

Основы анализа и визуализации данных для медиа

Лекции 1-2
2 сентября 2019

Рогович Татьяна
rogovich@gmail.com

О курсе

Преподаватель

Рогович Татьяна, rogovich@gmail.com

Ассистент

Филатов Михаил, mihfil1998@yandex.ru

Программа курса

ТВА

Wiki-страница курса

http://wiki.cs.hse.ru/%D0%9E%D1%81%D0%BD%D0%BE%D0%B2%D1%8B_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0_%D0%B8_%D0%B2%D0%B8%D0%B7%D1%83%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85_%D0%B4%D0%BB%D1%8F_%D0%BC%D0%B5%D0%B4%D0%B8%D0%B0_2019/2020

Чат курса в Telegram

<https://t.me/hse2019journalists>

Цели и задачи курса

- Развитие навыка чтения, интерпретации и оценки качества анализа и представления статистических данных
- Формирование навыка оценки качества визуализации таких данных и ее корректности
- Освоение инструмента визуализации количественной информации

Про что не будем говорить на курсе: про иллюстративную инфографику

АНАТОМИЯ ПЕТЕРБУРГА как устроен город

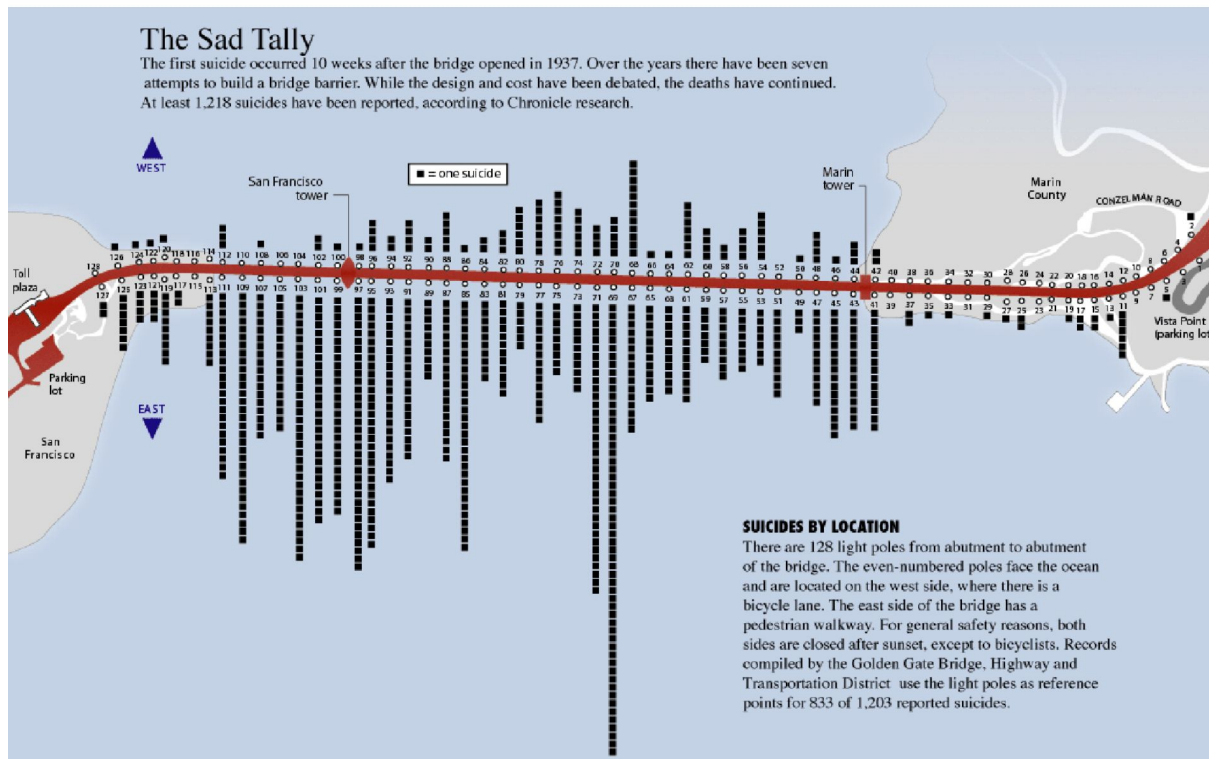
Исх. № 275
от 17-10-2014 г.

