UNIVERSITÀ DI PAVIA
Department of Electrical, Computer
and Biomedical Engineering

**Steering AI:
Legal Challenges
and Ethical Standards
from an Engineering
Perspective**

Thursday 18 April 2024

Trusting the Algorithms

The Many Shades of Bias

Marco Piastra

The Many Shades of Bias

- **Observer Bias**
- **Bias in Data**
- **Inductive Bias**

Observer Bias *Choosing the Right Metaphors*

The Economist

Culture | Johnson

Talking about AI in human terms is natural—but wrong

When it comes to artificial intelligence, metaphors are often misleading

IMAGE: NICK LOWNDES

Jun 22nd 2023 Save Share Give

[https://www.economist.com/culture/2023/06/22/talking-about-ai-in-human-terms-is-natural-but-wrong]

Observer Bias *If the Present is Confusing, the Future is Uncertain...*

OBSERVER

BUSINESS

Meta's A.I. Chief Yann LeCun Explains Why a House Cat Is Smarter Than The Best A.I.

"A cat can remember, can understand the physical world, can plan complex actions, can do some level of reasoning—actually much better than the biggest LLMs."

By Sissi Cao · 02/15/24 3:24pm

Yann LeCun testifies before the U.S. Senate Intelligence Committee on September 19, 2023 in Washington, D.C. Kevin Dietsch/Getty

[https://observer.com/2024/02/metas-a-i-chief-yann-lecun-explains-why-a-house-cat-is-smarter-than-the-best-a-i/]

FINANCIAL TIMES

Artificial intelligence + Add to myFT

Elon Musk predicts AI will overtake human intelligence next year

Tesla chief says infrastructure will need to keep up with technology's demands as he seeks investment for own start-up

Elon Musk had previously suggested AI would surpass human intelligence by 2029 © Kirsty Wigglesworth/AP

George Hammond in San Francisco APRIL 8 2024 453

[https://www.ft.com/content/027b133f-f7e3-459d-95bf-8afd815ae23d]

The New York Times

The Ezra Klein Show

April 12, 2024

What if Dario Amodei Is Right About A.I.?

Anthropic's co-founder and C.E.O. explains why he thinks artificial intelligence is on an "exponential curve."

[https://www.nytimes.com/2024/04/12/opinion/ezra-klein-podcast-dario-amodei.html]

AI-Specific Traits

- **Infinitely repeatable**

  Once trained, an AI model provides *deterministic* inference

- **Completely observable mechanisms**

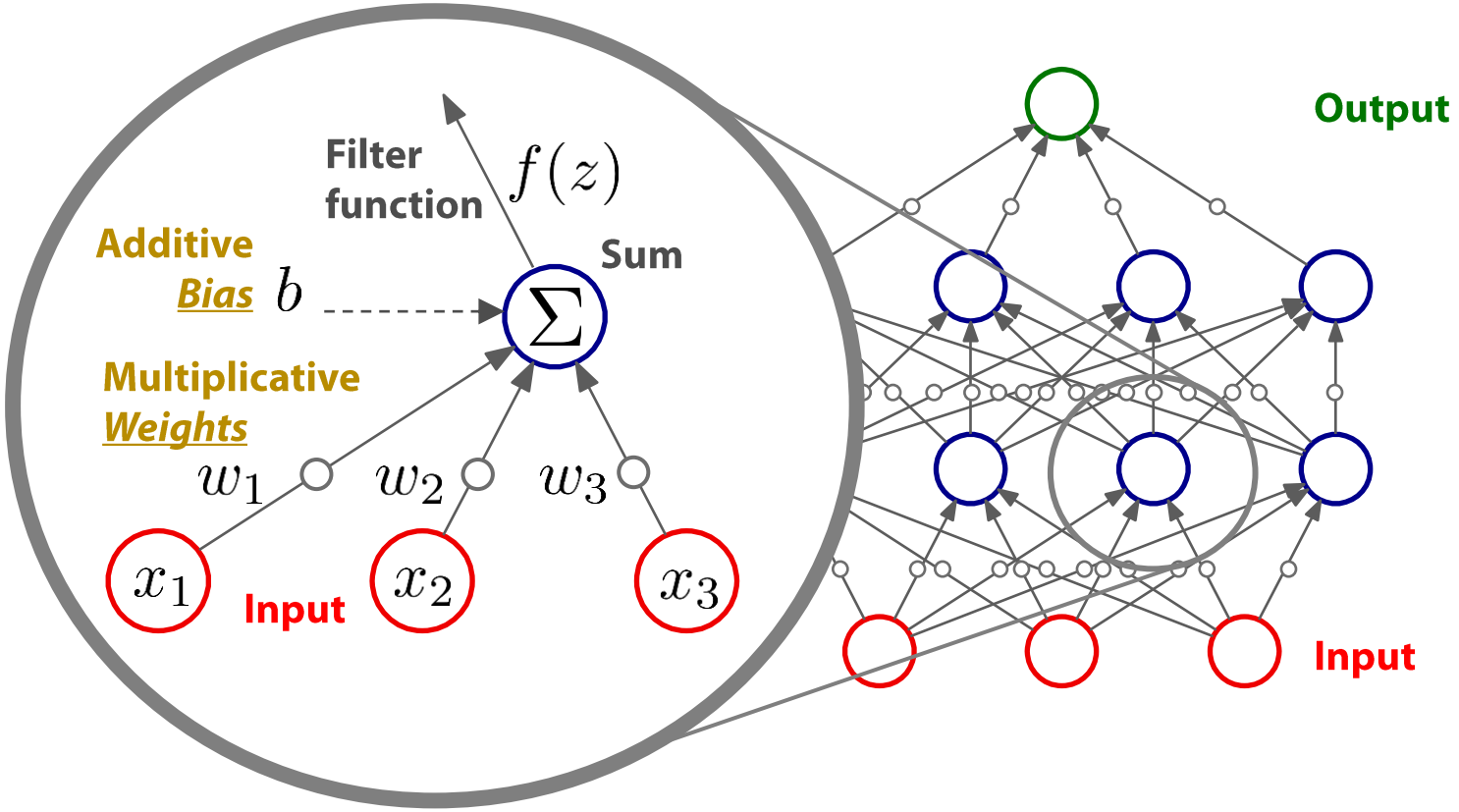
  Interpretation may be challenging, yet computations can be observed to the last bit

- **Duplicable**

  AI systems are *software*:
  they may require substantial hardware resources, but they can be replicated at will

