DATA, ARTIFICIAL INTELLIGENCE, AND ETHICS

PATRICK BOILY, IDLEWYLD ANALYTICS

JENNIFER SCHELLINCK, SYSABEE

Data Action Lab

Sysabee

DAVHILL Group

IDLEWYLD

Combined experience: 50+ university courses, 40+ workshops, 60+ projects, 35+ years.
Qualified for GoC A.I. Source List – EN578-180001/A (Band 1).

CONTENTS

1. Case Study in Digital Government and Citizen Data Use
2. Ethics
3. Artificial Intelligence and A.I. Ethics
4. Taking a step back
5. Impact of A.I./Automation on GoC and CRA
6. Appendix: a Brief Introduction to Neural Networks

HEADLINES

“AlphaGo vanquishes world’s top Go player marking A.I.’s superiority over human mind”
[*South China Morning Post*, May 27, 2017]

“A Japanese A.I. program just wrote a short novel, and it almost won a literary prize”
[*Digital Trends*, March 23, 2016]

“Elon Musk: Artificial intelligence may spark World War III”
[*CNET*, September 4, 2017]

“A.I. hype has peaked so what’s next?”
[*TechCrunch*, September 30, 2017]

CASE STUDY IN DIGITAL GOVERNMENT AND CITIZEN DATA USE

DANISH GOVERNMENT

WHY ETHICS AGAIN?

We all have a personal ethical system, don’t we?

- be honest
- be fair
- be objective
- be responsible
- be compassionate?
- etc.

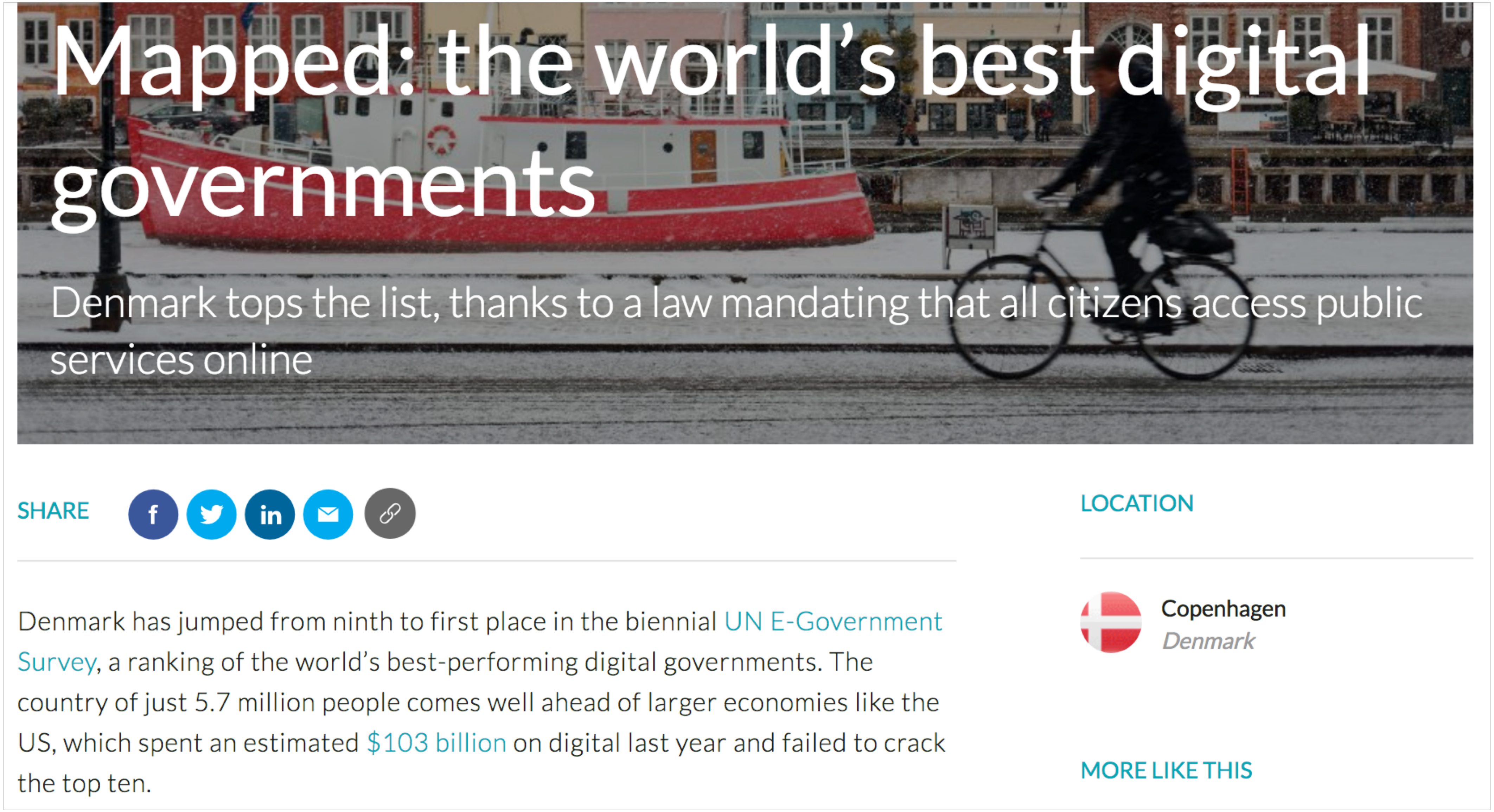
A PRAGMATIC STRATEGY

scenario + broader principle + judgment

COLD EVIL

Andrew Kimball defines "**cold evil**” as a systemic evil in which we are all complicit. As he writes in the essay *Cold Evil: Technology and Modern Ethics.*

"A synonym for the work “cold” is “distant,” and a vital component in the success of modern cold evil is the physical and psychic distance that technology creates between the doer and the deed."



Mapped: the world's best digital governments

Denmark tops the list, thanks to a law mandating that all citizens access public services online

LOCATION

Copenhagen
Denmark

Denmark has jumped from ninth to first place in the biennial UN E-Government Survey, a ranking of the world's best-performing digital governments. The country of just 5.7 million people comes well ahead of larger economies like the US, which spent an estimated $103 billion on digital last year and failed to crack the top ten.

MORE LIKE THIS

DANISH MEDICAL DATA PROJECT

The *Danish National Patient Registry* contains 68 million health observations on 6.2 million patients over a 15 yr time span (Jan ‘96 – Nov ‘10).

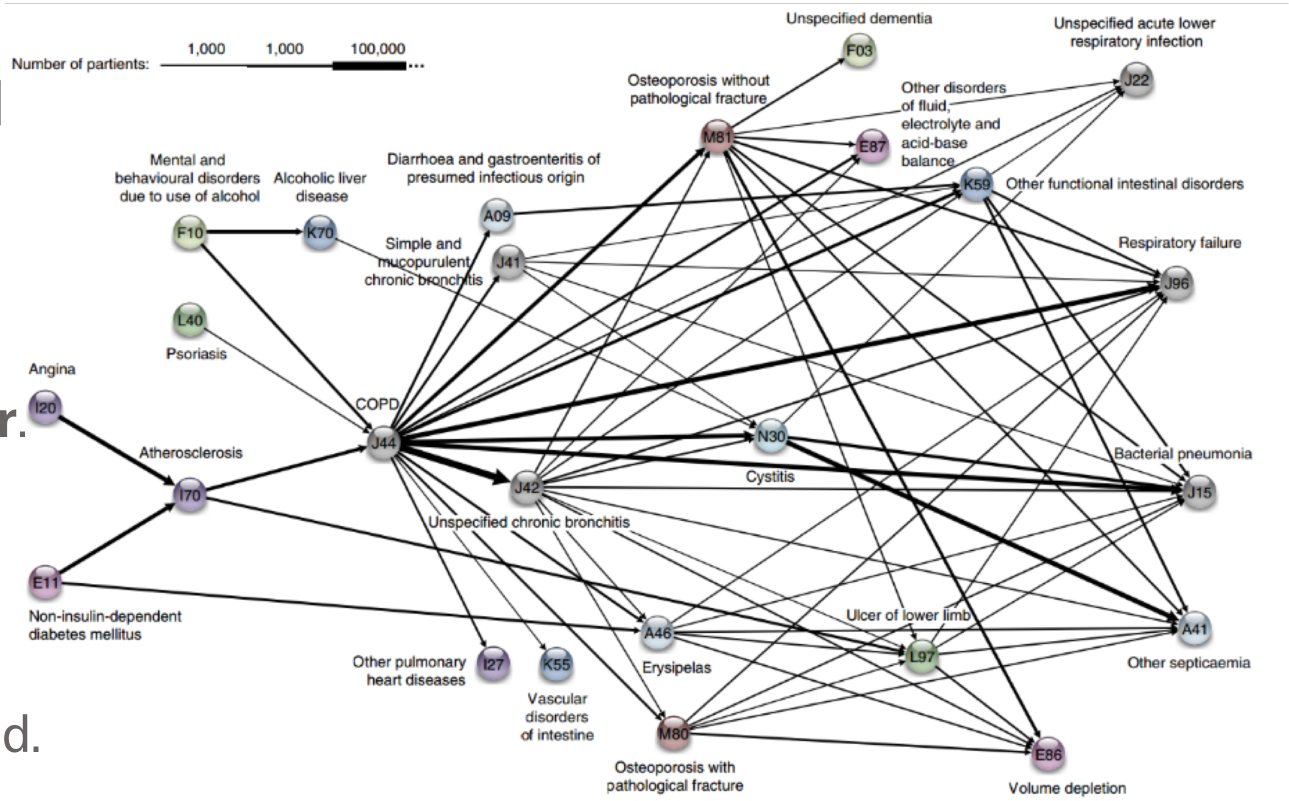
Objective: finding connections between different diagnoses and how the presence of a diagnosis at some point in time might allow for the prediction of another diagnosis at a later point in time.

Reference: Jensen, L.J., et al., *Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients*, Nature Communications

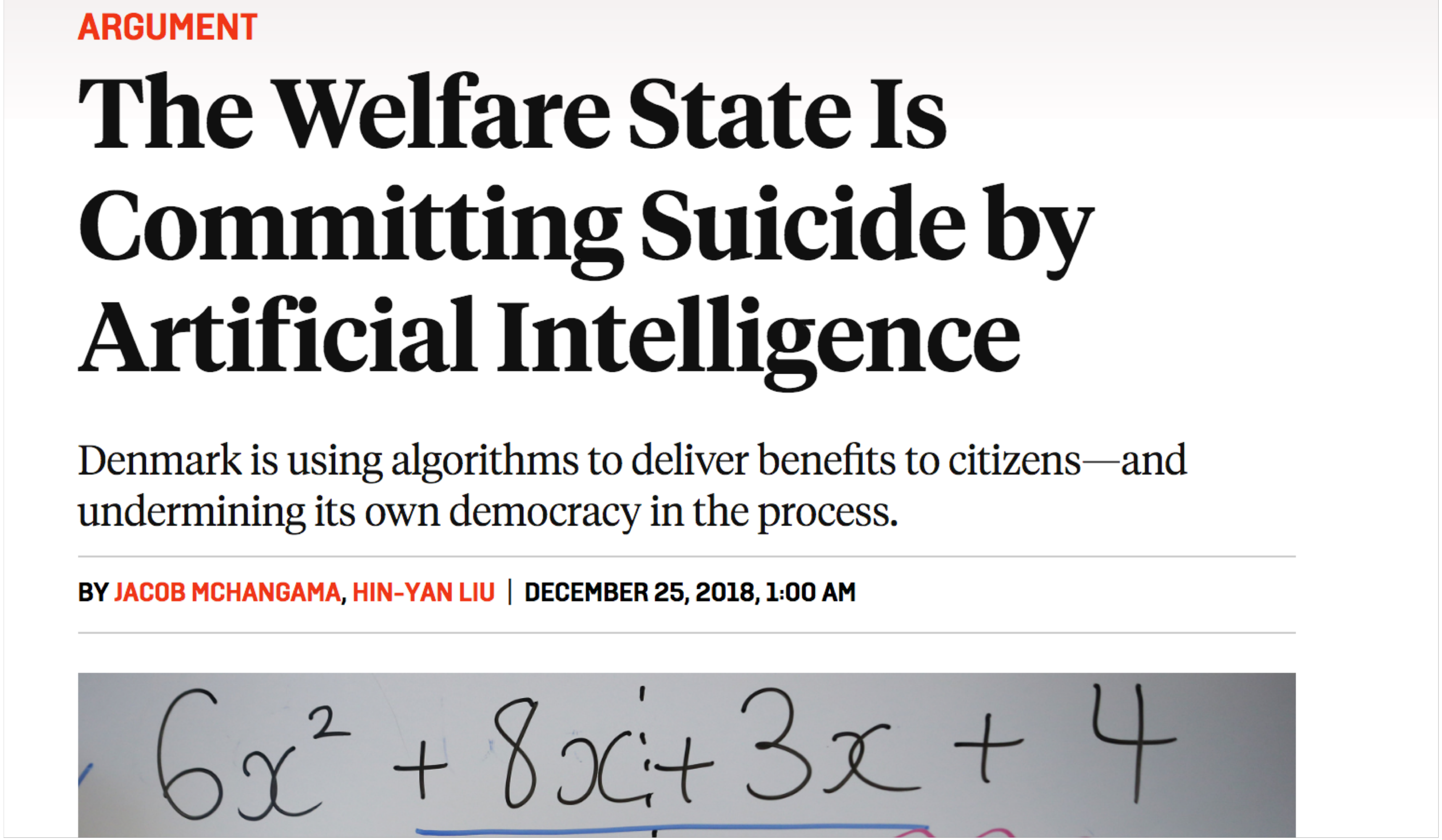
DANISH MEDICAL DATA – RESULTS

Massive dataset was reduced to 1,171 diagnoses **thoroughfares** centered around COPD, arthritis, cardiovascular disease, diabetes, and cancer:

- a diagnosis of **anemia** is typically followed months later by the discovery of **colon cancer**.
- **Gout** was identified as a step on the path toward cardiovascular disease.
- Chronic Obstructive Pulmonary Disease (**COPD**) is under-diagnosed and under-treated.



