Prediction and Causality:
How Can Machine Learning be Used for COVID-19?

Mark Crowley

Assistant Professor

Electrical and Computer Engineering

mcrowley@uwaterloo.ca

http://waterloo.ca/scholar/mcrowley/lab

@compthink

UNIVERSITY OF WATERLOO

WATERLOO.AI

WATERLOO ARTIFICIAL INTELLIGENCE INSTITUTE

Application/Experimental Domains

*To augment human **decision making** in complex domains and environments in a **dependable** and **transparent** way.*

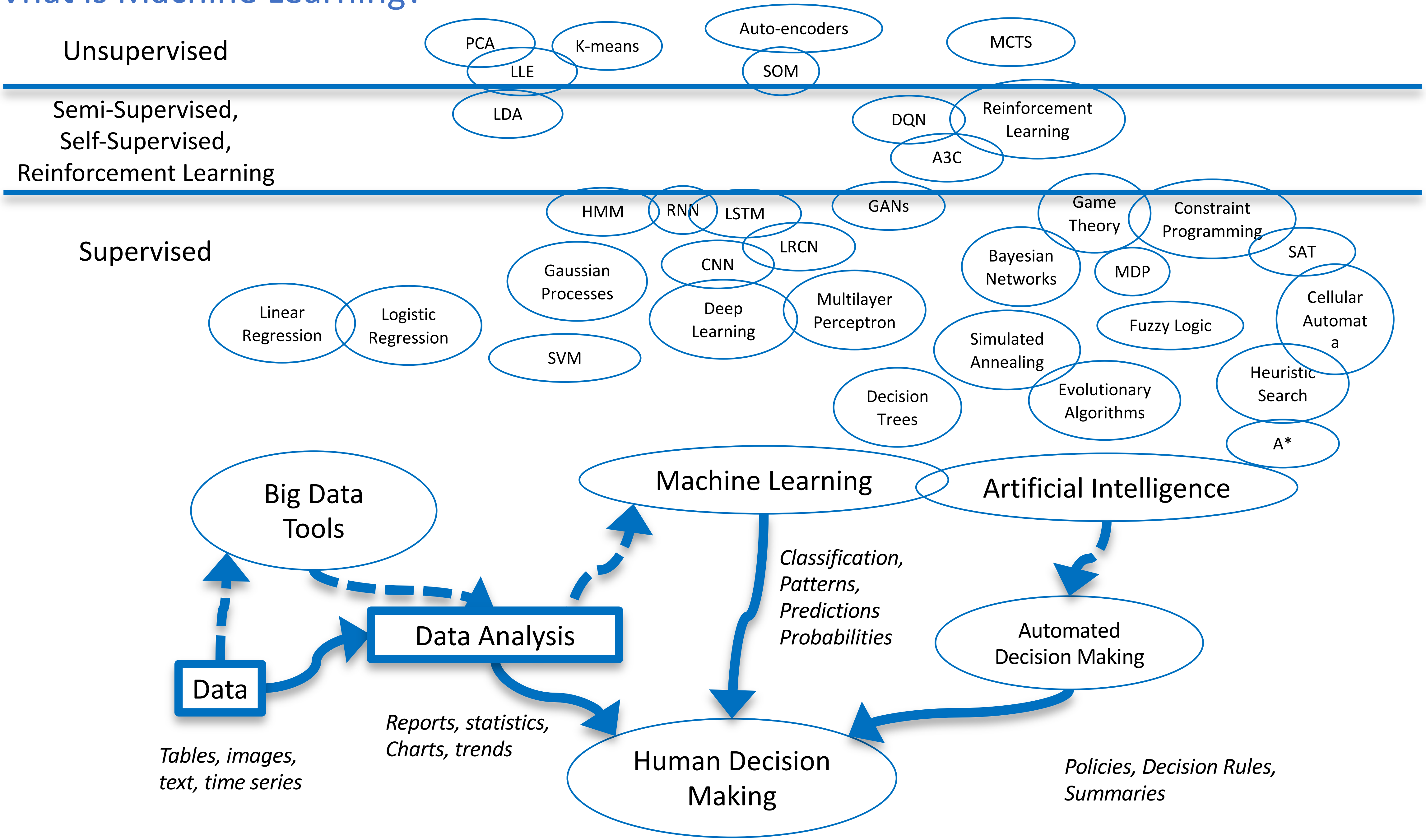
Domains

- **Medical Data Analysis**
- **Physics + ML**
- **Automotive**
- **Spatially spreading processes**
- **Natural Language Processing**

Methods

- **Reinforcement Learning**
- **Deep Learning**
- **Manifold Learning**
- **Dimensionality Reduction**
- **Streaming Anomaly Detection**

What is Machine Learning?



