



ОБРАЗОВАТЕЛЬНЫЙ  
ЦЕНТР МГТУ им. Н. Э. Баумана

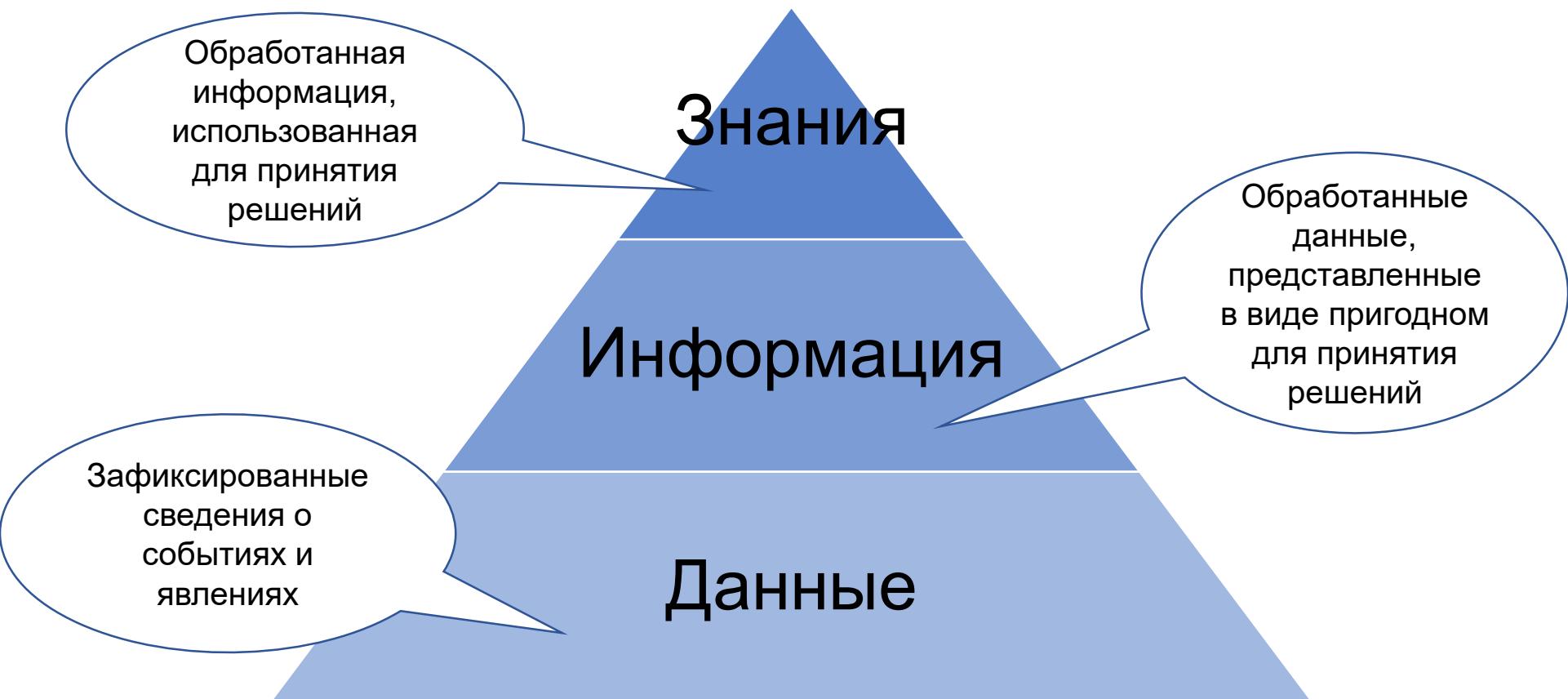
# Data Science

## Источники данных

Корпоративное обучение на базе  
Образовательного центра МГТУ им. Н. Э. Баумана  
под управлением МИЦ «Композиты России»  
Докладчик: Панфилов И.А. канд. техн. наук, доцент

Крупица информации дороже горы данных.  
Крупица знаний дороже горы информации.  
Крупица понимания дороже горы знаний.

Рассел Акофф





# Проблемы с ретроспективными данными

Формат данных, а также носитель данных может иметь принципиальное значение и вызывать массу проблем при работе с ним.

