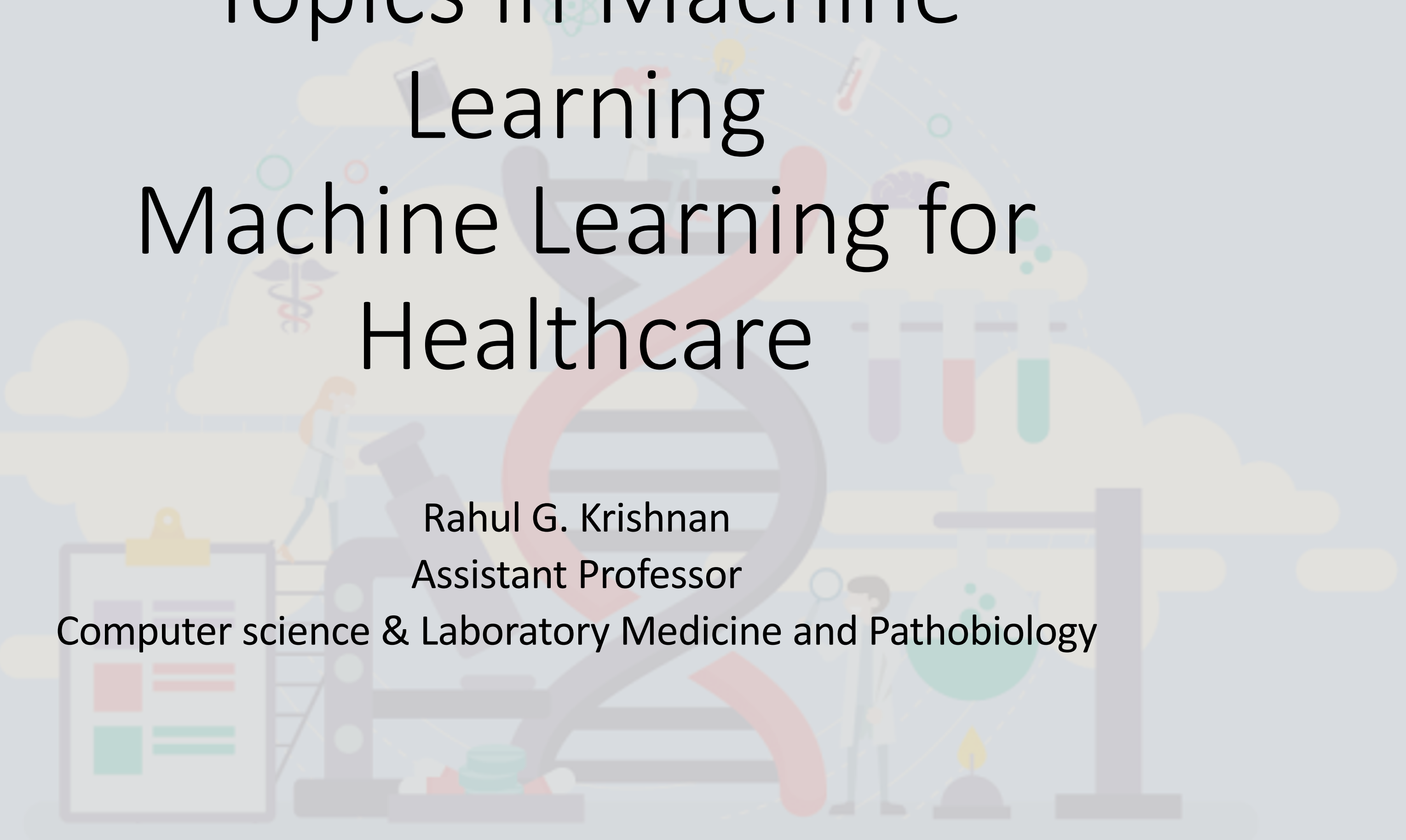
UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Topics in Machine Learning
Machine Learning for Healthcare



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

Outline

- Announcements
- Recap of risk stratification in pictures
- Health is a multi-scale problem
- Time series modeling
  - Where does time come into the picture?
  - Time-series data in the ICU
  - Time-series data in for chronic disease care and management
- What can machine learning do?
- Next week: Technical dive into time-series models

Announcements

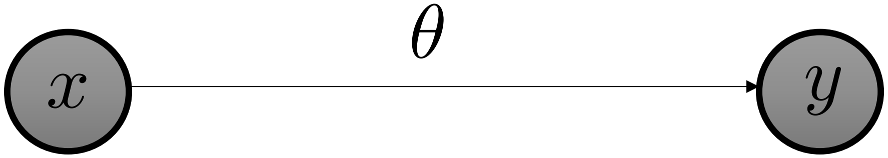
- In two weeks, project proposals [10% of grade] are due:
  - Start forming teams and brainstorming project ideas
  - Create an outline for your project proposal
  - Link: csc2541hf-2021.github.io/assignments/projectproposal
- In two weeks, we will start having student presentations [15% of grade]
  - Present in pairs, first to TAs, then to the class
  - Link should be active and contain details on where to sign up
  - First come first serve for papers
  - csc2541hf-2021.github.io/assignments/paperpresentation

