GoodLine
Оператор связи Кузбасса

Атвинта

Machine learning from scratch: myth or reality?

URL: http://goo.gl/V7mvD1

Dmitry Kozlov
Kemerovo

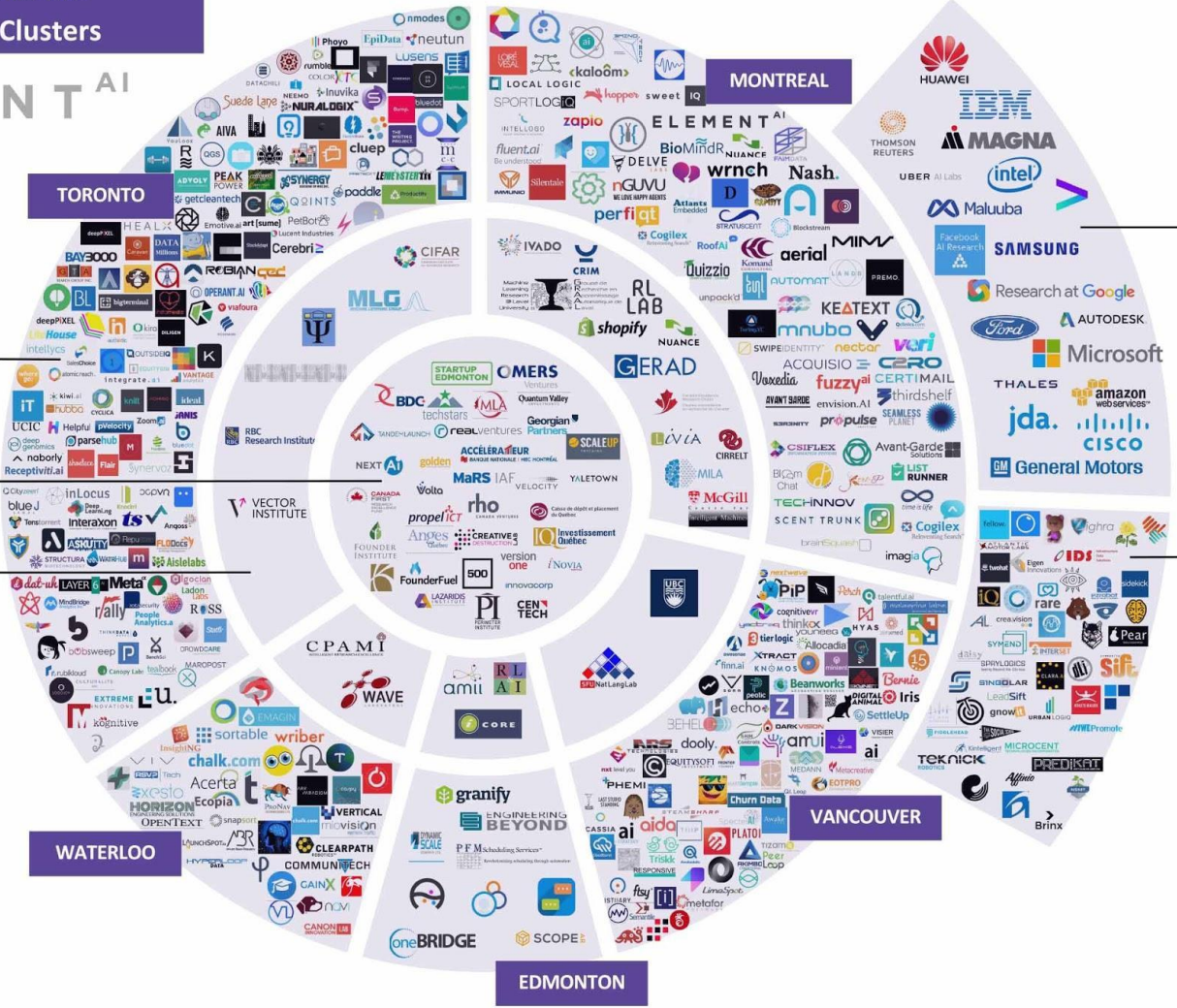
January 25, 2018

Data is the new Oil
We need to find it, extract it, refine it, distribute it and monetize it.

