Data Science: Roles and Skills

Grozin Vladislav, Constructor.io

About me

6+ years in Data Science; now work for Constructor.io

Mentor @ Open Data Science

Developed personalized recommender systems and search engines:

СИТИЛИНК
ЭЛЕКТРОННЫЙ ДИСКАУНТЕР

Ам
ВИНМАРКЕТ

simplewine

JOOM

faberlic

askona
ТЕРРИТОРИЯ ЗДОРОВОГО СНА

Кораблик
дарим улыбку детям

РИВ ГОШ
парфюмерия и косметика

ULMART.RU

Sulpak

AUTORUS
RACING WITH US

castorama

& many others

Plan

We will talk about:

1. What is Data Science, exactly?

2. Which roles exist in Data Science? How do they interact with each other within a company?

3. Which skills do I need? How can I improve them?

Data Science? What’s that?

Data Science

- “art of turning data into action”
- “field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data”
- “field that refers to the … processes and technologies that enable the … extraction of valuable knowledge and information from raw data”

In other words, science about how to:

- systematically extract value from data
- and build data-based products.

Sample practical applications of DS

- Personalized recommendations at online shops
- Mobile application that applies “clever” filters to images
- Page ranking at search engine
- Next track suggestions at online radio
- Fault prediction software at datacenters

Who Is Data Scientist

Umbrella term, may mean anything

- “Person who works with data and makes data-based decision” (duh)
- “Person employed to analyze and interpret complex digital data, such as the usage statistics of a website, especially in order to assist a business in its decision-making.
- “Person who is better at statistics than any programmer and better at programming than any statistician”

