КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Маша Шеянова, masha.shejanova@gmail.com

November 10, 2018

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.

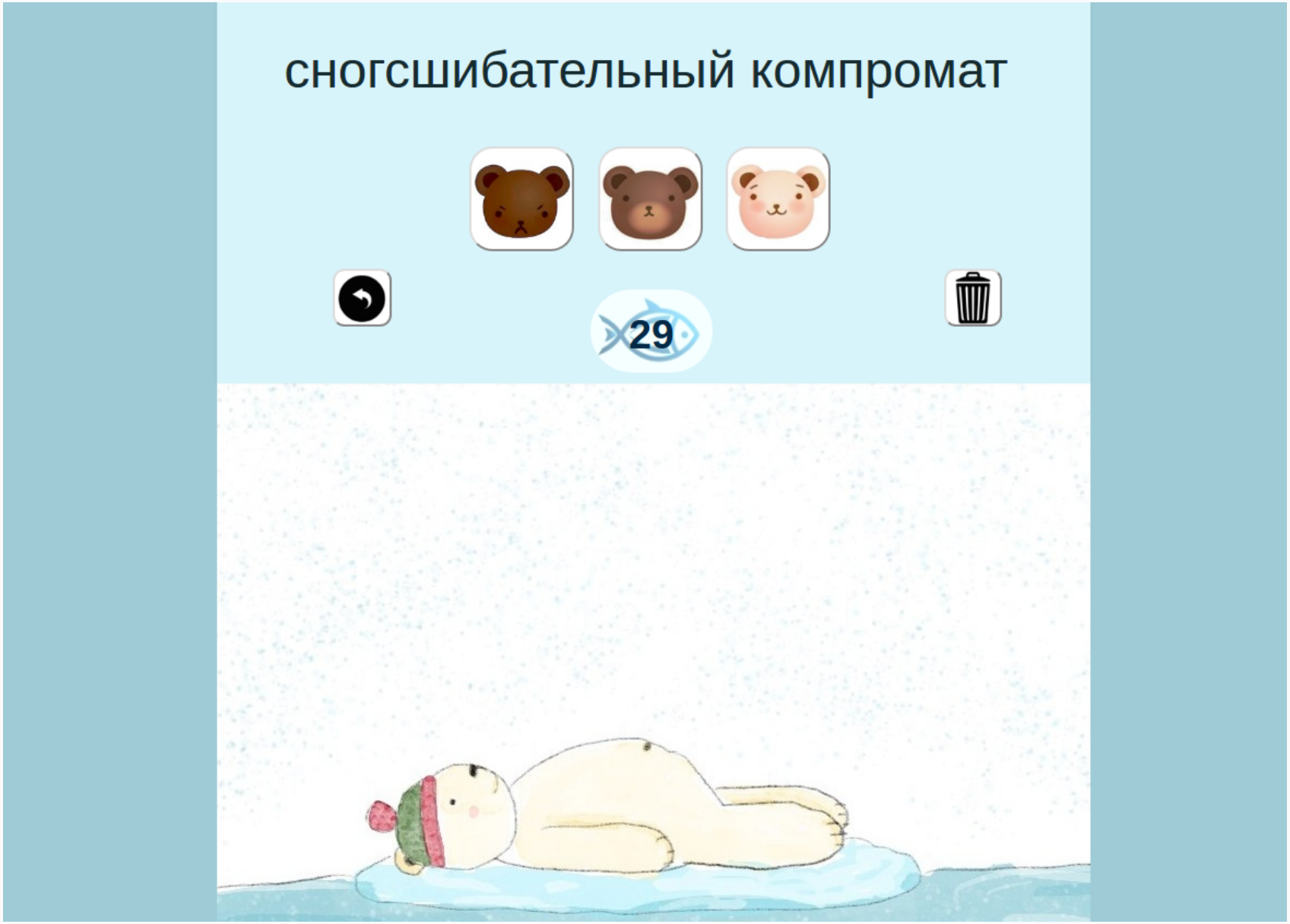
NB: 2 и 3 — очень разные вещи, хотя и то, и другое в русском называют ”компьютерной лингвистикой”

