Социальные медиа

Большие Данные.
Цели, задачи и парадигма анализа.

Макаренко А.В.

avm.science@mail.ru

Научно-исследовательская группа
«Конструктивная Кибернетика»
www.rdcn.ru

Институт проблем Управления
Российская академия наук
Москва, Россия

25 ноября 2015 г.

❑ План мини-курса «Социальные медиа»

Мини-курс «Социальные медиа»:

Лекция x.1. Большие Данные. Цели, задачи и парадигма анализа.

Лекция x.2. Дескриптивный и структурный анализ. Подходы и Методы.

Лекция x.3. Прикладные задачи анализа. Проблематика.

❑ План I лекции

1. Введение

2. Социальные медиа – цели и задачи анализа

3. Формальная схема площадок социальных медиа с точки зрения их анализа

4. Большие Данные – модный тренд или закономерная технология?

5. Переход: социальные медиа → Большие Данные → Большие Вычисления.

6. Структурирование и первичная формализация данных. Граф обработки.

7. Инструменты анализа социальных медиа.

8. Задания для самостоятельной проработки.

❑ Введение (объект и предмет мини-курса лекций)

Социальные медиа – это набор онлайновых Интернет технологий, которые позволяют пользователям общаться между собой.

Общение в социальных медиа принимает самые различные формы обменов:

- новостями, информацией, видео, фото, музыкой, линкам
- мнениями, опытом, знаниями, контактами, связями

Оперирование контентом происходит на идеологической и технологической базе *Web 2.0* (контент созданный и поддержанный пользователем).

Требуются эффективные подходы, методы и методики анализа:

- структур формирующихся
- процессов протекающих

внутри социальных медиа в привязке к реальной среде.

❑ Введение (о лекторе, общая информация)

- Основатель и руководитель научно исследовательской группы «Конструктивная Кибернетика»
- Старший научный сотрудник Института Проблем Управления РАН
- Кандидат технических наук по специальности «Системный анализ, управление и обработка данных»
- IEEE Signal Processing Society Membership
- IEEE Computational Intelligence Society Membership
- Сертифицированный эксперт-инструктор Wolfram Research Inc.

avm@rdcn.ru www.rdcn.ru

❑ Введение (о лекторе, научные интересы)

- анализ структуры сложных динамических процессов, предсказуемость
- обнаружение, классификация и диагностика не вполне наблюдаемых объектов (паттернов)
- синхронизация и самоорганизация в нелинейных и хаотических системах
- моделирование экономических, финансовых, социальных и биофизических систем и процессов
- исследования в области конвергенции Data Science, Big Data, Machine Learning, Nonlinear Dynamic и Network-Centric

❑ Введение (о лекторе, инженерные навыки)

- Data Mining, DSP, HPC
- Языки программирования: C/C++, R, Python, Wolfram Language
- Технологии: Cilk Plus, OpenMP, MPI, CUDA
- Библиотеки: Intel MKL, Intel MPP, …

❑ Социальные медиа – цели и задачи анализа

Базовые цели анализа:

1. Улучшение собственно социальных медиа;

2. Опосредованная эмпирическая социология;

3. Таргетированная реклама и маркетинг;

4. Информационная разведка (модель OSINT – open source intelligence);

….

❑ Социальные медиа – цели и задачи анализа

Базовый потенциал анализа:

1. Опосредованное проведение эмпирических социологических исследований (минимизация ошибки смещения мнения, вскрытие «внутреннего» мнения).
2. Скрытый характер проводимых исследований.
3. Глобальный характер данных, представленных в Web.
4. Возможность проведения многократных опросов, итеративно повышающих точность.
5. Возможность изучения табуированных тем.
6. Возможность выявления скрытых (маскируемых) тенденций.
7. Оперативность проведения исследований.
8. Сравнительно низкая стоимость исследований.
9. Возможность реализации сигнальных систем, функционирующих «на лету».

❑ Социальные медиа – цели и задачи анализа

Разнообразие информационных объектов социальных медиа

Сеть:

– livejournal.com
– facebook.com
– twitter.com
– youtube.com
– github.com
– Spoken Web (IBM)
– …

Сообщение:

– Текст
– Картинка
– Видео
– Звук

Инфо-блок:

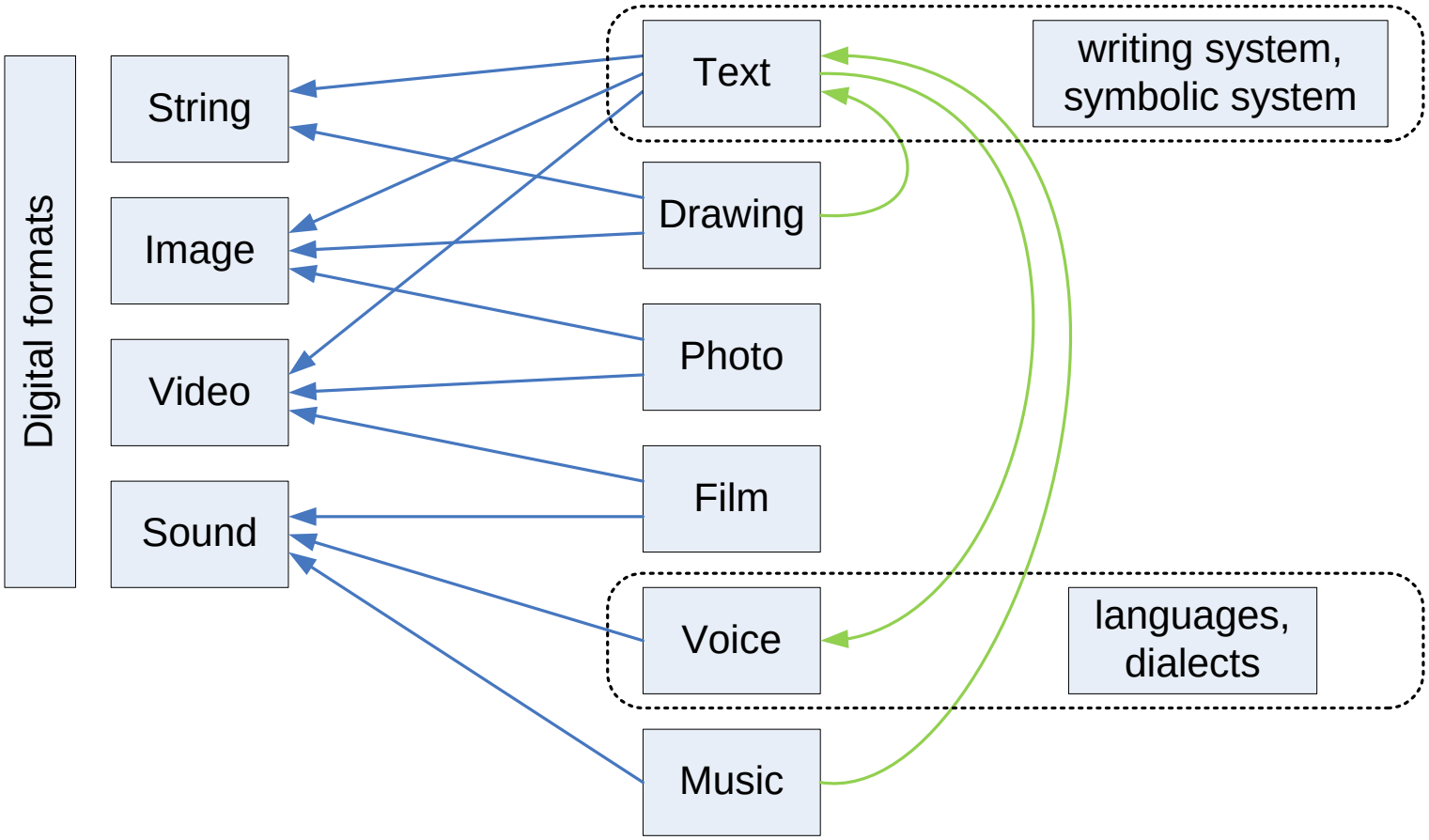
– Текст
– Чертёж
– Фото
– Фильм
– Голос
– Музыка

