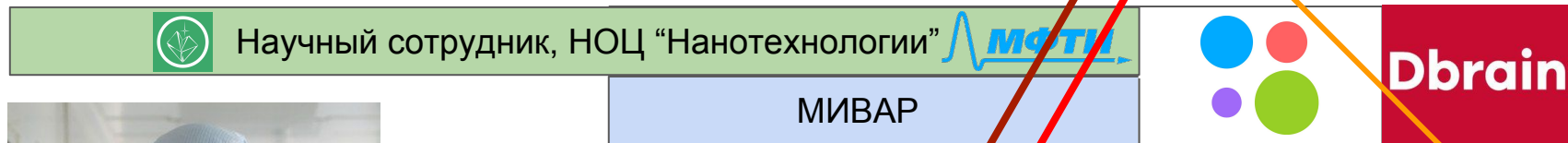
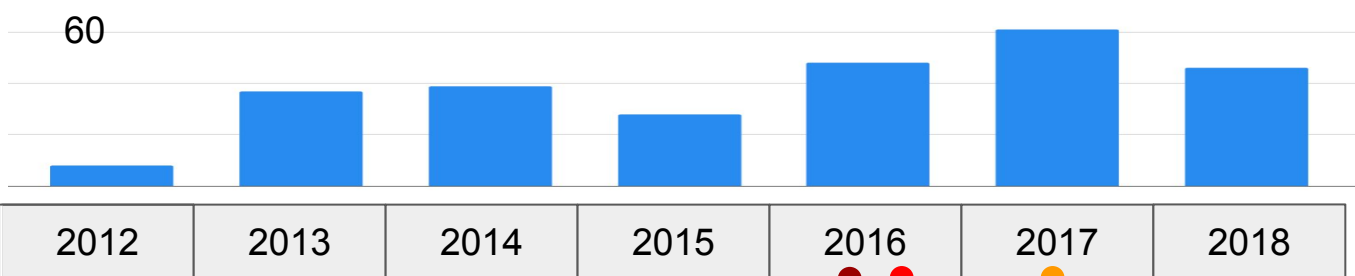


# DL-соревнования гуси, пайплайны, кулстори

Артур Кузин  
Lead Data Scientist, Dbrain

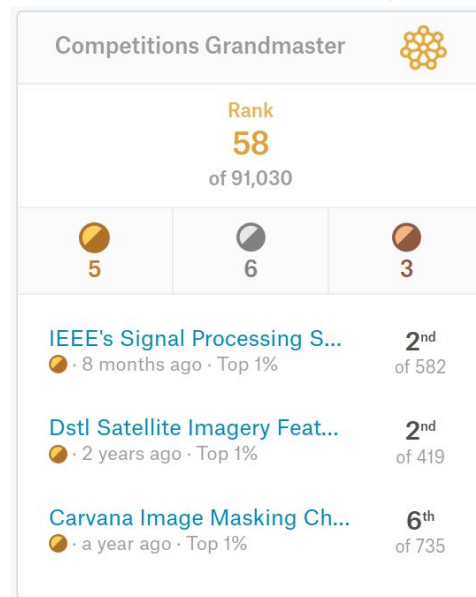
# Timeline



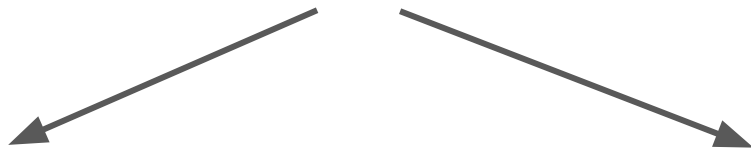
2008–2016  
50+ публикаций  
h-index: 9

Avito-2016: Распознавание  
марки и модели автомашин  
на изображениях  
**3 место**

Avito-2016:  
Распознавание категории  
объявления  
**1 место**



# Хочу быть дата саентистом, что делать?



Пройти курс(ы)  
(Andrew Ng, Воронцов,  
cs231n, OpenDataScience, etc)

