# КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Маша Шеянова, masha.shejanova@gmail.com
Саша Ершова, asershova@edu.hse.ru

December 16, 2016

НИУ ВШЭ

# Что такое ”компьютерная лингвистика”?

Есть лингвистика. Есть компьютеры. Что хорошего можно с этим сделать?

1. Можно делать корпуса и вспомогательные инструменты для теоретических лингвистов.
2. Computational linguistics: изучение языка при помощи формальных математических моделей, статистики и всего такого.
3. Natural language processing: автоматическое извлечение чего-нибудь из текста и автоматическое его порождение.

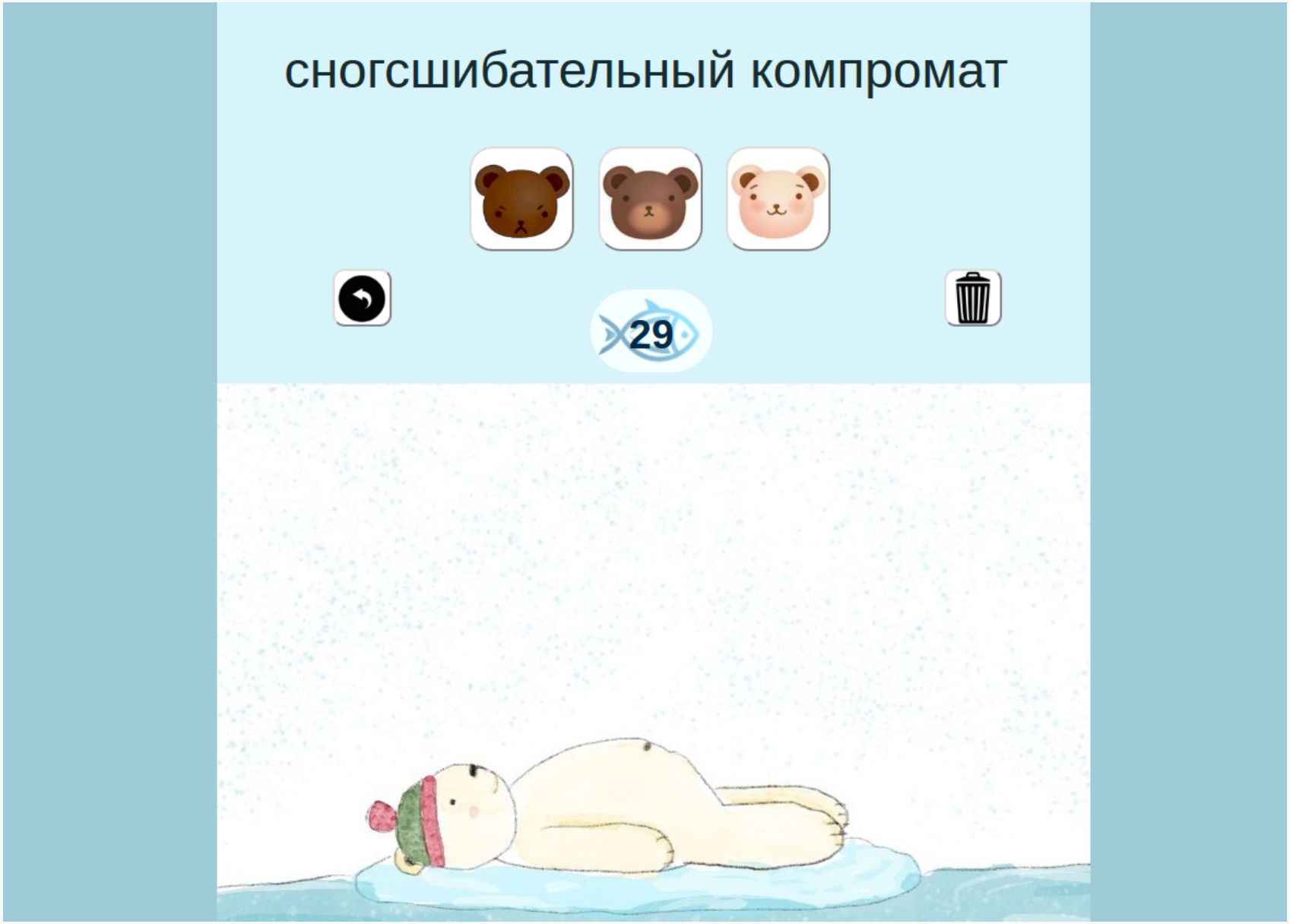
NB: 2 и 3 — очень разные вещи, хотя и то, и другое в русском называют ”компьютерной лингвистикой”