Disclaimer

- None of the material present in this course is intended as medical advice
- We are a long way away from these learning models being implemented in hospitals and routine clinical care

Risk stratification

- Example of supervised learning in healthcare
- Use clinical data to define covariates x
- Use data + domain knowledge to define y
- Build model to approximate risk



$p(y|x)$

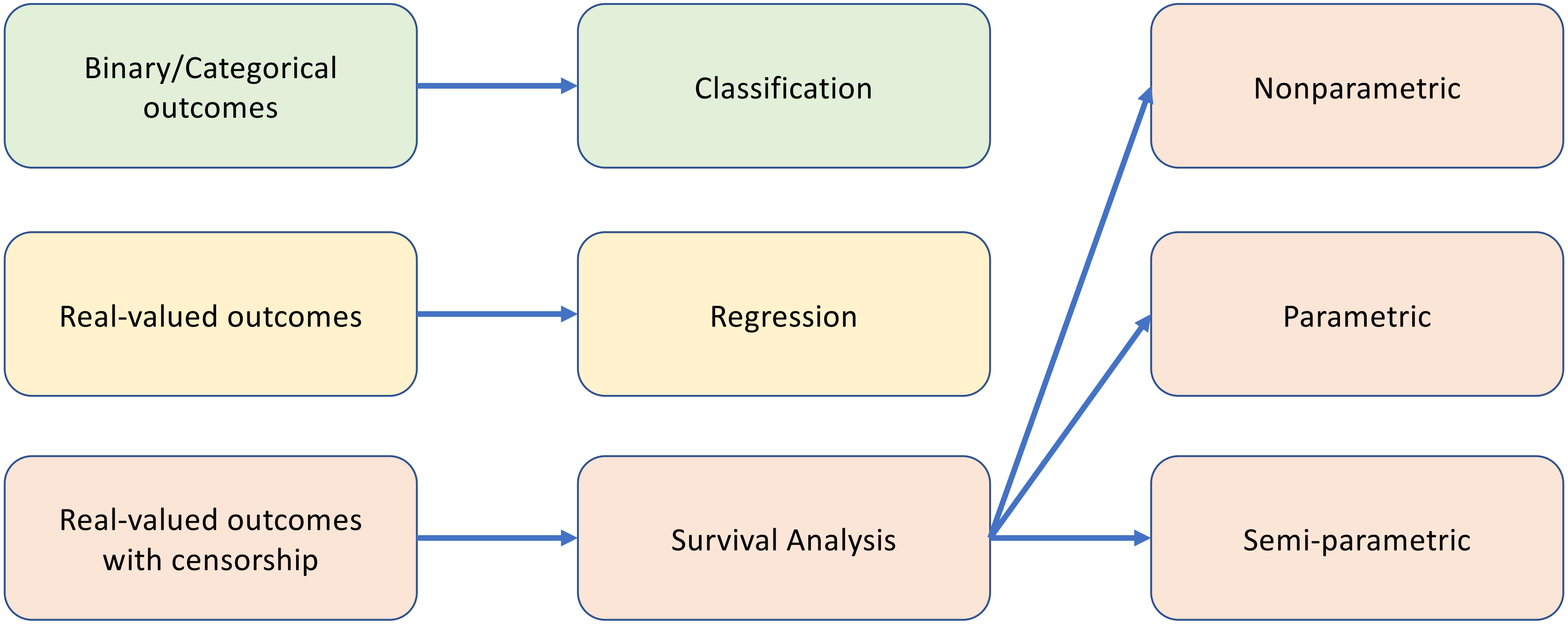
Patient features

Time to readmission

Chest X-ray image

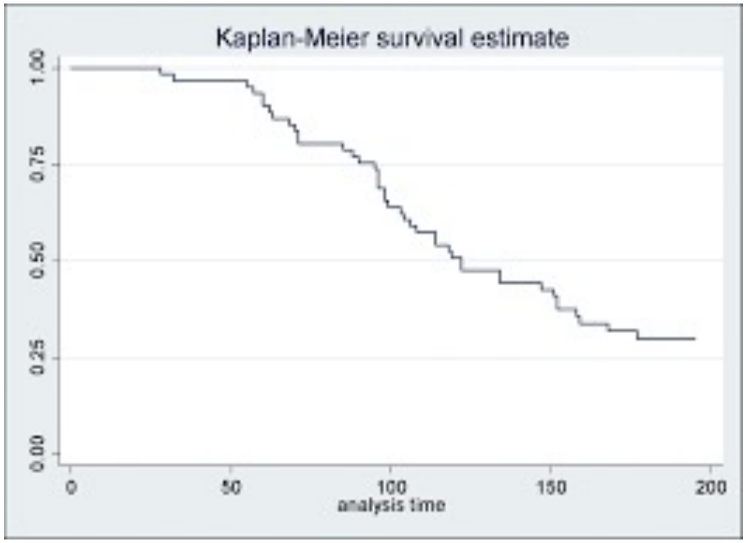
pneumonia/pneumothorax

Outcomes in risk stratification



Kaplan Meier estimator

- Derivation out of scope for this class
  - Survival analysis is a rich area of research and is often a course in and of itself
  - E.g. Lu Tian and Richard Olshen at Stanford



