Introduction, statistics review

Aaron Meyer

Machine learning & data-driven modeling in bioengineering

Lecture:

- Tuesdays/Thursdays, 10–11:50 am
- Boelter Hall 5249

Lab

- Varies by section—check the course schedule.

Lecture Slides

- Lecture slides will be posted on the course website.
- They will be finalized by the night before so you can print them out if you want.
- The slides are *not* everything, but will include space to fill out missing elements during class.

Textbook / Other Course Materials

- There is no textbook for this course.
- I will post related readings prior to each lecture.
- These will either broaden the scope of material covered in class, or provide critical background.
- I will make it clear if the material is more than optional.

Support / Office Hours

Prof. Meyer

- Thursday, 1–1:55 pm in Eng V 4121G
- I will usually also stick around after class and am happy to answer questions

TAs

- By appointment
- Will have regular times setup soon

Other

- Slack channels allow for quick help from both myself and the TAs
- Please use public/shared channels if possible, so others can see the questions

Learning Goals:

By the end of the course you will have an increased understanding of:

1. Critical Thinking and Analysis: Understand the process of identifying critical problems, analyzing current solutions, and determining alternative successful solutions.
2. Engineering Design: Apply mathematical and scientific knowledge to identify, formulate, and solve problems in the chosen design area.
3. Computational Modeling: Apply computational tools to solve and optimize engineering problems.
4. Communicate Effectively: Learn how to give an effective presentation. Understand how to communicate progress orally and in written reports.
5. Manage and Work in Teams: Learn to work and communicate effectively with peers to attain a common goal.

Practical Learning Objectives

By the end of the course you will learn how to:

1. Identify a question that can or cannot be solved by a modeling approach.
2. Determine the prerequisites to applying a modeling method.
3. Implement a number of different modeling methods to answer specific questions.
4. Critically assess modeling results.

Grade Breakdown

30% Final Project
20% Homework Assignments
30% Midterm
10% Class Participation
10% Lab Participation

Labs

- These are mandatory sessions.
- You will have an opportunity to get started on each week's implementation and/or work on your project.

Homework

- These will be a combination of a computational implementation and other problems.
- Each will help reinforce the material and provide hands-on experience by implementing what we learn in class.
- These are meant to challenge you to become comfortable applying the material.
  - Document your effort
  - Get started early
  - Seek answers to your questions in office hours, in lab, or on Slack!

Project

- You will take data from a scientific paper, and implement a machine learning method using best practices.
- A list of papers and data repositories is provided on the website as suggestions.
- Absolutely go find ideas that interest you!
- More details to come.
- First deadline will be in week 6 to pick a project topic.

Exams

- We will have a midterm exam on week 6.
- You will have a final project in lieu of a final exam.

Keys to Success

- Participate in an engaged manner with in-class and take-home activities.
- Turn in assignments on time.
- Work through activities, reading, and problems to ensure your understanding of the material.

If you do these three things, you will do well.

Introduction

How do we need to learn about the world?

- What is a measurement?
- What is a model?

Three things we need to learn about the world

- Measurements (data)
- Models (inference)
- Algorithms

Area of Focus

What we will cover spans a range of fields:

- Engineering (the data)
- Computational techniques (the algorithms)
- Statistics (the model)

Why do we need these things to learn about the world?

FILL IN

Why we need models - Can a biologist fix a radio?

Figure: Lazebnik et al, Cancer Cell, 2002

Why we need models - Can a biologist fix a radio?

