

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



Barclay R. Brown, Ph.D, ESEP
Assoc. Dir AI Research
Collins Aerospace
Barclay.brown@incose.net

- 
- **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?**

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

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



An Exercise in Image Recognition

Who Doesn't Love Trees?

Elm



Maple



Pine



Oak



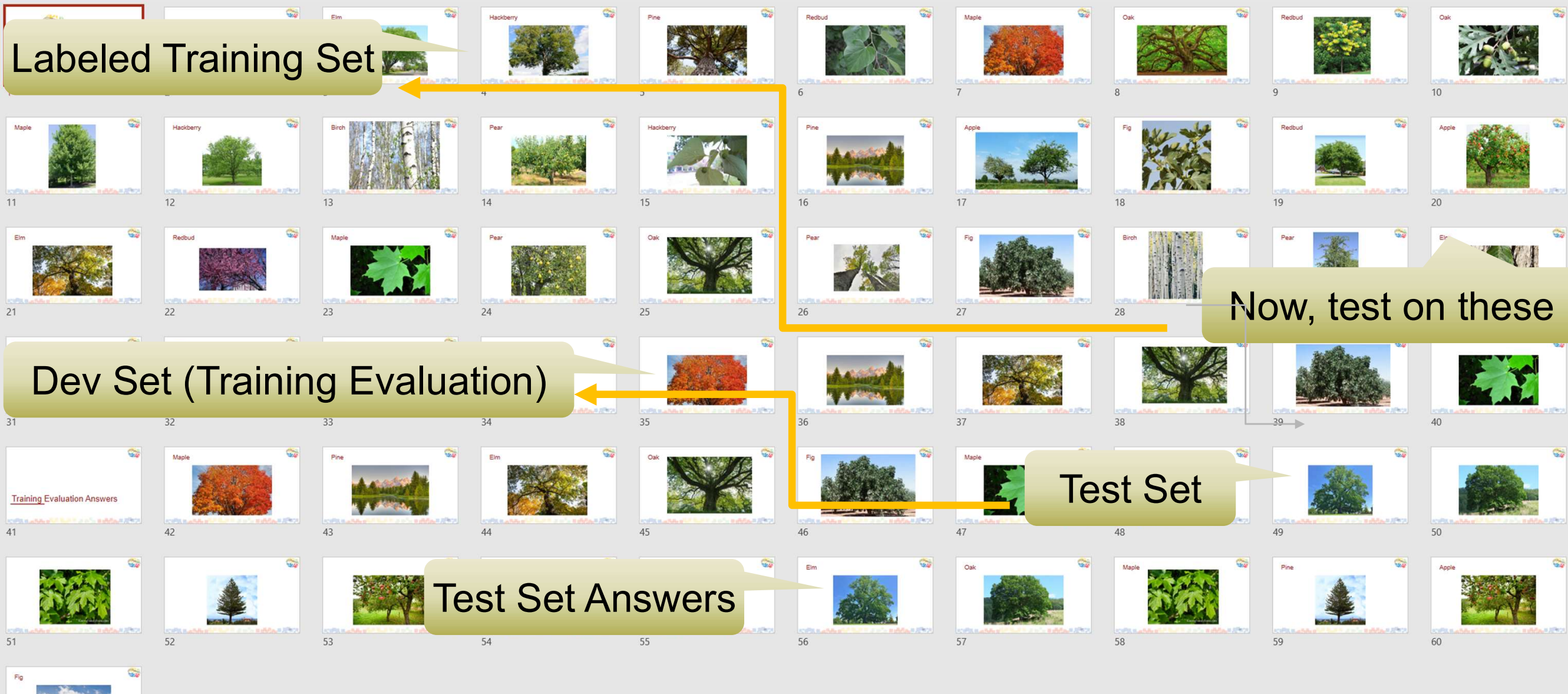








Developing Sympathy for the ML Model



Bottom Line: Systems Engineers must understand how training data works

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

MMD	MRN
FMD	FRN



<http://www.freestockphotos.biz/stockphoto/15411>, PD

<https://www.publicdomainpictures.net/en/view-image.php?image=209297&picture=nurse>, CC0

