AI and Social Justice: Beyond the Noise

Glen Berman and Charlotte Bradley

How this session will work

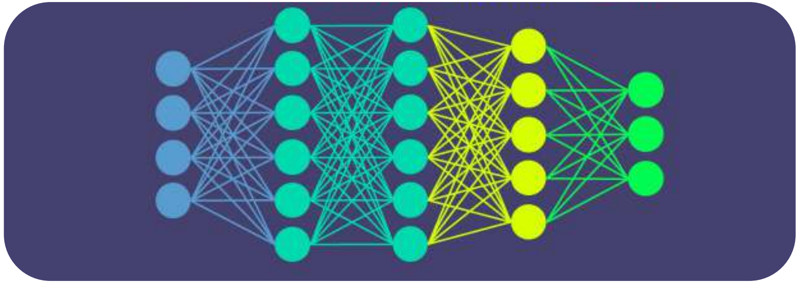
We’ll introduce **two frameworks**, and a case study for each framework.

Aiming to have lots of audience engagement (sorry).

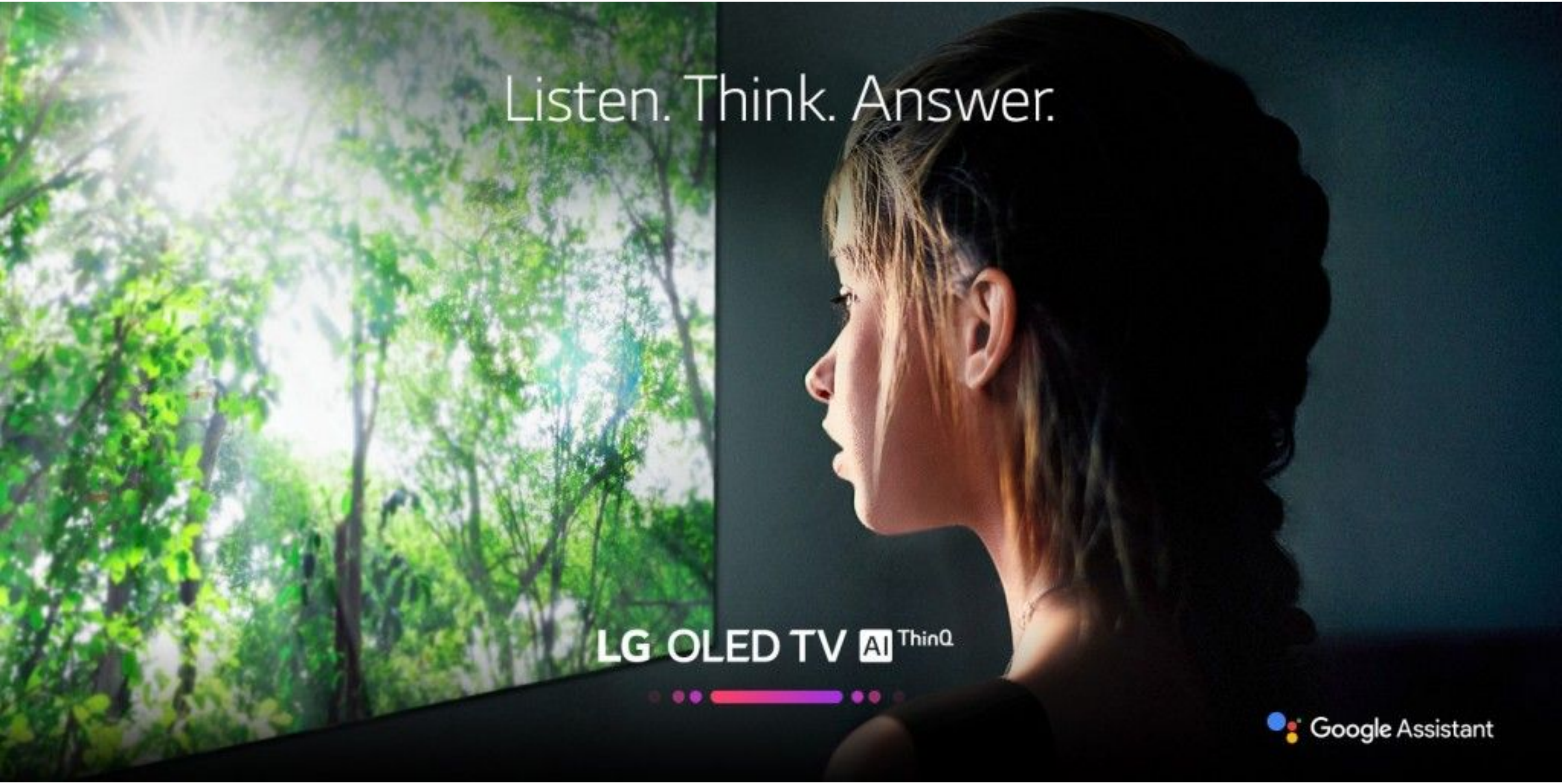
We’re making it all up.

