香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

DDA2001: Introduction to Data Science

Introduction

2023/01/10

Zicheng Wang

More Motivating Examples

Example 1

Blind Box Example

- Ten of you went to a blind box seller for the following blind box. You were told the chance to get a Harry Potter versus a Lord Voldemort is fifty-fifty.
- You all wish to get a Harry Potter.

Blind Box Example

- You each bought one blind box. When you opened it, you found…

- What do you think?

Blind Box Example

- If the seller told the truth, the probability that all ten of you got a Voldemort is < 0.001. → Probability

Blind Box Example

- If the seller told the truth, the probability that all ten of you got a Voldemort is < 0.001. → Probability
- Given the above observation, you are pretty confident that the seller was lying. What’s the true chance of getting a Harry Potter versus Voldemort? → Statistics

Blind Box Example

- If the seller told the truth, the probability that all ten of you got a Voldemort is < 0.001. → Probability
- Given the above observation, you are pretty confident that the seller was lying. What’s the true chance of getting a Harry Potter versus Voldemort? → Statistics
- How should the seller set the distribution of items in the boxes to balance the rarity and desirability of items, thereby increasing sales? → Optimization

Blind Box Example

- If the seller told the truth, the probability that all ten of you got a Voldemort is < 0.001. → Probability
- Given the above observation, you are pretty confident that the seller was lying. What’s the true chance of getting a Harry Potter versus Voldemort? → Statistics
- How should the seller set the distribution of items in the boxes to balance the rarity and desirability of items, thereby increasing sales? → Optimization
- Based on historical data, how should the seller improve his/her prediction on customer preferences and future sales trends? → Machine learning

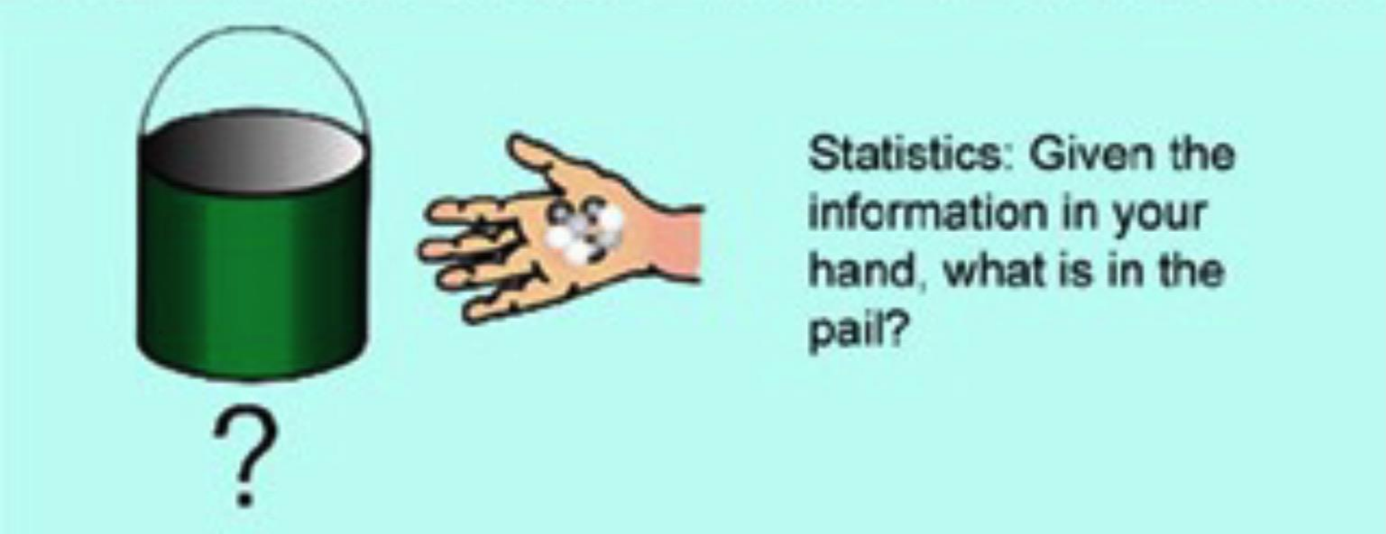
Probability

- A formality to make sense of the world in terms of uncertainty.
  - Numerical descriptions of how likely an event is to occur

?
Probability: Given the information in the pail, what is in your hand?

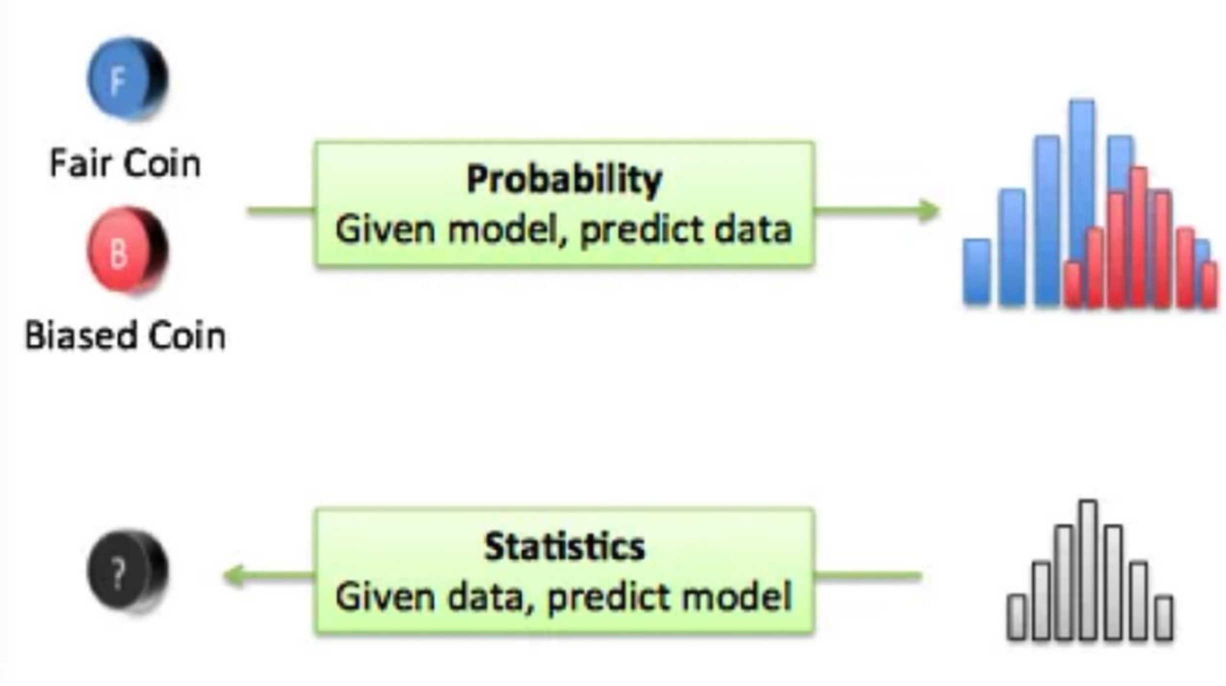
Statistics

- The discipline that extracts correct information from data



Statistics: Given the information in your hand, what is in the pail?

Probability and Statistics



Probability
Given model, predict data

Statistics
Given data, predict model

Questions

1. Given the data, you want to predict how the data is generated. Which theory you may use?
   a) Probability
   b) Statistics

Questions

1. Given the data, you want to predict how the data is generated. Which theory you may use?
   a) Probability
   b) Statistics

Questions

1. Given the data, you want to predict how the data is generated. Which theory you may use?
   a) Probability
   b) Statistics

2. Given the profit as a function of strategies, you want to maximize the profit. Which theory you may use?
   a) Simulation
   b) Optimization

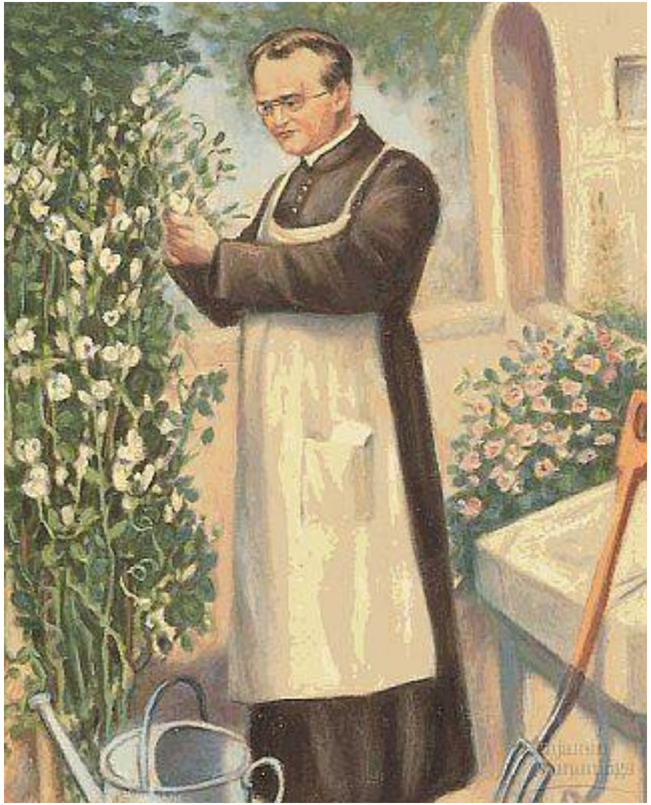
Questions

1. Given the data, you want to predict how the data is generated. Which theory you may use?
   a) Probability
   b) Statistics

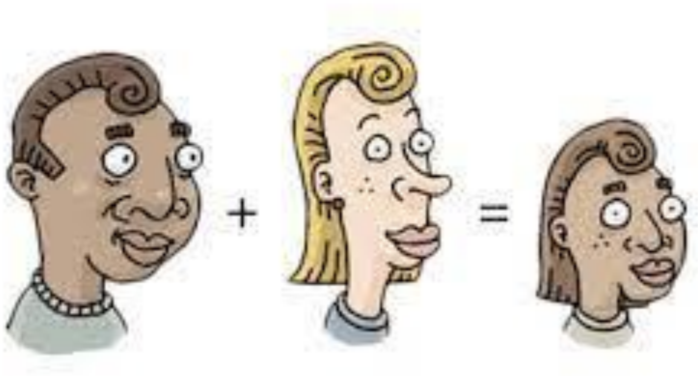
2. Given the profit as a function of strategies, you want to maximize the profit. Which theory you may use?
   a) Simulation
   b) Optimization