Смысл:

C.K. Ogden, I.A. Richards, The Meaning of Meaning, New York, 1964.

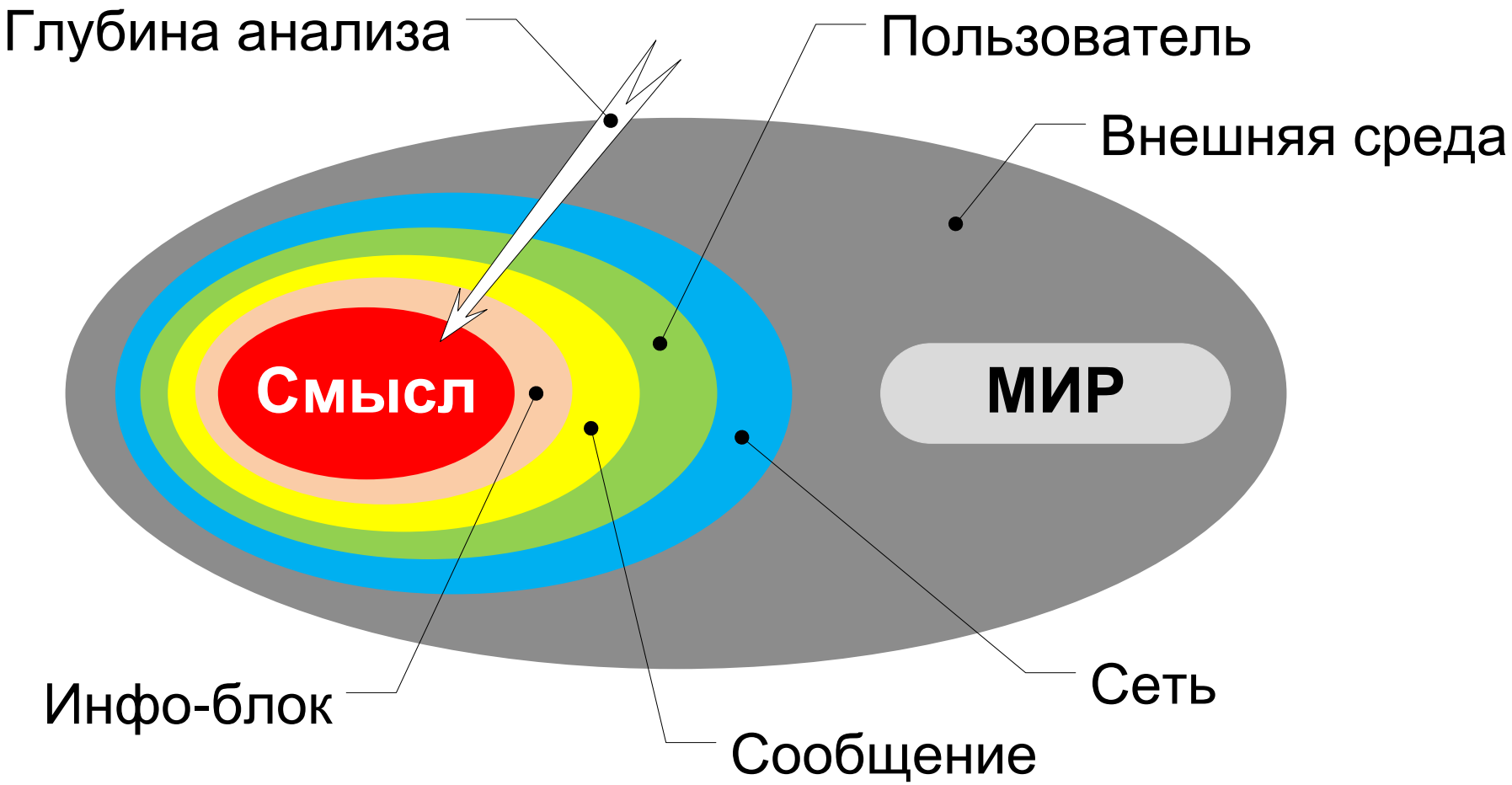
❑ Социальные медиа – цели и задачи анализа

Отношение между категориями и типом содержимого информационных объектов социальных медиа



