# From POC to Production in Minimal Time – Avoiding Pain in ML Projects

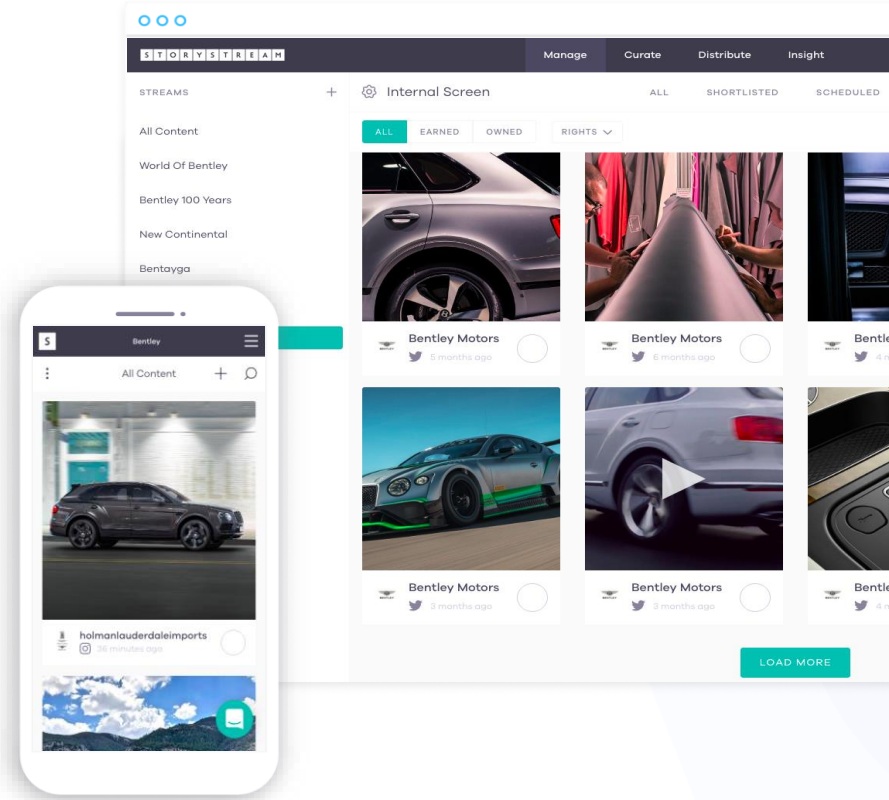

**Dr Janet Bastiman**
**@yssybyl**


"WELL, WHEN IT COMES DOWN TO ME AGAINST A SITUATION, I DON'T LIKE THE SITUATION TO WIN."

About StoryStream

## The world’s leading automotive content platform

StoryStream is a dedicated automotive content platform, trusted by some of the world’s leading car brands. Specifically created to help automotive brands provide a more relevant, engaging customer experience, fuelled with authentic content and designed for efficiently scaling content operations across global teams.

