Introduction to Bioinformatics

Topic 17
Motivating Example

Fall, 2023

Professor
Gregory R. Grant

Teaching Assistants
Chetan Vadali

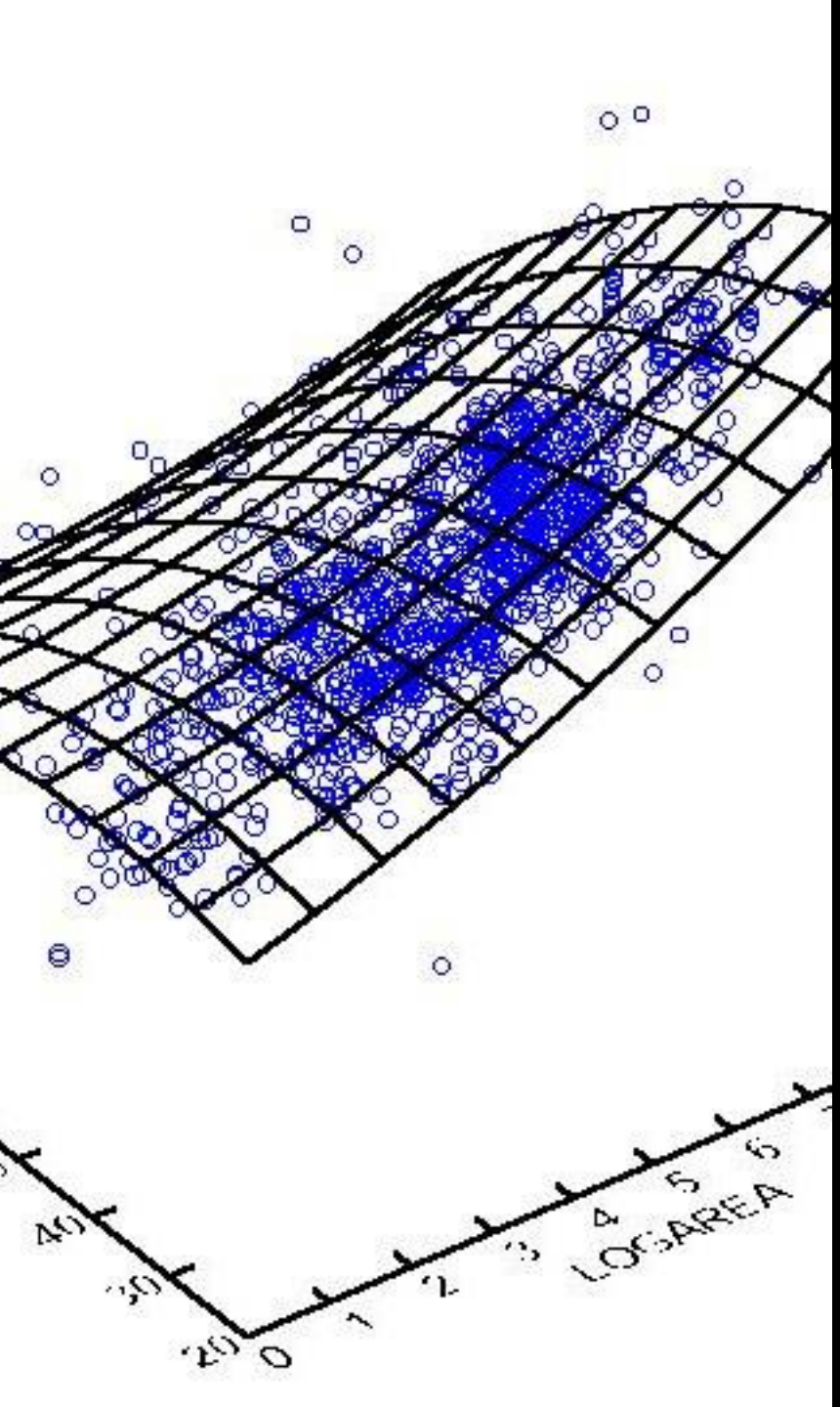
Gregory R. Grant
Genetics Department
ggrant@pennmedicine.upenn.edu

ITMAT Bioinformatics Laboratory
University of Pennsylvania

Motivating Example

Consider the following problem:

- Given a tissue sample, can you determine from its gene expression what time it was taken?
- If we have enough data where we know the truth, can we build a machine that does this with any sort of accuracy?



