



Introduction to

Competitive

Machine

Learning

Matt Motoki & Thomas Yokota

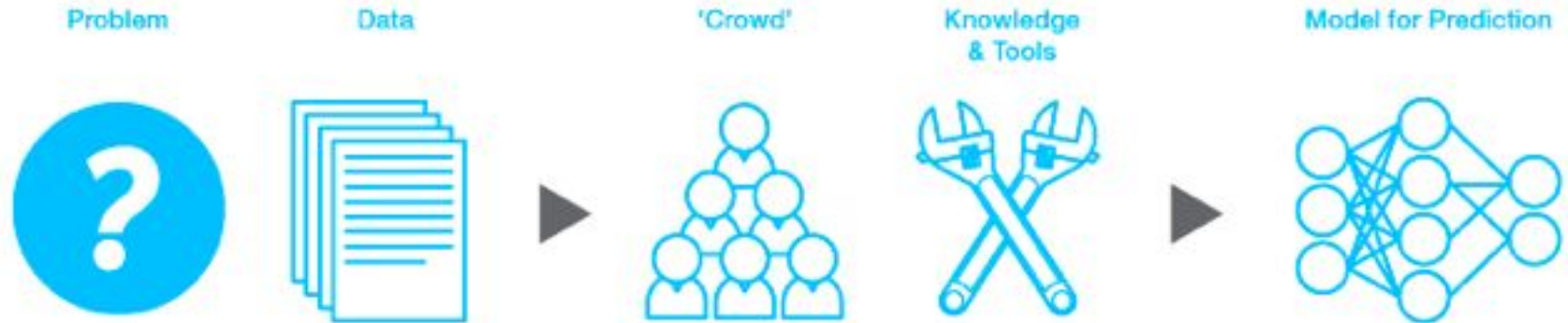
Overview



- **What** is Competitive Machine Learning?
- **Why** Compete?
- **Where** Can I Compete?
 - Netflix
 - ImageNet
 - Kaggle
- **How** Do I Start?
 - Kaggle Walkthrough
 - Titanic Competition

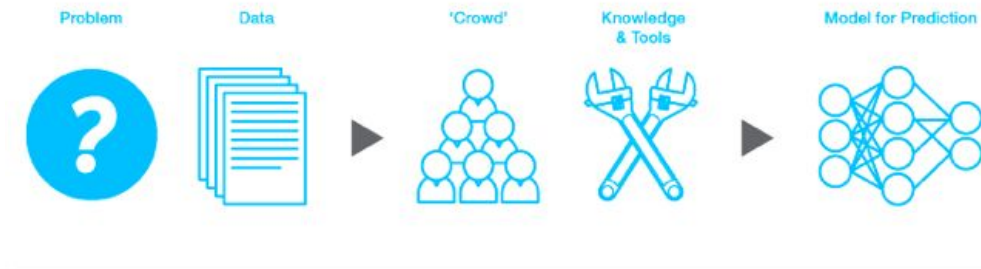
What is Competitive
Machine Learning?

What Is Competitive Machine Learning?



What Is Competitive Machine Learning?

1. Problem
 - a. Business
 - b. Research
2. Data
 - a. Messy
 - b. Disparate sources
 - c. Different formats (text, tabulated, images, etc.)
3. Crowd-sourced
 - a. Teaming up
4. Knowledge & Tools
 - a. Python, R, Keras, TF, C++, etc.
5. Models
 - a. Benchmarked (evaluation score)
 - b. Leaderboard