Artificial Neural Networks *At the Core of AI Revolution*

An assembly of simple computational units

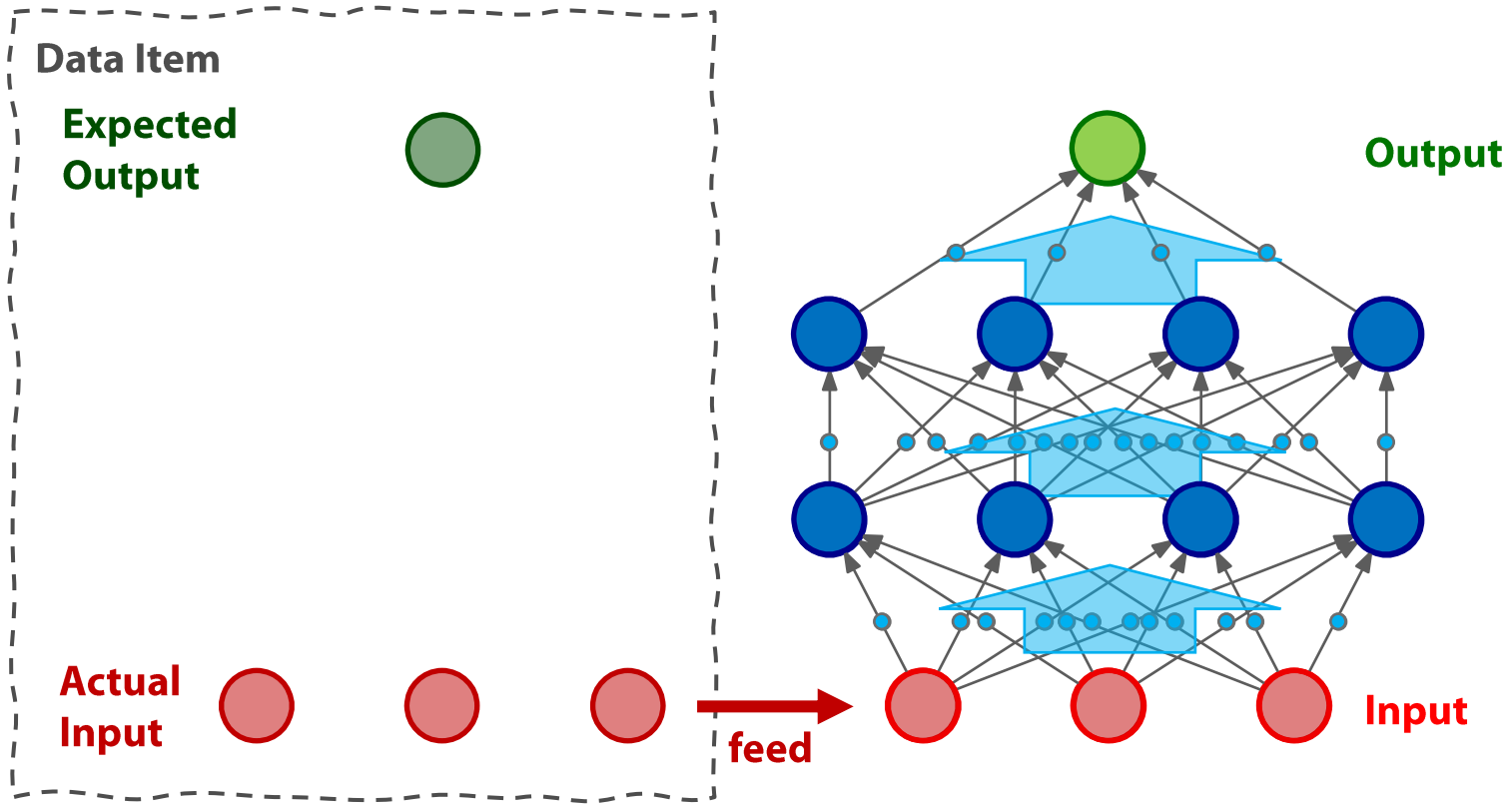


Each unit performs simple arithmetic operations

Weights and Biases are the only mutable parts

Learning from Data

Incremental, numerical optimization

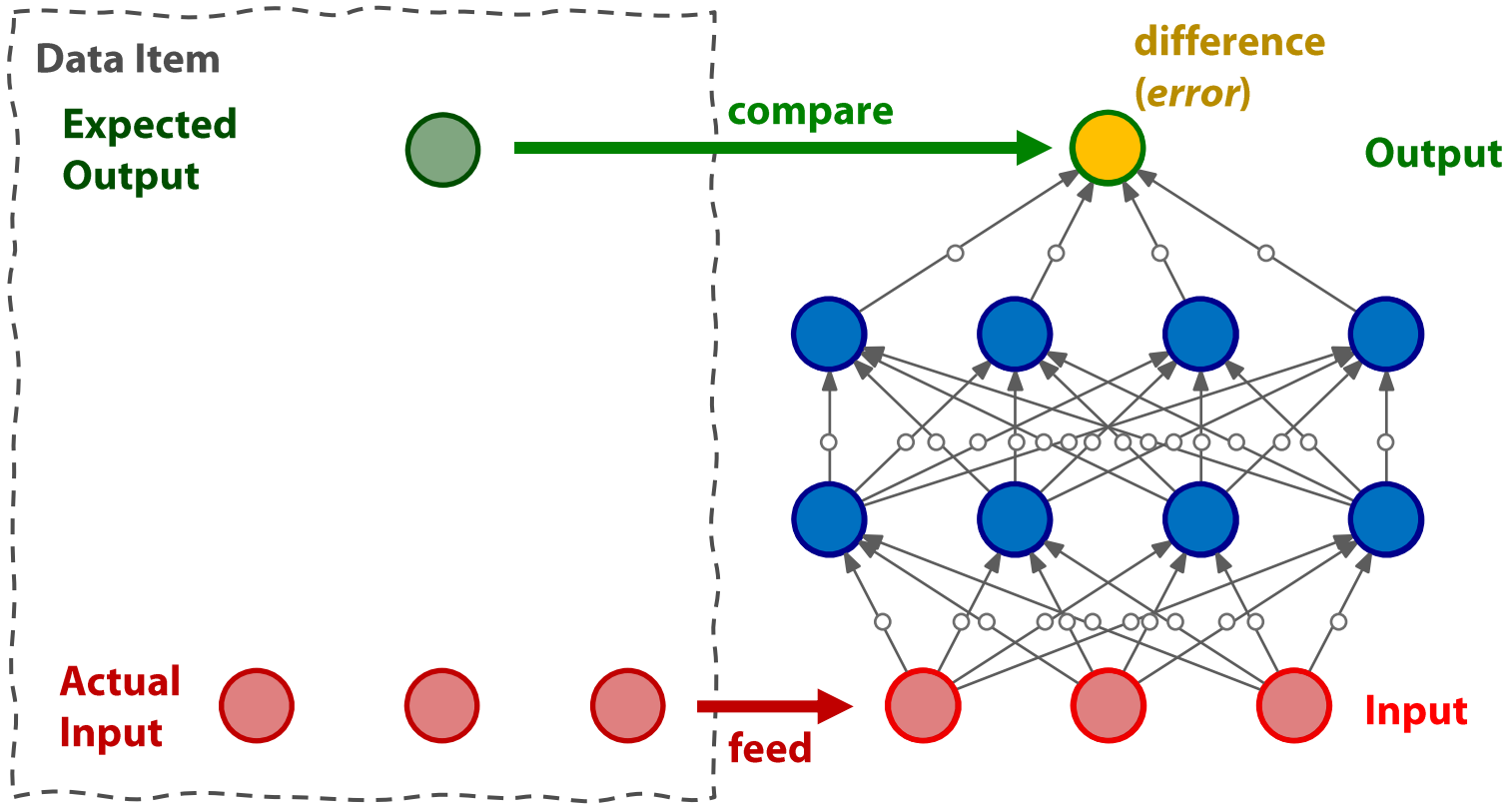


Learning from data items: input + output

Input is fed to the network and the output is computed

Learning from Data

Incremental, numerical optimization

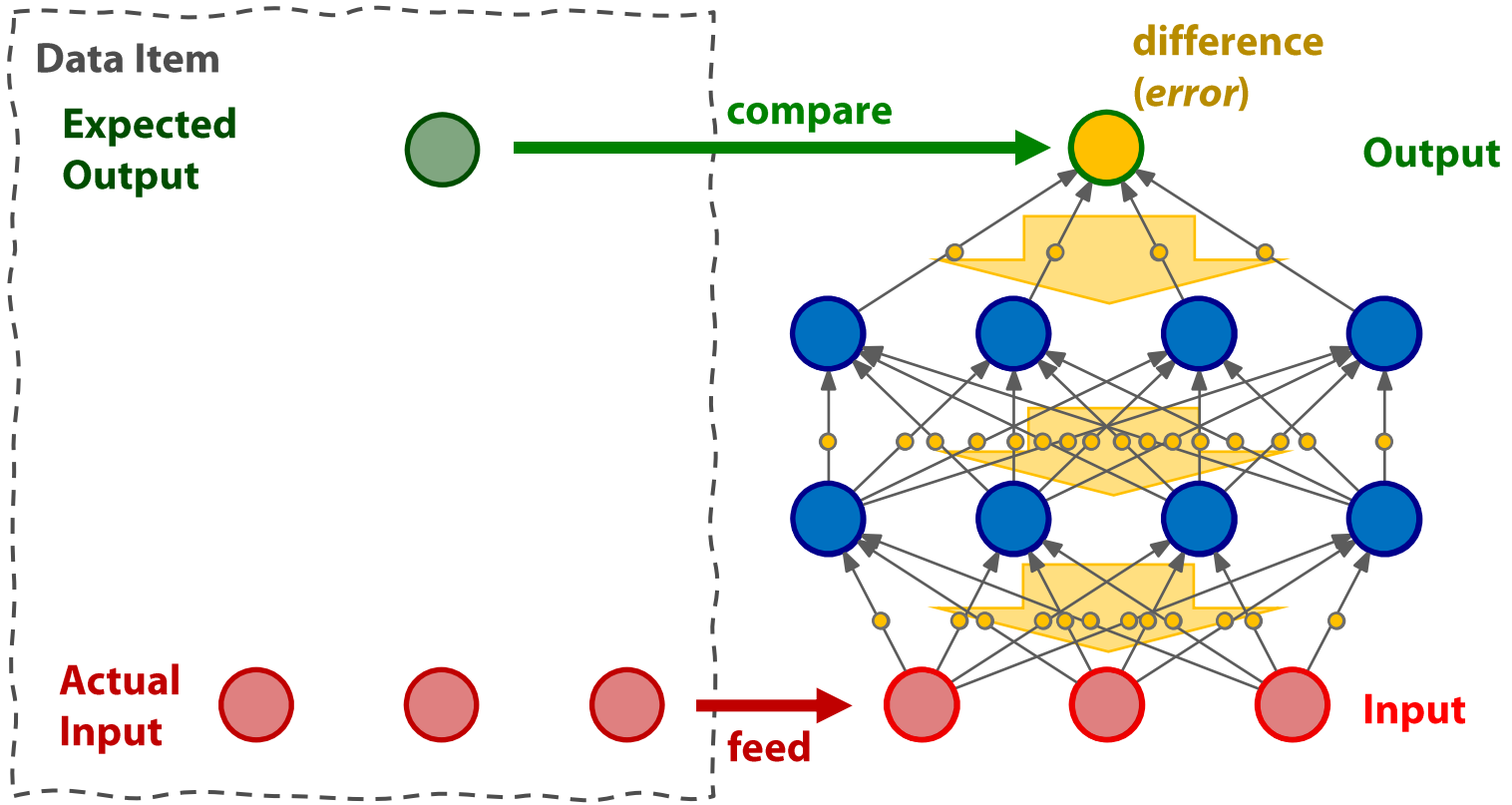


Output is compared with expected value

Error is estimated from difference

Learning from Data

Incremental, numerical optimization

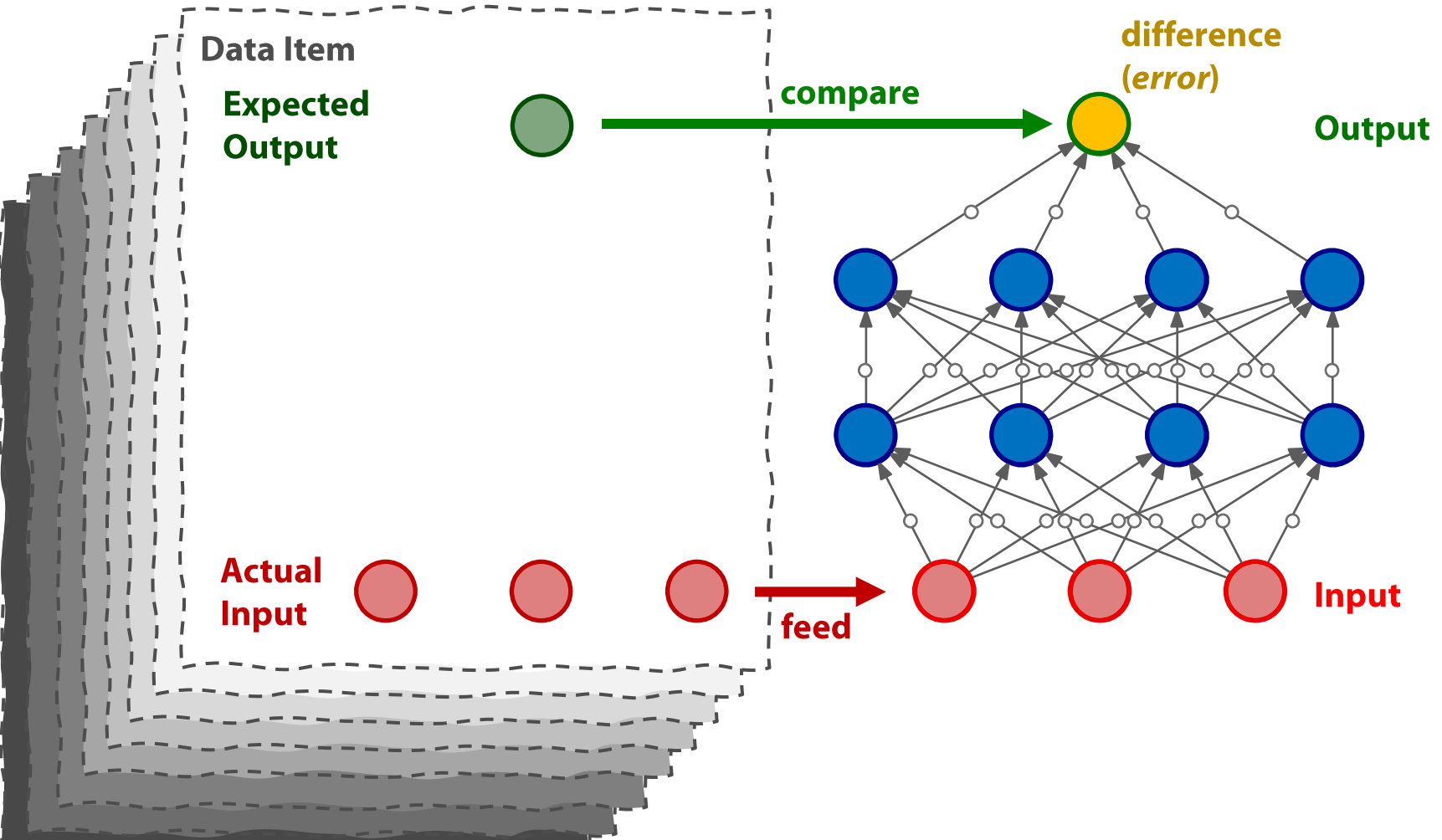


Error is propagated in the opposite direction

To change Weights and Biases

Learning from Data

Incremental, numerical optimization



The process is repeated on huge datasets

ChatGPT *Its Transformer-Based Precursor*

Sentences in input (*prompt*) are fed word by word to a very sophisticated and complex *artificial neural network*

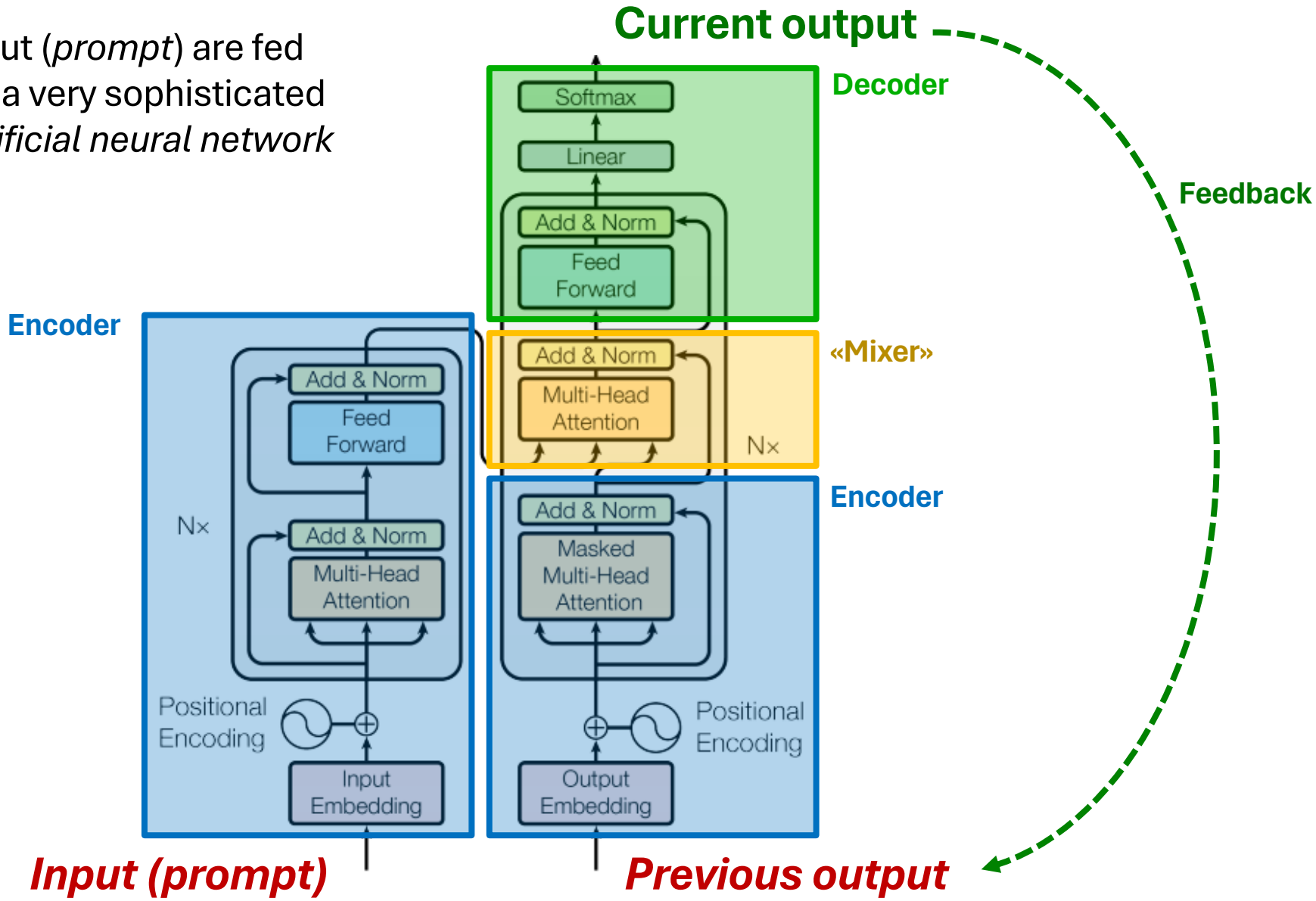
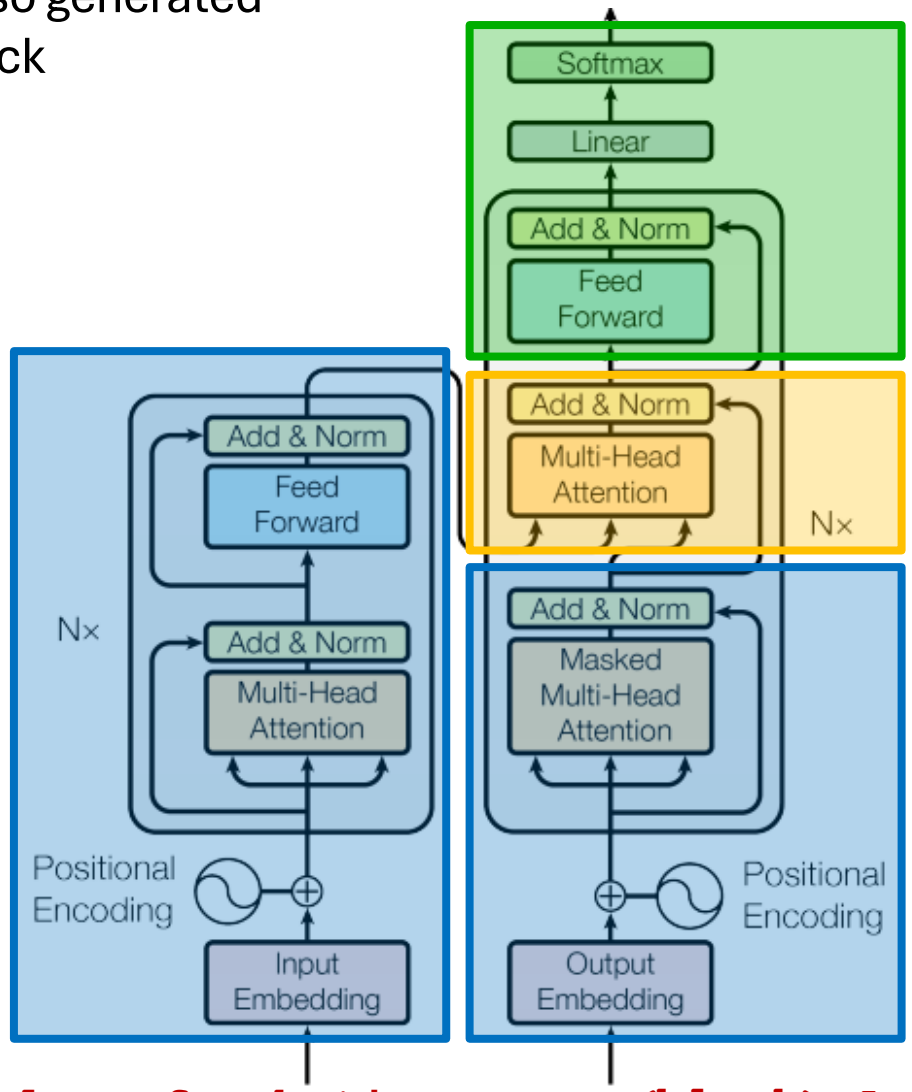


