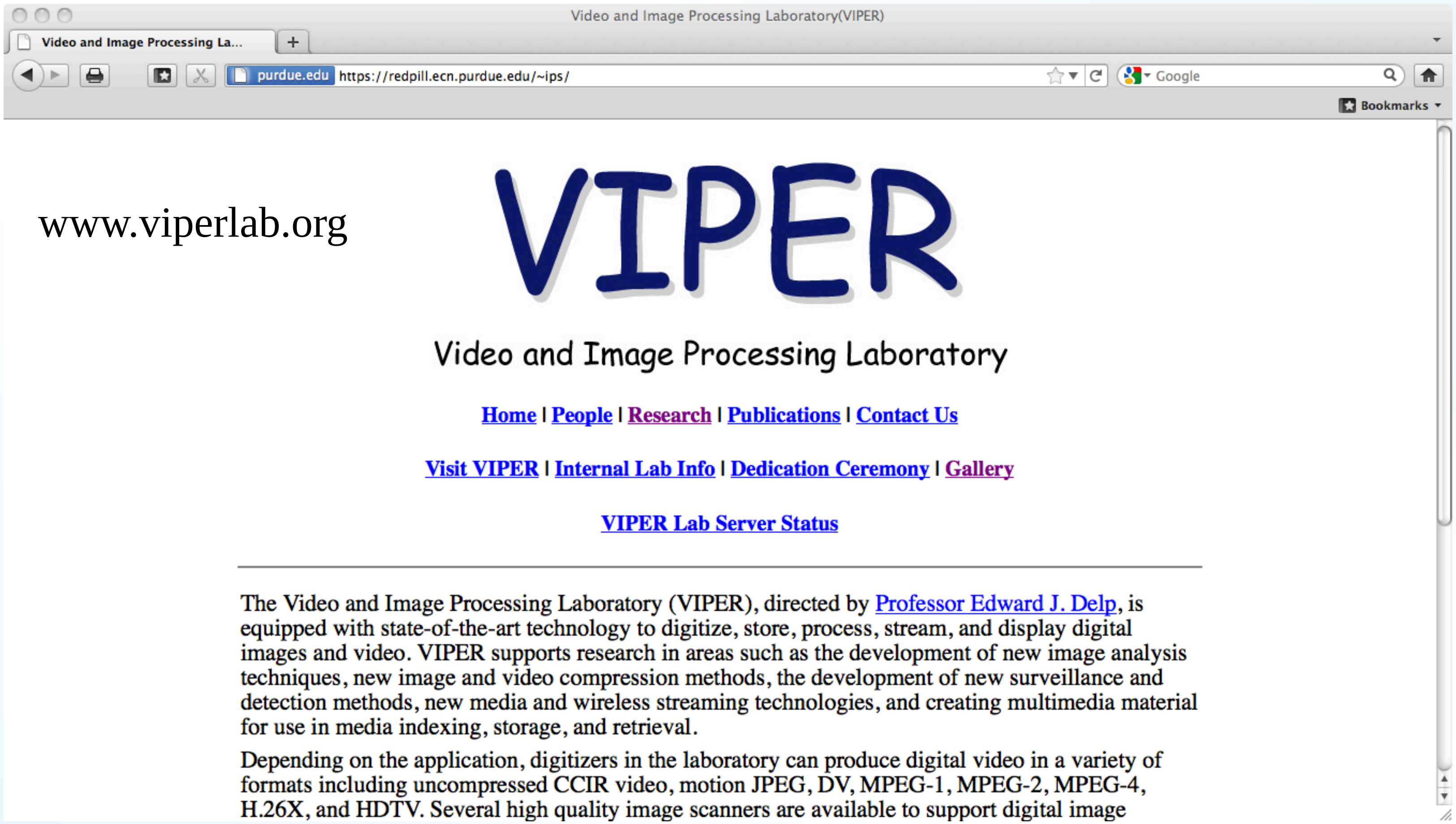
Cheap, Fast, and Low Power Deep Learning: I need it now!
(Please and Thank You)

Edward J. Delp

Video and Image Processing Laboratory (*VIPER*)
School of Electrical and Computer Engineering
School of Biomedical Engineering
Purdue University
West Lafayette, Indiana

email: ace@ecn.purdue.edu
https://engineering.purdue.edu/~ace/

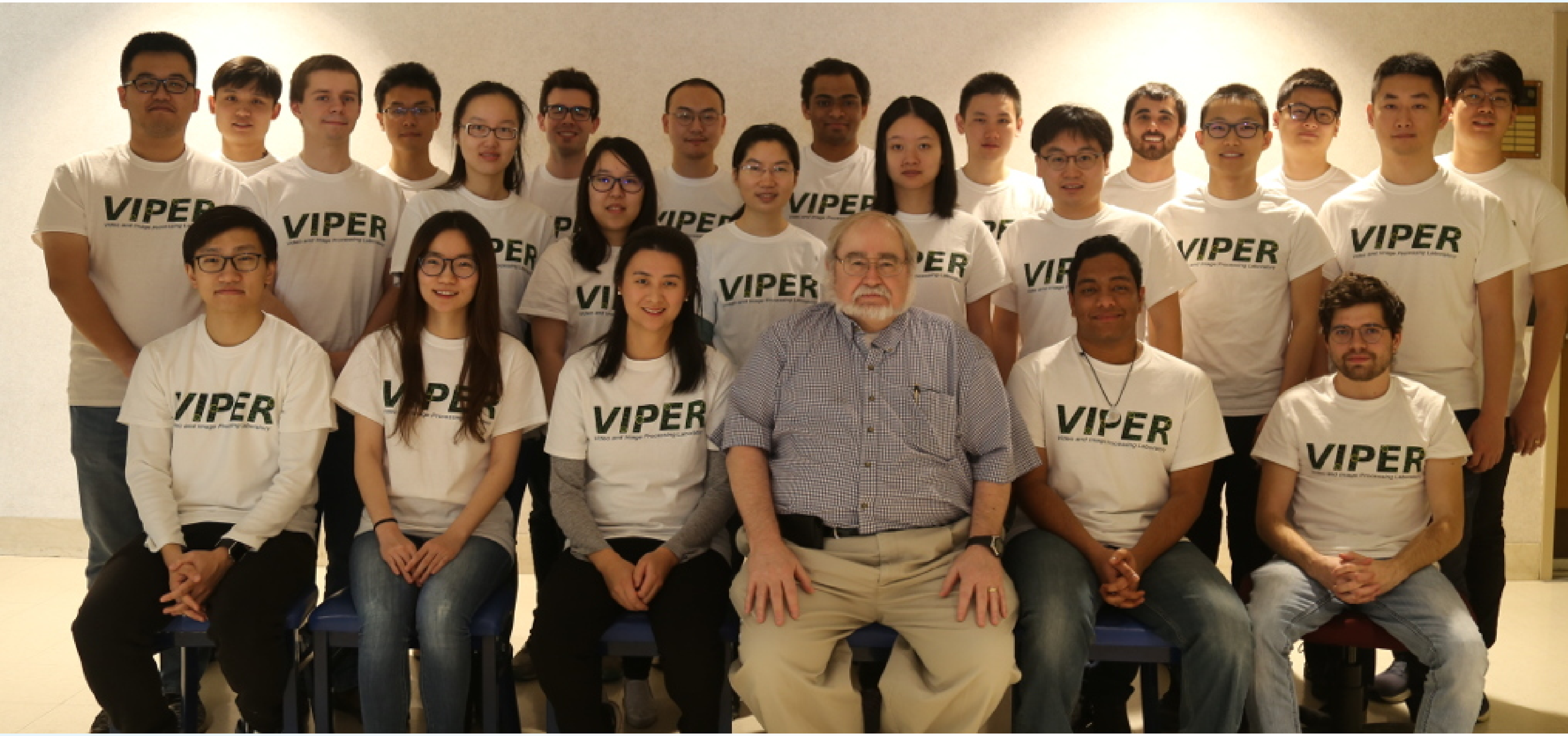
Video and Image Processing Laboratory (*VIPER*)



