PROАГРО ЛЕКТОРИЙ

Понятие, основные характеристики и источники больших данных в сельском хозяйстве

Демичев Вадим Владимирович,
кандидат экономических наук,
доцент кафедры статистики и кибернетики РГАУ-МСХА имени К. А. Тимирязева

Вопросы лекции

- Понятие большие данные.
- Формат больших данных.
- Методы обработки больших данных.
- Большие данные и Python.
- Источники больших данных в сельском хозяйстве.
- Возможности и трудности использования больших данных в сельском хозяйстве.

Понятие большие данные

Большие данные (Big Data) - это большие массивы данных, отличающиеся главным образом такими характеристиками, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа (ГОСТ Р ИСО/МЭК 20546-2021).

Массивы данных (data set, dataset) определены как идентифицируемая совокупность данных, к которой можно получить доступ или скачать в одном или нескольких форматах.

Понятие большие данные

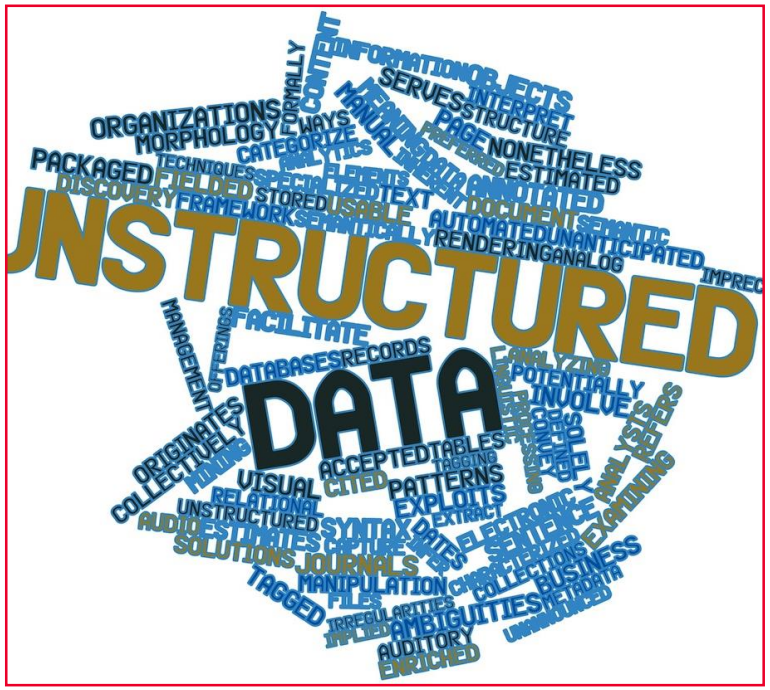
Разнообразие данных (data variety) - предполагает наличие диапазона форматов, логических моделей, временных шкал и семантики массива данных.
Скорость обработки данных (data velocity) - это скорость потока, с которой данные создаются, передаются, сохраняются, анализируются или визуализируются.
Вариативность данных (data variability) - изменения в скорости передачи, формате или структуре, семантике или качестве массива данных.
Достоверность данных (data veracity) предполагает полноту и/или точность данных.

