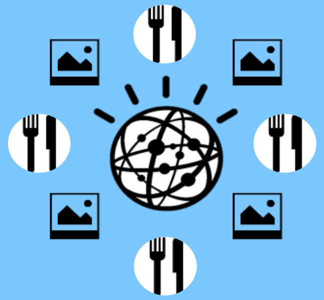
snap 

eat 

repEat 

a Food Recognition Engine for Dietary Logging

Michele Merler, Hui Wu, Rosario Uceda-Sosa, Quoc-Bao Nguyen, John R. Smith

IBM TJ Watson Research Center

Food Visual Recognition Team

John R Smith

Hui Wu

Michele Merler

Bao Nguyen

Rosario Uceda-Sosa

IBM TJ Watson Research Center - New York, USA

Outline

- Motivation
- System Architecture and Interface
- Image Recognition
- Conclusions and Future Directions

Motivation

Food Visual Recognition for Computer-Assisted Nutrition Logging

- **Exercise**, **sleep** and **nutrition** monitoring is essential for optimizing athletic **performance**
- Need to reduce friction (**manual**, **inaccurate**) to make nutrition monitoring fast and easy
- **Visual food recognition** greatly simplifies logging of meals using **context** and **content**
- Provides **accurate tracking of diet** and planning nutritional intake for achieving goals

Performance

Exercise

Sleep

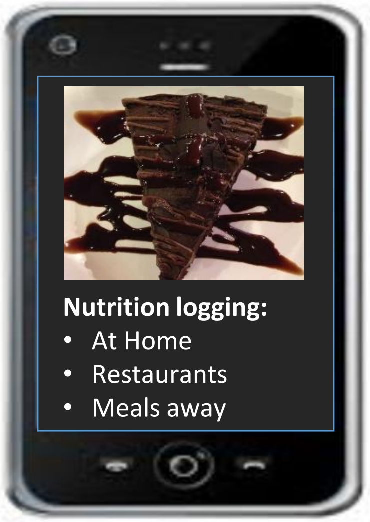
Nutrition

History | **Logging** | **Planning**

Watson Vision

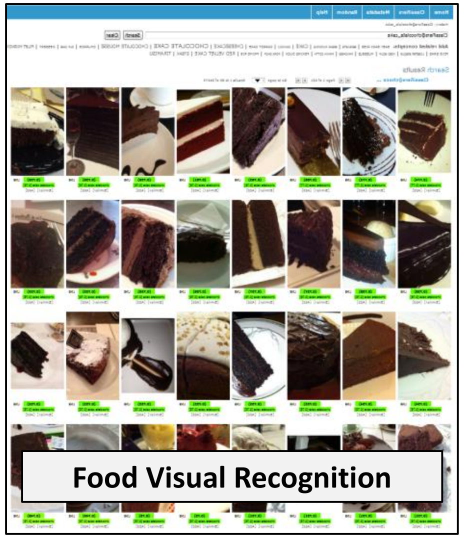
Context:
- Geo-Location
- Time of day
- Restaurant name
- Historical meals

Content:
- Photo
- Text
- Interaction



Unknown Photo →

← **Food Match & Nutrition Info**



Food matching:
- Fast, accurate
- Multi-modal
- Scalable

Food database:
- Food photos
- Nutrition info
- Menus
- User data

Leveraging *Context* for improving Food Recognition Accuracy

Known Menus (e.g., Restaurants)

Panera BREAD
Olive Garden ITALIAN KITCHEN
Applebee's
T.G.I. FRIDAYS
au bon pain
Denny's

Repeat Foods (e.g., Diet History)

Monday

Tuesday

Friday

Meal Times (e.g., Snack, Dessert)

Breakfast

Lunch

Dinner

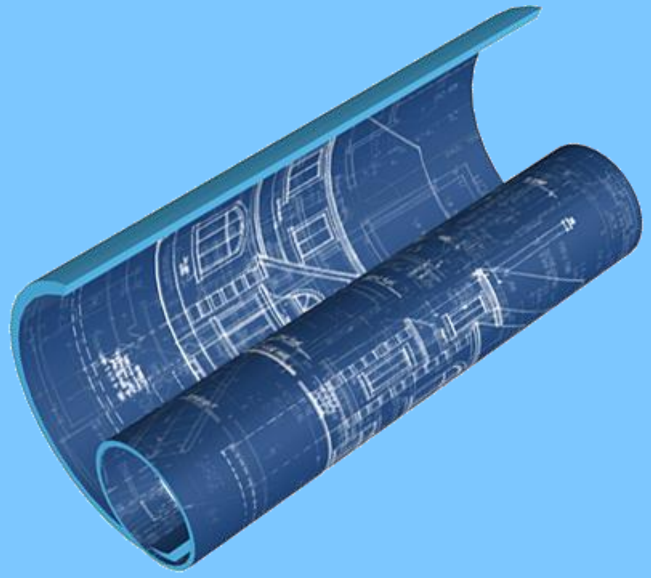
Cuisines (e.g., *Italian*)

