NAVAL
POSTGRADUATE
SCHOOL

Metacognition for Artificial Intelligence Systems:
an Approach to Safety and Desired Behavior in Complex Systems

2021 ICONS
16th International Conference on Systems

AI4SysSoS Special Track

20 April 2021

Dr. Bonnie Johnson
NPS Systems Engineering
bwjohnson@nps.edu

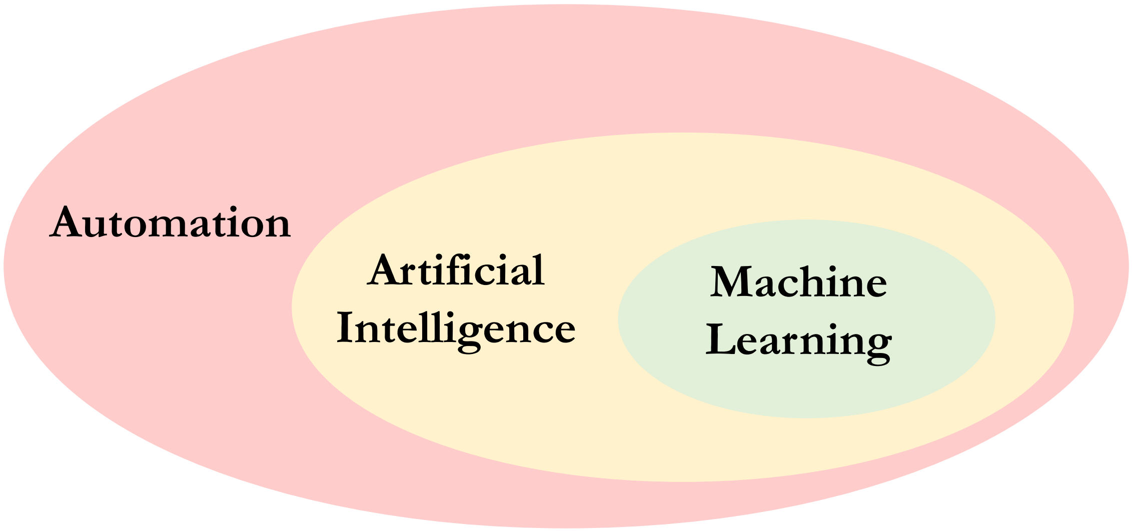
Advances in computational thinking and data science have led to a new era of artificial intelligence systems being engineered to adapt to complex situations and develop actionable knowledge. These learning systems are meant to reliably understand the essence of a situation and construct critical decision recommendations to support autonomous and human-machine teaming operations.

In parallel, the increasing volume, velocity, variety, veracity, value, and variability of data is confounding the complexity of these new systems – creating challenges in terms of their development and implementation. For artificial systems supporting critical decisions with higher consequences, safety has become an important concern. Methods are needed to avoid failure modes and ensure that only desired behavior is permitted.

What is AI?

Here's a good definition:

AI is a field that includes many different approaches with the objective of creating machines with intelligence[1]



Two Primary Types of AI[2]

Explicitly Programmed

Handcrafted Knowledge Systems[3]

- Think "if-then," but can be more complex
- Uses normal programming languages
- Can involve complex manually designed coding schemes for data / knowledge

