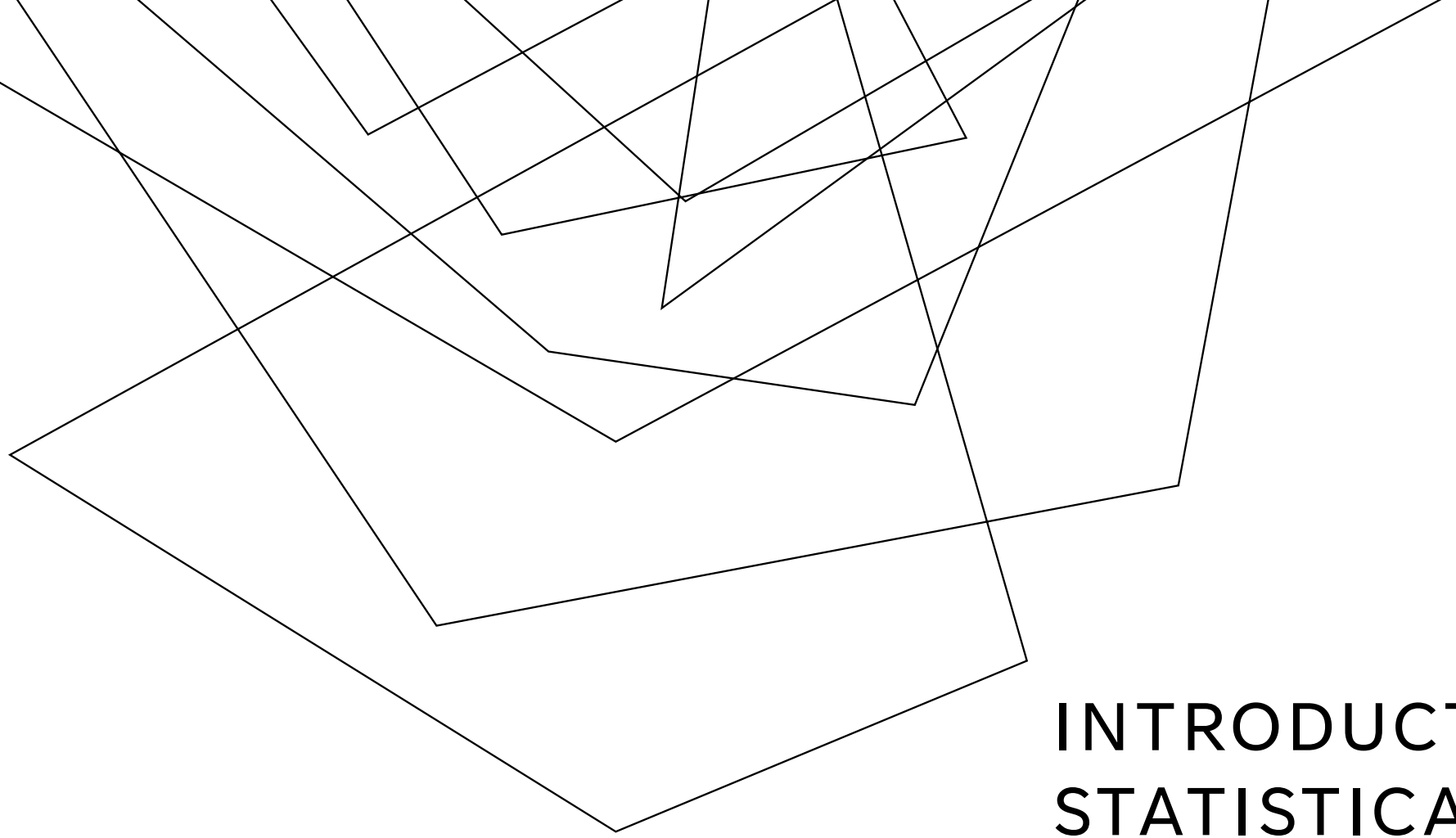


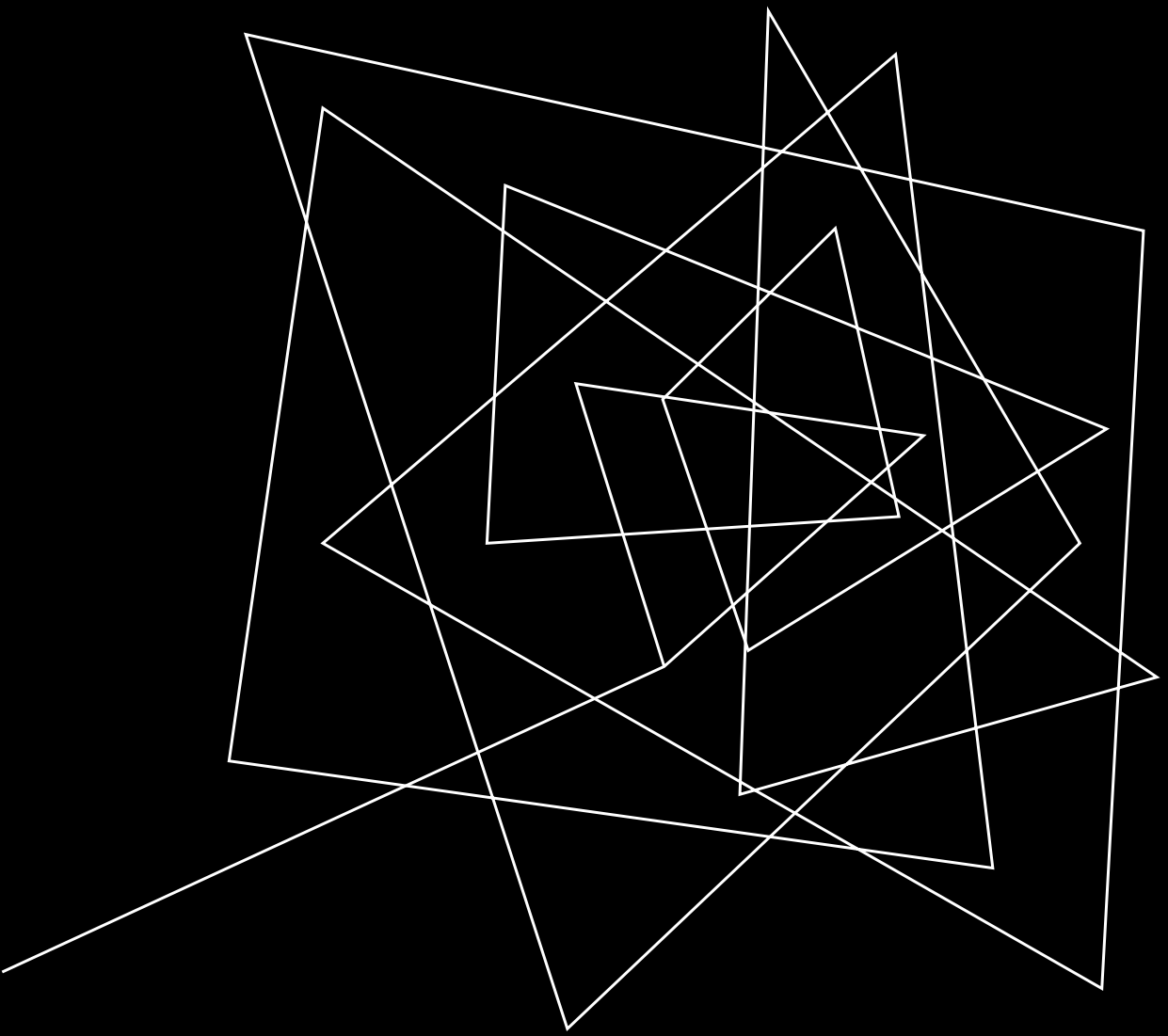
INTRODUCTION TO STATISTICAL INFERENCE

Nima Laal

2024 VIPER Summer School

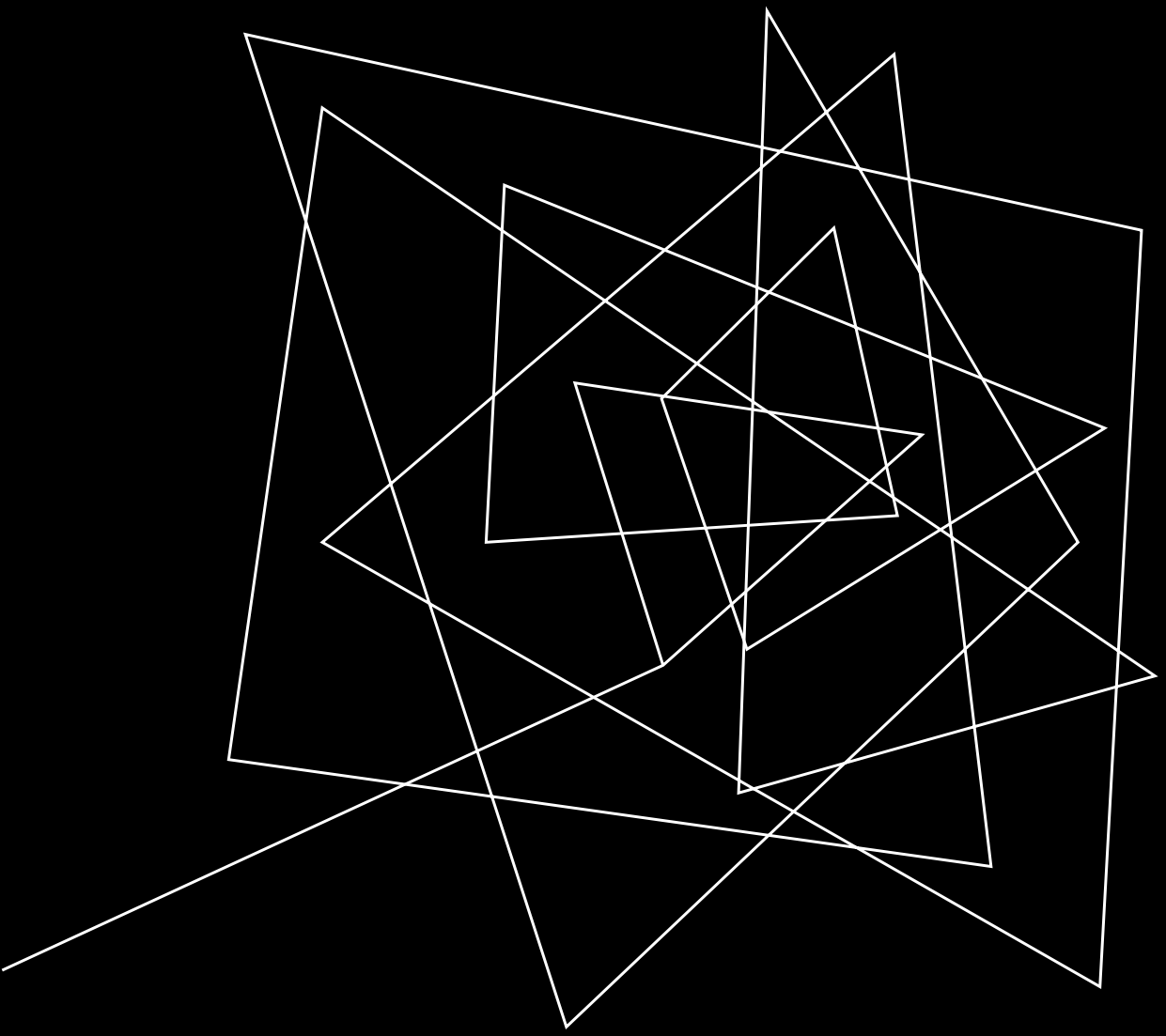
7/9/2024

VANDERBILT UNIVERSITY



THE NEED

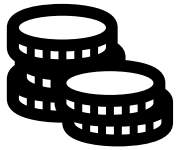
Everything is unknown!



THE PRINCIPLES

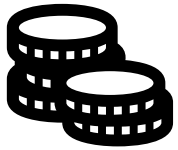
The principles behind probabilities are common sense.

COIN TOSS EXPERIMENT: Nothing but common sense!



- Data
  - 200 trials
  - 180 Heads & 20 Tails
- Question
  - Is the coin used for the experiment fair (equal chance of head over tail)?

COIN TOSS EXPERIMENT: Nothing but common sense!



- Data
  - 200 trials
  - 180 Heads & 20 Tails
- Question
  - Is the coin used for the experiment fair (no preference for head over tail)?
- Answer
  - The odds of observing 180 heads (and even more heads) in 200 trials using a fair coin is extremely low. So, the coin is most likely not fair.