Unsupervised

PCA

LLE

K-means

Auto-encoders

SOM

MCTS

Semi-Supervised,
Self-Supervised,
Reinforcement Learning

LDA

DQN

Reinforcement Learning

A3C

Supervised

HMM

RNN

LSTM

GANs

Game Theory

Constraint Programming

LRCN

CNN

Gaussian Processes

Bayesian Networks

MDP

SAT

Linear Regression

Logistic Regression

Deep Learning

Multilayer Perceptron

Cellular Automata

Fuzzy Logic

SVM

Simulated Annealing

Heuristic Search

Decision Trees

Evolutionary Algorithms

A*

Big Data Tools

Machine Learning

Artificial Intelligence

Data Analysis

Data

Tables, images, text, time series

Reports, statistics, Charts, trends

Classification, Patterns, Predictions Probabilities

Automated Decision Making

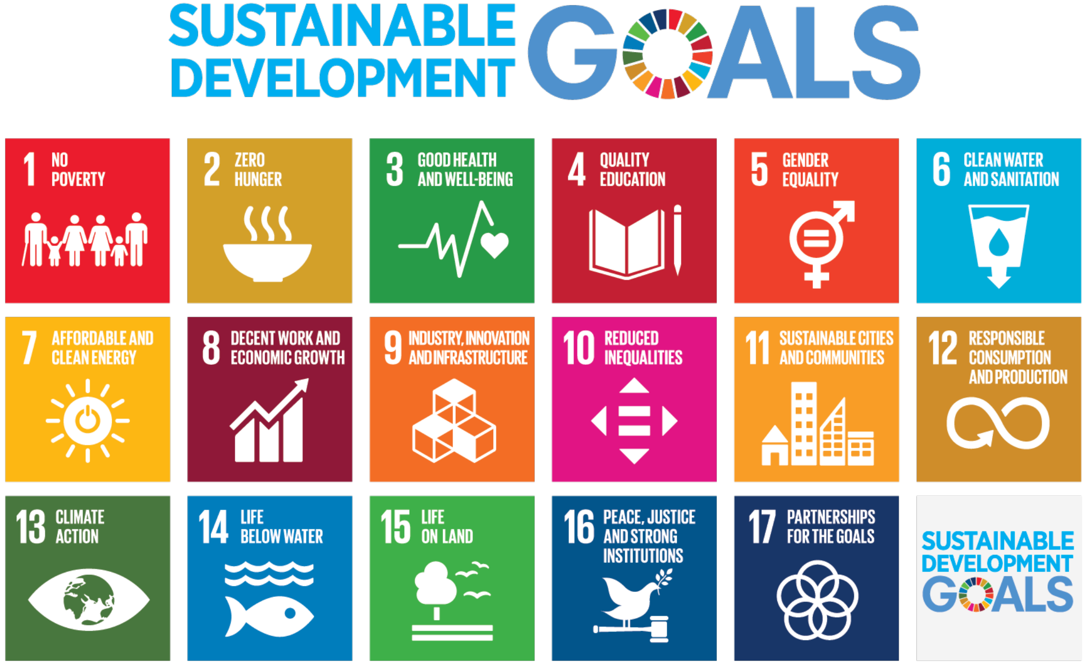
Human Decision Making

Policies, Decision Rules, Summaries

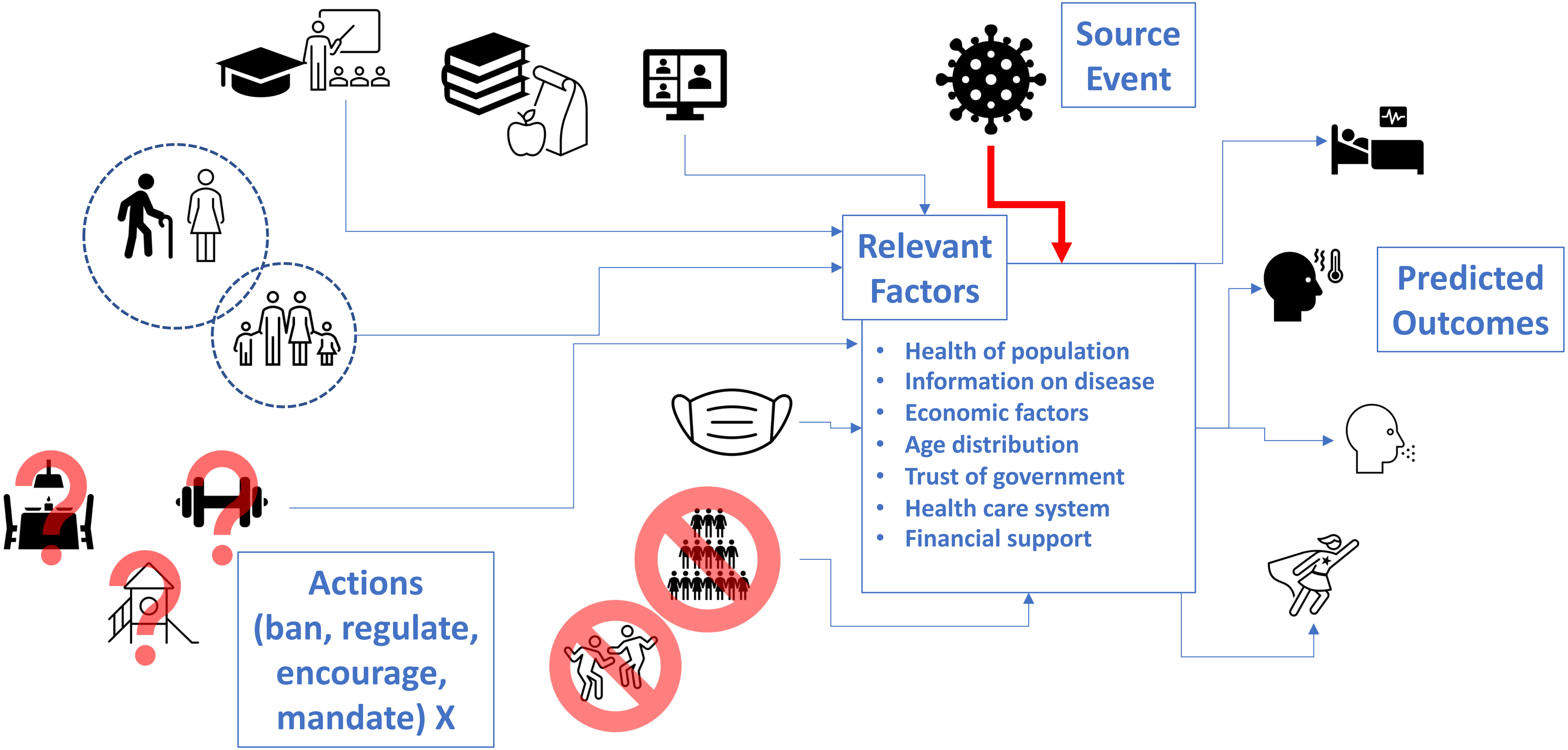
Nice Problems vs Important Problems

- Playing Chess - cool
- Classifying cat videos - surprising
- Alpha go – amazing!
- Text and speech recognition, translation - so handy!
- Advertisement Ranking – profit!!!
- …no seriously though, how about something important?
- Some of these can be important, but they also have low risk if you’re wrong.
- Update 2020:
  - *How about protein folding?*

Luckily, somebody made us a handy list


SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

What is COVID-19? (from an ML Researcher’s POV)



Source Event

Relevant Factors

- Health of population
- Information on disease
- Economic factors
- Age distribution
- Trust of government
- Health care system
- Financial support

Predicted Outcomes

Actions (ban, regulate, encourage, mandate) X

How AI and ML are being use right now

- *Spread Prediction:* Improve prediction spread models in one region given model of disease using in-region data.
  - Improving Prediction of ICU patient levels in Trillium/Peel region hospitals
- *Mutation:* Analyse genetic variation patterns in the disease as it mutates.
- *Logistics:* Predict and manage shortfalls and of hospital supplies such as PPE, cleaning supplies and basic medicines.
- *Anti-viral drug discovery:* aiding the search for promising chemicals to test in various stages of drug development to speed up the process.

