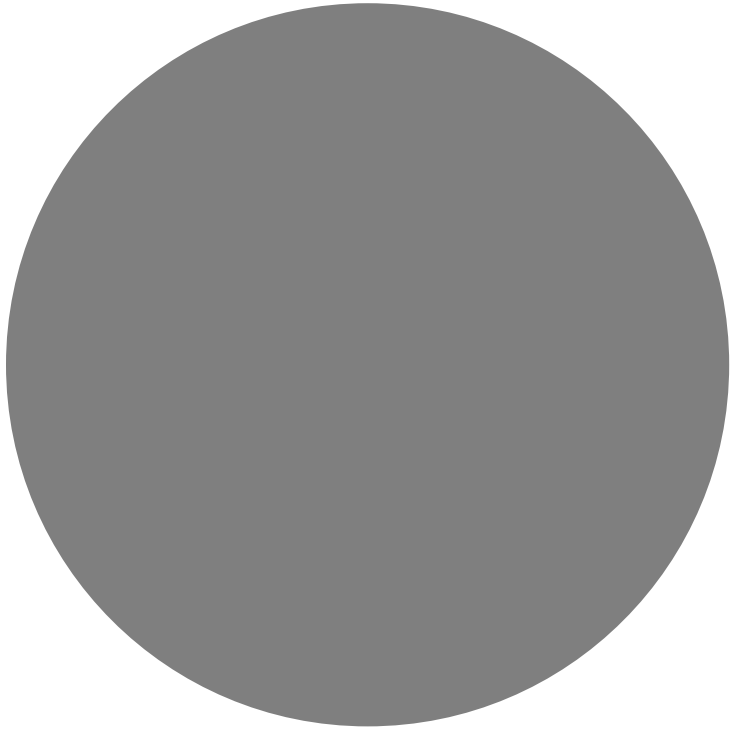
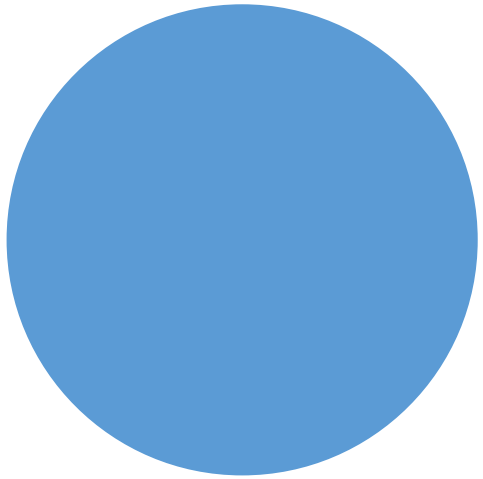
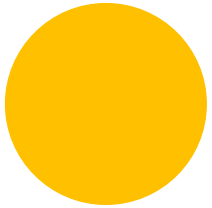
Introduction to Data Labeling for Machine Learning



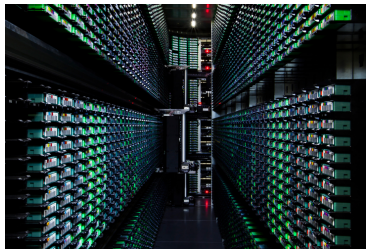
Nikolai Liubimov
Ph.D. in Computer Science
Founder & CTO at Heartex (www.heartex.ai)
AI-assisted data labeling solutions

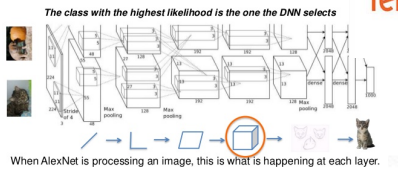
Deep learning®evolution

MLaaS

Machine Learning as a Service

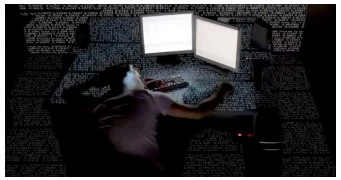


AlexNet (Krizhevsky et al. 2012)



When AlexNet is processing an image, this is what is happening at each layer.