Multiple Regression

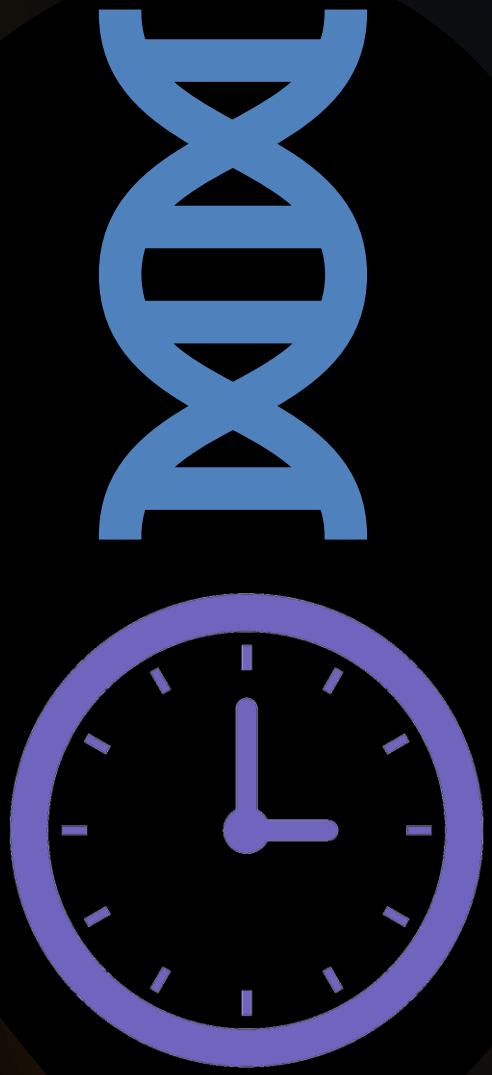
- On its face this is no different from the regression problems we're familiar with.
- You have some independent variables like BMI, average number hours sleep per night, average number of steps taken per day, average number of drinks per week.
- And some dependent variable, like life span.
- How well can we predict the dependent variable from the independent variables?

Model Building

- Make some assumptions about the form of the model
  - E.g. if there is one independent variable, that means the shape of the regression *curve*, linear, parabolic, sin wave, etc.
  - Or the shape of the regression *surface* if there are two variables.
  - Or the shape of a *hypersurface* in higher dimensions.

- Then we collect a bunch of data where the truth is known and use it to estimate the parameters of the model.
  - That's the "learning" part.

Gene Expression



- Trying to infer time from gene expression data fits right into this framework.
- Every gene is an independent variable. There are just a huge number of them.

$$X_1, X_2, \ldots, X_{30,000}$$

  - There's one dependent variable: T=Time of Day.
  - Consider the simplest possible model, linear in the gene expression values (the X_i's):

$$T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_{30,000} X_{30,000} + \varepsilon$$

Gene Expression

- We estimate the betas from data.

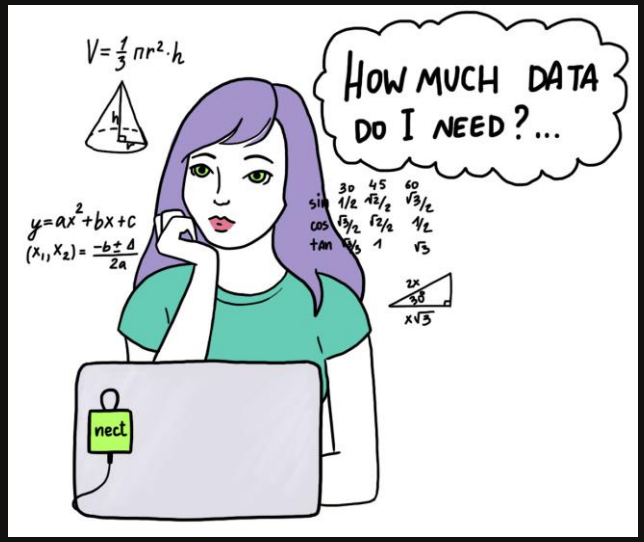
$$T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_{30{,}000} X_{30{,}000}$$

  - That's the "learning" part.
- Then from a new sample (new set of values for the X_i's) we plug in and get a value of T.
  - That's the machine.

