Machine Learning, Bias, and Hype

CSE 120, Winter 2020

Slides from Prof. Noah Smith (nasmith@cs.washington.edu)

Administrivia

- Assignments
  - Arrays & Elli: check off by the end of the day today
  - Word Guessing: checkoff by Tuesday (Feb 18)
  - Controlling Elli: submit by Friday (Feb 21)

- We’re done with new Processing material in lecture! 🥳
  - “Big Ideas” lectures for the rest of the quarter
  - See course calendar for a sneak peak – should be fun!

living computers museum + labs Report

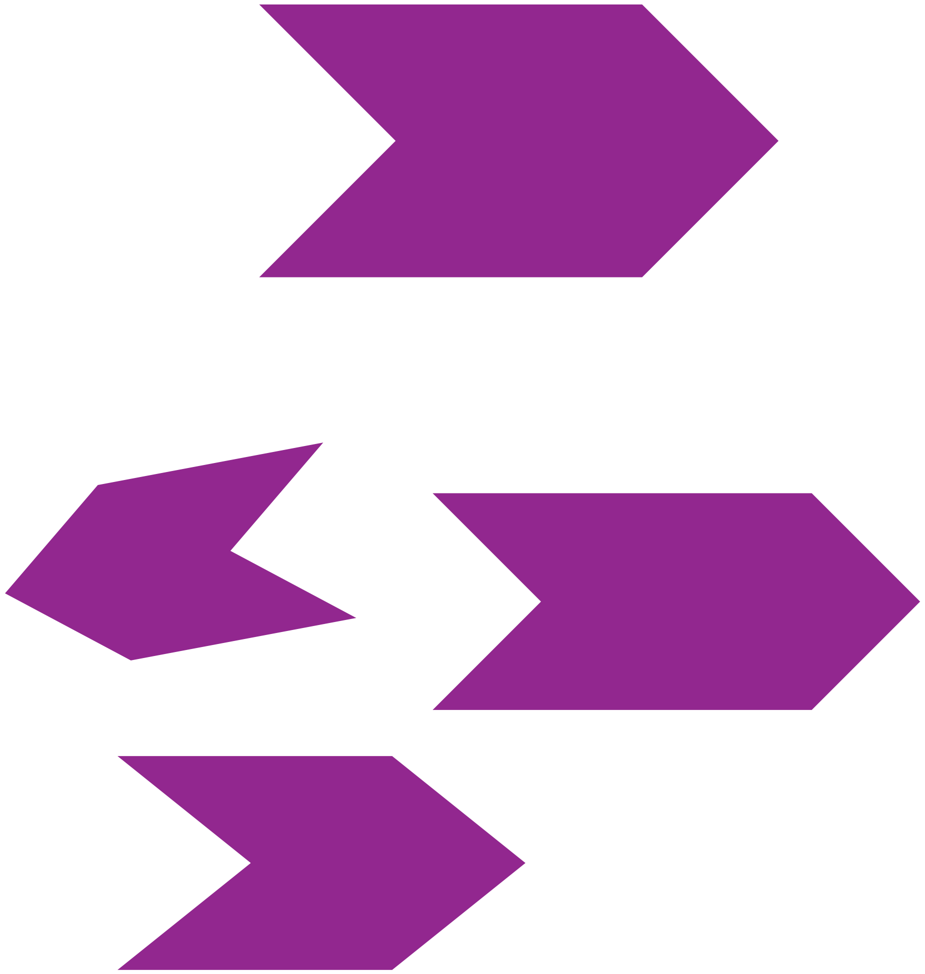
- Field trip out to the Living Computers: Museum + Labs in SoDo
  - Admission is paid for you!
  - Transportation: Link + walk, bus, drive
  - Go when you can: open Wed-Sun each week

- **Report:** PDF including photos and responses due Mar 2
  - Part 1: Favorite Exhibit
  - Part 2: Computer History
  - Part 3: Modern Tech Exhibit Reflection

