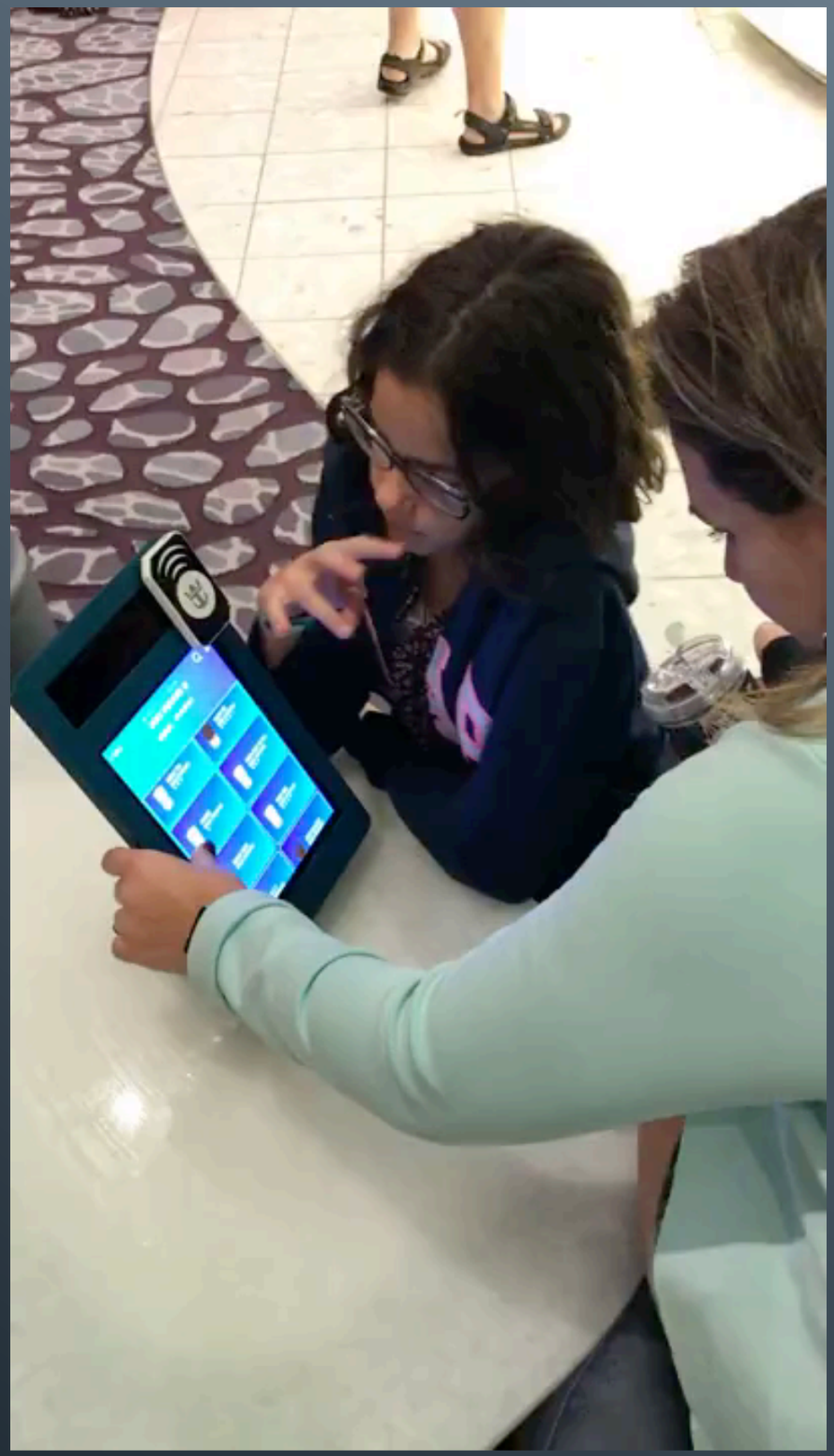
Deep Learning for detection on a phone:

how to stay sane
and build a pipeline you can trust

Irina Vidal Migallón
Viorama

Viorama: spil.ly

Irina Vidal Migallón



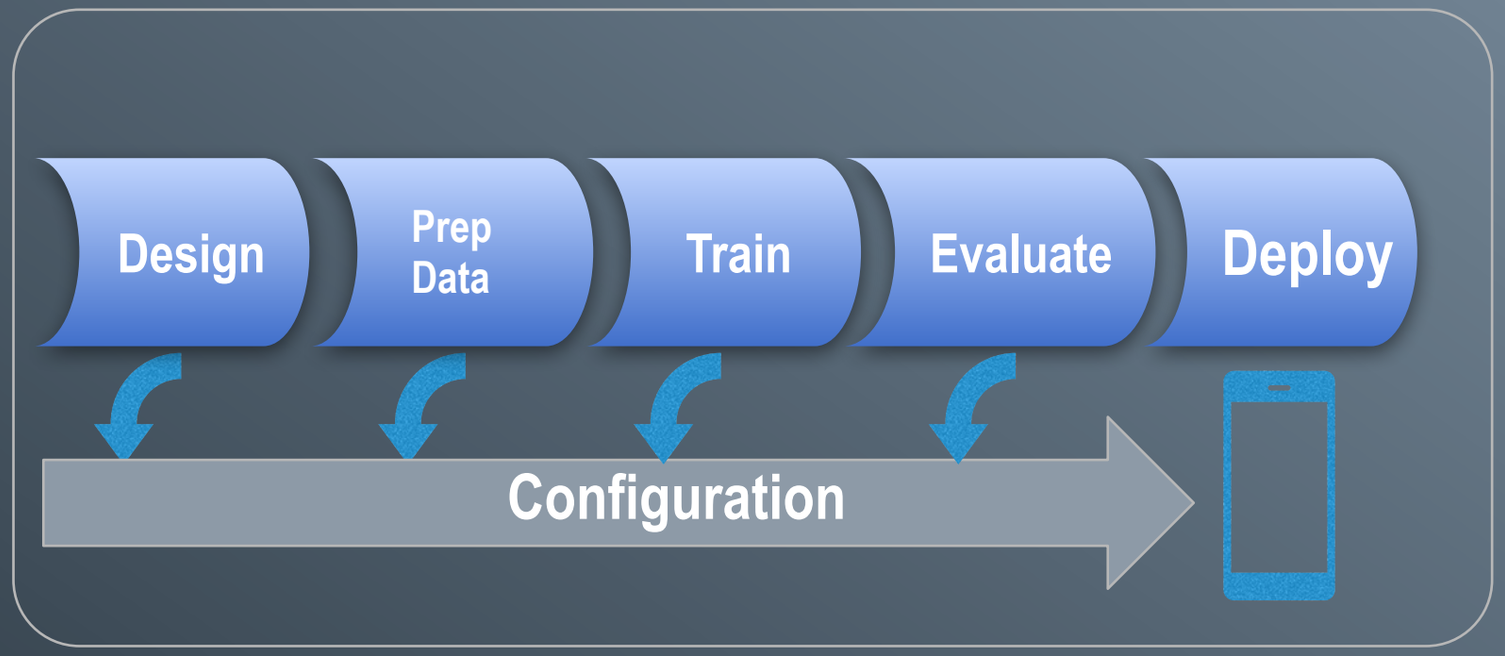
Viorama: spil.ly

Irina Vidal Migallón

"Work to deploy a Neural Network model on a phone starts **long before** you write the first line of native code."

– Sun Tzu (500 BC, apocryphal)

Which architecture?
What weights?
What deployment parameters?



We'll see:

- Pipeline phases
- How to avoid pitfalls

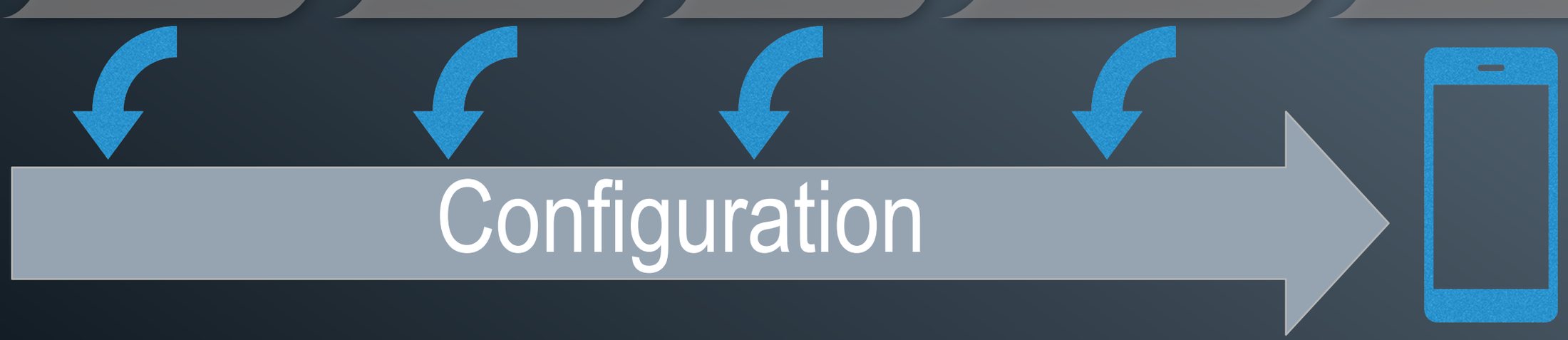
Design

Prep Data

Train

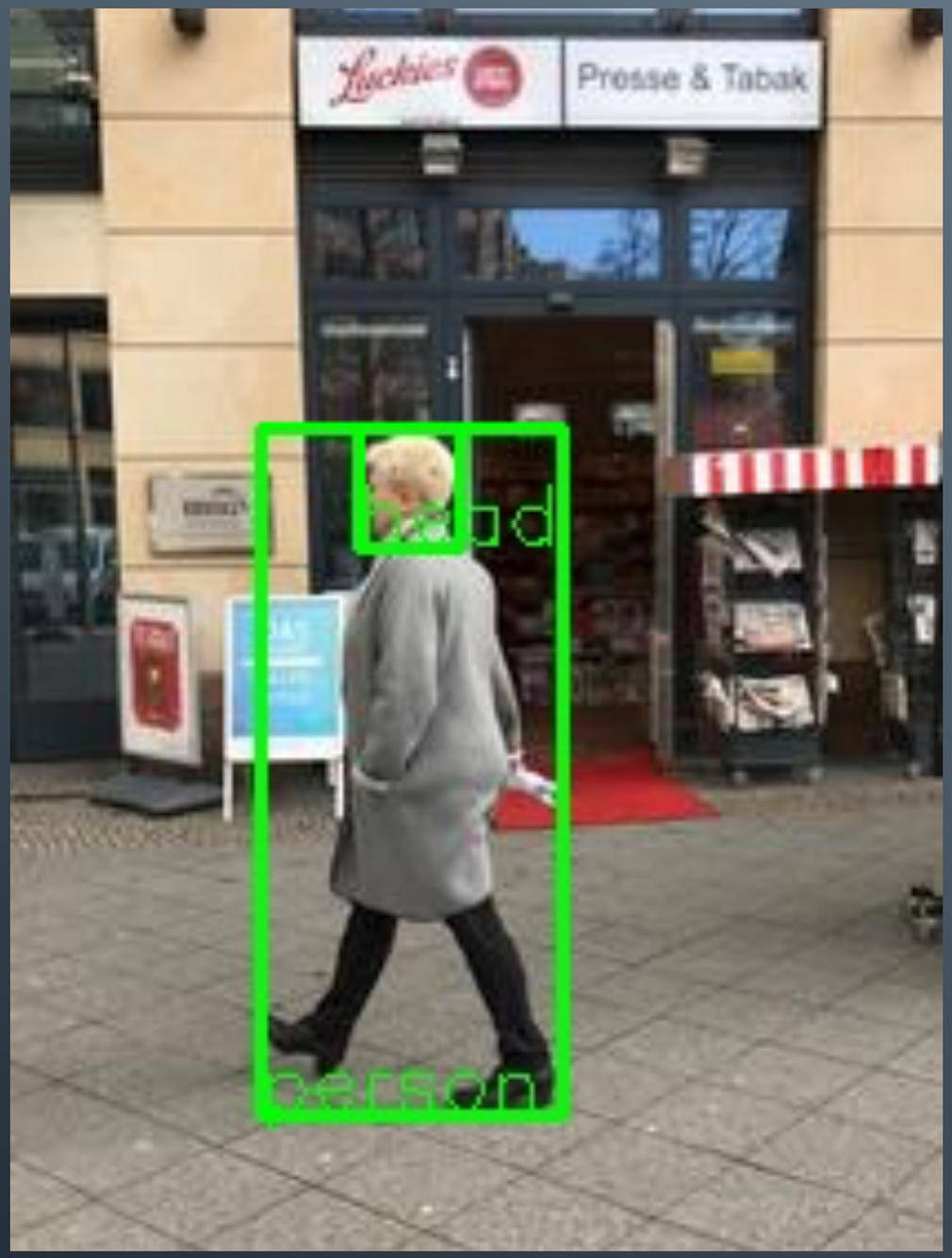
Evaluate

Deploy



Classifier

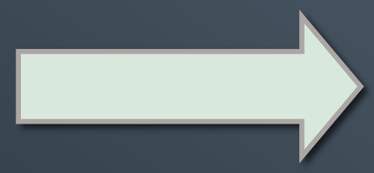
Detector



Detector

Classifier

Label proposal

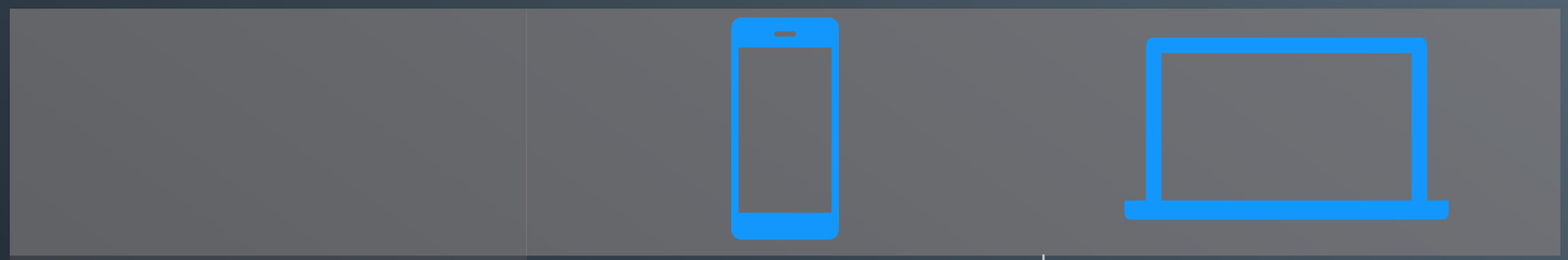


Localization

Region proposal

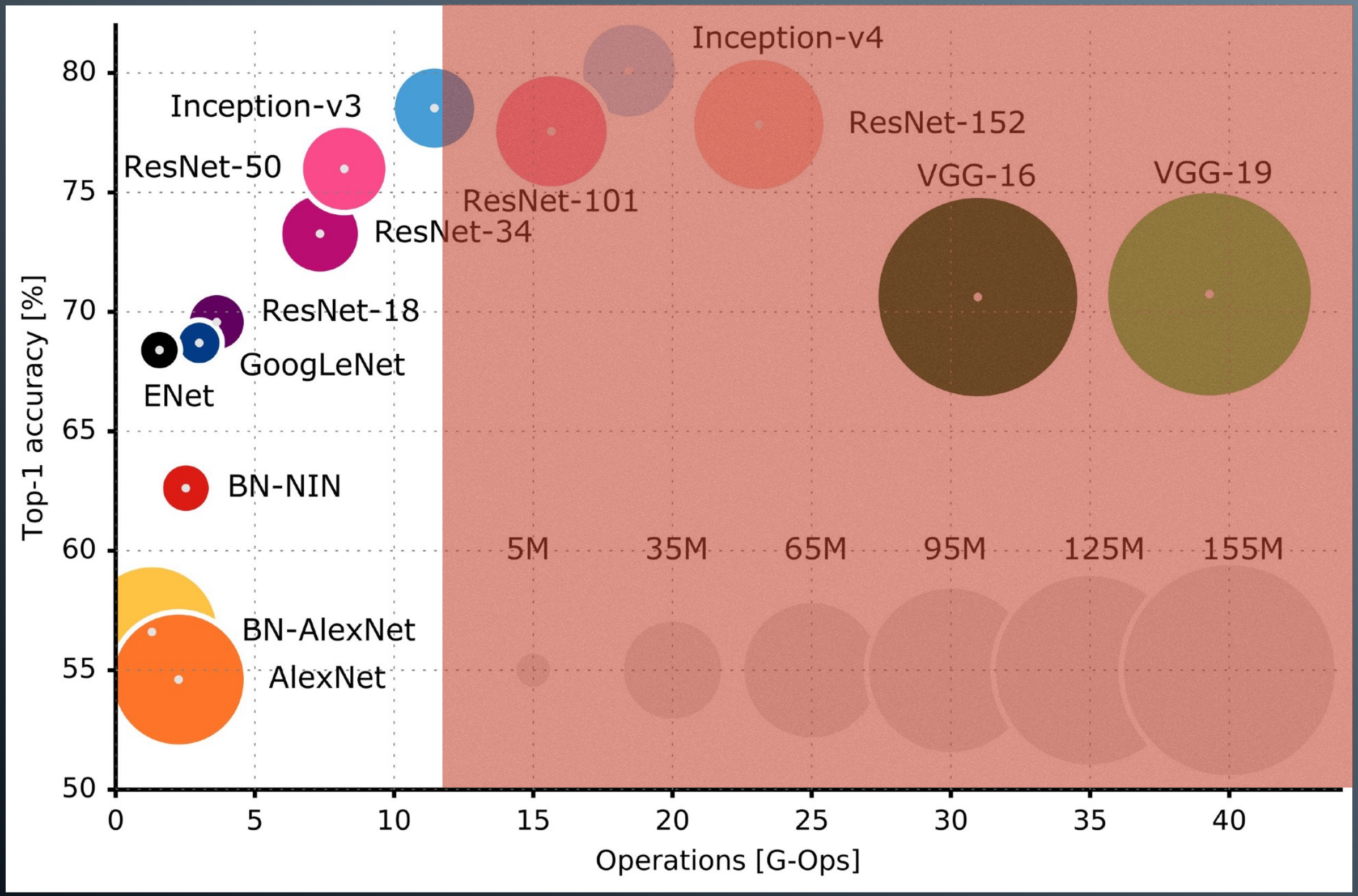
Constraints

Design



RAM	1 GB	32 GB
GFLOPS (GPU)	150	4000
Clock (GPU)	600 MHz	1500 MHz

40-60 FPS



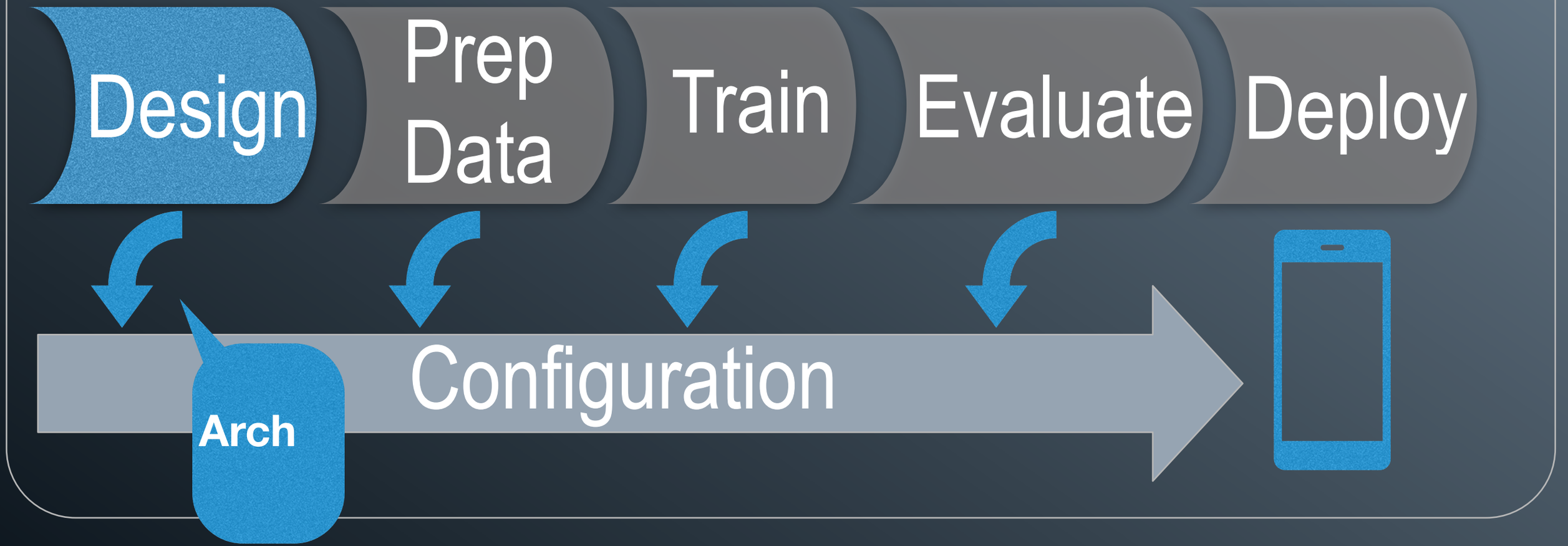
(A.Canziani, A. Paszke, E. Culurciello, 2016)

