# Quantitative Genomics and Genetics

## BTRY 4830/6830; PBSB.52201.03

*Lecture 2: Introduction to probability basics*

Jason Mezey

Jan 25, 2024 (Th) 8:40-9:55

# Times and Locations I

- Lectures are every Tues. / Thurs. 8:40-9:55AM - see class schedule (to be posted!)
- In-person lecture locations:
  - Ithaca: In-person lectures in Weill Hall 224 (Tues) and Weill Hall 226 (Thurs)
  - NYC: Belfer Building (69th between 1st and York) - 2nd or 3rd floor - a schedule will be posted on CANVAS later today (!!)
- Lectures will be recorded:
  - These will be posted along with slides / notes
  - In person lecture attendance is not required BUT I (strongly) encourage you to come to class.....

# Times and Locations II

- There is a REQUIRED computer lab
- **FIRST COMPUTER LAB WILL BE NEXT WEEK (Thurs. Feb 1 / Fri. Feb 2)**
- For those IN ITHACA (= Labs with Beulah):
  - Lab 1: 3:35-4:25PM on Thurs. (Mann Library B30A)
  - Lab 2: 9:05-9:55AM on Fri. (Mann Library B30A)
- For those IN NYC (= Labs with Sam!):
  - FRIDAYS (only!) 9-10AM in A-950 Auditorium, 1300 York Ave (9th floor)
  - If you have a problem with this lab time, please contact me (see following slides…)

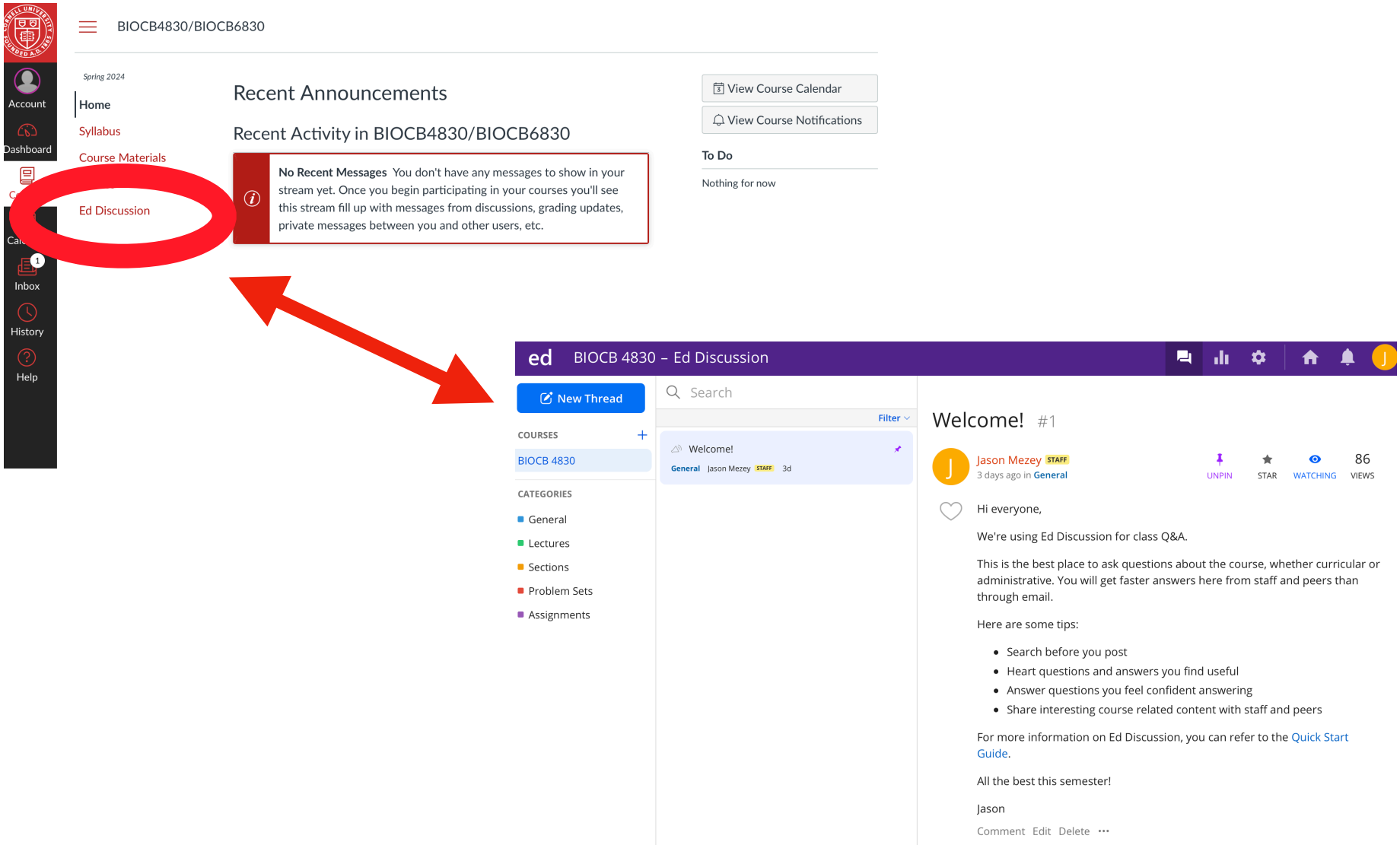
# Registering for the class

- If you can register for this class, please do so (even if you plan to audit!!)
- If you register, you may take this class for a grade (letter in Ithaca, Honors / HP / etc. at Weill), P/F, S/U, or Audit
- If you cannot register for some reason, you are still welcome to take the class (e.g., sit-in) and, if you do the work, we will grade it as if you are registered (!!)
- If you audit or do not register officially, while not required, I strongly recommend that you do the work for the class, (i.e. homework / exams / project / computer lab)
- My observation is that you are likely to be wasting your time if you do not do the work but I leave this up to you...