Dr. Brown's Magic Elixir

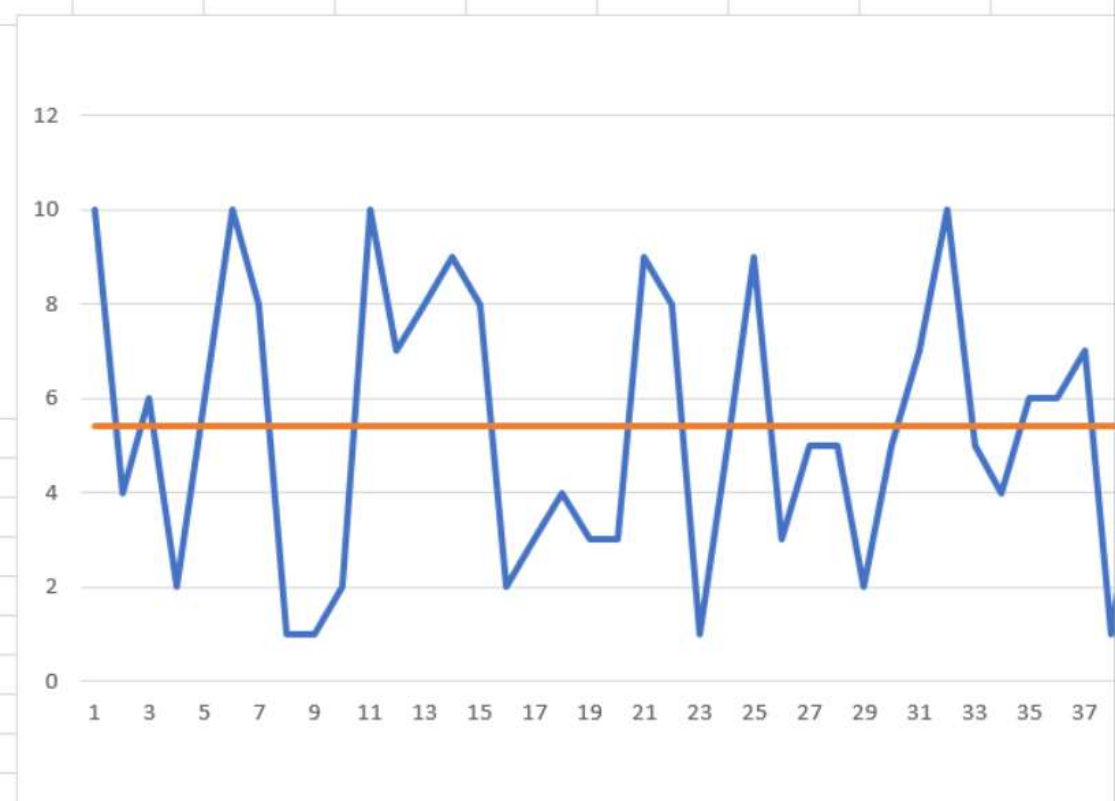
- Does it work?
- Chronic health problem
 - Have a normal level, and good/bad days
 - On bad days, take Dr. Brown's Elixir
 - Measure if feeling better:
 - Next Day
 - 2nd Day
- What happens?