- Sounds simple enough. What could go wrong?

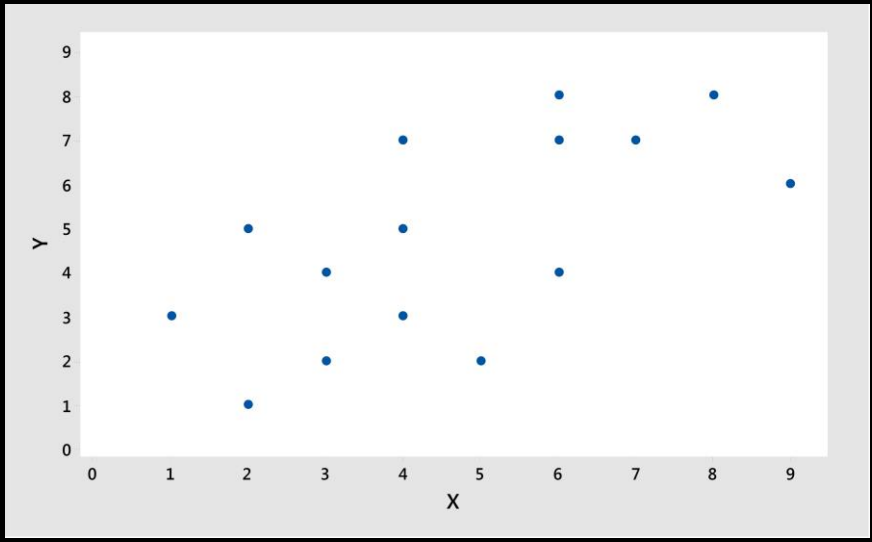
First Question

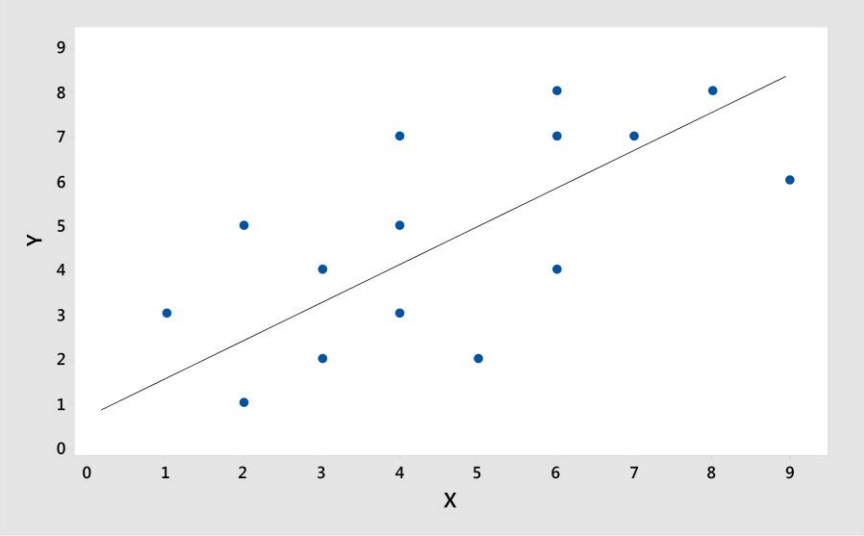
- How much data do we need to train a model with *N* independent variables?
- Does it depend on the number of variables?