Вспомогательные инструменты

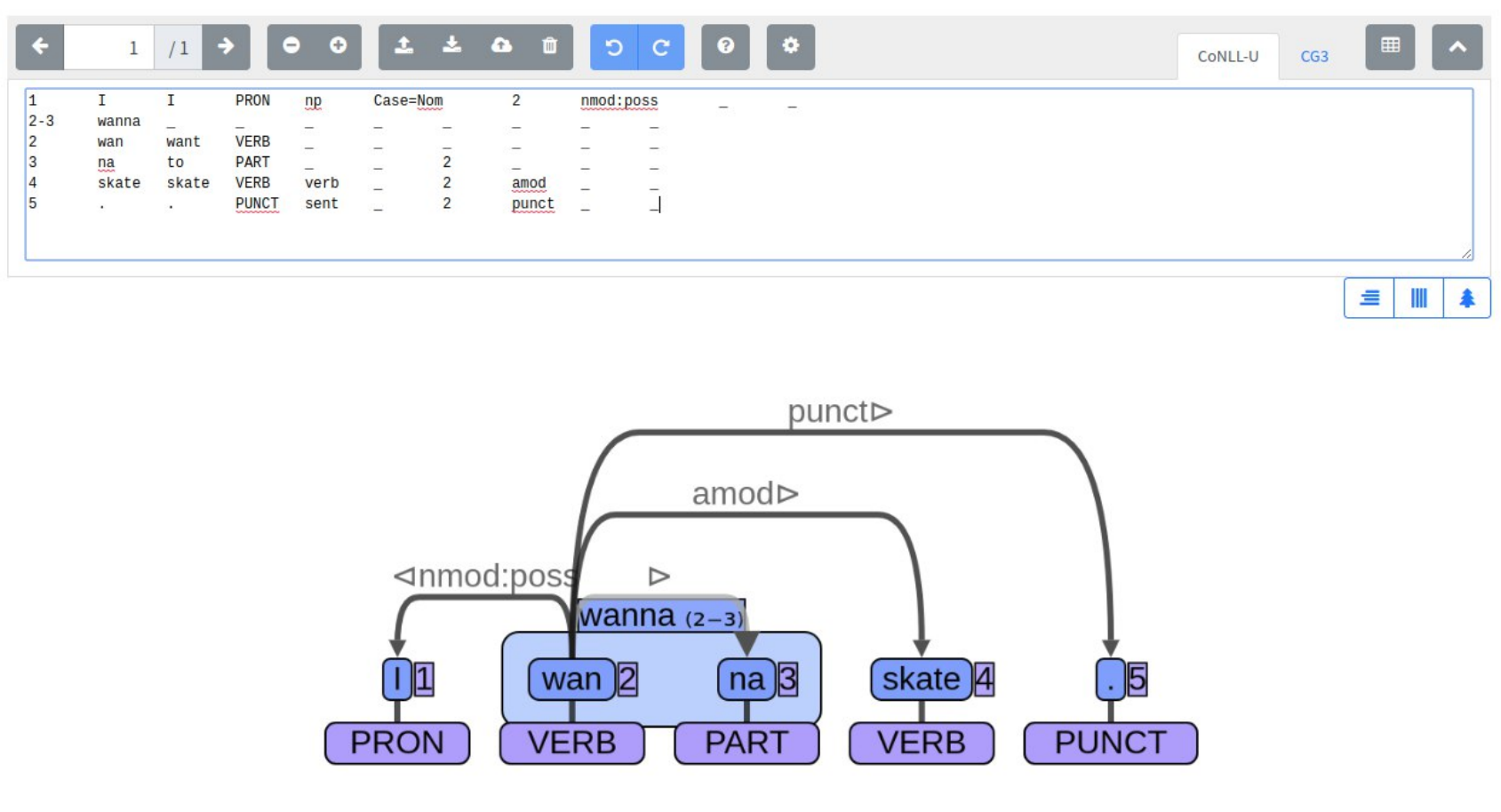
какие они бывают?