Ask questions.

Introduction



What is AI?


Listen. Think. Answer.
LG OLED TV AI ThinQ
Google Assistant

AI is not one thing

AI is experienced by people as a combination of:

- Machine Learning (= pattern recognition)
- Data
- Sensors
- Algorithms
- Infrastructure & Resources

Ask questions

- Sensors
  - What are they *actually* measuring? What assumptions do they rely on?
- Data
  - What are the categories of data? Why were they chosen?
  - What is missing from the data? What is missing from the category labels?
- Algorithms
  - Who created an algorithm? What was their intent?
  - How have the instructions been adjusted over time, by whom, in response to what?
- Infrastructure & Resources
  - What human and natural resources are required to build, run and maintain a system?

Case study: healthcare algorithms

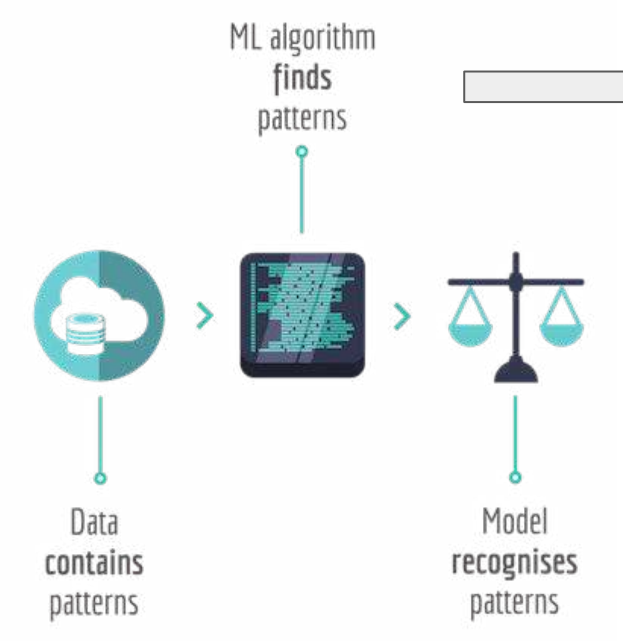
The problem: hospitals and healthcare systems are overstretched. They want to direct their resources to the people who most need them.

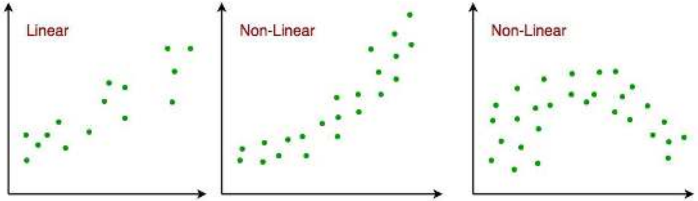
The solution: care coordination programs (e.g. early intervention or preventative medicine programs) -- try to help people before they become really sick.

The challenge: how can we identify which people are most likely to become really sick, and therefore who best to target with care condition programs?