Downsides to Competitive Machine Learning



What Competitive Machine Learning Lacks

- Problem Formulation
- Customer Interaction
- In-depth theory

Problems with Competitions

- Gains are incremental
- Addictive

Downsides to Competitive Machine Learning



What Competitive Machine Learning Lacks

- Problem Formulation
- Customer Interaction
- In-depth theory

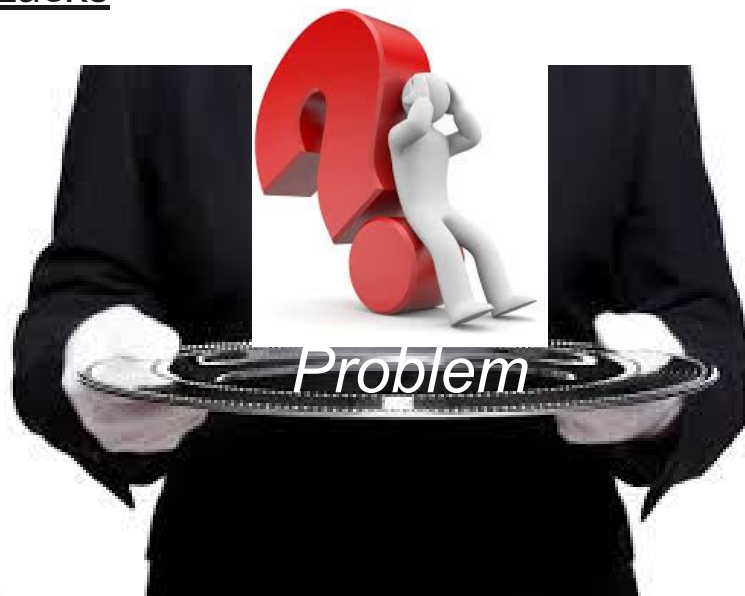
Problems with Competitions

- Gains are incremental
- Addictive

Downsides to Competitive Machine Learning

What Competitive Machine Learning Lacks

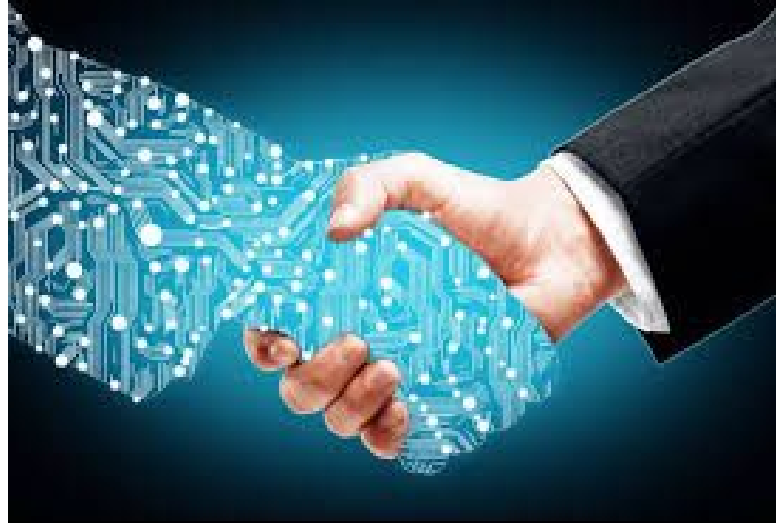
- Problem Formulation



Downsides to Competitive Machine Learning

What Competitive Machine Learning Lacks

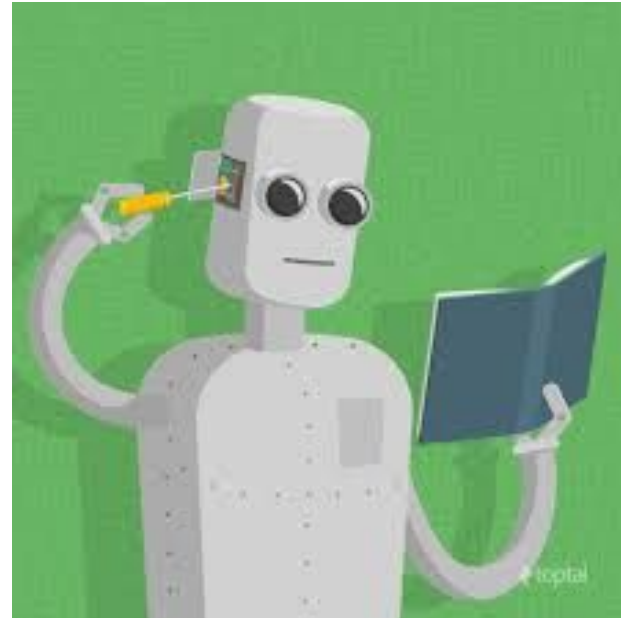
- Problem Formulation
- Customer Interaction



Downsides to Competitive Machine Learning

What Competitive Machine Learning Lacks

- Problem Formulation
- Customer Interaction
- In-depth theory



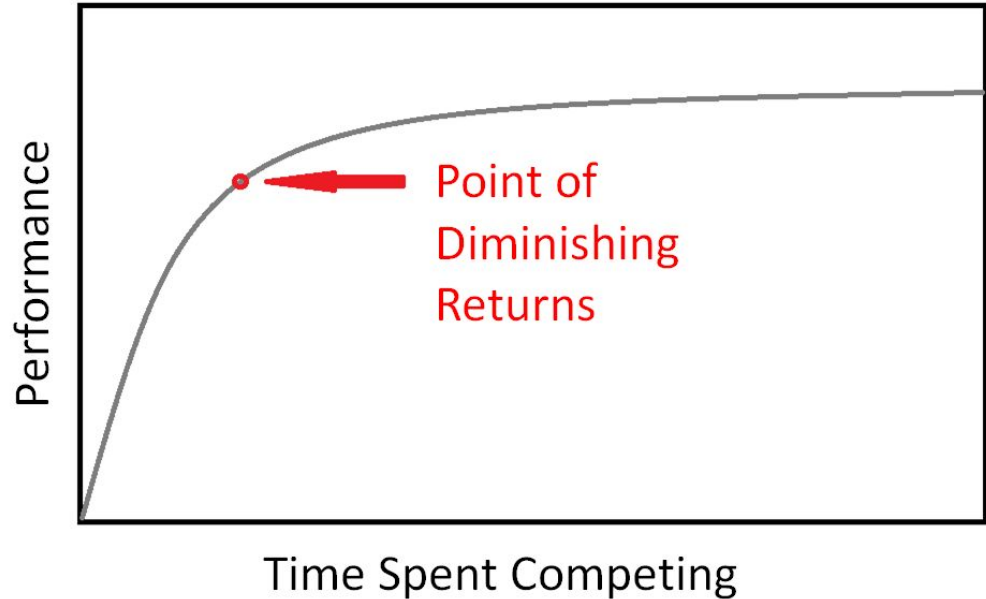
Downsides to Competitive Machine Learning

What Competitive Machine Learning Lacks

- Problem Formulation
- Customer Interaction
- In-depth theory

Problems with Competitions

- Gains are incremental



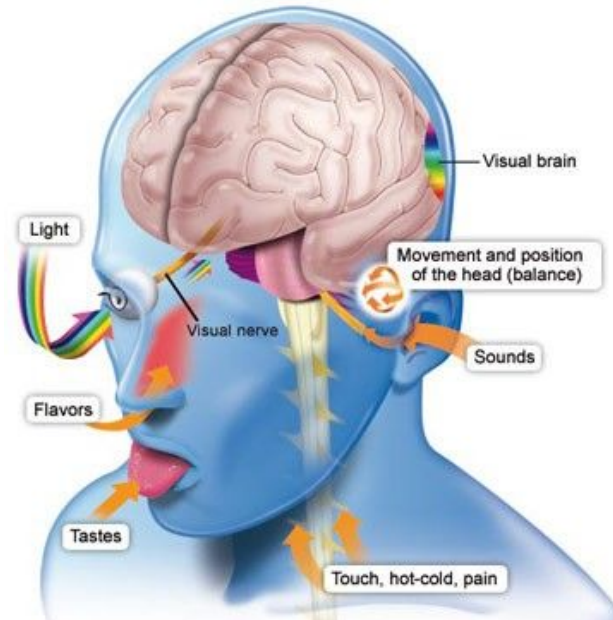
Downsides to Competitive Machine Learning

What Competitive Machine Learning Lacks

- Problem Formulation
- Customer Interaction
- In-depth theory

Problems with Competitions

- Gains are incremental
- **Addictive**



Why Compete?


