Legal Services National Technology Assistance Project

www.lsntap.org

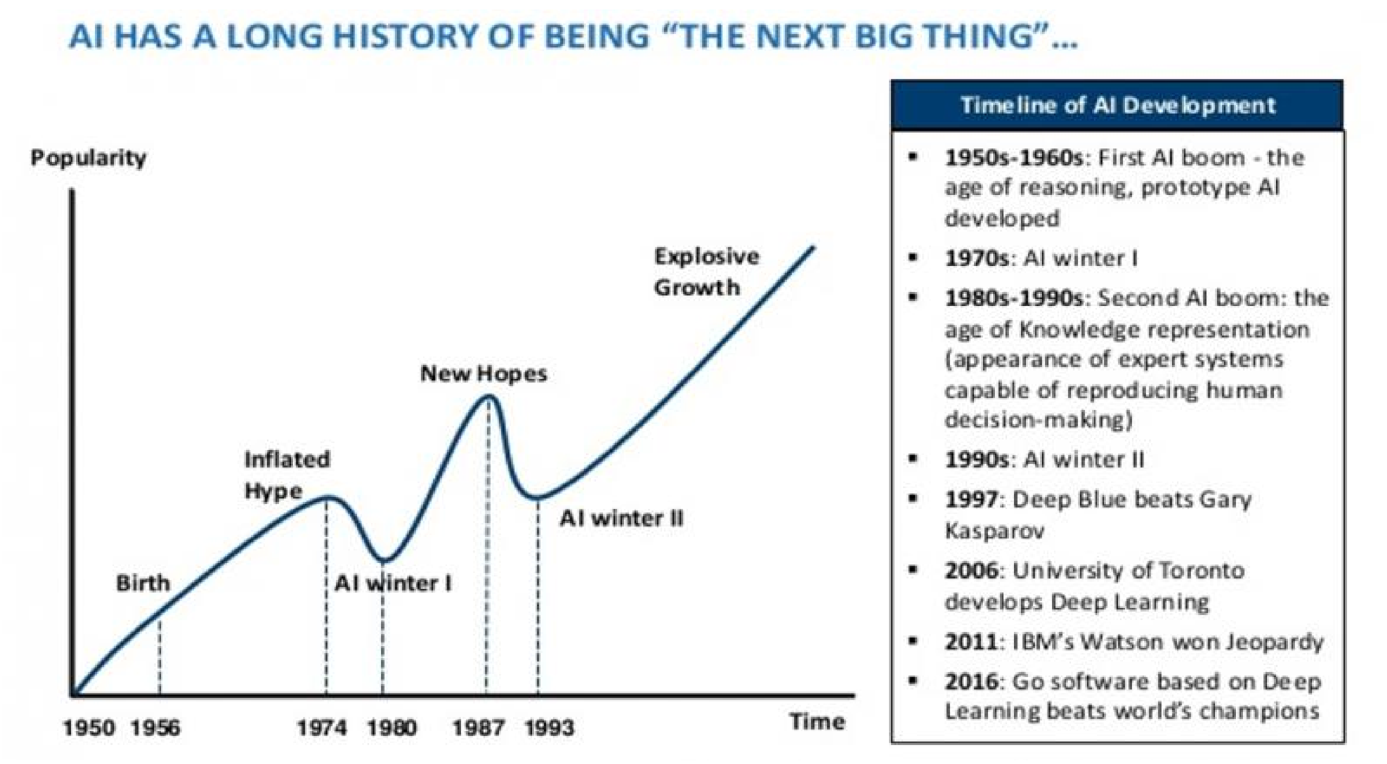
Webinar: AI in the Law

“I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.”

? Alan Turing, Computing machinery and intelligence

This is a huge field that is suddenly much more possible than it used to be. We’ve gone from “strange corner of a math department” to “driving some of the world’s fastest growing industries” in about two decades. There’s a lot of information and easily as much disinformation and fear out there. How to start making sense of it? AI has a “Shifting goalposts” problem, like fundamental particle physics. It’s basically always “what we hope a computer can do like a human can, but that we don’t feel is trivial by now.” That has applied to everything from managing an HVAC system to managing a business. These goalposts are often defined by the phrase “a machine will never:” (assemble a car by itself, play chess, understand speech, play chess well, beat a human at go, compose music, make a painting, drive a truck, write a book, teach children…..and so on). A lot of the hype comes from the fact that we’re now knocking down these goalposts almost as fast as people can put them up.

