Heritage Health Prize

A predictive modeling competition that's life & death

Anthony Goldbloom
Kaggle

a@kaggle.com @antgoldbloom

Agenda –

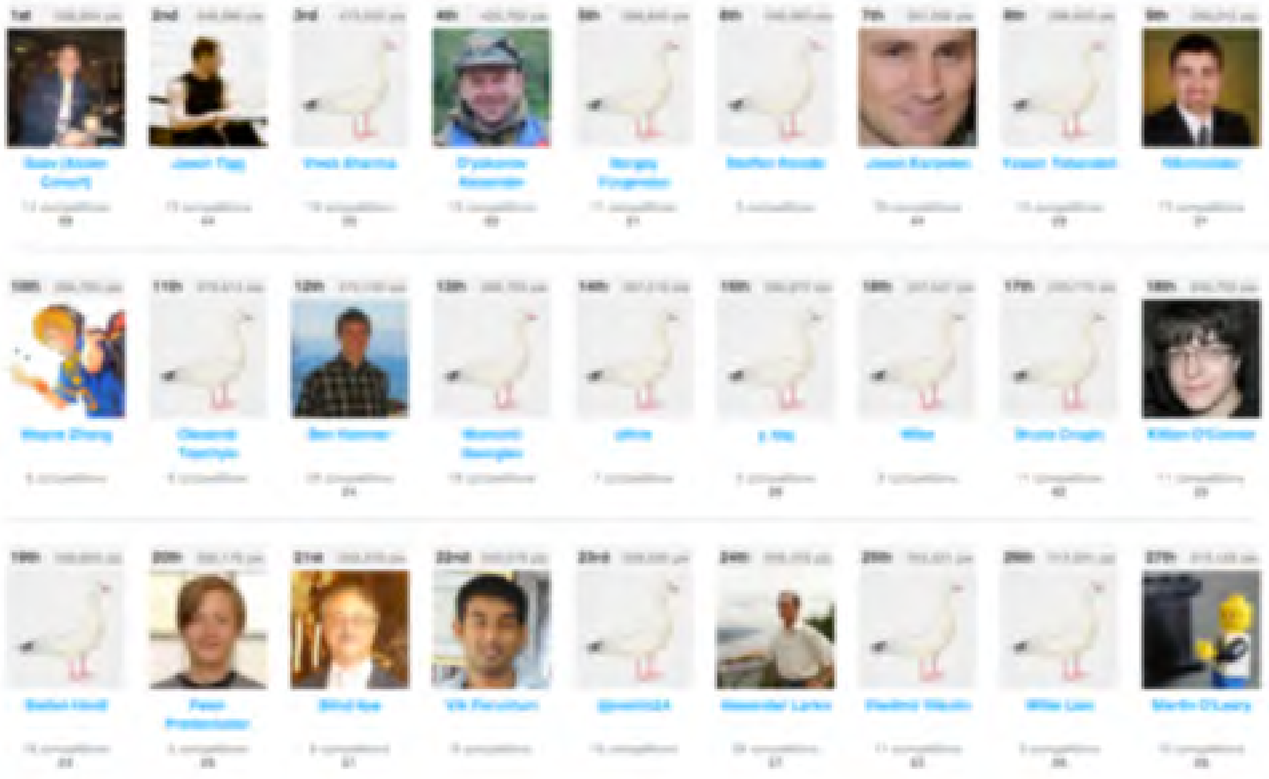
What is Kaggle

What are data science competitions

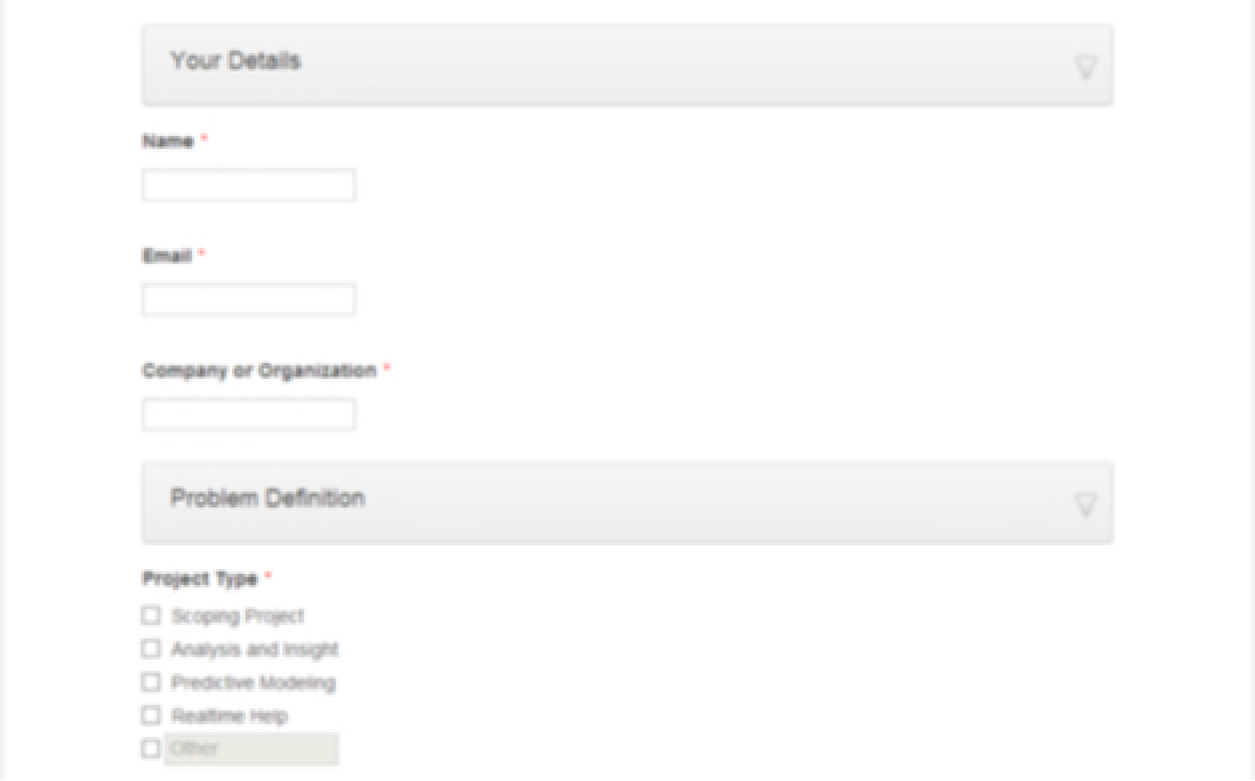
The Heritage Health Prize

Lessons Learned

Kaggle Connect is a marketplace with the world’s best, ranked



Kaggle Connect: submit your project and get matched with a data scientist



Your Details

Name *

Email *

Company or Organization *

Problem Definition

Project Type *

- □ Scoping Project
- □ Analysis and Insight
- □ Predictive Modeling
- □ Realtime Help
- □ Other

kaggle

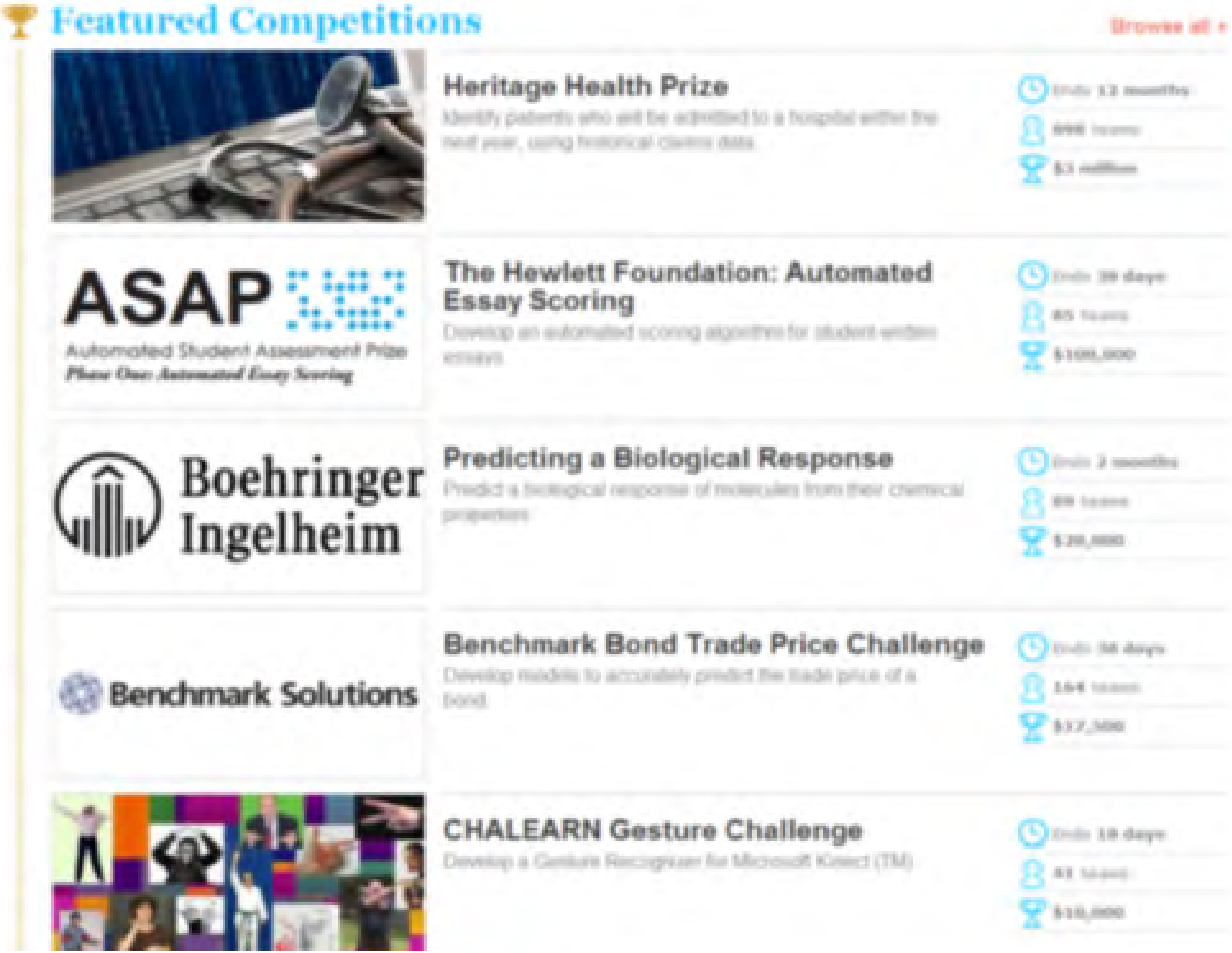
Agenda –

What is Kaggle

What are data science competitions

The Heritage Health Prize

Lessons Learned


ASAP
Automated Student Assessment Prize
Phase One: Automated Essay Scoring
Boehringer Ingelheim
Benchmark Solutions

Featured Competitions

Browse all »

Heritage Health Prize

Identify patients who will be admitted to a hospital within the next year, using historical claims data

