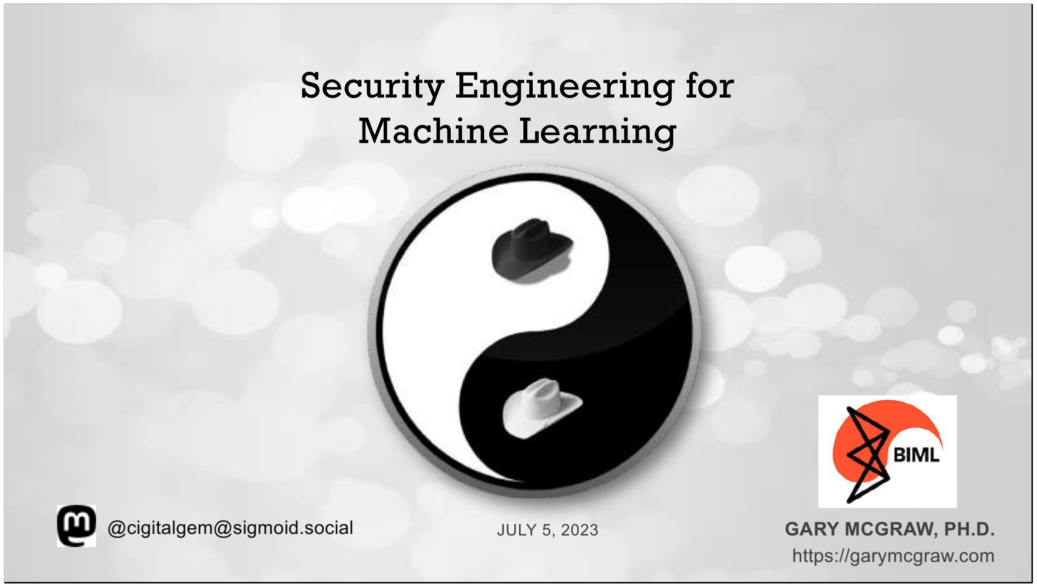


Security Engineering for Machine Learning

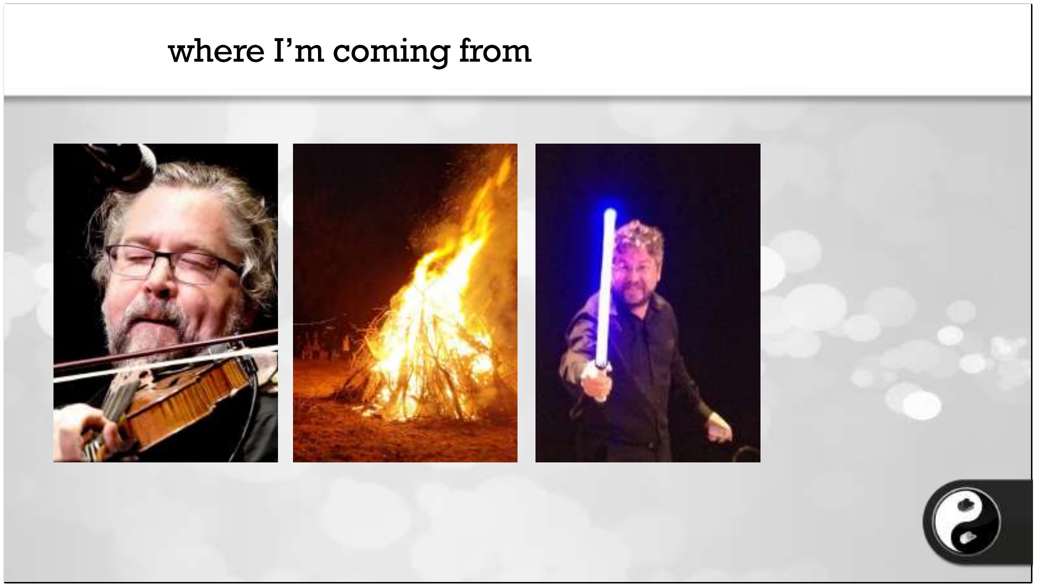
@cigitalgem@sigmoid.social

JULY 5, 2023

GARY MCGRAW, PH.D.
https://garymcgraw.com

PLEASE LIVE TWEET THIS TALK!

For more see https://garymcgraw.com



where I’m coming from

I AM AN OPTIMIST

I am the guy who gave the keynote this morning

Technology

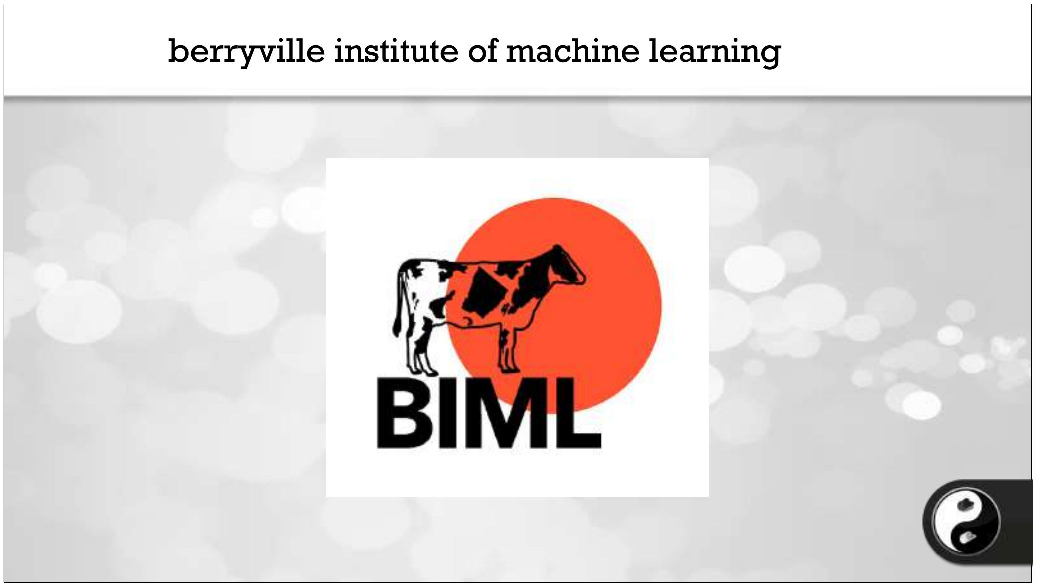
- Northern Virginia-Based Cigital to Synopsys (500 people)
- Invented the field of software security (12 books)
- alpha-geek who gives 20 talks a year
- Light saber

Music

- Carnegie Hall at 10 and 16. Suzuki.
- The Bitter Liberals
- Where’s Aubrey ($16,912)
- Sold out show at the Bright Box Saturday the 4th
- Funny faces while playing the violin

Life

- Clarke County on the river near Berryville,
- Living in the country
- Fiction reader, Art collector, Craft cocktail maker, Cook
- Solstice parties