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

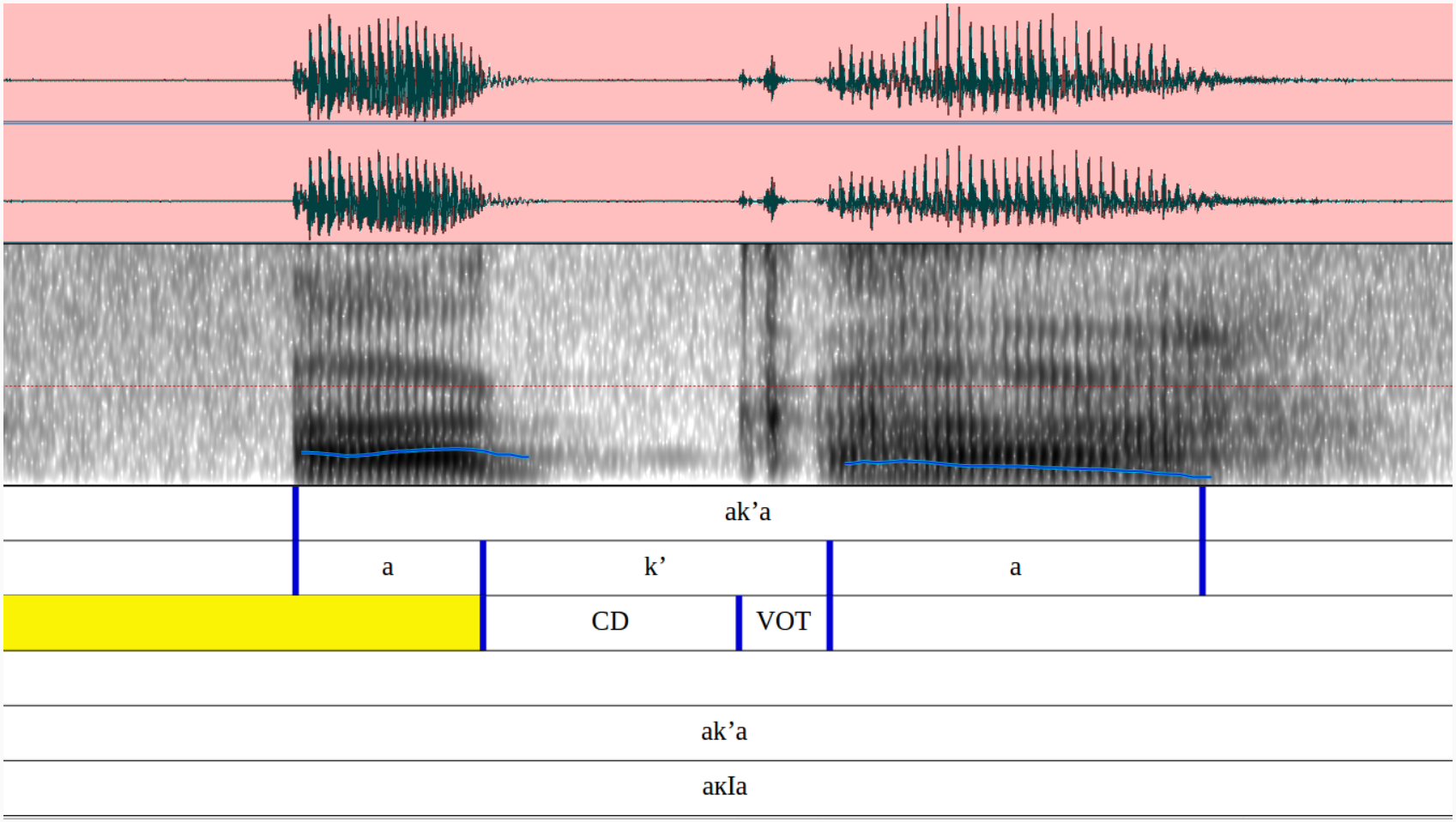
СБОР ДАННЫХ. SENTI GAME.



