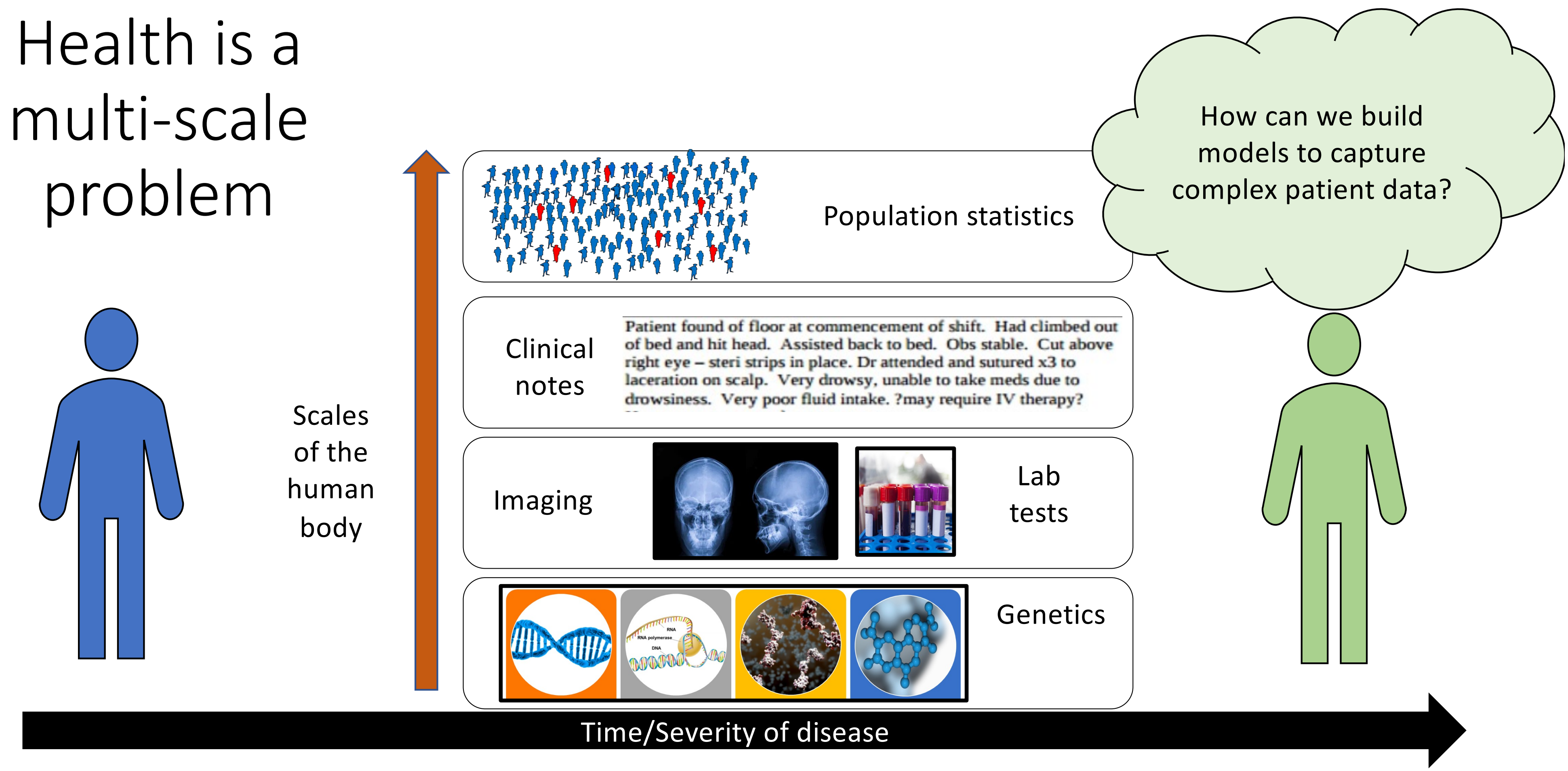
Topics in Machine Learning
Machine Learning for Healthcare

Rahul G. Krishnan
Assistant Professor
Computer science & Laboratory Medicine and Pathobiology

Outline

- Time series data in healthcare
  - Data in cardiology
  - Data in chronic disease care
  - Tasks for machine learning
  - Univariate time series models
  - Multi-variate time series models



Health is a multi-scale problem
Scales of the human body
Population statistics
How can we build models to capture complex patient data?
Clinical notes
Patient found of floor at commencement of shift. Had climbed out of bed and hit head. Assisted back to bed. Obs stable. Cut above right eye – steri strips in place. Dr attended and sutured x3 to laceration on scalp. Very drowsy, unable to take meds due to drowsiness. Very poor fluid intake. ?may require IV therapy?
Imaging
Lab tests
Genetics
Time/Severity of disease

Time in healthcare

- If you’re visiting the doctor just once, your visit may fall into one of the following:
  - Annual check up,
  - A minor issue that needs a referral,
  - A very severe issue (intensive trauma, late stage cancer) that is too late to be treated,
- In reality, **many** problems in healthcare involve time-varying (or longitudinal data).