Clinical Trial Results – Dr. Brown’s Elixir

RESULTS: % improved after bad day = **71%** % Improved by 2nd day = **92%**

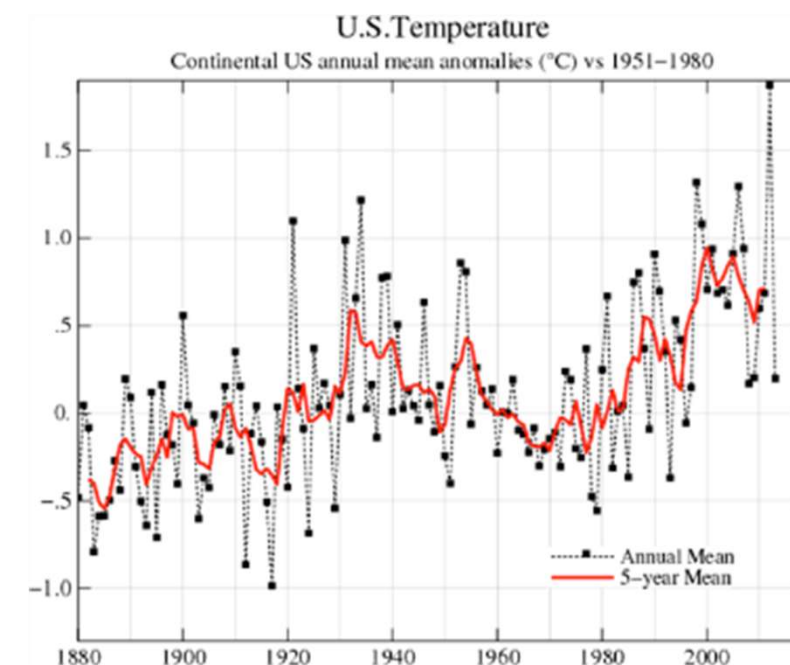
Day	Today's Change	Health Rating	Mean	Feeling bad yesterday (took Elixir)?	Feeling better today?	Feeling good tomorrow?	Feeling good today after feeling bad yesterday?	Feeling good tomorrow after feeling bad yesterday?	Feeling good either today or tomorrow after feeling bad yesterday?
1		5	5.56						
2	2.00	7	5.56	Y	Y	Y	Y	Y	Y
3	-1.00	6	5.56	-	-	Y	-	-	-
4	3.00	9	5.56	-	Y	-	-	-	-
5	-5.00	4	5.56	-	-	-	-	-	-
6	1.00	5	5.56	Y	Y	Y	Y	Y	Y
7	2.00	7	5.56	Y	Y	Y	Y	Y	Y
8	3.00	10	5.56	-	Y	-	-	-	-
9	-4.00	6	5.56	-	-	-	-	-	-
10	-4.00	2	5.56	-	-	-	-	-	-
11	1.00	3	5.56	Y	Y	Y	Y	Y	Y
12	2.00	5	5.56	Y	Y	-	Y	-	Y
13	-2.00	3	5.56	Y	-	Y	-	Y	Y



In trial after trial, over 2/3 felt better the next day and virtually all felt better by the 2nd day after taking the Elixir when they felt bad!

Regression to the Mean

- “Regression to the mean is a widespread statistical phenomenon with potentially serious implications for health care. It can result in wrongly concluding that an effect is due to treatment when it is due to chance. Ignorance of the problem will lead to errors in decision making.” (NIH)
- Francis Galton 1886, measured heights of children and parents
 - Taller parents had shorter children, while shorter parents had taller children!
 - Called it “regression to mediocrity”
- “In any series with complex phenomena that are dependent on many variables, where chance is involved, extreme outcomes tend to be followed by more moderate ones.” (Parrish)
- Danger: Leads to unfounded reasoning about causes
- Kahneman: announcers at Olympics gave reasons why a skier’s 2nd jump would be worse after a great first jump



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”
 - Nor, “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
- Ideally, include every variation that AI might see in deployed application
 - Colors, lighting, angles, backgrounds
 - Angle can change shape
 - Not all angles are likely

