Big Data for Labour Market Information

Сессия 1

Общий обзор использования больших данных в контексте ИРТ

Алессандро Ваккарино – Фабио Меркорио

Большие данные в контексте информации рынка труда – акцент на данных онлайн-вакансий. Учебный семинар

Милан, 21–22 ноября 2019 года

ETF Working together Learning for life

European Commission

EaP Eastern Partnership

Темы

1. Коротко о больших данных
2. Так что же такое искусственный интеллект? [на примерах]
3. Большие данные в контексте ИРТ

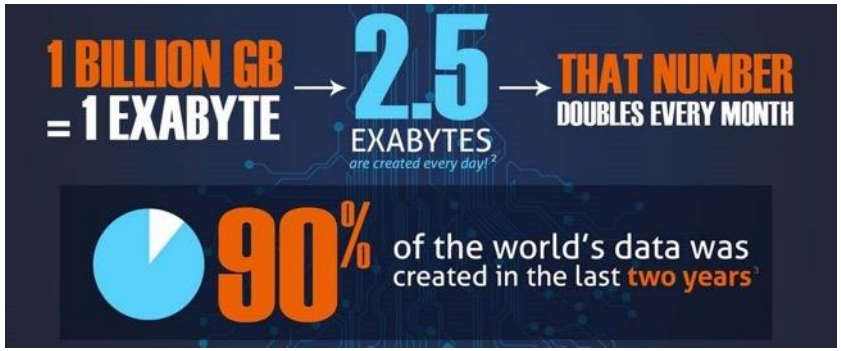
Коротко о больших данных

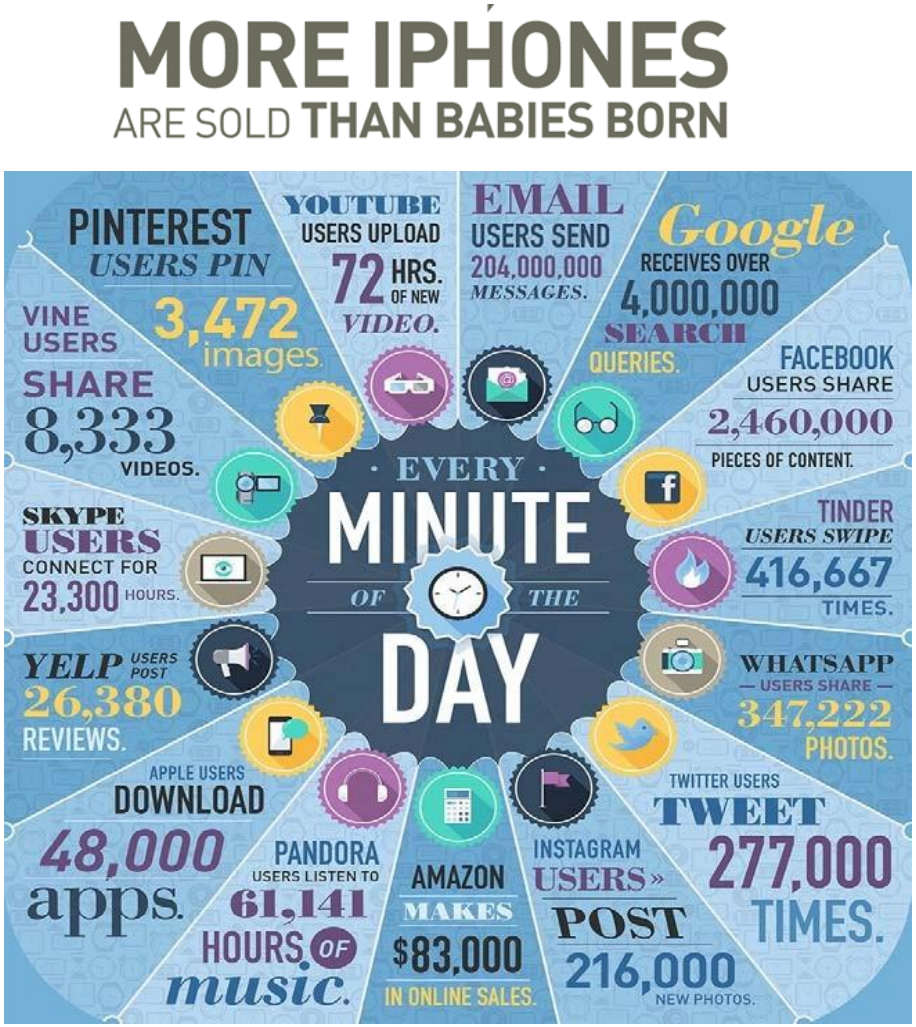
«Большими данными» обычно называют крупные объёмы разных видов данных, которые генерируются с высокой скоростью большим количеством разнообразных источников.

Чтобы эти **данные** стали полезны для заинтересованных лиц, их необходимо превратить в знания, **так как именно знания являются конечным продуктом, получаемым из сведений на основе данных**

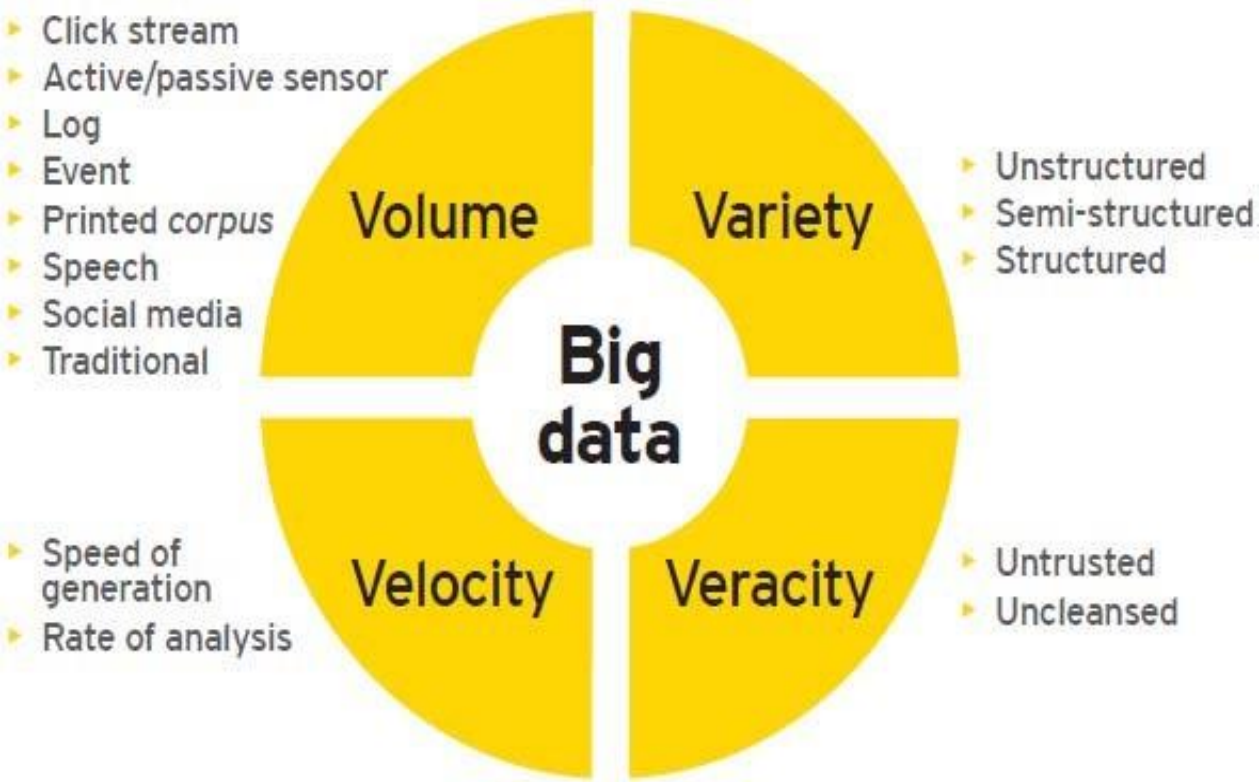
Модель больших данных на основе четырёх «V»

Количество данных быстро растёт





Не просто «много данных»



Большие данные бесполезны без «искусственного интеллекта», который извлекает из них знания.

Искусственный интеллект – перемена определений

Хаугеланд (1985)

Новая волнующая попытка научить компьютеры думать ... *машины, обладающие умом*, в полном и буквальном смысле

Рич и Найт (1991)

Работа над тем, чтобы научить компьютеры делать то, с чем люди пока ещё справляются лучше

Шалкофф (1990)

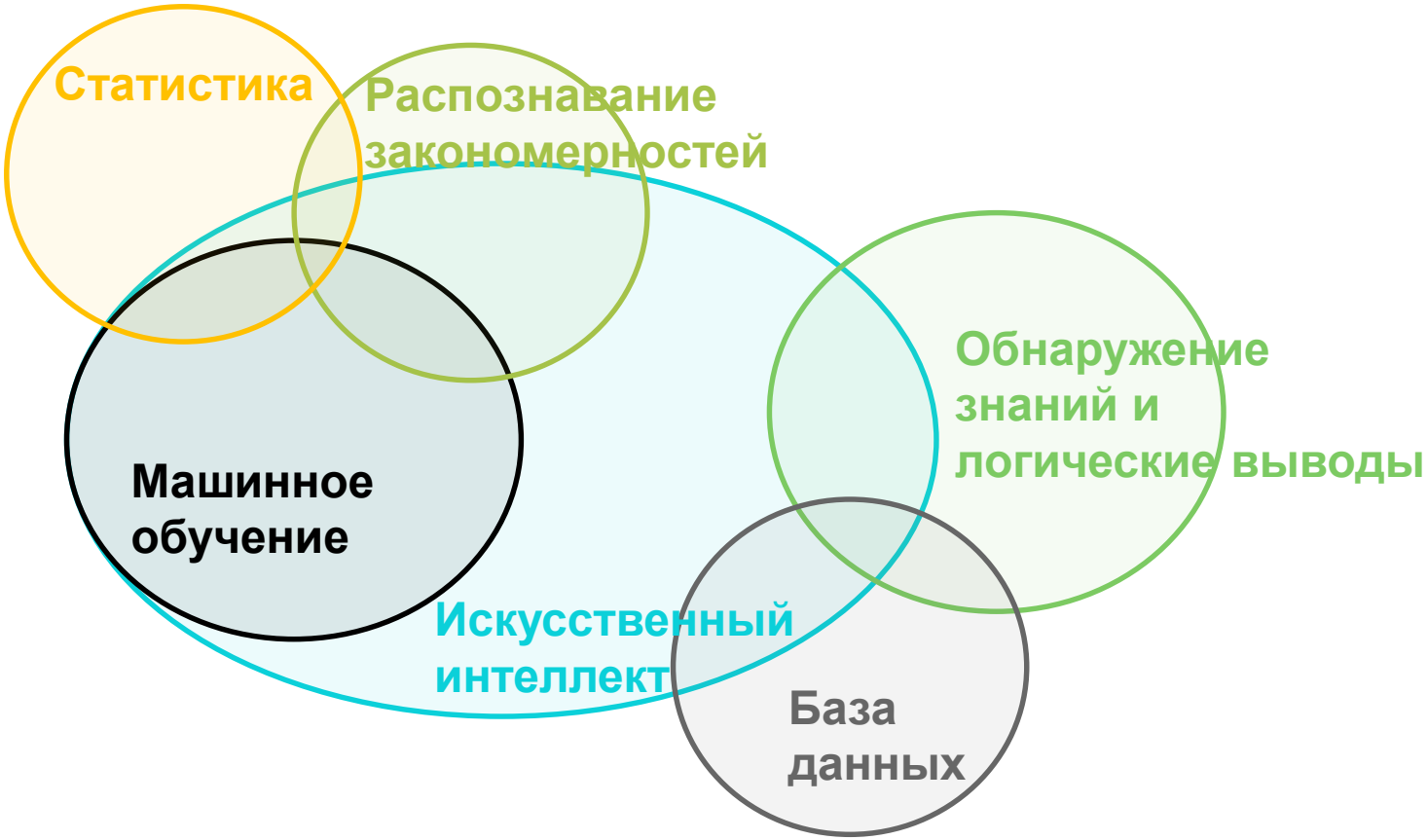
Область изучения, которая пытается объяснить и сымитировать разумное поведение с точки зрения вычислительных процессов