# Вспомогательные инструменты

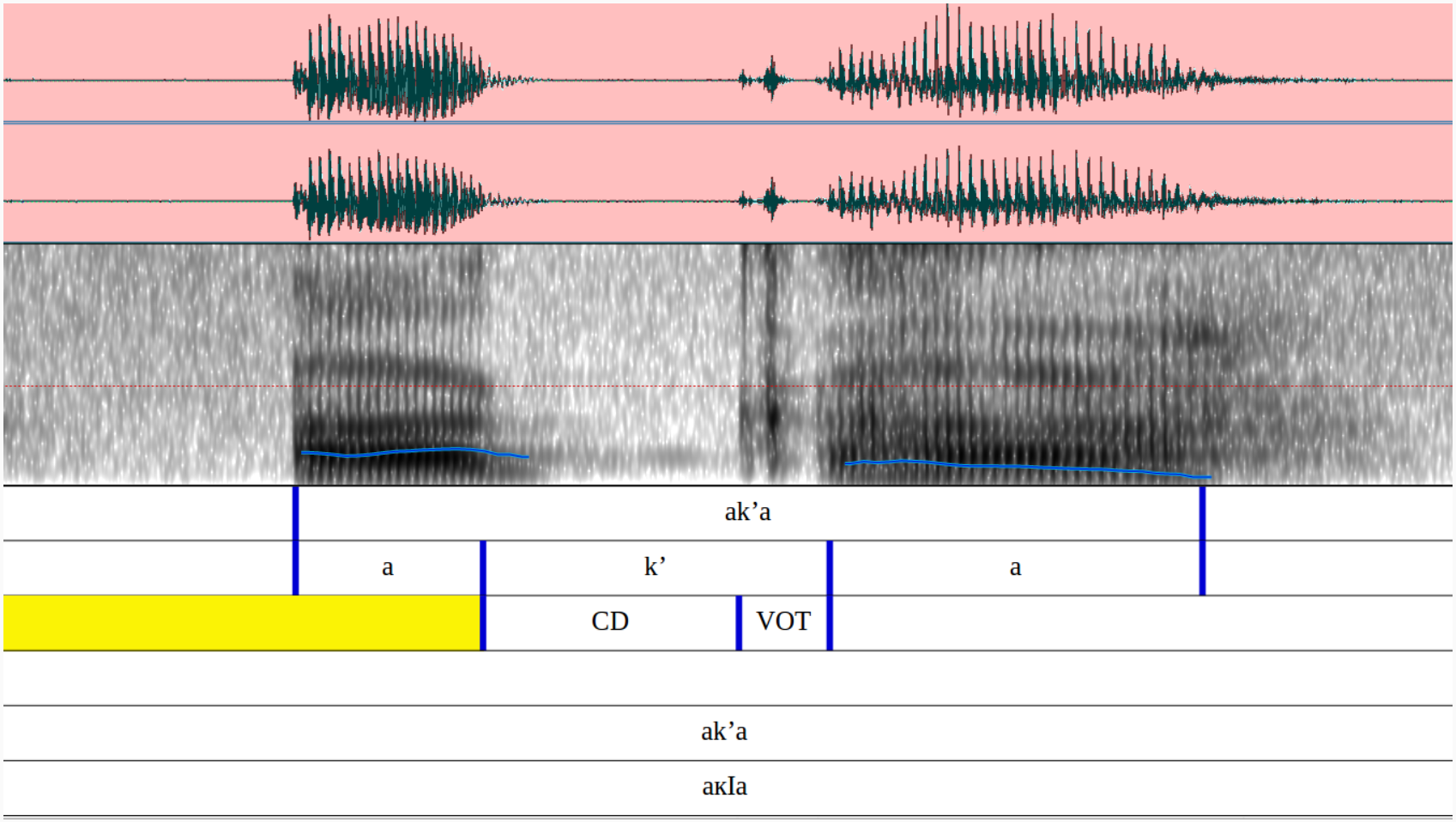
# какие они бывают?

