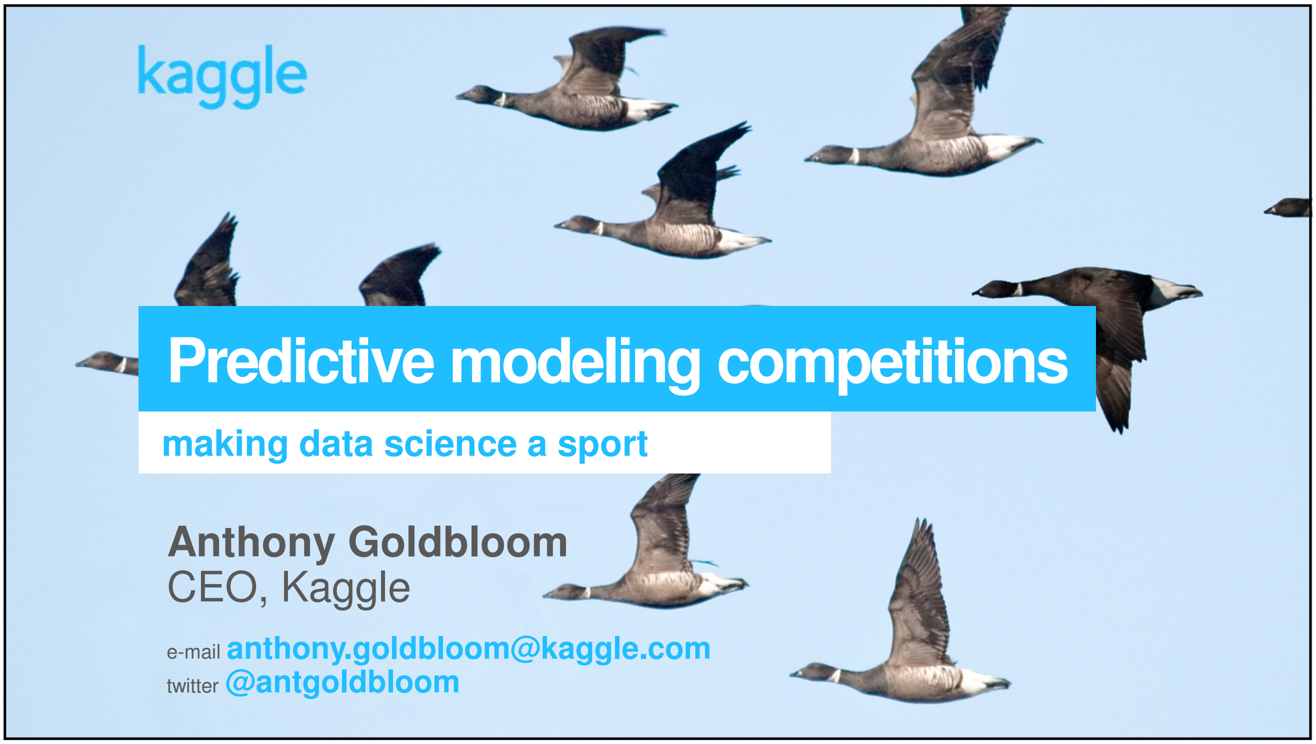


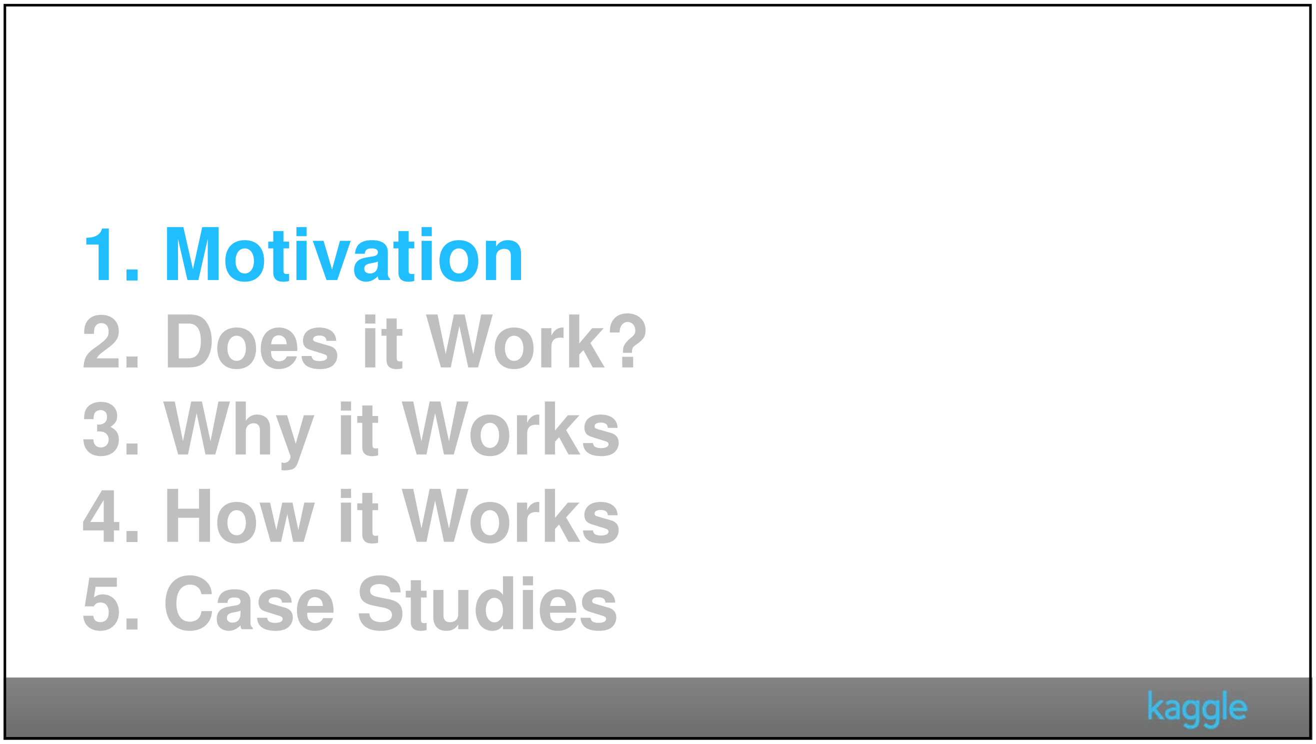
kaggle

Predictive modeling competitions

making data science a sport

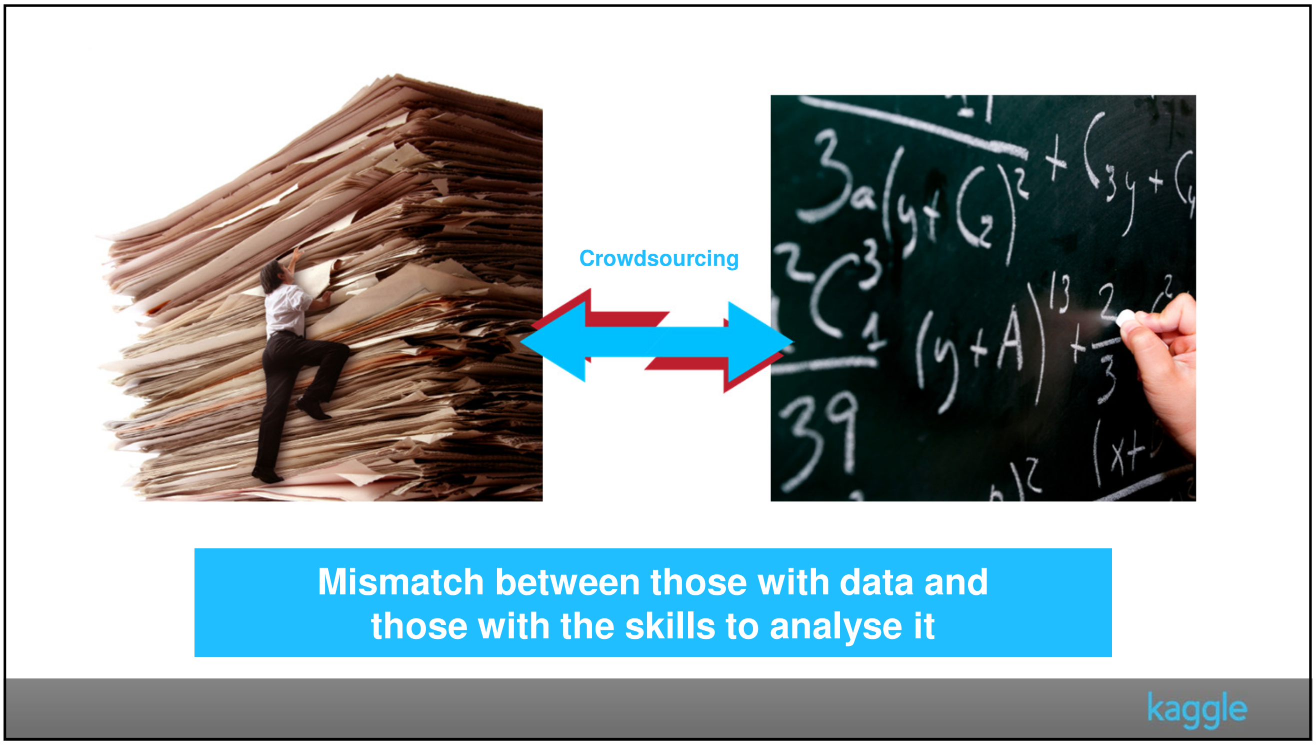
