

Part II - Time series understanding and prediction

Myra Spiliopoulou



INF

FACULTY OF
COMPUTER SCIENCE



The Knowledge Management & Discovery Lab @ U Magdeburg

Research focus:

- ▶ Machine Learning methods for streams and time series with gaps – prediction and feature contribution
- ▶ Parsimonious usage of data and features – cost-aware active feature acquisition methods
- ▶ Design of human-understandable solutions

Application areas:

- Treatment outcome prediction in clinical data
- Patient and clinic phenotyping on high-dimensional data
- Prediction in mHealth apps

- Before prediction: outcomes and phenotyping
- Formalizing the prediction task

2 Collections of timestamped data

- What for?
- Cohorts
- How to build a data collection?

3 One time series up close

- A time series and its zero position
- Missingness in time series

4 A choice of methods

- Neighbourhood-based prediction
- Learning in the latent space of predictors
- Gap prediction

5 Closing

Anomaly detection

Classification & Prediction	Description & Explanation	Anomaly detection
-----------------------------	---------------------------	-------------------

- ▶ Assign a label to a time series [Time series classification]
- ▶ Predict the value of a variable of interest using present values of other variables, and all the past [Stream classification]
- ▶ **Predict the next observation(s) from the past ones**
- ▶ Predict whether an event will occur [Survival analysis]
- ▶ Estimate the remaining time until an event [RUL]

Learning tasks on time series

Classification & Prediction

Description & Explanation

Anomaly detection

Classification & Prediction

- ▶ Assign a label to a time series [Time series classification]
- ▶ Predict the value of a variable of interest using present values of other variables, and all the past [Stream classification]
- ▶ **Predict the next observation(s) from the past ones**
- ▶ Predict whether an event will occur [Survival analysis]
- ▶ Estimate the remaining time until an event [RUL]

Description & Explanation

- ▶ **Understand how some variables predict the future of others**
- Understand how an intervention affects an outcome, and for whom

- ? What distinct 'groups' of individuals are there?
- ? How does each group respond to a (given) treatment?
- ? How does the treatment affect each group?

What characterizes each group?

A few words on phenotyping

For a disease/disorder/condition:

- ? What distinct 'groups' of individuals are there?
- ? How does each group respond to a (given) treatment?
- ? How does the treatment affect each group?

What characterizes each group?

Phenotype

the set of characteristics of a living thing, resulting from its combination of genes and the effect of its environment [BIOLOGY] ^a

all the observable characteristics of an organism that result from the interaction of its genotype (total genetic inheritance) with the environment. Examples of observable characteristics include behaviour, biochemical properties, colour, shape, and size. The phenotype may change constantly throughout the life of an individual because of environmental changes and the physiological and morphological changes associated with aging. Different environments can influence the development of inherited traits (as size, for example, is affected by available food supply) ... [GENETICS] ^b

^aOxford Advanced Learner's Dictionary,
<https://www.oxfordlearnersdictionaries.com/definition/english/phenotype> (2023-02-22)

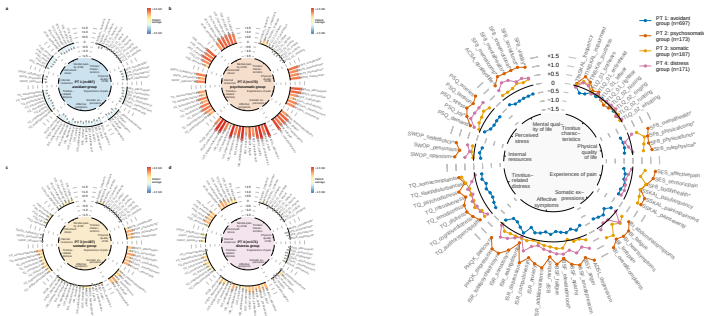
^bEncyclopaedia Britannica, <https://www.britannica.com/science/phenotype> (2023-02-22)

Phenotype $\equiv$ cluster over a feature space of observables ?

Phenotyping at baseline

[Niemann et al., 2020c]

Patients are partitioned in four phenotypes (by X-Means): radial barcharts (Figure to the left)³ & radial line chart as overview (Figure to the right)⁴



³

Figure 1. Radial barcharts visualizing the 4 phenotypes. **A** Phenotype 1 (PT1) characterises the patient subgroup with lowest health burden among all phenotypes. **B** PT2 represents the most suffering subgroup, with all of the psychosomatic and somatic measurement averages exceeding the population mean $+0.5$ standard deviations (SD). PT3 (**C**) exhibits above population average scores for somatic indicators whereas PT4 (**D**) is characterised by increased distress scores, including subjective stress and perceived quality of life. Bars are arranged in a circular layout. The height of a bar shows a feature's z-score normalised within-cluster average, and the grey line centred at the top of the bar illustrates the 95% confidence interval. The colour of a bar represents the difference of the within-cluster average from the overall patient average (PA), from -1.5 SD below PA (dark blue) to PA (yellow) and $+1.5$ SD above PA (bright red). Features were grouped into 9 categories defined by tinnitus experts. The categories are shown within the inner circle. See subsection Features for a description of each questionnaire and the extracted features.

⁴

Figure 2. Radial line chart juxtaposing the 4 phenotypes. In the chart, a point shows a feature's (z-score normalised) within-phenotype average. In each feature category (labels in inner circle), points are connected with line segments. Points and lines are coloured by cluster.

... on phenotypes

In a nutshell: Phenotypes refer to *observable* properties of *relevance* – here, of relevance to diagnostics and treatment.

Phenotypes can be built with a clustering mechanism. They must be verified from the application domain.

Phenotypes that respond differently to treatment are of particular interest.

⇒ It is of interest to predict the treatment outcome per phenotype and/or to use the phenotype as predictor variable for **treatment outcome**.

... on outcomes

In a nutshell: We refer to a 'treatment outcome', usually a score, or to 'treatment response' that can be positive (hopefully), none or negative (let's hope not).

Outcome prediction – Post-Learning step I⁵ [Niemann et al., 2020b]

Figure 4 from [Niemann et al., 2020b]

⁵ **Figure 4. SHAP analysis results for the best model (GBT, $i = 7$).** (A) Global feature importance based on the mean absolute magnitude of the SHAP values over all training instances. Values represent absolute change in log odds where higher values indicate higher feature importance. (B) Instance-individual SHAP values. A point represents the SHAP value for the feature depicted on the y-axis with respect to a single patient. The further a point from the vertical line at 0.0, the larger the attribution of the corresponding feature value to the model prediction. Vertically offset points depict regions of high density. Points are colored according to the actual feature value of the respective patient. (C) Combined SHAP feature attribution for all patients. Patients are ordered according to hierarchical clustering with complete linkage and $k = 5$. Blue horizontal lines depict the average sum of SHAP values of the cluster members.

... on outcomes and predictive features

In a nutshell: Of interest are the features that contribute to outcome/response prediction; including features that are in causal relationship to the outcome/response.

Phenotypes are of interest, too. But they must be first verified across centers.

16/89

Formalization for prediction

[Parmezan et al., 2019, Section 3]

Time series Z : sequence of observations $Z = (z_1, z_2, \dots, z_m)$ where $z_t \in R$

Two constraints: Z is discrete. Z is uniformly distributed over time ^a.

