

kaggle

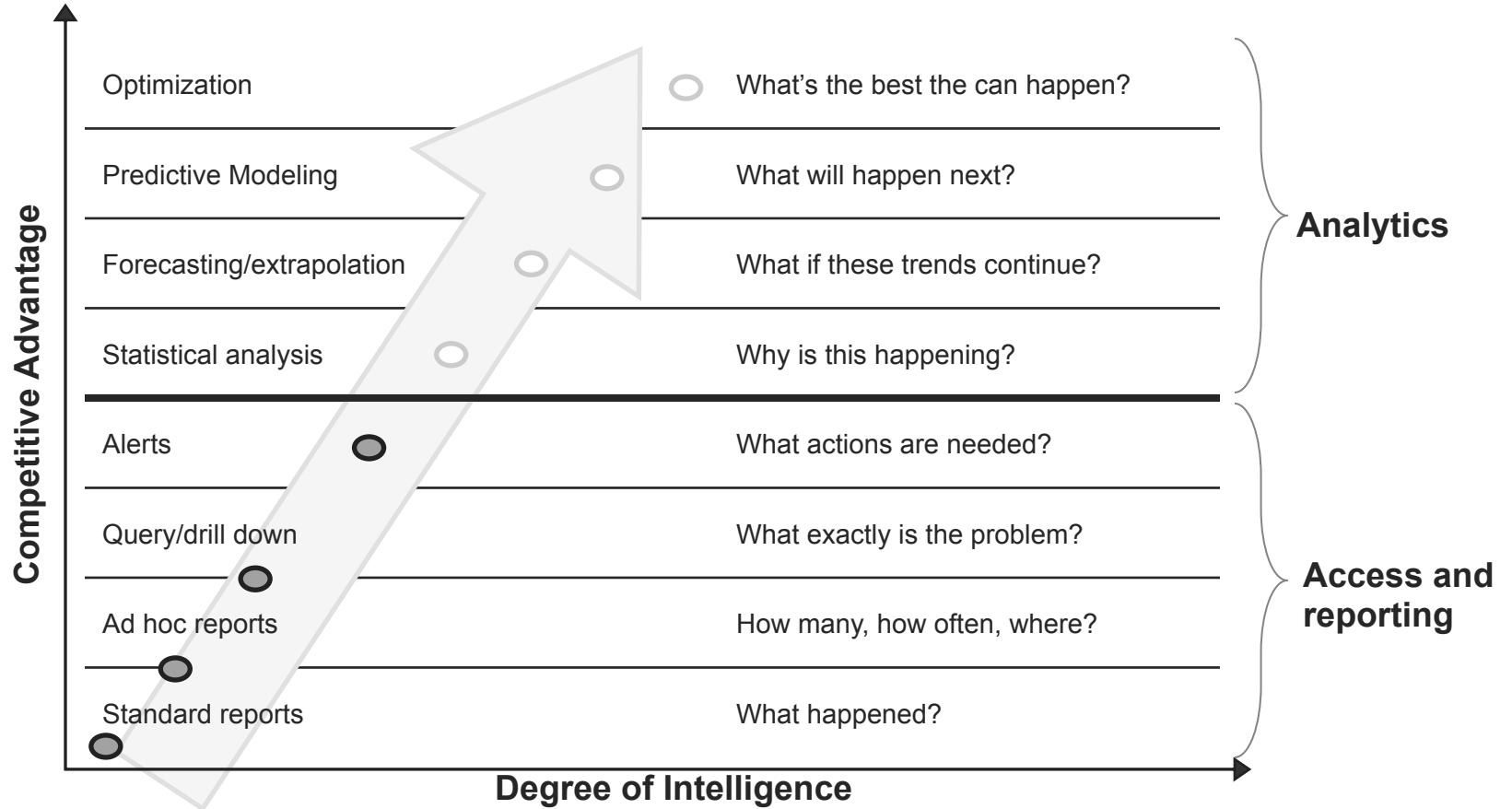


# Crowdsourcing Predictive Analytics: Using 60,000 Heads, without Losing Yours

William Cukierski, PhD  
will.cukierski@kaggle.com



# A predictive science food chain



# The unfortunate hype of predictive science

- Big data!
- Every second 6.2 quintillion exabytes of data are being collected
- Need shared vocabulary, shared scientific protocols
- Need to leverage
  - demographics
  - catastrophe models
  - predictive models
  - economic capital models
  - regulatory information
  - cell phone logs
  - satellite surveillance
  - Etc
  - Etc
  - Etc

CONSULTANTS SAY  
THREE QUINTILLION  
BYTES OF DATA ARE  
CREATED EVERY DAY.



# What do we do with big data?

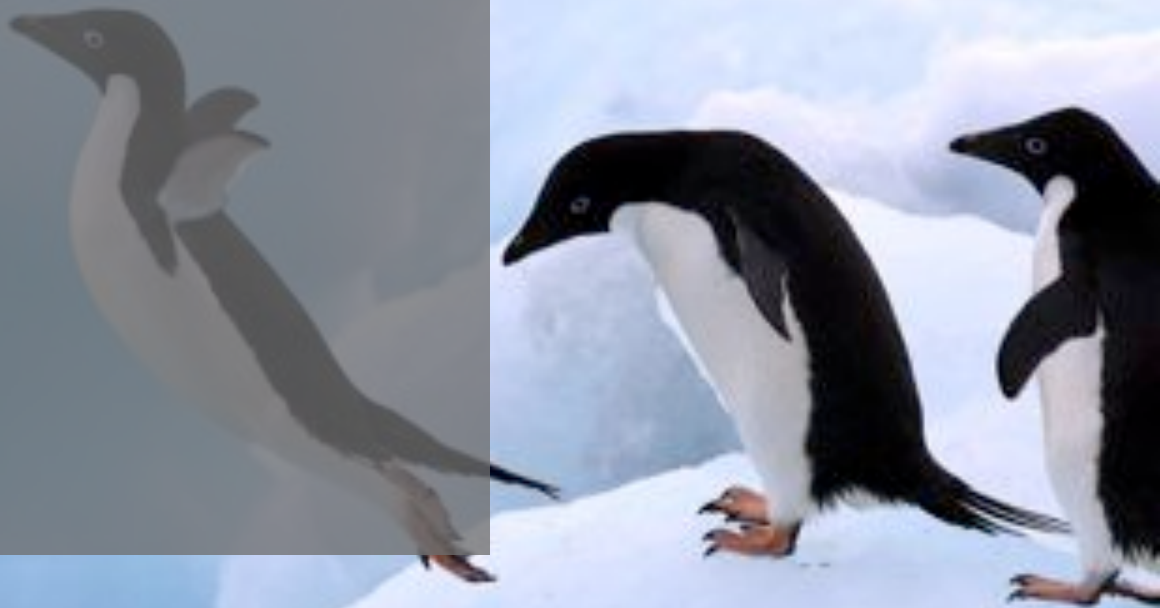
- Create committees, panels, consortiums, taxonomies
- Create acronyms for our committees, panels, consortiums, taxonomies
- Go to conferences to promote and learn about our acronym'd committees, panels, consortiums, taxonomies
- Promise to share, then hoard data and ideas until grant funding cycles make it safe
- And if time permits and the mood strikes?  
Actual work.

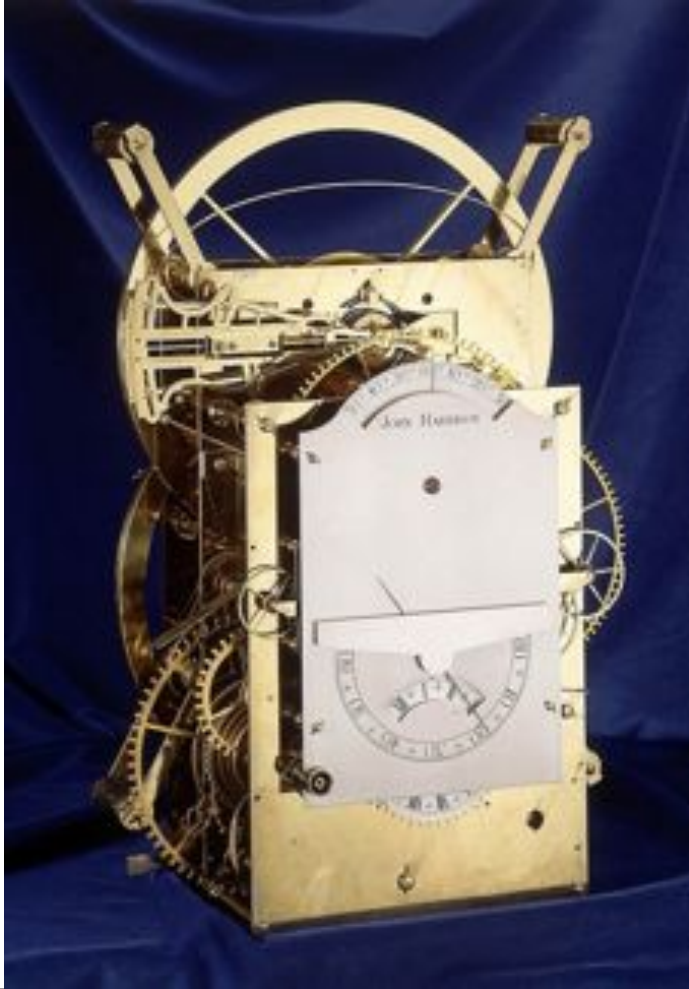
I'm ready to leave now.