# Getting on Canvas / emailing

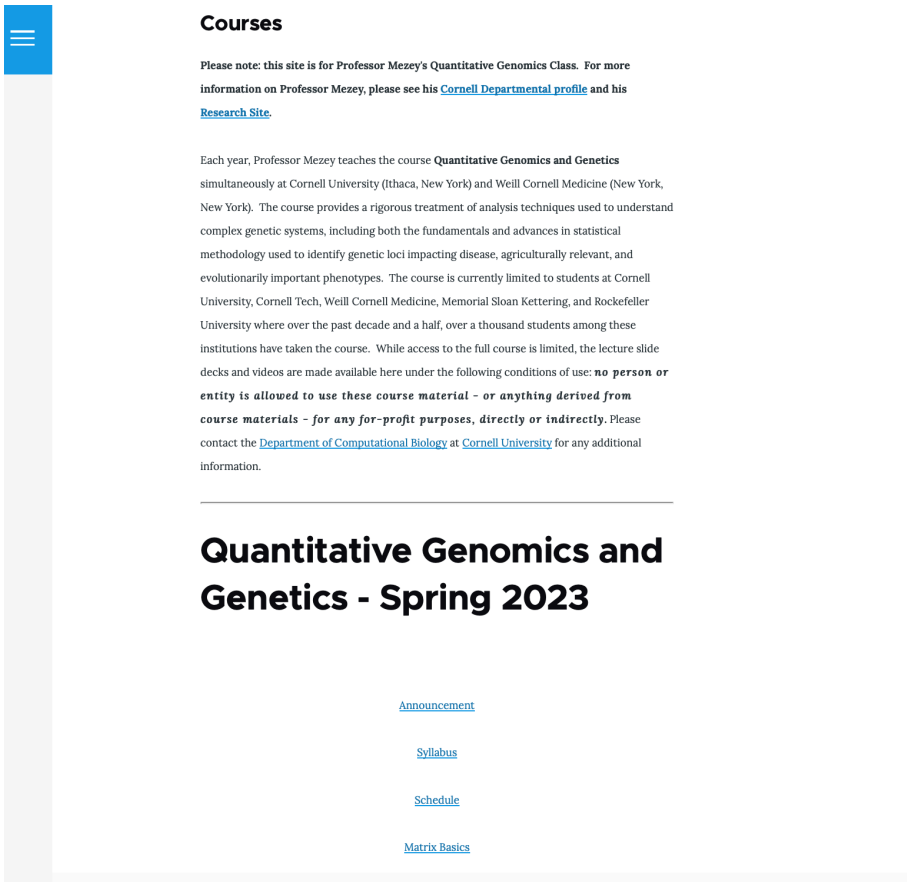
- We will use CANVAS for this class (for everything: posting, announcements, emailing, homework / work uploads, discussion posts, labs, etc.:)
    - If you are at Weill Cornell please register ASAP (!!) using your CWID at: https://request.canvas.cornell.edu
    - Please email me if you cannot sign up on the class Canvas for some reason (!!)
- ALL EMAIL for any aspect of the course must be sent through Canvas (we will stop answering direct emails after the first week of the course)
- PLEASE DON'T email Jason / Beulah / Sam's direct email after the first week (=we will ignore you - unless its an emergency…)
- We sent test email from Canvas last night (Weds, Jan 24) - if you received this email, you are good to go! If you ARE up on Canvas and DID NOT receive this email - please let me know (by Canvas email)

# Canvas posting / discussions



# Class Resource 11: Website

- An additional class website: https://mezeylab.biohpc.cornell.edu



## Courses

**Please note: this site is for Professor Mezey's Quantitative Genomics Class. For more information on Professor Mezey, please see his Cornell Departmental profile and his Research Site.**

Each year, Professor Mezey teaches the course **Quantitative Genomics and Genetics** simultaneously at Cornell University (Ithaca, New York) and Weill Cornell Medicine (New York, New York). The course provides a rigorous treatment of analysis techniques used to understand complex genetic systems, including both the fundamentals and advances in statistical methodology used to identify genetic loci impacting disease, agriculturally relevant, and evolutionarily important phenotypes. The course is currently limited to students at Cornell University, Cornell Tech, Weill Cornell Medicine, Memorial Sloan Kettering, and Rockefeller University where over the past decade and a half, over a thousand students among these institutions have taken the course. While access to the full course is limited, the lecture slide decks and videos are made available here under the following conditions of use: ***no person or entity is allowed to use these course material - or anything derived from course materials - for any for-profit purposes, directly or indirectly.*** Please contact the Department of Computational Biology at Cornell University for any additional information.

---

## Quantitative Genomics and Genetics - Spring 2023

Announcement

Syllabus

Schedule

Matrix Basics

- This has not yet been updated but currently has videos from last year and lecture slide decks from last year (=same content) - take a look!

# Summary of lecture 2: Introduction to probability basics

- In this class, we will be concerned with the most basic problem of quantitative genomics: how to identify genotypes where differences among individual genomes produce differences in individual phenotypes (i.e. genetic association studies)
- Today we will start with the basic question "why DNA?" discuss the rigorous conceptual set-up of probability and essential math concepts

# Motivating intro to prob & statistics: foundational biology concepts

- In this class, we will use ***statistical modeling*** to say something about *biology*, specifically the relationships between genotype (DNA) and phenotype
- Let's start with the biology by asking the following question: why DNA?
- The structure of DNA has properties that make it worthwhile to focus on...