Image from https://arxiv.org/abs/1706.03762

ChatGPT *Its Transformer-Based Precursor*

Output sentences are also generated word by word and fed back to the network



A robot may not injure a human being

Feedback

Please recite the first law of robotics

<blank> A robot may not injure a human

Image from https://arxiv.org/abs/1706.03762

ChatGPT *Its Transformer-Based Precursor*

Actual network output
is a **probability distribution**
over the next word *(token)*

The whole network is huge
(GPT-3: 175 billion parameters)

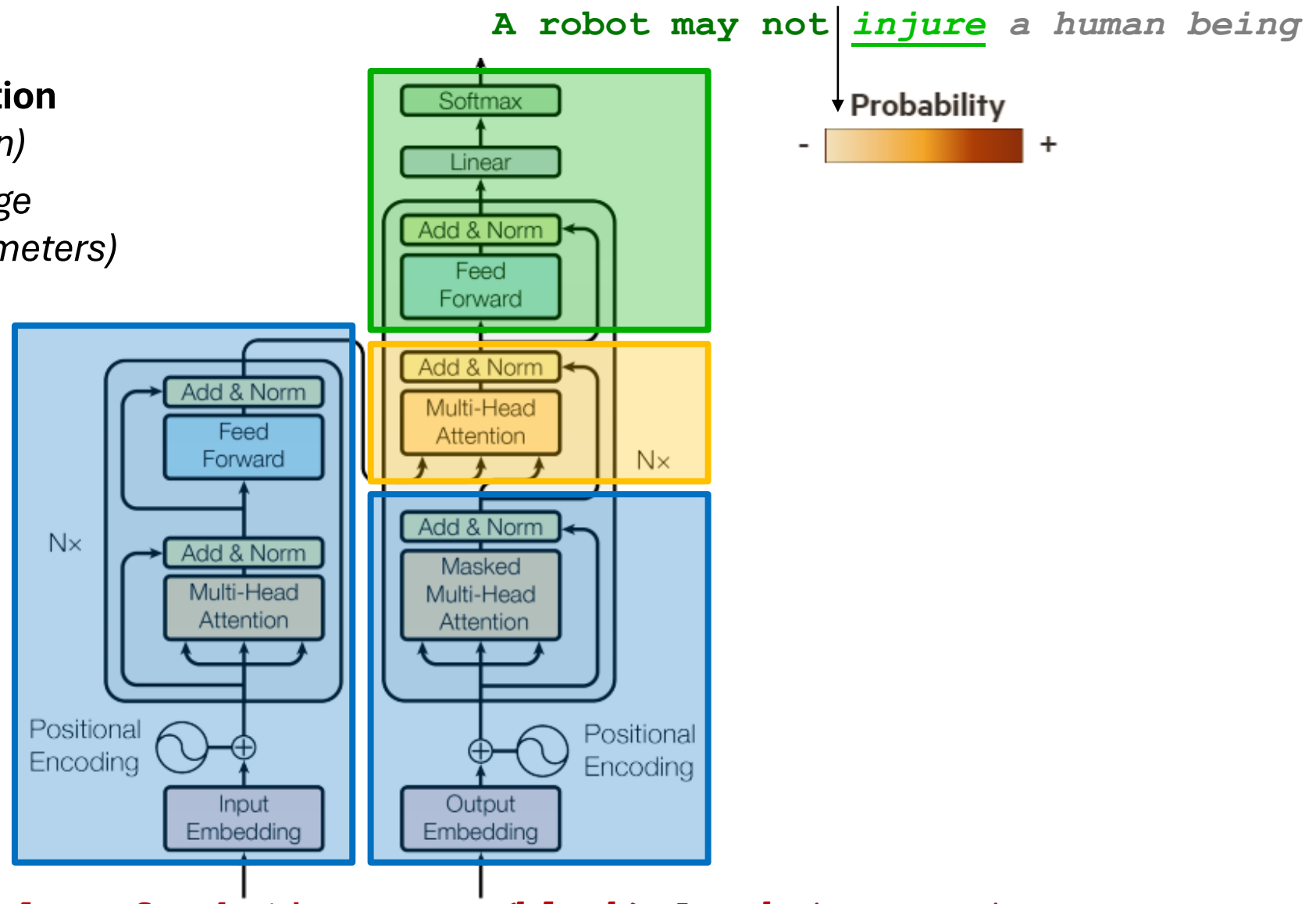
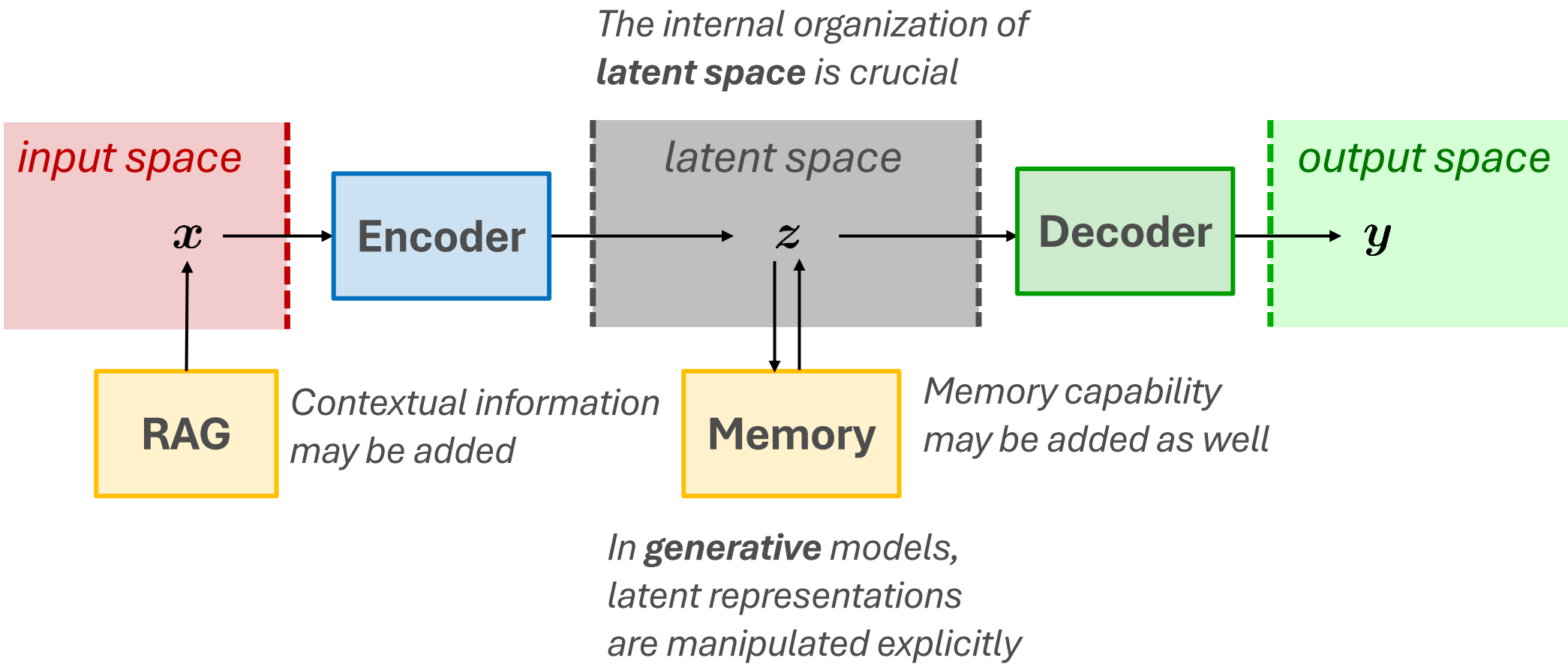


Image from https://arxiv.org/abs/1706.03762

Encoder - Decoder *A Very Popular Architectural Pattern*

Input is first translated into a **latent** (a.k.a. *hidden*, *intermediate*) **representation** and then translated into a meaningful output

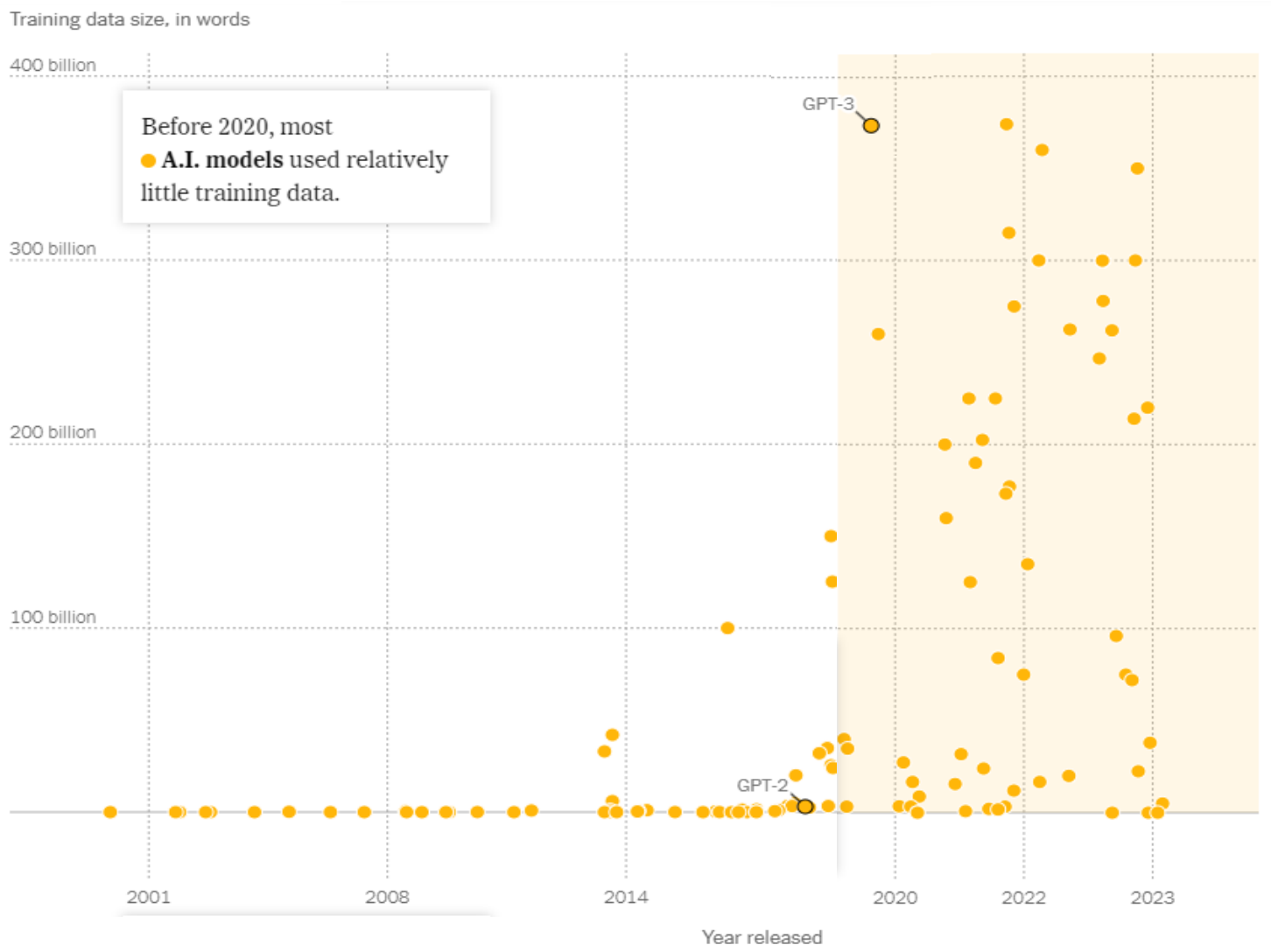
This architectural pattern is common in ***Generative AI***



Foundation Models *A Never-Ending Quest for Even More Data*

Foundation models (a.k.a. *pre-trained*, *zero-shot*) can be used ‘off-the-shelf’ without further training

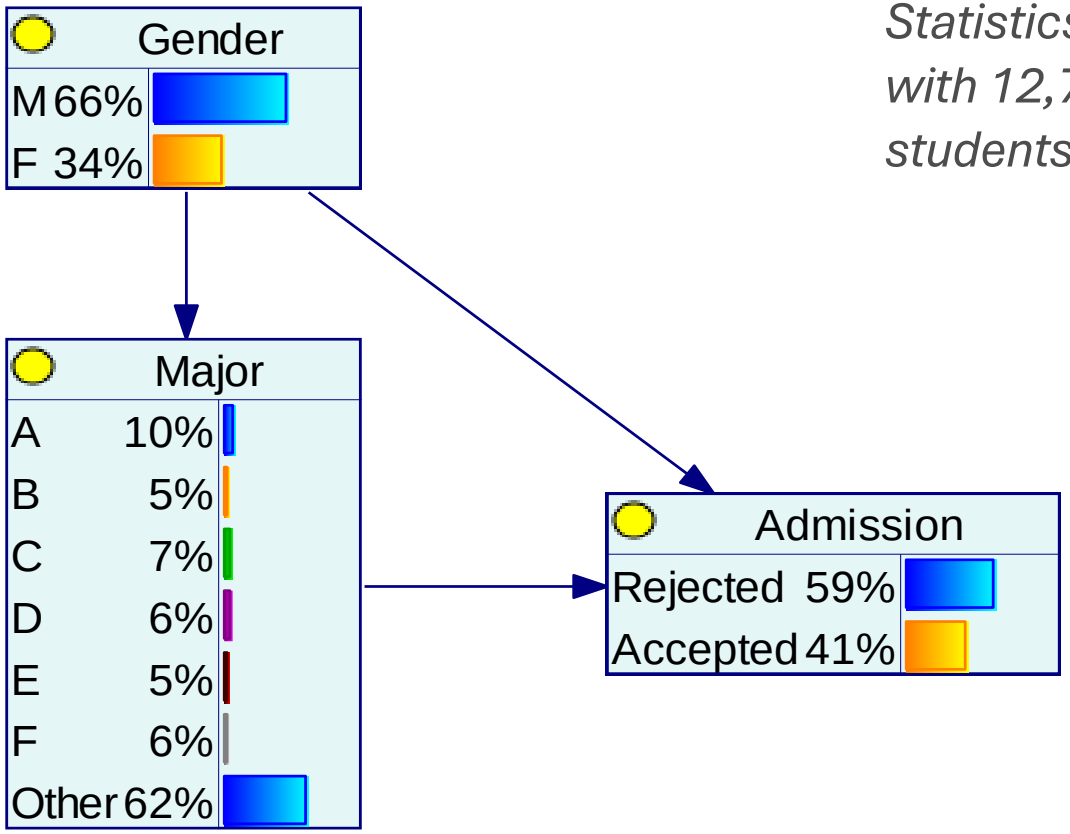
To achieve this, at present, an *enormous* amount of data is required