^aIf not: [Cismondi et al., 2013] describe solutions for the alignment of misaligned time series

Important properties of a time series

► **Deterministic vs Stochastic:**

- Deterministic $\leftarrow$ iff there is a mathematical function $f()$, such that the time series values are synthesized as $y = f(t)$
- Nondeterministic/Stochastic $\leftarrow y = f(t, \varepsilon)$, where ε is a random term

► **Stationary:** iff it develops around a constant average

A time series has a

- **Trend:** iff there is a long-term increase or decrease in the observed values, following an arbitrary increase (resp. decrease) pattern
- **Seasonality:** iff, next to a trend, there are also patterns that repeat in relatively constant intervals (aka: cyclic patterns)
- **Residue:** a 'component' that captures the short term fluctuations – which are not systematic and cannot be predicted.

An example time series

[Parmezan et al., 2019]

'Fig. 3(a) displays a real time series that expounds, in tons, the monthly chocolate production in Australia from January 1958 to December 1990. These measurements, provided by the Australian Bureau of Statistics, as well as all datasets adopted in this paper, are available at the ICMC-USP Time Series Prediction Repository [32].' [32] is: [Parmezan and Batista, 2014]

Figure 3a: the uppermost part of Figure 3 depicts the original time series of chocolate production, as described in the text

Formalization cntd.

[Parmezan et al., 2019, Section 3]

We can express a time series Z at timepoint t as a composition of the trend, seasonality and residual components.

Additive decomposition

$$Z_t = T_t + S_t + R_t$$

where all components use the same quantity unit as Z_t .

Multiplicative decomposition

$$Z_t = T_t \times S_t \times R_t$$

where component T_t uses the same quantity unit as Z_t and the other two components are modifiers over the trend component.

Decomposing the example time series

[Parmezan et al., 2019]

'Fig. 3(b) shows the trend, estimated using MA with 12 periods (monthly data), of the chocolate production time series;' MA:= Moving Average

Figure 3b: this part of the figure is directly below the original time series of chocolate production, and it captures the trend, as described in the text.

The example time series decomposed

[Parmezan et al., 2019]

Figure 3: complete figure, with the original time series, the trend component and two decompositions for seasonality and residue, as described in the text

Prediction workflow

[Parmezan et al., 2019, Section 4]

Figure 3: the 6 steps of the time series prediction process

Famous collections - and what they are (or not)

MIMIC

Covid-19 incidences

AirQuality dataset

AirQuality dataset

- ▶ Randomized clinical trials (RCT)
- ▶ Registries
- ▶ Experiments
- ▶ mHealth app recordings
- ▶ ...

AirQuality dataset

→

Temporal data as part of

- ▶ Randomized clinical trials (RCT)
- ▶ Registries
- ▶ Experiments
- ▶ mHealth app recordings
- ▶ ...

→ Cohorts

Cohorts

[Glenn, 2005]

The term “cohort”

Quoting [Glenn, 2005], page 2: “The term *cohort* originally referred to a group of warriors or soldiers, and the term is now sometimes used in a very general sense to refer to a number of individuals who have some characteristic in common.”

Cohorts

[Glenn, 2005]

The term “cohort”

Quoting [Glenn, 2005], page 2: “The term *cohort* originally referred to a group of warriors or soldiers, and the term is now sometimes used in a very general sense to refer to a number of individuals who have some characteristic in common.”

The term “cohort” in “cohort analysis”

Quoting [Glenn, 2005], page 2: “Here and in other literature on cohort analysis, however, the term is used in a more restricted sense to refer to those individuals (human or otherwise) who experienced a particular event during a specified period of time.

Cohorts

[Glenn, 2005]

The term “cohort”

Quoting [Glenn, 2005], page 2: “The term *cohort* originally referred to a group of warriors or soldiers, and the term is now sometimes used in a very general sense to refer to a number of individuals who have some characteristic in common.”

The term “cohort” in “cohort analysis”

Quoting [Glenn, 2005], page 2: “Here and in other literature on cohort analysis, however, the term is used in a more restricted sense to refer to those individuals (human or otherwise) who experienced a particular event during a specified period of time. The kind of cohort most often studied by social scientists is the human *birth cohort*, that is, those persons born during a given year, decade, or other period of time.”

Cohort Analysis

[Glenn, 2005]

The term “cohort analysis”

Quoting [Glenn, 2005], page 3: “The term *cohort analysis* is usually reserved for studies in which two or more cohorts are compared with regard to at least one dependent variable measured at two or more points in time.”

Cohort Analysis

[Glenn, 2005]

The term “cohort analysis”

Quoting [Glenn, 2005], page 3: “The term *cohort analysis* is usually reserved for studies in which two or more cohorts are compared with regard to at least one dependent variable measured at two or more points in time.”

Purposes of Cohort Analysis [Glenn, 2005], pages 1-2

- “Assessing the effects of aging”
- “Understand[ing] the sources and nature of social, cultural and political change.”

Cohort Analysis

[Glenn, 2005]

The term “cohort analysis”

Quoting [Glenn, 2005], page 3: “The term *cohort analysis* is usually reserved for studies in which two or more cohorts are compared with regard to at least one dependent variable measured at two or more points in time.”

Purposes of Cohort Analysis [Glenn, 2005], pages 1-2

- ⊙ “Assessing the effects of aging”
- ⊙ “Understand[ing] the sources and nature of social, cultural and political change.”

Counter-examples – [Glenn, 2005], page 3

- *Cross-sectional study*: Comparison of different groups of individuals with respect to some characteristic/variable – such a study “is conducted with data collected at one point in time, or, more accurately, within a short period of time.”
- *Panel study*: Comparison of the the attitudes of a group of individuals at two distinct timepoints – such a study “measures the characteristics of the same individuals at more than one point in time.”

Working with cohort data

Example: Study of Health in Pomerania (SHIP) [Völzke et al., 2011]

Profile of the Population-based longitudinal SHIP study:

- ▶ Selection criteria: main residence in Pomerania (Germany), age 20-79
- ▶ Cohorts and numbers

SHIP

- SHIP-0: n=4338, 1997-2001
- SHIP-1: n=3300, 2002-2006
- SHIP-2: n=2333, 2008-2012
- SHIP-3: ...

SHIP-TREND

- SHIP-TREND-0: n=4420, 2008-2012
- SHIP-TREND-1: ...

- ▶ Recordings: sociodemographics, lab tests, ultrasound, MRT, ...

Working with cohort data

Example: Study of Health in Pomerania (SHIP) [Völzke et al., 2011]

Profile of the Population-based longitudinal SHIP study:

- ▶ Selection criteria: main residence in Pomerania (Germany), age 20-79
- ▶ Cohorts and numbers

SHIP

- SHIP-0: n=4338, 1997-2001
- SHIP-1: n=3300, 2002-2006
- SHIP-2: n=2333, 2008-2012
- SHIP-3: ...

SHIP-TREND

- SHIP-TREND-0: n=4420, 2008-2012
- SHIP-TREND-1: ...

- ▶ Recordings: sociodemographics, lab tests, ultrasound, MRT, ...

Learning from population-based study data

the traditional way

- ▶ Formulate a hypothesis
- ▶ Select the data appropriate for this hypothesis: which waves, which strata, which variables?
- ▶ Perform a retrospective study on those data

Building a data collection

... from existing data

- An example case: collecting and analysing timestamped data for the rare disorder 'Charcot Foot' [Munson et al., 2014]

Disorders associated with Charcot Foot

[Munson et al., 2014]

Charcot Foot is a rare disease: the bones/joints get brittle and disintegrate.

