

# Welcome to STAT 547C, *Topics in Probability*

Instructor: Alexandre Bouchard  
Fall 2014

# Plan for today:

- Logistics.
- Why you should care about probability.
- First definitions.

# Logistics

# Contact & other logistic issues

- Web site: always check first!  
<http://www.stat.ubc.ca/~bouchard/courses/stat547c-fa2013-14/index.html>
- Textbook and other readings
- Hints and updates for assignments
- Office Hours: TBA, fill Doodle
- Contact:
  1. Piazza
  2. [bouchard@stat.ubc.ca](mailto:bouchard@stat.ubc.ca)

# Teaching assistant

- Seong-Hwan Jun
- Contact email:  
[s2jun.uw@gmail.com](mailto:s2jun.uw@gmail.com)
- Office hours: TBA, fill Doodle

# Homeworks (40%)

- Four main assignments (25%)
- ‘Exercises/Participation’ (15%): doing the short exercises, interacting in class, coming at office hours

# Exams

- In class midterm (20%)
- Finals:
  - Last day of class: 'Essay: what I have learned in the course' (10%)
  - Take home final/project (30%)

# Why this topic is important

- Fundamental tool in statistics and computer science
- Probability is arguably the best tool to model reality
- Computational power of randomness
- New foundations of Science ('Dawning of the age of stochasticity', D. Mumford)
- A very good investment for researchers



# Killer apps

- Probabilistic machine learning: Siri, autonomous cars, machine translation
- Physics: Determining the shape and ultimate fate of the Universe ('random fields'), quantum, stat. mech.
- Biology: Tree of life, DNA testing, folding@home
- Engineering: Compression for deep space communication ('LDPC'), design of polymers, Google
- But also: atomic bomb ('MC approximations'), financial crisis

# Core topics

- Applications of probability in statistics
- Formal treatment of the probability spaces and expectation and their properties (the language of probability)
- The ‘surprising challenges of composing r.v.s’
  - Asymptotics
  - Generating functions
- Conditioning
- Going beyond independence: Markov chains

# Additional potential topics

- Poisson processes
- Martingales
- Continuous time Markov chains

# Some highlights and examples of applications

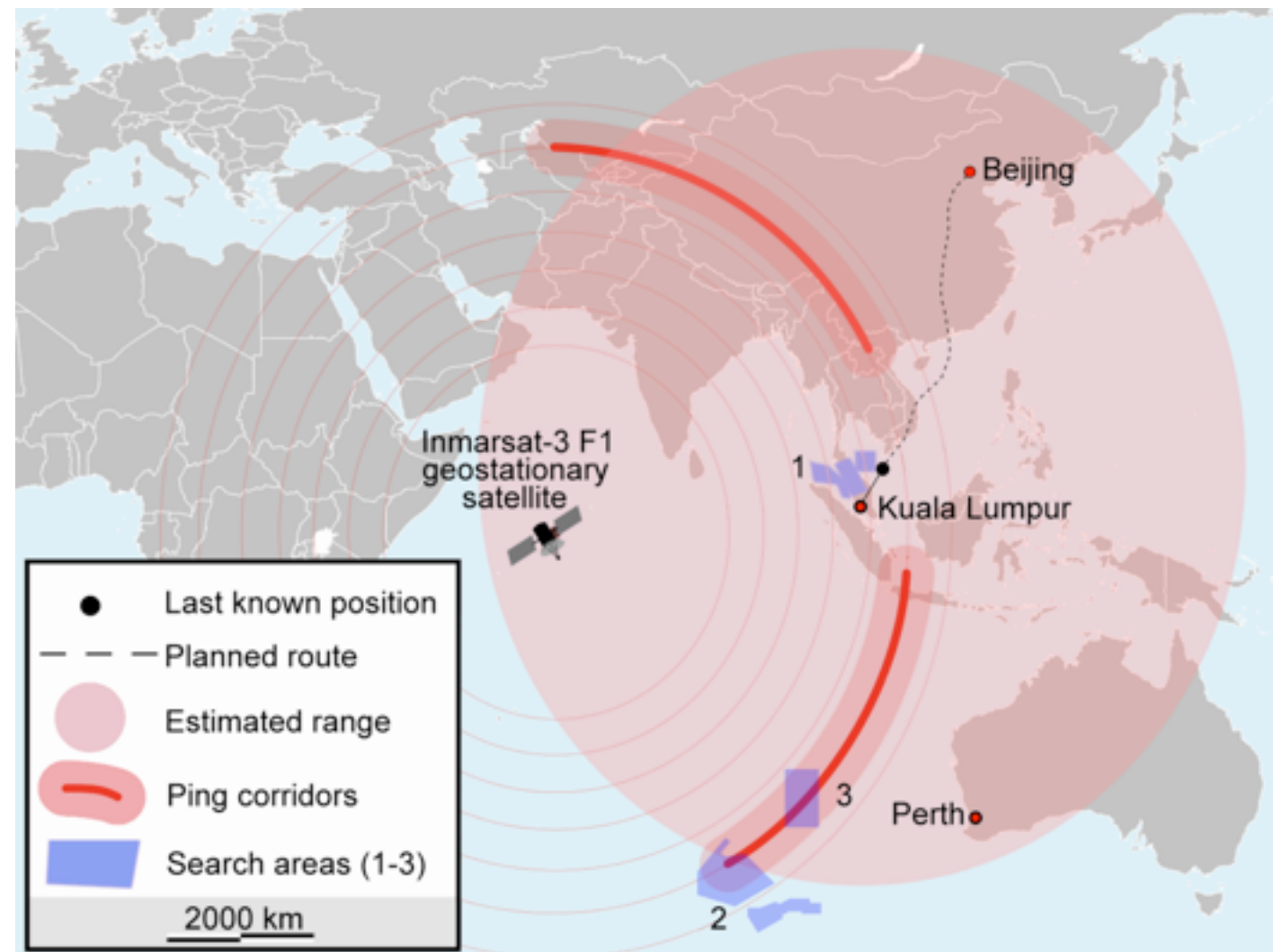
# The Search for Malaysia Airlines Flight 370