Data Science roles

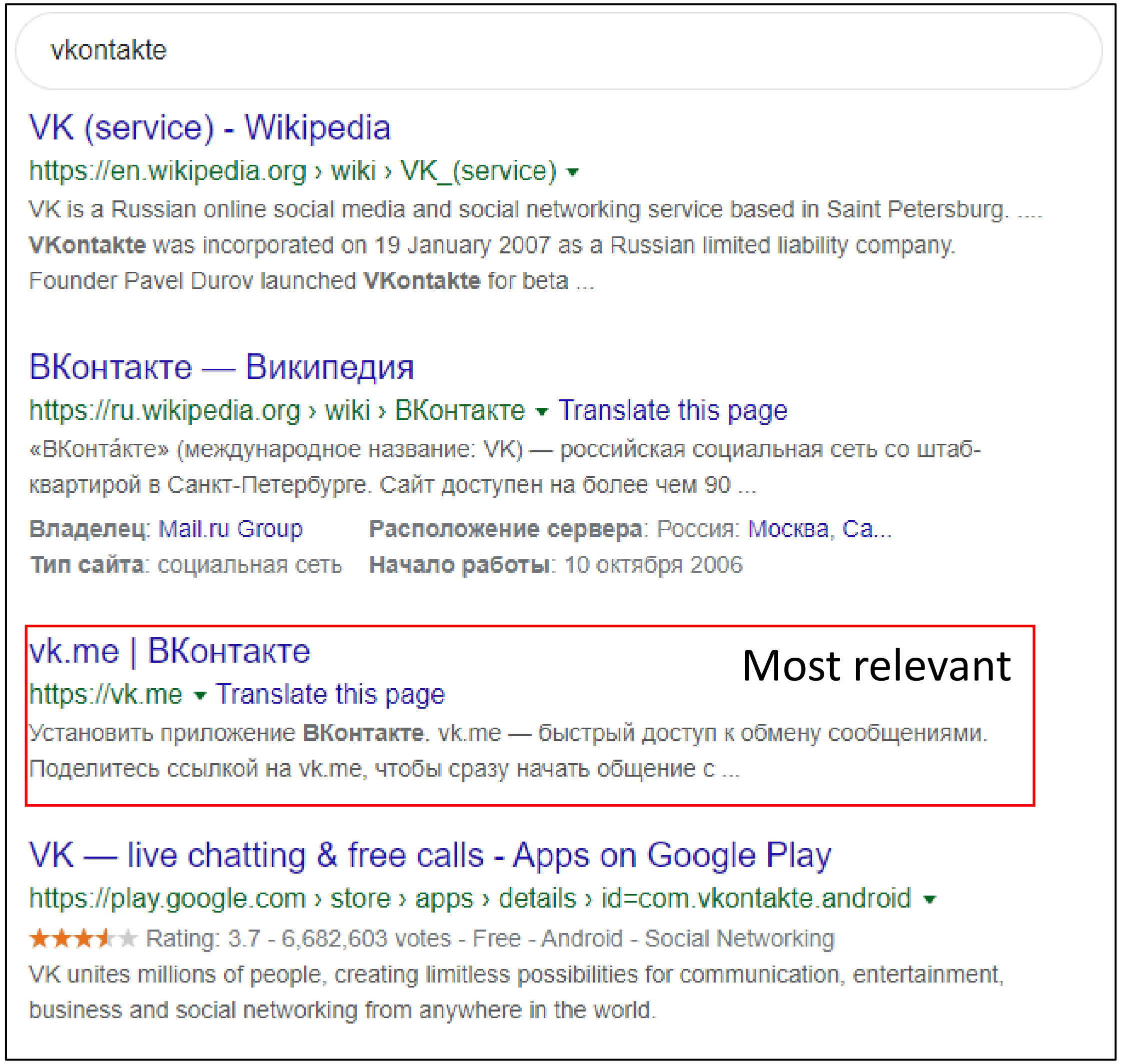
Case Study

We are improving our page search engine, Moogle

Now it works by matching words (i.e. ranks pages by number of cooccurring words)

We think that it does not work well :(

How can we leverage user data to improve it?



vkontakte

VK (service) - Wikipedia
https://en.wikipedia.org › wiki › VK_(service) ▾
VK is a Russian online social media and social networking service based in Saint Petersburg. ....
VKontakte was incorporated on 19 January 2007 as a Russian limited liability company.
Founder Pavel Durov launched **VKontakte** for beta ...

ВКонтакте — Википедия
https://ru.wikipedia.org › wiki › ВКонтакте ▾ Translate this page
«ВКонта́кте» (международное название: VK) — российская социальная сеть со штаб-квартирой в Санкт-Петербурге. Сайт доступен на более чем 90 ...
Владелец: Mail.ru Group **Расположение сервера**: Россия: Москва, Са...
Тип сайта: социальная сеть **Начало работы**: 10 октября 2006

vk.me | ВКонтакте
https://vk.me ▾ Translate this page
Установить приложение **ВКонтакте**. vk.me — быстрый доступ к обмену сообщениями.
Поделитесь ссылкой на vk.me, чтобы сразу начать общение с ...

Most relevant

VK — live chatting & free calls - Apps on Google Play
https://play.google.com › store › apps › details › id=com.vkontakte.android ▾
★★★★☆ Rating: 3.7 - 6,682,603 votes - Free - Android - Social Networking
VK unites millions of people, creating limitless possibilities for communication, entertainment, business and social networking from anywhere in the world.

1. Where is the data? What do we have?

We can ask our **Data Engineers**

1. Where is the data? What do we have?

We can ask our **Data Engineers**

- “The data is stored in Google Cloud”
- “We have user queries and their clicks in table format, but we also repartition it into per-user view”
- “Clicks are tracked using redirections URLs”

Data is the key component

1. Where is the data? What do we have?

We can ask our **Data Engineers**

- “The data is stored in Google Cloud” <- Data Storage
- “We have user queries and their clicks in table format, but we also repartition it into per-user view” <- General transformations (ETL)
- “Clicks are tracked using redirections URLs” <- Data Understanding / Data Stewardship

