Making decisions quickly (part 1)

Computational Cognitive Science 2014

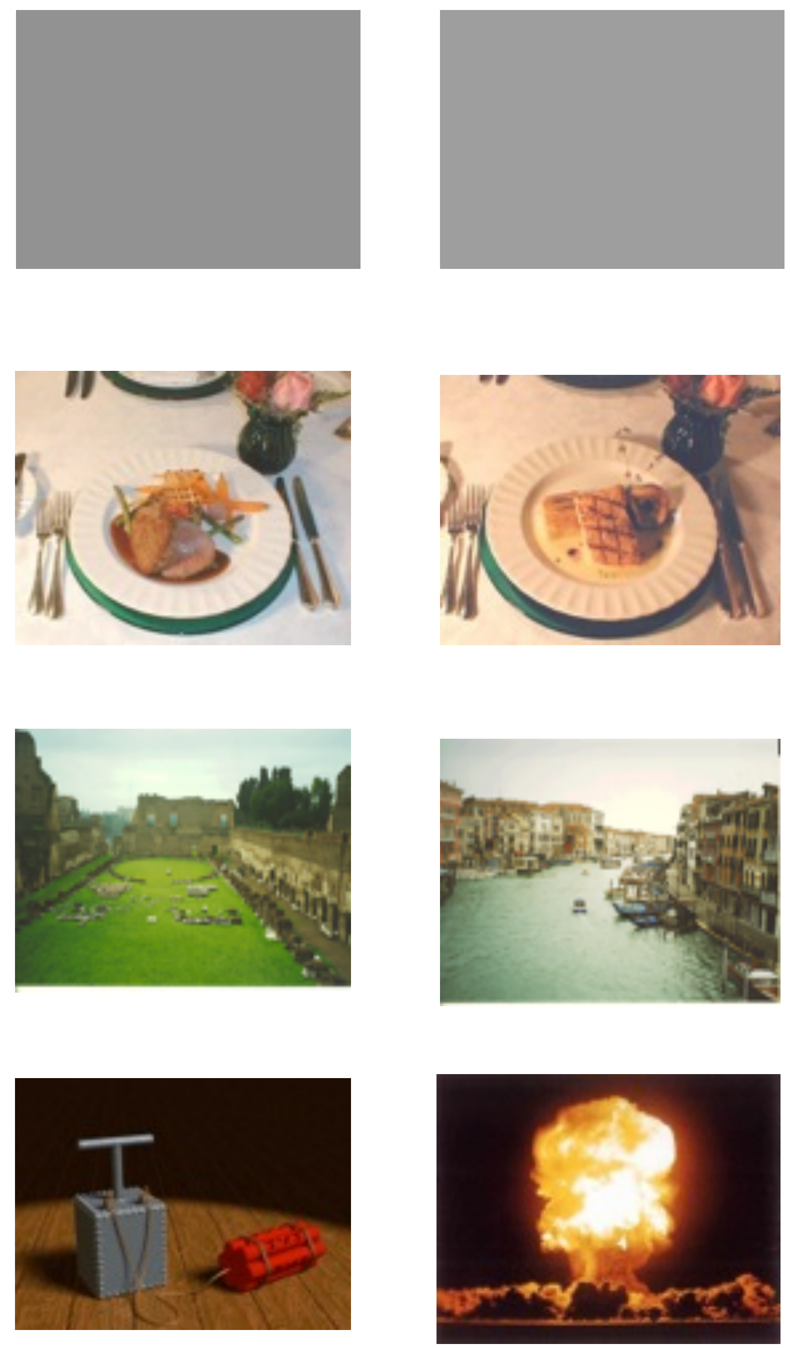
Dan Navarro

Overview of the lectures

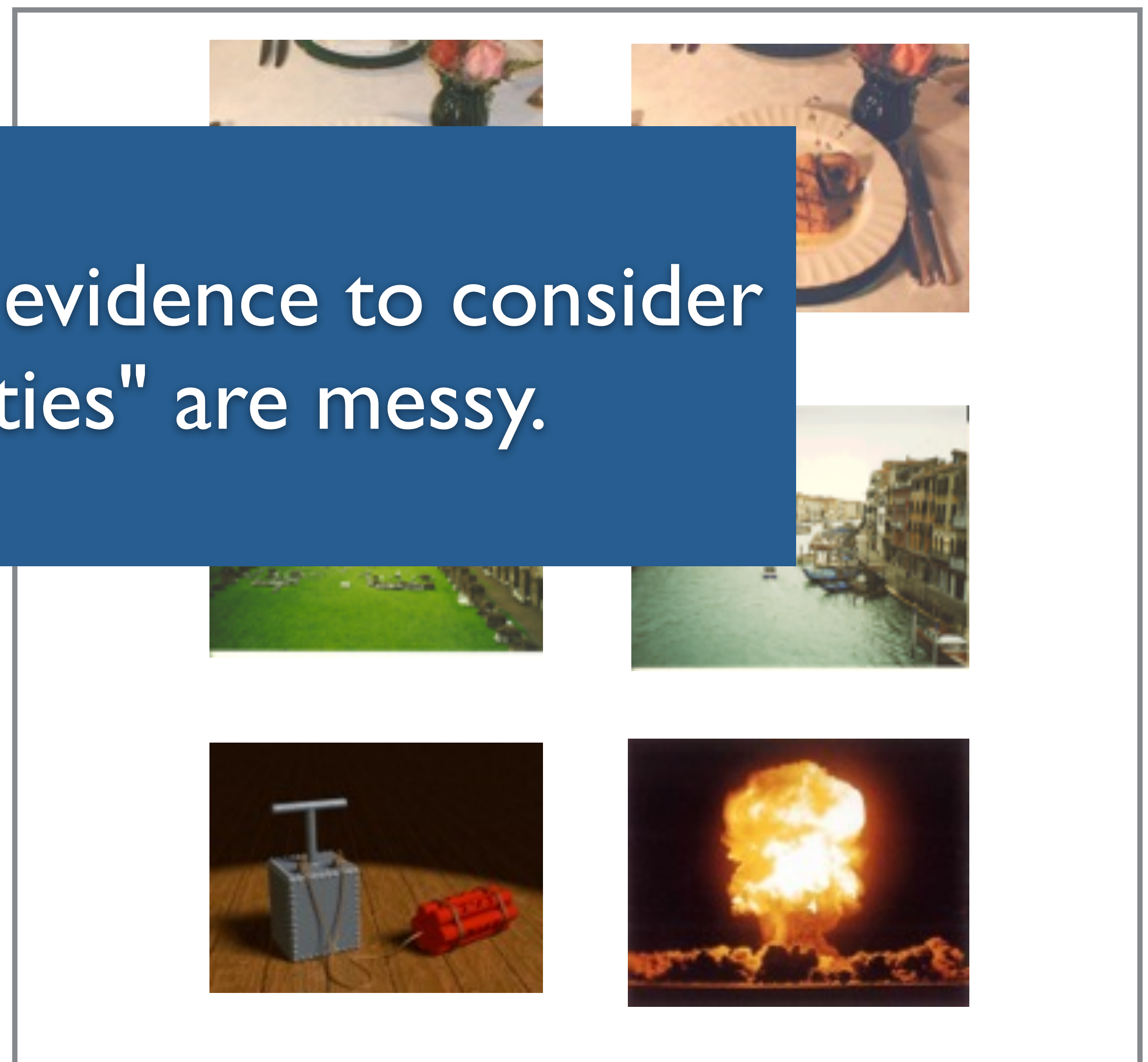
- This lecture:
  - Historical background: psychophysics
  - Introduction to signal detection theory
  - The utility of time and computation
  - Introduction to sequential sampling models

