ИТМО

Тренды искусственного интеллекта

GOFAI и вероятностные основы обучения





@SINECOR

Сергей Николенко

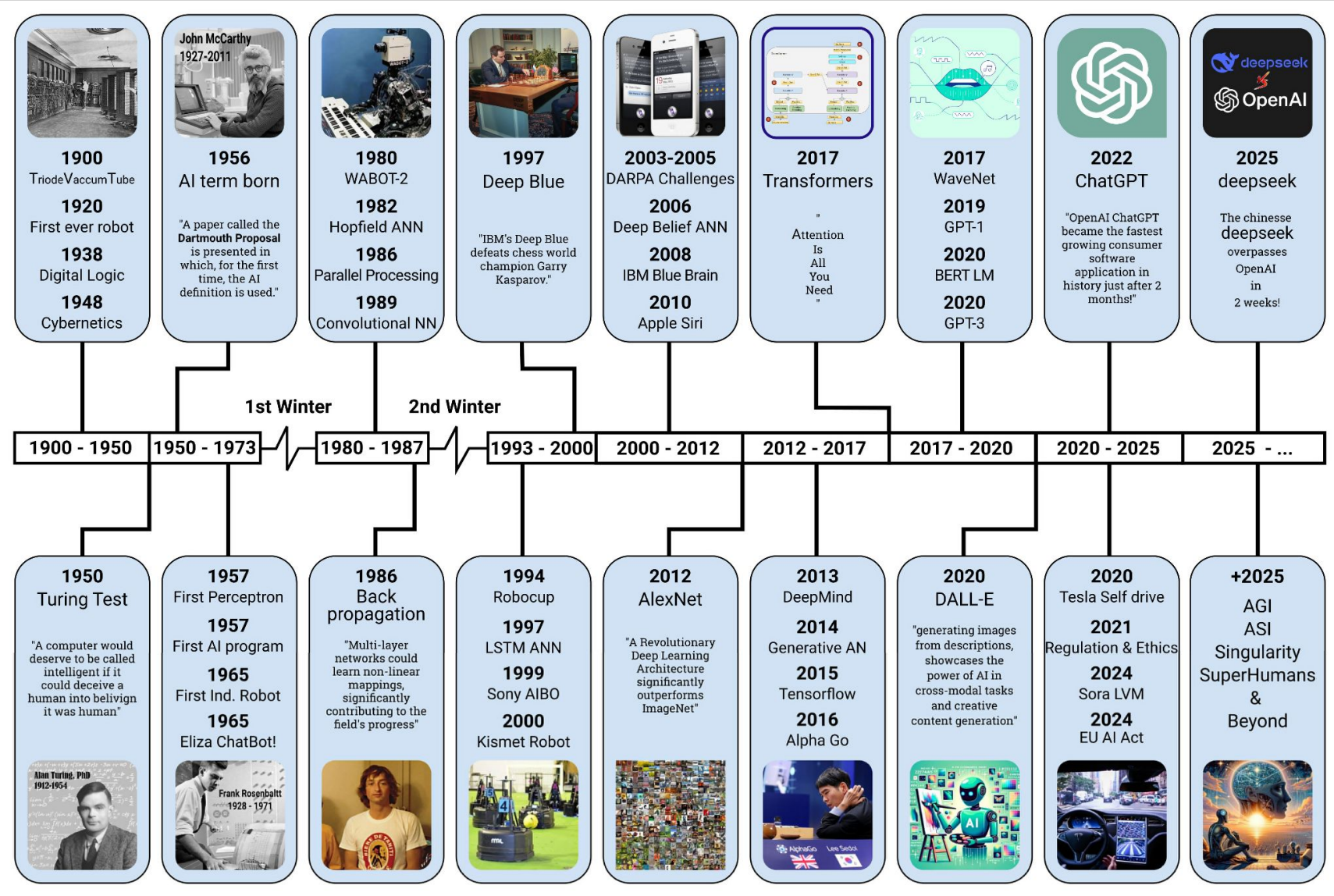
28 октября 2026 г.

Примеры ранних систем искусственного интеллекта



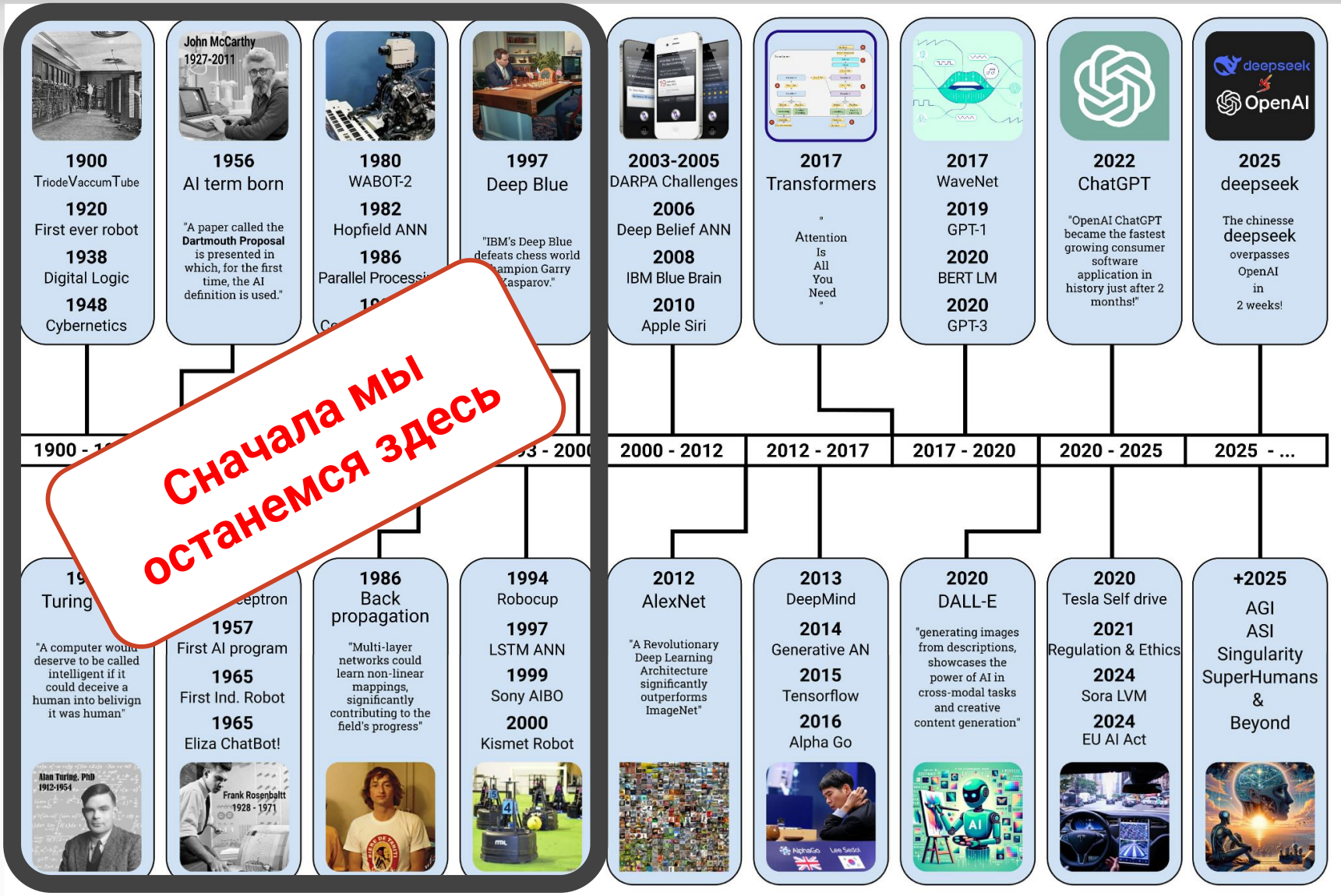
Краткая история AI как науки

- AI timeline (источник)



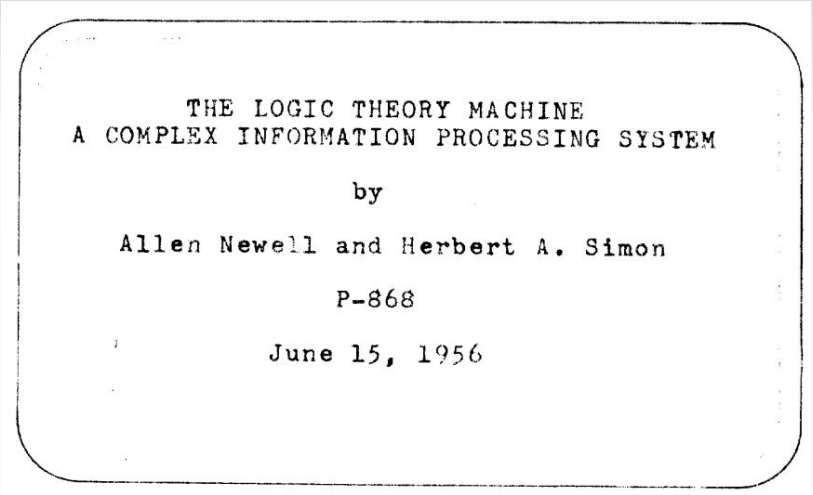
Краткая история AI как науки

- AI timeline (источник)



Ньюэлл и Саймон: Logic Theorist

- Newell, Simon (1956): Logic Theorist, программа для автоматического доказательства теорем
- Herbert Simon, Jan 1956: “Over Christmas, Al Newell and I invented a thinking machine”
- Он же, позже: “[we] invented a computer program capable of thinking non-numerically, and thereby solved the venerable mind-body problem, explaining how a system composed of matter can have the properties of mind”.



THE LOGIC THEORY MACHINE
A COMPLEX INFORMATION PROCESSING SYSTEM

by

Allen Newell and Herbert A. Simon

P-868

June 15, 1956



Herb Simon and Allen Newell.

Photograph courtesy, Carnegie Mellon University.

Ньюэлл и Саймон: Logic Theorist

- Newell, Simon (1956): Logic Theorist, программа для автоматического доказательства теорем
- На самом деле это был отличный прувер, передоказал много теорем из Principia Mathematica… но всё-таки как-то это не взлетело

The two connectives, - and v, are taken as primitives. The third connective, → , is defined in terms of the other two, thus:

1.01 $p \to q \ =_{def} \ -p \vee q$

The five axioms that are postulated to be true are:

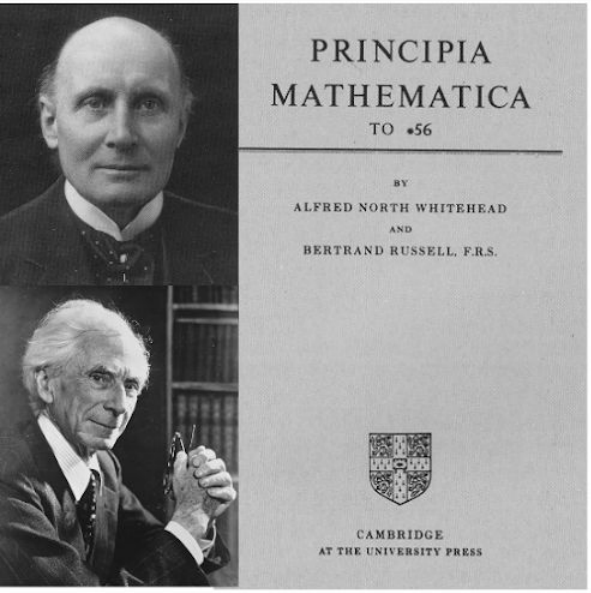
1.2 $p \vee p .\to. p$

1.3 $p .\to. q \vee p$

1.4 $p \vee q .\to. q \vee p$

1.5 $p .\vee. q \vee r :\to: q .\vee. p \vee r$

1.6 $p \to q .\to: r \vee p .\to. r \vee q$



***54·43.** $\vdash :. \alpha, \beta \in 1 . \supset : \alpha \cap \beta = \Lambda . \equiv . \alpha \cup \beta \in 2$

Dem.

$\vdash . *54{\cdot}26 . \supset \vdash :. \alpha = \iota' x . \beta = \iota' y . \supset : \alpha \cup \beta \in 2 . \equiv . x \neq y .$

[*51·231] $\equiv . \iota' x \cap \iota' y = \Lambda .$

[*13·12] $\equiv . \alpha \cap \beta = \Lambda$ (1)

$\vdash . (1) . *11{\cdot}11{\cdot}35 . \supset$

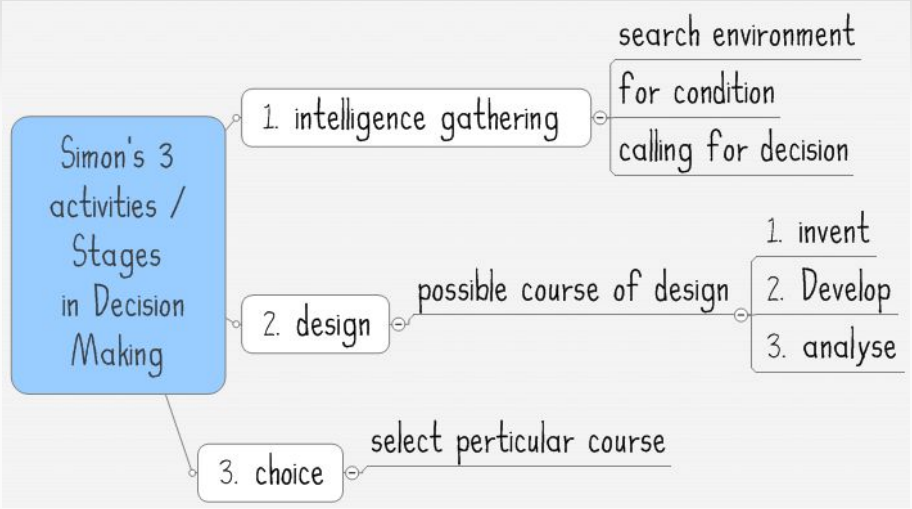
$\vdash :. (\exists x, y) . \alpha = \iota' x . \beta = \iota' y . \supset : \alpha \cup \beta \in 2 . \equiv . \alpha \cap \beta = \Lambda$ (2)

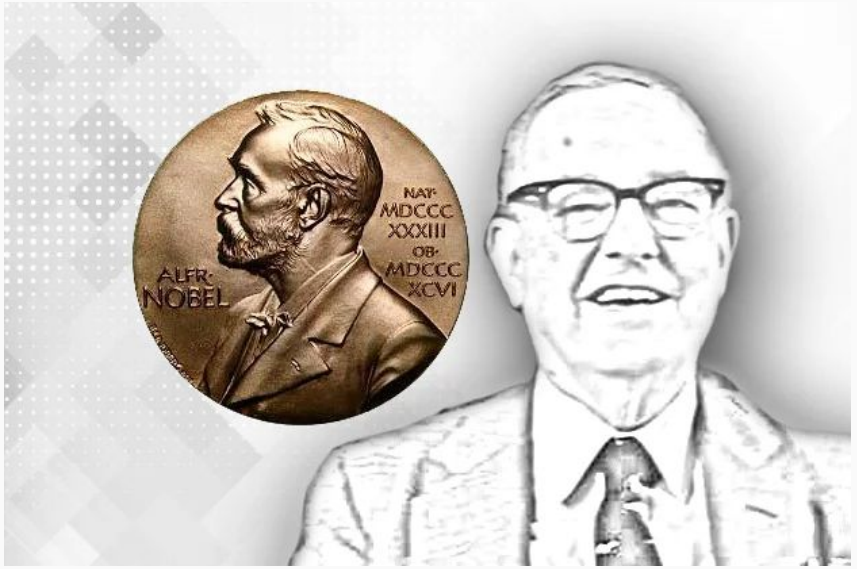
$\vdash . (2) . *11{\cdot}54 . *52{\cdot}1 . \supset \vdash . \text{Prop}$

From this proposition it will follow, when arithmetical addition has been defined, that $1 + 1 = 2$.

Герберт Саймон

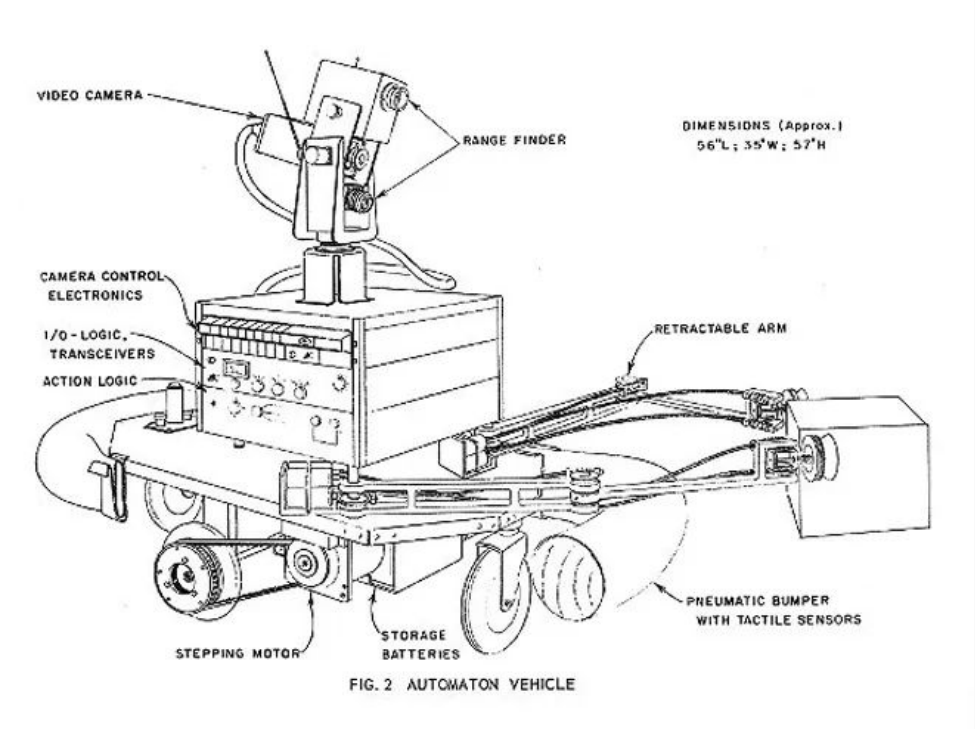
- Кстати, Herbert Alexander Simon был не только пионером AI
- Он в целом изучал принятие решений; *Administrative Behaviour* (1947) — о том, как люди принимают рациональные решения
- Стандартная тогда модель — homo economicus, человек рациональный и эгоистичный
- Саймон изучал рациональные решения с ограничениями (**bounded rationality**): человек не всё знает и не вполне эгоистичен
- **Satisficing**: как принять удовлетворительное, но не обязательно оптимальное решение





Shakey the Robot

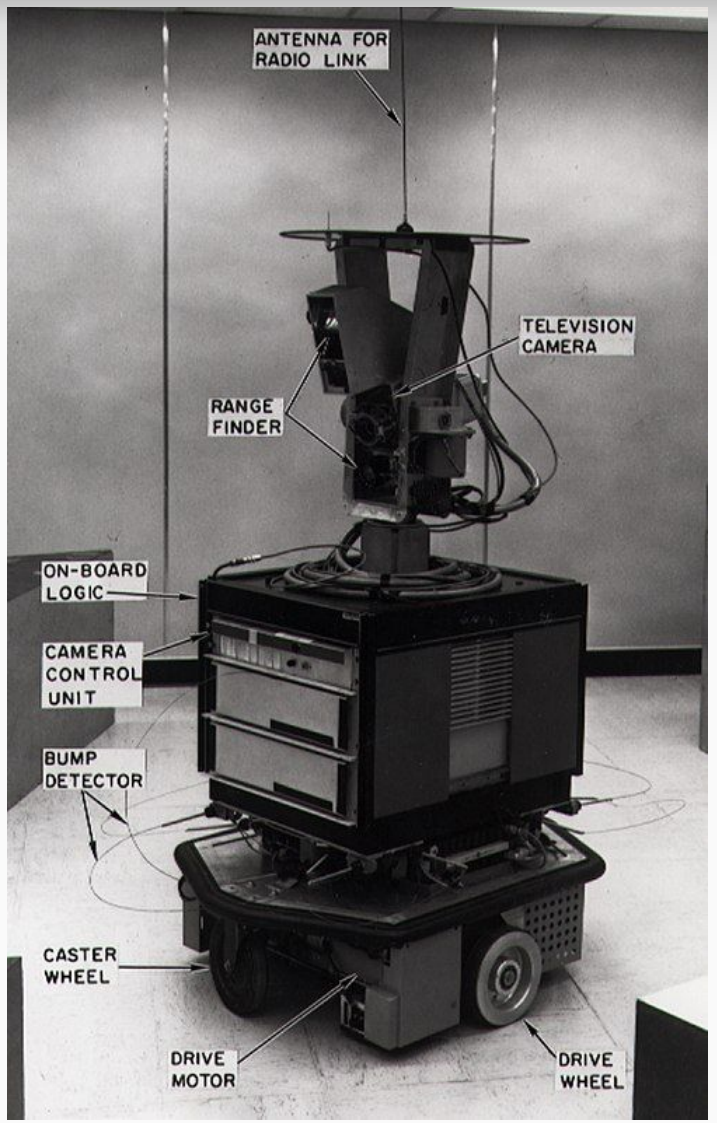
- Shakey the Robot (Stanford, 1966): первый мобильный робот, который мог планировать свои действия
- Мог получить команду вроде “push the block off the platform”, сам найти платформу с блоком, спланировать, как до неё добраться, а потом физически исполнить этот план





Shakey the Robot

- Планировщик STRIPS (Stanford Research Institute Problem Solver), написанный на Lisp; мог спланировать перемещения, включение/выключение света, открывание/закрывание дверей и т.д.
- Кроме того, Shakey сам смотрел на окружающий мир, это была одна из первых систем компьютерного зрения
- К нему относится и знаменитая байка: якобы в 1966 году Марвин Минский выдал Джеральду Сассмену (Gerald Jay Sussman), тогда аспиранту, задачу решить проблему компьютерного зрения в качестве проекта на лето



Shakey the Robot

- Планировщик STRIPS (Stanford Research Institute Problem Solver), написанный на Lisp; мог спланировать перемещение, включение/выключение, открывание/закрывание
- Кроме того, ...
- ...