Anthony Goldbloom
CEO, Kaggle

e-mail anthony.goldbloom@kaggle.com
twitter @antgoldbloom



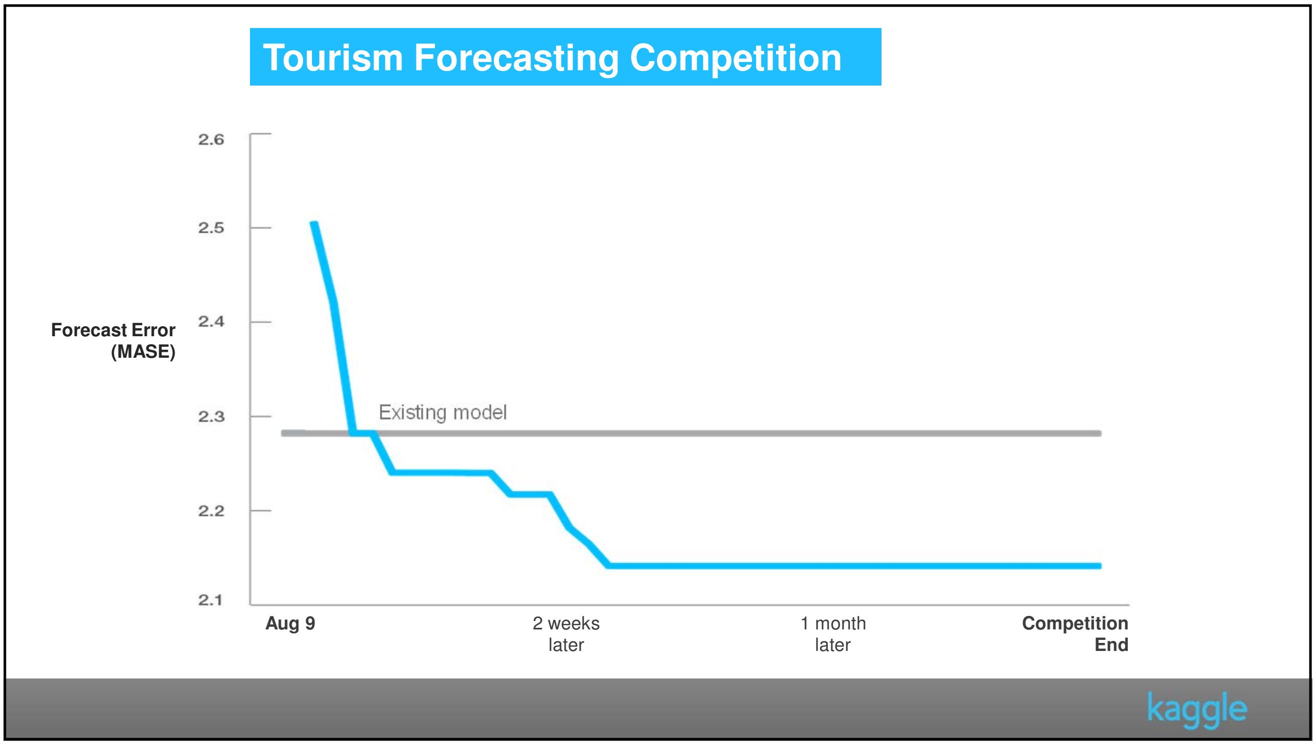
1. Motivation
2. Does it Work?
3. Why it Works
4. How it Works
5. Case Studies