- Next lecture
  - More on sequential sampling models
  - Applications of SSMs to cognitive science
  - Using SSMs in machine learning
  - Using SSMs in neuroscience

Many kinds of decisions

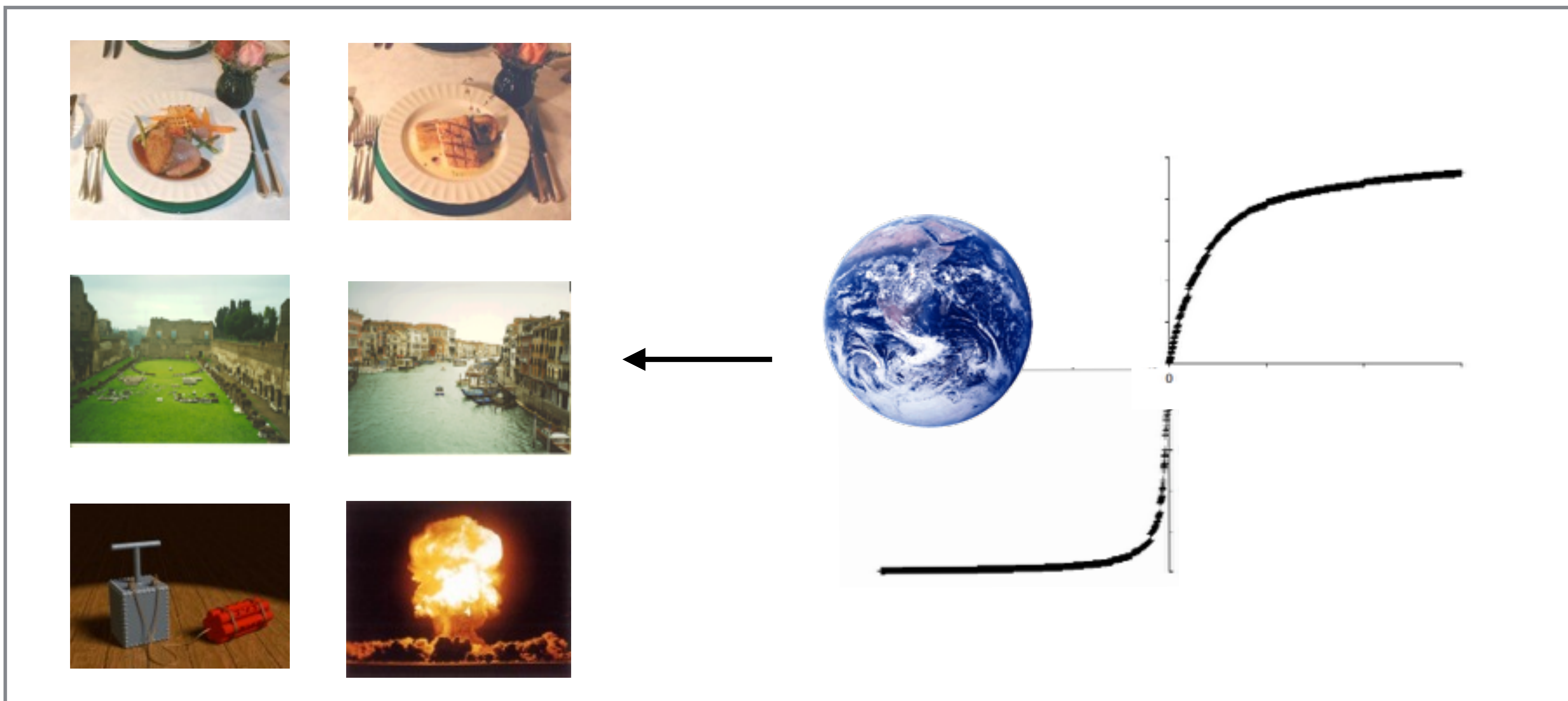


We’ve talked about complex choices

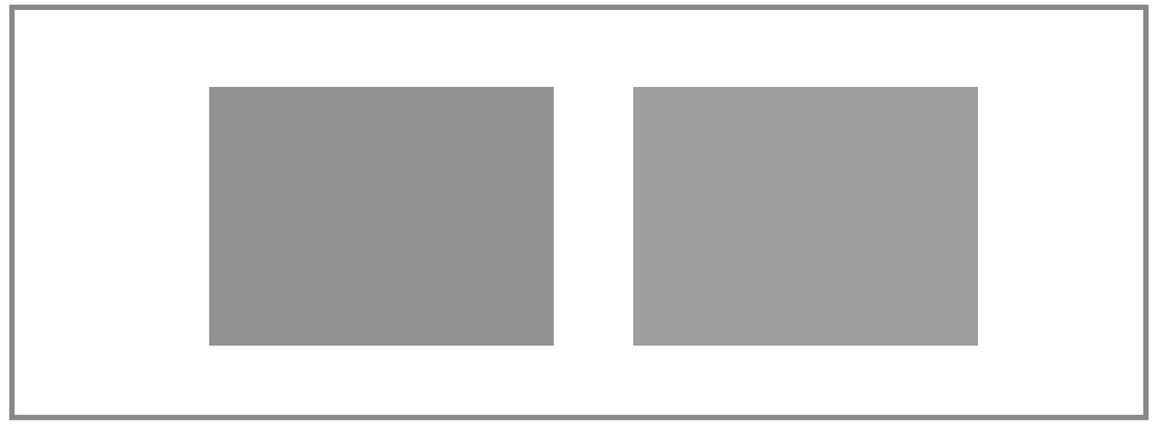


Many sources of evidence to consider & the "utilities" are messy.