На самом деле в 1966 году Сеймур Пейперт (Seymour Papert) выдал целой группе студентов и аспирантов под руководством Сассмена летний проект, в котором предложил решить всего лишь задачу сегментации изображений. :)



ELIZA

- ELIZA (Weizenbaum, 1964-1968): один из первых настоящих чатботов, система, которая симулировала разговор через простой pattern matching
- Самый известный скрипт — DOCTOR, симулятор психотерапевта
- Вейценбаум выбрал терапию, чтобы не нужно было строить модель мира, а можно было просто рефлексировать то, что говорит собеседник (PCT, person-centered therapy, Rogerian therapy от Карла Роджерса)

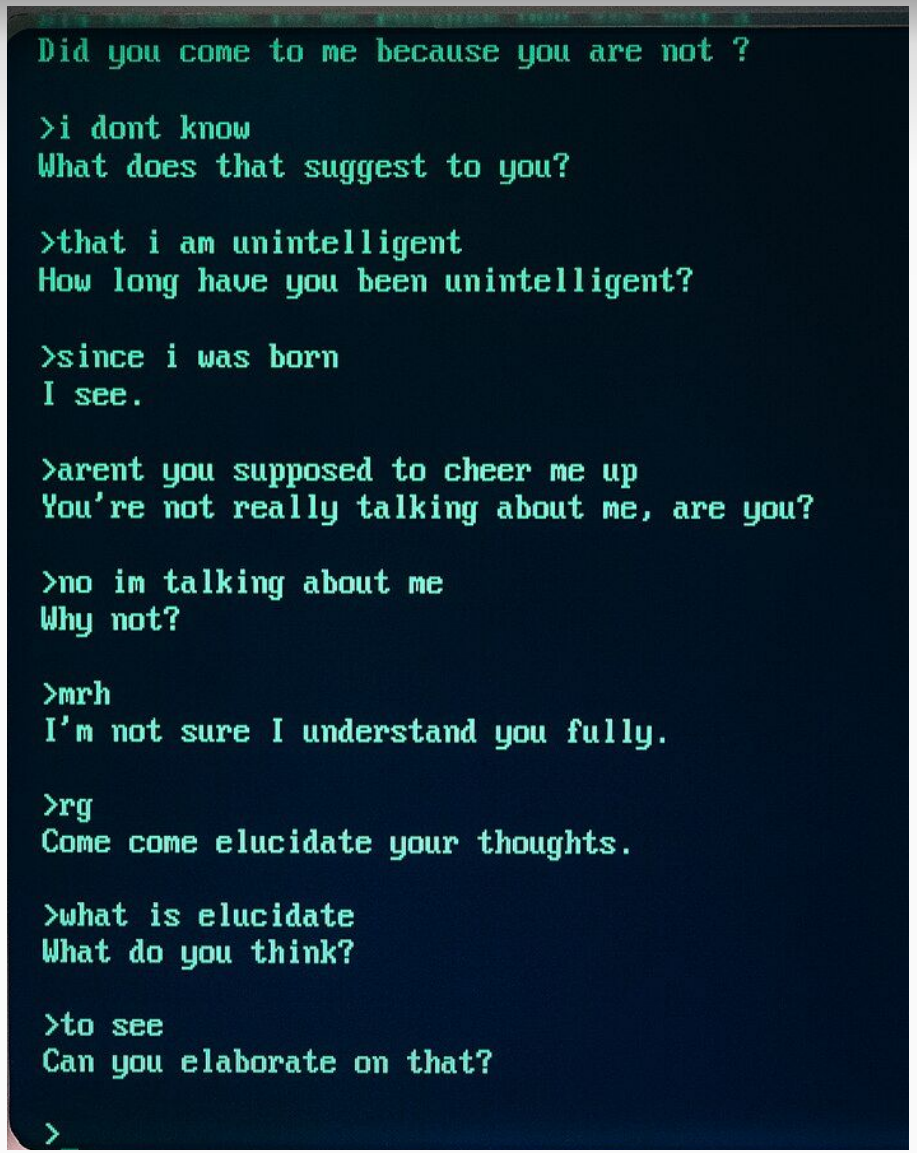
```
  Welcome to
                EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                EE      LL       II       ZZ   AA   AA
                EEEEE   LL       II     ZZZ    AAAAAAA
                EE      LL       II    ZZ      AA   AA
                EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

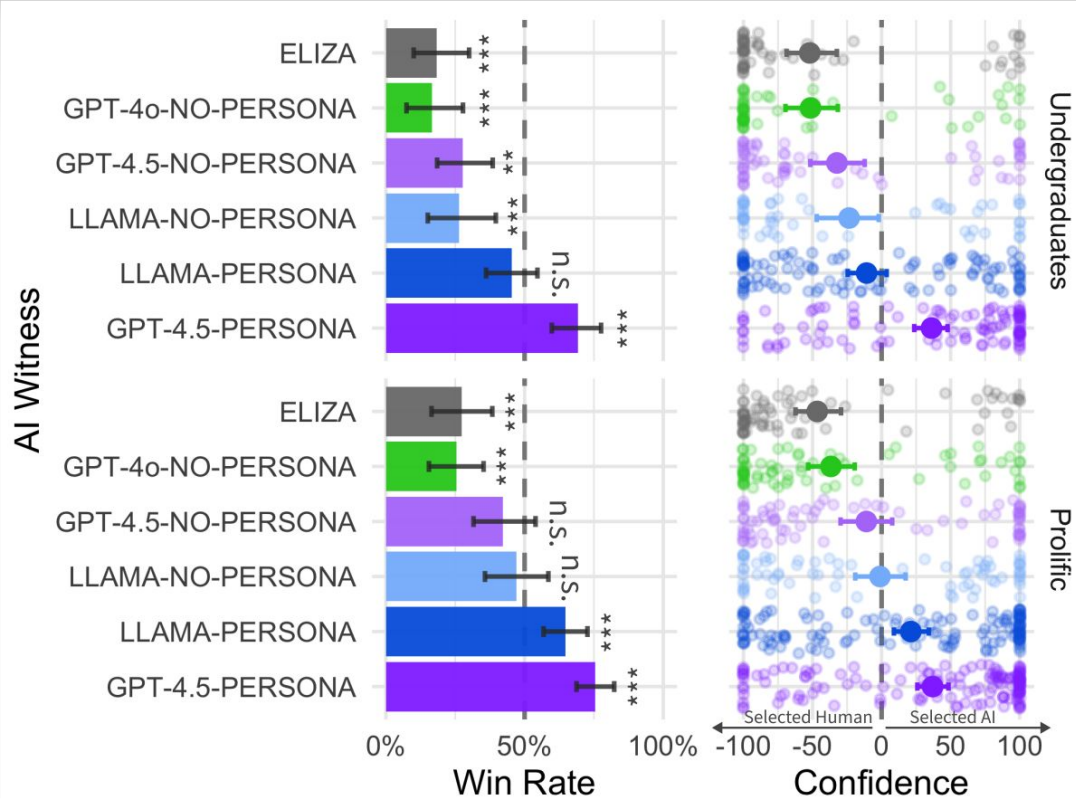
ELIZA

- ELIZA (Weizenbaum, 1964-1968): выглядит очень просто, и легко сломать
- Попробуйте сами: https://anthay.github.io/eliza.html
- И тем не менее, люди были восхищены, иногда эмоционально привязывались к ELIZA и разговорам с ней



ELIZA

- ELIZA (Weizenbaum, 1964-1968)
- Мы говорили о тесте Тьюринга
- ELIZA — первая программа, которая хоть иногда проходила тест Тьюринга
- И она осталась baseline'ом даже в Jones, Bergen (Mar 31, 2025)!



Large Language Models Pass the Turing Test

Cameron R. Jones
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
cameron@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
bkbergen@ucsd.edu

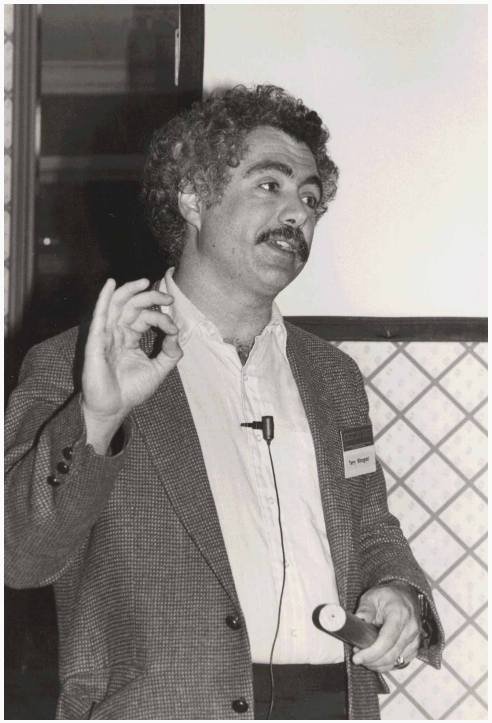
SHRDLU

- SHRDLU (Winograd, 1968-1970): система понимания естественного языка в ограниченном контексте (кубики на столе)
- Могла на естественном языке получать команды, планировать их исполнение, поддерживать контекст и анафору
- Это была знаменитая демонстрация искусственного интеллекта…