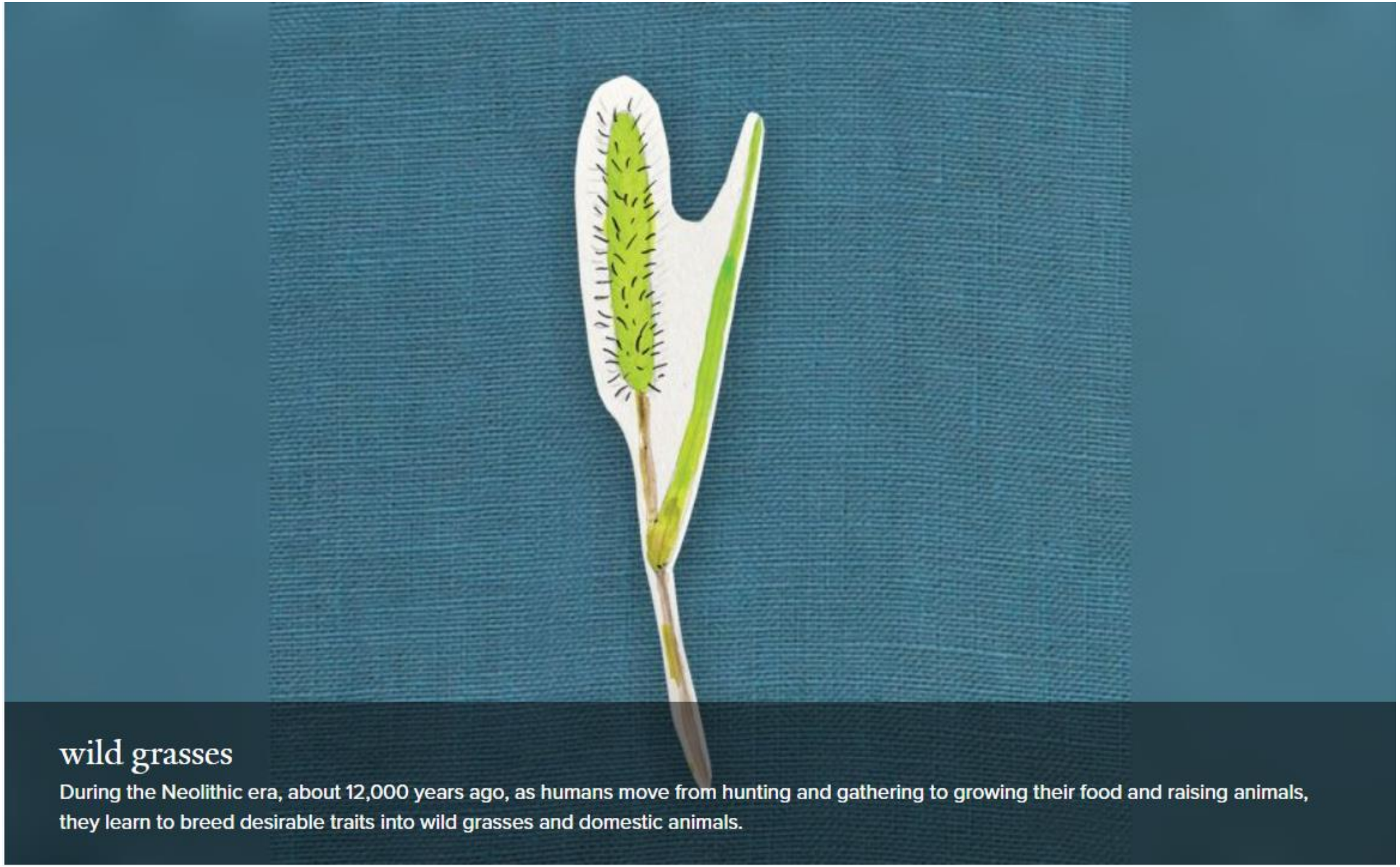
Example 2

The Victory of Data Science over Human Imagination



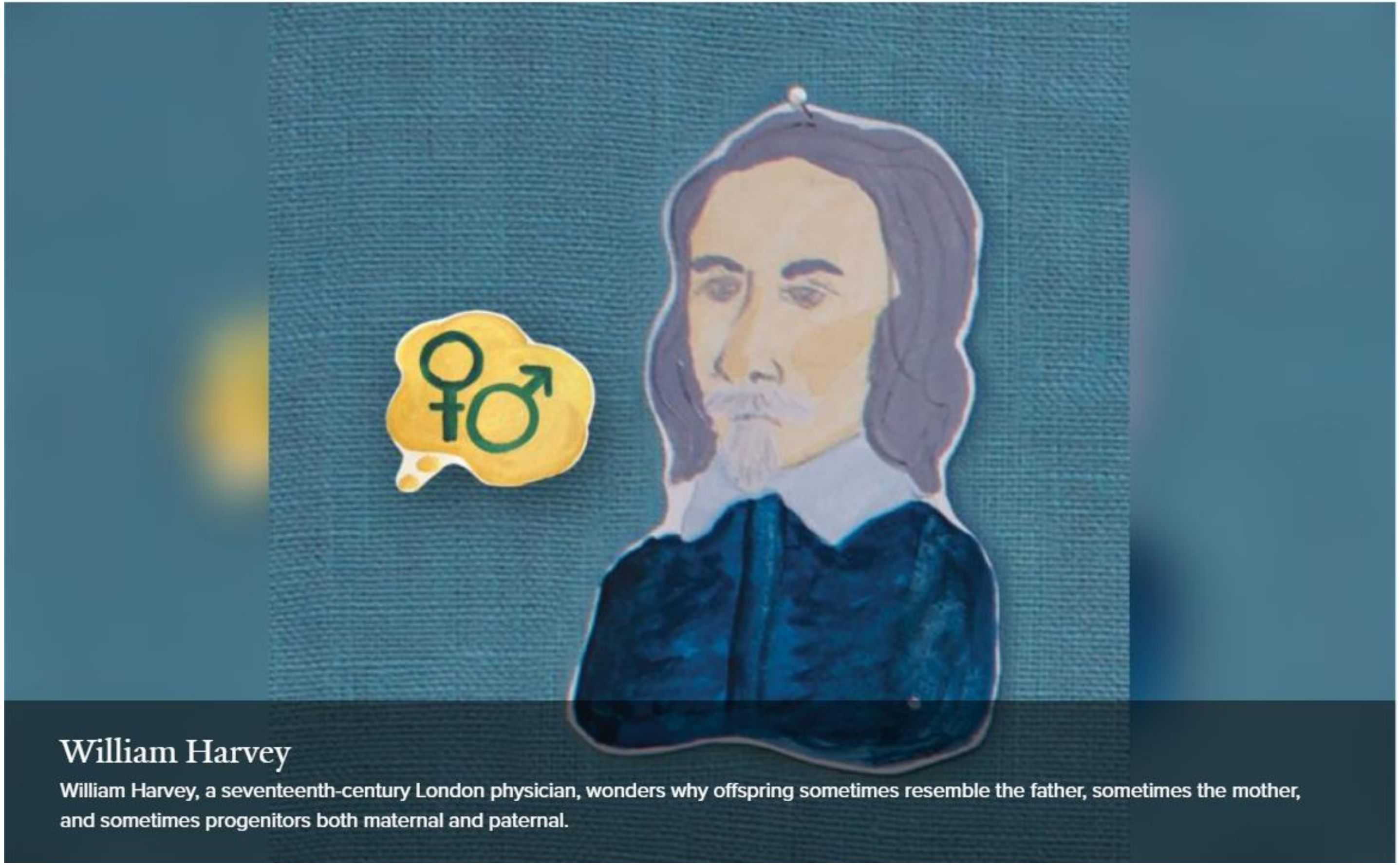
An Explanation for Heredity





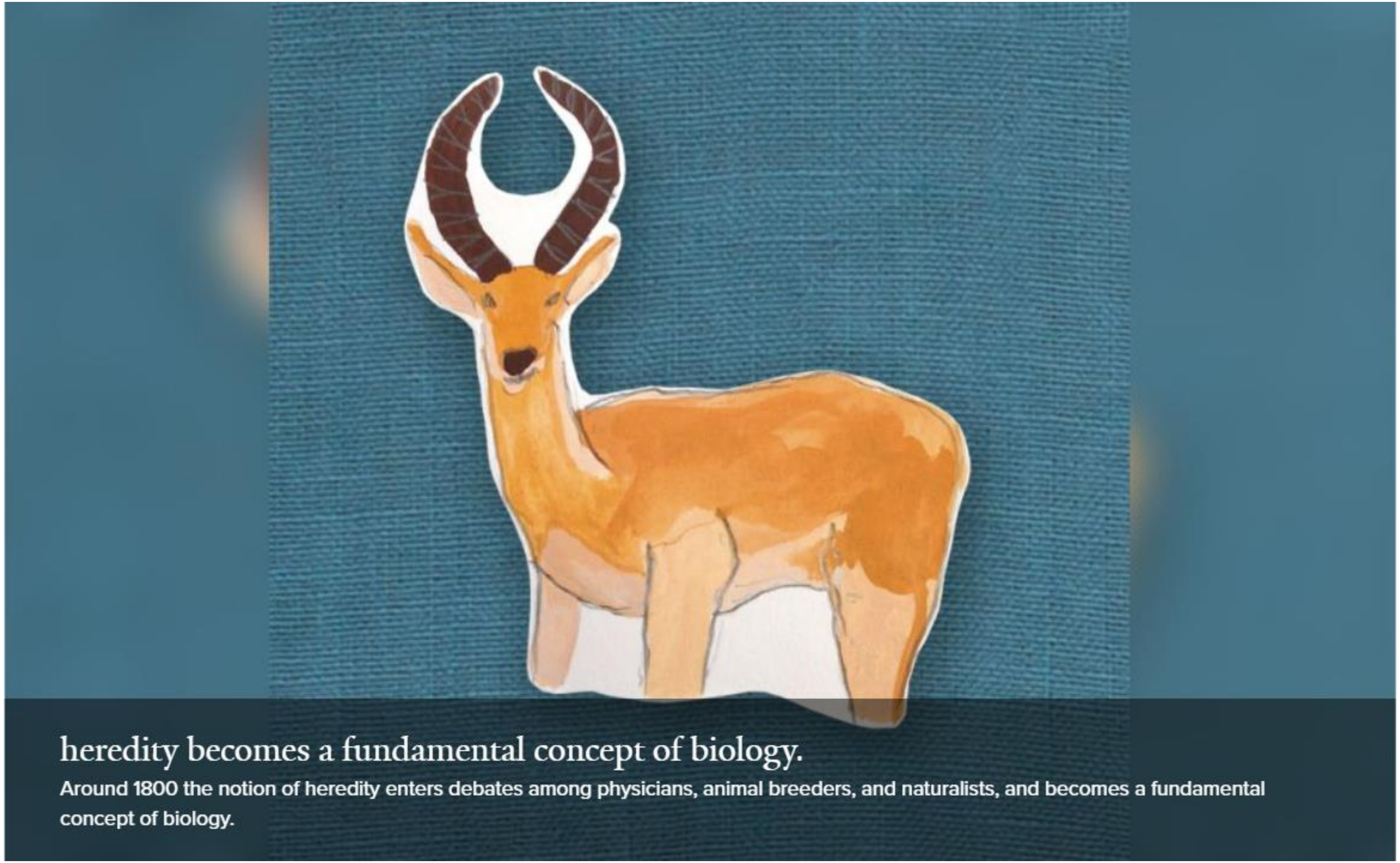
wild grasses

During the Neolithic era, about 12,000 years ago, as humans move from hunting and gathering to growing their food and raising animals, they learn to breed desirable traits into wild grasses and domestic animals.



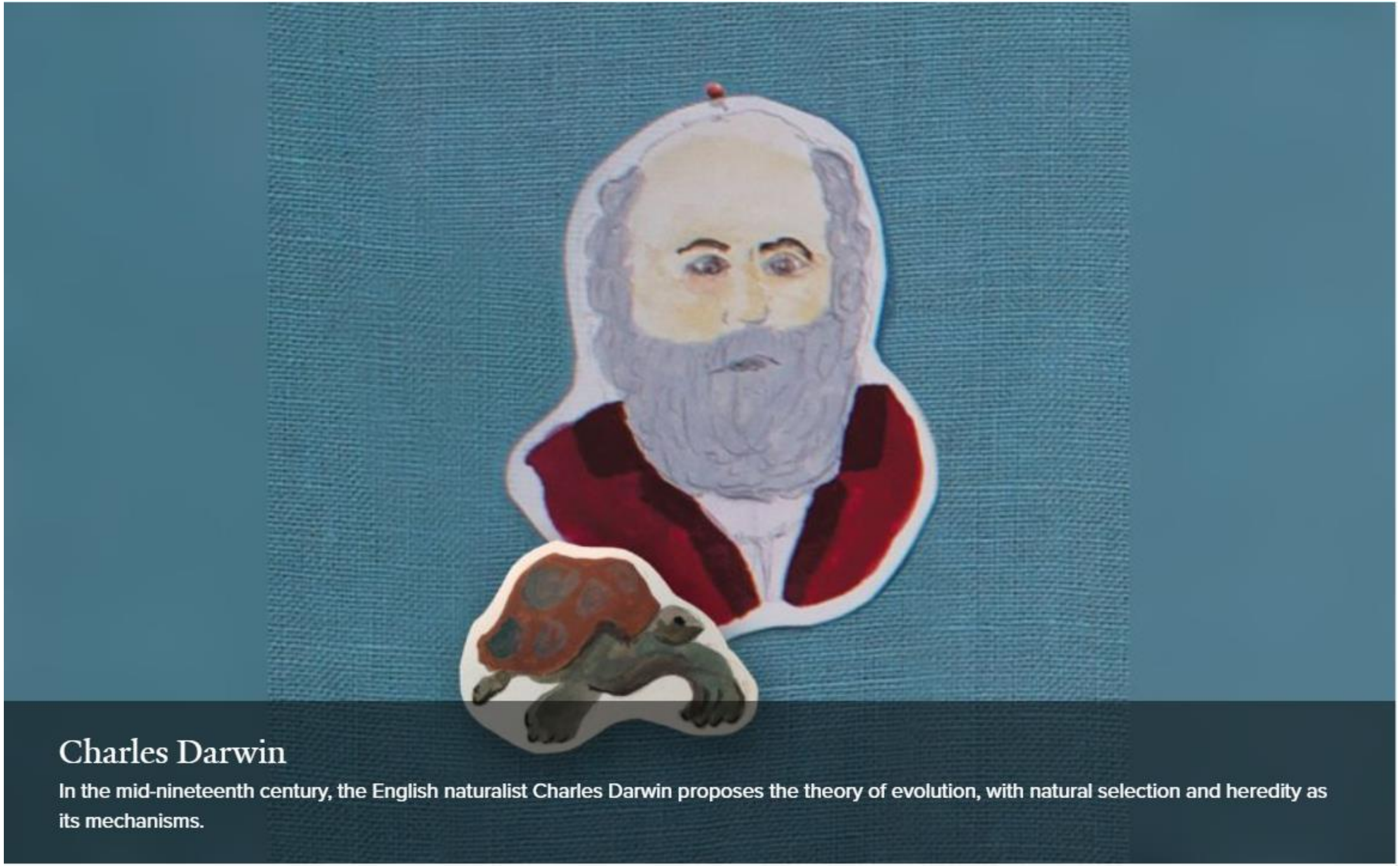
William Harvey

William Harvey, a seventeenth-century London physician, wonders why offspring sometimes resemble the father, sometimes the mother, and sometimes progenitors both maternal and paternal.



heredity becomes a fundamental concept of biology.

Around 1800 the notion of heredity enters debates among physicians, animal breeders, and naturalists, and becomes a fundamental concept of biology.

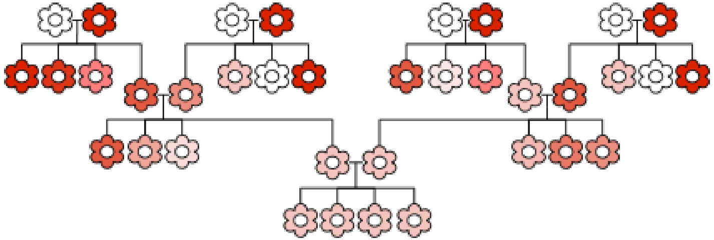


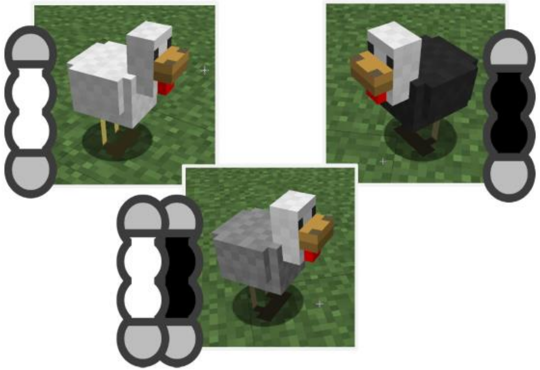
Charles Darwin

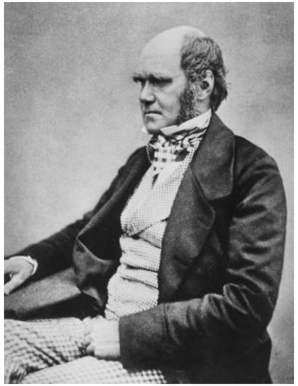