Ends 13 months
898 teams
$3 million

The Hewlett Foundation: Automated Essay Scoring

Develop an automated scoring algorithm for student-written essays

Ends 39 days
85 teams
$100,000

Predicting a Biological Response

Predict a biological response of molecules from their chemical properties

Ends 2 months
89 teams
$20,000

Benchmark Bond Trade Price Challenge

Develop models to accurately predict the trade price of a bond

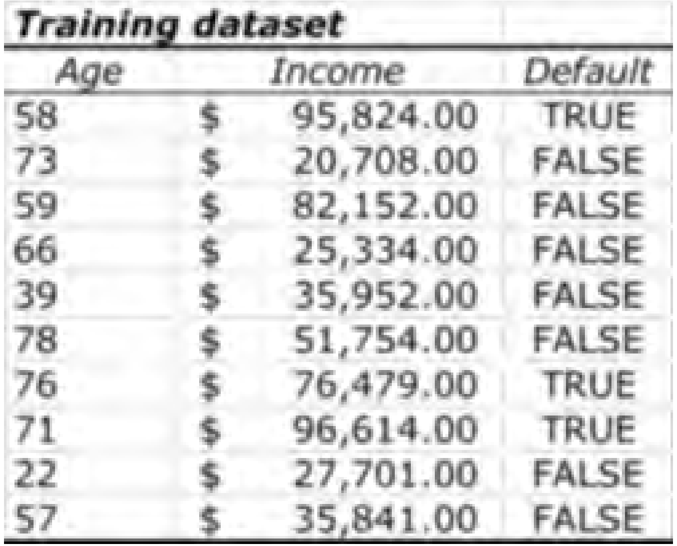
Ends 56 days
164 teams
$17,500

CHALEARN Gesture Challenge

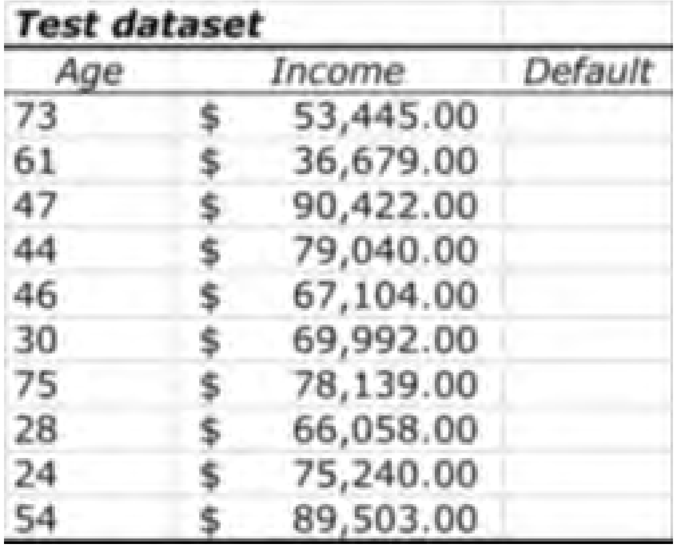
Develop a Gesture Recognizer for Microsoft Kinect (TM)

Ends 18 days
41 teams
$10,000

Competition Mechanics

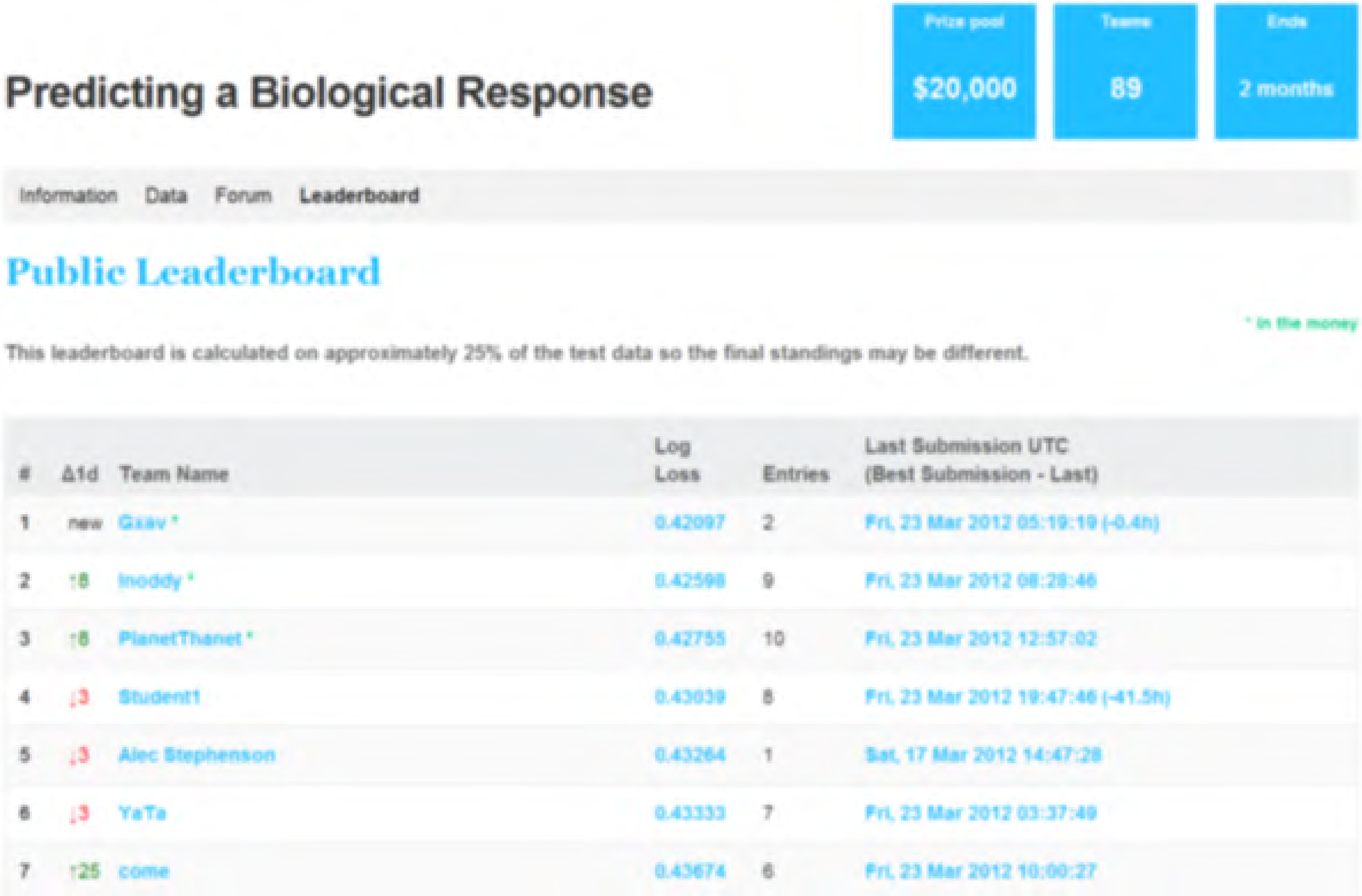


Training dataset		
Age	Income	Default
58	$ 95,824.00	TRUE
73	$ 20,708.00	FALSE
59	$ 82,152.00	FALSE
66	$ 25,334.00	FALSE
39	$ 35,952.00	FALSE
78	$ 51,754.00	FALSE
76	$ 76,479.00	TRUE
71	$ 96,614.00	TRUE
22	$ 27,701.00	FALSE
57	$ 35,841.00	FALSE



Test dataset		
Age	Income	Default
73	$ 53,445.00	
61	$ 36,679.00	
47	$ 90,422.00	
44	$ 79,040.00	
46	$ 67,104.00	
30	$ 69,992.00	
75	$ 78,139.00	
28	$ 66,058.00	
24	$ 75,240.00	
54	$ 89,503.00	

Competitions are judged on objective criteria



Predicting a Biological Response

Prize pool
$20,000

Teams
89

Ends
2 months

Public Leaderboard

* in the money

This leaderboard is calculated on approximately 25% of the test data so the final standings may be different.

#	Δ1d	Team Name	Log Loss	Entries	Last Submission UTC (Best Submission - Last)
1	new	Gxav *	0.42097	2	Fri, 23 Mar 2012 05:19:19 (-0.4h)
2	↑8	Inoddy *	0.42598	9	Fri, 23 Mar 2012 08:28:46
3	↑8	PlanetThanet *	0.42755	10	Fri, 23 Mar 2012 12:57:02
4	↓3	Student1	0.43039	8	Fri, 23 Mar 2012 19:47:46 (-41.5h)
5	↓3	Alec Stephenson	0.43264	1	Sat, 17 Mar 2012 14:47:28
6	↓3	YaTa	0.43333	7	Fri, 23 Mar 2012 03:37:49
7	↑25	come	0.43674	6	Fri, 23 Mar 2012 10:00:27