The world’s most valuable resource is no longer oil, but data

Top Players in the Canadian AI Clusters

ELEMENT AI

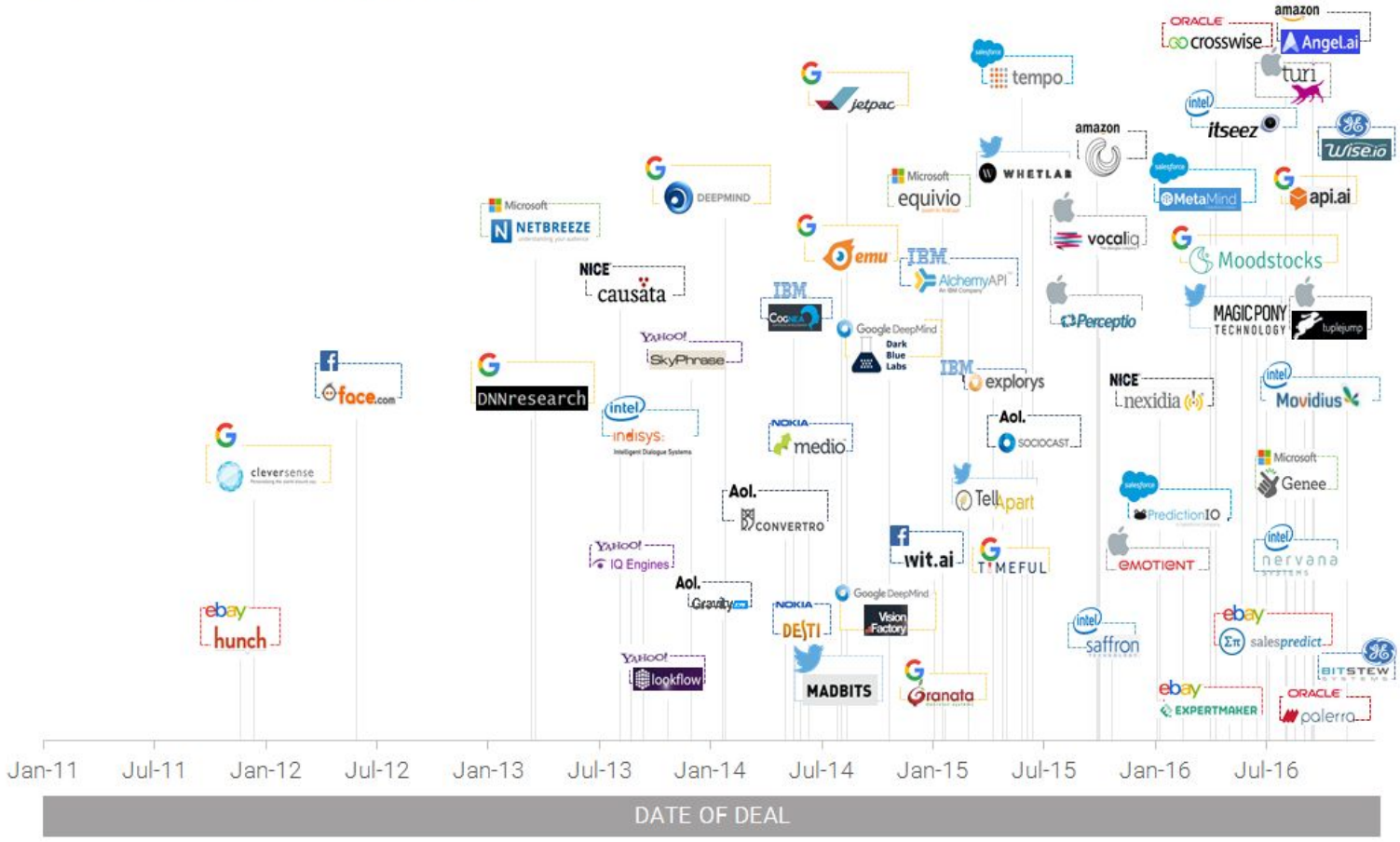


STARTUP & ENTERPRISE COUNT PER LOCATION	
195+	TORONTO
100+	VANCOUVER
90+	MONTREAL
50+	WATERLOO-KITCHENER
10+	EDMONTON
60+	ALL OTHERS