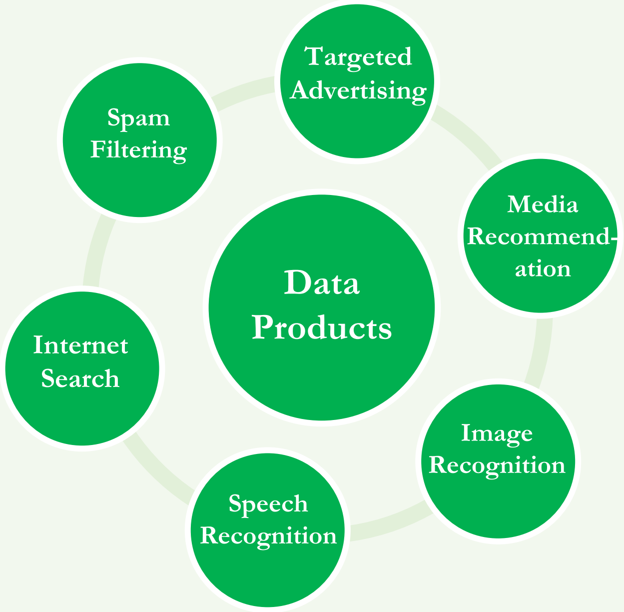
Learns from Data

Machine Learning Systems

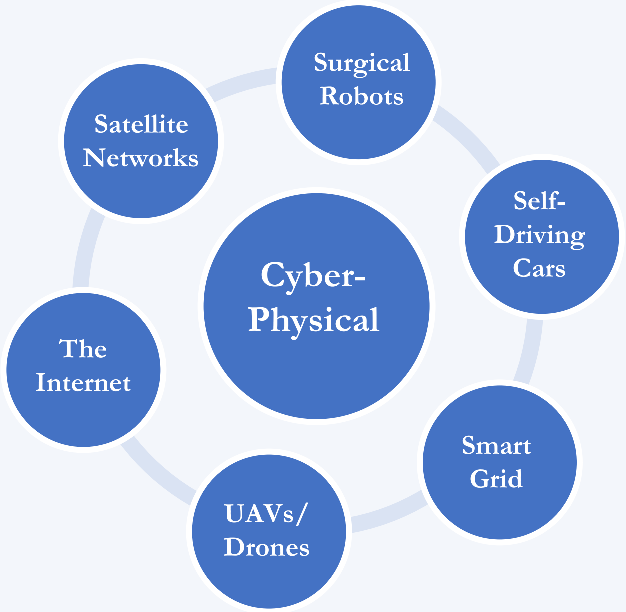
- The system is provided a large amount of data (many labeled examples)
- The system learns patterns by trial-and-error until it can predict the labeled examples
- Then, the "trained" system can be used (for prediction) given new data

1 – Melanie Mitchell. 2019. Artificial Intelligence – A Guide for Thinking Humans Picador: New York. – definition of AI as a field
2 – Barclay Brown. 2021. Presentation on AI Systems at the INCOSE International Workshop, January 2021. – description of two primary types of AI
3 – Greg Allen. 2020. Understanding AI Technology. Joint AI Center (JAIC) Report, US Dept of Defense – definition of handcrafted knowledge systems

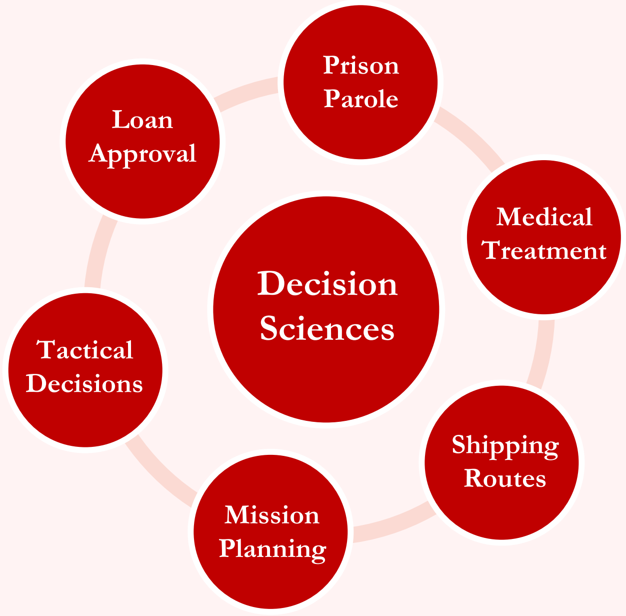
Three Types of AI System Application Domains



Data product systems use computers to generate information products.