# It's the same in all cells

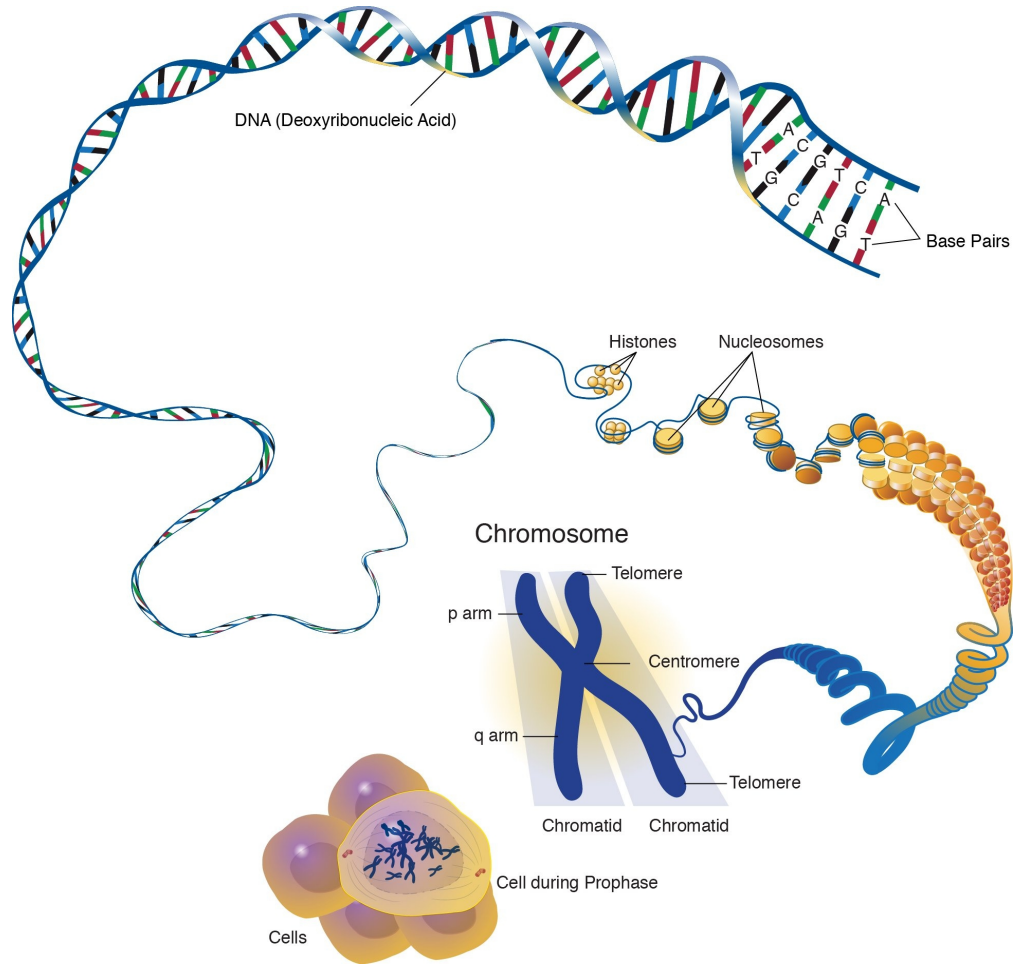
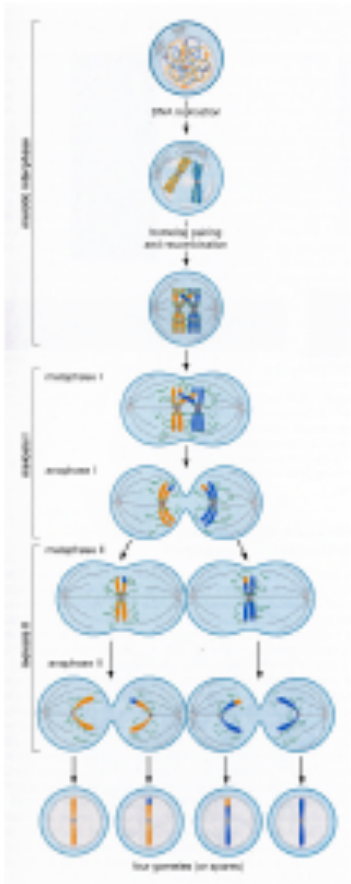
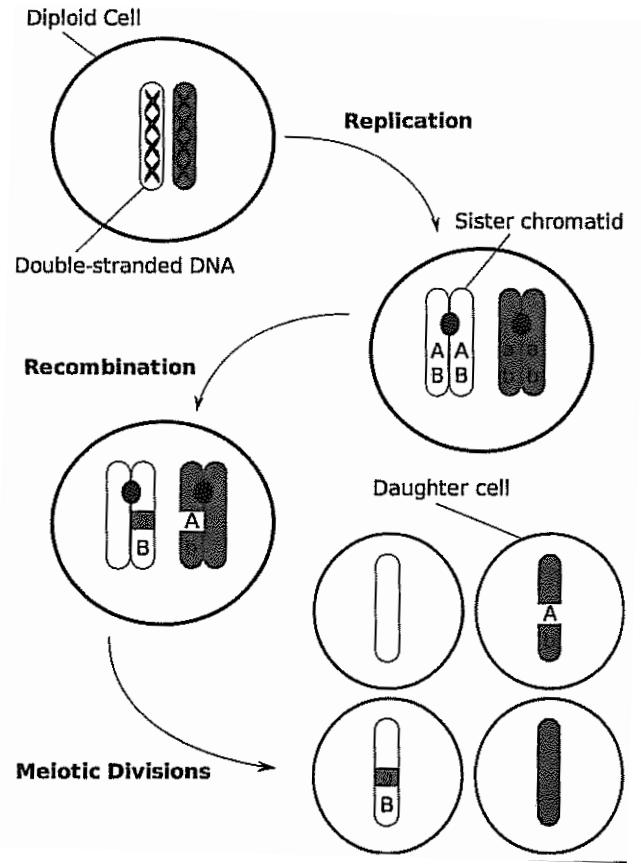



Figure 1: A simplified schematic showing genome organization in human cells. The DNA of a genome is located within the nucleus of a cell. The genome is organized in long strings that are tightly coiled around protein structures to form chromosomes. Each string is a double helix where the building blocks are A-T and G-C nucleotide pairs © *kintalk.org*.

with a few exceptions (e.g. cancer, immune system...)