In the mid-nineteenth century, the English naturalist Charles Darwin proposes the theory of evolution, with natural selection and heredity as its mechanisms.

Blending Inheritance

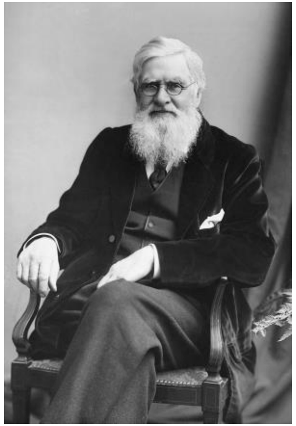
- The idea that the progeny receives an ‘average’ of the parent’s traits
- Red + White = Pink
- White + Black = Grey







Charles Darwin

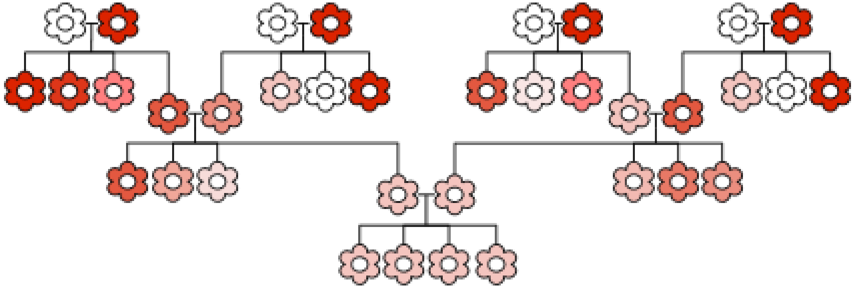


Alfred Wallace

A letter from Charles Darwin to Alfred Wallace

"I do not think you understand what I mean by the non-blending of certain varieties. It does not refer to fertility; an instance I will explain. I crossed the Painted Lady and Purple sweetpeas, which are very differently coloured varieties, and got, even out of the same pod, both varieties perfect but not intermediate. Something of this kind I should think must occur at least with your butterflies & the three forms of Lythrum; tho' those cases are in appearance so wonderful. I do not know that they are really more so than every female in the world producing distinct male and female offspring."