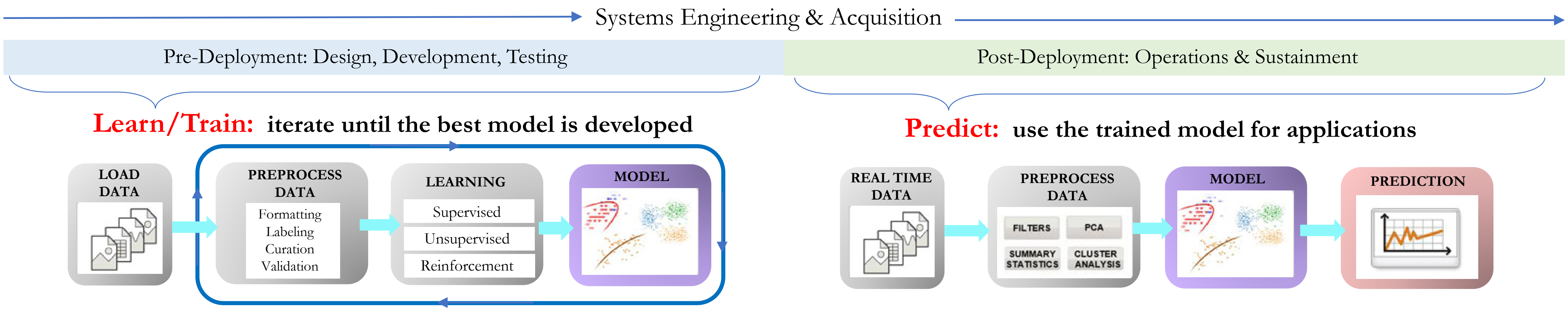
Cyber-physical systems include computer automation (often AI) and physical components.



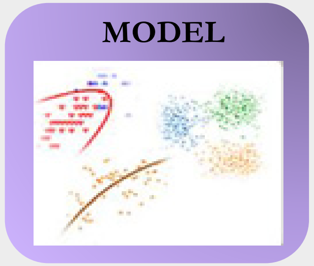
Decision science systems use computer algorithms to automate the process of making decision and advising plans and strategies.

Each application domain contains its own range of possible failure modes, and each will require tailored safety solution measures.

Machine learning systems introduce a new set of challenges



The machine learning "system" is the trained model



Characteristics of ML Systems:

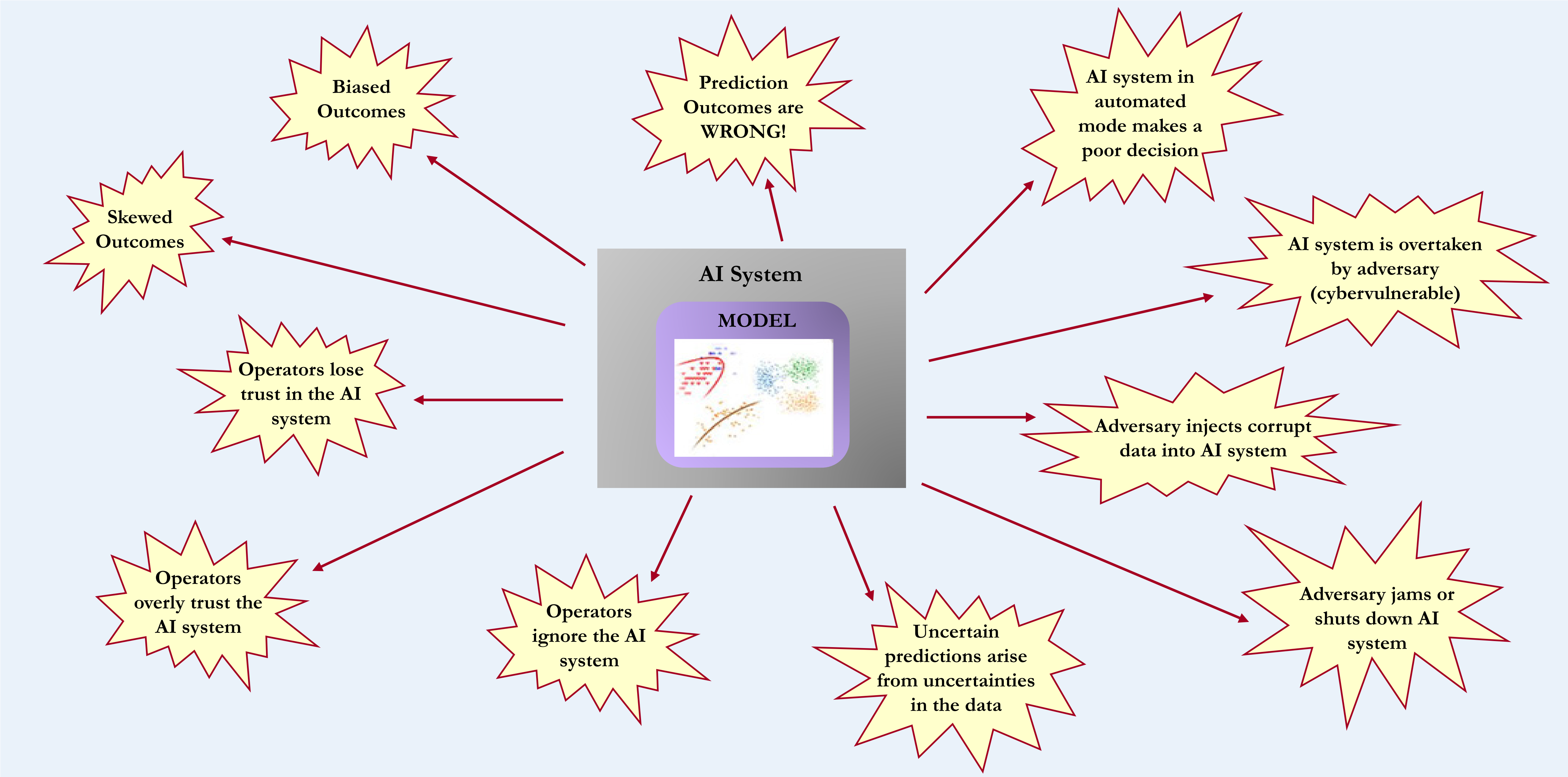
Non-Deterministic – ML is a technique that allows a computer to learn a task without being explicitly programmed. The ML system implements inductive inference on real-time or operational data sets after being trained. Therefore, ML system behavior leads to variability in results.