Time-series data in healthcare

- Population level:
  - Infection statistics for various diseases are tracked at the local, provincial, federal level
    - Used to inform and guide policy decisions
- Hospital level:
  - Weekly admission statistics to the emergency department are tabulated, tracked and forecast
    - Used to guide weekly staffing policies. e.g. nurse schedules
- Individual level:
  - Critical care
  - Chronic diseases

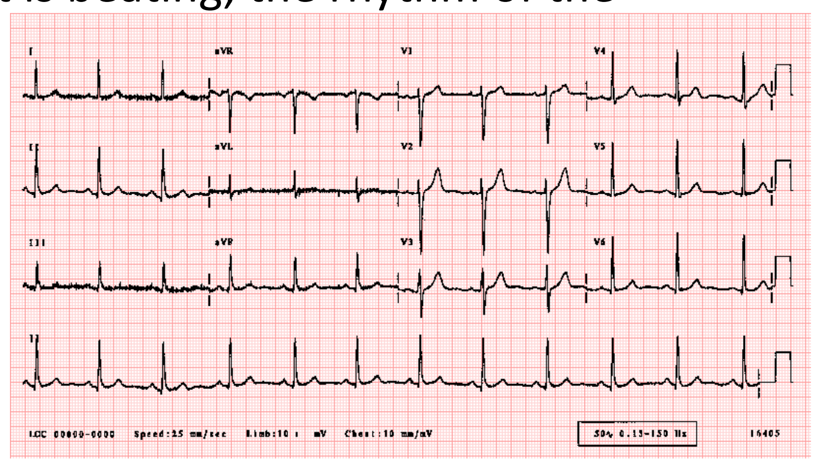
Patients in critical care units

Time-series data in critical care patients

- Often suffer from one or more severe conditions underlying the reason they are in the ICU,
- The goal of doctors in the ICU is often twofold:
  - Keep patient state stable
  - Treat the underlying disease burden
- Many different sensors, each tracking a different physiologic time-varying signal
- Many examples of data that are sampled and tracked at a high-frequency

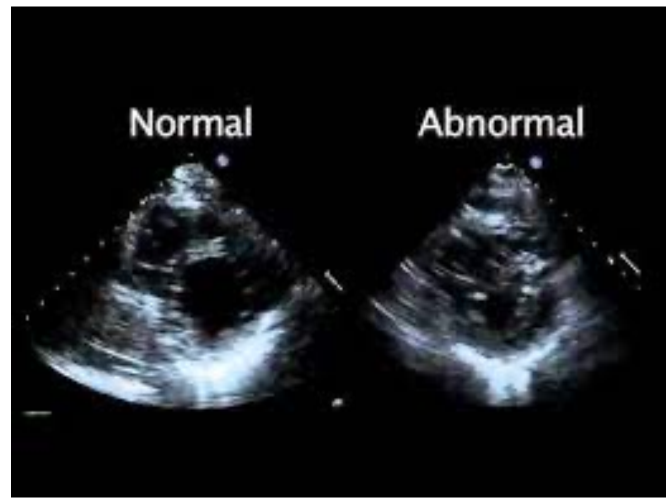
Physiological time-series data 1 [cardiology]

- Electrocardiogram:
  - A simple way to evaluate the functioning of the heart
  - Electrodes placed at different parts of the body and measure/interpret heart functioning
  - **Why does it work:** Natural electric impulses govern contractions of the heart. By measuring them, we can assess how fast it is beating, the rhythm of the heartbeat and the strength of the pulses
  - **Diseases:** Congestive heart failure
  - Type of data: continuous time



Physiological time-series data 2 [cardiology]

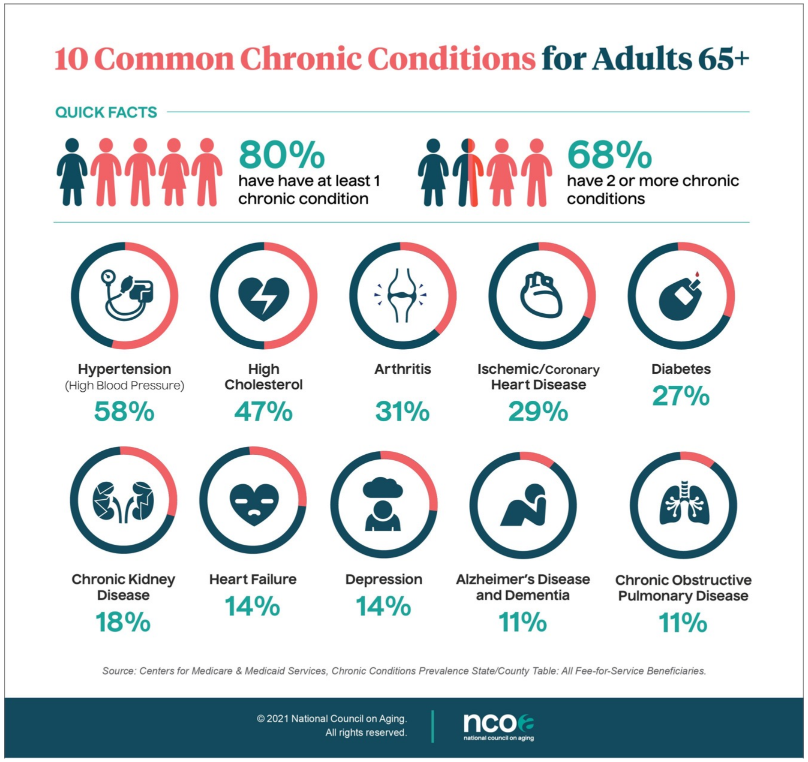
- Transthoracic **echocardiography** (TTE)
  - Widely used diagnostic tests in cardiology. Ultrasound of the heart
  - Characterize size and shape of the heart, pumping capacity, and the location of any tissue damage
- Type of data: video [time series of images]



Source: **Role of Echocardiography in the Intensive Care Unit: Overview of the Most Common Clinical Scenarios** ,Longobardo et. al, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6664324/

Patients suffering from chronic diseases

- Chronic diseases are defined broadly as conditions that last 1 year or more and:
  - Require ongoing medical attention
  - Limit activities of daily living
  - Both of the above
- The American Cancer Society views cancer as a chronic disease when the cancer can be controlled with treatment, becomes stable, or reaches remission.