Зарешать соревнование  
(kaggle, Topcoder, dataring,  
challenger.ai)

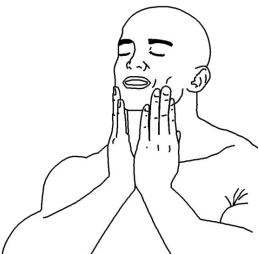
Слишком абстрактно  
Сложно найти мотивацию

Формализованная задача  
Конкуренция  
Тусовка  
Свое решение как проект

...



# Почему же это так весело?



67 ↑22 DataWookie

12676464579.14730

5







Wed, 09 Dec 2015 09:03:52

## Your Best Entry ↑

You improved on your best score by 811743125.21004.

You just moved up 56 positions on the leaderboard.

 Tweet this!

7	<span>▲32</span>	Human Analog		0.68159
8	<span>▲5</span>	dudemeister		0.67481
9	<span>▲1</span>	yanchen036		0.66545
10	<span>▼6</span>	Dimensionality Diabolists		0.66232
11	<span>▼6</span>	Adam Blazek		0.65954
12	new	Miha Skalic		0.65623







# Первая доза: запустить baseline

Overview Data **Kernels** Discussion Leaderboard Rules Team

---

All Mine

---

133			End-to-end baseline with U-net (keras) run 8 months ago by <a href="#">n01z3</a>
104			Full pipeline demo: poly -> pixels -> ML -> poly run 8 months ago by <a href="#">Konstantin Lopuhin</a>
63			[LB 0.42]Ultimate full solution (run on your HW) run 7 months ago by <a href="#">Sergey Mushinskiy</a>

