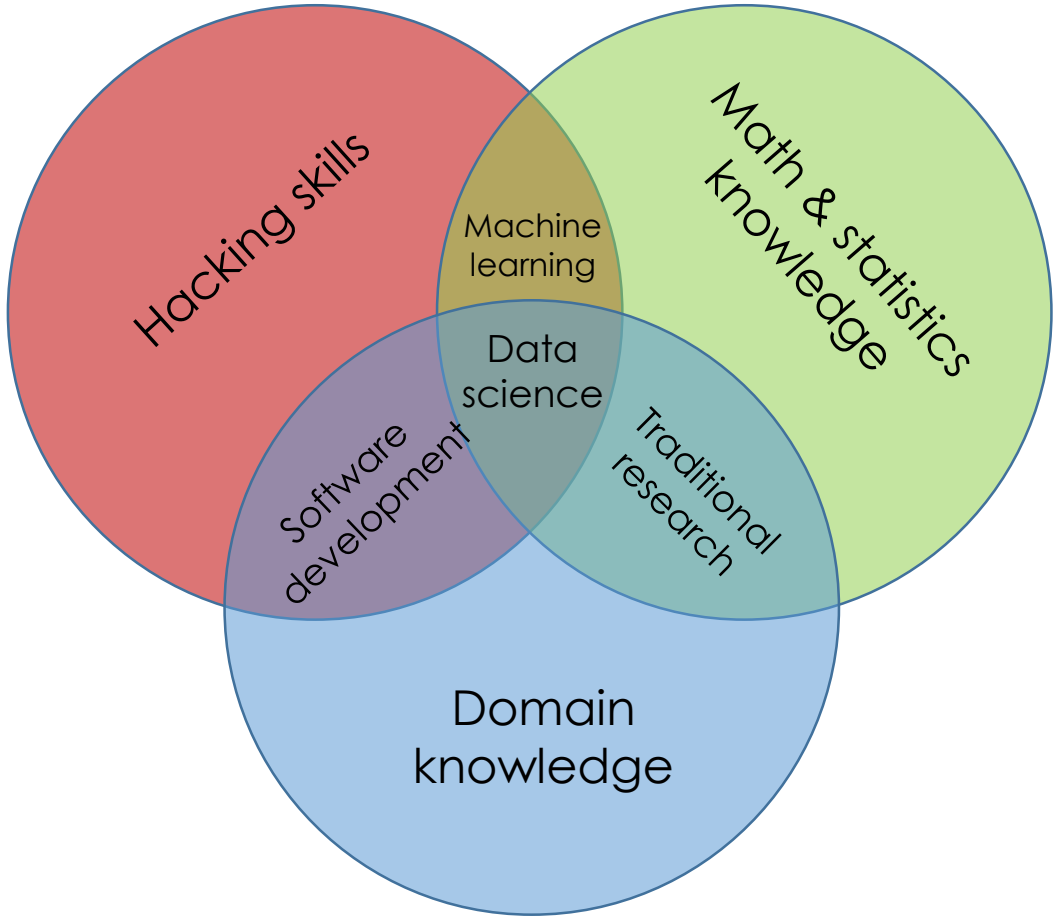
arcadia
software country

Data Science в Life Sciences

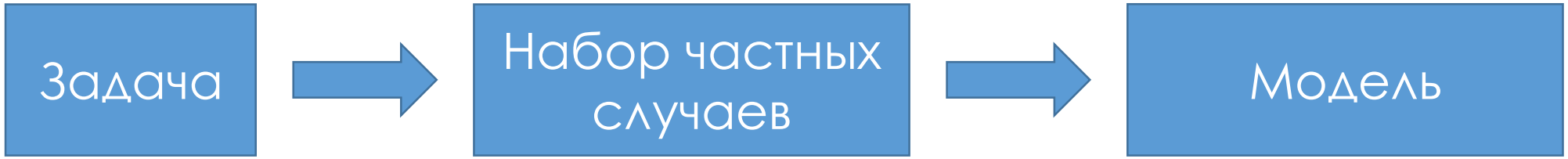
О себе

- 18 лет в IT
- Frontend, backend, DB – MS stack
- C++ -> .NET -> R/Python

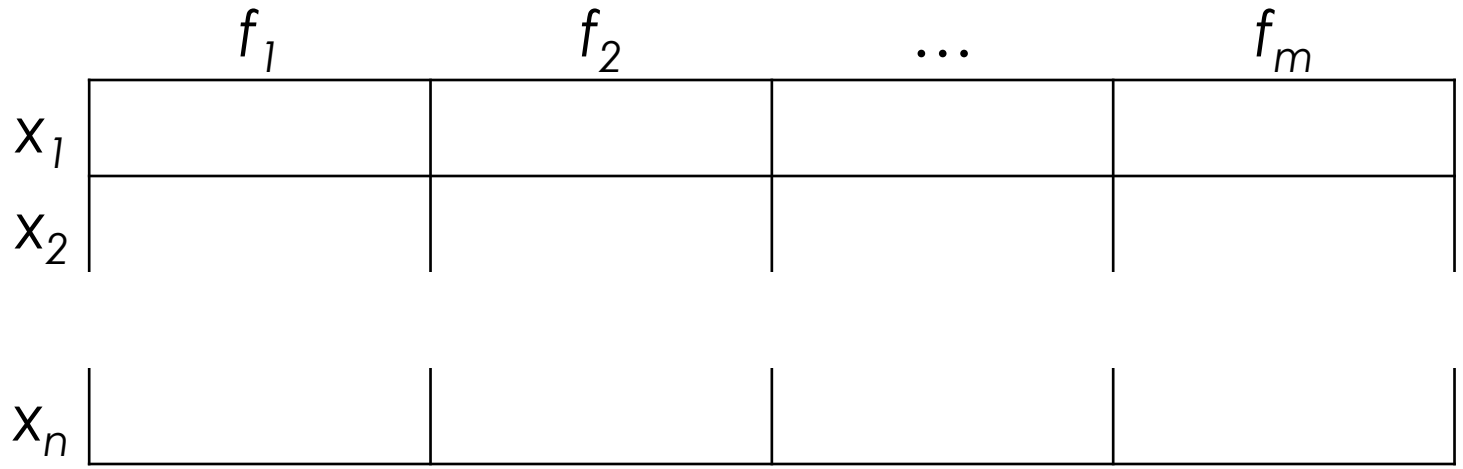
Что такое Data Science



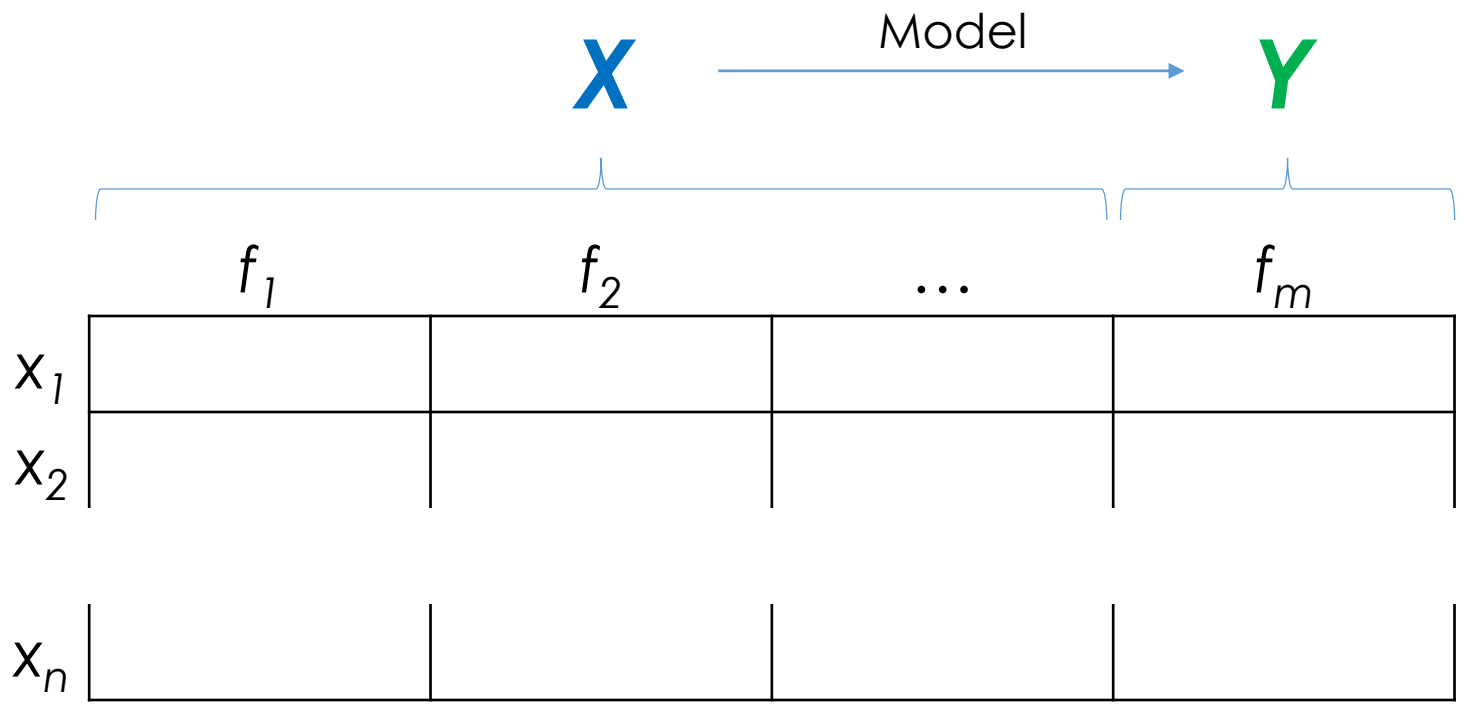
Machine Learning



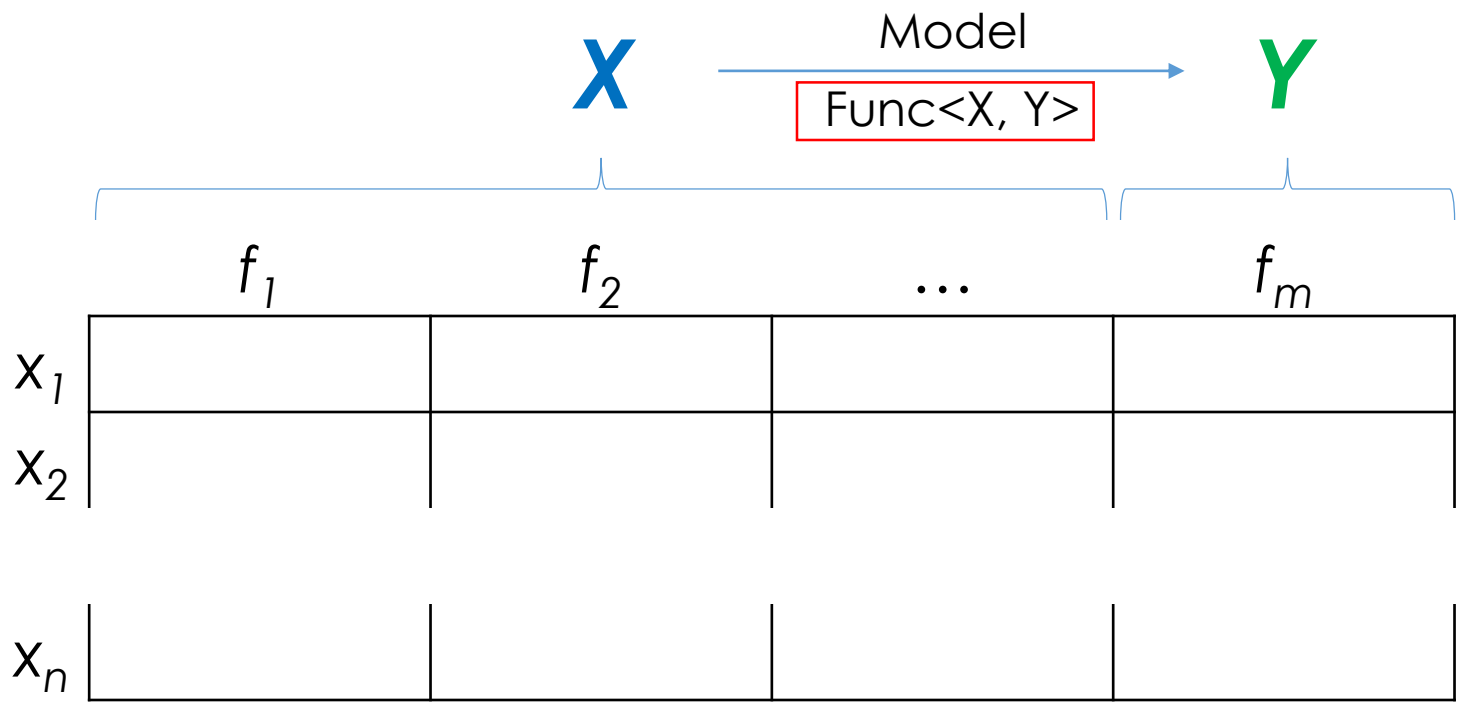
Machine Learning / Training set



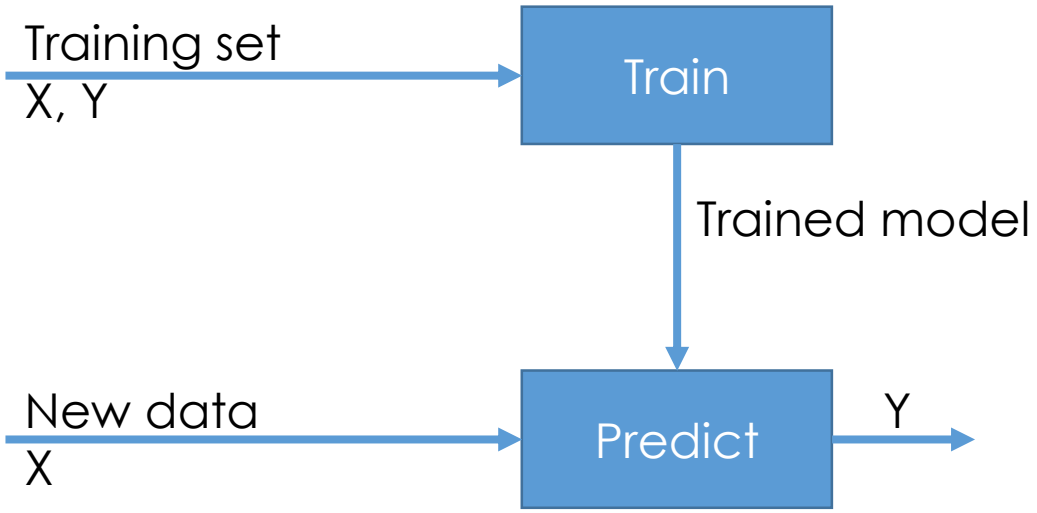
Machine Learning / Training set



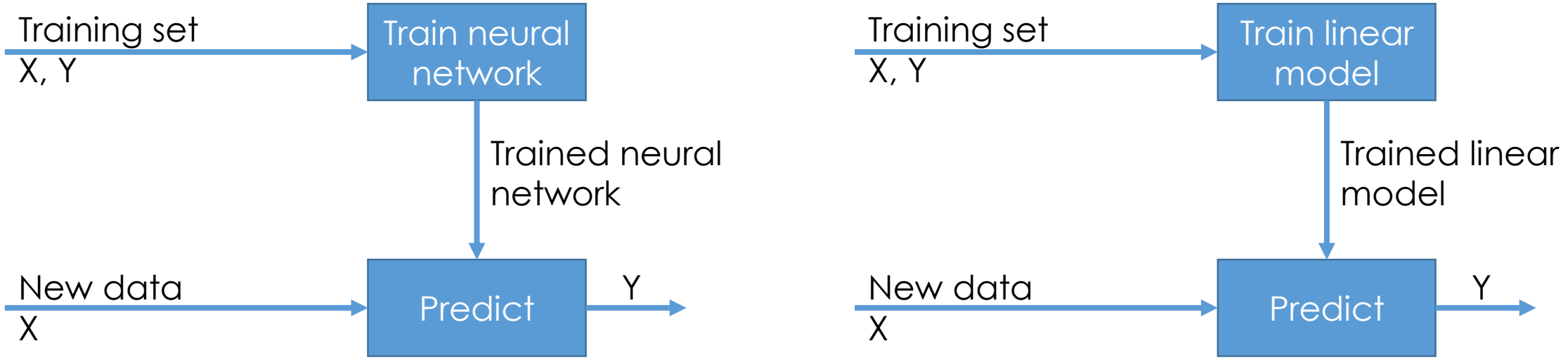
Machine Learning / Training set



Machine Learning



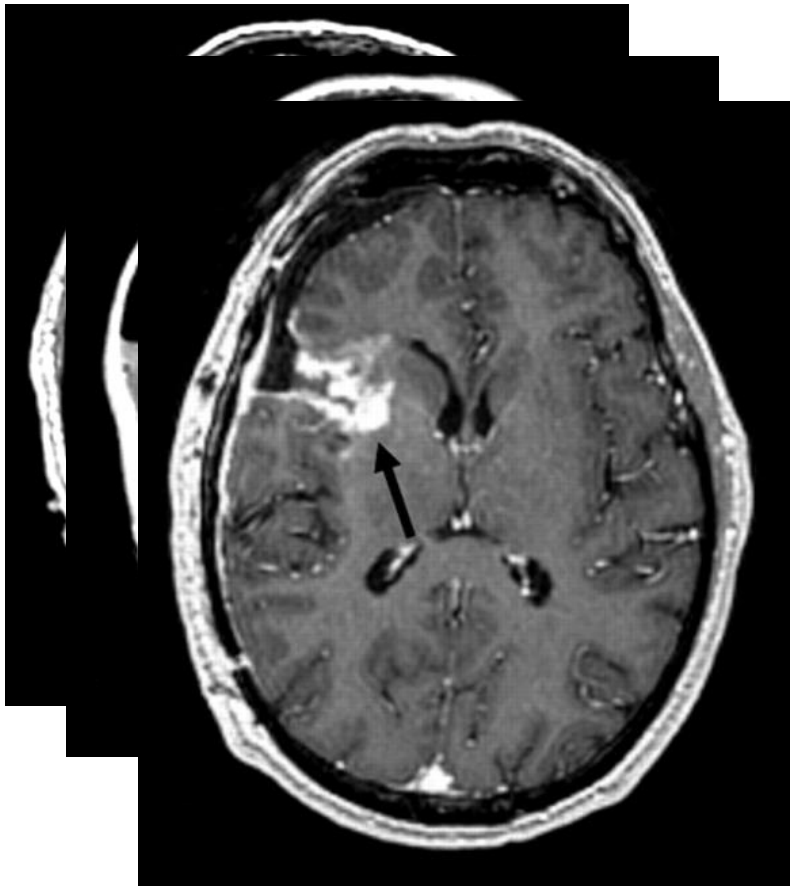
Machine Learning



Life sciences

- Agriculture, Fisheries & Food
- Anatomy & Morphology
- Behavioral Sciences
- Biology, Biochemistry and Biotechnology
- Biophysics
- Ecology, Evolution & Environment