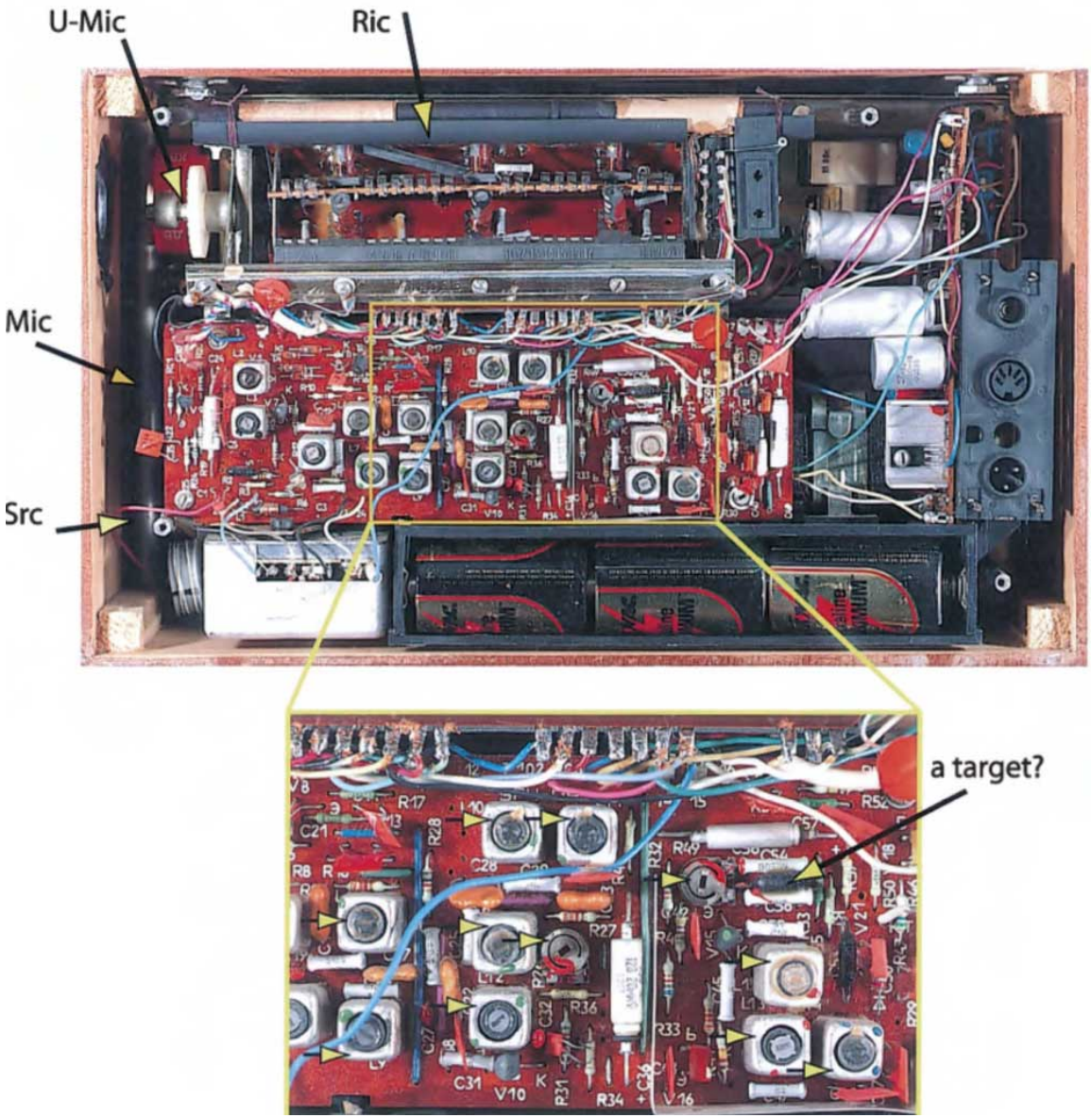


Figure: Lazebnik et al, Cancer Cell, 2002

Why we need models - Can a biologist fix a radio?

