Unsupervised learning - 2022

Basics of time series

Gleb Ivashkevich

whoami

Gleb Ivashkevich

doing **deep learning** - time series, satellite imagery

PhD in theoretical physics

6 years in **academia** doing numerical simulations

9 years in **data science** and **machine learning**

Y-DATA:

- Python for data processing (2018-2021)
- Advanced time series (2021-2022)


datarythmics **effimly**
data driven manufacturing efficiency

Time series

Time series are **sequences.**

- **various problems:** forecasting, segmentation, classification, event prediction, representation learning
- classical ML can be applied by windowing
- all the sequential **deep learning** blocks can be applied
  (RNN, CNN, combinations, incl. transformers → adaptations)

What this lecture is about?

- discussion of various time series problems
- practical considerations (windowed features, datetime indexes)
- segmentation example
- ~~ARIMA, SARIMA, etc.~~
- ~~econometrics, financial time series~~

Time series: definitions and examples

Time series: definition

→ time-ordered sequence of values (multivariate): $s_{\alpha}(t_k)$

→ $t_0 < t_1 < t_2 \ldots < t_N$

→ may be unevenly spaced,

→ very large Δt may be treated as a gap.

Time series: examples

- → EEG, ECG and other physiological signals (~1000-100 Hz),
- → motion sensors (~100-10 Hz),
- → factory sensors measurements (~10-10^{-1} Hz),
- → number of customers in a store per hour (10^{-1}-10^{-2} Hz).

Event stream: definition

→ a set of entities, having some attributes and indexed by some form of timestamp: $\mathbf{E}(t; \boldsymbol{a}_0, \boldsymbol{a}_1, \ldots)$

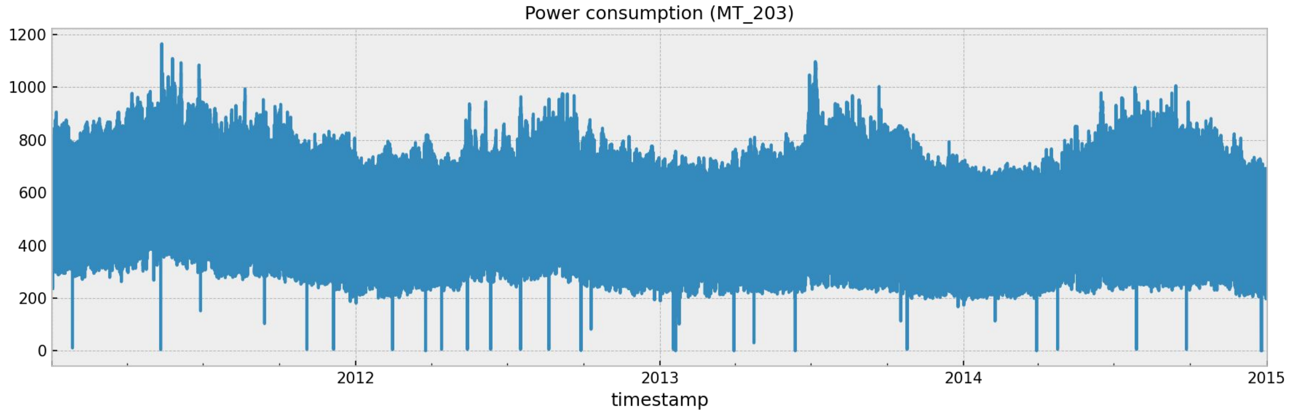
→ Individual events may be not related to each other,

→ aggregates from event stream may be represented as time series.

Example: power consumption

Electricity load [dataset]

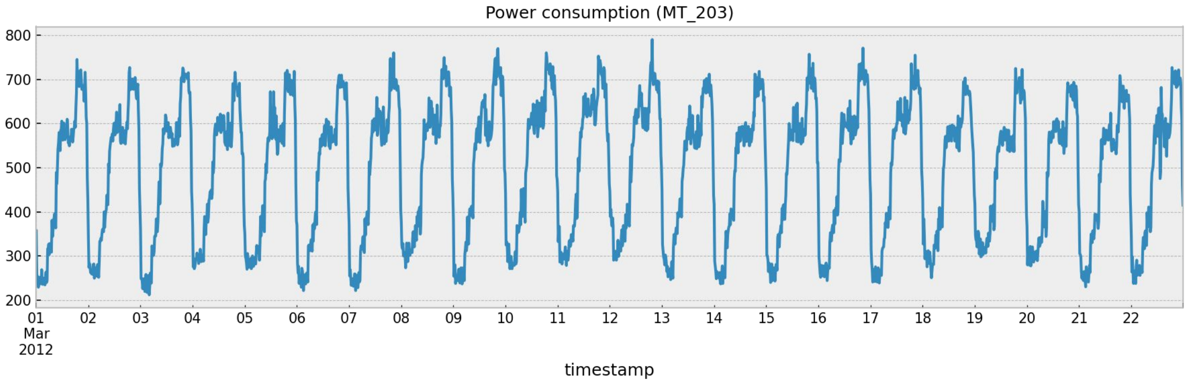
- 370 individual households
- 15 minutes sampling interval



Example: power consumption

Electricity load dataset

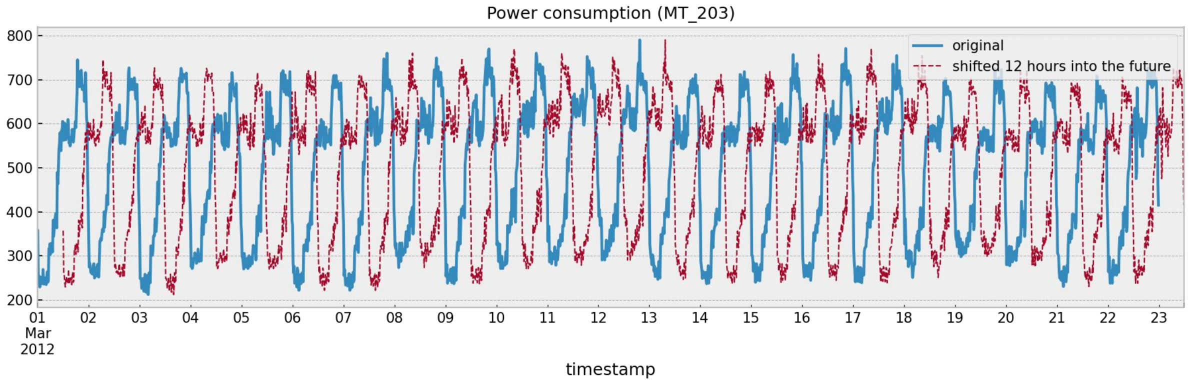
- 370 individual households
- 15 minutes sampling interval



Example: power consumption

Temporal structure?

