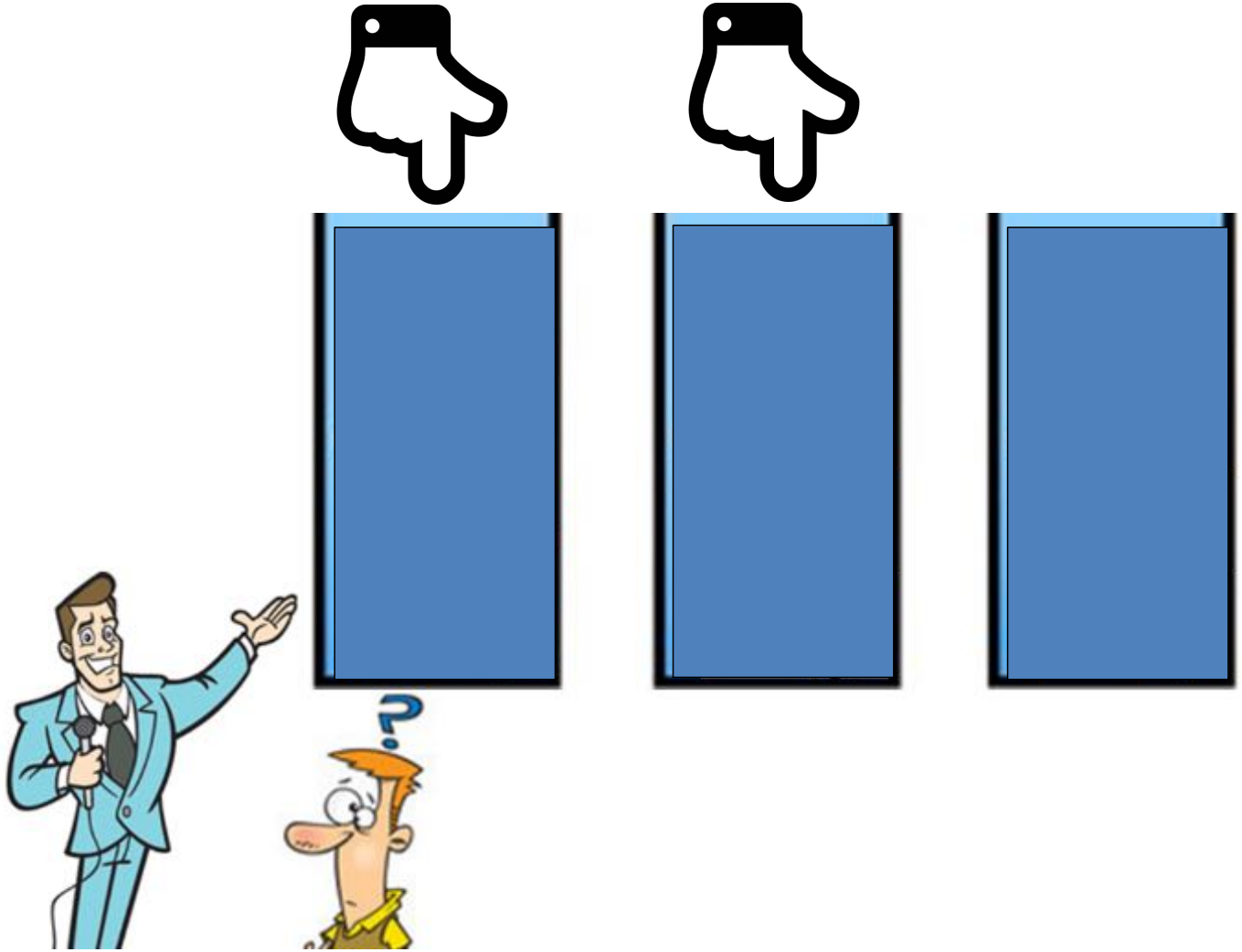
ВВЕДЕНИЕ В БИОИНФОРМАТИКУ
Лекция №2
Информация

Понятие информации. Способы измерения информации. ДНК как цифровой носитель информации. Системы счисления. Теория информации. Информационная энтропия. Сжатие информации. Теорема Котельникова. Теорема Шеннона-Хартли. Шифрование информации. Хранение информации. Источники больших данные в биомедицине. Проблемы передачи больших данных. Базы данных

Алексей Константинович Шайтан, к.ф.-м.н.

http://intbio.org
alex@intbio.org

Биологический факультет, МГУ имени М.В. Ломоносова, лекции для 2 курса



- В результате игрок получает **0.67 бита информации**

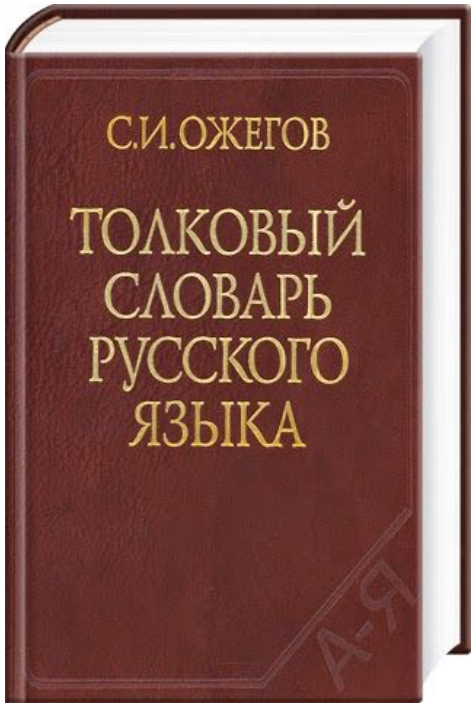
- Если бы ведущий открыл дверь в самом начале – только **0.58 бита информации**

---

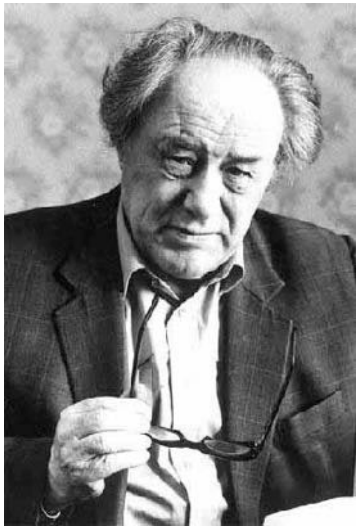
https://en.wikipedia.org/wiki/Monty_Hall_problem

Природа информации

Информация



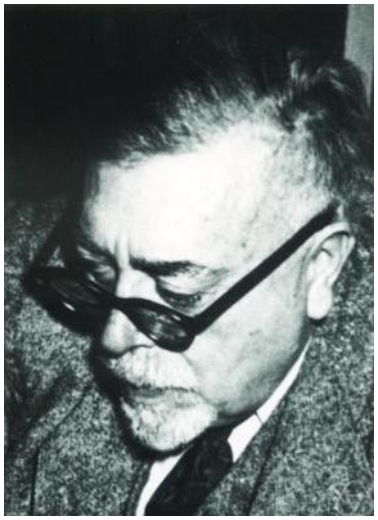
ИНФОРМАЦИЯ, 1. Сведения об окружающем мире и протекающих в нем процессах, воспринимаемые человеком или специальным устройством.



… универсального определения информации не только нет, но и быть не может из-за широты этого понятия.

Н.Н. Моисеев

Information is information, not matter or energy.



Norbert Wiener

http://www.aselibrary.ru/press_center/journal/irr/2007/number_3/number_3_6/number_3_6571/

Информация

iPhone 7 Plus Серебристый Ёмкость

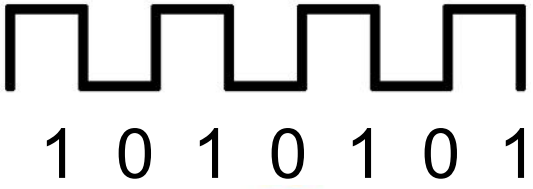
Теперь выберите ёмкость.

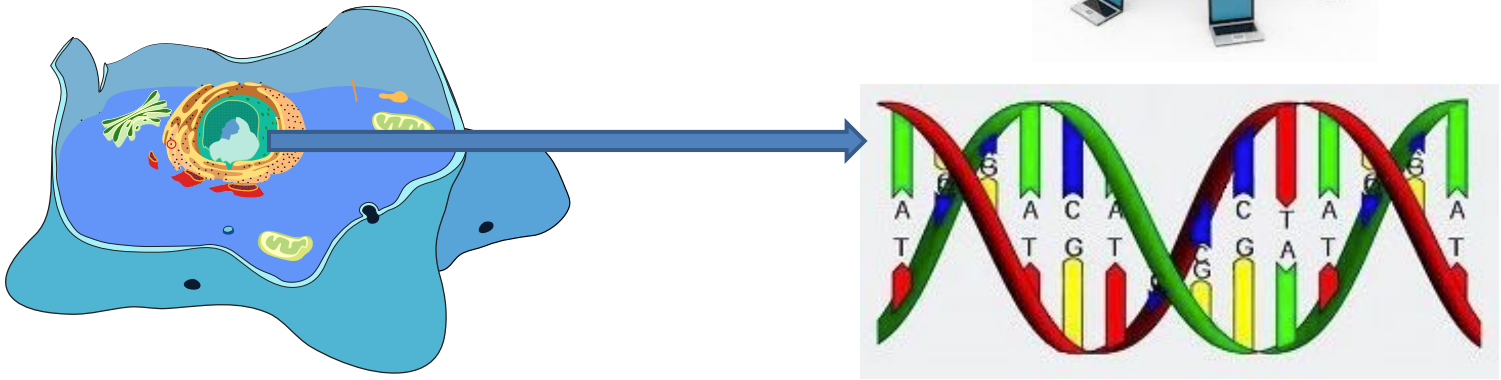
32 ГБ[1]	128 ГБ[1]
52 990.00 руб.	60 990.00 руб.
Доставка: на складе	Доставка: на складе

Живые системы и цифровые технологии

Аналоговые технологии

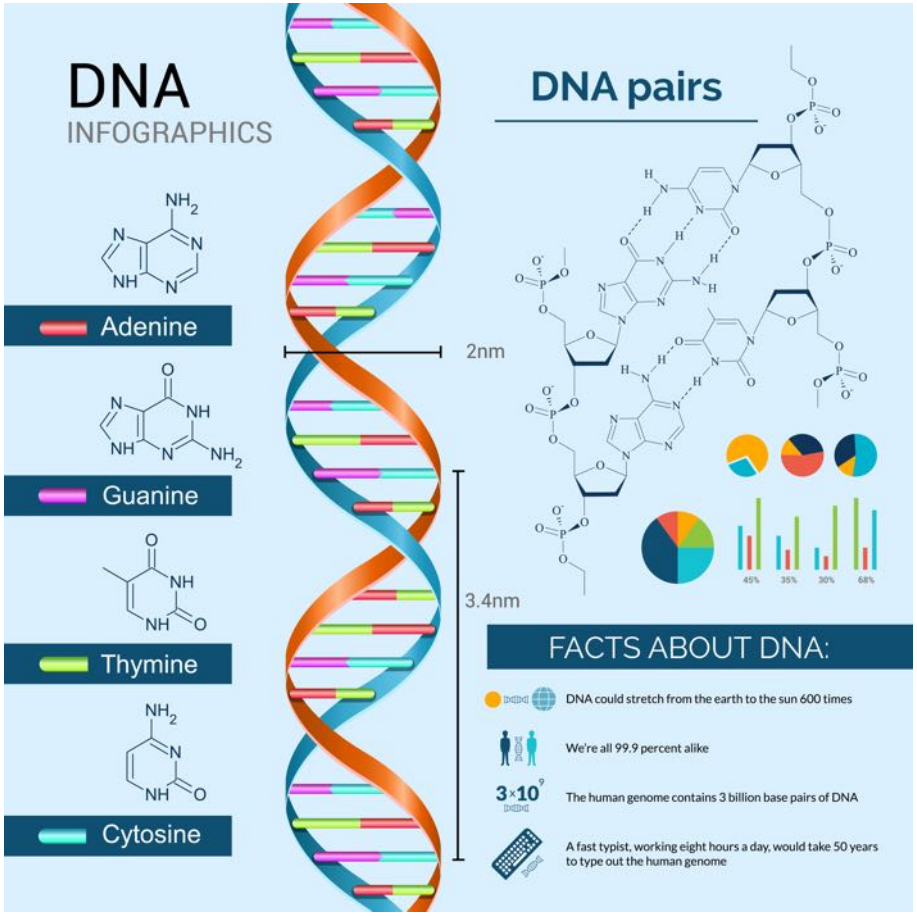
Цифровые технологии

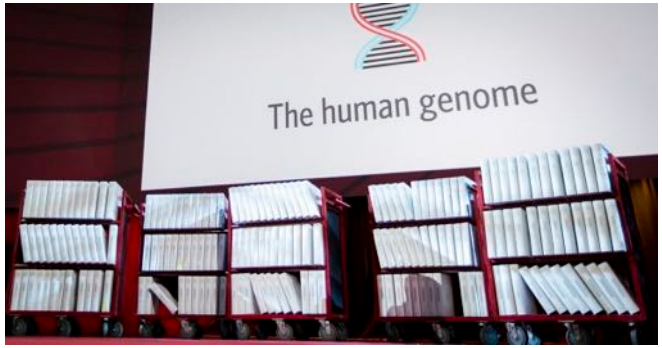


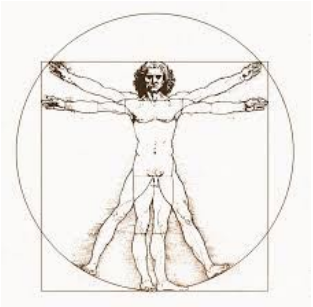


ДНК – цифровой код

Генетический код





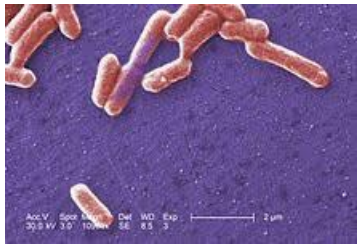


<u>Геном человека</u>

3 млрд нуклеотидов

262 тыс страниц

3 Гбайта

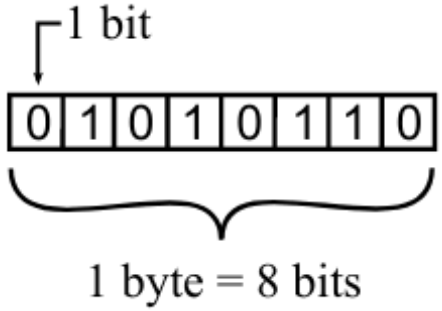


<u>Геном кишечной палочки</u>

4.6 млн нуклеотидов

400 страниц

4.6 Мбайт



A T G C

Двоичный код

Четвертичный код

Код может быть **позиционный** или **непозиционный**

Позиционный код активно используется в **системах счисления**

Соответствие цифр некоторых систем счисления				
Основание системы счисления	2	8	10	16
Зеленые ячейки — цифры системы счисления, желтые - числа.	0	0	0	0
	1	1	1	1
	10	2	2	2
	11	3	3	3
	100	4	4	4
	101	5	5	5
	110	6	6	6
	111	7	7	7
	1000	10	8	8
	1001	11	9	9
	1010	12	10	A
	1011	13	11	B
	1100	14	12	C
	1101	15	13	D
	1110	16	14	E
	1111	17	15	F

Целое число без знака x в b-ичной системе счисления представляется в виде конечной линейной комбинации степеней числа b[1]:

$x = \sum_{k=0}^{n-1} a_k b^k$, где a_k — это целые числа, называемые **цифрами**, удовлетворяющие неравенству $0 \leq a_k \leq b - 1$.