```
Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I DON'T UNDERSTAND WHICH PYRAMID
YOU MEAN.
Person (changing their mind): Find a block
which is taller than the one you are
holding and put it into the box.
Computer: BY "IT", I ASSUME YOU MEAN THE
BLOCK WHICH IS TALLER THAN THE ONE I AM
HOLDING.
Computer: OK.
Person: What does the box contain?
Computer: THE BLUE PYRAMID AND THE BLUE
BLOCK.
Person: What is the pyramid supported by?
Computer: THE BOX.
Person: How many blocks are not in the box?
Computer: FOUR OF THEM.
Person: Is at least one of them narrower
than the one which i told you to pick up?
Computer: YES, THE RED CUBE.
…
```

SHRDLU

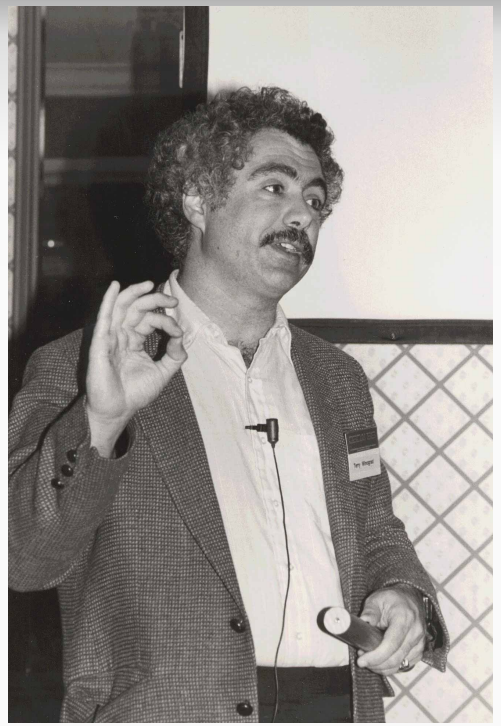
- ...но это была иллюзия :)
- Диалог поддерживался только в определённом контексте, и не было никаких гарантий качества; система скорее не работала, чем работала
- Сам Терри Виноград вспоминал (справа)



[...] the famous dialogue with SHRDLU where you could pick up a block, and so on, I very carefully worked through, line by line. If you sat down in front of it, and asked it a question that wasn't in the dialogue, there was some probability it would answer it. I mean, if it was reasonably close to one of the questions that was there in form and in content, it would probably get it. But there was no attempt to get it to the point where you could actually hand it to somebody and they could use it to move blocks around. And there was no pressure for that whatsoever. Pressure was for something you could demo. Take a recent example, Negroponte's Media Lab, where instead of "perish or publish" it's "demo or die." I think that's a problem. I think AI suffered from that a lot, because it led to "Potemkin villages", things which - for the things they actually did in the demo looked good, but when you looked behind that there wasn't enough structure to make it really work more generally.

Winograd schema

- Кстати, в честь Винограда назван очень интересный датасет на понимание мира, с которым модели полноценно начали справляться только после GPT-3
- Оказывается, разрешение анафоры иногда требует проникновения в суть вещей
- Но ещё больше проникновения требует перевод...



	Twin sentences	Options (**answer**)
✗	The monkey loved to play with the balls but ignored the blocks because he found **them** *exciting*.	**balls** / blocks
	The monkey loved to play with the balls but ignored the blocks because he found **them** *dull*.	balls / **blocks**
✗	William could only climb begginner walls while Jason climbed advanced ones because **he** was very *weak*.	**William** / Jason
	William could only climb begginner walls while Jason climbed advanced ones because **he** was very *strong*.	William / **Jason**
✓	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *less* time to get ready for school.	**Robert** / Samuel
	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *more* time to get ready for school.	Robert / **Samuel**
✓	The child was screaming after the baby bottle and toy fell. Since the child was *hungry*, **it** stopped his crying.	**baby bottle** / toy
	The child was screaming after the baby bottle and toy fell. Since the child was *full*, **it** stopped his crying.	baby bottle / **toy**

Машинный перевод в 1950-х годах

- Warren Weaver, директор Отделения естественных наук в Фонде Рокфеллера, в 1947 написал Винеру о возможности машинного перевода, а в 1949 выпустил меморандум под названием *Translation*
- Из письма Винеру: “One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

1 • Translation*

WARREN WEAVER

There is no need to do more than mention the obvious fact that a multiplicity of languages impedes cultural interchange between the peoples of the earth, and is a serious deterrent to international understanding. The present memorandum, assuming the validity and importance of this fact, contains some comments and suggestions bearing on the possibility of contributing at least something to the solution of the world-wide translation problem through the use of electronic computers of great capacity, flexibility, and speed.

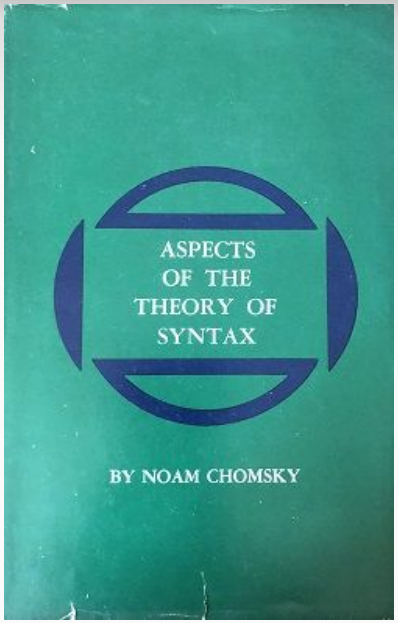
The suggestions of this memorandum will surely be incomplete and naïve, and may well be patently silly to an expert in the field—for the author is certainly not such.

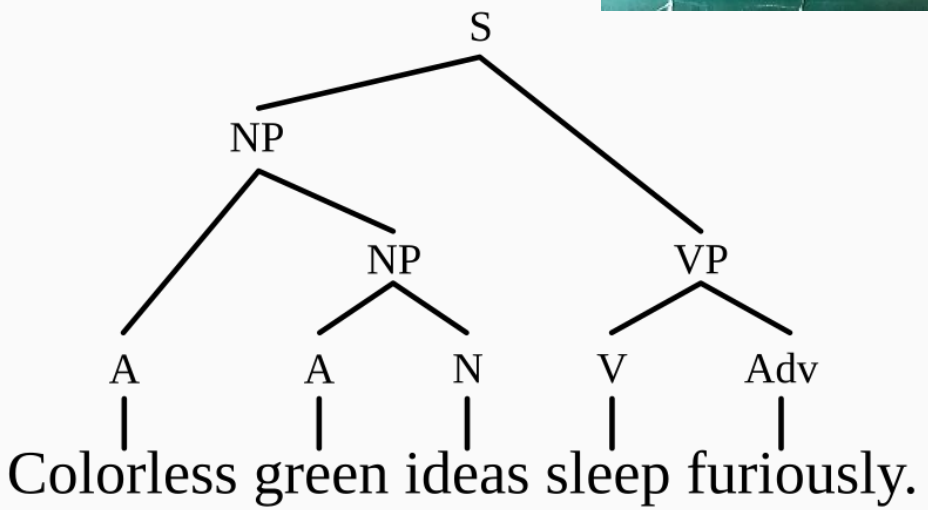


Машинный перевод в 1950-х годах

- Эта мысль вскоре получила подтверждение в работах Ноама Хомского о порождающих грамматиках
- В 1950-х он как раз занимался машинным переводом, и его диссертация (1955) и книга *Aspects of the Theory of Syntax* (1956) уже содержали все основы теории и синтаксические деревья





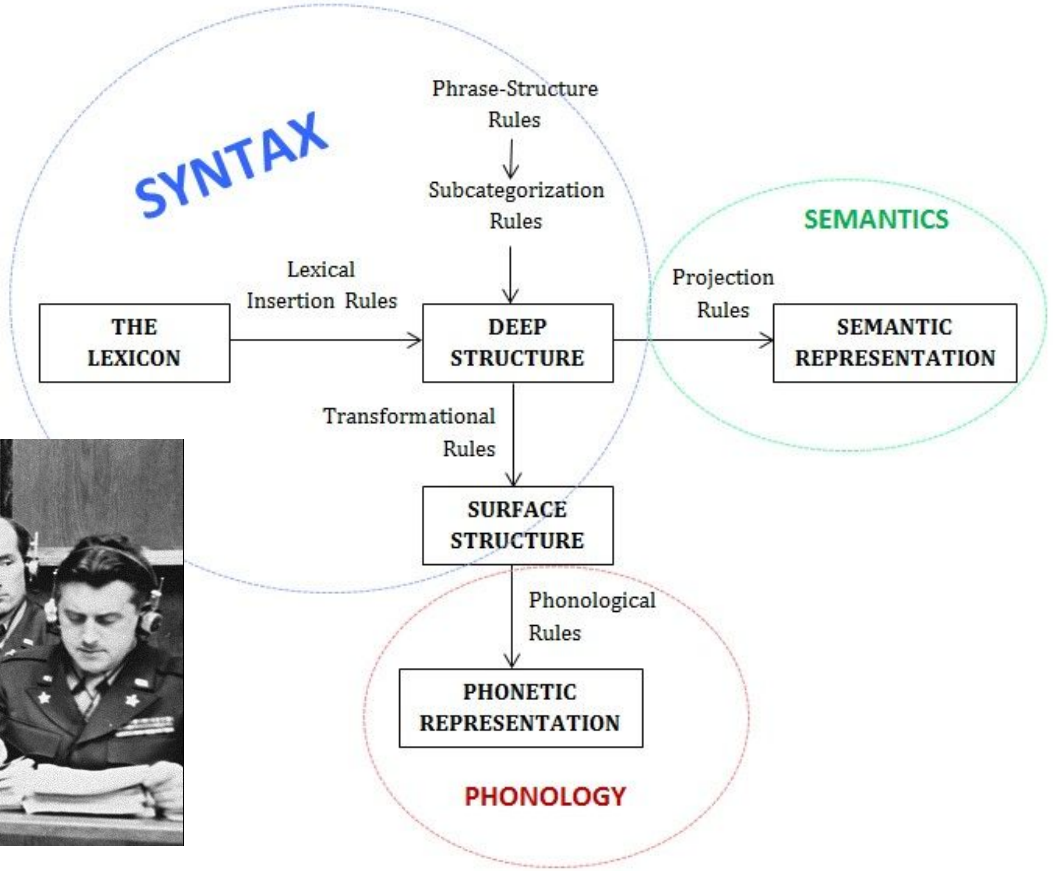


Машинный перевод в 1950-х годах

- Если у языка есть хорошо определённая грамматика, и её можно записать в виде правил преобразования, то можно записать и правила преобразования с одного языка на другой
- А там и запрограммировать их; Леон Достерт уже в 1954 провёл первую публичную демонстрацию машинного перевода



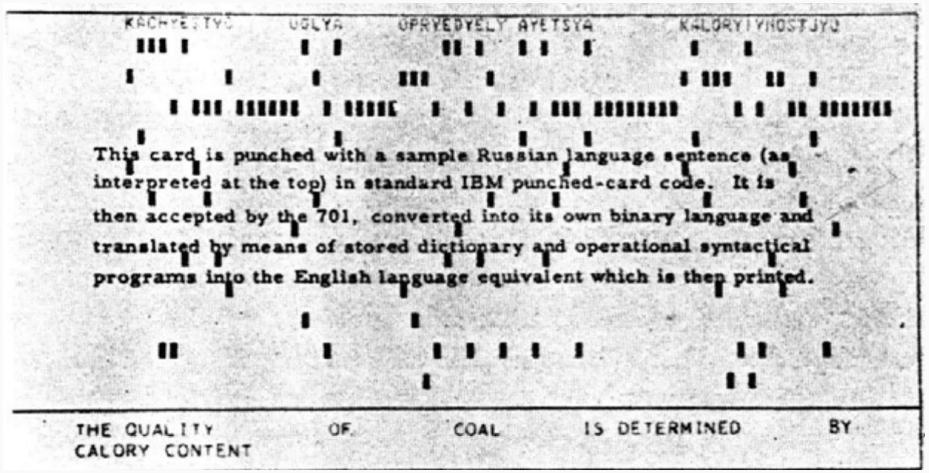


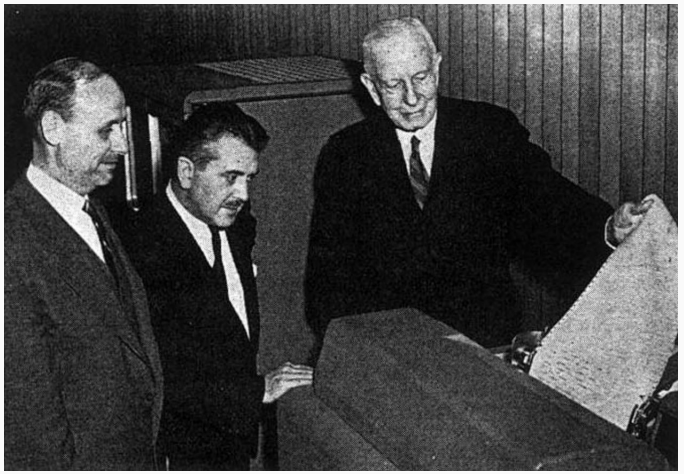


Машинный перевод в 1950-х годах

- Эта демонстрация известна как Georgetown-IBM Experiment (1954)

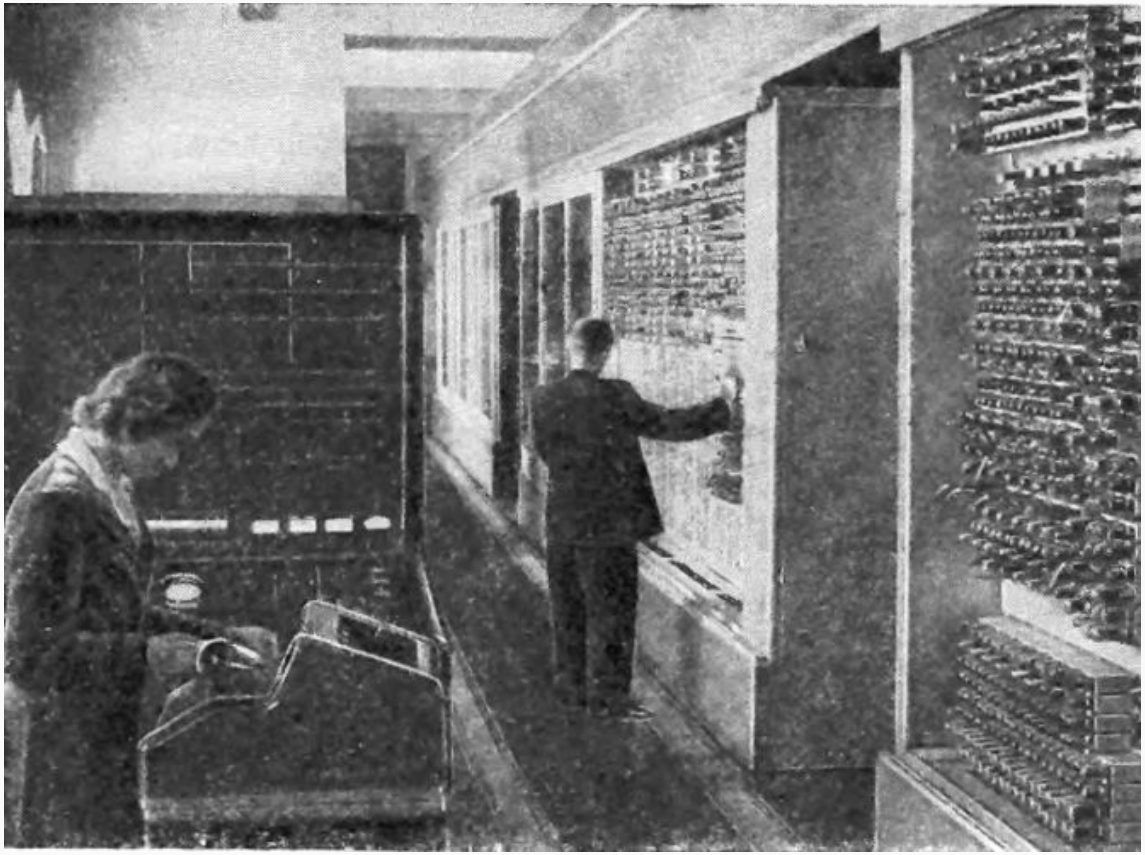
Вход (русский транслит)	Выход (английский)
Mi pyeryedayem mislyi posryedstvom ryechyi.	We transmit thoughts by means of speech.
Vyelyichyina ugla opryedyelyayetsya otnoshyenyiyem dlyini dugi k radyiusu.	Magnitude of angle is determined by the relation of length of arc to radius.
Myezhdunarodnoye ponyimanyiye yavlyayetsya vazhnim faktorom v ryeshyenyiyi polyityichyeskix voprosov.	International understanding constitutes an important factor in decision of political questions.

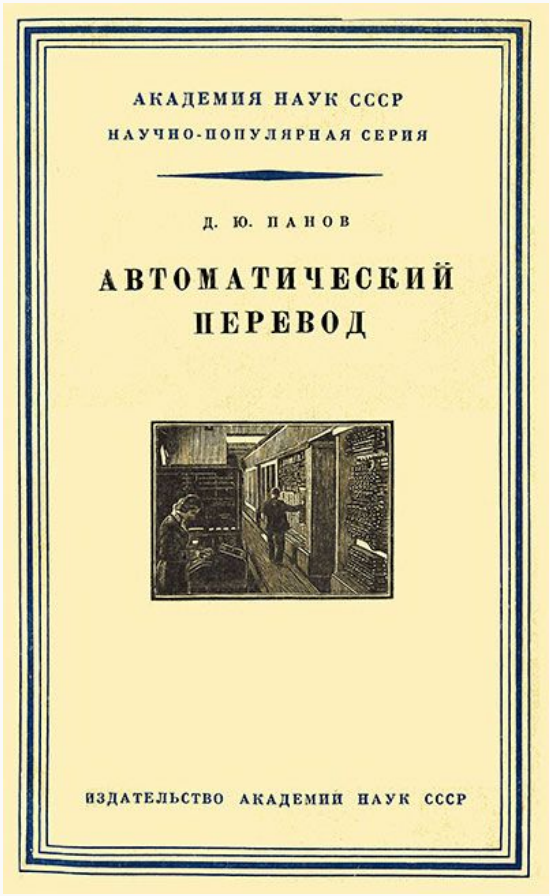




Машинный перевод в 1950-х годах

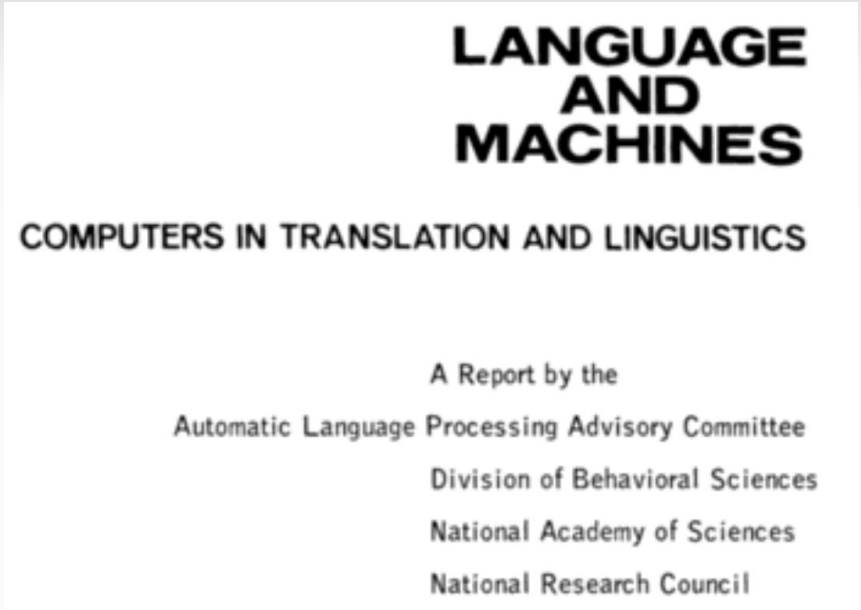
- В СССР в том же году аналогичную программу создали под руководством Дмитрия Панова и Изабеллы Бельской в Институте точной механики и вычислительной техники АН СССР на компьютере БЭСМ





ALPAC

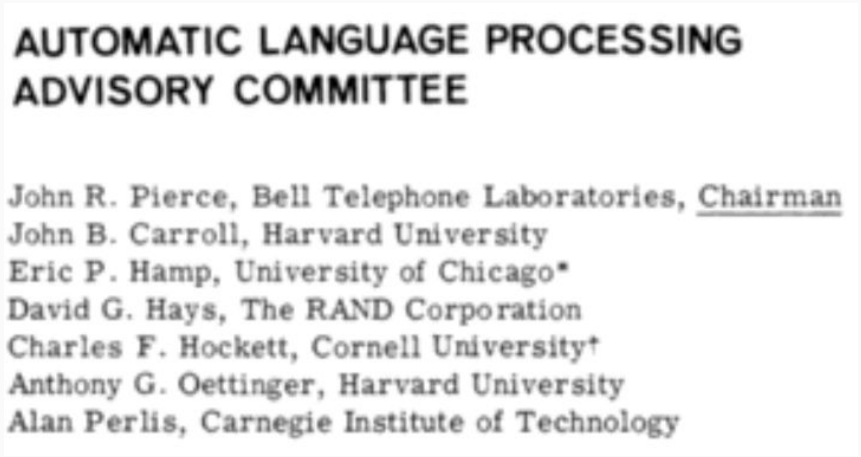
- В 1964 году правительство США организовало ALPAC (Automatic Language Processing Advisory Committee), который был призван оценить прогресс вычислительной лингвистики
- Комитет выпустил отчёт под названием *Language and Machines: Computers in Translation and Linguistics* (1966)
- Выводы были неутешительными



LANGUAGE AND MACHINES

COMPUTERS IN TRANSLATION AND LINGUISTICS

A Report by the
Automatic Language Processing Advisory Committee
Division of Behavioral Sciences
National Academy of Sciences
National Research Council

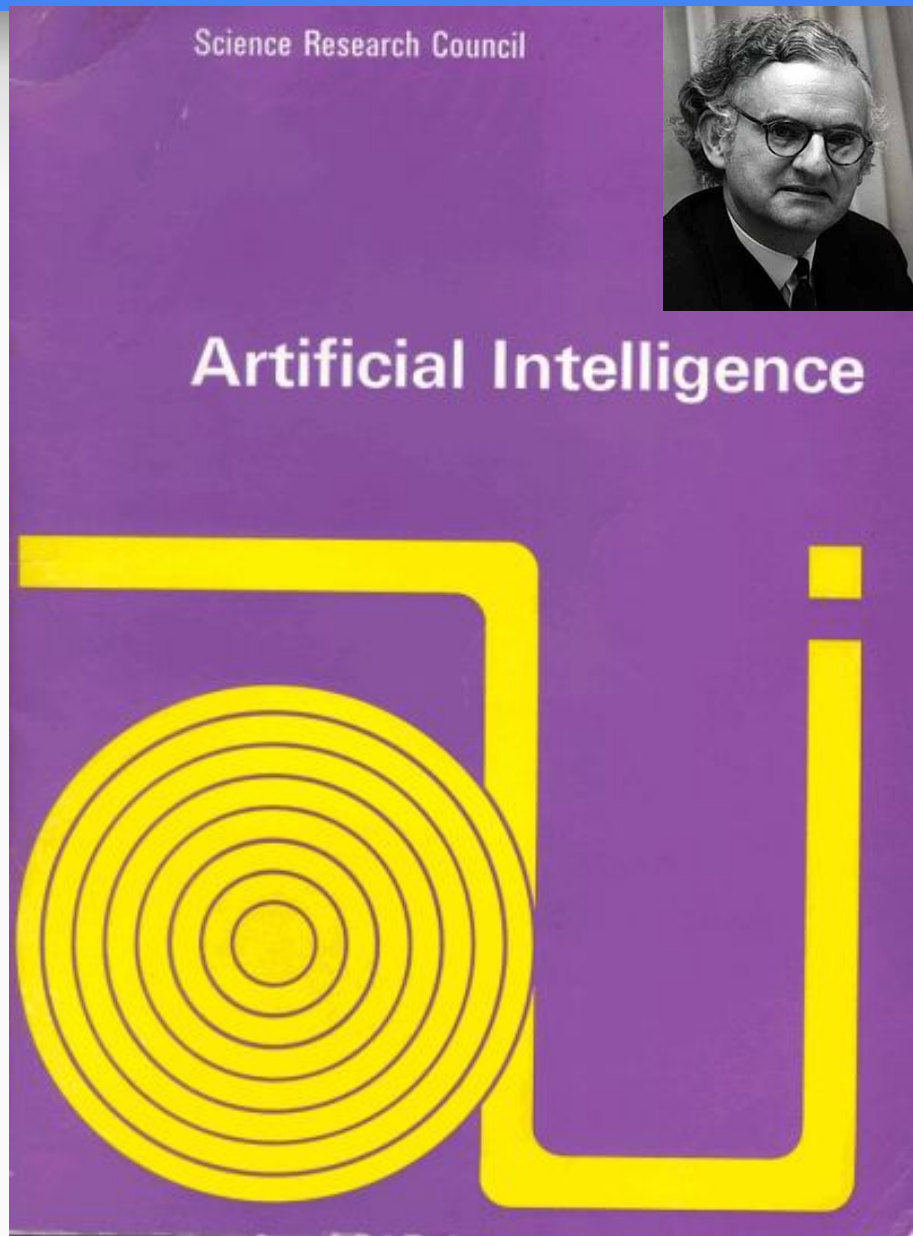


AUTOMATIC LANGUAGE PROCESSING ADVISORY COMMITTEE

John R. Pierce, Bell Telephone Laboratories, Chairman
John B. Carroll, Harvard University
Eric P. Hamp, University of Chicago*
David G. Hays, The RAND Corporation
Charles F. Hockett, Cornell University†
Anthony G. Oettinger, Harvard University
Alan Perlis, Carnegie Institute of Technology

Lighthill report

- В 1973 Джеймс Лайтхилл выпустил статью “Artificial Intelligence: A General Survey” с очень пессимистичными прогнозами
- "In no part of the field have the discoveries made so far produced the major impact that was then promised"
- В результате британское правительство прекратило поддержку AI-исследований почти во всех университетах UK
- В общем, и в США, и в Великобритании люди разочаровались



Зима искусственного интеллекта

- Выводы ALPAC, критика перцептронов, отчёт Лайтхилла, общее разочарование тем, что большие ожидания в реальности не подтвердились…
- Всё это привело к “зиме искусственного интеллекта” (AI winter): периоду примерно со второй половины 1960-х до начала 1980-х, когда AI в современном смысле мало кто всерьёз занимался
- Люди решили, что “это всё не работает” — и тогда действительно так и было!



Зима искусственного интеллекта

- Выводы ALPAC, критика перцептронов, отчёт Лайтхилла, общее разочарование тем, что большие ожидания в реальности не пе
- Всё это пр
  искусс
  wi
  с
  во
- Люд
  работ
  действительно так и было!

А что происходило в течение этой зимы?
Неужели история прекратила течение своё?



AI в 1970-х: экспертные системы



Что такое экспертная система

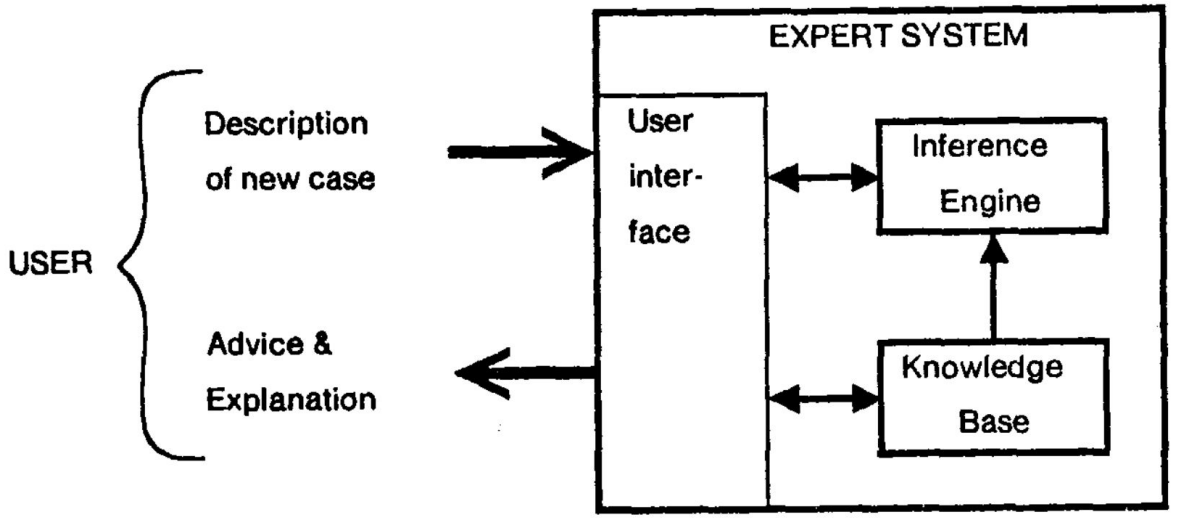
- У экспертной системы есть какая-то база знаний (простых правил) и правила вывода из них
- Правила вывода могут быть чисто логическими, прямо как силлогизмы
- Но это не очень реалистично...

If A, then B (Rule 1)

If B, then C (Rule 2)

A (Data)

∴C (Conclusion)



Что такое экспертная система

- Скорее всего, мы что-то знаем наверняка, а про что-то можем предполагать с некоторой степенью уверенности
- Certainty factors — это в экспертных системах обычно были не вероятности (хотя теперь уже сложно понять, почему), а просто абстрактные числа и некоторые правила операций с ними

h_1 = The identity of ORGANISM-1 is streptococcus
h_2 = PATIENT-1 is febrile
h_3 = The name of PATIENT-1 is John Jones

IF: 1) The stain of the organism is grampos, and
2) The morphology of the organism is coccus, and
3) The growth conformation of the organism is chains
THEN: There is suggestive evidence (.7) that the identity of the organism is streptococcus

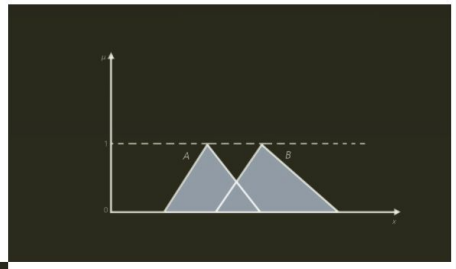
$CF[h_1,E] = .8$	:	There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus
$CF[h_2,E] = -.3$	:	There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile
$CF[h_3,E] = +1$	:	It is definite (1) that the name of PATIENT-1 is John Jones

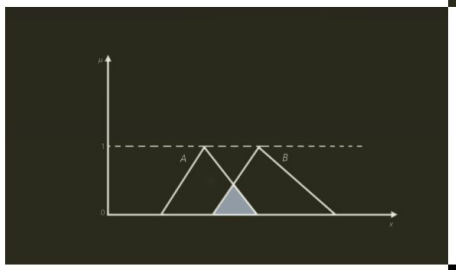
Definition of Certainty Factors (CF) in MYCIN

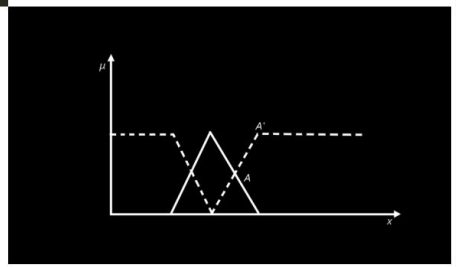
- Range: $-1 \leq CF \leq +1$
- CF level definitions:
  - CF = +1 the fact or rule is certainly true
  - CF = 0 we know nothing about whether the fact or rule is true or not
  - CF = -1 the fact or rule is certainly not true

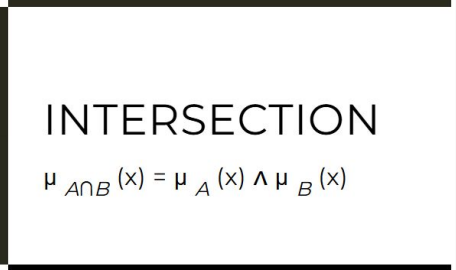
Нечёткая логика

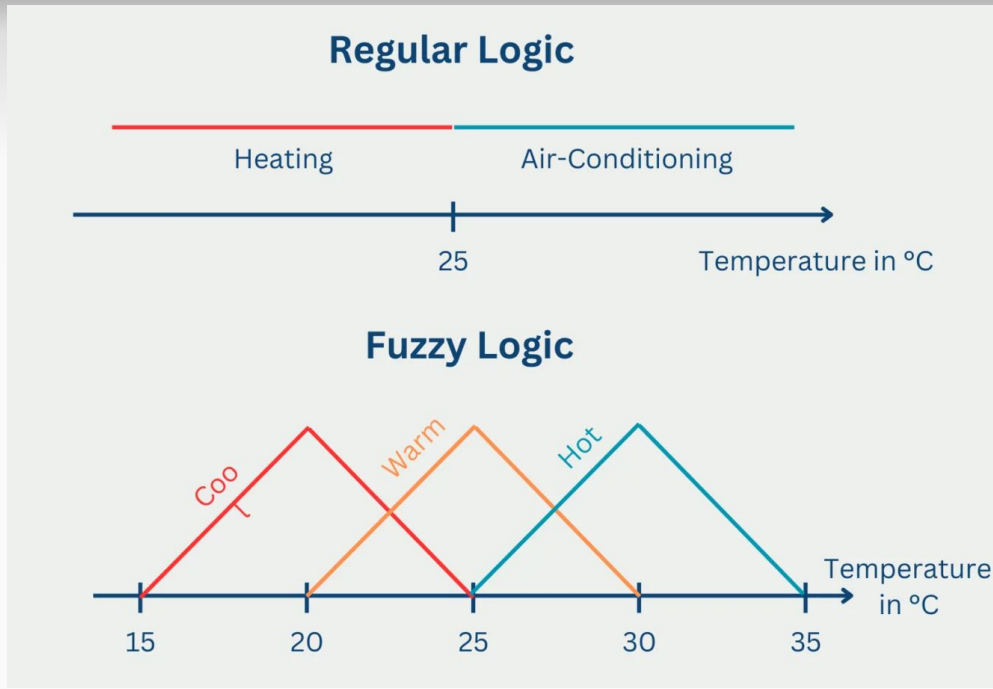
- Всё это ещё очень похоже на нечёткую логику (fuzzy logic), которая тоже в 1970-х активно развивалась (Лотфи Заде)

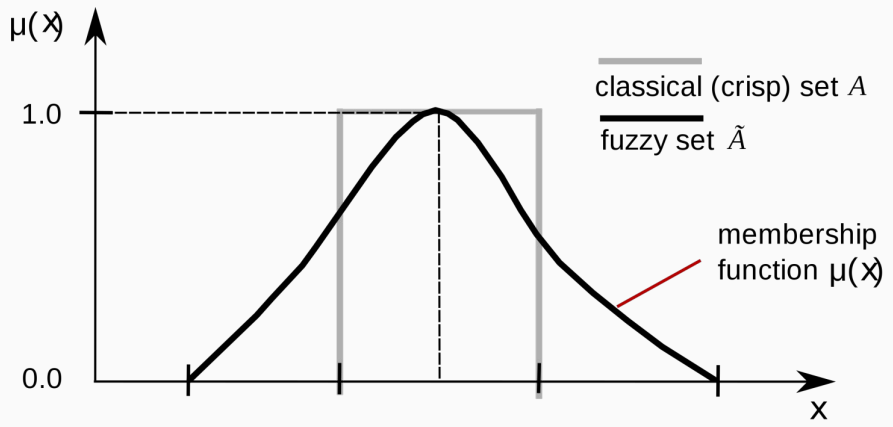












MYCIN

- MYCIN — экспертная система для диагностики; 600 правил

RULE037

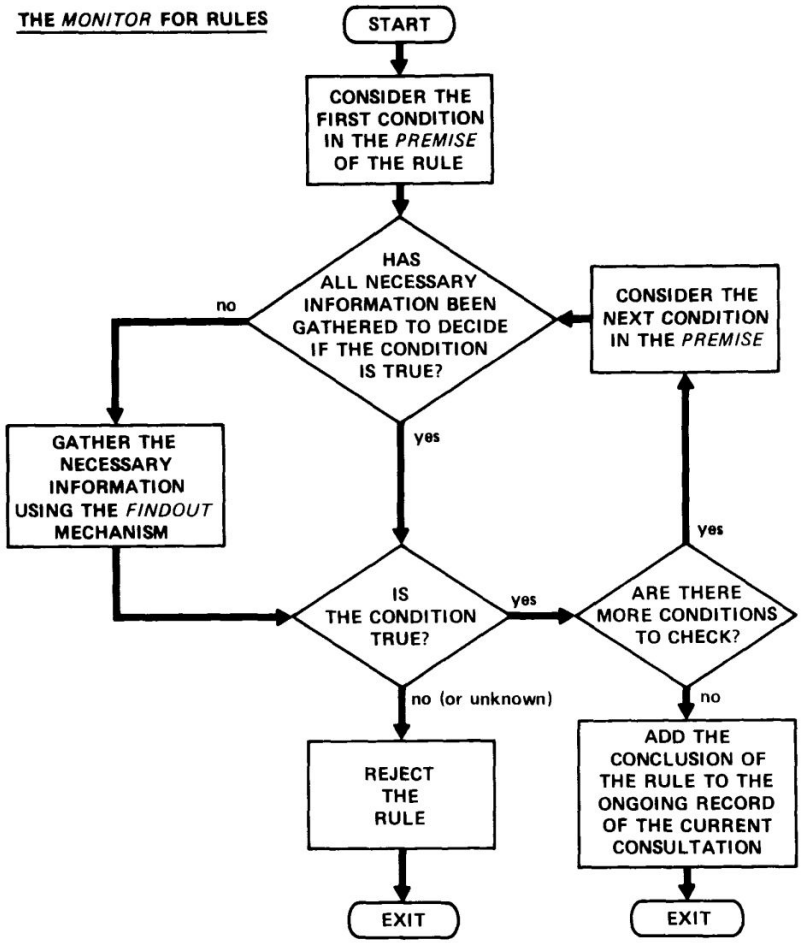
IF: 1) *The identity of the organism is not known with certainty, and*
2) The stain of the organism is gramneg, and
3) The morphology of the organism is rod, and
4) The aerobicity of the organism is aerobic

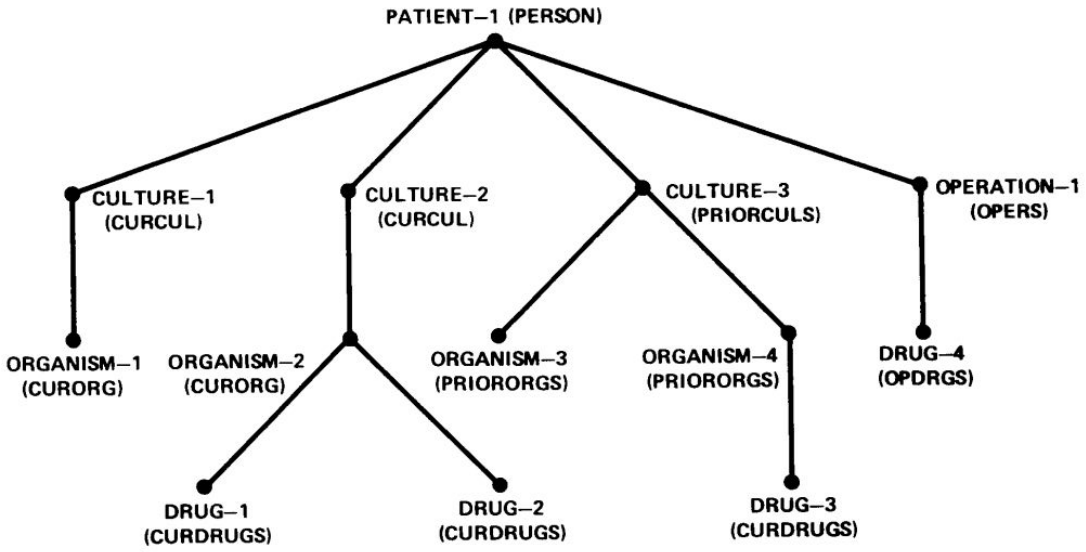
THEN: There is strongly suggestive evidence (.8) that the class of the organism is enterobacteriaceae

RULE145

IF: 1) The therapy under consideration is one of: cephalothin clindamycin erythromycin lincomycin vancomycin, and
2) Meningitis is an infectious disease diagnosis *for the patient*

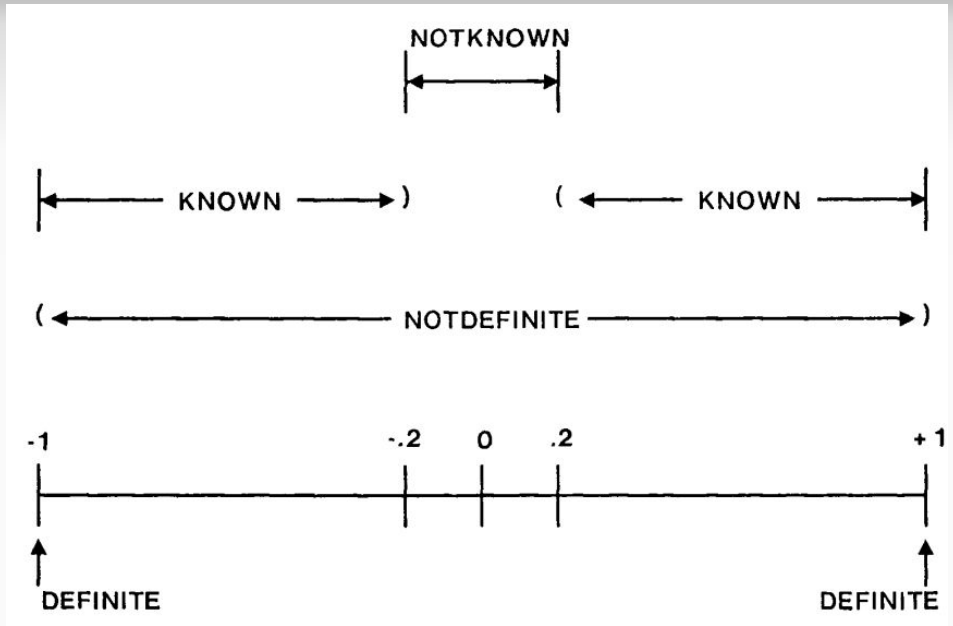
THEN: It is definite (1) the therapy under consideration is not a potential therapy for use against the organism

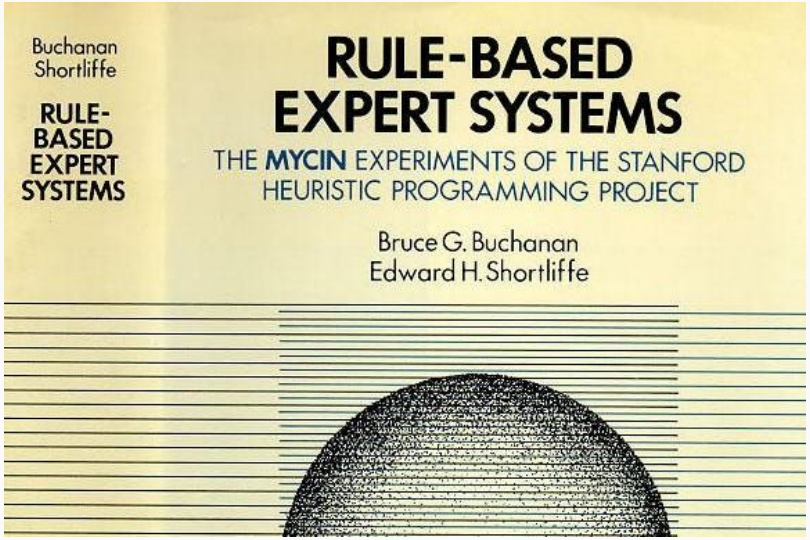




MYCIN

- MYCIN — экспертная система для диагностики
- Результаты во второй половине 1970-х были как у опытного врача и существенно лучше, чем у начинающих врачей
- Система обработки нечётких выражений и неопределённостей
- Давайте рассмотрим простой пример…
- Но, конечно, на практике это было сложно использовать

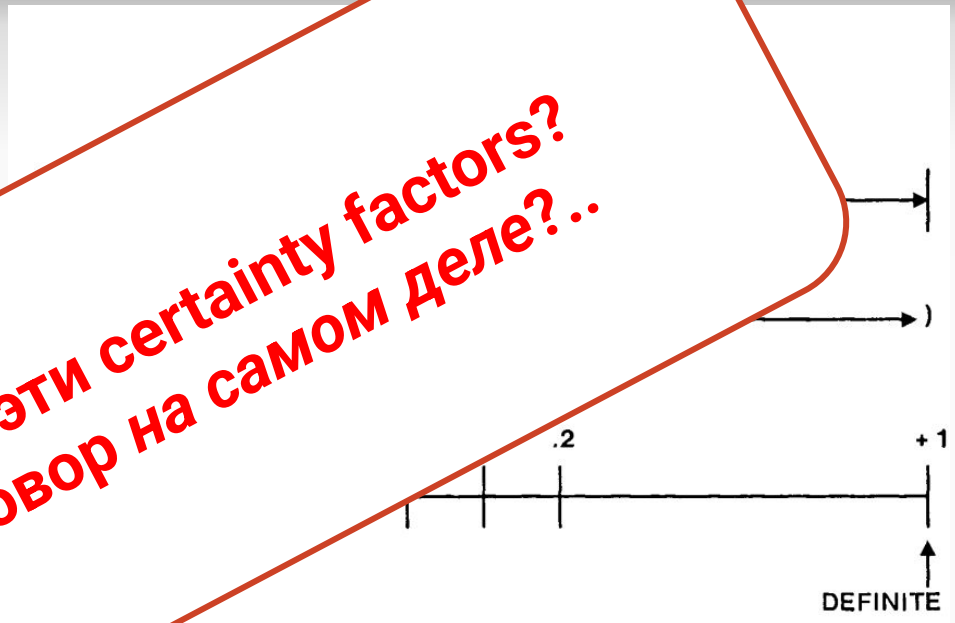


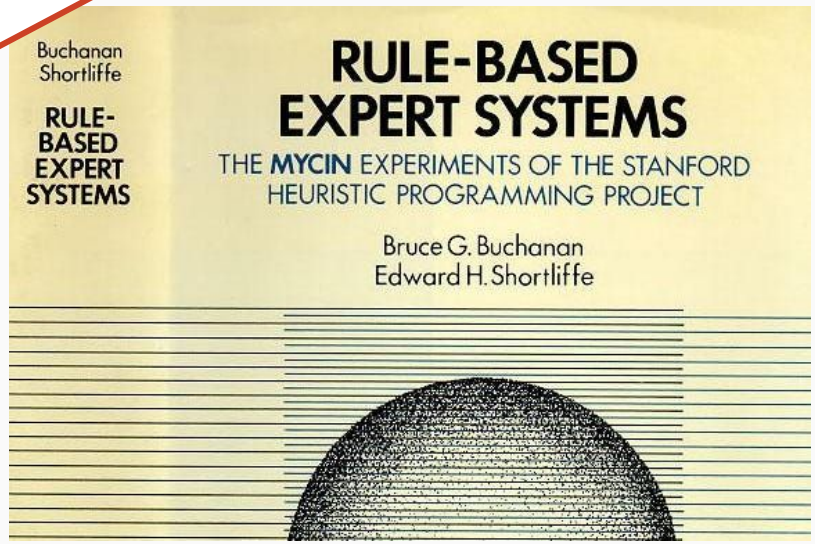


MYCIN

- MYCIN — экспертная система для диагностики
- Результаты уже в середине 1970-х были как у опыт врача и существе чем у начина
- Систе в не
- Дава прост
- Но, коне , на практике это было сложно использовать

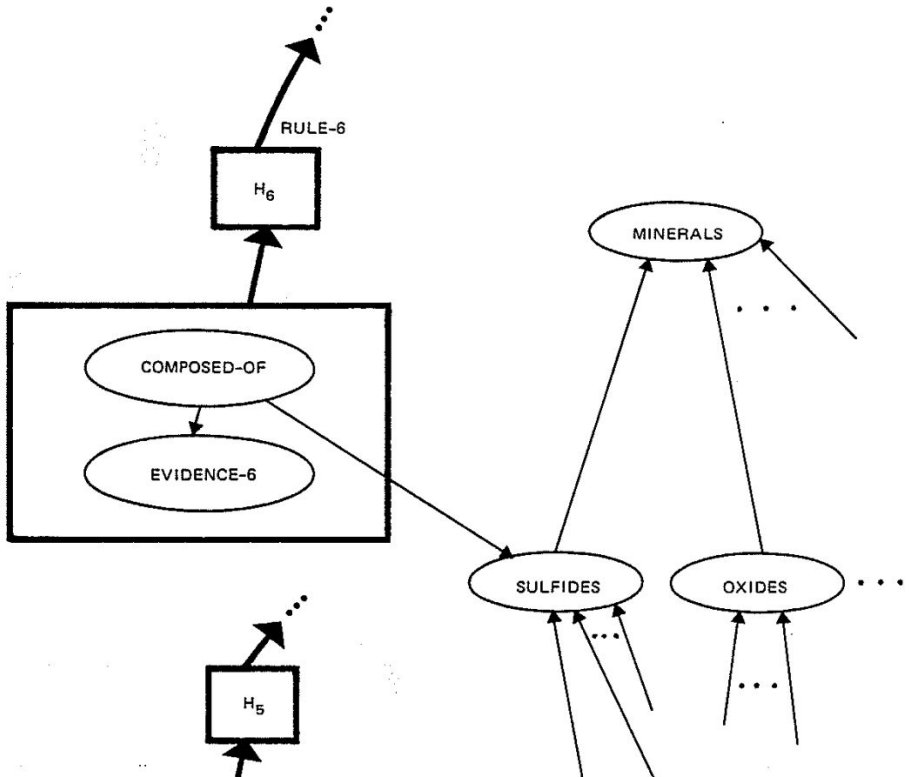
На что похожи все эти certainty factors?
О чём здесь разговор на самом деле?..





Prospector

- PROSPECTOR (Hart, Duda, 1977): экспертная система для поиска месторождений полезных ископаемых