Acknowledgements

VIPER

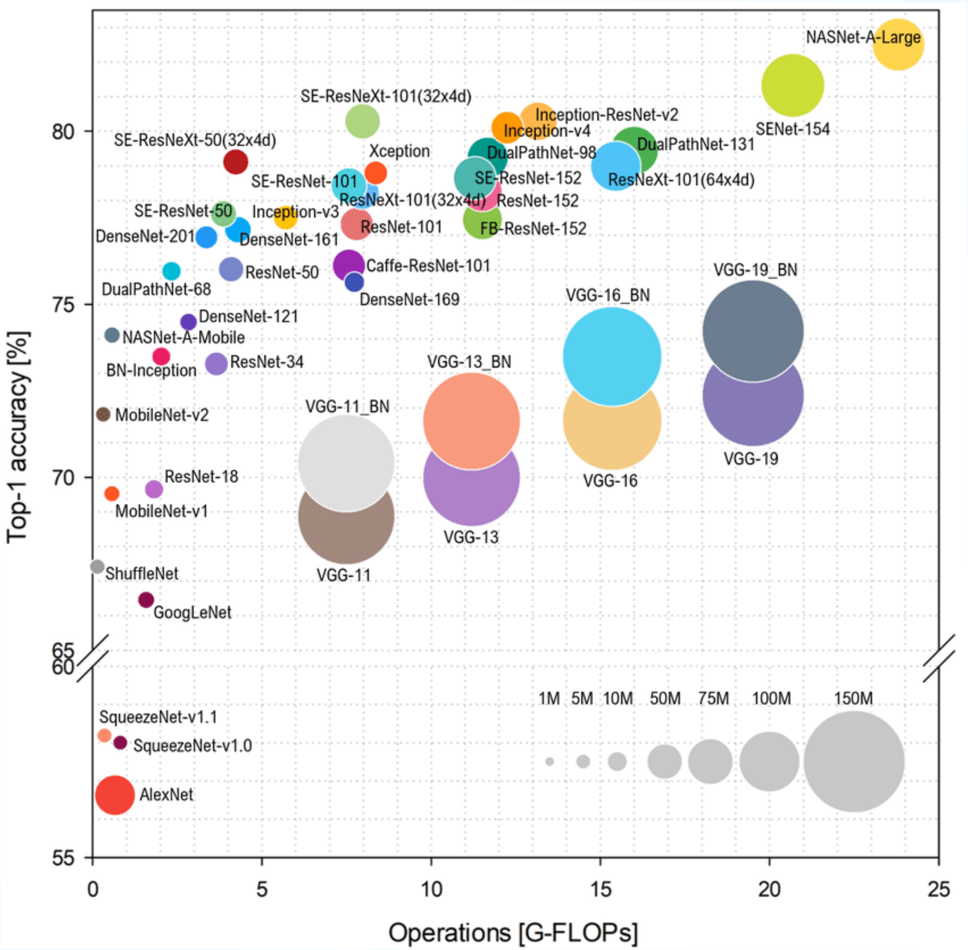
Video and Image Processing Laboratory



Outline

- **Introduction of Low Power Deep Learning**
- **Applications That Need Low Power Deep Learning**
  - **Health care monitoring**
  - **Biomedical image analysis**
  - **Image Based Phenotyping**

Deep Learning “Requirements”



Reference: S. Bianco, et al., Benchmark Analysis of Representative Deep Neural Network Architectures, IEEE Access, 2018

Low Power Deep Learning

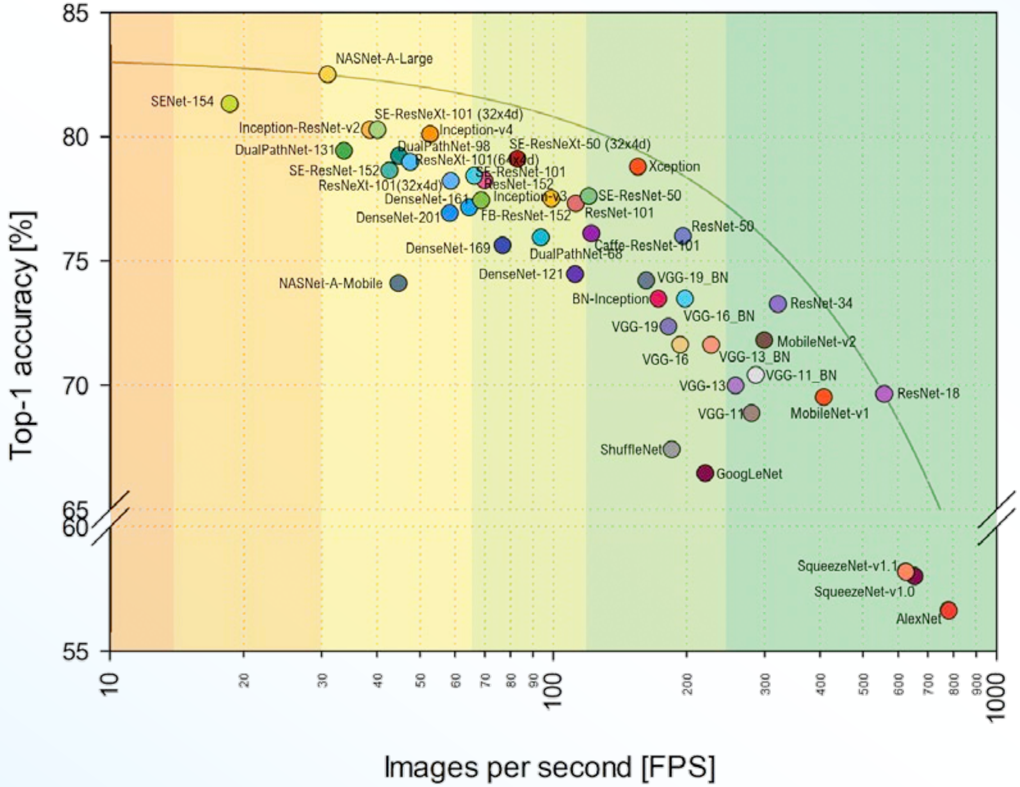
- **Current deep learning models are both power and memory intensive**

- **Need more analysis of computational cost (memory usage, inference time)**

After training...

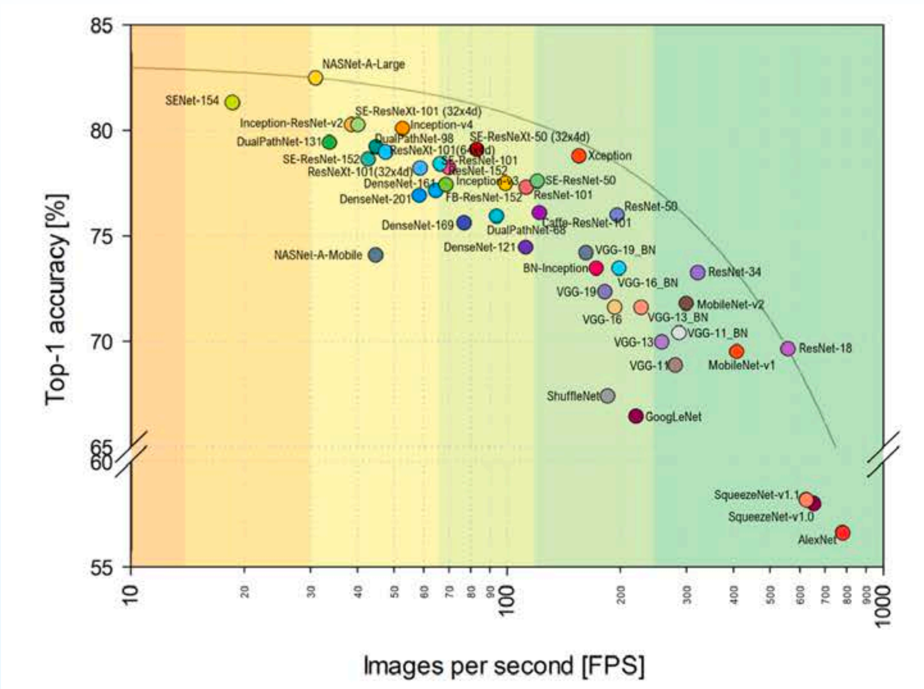
Images Per Second vs. Accuracy

- **Less computation hurts the performance**

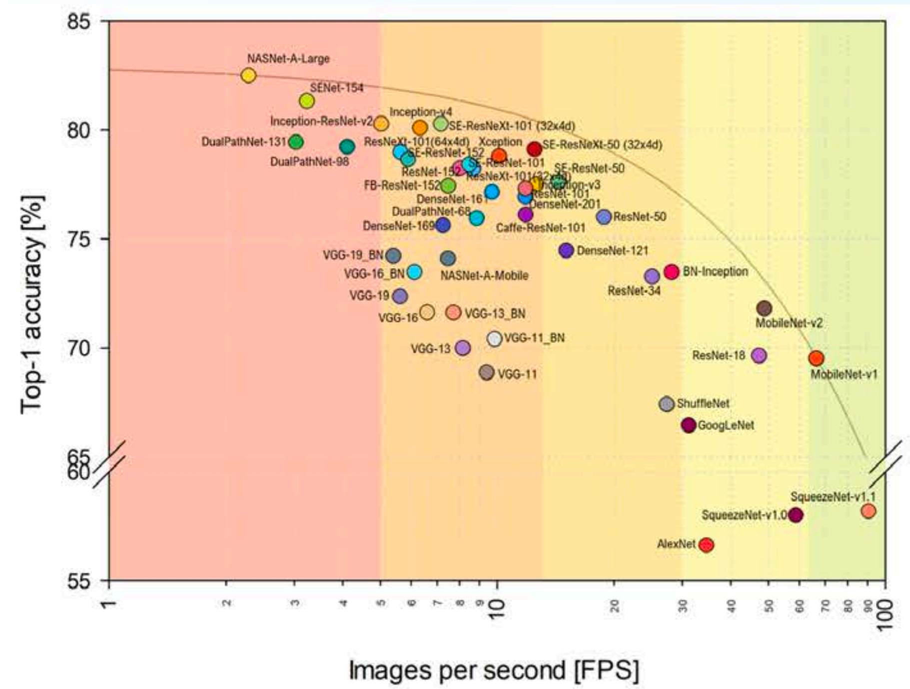


Reference: S. Bianco, et al., Benchmark Analysis of Representative Deep Neural Network Architectures, IEEE Access, 2018