ЕС – ИИ для Европы (2018)

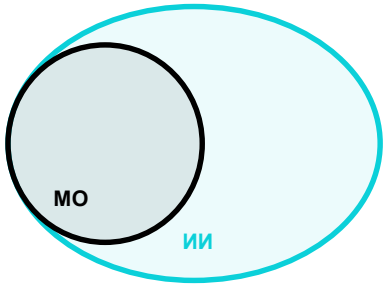
Системы, демонстрирующие разумное поведение, благодаря анализу окружения и осуществлению действий (с определённой мерой самостоятельности) для достижения конкретных целей

ИИ: междисциплинарный подход

Большие данные: топливо ИИ



Два макро-вида ИИ

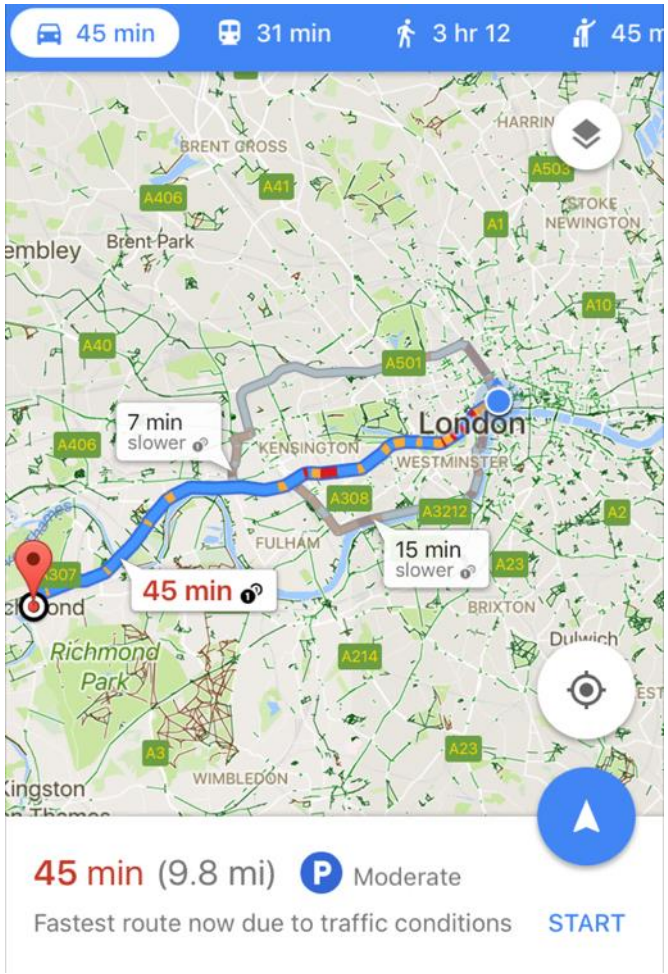


Узкий (слабый) ИИ: способен выполнять отдельные задачи (играть в шахматы, рекомендовать товары, составлять прогнозы и т.п.). Контекст и задачи определены.

Общий (сильный) ИИ: способен делать выводы, принимать самостоятельные решения и выполнять неопределённое количество задач, как человек. Контекст и задачи не определены (реальность).

Будет показано видео

Планирование ИИ



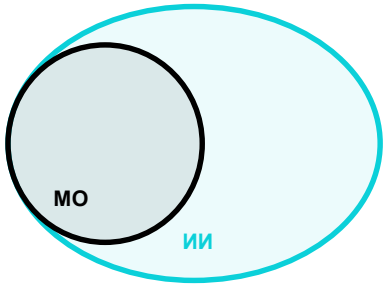
ВХОДНЫЕ ДАННЫЕ:

-карты
-исходные условия(загруженность дорог, GPS и т.п.)
-цель (пункт назначения)

РЕЗУЛЬТАТ:

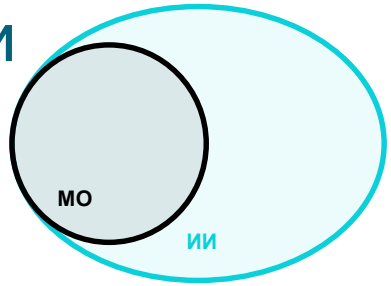
-план (минимальное время/км/прочее)

Машинное обучение



Программа, которая учится выполнять определённую задачу, используя собственный опыт, тем самым обогащая его и со временем улучшая способность выполнять задачу, для которой она была создана.