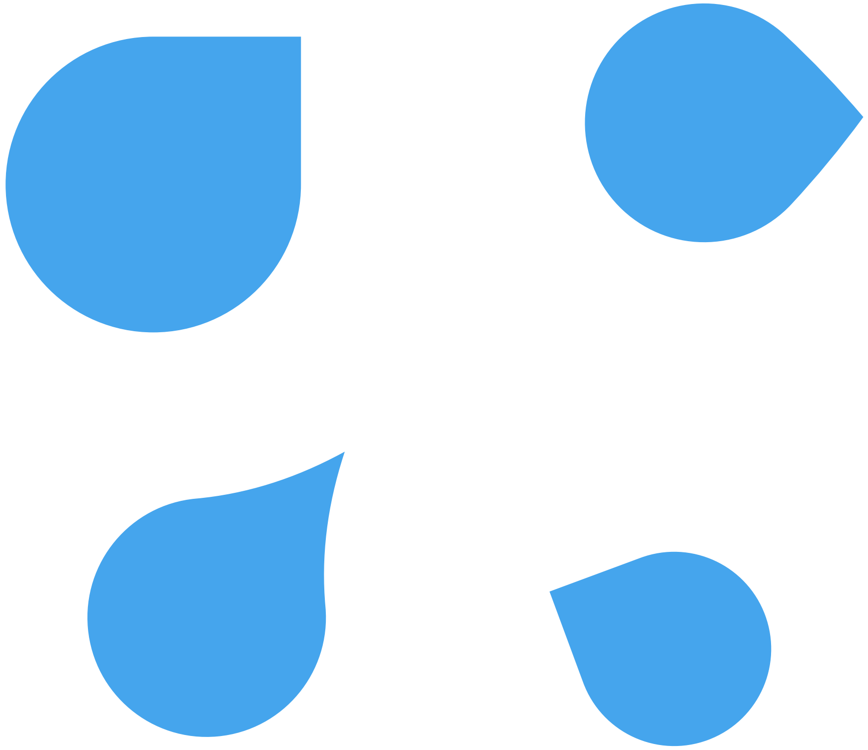
Outline

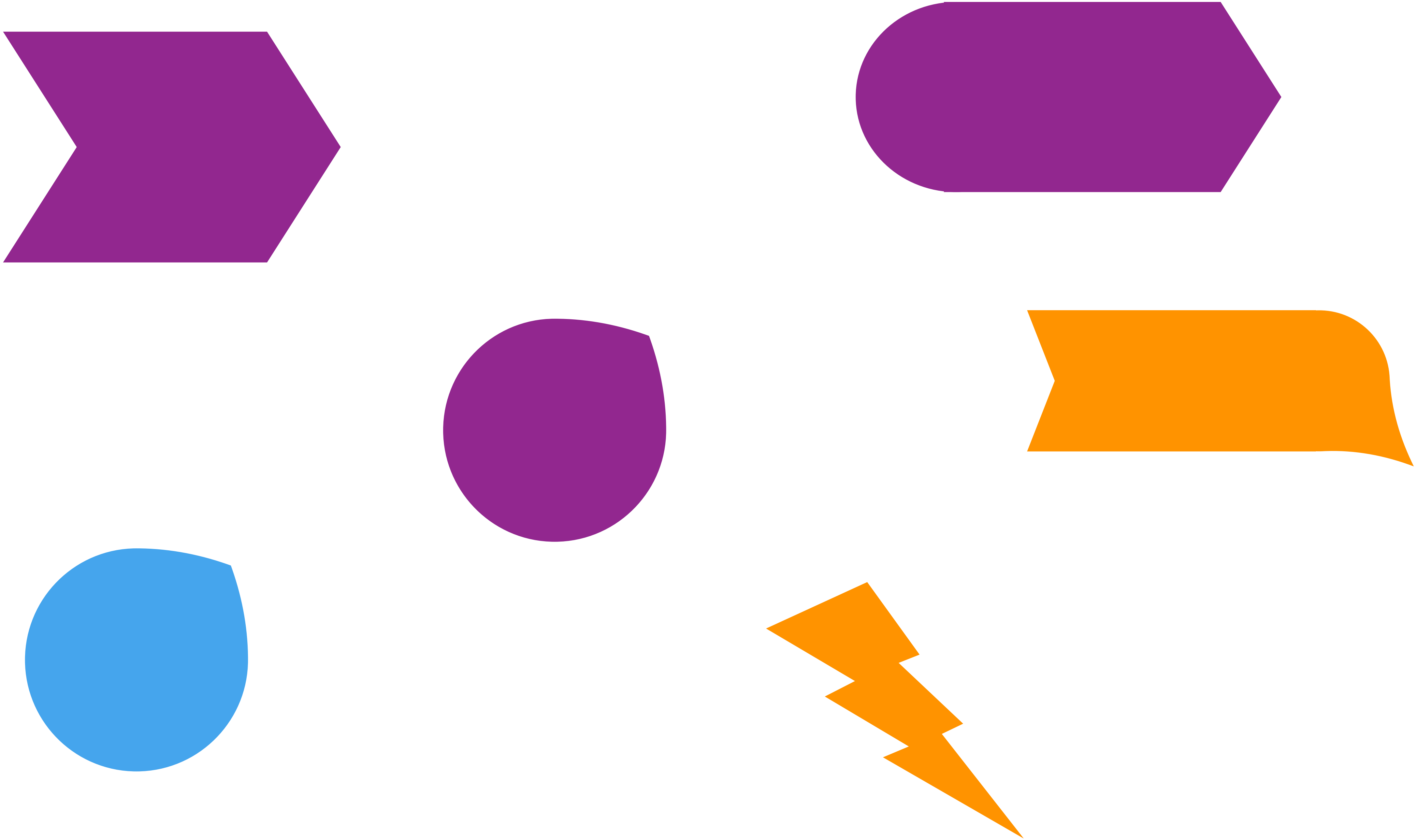
1. Basic introduction to machine learning
2. Bias in machine learning
3. Inoculation against AI hype