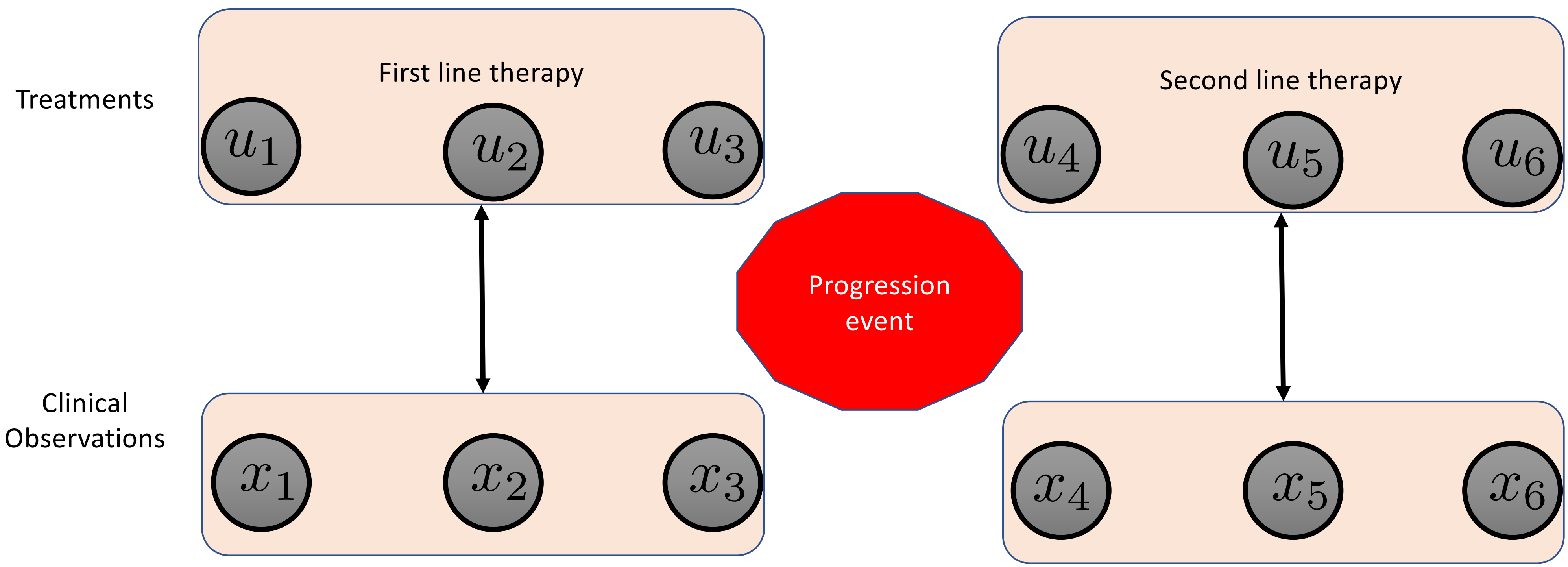
Chronic Disease Management – (1)

- Canonical picture that characterizes how healthcare data behave
- Interesting and useful structure in how chronic diseases are treated

Chronic Disease Management – (2)

- Treatments are often grouped across time
- Each line denotes an implicit plan that the clinician has on how to treat a patient
- The first line of therapy is generally what is recommended by clinical trials based on a match between patient characteristics and trial cohorts

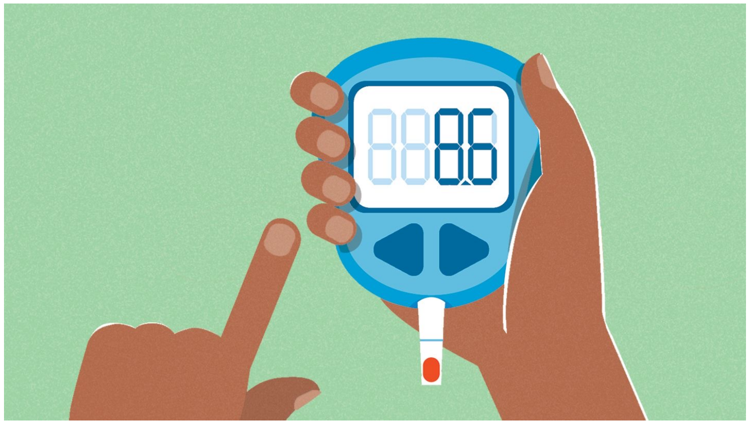
Chronology of chronic disease therapy



Progression events

- Progression events mark the failure of a line of therapy
  - Death
  - Patient did not respond
  - Patient cannot tolerate the medication
- Move onto the next line of medication
- Chronic disease care is personalized by care providers

Diabetes care and management

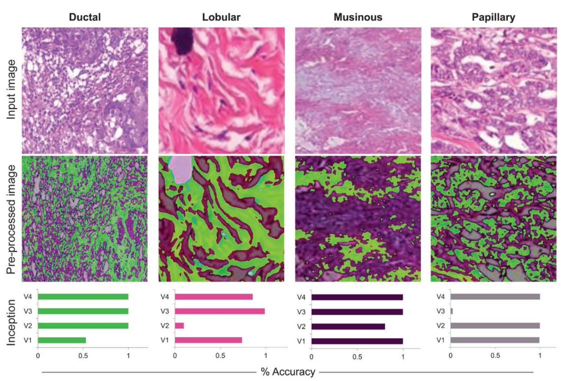


- Biomarkers:
  - Blood sugar (A1C) levels
- Interventions
  - First line: Metformin
  - Second line:
    - Combination therapy: Metformin + Sulfonylurea drug

Source: https://www.cadth.ca/sites/default/files/pdf/second_line_therapy_for_type_2_diabetes_in_brief_e.pdf

What does this mean for data?