Pizza

Pizza

Pizza

System Architecture and Interface



System Architecture

Snap Meal Photos

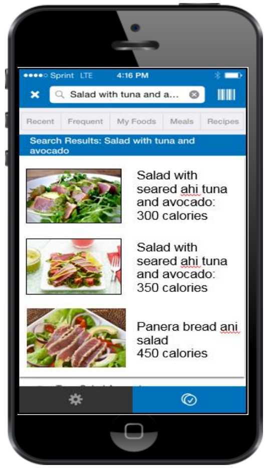
1 **In Context**
pics, restaurant

2 **In-the-wild**
pics

Context Information

Location, Restaurant, Menu

Nutrition Logging, Dietary Assistant



Recognized food category

Nutrition information

REST API

Food Visual Recognition and Analysis

Contextual Data (location, menu)

Food Semantic Hierarchy

Visual Models
Restau-rant 1
Restau-rant N
Wild

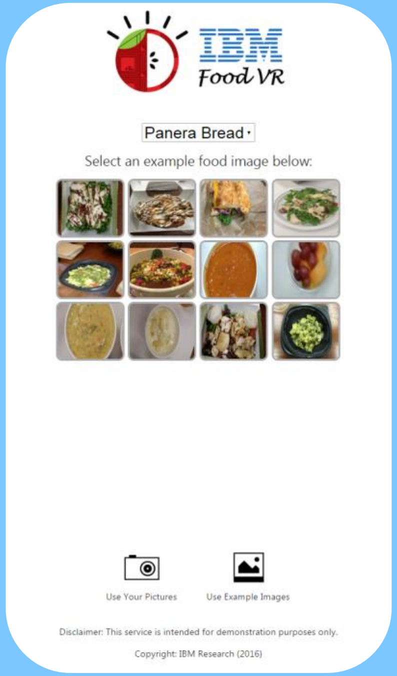
Nutritional info Database

Food Images Database

Client side

Server side

snap *eat* *repEat*



Demo

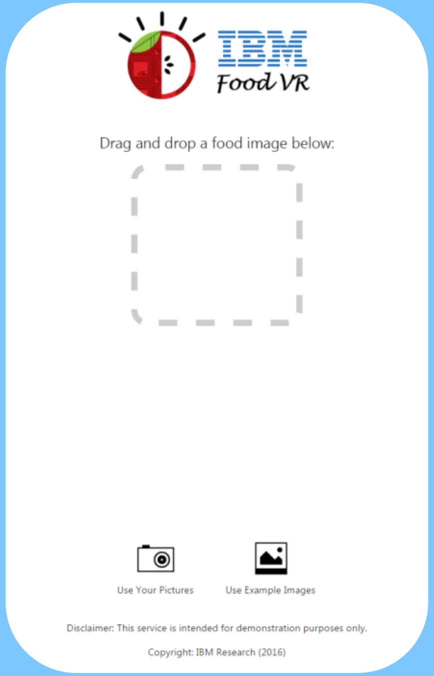


Image Recognition

Food Vs Not Food - Classifier

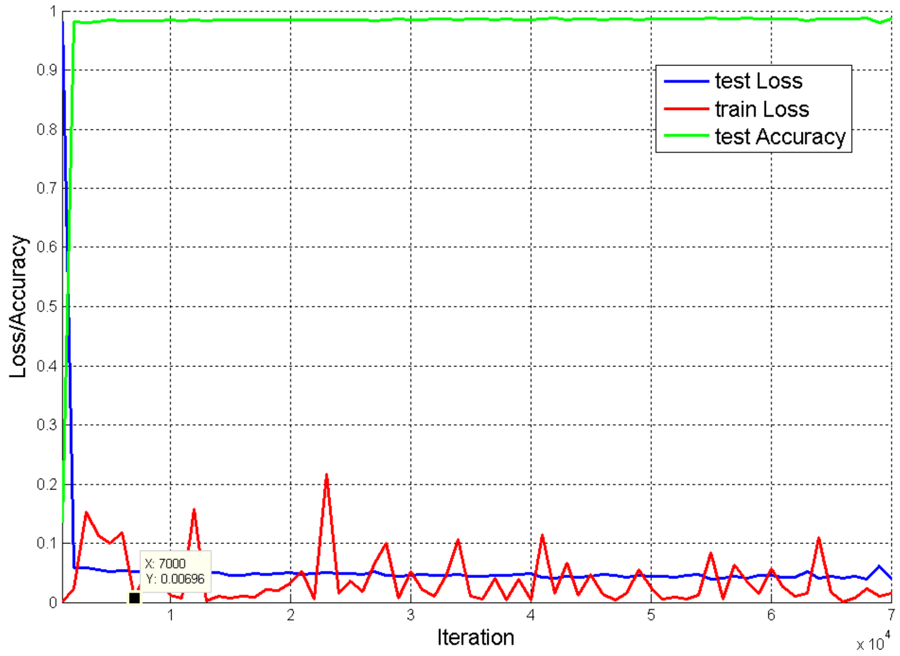
DATA

- Food vs Not-Food Dataset
  - Food
    - IBM food images
    - Tastespotting.com
    - Food.com
    - Food 101
    - UEC Food 256
    - Food 10K
    - UPMC_Food101
    - PFID
  - Not-Food
    - IBM non-food images
    - NUS Wide
    - SUN
    - ImageCLEF medical
    - Flickr images
- Training set 2.6M images
- Test set 660K images
- 43% Food, 57% Not-Food

