Санкт-Петербургский
Государственный
Политехнический
Университет

Институт прикладной
математики и механики

КАФЕДРА
ТЕЛЕМАТИКА

Введение в профессиональную деятельность

Лекция 6_3

. Что такое «наука о данных?»

21 апреля 2020 г.

Что обсуждали на прошлой лекции

Результаты физических измерений (**составшаяся** реальность) – носят вероятностный характер, поэтому доступное нам «научное» или количественное **описание реальности** по своей природе **информационно**, где $I = -\log(P)$

Физическая реальность **вычислима**, т.е. обладает свойствами, которые непосредственно выражаются через числовые отношения, Остался открытым вопрос, вычислима ли «метальная реальность» если число рассматривается как:

- информационная единица
- способ записи экстенциональных свойств («вес» объекта)
- код записи интенсиональных свойств («смысл» описания)

?

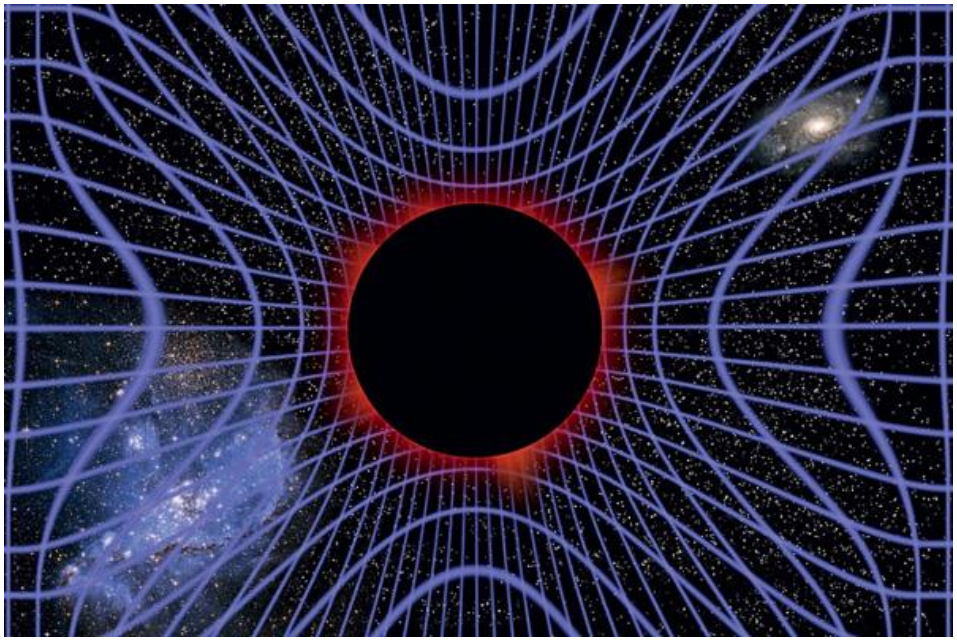
Также надо было ответить на вопросы:

- Как и чем характеризовать экстенционал (количественную характеристику) мозга ?
- Каким образом в компьютере можно сформировать интенсиональную (качественную) характеристику объектов реальности объекта ?
- Может ли компьютерная система, использующая цифровые модели, быть «непротиворечивой» и «целостной» ?

Наука о данных : почему абстрактные математические теории помогают понять, как устроен реальный мир ?

Современная физика в 17 веке началась с создания математического аппарата, который позволяет оперировать **числовыми данными**, с помощью других **числовых данных** (фундаментальных постоянных), но до конца не ясно, что они означают

Подобное выражается через подобное:



Ньютон «Общая арифметика»(1707г.). «Под числом мы подразумеваем не столько множество единиц, сколько абстрактное отношение какой-нибудь величины к другой величине такого же рода, взятой за единицу.

В законе Ньютона фигурируют силы тяготения, которые можно измерять количественно…. в выбранной системе единиц. Значения гравитационного потенциала можно изменить на любую постоянную величину — градиент останется тем же.

Проблема «данных» в Data Science

- Одному и тому же физическому полю могут соответствовать разные потенциалы. Например, к векторному потенциалу можно добавить любой постоянный вектор, а к скалярному — любое число
- Наука о данных (англ. data science или «даталогия» — datalogy) — раздел компьютерных наук, изучающий проблемы анализа, обработки и представления данных в цифровой форме в зависимости от контекста их получения.
- Метафора Big Data —термин одного из направлений компьютерных наук - Data Processing с учетом не только характеристики «**Объем** или Volume», но и **Скорости** изменения или Velocity и **Разнообразия** или Variety

Что такое «наука о данных» ?