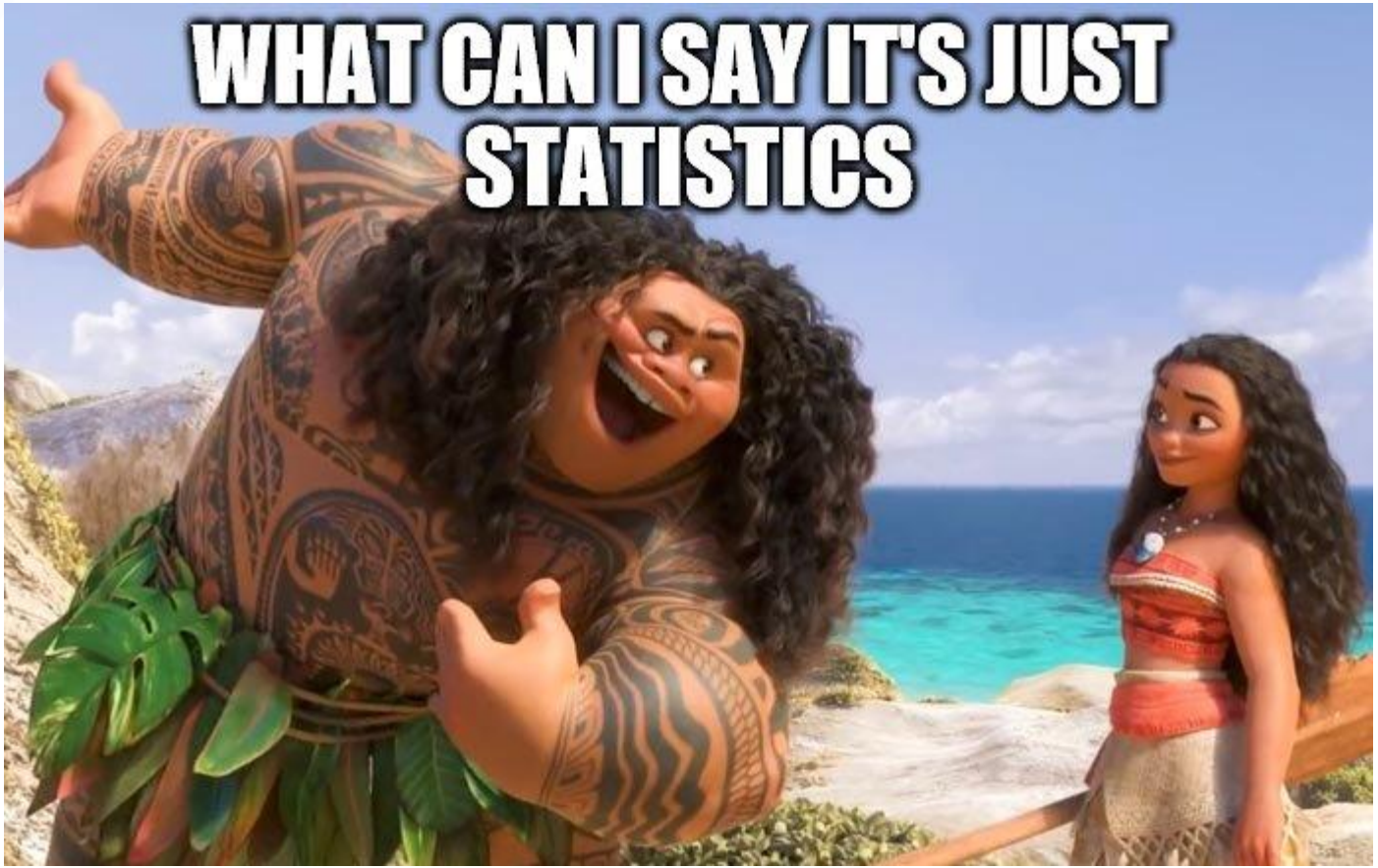



## The Core StoryStream Benefits

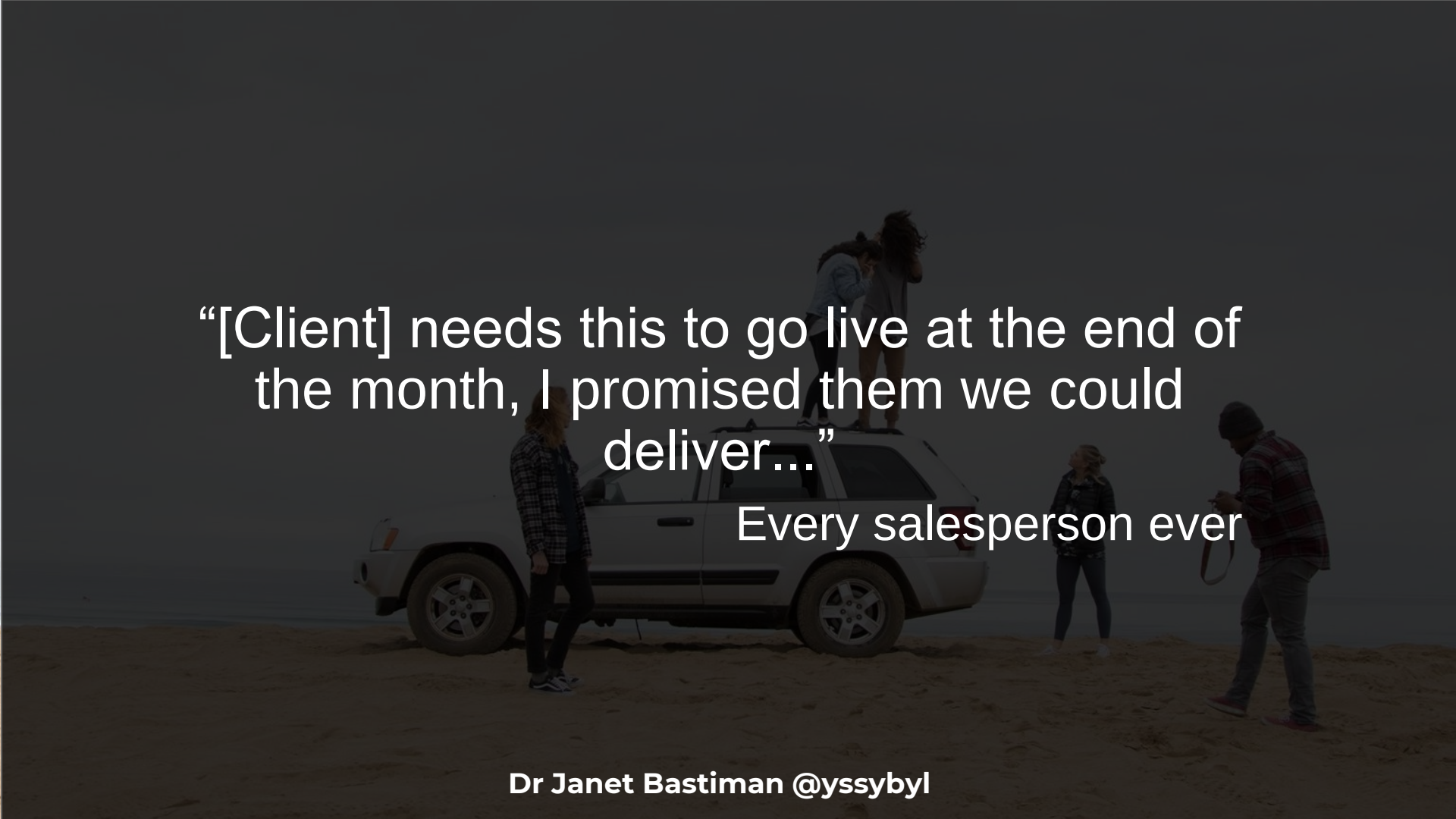
- Grow customer engagement and conversions by up to 25%
- Reduce content creation and management costs by up to 60%
- Provide a more authentic customer experience
- Understand your customer in a deeper way


GIRL
GUIDES


01001100
01000101
01000111
01001111




**Dr Janet Bastiman @yssybyl**

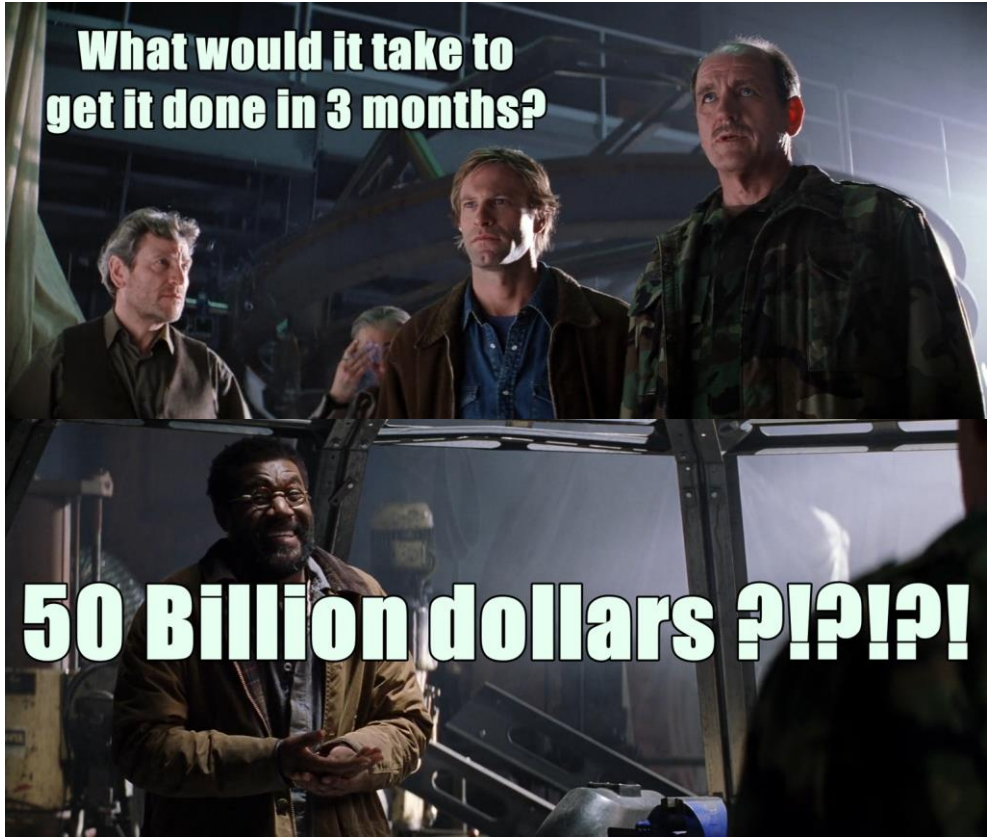

“[Client] needs this to go live at the end of the month, I promised them we could deliver...”
Every salesperson ever
Dr Janet Bastiman @yssybyl

# Project timings

- 35 models = 1050 days (one person linear)
- ~ 5 years for one person working Mon-Fri - who is allowed holidays :)
- 250 days with parallelisation of tasks and data upfront
- 150 days on worksheet, balanced by an increase in ongoing license