- Entomology
- Forestry
- Genetics & Heredity
- Immunology
- Mycology
- Paleontology
- Parasitology
- Pharmacology & Pharmacy
- Physiology
- Plant Sciences
- Toxicology
- Veterinary Sciences
- Virology
- Zoology
- …

Пример – компьютерная диагностика

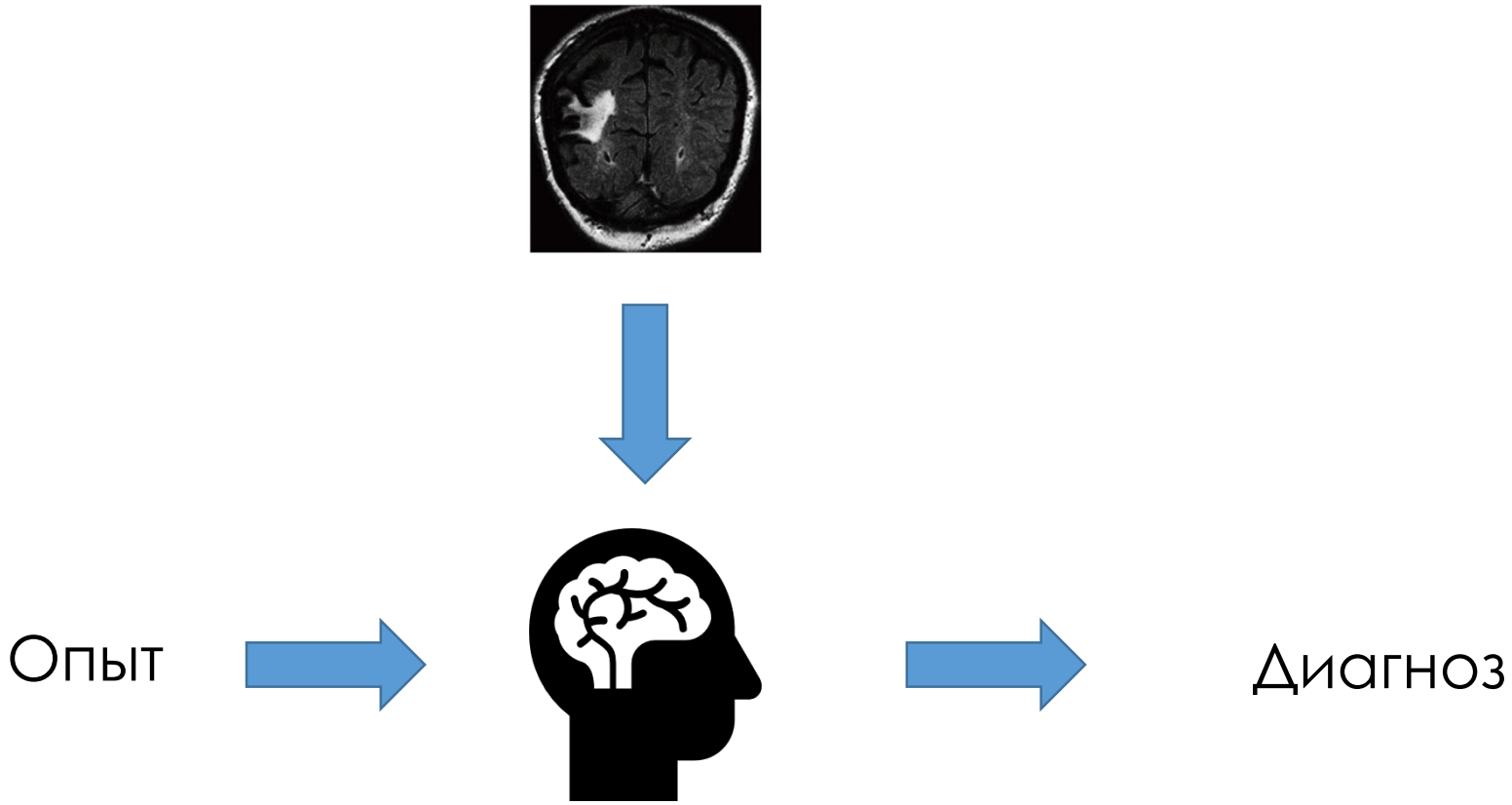


A

B

C

Пример – компьютерная диагностика



Классификация изображений

```
interface I3ClassTumorClassifier
{
    Tuple<float, float, float> Predict(Image image);
}
```

Вероятности принадлежности изображения к одному из трёх классов.
В сумме равны единице.

Классификация изображений

```
interface I3ClassTumorClassifier
{
    Tuple<float, float, float> Predict(Image image);
}
```

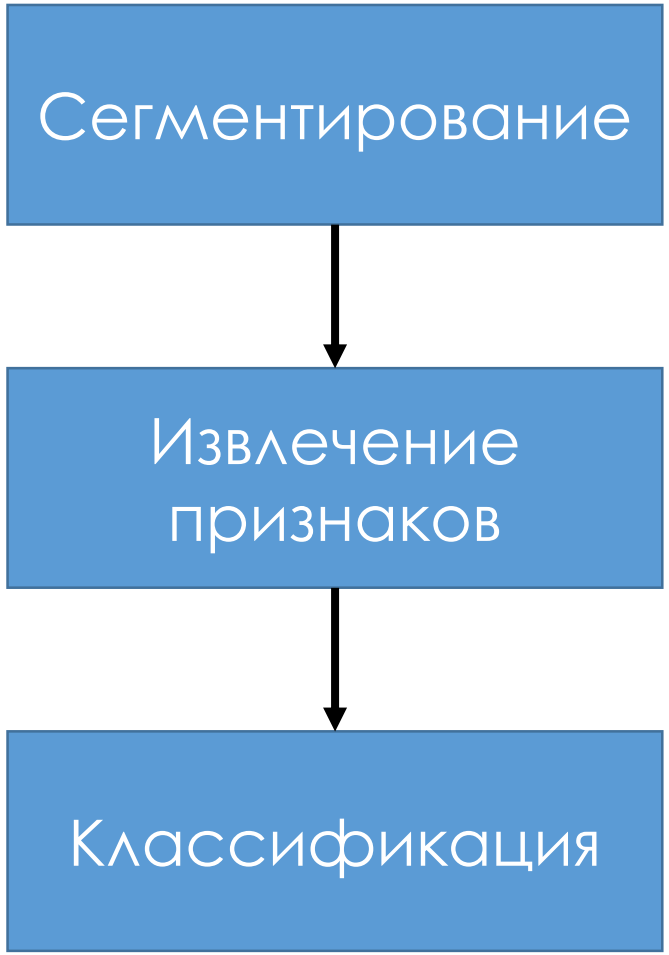
0.2 0.1 **0.7**

Вероятности принадлежности изображения к одному из трёх классов.
В сумме равны единице.