Case *N*=1

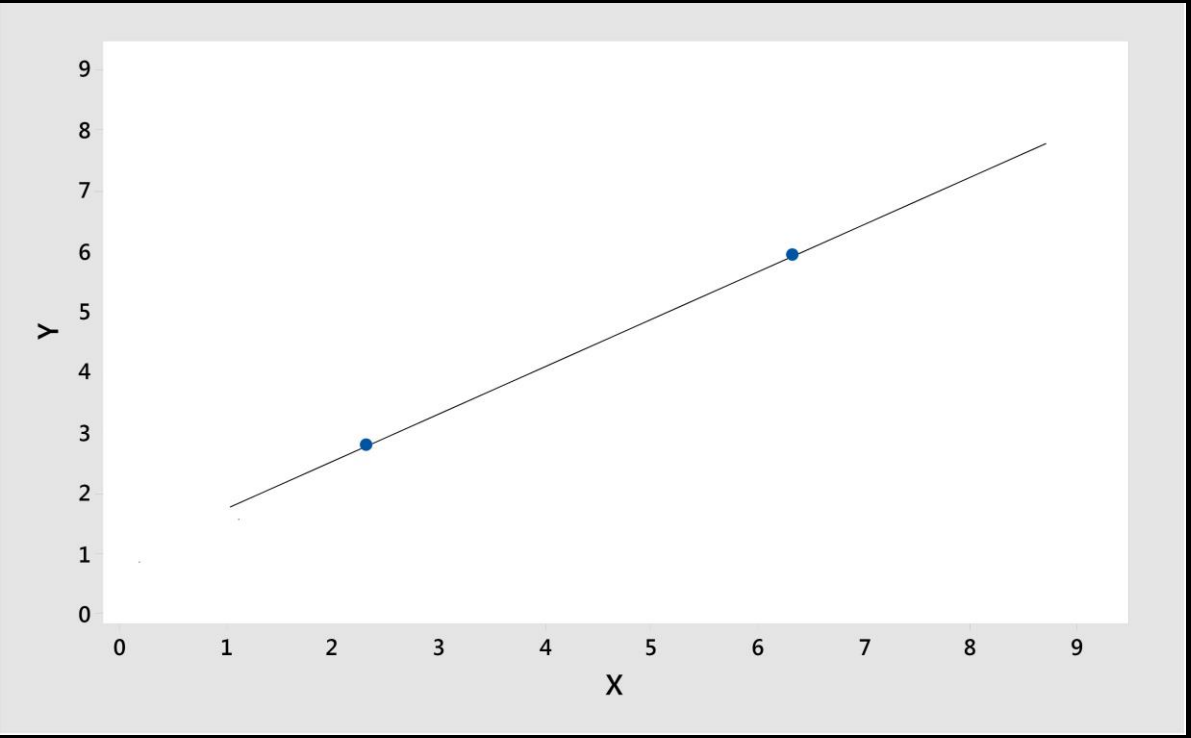
- If we have 1 independent variable, how much data do we need to train the model?
  - Remember, training the model means estimating the beta coefficients from data.
- If there is one independent variable the problem is one of fitting a straight line.