TensorFlow vs PYTORCH

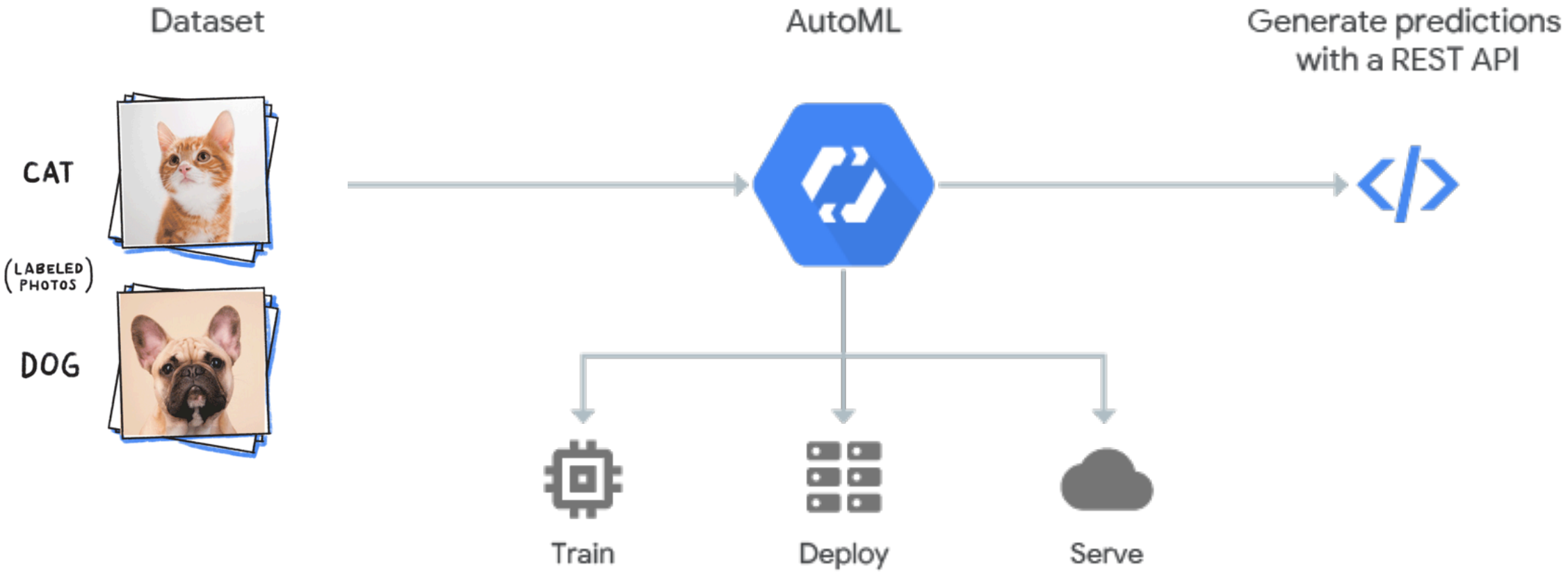


scikit learn

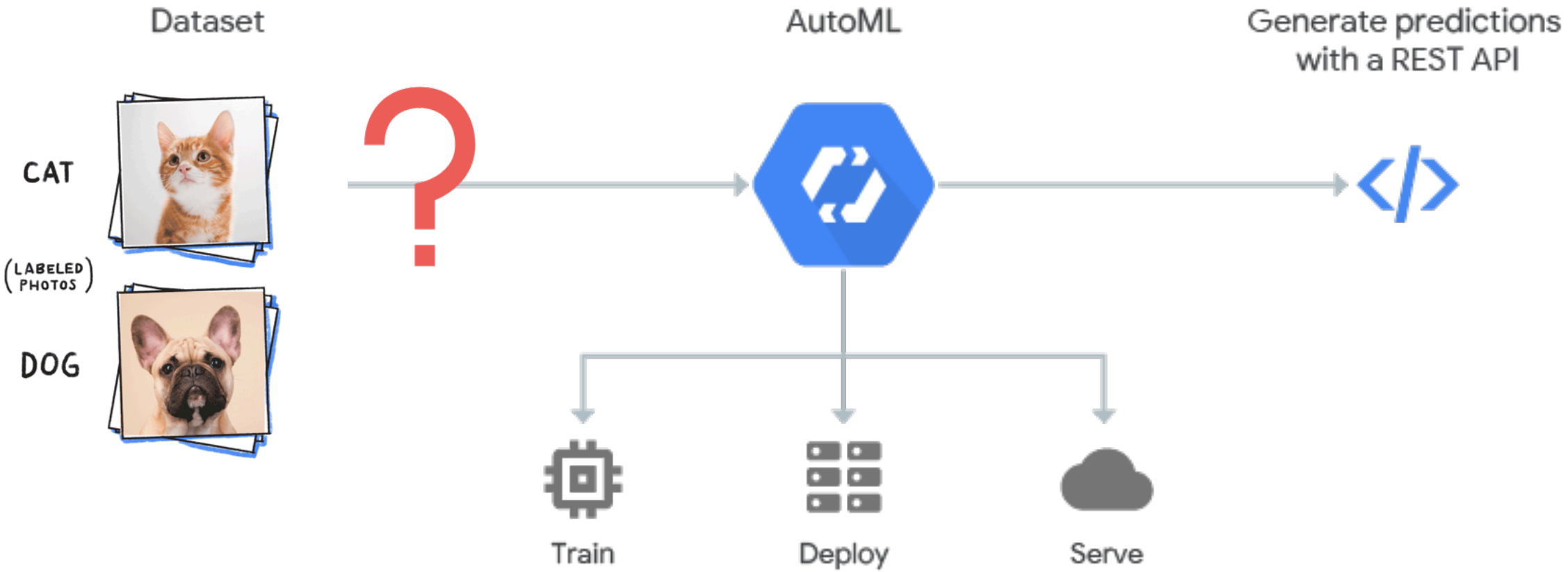
2006

2020

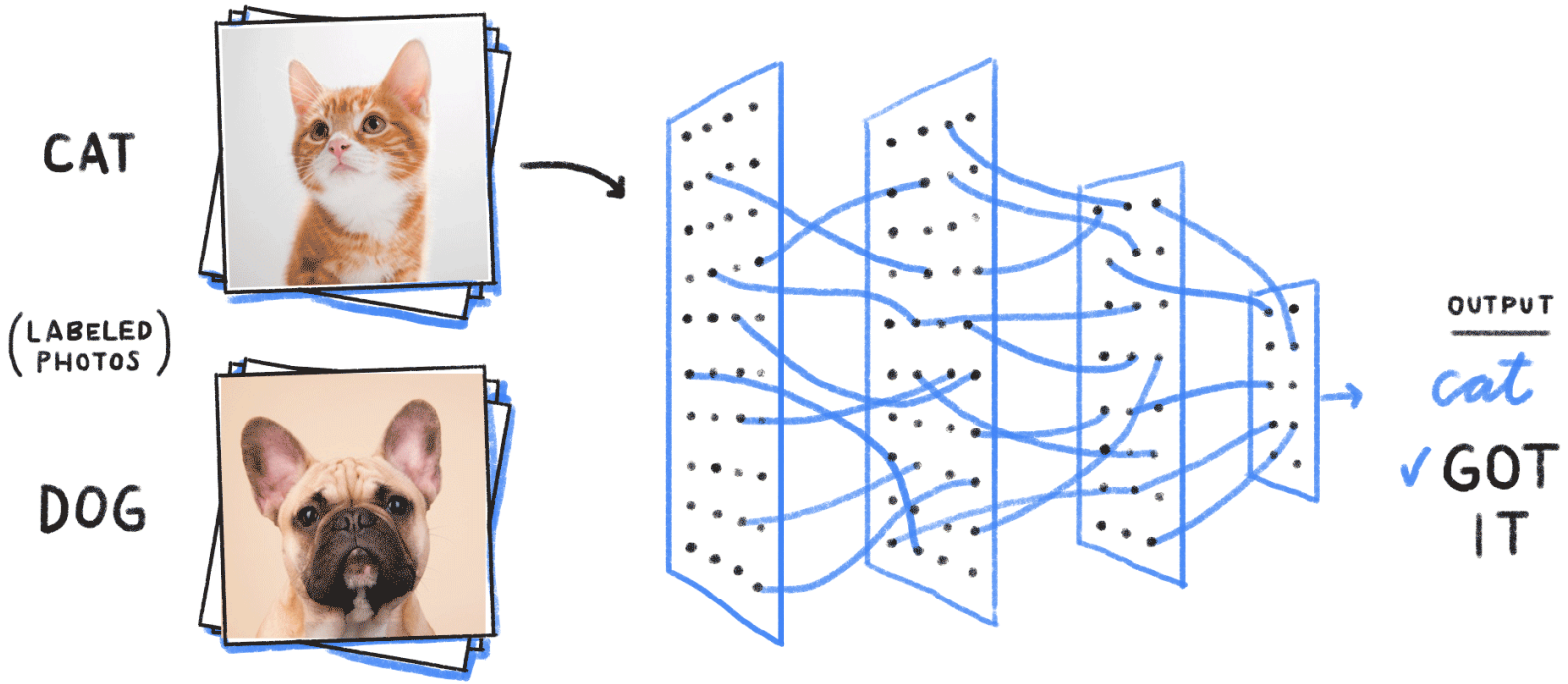
Machine learning as a Service