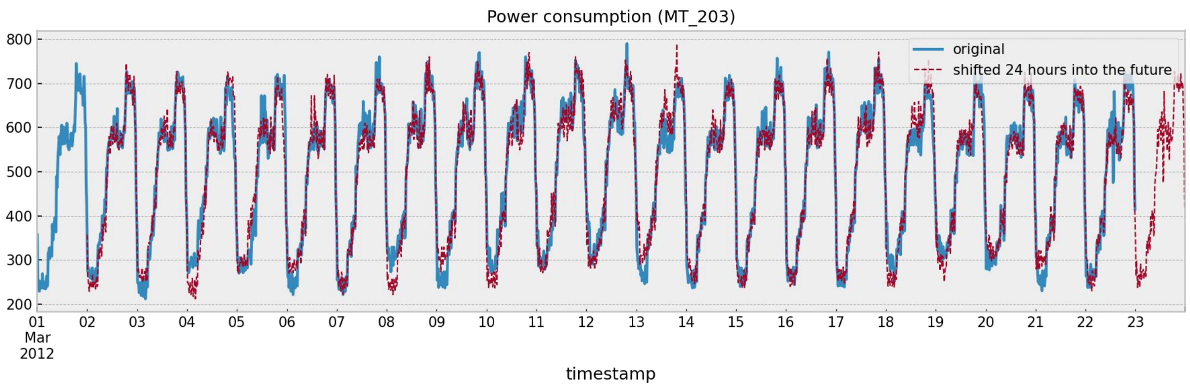
- shift the data: 12 hours



Example: power consumption

Temporal structure?

- shift the data: 24 hours



Notes

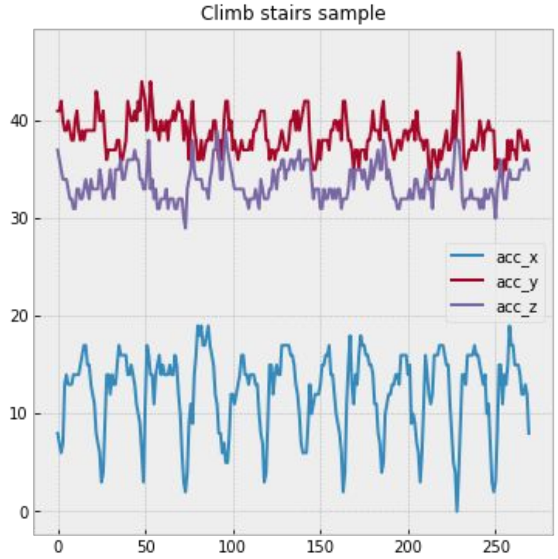
Structure:

- values at t_N depend on $t_{N\text{-}1}$, $t_{N\text{-}1}$ depend on $t_{N\text{-}2}$, …
- autocorrelation, multiple time scales

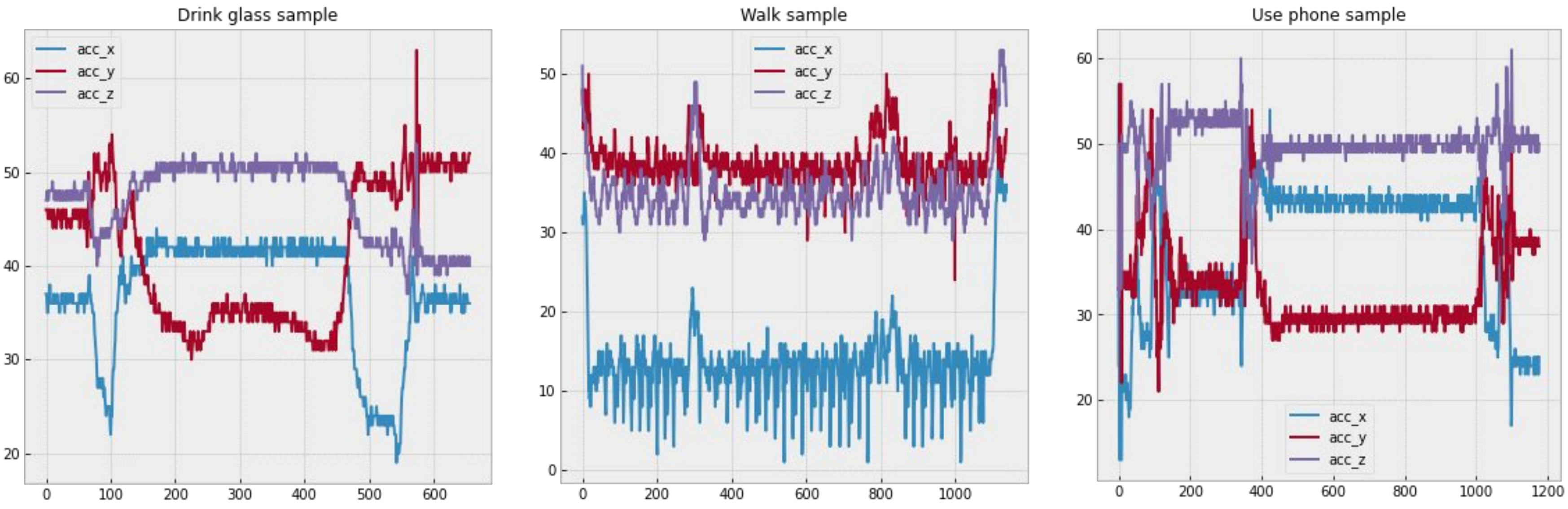
Example: activity recognition

ADL Recognition [dataset](#)