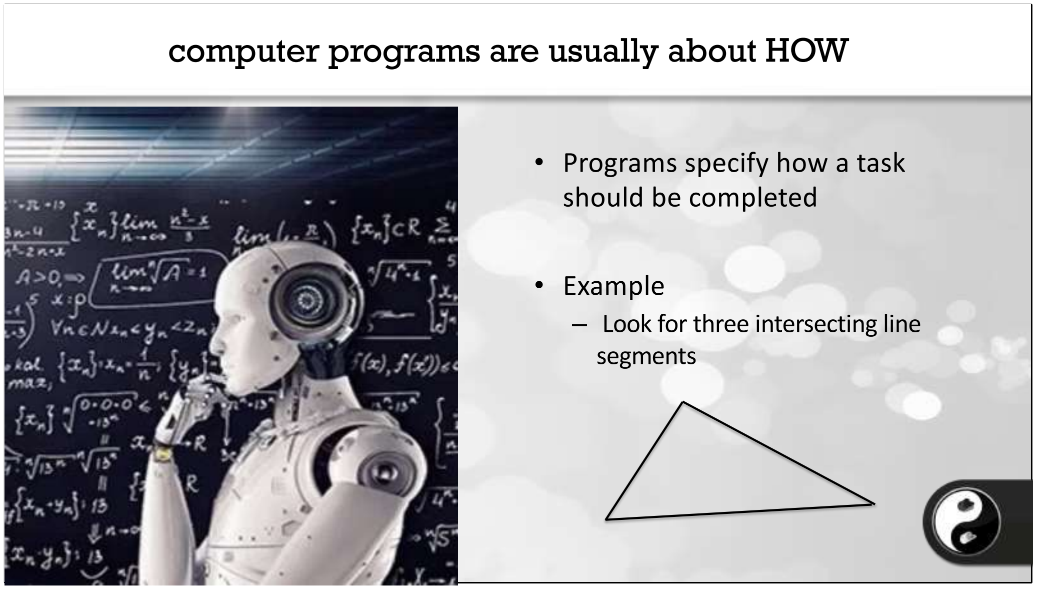


berryville institute of machine learning

Founded in January 2019, Our research at BIML focuses on three threads: building a taxonomy of known attacks on ML, exploring a hypothesis of representation and ML risk, and performing an architectural risk analysis (sometimes called a threat model) of ML systems in general.

See https://berryvilleiml.com

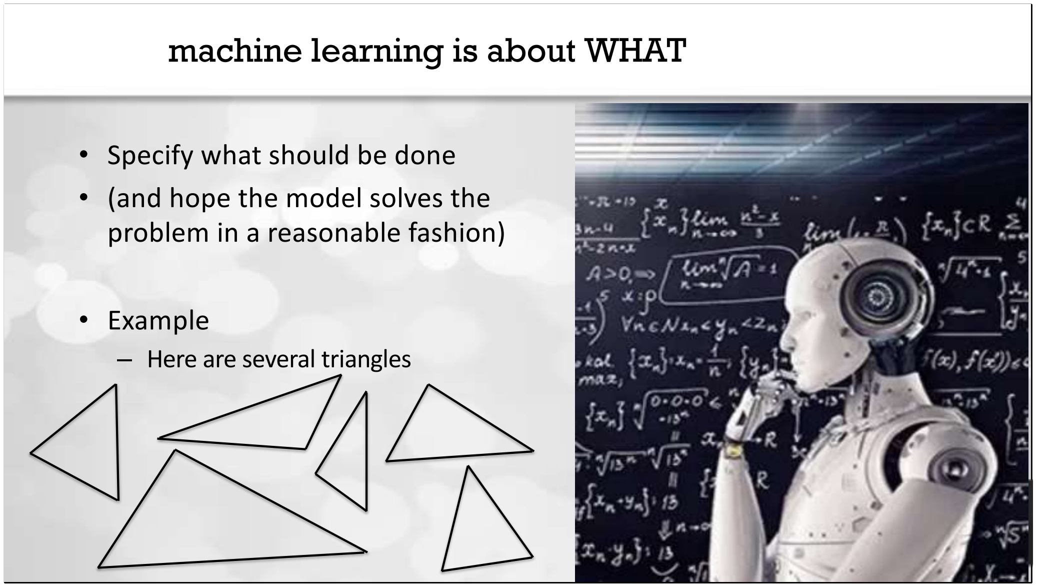
intro to ML and ML Security



computer programs are usually about HOW

- Programs specify how a task should be completed
- Example
  - Look for three intersecting line segments

Programs are brittle
Sometimes we don’t know how to perfectly describe HOW to do something



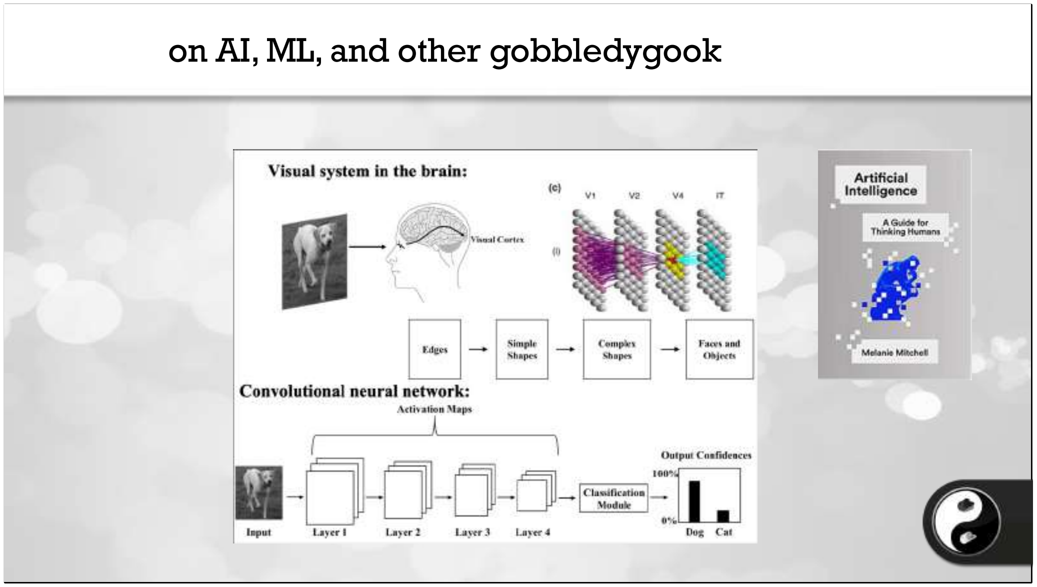
machine learning is about WHAT

- Specify what should be done
- (and hope the model solves the problem in a reasonable fashion)

- Example
  - Here are several triangles

ML models often “cheat” and solve a task through unintended means

WOLF-HUSKY-SNOW example

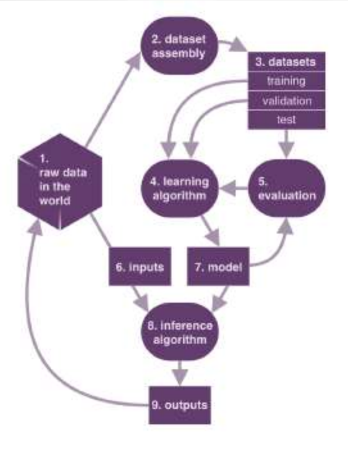


on AI, ML, and other gobbledygook

This is a classic picture processing model inspired by the brain.