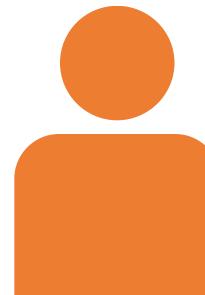


# Достоверность собранных данных

Документы  
на получение  
кредита



Документы  
на получение  
субсидии



70%



30%

60%



20%

75%



40%

80%



15%

# Манипуляции с данными



## Уровень ВВП в СССР и ведущих капиталистических странах в 1928–1939 годах

В долларах США 1990 года



Источник: The Maddison Project, 2018

# Пропуски в данных

	A	B	C	D	E	F	G	H	I	J
1	ОАО "КрАЗ", Корпус ХХ, Эл-ер XXXX, Для кластеризации ОА. 20XX г.									
2	Группа: Нет.									
3										
4	Сутки	АЭ: длит. <Сред.> (мин)	АЭ: кол-во <Сред.> (шт.)	АЭ: напр. ср. м. <Сред.> (В)	Время в недопитке <Сред.> (час)	Время в номинале <Сред.> (час)	Время в перепитке <Сред.> (час)	Время в тесте <Сред.> (час)	Время на голод. <Сред.> (час)	Кол-во доз АПГ в недопитке <Сумма> (шт.)
5	Год	2,22	0,342	57,26	11,89	4,95	6,28	0	1,17	165124
6	1 янв.	1,7	1	45	10,7	4,8	6,5	0	1,1	371
7	2 янв.		0		12,2	4,7	7,1	0		392
8	3 янв.		0		11,3	4,7	7,5	0		368
9	4 янв.		0		9,8	5,8	8,3	0		311
10	5 янв.	0,3	1	40	9	7,4	7,7	0		298
11	6 янв.	2,1	1	47	12,5	4,8	4,8	0	1,5	482
12	7 янв.	4,1	1	79	10,9	4,3	6	0	2,3	413
13	8 янв.		0		11,4	4,9	7,6	0		405
14	9 янв.		0		10,1	5,8	7,6	0		305
15	10 янв.		0		11,7	4,8	7,5	0		411
16	11 янв.		0		9,7	6,3	7,5	0		340
17	12 янв.	2,7	1	69	10,9	5	6,7	0	1,1	450
18	13 янв.	2,2	1	73	9,3	5,3	7	0	1,7	263
19	14 янв.	1,4	1	41	7,7	7,1	8,9	0		240

## ГРАФИК USD RUB ОНЛАЙН

[Как пользоваться графиком?](#)



График выглядит непрерывным, но это не так...

В каждые выходные дни торги на биржах останавливаются...

# Выбросы в данных



Михаил Шуклин | [vk.com/mikhailshuklin](https://vk.com/mikhailshuklin)

РМПР: коэф. <Сред.>	РМПР: кол-во ВИРА <Сред.> (шт)	РМПР: кол-во МАЙНА <Сред.> (шт)	Эл-лит: темпер <Сред.> (°C)	Металл: уровень <Сред.> (см)	Эл-лит: уровень <Сред.> (см)	Шум <Сред.> (В)
16,81	2,6	5,2	952	42,2	17,8	0,0255
15,6	6	9	955	40	17	0,035
18	2	5	959	40	17	0,036
18	1	0	959	40	17	0,034
19,8	3	2	957	40	18	0,034
21,8	4	8	957	40	17	0,035
17,1	2	12	954	41	19	0,035
18	4	8	957	41	18	0,037
17,5	3	3	958	41	20	0,035
15,6	2	3	1255	41	20	0,035
13,5	1	1	954	39	18	0,036
17,7	5	6	956	41	18	0,034
16,7	3	6	948	41	18	0,033
15,3	3	6	953	41	17	0,031
12,8	2	4	949	40	15	0,031
10,8	2	4	950	40	17	0,029
11,5	1	1	950	42	18	0,03

# Ошибки ручного ввода данных

## Запись

Введенные слова или показатели не те, что были в оригиналe.

«Красноярский край» → «Краснодарский край»

## Вставка

Появление дополнительного символа: 56,789 → 564,789.