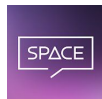
# EDA



Перед тем как решать задачу,  
посмотрите на данные к ней

(с) Евгений Нижибицкий

Data Fest<sup>2</sup> Minsk 2018: Изобразительные лики



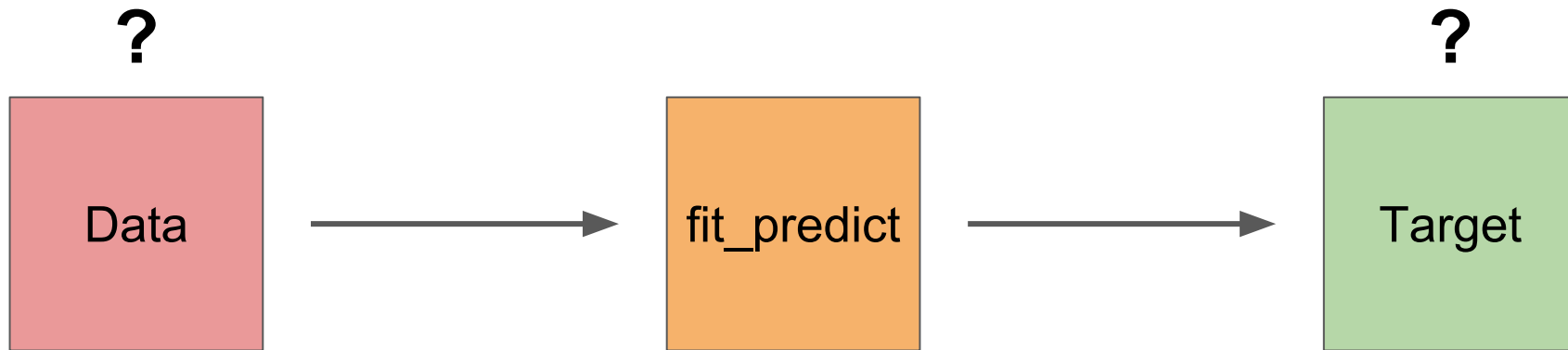
<https://youtu.be/Twli7zoiIPs>

История про лик в Kaggle Airbus Ship Detection



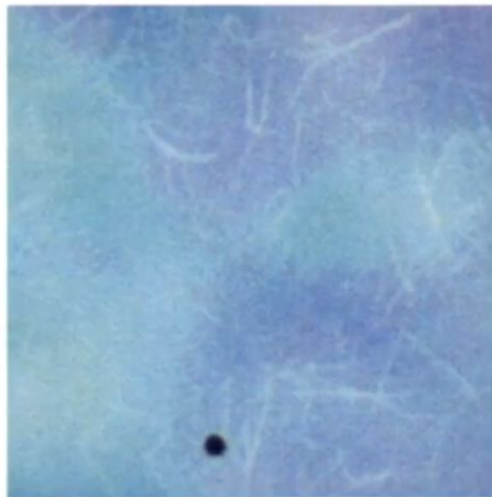
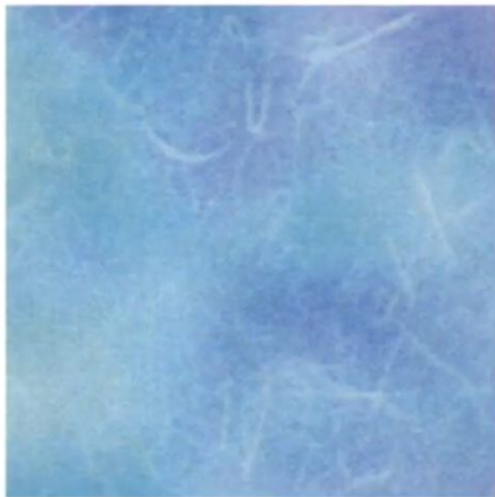
<https://youtu.be/MIbetMAnC04>

# Понимание задачи



Дайте мне ~~точку опоры~~ правильный таргет, и я ~~переверну землю~~ затащу компетишн (с) Архимед Стетхем

# Konica-Minolta - Pathology Segmentation





# Topcoder - Konica-Minolta-2

Rank	Handle	Provisional Rank
1	nizhib	1
2	n01z3	2
3	albu	3
4	codecrux	4
5	selim_sef	6
6	cannab	7
7	wleite	5
8	ipraznik	8
9	nofto	9
10	Mloody2000	11



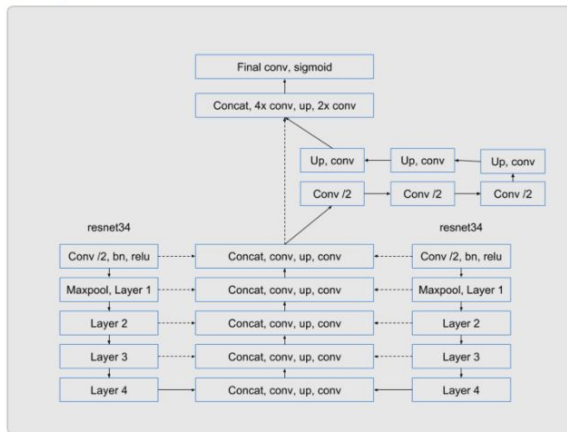
nizhib 5:53 PM

1. `cv2.findTransformECC(img, ref, np.eye(2, 3, dtype=np.float32), cv2.MOTION_EUCLIDEAN, criteria)` для выравнивания
2. `augment.ScaleAndCrop(scale=0.9858, padding=3)` для исправления кривых данных
3. Обычный юнет прямо как в учебнике, без весов
4. ???
5. Топ-1