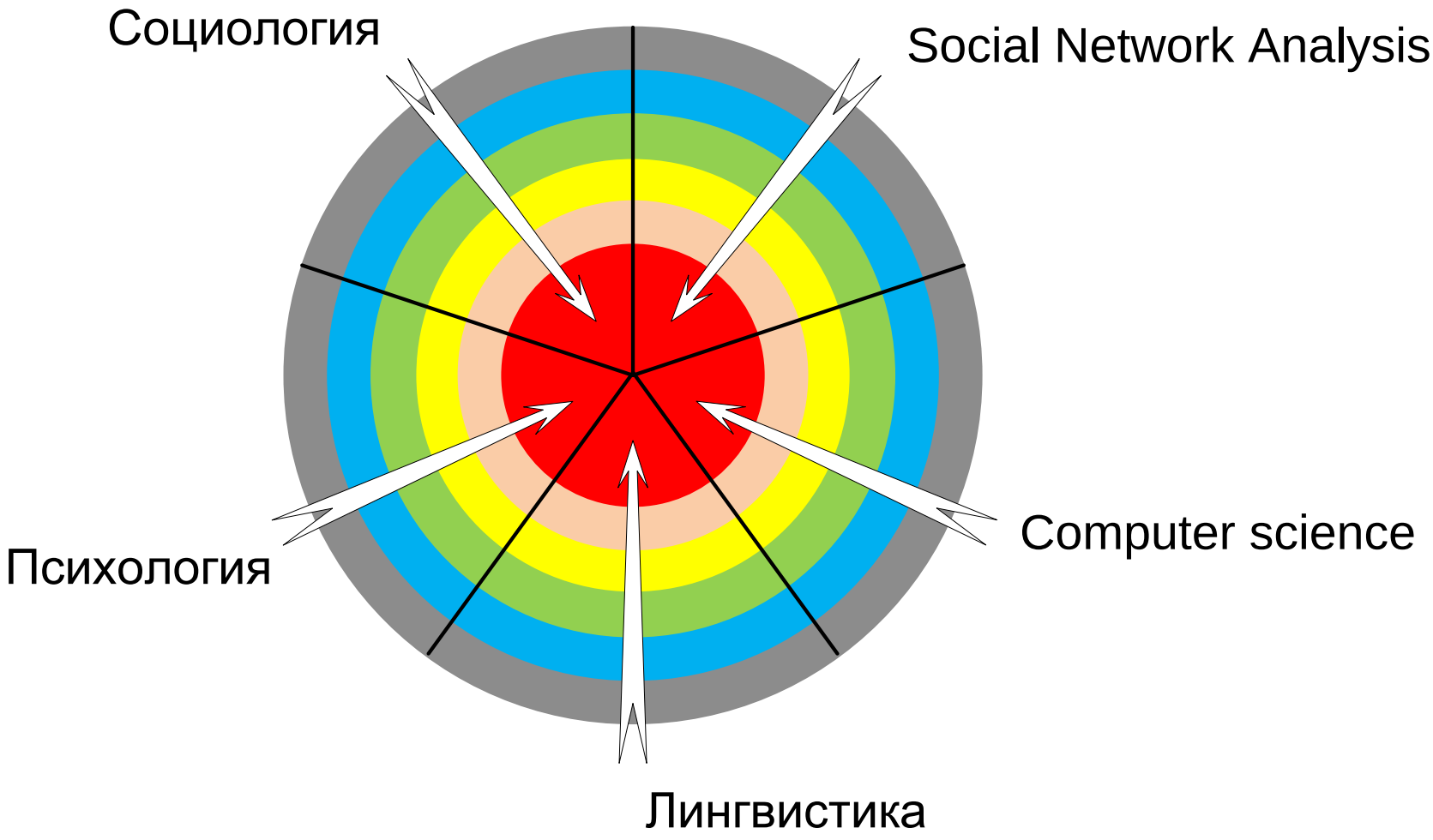
❑ Социальные медиа – цели и задачи анализа

Насколько сложно анализировать социальные медиа?



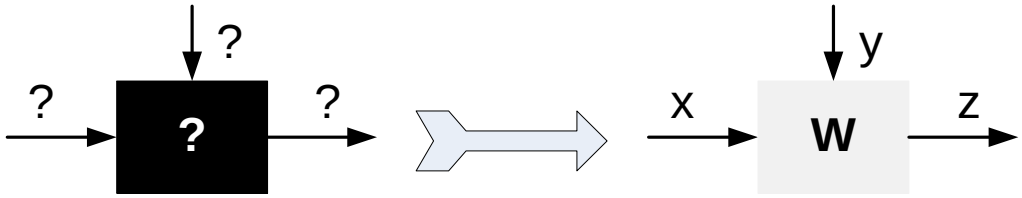
❑ Социальные медиа – цели и задачи анализа

Насколько сложно анализировать социальные медиа?

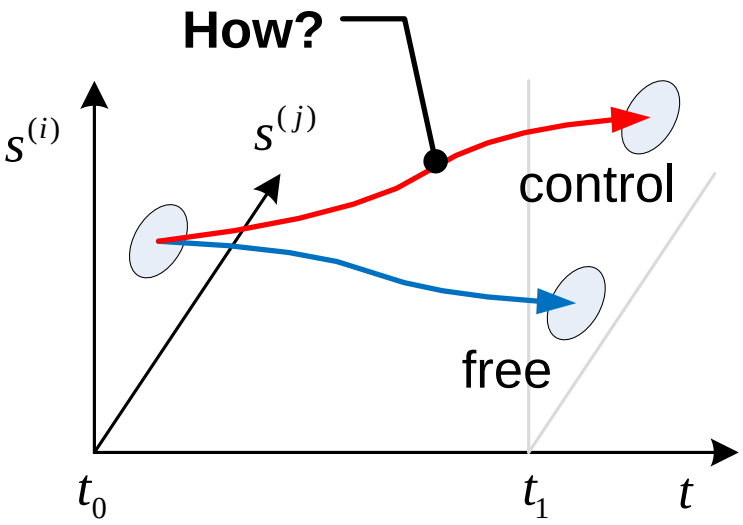


❑ Социальные медиа – цели и задачи анализа

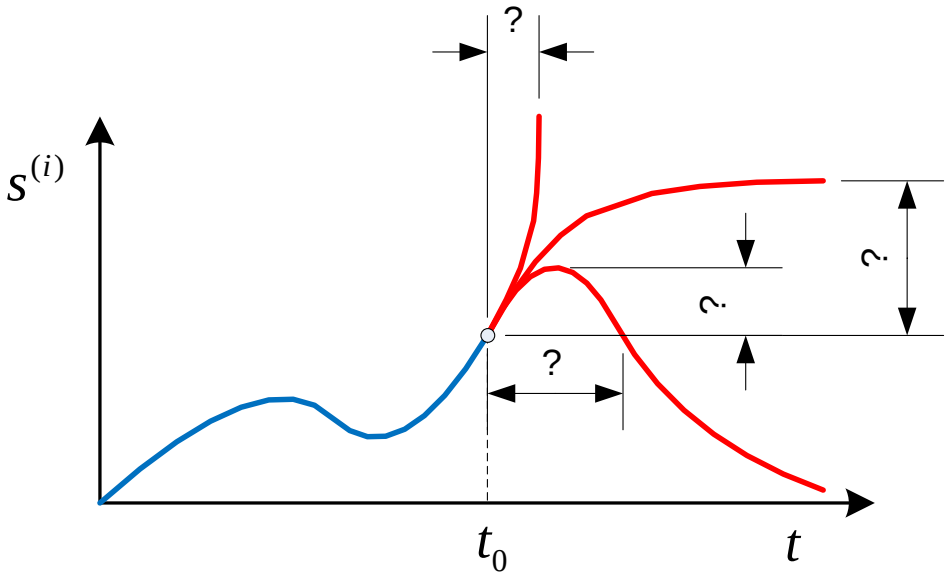
Формальные задачи анализа и математического моделирования