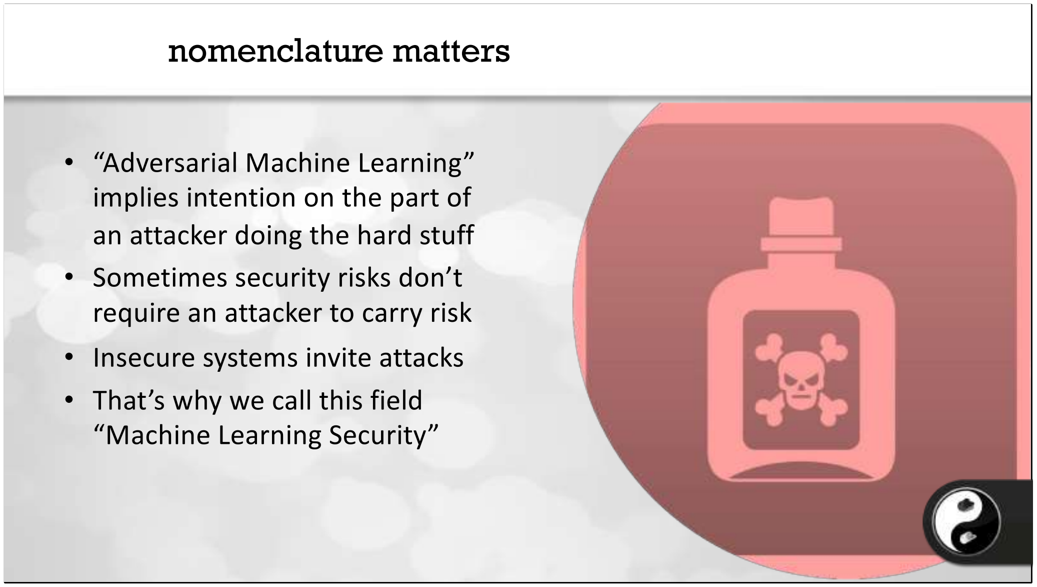
Thing is, the model is WAY WAY WAY WAY WAY more simple than a brain is. But it works.

a generic ML model



- Nine basic components
  - Processes are ovals
  - Collections are rectangles
- Arrows represent information flow
- We used this model to think about risks in each component

ML systems come in a variety of shapes and sizes, and frankly each possible ML design deserves its own specific ARA. For the purposes of this work, we describe a generic ML system in terms of its constituent components and work through that generic system ferreting out risks. The idea driving us is that risks that apply to this generic ML system will almost certainly apply in any specific ML system.

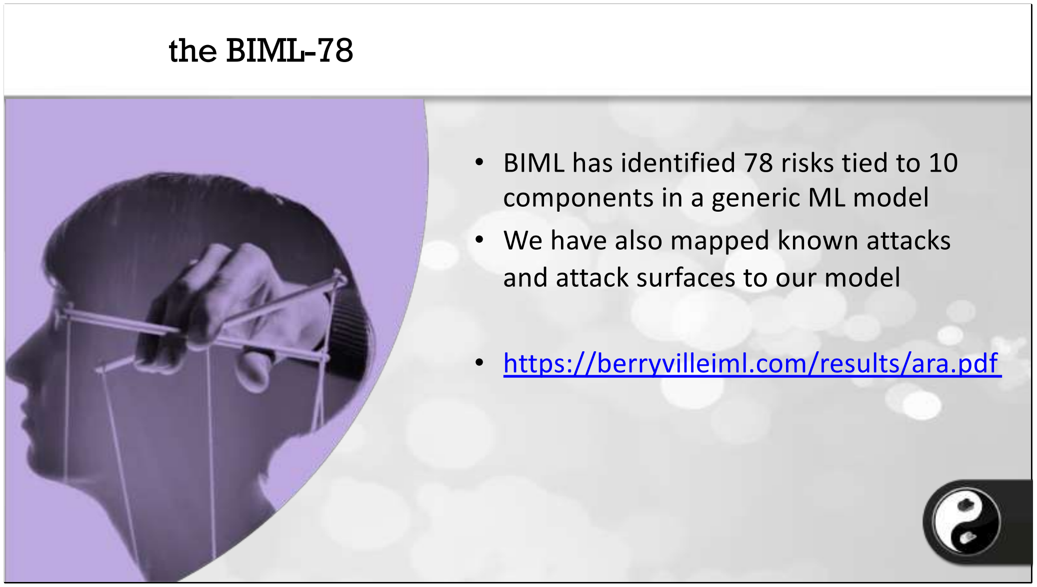


nomenclature matters

- “Adversarial Machine Learning” implies intention on the part of an attacker doing the hard stuff
- Sometimes security risks don’t require an attacker to carry risk
- Insecure systems invite attacks
- That’s why we call this field “Machine Learning Security”

ML risk analysis

An architectural risk analysis is more intense than an attack-based approach and is driven by a risk framework

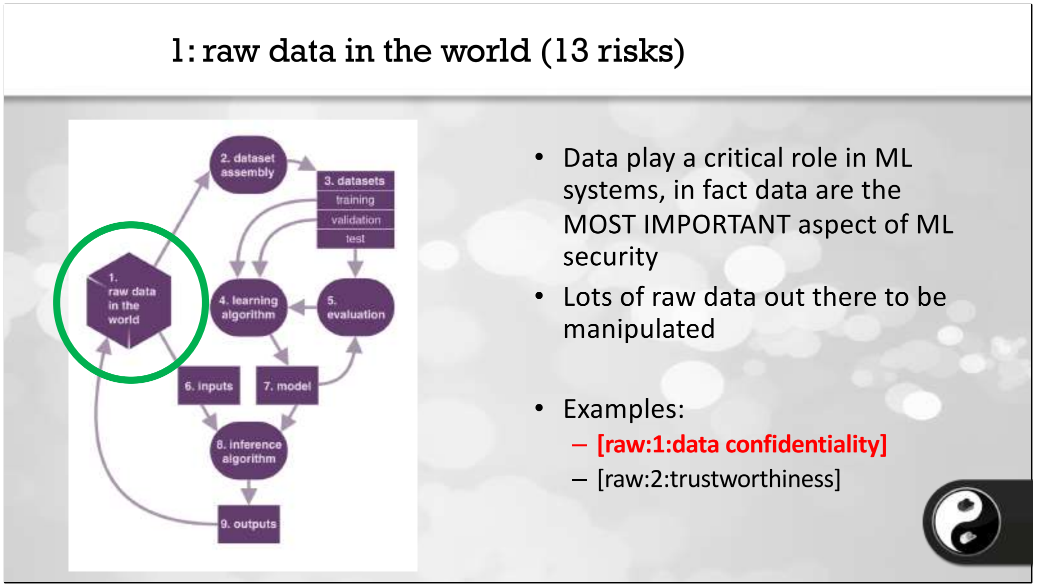


the BIML-78

- BIML has identified 78 risks tied to 10 components in a generic ML model
- We have also mapped known attacks and attack surfaces to our model

- https://berryvilleiml.com/results/ara.pdf

We'll fly through the 78 risks, introducing you to 10



1: raw data in the world (13 risks)