Inference Speed on Embedded System



Nvidia Titan X with 3840 cores
(Power consumption is approximately 250w)

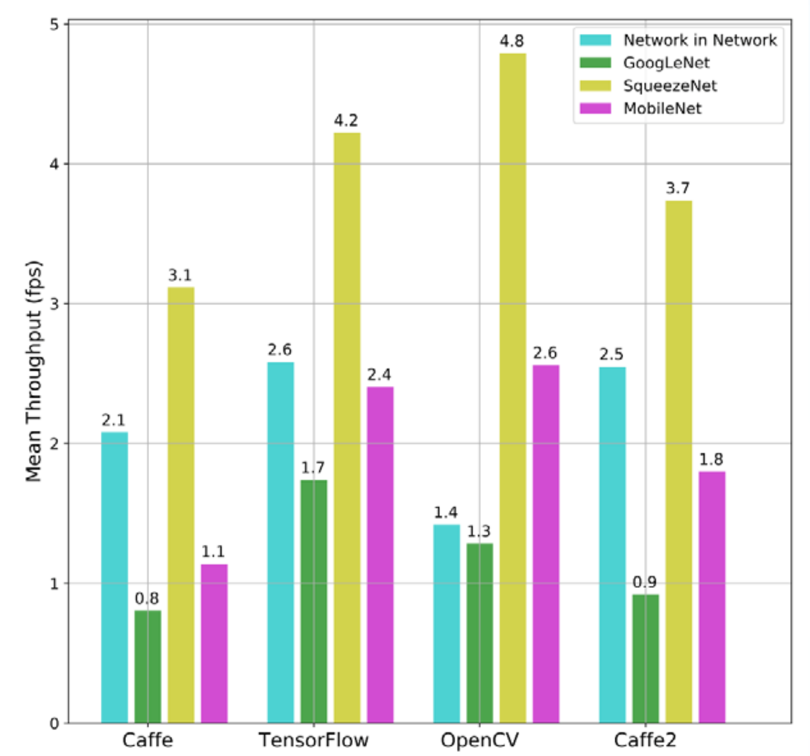


Nvidia Jetson TX1 board with 256 cores
(Power consumption is approximately 10w)

Reference: S. Bianco, et al., Benchmark Analysis of Representative Deep Neural Network Architectures, IEEE Access, 2018

Inference Speed on Non-GPU Device

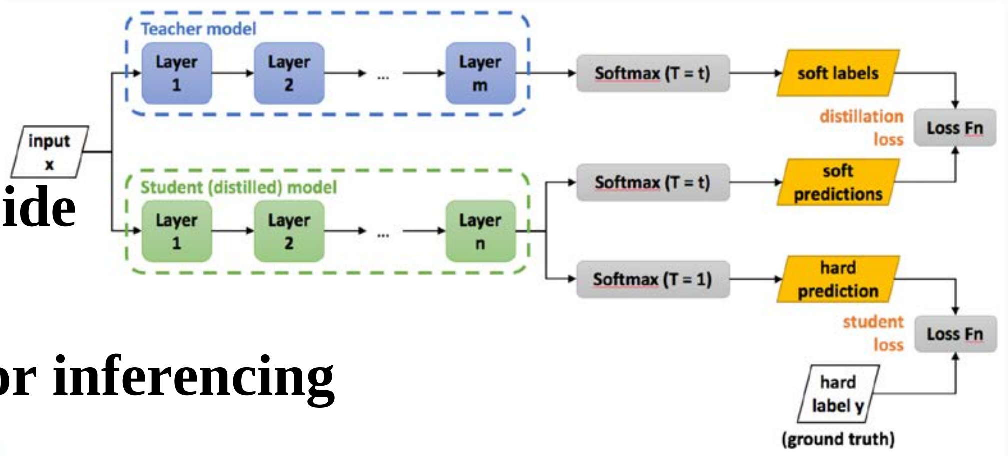
- **Frame per second on Raspberry Pi**
  - **Power consumption is approximately 6w**



Reference: D. Velasco-Montero, et al., Performance analysis of real-time DNN inference on Raspberry Pi, SPIE Real-Time Image and Video Processing 2018

Methods for Reducing the Complexity of Deep Learning Models

- **Pruning**
  - Weight pruning: make the model weights sparse
  - Structure pruning: remove the filter directly
- **Knowledge Distillation**
  - Use teacher model to guide smaller student model
  - Student model is used for inferencing



- **Quantization**
  - Use less precision for model weights
- **Adversarial attacks and brittleness?**

Applications That Need Low Power Deep Learning

- **Health care monitoring**
- **Biomedical image analysis**
- **Image Based Phenotyping**

TADA

technology assisted
dietary assessment

www.tadaproject.org

Health Care Monitoring

- **In 2015, the world spent $7.7 trillion on healthcare**
- **6 out of 10 leading causes of death in US are related to diet (e.g., cancer, diabetes)**
- **Understanding the dietary patterns behind these causes is of great importance**

Technology Assisted Dietary Assessment (TADA)

- **Traditional methods of tracking diet are inaccurate and labor-intensive**
  - **Consists of self-reporting and record keeping**
- **In recent years, researchers have leveraged mobile phones and the Internet to collect images of food**
- **These images are analyzed to extract nutritional information to monitor a person's diet**

Meal Image

Locate & Identify →

Peach

Ketchup

Coke

Milk

Hamburger

French Fries

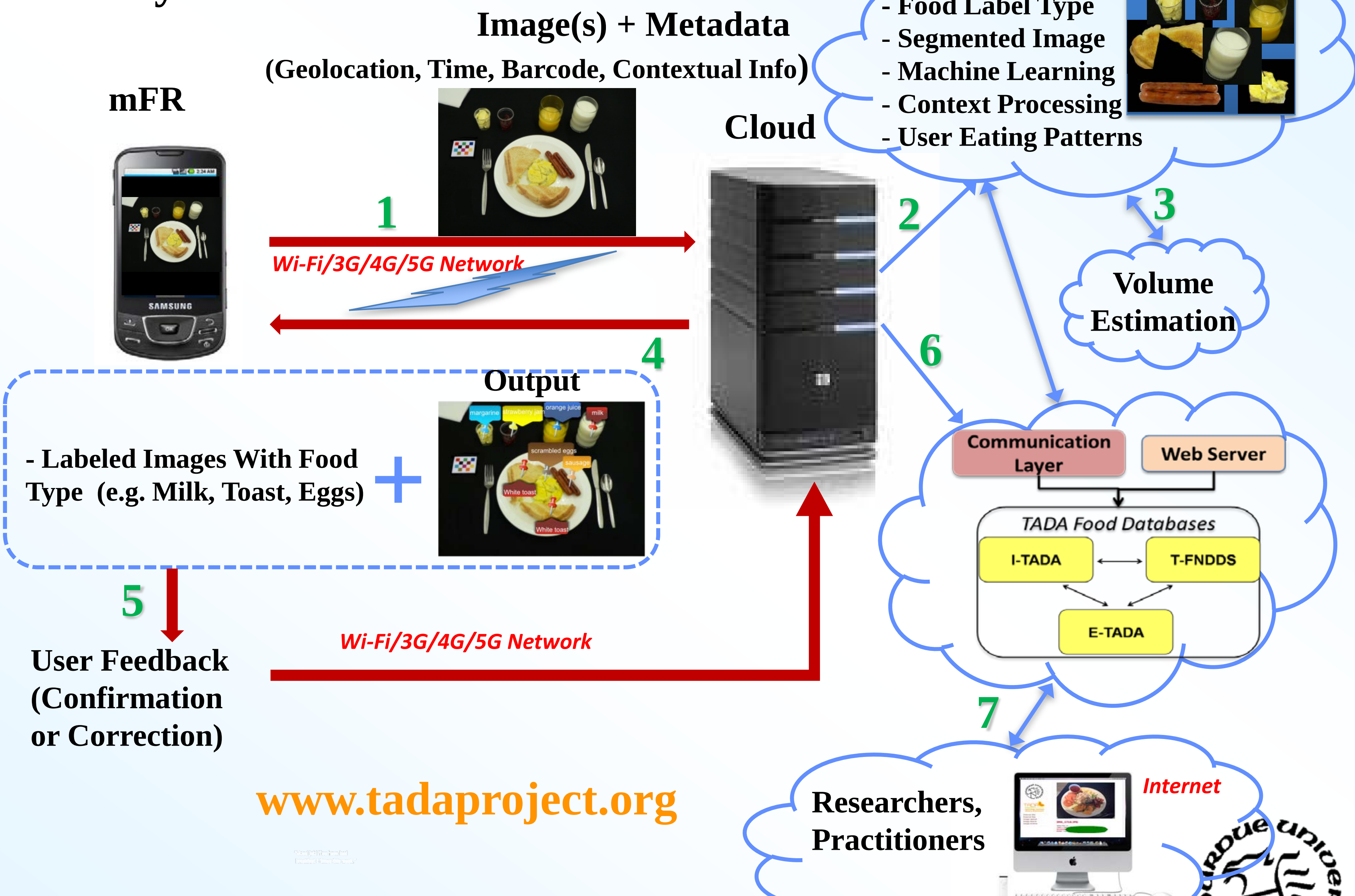
Sugar Cookie

User Studies

- **We have completed a total of more than 50 user studies**
  - **Free-living environment**
  - **More than 5000 participants**
  - **More than 400,000 images acquired**