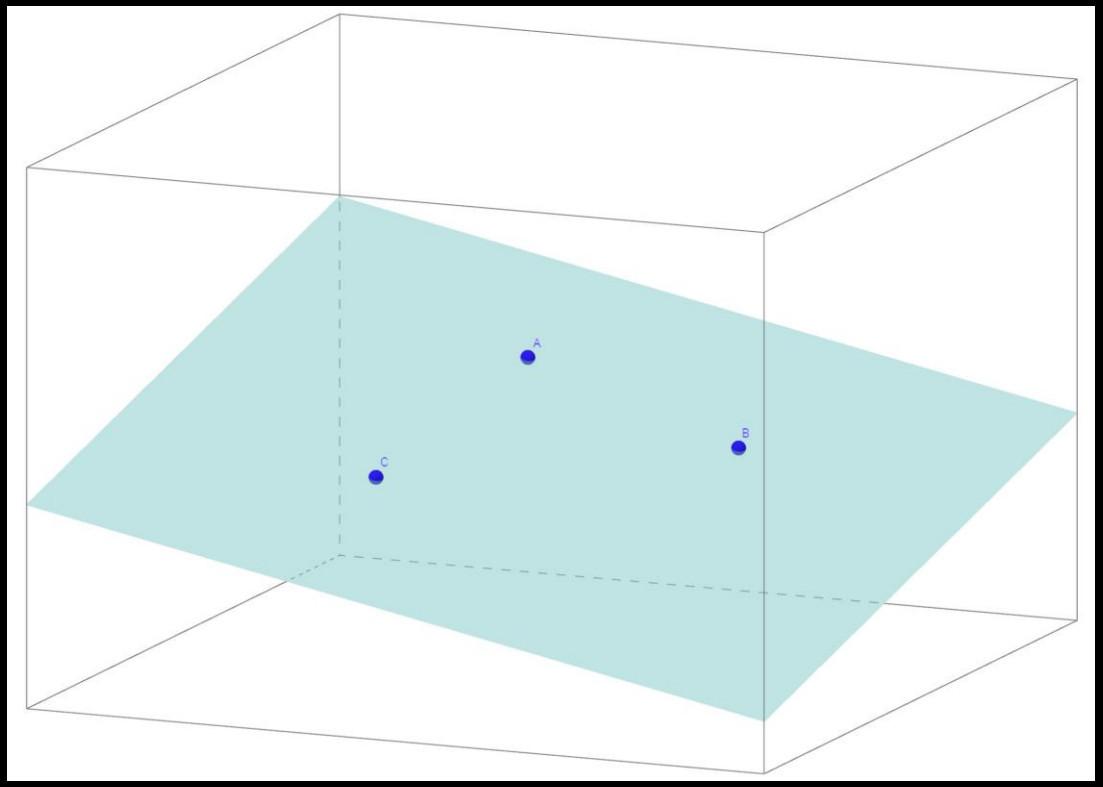
Case *N*=1

- What's the absolute minimum number of data points we'd need to fit a straight line?
- Clearly there'd have to be at least two, and they'd have to have two different *X* values.



Case *N*=2

- If we have 2 independent variable, how much data do we need to train the model?
- Now we're fitting a plane to points in three space.
- This cannot be done with less than three points.



In General

- In general, you cannot hope to train a model with *N* variables without at least *N*+1 subjects.
- Therefore, if we want to have any hope to train the model with gene expression data, we'd need data from at least 30,001 individuals.

$$T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_{30,000} X_{30,000}$$

- But that's unrealistic.
  - They may be doing that one day, but we'll be lucky to recruit and afford even 100 subjects today.
- So, what do we do?

Multiple Solutions

Several ways have been developed to reduce the number of independent variables. We'll look at two:

Use expert knowledge to focus attention on a small number of genes.	Dimensionality Reduction to reduce all genes down to a smaller set of *latent* variables.

There are several more we won't have time to talk about:

Regularization.	Factor Analysis.	Backwards Feature Elimination.	Decision Trees	Etc.

Expert Knowledge

- If there's prior expert knowledge available, it can be used to focus attention on a small number of genes.
- For example, one could just use genes known to be involved in the circadian clock.
- That's what this group did:

Research paper

Evaluation of mRNA markers for estimating blood deposition time: Towards alibi testing from human forensic stains with rhythmic biomarkers

Karolina Lech[a], Fan Liu[a], Katrin Ackermann[a,b], Victoria L. Revell[c], Oscar Lao[a,d], Debra J. Skene[c], Manfred Kayser[a,*]

Small Sample Size

- Their data consisted of only 12 individuals, sampled across time.

However, for the present analysis we used 18 two-hourly blood samples per each of 12 male participants (mean age ± SD = 23 ± 5 years) i.e. 216 samples in total. These samples spanned the first 36 h of the S/SD study (from 12:00 h Day 2 to 22:00 h Day 3), excluding the sleep deprivation condition (00:00 h Day 3 to 12:00 h Day 4).

Relied on Prior Expert Knowledge

- With 12 individuals they could not train the model on all genes.
- But they can get away with a more than 11 variables, because this is time-series data.
  - Each individual resulted in multiple observations from different times.

However, for the present analysis we used 18 two-hourly blood samples per each of 12 male participants (mean age ± SD = 23 ± 5 years) i.e. 216 samples in total. These samples spanned the first 36 h of the S/SD study (from 12:00 h Day 2 to 22:00 h Day 3), excluding the sleep deprivation condition (00:00 h Day 3 to 12:00 h Day 4).

