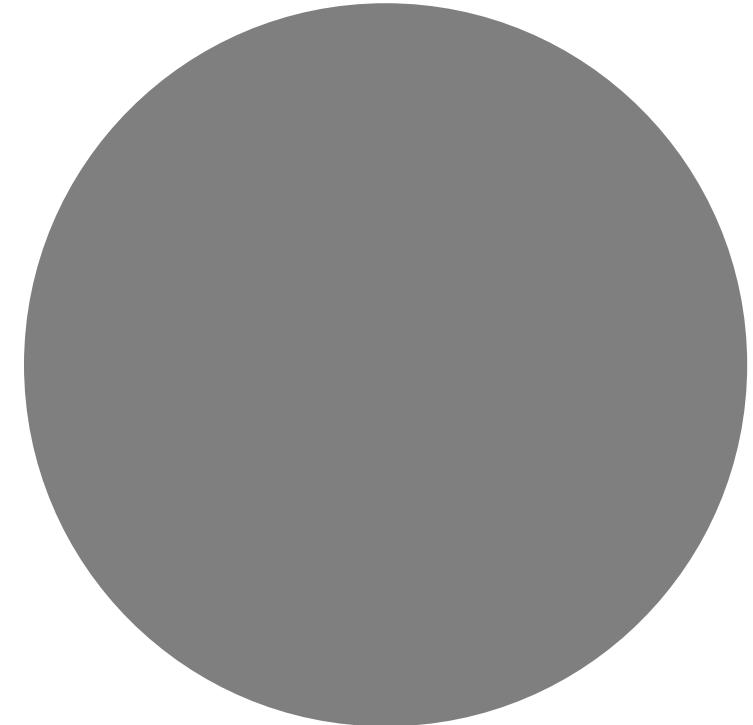
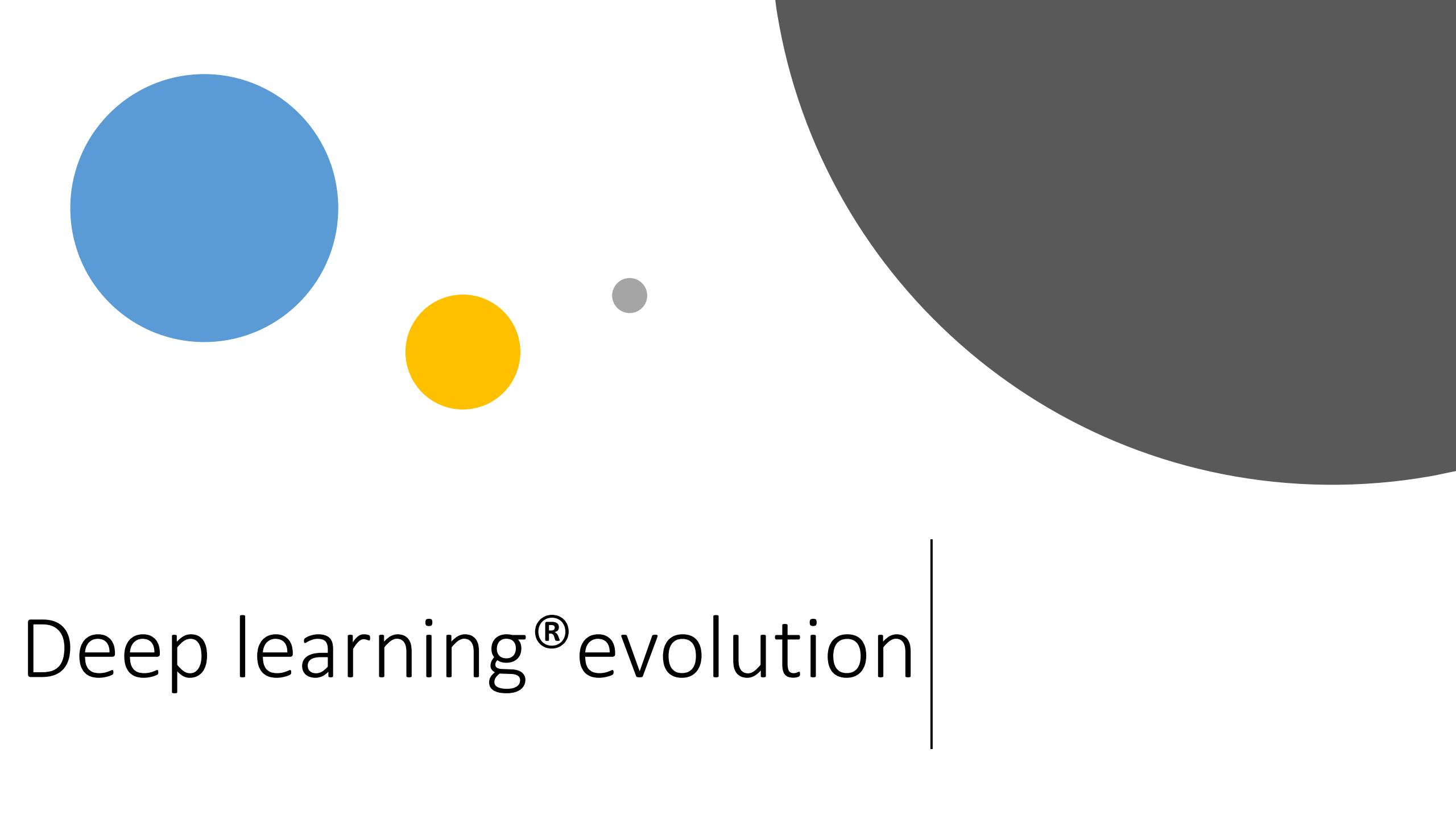


Introduction to Data Labeling for Machine Learning

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



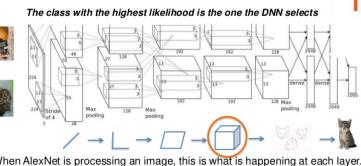


The background features a minimalist graphic composed of overlapping circles. A large blue circle is positioned in the upper left. To its right is a smaller yellow circle, and further right is a tiny gray circle. A large dark gray circle is partially visible on the far right edge of the frame.