Идентификация



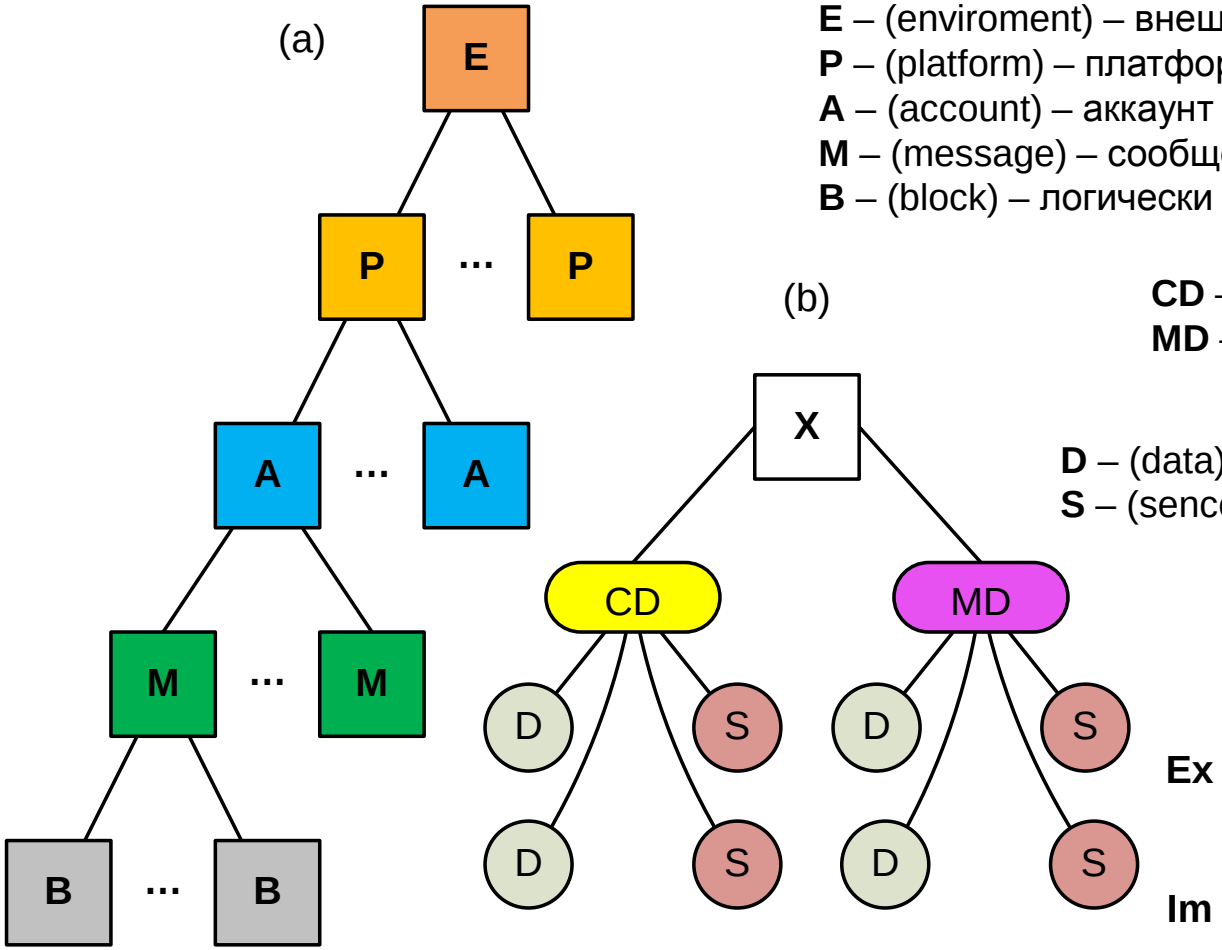
Управление



Предсказание

❑ Формальная схема площадок социальных медиа с точки зрения их анализа

Универсальная декомпозиция социальных медиа (объекты-контейнеры)



E – (enviroment) – внешняя среда;
P – (platform) – платформа отдельно взятого ресурса;
A – (account) – аккаунт пользователя;
M – (message) – сообщение пользователя;
B – (block) – логически целостная часть сообщения.

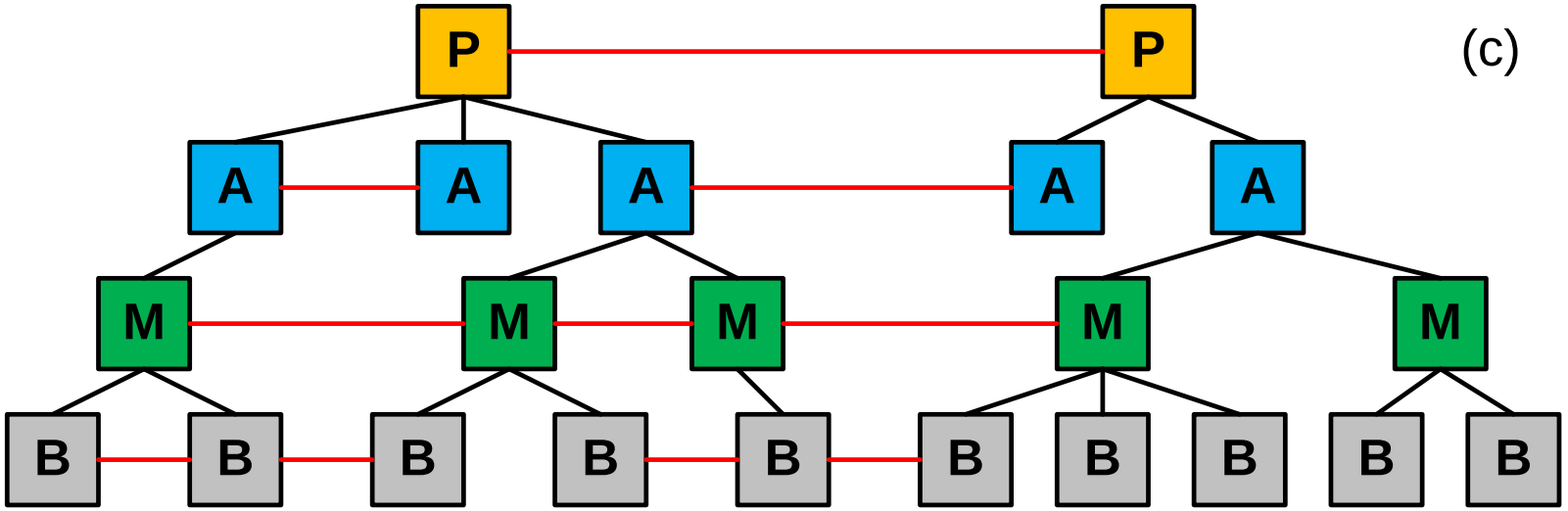
CD – (contents) – содержание;
MD – (metadata) – метаданные.

