Инструменты аналитика

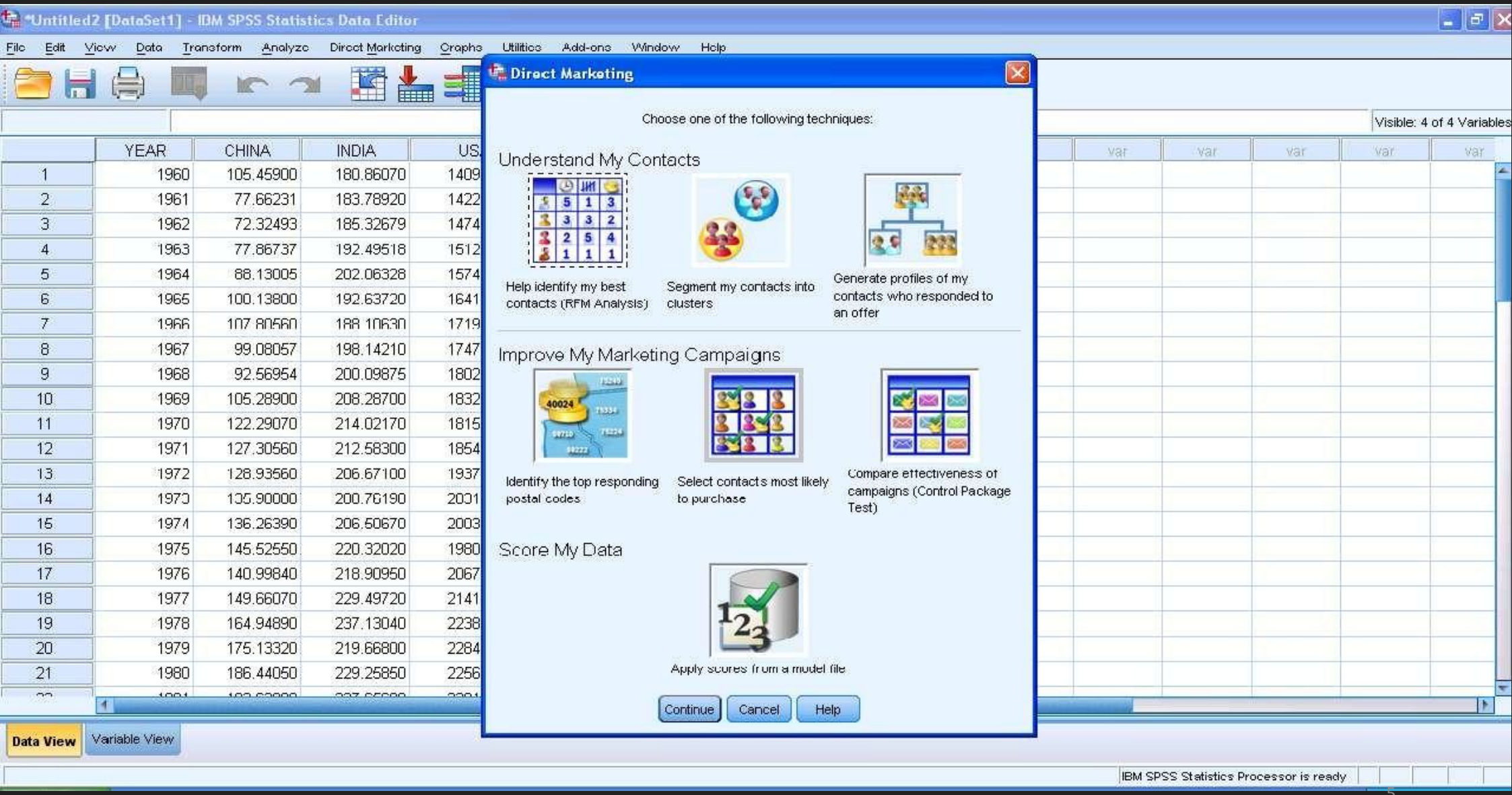
Сбор данных

- Scraping-Bot
- Scrapeworks
- Diggernaut
- ScrapingBee
- Scraper API

Обзор инструментов

- Excel
- SAS, SPSS
- MATLAB, Octave
- Онлайн платформы
- Hadoop, Spark
- Python, R
- Ноутбуки
- Библиотеки

Джон
Форман
МНОГО ЦИФР
Анализ больших данных
при помощи Excel
EDITOR'S
CHOICE
Знания — это сила, а знания, полученные
из больших данных, — большая сила.
альпина


*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 4 of 4 Variables
YEAR CHINA INDIA US
1 1960 105.45900 180.86070 1409
2 1961 77.66231 183.78920 1422
3 1962 72.32493 185.32679 1474
4 1963 77.86737 192.49518 1512
5 1964 88.13005 202.06328 1574
6 1965 100.13800 192.63720 1641
7 1966 107.80560 188.10630 1719
8 1967 99.08057 198.14210 1747
9 1968 92.56954 200.09875 1802
10 1969 105.28900 208.28700 1832
11 1970 122.29070 214.02170 1815
12 1971 127.30560 212.58300 1854
13 1972 128.93560 206.67100 1937
14 1973 135.90000 200.76190 2031
15 1974 136.26390 206.50670 2003
16 1975 145.52550 220.32020 1980
17 1976 140.99840 218.90950 2067
18 1977 149.66070 229.49720 2141
19 1978 164.94890 237.13040 2238
20 1979 175.13320 219.66800 2284
21 1980 186.44050 229.25850 2256
Data View
Variable View
IBM SPSS Statistics Processor is ready
Direct Marketing
Choose one of the following techniques:
Understand My Contacts
Help identify my best contacts (RFM Analysis)
Segment my contacts into clusters
Generate profiles of my contacts who responded to an offer
Improve My Marketing Campaigns
Identify the top responding postal codes
Select contacts most likely to purchase
Compare effectiveness of campaigns (Control Package Test)
Score My Data
Apply scores from a model file
Continue
Cancel
Help

MATLAB

File Edit Debug Desktop Window Help

Shortcuts How to Add What's New

Workspace

N...	Value	Class
a	3	double
phi	<1x401 double>	double
ro_1	<1x401 double>	double
ro_2	<1x401 double>	double
t	<1x101 double>	double
x	<31x21 double>	double
y	<31x21 double>	double