Agenda –

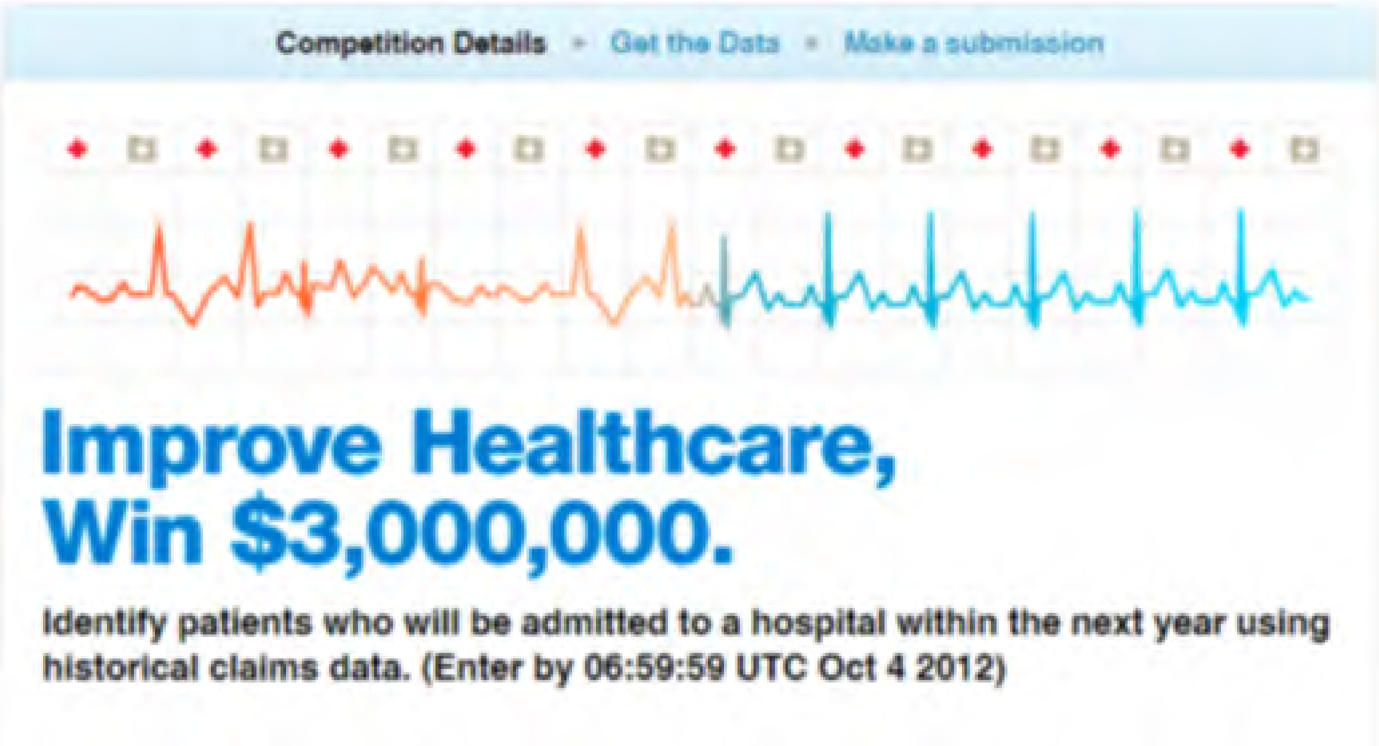
What is Kaggle

What are data science competitions

The Heritage Health Prize

Lessons Learned

The world’s largest ever data science competition

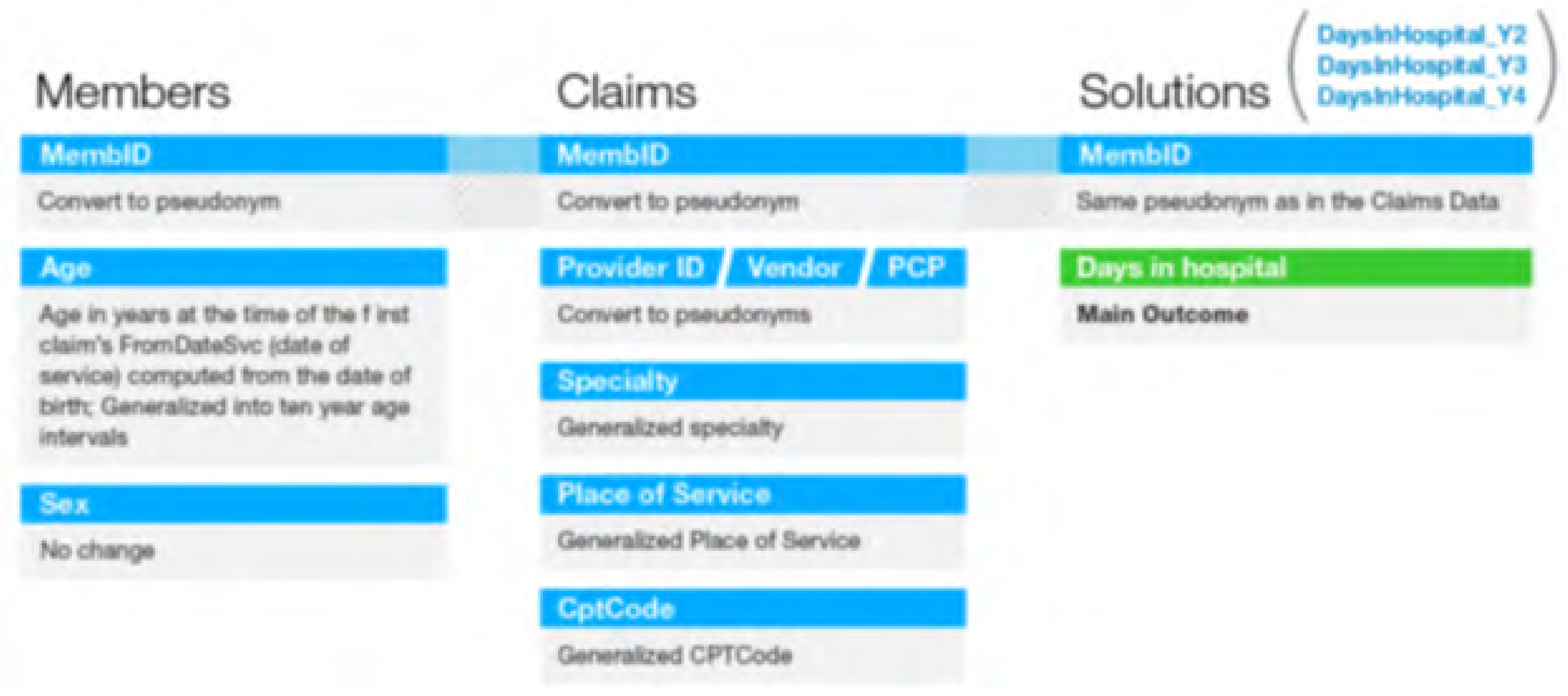


Competition Details » Get the Data » Make a submission

Improve Healthcare, Win $3,000,000.

Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)

Competition Mechanics



Members

MembID

Convert to pseudonym

Age

Age in years at the time of the f irst claim's FromDateSvc (date of service) computed from the date of birth; Generalized into ten year age intervals

Sex

No change

Claims

MembID

Convert to pseudonym

Provider ID / Vendor / PCP

Convert to pseudonyms

Specialty

Generalized specialty

Place of Service

Generalized Place of Service

CptCode

Generalized CPTCode

Solutions

DaysInHospital_Y2
DaysInHospital_Y3
DaysInHospital_Y4

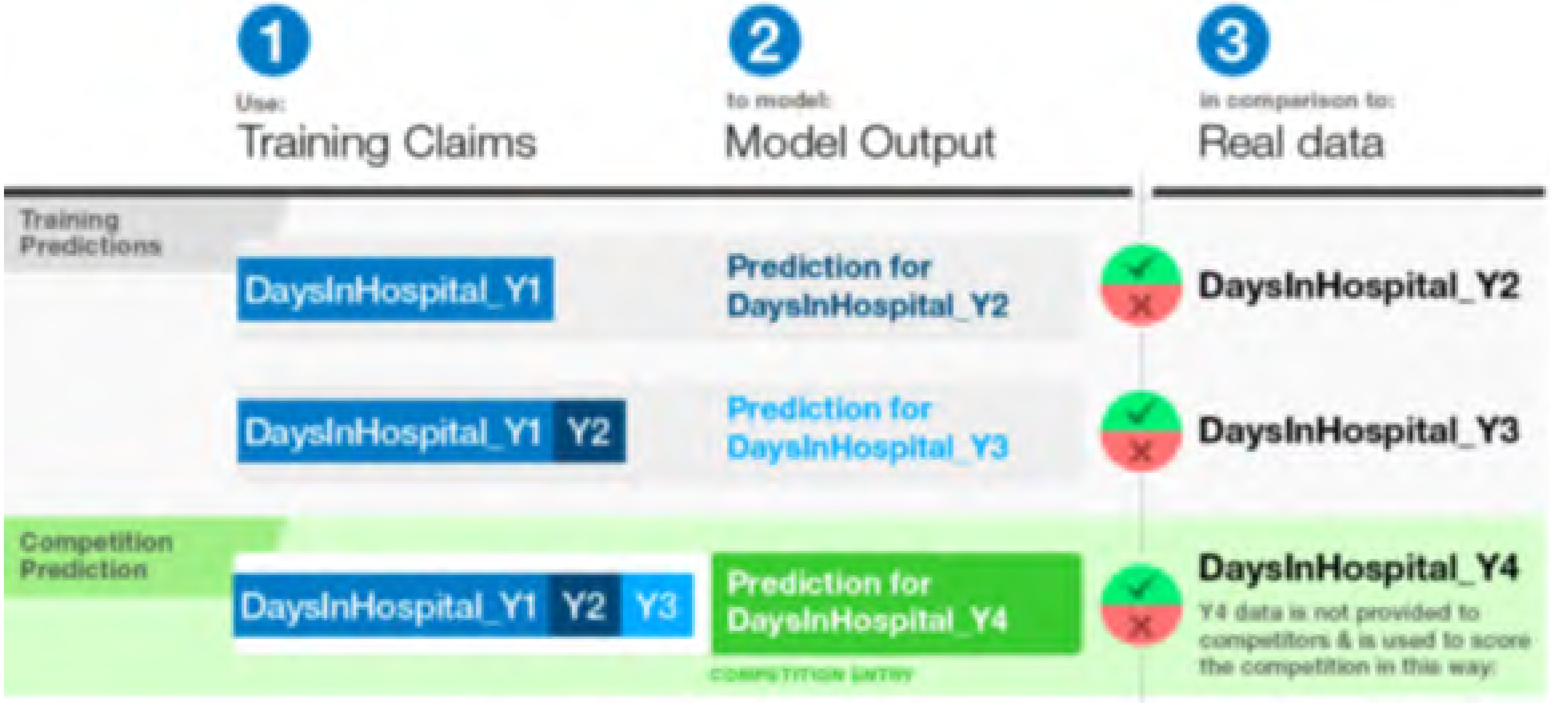
MembID

Same pseudonym as in the Claims Data

Days in hospital

Main Outcome

Competition Mechanics



1 Use: Training Claims

2 to model: Model Output

3 in comparison to: Real data

Training Predictions

DaysInHospital_Y1 → Prediction for DaysInHospital_Y2 ✓ ✗ DaysInHospital_Y2

DaysInHospital_Y1 Y2 → Prediction for DaysInHospital_Y3 ✓ ✗ DaysInHospital_Y3

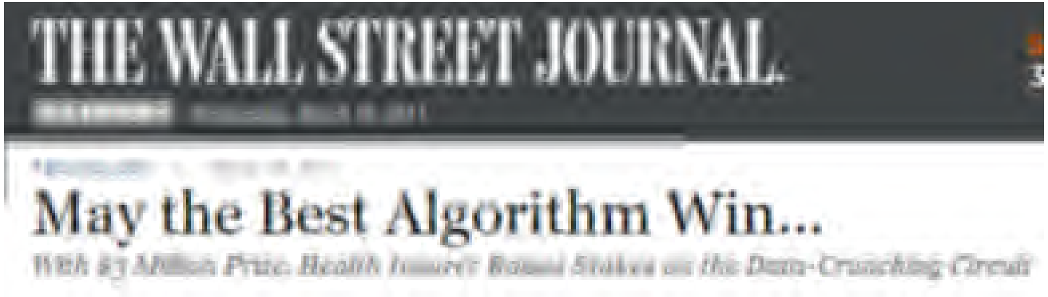
Competition Prediction

DaysInHospital_Y1 Y2 Y3 → Prediction for DaysInHospital_Y4 ✓ ✗ DaysInHospital_Y4

COMPETITION ENTRY

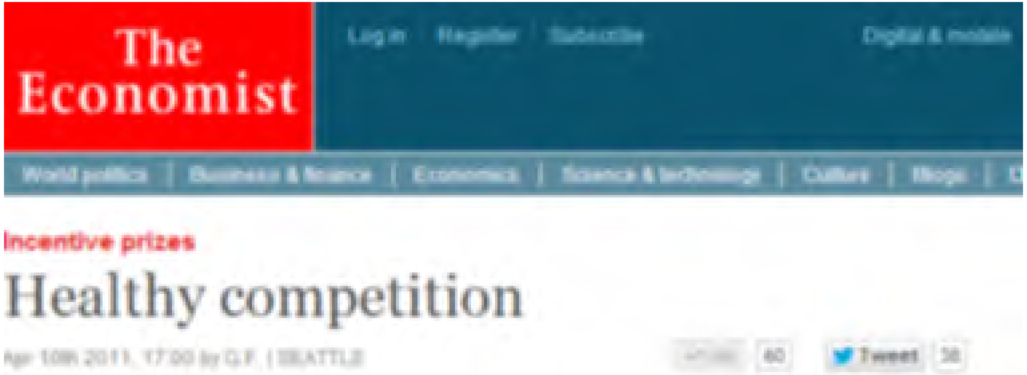
Y4 data is not provided to competitors & is used to score the competition in this way.

The prize launched on April 4th 2011 to lots of attention


THE WALL STREET JOURNAL.

May the Best Algorithm Win...