D – (data) – факты, сведения, показатели;
S – (sence) – смысл, значение, мнение.

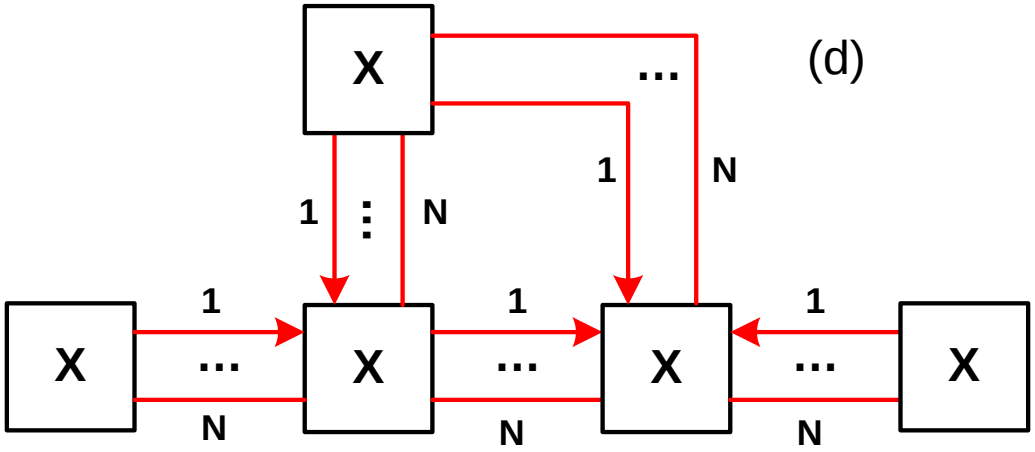
Ex – (explicit) – явно заданный (на уровне формальной семантической разметки);
Im – (implicit) – неявно заданный (выявляется с тем, или иным уровнем доверия).

❑ Формальная схема площадок социальных медиа с точки зрения их анализа

Универсальная декомпозиция социальных медиа (объекты и отношения)



(c)



(d)

Базовые отношения:

– Вхождение;

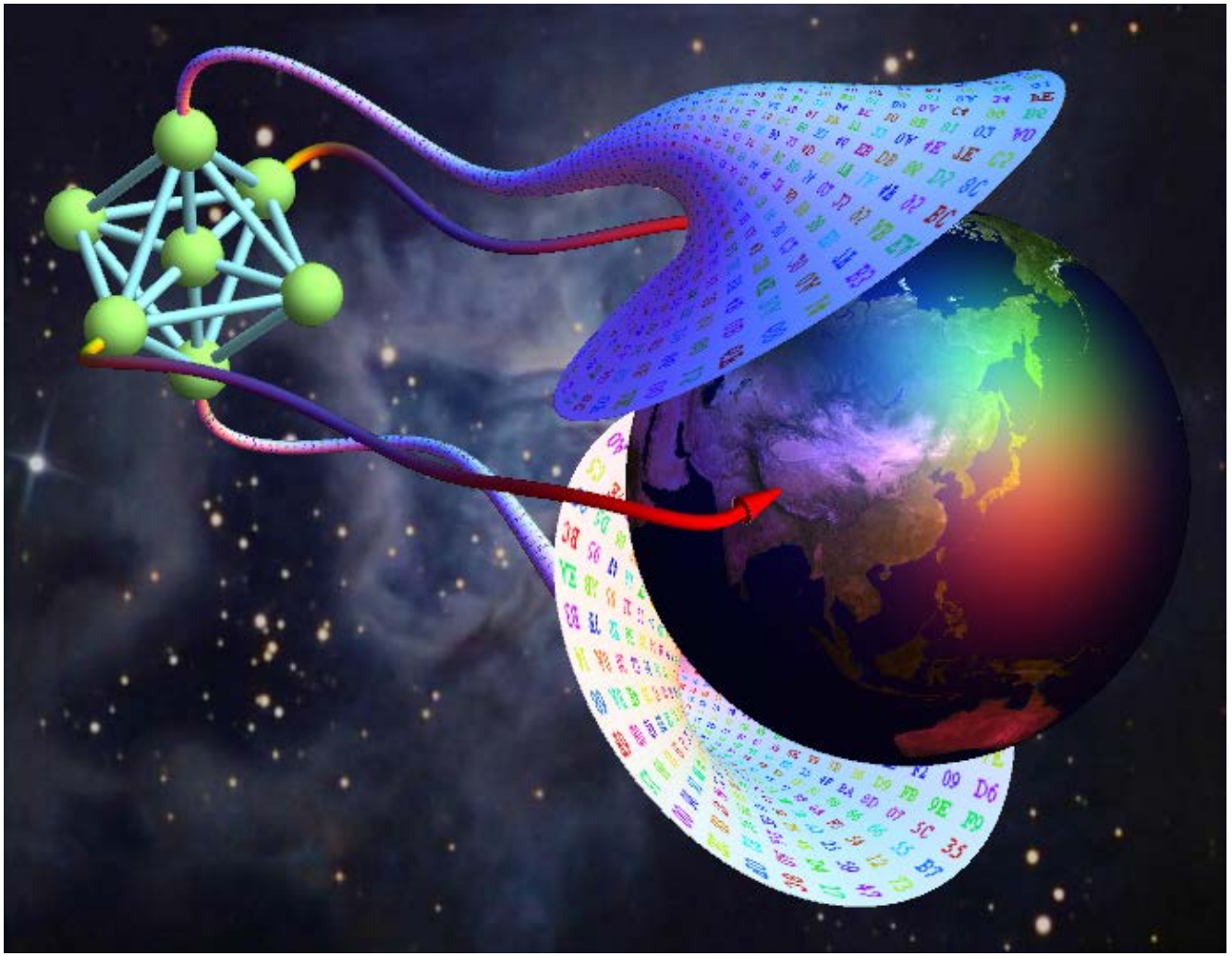
– Комментирование;

– Цитирование;

– Copy/Paste.

