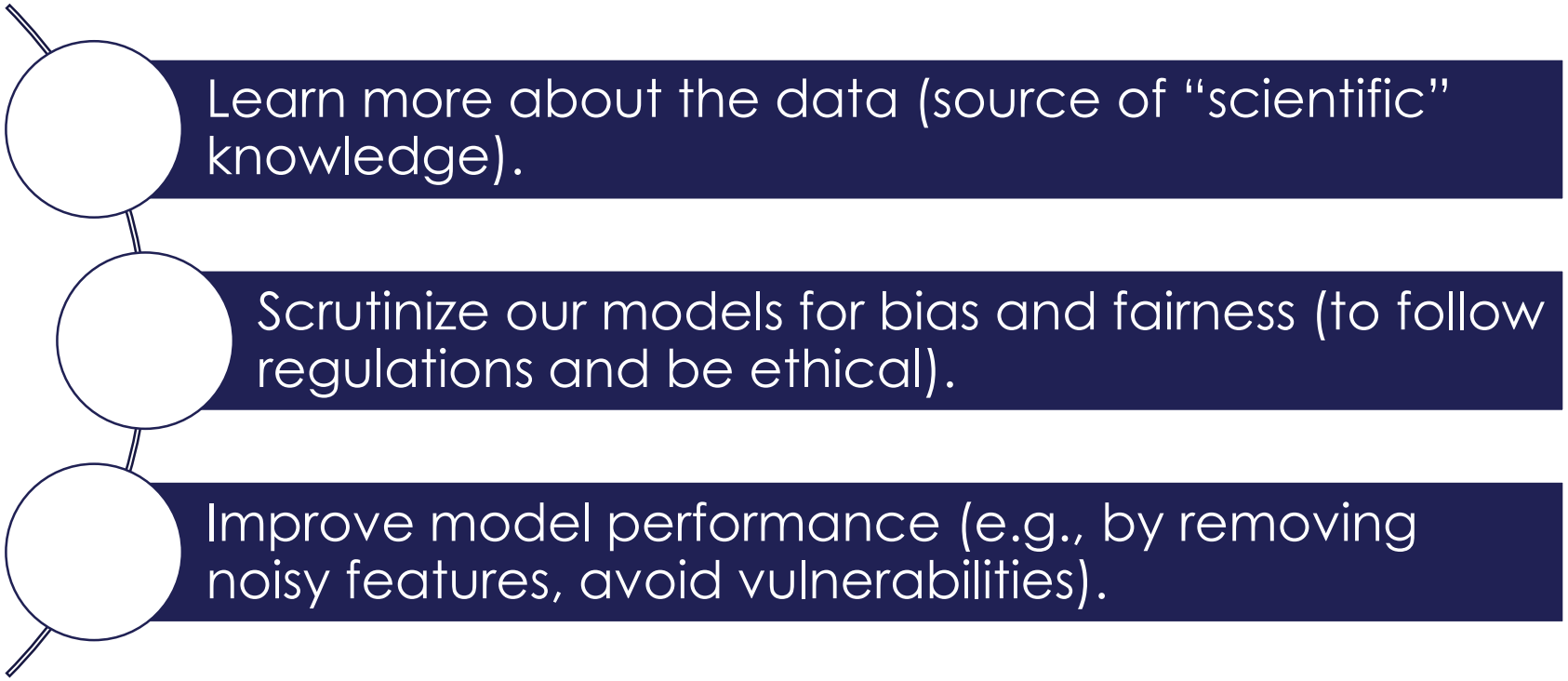
Importance of interpretability

Interpretability in machine learning

The ability to express in understandable terms, what the model has learned and the reasons that affect their output.

Importance of interpretability

Interpretability is a key element of trust for AI models. It allows us to:



Consequences

Black box models, which are inscrutable by design, or badly scrutinized white box models have led to serious societal problems that affect health, freedom, racial bias and safety.

Consequences

Organizations failing to scrutinize their models properly.

Lack of regulation ➔ no incentive to scrutinize models.