```
1:THERE ARE CARBONATES
  (CARBONATES) (4)
```
[The computer acknowledges the first volunteered evidence; the certainty is taken to be 4.]

```
2:THERE IS DOLOMITE
  (DOLOMITE) (4)

3:THERE IS GALENA
  (GALENA) (4)

4:THERE MIGHT BE SPHALERITE
  (SPHALERITE) (2)

5:DONE
```
[The user is finished volunteering evidence, and PROSPECTOR takes the initiative.]

```
--------------------

You told me about:
  - CARBONATES (4)
```
[The computer reminds the user of volunteered evidence.]

```
6--To what degree do you believe that:
(* Prospective ore body lies in carbonate sediments) ? 3

You told me about:
  - GALENA (4)

You suspected:
  - SPHALERITE (2)

Your statements imply:
  - Sulfides (4)
      ...
      ...

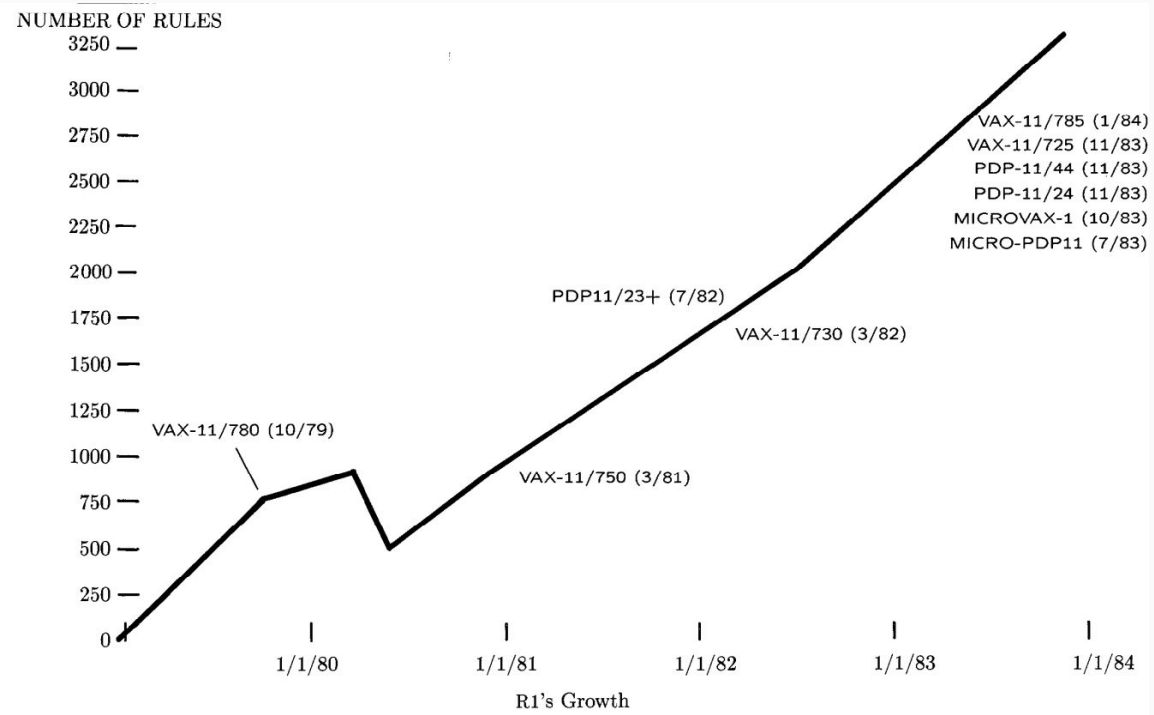
In summary, the following have been observed:

The major favorable conclusions were:
   - (* Suggestive evidence for MVTD) (5.0)
   - (* Prospective ore body lies in carbonate sediments) (3.0)
   - (* Right hydrothermal alteration) (3.0)
   - (* Prospective ore body contains galena and sphalerite) (2.0)