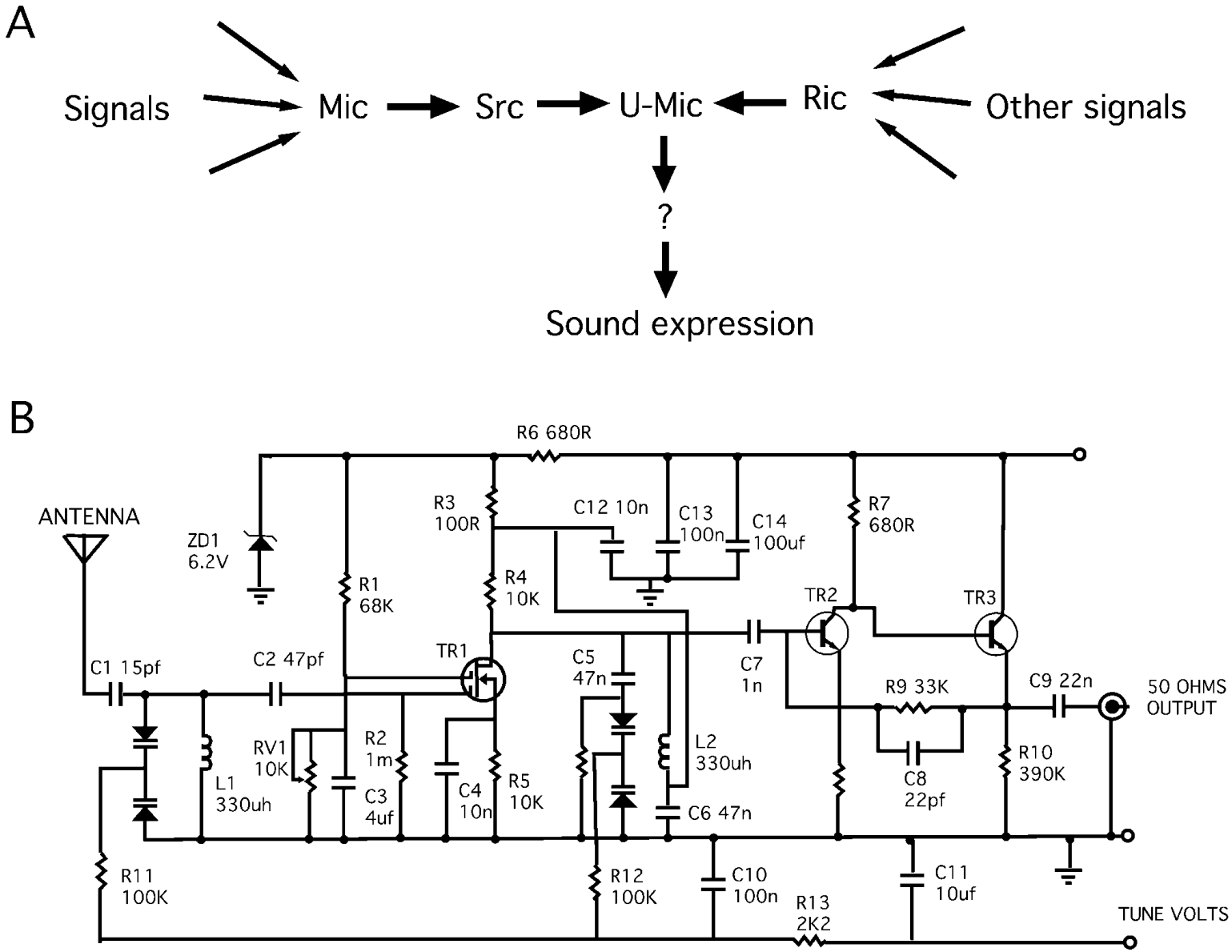
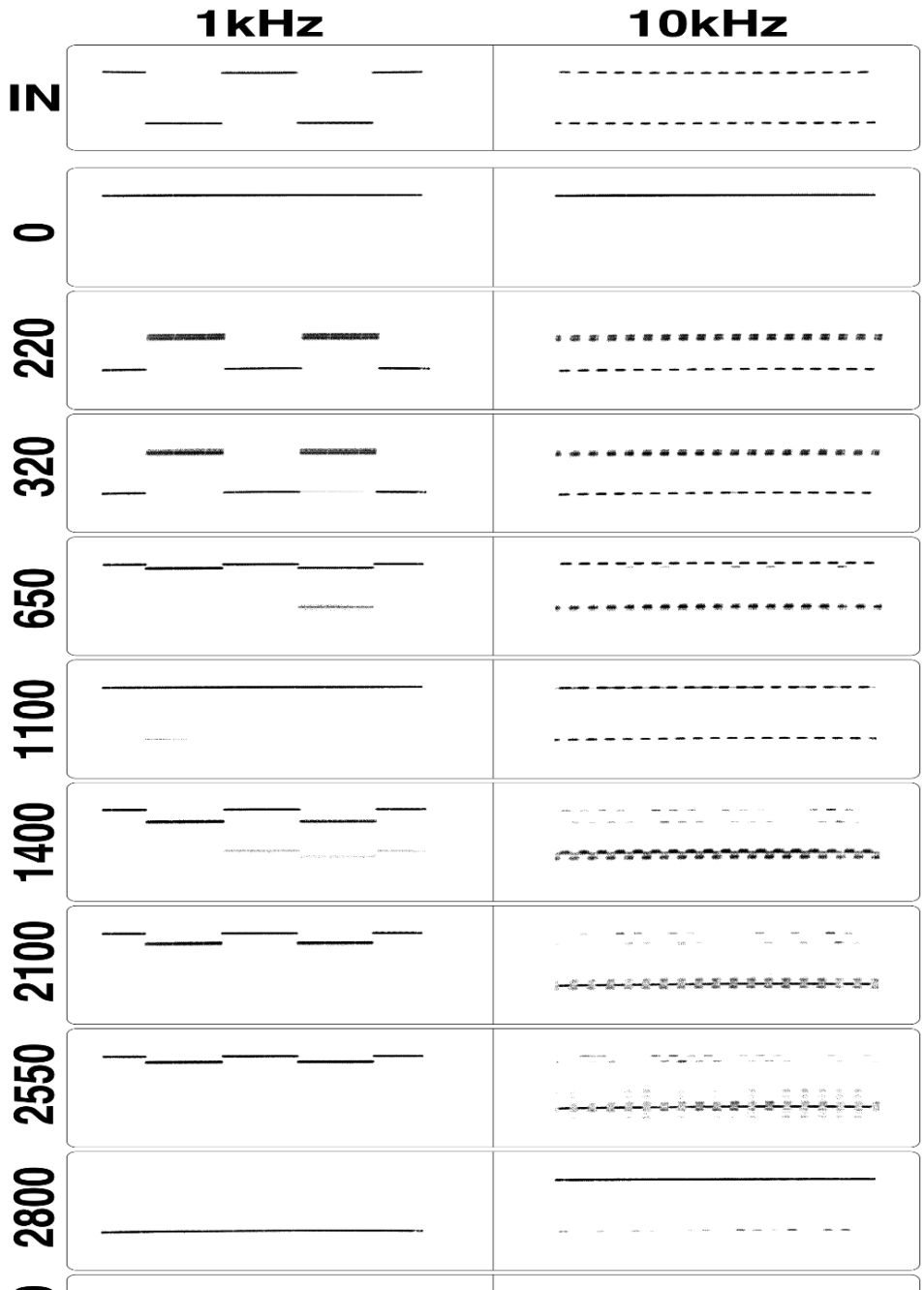


