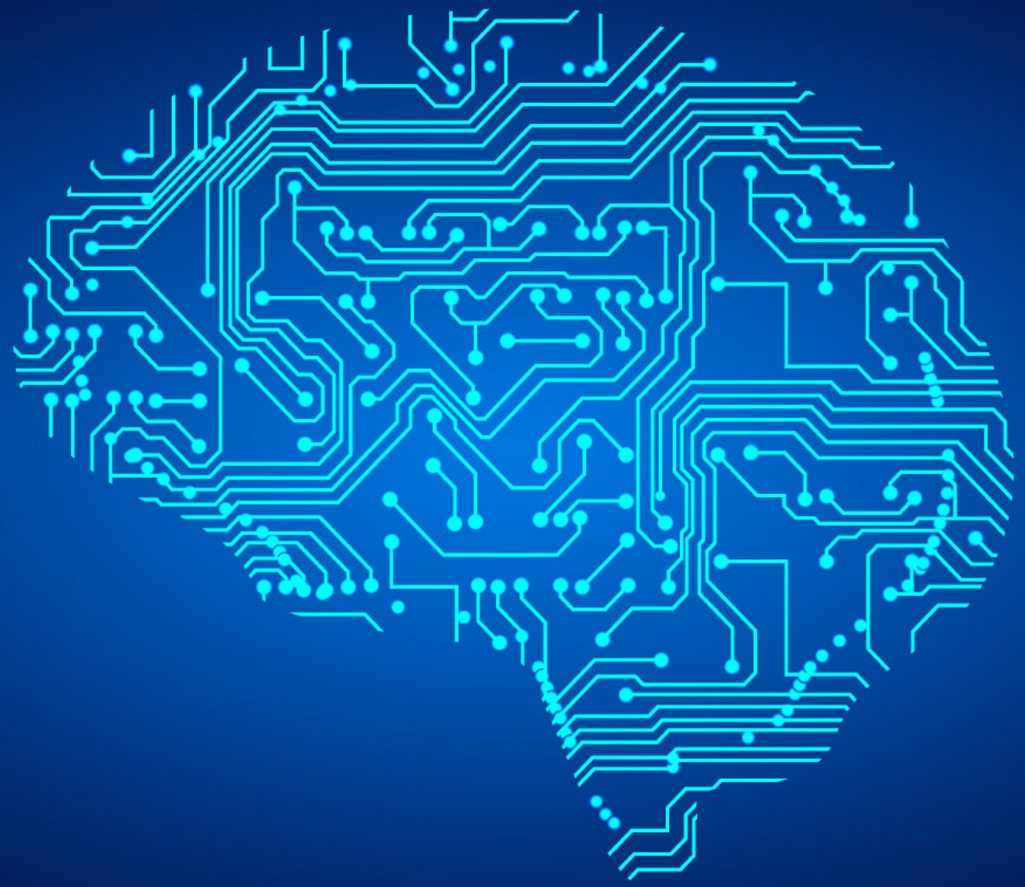
System Lab



DATA SCIENCE

ОЦКС Росатома

Москва. Январь 2018 г.

Дмитришин Юрий Михайлович
dmitrishin@system-lab.ru

ТЕРМИНОЛОГИЯ

DATA SCIENCE

DATA MINING

DATA SCIENCE AND DATA MINING

STATISTICAL ANALYSIS AND DATA MINING

MACHINE LEARNING

DEEP LEARNING

KNOWLEDGE DISCOVERY IN DATABASES

BIG DATA

ARTIFICIAL INTELLIGENCE

Рассматривая вопросы Data Science (Науки о данных), сначала необходимо ознакомиться с наиболее устоявшейся в мире англоязычной терминологией.

Область анализа и обработки данных интенсивно развивается, в связи с чем, встречается различная терминология, описывающая одно и тоже явление или сферу, или один термин, который может трактоваться по-разному. Например, в англоязычной литературе можно встретить различные термины и их сочетания, описывающие область интеллектуального анализа данных и являющихся достаточно близкими по значению.

Такая разнообразная терминология может содержать множество оттенков, которые определяются порой в каждом конкретном случае в зависимости от контекста.

Термины в русском языке: наука о данных, анализ данных, интеллектуальный анализ данных, глубинный анализ данных, машинное обучение, глубинное обучение, статистические методы анализа данных, большие данные, искусственный интеллект и.т.д.

BIG DATA БОЛЬШИЕ ДАННЫЕ

4V BIG DATA

VOLUME

VARIETY

VELOCITY

VERACITY

Анализ терминологии, достаточно полно описывающей область Data Science, позволяет отметить, что в ней часто в схожем контексте применяется и такой термин, как Big Data (рус. Большие данные).

