Санкт-Петербургский
Государственный
Политехнический
Университет

Институт прикладной
математики и механики

КАФЕДРА
ТЕЛЕМАТИКА

Введение в профессиональную деятельность

Лекция 10
Наука о данных (Data Science): объем vs точность

СПб,
24 апреля , 2018 г.

Что обсуждалось на предыдущей лекции:

Каждый бакалавр- выпускник должен уметь :

- решать не менее 300 прикладных задач
- анализировать данные методами Data Mining, машинного обучения, нейронных сетей
- тестировать ПО, верифицировать кодов, валидировать результаты вычислений
- применять компьютерные сети (стек TCP/IP, методы сетевого программирования на C/C++);

Почему так: Проблема «сложности **и** цикл «цифровой трансформации» знаний

Концепция «цифровой» науки

Компьютерные науки

Науки о природе и жизни

компьютерные науки

искусственный интеллект

«большие медицинские данные»

ЗНАНИЕ / ПОНИМАНИЕ

↙ в форме компьютерных моделей

↘ причин возникновения заболеваний и происходящих в организме биохимических процессов

Системы искусственного интеллекта - как «очки» для повышения «остроты» интеллекта человека

- Для интеллекта человека характерны две ключевые модальности - «знать и понимать»
  - **Знать** – это функция памяти, которая позволяет хранить данные. Эту функцию можно передать компьютерным системам.
  - **Понимать** – это функция выявления связей, интерпретации зависимостей и осознания причин.

- Искусственный Интеллект должен гармонично дополнить функции интеллекта человека средствами хранения и обработки данных, как оптические очки **повышают остроту** зрения, но **не заменяют функцию** зрения.
- Системы ИИ смогут выполнять роль интеллектуального «интерфейса» между «миром людей», наделенных знаниями и способностью понимать, и «миром машин», способных хранить, обрабатывать и агрегировать огромные объемы данных

Что мы добиваемся: на примере медицины

- Научиться «вычислять» персональную стратегию излечения конкретного пациента от характерных патологий, используя методы «машинного обучения для анализа больших объёмов медицинских данных, что позволяет
  - Снижать риски послеоперационных осложнений
  - Повышать точность диагностики заболеваний
  - Сокращать временя пребывания пациента в стационаре

Структура интеллектуальной системы управления состоянием пациента

«большие данные» о состоянии организма человека-пациента

мониторинг

терапия

геномные данные

нейронная сеть

нейронная сеть

нейронная сеть

прогноз процессов

прогноз процессов

прогноз процессов

диагноз

диагноз

управляющее корректирующее воздействие

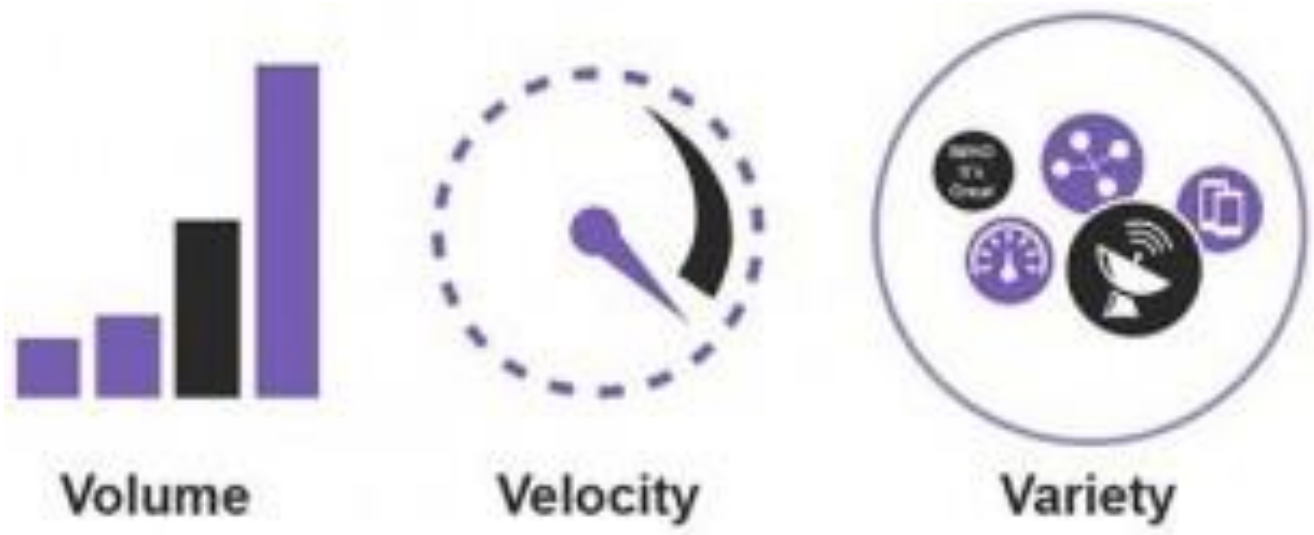
ТОП 50 Российских суперкомпьютеров

Текущий рейтинг

26-я редакция от 04.04.2017г.