Figure: Lazebnik et al, Cancer Cell, 2002

Comparisons

- Multiscale nature
  - Biology operates on many scales
  - Same is true for electronics
  - BUT electronics employ compartmentalization/abstraction to make understandable
- Component-wise understanding
  - Only provides basic characterization
  - Leads to "context-dependent" function

Machine learning can outperform a human for some tasks



Machine learning can outperform a human for some tasks

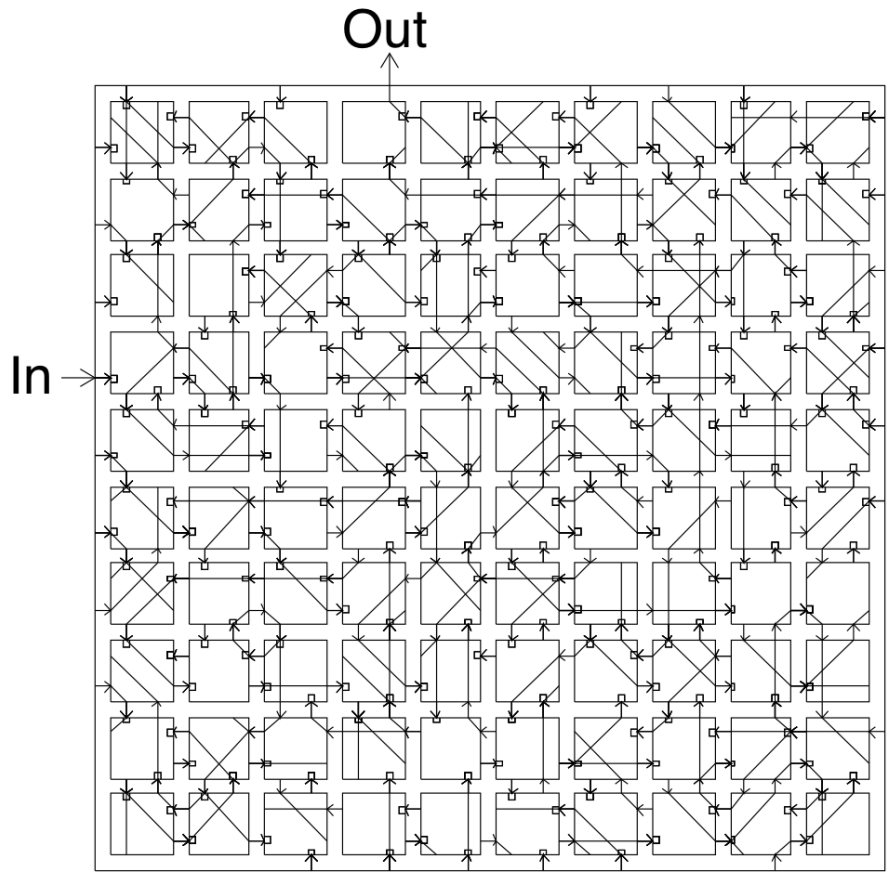


Figure: Thompson et al. Proc. 1st Int. Conf. on Evolvable Systems, 1996