- multiple activities, short samples
- 50 Hz sampling rate



Example: activity recognition



Notes

Structure:

- various patterns
- manual features, no windowing (data is already sliced)

TS problems: forecasting

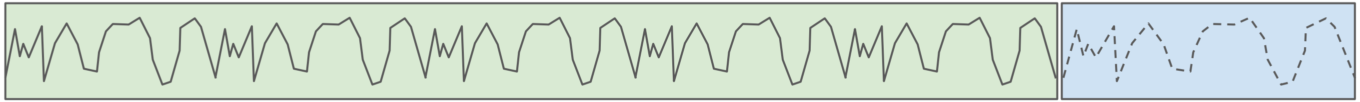
TS forecasting problem

Forecasting:

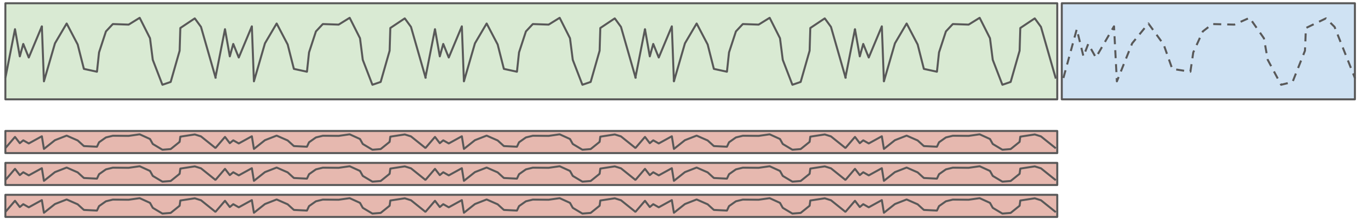
- estimate the **target time series in the future** using past data (endogenous)
- sometimes, you may know **something else besides target** (exogenous)
- depends strongly on **time scales** of relevant processes

TS forecasting setup

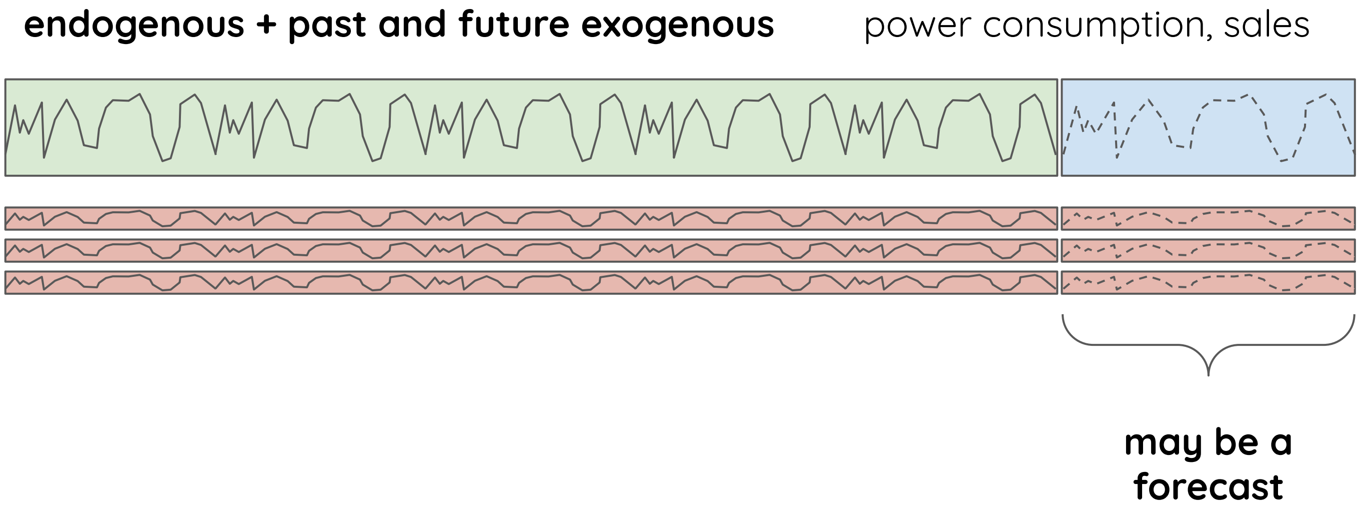
endogenous only weather time series, power consumption, sales



endogenous + past exogenous manufacturing



TS forecasting setup



TS forecasting: past and future

- future values may depend on past information
- they can depend on the future information as well
- you cannot forecast if you do not have information
- **no free lunch**

TS forecasting: past and future

Power consumption **tomorrow** depends on:

- consumption **today, yesterday**, etc.
- weather **tomorrow**,
- traffic **tomorrow**

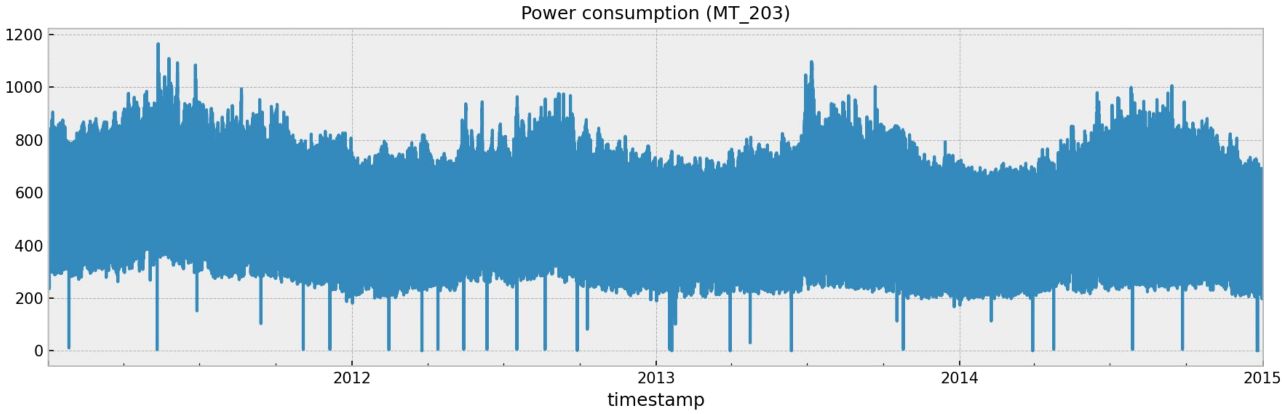
Time series: concepts

- stationarity: $X(t), X(t+1), \ldots \rightarrow X(t+h), X(t+1+h), \ldots$
- seasonality: season variations → calendar
- trend
- autocorrelation