❑ Большие Данные – модный тренд или закономерная технология?

Переход от Network-Centric к Big Data – яркий пример смены лидера технологий



www.rdcn.ru/estimation/2014/20140816.shtml

❑ Большие Данные – модный тренд или закономерная технология?

Что такое Big Data?

Большие Данные – это набор стратегий, методов и технологий связанных со сбором, хранением и обработкой наборов данных, отвечающих следующим условиям:

- Большой объём данных
- Высокая скорость поступления данных
- Существенная неструктурированность и гетерогенность поступающих данных

Важный аспект обработки Больших Данных – это глубина анализа данных, позволяющая осмыслить и понять суть явления.

❑ Большие Данные – модный тренд или закономерная технология?

Варианты проведения анализа социальных медиа:

- **Ручная обработка экспертом малых и средних выборок данных**

Ограничения:
- Низкая производительность и повторяемость
- Высокий уровень субъективизма
- Низкая обнаружительная способность

- **Автоматизированная обработка средних и больших выборок данных**

Ограничения:
- Существенные затраты на развёртывание и эксплуатацию системы
- Существенные затраты на разработку аналитических инструментов

- **Тотальный автоматический скрининг доступных наборов данных**

Ограничения:
- Высокие затраты на развёртывание и эксплуатацию системы
- Высокие временные затраты на проведение анализа
- Высокий риск получения результата: «гора родила мышь»

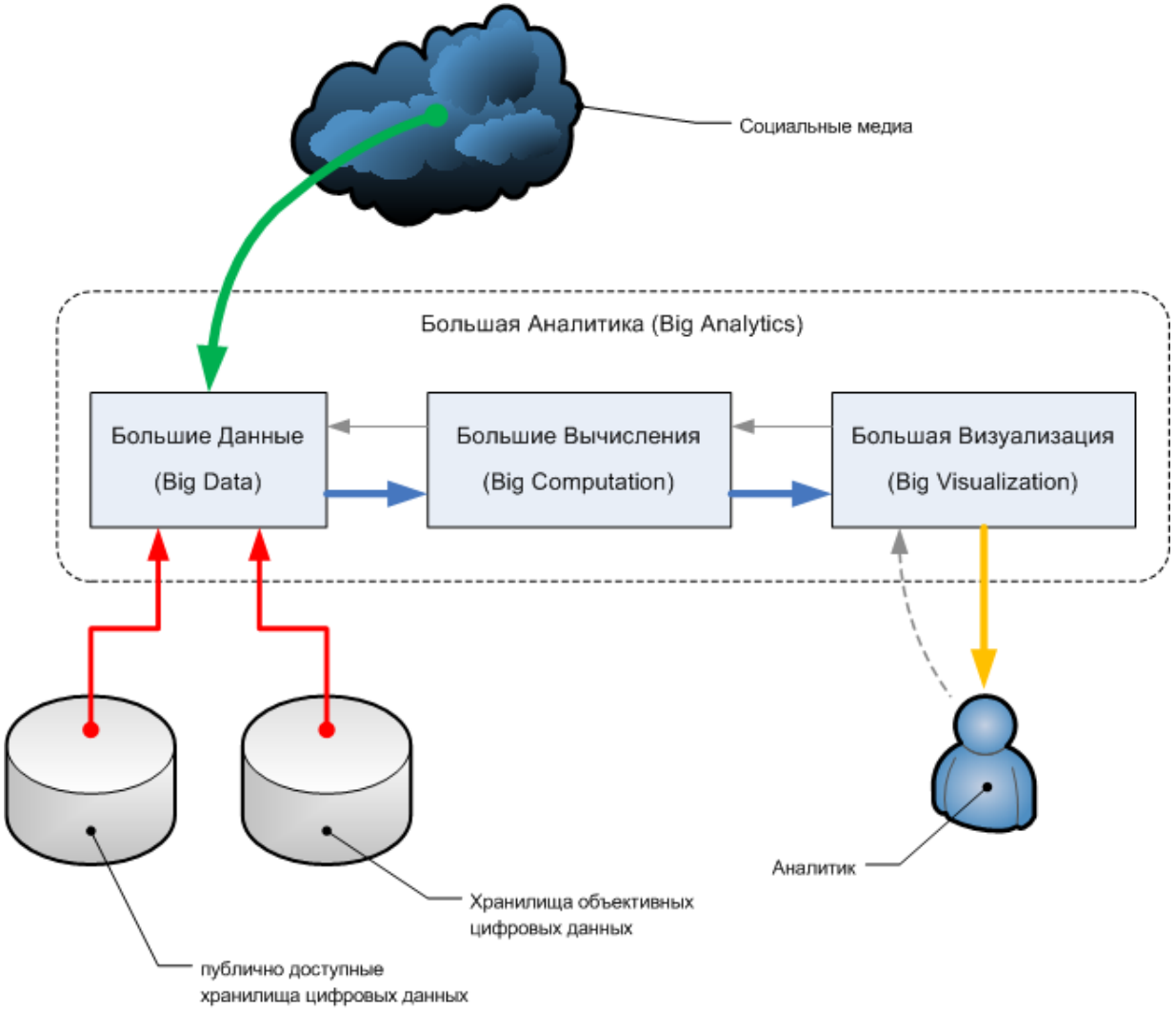
❑ Большие Данные – модный тренд или закономерная технология?

Особенности проведения Data Mining социальных медиа:

1. Гигантские объёмы цифровых данных (десятки и сотни терабайт) и высокий темп поступления новых данных (миллионы и десятки миллионов документов в сутки);
2. Слабая структурированность исходных данных (служебная информация перемешана с содержательной, отсутствие семантической идентификации блоков данных) и высокая разнородность данных (различные способы и форматы представления информации);
3. Новизна феномена социальных медиа, его слабая изученность с позиций психологии и социологии, слабая формализуемость и изученность данной предметной области;
4. Особые требования к алгоритмам и программному обеспечению: производительность (массовая параллельность, алгоритмы типа log N, N, N log N), устойчивость (отсутствие рекурсивных вызовов, вспомогательная память не более N) и организация (асинхронная среда с распределённой памятью, минимизация дисковых и коммуникационных операций).