- These findings supported Darwin’s theory of evolution because blending inheritance would mean that beneficial traits would have blended away long before natural selection could occur
- However, Charles Darwin did not collect enough data, nor did he analyze the collected data extensively enough to develop a new theory explaining heredity



Gregor Mendel's Principles of Inheritance

- The principle of uniformity
- The principle of segregation
- The principle of independent assortment

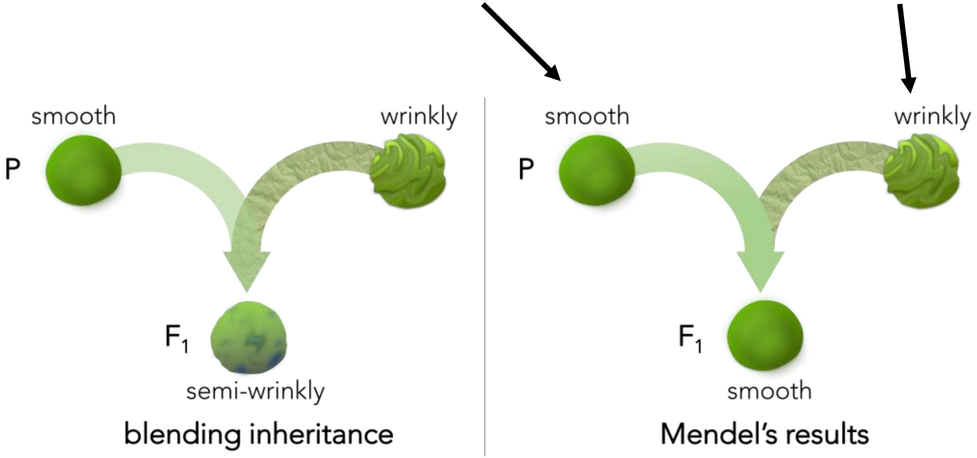
The Principle of Uniformity

- Before Mendel's experiments, most people believed that traits in offspring resulted from a blending of the traits of each parent

- However, when Mendel cross-pollinated one variety of purebred plant with another, these crosses would yield offspring that looked like either one of the parent plants, not a blend of the two

Dominant Trait

Recessive Trait



blending inheritance

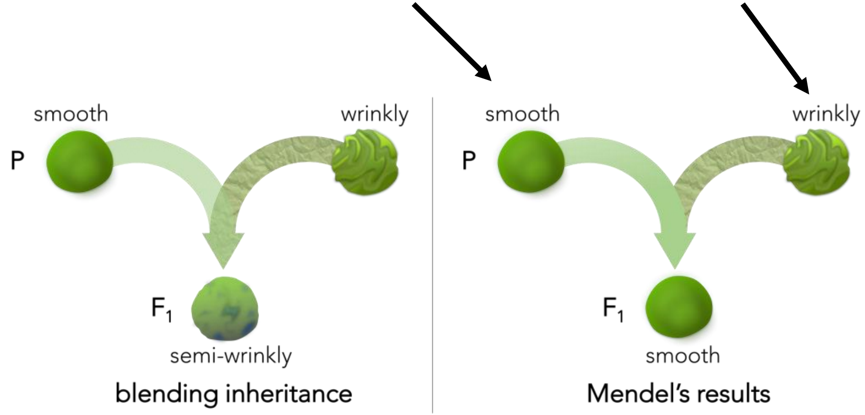
Mendel’s results

The Principle of Uniformity

- Pure Smooth + Pure Wrinkly = Smooth
- Dominant Trait V.S Recessive Trait
- A large sample size is required to make such conclusion (eight years of experiments)

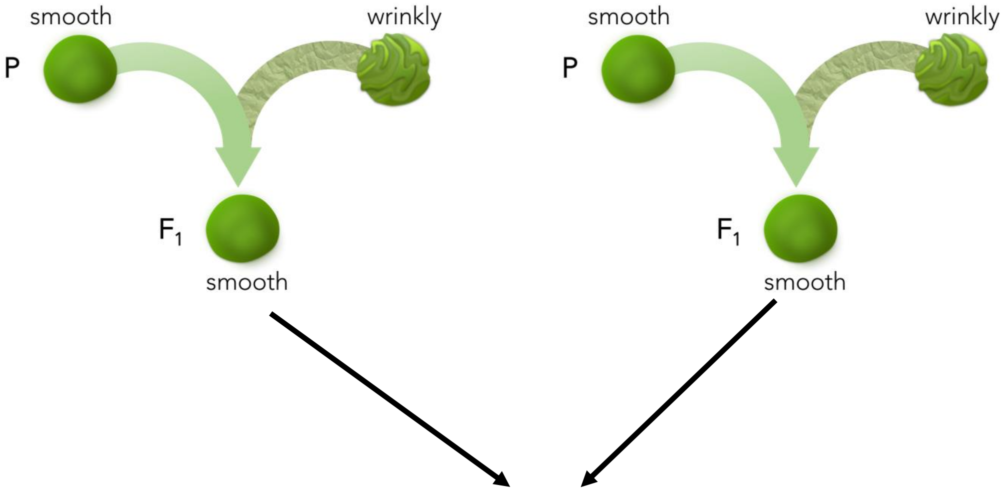
Dominant Trait

Recessive Trait



The Principle of Segregation

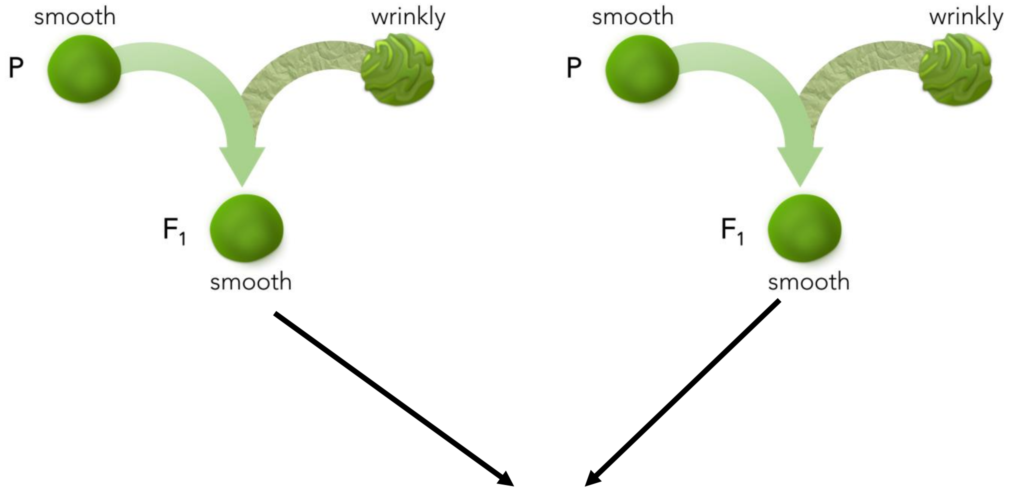
Whether recessive traits were hidden or eliminated?



Smooth? Wrinkly? Or Both?

The Principle of Segregation

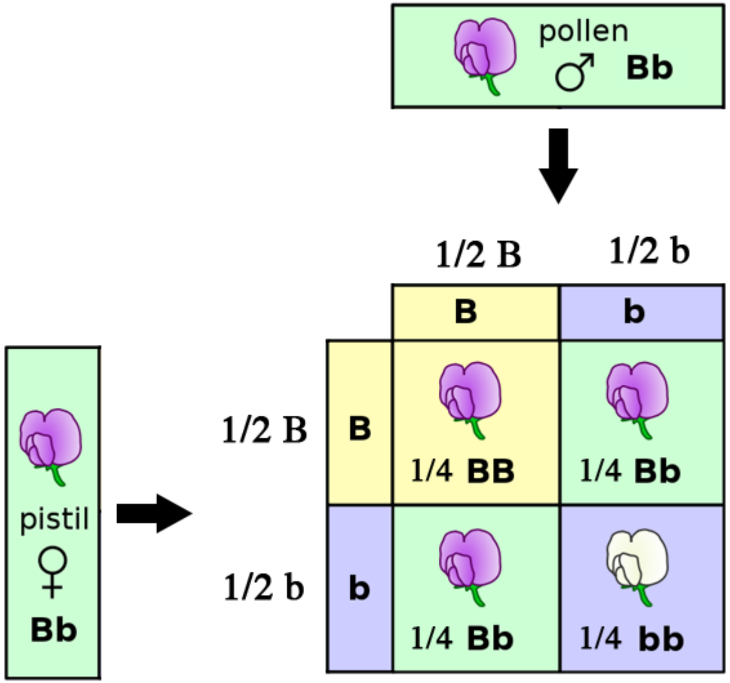
Whether recessive traits were hidden or eliminated?