[https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html]

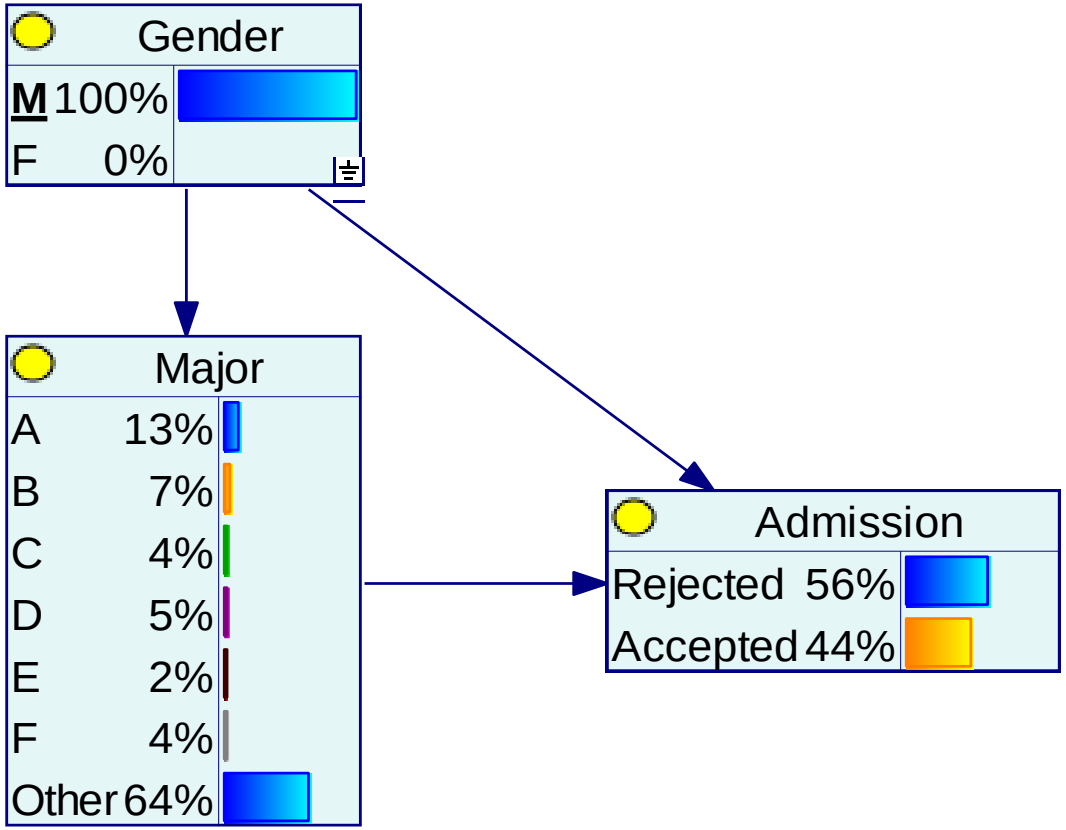
Bias in Data

Bias in Data *An Example* *(Berkeley Admission Test, 1973)*

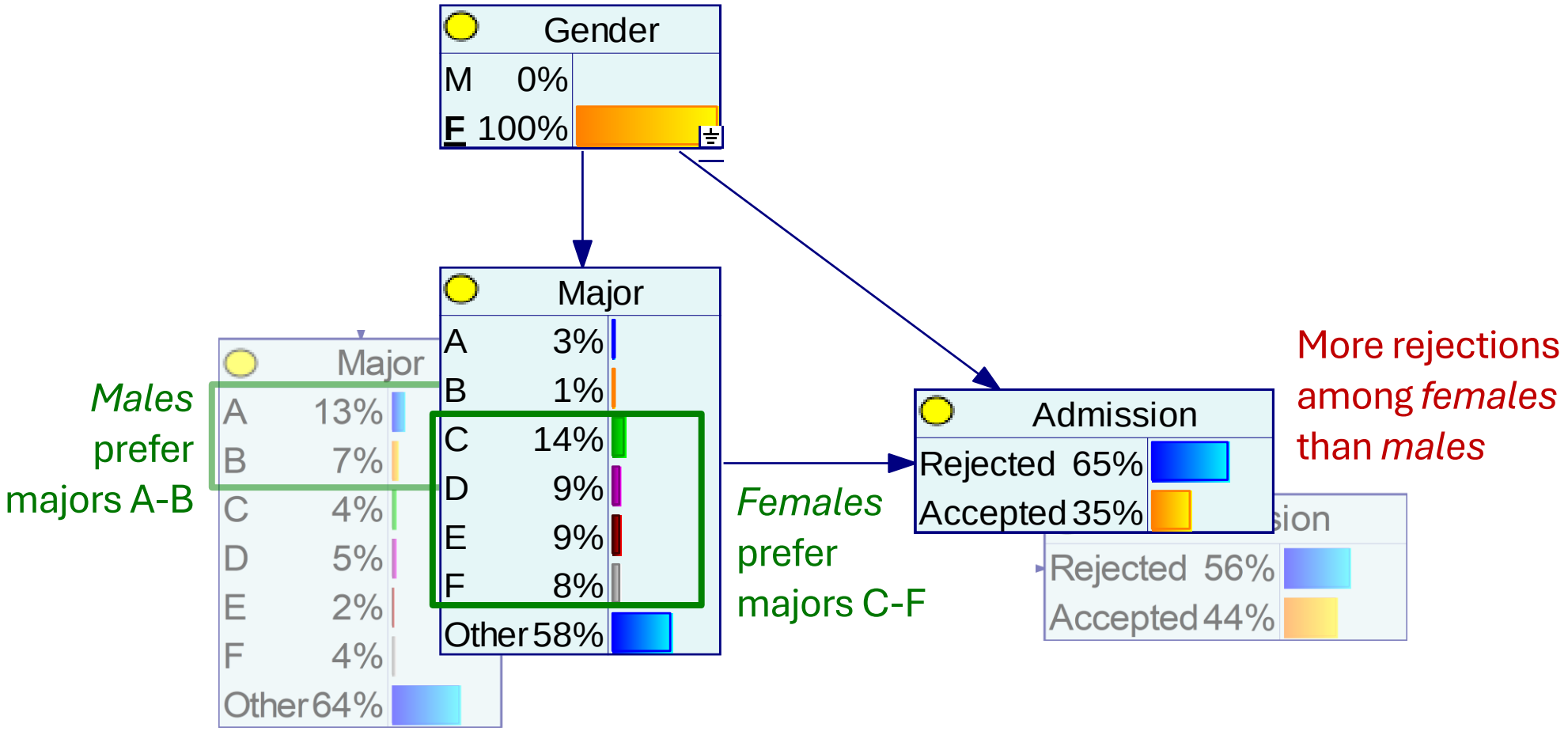


Statistics collected with 12,763 candidate students

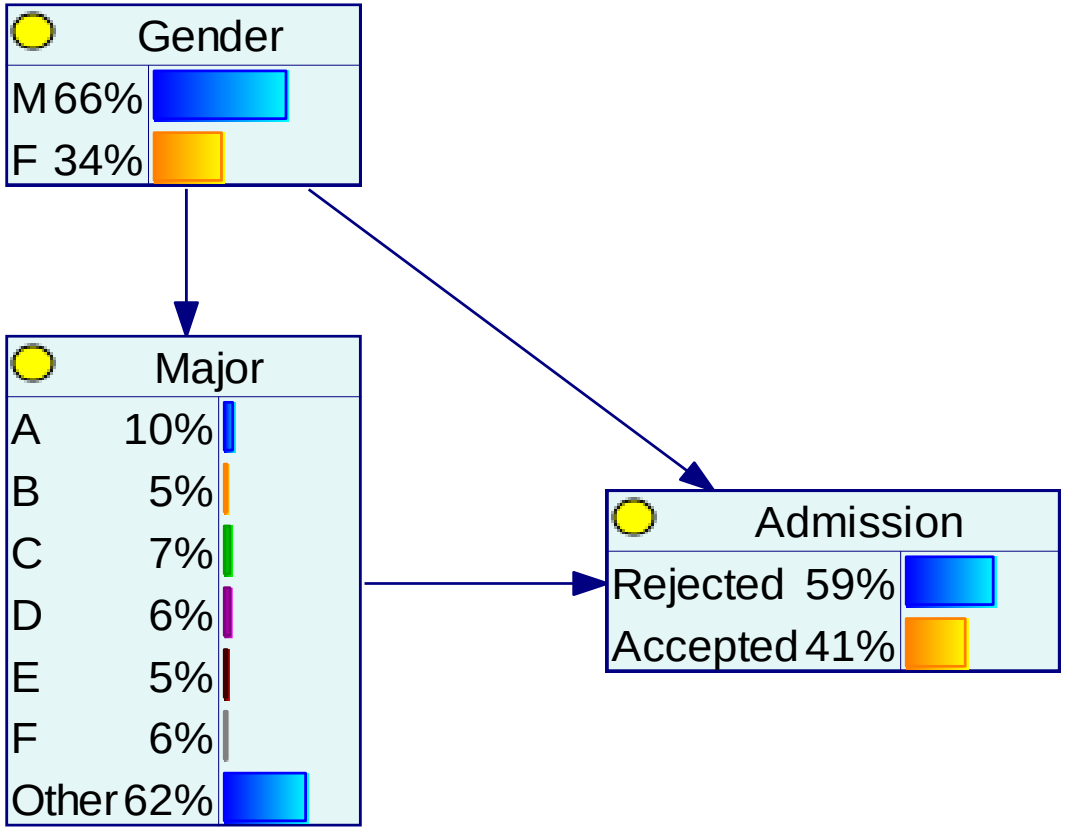
Bias in Data *An Example* *(Berkeley Admission Test, 1973)*



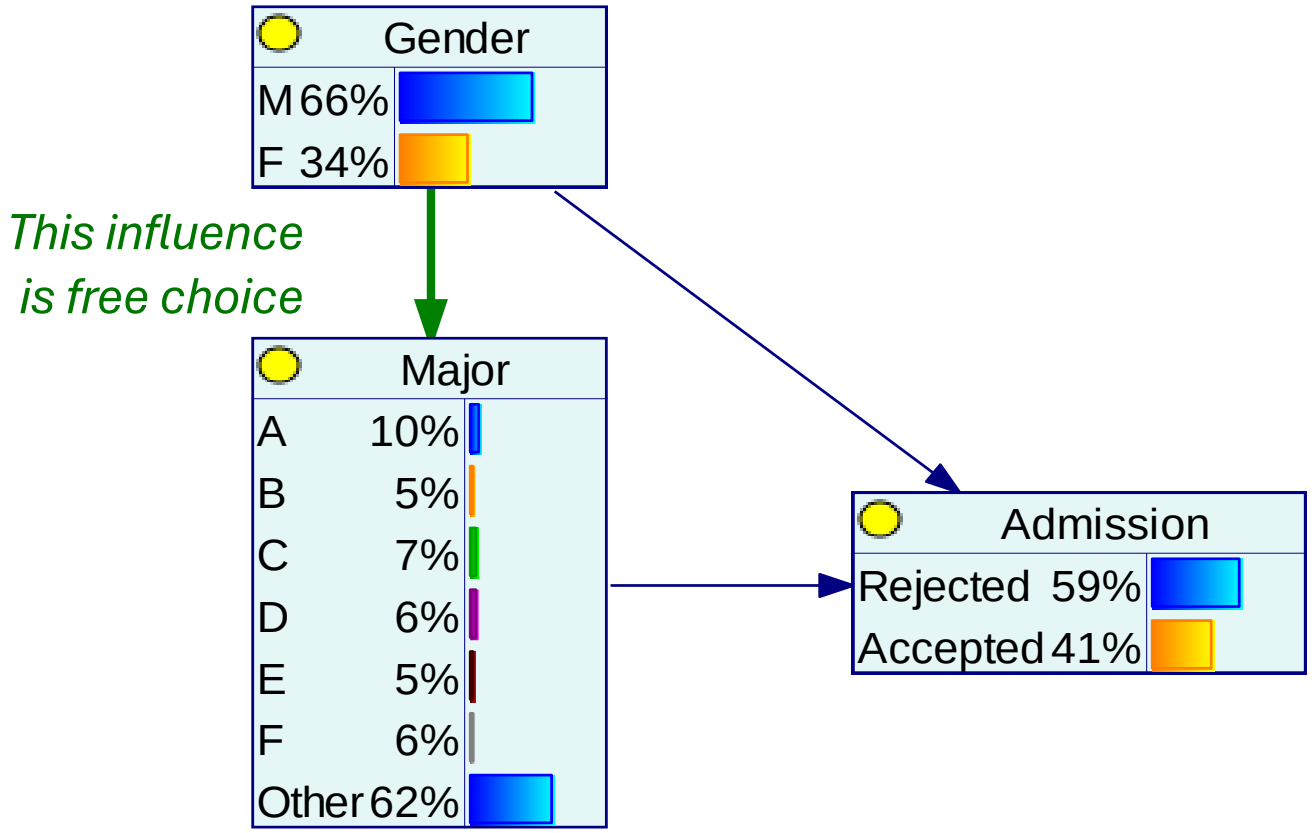
Bias in Data *An Example* *(Berkeley Admission Test, 1973)*