- Chronic disease care involves data collection at regular time intervals
  - Typically, intervals between data are a few weeks or months
- Data types:
  - Longitudinal lab-values and treatments
  - Genetics
  - Imaging



Tasks for machine learning

- Risk stratification with time-series data
  - All the same techniques we saw previously except our conditioning set x now comprises a time-series
- Pattern discovery in time-series data
  - K-means is easy to apply on static data
  - What about noisy, missing, time-varying data?
- Forecasting
  - Can we use statistical models to predict how a patient might evolve over time
  - Counterfactual reasoning is an important topic
    - Condition on aspects of the data that can change how observations behave over time

Challenges for machine learning

- Clinical decision making is multi-modal
- Frequency of observations and interventions can vary dramatically:
  - Intensive care unit: Observations and interventions happening in real-time
    - High-frequency data
  - Chronic disease management: Observations and interventions happen over the span of months or years
    - Low-frequency data
- Missingness is rampant
  - ICU: sensor noise
  - Chronic disease management: administrative errors, access to health insurance

Preprocessing for time-series data

- For static data:
  - Z-scoring
  - Min-max normalization
- For temporal data:
  - Normalization by standard reference measures (healthy values)
  - Log-transformation
  - Removing the mean of a time-series
  - Normalization to [-1 , 1]
  - Outlier removal
    - Not a good idea to remove if signal is in tails of the distribution
- Imputation for missing data:
  - Feed-forward imputation
  - Linear interpolation
  - Polynomial interpolation
  - We'll see more advanced imputation strategies later in the class

Learning problems with time-series data

- One of the best ways to learn about statistical models for time-series data is to know what you can do with them,
  - Unsupervised learning
    - Forecasting – predict time-series into the future
    - Identify and detect patterns and clusters in time-series
  - Supervised learning:
    - Make predictions from time-series
- Lets turn our attention to focus on forecasting
  - To do forecasting, we often need a *model of the time-series*
  - *We'll start with the task of modeling a single biomarker*

Univariate models of time series data

Time-series regression with time-series features

$$y_t = c + \theta_1 x_{t-1} + \theta_2 x_{t-2} + \ldots + \theta_3 x_{t-k}$$

- Treat time-series modeling as a *linear* regression problem
- x: features (potentially time-varying)
- y: outcome of interest
- But what if we had no other features?

ARIMA [**A**uto**R**Egressive **I**ntegrated **M**oving **A**verage]

$$y'_t = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t$$

- ARIMA(p,d,q) model
  - p: order of autoregressive part
  - d: degree of differencing
  - q: order of moving average

Pro: Very flexible model of time-series data!
Con: linear additive model

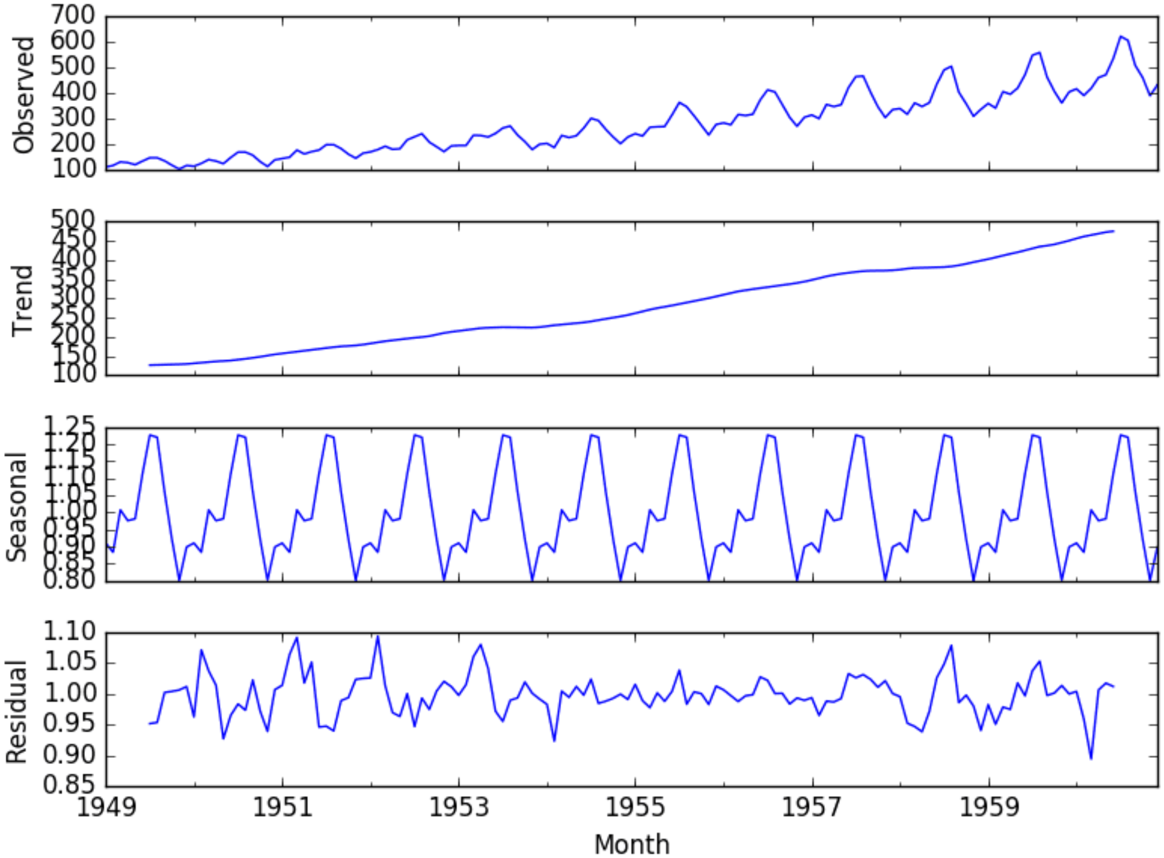
Reference: **Forecasting: Principles and Practice,** Rob J Hyndman and George Athanasopoulos