**Goal:** finding the location of the crash

**Question:** how to prioritize search

**How to reconcile several sources of partial info:**

- Last known position
- Fuel range
- Last satellite ping



<http://tinyurl.com/lhzrufa>

# Bayesian Search



1966: Palomares B-52 crash

Photo # NH 97221-KN Stern section of sunken USS Scorpion, 1986



1968: USS Scorpion  
disappearance



# Conditioning

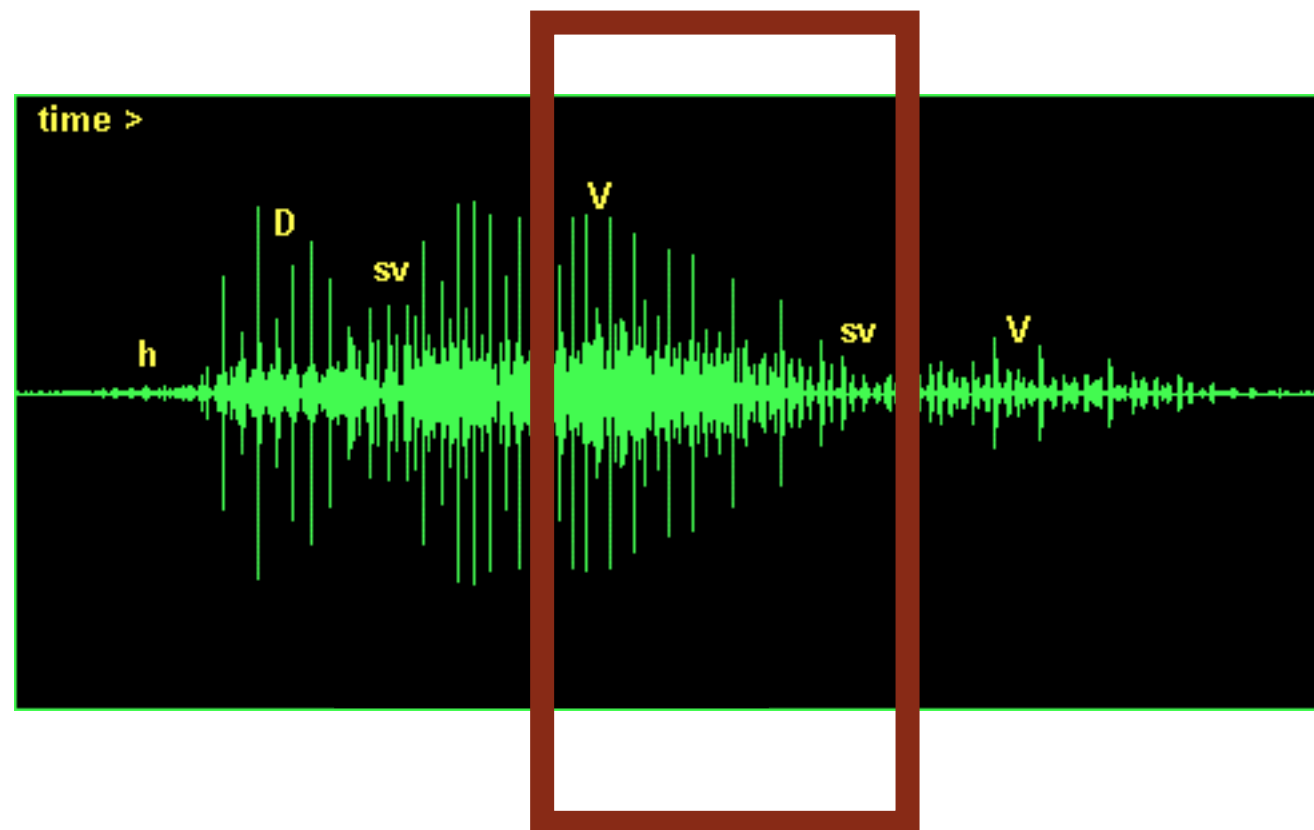
- Say you search in the square of highest success probability
- You find nothing
- What should you do next?
  - Note: even if the submarine is there, you might have missed it!
- Probability as a calculus of belief and uncertainty



Bayes theorem  
(Thomas Bayes),  
1763

<http://tinyurl.com/pcznhtml>

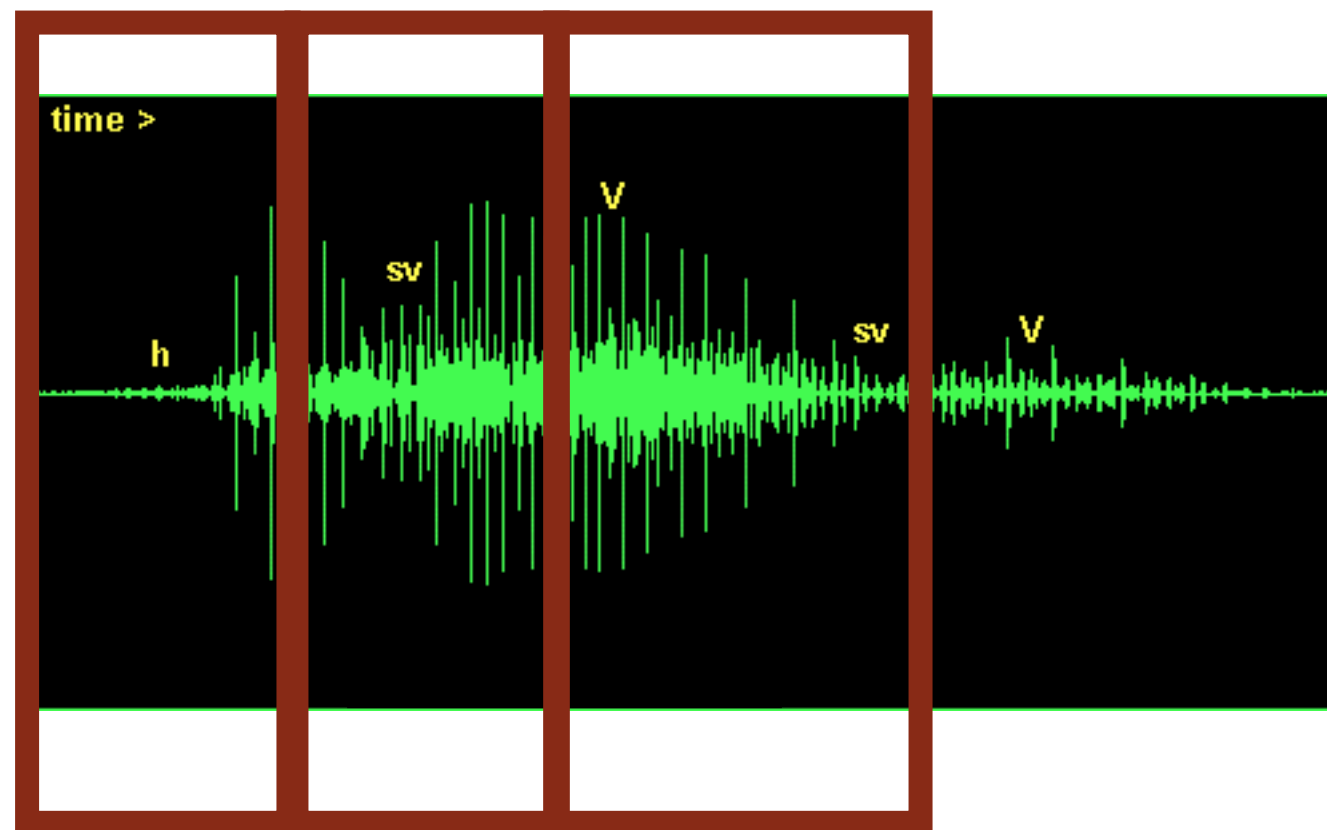
# *AI and machine learning:* Speech recognition



