



**Прогностика на основе ИИ
в мониторинге:
задачи и методы, мифы и компромиссы**

Докладчик: к.т.н., доц. Алексей Незнанов

Senior Data Scientist, компания «Шлюмберже»

старший научный сотрудник МЛ ИССА ФКН НИУ ВШЭ

2023-10-19



О докладчике

- Алексей Андреевич Незнанов
 - К.т.н., доцент
 - Член *IEEE*, РАИИ и НБМЗ
 - *Senior Data Scientist* в компании «Шлюмберже»
 - Старший научный сотрудник международной лаборатории интеллектуальных систем и структурного анализа ФКН НИУ ВШЭ (<http://cs.hse.ru/ai/issa>)
 - Консультант «маленького гида по большим данным» на Постнауке (http://postnauka.ru/author/a_neznanov)
 - Автор учебника, учебных пособий, 11 авторских курсов и более 65 научных публикаций





Motivation

- Мониторинг – это временные ряды событий и сенсоров:
 - Анализ подобных временных рядов – составляющая любых систем мониторинга
 - Режимы работы
 - Предвестники событий
 - Пропущенные значения и их восстановление
 - ...
- Мы наблюдаем тектонический сдвиг в методах и инструментах
 - За счёт чего стал возможным этот сдвиг?
 - Можно ли надеяться на практичность современных подходов и методов?
 - Насколько упростилась подготовка данных и интерпретация результатов?
 - И чего ожидать в дальнейшем?



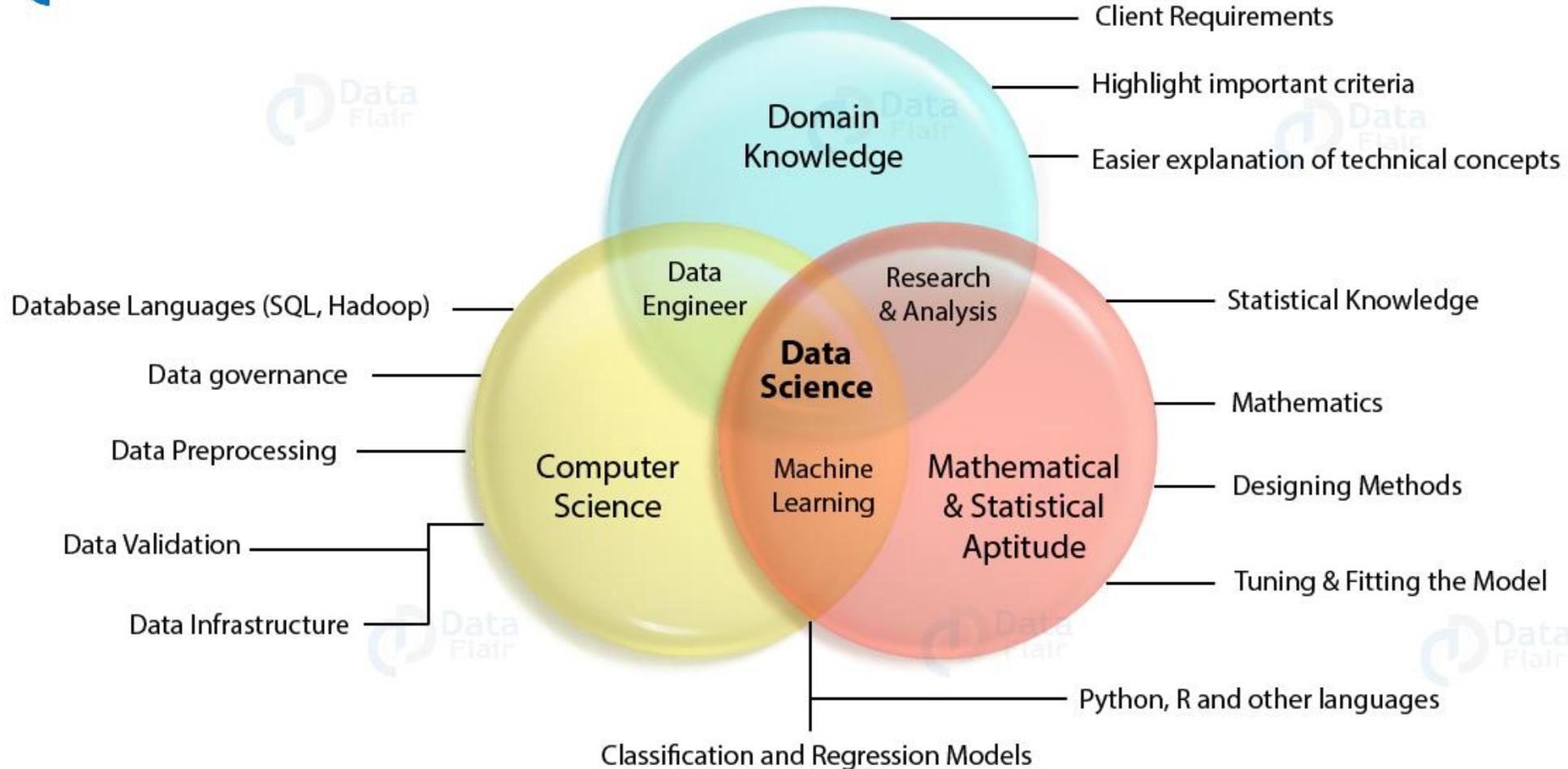
«Пререквизиты» любого анализа данных

«Мусор на входе – мусор на выходе» (“*Garbage in – garbage out*”)

- Качество данных определяет качество интеллектуальных методов анализа данных
 - Особенно при использовании комплексных процедур анализа и машинного обучения
- Управление данными
 - Управление мастер-данными (*MDM*) и управление качеством данных (*DQM*)
- Хоть чуть-чуть работы со «знанием»
 - Базовая онтология предметной области и интерпретация результатов анализа
 - Связь с мастер-данными
 - Подходы к гармонизации данных:
 - Идентификация, дедупликация, онтологизация, актуализация, версионирование, ...



Области знания и компетенции

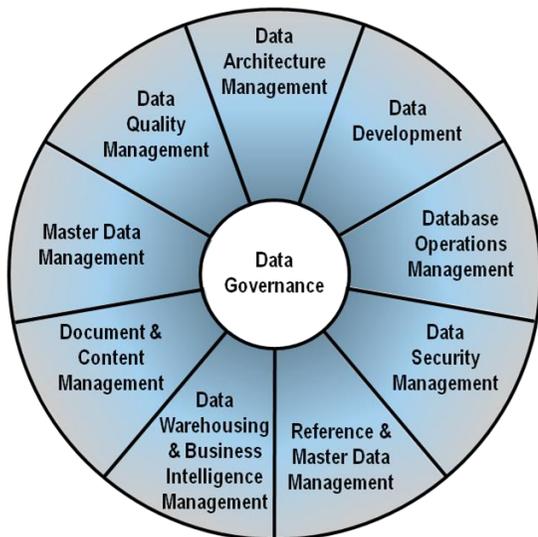
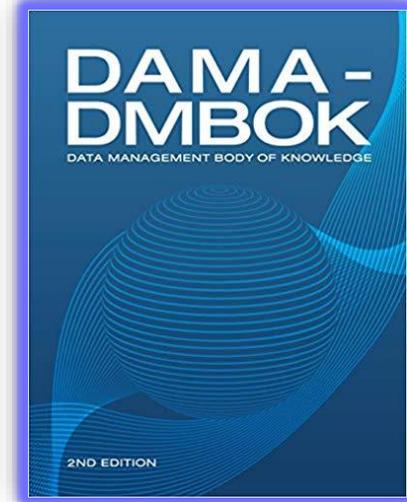


© 2019, data-flair (<http://data-flair.training/blogs/what-is-data-science/>)



Управление данными

- *DAMA-DMBOK: Data Management Body of Knowledge. 2nd ed. Technics Publications, 2017. 590 p.*
 - Свод знаний (BoK)! Готовится третье издание
- *DAMA International* (<http://dama.org/content/body-knowledge>)

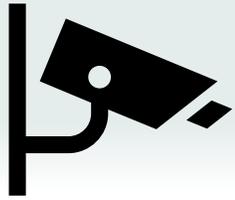
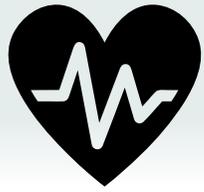


10 Data Management Functions



7 Elements





Временные ряды (ВР)

- ✓ Что это такое?
 - ✓ Что в них интересного?
 - ✓ Какие задачи ставятся?
 - ✓ В чём сложность анализа?
-



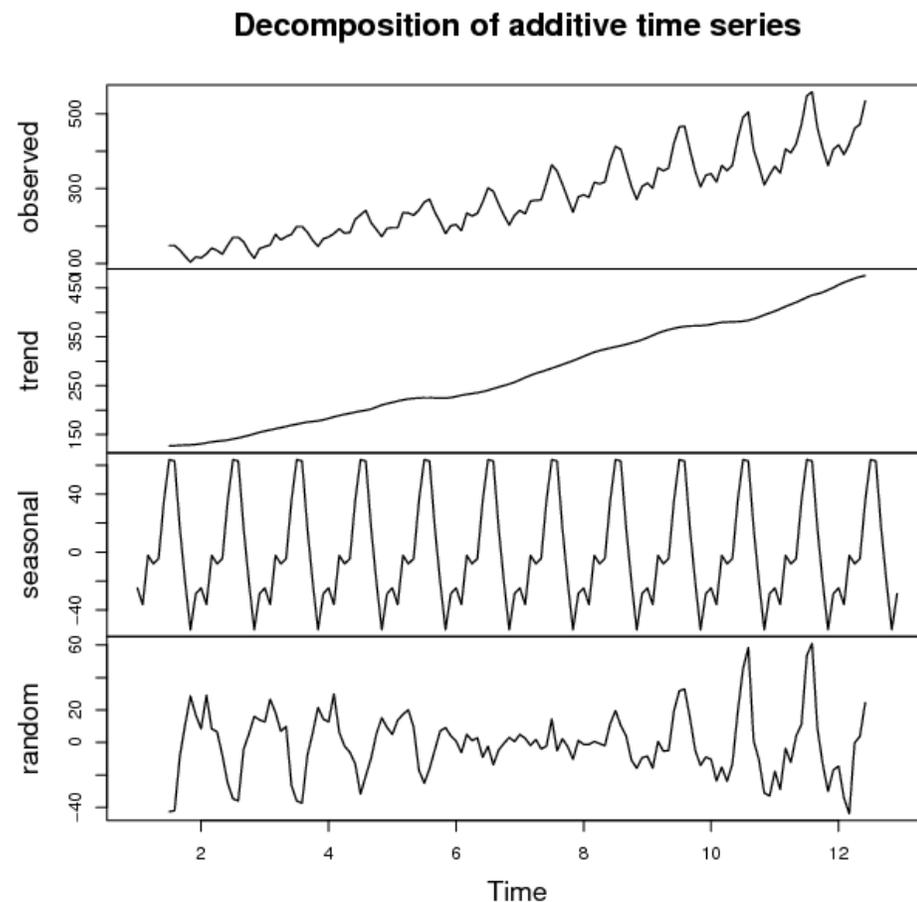
Базовые источники

- The Complete Guide to Time Series Analysis and Forecasting (<http://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775>)
 - Базовые принципы и иллюстрации
- Time Series Analysis Articles (<http://www.quantstart.com/articles/topic/time-series-analysis/>)
- С самого начала до ARIMA и GARCH (но финансовые приложения)
- Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice (<http://otexts.com/fpp3/>)
 - Один из лучших учебников с примерами на *R*



Как обычно думают о ВР

- Исходные данные, привязанные к отметкам времени
- Интересные свойства
 - Скользящее среднее
 - Сглаживание
 - Тренд
 - Сезонность
 - Автокорреляция
 - Периодические изменения
 - Запаздывание
 - ...



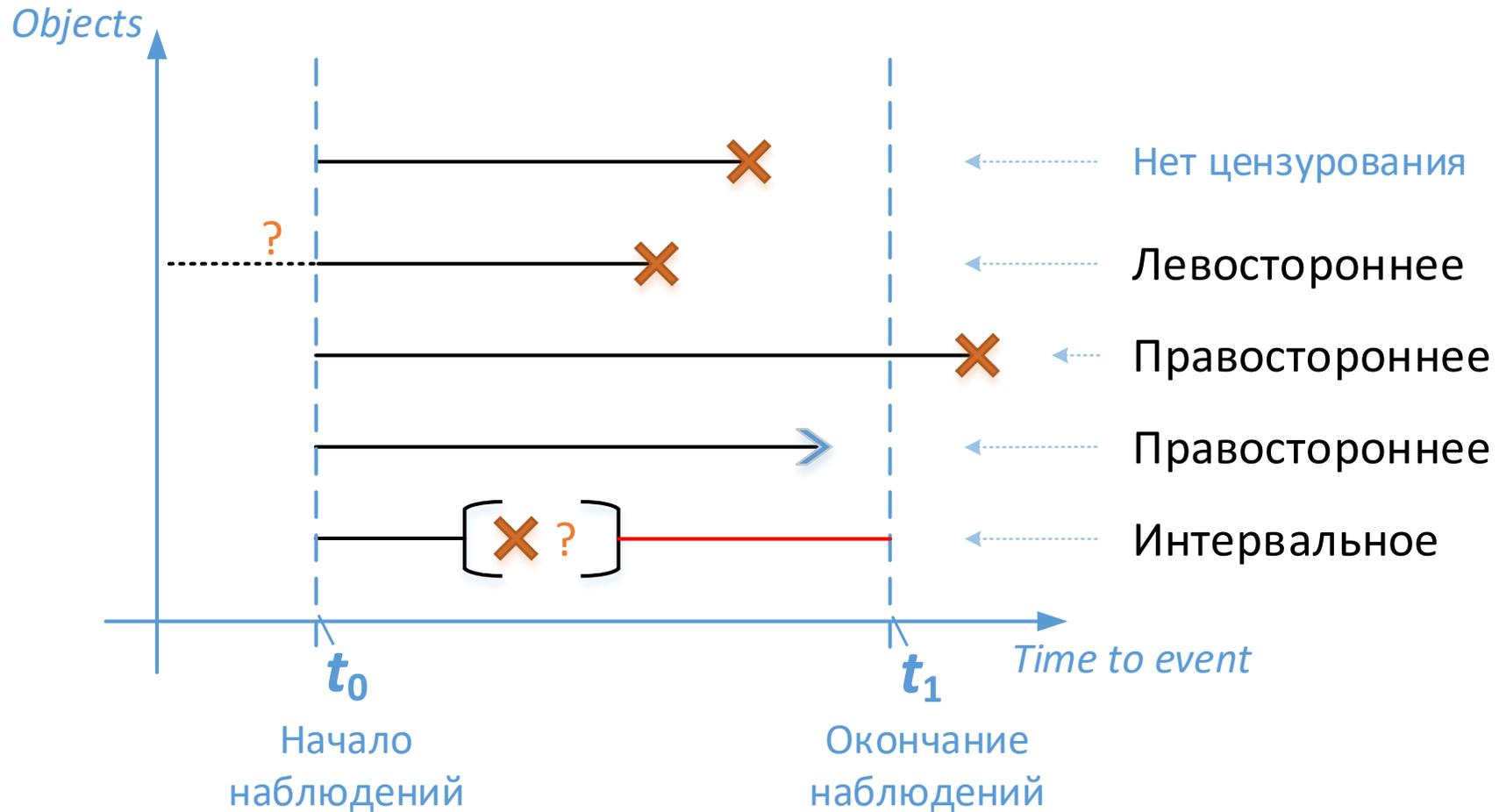


А как на практике?