Power consumption

Electricity load [dataset]

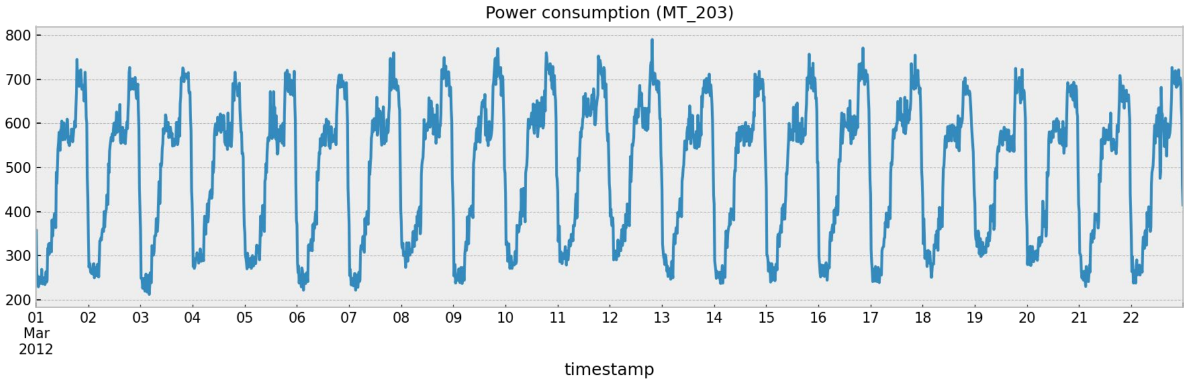
- 370 individual households
- 15 minutes sampling interval



Power consumption

Electricity load [dataset]

- 370 individual households
- 15 minutes sampling interval

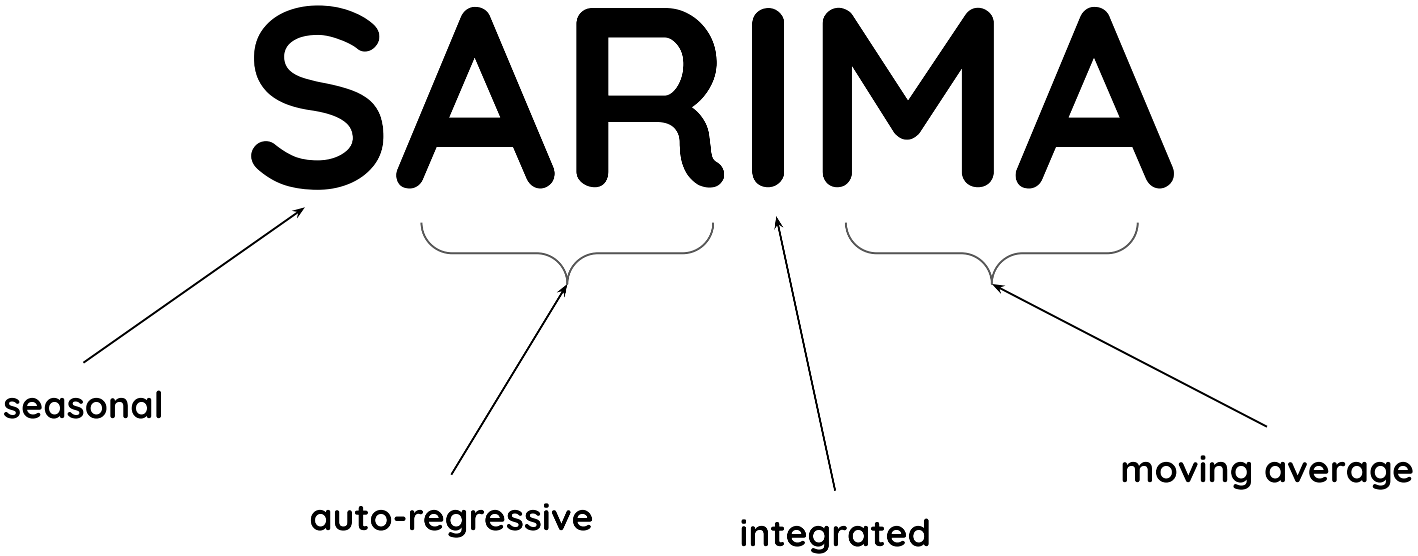


Classical models

- AR: autoregressive
- MA: moving average
- ARIMA: AR integrated moving average

More: Time Series: Autoregressive models AR, MA, ARMA, ARIMA

SARIMA


SARIMA
seasonal
auto-regressive
integrated
moving average

SARIMA: tools

`statstools`:

- a lot of time series functionality
- a lot of classical time series models
- convenient plotting

Classical models limitations

- linearity
- multiple seasonalities
- stationarity
- somewhat tricky
- **good baseline**

Cross-validation

Random split cannot be applied to time series

- use fixed split
- use rolling CV
- otherwise, autocorrelation will kill your model