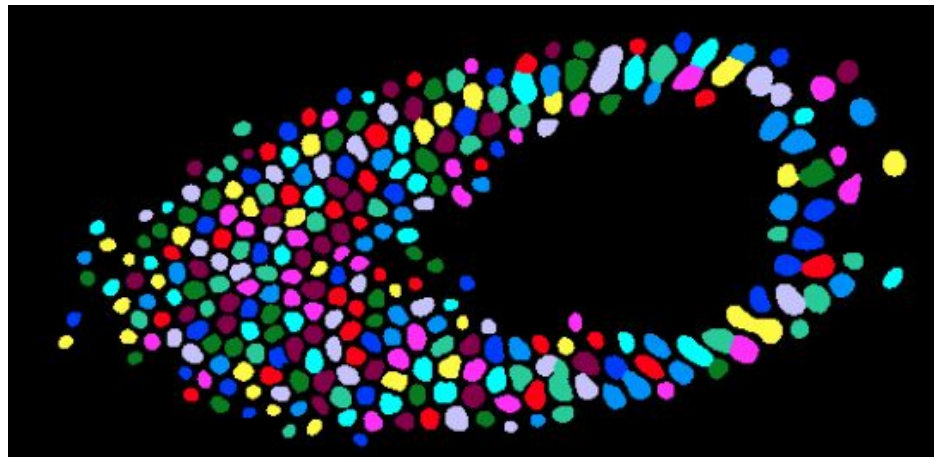
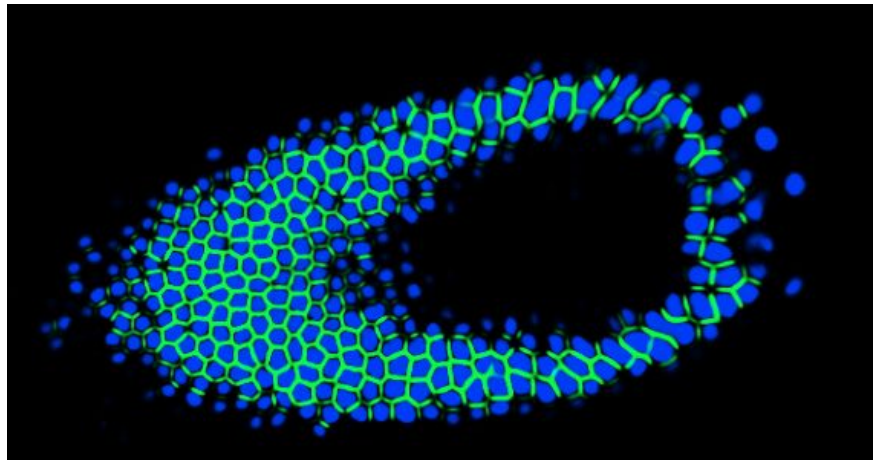