But this is a privacy, data and COVID conference:

- So I will focus on the possibilities for **using ML to predict and understand** the spread of the disease **using data about people, government interventions and policies.**

Causal Modelling

Significance Tests / Hypothesis Testing

- This is the bread and butter of much of science, but it's also highly restrictive
- “Controlled Experiment”, A/B Testing
- Its results can be very easily misunderstood, or worse, purposely misused
- Good for Critical Causal Questions:
  - Does the vaccine work? Do masks work?
  - Article about Human Challenge Studies
    - (https://www.theguardian.com/world/2020/sep/24/uk-covid-19-vaccine-trial-set-to-infect-healthy-volunteers-with-virus)
- Not so good for unknowns where we can’t run a trial:
  - Should we close restaurants and gyms?
  - Or should we have closed schools instead?
  - Or maybe we should have kept them all closed longer in the first place?

Causal Modelling

Causal Inference (Judea Pearl, 2018)

Pearl describes **the causal hierarchy**.
It is a general relationship between *data, labels, knowledge and causality:*

I. Association – passive observation

II. Intervention – active experiments, interventions

III. Counterfactuals – retrospective analysis, what-if scenarios

ML’s Dirty Not-So-Secret

- Machine Learning is **powerful**
  - it’s defeating world champions at their own games
  - it’s flying spaceships, driving cars
  - it’s optimizing addictive click-ability of social media apps *really really well*
  - ML is *even* (just this week?) solving the mind bogglingly complex challenge of *protein folding*
- BUT….it’s almost all associations!
- Remember the warning *“Correlation does not imply causation”*?
  - That applies to 90% of Machine Learning (Deep Learning too)
  - It turns out that correlation, or **association**, is easy with enough parameters
  - But what if we wanted more?

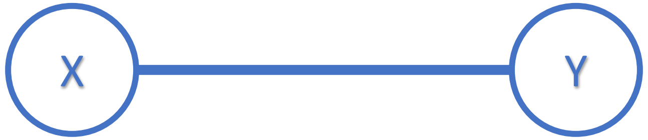
What ML is good at: Association (I)

_ _ _ _ _ _ _ _ _ _

Given many training pairs (x,y)…

$p(x,y)$ – are x and y related?

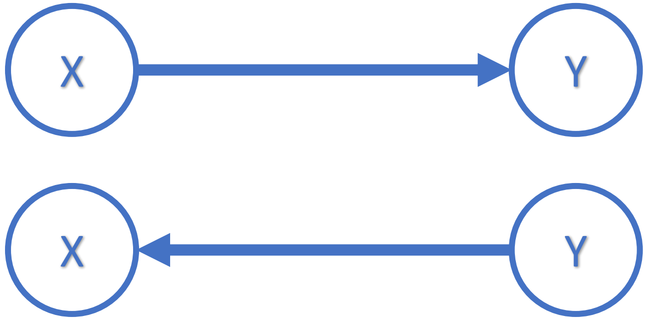
- Positively/negatively correlated



"If **x** goes up by 1, then **y** goes up by 3.74"

eg. High COVID-19 fatalities is correlated years over 70.

$p(x \mid y)$ or $p(y \mid x)$ – predict x or y after training on (x,y) pairs



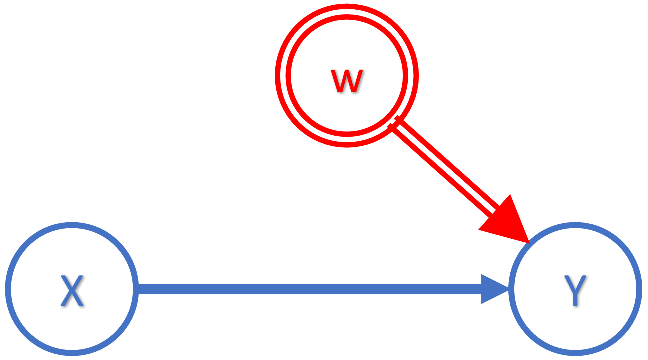
"If I know **x** I can predict **y** very well, and vice-versa."

eg. TODO example for not knowing the causal direction

What We Actually Need

Intervention (II)

$p(y|x, do(w))$ – if we do *w* to *x* what happens to *y*?

