MASTER DEGREE ONLINE

MASTER DEGREE
ONLINE

How to Win Data Analytics Competitions

Artem Volgin
Ekaterina Melianova

Types of Competitions

Machine Learning Competitions

- Goal to predict one variable based on the given data
- Large images, texts, time-series, table datasets
- Machine learning models
- Evaluation of submissions based on some objective metric

Analytics Competitions

- Different type of questions for each competition
- Surveys, administrative statistics, innovative sources of data
- Interpretable models
- Evaluation of submissions by the jury

MASTER DEGREE
ONLINE

Competitions Overview

MASTER DEGREE ONLINE

Kaggle ML & DS Survey 2019

Organizers: Kaggle

Dates: November 8, 2019 – December 3, 2019

Task: Tell a data story about a subset of the data science community represented in this survey, through a combination of both narrative text and data exploration.

Data: Survey of around 20,000 Kaggle users about their work, education, skills etc.

Submission: Story about Kaggle's PhD community using network analysis: network of participants, skills, and countries.

Result: 2nd place – $8,000

kaggle.com/artvolgin/exploring-phd-community-with-network-analysis

MASTER DEGREE ONLINE

DS4G - Environmental Insights Explorer

Organizers: Google

Dates: February 11, 2020 – March 24, 2020

Task: Develop a methodology to calculate an average annual historical emissions factor for the sub-national region.

Data: Remote sensing data about NO2 emissions, weather conditions, additional information from OpenStreetMap.

Submission: Methodology for calculating emissions factor using Spatial Panel Model.

Result: 1st place – $10,000

kaggle.com/katemelianova/ds4g-spatial-panel-data-modeling

MASTER DEGREE ONLINE

March Madness Analytics

Organizers: Google Cloud and National Collegiate Athletic Association

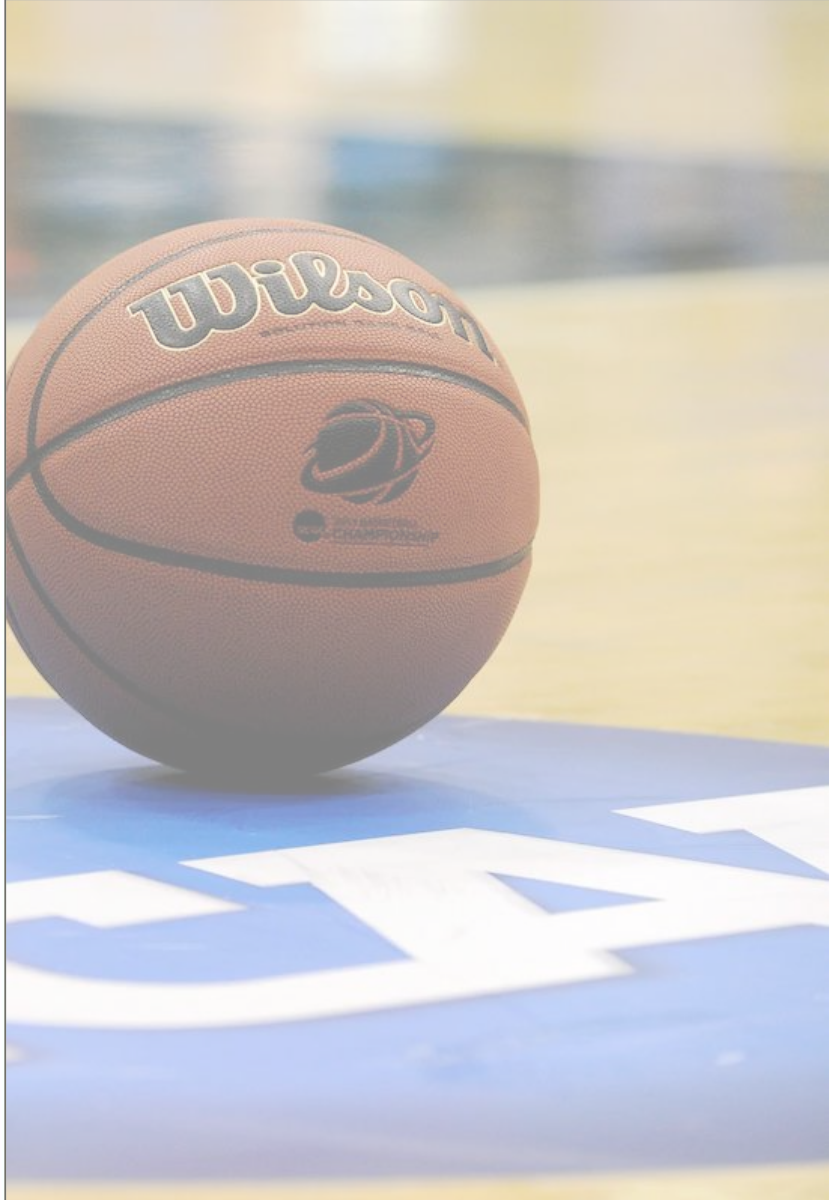
Dates: February 13, 2020 – April 30, 2020