2017-06-11
For updates see jfgagne.ai

RACE FOR AI: MAJOR ACQUIRERS IN ARTIFICIAL INTELLIGENCE

2011 - 2016 YTD (12/1/16)



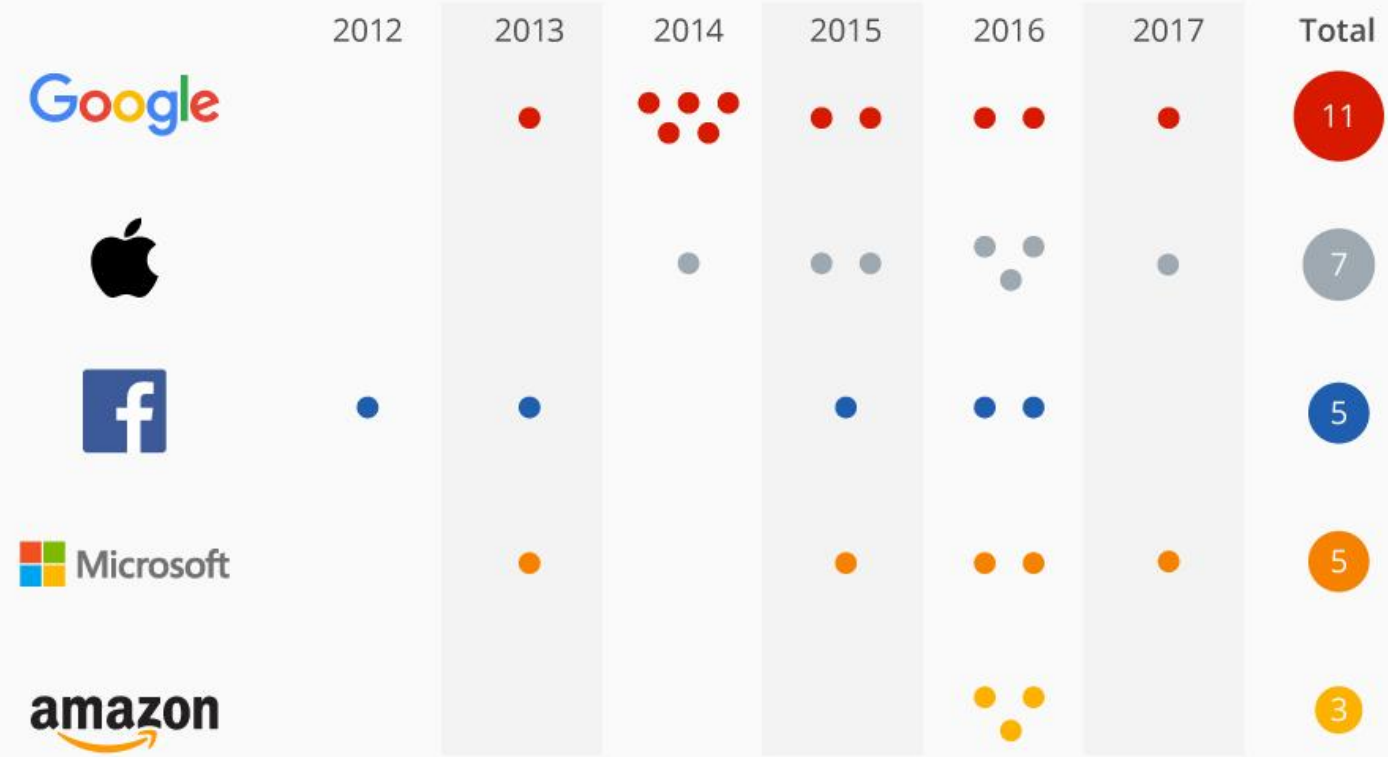
Artificial Intelligence: Most Active Corporate Investors