Небесная линия

Еще в начале строительства Петербурга дома ставили «по линиям и в один горизонт», то есть по периметру кварталов и одинаковой высоты. Здесь же на больших расстояниях друг от друга выросли многочисленные сооружения, сильно возвышающиеся над средним уровнем застройки.



Про что не будем говорить на курсе: Про то как графически вписывать статистические графики в контекст



Но при этом научимся
делать статистически
корректные графики, с
которыми потом смогут
работать дизайнеры.

Обратите внимание, что здесь все визуальные эффекты достигнуты исключительно техническими средствами, здесь нет дизайна ради дизайна.

[illegible]

Как этот курс будет полезен?

- Можно прокачать навыки, которые впоследствии пригодятся в научной, экономической или исследовательской журналистике (такой, где обычно работают эксперты области)
- Можно научиться технически визуализировать статистические данные. Мы не будем делать шедевры инфографики, но шедевры инфографики часто могут быть бессмысленны, если данные принесены в жертву красоте.
- Можно узнать важные вещи про психологию восприятия, которые пригодятся всем, кто работает с контентом.
- Поработаем с умением критически воспринимать информацию.

Структура курса

- Введение в статистику (разбираемся с базовыми понятиями и терминами, узнаем, что влияет на валидность исследований и выводов).
- Введение в визуализацию количественных данных (разбираемся с психологическими особенностями восприятия, поговорим об эстетике, научимся видеть ошибки и манипуляции с данными).
- Работаем с Tableau (учимся загружать и визуализировать различные данные, разбираемся, какие графики подходят для каких случаев, учимся делать небольшие “визуальные исследования”).

Структура курса

1 модуль

8 лекций (недели 1-4)

11 семинаров

Экзамен в сессию
(защита проекта)

Оценивание

1. Письменная контрольная работа по статистике (КР1)
2. Презентация плохой визуализации (ДЗ1)
3. Контрольная работа в Tableau (КР2)
4. Проект (Устная защита на экзамене во время сессии)

Окончательная оценка = Округление
 $(0.6 * (КР1 + КР2) + 0.15 * ДЗ1 + 0.25 * \text{Проект})$

Статистические данные и журналистика

Структура лекции

- Как журналист может использовать статистические данные и как это делать корректно?
- Зачем визуализировать данные?
- Что делать, если хочется развивать критическое мышление?
- Посмотрим примеры визуализаций на различных данных (и поговорим, где их брать)

Что такое статистика?

Это отрасль знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых статистических (количественных или качественных) данных.

Статистика в нашем курсе

1. Познакомимся с базовыми понятиями и релевантными терминами, достаточными, чтобы понимать смысл простых исследований.
2. В связке с теорией визуализации данных посмотрим, какие виды визуализаций лучше подходят для разных типов данных и как правильно их интерпретировать.

Статистические данные в журналистике

1. Использовать результаты исследований и опросов как инфоповод.
Нужно знать хотя бы основы статистики и научного метода, чтобы верно интерпретировать результаты исследования!

Пример: важна правильная интерпретация результатов исследования

ВОЗ не нашла доказательств вреда от микропластика в питьевой воде

<https://nplus1.ru/news/2019/08/22/no-harm>

ВОЗ: Микропластик в питьевой воде безопасен для здоровья  КОММЕНТАРИИ

<https://ru.euronews.com/2019/08/22/no-health-risk-from-microplastics-in-drinking-water-ru>

Пример: важен контекст (особенно для
подозрительных данных)

Reality Check: Is Malmo the 'rape capital' of Europe?

<https://www.bbc.com/news/uk-politics-39056786>

Пример: важен контекст (особенно для подозрительных данных)

Reality Check: Is Malmo the 'rape capital' of Europe?