Task: Tell a data story about college basketball through a combination of both narrative text and data exploration.

Data: Games results, Teams and players information, Records of play-by-play events during the games.

Submission: Assist networks between players, networks between teams based on the game results.

Result: No place – $0



kaggle.com/artvolgin/network-approach-to-basketball-analytics

MASTER DEGREE ONLINE

CDP - Unlocking Climate Solutions

Organizers: CDP

Dates: October 14, 2020 – December 2, 2020

Task: Develop a methodology for calculating KPIs that relate to the environmental and social issues. Discuss the intersection between them. Demonstrate whether city and corporate ambitions take these factors into account.

Data: Semi-structured surveys of city officials and corporate managers responsible for the climate change response.

Submission: Data Envelopment Analysis on obtained KPIs. Investigation of relationships between climate hazards, actions an co-benefits using Association Rules Mining. Exploration of cities reports and using Structural Topic Modelling.

Result: 2nd place – 25,000$

kaggle.com/katemelianova/cdp-a-path-to-efficient-and-sustainable-growth

MASTER DEGREE
ONLINE

The COVID-19 Symptom Data Challenge

Organizers: Facebook, Carnegie Mellon University, University of Maryland, Duke Margolis Center for Health Policy

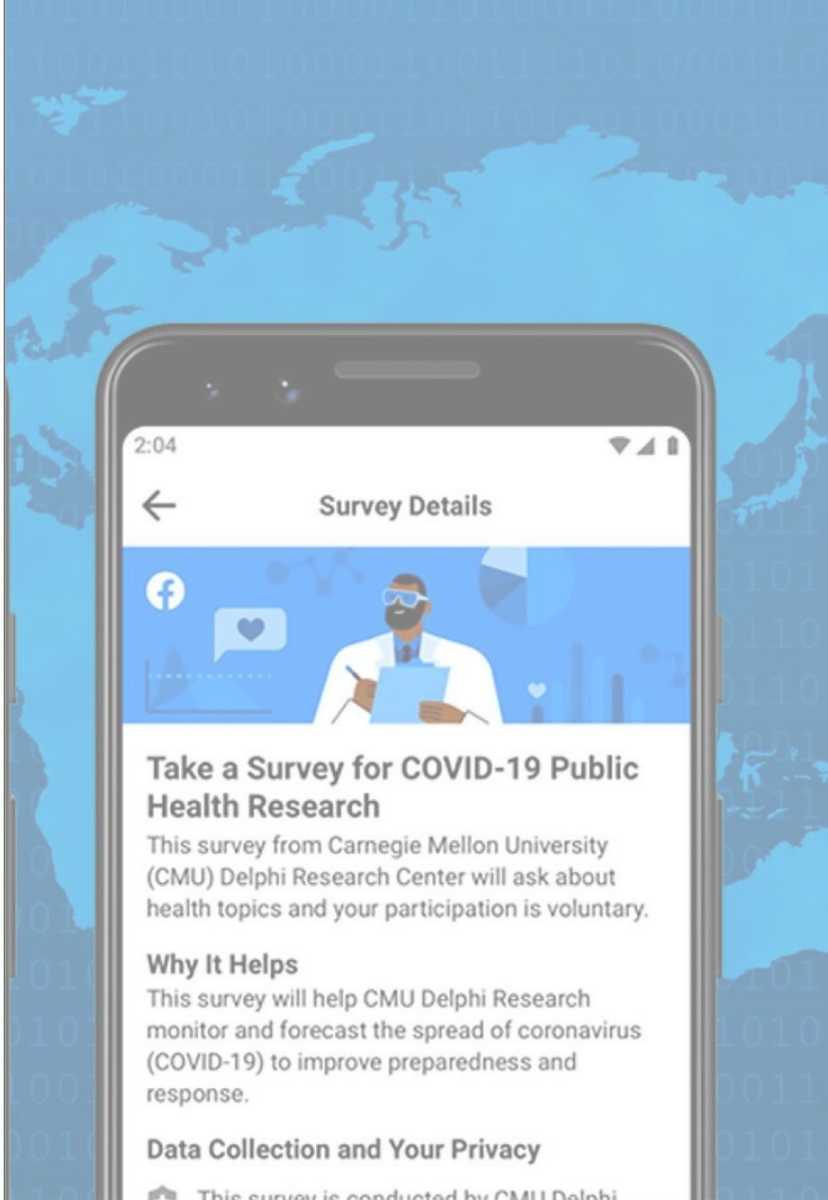
Dates: September 1, 2020 – November 29, 2020