❑ Переход: социальные медиа → Большие Данные → Большие Вычисления

Иерархия технологических систем анализа социальных медиа



❑ **Переход: социальные медиа → Большие Данные → Большие Вычисления**

Проблематика построения аналитических систем социальных медиа

Основные классы задач требующих решения:

1. Сбор, структурирование и хранение первичных разнородных данных;
2. Выявление паттернов, идентификация структурных и динамических свойств информационных процессов, онлайн сообществ, социальных групп и слоёв;
3. Понижение размерности данных и когнитивная визуализация;
4. Модели предметной области для возможностей изучения систем, явлений и процессов, прогнозирования и управления.

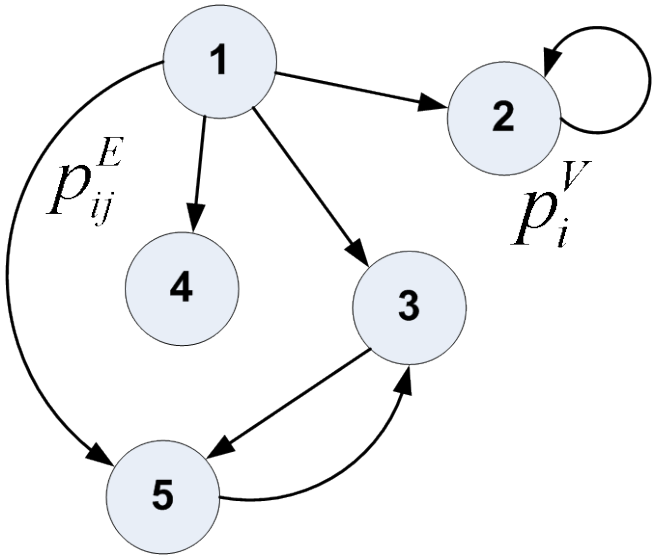
Основные направления задач требующих решения:

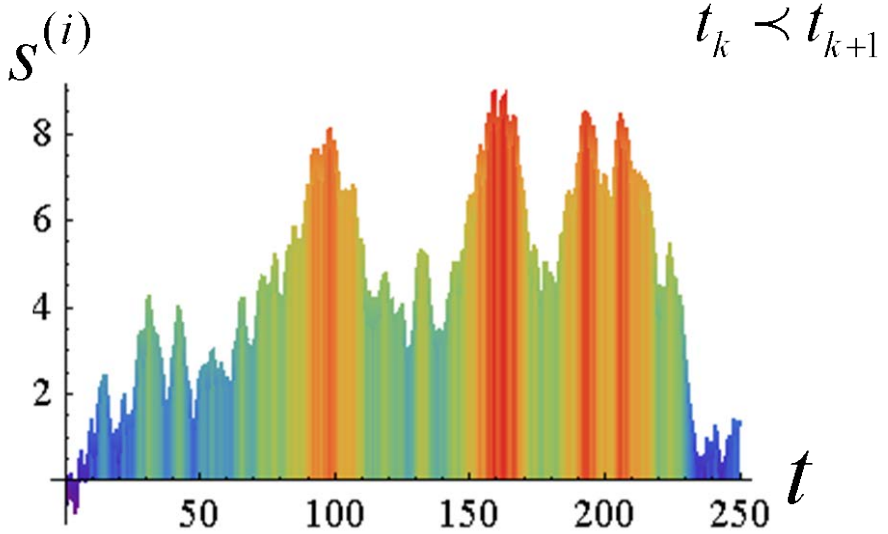
1. высоко-производительные вычислительные системы;
2. специализированные математические методы;
3. специализированное алгоритмическое и программное обеспечение.

❑ Структурирование и первичная формализация данных. Граф обработки.

Математические структуры для формального представления данных

$$\Gamma = \langle \mathrm{V}, \mathrm{E} \rangle \qquad \mathrm{R} = \left\{ \{\mathbf{s}_k\}_{k=1}^{K}, \{t_k\}_{k=1}^{K} \right\}, \mathbf{s} \in \mathrm{S} \subset \mathbb{R}^N, t \in \mathrm{T} \subset \mathbb{R},$$





$$\mathrm{M} = \{A_1, A_2, A_1, A_3, \ldots\}$$

$$\Gamma^t = \langle \mathrm{V}, \mathrm{E} \rangle \,|\, t \qquad \mathrm{M}^t = \{A_1, A_2, A_1, A_3, \ldots\} \,|\, t$$

❑ Структурирование и первичная формализация данных. Граф обработки.

Схема представления данных сети в формальном виде (базовая конструкция)

$$\texttt{Net} \Rightarrow \texttt{Account} \Rightarrow \texttt{Message} \Rightarrow \texttt{Block}$$

Net: $n_i \in \mathrm{N} \subset \mathbb{N}$

Account: $g_j \in \mathrm{G} \,|\, n_i \subset \mathbb{N}$

Message: $s_k \in \mathrm{S} \,|\, (g_j \,|\, n_i) \subset \mathbb{N}$

Block: $c_l \in C \,|\, \left(s_k \,|\, (g_j \,|\, n_i)\right) \subset \mathbb{N}$

t_m^a – время порождения объекта, *важнейший элемент* метаданных.

❑ Структурирование и первичная формализация данных. Граф обработки.

Схема программного представления данных сети (базовая конструкция)

(a)

#	ID of message	ID of messages's author	Time publication of message
1	MessageID (unsigned INT64)	UserAccountID (unsigned INT64)	TimeMessageCreate (Unix Time)
...	...	...	...
N	MessageID (unsigned INT64)	UserAccountID (unsigned INT64)	TimeMessageCreate (Unix Time)

(b)

#	ID of messages's author	URL of mesage's author
1	UserAccountID (unsigned INT64)	UserAccountURL (string)
...	...	...
N	UserAccountID (unsigned INT64)	UserAccountURL (string)

(c)

#	ID of message	URL of message
1	MessageID (unsigned INT64)	MessageURL (string)
...	...	...
N	MessageID (unsigned INT64)	MessageURL (string)

❑ Структурирование и первичная формализация данных. Граф обработки.

Построение графа обработки (базовые положения)

Граф обработки – набор логически связанных вычислительных процедур, которые должны быть применены к каждому объекту (набору) данных с целью решения задач анализа и достижения поставленных целей исследования.

$\Gamma = \langle V, E \rangle$ V – процедуры E – данные

Два полюса вычислительных процедур:

Изолированная по данным – для выполнения не требуется доступ к данным смежных объектов (размер документа, частота слов в документе, и т.п.) – хорошо реализуется через модель MapReduce, удобна для распараллеливания и конвейеризации.