Observed event times

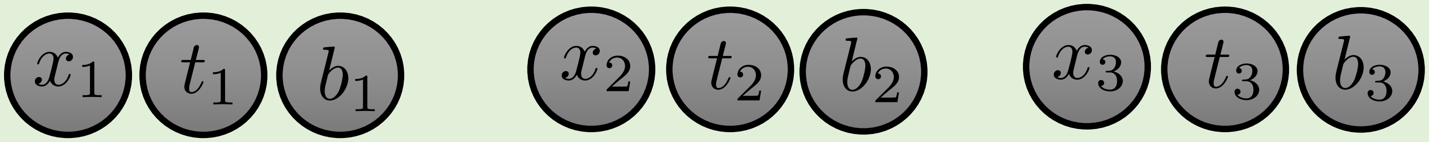
$$y_{(1)} < y_{(2)} < \cdots < y_{(D)}$$

$d_{(k)}$ = # events at this time

$n_{(k)}$ = # of individuals alive and uncensored

$$\widehat{S}_{K-M}(t) = \prod_{k:y_{(k)} \leq t} \left\{ 1 - \frac{d_{(k)}}{n_{(k)}} \right\}$$

Parametric survival analysis



Dataset (N=3)

- Given a dataset, the model parameters are learned via **maximum likelihood estimation**

$$p(T = t | x; \theta) = f(t)$$

Uncensored likelihood

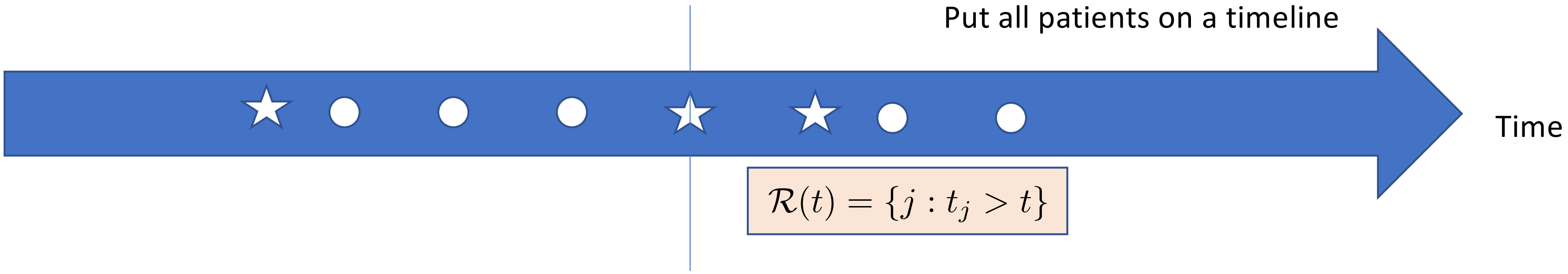
$$p(T > t | x; \theta) = S(t)$$

Censored likelihood

$$\sum_{i=1}^{N} b_i \log p(T = t_i | x_i; \theta) + (1 - b_i) \log p(T > t_i | x_i; \theta)$$

Maximize the following objective function to learn model parameters

Learning CoxPH with the partial likelihood



$$\mathcal{R}(t) = \{j : t_j > t\}$$

$$\mathcal{L}(\beta) = \sum_{i=1}^{K} b_i \log \frac{\exp(\beta^T X_i)}{\sum_{l \in \mathcal{R}(t_i)} \exp(\beta^T X_l)}$$

○ Censored event

☆ Event

Intuition: How likely are the features of this patient to explain their elevated risk of having the event occur now compared to all the individuals whose event occurs later!

Advances in machine learning for survival analysis

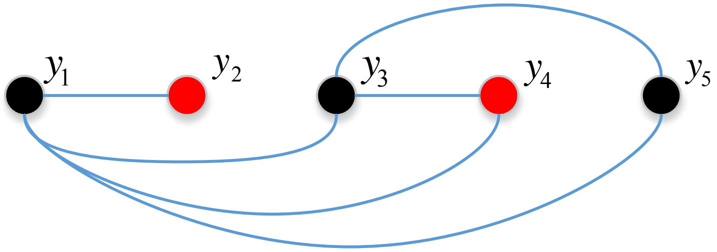
- DeepSurv, Katzman et. al, 2017
    - One of the readings for this week uses a deep neural network to parameterize the modification to the hazard function
    - Parameter estimation by taking derivatives of the
- Advanced reading Deep survival analysis, Ranganath et. al, 2016
    - What if x is very high dimensional?
    - Rather than condition on x directly, learn a latent representation of x while jointly modeling survival time

Evaluation in survival analysis

- Concordance index (aka C-statistic) – predicts how well the model ranks patients based on survival (i.e. predicts relative survival time)
- Equivalent to AUC (when there is no censoring)

$$\hat{c} = \frac{1}{num} \sum_{i:b_i = 0} \sum_{j:y_i < y_j} I[S(\hat{y}_j|X_j) > S(\hat{y}_i|X_i)]$$