With $3 Million Prize, Health Insurer Raises Stakes on the Data-Crunching Circuit


The Economist

Log in | Register | Subscribe

Digital & mobile

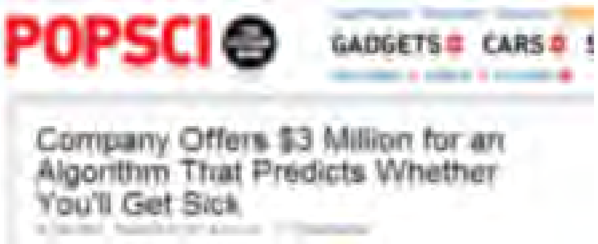
World politics | Business & finance | Economics | Science & technology | Culture | Blogs

Incentive prizes

Healthy competition

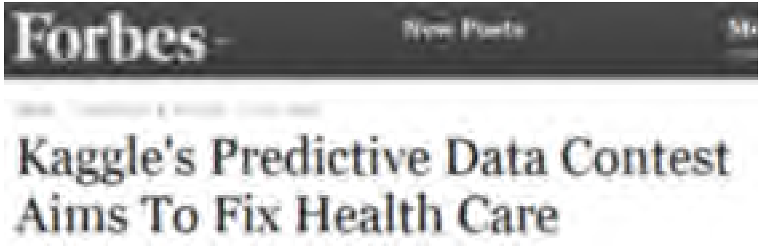
Apr 10th 2011, 17:00 by G.F. | SEATTLE

Tweet 38


POPSCI

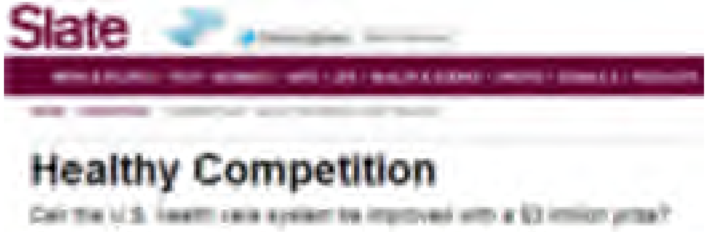
GADGETS | CARS

Company Offers $3 Million for an Algorithm That Predicts Whether You'll Get Sick


Forbes

New Posts

Kaggle's Predictive Data Contest Aims To Fix Health Care

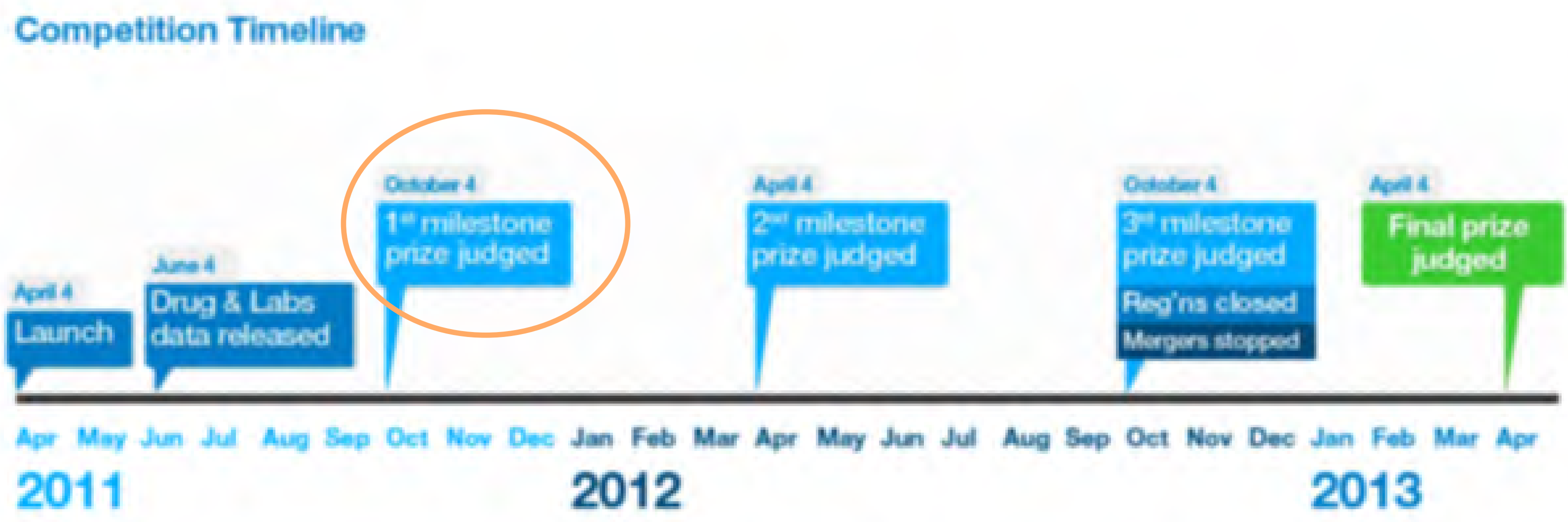

Slate

Healthy Competition

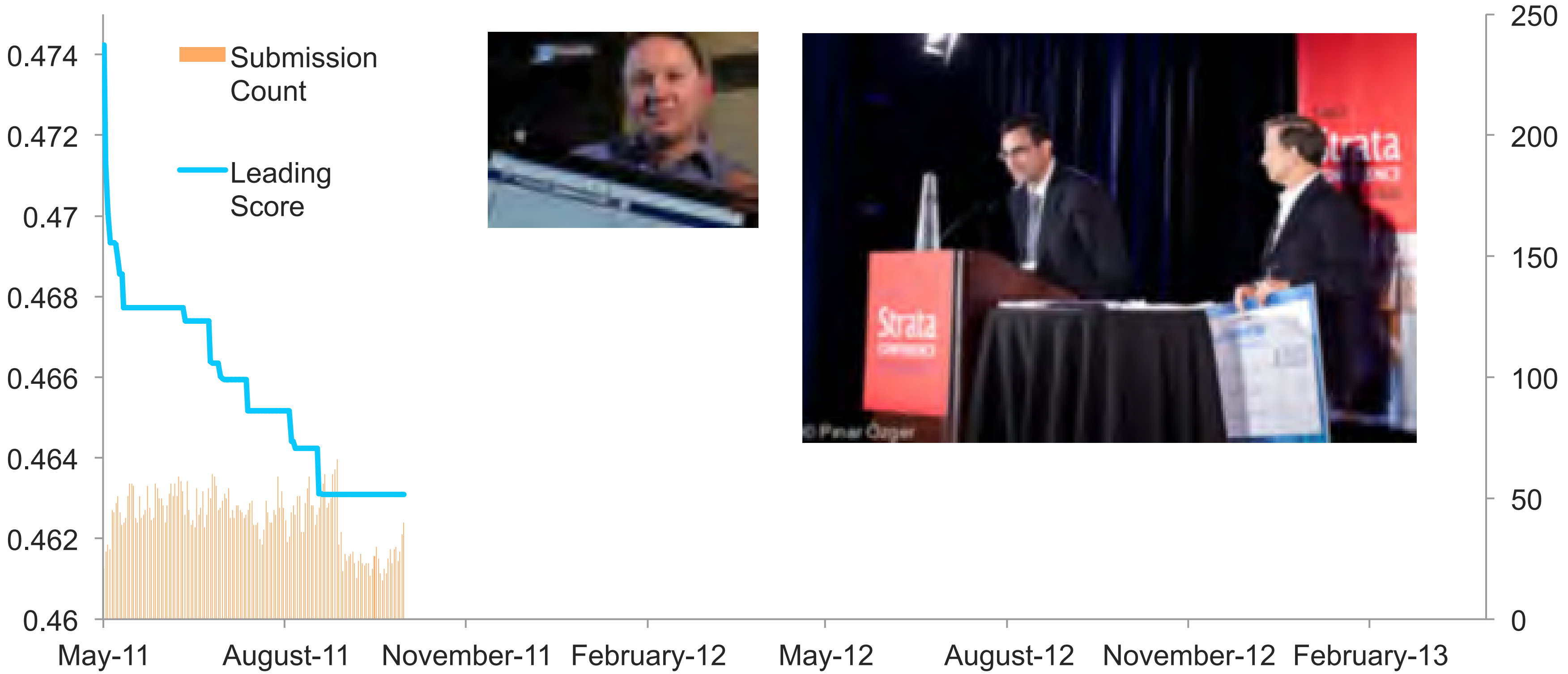
Can the U.S. health care system be improved with a $3 million prize?

kaggle

First milestone prize was awarded on October 4th 2011

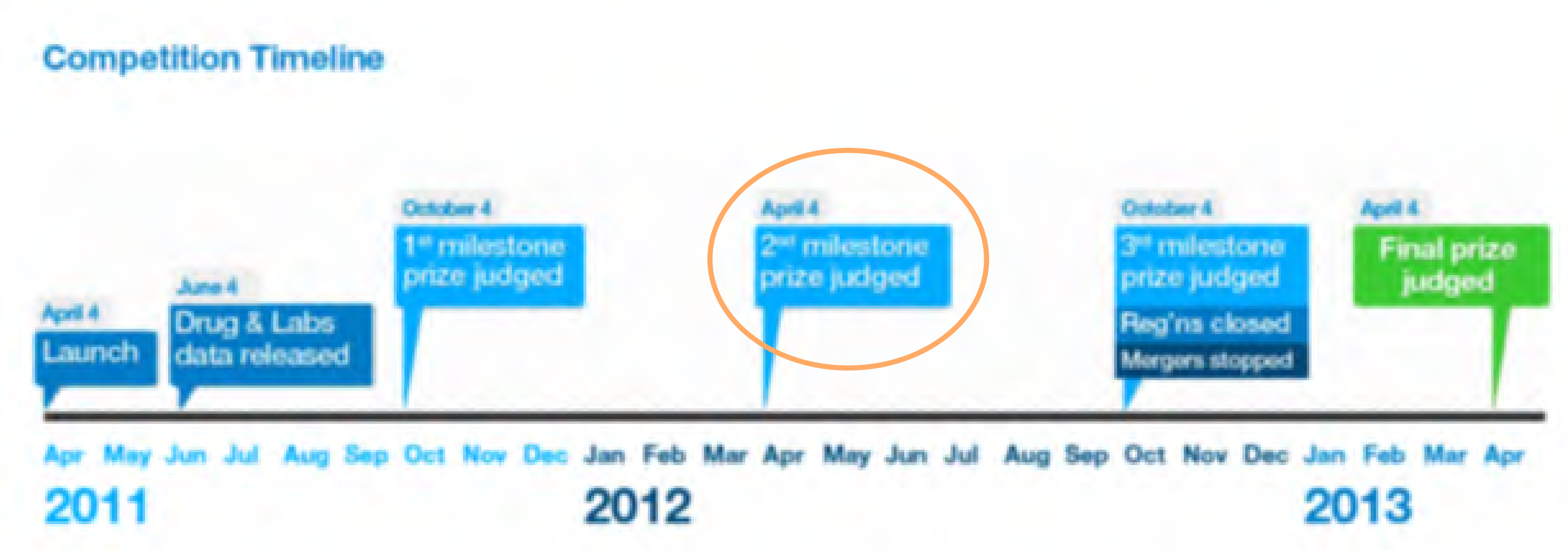

Competition Timeline
April 4
Launch
June 4
Drug & Labs data released
October 4
1st milestone prize judged
April 4
2nd milestone prize judged
October 4
3rd milestone prize judged
Reg'ns closed
Mergers stopped
April 4
Final prize judged
Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr
2011
2012
2013