???



# *AI and machine learning:* Speech recognition



How are ???

# Rational behavior and uncertainty

**General question:** how to **act** when

- we are facing uncertainty
- errors have different costs

**Examples:**

- fraud detection
- medical diagnosis
- spam classifiers

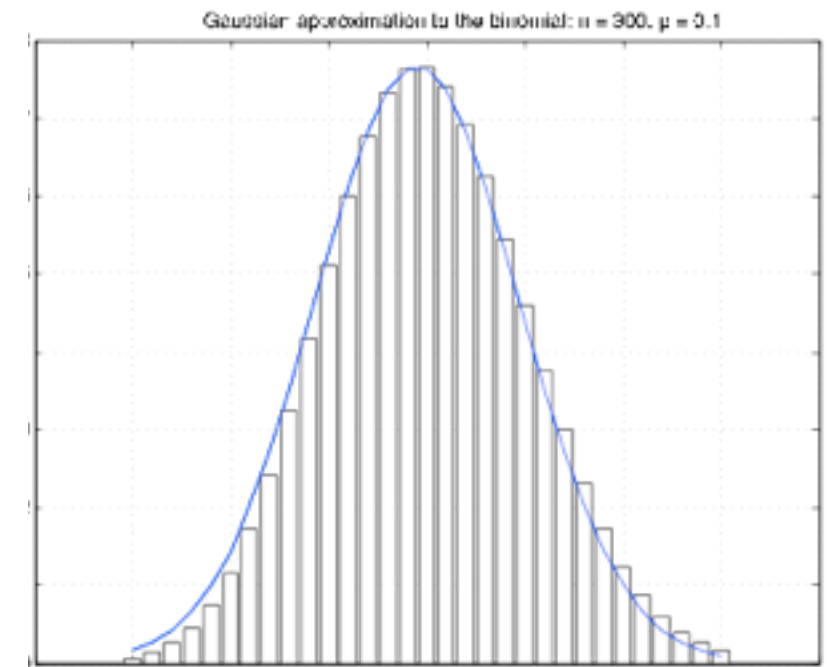
**Key tool:** *expected value*

# Surprising challenge

- Sums of random variables
  - Omnipresent in statistics
  - Taking the sum of variables is easy, so taking the sum of *random* variables should also be easy, right?
  - Not quite... consider for example the problem of computing the probability that the sum of 1000 coins is greater than 500.
    - Would have been hard in the pre-computer era
    - Generalized versions of this problem still hard with computer

# Limiting theory to the rescue

- Another surprise: sums of random variables can be approximated by something simple when large number of terms involved
- No matter what each  $X$  is!!! (almost)
- Also explains why we spend disproportionate amount of time on some specific types of random variables (normal, Poisson, ...)