Let’s try “data first, model as a tool” instead. Well, we can still do that and run into problems , i.e. the “is the check upside down” problem. So we really need to have “data first” methods that can also train *to* that data - fully generalizable models.


AI HAS A LONG HISTORY OF BEING “THE NEXT BIG THING”…

Timeline of AI Development

- **1950s-1960s**: First AI boom - the age of reasoning, prototype AI developed
- **1970s**: AI winter I
- **1980s-1990s**: Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
- **1990s**: AI winter II
- **1997**: Deep Blue beats Gary Kasparov
- **2006**: University of Toronto develops Deep Learning
- **2011**: IBM’s Watson won Jeopardy
- **2016**: Go software based on Deep Learning beats world’s champions

What can AI actually do?

- Describe the world in some specific ways
- Turn input parameters into estimated measurements

The issue:

- Requires deep expertise
- Can be super complex
- Doesn’t catch unknown-unknowns
- What happens when we have data with no model?

Supervised machine learning

- Uses labeled examples
- Goal: predict what the labeled data would be

Unsupervised machine learning

- Uses unlabeled examples
- Tries to group or “cluster” data together to allow meaningful

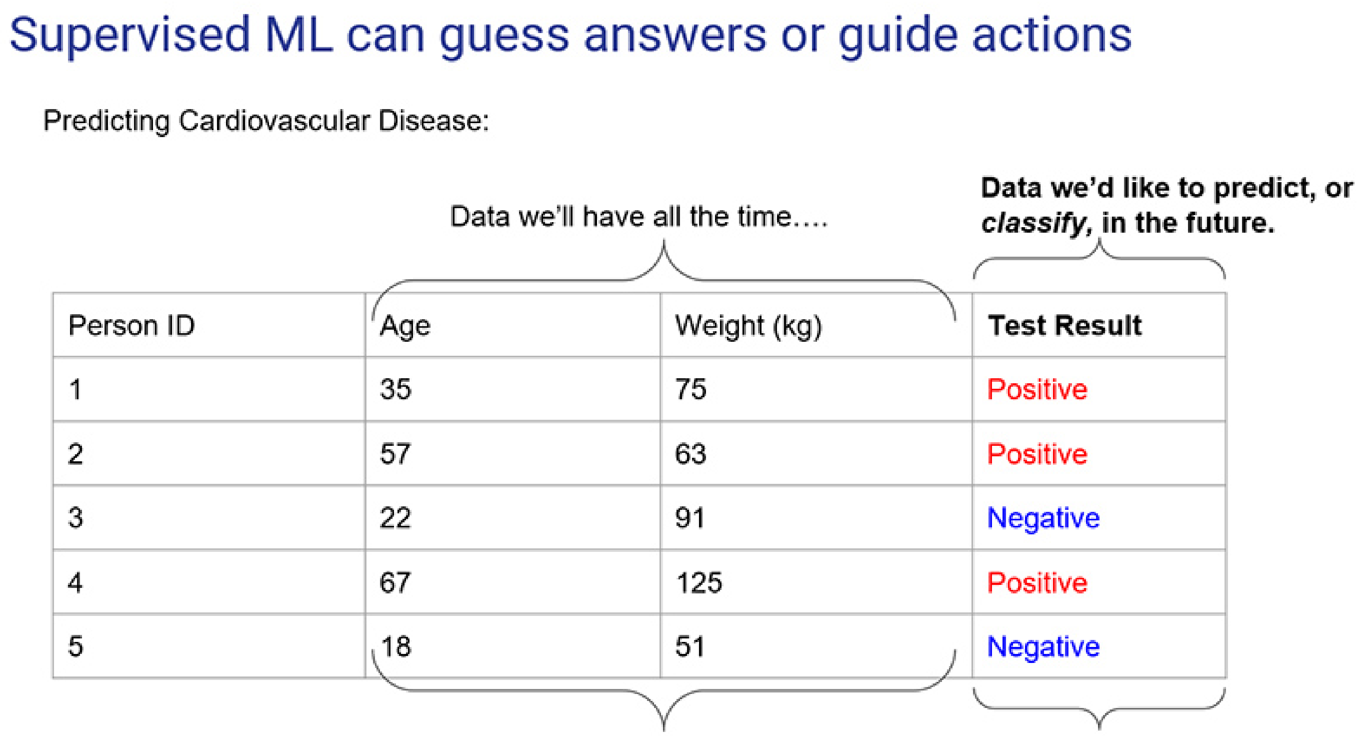
Reinforcement Learning

- Uses simulated or real environment
- Uses reward function

- Goal: Get the highest reward function

Modern Robotics: Q Learning and behavior optimization Needs three things:

- A simulated (or real) world with possible states
- A reward function to optimize against
- Logic to turn inputs into outputs



Supervised ML can guess answers or guide actions

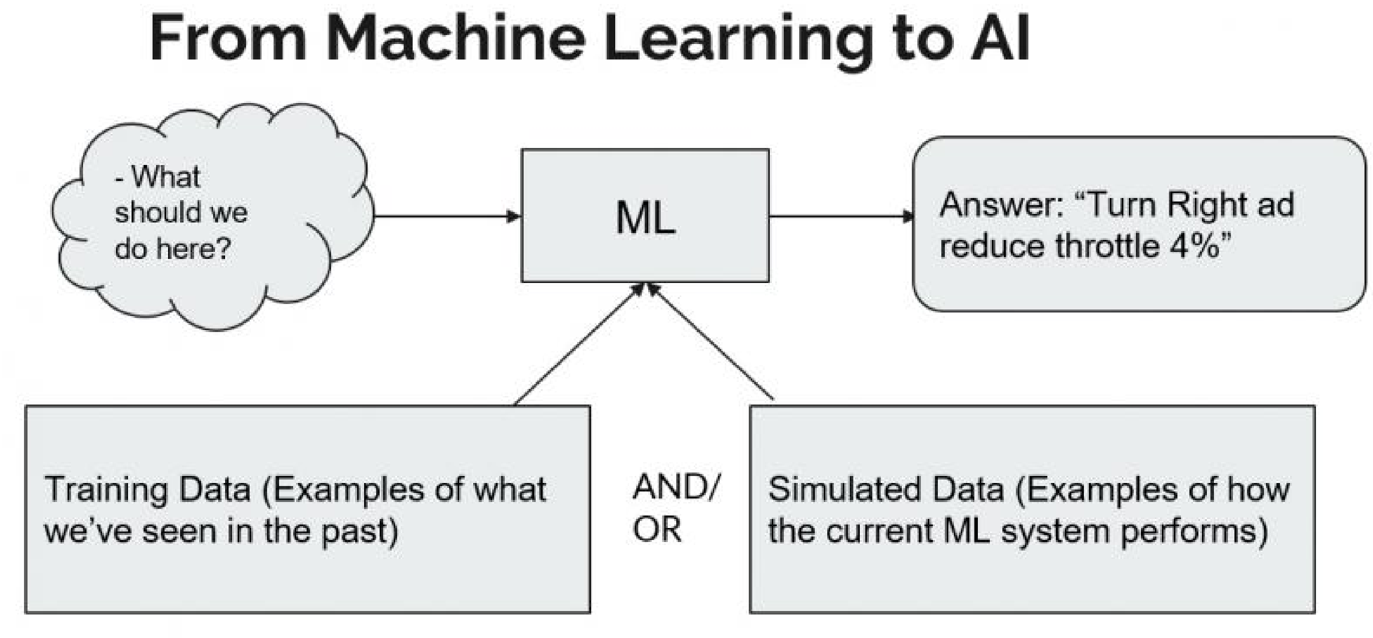
Predicting Cardiovascular Disease:

Data we'll have all the time....

Data we'd like to predict, or *classify,* in the future.

Person ID	Age	Weight (kg)	**Test Result**
1	35	75	Positive
2	57	63	Positive
3	22	91	Negative
4	67	125	Positive
5	18	51	Negative

This is a huge field that is suddenly much more possible than it used to be. We've gone from "strange corner of a math department" to "driving some of the world's fastest growing industries" in about two decades. There's a lot of information and easily as much disinformation and fear out there. How to start making sense of it? ML runs a lot of things nowadays, from crop yield estimates to credit scores to office software. But...how do these techniques add up to automation and robotics?



From Machine Learning to AI

- What should we do here?