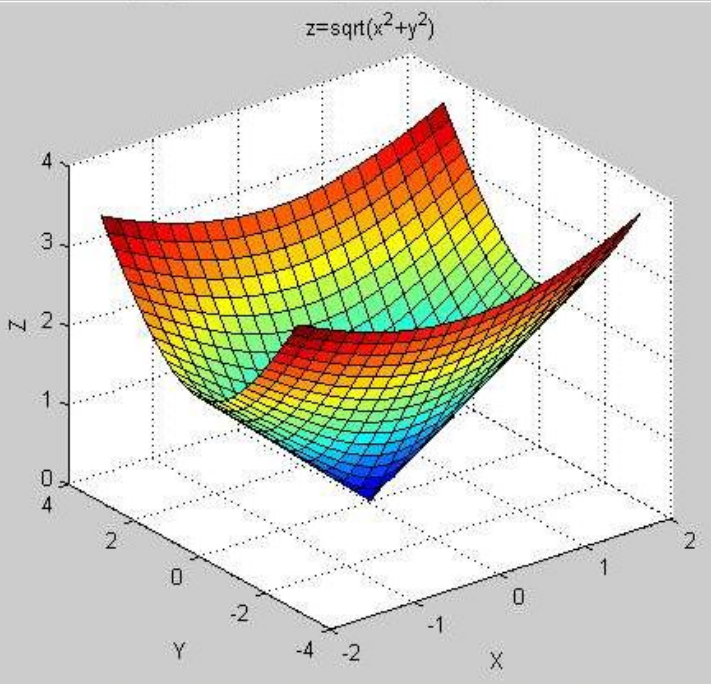
Current Directory Workspace

Command History

```
y=a*t.*(t.^2-1)./(t.^2+1);
plot (x,y,'c-');
grid on;
title('(x+a)a^2+(x-a)y^2=0, a=3');
xlabel('X');
ylabel('Y');
[x y]=meshgrid(-2:0.5:2, -3:0.5:3)
[x y]=meshgrid(-2:2, -3:3)
z=y.^2-x.^2
mesh(x,y,z)
[x y]=meshgrid(-2:0.2:2, -3:0.2:3);
z=sqrt(x.^2+y.^2);
surf(x,y,z);
grid on;
title('z=sqrt(x^2+y^2)');
xlabel('X');
ylabel('Y');
zlabel('Z');
```

Figure 1

File Edit View Insert Tools Desktop Window Help



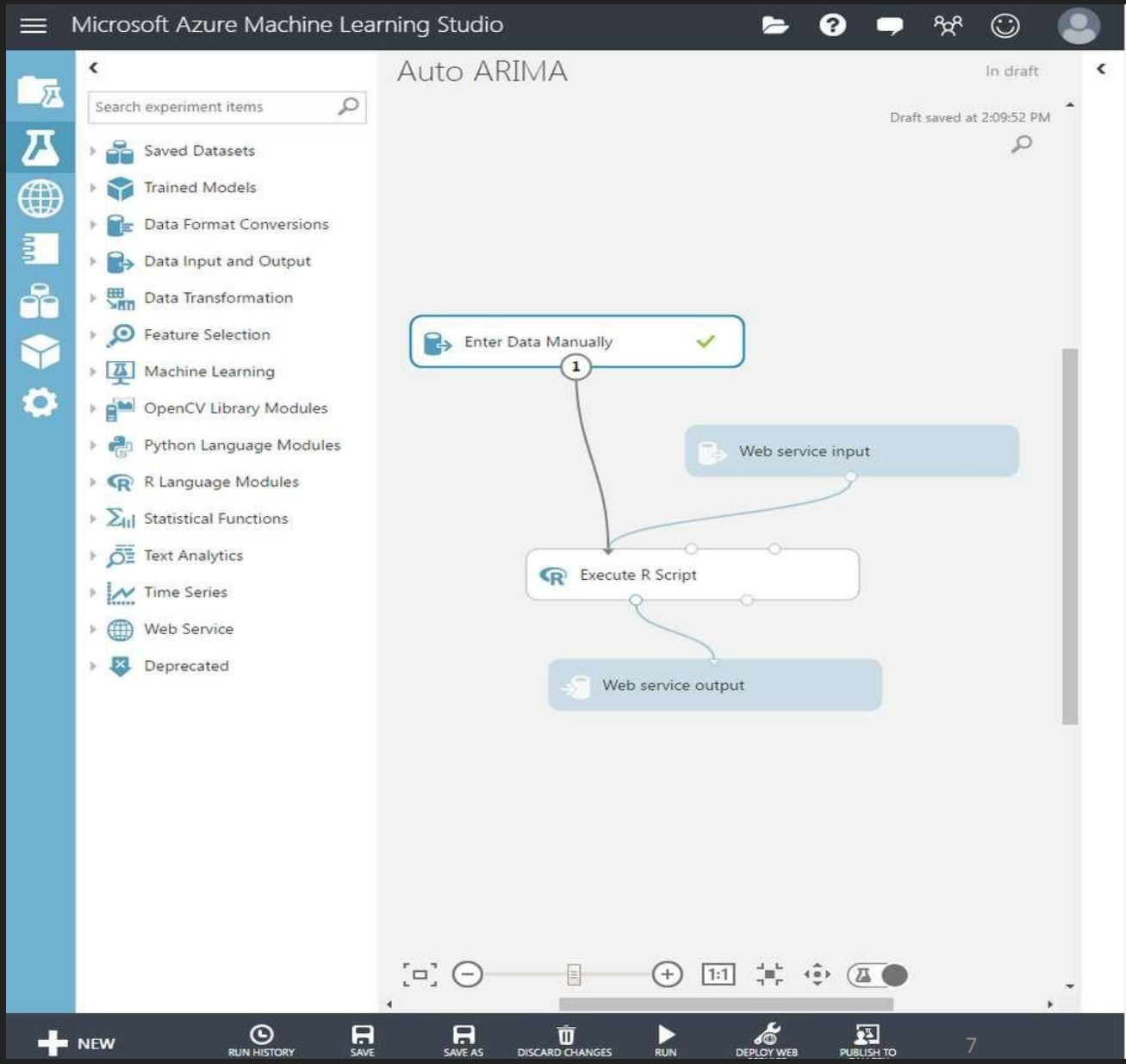

```
>> [x y]=meshgrid(-2:0.2:2, -3:0.2:3);
>> z=sqrt(x.^2+y.^2);
>> surf(x,y,z);
>> grid on;
>> title('z=sqrt(x^2+y^2)');
>> xlabel('X');
>> ylabel('Y');
>> zlabel('Z');
>>
```

Start

EN 21:12 25.06.2011

Облака

- Amazon AWS
- Microsoft Azure
- IBM Watson Analytics



IBM WatsonAnalytics

Georgia Henriot

Getting Started | Manage Data | Open Workbook

Welcome to Watson Analytics!

Explore our solutions by role

MARKETING | SALES | FINANCE | OPERATIONS | HR | IT

Enter a keyword to filter the list below, or to ask Watson a question about your data!

Start from Data

TOOL

EXPLORE YOUR DATA

The easy, beautiful way to find the stories in your data

TOOL

PREDICT AND EXPLAIN

Discover the drivers of behavior and results

TOOL

FORECAST

Start from a Story

TUTORIAL

GETTING STARTED WITH WATSON ANALTYTICS

Take a tour of Watson Analytics!

