Анализ данных и машинное обучение

Бернгардт Олег Игоревич
E-mail: berng@rambler.ru

Рекомендуемая литература

Основная:

1. Конспект лекций: https://github.com/berng/DAcourseBook
2. Часть лекций Малого ШАД: https://www.youtube.com/watch?v=c7y4G3eOEGg&list=PLO18kp5vWvhj79nJhWGEmt3vJZI_-y7A1

Дополнительная:

1. Маккинли У., Python и анализ данных // М.: ДМК Пресс, 2015.
2. Грас Дж., Data Science. Наука о данных с нуля//BHV, 2020, 416с
3. О'Нил К., Шатт Р., Data Science. Инсайдерская информация для новичков. Включая язык R. //Питер, 2019, 368с
4. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/ https://doi.org/10.18434/M32189
5. Бендат Дж., Пирсол А., Прикладной анализ случайных данных//М., Мир, 1989, 540с.
6. Hamilton, J.D. Time Series Analysis// Princeton University Press, 1994, 799c, https://doi.org/10.2307/j.ctv14jx6sm
7. Залманзон Л.А., Преобразования Фурье, Уолша, Хаара и их применение в управлении связи и других областях//М.,Наука, 1989. 496 с.
8. Ghojogh B., Crowley M., Karray F., Ghodsi A., Elements of Dimensionality Reduction and Manifold Learning //Springer, 2023, 606c, https://doi.org/10.1007/978-3-031-10602-6
9. Воронцов К. В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. 2024. http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Дополнительная литература

10. Chan S.H., Introduction to Probability for Data Science//Michigan Publishing, 2021, 690c
11. Bruce P., Bruce A., and Gedeck P., Practical Statistics for Data Scientists//O'Reilly, 2020, 342c
12. James G., Witten D., Hastie T., Tibshirani R., Taylor J., An Introduction to Statistical Learning with Applications in Python//Springer, 2023, 607c
13. Nield T., Essential Math for Data Science//O'Reilly, 2022, 332p.
14. Гонсалес Д., Лау С., Нолан Д., Изучаем Data Science: обработка, исследование, визуализация и моделирование данных с помощью Python//BHV, 2025, 560c (https://github.com/DS-100/textbook)

Аналитик данных

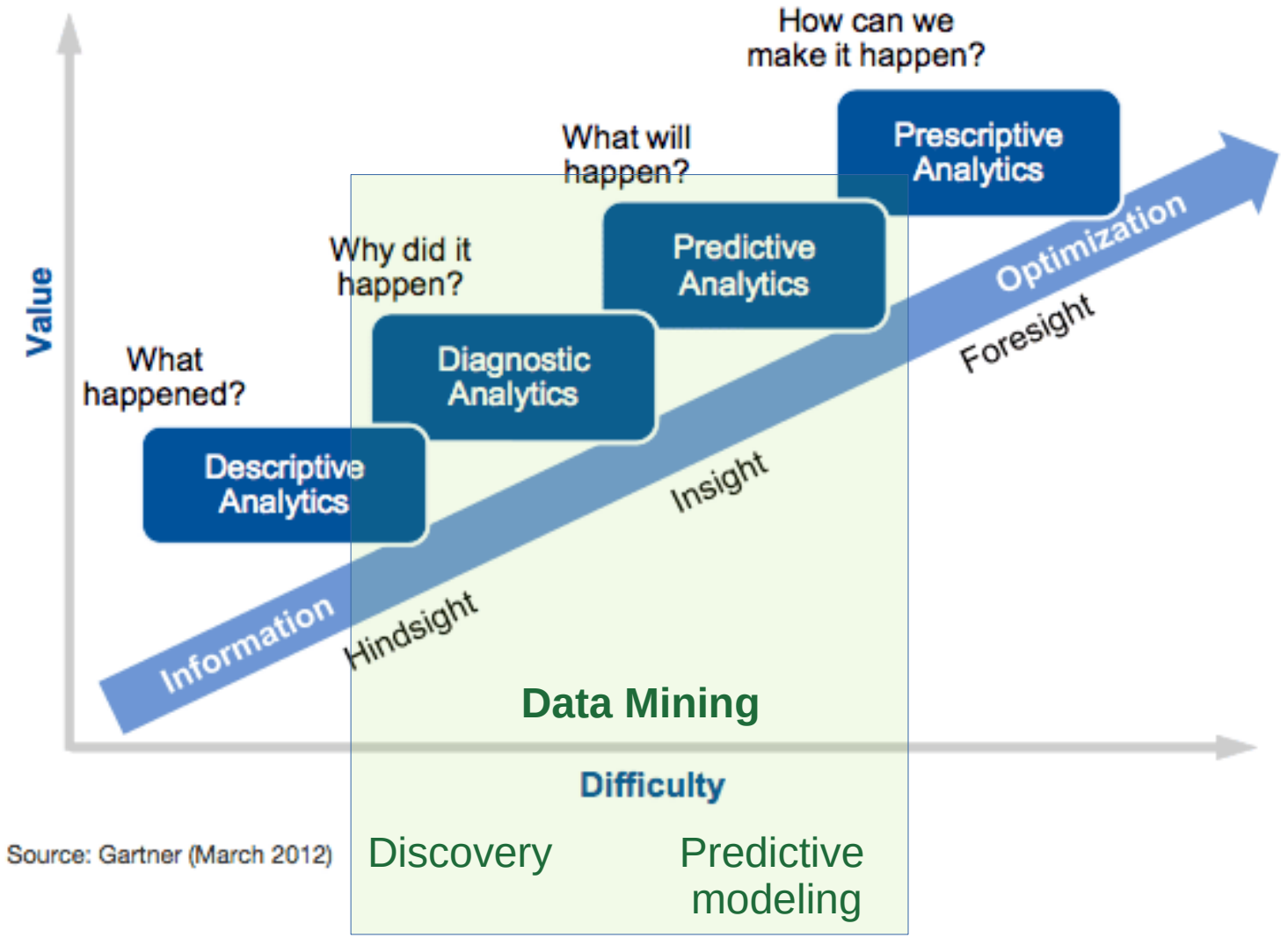
Аналитик данных - это тот, кто извлекает смысл из беспорядочных данных.

Аналитик данных должен иметь навыки в следующих областях:

- **Экспертиза в предметной области** - чтобы анализировать данные и делать выводы, относящиеся к их рабочему месту, аналитику необходимо обладать экспертными знаниями в этой предметной области.
- **Навыки программирования -** необходимо знать, какие программные продукты, языки программирования и библиотеки существуют обработки данных и получения из них аналитической информации.
- **Методы анализа -** необходимо знать, какие методы следует использовать для очистки данных, их обработки и получения из них аналитической информации и для проверки верности своих выводов.
- **Навыки визуализации** - необходимо обладать навыками визуализации данных и результатов, чтобы суммировать и представлять данные третьей стороне.
- **Объяснение результатов** - аналитик должен уметь объяснить свои выводы заинтересованному лицу или клиенту.

Виды анализа данных

Figure 2. Gartner Analytic Ascendancy Model



Виды анализа данных:

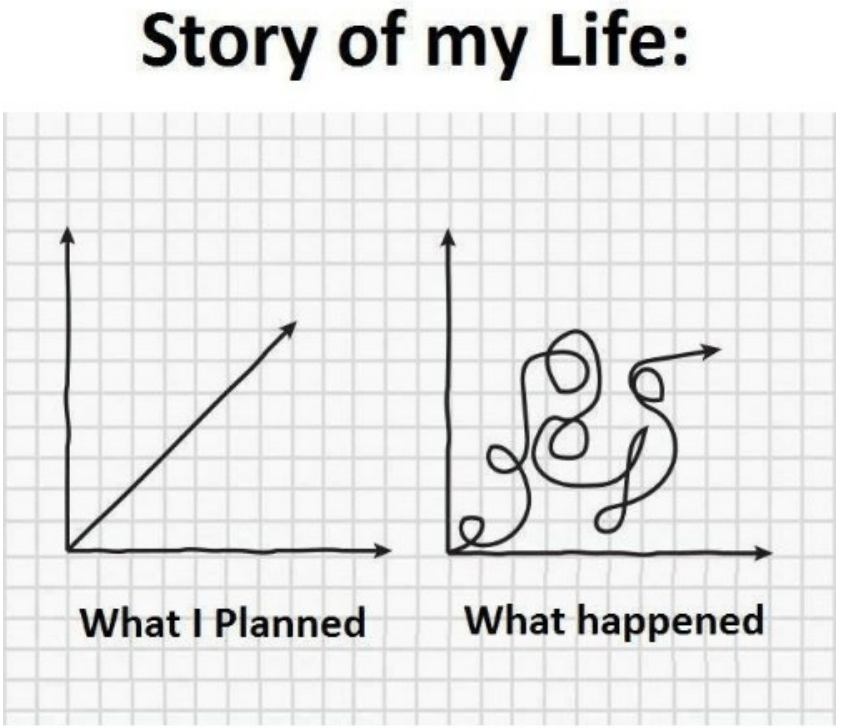
Описательный Диагностический Прогностический Рекомендательный

https://link.springer.com/article/10.1007/s00778-019-00539-

Описательный анализ

Описательный анализ - это самый простой и распространенный тип аналитики, который используют компании. Он суммирует и выделяет закономерности в текущих и исторических данных.

Описательная аналитика используется для визуализации ключевых показателей, которая позволяют отслеживать их тенденции. Описательная аналитика помогает понять, что происходит с данными в конкретный момент.



Обычно включает в себя агрегирование, первичный анализ и визуализацию данных.

Агрегирование данных – сбор, сортировка и форматрирование данных для упрощения их дальнейшего анализа.

Первичный анализ данных включает в себя определение типа данных, пределов их изменений, выявление ошибок в данных, выявление их качественного поведения во времени и пространстве (рост, убывание) и выявление качественных и количественных соотношений между различными частями данных.

Одним из способов представления результатов описательной аналитики является **визуализация данных**.

Диагностический анализ

Диагностический анализ обеспечивает более глубокий анализ взаимоотношений между данными и их динамики, чтобы ответить на вопрос: почему это произошло?

Часто диагностический анализ называют анализом причин.

Он включает использование таких подходов, как статистический и корреляционный анализ, спектральный анализ, регрессионный анализ.

При диагностическом анализе выявляются зависмости между различными данными, что и помогает найти ключевые причины наблюдаемых явлений, и ключевые параметры, ответственные за наблюдаемые изменения.