Deep learning®evolution



AlexNet (Krizhevsky et al. 2012)



vs PYTORCH



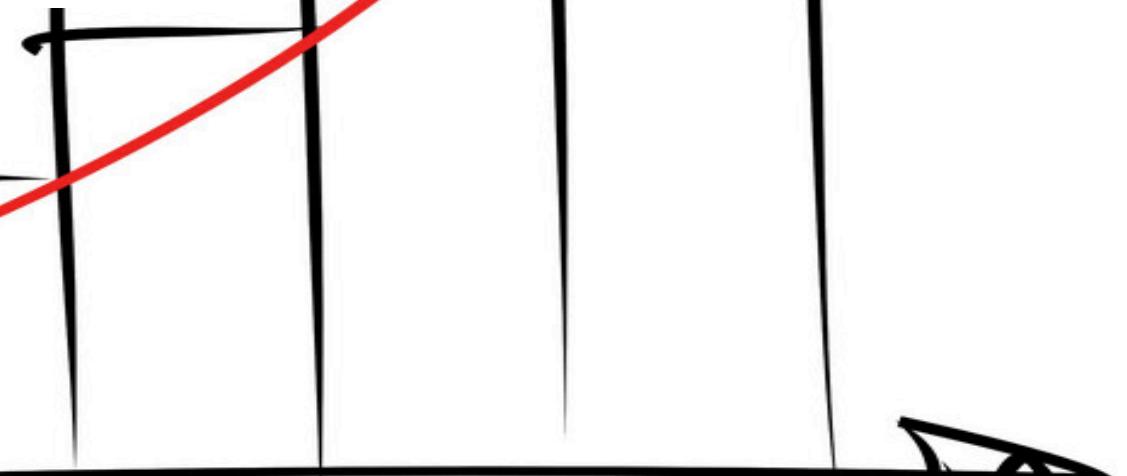
MLaaS

Machine Learning as a Service

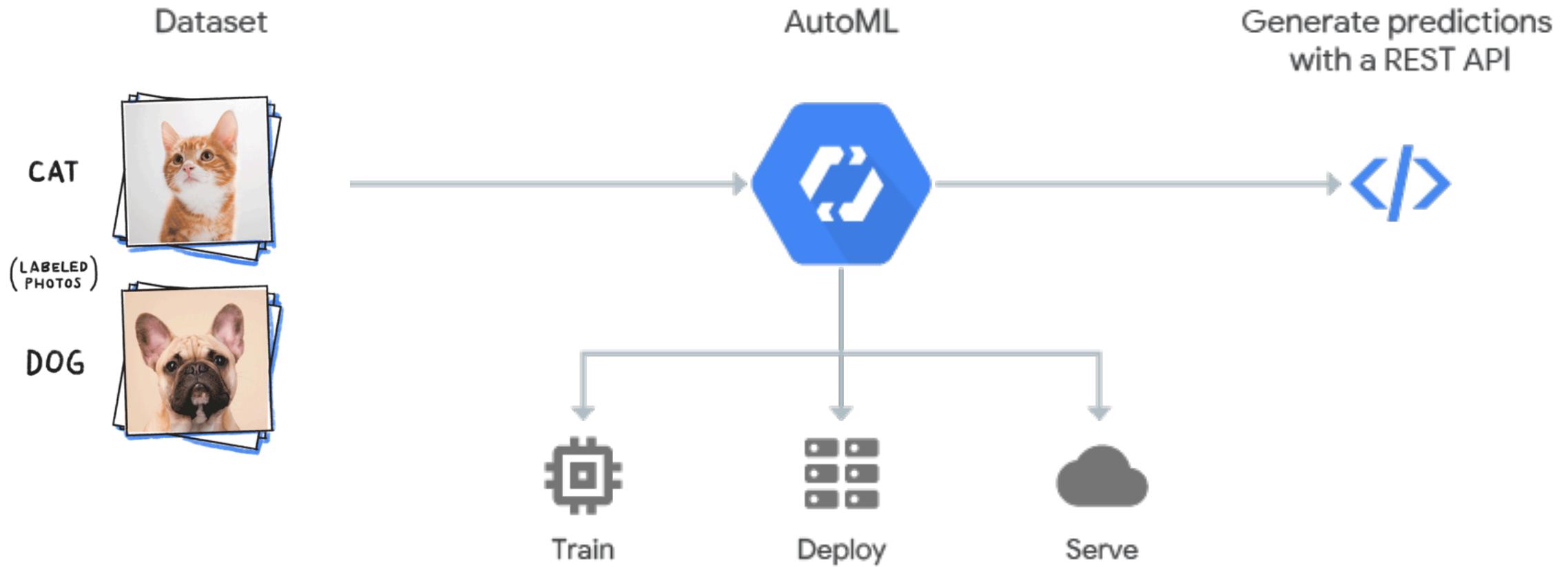


2006

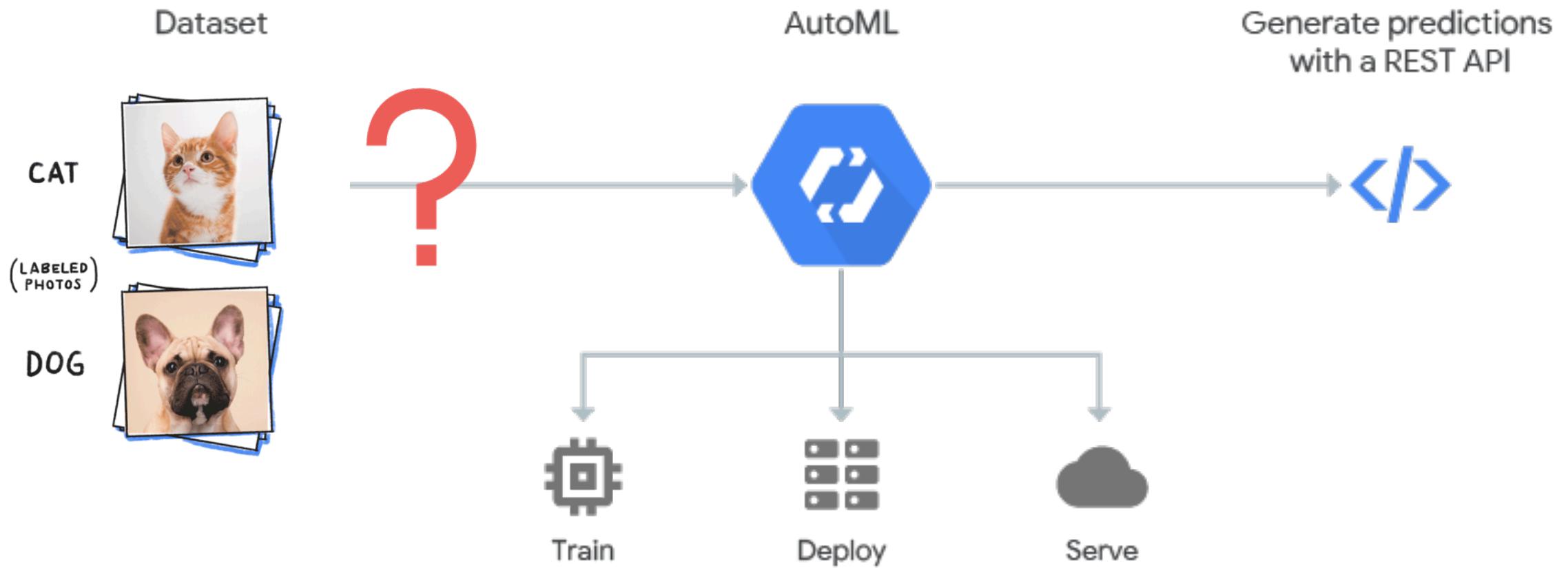
2020



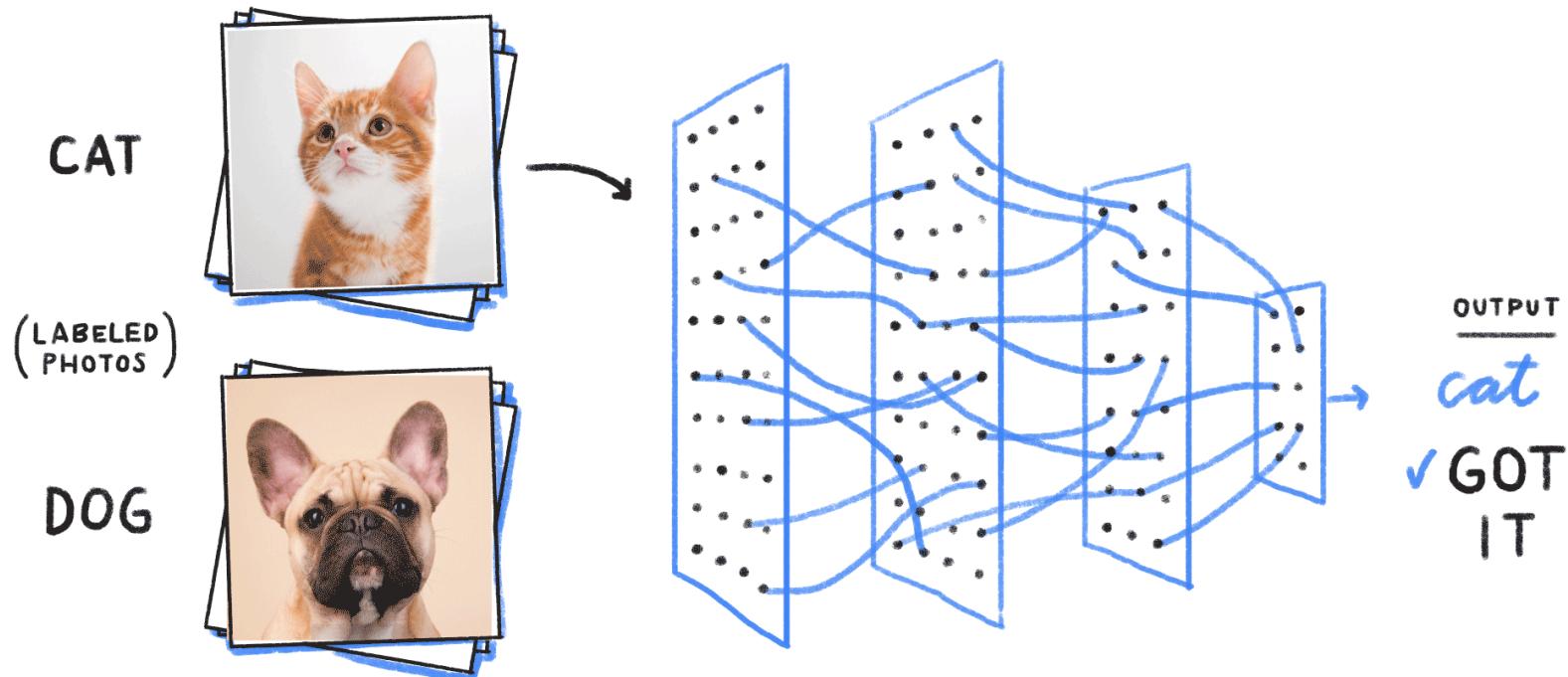
