January 16, 2025

Daniel Bogdoll

Mcity Data Engine

From Data Stream to AI Model

About Me

2011 - 2019 • M.Sc.
Computational Engineering Science

2016 • Research Intern
Autonomous Driving

2018 - 2021 • CEO & Co-Founder
Carpooling Startup

2020 - 2025 • PhD Candidate
Autonomous Driving

2024 - 2025 • Research Scholar
Data Engine

About Mcity


2021 • Expansion
Smart Intersections
2015 • Foundation
32-acre “mini-city”

About Mcity

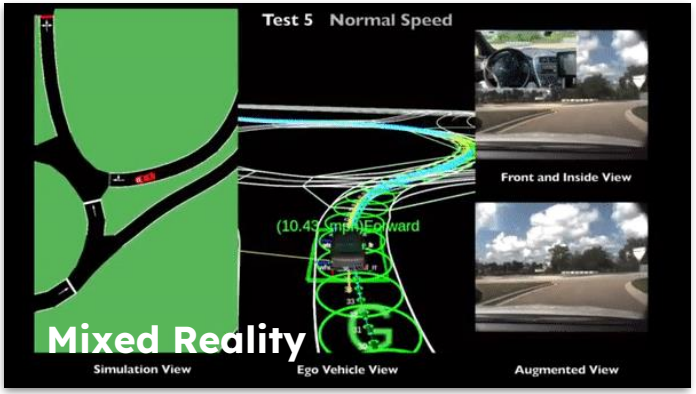
2015 • Foundation
32-acre “mini-city”

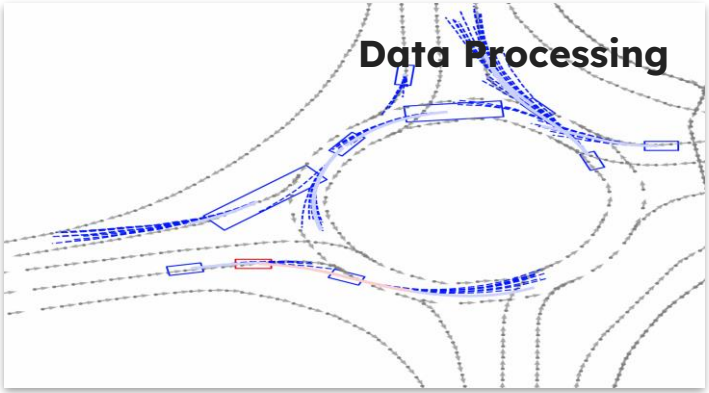
2021 • Expansion
Smart Intersections

2024 • Digital Twin
Mcity 2.0

About Mcity


Real World Data


Test 5 Normal Speed
Front and Inside View
Mixed Reality
Simulation View
Ego Vehicle View
Augmented View