Slimmer classifier

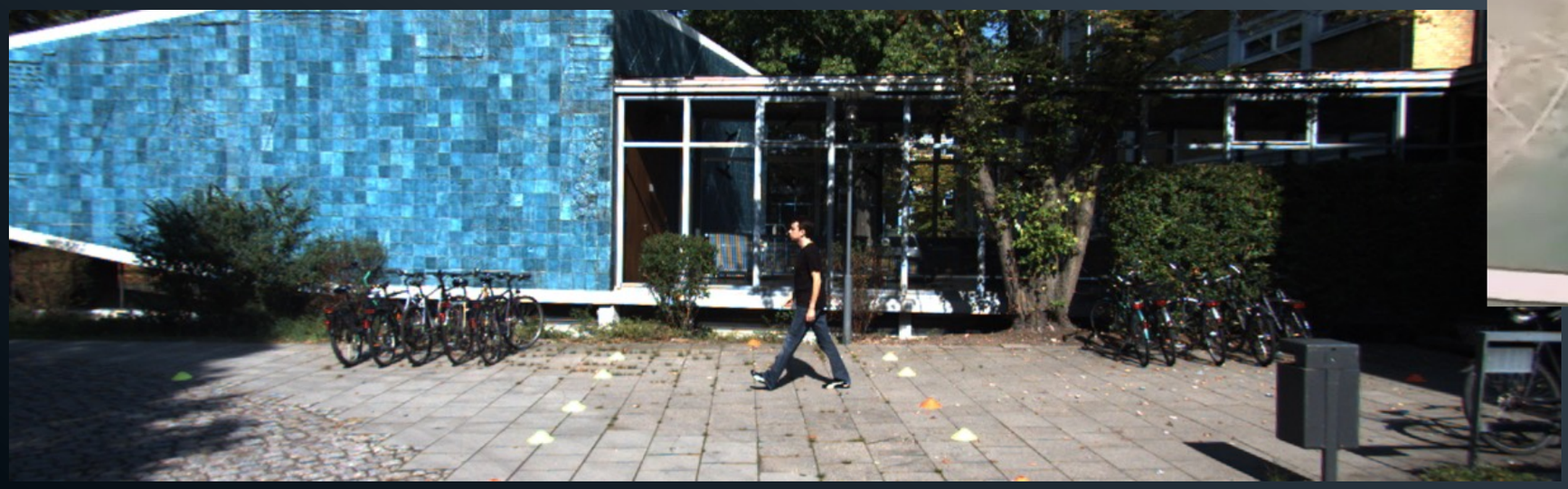
- Memory
  - Convolutional vs Fully Connected
- Inference time
  - Input & layer sizes

Slimmer detector

- What determines the size of your detector?
- Keep only what you need


Design
Prep Data
Train
Evaluate
Deploy
Configuration
Arch

Data



Viorama: spil.ly

Irina Vidal Migallón

- ✓ Intended use & format
- ✓ Annotation format
- ✓ Consistency

Free*

- MS COCO
- PASCAL VOC
- KITTI

*** Pay with your time instead.**

Data

Explore your data

Analyse your data:

...