Number of partients:
1,000
1,000
100,000
Unspecified dementia
F03
Unspecified acute lower respiratory infection
J22
Osteoporosis without pathological fracture
M81
E87
Other disorders of fluid, electrolyte and acid-base balance
K59
Other functional intestinal disorders
Mental and behavioural disorders due to use of alcohol
F10
Alcoholic liver disease
K70
Diarrhoea and gastroenteritis of presumed infectious origin
A09
Simple and mucopurulent chronic bronchitis
J41
Respiratory failure
J96
L40
Psoriasis
Angina
I20
COPD
J44
N30
Cystitis
Bacterial pneumonia
J15
Atherosclerosis
I70
J42
Unspecified chronic bronchitis
E11
Non-insulin-dependent diabetes mellitus
A46
Erysipelas
Ulcer of lower limb
L97
A41
Other septicaemia
Other pulmonary heart diseases
I27
K55
Vascular disorders of intestine
M80
Osteoporosis with pathological fracture
E86
Volume depletion

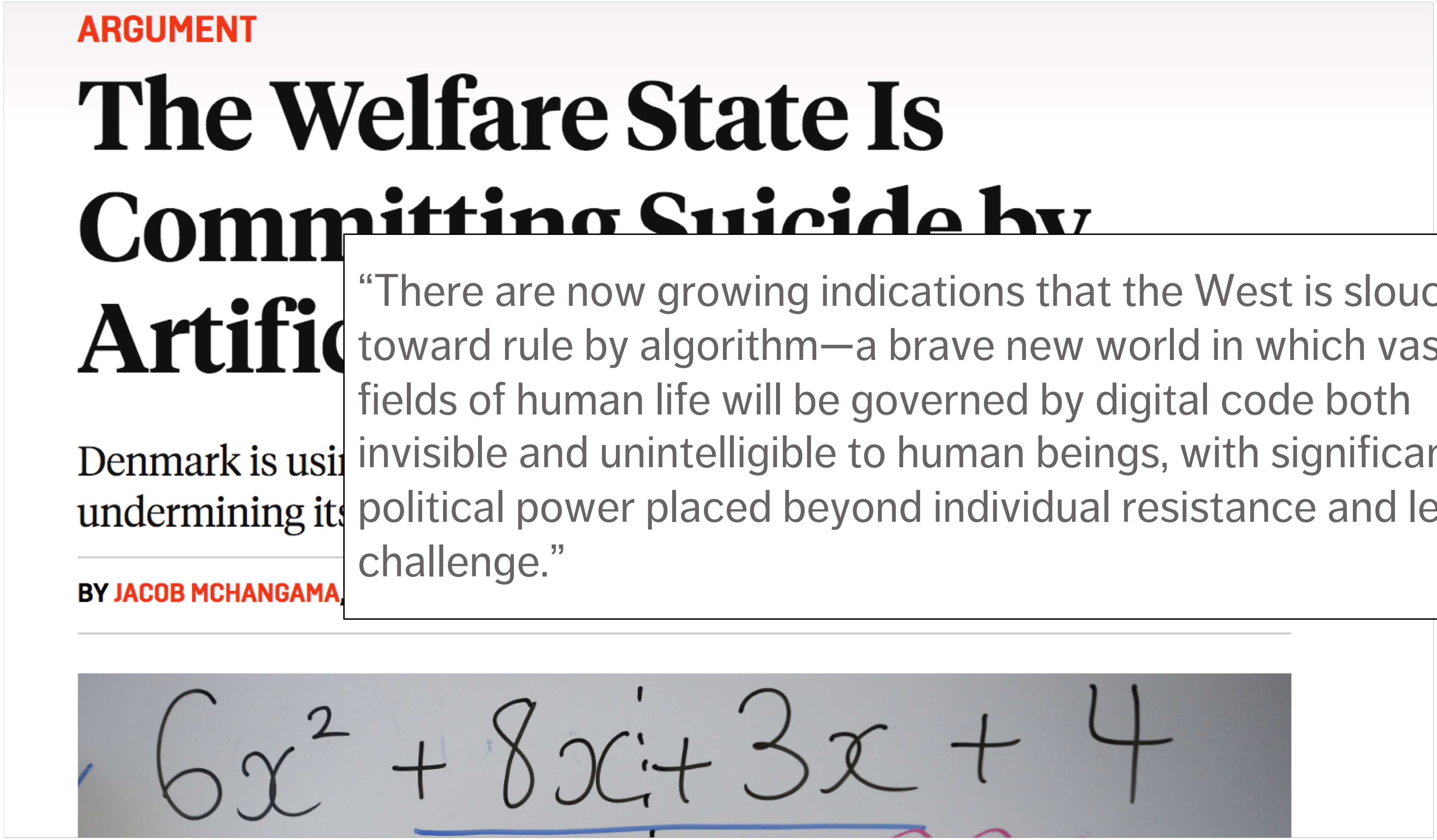


ARGUMENT

The Welfare State Is Committing Suicide by Artificial Intelligence

Denmark is using algorithms to deliver benefits to citizens—and undermining its own democracy in the process.

BY JACOB MCHANGAMA, HIN-YAN LIU | DECEMBER 25, 2018, 1:00 AM



ARGUMENT

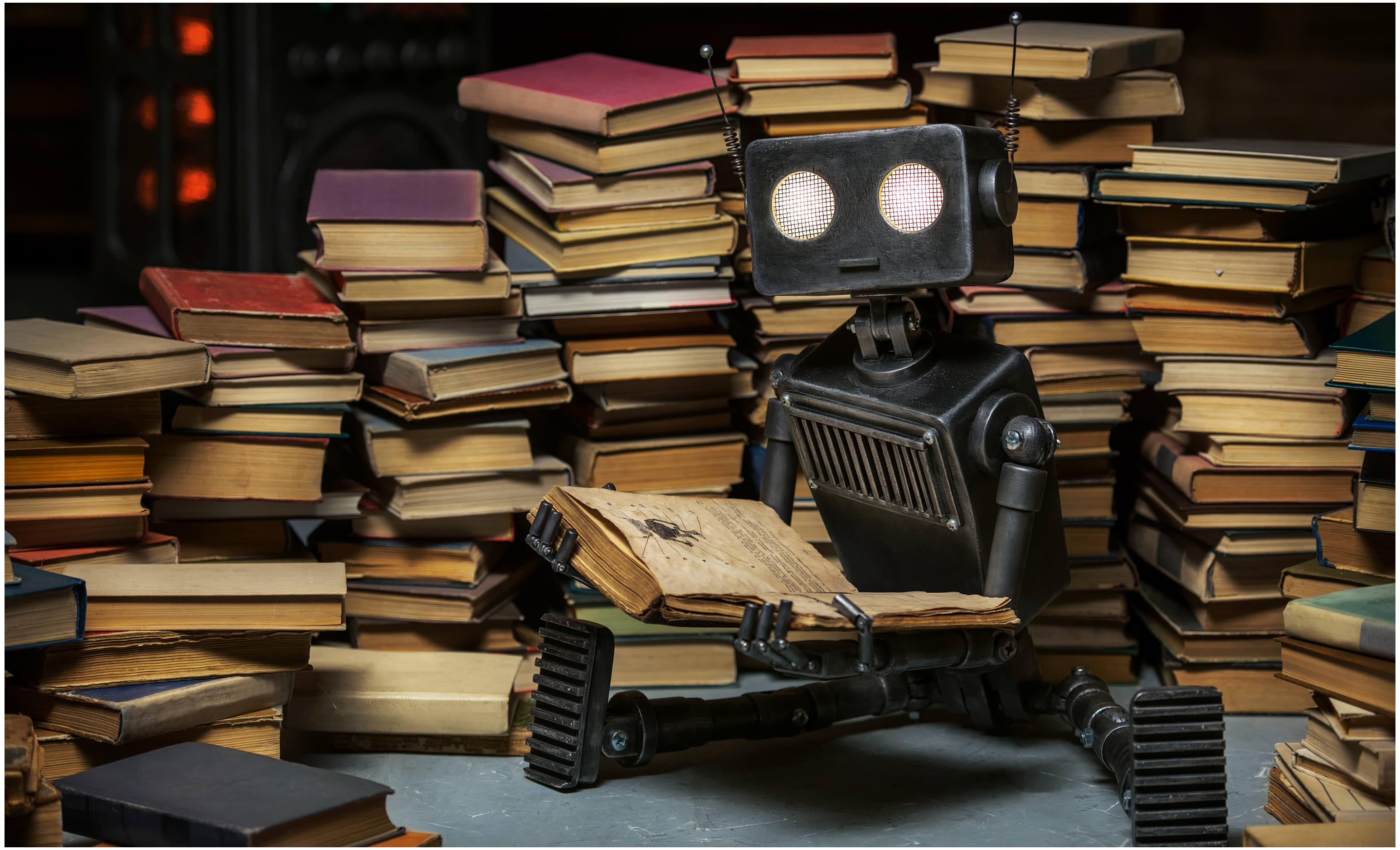
The Welfare State Is Committing Suicide by Artific

Denmark is usi
undermining its

BY JACOB MCHANGAMA,

> “There are now growing indications that the West is slouching toward rule by algorithm—a brave new world in which vast fields of human life will be governed by digital code both invisible and unintelligible to human beings, with significant political power placed beyond individual resistance and legal challenge.”

THE PROMISE AND THE THREAT



A.I. will amount to nothing!

A.I. will take over the world!

As is typical with ‘disruptive technology’, reality is hard to predict!

THE NEED FOR ETHICS

Formerly: "**Wild West**" mentality to data collection (and use). Whatever wasn't technologically forbidden was allowed.

Now: professional codes of conduct are being devised for data scientists (outline responsible ways to practice data science).

Additional responsibility for data scientists; but also **protection** against being hired to carry out questionable analyses.

Does your organization have a code of ethics for its data scientists? For its employees?

WHAT ARE ETHICS?

Broadly speaking, ethics refers to the **study** and **definition** of **right and wrong conducts:**

- “not [...] social convention, religious beliefs, or laws”. (R.W. Paul, L. Elder)

Influential *Western* ethical theories:

- Kant's **golden rule** (do onto others...), **consequentialism** (the ends justify the means), **utilitarianism** (act in order to maximize positive effect), etc.

Other influential ethical theories:

- **Confucianism**, **Taoism**, **Buddhism** (?), **OCAP**® principles, **Ubuntu**, etc.

[University of Virginia, Centre for Big Data Ethics, Law, and Policy]

ETHICS IN THE DATA CONTEXT

Data ethics questions:

- **Who**, if anyone, owns data?
- Are there **limits** to how data can be used?
- Are there **value-biases** built into certain analytics?
- Are there categories that should **not** be used in analyzing personal data?
- Should some data be **publicly available** to **all** researchers?

Analytically, the **general** is preferred to the **anecdotal** – decisions made on the basis of machine learning and A.I. (security, financial, marketing, etc.) may affect real beings in **unpredictable ways**.

BEST PRACTICES

“Do No Harm”: data collected from an individual **should not be used to harm** the individual.

Informed Consent:

- Individuals must **agree to the collection and use** of their data
- Individuals must have a **real understanding of what they are consenting to**, and of **possible consequences** for them and others

Respect “Privacy”: excessively hard to maintain in the age of constant trawling of the Internet for personal data.

BEST PRACTICES

Keep Data Public: data should be kept **public** (all? most? any?).

Opt-In/Opt-Out: Informed consent requires the ability to **opt out**.

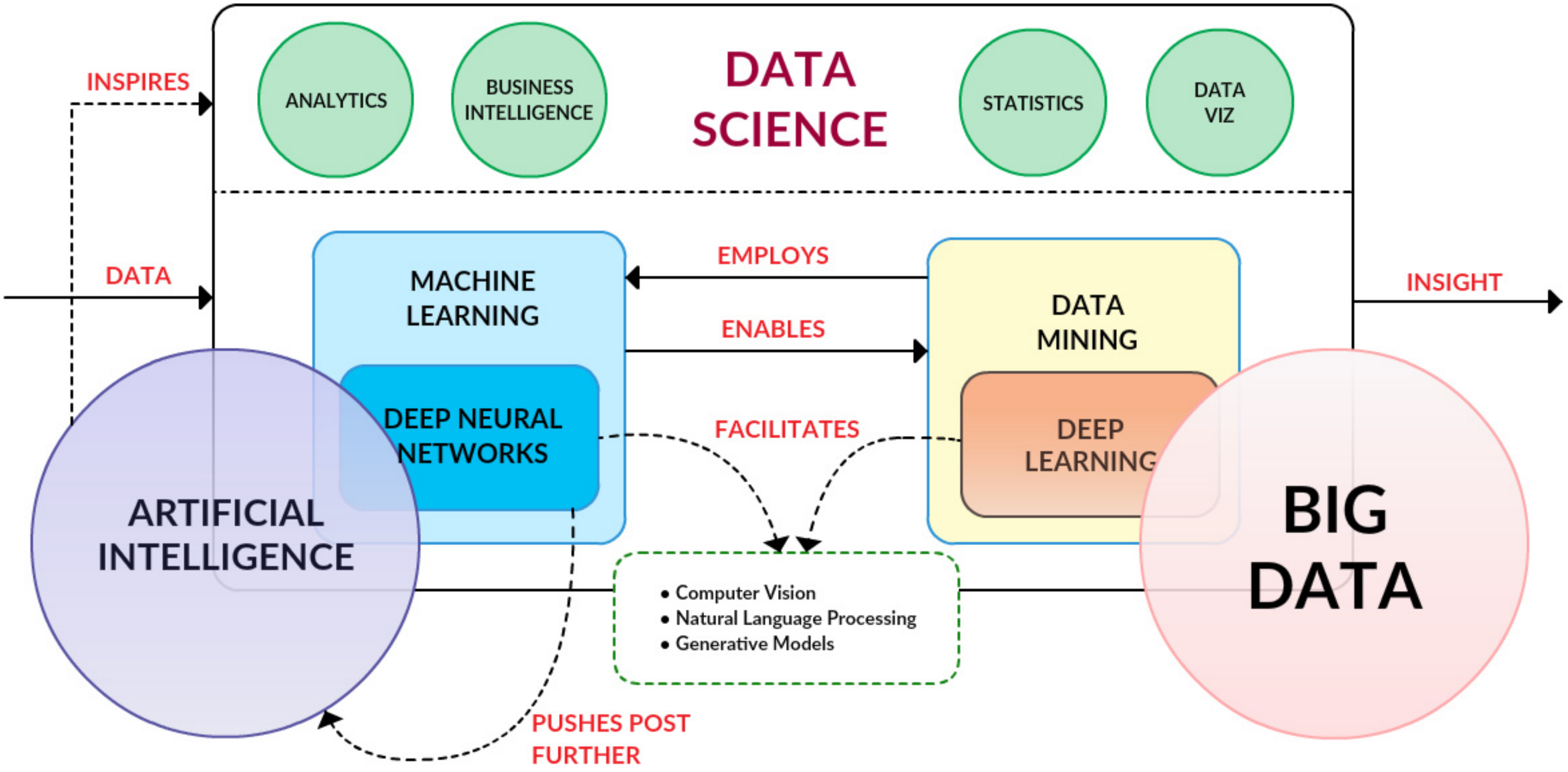
Anonymize Data: removal of id fields from data prior to analysis.

“Let the Data Speak”:

- no cherry picking
- importance of validation (more on this later)
- correlation and causation (more on this later, too)
- repeatability

ARTIFICIAL INTELLIGENCE

[Matthew Mayo, KDNuggets.com]



INSPIRES

DATA

INSIGHT

ANALYTICS

BUSINESS INTELLIGENCE

DATA SCIENCE

STATISTICS

DATA VIZ

MACHINE LEARNING

DEEP NEURAL NETWORKS

EMPLOYS

ENABLES

FACILITATES

DATA MINING

DEEP LEARNING

ARTIFICIAL INTELLIGENCE

BIG DATA

- Computer Vision
- Natural Language Processing
- Generative Models

PUSHES POST FURTHER

[D. Hoftsadter, *Gödel, Escher, Bach*, 1979]

WHAT IS ARTIFICIAL INTELLIGENCE (AI)?