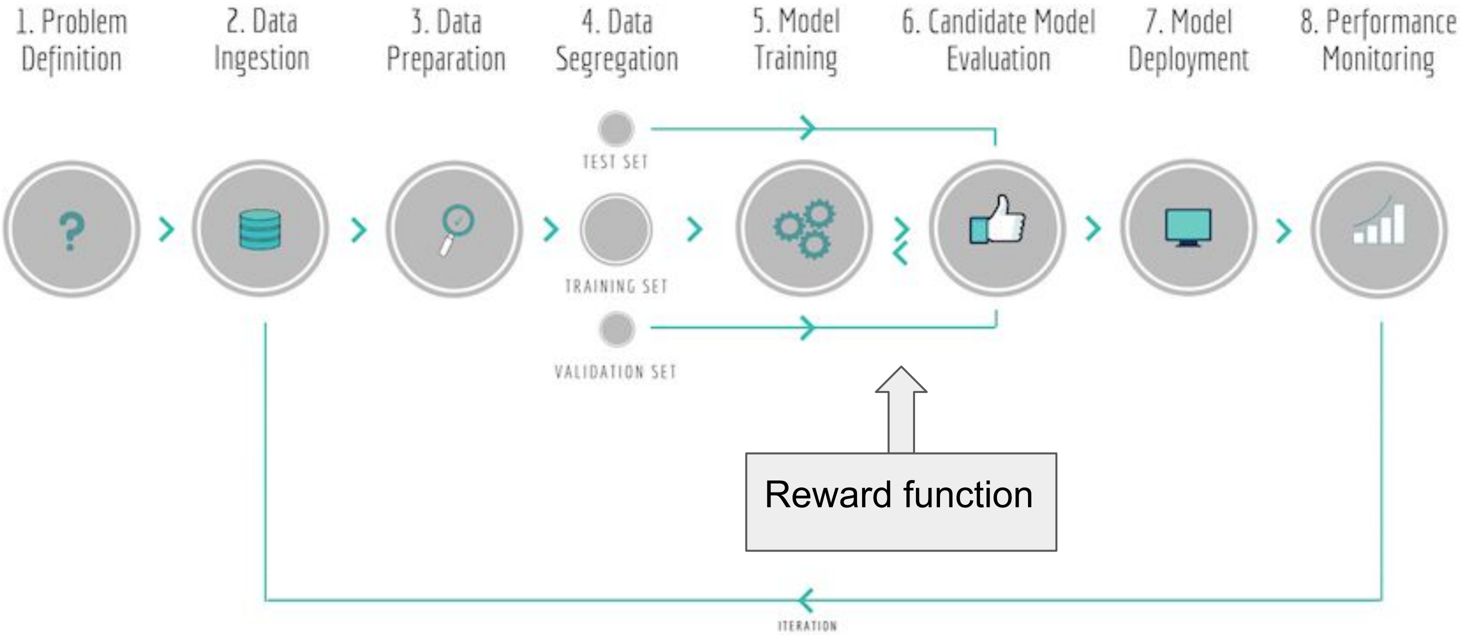
Seems like a good challenge for an AI system to try to address…
lets use ML!

Using machine learning





Using machine learning



Reward function

Case study: healthcare algorithms

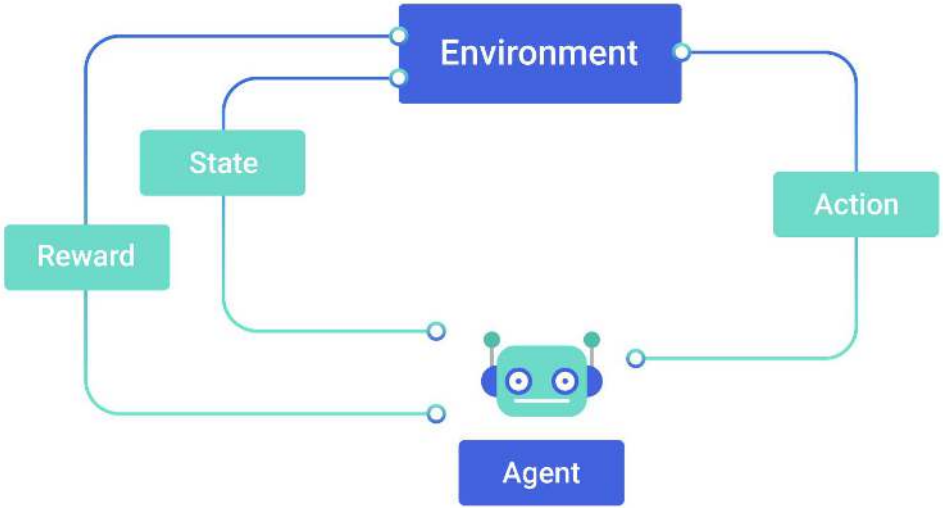
- Sensors
  - What are they *actually* measuring? What assumptions do they rely on? **Cannot measure actual health of people directly on a population scale. Have to rely on a proxy.**
- Data
  - What are the categories of data? Why were they chosen? **Primary category: healthcare expenses. Chosen because lots of data available on this.**
- Algorithms
  - Who created an algorithm? What was their intent? **Created as a commercial product, intended to predict patients at most risk of chronic ill-health.**
  - How have the instructions been adjusted over time, by whom, in response to what? **Deployed by numerous hospital institutions, used in different ways.**