При этом широта его употребления порой затрудняет однозначное толкование этого термина.

Активное тиражирование термина Big Data во многом вызвано сопровождением объективного процесса накопления сверхбольших объемов данных. Часто дискуссия о границах того, что является «действительно, большими данными», а что уже не является, сводится к тому, насколько дорогостоящая инфраструктура требуется для их поддержки.

Ряд экспертов сходится во мнении, что к «большим данным» относят конктретные технологии хранения и обработки данных.

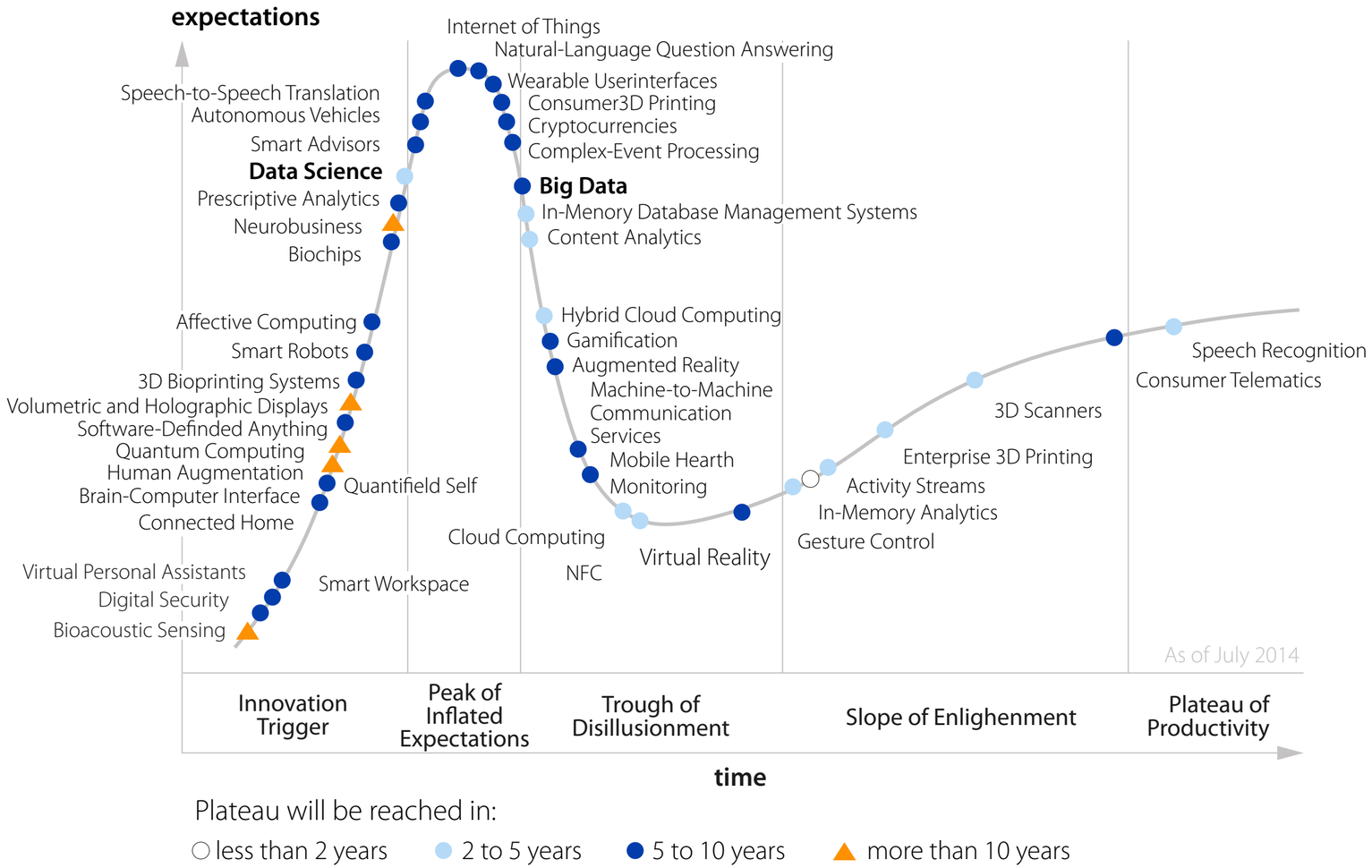
4V СВОЙСТВА BIG DATA

Дополнительное понимание термину Big Data придают четыре свойства, кратко сформулированные по четырем английским словам, начинающимся на букву v:

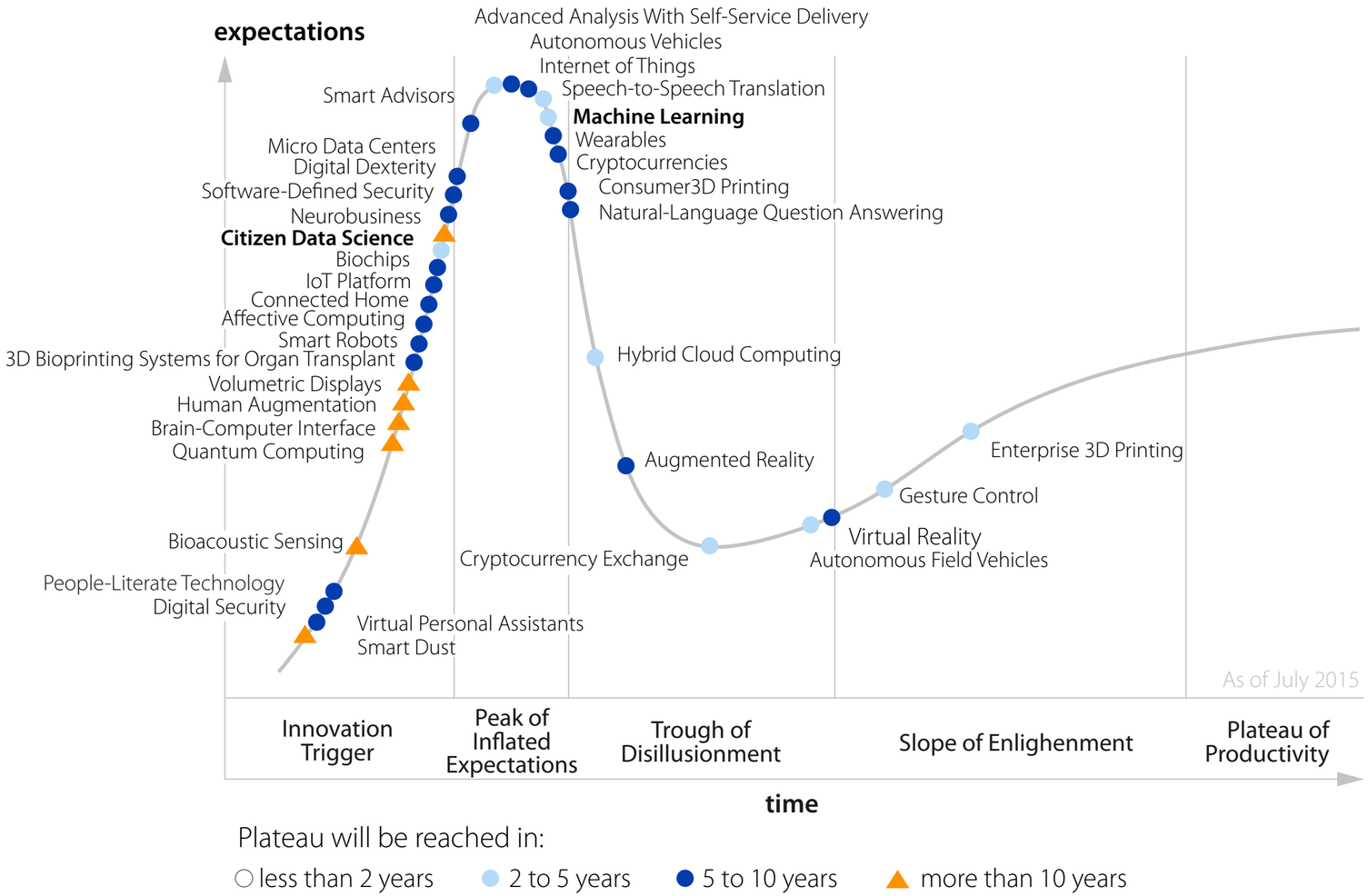
- **Volume (Объем)** — отражает значительный физический объем данных;
- **Variety (Разнообразие)** — показывает существенное разнообразие типов данных (например, структурированные, частично структурированные, неструктурированные), источников данных (внутренние, внешние, общественные) и их детальности;
- **Velocity (Скорость)** — демонстрирует скорость, с которой данные создаются и обрабатываются;
- **Veracity (Правдивость)** — определяет варьируемый уровень помех и ошибок в данных.

Свойство Volume часто наименее важное, и нет какого-либо обязательного требования к минимальному объему обрабатываемых данных в концепции Big Data.