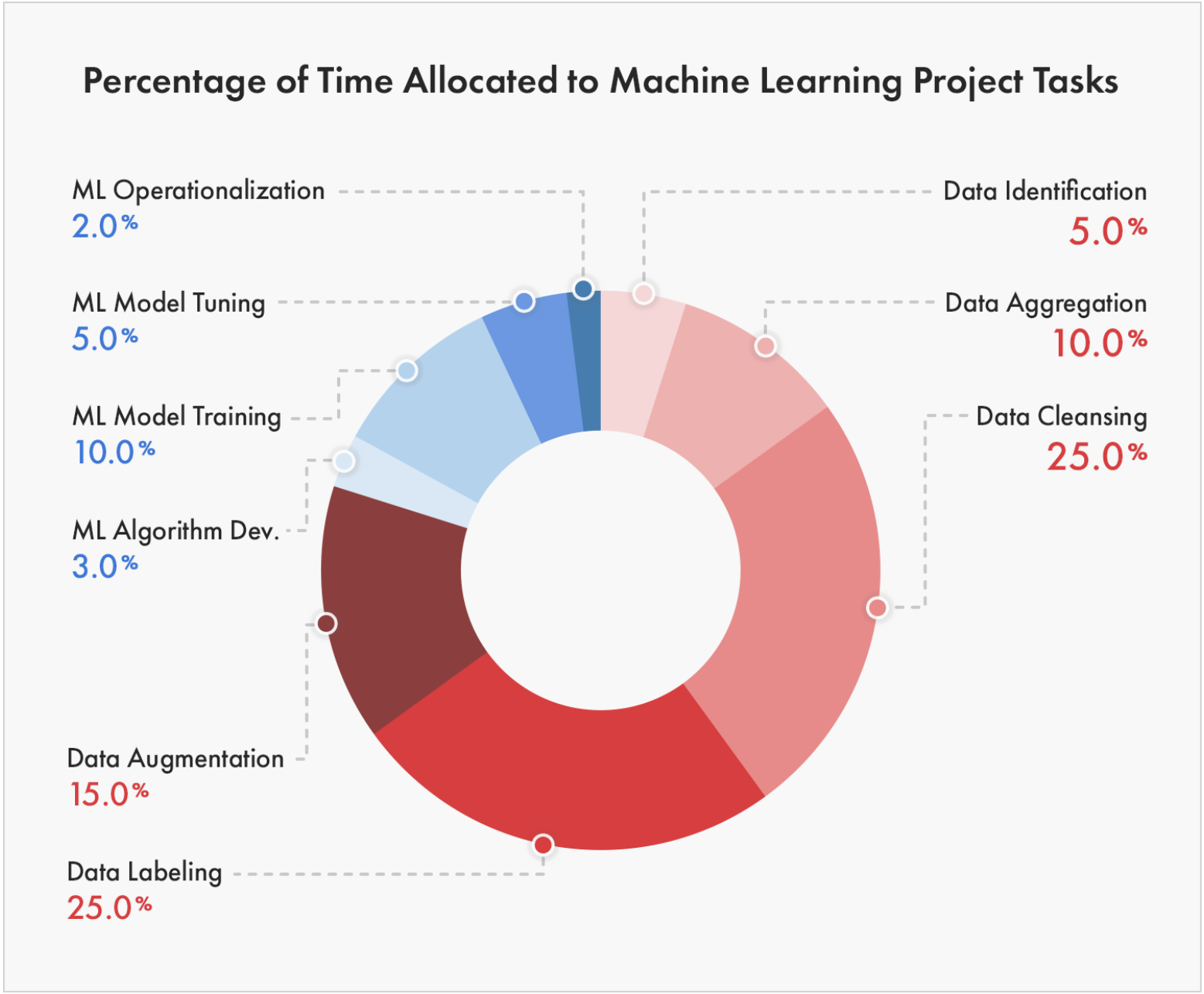
Machine learning as a Service



Neural networks

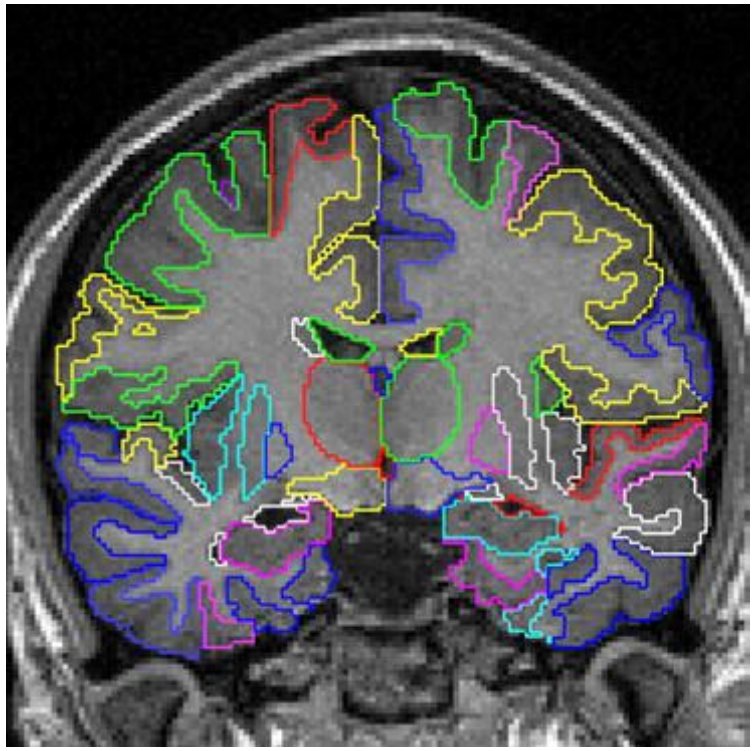


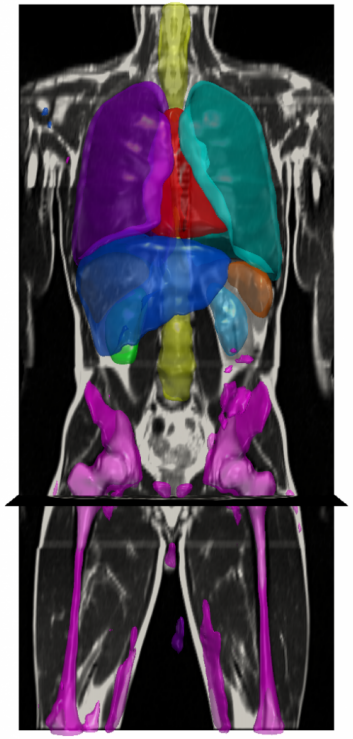
Learning is Labeling?