*Agenda:*

**Crowdsourcing**  
**What is Kaggle?**  
**Case Studies**  
**FAQs**



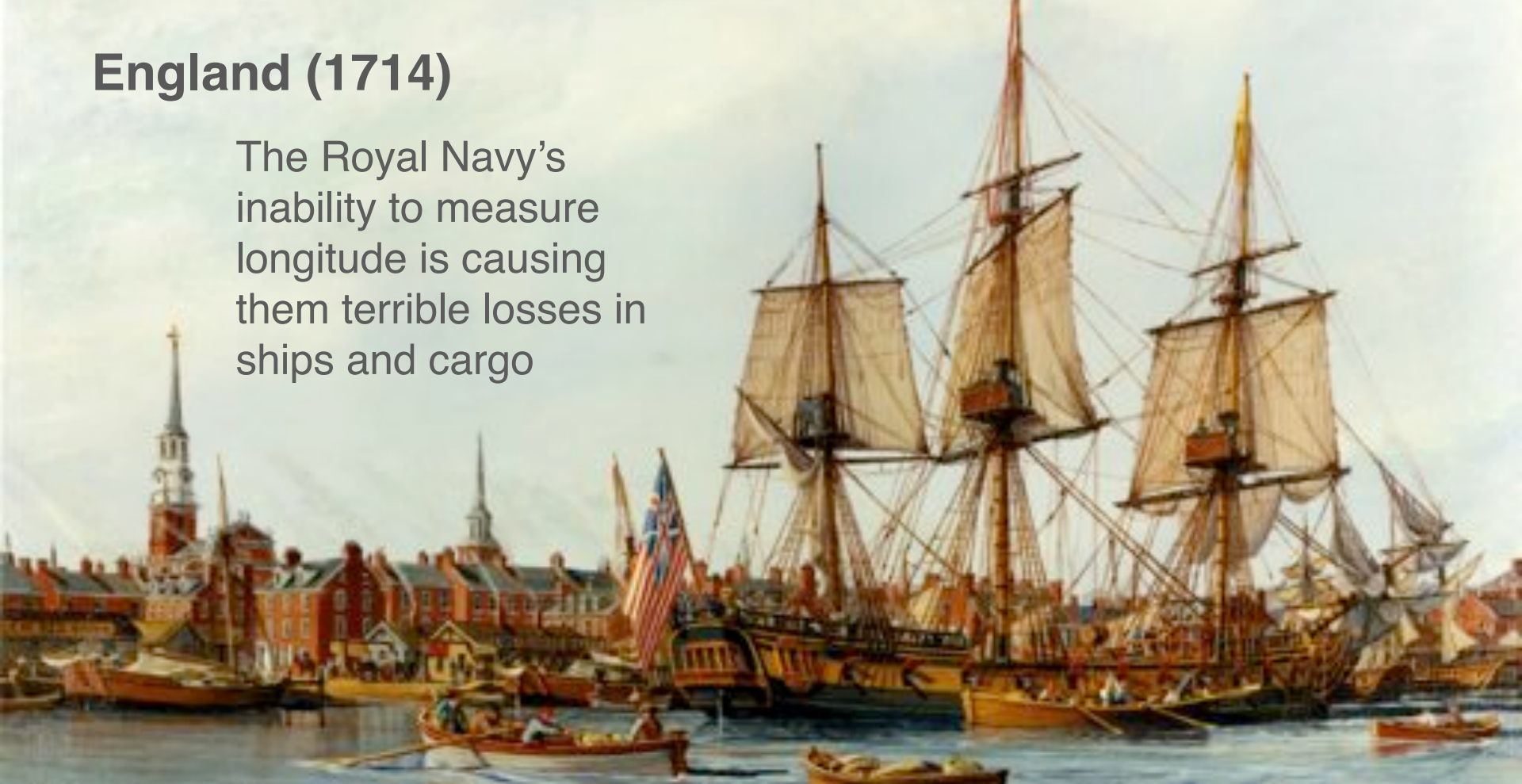


The uncanny efficiency of  
**Using prizes to  
induce the public**



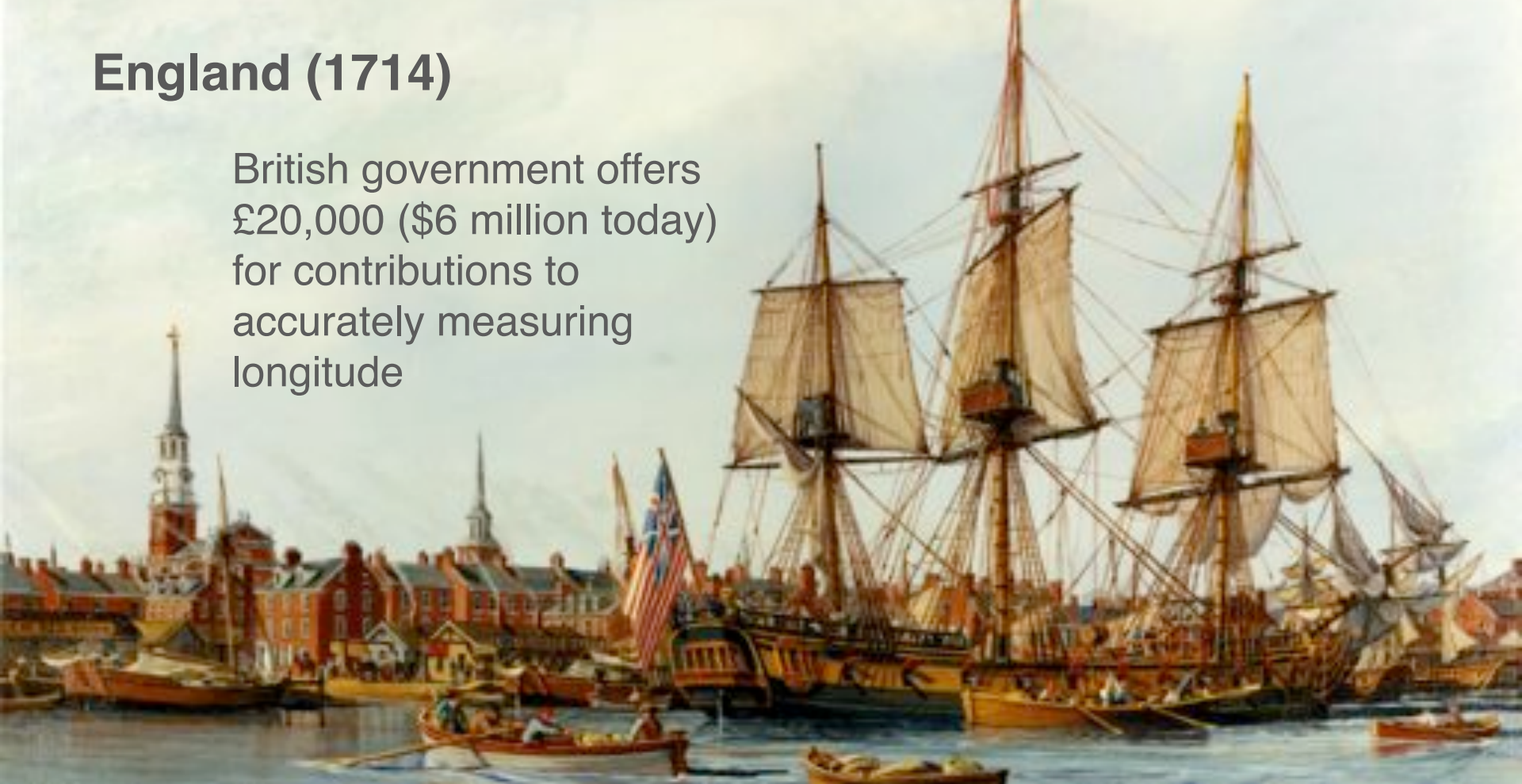
## England (1714)

The Royal Navy's inability to measure longitude is causing them terrible losses in ships and cargo



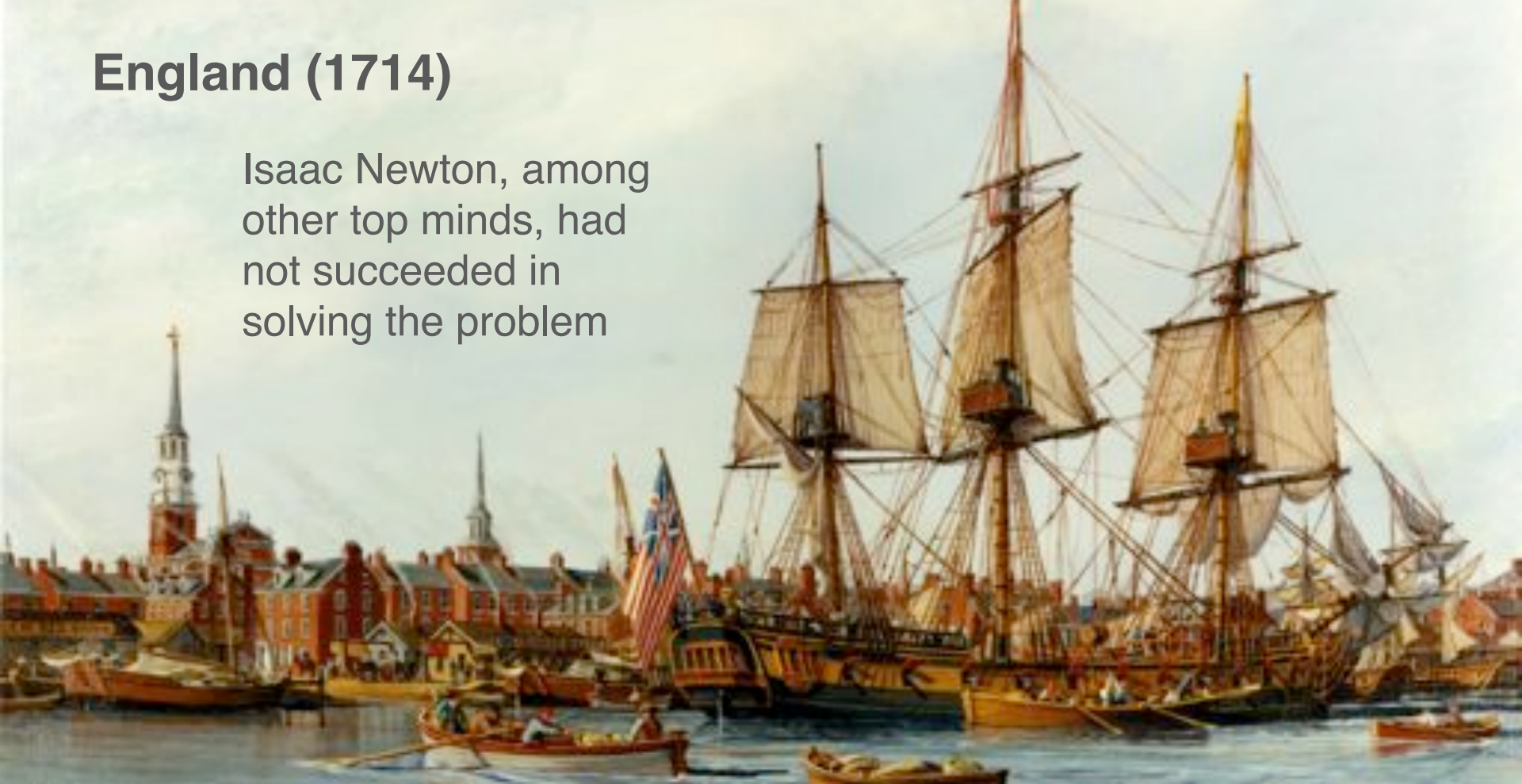
## England (1714)

British government offers  
£20,000 (\$6 million today)  
for contributions to  
accurately measuring  
longitude



## England (1714)

Isaac Newton, among other top minds, had not succeeded in solving the problem



## England (1714)

John Harrison, a cabinetmaker from Yorkshire, develops a clock that maintains accuracy on the seas, claims £14,315



**“No matter who you are, most of the smartest people work for someone else.”**

- Bill Joy, Sun Microsystems co-founder

300 years later...



## United States (2002)

US is experiencing a technology boom, the internet is going well, Public broadcasting is popular and well funded, “Statistician” named hottest job of 2002, life is generally okay



## United States (2002)

FOX announces \$1M prize to find the next great solo recording artists based on viewer voting





## United States (2002)

Causes irreparable damage to the U.S. GDP, reality television and “Next top” shows erode the intellectual capacity of Americans everywhere. Math is cool. But you know what’s really cool? Last night’s episode of Idol.

everything is going to be okay

# Ansari X-Prize

\$10M prize for the first non-government organization to launch a reusable manned spacecraft into space twice within two weeks

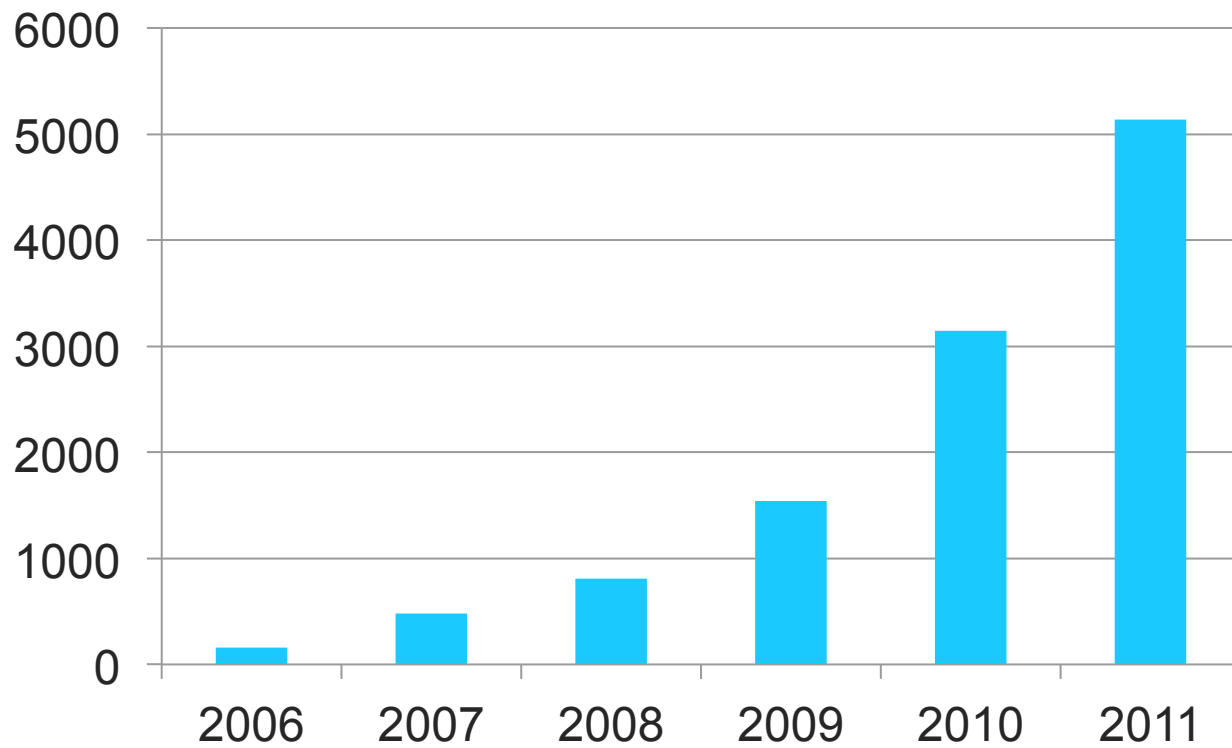


# Ansari X-Prize

Aeronautical experts from  
around the world  
collectively spend \$100M  
and solve the challenge



# Google Scholar Articles Mentioning “Crowdsourcing”





**Crowdfunding**

Financial contributions from online investors, sponsors or donors to fund the growth of new products, initiatives or enterprises.



**Crowd Creativity**

Seeking or creating ideas, content to design and develop original and creative content.



**Tools**

Enabling platforms and tools that support collaboration, communication and sharing among distributed groups of people.



**Distributed Knowledge**

Development of knowledge and/or information resources from distributed and distributed sources.



**Cloud Labor**

Leveraging of a distributed virtual labor pool, available via distributed SaaS or other tools from clients to complete.



**Open Innovation**

Use of diverse worlds of the entity or group to generate, develop and implement ideas.



*"On the Internet, nobody knows you're a dog."*



*“On the Internet, nobody knows you don’t have a PhD in statistics.”*



Which, let's be honest, is a nice way of saying...



# Two crowdsourcing flavors

Using the crowd to do **large, unwieldy, and highly distributed problems**  
("many hands make light work")



Ivory-billed Woodpecker, James John Audubon

Wikipedia  
Ornithology  
Open Source  
SETI  
iStockPhoto  
Mechanical Turk

Using the crowd to solve **singular, focused, difficult problems**  
("two heads are better than one")

Kaggle  
Innocentive  
DARPA  
IARPA  
NASA  
X-Prize Foundation



# Problems with many crowdsourcing initiatives

1. Recognition is rewarded subjectively
  - Leads to a high barrier to entry
2. Recognition is rewarded after proof of work
  - Leads to high sunk costs
3. Crowdsourcing is conflated with outsourcing
  - Ignores the closing gap between professionals and amateurs
4. Failure to appropriately divide complex tasks

## Overstock.com Offers \$1 Million For Improved Recommendations

BY E.B. BOYD | MAY 12, 2011

The online retailer is pulling a Netflix, dangling the promise of a rich reward--not to mention some serious bragging rights--to the team that increases customer purchases.



New Android  
Malware Is A Burglar's  
Best Friend



How To Be A Happy  
And Successful  
Creative Freelancer  
(Or Work With One)



Would You Recognize  
Yourself With A  
Completely  
Symmetrical Face?



- **Peer Review:** On or about April 16, 2012, a **peer review committee appointed by Sponsor** will select the top ten Semi-Finalists based on their **opinions** of the **expected** effectiveness at generating lift and novelty of design of each of the Entries. They will also attempt to ensure their selections reflect a diversity of approaches from among the Entries.”

## Overview

- Semi-  
machin  
recom  
interac  
on the  
more  
measu  
existin  
than S

***\*\*After careful review of all submitted Entries to the Reclab Prize on Overstock.com Contest, the Peer Review Committee has determined that no Entry met the effectiveness at generating lift and novelty of design to be selected for the Semi-Final Stage. Thank you to all the participating teams. At this time the Reclab Prize has concluded.\*\****

- **Final Stage:** Each of the Finalists will be given randomly-chosen 5% of sessions on the Overstock.com website for an additional three (3) weeks. The best performing Entry, as judged by increase in revenue per session over Sponsor’s existing algorithms, will be deemed the winner. **If no Finalist produces more than 1% more revenue per session than Sponsor’s existing algorithms**, then no prize will be awarded.