Machine learning can outperform a human for some tasks

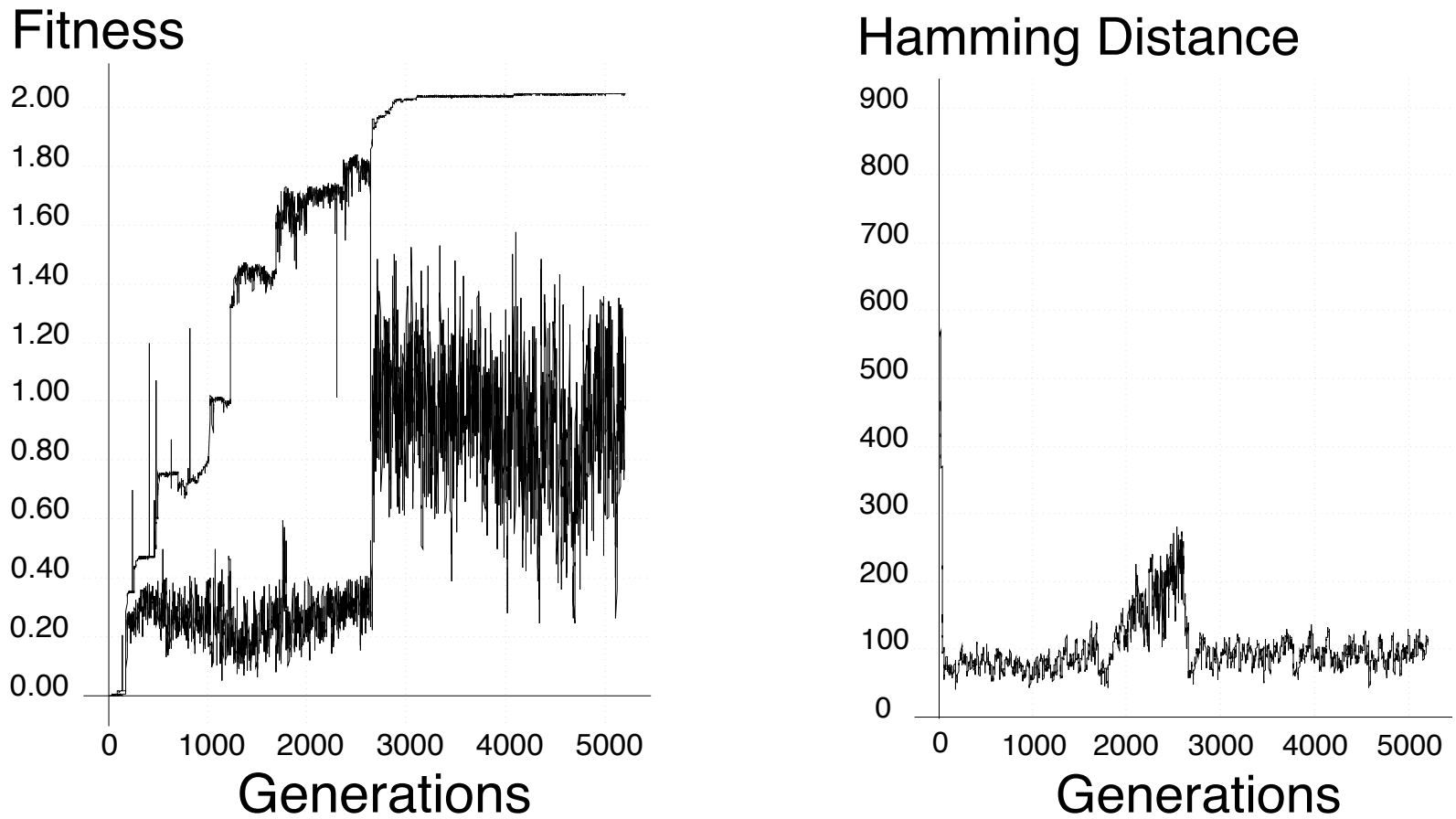


Figure: Thompson et al. Proc. 1st Int. Conf. on Evolvable Systems, 1996

Machine learning can outperform a human for some tasks

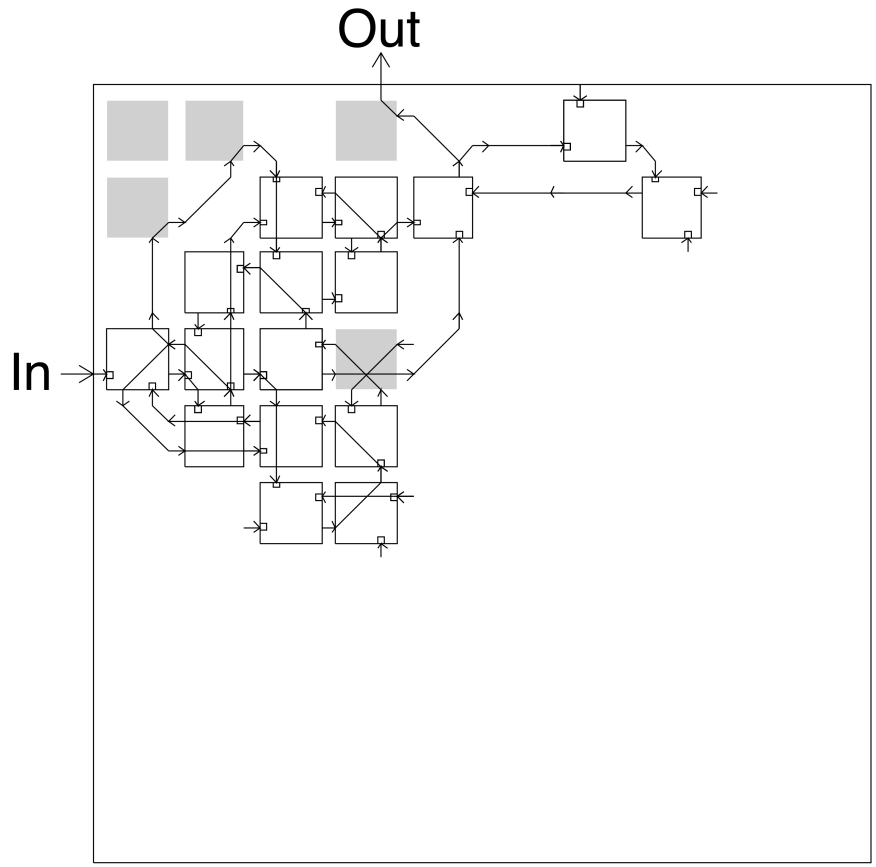


Figure: Thompson et al. Proc. 1st Int. Conf. on Evolvable Systems, 1996

Data

- What is a variable?
- What is an observation?
- What is N?

Types of variables

- Categorical
- Numerical/continuous
- Ordinal

Probability

Coin toss example

A set of trials: HTHHHTTHHTT

Two possibilities:

- Fair coin (heads 50%, tails 50%)