Data generation process

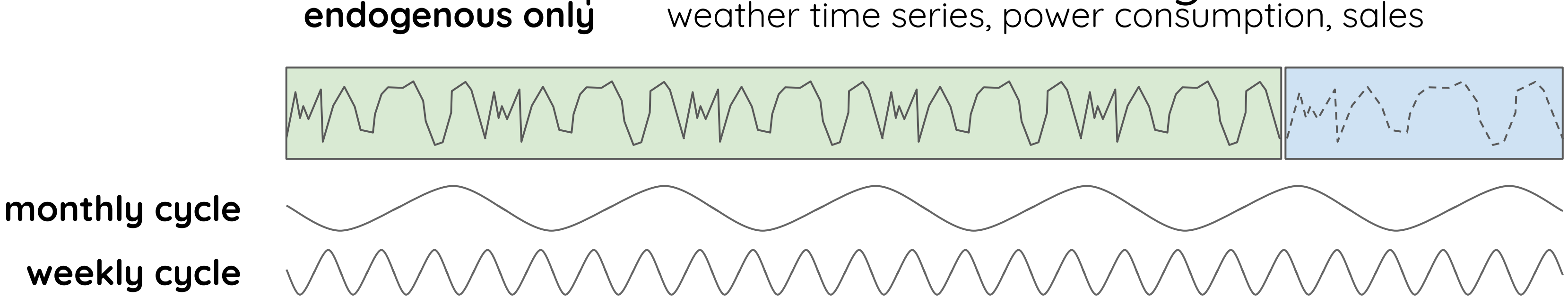
- all the underlying processes, which result in the observed data
- may be multilayered and non-linear
- not everything is known at inference time

Calendar information

Two main options:

- **one-hot encoded** (month, day of week, weekend/weekday, holidays, sale)

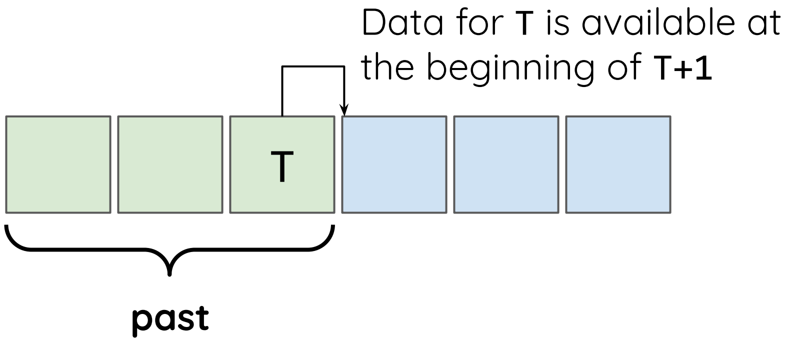
- **Fourier features:** explicit multi-seasonality

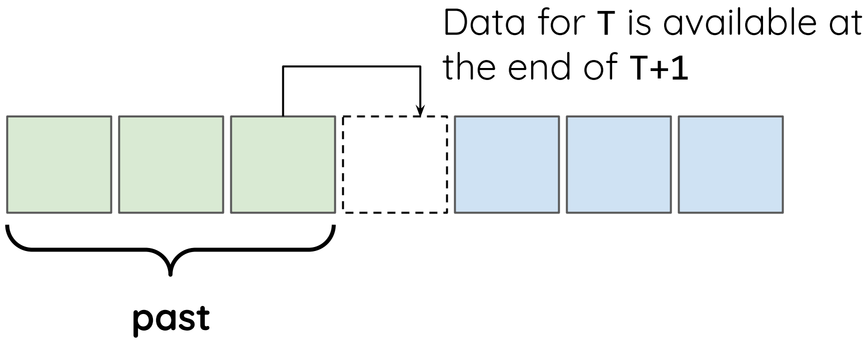


Production considerations

Forecasting windows:

- it is usually desirable to forecast on **regular intervals**
- based on data availability, you may need to **skip a window**





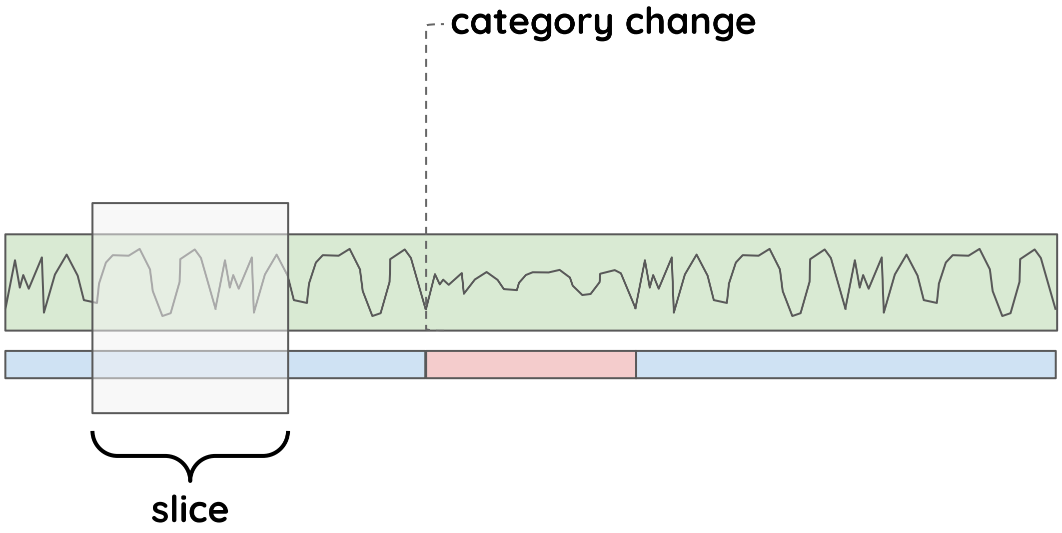
TS problems: classification and segmentation

Classification

Setup:

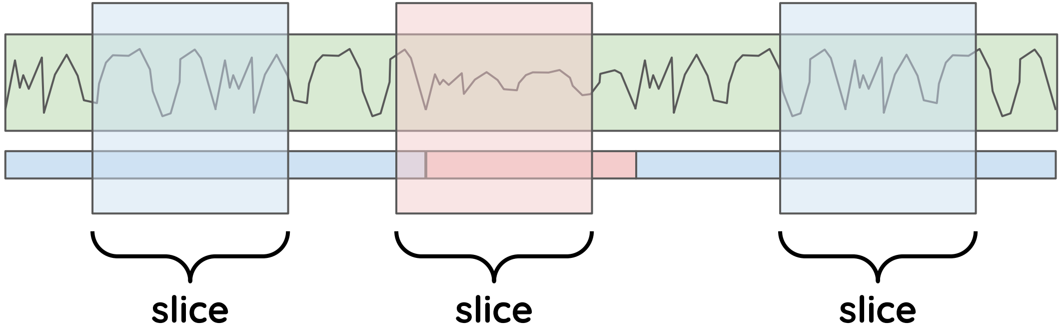
- given a **slice** of (usually multivariate) time series, get it's **category**
- **wide variety** of slice duration and typical time scales
- **patterns**, not long-term dependencies
- generally, **conceptually simpler** compared to forecasting

Patterns

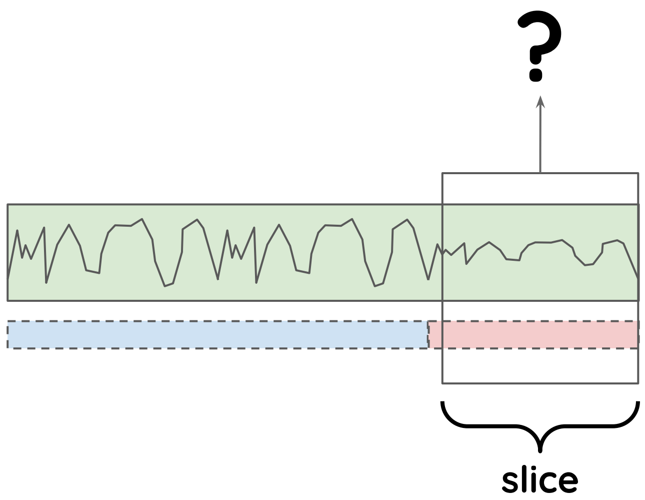


There are **no categories** if changes are **too gradual** and can be modeled with recurrent networks

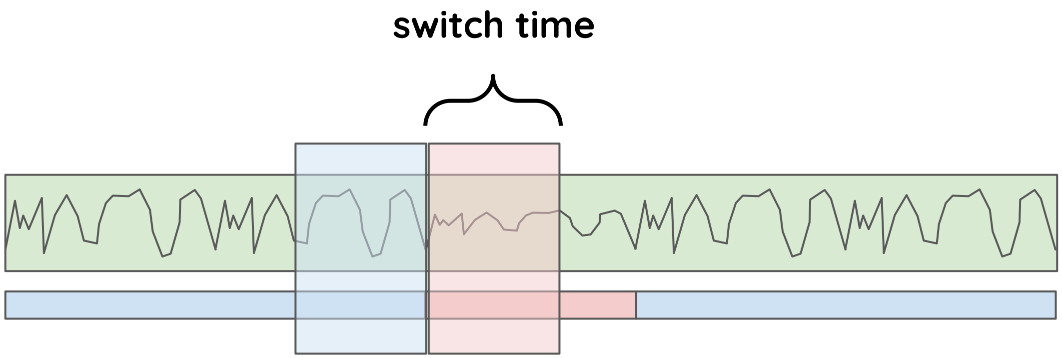
Training setup



Inference setup



Switch time



Switch time is about the size of the **classification window**. Must be **aligned** carefully to category duration time.

Classical

A lot of approaches:

- manually created windowed features + classical models
- **DTW** (dynamic time warping) as a distance measure
- etc.

