Linear Regression
Logistic Regression
Jackknife Regression
Density Estimation
Confidence Interval
Test of Hypotheses
Pattern Recognition
Clustering — (Unsupervised Learning)
Supervised Learning
Time Series
Decision Trees
Random Numbers
Monte-Carlo Simulation
Bayesian Statistics
Naive Bayes
Principal Component Analysis — (PCA)
Ensembles
Neural Networks
Support Vector Machine — (SVM)
Nearest Neighbors — (k-NN)
Feature Selection — (Variable Reduction)
Indexation / Cataloguing
(Geo-) Spatial Modeling

РАНХиГС
РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

«Большие данные» и «Наука о данных»

Использование инструментария «больших данных» в образовании

Дождиков Антон Валентинович, к. полит. наук,
директор центра аналитики образовательных данных
ФИРО РАНХиГС, dozhdikov-av@ranepa.ru

Recommendation Engine
Search Engine
Attribution Modeling
Collaborative Filtering
Rule System
Linkage Analysis
Association Rules
Scoring Engine
Segmentation
Predictive Modeling
Graphs
Deep Learning
Game Theory
Imputation
Survival Analysis
Arbitrage
Lift Modeling
Yield Optimization
Cross-Validation
Model Fitting
Relevancy Algorithm
Experimental Design

ЧТО ЭТО ТАКОЕ? ЗАЧЕМ? ПОЧЕМУ? ЧТО БУДЕТ ДАЛЬШЕ?

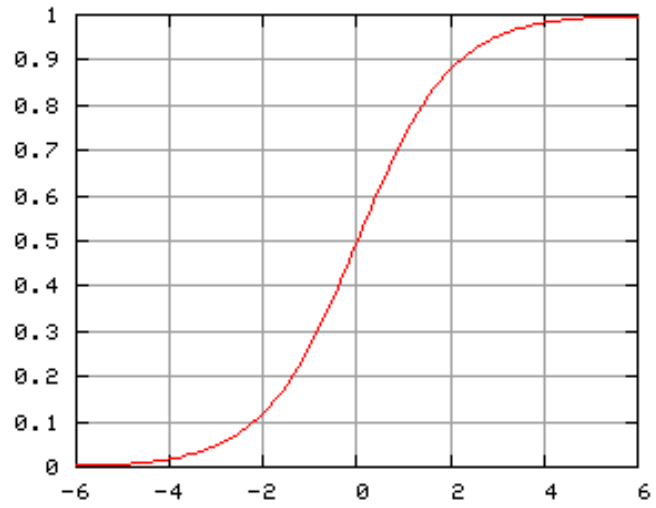
И ЧТО ПО ЭТОМУ ПОВОДУ ДУМАЮТ «КОТИКИ»? ХОТЯТ ЛИ ОНИ УЧИТЬСЯ?

Прогнозирование:

Можно ли предсказать какое-либо событие, основываясь на накопленном объеме информации и текущих данных?

Простой пример : «успех» или «неуспех» фильма в прокате ? Или успеха или провал на ЕГЭ? Признаки «успешности»?

Метод: логистическая регрессия (статистика, машинное обучение).Прогнозирования вероятности возникновения некоторого события (отклик Y) по значениям множества признаков (X1, X2… Xn)



Инструменты - языка Python:

- pandas - программная библиотека на языке Python для обработки и анализа данных
- sklearn (scikit-learn) - библиотека машинного обучения
- matplotlib - библиотека для визуализации данных двумерной и трёхмерной графикой

ДАНО: обучающая выборка (1000 строк) и тестовая выборка (400 строк)

	screens	budget	genre	director	scriptwriter1	scriptwriter2	age	time	box
0	100	46096480.0	0.509727	0.443680	1.171456	1.171456	12	115	0
1	117	57620600.0	0.940596	0.563948	0.742941	0.780971	0	98	0
2	315	121003260.0	0.940596	4.242373	1.662735	1.662735	16	115	1
3	47	46096480.0	0.509727	0.099780	0.773626	0.371797	12	101	0
4	188	57620600.0	1.260718	0.329743	0.329743	0.329743	6	90	0
...	...	...	...	...	...	...	...	...	...
995	266	37000000.0	0.509727	0.594595	0.594595	0.371797	18	98	0
996	1270	60000000.0	0.509727	1.316667	1.316667	0.371797	16	97	0
997	30	80000000.0	1.872024	0.200000	0.273121	0.371797	16	98	0
998	70	80000000.0	0.518263	0.013609	0.273121	0.371797	6	55	0
999	1226	125418200.0	1.872024	0.458113	1.291873	0.371797	18	100	0

1000 rows × 9 columns

	screens	budget	genre	director	scriptwriter1	scriptwriter2	age	time	box
0	818	75000000.0	1.037808	1.399977	1.466667	1.466667	16	97	0
1	195	80000000.0	1.260718	0.200000	0.273121	0.371797	12	75	0
2	56	80000000.0	1.872024	0.200000	1.206058	3.096553	18	94	0
3	1260	140000000.0	1.872024	3.813042	1.357143	1.357143	12	90	0
4	988	81521830.0	1.872024	0.210352	0.269866	0.269866	12	104	0
...	...	...	...	...	...	...	...	...	...
395	112	40000000.0	0.509727	0.020075	0.020075	1.316217	16	88	0
396	108	80000000.0	0.509727	0.200000	3.949617	0.371797	16	94	0
397	1695	120000000.0	1.872024	1.437529	0.566667	0.566667	16	105	0
398	1662	319000000.0	0.509727	0.407003	0.181818	0.371797	12	105	0
399	21	80000000.0	0.843633	0.200000	0.273121	0.371797	0	90	0

400 rows × 9 columns

Представлена информация по количеству экранов, на которых вышел фильм, жанру, режиссеру, двум сценаристам, возрастному рейтингу фильма, его длительности. Колонка «box» говорит о том что «фильм окупился в прокате» -1. Фильм «не окупился в прокате» - 0

Результаты прогнозирования:

```
#подключение библиотеки для вычисления метрик
from sklearn import metrics
```