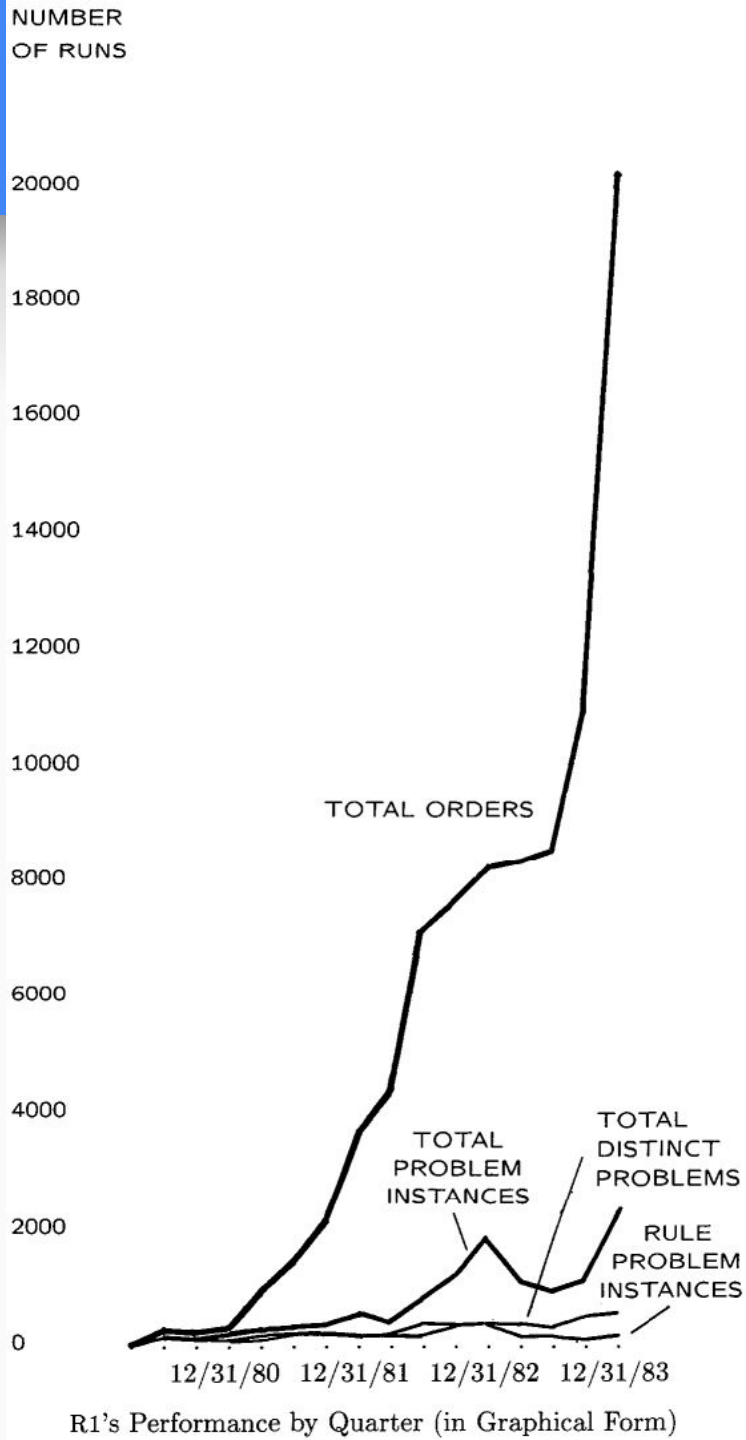
The major unfavorable conclusions were:
   - (* Right temperature of mineralization) (-3.996)
```

XCON/R1

- XCON/R1 (Bachant, McDermott, 1984): экспертная система для конфигурирования компьютеров VAX-11 в компании DEC
- Была постоянно в разработке, но реально применялась, экономила деньги, это был успех экспертных систем



R1's Growth



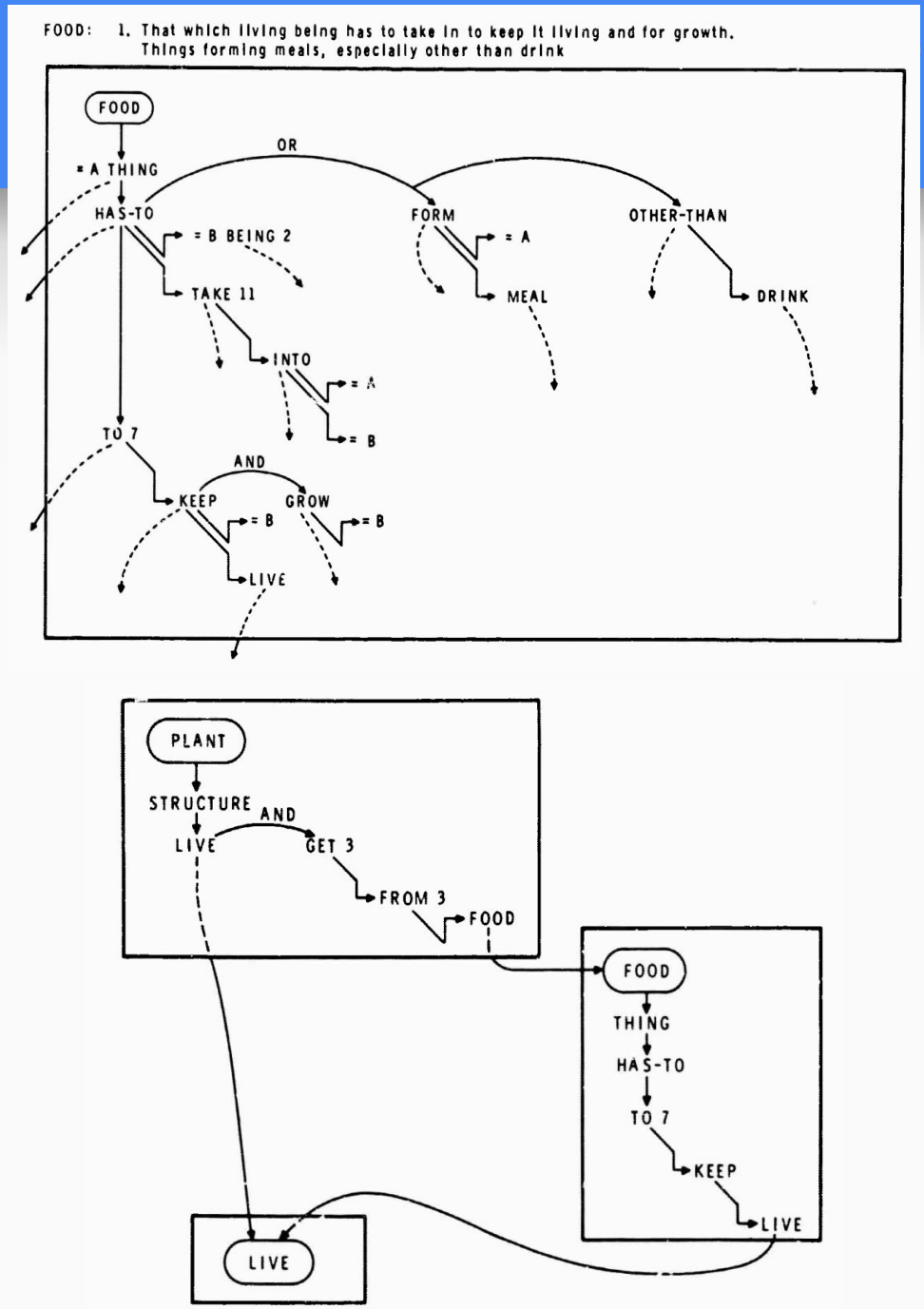
R1's Performance by Quarter (in Graphical Form)

Представление знаний



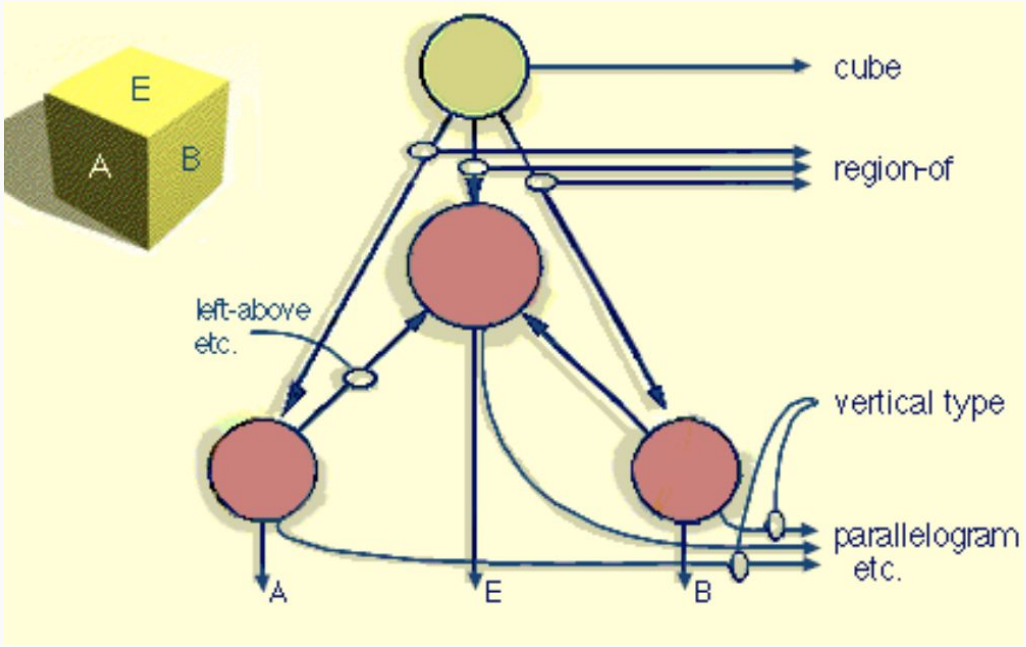
Представление знаний

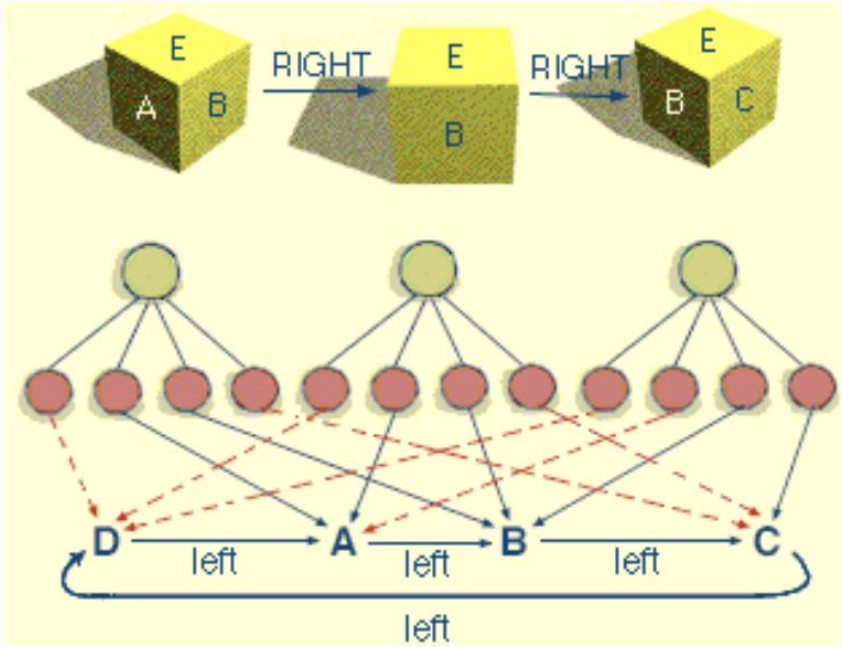
- Другое направление — представление знаний; в MYCIN “база знаний” — это просто набор правил, но этим, конечно, дело не исчерпывается
- Как нам представить знания человечества в машиночитаемой форме? Это вопрос до сих пор вполне актуальный
- Quillian (1968): *Semantic Memory*; здесь уже есть отношения вроде PART-OF, наследование свойств по иерархии и т.п.



Представление знаний

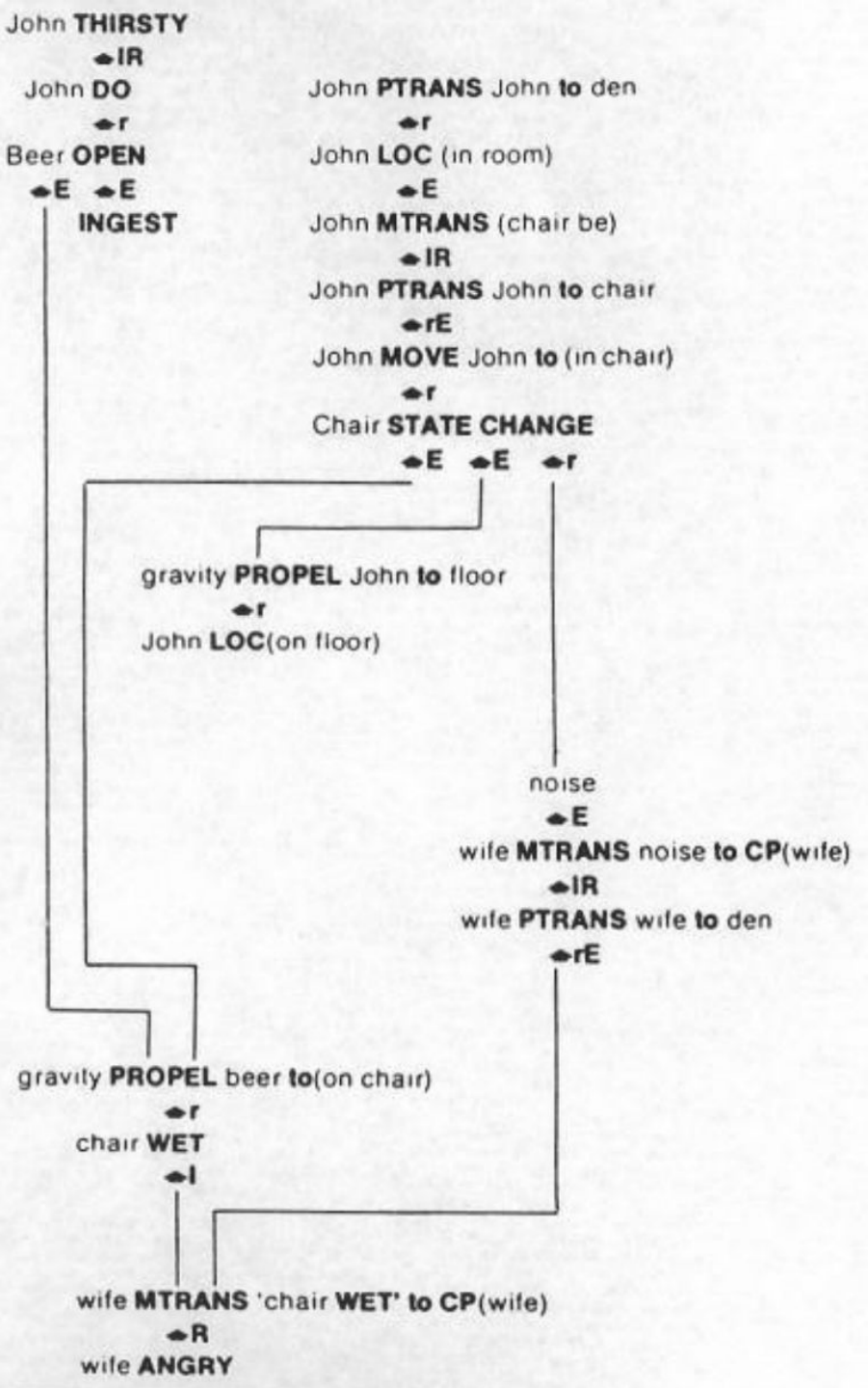
- Minsky (1974): *A Framework for Representing Knowledge*
- Фреймы, значения по умолчанию, преобразования, которые можно применять автоматически; довольно громоздкая система, но и цель была амбициозная: формализовать визуальные знания для компьютерного зрения





Представление знаний

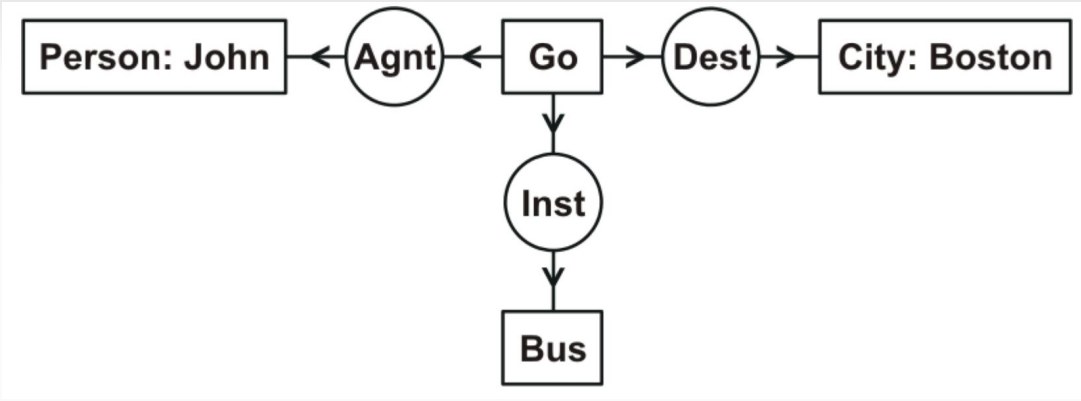
- Schank, Abelson (1977): *Scripts, Plans, Goals, and Understanding*
- Понимание мира через скрипты разных событий и их планирование
- Подход скорее с психологической стороны изначально (как человек это всё структурирует?), но оказавшийся полезным для AI

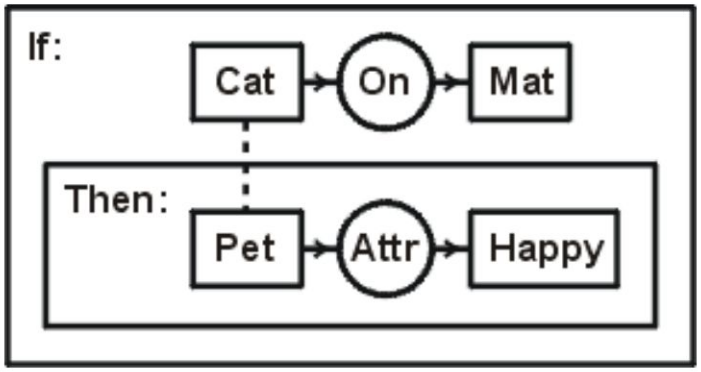


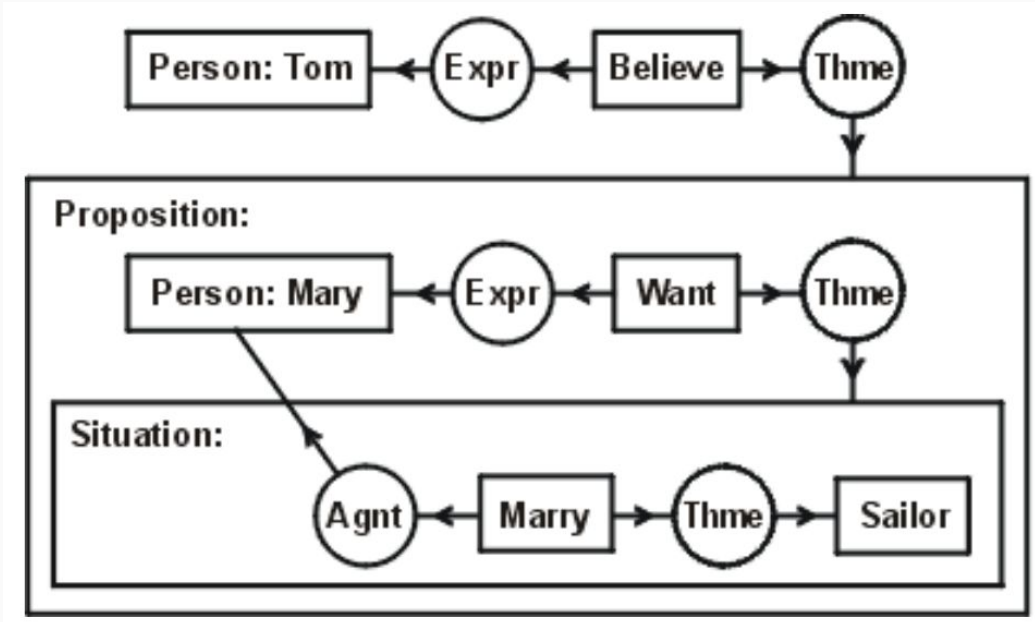
The above representation says, essentially this: John's thirst caused him to decide to **DO** an action that would result in a beer being opened which would enable him to drink it. What we have is: the **THIRST** initiated an **MBUILD** (left out) to open the beer which was the **REASON** that he did the action (whatever it was) that resulted in the beer being open. The beer being open enabled another action (spilling) that happened much later in the sequence of events. (States enable events but the events that are enabled can occur any time after the state is initially present just as long as the state continues to be present.)

Представление знаний

- John Sowa (1976): *Conceptual Graphs for a Database Interface*
- Прямоугольники — концепты, круги — концептуальные отношения
- Можно распространить на условные утверждения, утверждения об утверждениях

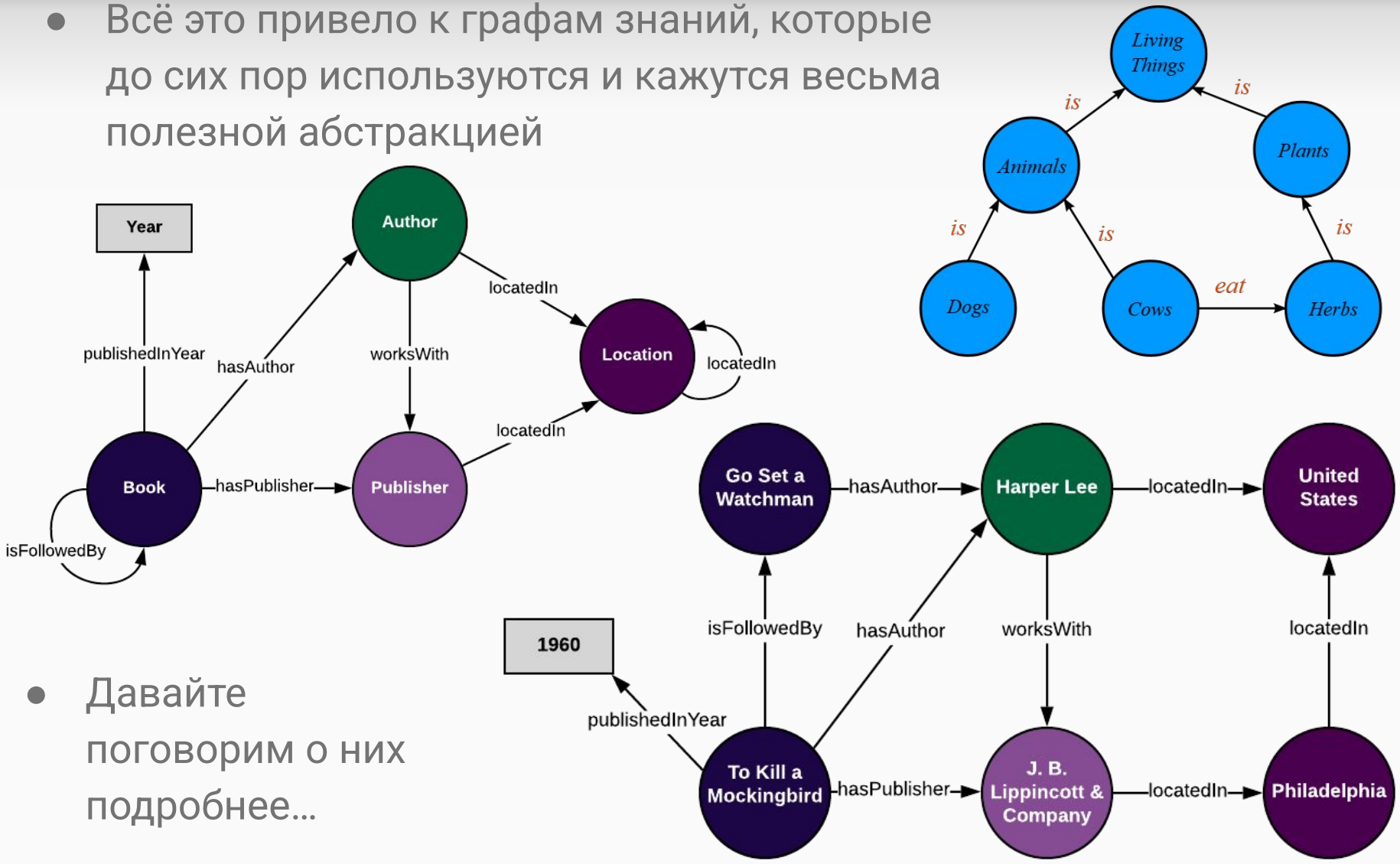






Графы знаний

- Всё это привело к графам знаний, которые до сих пор используются и кажутся весьма полезной абстракцией

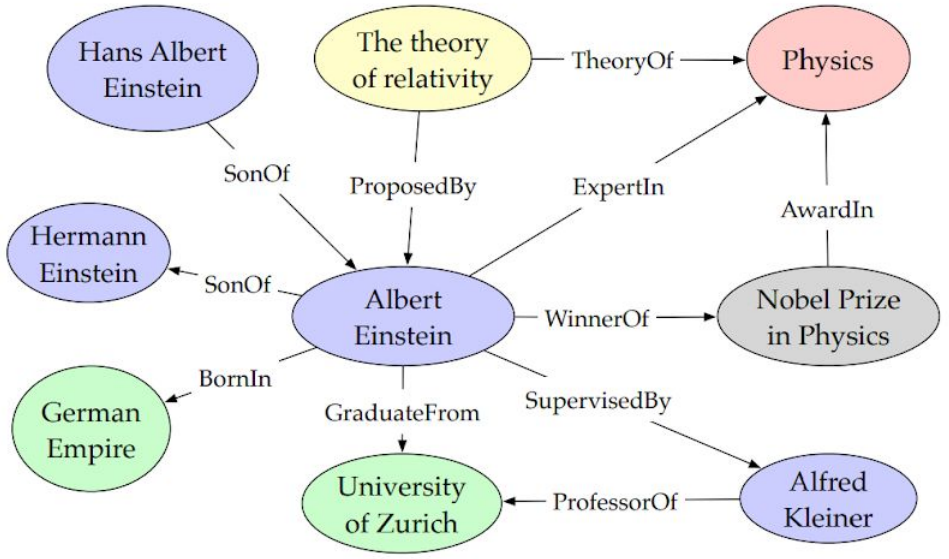


- Давайте поговорим о них подробнее...

Графы знаний

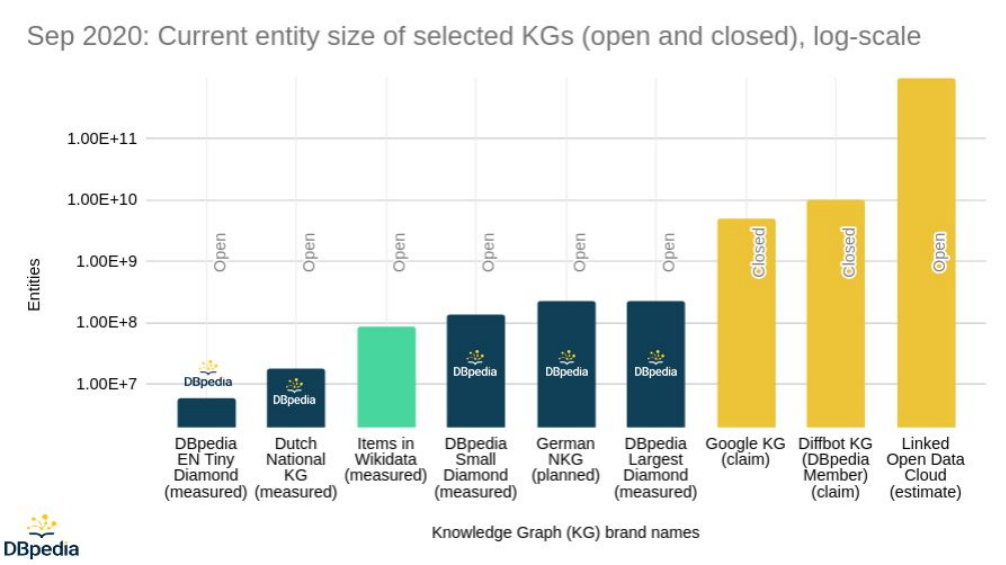
- Графы знаний (knowledge graphs) кажутся мне недооценёнными; это хорошо развитая область (Hogan et al., 2022; Ji et al., 2021; Heist et al., 2020; Yan et al., 2018), но кажется, что их можно было бы использовать чаще
- Вершины — объекты, рёбра — отношения между ними
- Объектов много (могут быть миллиарды), отношений мало (тысячи)

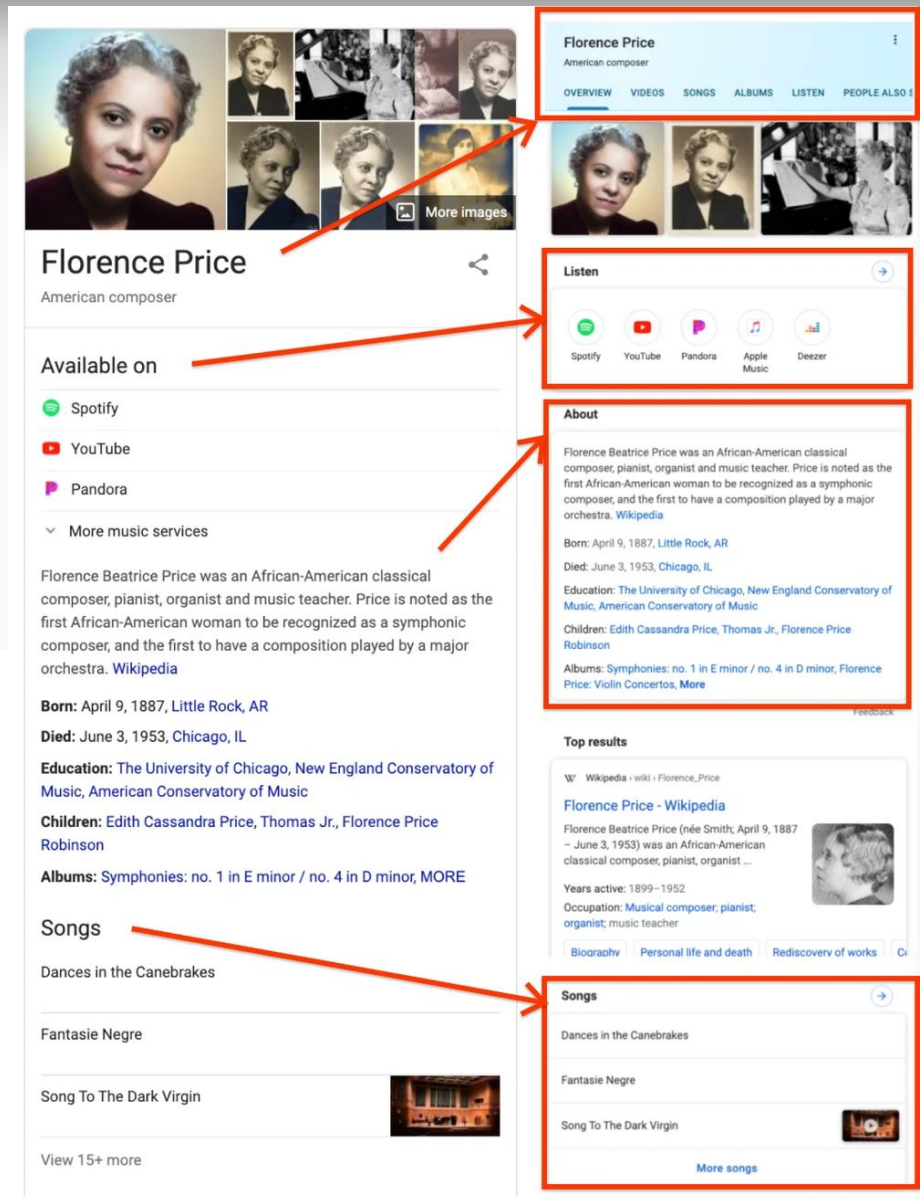
(Albert Einstein, **BornIn**, German Empire)
(Albert Einstein, **SonOf**, Hermann Einstein)
(Albert Einstein, **GraduateFrom**, University of Zurich)
(Albert Einstein, **WinnerOf**, Nobel Prize in Physics)
(Albert Einstein, **ExpertIn**, Physics)
(Nobel Prize in Physics, **AwardIn**, Physics)
(The theory of relativity, **TheoryOf**, Physics)
(Albert Einstein, **SupervisedBy**, Alfred Kleiner)
(Alfred Kleiner, **ProfessorOf**, University of Zurich)
(The theory of relativity, **ProposedBy**, Albert Einstein)
(Hans Albert Einstein, **SonOf**, Albert Einstein)



Графы знаний

- Например, Google Knowledge Graph — это структура, из которой происходят панели с фактами в Google
- DBpedia — граф знаний из Wikipedia; есть графы знаний с миллиардами вершин, и ими можно пользоваться



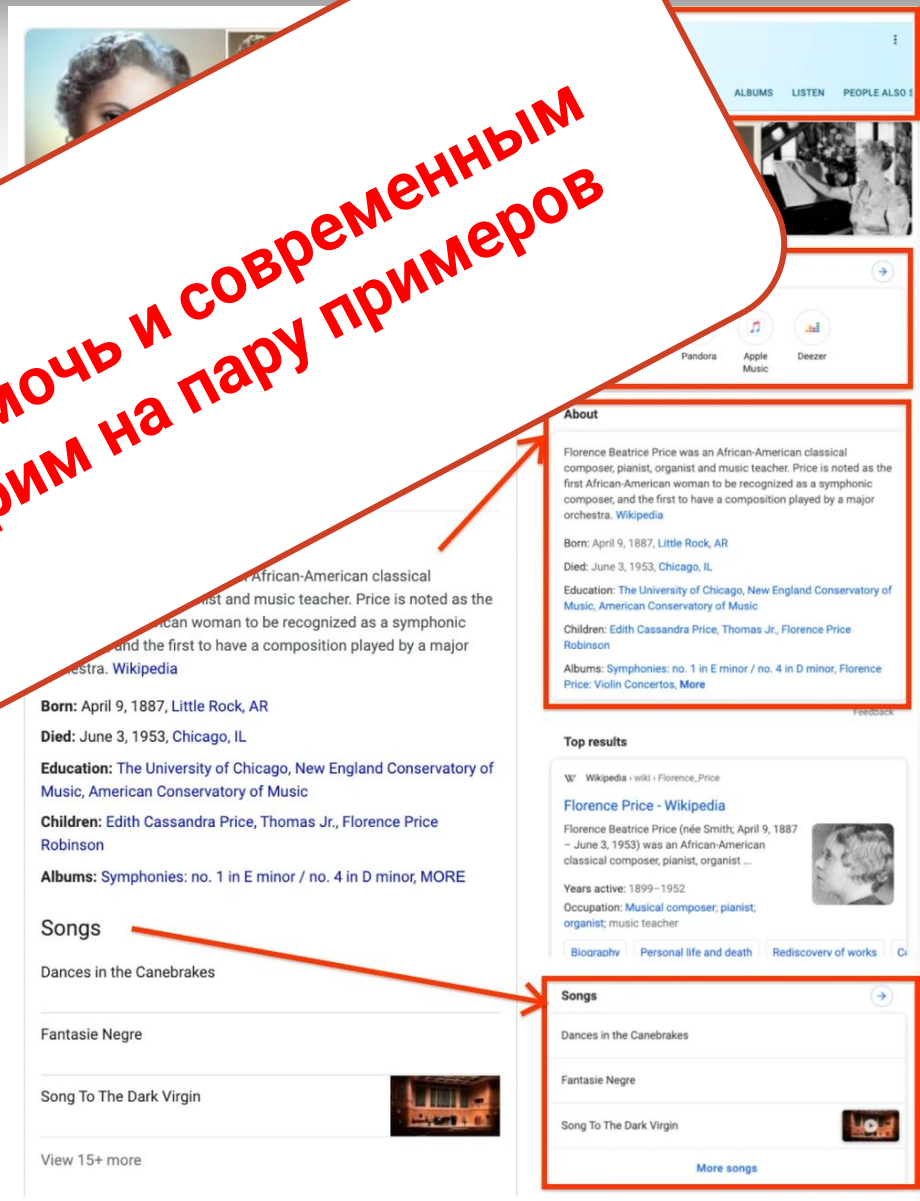


Графы знаний

- Например, Google Knowledge Graph — это структура, из которой происходят панели с фактами в Google
- DBpedia — граф зн
  Wikipedia; есть
  с миллио
  им

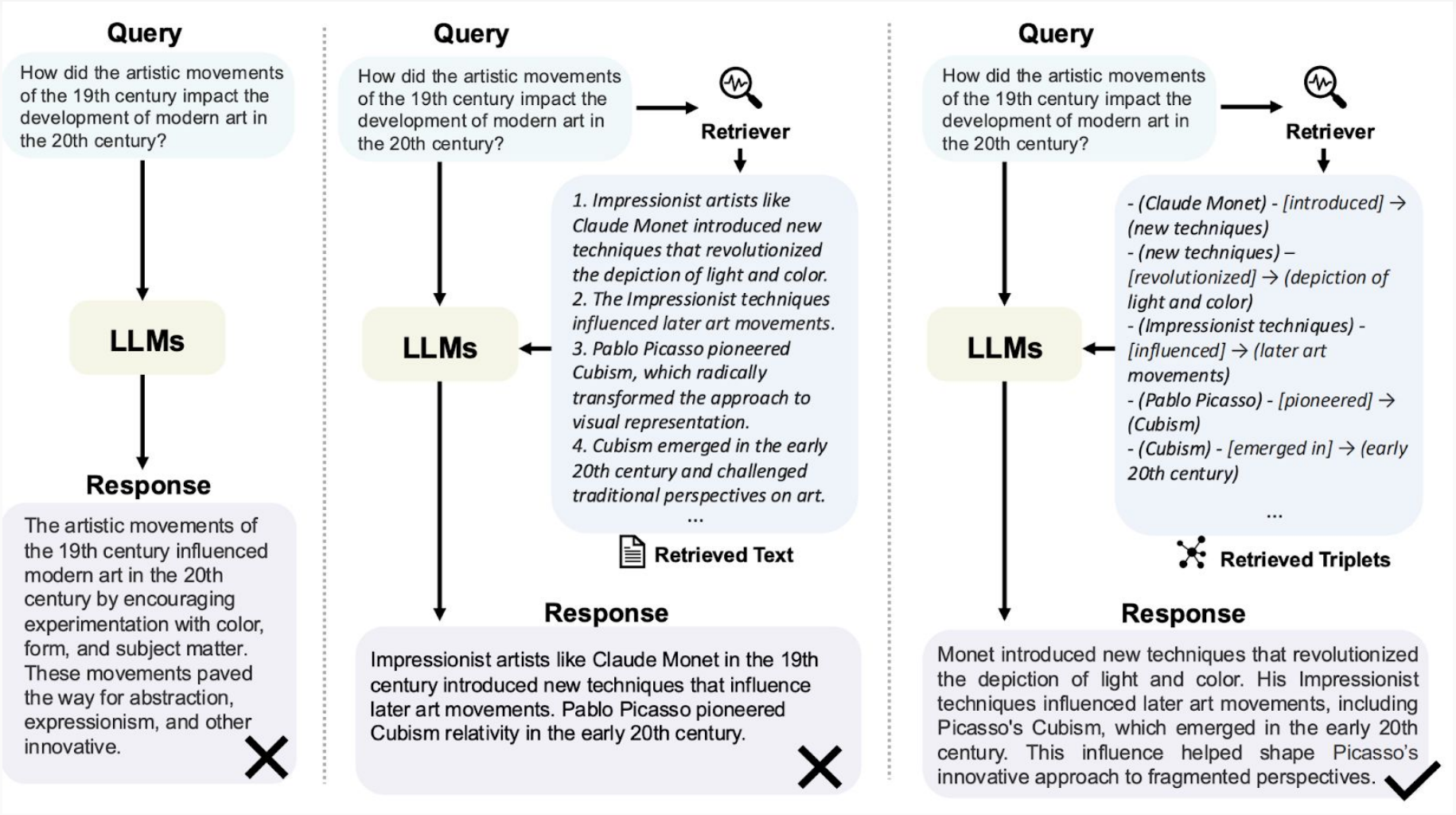
Графы знаний могут помочь и современным LLM! Давайте посмотрим на пару примеров





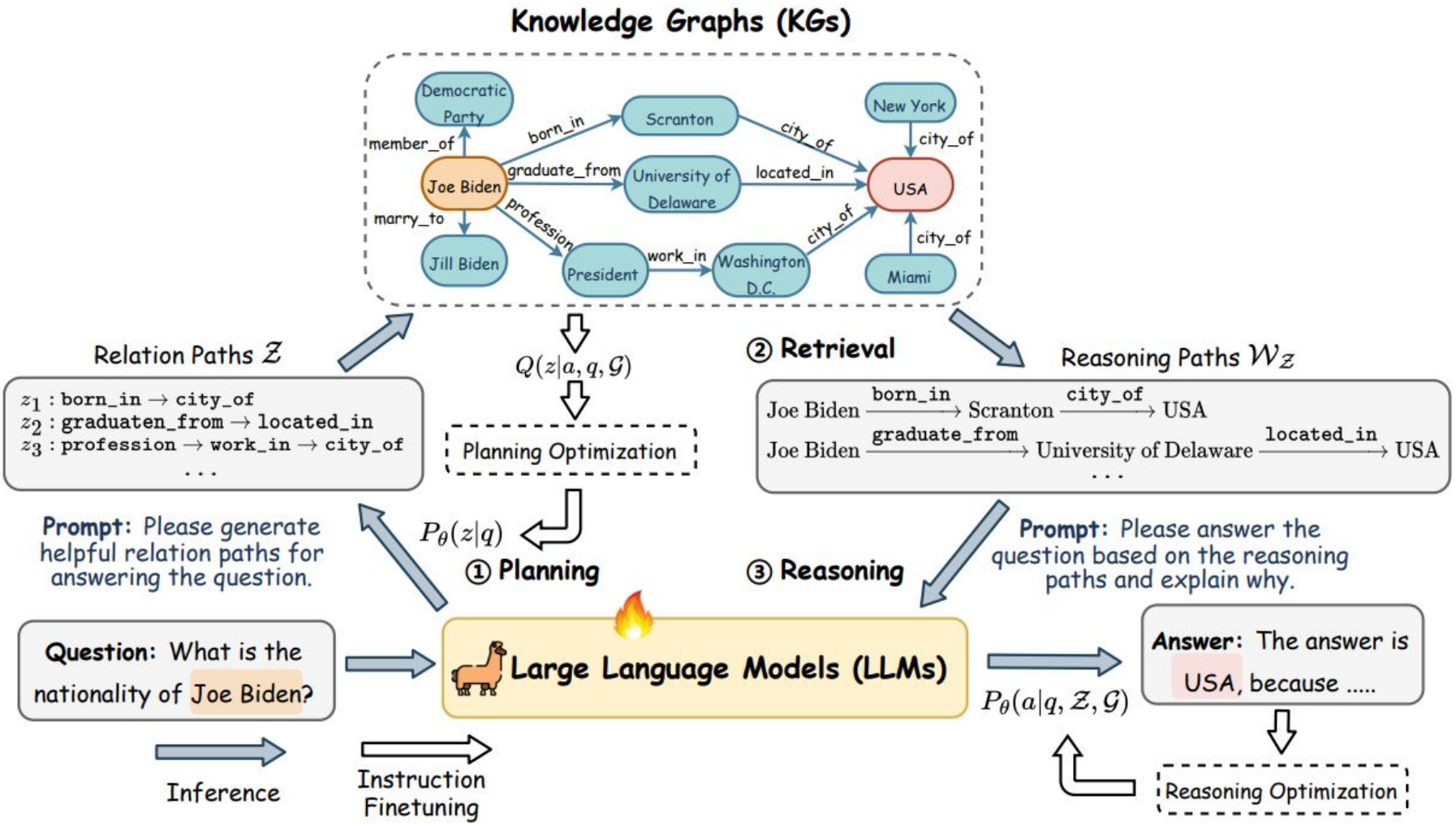
Графы знаний

- Peng et al. (2024): в контексте RAG (retrieval-augmented generation), если искать структурированные триплеты, ответы становятся точнее



Графы знаний

- Luo et al. (2024): reasoning on graphs (RoG)



Графы знаний

- Luo et al. (2024): reasoning on graphs (RoG)
- Графы знаний дают структурированную информацию, которая уже сразу предназначена для логического вывода и сбора данных

- Структура, которую даёт граф знаний, помогает лучше отвечать на вопросы и избегать галлюцинаций, в том числе если это делает LLM

Question	What type of government is used in the country with Northern District?
Answer	Parliamentary system
ChatGPT + CoT	To determine what type of government is used in a country with a Northern District, we need more specific information. The term "Northern District" does not specify the name of any particular country or provide enough context to identify its government type. Could you please provide additional details or clarify the specific country you are referring to?
Reasoning Paths	Northern District → location.administrative_division.first_level_division_of → Israel → government.form_of_government.countries → **Parliamentary system**
RoG	The type of government used in the country with the Northern District is a **Parliamentary system**. Explanations: 1. Northern District → location.administrative_division.first_level_division_of → Israel: The Northern District is a part of Israel. 2. Israel → government.form_of_government.countries → Parliamentary system: Israel has a Parliamentary system of government. So, by following the reasoning paths provided, we can conclude that the country with the Northern District, which is Israel, uses a Parliamentary system of government.

(a) RoG gathers additional knowledge to answer a question

Графы знаний

- Luo et al. (2024): reasoning on graphs (RoG)
- Графы знаний дают структурированную информацию, которая уже сразу предназначена для логического вывода и сбора данных

- Структура, которую даёт граф знаний, помогает лучше отвечать на вопросы и избегать галлюцинаций, в том числе если это делает LLM

Question	Where is the home stadium of the team who won the 1946 World Series championship?
Answer	Busch Stadium
ChatGPT + CoT	Step 1: Identify the team that won the 1946 World Series championship. The team that won the 1946 World Series championship was the St. Louis Cardinals. Step 2: Determine the home stadium of the St. Louis Cardinals during that time. The St. Louis Cardinals' home stadium during the 1946 season was Sportsman's Park. Therefore, the possible answer to the question "Where is the home stadium of the team who won the 1946 World Series championship?" is: - Sportsman's Park.
Reasoning Paths	1946 World Series → sports.sports_team.championships → St. Louis Cardinals → sports.sports_team.arena_stadium → **Busch Stadium** 1946 World Series → sports.sports_team.championships → St. Louis Cardinals → sports.sports_team.arena_stadium → Roger Dean Stadium
RoG	Based on the reasoning paths provided, the home stadium of the team who won the 1946 World Series championship is **Busch Stadium**. This conclusion is based on the following path: **1. 1946 World Series → sports.sports_team.championships → St. Louis Cardinals → sports.sports_team.arena_stadium → Busch Stadium** The reasoning path leads us to the St. Louis Cardinals as the team that won the 1946 World Series, and Busch Stadium is the stadium associated with the St. Louis Cardinals. Therefore, **Busch Stadium** is the home stadium of the team that won the 1946 World Series championship.