- Charcot Foot usually follows a bone injury.
- It often appears as followup of diabetes.
- Some risk factors are known, but the pathogenesis is not completely understood.

Disorders associated with Charcot Foot

[Munson et al., 2014]

Charcot Foot is a rare disease: the bones/joints get brittle and disintegrate.

- Charcot Foot usually follows a bone injury.
- It often appears as followup of diabetes.
- Some risk factors are known, but the pathogenesis is not completely understood.

Goal of the study is to identify novel associations between Charcot Foot and other disorders/diseases, paying particular emphasis on the temporal relationship in such associations.

Disorders associated with Charcot Foot

[Munson et al., 2014]

Charcot Foot is a rare disease: the bones/joints get brittle and disintegrate.

- Charcot Foot usually follows a bone injury.
- It often appears as followup of diabetes.
- Some risk factors are known, but the pathogenesis is not completely understood.

Goal of the study is to identify novel associations between Charcot Foot and other disorders/diseases,
paying particular emphasis on the temporal relationship in such associations.

The chase for Charcot Foot cases

- ▶ *Site of the study*: University of Michigan Health System (UMHS), encompassing three hospitals with six speciality centers (including a diabetes center with a podiatric clinic)
- ▶ *Complete dataset*: 1.6 million patients with 41.2 million ICD-9 codes (timestamped)

Disorders associated with Charcot Foot

[Munson et al., 2014]

Charcot Foot is a rare disease: the bones/joints get brittle and disintegrate.

- Charcot Foot usually follows a bone injury.
- It often appears as followup of diabetes.
- Some risk factors are known, but the pathogenesis is not completely understood.

Goal of the study is to identify novel associations between Charcot Foot and other disorders/diseases, paying particular emphasis on the temporal relationship in such associations.

The chase for Charcot Foot cases

- ▶ *Site of the study*: University of Michigan Health System (UMHS), encompassing three hospitals with six speciality centers (including a diabetes center with a podiatric clinic)
- ▶ *Complete dataset*: 1.6 million patients with 41.2 million ICD-9 codes (timestamped)
- ▶ *Candidates for Charcot Foot diagnosis*: “arthropathy associated with a neurological disorder” (ICD-9 code 713.5), amounting to

Disorders associated with Charcot Foot

[Munson et al., 2014]

Charcot Foot is a rare disease: the bones/joints get brittle and disintegrate.

- Charcot Foot usually follows a bone injury.
- It often appears as followup of diabetes.
- Some risk factors are known, but the pathogenesis is not completely understood.

Goal of the study is to identify novel associations between Charcot Foot and other disorders/diseases,
paying particular emphasis on the temporal relationship in such associations.

The chase for Charcot Foot cases

- ▶ *Site of the study*: University of Michigan Health System (UMHS), encompassing three hospitals with six speciality centers (including a diabetes center with a podiatric clinic)
- ▶ *Complete dataset*: 1.6 million patients with 41.2 million ICD-9 codes (timestamped)
- ▶ *Candidates for Charcot Foot diagnosis*: “arthropathy associated with a neurological disorder” (ICD-9 code 713.5), amounting to 388 patients.

Codes associated with Charcot Foot

[Munson et al., 2014]

Method

- ▶ **Reviewing by Experts** to separate among (1) well-known associations, (2) associations that were less known / had the potential to be novel, (3) uninformative associations – either because the ICD was unspecific⁷ or because it was a misdiagnosis⁸ that was later followed by the correct one, namely "Charcot Foot"
- ▶ **Investigation of the role of diabetes** by separating between patients with Charcot Foot and diabetes ($n=282$), and those with Charcoot Foot but without diabetes ($n=106$) and investigating the dominant associations
- ▶ **Ranking of the associations** on p-values and odds ratio
- ▶ **Testing the significance of the temporal relationship**, i.e. when another diagnosis precedes the 713.5 diagnosis, using binomial test and $p < 0.001$ ⁹

⁷unspecific ICD, e.g. "viral infection, not otherwise specified"

⁸misdiagnosis like "gout, not otherwise specified"

⁹The test was on whether the one ICD-9 code preceded the other in a non-random way.

Main Findings

[Munson et al., 2014]

676 (of 710) associations with p-value < 0.001; 603 with odds ratio > 5.0

- Some were not reportedly linked to Charcot Foot but can be associated to it on the basis of existing etiology models. (e.g. bladder disorder; diseases/disorders associated with neurotrophic influences)
- Some diagnoses could be explained by diabetes, e.g. obesity, peripheral neuropathy.
- Associations that did not fit to etiology models but had very high odds ratio were: alkalosis, pulmonary eosinophilia¹⁰, esophagean reflux¹¹

111 ICD-9 codes with significant temporal relationship to Charcot Foot

- Four of them followed Charcot Foot (327.23 "obstructive sleep apnea", 786.7 "abnormal chest sounds", 353.6 "phantom limb syndrome", 786.9 "nonspecific symptoms involving the chest and respiratory system")
- Alkalosis preceded Charcot Foot 100% of the times; pulmonary eosinophilia also preceded it (significantly).

¹⁰ Pulmonary eosinophilia may be treated with steroids; these may affect bone mineral density.

¹¹ Esophagean reflux might be associated to proton pump inhibitors; -//- -//- -//-

Building a data collection

... from existing data

- ✓ An example case: collecting and analysing timestamped data for the rare disorder 'Charcot Foot' [Munson et al., 2014]
- An example workflow for iterative construction and analysis of a cohort [Zhang et al., 2014]

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

- ▶ **Goal:** get new insights about a population of patients (e.g. all patients of the cardiology unit who have hypertension)
- ▶ **Parties involved:** team of physicians + team of technologists
- ▶ **Data:** EHR of hospital patients (timeseries of patient recordings)

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

- ▶ **Goal:** get new insights about a population of patients (e.g. all patients of the cardiology unit who have hypertension)
- ▶ **Parties involved:** team of physicians + team of technologists
- ▶ **Data:** EHR of hospital patients (timeseries of patient recordings)

Conventional workflow – from [Zhang et al., 2014] **with extensions**

At the beginning, there is a question/observation – a concrete phenomenon that must be explained (cf. use cases in [Zhang et al., 2014]).

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

- ▶ **Goal:** get new insights about a population of patients (e.g. all patients of the cardiology unit who have hypertension)
- ▶ **Parties involved:** team of physicians + team of technologists
- ▶ **Data:** EHR of hospital patients (timeseries of patient recordings)

Conventional workflow – from [Zhang et al., 2014] **with extensions**

At the beginning, there is a question/observation – a concrete phenomenon that must be explained (cf. use cases in [Zhang et al., 2014]).

1. The (team of) physician(s) devise one or more hypotheses.
2. **The physicians specify the cohort needed for the study of each hypothesis, possibly in interaction with a data analyst or DB expert.**
3. The DB expert writes scripts to create the cohort and extract the data.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

- ▶ **Goal:** get new insights about a population of patients (e.g. all patients of the cardiology unit who have hypertension)
- ▶ **Parties involved:** team of physicians + team of technologists
- ▶ **Data:** EHR of hospital patients (timeseries of patient recordings)

Conventional workflow – from [Zhang et al., 2014] **with extensions**

At the beginning, there is a question/observation – a concrete phenomenon that must be explained (cf. use cases in [Zhang et al., 2014]).