Can you guess what happened next?
Dr Janet Bastiman @yssybyl

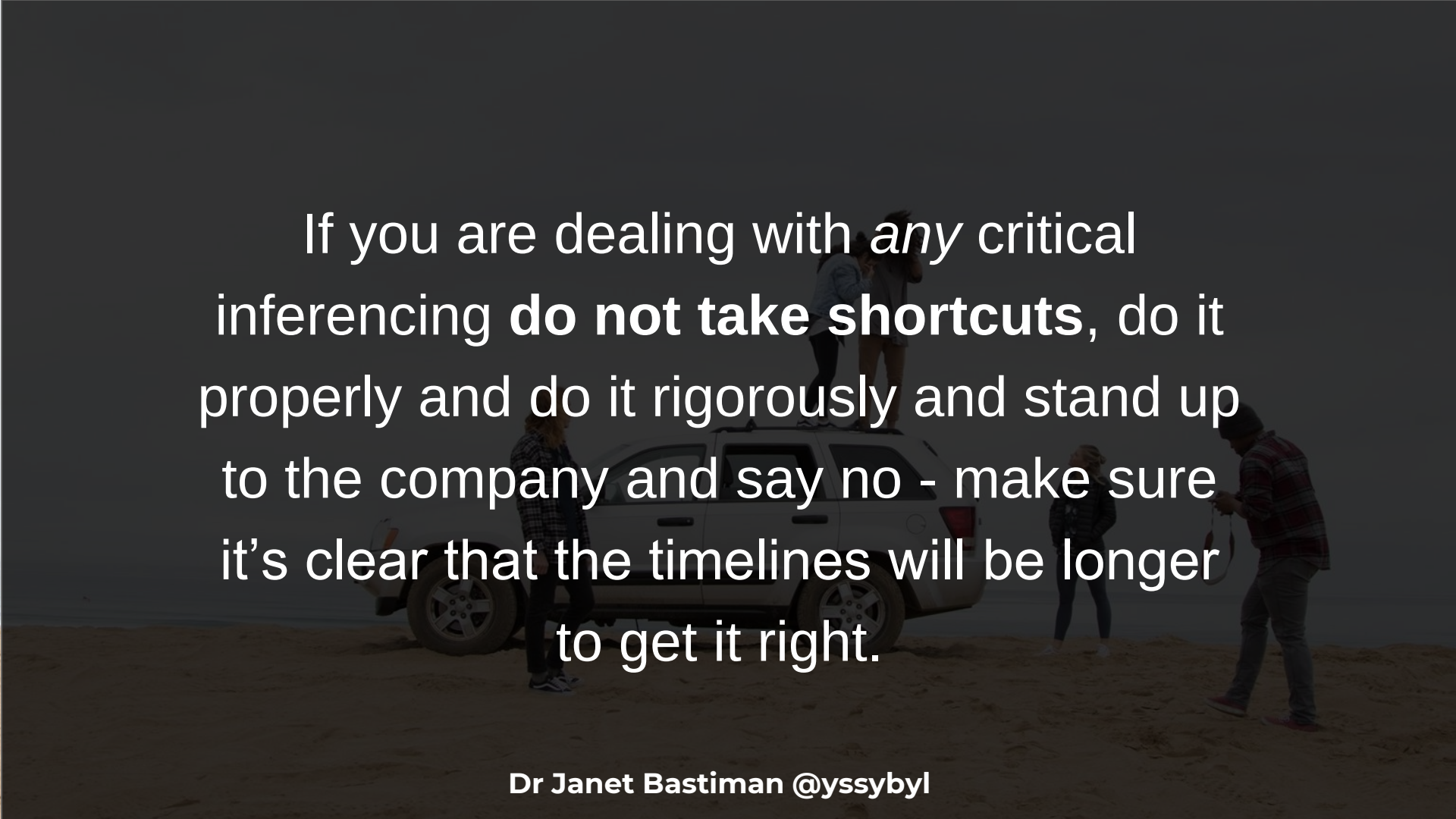
# What would it take to get it done in that time?




The Core (2003)
Paramount Pictures

**Dr Janet Bastiman @yssybyl**


“They don’t have any data to give us”
Dr Janet Bastiman @yssybyl



If you are dealing with *any* critical inferencing **do not take shortcuts**, do it properly and do it rigorously and stand up to the company and say no - make sure it’s clear that the timelines will be longer to get it right.

**Dr Janet Bastiman @yssybyl**

# Without Data ML is just a Random Result

- Legal public sources
  - https://github.com/awesomedata/awesome-public-datasets
  - https://www.kaggle.com/datasets
- Take your own pictures/videos
  - access/permission?
  - Slow and inconsistent
- Scrape the client site *with permission*

# How much data?

- Vision: 1000 images per output class but depends on complexity of the problem
- Time series: at least double the time period over which you are predicting, but be cautious of data becoming irrelevant
- Text: very variable depending on the problem
- This also changes if you already have pre-trained networks that you're updating

# What do you do with the Data?

**Financial Times | Future of Mobility**
**VIP Invitation**
Today at 09:58

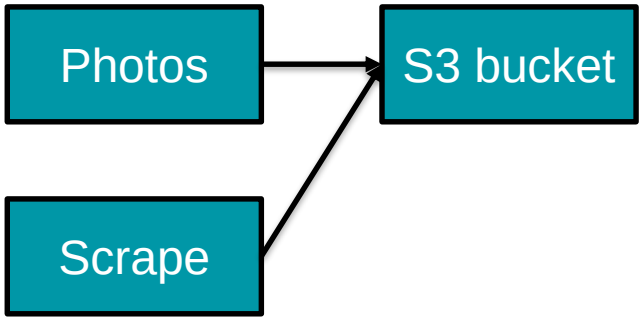
Dear Mr. Bastiman

I hope you are well.

- Selection bias
- Random Sampling
- Over coverage
- Undercoverage
- Measurement (Response) error
- Processing errors
- Participation bias


**Dr Janet Bastiman @yssybyl**

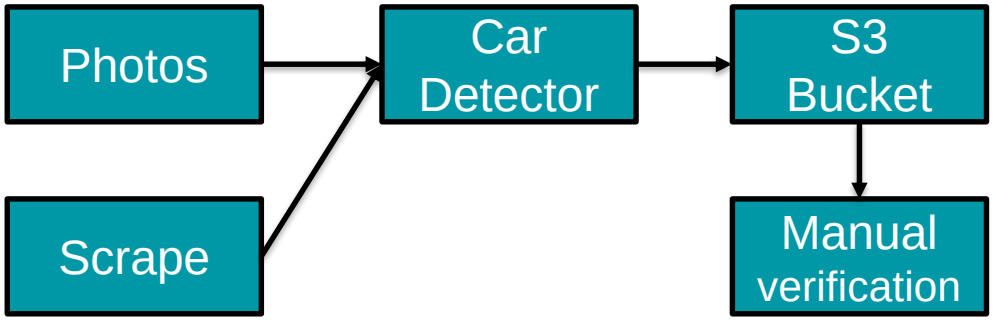
# What do you do with the Data?




- Unique filename
- source
- Set uuid (if multiple images of same car)
- Date taken

- S3 bucket per vehicle variant

# What do you do with the Data?




- Extra field for label
- S3 bucket name became mostly irrelevant

# Crowdsource labelling

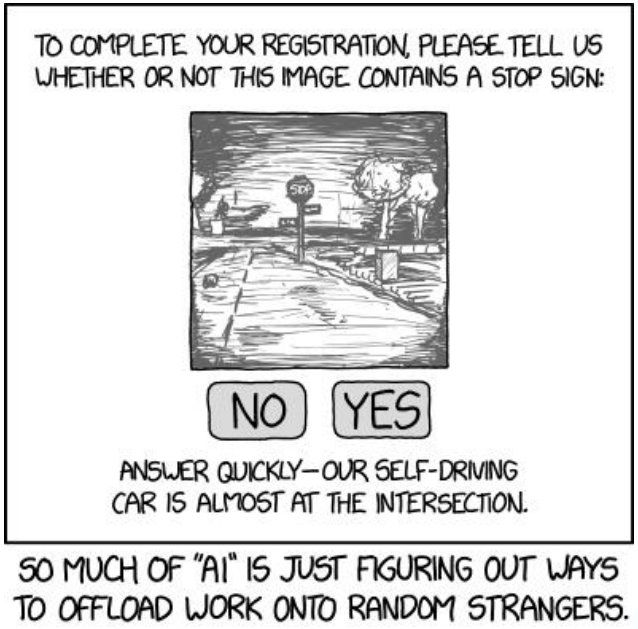
amazon mechanical turk

*Image Content:*

- Interior
- Exterior (more than half of the car visible)
- Close-up (less than half of the car visible)
- Close-up with text
- Obstructed
- Opening mechanism
- Wheel
- Unsure