- Временные ряды обычно:
 - Многомерные
 - Несколько параметров
 - Нерегулярные
 - Произвольные отметки времени
 - С цензурированными данными
 - Куски известных временных рядов не совпадают с жизненным циклом объектов
 - Пропуски
 - Часть данных неизвестна
 - Нечёткость
 - Данные введены и сохранены с погрешностью



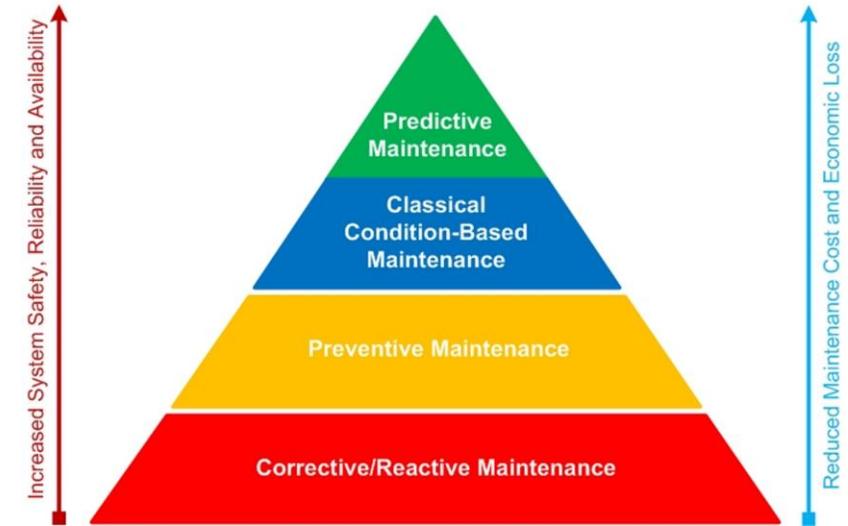
Иллюстрация цензурирования ВР





Прогностика

- *PHM – Prognostics and Health Management*
 - Но есть и другое популярное сокращение:
 - *PHM – Proportional Hazard Model* в анализе выживаемости
- Особенности
 - Оборудование [*equipment*] и сенсоры [*sensors*]
 - Остановки [*stops*] и отказы [*breakdowns*]
 - Принятие решений на основе данных
 - Данные – история работы оборудования
- Свежий альманах для понимания области:
 - *Maintenance Management – Current Challenges, New Developments, and Future Directions*, 2023 (<http://www.intechopen.com/books/11528>)



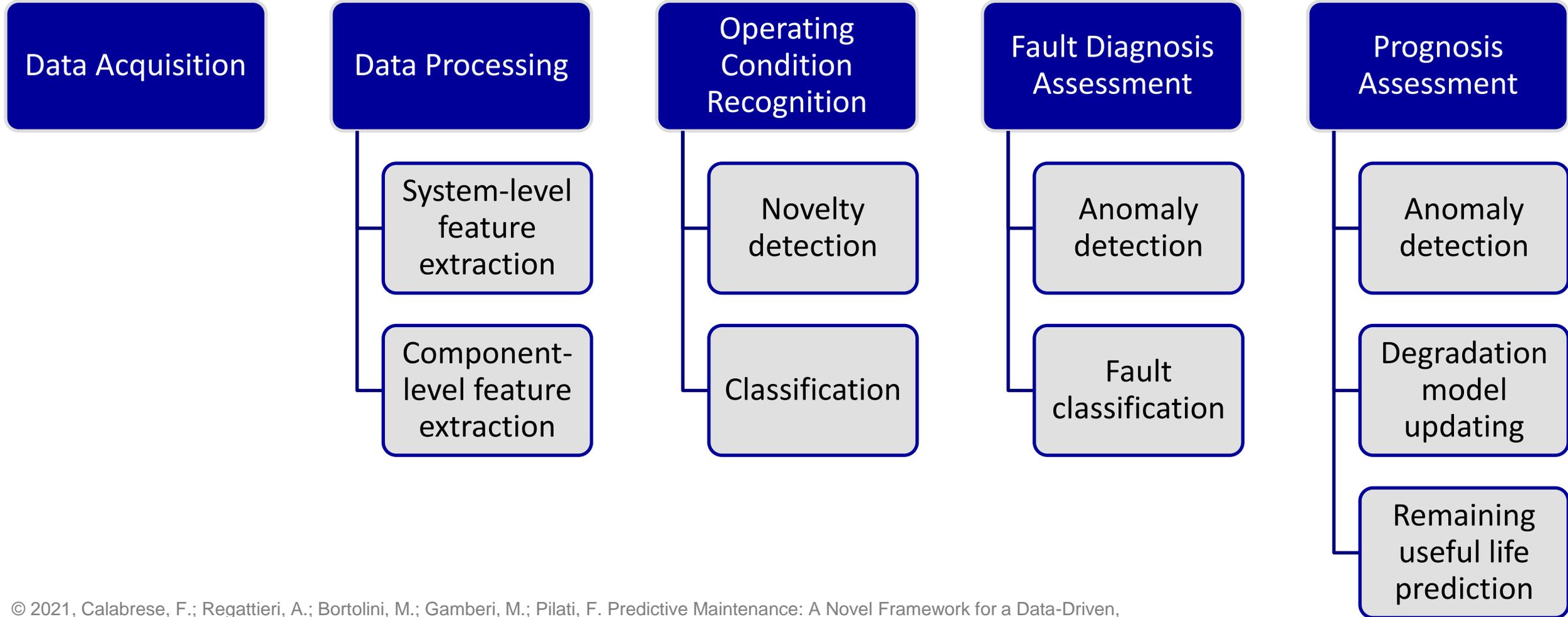


Многообразие задач в прогностике

- Прогнозирование ВР как такового [*forecasting*]
- Классификация/кластеризация ВР [*classification/clustering*]
- Обнаружение **аномалий** ВР [*anomaly detection*]
- Обнаружение **изменений в поведении** ВР [*change points detection*]
 - Крупная подобласть *Data Drift*
- Обнаружение **предвестников** событий [*precursors*]
 - Основа «раннего предупреждения», см. *Early warning system (EWS)*
- **Восстановление** значений ВР [*imputation*]
- Оценка **горизонта** прогнозирования [*horizon assessment*]
- ...



Пример классификации подзадач

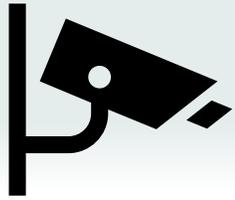


© 2021, Calabrese, F.; Regattieri, A.; Bortolini, M.; Gamberi, M.; Pilati, F. Predictive Maintenance: A Novel Framework for a Data-Driven, Semi-Supervised, and Partially Online Prognostic Health Management Application in Industries. Appl. Sci. 2021, 11, 3380



Тезисы

- С середины 2010-х годов произошёл тектонический сдвиг, причём не только в математике, но и в инструментах
- Многие новые методы и их реализации достойны внедрения
 - Для многих подзадач – сплошное машинное обучение
 - Далее раскроем
- Однако не стоит забывать о компромиссах
 - Качество/скорость
 - Качество/стоимость
- Также не стоит забывать о **качестве данных** и особенностях хранения
- В любом случае нужны адекватные **ETL-процедуры** и предобработка



Методы

- ✓ Подходы
 - ✓ Пути развития
 - ✓ Вехи
 - ✓ Интересные примеры
-



Современное представление о задачах

- Time Traveling with Data Science
 - Focusing on Change Point Detection in Time Series Analysis (Part 2)
(<http://www.iese.fraunhofer.de/blog/change-point-detection/>)

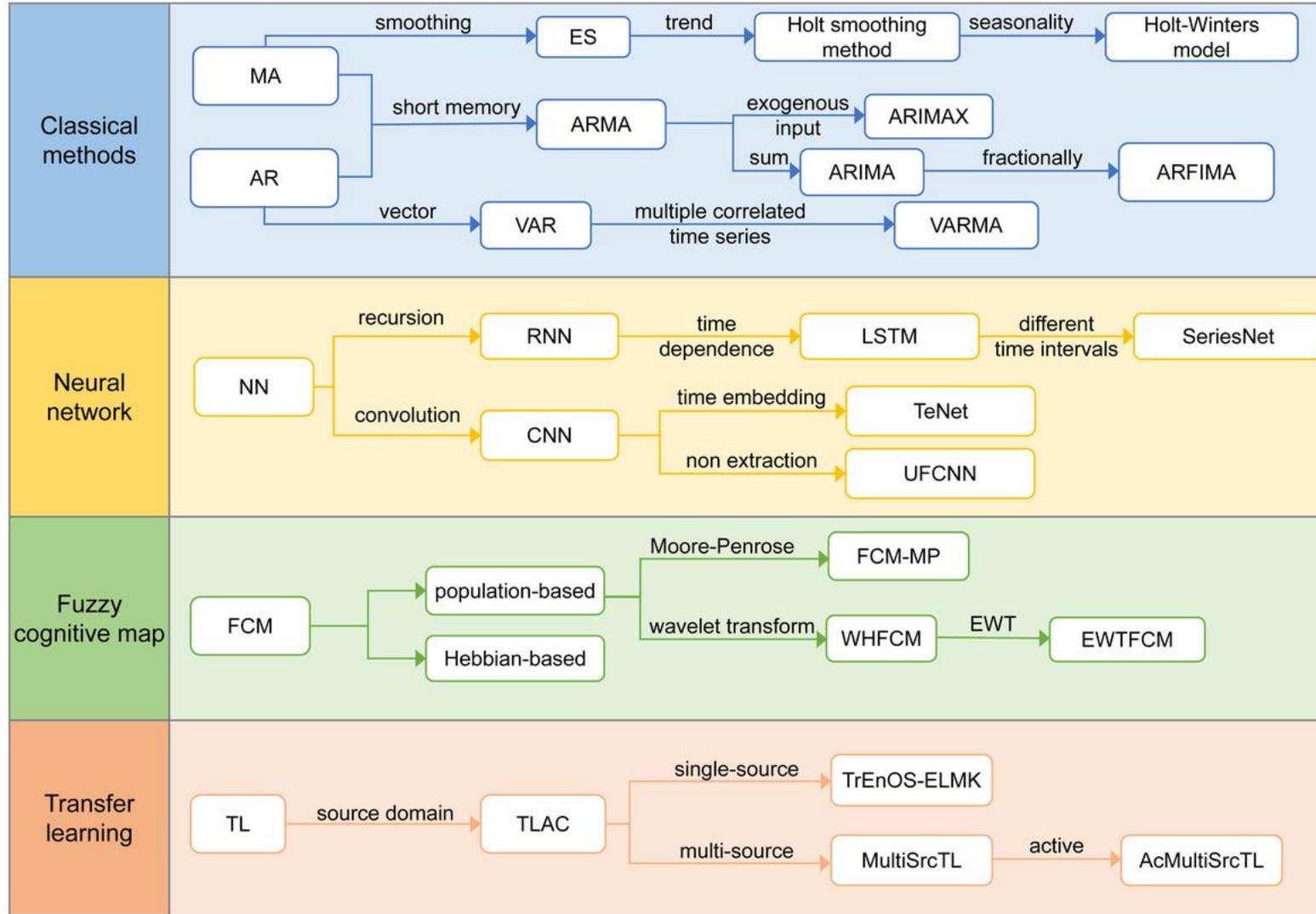


Многообразие подходов и методов

- Tong Y, Liu J, Yu L, Zhang L, Sun L, Li W, Ning X, Xu J, Qin H, Cai Q. 2022. Technology investigation on time series **classification** and **prediction**. PeerJ Computer Science 8:e982 (<http://peerj.com/articles/cs-982/>)