MODEL

- Fine-tuned Binary GoogleNet
- Converged pretty fast
- Picked model at 7K iteration

- base_lr: 0.001
- lr_policy: "step"
- stepsize: 320000
- gamma: 0.96
- max_iter: 10000000
- momentum: 0.9
- weight_decay: 0.0002

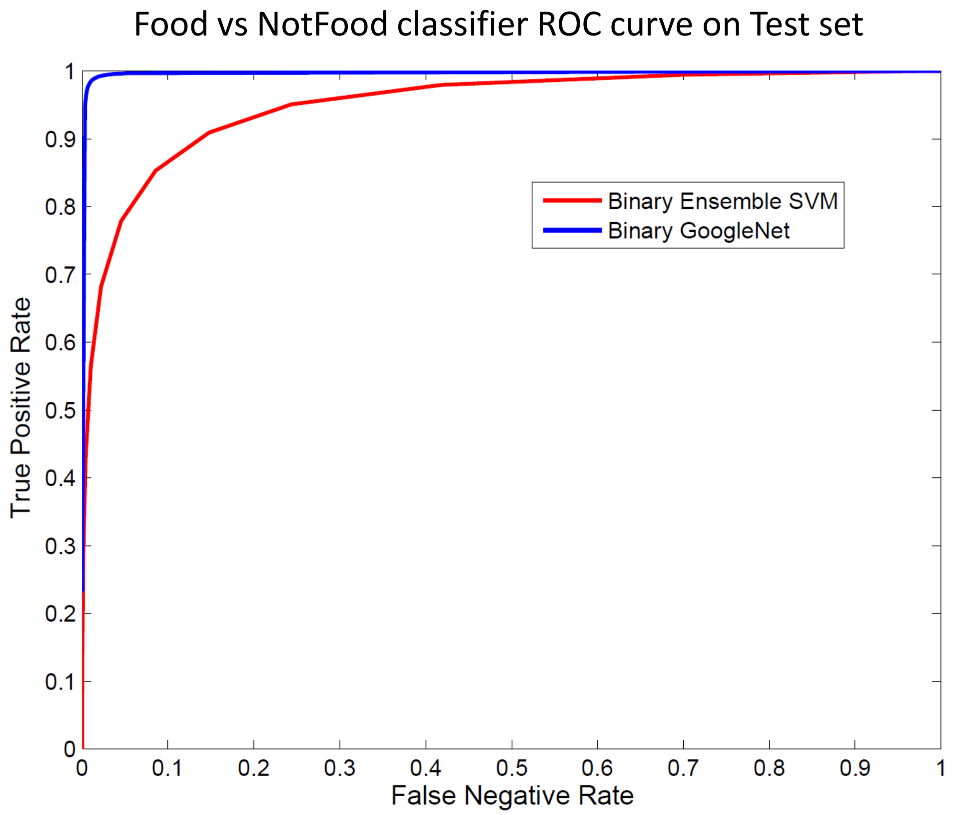


Food Filtering - Experiments

- Test set 660K images
  - 43% food
  - 57% not food
- Baseline: Ensemble SVM Food vs NotFood classifier
  - Best accuracy at 88.77% with t=0.45

- Binary GoogleNet has **98.95%** accuracy with t=0.55

Still ~7K errors!

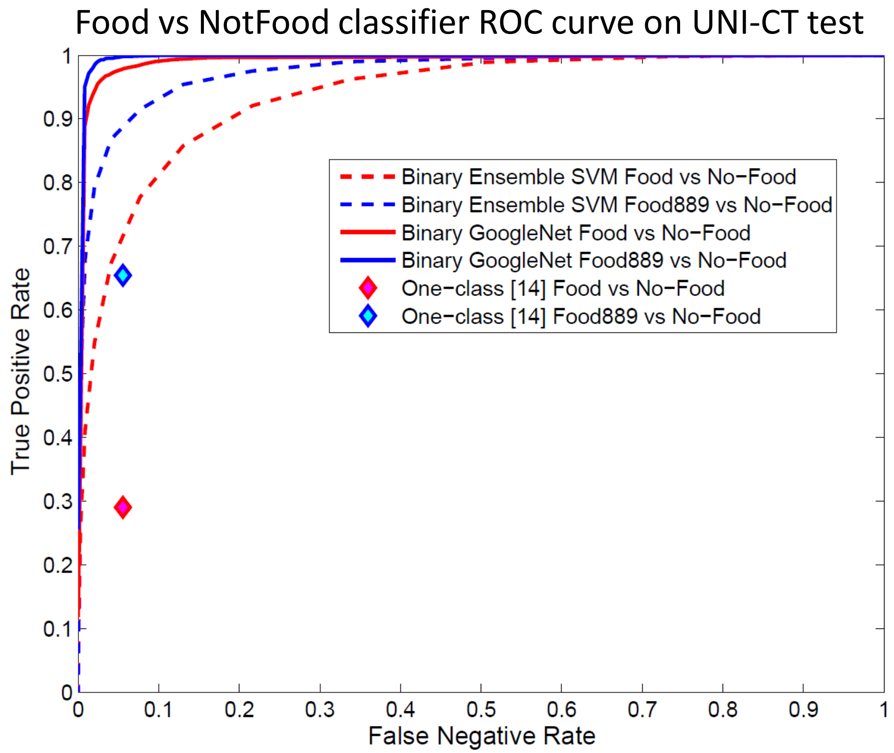


Food Filtering - Experiments

- UNI-CT Dataset http://iplab.dmi.unict.it/UNICT-FD889/
  - 3,583 Positive images of 889 foods (taken in restaurants with mobile)
  - 4,804 Positive food images (from Flickr)
  - 8,005 Negative images (from Flickr)