Формат больших данных

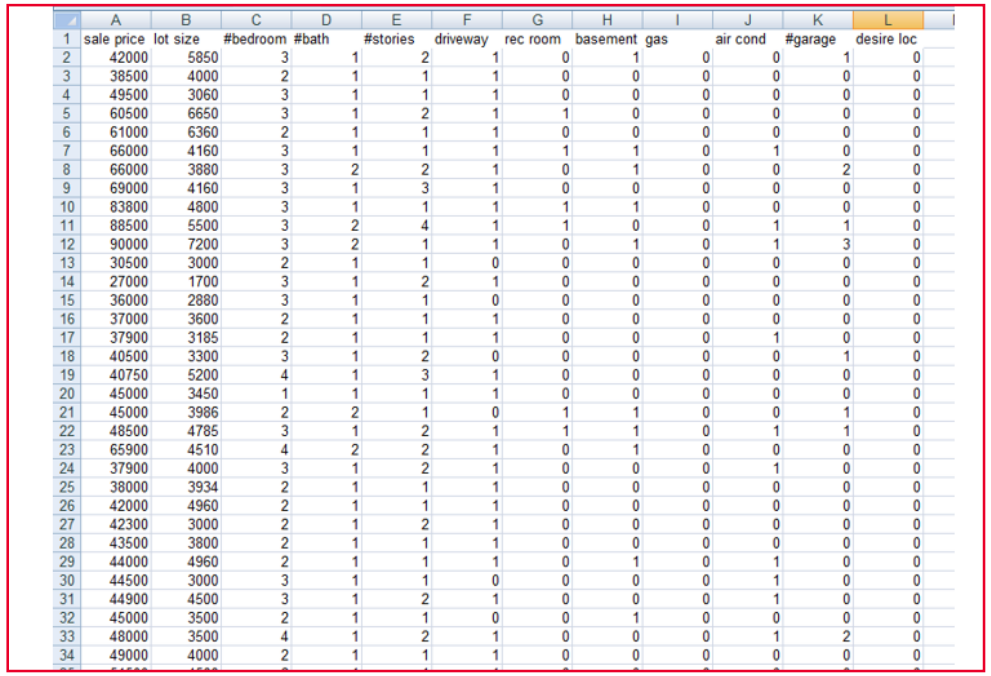
Структурированные данные

Неструктурированные данные

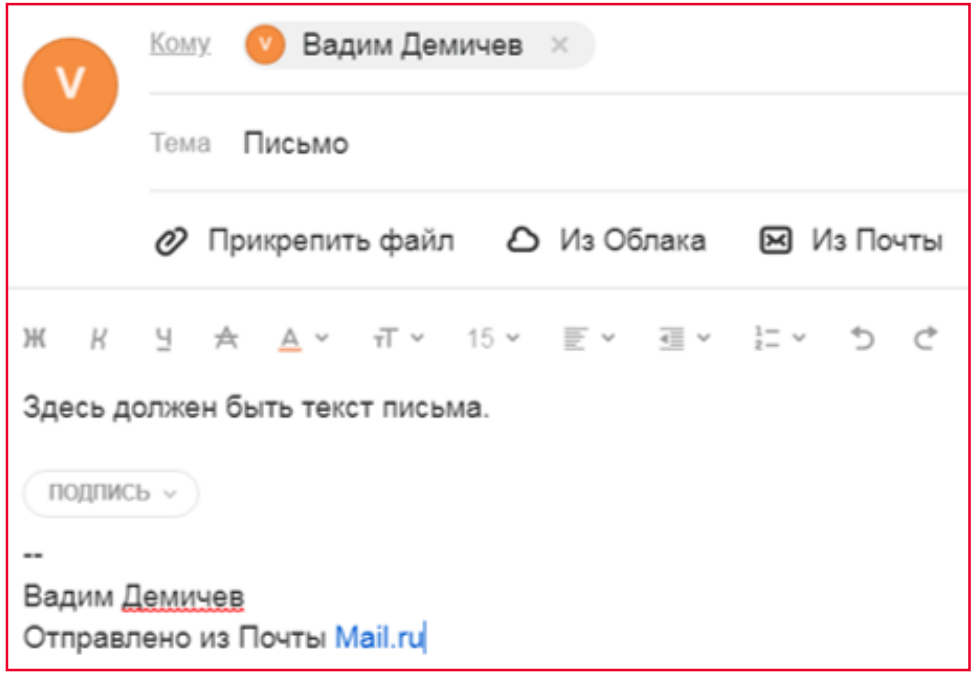
- На естественном языке
- Машинные
- Графовые
- Аудио-, видео- и графика
- Потоковые



