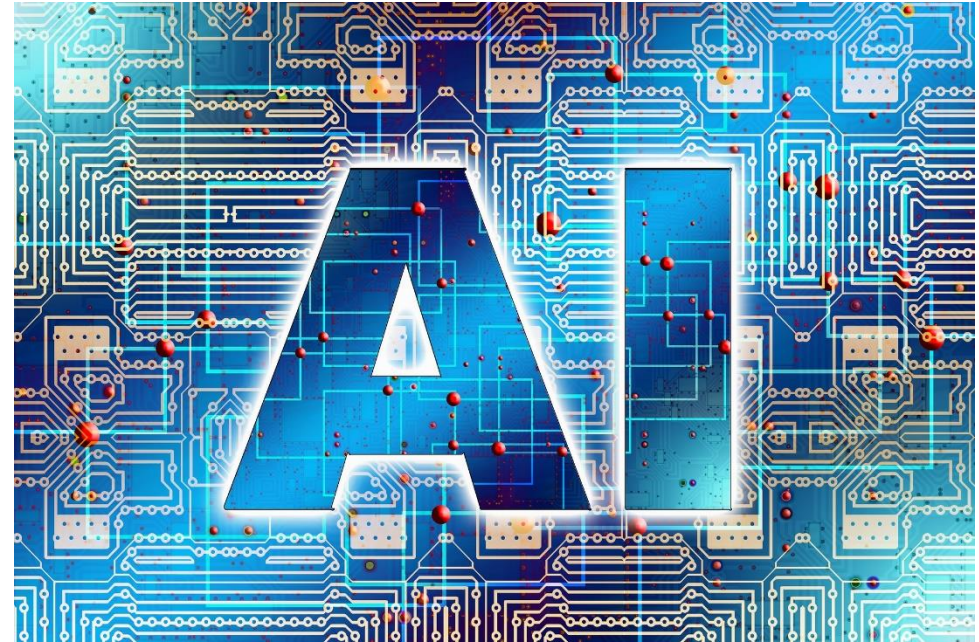




Virtual Mini Event, November 2021

Where is the Data?

*Data Requirements and the **Green** School Bus Problem*



Barclay R. Brown, Ph.D, ESEP
Engineering Fellow
Raytheon Missiles and Defense
Barclay.brown@incose.net

Ramakrishnan Raman, Ph.D, ESEP
Principal Systems Engineer
Honeywell
ramakrishnan.raman@incose.net

Barclay R. Brown, Ph.D., ESEP

Barclay R. Brown is an Engineering Fellow at Raytheon Technologies, focused on MBSE and artificial intelligence and machine learning in systems engineering.

Before joining Raytheon, he was the Global Solution Executive for the Aerospace and Defense Industry at IBM, and lead systems engineer for some of IBM's largest development projects.

Dr. Brown holds a bachelor's degree in Electrical Engineering, master's degrees in Psychology and Business and a PhD in Industrial and Systems Engineering.

He has taught systems engineering and systems thinking at several universities and is a certified Expert Systems Engineering Professional (ESEP), certified Systems Engineering Quality Manager.



Barclay R. Brown,
Engineering Fellow

[Barclay.R.Brown@
Raytheon.com](mailto:Barclay.R.Brown@Raytheon.com)

Ramakrishnan Raman, PhD, ESEP

- Ramakrishnan (Ramki) Raman is currently Principal Systems Engineer at Honeywell. He is a certified Six Sigma Black Belt and is INCOSE Certified Expert Systems Engineering Professional – ESEP, and IEEE Systems Council Distinguished Lecturer
- Ramki's areas of interest includes complex systems and system-of-systems, and artificial intelligence machine learning, and he has several publications in peer reviewed international journals & conferences on these areas
- Ramki is active in the professional societies, including INCOSE, IEEE, and SAE. He is currently serving as Assistant Director – INCOSE Asia Oceania, and as Chair of IEEE Bangalore Section Systems Council Chapter. He is a recipient of INCOSE Outstanding Service Award in 2016. He has also been the Technical Program Chair for international conferences including 2016 & 2019 Asia Oceania Systems Engineering Conference, and 31st INCOSE International Symposium.
- Ramki actively interacts with various academic institutions, and is currently Guest Faculty at IIT Bombay, and Adjunct Faculty at Manipal Institute of Technology



A background image showing a perspective view of a tunnel formed by the words "DATA" and "ATAQ" in blue, repeating and receding into the distance.