Data is the key component

2. Why our users are unhappy?

Let’s ask **Data Analysts**

2. Why our users are unhappy?

Let’s ask **Data Analysts**

- “We have looked into available data. Most users type short navigational queries”
- “Data shows that current ranking with #of cooccurring words do not handle it well”
- “Instead, we might want to put most clicked page from that query to the top of the list (i.e. rank by probability of click)”

2. Why our users are unhappy?

Let’s ask **Data Analysts**

- “We have looked into available data. Most users type short navigational queries” <- Insights
- “Data shows that current ranking with #of cooccurring words do not handle it well” <- BI
- “Instead, we might want to put most clicked page from that query to the top of the list (i.e. rank by probability of click)” <- ML objectives

3. Let’s improve lives of our users

ML Engineers, can you do something?

3. Let’s improve lives of our users

ML Engineers, can you do something?

- “We transformed and repartitioned data into format suitable for model training”
- “We have read papers, trained models that calculate click probability, and evaluated it (offline).”
- “We deployed model as a microservice and scheduled its daily updates. Now our users should be happy”

3. Let’s improve lives of our users

ML Engineers, can you do something?

- “We transformed and repartitioned data into format suitable for model training”

  <- Data Views / Application-specific ETL

- “We have read papers, trained models that calculate click probability, and evaluated it (offline).”

  <- Modelling

- “We deployed model as a microservice and scheduled its daily updates. Now our users should be happy”