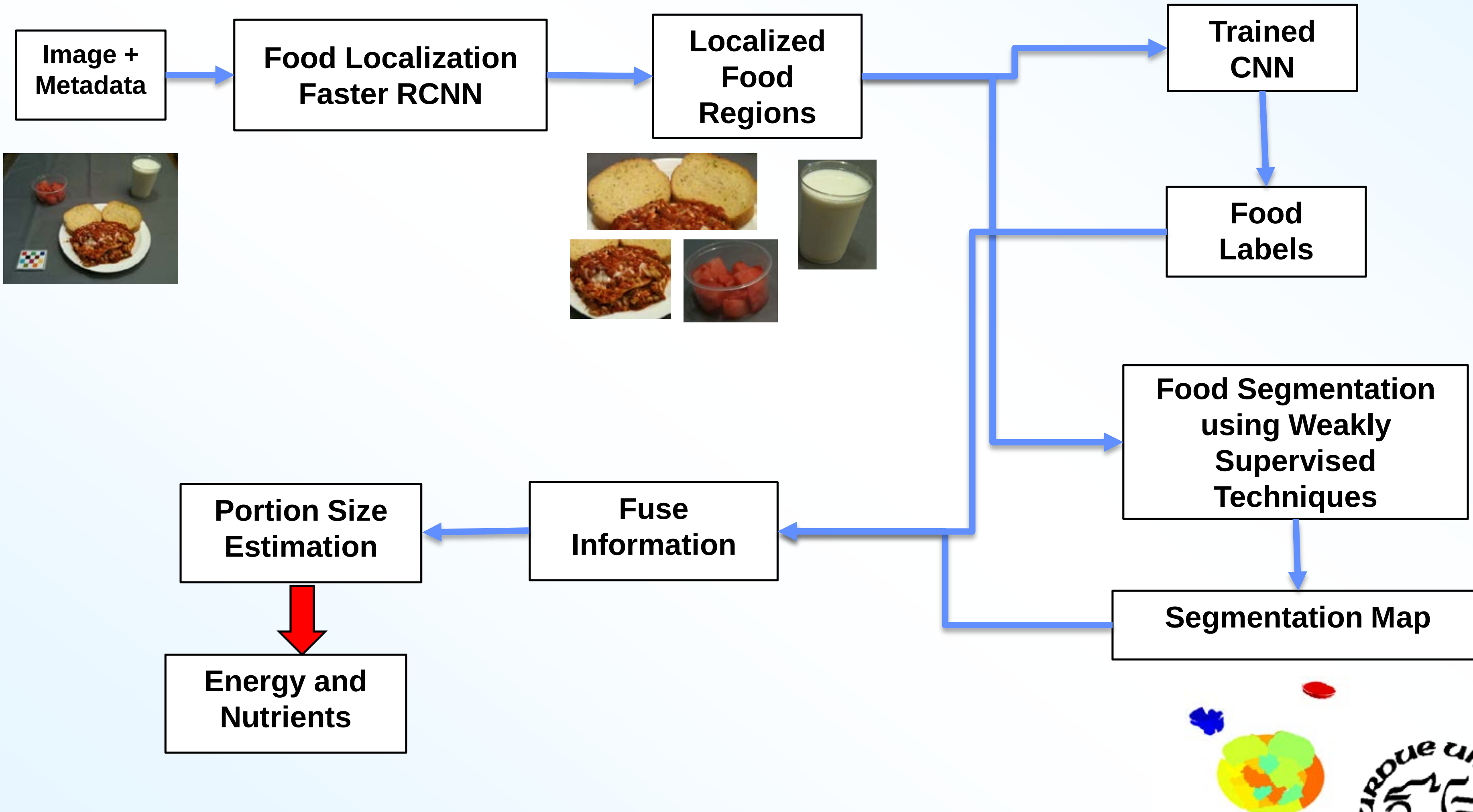
- **Each food image captures a real eating scene consists of multiple food items**

TADA System Overview

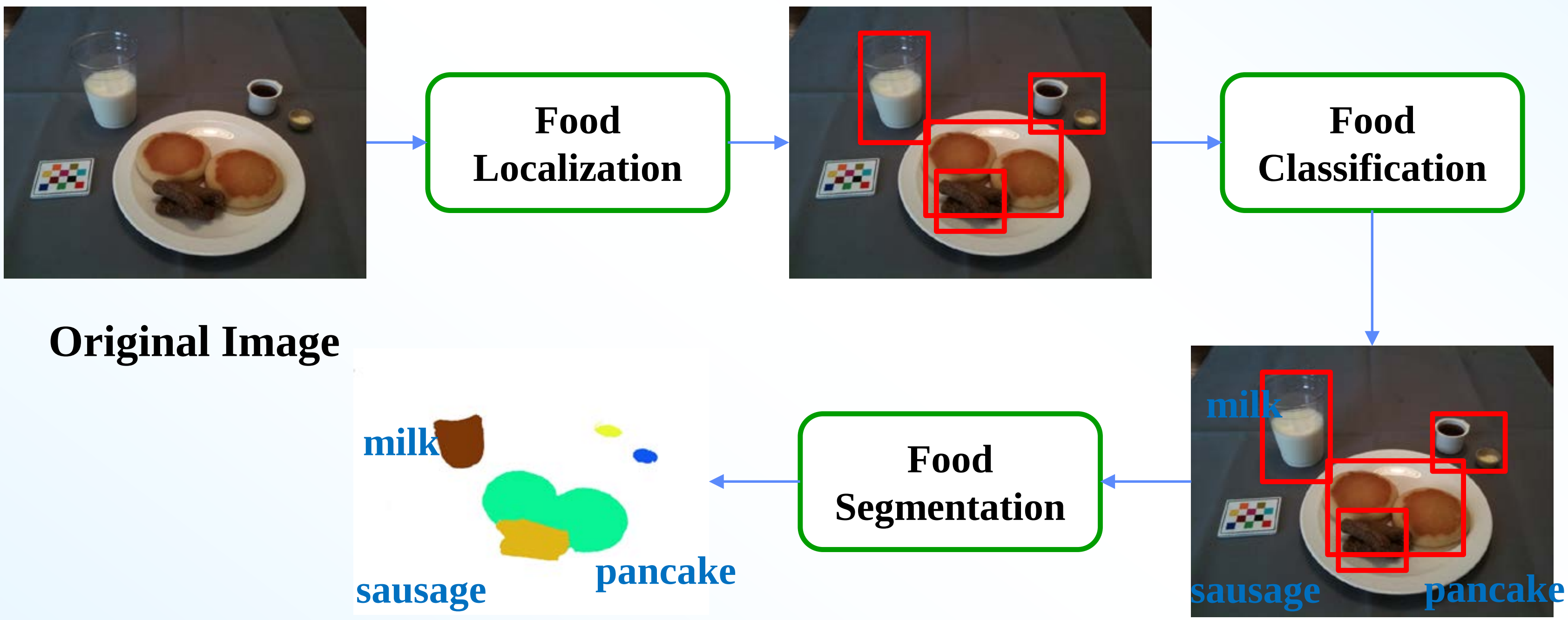


www.tadaproject.org

TADA Image Analysis System



TADA Image Analysis
Deep Learning Approach



Need for Low Power Computing

- **Our system is built on a cloud-based approach**
- **The relevant processing happens on a remote machine and real-time feedback to the user is difficult**
- **If a similarly performing system could be operated on commonly used mobile phones, the user could take more direct control of their diet and how it interacts with other factors**

Image Based Phenotyping

Definitions

- **Phenotype:**
  "any measurable characteristic or trait of a plant, and a result of combination of genes, environmental influence, and their interactions" or

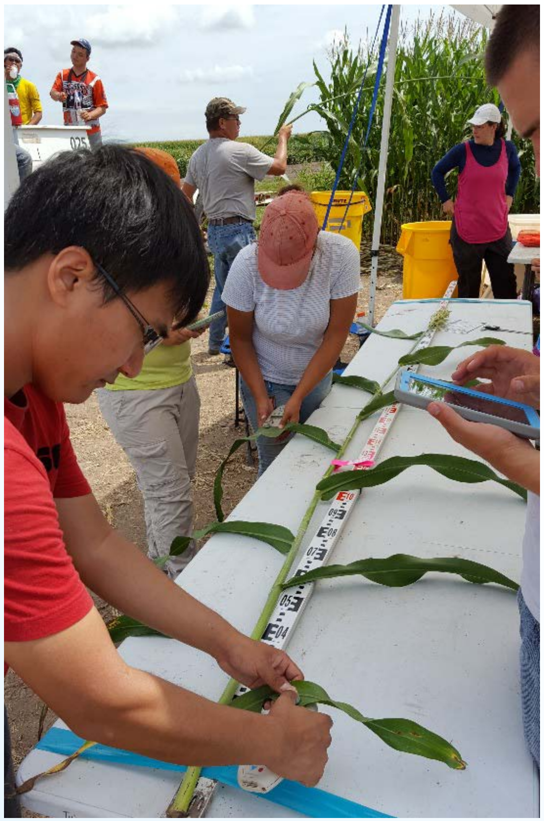
  "quantitative description of the plant's anatomical, ontogenetical, physiological and biochemical properties"