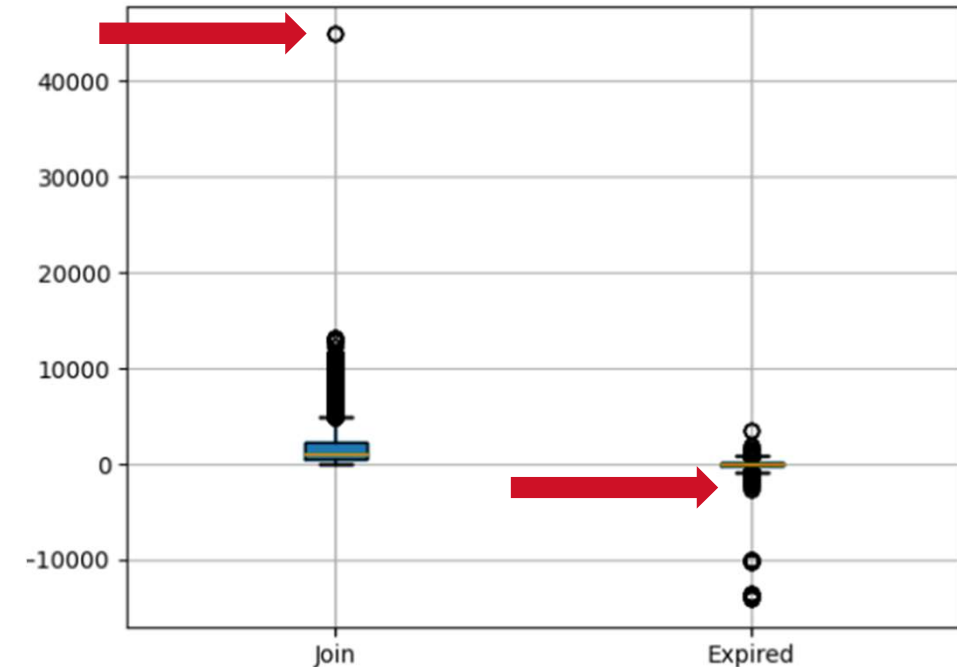
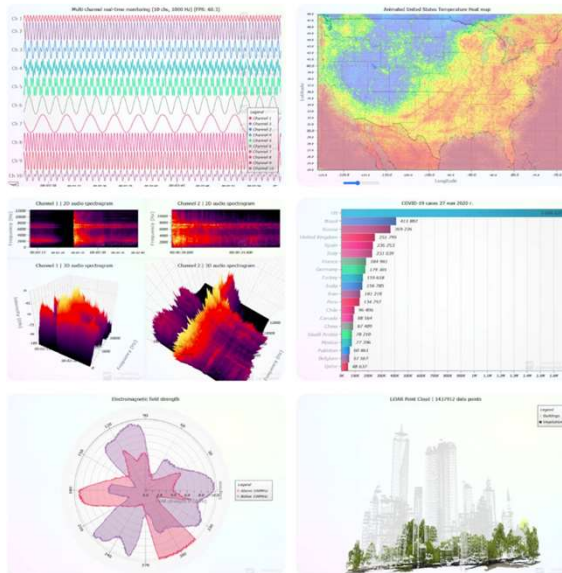