Task: Develop a novel analytic approach to enable earlier detection and improve situational awareness of the outbreak by public health authorities and the general public.

Data: Online survey of more than 10 million Facebook users about COVID symptoms, mask wearing, mental health etc.

Submission: Investigation of causal relationships between COVID-19 cases, people's behaviour, and governmental actions using Multilevel Vector Autoregression model.

Result: 2nd place – $25,000 and $5,000 for semi-final



github.com/artvolgin/symptom-covid-challenge

Towards a Better Submission

Team & Time

Team size
- 1 person is possible (but hard)
- 2 people is optimal
- 3+ people is too much

Required time
- Couple of week minimum
- The more time is better
- Do not be a perfectionist

Different expertise
- Statistical + Substantive
- Visualization + Writing
- Different statistical knowledge

Simultaneity
- EDA + Research on the topic
- Different parts of preprocessing
- Visualization + Writing

Organizers

Why do they organize such competitions?

1. To explore the data
2. To publicly promote their company
3. To hire new people

What do they NOT want to see from you?

1. Showcase of your knowledge in statistics
2. Obvious results (for them)
3. Bla-bla-bla

Organizers

What do they want?

DS4G - Environmental Insights Explorer → **Methodology to calculate the Emissions Factor**

1. Develop a methodology to calculate an average annual historical emissions factor for the sub-national region.
2. Recommendation for how the methodology could be applied to calculate the emissions factor for another area.
3. Additional points for calculating EF on smaller time slices and marginal emissions factors.

CDP - Unlocking Climate Solutions → **KPIs and something else (?)**

1. Develop a methodology for calculating KPIs that relate to the environmental and social issues.
2. Discuss the intersection between environmental issues and social issues.
3. Demonstrate whether city and corporate ambitions take these factors into account.

Competitors

Profile of participants

Most of the Participants

Students, Data engineers, Data scientists

Knowledge in Machine Learning

Usually without domain knowledge

Most of the Winners

Students, Data analysts, Academic researchers

Skills in analytical models, visualization, story-telling

Have domain knowledge in the area

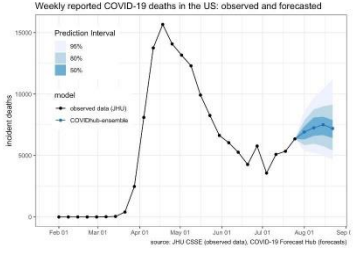
Competitors

Different background

The COVID-19 Symptom Data Challenge
Develop a novel analytic approach to enable earlier detection and improve situational awareness of the outbreak by public health authorities and the general public.

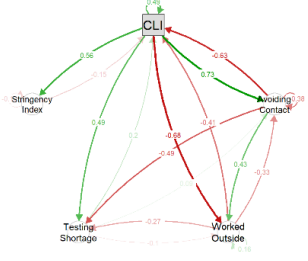
PhD students and professors in Computer Science and Epidemiology from the top US universities who specialized in time-series models and epidemiological forecasting