Классификация изображений

```
class Expert : I3ClassTumorClassifier
{
    public Tuple<float, float, float> Predict(Image image)
    {
        ???
    }
}
```

Классификация изображений



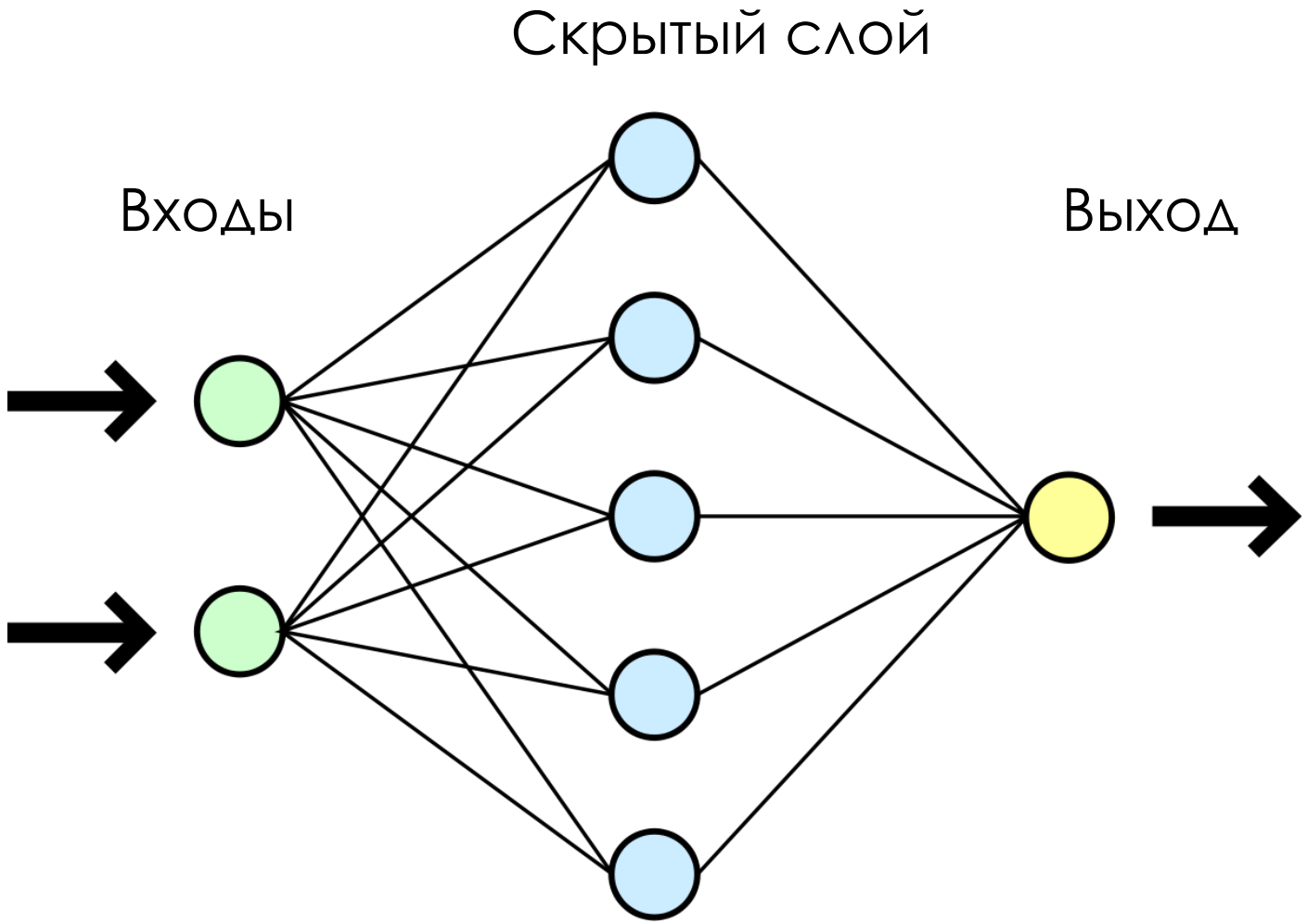
На что смотрим

Как представляем

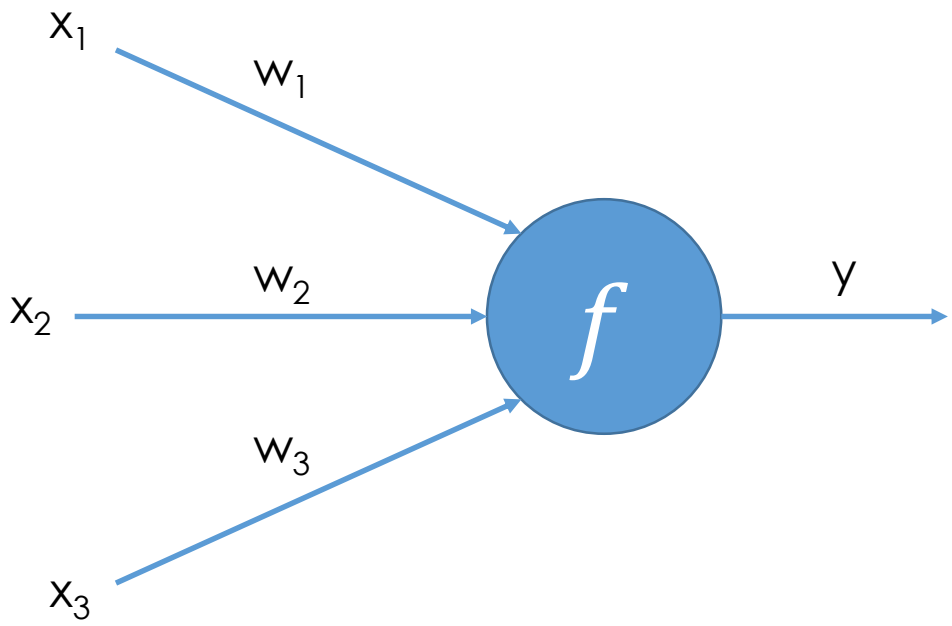
К чему относится


float[]
Нейросеть
float[]