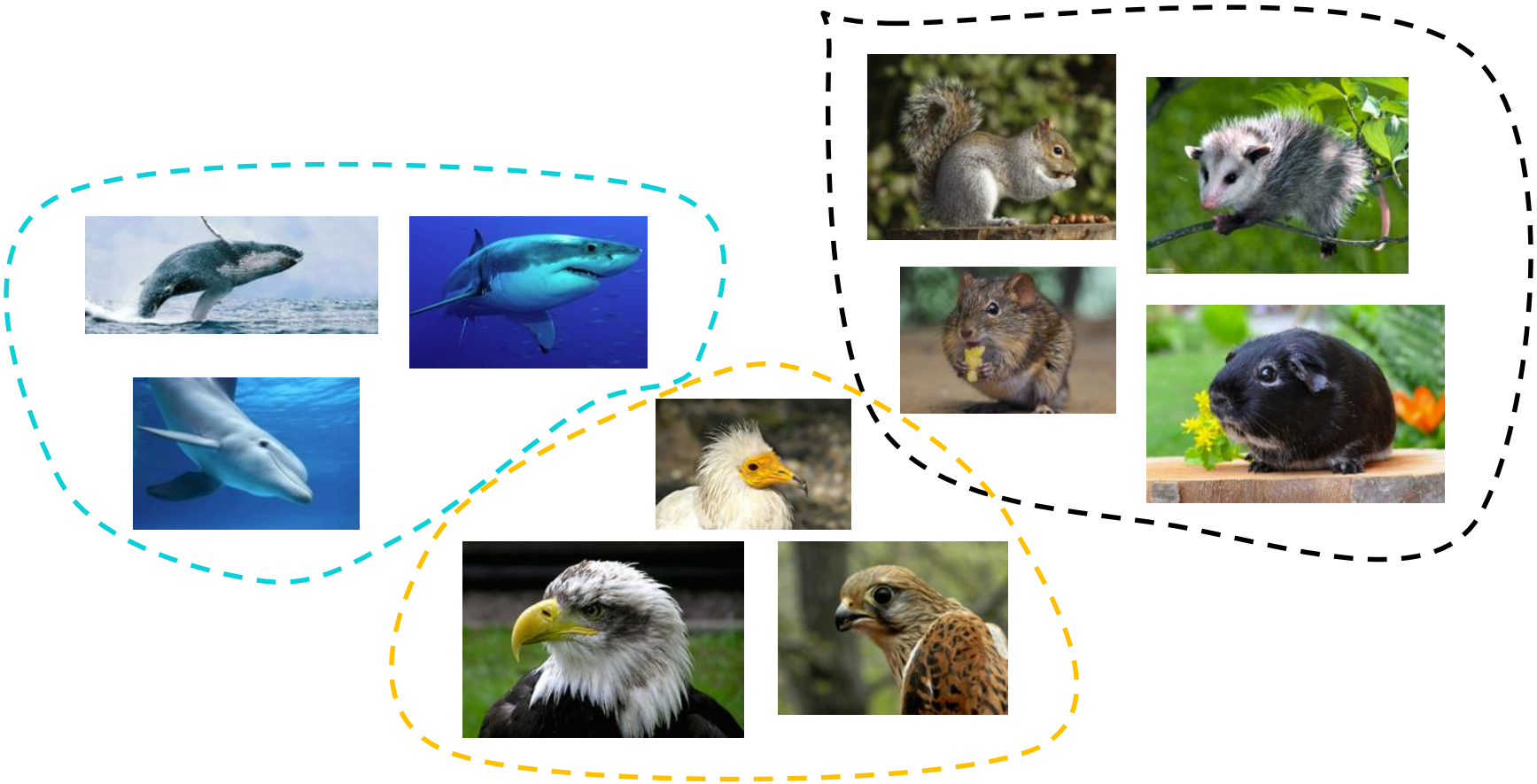
Машинное обучение: две категории



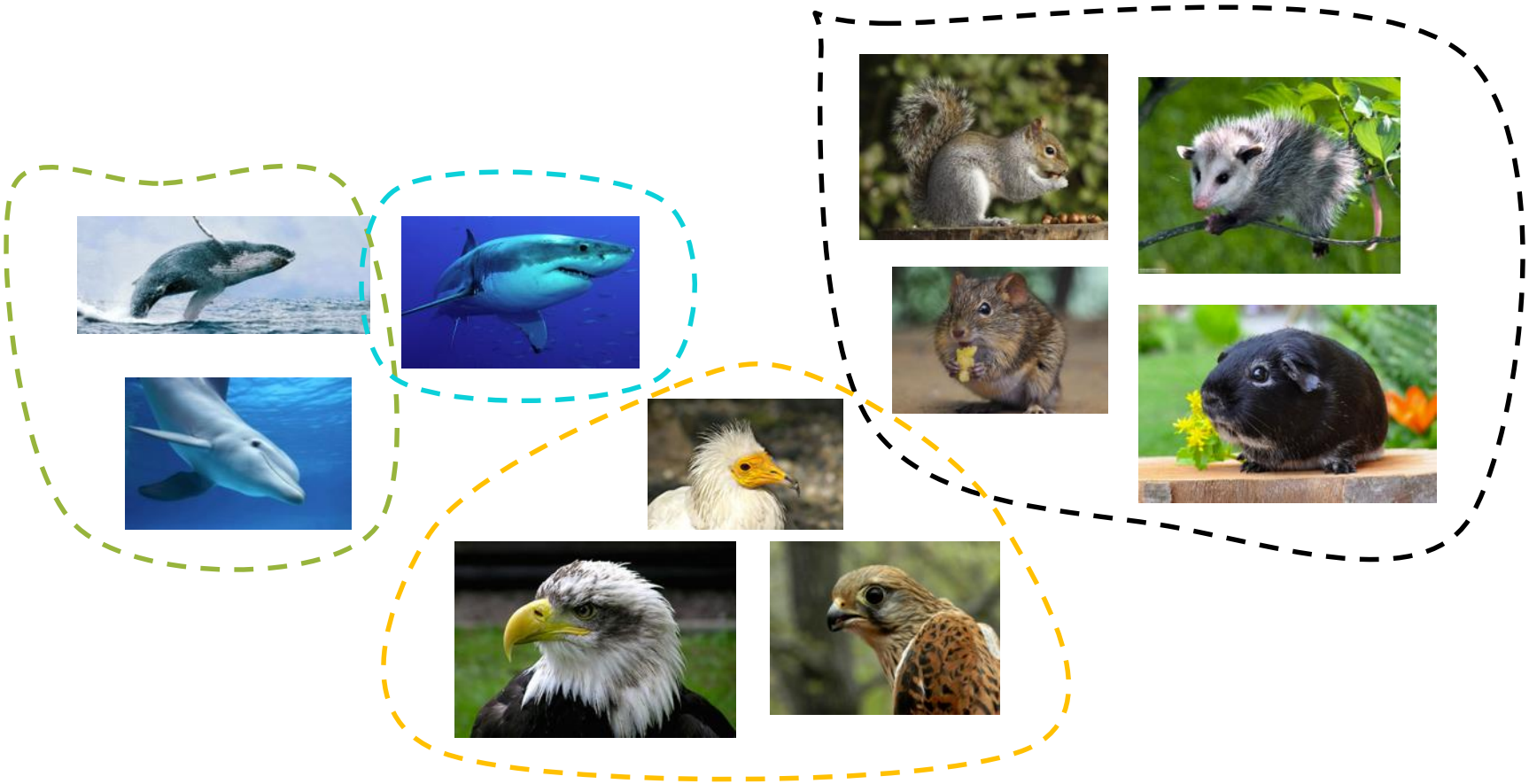
Без учителя

С учителем

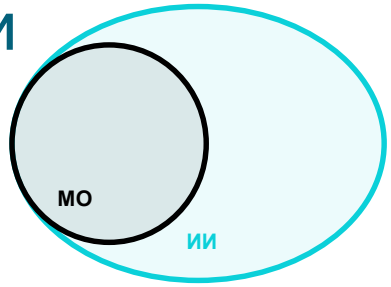
Обучение без учителя



Обучение без учителя



Машинное обучение: две категории



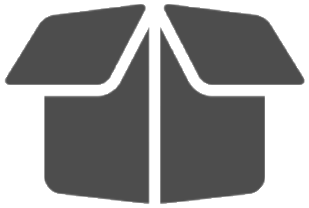
Без учителя:

Система классифицирует объекты, обладающие схожими и общими характеристиками (чертами), исходя из критерия сходства. Результаты могут отличаться в зависимости от критерия классификации

Обучение с учителем (фаза обучения)

Обучающий набор

(чем больше, тем лучше)



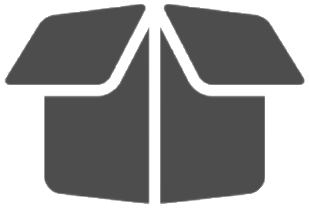
Алгоритм машинного обучения

Обучение с учителем (фаза оценки)

Белая акула (V)

Кит (X)

Тестовый набор

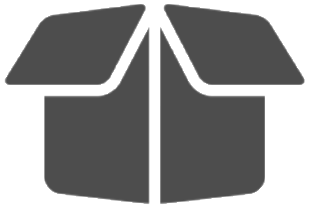


Алгоритм машинного обучения

Оценка: точность 92 %

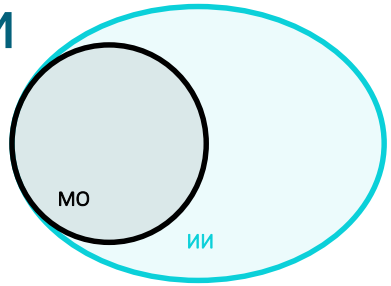
Обучение без учителя (в эксплуатации)

Акула или кит?



Алгоритм машинного обучения

Машинное обучение: две категории



С учителем

Система классифицирует объекты, обладающие схожими чертами на основе характеристик, обнаруженных на этапе обучения. Фаза тестирования просто позволяет узнать, насколько хорошо система прошла фазу обучения. Узнать, насколько хорошо система проявит себя на этапе эксплуатации (т.е. работая с новыми, незнакомыми объектами), нельзя

Обучение с учителем: проблема

Набор данных **должен быть**:

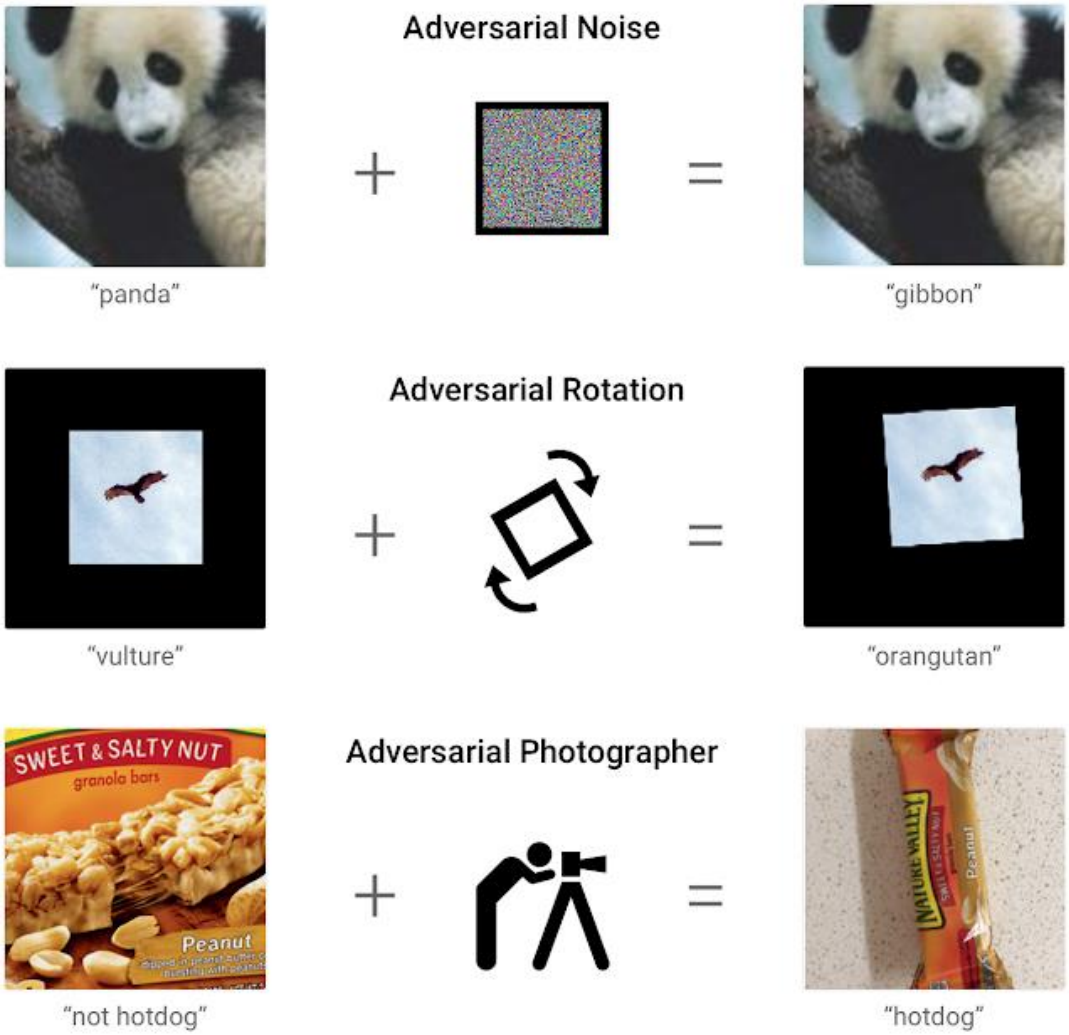
- большим
- иметь метки, проставленные экспертами в предметной области (контрольный набор данных)