→ Different Forecasting models of COVID-19



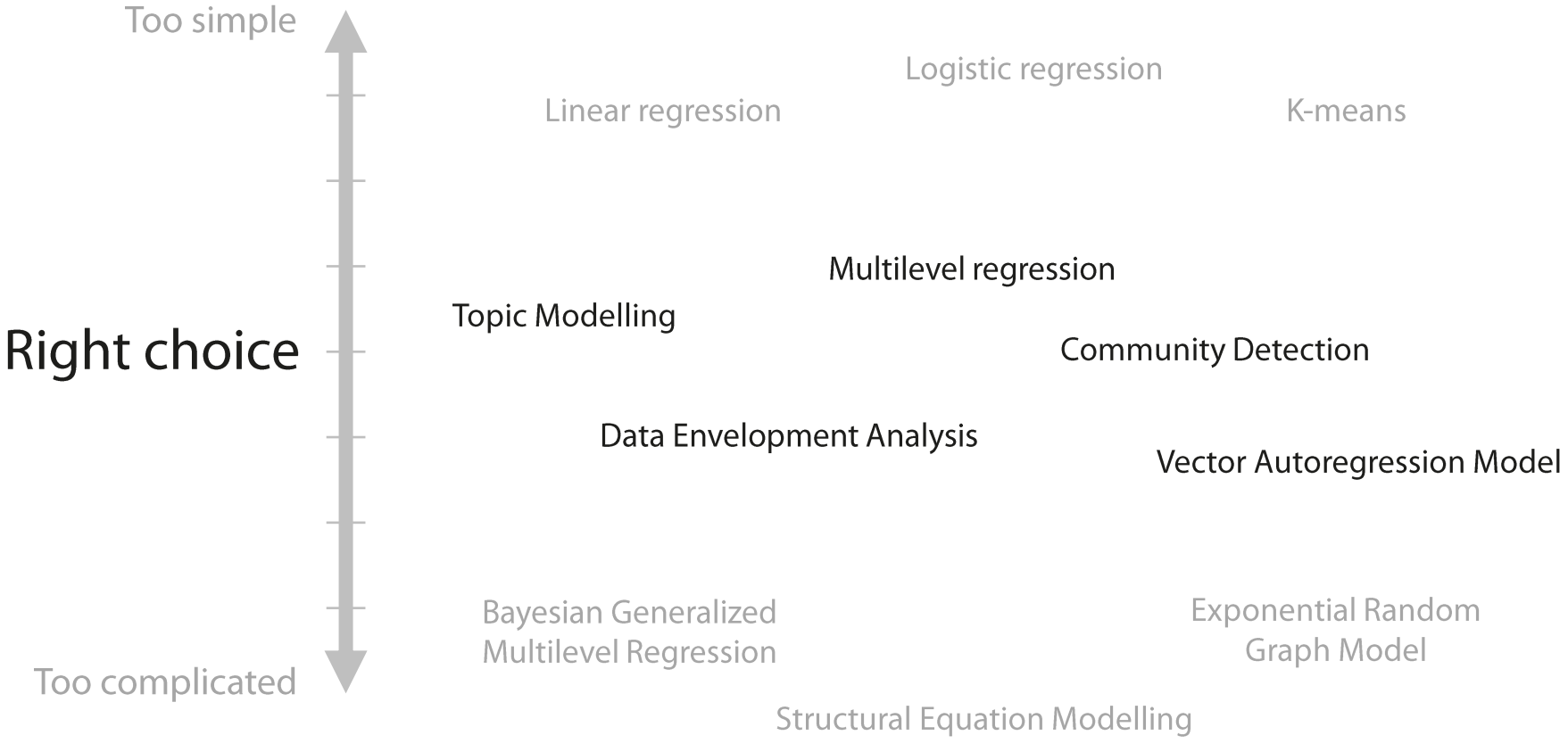
Two random Russian guys who had one course in time-series analysis