Working with AI

- Human-centred design
- What is ML good for users?
  - Personalisation (e.g. customising donation asks on an individual by individual basis)
  - Prediction (e.g. forecasting email performance)
  - Natural Language Processing (e.g. converting phone call messages to text)
- What is ML *not* good for users?
  - Predictability (i.e. “fuzzy prediction”)
  - Static or limited information (e.g. extracting credit card details from photos of credit cards)
  - Situations where the error cost is high (e.g. automated debt recovery emails from a government department, sent to financially vulnerable people)

Working with AI: Reward Functions

- Reward function, objective function, loss function



Example system: binary classifier

		Prediction	
		Positive	**Negative**
Reference	**Positive**	🙂 True Positive	😔 False Negative
	Negative	😔 False Positive	🙂 True Negative

Example system: different outcomes, different costs

		Prediction	
		Positive	Negative
Reference	Positive	😄 True Positive	😔 False Negative
	Negative	😢 False Positive	🙂 True Negative

Putting this into practice

We manage an email campaign list. We have data going back three years on all the emails we've sent to our members.

The problem we face is it is very hard to predict which issues individual email subscribers are most likely to be interested in. Our current work around is that we test emails to small subsets of our email list, and if they perform well, we then send them to our broader list. But even so, this means that many of our emails are only opened by 20% of our members, and actions are only taken by 6% of our members.

Exciting news: a team of data scientists have volunteered to work with us on developing a ML system that will predict which members are most likely to be interested in a given campaign, which will enable us to micro-target emails to specific members.

Example system: precision or recall?

- Precision - proportion of true positives correctly categorized out of all the true and false positives. Do we want to be *absolutely* correct?
- Recall - proportion of true positives correctly categorized out of all the true positives and false negatives. Or do we want to include as many options as possible?
- Systemic impact:
  - Access and inclusion
  - Impact over time
  - What if it was perfect?

Additional example: descriptive or normative?

Descriptive	**Normative**
Accurate when compared to human behaviour	Inaccurate when compared to human behaviour
Potentially racist/sexist/etc (like human behaviour)	*Potentially* neutral (unlike human behaviour)
Almost all ML systems are descriptive, but we often expect them to behave in normative ways.	

Links to resources

- Racial bias in healthcare algorithms: https://science.sciencemag.org/content/366/6464/447.full
- PAIR guidebook: https://pair.withgoogle.com/
- 3A Institute: https://3ainstitute.cecs.anu.edu.au/
- ML -- what's the process for using ML: https://towardsdatascience.com/not-yet-another-article-on-machine-learning-e67f8812ba86
- ML -- what is ML really: https://medium.com/hackernoon/the-simplest-explanation-of-machine-learning-youll-ever-read-bebc0700047c
- Descriptive vs normative -- breaking down bias in ML: https://podtail.com/en/podcast/ai-australia/bias-in-machine-learning-systems-with-katherine-ba/
- Algorithms as administrative mechanisms: https://journals.sagepub.com/doi/full/10.1177/2053951718757253