Why compete?

- Gain experience




Why compete?

- Gain experience
- Build a portfolio




Vladimir Iglovikov
Data Scientist at Lyft
San Francisco, California, United States
Joined 3 years ago · last seen in the past day
[Twitter](#) [LinkedIn](#)




Followers 162
Following 2


**Competitions Master**


[Home](#) [Competitions \(61\)](#) [Kernels \(7\)](#) [Discussion \(311\)](#) [Organizations \(1\)](#) [Followers \(162\)](#) [Contact User](#) [Follow User](#)


Competitions Master


Current Rank	Highest Rank
31 of 73,783	19

 4	 8	 6
--	--	---




Carvana Image Masking Ch... **1st**
 · 4 months ago · Top 1% of 735


Dstl Satellite Imagery Feat... **3rd**
 · a year ago · Top 1% of 419

Planet: Understanding the ... **7th**
 · 6 months ago · Top 1% of 938


Kernels Contributor

Unranked




 1	 0	 0
--	--	--


xgb 1114 **104**
 · a year ago votes


Jaccard(polygons, polygons... **1**
a year ago vote


Discussion Expert

Current Rank	Highest Rank
12 of 45,985	6

 28	 21	 128
---	---	--

This is insane discriminatio... **119**
 · 7 months ago votes

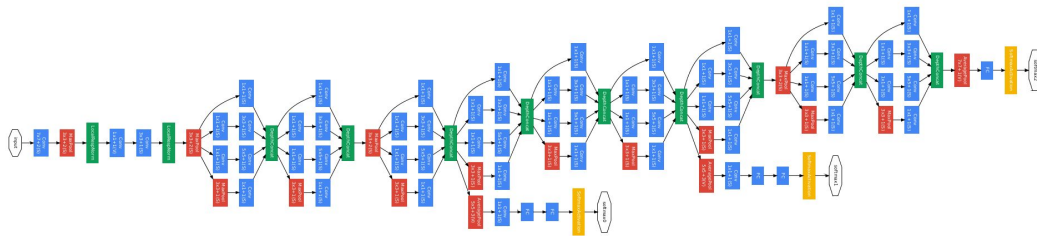
This is insane discriminatio... **76**
 · 7 months ago votes

This is insane discriminatio... **55**
 · 7 months ago votes

Why compete?

- Gain experience
- Build a portfolio
- **Learn new techniques**

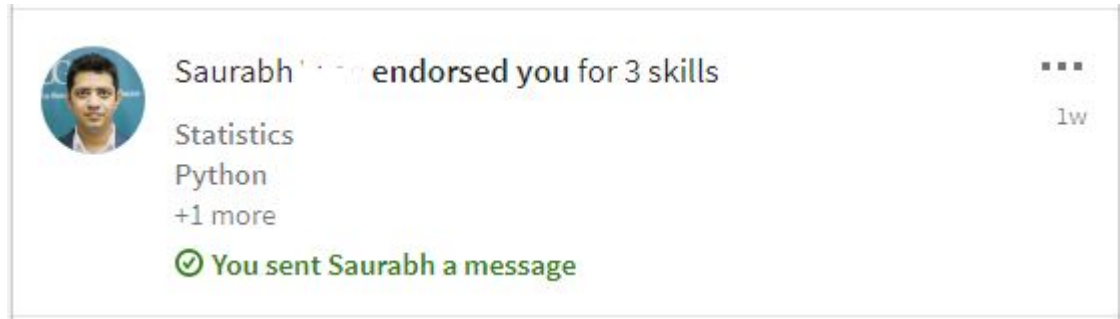
Google Inception Network



dmlc
XGBoost eXtreme Gradient Boosting

Why compete?

- Gain experience
- Build a portfolio
- Learn new techniques
- **Networking**



Why compete?

- Gain experience
- Build a portfolio
- Learn new techniques
- Networking
- Prizes



2018 Data Science Bowl

Find the nuclei in divergent images to advance medi...

Featured · 3 months to go · 🧬 biology

\$100,000

598 teams



Mercari Price Suggestion Challenge

Can you automatically suggest product prices to onl...

Featured · a month to go · 🛒

\$100,000

1,801 teams



Zillow Prize: Zillow's Home Value Predictio...

Can you improve the algorithm that changed the wo...

Featured · 14 days ago · 🏠 housing, real estate

\$1,200,000

3,779 teams



Passenger Screening Algorithm Challenge

Improve the accuracy of the Department of Homela...

Featured · a month ago · 🧑 terrorism, image data, ...

\$1,500,000

518 teams

Why compete?

- Gain experience
- Build a portfolio
- Learn new techniques
- Networking
- Prizes
- Job opportunities



Facebook V: Predicting Check Ins

Identify the correct place for check ins

1,212 teams · 2 years ago

Where Can I Compete?