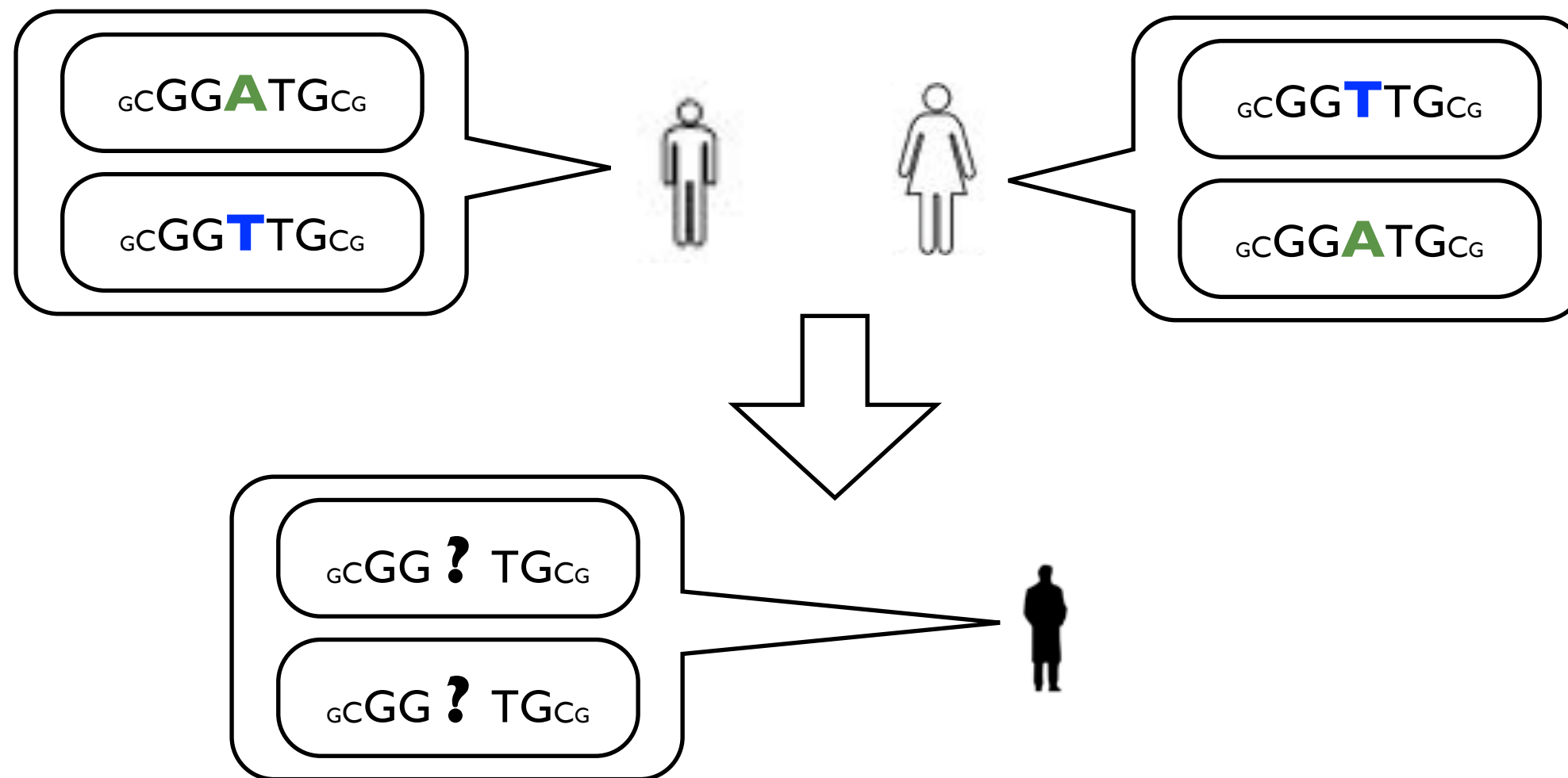
300 coins

# Beyond independence

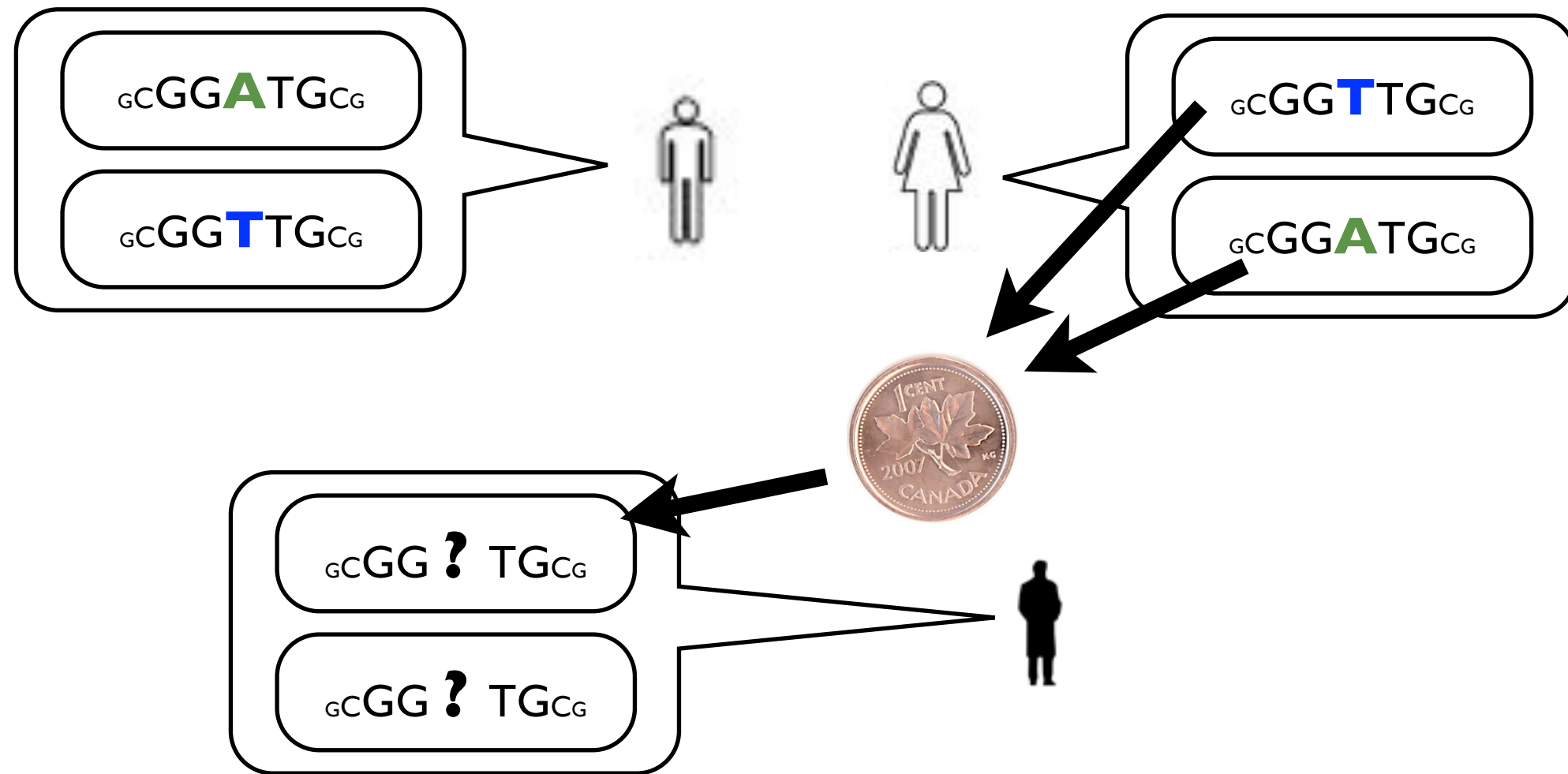
- Markov chain: independence of past and future given present
- Martingale: gambling model