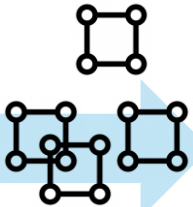

Data Processing


Simulation

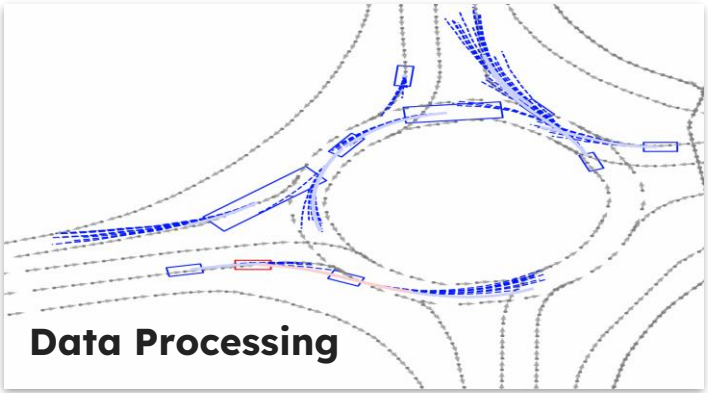
From Data to Model

Labeled Data

Neural Network

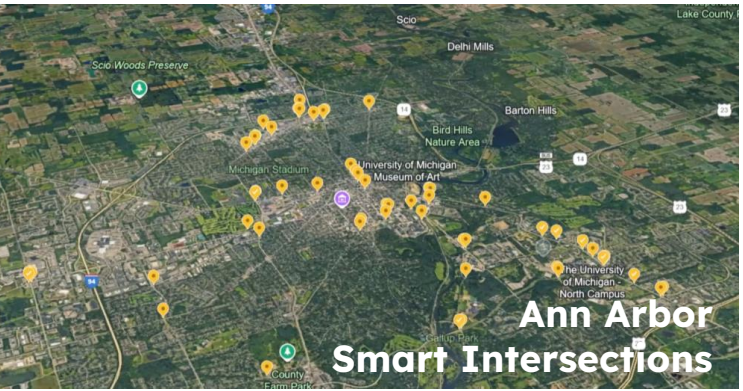


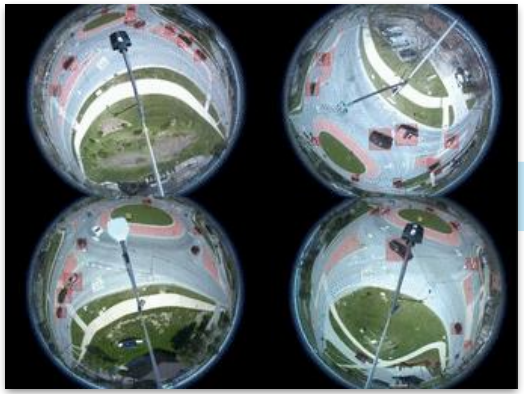
Abstract Data

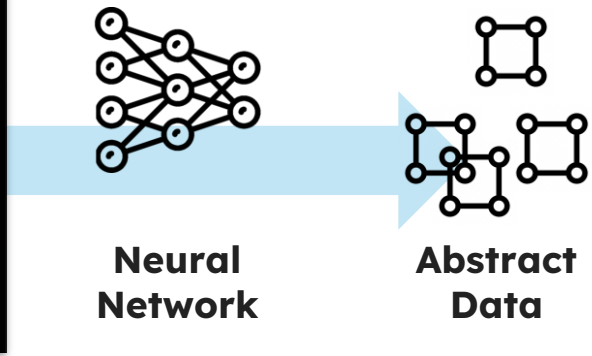


From Data to Model

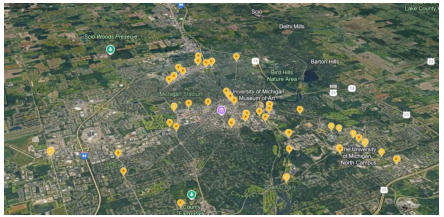
UNIVERSITY OF MICHIGAN







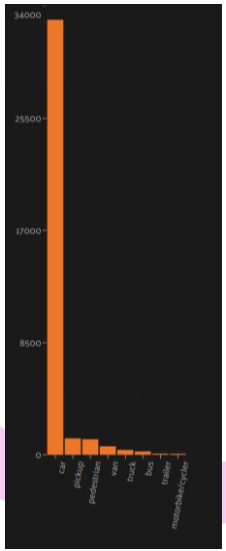
From Data to Model

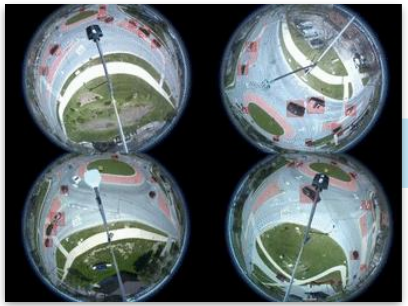


Raw Data

Random + Handpicked

Labeled Data



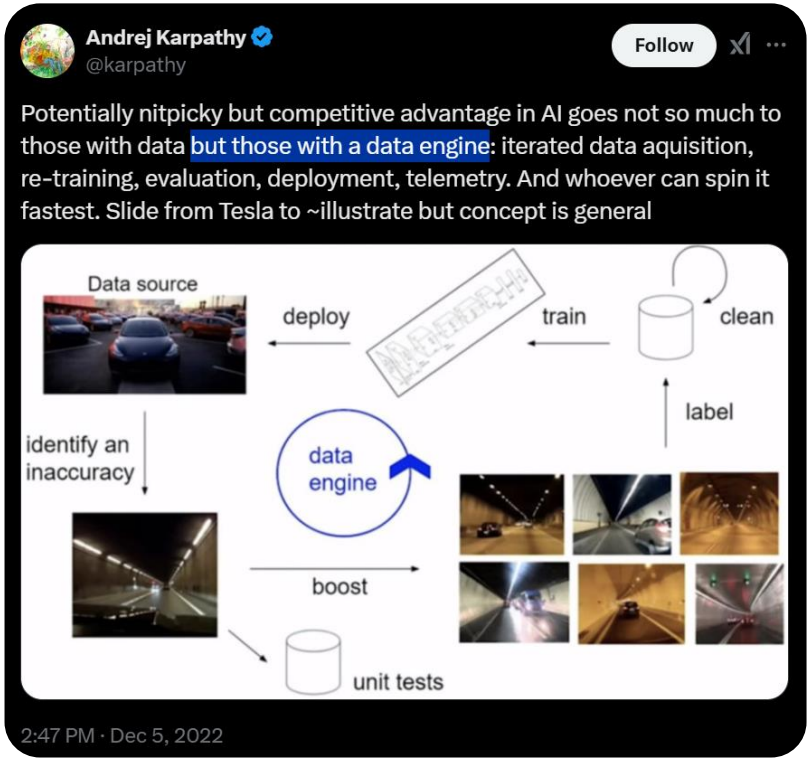


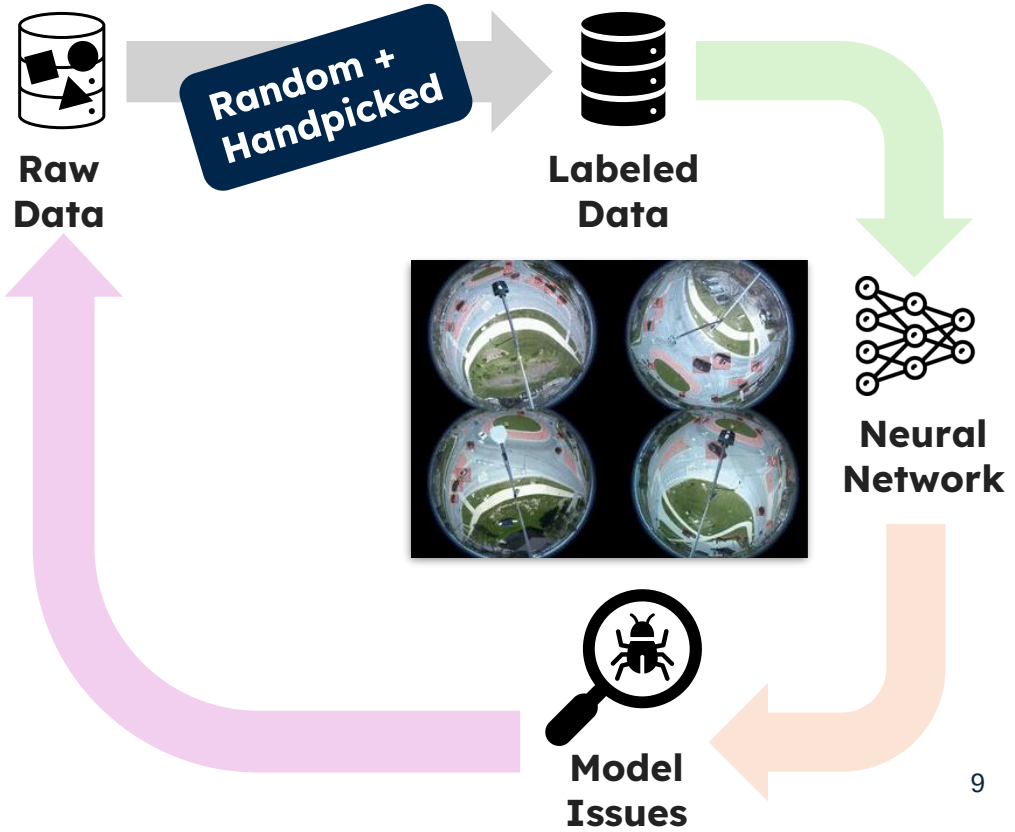
Neural Network

Abstract Data

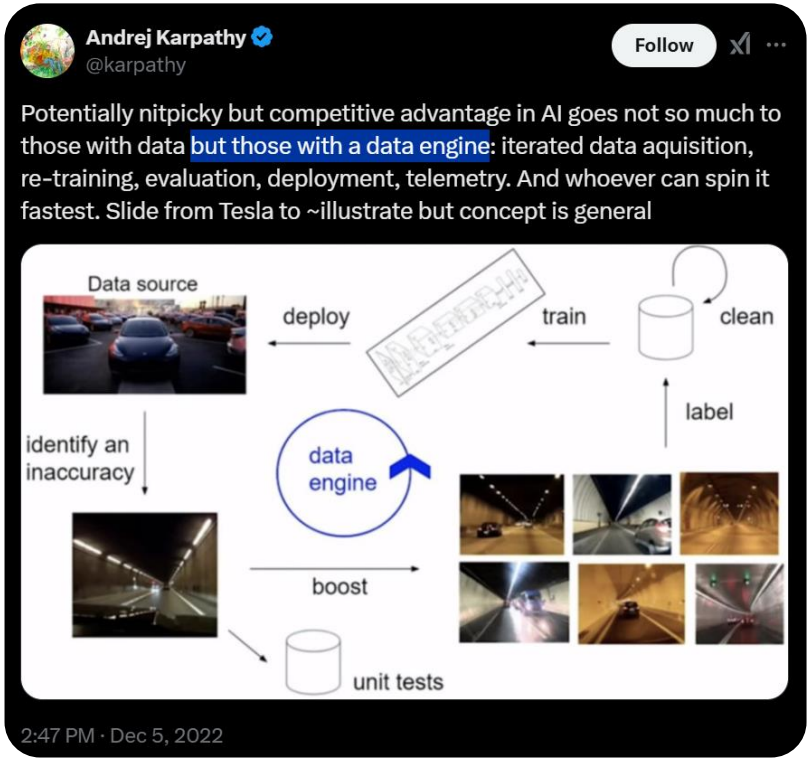
Model Issues

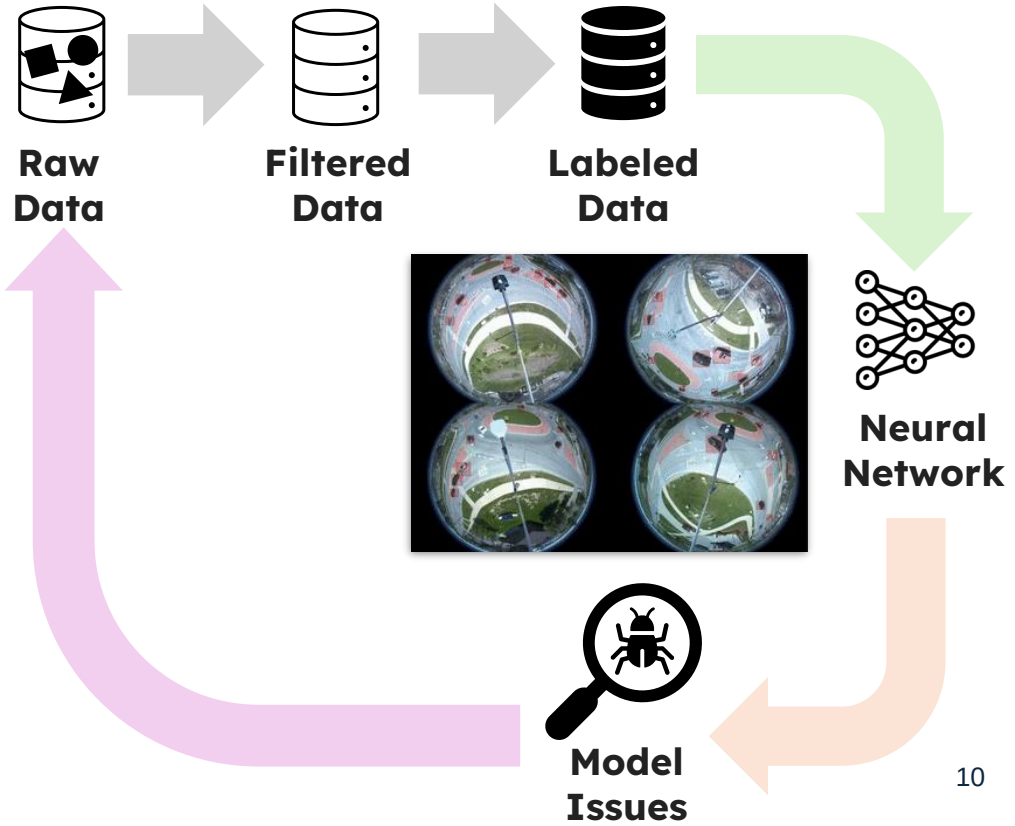
Data Engine


Andrej Karpathy
@karpathy
Follow
Potentially nitpicky but competitive advantage in AI goes not so much to those with data but those with a data engine: iterated data aquisition, re-training, evaluation, deployment, telemetry. And whoever can spin it fastest. Slide from Tesla to ~illustrate but concept is general
Data source
deploy
train
clean
label
identify an inaccuracy
data engine
boost
unit tests
2:47 PM · Dec 5, 2022