→ Model of causal relationships between variables



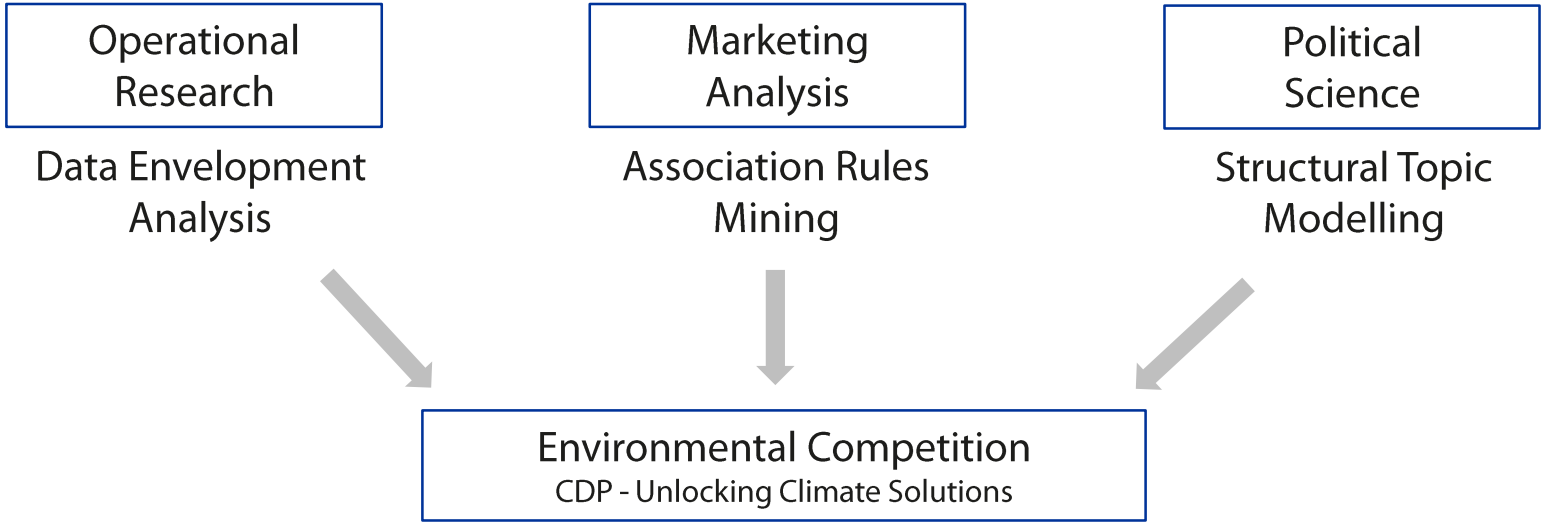
Methods

Complexity



Methods

Multidisciplinary approach



Programming

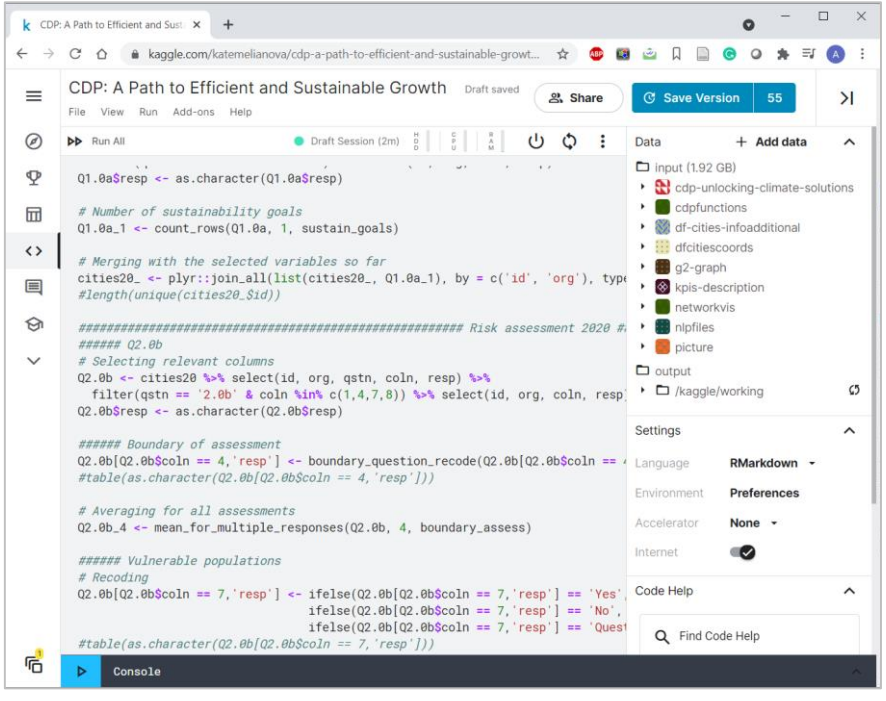
Python or R

1. Textual data
2. Large datasets
3. Complex data structures
4. Machine learning
5. More popular on Kaggle

1. Regression modelling
2. Network analysis
3. Time-series
4. Spatial analysis
5. Niche statistical methods

Programming

Kaggle Notebook



CDP - Unlocking Climate Solutions

2,000+ lines of code…
…and a few hours till the deadline

Version 50	4 months ago	CDP: A Pat...	608.7s	0 B	+1	-1
Version 49	4 months ago	CDP: A Pat...	562.8s	0 B	+2	-1
Version 48	4 months ago	CDP: A Pat...	5.9s	0 B	0	0
Version 47	4 months ago	CDP: A Pat...	642.1s	0 B	+8	-6
Version 46	4 months ago	CDP: A Pat...	588.4s	0 B	+13	-43
Version 45	4 months ago	CDP: A Pat...	105.6s	0 B	+6	-2
Version 44	4 months ago	CDP: Equit...	88.4s	0 B	+54	-49

Keep few days to transfer your code to Kaggle Notebook!