- **Phenotyping:**
  "characterizing the performance of the plants for desired trait(s)"

Traits

- **Some traits measured when phenotyping:**
  - **Color**
  - **Height**
  - **Architecture (shape)**
  - **Canopy temperature**
  - **Canopy aperture**
  - **Water/nitrogen use efficiency**
  - **Number of leafs**
  - **Total leaf area**
  - **Grain yield**
  - **Fluorescence intensity**

Traditional and Modern Phenotyping



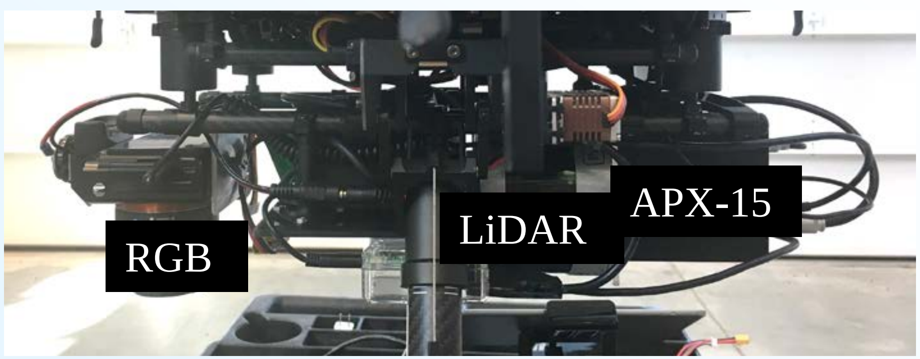
Traditional plant phenotyping

Automated phenotyping in the field

UAV imagery data

Sensor Platforms

- **Images are acquired from drones and the Phenorover**



Phenosorg RGB Data Processing Flow Chart

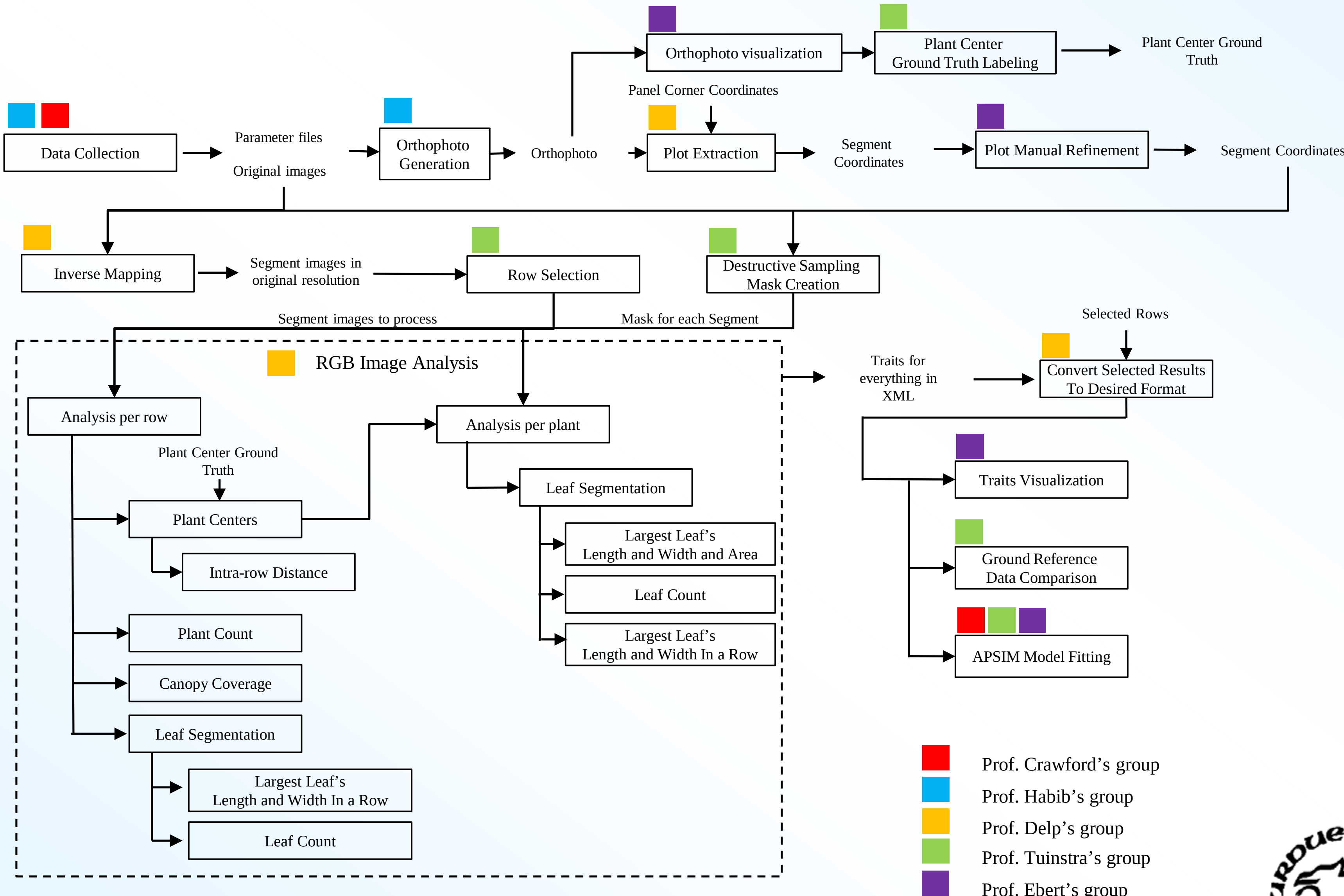
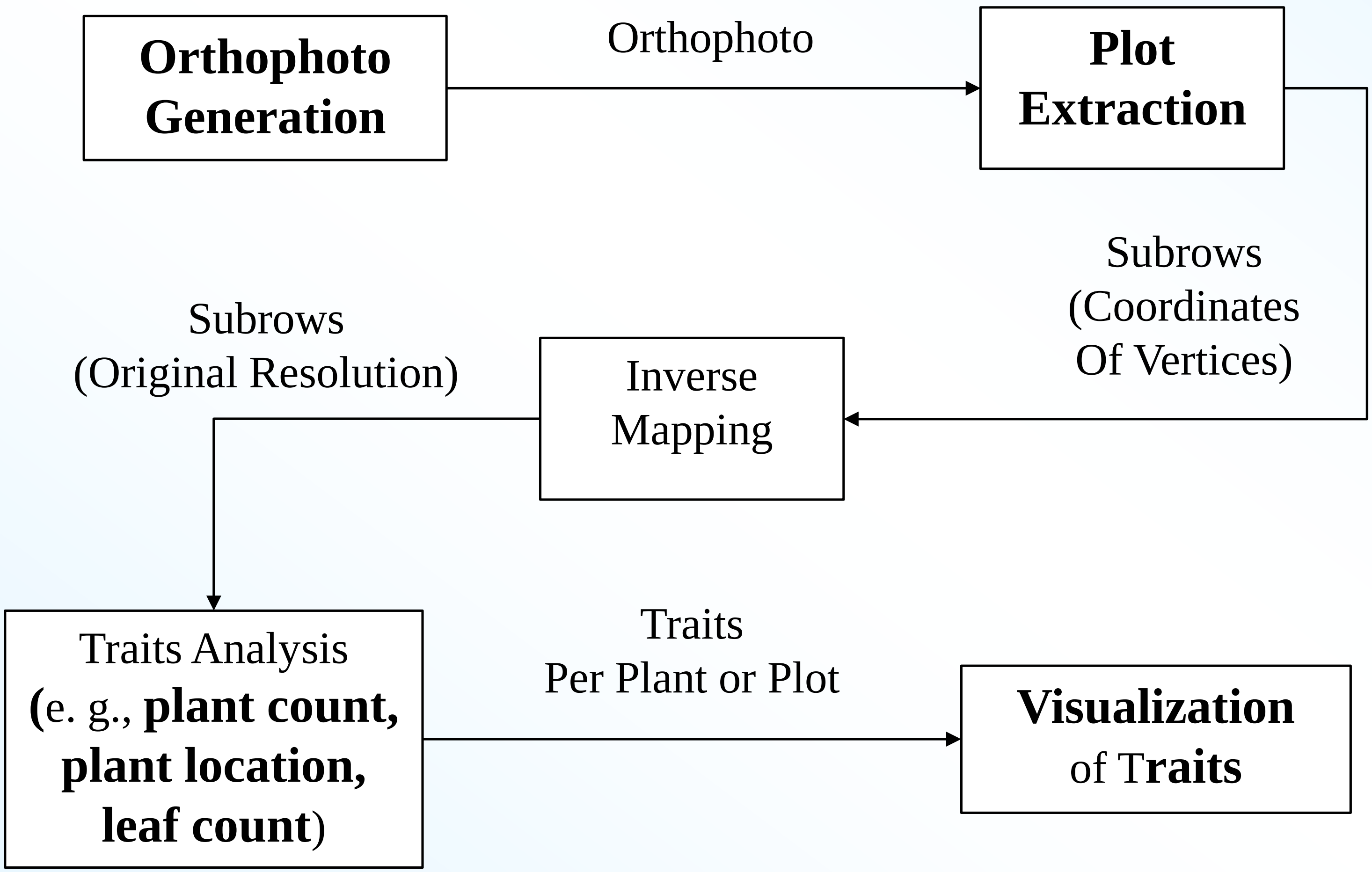
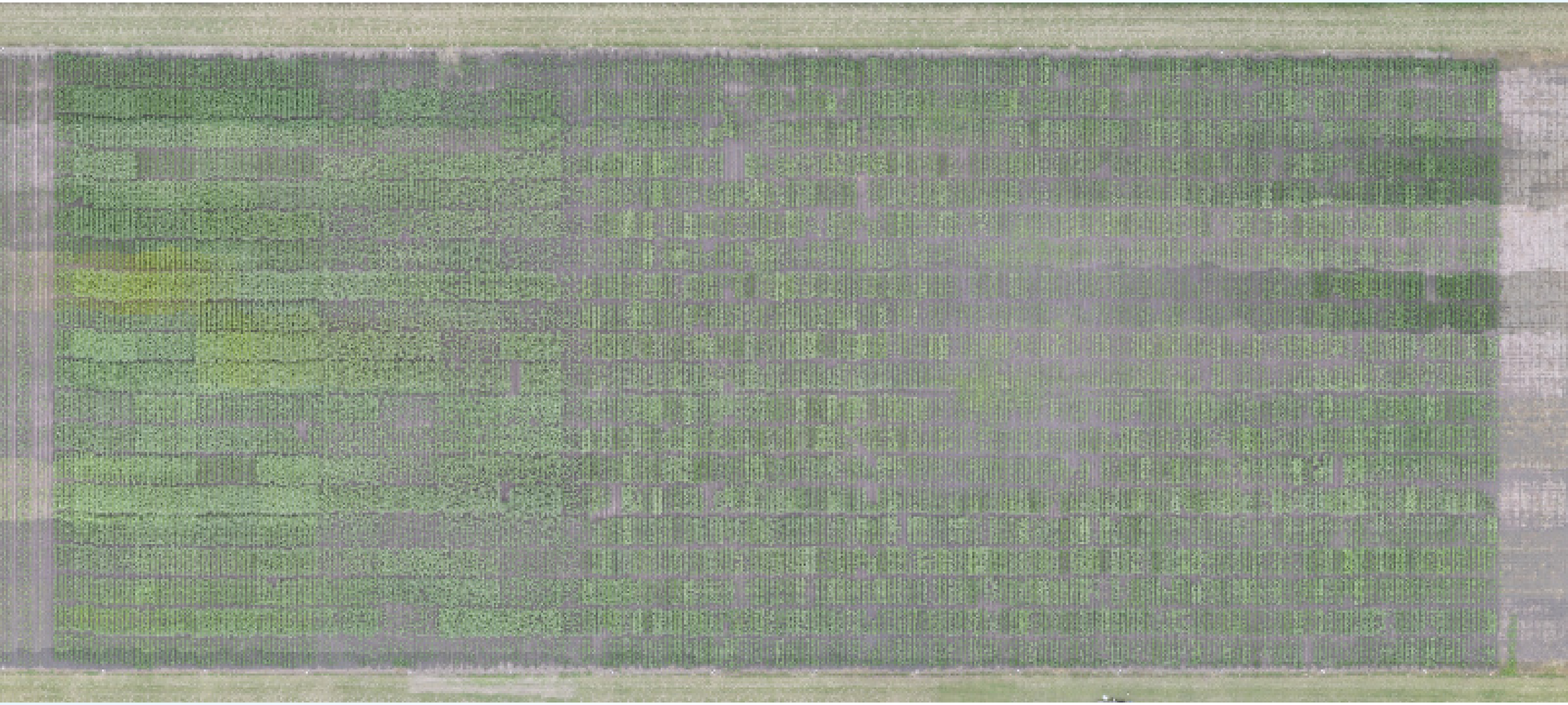