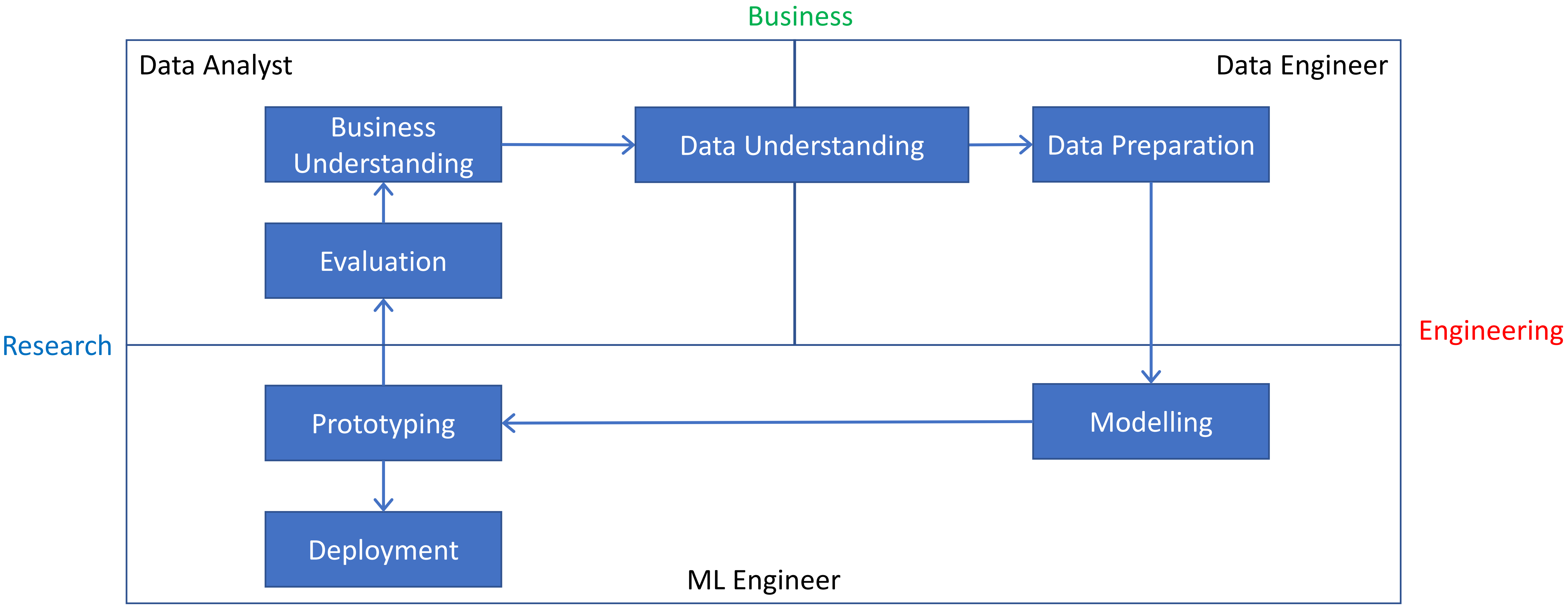
  <- Model deployment

4, 5, 6, ...

- Analysts check how new model affected our users
- Analysts refine our business / ML goal
- Data engineers collect and prepare new data (if needed)
- ML Engineers improve their models
- Analysts check how new model affected our users
- ... and so on

Data Science project cycle

(modified CRISP DM)



Data Engineer

aka Data Architect, Data Steward*, Database Engineer

“Data caretaker”, responsible for data availability and semantics, and transformation pipelines.

Requirements:

- Data transformation (ETL) tools, SQL
- Programming (Scala/Python), storage knowledge
- Domain knowledge

Engineering + Business understanding

*Data Steward is responsible only for data semantics