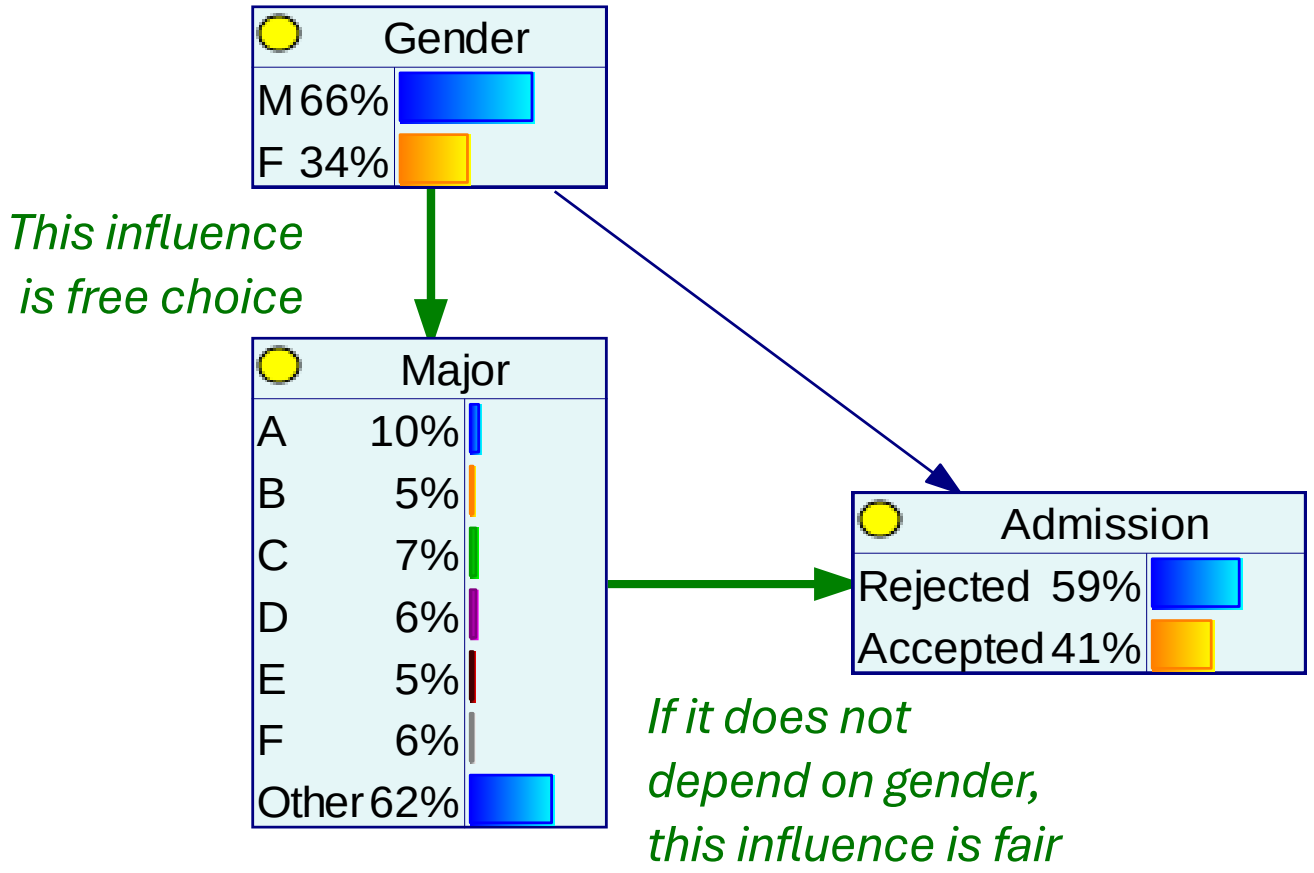
Bias in Data *An Example* *(Berkeley Admission Test, 1973)*



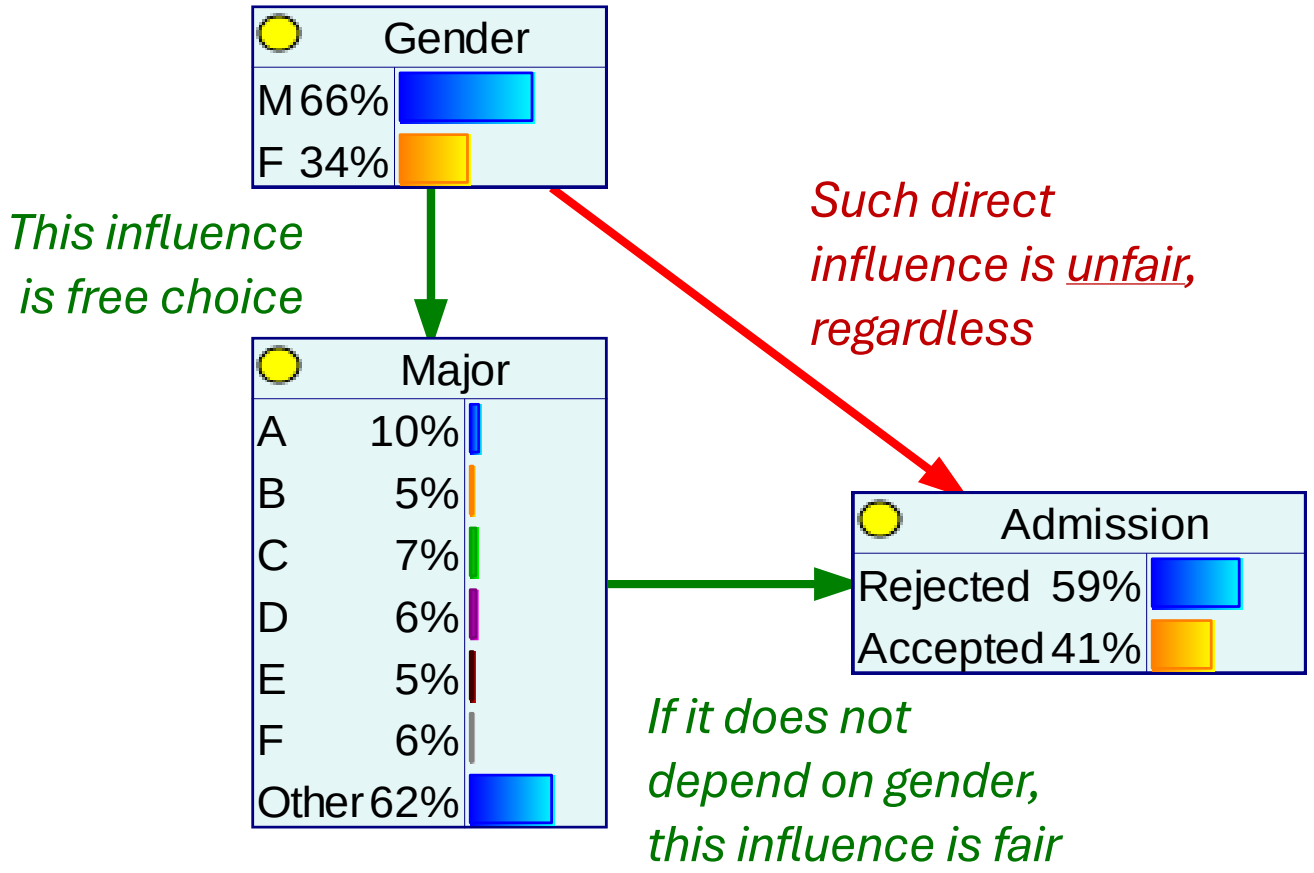
Bias in Data *An Example* *(Berkeley Admission Test, 1973)*



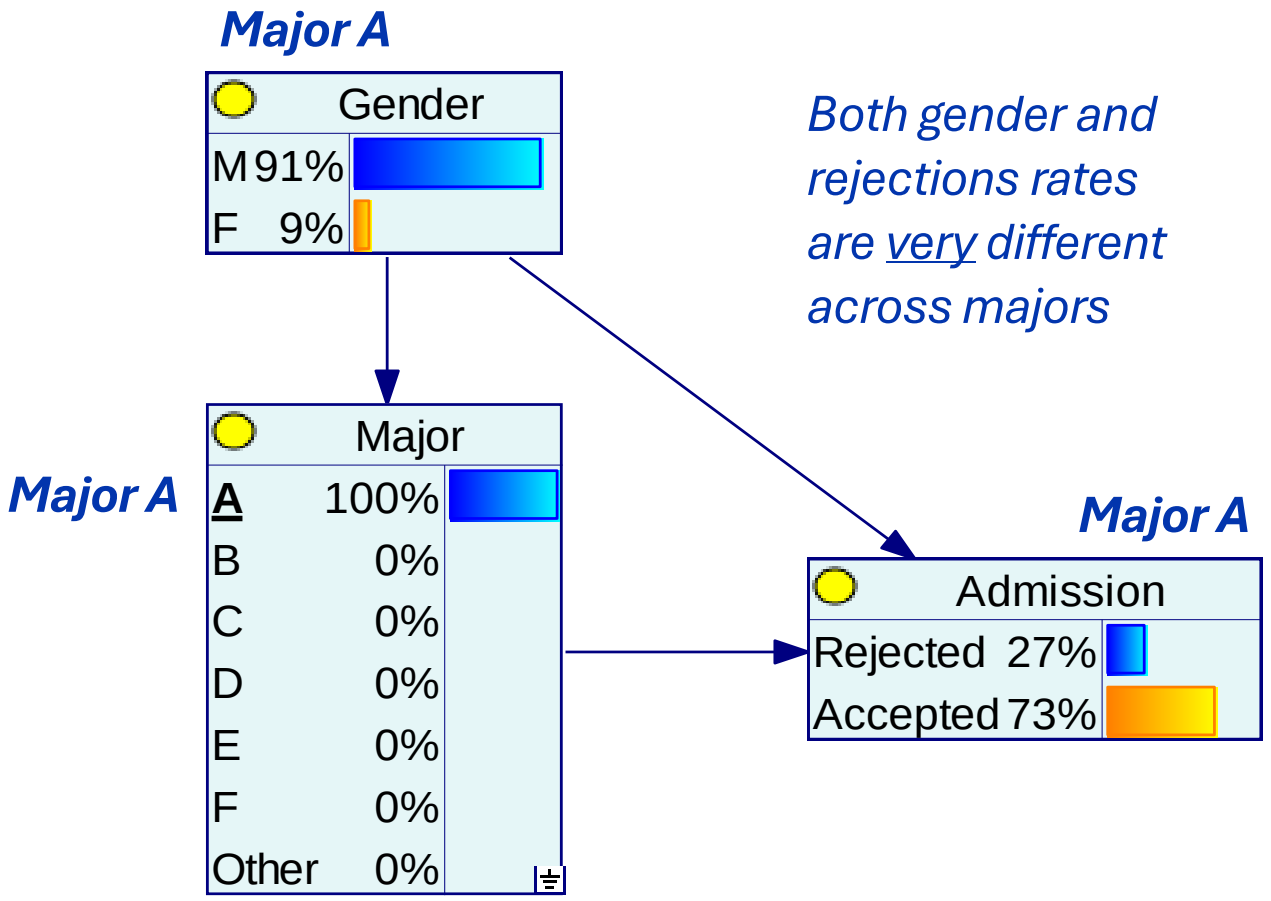
Bias in Data *An Example* *(Berkeley Admission Test, 1973)*



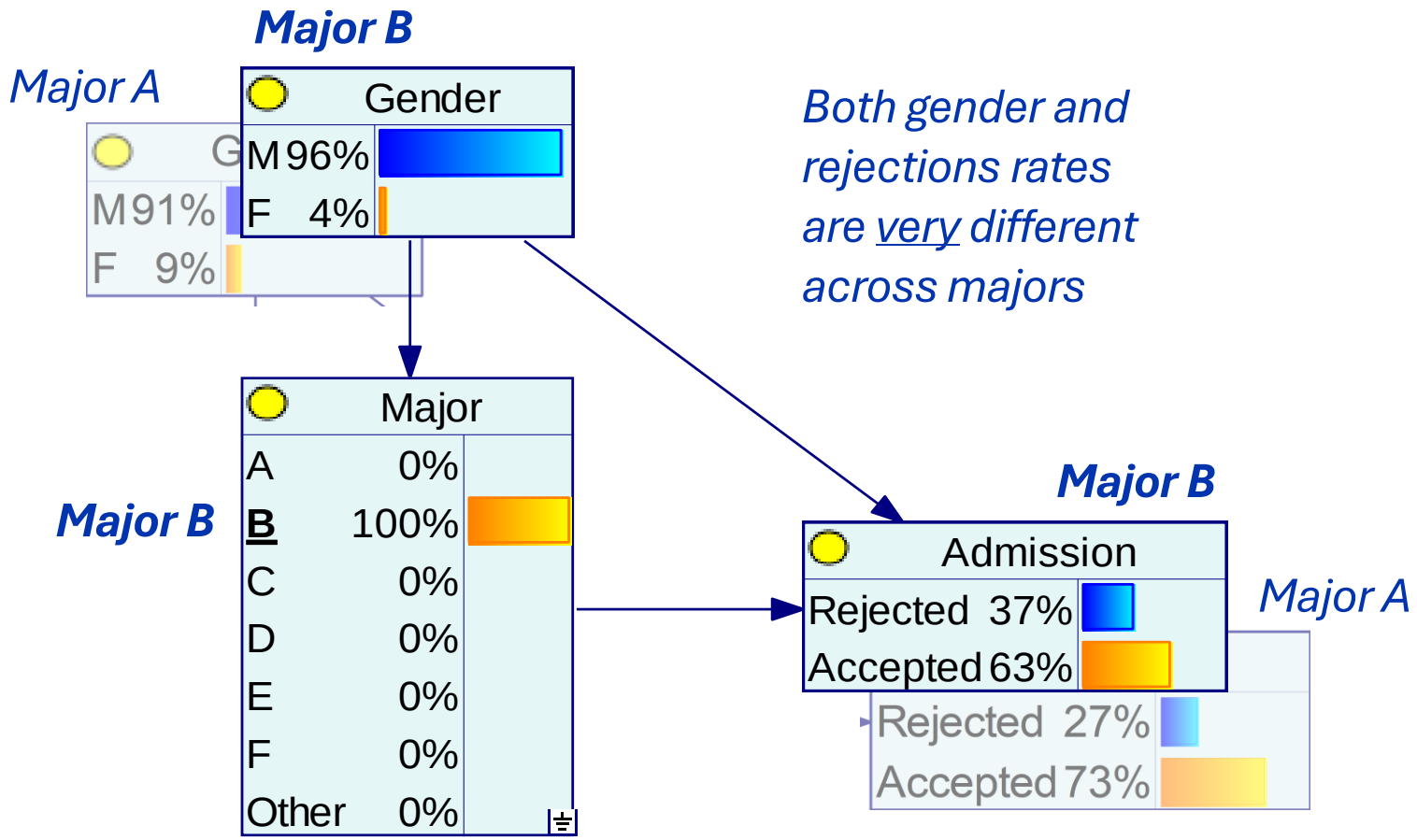
Bias in Data: *an Example* *(Berkeley Admission Test, 1973)*