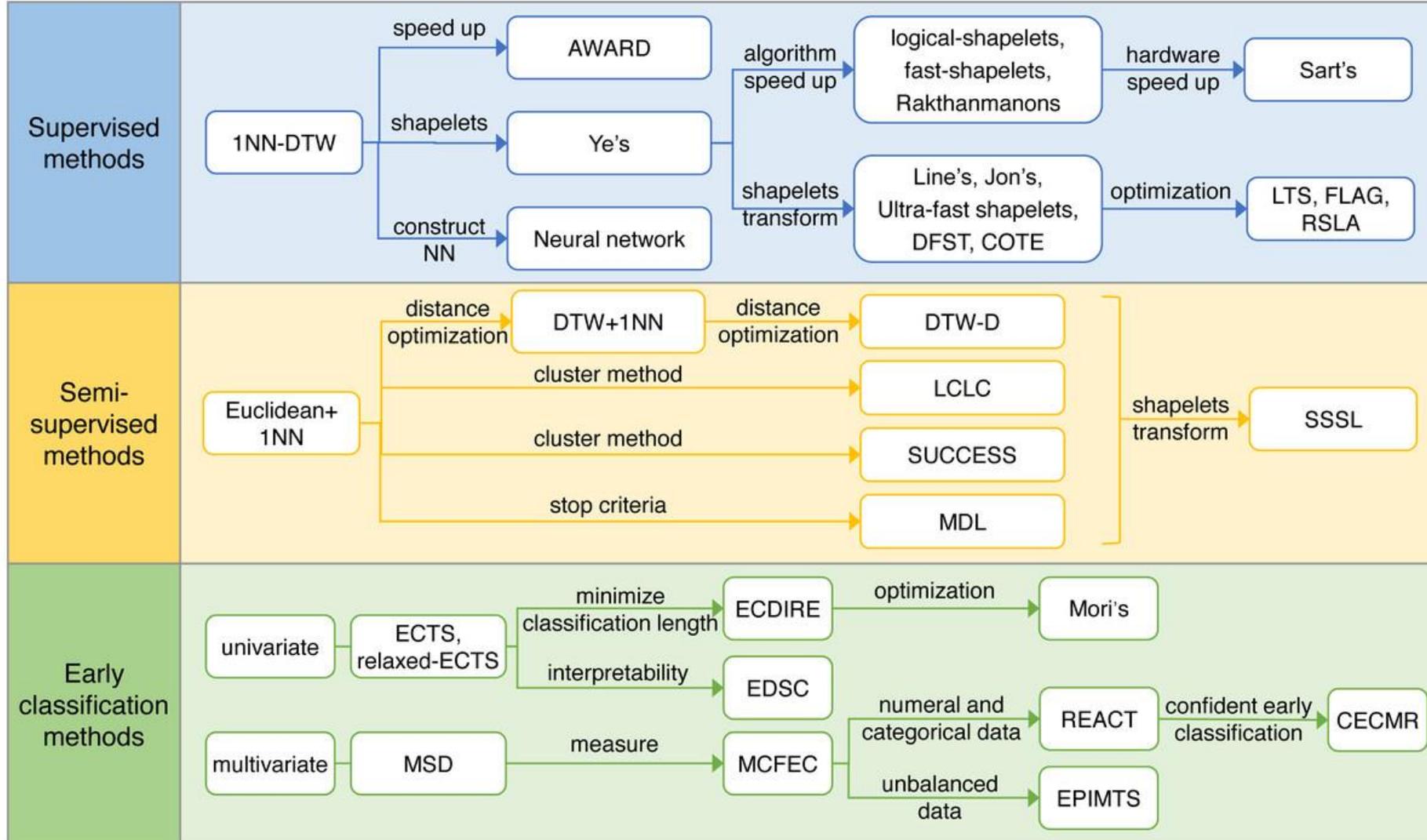


Развитие подходов к прогнозированию





Развитие подходов к классификации



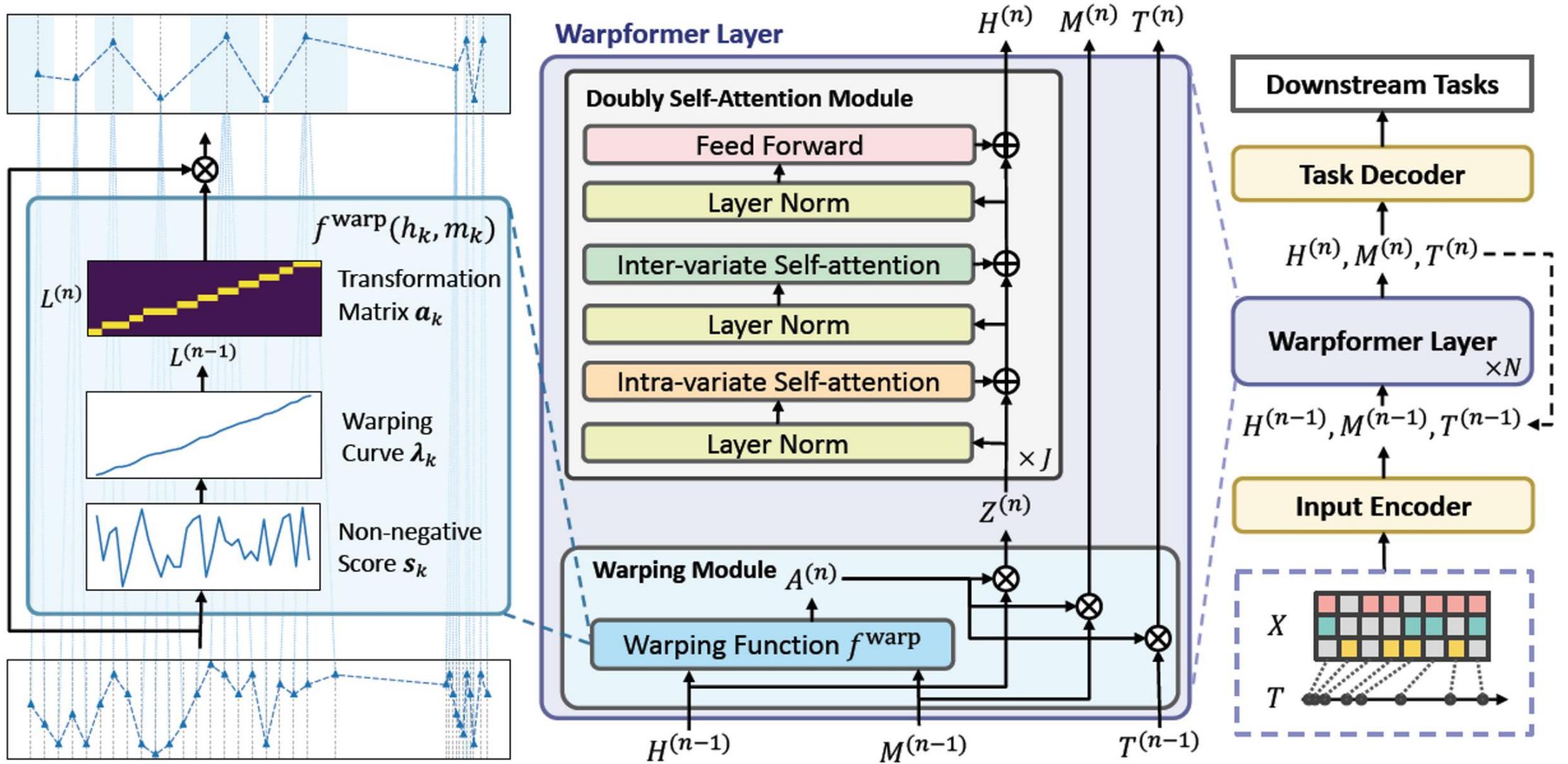


Работа с нерегулярными ВР

- Достаточно полный, но уже устаревший обзор:
 - *Narayan Shukla, Satya; Marlin, Benjamin M. A Survey on Principles, Models and Methods for Learning from **Irregularly Sampled Time Series**. eprint arXiv:2012.00168, 2020*
(<http://arxiv.org/abs/2012.00168>)
- Пример последних достижений:
 - *Zhang, Jiawen; Zheng, Shun; Cao, Wei; Bian, Jiang; Li, Jia. **Warpformer**: A Multi-scale Modeling Approach for Irregular Clinical Time Series. eprint arXiv:2306.09368, 2023*
(<https://arxiv.org/abs/2306.09368>)



Архитектура нейромодели *Warpformer*





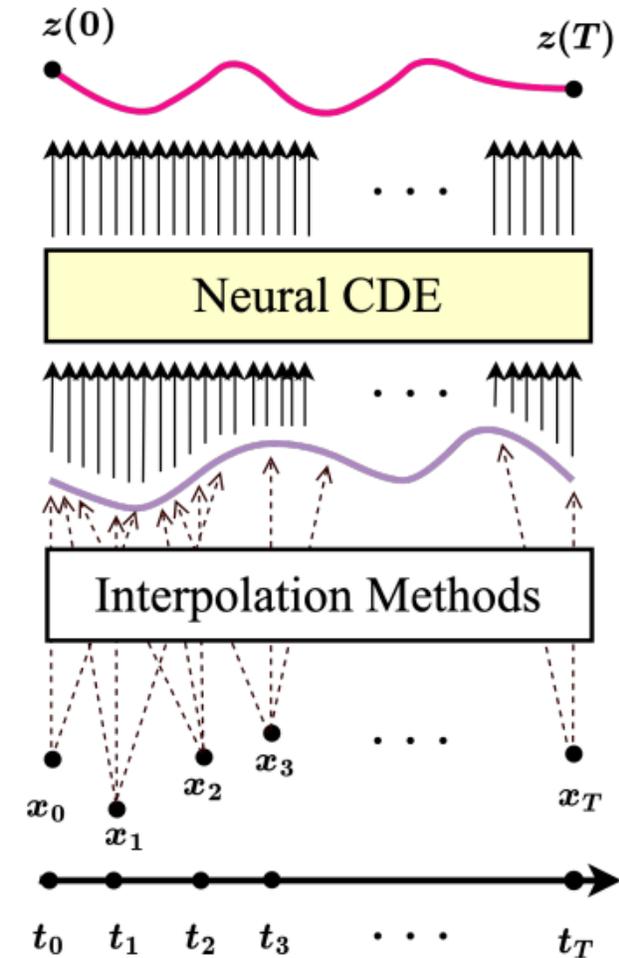
«Стандарт» в предсказании аномалий

- От случайных лесов [*RF*] до длинной краткосрочной памяти [*LSTM*]
- Пример:
 - Xu X., Zhao H., Liu H., Sun H. **LSTM-GAN-XGBOOST** Based Anomaly Detection Algorithm for Time Series Data. 2020 11th International Conference on Prognostics and System Health Management (PHM-2020 Jinan), Jinan, 2020, pp. 334-339



Свежие достижения в предсказании аномалий

- *Precursor-of-Anomaly (PoA) Detection*
 - Jhin, Sheo Yon ; Lee, Jaehoon ; Park, Noseong.
Precursor-of-Anomaly Detection for Irregular Time Series.
eprint arXiv:2306.15489, 2023. (<http://arxiv.org/abs/2306.15489>)
 - Репозиторий (<http://github.com/sheoyon-jhin/PAD>)
- С чем сравниваем?
 - *LSTM, LSTM-VAE, USAD*
- Что предлагаем?
 - *Neural controlled differential equations (NCDEs) +*
 - *Multi-task learning (MTL)*





Систематизация достижений ИИ для VR

- *AI for Time Series (AI4TS) Papers, Tutorials, and Surveys*
(<http://github.com/qingsongedu/awesome-AI-for-time-series-papers>)
 - Подборка журналов:
 - *CACM, PIEEE, TPAMI, TKDE, TNNLS, TITS, TIST, SPM, JMLR, JAIR, CSUR, DMKD, KAIS, IJF, arXiv(selected), etc.*
 - Подборка материалов конференций:
 - *Machine Learning: NeurIPS, ICML, ICLR*
 - *Data Mining: KDD*
 - *Artificial Intelligence: AAAI, IJCAI*
 - *Data Management: SIGMOD, VLDB, ICDE*
 - *Misc (selected): WWW, AISTAT, CIKM, ICDM, WSDM, SIGIR, ICASSP, CVPR, ICCV, etc.*



Обнаружение изменения поведения ВР

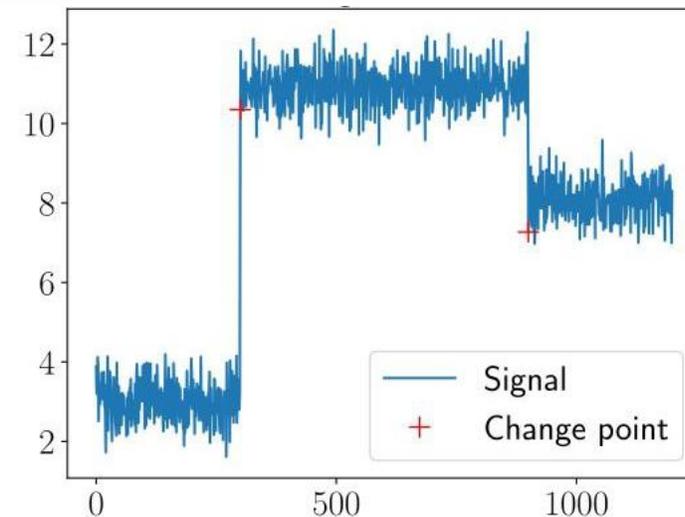
- *Truong C., Oudre L., Vayatis N. **Selective review of offline change point detection methods.** Signal Processing, 167:107299, 2020.*
- Базовые библиотеки:
 - ***ruptures** – a Python library for off-line change point detection*
(<http://github.com/deepcharles/ruptures/>)
 - ***stumpy** – a powerful and scalable Python library for time-series segments similarity analysis* (<http://github.com/TDAmeritrade/stumpy/>)
 - ...