- Biased (heads 60%, tails 40%)

Distributions

We've already been talking about these! Distributions describe the range of probabilities that exist for all possible outcomes.

Other Probabilities

Conditional probability The measure of an event given that another event has occured.

Marginal distribution The probability distribution regardless of other observations/factors.

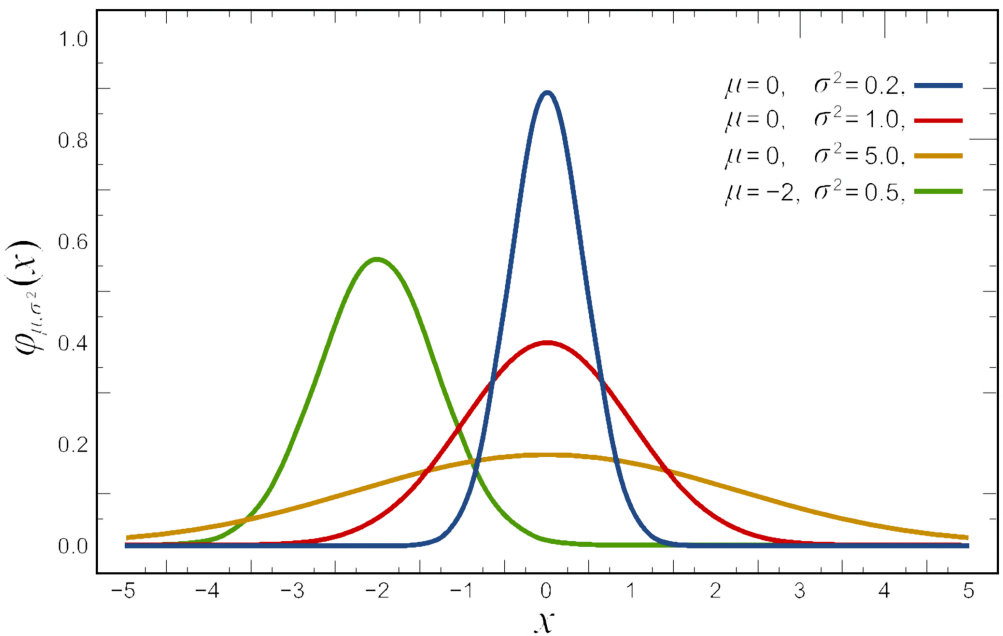
Joint probability In a multivariate probability space, the distribution for more than one variable.

Complementary event The probability of an event not occuring.