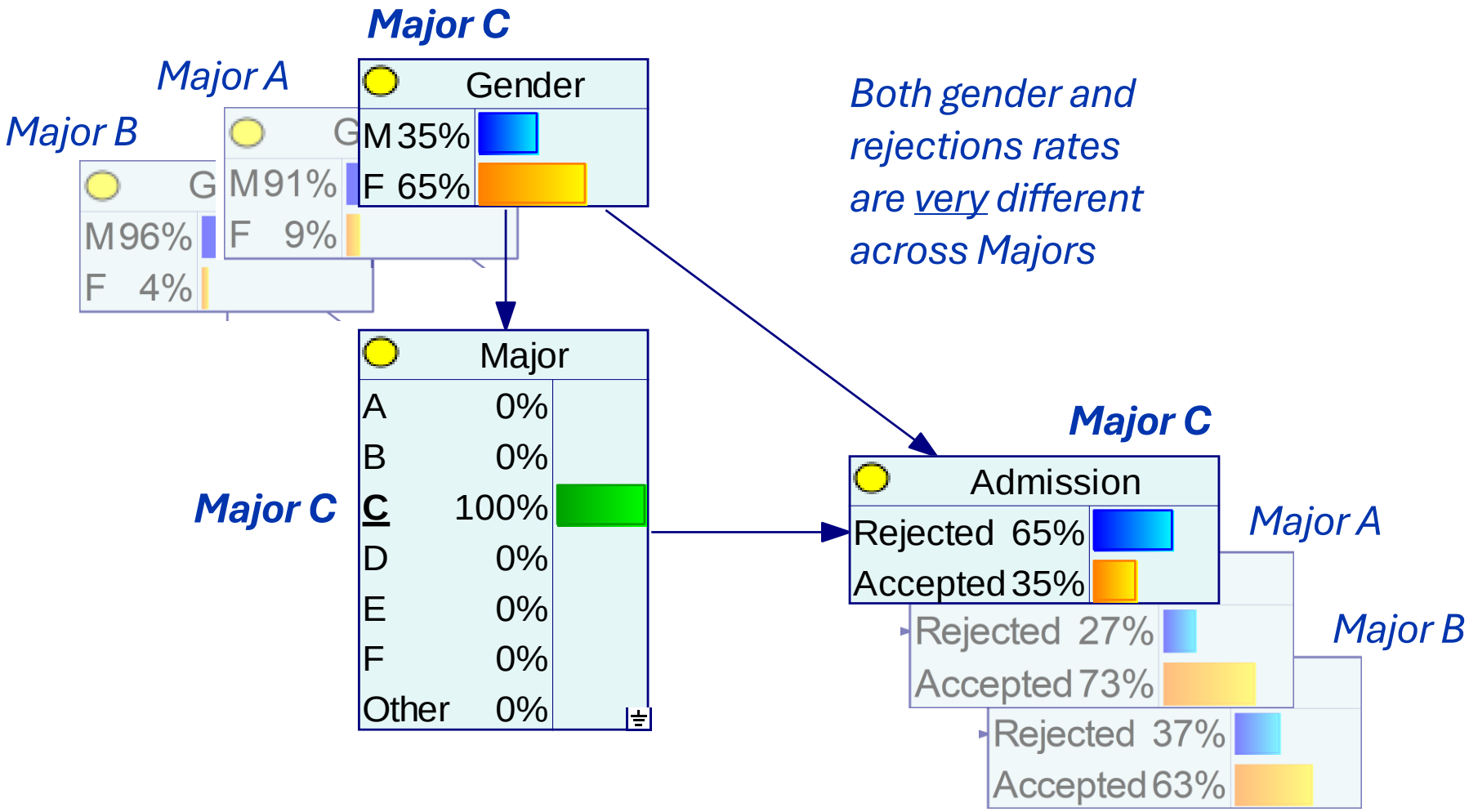
Bias in Data *An Example* *(Berkeley Admission Test, 1973)*



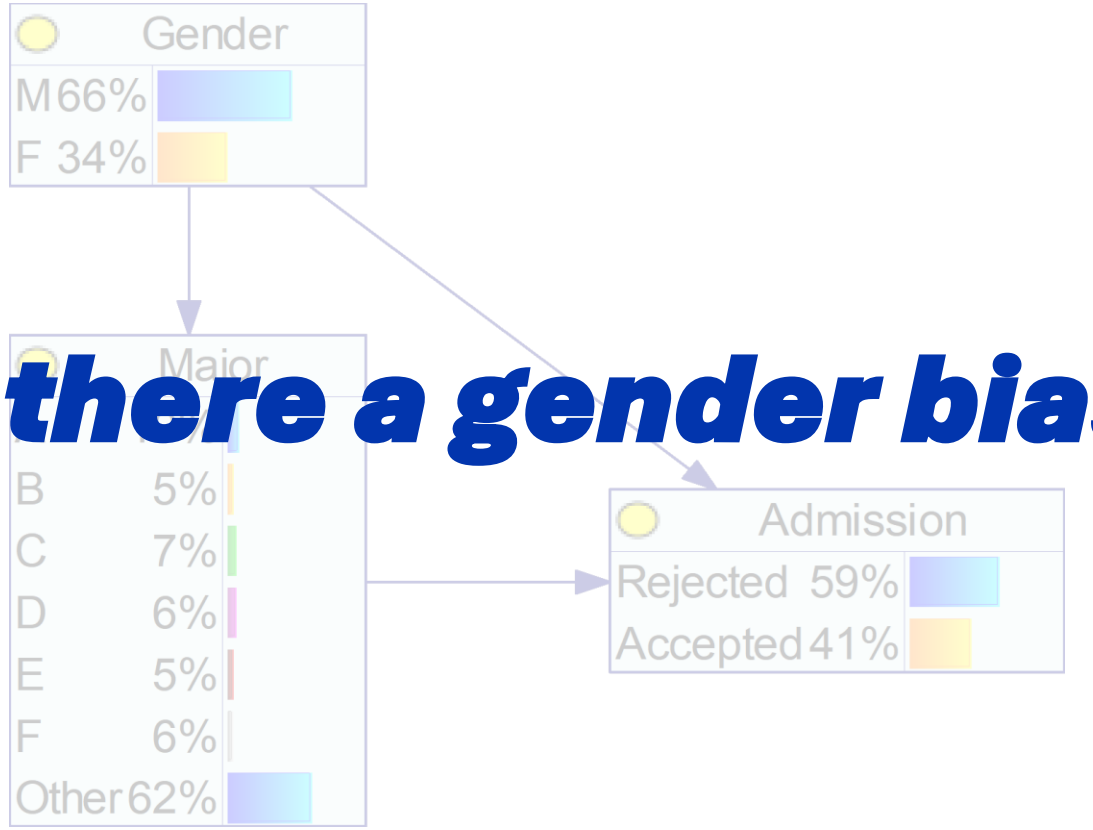
Bias in Data *An Example* *(Berkeley Admission Test, 1973)*



Bias in Data *An Example* *(Berkeley Admission Test, 1973)*



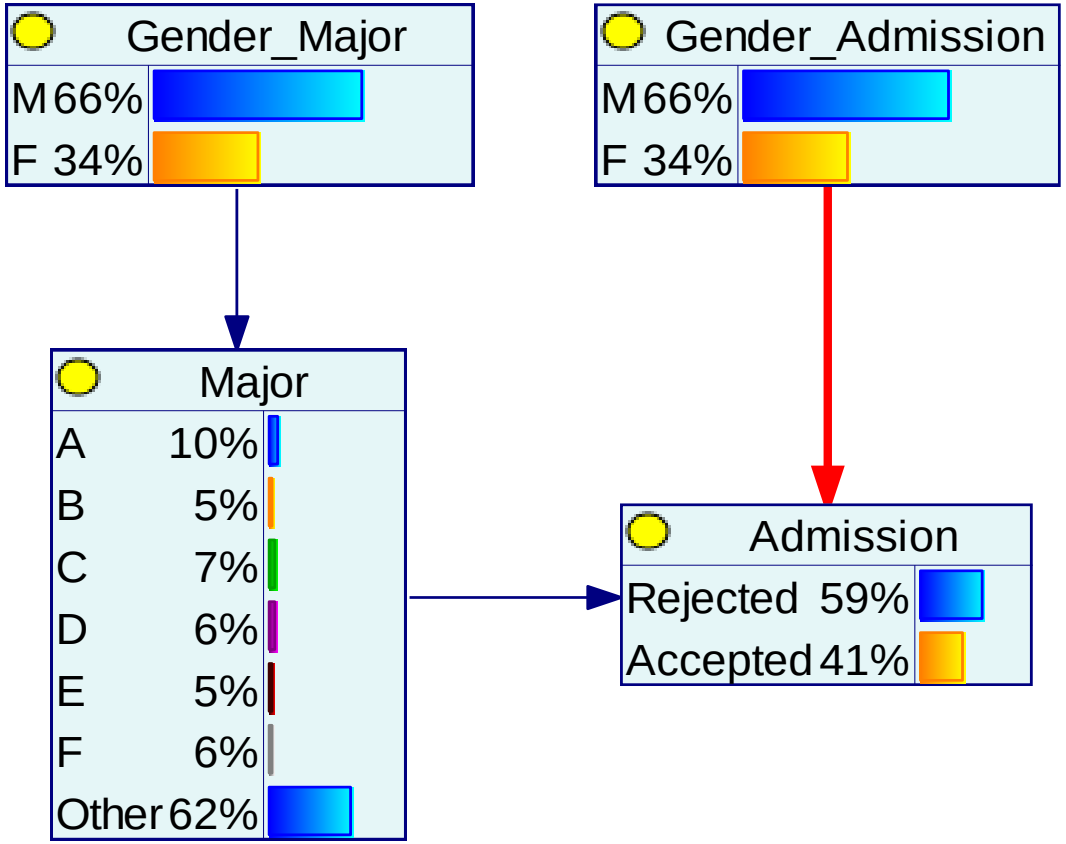
Bias in Data *An Example* *(Berkeley Admission Test, 1973)*



Is there a gender bias?

An Impossible Experiment

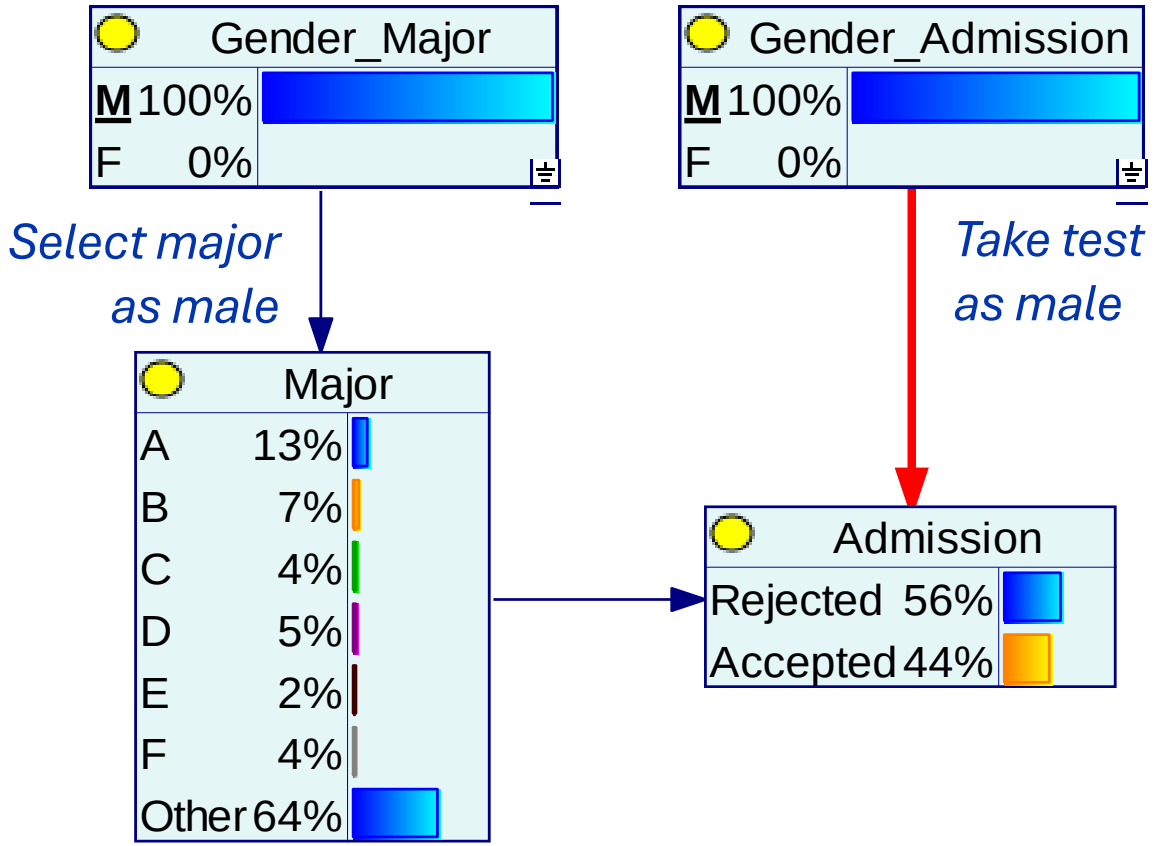
Selecting a major as gender X and taking test as gender Y



Path-Specific Counterfactual Fairness
[S. Chiappa, 2019]