*Image Quality:*

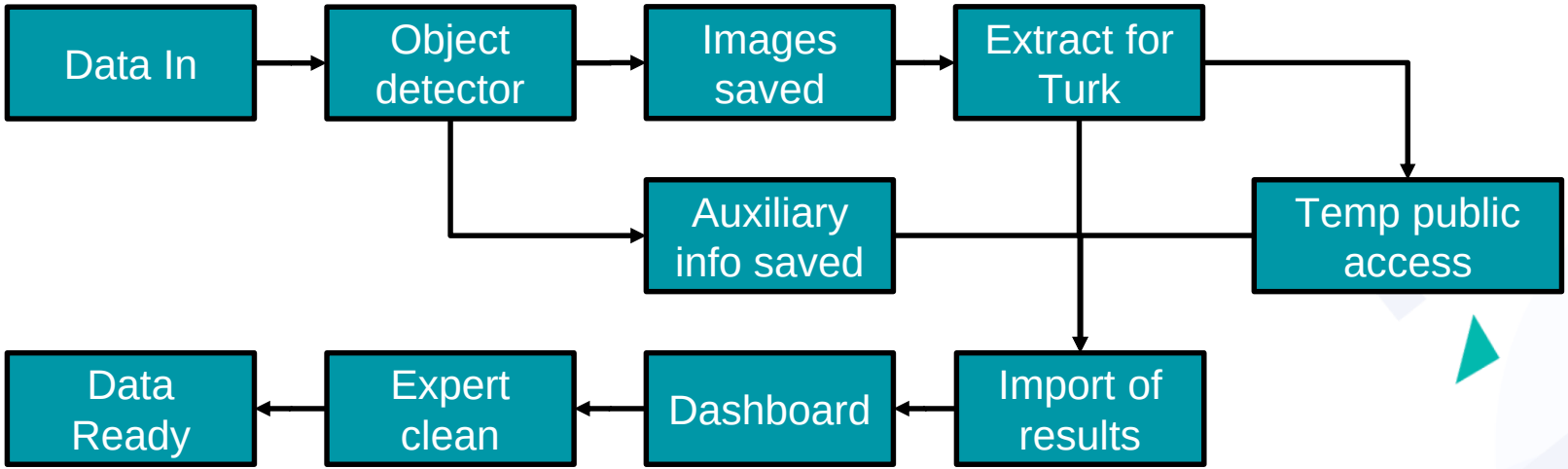
- Good
- Blurry
- Low resolution
- Distorted
- Other bad quality/Unsure




https://xkcd.com/1897/

**Dr Janet Bastiman @yssybyl**

# Data Pipeline

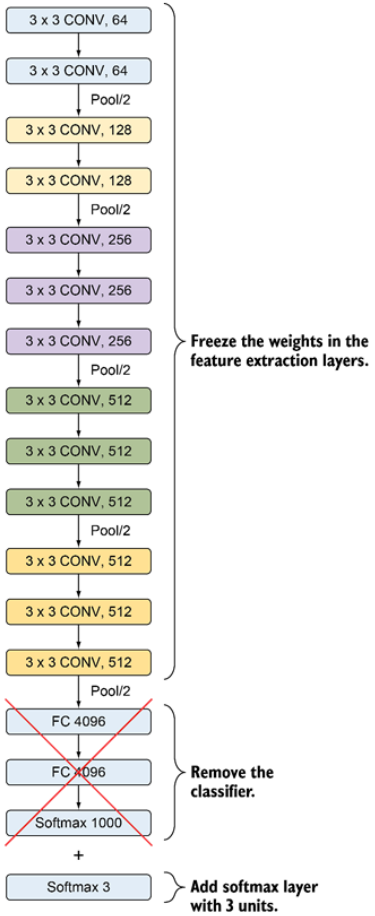



**Dr Janet Bastiman @yssybyl**

# Transfer Learning

- Use transfer learning - fix most of the weights of a good network and adapt the last few layers
- Fast and easy retraining and works with smaller data sets in a variety of fields

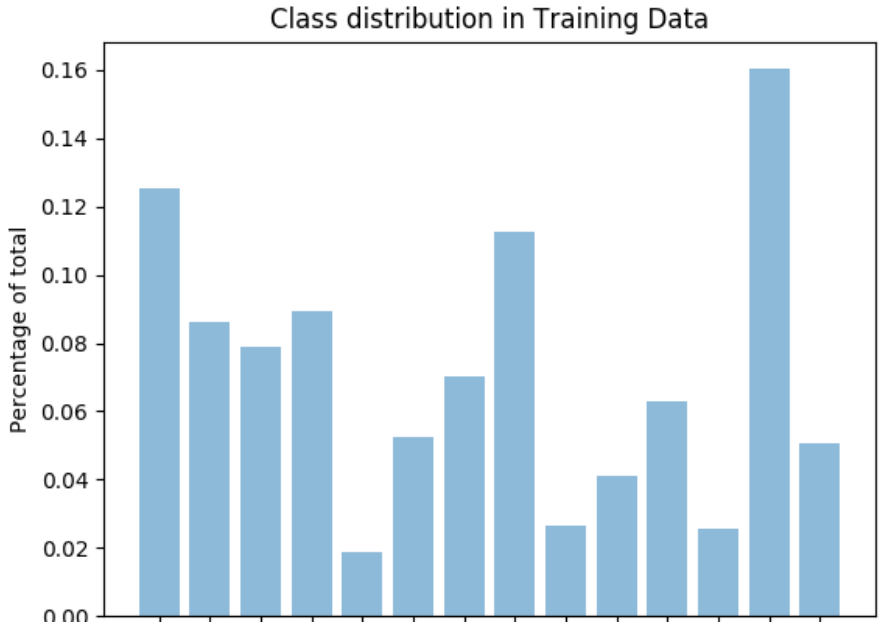
- (image) https://arxiv.org/abs/1903.02196
- (series) https://arxiv.org/abs/1907.01332
- (audio) https://arxiv.org/abs/1909.07526