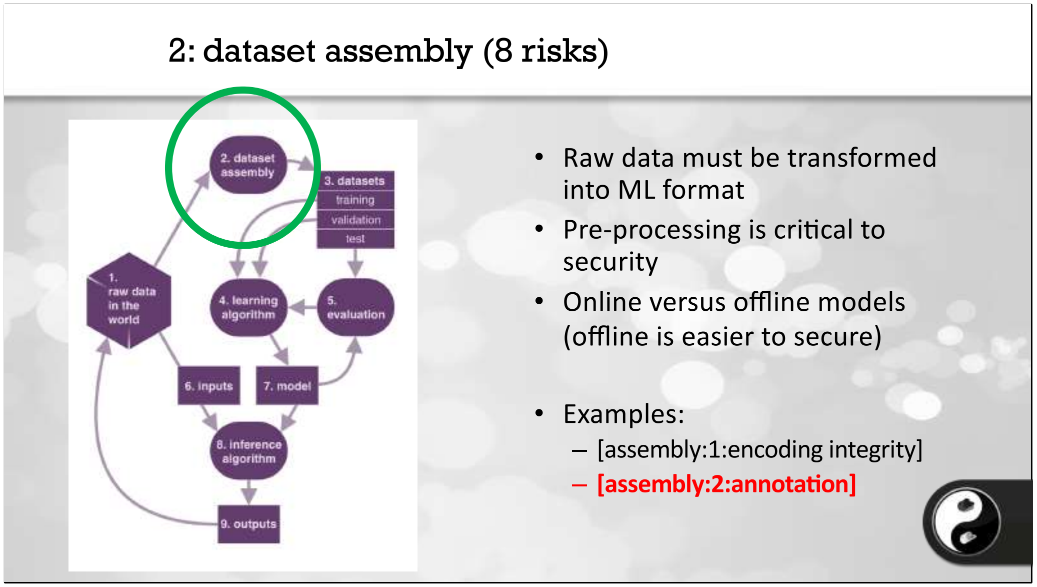
- Data play a critical role in ML systems, in fact data are the MOST IMPORTANT aspect of ML security
- Lots of raw data out there to be manipulated

- Examples:
  - **[raw:1:data confidentiality]**
  - [raw:2:trustworthiness]

1. An ML system that is trained up on confidential or sensitive data will have some aspects of those data built right into it through training. Attacks to extract sensitive and confidential information from ML systems (indirectly through normal use) are well known.

2. Data sources are not always trustworthy, suitable, and reliable. How might an attacker tamper with or otherwise poison raw input data? What happens if input drifts, changes, or disappears?

There are eleven more of these risks in the paper.



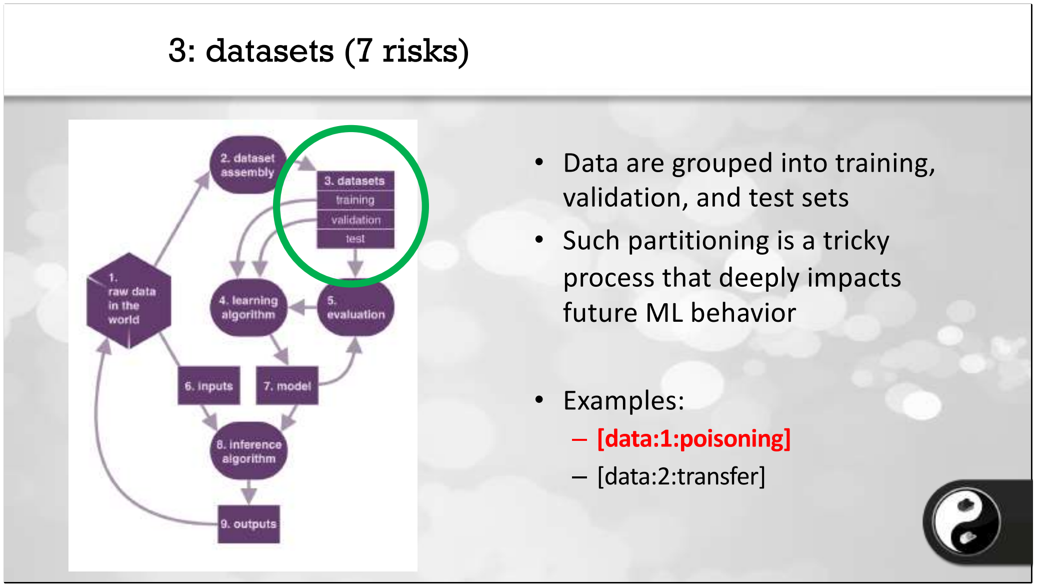
2: dataset assembly (8 risks)

- Raw data must be transformed into ML format
- Pre-processing is critical to security
- Online versus offline models (offline is easier to secure)
- Examples:
  - [assembly:1:encoding integrity]
  - **[assembly:2:annotation]**

1. Encoding integrity can be both introduced and exacerbated during pre-processing. Does the pre-processing step itself introduce security problems? Bias in raw data processing can impact ethical and moral implications.

2. The way data are “tagged and bagged” (or annotated into features) can be directly attacked, introducing attacker bias into a system. An ML system trained up on examples that are too specific will not be able to generalize well. Much of the human engineering time that goes into ML is spent cleaning, deleting, aggregating, organizing, and just all-out manipulating the data so that it can be consumed by an ML algorithm.

There are six more of these risks in the paper!



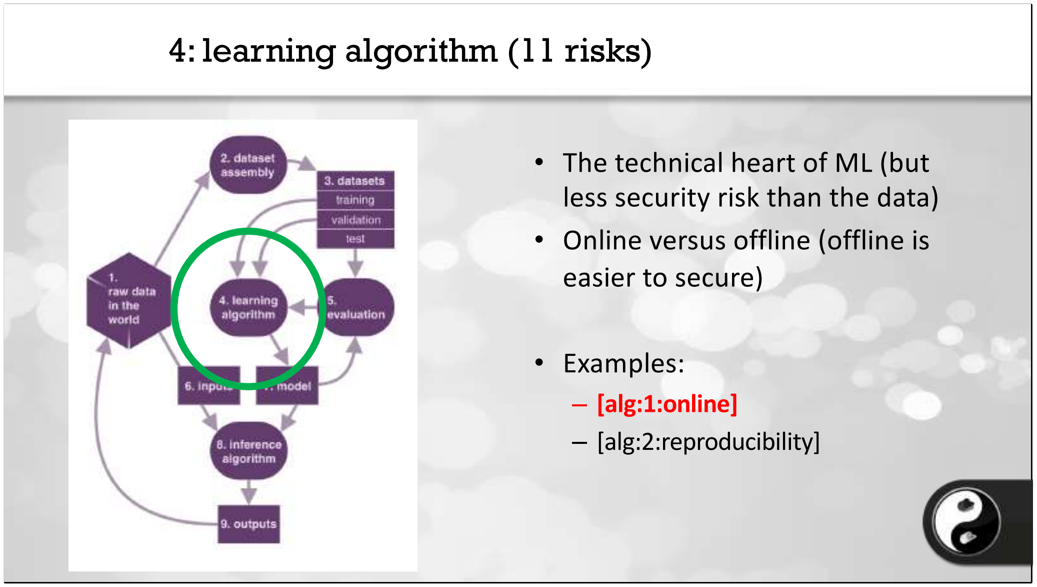
3: datasets (7 risks)

- Data are grouped into training, validation, and test sets
- Such partitioning is a tricky process that deeply impacts future ML behavior
- Examples:
  - **[data:1:poisoning]**
  - [data:2:transfer]

1. All of the first three components in our generic model (raw data in the world, dataset assembly, and datasets) are subject to poisoning attacks whereby an attacker intentionally manipulates data in any or all of the three first components, possibly in a coordinated fashion, to cause ML training to go awry. Recall Microsoft TAY.