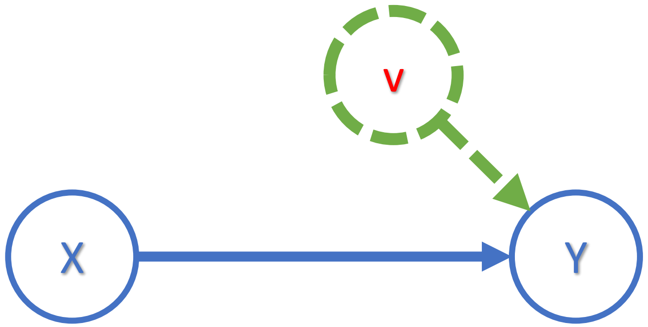


"If you prick us, do we not bleed?"

eg. The probability the patient will develop COVID-19 (y) that exposed to conditions x and given a vaccine (w).

Counterfactuals (III)

$p(y_v|x, y)$ – counterfactual…



"So, I know that we observed **x** and **y**, but *what if* we had seen **v** instead? Would **y** still have happened?"

eg. What if they had just worn masks…

ML Methods for Inferring Causality

- Causal Hierarchy Theorem:
  - In most cases, in order to answer questions about Level you need to have a model for the lower Level.
- Learn Causal direction from data
  - functional decomposition – compare to Gaussian
  - Natural shocks as interventions
- Propensity Score Matching
  - Finding subpopulations with similar features but some intervention changes
  - Using the expected outcome vs the actual outcome to make a case for a causal relationship among factor

- Any hope for counterfactuals (III)?
  - One step at a time, a full (II) model is needed first.
- Causal graphs and Structural Equation Models allow us to
  - work out the right way to calculate probabilities
  - whether it is possible at all with the given information
- But you really need to
  - get *all the data in one place* and
  - know all these things about it
- This makes privacy a challenge…

Privacy Concerns

- Canada’s COVID App
  - There will be other talks on this
  - Privacy is the highest possible standard, maybe too high?
  - Canada’s app is so private and secure that the government can’t post analyze the data, can’t force people to enter their diagnosis codes.
  - It also only does one step of tracing at a time.
    - But in reality, there is enough information in the network to inform people multiple degrees of separation out, if we allowed it.
- To do more with data…
  - we need more data and we need to know where it came from
  - and who it came from, at least to the extent that we have information like neighbourhoods, activities, interactions with others

Let a Thousand Experiments Bloom

The global response to COVID in each country, region and city is like a **thousand different experiments** being carried out at once on the same problem.

- But! These experiments are not “controlled experiments” in the statistical sense.
- We cannot just take the results from Sweden (No Mask Mandate) and make inferences about a Province in Canada without
  - Matching the factors in common in both populations
  - Analysing the differences and adding that to our model
  - Using a causal graph structure to work out which questions can be answered

Using the Global Unintentional Experiment

Challenge:

- Can we collect together all the data about interventions, in different countries and regions?
- Can we combine all the demographic, economic, even socio-political information we have?
- And then, can we learn a *useful* model that explains the differences in outcomes in different regions of the world based on those factors?