Winners include an IBM consultant and a hedge fund trader



Meet Phil, the IBMer who was part of the team that won the first milestone prize

Second milestone prize was announced on April 4th 2012



Competition Timeline

April 4
Launch

June 4
Drug & Labs data released

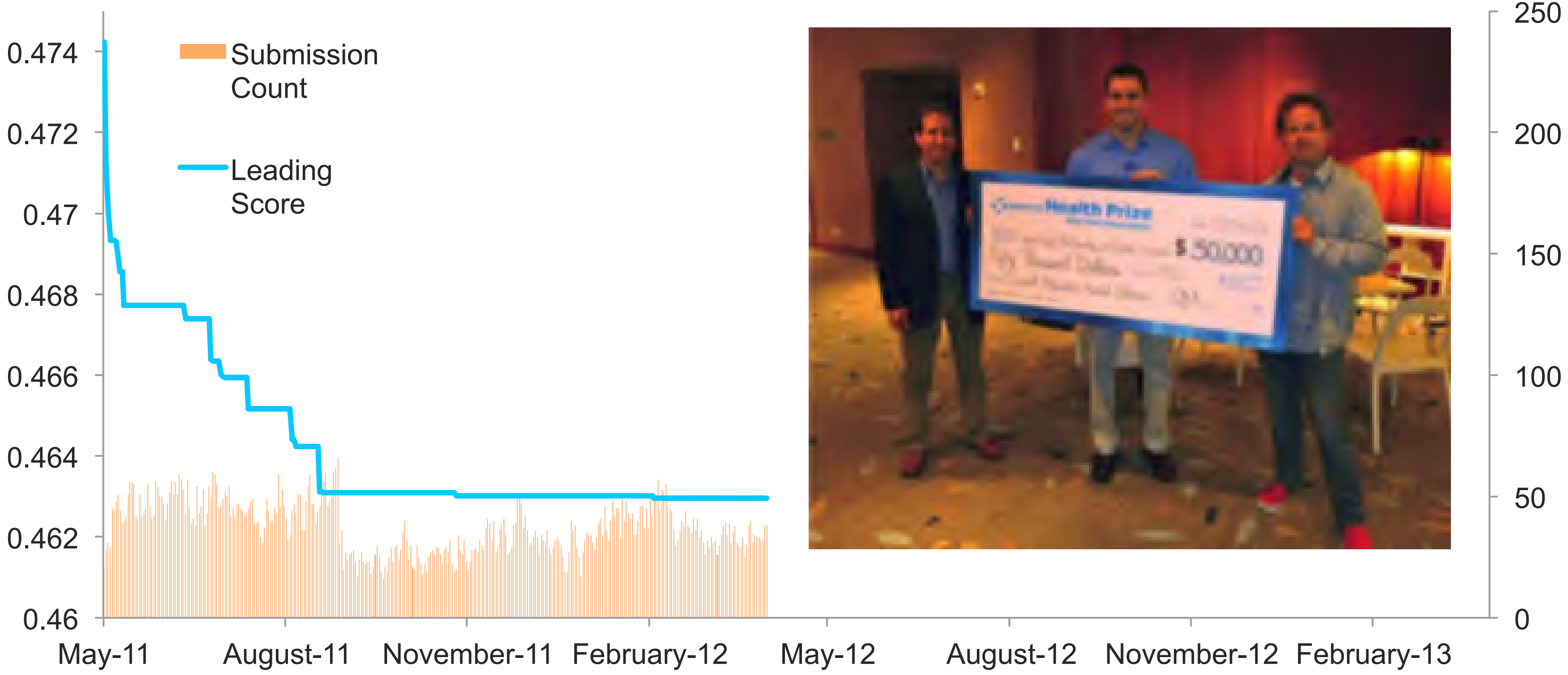
October 4
1st milestone prize judged

April 4
2nd milestone prize judged

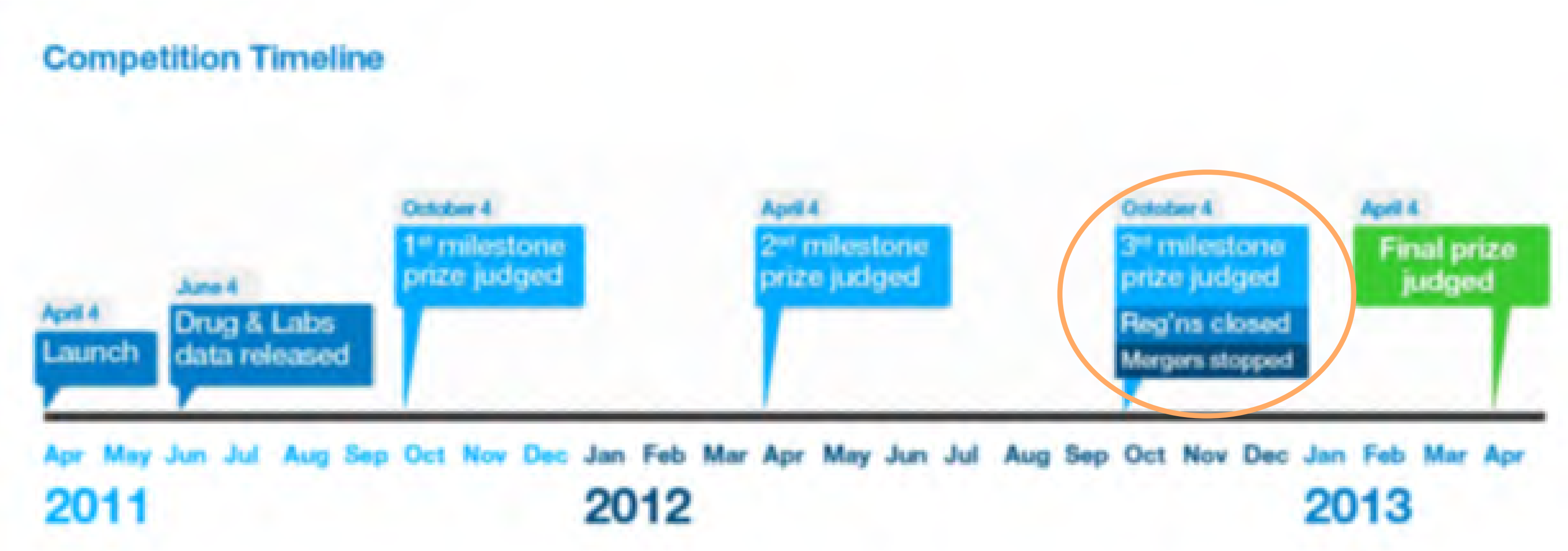
October 4
3rd milestone prize judged
Reg'ns closed
Mergers stopped

April 4
Final prize judged

Dave & Phil win milestone prize 2 as well



Third milestone prize was announced on October 4th 2012



Competition Timeline

April 4
Launch

June 4
Drug & Labs data released

October 4
1st milestone prize judged

April 4
2nd milestone prize judged

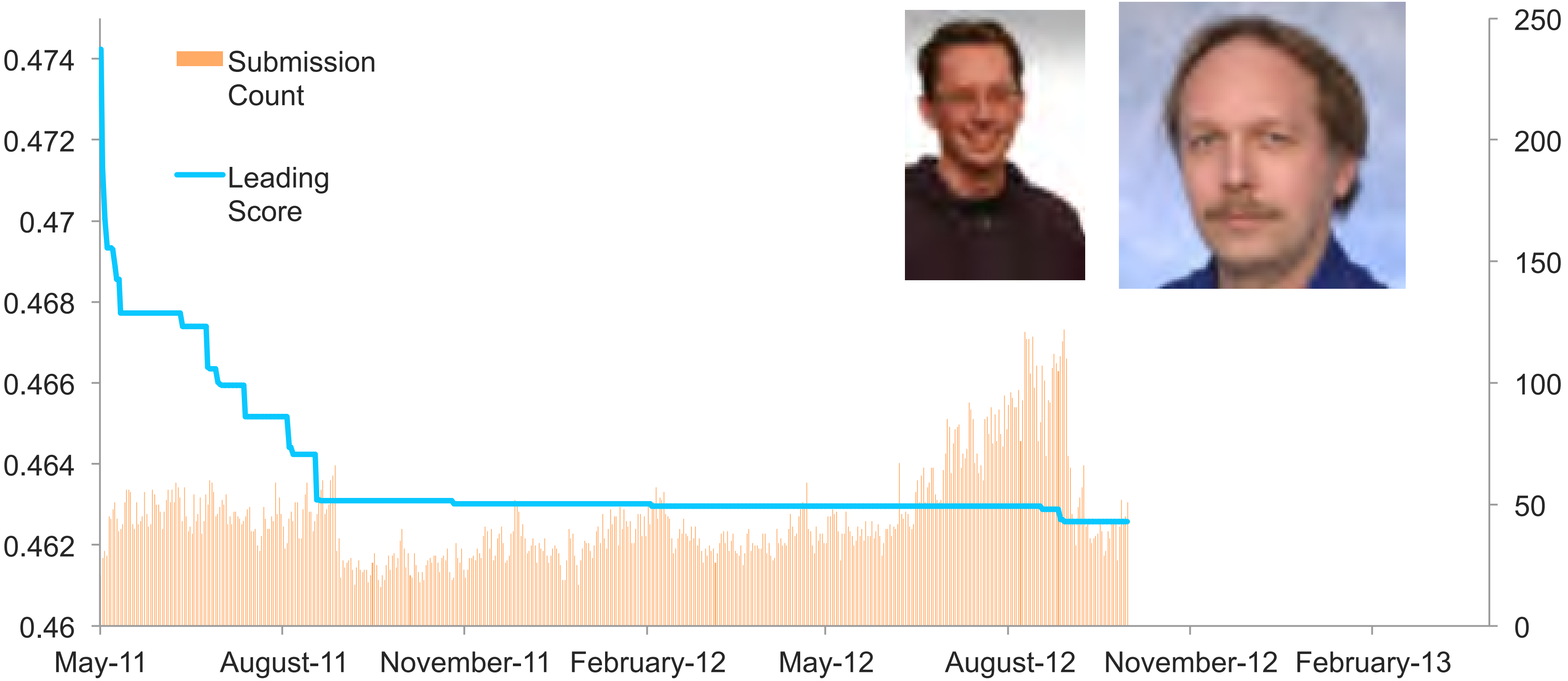
October 4
3rd milestone prize judged
Reg'ns closed
Mergers stopped