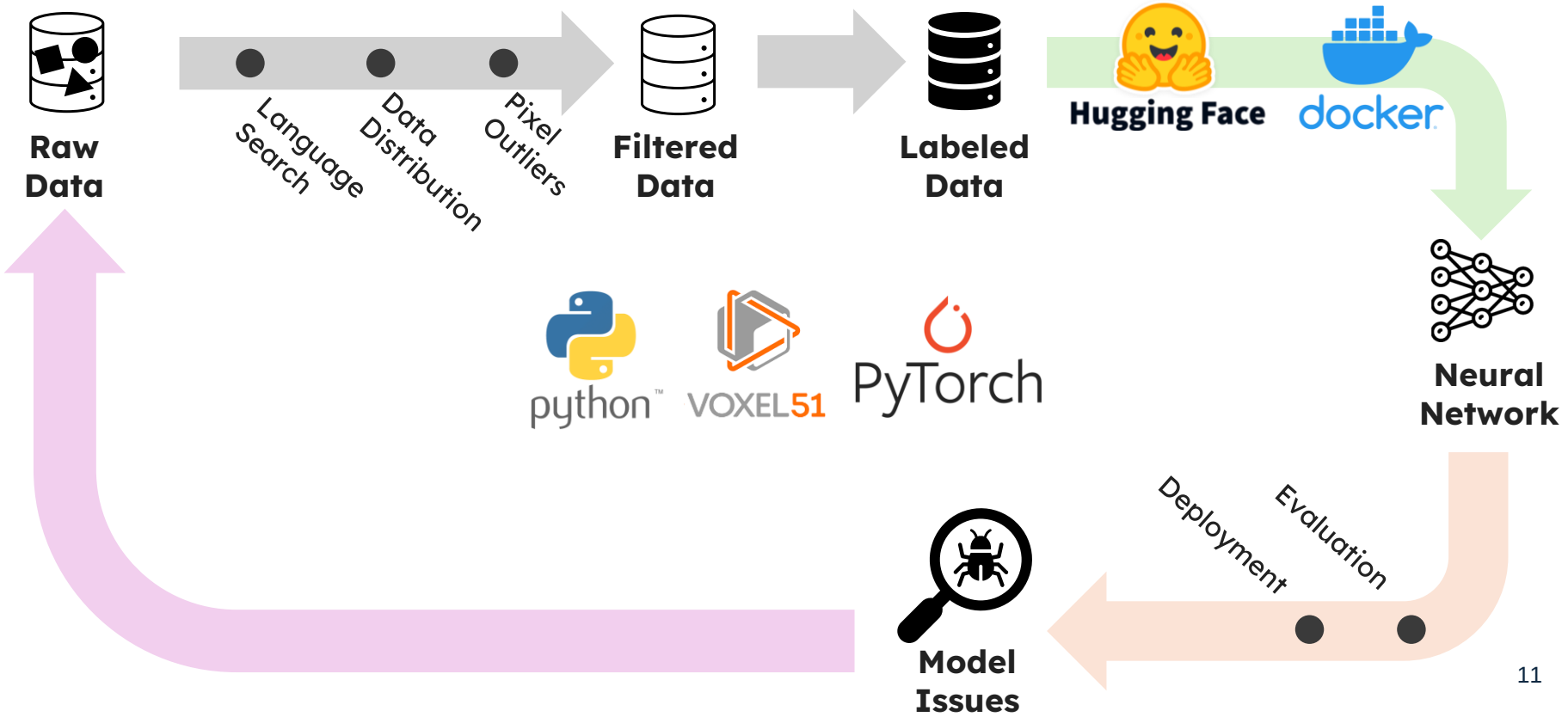

Raw Data
Random + Handpicked
Labeled Data
Neural Network
Model Issues

Data Engine

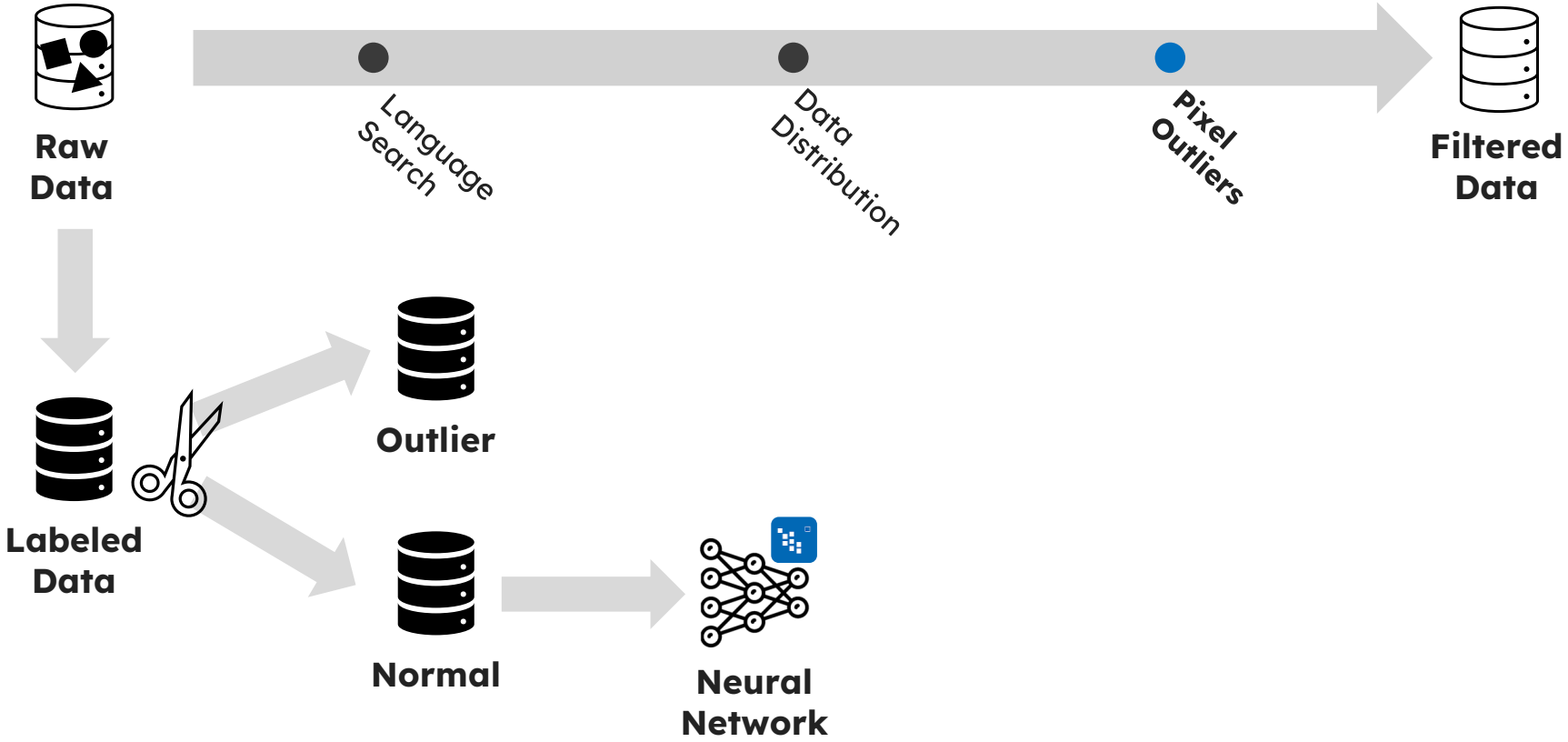

Andrej Karpathy
@karpathy
Follow
Potentially nitpicky but competitive advantage in AI goes not so much to those with data but those with a data engine: iterated data aquisition, re-training, evaluation, deployment, telemetry. And whoever can spin it fastest. Slide from Tesla to ~illustrate but concept is general
Data source
deploy
train
clean
label
identify an inaccuracy
data engine
boost
unit tests
2:47 PM · Dec 5, 2022