Over 75% of what people watch comes from a recommendation

NETFLIX

Background

- Improve Netflix's Cinematch algorithm by 10% (RMSE).
- 100 million movie ratings, 480,189 movies and 17,770 users
- Oct 2006 – Sep 2009
- *\$1,000,000 prize*

NETFLIX

Background

- Improve Netflix's Cinematch algorithm by 10% (RMSE).
- 100 million movie ratings, 480,189 movies and 17,770 users
- Oct 2006 – Sep 2009
- \$1,000,000 prize

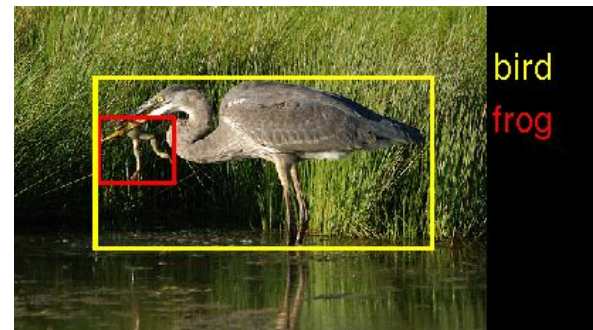
Winning Solution

- Matrix factorization
- Neural Networks
- Gradient Boosted Trees



ImageNet

Large Scale Visual Recognition Challenge

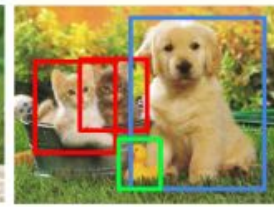


**Classification
+ Localization**



CAT

Object Detection

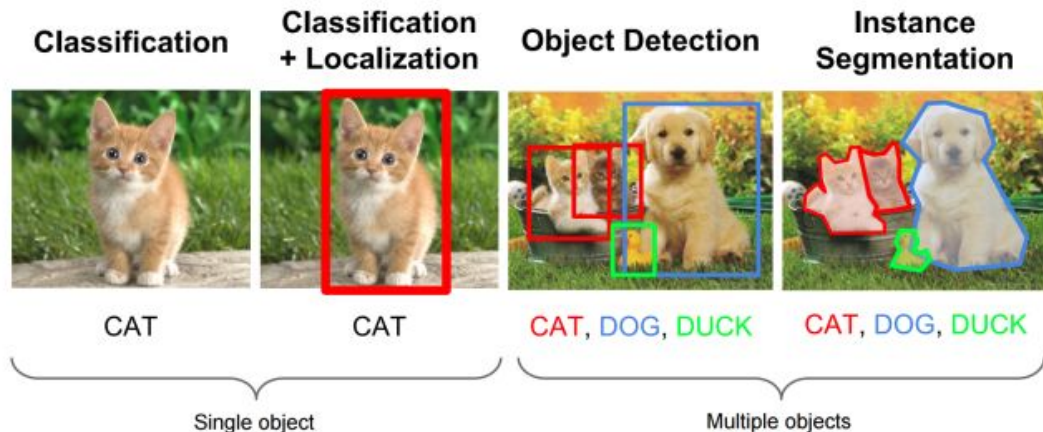


CAT, DOG, DUCK



Background

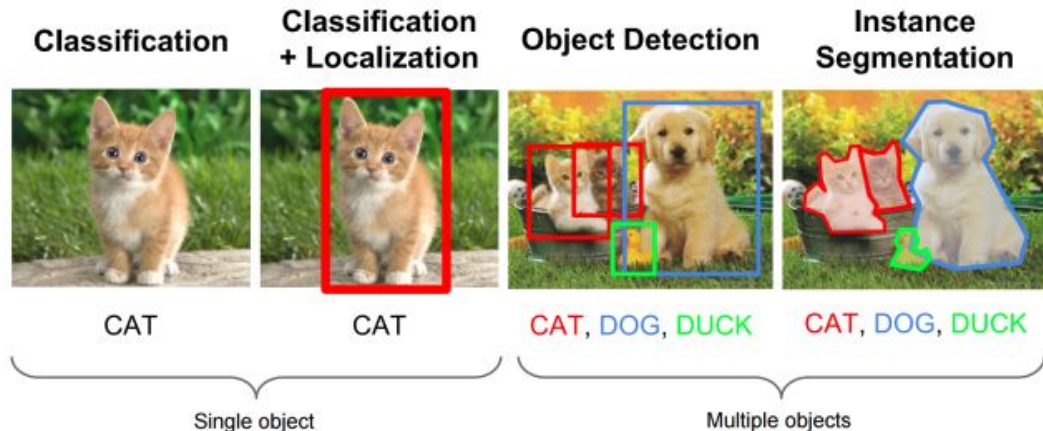
- Started in 2010
- 14,197,122 images
- 154 GB compressed