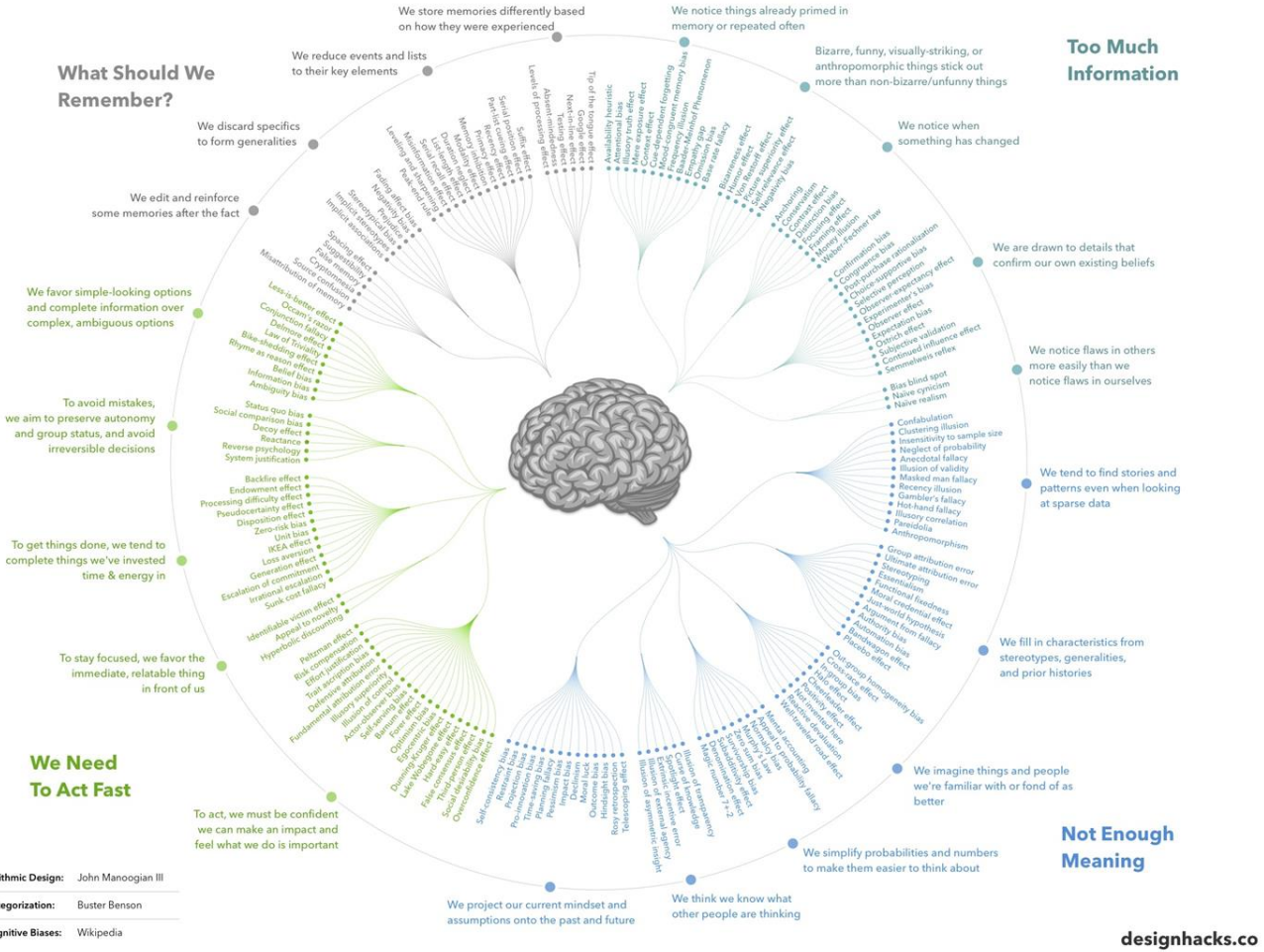

Deep Learning for Vision Systems, Mohamed Elgendy

**Dr Janet Bastiman @yssybyl**

# Unbalanced Data




**Dr Janet Bastiman @yssybyl**

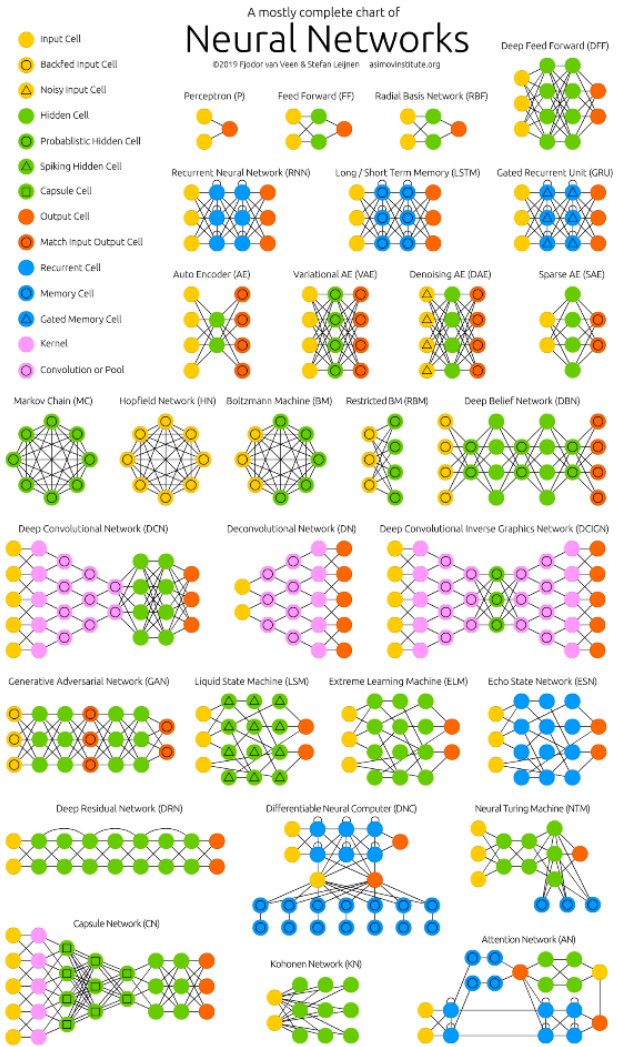



https://www.designhacks.co/products/cognitive-bias-codex-poster

# Stand on the shoulders of giants...

- For some problems CNNs are robust to noisy labels and up to 20 time noise to real labels can still give business level accuracy https://arxiv.org/pdf/1705.10694.pdf
- Find the right architecture

http://www.asimovinstitute.org/neural-network-zoo/



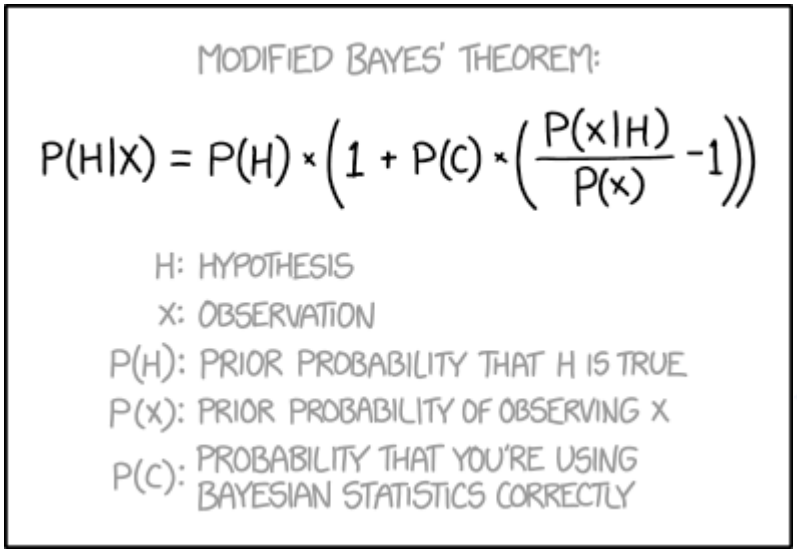

**Dr Janet Bastiman @yssybyl**

# Go old school

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\frac{\left\|\mathrm{x}_i - \mathrm{x}_j\right\|^2}{2\sigma^2}\right)$$