Image Analysis System



Orthophoto

07/15/2015



Example of a fully rectified image of the entire field (orthomosaic)

Counting Plants
With Deep Learning

Dataset

1,240 cigars ⇒ 1,240 images

06/21/2016

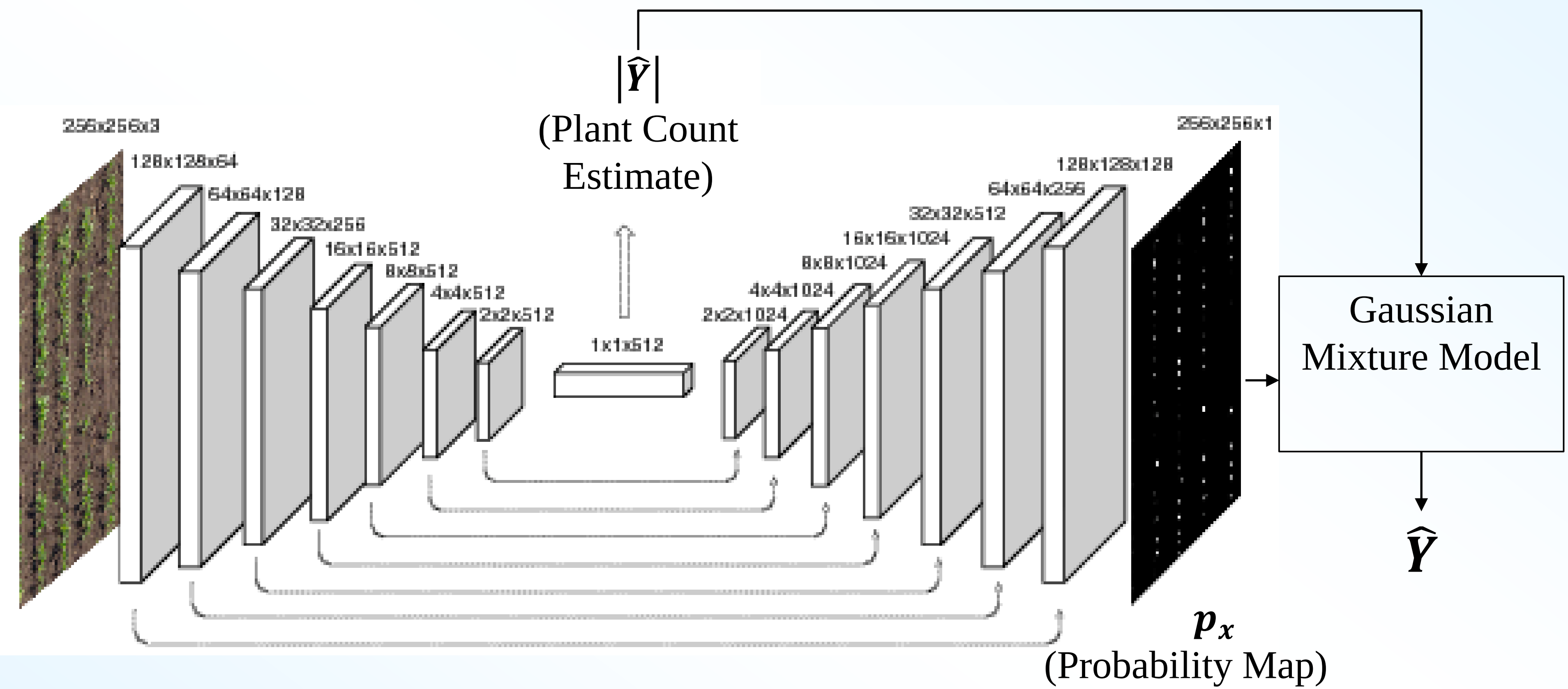
- **We extract one-row plots with our plot extraction tool**

Plant Location Using Fully Convolutional Networks

- A Fully Convolutional Network (FCN) is any network that only contains convolutional layers
- This means the output of the FCN is another image

- The architecture of a FCN can be designed such that the output is of the same size as the input
- This is used in U-Net, a popular FCN architecture, to perform pixelwise segmentation