- 2 evaluation settings:
  - Food889 (positive) vs No-Food (Negative Flickr)
  - Food (positive Flickr) vs No-Food (Negative Flickr)

- Baseline: one class SVM from Farinella et al. [14]



Method	One-Class SVM [14]	Binary Ensemble SVM	Binary Fine-Tuned GoogleNet
Food889 True Positives Rate	0.6543	0.8685	**0.9711**
Flickr Food True Positives Rate	0.4300	0.6744	**0.9417**
Flickr No-Food True Negative Rate	0.9444	0.9589	**0.9817**
Overall Accuracy	0.9202	0.9513	**0.9808**

[14] G. M. Farinella, D. Allegra, F. Stanco, and S. Battiato. *On the exploitation of one class classification to distinguish food vs non-food images*. In New Trends in Image Analysis and Processing ICIAP **MaDiMa Workshop**, 2015.

How many foods need to be distinguished?

- In 2010, 85k different products were identified in US food chains[1]
- Most nutrition databases glean data from USDA, manufacturers and restaurant chains. Commercial database sizes range from 10k to 700k, but size is deceptive and too many options make logging food almost impossible
- Some databases are NOT curated (they include duplicates, unverified user entries, multiple entries per different portions of the same item, etc.). Most scientific, curated, comprehensive databases have 50k-80k entries
- Nutritionix[2] is the largest curated database, with 620k entries ('Spaghetti Marinara' produces over 3000 matches!)

	Approx size (US)	**Sample sources of data**
Restaurant menu items	27K	Restaurant sites (by law) (1800 large chains x 150 menu items)
Brand foods	25K	Manufacturer sites (by law)
Dishes in-the-wild	10K	USDA (9114 entries as of today)
Simple Ingredients	10K	Ingredient computation databases (Wolfram Alpha)

How many images for 70k categories?

Between 5 – 7 million
30-300 images per dish
AND abstract categories
Averaging 100 images per dish.

1. Weng Ng, Popkin: "**Monitoring foods and nutrients sold and consumed in the United States: Dynamics and Challenges**", http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289966/
2. https://www.nutritionix.com/

Food Recognition : Evaluation Datasets

Food in the wild

- **Food-101** [7]
  - 101 classes
  - 1,000 images per class
- **Food 500 (ours)**
  - 508 classes
  - 290 images per class

Food-101 Images

Food in context

- **6-Chain** (ours)
  - ~ 50 classes / chain
  - ~10 image / class
  - Images from Applebee's, Denny's, Olive Garden, Panera Bread, and TGI Fridays

6-Chain Images

- Random splits: 75% for training, 25% for testing
- Evaluation metric: Fine-grained classification accuracy

[7] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014.
https://www.vision.ee.ethz.ch/datasets_extra/food-101/

Context-based Food Recognition (top 1 accuracy)

- Performance of Deep Learning Food Recognition Models on Restaurant Chains food
- Each Restaurant chain is evaluated independently