FREQUENTIST APPROACH

- Philosophy: What we observe is a part of a larger set of outcomes. We merely observe random draws from many possible (perhaps infinitely many) outcomes.
- Assigning probabilities to data
- Models are fixed, not probabilistic

Ronald Fisher
(source: Wikipedia)

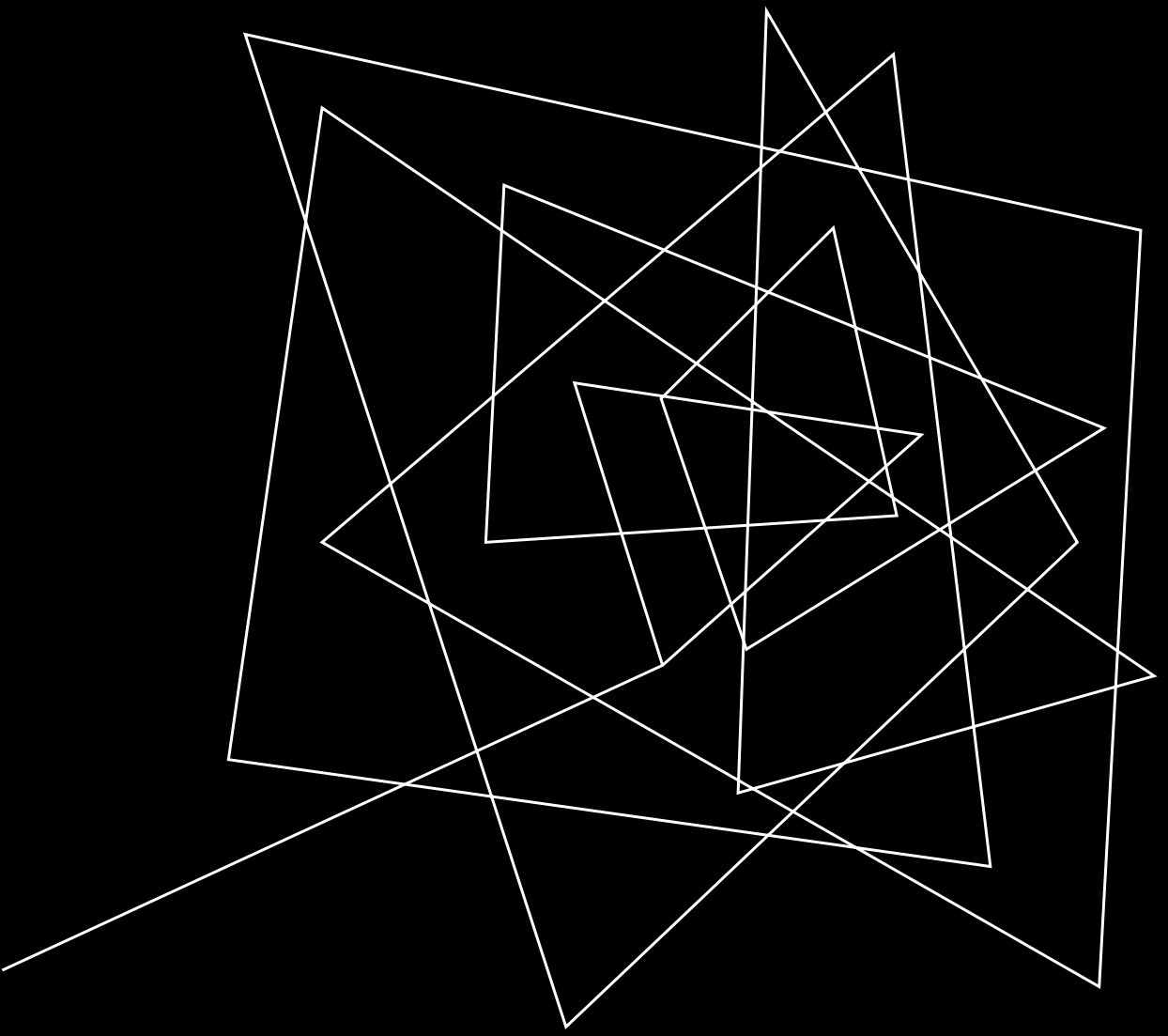
BAYESIAN APPROACH

- Philosophy: What we observe is what it is, and there is nothing random about it. Data is used to update our prior information about a model.
- Data is fixed, not probabilistic
- Assigning probabilities to models instead of data

Thomas Bayes
(source: Wikipedia)

Which one should you use?

You need to use both to convince everyone about your claims. Often, you do not even have a choice! You are forced to use either or both!



THE PROCESS

How to get from data to inferred parameters of models that aim to describe the data?

Step 0) Think about a model that describes data and can answer your questions about the system that generated the data

data = [uninteresting things] + [interesting things]

data = noise + signal

Step 1) Find the likelihood function

$$p\left(\underbrace{d}_{\text{data}} \middle| \underbrace{\theta_M}_{\text{parameters of model } M} \right)$$

Step 2) Choose either a frequentist or a Bayesian approach

Step 3) If a frequentist approach is chosen, you most likely need to explore the *maximum-likelihood* conditions and find optimal estimators.

Step 3) If a Bayesian approach is chosen, you update your *prior* information, using data, by following the *Bayes theorem*

$$\underbrace{p\left(\theta_M \middle| d\right)}_{\text{posterior (updated prior)}} = \frac{\overbrace{p\left(d \middle| \theta_M\right)}^{\text{likelihood}} \overbrace{p\left(\theta_M\right)}^{\text{prior}}}{\underbrace{\int d\theta_M \left\{ p\left(d \middle| \theta_M\right) p\left(\theta_M\right) \right\}}_{\text{evidence (normalization)}}}$$

BAYESIAN PRIOR PROBABILITY

The prior probability is an integral part of the Bayesian approach. It reflects how much you are willing to constrain the model parameters before looking at data. Any probability, even the improper ones, can serve as a prior.

This is your choice!