kaggle

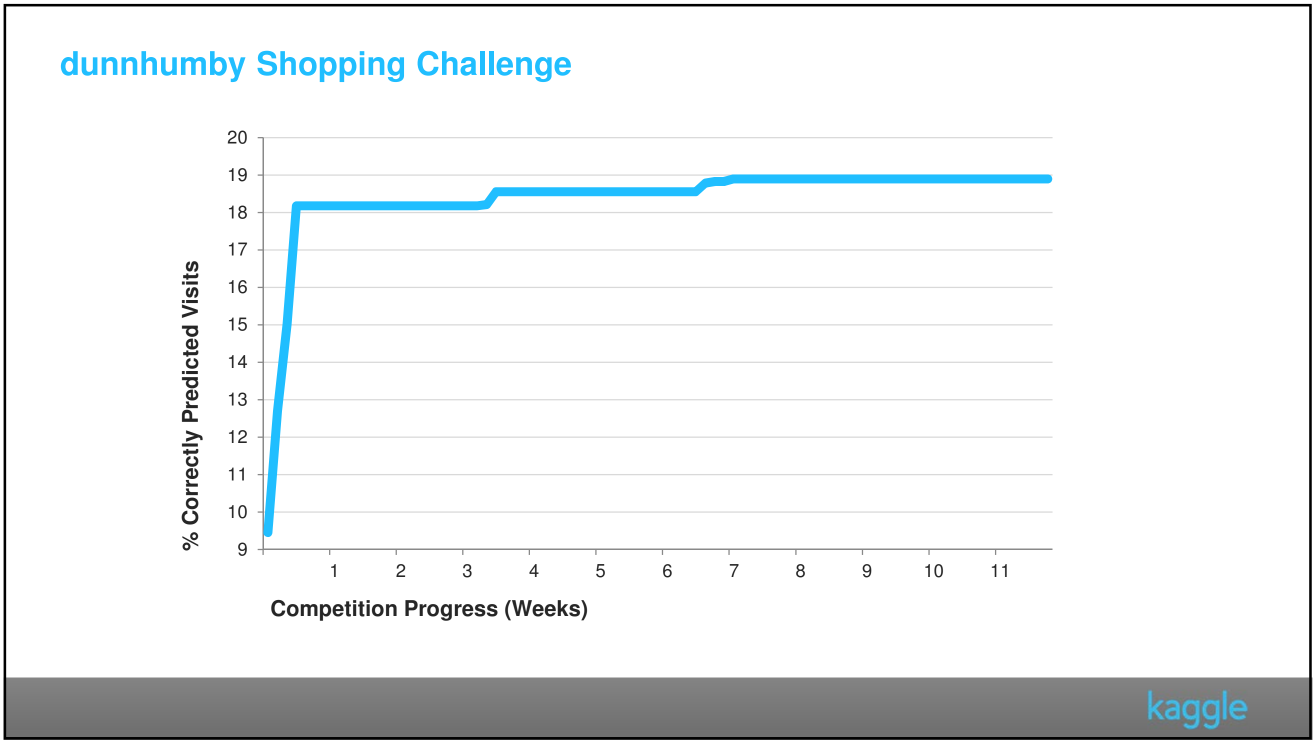


Mismatch between those with data and those with the skills to analyse it

1. Motivation
2. **Does it Work?**
3. Why it Works
4. How it Works
5. Case Studies

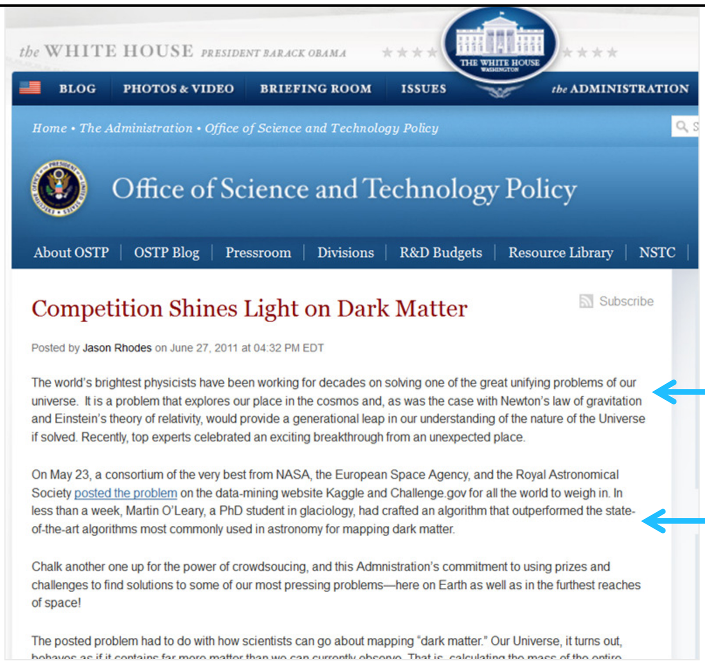


Tourism Forecasting Competition



dunnhumby Shopping Challenge

1. Motivation
2. Does it Work?
3. Why it Works
4. How it Works
5. Case Studies


the WHITE HOUSE PRESIDENT BARACK OBAMA
THE WHITE HOUSE WASHINGTON
BLOG
PHOTOS & VIDEO
BRIEFING ROOM
ISSUES
the ADMINISTRATION
Home • The Administration • Office of Science and Technology Policy
Office of Science and Technology Policy
About OSTP
OSTP Blog
Pressroom
Divisions
R&D Budgets
Resource Library
NSTC
Subscribe

Competition Shines Light on Dark Matter

Posted by Jason Rhodes on June 27, 2011 at 04:32 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 crowdsoucing, and this Admnistration'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 furthest reaches of space!

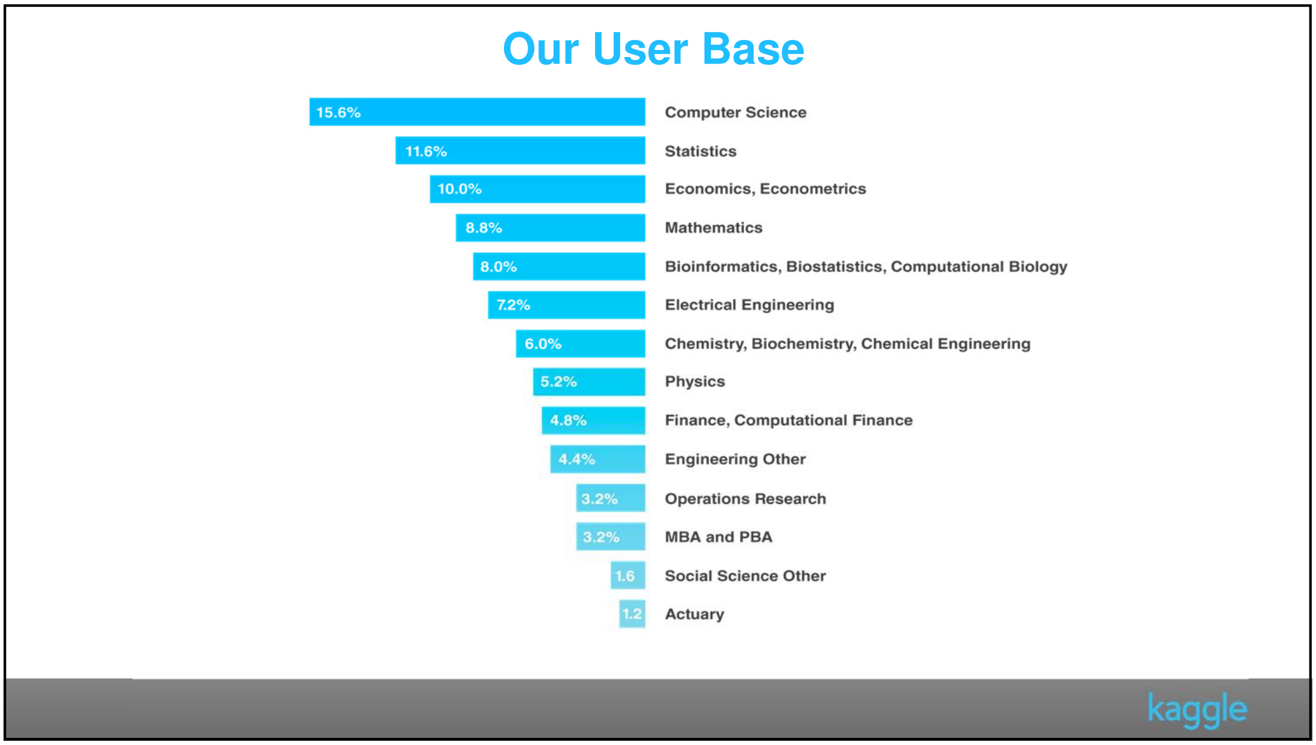
The posted problem had to do with how scientists can go about mapping "dark matter." Our Universe, it turns out, behaves as if it contains far more matter than we can currently observe. That is, calculating the mass of the entire