Stats

Track:

- Subsets
- Augmentations

Design

Prep Data

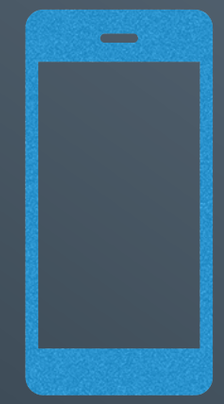
Train

Evaluate

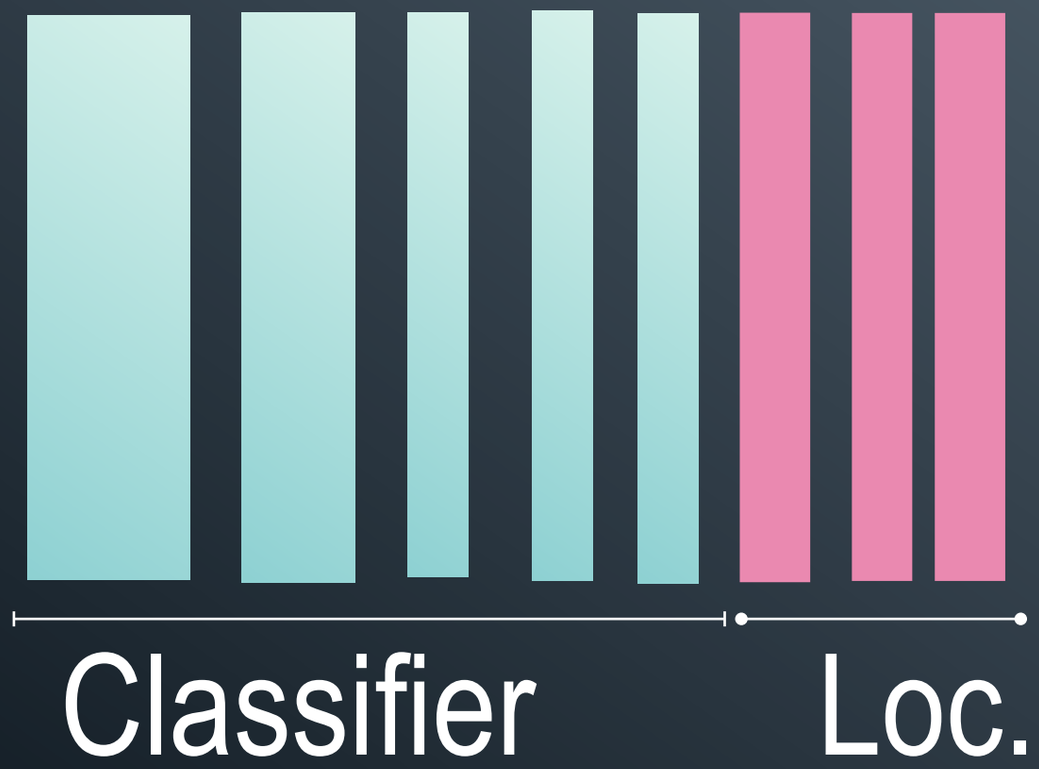
Deploy

Co ation

Arch
Stats



Training

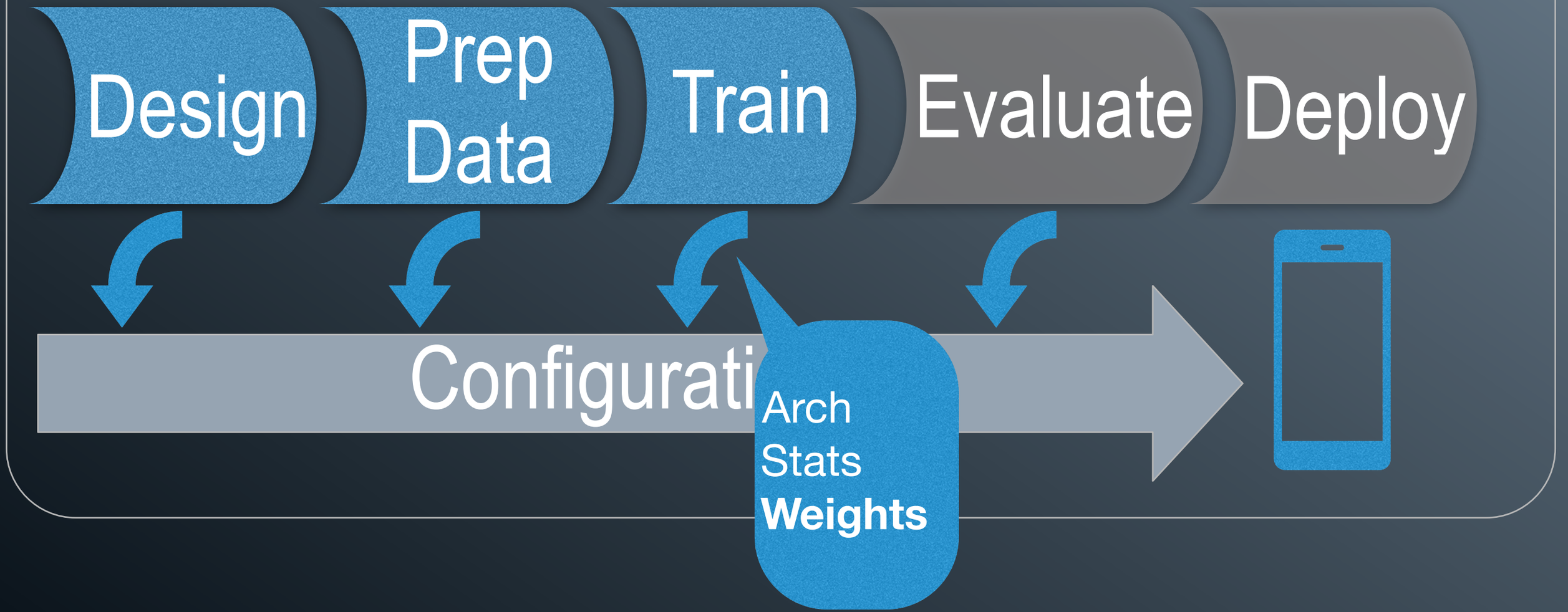