April 4
Final prize judged

Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr

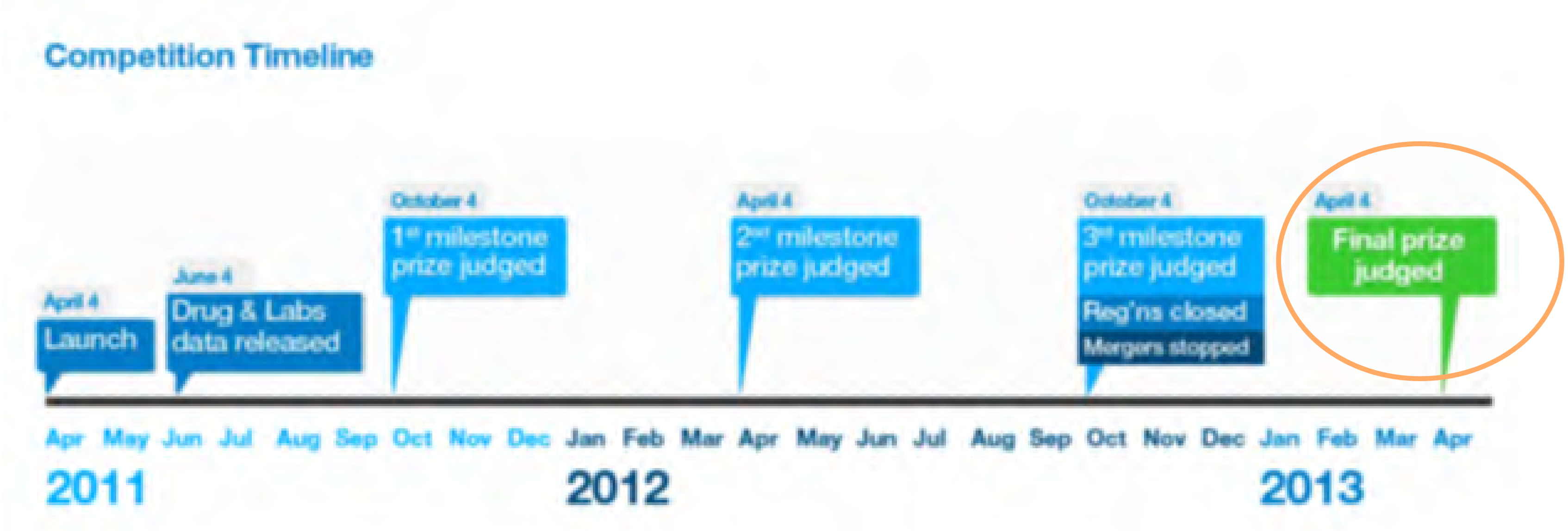
2011 2012 2013

Third awarded to a BI consultant and a software engineer



kaggle

The competition closed on April 4th 2013



Competition Timeline

April 4
Launch

June 4
Drug & Labs data released

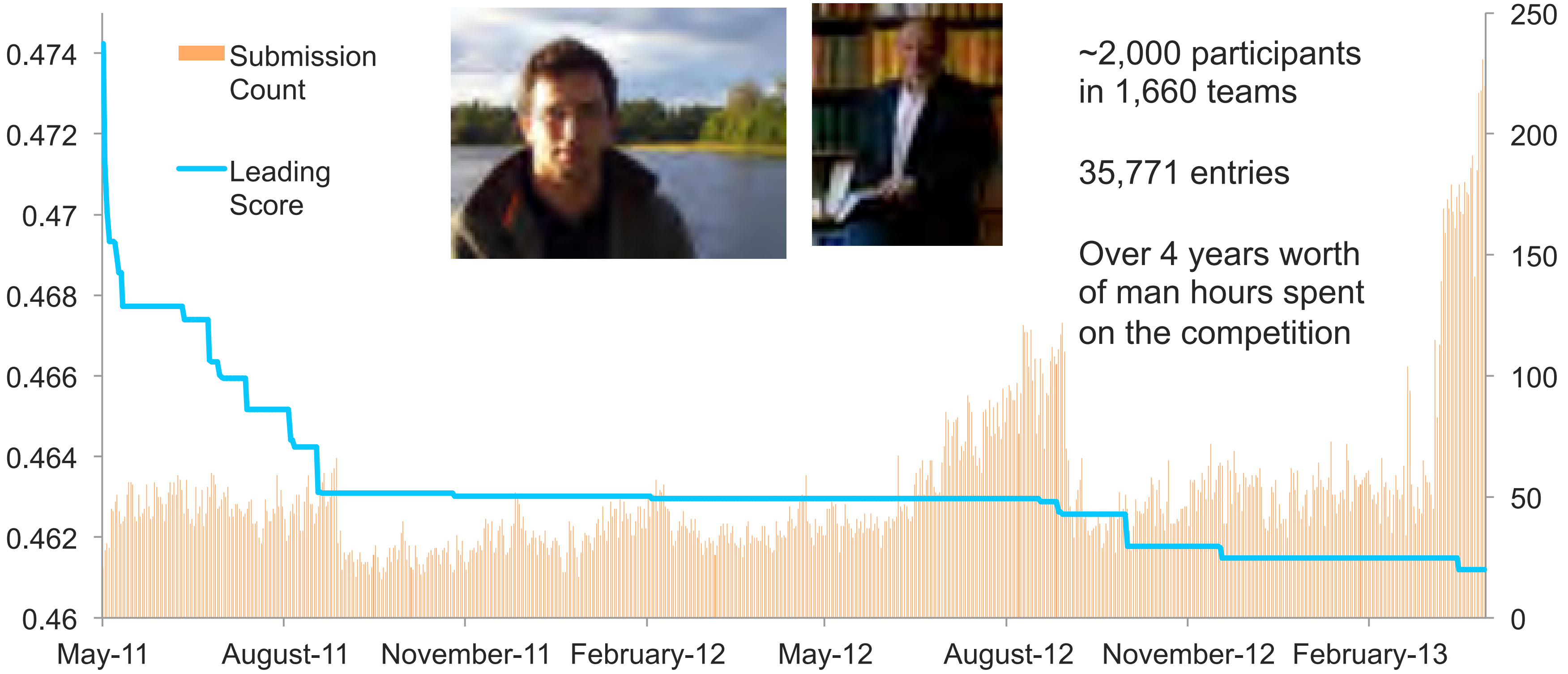
October 4
1st milestone prize judged

April 4
2nd milestone prize judged

October 4
3rd milestone prize judged
Reg'ns closed
Mergers stopped

April 4
Final prize judged

Final prize will be announced at the in DC on June 4th



~2,000 participants in 1,660 teams

35,771 entries

Over 4 years worth of man hours spent on the competition

Agenda –

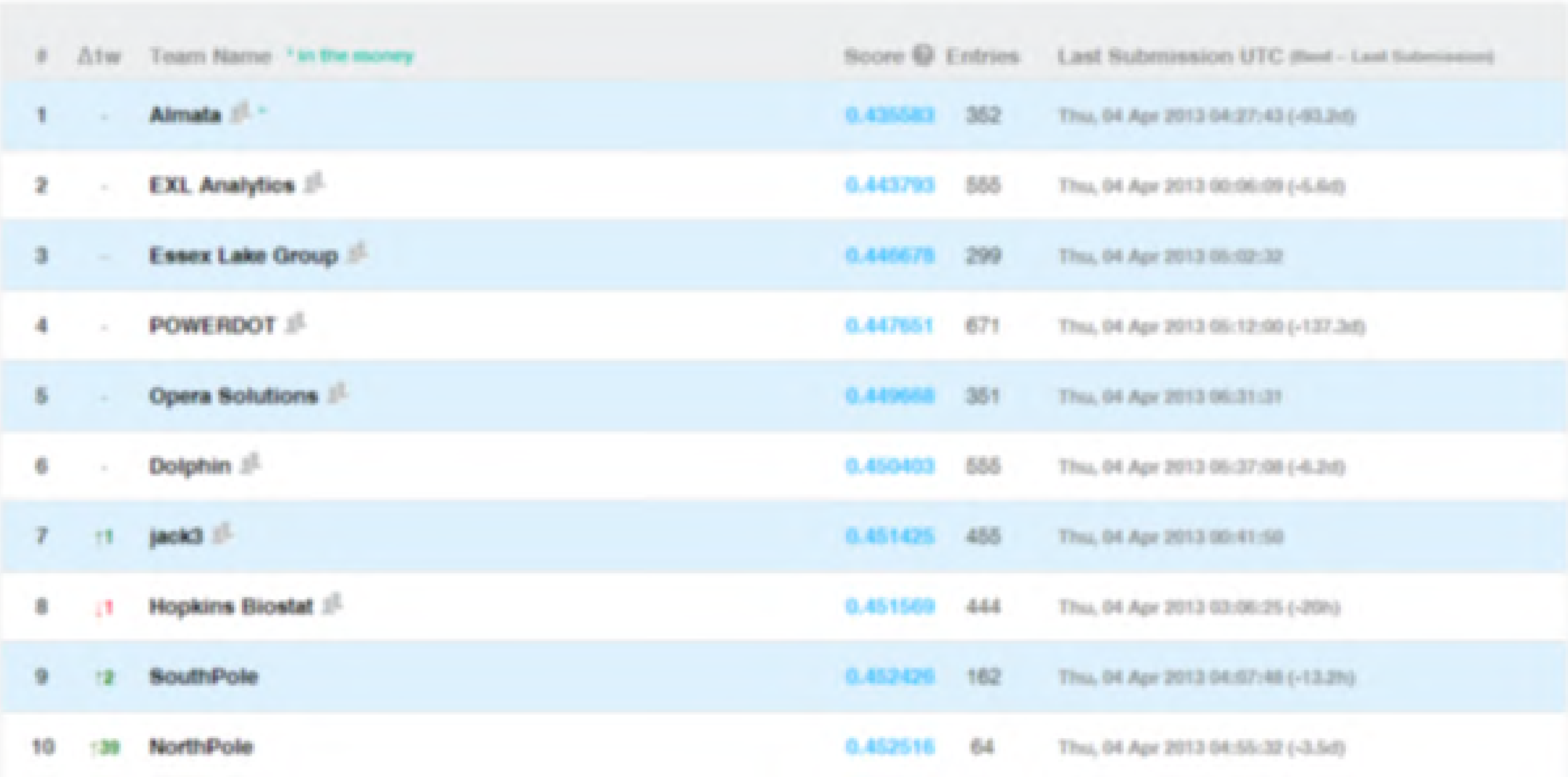
What is Kaggle

What are data science competitions

The Heritage Health Prize

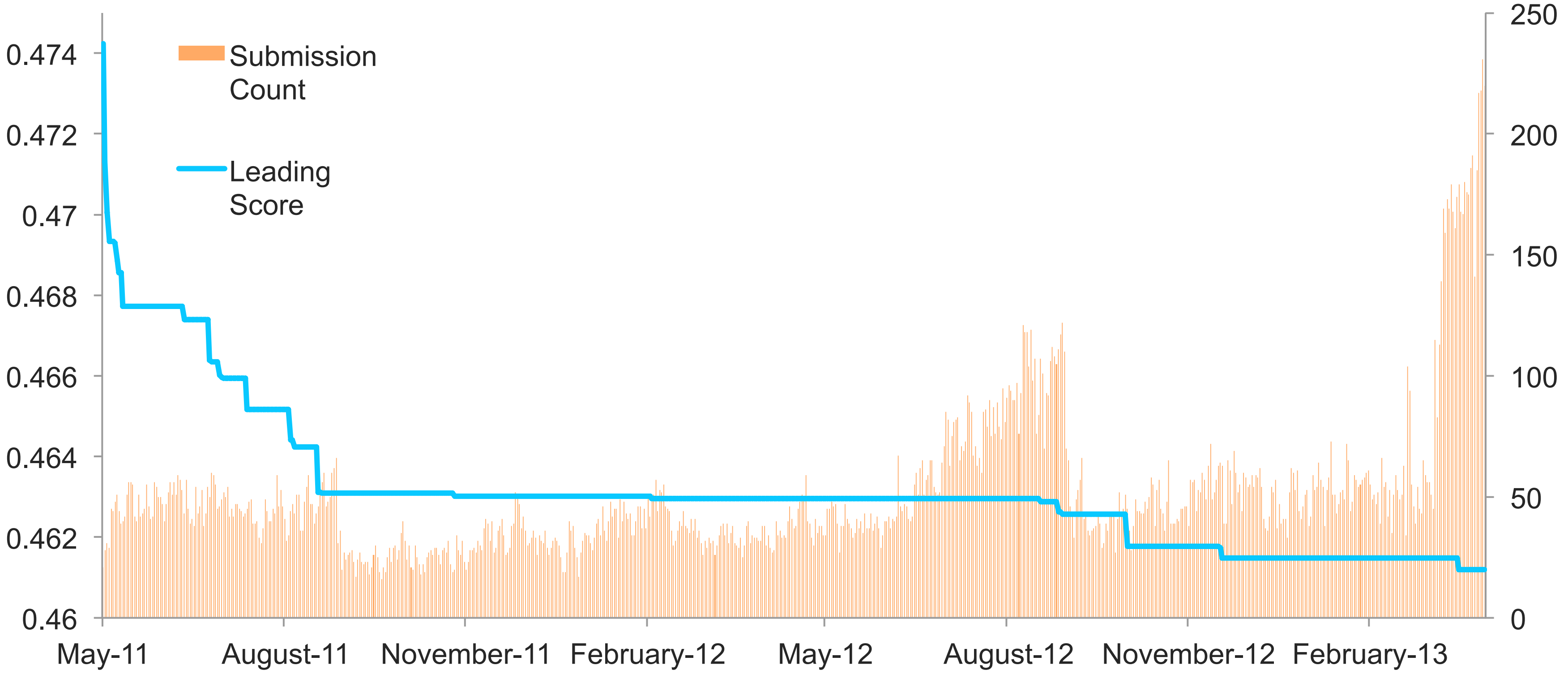
Lessons Learned

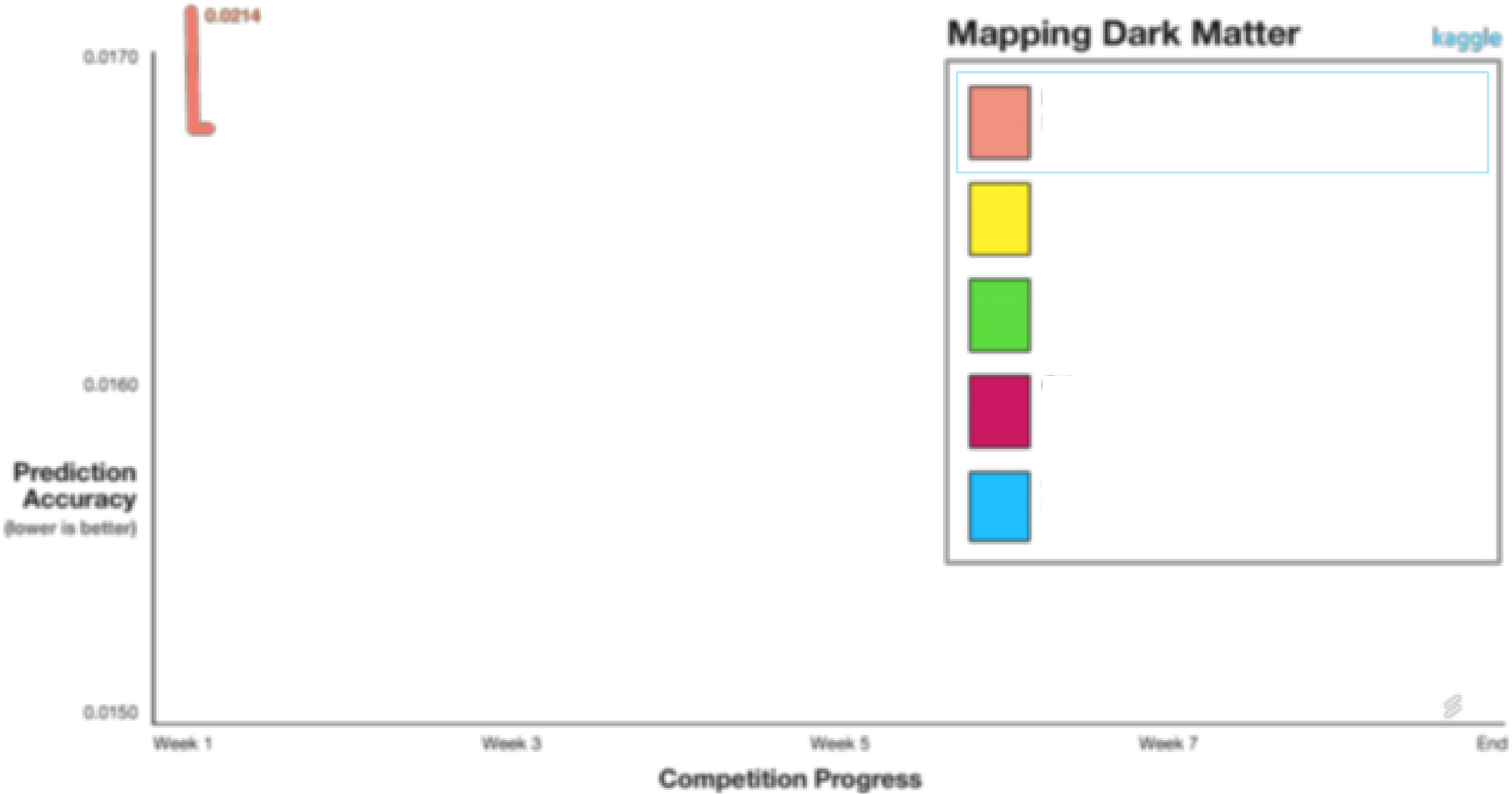
Top entries bunch at around the same level of predictive accuracy

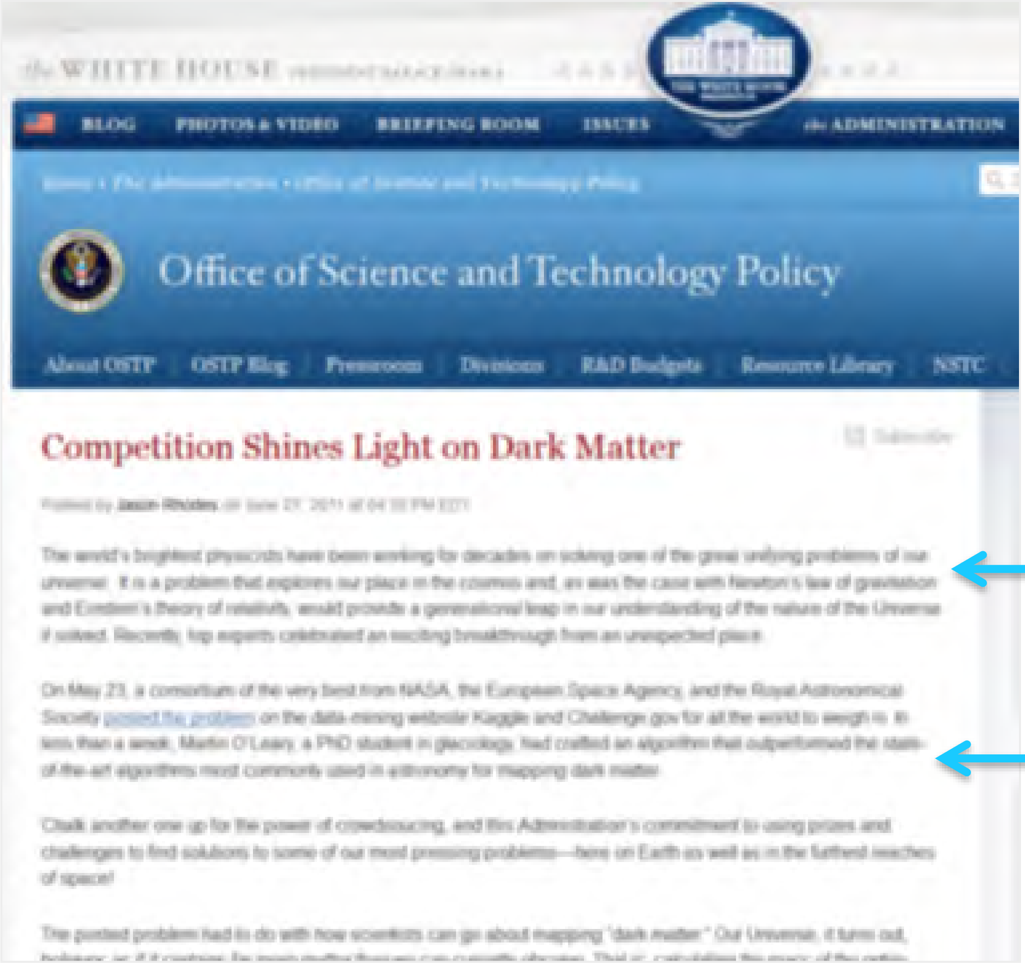


#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	-	Almata *	0.435583	352	Thu, 04 Apr 2013 04:27:43 (-93.2d)
2	-	EXL Analytics	0.443793	555	Thu, 04 Apr 2013 00:06:09 (-5.6d)
3	-	Essex Lake Group	0.446678	299	Thu, 04 Apr 2013 05:02:32
4	-	POWERDOT	0.447651	671	Thu, 04 Apr 2013 05:12:00 (-137.3d)
5	-	Opera Solutions	0.449658	351	Thu, 04 Apr 2013 06:31:31
6	-	Dolphin	0.450403	555	Thu, 04 Apr 2013 05:37:08 (-6.2d)