# How we have attempted to solve these problems

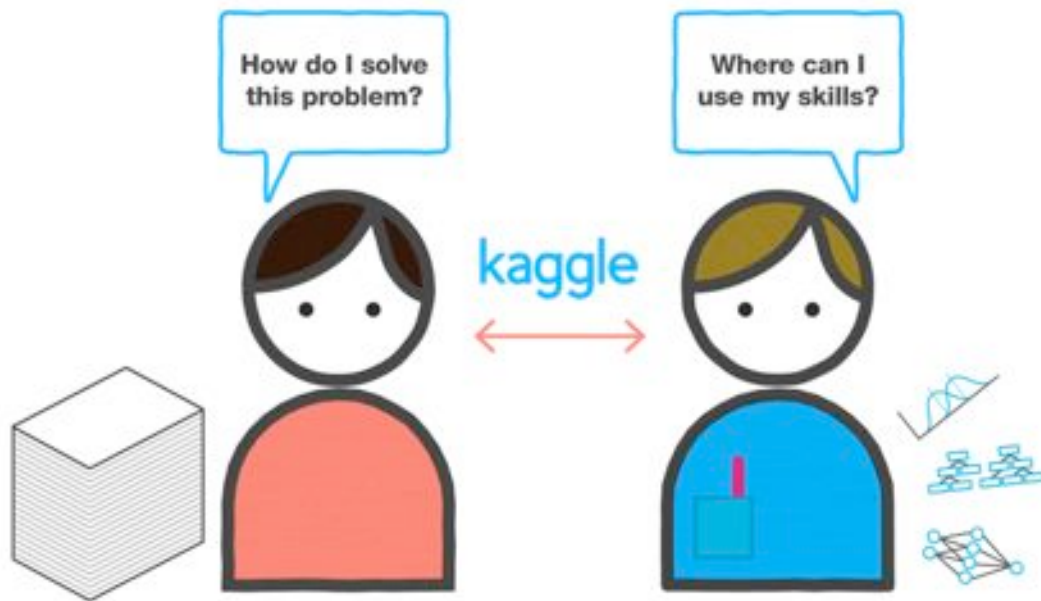
1. Recognition is entirely objective
2. Prizes are established up front and dependent on relative rank, not absolute performance
3. We prioritize the professional interests of our users:
  - Compensation
  - Meritocratic access to job opportunities
  - Education (“learning by doing”, “learning by necessity”)
  - At-cost partnerships with research groups

# Disparity: “Many hands make light work”



**High-quality, high-resolution, digital photos had become ubiquitous, but stock agencies still treated them as a scarce resource.**

# Disparity: “Two heads are better than one”

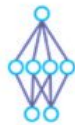


**A mismatch exists between those with data and those with the skills to analyze it**



## We strive NOT to:

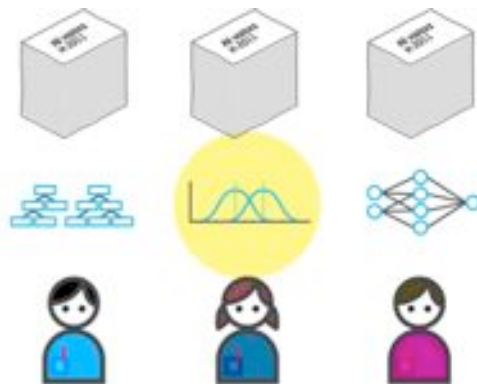
- Be an outsourcing company
- Be the next cloud-based, Hadoop-ready, unstructured-data, scalable, NoSQL, enterprise, insight-leveraging, big-data, analytics platform
- Displace actuaries (or any domain experts)
- Be intimidated by domain biases, presuppositions, or challenges with bad reputations
- Be slow



Why competitions suit

# Predictive scientific problems

# Theory 1: Diversity of Approaches



## Attacking from all sides

There are countless ways to solve any predictive modeling problem. No one person can try them all. By exposing the problem to a large number of participants, all trying different techniques, competitions can very quickly advance the frontier of what's possible using a given dataset.



# Theory 2: Diversity of People

This problem can only be solved by an 8<sup>th</sup>-order kernel projection onto an orthonormal space of homoscedastic eigentensors



The boss is going to have my neck if I can't get this Hadoop iPhone app ready in time for BigDataFest



I'm making an Excel VBA script to access our Oracle database and find the mean of the revenue column!

Data science (noun): Statistics done wrong



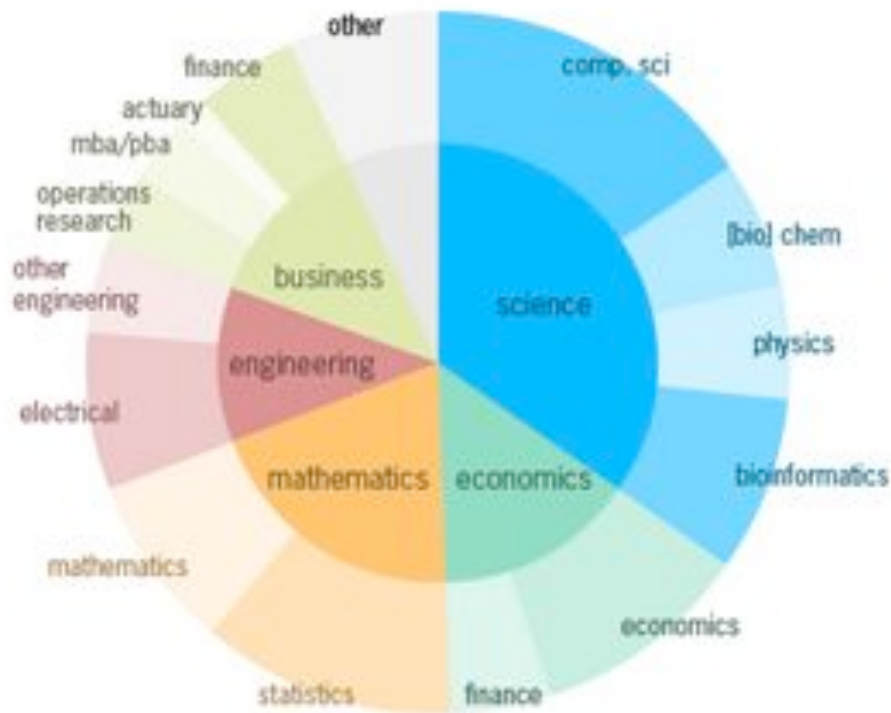
REV. T. BAYES



# 50,000+ registered data scientists



# Diverse Skills



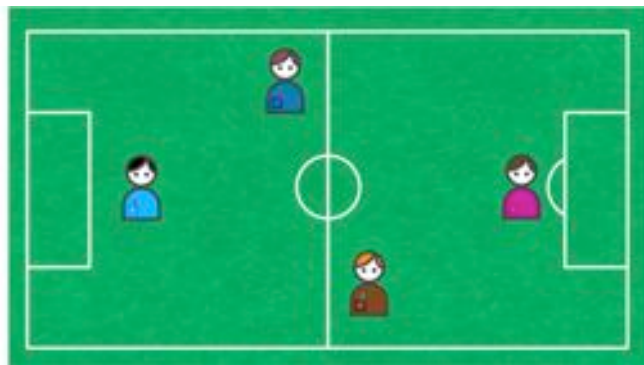
**Our community of data scientists comprises thousands of PhDs** from quantitative fields such as computer science, statistics, econometrics, maths and physics. They come from over 100 countries and 200 universities.

In addition to the prize money and data, they use Kaggle to meet, network and collaborate with experts from related fields.

**“... our interests are more diverse than our business cards would have one believe.”**

- Jeff Howe, *Crowdsourcing*

# Theory 3: Competition Dynamics



## Competitive pressure

drives participants to keep trying new ideas. Real-time feedback is given on a live leaderboard, so when somebody makes a breakthrough, others revise their own algorithms to outdo the leader's performance. This leapfrogging continues until participants reach the full extent of what is possible.





# The leaderboard drives participation

Public Leaderboard

Private Leaderboard

- Objective, meritocratic
  - Reduces the sunk cost dilemma
- Encourages leapfrogging
- Frames a range of acceptable performance

This leaderboard is calculated on approximately 53% of the test data. The final results will be based on the other 47%, so the final standings may be different.  
Reminder: It's against the rules to make submissions through multiple accounts. Contact us if you notice any 'sock-puppets'.

\* in the money

#	Δ1w	Team Name	MAP@3	Entries	Last Submission UTC (Best Submission - Last)
1	-	ACMClass@SJTU *	0.44153	253	Fri, 01 Jun 2012 23:22:46 (-8.6h)
2	↑5	Shanda Innovations *	0.43959	121	Fri, 01 Jun 2012 23:55:19 (-0.1h)
3	-	Steffen Rendle *	0.42909	82	Fri, 01 Jun 2012 23:38:21
4	↑1	FICO Model Builder	0.42811	138	Fri, 01 Jun 2012 23:19:20 (-2.1h)
5	↓3	Medrr	0.42657	267	Fri, 01 Jun 2012 11:08:34 (-0.6h)
6	↓2	SYSU_Wargreymon	0.42644	77	Fri, 01 Jun 2012 15:58:13 (-0.3h)
7	new	mmmsoldier	0.42241	20	Wed, 30 May 2012 17:34:27
8	new	lolirush	0.42116	13	Thu, 31 May 2012 06:53:23 (-37.5h)
9	↑71	BBCC	0.41427	124	Fri, 01 Jun 2012 23:49:01 (-1.3h)

# Performance is relative – 100m Dash

**Usain Bolt**  
9.63 seconds



Sources: "The Complete Book of the Olympics" by David Wallechinsky and Jaime Loucky, International Olympic Committee; Amateur Athletic Association; Photographs: Chang W. Lee/The New York Times, Getty Images, International Olympic Committee  
<http://www.nytimes.com/interactive/2012/08/05/sports/olympics/the-100-meter-dash-one-race-every-medalist-ever.html>

# Theory 4: Good Will Hunting

WIRED MAGAZINE: 16.03

TECH | Biz | MEDIA

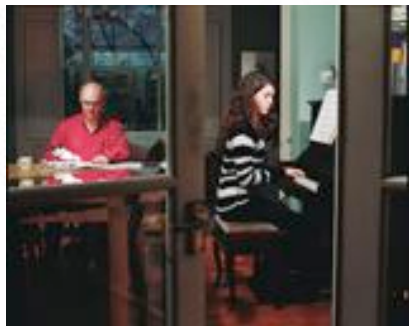
## This Psychologist Might Outsmart the Math Brains Competing for the Netflix Prize

By Jordan Ellenberg | 02.25.08



**At first, it seemed some geeked-out supercoder was going to make an easy million.**

In October 2006, Netflix announced it would give a cool seven figures to whoever created a movie-recommending algorithm 10 percent better than its own. Within two weeks, the DVD rental company had received 169 submissions, including three that were slightly superior to Cinematch, Netflix's recommendation software. After a month, more than a thousand programs had been entered, and the top scorers were almost halfway to the goal.



# Crowdsourcing is *not* replacing domain knowledge

## Domain Expertise

Domain Expertise

Data  
Expertise

Domain +  
Data  
Expertise

Problem

Data

'Crowd'

Knowledge  
& Tools

Model for Prediction



## But domain knowledge appears less important than we thought...

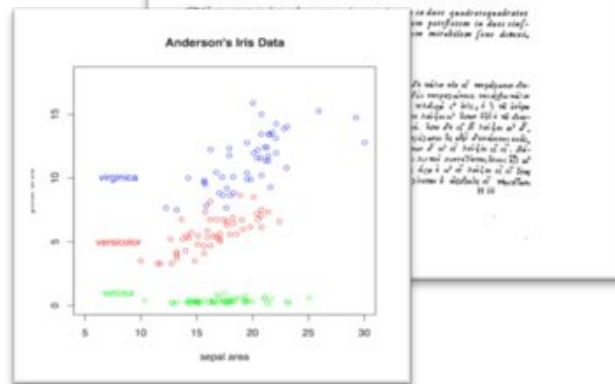
- Karim Lakhani – InnoCentive study








- 166 posted problems, 26 research labs, 4 year timespan
  - The more diverse the problem-solving population, the more likely the problem is to be solved
  - No significant correlation between prize amount and a problem's likelihood of being solved
  - The further the problem was from a solver's expertise, the more likely he or she was to solve it
- On Kaggle:
    - We observe less domain expertise, more “data literacy” skills dominating

# Running a market of competitions is not yet a science (but we're trying to get there...)



- **Problem too easy?** (Iris data) Saturated leaderboard, decimal-place showdown
- **Problem too hard?** (Fermat's Last Thm.) discouragement
- **Too few problems?** competitions hypercompetitive, beginners lose interest
- **Too many problems?** crowd is overwhelmed, participation spread thinly
- **Data too large?** people complain
- **Data too small?** people complain



Featured Competitions		<a href="#">Browse all</a>
	<b>Heritage Health Prize</b> Identify patients who will be admitted to a hospital within the next year, using historical claims data.	Ends 6 months 1339 teams \$3 million
	<b>Merck Molecular Activity Challenge</b> Help develop safe and effective medicines by predicting molecular activity.	Ends 18 days 184 teams \$40,000
	<b>U.S. Census Return Rate Challenge</b> Predict census mail return rates.	Ends 23 days 97 teams \$25,000
	<b>Job Recommendation Engine Challenge</b> Predict which job users will apply to.	Ends 8.1 days 72 teams \$20,000
	<b>Predict Closed Questions on Stack Overflow</b> Predict which new questions asked on Stack Overflow will be closed.	Ends 10 days 117 teams \$20,000

A diverse range of organizations have used Kaggle to

# Improve the state of the art

Research Competitions		<a href="#">Browse all</a>
	<b>Global Energy Forecasting Competition 2012 - Load Forecasting</b> A hierarchical load forecasting problem: backcasting and forecasting hourly loads (in MW) for a US utility with 20 zones.	Ends 33 days 42 teams \$7,500
	<b>Global Energy Forecasting Competition 2012 - Wind Forecasting</b> A wind power forecasting problem: predicting hourly power generation up to 48 hours ahead at 7 wind farms.	Ends 33 days 48 teams \$7,500

# Public Competitions



**Titanic: Machine Learning from Disaster**  
Getting Started Competition, with tutorials in Excel, Python and Introduction to Random Forests.