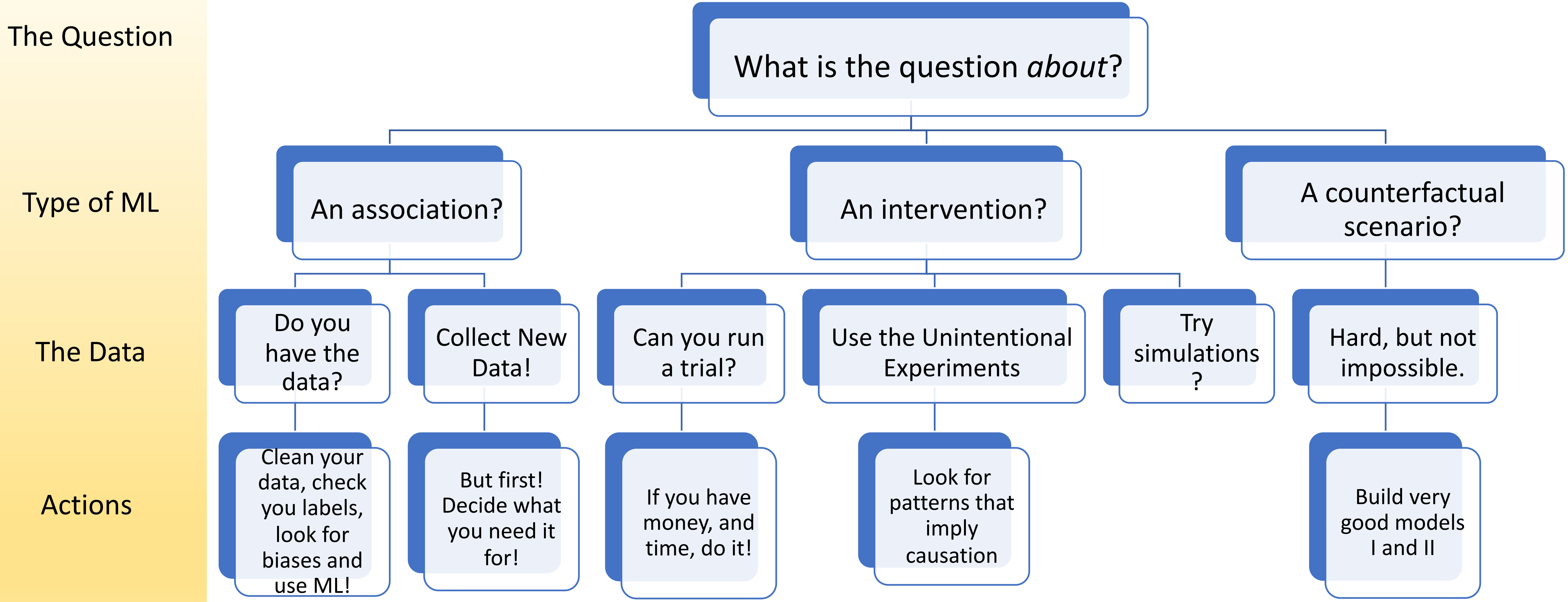
Doing this well requires us to be able to answer all three types of questions.

What is Learnable?

Can we learn about the effectiveness of any policy interventions?   … It depends…

- Mask mandate, social circles, curfews and travel perimeters…
  - It’s not clear we actually *can* collect the data on an individual level?
  - So, results will be aggregate only. → That’s fine, but it needs to be part of the model.
- Phone Proximity Apps
  - We can measure installations regionally. What about locally…by area code?
  - Correlations with education/marketing programs, privacy concerns, trust of government, alternate organizational structures (religion, community)
  - Measure population perspective on app with polls, locally.
  - Combine local data on all these with positive tests, probability of app installation (by age, user type, other information) try to infer how much it is working given what we know.
  - What part of the puzzle is missing?

Is Your Question Learnable?



The Question

What is the question *about*?

Type of ML

An association?

An intervention?

A counterfactual scenario?

The Data

Do you have the data?

Collect New Data!

Can you run a trial?

Use the Unintentional Experiments

Try simulations ?

Hard, but not impossible.

Actions

Clean your data, check you labels, look for biases and use ML!

But first! Decide what you need it for!

If you have money, and time, do it!

Look for patterns that imply causation

Build very good models I and II

Conclusion

- Machine Learning is not a silver bullet to get better predictions or policies:
  - The right question needs to be asked
  - The data needs to align with the question
  - Each method/algorithm/technology are good at specific types of questions and most of those are *associational* only.

Conclusion

- Factors Influencing Spread Reduction:
  - So many! Simple model might be best knowing how much data is missing
  - But the Unintentional Experiments of the world offer an opportunity for learning more if it can be used right

- Possible Influence Factors on Vaccine Effectiveness
  - Medical and Epidemiological factors
    - The usual plus : which strain? What about mutation?
  - Logistical factors (all apply nationally, regionally, locally)
    - Political will
    - Ability to organize vaccine distribution
  - Societal factors (all apply nationally, regionally, locally)
    - Willingness of public to take it (what parts of the public? When?