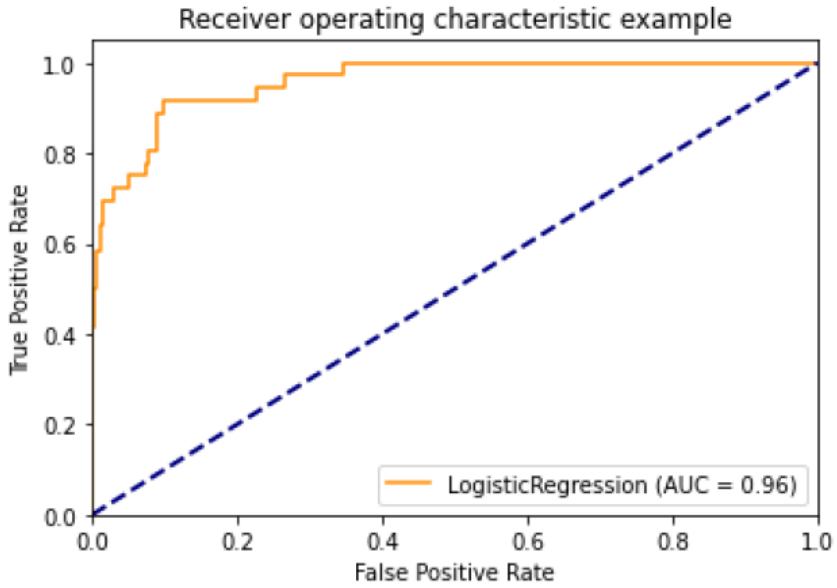
```
print('Точность предсказаний:',metrics.accuracy_score(Y_pr
print('Количество предсказанных верно из 400:', metrics.ac
```

```
Точность предсказаний: 0.9175
Количество предсказанных верно из 400: 367
```

```
#подключим библиотеки для визуализации
import matplotlib.pyplot as plt
%matplotlib inline
metrics.plot_roc_curve(reg, X_test, Y_true, color='darkorange')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic example')
plt.legend(loc="lower right")
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.show()
```

Точность предсказаний в 0.92 – не предел. Необходим более сложный алгоритм, учет не 8 наборов предикторов а примерно 20-23 и полные данные. В текущей выборке 30% данных – неполные, нет данных по бюджету фильма.

А если у нас сотни тысяч и миллионы записей? Полных, с отсутствием погрешностей, нормализованных? Как результаты ЕГЭ?



Для сферы образования:

Данные	Направления использования
Портфолио обучающегося, результаты ОГЭ, оценки по предметам, другие количественные оценочные характеристики	Прогноз результатов ЕГЭ и других экзаменов Обучающийся видит, где у него пробелы в знаниях и навыках, что нужно подтянуть
Выбор обучающихся предметов ЕГЭ и результаты ЕГЭ, данные по образовательной миграции из региона	Подготовка предложений по оптимизации распределения контрольных цифр приема в вузы
Образовательные данные и учебные результаты студентов; данные по миграции; запросы работодателей и базы вакансий	Планировать реализацию сетевых ОП. Подбирать места для практики и готовить специалиста под конкретный набор требований. Гибкость и персонализированность образовательных программ
Данные по использованию обычных учебников Параметрическая информация электронных образовательных ресурсов	Какой учебник дает лучшие результаты по каким направлениям?
Данные по преподавателям и результатам их работы	Объективная и беспристрастная оценка. Без бюрократии и отчетов.
….	….

Немного о терминах:

Data Science - прикладная наука о данных на стыке: математики, статистики и информатики. В качестве основы для анализа можно использовать любые достоверные данные. Если есть статистически важные закономерности, то методы машинного обучения их «вытащат».

1. Большие данные.

data science обрабатывает терабайты разнообразных данных, чья скорость поступления постоянно растет. Это может быть разрозненная, неформализованная, бессистемная информация, которую нужно чистить, обрабатывать и структурировать. Обычные аналитические методы не справляются. Качество данных в data science играет решающую роль. Сбор, обработка и «чистка» данных занимает до 80%, а иногда и более времени. Каторжный труд.

2. Использование математических методов.

Все методы data science пришли из прикладной математики и естественных наук. Методы универсальны: как при исследовании звезд, ядерных реакций, так и при данных о людях (people data). Наличие глубокого знания предметной области ускоряет подбор нужного метода и настройку инструмента. Необходимы: математический анализ, линейная алгебра, теория вероятностей и математическая статистика, программирование на Python

3. Компьютерный анализ данных и машинное обучение.

Компьютеры обрабатывают миллионы строк данных и в автоматическом режиме. На домашнем ПК— не хватит ни памяти, ни мощности оборудования. Только учебные примеры и тысячи строк. Сотни тысяч и миллионы – добро пожаловать на сервер или супекомпьютер.

А если совсем просто? На котиках

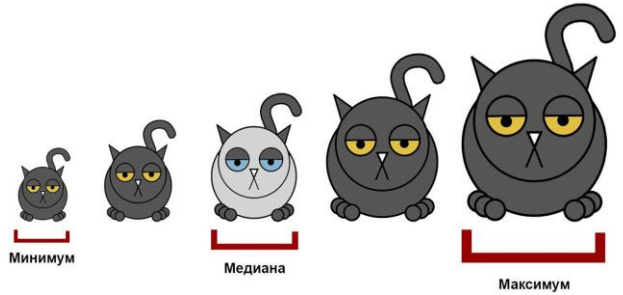
Чем отличается социология от «больших данных»?

Социология и статистика как правило работают с выборочной совокупностью (выборкой) – это часть единиц генеральной совокупности, отобранная специальным методом и предназначенная для характеристики всей генеральной совокупности.

«Котики бывают/котики будут…(признаки)»

Big Data работает сразу со всей «генеральной совокупностью» – это совокупность всех без исключения единиц изучаемого объекта, всех единиц, которые соответствуют цели исследования. Причем эта совокупность может изменяться (пополняться) в ходе исследования.