РАЗМЕТКА ДАННЫХ. UD-ANNOTATRIX.



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, как меняется значение слова (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:

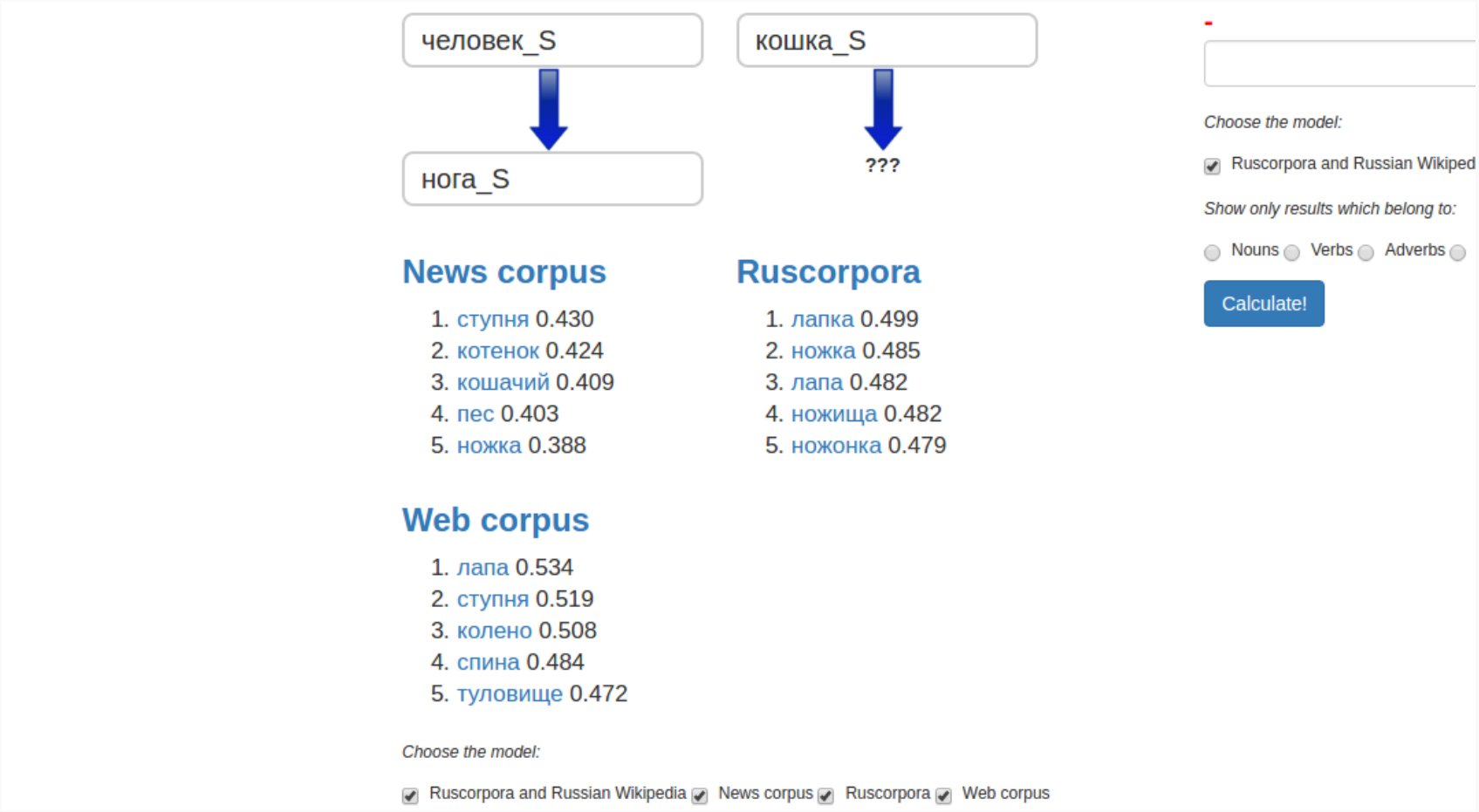
- много текстов, чтобы картинка была репрезентативной
- посчитать в этих текстах взаимную встречаемость слов друг с другом
- найти слова, которые могут заменить друг друга и слова, у которых нет общих контекстов