MARKETING

IMPROVE CAMPAIGN EFFECTIVENESS

Understand the drivers of campagin success

HR

RETAIN YOUR TEAM

Identify high risk employees

TUTORIAL

WORKING WITH DATA

Get more from your data with Watson Analytics

HR

PREVENTING EMPLOYEE ATTRITION

Identify the causes of attrition before its too late

SALES

FIND PATTERNS IN WINS AND LOSSES

What combination of factors leads to a win?

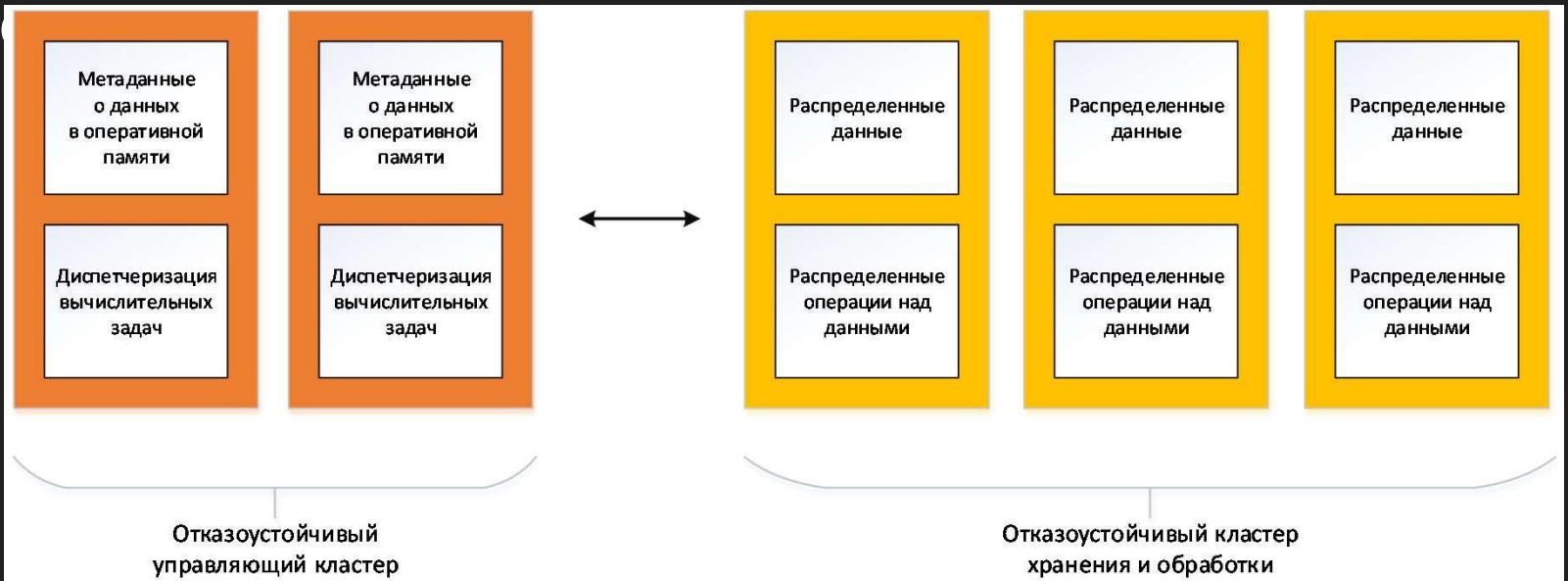
HR

MARKETING

SALES

Зачем нужен Hadoop

- В 10 раз дешевле СХД
- Вычисления и данные — в одном месте
- Вместе с удешевлением HDD устроил рев