Формат больших данных



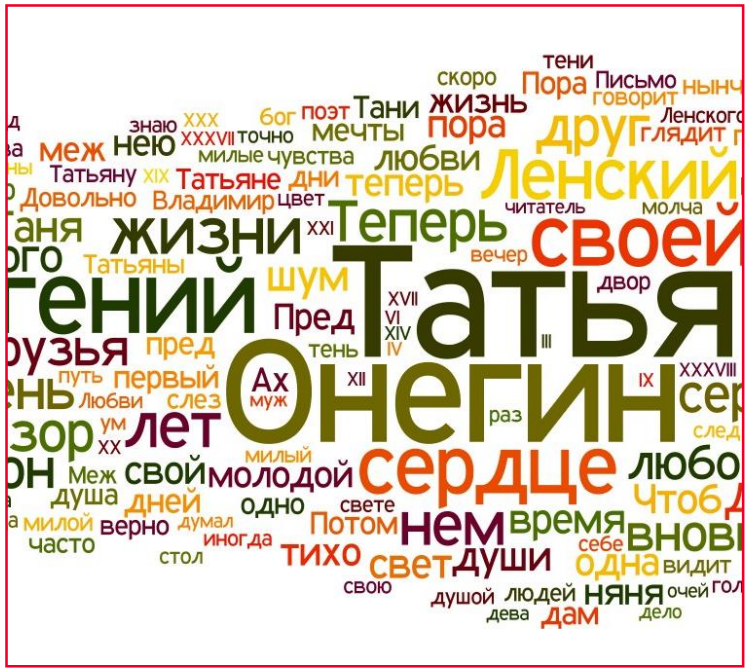
	A	B	C	D	E	F	G	H	I	J	K	L
1	sale price	lot size	#bedroom	#bath	#stories	driveway	rec room	basement	gas	air cond	#garage	desire loc
2	42000	5850	3	1	2	1	0	1	0	0	1	0
3	38500	4000	2	1	1	1	0	0	0	0	0	0
4	49500	3060	3	1	1	1	0	0	0	0	0	0
5	60500	6650	3	1	2	1	1	0	0	0	0	0
6	61000	6360	2	1	1	1	0	0	0	0	0	0
7	66000	4160	3	1	1	1	1	1	0	1	0	0
8	66000	3880	3	2	2	1	0	1	0	0	2	0
9	69000	4160	3	1	3	1	0	0	0	0	0	0
10	83800	4800	3	1	1	1	1	1	0	0	0	0
11	88500	5500	3	2	4	1	1	0	0	1	1	0
12	90000	7200	3	2	1	1	0	1	0	1	3	0
13	30500	3000	2	1	1	0	0	0	0	0	0	0
14	27000	1700	3	1	2	1	0	0	0	0	0	0
15	36000	2880	3	1	1	0	0	0	0	0	0	0
16	37000	3600	2	1	1	1	0	0	0	0	0	0
17	37900	3185	2	1	1	1	0	0	0	1	0	0
18	40500	3300	3	1	2	0	0	0	0	0	1	0
19	40750	5200	4	1	3	1	0	0	0	0	0	0
20	45000	3450	1	1	1	1	0	0	0	0	0	0
21	45000	3986	2	2	1	0	1	1	0	0	1	0
22	48500	4785	3	1	2	1	1	1	0	1	1	0
23	65900	4510	4	2	2	1	0	1	0	0	0	0
24	37900	4000	3	1	2	1	0	0	0	1	0	0
25	38000	3934	2	1	1	1	0	0	0	0	0	0
26	42000	4960	2	1	1	1	0	0	0	0	0	0
27	42300	3000	2	1	2	1	0	0	0	0	0	0
28	43500	3800	2	1	1	1	0	0	0	0	0	0
29	44000	4960	2	1	1	1	0	1	0	1	0	0
30	44500	3000	3	1	1	0	0	0	0	1	0	0
31	44900	4500	3	1	2	1	0	0	0	1	0	0
32	45000	3500	2	1	1	0	0	1	0	0	0	0
33	48000	3500	4	1	2	1	0	0	0	1	2	0
34	49000	4000	2	1	1	1	0	0	0	0	0	0



Структурированные данные – часто хранятся в таблицах, базах данных и таблицах Excel.
Неструктурированные данные сложно подогнать под конкретную модель данных, так как их контент зависит от контекста или имеет переменный характер. Примером неструктурированных данных может служить – обычное сообщение электронной почты.