«Все котики были, есть прямо сейчас и возможно будут…(признаки)»



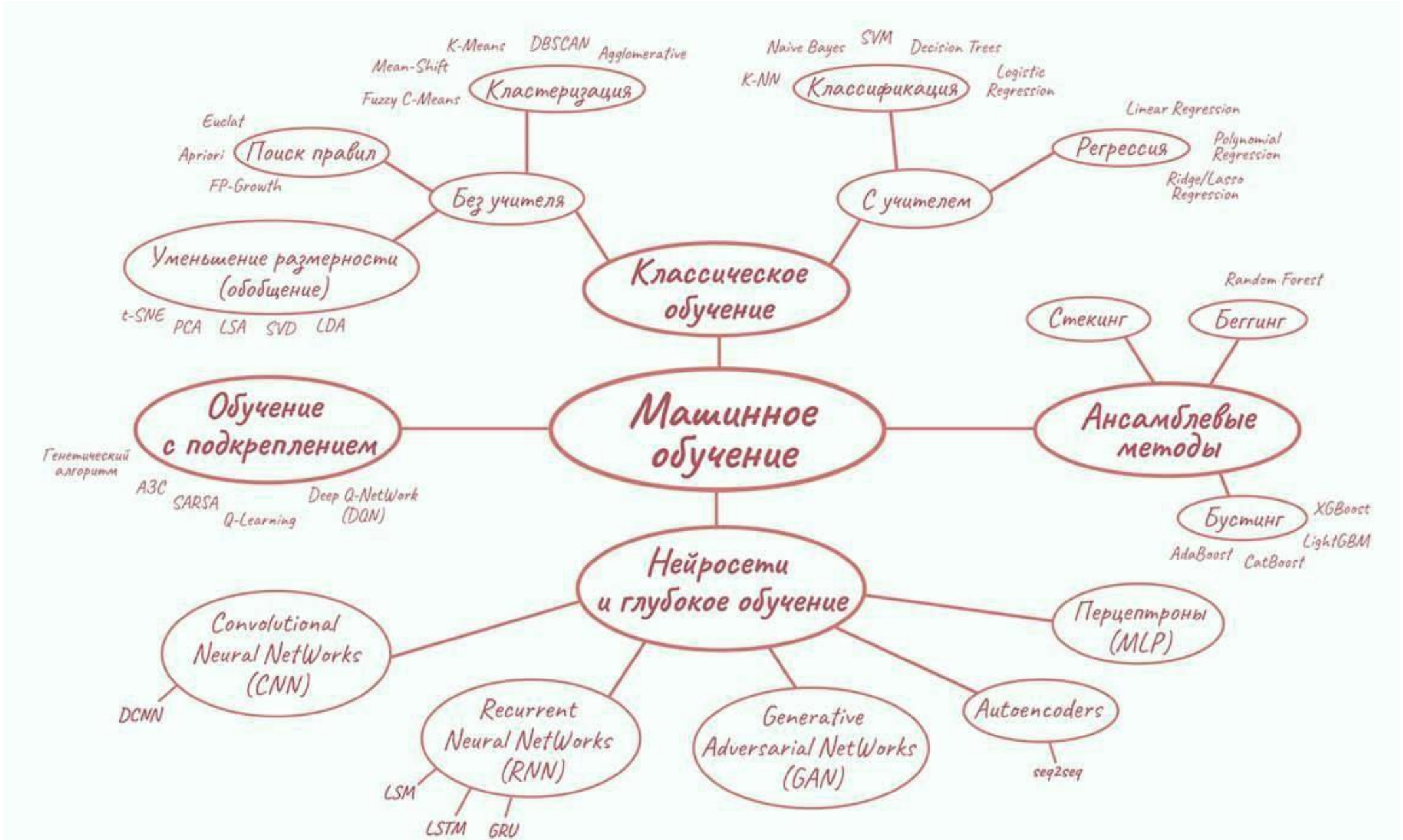
Исследовательский алгоритм Data Science:

1. Описание изучаемых объектов (предметной области)
2. Выявление переменных: ключевых признаков, «предикторов» и «откликов»
3. Формулировка гипотез (предположений)
4. Сбор и подготовка (первичная обработка, «чистка», «нормализация») данных
5. Разведочный анализ данных. Простые аналитические методы. Частичное подтверждение гипотез.
6. Очистка данных от «шумов» и «аномалий».
7. Преобразование данных (подготовка к визуализации)
8. Построение математических моделей, вывод целевых формул, создание нейросети
9. Интерпретация данных. Выводы. Подтверждение, опровержение, корректировка исходных гипотез.

Итоговым результатом работы может компьютерный алгоритм, который принимает новые данные и прогнозирует будущее с определенной точностью.

Методы Big Data

- **Краудсорсинг** — сбор, первичный анализ, обогащение данных силами многих людей.
- **Парсинг** — автоматизированный сбор и систематизация информации из открытых источников в сети с помощью скриптов (веб-скрейпинг).
- **Автоматизированный сбор данных** приложений (телеметрия, логи, история), пользовательского выбора, предпочтений и people data – методы фиксации и сбора данных, генерируемых информационными образовательными системами в процессе их эксплуатации и взаимодействия с пользователями
- **Data Mining** - обучение ассоциативным правилам , классификация (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным), кластерный анализ, регрессионный анализ; технология добычи новой значимой информации из большого объема неструктурированных данных
- **Смешение и интеграция данных** — приведение данных из разных источников к одному виду, уточнение и дополнение данных.
- **Обработка данных -** нормализация, приведение к единому формату
- Проверка данных - на полноту, на наличие выбросов-аномалий
- **Машинное обучение и нейронные сети** — создание программ, которые умеют анализировать и принимать решения, выстраивая логические связи.
- **Предиктивная аналитика** - предсказание будущего на основе собранных и исторических данных
- **Имитационное моделирование** — построение моделей на основе больших данных, которые помогают провести эксперимент в виртуальной реальности
- **Статистический анализ** — подсчет данных по формулам и выявление в них тенденций, сходств и закономерностей, A/B-тестирование и анализ временных рядов
- **Визуализация** — представление больших данных и результатов их анализа в виде удобных графиков и схем (интерактивных дашбордов), понятных человеку для принятия решений
- **Распознавание** образов, текста, аудио, видео («компьютерное зрение»);
- **Пространственный анализ**— класс методов, использующих топологическую, геометрическую и географическую информацию в данных.
- …