# Genetics: inheritable diseases

## Randomness in inheritance :

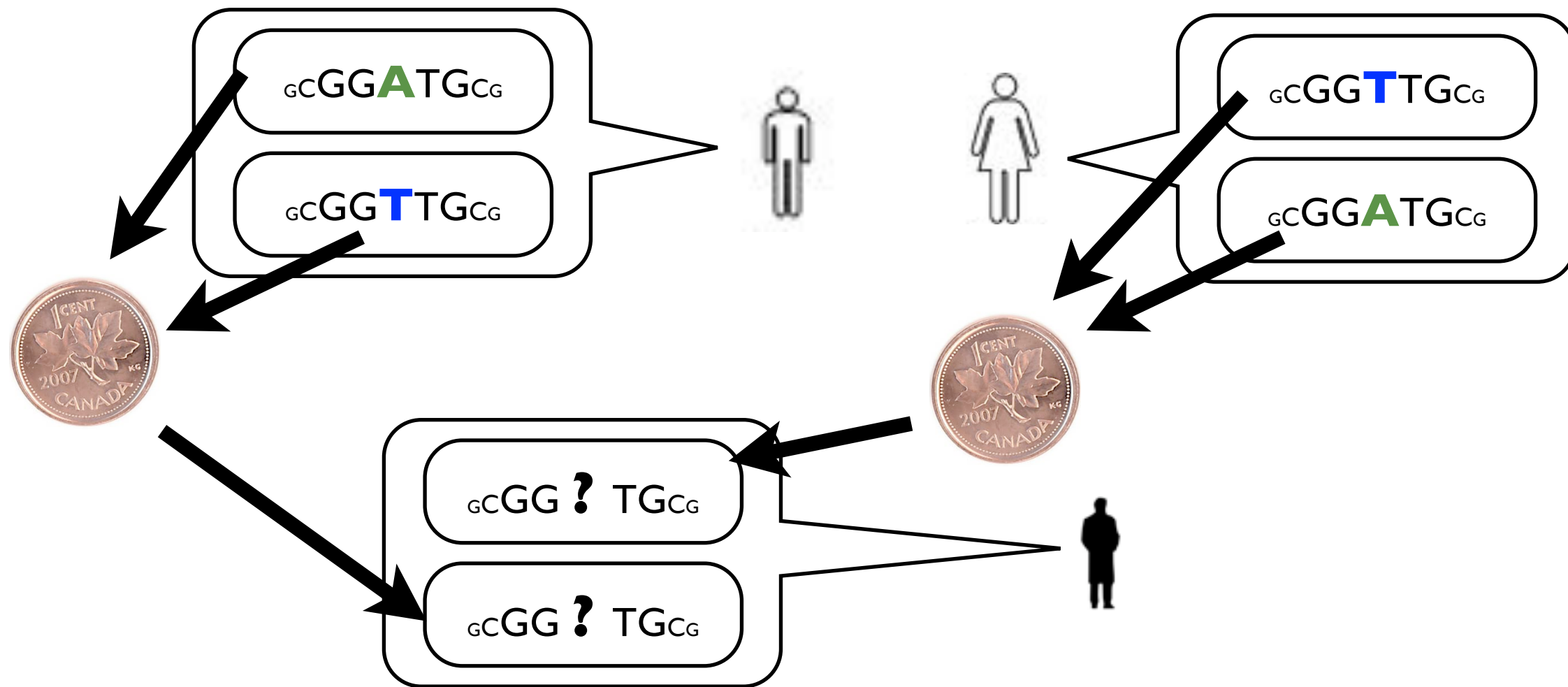


# Genotype inheritance



I) Flip a fair coin to decide if you inherit mom's T or A

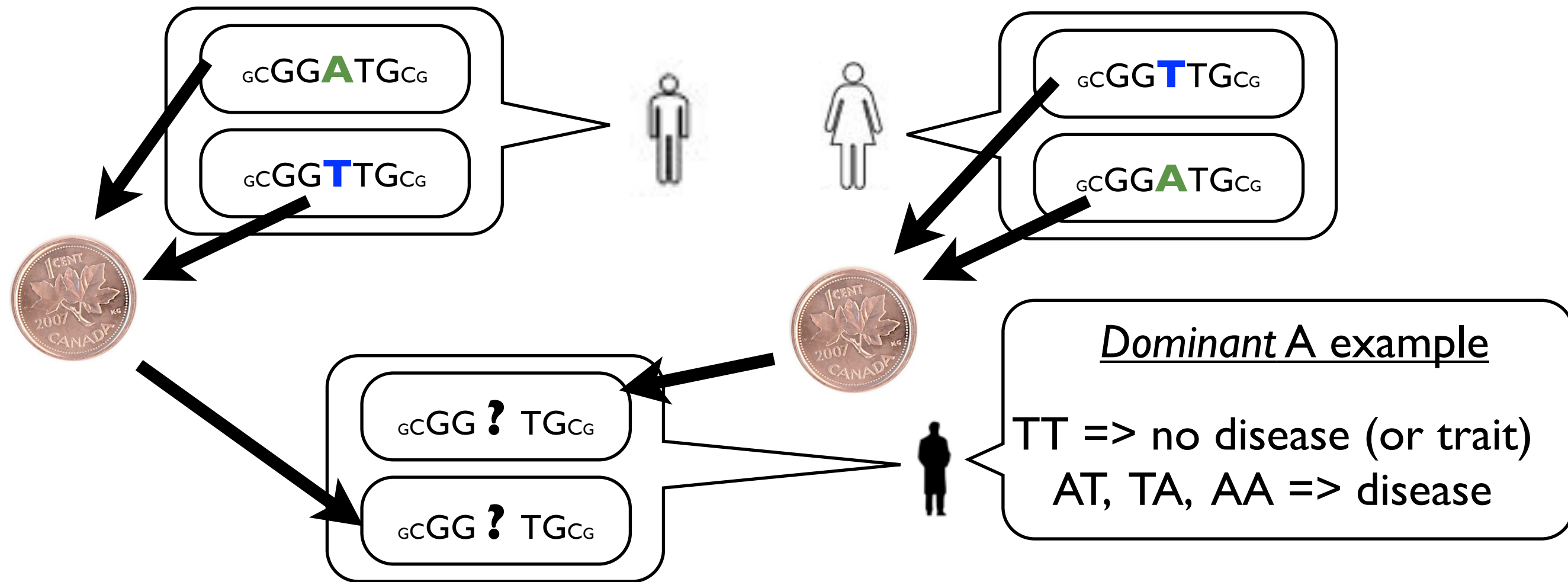
# Genotype inheritance



- 1) Flip a fair coin to decide if you inherit mom's T or A
- 2) Flip another fair coin to decide if you inherit dad's T or A