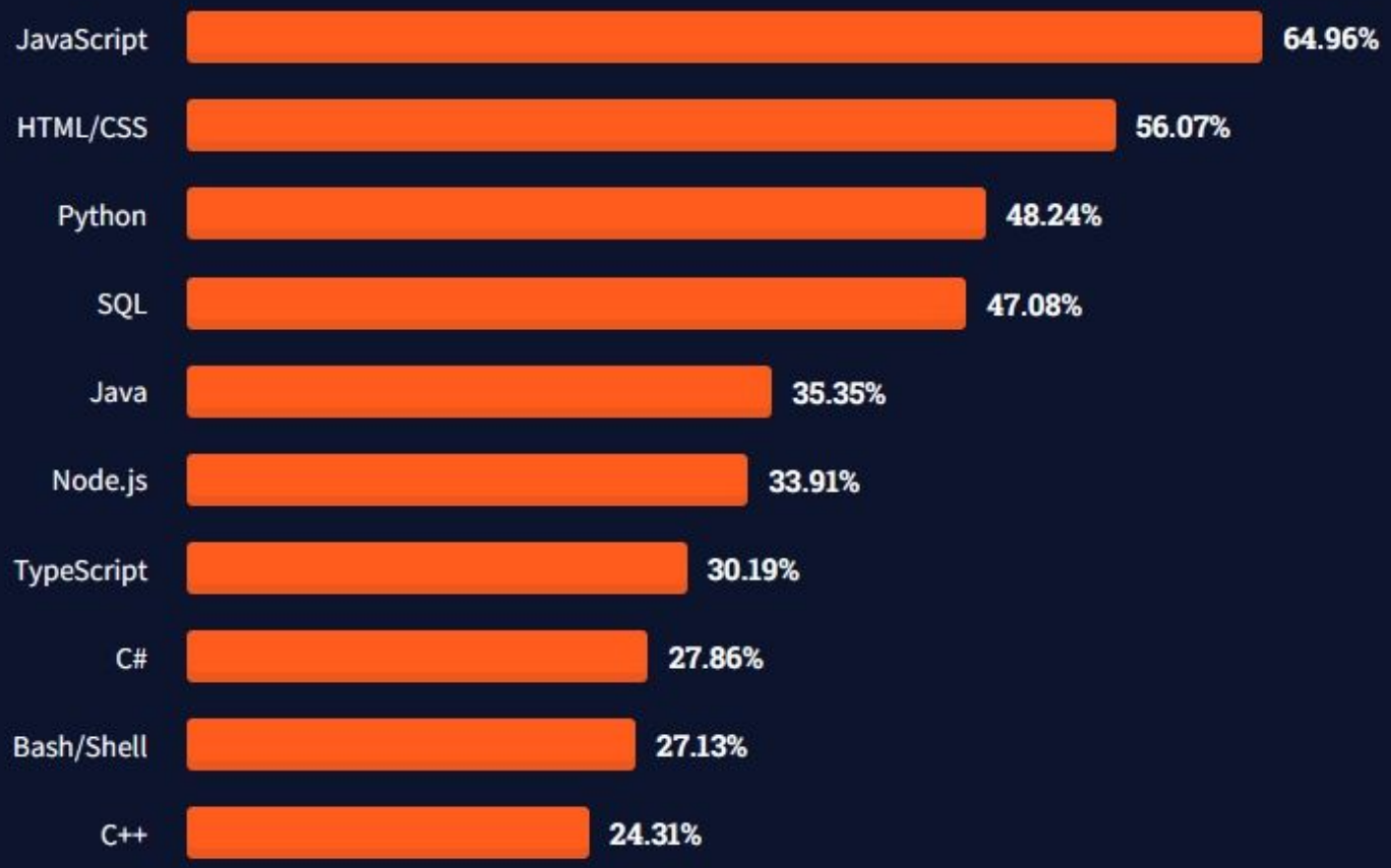
Вычисления на кластере

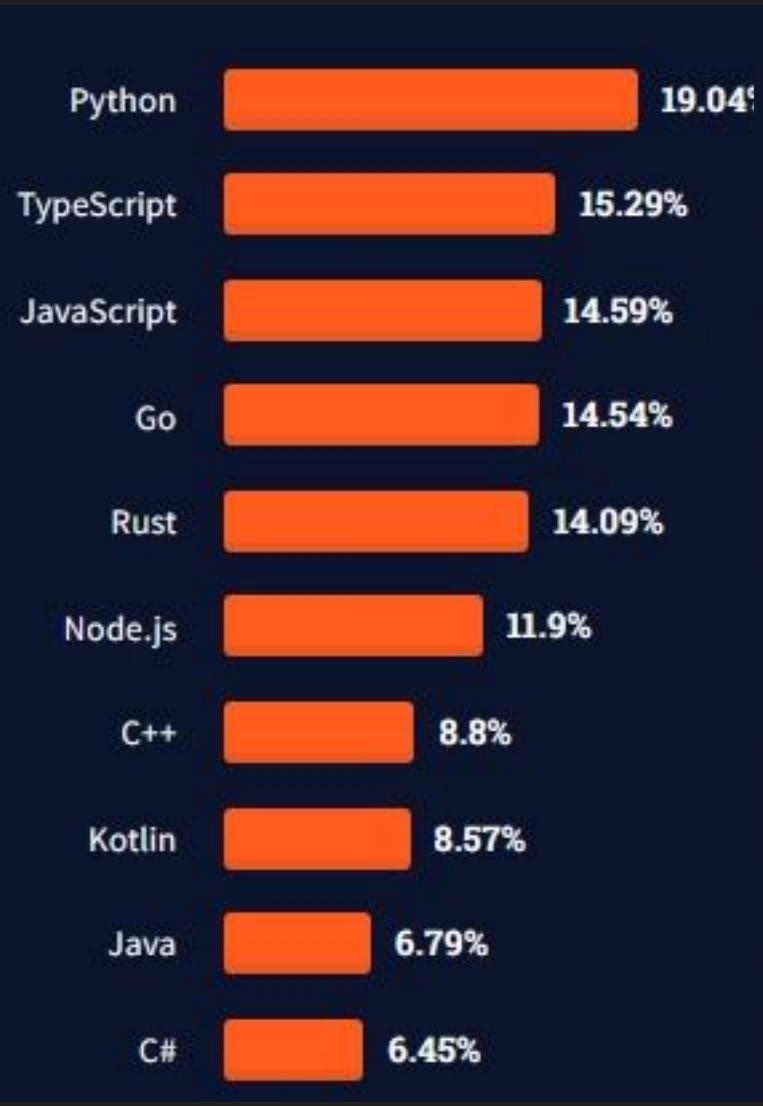
- Единый шедуллер, разделяющий ресурсы между всеми
- Типичные потребители:
  - Базы данных (HBASE)
  - SQL-like инструменты (HIVE)
  - Map Reduce операции
  - Spark (in-memory)
- Крайне важно грамотно организовать параллельность

Python

- Все становятся программистами
- Преимущества Python
  - Простота и удобство, легкость в освоении
  - Расширяемость, огромное количество библиотек и примеров
  - Data Science стек, парсинг сайтов, веб-сервисы…
- Недостатки Python
  - Медленнее компилируемых языков e.g. C++, Java
  - Не подходит для мобильной разработки
  - Не всегда лучшее решение для enterprise

All Respondents | Professional Developers




Python
19.04
TypeScript
15.29%
JavaScript
14.59%
Go
14.54%
Rust
14.09%
Node.js
11.9%
C++
8.8%
Kotlin
8.57%
Java
6.79%
C#
6.45%


Rust 86.98% 13.02%
Clojure 81.12% 18.88%
TypeScript 72.73% 27.27%
Elixir 72.11% 27.89%
Julia 70.69% 29.31%
Python 67.83% 32.17%
Dart 63.77% 36.23%
Swift 63.56% 36.44%
Node.js 63.22% 36.78%
Go 62.74% 37.26%
F# 62.44% 37.56%
C# 61.96% 38.04%
SQL 61.83% 38.17%
Kotlin 61.55% 38.45%
JavaScript 61.51% 38.49%
HTML/CSS 60.19% 39.81%
Crystal 59.44% 40.56%
Bash/Shell 56.34% 43.66%
LISP 55.29% 44.71%
Erlang 53.76% 46.24%
Ruby 53.24% 46.76%
Haskell 53.00% 47.00%
Scala 52.28% 47.72%
APL 50.75% 49.25%

Рейтинг языков программирования по количеству вакансий

Язык программирования	Количество вакансий для языка программирования
SQL	28729
JavaScript	15903
Python	15027
Java	11362
C#	6697
PHP	6539
C++	5612
TypeScript	4227
Kotlin	3012
Golang	1560
Swift IOS	1457
Ruby	929
Scala	867
Objective-C	686
ABAP	631
Delphi	615
Perl	585
Assembler	357
Dart	246
Rust	165
Elixir	98
Erlang	93
Haskell	43
F#	15

О языке

- 1991 год рождения, Нидерланды
- Основан на ABC, который основан на SETL, 1969

C++ – 1983, C – 1973
R – 1993, S – 1976
JavaScript – 1995
SQL – 1979

Красивое лучше, чем уродливое. Явное лучше, чем неявное.
Простое лучше, чем сложное.
Сложное лучше, чем запутанное. Плоское лучше, чем вложенное. Разреженное лучше, чем плотное.
Читаемость имеет значение.
Особые случаи не настолько особые, чтобы нарушать правила.
При этом практичность важнее безупречности. Ошибки никогда не должны замалчиваться.
Если они не замалчиваются явно.
Встретив двусмысленность, отбрось искушение угадать. Должен существовать один и, желательно, только один очевидный способ сделать это.
Хотя он поначалу может быть и не очевиден, если вы не голландец.
Сейчас лучше, чем никогда.
Хотя никогда зачастую лучше, чем прямо сейчас. Если реализацию сложно объяснить — идея плоха.
Если реализацию легко объяснить — идея, возможно

Основные свойства

- Python – интерпретируемый язык
- CPython – основная реализация интерпретатора, написан на C
- Динамическая типизация
- «white space» играет роль
- Установка модулей через пакетный менеджер (pip, conda)
- 130 000 различных модулей (март 2021)

Hello, world!