Kaggle's Dark Matter Competition
on the White House blog

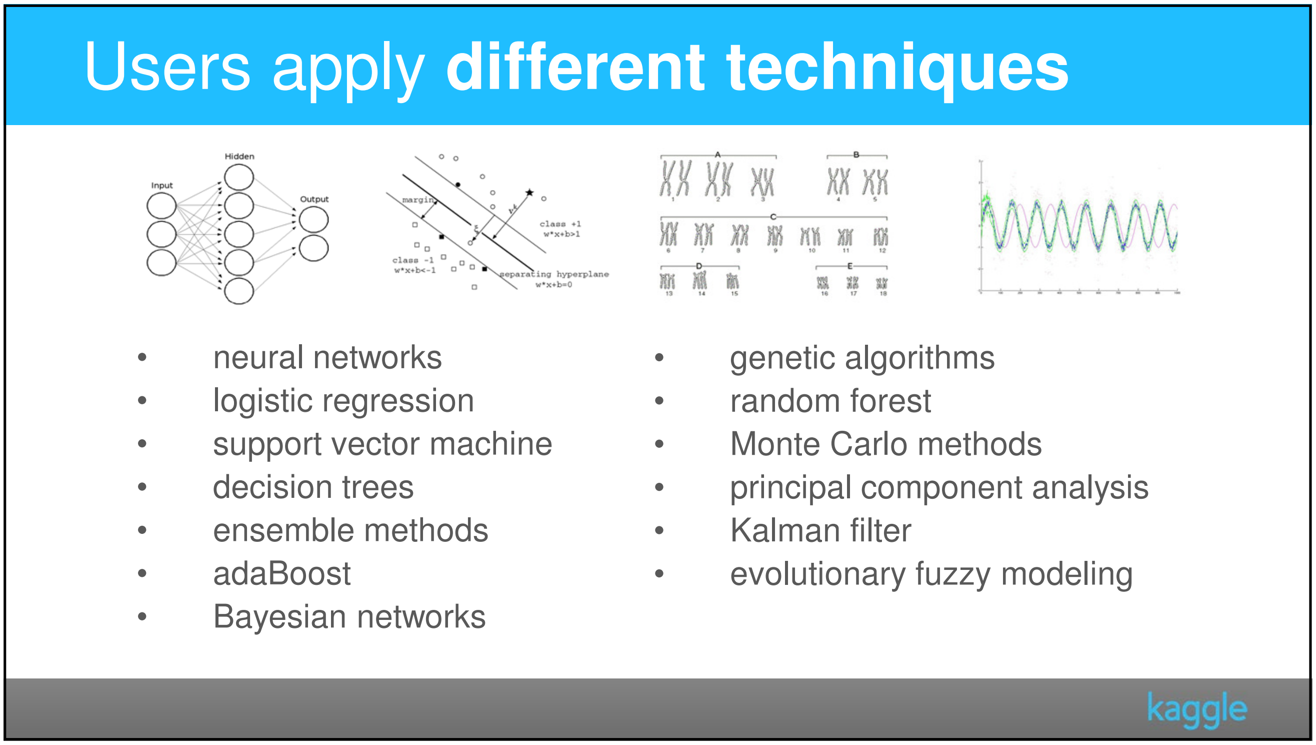
"The world's brightest physicists have been working for decades on solving one of the great unifying problems of our universe"

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

User base: ~16,000 registered data scientists



Our User Base



Users apply **different techniques**

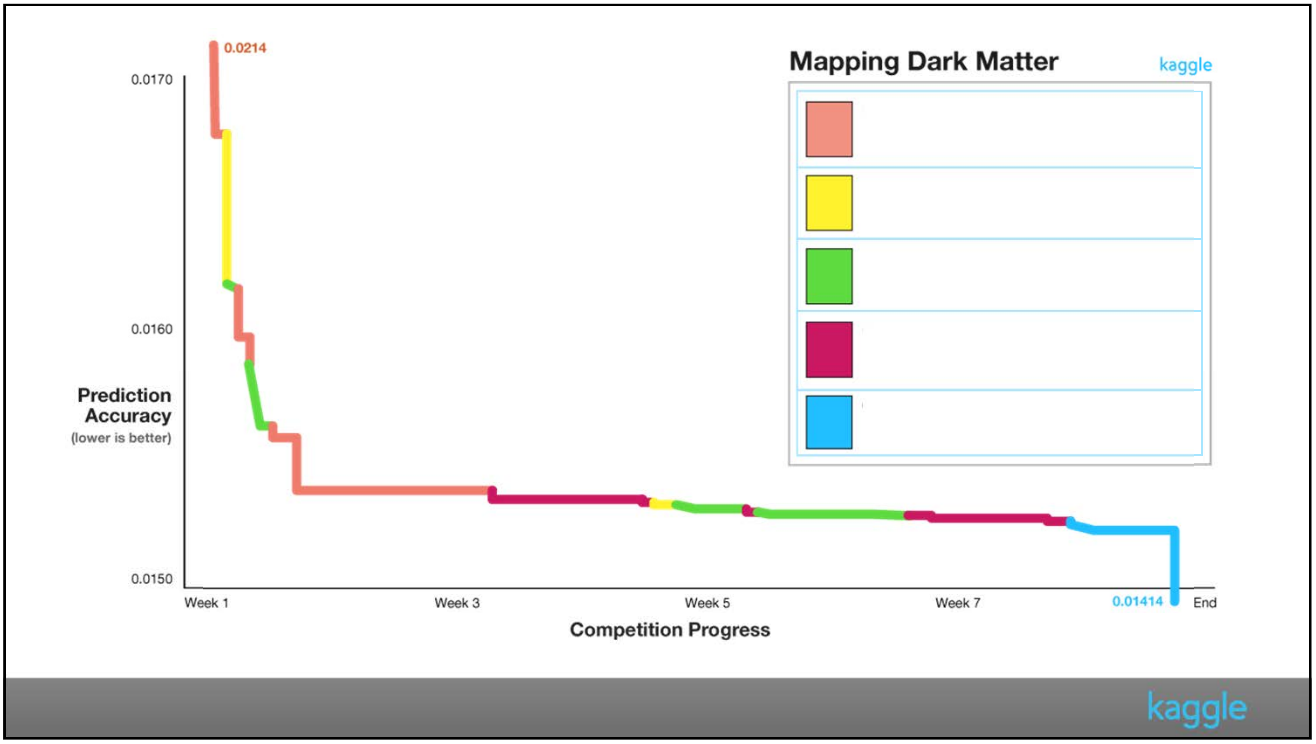
Hidden
Input
Output
margin
class +1
w*x+b>1
class -1
w*x+b<-1
separating hyperplane
w*x+b=0
A
B
C
D
E