Raw Data
Filtered Data
Labeled Data
Neural Network
Model Issues

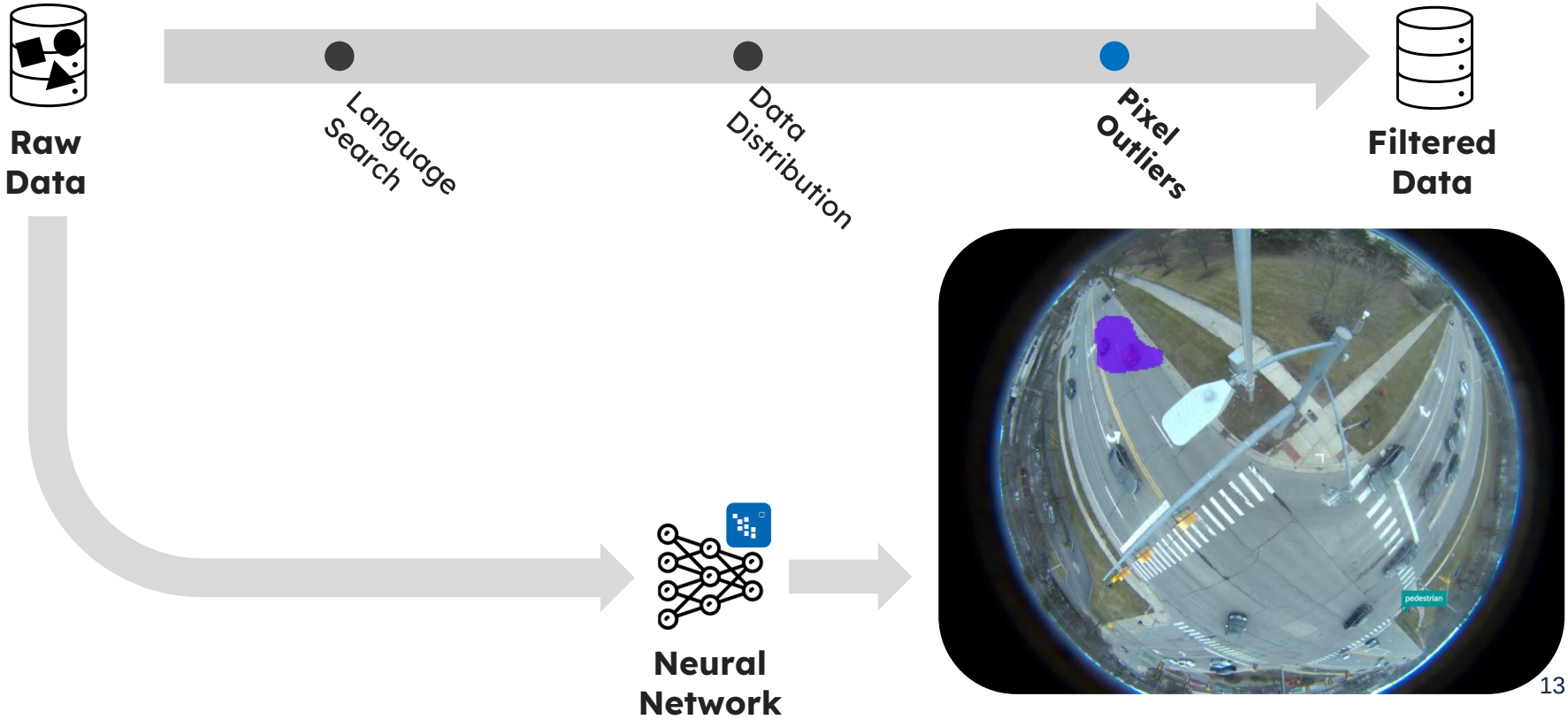
Mcity Data Engine



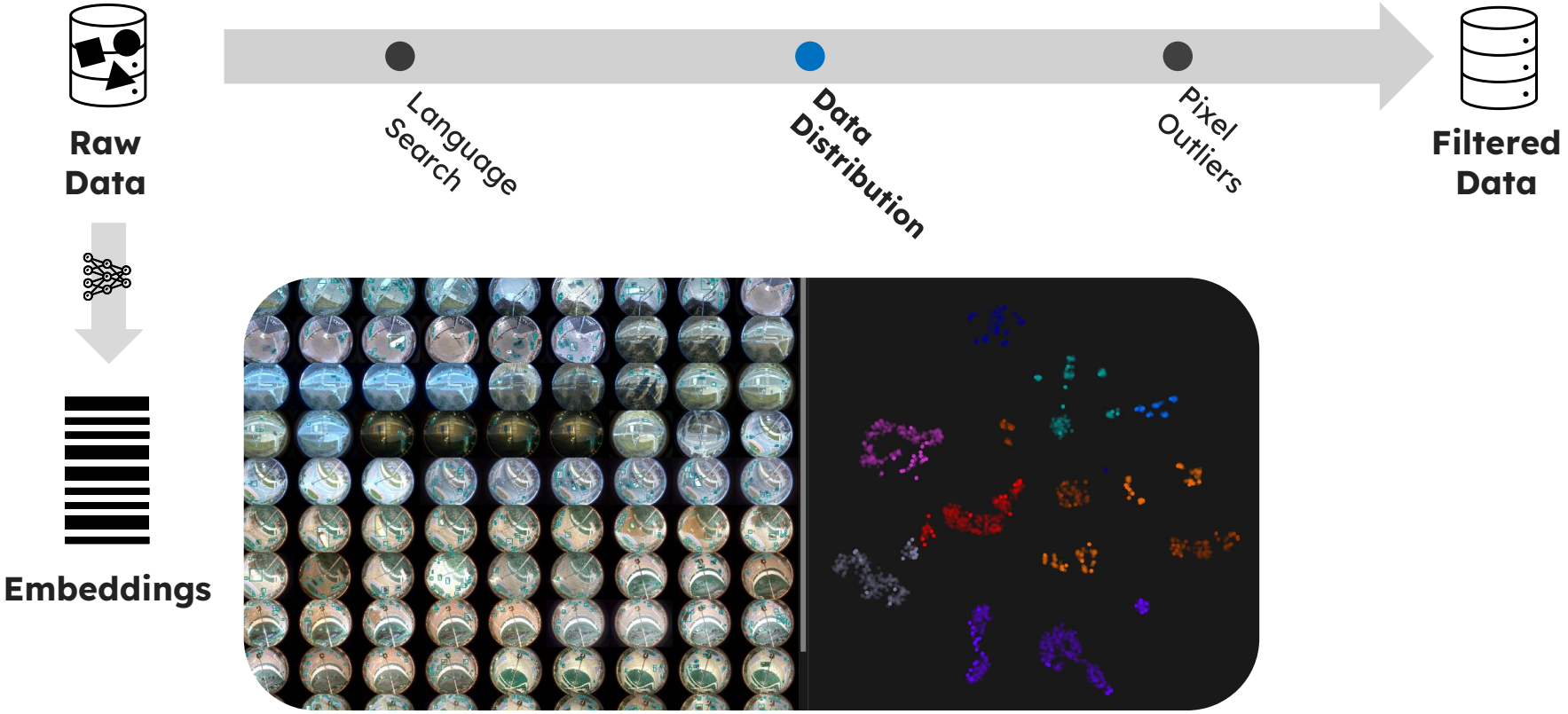
Outlier Detection



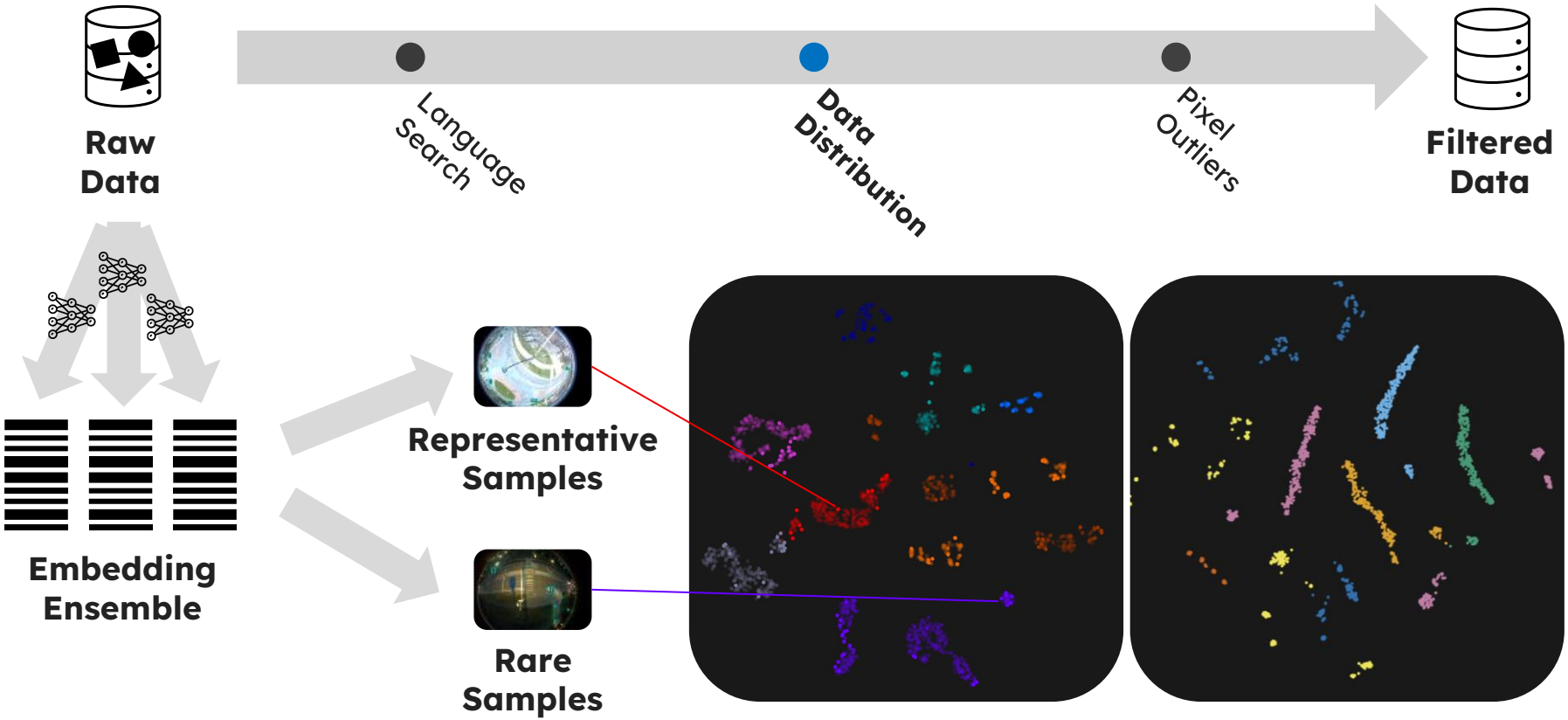
Outlier Detection



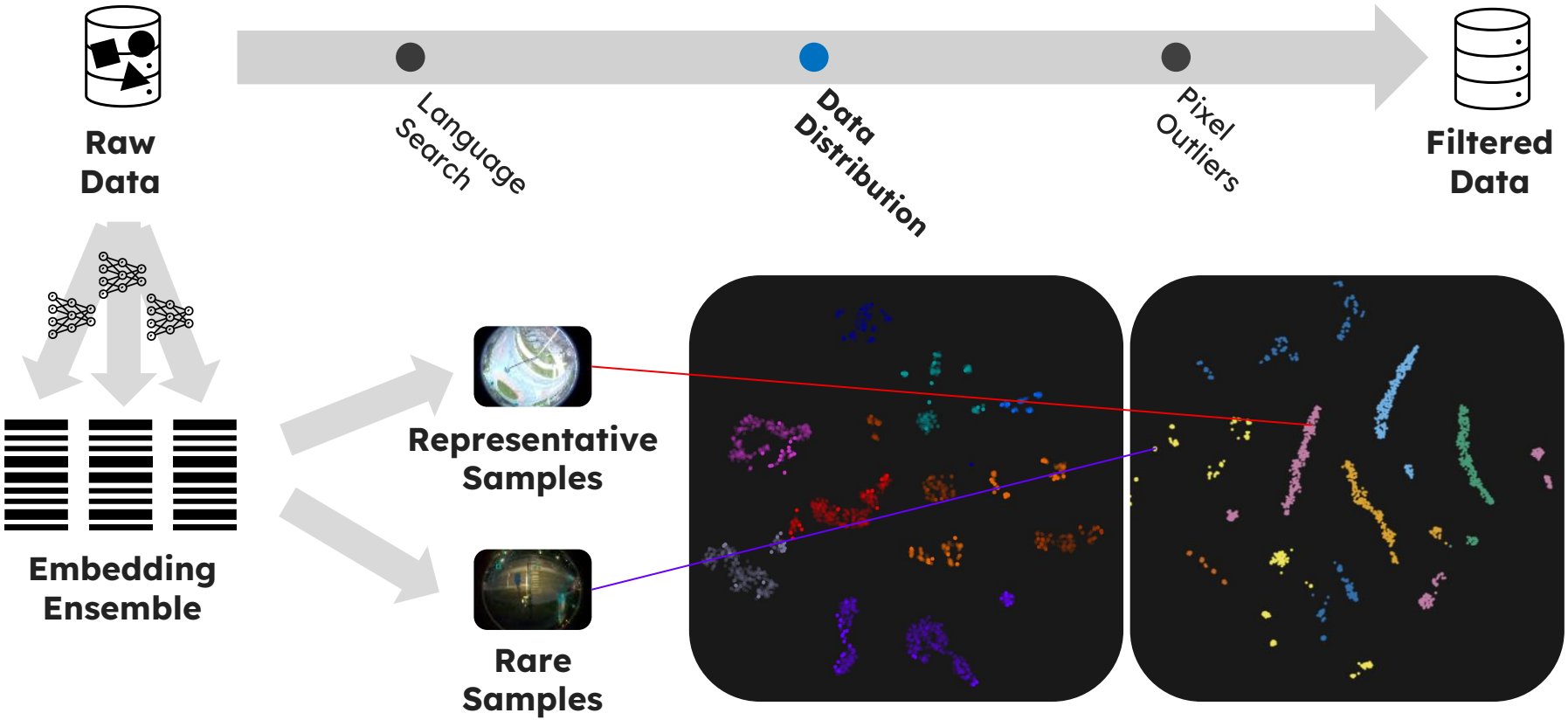
Data Distribution



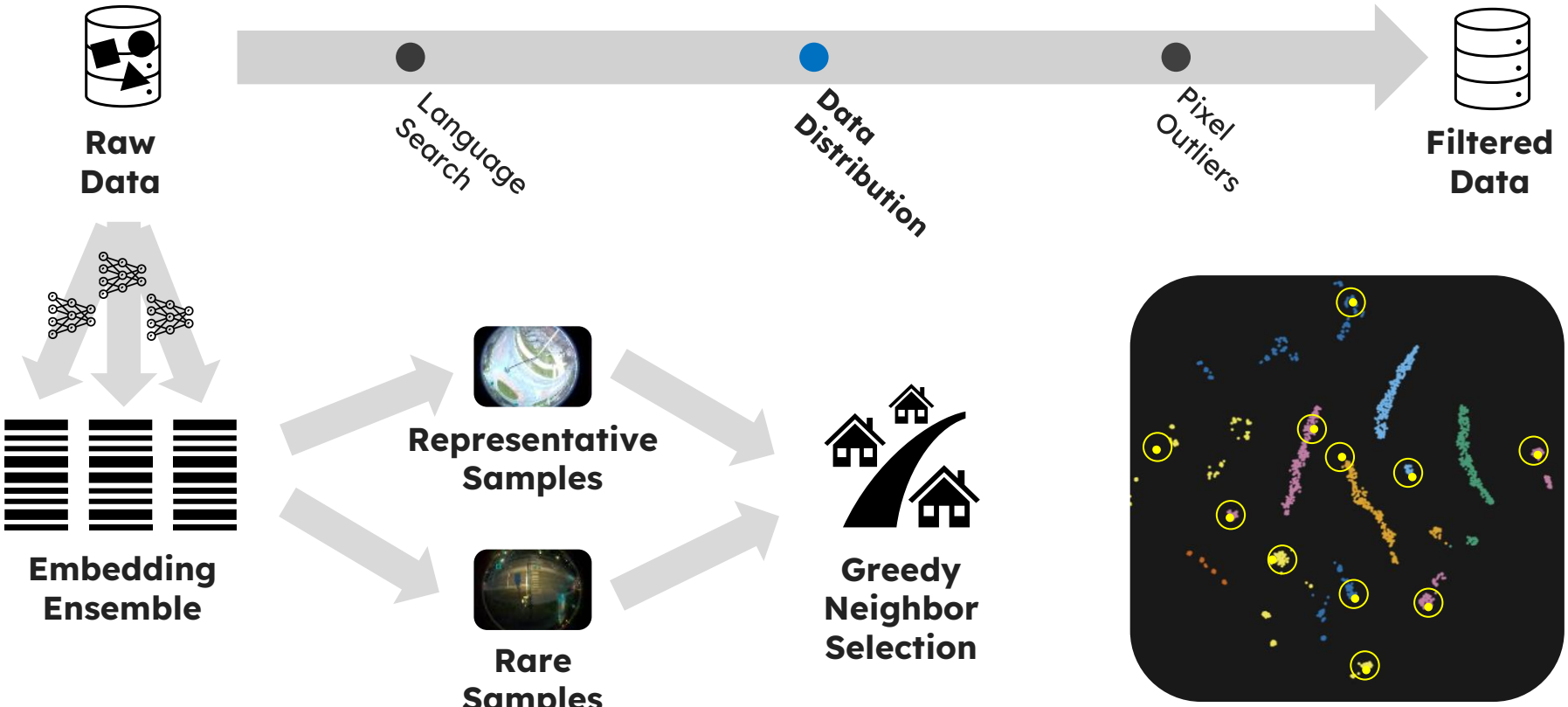
Data Distribution



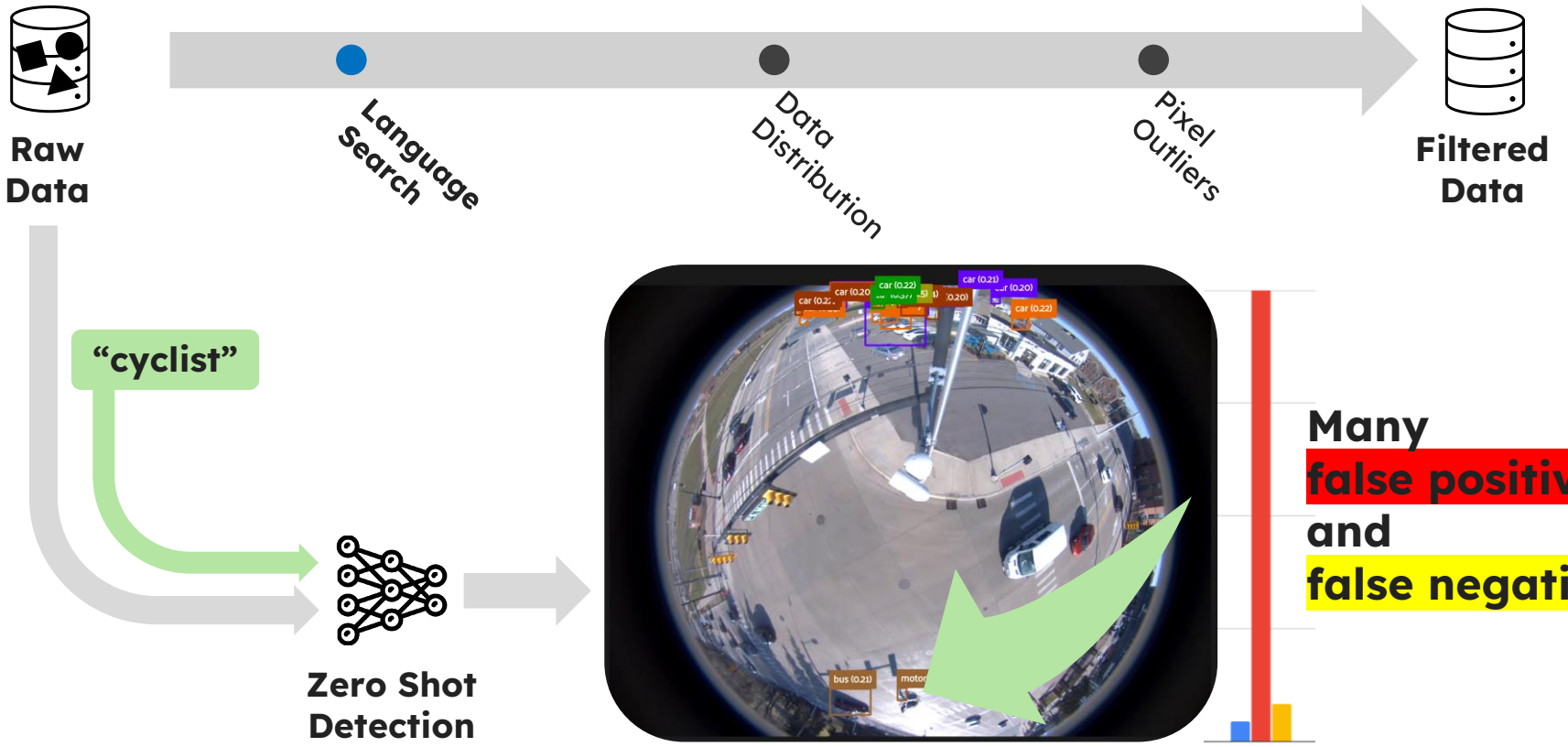
Data Distribution



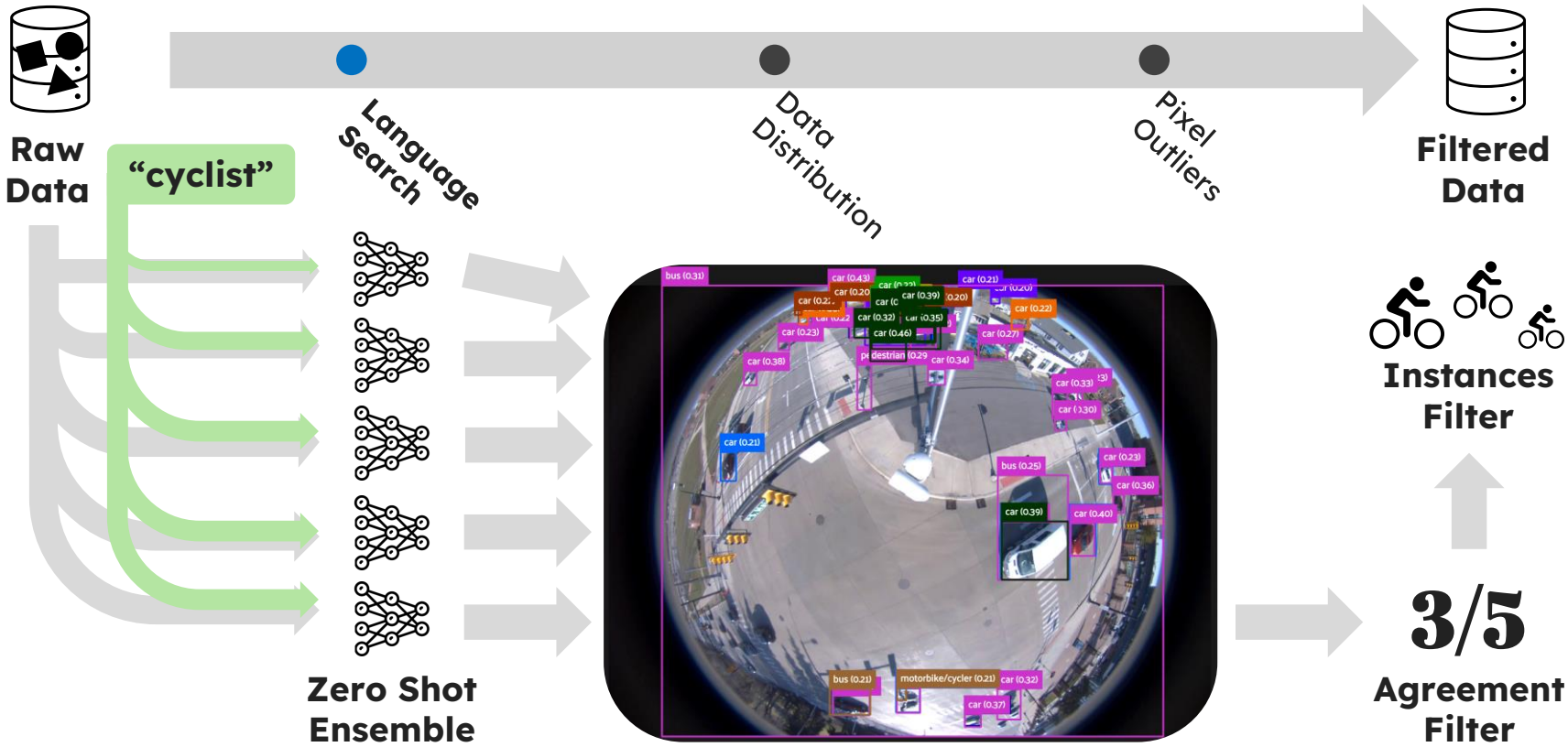