Полносвязная по данным – для выполнения требуется доступ к данным всех объектов выборки (поиск цитирований, построение компонент связности графа комментирования, и т.п.) – требует моделей матрично-графовых вычислений, сложно распараллеливается, является высоконагруженной вычислительной задачей.

❑ Инструменты анализа социальных медиа

- **Промышленные системы**

Масштабируемость, Стабильность, Производительность, Драйвера

  - Hadoop
  - Spark
  - …

- **Прототипирование**

Производительность, Библиотеки, Драйвера, Визуализация

  - Python
  - R
  - …

- **Аналитика**

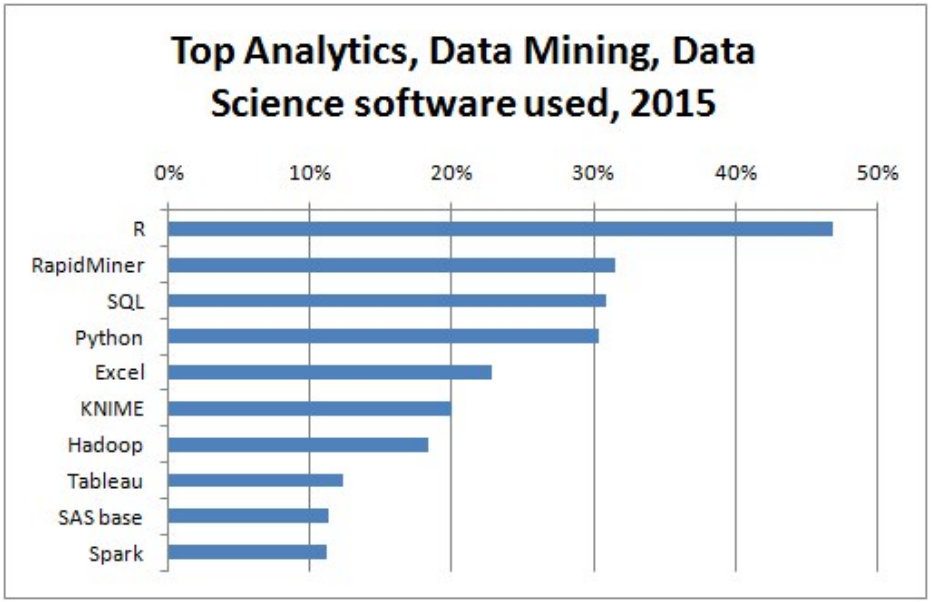
Алгоритмы, Драйвера, Визуализация, Производительность

  - R
  - Python
  - Weka, RapidMiner
  - Wolfram Mathematica, MatLAB, SPSS, SAS
  - …

❑ Инструменты анализа социальных медиа

Аналитика и язык R

Language Rank	Types	Spectrum Ranking
1. Java		100.0
2. C		99.9
3. C++		99.4
4. Python		96.5
5. C#		91.3
6. R		84.8
7. PHP		84.5
8. JavaScript		83.0
9. Ruby		76.2
10. Matlab		72.4



R #6 in IEEE 2015 Top Programming Languages, Rising 3 Places

http://blog.revolutionanalytics.com/2015/07/ieee-2015-rankings.html

R leads RapidMiner, Python catches up, Big Data tools grow, Spark ignites

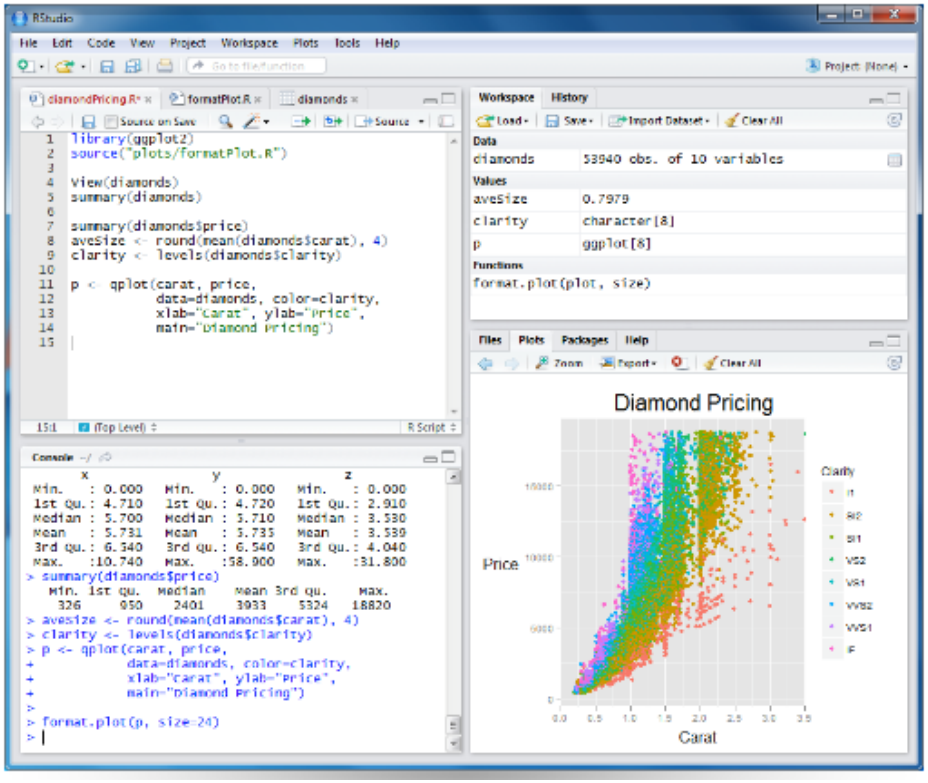
http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html

Язык R – является языком предметной области (DSL) для обработки данных, реализован поверх классической архитектуры интерпретатора-компилятора Scheme.

Скачать: https://cran.r-project.org/, актуальная версия: 3.2.2.

❑ Инструменты анализа социальных медиа

Среда программирования RStudio



– Бесплатна;

– Мультиплатформенна;

– Автодополнение кода;

– Навигация и формат кода;

– Подсветка кода (слабовато!);

– Работа с проектами;

– Исполнение и отладка кода;

– Доступ к R консоли;

– Доступ к переменным;

– Визуализация и Интерактив.

Скачать: https://www.rstudio.com/products/rstudio/download/, актуальная версия: 0.99.

❑ Задачи для самостоятельного решения

Теоретического плана (отчёт в электронном виде):

1. Расписать и обосновать примерами возможные отношения между категориями и типом содержимого информационных объектов социальных медиа.

Практического плана:

1. Установить R.
2. Установить RStudio.
3. Изучить возможности R, как интерактивного калькулятора.