What are the **essential qualities and skills** of an intelligence?

- provides flexible responses in various scenarios
- takes advantage of lucky circumstances
- makes sense out of contradictory messages
- recognizes the relative importance of a situation’s elements
- finds similarities between different situations
- draws distinctions between similar situations
- comes up with new ideas from scratch or by re-arranging previous known concepts

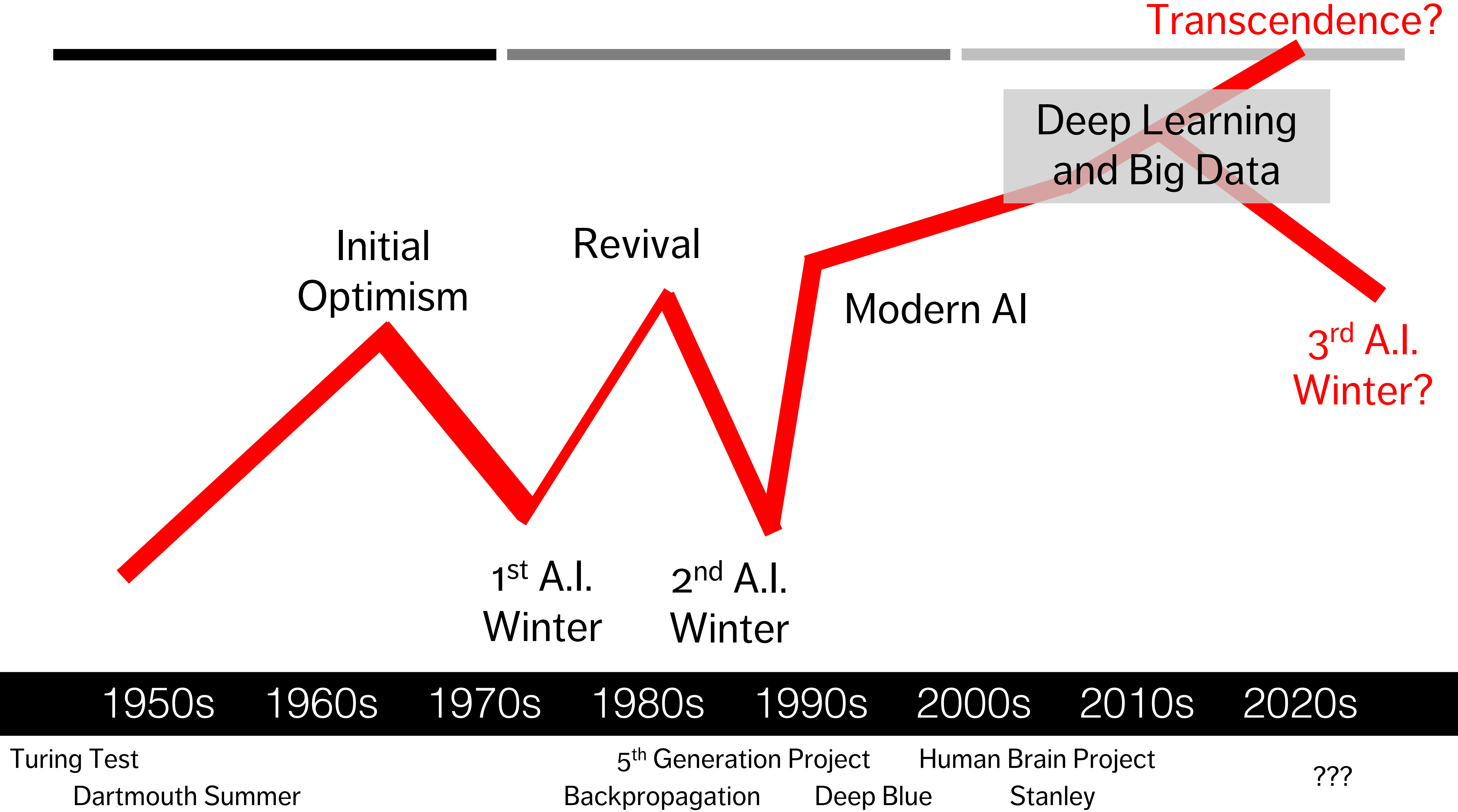
WHAT IS ARTIFICIAL INTELLIGENCE (AI)?

AI research is defined as the study of **intelligent agents**: any device that perceives its environment and takes actions that maximize its chance of success at some goal.

[Artificial Intelligence, *Wikipedia*]

Examples

- *Expert Systems*: TurboTax, WebMD, technical support, insurance claim processing, air traffic control, etc.
- *Decision-Making*: Deep Blue, auto-pilot systems, ”smart” meters, etc.
- *Natural Language Processing*: machine translation, Siri, named-entity recognition, etc.
- *Recommenders*: Google, Expedia, Facebook, LinkedIn, Netflix, Amazon, etc.
- *Content generators*: music composer, novel writer, animation creator, etc.
- *Classifiers*: facial recognition, object identification, fraud detection, etc.



Transcendence?

Deep Learning and Big Data

Initial Optimism

Revival

Modern AI

1st A.I. Winter

2nd A.I. Winter

3rd A.I. Winter?

1950s 1960s 1970s 1980s 1990s 2000s 2010s 2020s

Turing Test

Dartmouth Summer

5th Generation Project

Backpropagation

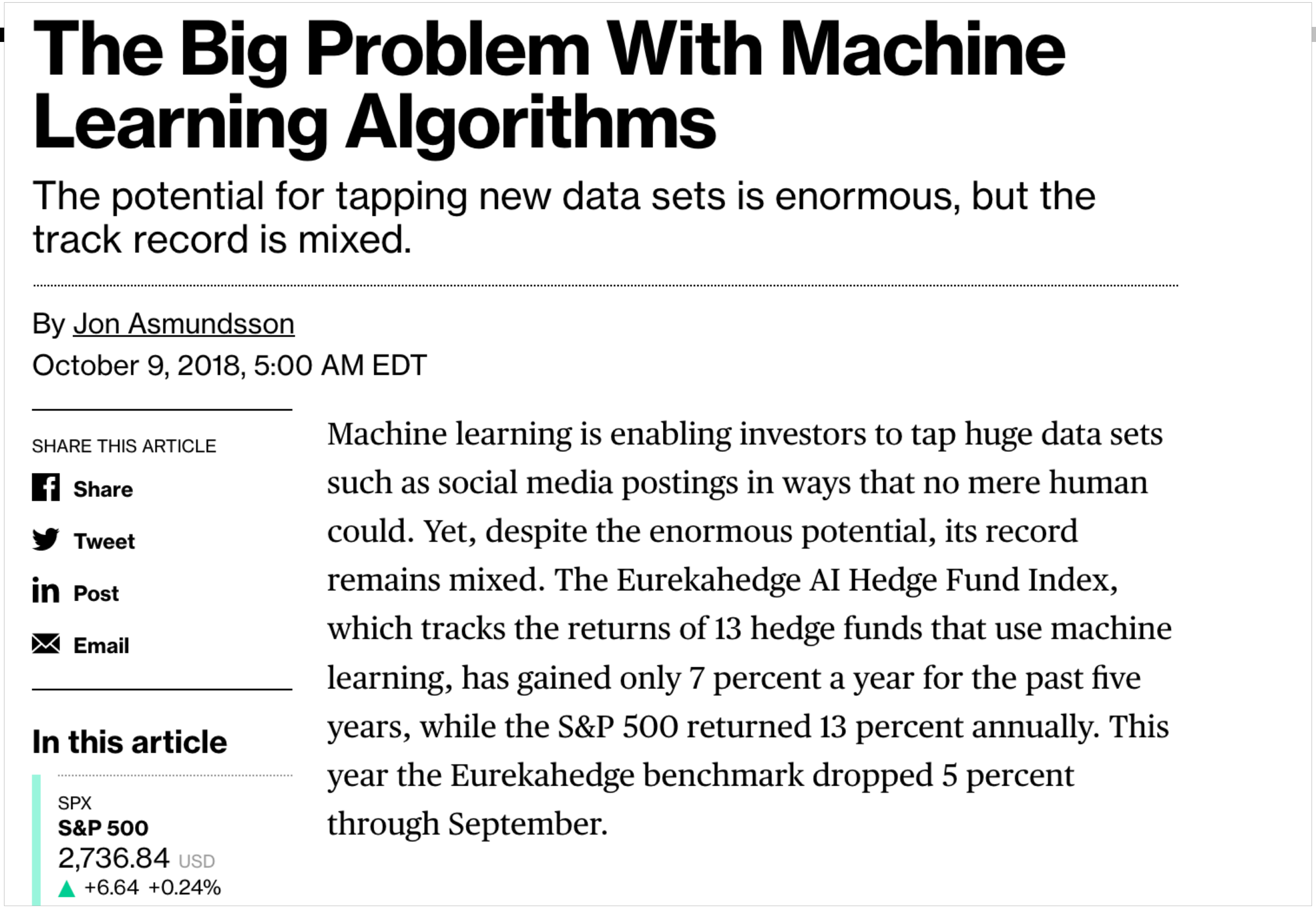
Deep Blue

Human Brain Project

Stanley

???

A.I. ETHICS



The Big Problem With Machine Learning Algorithms

The potential for tapping new data sets is enormous, but the track record is mixed.