Nonlinear models of univariate time-series data

$$y_t = f(x_{t-1}, \ldots, x_{t-k}; \theta)$$

- Very general formulation for a broad class of time series problems with nonlinear models
- Theta represent the parameters of this model
- Next, we'll study a single case study of the use of a such a non-linear model to make predictions from electrocardiogram data

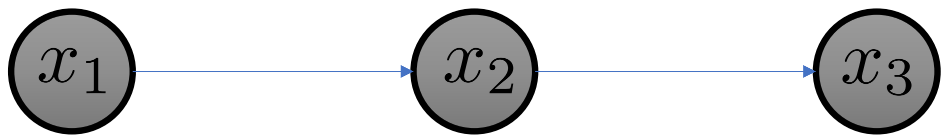
General rule: Decompose time-series



When you think about modeling time-series data, think about trends and patterns that exist and how to design models to capture different variation.

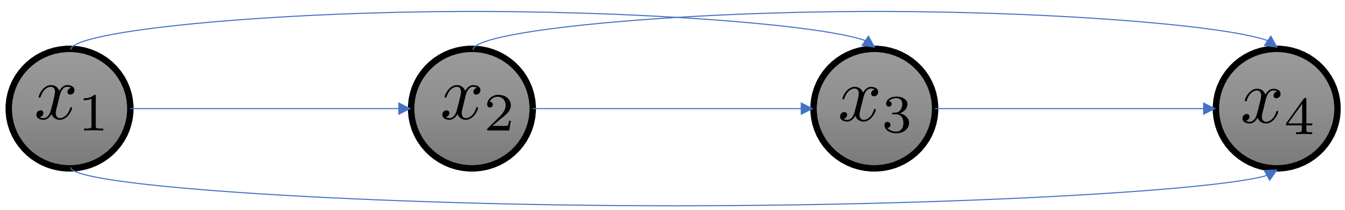
Multivariate models of time series data

First-order Markov models



$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$$

K-gram models

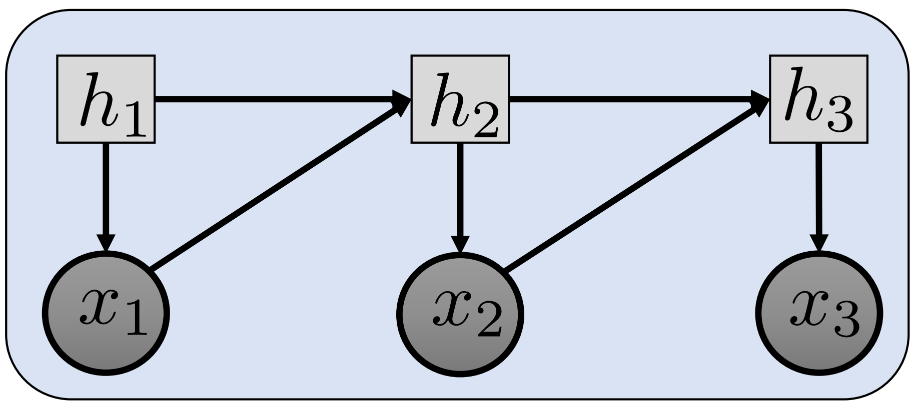


$$p(x_1, x_2, \ldots, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_{1\ldots2})p(x_4|x_{1\ldots,3})$$

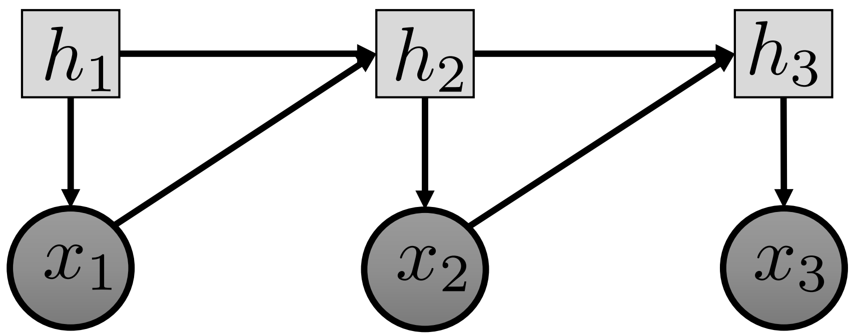
Recurrent Neural Networks

- Auto-regressive sequential models of data
- Forward recurrent neural network model
  - Each **hidden state** summarizes all the variables in the **past**

$$p(x_1, x_2, x_3) = p(x_1|h_1)\hat{p}(h_2|h_1)p(x_2|h_2)\hat{p}(h_3|h_2)p(x_3|h_3)$$