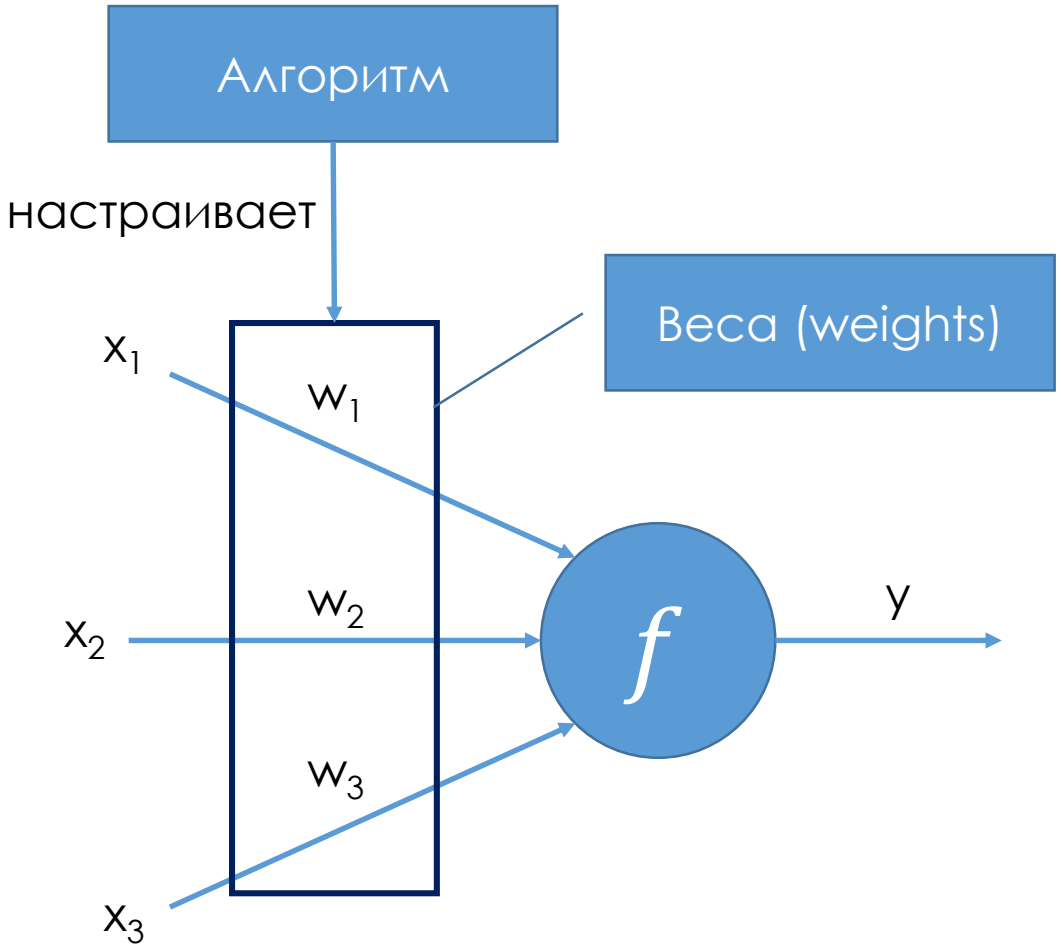
Перцептрон



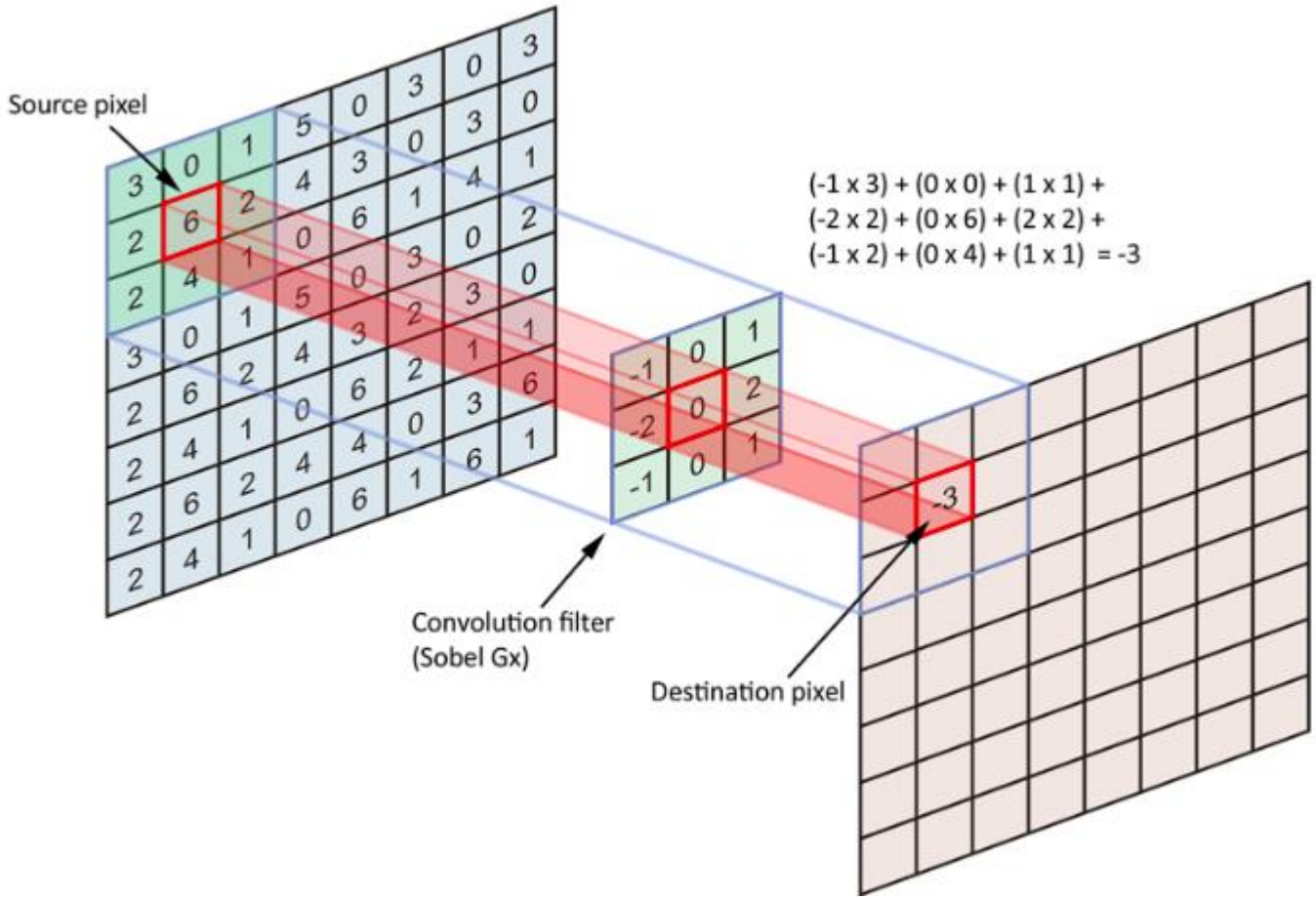
Модель нейрона



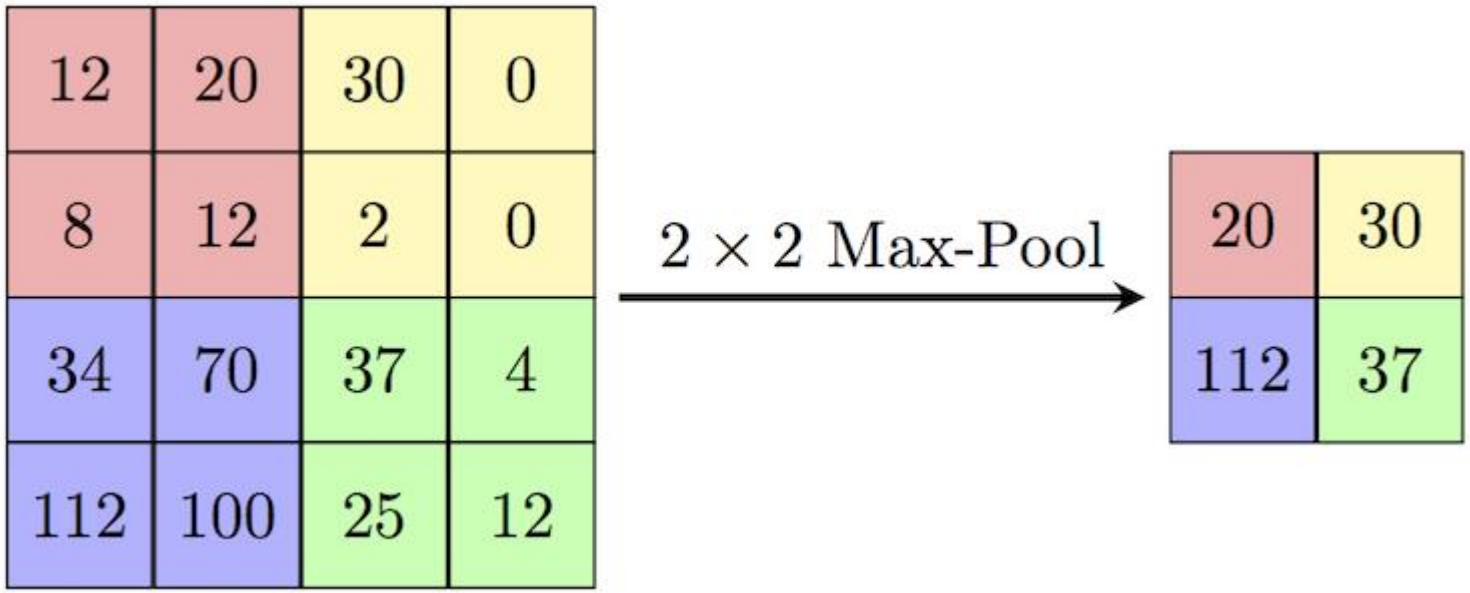
Модель нейрона



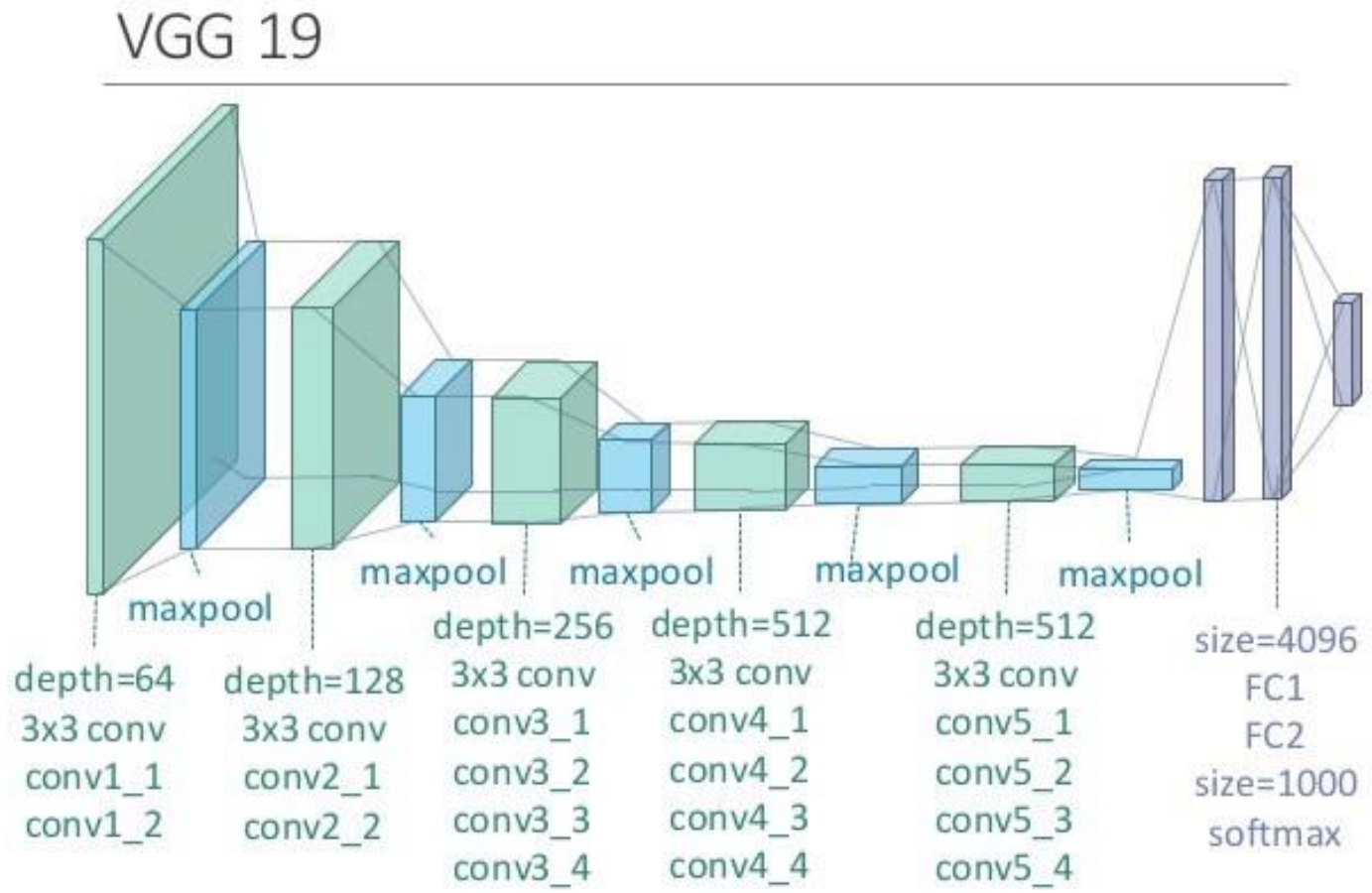
Свёрточная нейросеть / Свёртка



Свёрточная нейросеть / Субдискретизация



Свёрточная нейросеть

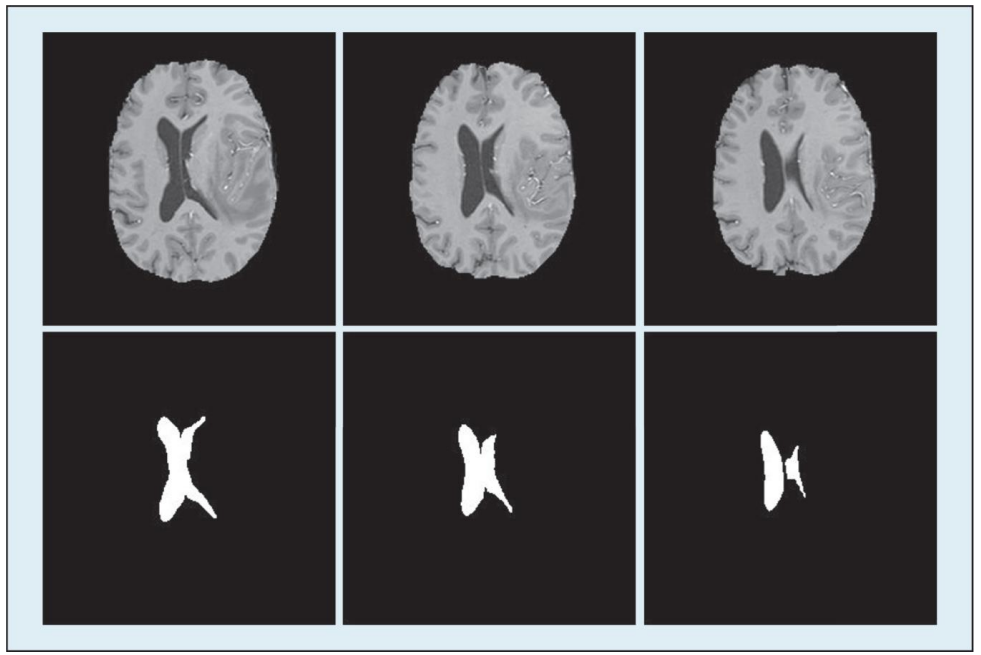


Пример – компьютерная диагностика

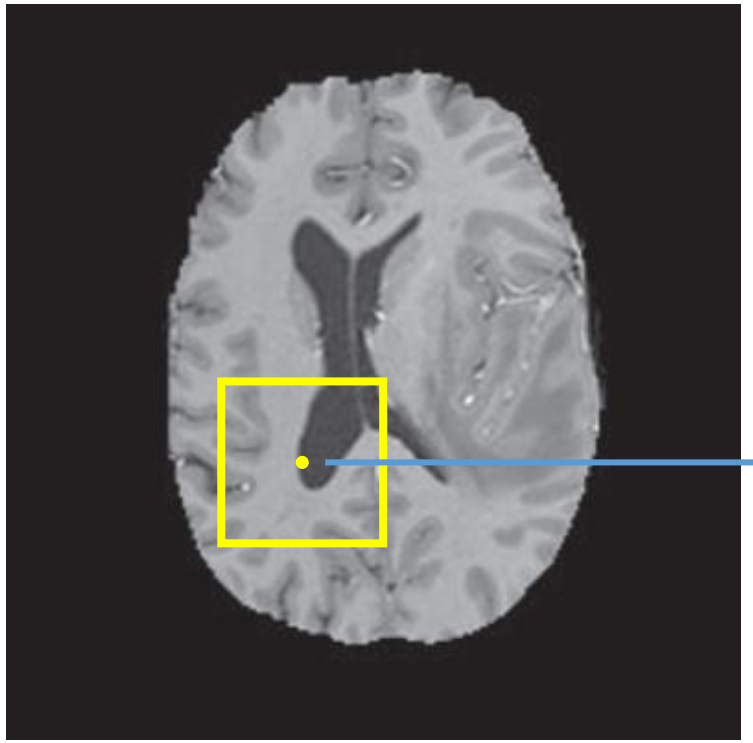
- Сегментирование: на что именно смотрим?
- Извлечение признаков: как представляем?
- Классификация: к какому классу относится представление?

```
class Mask
{
    private float [,] mask;
}

interface ISegmenter
{
    Mask GetMask(Image image);
}
```