**Digit Recognizer**  
Classify handwritten digits in this "Getting Started" competition.



**Practice Fusion Analyze This! 2012 - Prediction Challenge**  
Start digging into electronic health records and submit your ideas for the most promising, impactful or interesting predictive modeling competitions



**Facebook Recruiting Competition**  
Data Scientist at Facebook  
Multiple Locations



**ICFHR 2012 - Arabic Writer Identification**  
Identify which writer wrote which documents.



**Semi-Supervised Feature Learning**  
There's been a lot of recent work done in unsupervised feature learning for classification and there are a ton of older methods that also work well. The purpose of this competition is to find out which of these methods work best on relatively large-scale high dimensional learning tasks.



**Automated Essay Scoring**  
Develop an automated scoring algorithm for student-written essays.



**EMC Data Science Global Hackathon (Air Quality Prediction)**  
Build a local early warning systems to accurately predict dangerous levels of air pollutants on an hourly basis.



**Photo Quality Prediction**  
Given anonymized information on thousands of photo albums, predict whether a human evaluator would mark them as 'good'.



**CHALEARN Gesture Challenge 2**  
Develop a Gesture Recognizer for Microsoft Kinect (TM)



**EMC Israel Data Science Challenge**  
Match source code from a project



**Predicting a Biological Response**  
Predict a biological response of molecules from their chemical properties



**Job Recommendation Engine Challenge**  
Predict which jobs users will apply to



**The Hewlett Foundation: Short Answer Scoring**  
Develop a scoring algorithm for student-written short-answer responses.



**Practice Fusion Analyze This! 2012 - Open Challenge**  
Start digging into electronic health records and submit your creative, insightful, and visually striking analyses.



**Harvard Business Review 'Vision Statement' Prospect**  
Your Analysis and/or Visualization featured in the Harvard Business Review



**Data Mining Hackathon on BIG DATA (7GB) Best Buy mobile web site**  
Predict which BestBuy product a mobile web visitor will be most interested in based on their search query or behavior over 2 years (7 GB).



**KDD Cup 2012, Track 1**  
Predict which users (for information sources) one user might follow in Tencent Weibo.



**R Package Recommendation Engine**  
The aim of this competition is to develop a recommendation engine for R libraries (or packages). (R is opensource statistics software)



**Heritage Health Prize**  
Identify patients who will be admitted within the next year, using historical claims data.



**Practice Fusion Diabetes Classification**  
Identify patients diagnosed with Type 2 Diabetes

**Algorithmic Trading Challenge**  
Develop new models to accurately predict the market response to large trades.



**Million Song Dataset Challenge**  
Predict which songs a user will listen to.



**Competition 2012 - Load Forecasting**  
A hierarchical load forecasting problem: backcasting and forecasting hourly loads (in kW) for a US utility with 20 zones.



**Benchmark Bond Trade Price Challenge**  
Develop models to accurately predict the trade price of a bond.



**Give Me Some Credit**  
Improve on the state of the art in credit scoring by predicting the probability that somebody will experience financial distress in the next two years.



**Data Mining Hackathon on (20 mb) Best Buy mobile web site - ACM SF Bay Area Chapter**  
Getting Started - Predict which Xbox game a visitor will be most interested in based on their search query. (20 MB)



**EMI Music Data Science Hackathon - July 21st - 24 hours**  
Can you predict if a listener will love a new song?



**CPROD: Consumer PRODUCTS contest #1**  
Identify product mentions within a largely user-generated web-based corpus and disambiguate mentions against a large product catalogue.



**Global Energy Forecasting Competition 2012 - Wind Forecasting**  
A wind power forecasting problem: pre hourly power generation up to 48 hour: wind farms



**Don't Overfit!**  
With nearly as many variables as training cases, what are the best techniques to avoid disaster?



**Mapping Dark Matter**  
Supported by NASA and the Royal Astronomical Society. A cosmological image analysis competition to measure the small distortion in galaxy images caused by dark matter. The prize is an expenses paid visit to the NASA Jet Propulsion Laboratory (JPL).



**Wikipedia's Participation Challenge**  
This competition challenges data-mining experts to build a predictive model that predicts the number of edits an editor will make five months from the end date of the training dataset.



**Raising Money to Fund an Organizational Mission**  
Help worthy organizations more efficiently target and recruit loyal donors to support their causes.



**Detecting Insights in Social Commentary**  
Principal Data Scientist at Imperium Corporation  
Wood City, CA



**Psychopathy Prediction Based on Twitter Usage**  
Identify people who have a high degree of Psychopathy based on Twitter usage.



**What Do You Know?**  
Improve the state of the art in student evaluation by predicting whether a student will answer the next test question correctly.



**Claim Prediction Challenge (Allstate)**  
A key part of insurance is charging each customer the appropriate price for the risk they represent. Risk varies widely from customer to customer, and a deep understanding of different risk factors helps predict the likelihood and cost of insurance claims. The goal of this competition is to better predict Bodily Injury Liability Insurance claim payments based on the characteristics of the insured customer's vehicles.



**Stay Alert! The Ford Challenge**  
Driving while not alert can be deadly. The objective is to design a classifier that will detect whether the driver is alert or not alert, employing data that are acquired while driving.



**Predict Closed Questions on Stack Overflow**  
Predict which new questions asked on Stack Overflow will be closed



**Follow the Money: Investigative Reporting Prospect**  
Find hidden patterns, connections, and ultimately compelling stories in a treasure trove of data about US federal campaign contributions



**RTA Freeway Travel Time Prediction**  
This competition requires participants to predict travel time on Sydney's M4 freeway from past travel time observations.



**Eye Movements Verification and Identification Competition**  
Determine how people may be identified based on their eye movement characteristics.



**Predict HIV Progression**  
This contest requires competitors to predict the likelihood that an HIV patient's infection will become less severe, given a small dataset and limited clinical information.



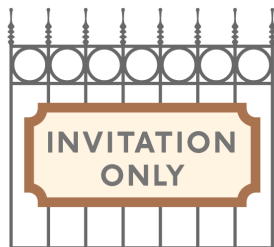
**Don't Get Predicted if a car**



# Competition Types



Public



Private



Recruitment



Prospect



Boehringer  
Ingelheim



**\*6000+**

w/ 1,776 characteristics each



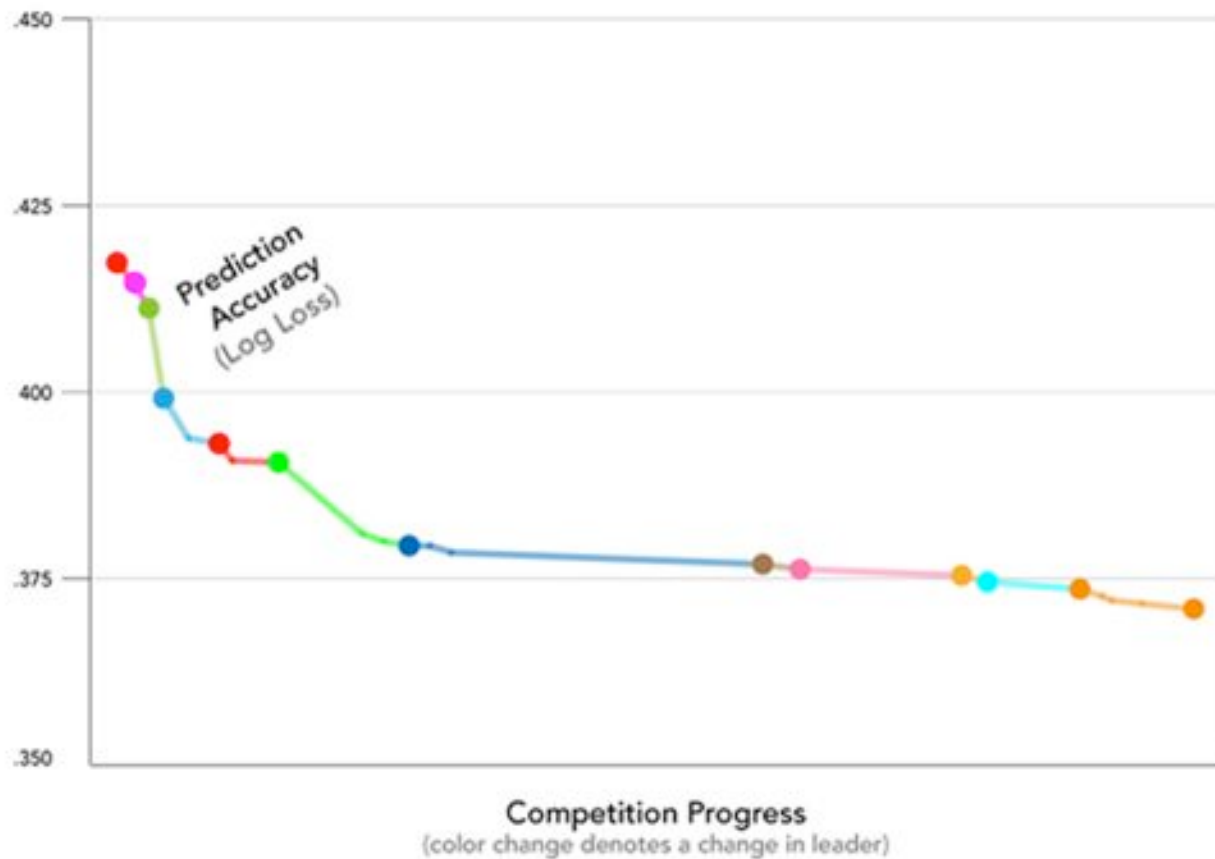
796 entrants

703 teams

8,841 submissions

over 91 days

25.6% improvement over  
previous accuracy benchmark



dunnhumby

## Predicting Grocery Shoppers' Spending Habits

**Grocery shopping: we all have to do it, but can you predict it?** Dunnhumby, a U.K. firm that does analytics for supermarket chains, was looking to build a model to predict when supermarket shoppers will next visit the store and how much they will spend.



537 players in 287 teams + \$10k in prize money = 208% improvement to prediction

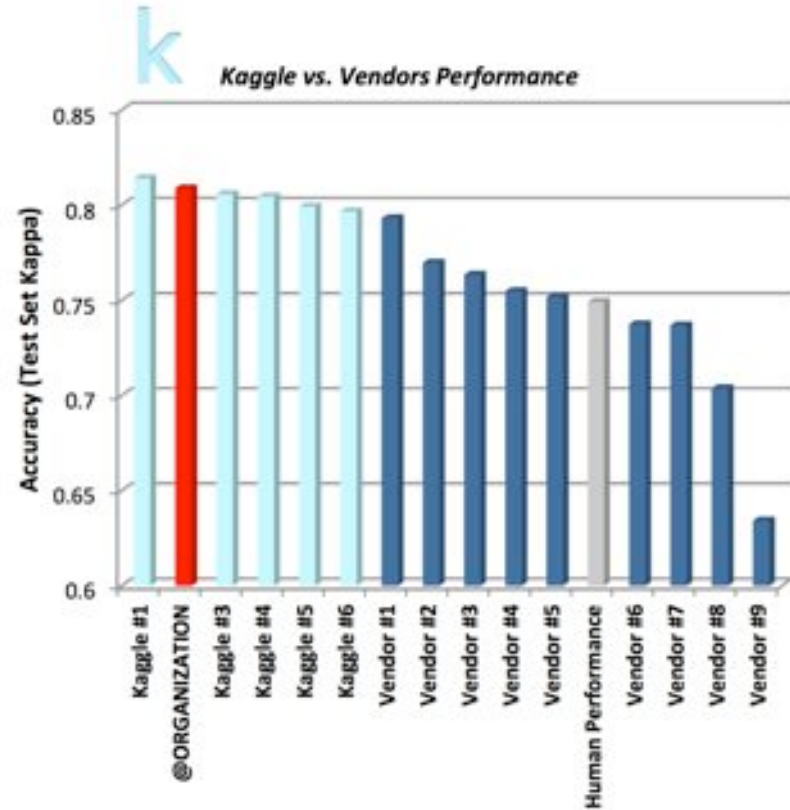


Automated Student Assessment Prize  
*Phase One: Automated Essay Scoring*

### SAMPLE ESSAY PROMPT

We all understand the benefits of laughter. For example, someone once said, “Laughter is the shortest distance between two people.”

Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter was one element or part.

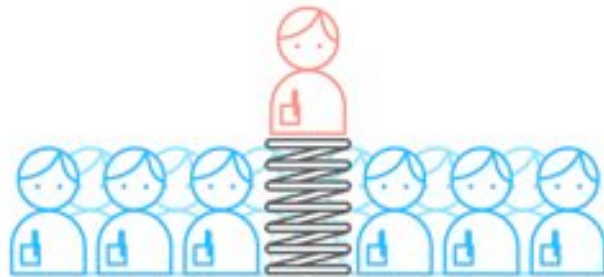
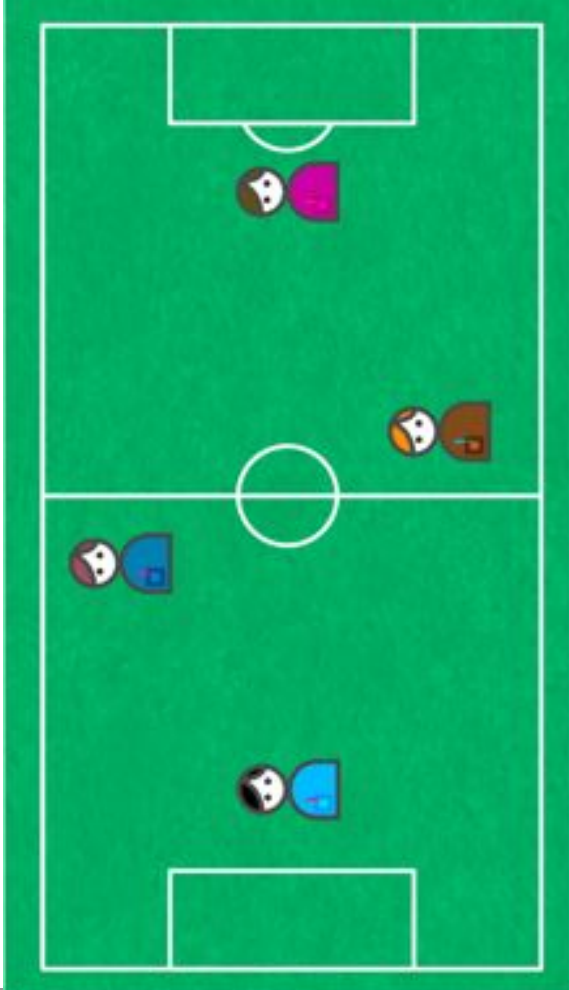