Data Analyst

aka (Product) Data Scientist, BI Specialist

“Data detective”, responsible for visualization, hypothesis checking, and transforming business objectives into ML goals.

Requirements:

- Statistics, visualization tools
- SQL, basic programming skills
- Strong business understanding and domain knowledge

Research + Business understanding

Machine Learning Engineer

aka Machine Learning Specialist, Data Mining Engineer/Specialist, Applied Research Scientist

“Data wizard”, responsible for application-specific ETL, model training, offline evaluation and model deployment*

Requirements:

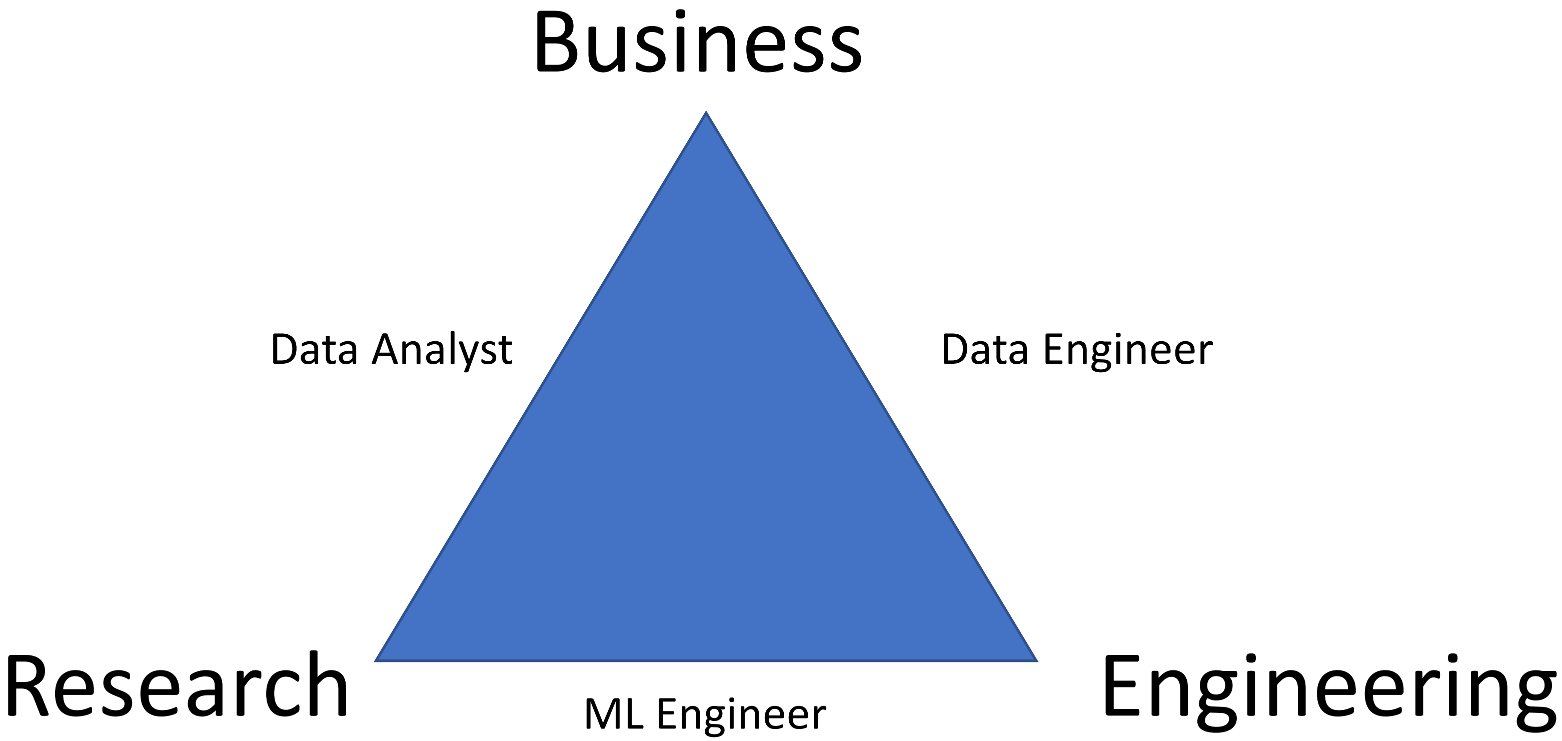
- Machine learning modelling
- Math, basic statistics
- Programming knowledge