N	Место	Кол-во CPU/ядер	Архитектура (тип процессора / сеть)	Производительность (Tflop/s) Linpack	Производительность (Tflop/s) Пиковая	Разработчик
1	Москва Московский государственный университет имени М.В.Ломоносова 2016 г.	1472/42688	узлов: 1472 (Xeon E5-2697v3 [Acc: Tesla K40M] 2.6 GHz 64 GB RAM) сеть: FDR Infiniband/FDR Infiniband/Gigabit Ethernet	2,102.00	2,962.30	Т-Платформы
2	Москва Московский государственный университет имени М.В.Ломоносова 2012 г.	12422/82468	узлов: 4160 (2xXeon 5570 2.93 GHz 12 GB RAM) узлов: 777 (2xXeon E5630 [Acc: 2xTesla X2070] 2.53 GHz 12 GB RAM) узлов: 640 (2xXeon 5670 2.93 GHz 24 GB RAM) узлов: 288 (2xXeon E5630 [Acc: 2xTesla X2070] 2.53 GHz 24 GB RAM) узлов: 260 (2xXeon 5570 2.93 GHz 24 GB RAM) узлов: 40 (2xXeon 5670 2.93 GHz 48 GB RAM) узлов: 30 (2xPowerXCell 8i 3.2 GHz 16 GB RAM) узлов: 4 (4xXeon E7650 2.26 GHz 512 GB RAM) сеть: Infiniband QDR/Gigabit Ethernet/Gigabit Ethernet	901.90	1,700.21	Т-Платформы
3	Санкт-Петербург Суперкомпьютерный центр Санкт-Петербургский политехнический университет 2017 г.	1468/20552	узлов: 623 (2xXeon E5-2697v3 2.6 GHz 64 GB RAM) узлов: 56 (2xXeon E5-2697v3 [Acc: 2x NVIDIA K40] 2.6 GHz 64 GB RAM) узлов: 36 (2xXeon E5-2697v3 2.6 GHz 128 GB RAM) узлов: 8 (2xXeon E5-2697v3 [Acc: NVIDIA K1] 2.6 GHz 128 GB RAM) узлов: 8 (2xXeon E5-2697v3 [Acc: NVIDIA K2] 2.6 GHz 128 GB RAM) узлов: 3 (2xXeon E5-2697v3 2.6 GHz 256 GB RAM) сеть: FDR Infiniband/Gigabit Ethernet/Gigabit Ethernet	715.94	1,015.10	Группа компаний РСК
4	Москва МСЦ РАН 2016 г.	416/28704	узлов: 208 (2xXeon E5-2690 [Acc: 2x Xeon Phi 7110X] 2.9 GHz 80 GB RAM) сеть: FDR Infiniband/Gigabit Ethernet/Fast Ethernet	383.21	523.83	Группа компаний РСК

Особенности «Науки о данных « - volume, velocity, variety



Проблема VVV – в качестве базового принципа обработки «больших данных» указывают на требование **горизонтальной масштабируемости**, путем распределения алгоритмов обработки на сотни и тысячи компьютерных узлов, без потери производительности вычислений.

От «данных большого объема» к математике «больших данных «

- Классическая математика описывает то, что «однородно и делимо». Но не все объекты реальности являются таковыми. Нужны новые методы изучения закономерностей, описывающих объекты окружающего нас мира как ЦЕЛОСТНОЙ СИСТЕМЫ:
  - Данные представимы множеством точек **топология** которых отражает физический феномен. Однако, феномен может быть один – а «данных» может быть много.
  - **Нужна математика больших данных которая** позволят строить математический объект (группу или кольцо) в контексте целевых условий как топологический инвариант.

Топология, топологическое пространство, топологический инвариант

- Топология изучает свойства метрических пространств, которые остаются неизменными при непрерывных деформациях
- Топологи́ческое простра́нство — множество с дополнительной структурой определённого типа (топологией)

Сколько данных нужно , чтобы «закодировать» мир ?