C++

```
#include <iostream>
using namespace std;

int main()

{
  // print output to user
  cout << "Hello, world!" << endl;
  return 0;
}
```

Python

```
print("Hello world!")
```

Python

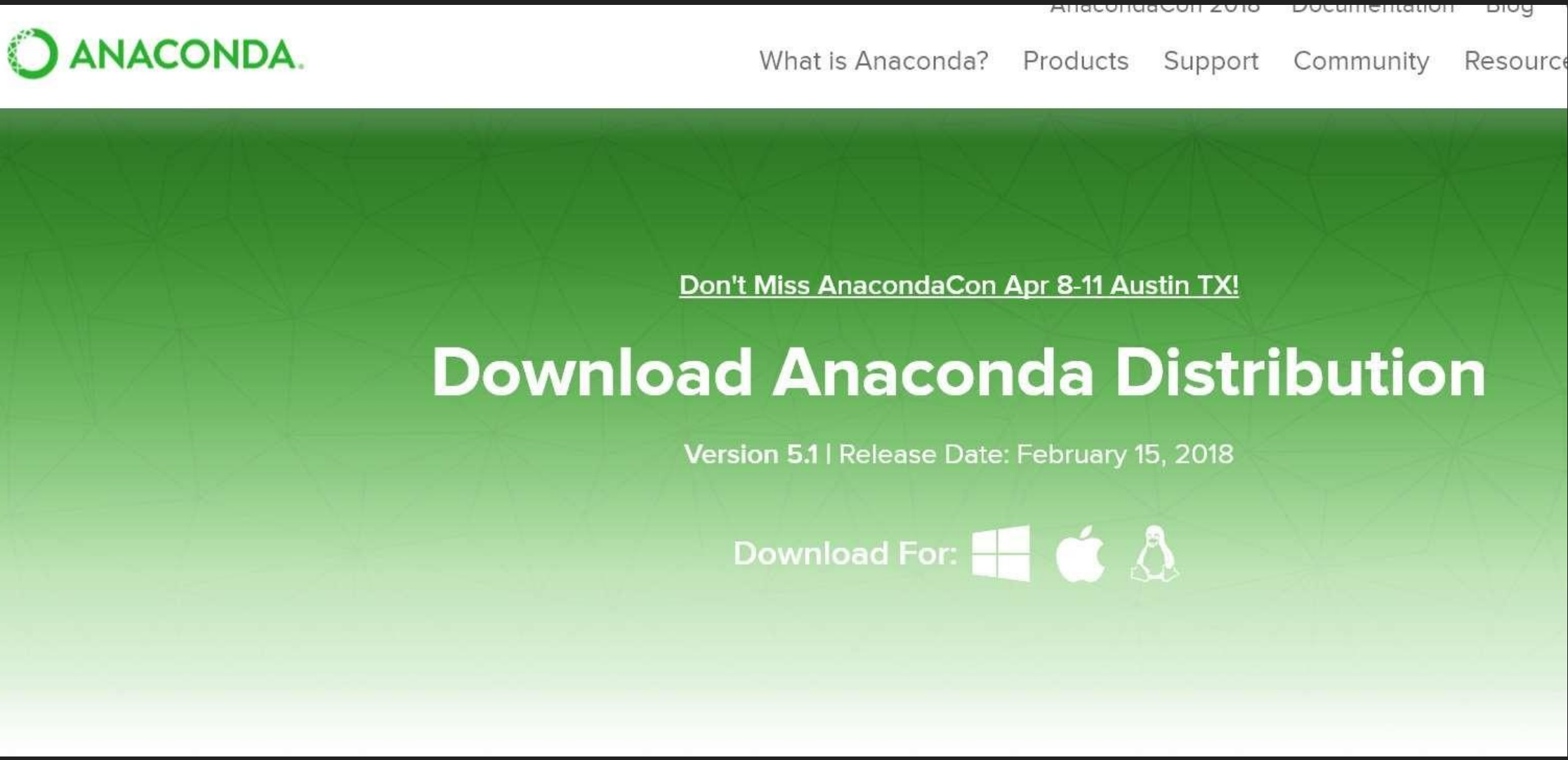
- Numpy
- Scipy
- Pandas
- Matplotlib
- Scikit-learn

- Ну и тысячи других

Модели, фреймворки

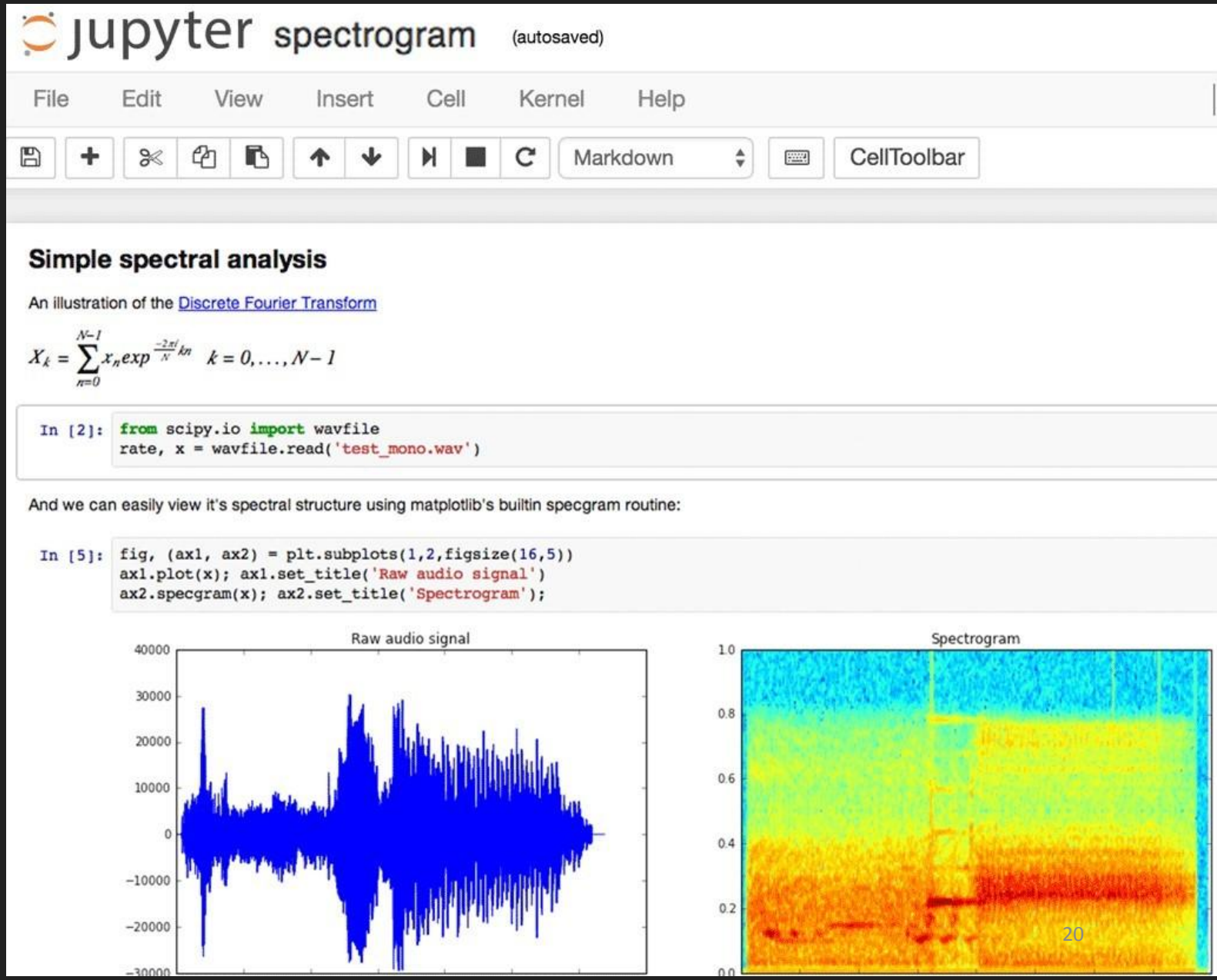
- Градиентный бустинг
  - XGBoost
  - Catboost
  - LightGBM
- Нейросети
  - Keras
  - Caffe
  - TensorFlow
  - Theano
  - PyTorch