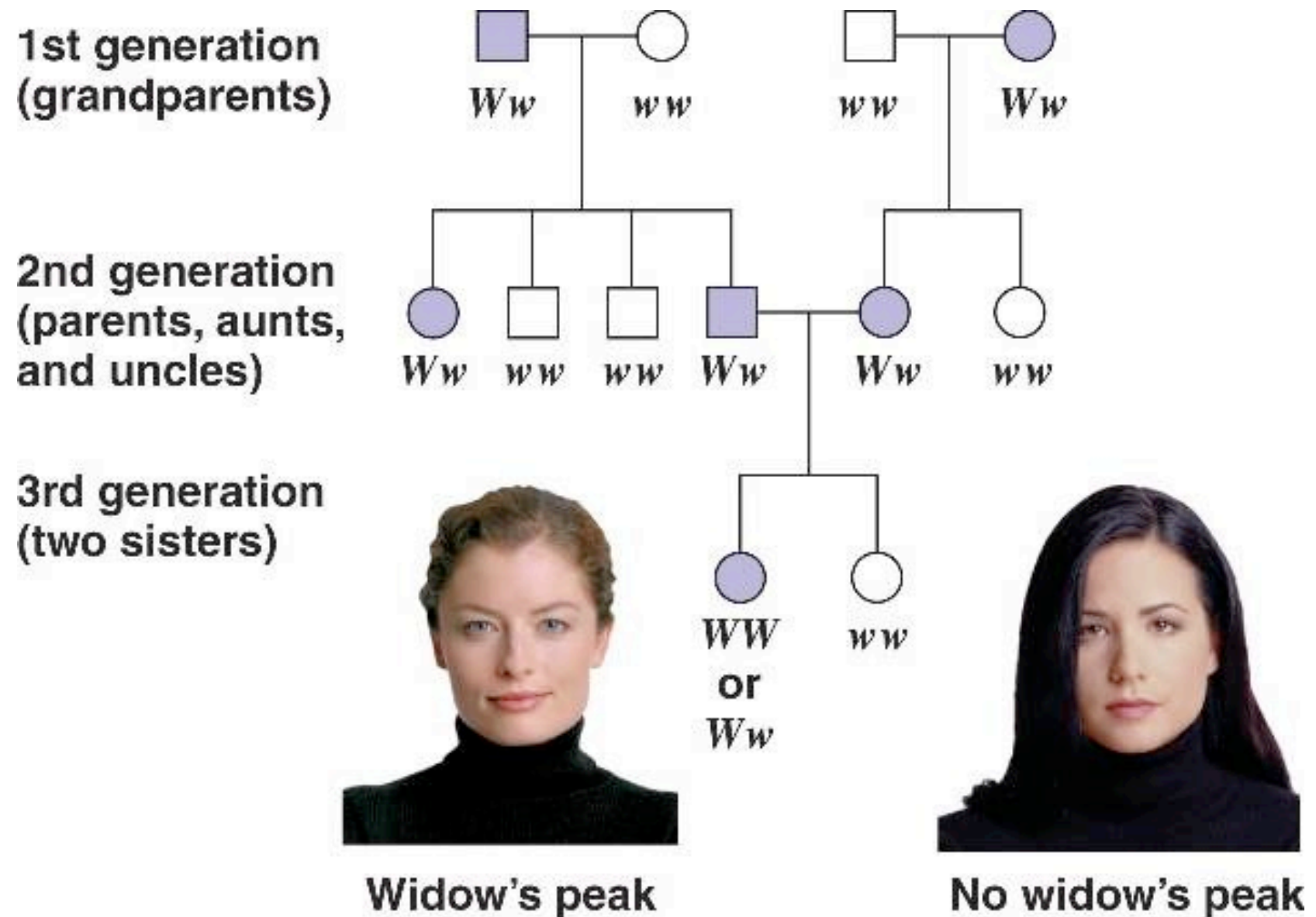
# Genotype inheritance



- 1) Flip a fair coin to decide if you inherit mom's T or A
- 2) Flip another fair coin to decide if you inherit dad's T or A

# Larger family trees

- A larger example where  $W$  is dominant over  $w$
- Goals:
  - genetic counseling
  - finding genetic factor of diseases / traits
- Complication factor
  - incomplete data