Research + Engineering

*In some domains (low-latency, low-memory constraints) Data Engineers are responsible for the deployment

How do I polish my skills?

Basic skills



Business

Data Analyst

Data Engineer

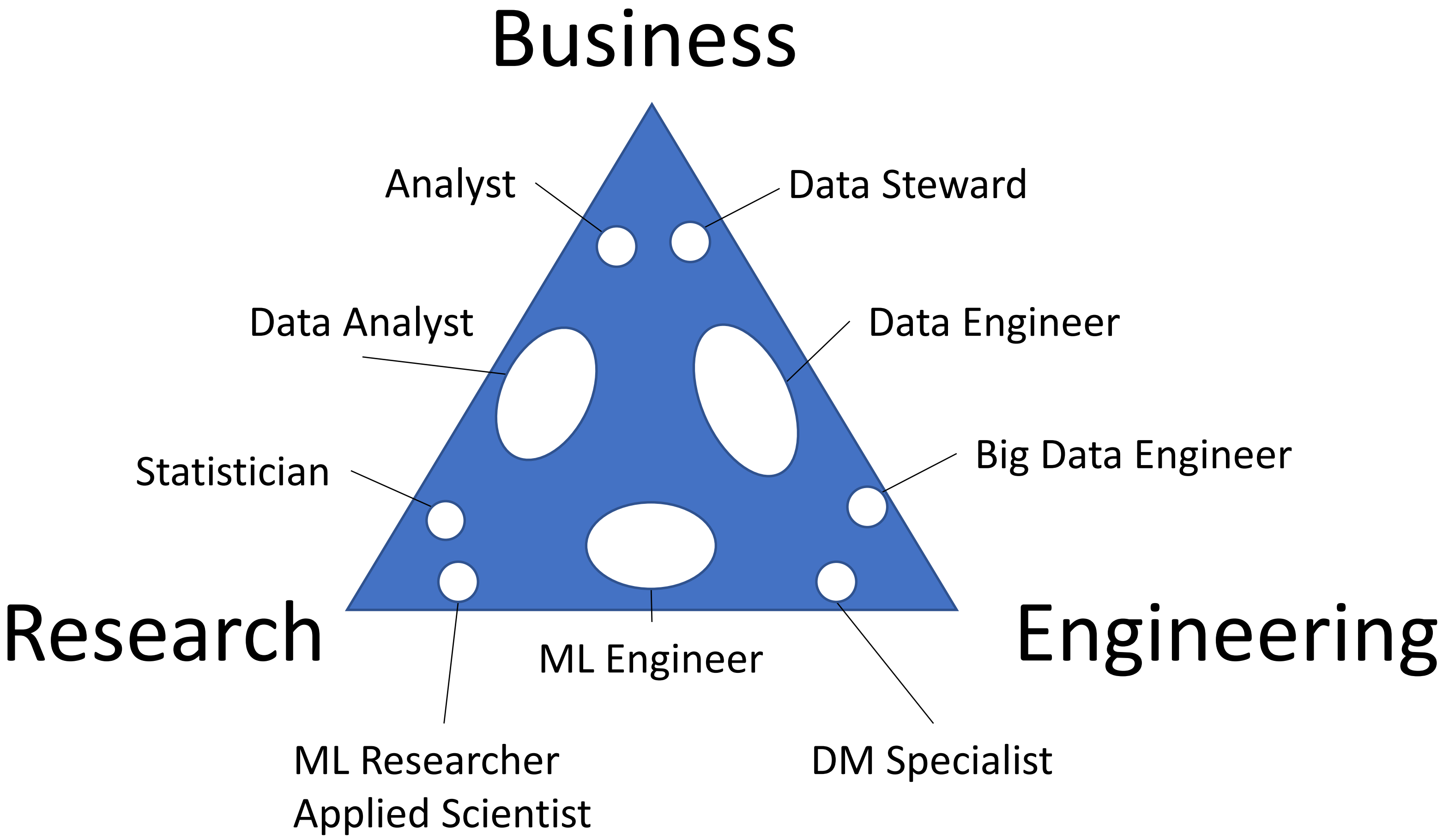
Research

ML Engineer

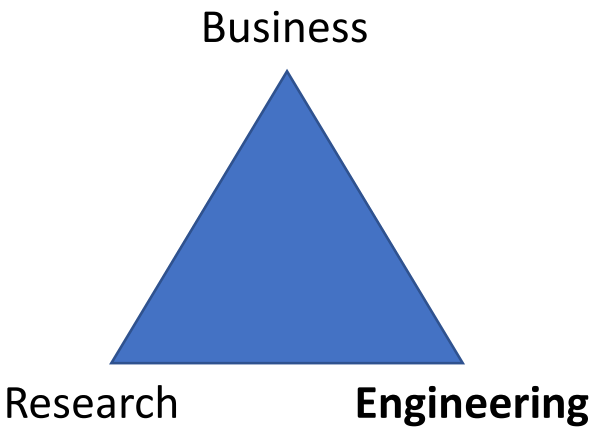
Engineering

Basic skills

Each specific project / task / role in Data Science occupies a region within the triangle



How do I polish my skills?



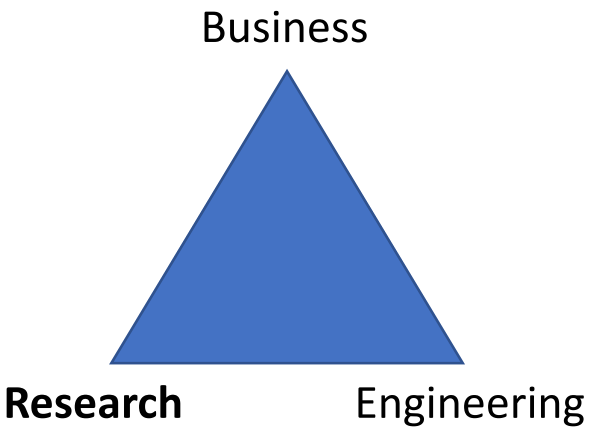
Engineering

Programming
Data storages, databases
Libraries, tools

How to improve:

- Collaborative projects (~open source)
- Competitions (Top Coder)
- Pet Projects

How do I polish my skills?