Что может меняться в ВР (2)

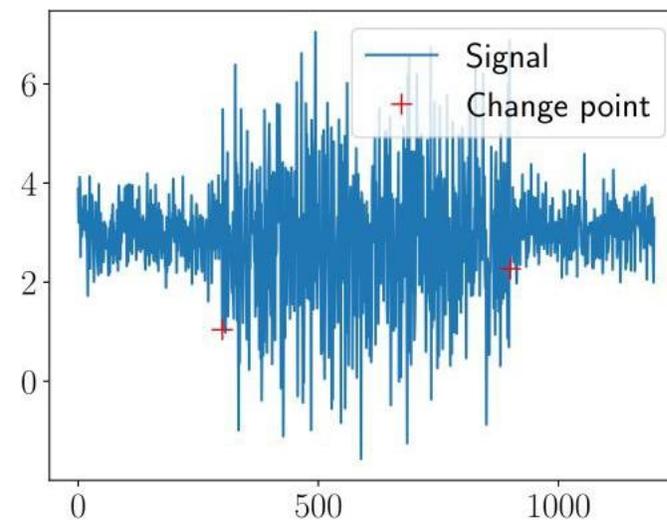
- Тренд/среднее значение [*mean*]

- Сдвиг



- Разброс/дисперсия [*variance*]

- Разбалансировка

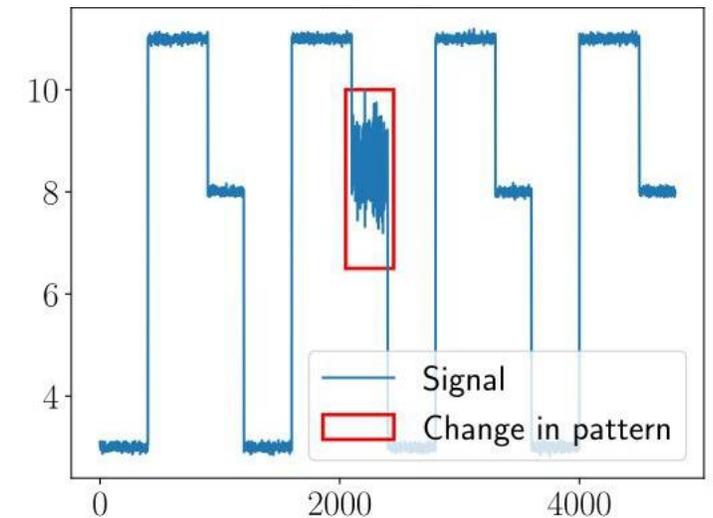
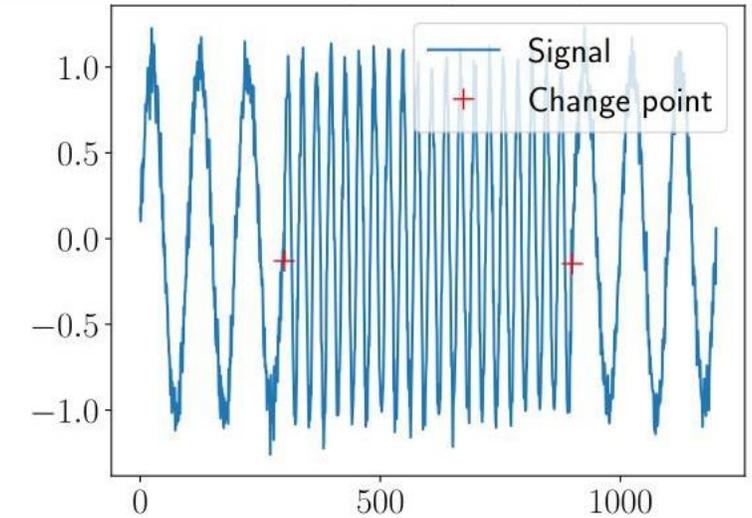




Что может меняться в ВР (1)

- Периодичность/частота [*periodicity*]
 - Наполненность

- Шаблон [*pattern*]
 - Повторяющиеся/«выдающиеся» особенности поведения





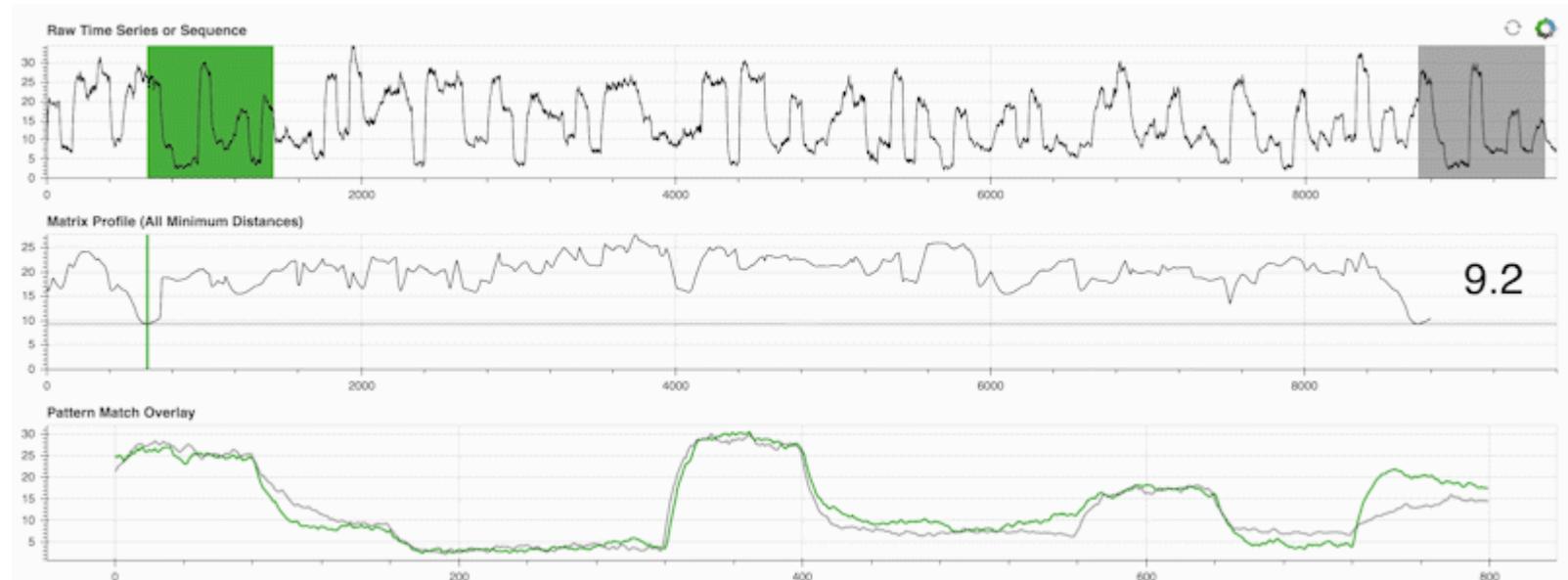
Матричные профили

- *Matrix Profile (MP)* – символ одной из «практичных революций»
 - *The UCR Matrix Profile Page* (<https://www.cs.ucr.edu/~eamonn/MatrixProfile.html>)
- 2016 год
 - *Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets*. Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, Eamonn Keogh. *IEEE ICDM 2016*.
- Настоящее время
 - *Matrix Profile XXVIII: Discovering Multi-Dimensional Time Series Anomalies with K of N Anomaly Detection*. Sadaf Tafazoli and Eamonn Keogh. *SIAM SDM 2023*



Наиболее популярная реализация *MP*

- **Stumpy** (<http://github.com/TDAmeritrade/stumpy>)
 - Универсальная реализация
 - Но не самая эффективная для частных задач!
 - Matrix profile understanding (http://stumpy.readthedocs.io/en/latest/Tutorial_The_Matrix_Profile.html)





А как начать использовать?

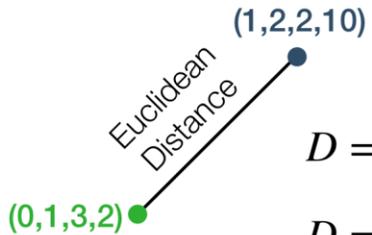
- Хорошее введение – *Mueen A., Keogh E. Time Series Data Mining Using the Matrix Profile: A Unifying View of Motif Discovery, Anomaly Detection, Segmentation, Classification, Clustering and Similarity Joins. 2017*
(http://www.cs.ucr.edu/~eamonn/Matrix_Profile_Tutorial_Part1.pdf)



Пример вычисления простейшего *MP*

- Сходство частей ВР можно считать по-разному...
 - Евклидово расстояние между окнами как простейший пример

Euclidean Distance



$$D = \sqrt{(1 - 0)^2 + (2 - 1)^2 + (2 - 3)^2 + (2 - 10)^2}$$

$$D = \sqrt{67}$$

#Back2School

Pairwise Euclidean Distance



#DistanceProfile



ВР с пропусками данных

- Восстановление пропусков и работа на данных с пропусками – важнейшее направление исследований с точки зрения практически
- **PyPOTS**: a Python toolbox for data mining on Partially-Observed Time Series (<http://github.com/WenjieDu/PyPOTS/>)
- **SAITS**: Self-Attention-based Imputation for Time Series, 2023 (<http://github.com/WenjieDu/SAITS>)



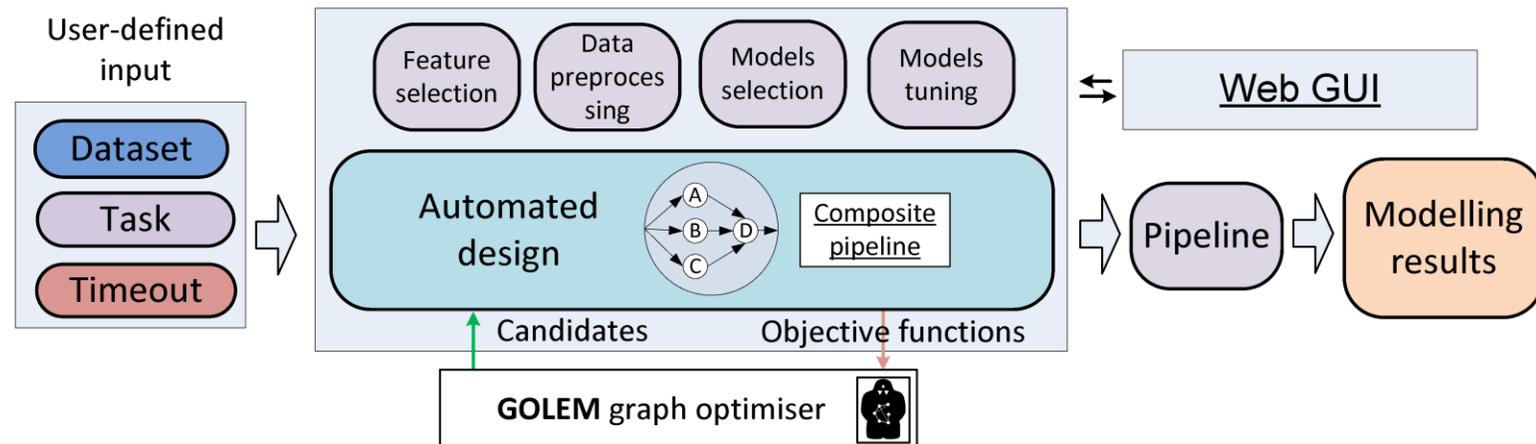


Прогнозирование – FEDOT

- **FEDOT** – российская библиотека *AutoML* для ВР на основе **генетических алгоритмов**

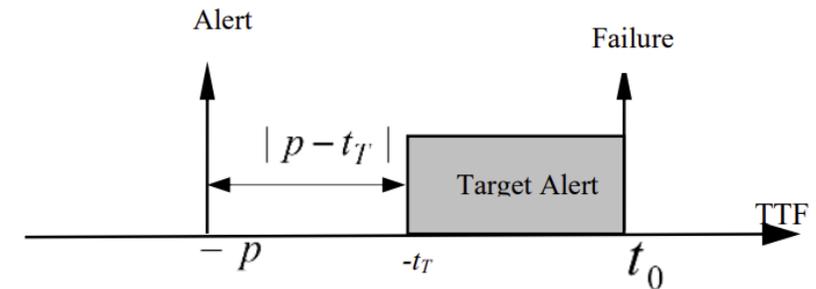


- *FEDOT is an open-source framework for automated modeling and machine learning (AutoML) problems* (<https://github.com/aimclub/FEDOT>)
- *Time series forecasting with FEDOT. Guide* (http://github.com/ITMO-NSS-team/fedot-examples/blob/main/notebooks/latest/3_intro_ts_forecasting.ipynb)



♥ А приложения напрямую к РНМ?

- Отстаём!
 - Но активно работаем...
- Пример уже классической попытки:
 - *Ducoffe, M., Haloui, I., Sen Gupta, J. **Anomaly Detection on Time Series With Wasserstein Gan Applied to PHM.** IJPHM Special Issue on PHM Applications of Deep Learning & Emerging Analytics, International Journal of Prognostics and Health Management, 10(4), 2019.*
- Как оценить результат?
 - *Yang, C.; Zou, Y.; Liu, J.; Mulligan, K.R. **Predictive model evaluation for PHM.** International Journal of Prognostics and Health Management, 5(2), 2014*





Большие языковые модели

- **LLM** – *Large Language Model*
 - *Microsoft + OpenAI (ChatGPT, ...) vs Google + DeepMind (Bard, ...) vs Meta* (LLaMA, ...)*
 - Хотя там все уже сменили партнёров и перекупили друг друга...
 - Многочисленные попытки внедрения
 - Дообучение и интеграция с корпоративными информационными системами
 - Обсуждения «разумности»!
- Объединение с онтологиями и логическим выводом



* Организация, деятельность которой запрещена на территории Российской Федерации



С текстами более-менее понятно

- Очевидные приложения в написании, редактировании, стилизации текстов
 - См. *Microsoft Office 365*:

The screenshot displays the Microsoft Word 365 interface. The top ribbon is set to the 'Review' tab, showing options for Spelling and Grammar, Editor, Read Aloud, Check Accessibility, Language, New Comment, Delete, Previous, Show Comments, Tracking, Accept, Compare, Protect, Hide Ink, and Linked Notes. The main document area contains text about Large Language Models (LLMs). A context menu is open over the phrase 'large quantities', offering suggestions: 'Vocabulary', 'More specific adjectives are clearer and add impact', 'copious quantities', 'enormous quantities', and 'massive quantities'. On the right, the 'Editor' pane is visible, showing an 'Editor Score' of 83% with a progress bar, a 'Formal writing' style selector, and a 'Corrections' section listing 3 spelling and 2 grammar errors. Below that, 'Refinements' for Clarity, Conciseness, and Formality are all checked. The status bar at the bottom indicates 'Page 1 of 1', 'Column: 62', '400 words', 'English (United States)', and 'Text Predictions: On'.