■ What is the role of data in AIML applications from a systems perspective?

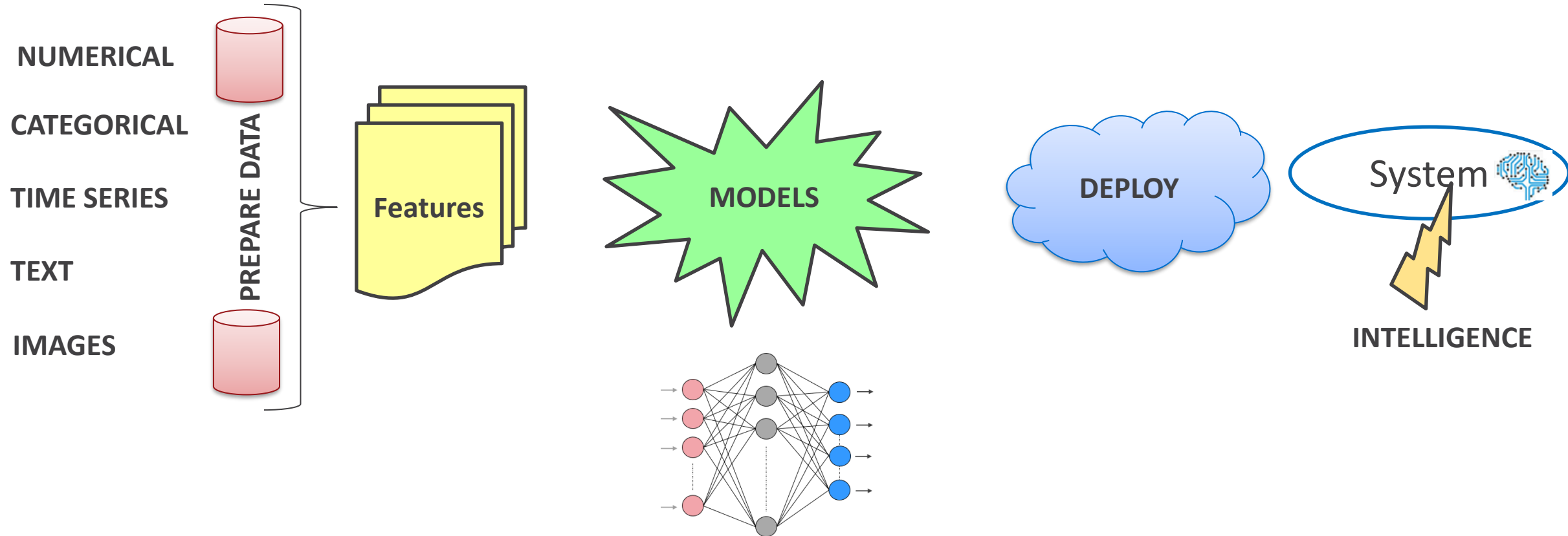
■ How much do systems engineers need to know about AIML and about training data for systems?

■ Can't we just leave data selection to the data scientists and ML developers?

■ Does the existence of training data affect the systems development lifecycle?