- Обёртки для языков
- Применение из консоли
- Параллельное обучение
- Параллельное применение
- На одной машине и на кластере
- Исполнение на CPU, GPU
- Поддержка Windows/Mac/Linux
- Поддержка ARM


ANACONDA.
What is Anaconda?
Products
Support
Community
Resource
Don't Miss AnacondaCon Apr 8-11 Austin TX!
Download Anaconda Distribution
Version 5.1 | Release Date: February 15, 2018
Download For:

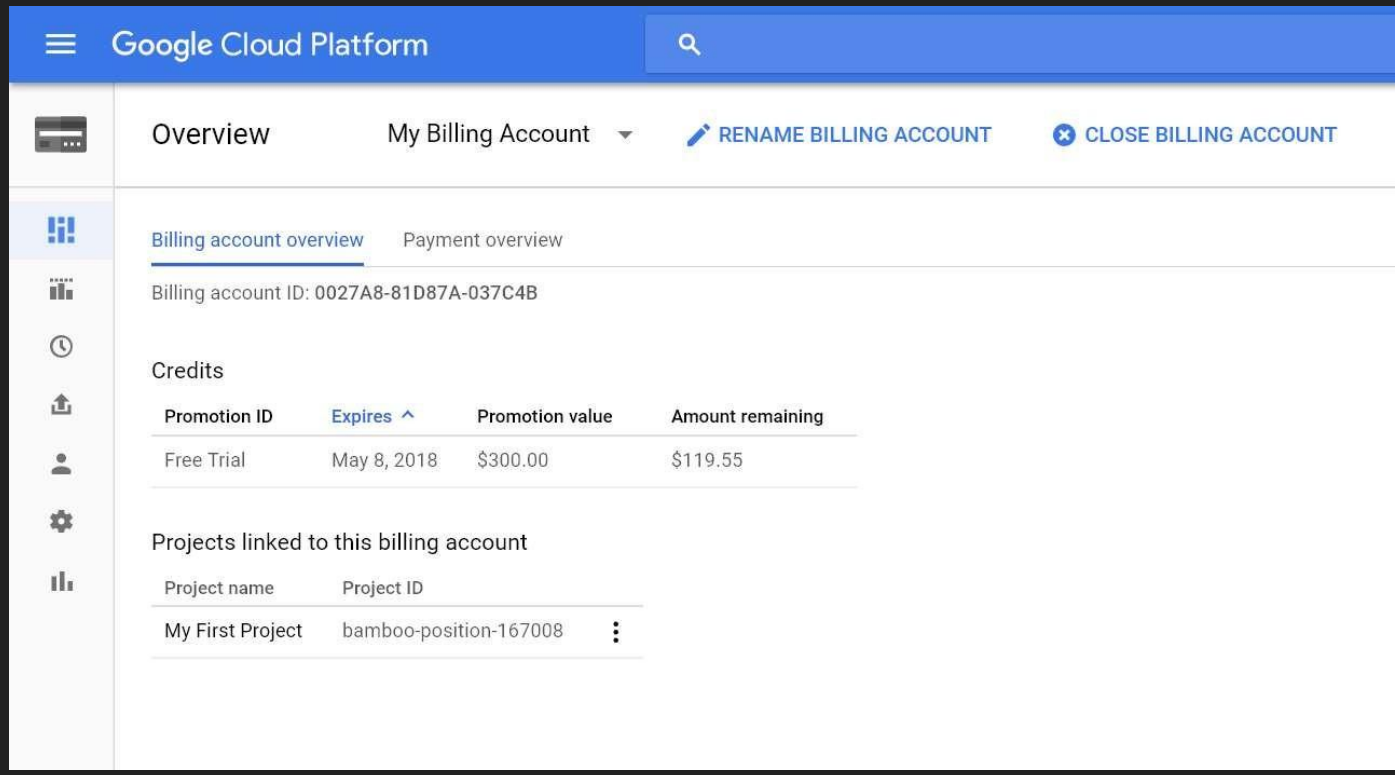
IPython и Jupyter

- IPython — интерактивная консоль python’а
- Jupyter — популярный аналитчиески й ноутбук