Машинное обучение
Классическое обучение
Без учителя
Кластеризация
K-Means
DBSCAN
Agglomerative
Mean-Shift
Fuzzy C-Means
Поиск правил
Euclat
Apriori
FP-Growth
Уменьшение размерности (обобщение)
t-SNE
PCA
LSA
SVD
LDA
С учителем
Классификация
Naive Bayes
SVM
Decision Trees
K-NN
Logistic Regression
Регрессия
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
Обучение с подкреплением
Генетический алгоритм
A3C
SARSA
Q-Learning
Deep Q-NetWork (DQN)
Ансамблевые методы
Стекинг
Беггинг
Random Forest
Бустинг
XGBoost
LightGBM
AdaBoost
CatBoost
Нейросети и глубокое обучение
Convolutional Neural NetWorks (CNN)
DCNN
Recurrent Neural NetWorks (RNN)
LSM
LSTM
GRU
Generative Adversarial NetWorks (GAN)
Autoencoders
seq2seq
Перцептроны (MLP)

Образовательные данные

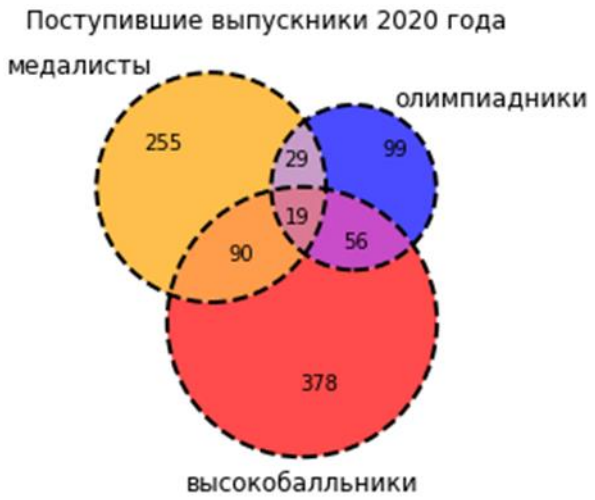
Данные об образовательных результатах обучающихся, развития образовательных систем, условиях образования (образовательных программах, особенностях образовательной среды, образовательных методиках.)

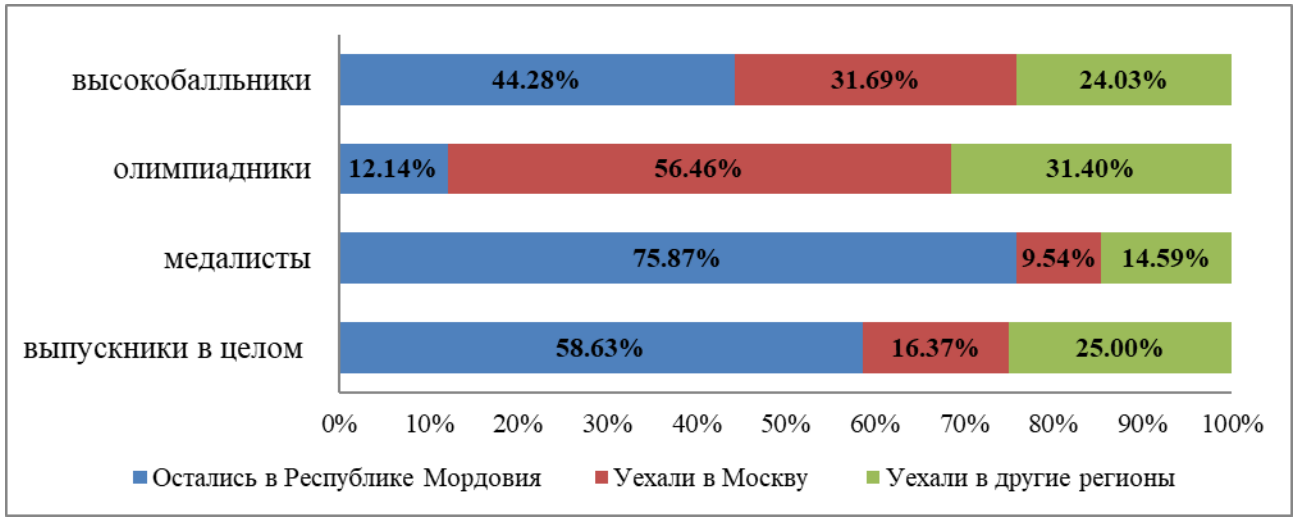
- Прогнозирование будущего поведения учащихся в процессе обучения
- Обнаружение или улучшение моделей предметной области
- Изучение разных эффектов поддержки обучения, экспериментов
- Расширение научных знаний об обучении и его составляющих

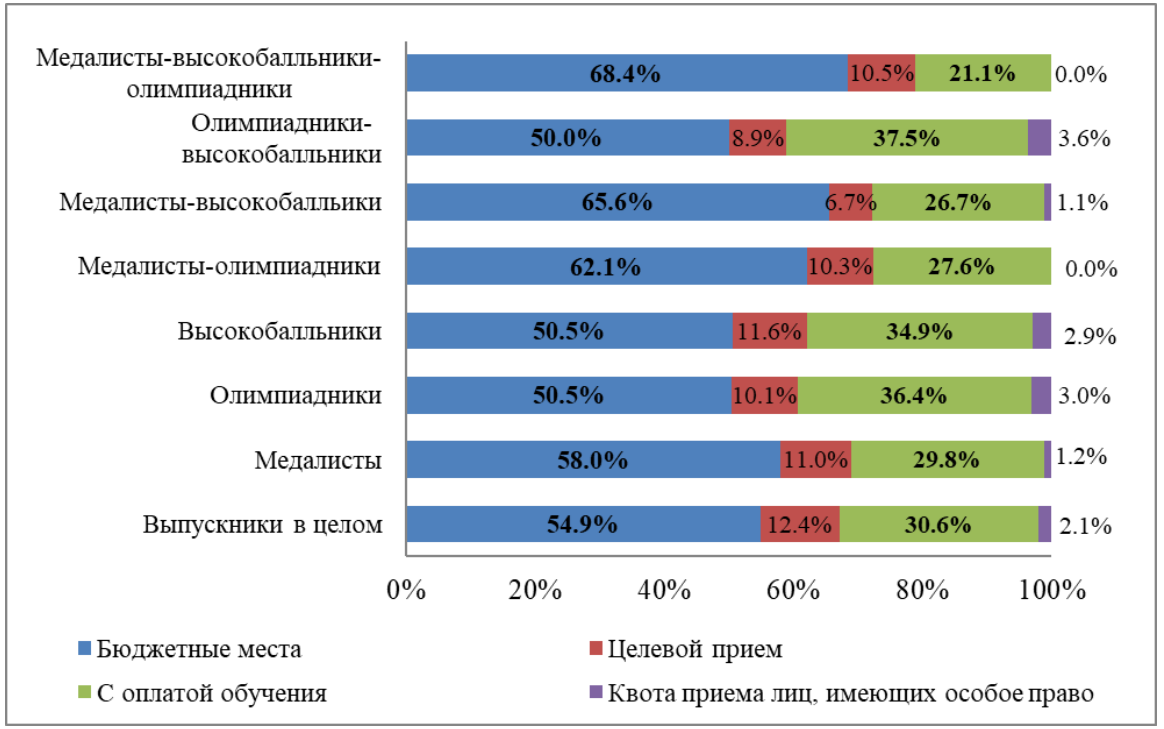
Примеры исследовательских тем и проектов:

- Образовательная миграция
- Оценка качества образования и качества управления образованием
- Оценка образовательного неравенства

Образовательная миграция различных категорий обучающихся (на примере одного региона)







Сочетание со статистикой и миграционными данными

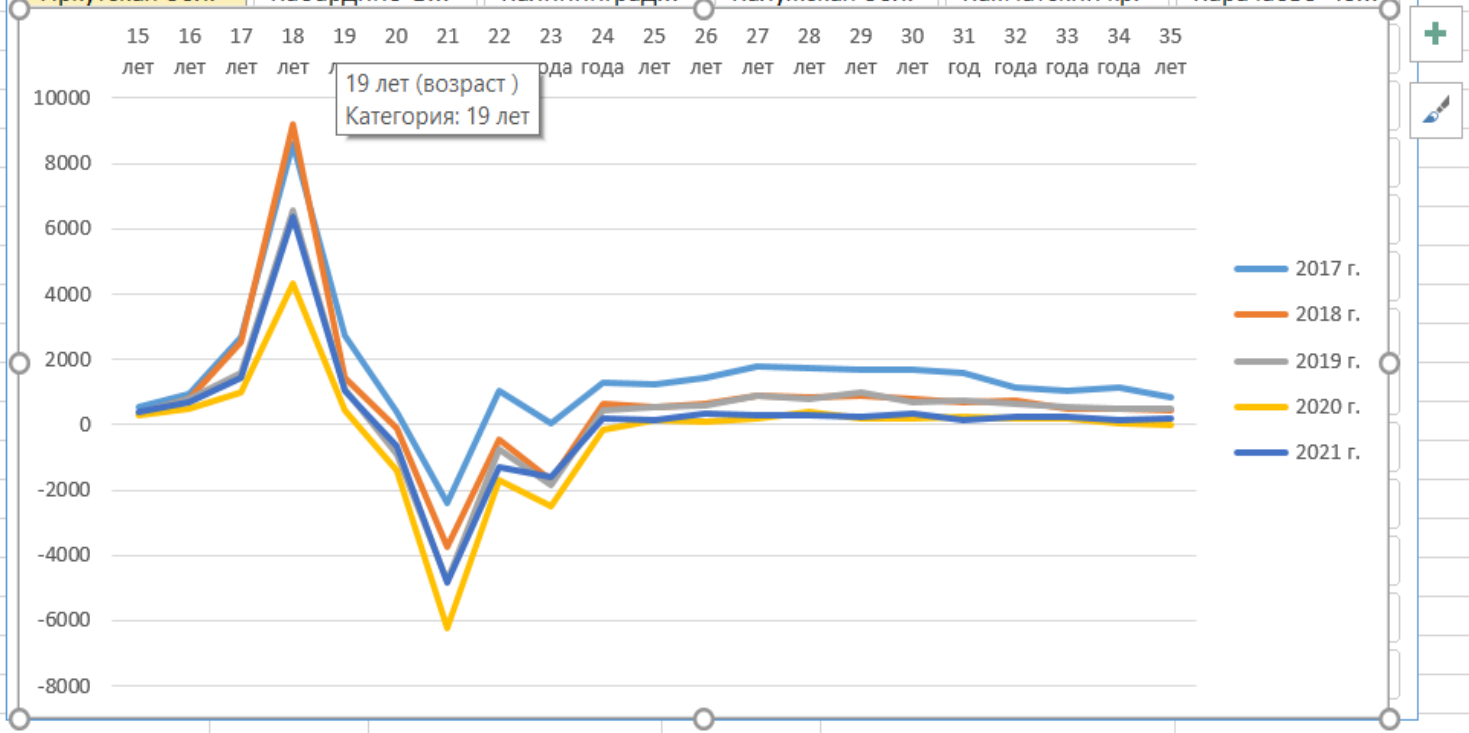
Территории	г. Санкт-Петербург
показатель	(Все)

показатель: выбывшие | прибыв...