Black = uncensored
Red = censored



Other ways to evaluate models

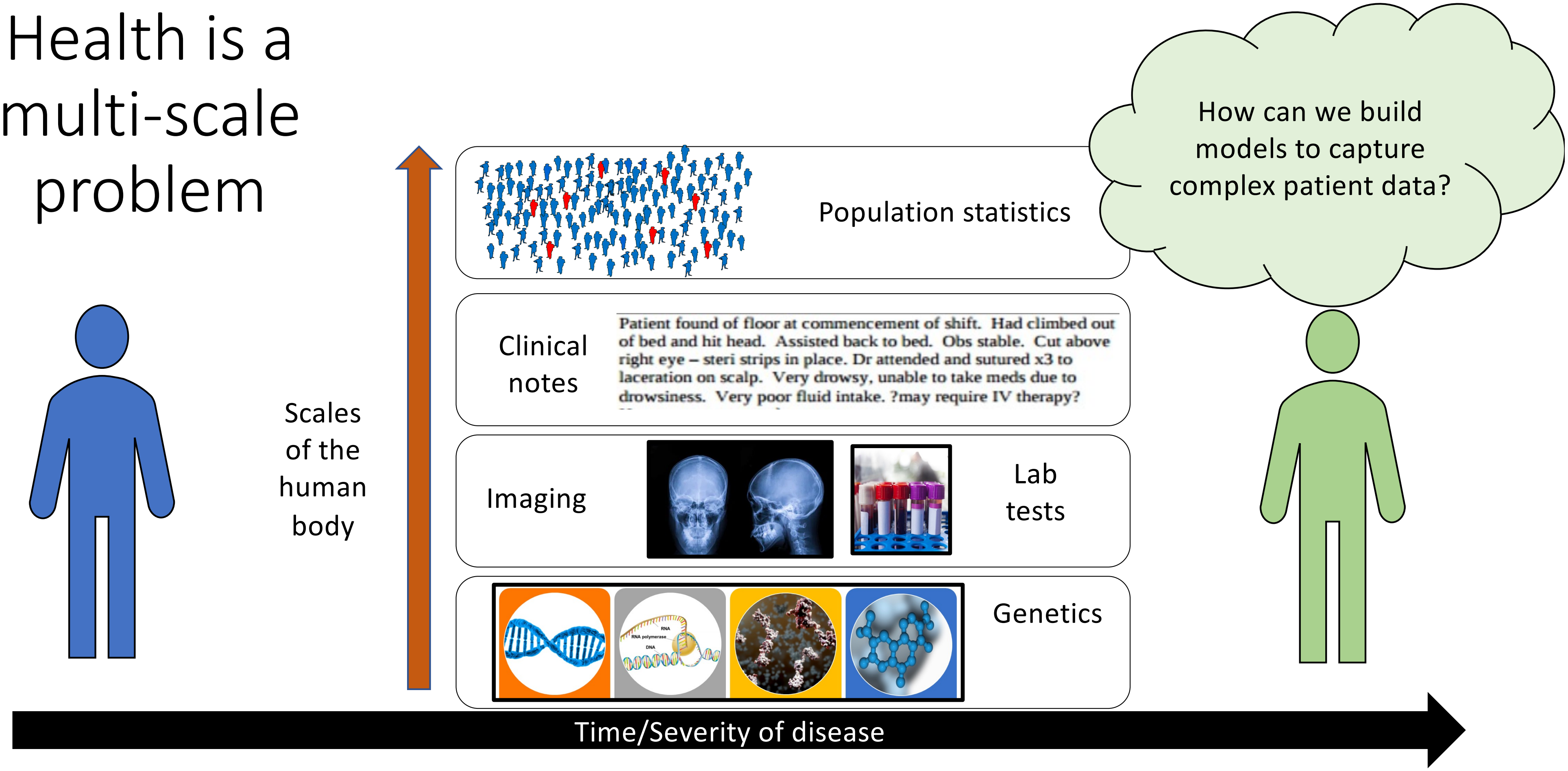
- Mean squared error [for just those who are uncensored]
- Held out likelihood (censored + uncensored)

Application to risk stratification for diabetes from electronic health records

- Predict the onset of diabetes from patient covariates
- Use to create personalized interventions for patients
  - Request they see care provider for personalized treatment plan
  - Modifications to diet with a smartphone app
- Advantages (wrt traditional risk stratification models):
  - Scalable (not limited to point of care)
- Limitations:
  - Covariate shift across time
  - Covariate shift across populations
  - Hidden confounding

Questions?

Health is a multi-scale problem



Scales of the human body

Population statistics

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

How can we build models to capture complex patient data?