7	↑1	jack3	0.451425	455	Thu, 04 Apr 2013 00:41:50
8	↓1	Hopkins Biostat	0.451569	444	Thu, 04 Apr 2013 03:06:25 (-20h)
9	↑2	SouthPole	0.452426	162	Thu, 04 Apr 2013 04:07:48 (-13.2h)
10	↑39	NorthPole	0.452516	64	Thu, 04 Apr 2013 04:55:32 (-3.5d)

There’s only so much information you can extract from a data set




Mapping Dark Matter
kaggle
0.0214
0.0170
0.0160
0.0150
Prediction Accuracy
(lower is better)
Week 1
Week 3
Week 5
Week 7
End
Competition Progress



Office of Science and Technology Policy

Competition Shines Light on Dark Matter

Posted by Jason Rhodes on June 27, 2011 at 04:50 PM EDT

The world's brightest physicists have been working for decades on solving one of the great unifying problems of our universe. It is a problem that explores our place in the cosmos and, as was the case with Newton's law of gravitation and Einstein's theory of relativity, would provide a generational leap in our understanding of the nature of the Universe if solved. Recently, top experts celebrated an exciting breakthrough from an unexpected place.

On May 23, a consortium of the very best from NASA, the European Space Agency, and the Royal Astronomical Society posted the problem on the data-mining website Kaggle and Challenge.gov for all the world to weigh in. In less than a week, Martin O'Leary, a PhD student in glaciology, had crafted an algorithm that outperformed the state-of-the-art algorithms most commonly used in astronomy for mapping dark matter.