blickets

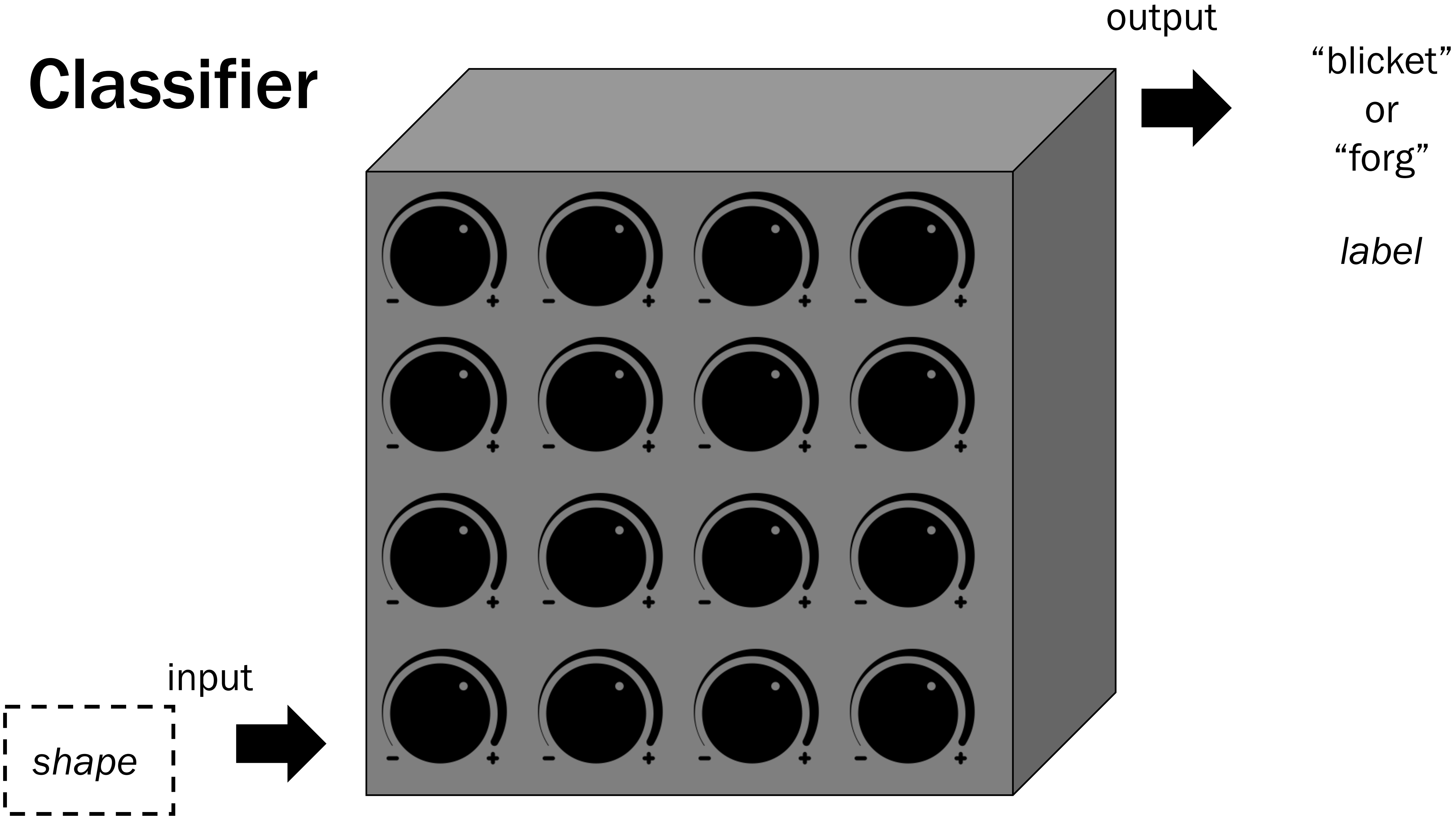


forgs

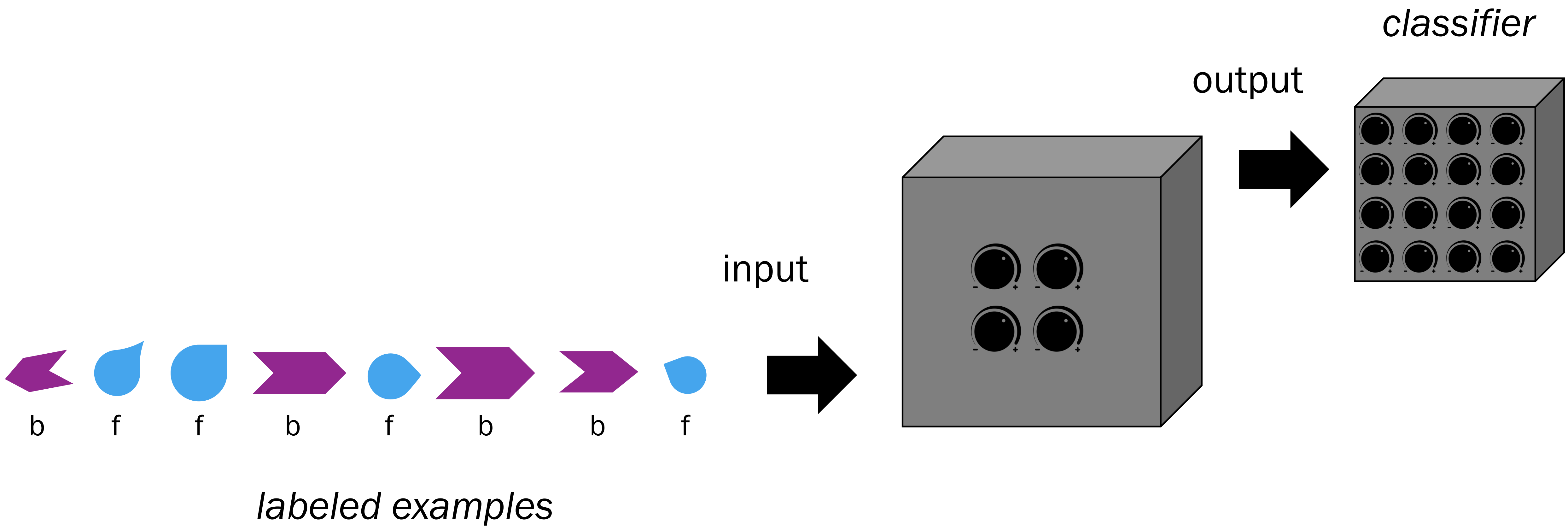




Classifier



Supervised Learner



Some Secrets about Supervised Learning

- The data matter **a lot**

- How we represent the data as an “input” matters **a lot**

- Sources of error in generalizing to new (non-training) examples:
  - Flaws in our representation of the problem (“irreducible”)
  - Assumptions made by a learning algorithm (“bias”)
  - Randomness/noise in the data (“variance”)

- There is a **tradeoff** between bias and variance!

On Bias

- Bias is prejudice or preference held prior to exposure to evidence (held by a human or a program)

- Learners cannot generalize without (inductive) bias!