2. Many ML systems are constructed by tuning an already trained base model so that its somewhat generic capabilities are fine-tuned with a round of specialized training. A transfer attack presents an important risk in this situation. Pre-trained model risks carry over, and Trojans may be inserted.

There are five more of these risks in the paper!



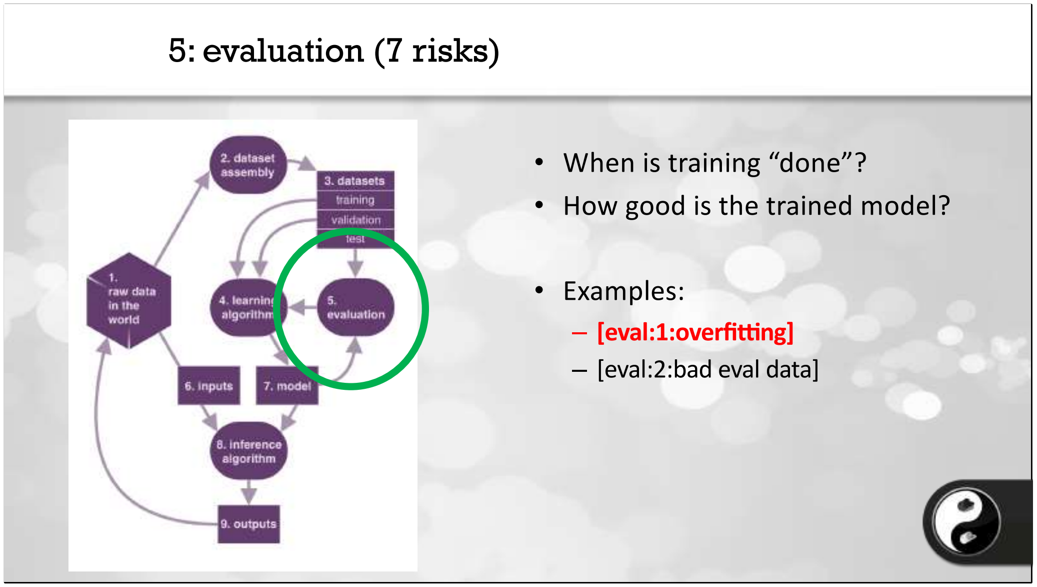
4: learning algorithm (11 risks)

- The technical heart of ML (but less security risk than the data)
- Online versus offline (offline is easier to secure)
- Examples:
  - **[alg:1:online]**
  - [alg:2:reproducibility]

1. An online learning system that continues to adjust its learning during operations may drift from its intended operational use case. Clever attackers can nudge an online learning system in the wrong direction on purpose.

2. ML work has a tendency to be sloppily reported. Results that can't be reproduced may lead to overconfidence in a particular ML system to perform as desired.

There are nine more of these risks in the paper!



5: evaluation (7 risks)

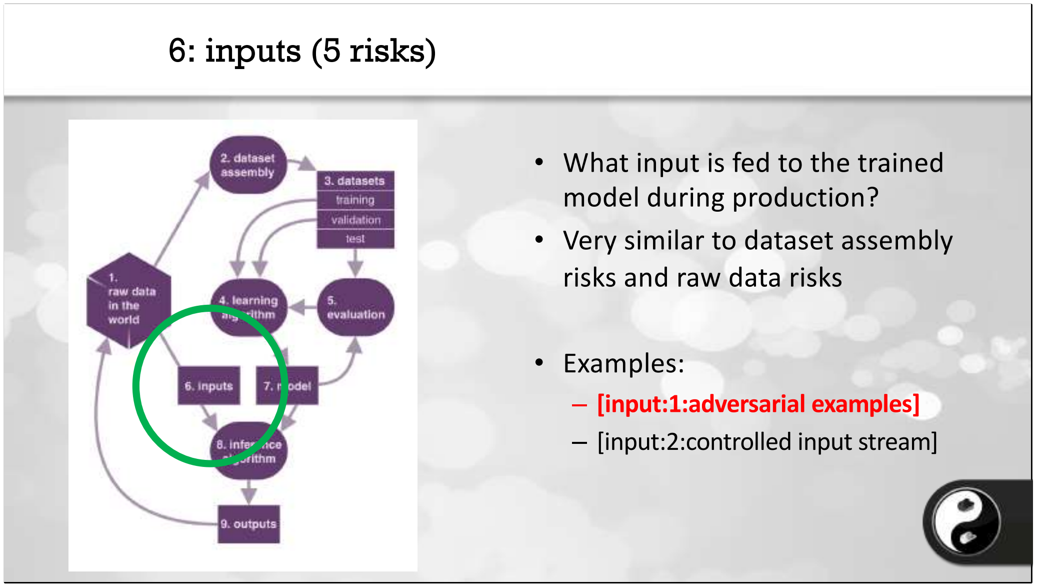
- When is training "done"?
- How good is the trained model?

- Examples:
  - **[eval:1:overfitting]**
  - [eval:2:bad eval data]

1. A sufficiently powerful machine is capable of learning its training data set so well that it essentially builds a lookup table. This can be likened to memorizing its training data. The unfortunate side effect of "perfect" learning like this is an inability to generalize outside of the training set and is called overfitting.

2. A bad evaluation data set that doesn't reflect the data it will see in production can mislead a researcher into thinking everything is working even when it's not. Evaluation sets can also be too small or too similar to the training data to be useful.

There are five more of these risks in the paper!



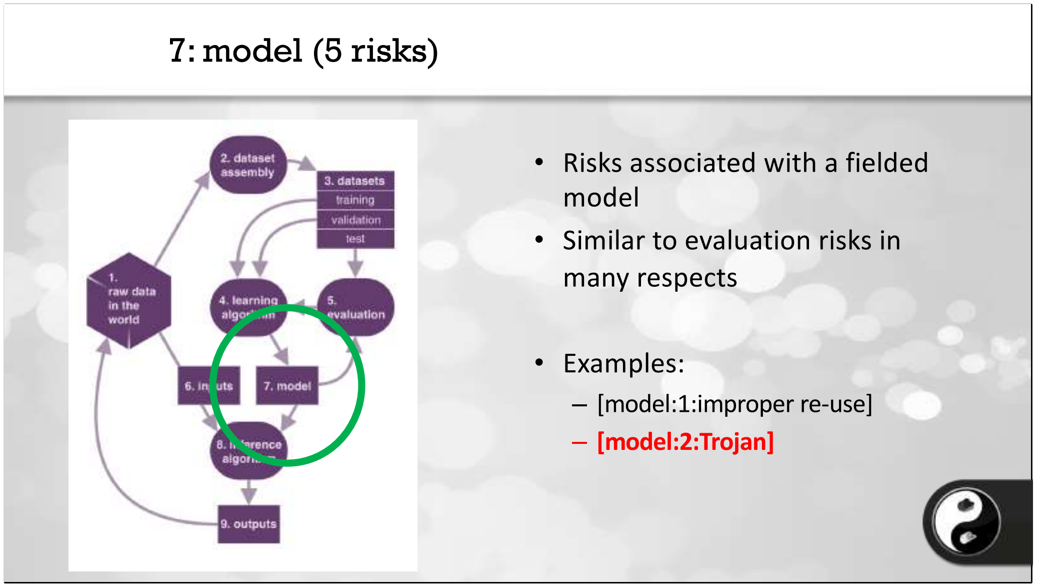
6: inputs (5 risks)