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

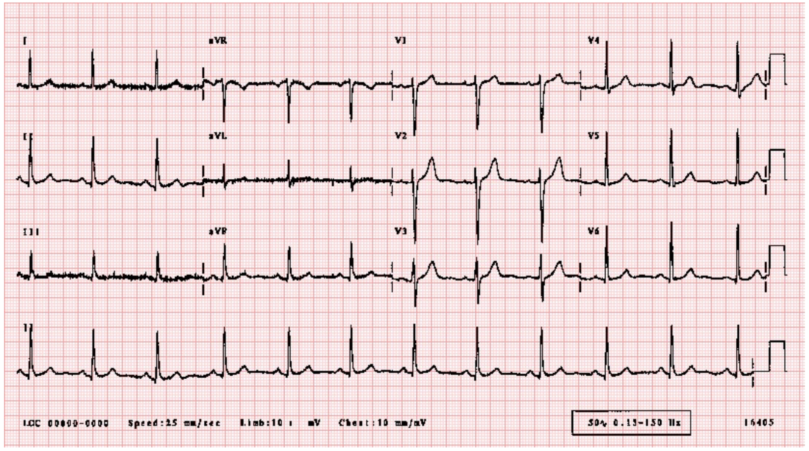
Patients in critical care units

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



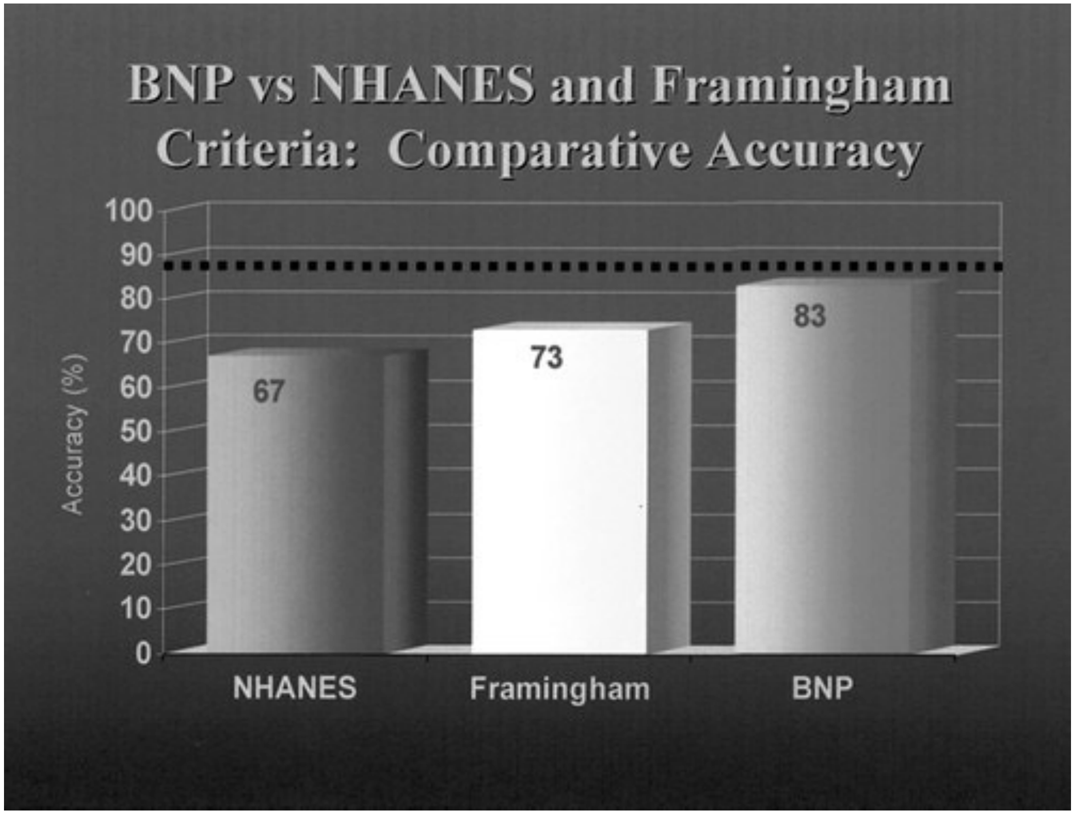
Physiological time-series data 2 [cardiology]

- An ECG is often (but not always) abnormal in patients w heart failure
- B-type natriuretic peptide (BNP) belongs to a family of protein hormones called *natriuretic peptides*.
  - Regulate circulation of the blood causing them to dilate
  - Also have an effect on the kidneys
  - Play a role in reducing the heart's workload
  - Helps the body handle congestive heart failure
- Normal BNP is useful to **rule out** hypotheses
- High BNP can be informative but not always conclusive
- Type of data: discrete time

Source: https://www.health.harvard.edu/newsletter_article/bnp-an-important-new-cardiac-test