Normal Distributions

- μ: center of the distribution
- σ: standard deviation
- σ^2: variance

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



Normal Distributions

For a *standard* normal distribution ($\mu = 0$, $\sigma = 1$):

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Area between:

- One standard deviation: 68%
- Two stdev: 95%
- Three stdev: 99.7%

You can normalize any normal distribution to the standard normal.

Other Distributions

Normal Distribution Describes many naturally observed variables and has statistics mean and standard deviation

Exponential Distribution Describes the time between events in Poisson Processes

Poisson Stochastic process that counts events in some time frame

Rayleigh Measure of vector magnitude when orthogonal directions are independent

Gamma Used in Bayesian statistics, often for modeling waiting times

Beta Random variables limited to intervals of finite length (e.g. Allele frequency in population genetics)

Bernoulli From binary Bernoulli trial, like a coin flip, describes the probability of observing a single event on next flip

Binomial Extension of the bernoulli trial, describes the # of successes in a sequence of n-independent binary trials

Multinomial Generalization of Binomial for more than two states.

Distribution moments

The moments of a distribution describe its shape:

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n \, f(x) \, \mathrm{d}x$$

First	Mean
Second	Variance
Third	Skewness
Fourth	Kurtosis

- Essential properties to determining how a set of data will behave during analysis
  - How might your measurements need to change with changes in variance?
  - What are these values for a normal distribution?

https://gregorygundersen.com/blog/2020/04/11/moments/

Sample statistics

If we sampled a number of times ($n = 3$, say) many times, we could build a **sampling** distribution of the statistics (e.g., one for the **sample** mean and one for the **sample** standard deviation).

General properties of sampling distributions:

1. The sampling distribution of a statistic often tends to be centered at the value of the population parameter estimated by the statistics
2. The spread of the sampling distributions of many statistics *tends* to grow smaller as sample size n increases
3. As n increases, sampling distributions tends towards normality. If a process has mean μ and standard deviation σ, them the sample mean $= \mu$ and the sample standard deviation $= \sigma/\sqrt{n}$

Sample statistics

- This means that as n increases, the better estimate μ_x is of μ.
  - Sample standard deviation is the standard deviation of the mean.
- When a population distribution is normal, the sampling distribution of the sample statistic is also normal, regardless of n.
- And the central limit theorem states that the sampling distribution can be approximated by a normal distribution when the sample size, n, is sufficiently large.
- Rule of thumb is that $n = 30$ is sufficiently large, but there are times when smaller n will suffice. Greater n is required with higher skew.

Hypothesis Testing

In hypothesis testing, we state a null hypothesis that we will test; if its likelihood is less than some value, then we reject it.

For example:

- H_0: A particular set of points come from a normal distribution with mean μ and variance σ.
- H_0: Two sets of observations were sampled from distributions with different means.

T-distribution

When n is small use the t-distribution with $n-1$ degrees of freedom.

- H_0: Assume $\mu = \mu_0$ then calculate t.

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

  - Can think of t designed to be z/s, where it's sensitive to the magnitude of the difference to the alternate hypothesis and scaled to control for the spread.
- When comparing the differences between two means: (null hypothesis the means are the same, variances/sizes assumed equal).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2_{X_1} + s^2_{X_2}}{n}}}$$