TIME

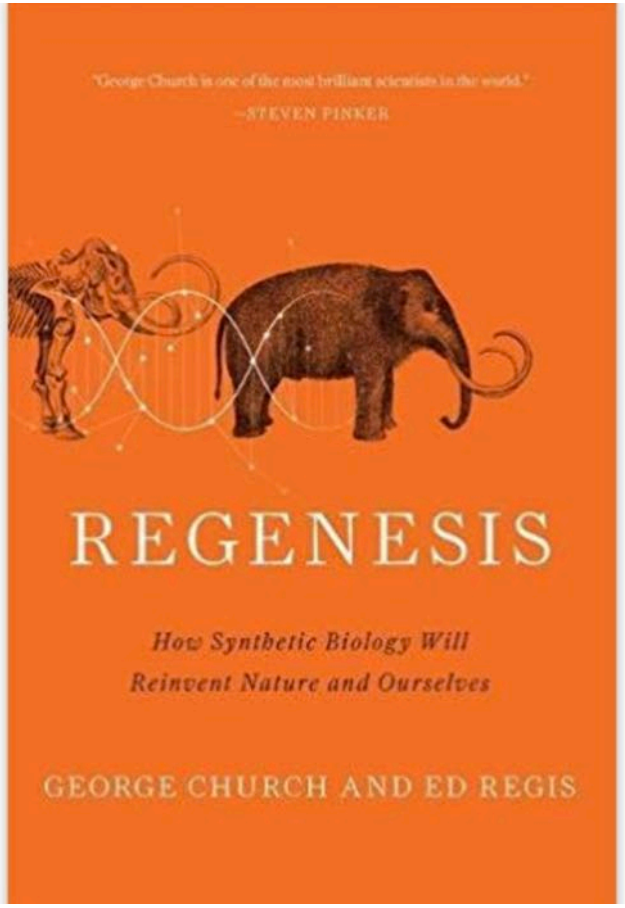
SCIENCE

The First Book To Be Encoded in DNA

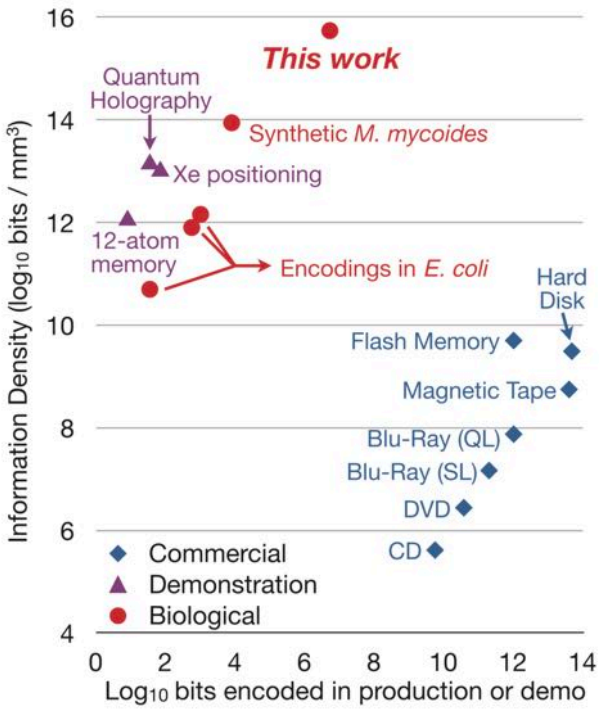
Two Harvard scientists have produced 70 billion copies of a book in DNA code --and it's smaller than the size of your thumbnail.

Lisa Poole / AP FILE

In his lab at the Harvard Medical School in Boston, George Church,

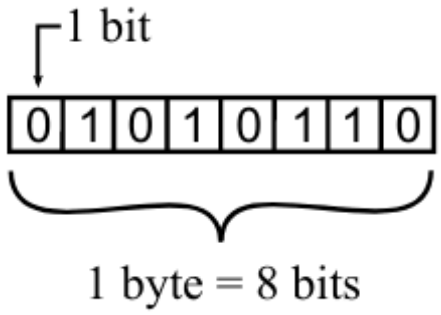


~1 Петабайт на мм^3



Church, G. M., Gao, Y., & Kosuri, S. (2012). *Science, 337(6102), 1628–1628.*

Измерение информации

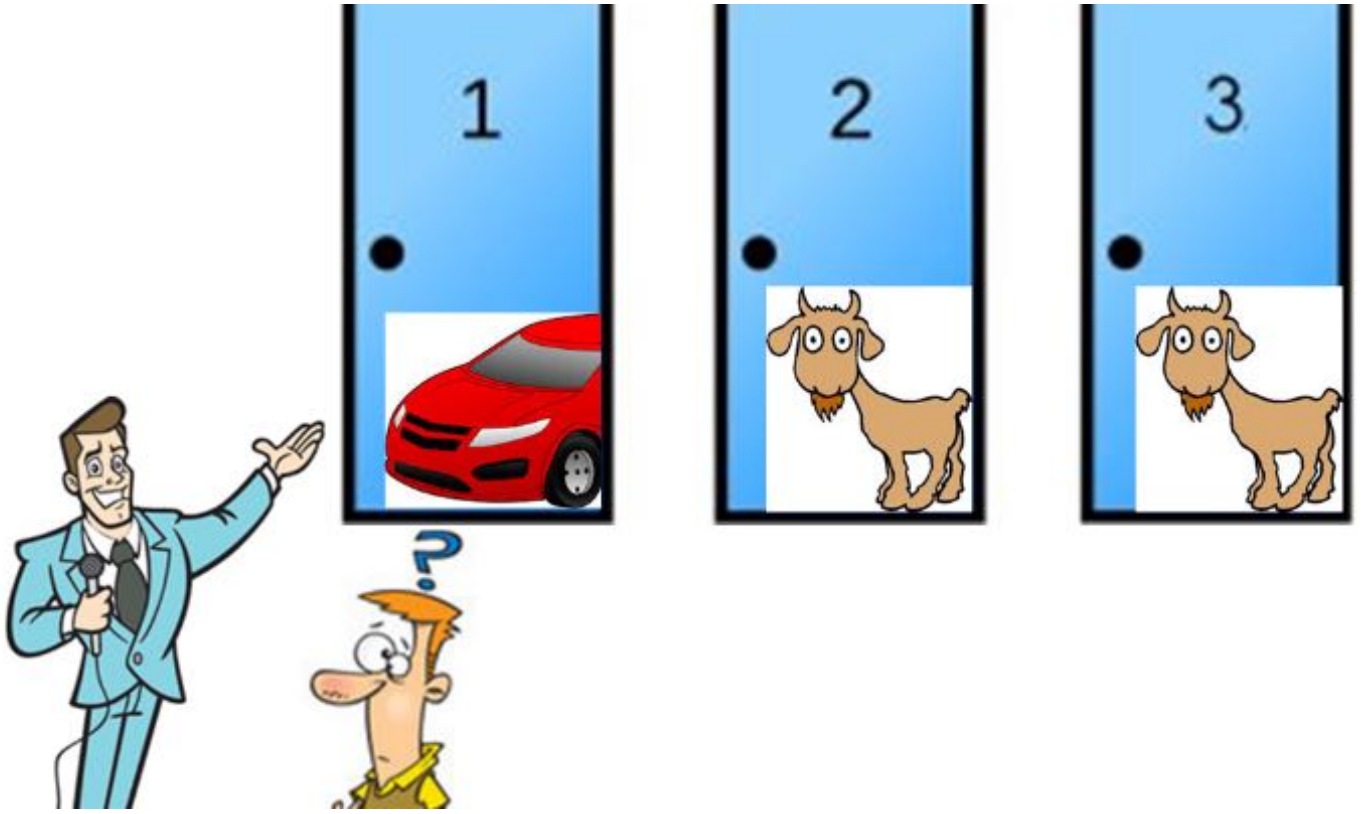


Измерения в байтах				
ГОСТ 8.417—2002			Приставки СИ	
Название	Обозначение	Степень	Название	Степень
байт	Б	10^0	-	10^0
килобайт	кбайт	10^3	кило-	10^3
мегабайт	Мбайт	10^6	мега-	10^6
гигабайт	Гбайт	10^9	гига-	10^9
терабайт	Тбайт	10^{12}	тера-	10^{12}
петабайт	Пбайт	10^{15}	пета-	10^{15}
эксабайт	Эбайт	10^{18}	экса-	10^{18}
зеттабайт	Збайт	10^{21}	зетта-	10^{21}
иоттабайт	Ибайт	10^{24}	иотта-	10^{24}

Decimal	Binary
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001

Вероятность того, что машина за этой дверью:

1/3 1/3 1/3

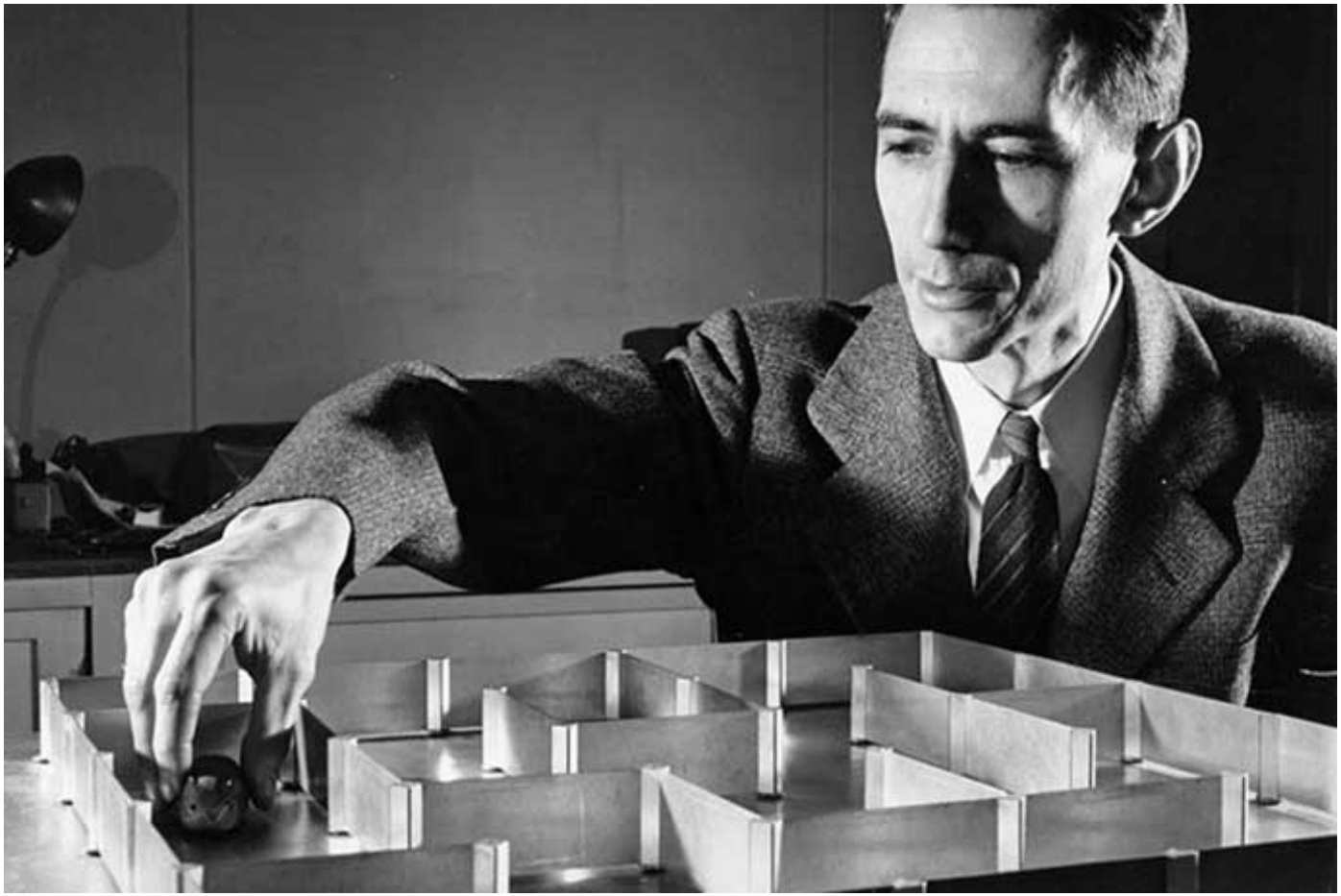


- Сколько информации нужно, чтобы закодировать положении машины?