Machine learning as a Service



Machine learning as a Service



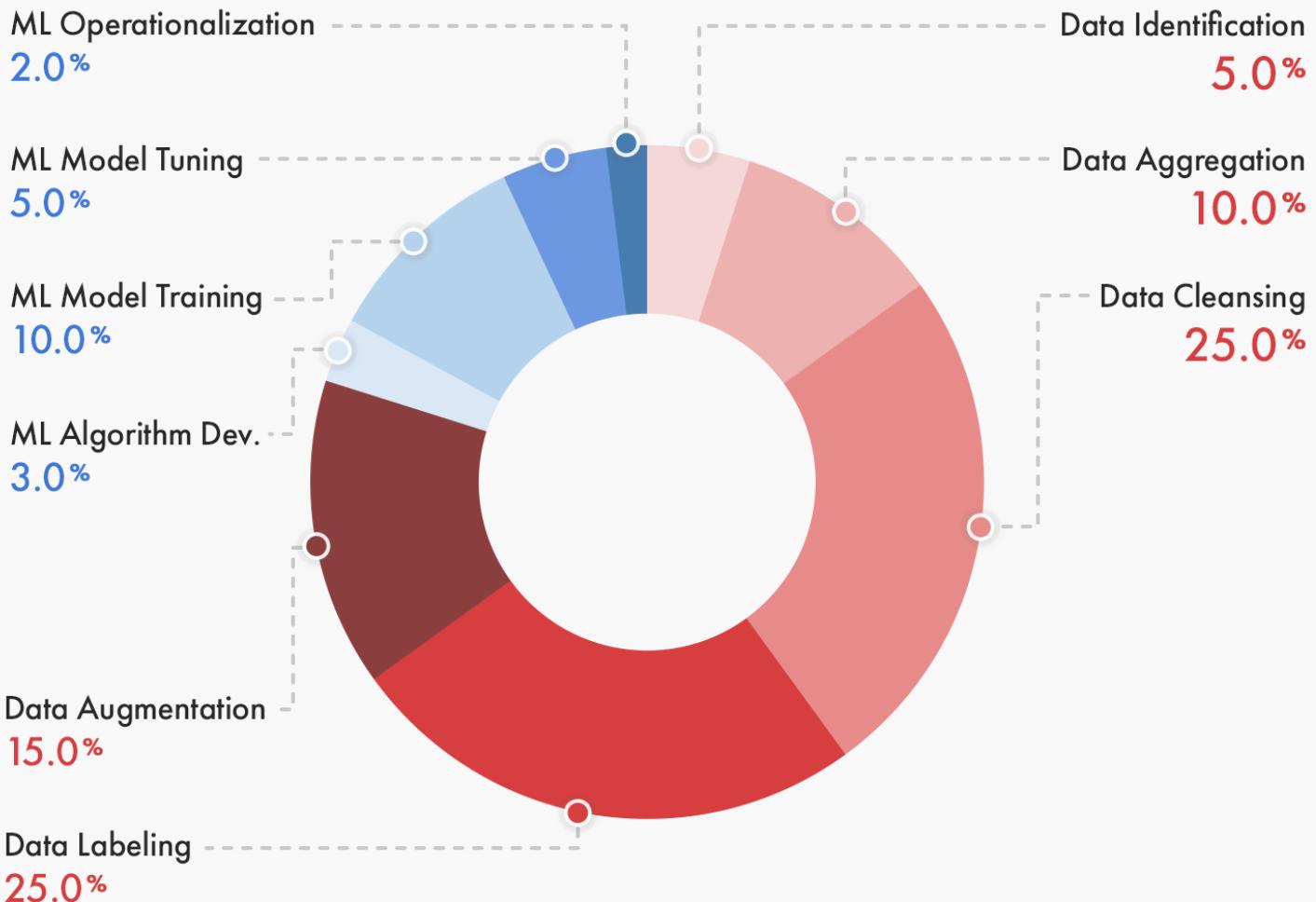
Neural networks



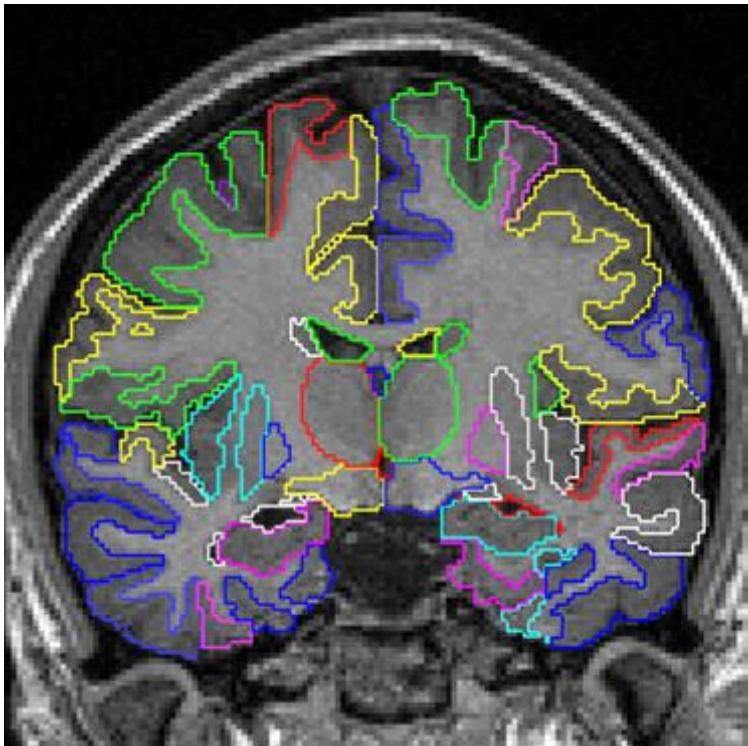
A photograph of two young children, a boy and a girl, sitting at a table. The boy, on the left, has curly brown hair and is wearing a blue t-shirt. He is looking down at something on the table. The girl, on the right, has long dark hair and is wearing a pink shirt. She is smiling and also looking down at the same thing. In the bottom right corner of the image, there is a small glass filled with a yellow liquid, possibly juice.

Learning is Labeling?