albu 6:15 PM

wnet.png

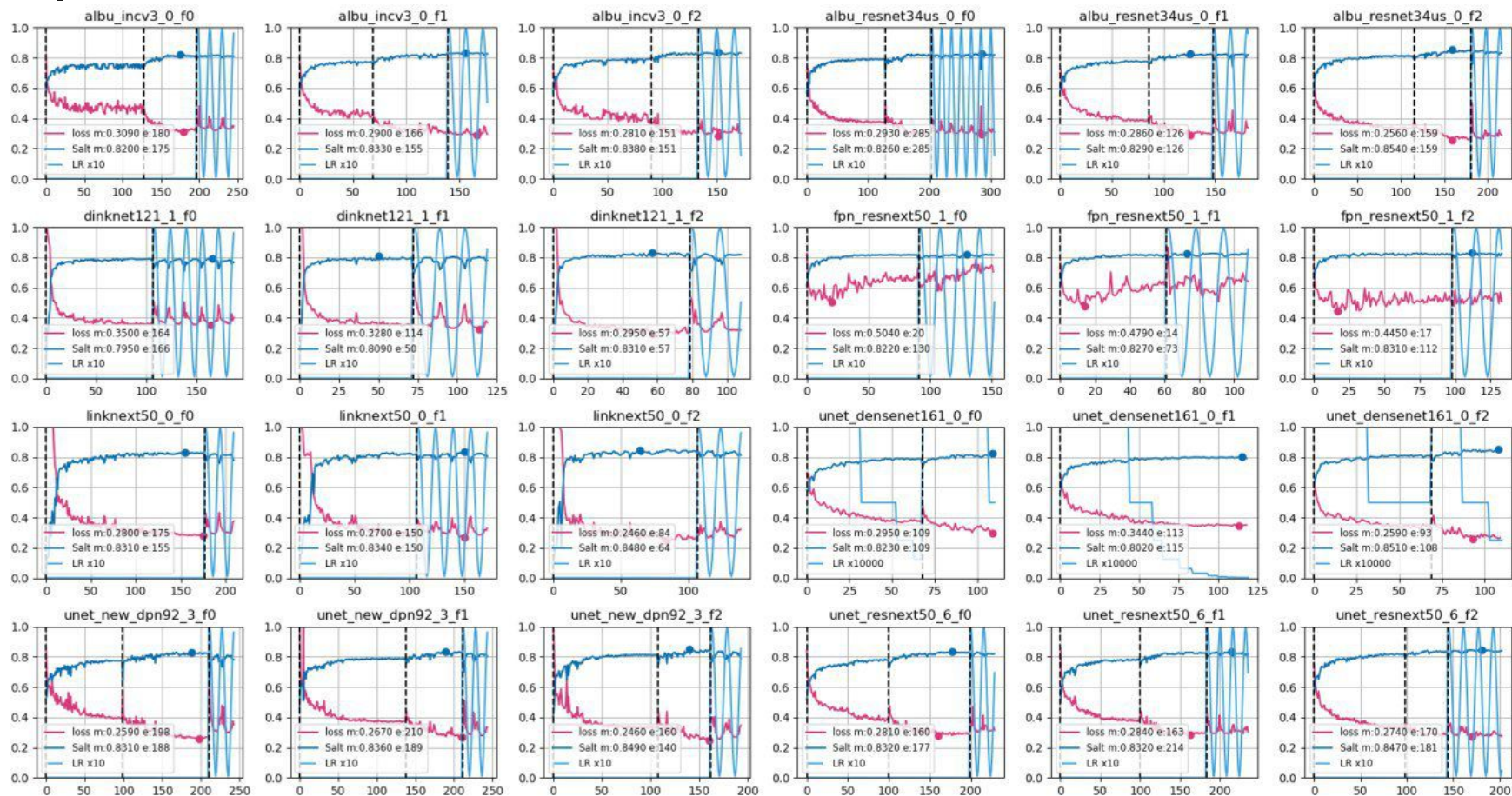


# Kaggle - Data Science Bowl 2018



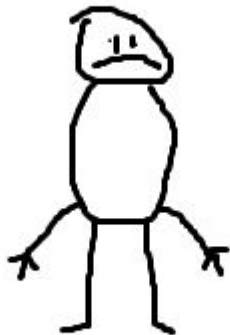
<https://www.kaggle.com/c/data-science-bowl-2018/discussion/54741>

# Pipeline



# Необходимые атрибуты хорошего пайплайна

- Config запуска
- Логирование
- Модульность
- Воспроизводимость
- Переиспользуемость



# Есть решение - catalyst!



Сергей Колесников  
Senior Data Scientist  
Dbrain

~~<https://github.com/Scitator/pytorch-common>~~



~~<https://github.com/Scitator/prometheus>~~



<https://github.com/Scitator/catalyst>

## Features

- Universal train/inference loop.
- Key-values storages.
- Data and model usage standardization.
- Configuration files - yaml for model/data hyperparameters.
- Loggers and Tensorboard support.
- Reproducibility - even source code will be saved to logs.
- OneCycle lr scheduler and LRFinder support.
- FP16 support for any model.
- Corrected weight decay (AdamW).
- N-best-checkpoints saving (SWA).
- Training stages support - run whole experiment by one command.
- Logdir autonaming based on hyperparameters - make hyperopt search easy again.
- Callbacks - reusable train/inference pipeline parts (and you can write your own if needed).
- Well structured, so you can just grab a part to your project.