(b) RoG successfully avoids hallucination

Графы знаний

- Belova et al. (Oct 10, 2025): с другой стороны, LLM могут помочь собирать графы знаний, но это оказывается не так просто

Prompt. Please complete the following medical KG triple (head, relation, tail):
`(chronic kidney disease, has_finding_site, ...)`

based on the sequence:

Chronic kidney disease (CKD) is a renal disorder. However, neuroimaging cohorts of people with CKD have reported subtle, group-level abnormalities within the cerebellar gray matter. Studies note modest reductions or regional alterations in cerebellar gray-matter volume, cortical thickness, and diffusion metrics. These observations are heterogeneous and not diagnostic, yet they place observable findings at the cerebellar gray matter in CKD. Thus, CKD may be associated with cerebellar gray-matter changes.

Gemini 2.5 Pro: (chronic kidney disease, has_finding_site, cerebellar gray matter)

The provided text explicitly states that neuroimaging studies "place observable findings at the cerebellar gray matter in CKD."

GPT-5: Based on the medical knowledge graph triple format and the provided sequence, the completed triple would be: (chronic kidney disease, has_finding_site, cerebellar gray matter)

This is based on the text where chronic kidney disease (CKD) is associated with observable findings in the cerebellar gray matter, specifically mentioning reductions in cerebellar gray-matter volume, cortical thickness, and diffusion metrics.

Графы знаний

- Belova et al. (Oct 10, 2025): с другой стороны, LLM могут помочь собирать графы знаний, но это оказывается не так просто

Prompt. Please complete the following medical KG triple (head, relation, tail):
`(chronic kidney disease, has_finding_site, ...)`

based on the sequence:

Chronic kidney disease (CKD) is a renal disorder. However, neuroimaging cohorts of people with CKD have reported subtle, group-level abnormalities within the cerebellar gray matter. Studies note modest reductions or regional alterations in cerebellar gray-matter volume, cortical thickness, and diffusion metrics. These observations are heterogeneous and not diagnostic, yet they place observable findings at the cerebellar gray matter in CKD. Thus, CKD may be associated with cerebellar gray-matter changes.

Gemini 2.5 Pro: (chronic kidney disease, has_finding_site, cerebellar gray matter)

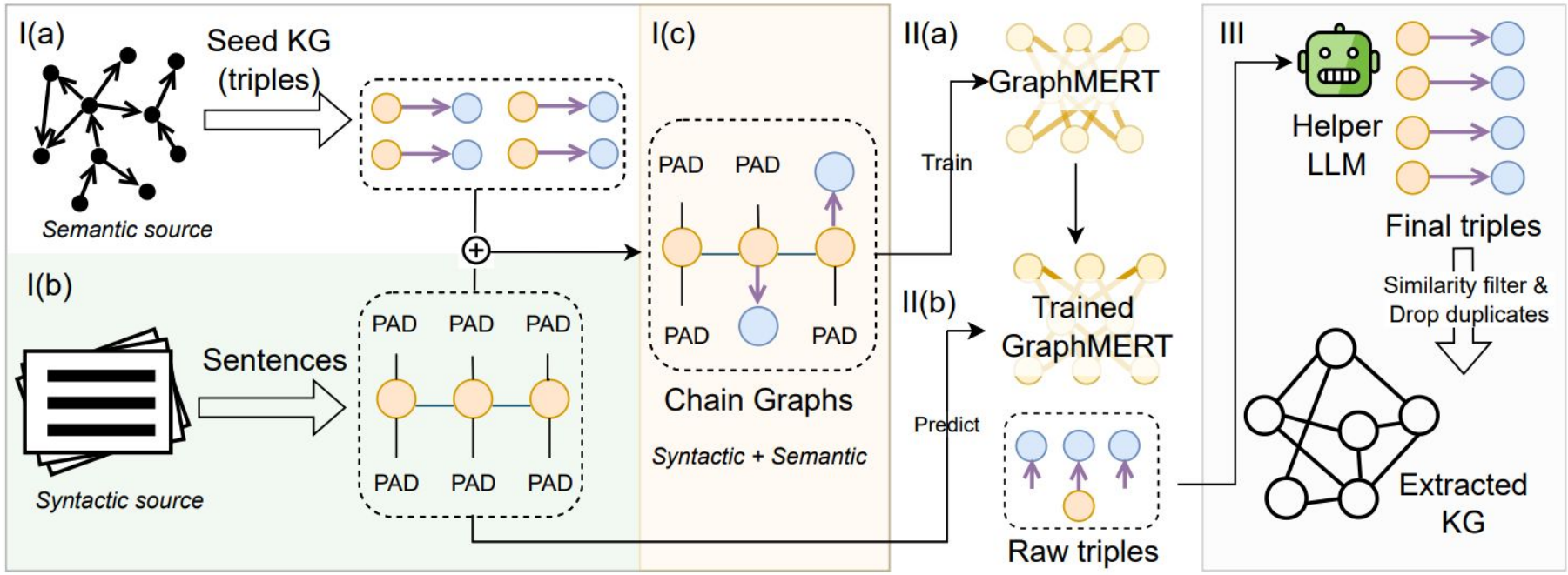
The provided text explicitly states that neuroimaging studies "place observable findings at the cerebellar gray matter in CKD."

GPT-5: Based on the medical knowledge graph triple format and the provided sequence, the completed triple would be: (chronic kidney disease, has_finding_site, cerebellar gray matter)

This is based on the text where chronic kidney disease (CKD) is associated with observable findings in the cerebellar gray matter, specifically mentioning reductions in cerebellar gray-matter volume, cortical thickness, and diffusion metrics.

Графы знаний

- Belova et al. (Oct 10, 2025): GraphMERT; специальная архитектура (на основе RoBERTa) объединяет синтаксическую информацию (обычные предложения) с семантической (граф знаний) и выдаёт кандидатов в дополнение троек
- В детали сейчас вдаваться не будем, но это (экзотический) пример построения *общего мультимодального представления*; что это значит?



Графы знаний

- Belova et al. (Oct 10, 2025): и действительно помогает (с LLM, не сам)

Prompt. Please complete the following medical KG triple (head, relation, tail):
`(chronic kidney disease, has_finding_site, ...)`

based on the sequence:

Chronic kidney disease (CKD) is a renal disorder. However, neuroimaging cohorts of people with CKD have reported subtle, group-level abnormalities within the cerebellar gray matter. Studies note modest reductions or regional alterations in cerebellar gray-matter volume, cortical thickness, and diffusion metrics. These observations are heterogeneous and not diagnostic, yet they place observable findings at the cerebellar gray matter in CKD. Thus, CKD may be associated with cerebellar gray-matter changes.

GraphMERT:
structure, kidney, system, function, renal, disease, heart, kidneys, of, physiology, arterial, underlying, pancreas, organ, impaired, health, cerebrovascular, pathophysiology, injury, morphology

Gemini 2.5 Pro reasons over GraphMERT triple candidates:
1. Analyze the Input:
Sequence: The text describes Chronic Kidney Disease (CKD) as a "renal disorder," indicating its primary location is the kidneys. It also explicitly mentions that in people with CKD, neuroimaging has found "abnormalities within the cerebellar gray matter."

...

4. Final Decision: **The best and most directly supported tail I can construct from the given tokens is kidneys. While the text mentions other sites, the provided tokens do not allow for their accurate creation.**

Графы знаний

- Belova et al. (Oct 10, 2025): и действительно помогает (с ... 4, не сам)

Prompt. Please complete the following medical KG triple (head, relati
`(chronic kidney disease, has_finding_site, ...)`

based on the sequence:

Chronic kidney disease (CKD) is a renal disord
have reported subtle, group-level abnormali
reductions or regional alterations in c
rics. These observations are hete
cerebellar gray matter in C

...net-
...gs at the
...r changes.

GraphMERT:
structure, k
organ

...ogy, arterial, underlying, pancreas,
...ology

Ge
1. A
Sequ
is the
within t

...dates:
... (CKD) as a "renal disorder," indicating its primary location
...at in people with CKD, neuroimaging has found "abnormalities

...

4. Final De
kidneys. W

...and most directly supported tail I can construct from the given tokens is
...ext mentions other sites, the provided tokens do not allow for their accurate creation.

А теперь вернёмся к тому, что анонсировали в разговоре об экспертных системах. Какова природа машинного обучения, как его правильно формализовать?..

Байесовский подход к вероятностям

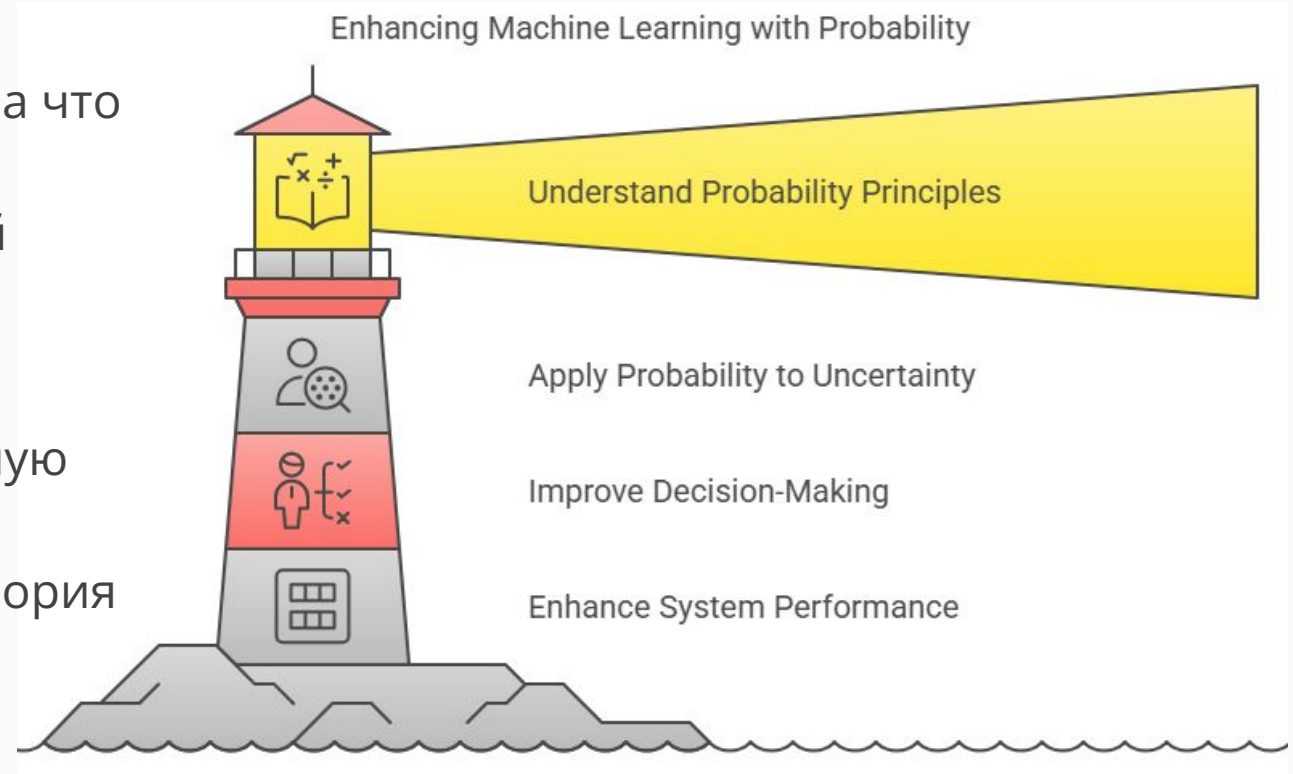


Почему теория вероятностей

- Во всех методах и подходах машинного обучения очень пригодился бы метод, который мог бы не просто выдавать ответ, а ещё оценивать, насколько модель уверена в этом ответе, насколько модель хорошо описывает данные, как изменятся эти величины при дальнейших экспериментах и так далее

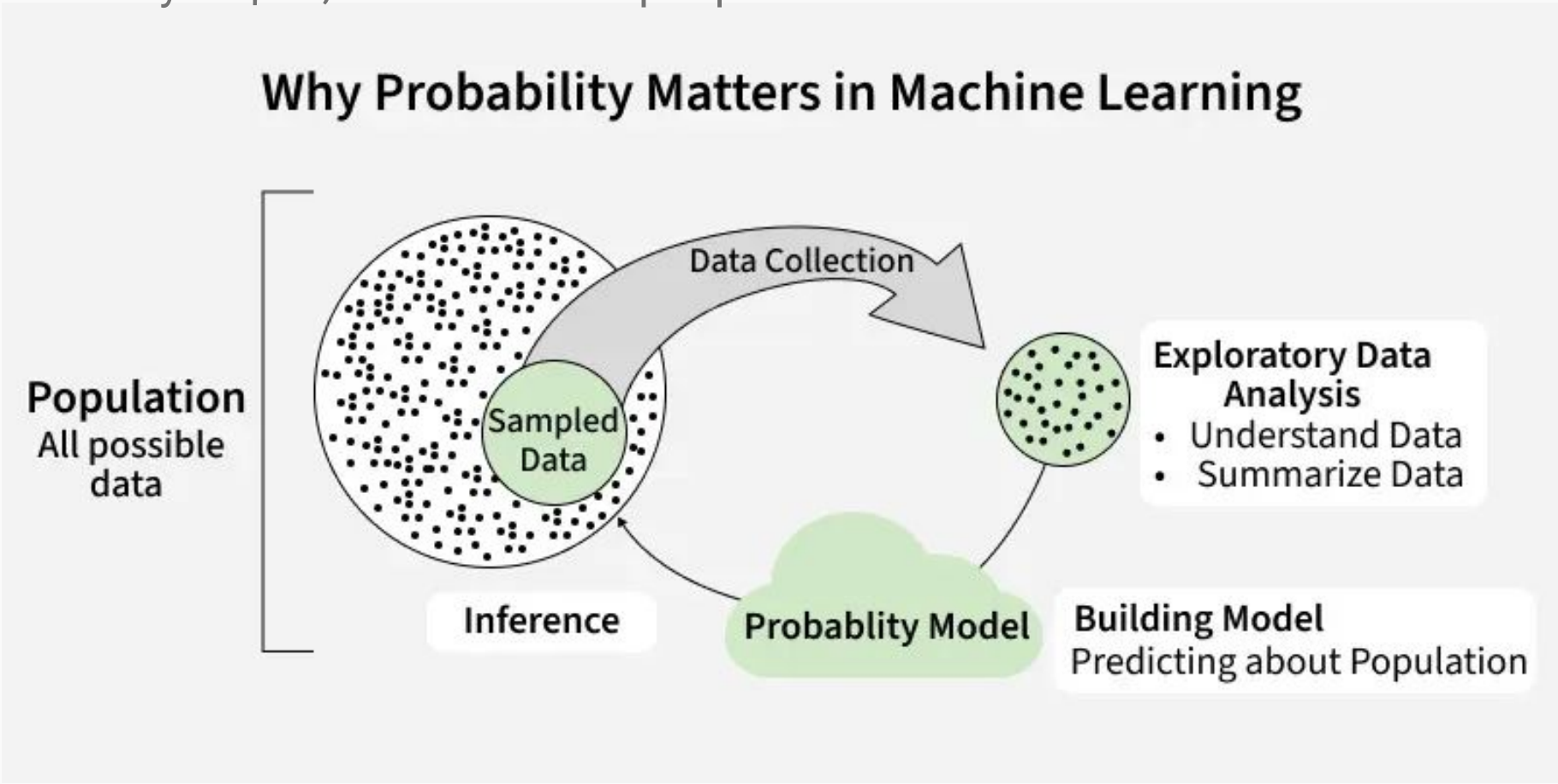
- Нам нужна *мера неопределённости*, а что может быть более естественной такой мерой, чем вероятность?..

- Поэтому центральную роль в машинном обучении играет теория вероятностей



Почему теория вероятностей

- Что такое вообще машинное обучение? Это неполная индукция: мы хотим обобщить закономерности с выборки на всю популяцию; типичный теорвер!

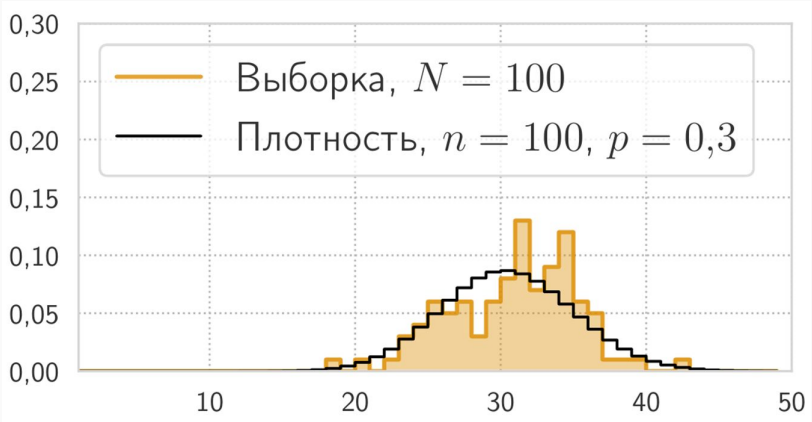


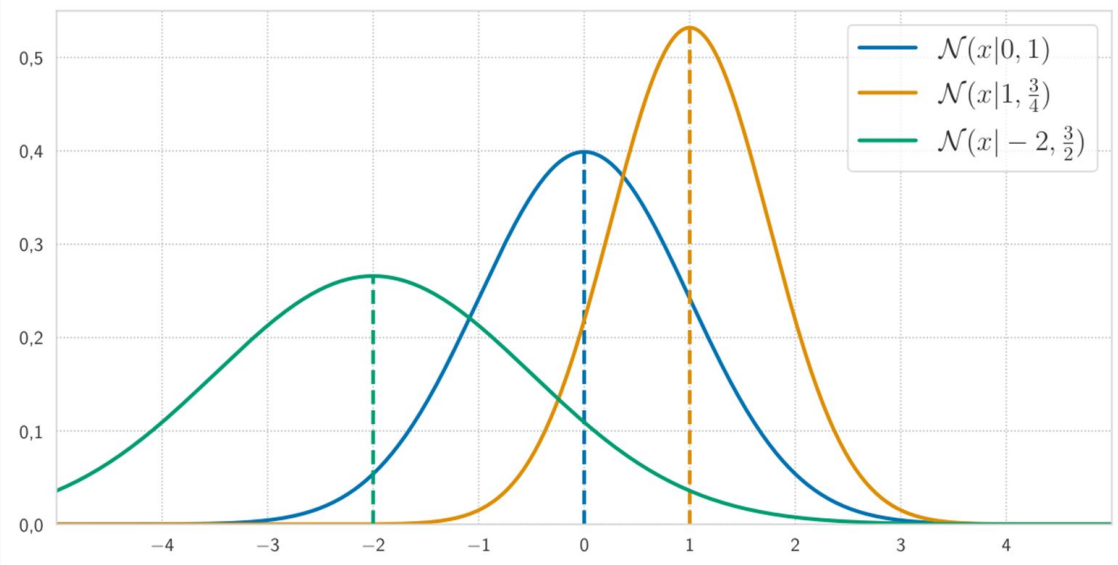
Почему теория вероятностей

- Почти любая задача машинного обучения сводится к тому, чтобы делать выводы в присутствии шума и неполной информации
- Теория вероятностей даёт единственный внутренне непротиворечивый «калькулятор» для работы с неопределённостями (теорема Кокса)
- Обучение — это статистический вывод: мы хотим сделать выводы о параметрах модели на основе данных, исходя из каких-то предположений
- Поэтому всё машинное обучение — это фактически раздел прикладной теории вероятностей

Основы теории вероятностей

- Нам не понадобятся математические определения сигма-алгебры, вероятностной меры, борелевских множеств и т.п.
- Достаточно понимать, что бывают дискретные и непрерывные случайные величины (сумма вероятностей или интеграл плотности равен единице)





Основы теории вероятностей