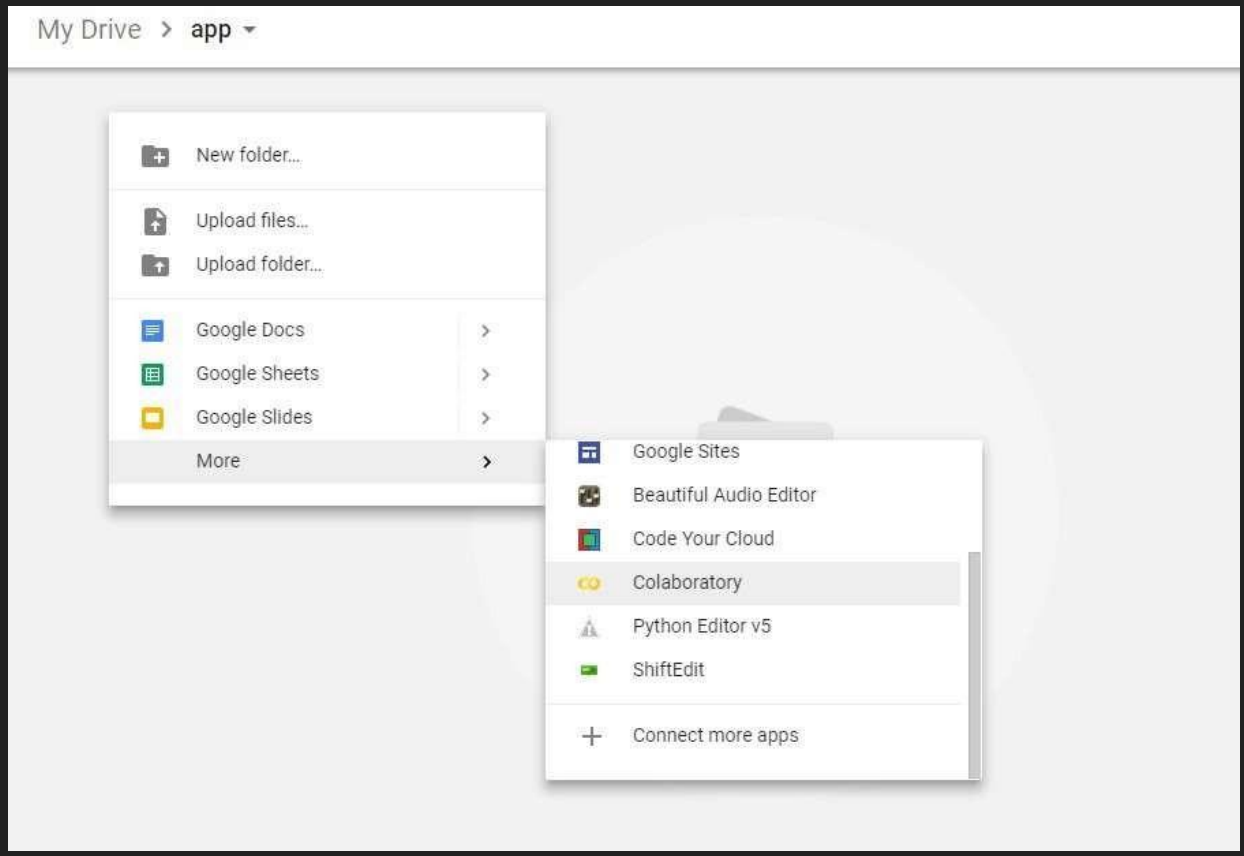
Облака

- Google Cloud Platform

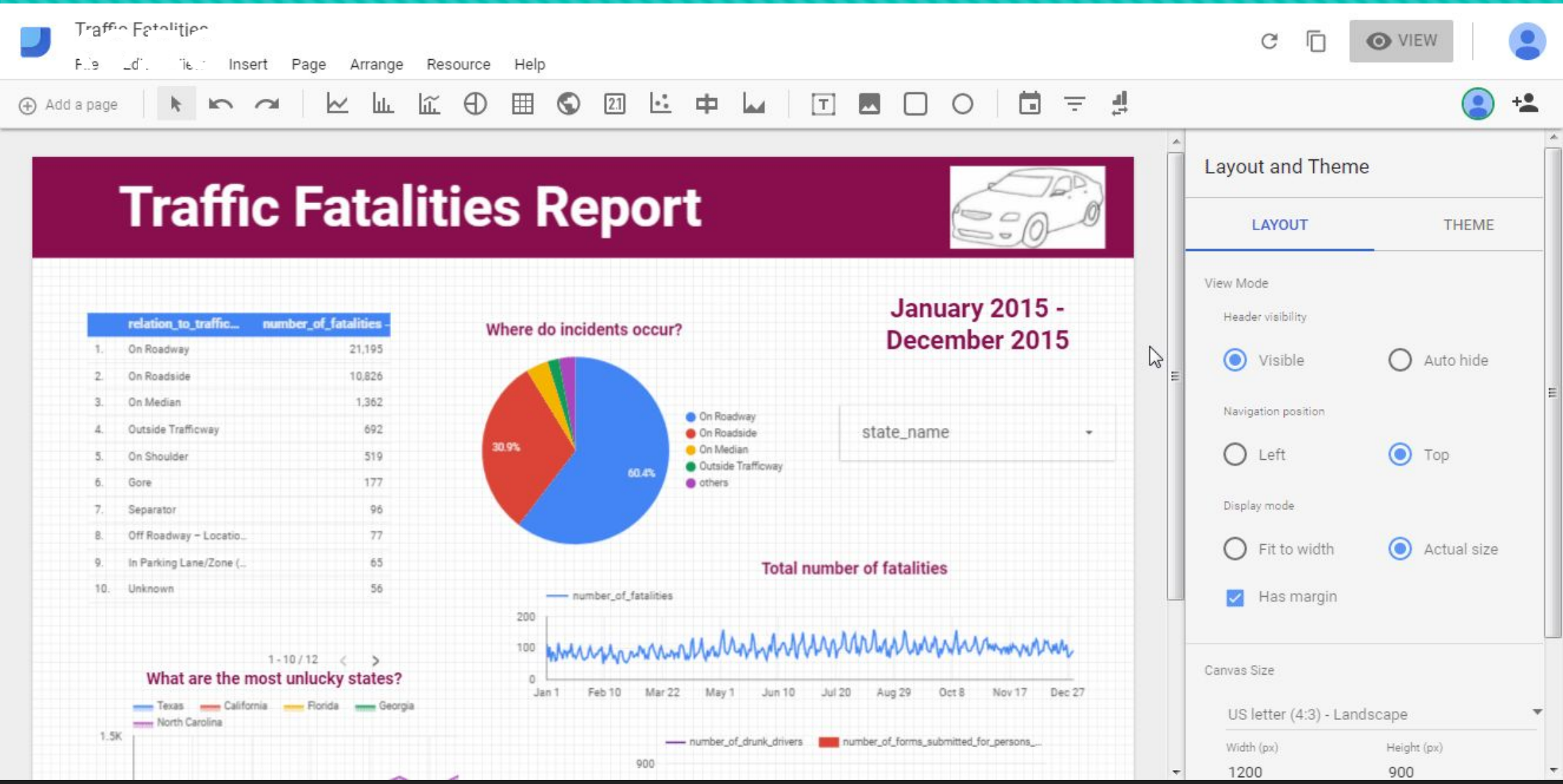


Облака

- Google Colab



Tableau, Power BI, Google Data Studio: простая онлайн-визуализация без



Программа на Python

problems_test.py

```
1  # coding=utf-8
2  # Copyright 2018 The Tensor2Tensor Authors.
3  #
4  # Licensed under the Apache License, Version 2.0 (the "License");
5  # you may not use this file except in compliance with the License.
6  # You may obtain a copy of the License at
7  #
8  #     http://www.apache.org/licenses/LICENSE-2.0
9  #
10 # Unless required by applicable law or agreed to in writing, software
11 # distributed under the License is distributed on an "AS IS" BASIS,
12 # WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 # See the License for the specific language governing permissions and
14 # limitations under the License.
15
16 """tensor2tensor.problems test."""
17
```