# Recruiting competitions



**Want an interview at Facebook? Facebook will review the top entries in the competition and offer you an interview if they like what they see.**

- Within a hour of posting, competition page had 750 simultaneous unique users
- 422 individuals competed
- 1 hired, several in consideration, many are moving through the interview process



Competition dynamics

**Give insight into the data**

# Anatomy of a public competition

Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
0.64	78023	Household	229.9	Brazil
0.72	12340	Audio	19.95	Mexico
0.41	31240	Computer	6.99	Taiwan
1.94	54323	Hardware	11.99	Taiwan
0.023	92356	Household	2.05	USA
0.08	78023	Computer	99.99	USA
2.09	12340	Computer	129.99	China
1.1	31240	Audio	18.99	China



# Anatomy of a public competition

Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
0.64	78023	Household	229.9	Brazil
0.72	12340	Audio	19.95	Mexico
0.41	31240	Computer	6.99	Taiwan
1.94	54323	Hardware	11.99	Taiwan
0.023	92356	Household	2.05	USA
0.08	78023	Computer	99.99	USA
2.09	12340	Computer	129.99	China
1.1	31240	Audio	18.99	China

Solution  
"Ground Truth"

Training

Test

# Anatomy of a public competition

Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
0.64	78023	Household	229.9	Brazil
?	12340	Audio	19.95	Mexico
?	31240	Computer	6.99	Taiwan
?	54323	Hardware	11.99	Taiwan
?	92356	Household	2.05	USA
?	78023	Computer	99.99	USA
?	12340	Computer	129.99	China
?	31240	Audio	18.99	China

Solution "Ground Truth"

Training

Test

# Anatomy of a public competition



Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
0.64	78023	Household	229.9	Brazil
0.03	12340	Audio	19.95	Mexico
1.298	31240	Computer	6.99	Taiwan
0.94	54323	Hardware	11.99	Taiwan
0.04	92356	Household	2.05	USA
0.36	78023	Computer	99.99	USA
1.2	12340	Computer	129.99	China
0.02	31240	Audio	18.99	China

Training

Test

Submission

# Anatomy of a public competition

 Public Leaderboard  
 Private Leaderboard

Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
0.64	78023	Household	229.9	Brazil
0.03	12340	Audio	19.95	Mexico
1.298	31240	Computer	6.99	Taiwan
0.94	54323	Hardware	11.99	Taiwan
0.04	92356	Household	2.05	USA
0.36	78023	Computer	99.99	USA
1.2	12340	Computer	129.99	China
0.02	31240	Audio	18.99	China

Training

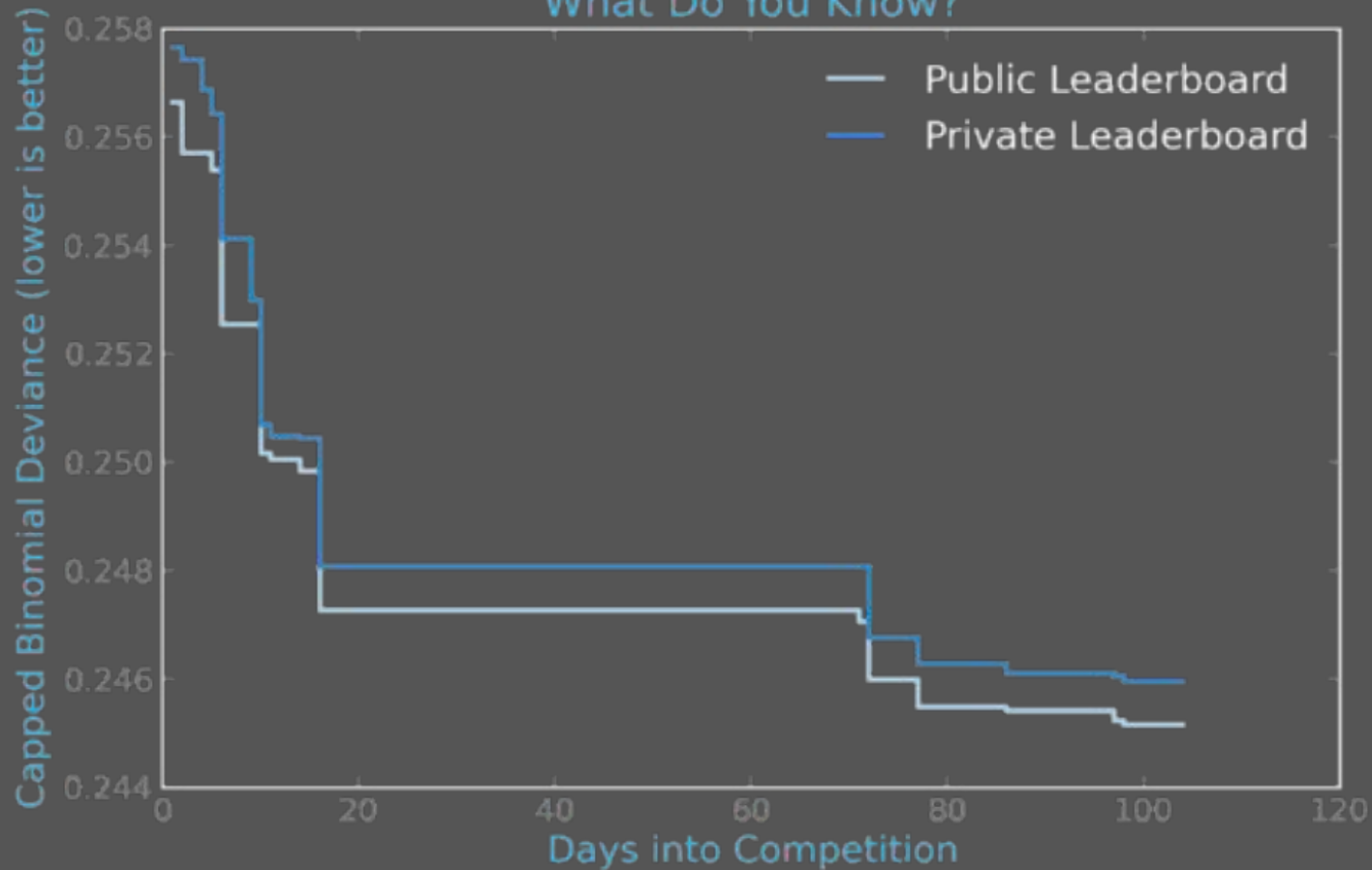
Test

Submission

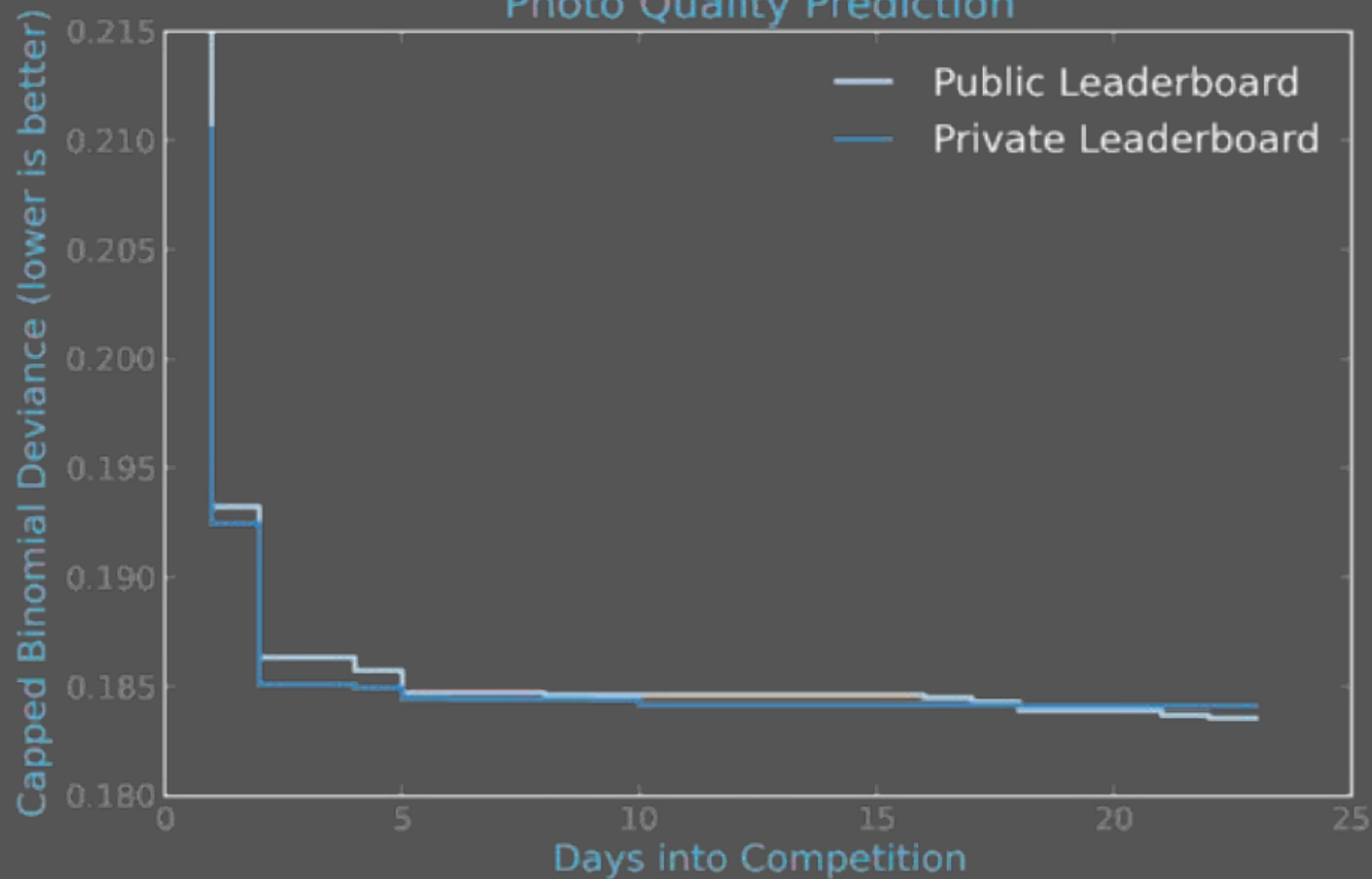
## Dynamics help us infer:

- The orthogonality of approaches
- How much duplicated work is being done
- The extent of overfitting
- How close participants are to the predictive “frontier”

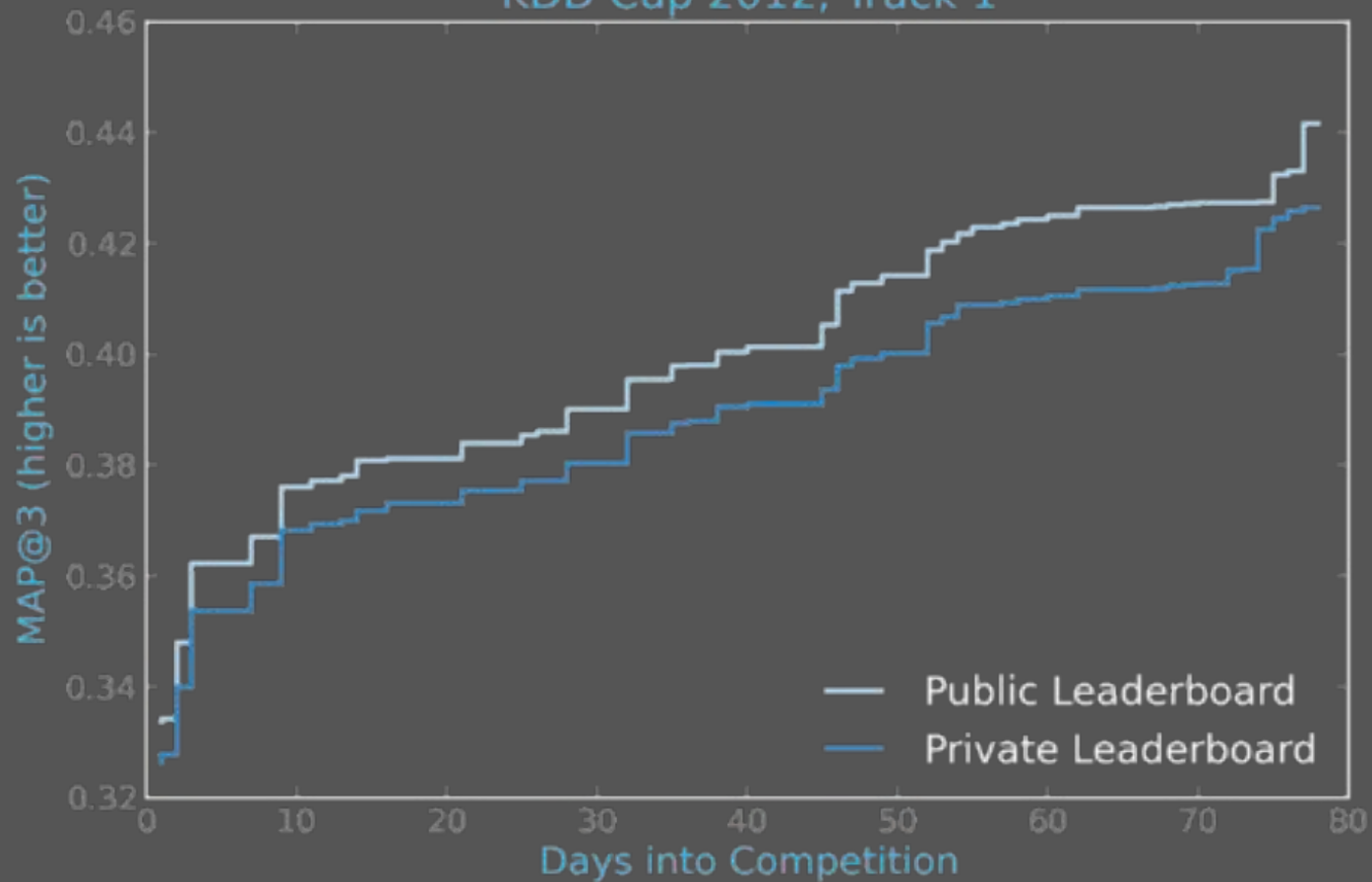
## What Do You Know?