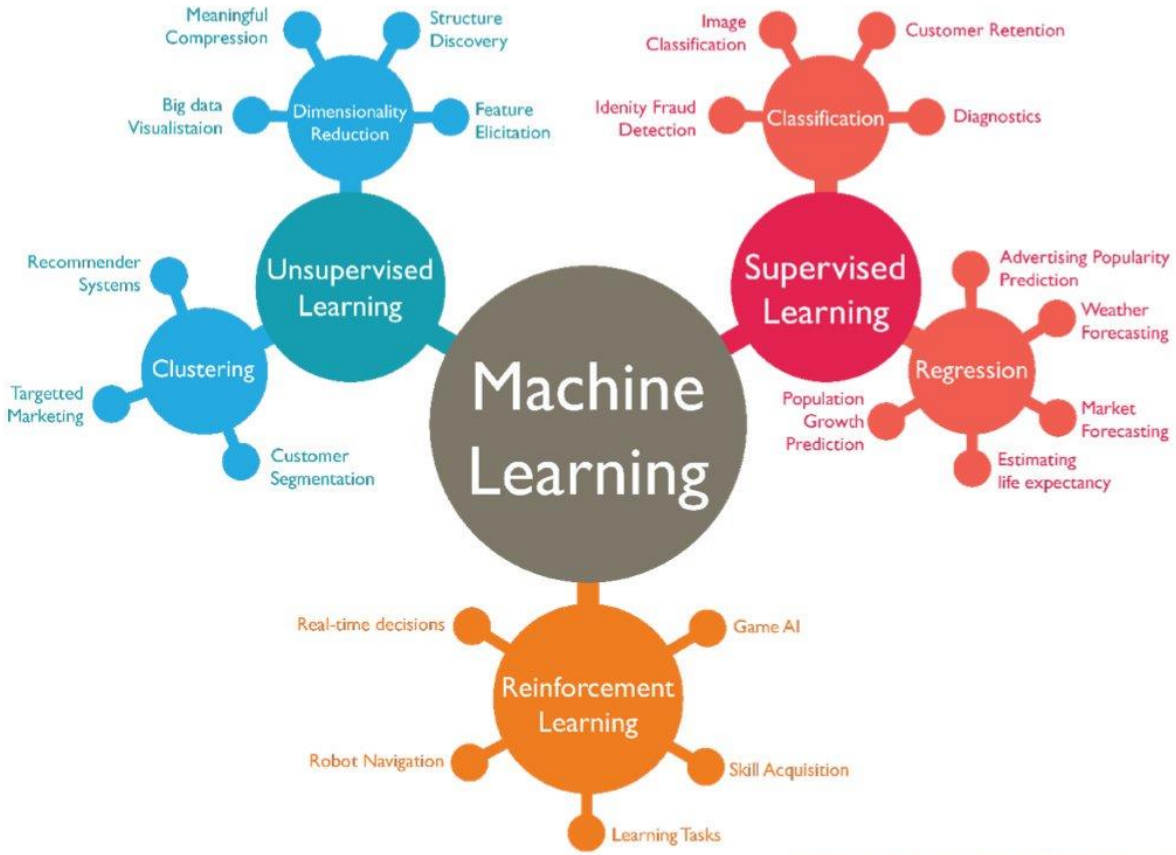



https://xkcd.com/2059/

Reduce the dimensionality of the problem and use Bayesian approach, KNN or SVM

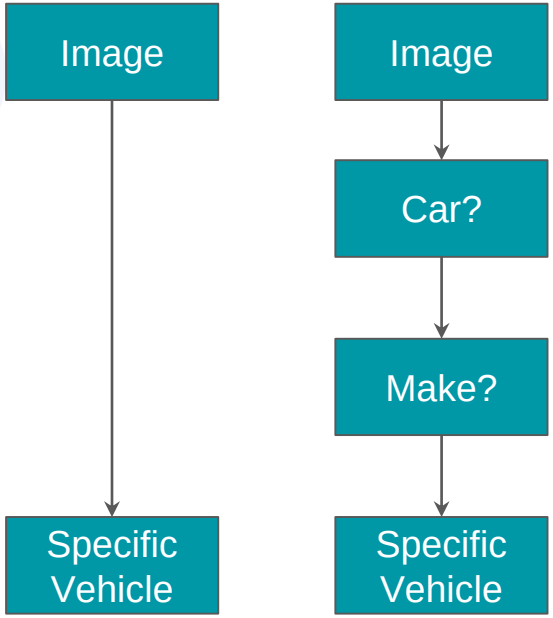
**Dr Janet Bastiman @yssybyl**

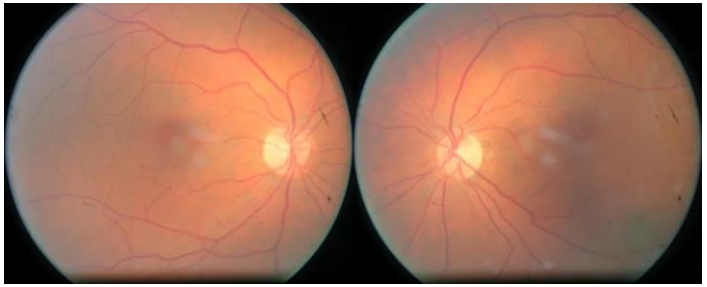
# TYPES OF ML/DL




source nvidia via @mikequindazzi

# Simplify the problem



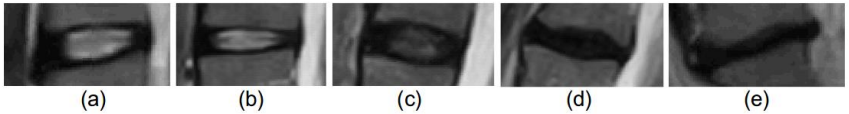
Removal of camera artefacts in eye images to make detection easier - Jeffrey De Fauw
http://blog.kaggle.com/2015/08/10/detecting-diabetic-retinopathy-in-eye-images/

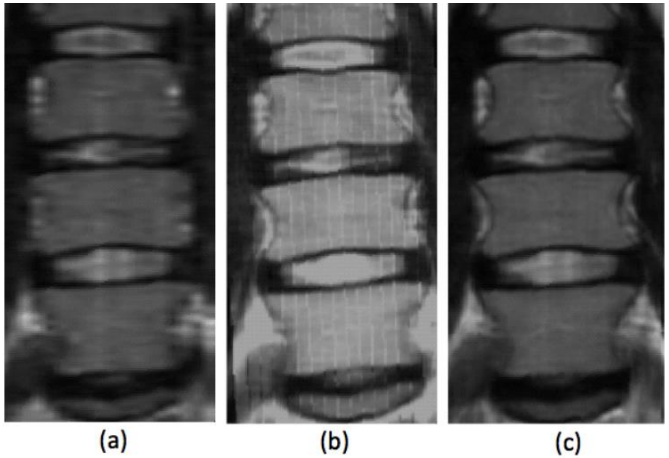
Removal of Doppler effect on moving source using fractional octave band shifting, F Mobley
https://asa.scitation.org/doi/pdf/10.1121/2.0000578?class=pdf

$$\Delta n = -\mathrm{r}[log_2(1 - M\cos\theta\sin\varphi)]$$

**Dr Janet Bastiman @yssybyl**

# Get every last drop from what you have




Statistical anatomical modelling for efficient and personalised spine biomechanical models - I Castro Mateos PhD thesis

Have a toolkit of augmentation approaches but choose what's relevant to your needs...

**Dr Janet Bastiman @yssybyl**

# Augmentation - detail

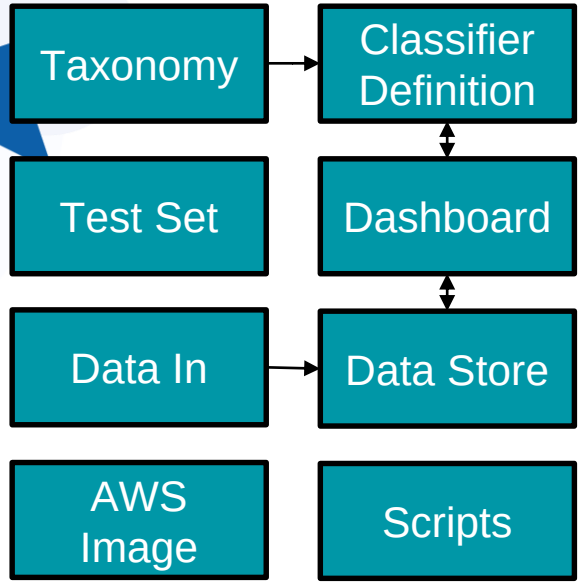
- Flip L/R U/D
- Rotations
- Reduce or enlarge bounding box coordinates by N%
- Add occlusions https://www.umbc.edu/rssipl/people/aplaza/Papers/Journals/2019.GRSL.Occlusion.pdf
- Change hue saturation and value of colours in the image https://arxiv.org/pdf/1902.06543.pdf
- Copypairing - https://arxiv.org/abs/1909.00390#