2011-2016YTD (as of 6/15/2016)

Investor	Rank	Select Investments
Intel Capital	1	DataRobot, prelert, lumiata, MAANA, incoming, saffron, PERFANT, EMOTIENT, Reflektion, Parallel Machines, MindMeld, smartzip, api.ai, indisys, COLDLIGHT
Google Ventures	2	BUILDING ROBOTICS, clarifai, KENSHO, FRAMED, ZEPHYR HEALTH, Unbabel, tamr, MindMeld, Orbital Insight, Predilytics
GE Ventures	3	SIGHT MACHINE, ARTERYS, AYASDI, BITSTEW SYSTEMS, MedAware, PingThings, stem, Predixion, MAANA
Samsung Ventures	4	vicarious, sentiance, Maluuba, IDIBON, jibo, MindMeld, AUTOMATED INSIGHTS
Bloomberg Beta	4	deep genomics, context relevant, AVISO, howdy., DOMINO, Orbital Insight, DigitalGenius, DIFFBOT
In-Q-Tel	6	MindMeld, Digital Reasoning, CYLANCE, DOMINO, celect, INTERSET
Tencent	7	DIFFBOT, CLOUDMEDX, SCALED INFERENCE, iCarbonX, skymind
Nokia Growth Partners	8	rocketfuel, WorkFusion, rapidminer, indix.
Microsoft Ventures	8	BUILDING ROBOTICS, NEURA, insidesales, CrowdFlower
Qualcomm Ventures	8	clarifai, Predilytics, Welltok, tempo
Salesforce Ventures	8	DigitalGenius, insidesales, 6sense
AXA Strategic Ventures	8	NEURA, BLUEBEATS, medlanes, pricemethod
New York Life Insurance Company	8	context relevant, DataRobot, Skycure, captricity

Google Leads the Race for AI Domination

Number of Artificial Intelligence startups acquired since 2012 (as of March 24, 2017)



@StatistaCharts

Source: CB Insights

statista

Applications of machine learning in real life

- Fraud Detection
- Customer churn prediction
- Credit scoring
- Image recognition system
- Recommender system
- Anomaly detection
- Network analysis
- Cluster analysis
- Natural Language Processing
- Audio, Speech recognition
- etc.

Vinci
Yandex
Gosu.ai
Rambler
VisionLabs
Avito
Mail.ru
Kaspersky
Fabby
Sberbank
Kuznech
DoubleData
NTechLab
Prisma

Зачем?

- Возможность получить интересную работу и сложные задачи
- Развитие интуиции, собственная оценка событий и фактов
- Общие подходы к решениям задач в различных прикладных областях
- Применение в реальных практических задачах