And we had some hints that a lot of this "complexity" is in the world… simple "sampling" processes reproduce prospect curves



So let’s talk about simple decisions.

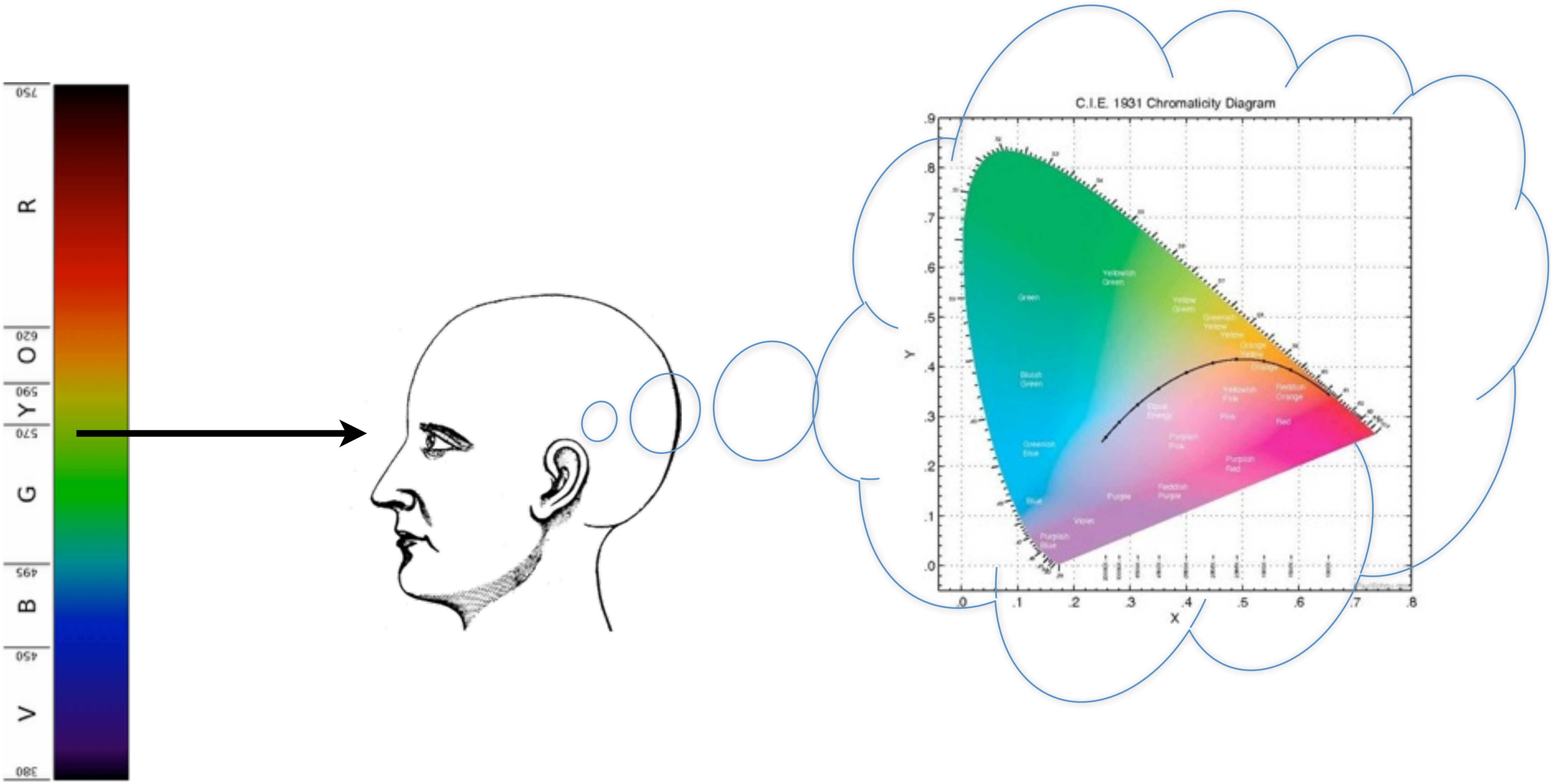


Because the decisions are simple, this is the more tractable case: EU theory and prospect theory both make the vacuous “pick the darker one” prediction.