An Impossible Experiment

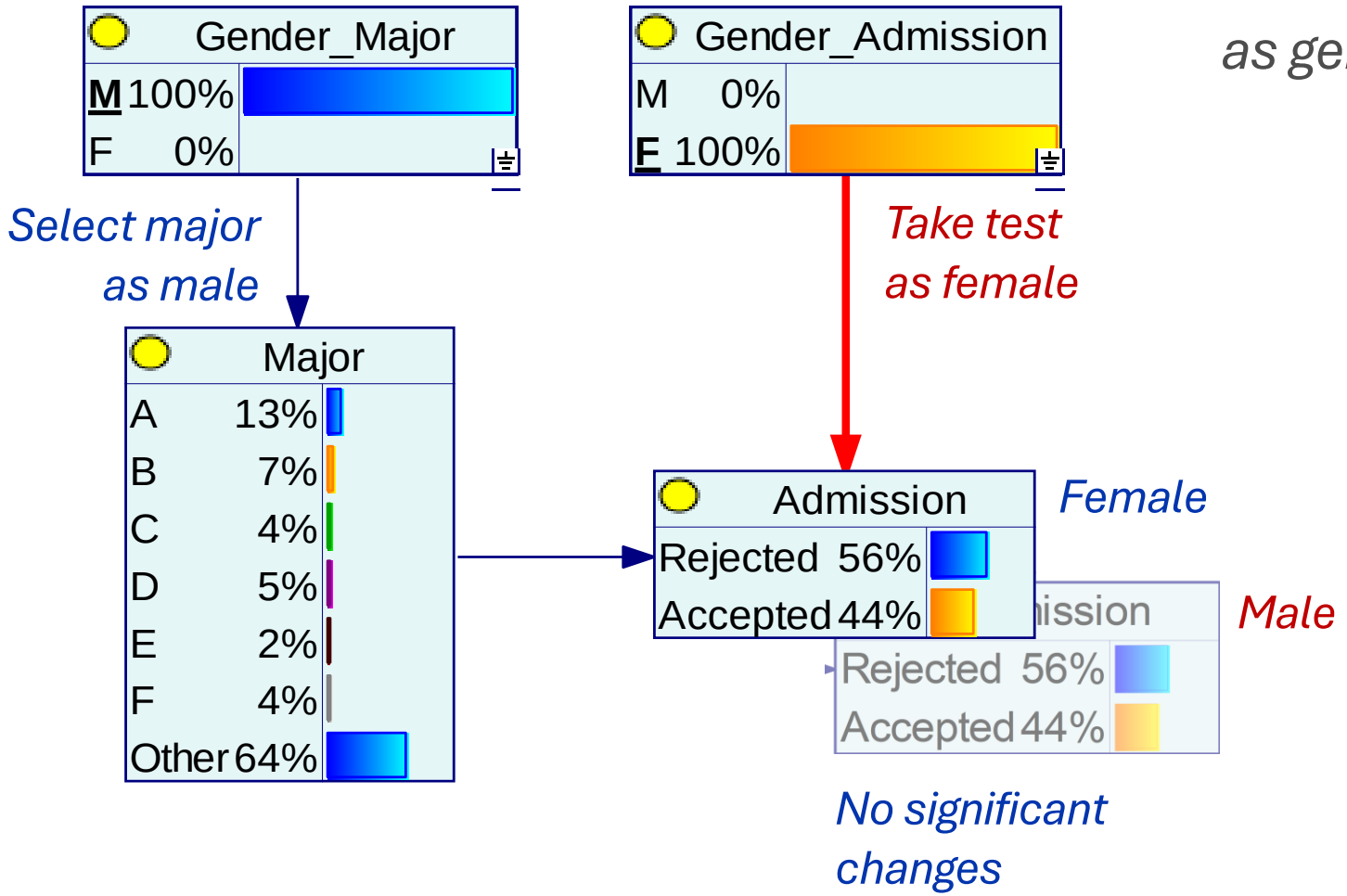
Selecting a major as gender X and taking test as gender Y



Path-Specific Counterfactual Fairness
[S. Chiappa, 2019]

An Impossible Experiment

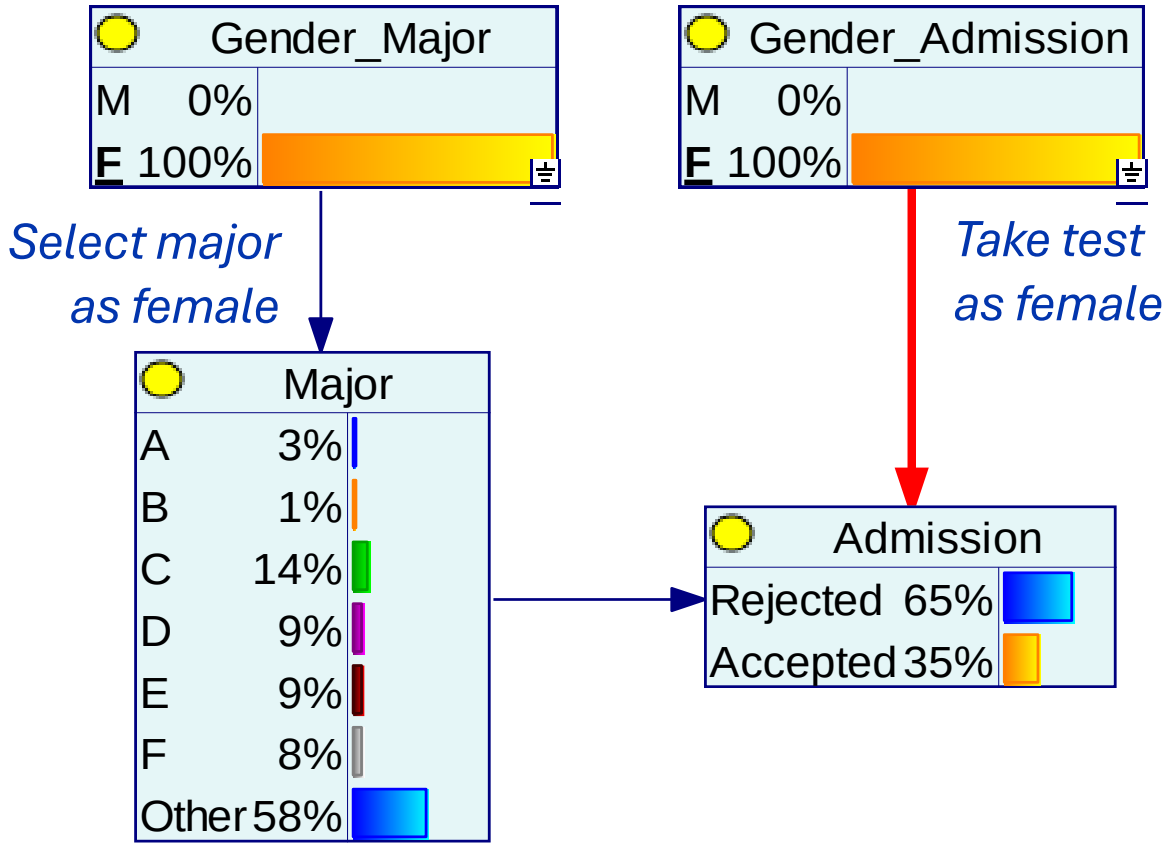
Selecting a major as gender X and taking test as gender Y



Path-Specific Counterfactual Fairness
[S. Chiappa, 2019]

An Impossible Experiment

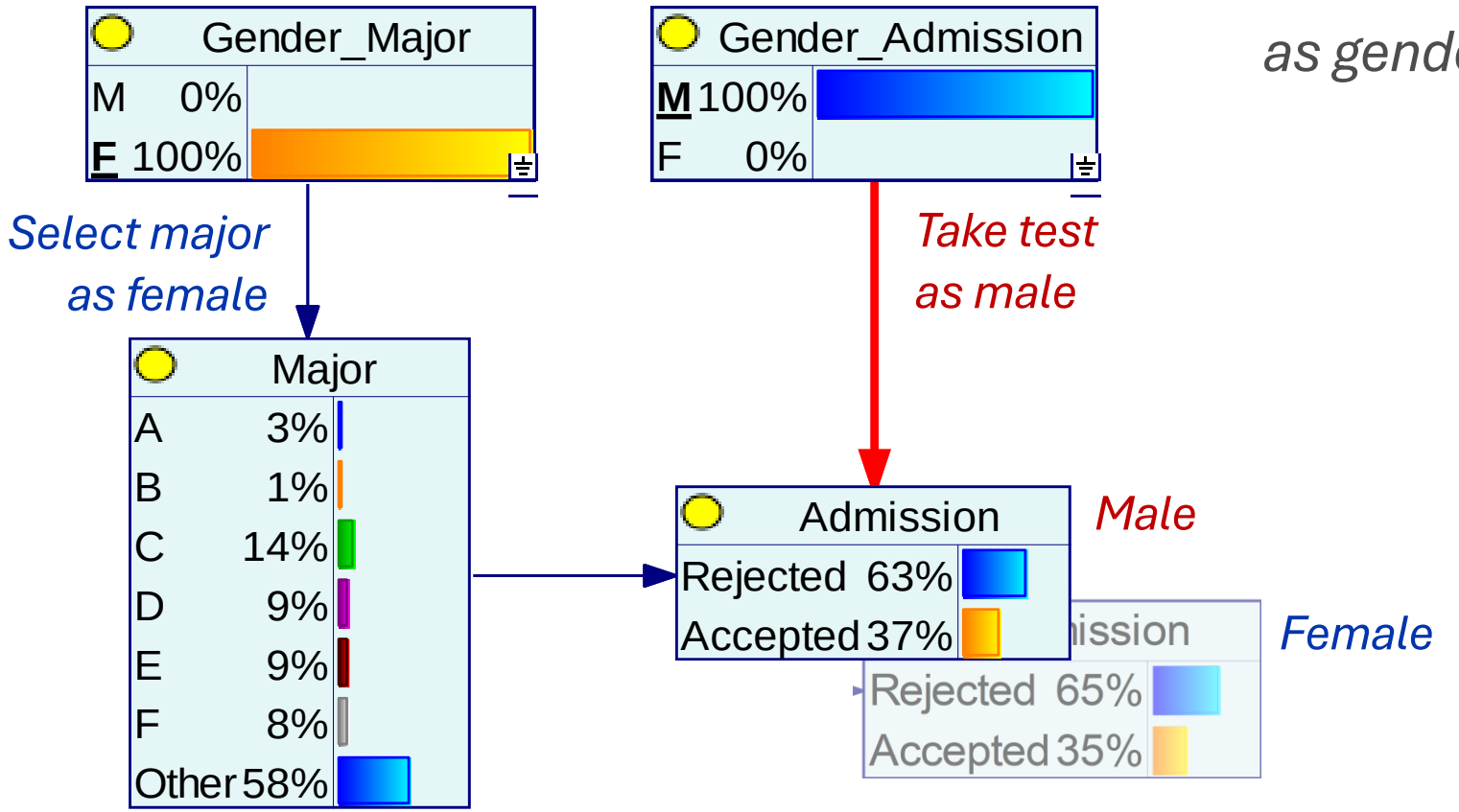
Selecting a major as gender X and taking test as gender Y



Path-Specific Counterfactual Fairness
[S. Chiappa, 2019]

An Impossible Experiment

Selecting a major as gender X and taking test as gender Y

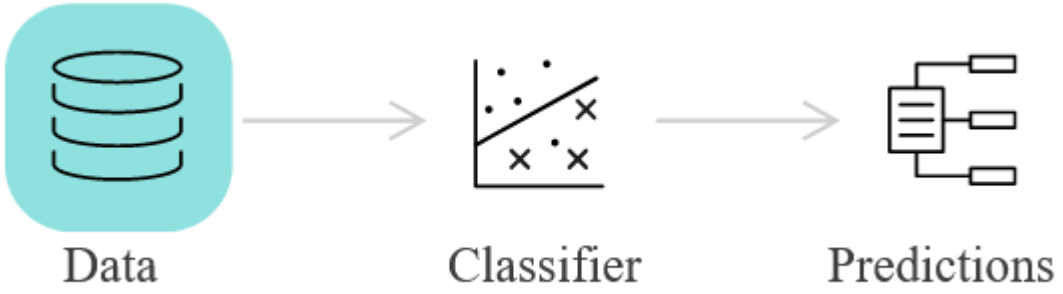


Path-Specific Counterfactual Fairness
[S. Chiappa, 2019]

Mitigating Bias *Via Counter-Bias*

- **Pre-processing data**

  Generate new data by applying appropriate transformations

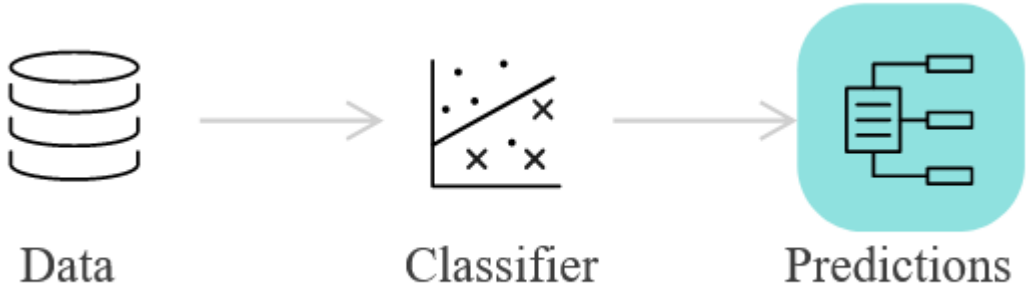


- **Altering the algorithm**

  Modify the training process to compensate for biased predictions

- **Post-processing outcomes**

  Change algorithm's predictions to address biased outputs



[Images from https://aif360.res.ibm.com/mitigate]

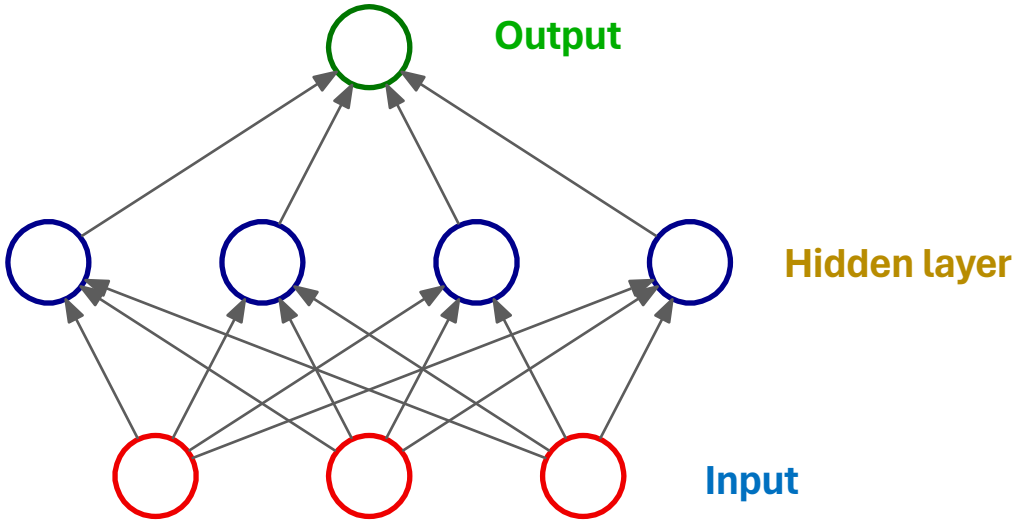
Inductive Bias

Basic Architecture *A Universal Pattern*

There is a well-known mathematical theorem [Cybenko, 1989; Hornik, 1991; Leshno et al. 1991] saying that, once trained, any neural networks could be translated into an equivalent one

With a much simpler architecture

The hidden layer may be as large as required (e.g., billions of units)



What is the difference, then?