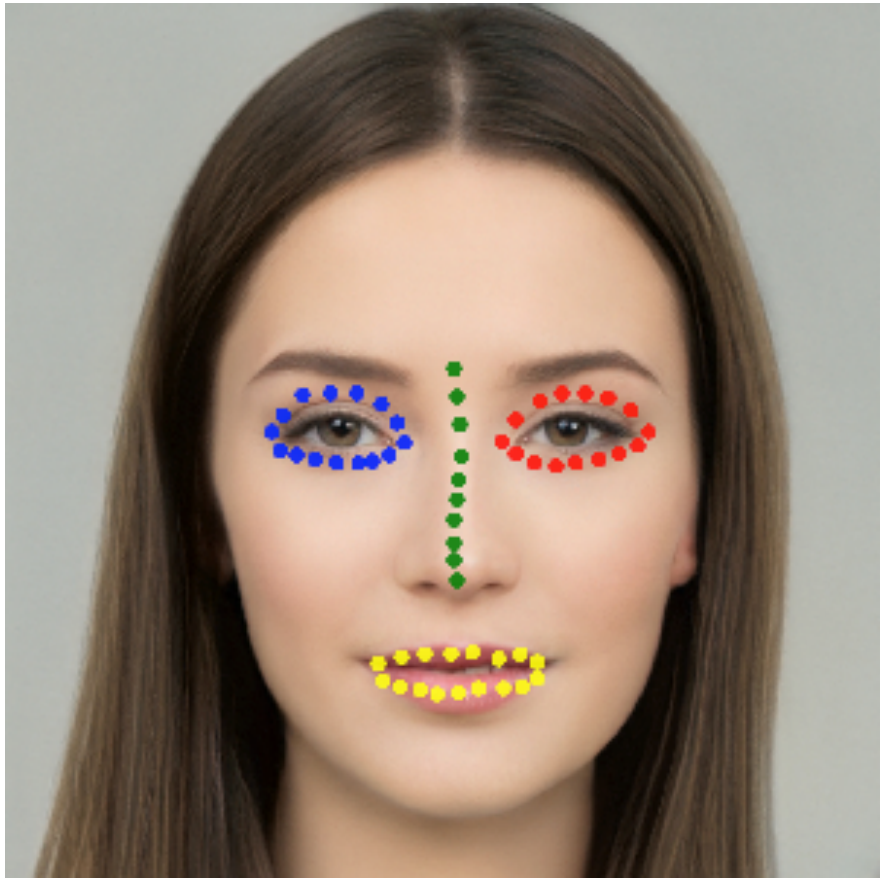
Cognylitica Research. Data engineering, Preparation and Labeling for AI 2019