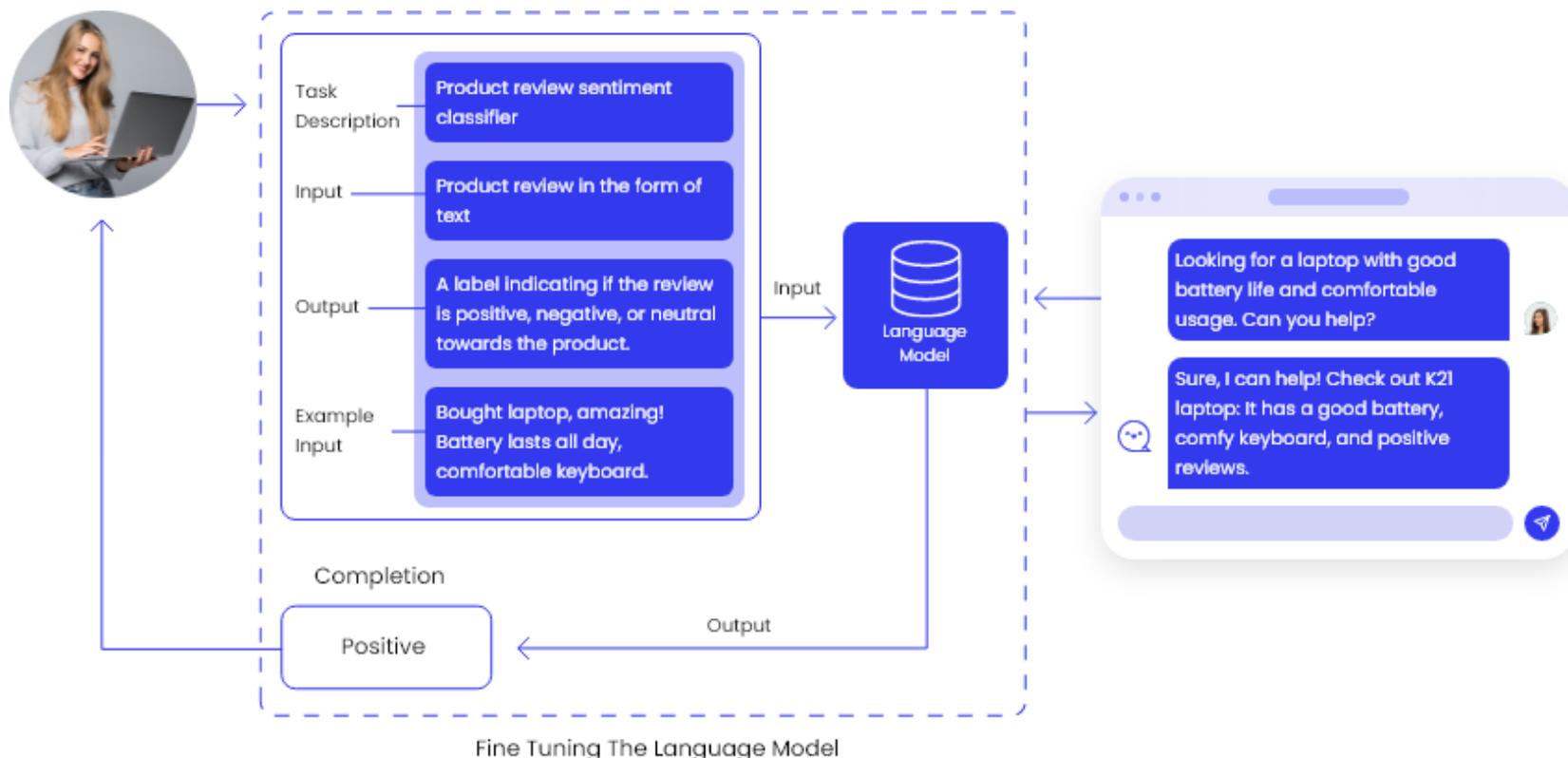


Мультимодальные модели и промты

- Промт-инжиниринг [*prompt engineering*]

- Новая область
- Сразу несколько профессий

- *Prompt Engineering Guide*
(<http://www.promptingguide.ai>)



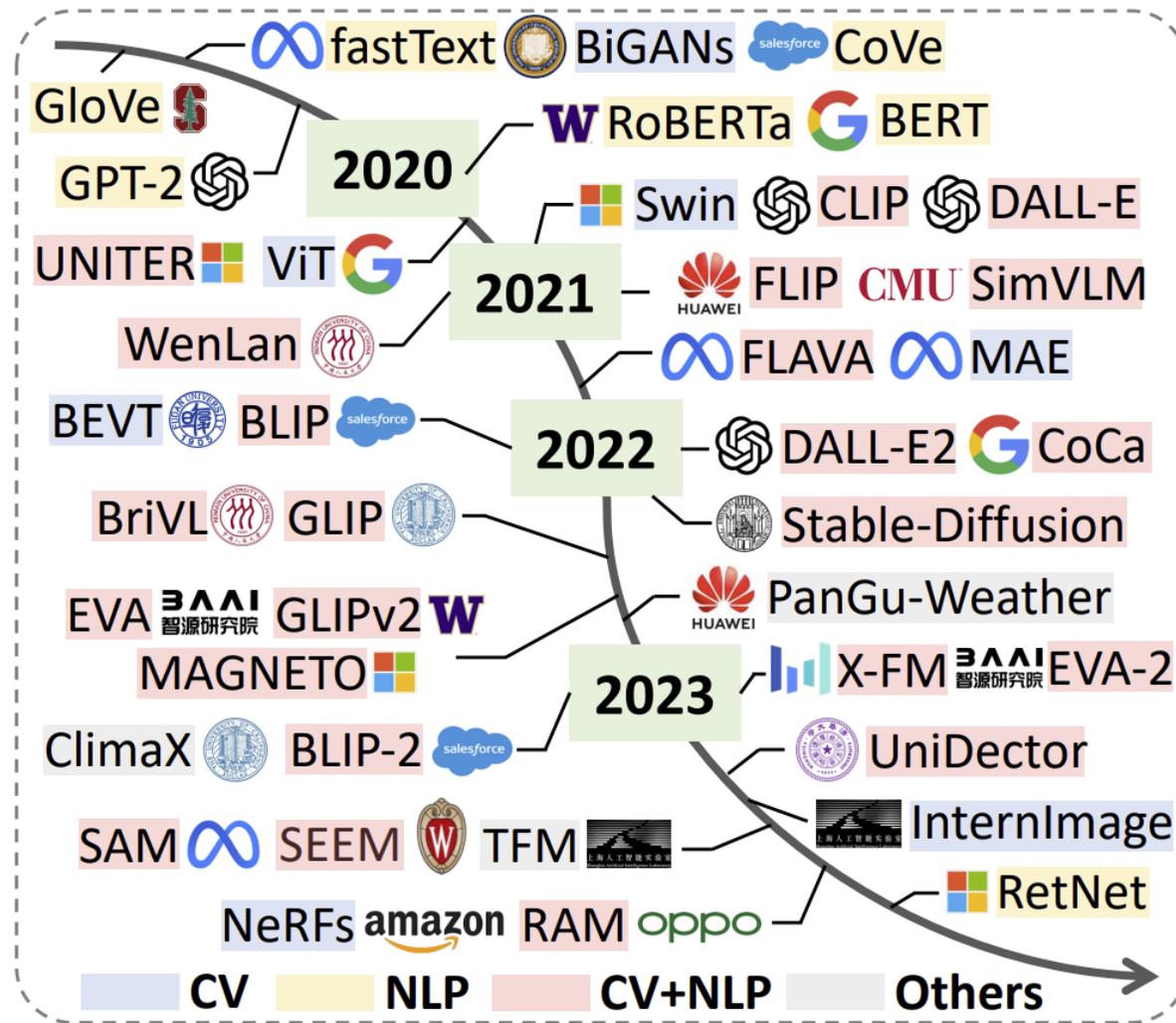
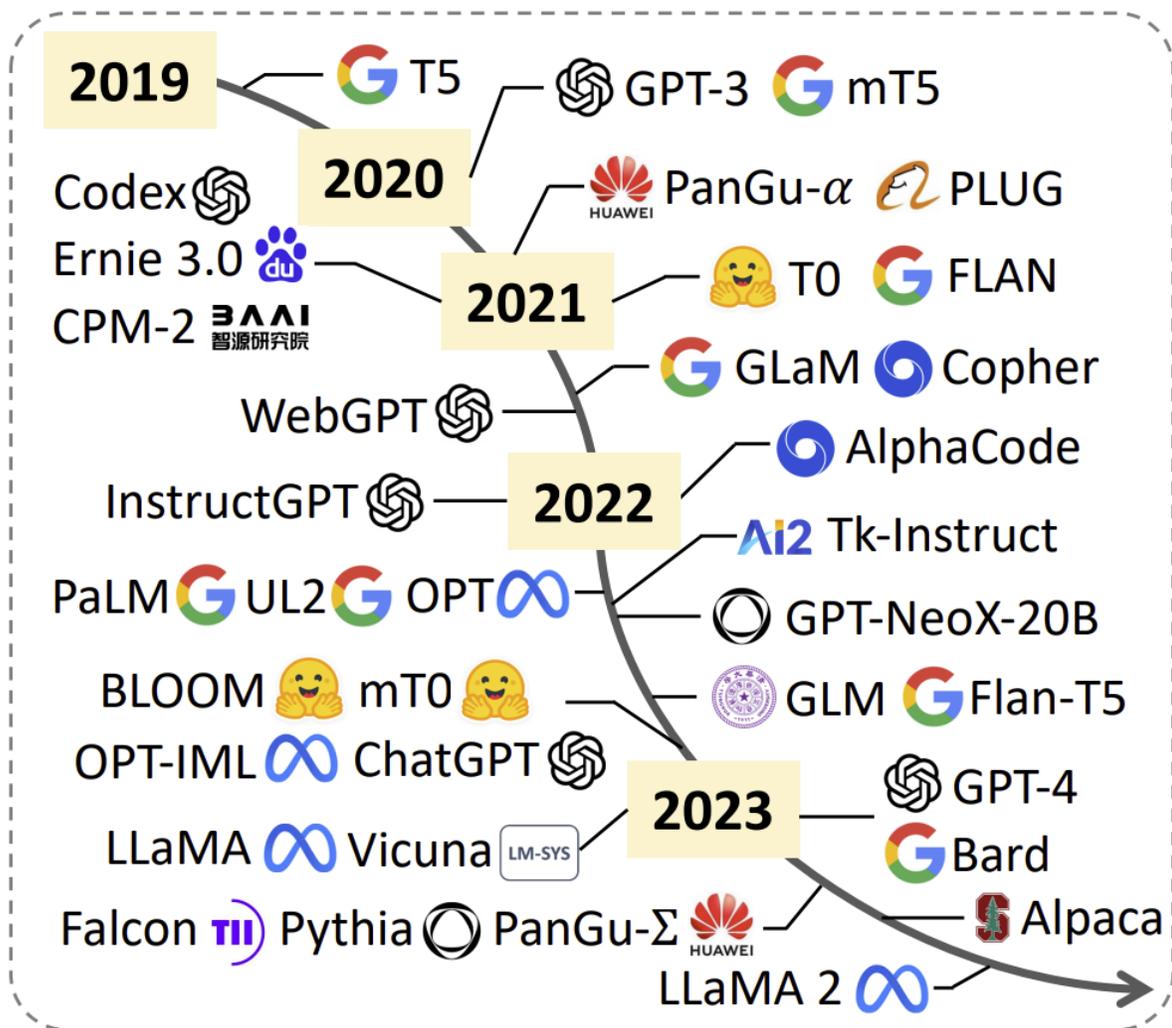


LLM для анализа ВР и темпоральных данных?

- *Large (Language) Models and Foundation Models (LLM, LM, FM) for Time Series and Spatio-Temporal Data* (<http://github.com/qingsongedu/awesome-timeseries-spatiotemporal-lm-llm>)
 - Исследуются также и пространственно-временные данные
 - *Domain models including spatio-temporal graphs (STG), temporal knowledge graphs (TKG)*
 - Кроме *LLM* интересны *pre-trained foundation models (PFM)*
 - Не знаю ни одного короткого адекватного перевода ☹️
 - «Базовая модель» – совсем плохо...