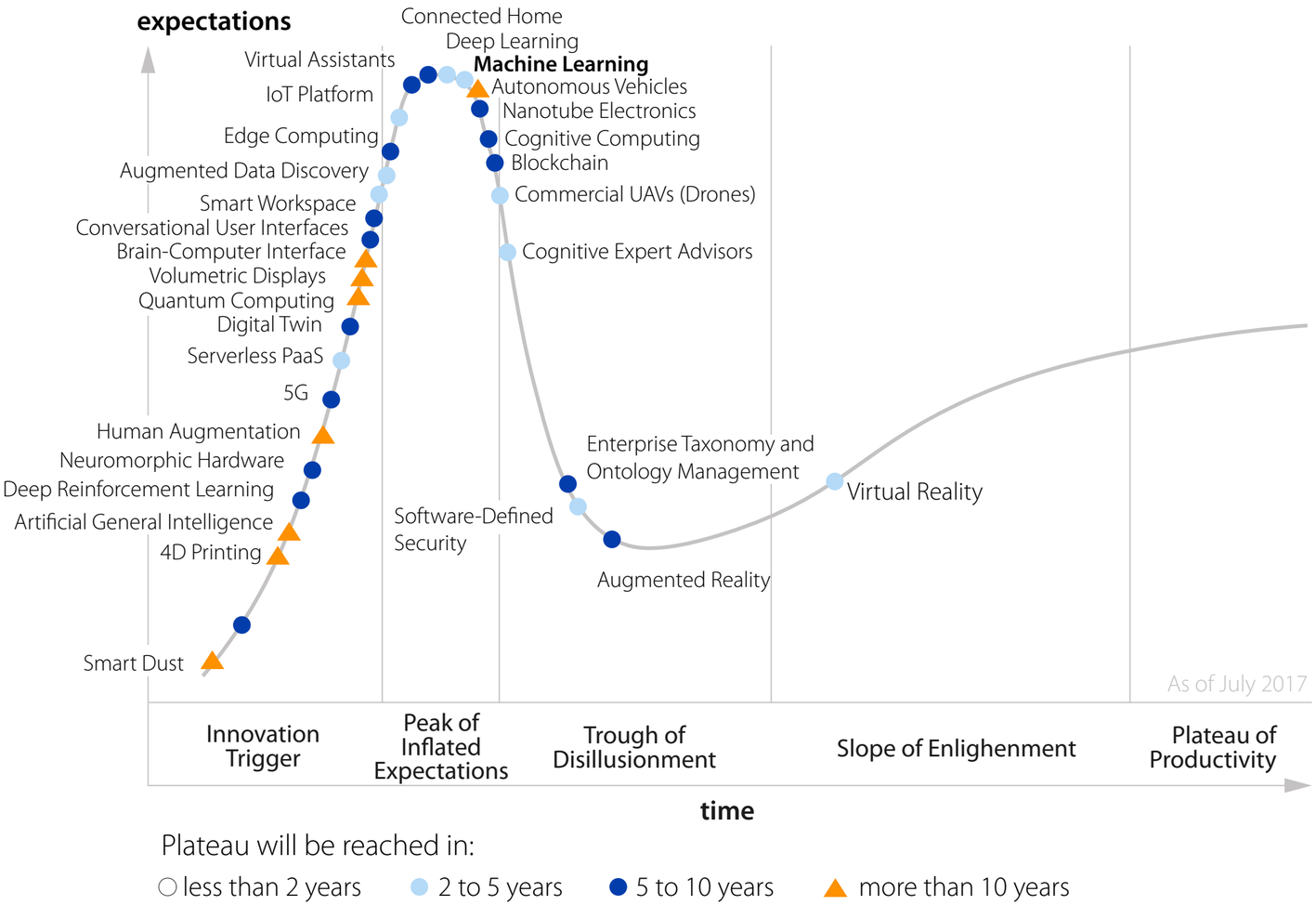
ЦИКЛ ЗРЕЛОСТИ ТЕХНОЛОГИИ (HYPE CYCLE) GARTNER 2014

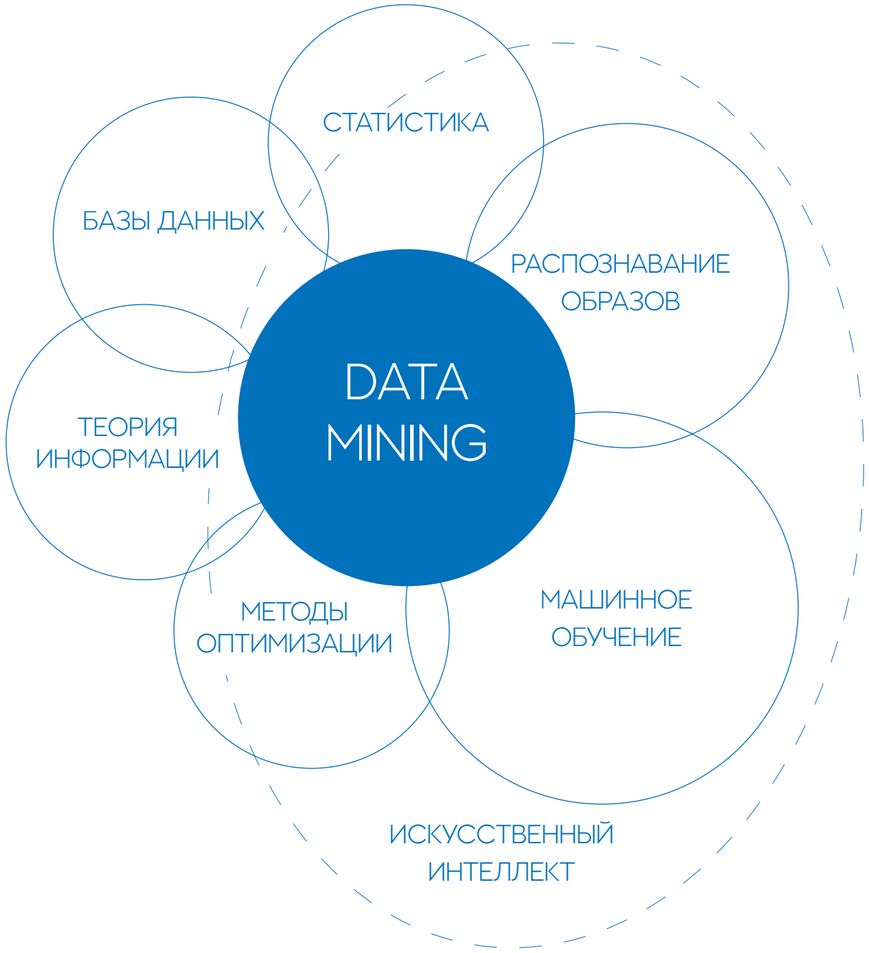


ЦИКЛ ЗРЕЛОСТИ ТЕХНОЛОГИИ (HYPE CYCLE) GARTNER 2015



ЦИКЛ ЗРЕЛОСТИ ТЕХНОЛОГИИ (HYPE CYCLE) GARTNER 2017





DATA MINING

Существующие массивы данных характеризуются не только значительным объемом и регулярной пополняемостью, но