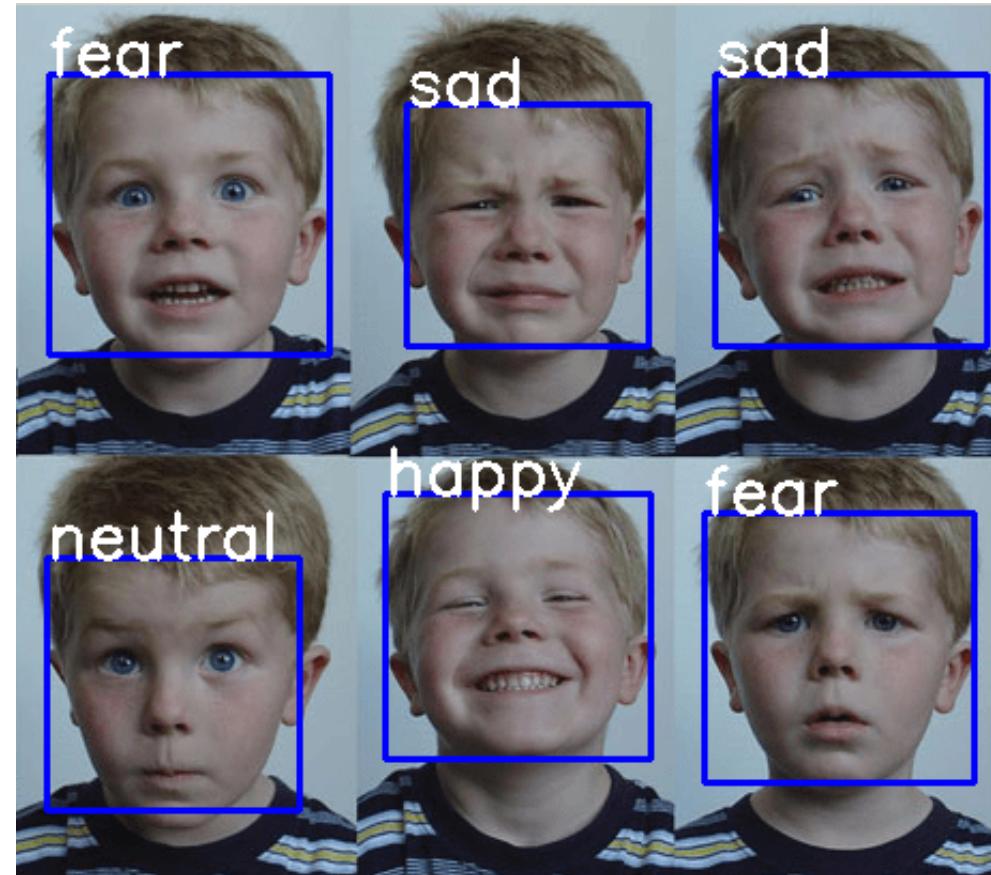
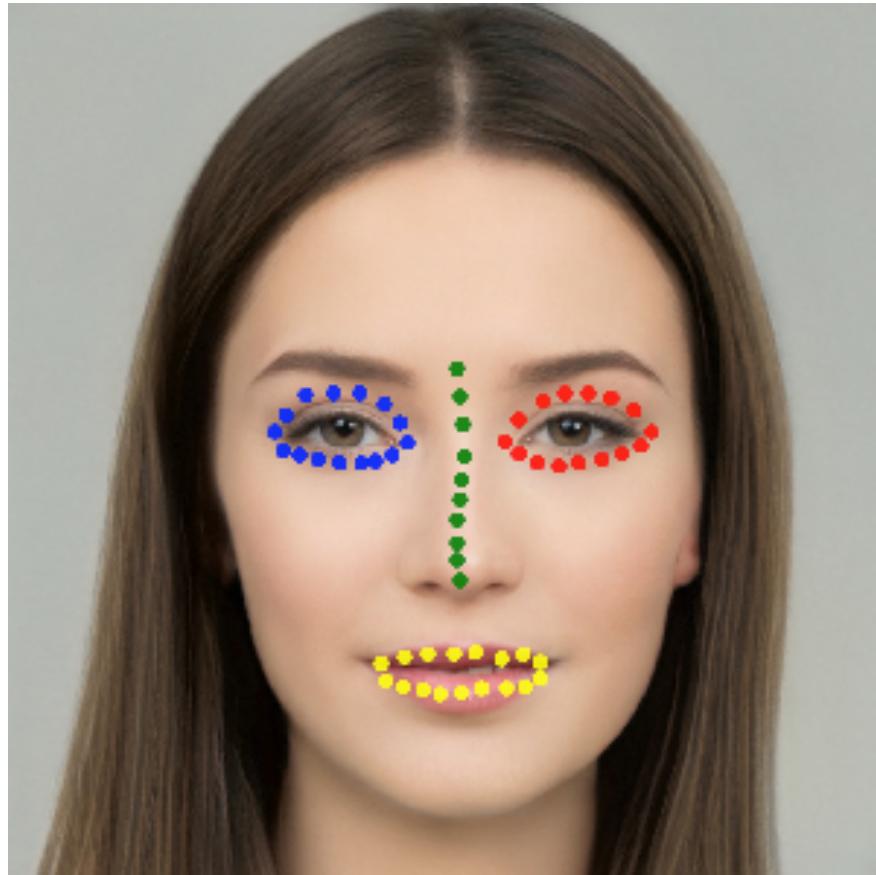
Percentage of Time Allocated to Machine Learning Project Tasks