Intimately Connected to Data – ML systems "emerge" or are generated through the process of learning on training data sets. They are a product of the quality, sufficiency, and representativeness of the data. They are intimately connected and wholly dependent on their training data.

Complex – ML systems can exhibit complex behavior due to deep learning (the ML system consists of networks of many learning sub-components) and complex mathematical operations involving very large datasets and computations. The complex (unexpected) behavior can emerge.

Intimately Connected to Context – During operations, the behavior of ML systems is highly dependent on the context, or operational situation. Uncertainty in data representations of situational awareness, will lead to ML system prediction error. Complexity in the operational situation will lead to complex ML system operations.

Failure Modes



Consequences

Two types of AI systems according to the severity of their failure consequences

Type A ***Safety is Paramount***	**Type B** ***Safety is Less Important***
Applications in which AI system model predictions are used to support consequential decisions that can have a profound effect on people's lives	Applications in which AI system model predictions are used in settings of low consequence and large scale that have minimal effects on people's lives

Root Causes

Systems Engineering & Acquisition

Pre-Deployment: Design, Development, Testing	Post-Deployment: Operations & Sustainment

Bias in the training data sets

Incompleteness---data sets don't represent all scenarios

Rare examples – data sets don't include unusual scenarios

Corruption in the training data sets

Mis-labeled data

Mis-associated data

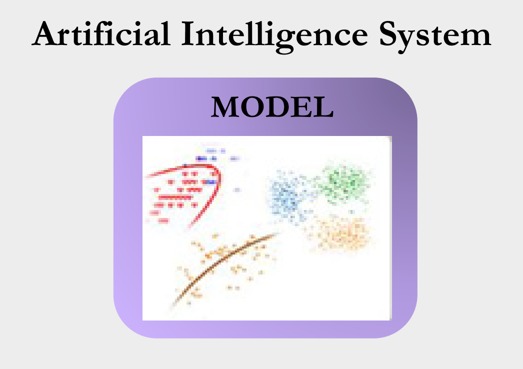
Poor validation methods (is there criteria for deciding how much training data is good enough?)

Poor data collection methods

Underfitting in the model – when the model is not capable of attaining sufficiently low error on the training data

Cost function algorithm errors – when trained model is optimized to the wrong cost function

Wrong algorithm – when the training data is fit to the wrong algorithmic approach (regression neural network, etc.)