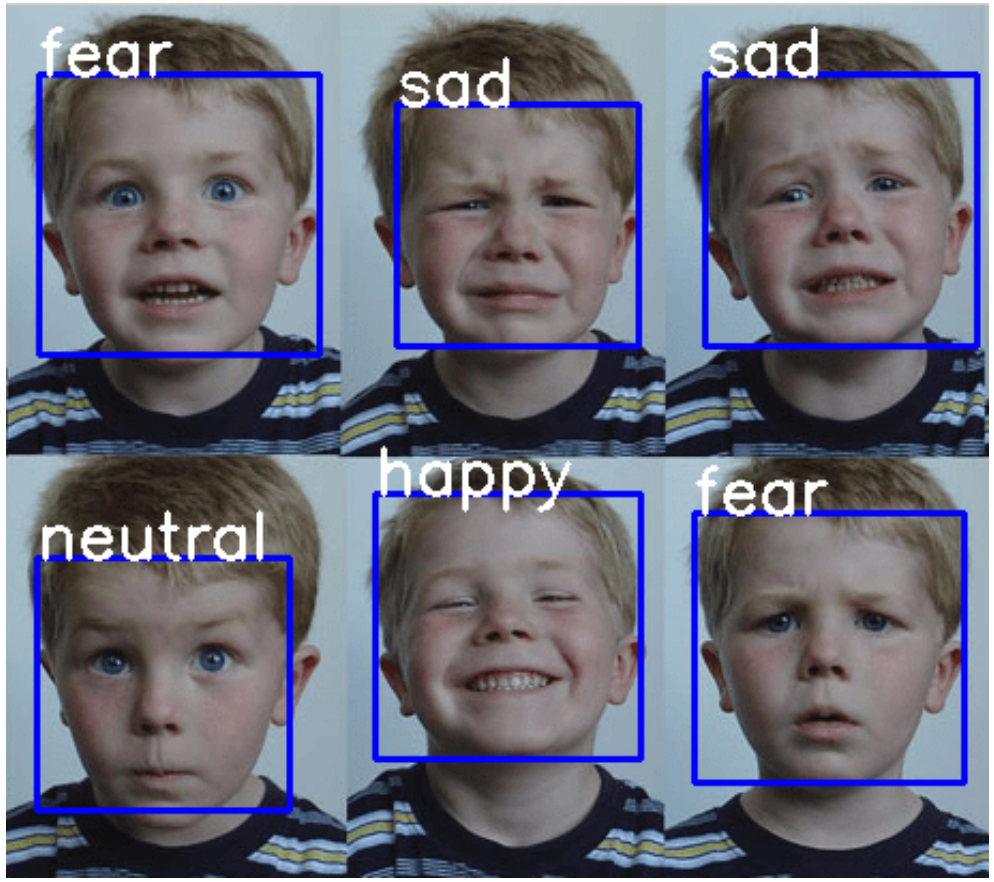
Medical imaging





Facial analysis & Emotion recognition





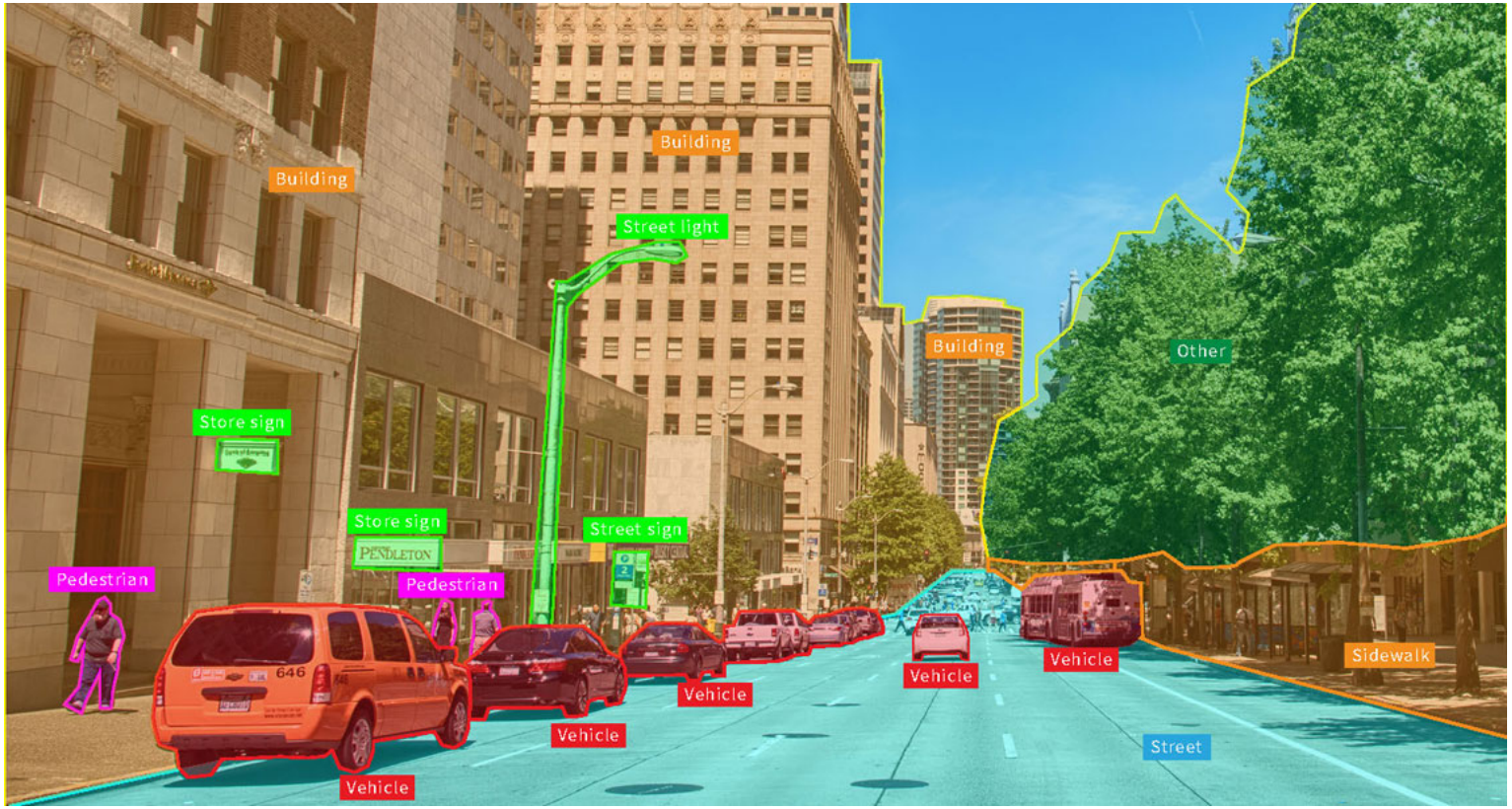
E-commerce & Retail

Christian Dior Lady Di
Bag Canage white lea
Women's Bag Crossbo

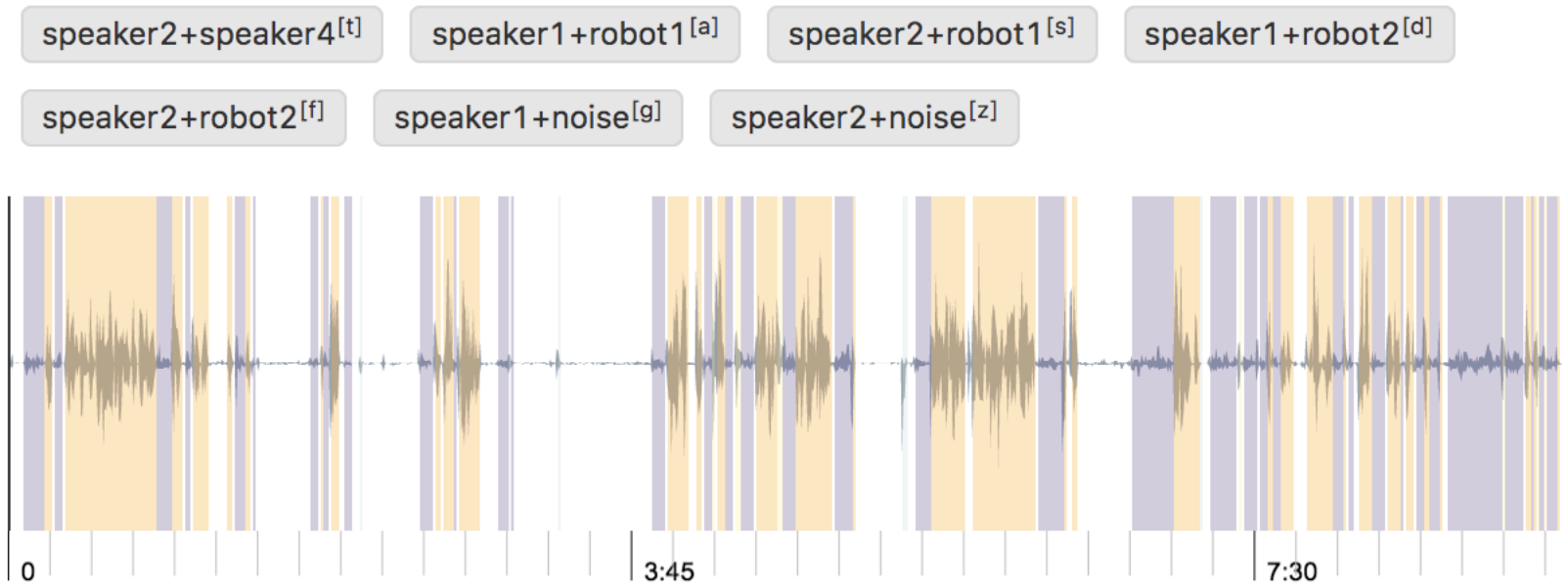
- one features fashiona pattern, another one
- long sturdy strap in s can't be removable, m 119cm
- 10.2" wide x 8.3" tall x weight: 1.35lb