## Удаление

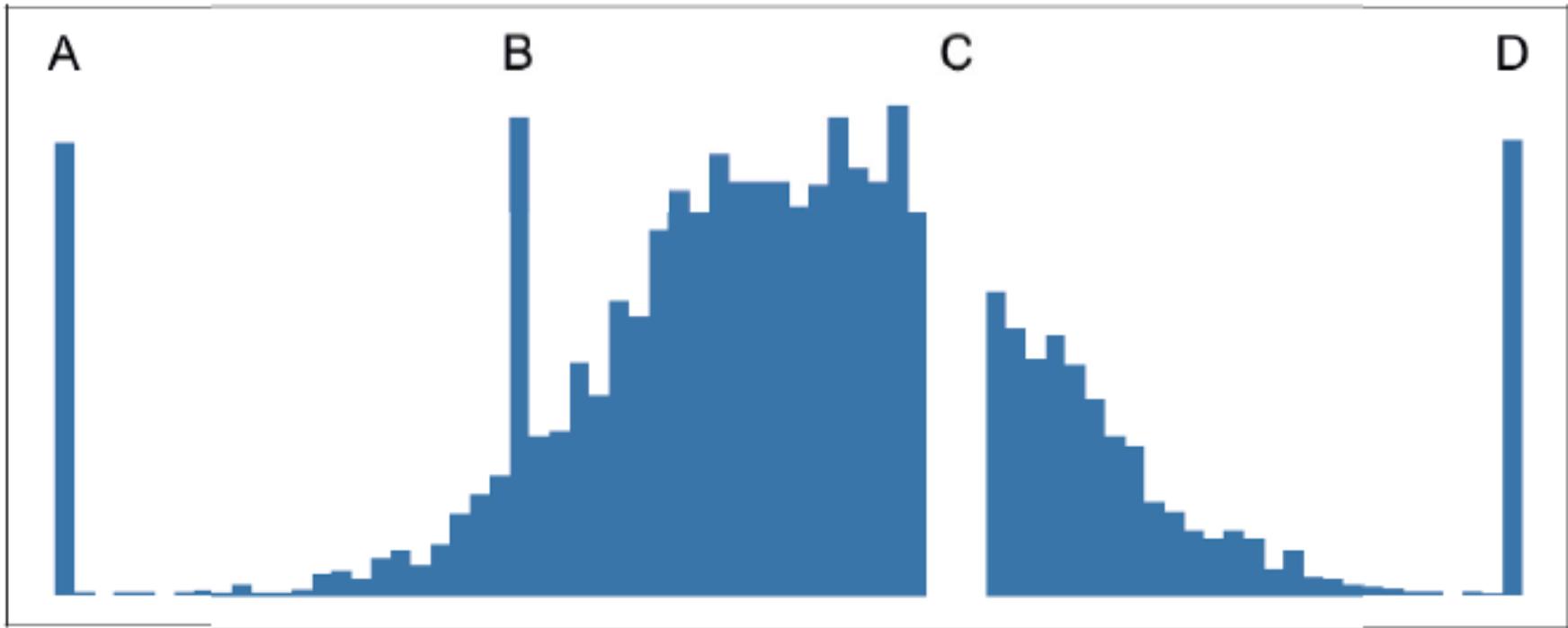
Один или несколько символов теряются: 56,789 → 56,89.

## Перемена мест

Два или более символов меняются местами: 56,789 → 56,798.

# Актуальность данных

	Факт															Oценка	
Показатель	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Объем валового регионального продукта в ценах текущих лет, млн. руб.	214663	239420	230995	272727	365454	439737	585882	734155	737951	749195	1055525	1170827	1183228	1256934	1423247	1579105	1 687 813
Индекс физического объема ВРП (в постоянных ценах), в процентах к предыдущему году	105,1	106,3	104,0	105,6	106,4	103,3	104,4	106,0	104,6	98,5	105,8	105,7	105,8	102,9	101,0	97,8	98,57
Инвестиции в основной капитал за счет всех источников финансирования в ценах	25457	33758	31759	37196	49089	71388	92587	120833	204171	247789	266910	308588	381657	376903	363996	394410	419059
Темп роста объема инвестиций в основной капитал, % к предыдущему году в сопоставимых	143,7	111,6	87,7	109,7	118,8	129,5	116,6	113,0	141,4	118,5	110,8	114,9	117,2	96,1	92,9	95,8	100,90



Примеры типов ошибок, которые можно выявить с помощью простой гистограммы:

- А — значения по умолчанию, такие как –1, 0 или 1/1/1900;
- В — неправильный ввод или повтор данных;
- С — пропущенные данные;
- Д — значения по умолчанию, такие как 999

# Требования по качеству данных:

## **Доступность**

У аналитика должен быть доступ к данным. Это предполагает не только разрешение на их получение, но также наличие соответствующих инструментов, обеспечивающих возможность их использовать и анализировать.

## **Точность**

Данные должны отражать истинные значения или положение дел. Например, показания неправильно настроенного термометра, ошибка в дате рождения или устаревший адрес — это все примеры неточных данных.

## **Взаимосвязанность**

Должна быть возможность точно связать одни данные с другими. Например, заказ клиента должен быть связан с информацией о нем самом, с товаром или товарами из заказа, с платежной информацией и информацией об адресе доставки. Этот набор данных обеспечивает полную картину заказа клиента. Взаимосвязь обеспечивается набором идентификационных кодов или ключей, связывающих воедино информацию из разных частей базы данных .

# Требования по качеству данных:

## *Полнота*

Под неполными данными может подразумеваться как отсутствие части информации (например, в сведениях о клиенте не указано его имя), так и полное отсутствие единицы информации (например, в результате ошибки при сохранении в базу данных потерялась вся информация о клиенте).

## *Непротиворечивость*

Данные должны быть согласованными. Например, адрес конкретного клиента в одной базе данных должен совпадать с адресом этого же клиента в другой базе. При наличии разногласий один из источников следует считать основным или вообще не использовать сомнительные данные до устранения причины разногласий.

## *Однозначность*

Каждое поле, содержащее индивидуальные данные, имеет определенное, недвусмысленное значение. Четко названные поля в совокупности со словарем базы данных (подробнее об этом чуть позже) помогают обеспечить качество данных.

# Требования по качеству данных:

## **Релевантность**

Данные зависят от характера анализа. Например, исторический экскурс по ценам на недвижимость в Москве может быть интересным, но при этом не иметь никакого отношения к анализу спроса на уголь в Китае.