- Корпуса
- Словари
- Инструменты сбора данных
- Программы для анализа данных (анализ звука: Praat, анализ морфологии: Fieldworks)

# СБОР ДАННЫХ. SENTI GAME.



# АНАЛИЗ ДАННЫХ. PRAAT.



# COMPUTATIONAL LINGUISTICS

# что сюда входит?

В принципе, это любые лингвистические исследования, где нужно что-то посчитать, например:

- посмотреть, от чего возникают дырки в парадигмах
- доказать, что вид в русском — это континуум
- посмотреть, какие слова ближе по значению, а какие дальше (этим умеет заниматься дистрибутивная семантика)

# ДИСТРИБУТИВНАЯ СЕМАНТИКА. ЧТО ЭТО.

Что мы хотим:

- формальный способ считать лексическую близость
- глобально: научить компьютер извлекать смыслы из текста

Как делать это автоматически?

Дистрибутивная гипотеза: значения слов полностью определяются их контекстами. Слова с похожими типичными контекстами имеют схожее значение.

# ДИСТРИБУТИВНАЯ СЕМАНТИКА. КАК ЭТО РАБОТАЕТ.

Нам нужно:

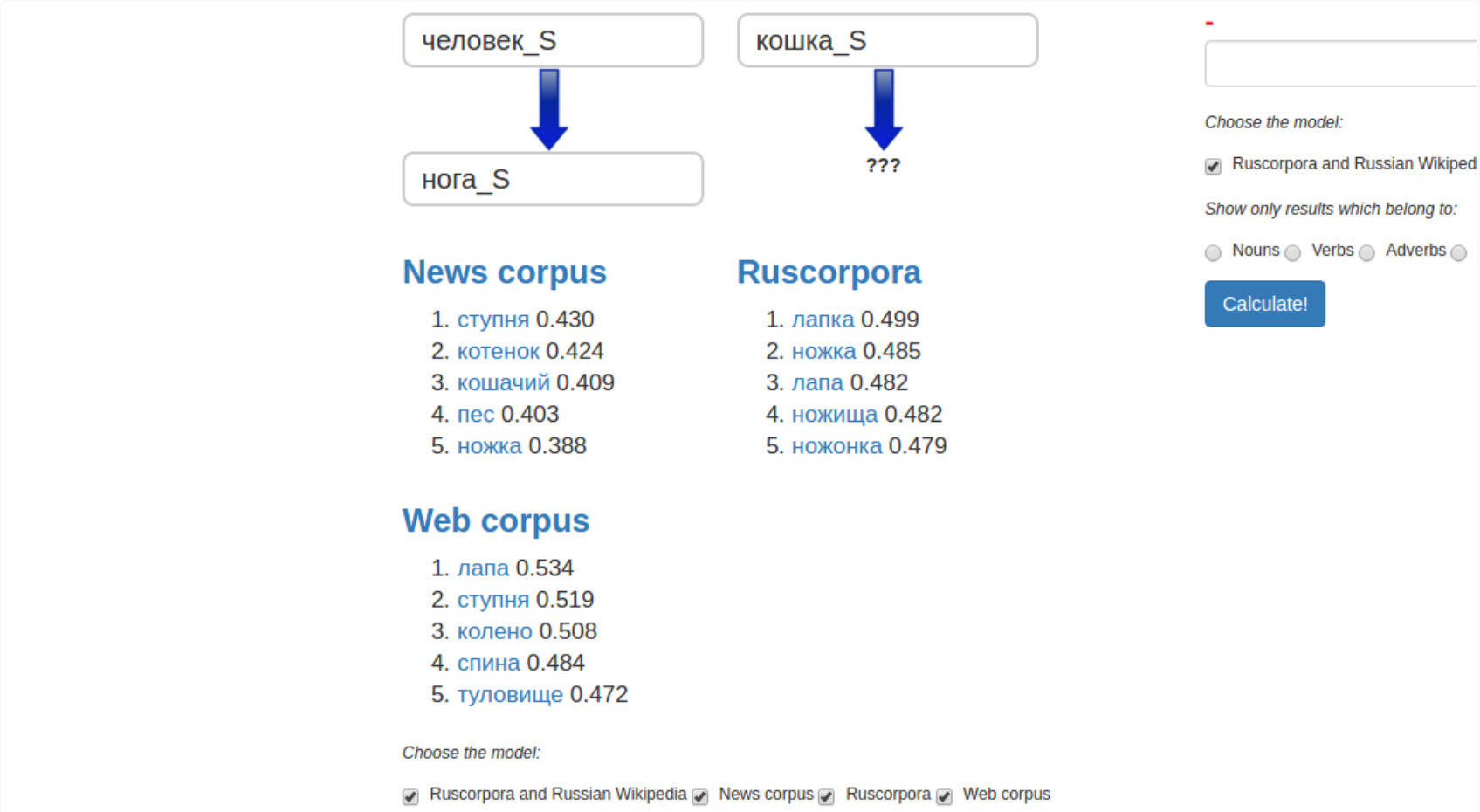
- много текстов, чтобы картинка была репрезентативной
- посчитать в этих текстах взаимную встречаемость слов друг с другом
- найти слова, которые могут заменить друг друга и слова, у которых нет общих контекстов