1.5849625007211563... бит

Теория информации

“the father of information theory"



Claude Elwood Shannon
(April 30, 1916 – February 24, 2001)

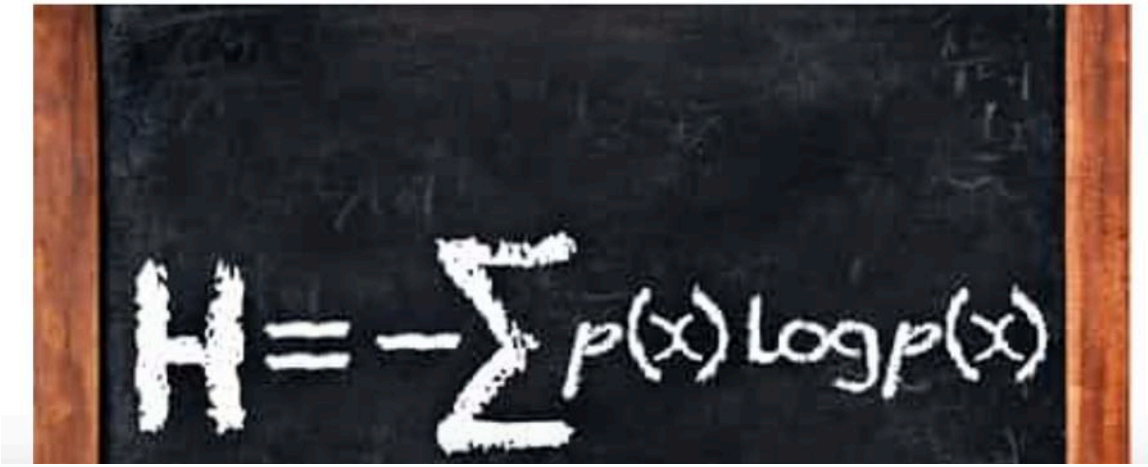
https://youtu.be/z2Whj_nL-x8

Теория информации

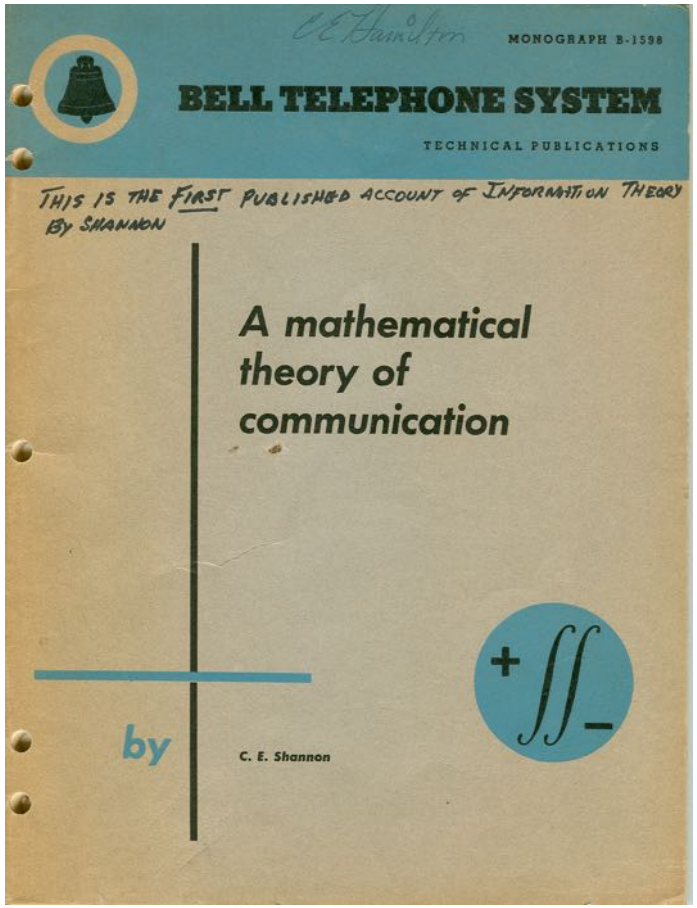
Science A short history of equations

Without Claude Shannon's information theory there would have been no internet

It showed how to make communications faster and take up less space on a hard disk, making the internet possible



The Guardian



Введена мера информации(!)

кг, метр, секунда + БИТ

1948

https://www.khanacademy.org/computing/computer-science/informationtheory/info-theory/v/intro-information-theory

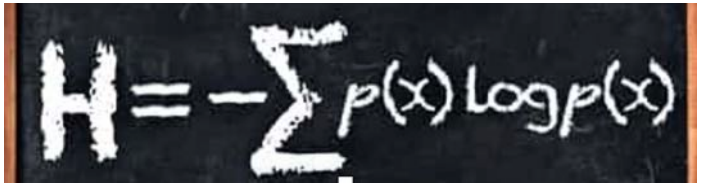
Информационная энтропия

Клод Шеннон предположил, что прирост информации равен утраченной неопределённости, и задал требования к её измерению:

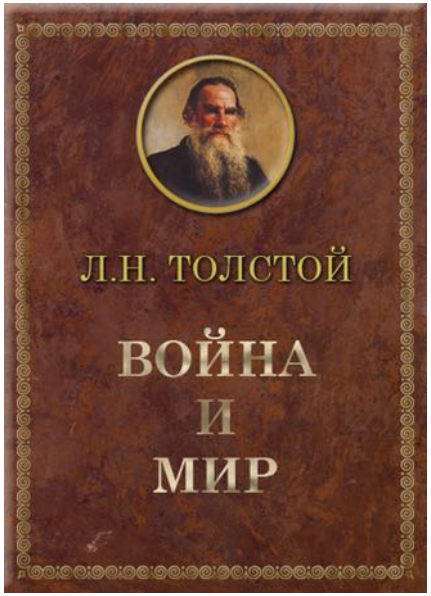
1.мера должна быть непрерывной; то есть изменение значения величины вероятности на малую величину должно вызывать малое результирующее изменение функции;

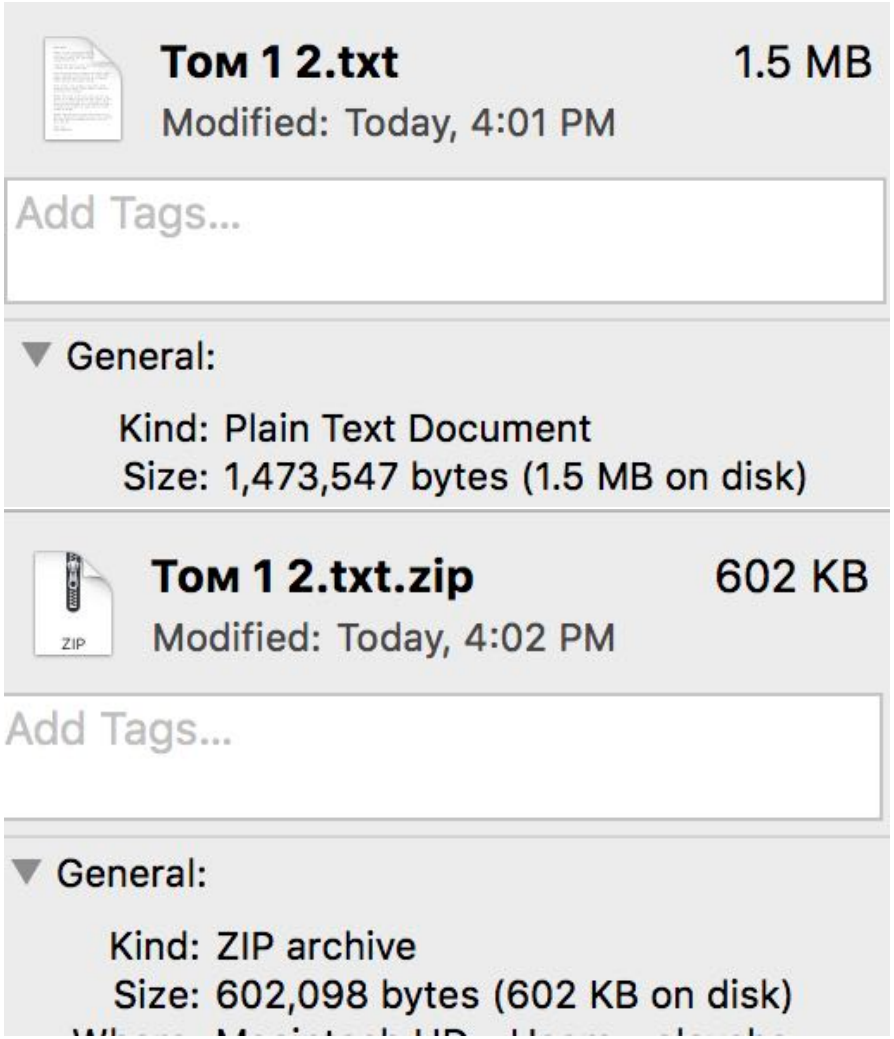
2.в случае, когда все варианты (буквы в приведённом примере) равновероятны, увеличение количества вариантов (букв) должно всегда увеличивать значение функции;

3.должна быть возможность сделать выбор (в нашем примере букв) в два шага, в которых значение функции конечного результата должно являться суммой функций промежуточных результатов.



Сжатие информации





Буква в тексте не независимы – одни встречаются чаще других в разных контекстах.

Сжатие информации

Кодирование длин серий

```
WWWWWWWWWWWWBWWWWWWWWWWWWBBBWWWWWWWWWWWWWWWWWWWWWWWWBWWWWWWWWWWWWWW
```

Посчитаем количество повторяющихся символов:

1. 12 символов «W»;
2. 1 символ «B»;
3. 12 символов «W»;
4. 3 символа «B»;
5. 24 символа «W»;
6. 1 символ «B»;
7. 14 символов «W».

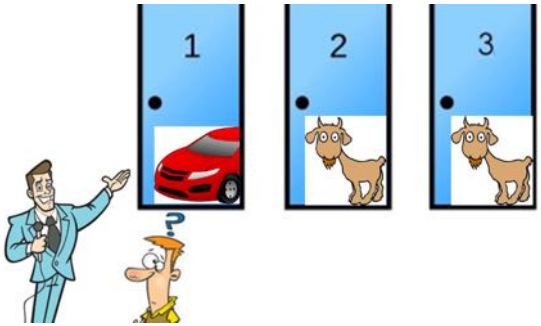
Итого найдено 7 серий. Заменим серии на число повторов и сам повторяющийся символ:

```
12W1B12W3B24W1B14W
```

https://ru.wikipedia.org/wiki/Кодирование_длин_серий

Информационная энтропия

$$H = -\sum p(x) \log p(x)$$

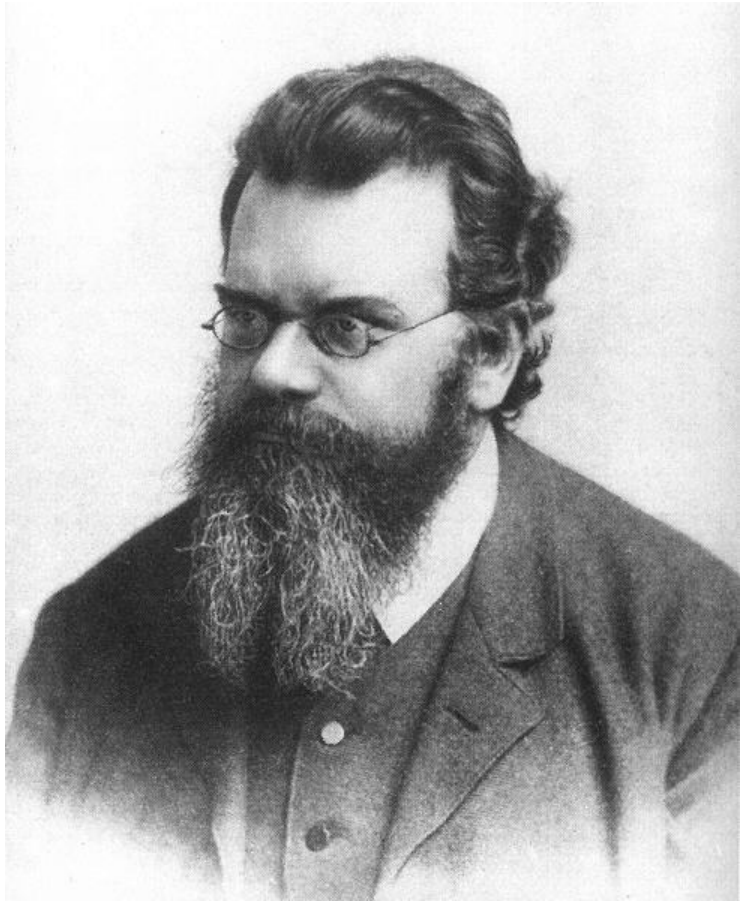


$$H = -\frac{1}{3}log_2\frac{1}{3} - \frac{1}{3}log_2\frac{1}{3} - \frac{1}{3}log_2\frac{1}{3} = -log_2\frac{1}{3}$$

1.5849625007211563... бит

Энтропия

Информационная и физическая энтропии имеют глубинную связь



Ludwig Eduard Boltzmann

$$H \stackrel{\text{def}}{=} \int P(\ln P)\, d^3v = \langle \ln P \rangle$$

В термодинамике и кинетической теории, H - теорема, полученная Больцманом в 1872 году, описывает неубывания энтропии идеального газа в необратимых процессах, исходя из уравнения Больцмана.

$$S \stackrel{\text{def}}{=} - NkH,$$

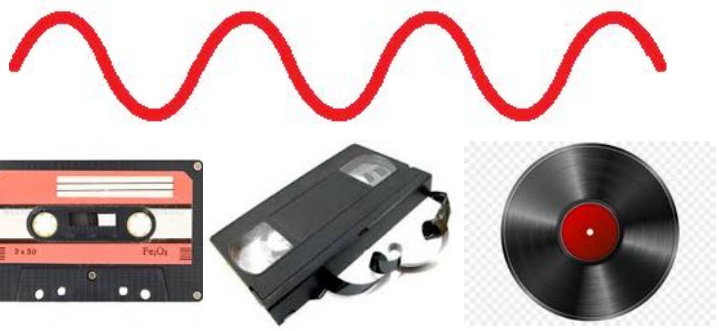
https://ru.wikipedia.org/wiki/H-теорема

Живые системы и цифровые технологии - аналогии

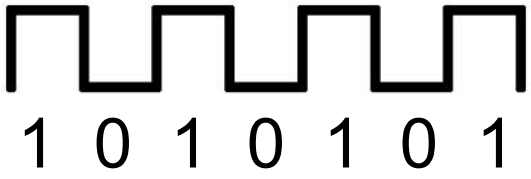
От программирования компьютеров к программированию живых систем

