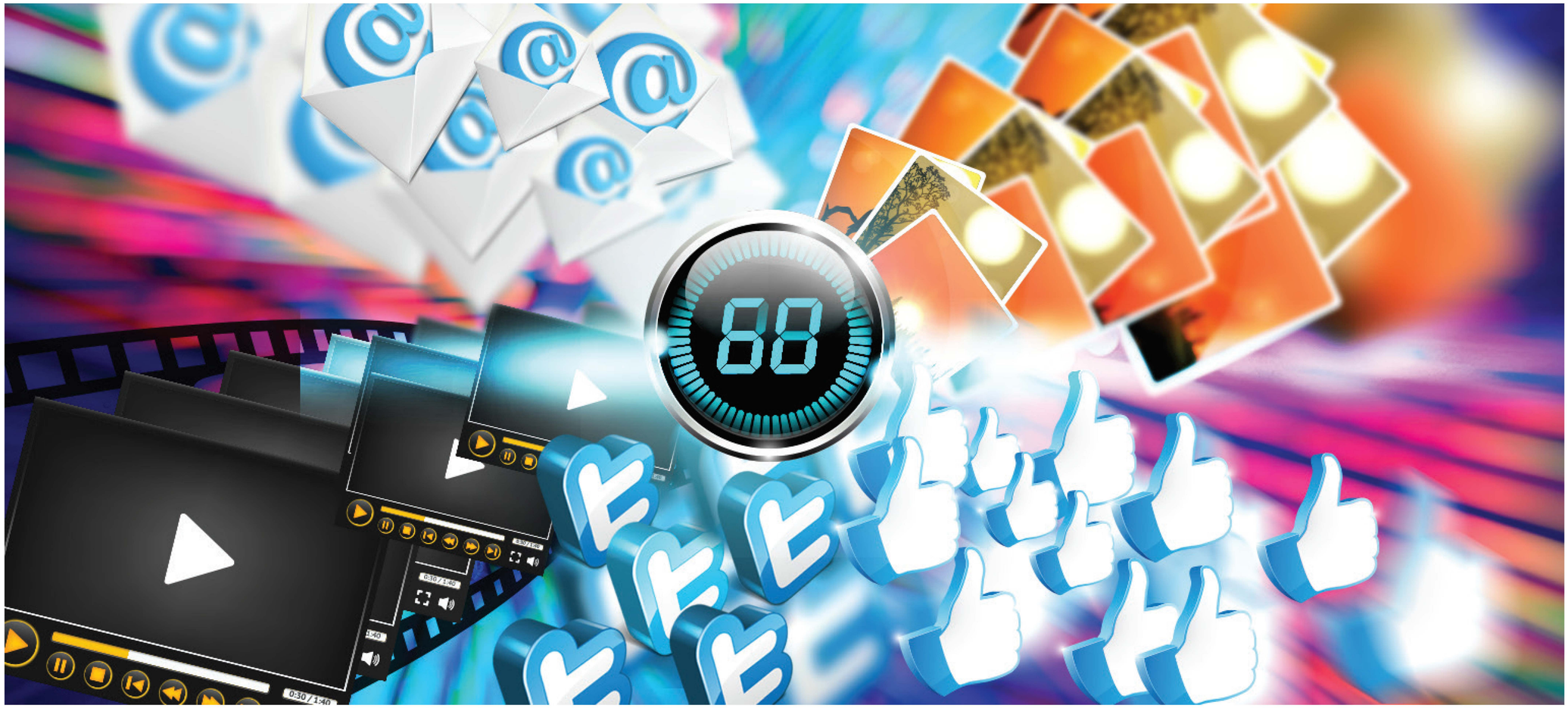
88

Причины появления больших данных (Big Data)

- Значительное увеличение объемов данных по сравнению с возможностью их обработки.
- Сейчас объемы данных превышают объемы доступных носителей для хранения информации.

Признаки больших данных (Big Data)

СПбГУ

- **Volume.** Действительно большие.
- **Variety.** Слабо структурированные и очень разнородные.
- **Velocity.** Требуют быстрой обработки.
- **Value.** Позволяют получить значимые результаты анализа данных.

Классификация объемов данных

СПбГУ

- **Большие наборы данных:**
  от тысячи мегабайт до сотен гигабайт.
- **Огромные наборы данных:**
  от тысячи гигабайт до нескольких терабайт.
- **Big Data:**
  от нескольких терабайт до сотен терабайт.
- **Extremely Big Data:**
  от тысячи терабайт.

Отправитель

Электронное письмо

Получатель

Разнородность структур данных

СПбГУ

Структурированная информация:

- таблицы с известными типами данных.

Полуструктурированная информация:

- XML-документы и XSD-схемы.

Неструктурированная информация:

- текстовые документы;
- видеоконтент;
- аудиоконтент.

Оперативная обработка данных

СПбГУ

- Применение инструментов и технологий для распараллеливания вычислений (Map-Reduce+Hadoop).
- Использование методов приближенной обработки данных.
- Использование NoSQL DB.

Технология Map-Reduce

СПбГУ

- Технология Map-Reduce — модель для распределенных вычислений.
- Принцип работы:
  - распределение входных данных на рабочие узлы распределенной системы для предварительной обработки (Map);
  - объединение уже предварительно обработанных данных (Reduce).

Hadoop

СПбГУ

- Технология Map-Reduce используется в инструменте Hadoop.
- Данные хранятся в нескольких копиях и распределяются по тысячам узлов.
- Система практически неограниченно масштабируется.
- Каждый узел является сервером и хранения, и обработки
- Обработка данных ведется в массивно-параллельном режиме
- Данные хранятся в нескольких копиях и отказ узла не ведет к потере данных
- Система практически неограниченно масштабируется

Приближенная обработка данных

СПбГУ

- Ограничение во времени обусловливает применение приближенных методов обработки больших данных.
- Для некоторых задач точное выполнение не имеет смысла.
- Используется метод случайных выборок относительно небольшого объема.
- Результаты, полученные только на основе выборки, будут приближенными, их точность зависит от размеров выборки.

Internet of Things - примеры

- авиалайнеры;
- автомобили;
- одежда;
- бытовые приборы;
- медицинские инструменты;
- парковки и т. д.

Internet of Things - новые сферы применения

СПбГУ

Интернет стал проникать в ранее недоступные сферы.

- Пациенты начинают глотать интернет-устройства, позволяющие диагностировать некоторые заболевания и выявлять их причины.

- Миниатюрные интернет-датчики закрепляют на животных, растениях и геологических объектах.

Internet of Things в контексте аналитической обработки

СПбГУ

- Интернет вещей радикальным образом увеличивает объемы данных.
- Чем больше генерируется данных, тем больший объем знаний можно получить в итоге.
- На основе большего объема знаний можно принимать решения о дальнейшем поведении системы в целом.

Данные

СПбГУ

- Данные — это сырой материал, который может превратиться в полезную информацию в контексте определенной задачи.
- Чем больше объем накопленных данных, тем больше можно выявить закономерностей и тенденций развития.

Обработка данных

СПбГУ

- Агрегированная информация используется для получения знаний не более чем на 10%.
- Аналитическая обработка данных, как правило, осуществляется в рамках отдельной локальной задачи.