By Jon Asmundsson
October 9, 2018, 5:00 AM EDT

SHARE THIS ARTICLE

Share

Tweet

Post

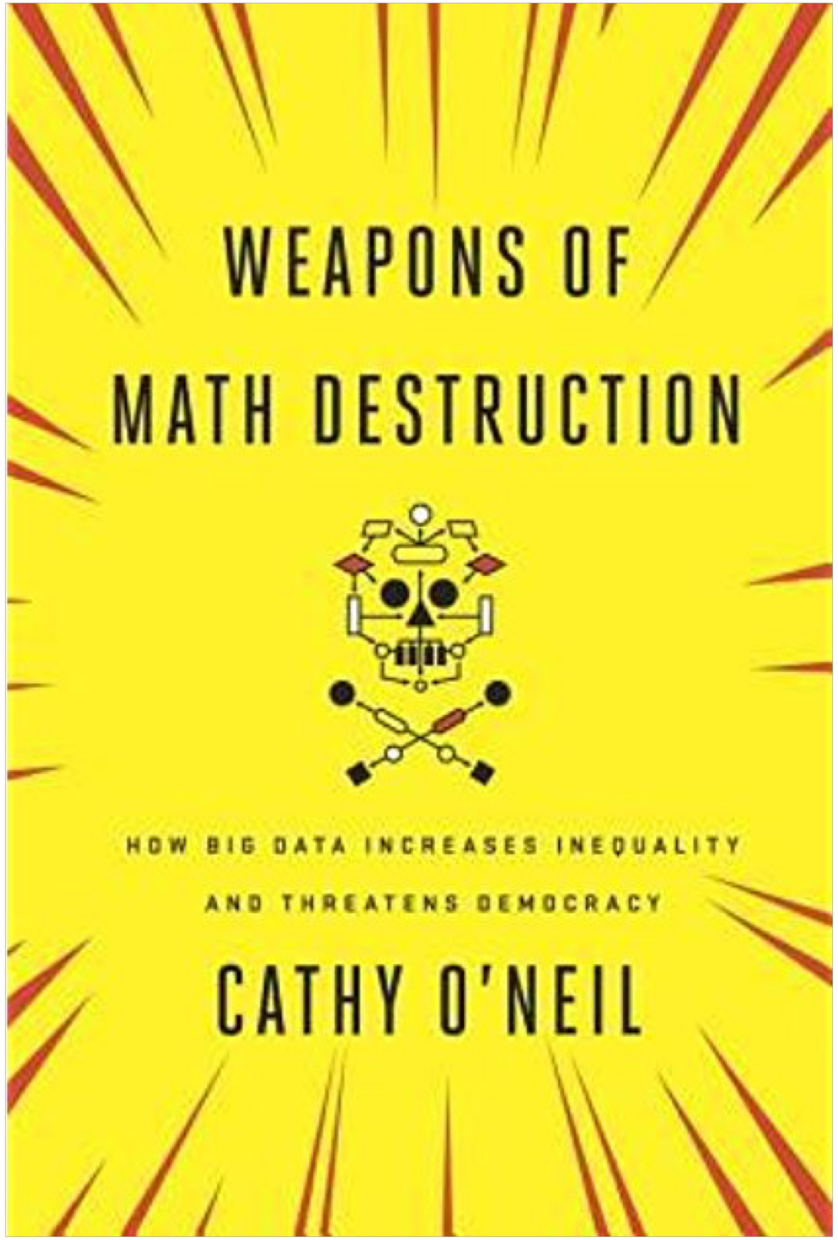
Email

In this article

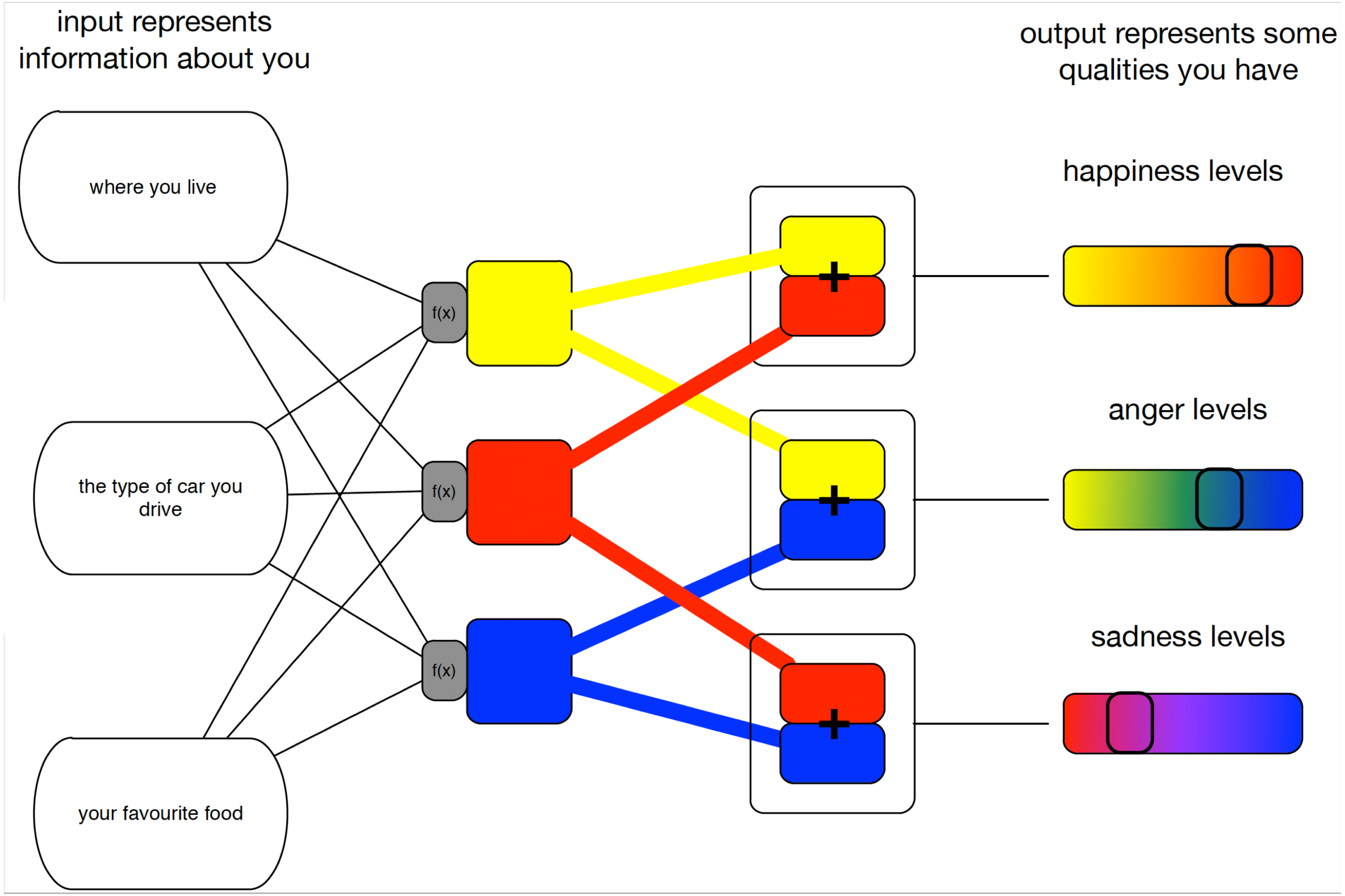
SPX
S&P 500
2,736.84 USD
▲ +6.64 +0.24%

Machine learning is enabling investors to tap huge data sets such as social media postings in ways that no mere human could. Yet, despite the enormous potential, its record remains mixed. The Eurekahedge AI Hedge Fund Index, which tracks the returns of 13 hedge funds that use machine learning, has gained only 7 percent a year for the past five years, while the S&P 500 returned 13 percent annually. This year the Eurekahedge benchmark dropped 5 percent through September.

In her book on the subject of data power, Dr. Cathy O’Neil has a number of cautionary examples and tales.



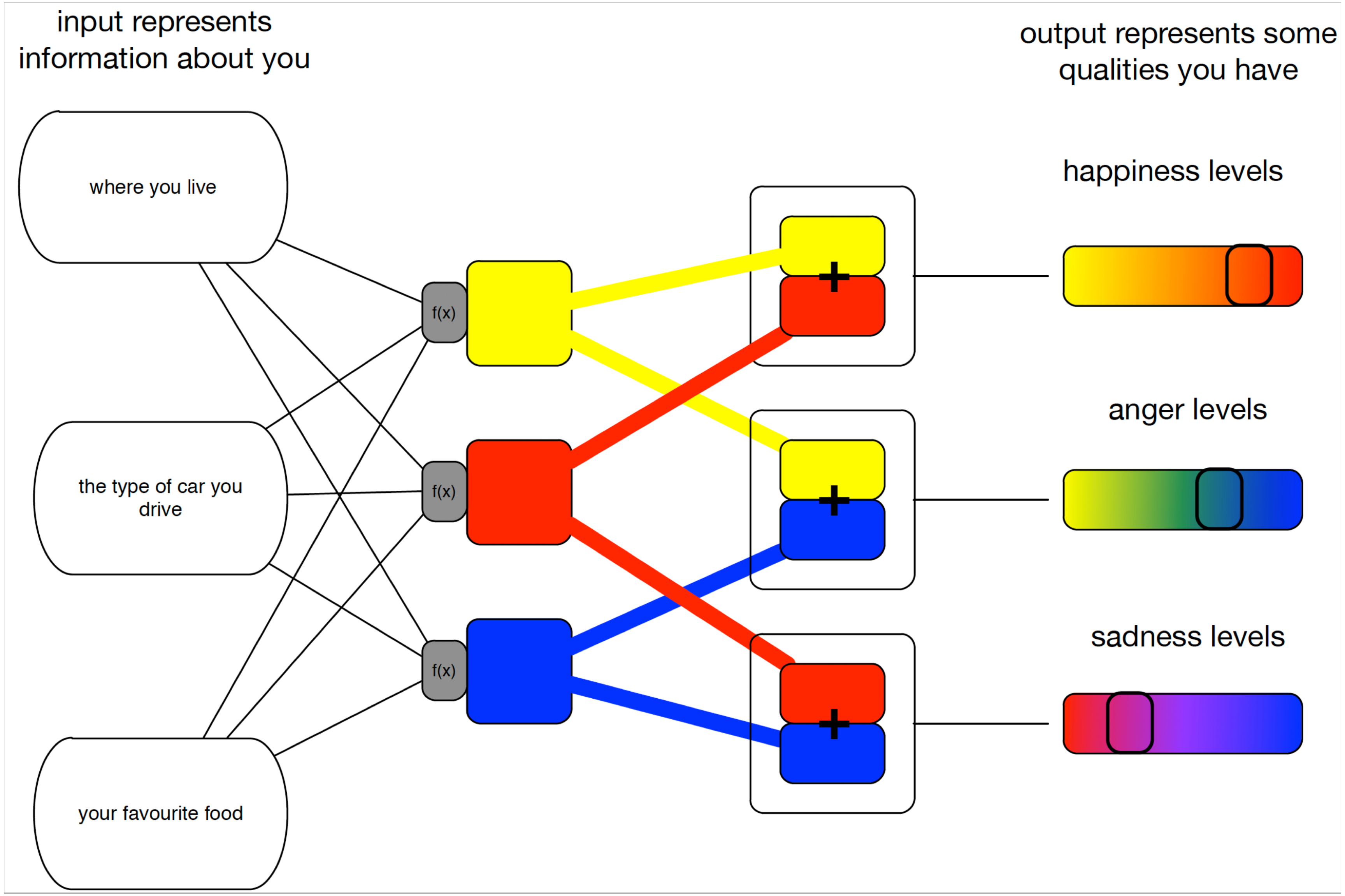
Predicting your emotional qualities based on some information about you:


input represents information about you
output represents some qualities you have
where you live
the type of car you drive
your favourite food
f(x)
f(x)
f(x)
+
+
+
happiness levels
anger levels
sadness levels

How do we get the training data for this?

Self-reports of people in your organization?

Will this data be accurate?

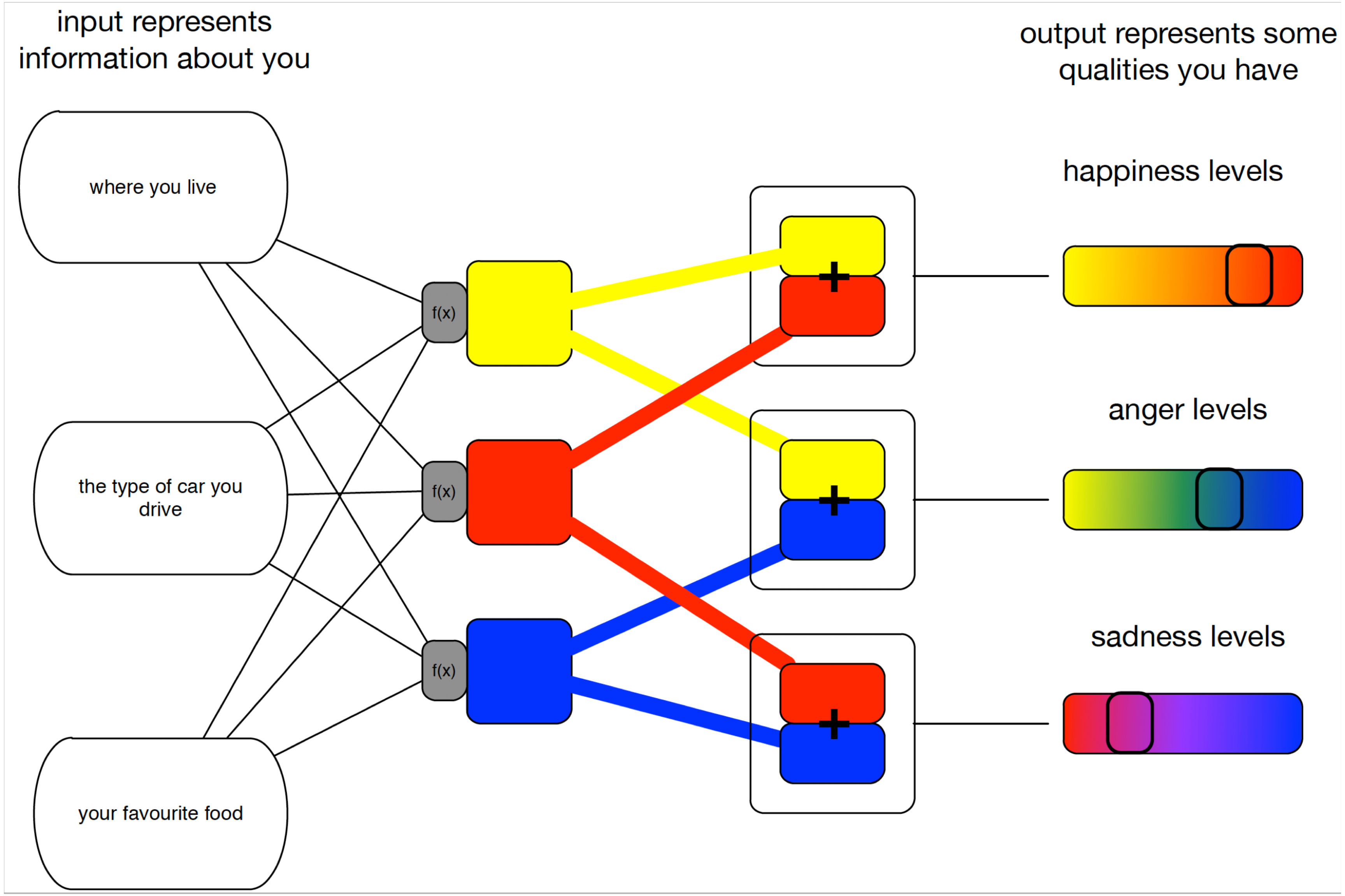

input represents information about you
output represents some qualities you have
where you live
the type of car you drive
your favourite food
f(x)
f(x)
f(x)
+
+
+
happiness levels
anger levels
sadness levels

How precise will the results be?

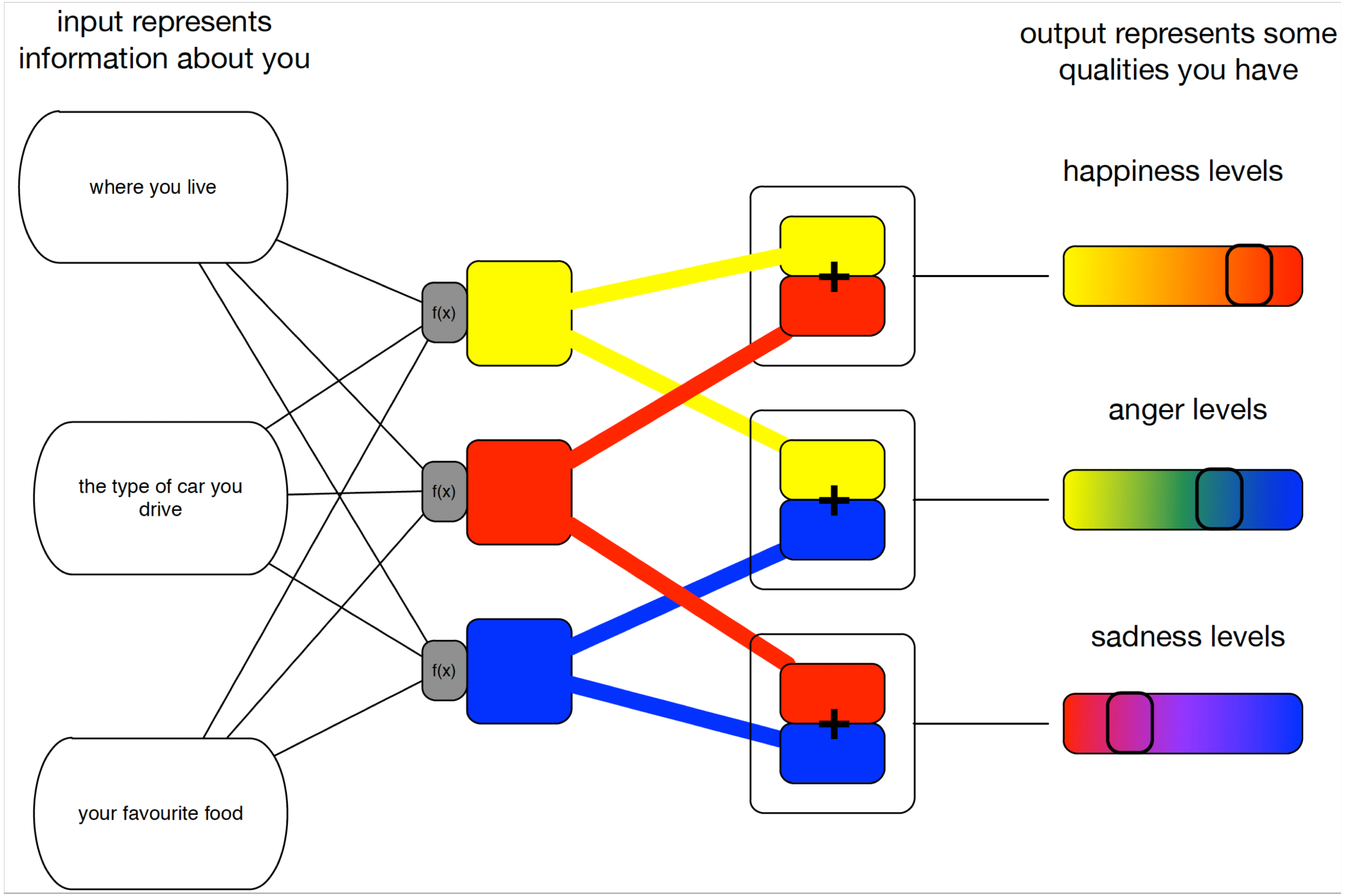
How often will it say my happiness levels are high when they are in the middle?

Does this matter?