и содержат порой в себе скрытые данных и закономерности.

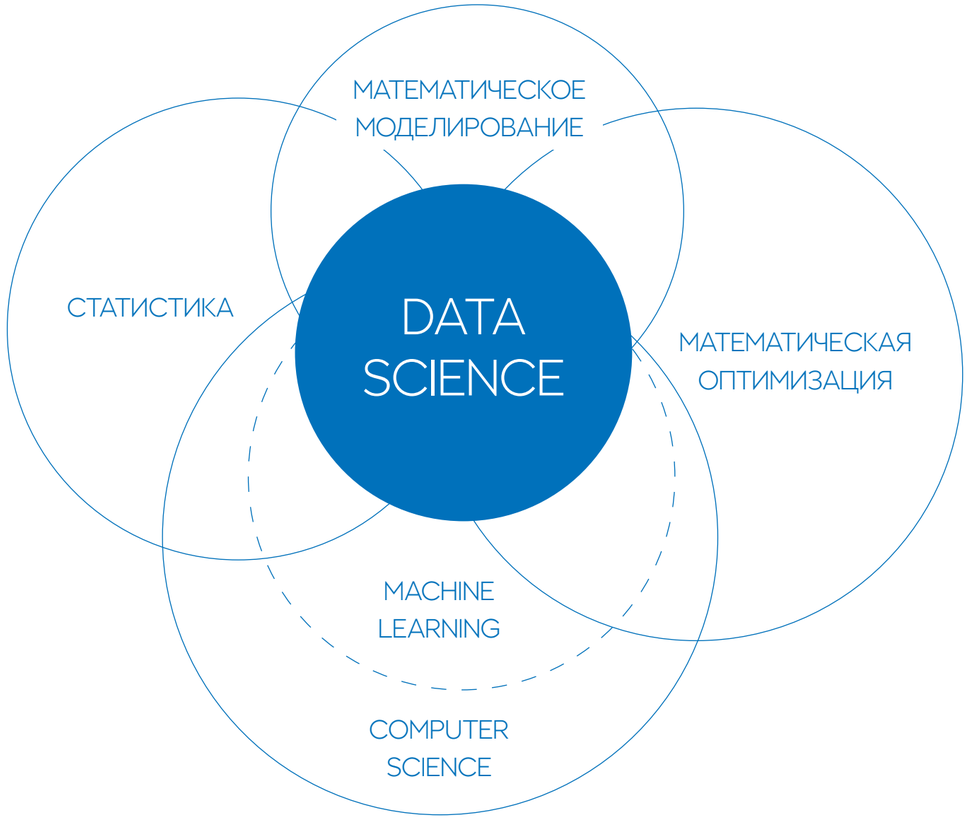
Процесс поиска в этих закономерностей в качестве чего-то ценного, стал сравним с работой на горнорудных предприятиях, где в многотонных завалах руды осуществляется поиск (добыча) драгоценных металлов или камней, полезный выход которых может исчисляться граммами.