<https://www.bbc.com/news/uk-politics-39056786>

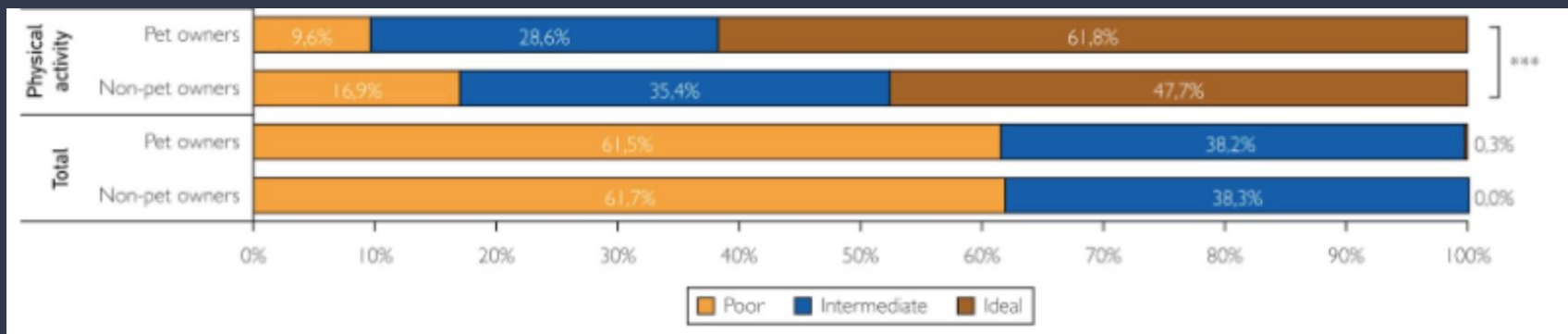
Sweden approves new law recognising sex without consent as rape

<https://www.bbc.com/news/world-europe-44230786>

Пример: нужно подвергать сомнению выводы авторов

Собачники отличились здоровьем сердца и сосудов

<https://nplus1.ru/news/2019/08/23/cvd-health-in-dog-owners>



Пример: нужно подвергать сомнению
источники информации

E-cigarettes: Can They Help People Quit?

<https://www.bbc.co.uk/programmes/p03hnfb6>

... и валидность опросов общественного мнения

The 1936 *Literary Digest* Poll

(нерепрезентативная выборка)

https://en.wikipedia.org/wiki/The_Literary_Digest

The 1992 UK general election

(shy Tories, social acceptability)

<https://www.itv.com/news/2017-04-18/times-the-election-polls-got-it-wrong/>

... иногда мы все равно, к сожалению, попадем
впросак

Grad Student Accused Of Faking Data In Gay Marriage Study

https://io9.gizmodo.com/grad-student-accused-of-faking-data-in-highly-publicize-1705814532?trending_test_two_e&utm_expid=66866090-68.hhyw_lmCRuCTCq0l2RHHtw.4&utm_referrer=http%3A%2F%2Fio9.com%2Ftag%2Fscience-scandals%3Ftrending_test_two_e

<https://www.onlineuniversities.com/blog/2012/02/the-10-greatest-cases-of-fraud-in-university-research/>

Статистические данные в журналистике

1. Использовать результаты исследований и опросов как инфоповод.
Нужно знать хотя бы основы статистики и научного метода, чтобы верно интерпретировать результаты исследования!
2. Самостоятельно собирать данные для историй.
Нужно знать и соблюдать определенные правила, чтобы такие данные были валидны.
The Data Journalism Handbook (2012)

Пример: data crowdsourcing

How many Olympic tickets did you get?
Here's our readers' results

We asked how you had fared in the London 2012 ticket ballot.
Here's our analysis of the information you gave us

<https://www.theguardian.com/sport/london-2012-olympics-blog/2011/jun/03/london-2012-olympic-ticket-crowdsourcing-result>

How much were you prepared to spend

Range	% chance of getting tickets
£1-£249	35.80%
£250-499	40.00%
£500-£749	42.90%
£750-£999	47.30%
£1,000-£1,499	53.50%
£1,500-£1,999	57.10%
£2,000-£2,499	59.40%
£2,500-£2,999	67.70%
£3,000-£3,999	60.20%
£4,000-£4,999	62.30%
£5,000 plus	83.10%

Пример: data crowdsourcing

Crowdsourcing the Sarah Palin emails: user guide

24,000 pages of Sarah Palin's emails have been released. Here's how to get involved in going through them