Research

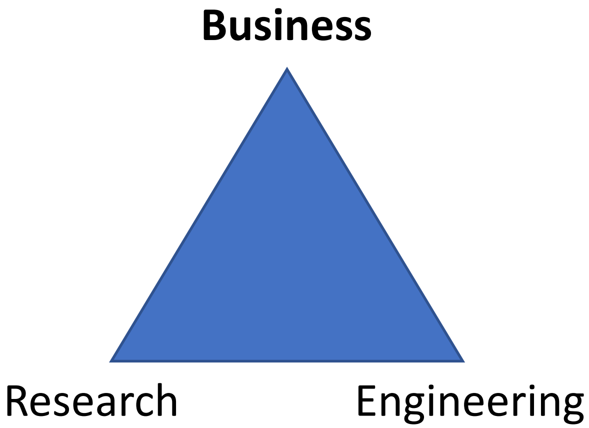
Statistics, probability theory

Machine learning

How to improve:

- ML competitions (Kaggle), Hackathons
- MOOCs
- Read journal articles, implement new methods

How do I polish my skills?



Business

Domain knowledge
Visualization
Data presentation
Analytics

How to improve:

- EDA (@Kaggle), hackathons
- Actual work ;)
- Pet Projects

How to improve my DS skills

Step by step guide

Step by step guide

Step by step guide

1. Pass an online MOOC (Coursera)

Step by step guide

1. Pass an online MOOC (Coursera)
2. Join http://ods.ai/ and its Slack

Step by step guide

1. Pass an online MOOC (Coursera)
2. Join http://ods.ai/ and its Slack

Open Data Science

Largest DS Community in Europe
Regular Conferences & Meetups

Step by step guide

1. Pass an online MOOC (Coursera)
2. Join http://ods.ai/ and its Slack
3. Participate in data-breakfast at #*_meetings_siberia*

Step by step guide

1. Pass an online MOOC (Coursera)
2. Join http://ods.ai/ and its Slack
3. Participate in data-breakfast at *#_meetings_siberia*
4. Join channel *#kaggle_crackers* and find a team to participate in Kaggle

Step by step guide

1. Pass an online MOOC (Coursera)
2. Join http://ods.ai/ and its Slack
3. Participate in data-breakfast at *#_meetings_siberia*
4. Join channel *#kaggle_crackers* and find a team to participate in Kaggle
5. Read papers and publish summaries at *#article_essence*

Step by step guide

1. Pass an online MOOC (Coursera)
2. Join http://ods.ai/ and its Slack
3. Participate in data-breakfast at *#_meetings_siberia*
4. Join channel *#kaggle_crackers* and find a team to participate in Kaggle
5. Read papers and publish summaries at *#article_essence*
6. Choose a problem and build your own solution with *#ods_pet_projects*

Step by step guide

1. Pass an online MOOC (Coursera)
2. Join http://ods.ai/ and its Slack
3. Participate in data-breakfast at *#_meetings_siberia*
4. Join channel *#kaggle_crackers* and find a team to participate in Kaggle
5. Read papers and publish summaries at *#article_essence*
6. Choose a problem and build your own solution with *#ods_pet_projects*
7. *…*

Step by step guide

1. Pass an online MOOC (Coursera)
2. Join http://ods.ai/ and its Slack
3. Participate in data-breakfast at *#_meetings_siberia*
4. Join channel *#kaggle_crackers* and find a team to participate in Kaggle
5. Read papers and publish summaries at *#article_essence*
6. Choose a problem and build your own solution with *#ods_pet_projects*
7. *…*
8. *Profit!*

Thanks!

Vlad Grozin

- vlad@constructor.io
- http://rampeer.github.io/
- ODS.ai : @rampeer

Leave your

feedback