Price: $569

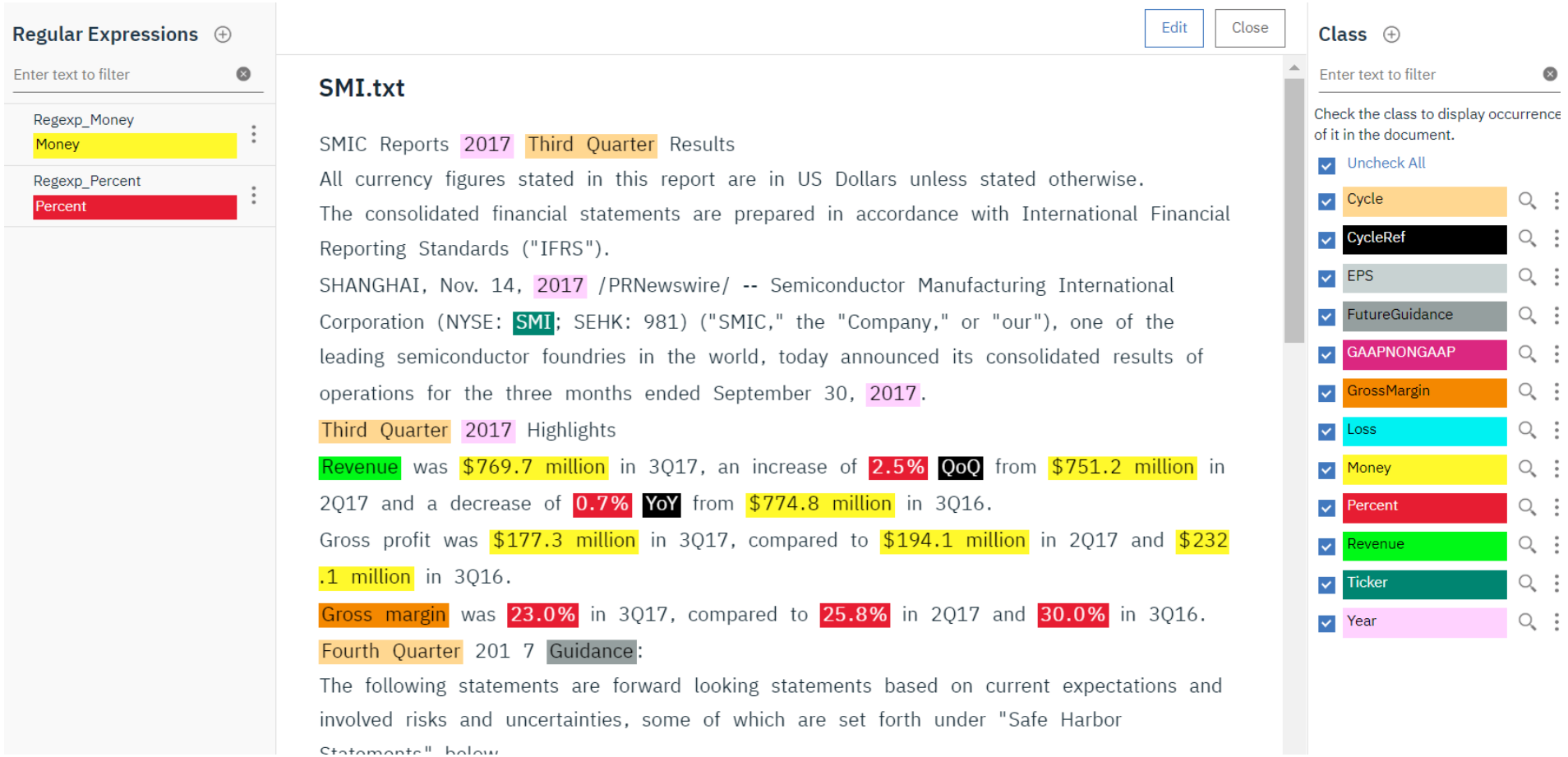
Autonomous vehicles



Speech analysis & conversational analytics



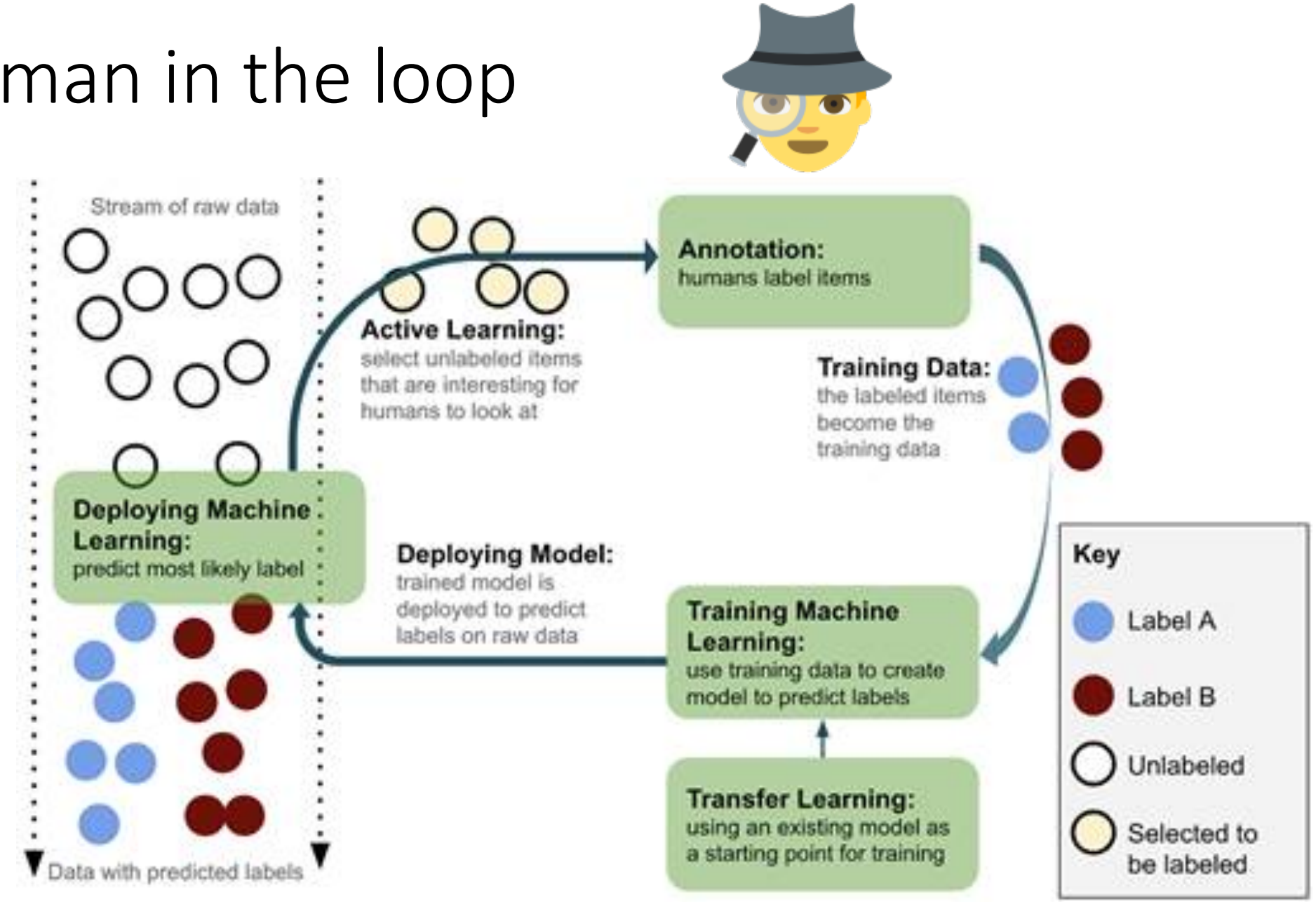
Document processing



Data labeling approaches

Approach	Description	Pros	Cons
Internal labeling	Assignment of tasks to an in-house data science team	✓ High accuracy ✓ Track progress	✗ Takes much time & money
Crowdsoursing	Cooperation with freelancers and labeling workforce from 3rd party services	✓ Fast results	✗ Quality of work can suffer ✗ Not easy to manage
Synthetic labeling & data programming	Generating and programmatically labeling data using scripts	✓ Very fast results ✓ No manual work	✗ Lower quality dataset ✗ Needs computational power

Human in the loop



What are the challenges?

Labeling cost money (0.1$ per image annotation results to >1M$ for ImageNet)

=> need to pay less money for the same modeling quality

Human-in-the-loop cycle is long (recovering model errors only after it is deployed in production)