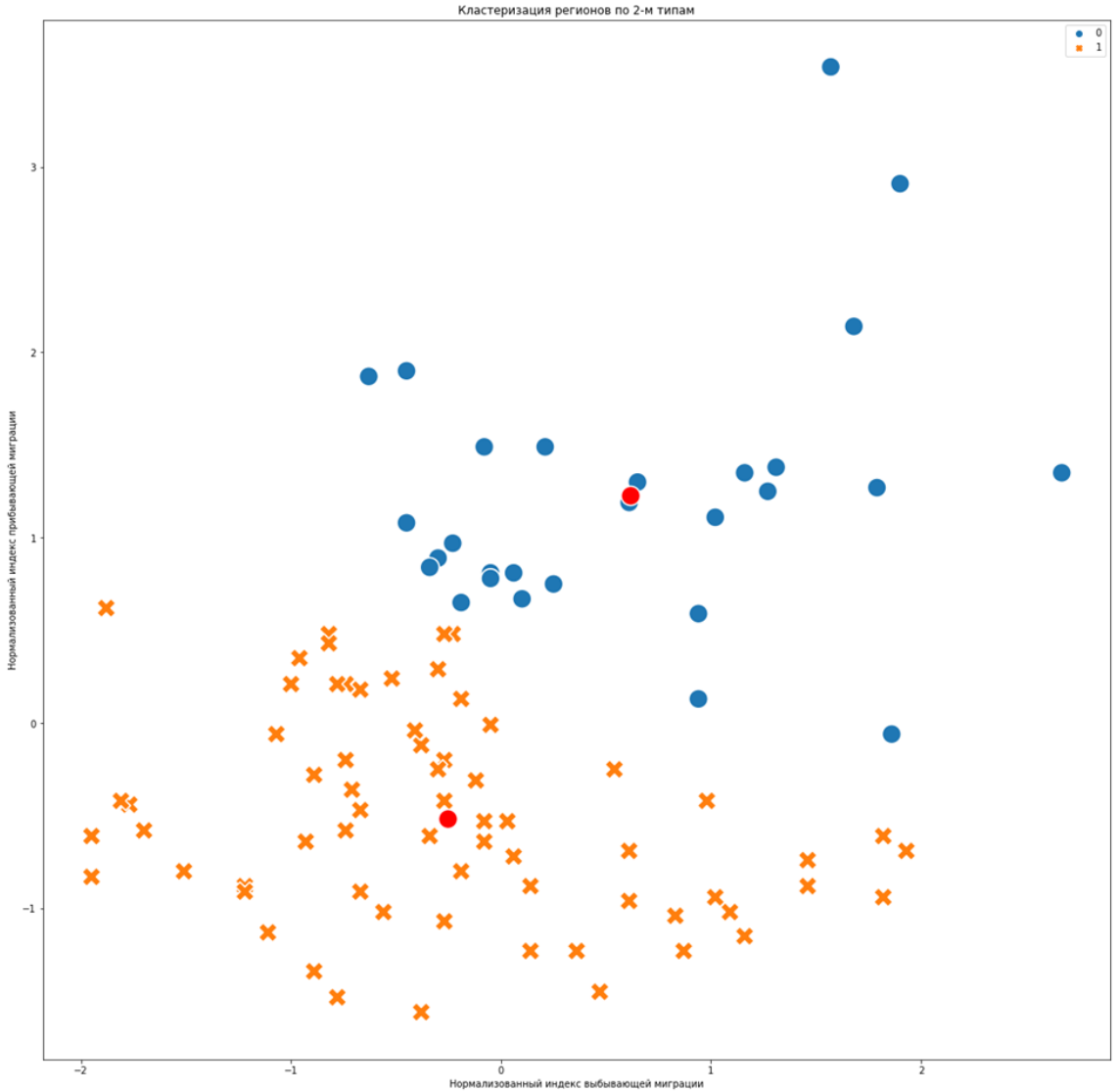
возраст/период	2017 г.	2018 г.	2019 г.	2020 г.	2021 г.
15 лет	554	333	354	283	372
16 лет	945	800	769	514	685
17 лет	2661	2520	1608	979	1461
18 лет	8606	9203	6557	4316	6342
19 лет	2726	1449	1076	432	1021
20 лет	375	-115	-921	-1404	-628
21 год	-2386	-3717	-4803	-6218	-4833
22 года	1021	-428	-745	-1697	-1277
23 года	38	-1711	-1864	-2503	-1613
24 года	1296	642	426	-133	180
25 лет	1264	524	551	130	143
26 лет	1453	620	596	115	325
27 лет	1772	875	875	204	313
28 лет	1730	861	800	418	317
29 лет	1666	909	969	204	261
30 лет	1672	799	713	210	354
31 год	1601	684	752	249	140
32 года	1128	728	622	199	266
33 года	1036	472	543	213	252
34 года	1151	502	482	61	161
35 лет	817	467	475	-3	177
Общий итог	**31126**	**16417**	**9835**	**-3431**	**4419**

Территории

Алтайский кр.	Амурская обл.	Архангельска...	Архангельска...	Астраханская ...	Белгородская...
Брянская обл.	Владимирска...	Волгоградска...	Вологодская ...	Воронежская ...	г. Москва
г. Санкт-Пете...	г. Севастополь	Дальневосточ...	Еврейская авт...	Забайкальски...	Ивановская о...
Иркутская обл.	Кабардино-Б...	Калининград...	Калужская обл.	Камчатский кр.	Карачаево-Че...