## **Надежность**

Данные должны быть одновременно полными (то есть содержать все сведения, которые вы ожидали получить) и точными (то есть отражать достоверную информацию).

## **Своевременность**

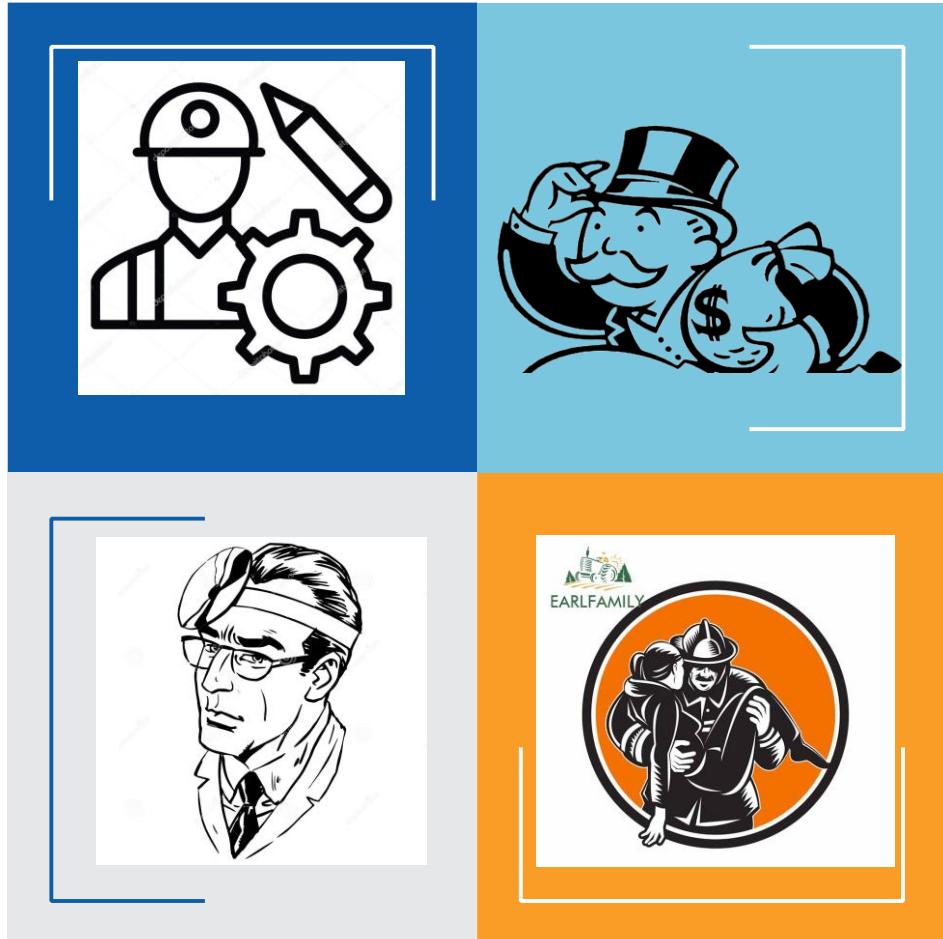
Между сбором данных и их доступностью для использования в аналитической работе всегда проходит время. На практике это означает, что аналитики получают данные как раз вовремя, чтобы завершить анализ к необходимому сроку. Например, если у компании время ожидания при работе с хранилищем данных составляет до одного месяца. При такой задержке данные становятся практически бесполезными (при сохранении издержек на их хранение и обработку), их можно использовать только в целях долгосрочного стратегического планирования и прогнозирования.

# Экспертное знание



# Предметный эксперт

**ЭКСПЕРТ.** (лат. *expertus* – опытный) в широком смысле специалист в **определенной** области, привлекаемый для исследования, консультирования, выработки суждений, заключений, предложений, проведения **экспертизы**