1. The (team of) physician(s) devise one or more hypotheses.
2. **The physicians specify the cohort needed for the study of each hypothesis, possibly in interaction with a data analyst or DB expert.**
3. The DB expert writes scripts to create the cohort and extract the data.
4. Data analysts build models **according to the instructions of the physicians, e.g. on age and gender adjustment.**
5. Physicians become a **presentation**/visualization of the model(s) and check whether their hypothesis is supported.
6. If necessary, GOTO 2.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Expanding the workflow to incorporate ...

[Zhang et al., 2014]

- *Early cohort definition:* The physicians must be able to define a cohort themselves in an ad hoc way, whenever they see fit (cf. steps 2 and 3 of the conventional workflow).
- *Flexible visualization:* The physicians must be able to inspect the cohort in different ways, without having to ask the technologists.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Expanding the workflow to incorporate ...

[Zhang et al., 2014]

- ▶ *Early cohort definition*: The physicians must be able to define a cohort themselves in an ad hoc way, whenever they see fit (cf. steps 2 and 3 of the conventional workflow).
- ▶ *Flexible visualization*: The physicians must be able to inspect the cohort in different ways, without having to ask the technologists.
- ▶ *Flexible analysis*: The physicians must be able to invoke analytics modules and use them to perform analytics tasks without having to ask the technologists.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Expanding the workflow to incorporate ...

[Zhang et al., 2014]

- ▶ *Early cohort definition*: The physicians must be able to define a cohort themselves in an ad hoc way, whenever they see fit (cf. steps 2 and 3 of the conventional workflow).
- ▶ *Flexible visualization*: The physicians must be able to inspect the cohort in different ways, without having to ask the technologists.
- ▶ *Flexible analysis*: The physicians must be able to invoke analytics modules and use them to perform analytics tasks without having to ask the technologists.
- ▶ *Cohort refinement and expansion*: The physicians must be able to modify themselves the cohort, i.e. the choice of patients and the choice of variables on them (cf. steps 6 and 1 of the conventional workflow).

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Expanding the workflow to incorporate ...

[Zhang et al., 2014]

- ▶ *Early cohort definition*: The physicians must be able to define a cohort themselves in an ad hoc way, whenever they see fit (cf. steps 2 and 3 of the conventional workflow).
- ▶ *Flexible visualization*: The physicians must be able to inspect the cohort in different ways, without having to ask the technologists.
- ▶ *Flexible analysis*: The physicians must be able to invoke analytics modules and use them to perform analytics tasks without having to ask the technologists.
- ▶ *Cohort refinement and expansion*: The physicians must be able to modify themselves the cohort, i.e. the choice of patients and the choice of variables on them (cf. steps 6 and 1 of the conventional workflow).
- ▶ *Iterative analysis*: Cohort definition, visualization, analysis, refinement and expansion may need to be performed repeatedly, on the results of the previous iterations.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Expanding the workflow to incorporate ...

[Zhang et al., 2014]

- ▶ *Early cohort definition*: The physicians must be able to define a cohort themselves in an ad hoc way, whenever they see fit (cf. steps 2 and 3 of the conventional workflow).
- ▶ *Flexible visualization*: The physicians must be able to inspect the cohort in different ways, without having to ask the technologists.
- ▶ *Flexible analysis*: The physicians must be able to invoke analytics modules and use them to perform analytics tasks without having to ask the technologists.
- ▶ *Cohort refinement and expansion*: The physicians must be able to modify themselves the cohort, i.e. the choice of patients and the choice of variables on them (cf. steps 6 and 1 of the conventional workflow).
- ▶ *Iterative analysis*: Cohort definition, visualization, analysis, refinement and expansion may need to be performed repeatedly, on the results of the previous iterations.

i.e. foster interaction between physician and system in a complete workflow, taking the technologists out of the workflow.

Building a data collection from existing data DOs & DONTs

[1] DO write down with indelible ink:

- ▶ the research question
- ▶ the kind of data you need [inclusion criteria]
- ▶ the data you need to exclude [exclusion criteria]
- ▶ the target variable(s) and the algorithm that computes the ground truth
- ▶ the evaluation procedure

DOs & DONTs

[1] DO write down with indelible ink:

- ▶ the research question
- ▶ the kind of data you need [inclusion criteria]
- ▶ the data you need to exclude [exclusion criteria]
- ▶ the target variable(s) and the algorithm that computes the ground truth
- ▶ the evaluation procedure

After collecting the data:

[2] DO crosscheck:

- ? whether the starting point, end point and length of each and all time series agrees with your research question
 - ? whether the ground truth can be computed for each time series
 - ? whether you want all the diseases, medications etc that are in the data
- If NOT, then goto [1].

Building a data collection from existing data DOs & DONTs

[1] DO write down with indelible ink:

- ▶ the research question
- ▶ the kind of data you need [inclusion criteria]
- ▶ the data you need to exclude [exclusion criteria]
- ▶ the target variable(s) and the algorithm that computes the ground truth
- ▶ the evaluation procedure

After collecting the data:

[2] DO crosscheck:

- ? whether the starting point, end point and length of each and all time series agrees with your research question
 - ? whether the ground truth can be computed for each time series
 - ? whether you want all the diseases, medications etc that are in the data
- If NOT, then goto [1].

- ! DONT modify/refine/adjust the research questions
- ! DONT change the definition of the ground truth
- ! DONT add/remove time series

Takeaway message: Treat the data collection as a cake recipe. Decide what it should look like, prepare the ingredients and make the cake. If it feels like necessary to do modifications, never add/remove ingredients!
Go for the next cake!

Modeling time

Given is a time unit, and a sequence of time points $t_0, t_1, t_2, \dots, t_n$ where the step between two points corresponds to one time unit, and t_0 is the zeroth time point.

A time series with *absolute time points* has the form $T = v_{t_0}, \dots, v_{t_n}$ where $v_{\cdot}$ is a vector in the multidimensional feature space $F_1 \times \dots \times F_k$.

Each of $F_1, \dots, F_k$ contains also the special value NULL.

Alignment at zero

For two time series T_A, T_B , if the t_0 of T_A is the same absolute time point as the t_0 of T_B , then T_A, T_B are *aligned at zero*.

Example 1 from [Chang et al., 2021]

Figure 2 from [Chang et al., 2021]

Sampling rate

Two time series may have different *sampling rates* (and thus time units), even if they belong to the same entity.

[Unnikrishnan, 2017]

[Unnikrishnan, 2017]

- The time units are retained; the absolute time point positions are not.
- Information about seasonality is lost.

Modeling time without time units

A time series with absolute time points $T = v_{t_0}, \dots, v_{t_n}$ where v_- can be mapped to the ordered sequence $T_{seq} = v_0, v_1, \dots, v_{n'}$, where $n' \leq n$ and v_- is a vector in the multidimensional NULL-free feature space $F'_1 \times \dots \times F'_k$ ¹².

¹² $F'_i \leftarrow F_i \setminus \{NULL\}$, for each $i = 1, \dots, k$

Modeling time without time units

A time series with absolute time points $T = v_{t_0}, \dots, v_{t_n}$ where v_- can be mapped to the ordered sequence $T_{seq} = v_0, v_1, \dots, v_{n'}$, where $n' \leq n$ and v_- is a vector in the multidimensional NULL-free feature space $F'_1 \times \dots \times F'_k$ ¹².