Uncertainty/error in operational datasets

Corruption in operational datasets

Inaccuracy in the algorithm model (prediction error)

Operational complexity that overwhelms the AI system

Overfitting – when the model presents a very small error on the training data but fails to generalize, i.e., fails to perform as well on new examples; the model is "overfit" to the training data

Lack of explainability

Trust issues

Operator-induced error

Adversarial attacks – hacking, deception, inserting false data, controlling automated systems

AI System Safety: Four Types of Solution Strategies

Systems Engineering & Acquisition Lifecycle

Pre-Deployment: Design, Development, Testing

Post-Deployment: Operations & Sustainment

1. Inherently Safe Design

Focus: ensuring robustness against uncertainty in the training data sets

- Interpretability – ensuring designers understand the complex AI and ML systems that are produced from the data training process
- Causality – reducing uncertainty by eliminating non-causal variables from the model

2. Safety Reserves

Focus: achieving safety through additive reserves, safety factors, and safety margins – through training data set validation

- Validating training data sets – eliminating uncertainty in the data sets; ensuring data sets are accurate, representative, sufficient, bias-free, etc.
- Increasing/improving model training process – ensuring adequate time and resources are provided for training and validation process

3. Safe Fail

Focus: system remains safe when it fails in its intended operation

- Human operation intervention – the operation of AI systems should allow for adequate human-machine interaction to allow for system overrides and manual operation
- Metacognition – the AI system can be designed to recognize uncertainty in predicted outcomes or possible failure modes and then alert operators and revert to a manual operation mode
- Explainability/Understandability/Trust-worthy

4. Procedural Safeguards

Focus: measures beyond ones designed into the system; measures that occur during operations

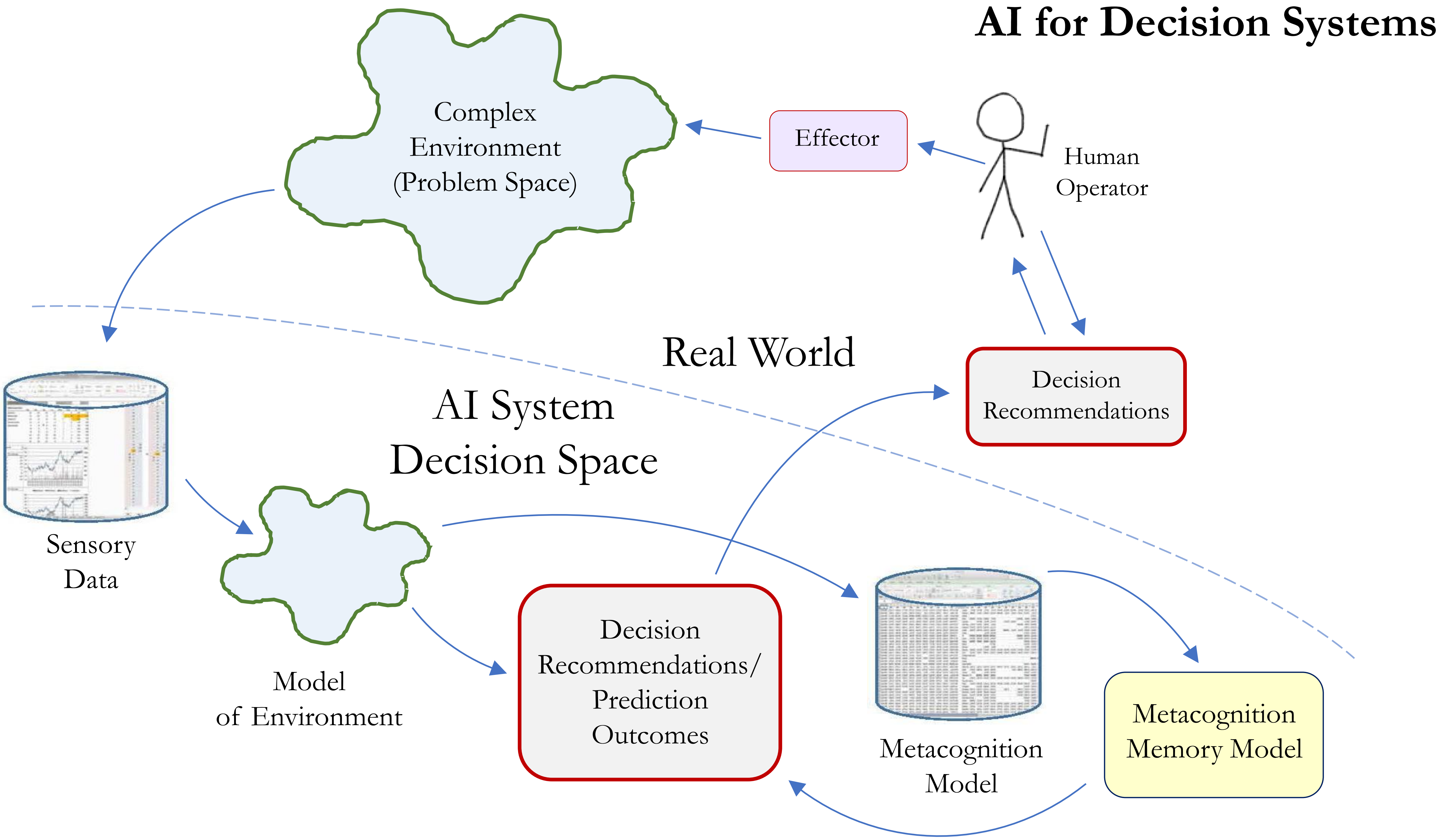
- Audits, training, posted warnings, on-going evaluation