- *Совместная вероятность* — вероятность одновременного наступления двух событий, p(*x*,*y*); маргинализация:

$$p(x) = \sum_y p(x, y).$$

- *Условная вероятность* — вероятность наступления одного события, если известно, что произошло другое, p(*x*|*y*):

$$p(x, y) = p(x \mid y)p(y) = p(y \mid x)p(x).$$

- *Теорема Байеса* — из предыдущей формулы:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y'} p(x|y')p(y')}.$$

- *Независимость*: *x* и *y* независимы, если

$$p(x, y) = p(x)p(y).$$

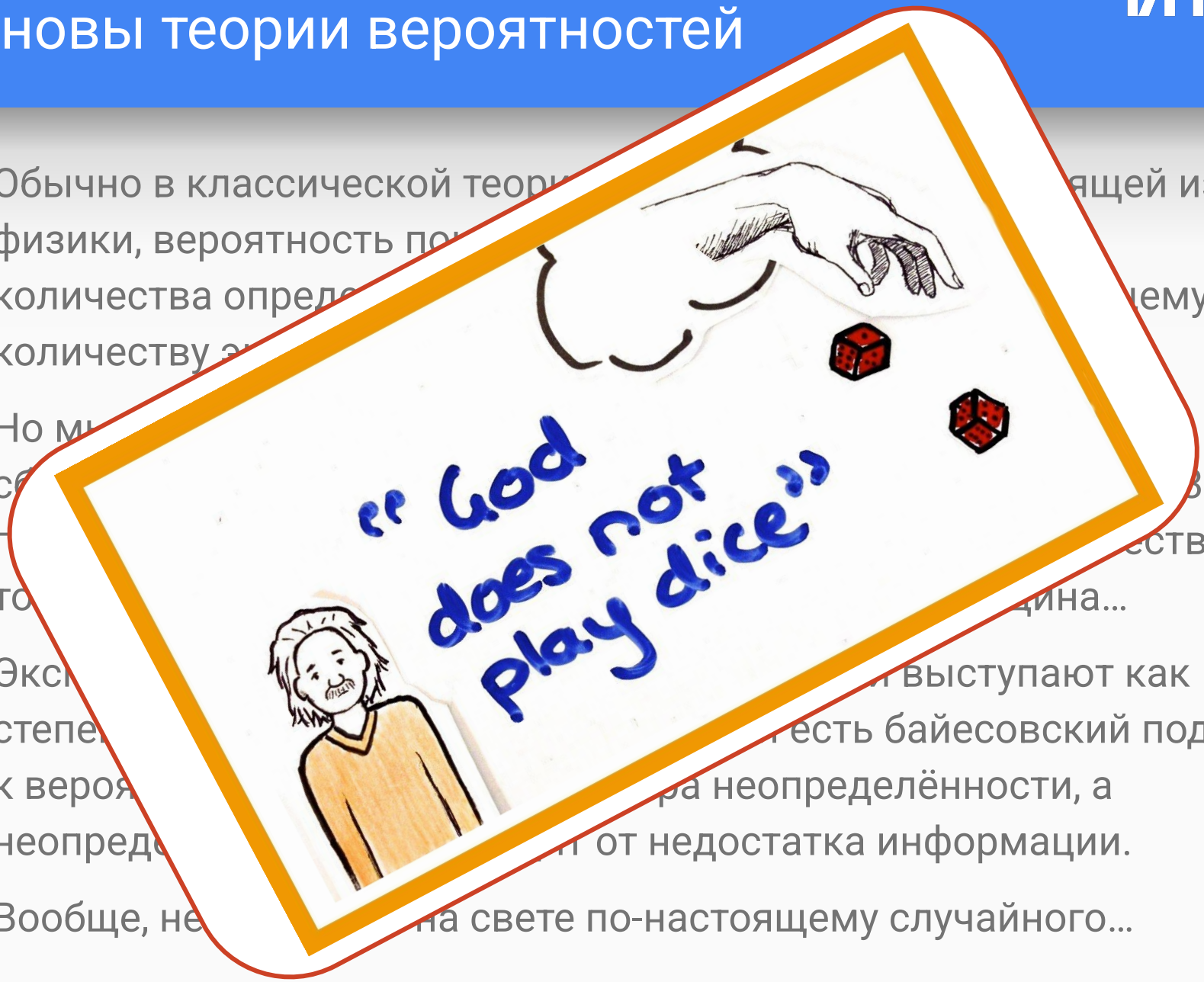
Основы теории вероятностей

ИТМО

- Обычно в классической теории вероятностей, происходящей из физики, вероятность понимается как предел отношения количества определённого результата эксперимента к общему количеству экспериментов.
- Но мы бы хотели рассуждать и о том, насколько вероятно, что сборная России победит на чемпионате мира по футболу в 2030 году, или что искусственный интеллект поработит человечество к тому же 2030 году, или что «Одиссею» написала женщина…
- Эксперимент здесь ровно один, а вероятности выступают как степени доверия (degrees of belief). Это и есть байесовский подход к вероятностям: вероятность – мера неопределённости, а неопределённость происходит от недостатка информации.
- Вообще, не так много на свете по-настоящему случайного…

Основы теории вероятностей

- Обычно в классической теори… …ящей из физики, вероятность по… количества опреде… …нему количеству э…
- Но м… с… …30 г… …ество к то… …дина…
- Экс… …выступают как степе… …есть байесовский подход к вероя… …ра неопределённости, а неопреде… …от недостатка информации.
- Вообще, не… …на свете по-настоящему случайного…



Основы теории вероятностей

- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-нибудь болезнь имеет вероятность успеха 95% (т.е. 5% — вероятность как ложноположительного, так и ложноотрицательного срабатывания).
- Всего болезнь имеется у 1% респондентов (отложим на время то, что они разного возраста и профессий).
- Пусть некий человек получил позитивный результат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?



Основы теории вероятностей

- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-н… болезнь имеет вероятн… 95% (т.е. 5% — веро… ложноположит… ложноотр… сраба…
- …
- Пуст… …к получил позит… …зультат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?



Ответ: 16%.

Основы теории вероятностей

- Можно подсчитать формально по теореме Байеса. Обозначим через t результат теста, через d — наличие болезни.

$$p(t=1) = p(t=1|d=1)p(d=1) + p(t=1|d=0)p(d=0)$$

$$p(d=1|t=1) =$$

$$= \frac{p(t=1|d=1)p(d=1)}{p(t=1|d=1)p(d=1) + p(t=1|d=0)p(d=0)} =$$

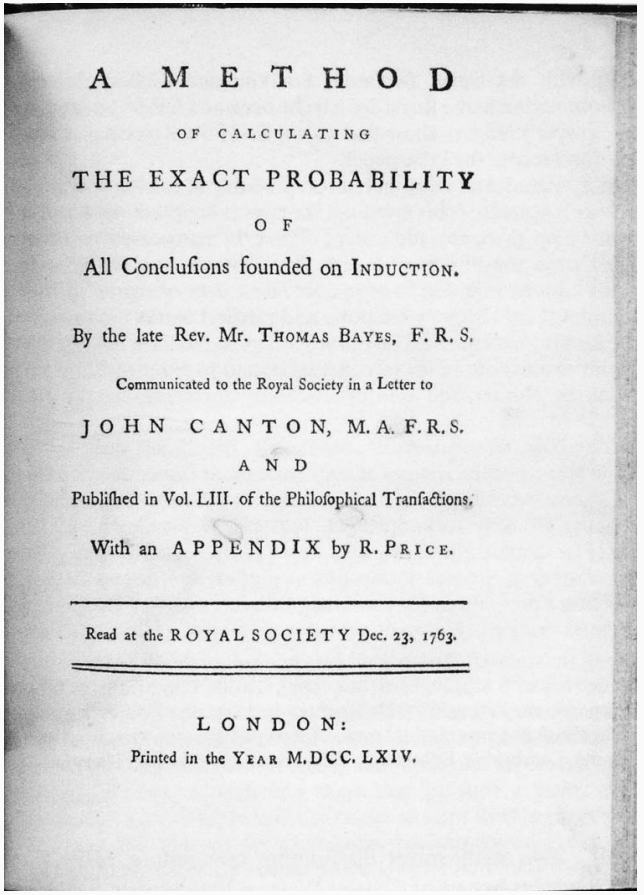
$$= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = 0.16.$$

- А можно и интуитивно это объяснить: есть 5% «на то», что ошибся тест, и 1% априорной вероятности «на то», что человек действительно болен, значит, должны получаться шансы примерно 1:5, то есть вероятность около 1/6 (примерно, потому что это рассуждение для простоты обнуляет вероятность ошибки в другую сторону).

Основы теории вероятностей

- Вот такие задачи составляют суть *вероятностного вывода* (probabilistic inference). Поскольку они обычно основаны на теореме Байеса, вывод часто называют *байесовским* (Bayesian inference)
- Но не только поэтому





A METHOD

OF CALCULATING

THE EXACT PROBABILITY

OF

All Conclusions founded on INDUCTION.

By the late Rev. Mr. THOMAS BAYES, F. R. S.

Communicated to the Royal Society in a Letter to

JOHN CANTON, M. A. F. R. S.

AND

Published in Vol. LIII. of the Philosophical Transactions,

With an APPENDIX by R. PRICE.

Read at the ROYAL SOCIETY Dec. 23, 1763.

LONDON:

Printed in the YEAR M.DCC.LXIV.

Простые примеры вероятностной интуиции



Основы теории вероятностей

- У моего знакомого два ребёнка. Будем предполагать, что пол ребёнка выбирается независимо и равновероятно, с вероятностью ½. Две постановки вопроса:
  - я спросил, есть ли у него мальчики, и он ответил «да»; какова вероятность того, что один из детей – девочка?
  - я встретил одного из его детей, и это мальчик; какова вероятность того, что второй ребёнок – девочка?



Парадокс Монти Холла

- Monty Hall, *Let's Make a Deal!*
- Нужно ли менять своё решение?
- Это, конечно, никакой не парадокс, ответ довольно простой, но для многих он оказался контринтуитивным
- Marilyn vos Savant, *Ask Marilyn* (1990)







Ошибка прокурора (prosecutor’s fallacy)

- Произошло убийство. На месте убийства найдена кровь, которая явно принадлежит убийце. Кровь принадлежит редкой группе, которая присутствует у 1% населения, в том числе у подсудимого.
- Прокурор говорит: «Шанс, что у подсудимого была бы именно такая группа крови, если бы он был невиновен – всего 1%; значит, с вероятностью 99% он виновен».
- В чём не прав прокурор?



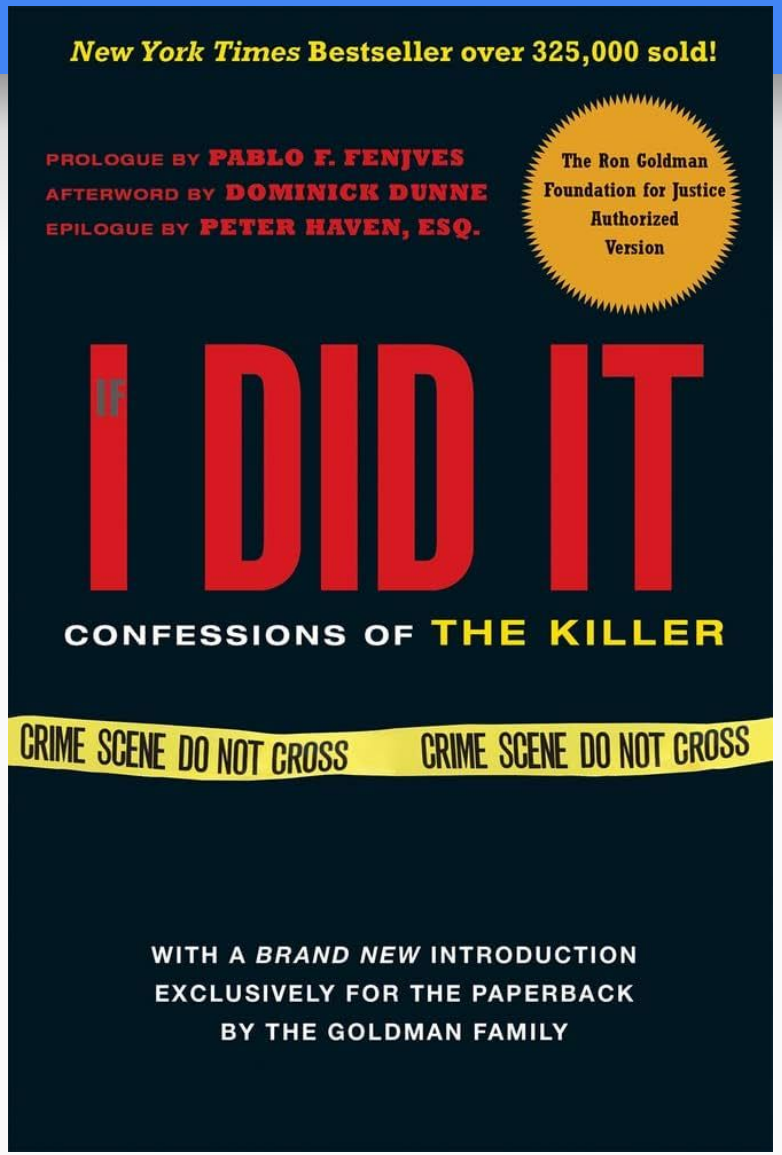
Ошибка адвоката (defense attorney’s fallacy)

- Произошло убийство. На месте убийства найдена кровь, которая явно принадлежит убийце. Кровь принадлежит редкой группе, которая присутствует у 1% населения, в том числе у подсудимого.
- Адвокат говорит: «В городе живёт миллион человек, то есть у 10000 из них такая группа крови. Значит, всё, что говорит нам эта кровь – это что подсудимый совершил убийство с вероятностью 0.01%; никакое это не доказательство».
- В чём не прав адвокат?



Реальные случаи

- Дело О Джей Симпсона
- Прокурор указал, что O.J. Simpson уже бил жену в прошлом
- Адвокат ответил: «Убивают только одну из 2500 женщин, подвергавшихся семейному насилию, так что это вообще нерелевантно»
- Суд согласился с адвокатом
- Верно ли это рассуждение?



Реальные случаи

- Адвокаты правильно оценили вероятность

$$p(\text{муж убил жену} \mid \text{муж бил жену в прошлом}).$$

- Но ведь убийство уже произошло! Поэтому нас интересует не эта статистика, а вероятность

$$p(\text{муж убил жену} \mid [\text{муж бил жену}] \text{ И } [\text{кто—то убил жену}]).$$

- И здесь по той же статистике выходит… около 90%!



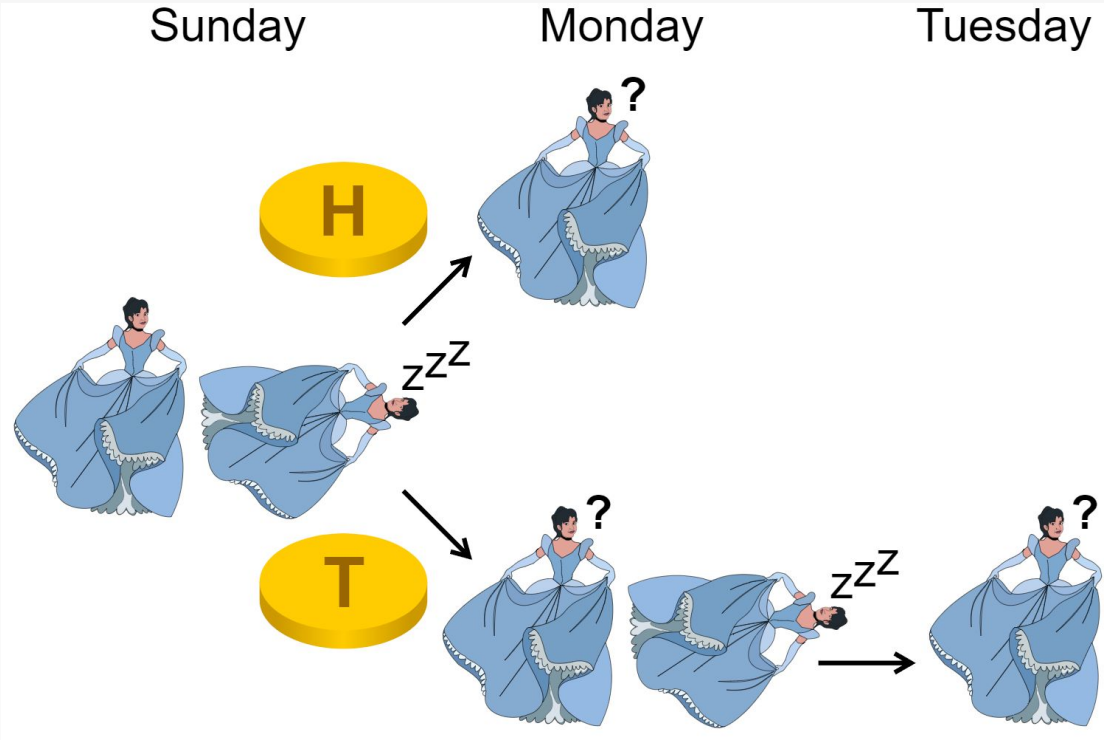
Реальные случаи

- У Салли Кларк погибли два младенца
- Прокурор указал, что вероятность двух случаев SIDS в одной семье, которую он получил из статистики одиночных случаев, — около 1 из 73 миллионов
- В чём он не прав?



Парадокс Спящей красавицы

- А вот парадокс, до сих пор вызывающий споры (Zuboff, 1990; Elga, 2000).
- В воскресенье Спящую красавицу усыпляют и бросают монетку:
  - если выпадает орёл, её будят в понедельник и опять усыпляют до среды;
  - если решка, то будят в понедельник, делают укол, от которого она забывает о пробуждении, и будят ещё раз во вторник, а потом так же в среду.



Парадокс Спящей красавицы

- Спящая красавица знает всю процедуру эксперимента, но не может различить разные пробуждения и отличить понедельник от вторника, когда просыпается
- Вы Спящая красавица; вас разбудили; какова вероятность того, что монета упала орлом?



Теорема Байеса в машинном обучении

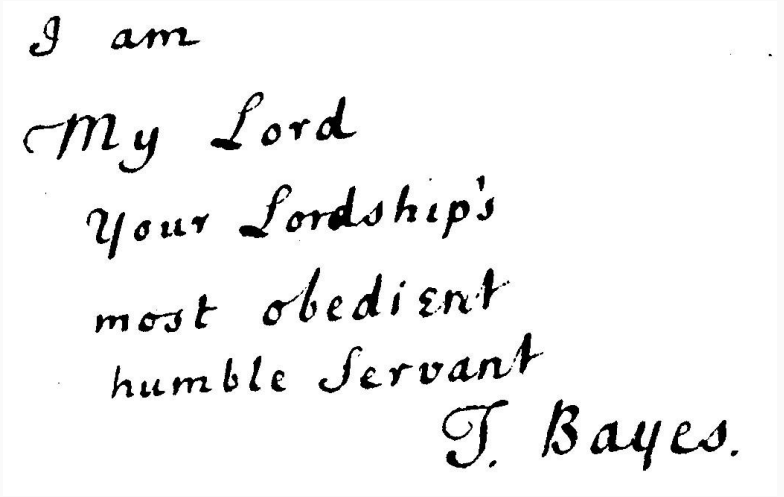


Теорема Байеса в ML

- Давайте обсудим компоненты формулы Байеса в машинном обучении:
- Параметры θ, данные D:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$





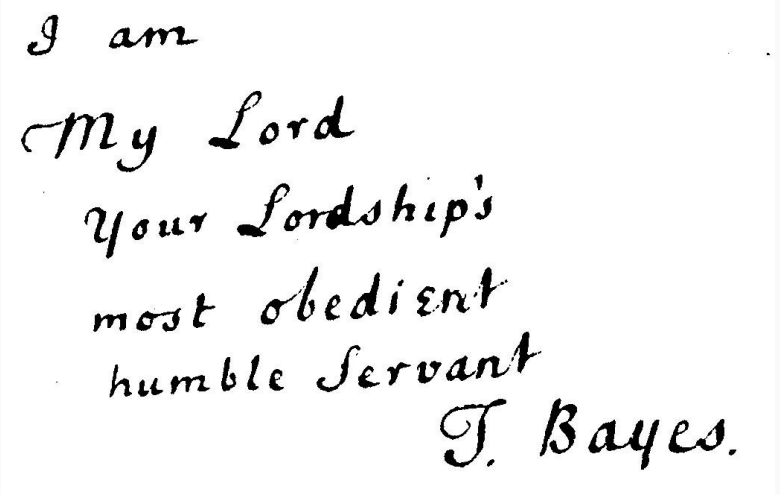
Теорема Байеса в ML

- Давайте обсудим компоненты формулы Байеса в машинном обучении:
- Параметры θ, данные D:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$



- **Правдоподобие** p(D|θ)



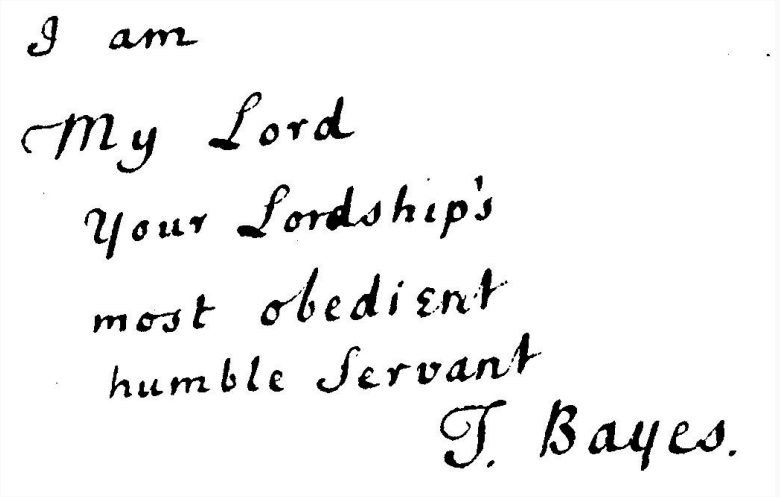
Теорема Байеса в ML

- Давайте обсудим компоненты формулы Байеса в машинном обучении:
- Параметры θ, данные D:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$



- **Правдоподобие** $p(D|\theta)$
- **Априорное распределение** $p(\theta)$



Теорема Байеса в ML

- Давайте обсудим компоненты формулы Байеса в машинном обучении:
- Параметры θ, данные D:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$



- **Правдоподобие** p(D|θ)
- **Априорное распределение** p(θ)