# Наигрался один? Есть результат? Объединяйся в команду!



**zfturbo** 5:57 PM

Если кто есть в ~ТОП50, кто хочет за золото побороться, пишите.  
Ансамбли заходят очень хорошо. Остался один день на объединение.



**vilmar** 6:39 PM

Если кто ниже объединиться хочет - я тоже открыт для предложений, а то вернулся к задачке, но пока на валидации 0.9961 🐸



**3 replies** Last reply 15 days ago

# Командное взаимодействие



## Общий git



easygold / amazon\_from\_space

Use satellite data to track the human footprint in the Amazon rainforest

Name	Last commit
data	Add flat sample test
albu	Merge remote-tracking branch 'origin/master'
alno	New ensemble
kostia	final heuristics notebook
n01z3	resnet 50 predict
nizhib	Add test predicting script
romul	start train vgg16 with batch normalization
.gitignore	Add TTA predictor
prepare.py	Add flat sample test
prepare_splits.py	rm extra blank line



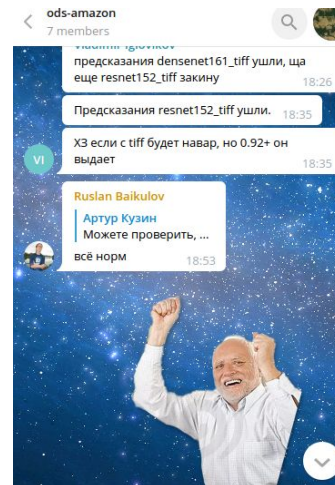
## Общий формат предсказаний

```
df.to_hdf(out_path, 'w', index_label='images')
```

```
amazon_from_space > predictions > nizhib
```



## Общий чатик



Окей, я дошел до конца, потюнив чужое решение.  
Что я вынес?

- Опыт запуска чужого кода
- Привычку смотреть в данные
- Понимать задачу и подбирать таргет
- Ведение экспериментов
- Командное взаимодействие
- Свое решение как проект





Как это помогает компаниям?

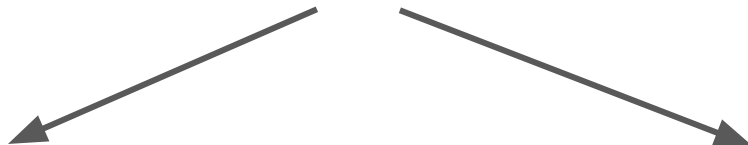
# Свой компетишкен с бигдатоу и призами




## Avito Duplicate Ads Detection


Can you detect duplicitous duplicate ads?