● **[Click here to see the Sarah Palin emails and get started](https://www.theguardian.com/world/datablog/2011/jun/10/crowdsource-sarah-palin-emails)**

<https://www.theguardian.com/world/datablog/2011/jun/10/crowdsource-sarah-palin-emails>

Статистические данные в журналистике

1. Использовать результаты исследований и опросов как инфоповод.
Нужно знать хотя бы основы статистики и научного метода, чтобы верно интерпретировать результаты исследования!
2. Самостоятельно собирать данные для историй.
Нужно знать и соблюдать определенные правила, чтобы такие данные были валидны.
The Data Journalism Handbook (2012)
3. Использовать данные для визуализаций в качестве бэкграунда или основной истории.
Нужно знать и как правильно визуализировать такие данные, не исказив их, и психологию восприятия визуальной информации.

The Changing Nature of Middle-Class Jobs

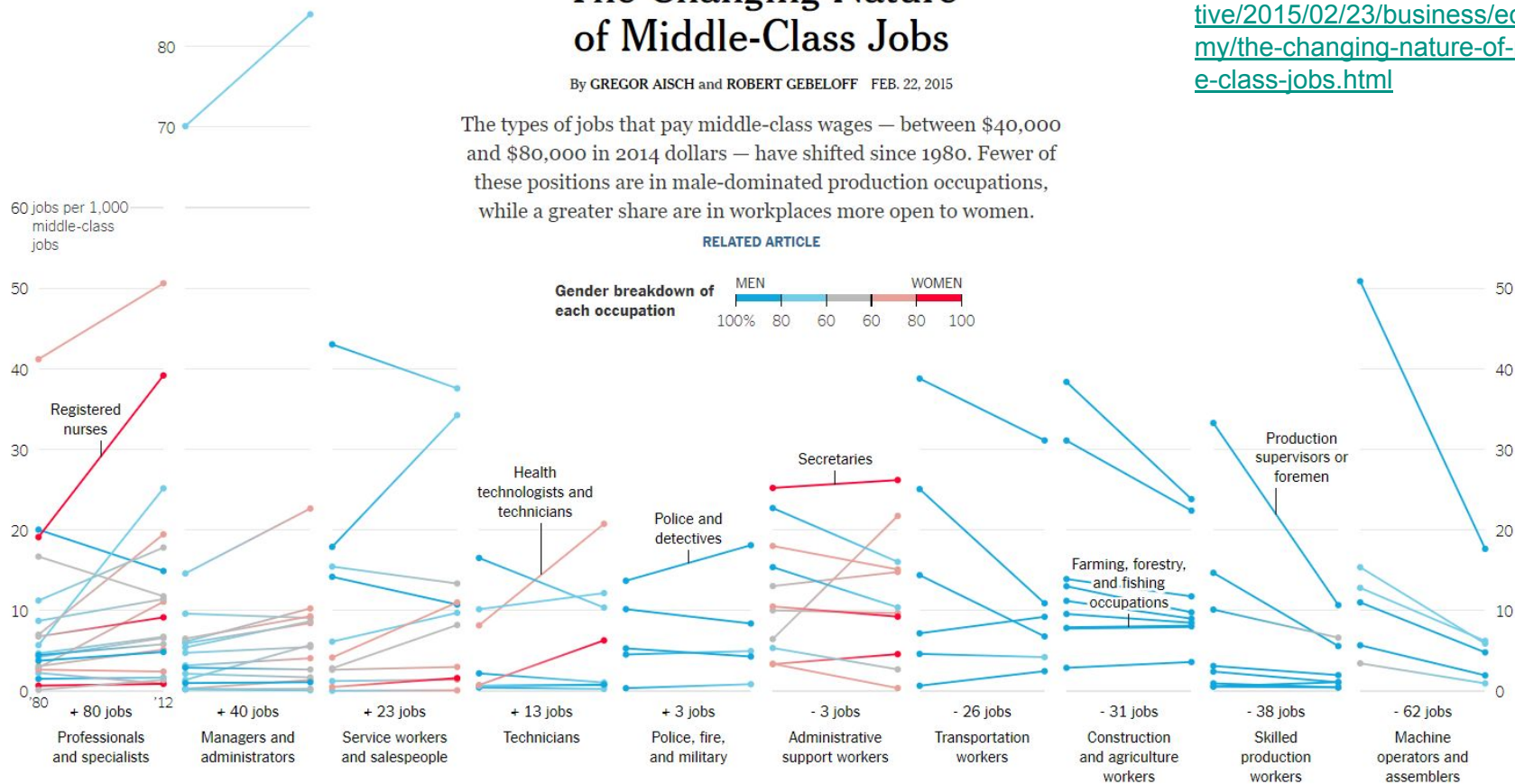
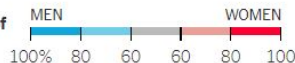
By GREGOR AISCH and ROBERT GEBELOFF FEB. 22, 2015