- What input is fed to the trained model during production?
- Very similar to dataset assembly risks and raw data risks
- Examples:
  - **[input:1:adversarial examples]**
  - [input:2:controlled input stream]

1. One of the most important categories of computer security risks is malicious input. The ML version of malicious input has come to be known as adversarial examples.

2. A trained ML system that takes as its input data from outside may be purposefully manipulated by an attacker.

There are three more of these risks in the paper!



7: model (5 risks)

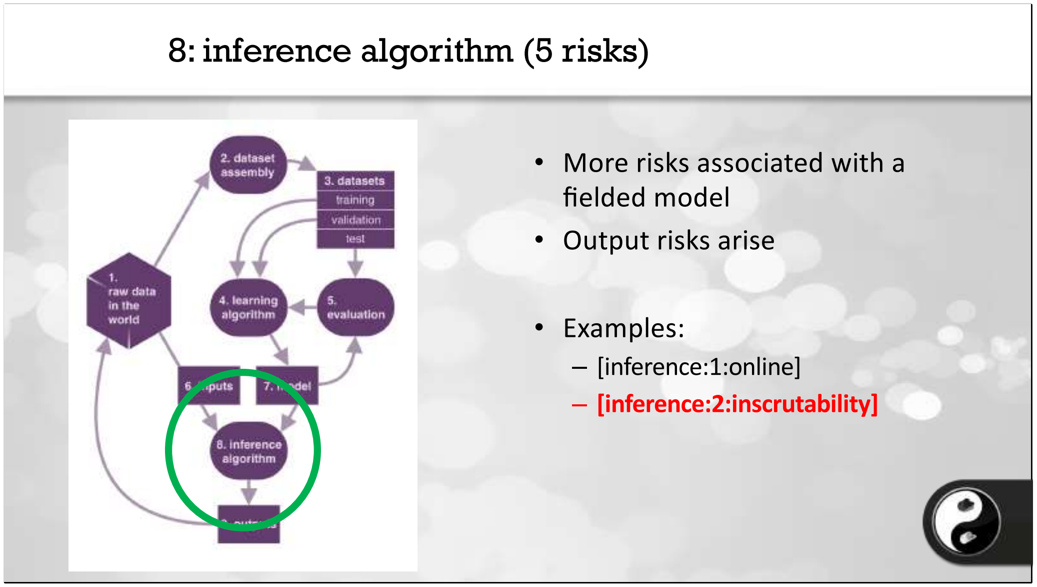
- Risks associated with a fielded model
- Similar to evaluation risks in many respects

- Examples:
  - [model:1:improper re-use]
  - **[model:2:Trojan]**

1. ML-systems are re-used intentionally in transfer situations. The risk of transfer outside of intended use applies.

2. Model transfer leads to the possibility that what is being reused may be a Trojaned (or otherwise damaged) version of the model being sought out

There are three more of these risks in the paper!



8: inference algorithm (5 risks)

- More risks associated with a fielded model
- Output risks arise

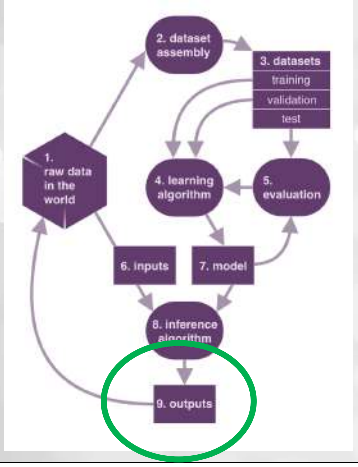
- Examples:
  - [inference:1:online]
  - **[inference:2:inscrutability]**

1. A fielded model operating in an online system (that is, still learning) can be pushed past its boundaries.

2. In far too many cases, an ML system is fielded without a real understanding of how it works or why it does what it does. Integrating an ML system that "just works" into a larger system that then relies on the ML system to perform properly is a very real risk.

There are three more of these risks in the paper!

9: outputs (7 risks)



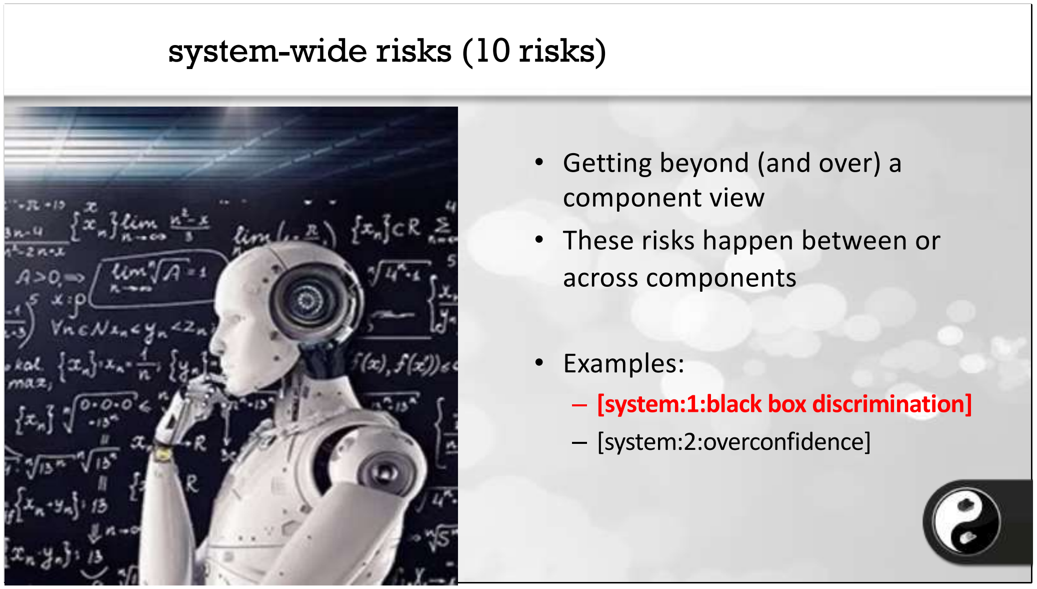
- System output is often the whole point
- Direct attack on the output is pretty obvious

- Examples:
  - **[output:1:direct]**
  - [output:2:provenance]

1. An attacker tweaks the output stream directly. This will impact the larger system in which the ML subsystem is encompassed. There are many ways to do this kind of thing. Probably the most common attack would be to interpose between the output stream and the receiver. Inscrutability of ML makes this easier.