- Для анализа больших объемов данных применяются различные методы «математики больших данных»:
  - нейронные сети — модели, построенные по принципу организации и функционирования биологических нейронных сетей;
  - методы предсказательной аналитики,
  - статистики и Natural Language Processing (направления искусственного интеллекта и математической лингвистики, изучающего проблемы компьютерного анализа и синтеза естественных языков).
  - методы, привлекающие экспертов, или краудсорсинг

Новые подходы

- NoSQL базы данных и СУБД которые не подразумевают внутренних связей между хранящимися данными, а основаны использование хеш функции «ключ-значение».

  СУБД Caché (произносится: «кашэ́») — иерархическая СУБД, позиционирующаяся мультимодельная, т.е. одновременный доступ к данным как «объекту» по значению и «реляционной сущности» - структуре данных как таблице.

В рамках концепции Data Science вычисления физический процесс, поэтому то надо знать ответы на вопросы:

- Какие ресурсы и законы лежит в основе формулы физической реальности :

  реальность= материя (вещество + энергия) + «данные»

- Являются ли «данные» атрибутом физической или ментальной реальности ?
- Что есть результат процесса обработки – новый физический процесс, новый объект, новая информация или новые данные ?
- Как, организуя «правильную» последовательность вычислений, «**быстро**» получить «**точный**» результата ?

Нужна «новая» информационная модель реальности:

Существуют явления, которые нельзя объяснить в рамках классической физики, которая изучает реальность, разделяя ее на части, силы и энергии. Это :

- поведение мелких рыбешек или птиц, образующих косяки или стаи;
- поведение толпы людей, например на концерте популярного ансамбля;
- явления предшествующие землетрясению: поведение животных, заряженные частицы в атмосфере, свечение воды, образование облаков и пр.;
- кооперативные явления в природе и обществе, самоорганизация, т.н. «фликкер-шум» и пр.;

и др.……

информационная «ЗАПУТАННОСТЬ»: синхронизация и когерентность «Больших Данных»

Случай 1: политическая «полигамия» - синхронизация на основе **общих интересов**

Случай 2: временная синхронизация на основе **общего визуального поля** с дирижёром

Между информацией и данными находится **«субстрат-посредник»**

Случай 3: когерентная «моногамия» на основе **общего звукового пространства**

Концепция науки о данных, которая изучает

Науки – наука о «данных»

Науки о материи и жизни

Математика и цифровые данные

Интеллект или субстрат-посредник

«большие данные»

ЗНАНИЕ / ПОНИМАНИЕ

↙ в форме цифровых моделей

↘ причин и процессов

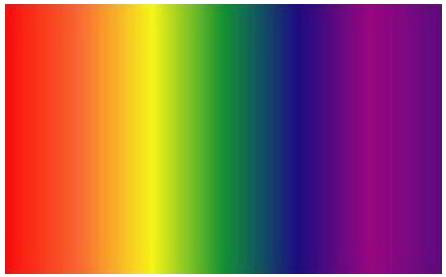
Интеллекта - посредник между миром процессов и миром данных

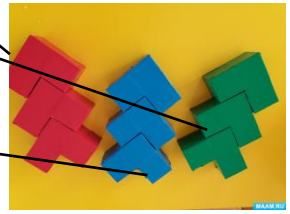
- Для интеллекта человека характерны две ключевые модальности - «знать и понимать»
  - **Знать** – это функция памяти, которая позволяет хранить данные. Эту функцию можно передать компьютерным системам.
  - **Понимать** – это функция выявления связей, интерпретации зависимостей и осознания причин.

- 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

5Когнитивный синтез «больших данных»