# Photo Quality Prediction

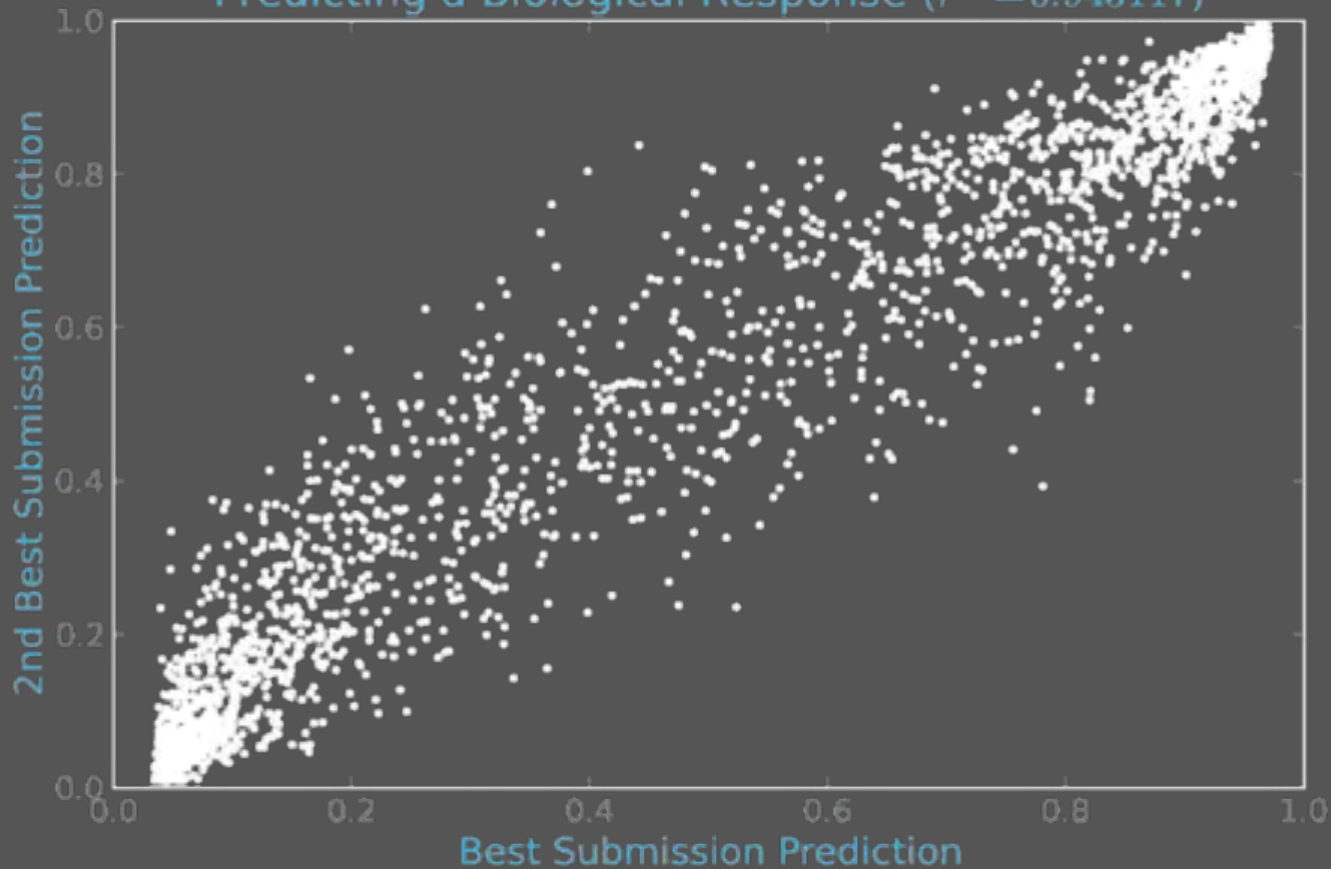


## KDD Cup 2012, Track 1

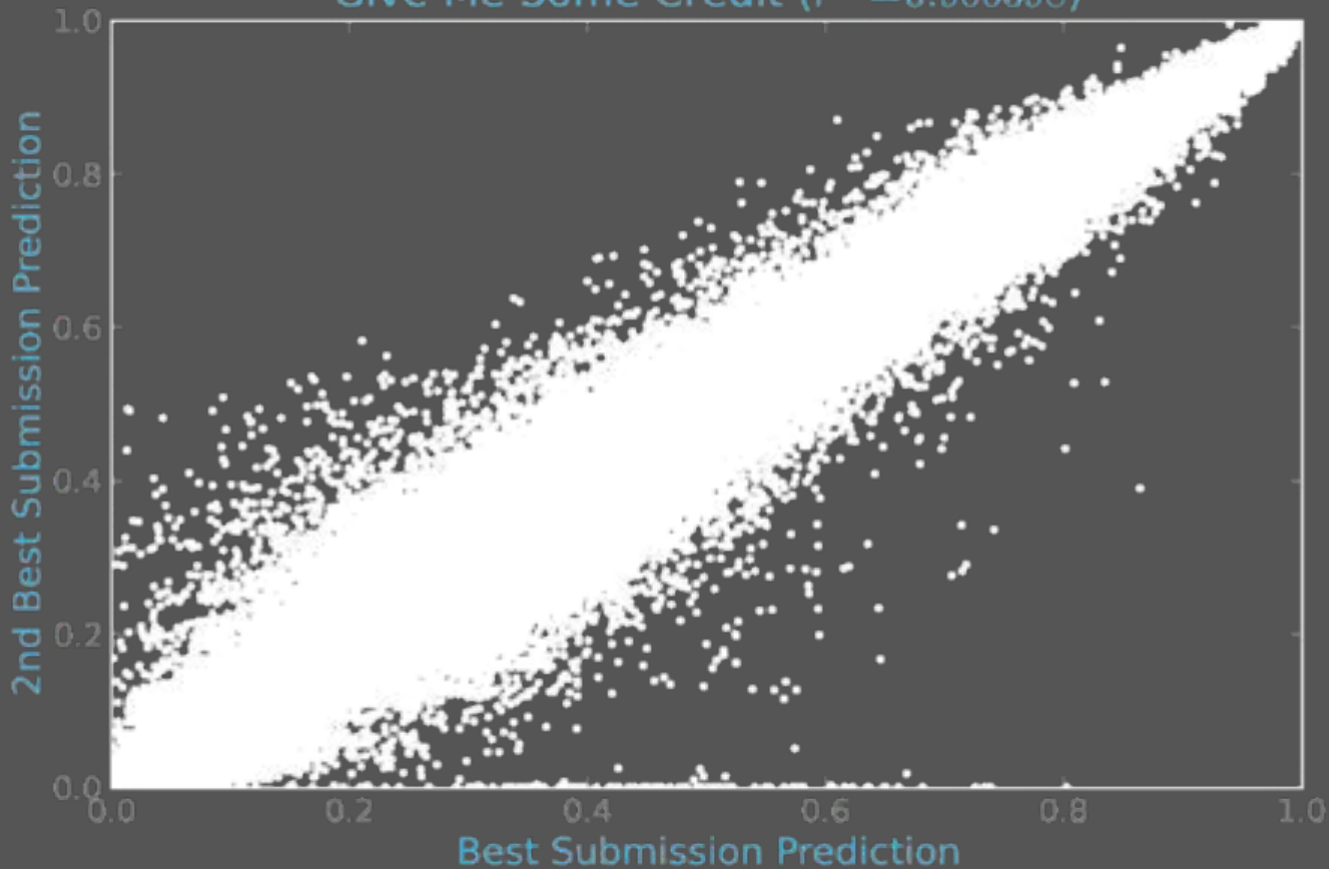




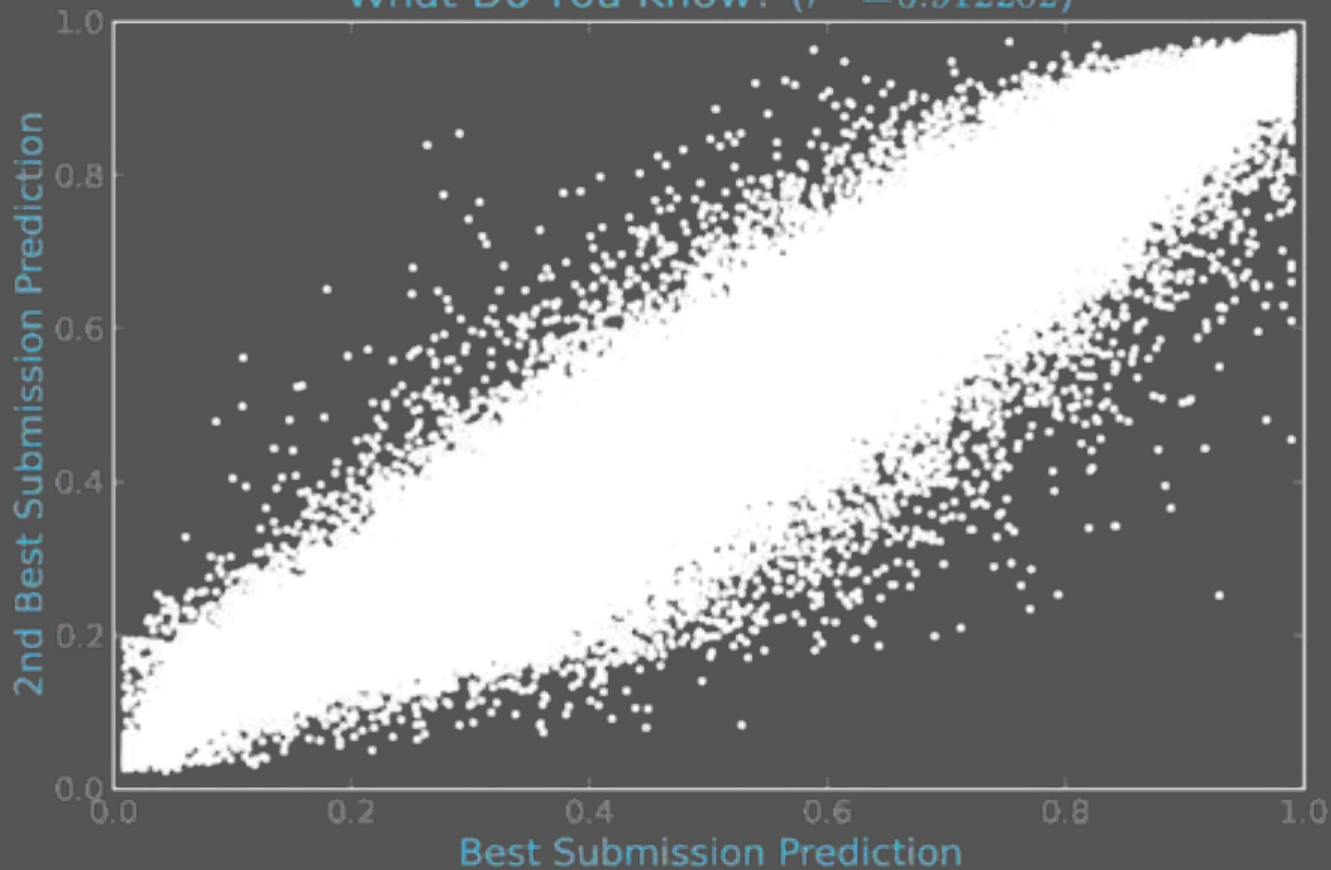
Predicting a Biological Response ( $r^2 = 0.943117$ )



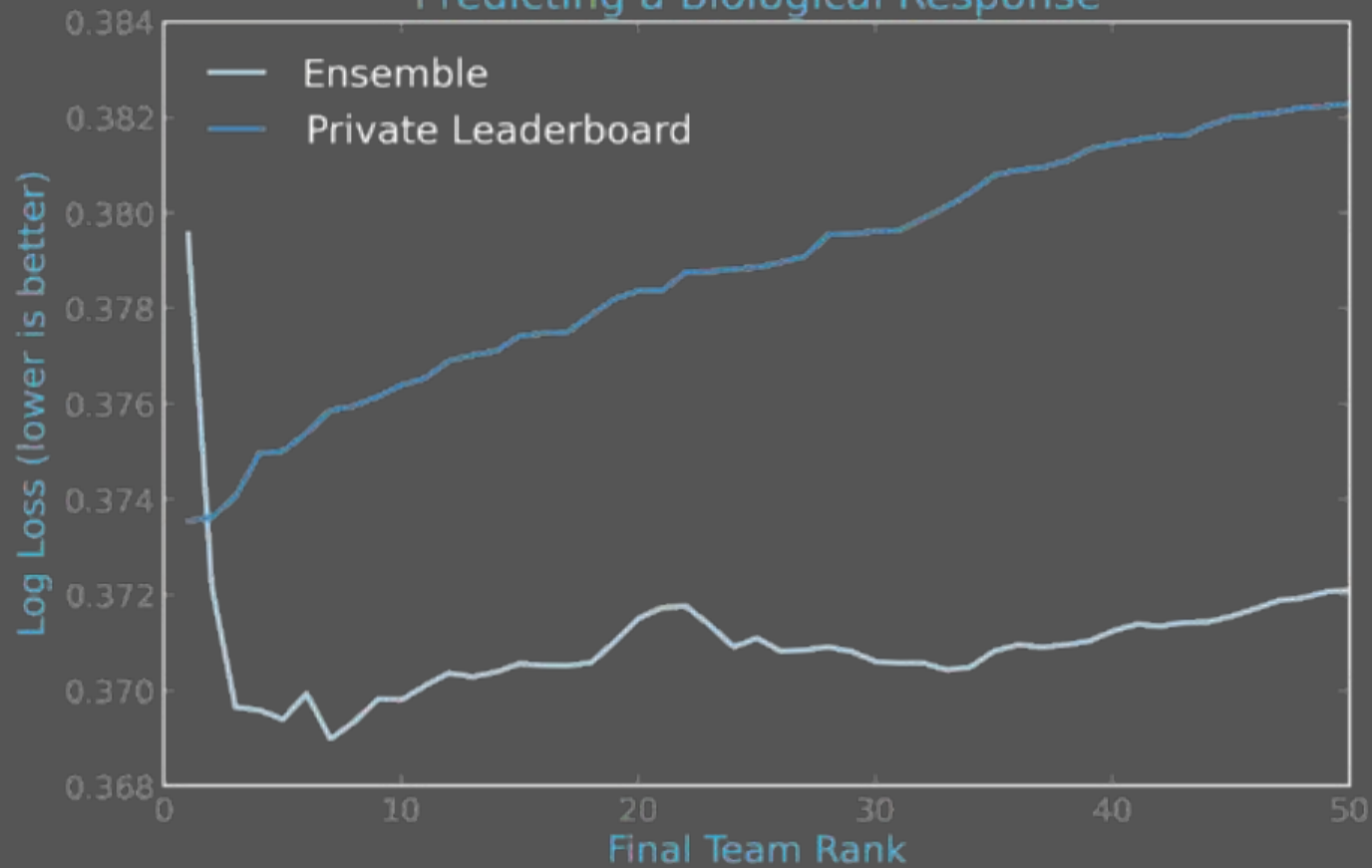
Give Me Some Credit ( $r^2 = 0.966398$ )