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

- genetic algorithms
- random forest
- Monte Carlo methods
- principal component analysis
- Kalman filter
- evolutionary fuzzy modeling

kaggle

Not MIT, not SAS … UoL?

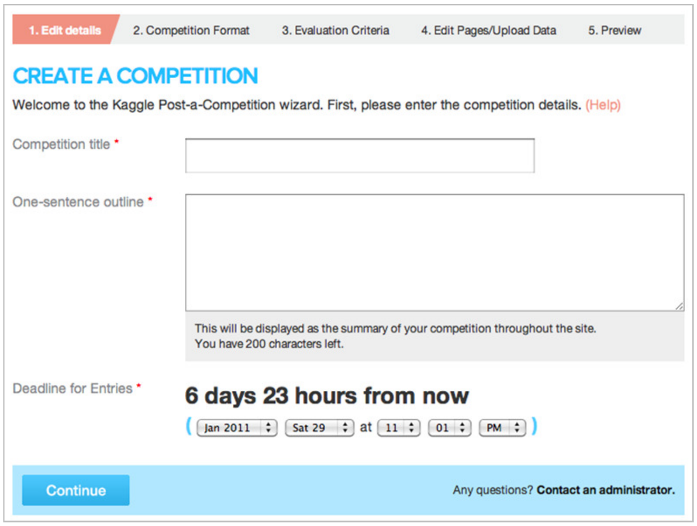


1. Motivation
2. Does it Work?
3. Why it Works
4. **How it Works**
5. Case Studies

1
Upload

2
Submit

3
Evaluate &
Exchange



CREATE A COMPETITION

Welcome to the Kaggle Post-a-Competition wizard. First, please enter the competition details. (Help)

Competition title *

One-sentence outline *

This will be displayed as the summary of your competition throughout the site.
You have 200 characters left.

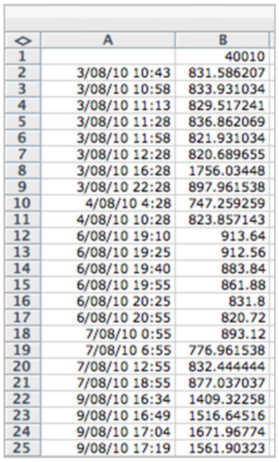
Deadline for Entries *

6 days 23 hours from now

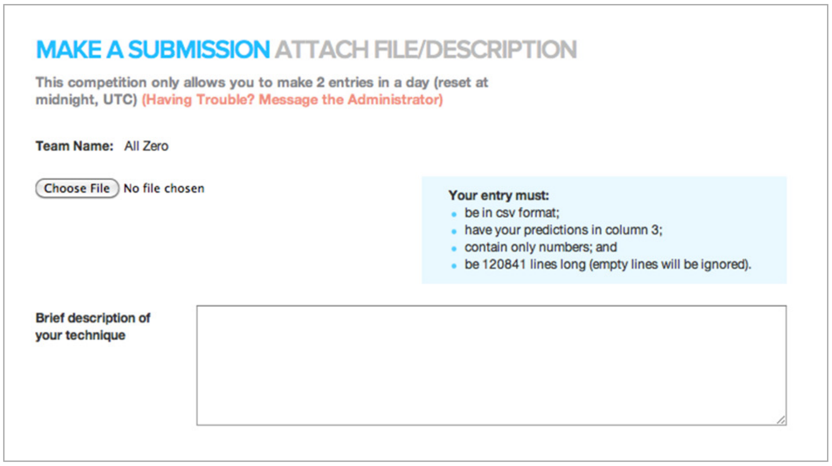
(Jan 2011 Sat 29 at 11 01 PM)

Any questions? **Contact an administrator.**

Use the wizard to post a competition



	A	B
1		40010
2	3/08/10 10:43	831.586207
3	3/08/10 10:58	833.931034
4	3/08/10 11:13	829.517241
5	3/08/10 11:28	836.862069
6	3/08/10 11:58	821.931034
7	3/08/10 12:28	820.689655
8	3/08/10 16:28	1756.03448
9	3/08/10 22:28	897.961538
10	4/08/10 4:28	747.259259
11	4/08/10 10:28	823.857143
12	6/08/10 19:10	913.64
13	6/08/10 19:25	912.56
14	6/08/10 19:40	883.84
15	6/08/10 19:55	861.88
16	6/08/10 20:25	831.8
17	6/08/10 20:55	820.72
18	7/08/10 0:55	893.12
19	7/08/10 6:55	776.961538
20	7/08/10 12:55	832.444444
21	7/08/10 18:55	877.037037
22	9/08/10 16:34	1409.32258
23	9/08/10 16:49	1516.64516
24	9/08/10 17:04	1671.96774
25	9/08/10 17:19	1561.90323



MAKE A SUBMISSION ATTACH FILE/DESCRIPTION

This competition only allows you to make 2 entries in a day (reset at midnight, UTC) (Having Trouble? Message the Administrator)