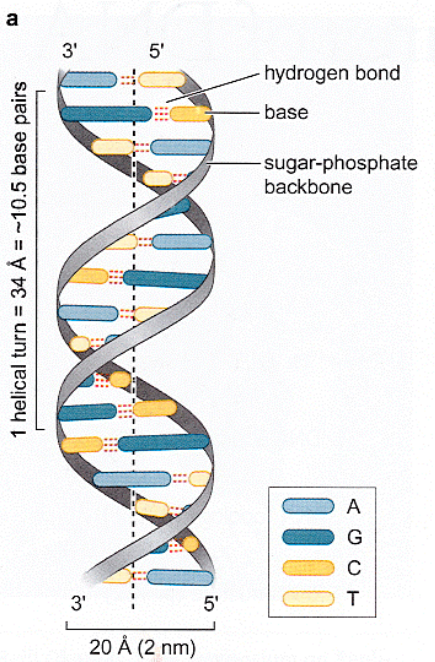
# It's passed on to the next generation

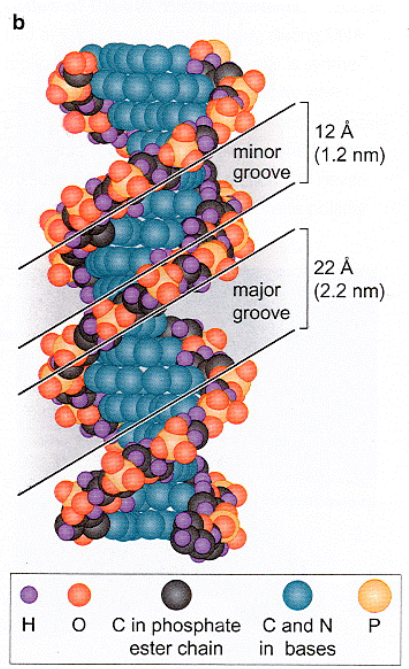


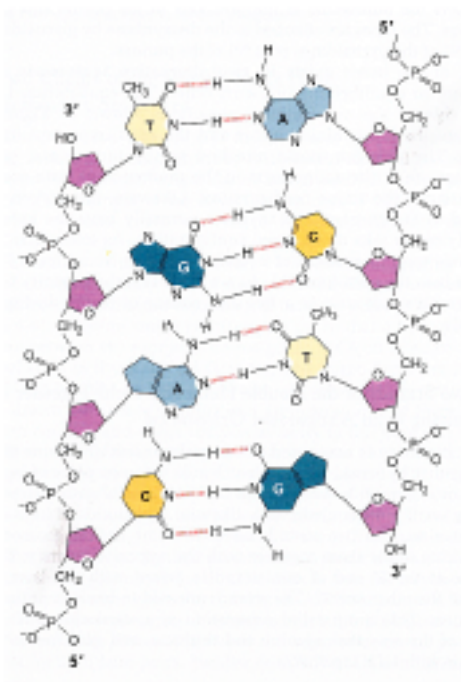


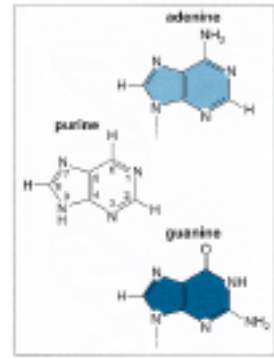

Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004

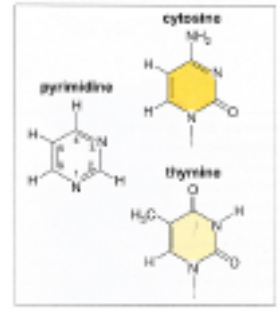
# It has convenient structure for quantifying differences



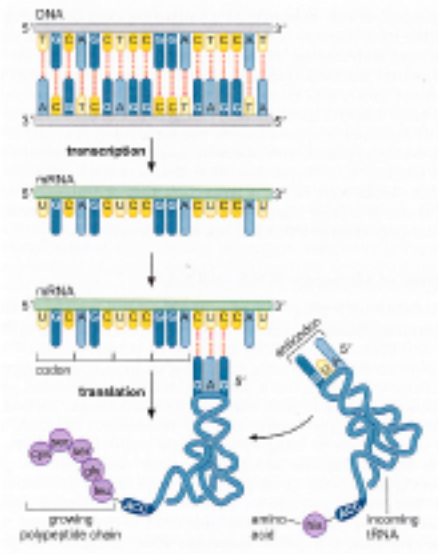




Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004

# It's almost the same in each individual in a species

1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....

# It's responsible for the construction and maintenance of organisms



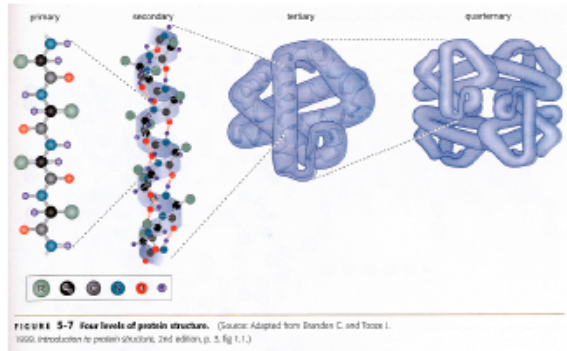




FIGURE 5-7 Four levels of protein structure. (Source: Adapted from Branden C. and Tooze J. 1999. Introduction to protein structure, 2nd edition, p. 5, fig 1.1.)

Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004

Note: other regions of genomes can impact phenotypes...

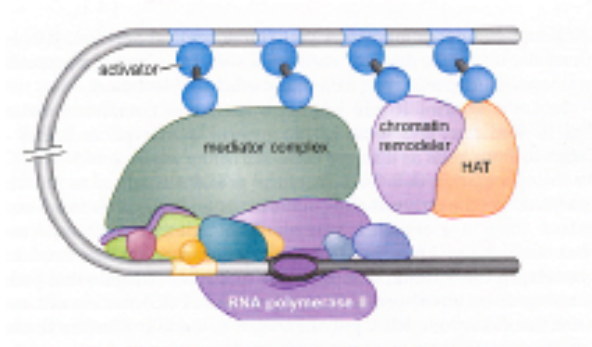



**FIGURE 12-16 Assembly of the pre-initiation complex in presence of Mediator, nucleosome modifiers and remodelers, and transcriptional activators.** In addition to the general transcription factors shown in Figure 12-15, transcriptional activators bound to sites near the gene recruit nucleosome modifying and remodeling complexes, and the Mediator Complex, which together help form the pre-initiation complex.

# Statistics and probability I

- **Quantitative genomics** is a field concerned with the ***modeling*** of the relationship between *genomes* and *phenotypes* and using these models to ***discover*** *and* ***predict***
- We will use frameworks from the fields of probability and statistics for this purpose
- Note that this is not the only useful framework (!!) - and even more generally - mathematical based frameworks are not the only useful (or even necessarily "the best") frameworks for this purpose

# Statistics and probability II

- A non-technical definition of probability:
  a mathematical framework for modeling under uncertainty
- Such a system is particularly useful for modeling systems where we don't know and / or cannot measure critical information for explaining the patterns we observe
- This is exactly the case we have in quantitative genomes when connecting differences in a genome to differences in phenotypes

# Statistics and probability III

- We will therefore use a probability framework to model, but we are also interested in using this framework to discover and predict
- More specifically, we are interested in using a probability model to identify relationships between genomes and phenotypes using DNA sequences and phenotype measurements (=Data)
- For this purpose, we will use the framework of *statistics*, which we can (non-technically) define as a system for interpreting data for the purposes of prediction and decision making given uncertainty

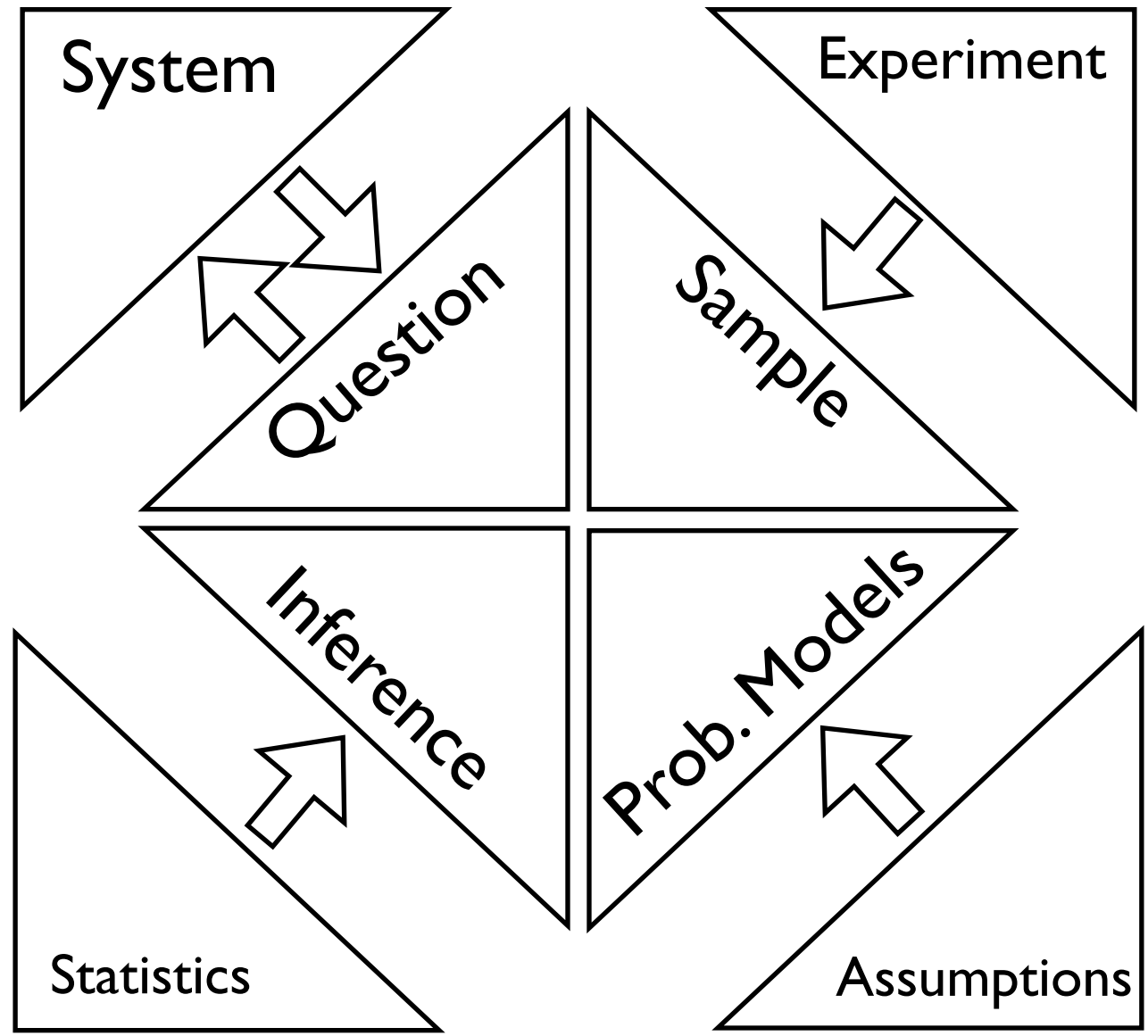
# Definitions: Probability / Statistics

- **Probability** (non-technical def) - a mathematical framework for modeling under uncertainty