Not surprisingly, there's more to it than this

A very brief primer on psychophysics

Physical quantities vs subjective ones



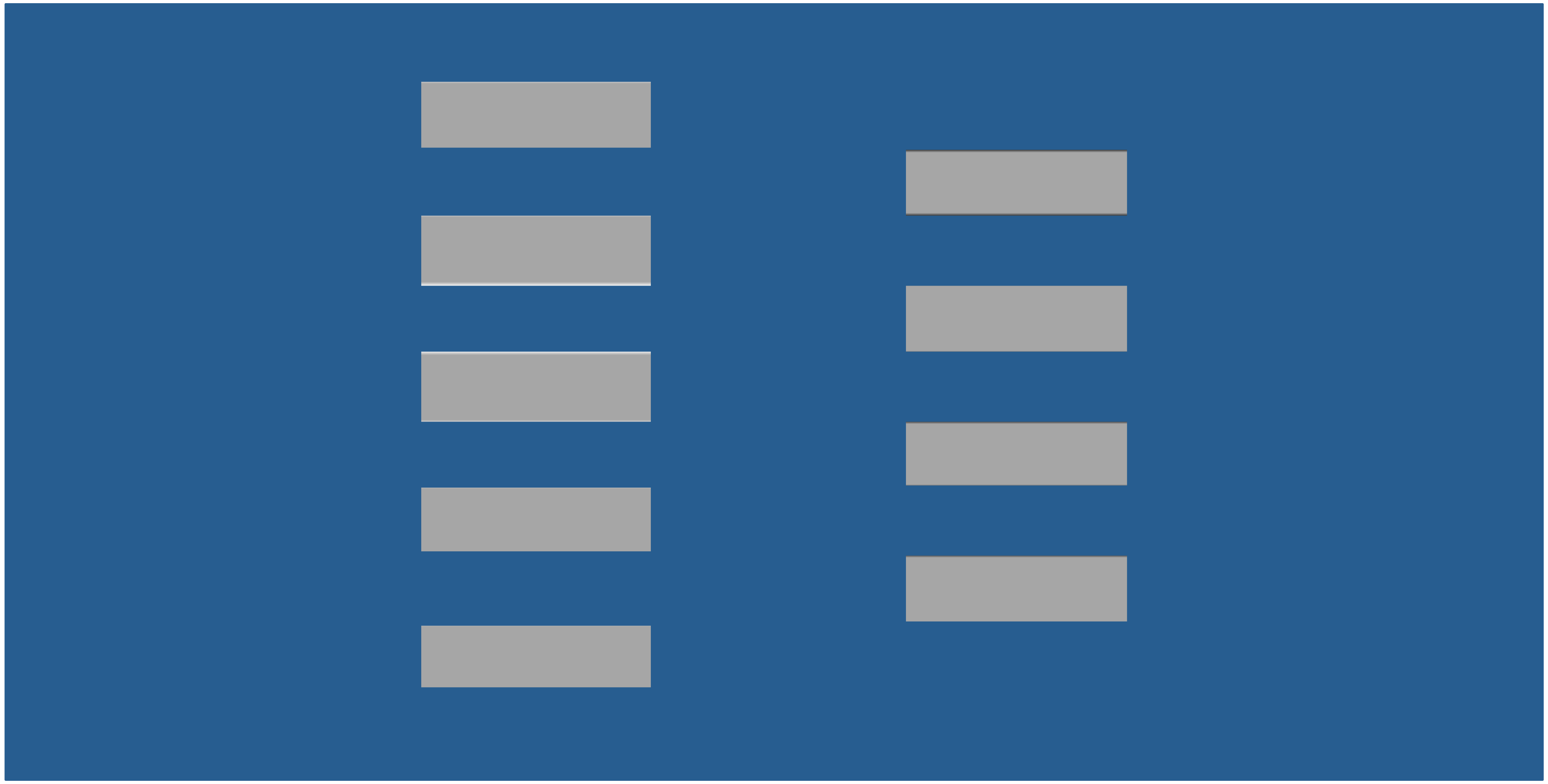
Physical dimension of wavelength

Subjective colour space is very different

Subjective "brightness" is not the same thing as objective "luminance"

White's illusion: the grey rectangles are the same colour.

Subjective "brightness" is not the same thing as objective "luminance"



Really!

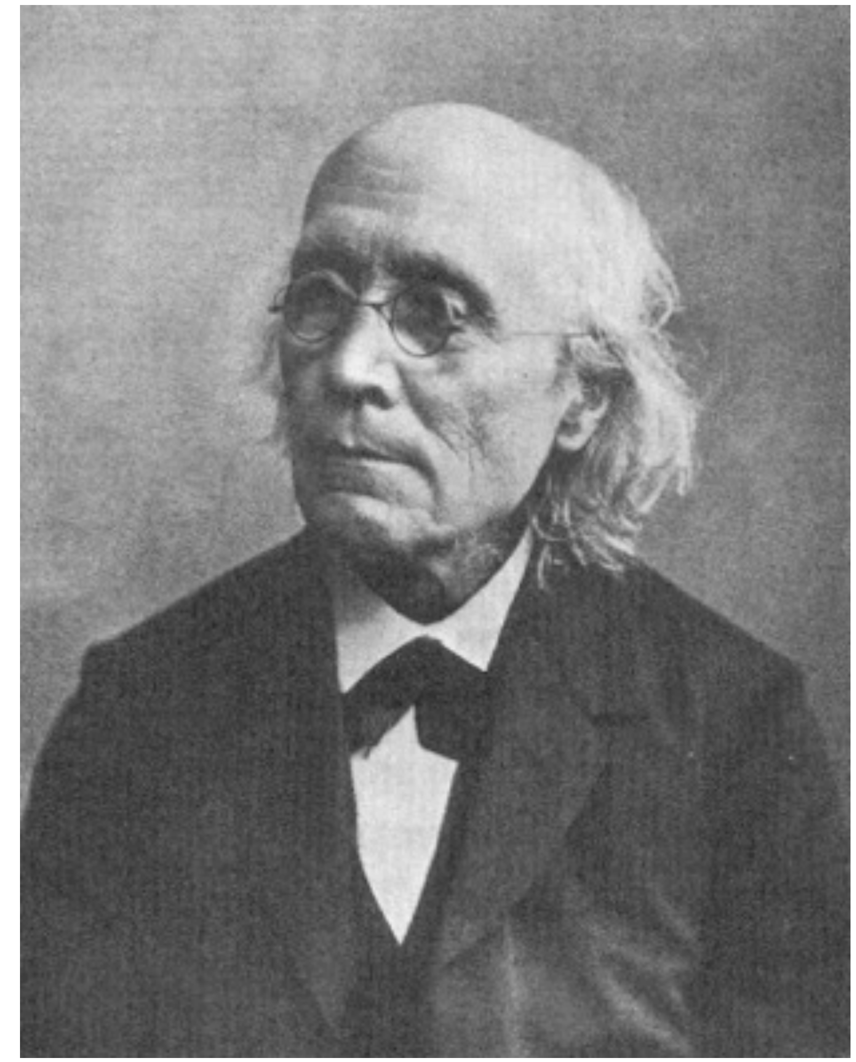
The illusion isn’t the main point

The point is that we can’t just assume that people see colours in the “obvious” way… and if not, what should we assume about how people see these colours?