Готово! Мы прекрасны и можем

- находить слова, близкие по значению к данному
- строить семантические пропорции
- строить семантические визуализации

# ДИСТРИБУТИВНАЯ СЕМАНТИКА. ЭТО РАБОТАЕТ!

На rusvectores можно найти слова, наиболее близкие к данному, построить семантическую пропорцию и многое другое.



# NATURAL LANGUAGE PROCESSING

# ПРИМЕРЫ

- Спеллчекеры
- Машинный перевод
- Text mining
- Speech recognition и OCR
- Когнитивные технологии: боты, weak AI, seq2seq-нейросети

# Машинный перевод. Corpus-based.

У нас есть параллельные корпуса, то есть корпуса, где каждое предложение одного языка сопоставлено с предложением другого. С их помощью мы учим компьютер переводить предложения пользователя.

| Английский | Японский |
|---|---|
| How much is that red umbrella? | Ano akai kasa wa ikura desu ka. |
| How much is that small camera? | Ano chiisai kamera wa ikura desu ka. |

Corpus-based бывает:

- Statistical
- Example-based

# Машинный перевод. Rule-based.

Параллельные корпуса не используются. Часть информации хранится в словарях, часть прописана в правилах.

Как это работает в Apertium

- словари:
  - билингвальные: лексические соответствия
  - монолингвальные: парадигмы
- правила:
  - лексический выбор: сложно → difficult, complicated или complex?
  - разрешение морфологической омонимии
  - изменение структуры

# Машинный перевод. Rule-based vs. Corpus-based.

Corpus-based:

- широко используется сейчас (Google, Яндекс)
- требует параллельные корпуса: чем больше, тем лучше
- в принципе, не требует лингвистических знаний

Rule-based:

- сейчас всё больше уступает статистическому, НО
- может применяться при отсутствии больших корпусов → можно работать с малыми языками!
- их можно постепенно улучшать
- требует лингвистических знаний

# TEXT MINING

Автоматическое извлечение информации для:

- категоризации текстов
- информационного поиска
- извлечения информации

# Speech recognition и OCR

OCR — Optical Character Recognition — извлечение текста из картинки.

Speech recognition — извлечение текста из аудиозаписи.

Зачем нам это, если можно просто взять и послушать/почитать?

# Speech recognition и OCR

OCR — Optical Character Recognition — извлечение текста из картинки.

Speech recognition — извлечение текста из аудиозаписи.

Зачем нам это, если можно просто взять и послушать/почитать?

- **Невероятно много информации.**
- Возможность ”на лету” проделывать с извлечённым текстом ещё какие-нибудь операции.

Например, машинный перевод надписей на улице.