При этом проводится проверка различных гипотез и отбор наиболее правдоподобных из них

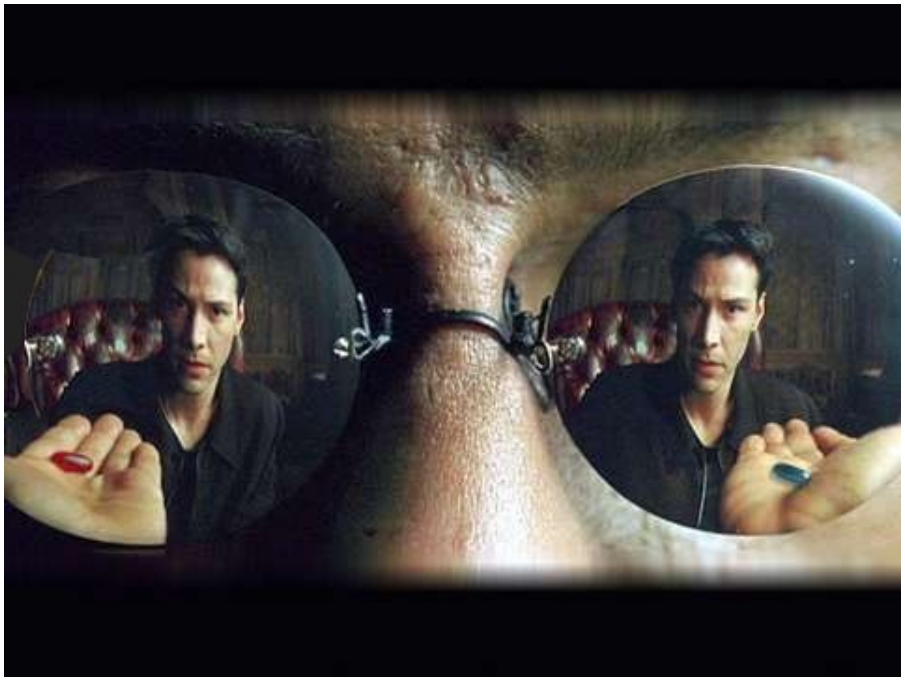
Прогностический анализ

Прогностическая аналитика – создание эффективной модели процессов, учитывающей реальные данные. Так называемое решение прямой задачи – прогноз ключевой величины по историческим и/или получаемым в текущем времени экспериментальным данным. Используется, чтобы предсказать, что произойдет дальше по известным значениям параметров.

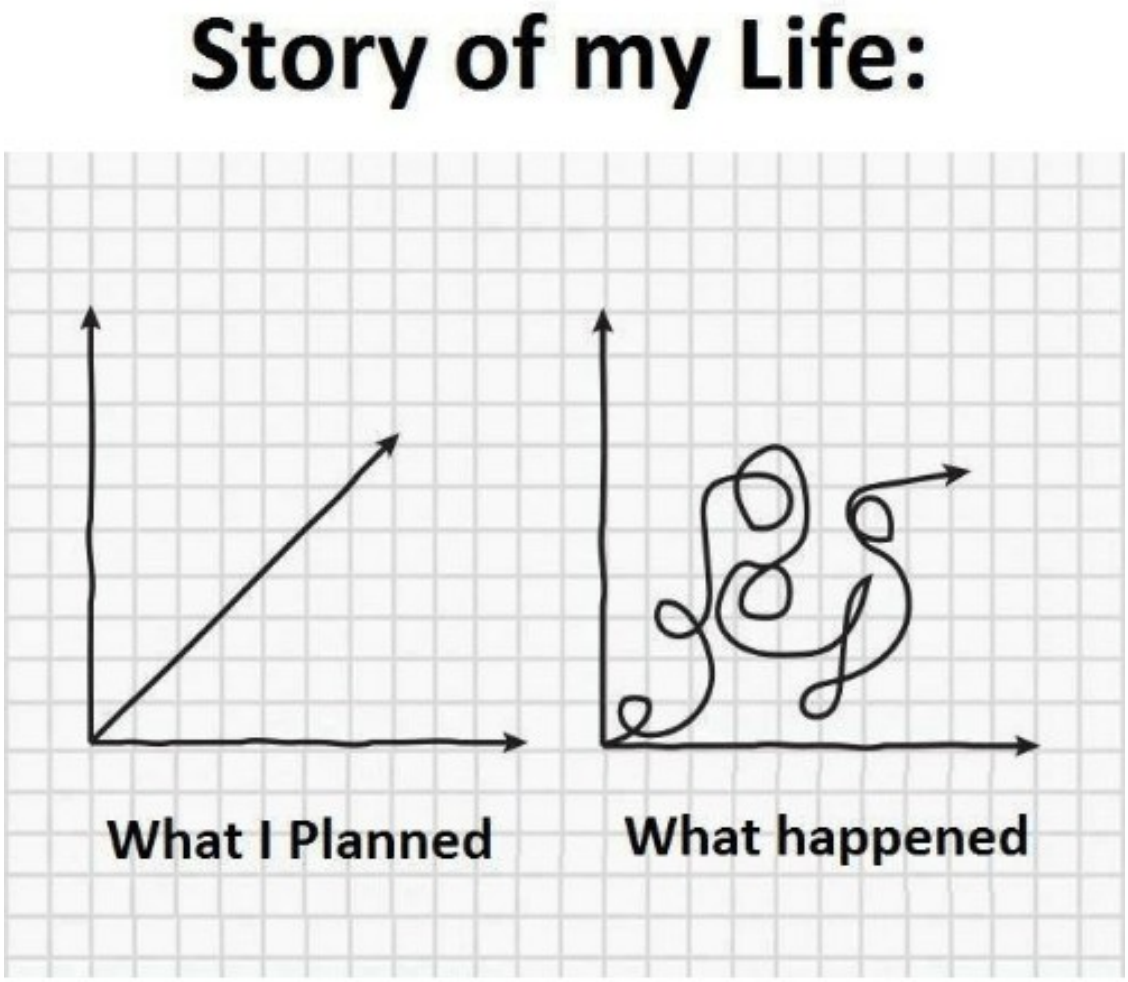
Обычно для этого используются теоретические, эмпирические, математические или обучающиеся модели, наиболее адекватно описывающие существующие данные. В последнее время широко применяются модели машинного обучения типа нейронных сетей.

Рекомендательный анализ

Рекомендательный анализ позволяет дать рекомендации, что можно изменить в текущих процессах, чтобы получить необходимое вам значение ключевого параметра в будущем. Он предлагает различные варианты действий и описывает возможные последствия для каждого из них. В математике часто называется решением обратной задачи – по ожидаемому значению ключевого параметра определить, какими должны быть исходные значения параметров.



ОПИСАТЕЛЬНЫЙ АНАЛИЗ



Основные шаги описательного анализа

Определение метрик: определить ключевые параметры (метрики), которые вы хотите изучать.

Определение и поиск необходимых данных: найти данные, необходимые для получения (расчета) желаемых параметров (метрик).

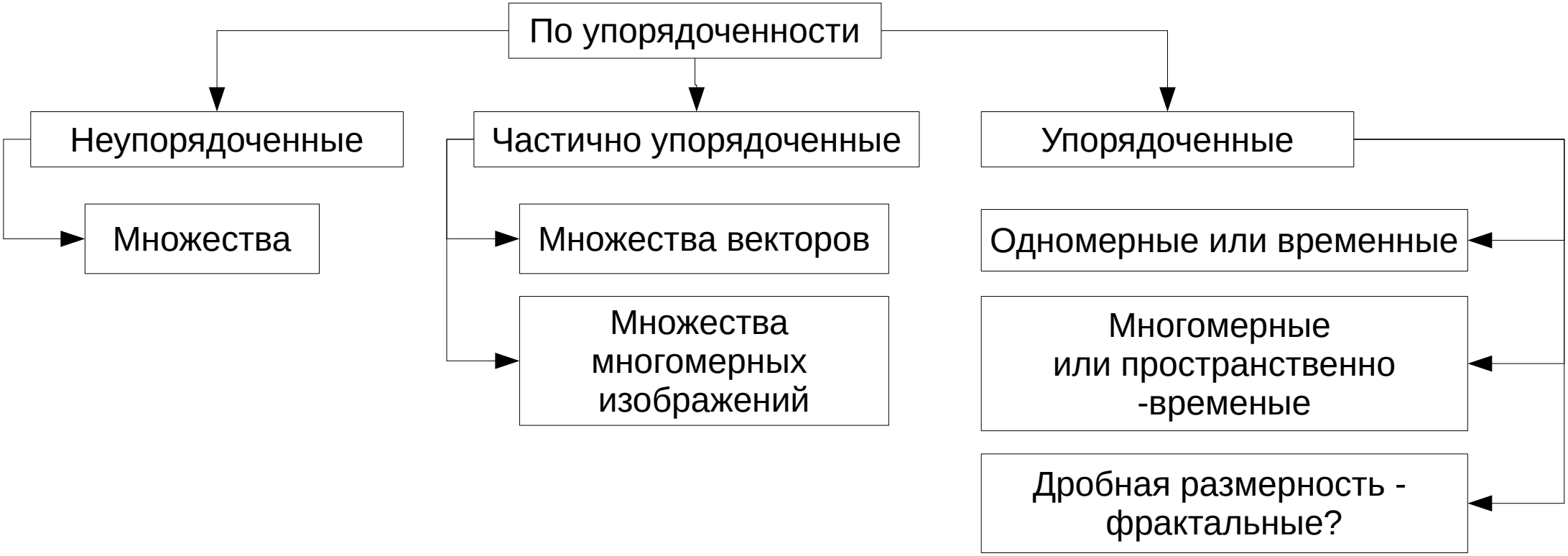
Извлечение и подготовка данных(переменных): если данные поступают из нескольких источников, извлечение, объединение и подготовка данных для анализа является трудоемким процессом. Этот шаг может включать в себя очистку данных для устранения несоответствий и ошибок в данных из разных источников, а также преобразование данных в формат, подходящий для инструментов анализа.

Анализ данных: описательная аналитика часто включает применение основных математических операций к одной или нескольким переменным, а тажке расчеты по ним необходимых метрик.

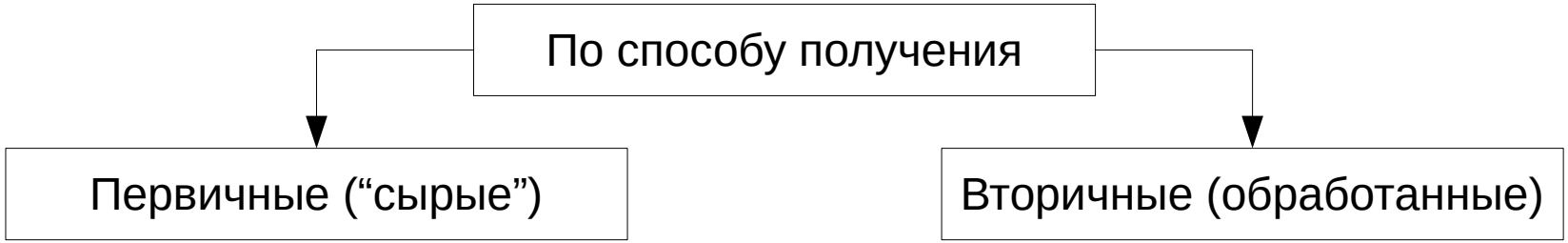
Представление данных: представление данных в понятных визуальных формах часто упрощает понимание заинтересованными сторонами.

