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
WHAT CAN I SAY IT'S JUST STATISTICS

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

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

Dr Janet Bastiman @yssybyl
What would it take to get it done in that time?

What would it take to get it done in 3 months?

50 Billion dollars?!?!

The Core (2003)
Paramount Pictures

Dr Janet Bastiman @yssybyl
“They don’t have any data to give us”
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://github.com/awesomedata/awesome-public-datasets)
  - [https://www.kaggle.com/datasets](https://www.kaggle.com/datasets)

- Take your own pictures/videos
  - access/permission?
  - Slow and inconsistent

- Scrape the client site *with permission*

Dr Janet Bastiman @yssybyl
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

Dr Janet Bastiman @yssybyl
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

Dr Janet Bastiman @yssybyl
What do you do with the Data?

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

Dr Janet Bastiman @yssybyl
Crowdsourcing labelling

**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

**Text:**
TO COMPLETE YOUR REGISTRATION, PLEASE TELL US WHETHER OR NOT THIS IMAGE CONTAINS A STOP SIGN:

NO YES

ANSWER QUICKLY—OUR SELF-DRIVING CAR IS ALMOST AT THE INTERSECTION.

SO MUCH OF “AI” IS JUST FIGURING OUT WAYS TO OFFLOAD WORK ONTO RANDOM STRANGERS.

https://xkcd.com/1897/

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

Deep Learning for Vision Systems, Mohamed Elgendy

Dr Janet Bastiman @yssybyl
Unbalanced Data

Class distribution in Training Data

Dr Janet Bastiman @yssybyl
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

- Find the right architecture

Dr Janet Bastiman @yssybyl
Go old school

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

Dr Janet Bastiman @yssybyl
Simplify the problem

Removal of camera artefacts in eye images to make detection easier - Jeffrey De Fauw

Removal of Doppler effect on moving source using fractional octave band shifting, F Mobley

\[ \Delta n = \tau \log_2 \left( 1 - M \cos \theta \sin \phi \right) \]

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
- Change hue saturation and value of colours in the image

Dr Janet Bastiman @yssybyl
Infrastructure

- Taxonomy
- Classifier Definition
- Test Set
- Dashboard
- Data In
- Data Store
- AWS Image
- Scripts
- DockerHub Setup
- Codeship Project
- GitHub Setup
- Notification
- Slack
- Email
- Template

Dr Janet Bastiman @yssybyl
Cloud Formation

Dr Janet Bastiman @yssybyl
Stack Automation

Add new container → Start stack → Run stack test harness → Create docs → Compare results

Human investigation → No → Better?

Live → Update CF → Yes

Dr Janet Bastiman @yssybyl
Automatic Documentation

LaTeX templates → Pweave → .tex files and images → Run LaTeX → Email to team → Save with model files → Convert to PDF → If live, save in live docs

\LaTeX\ 2\varepsilon

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