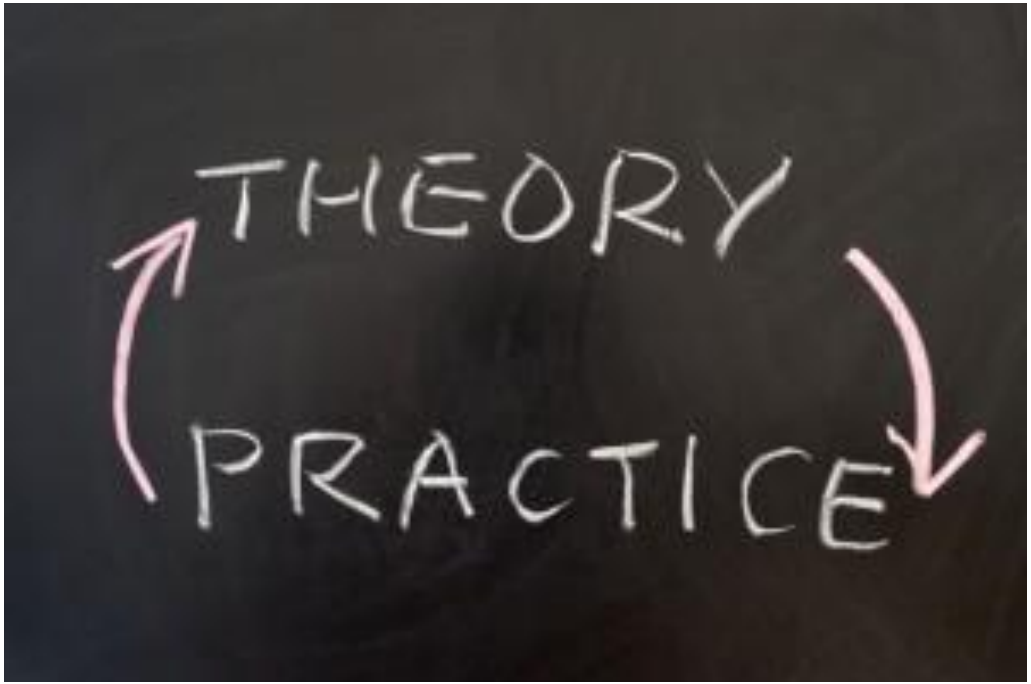
Что важно для старта?

- Мотивация, фокус и желание
- Английский язык (GitHub, arXiv, YouTube, Coursera, Google, etc.)
- Задавать вопросы на английском языке в Google
- Хотя бы один язык программирования (Python, R, C++, C, Java, ~~Matlab~~, etc.)
- ~~Windows~~, ~~macOS~~, Ubuntu
- Поддержка сообщества

Какие бывают данные?

- Табличные данные
- Временные ряды
- Изображения
- Видео
- Текст
- Звук
- Другие...

С чего начать?



С чего начать?

1. **Начать с практики**
   a. Столкнуться с проблемами
   b. Найти решение в теории
   c. Применить решение или вернуться к пункту a)
2. **KISS principle** “Keep it simple, stupid”
3. **Линейные модели** (Linear regression, Logistic Regression, Ridge regression, Lasso, SVM, Naive Bayes, etc.)

Что нужно помнить?

- Время ограничено, в том числе на обучение
- Необходимо декомпозировать сложные задачи
- Проще начать с хорошо изученных областей машинного обучения
- Помнить свою цель обучения, выбирая образовательную траекторию

Какие инструменты?

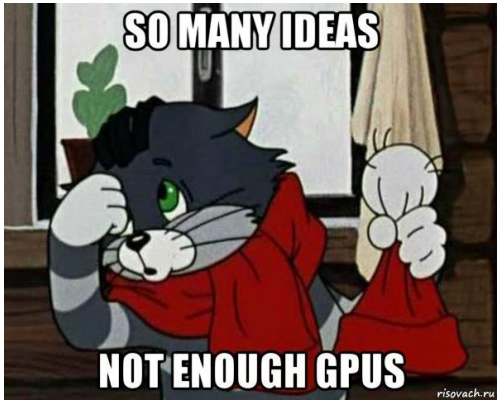
- Искать популярные инструменты на **GitHub**
- Табличные данные (Pandas)
- Линейный модели (Scikit-learn)
- Градиентный бустинг (LightGBM, CatBoost, XGBoost)
- Нейронные сети (Tensorflow, Keras, PyTorch, Caffe, MXNet)
- Оптимизация гиперпараметров (Hyperopt)
- Визуализация (Seaborn, Plotly, Bokeh, Matplotlib)

Какие ресурсы нужны?