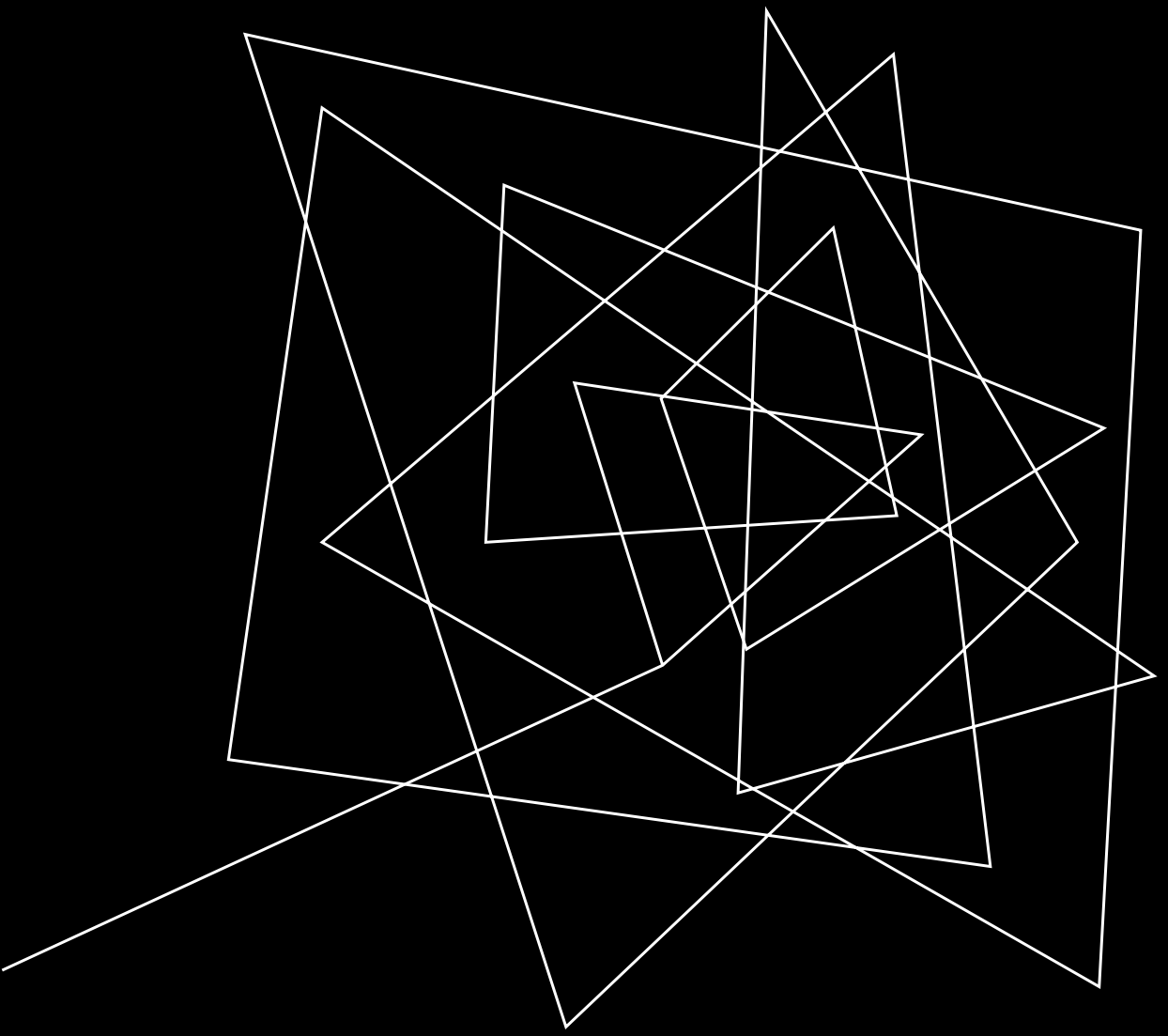
BAYESIAN PRIOR PROBABILITY: Examples

$$\text{uniform: } \theta_{\min} < \theta < \theta_{\max}$$

$$\log\text{-uniform: } \log\theta_{\min} < \log\theta < \log\theta_{\max}$$

$$\text{normal: } p(\theta_1, \theta_2) = \text{Normal}(\text{mean} = \theta_1, \text{cov} = \theta_2)$$

Step 4) Assess how well the inferred model describes the data (hypothesis testing, model selection, etc.).



EXAMPLE 1: PTA DATA ANALYSIS

STEP 0: WHAT IS THE MODEL?

$$r(t) = M\varepsilon + Fa + w$$

Data (Timing Residuals) = Timing Model Contribution + GW Contribution + White Noise

STEP 1: WHAT IS THE LIKELIHOOD FUNCTION?

$$w = r - M\varepsilon - Fa$$

$$w \sim \text{Normal}\left(\text{mean} = 0, \text{Cov} = N\right)$$

$$p\left(r \middle| \text{Model}\right) = \frac{\exp\left\{-\frac{1}{2}\left(r - M\varepsilon - Fa\right)^T N^{-1}\left(r - M\varepsilon - Fa\right)\right\}}{\sqrt{\det\left(2\pi N\right)}}$$

STEP 2: TAKE A SIDE: FREQUENTIST

$$T = [M, F];\ b = [\varepsilon, a];\ \frac{\partial \ln p(r|\text{Model})}{\partial b} = 0$$

$$\therefore \quad \hat{b} = \left(T^T N^{-1} T\right)^{-1} T^T N^{-1} r$$

maximum-likelihood estimator

STEP 2: TAKE A SIDE: BAYESIAN

$$B = \left\langle bb^T \right\rangle_{\text{realization}}$$

$$p(b,B) = \text{Normal}(\text{mean} = 0, \text{Cov} = B)\, p(B)$$

$$p(b,B|r) = p(B)\frac{\exp\left\{-\frac{1}{2}\left[(Tb-r)^T N^{-1}(Tb-r) + b^T B^{-1} b\right]\right\}}{\sqrt{\det(2\pi N)\det(2\pi B)}}$$

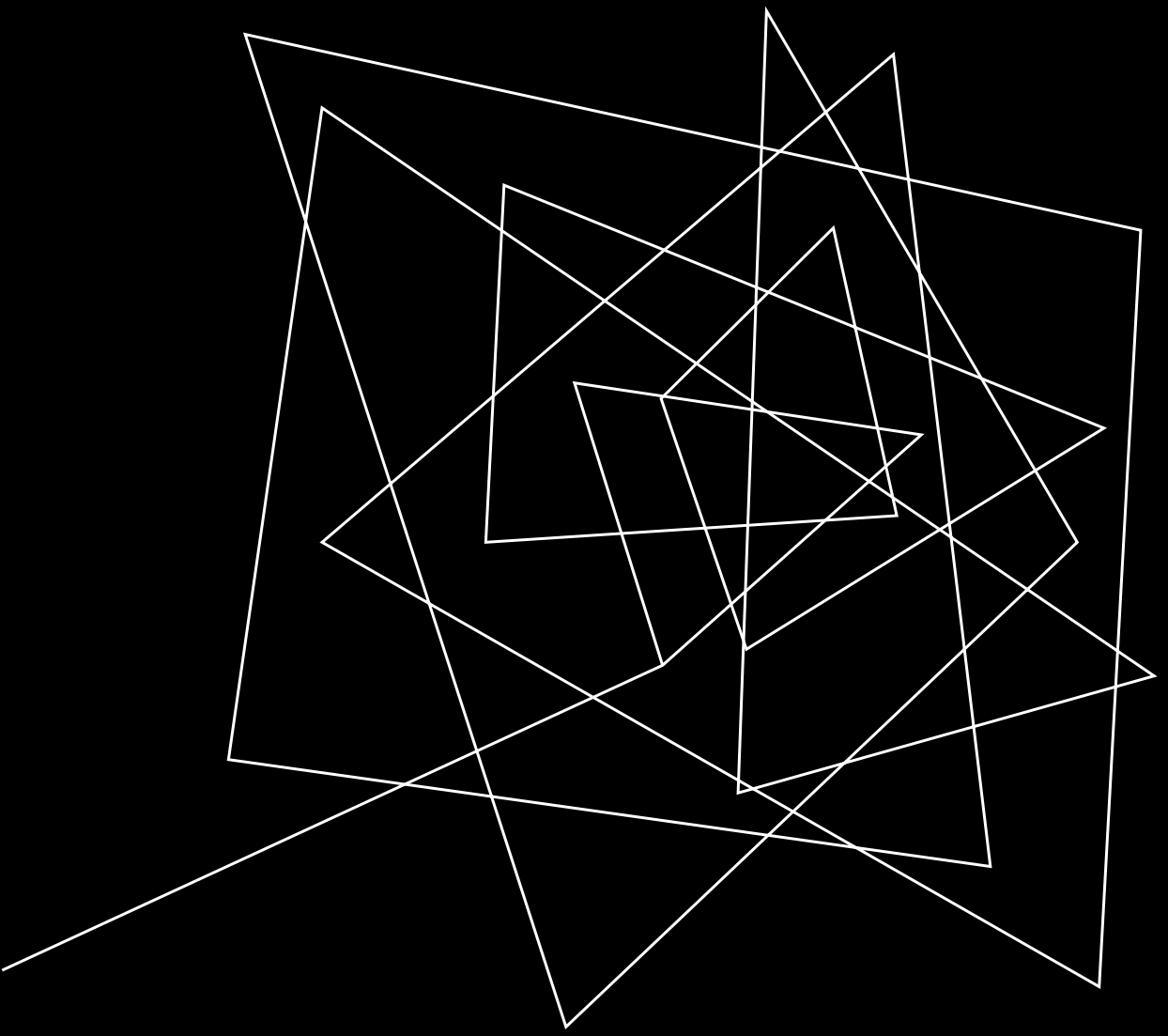
STEP 2: TAKE A SIDE: BAYESIAN (CONTINUED)