<https://pixabay.com/photos/kid-dog-outdoors-little-girl-5718703/>

Data Diversity Metrics

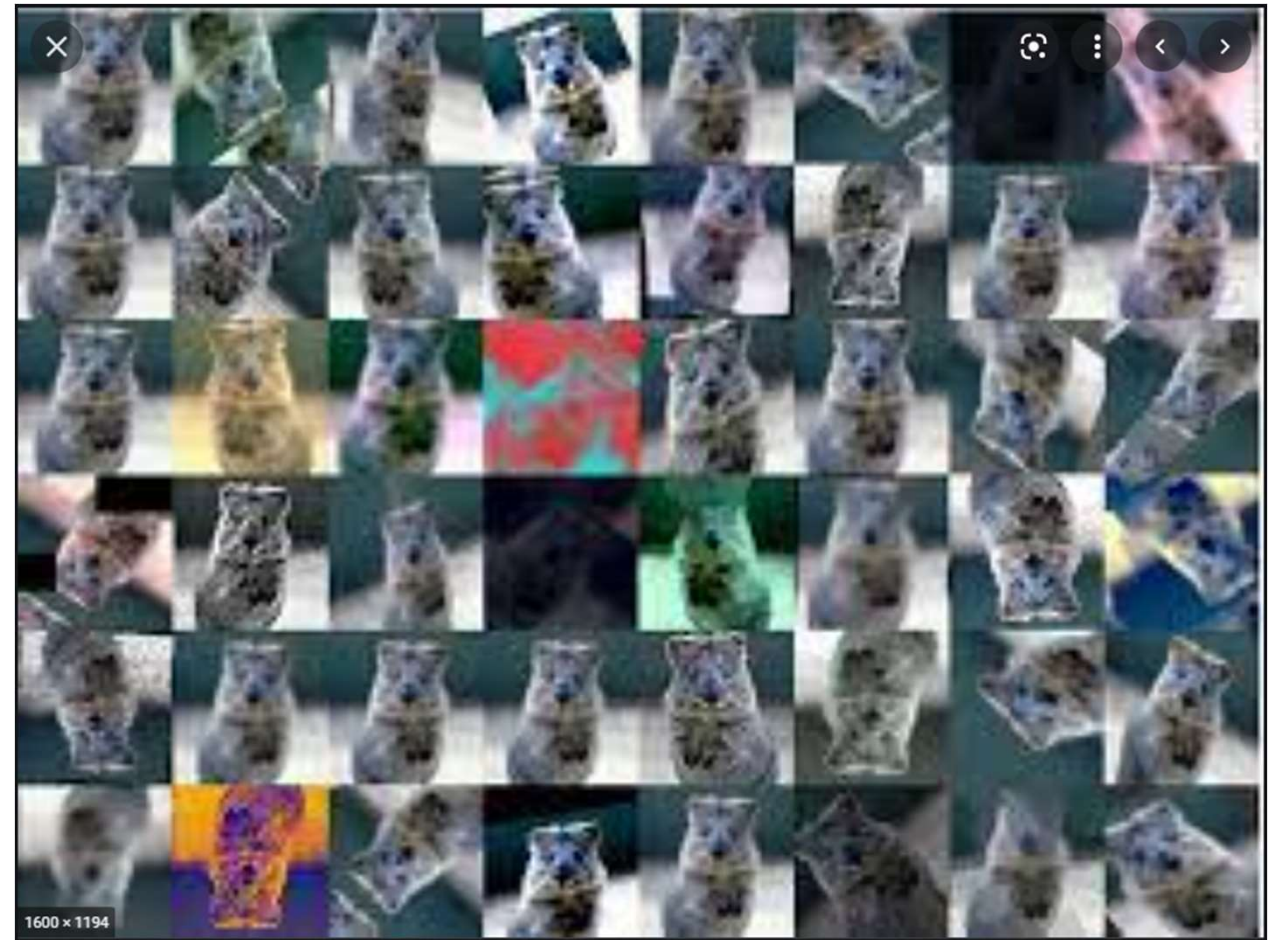
- Quick frequency table
- Good old mean, median, mode
 - Mean = 1718
 - Median = 996
 - Mode = 0 (count 464)
- Reasonableness ranges
 - 32 years is 11,680
 - Temperature on earth < 134°F
 - Human body weight < 1311 lbs.
- Boxplot
 - Shows median, quartiles, outliers
- Consider multiple visualizations
- Become a data detective
- What to do with bad or missing data?
 - Drop the entire record?
 - Set to zero?
 - Set to mean of other values?
 - Interpolate?

```
mem.dayssincejoin.value_counts(bins=10, dropna=False)
(-41.818999999999996, 4484.8]      18961
(4484.8, 8966.6]                  1711
(8966.6, 13448.4]                 268
(40339.2, 44821.0]                1
(13448.4, 17930.2]                0
(17930.2, 22412.0]                0
(22412.0, 26893.8]                0
(26893.8, 31375.6]                0
(31375.6, 35857.4]                0
(35857.4, 40339.2]                0
```



A – Augmentation of Data

- AI does not generalize as we do
 - Counter-intuitive from human perspective; child can generalize with VERY few examples
- Ideally, include every variation that AI might see in deployed application
 - Colors, lighting, angles, backgrounds
 - Angle can change shape
 - Not all angles are likely
- What augmentation might be **required** for this application?



<https://www.google.com/url?sa=i&url=https%3A%2F%2Falgorithmia.com%2Fblog%2Fintr-education-to-dataset-augmentation-and-expansion&psig=AOvVaw3gnEhAxPR9Ys3nm4KleqP&ust=1632917656856000&source=images&cd=vfe&ved=0CA0Q3YKBAhckEwj9Zf0qHzAhUAAAAAHQAAAAAQAw>

Augmentation Metrics

- Augmentation to enhance “color-blindness”
 - Green school bus, and orange, and purple, and pink, and...
 - In effect, “don’t pay attention to the color of the bus”
- Augmentation to better represent reality
 - Upside down cats: yes
 - Upside down trucks: probably not
 - Upside down ships: no thanks
- Augmentation of viewpoint
 - Consider viewpoints in the data
 - From what angles will images need to be recognized?
- Image segments, partial views, occluded views
 - Can simulate by cutting up complete images



Will the real Ferrari F355 Spider please rev up?