Data Distribution



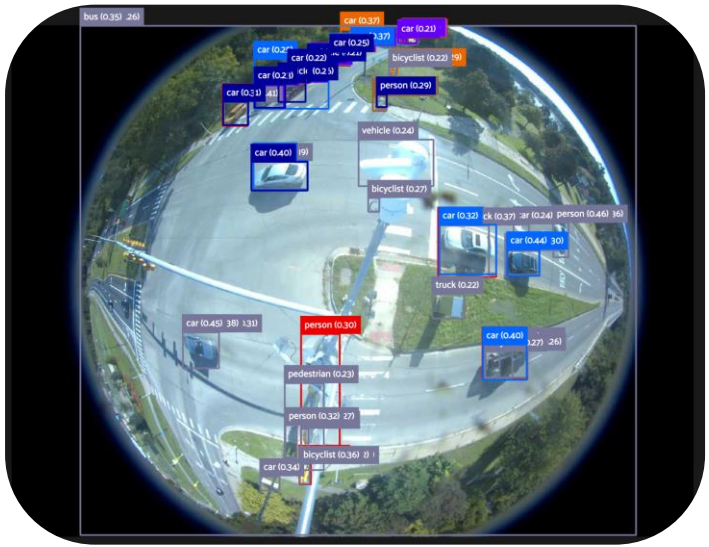
Language Search

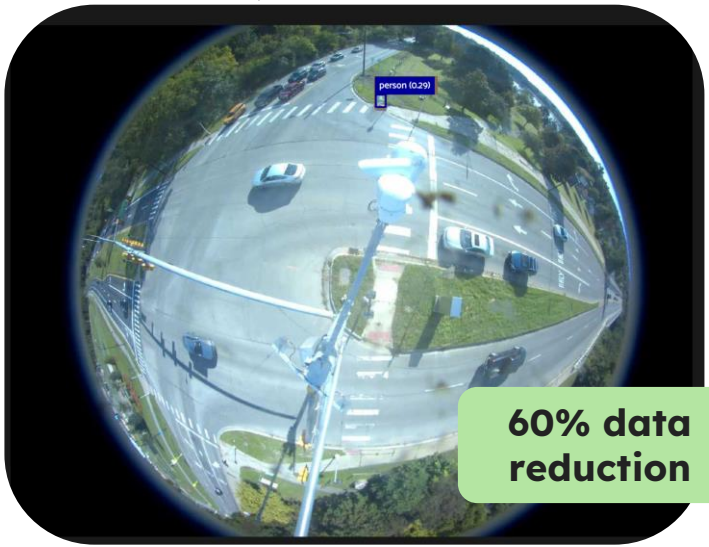


Language Search



Language Search





60% data reduction

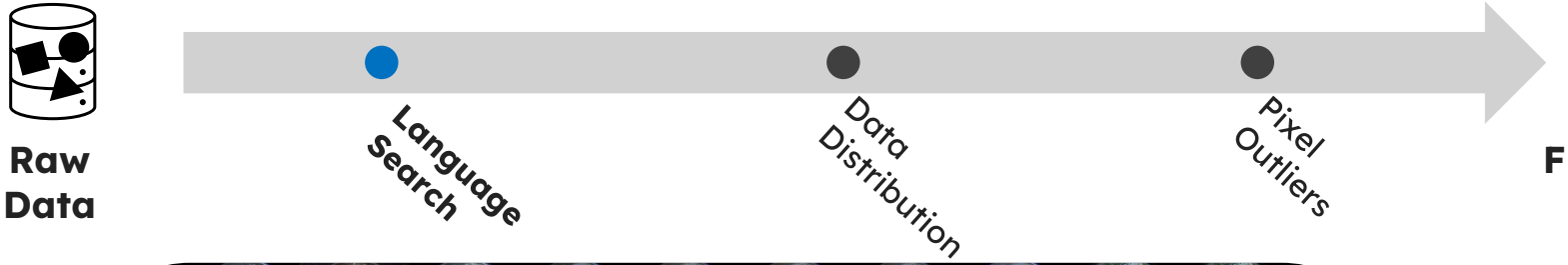
Instances Filter

3/5

Agreement Filter

Language Search

UNIVERSITY OF MICHIGAN



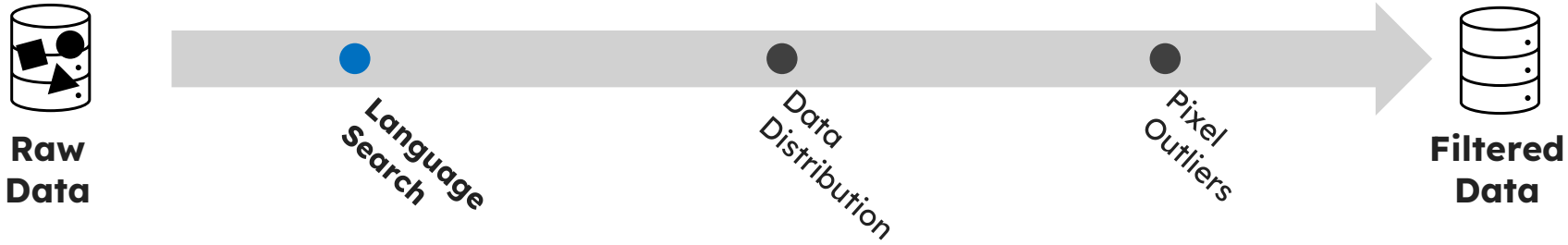
99% data reduction

Instances Filter

3/5

Agreement Filter

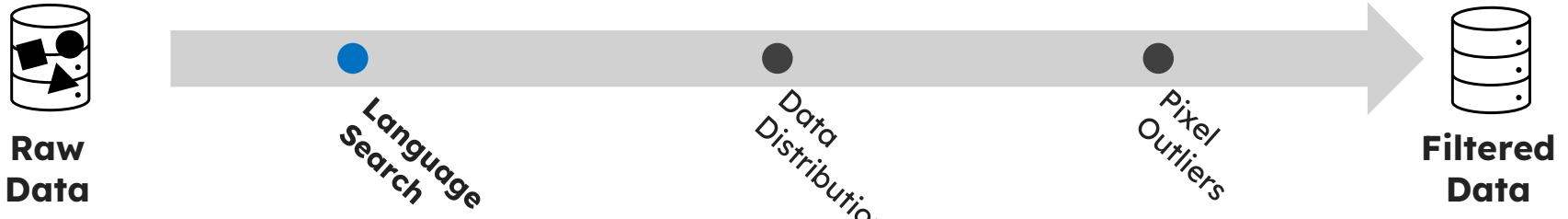
Language Search