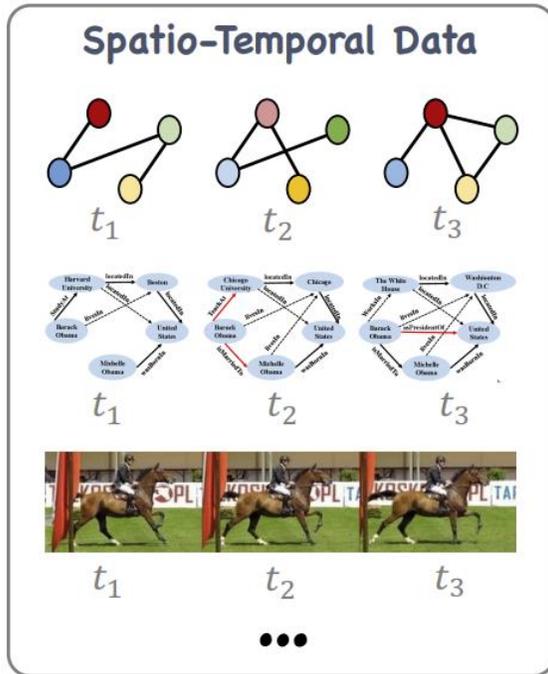
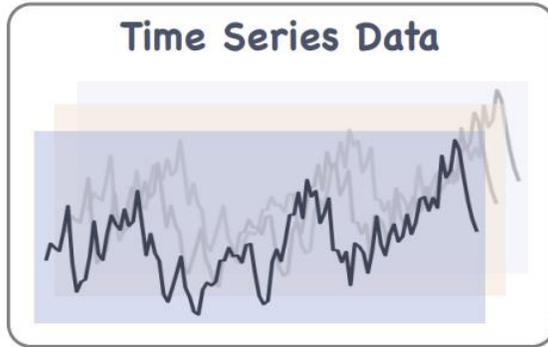


Иллюстративные LLM и их изводы



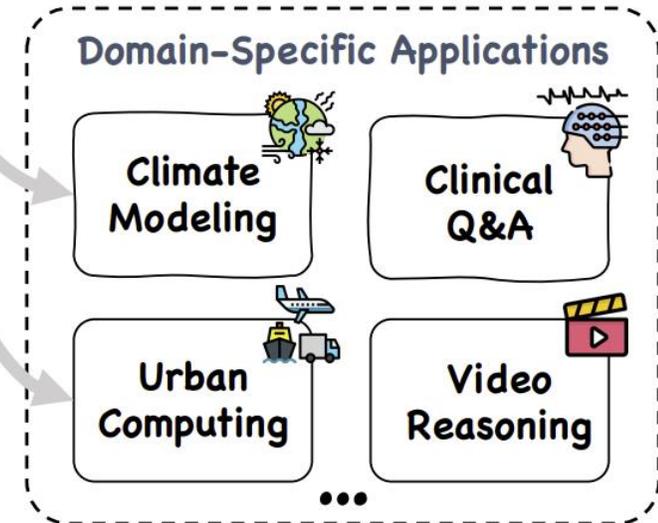
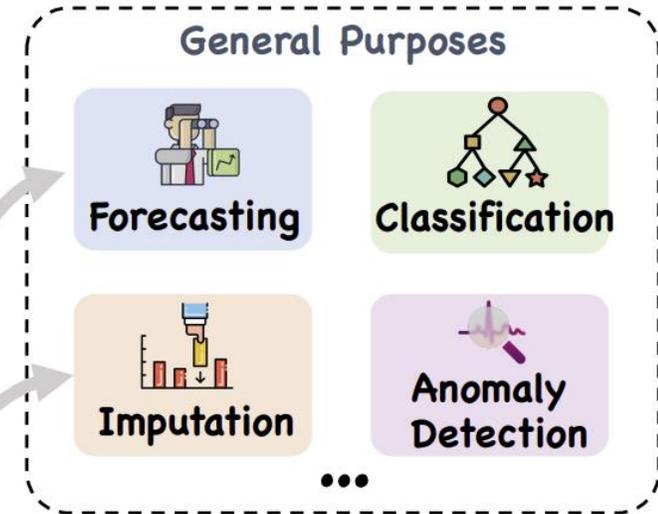
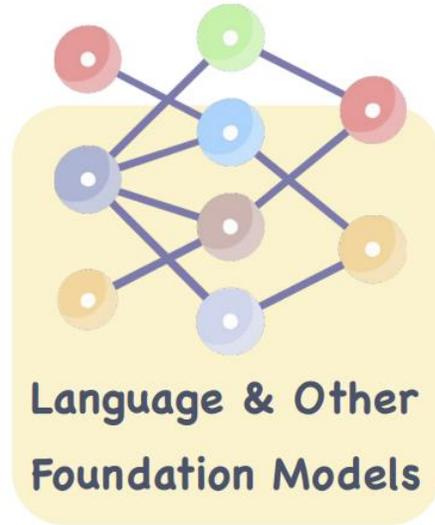


Трактовка темпоральных данных в *LLM*



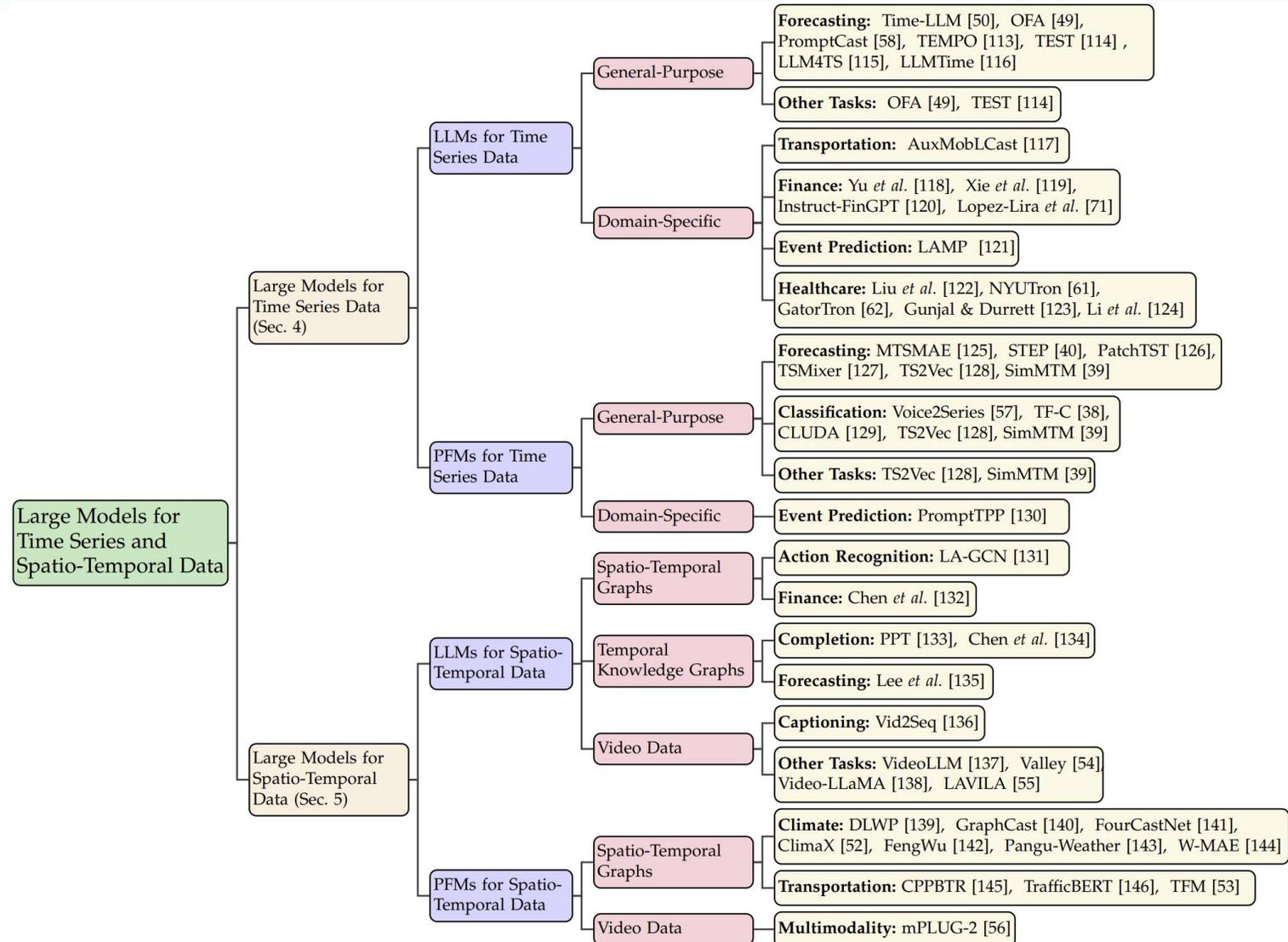
Training

Repurposing





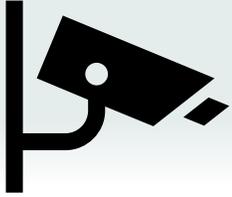
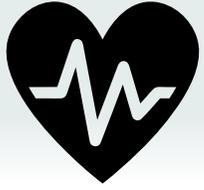
Таксономия больших моделей





Практичность *LLM*

- Если вдруг подойдёт готовый инструмент – замечательно
- А если нет?
 - Обучение *LLM* с нуля очень дорого
 - Оценка одного прогона обучения *ChatGPT* в 2023 году > 50 млн долларов, в 2024 году прогнозируют > 1 млрд долларов
 - Можно взять чужую модель и дообучить
 - Очень популярное направление исследований
 - Но предсказать эффект внедрения очень тяжело...
- Но даже попробовать сложно – нужна трансдисциплинарная команда



DDDM meets PHM

- ✓ *DDDM – Data-driven Decision Management (and Making)*
- ✓ Решения, основанные на данных (и знаниях!)
- ✓ Суть подхода
- ✓ Инструменты



Data-driven approach в целом

- **Подход на основе данных**
 - По другому не бывает! Но теперь у нас «цифровизация»...
 - На самом деле уже в 2010 году в США *DDDM* использовало ~30% производств
 - *Brynjolfsson E., McElheran K. The Rapid Adoption of Data-Driven Decision-Making // American Economic Review, 106(5), 2016, 133-139.*
- **Хайп и мода**
 - Данные уже давно – важный **актив организации** («новая нефть»)
 - Но **уровень использования данных** при принятии решений до сих пор остаётся **неудовлетворительным**
 - По различным опросам, так считают около **70-80% руководителей**
 - По другим опросам, примерно **40% бизнес-инициатив** проваливаются из-за плохого качества данных при принятии ключевых решений
 - Появился даже специальный термин “*data-driven disaster*”!



Data-driven Decision Management (DDDM)

- Управление принятием решений на основе данных
 - *Your Quick-Start Guide to Data-Driven Decision Making* (<http://www.smartsheet.com/data-driven-decision-making-management>)
 - *Weigert T. Data-Driven Decision Making: An Adoption Framework. MIT, 2017* (<http://dspace.mit.edu/handle/1721.1/111450>)
 - *Garofalo J. What Drives Executives to Adopt Data Driven Decision Making? 2017* (<http://www.idg.com/blog/what-drives-executives-to-adopt-data-driven-decision-making/>)
 - *Crosby P. How to Implement Data-Driven Decision Making in Your Organization. 2019* (<http://theuncommonleague.com/blog/data-driven-decision-making>)
- Когда получается?
 - Нормально работает ИТ-отдел (ведь нужно обрабатывать данные)
 - Много образованных сотрудников (ведь нужно думать)
 - Филиальная сеть (возникает дополнительная потребность в агрегации данных)
 - Есть кому внедрять *DDD* (очевидно)



DDDM – Процесс

- *Workflow* в зависимости от методологии
 - Стандартные «цепочки обработки» [*pipeline*]
- *Experiment Design (DoE)* как основа любой методологии
 - Статистика!
- Связь наиболее популярных методологий с доступными инструментами
 - Какие бывают с «затыки»?
 - Насколько они неожиданны?
 - Как решение проблем влияло на понимание/применение методологии?
- **Мониторинг становится частью *DDDM*!**



DDDM – Отдача

- Любое системное внедрение *DDDM* позволяет получить кумулятивный эффект:
 - Рост качества данных
 - Оптимизация затрат на аналитические платформы и сервисы
 - Повышение качества оперативного управления
 - Перевод большого числа источников данных из «серой» зоны в «белую»
 - Упрощения доступа к историческим данным
 - Уточнение политик безопасности
- Любое внедрение **Искусственного Интеллекта (ИИ)** – это на **50% *DDDM*!**
 - В основном из-за принципа “*Garbage In – Garbage Out*” и нормальных условий для машинного обучения



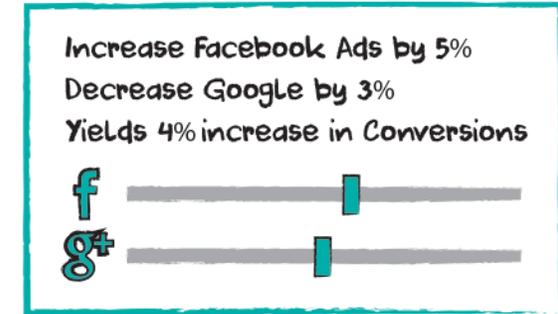
DDDM – Основные виды отчётов

- **Описательные** (эксплораторные)
 - Текущая ситуация
- **Диагностические** (мониторинговые)
 - Отслеживание изменений
- **Предсказывающие** (предиктивные)
 - Предсказание изменений
- **Предписывающие** (прескриптивные)
 - Выявление причин изменений и их взаимосвязи
- **Комбинированные**

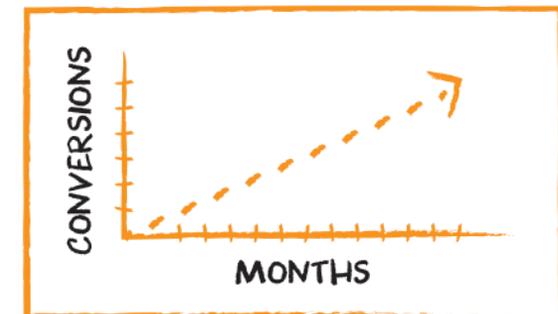
DESCRIPTIVE



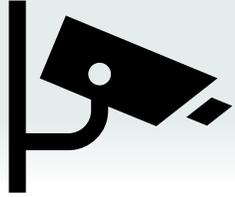
PRESCRIPTIVE



DIAGNOSTIC



PREDICTIVE



Инфраструктурные инструменты прогностики

- ✓ Хранилища данных
- ✓ Как дополнять/изменять
- ✓ Интеграция инструментов
- ✓ Визуализация