# AI

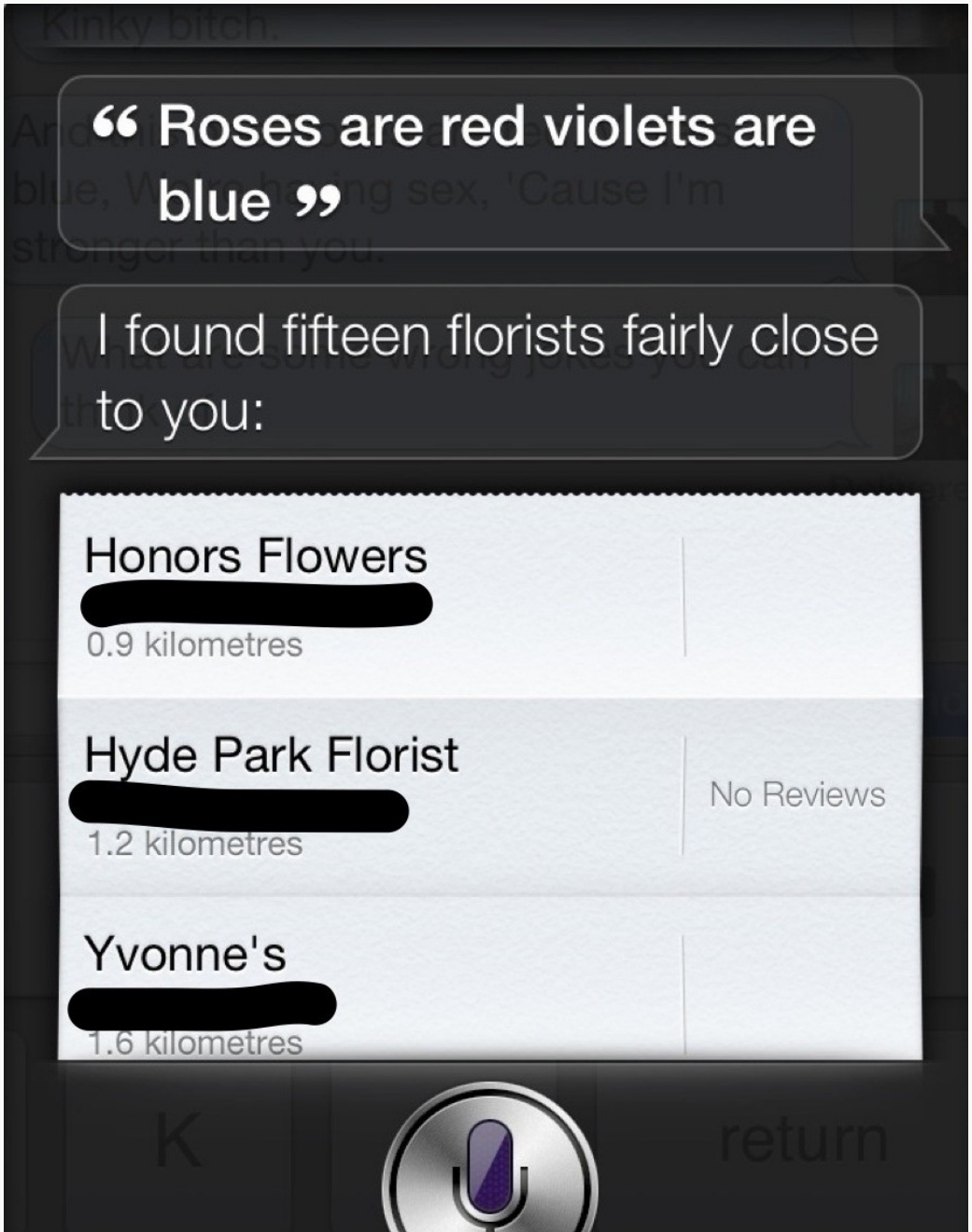
AI (Artificial Intelligence) — strong vs. weak.