ТИПЫ ДАННЫХ

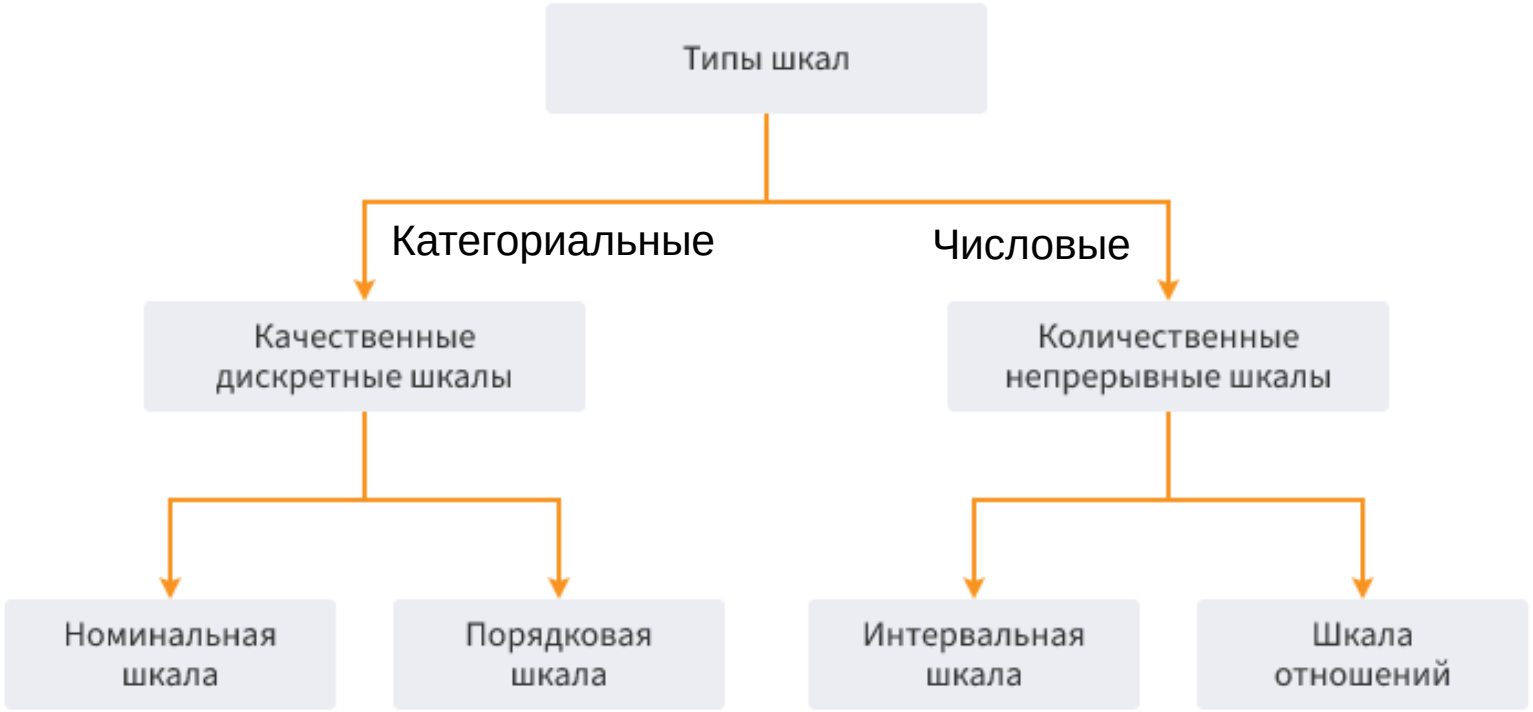
Классификация данных по мере упорядоченности



Классификация по способу получения



Классификация по типу шкалы



Свойства \ Тип шкалы	Номинальная	Порядковая	Интервальная	Отношений
Идентифицируемость	x	x	x	x
Величина (магнитуда)		x	x	x
Равенство интервалов			x	x
Абсолютный ноль				x

https://wiki.loginom.ru/articles/scale-

Качественные дискретные шкалы

Номинальная шкала (категориальная, наименований) — это шкала измерения, которая используется для идентификации. Она является самой «слабой» из четырех видов шкал в смысле возможности обработки данных. Она присваивает номера атрибутам для удобства идентификации, но может использоваться только как метка. Единственный вид статистического анализа, который можно выполнить с использованием номинальной шкалы, это вычисление процентных долей и частот. Данные в номинальной шкале можно проанализировать графически с помощью гистограммы и круговой диаграммы. Например, если измерить атрибут «Товар» в номинальной шкале, то она будет выглядеть так: 1 — мороженное; 2 — соки; 4 — выпечка. При этом значения шкалы не определяют какого-либо приоритета между товарами, а просто идентифицируют их. Очевидно, что такая шкала может использоваться только для самого просто анализа.

Порядковая шкала (ординальная, ранговая) — предполагает ранжирование (упорядочивание) значений переменной в зависимости от масштабирования. Атрибуты в порядковой шкале обычно располагаются в порядке возрастания или убывания. Порядковая шкала может быть использована в исследованиях рынка, рекламы и опросов удовлетворенности клиентов. Она использует квалификаторы, такие как «очень», «высоко», «больше», «меньше» и т. д. В порядковой шкале можно использовать для статистического анализа такие статистики как медиана, но не среднее значение. Существуют и другие виды анализа, которые могут быть проведены с использованием порядковой шкалы. Например, компания-разработчик ПО может провести опрос пользователей для оценки нового приложения в шкале: «Отлично», «Очень хорошо», «Хорошо», «Плохо», «Очень плохо». Атрибуты в этом примере перечислены в порядке убывания.

Количественные непрерывные шкалы

Интервальная шкала (разностей) — это шкала, в которой уровни упорядочены, а интервалы между ними равны. Её можно рассматривать как расширение порядковой шкалы. Основным отличием является свойство равных интервалов. Интервальная шкала не только позволяет однозначно определить, какое значение больше (меньше), но и на сколько. Кроме того, в отличие от порядковой и номинальной шкал, в интервальной могут выполняться арифметические операции. Типичным примером является измерение температуры по шкале Фаренгейта. Интервальную шкалу можно использовать при расчете среднего значения, медианы, моды, стандартного отклонения и других статистик.

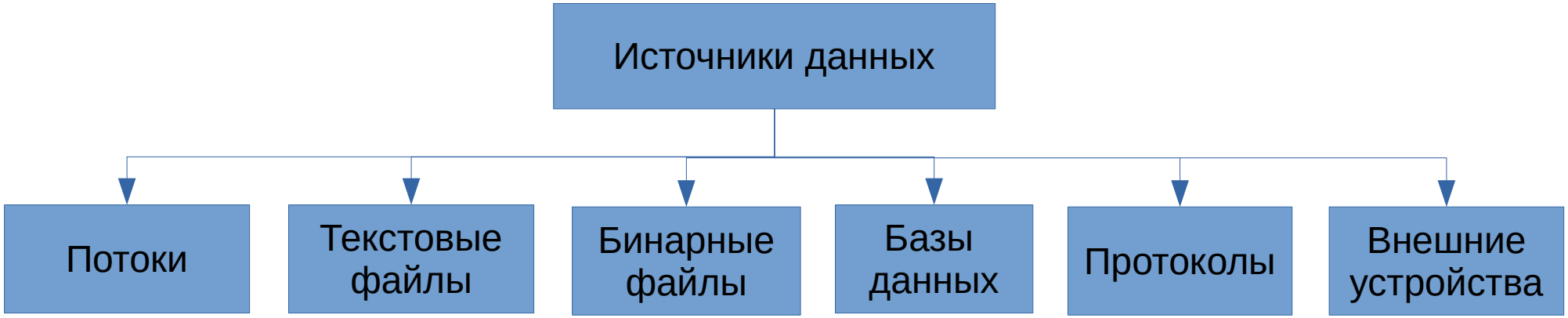
Шкала отношений (абсолютная) является «наивысшим» уровнем представления данных. Она может рассматриваться как расширение интервальной шкалы, и следовательно, удовлетворяет четырем свойствам шкалы измерения: идентифицируемостью, величиной, равноинтервальностью и наличием абсолютного нуля. Примерами шкал отношения являются длина, вес, время и т. д. В исследованиях рынка примерами шкалы отношений являются цена, количество клиентов, суммы продаж и т. д. Она широко используется в маркетинге и рекламе. Шкала отношений совместима со всеми методами статистического анализа и может использовать как показатели центральной тенденции (среднее значение, медиана, мода и т. д.), так и разброса значения (дисперсии, размаха, стандартного отклонения и т. д.).