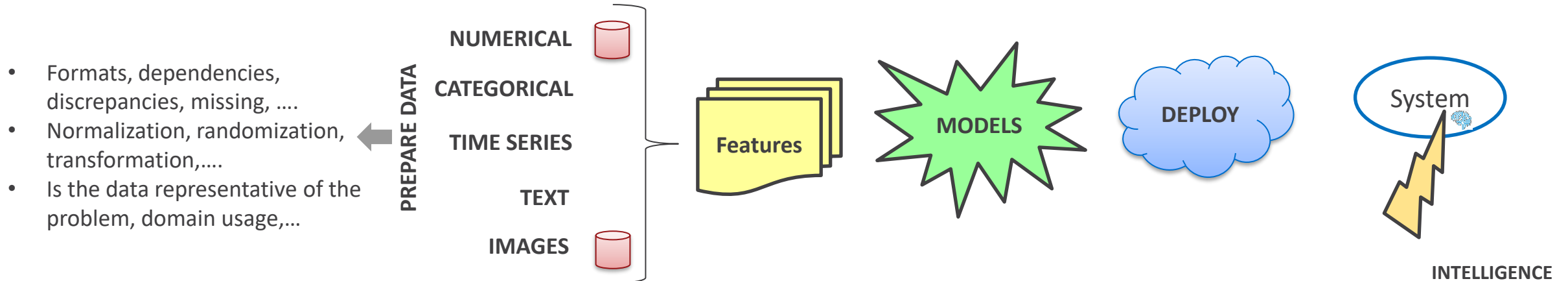
<https://pixabay.com/photos/data-amount-of-data-word-2723105/>

Context: Generic Workflow



But it's more than what meets the eye...

- What is the intelligence that is required for the system? What are the desired behaviors? What are the desired performance levels? What are the acceptable tolerances for errors/ false-positives/ false negative etc.?
- What is the source data on which the intelligence can be built? What are the data types? What are the restrictions on usage of the data? What should be the training, validation & test data sets? What could be potential features?
- What are the potential approaches to build the intelligence model? Neural Networks, SVM, Decision tree ... etc.? What are the specific topologies/ parameters that need to be decided/ configured?



Quick Quiz:

Training data in AIML applications is most similar to what in traditional software development:

- a. Test data
- b. Configuration settings
- c. Compiled object code
- d. Source code
- e. Database

Als are dangerous because they get things wrong for unknown reasons

Data Requirements: *The Green School Bus Problem*

AI image recognition is taught to identify military vehicles and differentiate civilian vehicles



<https://pixabay.com/photos/tank-panzer-battle-tank-gun-2729903/>, free



<https://pixabay.com/photos/military-lmtv-defense-afghanistan-165448/>, free



<https://pixabay.com/photos/us-army-united-states-army-humvee-2526752/>, free



<https://pixabay.com/photos/us-army-united-states-army-oshkosh-2526749/>, free



<https://pixabay.com/photos/transport-traffic-vehicle-bus-4405087/>, free



<https://pixabay.com/photos/suv-car-vehicle-jeep-travel-1353451/>, free

Now, into the field of view wanders this:

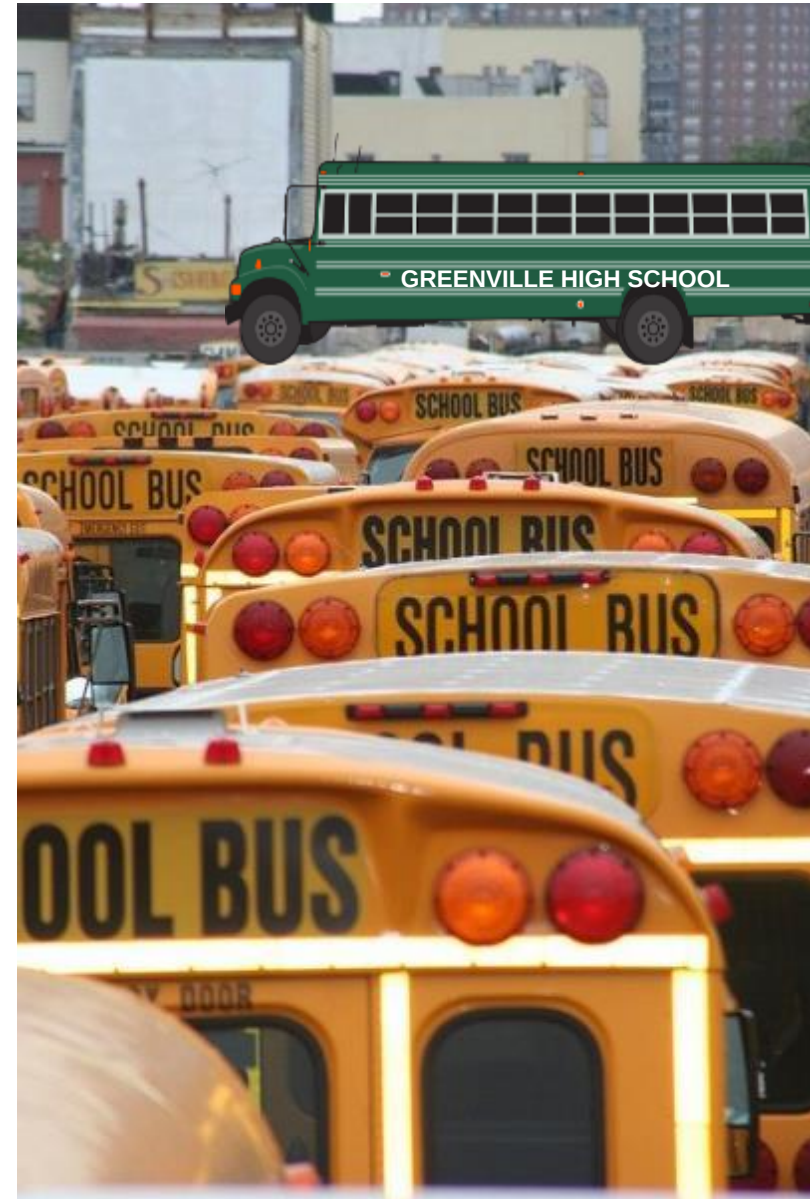


<https://pixabay.com/vectors/green-bus-bus-green-vehicle-auto-3749394/>, FREE

Training data is more like source code to an AI ML System

How will the green school bus be identified?

- a) School bus
- b) Tank
- c) Aston Martin DB9
- d) Military Truck
- e) Sort of like a school bus, but a color never seen before
- f) Image not recognized



<https://pixabay.com/photos/school-bus-america-vehicles-school-600270/>

Imagining the Nightmare Headline



Who is at fault?

- a) Someone else (not me)
- b) Systems engineer
- c) Data scientist
- d) ML developer
- e) Tester
- f) Program manager

Data and Bias

Systems Thinking: use an **unreal world** to counter bias

- Bias in the **world** vs. bias in the **data**
Application: identify male nurses and female nurses in photos
- In the world: 93% of nurses are female
- Should data consist of 93% female nurse photos?
- A “real world” dataset might do poorly on male nurses
- Expand application to all hospital personnel
 - Among MDs, 50% are female
 - In our **real world** set, most men are doctors and most women are nurses!
 - Likely misclassification of male nurses as doctors and female doctors as nurses!
- Better dataset would be 50/50 on both m/f doctors and nurses
- Can use conditional probability to reason about data, e.g. $P(\text{female} \mid \text{nu})$

	MRN
MMD	
FMD	FRN



Toward a Data Requirements Approach

- New kind of requirements, similar to the “-ilities” (reliability, maintainability, dependability, safety, security)
- Must develop specific language to describe data requirements
 - Not “System shall use the right data...”
 - Just like not “System shall be easy to use”
 - Or, “System shall be explainable”