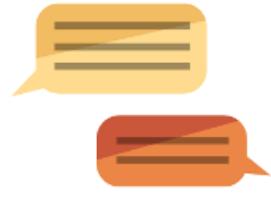
# Доступ к эксперту



# Извлечение информации из эксперта



# ТОП 10 Soft Skills



Коммуникация



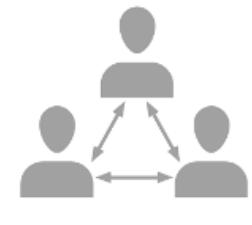
Мотивация



Лидерство



Ответственность



Командная  
работа



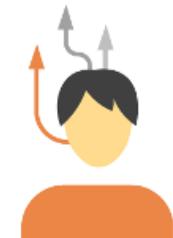
Решение  
проблем



Принятие  
решений



Работоспособность и  
стессоустойчивость



Гибкость



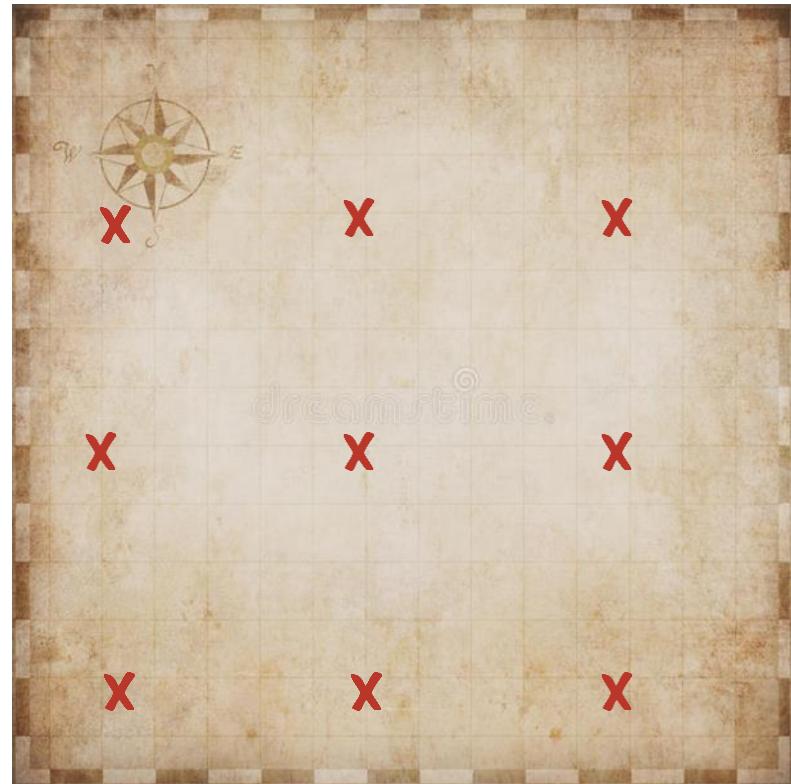
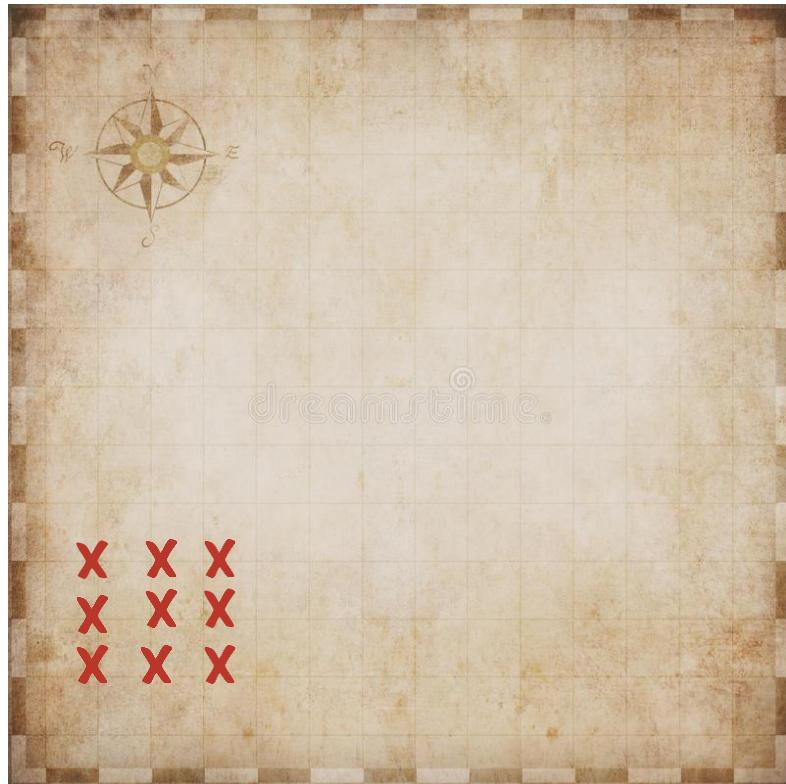
Умение вести  
переговоры

# Экспериментальные данные



Количество решенных задач $x_i$	Абсолютная частота $n_i$	Накопленная абсолютная частота	Относительная частота $f_i$	Накопленная относительная частота
8	5	5	0,10	0,10
10	10	15	0,20	0,30
11	15	30	0,30	0,60
12	11	41	0,22	0,82
14	2	43	0,04	0,86
16	3	46	0,06	0,92
17	4	50	0,08	1,00