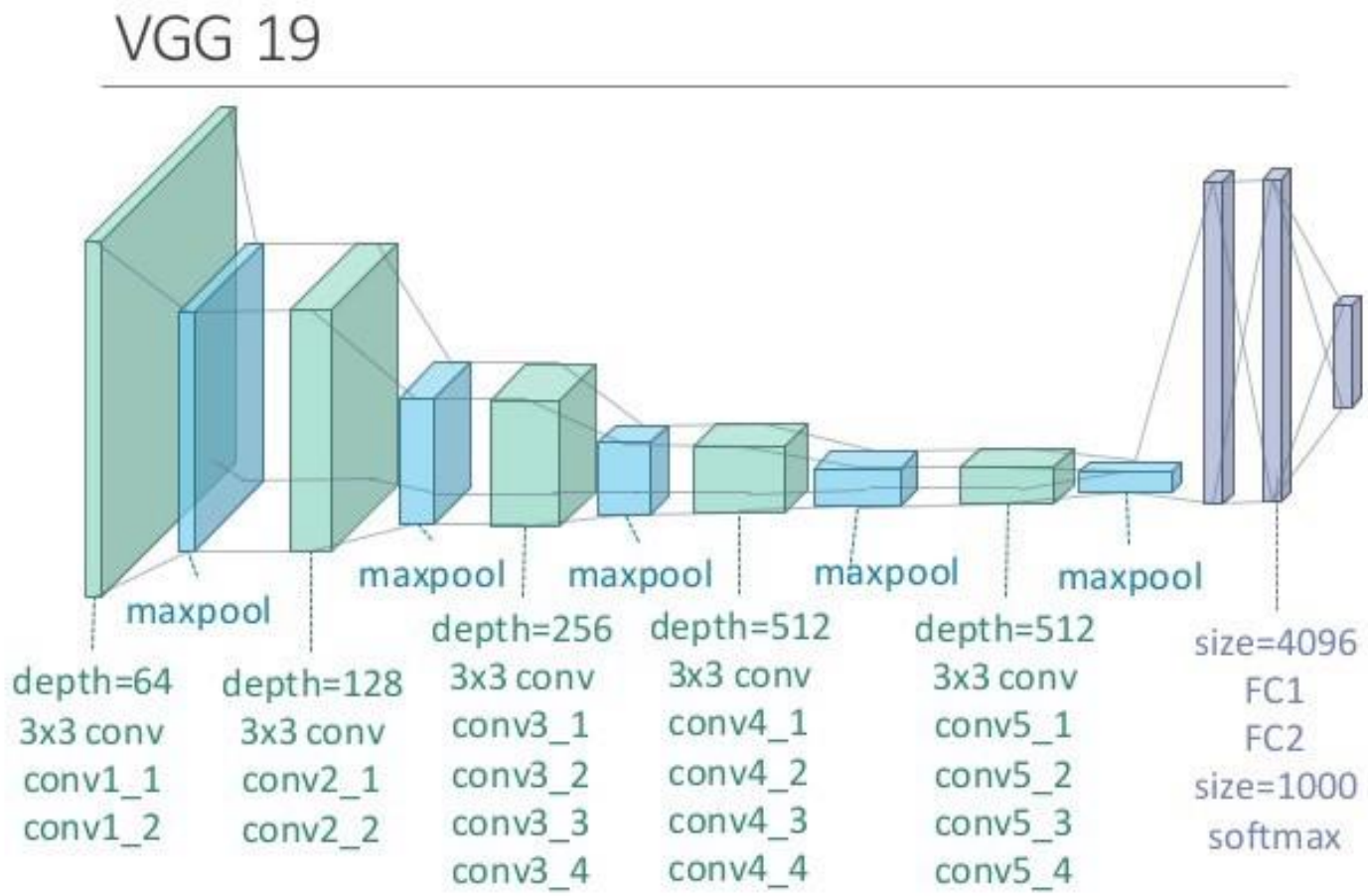


Классификация пикселей

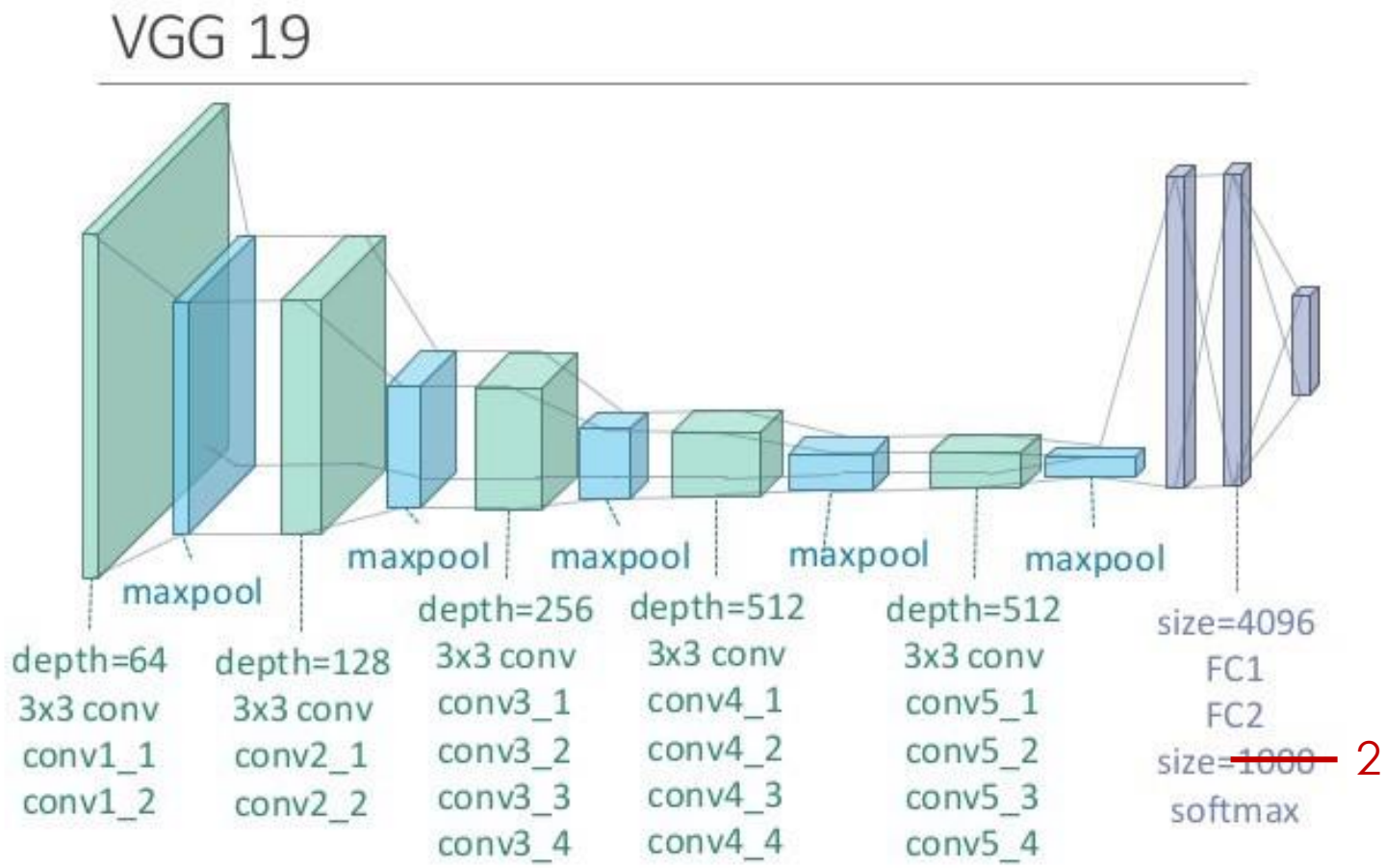


0.2	0.8

Классификация пикселей

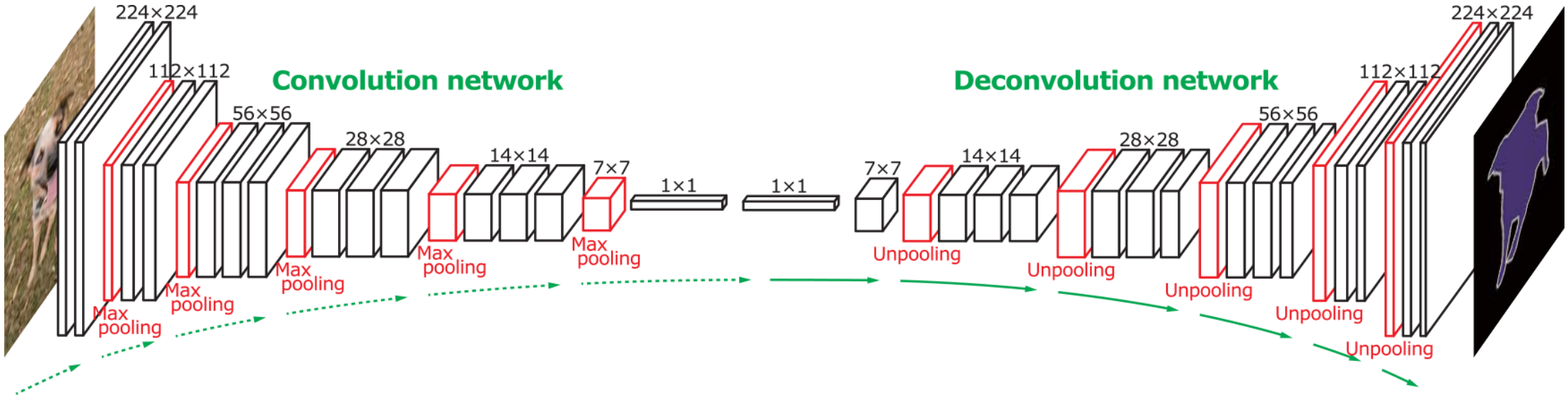


Классификация пикселей



Классификация пикселей

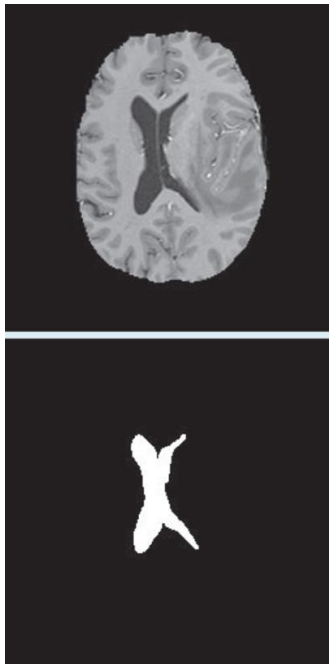
Fully Convolutional Network

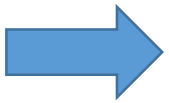


http://cvlab.postech.ac.kr/research/deconvnet/

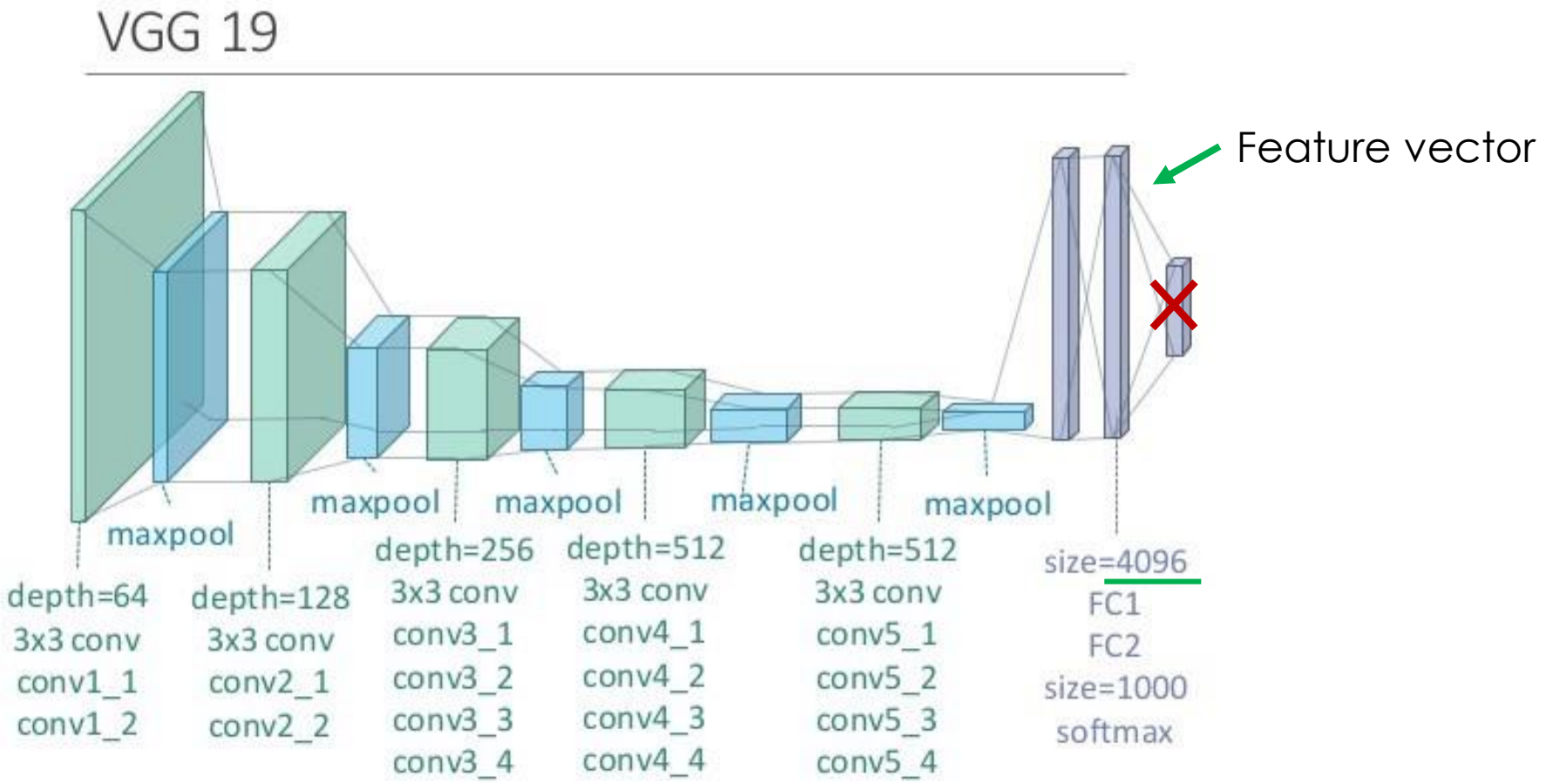
Пример – компьютерная диагностика

- Сегментирование: на что именно смотрим?
- Извлечение признаков: как представляем?
- Классификация: к какому классу относится представление?



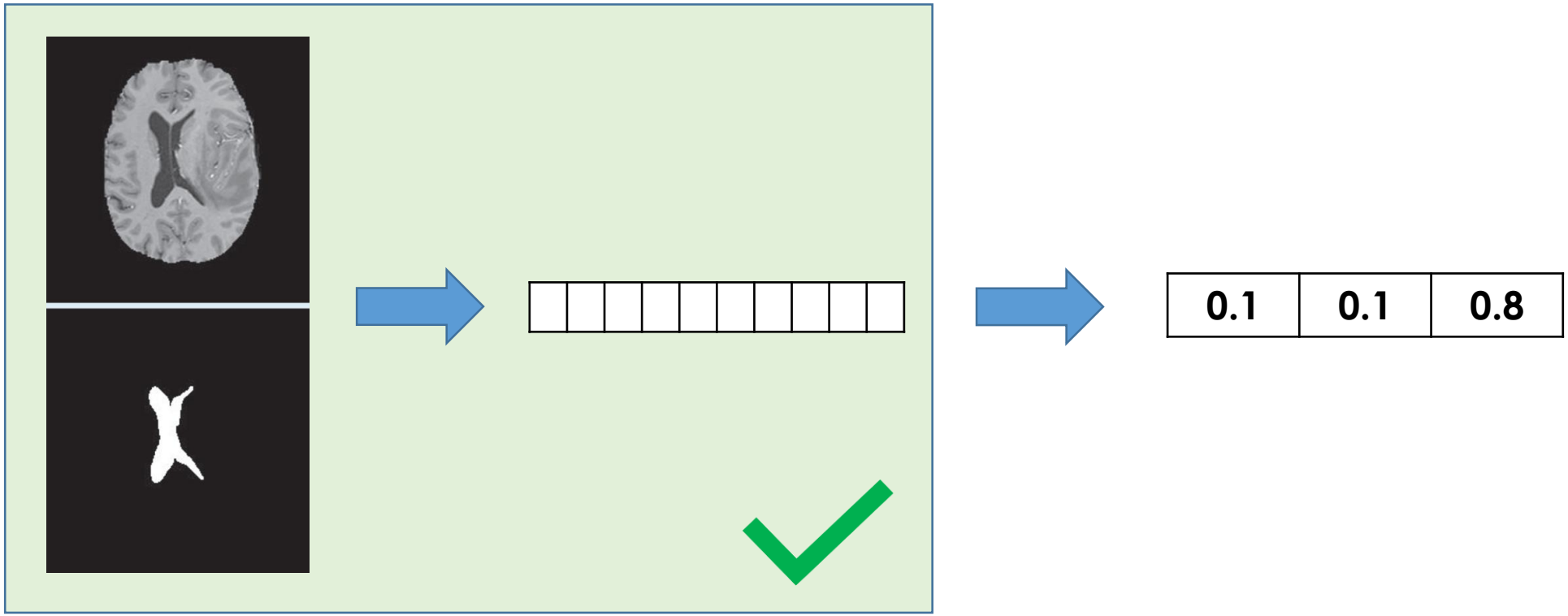


Пример – компьютерная диагностика



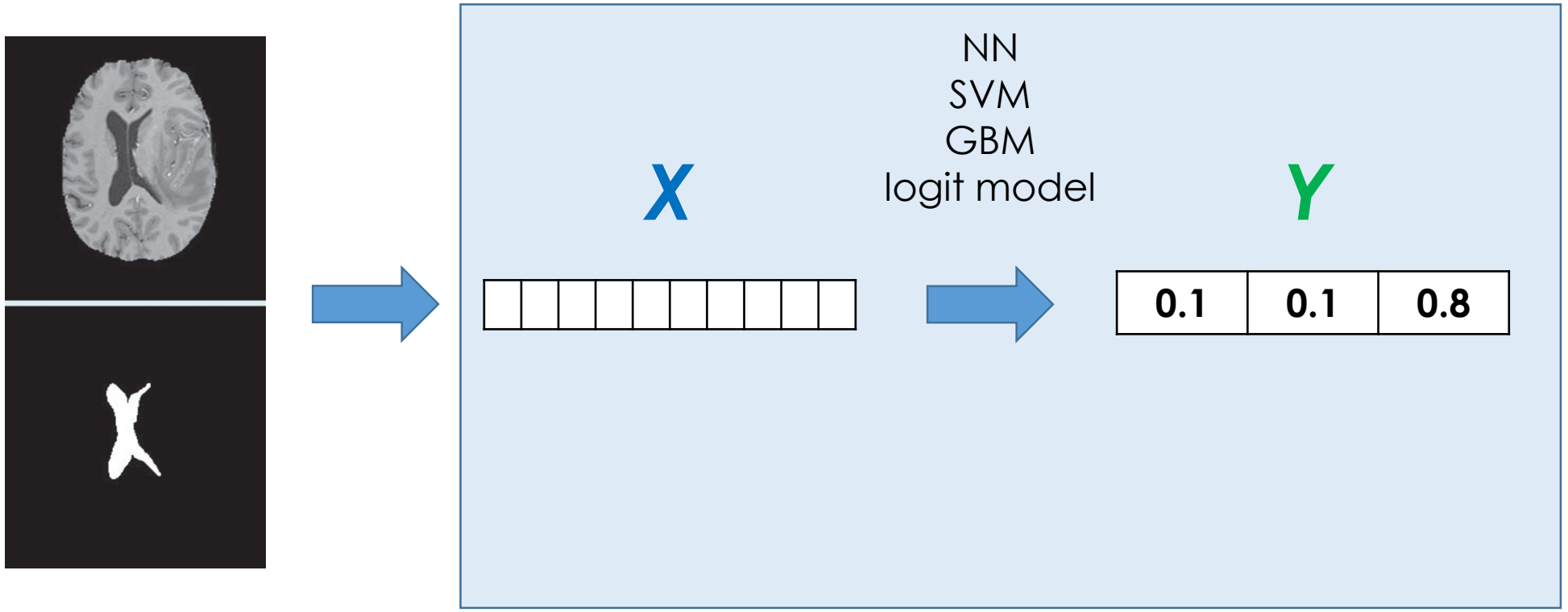
Пример – компьютерная диагностика

- Сегментирование: на что именно смотрим?
- Извлечение признаков: как представляем?
- Классификация: к какому классу относится представление?