Background

- Started in 2010
- 14,197,122 images
- 154 GB compressed

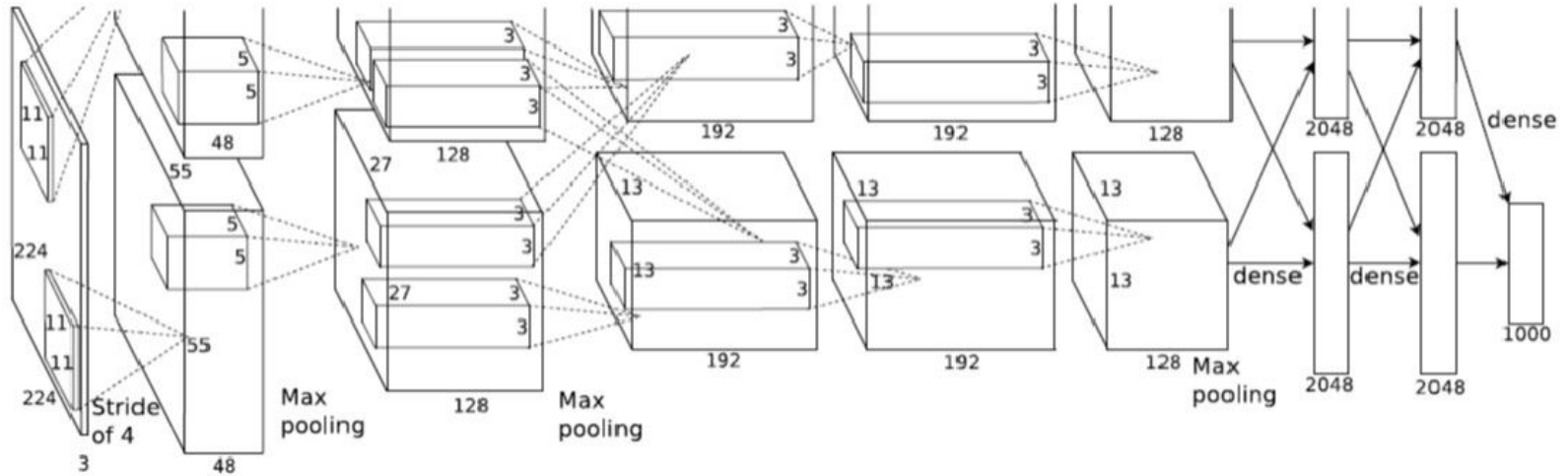


Tasks

- [Object localization](#) for 1000 categories.
- [Object detection](#) for 200 fully labeled categories.
- [Object detection from video](#) for 30 fully labeled categories.



2012: Alexnet started the deep learning craze.





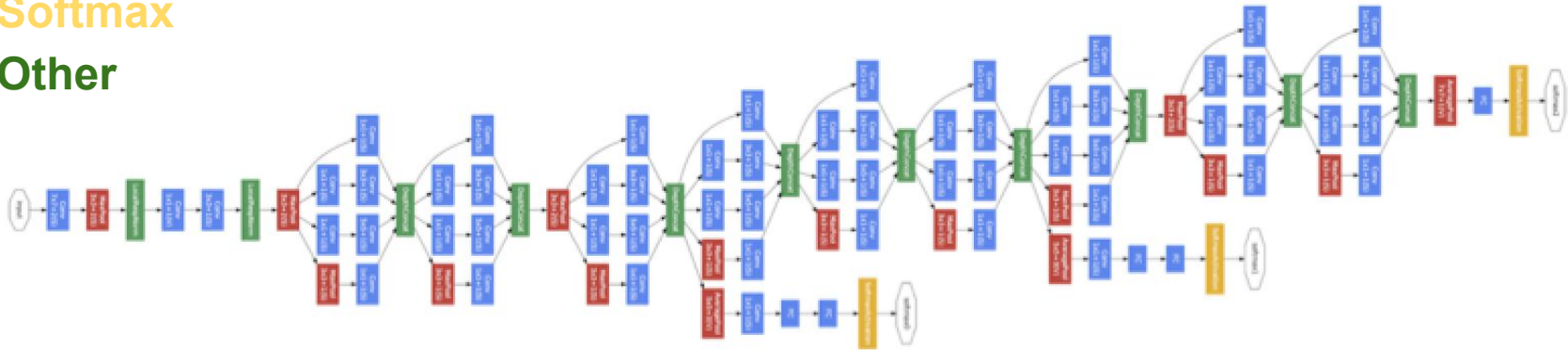
2014: Google introduced the Inception architecture