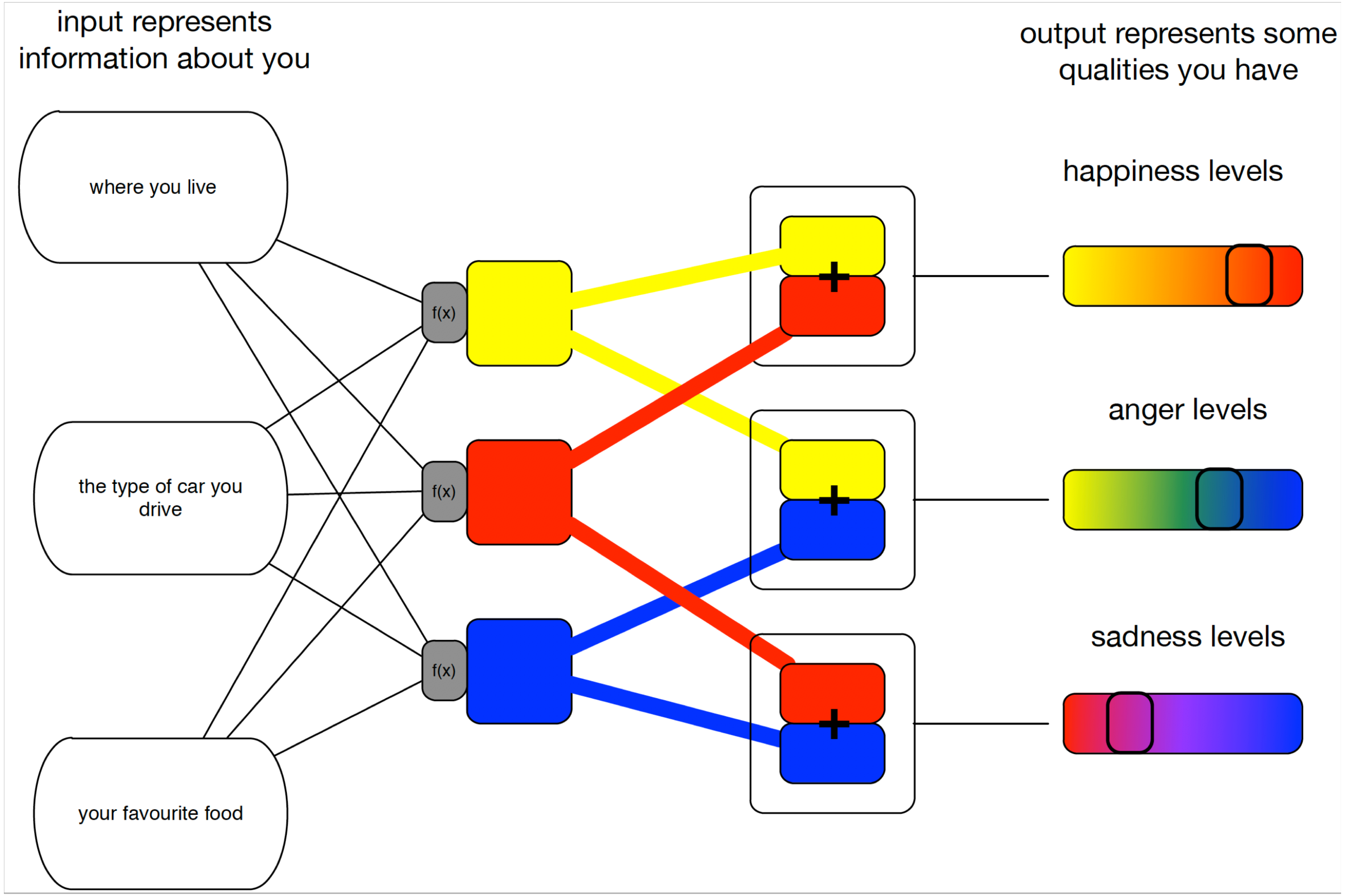
Will the results have the same level of precision for all subgroups?


input represents information about you
output represents some qualities you have
where you live
the type of car you drive
your favourite food
f(x)
f(x)
f(x)
+
+
+
happiness levels
anger levels
sadness levels

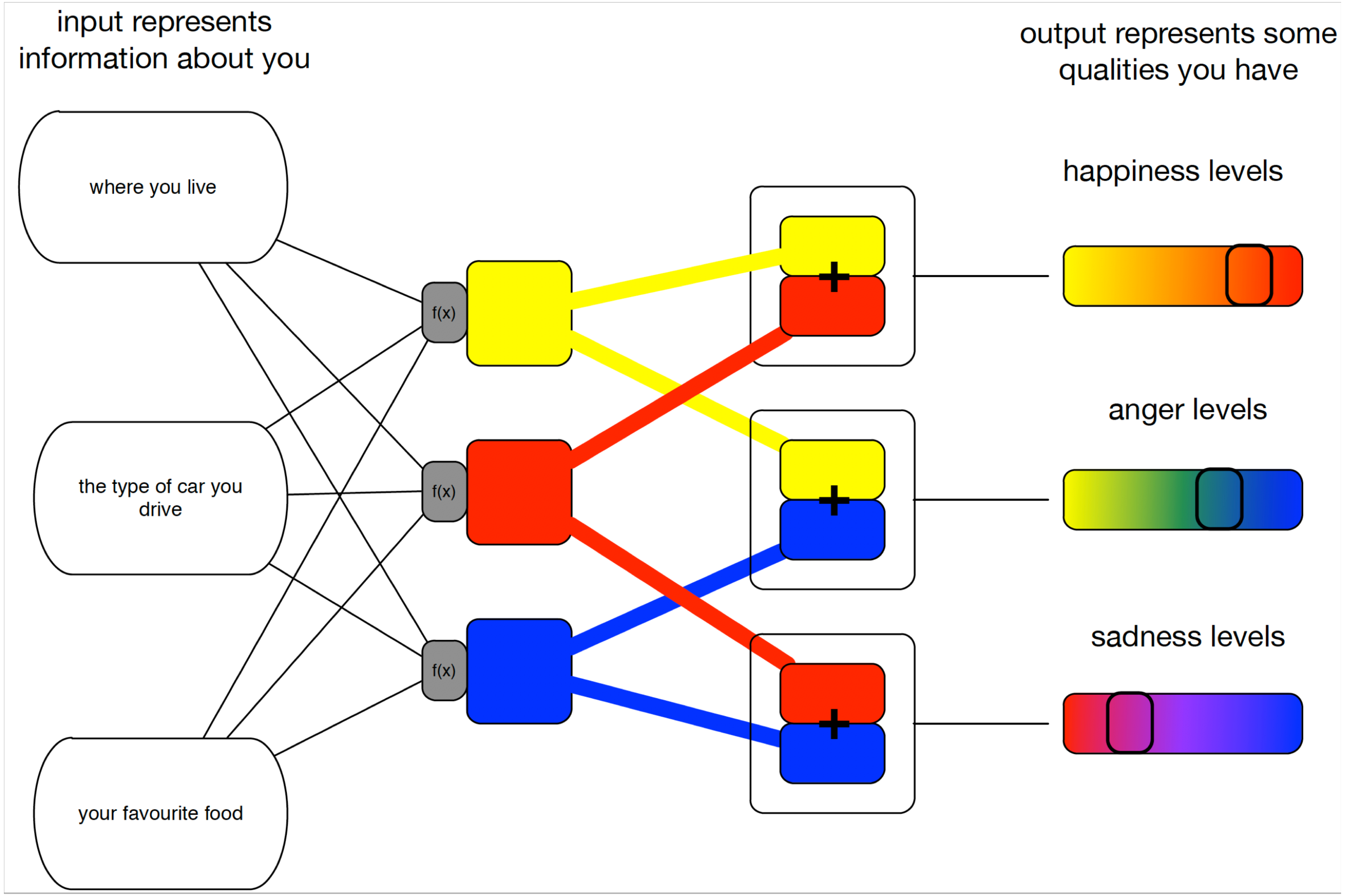
What happens if we don’t know what kind of car you drive?


input represents information about you
output represents some qualities you have
where you live
the type of car you drive
your favourite food
f(x)
f(x)
f(x)
+
+
+
happiness levels
anger levels
sadness levels

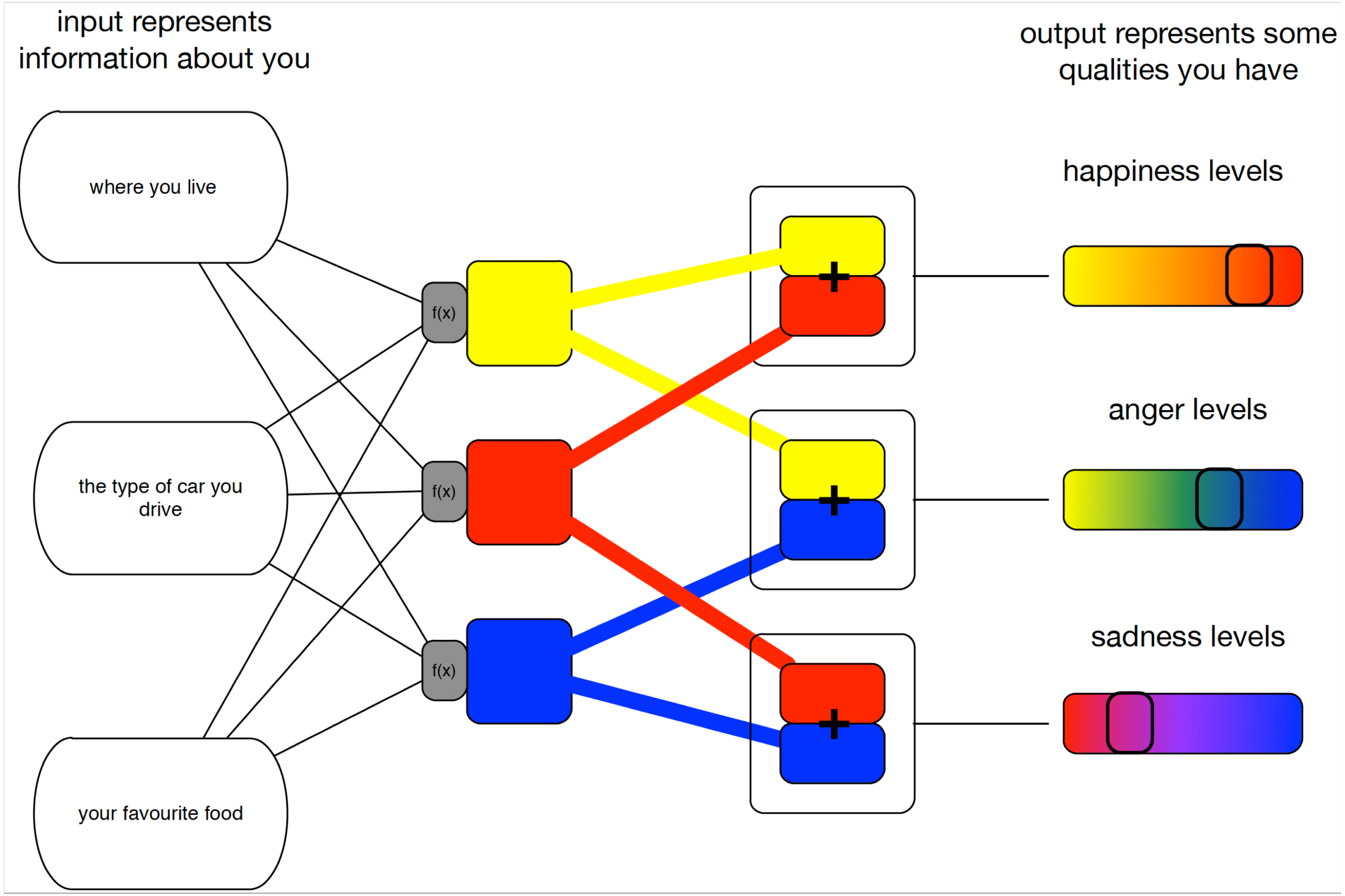
What happens if we mix up the order of the inputs by mistake?


input represents information about you
output represents some qualities you have
where you live
the type of car you drive
your favourite food
f(x)
f(x)
f(x)
+
+
+
happiness levels
anger levels
sadness levels

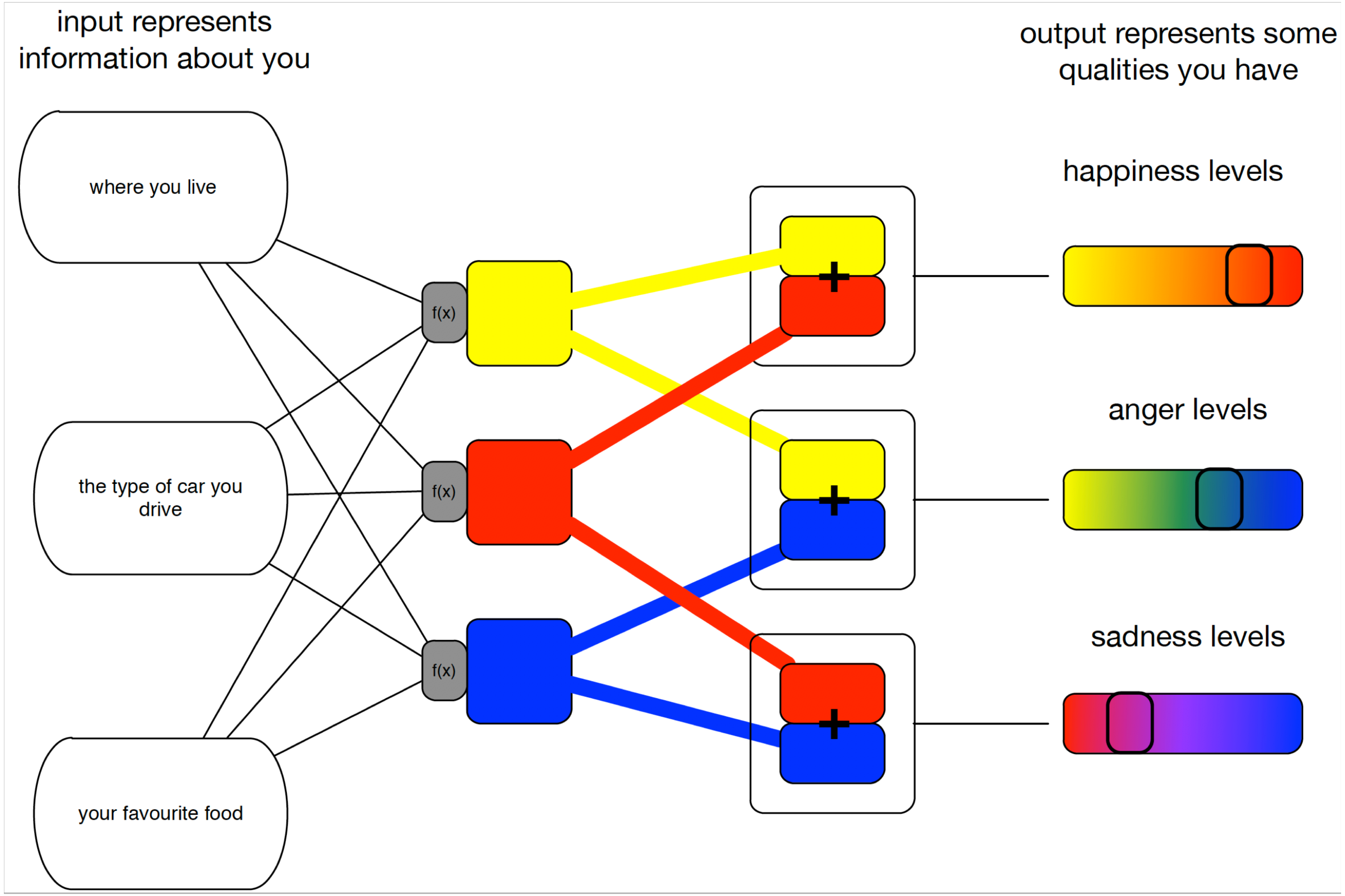
What happens if the information about where you live isn't very precise?


input represents information about you
where you live
the type of car you drive
your favourite food
f(x)
f(x)
f(x)
+
+
+
output represents some qualities you have
happiness levels
anger levels
sadness levels

If you made decisions about people based on this output (e.g. whether or not to hire someone), how would you explain it to them?


input represents information about you
output represents some qualities you have
where you live
the type of car you drive
your favourite food
f(x)
f(x)
f(x)
+
+
+
happiness levels
anger levels
sadness levels

What happens if we try to repurpose this to learn about cats instead of people? We can still feed in input and get output, right?


input represents information about you
output represents some qualities you have
where you live
the type of car you drive
your favourite food
f(x)
f(x)
f(x)
+
+
+
happiness levels
anger levels
sadness levels

MODEL ASSESSMENT AND VALIDITY

Models should be **current**, **useful**, and **valid**.