**strong AI** — настоящий мыслящий искусственный интеллект, неотличимый от человека.

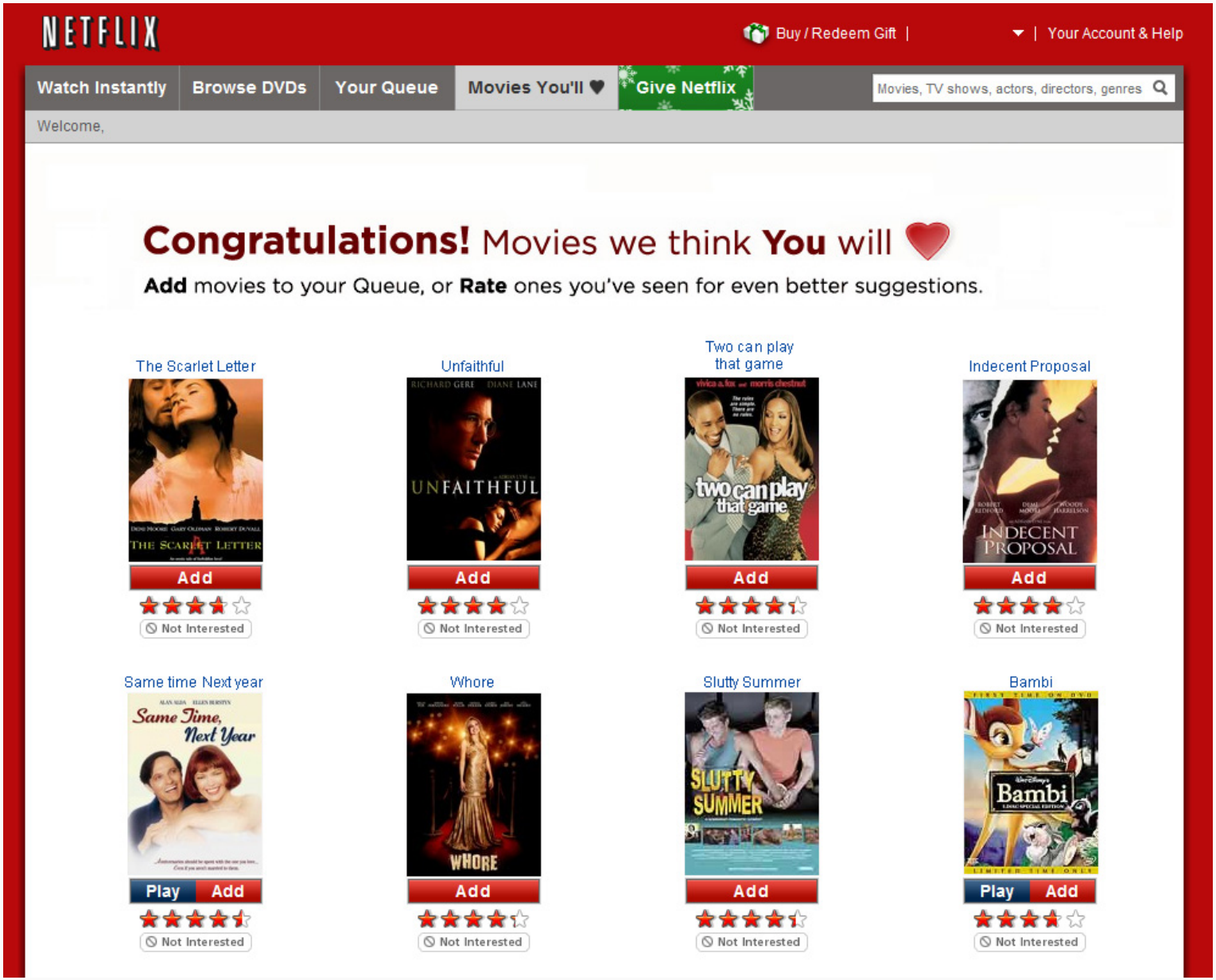
**weak AI** — штука, которая умеет выполнять некоторые когнитивные задачи, которыми обычно занимается человек.