Задание: Определить типы шкал

Пол	1 = мужской
	2 = женский
Семейное положение	1 = холост/не замужем
	2 = женат/замужем
	3 = вдовец/вдова
	4 = разведен(а)
Курение	1 = некурящий
	2 = изредка курящий
	3 = интенсивно курящий
	4 = очень интенсивно курящий
Месячный доход	1 = до 3000 DM
	2 = 3001 - 5000 DM
	3 = более 5000 DM
Коэффициент интеллекта (I.Q.)	
Возраст (лет)	

БАЗОВЫЕ МЕТОДЫ АГРЕГАЦИИ

АГРЕГАЦИЯ НА ПИТОН




```
import sys
f=sys.stdin
for str1 in f:
 print(str1)
```

Вызов:

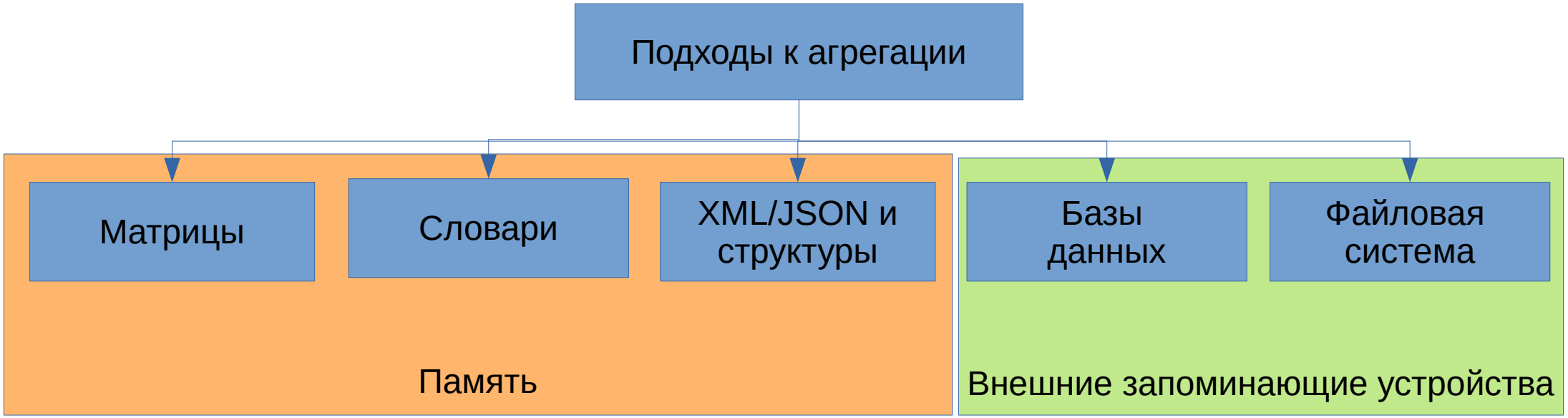
#python3 a.py < filename.txt

```
f=open('filename.txt')
for str1 in f:
 print(str1)
```

Вызов:

#python3 a.py

АГРЕГАЦИЯ НА ПИТОН



Через словарь

```
import sys
f=open('file1.txt')
DICT=dict()
for i in f:
 d=i.split()
 DICT[d[0]]=i
f.close()
f=open('file2.txt')
DICT2=dict()
for i in f:
 d=i.split()
 DICT2[d[0]]=i
f.close()
```

```
for i in DICT:
 if i in DICT2:
  print(i,':',DICT[i],DICT2[i])
```

Через матрицу

```
import sys
import numpy as np
f=open('file1.txt')
DICT=np.zeros(5)
for i in f:
 d=i.split()
 DICT[int(d[0])]=int(d[1])
f.close()
f=open('file2.txt')
DICT2=np.zeros(5)
for i in f:
 d=i.split()
 DICT2[int(d[0])]=int(d[1])
f.close()
```

```
for i in range(5):
  print(i,':',DICT[i],DICT2[i])
```

ПРИВЕДЕНИЕ К ОДИНАКОВОМУ МАСШТАБУ

Прореживание

```
f=open('f1.txt')
D=dict()
for i in f:
 d=i.split()
 D[str((int(float(d[0])*100))/100)]=i
f.close()

f=open('f2.txt')
D2=dict()
for i in f:
 d=i.split()
 D2[str(d[0])]=i
f.close()

for i in D:
 if i in D2:
  print('fnd',i,D[i],D2[i])
```

Интерполяция

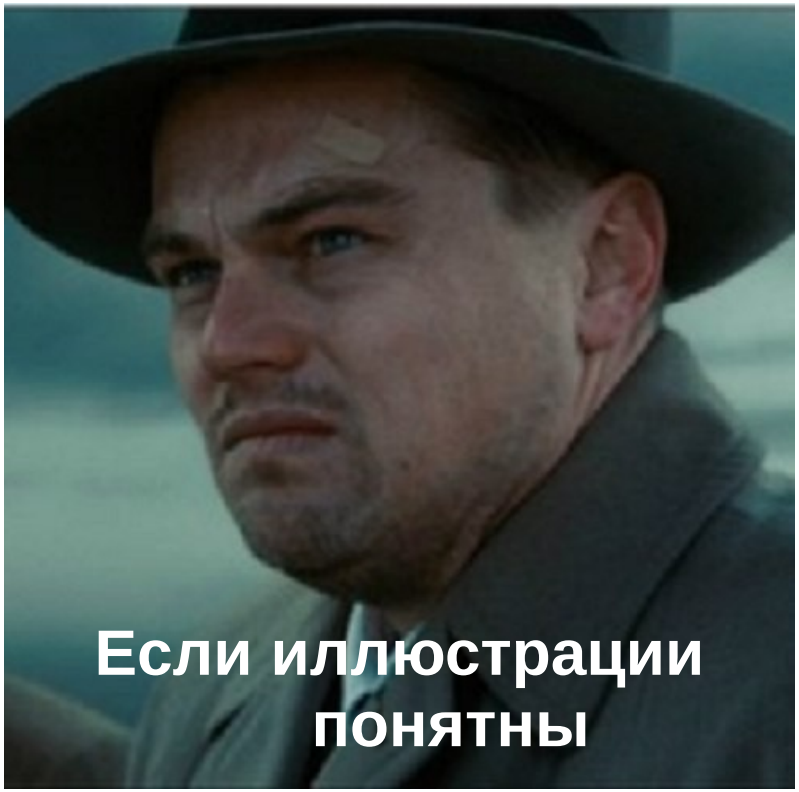
```
f=open('f1.txt')
D=dict()
for i in f:
 d=i.split()
 D[str(float(d[0]))]=i
f.close()

f=open('f2.txt')
D2=dict()
for i in f:
 d=i.split()
 D2[str(d[0])]=i
f.close()

for i in D:
 idx2=str((int(float(i)*100))/100)
 if idx2 in D2:
  print(D[i],D2[idx2])
```

МЕТОДЫ ВИЗУАЛИЗАЦИИ

В 4 раза лучше человек выполняет инструкцию, если она содержит иллюстрации.



Визуализация данных

90% информации человек воспринимает через зрение;

70% сенсорных рецепторов находятся в глазах,

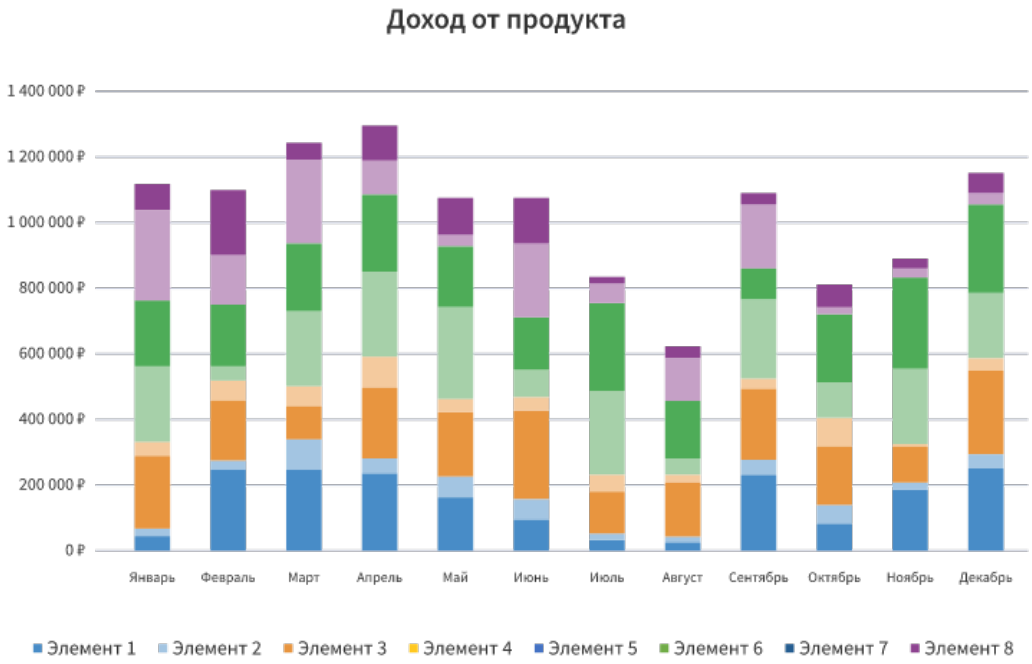
около половины нейронов головного мозга человека задействованы в обработке визуальной информации;

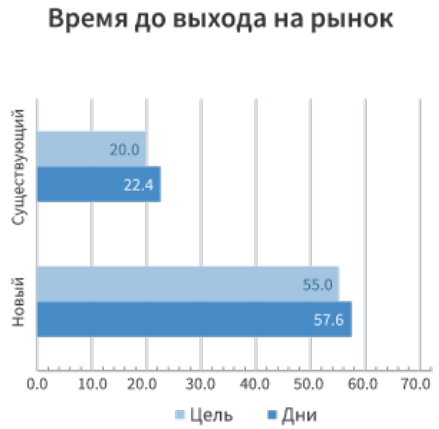
- Общая
- Специальная
- Концептуальная
- Стратегическая
- Метафорическая
- Комбинированная

https://yvision.kz/post/

Общая визуализация

Общая визуализация - представление количественной информации в традиционнй схематической форме. Круговые и линейные диаграммы, гистограммы и спектрограммы, таблицы и различные точечные графики.