Metacognition is a solution strategy that promotes self-awareness within the artificial intelligence system to understand its external and internal operational environments and use this knowledge to identify potential failures and enable self-healing and self-management for safe and desired behavior.

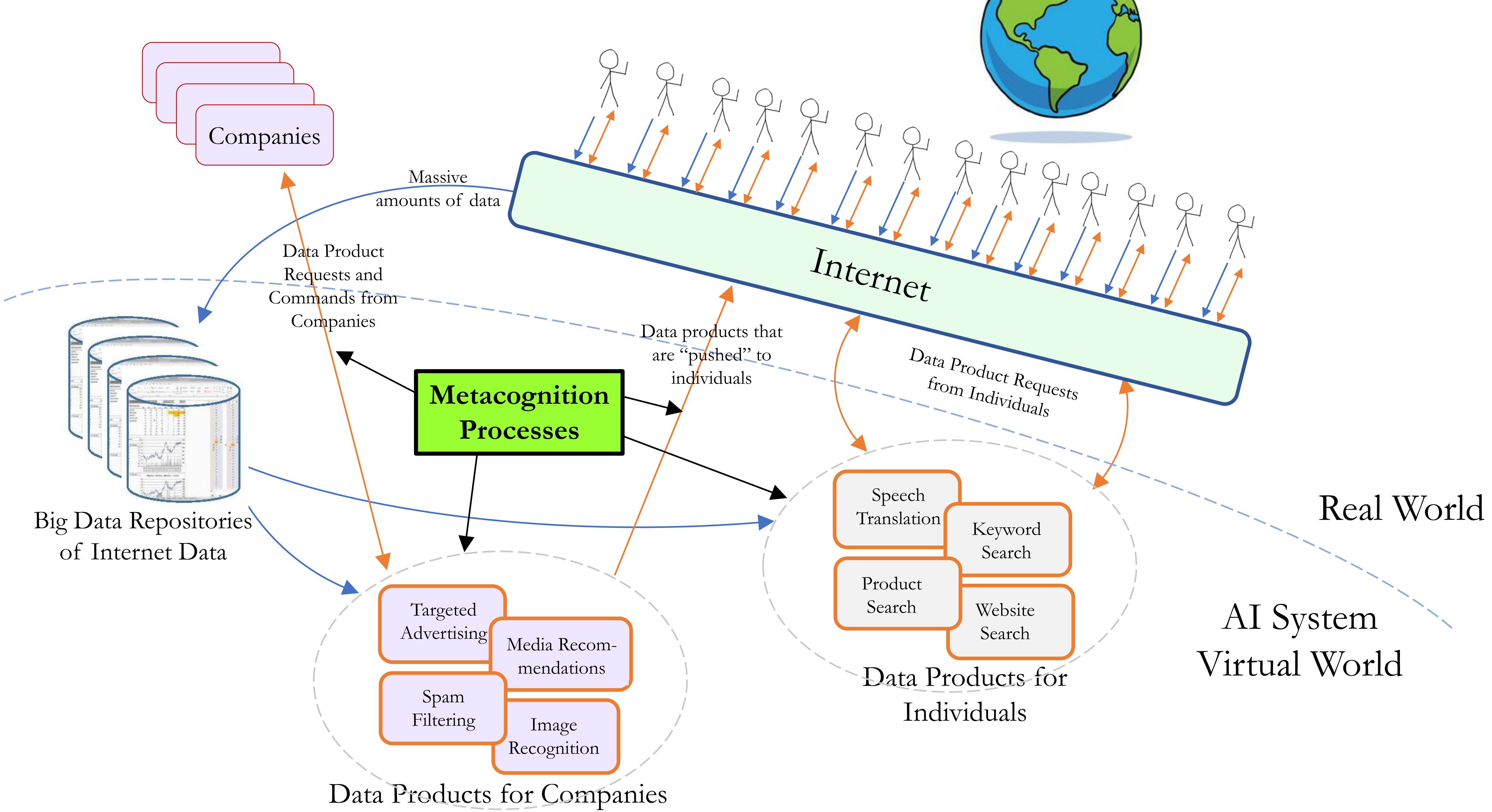
Metacognition as a safety measure

Metacognition Capabilities

1. Evaluating level of uncertainty in knowledge
2. Evaluating level of uncertainty in AI outputs
3. Failure self-predictions
4. Anomaly detection
5. Identification of new or unfamiliar situation
6. Evaluation of situation complexity
7. Constructionist learning: self-sufficient locus of control
8. Identification/prediction of high-risk courses of action
9. Identification/prediction of undesirable emergent behavior
10. Prediction of poor performance
11. Development of metacognitive memory
12. Evaluation of historical safety risks, failures, error, poor performance
13. Evaluation of contextual complexity, uncertainty, and unfamiliarity
14. Evaluation of individual component failures


AI for Decision Systems
Complex Environment (Problem Space)
Effector
Human Operator
Real World
Decision Recommendations
AI System Decision Space
Sensory Data
Model of Environment
Decision Recommendations/ Prediction Outcomes
Metacognition Model
Metacognition Memory Model

AI for Data Products



NAVAL POSTGRADUATE SCHOOL

Wrap Up

- AI/ML has huge potential for many diverse applications (data products, cyber-physical, decision sciences)
- AI systems present new types of safety risks: failure modes, consequences, root causes
- AI safety must be implemented throughout the systems engineering lifecycle
- Metacognition is an AI system safety strategy that must be engineered into systems and implemented during operations.
- Many exciting research opportunities!

I welcome collaboration!

Dr. Bonnie Johnson
Naval Postgraduate School
bwjohnson@nps.edu

References

Allen, G. 2020. Understanding AI technology. Joint Artificial Intelligence Center (JAIC) Report, US Department of Defense.