Team Name: All Zero

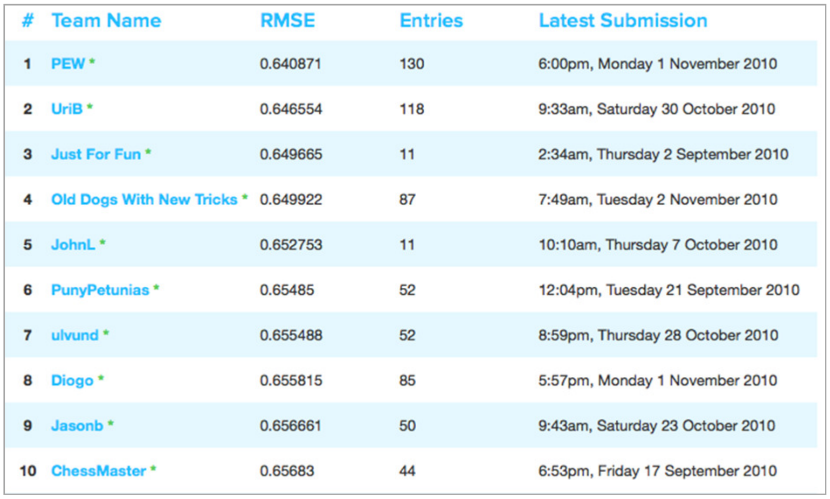
No file chosen

Your entry must:

- be in csv format;
- have your predictions in column 3;
- contain only numbers; and
- be 120841 lines long (empty lines will be ignored).

Brief description of your technique

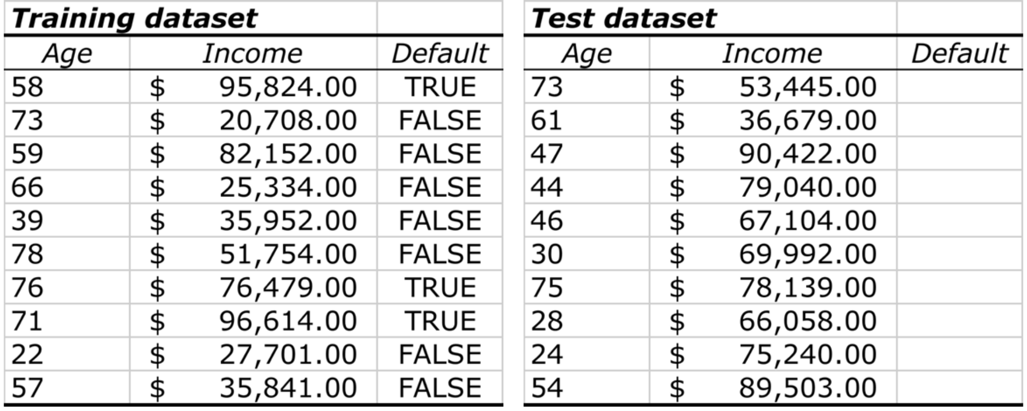
Participants make their entries



#	Team Name	RMSE	Entries	Latest Submission
1	PEW *	0.640871	130	6:00pm, Monday 1 November 2010
2	UriB *	0.646554	118	9:33am, Saturday 30 October 2010
3	Just For Fun *	0.649665	11	2:34am, Thursday 2 September 2010
4	Old Dogs With New Tricks *	0.649922	87	7:49am, Tuesday 2 November 2010
5	JohnL *	0.652753	11	10:10am, Thursday 7 October 2010
6	PunyPetunias *	0.65485	52	12:04pm, Tuesday 21 September 2010
7	ulvund *	0.655488	52	8:59pm, Thursday 28 October 2010
8	Diogo *	0.655815	85	5:57pm, Monday 1 November 2010
9	Jasonb *	0.656661	50	9:43am, Saturday 23 October 2010
10	ChessMaster *	0.65683	44	6:53pm, Friday 17 September 2010