Convolution

Pooling

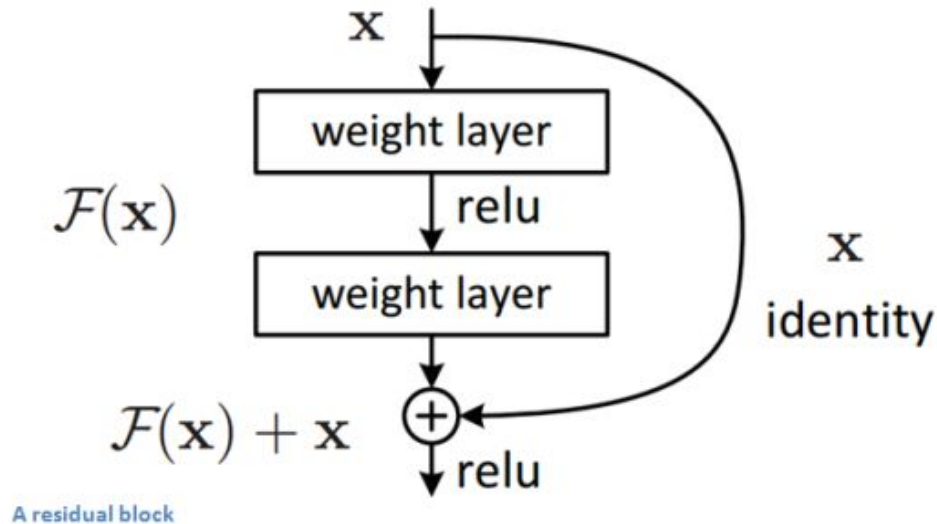
Softmax

Other



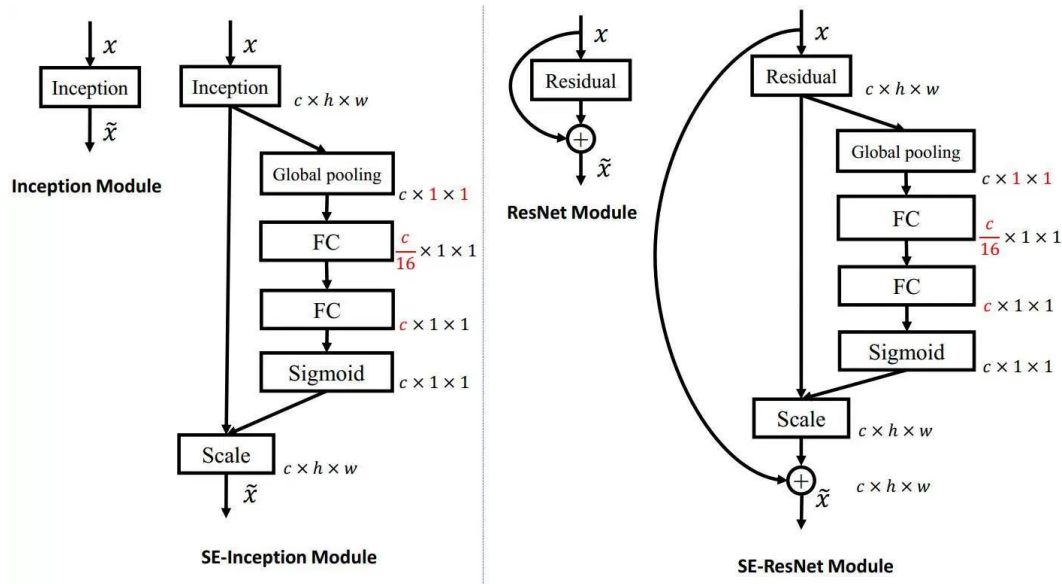


2015: Microsoft research introduced **Deep Residual Networks (ResNet)**

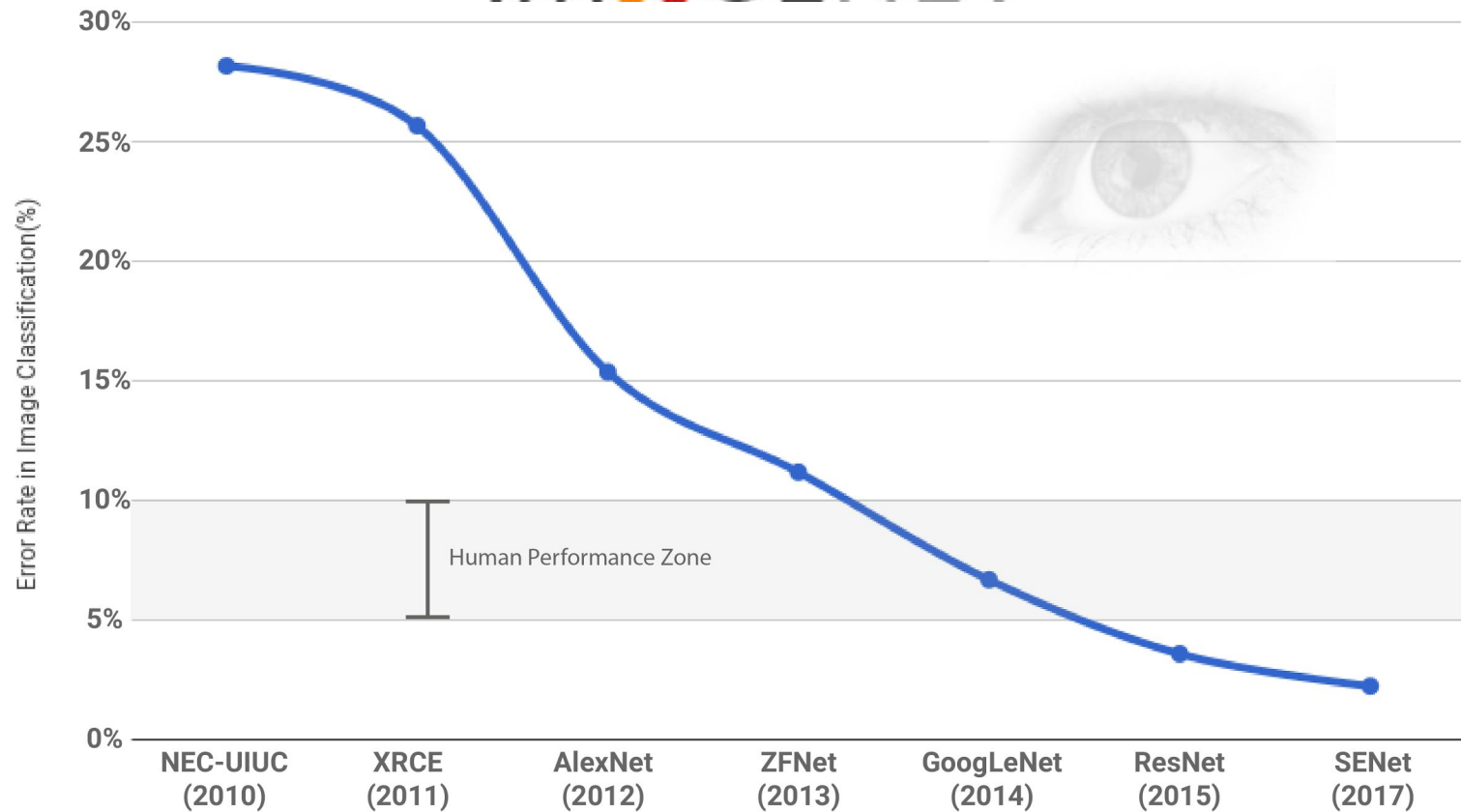




2017: Momenta introduced Squeeze-and-Excitation Networks (SENet)