\$20,000 · 548 teams · a year ago









**u1234x1234**  
Chelyabinsk, Russia  
Joined 3 years ago · last seen 8 days ago  
[Q](#) [T](#) [in](#)



Competitions Master

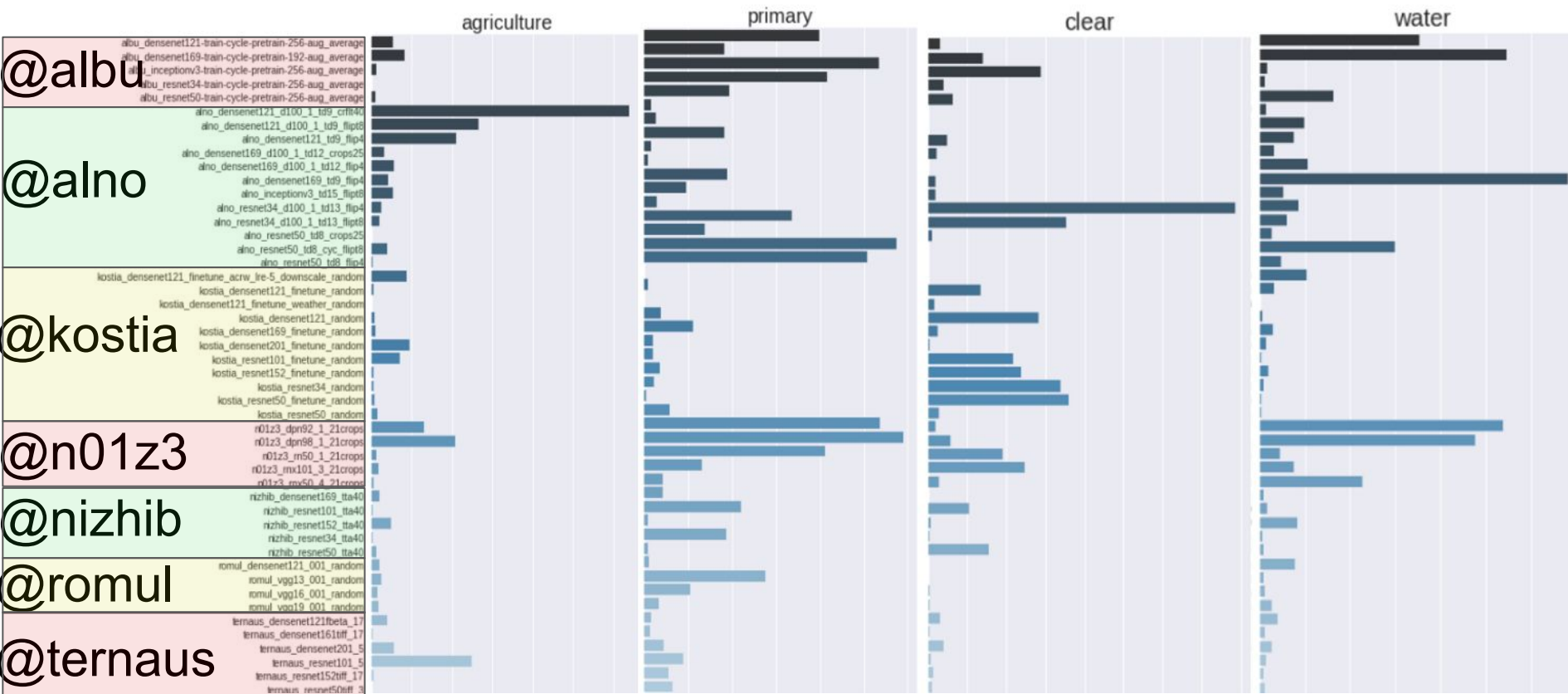
[Home](#) [Competitions \(23\)](#) [Discussion \(14\)](#) [Contact User](#)

Competitions Summary

 <p>Competitions Master</p>	<p>Rank <b>16</b> of 47,430</p> <div><p>3</p><p>0</p><p>6</p></div>	<p>Competitions: 20 Solo: 18 (90%) Team: 2 (10%)</p>
---	--	--

- Формализация задача
- Инсайты из данных

# Model Importance



# Обзор лучших решений

24



**3rd place Solution**

[ZFTurbo](#) a month ago

29



**summary of top-15 methods**

[Heng CherKeng](#) 2 months ago

1



**Pytorch Kaggle Start Kit**

[Brendan Fortuner](#) a month ago

15



**New to 14th in 1 week**

[To Train Them Is My Cause](#) 2 months ago

17



**let's repeat the top results**

[Heng CherKeng](#) 2 months ago

28



**San Francisco. August 8. Meetup. Team ods.ai 7th place solution.**

[Vladimir Iglovikov](#) 2 months ago

86



**My brief overview of my solution**

[bestfitting](#) 2 months ago

Этому где-то учат?

<https://ru.coursera.org/learn/competitive-data-science>

Что порешать?

<http://mltrainings.ru>

Где обсуждение?

<http://ods.ai>