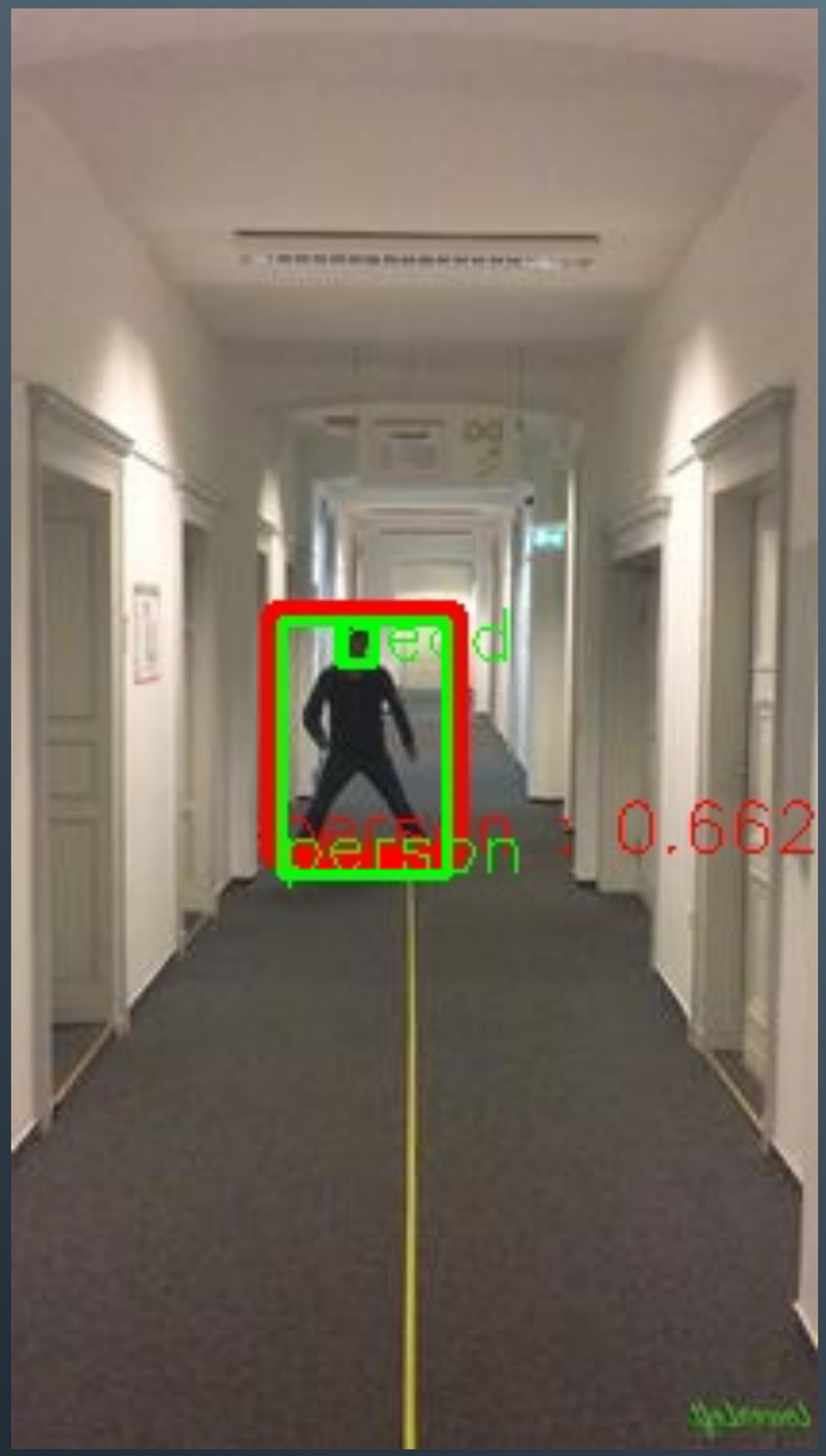
- Pre-trained models
- Transfer learning

Training

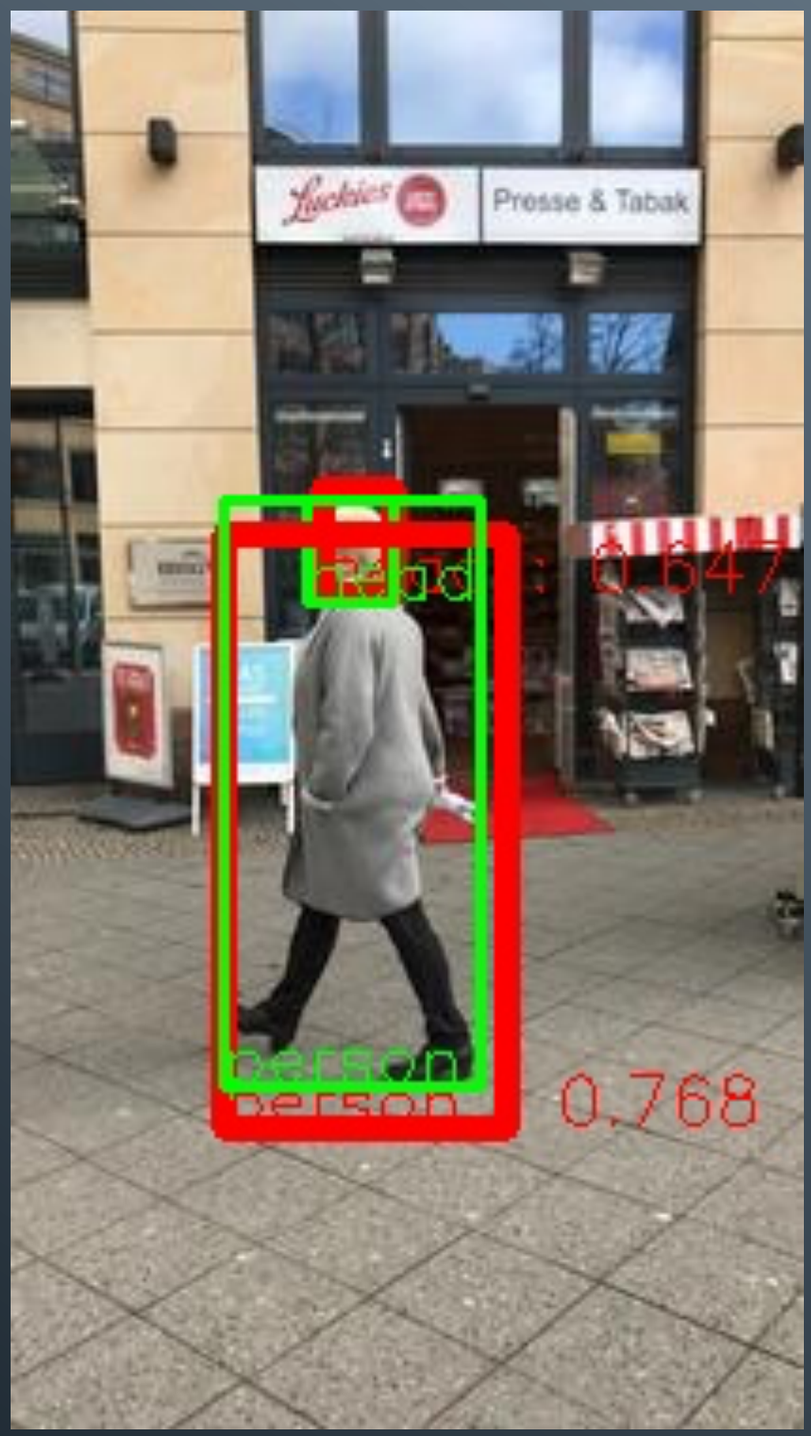
- Track, track, track

- *Any* tool - just do it.
  - 💡you can feed CSVs to any other tool!