- Информация в компьютере записывается в виде «нулей и единиц»

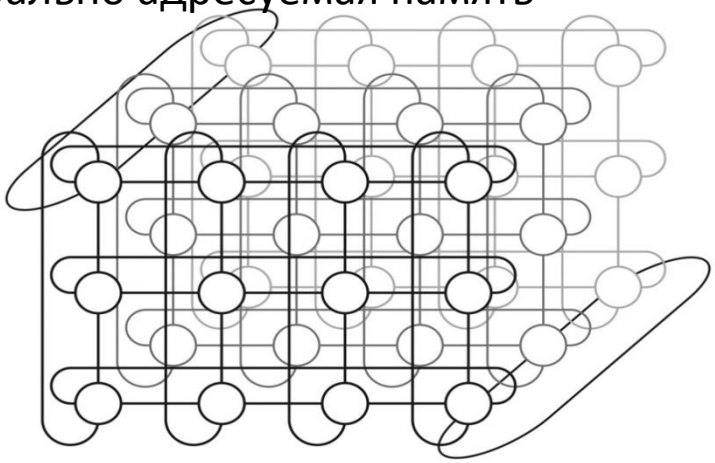
Так, чтобы сохранить фотографию объемом 3 мегабайта надо

$3x8x2^{20}$ т.е. больше 25 миллионов ячеек памяти компьютера

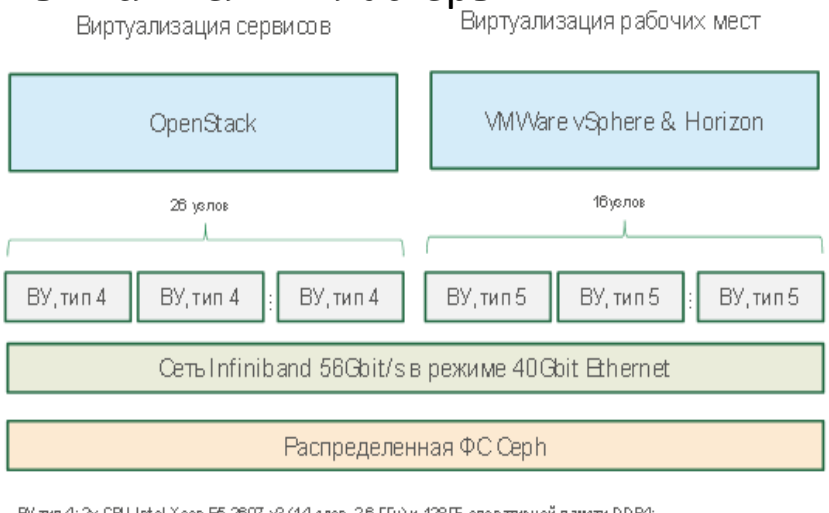
- Данные для обработки могут храниться в оперативной памяти или на «жестком диске». Объем хранимых данных практическим ни чем не ограничен, но ограничена скорость доступа к хранимым данным.
- Метод «LogLog обработки» - преобразование массива входных данных произвольной длины в (выходную) битовую строку фиксированной длин:
  - $\log_2(\log_2(\ 1000\ 000\ 000)) = 4.9$ Каждой записи присваиваться т.н. хэш-значение, т.е. вычисляется свертка. Возвращаемые хеш-функцией значения (выходные данные) менее разнообразны, чем значения входного массива (входные данные). Пример - хеш-функция остаток от деления входных данных на M. Не всегда, если записи совпадают, то и их хеш-значения тоже совпадают.

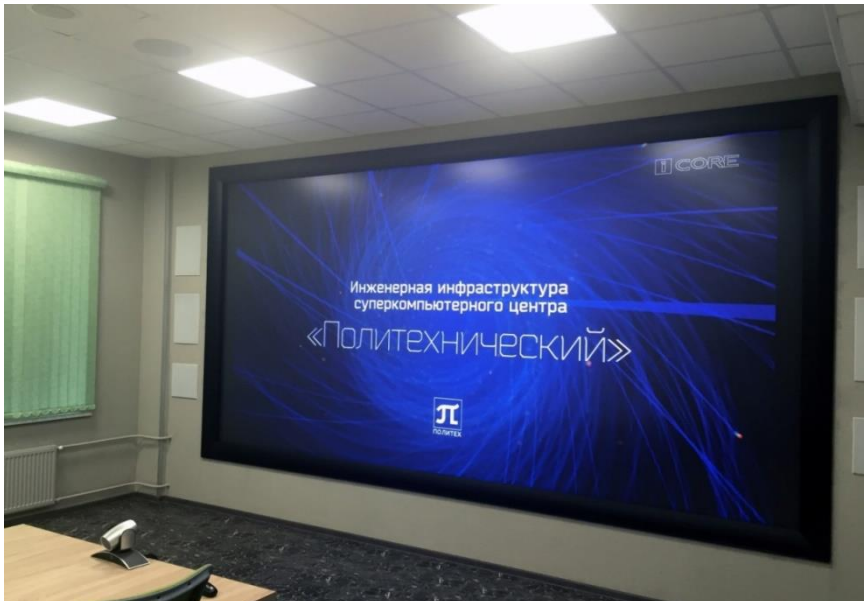
Математическое моделирование – это построение «хэш функций» на основе данных, полученных для разных масштабов описания физической реальности