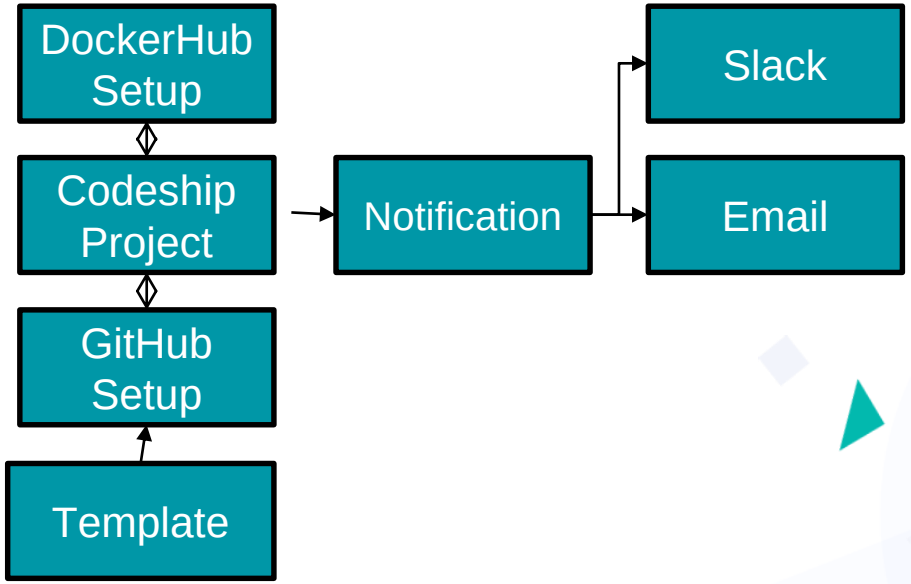


```
# Random flip augmentation
def im_flip(image, flip_prob=0.5):
    if np.random.binomial(1, flip_prob) == 1:
        image = np.fliplr(image)
    return image
```

**Dr Janet Bastiman @yssybyl**

# Infrastructure




**Dr Janet Bastiman @yssybyl**

# Cloud Formation

```
  ai-yolo-bb:
    Image: ai-yolo-bb
    Port: 9722
    Tag: 6ce337e844d2309d8c038c554e21f469ce52cdcf
    DesiredCount: 1
    ServiceName: ai-yolo-bb
```

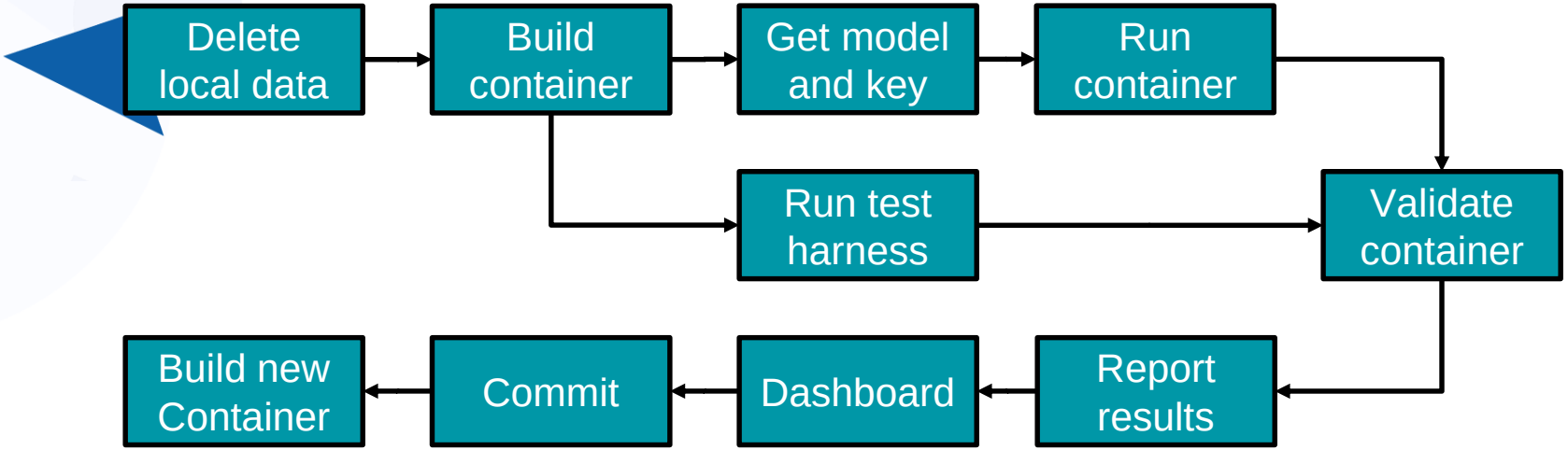
```
AIYoloBBTaskDefinition:
  Type: AWS::ECS::TaskDefinition
  DependsOn: LogGroup
  Properties:
    Family: !Join [ "-", [!FindInMap [Skills, ai-yolo-bb, Image], !FindInMap [StackMetaData, StackName, Slug]] ]
    Cpu: 1024
    Memory: 2048
    NetworkMode: awsvpc
    RequiresCompatibilities:
      - FARGATE
    ExecutionRoleArn: !Ref ExecutionRole
    ContainerDefinitions:
      - Name: !Join [ "-", [!FindInMap [Skills, ai-yolo-bb, Image], !FindInMap [StackMetaData, StackName, Slug]] ]
        RepositoryCredentials:
          CredentialsParameter: !FindInMap [StackMetaData, StackName, SecretsARN]
        Environment:
          - Name: AI_ENV
            Value: live
        Image: !Join ["", ["docker.io/storystream/", !FindInMap [Skills, ai-yolo-bb, Image], ":", !FindInMap [Skills, ai-yolo-bb
        PortMappings:
          - ContainerPort: !FindInMap [Skills, ai-yolo-bb, Port]
        LogConfiguration:
          LogDriver: awslogs
          Options:
            awslogs-group: !Ref LogGroup
            awslogs-region: !Ref AWS::Region
            awslogs-stream-prefix: !FindInMap [Skills, ai-yolo-bb, Image]
```

```
AIYoloBBServiceDiscovery:
  Type: AWS::ServiceDiscovery::Service
  Properties:
    Name: !FindInMap [Skills, ai-yolo-bb, ServiceName]
    HealthCheckCustomConfig:
      FailureThreshold: 1
    DnsConfig:
      DnsRecords:
        - Type: A
          TTL: 60
      NamespaceId: !GetAtt LocalNamespace.Id

AIYoloBBService:
  Type: AWS::ECS::Service
  DependsOn: [PrivateSubnetOneRouteTableAssociation, PrivateSubnetTwoRouteT
  Properties:
    ServiceName: !Join [ "-", [!FindInMap [Skills, ai-yolo-bb, Image], !Fin
    Cluster: !Ref ECSCluster
    LaunchType: FARGATE
    DesiredCount: !FindInMap [Skills, ai-yolo-bb, DesiredCount]
    ServiceRegistries:
      - RegistryArn: !GetAtt AIYoloBBServiceDiscovery.Arn
    NetworkConfiguration:
      AwsvpcConfiguration:
        AssignPublicIp: DISABLED
        SecurityGroups:
          - !Ref FargateContainerSecurityGroup
        Subnets:
          - !Ref PrivateSubnetOne
          - !Ref PrivateSubnetTwo
    TaskDefinition: !Ref AIYoloBBTaskDefinition
```