Medical imaging



Facial analysis & Emotion recognition



E-commerce & Retail

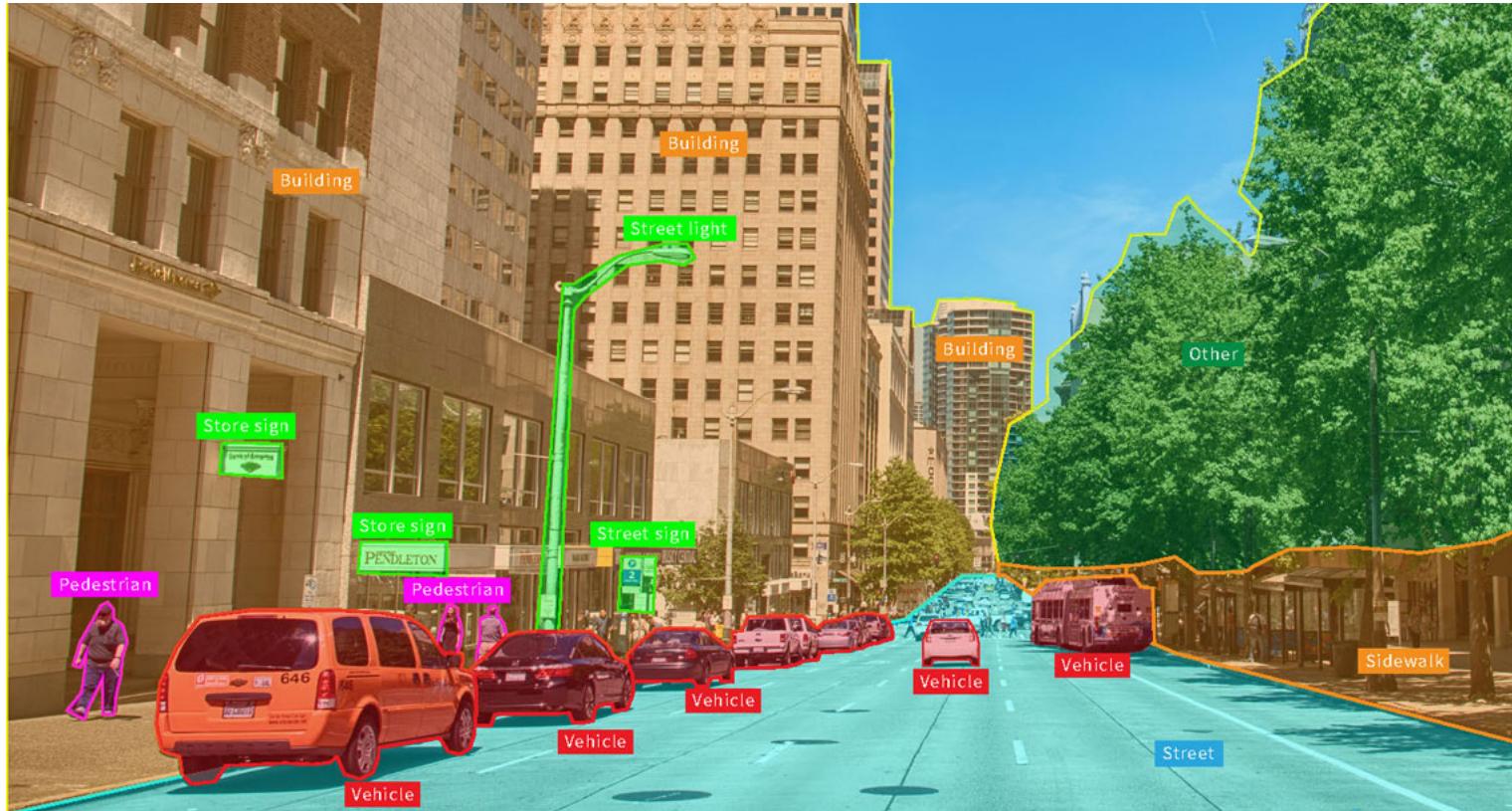


Christian Dior Lady Di
Bag Canage white lea
Women's Bag Crossbo

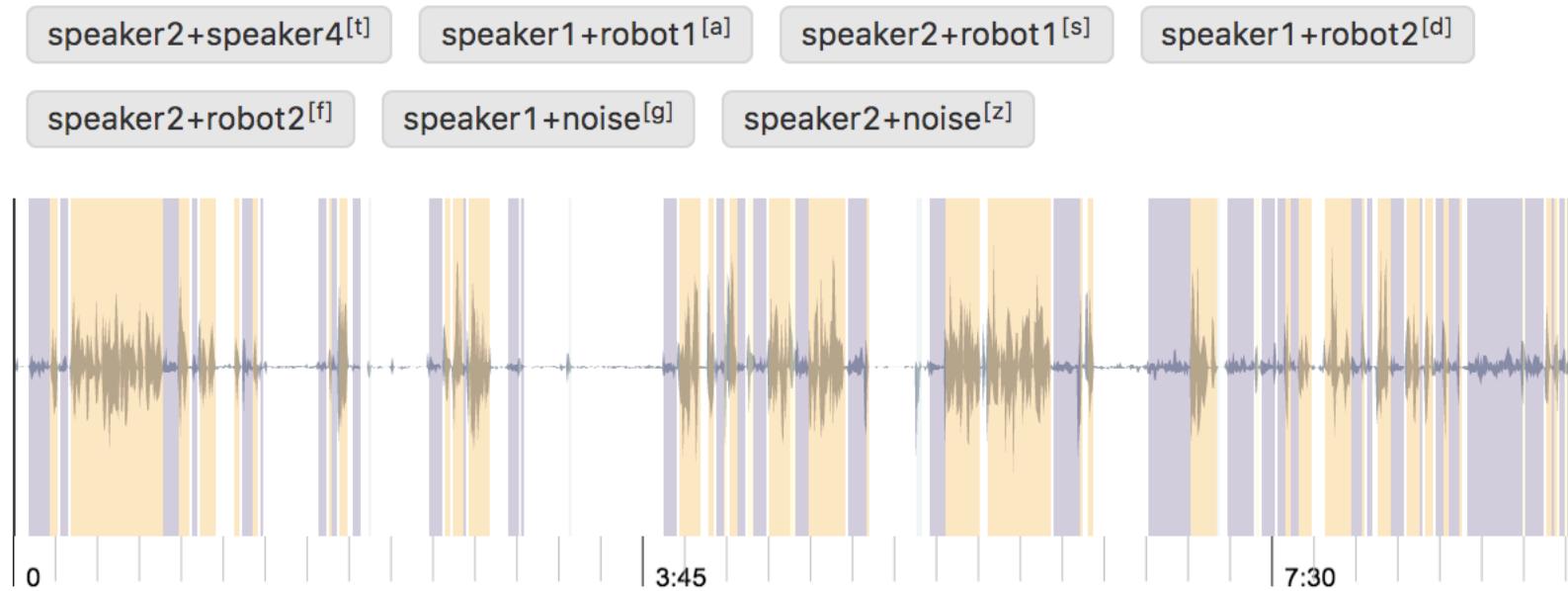
- one features fashiona pattern, another one :
- long sturdy strap in s can't be removable, m 119cm
- 10.2" wide x 8.3" tall x weight: 1.35lb

Price: \$569

Autonomous vehicles



Speech analysis & conversational analytics



Document processing

Regular Expressions (+)

Enter text to filter (x)

- Regexp_Money
 - Money (Yellow)
- Regexp_Percent
 - Percent (Red)

SMI.txt

SMIC Reports **2017** **Third Quarter** Results

All currency figures stated in this report are in US Dollars unless stated otherwise.

The consolidated financial statements are prepared in accordance with International Financial Reporting Standards ("IFRS").

SHANGHAI, Nov. 14, **2017** /PRNewswire/ -- Semiconductor Manufacturing International Corporation (NYSE: **SMI**; SEHK: 981) ("SMIC," the "Company," or "our"), one of the leading semiconductor foundries in the world, today announced its consolidated results of operations for the three months ended September 30, **2017**.

Third Quarter **2017** Highlights

Revenue was **\$769.7 million** in 3Q17, an increase of **2.5%** **QoQ** from **\$751.2 million** in 2Q17 and a decrease of **0.7%** **YoY** from **\$774.8 million** in 3Q16.