$$p\left(b,B\middle|r\right)=p\left(B\right)\frac{\exp\left\{-\frac{1}{2}\left[\left(Tb-r\right)^{T}N^{-1}\left(Tb-r\right)+b^{T}B^{-1}b\right]\right\}}{\sqrt{\det\left(2\pi N\right)\det\left(2\pi B\right)}}$$

$$B=\begin{bmatrix}0 & 0\\ 0 & \rho^{2}\end{bmatrix};\ P\left(\rho\right)\propto\frac{1}{\rho}$$

VANDERBILT UNIVERSITY

YUK...



THE MACHINERY OF BAYESIAN STATISTICS

Monte Carlo simulation is the machinery of Bayesian statistics. It makes the Bayesian statistics practical!


THIS IS A THOUSAND MONKEYS
WORKING AT A THOUSAND
TYPEWRITERS.

TURN CHAOS INTO ORDER

- The idea: You give a thousand monkeys a dictionary, and they will eventually write Shakespeare
- The monkeys are your computer processing units
- Dictionary is your fully-expressed bayes theorem, Monte Carlo algorithms, and the condition of detailed balance
- The words in the dictionary are random numbers
- Shakespeare is the posterior probability of the model parameters

DETAILED BALANCE: Net flux of probability is zero out of every point in the parameter space

$$p\left(\underbrace{x_0}_{\text{starting point}} \middle| \underbrace{x_1}_{\text{next point}}\right) p(x_0) = p(x_1|x_0)\, p(x_1)$$

Detailed balance guarantees that the posterior probability is eventually reached if every point in the parameter space is accessible (the posterior is ergodic).

DETAILED BALANCE (CONTINUED)

$$\underbrace{p\left(x_0 \middle| x_1\right)}_{\text{unknown}} = \underbrace{q\left(x_0 \middle| x_1\right)}_{\text{known}} \min\left(1, \underbrace{\alpha}_{\text{weight}} = \frac{p\left(x_1\right) q\left(x_1 \middle| x_0\right)}{p\left(x_0\right) q\left(x_0 \middle| x_1\right)}\right)$$

- *min* is used to ensure probability does not exceed 1.
- The choice for weight is to ensure the detailed balance is satisfied no matter what!

$$\text{Let } \alpha = \frac{p(x_1)q(x_1|x_0)}{p(x_0)q(x_1|x_0)} < 1$$

$$p(x_0|x_1) = q(x_0|x_1)\left[\frac{p(x_1)q(x_1|x_0)}{p(x_0)q(x_0|x_1)}\right]$$

$$p(x_1|x_0) = q(x_1|x_0)$$

$$\therefore p(x_0|x_1)p(x_0) = q(x_1|x_0)\frac{p(x_1)}{p(x_0)}p(x_0) = p(x_1|x_0)p(x_1)$$

MARKOV CHAIN MONTE CARLO

$$\alpha = \frac{p(x_1)q(x_1|x_0)}{p(x_0)q(x_0|x_1)}$$

- The ratio α is called the Hastings ratio.
- Proposed pints which result in a Hastings ratio smaller than a uniform random number between 0 and 1 are rejected. This is the criteria to reject or accept a new point.
- If the proposed distribution, q, is symmetric, the algorithm is called Metropolis algorithm.
- The Bayesian evidence is no longer needed to obtain posterior probability!

GIBBS SAMPLING: α=1.

If you know all the conditional probabilities, and you know how to sample from them directly, you can use Gibbs sampling, where α=1.

$$p\left(\theta_1 \middle| \theta_2, \theta_3, \ldots, \theta_n, \text{data}\right) = p_1$$

$$p\left(\theta_2 \middle| \theta_1, \theta_3, \ldots, \theta_n, \text{data}\right) = p_2$$