Data can be used in conjunction with existing models to come to some conclusions, or can be used to update the model itself.

At what point does one determine that the current data model is **out-of-date** or is **not useful anymore**?

Past successes can lead to **reluctance** to re-assess and re-evaluate a model.

TAKING A STEP BACK

GOAL OF ADDING AUTOMATION/AI TO A SYSTEM?

~~To be trendy?~~

To increase or add to the capabilities of the system?

To increase the power of the system?

To make the system better?

Better how? For whom?

HOW CAN AUTOMATION SUCCEED? FAIL?

Succeed:

- Increase efficiency, effectiveness, consistency, reliability, speed

Fail:

- Decrease functionality or capability of the system in some way
- Decrease flexibility and options
- Decrease finesse
- Decrease autonomy of people participating in the system
- Decrease dignity? Increase alienation or objectification?

HOW CAN A.I. SUCCEED? FAIL?

Succeed:

- Come to conclusions/make predictions more accurately than humans
- Come to novel conclusions, discover new knowledge that is beneficial in some way
- Perform functions that humans don't want to carry out (e.g. menial, repetitive functions)
- Increase the extent of our current capabilities, in range and degree (e.g. situational awareness)

Fail:

- Come to erroneous or inaccurate conclusions, which are then acted upon
- Decrease the ability to be transparent or trustworthy; decrease flexibility
- Increase the potential for abuse? Of all participants? Of vulnerable participants?

RESULTS OF SUCCESS? FAILURE?

Results of success:

- System becomes more capable (efficient, powerful, etc.), fair, consistent, trustworthy
- System increases agency, capabilities and dignity of participants, empowers participants
- System becomes more secure, less vulnerable to exploitation

Results of failure:

- System loses desired functionality, ceases to operate well or to be equitable and fair
- People lose trust or confidence in the system
- People reject, rebel against, defy or refuse to participate in the system
- System becomes (more) vulnerable to exploitation or misuse

DATA: FACTORS LEADING TO SUCCESS

Representative data

Appreciation for the **limitations** of the data obtained, in terms of proper uses and relevant conclusions

Consent respected, including circumstances in which data is re-purposed

Respect for **privacy**

DATA: FACTORS LEADING TO FAILURE

Issues relating to:

- unavailable data, missing data, complex and unstructured data, mis-transformed data, analyst training, unconscious and conscious biases (gender, racial, social and data-driven)

Ignorance or disregard for **ethical use of data** – note this is not a technological issue, rather it is simply being made more relevant due to the power and reach of technology.

ALGORITHMS: FACTORS LEADING TO SUCCESS

Acceptable rate of false positives and false negatives for the CRA and the clients

Explainability (neural nets?), monitoring and algorithmic **renewal/adjustment** standards (new data meets older models)

Model **validity**

Appropriate model choice

ALGORITHMS: FACTORS LEADING TO SUCCESS

Accessibility/transparency (access to the algorithm that is being used)

Finesse

Granularity of results

Correct use and **interpretation** of results

ALGORITHMS: FACTORS LEADING TO FAILURE

Mismatch between algorithm and data

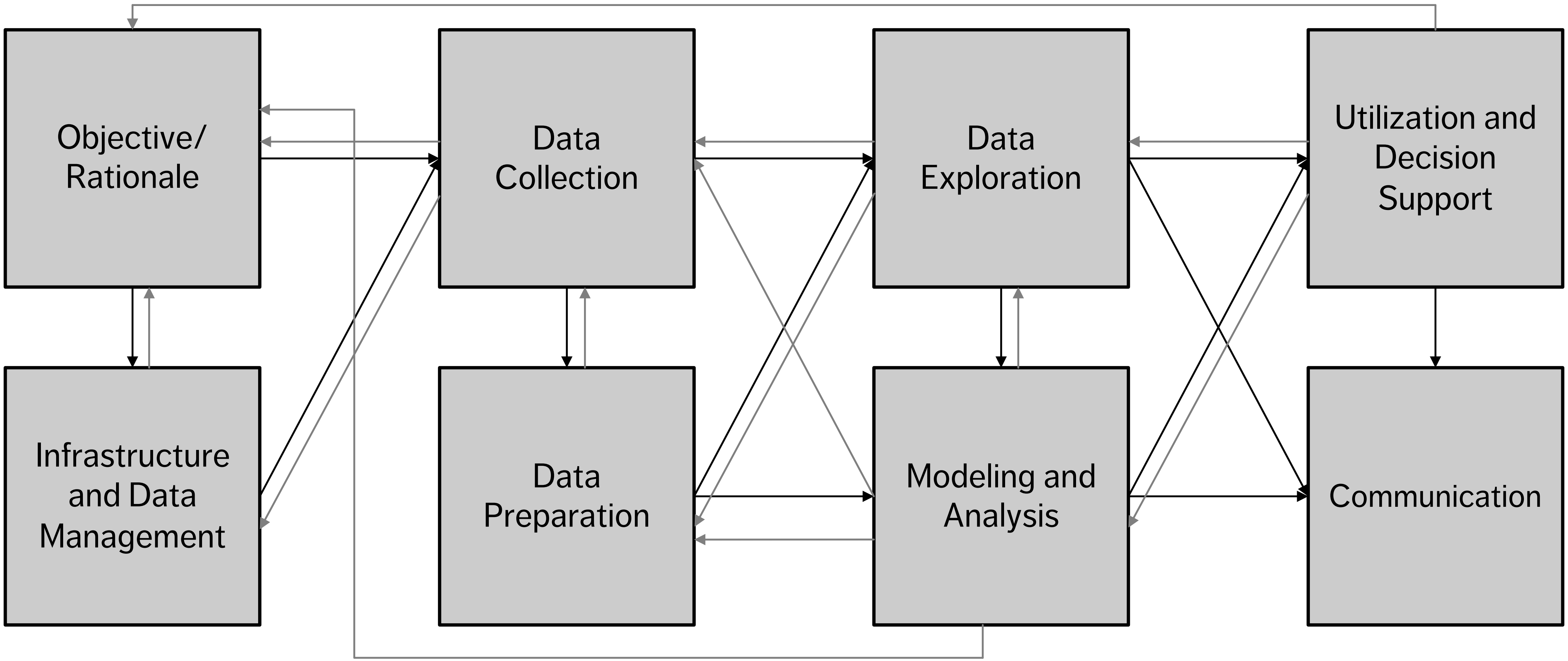
Inadequate data to power algorithm, lead to ‘good’ results

Lack of understanding of what the results of the algorithm mean

THE DATA SCIENCE "WORKFLOW"

- Objective/ Rationale
- Data Collection
- Data Exploration
- Utilization and Decision Support
- Infrastructure and Data Management
- Data Preparation
- Modeling and Analysis
- Communication

THE DATA SCIENCE "WORKFLOW"



THE DATA ANALYSIS PROCESS

A **large number of analytical models** have to be generated before a final selection can be made.

Iterative process: feature selection and data reduction may require numerous visits to domain experts before models start yielding promising results.

Domain-specific knowledge has to be integrated in the models in order to beat random classifiers and clustering schemes, **on average**.

IMPACT OF AI/AUTOMATION ON GOC AND CRA

COMPLIANCE

Loss of reputation/trust (CRA & GOC) – withholding information in future similar interactions (refusal to grant data sharing consent)

Reduced compliance

Inequitable administrative/program outcomes – original data is incomplete

Financially costly errors in program administration

Negative privacy impact on Canadians

SERVICE

Reduced Agency service ratings

Impact on self-service channels (Digital Government) – decline in trust

Increase in call centre volume (public unlikely to distinguish between front line automation and web-based applications)

APPENDIX: A BRIEF INTRODUCTION TO NEURAL NETWORKS

[Yes, Androids Do Dream of Electric Sheep, The Guardian UK]

NEURAL NETWORKS IN A NUTSHELL

A trained **Artificial Neural Network** (ANN) is a function that maps inputs to outputs:

- receive input(s)
- compute values
- provide output(s)

ANNs use a Swiss-army-knife approach to things (**plenty of options, but it's not always clear which one should be used**).

The user does not need to decide much about the function or know much about the problem space in advance (**quiet model**).

NEURAL NETWORKS IN A NUTSHELL

Algorithms allow ANNs to **learn** (i.e. generate the function and its internal values) **automatically**.

ANNs can be used for:

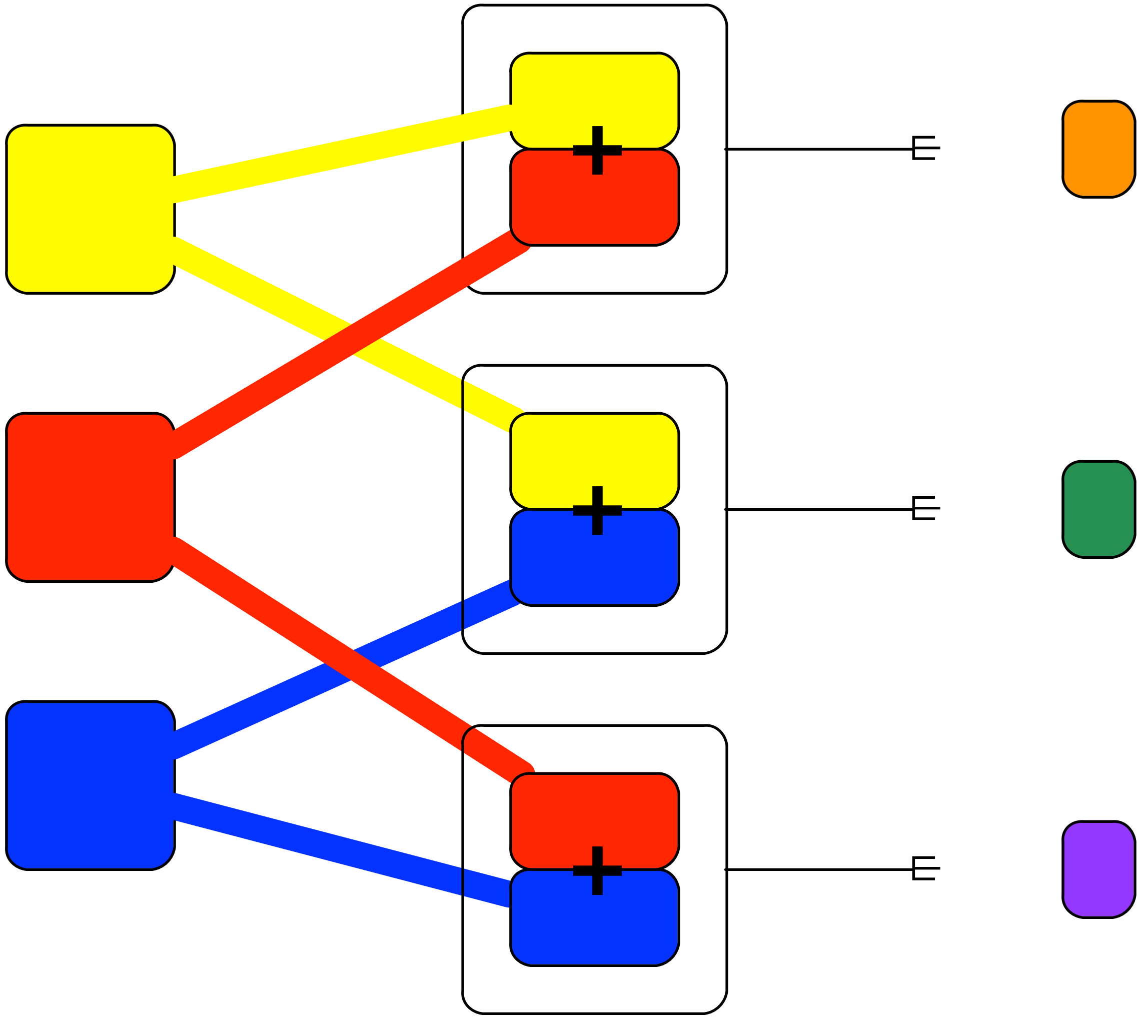
- supervised learning (**multi-layered feedforward neural networks**)
- unsupervised learning (**self-organizing maps**)
- reinforcement learning.