Psychophysical laws

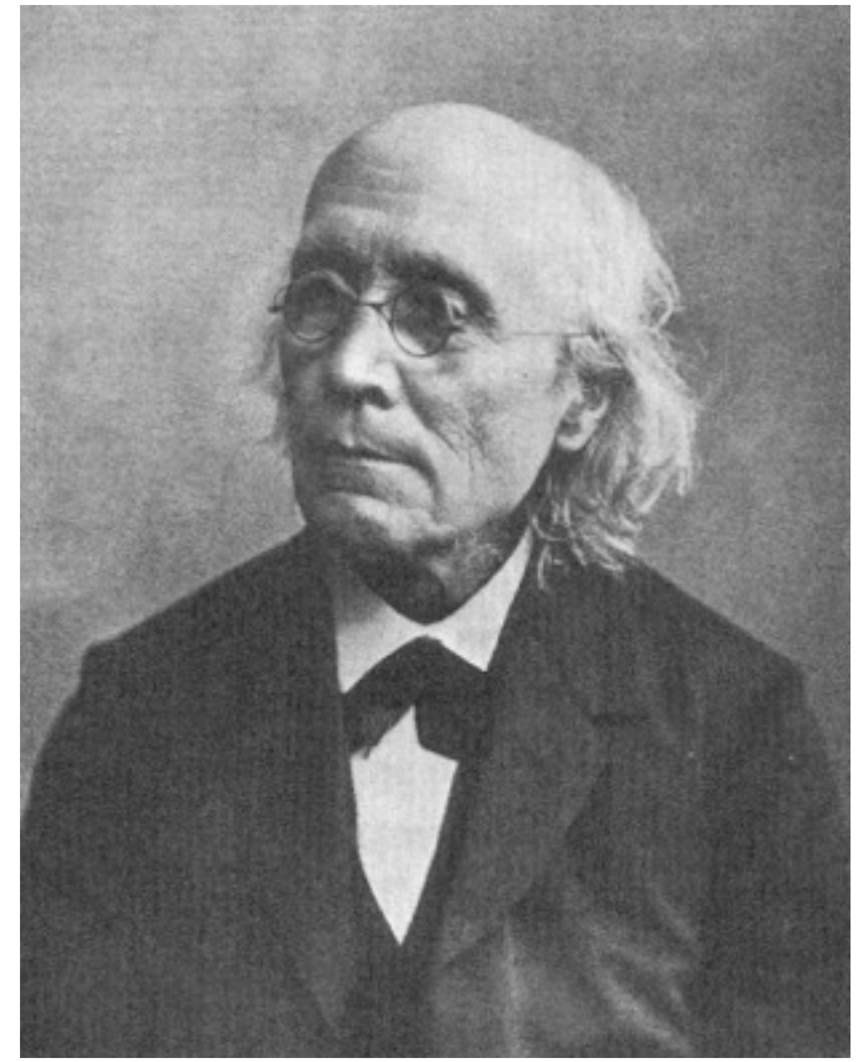
- Definition: modelling the relationship between a subjective quantity ψ (e.g., "brightness") and corresponding objective quantity ϕ (e.g., "luminance")

- This is an old problem, arguably the first topic studied in modern experimental psychology (Fechner 1860)