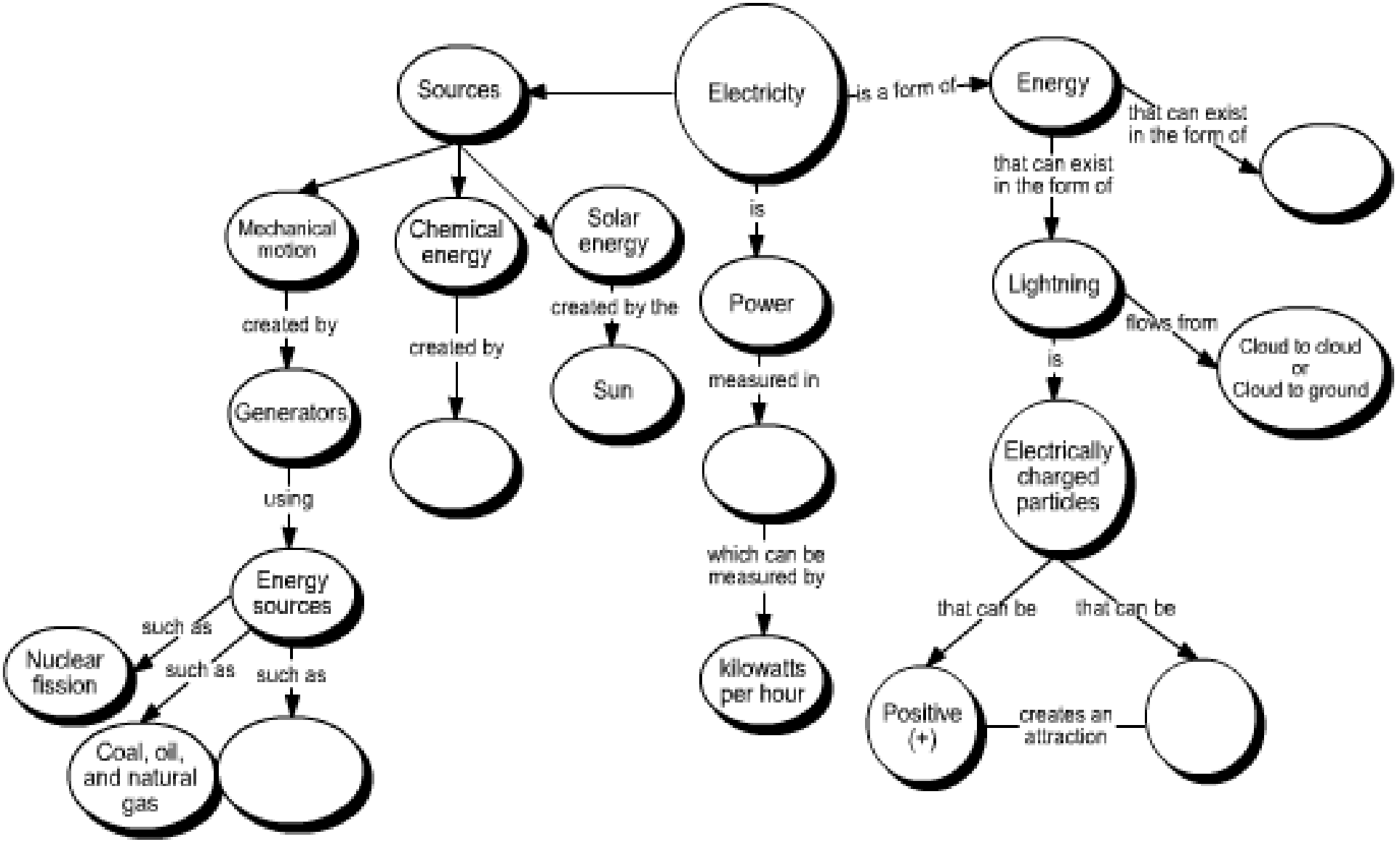


Специальная визуализация

Специальная визуализация — специфические формы представления информации - карты и полярные графики, графики с параллельными осями, диаграммы Эйлера и др.

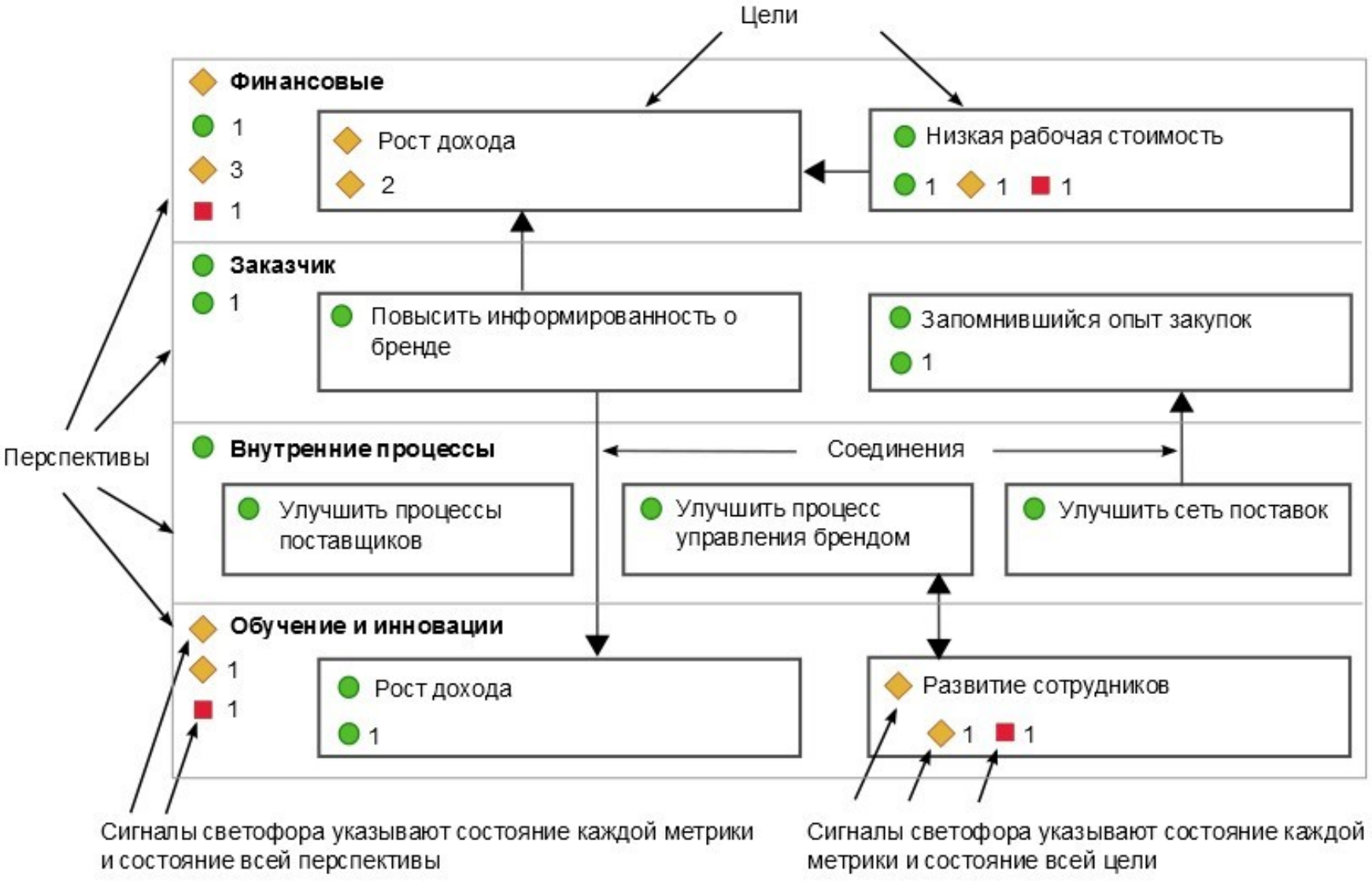
Концептуальная визуализация

Концептуальная визуализация — позволяет разрабатывать сложные концепции, отношения между объектами, идеи и планы с помощью концептуальных карт и других подобных видов диаграмм.



Стратегическая визуализация

Стратегическая визуализация — переводит в визуальную форму различные данные об аспектах работы организаций и бизнес-процессах. Это всевозможные диаграммы производительности, жизненного цикла и графики структур организаций.



https://www.ibm.com/docs/ru/cognos-tm1/10.2.2?topic=scorecarding-strategy-map

Метафорическая визуализация

Метафорическая визуализация — используется для представления информации в виде геометрических фигур и их композиций (например, значения признака представляются кругами разного размера).

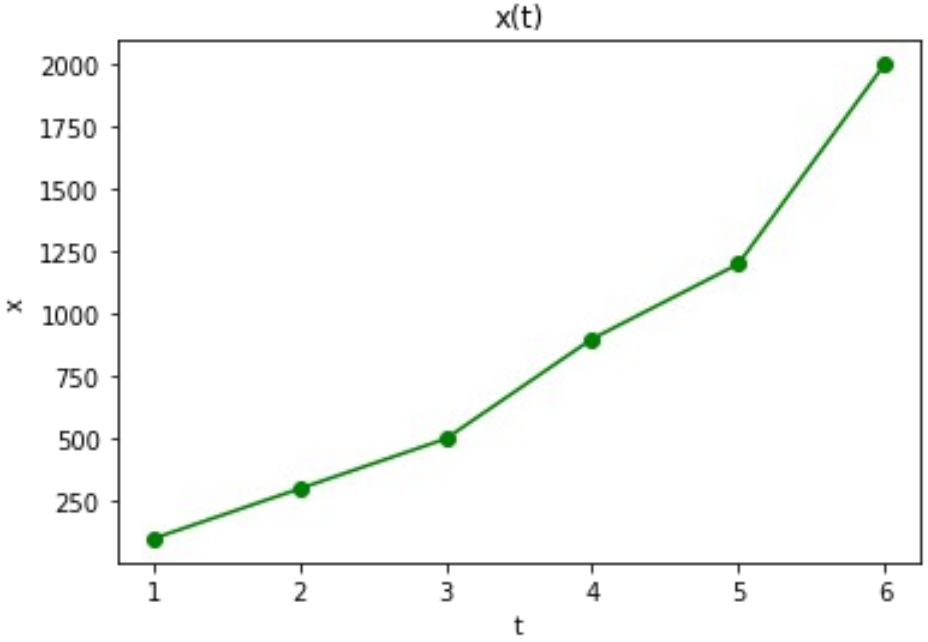
Комбинированная визуализация позволяет объединить несколько сложных представлений в одну схему.

БАЗОВАЯ ВИЗУАЛИЗАЦИЯ НА ПИТОН

Визуализация на Питоне

Линейный график

```
from matplotlib import pyplot as plt
t=[1,2,3,4,5,6]
res=[100,300,500,900,1200,2000]
plt.plot(t,res,color='green',marker='o',linestyle='solid')
plt.title("x(t)")
plt.ylabel("x")
plt.xlabel("t")
```

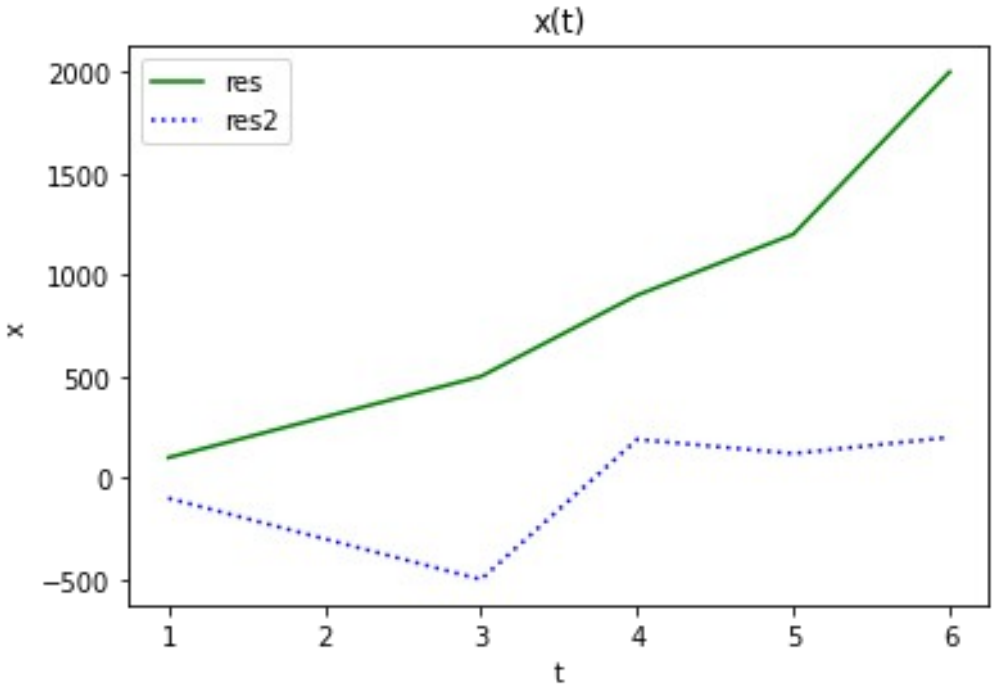


Задание.
Построить 3 линейных графика по файлу t,x,y,z:
x(t), y(t), z(x)