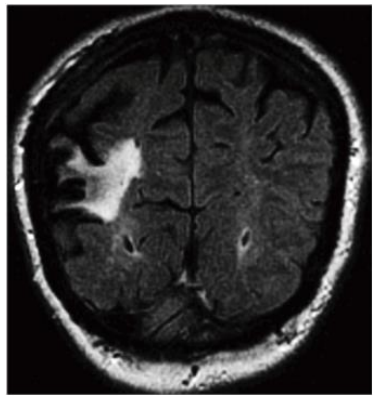
Пример – компьютерная диагностика

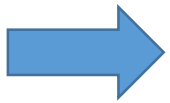
- Сегментирование: на что именно смотрим?
- Извлечение признаков: как представляем?
- Классификация: к какому классу относится представление?

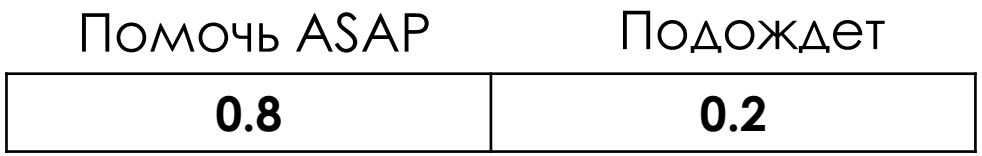


Пример – компьютерная диагностика

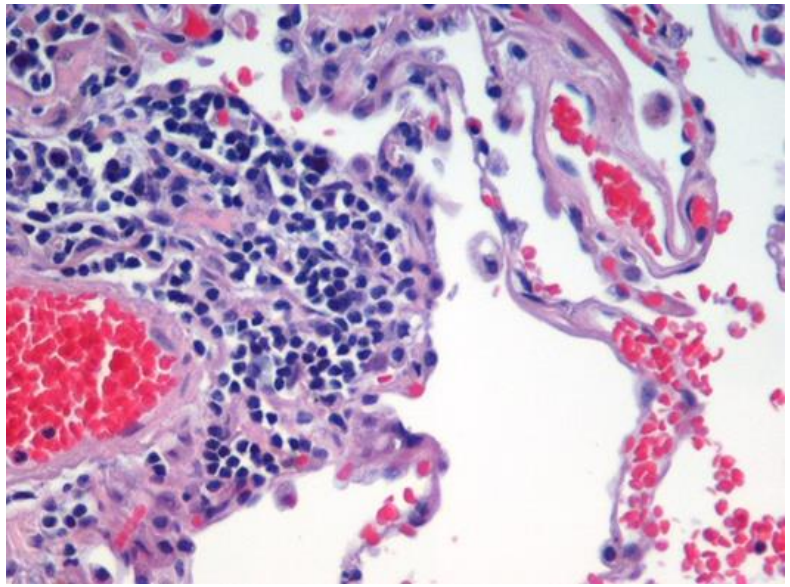
- Более простая форма – триаж
  - Классы – степени критичности
  - До поступления к эксперту







Пример – компьютерная диагностика



0.1	0.1	0.8

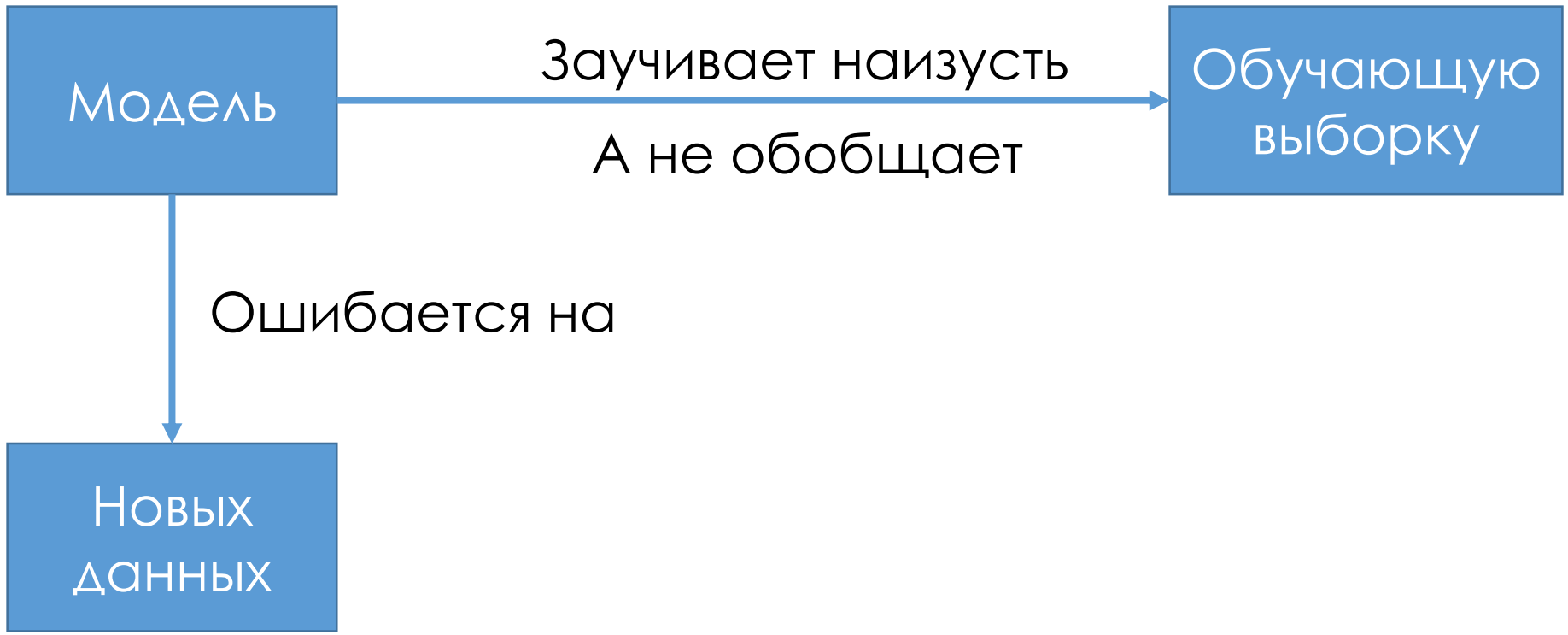
0.1	0.2	0.7

0.2	0.7	0.1

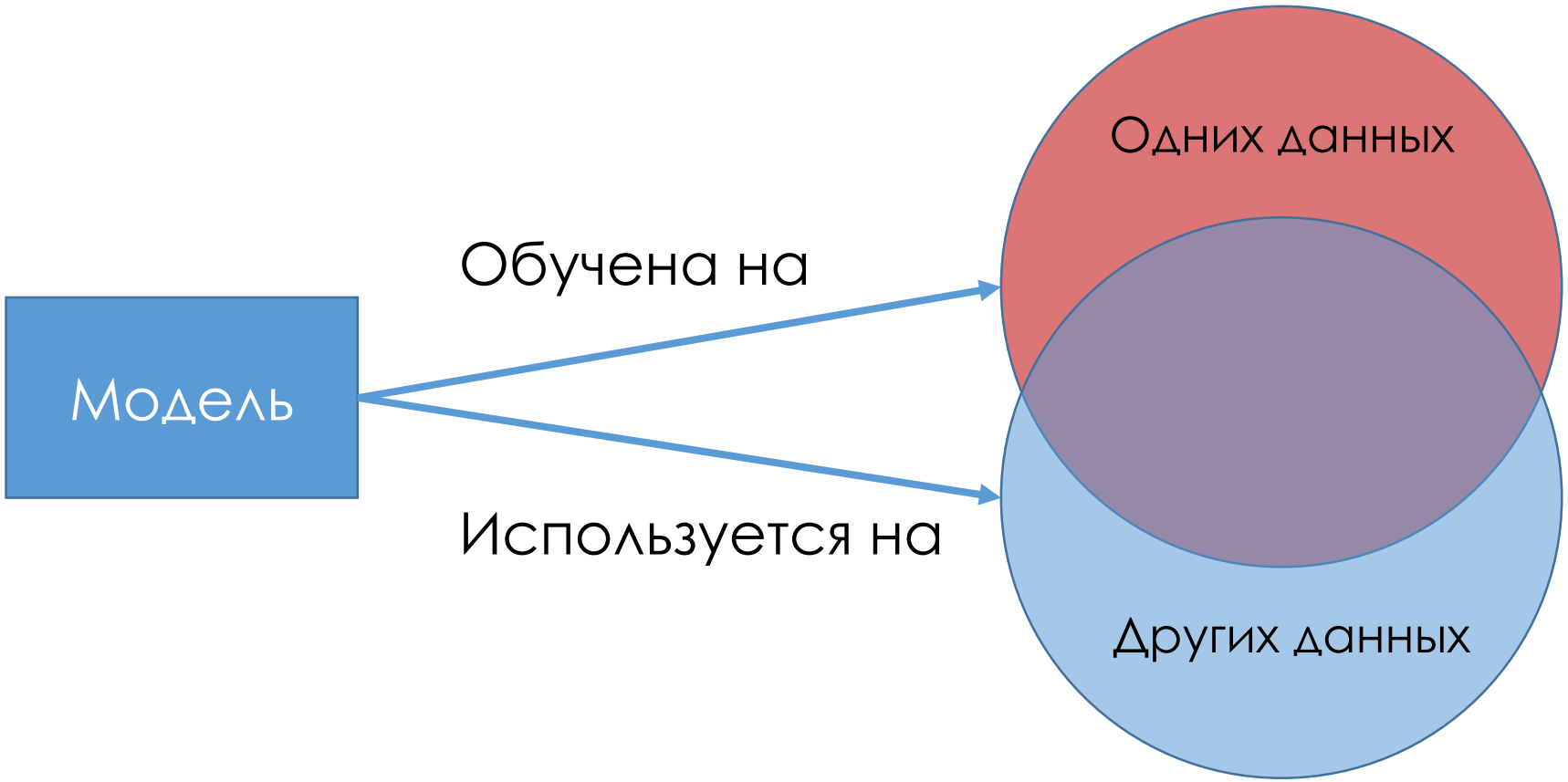
(2004) T Cells Cause Lung Damage in Emphysema. PLoS Med 1(1): e25. https://doi.org/10.1371/journal.pmed.0010025

arcadia
software country

Проблемы: переобучение



Проблемы: использование не по назначению



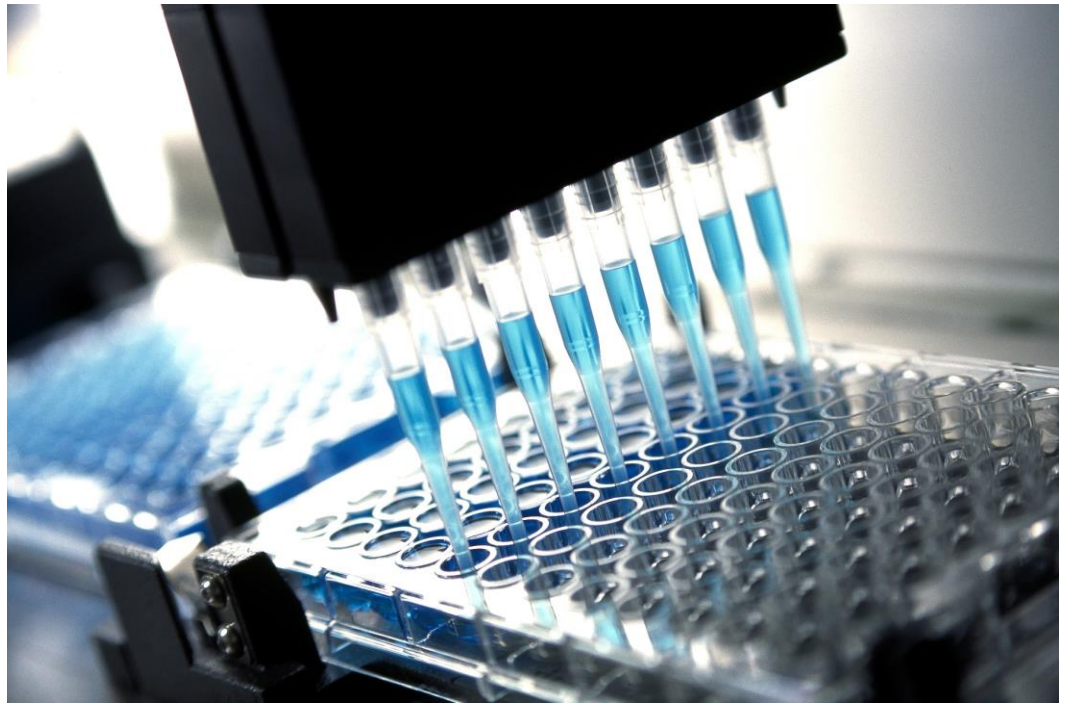
Пример – разработка лекарств

```
var candidateCompound = Lab.GetNextCompound();
var bindResult = candidateCompound.TryBindTo(Body.SomeTargetProtein);

if (!bindResult.Successful)
{
    Console.WriteLine("Time and money has been wasted.");
}
```