Consequences of collapsing time series to ordered sequences

- ▶ All time series become aligned at zero.
- ▶ The more gaps the original time series had, the shorter becomes the collapsed ordered sequence.
- ▶ Information about seasonality, periodicity and circadian phenomena is lost.

¹² $F'_i \leftarrow F_i \setminus \{NULL\}$, for each $i = 1, \dots, k$

... Part II - Time series unders

Misalignment and missingness in TS

[Cismondi et al., 2013]

¹³The figures are remakes of Figure 1 from [Cismondi et al., 2013]

Misalignment and missingness in TS

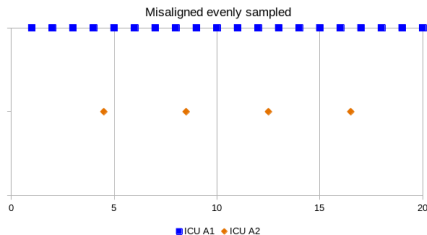
[Cismondi et al., 2013]

Time series context: Patient monitoring over a period of time T ¹³

Misaligned evenly sampled as in

ICU [Cismondi et al., 2013]

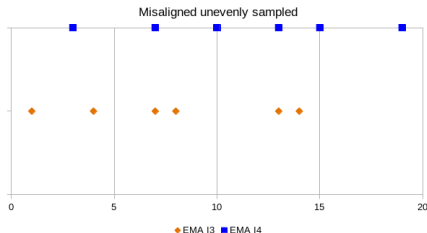
- Blood pressure: hourly
- Lab tests: every 4 hours



Misaligned unevenly sampled as in

TrackYourTinnitus [Probst et al., 2017]

- 8 questions, all answers optional
- each one answered independently
- triggered more than once a day



¹³The figures are remakes of Figure 1 from [Cismondi et al., 2013]

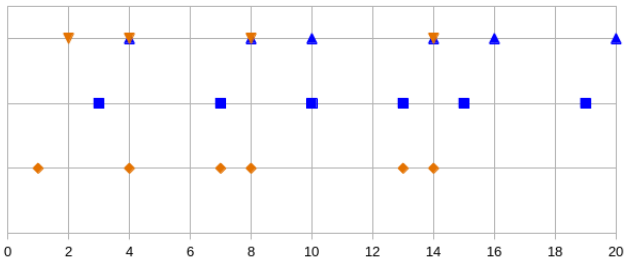
Alignment of misaligned unevenly sampled data: Gridding

[Cismondi et al., 2013]

- A sampling rate is specified and a grid defined on the time axis.
 - Each value of *each variable* is shifted to occupy the closest grid position^a
- ⇒ all variables get values **or NULLs** at the same positions – of the grid.

^aIf more than one values fall into the same grid position, a predefined mechanism is used to determine the final value at the grid position.

Misaligned unevenly sampled TS: aligned on a grid of 2 time units



Alignment of misaligned unevenly sampled data: Templating

[Cismondi et al., 2013]

- One of the variables is chosen as template – usually the one with the highest sampling rate.
The time points at which this variable is sampled are used as grid positions.
- Each value of each other variable is shifted as in gridding.

Missing and/or Observed [NOT] at Random

[Cismondi et al., 2013]

Let D be a dataset with n variables, let Y be the one whose missingness we investigate, and X the remaining variable(s).

For each variable, we have observed values and missing values, i.e.

$$Y = \{Y_{obs}, Y_{miss}\}$$

$$\forall X \neq Y : X = \{X_{obs}, X_{miss}\}$$

Missing and/or Observed [NOT] at Random

[Cismondi et al., 2013]

Let D be a dataset with n variables, let Y be the one whose missingness we investigate, and X the remaining variable(s).

For each variable, we have observed values and missing values, i.e.

$$Y = \{Y_{obs}, Y_{miss}\}$$

$$\forall X \neq Y : X = \{X_{obs}, X_{miss}\}$$

MAR, MCAR, MNAR for Y

- ▶ MAR: Probability of Y_{miss} is independent of Y ; can depend on X .
- ▶ MCAR: Probability of Y_{miss} is independent of Y and of X .
- ▶ MNAR: Probability of Y_{miss} depends on (present/absent values of) Y ; can depend on X .

Missing and/or Observed [NOT] at Random

[Cismondi et al., 2013]

Let D be a dataset with n variables, let Y be the one whose missingness we investigate, and X the remaining variable(s).

For each variable, we have observed values and missing values, i.e.

$$Y = \{Y_{obs}, Y_{miss}\} \quad \forall X \neq Y : X = \{X_{obs}, X_{miss}\}$$

MAR, MCAR, MNAR for Y

- ▶ MAR: Probability of Y_{miss} is independent of Y ; can depend on X .
- ▶ MCAR: Probability of Y_{miss} is independent of Y and of X .
- ▶ MNAR: Probability of Y_{miss} depends on (present/absent values of) Y ; can depend on X .

Dependency of Y on X

- ▶ ONAR: Likelihood of Y_{miss} depends on X
- ▶ OAR: Likelihood of Y_{miss} does not depend on X

$$Y_{miss} \perp\!\!\!\perp X$$

$$Y_{miss} \not\perp\!\!\!\perp X$$

i.e. no difference on X_{obs} between Y_{miss} and Y_{obs}

[Cismondi et al., 2013]

The missing values dilemma: Impute the missing values, in order to increase statistical power, **vs** leaving the NaNs as they are, because there are significant differences between observed and unobserved data

Classifying TS on missingness

[Cismondi et al., 2013]

The missing values dilemma: Impute the missing values, in order to increase statistical power, **vs** leaving the NaNs as they are, because there are significant differences between observed and unobserved data



Classify-before-imputing

Classification Level 2: Recoverability

Generalizing from [Cismondi et al., 2013]

Is there a (set of) variable(s) X so that $Y_{miss} \perp\!\!\!\perp X$?

- Identify a set of variables X that can distinguish between Y_{miss} and Y_{obs} ^a
- Check whether X is empty:
 - ▶ If YES, then mark Y_{miss} as 'recoverable' ^b
 - ▶ ELSE mark Y_{miss} as 'non-recoverable' and do not attempt to derive them. ^c

^aThis is a learning problem.

^bThen, Y_{miss} & Y_{obs} have the same behaviour given X , so X and Y_{obs} can recover Y_{miss} .

^cThis means that X dictated the decision to not sample Y at the location(s) of Y_{miss} .

Classification Level 2: Recoverability

Generalizing from [Cismondi et al., 2013]

Is there a (set of) variable(s) X so that $Y_{miss} \perp\!\!\!\perp X$?

- Identify a set of variables X that can distinguish between Y_{miss} and Y_{obs} ^a
- Check whether X is empty:
 - ▶ If YES, then mark Y_{miss} as 'recoverable' ^b
 - ▶ ELSE mark Y_{miss} as 'non-recoverable' and do not attempt to derive them. ^c

^aThis is a learning problem.

^bThen, Y_{miss} & Y_{obs} have the same behaviour given X , so X and Y_{obs} can recover Y_{miss} .

^cThis means that X dictated the decision to not sample Y at the location(s) of Y_{miss} .

