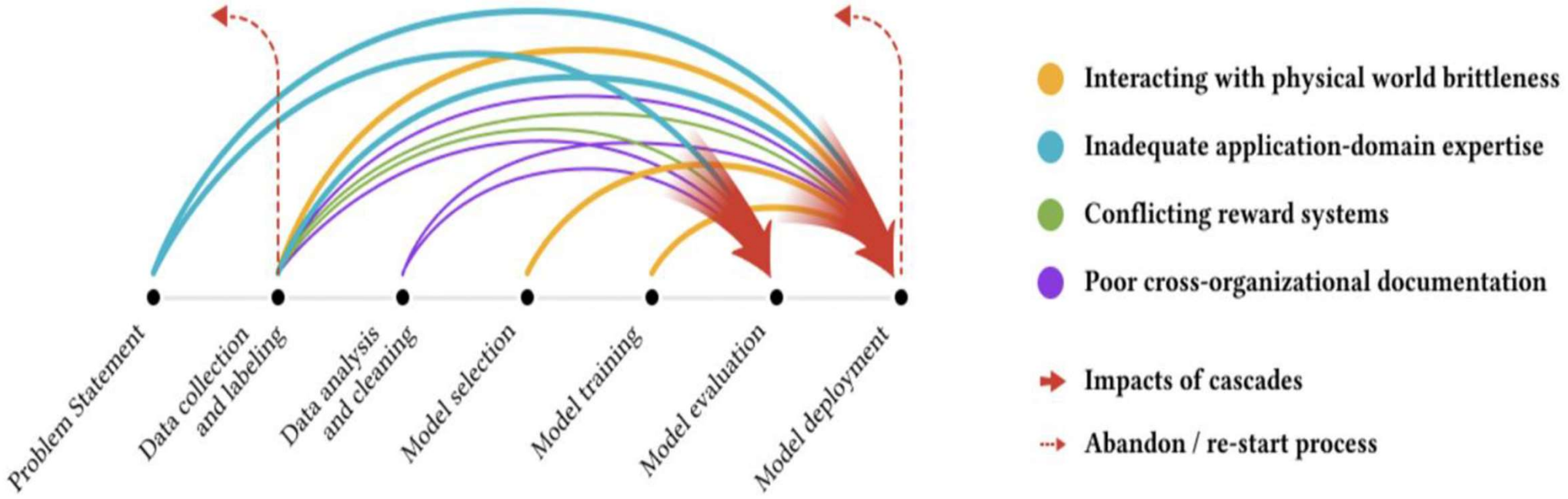
Лекция. Предобработка данных. Технология построения экспертных систем

Данные

- Качество данных
- Прозрачность данных
- Инструменты и примеры решений

Каскады данных



Эффет качественной очистки данных

Точность модели на «сырых» данных:

- при использовании библиотек DS (например Scikit Learn) – 0.7
- при изменении параметров и оптимизации модели – 0.73

Точность модели на «очищенных» данных:

- при использовании библиотек DS (например Scikit Learn) – 0.9
- при изменении параметров и оптимизации модели – 0.93

https://sigmod2016.org/sigmod_tutorial1.shtml

От оптимизации модели к очистке данных

	Обнаружение дефектов стали	**Солнечные панели**	**Инспекция поверхностей**
Базовый показатель	**76.2%**	**75.68%**	**85.05%**
Оптимизация модели	**+0%(76.2%)**	**+0.04%(75.72%)**	**+0%(85.05%)**
Очистка данных	**+16.9%(93.1%)**	**+3.06%(78.74%)**	**+0.4%(85.45%)**

Плохие данные

Лучшая модель

Плохие результаты

Управление данными

- **Данные разбросаны по разным системам хранения.**
- **Не всегда очевидно, кто владеет данными**
- **Проблемы с поиском и доступом к данным**
- **Тратится 25-50% времени только на поиск и оценку найденных данных**

- **Тестирование данных**
- **Наличие хорошего Каталога данных**
- **Рассмотрение данных как самостоятельного продукта**