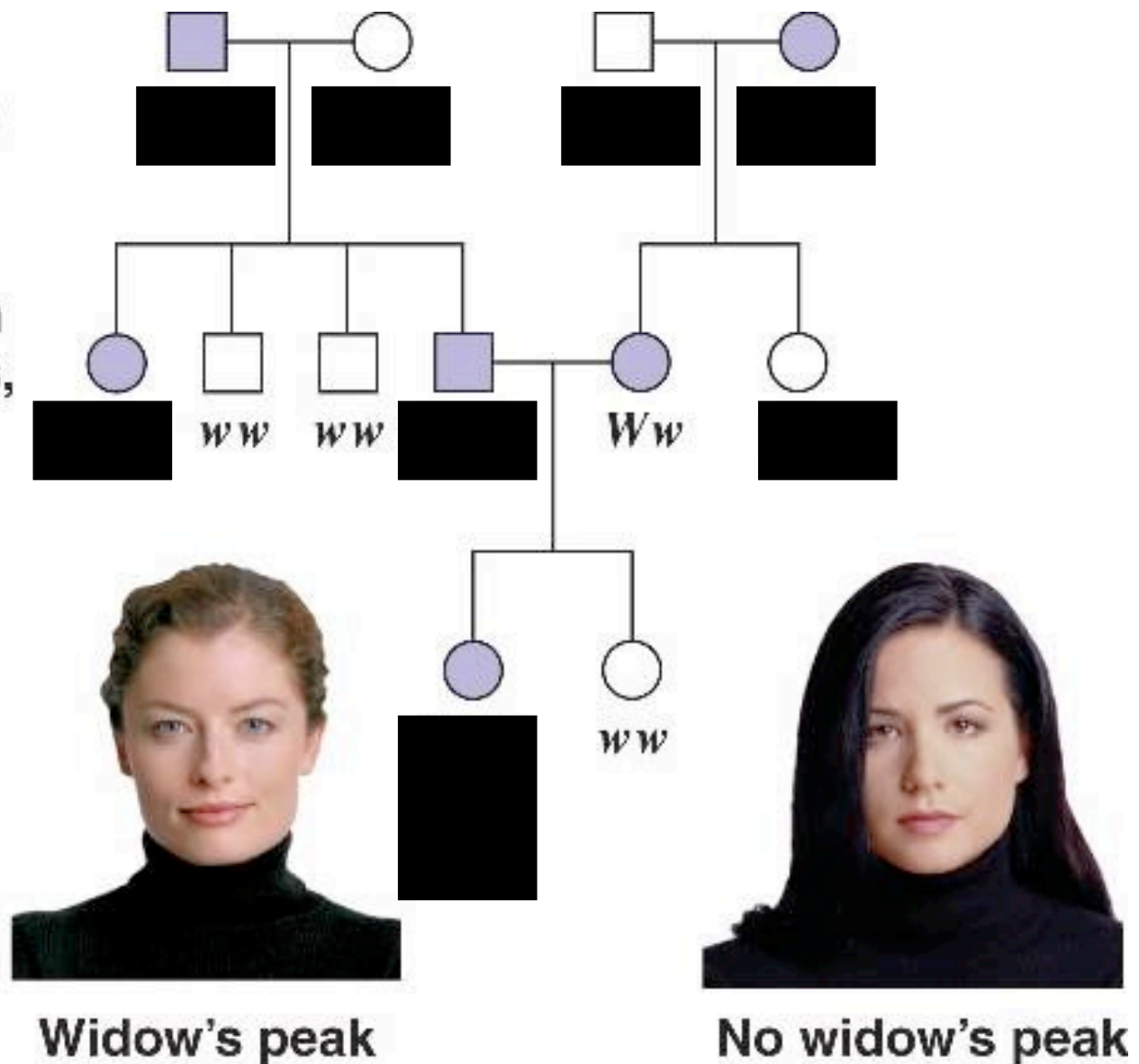
# Larger family trees

- A larger example where  $W$  is dominant over  $w$
- Goals:
  - genetic counseling
  - finding genetic factor of diseases / traits
- Complication factor
  - incomplete data

1st generation  
(grandparents)

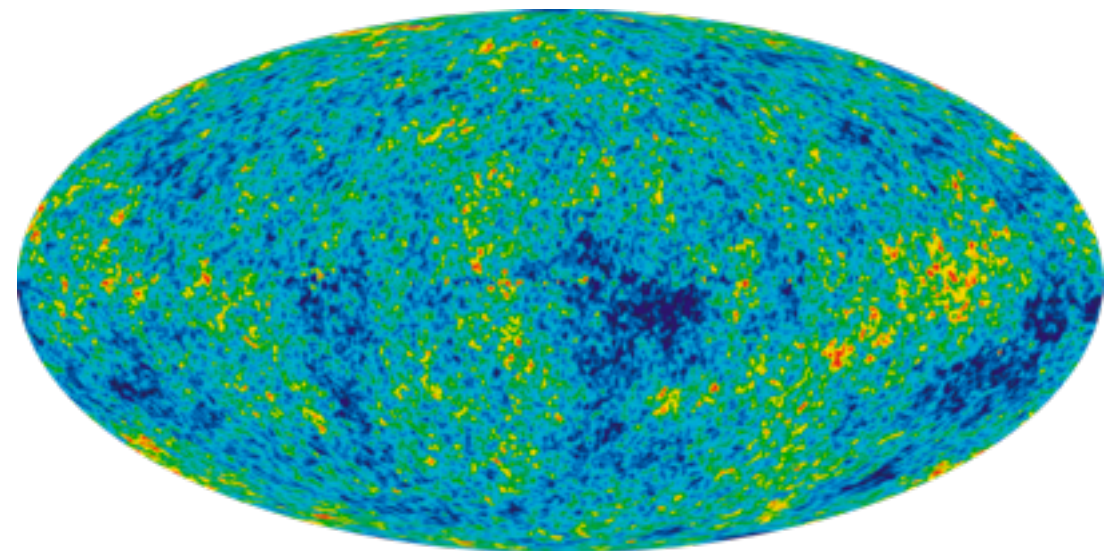
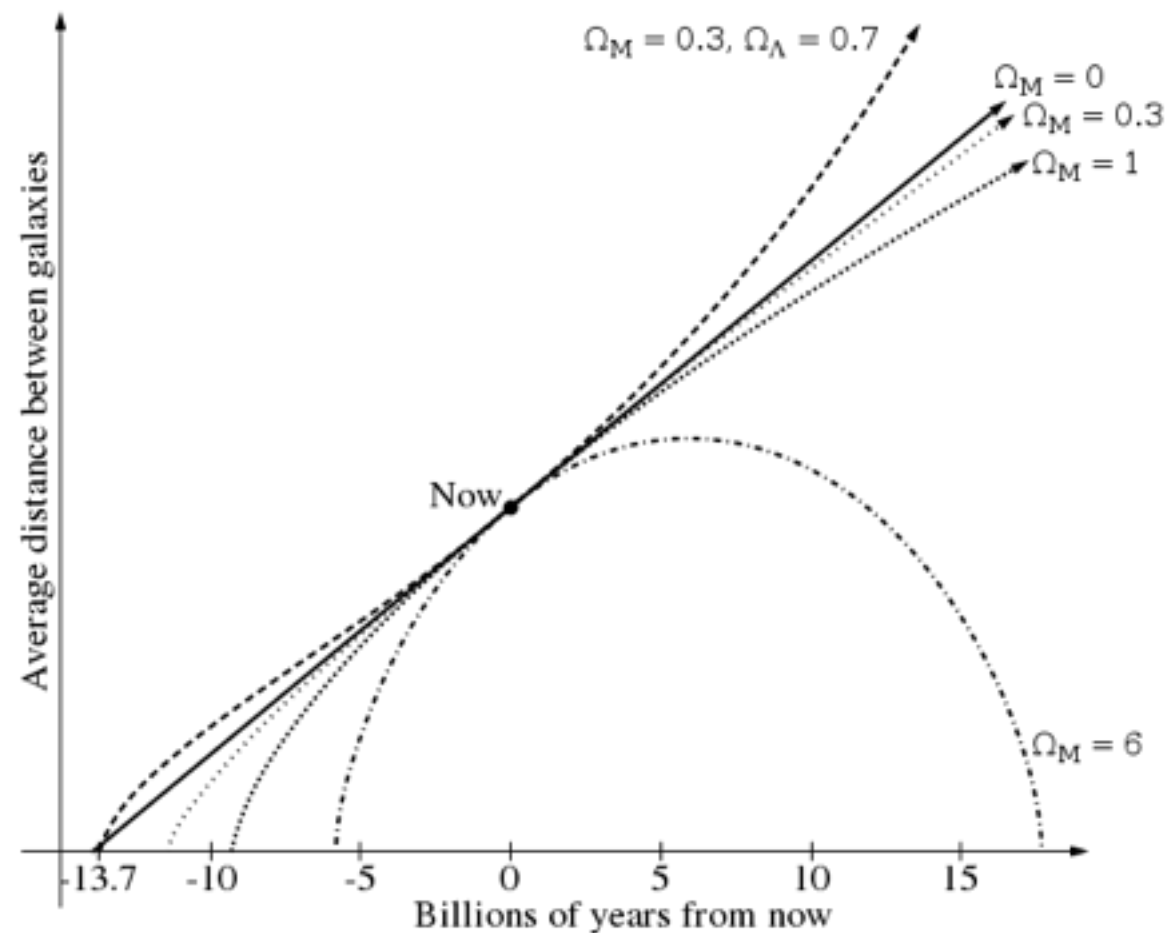
2nd generation  
(parents, aunts,  
and uncles)