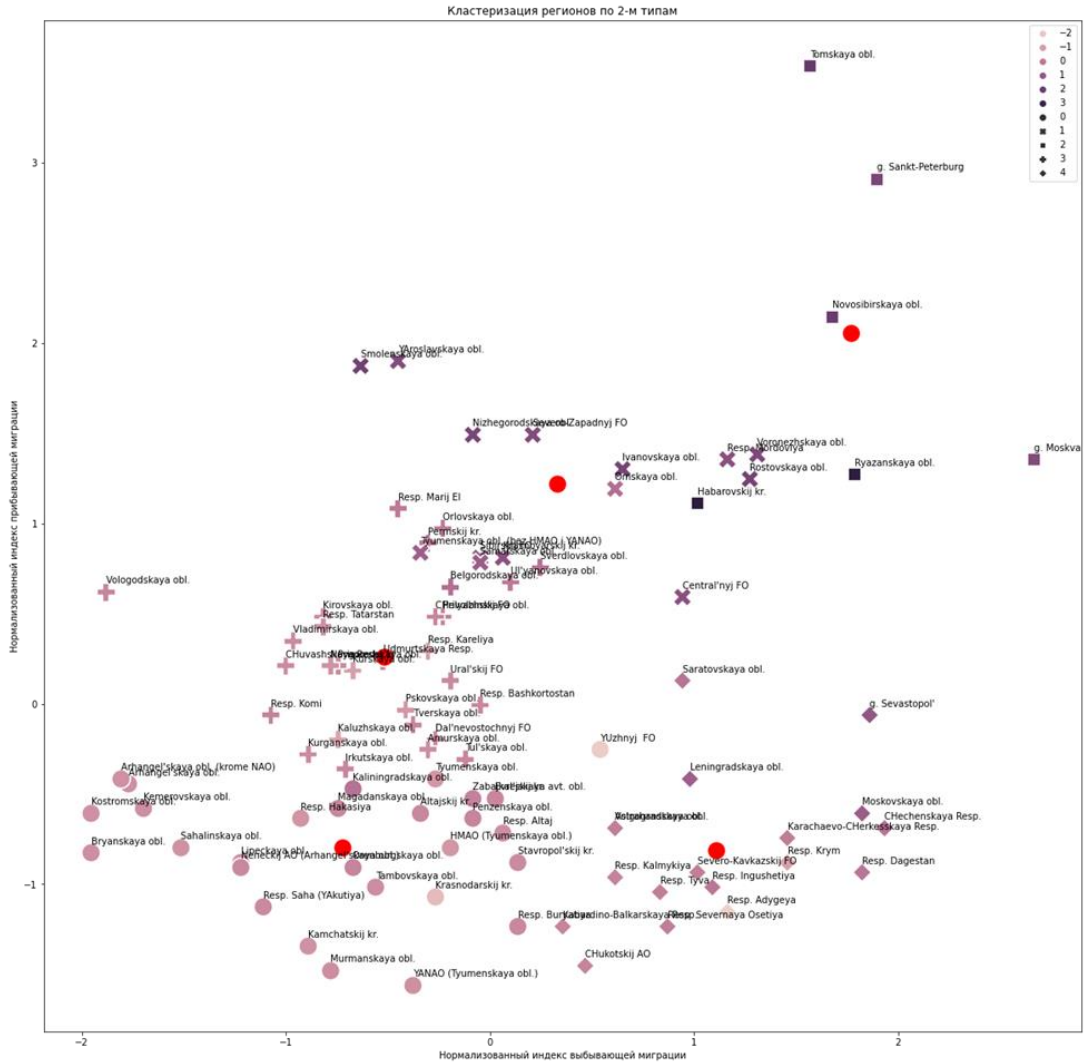


Кластеризация регионов по 2-м типам методом К-средних.

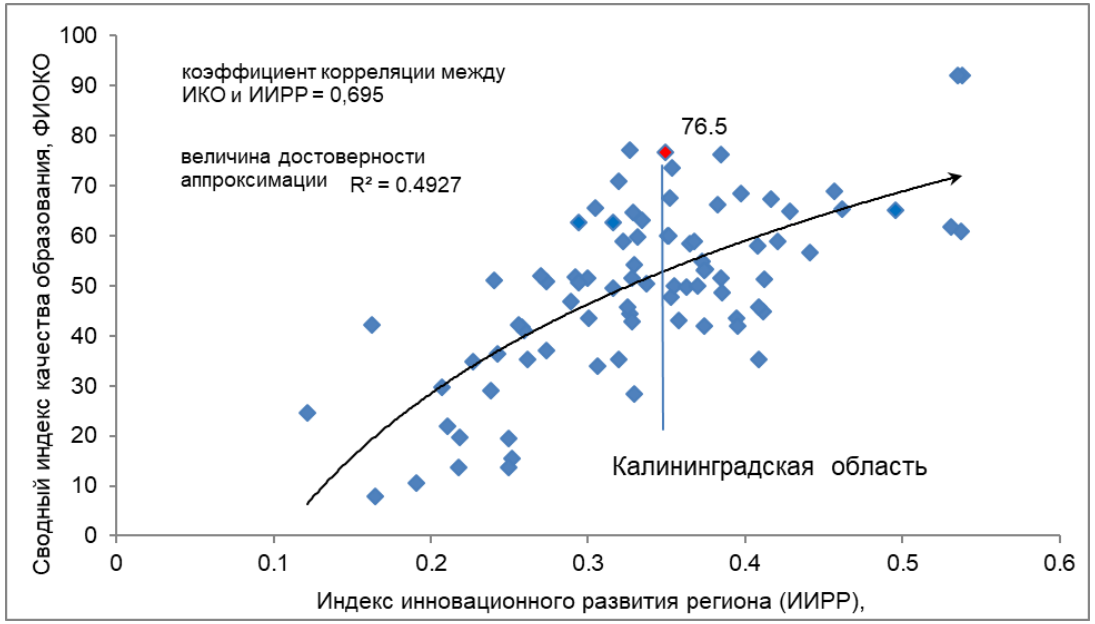


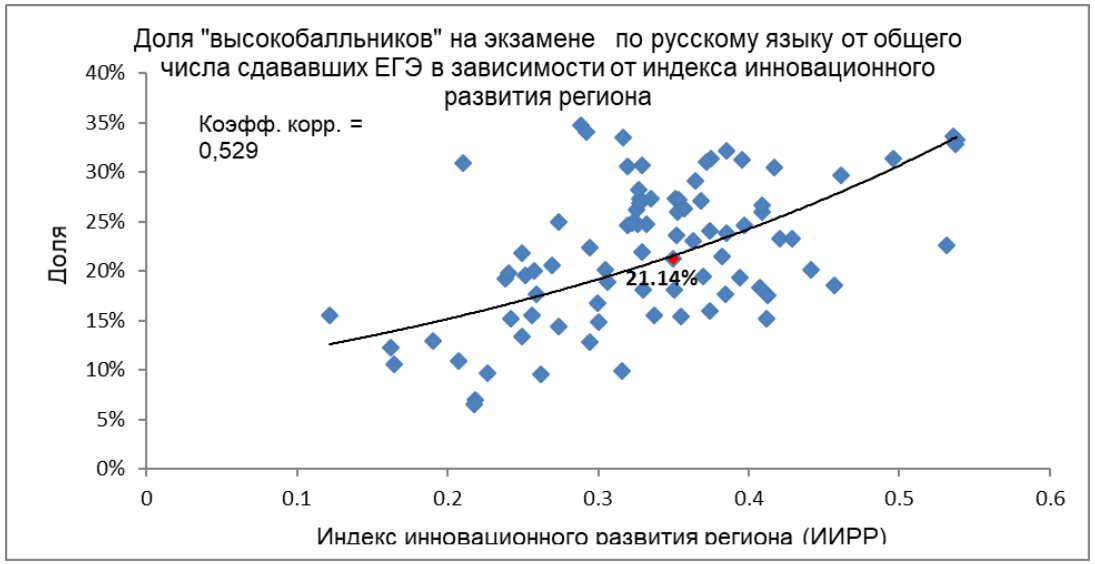
3DPlot.html

Кластеризация регионов по 5-м типам методом К-средних.