Тестирование данных

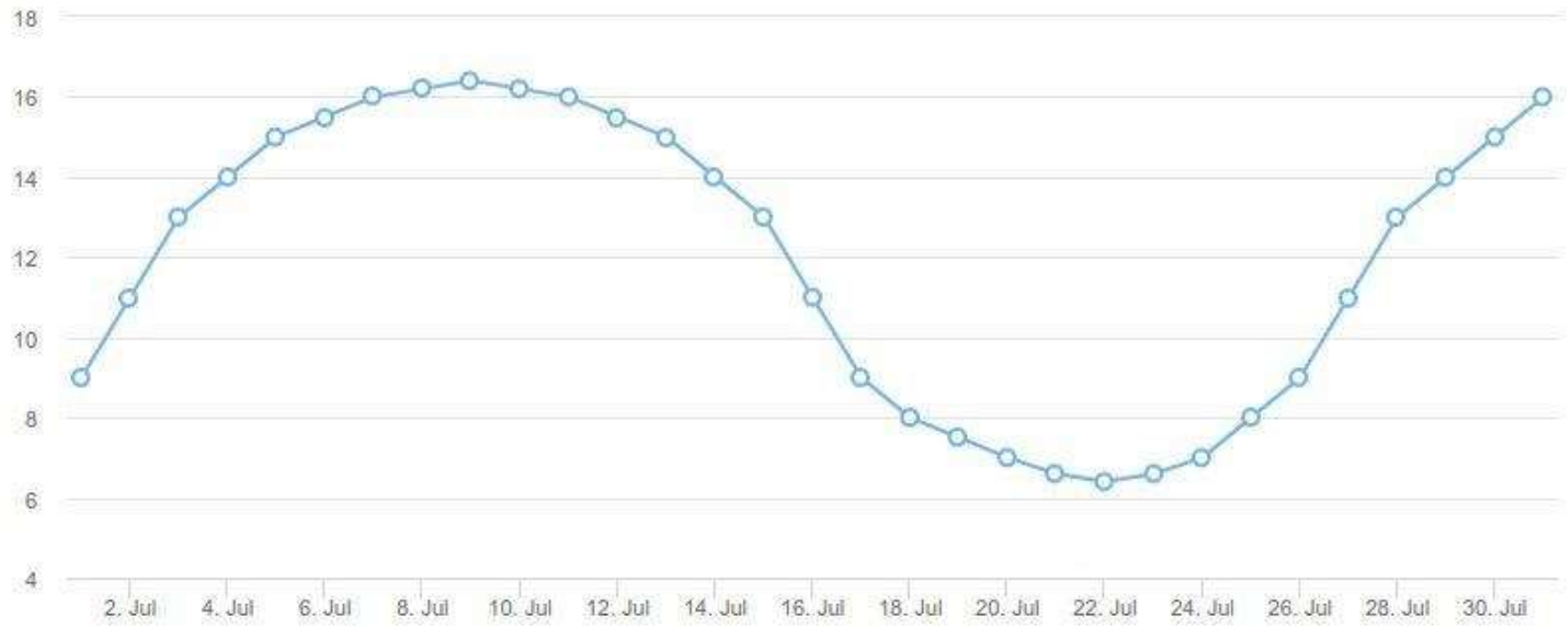
Стандартные проверки данных:

- **Дублирование**
- **Пропущенные значения**
- **Синтаксические ошибки**
- **Ошибки форматирования**
- **Семантические ошибки**
- **Целостность**

Расширенные методы:

- **Проверки распределения**
- **Критерий Колмогорова-Смирнова**
- **Критерий хи-квадрат**
- **Автоматический поиск аномалий**
- **Автоматическая генерация ограничений**

Автоматический поиск аномалий




20
18
16
14
12
10
8
6
4
2. Jul
4. Jul
6. Jul
8. Jul
10. Jul
12. Jul
14. Jul
16. Jul
18. Jul
20. Jul
22. Jul
24. Jul
26. Jul
28. Jul
30. Jul

Предсказание

Предсказание (forecast, prediction) — это по определению сообщение о некотором событии, которое непременно произойдет в будущем

Доверительный интервал

Доверительным называют интервал, который покрывает неизвестный параметр с заданной надежностью
Доверительный интервал — на сколько можно ошибиться в предсказании

- **Наивная**

$Y(t+1)=Y(t)$

Предсказания для каждого горизонта соответствуют последнему наблюдаемому значению.

Данная модель не должна использоваться для предсказаний, только для сравнения с другими.

Скользящее среднее

$$SMA_t = \frac{1}{n} \sum_{i=0}^{n-1} p_{t-i}$$

Предсказания численно равны среднему арифметическому значений исходной функции за установленный период. Отлично подходит под данные с выраженной сезональностью.

Линейная регрессия

- **Линейная регрессия** — регрессионная модель зависимости одной переменной y от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости.