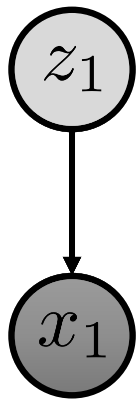


Recurrent neural networks in action

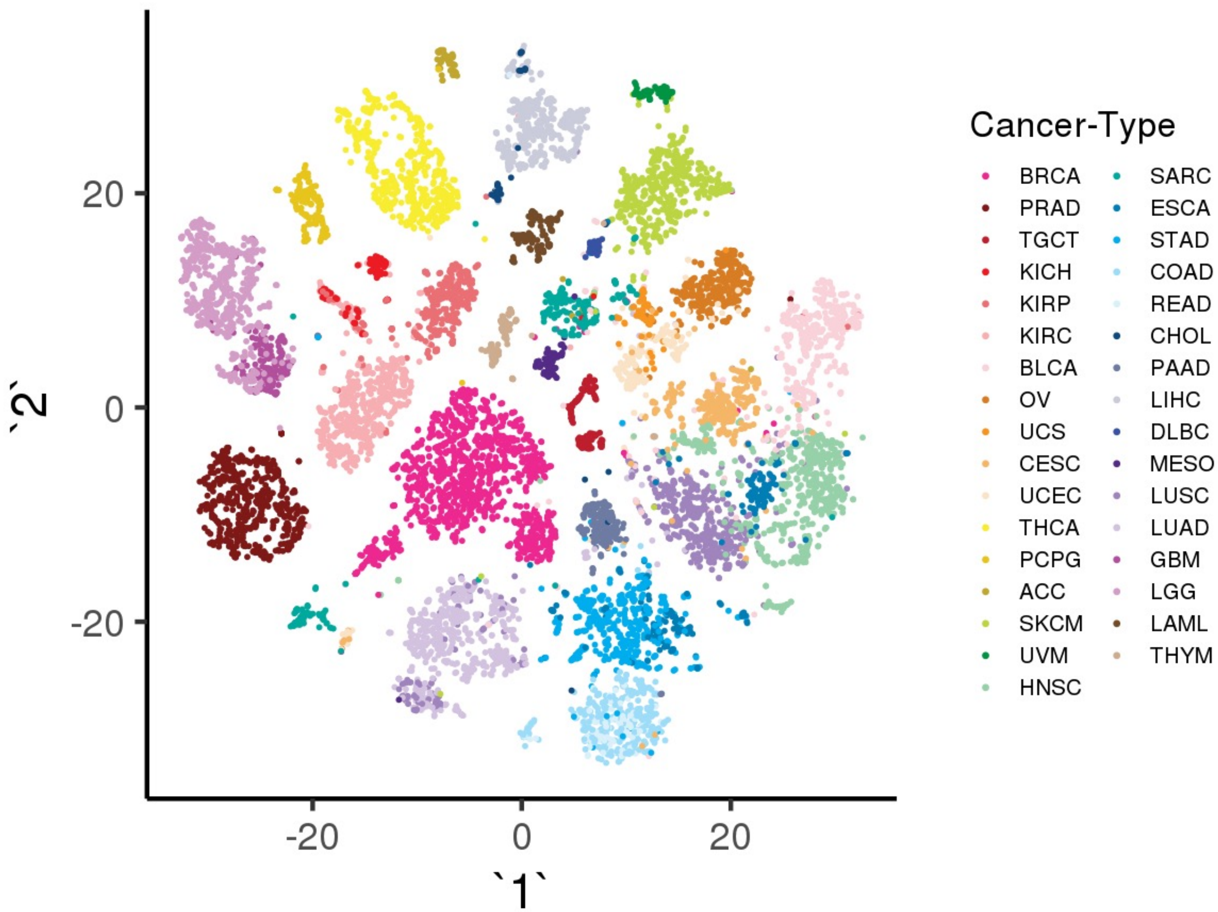


- Widely used for time-series modeling
- The parameterization of the functions that control how h behaves dictate the type of recurrent neural networks:
  - Long short-term memory (LSTM)
  - Gated recurrent units (GRU)

Latent factor models

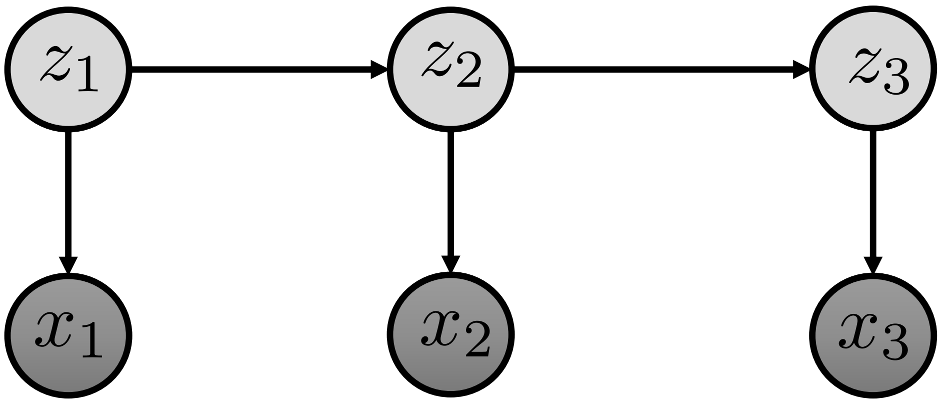


- Unsupervised models of (often high-dimensional data)
- Z: unobserved latent variation (often lower dimensional) than X (observed data)
- You may have encountered many variations of latent factor models:
  - Linear models:
    - Probabilistic PCA
    - Factor analysis
  - Non-linear models
    - Variational autoencoders



Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders
Way et. al , PSB 2014

State space models



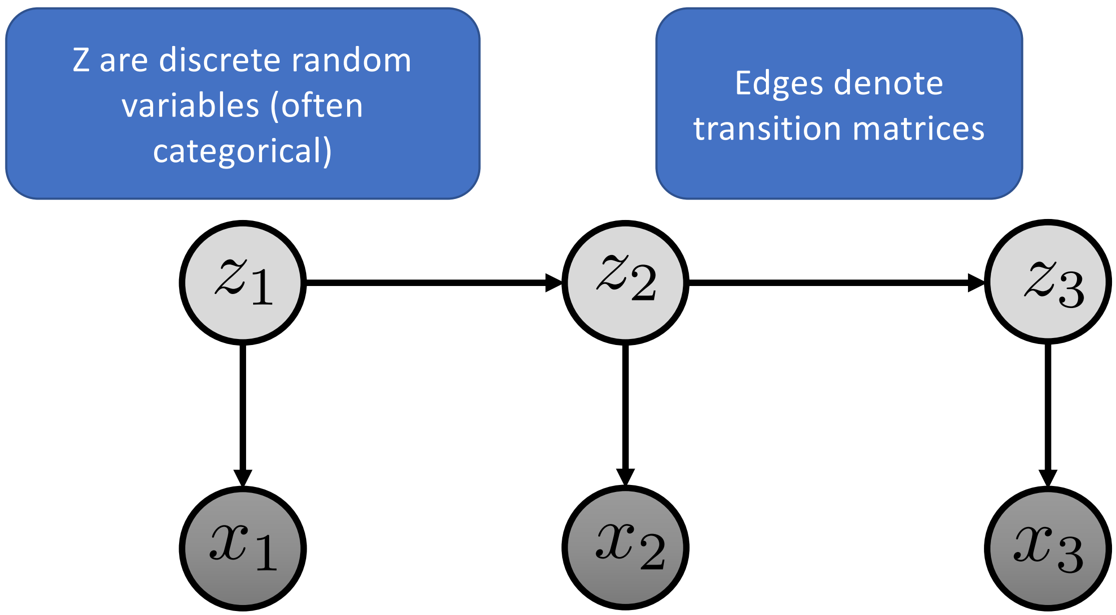
There are many different varieties of state space models.

Each one makes different assumptions on how the probabilities behave and are transformed.

$$p(x_1, x_2, x_3) = \int_{z_1, z_2, z_3} p(x_1, x_2, x_3, z_1, z_2, z_3)$$

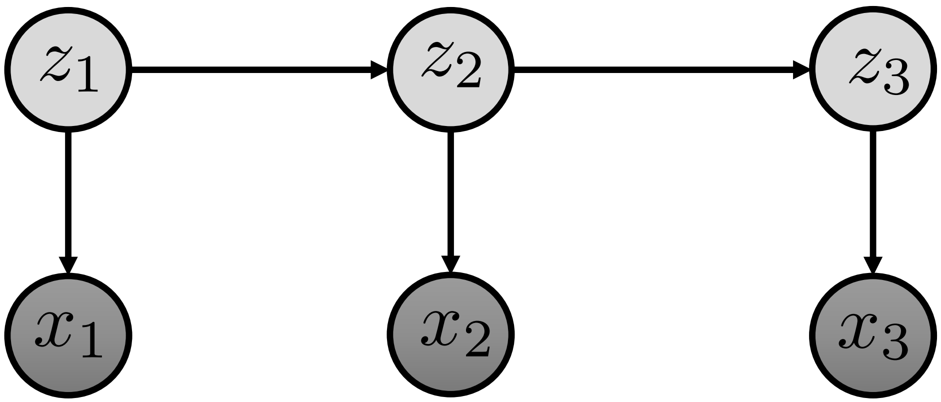
$$= \int_{z_1, z_2, z_3} p(z_1)p(z_2|z_1)p(z_3|z_2) \prod_{k=1}^{3} p(x_k|z_k)$$

Hidden Markov Model



Linear Gaussian State Space Model

Z are continuous valued random variables (Gaussian)