$$\vdots$$

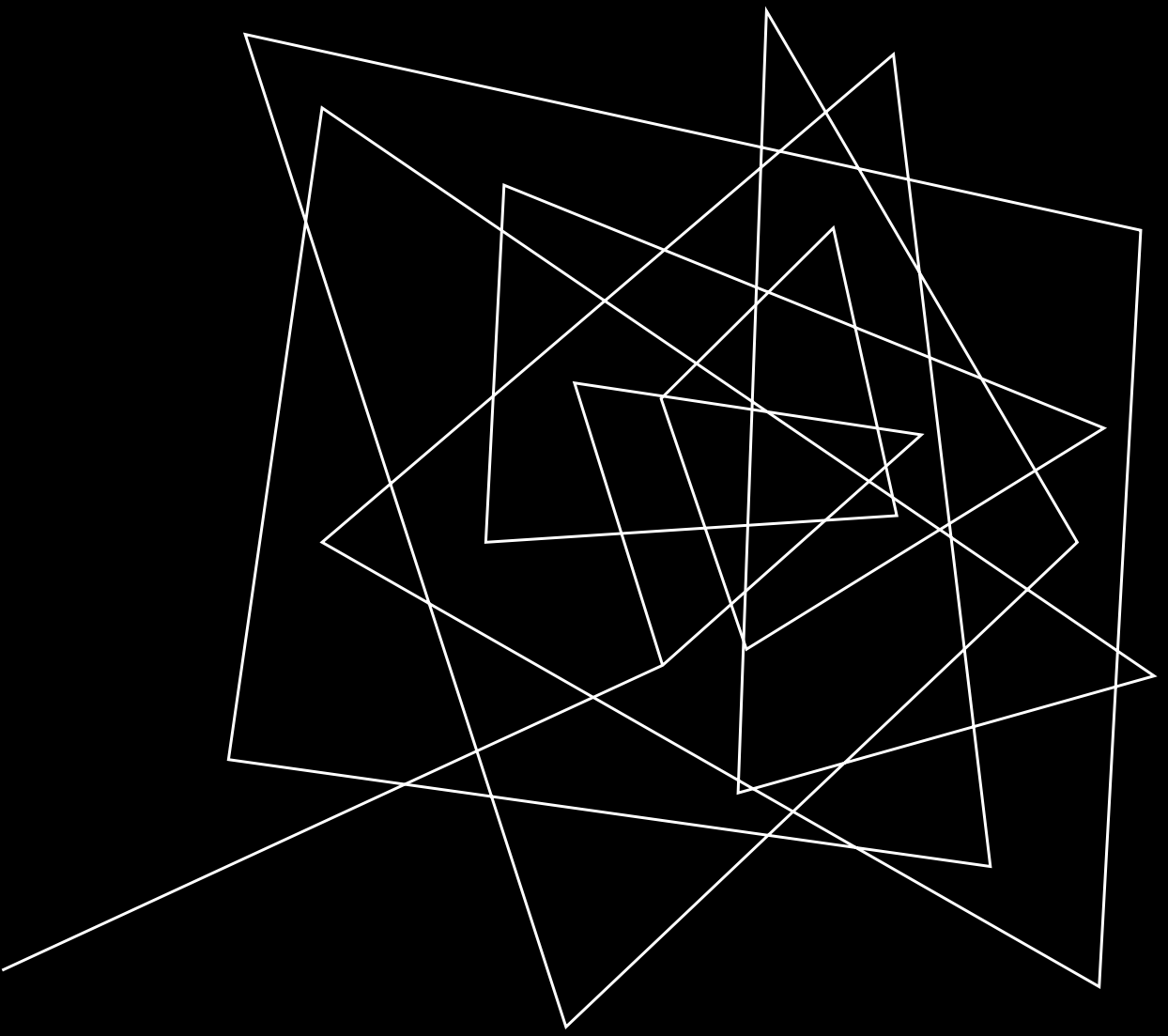
$$p\left(\theta_n \middle| \theta_1, \theta_2, \ldots, \theta_{n-1}, \text{data}\right) = p_n$$

THIS IS THE LAST TIME I WILL TALK ABOUT MONKEYS!

- The smarter the monkeys, the faster and more accurate the results will be to Shakespeare's. Your implementation of Monte Carlo simulation matters!
- Gibbs sampling is like giving monkeys the actual Shakespeare, but with page numbers randomized: all they need to do is to get the ordering right.
- The more monkeys you have, the better. GPUs as well as CPUs with high core-count will give you that.


"IT WAS THE BEST OF TIMES,
IT WAS THE BLURST OF TIMES..."
YOU STUPID MONKEY!

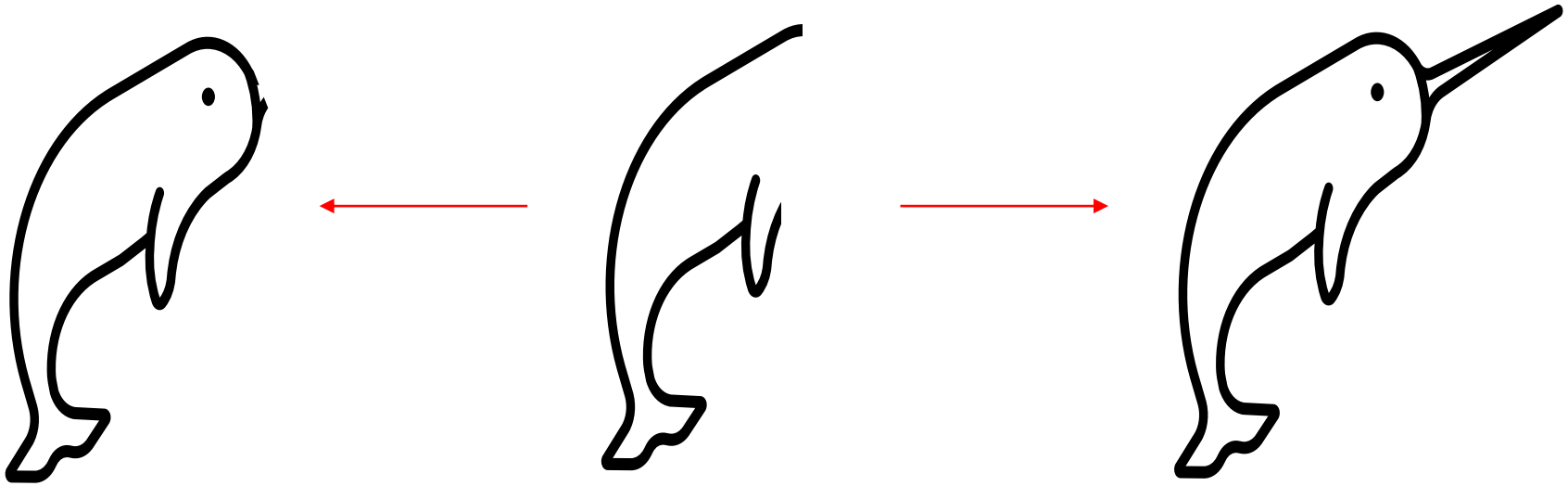
Different flavors of MCMC Demo



MODEL SELECTION & HYPOTHESIS TESTING

Sometimes, all that people care about is just a single number that summarizes your research! Model selection techniques give you that number.

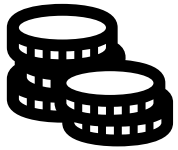
WHICH MODEL FITS YOUR DATA? Your data is rarely informative enough to make things obvious!



FREQUENTIST HYPOTHESIS TESTING: P-VALUE

- What is the probability that you obtain your data and more extreme versions of your data just by randomness nature of the data itself?
- The null-hypothesis is compared to an alternative hypothesis
- You need to choose a threshold for rejecting/accepting the null-hypothesis. This threshold is referred to as the significance level. Significance level of 5% is a common choice.
- If the probability of obtaining a data as extreme as yours is less than the significance level, you reject the null-hypothesis. Otherwise, you accept it.

COIN TOSS EXPERIMENT: REVISITING



- Data: 200 trials, 180 Heads & 20 Tails
- Question: Can this data be generated by random chance while the coin is fair given significance level of 5%?

$$p = \underbrace{\sum_{k=180}^{200}}_{\text{The sum is over data and more extreme versions of it}} \text{Binomial}\left(n = 200, k = k, \quad \underbrace{p = \frac{1}{2}}_{\text{null-hypothesis}} \right)$$

BAYESIAN MODEL SELECTION: The final boss of PTA detection statistics!