Формально для Data Mining могут быть даны различные определения, не претендующие на исключительную полноту. Вот некоторые из них.

Data Mining — это:

- процесс обнаружения в базах данных нетривиальных и практически полезных закономерностей
- процесс выделения, исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур с целью достижения преимуществ в бизнесе.

DATA SCIENCE

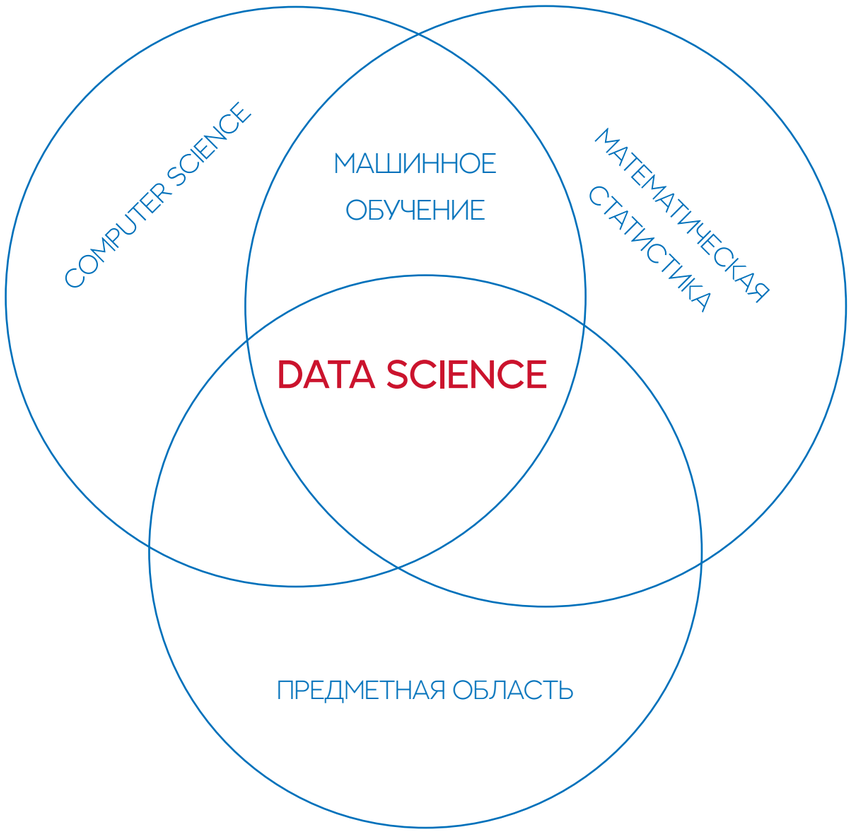


Data Science — это набор конкретных дисциплин из разных направлений, отвечающих за анализ данных и поиск оптимальных решений на их основе. Раньше этим занималась только математическая статистика и прикладная математика, затем начали использовать машинное обучение и искусственный интеллект, которые в качестве методов анализа данных к математической статистике добавили Computer science.

Благодаря анализу большого объема данных получается эффективнее принимать управленческие решения. Пользу от анализа данных можно извлечь во всех более-менее прикладных областях, где есть достаточно данных. Для того, чтобы понять как применить анализ данных к предметной области, необходимо в ней разбираться.

Знания — это сила, а знания, полученные из больших данных,— большая сила.