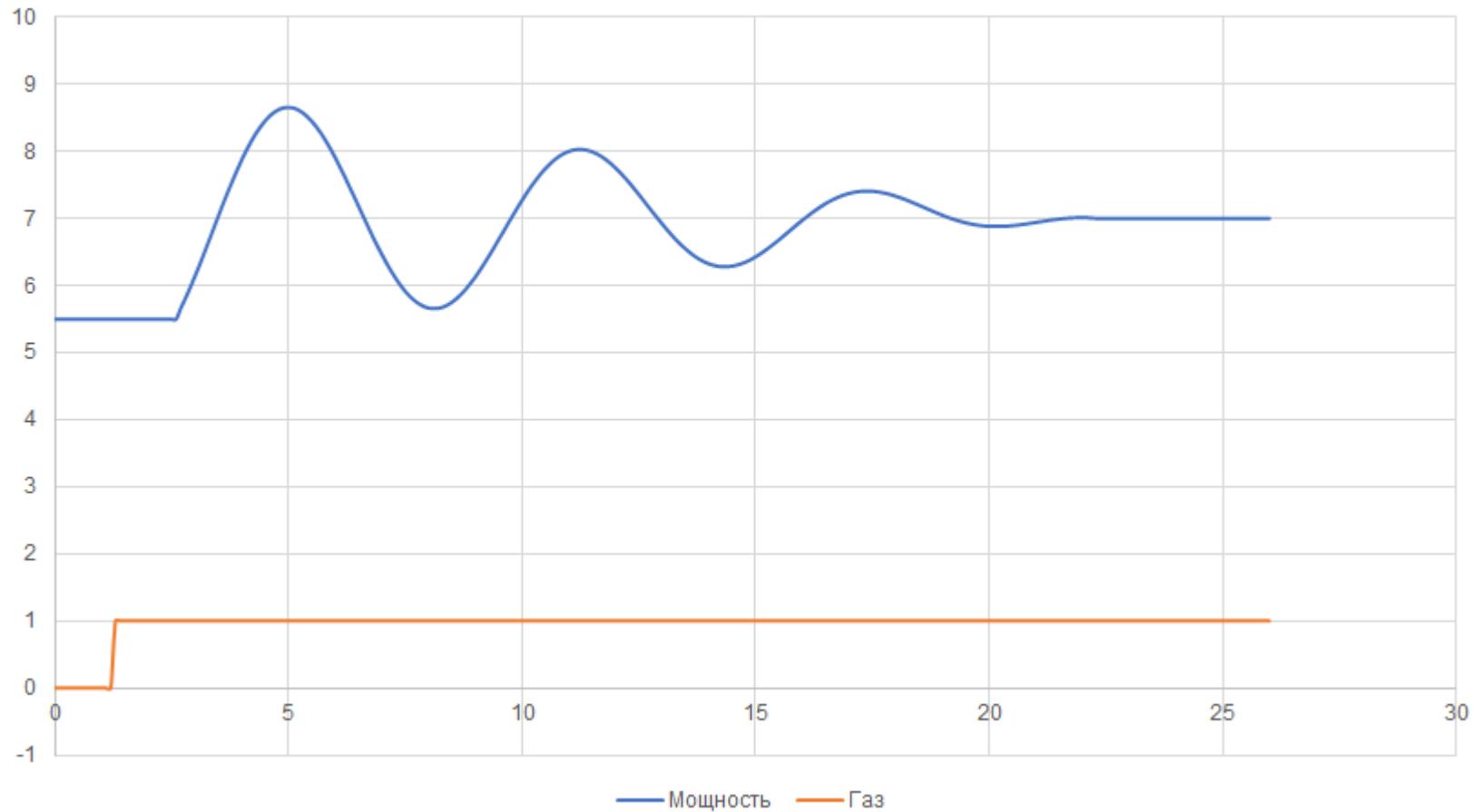
# Планирование эксперимента



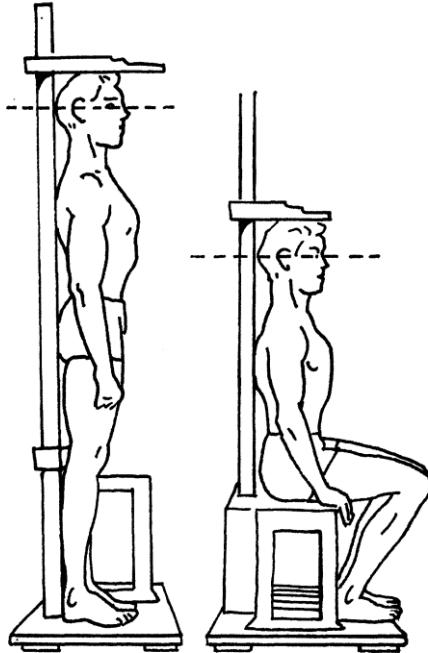
# Активный и пассивный эксперимент



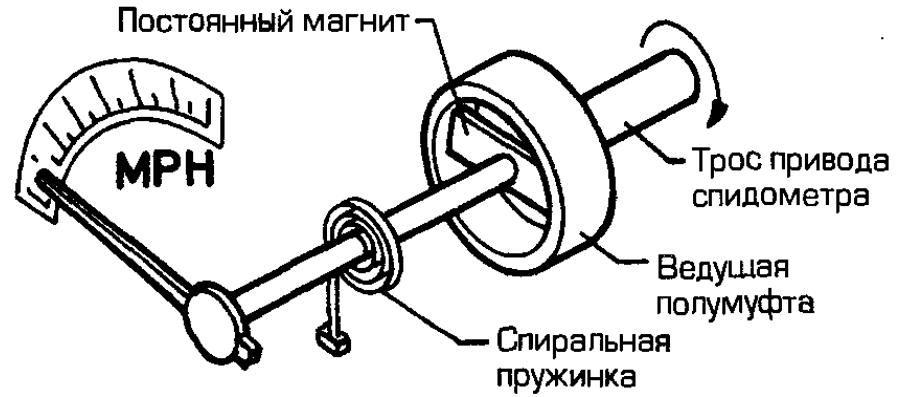
## Активный эксперимент, единичное воздействие