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

Чем занимается NLP?

- Спеллчекеры
- Машинный перевод
- Text mining (например, извлечение фактов)
- Speech Recognition и Optical Character Recognition
- чатботы (например, Алиса)

МАШИННЫЙ ПЕРЕВОД. ОБЗОР.

Машинный перевод (machine translation, MT). Служит, в первую очередь, для перевода технической документации и как вспомогательный инструмент человека-переводчика.

- Основанные на корпусах:
  - **Статистический** (SBMT)
  - **Нейронный** (NMT)
  - **Example-based** (EBMT)
- **Правиловый** (RBMT). Использует лингвистические знания человека для создания адекватной языковой модели.
- **Гибридный** (HMT). Не один подход, а разнородный кластер.

Машинный перевод. Corpus-based.

У нас есть параллельные корпуса:

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.

С их помощью мы учим компьютер переводить предложения пользователя.

- **Статистический**: Учитывает вероятность того, что строка A является переводом строки B и вероятность появления строки A в целевом языке.
- **Neural**: Нечто похожее, но на нейросетях.
- **Example-based**: Не использует статистику. Переводы строятся на основе пропорциональных аналогий.

ПРАВИЛОВЫЙ МАШИННЫЙ ПЕРЕВОД

Правиловый перевод подразделяется на:

- **Dictionary-based** (direct) method – наивный подход. Использует прямые словарные соответствия между исходным и целевым языками. Не учитывает грамматическую структуру текста. Самый ранний.
- **Interlingua** method использует абстрактное представление, не привязанное к конкретному языку. Хорош для многоязыковых (multilingual) систем.
- **Transfer** method: текст сначала преобразуется в проекцию, близкую к исходному языку, затем из неё – в проекцию, ориентированную на целевой язык. Включает в себя:
  - **deep transfer**: каждое предложение имеет дерево разбора;
  - **shallow transfer**: оперирует частями предложения (chunks).

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

ГИБРИДНЫЙ ПОДХОД

Делится на две большие группы:

- Multi-engine: применяется одновременно несколько подходов, результат сравнивается, выбирается лучший кандидат.
- Single-engine:
  - Статистический перевод, модифицированный правилами (e.g. использовать знания о морфологии, о синтаксисе; factor-based).
  - Правиловый перевод, использующий статистические методы (e.g. предобработка, взвешивание правил).

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.

ПОРОЖДЕНИЕ ТЕКСТА

http://kingjamesprogramming.tumblr.com/ — марковскую цепь натренировали на библии и учебнике программирования:

> “The LORD commanded my lord to give the user the ability to manipulate inexact quantities”

> “The theme of computers being viewed not merely as logic devices but as the servants of Pharaoh”

> “the LORD is greater than or equal to the expression involving ?y”