D – Distribution of Data

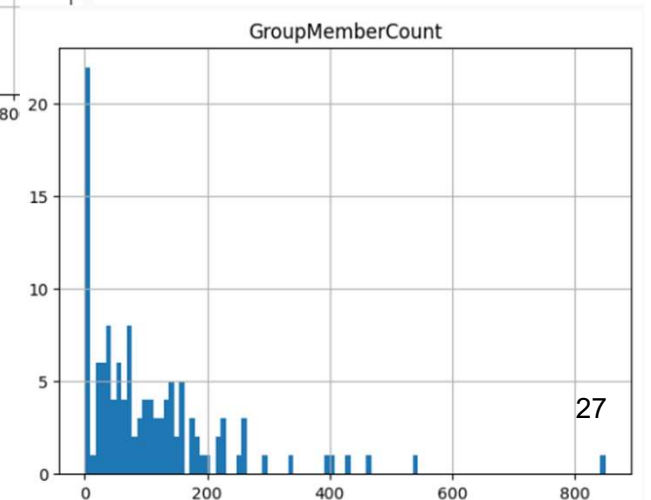
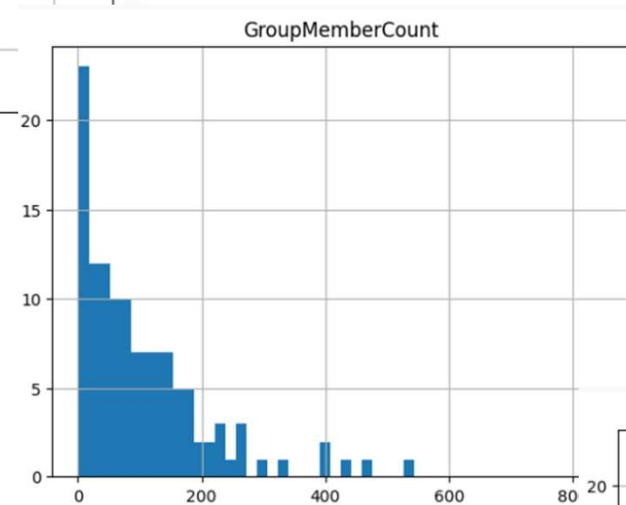
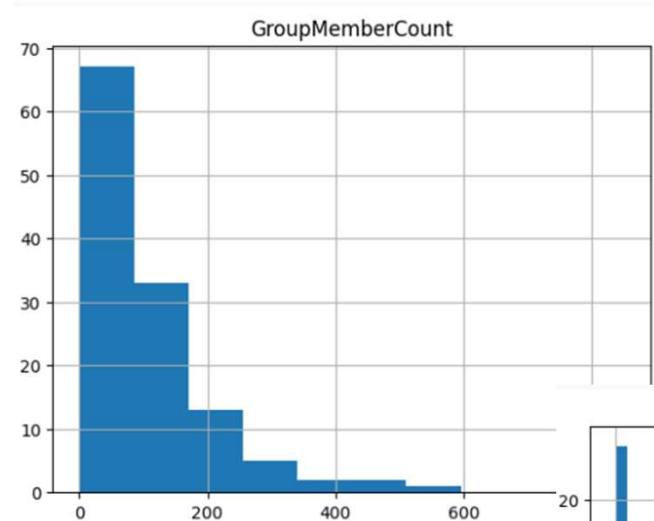
- Must have enough images of each required category to train system well
- Must not introduce bias due to insufficient training data
- Evaluate by measuring accuracy on various classes



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

Distribution Metrics

- Classic histogram
 - Number of bins can be revealing
- Don't forget per-capita
- Which cities have the most cancer deaths? (The largest ones)
- Check for bias in the data and evaluate based on the application



Revisiting Nurses and Doctors

- Nurses outnumber docs 3:1
- 93% of nurses are women
- 50 % docs are women
- Pick a woman at random
 - 84% chance a nurse
 - 16% chance a doc
- Pick a man at random
 - 72% chance a doc
 - 28% chance a nurse
- Pick a person at random
 - 75% chance a nurse
 - 25% chance a doc
 - 70% chance a female nurse

If an “AI” system simply predicted that EVERYONE is a female nurse, it would be correct

70%

Of the time.