Формат больших данных

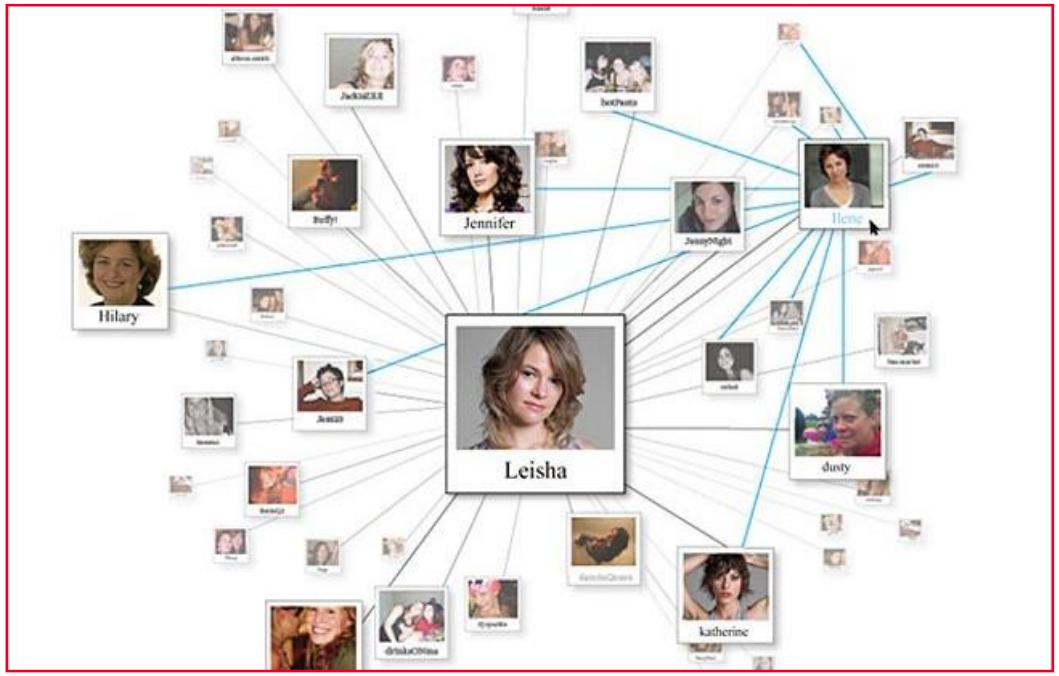
Данные на естественном языке являются разновидностью неструктурированных данных; обработка таких данных достаточна сложна и требует знания как лингвистики, так и методов анализа данных.



Формат больших данных

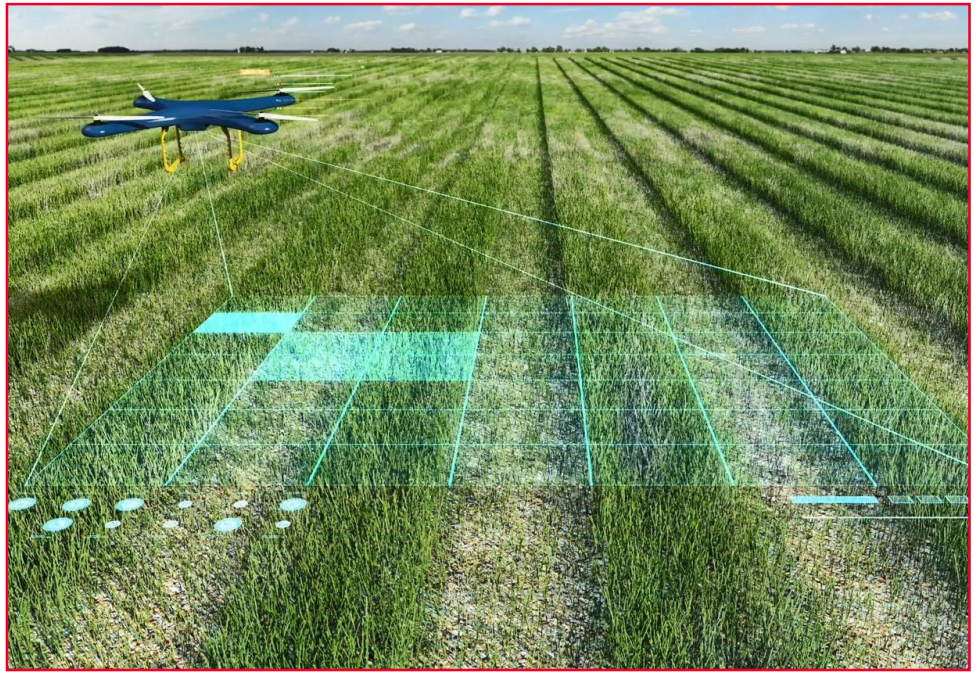
К машинным данным относится информация, автоматически генерируемая компьютером, процессором, приложением или устройством без участия человека. Этот тип данных во многом формируется в сфере так называемого интернета вещей.

Формат больших данных



Графовые данные предполагают представление данных в виде графа. Граф – математическая структура для моделирования попарных отношений между объектами.