Популярные СУБД для ВР

- *DB-Engines Ranking of Time Series DBMS*
(<http://db-engines.com/en/ranking/time+series+dbms>)

Rank			DBMS	Database Model	Score		
Oct 2023	Sep 2023	Oct 2022			Oct 2023	Sep 2023	Oct 2022
1.	1.	1.	InfluxDB +	Time Series, Multi-model ⓘ	29.74	-1.53	+0.16
2.	2.	2.	Kdb +	Multi-model ⓘ	8.39	-0.56	+0.71
3.	3.	↑ 4.	Prometheus	Time Series	7.28	-0.34	+0.97
4.	↑ 5.	↑ 5.	TimescaleDB	Time Series, Multi-model ⓘ	5.38	-0.02	+0.19
5.	↓ 4.	↓ 3.	Graphite	Time Series	5.23	-0.22	-1.50
6.	6.	↑ 9.	DolphinDB	Time Series, Multi-model ⓘ	3.94	-0.11	+1.92
7.	7.	↓ 6.	Apache Druid	Multi-model ⓘ	3.10	-0.10	+0.56
8.	8.	↓ 7.	RRDtool	Time Series	3.06	-0.08	+0.53
9.	9.	↑ 13.	TDengine +	Time Series, Multi-model ⓘ	2.87	+0.25	+1.29
10.	10.	10.	QuestDB +	Time Series, Multi-model ⓘ	2.16	-0.23	+0.41
11.	↑ 12.	↓ 8.	OpenTSDB	Time Series	2.00	-0.11	-0.50
12.	↓ 11.	12.	GridDB +	Time Series, Multi-model ⓘ	1.96	-0.21	+0.32
13.	13.	↓ 11.	Fauna	Multi-model ⓘ	1.89	+0.20	+0.19
14.	14.	↑ 17.	VictoriaMetrics +	Time Series	1.28	-0.08	+0.37
15.	↑ 16.	↑ 19.	M3DB	Time Series	1.26	+0.19	+0.48
16.	↑ 18.	↑ 20.	Apache IoTDB	Time Series	1.18	+0.20	+0.53
17.	↓ 15.	↓ 14.	Amazon Timestream	Time Series	1.14	+0.02	-0.16
18.	↓ 17.	↓ 16.	eXtremeDB +	Multi-model ⓘ	0.93	-0.09	+0.02
19.	19.	↓ 18.	KairosDB	Time Series	0.90	-0.04	+0.05
20.	↑ 21.	↓ 15.	CrateDB +	Multi-model ⓘ	0.86	+0.05	-0.27



Актуальные открытые хранилища данных

- **Timescale**

- Postgres for time-series data (<http://www.timescale.com>)



Timescale

- InfluxData **InfluxDB**

- Time series data with a single, purpose-built database (<http://www.influxdata.com>)



influxdata[®]

- **OpenTSDB**

- The Scalable Time Series Database (<http://opentsdb.net>)



OPENTSDB

- Apache **Pinot**[™]

- Realtime distributed OLAP datastore (<http://pinot.apache.org>)
 - Timestamp Index (<http://docs.pinot.apache.org/basics/indexing/timestamp-index>)

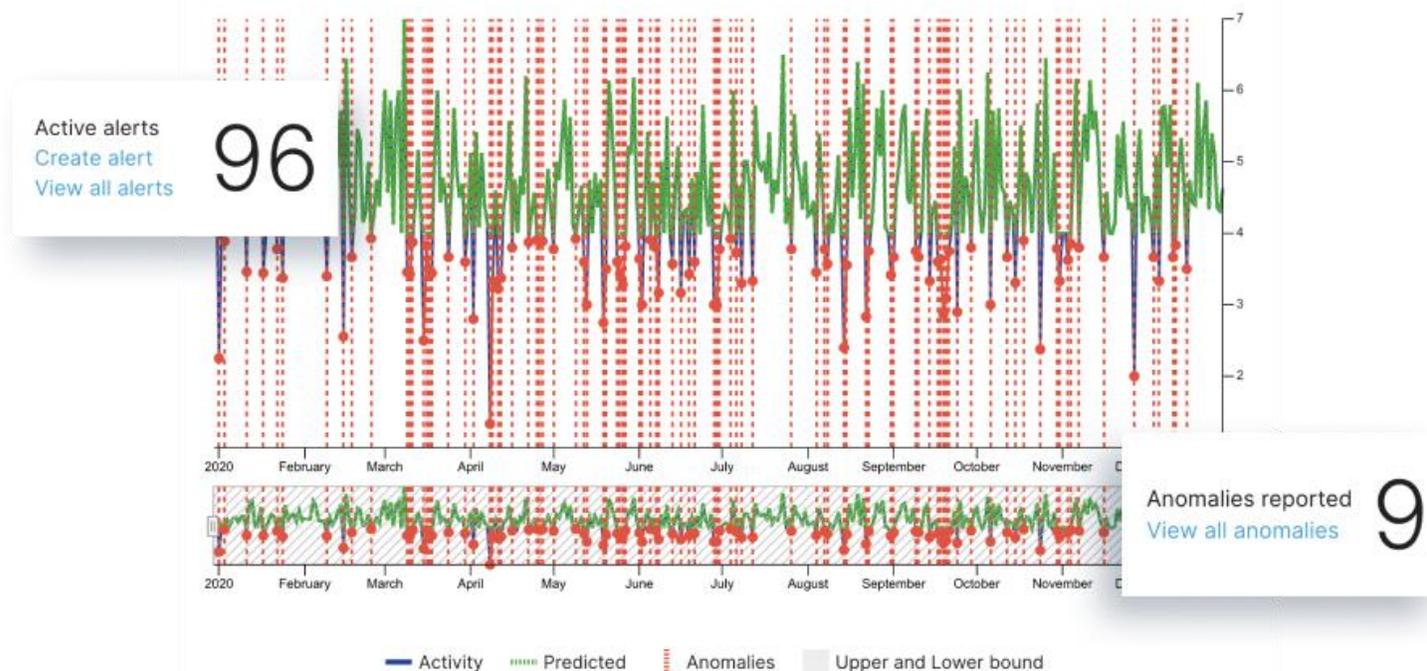


pinot

Мониторинг в реальном времени

- *ThirdEye* – an integrated tool for realtime monitoring of time series and interactive root-cause analysis (<http://github.com/startreedata/thirdeye>)
 - Настройка к Pinot

star[⚡]tree
THIRD EYE



Доработка методов анализа ВР

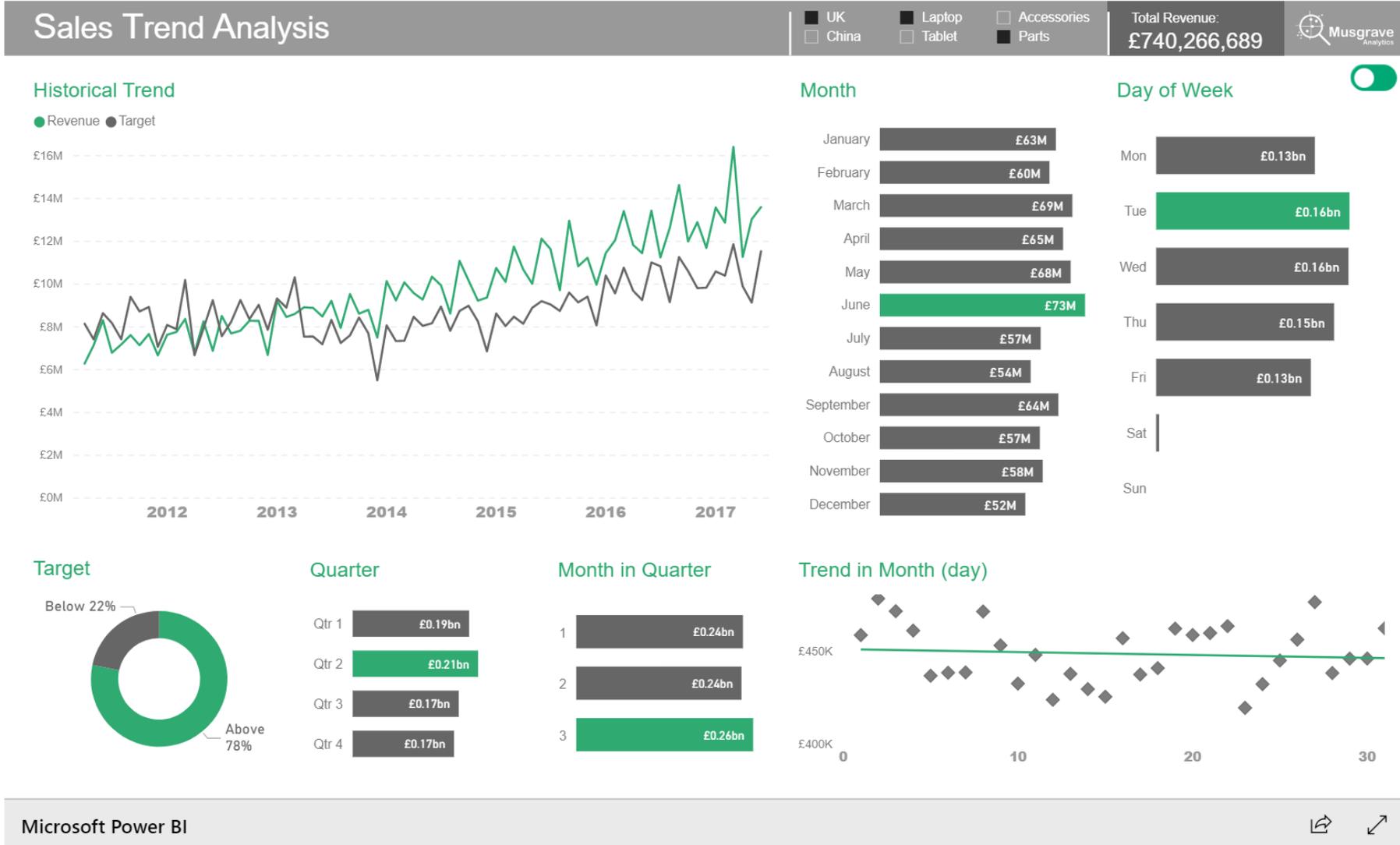
- Основные языки наук о данных и их библиотеки
- Инструментарий *Python* + *SciKit*
 - *Time Series for scikit-learn People (Part I): Where's the X Matrix?*
(<http://www.ethanrosenthal.com/2018/01/28/time-series-for-scikit-learn-people-part1/>)
 - *Time Series for scikit-learn People (Part II): Autoregressive Forecasting Pipelines*
(<http://www.ethanrosenthal.com/2018/03/22/time-series-for-scikit-learn-people-part2/>)
- Инструментарий *R*
 - *Using R for Time Series Analysis* (<http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html>)

Маленькие или большие данные?

- Большие данные позволяют получить новые знания из-за повышения детализации
 - Но всегда ли это надо?
- Но многие реальные задачи анализа событий прекрасно решаются в *Microsoft Excel / Power BI* на собственном ноутбуке
 - Анализ полумиллиона отзывов клиентов финансовых организаций – *Analyzing half a million consumer complaints* (<http://chandoo.org/wp/analyzing-consumer-complaints-1/>)
 - Примеры анализа продаж – *Sales Trend Analysis Dashboard* (<http://www.musgraveanalytics.com/sales-trend-analysis-dashboard>)
 - По опыту, при очищенных данных такой отчёт делается с нуля примерно за полдня