Relied on Prior Expert Knowledge

- They reduced the set of independent variables to 21 Clock and Clock controlled genes, which were determined from a previous study.

Recently, we analysed the expression of 12 well-known clock and clock-related genes [39] and of 9 candidate clock-controlled genes [40] and measured the concentration of melatonin and cortisol, in blood samples drawn from 12 individuals around the clock at 2 h intervals for 48 h under controlled conditions of a sleep/sleep deprivation (S/SD) study protocol, and under a separate constant routine study protocol (CR) [39–41]. Data analysis in these previous biologically-motivated studies focused on identifying diurnal and circadian genes, understanding their biological function, and assessing the influence of sleep and no-sleep on gene expression.

The Model

- "Multinomial" means there's more than one predictor (independent) variable.
- "Logistic" regression just means the dependent variable is a categorical rather than a numerical variable.

2.2. Model building and time predictions

Prediction models were constructed based on multinomial logistic regression, where the *ACTB*-normalised expression levels of the genes and the concentration values of the hormones were considered as the predictors, and the multinomial time categories as the response variable,

- This group started with a simpler problem dividing the day into just three time-categories, so the dependent variable just has three values, 0, 1 or 2.

1. morning/noon
2. afternoon/evening
3. night/early morning

Independent Variables

- Expression of the 21 genes were the independent variables.

2.2. Model building and time predictions

Prediction models were constructed based on multinomial logistic regression, where the *ACTB*-normalised expression levels of the genes and the concentration values of the hormones were considered as the predictors, and the multinomial time categories as the response variable,

The Dependent Variable

Categorical variable

1. morning/noon
2. afternoon/evening
3. night/early morning

2.2. Model building and time predictions

Prediction models were constructed based on multinomial logistic regression, where the *ACTB*-normalised expression levels of the genes and the concentration values of the hormones were considered as the predictors, and the multinomial time categories as the response variable,

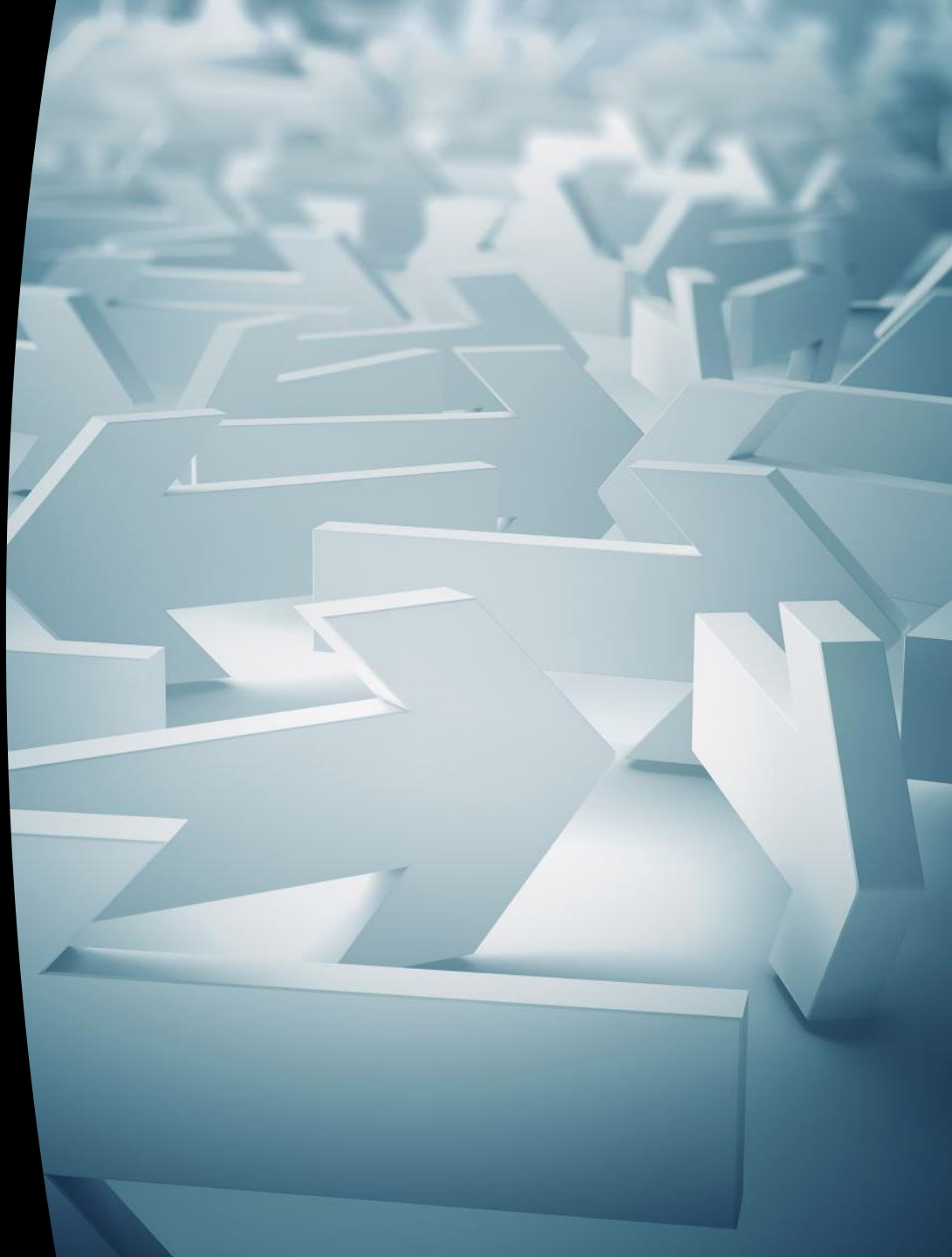
Did it Work?