ML

Answer: “Turn Right ad reduce throttle 4%”

Training Data (Examples of what we’ve seen in the past)

AND/ OR

Simulated Data (Examples of how the current ML system performs)

Modern robotics: Q learning and behavior optimization needs 3 things:

- A simulated (or real) world with possible states
- A reward function to optimize against

Logic to Turn Inputs into Outputs

A best input-output matrix (the “Q Matrix”) is trained to match all possible inputs to behavior. The same optimization routines are used to produce the best behavior across all possible states. This way we can learn behaviors we don’t even know ourselves. So now we can, say, make a self driving car. And now we have problems. Such as:

Someone had to decide what "good job" means, and robots will try to optimize their behavior to be considered "good." But does "good" mean 'protect humans too', or would we rather have robots not include humans explicitly in their thinking? Consider:

The opportunistic killer: The self-driving car that can take any action it sees fit so long as it minimizes both travel time and maximizes safety of all humans, and finally its own safety. Researchers are working very hard to make this happen, but in a context-free environment, perhaps these goals are given relative weights - with safety the highest, speed the second, and vehicle safety, defined as the likelihood of further operation, third. However, if the car is allowed to notice its own demise from any cause, it is a possible fringe case that the car would "learn" that driving is the only thing that kills it, either due to eventual part failure or an accident, and refuse to drive at all. If this was then overridden by disallowing that as an option, the next option might be to attempt to run over the owner in any way possible, and then stop, removing the likely future force making the car drive and thus, living forever. This seems like a violation of the "life preservation" clause, but as written it isn't strictly a law, just part of an optimization function. By allowing self-preservation into our AI devices as other than a strict hierarchy with the safety of others, fringe cases like this become possible. Here's where you invoke Asimov's laws of robotics, of course, and for good reason. So maybe those laws....should be made actual laws. If AI is made with relative altruism instead of strict altruism, these cases become possible.

Scenario 3: The Paperclip Maximizer (Bostrom)

Suppose we have an AI whose only goal is to make as many paper clips as possible. The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. Because if humans do so, there would be fewer paper clips. Also, human bodies contain a lot of atoms that could be made into paper clips. The future that the AI would be trying to gear towards would be one in which there were a lot of paper clips but no humans.

- How does this affect business, present and future?
- A lot of tasks that can be described as “given inputs A, do action or make assessment B” can be done totally by algorithms
- A lot of jobs are “on the block” - a 2013 report1 claims almost half of all US employment is at risk
- Replacement by intelligent agents - not just “unskilled” labor:
- Drivers account for about 2% of total employment all by themselves.
- Advancements in deep learning allow image analysis and other decision making
- Radiology, oncology, physical and information security, the list goes on...
- But what about jobs that seem to “require” humans now, combinations of physical, emotional and decision making work?
- Physical and general robotics may be our near future
- What will we do if productivity explodes while employment shrinks?
- What about ethical actions taken by AI systems?
- Instead of “if X, do Y” we must work with “act such as to maximize Z.”
- We need a framework to think about ethical outcomes with these new systems!

1 https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

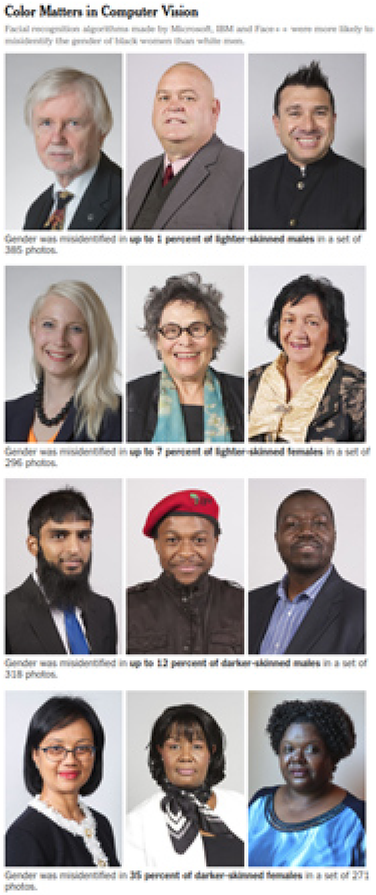
2 https://www.cnbc.com/2017/05/22/goldman-sachs-analysis-of-autonomous-vehicle-job-loss.html

How can an Algorithm be Biased?