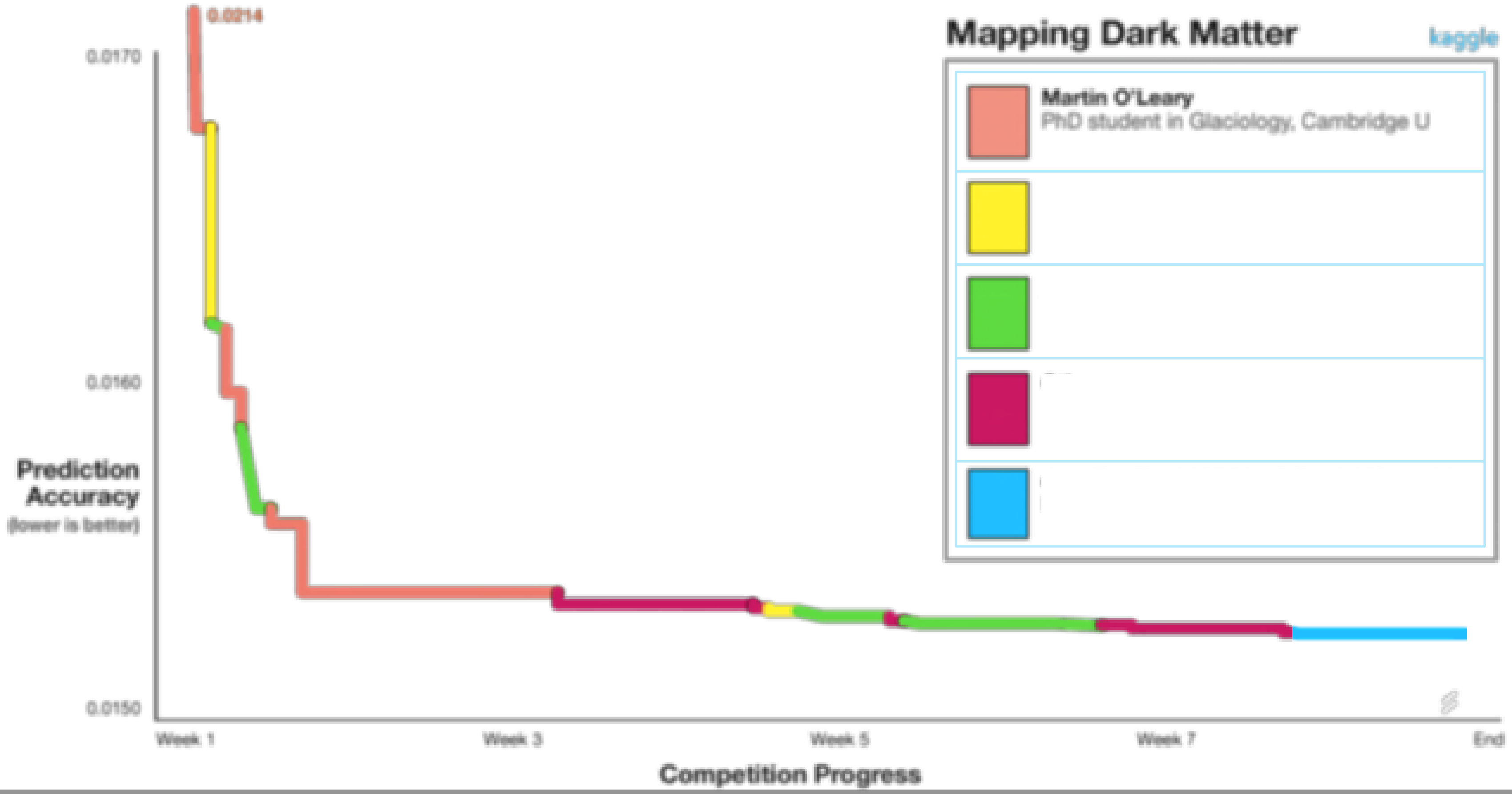
Chalk another one up for the power of crowdsourcing, and this Administration's commitment to using prizes and challenges to find solutions to some of our most pressing problems—here on Earth as well as in the farthest reaches of space!

The posted problem had to do with how scientists can go about mapping "dark matter." Our Universe, it turns out,

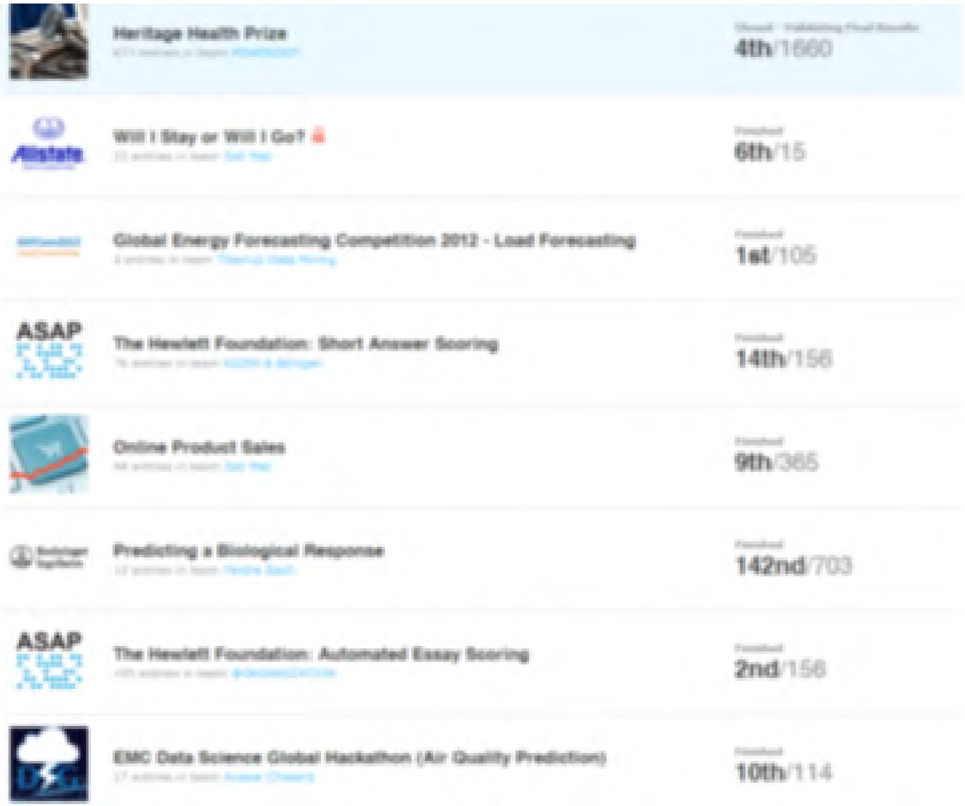
Kaggle's Dark Matter Competition
on the White House blog

"The world's brightest physicists have been working for decades on solving one of the great unifying problems of our universe"

"In less than a week, Martin O'Leary, a PhD student in glaciology, outperformed the state-of-the-art algorithms"

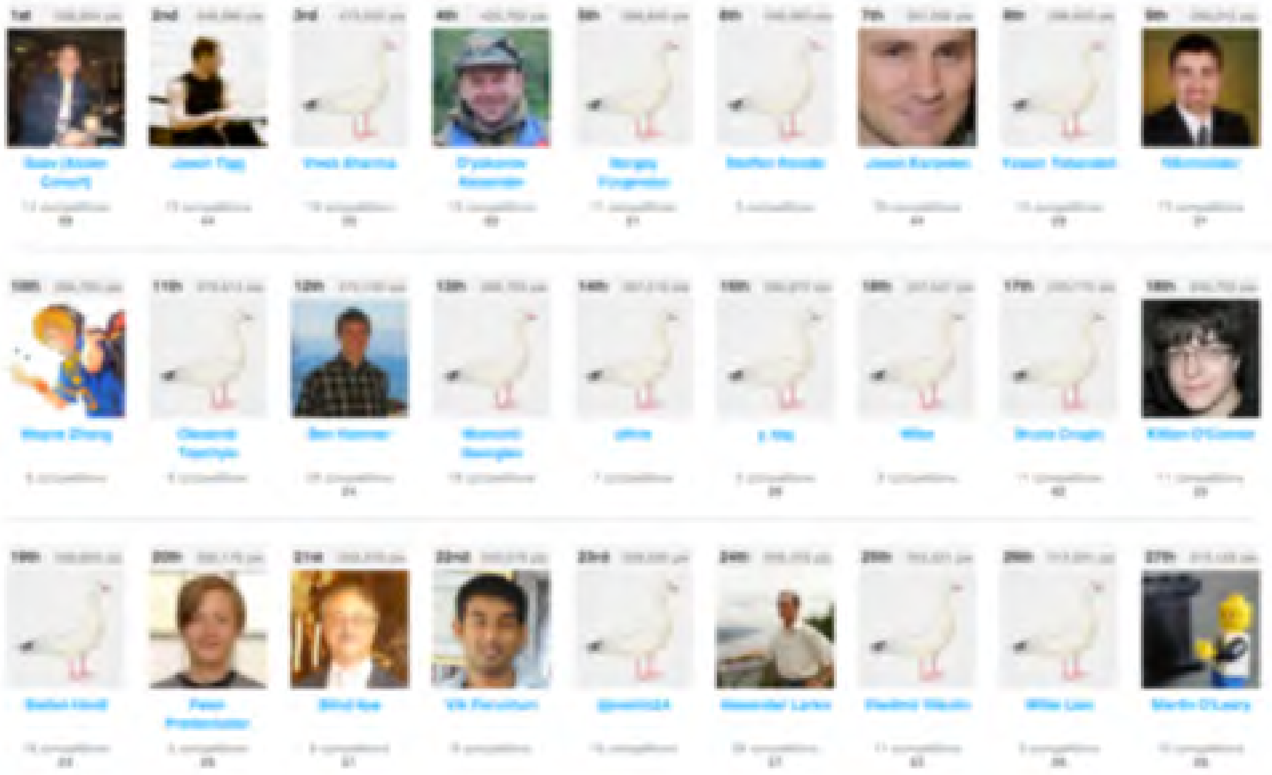

0.0214
0.0170
0.0160
0.0150
Prediction Accuracy
(lower is better)
Week 1
Week 3
Week 5
Week 7
End
Competition Progress
Mapping Dark Matter
kaggle
Martin O'Leary
PhD student in Glaciology, Cambridge U

Phil Brierley has performed well in a huge range of problems



Heritage Health Prize [illegible]	Closed - Validating Final Results 4th/1660
Will I Stay or Will I Go? [illegible]	Finished 6th/15
Global Energy Forecasting Competition 2012 - Load Forecasting [illegible]	Finished 1st/105
The Hewlett Foundation: Short Answer Scoring [illegible]	Finished 14th/156
Online Product Sales [illegible]	Finished 9th/365
Predicting a Biological Response [illegible]	Finished 142nd/703
The Hewlett Foundation: Automated Essay Scoring [illegible]	Finished 2nd/156
EMC Data Science Global Hackathon (Air Quality Prediction) [illegible]	Finished 10th/114

Kaggle Connect provides the people – many of the world's best, ranked



1st [illegible]

2nd Jason Tigg

3rd Vivek Sharma

4th D'yakonov Alexander

5th Sergey Yurgenson

6th Steffen Rendle

7th [illegible]

8th Yasser Tabandeh

9th [illegible]

10th Wayne Zhang

11th [illegible]

12th Ben Hamner

13th Momchil Georgiev

14th [illegible]

15th [illegible]

16th [illegible]

17th Bruce Cragin

18th Killian O'Connor

19th [illegible]

20th Peter Prettenhofer

21st [illegible]

22nd Vik Paruchuri

23rd [illegible]

24th Alexander Larko

25th Vladimir Nikulin

26th [illegible]

27th Martin O'Leary