FCN Architecture



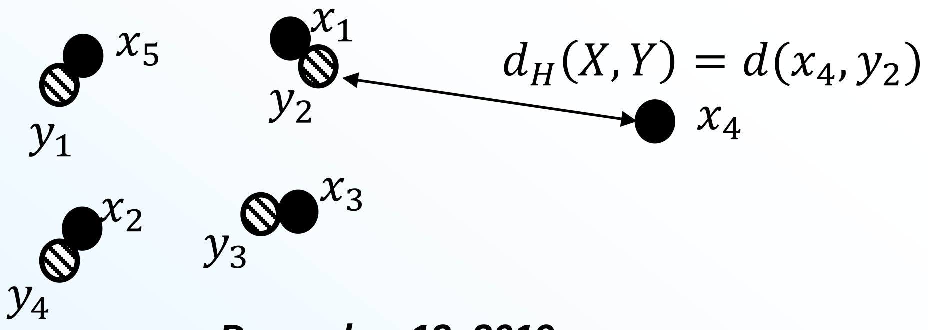
- **It accepts images of any size (bigger than 256x256)**

Cost Function For Plant Location

- **The l_1 norm does not work well for localization tasks**

- **Let $X = \{x_1, \dots, x_{|\hat{Y}|}\}$ be the set of ground truth points, and let $Y = \{y_1, \dots, y_{|Y|}\}$ be the set of estimated locations**
- **A metric that measures the similarity between X and Y is the Hausdorff Distance:**

$$\boldsymbol{d_H}\,(\boldsymbol{X},\boldsymbol{Y}) = \max\{\max_{x\in X}\ \min_{y\in Y} d(x,y)\,, \max_{y\in Y} \min_{x\in X} d(x,y)\}$$



Plant Location Cost Function

- **In fact, our cost metric is a little more complicated to deemphasize outliers and make it differentiable with respect to the output of the FCN, $\boldsymbol{p} \in [0, 1]$:**

$$d_{WH}\,(p, Y) = \frac{1}{\sum_{x\in\Omega} p_x} \sum_{x\in\Omega} p_x \min_{y\in Y} d(x, y) +$$

$$\frac{1}{|Y|} \sum_{x\in\Omega} \min_{x\in\Omega} \frac{d(x,y)}{p_x^{\alpha}}$$

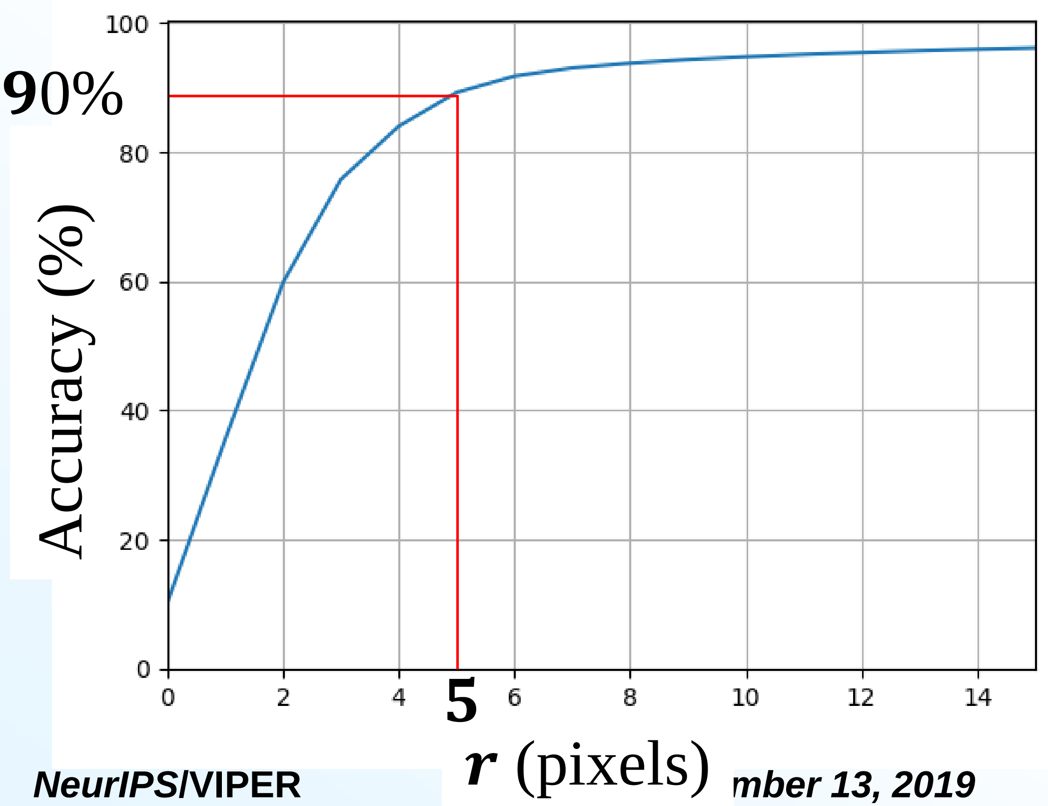
- **The parameter α balances precision and recall**
- **This is similar to the Average Hausdorff Distance**

$$d_{AH}\,(X, Y) = \frac{1}{|X|} \sum_{x\in X} \min_{y\in Y} d(x, y) + \frac{1}{|Y|} \sum_{y\in Y} \min_{x\in X} d(x, y)$$

Results (Metrics)

- **We report the following metrics in 256 × 256 images**
  - **Accuracy (% of estimations at ≤ r pixels to a plant)**
  - **Average Hausdorff Distance**
  - **Mean Average Percent Error**

$$\text{MAPE} = 100\frac{\overline{|\hat{x} - C|}}{C}$$



Metric	
AHD	8.8 px
MAPE	4 %

Regression Term

- **We modified the regression term in our cost function**

$$d_{WHD}(p,Y) = \frac{1}{\sum_{x\in\Omega} p_x} \sum_{x\in\Omega} p_x \min_{y\in Y} d(x,y) + \frac{1}{|Y|} \sum_{x\in\Omega} \min_{x\in\Omega} \frac{d(x,y)}{p_x^{\alpha}} + L_{reg}(|\hat{Y}| - |Y|)$$

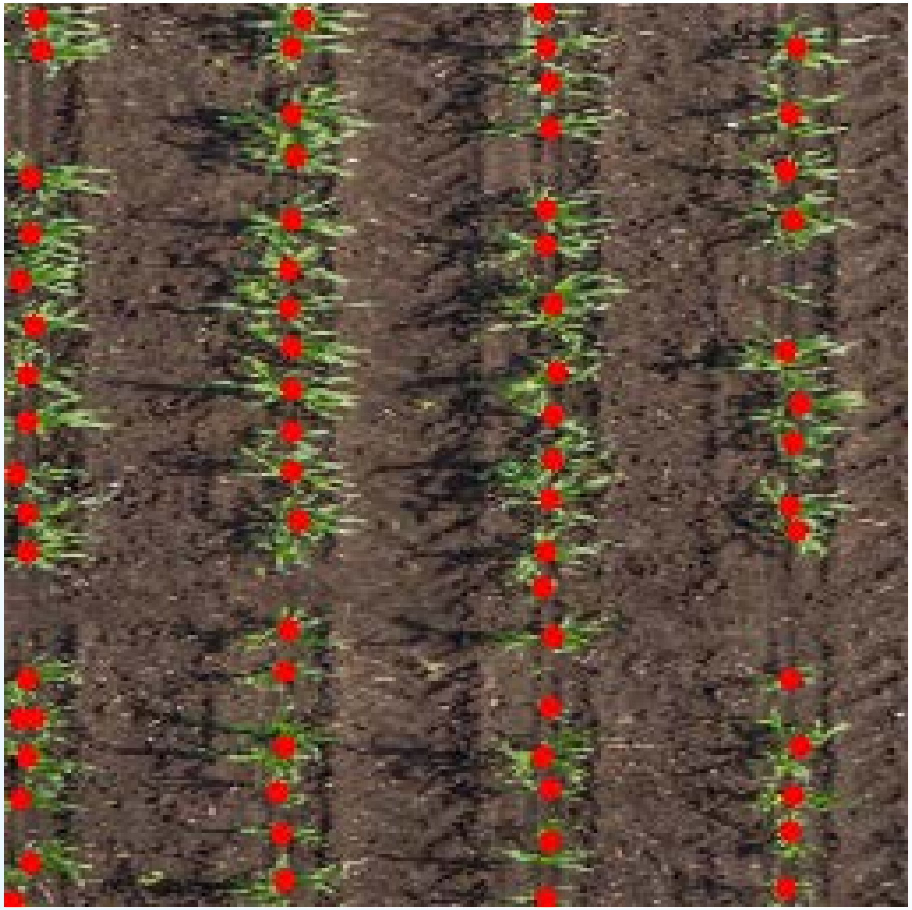
  - **We use Huber loss (also known as smooth L1)**