Визуализация на Питоне

Несколько графиков

```
from matplotlib import pyplot as plt
t=[1,2,3,4,5,6]
res=[100,300,500,900,1200,2000]
res2=[-100,-300,-500,190,120,200]
plt.plot(t,res,'g-',label='res')
plt.plot(t,res2,'b:',label='res2')
plt.title("x(t)")
plt.ylabel("x")
plt.xlabel("t")
plt.legend()
```



Задание.
Построить 3 линейных графика по файлу t,x,y,z:

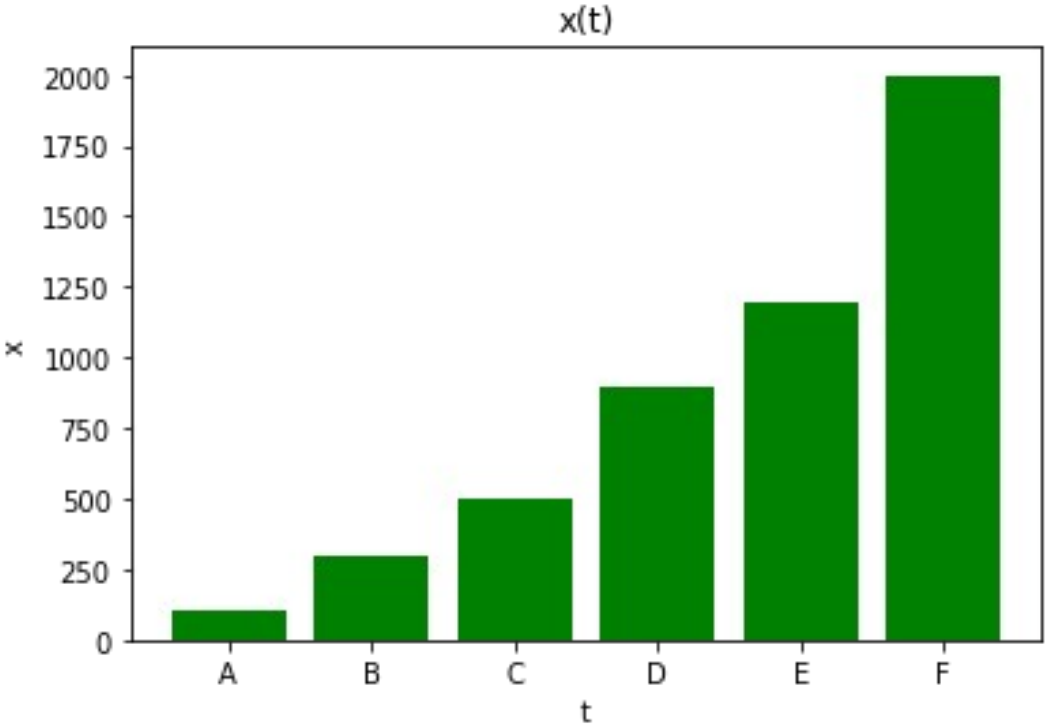
x(t), y(t), z(x)

на одном графике, с подписями

Визуализация на Питоне

Столбчатая диаграмма

```
from matplotlib import pyplot as plt
t=["A","B","C","D","E","F"]
res=[100,300,500,900,1200,2000]
plt.bar(t,res,color='green')
plt.title("x(t)")
plt.ylabel("x")
plt.xlabel("t")
```

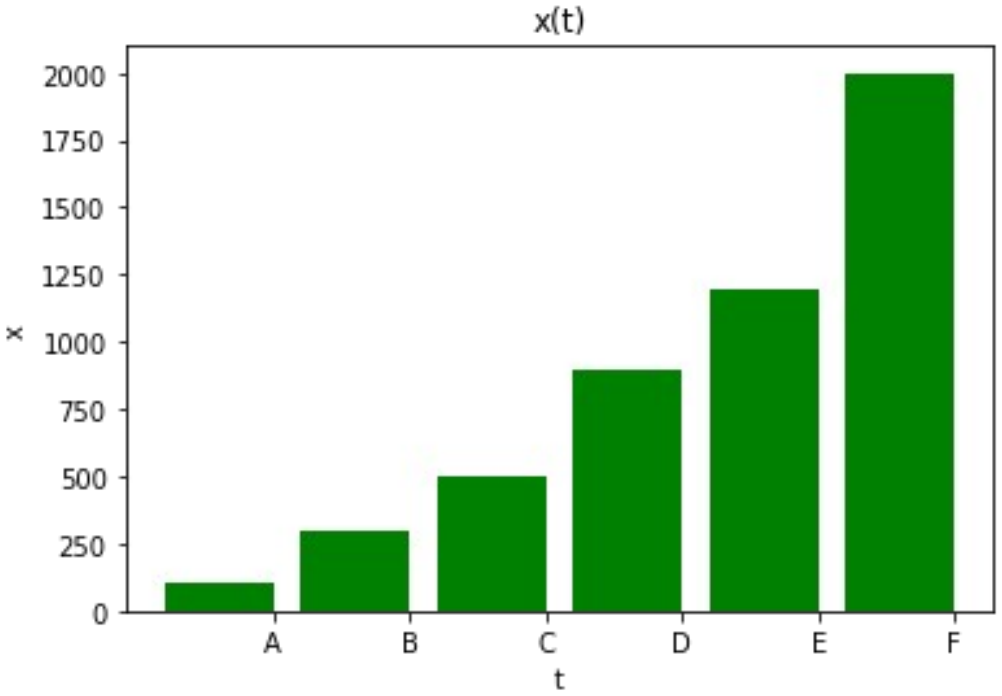


Задание.
Построить 3 столбчатых диаграммы по файлу t,x,y,z:
x(t), y(t), z(x)

Визуализация на Питоне

Сдвижка и переназначение тиков

```
from matplotlib import pyplot as plt
t=["A","B","C","D","E","F"]
res=[100,300,500,900,1200,2000]
xs=[i+0.1 for i,_ in enumerate(t)]
plt.bar(xs,res,color='green')
plt.xticks([i+0.5 for i,_ in enumerate(t)],t)
plt.title("x(t)")
plt.ylabel("x")
plt.xlabel("t")
```



Задание.
Построить столбчатую диаграмму по файлу t,x:

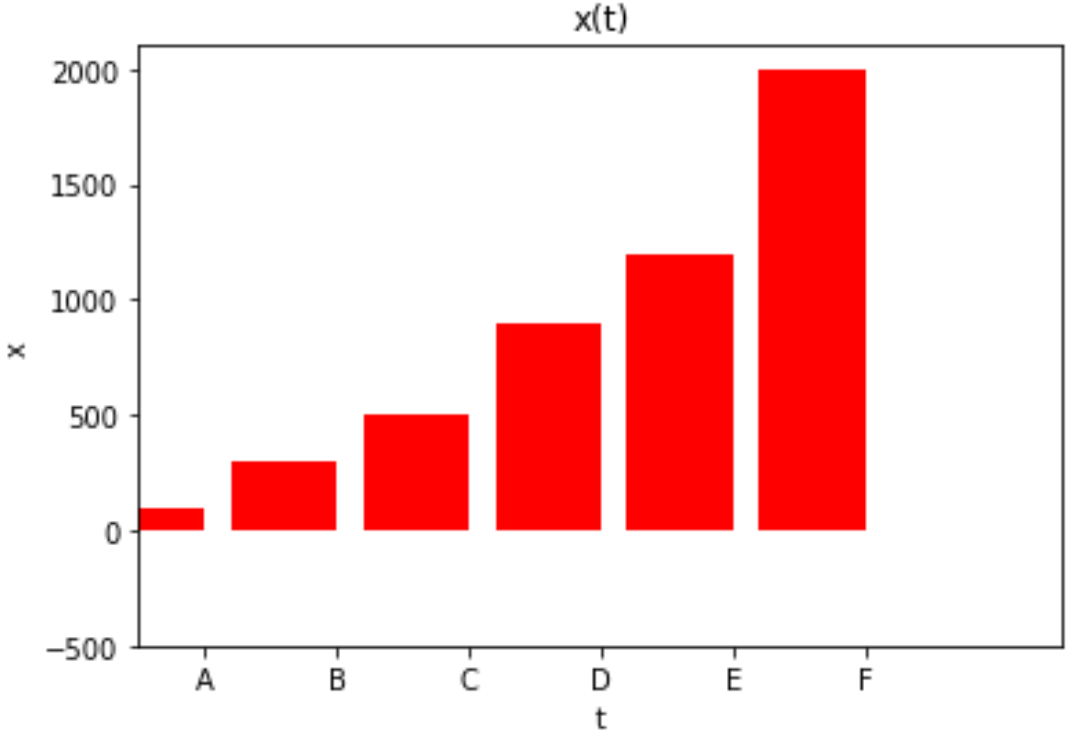
x(t)

Подписи взять из другого файла t,Z

Визуализация на Питоне

Смещение осей

```
from matplotlib import pyplot as plt
t=["A","B","C","D","E","F"]
res=[100,300,500,900,1200,2000]
xs=[i+0.1 for i,_ in enumerate(t)]
plt.bar(xs,res,color='red')
plt.xticks([i+0.5 for i,_ in enumerate(t)],t)
plt.title("x(t)")
plt.axis([0,7,-500,2100])
plt.ylabel("x")
plt.xlabel("t")
```



Задание.
Построить столбчатую диаграмму по файлу t,x:

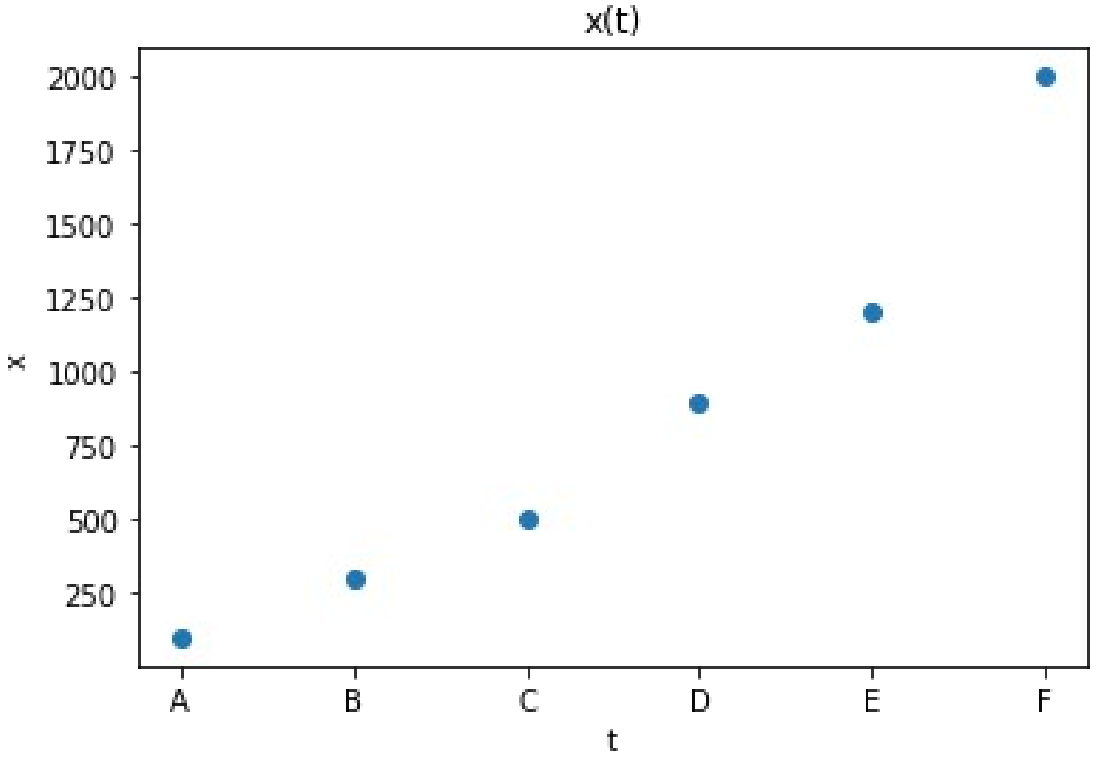
x(t)

Подписи взять из другого файла t,Z

Диаграмма рассеивания (scatter plot)

Scatterplot

```
from matplotlib import pyplot as plt
res=[100,300,500,900,1200,2000]
res2=[-100,-300,-500,190,120,200]
plt.scatter(t,res)
plt.title("x(t)")
plt.ylabel("x")
plt.xlabel("t")
```

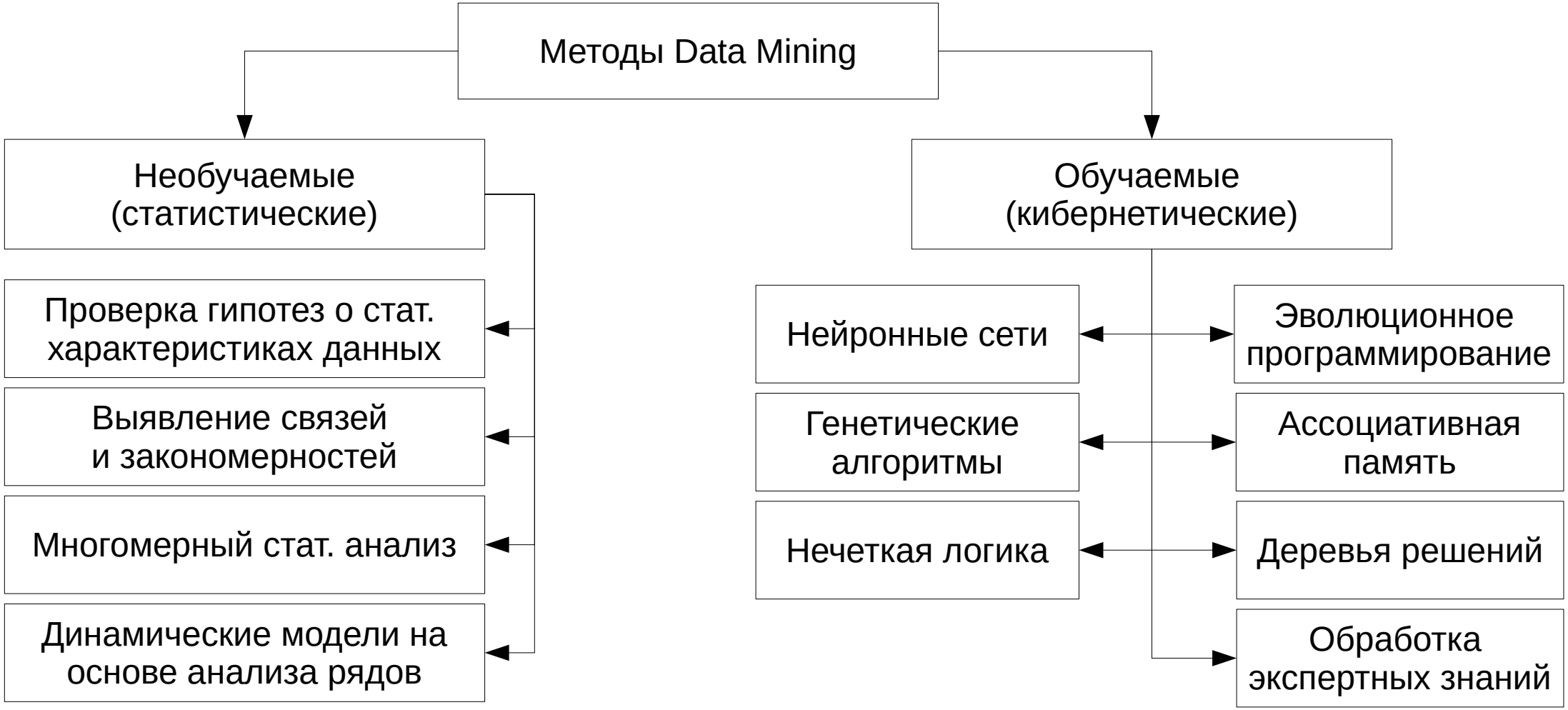


Задание.
Построить scatterplot по файлу t,x:

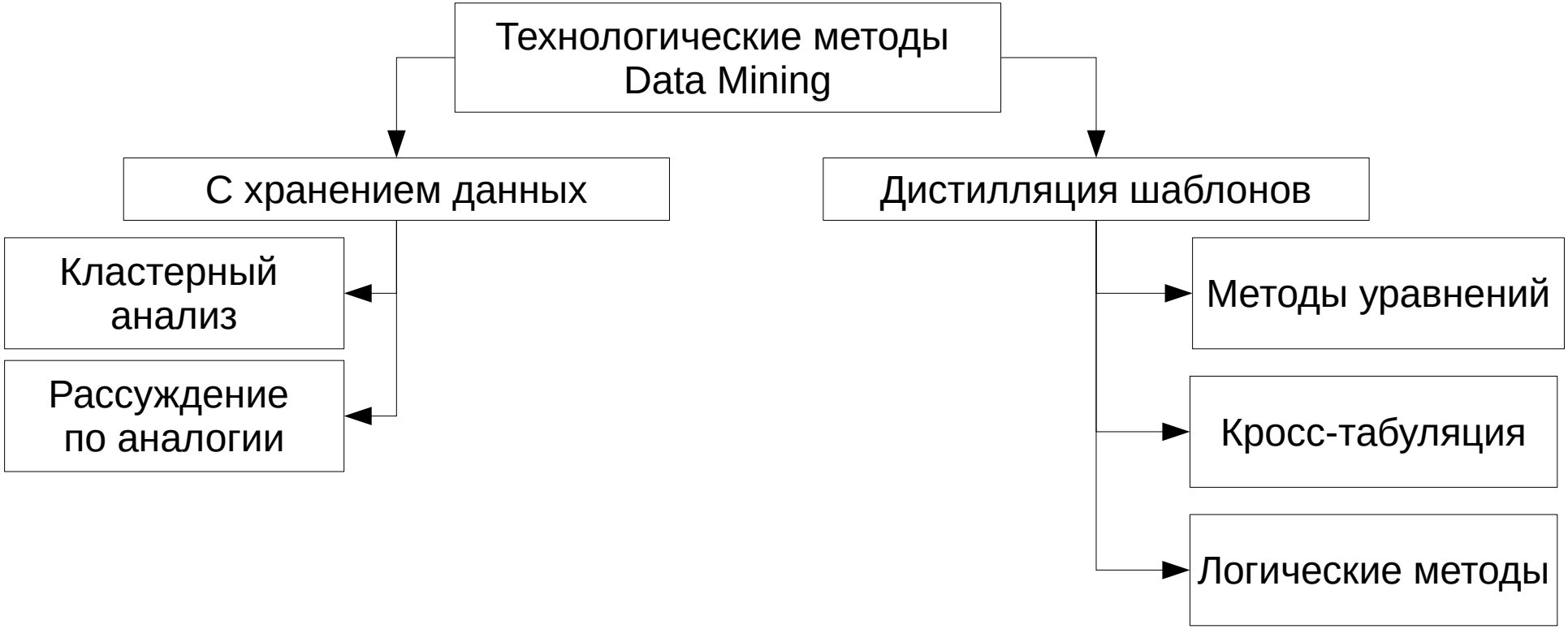
y(x)

ДИАГНОСТИЧЕСКИЙ И ПРОГНОСТИЧЕСКИЙ АНАЛИЗ


Этапы Data Mining
1.Свободный поиск
2.Прогностическое моделирование
3.Анализ исключений
Выявление закономерностей условной логики (conditional logic)
Выявление закономерностей ассоциативной логики (associations and affinities);
выявление трендов и колебаний (trends and variations).
предсказание неизвестных значений (outcome prediction)
прогнозирование развития процессов (forecasting).
анализ исключений и аномалий, выявленные в закономерностях
Улучшение моделей
Очистка данных
Диагностический анализ
Прогностический анализ