AI

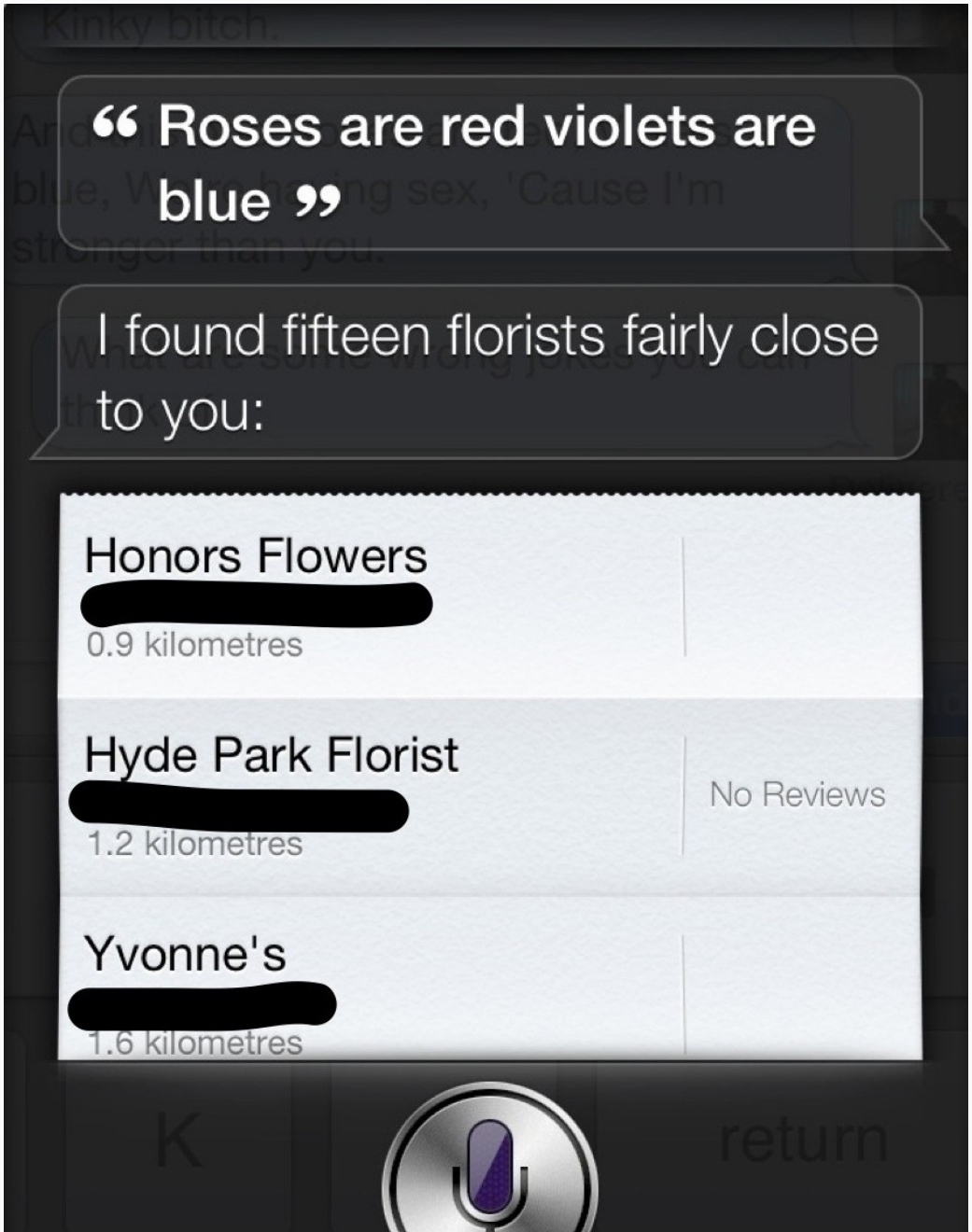
AI (Artificial Intelligence) — strong vs. weak.