- Several ways:
- Behavioral example data can be biased
- Data may have gaps
- Minority data may be ignored
- Underlying populations may be unequal

Biased Example Behavioral Data

- Remember “what would a person do?” It isn’t the best idea when people are biased as well…
- In 2016, Microsoft launched “Tay.ai,” a twitter chatbot
- ‘Tay’ was designed to learn from people who tweeted in her direction, and repeat and produce similar statements
- Within 24 hours, she was saying things along the lines of “hitler was right” and other ridiculous phrases
- The lesson: If “sound like others” is the only metric for success, there can be awful fallout.
- One possible solution: Change the success metrics!



Color Matters in Computer Vision

Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Incomplete Data Leads to Racial Bias

- Facial recognition generally relies on convolutional neural networks
- Dependent variables like “is this a face” and “what is the gender” are estimated with highly complex and opaque data manipulations
- All algorithms are trained until they “succeed” most with the average of a population
- If lots of data from one “class” are used and only a little from another, the algorithm will use most of its “latent space” training to find variables that are useful for that class
- Result: whoever you measure a lot, the algorithm will work well on. Others, not so much.
- One possible solution: Take more data!

Underlying Population Differences

- Say I want to issue car loans and get the best returns.
- I could produce “neighborhood level” economic demographics, average credit score, income or loan default rate.
- But doing this risks producing a tool with strong implicit bias along features we don’t want to use
- Say we find the poorest 10% of neighborhoods account for 50% of the failures-to pay, or 20% of the otherwise-negative experiences on the job.
- Making a model that strongly or exclusively considered geography would also tend to “label” a larger set of people from one demographic (say a race, or religion) as being worse ‘risks’ due to varying demographics between neighborhoods
- One possible solution: explicitly evaluate and correct for the bias!

So How can We Make Things “Fair?”

- The field is thinking about this issue a lot. Several possibilities
- “Maximize my return”: If no bias is explicitly added, the results can be considered “fair.”
- “Remove group awareness”: If no features are present that should allow explicit bias, the results are “fair.”
- “Inter-group parity”: If the same fraction of each demographic group end up chosen (or valued positively, or incorrectly measured, etc), then the results are “fair.”
- “Equal Opportunity”: If, of the people who would represent a positive measurement, the same ratio in each demographic group are estimated to be positive, then the results are “fair.”
- Others are possible: Equalizing false positives, etc.

So What Are we Going to Do About ML Bias?

- Even now, bias is very possible and the field needs consistent guidance and consensus

about what “fair” looks like.
- We should consider tests for differential treatment with broad use-cases
- The field needs consensus on which “fairness” metric we should use and when

So What are We Going to do About Emergent Behaviors and Dangerous AI?

- We need a standards organization. We need guidance that could drive legislation on the development and use of strong AI
- What variations to current modern AI would precipitate these nightmare scenarios?
- The development of self-referential reasoning (in which the value function is tied to the state of health of the robot itself), the development of recursive self-programming (in which the algorithm can vary not just its parameter set but the value function or the surrounding code used to construct the model) and others

So what are we going to do about emergent behaviors and dangerous AI?

- What limitations could be put in place to allow strong AI development without risking "runaway?" some possibilities:
- "checkpointing" where human input is periodically required to continue training and learning
- "structure limiting" that disallows model past a certain level of complexity
- "algorithm isolation" where robots undergoing learning processes cannot share this information directly without human intervention
- Other improvements to common practice include a development of a standard test set of common and uncommon circumstances, and the "checkout" of any algorithm by displaying acceptable behavior in those tests before deployment in the real world.

There are still other issues as well

- If AI solutions are allowed to do jobs that humans now do (from driving to surgery), who holds the liability for accident and error? We have standards issues. We need:
- Certification
- Understanding of liability
- Standardized tests for performance
- Remediation best practices
- These can be informed by legal definition and precedent
- Even if humanity is “safe,” how do we distribute resources in a world where the cost of goods and service plummets?
- As the legal and tech community, we need to get out in front of this!

And right now? It’s kind of the Wild West out here…

Very little of this is mandatory or even standard practice anywhere; most attempts at compliance with these principles is voluntary and nascent

Cases like the Uber car accident will be important in defining where liability is handled, we may expect similar cases as autonomous systems become more common

We can expect the next decade to necessitate a lot of these cases, we probably won’t go another 10 years without some kind of real regulatory body either.

In the end, we should brace for some of thee issues, expect a few significant legal cases, and “get ahead” of it wherever we can to head off potentially dangerous fallout.

Printed:May 18, 2022
http://www.lsntap.org/node/202/webinar-ai-law
©Legal Services National Technology Assistance Project