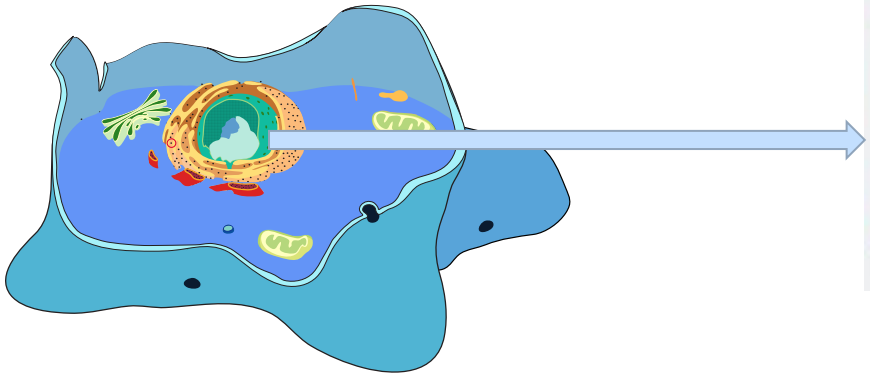
Живые системы и цифровые технологии

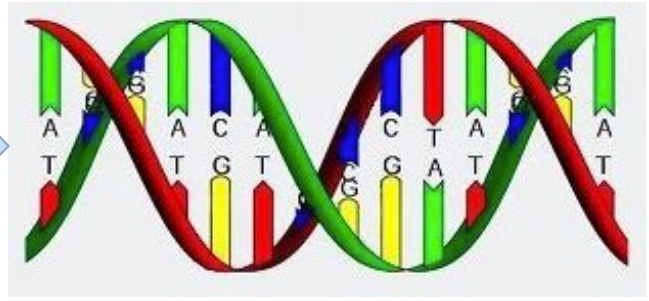
Аналоговые технологии



Цифровые технологии

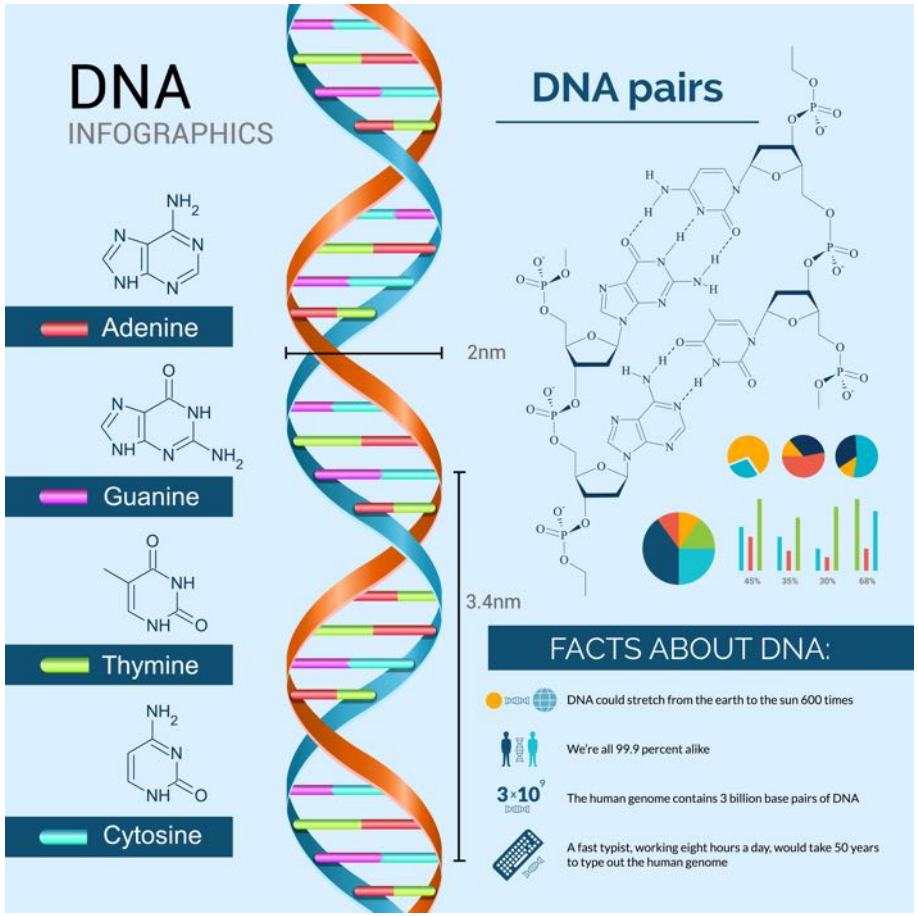


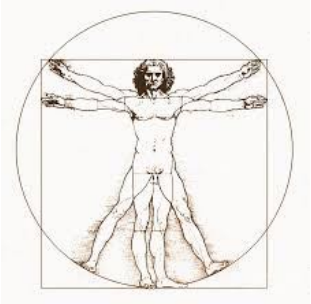


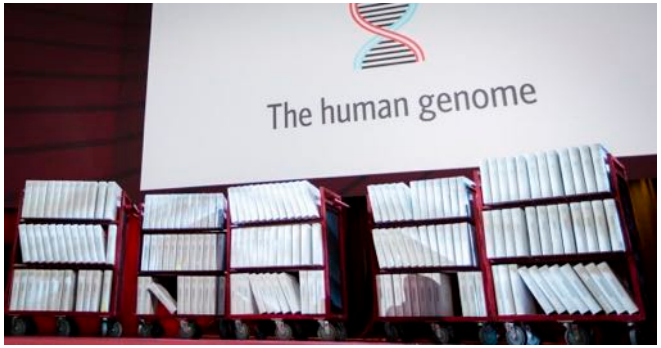


ДНК – цифровой код

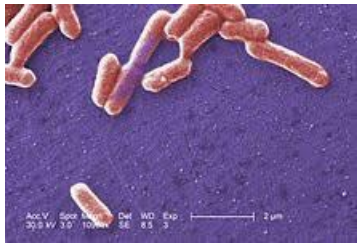
Генетический код





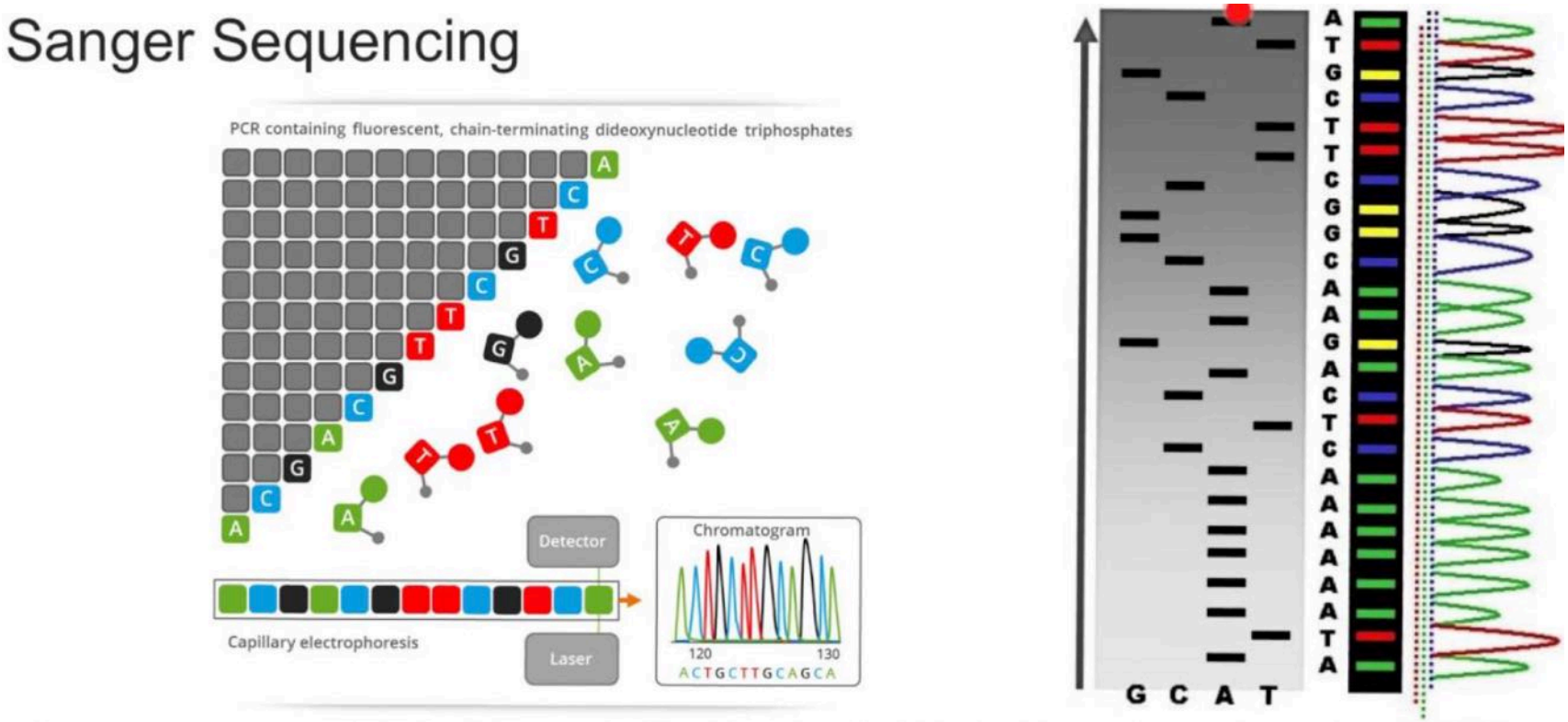


<u>Геном человека</u>
3 млрд нуклеотидов
262 тыс страниц
3 Гбайта



<u>Геном кишечной палочки</u>
4.6 млн нуклеотидов
400 страниц
4.6 Мбайт

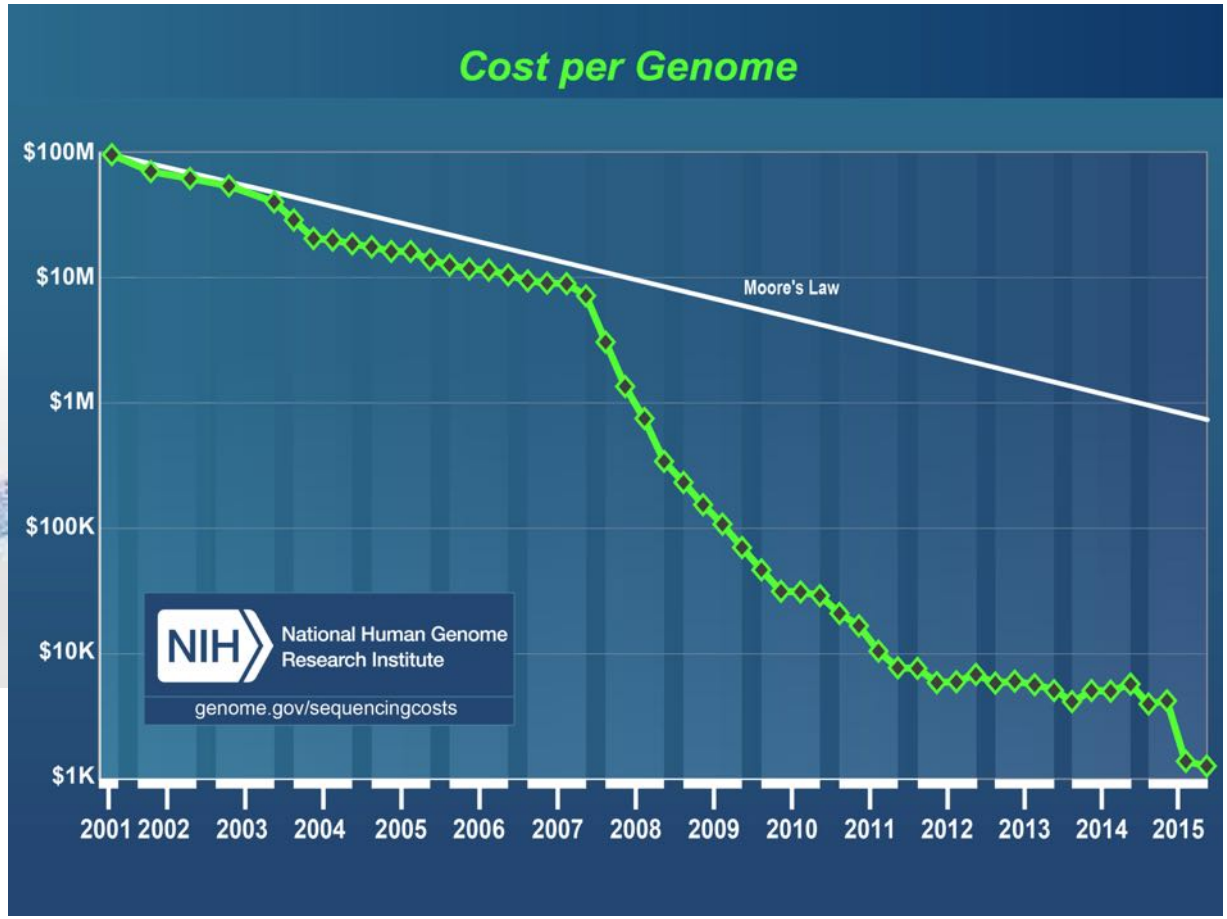
Секвенирование ДНК (метод Сэнгера)



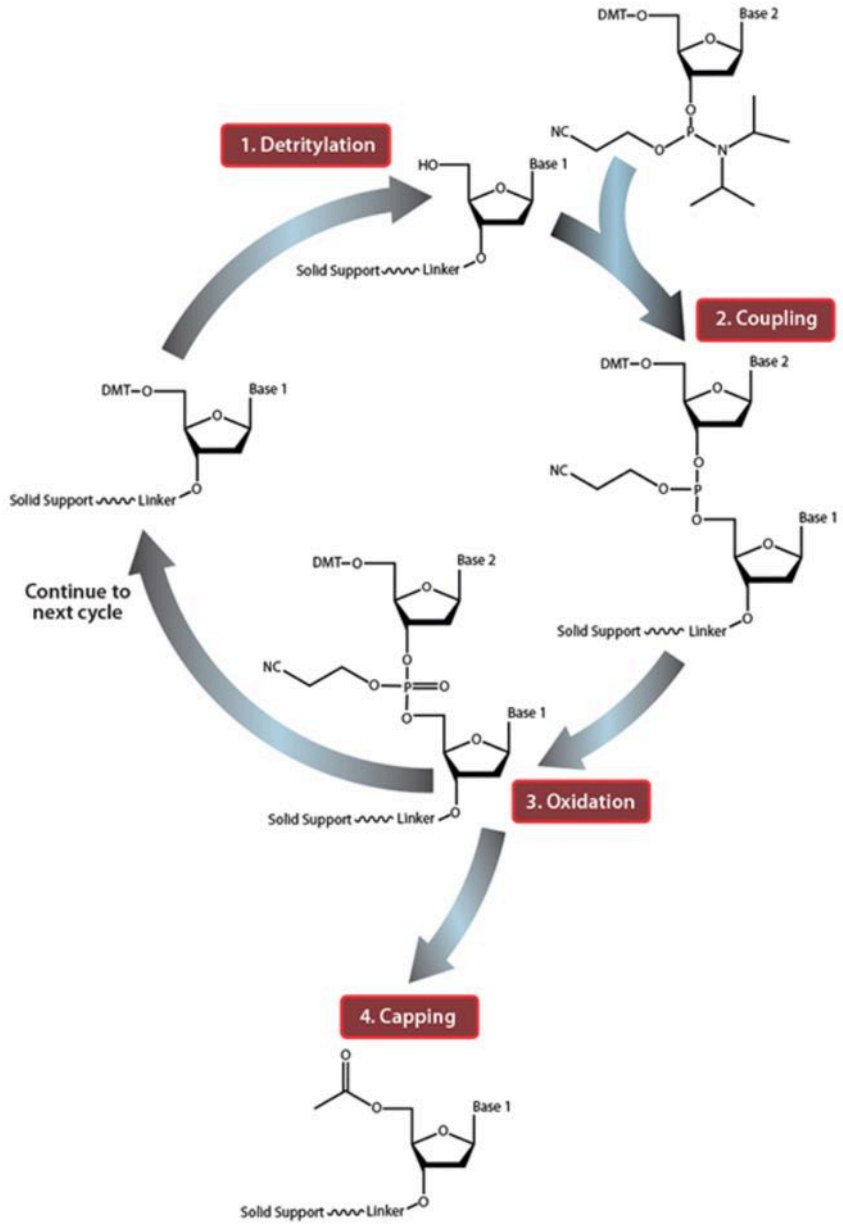
Sanger sequencing uses ddNTPs (dideoxynucleotide triphosphates) which do not have a free 3' OH mixed in with dNTPs
Whenever the DNA polymerase incorporates a ddNTP it won't be able to add any other nucleotides
Then gel electrophoresis is used to separate the DNAvv