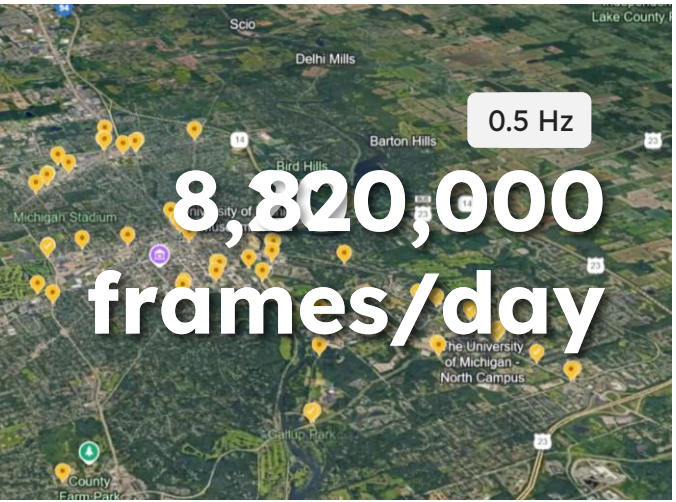


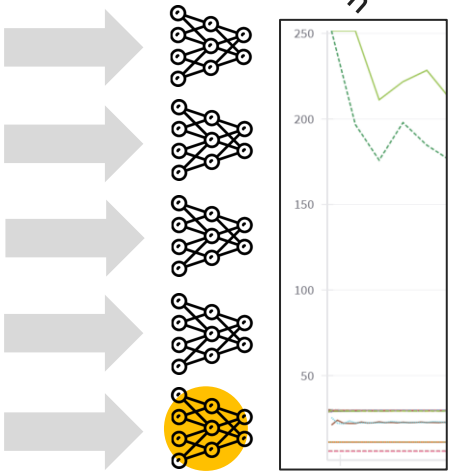
3/5

Agreement Filter

Language Search



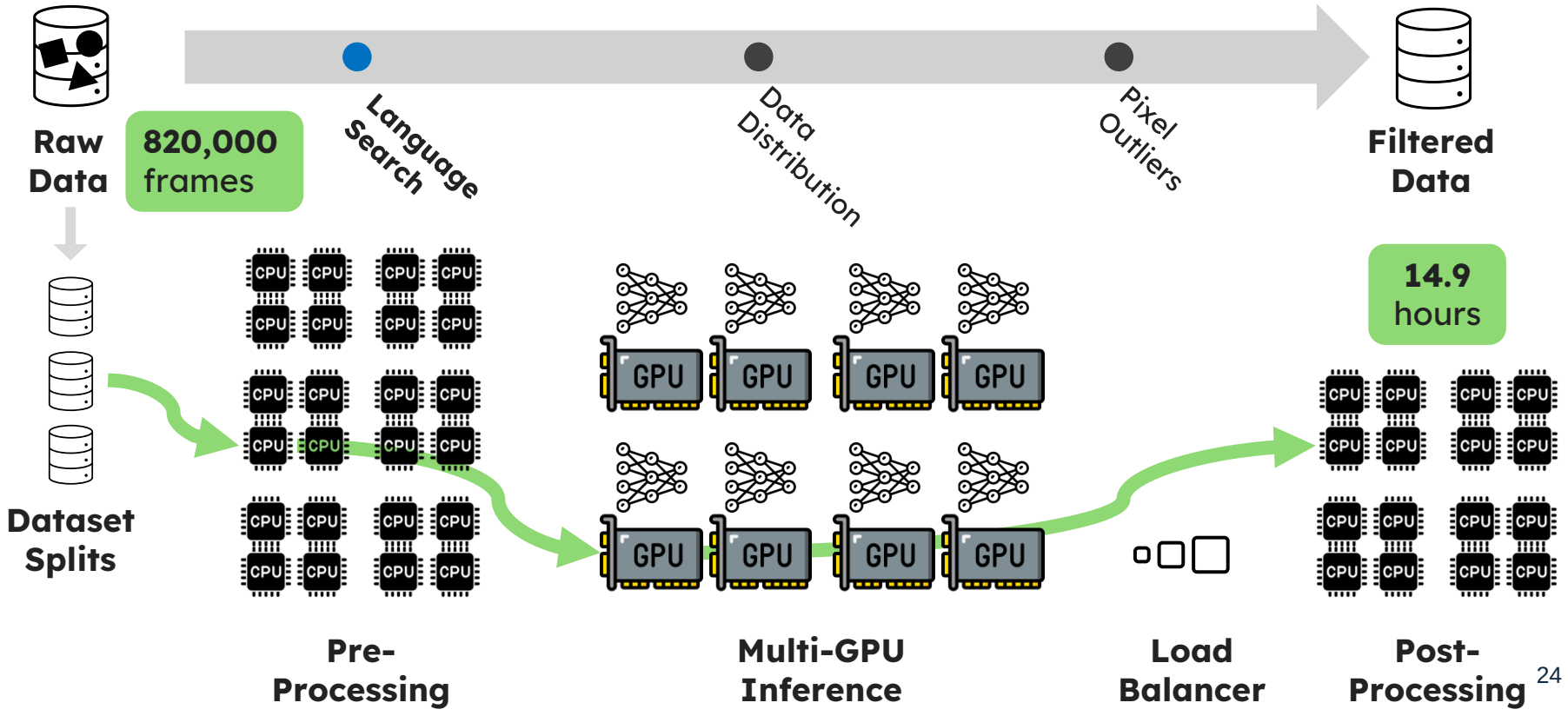




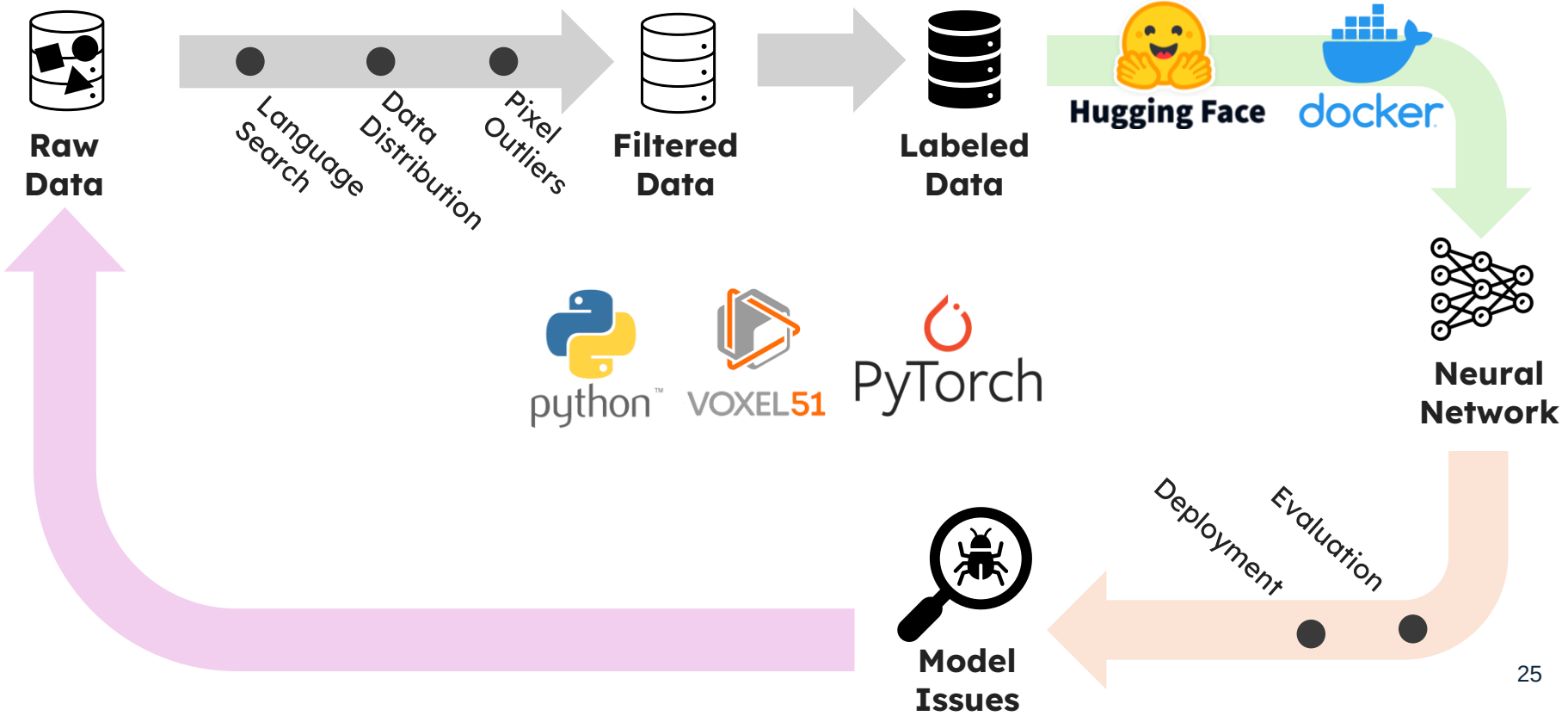
Zero Shot Ensemble

Google OWLv2
6 frames/second

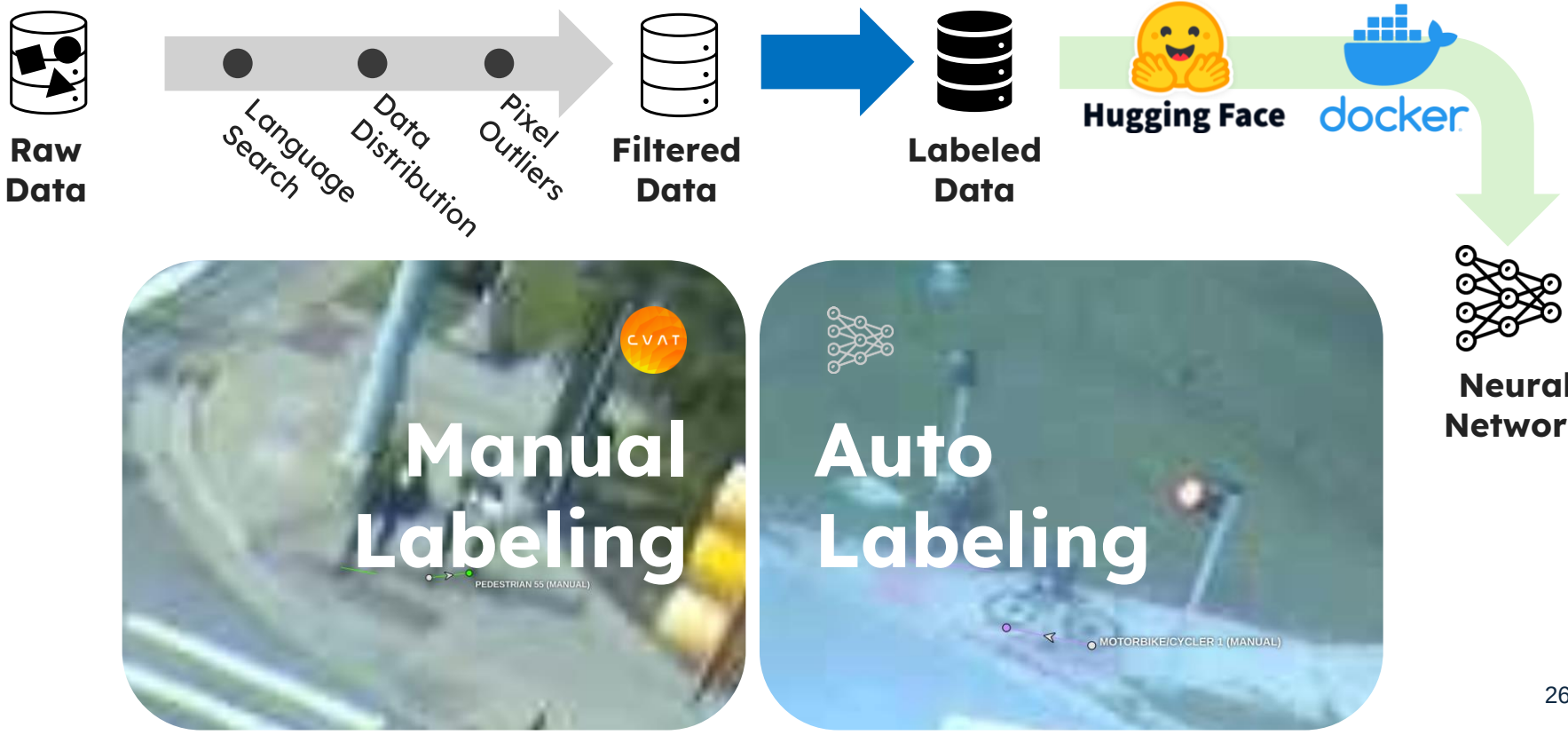
Language Search



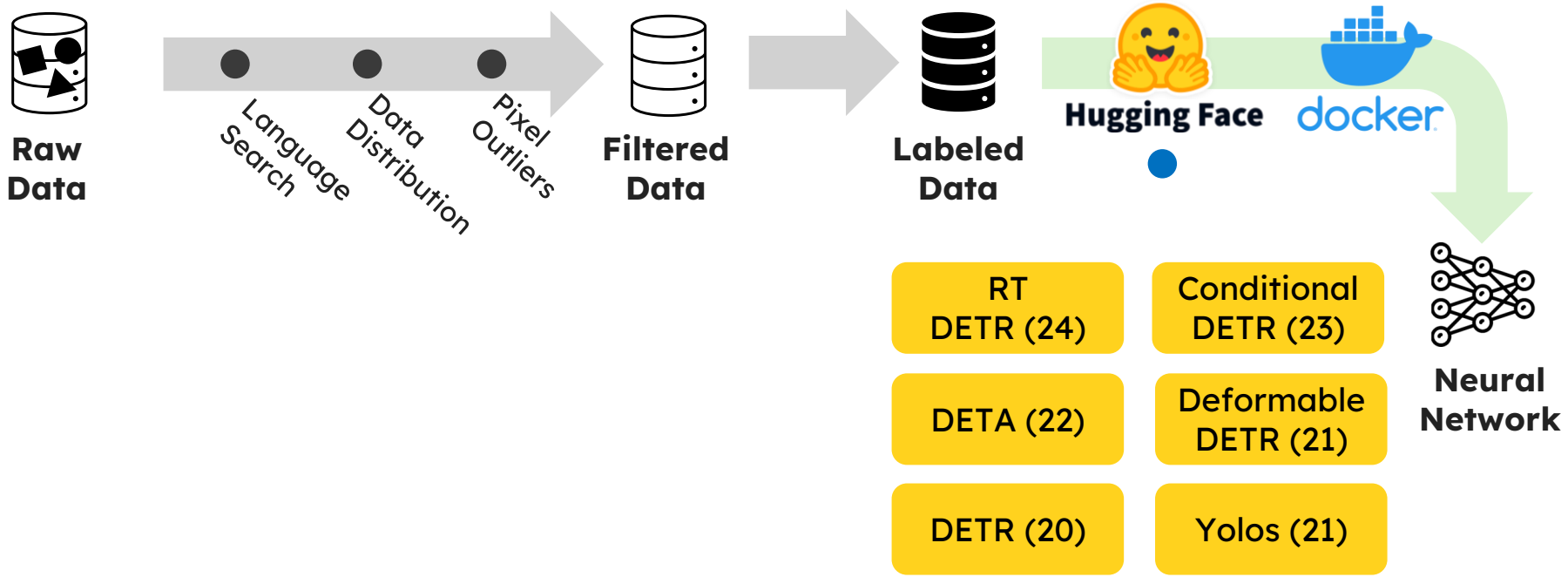
Mcity Data Engine



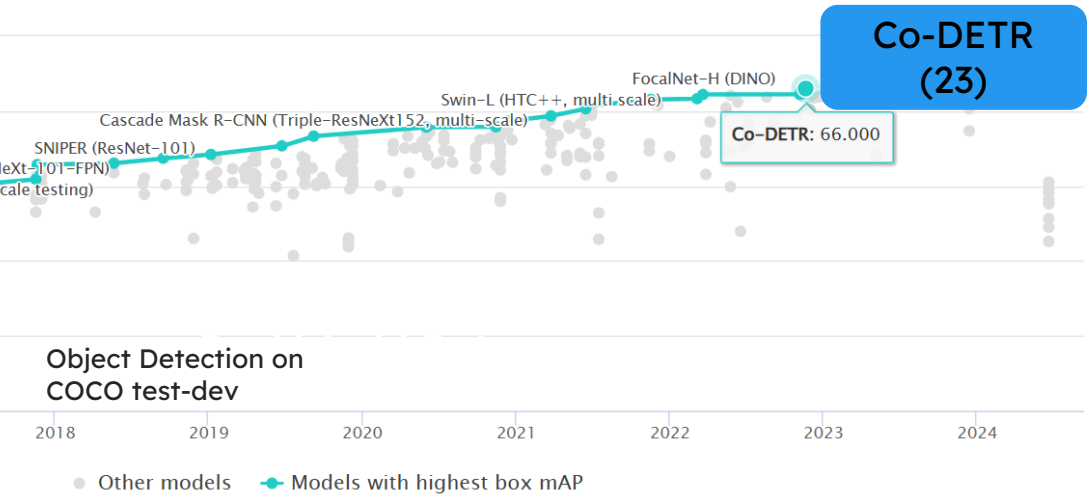
Data Labeling



Model Training



Model Training