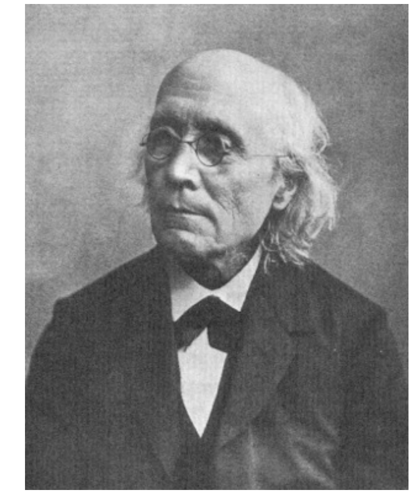


Psychophysical laws

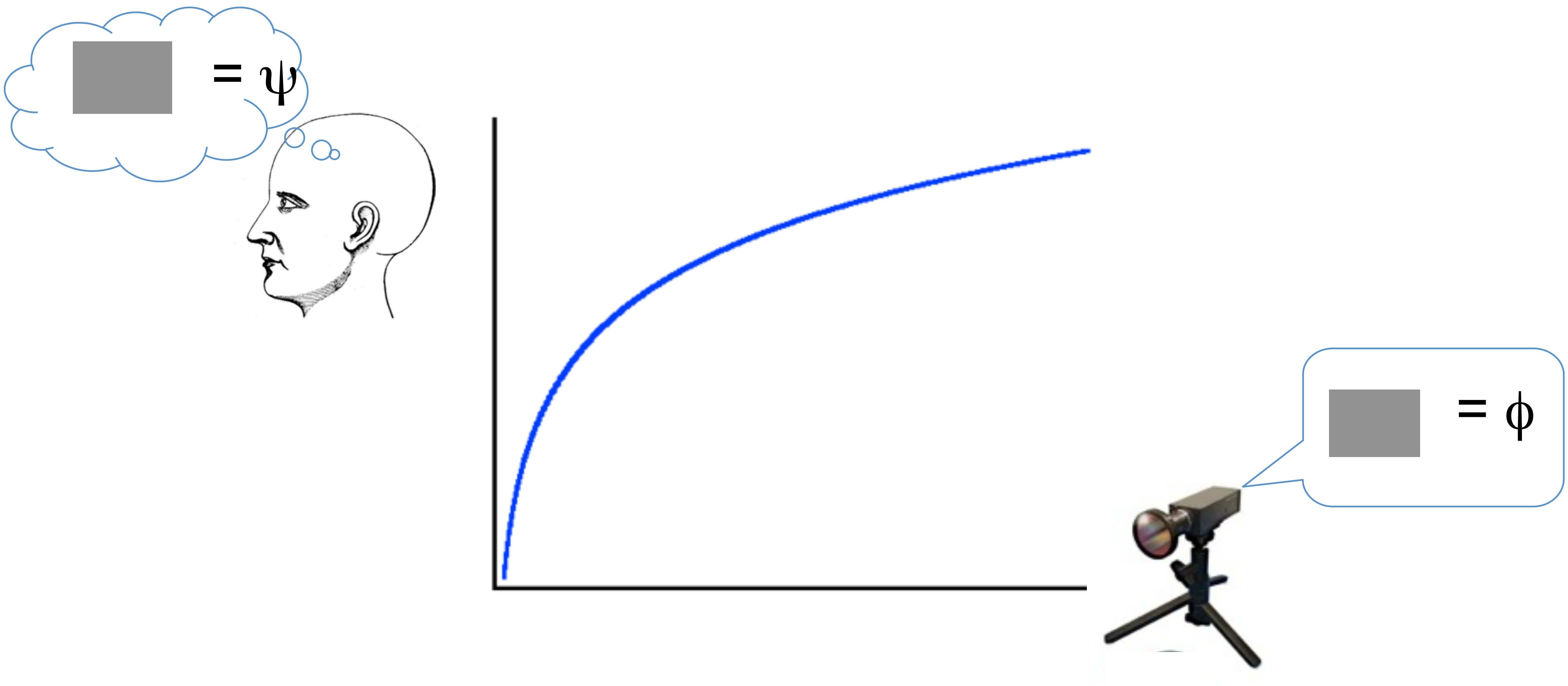
- The relationship is typically logarithmic, or very nearly so
  - i.e., $\psi = k \log \phi$
- This "Weber-Fechner" law remained the best general model for psychophysical relationships for almost a century (until Stevens, 1956).



The psychophysical idea



- Nonlinear law relating objective to subjective magnitudes, $\psi = k \log \phi$



How do you show that $\psi = k \log \phi$?
An example of a psychophysics experiment

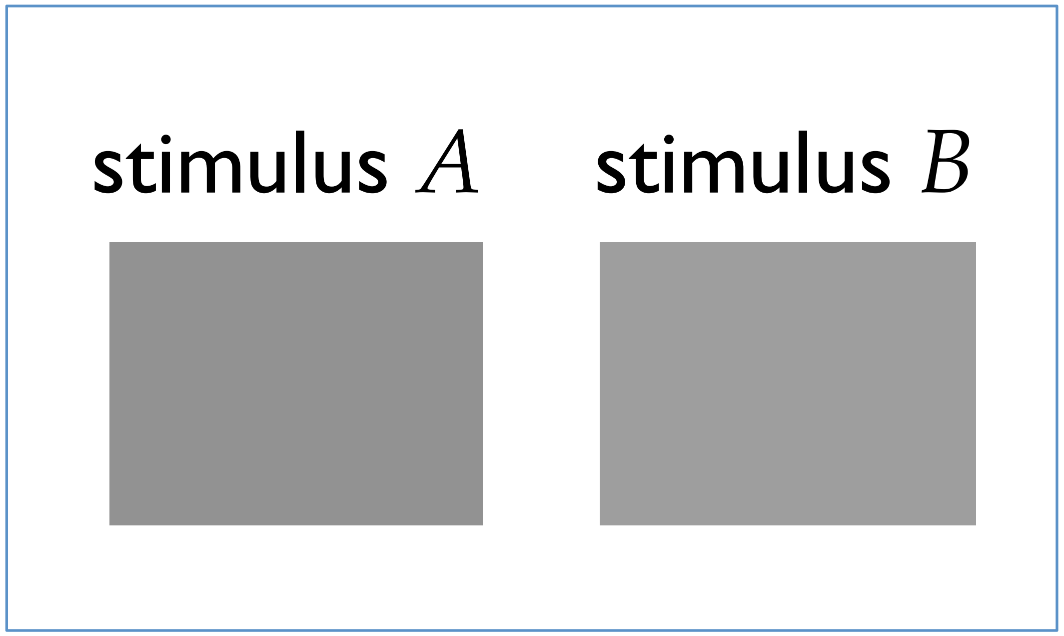
Use people's decisions to learn about their visual perception!

- The "method of right and wrong cases"
  - Give people two stimuli, A and B
  - Ask them to decide if A>B or B>A.

