

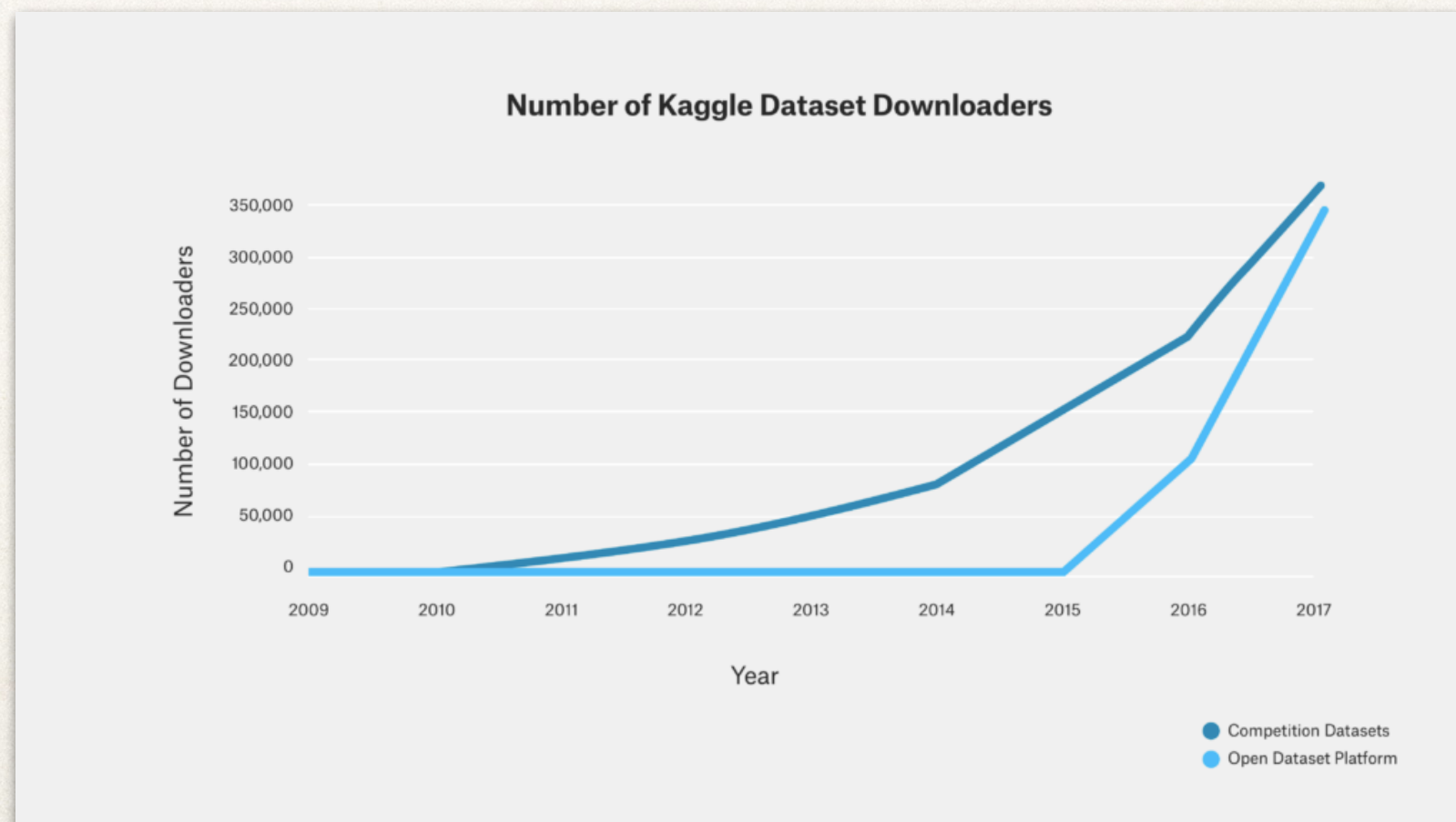
How to compete on Kaggle

Valentin Kuznetsov, Cornell University

SOSC 2018

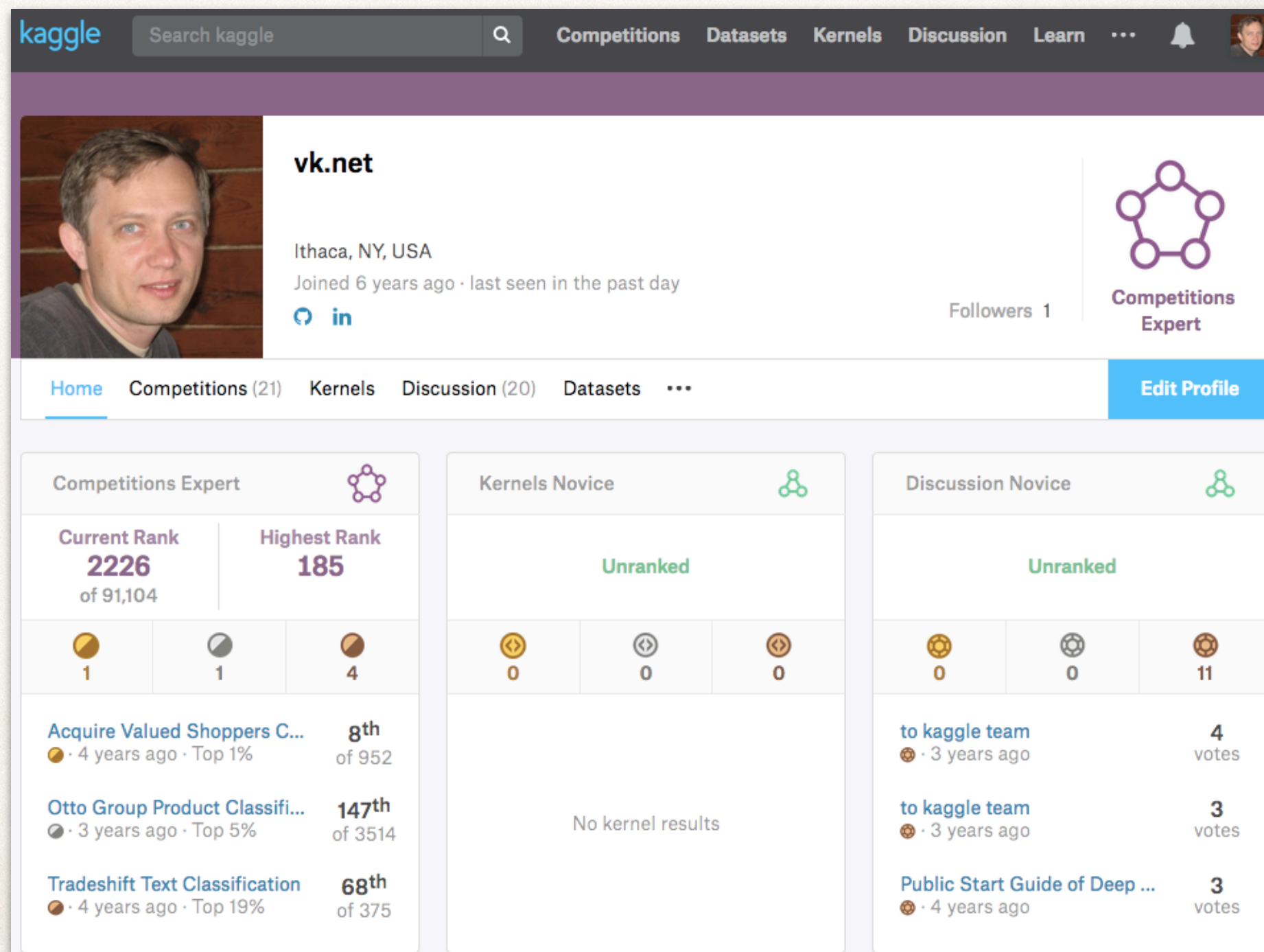
kaggle.com

- ❖ It is open platform for Data Scientist to compete over published datasets
- ❖ In 2017: 120K DataScientists compete in 44 competitions, a total prize sum was \$4.75M+, 600K new users joined, 1.3M total users