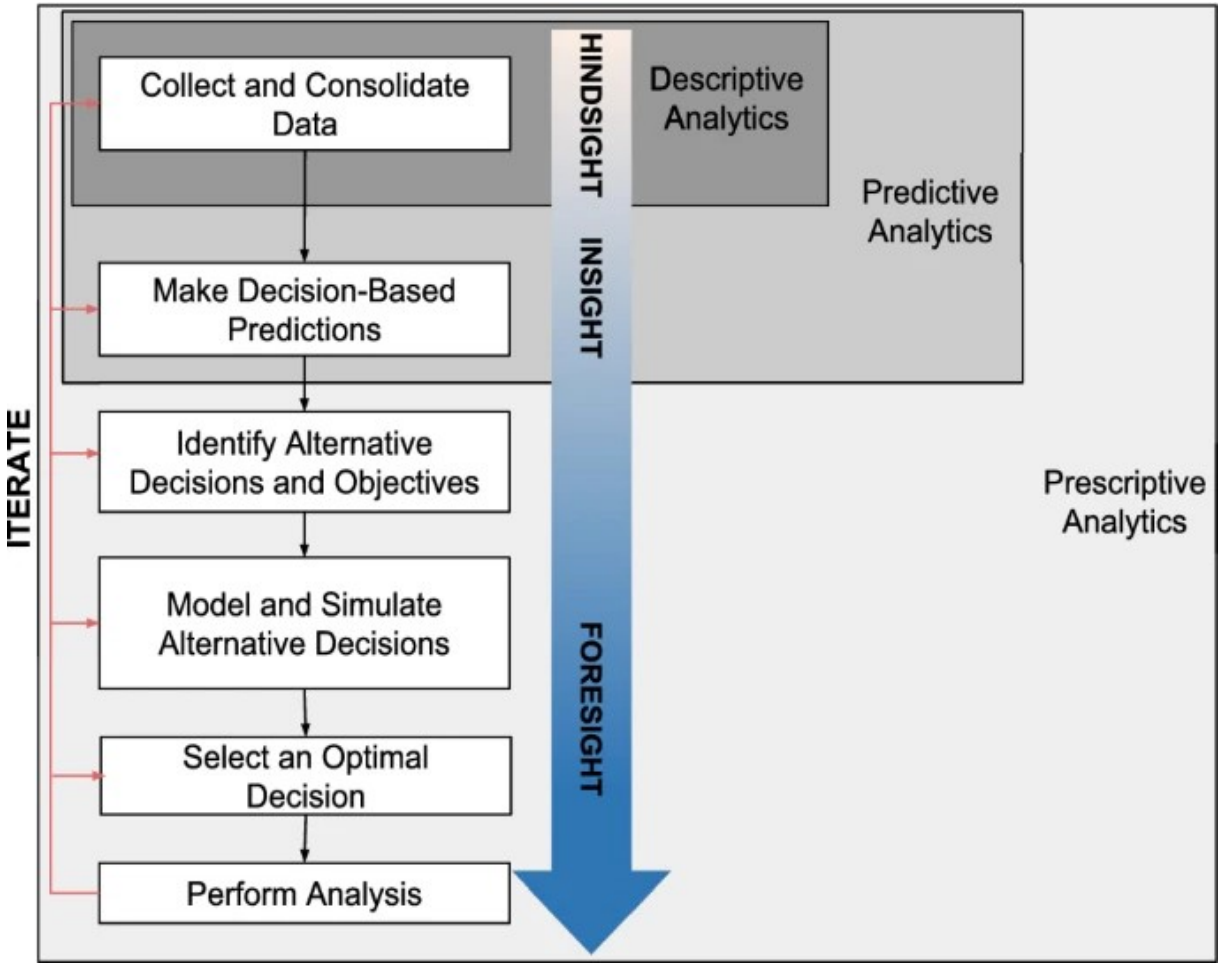

Методы Data Mining
Необучаемые (статистические)
Обучаемые (кибернетические)
Проверка гипотез о стат. характеристиках данных
Выявление связей и закономерностей
Многомерный стат. анализ
Динамические модели на основе анализа рядов
Нейронные сети
Генетические алгоритмы
Нечеткая логика
Эволюционное программирование
Ассоциативная память
Деревья решений
Обработка экспертных знаний

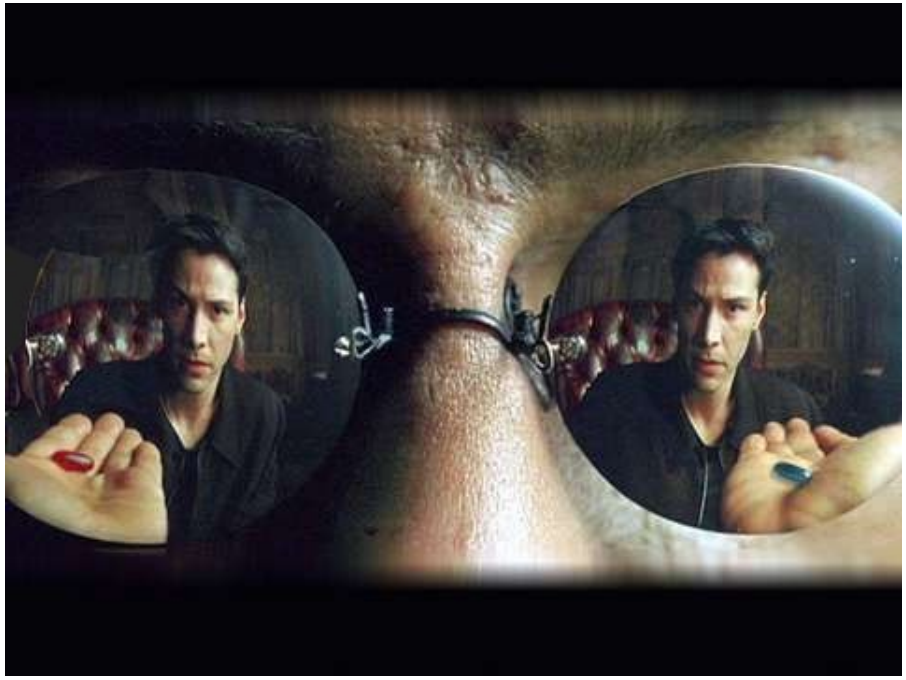

Методы Data Mining
Описательные
Кластерный анализ
Выявление связей и закономерностей
Поиск паттернов и ассоциативных правил
Построение регрессионых моделей
Многомерный стат. анализ
Динамические модели на основе анализа рядов
и др.
Прогностические
Регрессионный прогноз
Классификация
Прогноз
и др.



https://intuit.ru/studies/courses/6/6/lecture/162?
page=3

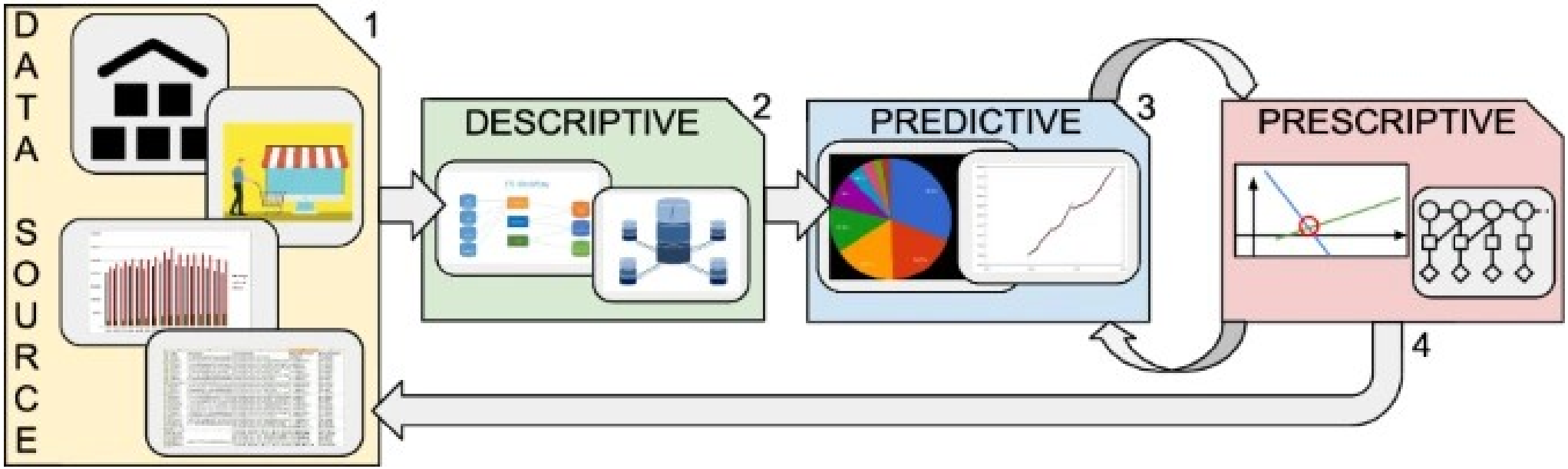
Рекомендательная аналитика





https://link.springer.com/article/10.1007/s00778-019-00539-

Рекомендательная аналитика



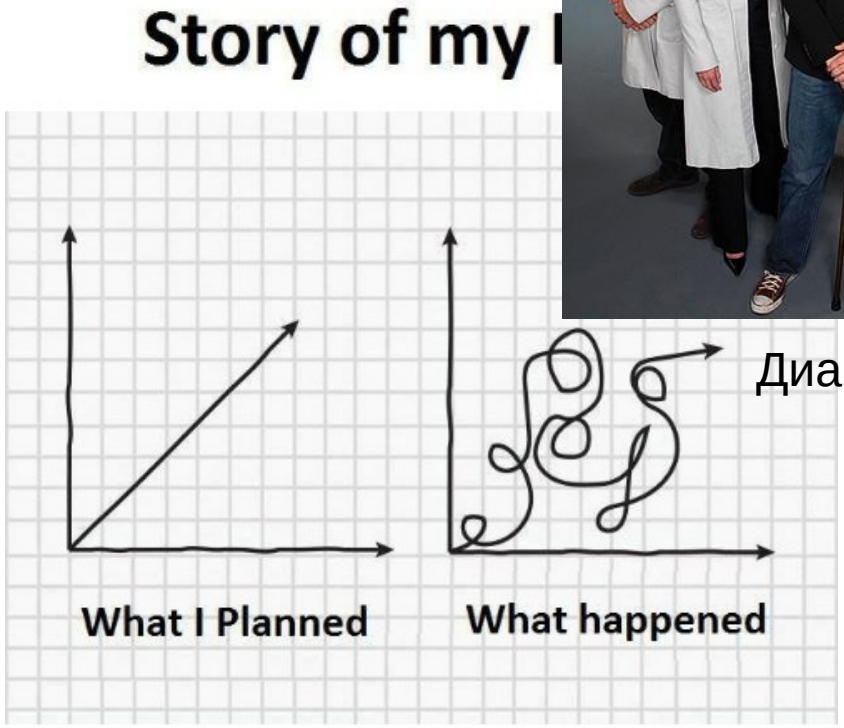
https://link.springer.com/article/10.1007/s00778-019-00539-

Анализ данных

Рекомендательный

Прогностический

Диагностический



Описательный