- Для анализа небольших табличных данных (Pandas, Scikit-learn, XGBoost, LightGBM, etc): Ноутбук / Домашний компьютер с SSD, RAM >= 4-8 GB, CPU >= 2

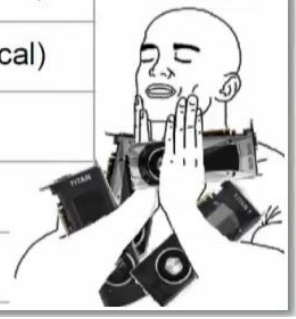
- Для нейронных сетей, анализа текста, изображений и аудио - нужны видеокарты (**GPU**) от Nvidia

Какие ресурсы нужны для DL?



Hardware

n01z3-dl1	i7-6700K, 64Gb	2x TITAN X (Maxwell)
n01z3-dl2	i7-5930K, 64Gb	3x 1080Ti
n01z3-home	i5-6600, 32Gb	1x 1080Ti
dictator-black0	i7-6850K, 128Gb	2x 1080Ti
nizhib-rambler	2xE5-2630 v4, 256Gb	2x Tesla P40
nizhib-2	i7-6700K, 64Gb	2x TITAN X (Pascal)
ternaus-1	i7-4790K, 32Gb	2x TITAN X (Pascal)
ternaus-2	i7-5930K, 64Gb	4x 1080Ti
romul		1x 1080Ti
...	...	...



Какую IDE выбрать?

- Jupyter Notebook
- PyCharm
- Vim
- Любую, с которой вы уже знакомы и хорошо ориентируетесь

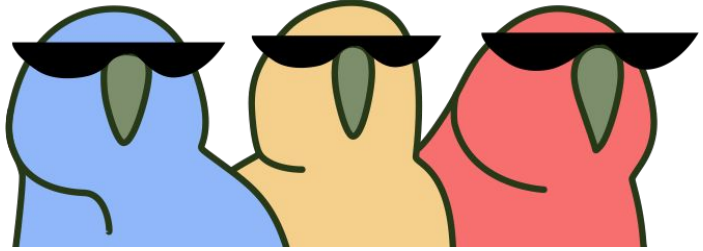
Что делать потом?

- Постоянно учиться и узнавать новое
- Вспоминать лучшие наработки прошлого на практике
- Погружаться в детали и научные статьи, если есть необходимость модификации метода или параметров

Что пригодится?

- Линейная алгебра
- Комбинаторика
- Дискретная математика
- Теория вероятности
- Математический анализ
- Методы оптимизации
- Дифференциальные уравнения
- Структуры данных
- Визуализация данных
- Теория графов, алгоритмы на графах


ПРИСТЕГНИТЕ РЕМНИ-
ВЗЛЕТАЕМ!



Open Data Science

- **Крупнейшее** русскоязычное Data Science сообщество, существует с 2015 года
- Количество участников на данный момент: **10014**
- https://youtu.be/yPKu2vE4UqM?t=2h45m55s
- Регистрация: http://ods.ai
- Блог на хабре: https://habrahabr.ru/company/ods/

Что нужно знать про ODS?

- История сообщений с **2015 года**! (Поиск по ключевым словам, каналам и авторам в Slack)
- Встречи, конференции, Data Science завтраки, тренировки, соревнования, вакансии, (#meetings, #kaggle_crackers, #deep_learning, #nlp, #proj_*,etc.)
- Есть каналы и информация по всем темам так или иначе связанным с машинным обучением и анализом данных

Что нужно знать про ODS?

- Обязательно стоит задавать вопросы в соответствующих тематических каналах (правильный вопрос - это больше половины ответа)
-  Будьте осторожны, ODS затягивает 

Что нужно знать про ODS?

- Ежегодный Data Fest http://datafest.ru/
- Большое количество специалистов из лучших IT-компаний России всегда готовы ответить на Ваши вопросы и **бесплатно**
- Несколько запусков бесплатного массового курса по машинному обучению ML Course ODS (участники сообщества делятся опытом с начинающими)

Что нужно знать про ODS?