Инструменты работы с «большими данными» -





Есть много неструктурированных данных разных типов

Есть кластер –много связанных в сеть однотипных компьютеров

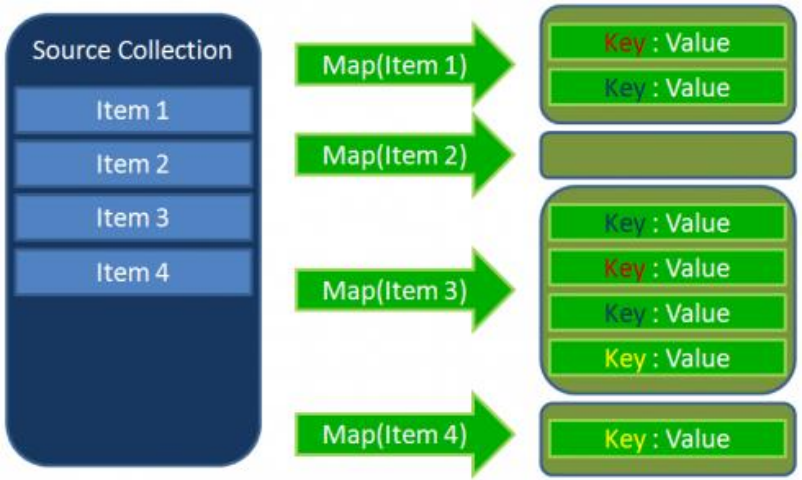
которые синхронизированы (упорядочены) по типам данных

Принцип: «подобное – подобным»

Как работать с «Большими данными» - модель MapReduce

- MapReduce — двух шаговая «Map» и «Reduce» модель вычислений, используемая для обработки **больших**, вплоть до нескольких петабайт, наборами данных, с помощью **большого** числа компьютеров, организованных в компьютерный кластер.

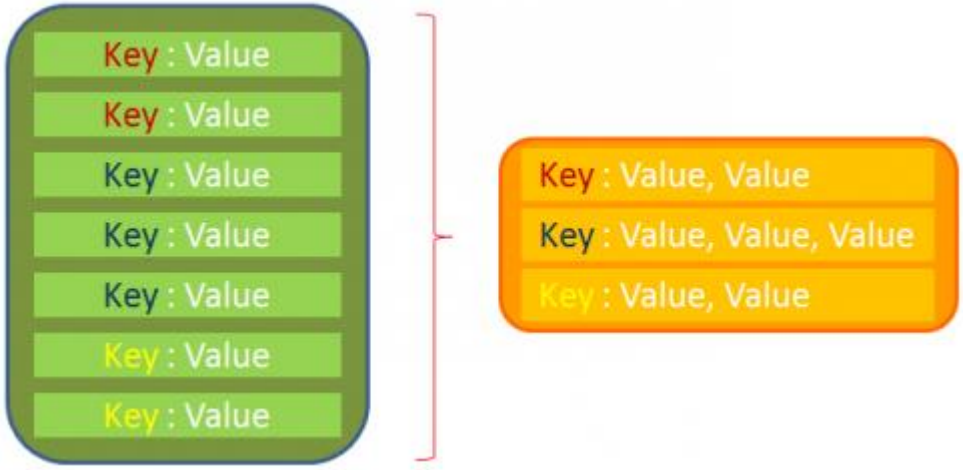
Исходный набор данных



Один из компьютеров кластера - master node получает входные данные, разделяет их на части и передает другим компьютерам (рабочим узлам — worker node) для последующей обработки

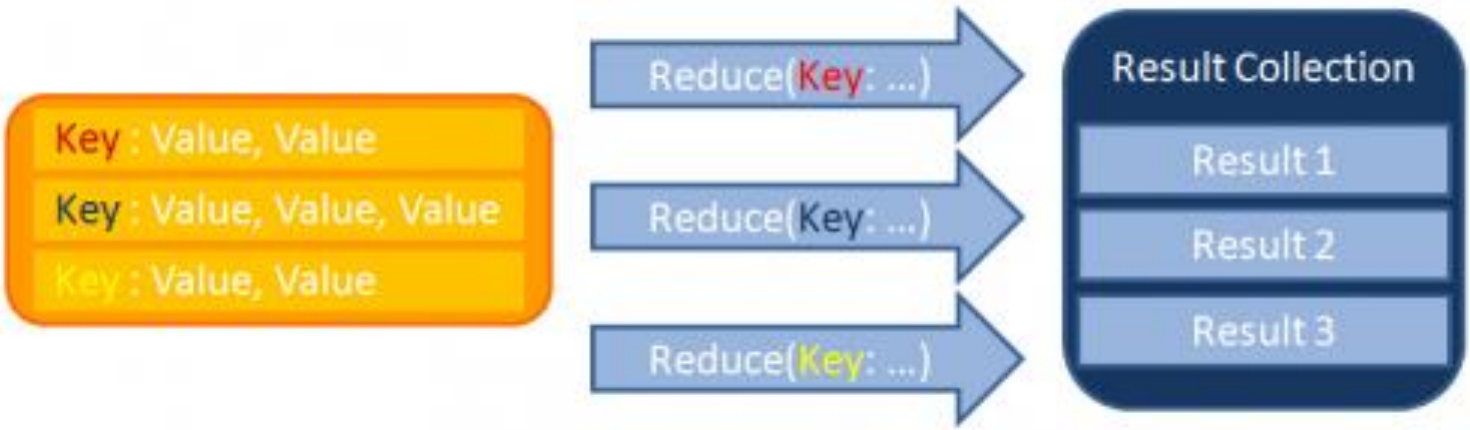
Шаг MAP: Классификация объектов по «ключам»

Создаются новые экземпляры объектов, где все значения (value) сгруппированы по «коду» или по ключу.