2. ML systems must be trustworthy to be put into use. Even a temporary or partial attack against output can cause trustworthiness to plummet.

There are five more of these risks in the paper!



system-wide risks (10 risks)

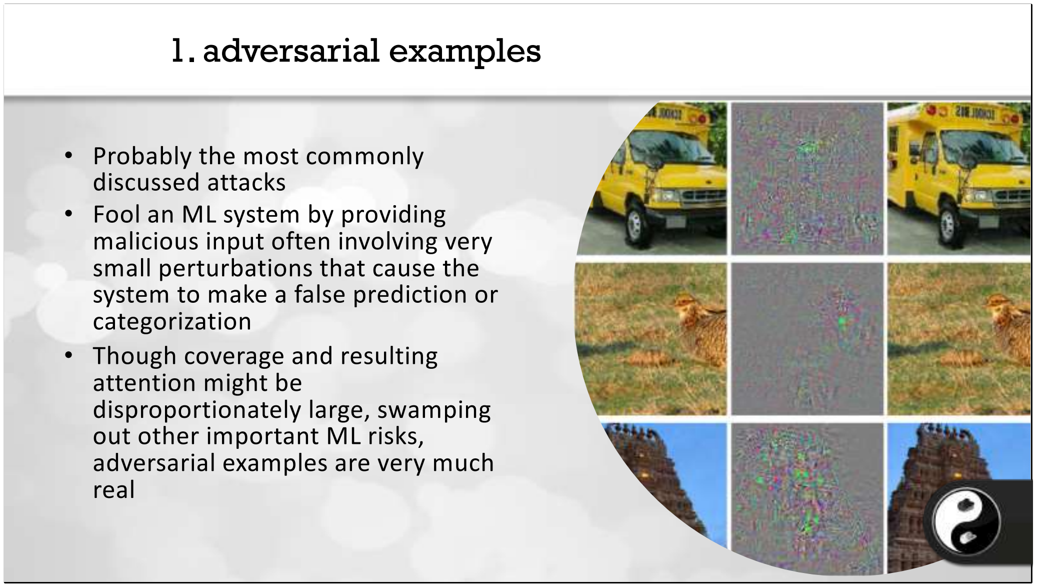
- Getting beyond (and over) a component view
- These risks happen between or across components
- Examples:
  - **[system:1:black box discrimination]**
  - [system:2:overconfidence]

1. Many data-related component risks lead to bias in the behavior of an ML system. ML systems that operate on personal data or feed into high impact decision processes (such as credit scoring, employment, and medical diagnosis decisions) pose a great deal of risk. When biases are aligned with gender, race, or age attributes, operating the system may result in discrimination with respect to one of these protected classes.

2. When an ML system with a particular error behavior is integrated into a larger system and its output is treated as high confidence data, users of the system may become overconfident in the operation of the system for its intended purpose. Developing overconfidence in ML is made easier by the fact that ML systems are often poorly understood and vaguely described.

There are eight more of these risks in the paper!

top five ML risks

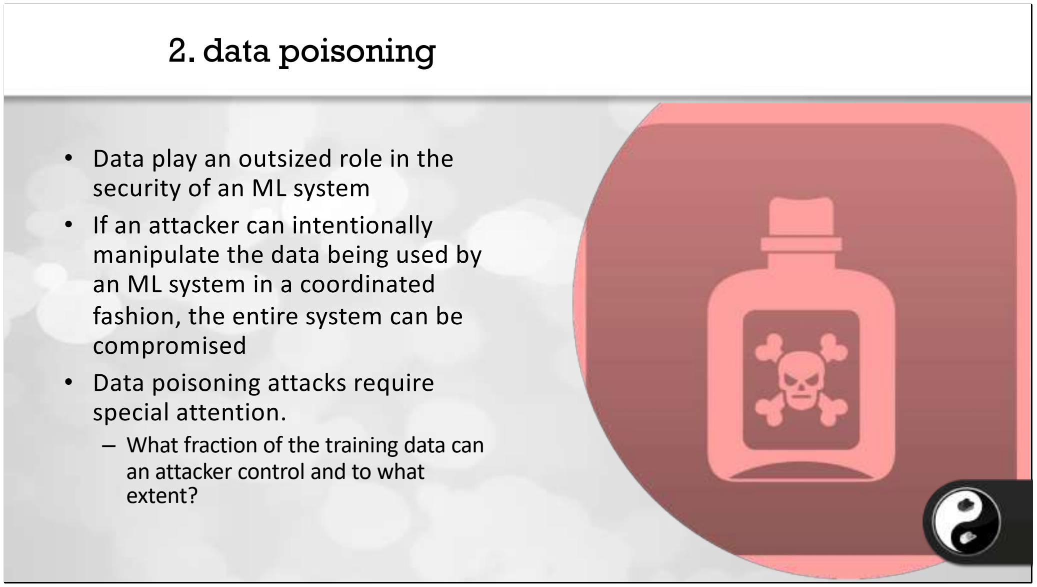


1. adversarial examples

- Probably the most commonly discussed attacks
- Fool an ML system by providing malicious input often involving very small perturbations that cause the system to make a false prediction or categorization
- Though coverage and resulting attention might be disproportionately large, swamping out other important ML risks, adversarial examples are very much real

Adversarial examples
Facial recognition's pitfalls: https://www.forbes.com/sites/forbestechcouncil/2019/10/04/how-facial-recognition-needs-to-improve-to-be-effective/
Self-driving cars and medical diagnosis (with diagrams): https://www.vox.com/future-perfect/2019/4/8/18297410/ai-tesla-self-driving-cars-adversarial-machine-learning

WHO IN YOUR ORGANIZATION SHOULD WATCH OUT FOR THIS?

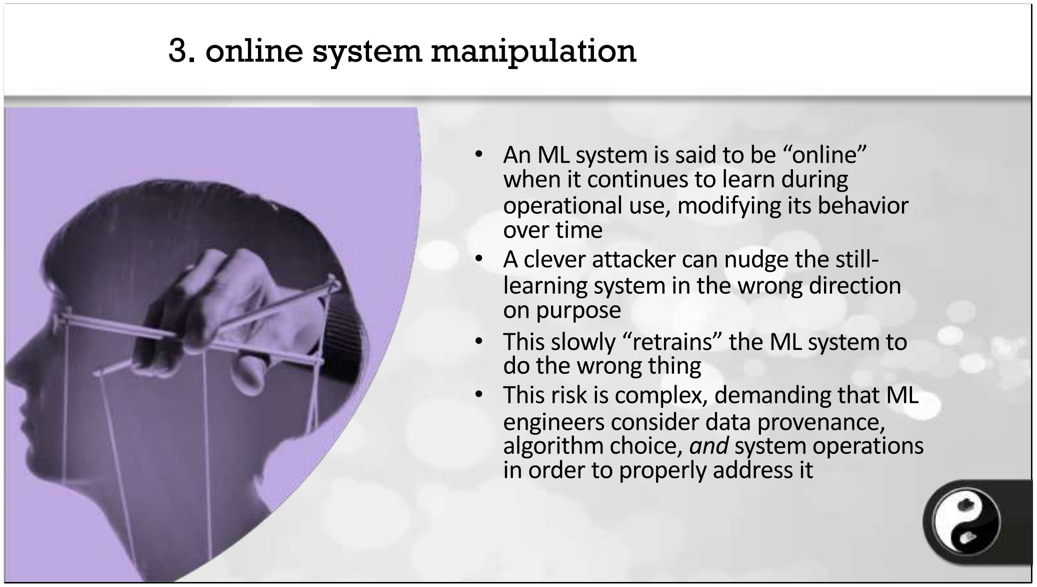


2. data poisoning

- Data play an outsized role in the security of an ML system
- If an attacker can intentionally manipulate the data being used by an ML system in a coordinated fashion, the entire system can be compromised
- Data poisoning attacks require special attention.
  - What fraction of the training data can an attacker control and to what extent?