<https://www.nytimes.com/interactive/2015/02/23/business/economy/the-changing-nature-of-middle-class-jobs.html>

The types of jobs that pay middle-class wages — between \$40,000 and \$80,000 in 2014 dollars — have shifted since 1980. Fewer of these positions are in male-dominated production occupations, while a greater share are in workplaces more open to women.

RELATED ARTICLE

Gender breakdown of each occupation



<https://archive.nytimes.com/www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html?>

Four Ways to Slice Obama's 2013 Budget Proposal

Explore every nook and cranny of President Obama's federal budget proposal.

All Spending

Types of Spending

Changes

Department Totals

\$921.6 billion

Health and Human Services

\$885.3 billion

Social Security

\$620.3 billion

Defense

\$519.5 billion

Treasury

*Empty circles
show income.*

Зачем визуализировать данные?

Визуализация данных решает три главные задачи:

Задача визуализации	Цель
Разведка: поиск важных фактов	Обнаружение
Создание смысла: ищем смысл в данных	Понимание
Коммуникация: передача информации другим	Принятие информированных решений

Грамотная визуализация, которая учитывает психологию восприятия помогает решить эти задачи эффективно.

1. Легко и быстро сравнивать значения

Sales

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493
International	574	636	673	593	644	679	593	139	599	583	602	690
	\$2,557	\$2,979	\$3,266	\$2,876	\$3,218	\$3,517	\$2,975	\$2,773	\$3,537	\$3,322	\$3,585	\$4,183

1. Легко и быстро сравнивать значения

Sales

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493
International	574	636	673	593	644	679	593	139	599	583	602	690
	\$2,557	\$2,979	\$3,266	\$2,876	\$3,218	\$3,517	\$2,975	\$2,773	\$3,537	\$3,322	\$3,585	\$4,183

Sales

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493
International	574	636	673	593	644	679	593	139	599	583	602	690
	\$2,557	\$2,979	\$3,266	\$2,876	\$3,218	\$3,517	\$2,975	\$2,773	\$3,537	\$3,322	\$3,585	\$4,183

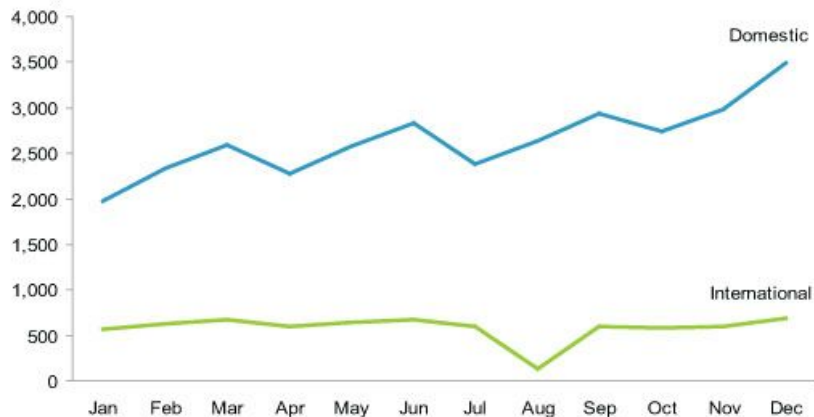
1. Легко и быстро сравнивать значения

Sales

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Domestic	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493
International	574	636	673	593	644	679	593	139	599	583	602	690
	\$2,557	\$2,979	\$3,266	\$2,876	\$3,218	\$3,517	\$2,975	\$2,773	\$3,537	\$3,322	\$3,585	\$4,183