Consequences

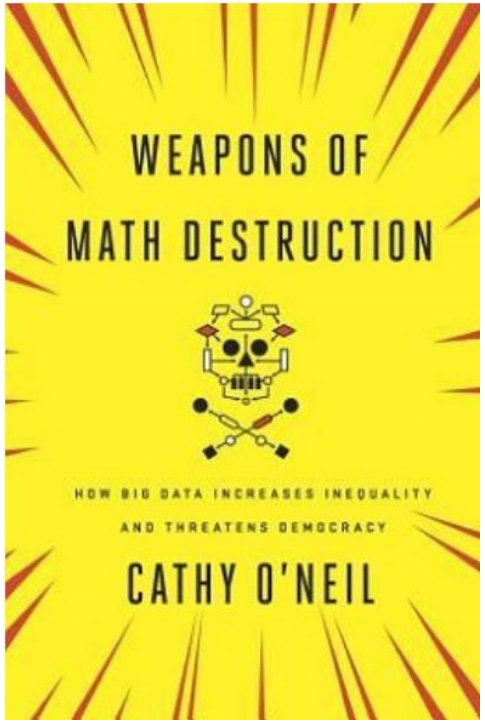
Examples	Models that use illegal data sources (e.g., psychology tests for recruitment).
	Models created to optimize a certain metric (i.e., engagement) without analyzing further impact promote misinformation.
	Models that discriminate on ethnicity or gender.

Consequences

Examples of badly scrutinized algorithms (in insurance, advertising, education, policing), that led to decisions that harmed the poor, reinforced racism, and amplified inequality.

Algorithms are opaque, hence difficult to contest.

Algorithms are scalable ➔ amplify inherently bias to affect larger populations.



Consequences

Facebook and twitter algorithms promote misinformation and content that spread anger.

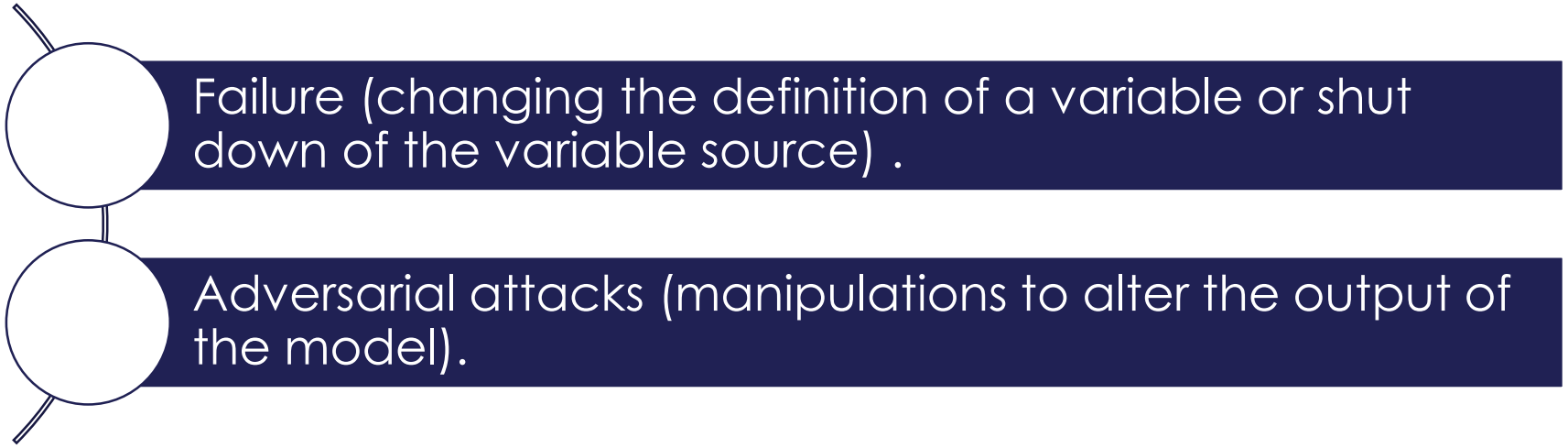
Instagram algorithm affects young teenager's mental well being.

https://www.wsj.com/articles/the-facebook-files-11631713039

Consequences

Not knowing what drives the model outputs can affect the organization.

Models that rely mostly on a few variables are more vulnerable to:



Importance of interpretability

Interpretability helps us:	Increase our knowledge.
	Improve model performance.
	Make models unbiased and fair.
	Follow the law and regulations.
	Decrease the possibility of unwanted effects.
	Protect the organization from adversarial attacks or profit loss.

Train In Data

THANK YOU

www.trainindata.com