Диаграмма о Data Science Дрю Конвея



DATA SCIENCE

Наука о данных охватывает три отдельные, но пересекающиеся сферы:

- навыки специалиста по математической статистике, умеющего моделировать наборы данных и извлекать из них основное;
- навыки специалиста в области компьютерных наук, умеющего проектировать и использовать алгоритмы для эффективного хранения, обработки и визуализации этих данных;
- экспертные знания предметной области, полученные в ходе традиционного изучения предмета,— умение как формулировать правильные вопросы, так и рассматривать ответы на них в соответствующем контексте.

С учетом этого я рекомендовал бы рассматривать науку о данных не как новую область знаний, которую нужно изучить, а как новый набор навыков, который вы можете использовать в рамках хорошо знакомой вам предметной области.

Плас Дж. Вандер «Python для сложных задач: наука о данных и машинное обучение» 2018 г.

МАШИННОЕ ОБУЧЕНИЕ

КЛАССИЧЕСКИЕ ГРУППЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

КЛАССИФИКАЦИЯ

РЕГРЕССИЯ

ОБУЧЕНИЕ С УЧИТЕЛЕМ

КЛАСТЕРИЗАЦИЯ

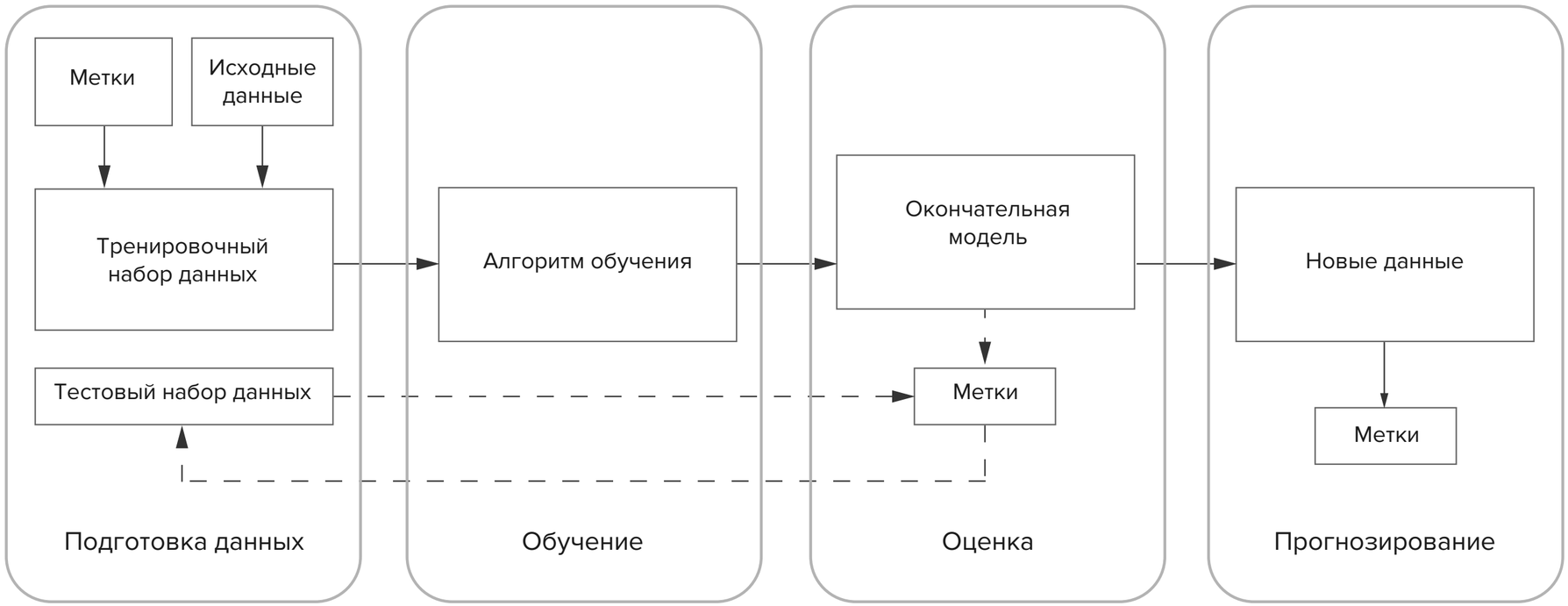
ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