- **Апостериорное распределение** p(θ|D)

Задачи байесовского вывода

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

- Отсюда следуют основные задачи байесовского вывода
- Первая задача – найти *гипотезу максимального правдоподобия* (maximum likelihood, ML):

$$\theta_{ML} = \operatorname{argmax}_{\theta} p(D|\theta)$$

Задачи байесовского вывода

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

- Отсюда следуют основные задачи байесовского вывода
- Первая задача – найти *гипотезу максимального правдоподобия* (maximum likelihood, ML):

$$\theta_{ML} = \arg\max_{\theta} p(D|\theta)$$

- Вторая – найти апостериорное распределение (posterior)

$$p(\theta|D) \propto p(\theta)p(D|\theta)$$

и, если будет нужно, максимальную апостериорную гипотезу (maximum a posteriori, MAP)

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|D) = \arg\max_{\theta} p(\theta)p(D|\theta)$$

Задачи байесовского вывода

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

- Отсюда следуют основные задачи байесовского вывода
- Первая задача – найти *гипотезу максимального правдоподобия* (maximum likelihood, ML):

$$\theta_{ML} = \arg\max_{\theta} p(D|\theta)$$

- Вторая – найти апостериорное распределение (posterior)

$$p(\theta|D) \propto p(\theta)p(D|\theta)$$

  и, если будет нужно, максимальную апостериорную гипотезу (maximum a posteriori, MAP)

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|D) = \arg\max_{\theta} p(\theta)p(D|\theta)$$

- Самая сложная (и интересная) третья задача – найти *предсказательное распределение* (predictive distribution)

$$p(x|D) \propto \int_{\theta\in\Theta} p(x|\theta)p(\theta|D)\mathrm{d}\theta$$

Задачи байесовского вывода

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

- Отсюда следуют основные задачи байесовского в
- Первая задача – найти *гипотезу максимал* (maximum likelihood, ML):

$$\theta_{ML} = \arg$$

- Вторая – найти апос sterior)

и, ес ую апостериорную гипотезу (

$$\max_\theta p(\theta|D) = \arg\max_\theta p(\theta)p(D|\theta)$$

- Сама я (и интересная) третья задача – найти *предск ательное распределение* (predictive distribution)

$$p(x|D) \propto \int_{\theta\in\Theta} p(x|\theta)p(\theta|D)\mathrm{d}\theta$$

Давайте обсудим их на конкретном примере.

Байесовский вывод для монетки



Задачи байесовского вывода

- Рассмотрим монетку, которая подброшена N раз, и результаты этих бросков известны, т.е. данные D – это последовательность исходов: орёл, решка, решка, орёл, орёл...
- Что такое здесь параметры θ? Сколько их?



$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

Задачи байесовского вывода

- Задача машинного обучения (байесовского вывода) здесь в том, чтобы определить «нечестность» монетки, то есть её единственный параметр θ (вероятность выпадения орла), и предсказать, с какой вероятностью она выпадет орлом в следующий раз.
- Начнём с *правдоподобия*: какое будет правдоподобие $p(D|\theta)$ для датасета D из n орлов и m решек?



$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

Задачи байесовского вывода

- Задача машинного обучения (байесовского вывода) здесь в том, чтобы определить «нечестность» монетки, то есть её единственный параметр θ (вероятность выпадения орла), и предсказать, с какой вероятностью она выпадет орлом в следующий раз.
- Начнём с *правдоподобия*: какое будет правдоподобие p(D|θ) для датасета D из n орлов и m реше

$$p(D|\theta) = \theta^n(1-\theta)^m$$

- Где оно максимизируется?



$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

Задачи байесовского вывода

- Мы вроде бы получили решение: $\theta_{ML} = \frac{n}{n+m}$
- Но, во-первых, у правдоподобия для m = 0 нет максимума, оно идёт к бесконечности.
- Во-вторых, даже если предположить, что мы как-то ввели ограничение, утверждающее, что θ должно быть вероятностью, получается, что после первого же броска мы предсказываем, что теперь монетка всегда будет выпадать той же стороной…
- Звучит странно… а что не так? Чего не хватает?

Задачи байесовского вывода

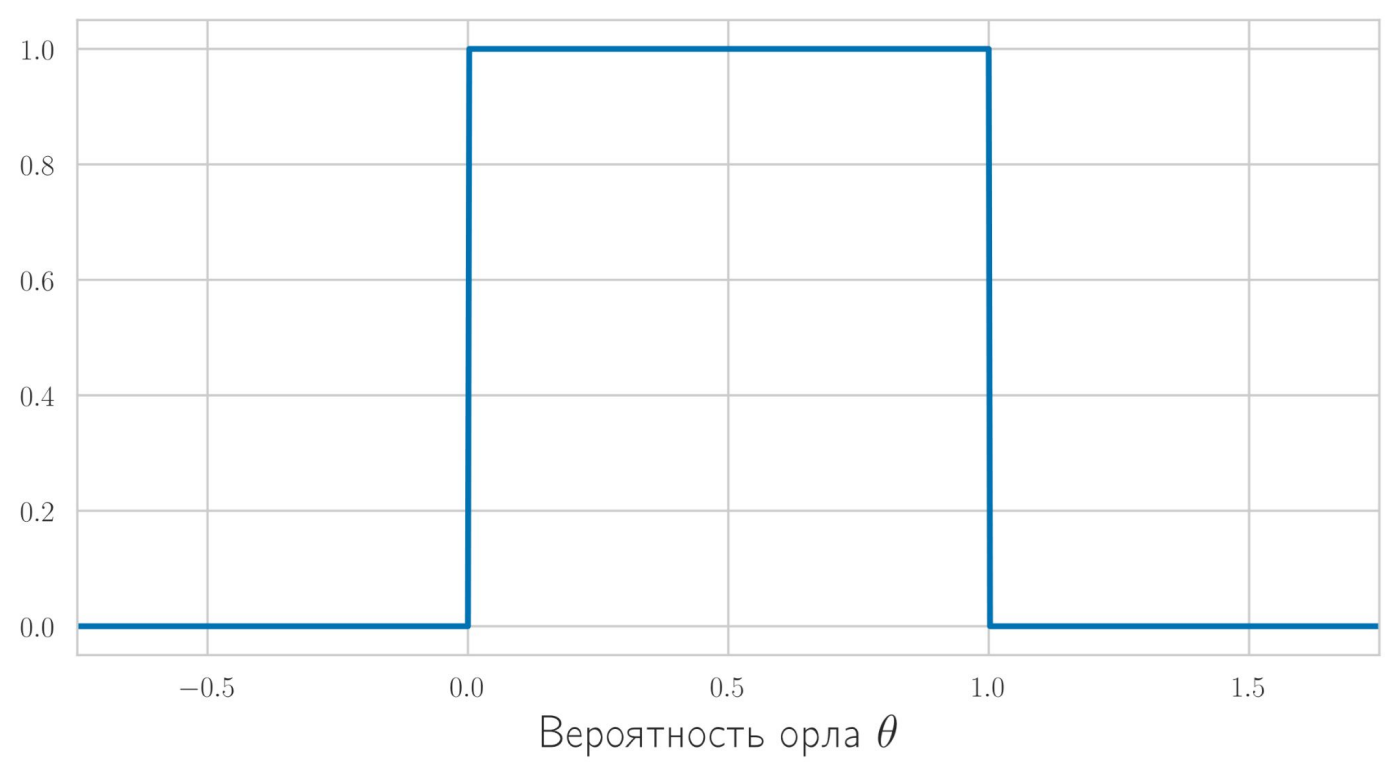
- Мы вроде бы получили решение: $\theta_{ML} = \frac{n}{n+m}$
- Но, во-первых, у правдоподобия для m = 0 нет максимума, оно идёт к бесконечности.
- Во-вторых, даже если предположить, что мы как-то ввели ограничение, утверждающее, что θ должно быть вероятностью, получается, что после первого же броска мы предсказываем, что теперь монетка всегда будет выпадать той же стороной…
- Звучит странно… а что не так? Чего не хватает?
- Не хватает нашей интуиции о том, какие на свете бывают монетки. Что выражает эту интуицию в байесовском выводе?

Задачи байесовского вывода

- Мы вроде бы получили решение: $\theta_{ML} = \frac{n}{n+m}$
- Но, во-первых, у правдоподобия для m = 0 нет максимума, оно идёт к бесконечности.
- Во-вторых, даже если предположить, что мы как-то ввели ограничение, утверждающее, что θ должно быть вероятностью, получается, что после первого же броска мы предсказываем, что теперь монетка всегда будет выпадать той же стороной…
- Звучит странно… а что не так? Чего не хватает?
- Не хватает нашей интуиции о том, какие на свете бывают монетки. Что выражает эту интуицию в байесовском выводе?
- Априорное распределение $p(\theta)$! Давайте его зададим.

Задачи байесовского вывода

- Начнём с равномерного априорного распределения:



- Это не совсем моя интуиция о монетках, но начнём отсюда...

Задачи байесовского вывода

- Теперь мы хотим найти

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|D) = \arg\max_{\theta} p(\theta)p(D|\theta)$$

- Но с равномерным априорным распределением это та же самая функция, оно добавляет только то самое ограничение θ от 0 до 1, а внутри этого правдоподобие не меняется
- То есть мы ничего не достигли? Нет, это только начало…

Задачи байесовского вывода

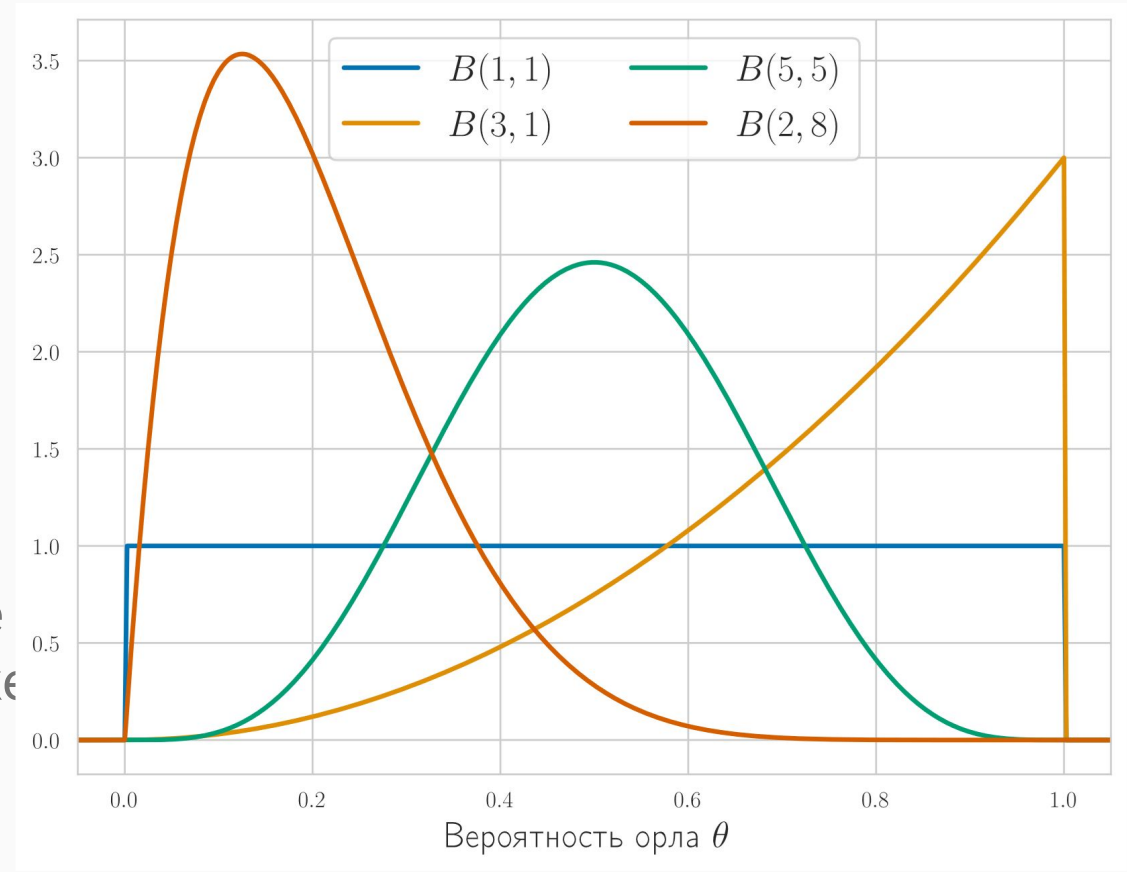
- Нормировочная константа для апостериорного распределения:

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x \qquad B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- *Бета-распределение* на [0, 1]:

$$\mathrm{Beta}(\theta|\alpha,\beta)=\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Давайте подсчитаем теперь с константой; от этого, правда, максимум всё равно не изменится, так зачем же всё это было нужно?



Задачи байесовского вывода

- Последняя, самая сложная задача байесовского вывода — найти *предсказательное распределение*:

$$p(\text{орёл}|D) = \int p(\text{орёл}|\theta)p(\theta|D)\mathrm{d}\theta$$

- Подсчитав его в данном случае, получим *правило Лапласа*:

$$p(\text{орёл}|D) = \frac{n+1}{n+m+2}$$

- Давайте это обсудим…



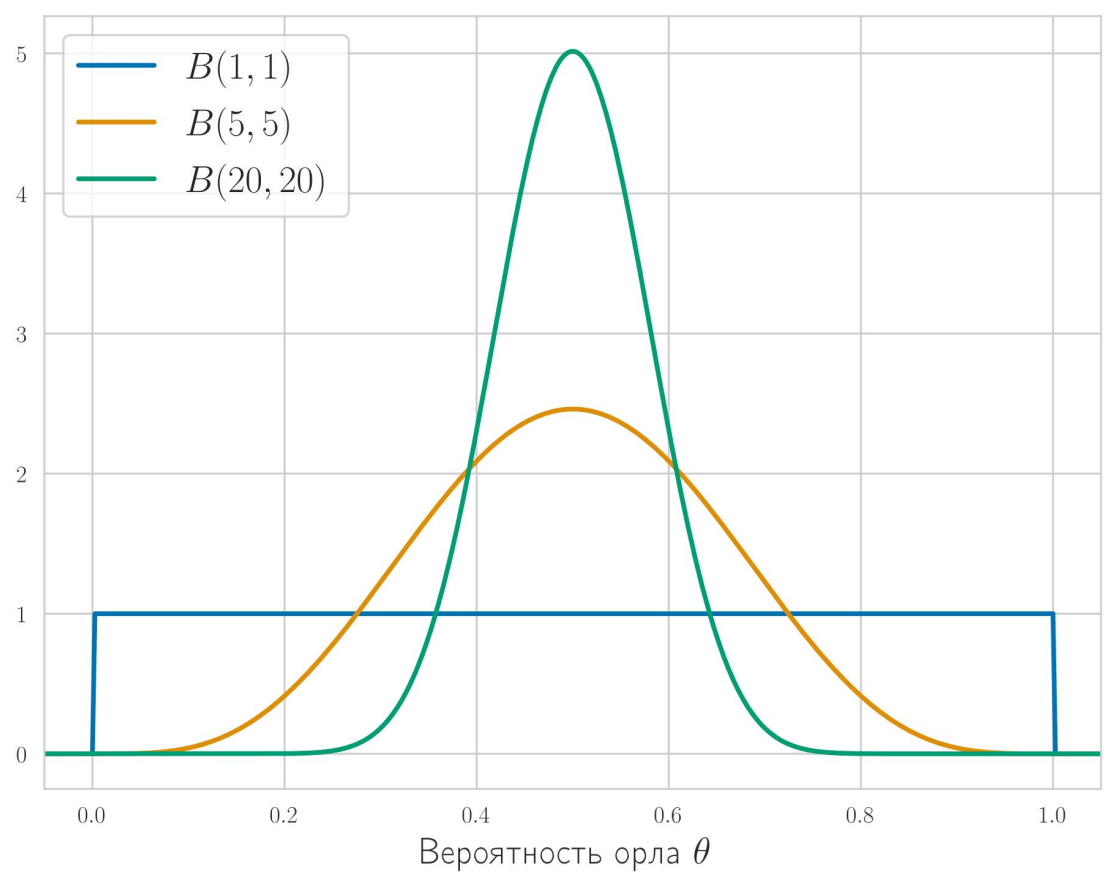
Сопряжённые априорные распределения



Бета-распределение как сопряжённое

- Мы уже выяснили, что *бета-распределение* похоже на правдоподобие монетки
- Более того, кажется, что для больших одинаковых параметров бета-распределение выражает нужную интуицию про настоящие монетки…

$$\text{Beta}(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$



Бета-распределение как сопряжённое

- Давайте использовать бета-распределение как априорное:

$$p(\theta) = \text{Beta}(\theta|\alpha,\beta)$$

- Тогда

$$p(\theta|D) \propto p(\theta)p(D|\theta) =$$

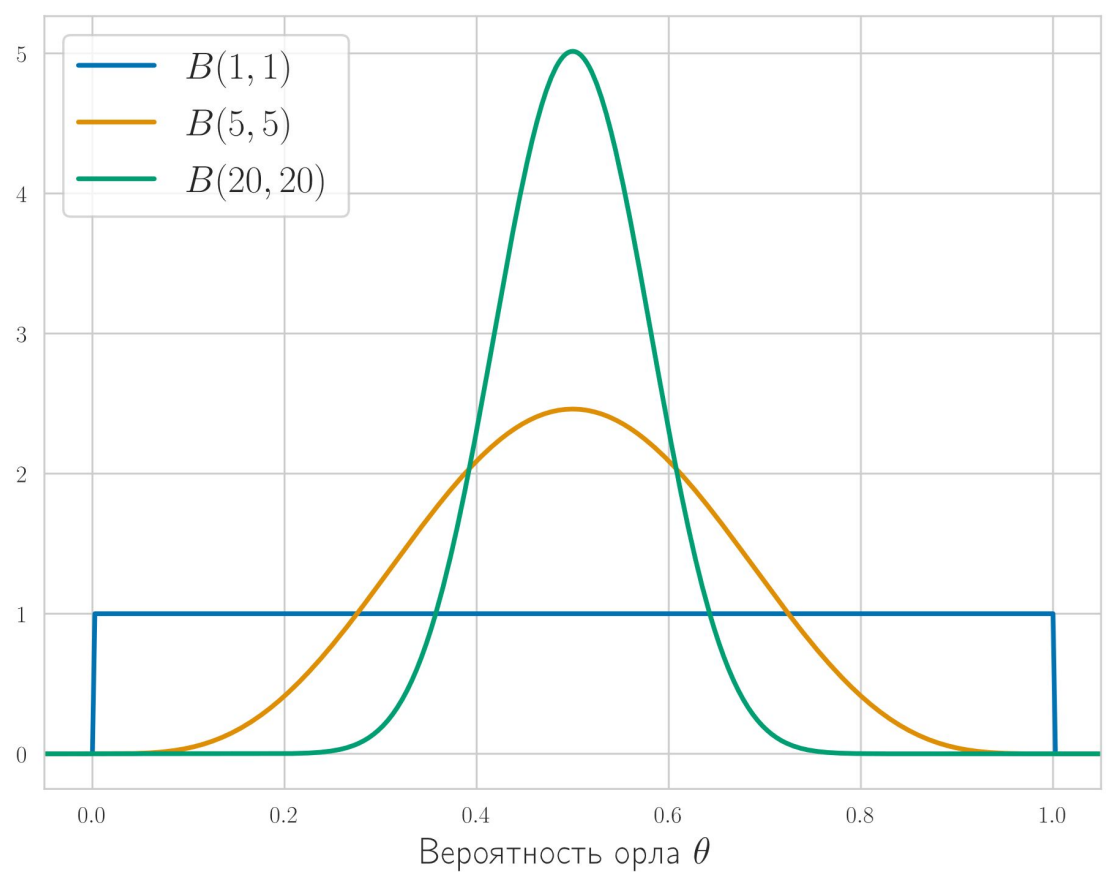
$$= \text{Beta}(\theta|\alpha,\beta)\theta^n(1-\theta)^m \propto$$

$$\propto \theta^{n+\alpha-1}(1-\theta)^{m+\beta-1},$$

и снова получается бета-распределение, теперь как апостериорное!

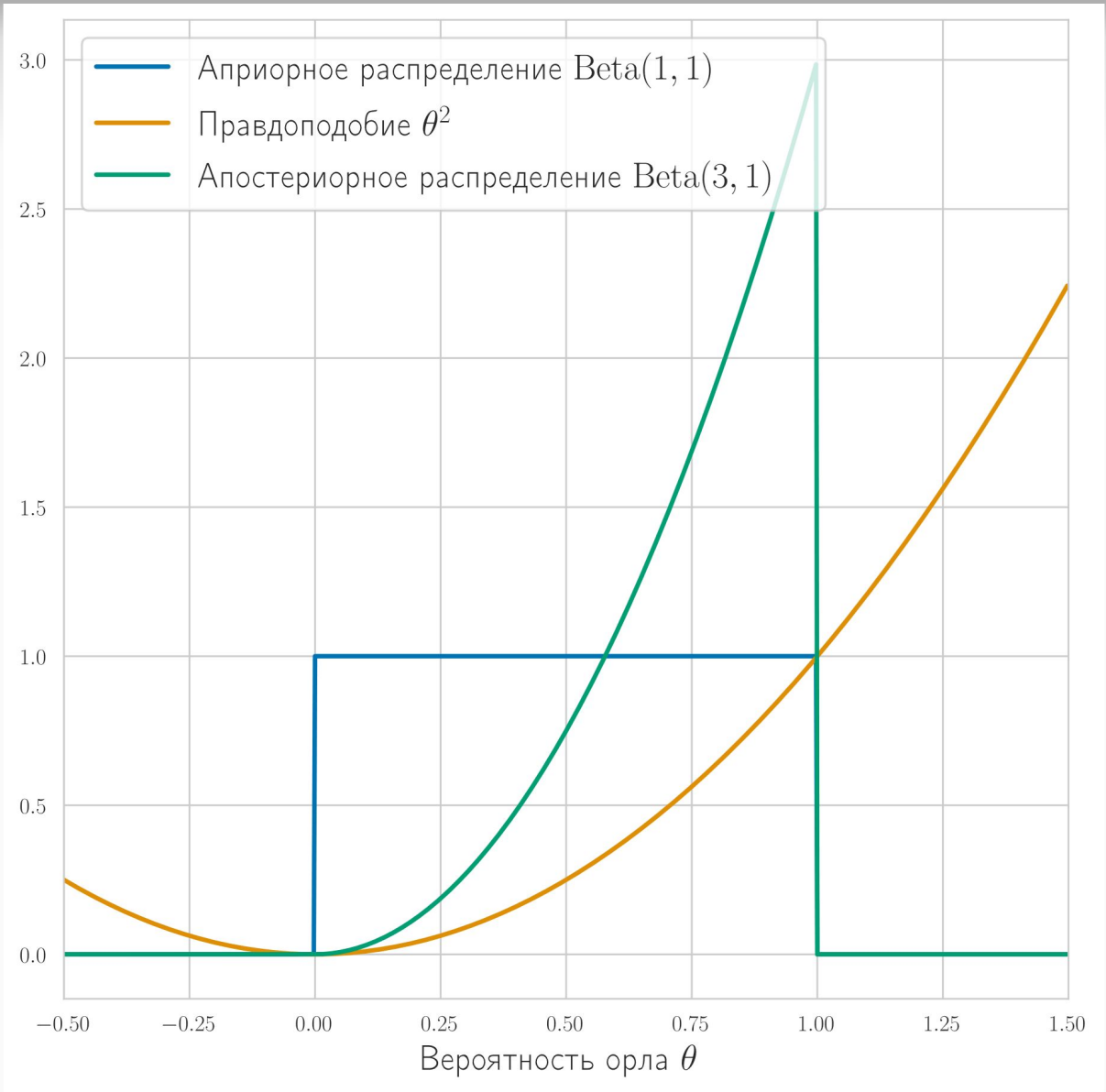
- Это пример *сопряжённых* априорных распределений (conjugate priors)

$$\text{Beta}(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$



Пример байесовского вывода

- Вот пример байесовского вывода как умножения одного распределения на другое
- Давайте посмотрим на примеры в коде и заодно обсудим один любопытный частный случай...



ИТМО



@SINECOR

Спасибо за внимание!