3rd generation  
(two sisters)



# Astrophysics: Estimating the age and faith of the Universe

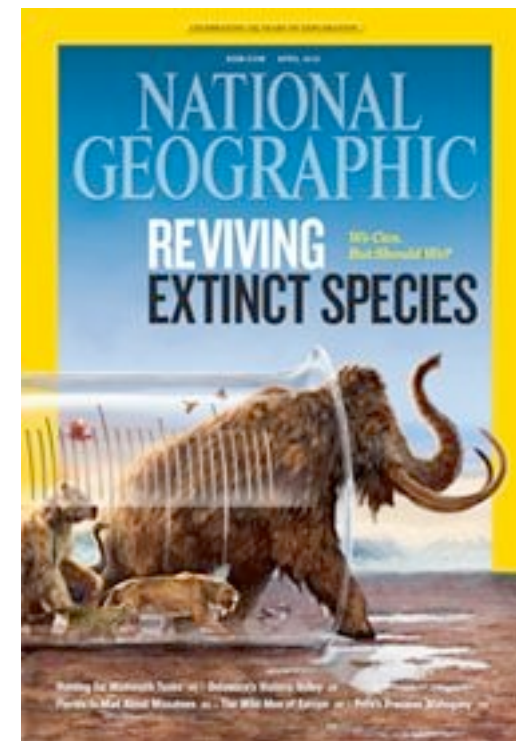
- **Goals:** finding the Universe's
  - age
  - density ( $\Rightarrow$  faith)
- **Data:** Cosmic Microwave Background (CMB): remnants of Big Bang
  - Detailed map from the Planck satellite
- Age, Physical constants  $\Rightarrow$  known *distribution* on CMP
- Invert using Bayes' rule





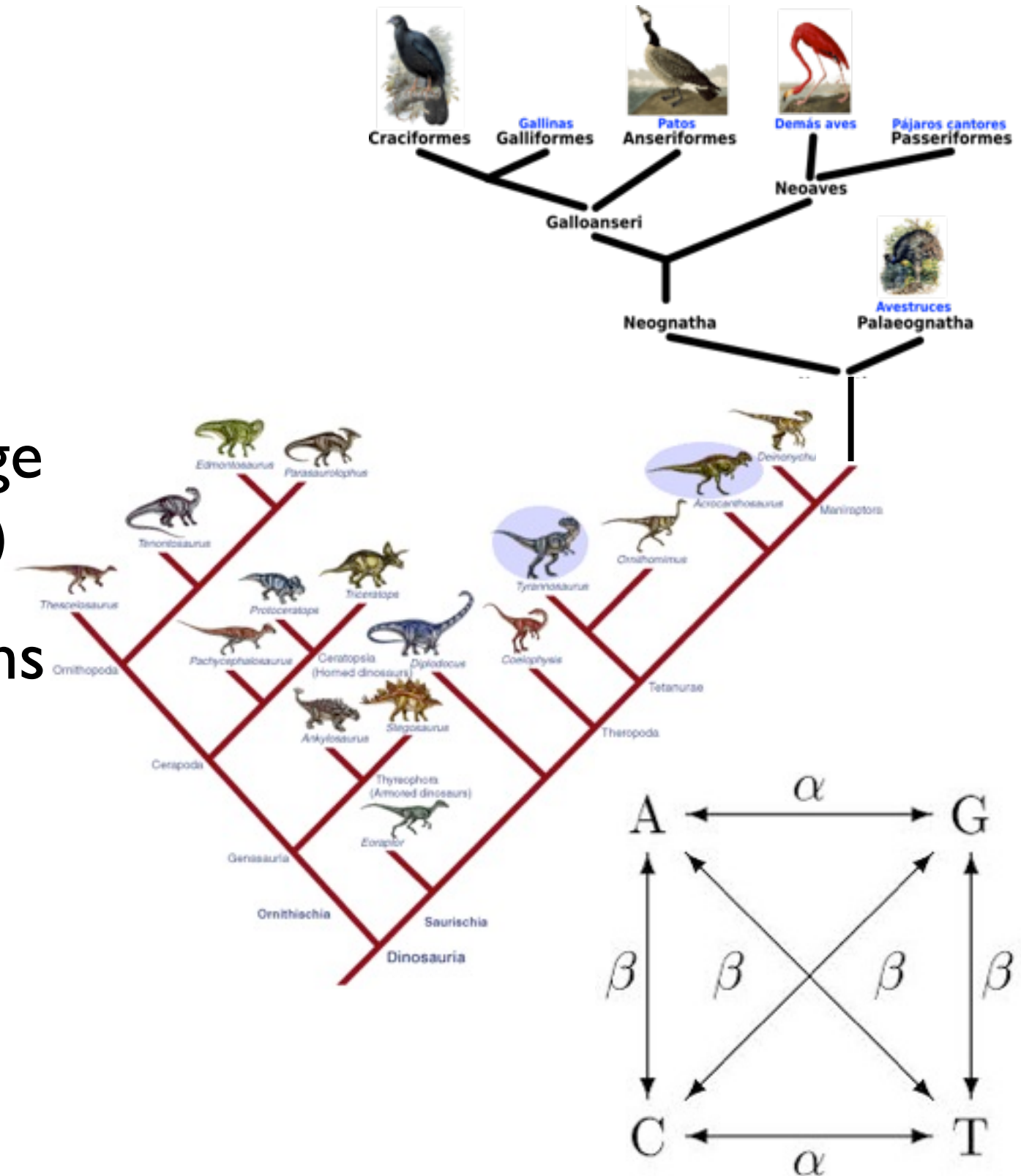
# *Phylogenetics: Reconstruction of ancient species*

- **Goals:**
  - better understand ancient species
  - revive them?
- **Data:** fossil DNA
- Limitation: degrades after few 1000s years
- Are dinosaurs' genomes completely lost?



# Phylogenetic tree

- **Idea:** use the genomes from the descendants of dinosaurs (modern birds)
- We know how DNA change over time (probabilistically)
- Marginalization of unknowns (as in family tree example)
- Additional challenge: structure of tree is unknown



# Foundations