My kaggle profile

- ❖ My goal is to learn ML / DL / AI and not prizes
 - ❖ my main source of DataScience
- ❖ I competed alone in my free time apart from regular job, teachings, student projects, family, etc.
 - ❖ turns out it is much tougher to compete alone since amount of info, data, training, ideas significantly increases
- ❖ I was mostly active around 2015



The screenshot shows a Kaggle profile for a user named 'vk.net'. The profile includes a profile picture, location (Ithaca, NY, USA), and join date (6 years ago). The user is a 'Competitions Expert' with 1 follower. The profile is divided into three main sections: Competitions, Kernels, and Discussion. The Competitions section shows a current rank of 2226 (out of 91,104) and a highest rank of 185. It lists three recent competitions: 'Acquire Valued Shoppers C...' (8th of 952), 'Otto Group Product Classifi...' (147th of 3514), and 'Tradeshift Text Classification' (68th of 375). The Kernels section shows 'Unranked' status and 'No kernel results'. The Discussion section shows 'Unranked' status and three recent discussions: 'to kaggle team' (4 votes), 'to kaggle team' (3 votes), and 'Public Start Guide of Deep ...' (3 votes).

vk.net
Ithaca, NY, USA
Joined 6 years ago · last seen in the past day
Followers 1
Competitions Expert

Competitions Expert

Current Rank	Highest Rank
2226 of 91,104	185

Competitions

Competition	Rank
Acquire Valued Shoppers C... 4 years ago · Top 1%	8th of 952
Otto Group Product Classifi... 3 years ago · Top 5%	147th of 3514
Tradeshift Text Classification 4 years ago · Top 19%	68th of 375

Kernels Novice

Unranked
















No kernel results

Discussion Novice

Unranked

Discussion	Votes
to kaggle team 3 years ago	4 votes
to kaggle team 3 years ago	3 votes
Public Start Guide of Deep ... 4 years ago	3 votes

Kaggle competitions

	Acquire Valued Shoppers Challenge Predict which shoppers will become repeat buyers <i>Featured</i> · 4 years ago	  8/952 Top 1%
	Otto Group Product Classification Challenge Classify products into the correct category <i>Featured</i> · 3 years ago · 📁 internet, tabular data	  147/3514 Top 5%
	Tradeshift Text Classification Classify text blocks in documents <i>Featured</i> · 4 years ago	  68/375 Top 19%
	National Data Science Bowl Predict ocean health, one plankton at a time <i>Featured</i> · 3 years ago · 📁 oceanography, image data, multiclass classification	  90/1049 Top 9%
	Homesite Quote Conversion Which customers will purchase a quoted insurance plan? <i>Featured</i> · 2 years ago · 📁 tabular data, binary classification	  102/1764 Top 6%





































❖ [Homesite](#) competition

1,764	1,939	36,387
Teams	Competitors	Entries

Homesite dataset

- ❖ Using an anonymized database of information on customer and sales activity, including property and coverage information, Homesite is challenging you to predict which customers will purchase a given quote. Accurately predicting conversion would help Homesite better understand the impact of proposed pricing changes and maintain an ideal portfolio of customer segments.
- ❖ This dataset represents the activity of a large number of customers who are interested in buying policies from Homesite. Each QuoteNumber corresponds to a potential customer and the QuoteConversion_Flag indicates whether the customer purchased a policy.
- ❖ The provided features are anonymized and provide a rich representation of the prospective customer and policy. They include specific coverage information, sales information, personal information, property information, and geographic information. Your task is to predict QuoteConversion_Flag for each QuoteNumber in the test set.
- ❖ Train sample: 299 columns (28 categorical variables), 260K rows (200MB); test sample 174K rows (131MB)