Формат больших данных



Аудио-, видео- и графика – наиболее трудно анализируемые типы данных. Задача, кажущаяся простой для человека (например, распознание объекта на картинке), оказывается сложной для компьютера.

Формат больших данных

Потоковые данные – могут принимать любую из перечисленных форм, но имеют одну отличительную особенность – данные поступают в систему при наступлении определенных событий, а не загружаются в базы данных большими массивами.

Методы обработки больших данных

- **Нейросетевой анализ** основан на построении математической модели нейронной сети. Нейронная сеть представляет собой метод в искусственном интеллекте, который учит компьютеры обрабатывать данные таким же способом, как и человеческий мозг.

- **Интеллектуальный анализ данных (Data Mining)** – это процесс обнаружения закономерностей в наборе данных и прогнозирования возможных значений изучаемого показателя; также известен как обнаружение знаний в базах данных.

- **Машинное обучение** представляет собой процесс обучения компьютера подобно тому, как это делают люди. Благодаря машинному обучению компьютеры учатся определять вероятности и делать прогнозы на основе данных.

Большие данные и Python

Pandas - программная библиотека на языке Python для обработки и анализа данных.
Matplotlib - библиотека для визуализации данных двумерной (2D) и 3D графикой.
NumPy - библиотека, добавляющая поддержку больших многомерных массивов и матриц, вместе с большой библиотекой высокоуровневых (и очень быстрых) математических функций для операций с этими массивами.
Scikit-learn – библиотека для задач Data Science и Machine Learning. Содержит функции и алгоритмы для машинного обучения: классификации, прогнозирования или разбивки данных на группы.

Большие данные и Python

Seaborn, Statmodels, TensorFlow, Keras, PyMC3, Plotly, Altair, Geoplotlib, Gensin, NLTK, Natasha, BeautifulSoup, Feather, Ibis, ParaText, Bcolz, Blaze, Xarray, Dask.

Источники больших данных в сельском хозяйстве

- Датчики, установленные на растения или технику, формирующие поток данных.
- Применение цифровой ушной бирки для КРС, позволяющее контролировать температуру и активность животного.
- «Умные» теплицы, где температура, уровень влажности и другие параметры регулируются автоматически.
- «Умные» фермы используют интернет-вещей, чтобы отслеживать передвижение и здоровье животных.
- Беспилотное управление комбайнами, тракторами и другой техникой.
- Использование дронов, контролирующих уровень влажности, света и силу ветра, а также осуществляющих мониторинг роста растений, распространение сорняков и других параметров.
- Технологии интернета вещей (IoT) - использование специальных метеостанций, с помощью сенсоров собирающих данные (температура, влажность).

Возможности и трудности использования больших данных в сельском хозяйстве

Возможности/ Трудности	Вид возможности/ трудности	Описание возможности/трудности
Возможности	Функциональные	- Выявление новых зависимостей, моделей поведения, рисков, классификаций - Предиктивная аналитика позволит осуществлять более точное прогнозирование - Прескриптивная аналитика позволит повысить качество принимаемых управленческих решений
	Экономические	- Повышение производительности труда и снижение издержек - Улучшение операционной эффективности через автоматизацию процессов - Оптимальное бизнес-планирование производства продукции (потребность в кормах, удобрениях, средствах защиты и так далее) - Улучшение взаимодействия между поставщиками, производством и потребителями
	Экологические	- Рациональное использование ресурсов - Минимизация отходов производства
	Социальные	- Трансформация сельского хозяйства, основанного на традиционных технологиях, к цифровому сельскому хозяйству - Появление новых, современных рабочих мест
	Технологические	- Способность иметь дело с данными большой величины, из разнородных источников, формирующимися в режиме онлайн или около этого, с дальнейшей возможностью анализировать их и принимать управленческие решения
Трудности	Организационные	- Децентрализация сбора и хранения данных - Контроль данных при множестве пользователей - Монетизация больших данных
	Социальные	- Убедить крупный и средний бизнес в ценности больших данных и обоснованности вложение средств в такого рода инновации - Изучение этических последствий больших данных в сельском хозяйстве - Поиск специалистов, способных анализировать большие данные
	Технологические	- Обеспечение технической возможности реализации «7V»

PROАГРО
ЛЕКТОРИЙ

Спасибо за внимание!