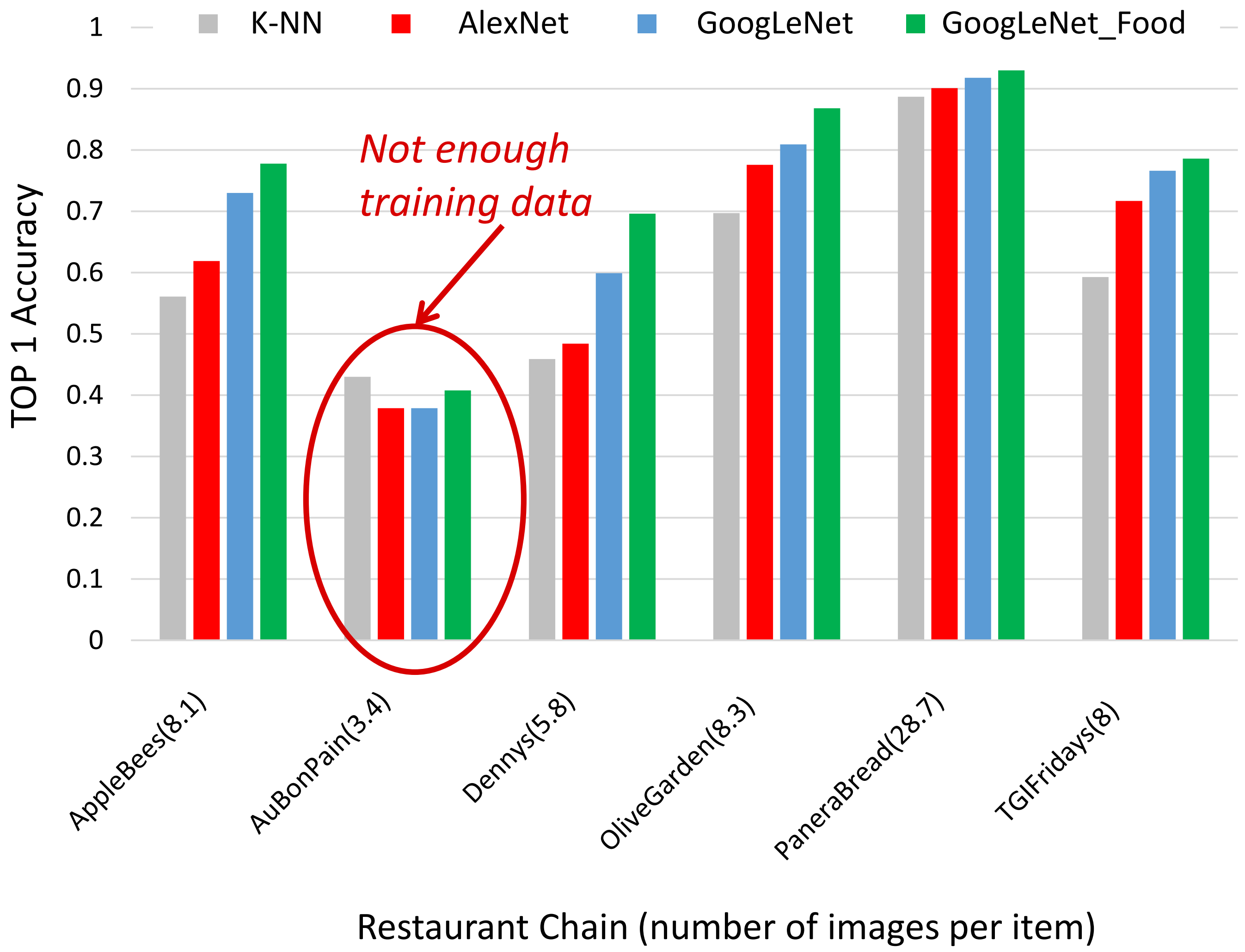
K-NN: based on fc7 features from AlexNet [26]

AlexNet: finetuned on restaurant chain training set

GoogLeNet [36] : finetuned on Restaurant chains training set, similar to im2calories [30]

GoogLeNet$_{Food}$: two finetuning steps, first n subset of Food vs Not-food dataset, then Restaurant chains training set

Restaurant	# Classes	# Images	# Images per class
Applebee's	50	405	8
Au Bon Pain	43	146	3
Denny's	56	325	6
Olive Garden	55	457	8
Panera Bread	79	2,267	28
TGI Fridays	54	432	8



[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS* 2012
[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CVPR* 2015
[30] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy. Im2calories: towards an automated mobile vision food diary. *ICCV* 2015

Context-based Food Recognition (top 3 accuracy)

- Performance of Deep Learning Food Recognition Models on Restaurant Chains food
- Each Restaurant chain is evaluated independently

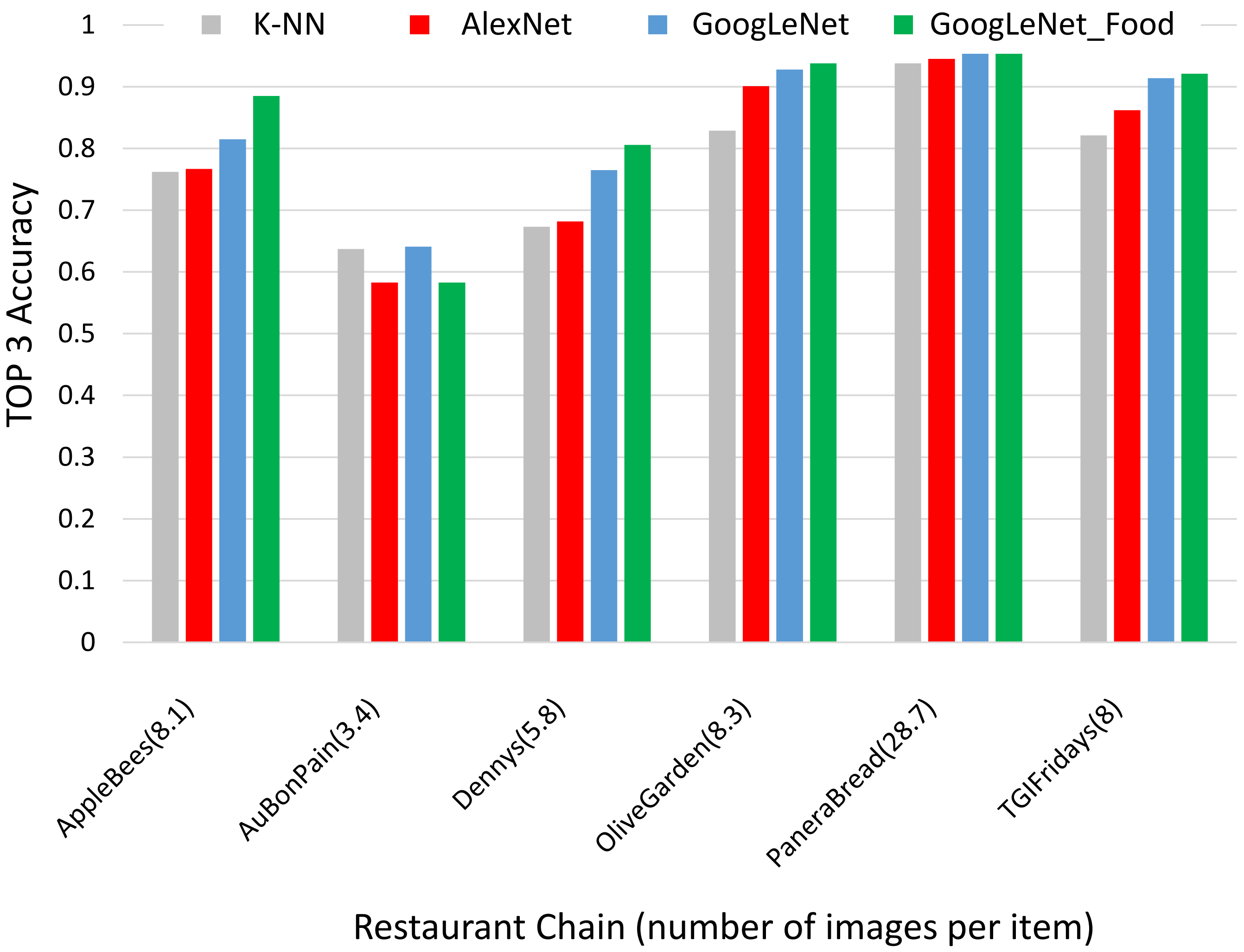
K-NN: based on fc7 features from AlexNet [26]