https://www.youtube.com/watch?v=593zWZNwbJI

Технологии чтения и записи ДНК

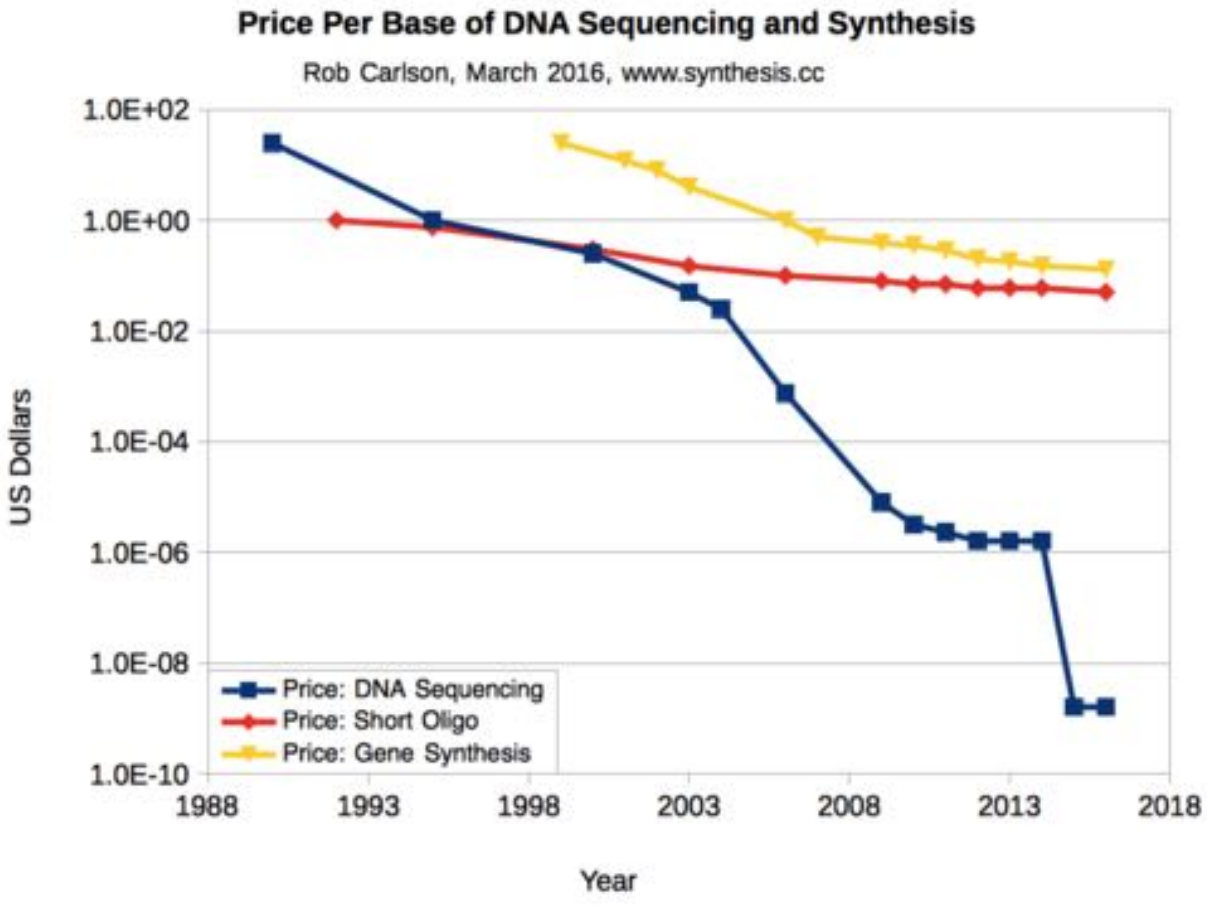


Синтез ДНК



https://www.sigmaaldrich.com/technical-documents/articles/biology/dna-oligonucleotide-synthesis.html

Удешевление синтеза ДНК



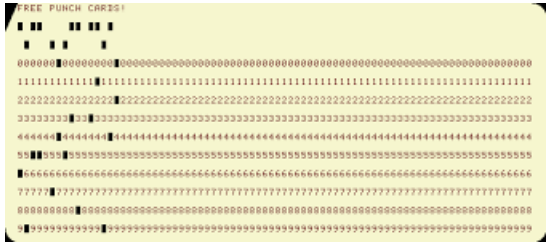
*All prices are for USD and Canadian dollars.

Description	Estimated TAT	Price (USD)
Standard Gene (≤500 bp)	6 days	**$125.00**
Standard Gene (501–1,000 bp)	6 days	**$0.35**/bp
Long Standard Gene (1,001–3,000 bp)	3–5 weeks	**$0.35**/bp

Gene Synthesis Price List - Eurofins Genomics
https://www.eurofinsgenomics.com/en/products/gene-synthesis/price-list/

About this result Feedback

От программирования компьютеров к программированию жизни

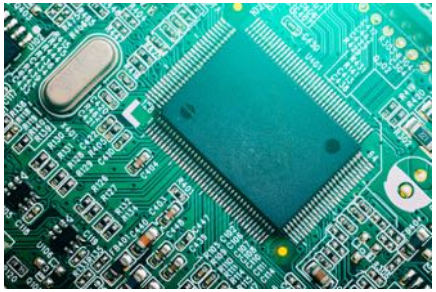


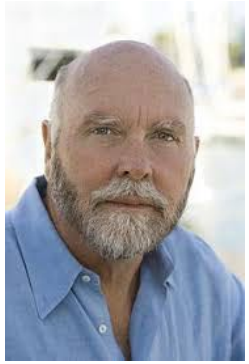
Ядро ОС Linux
~100 Мб

????

Аналогии компьютеров и живых клеток

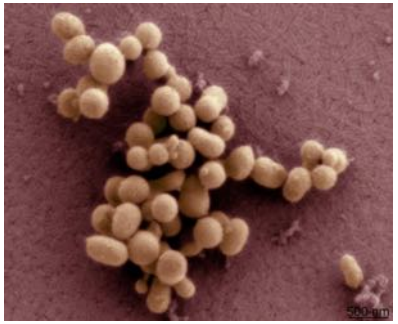
- Software boots into hardware
- Genetic code boots into cells





Craig Venter

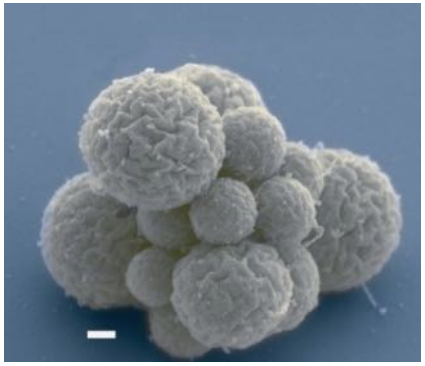
2010 *Mycoplasma mycoides* JCVI-syn1.0



Human	~20,000–25,000
Escherichia coli (K12 strain)	~4500
Syn 1.0	901
*Mycoplasma genitalium**	525
Syn 3.0	473

Первый искусственный геном (~$40 млн)

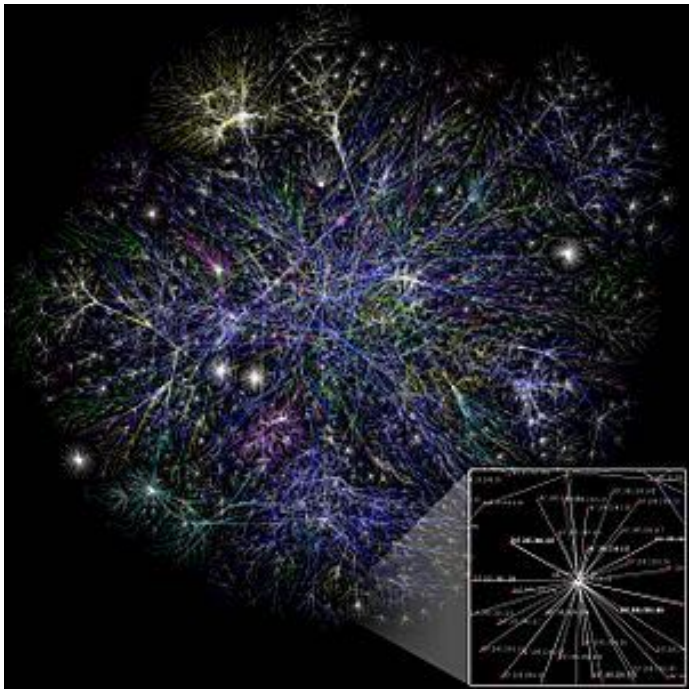
2016 *Syn3.0*



Минимальный геном

От программирования компьютеров к программированию жизни

~100 млрд. (10^11) нейронов
Каждый нейрон имеет
~7000 синаптических связей



23 млрд. устройств
подключенных к интернету

Сложные инженерные системы

~2500 лет
д.н.э.

Boeing 747
6 млн частей

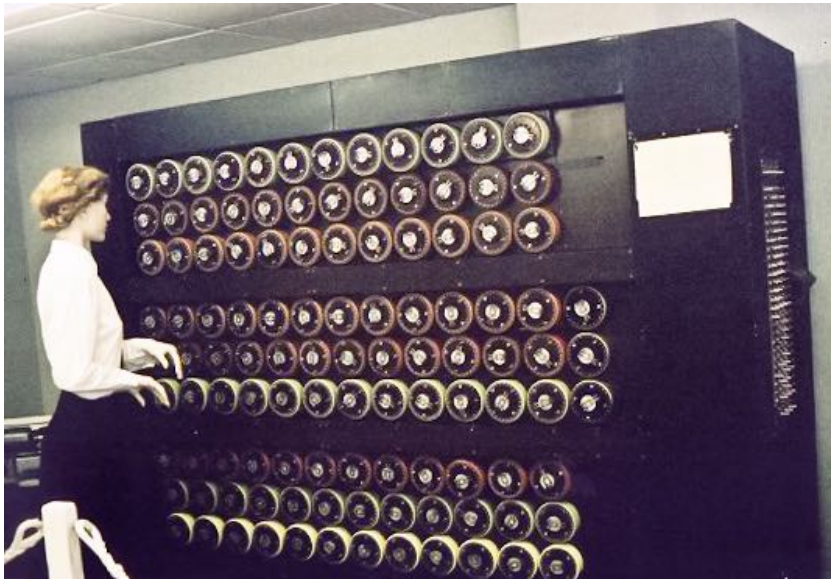
Boeing 777
Первый самолет спроектированный полностью на компьютере

Intel Xeon Phi CPU
8*10^9 транзисторов

Передача информации

Передача информации

- Комплексная область: **теоретические**, **практические**, физические аспекты
- Вопросы сжатия данных
- Вопросы надежности
- Вопросы **шифрования и защиты данных** (особенно в медицине и биологии)



https://en.wikipedia.org/wiki/The_Imitation_Game

Аналоговый сигнал

Цифровой сигнал

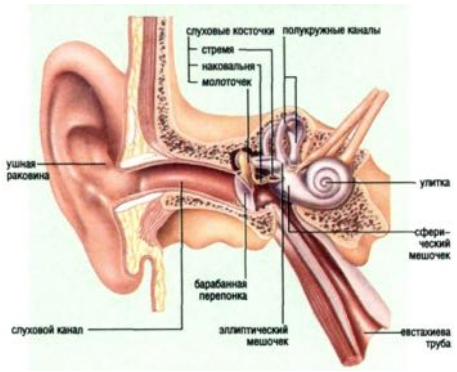
Как конвертировать?

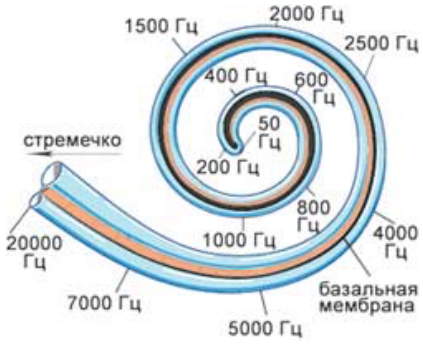
Передача информации

Связь частоты сигнала и пропускной способности

1908 - 2005

1889 – 1976





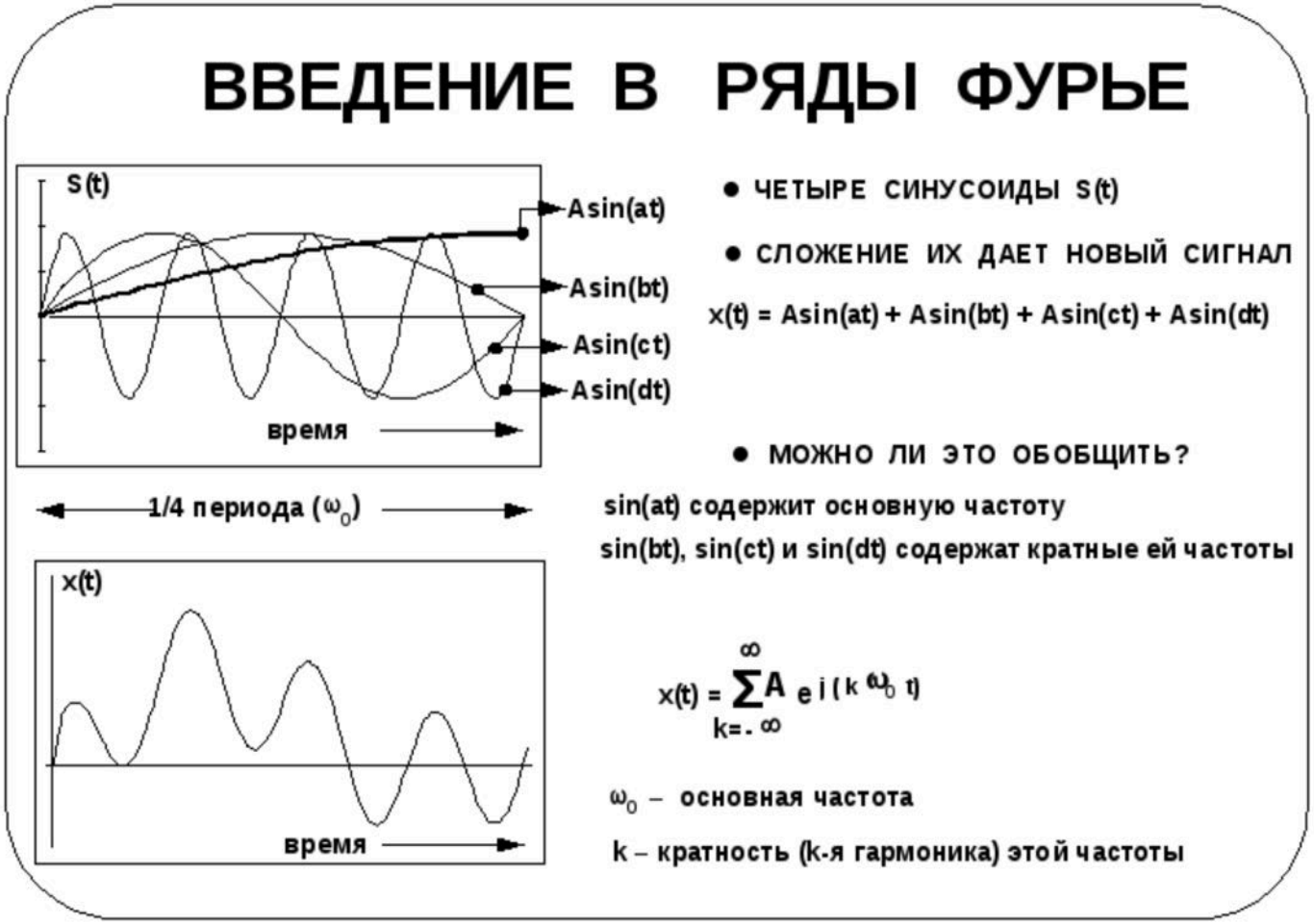
Теорема Котельникова-(Найквиста-Шенона)

