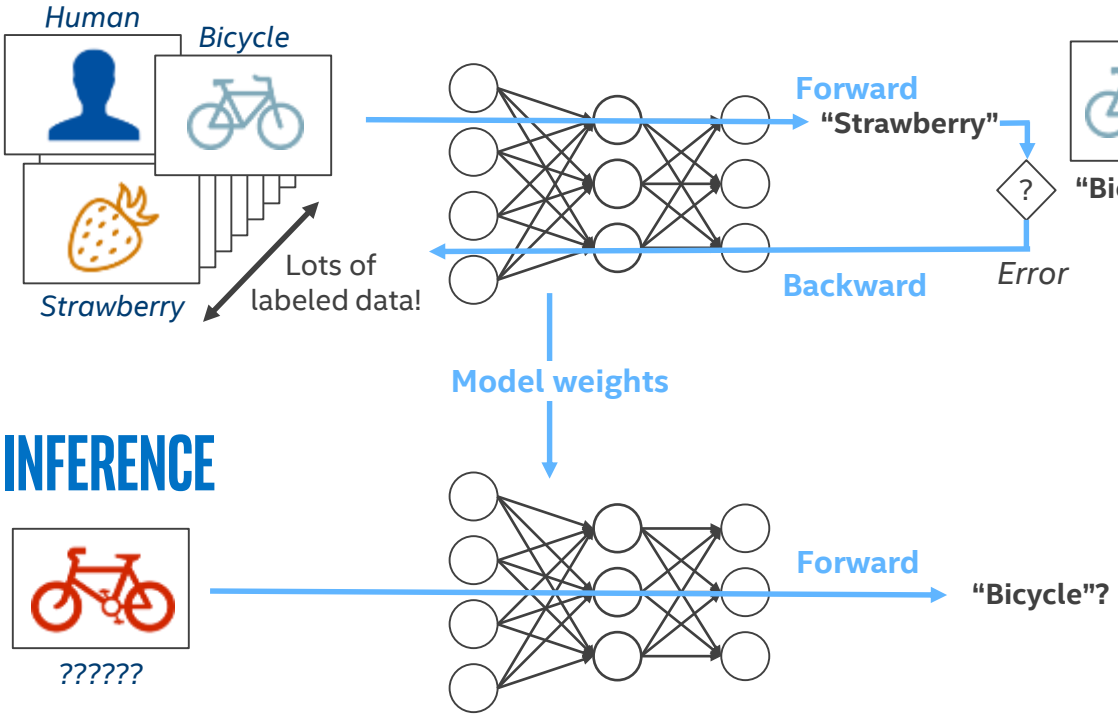
INFERENCE AT THE EDGE / AI ON PC

Outline: What students are expected to learn ?
AI on PC

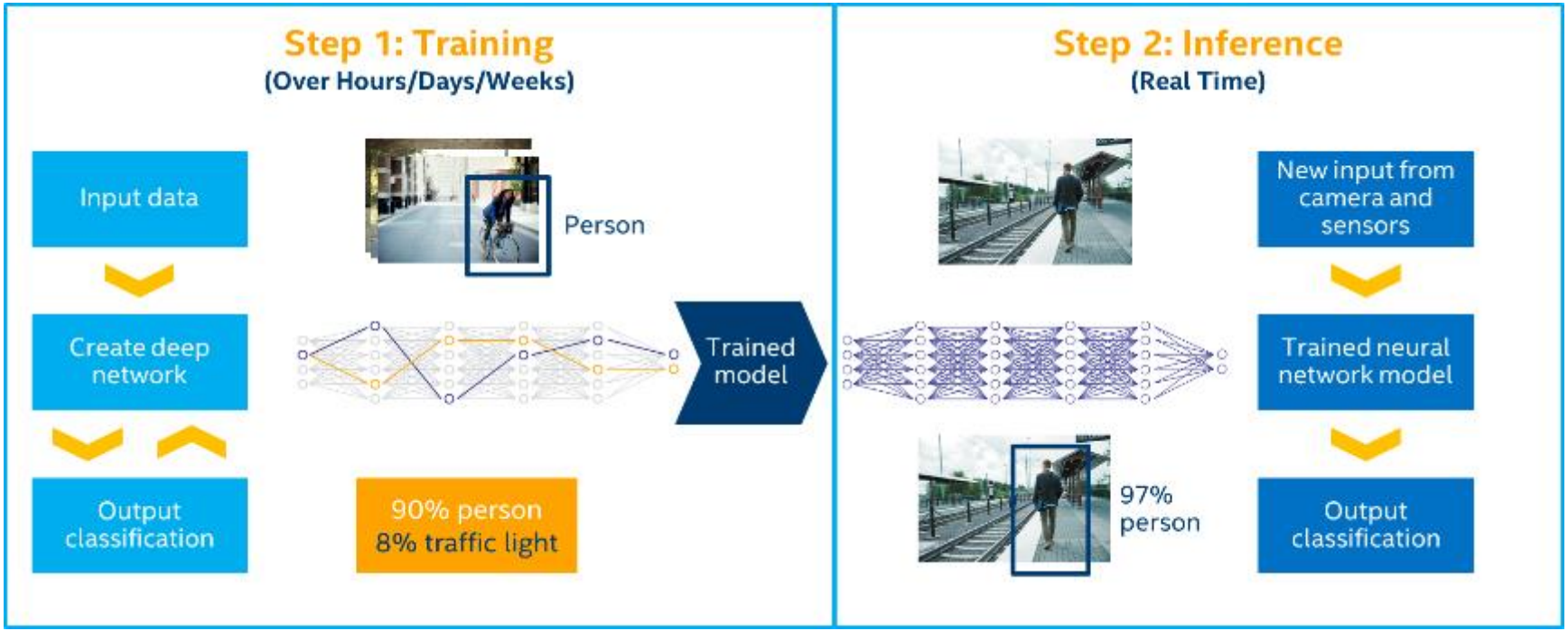
- What is inference ?
- What is Edge computing ?
- What is inference at the Edge ?
- Why inference at the Edge ?
- AI on PC Use Cases
- Summary