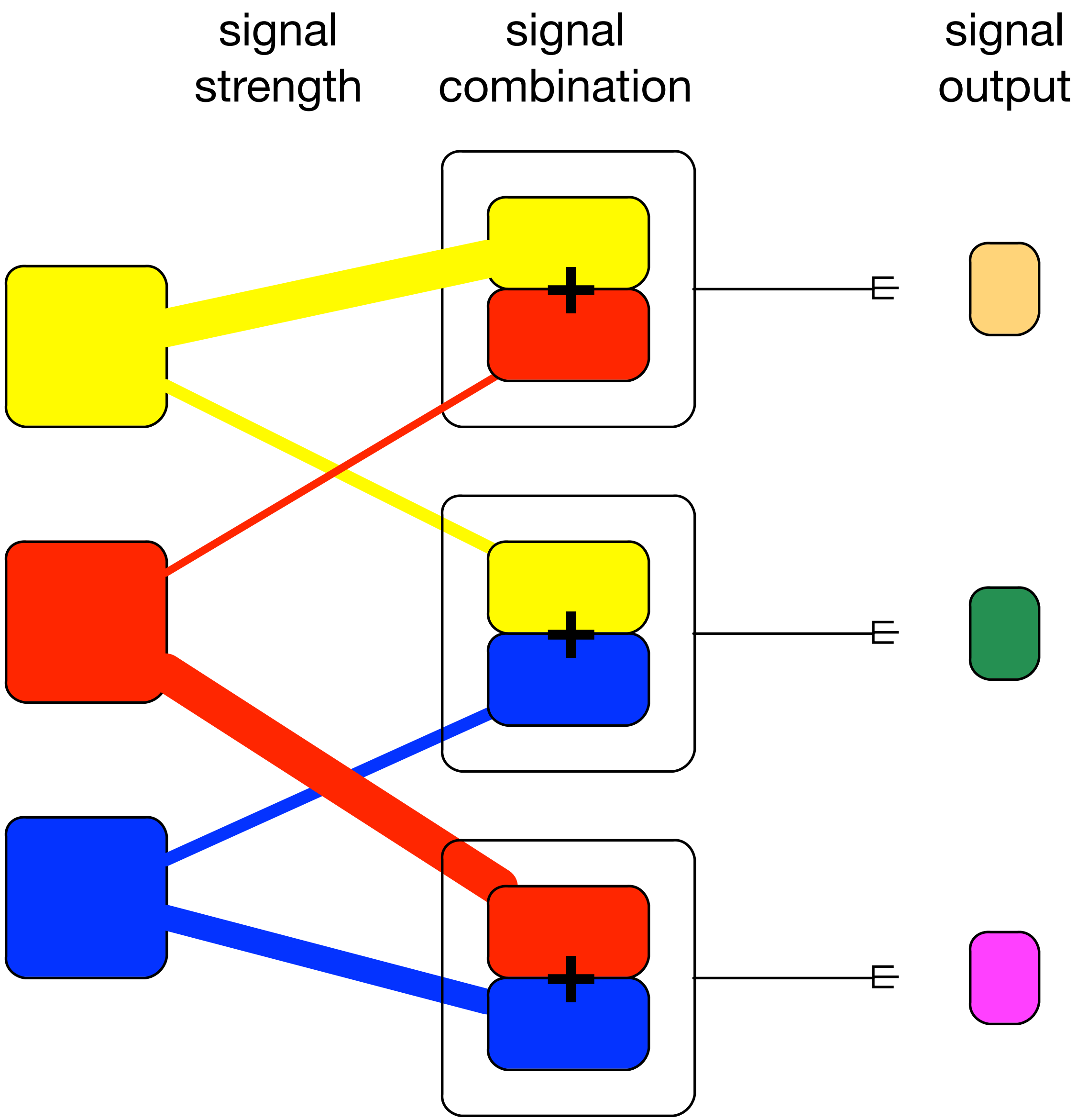
Technically, the only requirement is the ability to minimize a cost function (**optimization**).