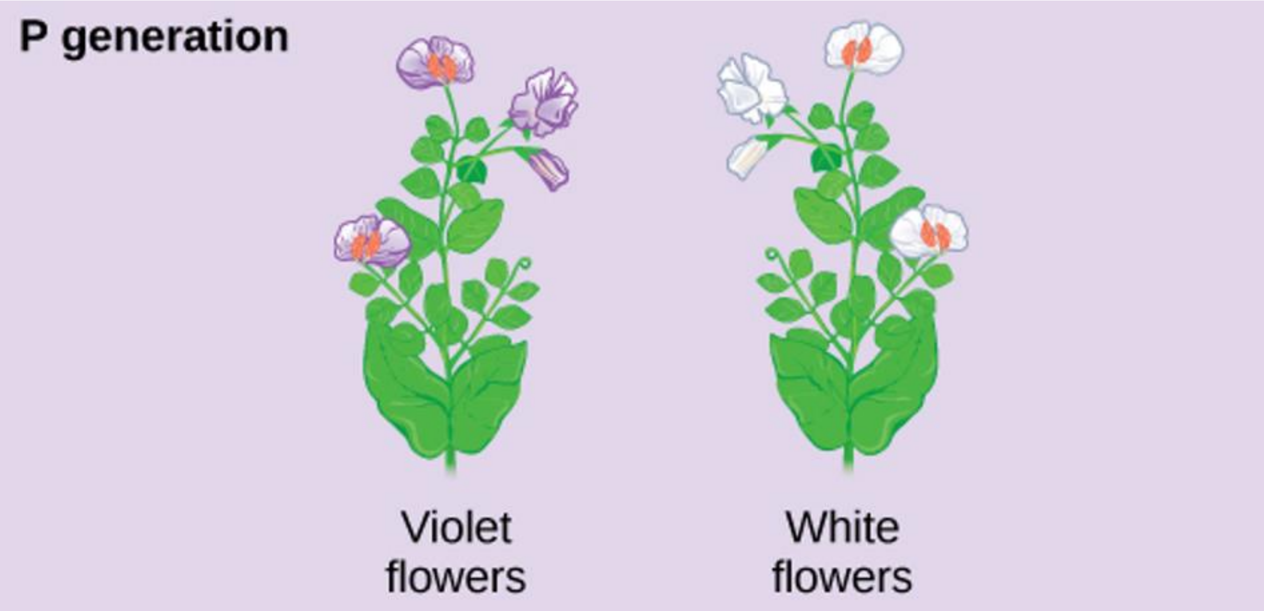


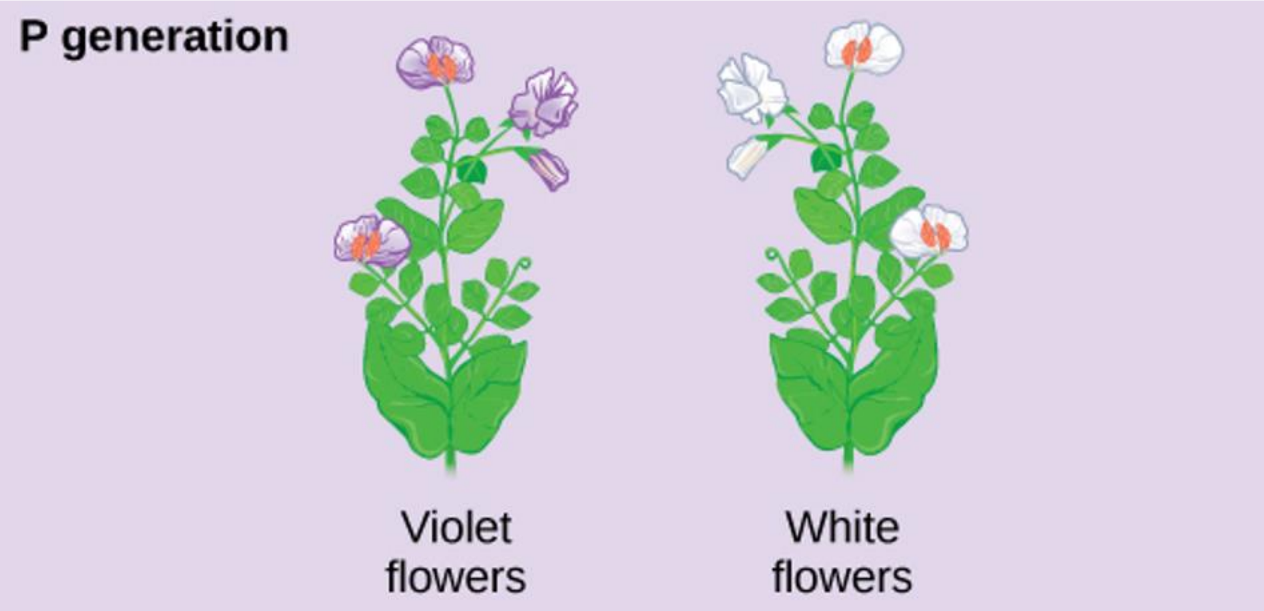
Smooth? Wrinkly? Or Both?