Оценка качества образования и качества управления образованием





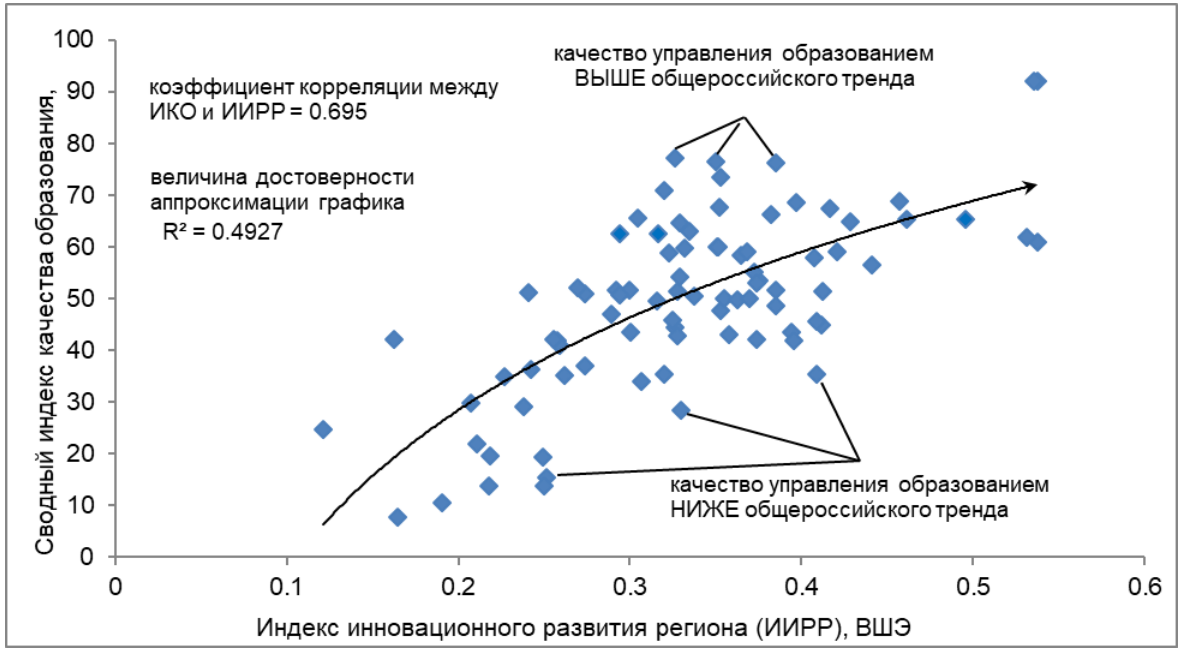
Коэффициенты корреляции между результатами ЕГЭ и ИИРР и ИКО

Предмет ЕГЭ	Когорта успешности	Корреляция с ИИРР	Корреляция с ИКО
Русский язык	«Высокобалльники», 81+	0,529	0,558
	«Не сдавшие», 27 -	-0,461	-0,642
Профильная математика	Высокобалльники», 81+	0,698	0,724
	«Не сдавшие», 27-	-0,566	-0,624
Физика	«Высокобалльники», 81+	0,632	0,718
	«Не сдавшие», 35-	-0,586	-0,722

Оценка образовательного неравенства в регионах

Коэффициенты корреляции доли низкобалльников ЕГЭ по региону от ИСЭПР, ИИРР и ИКО

Предметы ЕГЭ	корр. доли от ИСЭПР	корр. доли от ИИРР	корр. доли от ИКО
Биология, 2020	-0.32	-0.45	-0.56
Биология, 2019	-0.36	-0.48	-0.63
ИКТ, 2020	-0.40	-0.50	-0.68
ИКТ, 2019	-0.46	-0.53	-0.71
История, 2020	-0.38	-0.57	-0.69
История, 2019	-0.41	-0.52	-0.68
Матем проф, 2020	-0.36	-0.50	-0.61
Матем проф, 2019	-0.42	-0.57	-0.62
Обществоз., 2020	-0.35	-0.50	-0.60
Обществоз., 2019	-0.40	-0.52	-0.74
Русский язык, 2020	-0.41	-0.57	-0.76
Русский язык, 2019	-0.38	-0.46	-0.64
Физика, 2020	-0.37	-0.54	-0.71
Физика, 2019	-0.44	-0.54	-0.70
Химия, 2020	-0.15	-0.16	-0.11
Химия, 2019	-0.08	-0.10	-0.18



Перспективы:

1. нейросети, искусственный интеллект, машинное обучение;
2. экспертные системы поддержки принятия решений, мониторинга;
3. человеко-машинные и полностью машинные системы организации и управления образовательной деятельностью и обеспечивающими процессами

Для сферы образования:

- Анализ и визуализация образовательных данных и обеспечивающих образовательную деятельность процессов
- Предоставление обратной связи для оценки и улучшения эффективности работы преподавателей
- Адресные, персонализированные рекомендации для обучающихся
- Прогнозирование успеваемости учащихся
- Моделирование образовательной деятельности и образовательных маршрутов обучающихся
- Выявление нежелательного поведения обучающихся
- Группирование учащихся по категориям и признакам
- Анализ социальных сетей обучающихся
- Создание программного обеспечения и контента для курсов
- Планирование и реализация сетевых (между организациями) образовательных программ, персонализированных программ с учетом особенностей каждого

ГДЕ УЧИТЬСЯ?

Бесплатно и со скидкой

Открытое образование

stepik

coursera

Сложно… 2-3 курс матфака
Большой отсев… 80-95%
Системы СДО «висят»…
Требуется МНОГО времени и практики