AlexNet: finetuned on restaurant chain training set

GoogLeNet [36] : finetuned on Restaurant chains training set, similar to im2calories [30]

$\textbf{GoogLeNet}_{\textbf{Food}}$: two finetuning steps, first n subset of Food vs Not-food dataset, then Restaurant chains training set

Restaurant	# Classes	# Images	# Images per class
Applebee's	50	405	8
Au Bon Pain	43	146	3
Denny's	56	325	6
Olive Garden	55	457	8
Panera Bread	79	2,267	28
TGI Fridays	54	432	8



[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS* 2012
[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CVPR* 2015
[30] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy. Im2calories: towards an automated mobile vision food diary. *ICCV* 2015

Context-based Food Recognition (Category level accuracy)

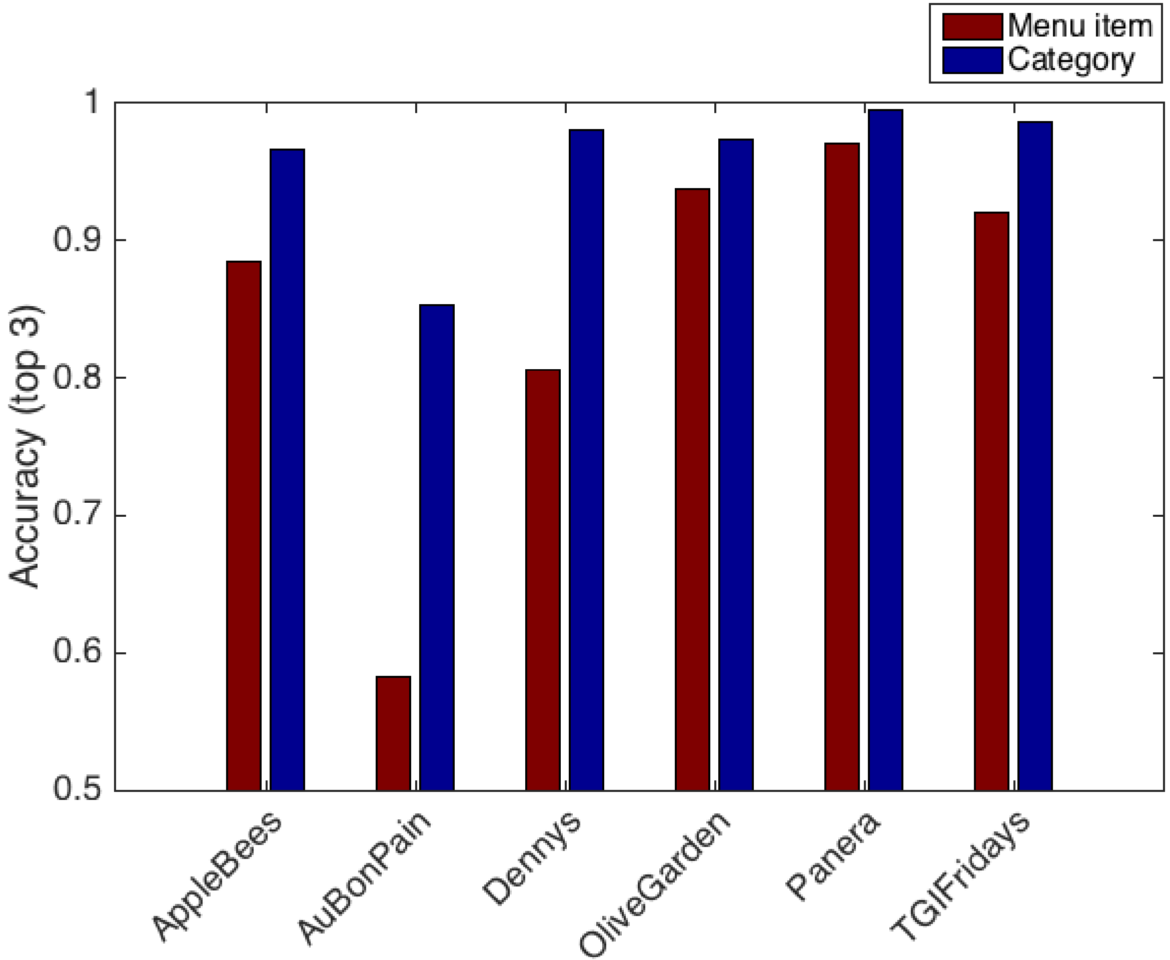
- Most recognition errors result from visually similar dish items in the same category
- E.g., even if the system fails to recognize the specific type of soup, it still recognizes that it is a soup
- Idea*: incorporate hierarchical taxonomic information in learning process