U.S. Dollars
(thousands)

Sales



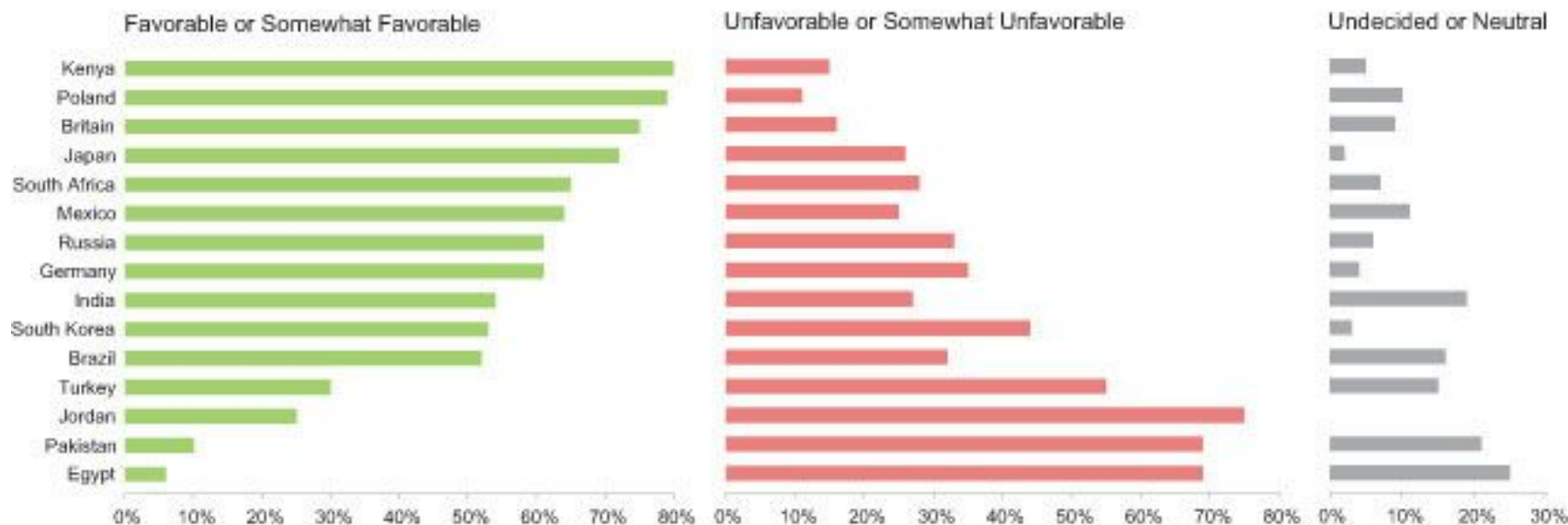
2. Видеть общую картину

Favorable or Unfavorable View of the U.S.

Brazil: % with somewhat or very favorable opinion of the U.S.:	52%
Brazil: % with somewhat or very unfavorable opinion of the U.S.:	32%
Mexico: % with somewhat or very favorable opinion of the U.S.:	64%
Mexico: % with somewhat or very unfavorable opinion of the U.S.:	25%
Britain: % with somewhat or very favorable opinion of the U.S.:	75%
Britain: % with somewhat or very unfavorable opinion of the U.S.:	16%
Germany: % with somewhat or very favorable opinion of the U.S.:	61%
Germany: % with somewhat or very unfavorable opinion of the U.S.:	35%
Russia: % with somewhat or very favorable opinion of the U.S.:	61%
Russia: % with somewhat or very unfavorable opinion of the U.S.:	33%
Poland: % with somewhat or very favorable opinion of the U.S.:	79 %
Poland: % with somewhat or very unfavorable opinion of the U.S.:	11%
South Africa: % with somewhat or very favorable opinion of the U.S.:	65%
South Africa: % with somewhat or very unfavorable opinion of the U.S.:	28%
Kenya: % with somewhat or very favorable opinion of the U.S.:	80%
Kenya: % with somewhat or very unfavorable opinion of the U.S.:	15%
India: % with somewhat or very favorable opinion of the U.S.:	54%
India: % with somewhat or very unfavorable opinion of the U.S.:	27%
Japan: % with somewhat or very favorable opinion of the U.S.:	72%
Japan: % with somewhat or very unfavorable opinion of the U.S.:	26%
South Korea: % with somewhat or very favorable opinion of the U.S.:	53%
South Korea: % with somewhat or very unfavorable opinion of the U.S.:	44%
Egypt: % with somewhat or very favorable opinion of the U.S.:	6%
Egypt: % with somewhat or very unfavorable opinion of the U.S.:	69%
Pakistan: % with somewhat or very favorable opinion of the U.S.:	10%
Pakistan: % with somewhat or very unfavorable opinion of the U.S.:	69%
Turkey: % with somewhat or very favorable opinion of the U.S.:	30%
Turkey: % with somewhat or very unfavorable opinion of the U.S.:	55%
Jordan: % with somewhat or very favorable opinion of the U.S.:	25%
Jordan: % with somewhat or very unfavorable opinion of the U.S.:	75%