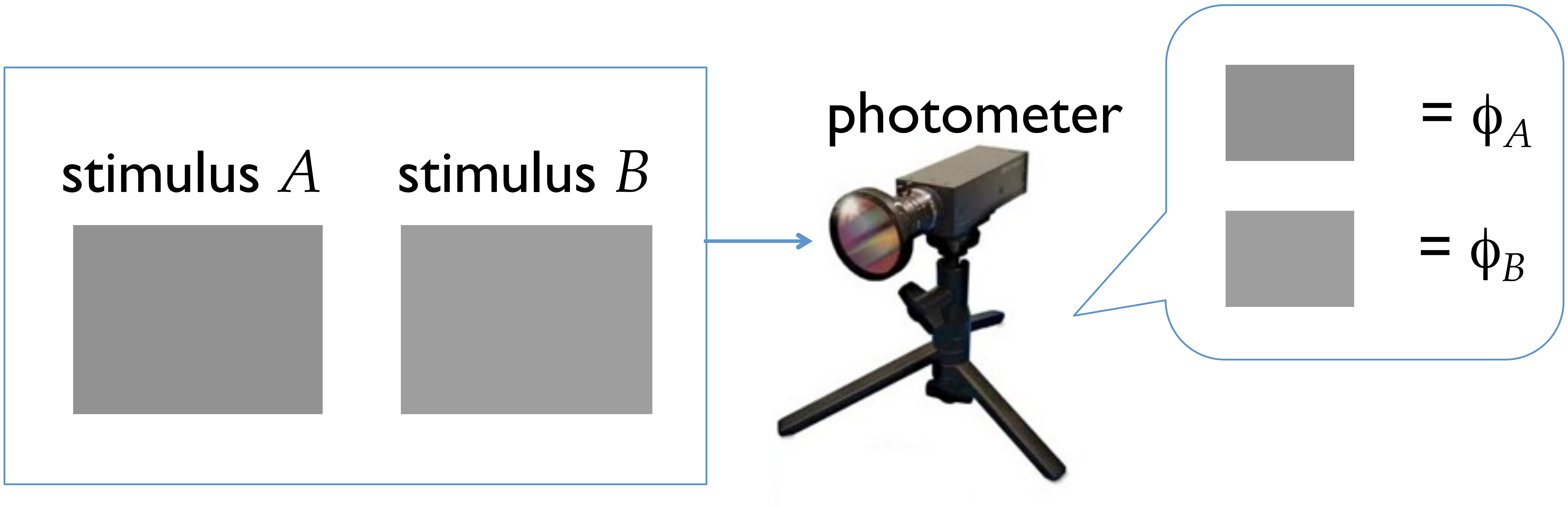
A. Use people's decisions to learn about their visual perception!

- The "method of right and wrong cases"
  - Give people two stimuli, A and B
  - Ask them to decide if A>B or B>A.

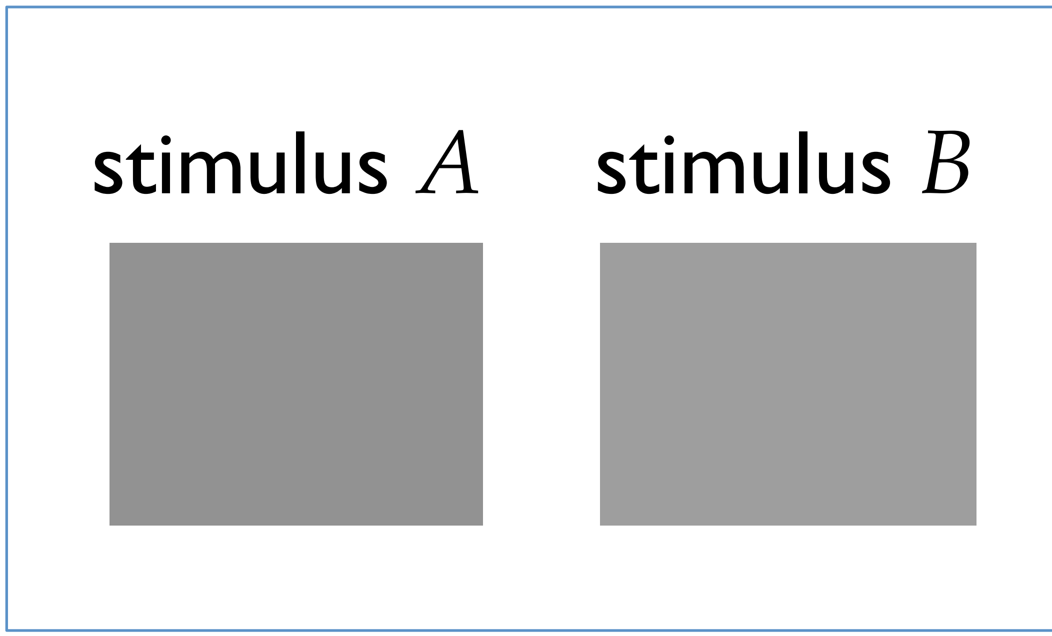
- Goal:
  - Infer the subjective difference $\psi_A - \psi_B$ from the choice probability P(A>B), given that the two objective magnitudes φ_A and φ_B are known