Отлично предсказывает данные, обладающие выраженным трендом

Какие модели еще можно использовать

- Экспоненциальное сглаживание
- ARIMA, SARIMA
- GARCH
- TBATS
- Prophet
- NNETAR
- LSTM

Mean Square Error — средний квадрат ошибки определения какой-либо величины, квадратный корень из MSE есть среднеквадратическое отклонение определяемой величины от её математического ожидания.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Выборочная дисперсия — это среднее арифметическое квадратов отклонений всех вариант выборки от её средней

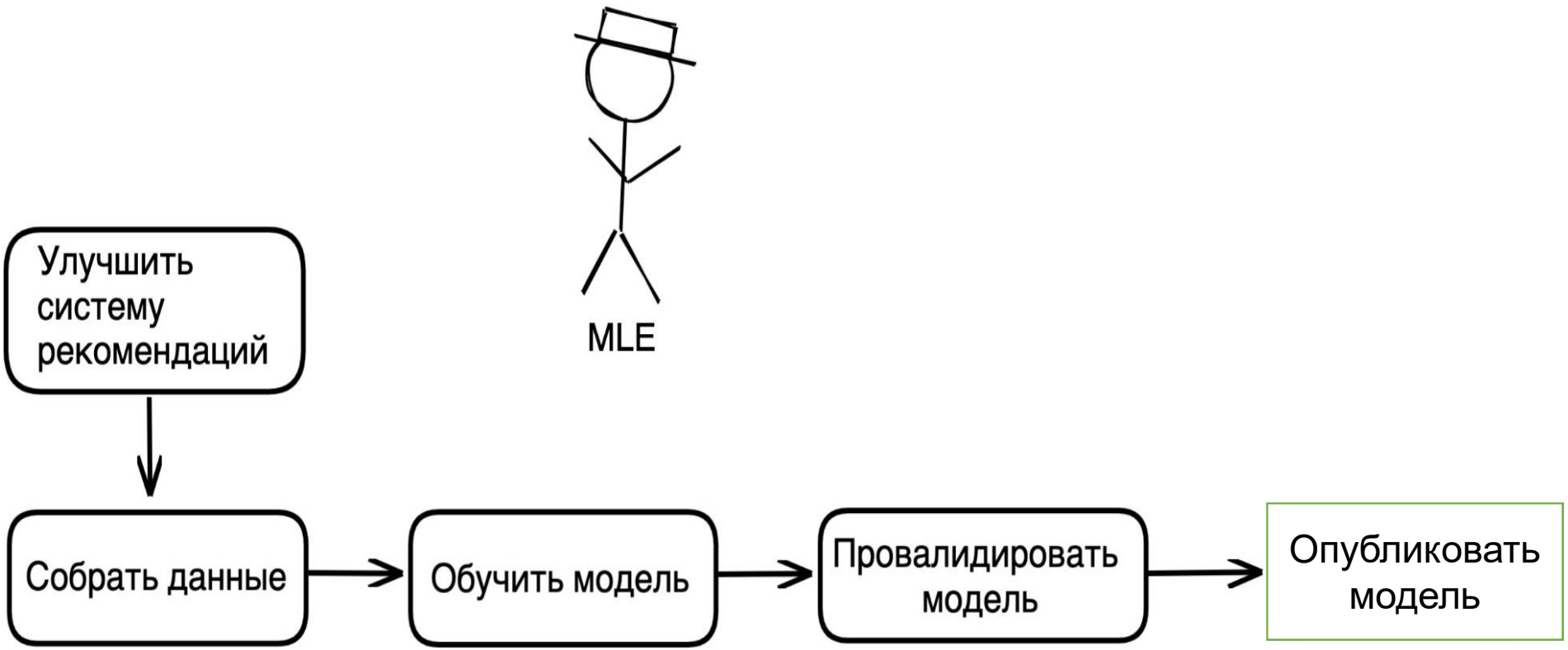
$$\sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n}$$

Распределение Стьюдента используется, например, в t-критерии Стьюдента для оценки статистической значимости разности двух выборочных средних при построении доверительного интервала для математического ожидания

Позволяет получить доверительный интервал с нужной точностью при малых объемах выборки

Коэффициенты (critical values) не зависят от данных и можно использовать табличные значения


200k
100k
0
-100k


MLE
Улучшить систему рекомендаций
Собрать данные
Обучить модель
Провалидировать модель
Опубликовать модель