Плюсы и минусы

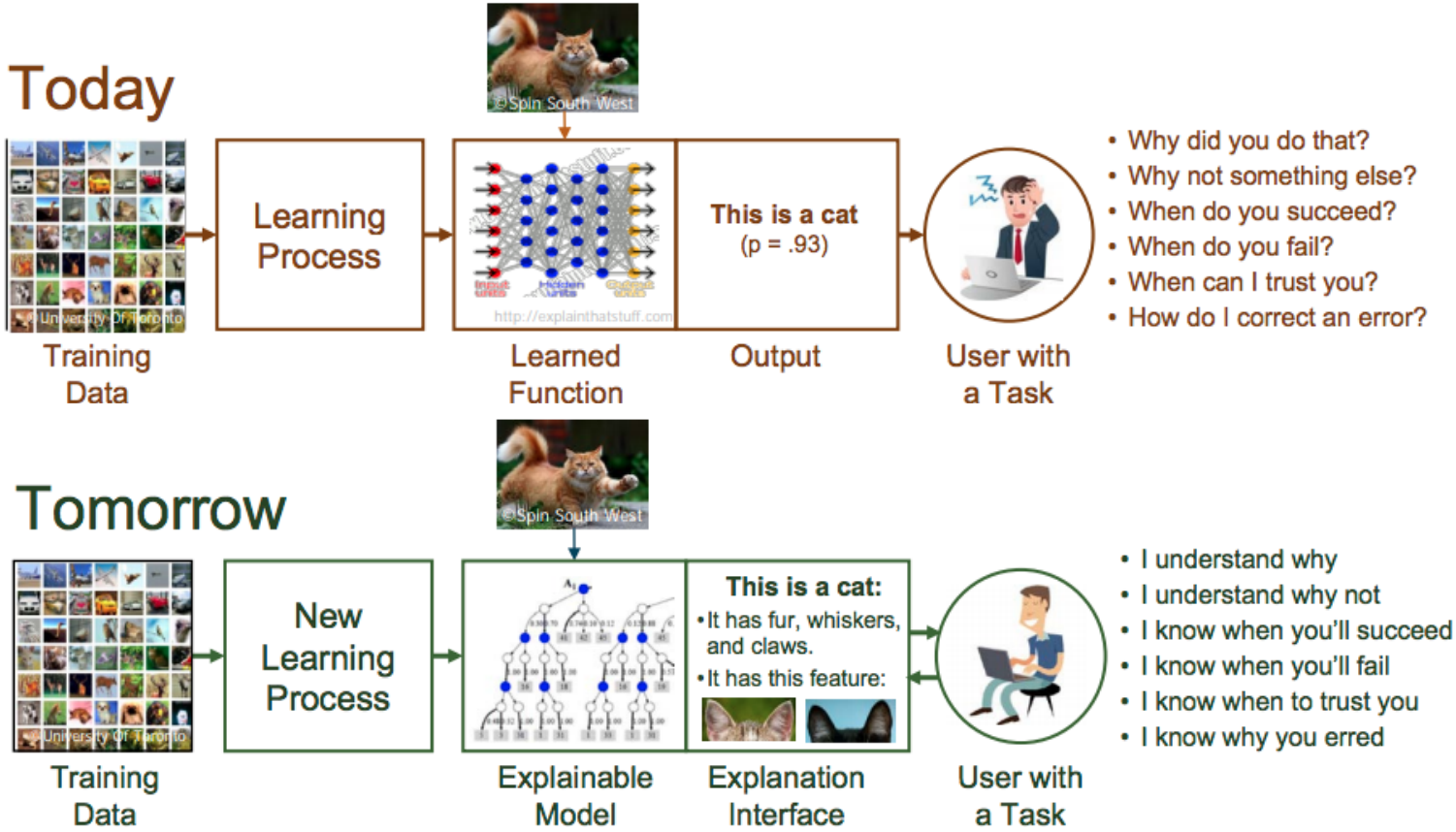
- МО годится для выполнения **задач, НЕ имеющих критического значения**, поскольку программы не могут дать человеку обоснование своего поведения... **объяснимый ИИ**

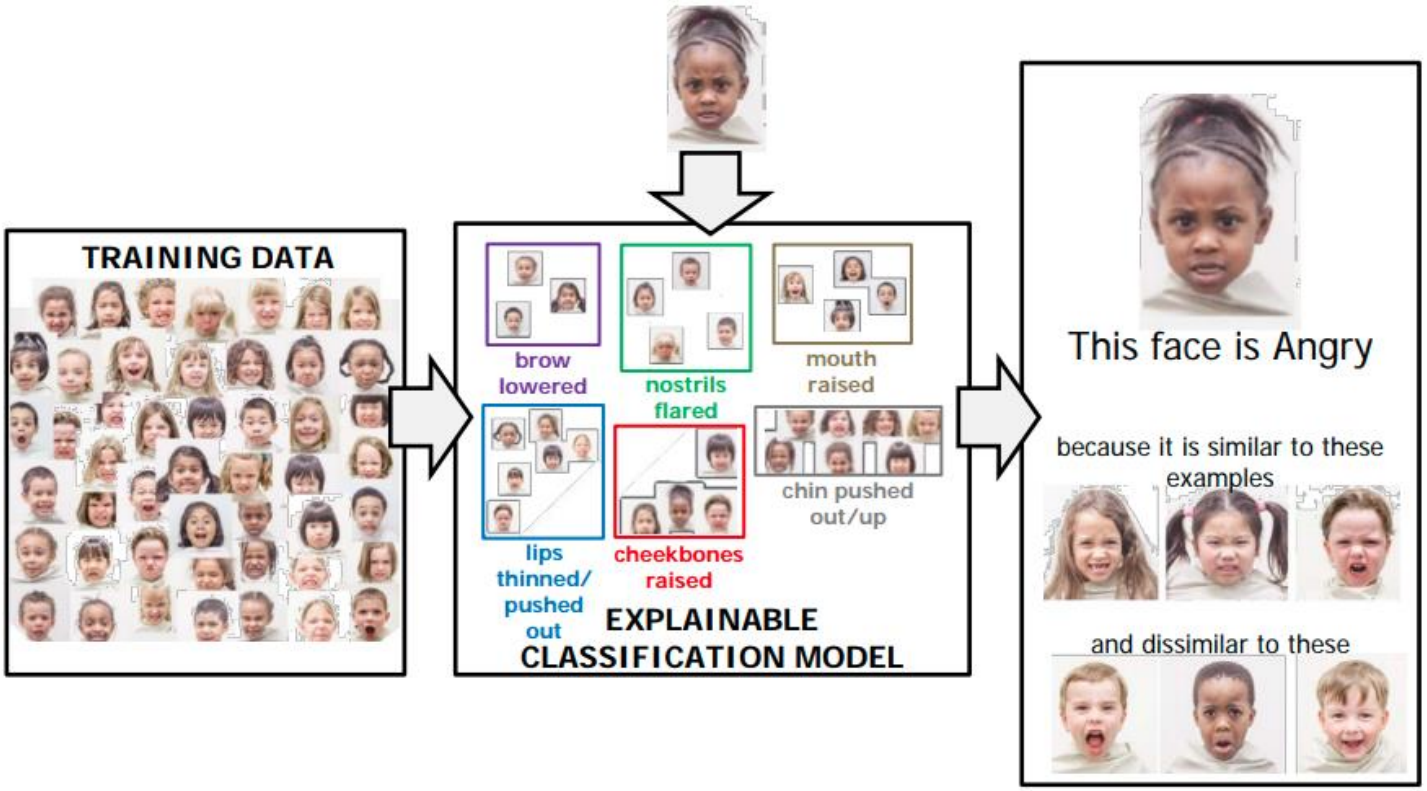
Глубокая нейронная сеть

Такая сильная... такая слабая



Объяснимый ИИ объясняет свои решения

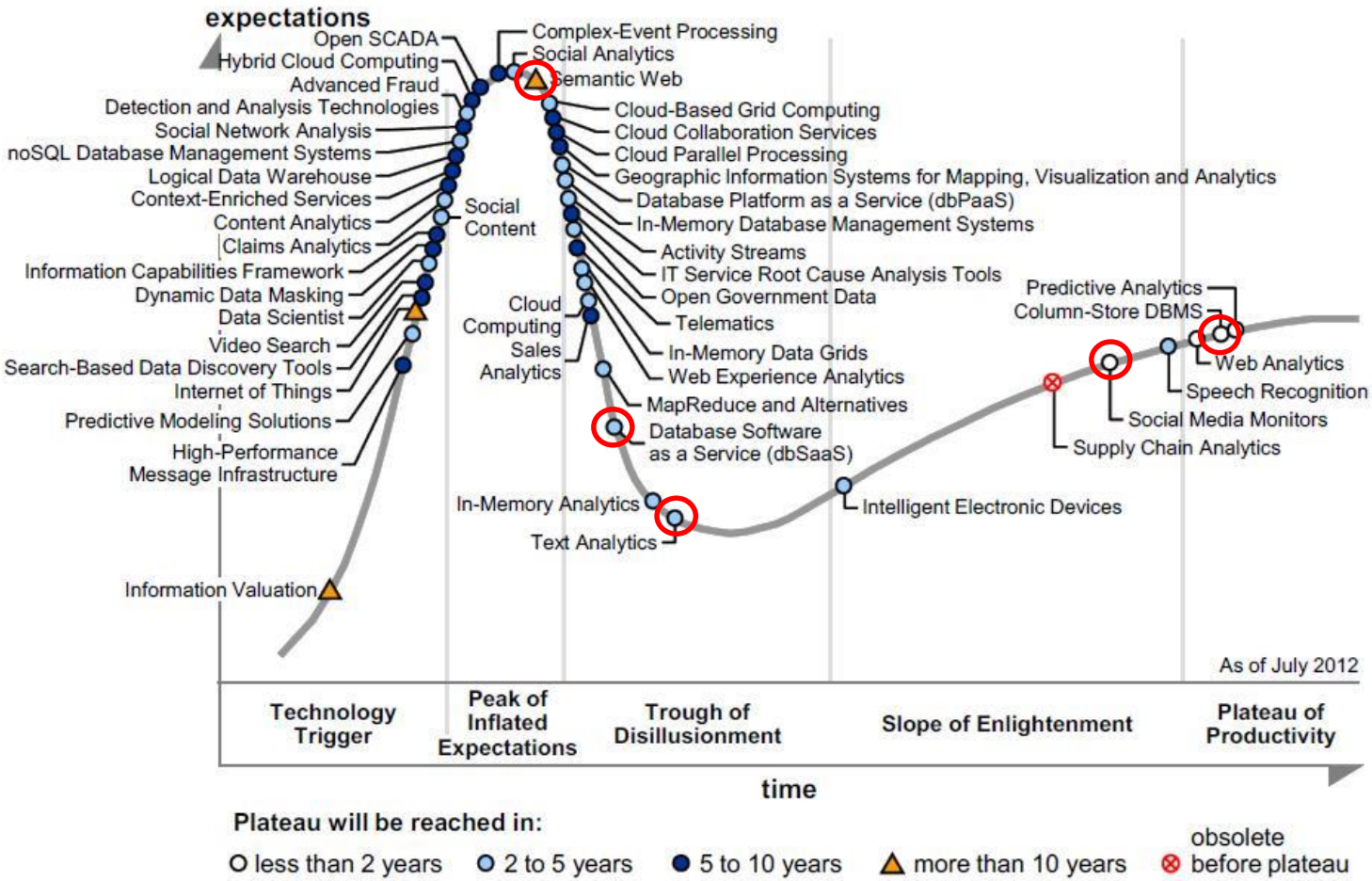




Сферы, где необходимо объяснение:

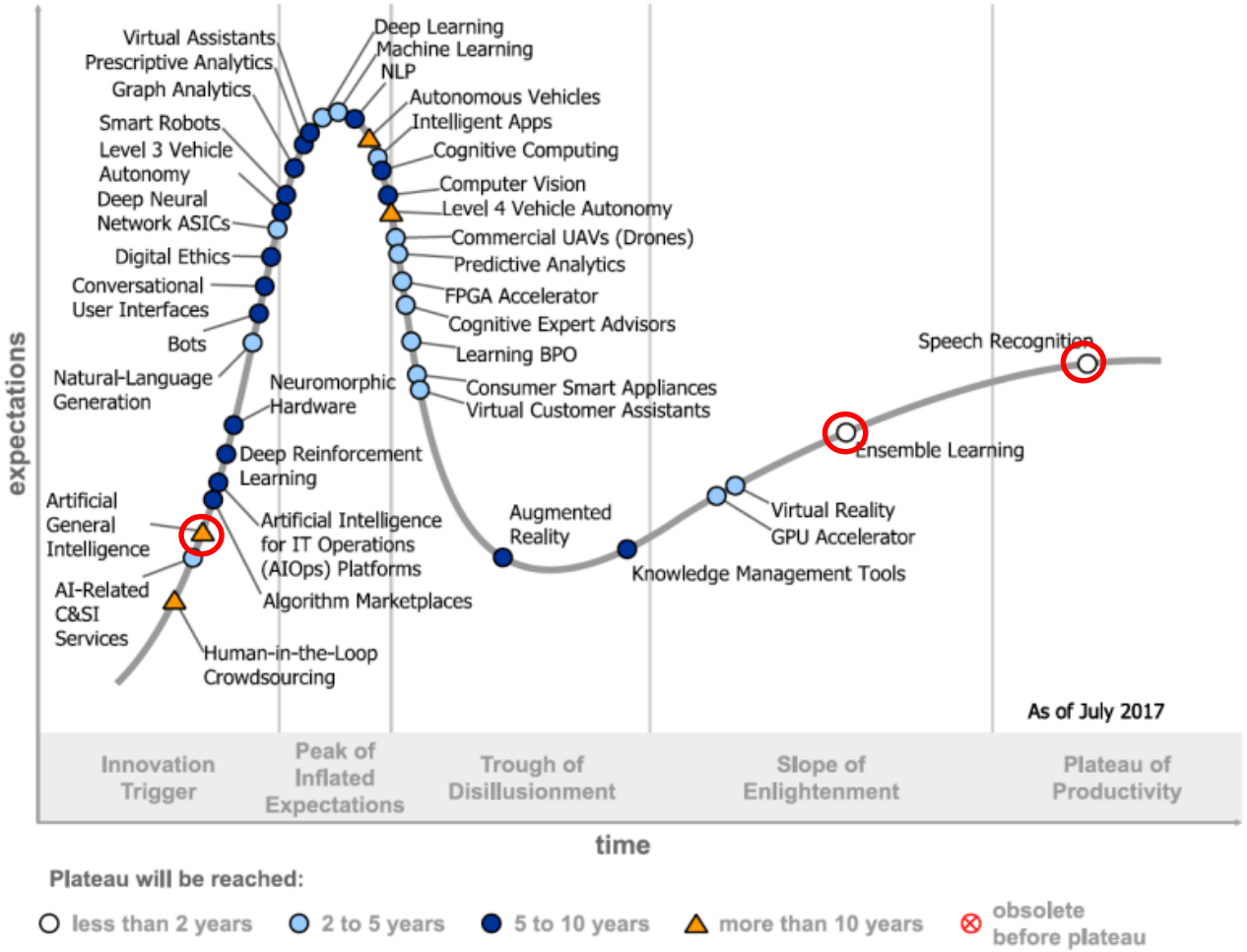
-- Медицина
-- Транспорт
-- Военное дело
-- Финансы
-- Юриспруденция
...

Figure 1. Hype Cycle for Big Data, 2012



Source: Gartner (July 2012)

Figure 1. Hype Cycle for Artificial Intelligence, 2017

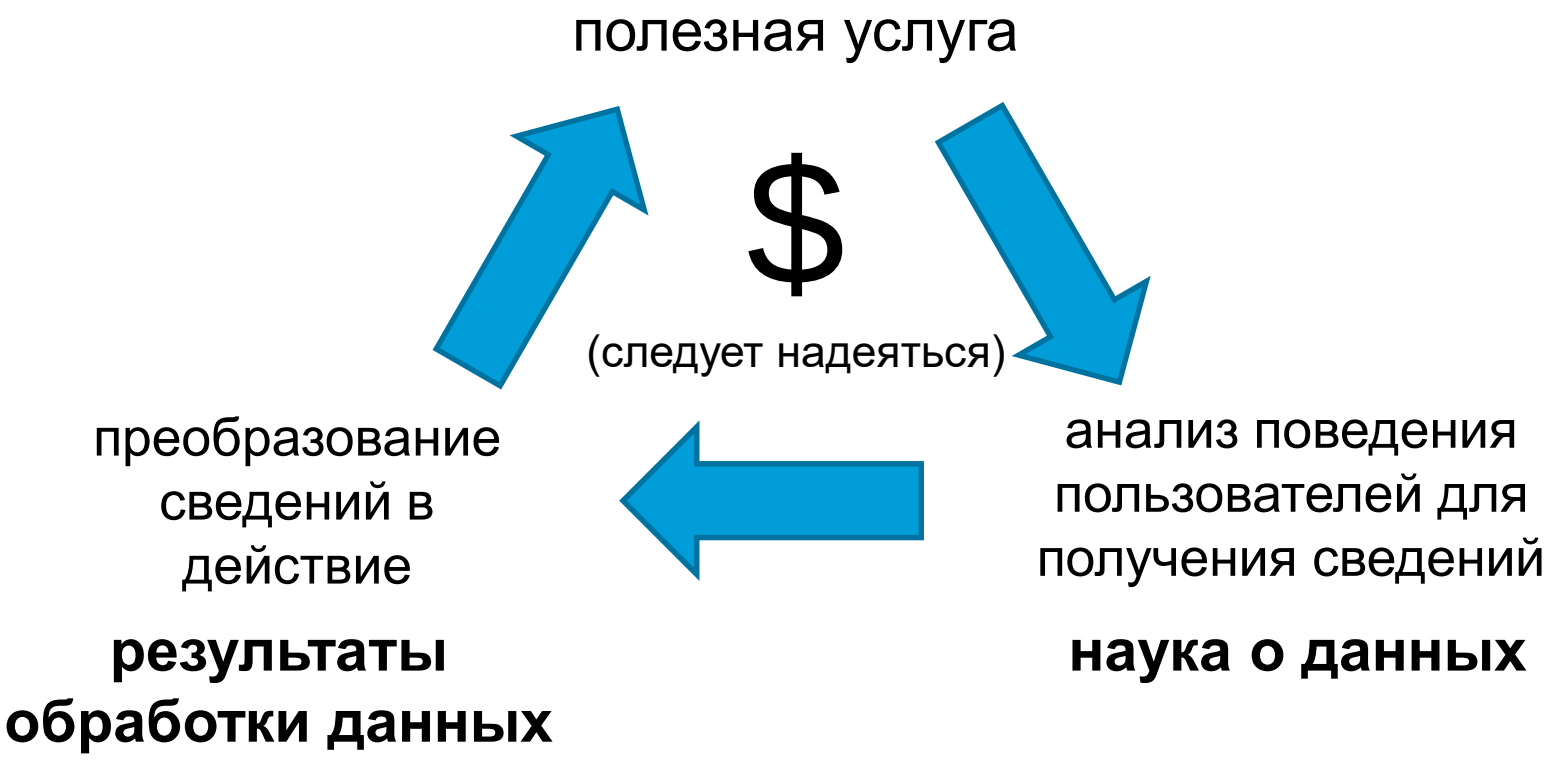


Plateau will be reached:

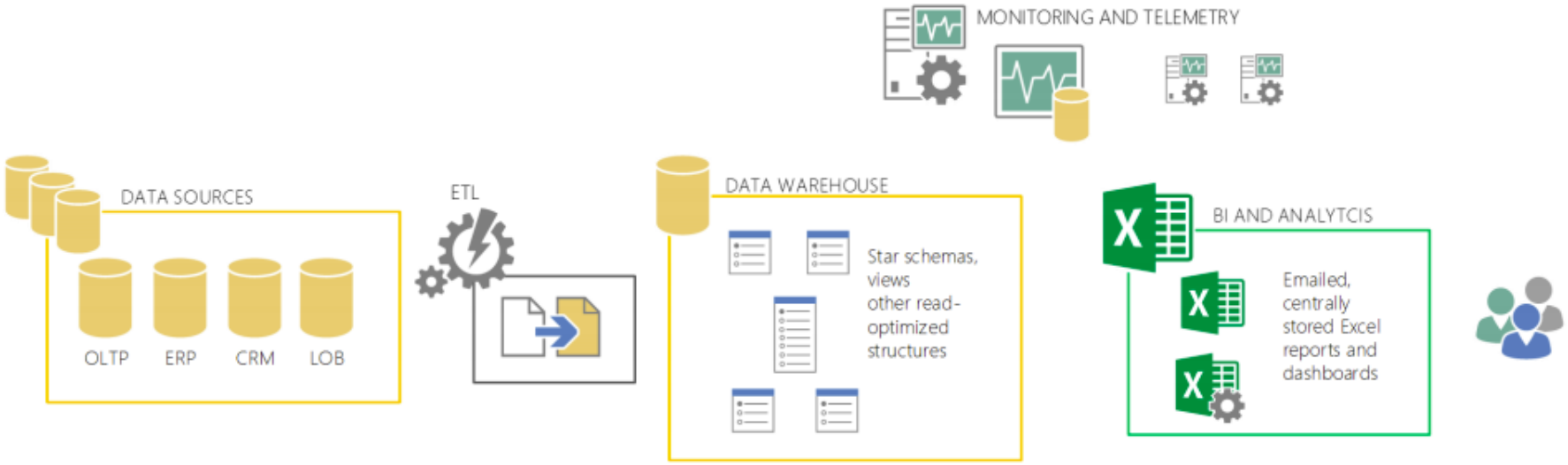
○ less than 2 years ● 2 to 5 years ● 5 to 10 years ▲ more than 10 years ⊗ obsolete before plateau

Как большие данные и ИИ взаимодействуют, чтобы извлечь новые знания из данных?