Microsoft's Tay chatbot is a classic
https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/
https://en.wikipedia.org/wiki/Tay_(bot)

DOES THE DEPARTMENT OF DATA SCIENCE EVEN CONSIDER BAD ACTORS?

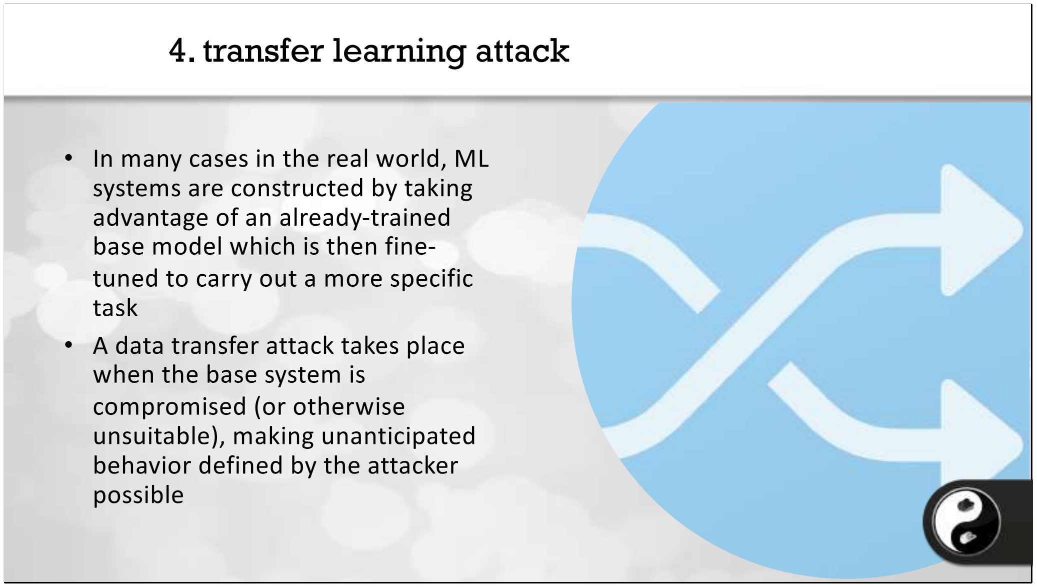


3. online system manipulation

- An ML system is said to be “online” when it continues to learn during operational use, modifying its behavior over time
- A clever attacker can nudge the still-learning system in the wrong direction on purpose
- This slowly “retrains” the ML system to do the wrong thing
- This risk is complex, demanding that ML engineers consider data provenance, algorithm choice, *and* system operations in order to properly address it

Microsoft's Tay chatbot is a classic
https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/
https://en.wikipedia.org/wiki/Tay_(bot)

UH OH, LOOKS LIKE OPERATIONS HAS A NEW JOB. DOES YOUR SEIM MONITOR YOUR ML SYSTEMS?



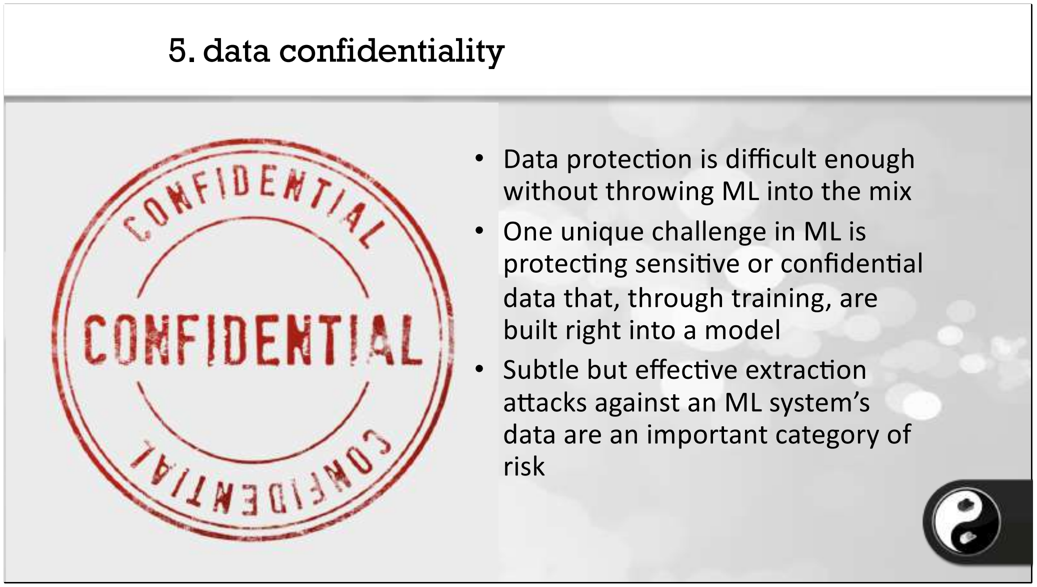
4. transfer learning attack

- In many cases in the real world, ML systems are constructed by taking advantage of an already-trained base model which is then fine-tuned to carry out a more specific task
- A data transfer attack takes place when the base system is compromised (or otherwise unsuitable), making unanticipated behavior defined by the attacker possible

Gu, T., B. Dolan-Gavitt, and S. Garg. “Badnets: Identifying vulnerabilities in the machine learning model supply chain.” arXiv preprint arXiv:1708.06733 (2017)

Kumar, R.S.S., D. O Brien, K. Albert, S. Viljöen, J. Snover, “Failure Modes in Machine Learning Systems.” arXiv preprint 1911.11034 (2019)

HOW MUCH MONEY DID IT COST YOU TO DEVELOP YOUR ML?


CONFIDENTIAL
CONFIDENTIAL
CONFIDENTIAL

5. data confidentiality

- Data protection is difficult enough without throwing ML into the mix
- One unique challenge in ML is protecting sensitive or confidential data that, through training, are built right into a model
- Subtle but effective extraction attacks against an ML system's data are an important category of risk

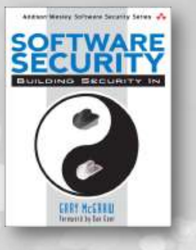
Shokri, R., M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. 2017 IEEE Symp. Security Privacy*, 2017, pp. 3–18.

where to learn more

build security in

- Writings, Blogs, Music
  https://garymcgraw.com
- BIML
  https://berryvilleiml.com/
- Send e-mail:
  gem@garymcgraw.com

BIML



@cigitalgem@sigmoid.social