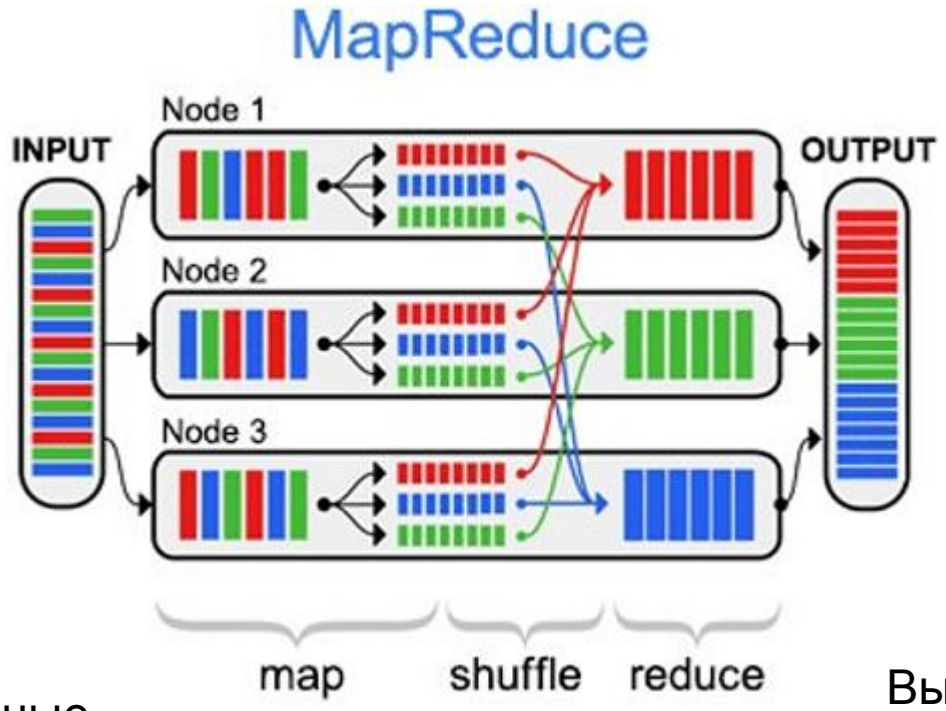
Шаг Reduce

Возвращает новый экземпляр объекта, который включен в результирующую коллекцию



На Reduce-шаге происходит свёртка предварительно обработанных данных — для этого необходимо, чтобы все результаты предварительной обработки с одним конкретным значением ключа обрабатывались одним рабочим узлом кластера в один момент времени.
(аспекты когерентности и синхронизации)

Итак, модель распределённых вычислений над очень большими наборами данных с использованием компьютерных кластеров :

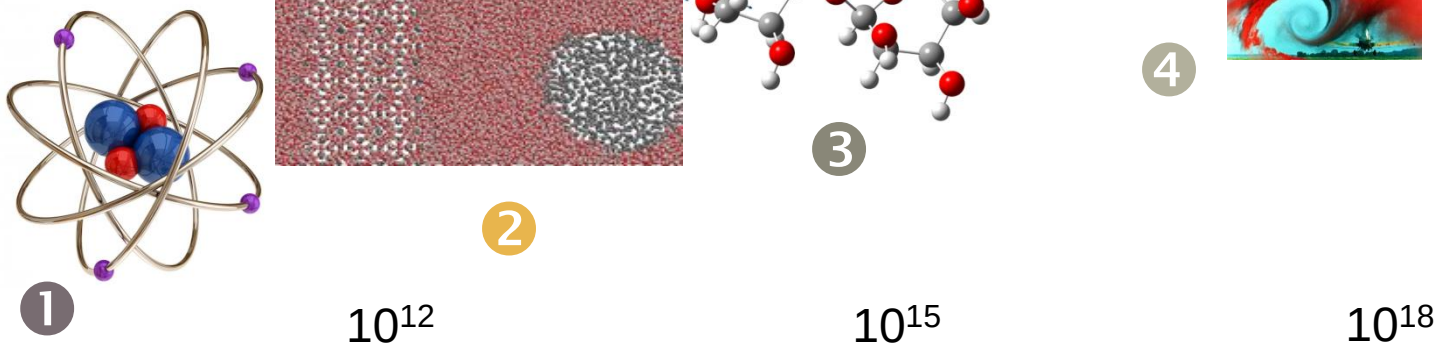


Входные неструктурированные данные

Выходные данные структурированные по ключам (цветам)