Visualization

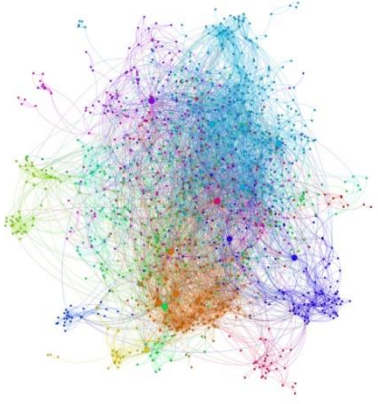
General recommendations

1. Only important graphs
2. One picture – one story
3. Consistent style
4. Interactive
5. No memes

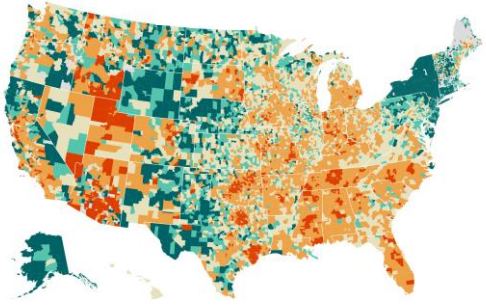
Visualization

Types of visualization to consider

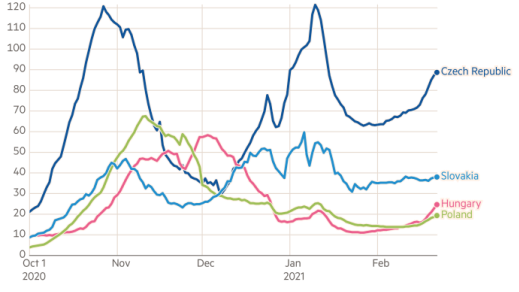
Networks



Maps

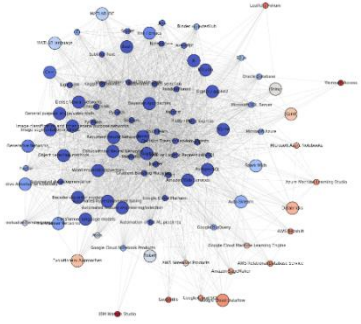


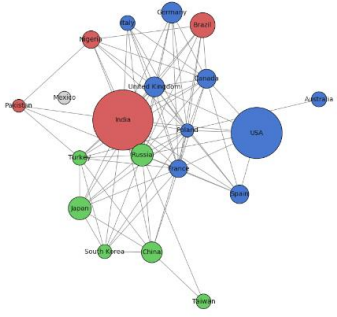
Time-series

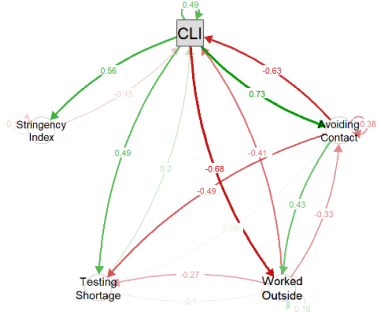


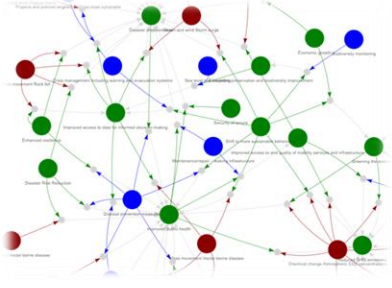
Visualization

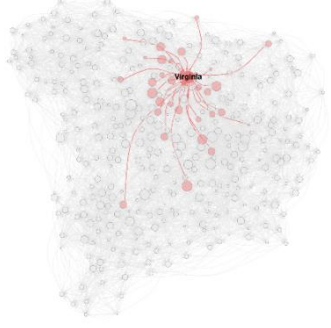
Tunnel vision problem

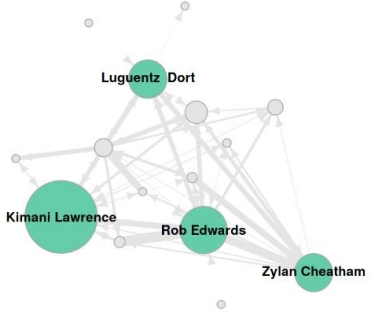












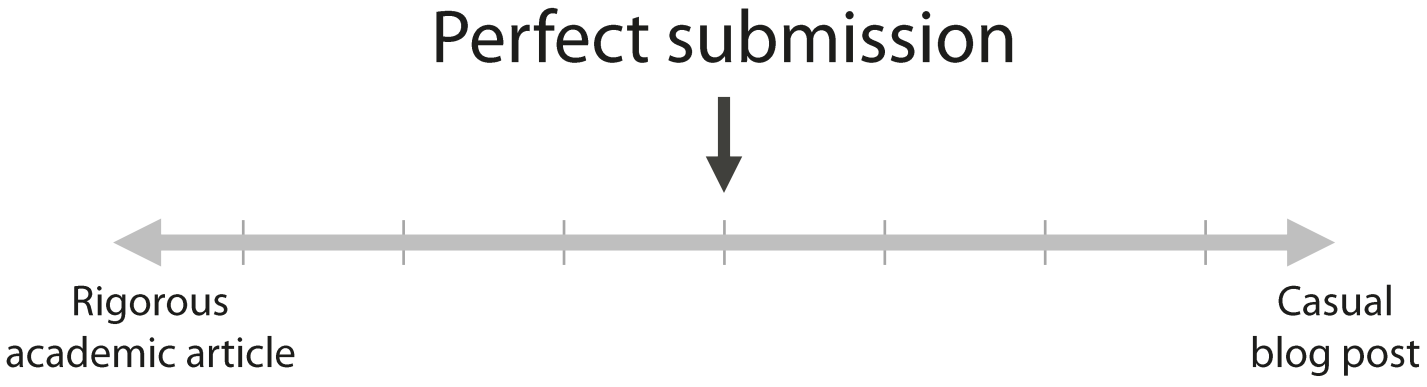
Writing

General recommendations

1. Good language
2. Story-telling
3. Domain vocabulary
4. References
5. Summary

Writing

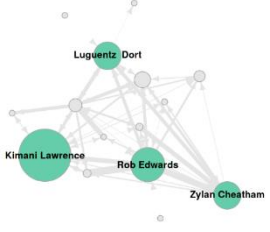
Overall narrative



Writing

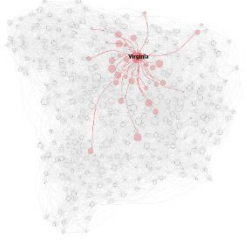
Importance of writing

March Madness Analytics
Your challenge is to tell a data story about college basketball through a combination of both narrative text and data exploration.