Approach of [Cismondi et al., 2013] for classification level 2

- Split the data (excluding Y) into training set (,validation set) and test set
- Create a (target) variable Y' that takes the value 0 for each observed Y_{obs} and the value 1 for each empty location, i.e. Y_{miss}
- Build up a set of fuzzy rules that distinguish between the two values of the target variable
- Apply the set of rules on the held-out dataset

Missing values: to recover or not to recover ? Example cntd

Assume that the data are EMA (diary entries), inserted after a reminder from an mHealth app; the user decides to ignore the reminder (Y/N), and this decision is not transparent $\Rightarrow$ Classification level 1: TrulyMissing

Can we recover the missing values of I4 in the cases where I3 is recorded?

A simplistic classification rule: IF $I3 \geq 60$ THEN $Y'_{miss}=1$ (some users)
 $\Rightarrow$ missing values of I4 are not recoverable

Ecological Momentary Assessments

... in apps for wellness and health/disease monitoring

One of the UNITI apps – for the RCT participants (and not only) ¹⁴

¹⁴H2020 project on the 'Unification of Treatments and Interventions for Tinnitus Patients' (2020-2023)

BUT ...

'Most mHealth entries are between baseline and final visit, but outliers exist'

15

¹⁵by Anne Kleinau & Simon Flügel, master students at Univ Magdeburg

Neighbourhood-augmented prediction on 'data-sparse' TS

[Unnikrishnan et al., 2019, Unnikrishnan et al., 2020], [Unnikrishnan et al., 2021],

[Unnikrishnan et al., 2023]

Neighbourhood-augmented prediction on 'data-sparse' TS

[Unnikrishnan et al., 2019, Unnikrishnan et al., 2020], [Unnikrishnan et al., 2021],

[Unnikrishnan et al., 2023]

Core idea: Learn a predictor on the data of the entity itself and augment this predictor with information from the entity's neighbours

Neighbourhood-augmented prediction on 'data-sparse' TS

[Unnikrishnan et al., 2019, Unnikrishnan et al., 2020], [Unnikrishnan et al., 2021],

[Unnikrishnan et al., 2023]

Core idea: Learn a predictor on the data of the entity itself and augment this predictor with information from the entity's neighbours

Gap treatment: Treated during the vectorisation

Sparsity treatment: Separate between 'short' and 'long' entities; exploit the long entities to learn on the short ones



Neighbourhoods for prediction

RQ 1: How to define similarity between entities on their time series?

RQ 2: How to exploit similarity for prediction?

RQ 3: How to identify and exploit only those neighbours of an entity that contribute to better prediction?

Neighbourhood-augmented prediction

[Unnikrishnan et al., 2021]

RQ1: How to define similarity between entities on their time series?

Multi-layered representation of an entity

[Unnikrishnan et al., 2020]

- ① Static component $\leftarrow$ vector of registration data (no missing values)
- ② Dynamic component $\leftarrow$ vector of EMA features
 - value of each EMA feature for each time unit
 - for all time units within the temporal segment used for learning
 - after aligning at zero and collapsing consecutive observations into adjacent time units

Re-defining the grid for the dynamic component

Hyperparameters to tune:

- ▶ Grid *Coarsening*^a
- ▶ Length of the temporal segment used for learning^b
- ▶ Position of the segment

[Begin, Mid, End]

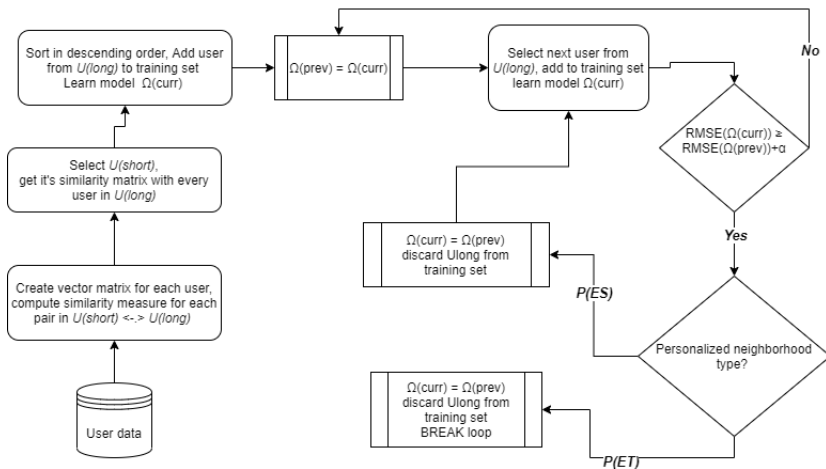
^aFor EMA recorded many times a day, the coarsened grid unit can become 'day' or 'week'.

^bPart or all of the time series

Neighbourhood-augmented prediction

[Unnikrishnan et al., 2021]

RQ3: How to identify and exploit only those neighbours of an entity that contribute to better prediction?



Neighbourhood-augmented prediction – New results

Three patient empowerment studies

- ★ Bulgarian study (diabetes melitus): $n = 10$ users contributing 387 EMA
 - ★ Spanish study (diabetes type II): $n = 12$ users contributing 650 EMA
- and German study on tinnitus, description in [Unnikrishnan et al., 2019]

Dataset	#Short entities	RMSE		
		Early termination	Exhaustive search	Global model
Bulgaria	5 of 10	25.36	23.18	23.16
Spain	8 of 12	14.71	13.40	30.55
Germany	260	13.9	19.6	23.87



The benefits of early termination (when expanding the neighbourhood) increase with the proportion of short entities.

[Unnikrishnan et al., 2023]

- ▶ 83-90% of users are better predicted through their neighbourhood,
- ▶ better performance by 13%-15%
- ▶ on two datasets (an observational study dataset and the UNITI rct dataset)

... on neighbourhood-augmented prediction

Takeaway message: Neighbourhood-based prediction of an entity's future demands less resources than a global predictor and achieves comparable performance. But the selection of the similar users must be done intelligently; not all similar users are informative towards an entity's predictor.

Next: How to find the rules that govern an entity's evolution?

Granger causality

[Rodrigues and Andrade, 2013]

Informally speaking:

Assume two time series Y and X . We want to predict X .

Restricted model: We predict $Y(t+1)$ using only the past of Y .

Unrestricted model: We check whether adding knowledge from X until and including t improves this prediction.

If prediction is improved by the unrestricted model, then X granger-causes Y .

Granger causalities on 'data-sparse' TS

[Jamaludeen et al., 2021]

Granger causalities on 'data-sparse' TS

[Jamaludeen et al., 2021]

Core idea: Map each multivariate time series on a set of cross-variable sequences, where the antecedent granger-causes the consequent

Gap treatment: Each variable contributes with all its occurrences; no alignment, since the precedence relation must be preserved

Sparsity treatment: Each entity/participant supports some of the rules, some entities support more rules



Granger-causal relationships among EMA

RQ 1: To what extent does the past status of one of the EMA variables cause/predict the current status of another EMA variable within the same day?

RQ 2: How to find correlations among the inferred $\mathcal{G}$ -causal relationships involving the EMA-related variables?

RQ 3: How to associate the $\mathcal{G}$ -causal relationships among the EMA-related variables to answers given to the registration questionnaires?