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

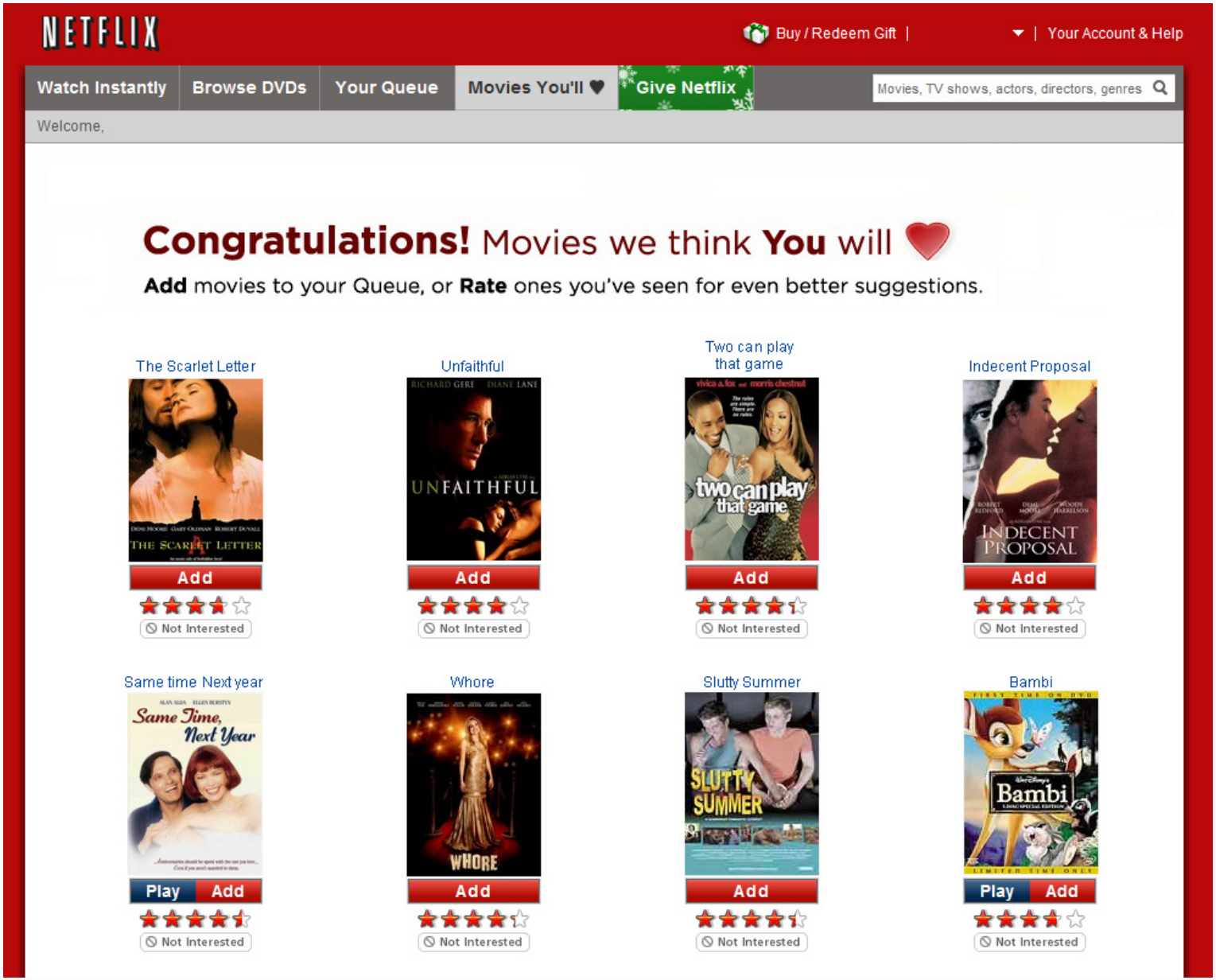
weak AI — штука, которая умеет выполнять некоторые когнитивные задачи, которыми обычно занимается человек.

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

AI

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, OCR и порождения текста.

Полным знанием о том, что происходит внутри нейросетки, не обладает никто.

Спасибо за внимание!