«любую функцию F(t), состоящую из частот от 0 до f, можно непрерывно передавать с любой точностью при помощи чисел, следующих друг за другом через 1/(2f) секунд

44.1 кГц – частота дискретизации при записи звука

---

https://ru.wikipedia.org/wiki/Теорема_Котельникова



Говоря шире, теорема Котельникова утверждает, что непрерывный сигнал $x(t)$ можно представить в виде интерполяционного ряда:

$$x(t) = \sum_{k=-\infty}^{\infty} x(k\Delta)\,\mathrm{sinc}\left[\frac{\pi}{\Delta}(t - k\Delta)\right],$$

где $\mathrm{sinc}(x) = \sin(x)/x$ — функция sinc. Интервал дискретизации удовлетворяет ограничениям $0 < \Delta \leqslant \dfrac{1}{2f_c}$. Мгновенные значения данного ряда есть дискретные отсчёты сигнала $x(k\Delta)$.

Передача информации

Связь частоты сигнала и пропускной способности

1888 – 1970

$$C = B \log_2 \left(1 + \frac{S}{N}\right),$$

где

C — пропускная способность канала, бит/с;

B — полоса пропускания канала, Гц;

S — полная мощность сигнала над полосой пропускания, Вт или В²;

N — полная шумовая мощность над полосой пропускания, Вт или В²;

S/N — отношение мощности сигнала к шуму (SNR).

Теорема Шеннона-Хартли

https://ru.wikipedia.org/wiki/Теорема_Шеннона_—_Хартли

Передача информации

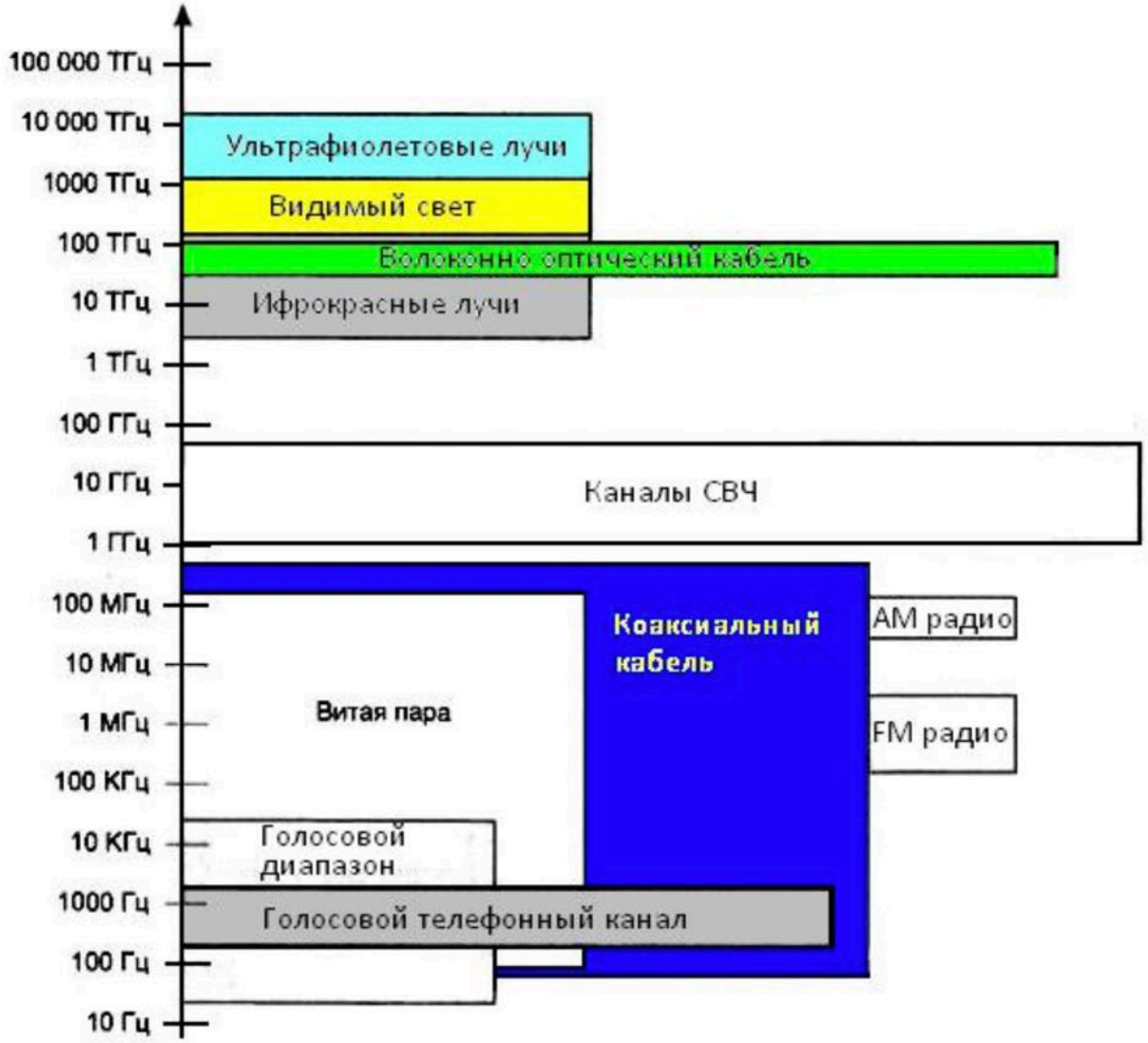
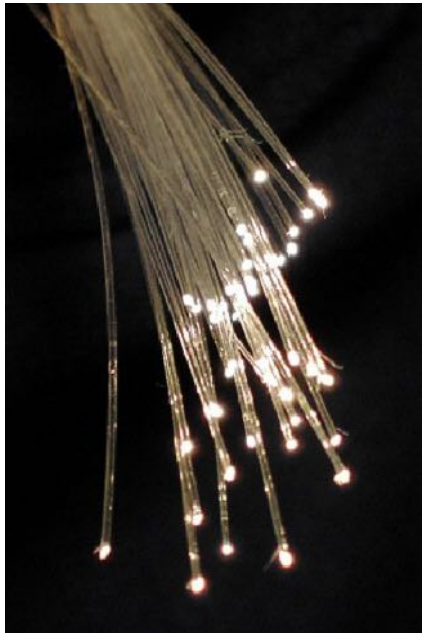