Competitions are judged based on predictive accuracy

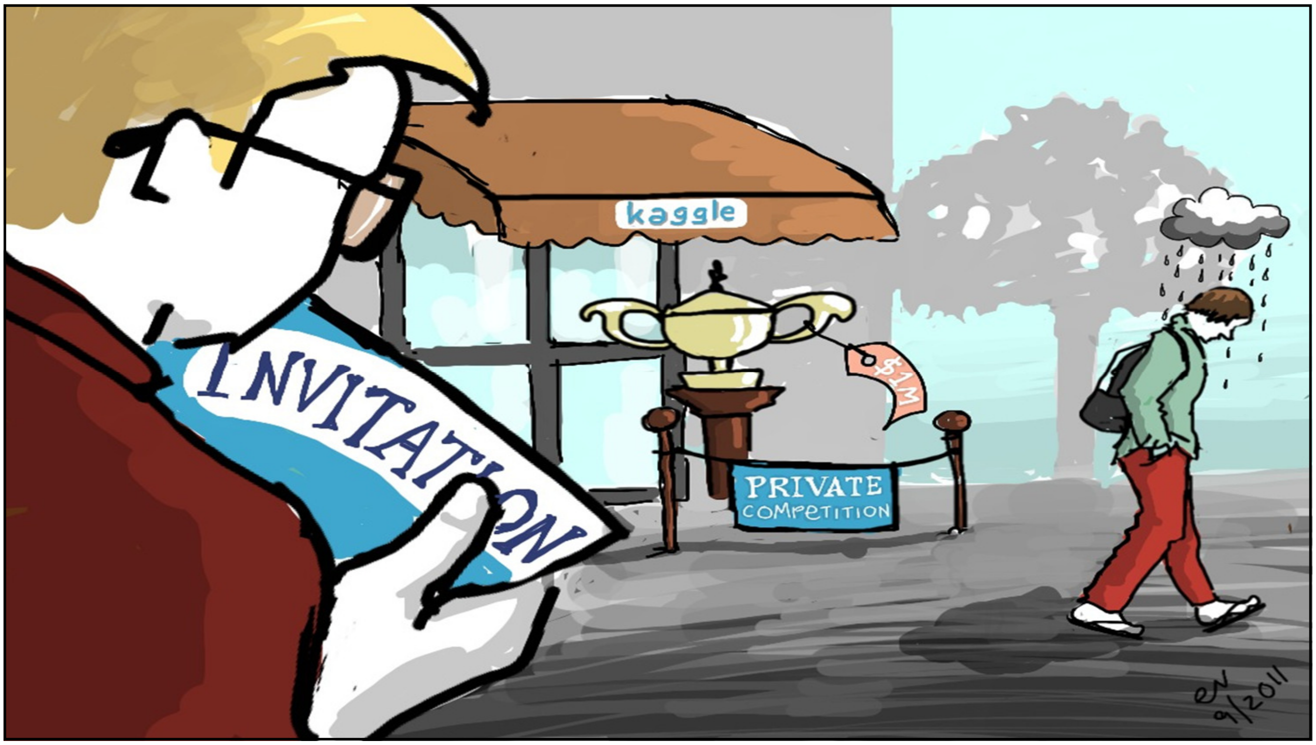
Competition Mechanics



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

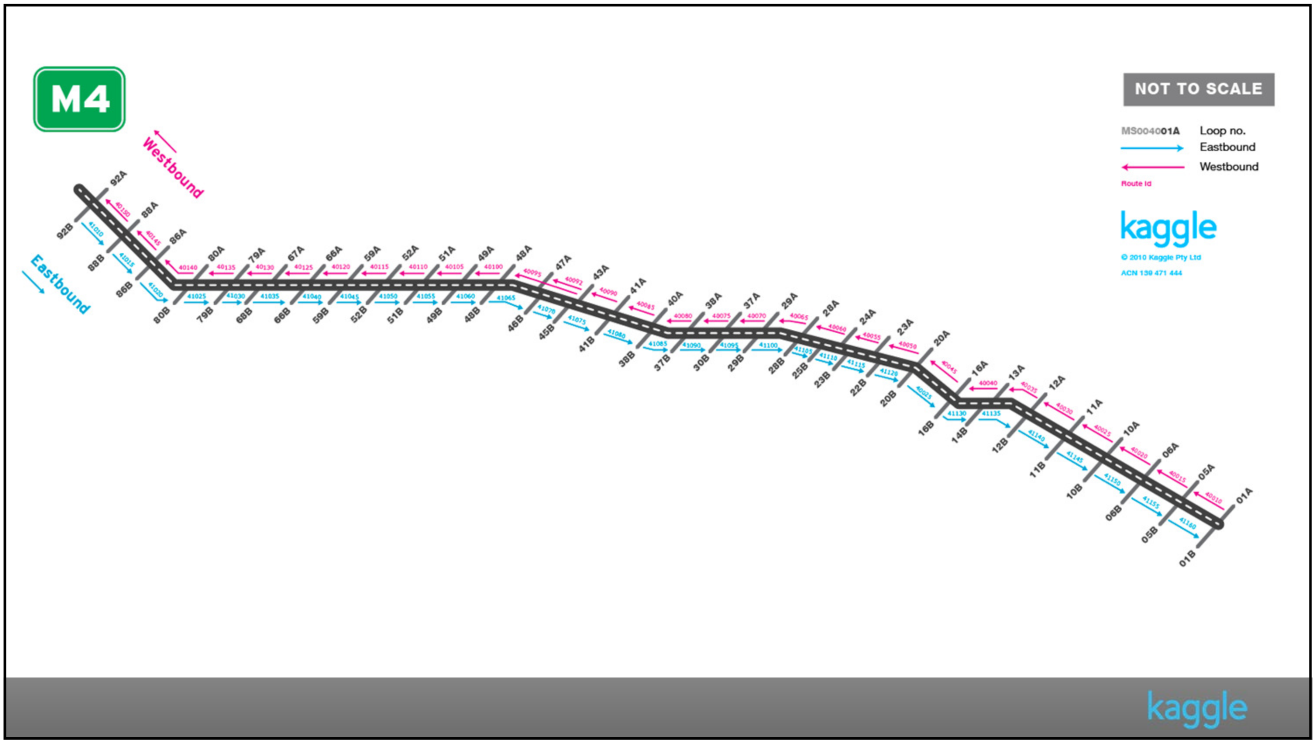
Test dataset		
Age	*Income*	*Default*
73	$ 53,445.00	
61	$ 36,679.00	
47	$ 90,422.00	
44	$ 79,040.00	
46	$ 67,104.00	
30	$ 69,992.00	
75	$ 78,139.00	
28	$ 66,058.00	
24	$ 75,240.00	
54	$ 89,503.00	

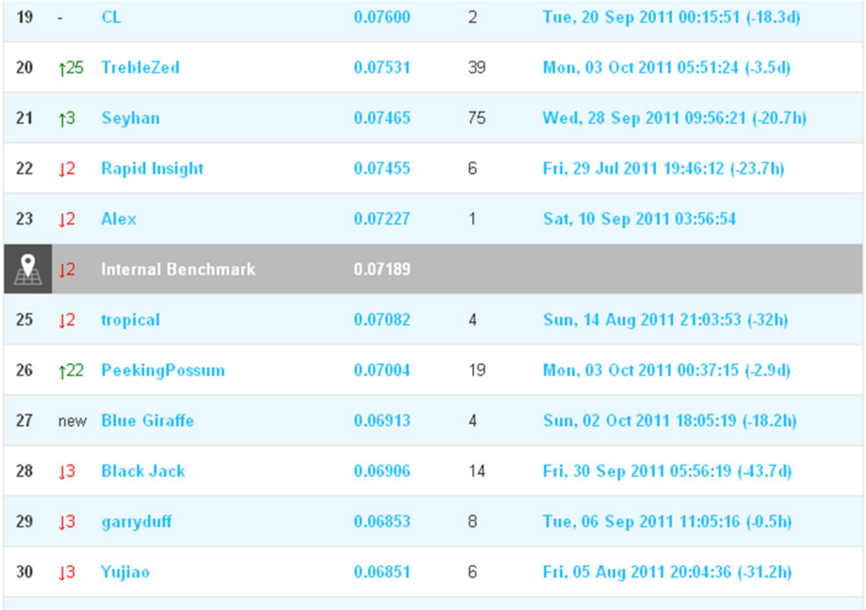
Competitions are judged on objective criteria



1. Motivation
2. Does it Work?
3. Why it Works
4. How it Works
5. Case Studies

Benchmarking





19	-	CL	0.07600	2	Tue, 20 Sep 2011 00:15:51 (-18.3d)
20	↑25	TrebleZed	0.07531	39	Mon, 03 Oct 2011 05:51:24 (-3.5d)
21	↑3	Seyhan	0.07465	75	Wed, 28 Sep 2011 09:56:21 (-20.7h)
22	↓2	Rapid Insight	0.07455	6	Fri, 29 Jul 2011 19:46:12 (-23.7h)
23	↓2	Alex	0.07227	1	Sat, 10 Sep 2011 03:56:54
	↓2	Internal Benchmark	0.07189		
25	↓2	tropical	0.07082	4	Sun, 14 Aug 2011 21:03:53 (-32h)
26	↑22	PeekingPossum	0.07004	19	Mon, 03 Oct 2011 00:37:15 (-2.9d)
27	new	Blue Giraffe	0.06913	4	Sun, 02 Oct 2011 18:05:19 (-18.2h)
28	↓3	Black Jack	0.06906	14	Fri, 30 Sep 2011 05:56:19 (-43.7d)
29	↓3	garryduff	0.06853	8	Tue, 06 Sep 2011 11:05:16 (-0.5h)
30	↓3	Yujiao	0.06851	6	Fri, 05 Aug 2011 20:04:36 (-31.2h)