- Monte Carlo simulation does not care about the Bayesian evidence because all you need is Hastings ratio and bunch of random numbers. But, Bayesian evidence is extremely valuable: you need it to compare different models.

$$z_M = \int d\theta_M \left\{ p\left(d|\theta_M\right) p\left(\theta_M\right) \right\}$$

$$O_{MM'} = \underbrace{\left[\frac{Z_M}{Z_{M'}}\right]}_{\text{Bayes factor}} \underbrace{\left[\frac{p\left(\theta_M\right)}{p\left(\theta_{M'}\right)}\right]}_{\text{Prior ratio}}$$

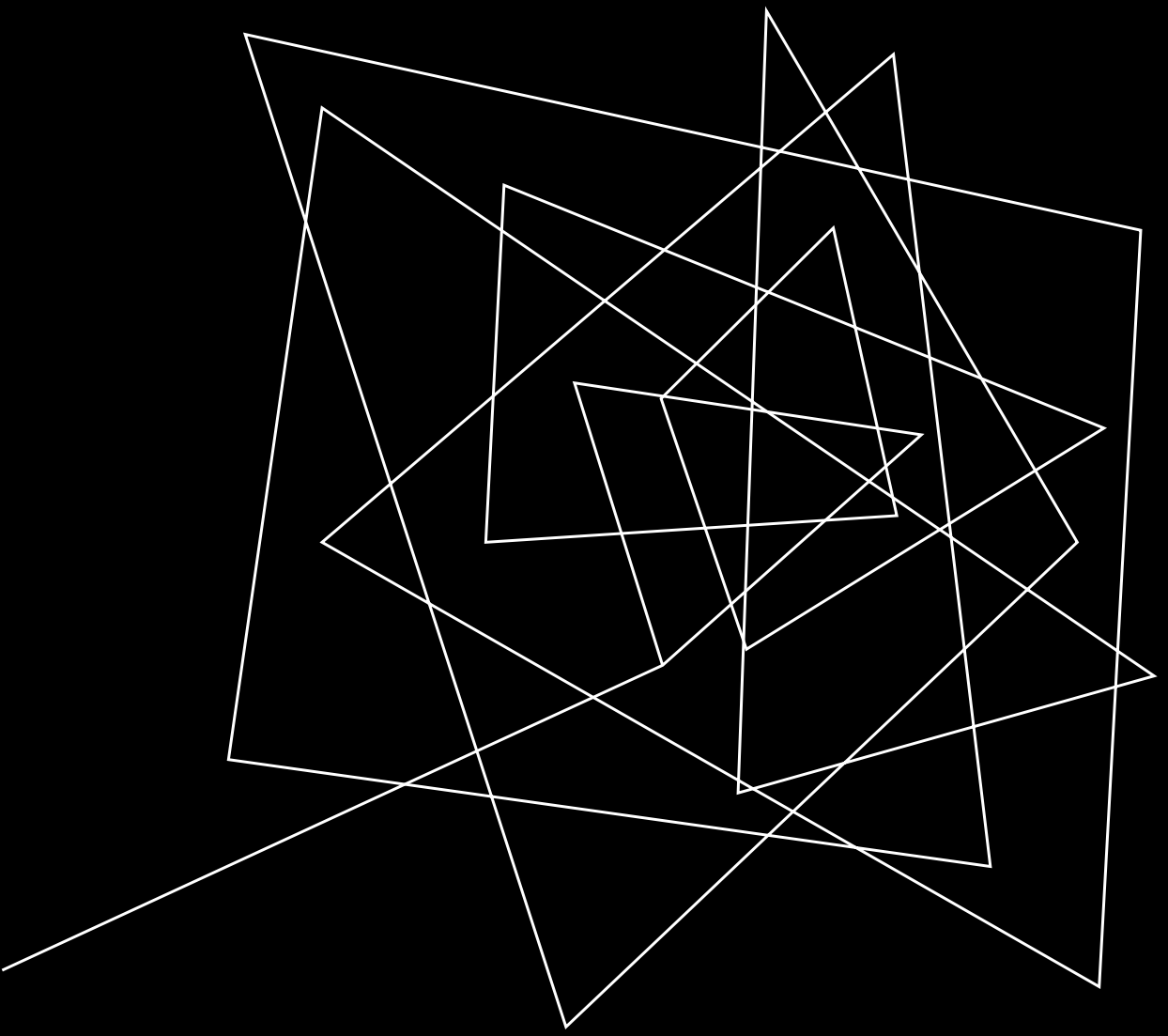
Odds ratio

ODDS RATIO: THE MEANING

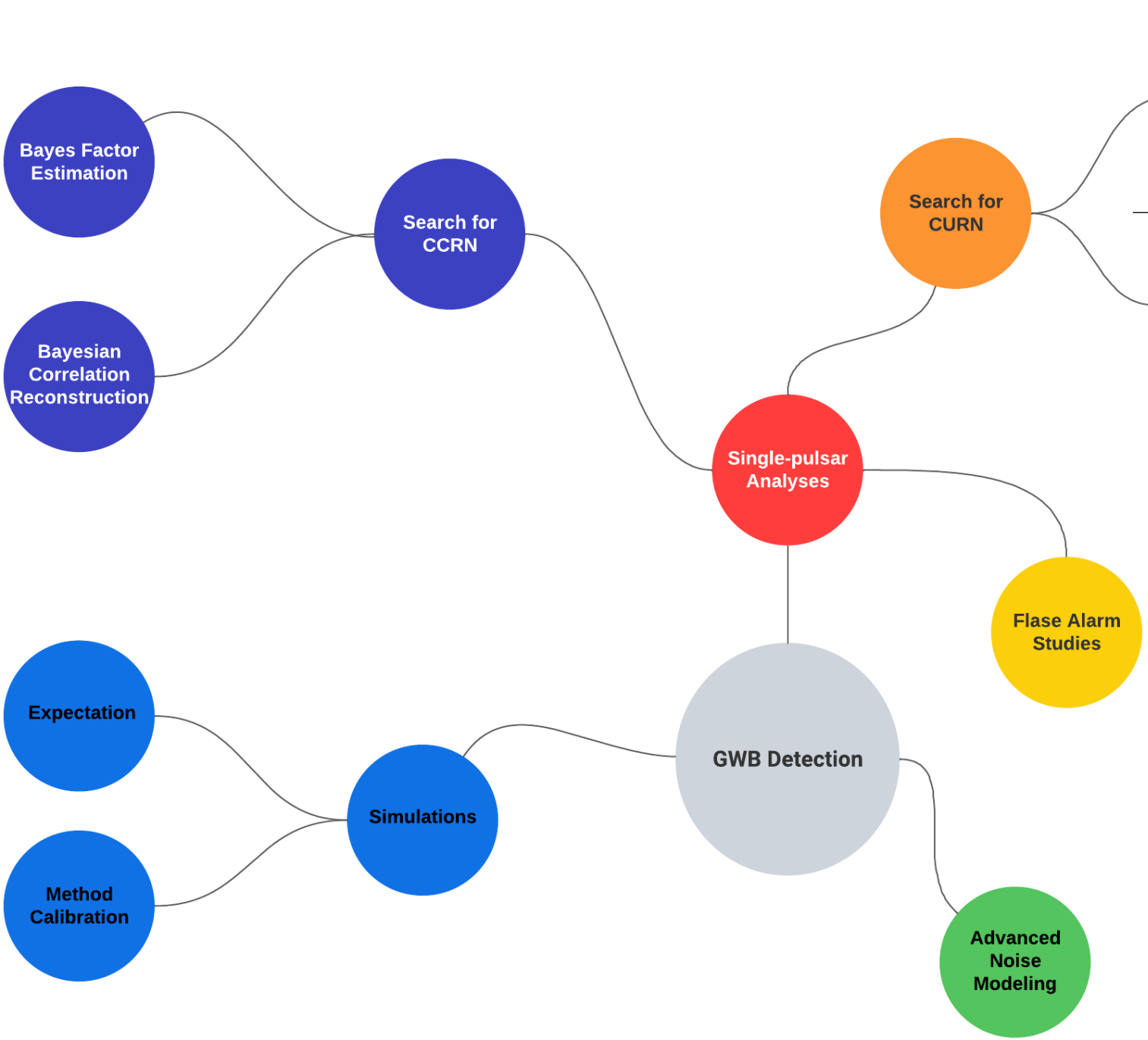
- The odds ratio tells you the odds of a particular model being better than another model in explaining the data. What else do you want?
- How large should it be to be say, with confidence, that a certain model is better than another model? You simulate data sets and determine this for your specific case. THERE IS NO FIXED THRESHOLD.
- The prior preference for a given model affects the odds ratio. You may not like this, but I do!