We had

- ✓ Different types of Networks
- ✓ Nice visualizations
- ✓ Exponential Random Graph Model

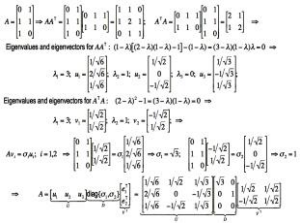


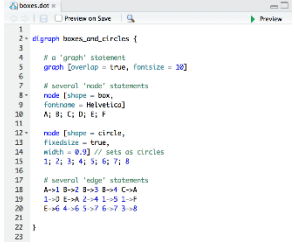
but

- Poor story-telling
- Absence of the domain knowledge

No place for us in the end :(

Your Submission





Methods + Programming + Writing + Visualization

Summary

How to Win?

1. Have a great team and a lot of time
2. Understand what organizers really want
3. Know your competitors
4. Utilize reasonably complicated methods
5. Choose R and handle Kaggle Notebook
6. Fascinate judges with network pictures
7. Practice your writing skills

Should You Participate?

If you participate

1. Methodological experience
2. Knowledge in the substantive area
3. Story-telling skill
4. Consulting experience
5. Coding practice
6. Pet-project

+

If you **win**

7. Line in your CV
8. Fame (for a couple of days)
9. Some money

MASTER DEGREE ONLINE

How to Win Data Analytics Competitions

Artem Volgin
art.volgin@gmail.com

Ekaterina Melianova
melianova95@gmail.com