# STRONG AI

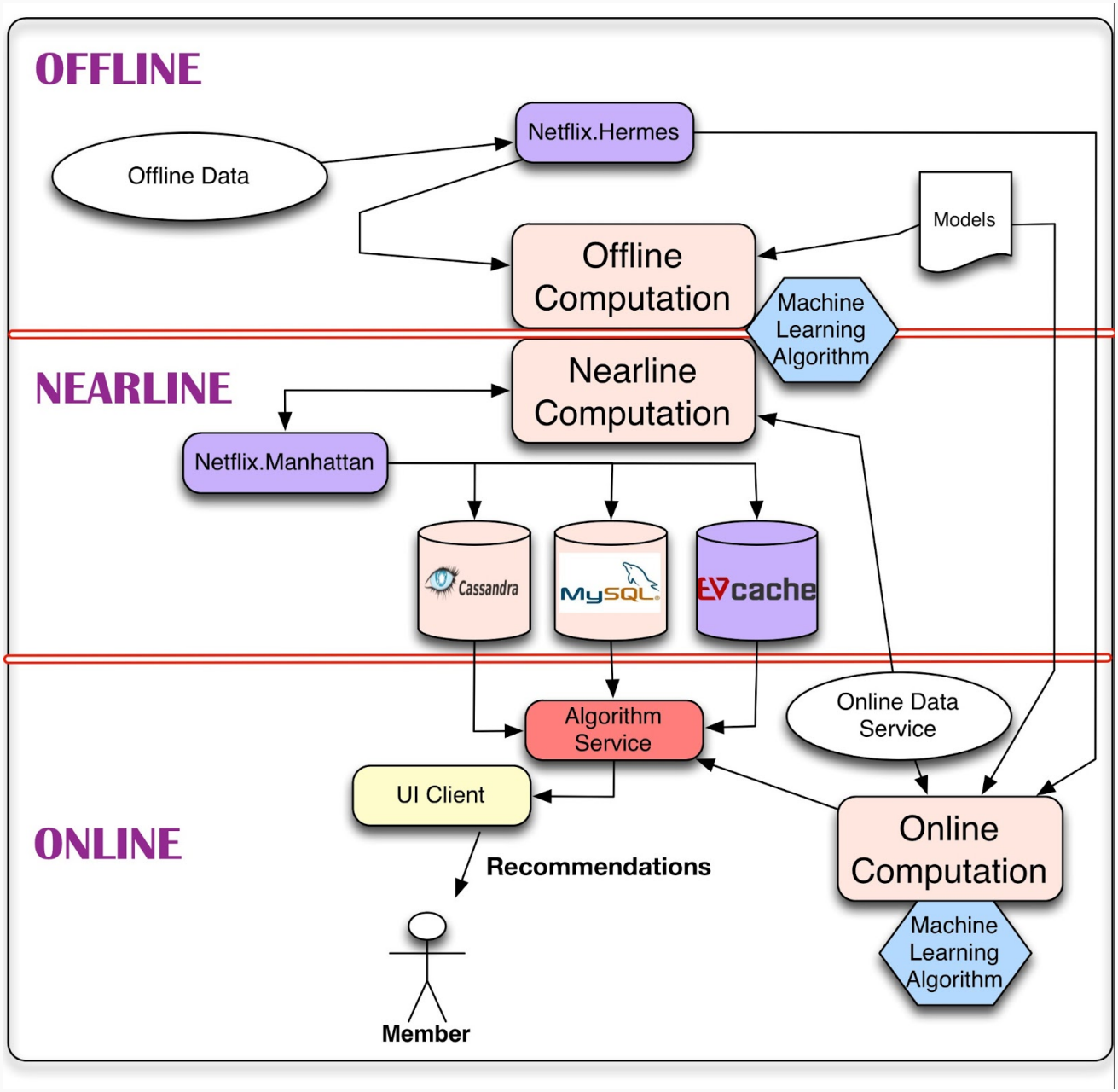
# WEAK AI


“ Roses are red violets are blue ”
I found fifteen florists fairly close to you:
Honors Flowers
[redacted]
0.9 kilometres
Hyde Park Florist
[redacted]
1.2 kilometres
No Reviews
Yvonne's
[redacted]
1.6 kilometres

# WEAK AI


NETFLIX
Buy / Redeem Gift |
| Your Account & Help
Watch Instantly
Browse DVDs
Your Queue
Movies You'll ♥
Give Netflix
Movies, TV shows, actors, directors, genres
Welcome,
Congratulations! Movies we think You will ♥
Add movies to your Queue, or Rate ones you've seen for even better suggestions.
The Scarlet Letter
Add
Not Interested
Unfaithful
Add
Not Interested
Two can play that game
Add
Not Interested
Indecent Proposal
Add
Not Interested
Same time Next year
Play
Add
Not Interested
Whore
Add
Not Interested
Slutty Summer
Add
Not Interested
Bambi
Play
Add
Not Interested

# WEAK AI


OFFLINE
Offline Data
Netflix.Hermes
Models
Offline Computation
Machine Learning Algorithm
NEARLINE
Nearline Computation
Netflix.Manhattan
Cassandra
MySQL
EVcache
Online Data Service
Algorithm Service
UI Client
ONLINE
Online Computation
Recommendations
Machine Learning Algorithm
Member

Strong AI пока не существует, но его хотят, боятся и ищут в существующих программах при помощи теста Тьюринга.

Что не так с тестом Тьюринга?

Weak AI есть вообще практически везде.

# НЕЙРОСЕТИ

Нейросеть — это магический способ решения лингвистических (и не только) проблем. Она смотрит на данные и даёт правильные (обычно) ответы на вопросы про эти данные.

Только надо ~~выбрать правильное заклинание~~ правильно её сконфигурировать — это обычно самое сложное.