Comparison to risk scores

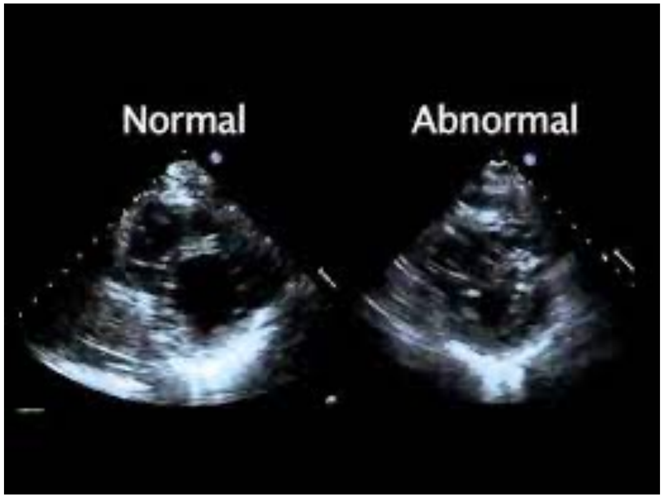
“single BNP level was more accurate than both the National Health and Nutrition Examination Score and Framingham, arguably the two criteria most commonly used to diagnose CHF”



Source: https://www.ahajournals.org/doi/10.1161/01.CIR.0000019121.91548.C2

Physiological time-series data 3 [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]

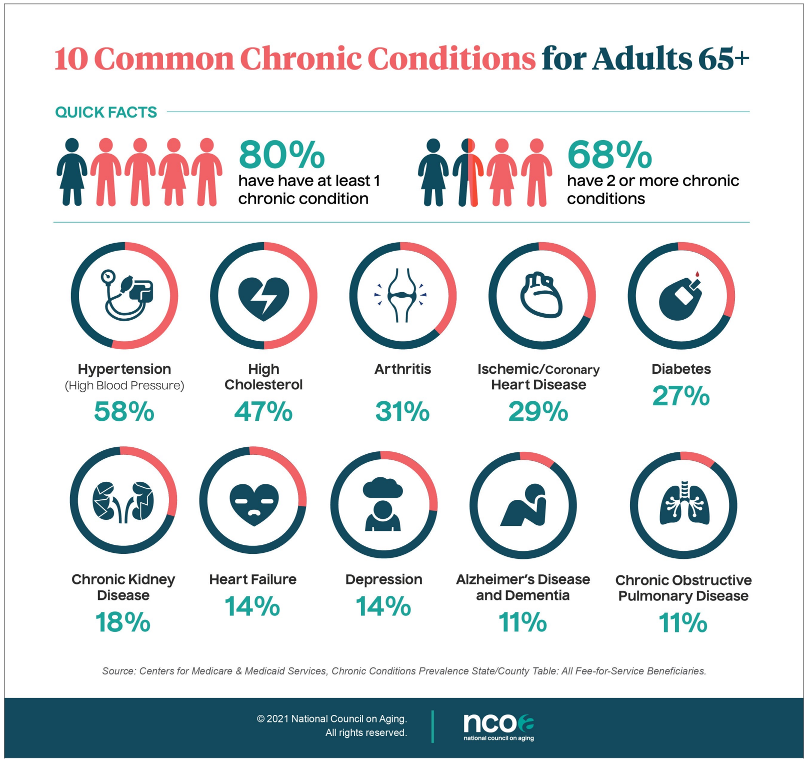


Source: **Role of Echocardiography in the Intensive Care Unit: Overview of the Most Common Clinical Scenarios** ,Longobardo et. al, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6664324/

Questions?

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.