Пример: *Sales Trend Analysis Dashboard*

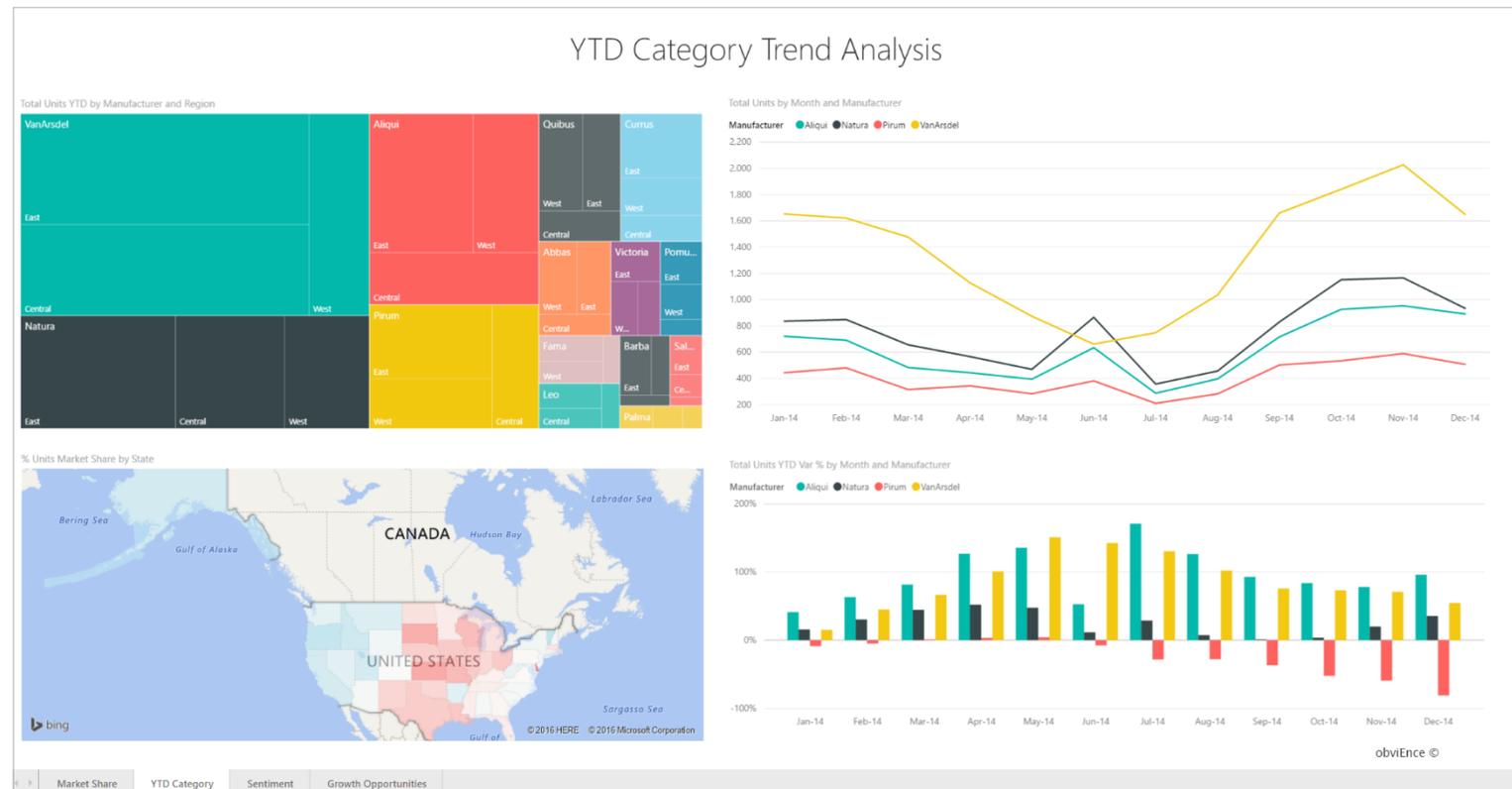
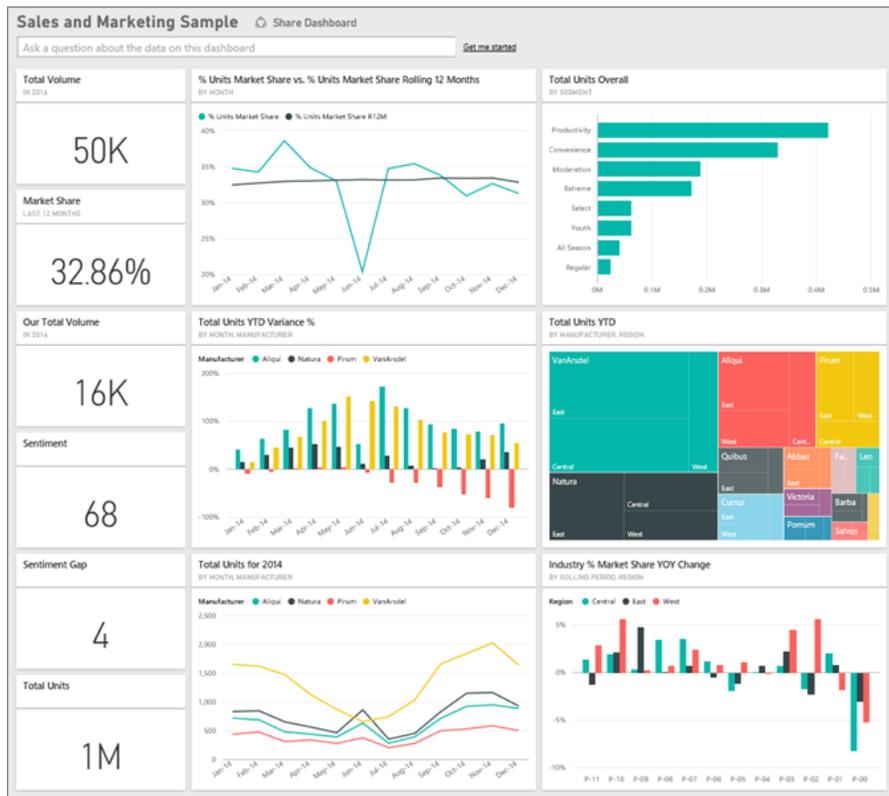




Пример от самой Microsoft (+ obviEnce)

- Sales and Marketing sample for Power BI: Take a tour

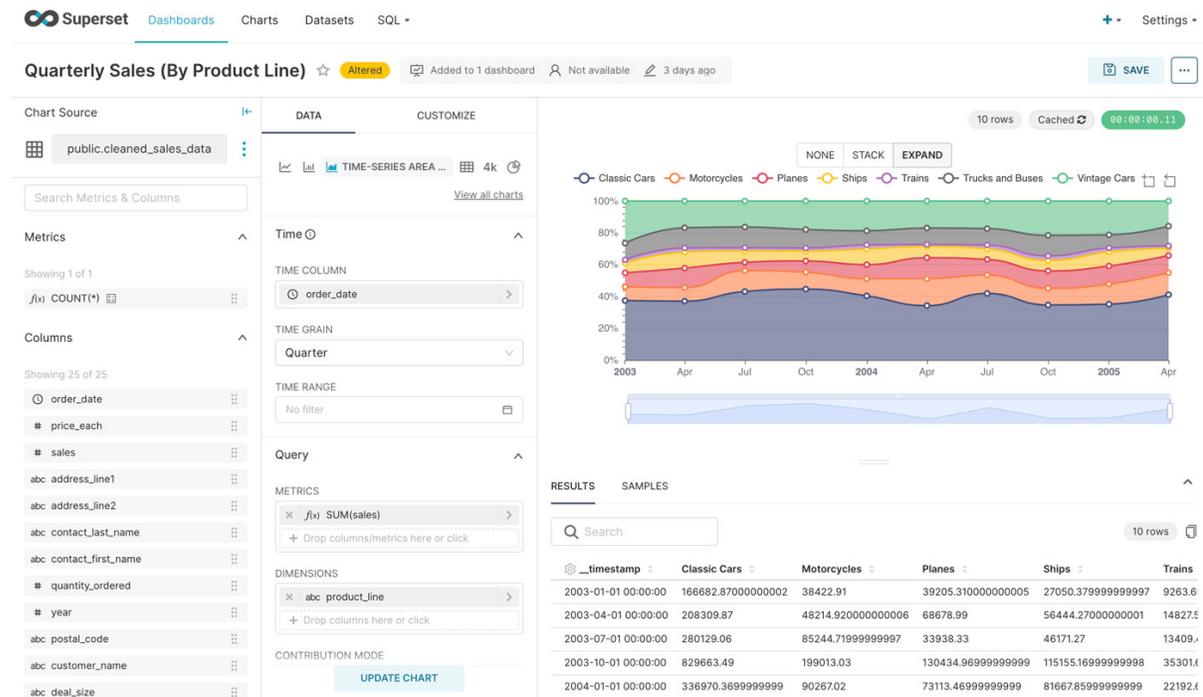
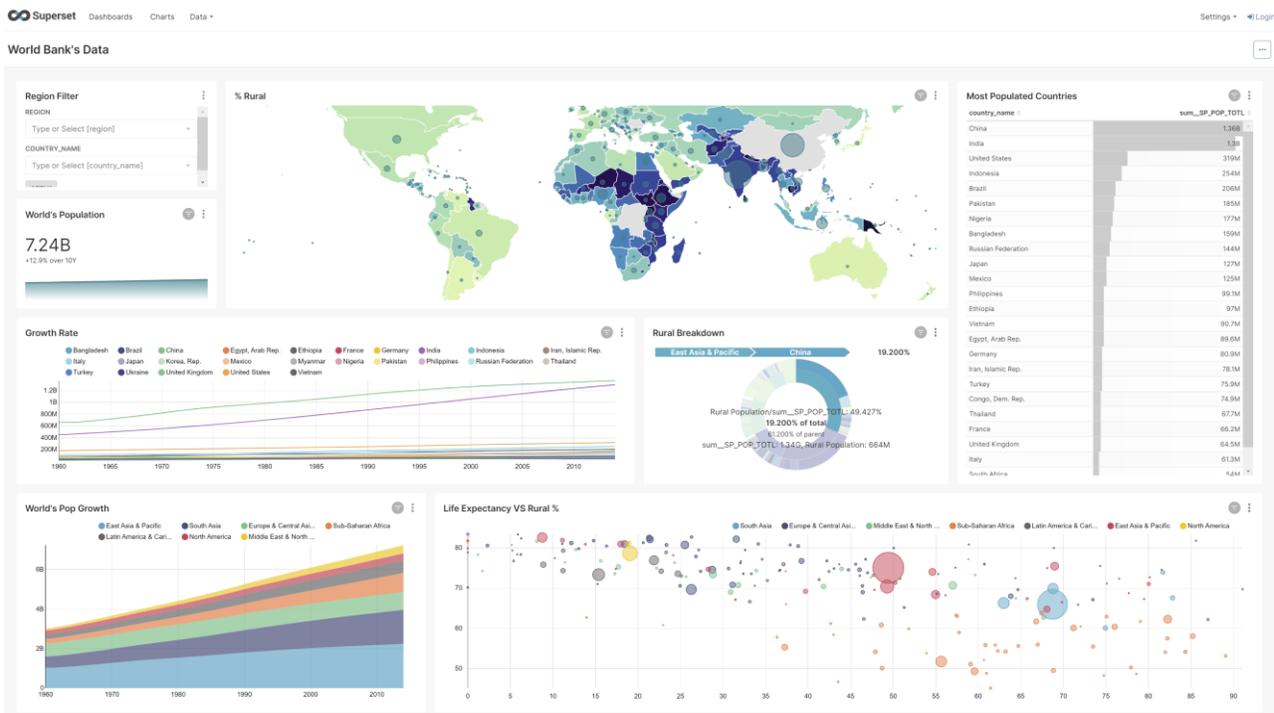
(<http://docs.microsoft.com/en-us/power-bi/create-reports/sample-sales-and-marketing>)





Инструменты с открытым исходным кодом

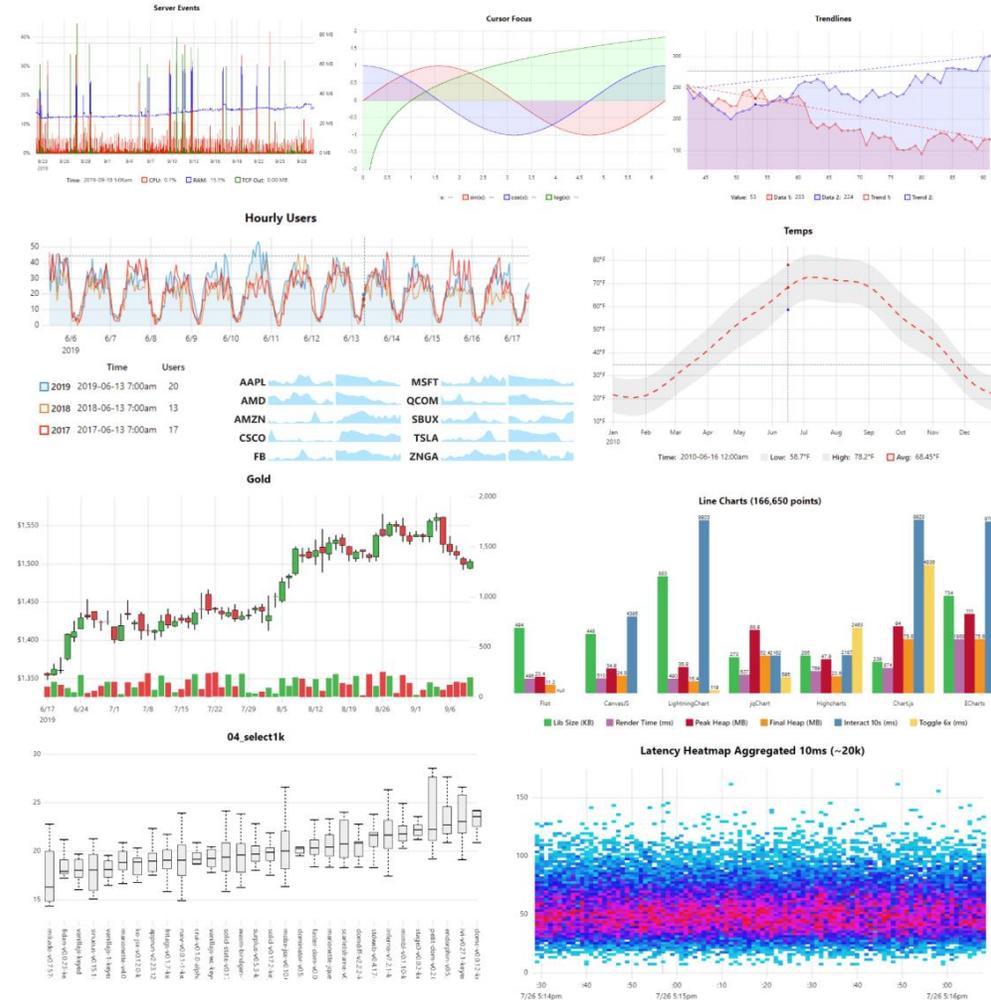
- **Apache Superset** (<http://superset.apache.org>)
 - Интерактивная визуализация данных на основе проекта *ECharts* и расширения *SQL*
 - Обычное ограничение на число точек в «карте» – 50 000





Визуализация больших ВР

- Для своей «веб-морды»?
- **μPlot** (<http://github.com/leoniya/uPlot>)
 - A fast, memory efficient Canvas 2D based chart for plotting time series, lines, areas, ohlc & bars
 - From a cold start it can create an interactive chart containing **150 000 data points in 90 ms**, scaling linearly at **~31 000 pts/ms**
- Но очень плохая документация...





Но это ещё не конец?!

Вопросы? Замечания? Предложения?

● Контакты:

● к.т.н., доц. **Незнанов Алексей Андреевич**

- Старший научный сотрудник международной лаборатории интеллектуальных систем и структурного анализа ФКН НИУ ВШЭ
- E-mail: alex.neznanov@gmail.ru
- Web-site: <http://hse.ru/staff/aneznanov>
- Blog: <http://siberianshamanssongs.blogspot.ru> (RU)