- Put another way: if you eliminate all bias, your model will be extremely *flexible* and will tend to be extremely sensitive to the particular training instances.
    - Result: higher variance, unless there's "enough" data

Examples of Bias

Input	Output	Result
image of tank	American or Russian?	
tweet	abusive?	
speech stream	sequence of words	
details about person convicted of a crime	sentence length	
two English sentences	semantic relationship (entailment, contradiction, ...)	
product reviews	sentiment of author	

Examples of Bias

Input	Output	Result
image of tank	American or Russian?	clear/blurry
tweet	abusive?	
speech stream	sequence of words	
details about person convicted of a crime	sentence length	
two English sentences	semantic relationship (entailment, contradiction, ...)	
product reviews	sentiment of author	

Examples of Bias

Input	Output	Result
image of tank	American or Russian?	clear/blurry
tweet	abusive?	AAVE
speech stream	sequence of words	
details about person convicted of a crime	sentence length	
two English sentences	semantic relationship (entailment, contradiction, …)	
product reviews	sentiment of author	

Examples of Bias

Input	Output	Result
image of tank	American or Russian?	clear/blurry
tweet	abusive?	AAVE
speech stream	sequence of words	only worked for men
details about person convicted of a crime	sentence length	
two English sentences	semantic relationship (entailment, contradiction, ...)	
product reviews	sentiment of author	

Examples of Bias

Input	Output	Result
image of tank	American or Russian?	clear/blurry
tweet	abusive?	AAVE
speech stream	sequence of words	only worked for men
details about person convicted of a crime	sentence length	longer sentences for minorities
two English sentences	semantic relationship (entailment, contradiction, ...)	
product reviews	sentiment of author	

Examples of Bias

Input	Output	Result
image of tank	American or Russian?	clear/blurry
tweet	abusive?	AAVE
speech stream	sequence of words	only worked for men
details about person convicted of a crime	sentence length	longer sentences for minorities
two English sentences	semantic relationship (entailment, contradiction, ...)	“cat” ⟶ contradiction
product reviews	sentiment of author	

Examples of Bias

Input	Output	Result
image of tank	American or Russian?	clear/blurry
tweet	abusive?	AAVE
speech stream	sequence of words	only worked for men
details about person convicted of a crime	sentence length	longer sentences for minorities
two English sentences	semantic relationship (entailment, contradiction, ...)	“cat” ⟶ contradiction
product reviews	sentiment of author	fails on political speech

Where does bias come from?

1. The real-world process that produced the labels, or the data sample, might be biased.
   Just because something comes from data, that doesn't mean it's "fair" or "unbiased"!
2. The design/definition of the task might encode bias.
3. The design of the program itself might encode bias.
4. Deployed systems that affect their own future inputs can create feedback loops and exacerbate their own biases.

Disparate Impact

- US law (hiring and housing): 80% rule

  Informally: your rate of hiring women (for instance) must be at least 80% of your rate of hiring men.

- Can we just hide the sex attribute from the learner?

No!

- There are many alternative definitions of fairness.
- Open question: can we guarantee high accuracy and still be unbiased?