Item: triple bacon burger
Estimated: mushroom swiss burger
Category: Burger

Item: black bean soup
Estimated: turkey chili
Category: Soup

Item: strawberry fields salad
Estimated: Yucatan Chicken Salad
Category: Salad

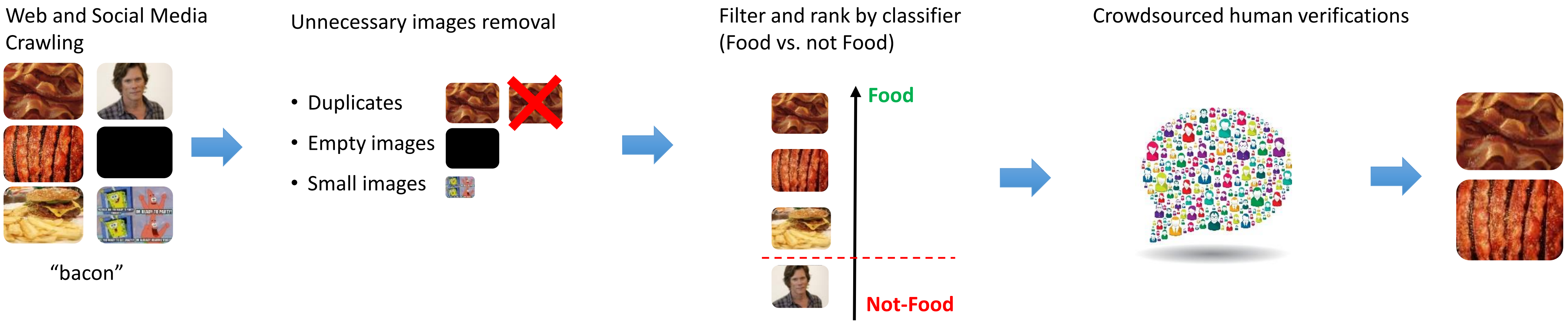
Item: sesame seed bagel
Estimated: everything bagel
Category: Bagel

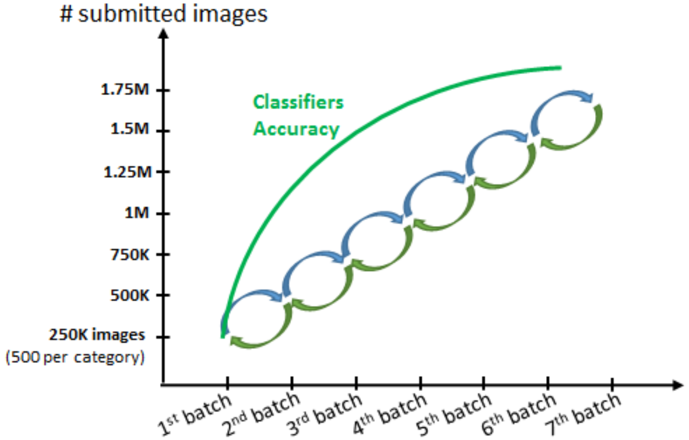


* Hui Wu, Michele Merler, Rosario Uceda-Sosa, John Smith, Learning to Make Better Mistakes: Semantics-aware Visual Food Recognition. *ACM Multimedia* 2016

Food “in the wild” Dataset Curation

- Building a large-scale food image database
- Enables accurate food visual recognition and nutrition logging in real world settings