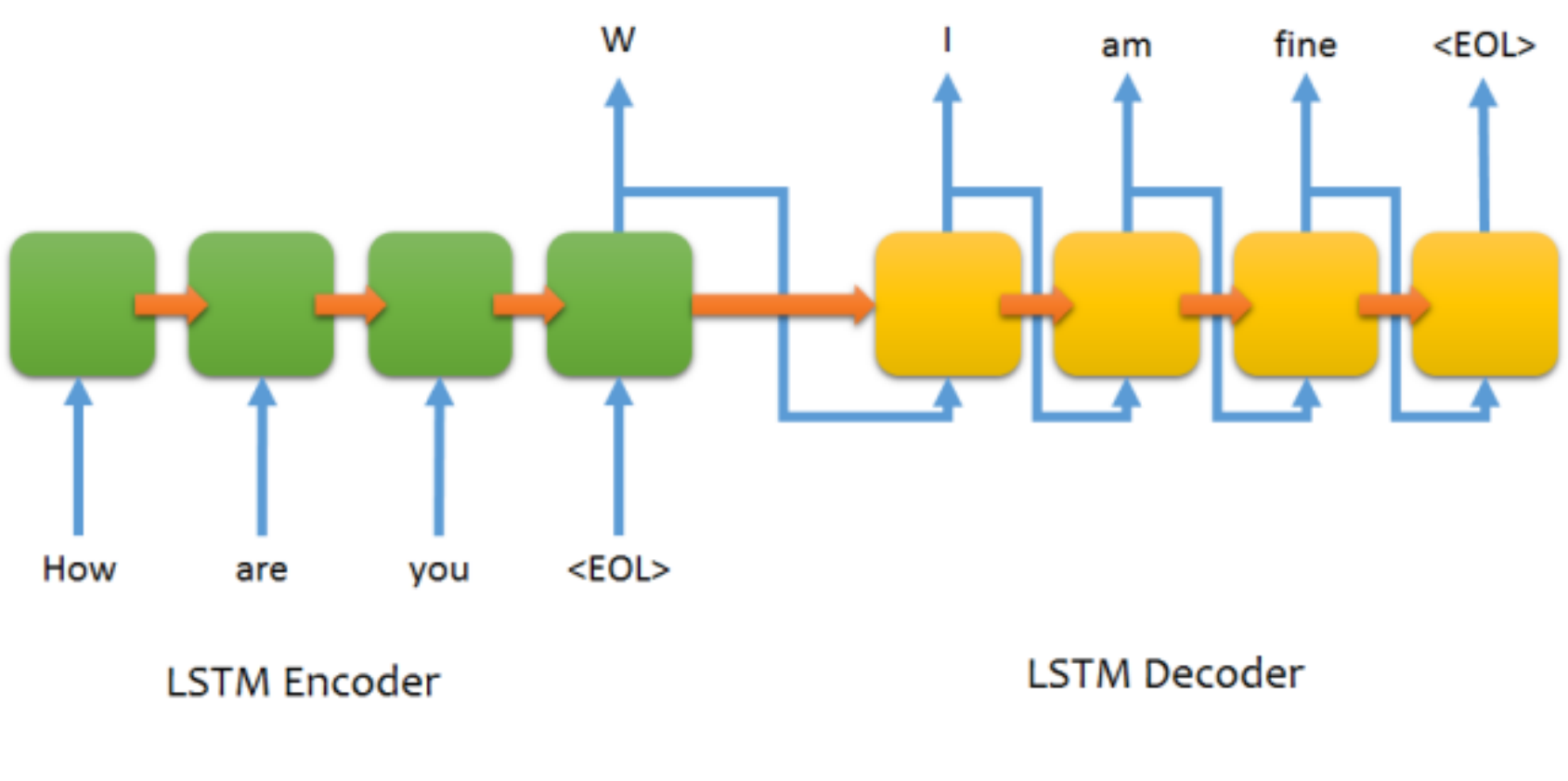
НС используются, в частности, для speech recognition и OCR.

Полным знанием о том, как работают нейросети, не обладает никто.

# SEQ2SEQ-НЕЙРОСЕТИ

Нейросеть вида ”sequence to sequence” принимает на вход некоторую последовательность чего угодно (чисел, пикселей, символов) и порождает другую последовательность чего угодно, сооветствующую первой.

# SEQ2SEQ-НЕЙРОСЕТИ



# Спасибо за внимание!