- **Statistics** (non-technical def) - a system for interpreting data for the purposes of prediction and decision making given uncertainty

*These frameworks are particularly appropriate for modeling genetic systems, since we are missing information concerning the complete set of components and relationships among components that determine genome-phenotype relationships*

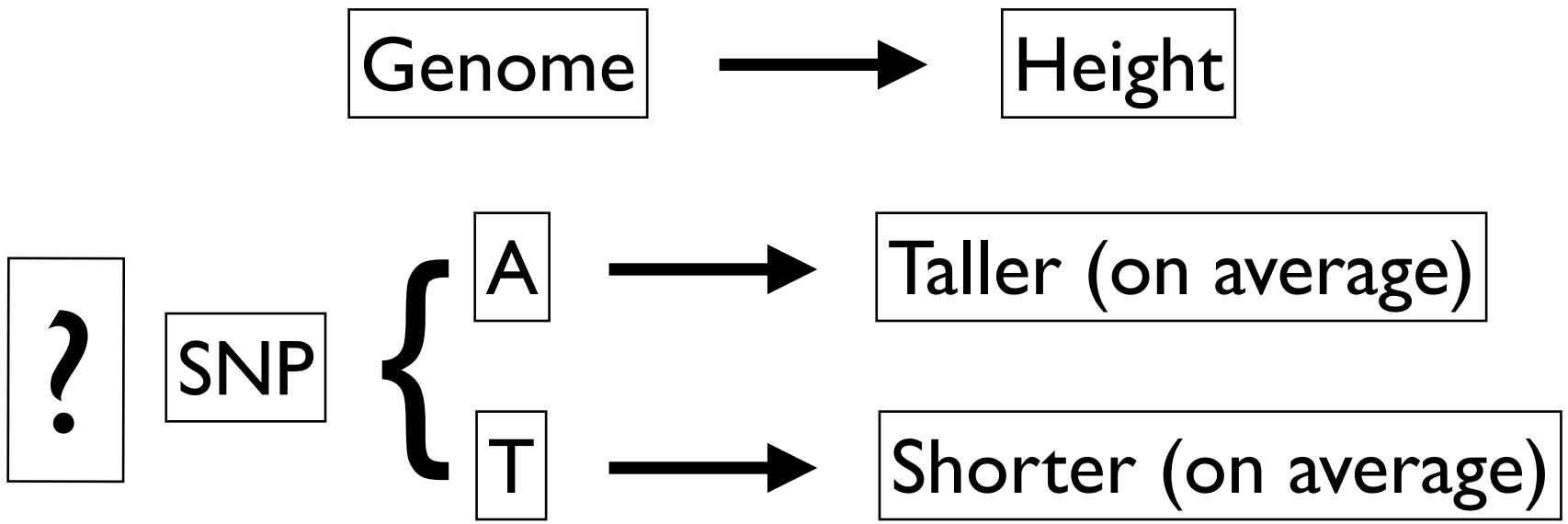
# Conceptual Overview



# Starting point: a system

- **System** - a process, an object, etc. which we would like to know something about

- Example: Genetic contribution to height



# Starting point: a system

- **System** - a process, an object, etc. which we would like to know something about

- Examples: (1) coin, (2) heights in a population

Coin - same amount of metal on both sides?

Heights - what is the average height in the US?

# Experiments (general)

- To learn about a system, we generally pose a specific question that suggests an experiment, where we can extrapolate a property of the system from the results of the experiment
- Examples of "ideal" experiments (System / Experiment):
  - SNP contribution to height / directly manipulate A -> T keeping all other genetic, environmental, etc. components the same and observe result on height
  - Coin / cut coin in half, melt and measure the volume of each half
  - Height / measure the height of every person in the US

# Experiments (general)

- To learn about a system, we generally pose a specific question that suggests an experiment, where we can extrapolate a property of the system from the results of the experiment
- Examples of "non-ideal" experiments (System / Experiment):
  - SNP contribution to height / measure heights of individuals that have an A and individuals that have a T
  - Coin / flip the coin and observe "Heads" and "Tails"
  - Height / measure some people in the US

# Experiments and Outcomes

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe
- **Experiment Outcome** - a possible result of the experiment
- Example (Experiment / Outcomes):
  - Coin flip / "Heads" or "Tails"
  - Two coin flips / HH, HT, TH, TT
  - Measure heights in this class / 1.5m, 1.71m, 1.85m, …

# Sets / Set Operations / Definitions

- **Set** - any collection, group, or conglomerate
- **Element** - a member of a set
- A Special Set: **Empty Set** ($\emptyset$) $\equiv$ the set with no elements (the empty set is unique and is sometimes represented as { }).
- Set Operations:

  **Union** ($\cup$) $\equiv$ an operator on sets which produces a single set containing all elements of the sets.

  **Intersection** ($\cap$) $\equiv$ an operator on sets which produces a single set containing all elements common to all of the sets.

- Important Definitions:

  **Element of** ($\in$) $\equiv$ an object within a set, e.g. $H \in \{H, T\}$

  **Subset** ($\subset$) $\equiv$ a set that is contained within another set, e.g. $\{H\} \subset \{H, T\}$

  **Complement** ($\mathcal{A}^c$) $\equiv$ the set containing all other elements of a set other than $\mathcal{A}$, e.g. $\{H\}^c = \{T\}$.