The Principle of Segregation

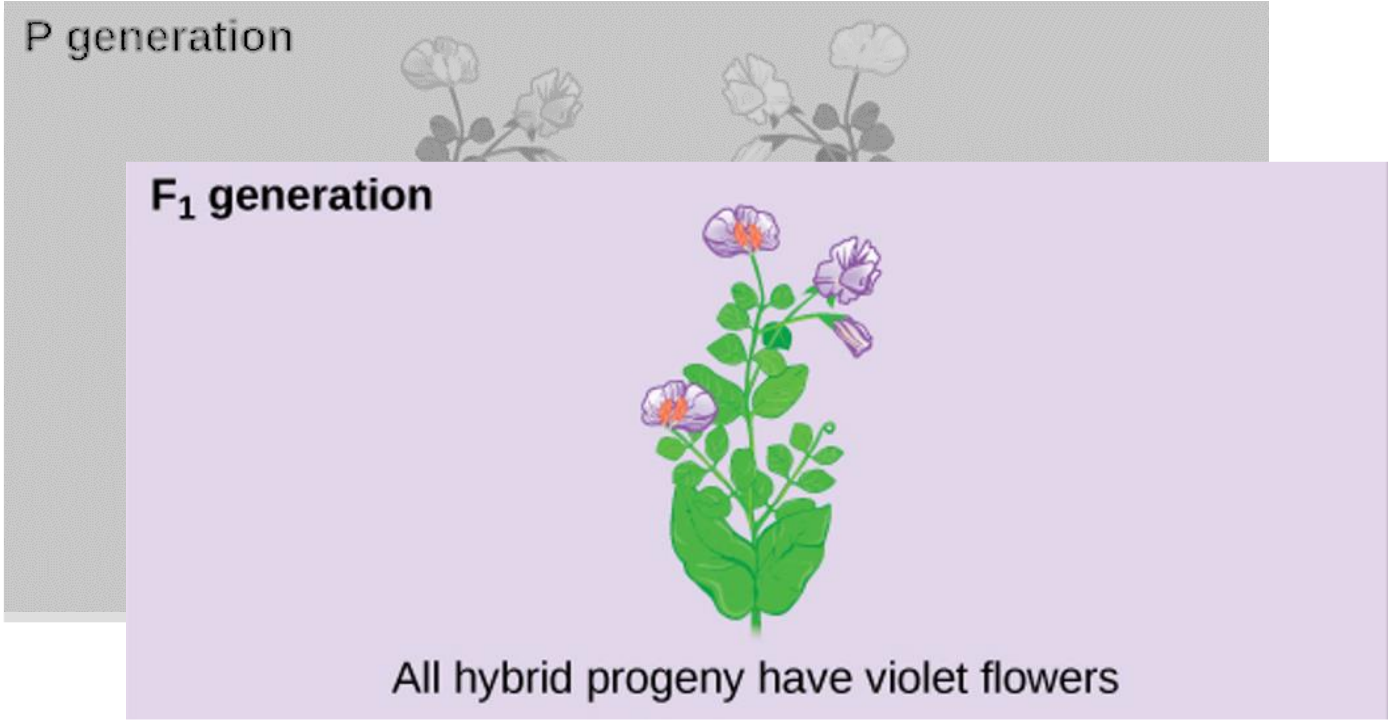
A Probability Model

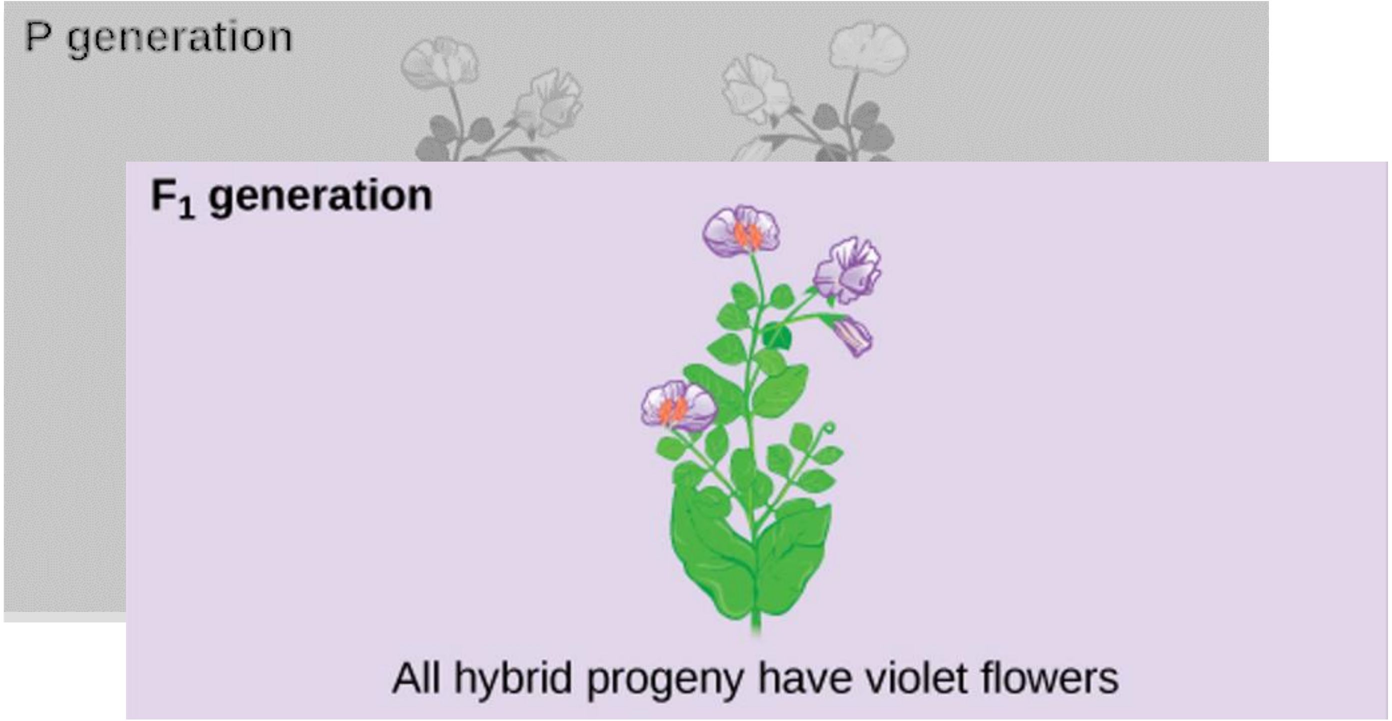



P generation
Violet flowers
White flowers

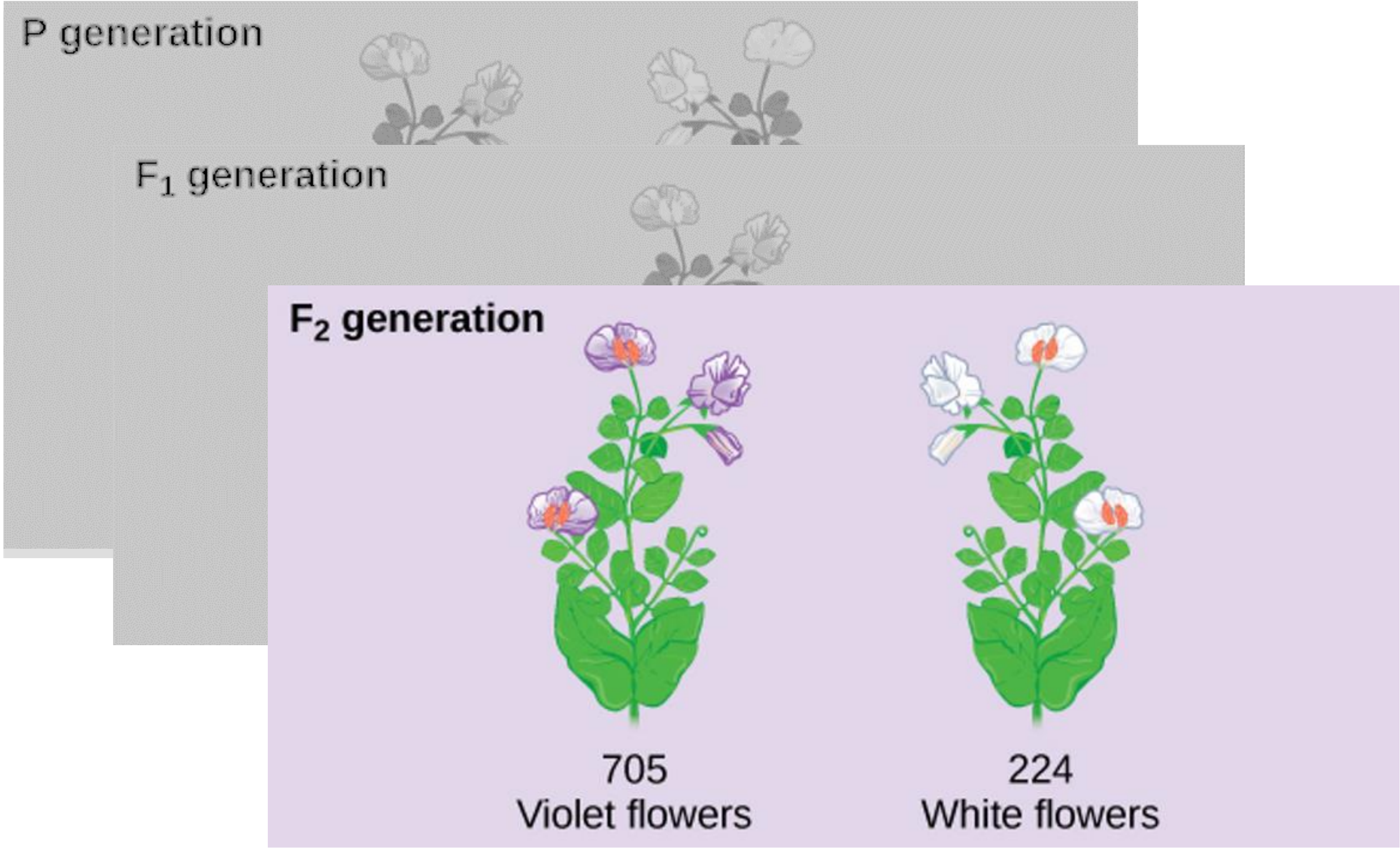


Hybridization of true-breeding plants

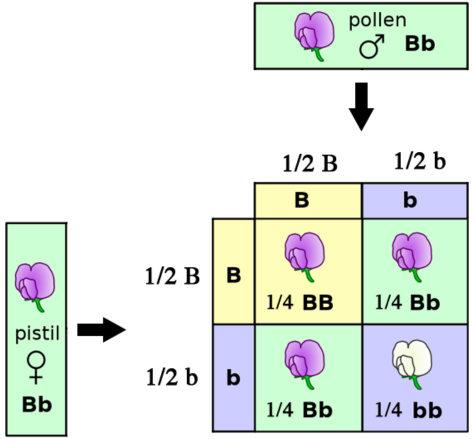

P generation
F1 generation
All hybrid progeny have violet flowers

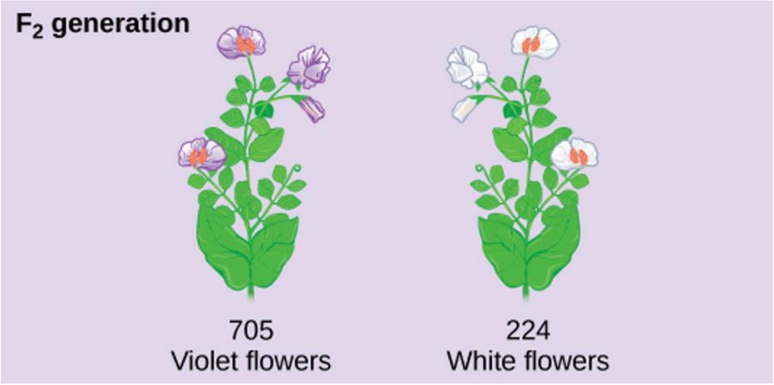


Self-fertilization of hybrid plants


P generation
F_1 generation
F_2 generation
705
Violet flowers
224
White flowers

- Statistics: Test Credibility
- Model Prediction: 75% of descendants are violet
- Experiment Data: 75.9% of descendants are violet



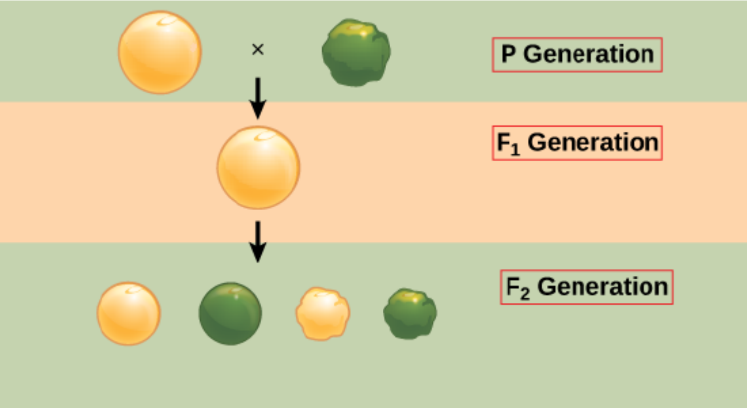


The Principle of Independent Assortment

- What happens when two plants that are each hybrid for two traits are crossed?
- Model 1: Two traits are dependent
- Model 2: Two traits are independent

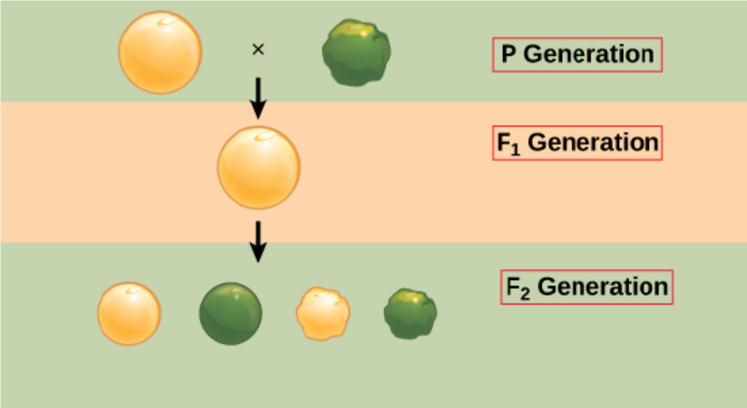
The Principle of Independent Assortment

Model 1



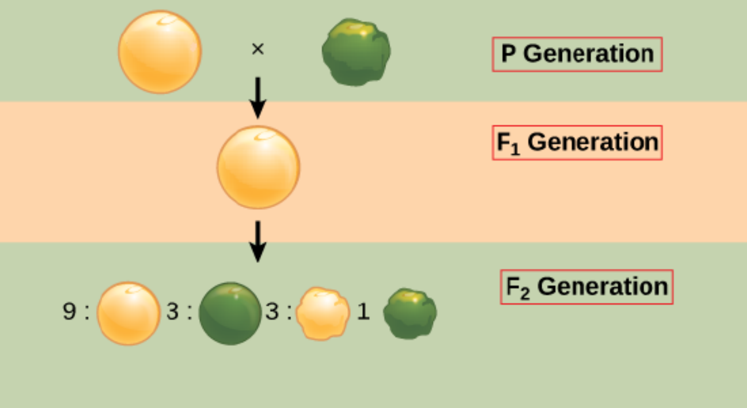
3 : 0 : 0 : 1

Model 2



9 : 3 : 3 : 1

The Principle of Independent Assortment



- 315 plants with round, yellow seeds
- 108 plants with round, green seeds
- 101 plants with wrinkled, yellow seeds
- 32 plants with wrinkled, green seeds

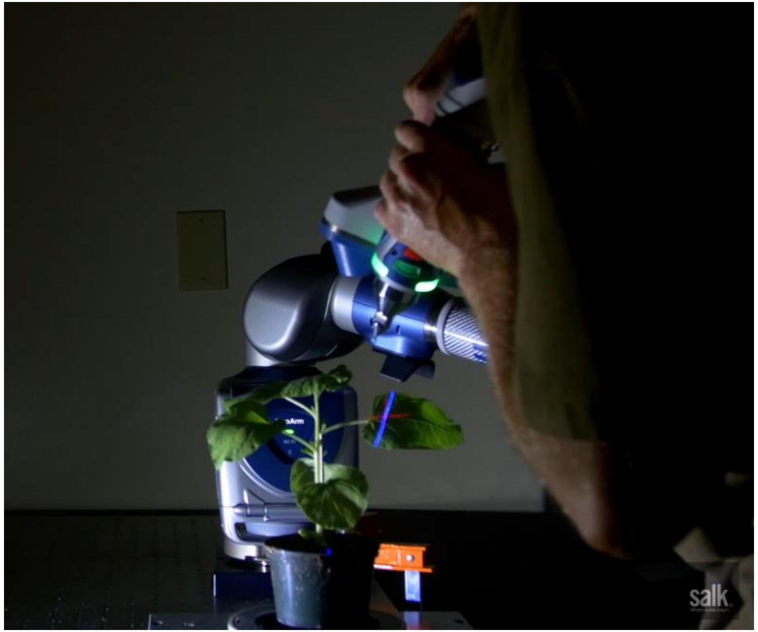
Gregor Mendel's Principles of Inheritance

- The principle of uniformity
  - Exceptions: Incomplete dominance
- The principle of segregation
  - Exceptions: Nondysjunction
- The principle of independent assortment
  - Exceptions: Linkage

- Gregor Mendel was luck that none of the exceptions occurred
- The most valuable part of his study resides in his data science methodology
- If he had access to today's technology, he could discover much more

Generated with Stable Diffusion 2.1.