10 Common Chronic Conditions for Adults 65+

QUICK FACTS

80% have have at least 1 chronic condition

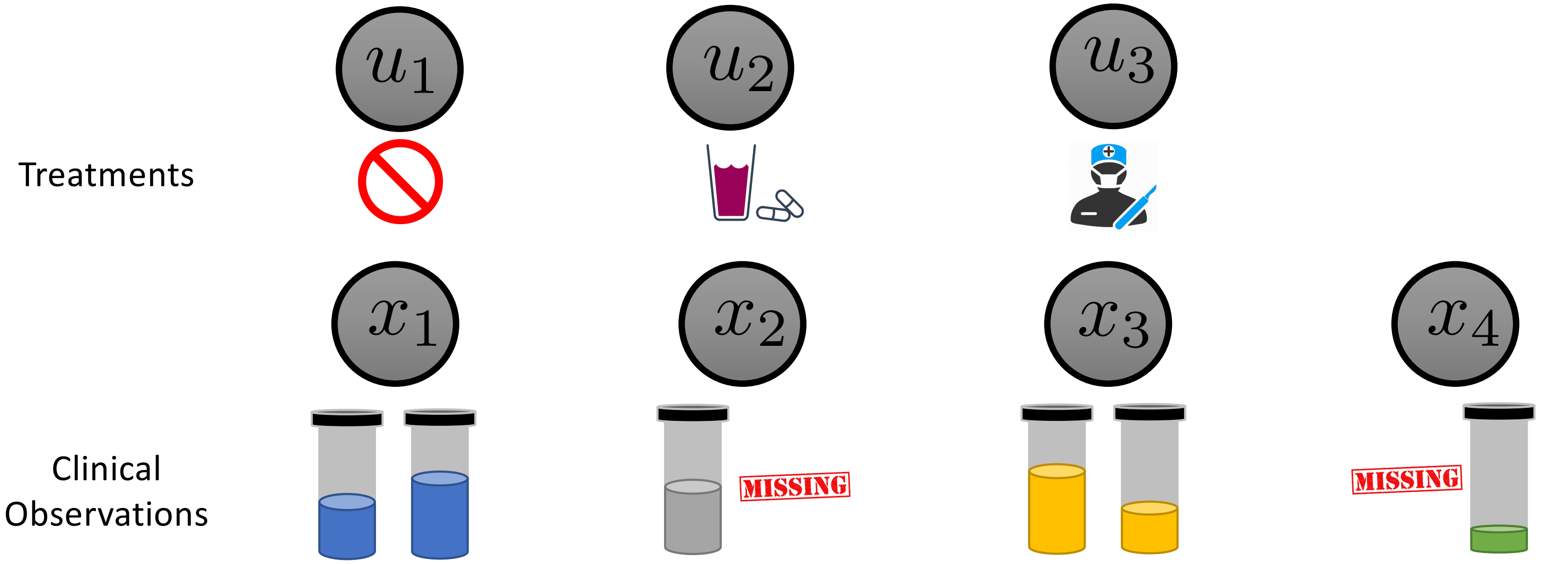
68% have 2 or more chronic conditions

Hypertension (High Blood Pressure)	58%
High Cholesterol	47%
Arthritis	31%
Ischemic/Coronary Heart Disease	29%
Diabetes	27%
Chronic Kidney Disease	18%
Heart Failure	14%
Depression	14%
Alzheimer's Disease and Dementia	11%
Chronic Obstructive Pulmonary Disease	11%

Source: Centers for Medicare & Medicaid Services, Chronic Conditions Prevalence State/County Table: All Fee-for-Service Beneficiaries.

ncoa national council on aging

Chronic Disease Management – (1)

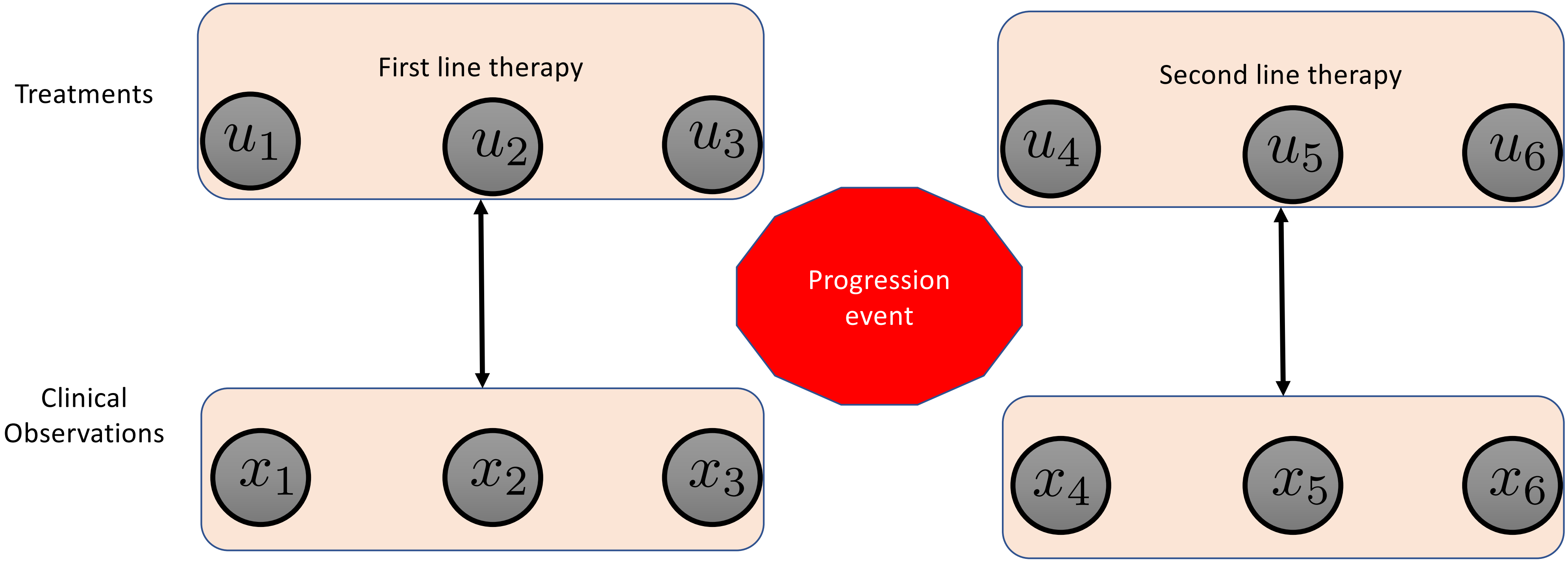


- Canonical picture that characterizes how healthcare data behave
- Interesting and useful structure in how chronic diseases are treated