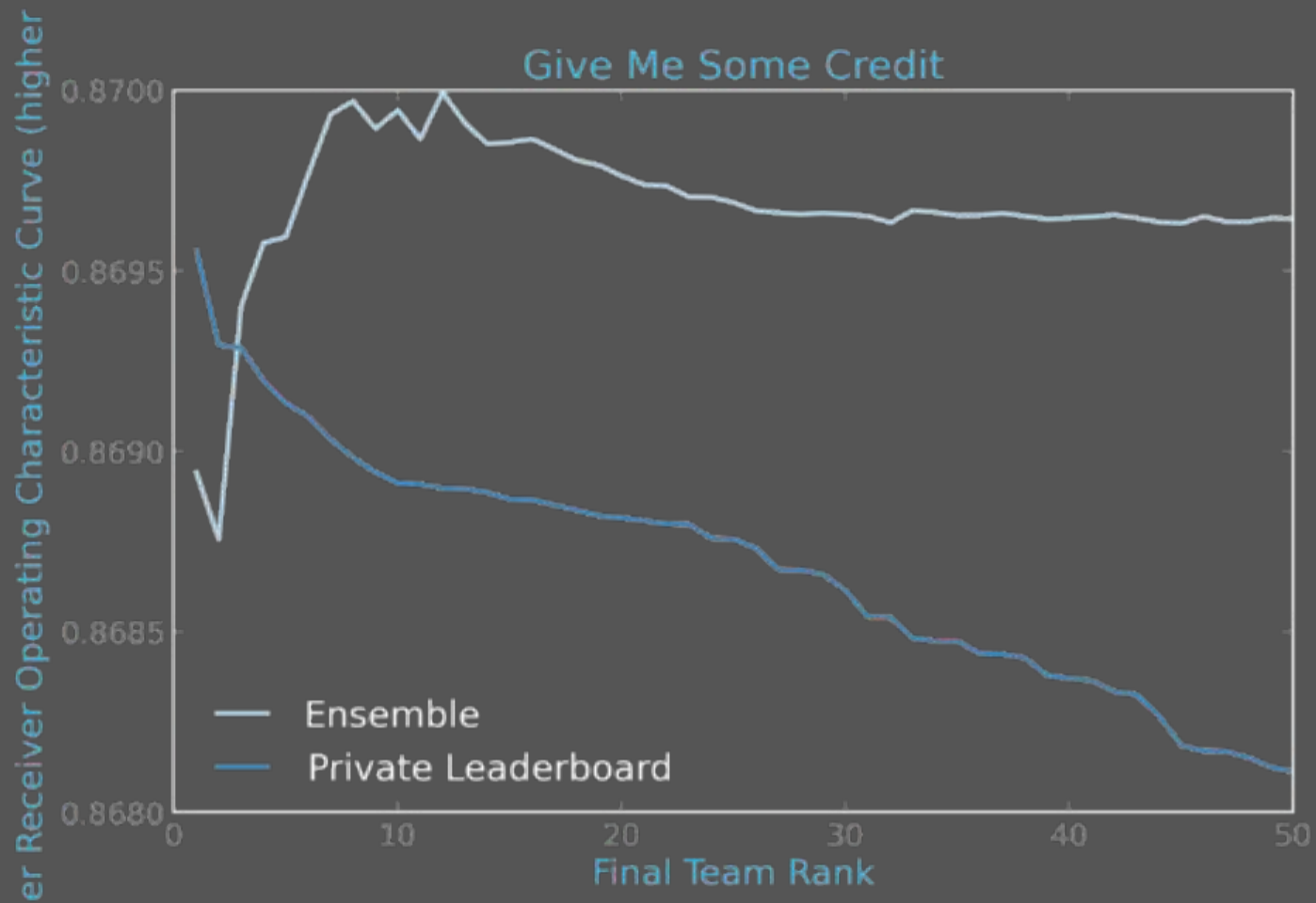
What Do You Know? ( $r^2 = 0.912262$ )



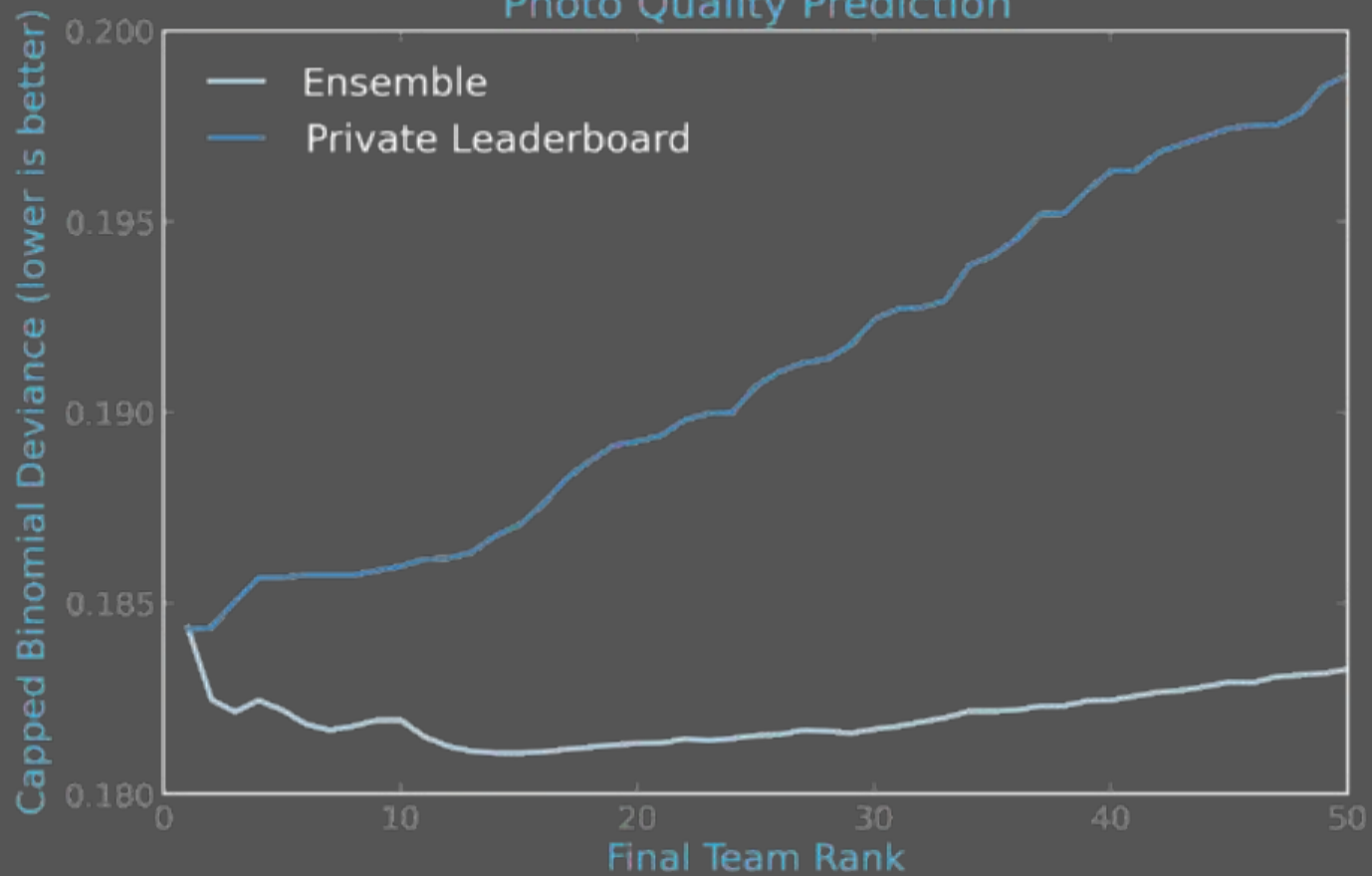
## Predicting a Biological Response



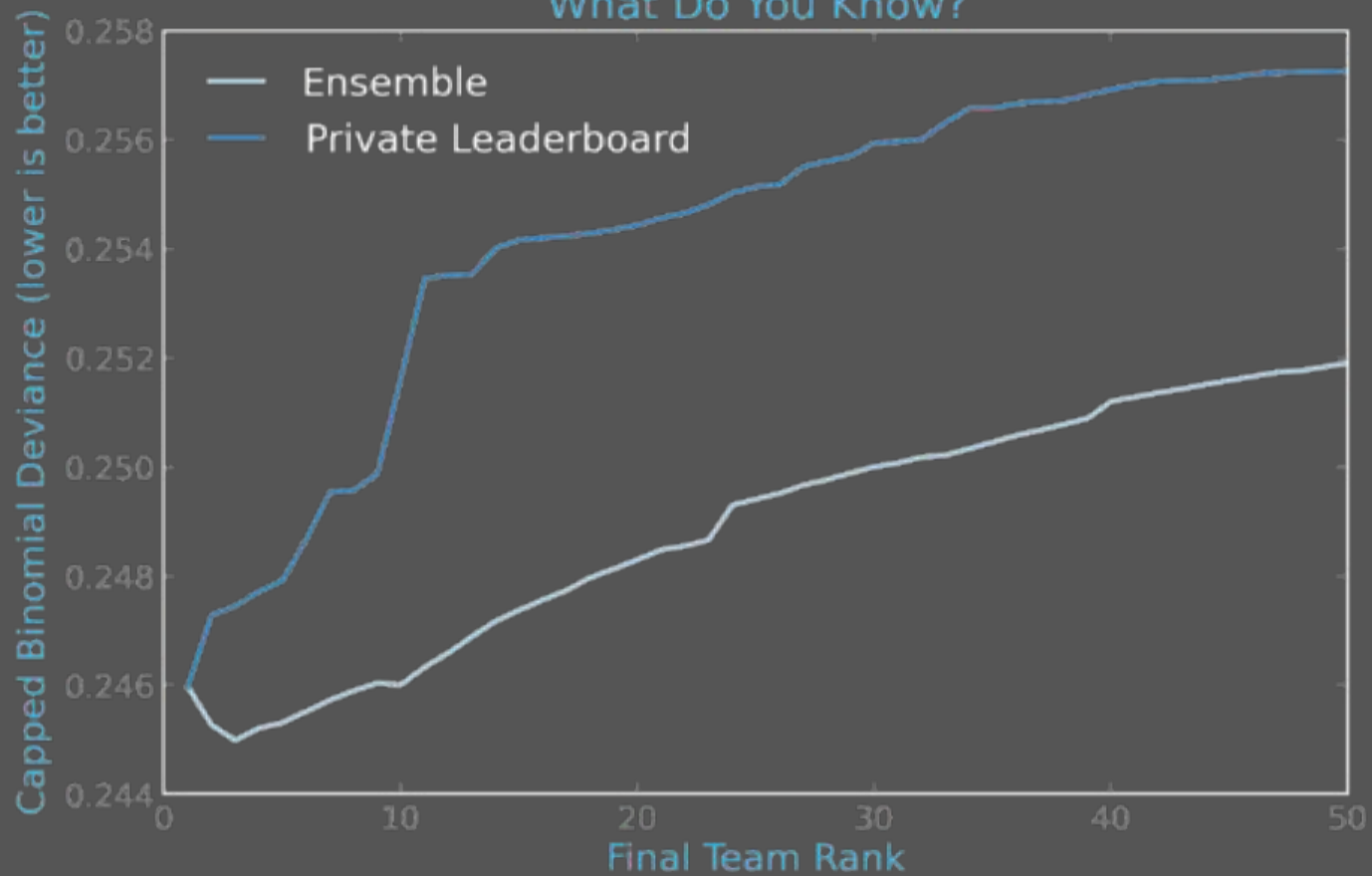
# Give Me Some Credit



## Photo Quality Prediction



## What Do You Know?





What are the

# Risks Involved?



**Q: What if the data is  
proprietary?**

# Private Competitions



## Will I Stay or Will I Go?

Predict which of our current customers will stay insured with us for an entire policy term.

Ends 43 days

12 teams

USD

## Who gets invited?

A liquid market of competitors and competitions enables ranking of participants (Like chess, golf, etc.)

We are refining our methodology to find our strongest members for private competitions



**Q: Is anonymized data  
anonymous in the hands of  
the seething masses?**



## Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge

Arvind Narayanan

Elaine Shi

Benjamin I. P. Rubinstein

This paper describes the winning entry to the 2011 IJCNN Social Network Challenge run by Kaggle.com. The goal of the contest was to promote research on real-world link prediction, and the dataset was a graph obtained by crawling the popular Flickr social photo sharing website, with user identities scrubbed. By de-anonymizing much of the competition test set using our own Flickr crawl, we were able to effectively game the competition. Our attack represents a new application of de-anonymization to gaming machine learning contests, suggesting changes in how future competitions should be run.

We introduce a new simulated annealing-based graph matching algorithm for the seeding step of de-anonymization. We also show how to combine de-anonymization with link prediction—the latter is required to achieve good performance on the portion of the test set not de-anonymized—for example by training the predictor on the de-anonymized portion of the test set, and combining probabilistic predictions from de-anonymization and link prediction.

### I. INTRODUCTION

KAGGLE.COM—a platform for machine learning competitions—ran the IJCNN 2011 Social Network Challenge for 9 weeks from Nov 8, 2010 through Jan 11, 2011 [18]. The goal of the Social Network Challenge was to promote research on link prediction. The contest dataset was created by crawling a large online social network and partitioning the obtained edge set into a large training set and a smaller test set of edges augmented with an equal number of fake edges. Challenge entries were required to be probabilistic predictions on the test edge set. Node identities

prior work studied de-anonymizing complete snapshots of social networks [26]. We achieve this by focusing on nodes with high in-degrees for “seeding” the de-anonymization process. As we explain in Section III-A, the set of high in-degree nodes is (approximately) preserved even in a snapshot obtained from a partial crawl.

