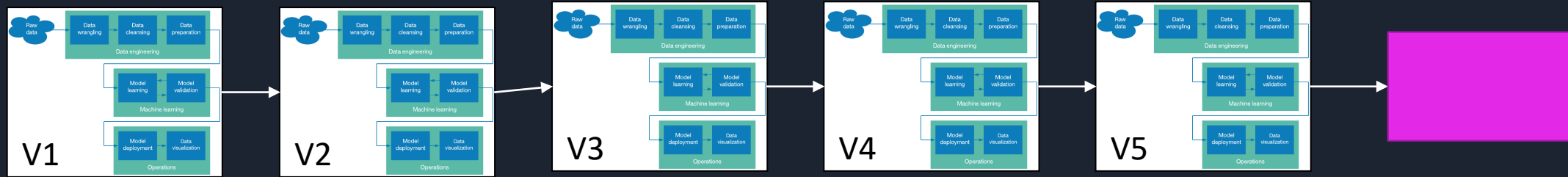
Manager: What parts of my team need retraining? (i.e. less productive)

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

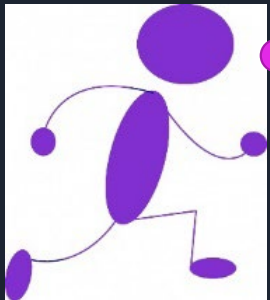
Analyst: If you are interested in X, others were interested in Y and Z.

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

# 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





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.

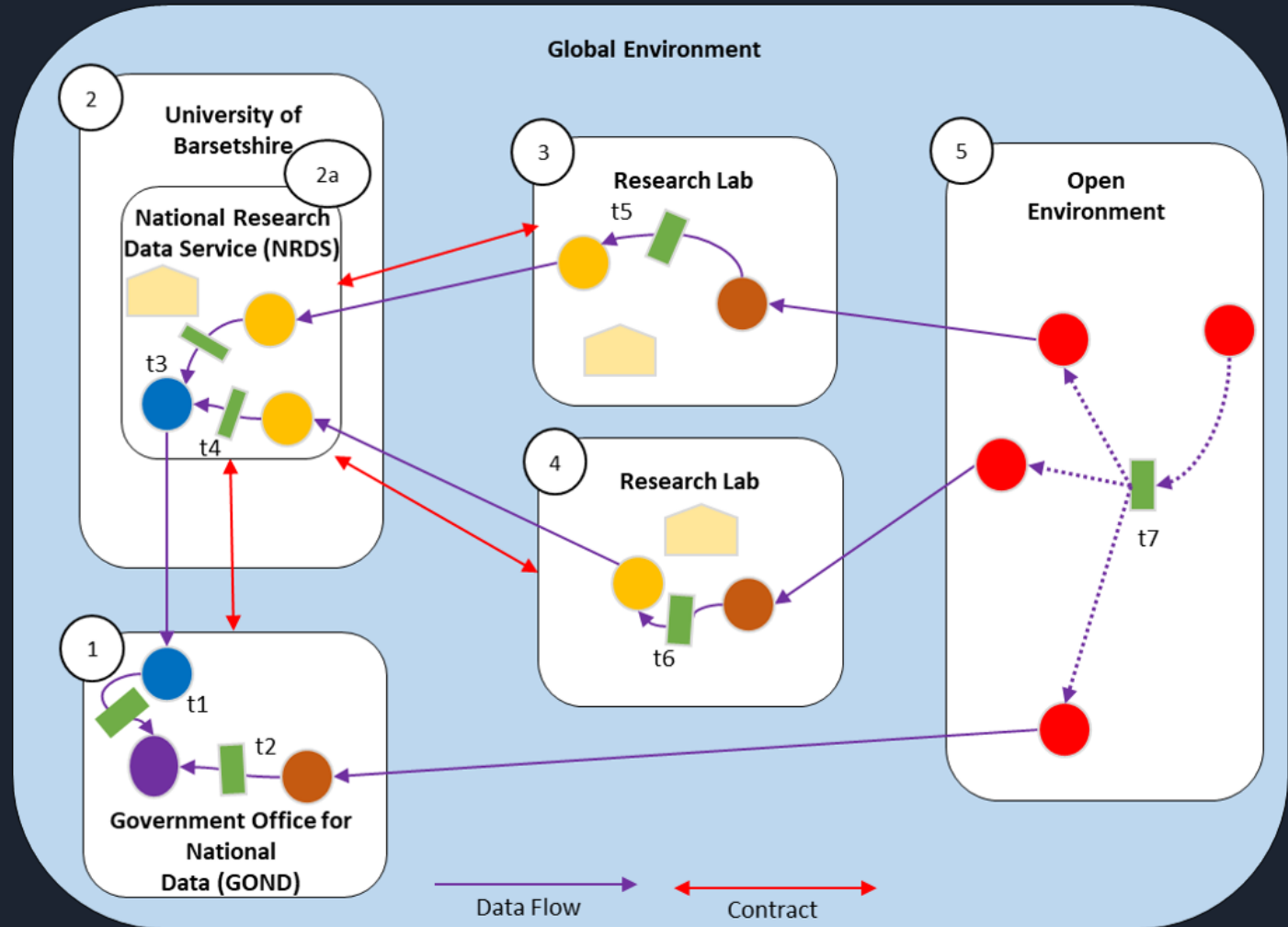
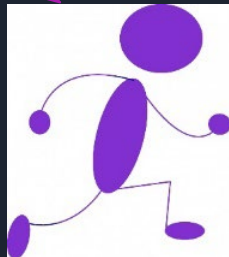
WIRED What Happens When You Track Your Boyfriend on Strava

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



Don't give back (Name, Disease) associations

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

# 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!*