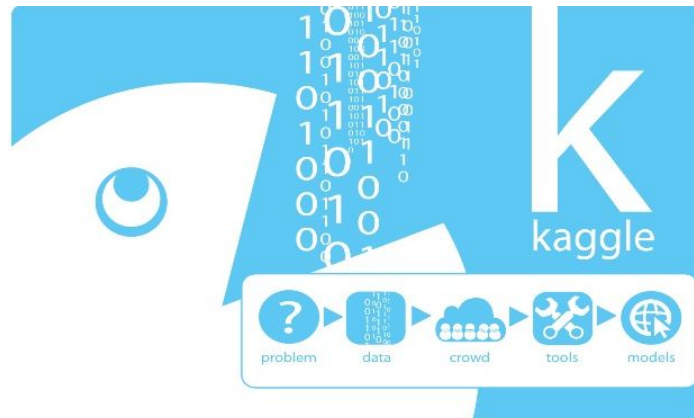
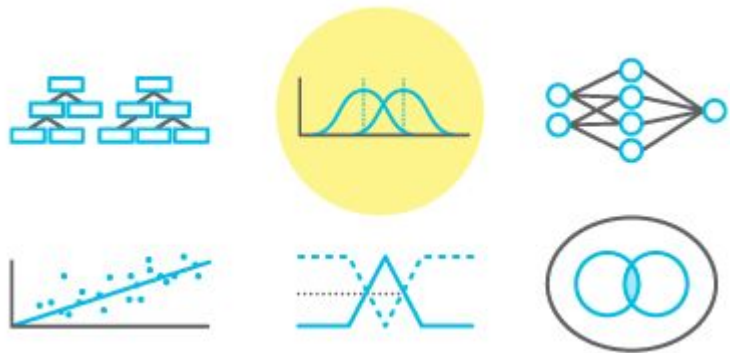
IMGENET





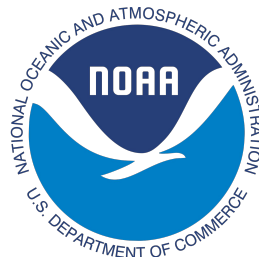
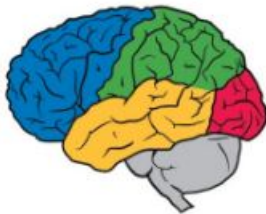
kaggle

<http://www.kaggle.com>



kaggle™

- Created in 2010
- Acquired by Google in 2017
- Competitions support advances in **machine learning** and **AI** in both **industry** and **research**
- Clients include



kaggle Walkthrough

The Home of Data Science & Machine Learning

Kaggle helps you learn, work, and play



Create an account

or

Host a competition



Competitions ›

Climb the world's most elite machine learning leaderboards

Datasets ›

Explore and analyze a collection of high quality public datasets

Kernels ›

Run code in the cloud and receive community feedback on your work

✕ Dismiss

General

InClass

Hosted

Sort by

Grouped

All Categories

Search competitions



6 Entered Competitions



2018 Data Science Bowl

Find the nuclei in divergent images to advance medical discovery

Featured · 3 months to go · biology



\$100,000
773 teams



Mercari Price Suggestion Challenge

Can you automatically suggest product prices to online sellers?

Featured · 24 days to go ·



\$100,000
1,912 teams



Statoil/C-CORE Iceberg Classifier Challenge

Ship or iceberg, can you decide from space?

Featured · 5 days ago · weather, shipping, image data, binary classification



\$50,000
3,343 teams



Corporación Favorita Grocery Sales Forecasting

Can you accurately predict sales for a large grocery chain?

Featured · 13 days ago · food and drink, tabular data, regression, future predict...



\$30,000
64/1675



Recruit Restaurant Visitor Forecasting

Predict how many future visitors a restaurant will receive

Featured · 9 days to go ·



\$25,000
1,924 teams



Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Getting Started · 2 years to go · tutorial, tabular data, binary classification



Knowledge
9,562 teams

 Featured Prediction Competition

2018 Data Science Bowl

Find the nuclei in divergent images to advance medical discovery



Passion. Curiosity. Purpose.

\$100,000
Prize Money



Booz Allen Hamilton · 774 teams · 3 months to go (2 months to go until merger deadline)

Presented by

Booz Allen | Hamilton & Kaggle

[Overview](#)

[Data](#)

[Kernels](#)

[Discussion](#)

[Leaderboard](#)

[Rules](#)

Overview

Description

Evaluation

Prizes

About

Timeline

Spot Nuclei. Speed Cures.

Imagine speeding up research for almost every disease, from lung cancer and heart disease to rare disorders. The 2018 Data Science Bowl offers our most ambitious mission yet: create an algorithm to automate nucleus detection.

We've all seen people suffer from diseases like cancer, heart disease, chronic obstructive pulmonary disease, Alzheimer's, and diabetes. Many have seen their loved ones pass away. Think how many lives would be transformed if cures came faster.

2018 Data Science Bowl

Find the nuclei in divergent images to advance medical discovery



\$100,000
Prize Money



Booz Allen Hamilton · 774 teams · 3 months to go (2 months to go until merger deadline)

Presented by
Booz Allen | Hamilton | Kaggle

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#)

Overview

Description

Evaluation

Prizes

About

Timeline

This competition is evaluated on the mean average precision at different intersection over union (IoU) thresholds. The IoU of a proposed set of object pixels and a set of true object pixels is calculated as:

$$IoU(A, B) = \frac{A \cap B}{A \cup B}.$$

The metric sweeps over a range of IoU thresholds, at each point calculating an average precision value. The threshold values range from 0.5 to 0.95 with a step size of 0.05: (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95). In other words, at a threshold of 0.5, a predicted object is considered a "hit" if its intersection over union with a ground truth object is greater than 0.5.

Competition Data

 stage1_sample_submis...

stage1_sample_submission.csv.zip 2.62 KB

 Download

 stage1_test.zip

 stage1_train.zip

 stage1_train_labels....

Data Description

This dataset contains a large number of segmented nuclei images. The images were acquired under a variety of conditions and vary in the cell type, magnification, and imaging modality (brightfield vs. fluorescence). The dataset is designed to challenge an algorithm's ability to generalize across these variations.

Each image is represented by an associated `ImageId`. Files belonging to an image are contained in a folder with this `ImageId`. Within this folder are two subfolders:

- `images` contains the image file.
- `masks` contains the segmented masks of each nucleus. This folder is only included in the training set. Each mask contains one nucleus. Masks are not allowed to overlap (no pixel belongs to two masks).