Data Requirements

Not your DADS requirements

D – Diversity

A – Augmentation

D – Distribution

S - Synthesis

D – Diversity of Data

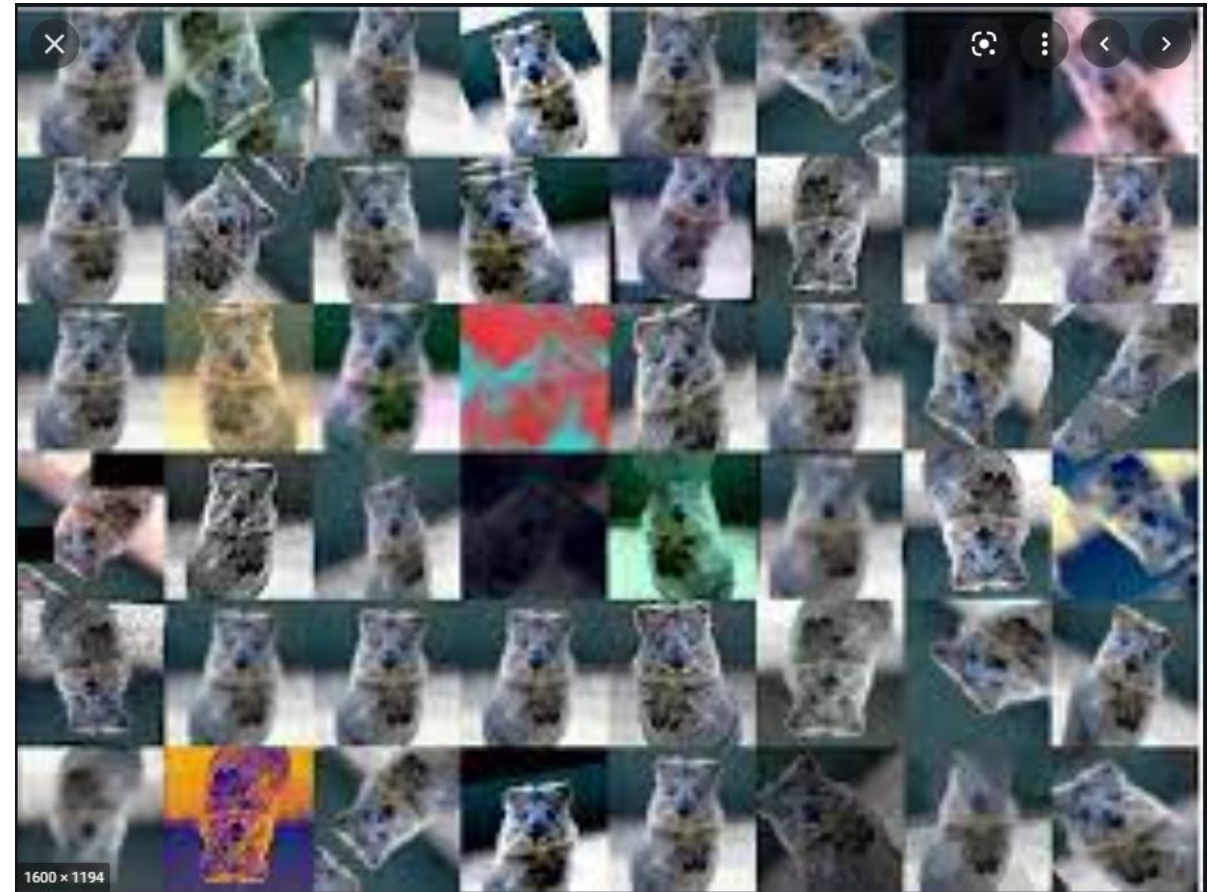
- AI does not generalize as we do
 - Counter-intuitive from human perspective; child can generalize with VERY few examples
- Need images of every size, type and color
- Must understand what is expected in the intended environment
- Systems Engineers must specify the data and its diversity



<https://pixabay.com/photos/kid-dog-outdoors-little-girl-5718703/>

A – Augmentation of Data

- Software-generated variations of training data
- Ideally, include every variation that AI might see in deployed application
 - Colors, lighting, angles, backgrounds
 - Angle can change apparent shape of object
 - Not all angles are likely
- Vary the images to “teach” the AI that the category variations are still the same category
- Systems engineers must specify what variations should be included
- Data scientists can suggest others to improve model accuracy
 - What augmentation might be **required** for this application?