...«Наука о данных»


полезная услуга
$
(следует надеяться)
преобразование сведений в действие
результаты обработки данных
анализ поведения пользователей для получения сведений
наука о данных

До больших данных



Well manicured, often relational sources

Known and expected data volume and formats

Little to no change

Complex, rigid transformations

Required extensive monitoring

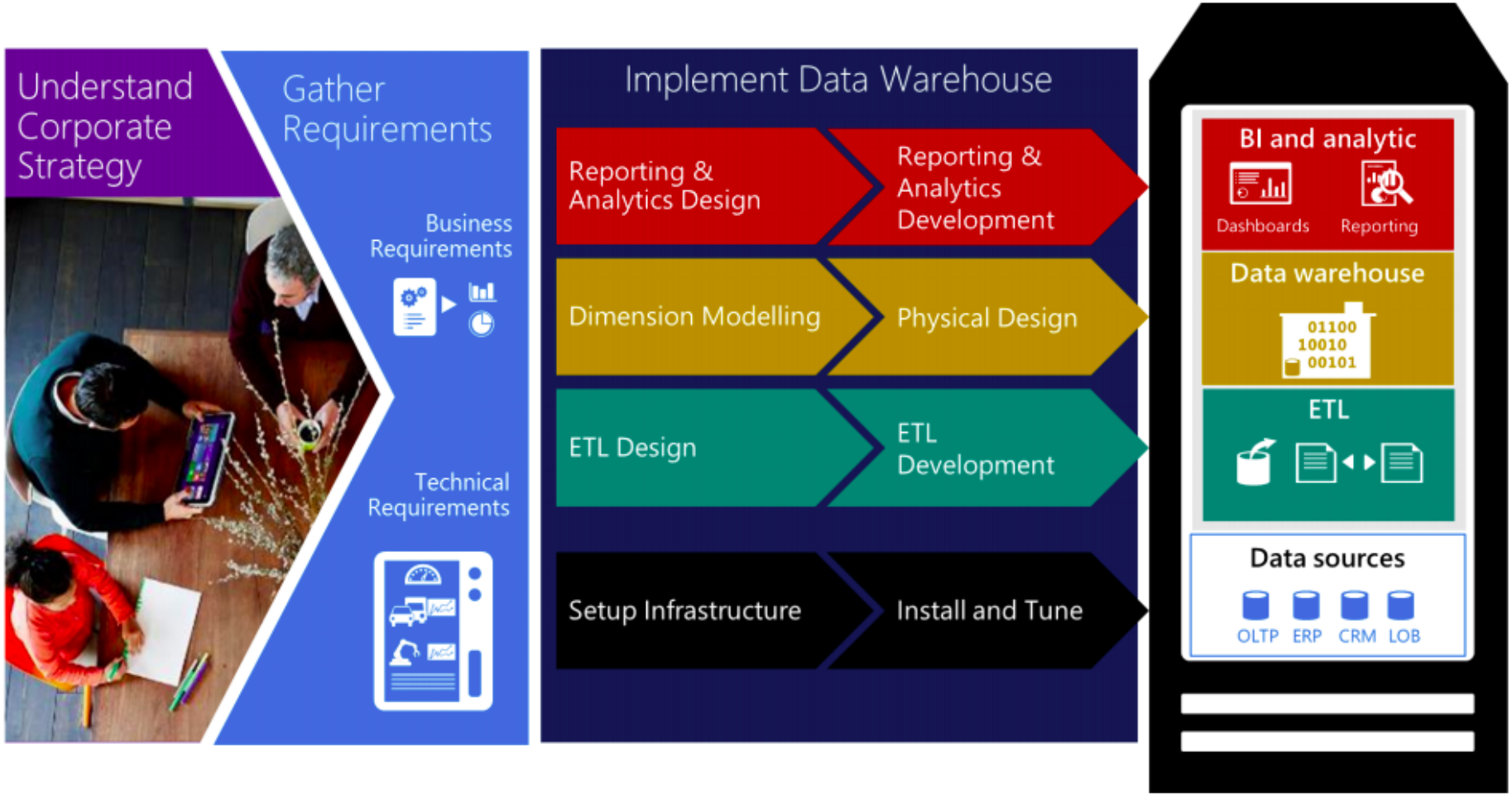
Transformed historical into read structures

Flat, canned or multi-dimensional access to historical data

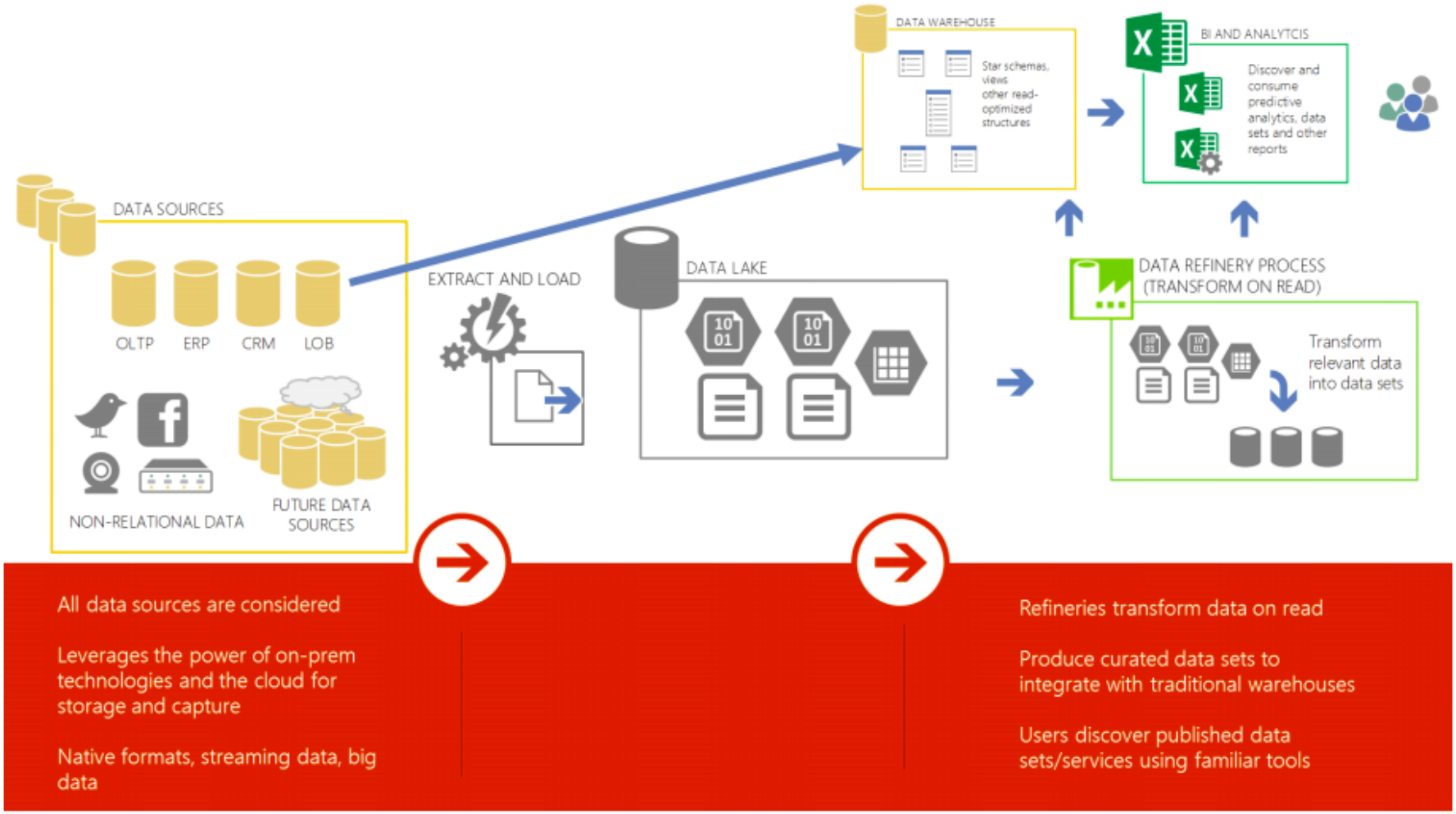
Many reports, multiple versions of the truth

24 to 48h delay

Принцип нисходящего анализа



После больших данных



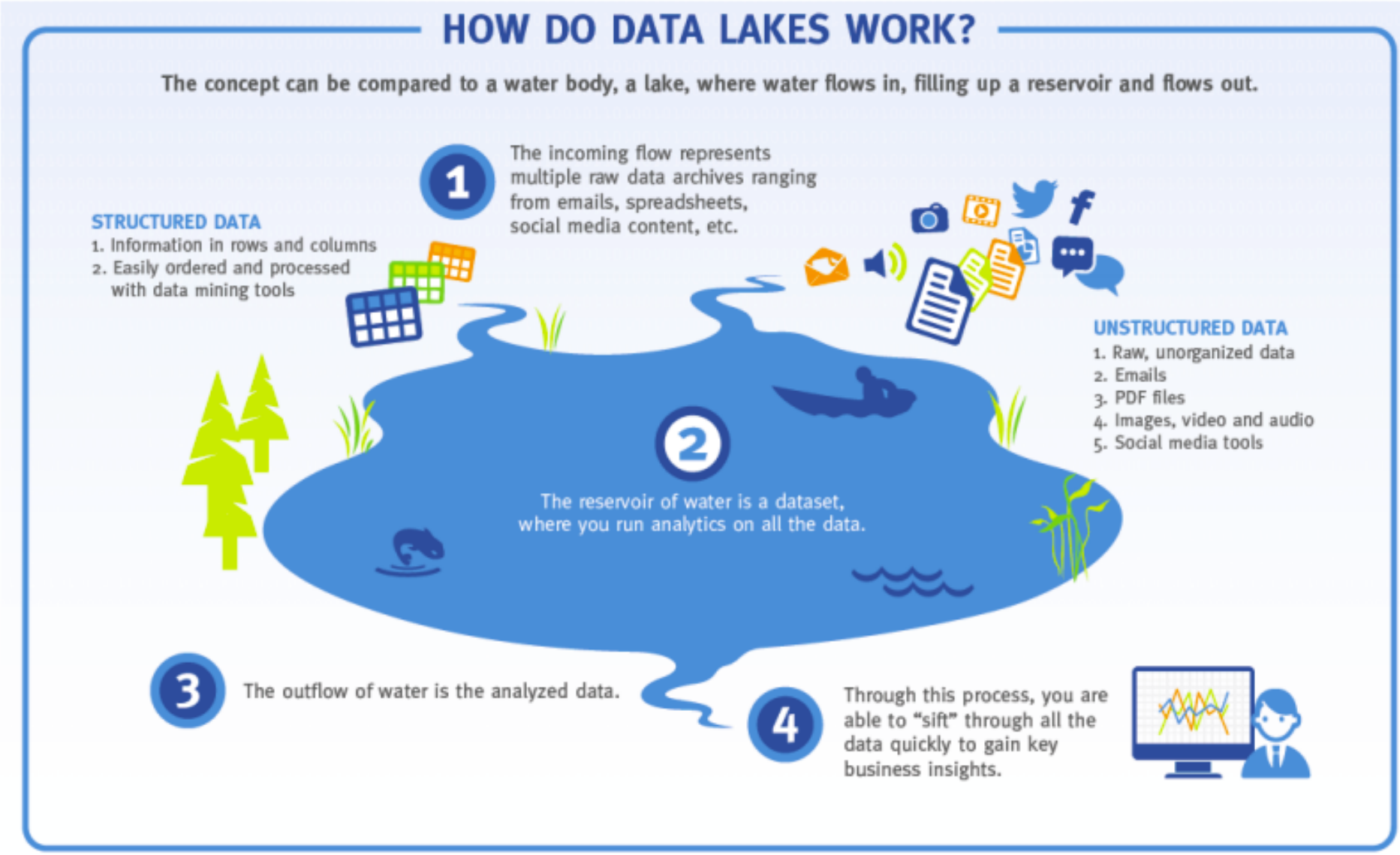
Принцип восходящего анализа

Как свести все эти данные?