Рис. 1. Полосы пропускания линий связи и популярные частотные диапазоны

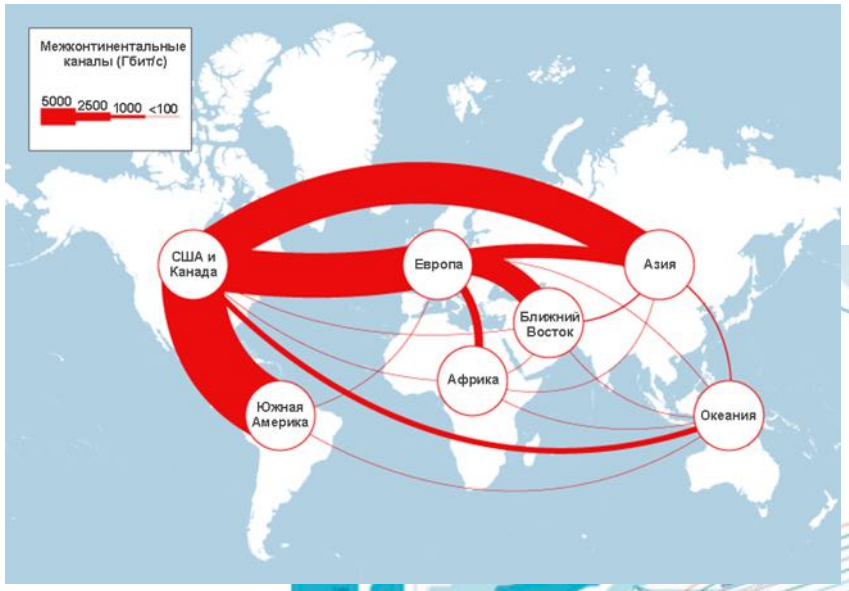


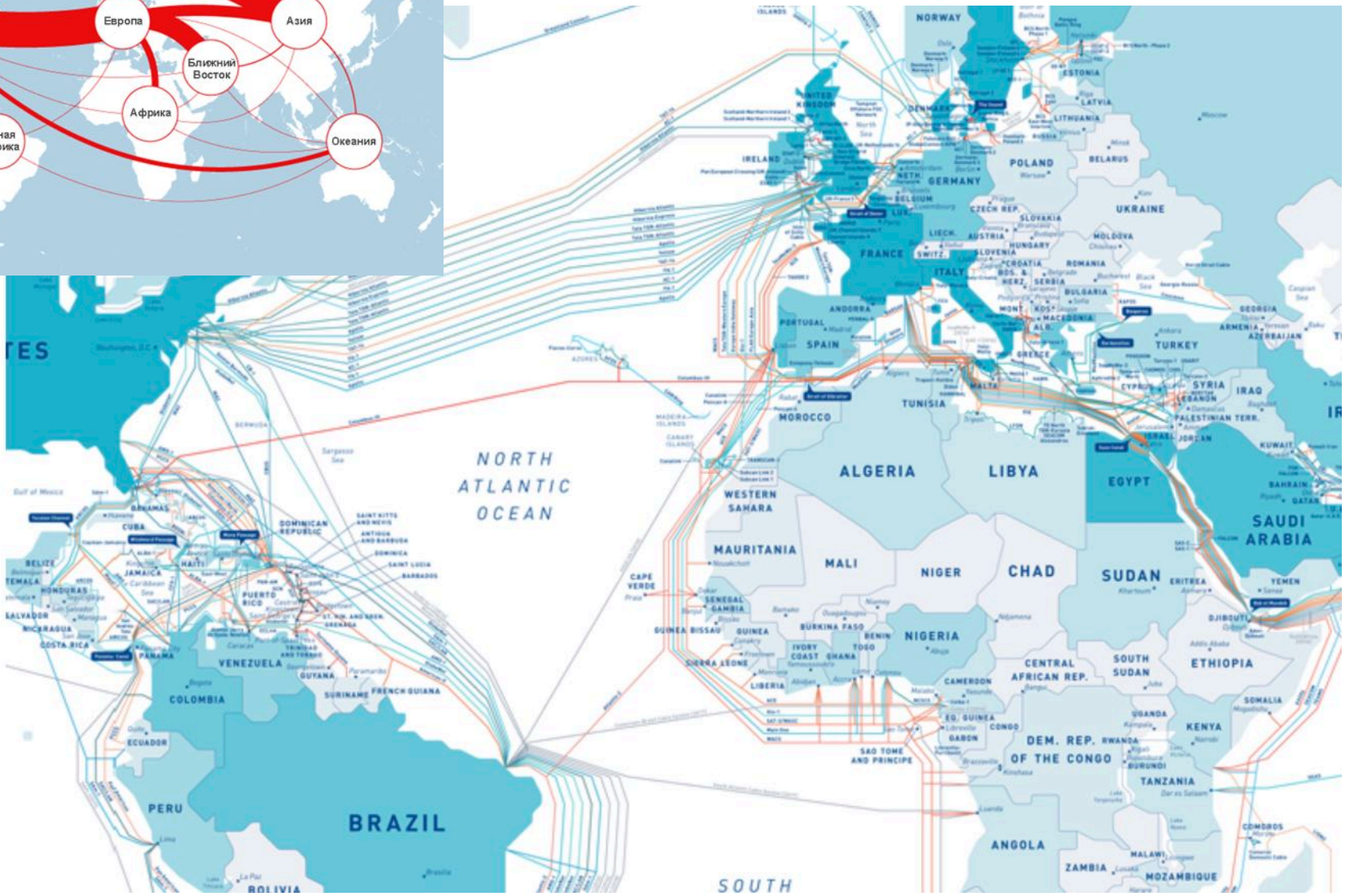
Оптоволокно

Антенны КНЧ

https://ru.wikipedia.org/wiki/Связь_с_подводными_лодками

Каналы связи



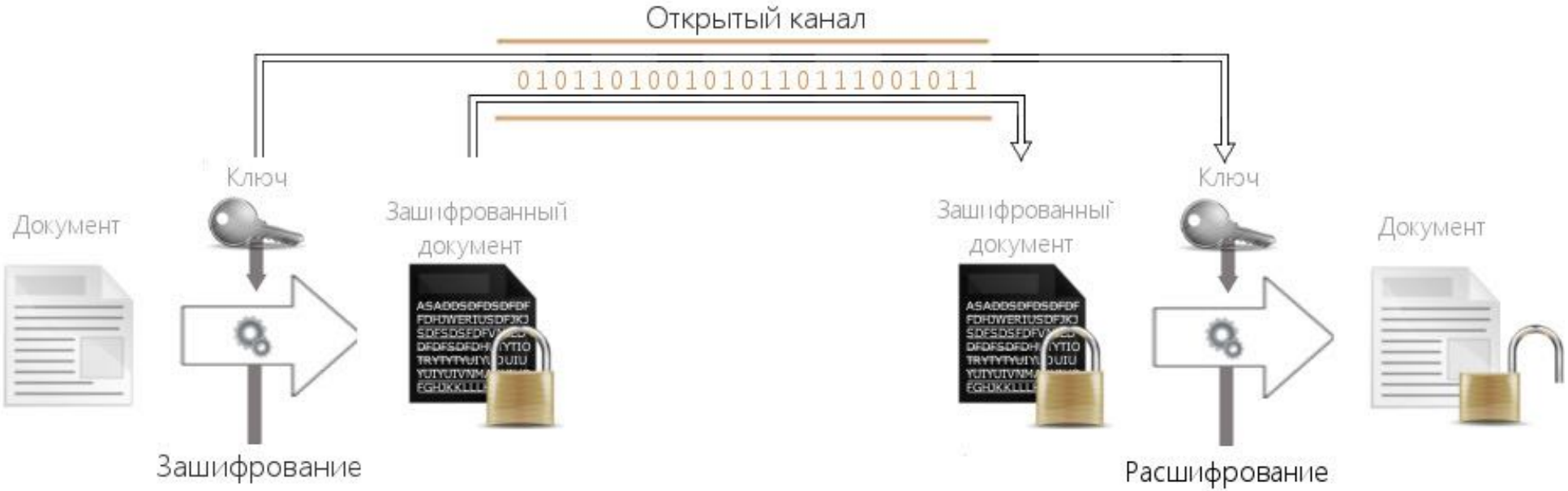


Карта подводных кабелей

---

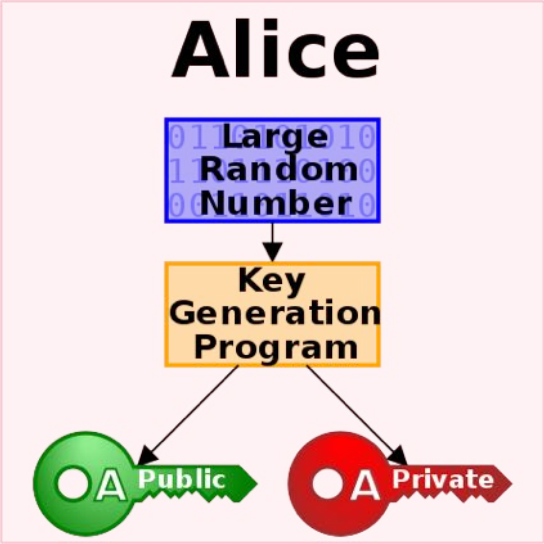
https://habrahabr.ru/company/rootwelt/blog/305634/

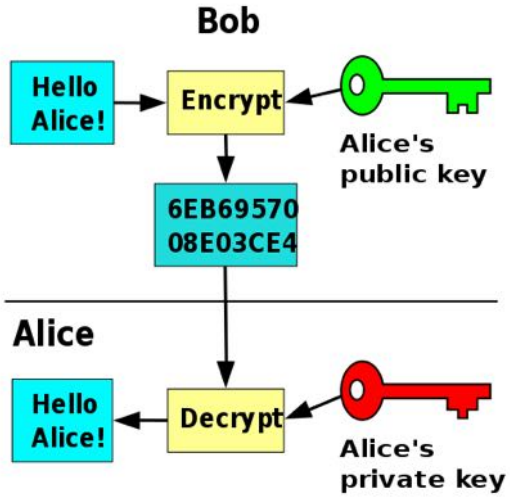
Шифрование информации



https://en.wikipedia.org/wiki/Public-key_cryptography

Криптосистемы с открытым ключом





Необратимая Хэш функция

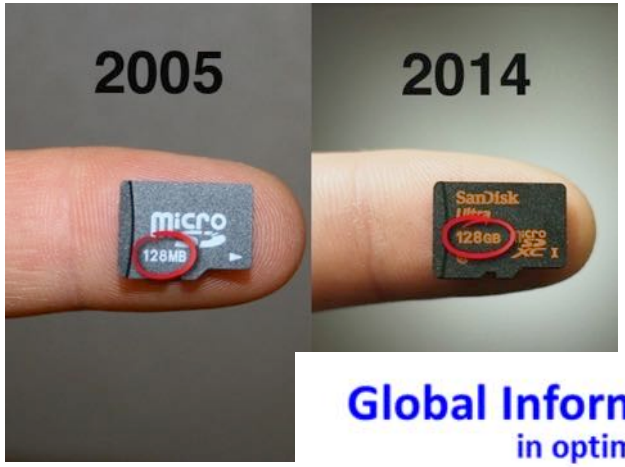
```
mbptb:~ alexsha$ md5 -s 'Hello world!!!'
MD5 ("Hello world!!!") = 87ee732d831690f45b8606b1547bd09e
```

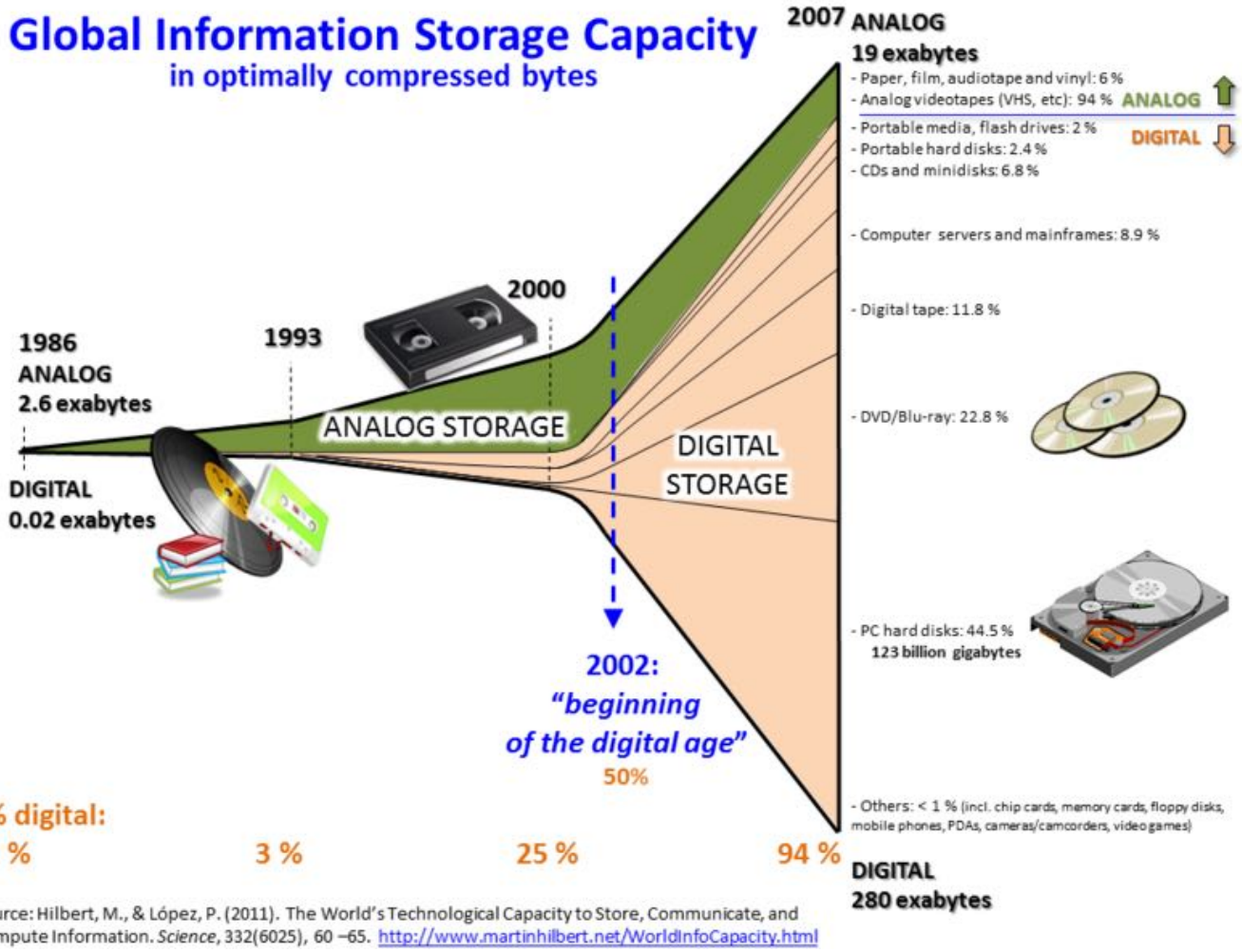
https://en.wikipedia.org/wiki/Public-key_cryptography

Хранение информации

Хранение информации

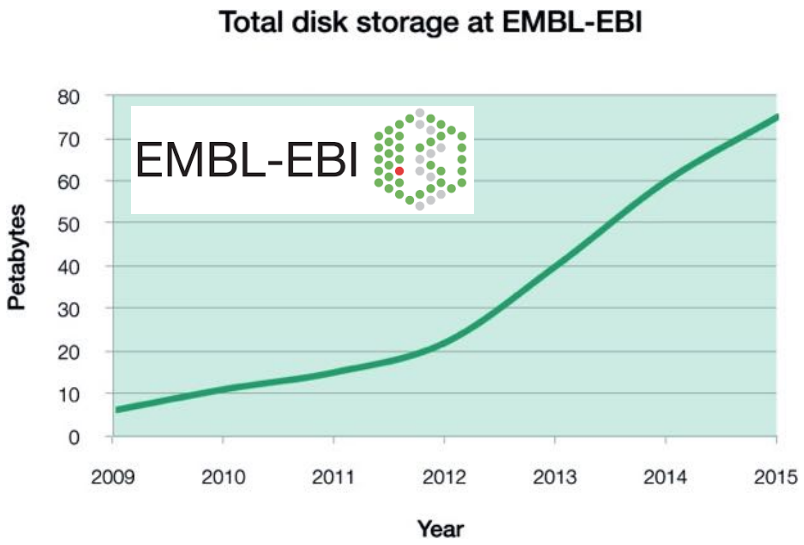
Хранение информации





Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Большие данные в биомедицине



NIH Big Data to Knowledge (BD2K)

2013-2021
~$400 млн

Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

Big Data: Astronomical or Genomical?, PLOS 2015

Источники больших данных в биомедицине

- Омиксные технологии
  - Секвенирование, геномика, транскриптомика, протеомика, метаболомика и т.д.
  - Коннектом мозга
- Медицинская информация
  - Электронные медицинские карты, результаты клинических исследований и т.д.
  - Медицинские изображения, МРТ и т.д.
- Структурная биология и моделирование
  - Данные с лазеров на свободных электронах (XFEL)
  - Моделирование структуры и динамики белков.

Данные секвенирования, пример
Геномы раковых опухолей

International Cancer Genome Consortium

Геном человека ~ 3.3 Gb
x100 секвенирование ~300Gb

Malignant Lymphoma Germany	Melanoma Brazil	Nasopharyngeal Cancer China
Non-Hodgkin Lymphoma Mexico	Oral Cancer India	Ovarian Cancer Australia

25000 образцов опухолей

Международный проект, данные распределены по миру

File ID	Donor	Repository	Project	Study	Data Type	Strategy	Format	Size
FI9995	DO217962	PCAWG - London, PCAWG - Barcelona, Collaboratory - Toronto, EGA - Hinxton	BRCA-EU	PCAWG	Aligned Reads	WGS	BAM	128.72 GB
FI9994	DO217962	PCAWG - London, PCAWG - Barcelona, Collaboratory - Toronto, EGA - Hinxton	BRCA-EU	PCAWG	Aligned Reads	WGS	BAM	107.27 GB
FI9974	DO46390	PCAWG - London, PCAWG - Barcelona, EGA - Hinxton, Collaboratory - Toronto, AWS - Virginia	OV-AU	PCAWG	Aligned Reads	WGS	BAM	134.05 GB
FI9973	DO46390	PCAWG - London, PCAWG - Barcelona, Collaboratory - Toronto, AWS - Virginia, EGA - Hinxton	OV-AU	PCAWG	Aligned Reads	WGS	BAM	101.45 GB
FI9956	DO222303	PCAWG - Chicago (TCGA), PDC - Chicago	DLBC-US	PCAWG	Aligned Reads	WGS	BAM	202.00 GB
FI9955	DO222303	PCAWG - Chicago (TCGA), PDC - Chicago	DLBC-US	PCAWG	Aligned Reads	WGS	BAM	100.77 GB

Централизованные репозитории омиксных данных

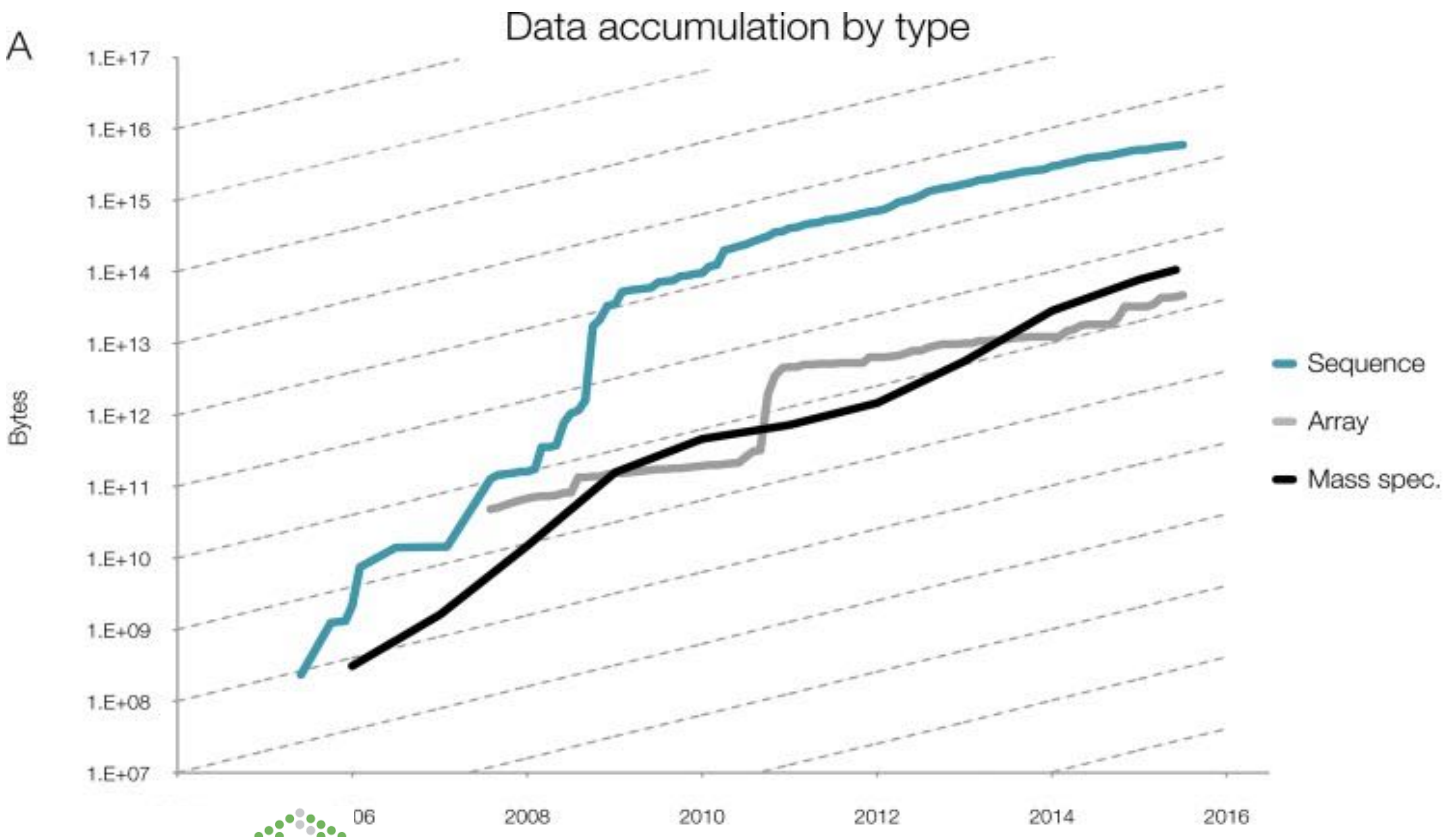
National Center for Biotechnology Information
NCBI