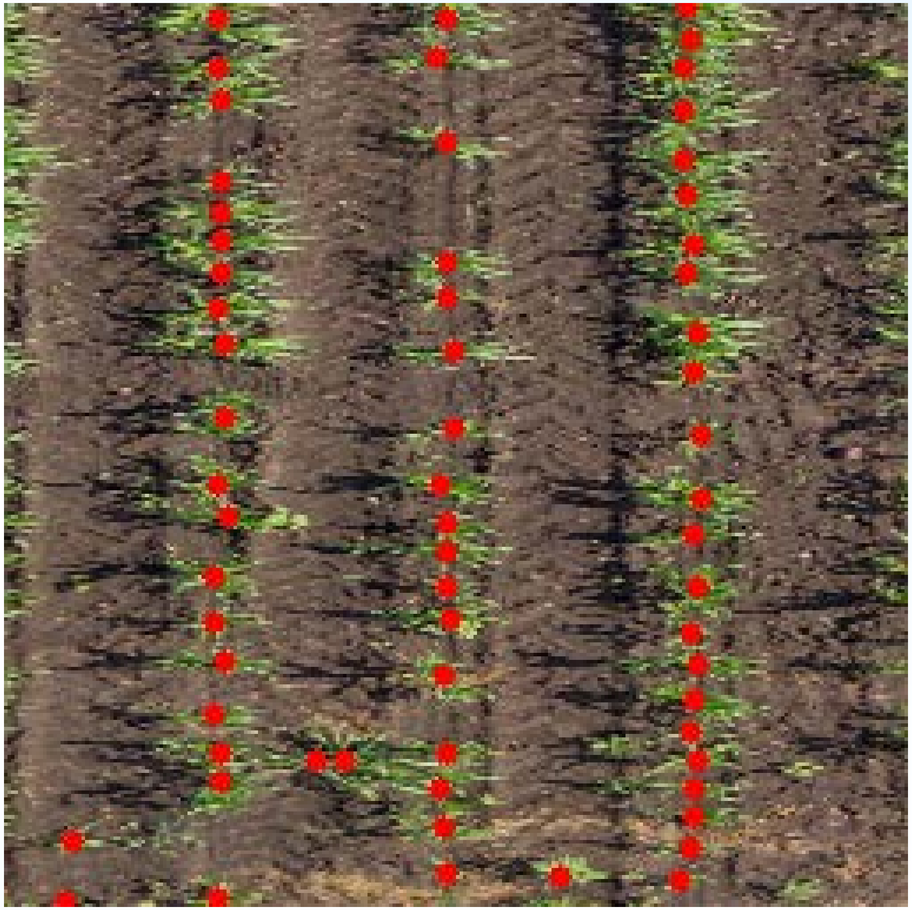
$$\mathcal{L}_{\text{reg}}(x) = \begin{cases} 0.5x^2, & \text{for} |x| < 1 \\ |x| - 0.5, & \text{for} |x| \geq 1 \end{cases}$$

  - **Advantages: Differentiable at the origin, and does not over-penalize outliers**

- **We achieved an F-score: 93 % (at 5cm of error)**

Results: 2016 Data

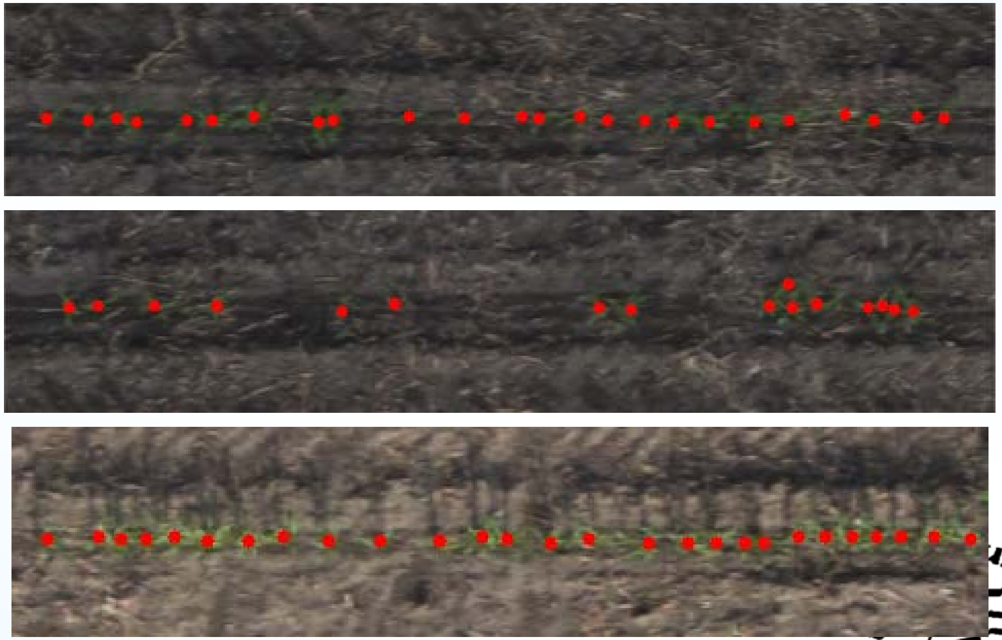




Estimated plant centers (in red)

Results: 2017 Data

- **Trained our model using ground truth of the Inbred Calibration Panel by Neal Carpenter**
- **Mean Average Percent Error: 12%**
- **Mean Average Error: 2.6 plants per row**
- **F-score: 93 % (at 5cm error)**



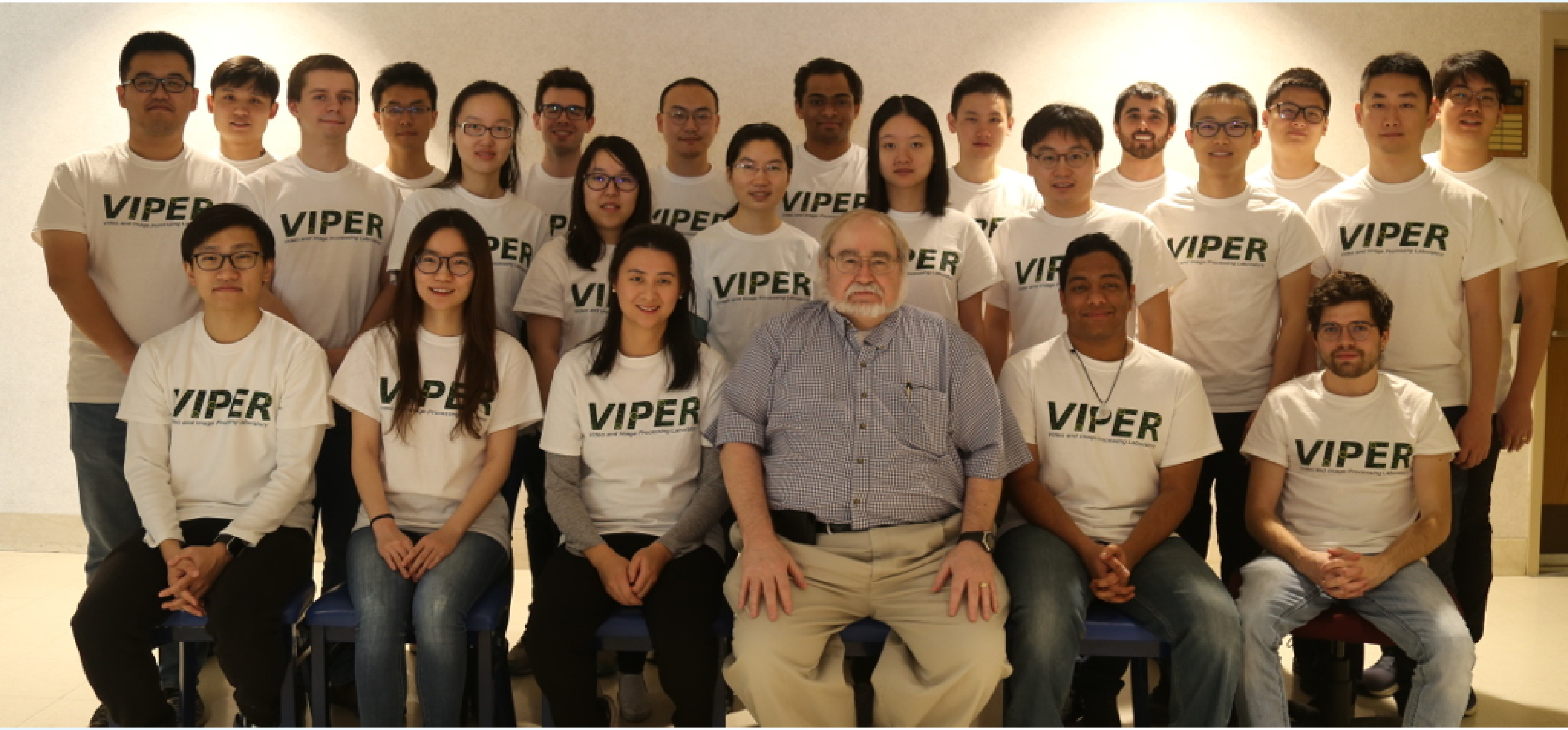
Testing results on the Hybrid Calibration Panel (2017/06/09)

Need for Low Power Computing

- End-users may not have the computational resource for fast inferencing
- Need for re-training
- Farmers need portable devices to get the results
- Edge-based computing is needed in this situation

VIPER

Video and Image Processing Laboratory



Acknowledgements

Cheap, Fast, and Low Power Deep Learning: I need it now!
(Please and Thank You)

Edward J. Delp

Video and Image Processing Laboratory (*VIPER*)
School of Electrical and Computer Engineering
School of Biomedical Engineering
Purdue University
West Lafayette, Indiana

email: ace@ecn.purdue.edu
https://engineering.purdue.edu/~ace/