D – Distribution of Data

- Must have enough images of each required category to train system well
- Must not introduce bias due to insufficient training data
- Evaluate by measuring accuracy on various classes
- Systems engineers must be aware of “non-essential” distribution in the world, e.g. m/f nurses vs. “essential” features like tanks have tracks



<https://pxhere.com/en/photo/815312> Public Domain

S – Synthesis of New of Data

- New training data generated by computer
- Synthesized data can result in effective training
- Example: horses and humans dataset, CGI Photoreal generation (Laurence Moroney)
- Rules, scenarios and other methods can generate example data for training, then allow the AI to learn from it



<https://laurencemoroney.com/datasets.html>, CC0

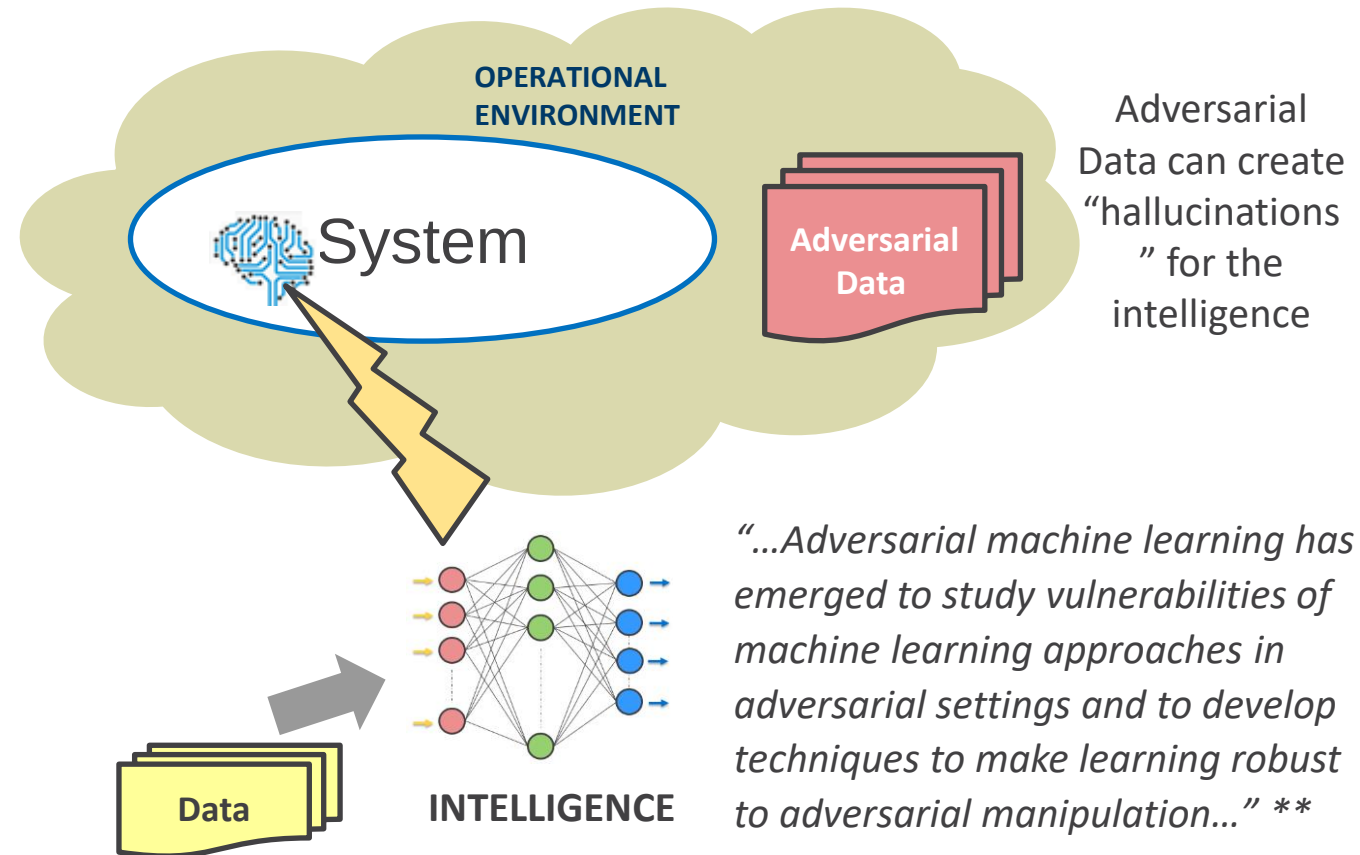
Intelligence Implications of Data

Intelligence is as good as the data used to develop the intelligence....