  **Disjoint Sets** $\equiv$ sets with no elements in common.

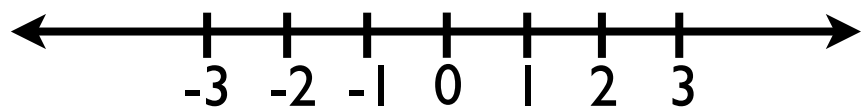
# Some Special Sets

- The following sets have properties that align with our intuitive conception about how we represent and use groups
  - The **Natural Numbers** and the **Integers:**

$$\mathbb{N} = \{1, 2, 3, ...\}$$

$$\mathbb{Z} = \{... - 3, -2, -1, 0, 1, 2, 3, ...\}$$

  - The **Reals:**

$\mathbb{R} =$ 


- Note that these sets are infinite (although they represent two different "sizes" of infinite: countable and uncountable), where we often make use of the following symbols in both cases:

$$-\infty > x > \infty$$

# Sample Spaces

- **Sample Space** ($\Omega$) - set comprising all possible outcomes associated with an experiment
- (Note: we have not defined a **Sample** - we will do this later!)
- Examples (Experiment / Sample Space):
  - "Single coin flip" / {H, T}
  - "Two coin flips" / {HH, HT, TH, TT}
  - "Measure Heights" / any actual measurement OR we could use $\mathbb{R}$
- **Events** - a subset of the sample space
- Examples (Sample Space / Examples of Events):
  - "Single coin flip" / $\emptyset$, {H}, {H, T}
  - "Two coin flips" / {TH}, {HH, TH}, {HT, TH, TT}
  - "Measure Heights" / {1.7m}, {1.5m, ..., 2.2m} OR [1.7m], (1.5m, 1.8m)

# Functions

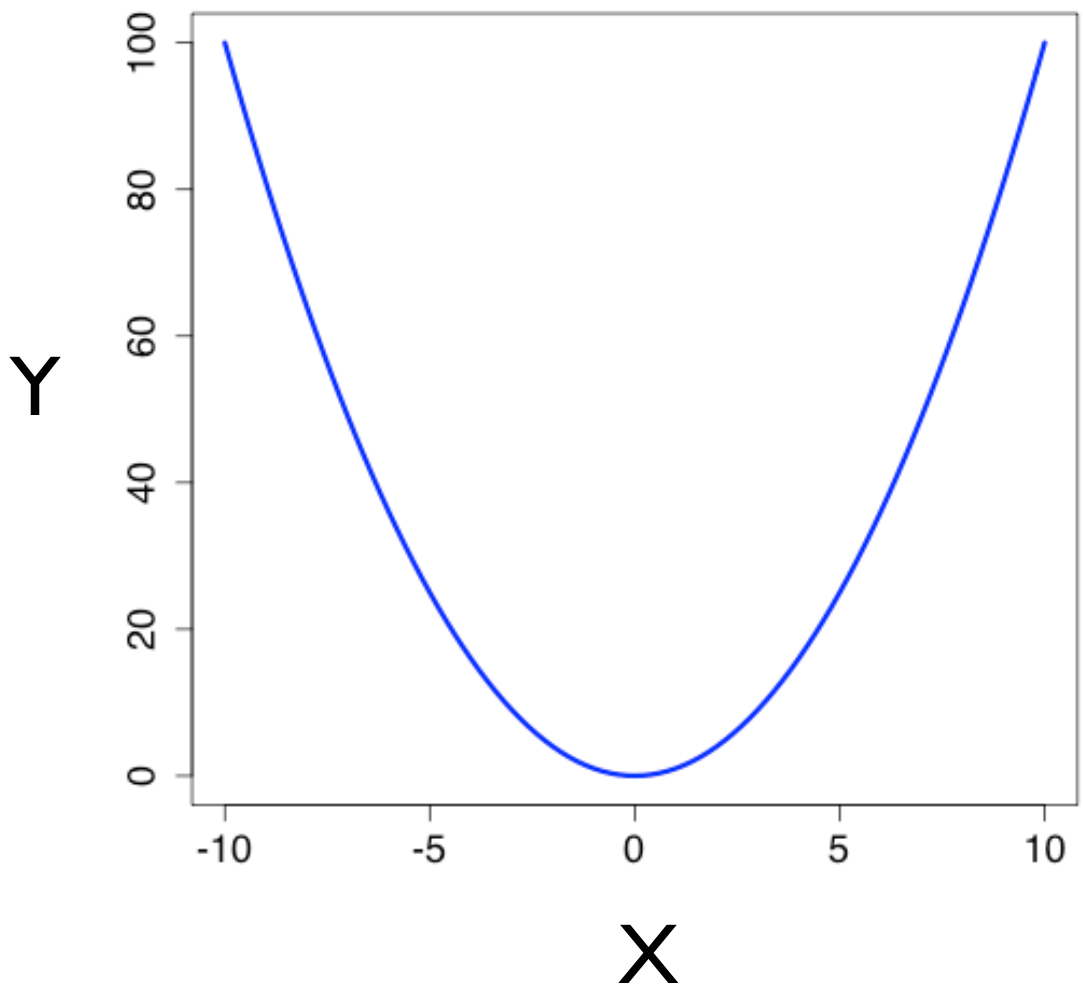
- Now that we have formalized the concept of a sample space, we need to define what "probability" means
- To do this, we need the concept of a mathematical function

- **Function** (formally) - a binary relation between every member of a domain to exactly one member of the codomain