# Прямые и косвенные эксперименты



Измеряем рост, см



Измеряем скорость авто, км/ч

# Измерительные шкалы

- 1 Номинальные шкалы (категориальные)**  
Номера телефонов, цвет автомобиля, фамилия
- 2 Порядковые шкалы (ранговые)**  
Школьные оценки, баллы Яндекс пробки, места на соревнованиях
- 3 Шкалы расстояний (разностей)**  
Температура, летоисчисление
- 4 Шкалы отношений (дискретные и непрерывные)**  
Расстояние, вес, объем, измерение материалов
- 5 Абсолютная шкала**  
математика

Точность

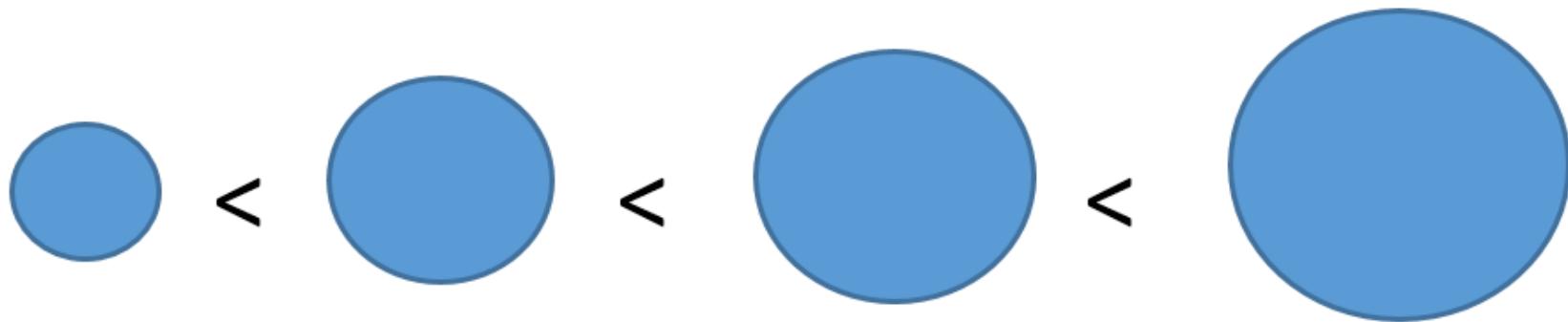
# Номинальная шкала

8 800 200 33 44 + 8 750 500 66 77 = **X**

8 800 200 33 44 **≠** 8 750 500 66 77 **✓**

Пол	Цвет авто	Номер дома	Номер трамвая
Муж	<b>Красный</b>	2	3
Жен	<b>Красный</b>	4	4
Муж	<b>Черный</b>	18	3
Муж	<b>Синий</b>	11	3
Всего мужчин 3 Всего женщин 1	<b>Красных авто больше чем синих</b>	Домов на четной стороне улицы больше	Интервал движения трамвая №3 короче

# Ранговые шкалы



Свободная  
дорога

Дорога займет  
в 1,4 раза больше  
времени

Дорога займет  
в 2 раза больше  
времени



ЭТО НЕПРАВДА (((((

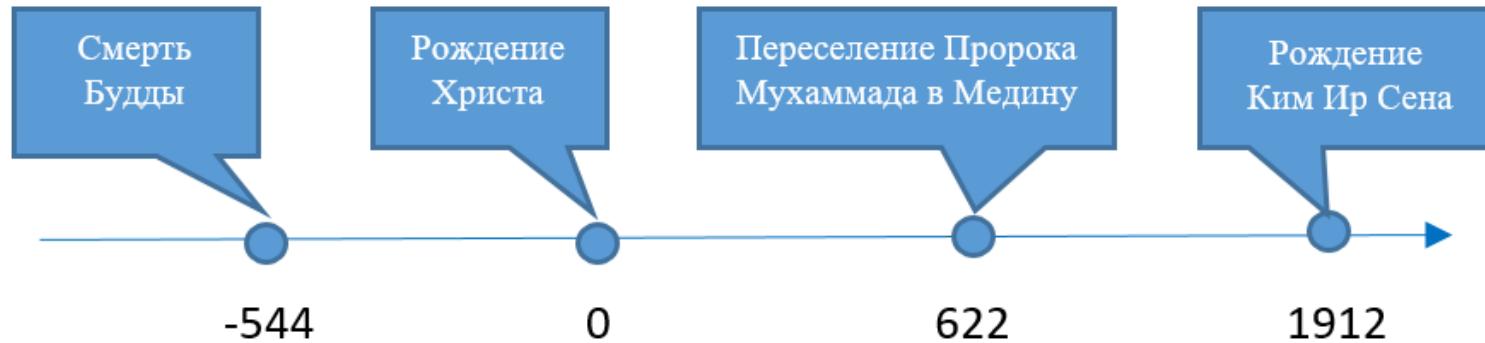
№	Страна				
1	США	39	41	33	113
2	Китай	38	32	18	88
3	Япония	27	14	17	58
4	Великобритания	22	21	22	65
5	Россия	20	28	23	71