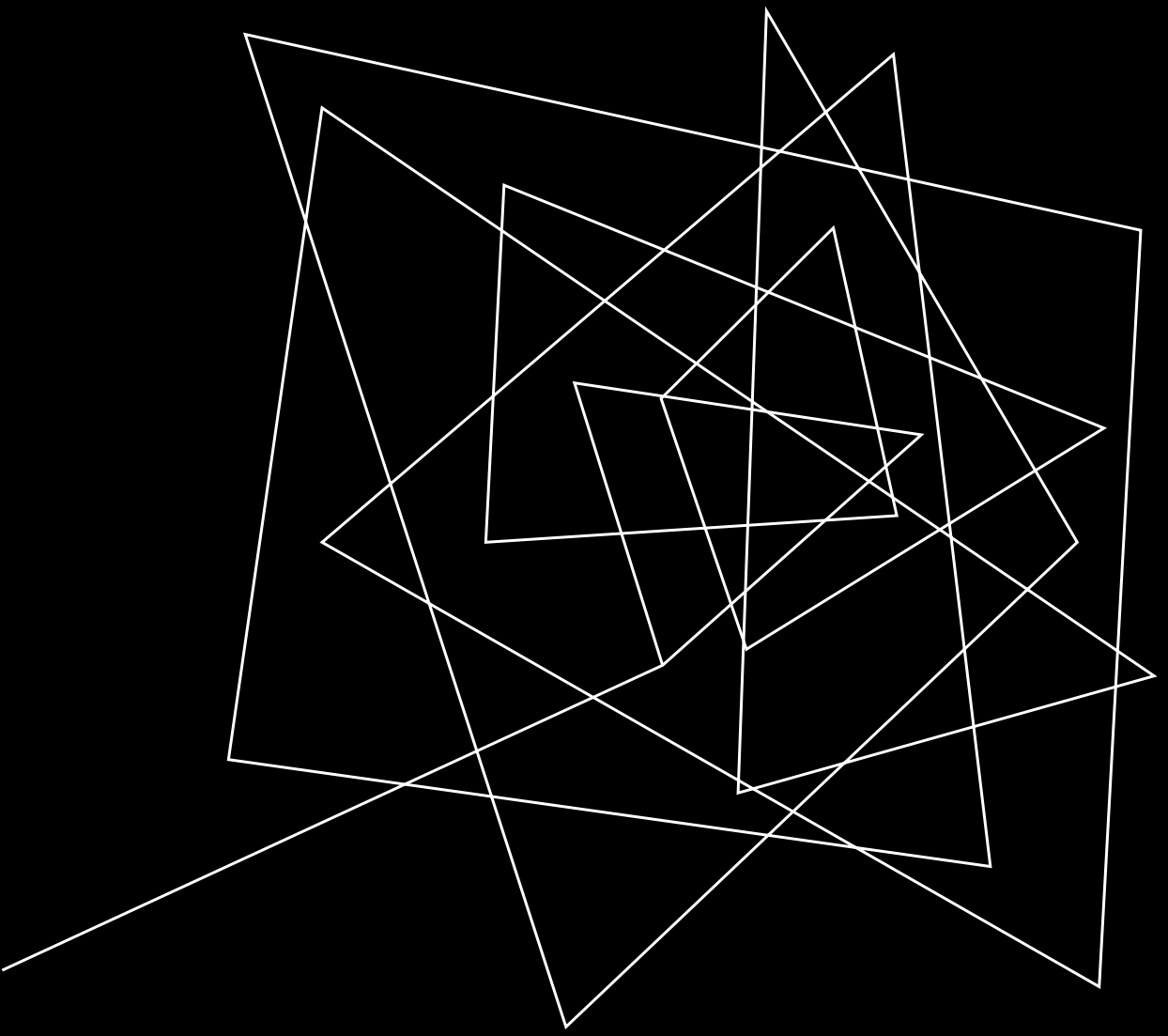
ODDS RATIO: HOW TO ESTIMATE

- It is a nightmare to estimate. For models with small number of parameters, Monte Carlo simulations can be used to evaluate the integral of the Bayesian evidence. Nested sampling is capable of estimating the Bayesian evidence.
- For models with many parameters (e.g., PTAs), you need fancy algorithms. Hyperspace sampling and thermodynamic integration are two examples.



PTA GWB DETECTION


Bayes Factor Estimation
Bayesian Correlation Reconstruction
Search for CCRN
Frequentist Cross Correlation Analyses
Search for CURN
Spectrum Fitting
Single-pulsar Analyses
Flase Alarm Studies
Expectation
Method Calibration
Simulations
GWB Detection
Advanced Noise Modeling



RESOURCES

RESOURCES

- Bayesian Data Analysis
- The Nanohertz Gravitational Wave Astronomer
- Detection methods for stochastic gravitational-wave backgrounds: a unified treatment