Segmentation

- similar to classification in setup (slices, etc.)
- unsupervised

TS problems:
time-to-event

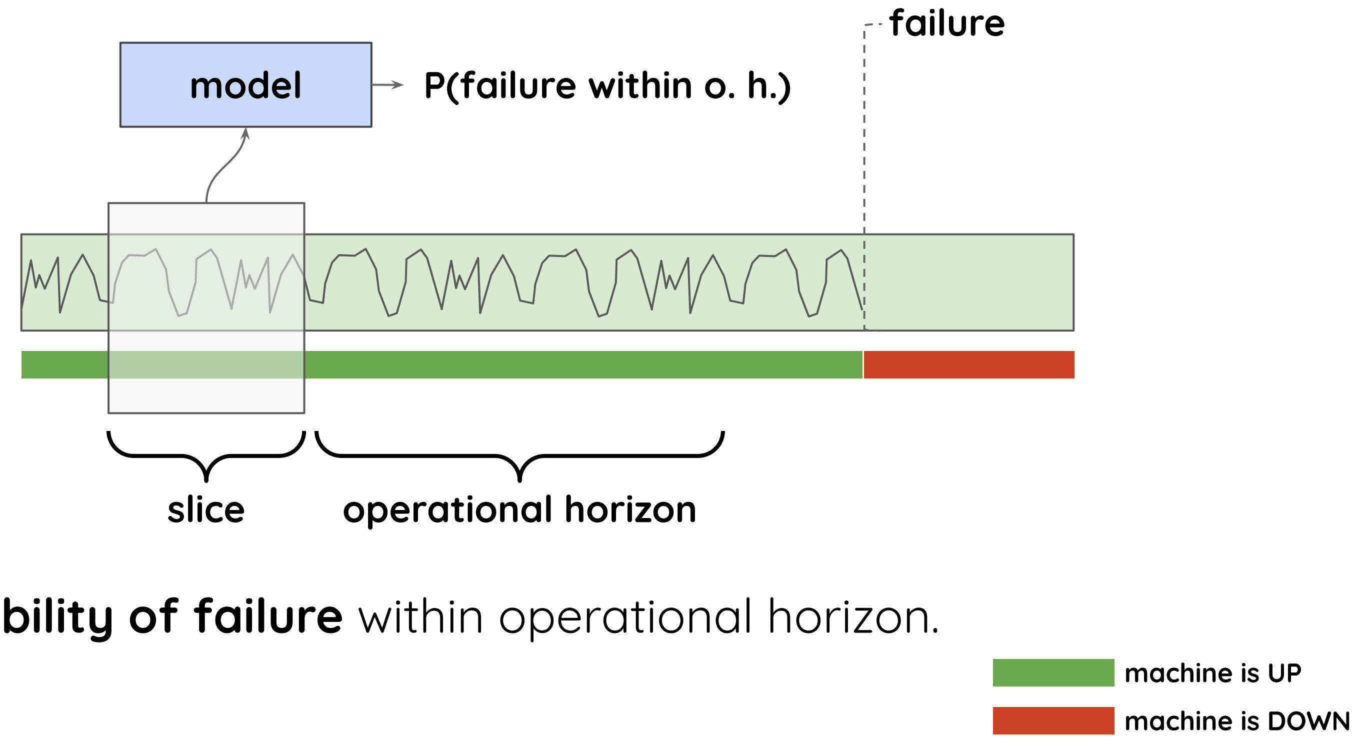
TTE and PdM

Typical scenario:

- **equipment,** vehicles, etc. fails from time to time
- **sensors** provide time series data (often used for other reasons)
- **failures data** is collected as well
- can we **predict failures using sensors data?**

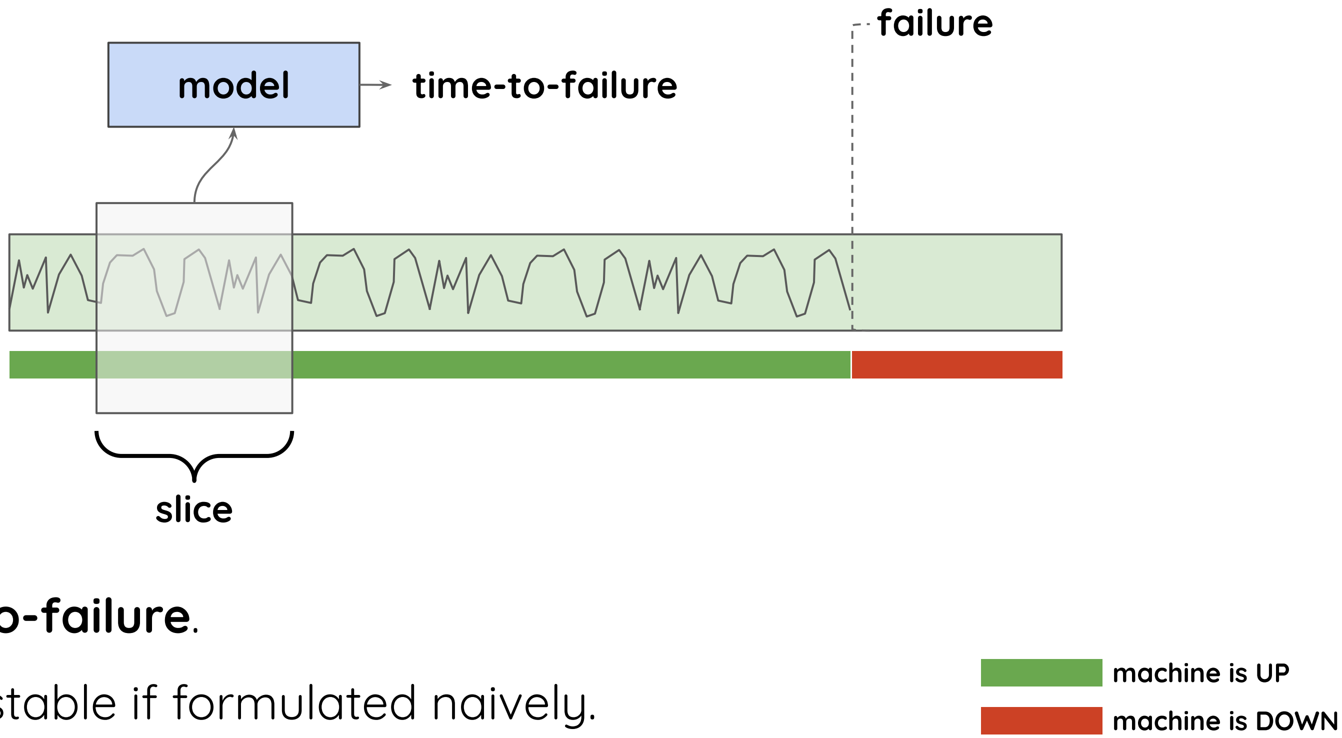
Value: improved **operational efficiency**

Setup: probability



Predict **probability of failure** within operational horizon.

Setup: TTE



Predict **time-to-failure**.

Way more unstable if formulated naively.

TTE and PdM

Naive formulation:

- create some windowed features/use deep learning model
- train a classification model
- rolling predictions

TTE and PdM

When formulated naively:

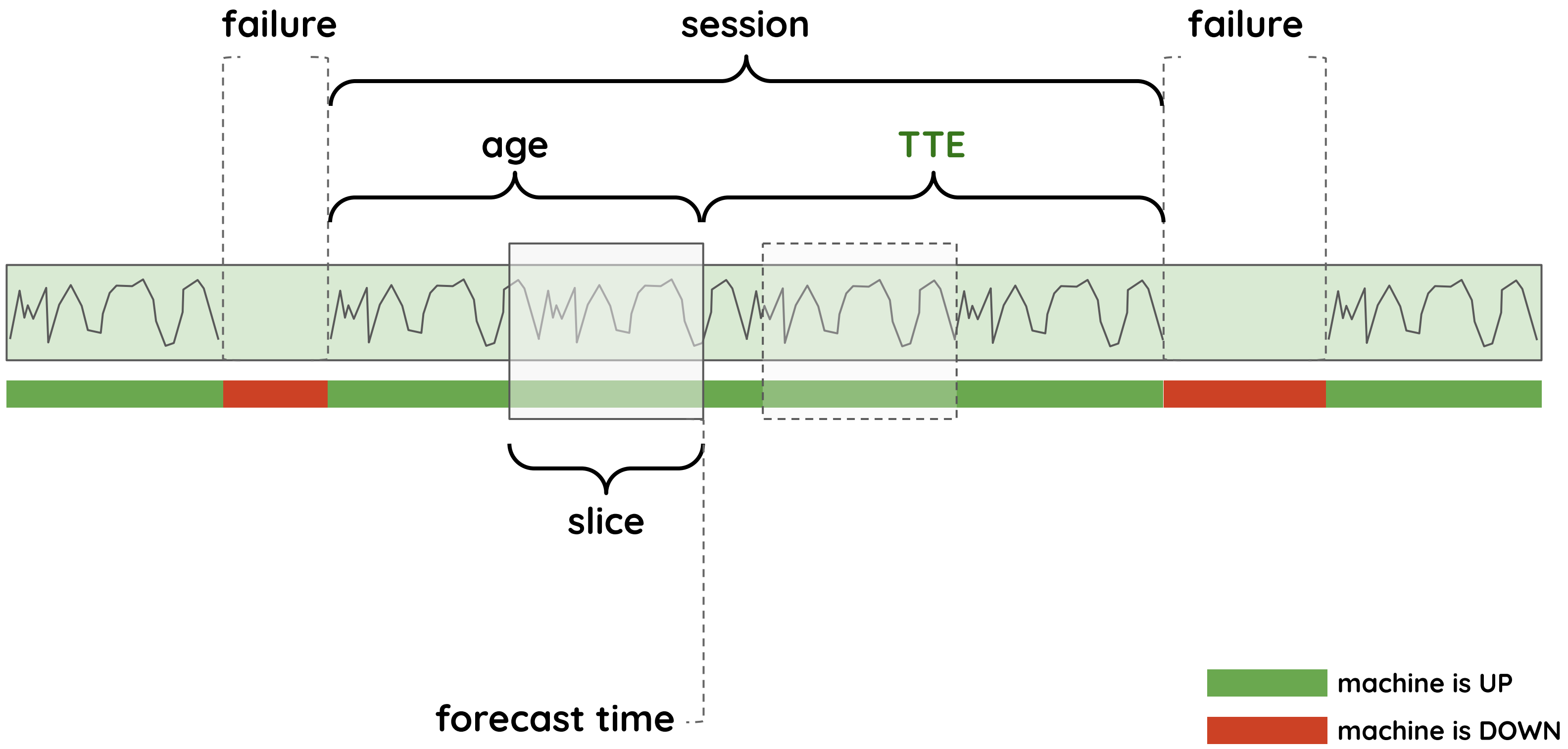
- failure probability over a **single** o. h. may be **not enough**: no planning beyond o. h.
- **hard to communicate**
- no intrinsic **risk** concept

TTE and PdM

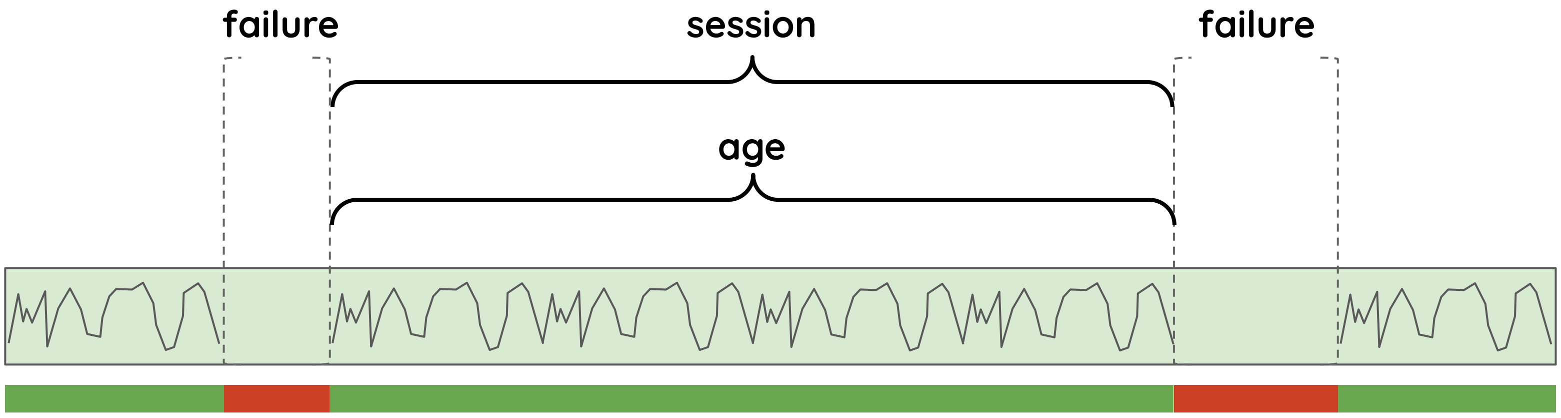
Solution:

- **survival analysis**
- well known in medicine and other domains
- has intrinsic **risk** concept
- can be married with **deep learning**

PdM data setup: slices

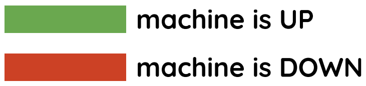


PdM data setup: sessions



One vector of covariates for entire session.

No need for time varying covariates.



Realistic PdM

Some considerations:

- model each type of failure **separately**
  (slices/sessions ended with a different failure are censored)
- session-based analysis for **post-mortem analysis**
- try session-based models for real-time predictions
  with **expanding windows** (may work for frequent failures)

TS problems:
Representation learning

Representations for t.s.

When:

- highly dimensional time series with complex patterns
- barely interpretable

Why:

- denser
- hopefully, provide some insights into structure
- simplify forecasting, classification and t.t.e.: substitute for pre-training

Representations for t.s.

Applications:

- manufacturing data
- molecular dynamics data
- various medical data

Tools

Pandas

For preprocessing and feature calculation:

- datetime operations
- resampling, rolling
- shifts

tsfresh

Features:

- simple features (statistics)
- entropy, energy, SNR
- zero-crossings, symmetry, etc.

questions?