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

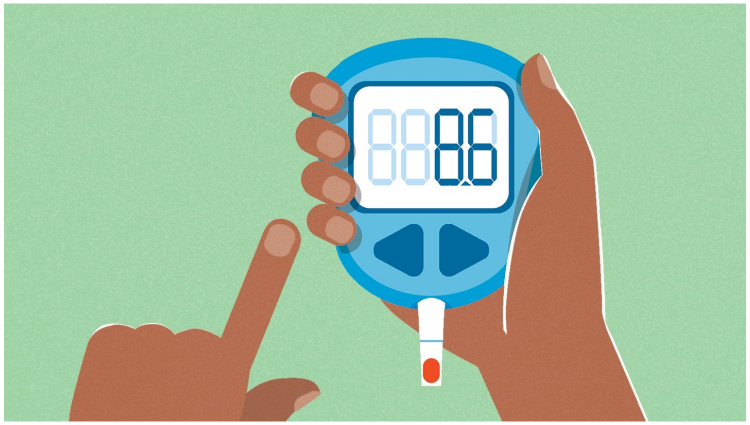


Progression events

- 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

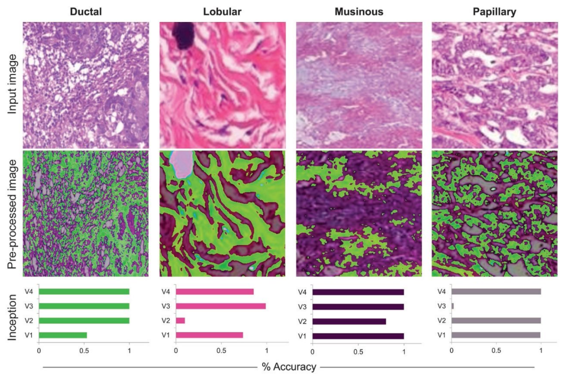
- Biomarkers:
  - Blood sugar (A1C) levels
- Interventions
  - First line: Metformin
  - Second line:
    - Combination therapy: Metformin + Sulfonylurea drug



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

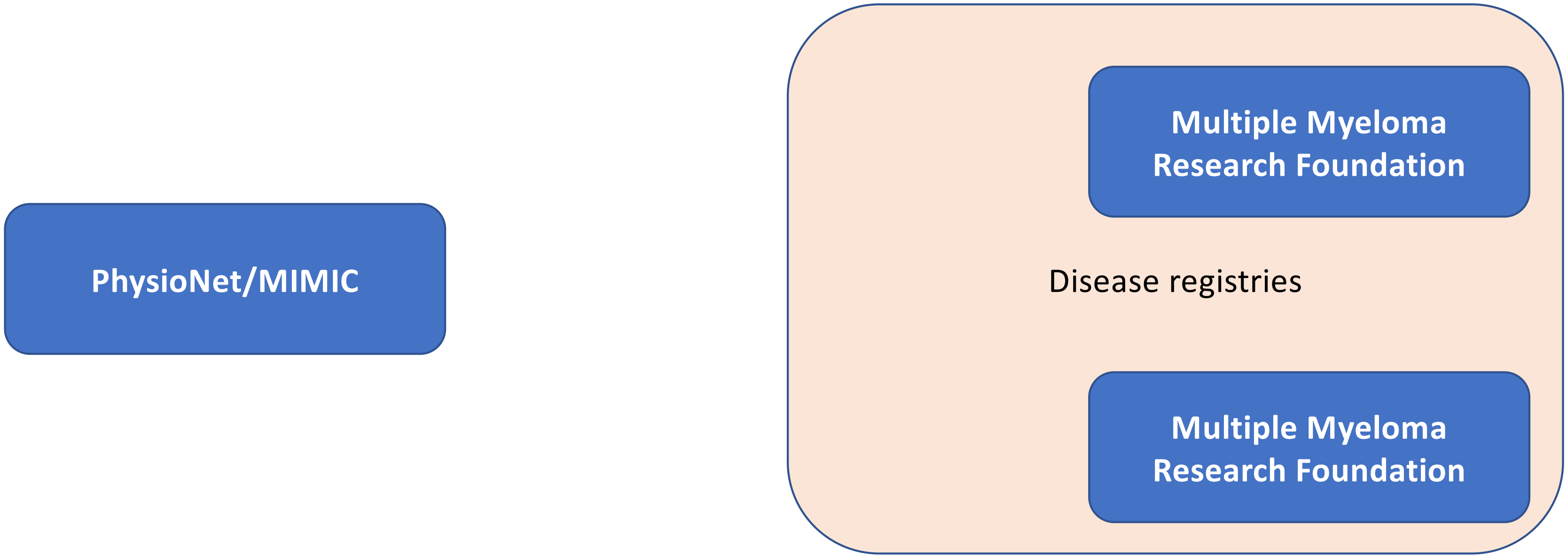
What does this mean for data?

- 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



Questions?

Time-series datasets



PhysioNet/MIMIC

Multiple Myeloma Research Foundation

Disease registries

Multiple Myeloma Research Foundation

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

Discussion on data

- Start time of data = start time of disease
  - Is this correct?
- End time of data = end time of disease
  - Is this correct?

Keep an eye out for left and right censorship!

Questions?

Wednesday: deep dive into technical topics for time-series analysis

Friday: open tutorial session for project brainstorming (we’ll split up the classroom into three groups)