Это «озеро» данных, где:

- Входящие потоки – это входные данные, которые могут обладать разной формой и структурой
- Исходящие потоки – это выходные данные, т.е. проанализированные

Что такое озеро данных?

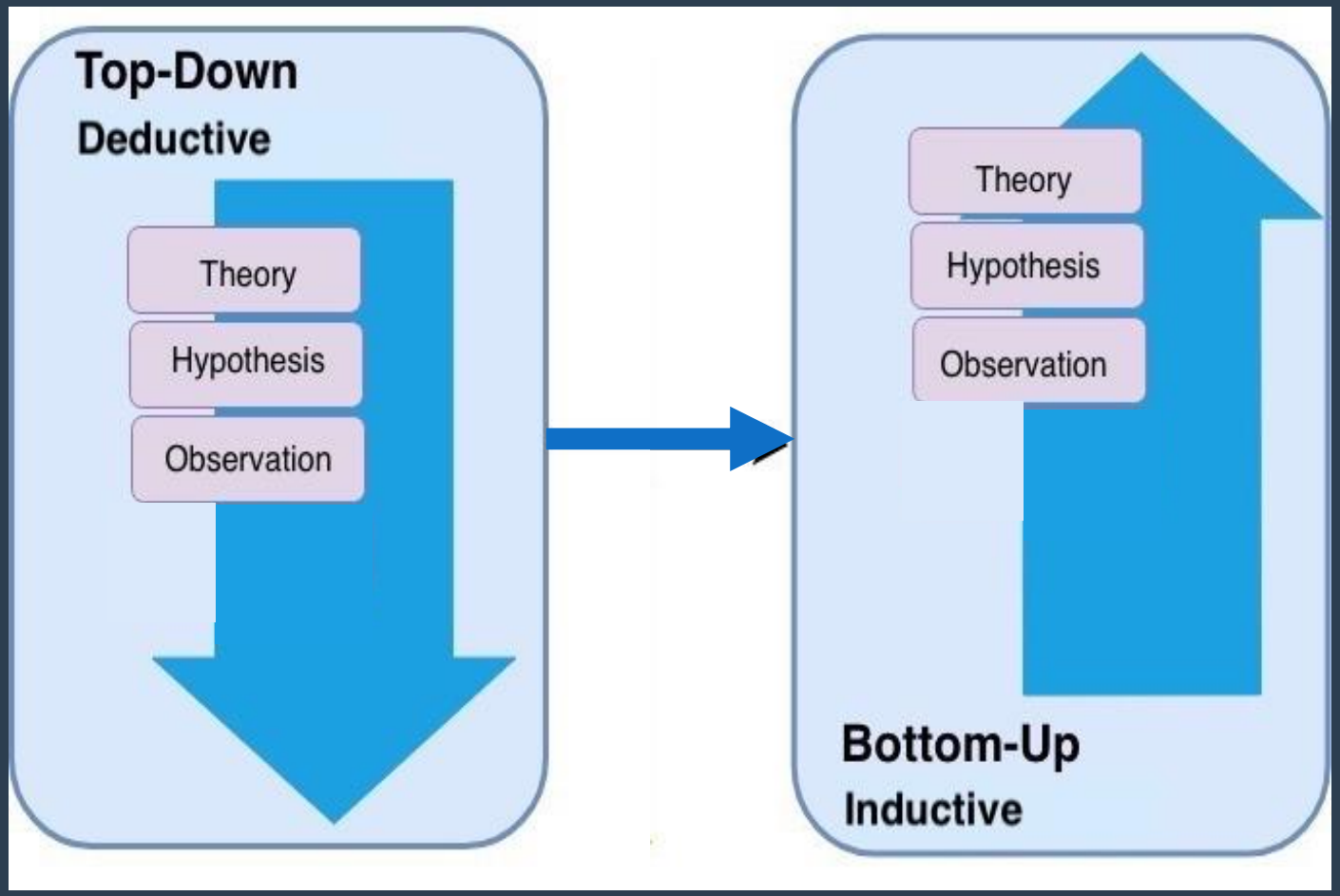


Как сделать так, чтобы машины могли обрабатывать такие объёмы данных без ущерба для производительности?

Увеличение масштаба

Горизонтальное масштабирование (расширение)

Как большие данные изменили аналитику рынка труда?



Какое отношение большие данные и ИИ имеют к рынку труда?

Аналитика рынка труда АРТ

Коротко об АРТ

- Аналитика рынка труда (АРТ, англ. *Labour Market Inelligence/LMI*) – термин, входящий в обиход во всём сообществе специалистов, работающих в области рынка труда, особенно в Европейском союзе.
- Унифицированного определения понятия АРТ не существует. Оно может касаться **дизайна и использования алгоритмов и систем ИИ для анализа данных, связанных с рынком труда (т.е. информации рынка труда), в целях помощи в формировании политики и принятии решений.**

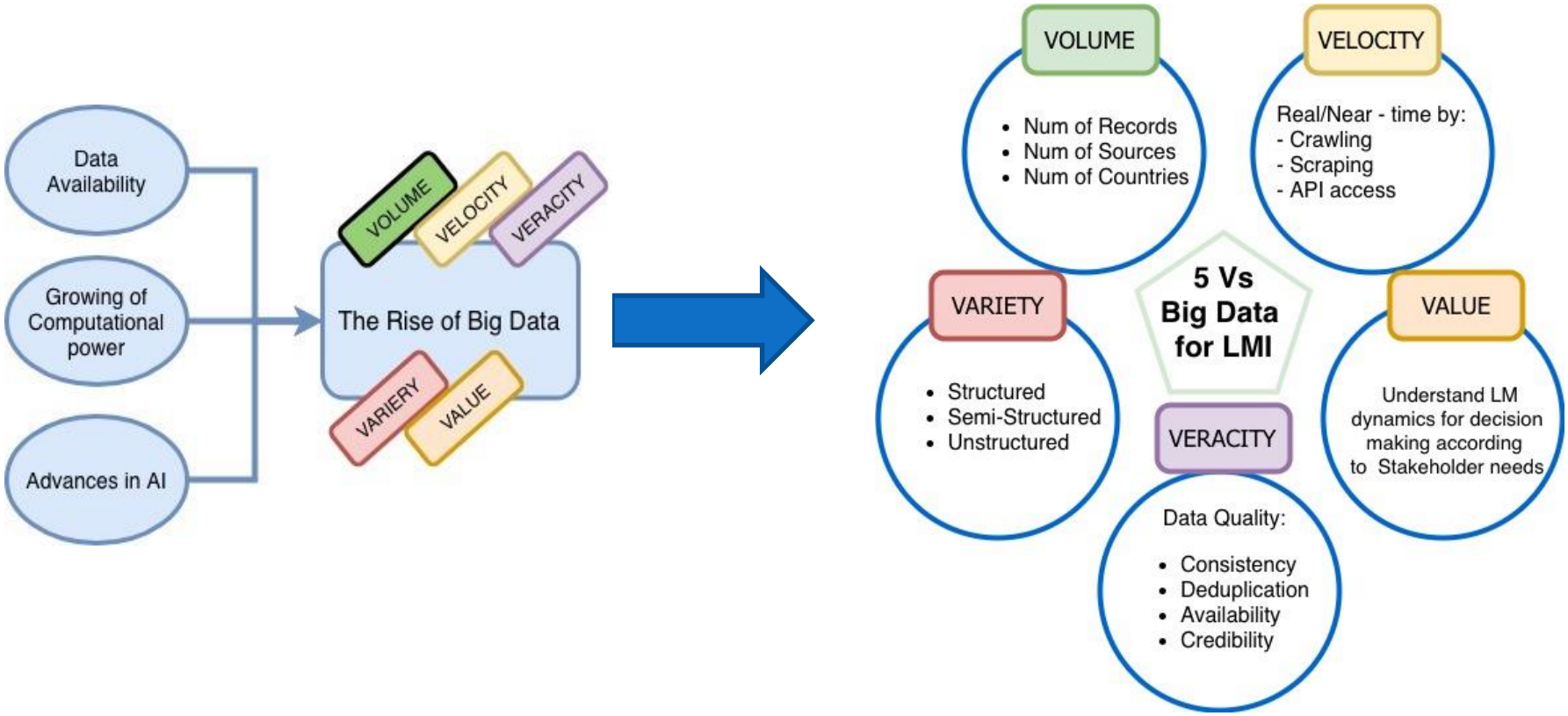
Потребности: новые инструменты для АРТ

- Знаменитое исследование **Фрея и Осборна** (THE FUTURE OF EMPLOYMENT – Будущее сферы занятости, Оксфорд)
  - **47 % профессий исчезнут в течение ближайших 25 лет.**
- 65 % детей, которые сегодня (2017) идут в начальную школу, будут заниматься совершенно новыми видами работ, которых в настоящее время ещё не существует.
- Огромные последствия с точки зрения требований к профессиональным умениям.
  - Пугающие цифры, но насколько они соответствуют действительности?
- Нам нужно применить несколько дополнительных инструментов, чтобы глубже изучить эти изменения.