Homesite leaderboard

1	—	KazAnova Faron clobber	  	0.97024
2	—	Frenchies	  	0.97018
3	▲1	New Model Army CAD & QuY	   	0.97001
4	▼1	Gilberto Leustagos Stanislav	  	0.96988
5	—	The Northern Hemisphere	   	0.96983
6	▲1	victor, clustifier & adam	  	0.96968
7	▼1	monkeys rising	 	0.96961
8	—	A Few with NO Clue	   	0.96960
9	▲2	Daniel FG		0.96959
10	▼1	VinaKago	   	0.96956
100	▼11	BMX	 	0.96793
101	▲23	Overfitters	 	0.96792
102	▲101	vk.net		0.96792

1st place

-0.49%



All Zeros Benchmark



0.50000

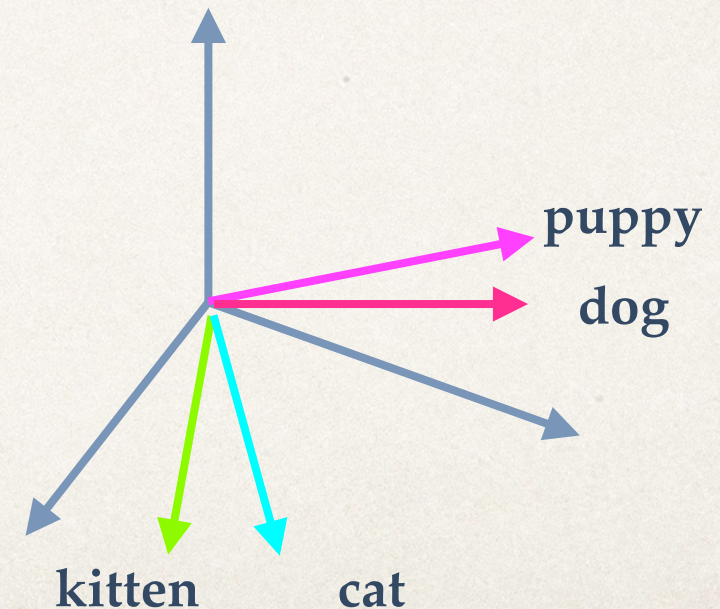
Recipe

- ❖ Node/environment setup
 - ❖ introduction to Anaconda
- ❖ Data exploration
 - ❖ introduction to R
- ❖ System limitations
 - ❖ issues with python, R and others
- ❖ Data preprocessing
 - ❖ intro to Python tools, common format, data scaling, normalization, working with NAs, etc.
- ❖ Training and modeling
- ❖ Reaching the limit
 - ❖ embeddings, stacking, etc.

Word embedding

- ❖ A way to capture multi-dimensional relationships between categories
 - ❖ e.g. Sun and Sat may have similar effect while other days may be treated independently
 - ❖ you define a dimension of word vector up-front
 - ❖ it projects categorical variables into another phase space, e.g. days may be sunny or rainy, season or off season; all of these features are hidden from original data representation
- ❖ Use NN or other ML algorithms to train the model to find best representation of embedded variables

puppy	[0.9, 1.0, 0.0]
dog	[1.0, 0.2, 0.0]
kitten	[0.0, 1.0, 0.9]
cat	[0.0, 0.2, 1.0]



Embeddings recipe

- ❖ Identify categorical variables and order them
- ❖ Define embedded matrix and cardinality of categorical variable
- ❖ Perform one-hot-encoding
- ❖ Train Neural Network model
- ❖ Extract NN weights (embeddings matrix)
- ❖ Plug embeddings matrix into regular ML model instead of categorical variable
- ❖ Train ML model with embeddings matrices