Gross profit was **\$177.3 million** in 3Q17, compared to **\$194.1 million** in 2Q17 and **\$232.1 million** in 3Q16.

Gross margin was **23.0%** in 3Q17, compared to **25.8%** in 2Q17 and **30.0%** in 3Q16.

Fourth Quarter **2017** **Guidance:**

The following statements are forward looking statements based on current expectations and involved risks and uncertainties, some of which are set forth under "Safe Harbor [Statements](#)" below.

Edit **Close**

Class (+)

Enter text to filter (x)

Check the class to display occurrence of it in the document.

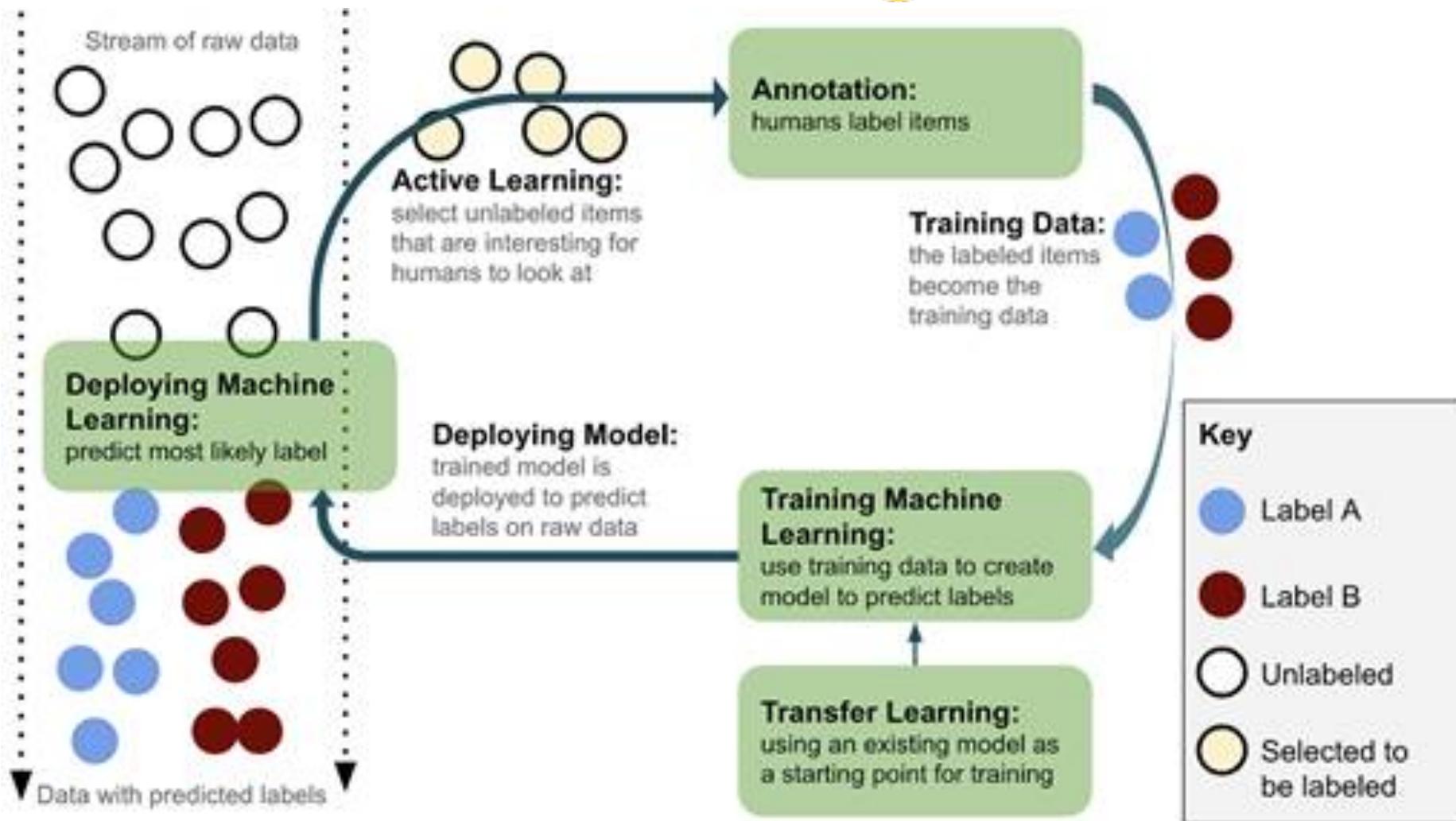
Uncheck All

- Cycle** (Orange)
- CycleRef** (Black)
- EPS** (Grey)
- FutureGuidance** (Dark Grey)
- GAAPNONGAAP** (Pink)
- GrossMargin** (Orange)
- Loss** (Cyan)
- Money** (Yellow)
- Percent** (Red)
- Revenue** (Green)
- Ticker** (Teal)
- Year** (Light Blue)

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
Crowdsourcing	Cooperation with freelancers and labeling workforce from 3 rd 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**