- **Function** (informally) - **?**

# Example of a function

$$Y = X^2$$



# Probability functions (intuition)

- **Probability Function** (intuition) - we would like to construct a function that assigns a number to each event such that it matches our intuition about the "chance" the event will happen (as a result of an experiment)
- To be useful, we need to assign a number not just to each individual ELEMENT of the sample space but to every EVENT
- To accomplish this, we will need the concept of a **Sigma Algebra** (or **Sigma Field**)
- What's more, we need to make sure the function that we use to assign these numbers adheres to a specific set of "rules" (axioms)

# Sample Spaces / Sigma Algebra

- **Sigma Algebra** ($\mathcal{F}$) - a collection of events (subsets) of $\Omega$ of interest with the following three properties: **1**. $\emptyset \in \mathcal{F}$ , **2**. $\mathcal{A} \in \mathcal{F}$ then $\mathcal{A}^c \in \mathcal{F}$ , **3**. $\mathcal{A}_1, \mathcal{A}_2, ... \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} \mathcal{A}_i \in \mathcal{F}$

  *Note that we are interested in a particular Sigma Algebra for each sample space...*

- Examples (Sample Space / Sigma Algebra):
  - {H,T} / $\emptyset, \{H\}, \{T\}, \{H, T\}$
  - {HH, HT,TH,TT} /

$$\emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\},$$
$$\{TH, TT\}\{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{TH, HT, TT\}\{HH, TH, HT, TT\}$$

  - $\mathbb{R}$ / more complicated to define the sigma algebra of interest (see next slide…)

# The (appropriate) Sigma Algebra on the Reals

- For probability, we need an appropriate Sigma Algebra on the Reals (remember there are many possible Sigma Algebra!)
- Interestingly, this Sigma Algebra does not include all subsets of the reals
- One problem is this would include "more sets than we need" for what we need in probability
- Another problem is these subsets include "non-measurable sets" such that if they were included, we could not define a probability measure (!!)
- A way of describing the appropriate Sigma Algebra for the Reals is all open and closed intervals (where *a* and *b* may be any number) and all unions and intersections of these intervals:

$$[a, b], (a, b], [a, b), (a, b)$$

- It seems like these should include all subsets of the Reals, but they don't...

# Review: Probability functions I

- **Probability Function -** maps a Sigma Algebra of a sample to a subset of the reals:

$$Pr : \mathcal{F} \rightarrow [0, 1]$$

- Not all such functions that map a Sigma Algebra to [0,1] are probability functions, only those that satisfy the following Axioms of Probability (where an axiom is a property assumed to be true):

1. For $\mathcal{A} \subset \Omega, Pr(\mathcal{A}) \geqslant 0$
2. $Pr(\Omega) = 1$
3. For $\mathcal{A}_1, \mathcal{A}_2, ... \subset \Omega$, if $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ (disjoint) for each $i \neq j$: $Pr(\bigcup_i^\infty \mathcal{A}_i) = \sum_i^\infty Pr(\mathcal{A}_i)$

- Note that since a probability function takes sets as an input and is restricted in structure, we often refer to a probability function as a *probability measure*

# Probability function: example 1

- For "two coin flips" a probability function will assign a probability to each subset of the Sigma Field:

$$\emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\},$$
$$\{TH, TT\}\{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{TH, HT, TT\}\{HH, TH, HT, TT\}$$

- We could define a probability function as follows:

$$Pr(\emptyset) = 0$$

$$Pr(\{HH\}) = 0.25, Pr(\{HT\}) = 0.25, Pr(\{TH\}) = 0.25, Pr(\{TT\}) = 0.25$$

$$Pr(\{HH, HT\}) = 0.5, Pr(\{HH, TH\}) = 0.5, Pr(\{HH, TT\}) = 0.5,$$

$$Pr(\{HT\,TH\}) = 0.5, Pr(\{HT, TT\}) = 0.5, Pr(\{TH, TT\}) = 0.5,$$

$$Pr(\{HH, HT, TH\}) = 0.75, \quad \text{etc.} \quad Pr(\{HH, HT, TH, TT\}) = 1.0$$

- Not that this is one possible probability model - what other possible probability models could be assumed for this system / experiment?

# Probability function: example II

- The following is (one example) of a probability function (on the sigma algebra) for the two coin flip experiment:

$$Pr(\emptyset) = 0$$

$$Pr(\{HH\}) = 0.25, Pr(\{HT\}) = 0.25, Pr(\{TH\}) = 0.25, Pr(\{TT\}) = 0.25$$

$$Pr(\{HH, HT\}) = 0.5, Pr(\{HH, TH\}) = 0.5, Pr(\{HH, TT\}) = 0.5,$$

$$Pr(\{HT, TH\}) = 0.5, Pr(\{HT, TT\}) = 0.5, Pr(\{TH, TT\}) = 0.5,$$

$$Pr(\{HH, HT, TH\}) = 0.75, \quad \text{etc.} \quad Pr(\{HH, HT, TH, TT\}) = 1.0$$

- The following is an example of a function (on the sigma algebra) of the two coin flip experiment but is not a *probability function*:

$$\not{Pr}(\emptyset) = 0$$

$$\not{Pr}(\{HH\}) = 0.25, \not{Pr}(\{HT\}) = 0.25, \not{Pr}(\{TH\}) = 0.25, \not{Pr}(\{TT\}) = 0.25$$

$$\not{Pr}(\{HH, HT\}) = 0.5, \not{Pr}(\{HH, TH\}) = 0.5, \not{Pr}(\{HH, TT\}) = 1.0,$$

$$\not{Pr}(\{HT, TH\}) = 0, \not{Pr}(\{HT, TT\}) = 0.5, \not{Pr}(\{TH, TT\}) = 0.5,$$

$$\not{Pr}(\{HH, HT, TH\}) = 0.75, \quad \text{etc.} \quad \not{Pr}(\{HH, HT, TH, TT\}) = 1.0$$

# That's it for today

- Next lecture, we will continue our discussion of probability by introducing concept of conditional probability and random variables!