Crowder, J., Friess, S. 2011. Metacognition and metamemory concepts for AI systems. In: Athens: the Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 1-6.

Crowder, J., Friess, S. 2012. Extended metacognition for Artificially Intelligent Systems (AIS): artificial locus of control and conitive economy. In: Athens: the Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 1-6.

Crowder, J., Carbone, J. 2014. Eliminating cognitive ambiguity with knowledge relativity threads. In: Athens: The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 1-7.

Faria, J. 2017. Non-determinism and failure modes in machine learning. In: 2017 IEEE 28th International Symposium on Software Reliability Engineering Workshops, 310 – 316.

Faria, J. 2018. Machine learning safety: an overview. In: Proceedings of the 26th Safety-Critical Systems Symposium. York, UK, 4-18.

Gunning, D., Aha, D. 2019. DARPA's Explainable Artificial Intelligence Program. AI Magazine 40(2), 44-58.

Johnson, M., Bradshaw, J., Feltovich, P. 2018. Tomorrow's human-machine design tools: from levels of automation to interdependencies. Journal of Cognitive Engineering and Decision-Making, 12(1), 77-82.

Michael, C., Acklin, D., Scheuerman, J. 2020. On interactive machine learning and the potential of cognitive feedback. In: arXiv:2003.10365v1.

Mitchell, M. 2019. Artificial Intelligence – A Guide for Thinking Humans. Picador: New York.

Nushi, B., Kamar, E., Horvitz, E. 2018. Towards accountable AI: hybrid human-machine analyses for characterizing system failure. In: Sixth Annual Conference on Human Computation and Crowdsourcing (HCOMP), 126-135.

Varshney, K. 2016. Engineering safety in machine learning. In: Information Theory and Applications Workshop (ITA), La Jolla, CA, USA, 1-5. doi: 10.1109/ITA.2016.7888195.

Varshney, K., Alemazdeh, H. 2017. On the safety of machine learning: cyber-physical systems, decision sciences, and data products. Big Data 5(3), 246-255.

Welling, M. 2015. Are ML and statistics complimentary? In: IMS-ISBA Meeting on Data Science in the Next 50 Years, December.