kaggle

Untouched problems

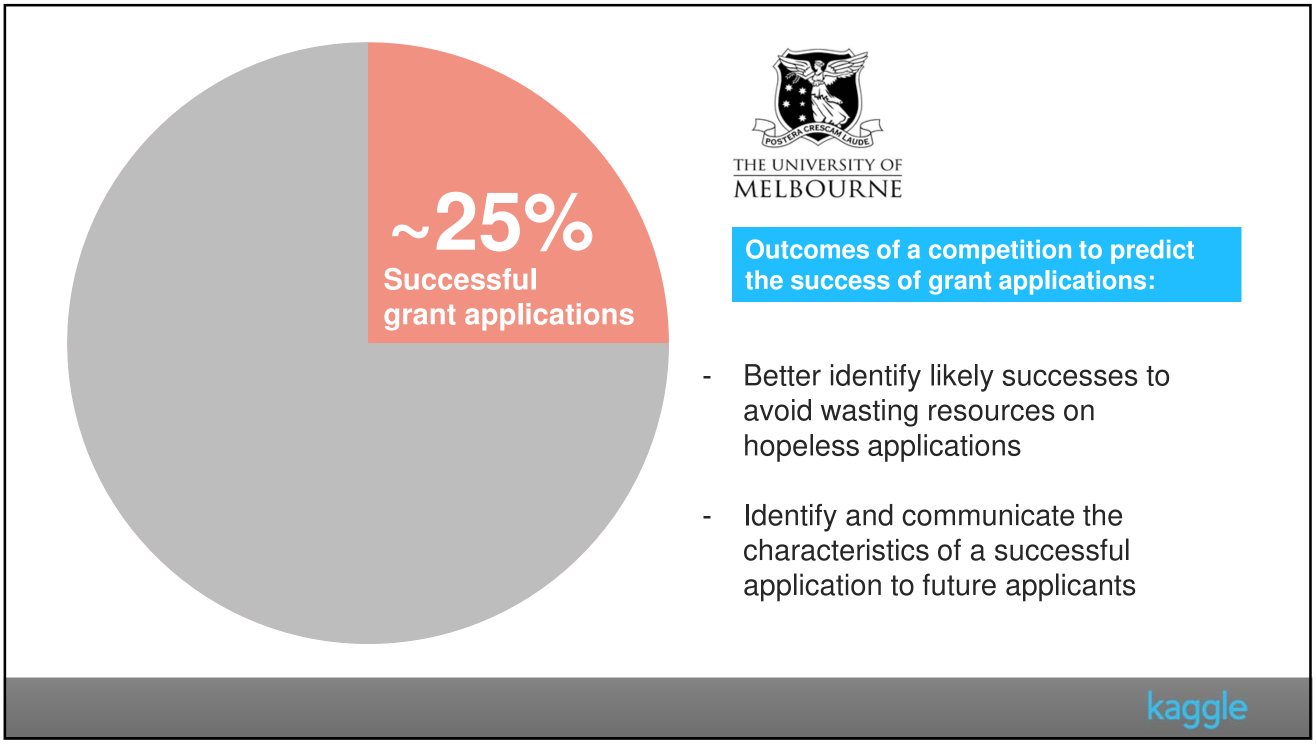
kaggle

HERITAGE PROVIDER NETWORK
HEALTH PRIZE

2011
$3 million prize

HERITAGE PROVIDER NETWORK
Health Prize

kaggle



~25%
Successful
grant applications

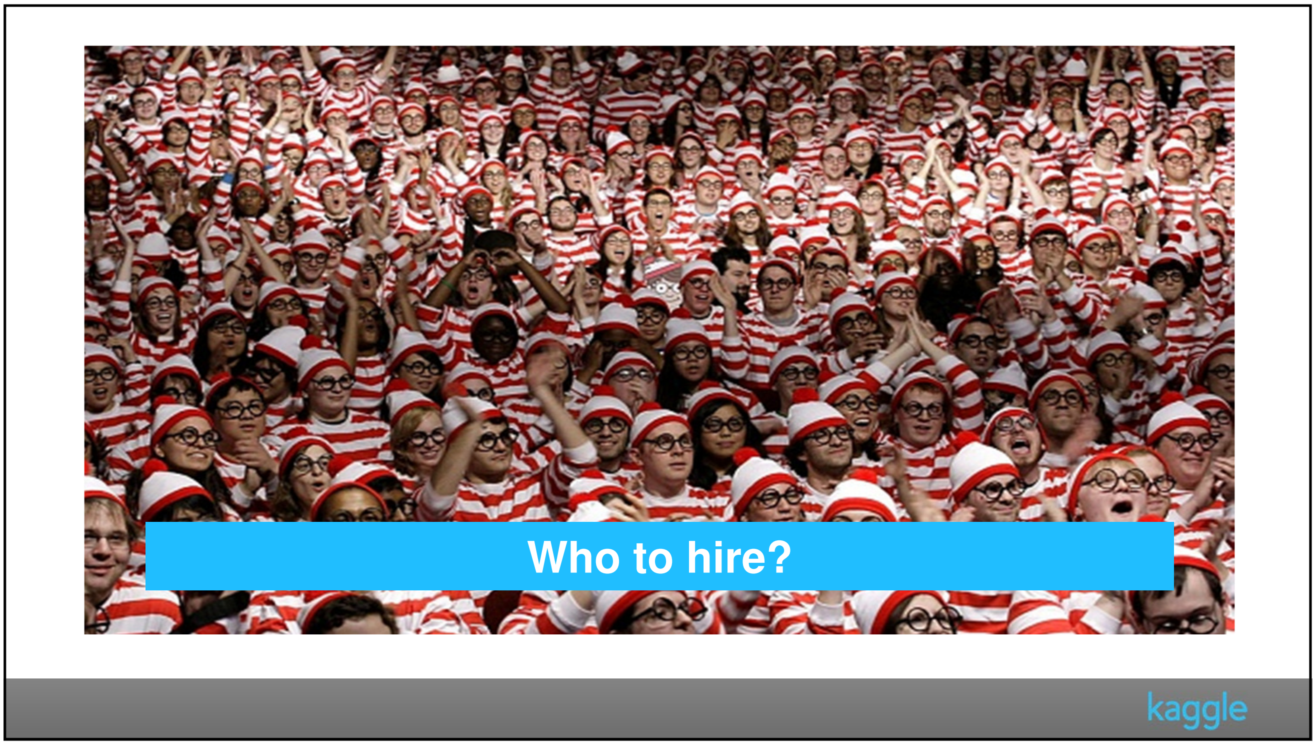
POSTERA CRESCAM LAUDE

THE UNIVERSITY OF
MELBOURNE

Outcomes of a competition to predict the success of grant applications:

- Better identify likely successes to avoid wasting resources on hopeless applications
- Identify and communicate the characteristics of a successful application to future applicants

kaggle



Who to hire?

“Hey, no problem!”

Branding: “we do analytics”

kaggle.com

What could the world’s best analysts find in your data?

e-mail anthony.goldbloom@kaggle.com
phone +1 650 283 9781

Photo by gidzy, www.flickr.com/photos/gidzy