2. Видеть общую картину

Current World Opinions About the U.S.A



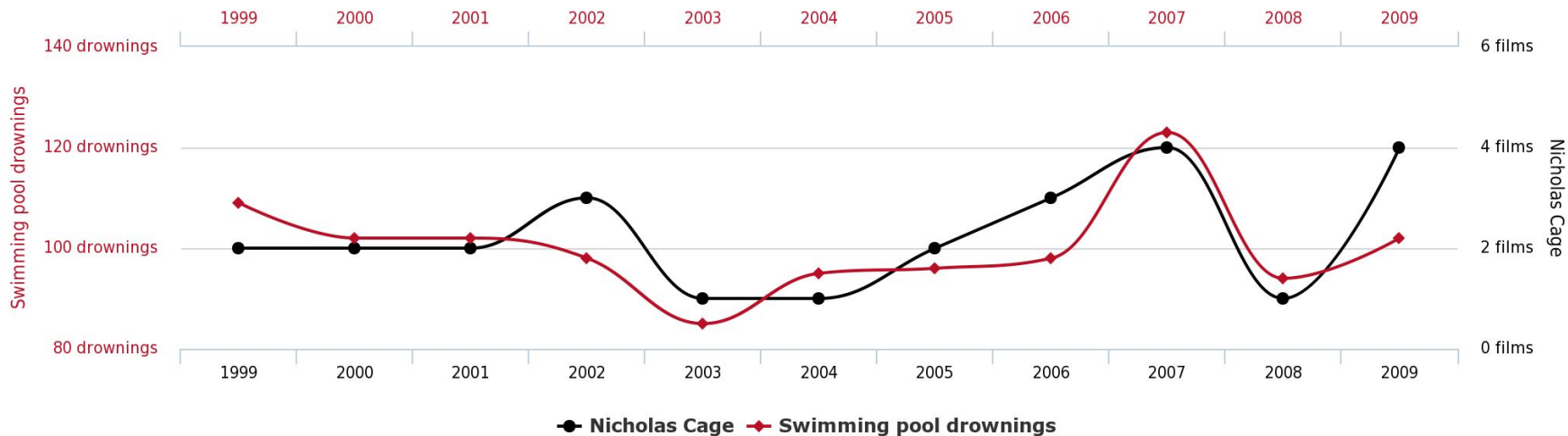
Source: 2004 study conducted by the Pew Research Center, as reported by the PBS television program NOW.