=> mitigate the risk of execution failures at early stages

Labeling is not accurate

=> monitoring & continuously control modeling quality over labeling process

Building tool for every labeling task is tedious

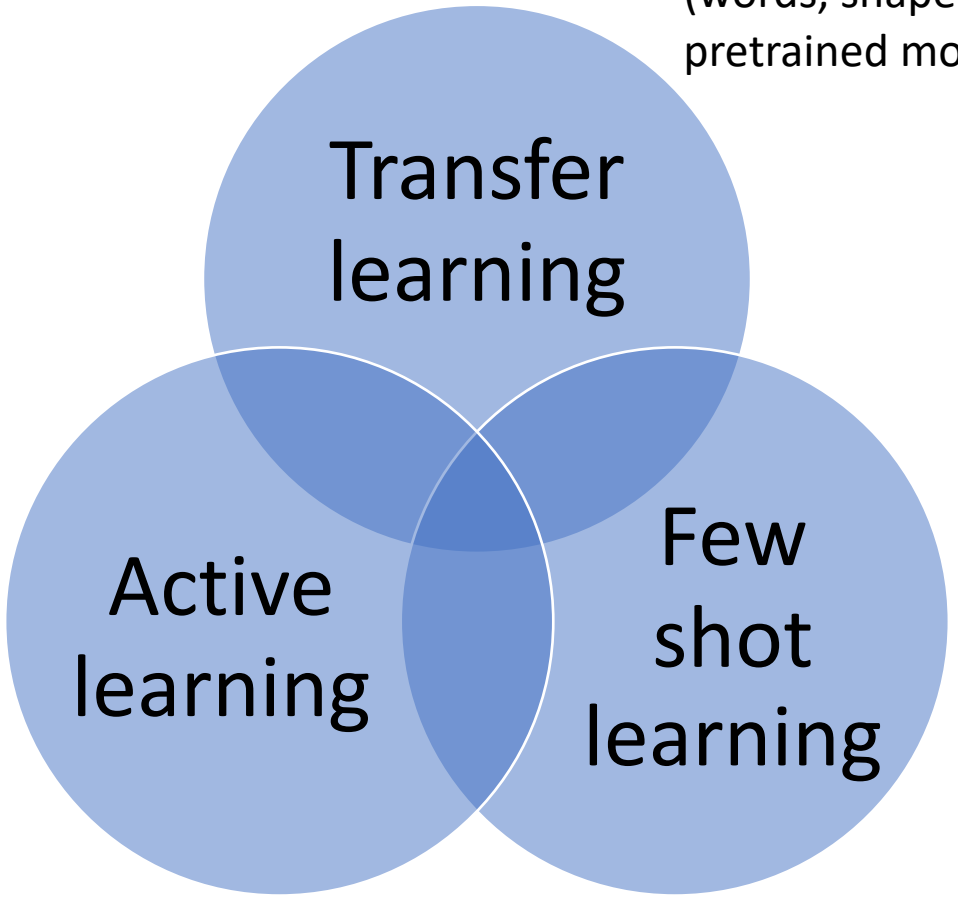
=> flexible configuration tools

We can’t crowdsource private data

=> depersonalize data before annotation

Reducing labeling cost

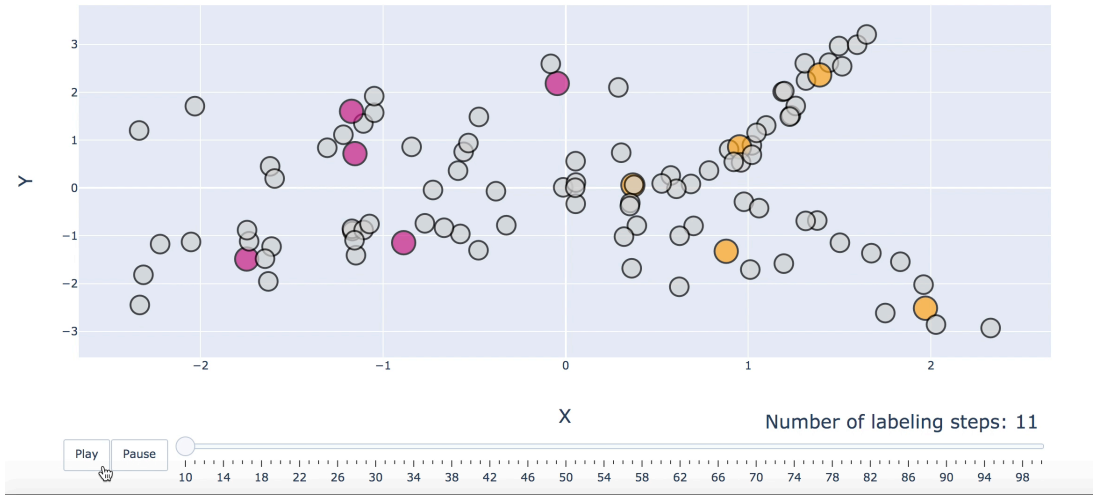
Capture low-level semantics (words, shapes, etc.) from pretrained models



Pick only most informative examples

Quickly adapt to unseen tasks

Learn faster with smarter data labeling



Try to pick and label samples to promote:

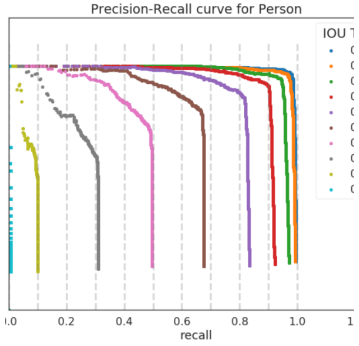
- Uncertainty (class borderline)
- Informational density
- Diversity

Quality control & assessment

Annotator consensus: how labelers are agreed with each other on specific item