What is inference?



Inference is the stage in which a trained **model** is used to **infer**/predict the testing samples and comprises of a similar forward pass as training to predict the values. Unlike training, it doesn't include a backward pass to compute the error and update weights

Training and inference workflow



What is edge computing?

- Edge devices are at the periphery of a network. The data is processed locally and decisions can be made in place rather than sending it to the cloud for computation.
- An example of edge devices is PC.
- PCs as edge computing devices deliver richer human-computer through workflows emphasizing image detection, classification and tracking.

What is inference at the edge?

- Inference at the edge is the process of performing computations on custom or specialized trained AI models in systems where limitations for size, power, and real-time performance are required to ensure success.
- Inference at the edge supports real-time analytics and decision-making. One example being predictive maintenance. Another example being gaming on PC.
- Inference at the edge requires AI models that are specially tuned to the above-mentioned constraints. Models such SqueezeNet, for example, are tuned for image inferencing on PCs and embedded devices.

Why inference at the edge?

Business Imperatives and technical constraints drive demands for Inference at the edge due to:

- Requirements to overcome hurdles in managing the volumes of data, timeliness of data processing, and real-time optimization.
- In the case of usages on the PCs these constraints arise when mission-critical applications in health care, for example, require accurate, timely, scalable, and automated solutions.

Why inference at the edge (cont.)

1. **Bandwidth and Latency**

   Applications that demand near instantaneous inference can not function properly with the latency and bandwidth bottlenecks.

2. **Security and Decentralization**

   Commercial servers are prone to attacks and hacks.

3. **Job Specific Usage (Customization)**

   Each work station has a different set of objects, and requires customized inferencing. Centralized decision-making in the cloud would be prohibitive.

USE CASES OF AI ON PC

Example: Intruder Detection Solution

- Detects any number of objects entering a defined space.
- Alerts you when someone enters your predefined restricted area.

For additional details please go to https://software.intel.com/en-us/iot/reference-implementations/intruder-detector

Example: real-time people counter

- Real-time people counter on the PC
- Smart video applications using models and inference to run single-class object detection

For additional details please go to https://software.intel.com/en-us/iot/reference-implementations/people-counter-system

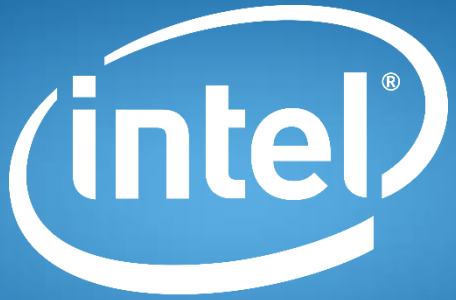
SUMMARY

Summary

- **Inference** refers to the process of inferring things about the world by applying your model to new data. In the context of machine learning refers to the process of taking a model that's already been trained and using that trained model to make useful predictions.
- **Edge computing** devices include PCs, etc.
- **Inference at the Edge** refers to the process of pushing inference models to the edge devices and perform such computations locally, timely and independent of access to network or cloud resources.
- Form more info on Intel AI on PC, check the links below:

  https://software.intel.com/en-us/ai-academy/ai-on-pc and

  https://devmesh.intel.com/



experience
what's inside™