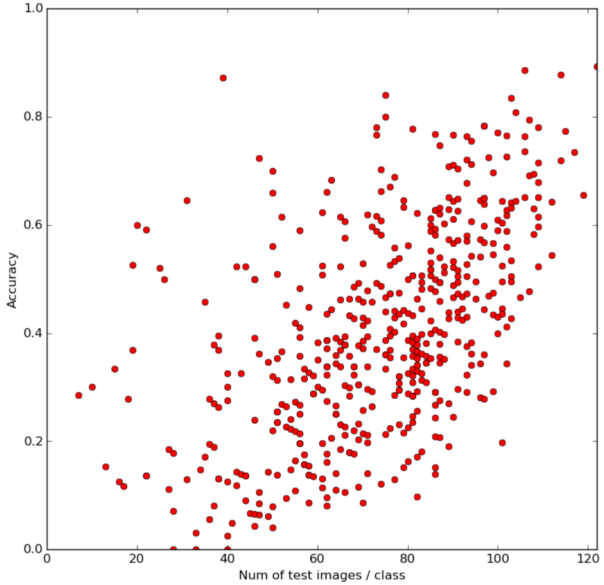
Comparison to existing datasets

	Dataset	Number of Classes	Number of Images/Class	Number of Images	Food Ontology
NOT-IBM	**UEC Food 256 [22]**	256	89	31,651	None
NOT-IBM	**Geolocalized [40]**	3,852	30	117,504	None
NOT-IBM	**Food-101 [7]**	101	1000	101,100	None
NOT-IBM	**ETHZ Food 101 [37]**	101	1000	101,100	None
IBM	**Food 500**	**508**	**290**	**148,408**	**Yes**
IBM	**Food 3,000 (ongoing)**	3000	500	1.5M	**Yes**

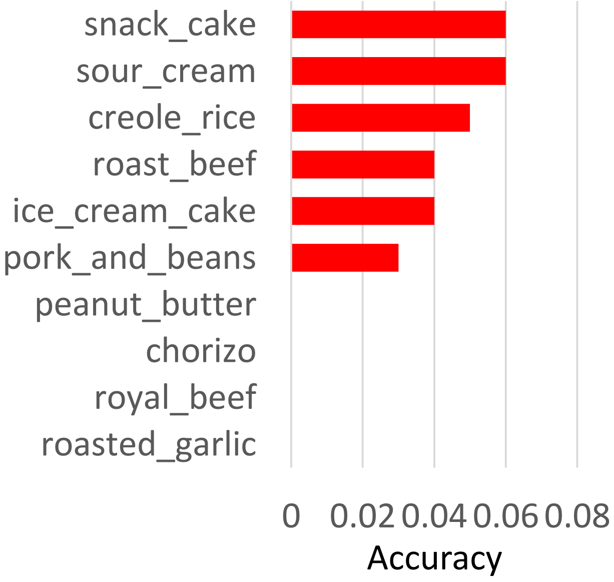
500 Foods "in the wild" Classification

Model: GoogleNet pretrained on Imagenet and finetuned on given dataset

Dataset	Accuracy (top 1)
Food 101 [Martinel ICCV15]	79
Food 101 (ours)	69.64
Food 500 (ours)	40.37



Worst Categories



Creole rice

VS

Jambalaya

Roast beef

VS

Pastrami

Beef vindaloo

VS

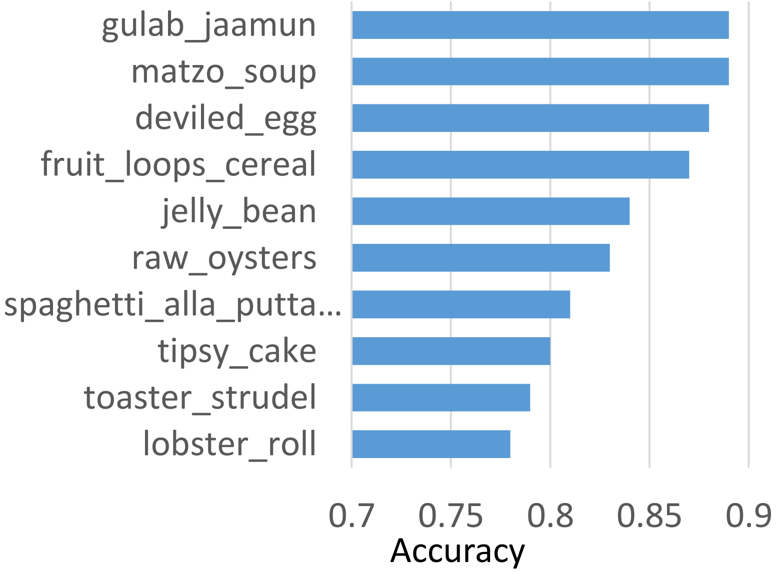
Rogan josh

Peanut butter

VS

Fudge

Best Categories



Conclusions

Conclusions and Future Directions

- Created end-to-end food recognition API that can recognize pictures of food in restaurants and “in the wild”
- Tested state of the art on largest food image dataset with ~150K images of 500 food categories organized in a hierarchical taxonomy
- Context matters
- Amount and quality of training images matter

FUTURE DIRECTIONS

- More data
  - expand “wild” dataset to 1-3K categories and 1-2M images
  - expand Restaurant chains dataset by adding more restaurants
- Food portion estimation “in the wild” will require food segmentation, depth and volume estimation
- Incorporate other types of context (diet history, meal time, local cuisine)

snap *eat* *repEat*

Check out our related work!

Hui Wu, Michele Merler, Rosario Uceda-Sosa, John Smith

Learning to Make Better Mistakes: Semantics-aware Visual Food Recognition

ACM Multimedia Poster Session – Monday Oct 17th 14.00 – 17.00

Questions?

snap *eat* *repEat*