- Not well enough to put somebody in prison.
- But they claim to have advanced the field with a proof of principle.
- This was 2016, people are still working on the problem.

Our data best support a model that by using these five molecular biomarkers estimates three time categories, i.e. night/early morning, morning/noon, and afternoon/evening with prediction accuracies expressed as AUC values of 0.88, 0.88, and 0.95, respectively. For the first time, we demonstrate the value of mRNA for blood deposition timing and introduce a statistical model for estimating day/night time categories based on molecular biomarkers, which shall be further validated with additional samples in the future. Moreover, our work provides new leads for molecular approaches on time of death estimation using the significantly rhythmic mRNA markers established here.

Approach #2: Dimensionality Reduction

- A second approach is called Principal Components Regression (PCR).
- It starts by doing the dimensionality reduction on the independent variables.
- Specifically, principal components analysis (PCA) on the X_i's.
- It then uses some small number of the first principal components $PC_1, PC_2, \ldots, PC_N$ as latent variables to use as independent variables in place of the large number of X_i's.



Principal Components Regression

- The PCA approach was taken by this group.

CLINICAL MEDICINE

The Journal of Clinical Investigation

High-accuracy determination of internal circadian time from a single blood sample

Nicole Wittenbrink,[1,2] Bharath Ananthasubramaniam,[1,3] Mirjam Münch,[1,4,5] Barbara Koller,[1] Bert Maier,[1] Charlotte Weschke,[1] Frederik Bes,[4,5] Jan de Zeeuw,[5,6] Claudia Nowozin,[4,5] Amely Wahnschaffe,[4,5] Sophia Wisniewski,[4,5] Mandy Zaleska,[6] Osnat Bartok,[7] Reut Ashwal-Fluss,[7] Hedwig Lammert,[8] Hanspeter Herzel,[3] Michael Hummel,[8] Sebastian Kadener,[7,9] Dieter Kunz,[4,5,6] and Achim Kramer[1]

ZeitZeiger

extracts a time-telling gene set by means of supervised sparse principal component analysis.

Machine Learning

We'll revisit this motivating example when we talk about machine learning in more detail.

- Two takeaways:

There are two basic types of machine learning problems

1. Categorical prediction:
   - E.g. Is this a picture of a cat?
2. Numerical prediction:
   - E.g. Predict the lifespan of somebody who smokes *N* cigarette's a day.

Settling on which independent (predictor) variables to use is a big part of the problem.

- This is called "variable selection" and could be a semester topic in its own right.