Granger Causal relationships among EMA [Jamaludeen et al., 2021]

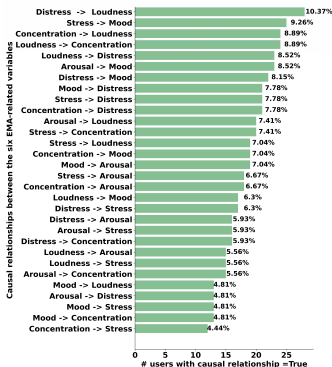
RQ1. Finding Circadian Conditional Granger Causal Relationships^{16 17}

- For each user u and EMA variable X_i , test whether $X_i^u \xrightarrow{\text{ccgCauses}} X_j^u$ by building

- the unrestricted model which includes the K lagged variables for lags up to p .
- the restricted model, where X_i^u is excluded

with the (circadian) constraint that only values within the same day are considered.

- For each circadian conditional granger causal relationship (CCGC), count its support among the users



¹⁶The time series of some users are excluded after testing on stationarity and normality.

¹⁷Fig. 1. Number of users with the plotted $\mathcal{G}$ -causal relationships true. Percentages indicate the frequency of the corresponding $\mathcal{G}$ -causal relationship of the total number of users (270 users).

CCGC-relationships among EMA

[Jamaludeen et al., 2021]

RQ2. Identifying associations among CCGC-Relationships ¹⁸

- 1 Treat each CCGC-relationship as an item
- 2 Build pairs of CCGC-relationships and derive association rules
- 3 Compute the confidence and lift for each association rule

CCGC-Relationship A	Co-occurring CCGC-relationship B	Confidence ($A \rightarrow B$)	Lift ($A \rightarrow B$)
Mood $\xrightarrow{\text{ccgCauses}}$ Loudness	Mood $\xrightarrow{\text{ccgCauses}}$ Distress	0.54	6.92
Loudness $\xrightarrow{\text{ccgCauses}}$ Stress	Distress $\xrightarrow{\text{ccgCauses}}$ Stress	0.4	6.35
Stress $\xrightarrow{\text{ccgCauses}}$ Loudness	Stress $\xrightarrow{\text{ccgCauses}}$ Distress	0.47	6.09
Loudness $\xrightarrow{\text{ccgCauses}}$ Arousal	Loudness $\xrightarrow{\text{ccgCauses}}$ Stress	0.33	6
Loudness $\xrightarrow{\text{ccgCauses}}$ Arousal	Distress $\xrightarrow{\text{ccgCauses}}$ Arousal	0.33	5.62
Loudness $\xrightarrow{\text{ccgCauses}}$ Mood	Distress $\xrightarrow{\text{ccgCauses}}$ Mood	0.41	5.05

¹⁸Table II. Most frequent co-occurring CCGC-relationships

CCGC-relationships among EMA

[Jamaludeen et al., 2021]

RQ3. Linking CCGC-Relationships to the registration data¹⁹

- 1 Consider each question (from a registration questionnaire) and each answer as an item
- 2 and build all rules that have a (question,answer)-pair in the antecedent and a CCGC-relationship in the antecedent
- 3 Perform rule filtering, subject to support, confidence and lift constraints

CCGC-Relationship c	$ U_c / U $	Top-12 predictive Features at registration (Question, Answer)	Confidence $(c \rightarrow (q,a))$	Prior Probability (q,a)	Lift $(c \rightarrow (q,a))$
Mood $\xrightarrow{\text{ccgCauses}}$ Arousal	7.04%	TSCHQ.q28="Usually"	0.21	0.04	5.17
		TSCHQ.q20="YES"	0.21	0.05	4.37
Distress $\xrightarrow{\text{ccgCauses}}$ Mood	8.15%	TSCHQ.q14="noise"	0.18	0.04	4.09
Distress $\xrightarrow{\text{ccgCauses}}$ Arousal	5.93%	TSCHQ.q15="very high frequency"	0.25	0.07	3.55
Concentration $\xrightarrow{\text{ccgCauses}}$ Mood	7.04%	TSCHQ.q26="YES"	0.21	0.06	3.55
		TSCHQ.q28="Sometimes"	0.21	0.06	3.55
Arousal $\xrightarrow{\text{ccgCauses}}$ Concentration	5.56%	TSCHQ.q29="NO"	0.27	0.08	3.43
Loudness $\xrightarrow{\text{ccgCauses}}$ Arousal	5.56%	TSCHQ.q3="Both Sides"	0.33	0.10	3.33
Loudness $\xrightarrow{\text{ccgCauses}}$ Distress	8.52%	worstsymptom ="Because of the tinnitus I am more sensitive to environmental noises."	0.30	0.09	3.29

¹⁹TABLE: Rules sorted on lift, top-7 rules selected

... on the rules of entity evolution

Takeaway message: Granger causalities shed some light on how entities evolve. But conventional GC cannot cover the complex temporal relationships among the (usually misaligned and gappy) time series of mHealth applications. Rather, associations among the (circadian) (conditional) granger-causal relationships are needed to capture entity evolution.

But: the relationships among variables may evolve as well. e.g. during human learning !
[Jamaludeen et al., 2022]

Learning from little timestamped data

- ✓ Prediction on time series with gaps: What to learn from the neighbours?
- ✓ Exploiting predictors for interpretation: What to learn from neighbours in a latent space?
- Understanding time series with gaps: When will the gap close?

Key concepts for gaps in EMA time series [Schleicher et al., 2022a]

- ★ **Adherence:** 'the extent to which a person's behaviour . . . corresponds with agreed recommendations' [1]²⁰
- ★ **Engagement:** 'efforts by a user to start and continue with an intervention' [2]²¹
- ★ **Compliance:** '. . . main difference is that adherence requires the patient's agreement to the recommendations' [1]
- ★ **Dropout attrition:** 'the phenomenon of losing participants to follow-up (eg, participants do not return to fill in follow-up questionnaires)' [3]²²
- ★ **Nonusage attrition:** 'the phenomenon of nonusage' [3]
- ★ **Phases of attrition** [3]
 - ▶ I – *Curiosity plateau* is 'the initial phase where the users are interested in exploring a new technology',
 - ▶ II – *Attrition phase:* 'the users start to reject the usage, . . .'
 - ▶ III – *Stable use phase* 'where only the hardcore users remain [and] will continue to use the application for a long time'

²⁰[1] World Health Organization and others: Adherence to long-term therapies: evidence for action. World Health Organization (2003)

²¹[2] Cavanagh, K.: Turn on, tune in and (don't) drop out: engagement, adherence, attrition, and alliance with internet-based interventions. Oxford guide to low intensity CBT interventions 227, 232 (2010)

²²[3] Eysenbach, G.: The law of attrition. J. of Medical Internet Research 7(1), e402 (2005)

Predicting the gaps

[Schleicher et al., 2022a]

Core idea: Gaps are modeled as time series elements that have a position and a length

Gap treatment: Prediction of gap occurrence, length and/or end

Sparsity treatment: The gaps are filled with 'gap elements'



Gap length prediction when the attrition rates change

RQ1: How to trace changes in the attrition rates?

RQ2: How to predict the transition from one attrition phase to the next?

RQ3: How to predict the gap length, once a gap is there?

[Schleicher et al., 2022a]

- 1 Change point detection on user activity
- 2 Prediction of the phase transition modelled as 3-class separation task
- 3 Assignment of the gaps in the training part of the time series to bins
- 4 In the test data: prediction of a gap's length by
 - deducing the phase – XBoost classifier, and
 - the bin to which the gap belongs – lazy classification task with 1NN, using Dynamic Time Warping to rank a time series neighbours

[Schleicher et al., 2022b]

- 1 Map a user's time series into a sequence of sequences: data + gaps
- 2 Bin the gap lengths
- 3 Add the binned gaps into the series of sequences
- 4 Cluster the sequences on similarity
- 5 Map each user into a vector of cluster IDs
- 6 Place each vector in a recommender matrix – – training part only
- 7 At the beginning of each gap, recommend the gap bin size (with MF)

If there are too few data for most users: See neighbourhood-augmented predictors.