The second stage dataset will contain images from unseen experimental conditions. To deter hand labeling, it will also contain images that are ignored in scoring. The metric used to score this competition requires that your submissions are in run-length encoded format. Please see the evaluation page for details.

As with any human-annotated dataset, you may find various forms of errors in the data. You may manually correct errors you find in the training set. The dataset will not be updated/re-released unless it is determined that there are a large number of systematic errors. The masks of the stage 1 test set will be released with the release of the stage 2 test set.

File descriptions

- `/stage1_train/*` - training set images (images and annotated masks)
- `/stage1_test/*` - stage 1 test set images (images only, you are predicting the masks)

Public

Your Work

Favorites

Sort by Hotness

Outputs

Languages

Types

Search kernels



1



Optimizing Computer Vision Segmentation

7h ago 0.182



Py

0

336



Keras U-Net starter - LB 0.277

8d ago 0.277 tutorial



Py

37

13



Fast, tested RLE and input routines

7h ago pipeline code, image data



Py

7

97



Nuclei Overview to Submission

12d ago 0.154 beginner, eda, image processing, data visualization, cnn



Py

0

51



Data augmentation and Tensorflow U-Net

6d ago image processing



Py

7

35



Basic Pure Computer Vision Segmentation (LB 0.229)

9d ago 0.229 biology, beginner, image processing



Py

2

20



Nucleus Masking with TensorFlow Encoder

10d ago



Py

0

94 topics and kernels

[Subscribe](#)

Sort by

Hotness

All

Mine

Upvoted

Topics & Kernels

Search topics



15



Thread to post data quality issues

[William Cukierski](#) 11 days ago

last comment by

[Anne Carpenter](#) 10h ago

33

27



Availability of the images

[Anne Carpenter](#) 9 days ago

last comment by

[YQM0nk3y](#) 1d ago

9

16



Official External Data Thread

[William Cukierski](#) 13 days ago

last comment by

[Wesley Goi](#) 3h ago

19

80



My notebook

[Allen Goodman](#) 5 days ago

last comment by

[Allen Goodman](#) 3h ago

43

336



Keras U-Net starter - LB 0.277

[Kjetil Åmdal-Sævik](#) last run 12 days ago

last comment by

[Yi Wei](#) 1d ago

37

13



Fast, tested RLE and input routines

[Sam Stainsby](#) last run 4 days ago

last comment by

[Sam Stainsby](#) 7h ago

7

57



[pytorch starter kit] -LB 0.311

[Heng CherKeng](#) 11 days ago

last comment by

[YashKatariya](#) 13h ago

37

Public Leaderboard

Private Leaderboard

This leaderboard is calculated with all of the test data.

[Raw Data](#) [Refresh](#)

■ In the money ■ Gold ■ Silver ■ Bronze

#	Δ1w	Team Name	Kernel	Team Members	Score ?	Entries	Last
1	new	Allen Goodman (not prize elig...			0.634	13	3h
2	—	Malong Tech.			0.508	22	1d
3	▼ 2	ZFTurbo			0.462	12	1d
4	—	Tim Hochberg			0.454	9	4d
5	▼ 2	outrunner			0.425	7	9d
6	▲ 1	Tran Dang Dinh Ang			0.425	22	2d
7	▲ 105	Yuanfang Guan			0.422	2	5d
8	▲ 6	Diogo			0.416	17	1d

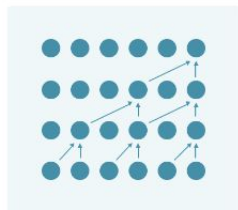


Hands-On Data Science Education

Learn the basics to confidently start a new career or upgrade your skills.



Tracks <https://www.kaggle.com/learn/overview>



Machine Learning

Machine learning is the hottest field in data science, and this track will get you started quickly.



R

Learn the language designed for data analysis. This track includes data set-up, machine learning and data visualization.



Data Visualisation

Visualisation is one of the most versatile skills in data science. Make insightful and beautiful graphics to see what's happening in any dataset.



Deep Learning

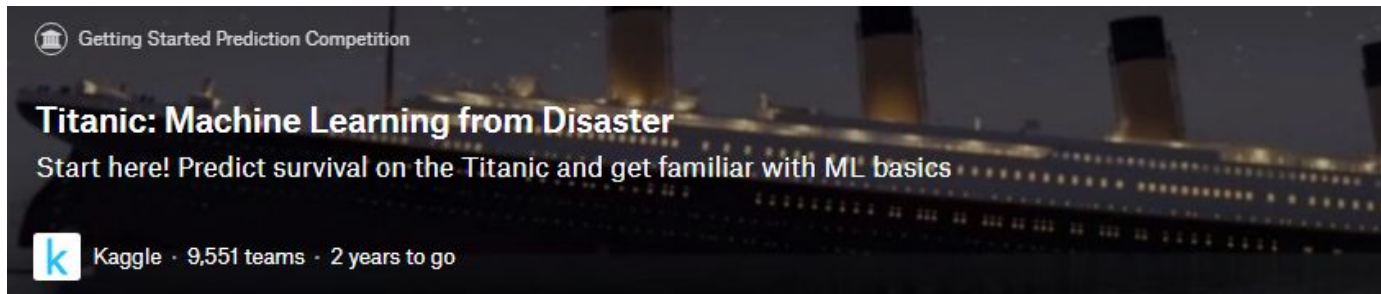
Learn TensorFlow to take Machine Learning to the next level. Your new skills will amaze you.

Q&A



Titanic: Machine Learning from Disaster

- Beginner friendly binary classification problem
- Predict whether a person will survive
- Excellent tutorials in python and R
- Join the Local Team!



Titanic Competition