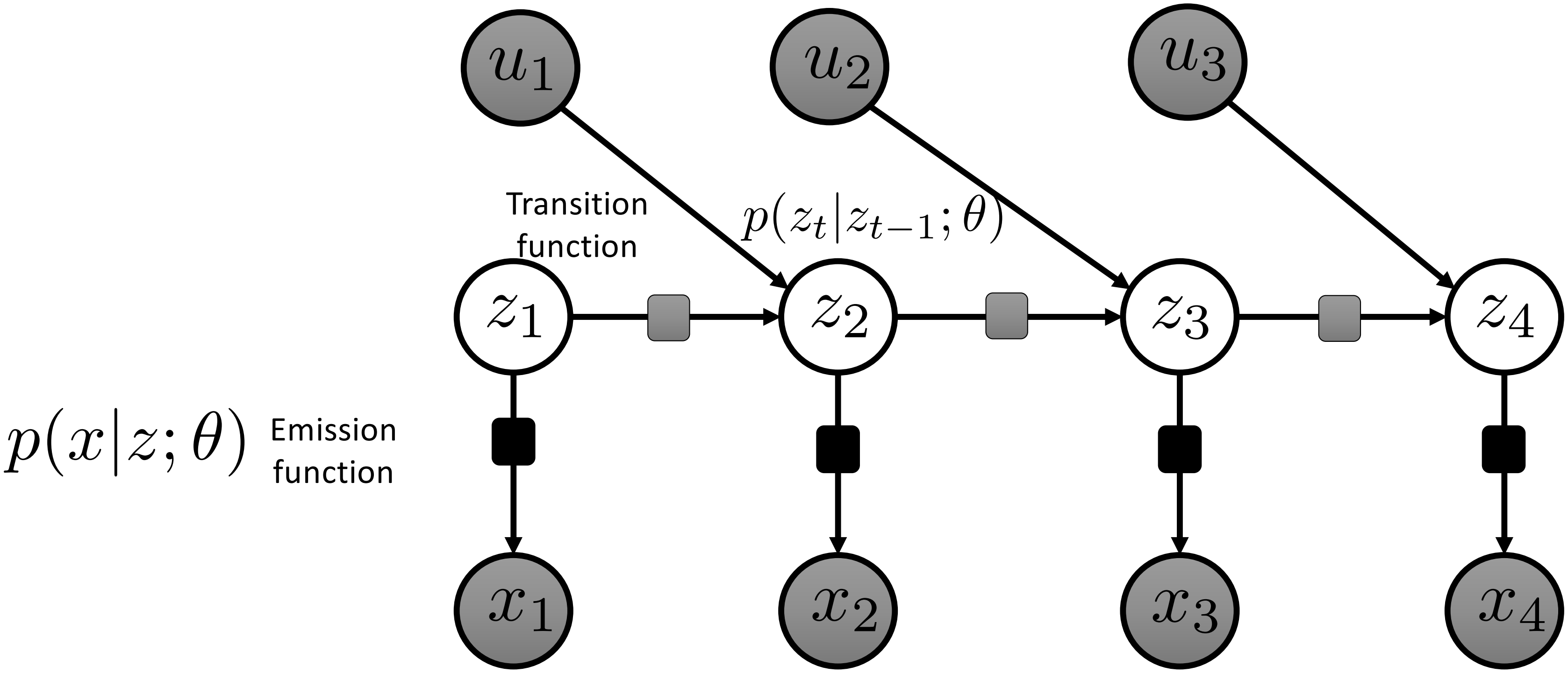


$$z_t = \mathcal{N}(\mu_t, \sigma)$$

$$\mu_t = W z_{t-1} + b$$

$$\sigma = C$$

Deep Markov Models



Structured Inference Networks for Nonlinear State Space Models, **RGK,** US, DS, AAAI 2017

Learning time-series models

Univariate time-series

Regression

ARIMA

Nonlinear regression via conv. nets

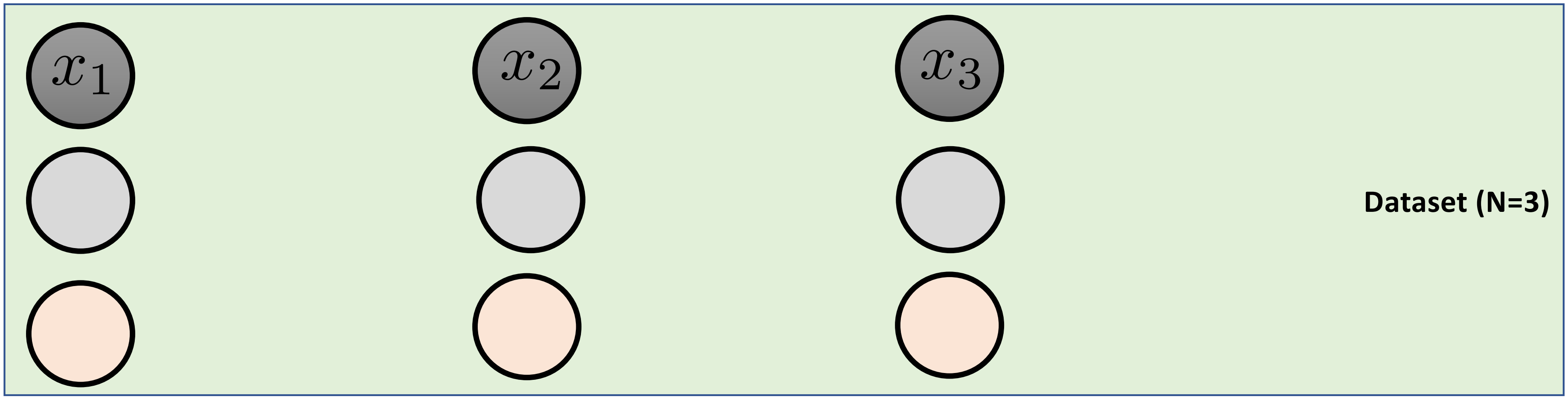
Multivariate time series

K-order Markov models

Recurrent neural networks

State space models

Learning via maximum likelihood estimation



- Model parameters are learned via **maximum likelihood estimation**

$$\mathcal{L}(x_1, \ldots, x_T; \theta) = \log p(x_1, \ldots, x_T; \theta)$$

Score function
(high is good, low is bad)

$$\theta = \arg\max_{\theta} \sum_{i=1}^{N} \mathcal{L}(x_1^i, \ldots, x_T^i; \theta)$$

Solve this optimization problem to **learn** the model. Often formulated as a minimization of the negative of the log-likelihood function

Recipes for learning via maximum likelihood estimation

- Usually:
  - Write down the log likelihood as a function of the model parameters
  - Use stochastic gradient ascent to maximize log likelihood of observed data to learn parameters
- For latent variable models:
  - If the posterior distribution is tractable, often can write the log-likelihood in closed form or obtain an unbiased estimate via Monte-Carlo sampling
  - Else: approximate inference
    - Variational inference
    - Markov Chain Monte Carlo

Evaluation of time-series models

- Mean-squared error
  - Forecasting on training data
  - Forecasting on held-out data
- Held-out log likelihood
- Introspection of model parameters