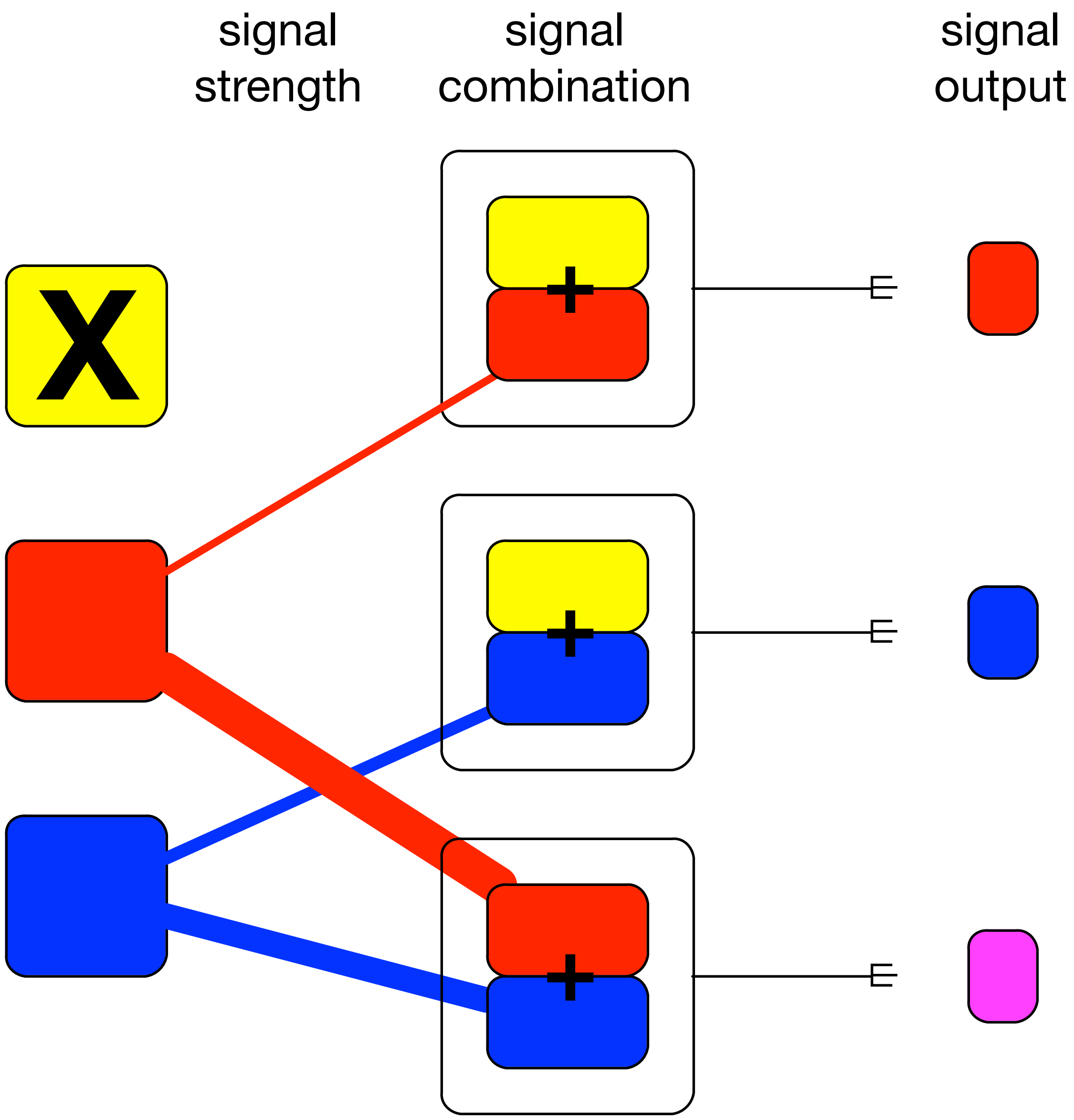
signal strength

signal combination

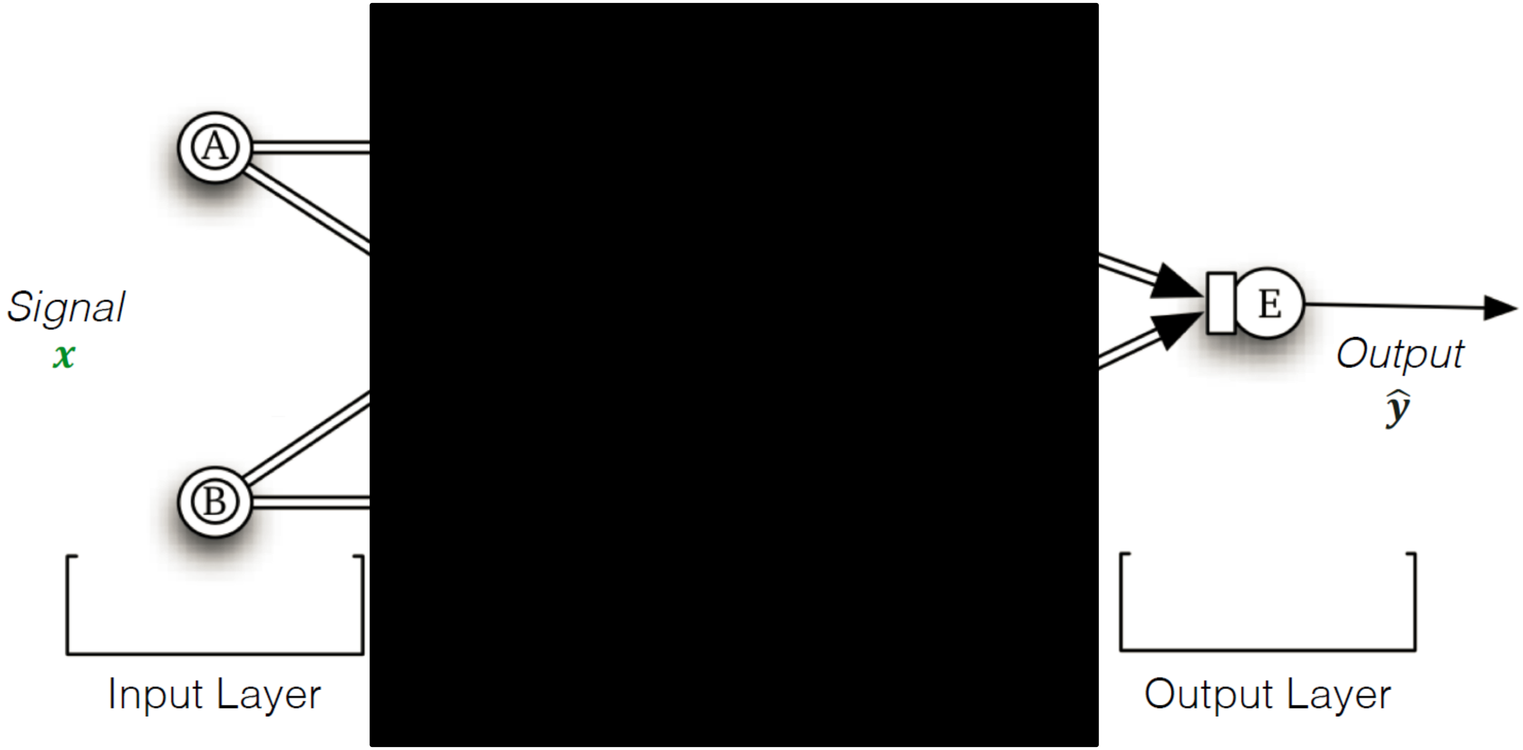
signal output



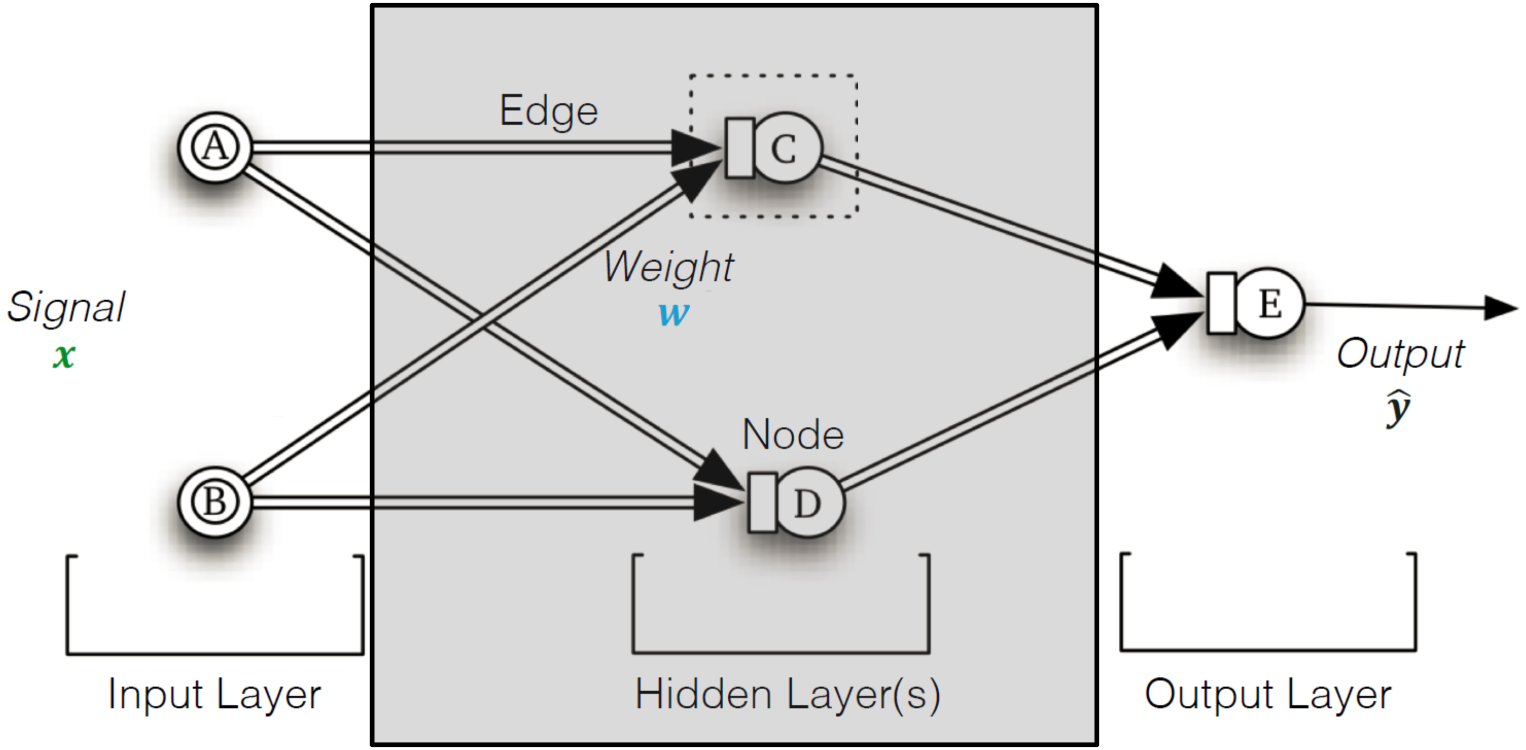

signal strength
signal combination
signal output
+
+
+


signal strength
signal combination
signal output
X
+
+
+

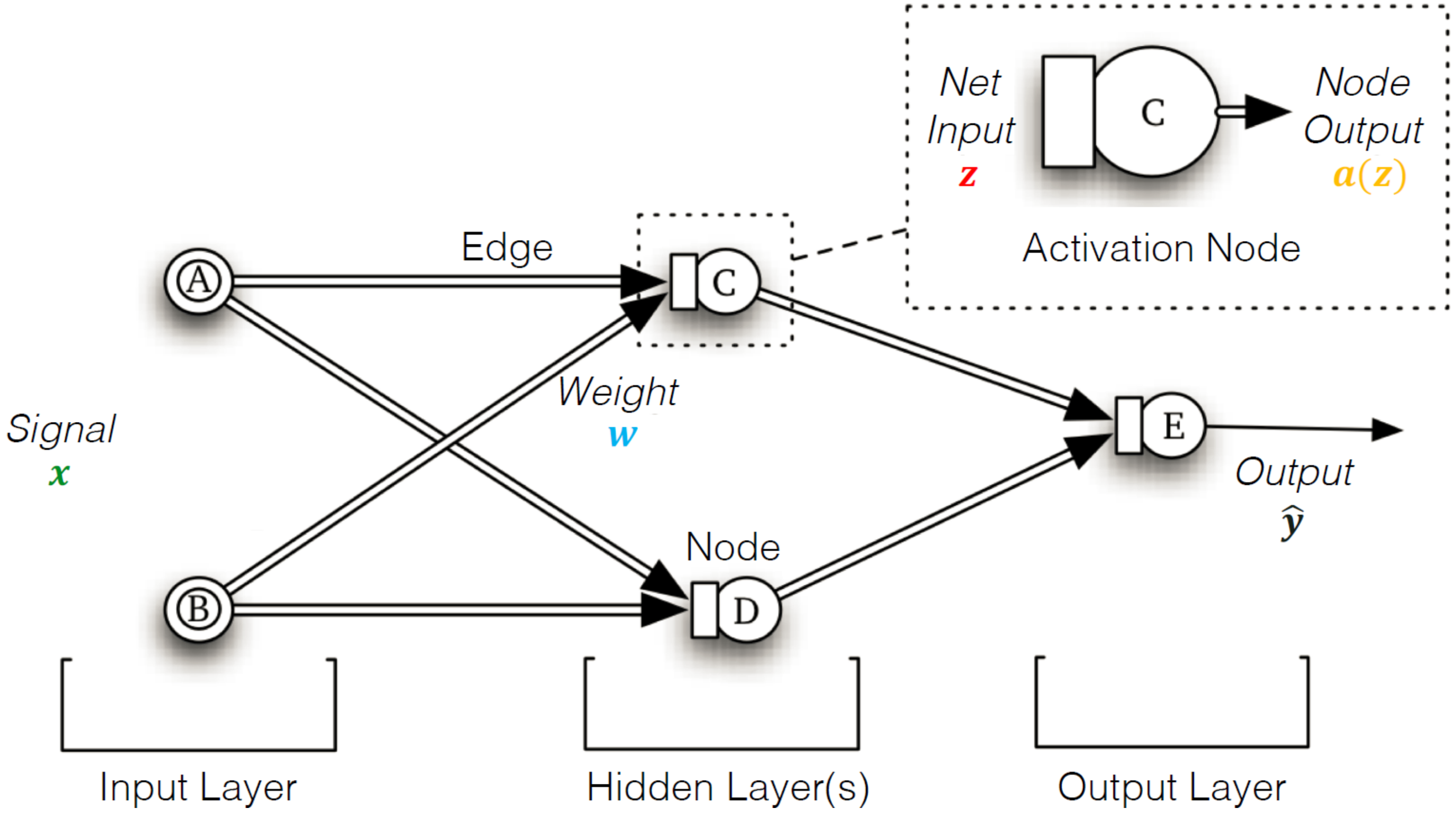
NETWORK TOPOLOGY AND TERMINOLOGY



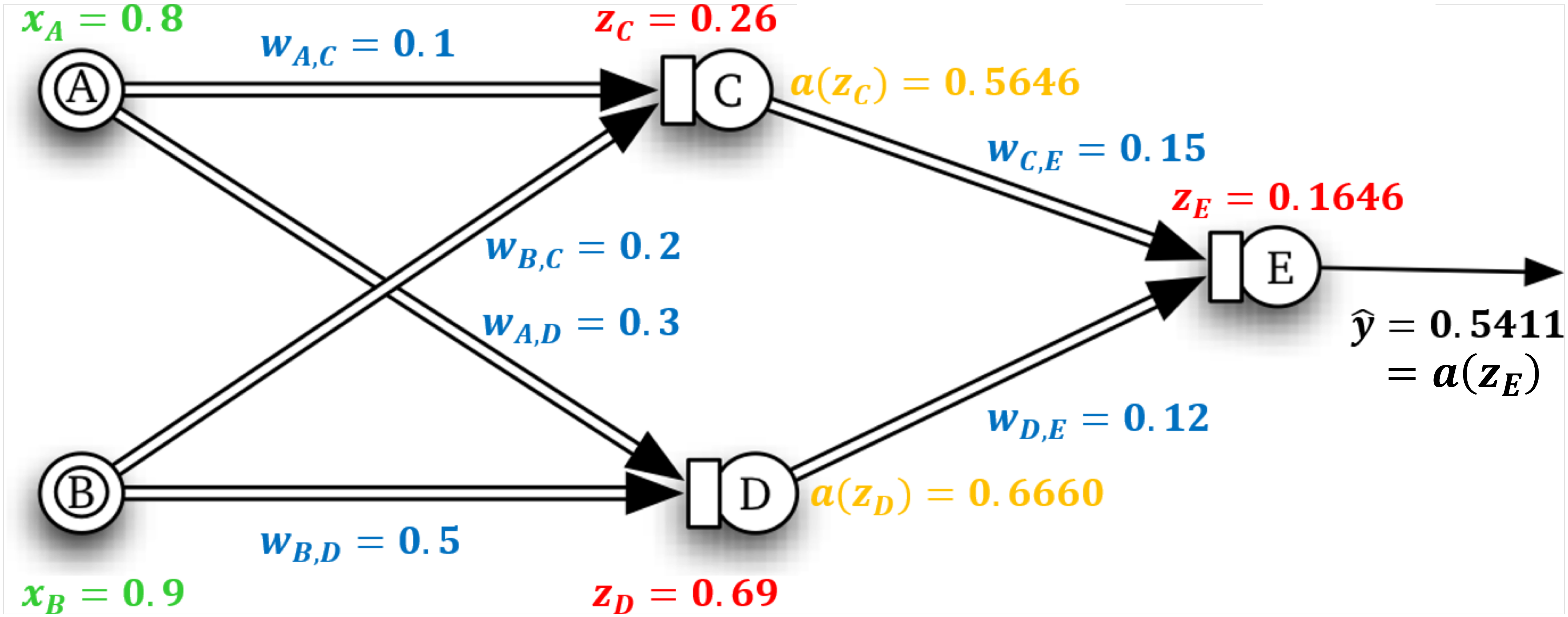
NETWORK TOPOLOGY AND TERMINOLOGY



NETWORK TOPOLOGY AND TERMINOLOGY



FEED FORWARD NETWORK



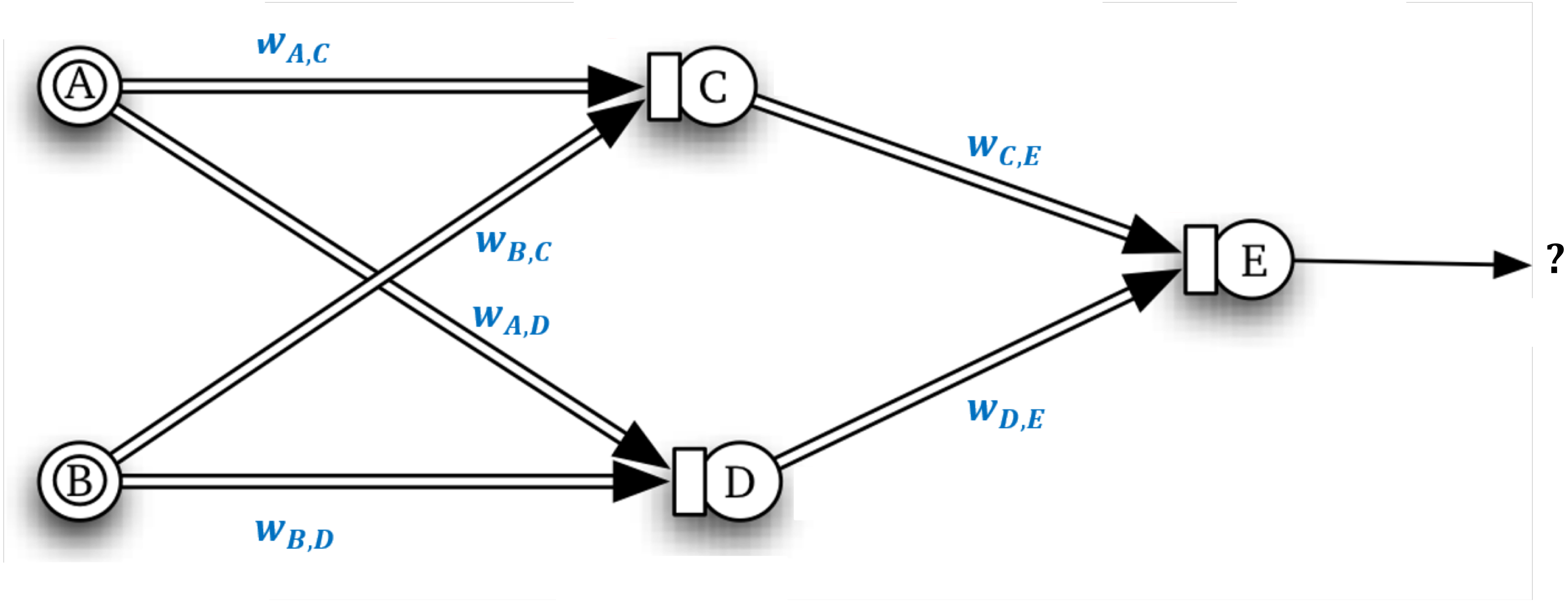
HOW TO TRAIN YOUR NEURAL NETWORK

Given a signal, an ANN can produce an output, as long as the weights are specified.

For **supervised** learning tasks (i.e. when an ANN attempts to emulate the results of training examples), simply picking weights at random is a failing proposition.

Backward propagation is a method to optimize the choice of the weights against an error function $R(\mathbf{W})$.

FEED FORWARD NETWORK



PROS AND CONS

ANNs can be quite **accurate** when making predictions – better than other algorithms with a proper set up.

ANNs often work when other things fail:

- when the relationship between attributes is **complex**
- when there are a lot of dependencies/**nonlinear relationships**
- **messy**, highly connected inputs (images, text and speech)

ANNs are relatively easy to set up (with available packages).

ANNs degrade gracefully (important in robotics).

PROS AND CONS

ANNs are relatively slow (creating and using) and prone to overfitting (may require **large/diverse** training set).

ANNs usually do not provide good interpretation (unlike decision trees or logistic regression, say). Can you live with that?

No algorithms for selecting the optimal network topology.

Even when they do perform better than other options, ANNs may not perform that much better due to **No Free-Lunch Theorems**; and they're susceptible to various forms of **adversarial attacks**.

DEEP LEARNING NETWORKS

Deep Learning networks are simply ANNs with **a large number of hidden layers** and various types of nodes and connections.

Types:

- Convolution Neural Networks: handwritten digit recognition, 99.7% accuracy in 2013; self-driving cars
- Recurrent Neural Networks: natural language processing (speech recognition, machine translation, etc.)
- Autoencoders
- Restricted Boltzmann Machines: BellKor's Pragmatic Chaos, Netflix Prize, 2009

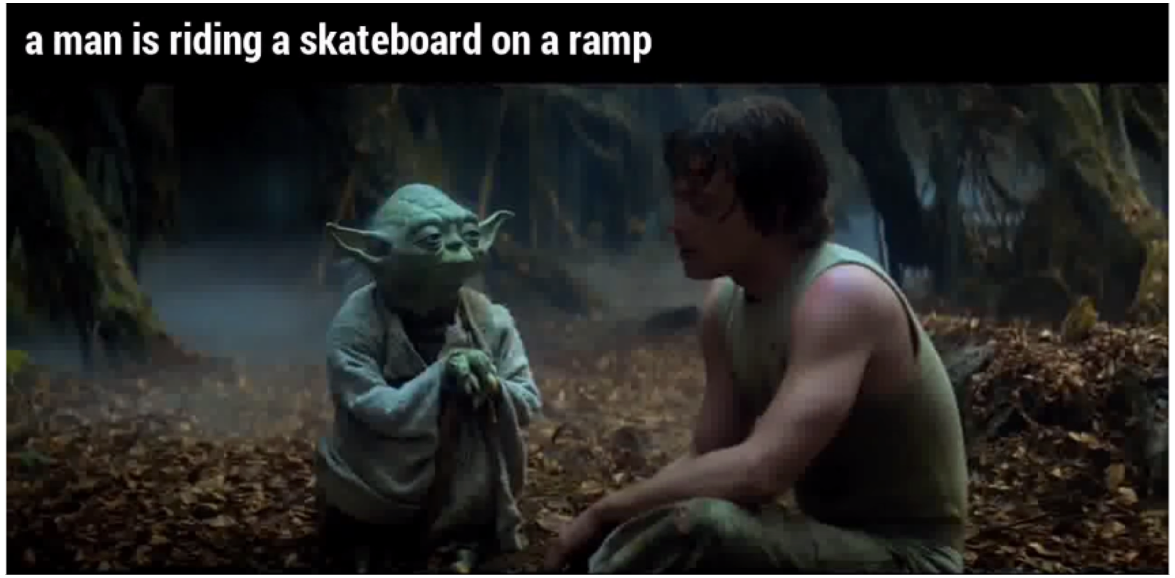
[@samim, medium.com]

CONS (REPRISE)

Require **large**, **diverse**, and **correctly labeled** training sets.

Accurate on average, but they can still be **spectacularly** wrong.

They can be “hacked” (NFLT).


a man is riding a skateboard on a ramp

How do we align autonomous AI goals with human values?

Humans don’t need that much labeled data to make decisions: so **what’s really going on under the hood**? (3rd AI Winter?)

SELECTED REFERENCES

REFERENCES

Mayer-Schönberger, V. and Cukier, K. [2013], *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan/Houghton Mifflin Harcourt.

Mayer-Schönberger, V. [2009], *Delete: The Virtue of Forgetting in the Digital Age*, Princeton University Press.

Data Science Association, *Data Science Code of Professional Conduct*.

Chen, M. [2013], *Is 'Big Data' Actually Reinforcing Social Inequalities?*, The Nation.

Shin, L. [2013], *How the New Field of Data Science is Grappling With Ethics*, SmartPlanet.

Schutt, R. and O'Neill, C. [2013], *Doing Data Science: Straight Talk From the Front Line*, O'Reilly.

O'Neill, C. [2016], *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown.

REFERENCES

Chang, R.M., Kauffman, R.J., Kwon, Y. [2014], *Understanding the paradigm shift to computational social science in the presence of big data*, Decision Support Systems, 63:67–80, Elsevier.

Hurlburt, G.F., Voas, J. [2014], *Big Data, Networked Worlds*, IEEE Computer Society.

Introna, L.D. [2007], *Maintaining the reversibility of foldings: Making the ethics (politics) of information technology visible*, Ethics and Information Technology, 9:11–25, Springer.

Floridi, L. [2011], *The philosophy of information*, Oxford University Press.

Floridi, L. (ed) [2006], *The Cambridge handbook of information and computer ethics*, Cambridge University Press, 2006.

Big Data & Ethics

Mason, H. [2012], What is a Data Scientist?, Forbes.

[Welch Labs]

NEURAL NETWORKS DEMYSTIFIED

1. Data and Architecture – nodes and edges, weights and activation functions
2. Forward Propagation – from the input layer to the output layer…
3. Gradient Descent
4. Backpropagation – … back the other way
5. Numerical Gradient Checking
6. Training – … and back again, and again, and again.
7. Overfitting, Testing, and Regularization