Closing

We have seen key concepts:

- ✓ Key terms: time series, outcome, phenotype, cohort
- ✓ Formalisation of the prediction task on time series
- ✓ Missingness

We have discussed general methods:

- ✓ Workflow for prediction
- ✓ Workflows for cohort construction
- ✓ Alternatives for the alignment of a time series at zero
- ✓ Dealing with missingness through gridding and templating

Closing

We have seen key concepts:

- ✓ Key terms: time series, outcome, phenotype, cohort
- ✓ Formalisation of the prediction task on time series
- ✓ Missingness

We have discussed general methods:

- ✓ Workflow for prediction
- ✓ Workflows for cohort construction
- ✓ Alternatives for the alignment of a time series at zero
- ✓ Dealing with missingness through gridding and templating

For the case of

- ▶ only few time series
- ▶ consisting mainly of empty space

...we have seen specialized methods

- ✓ Neighbourhood-based augmentation
- ✓ Predictors in a latent space of patterns
- ✓ Ways of predicting the length of a gap

We have skipped:

- all of imputation (MICE etc)

- all of imputation (MICE etc)
- all of deep learning for time series

We have skipped:

- all of imputation (MICE etc)
- all of deep learning for time series → see Part III tomorrow!

- time series on human/animal experiments – a category by themselves

We have skipped:

- all of imputation (MICE etc)
- all of deep learning for time series → see Part III tomorrow!
- time series on categorical data
- time series on human/animal experiments – a category by themselves

How to use these materials?

- Use the methods and insights of sections 1 and 2 to build cohorts
- You now have more options of dealing with missingness:
 - ▶ align or classify the missingness
 - ▶ predict the gaps
- For short time series: better use its neighbours rather than a global model, but choose the neighbours carefully

The KMD team



Prof. Myra
Spiliopoulou



Christian
Beyer



Noor
Jamaludeen



Miro
Schleicher



Saijal
Shahania



Clara
Puga



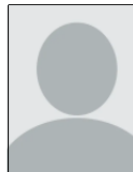
Hafez
Kader Omar



Anne
Rother



Vishnu
Unnikrishnan



Maik
Büttner

THANK YOU !

Bibliography II



Jamaludeen, N., Unnikrishnan, V., Brechmann, A., and Spiliopoulou, M. (2022). Discovering instantaneous granger causalities in non-stationary categorical time series data. In *20th Int. Conf. on Artificial Intelligence in Medicine (AIME)*. Springer.



Jamaludeen, N., Unnikrishnan, V., Pryss, R., Schobel, J., Schlee, W., and Spiliopoulou, M. (2021). Circadian conditional granger causalities on ecological momentary assessment data from an mhealth app. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 354–359.



Lin, Q., Zhang, D., Huang, X., Ni, H., and Zhou, X. (2012). Detecting wandering behavior based on gps traces for elders with dementia. In *2012 12th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 672–677. IEEE.



Munson, M. E., Wrobel, J. S., Holmes, C. M., and Hanauer, D. A. (2014). Data mining for identifying novel associations and temporal relationships with charcot foot. *Journal of Diabetes Research*, 2014.

Bibliography III



Niemann, U., 8th International Conference on Data Science, M. S. I., (DSAA)Iopoulou, A. A., Malanowski, J., Kellersmann, J., Szczepanski, T., Klose, S., Dedonaki, E., Walter, I., Ming, A., and Mertens, P. R. (2020a). Plantar temperatures in stance position: A comparative study with healthy volunteers and diabetes patients diagnosed with sensoric neuropathy. *EBioMedicine*, 54:102712.



Niemann, U., Boecking, B., Brueggemann, P., Mebus, W., Mazurek, B., and Spiliopoulou, M. (2020b). Tinnitus-related distress after multimodal treatment can be characterized using a key subset of baseline variables. *PLOS ONE*, 15(1):1–18.



Niemann, U., Brueggemann, P., Boecking, B., Rose, M., Spiliopoulou, M., and Mazurek, B. (2020c). Phenotyping chronic tinnitus patients using self-report questionnaire data: cluster analysis and visual comparison. *Scientific Reports*, 10(1):16411.



Parmezan, A. R. S. and Batista, G. E. A. P. A. (2014). ICMC-USP time series prediction repository. Technical report, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brasil.



Parmezan, A. R. S., Souza, V. M. A., and Batista, G. E. A. P. A. (2019). Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Information Sciences*.



Probst, T., Pryss, R. C., Langguth, B., Rauschecker, J. P., Schobel, J., Reichert, M., Spiliopoulou, M., Schlee, W., and Zimmermann, J. (2017). Does tinnitus depend on time-of-day? an ecological momentary assessment study with the “trackyourtinnitus” application. *Frontiers in Aging Neuroscience*, 9:253.



Rodrigues, J. and Andrade, A. (2013). Instantaneous granger causality with the hilbert-huang transform. *International Scholarly Research Notices*. 2013.

Bibliography VI



Unnikrishnan, V., Beyer, C., Matuszyk, P., Niemann, U., Pryss, R., Schlee, W., Ntoutsis, E., and Spiliopoulou, M. (2019). Entity-level stream classification: exploiting entity similarity to label the future observations referring to an entity. *International Journal of Data Science and Analytics*, pages 1–15.



Unnikrishnan, V., Schleicher, M., Puga, C., Pryss, R. P., Vogel, C., Schlee, W., and Spiliopoulou, M. (2023). A similarity-guided framework for error-driven discovery of patient neighbourhoods in ema data. In *Proceedings of the Symposium on Intelligent Data Analysis (IDA 2023)*, Louvain-la-Neuve, Belgium.

Bibliography VII



Unnikrishnan, V., Shah, Y., Schleicher, M., Fernández-Viadero, C., Strandzheva, M., Velikova, D., Dimitrov, P., Pryss, R., Schobel, J., Schlee, W., and Spiliopoulou, M. (2021). Love thy neighbours: A framework for error-driven discovery of useful neighbourhoods for one-step forecasts on EMA data. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 295–300. IEEE.



Unnikrishnan, V., Shah, Y., Schleicher, M., Strandzheva, M., Dimitrov, P., Velikova, D., Pryss, R., Schobel, J., Schlee, W., and Spiliopoulou, M. (2020). Predicting the health condition of mhealth app users with large differences in the number of recorded observations - where to learn from? In Appice, A., Tsoumakas, G., Manolopoulos, Y., and Matwin, S., editors, *DS*, volume 12323 of *Lecture Notes in Computer Science*, pages 659–673. Springer.

Bibliography VIII



Völzke, H., Alte, D., Schmidt, C. O., Radke, D., Lorbeer, R., Friedrich, N., Aumann, N., Lau, K., Piontek, M., Born, G., et al. (2011). Cohort profile: the Study of Health In Pomerania. *Int. J. of Epidemiology*, 40(2):294–307.



Zhang, Z., Gotz, D., and Perer, A. (2014). Iterative cohort analysis and exploration. *Information Visualization (Info Vis)*, pages 1–19.