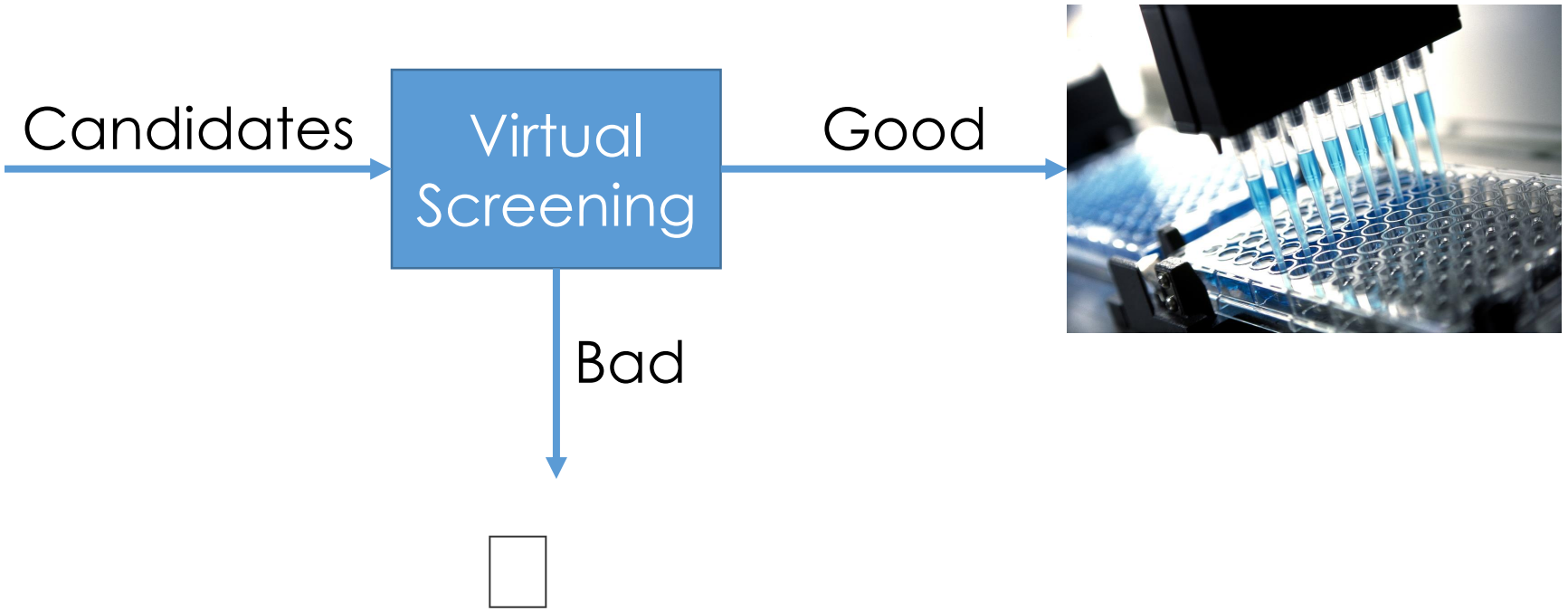
Пример – разработка лекарств

- Найти действующее вещество с нужными свойствами
  - Связывается с правильным белком
  - Не связывается с неправильными
- Перебирать все – слишком долго и дорого (10^{60} веществ)

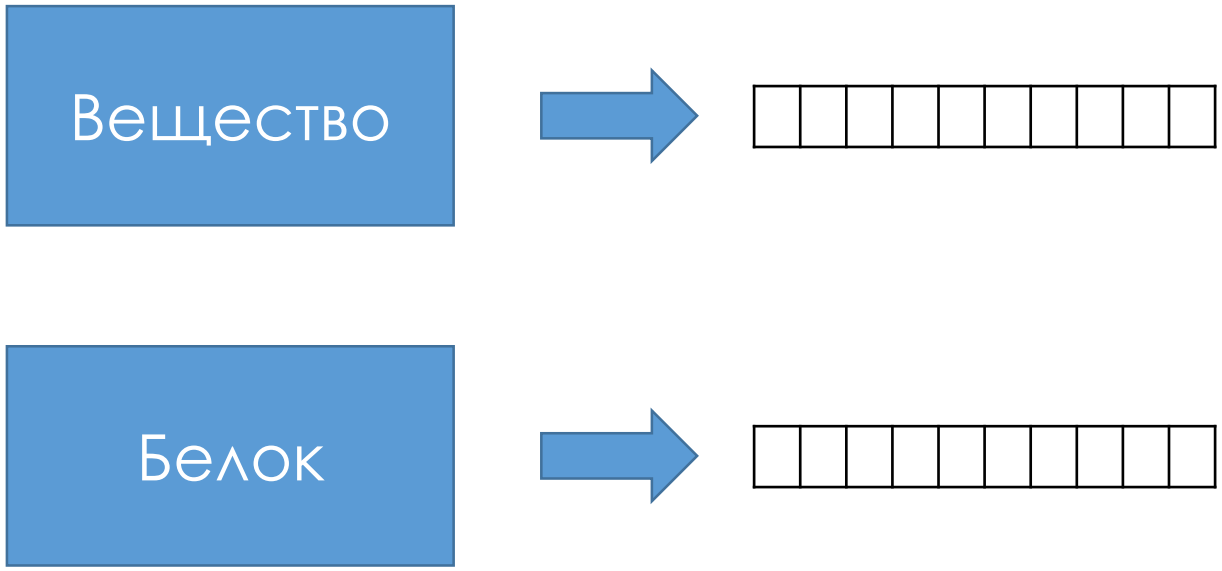


Пример – разработка лекарств

- Предсказать результат до эксперимента
  - Отбросить обречённые вещества
  - Сосредоточиться на перспективных



Пример – разработка лекарств



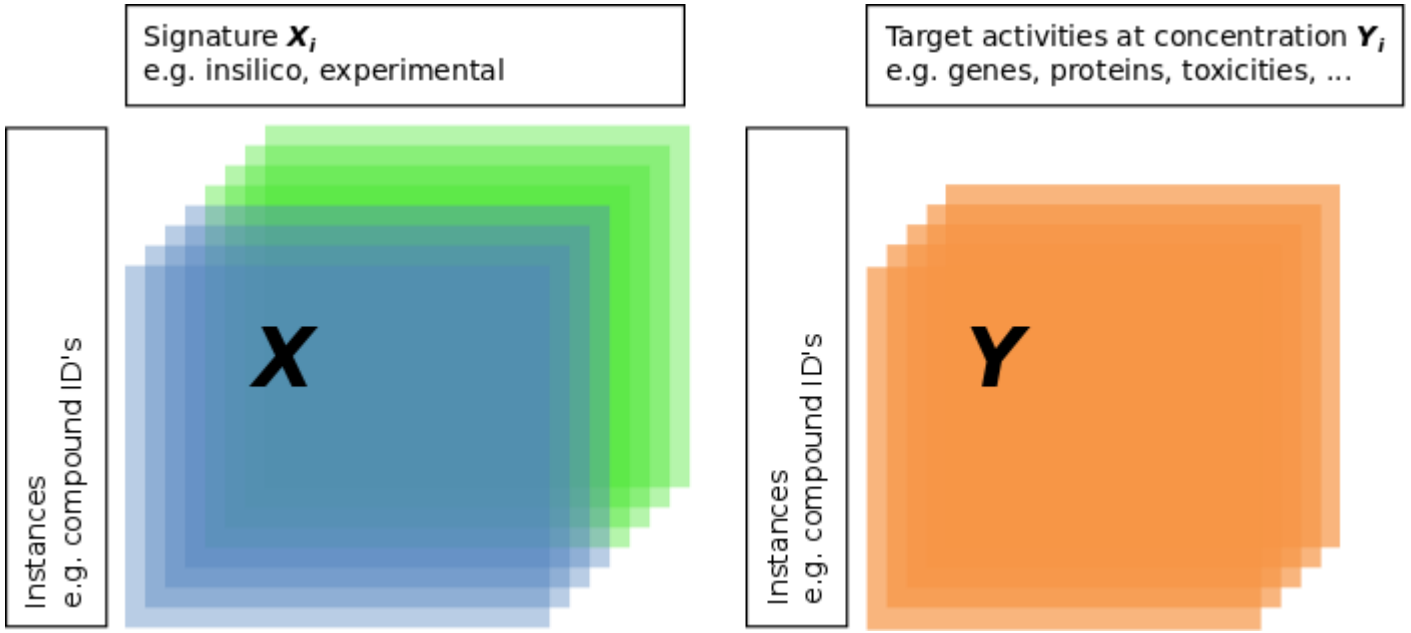
Пример – разработка лекарств

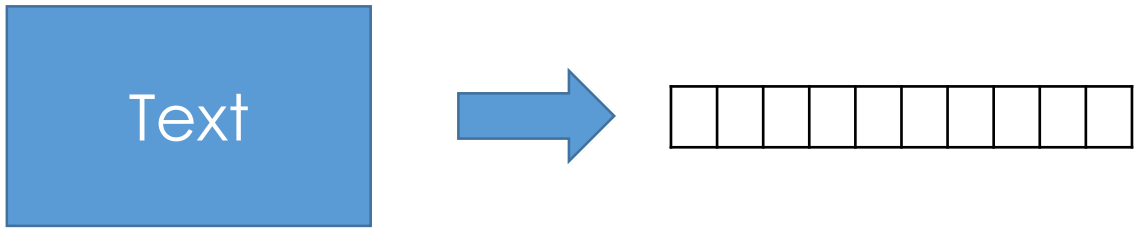
- Biosignature Based Drug Design
  - Известна способность некоторых веществ связываться с некоторыми белками
  - Предсказать эту способность для других комбинаций вещество-белок

	Белок 1	Белок 2	…	Белок m
Вещество 1	Модель 1	Модель 2	…	Модель m
Вещество 2				
…				
Вещество n				

Пример – разработка лекарств

- Biosignature Based Drug Design
  - Известна способность некоторых веществ связываться с некоторыми белками
  - Предсказать эту способность для других комбинаций вещество-белок




Text

Интеллектуальный анализ текста (Text mining)

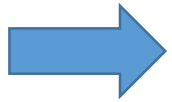
- Классификация документов
- Кластеризация документов
- Извлечение именованных сущностей
- Анализ тональности
- Суммаризация текста
- Извлечение отношений

Тематическое моделирование

Document 1

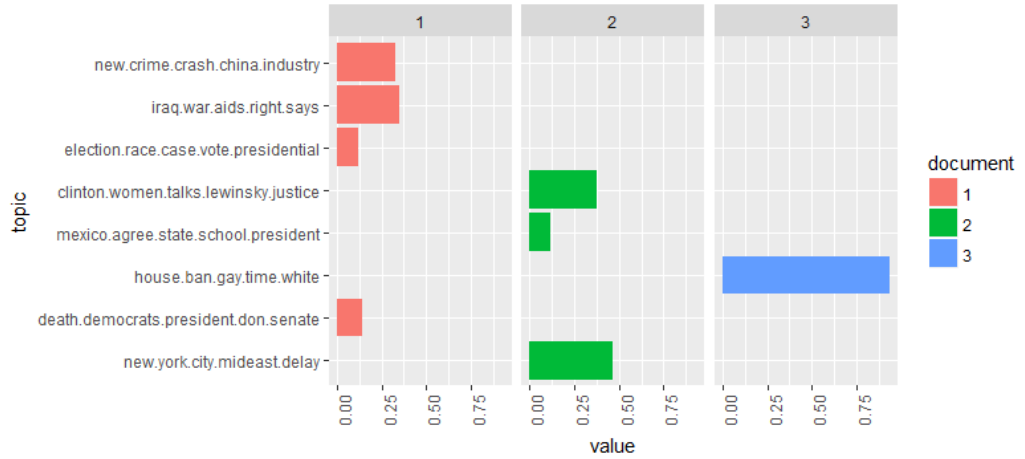
…

Document n



Filter

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	new	death	house	mexico	democratic
2	york	democrats	ban	agree	election
3	city	president	gay	state	democrats
4	mideast	don	time	school	party
5	delay	senate	white	president	bloomberg
6	care	penalty	kills	kansas	use
7	health	resigns	china	system	issue
8	texas	study	class	cut	presidential
9	ethics	people	saudi	pakistan	air
10	state	cancer	web	senator	candidates