- Gregor Mendel spent years tediously observing and measuring pea plant traits by hand
- Today, botanists can track the traits, or phenotypes, of hundreds or thousands of plants much more quickly, with automated camera systems
- Now, researchers have helped speed up plant phenotyping even more, with machine-learning algorithms that teach a computer system to analyze three-dimensional shapes of the branches and leaves of a plant

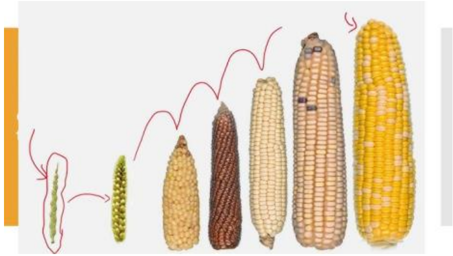


A Salk technician 3D scanning a plant

Breeding Strategies

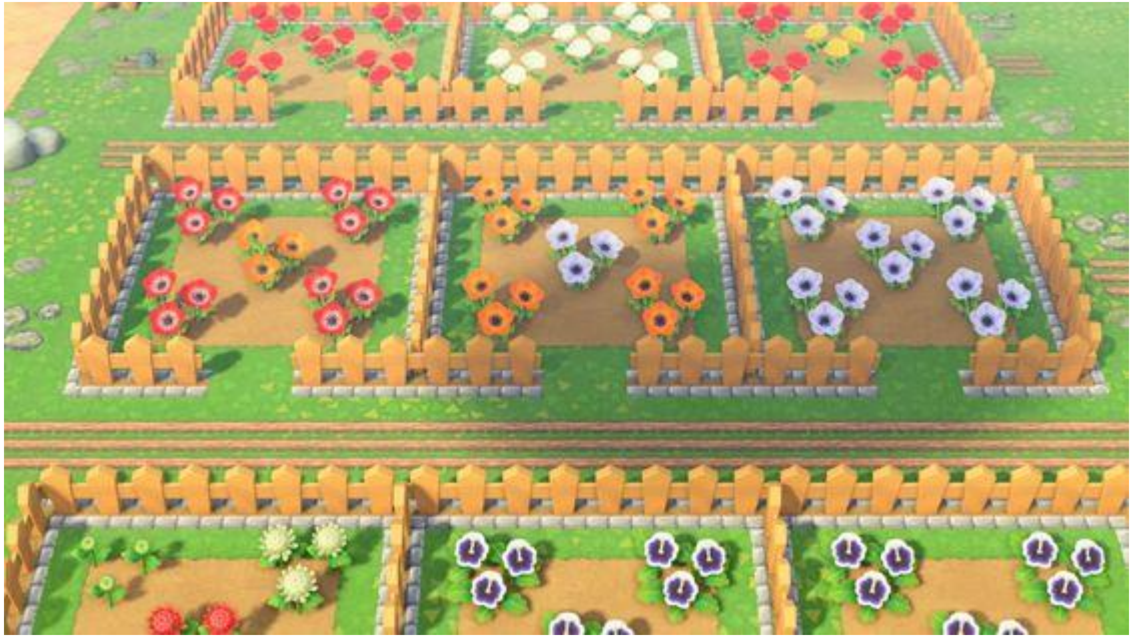
- Gregor Mendel's Principles of Inheritance paves the way for optimization in animal and plant breeding

- How to effectively deliver new varieties needed by farmers?



A Toy Example

The Optimal Garden Layout



The Optimal Garden Layout

- Flowers have a chance to reproduce once per day
- If a flower rolls "successful" at reproduction, it will then look a valid partner (same species) in the 3x3 space around itself to breed with, and then produce an offspring.
- If it does not find a valid partner, it will make a cloned offspring
- Breeding takes precedence over cloning whenever possible
- When a flower is successful, it (and its partner) are marked as invalid or "locked“ and cannot produce until the next day

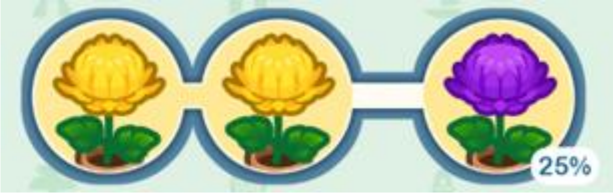
https://yuexr.github.io/acnh/garden_layouts

The Optimal Garden Layout

- Breeding is 2 flowers mixing genes to make 1 offspring. Cloning is 1 flower making 1 exact copy of itself

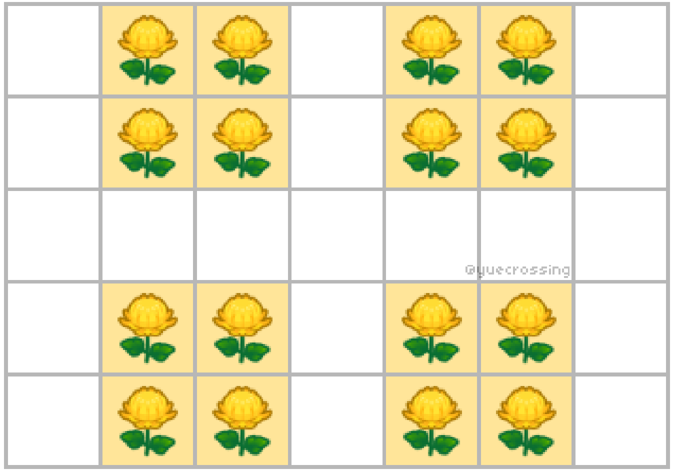
- Suppose you have a 5-by-7 garden and you would like to breed two yellow flowers to get violet flowers

- What is the optimal garden layout to produce violet flowers using yellow flowers?



The Optimal Garden Layout

- It is an integer program (very, very hard!)
- You will learn more about optimization algorithms to solve this kind of problem (MAT3007)
- An alternative approach would be to employ machine learning, deep learning, or neural networks to address this problem (DDA3020, DDA4210)

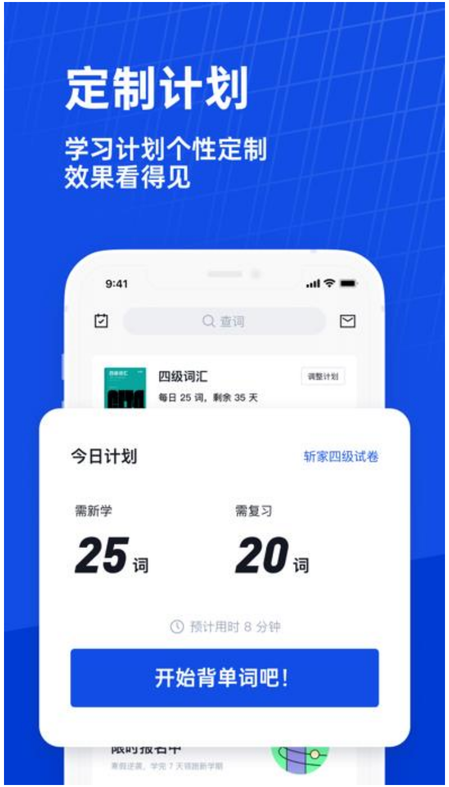


Example 3

Enhancing Student Learning Outcomes

Goal: Create a mobile app that enhances language learning for students—making the process *personalized, fast, enjoyable*!

Any ideas?



定制计划

学习计划个性定制
效果看得见

9:41

查词

四级词汇 调整计划
每日 25 词，剩余 35 天

今日计划 斩家四级试卷

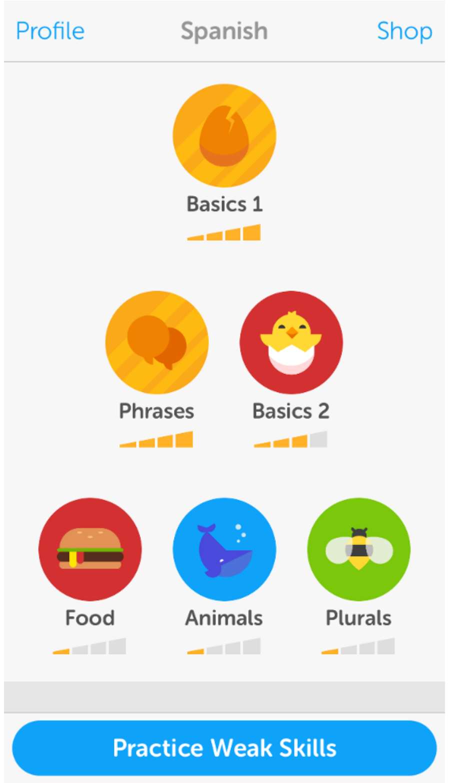
需新学 25 词

需复习 20 词

预计用时 8 分钟

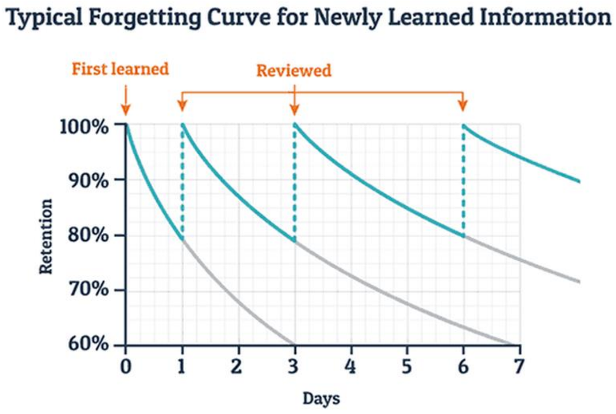
开始背单词吧！

限时报名中



How a person forgets newly learned information?