EMBL
European Molecular Biology Laboratory
EBI
European Bioinformatics Institute

DDBJ
DNA Data Bank of Japan
30th SINCE 1987

China National GeneBank
CNGB
国家基因库

华大基因
BGI

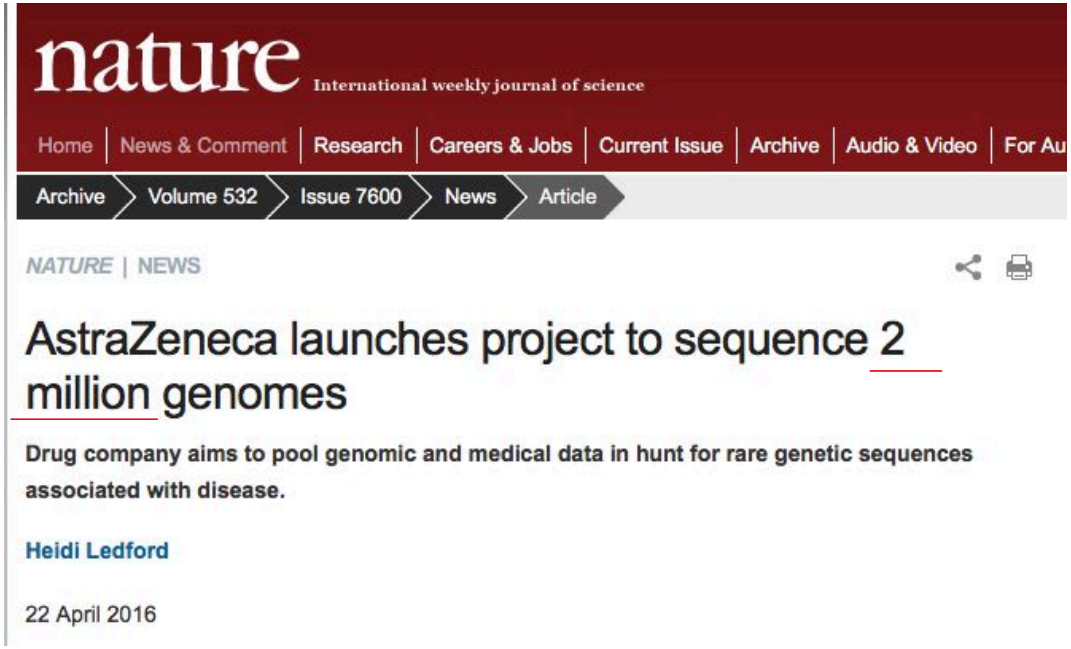


EMBL-EBI
European Bioinformatics Institute

Genomes en masse

برنامج الجينوم البشري السعودي
Saudi Human Genome Program

5 years ~ 100 000 genomes

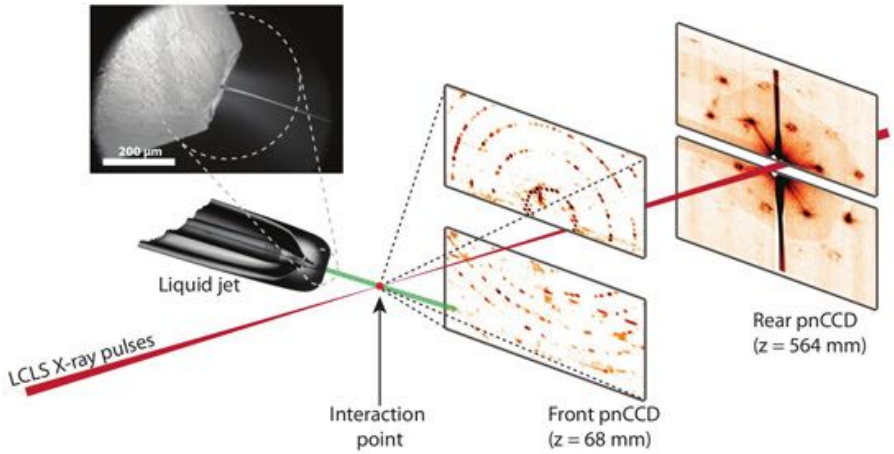


华大基因
BGI

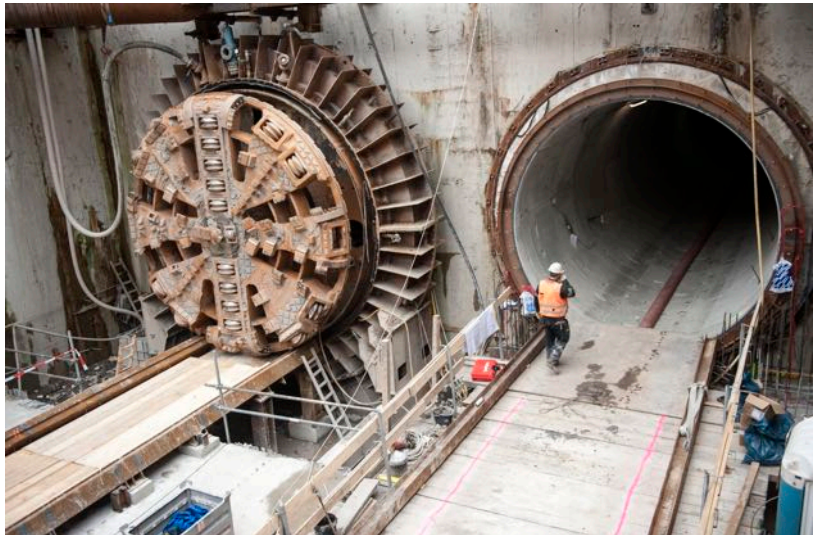
CNGB China National GeneBank 国家基因库

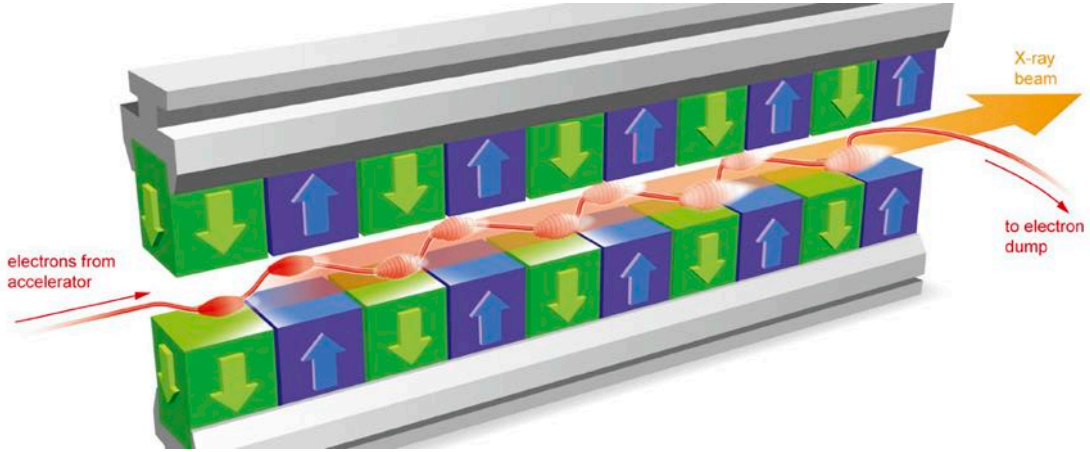
Структурная биология и моделирование

European XFEL, Hamburg



27000 импульсов в секунду





Передача больших данных

Выделенные научно-образовательные сети 100Gbit/s

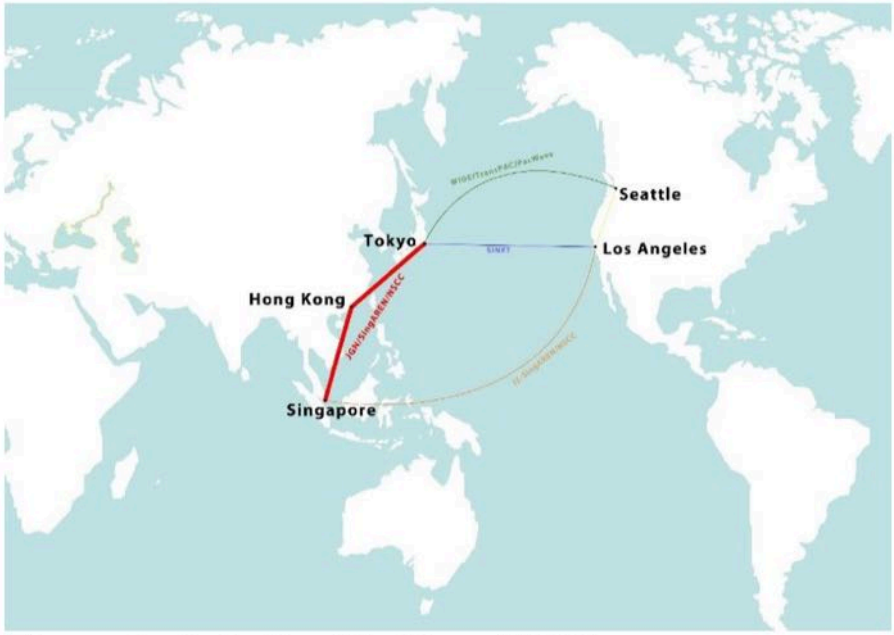


Figure 1: Asia Pacific Ring (APR)

INTERNET2

Программные решения

aspera
an IBM® company

globus toolkit

Базы данных

Базы данных

- Реляционные базы данных, объектно-ориентированные, RDF
- Системы управления базами данных СУБД
- Языки и стандарты SQL, SPARQL, RDF

Larry Ellison

Larry Ellison in 2016

Oracle Corporation

Реляционные базы данных

Клиент

Id _кл	Фамилия	Имя	Отчество
15	Иванов	Иван	Иванович
16	Петров	Петр	Петрович
17	Николаев	Николай	Николаевич

Товар

Id_тов	Название
1	Шкаф
2	Стул
3	Стол

Заказ

Id_зак	Клиент	Товар	Дата	Количество
1	15	1	15.09.2003	1
2	17	1	17.09.2003	2
3	15	2	20.09.2003	12

SQL

Целостность данных

Транзакции

Атомарность

Изолированность

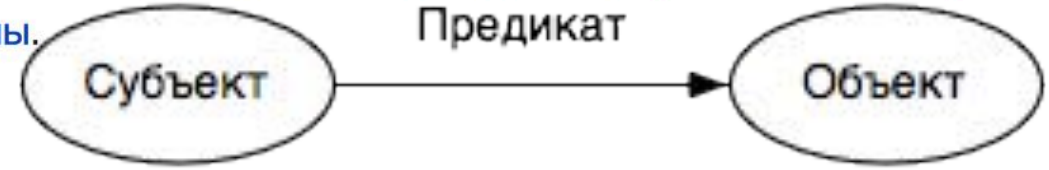
Надежность

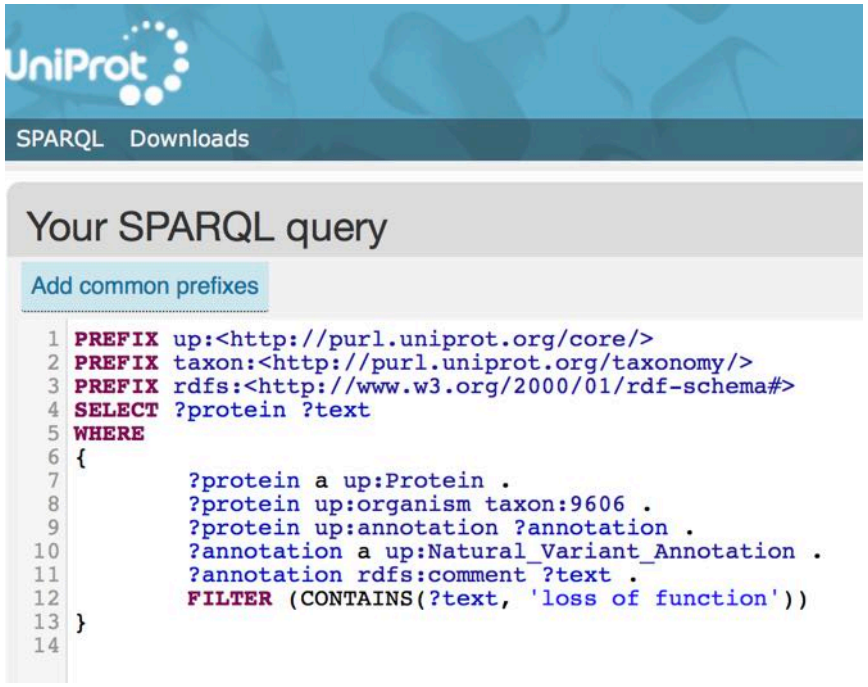
Соответствие требованиям ACID

https://aws.amazon.com/ru/relational-database/

Графовые базы данных

Resource Description Framework (RDF, «среда описания ресурса»[1]) — это разработанная консорциумом Всемирной паутины модель для представления данных, в особенности — метаданных[2]. RDF представляет *утверждения* о *ресурсах* в виде, пригодном для машинной обработки. RDF является частью концепции семантической паутины.





Язык SPARQL

Select all human UniProt entries with a sequence variant that leads to a 'loss of function'

https://ru.wikipedia.org/wiki/Resource_Description_Framework/

Словари, JSON

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

Спасибо за внимание!