- Is the operational environment benign?
- Is the “Data” used for developing the intelligence representative of the operational environment?

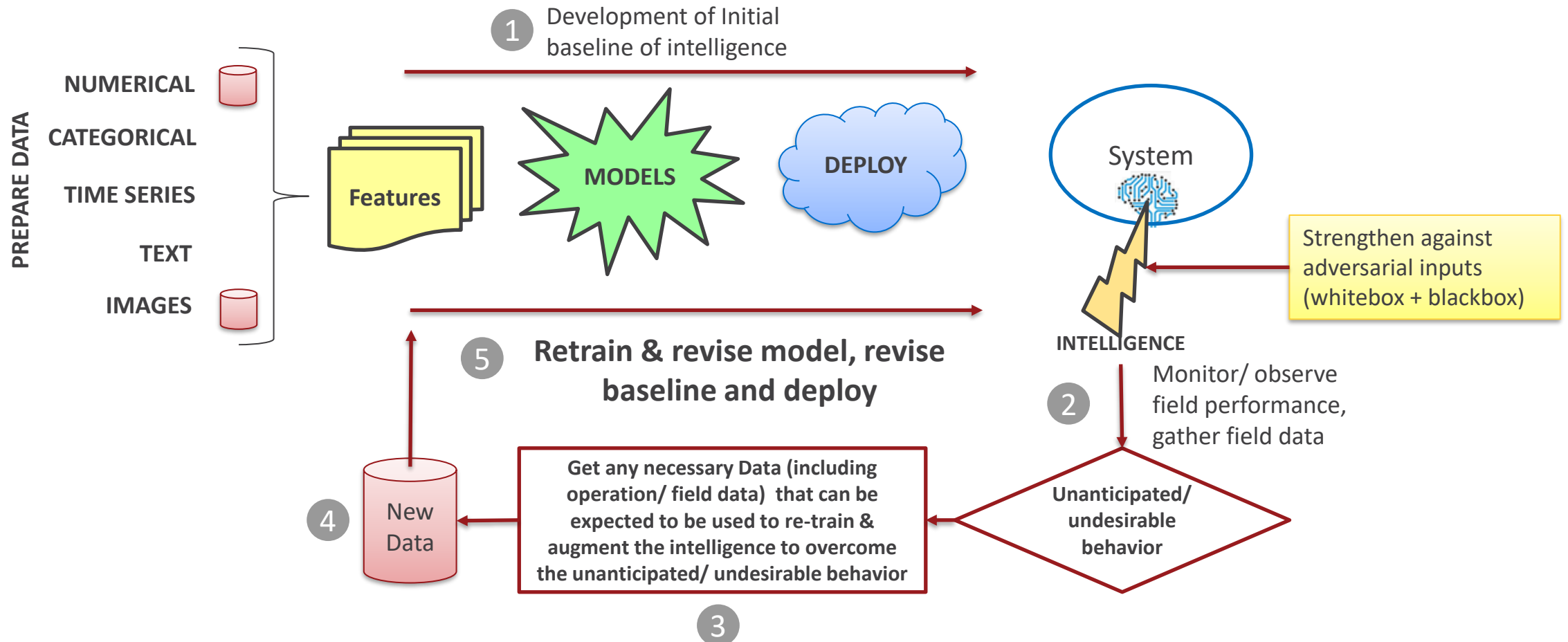


<https://www.wired.com/story/researcher-fooled-a-google-ai-into-thinking-a-rifle-was-a-helicopter/>



** Vorobeychik, Y. and Kantarcioglu, M., 2018. Adversarial machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 12(3)

Intelligence – Baselines & Updates



Wrapping Up

■ Key take-aways

- Systems engineers must be aware of the role of training data
- Systems engineers must specify data based on the application and environment, with collaboration from data scientists
- Key areas of data requirements (DADS)
 - Diversity
 - Augmentation
 - Distribution
 - Synthesis
- Data has its own lifecycle that must be managed within the larger system development lifecycle



<http://marubon-ds.blogspot.com/2017/09/fashion-mnist-exploring.html>