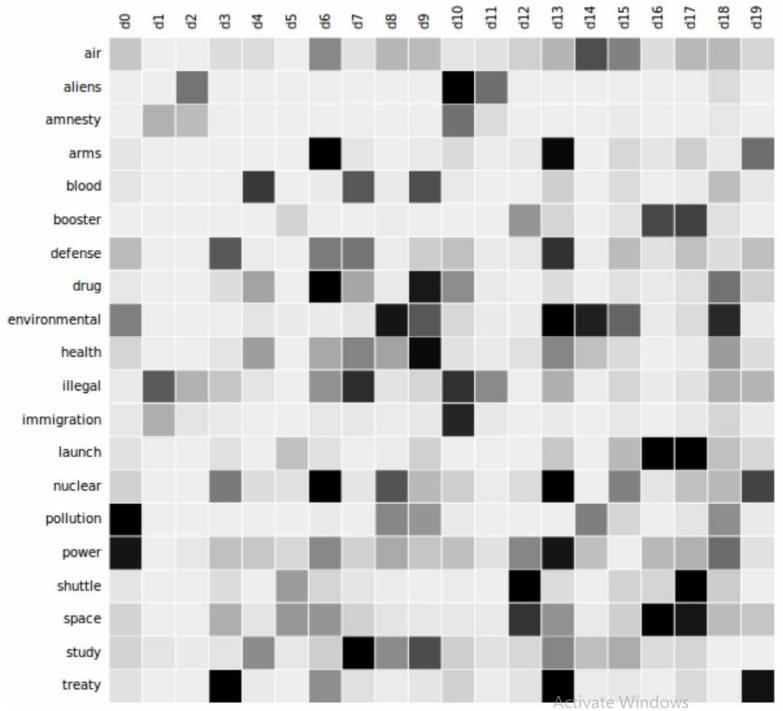
Представление документов

Document 1

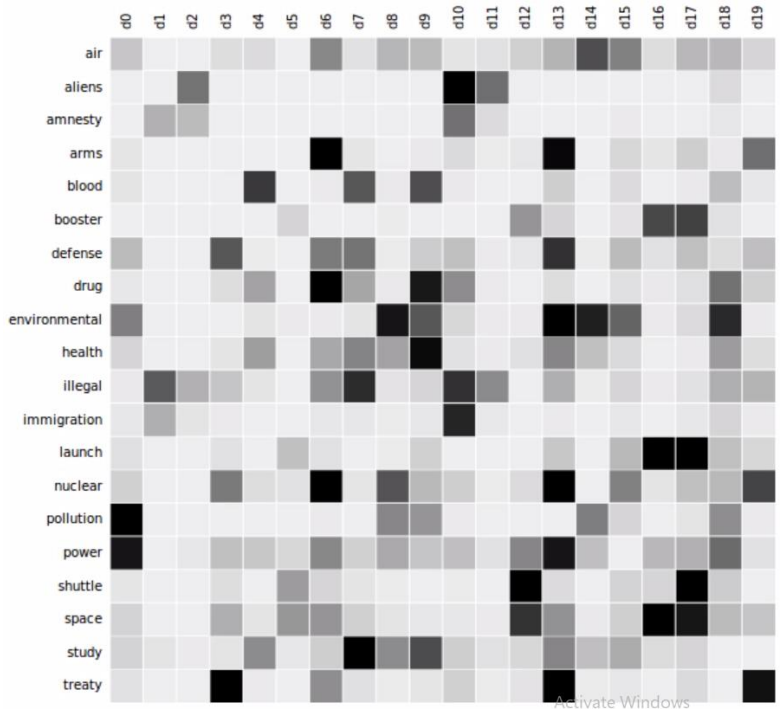
…

Document n

Document-term matrix



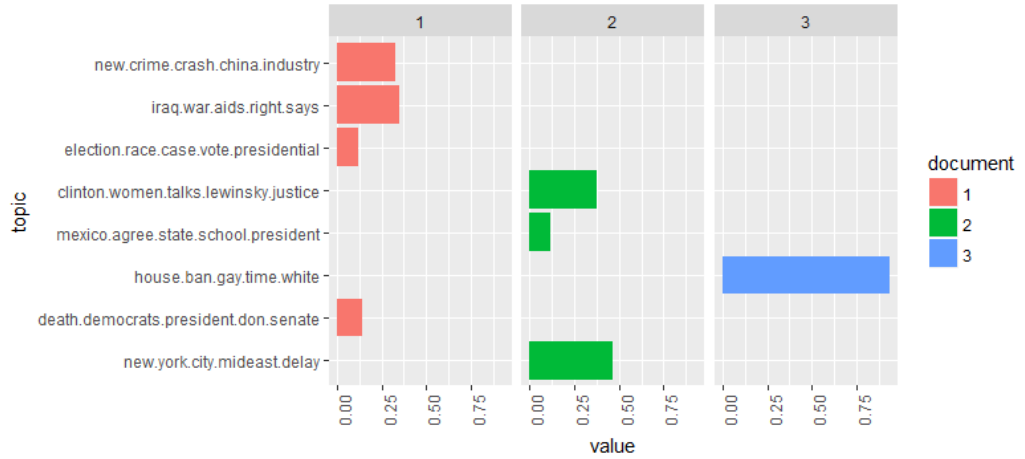
Тематическое моделирование



Latent Dirichlet Allocation

Filter

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	new	death	house	mexico	democratic
2	york	democrats	ban	agree	election
3	city	president	gay	state	democrats
4	mideast	don	time	school	party
5	delay	senate	white	president	bloomberg
6	care	penalty	kills	kansas	use
7	health	resigns	china	system	issue
8	texas	study	class	cut	presidential
9	ethics	people	saudi	pakistan	air
10	state	cancer	web	senator	candidates



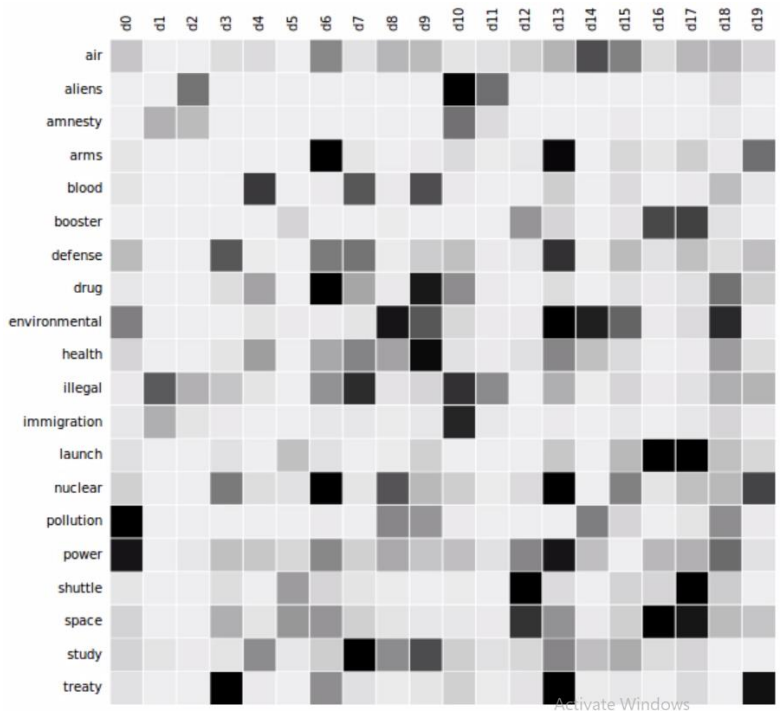
arcadia software country

Тематическое моделирование

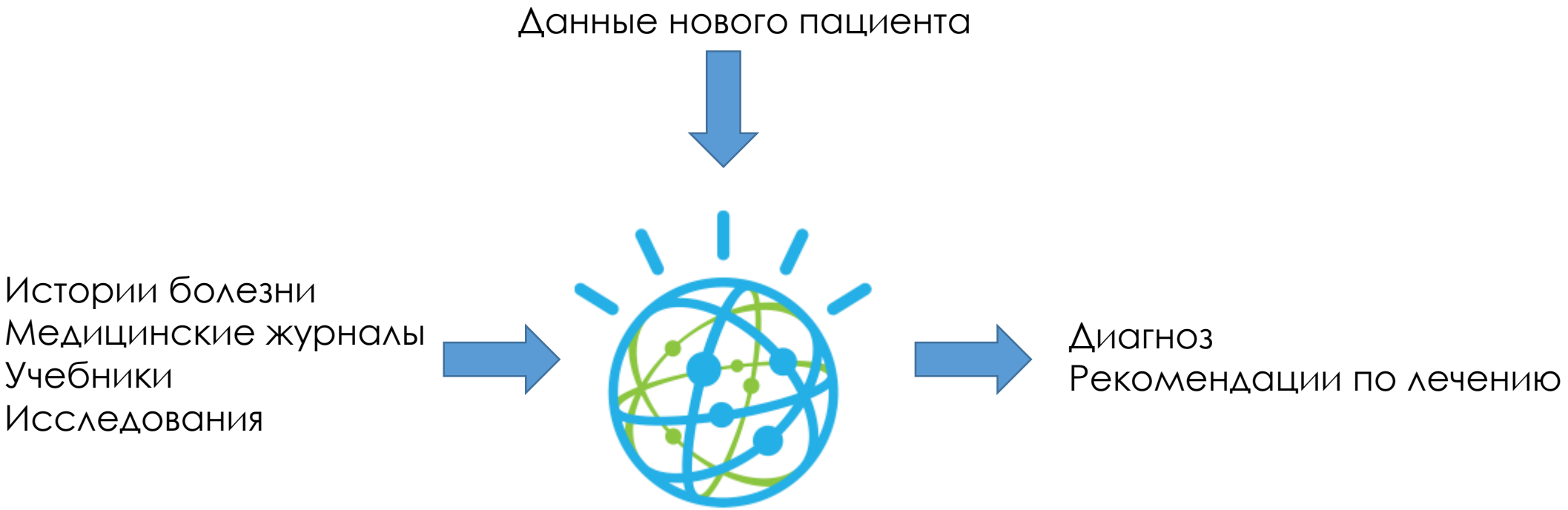
Corrective Action
Preventive Action

CAPA 1
…
CAPA n

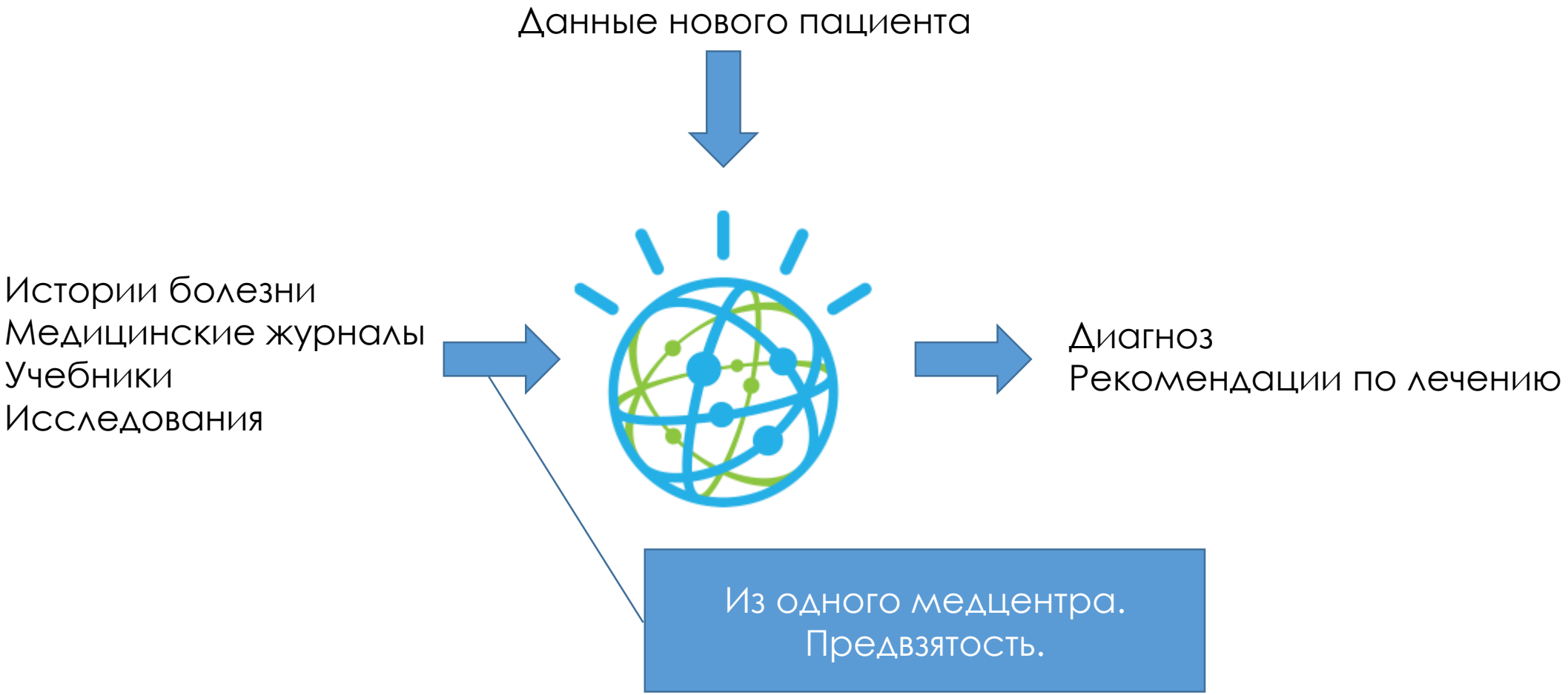
Document-term matrix



IBM Watson for Oncology



IBM Watson for Oncology



Инструмент

- NumPy, SciPy, Pandas
- StatsModels
- Matplotlib, Seaborn, Plotly, Bokeh
- Scikit-learn
- XGBoost, LightGBM
- TensorFlow, Keras

- dplyr, readr, data.table, xts
- ggplot2, plotly
- caret
- gbm, XGBoost, randomForest
- TensorFlow, Keras

Инструмент

- ML.NET - https://github.com/dotnet/machinelearning/
- Microsoft Cognitive Toolkit - https://www.microsoft.com/en-us/cognitive-toolkit/
- Microsoft Cognitive Services - https://azure.microsoft.com/en-us/services/cognitive-services/

- Accord.NET - https://github.com/accord-net/framework
- Encog - https://github.com/encog/encog-dotnet-core
- numl - https://github.com/sethjuarez/numl
- SharpLearning - https://github.com/mdabros/SharpLearning

Инструмент

- Линейная алгебра
- Математическая статистика
- Методы оптимизации
- Структуры данных
- Теория вероятностей
- Многомерный анализ

- machinelearning.ru
- Курс «Машинное обучение» 2014 – К.В. Воронцов.

- https://www.youtube.com/results?search_query=машинное+обучение

Вопросы