- Канал #welcome и #career - здесь вы можете узнать биографию и карьеру многих участников ODS
- #edu_books, #edu_coursees
- Тренировки по машинному обучению #mltrainings_beginners

#_meetings_siberia in ODS

- Сибирская ячейка ODS, каналы: #_meetings_siberia, #_meetings_tomsk (Новосибирск (ЦФТ, 2ГИС, etc), Томск, Барнаул давно и активно встречаются, устраивают совместные завтраки, митапы и конференции)
- Календарь в Новосибирске https://goo.gl/RrSAa4
- Meetup ODSS CFT 16.12.17

#_meetings_siberia in ODS

ivankomarov 6:42 PM

Было бы клёво! Мы хотим всех подтянуть сибиряков...

1

Тренировки по машинному обучению в Yandex

- Анонс новых тренировок: https://events.yandex.ru/events/mltr
- Видео с прошедших тренировок: https://www.youtube.com/channel/UCeq6ZIlvC9SVsfhfKnSvM9w)
- Календарь соревнований: http://mltrainings.ru/

- Платформа для соревнований по машинному обучению мирового уровня с обсуждением задач и общим рейтингом участников

как решать kaggle™?

- Решать вместе
- Быстрые проверки гипотез, больше экспериментов
- Фокус на целевой метрике
- Учиться на сложных примерах
- Расширять кругозор
- Автоматизировать повторяющиеся операции
- Собирать коллекцию трюков

Полезные ссылки

- Тренировки по машинному обучению
- Видео с тренировок по машинному обучению
- https://www.coursera.org/learn/competitive-data-science

Полезные ссылки

- Machine Learning
  https://www.coursera.org/specializations/aml
- Reinforcement learning (#reinfocement_learnin ODS):
  https://www.youtube.com/watch?v=PtAIh9KSnjo
  https://www.coursera.org/learn/practical-rl
  https://www.edx.org/course/reinforcement-learning-explained-microsoft-dat257x
  http://rll.berkeley.edu/deeprlcourse/
  https://www.youtube.com/watch?v=2pWv7GOvuf0

Полезные ссылки

- Natural Language Processing (#nlp in ODS):
  http://web.stanford.edu/class/cs224n/
  https://www.youtube.com/watch?v=OQQ-W_63UgQ
  https://www.coursera.org/learn/language-processing
  http://deephack.me/
- Self-driving cars (#self_driving in ODS):
  https://www.udacity.com/courses/self-driving-car
  https://selfdrivingcars.mit.edu/

Полезные ссылки

- Deep Learning (#deep_learning in ODS):
  http://vision.stanford.edu/teaching/cs231n/
  https://www.coursera.org/specializations/deep-learning
  https://www.youtube.com/playlist?list=PLC1qU-LWwrF64f4QKQT-Vg5Wr4qEE1Zxk
  https://www.youtube.com/watch?v=Am82yvUSwRE
  http://vision.stanford.edu/teaching/cs131_fall1718/
  https://www.youtube.com/watch?v=p5SjqD7Ut4Y&list=PLbwKcm5vdiSYL_yEwQ6JIICBA4dMtHNxo

Полезные ссылки

- Big Data (#big_data in ODS)
  http://mattturck.com/wp-content/uploads/2017/05/Matt-Turck-FirstMark-2017-Big-Data-Landscape.png

  https://www.coursera.org/learn/big-data-essentials

  https://www.coursera.org/courses?languages=en&query=Yandex

Полезные ссылки

- Разбор лучших решений Kaggle:
  http://ndres.me/kaggle-past-solutions/
  https://www.kaggle.com/wiki/PastSolutions
  http://www.chioka.in/kaggle-competition-solutions/
- Блог Александра Дьяконова
- Беседы с гуру Data Science
- https://github.com/rushter/data-science-blogs

Полезные ссылки

- Крупнейшие научные конференции: NIPS, ICML, CVPR, ICCV, KDD
- Видео: NIPS, ICML, CVPR+ICCV, KDD

Школы анализа данных: Yandex, Mail.ru

- https://yandexdataschool.ru/
- https://sphere.mail.ru

Вопросы?

Open Data Science

dmitry.f.kozlov@gmail.com
Telegram: @dfkozlov