# Шкала расстояний

$10^{\circ}\text{C} \rightarrow 20^{\circ}\text{C}$  температура увеличилась в 2 раза



$10^{\circ}\text{C} \rightarrow 20^{\circ}\text{C}$  температура увеличилась на  $10^{\circ}\text{C}$



$$1980 + 1985 = 3965$$



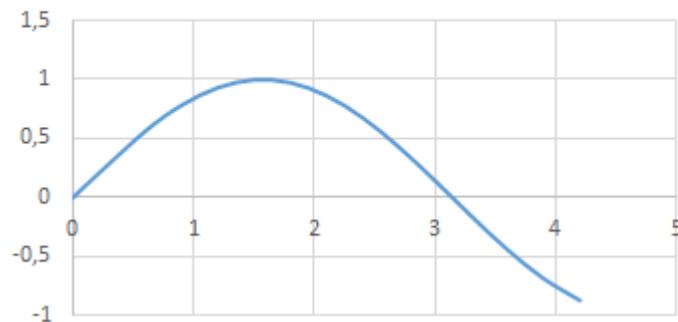
$$1980 + 1985 = -5$$



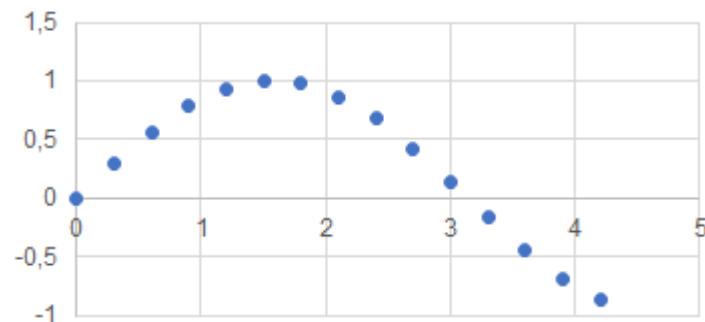
# Шкалы отношений

Непрерывные		Дискретные	
Расстояние	3.750 км; 11.999 км	Количество машин	12; 2; 100; 33
Вес	250 г; 0.001 г; 15.05 г	Число поездок	1; 14; 30
Деньги	0 руб.; 11 евро; 12.40 \$	Объем воды в бутылках	0.5 л; 4 л; 3.5 л

Непрерывные значения

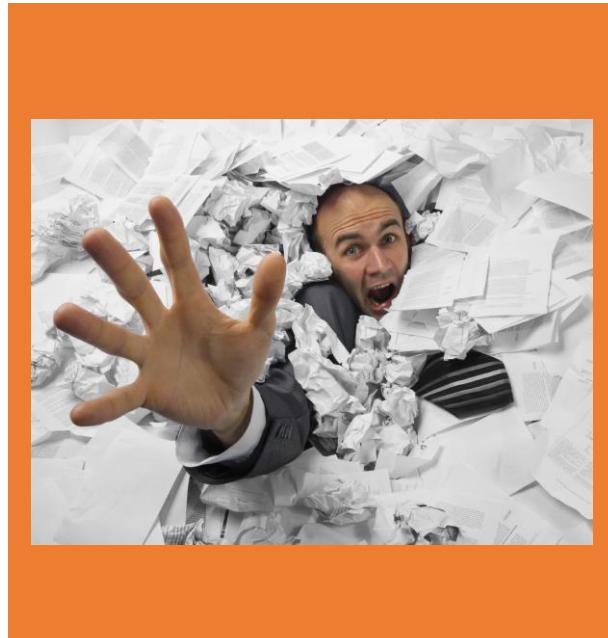


Дискретные значения



# Полезные ссылки

- 1** <https://archive.ics.uci.edu/ml/index.php>  
Репозиторий задач машинного обучения
- 2** <https://datasetsearch.research.google.com/>  
Data Set Search от Google
- 3** <https://www.kaggle.com/>  
Профессиональный портал для Data Scientists
- 4** <https://dorozhniij.com/opendata>  
Портал открытых данных России
- 5** <https://ods.ai/>  
Ещё один большой портал
- 6** <http://aisori-m.meteo.ru/waisori/index.xhtml?idata=5>  
Архив метеоданных





**edu.bmstu.ru**

+7 (495) 120-30-75

E-mail: [edu@bmstu.ru](mailto:edu@bmstu.ru)

Москва, ул. 2-я Бауманская,  
дом 5, стр. 1