Second, we formulate seed identification—the first step of de-anonymization—as a combinatorial optimization problem, specifically *weighted graph matching*, in contrast to the pattern search approaches of [6] and [26]. We then show how to use simulated annealing to solve this problem. Since our formulation makes no assumptions specific to the de-anonymization context, our solution is broadly applicable to the weighted graph matching problem.

Third, our winning entry, which yielded a combined test Area Under Curve (AUC) of 0.981, made use of a novel combination of standard link prediction with de-anonymization to game a popular link prediction contest. Moreover the link prediction component of our entry was advantaged by training on the de-anonymized portions of the test set. While previous applications of de-anonymization have been to privacy attacks [27], [25], to the best of our knowledge this is the first application of de-anonymization to gaming a machine learning contest.

The success of our approach has important consequences for future machine learning contests particularly in social network analysis. We argue that while appropriate contest rules should be used to disincentivize gaming through de-

## Graph-based Features for Supervised Link Prediction

William Cukierski, Benjamin Hamner, Bo Yang

**Abstract**—The growing ubiquity of social networks has spurred research in link prediction, which aims to predict new connections based on existing ones in the network. The 2011 IJCNN Social Network challenge asked participants to separate real edges from fake in a set of 8960 edges sampled from an anonymized, directed graph depicting a subset of relationships on Flickr. Our method incorporates 94 distinct graph features, used as input for classification with Random Forests. We present a three-pronged approach to the link prediction task, along with several novel variations on established similarity metrics. We discuss the challenges of processing a graph with more than a million nodes. We found that the best classification results were achieved through the combination of a large number of features that model different aspects of the graph structure. Our method achieved an area under the receiver-operator characteristic (ROC) curve of 0.9695, the 3rd best overall score in the competition and the best score which did not de-anonymize the dataset.

### I. INTRODUCTION

Directed graphs encapsulate relationships in social networks, with nodes representing members of the network and edges signifying the relations between them. Link prediction, the task of forecasting new connections based on existing ones, is a topic of growing importance as digital networks grow in size and ubiquity [1], [2], [3], [4]. The study of network dynamics has numerous applications. Marketers would like to recommend products or services based on existing preferences or contacts. Social networking web-

there are numerous reasons for friendship on a photo sharing site. It may be that two users are friends in real life, or they may share interest in a common subject matter, or they may share interest in a common style of photography. Recognizing the disparate meanings of graph edges leads to new interpretations of traditional link prediction methods.

Our approach to the IJCNN Social Network Challenge follows a classical paradigm in supervised learning, starting with feature extraction, then preprocessing, and lastly repeated classification using the posterior probabilities from Random Forests [5]. Instead of presenting a single novel methodology for link prediction, the foremost contribution of this paper is in the breadth and variety of techniques incorporated into the feature extraction step. Sec. II describes this process in detail, starting with subgraph extraction, descriptions of the features, and finally, meta approaches to make valuable, new predictors from these features. Aspects of the work which are novel, such as the application of Bayesian Sets and the development of the three-problem approach, are discussed in more detail at the end of the section.

### II. FEATURE EXTRACTION

We now introduce notation used throughout the paper. A graph  $G$  is a set of  $N$  vertices and directed edges  $(V, E)$ , with associated adjacency matrix  $A$ . When considering whether a specific edge  $A_{ij}$  is real or fake, we label the outbound

1102.4374v1 [cs.CR] 22 Feb 2011

**A: Privacy and utility trade off.  
Greater extremes of  
anonymization lead to less  
useful models.**

**Q: Do competitions  
always lead to grotesquely  
complicated models?**

## GEAR &amp; GADGETS / PRODUCT NEWS &amp; REVIEWS

## Netflix never used its \$1 million algorithm due to engineering costs

Netflix never used the recommendation improvements produced by its \$1 million ...

by Casey Johnston - Apr 13 2012, 2:25pm PDT

[BIG DATA](#) [BUSINESS](#) [DEVELOPMENT](#) [IT](#) [MEDIA INNOVATION](#)



Netflix awarded a \$1 million prize to a developer team in 2009 for an algorithm that increased the accuracy of the company's recommendation engine by 10 percent. But today it doesn't use the million-dollar code, and has no plans to implement it in the future, Netflix [announced](#) on its blog Friday. The post goes on to explain why: a combination of too much engineering effort for the results, and a shift from movie recommendations to the "next level" of personalization caused by the transition of the business from mailed DVDs to video streaming.

Netflix notes that it does still use two algorithms from the team that won the first Progress Prize for an 8.43 percent improvement to the recommendation engine's root mean squared error (the full \$1 million was [awarded](#) for a 10 percent improvement). But the increase in accuracy on the winning improvements "did not seem to justify the engineering effort needed to bring them into a production environment," the blog post said. By that time, the company had moved on anyway.

When Netflix announced the contest to improve the service in 2007, its business was centered on DVDs, which are dealt with by customers in periods of days or weeks and provide little granular data. Now that Netflix's primary offering is streaming, it has access to much more information: Streaming members are looking for something great to watch right now, they can sample a few

### TOP FEATURE STORY ▾



FEATURE STORY (2 PAGES)

## Transportation innovation: How Lyft and SideCar are changing commuting

Smartphone apps + "ride-sharing" = travel revolution? Two startups say yes.



### STAY IN THE KNOW WITH ▾



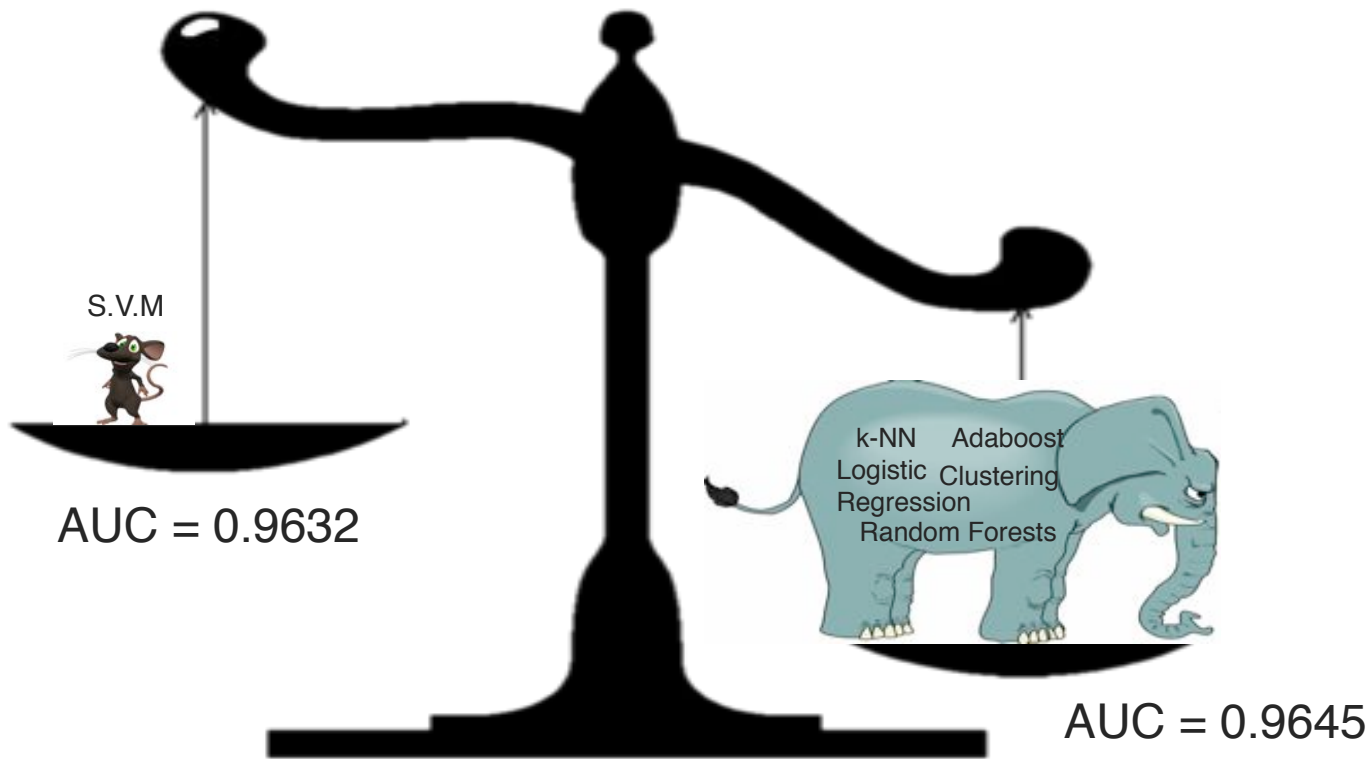
### LATEST NEWS ▾

EXCESSIVE FORCE

Con accused of tackling 15-year-old in



# The (Ensembling) Elephant in the room



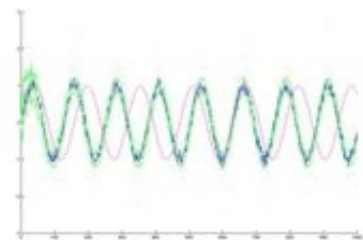
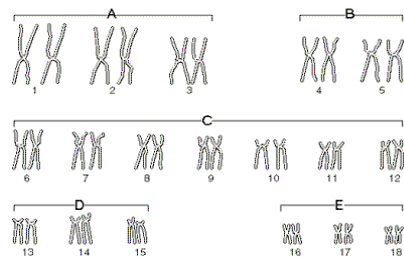
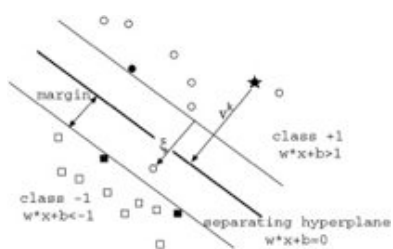
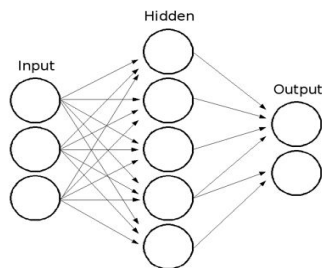
As the number and competitiveness of competitions increases, will non-specific ensemble approaches dominate the solution landscape?



**A: It's perhaps the wrong question to ask. The pieces are just as valuable as the whole.**

**Q: Accuracy isn't everything.  
Isn't this a rigged game?**

# Depth and breadth



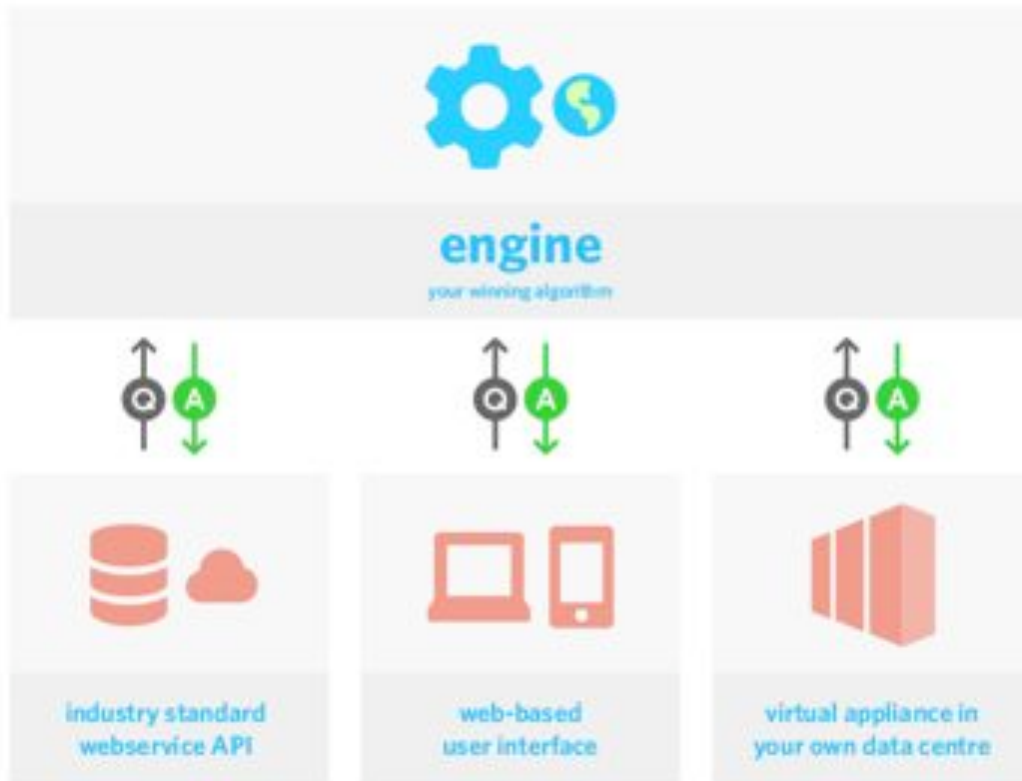
- neural networks
- logistic regression
- support vector machines
- decision trees
- ensemble methods
- adaBoost
- Bayesian networks

- **feature extraction**
- genetic algorithms
- random forests
- Monte Carlo methods
- Stochastic gradient boosting
- Kalman filters
- evolutionary fuzzy modeling

**A: Accuracy is only the headline.  
The story is in the depth and  
breadth of approaches.**

**Q: It's great that the winning solution employed a 6000-dimension nonlinear kernel. Just give us the answers.**

# Kaggle Engine





# Our experience crowdsourcing predictive models

## Better Results

- Every competition we've hosted has **beaten existing benchmarks**
- Commercial benchmarks have been **improved by an average of 40%**

## Faster Timelines

- **Results are typically achieved in weeks**, often improving on benchmarks that reflect years of work

## Added Certainty

- **Knowing what is possible** with existing data is a rare luxury in scientific research

## Reduced Cost

- **Significantly less expensive** than traditional alternatives



kaggle.com

**What can the world's best  
data scientists find in your data?**

e-mail [will.cukierski@kaggle.com](mailto:will.cukierski@kaggle.com)

phone +1 415 309 0069