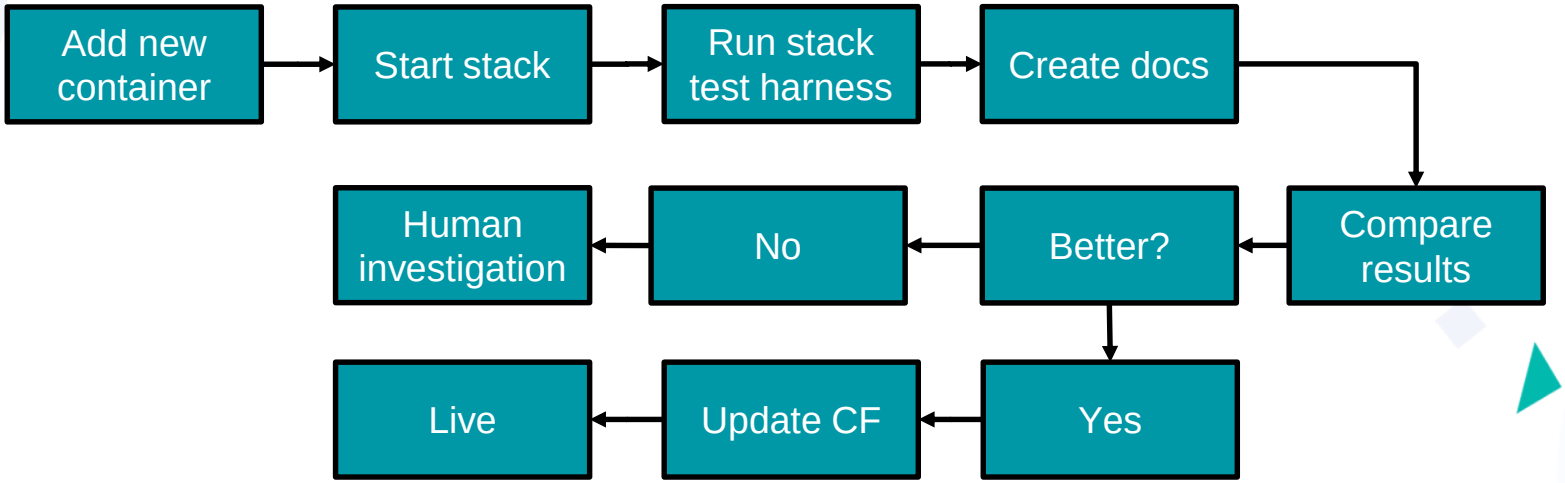
**Dr Janet Bastiman @yssybyl**

# Automation



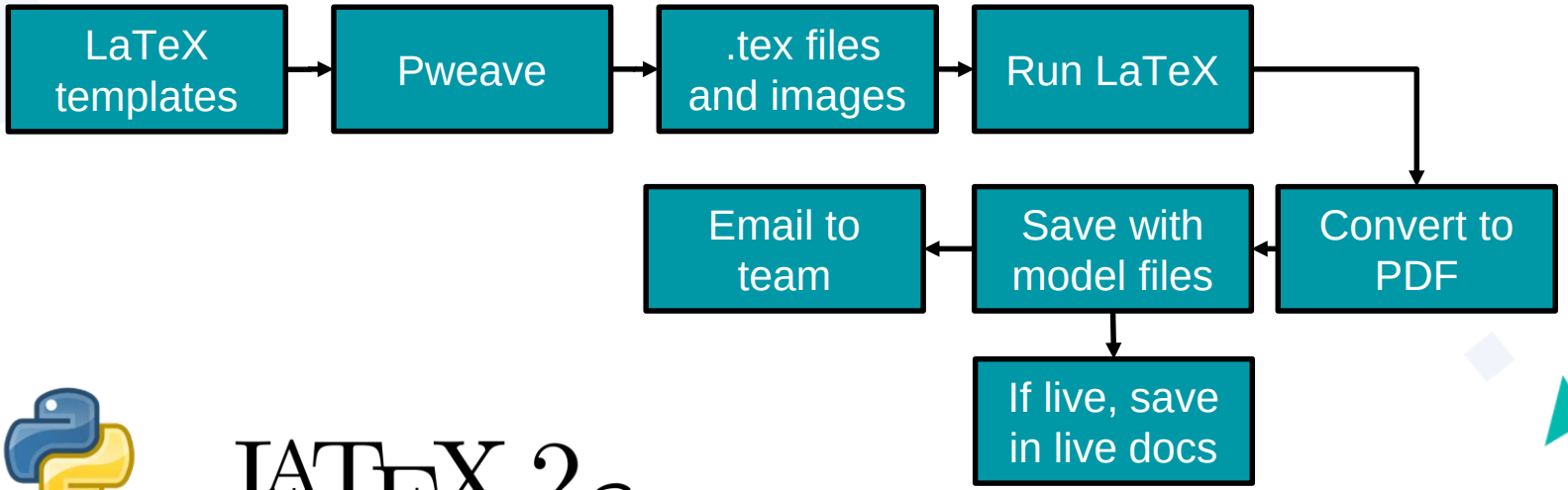

**Dr Janet Bastiman @yssybyl**

# Stack Automation




**Dr Janet Bastiman @yssybyl**

# Automatic Documentation




python

LaTeX 2ε

**Dr Janet Bastiman @yssybyl**

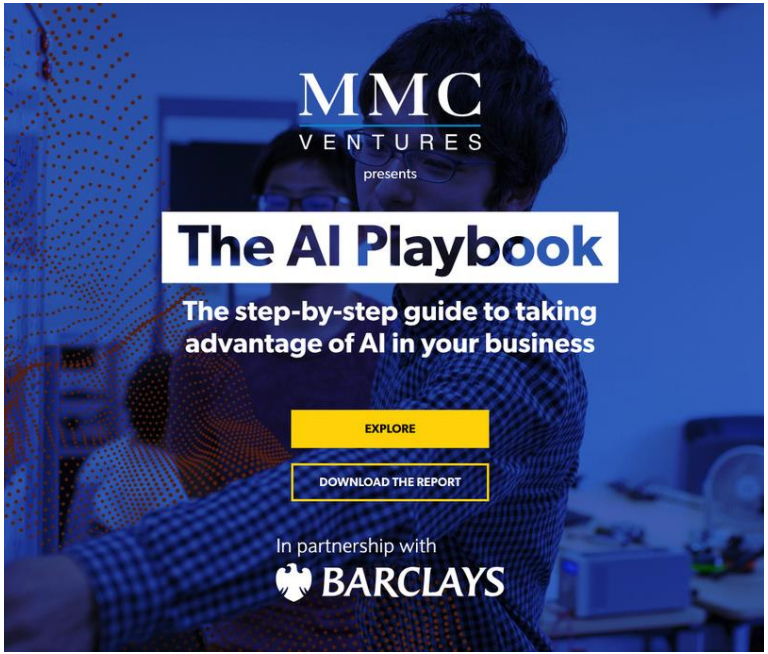
# Did we make it?

- Some really difficult images
- Only expected images were given
- Where it was wrong it was (mostly) sensibly wrong

- Client happy
- Cool automated system


**Dr Janet Bastiman @yssybyl**

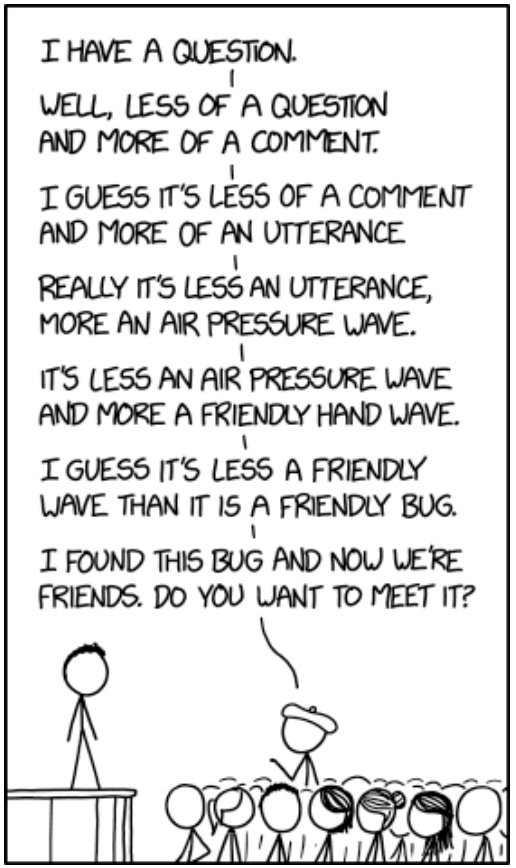
# The Playbook




**ai-playbook.com**

**Dr Janet Bastiman @yssybyl**

# Thank You




https://xkcd.com/2191/

**Dr Janet Bastiman @yssybyl**