This is known as the *forgetting curve* in psychology.



Observation 1: The forgetting curve jumps when reviewing but decays afterwards

How to model the forgetting curve?

We can use, for example, an exponential model to describe the **probability** of recalling a word:

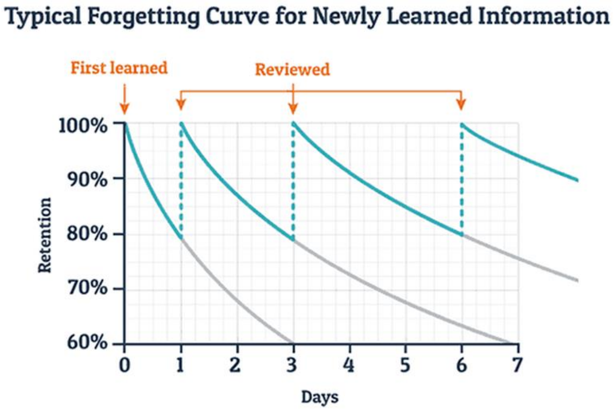
$$\mathbb{P}\{q(t) = 1\} = \exp(-r(t) \cdot (t - t_{last}))$$

where:

- $q(t)$: a binary variable indicating if the user recall the word at time t
- t_{last}: the time of last review
- $r(t) \geq 0$: forgetting rate, which itself depends on many factors. For example,

  $$r(t) = r_0 - a \cdot n(t)$$

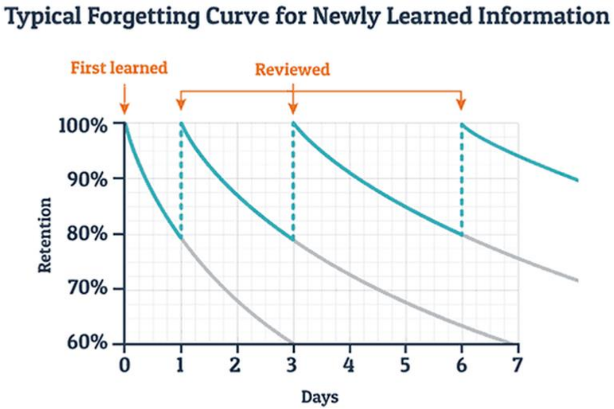
  - Here $n(t)$ represents the number of repetitions, and r_0, a vary among users



How to learn the forgetting curve from data?

$$\mathbb{P}\{q(t) = 1\} = \exp(-r_0 - a \cdot n(t) \cdot (t - t_{last}))$$

- Samples of words
- Records of last time that the user review the word t_{last} and the number of repetitions n
- Records that whether the user recalls the word at a later time t
- Using **logistic regression**, we can estimate r_0 and a!
  - This enables *personalization*!



What if the forgetting curve is more complex?

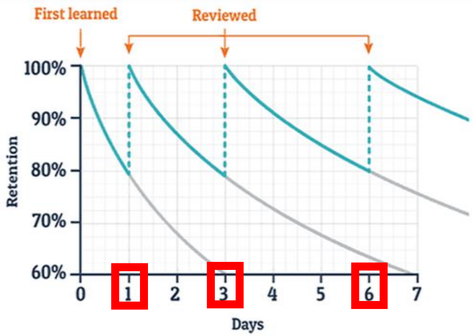
- Recall

$$\mathbb{P}\{q(t) = 1\} = \exp(-r(t) \cdot (t - t_{last}))$$

- The forgetting rate function $r(t)$ can depend on lots of factors with complicated probability models:
  - Language: English, French, Spanish,…
  - Words' difficulty level: CET4/6, TOEFL/IELTS, GRE/GMAT/LSAT,…
  - Time for reviewing: early morning, late night,…
  - User background: students, working class, children,…
  - ……
- With modern **machine learning**, we can take account of all these factors and learn complex functional relationships, e.g., Neural Hawkes process

Model trade-off

- Observation 2: Reviewing material at intervals over time enhances long-term memory



  - “Fast”: To avoid forgetting, we want to flatten the curve as much as we can → increase the number of repetitions
  - “Enjoyable”: the user would be painful if repeating the same task too many times → decrease the number of repetitions

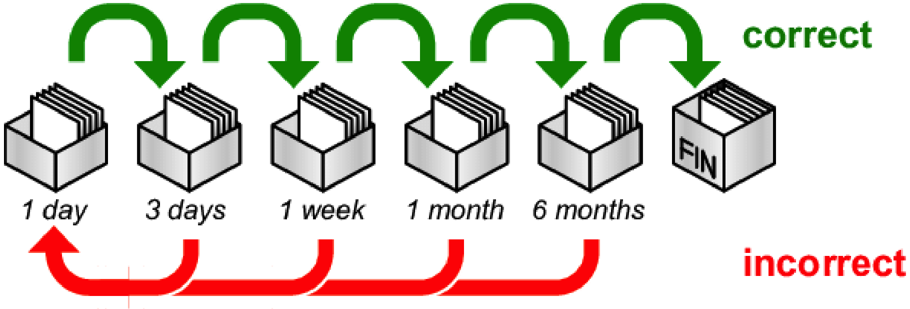
How do we find the *optimal* time for reviewing?

- We need an **optimization** model!
- *Decision*: the time points that remind the user to review
- *Objective*: Maximizes the probability of memorizing words while minimizes pain of repetition
- *Constraints*: Duration of study, maximum number of repetitions, etc.

- Solving this mathematical optimization yields the optimal decision!

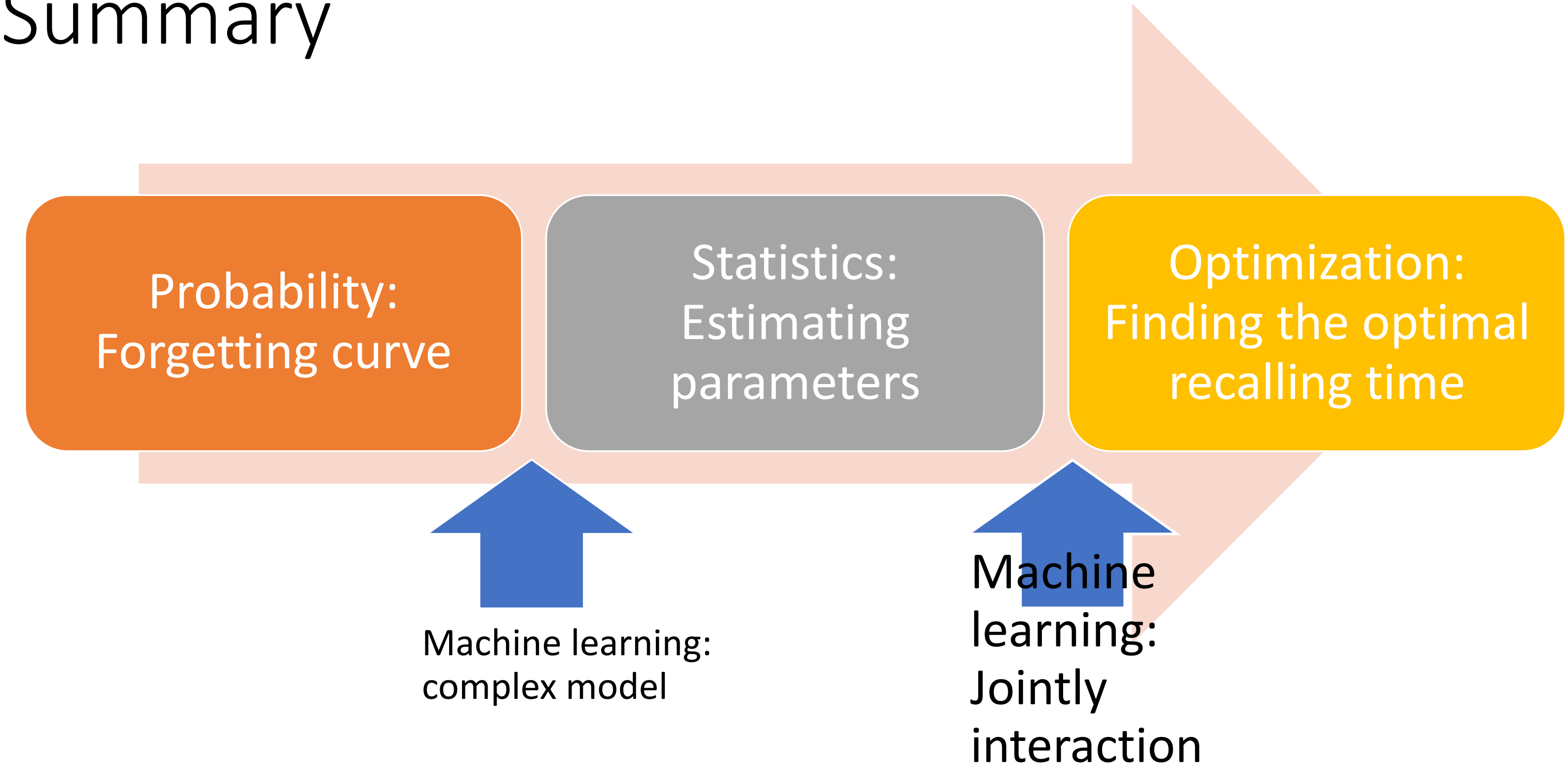
Learning and decision-making on the fly

- What if we do not have data for a new user?



- **Reinforcement learning** designs “smart” interactions between the APP and the user, so that it can jointly learn the user’s forgetting curve while optimize the learning outcomes

Summary



Probability: Forgetting curve

Statistics: Estimating parameters

Optimization: Finding the optimal recalling time

Machine learning: complex model

Machine learning: Jointly interaction