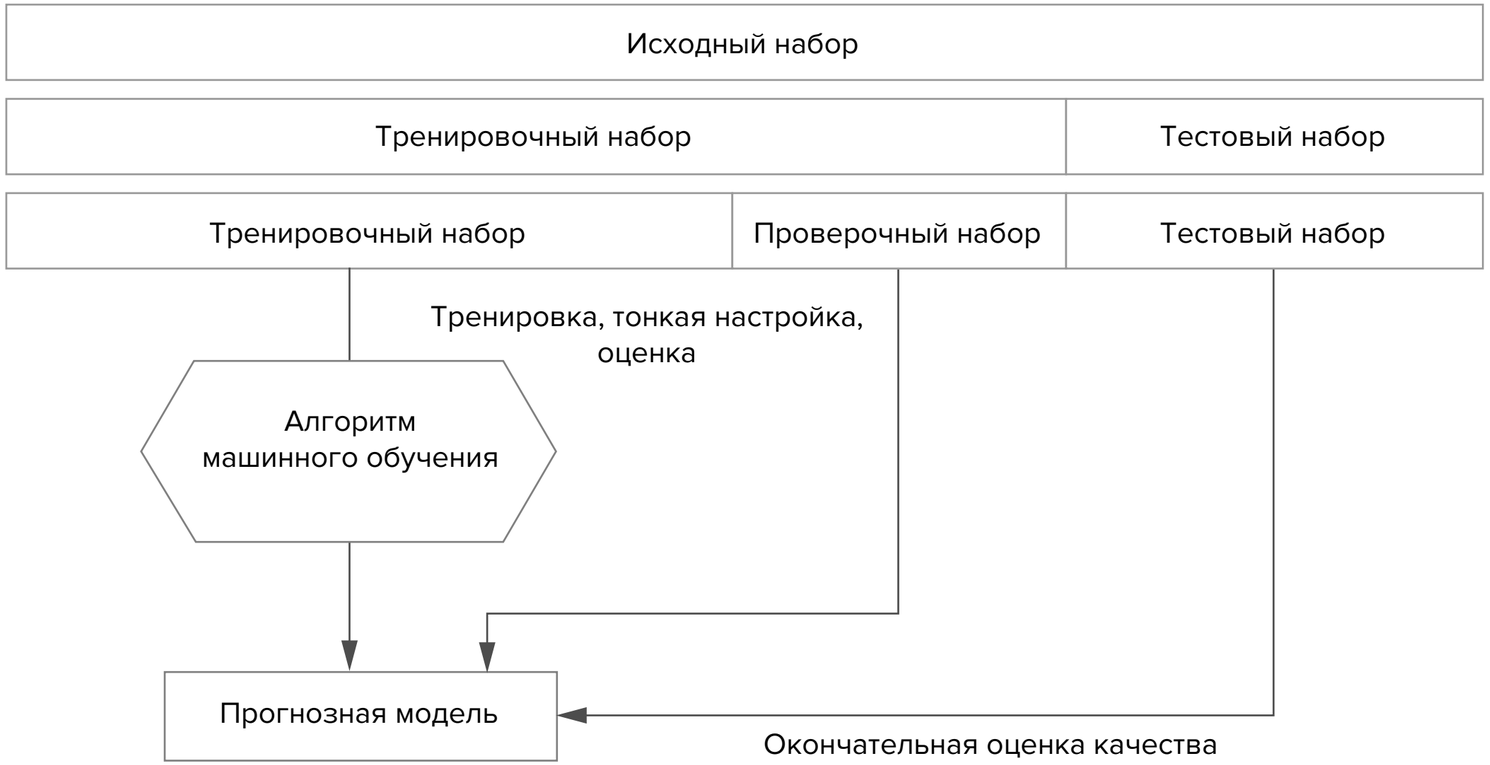
Задача классификации состоит в том, что требуется определить к какому из известных классов относятся исследуемые объекты, то есть классифицировать эти объекты. Причем каждый из этих объектов имеет некоторое количество характеристик. Например рассматривается большое количество объектов имеющих несколько фиксированных характеристик и требуется дать ответ «да» или «нет» по каждому из этих объектов. На первом этапе решения таких типов задач выделяется обучающая выборка. На основании обучающей выборки строится модель определения значений зависимой переменной (функцией классификации или функции регрессии). На втором этапе построенную модель применяют к анализируемым объектам.

Задача кластеризации состоит в разделении исследуемого множества объектов на группы «похожих» объектов, называемых кластерами. Слово кластер переводится как сгусток, пучек, группа.

ПРОЦЕСС ПОСТРОЕНИЯ СИСТЕМ МАШИННОГО ОБУЧЕНИЯ



ОЦЕНКА И ТОНКАЯ НАСТРОЙКА МОДЕЛЕЙ В МАШИННОМ ОБУЧЕНИИ



System Lab