Effect size

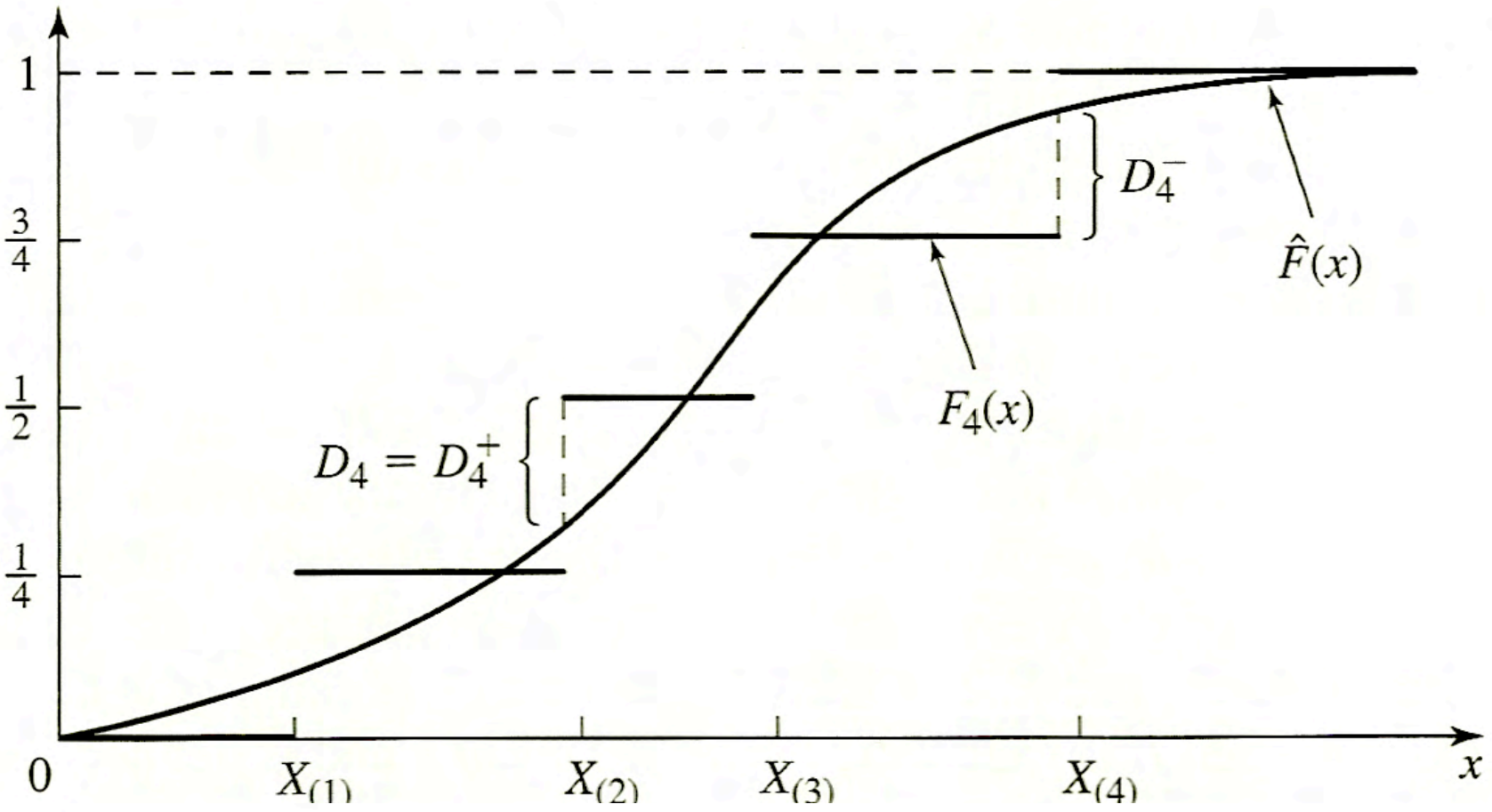
- The scalar factor scales the t-value
  - If using a direct guassian, the estimation of the mean scales with $1/\sqrt{n}$
  - Then p-values become significant even though the differences in means is small
- Exercise caution and report the effect size
  - For example, a 1% or 50% difference in the means

Kolmogorov-Smirnov Test

- Comparison of an empirical distribution function with the distribution function of the hypothesized distribution.
- Does not depend on the grouping of data.
- Relatively insensitive to outlier points (i.e., distribution tails).

Kolmogorov-Smirnov Test

- K-S test is most useful when the sample size is small
- Geometric meaning of the test statistic:



Kolmogorov-Smirnov Test

Test statistic:

$$D_n^+ = \max_{1 \le i \le n} \left(\frac{i}{n} - \hat{F}(X_{(i)}) \right)$$

$$D_n^- = \max_{1 \le i \le n} \left(\hat{F}(X_{(i)} - \frac{i-1}{n}) \right)$$

$$D_n = \max \left(D_n^+, D_n^- \right)$$

Not expressed in one equation with absolute value because distance is assessed from opposite ends for each.
How is this then converted to a p-value?

Graphical Analysis

- Plotting a distribution is often more informative than a goodness-of-fit test.
- Not only assesses deviation, but can explain where it occurs.
- Many variants:
  - Q-Q plot
  - P-P plot
  - Histogram with fitted distribution

Testing errors

- Type I error: error of rejecting H_0 when it is true (false positive)
- Type II error: not rejecting H_0 when it is false (false negative)
- Alpha: significance level in the long run H_0 would be rejected this amount of the time falsely (i.e., we are willing to accept x fraction of false positives)

Beware of goodness-of-fit tests because they are unlikely to reject *any* distribution with little data, and are very sensitive to the smallest systematic error with lots of data.

Multiple hypotheses

We want to test whether the gene expression between two cells differs greater than chance alone. We test the two samples with a p-value cutoff of 0.05:

- How many false positives would we expect after testing 20 genes?
- How about 1000 genes?

What about false negatives?

What does this mean when it comes to hypothesis testing?

Further Reading

- Computer Age Statistical Inference, Chapters 1 and 2
- `scipy.stats`