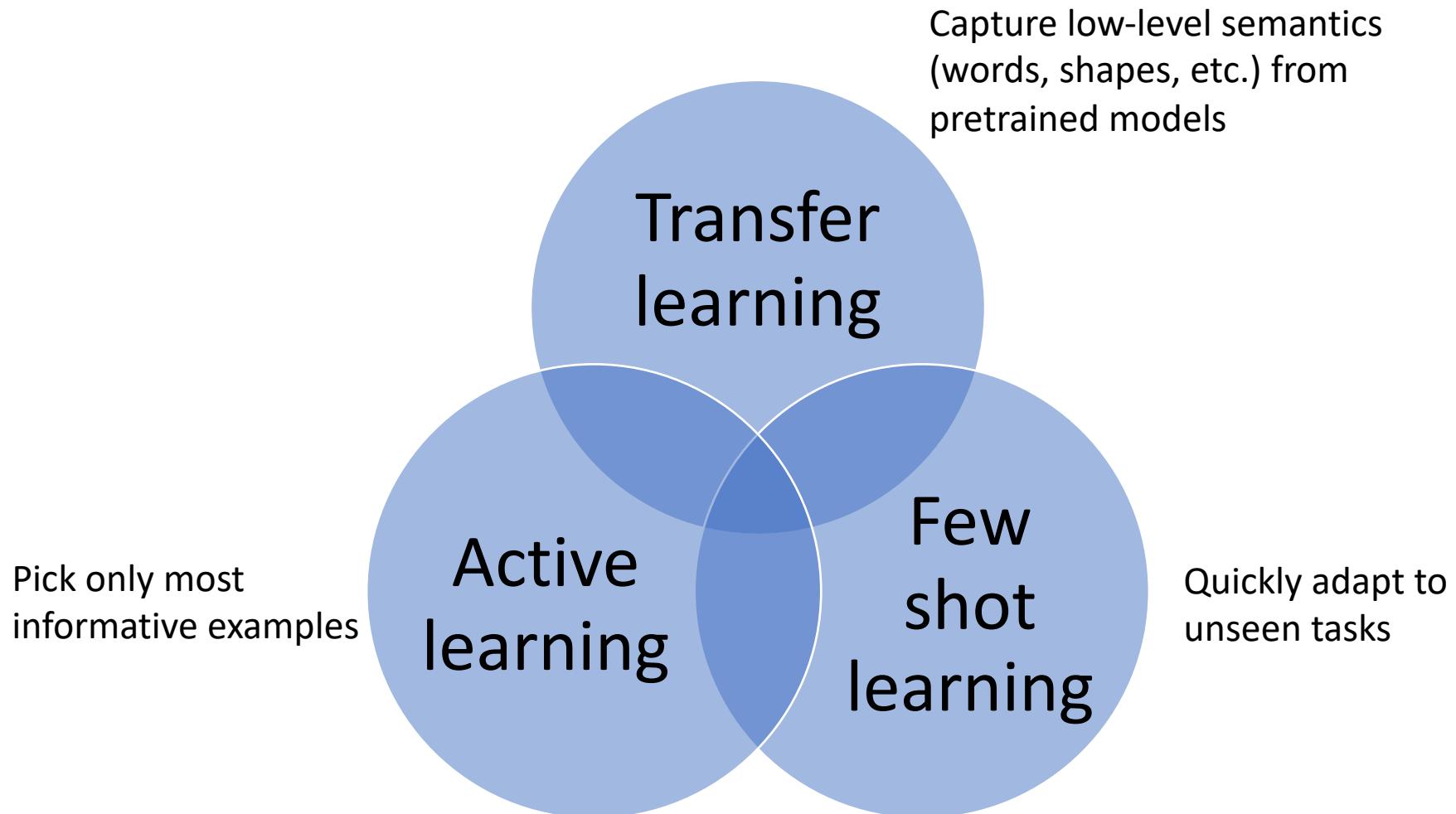
=> flexible configuration tools



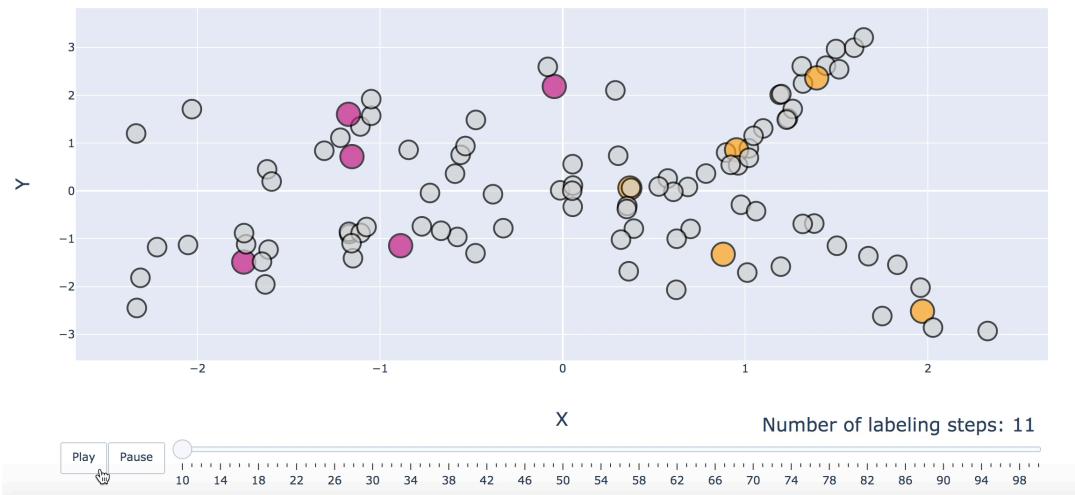
**We can't crowdsource
private data**

=> depersonalize data before
annotation

Reducing labeling cost



Learn faster with smarter data labeling



Try to pick and label samples to promote:

- Uncertainty (class borderline)
- Informational density
- Diversity

Quality control & assessment



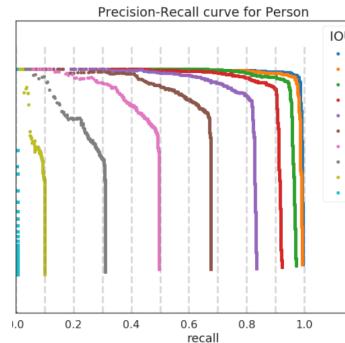
Annotator consensus:
how labelers are agreed
with each other on
specific item



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)

A dense word cloud centered around the term "Federated learning". The words are arranged in a circular pattern, with "Federated learning" at the top center. Other prominent words include "Data privacy", "GDPR", "K-anonymity", "Differential privacy", "Homomorphic encryption", and "Federated learning". Each word is surrounded by smaller, related terms in a repeating color scheme of blue, green, red, and orange.

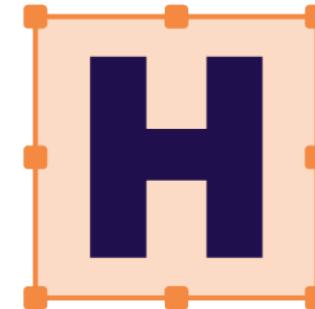
Thank you!



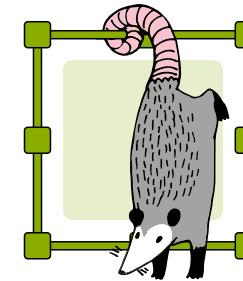
Nikolai Liubimov

nik@heartex.ai

linkedin.com/in/liubimov/



heartex.ai



labelstud.io