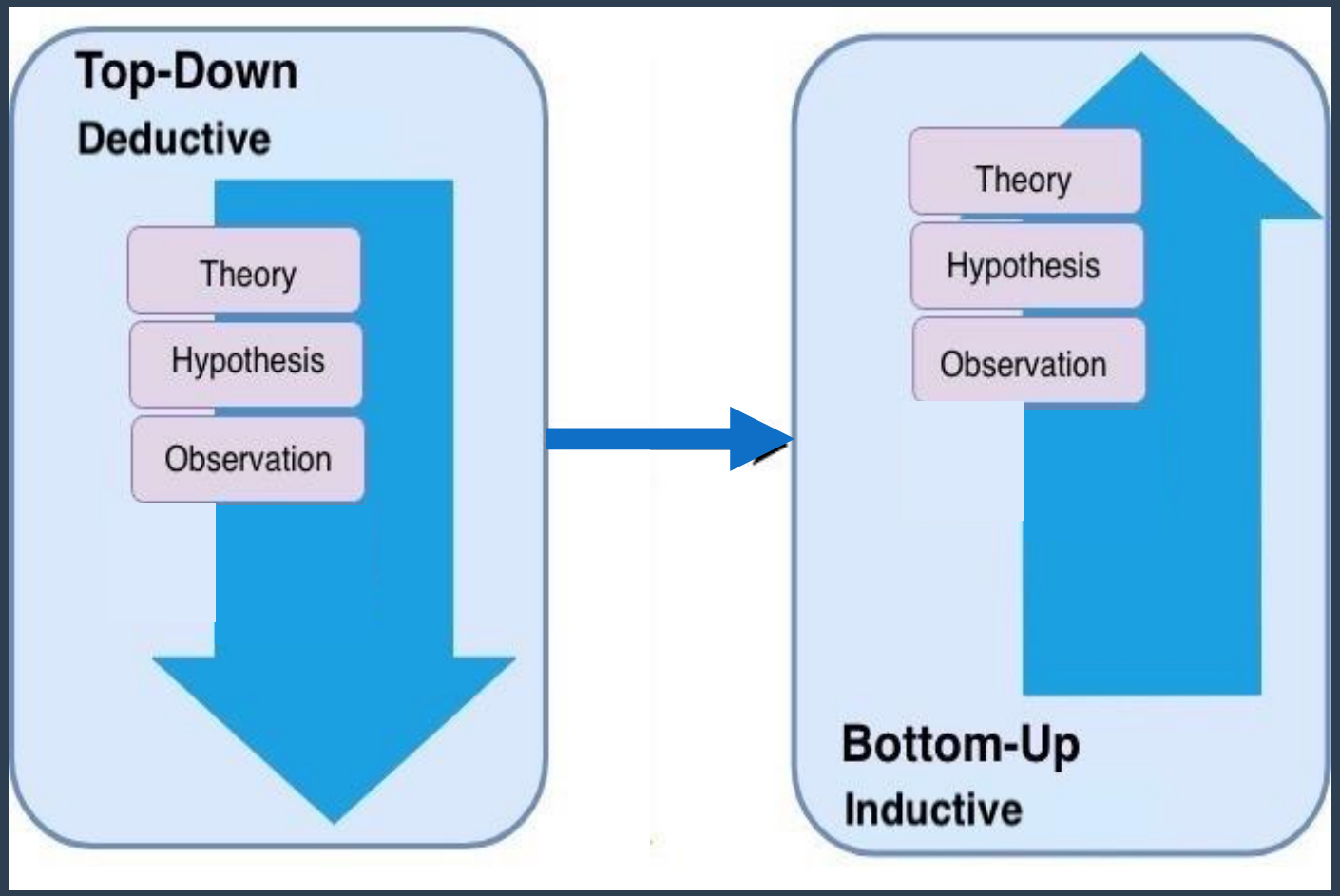
Почему для АРТ следует использовать большие данные?

- Дефицит данных о потребностях работодателей в профессиональных умениях
- Традиционные методы:
  - дорогие,
  - чреваты отставанием во времени,
  - концентрируются на конкретных видах умений,
  - исследования негибкие и требуют времени.
- Инструменты прогнозирования для определения наиболее релевантных тенденций
  - Но инструменты прогнозирования всегда недостаточно точны в отношении требуемых профессиональных умений и характеристик профессий будущего.
- Прогнозирование потребностей в профессиональных умениях
- Полезно:
  - для понимания реальных потребностей рынка
  - для осознанных решений в отношении карьерной мобильности и выбора направления обучения
  - для точной подстройки предложений в сфере обучения

5 «V» больших данных в контексте АРТ



Как большие данные изменили аналитику рынка труда?



Знаменуют ли большие данные прорыв в сфере рынка труда?

Аналитике рынка труда
могут содействовать три главных
источника данных:

(1) статистические источники
(2) административные источники
(3) интернет-источники (большие
данные для АРТ)

К чему идет рынок труда?

ВЫЗОВЫ РТ

1. Эволюция профессиональных умений
2. Новые нарождающиеся профессии
3. Автоматизация работы и замена человека
4. Мобильность

Потребности РТ

1. Обновление информации (почти в реальном времени)
2. Решения на основе данных (пусть говорят данные)
3. Возможность предугадывать тенденции

Знания приобретают критическое значение для разных игроков и разработчиков политики рынка труда, давая возможность понимать его динамику и тенденции

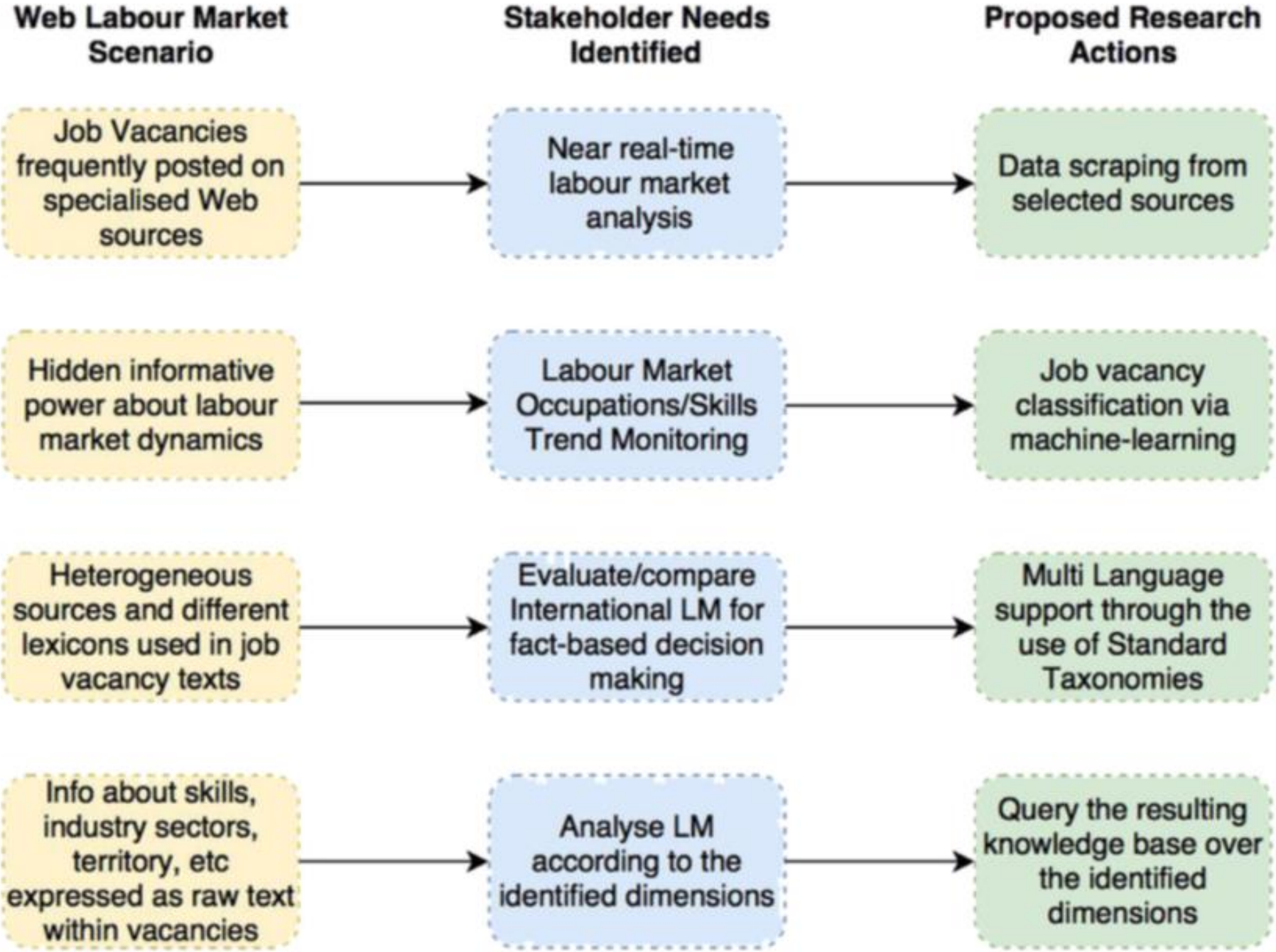

Web Labour Market Scenario
Stakeholder Needs Identified
Proposed Research Actions
Job Vacancies frequently posted on specialised Web sources
Near real-time labour market analysis
Data scraping from selected sources
Hidden informative power about labour market dynamics
Labour Market Occupations/Skills Trend Monitoring
Job vacancy classification via machine-learning
Heterogeneous sources and different lexicons used in job vacancy texts
Evaluate/compare International LM for fact-based decision making
Multi Language support through the use of Standard Taxonomies
Info about skills, industry sectors, territory, etc expressed as raw text within vacancies
Analyse LM according to the identified dimensions
Query the resulting knowledge base over the identified dimensions

Table 1 Main characteristics for LM Data Sources

LM Source Type	*Data Type*[2]	*Generation Rate*	*Data Model Paradigm*	*Quality*	*Coverage*	*Analysis Paradigm*	*Believability*	*Value*
Statistical	Structured	Periodically	Relational	Owner's responsibility	Owner's responsibility	Top Down & Model Based	Owner's responsibility	intrinsic
Administrative	Structured or Semi-structured	Periodically	Relational	Owner's responsibility	Owner's responsibility & User's responsibility	Top Down & Model Based	Owner's responsibility & User's responsibility	intrinsic
Web	Structured, Semi-structured or Unstructured	Near-real-time or real-time	Relational and Non Relational (NoSQL)	User's responsibility	User's responsibility	Bottom up & Data Driven	User's responsibility	extrinsic

Table 2 Most significant limitations of Big Data architectures

Issue (most significant)	Caused by	Conceptual Blocks of Big Data Architectures
Schema-free data are out: only structured data sources can be manipulated. Roughly, this means that only data that obey a rigid, well-defined data model can be handled, to the exclusion of all "unstructured" data, such as free text, comments and Web content in general.	Variety	Data ingestion; NoSQL models;
No adaptability to change: the addition of a new source requires the whole process to change, and this makes it difficult to scale the architecture over multiple (albeit structured) sources.	Variety, Velocity	Data lake
Rigid ETL: the procedures that transform content from source formats to target formats have to be precisely written to fit the desired data structure (e.g., data warehouse).	Variety	Schema free; data-driven approach (bottom-up rather than top-down)
Time consuming: the larger the volume of data to be processed, the greater the time needed to complete the process. ETL procedures are usually high time and memory consumers, as they need to "scan" all the data sources at any time to transform source data.	Volume, Variety, Velocity	Scale-out rather than scale-up