Modeling consensus: how model predictions are agreed with annotator labels

Ground truth check: how well labelers perform on tasks with hidden answers (”honeypots”)



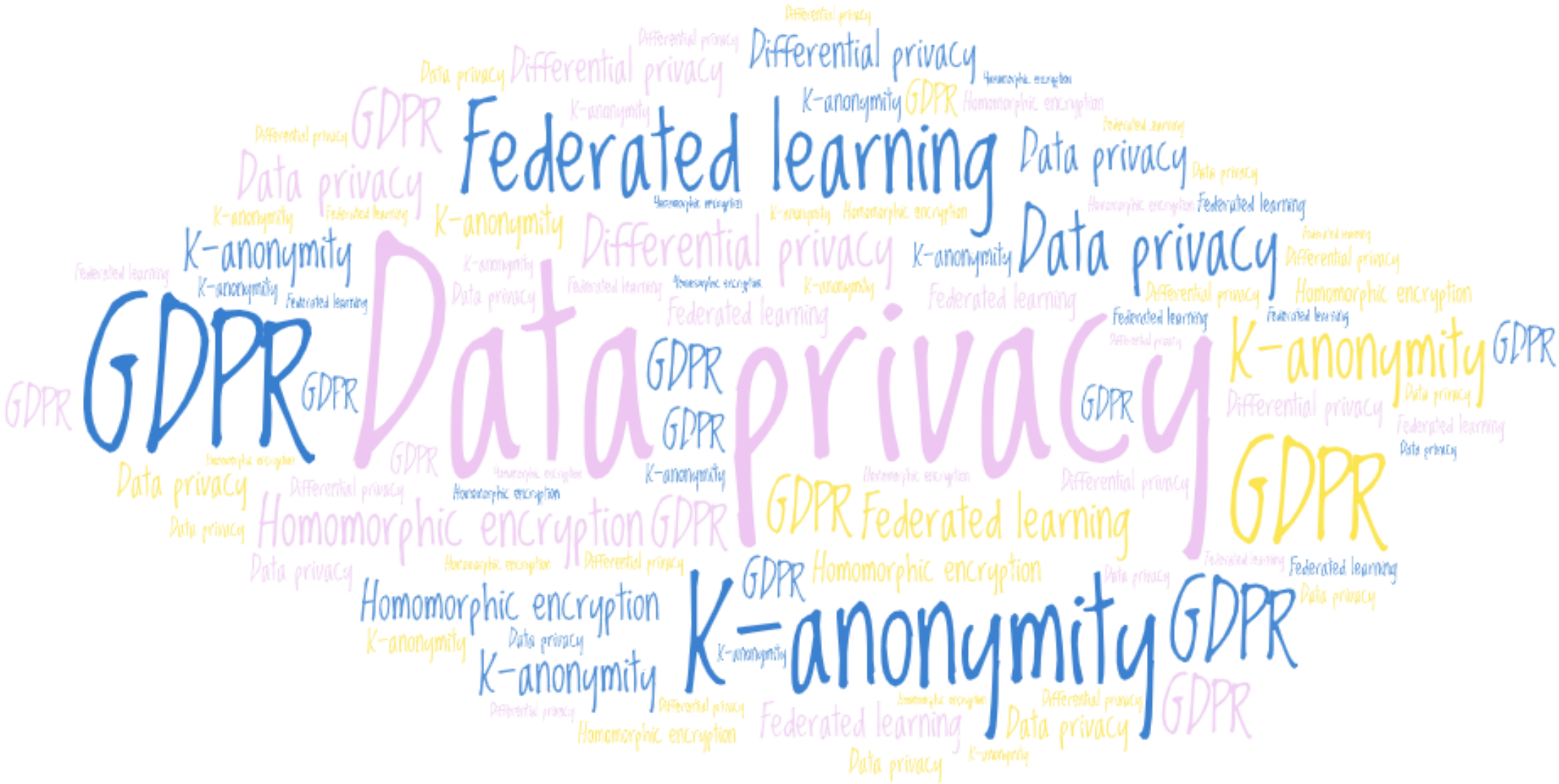
Modeling quality: measuring model performance on holdout datasets (cross-validation)

Building labeling tool is like building web site

- Hypertext-like configuration language

```
<View>
    <Choices name="Pets" toName="Image" choice="single">
        <Choice value="Cat"></Choice>
        <Choice value="Dog"></Choice>
    </Choices>
    <Image name="Image" value="$image_url"></Image>
</View>
```

- Easy to embed in any ecosystems with ongoing annotations (e.g. user clicks or medical imagery)


Data privacy
GDPR
Federated learning
K-anonymity
Differential privacy
Homomorphic encryption

Thank you! ☺

Nikolai Liubimov
nik@heartex.ai
linkedin.com/in/liubimov/

heartex.ai

labelstud.io