«Размер» имеет значение – структуры физических объектов имеют разный «вид» на разных масштабах

1. взаимодействие свободных квантовых объектов и атомов
2. упорядоченные атомарные структуры
3. молекулярные структуры
4. макро процессы, материалы и конструкции



Характеристики компьютера

10^{12} 10^{15} 10^{18}

Число машинных операций

Характеристики объекта

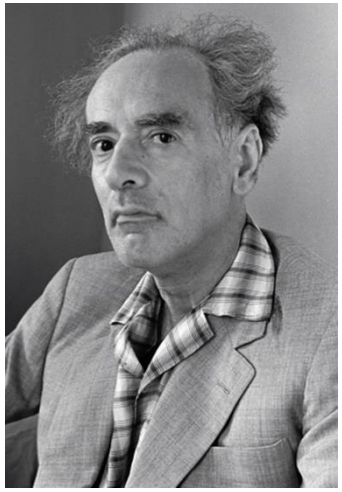
10^{1} 10^{2} 10^{3} 10^{4} 10^{5} 10^{6}

Число вычислительных ядер в компьютере

10^{2} 10^{3} 10^{4} 10^{5} 10^{6} 10^{7} 10^{8} 10^{9}

Число «компонент», учитываемых в модели объекта

В физике определен «теоретический минимум Ландау»



- 11 экзаменов по основным разделам математики и теоретической физики. Теоретический минимум включает в себя следующие экзамены: Математика-I, Механика, Теория поля, Математика-II, Квантовая механика, Квантовая электродинамика, Статистическая физика-I, Механика сплошных сред, Электродинамика сплошных сред, Статистическая физика-II и Физическая кинетика.

Должен быть и теоретический минимум, характеризующих суть «Компьютерных наук»

- **компьютерная математика** (*computer math*)
- **проектирование**, вычислительная техника (*computer engineering*)
- **программирование** (*computer programming*)
- **искусственный интеллект** (*artificial intelligence*)
- **робототехника** (*robotics*)

Вариант 1: Теоретические основы КН

- **Теория информации** (**information theory**) – научная дисциплина, основанная на понятии количество информации, которое может быть передано от передатчика к приёмнику с учетом влияния шума (*noise level*) и других характеристик среды передачи;
- **Теория автоматов** (automata theory) – научная дисциплина, изучающая абстрактные вычислительных устройств, или “машины вычислений”.
- **Теория алгоритмов** (theory of algorithms**)** – математическая дисциплина, изучающая алгоритмы и их общие свойства.
- **Теория «машинного обучения»** - научная дисциплина, изучающая возможности организации вычислений на основе «феномена обучения» и самоконфигурации, без задания в явном виде алгоритма в форме текста программы.
- Математика «больших данных» и квантовых компьютеров

а также: численный анализ, теория графов, теория вероятности, эволюционные вычисления, генетические алгоритмы, исследование операций, теория систем….

Заключение

- В основе науке о данных или DataScience лежит формула :

  реальность= материя (вещество + энергия) + «данные»

- При этом «данные» рассматриваются не только «по значению», но и с учетом таких характеристик как «**Объем** Volume», «**Скорость** изменения или Velocity» и «**Разнообразия** или Variety»

Задание-вопрос: что происходит с данными в этой программе построенной на основе модели MapReduce

```
// Функция, используемая рабочими нодами на Map-шаге
// для обработки пар ключ-значение из входного потока
void map(String name, String document):
    // Входные данные:
    //   name - название документа
    //   document - содержимое документа
    for each word w in document:
        EmitIntermediate(w, "1");

// Функция, используемая рабочими нодами на Reduce-шаге
// для обработки пар ключ-значение, полученных на Map-шаге
void reduce(String word, Iterator partialCounts):
    // Входные данные:
    //   word - слово
    //   partialCounts - список группированных промежуточных результатов. Количество записей в partialCounts и есть
    //     требуемое значение
    int result = 0;
    for each v in partialCounts:
        result += parseInt(v);
    Emit(AsString(result));
```