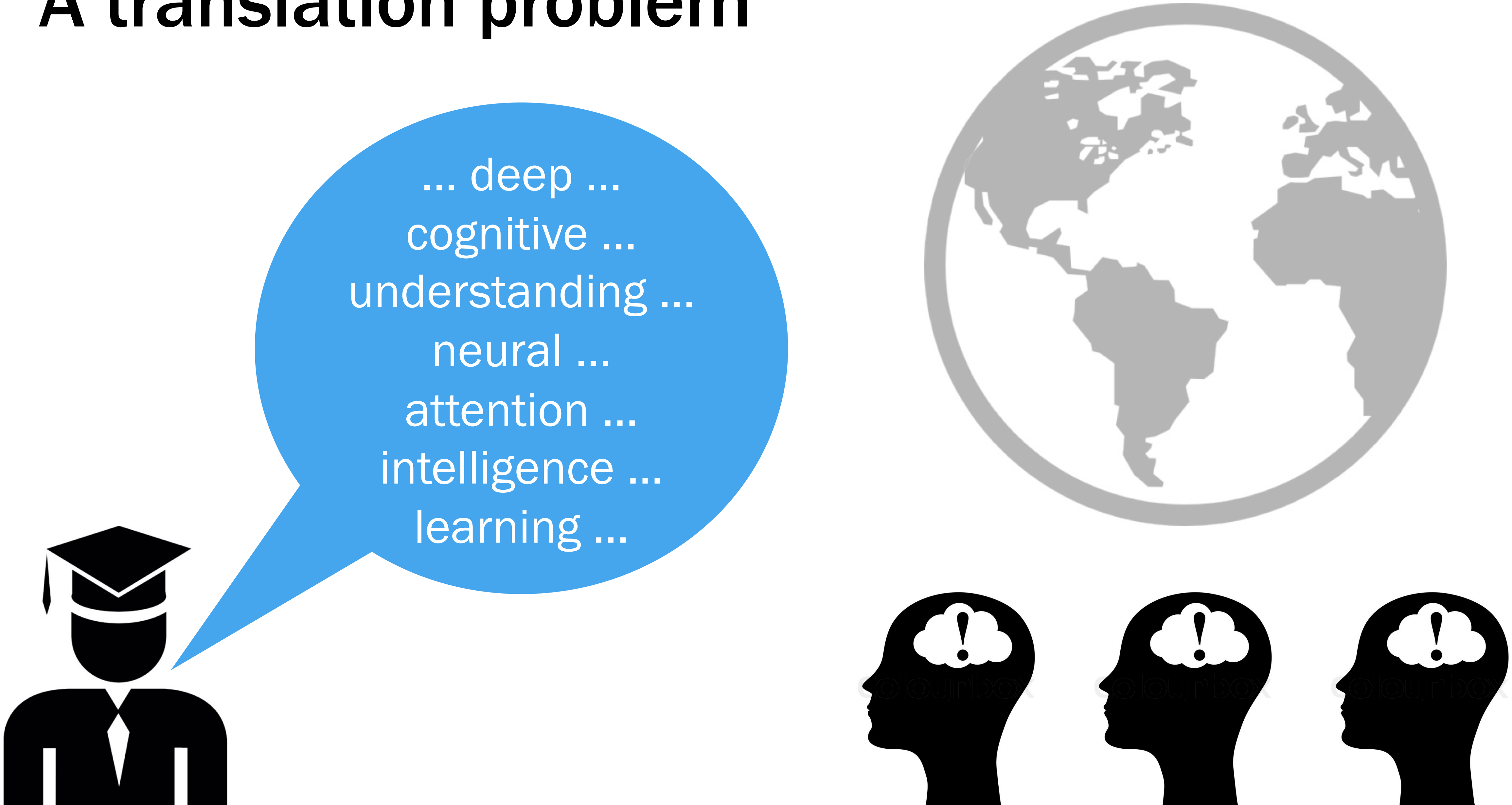
We aren’t aware of all the biases!

- Typically we measure the accuracy of learned programs: what proportion of inputs do they correctly label, in a held-out test set?
  - Sometimes we look at accuracy for particular subcategories.

- We don’t always know which biases to look for!

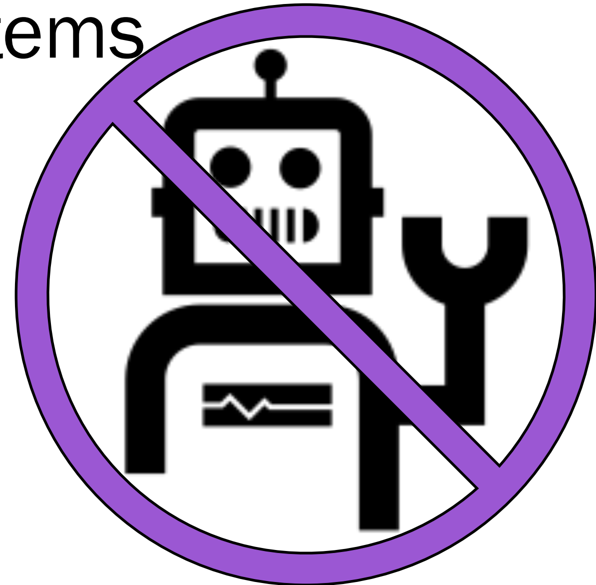
Inoculation against Hype

A translation problem



Tips

- ✓ "Human level performance" has a *very narrow* meaning
- ✓ "95% accuracy" was measured only on a *specific* type of input
- ✓ Ask about the data and computation requirements (i.e., cost)
- ✓ Researchers' benchmarks are *not* real-world systems
- ✓ Do not trust anthropomorphic descriptions of systems



Learn More

- *A Course in Machine Learning*, by Hal Daumé III. http://ciml.info
- CSE 416 or 446