Глобально адресуемая память

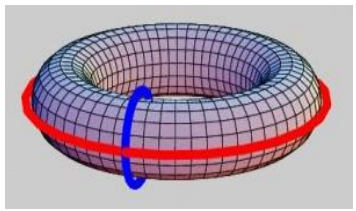
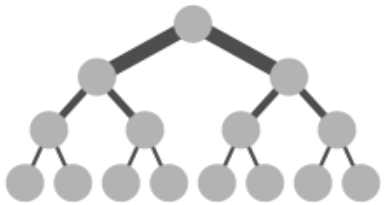


Многоядерные многопотоковые вычислители - кластеры





Эко-система СКЦ «Политехнический»

1. Вычислительная система с глобальной памятью	2. Кластер с графическими ускорителями CPU/GPGPU	3. Система с ультравысокой многопоточностью	4. Среда облачных вычислений
Тороидальная топология межузловой связи	Гибридная кластерная архитектура, топология связи узлов «толстое дерево»	Связь узлов на основе технологии InfiniBand FDR	Динамическая архитектура класса IaaS
16 узлов, МП x3 AMD Opteron 6380 3072 ядра; 12288 ГБ RAM общей памяти; пиковая производительность **31 ТФлопс**	668 вычислительных узлов МП x2 Intel Xeon 2697 , 56 узлов МП ускоритель NVIDIA K40 Всего 1336 МП ч86, 18704 ядра; 112 GPGPU ускорителей, пиковая производительность **938 ТФлопс**	256 узлов x8 Intel Xeon Phi 5120D Всего: 15360 ядер; 61440 потоков; 2048 ГБ RAM Пиковая производительность **259 ТФлопс**	44 вычислительных узла x2 Intel Xeon 2697, Всего 1232 ядра; 5632 ГБ RAM; пиковая производительность **51 ТФ**
n-Top 			

Сетевая инфраструктура связи узлов , и оперативное хранилища данных 1 ПБ

Предельные возможности современных «компьютеров»

Так, минимально возможные линейные размеры канала транзистора (x_{min}) и максимальная рабочая частота ($f_{max} = 1/t_{min}$) ограничиваются известными соотношениями неопределенностей Гейзенберга (h – постоянная Планка):

$\Delta x * \Delta p = x_{min} * \Delta p \geq h/2\pi$, где $\Delta p = \sqrt{2} * m * E_{bit} = \sqrt{2} * m * kT * \ln 2$

$\Delta t * \Delta E = t_{min} * \Delta E \geq h/2\pi$, где $\Delta E = kT * \ln 2$

при Т=300 К, получим x_{min} = 1.5 нм, а $t_{min} = 4*10^{-14}$ с .

Если создать микропроцессор, у которого одновременно будет и самая большая плотность упаковки (определяемая x_{min}) и максимально возможная частота (определяемая t_{min}), и оценить, какая при таких условиях должна выделяться мощность *P* на единицу площади, то при Т=300 К, получим:

$$P = E/t = n_{max} E_{bit}/t_{min} = kT * \ln 2/((x_{min})^2 * t_{min}) = 3.2 * 10^6 \text{ Вт/см}^2,$$

где n_{max} – число транзисторов, которые занимают площадь, определяемую линейными размерами канала (x_{min}).

Что важно понимать

"Я все больше и больше склоняюсь к мысли, что нельзя продвинуться дальше, используя теории, строящиеся на континууме".

А. Эйнштейн

Среди всех понятий физики время оказывает наибольшее сопротивление свержению мира идеального континуума в мир дискретности, информации, битов…

Дж. А. Уилер

Вывод. Результат вычислений это новый объект !?