Программа на Python

```
21
22  # Dependency imports
23
24  from tensor2tensor import problems
25
26  import tensorflow as tf
27
```

Базовые алгоритмические конструкции

```
a = 1
b = 2
c = a + b
d = a – b
print(c)
print(d)
```

Оператор условия

```
if a > b:
c = a
else:
  c = b
```

Оператор условия

```
if a > 0:
  c = a
elif a == 0:
  c = b
else:
  c =
  d
```

Циклы

```
# while
m = 0
while m < 10:
  m = m + 1
  print(m)

# for
for n in range(1, 10):
  print(n)
```

Функции

```
# пример определения и вызова
функции  def time(hour, minute=0):
  return("Время: %i часов %i минут" % (hour,
  minute))

time(8)
time(9, 20)
time(minute=5, hour=10)
```

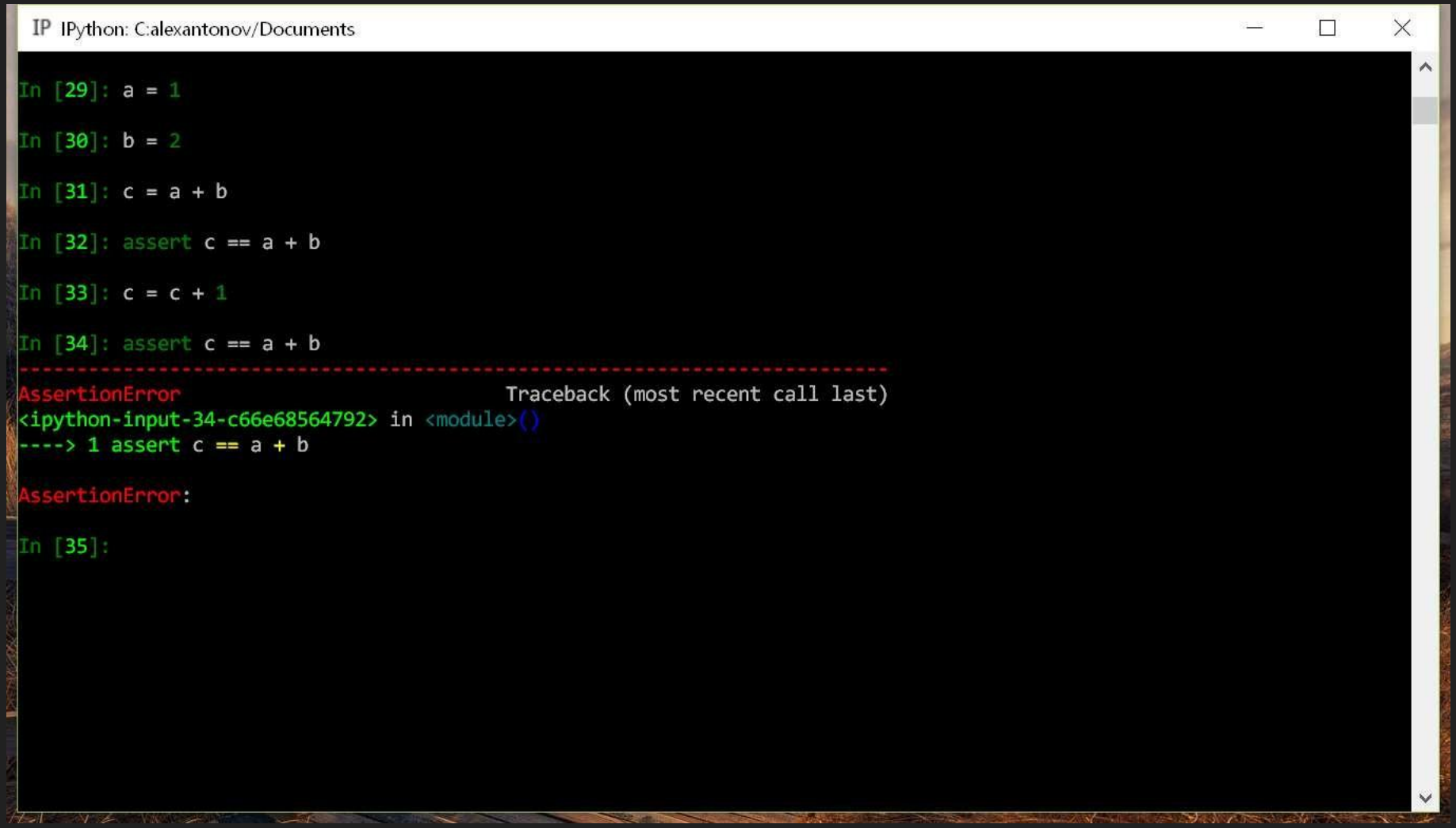
Исключения


IP IPython: C:\alexantonov/Documents
In [14]: try:
...: a = 1
...: b = 0
...: c = a/b
...: except ZeroDivisionError:
...: print("Ошибка деления на ноль")
...: except:
...: print("Другая ошибка")
...:
...:
Ошибка деления на ноль
In [15]: try:
...: a = 1
...: b = 0
...: c = a/d
...: except ZeroDivisionError:
...: print("Ошибка деления на ноль")
...: except:
...: print("Другая ошибка")
...:
...:
Другая ошибка
In [16]:

Исключения


IP IPython: C:\alexantonov/Documents
In [14]: try:
...: a = 1
...: b = 0
...: c = a/b
...: except ZeroDivisionError:
...: print("Ошибка деления на ноль")
...: except:
...: print("Другая ошибка")
...:
...:
Ошибка деления на ноль
In [15]: try:
...: a = 1
...: b = 0
...: c = a/d
...: except ZeroDivisionError:
...: print("Ошибка деления на ноль")
...: except:
...: print("Другая ошибка")
...:
...:
Другая ошибка
In [16]:

Исключения


IP IPython: C:alexantonov/Documents
In [29]: a = 1
In [30]: b = 2
In [31]: c = a + b
In [32]: assert c == a + b
In [33]: c = c + 1
In [34]: assert c == a + b
AssertionError Traceback (most recent call last)
<ipython-input-34-c66e68564792> in <module>()
----> 1 assert c == a + b
AssertionError:
In [35]:

Типы данных

- целое число **int** и **long**
- число с плавающей точкой **float**
- логический **bool** (True или False)
- строка **string**

Типы данных

- кортеж **tuple** p = 1, "Winter", True
- список **list** l = [1, 2, 3, 4]
- словарь **dict** d = {1: 'one', 2: 'two', 3: 'three', 4: 'four'}