RT DETR (24)

Conditional DETR (23)

DETA (22)

Deformable DETR (21)

DETR (20)

Yolos (21)

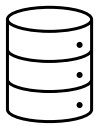
Model Training

Raw Data

Language Search

Data Distribution

Pixel Outliers



Filtered Data

Labeled Data

Hugging Face

docker

Neural Network

Co-DETR (23)

RT DETR (24)

Conditional DETR (23)

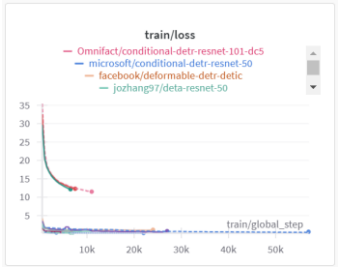
DETA (22)

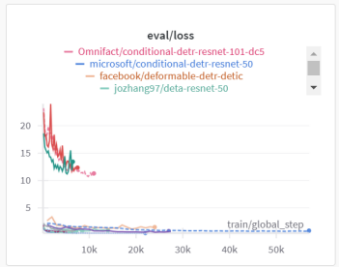
Deformable DETR (21)

DETR (20)

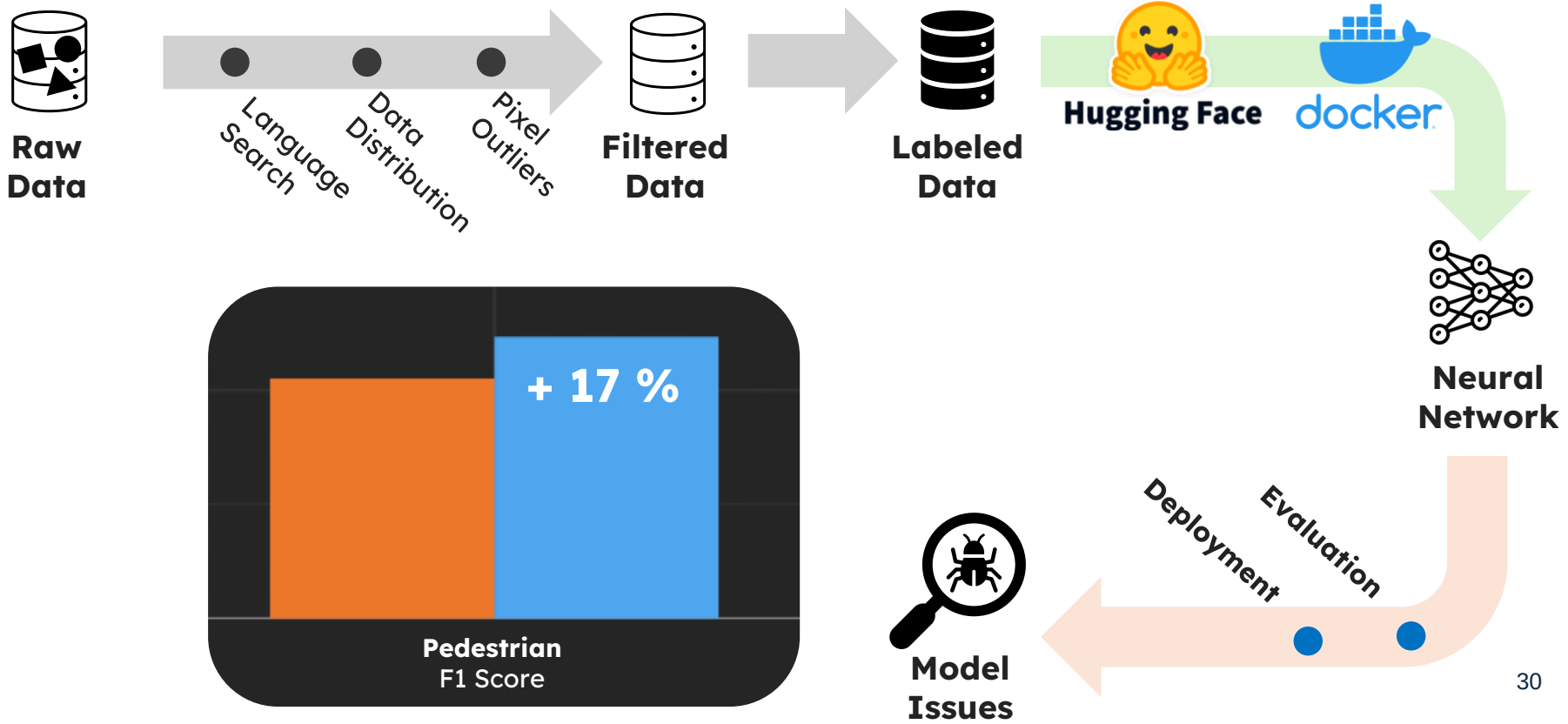
Yolos (21)

Weights & Biases

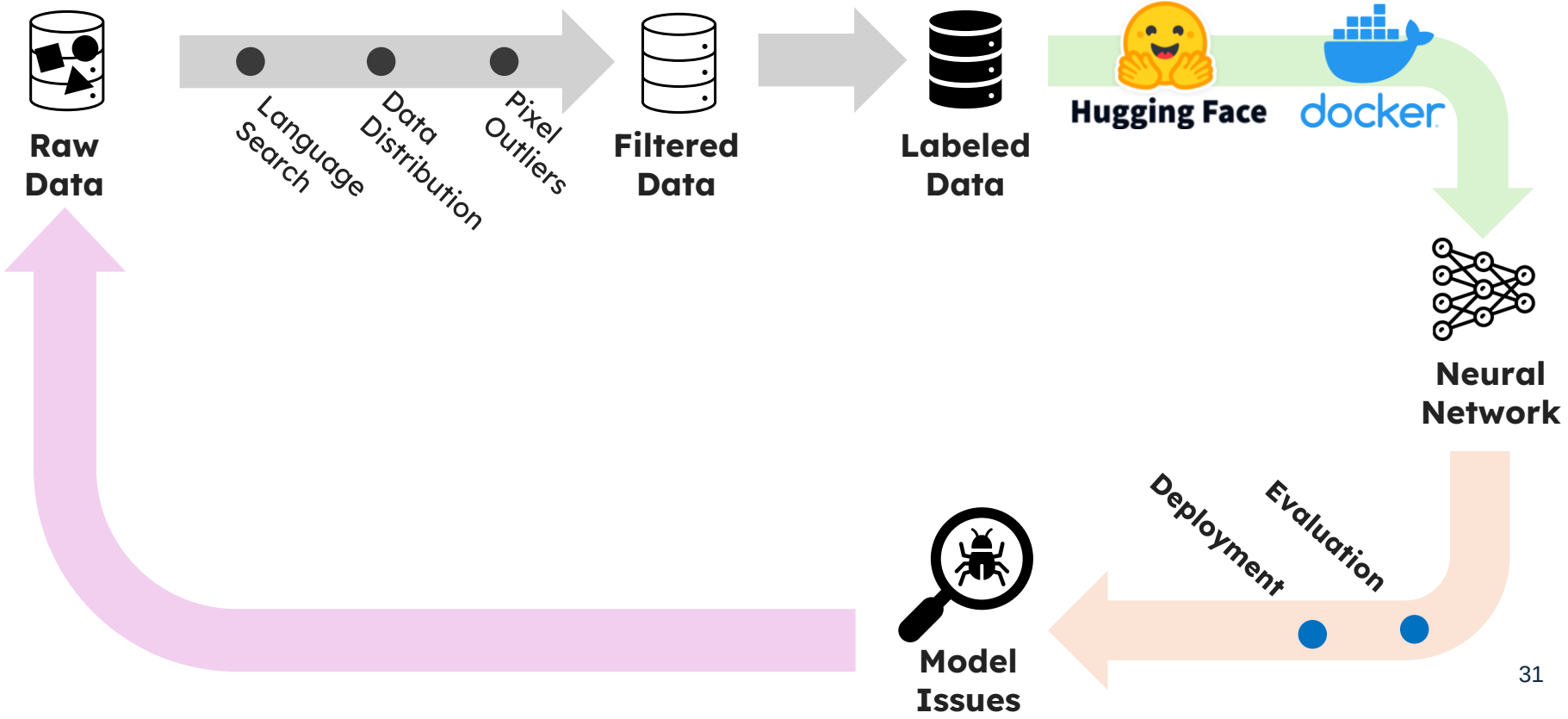




Mcity Data Engine

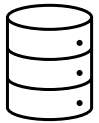


Mcity Data Engine



Mcity Data Engine

Raw Data



Filtered Data

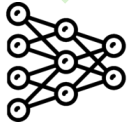
Labeled Data

Hugging Face

docker

Open Source Release
February 2025

github.com/mcity/mcity_data_engine



Neural Network

Model Issues

Mcity Data Engine

Linked in

Daniel Bogdoll

Thank you!
Questions?