So, any effective system must be better than that.

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

Data Synthesis Metrics

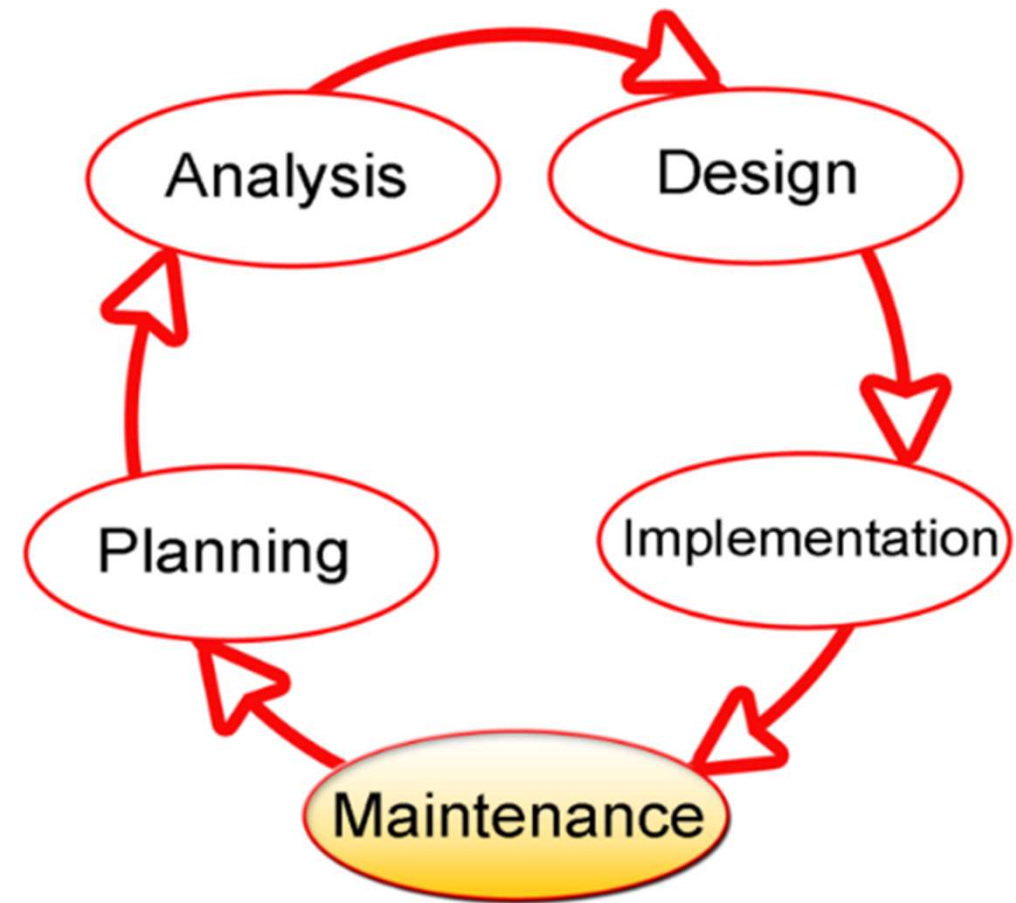
- Evaluate whether addition of synthetic data increases accuracy or not
- Include synthetic data is training data, but not in test data
 - No point in assessing whether system can identify synthetic images



Fish? Bird? Mammal?

Training Data and the Lifecycle

- Apply systems development lifecycle to Data used in training
- Requirements / Planning
 - What data is needed
- Analysis
 - Coverage, negative examples, adversarial, edge cases, bias in data
- Design of Data
 - How to use the data
 - Augment, Synthesize
- Implementation



<https://commons.wikimedia.org/wiki/File:SDLC-Maintenance-Highlighted.png> CC