Design
Prep Data
Train
Evaluate
Deploy
Configurati
Arch
Stats
Weights

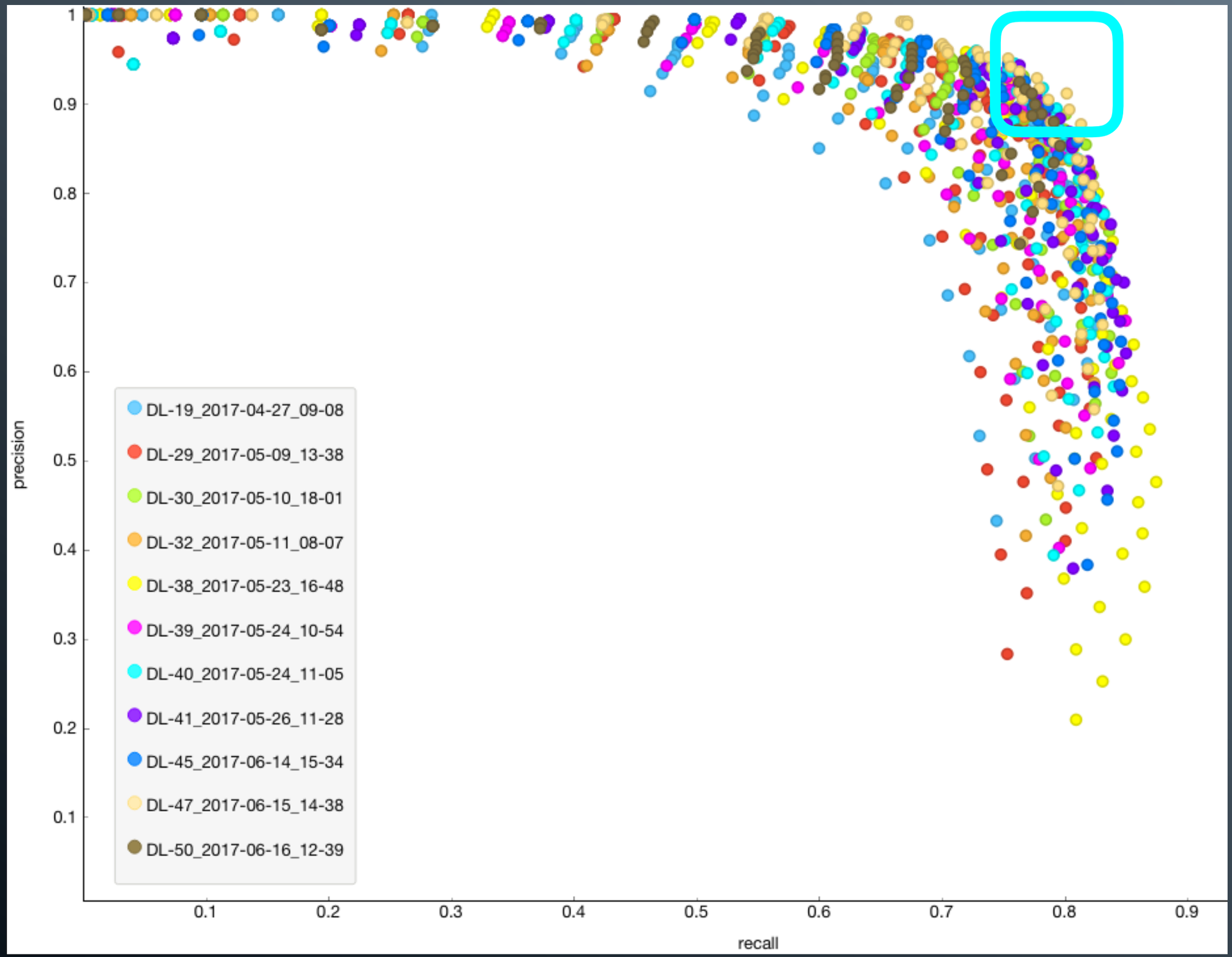


Detection
Ground truth



Detection
Ground truth

Evaluate



Viorama: spil.ly

Irina Vidal Migallón

Evaluate

- Architecture
- Datasets used
- Augmentations
- Hyperparameters
- Deployment parameters