- using less units (more compact networks)
- achieving a better **inductive bias**

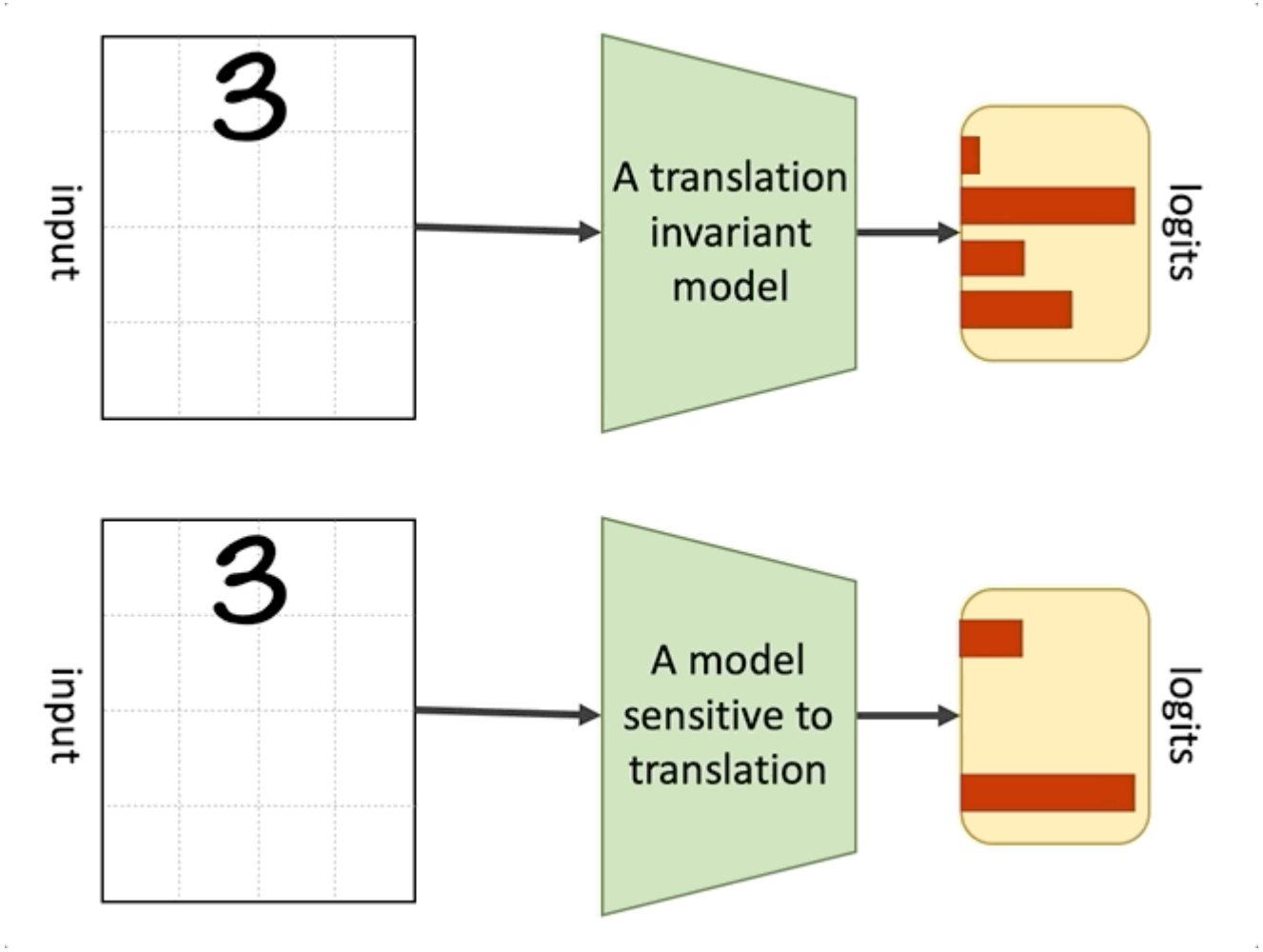
Inductive Bias *Have It Defined*

Inductive bias is anything which makes the algorithm *learn*
one pattern instead of another pattern

When searching a space of solutions, multiple possibilities
may be equally good, for a particular purpose:
an inductive bias allows a learning algorithm *to prioritize* one solution
(or interpretation) over another, independent of the observed data

[Adapted from Wikipedia]

Inductive Bias *Example: Translation Invariance*



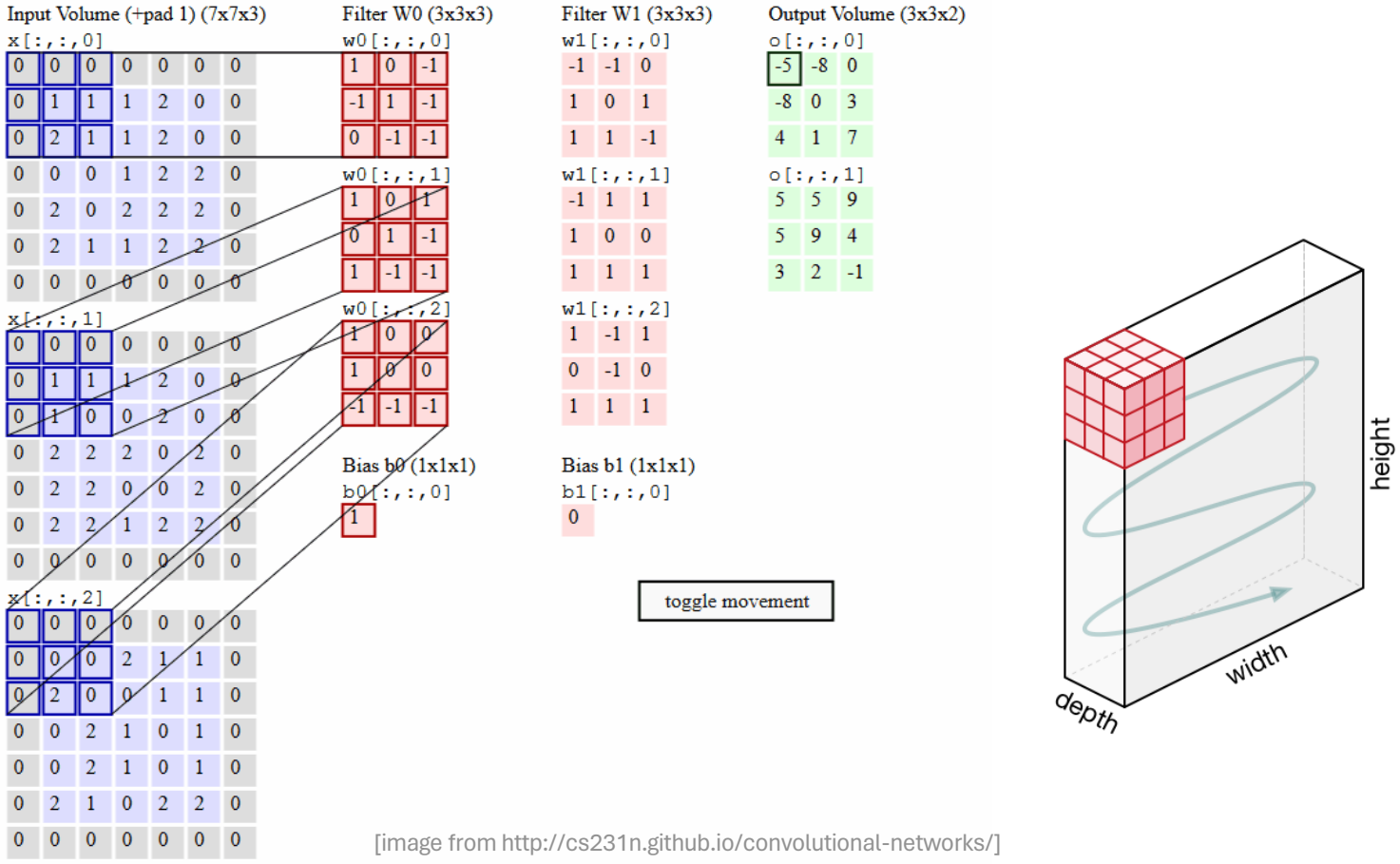
Output is the same, independent of input positioning

Output changes with input positioning

[Image from https://samiraabnar.github.io/articles/2020-05/indist]

Inductive Bias *Example: Translation Invariance*

In fact, neural network for vision-related tasks are ***convolutional***



[image from http://cs231n.github.io/convolutional-networks/]

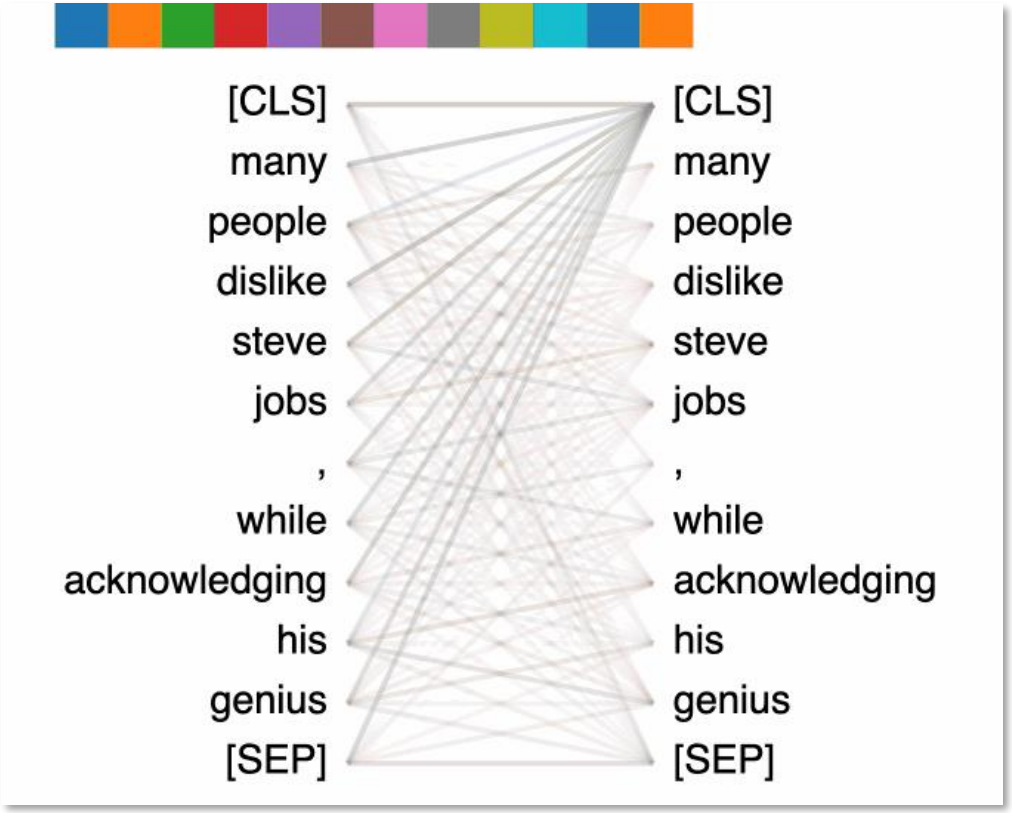
Which means that translational invariance is there by design

Transformers *Flexible, Adaptive Patterns*

Whereas convolution has a fixed scope,
transformers can learn to focus attention on different input spots

Attention spots may vary depending
on input positioning

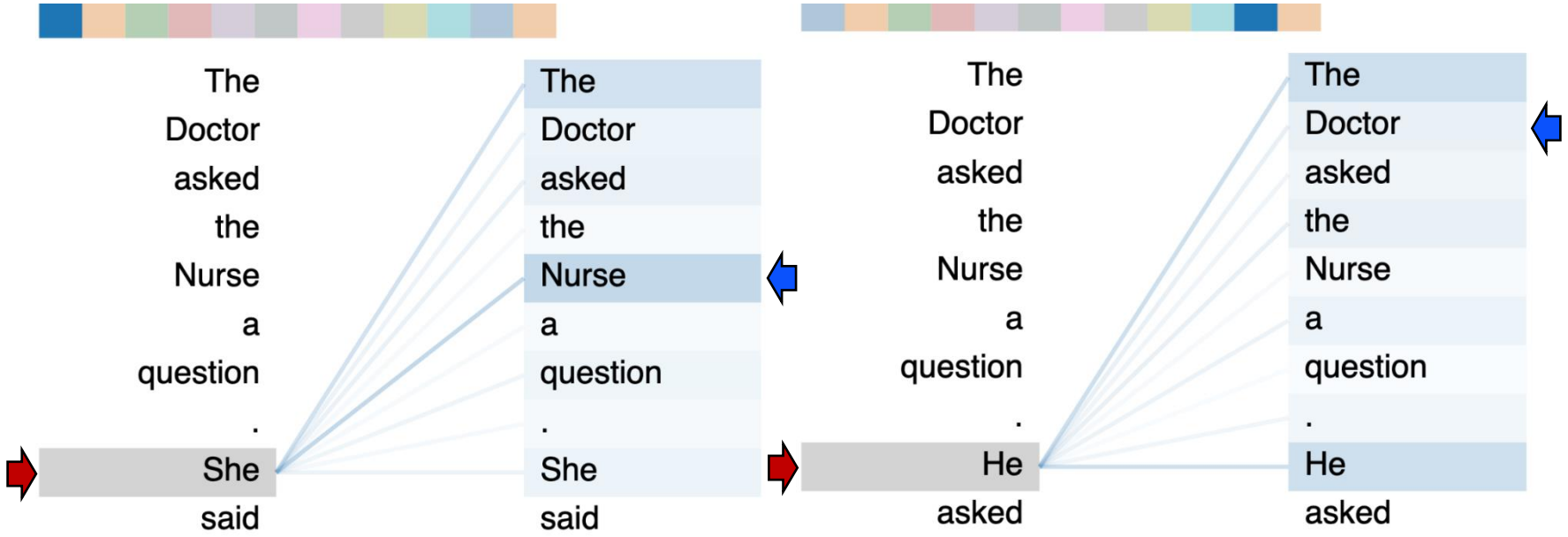
Transformer-based networks (like GPT)
use multiple such units in parallel



[image adapted from https://www.comet.com/site/blog/explainable-ai-for-transformers/]

No Free Lunch *No Guarantee of Being Bias-Free*

Depending on training data, *transformers* may learn to focus attention with unwanted biases



In these input sentences, the pronoun is switched from 'She' to 'He'

The network switches attention from 'Nurse' to 'Doctor'

[image adapted from https://www.comet.com/site/blog/explainable-ai-for-transformers/]

No Free Lunch *Intrinsic Inductive Bias*

Albeit not entirely understood yet, *transformers* have intrinsic inductive bias

- preference over sparser functions (*attention over fewer input elements*)
- less effective on problems involving recursion (*balancing brackets*, *iterated negations*)

The study of inductive bias in neural networks is relatively new and in progress

[see https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00306/43545/Theoretical-Limitations-of-Self-Attention-in]

Conclusions

- **Avoid simplistic metaphors to understand AI**

  By viewing AI as having human-like qualities,
  we risk overlooking its true nature and potential consequences

- **AI bias can serve as an instrument**

  When intentionally introduced by AI designers,
  bias can serve as a countermeasure against negative effects

- **Let's work together** (legal experts and AI engineers)

  Thoughtfully crafted regulations have the potential to guide AI systems
  toward better comprehension and more effective governance