3. Видеть связи в данных

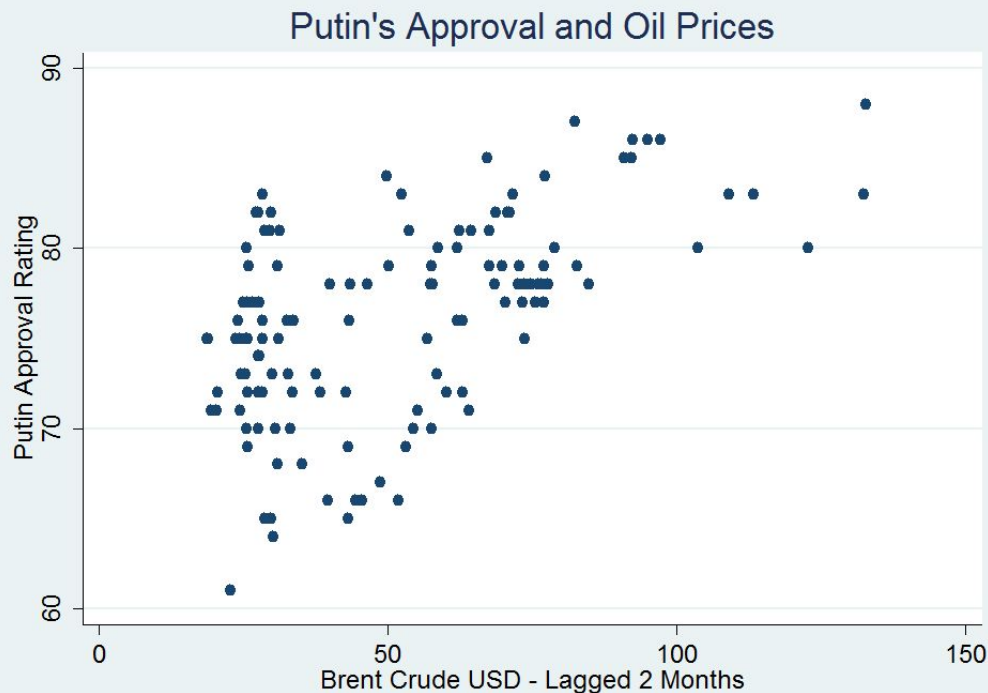
Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



3. А теперь серьезно – видеть связи в данных



<http://causalloop.blogspot.com/2012/01/domestic-crisis-politics-of-russian.html>

Мифы о визуализации

- Визуализируем данные только потому, что есть некоторые люди, которые лучше визуально воспринимают информацию.
- Визуализируем данные, потому что есть люди, у которых сложности с математикой и числами.

Нет. Просто некоторые виды информации лучше всего воспринимаются визуально всеми людьми.

- Визуализируем, чтобы привлечь внимание, но при этом жертвуем информативностью.

Нет, если информация не принесена в жертву красоте.

- Графика - лучший способ рассказать историю.

Необязательно. Многие истории все еще прекрасно чувствуют себя, написанными словами. Графика может быть бэкграундом.

Что делать, если я хочу развивать критическое мышление?

Бесплатно

- Читать N+1 (можно фильтровать по исследованиям в социальных науках, например).
<https://nplus1.ru>
- Или смотреть/слушать проект N+1 и “Эхо Москвы” “На пальцах”
- Слушать подкасты про статистику и исследования
Freakonomics
<http://freakonomics.com/>

More or Less
<https://www.bbc.co.uk/programmes/b006qshd>

Planet Money
<https://www.npr.org/sections/money/>
- Учить статистику (например, для поведенческих наук)
- Находить как можно больше плохих примеров и уметь их анализировать

Где искать данные и что с ними делать?

Смотрим примеры

Визуализации, сделанные на
открытых данных

Eurozone Debt Web

<https://www.bbc.co.uk/news/business-15748696>

Бюджетная информация

<https://offenerhaushalt.de>

Анализ данных о голосовании

<http://trv-science.ru/2011/12/20/matematika-vyborov-2011/>

Gapminder

[https://www.gapminder.org/tools/#\\$chart-type=bubbles](https://www.gapminder.org/tools/#$chart-type=bubbles)

Где искать данные и что с ними делать?

Смотрим примеры

Визуализации, сделанные на данных,
собранных из интернета

Данные из системы Scopus

<http://econrank.sk/>

Twitter

<https://onemilliontweetmap.com>

Анализ сети актеров

<https://katie-truong.github.io/network-analysis/IMDB/>

Где искать данные и что с ними делать?

Смотрим примеры

Готовые инструменты

Гугл-тренды

<https://trends.google.com/trends/?geo=US>

World Values Survey Online Tool

<http://www.worldvaluessurvey.org/WVSONline.jsp>

Главные вопросы при работе с данными

Про все эти вещи поговорим на
следующей лекции

- Как были собраны данные?
- Как была сделана выборка?
- Как операционализованы переменные?
- Как интерпретировать результаты?