Design

Prep Data

Train

Evaluate

Deploy

Configuration

Arch
Stats
Weights
Params

Design

Prep Data

Train

Evaluate

Deploy

Configuration

Arch
Stats
Weights
Params

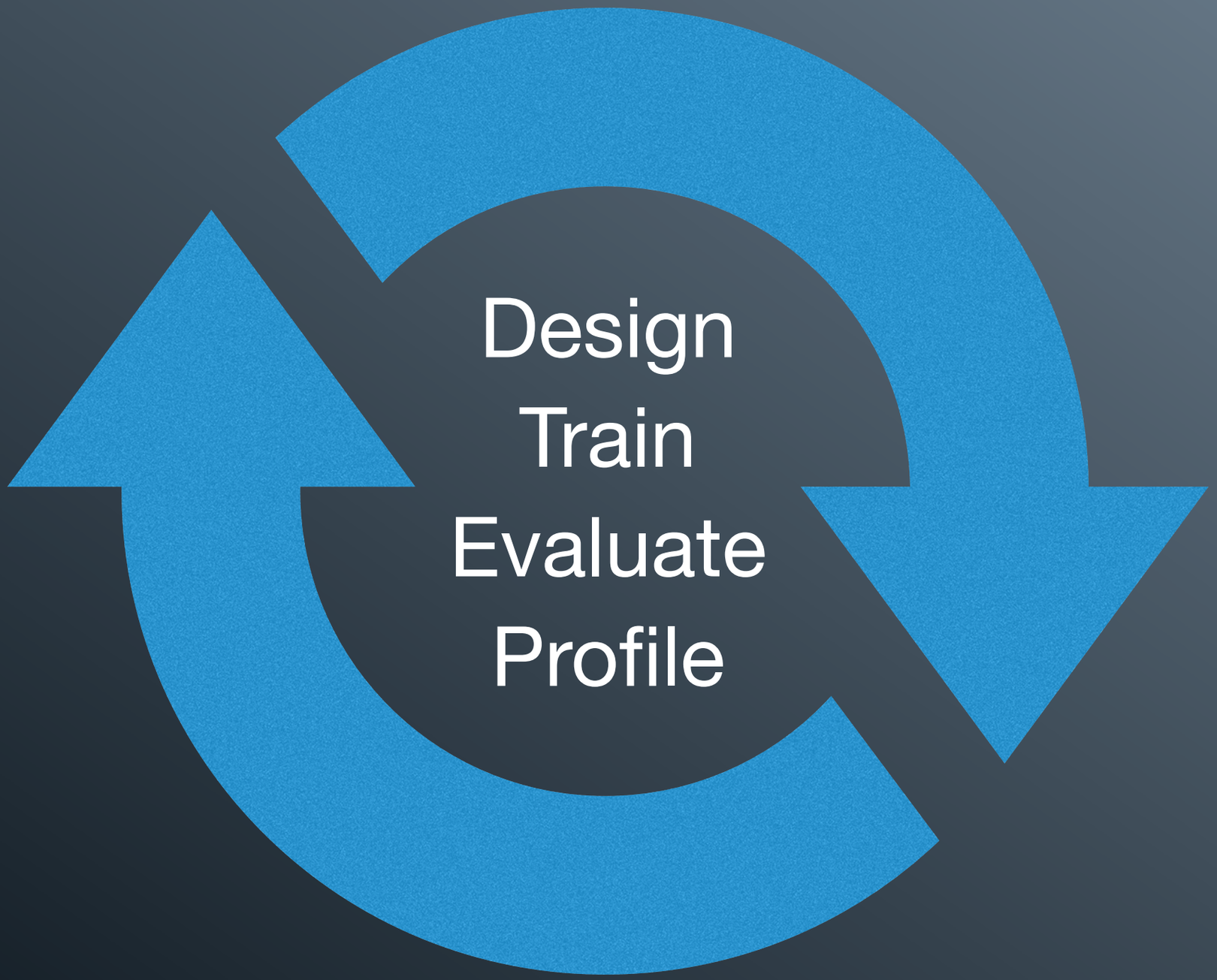
Deploy

TensorFlow

Caffe2

- Frameworks announced left & right

- Check for ops available


Design
Train
Evaluate
Profile

Avoiding Pitfalls

Facepalm Inception by zantaff.deviantart.com

Avoiding Pitfalls

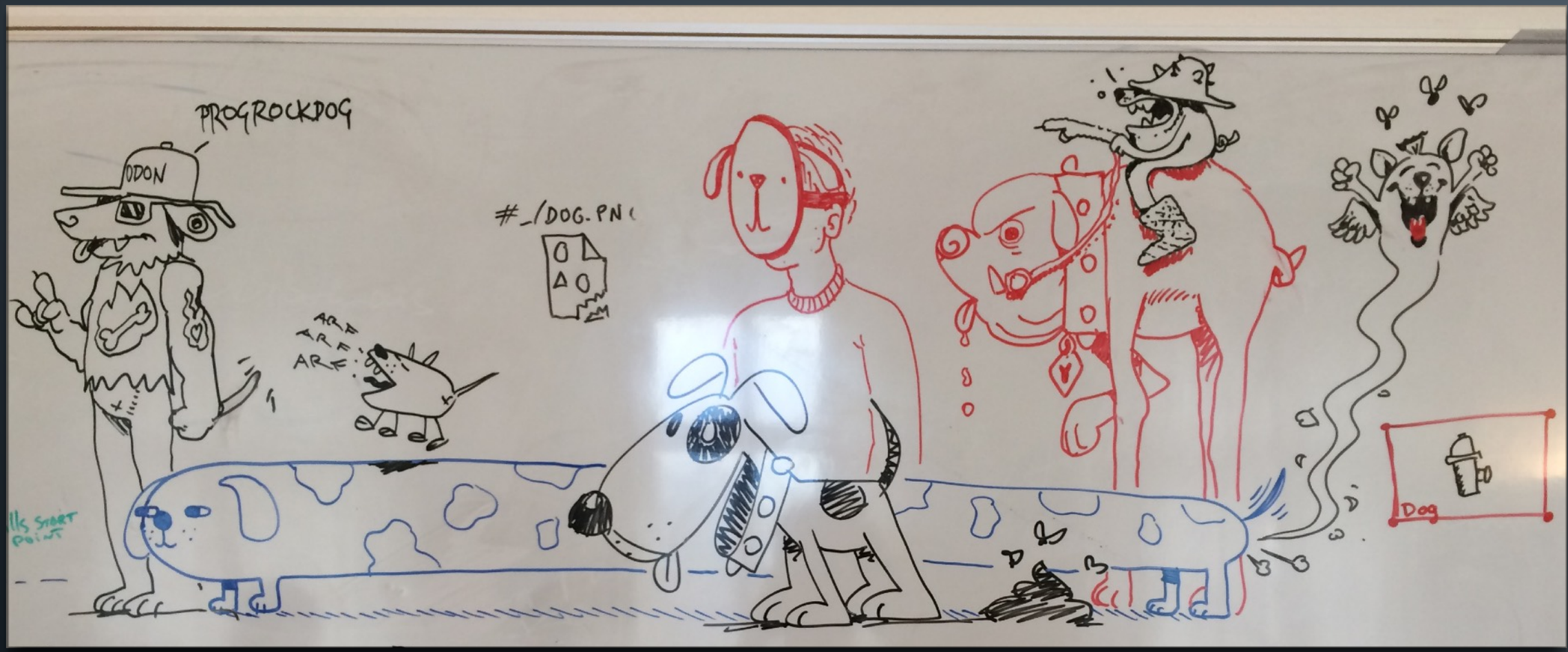
- ✓ Datasets:
  - ✓ Consistent annotations
    - ✓ Define detection
    - ✓ Define box
    - ✓ Define person

Avoiding Pitfalls

- ✓ Evaluation: visual and metric
- ✓ Automatic configuration
- ✓ Tracking experiments
- ✓ Early profiling on device
- ➡ Feedback loop to design

Summary

- Tips for models targeting mobile devices
- Configuration-driven pipeline
- Avoiding pitfalls!



Thank you!
Any questions?