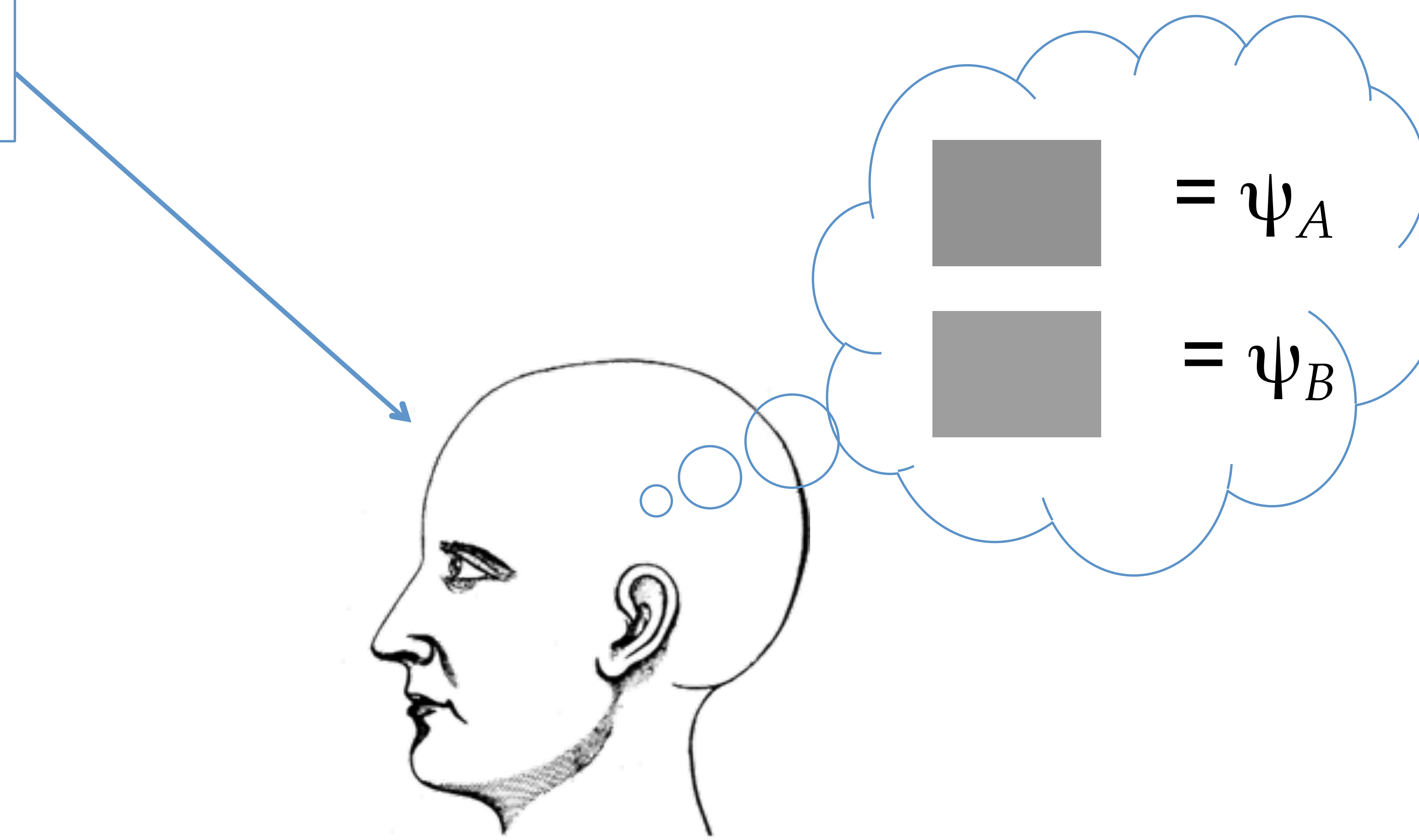
Here are the stimuli people need to choose between



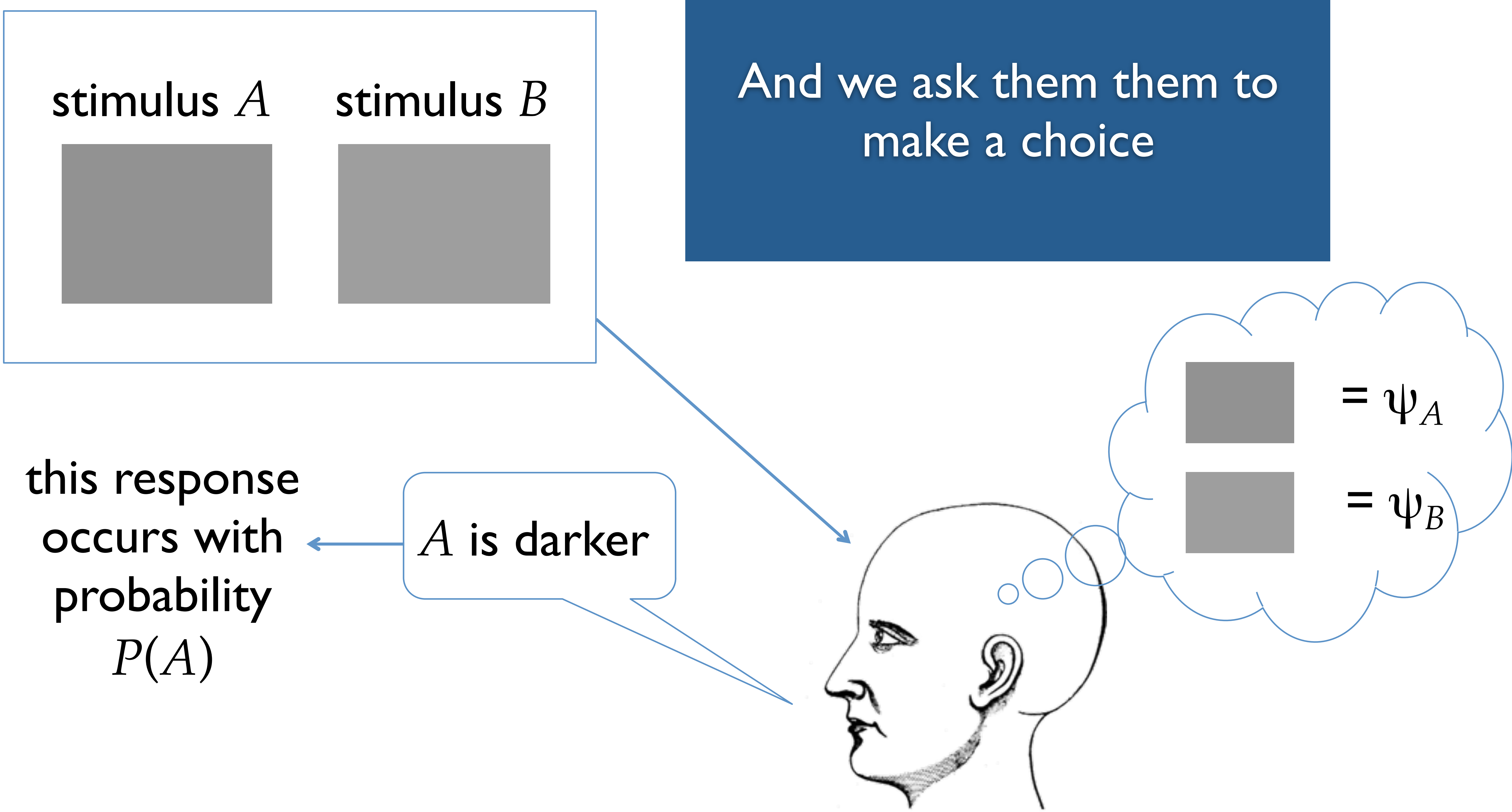
We use a measuring device to determine physical magnitudes

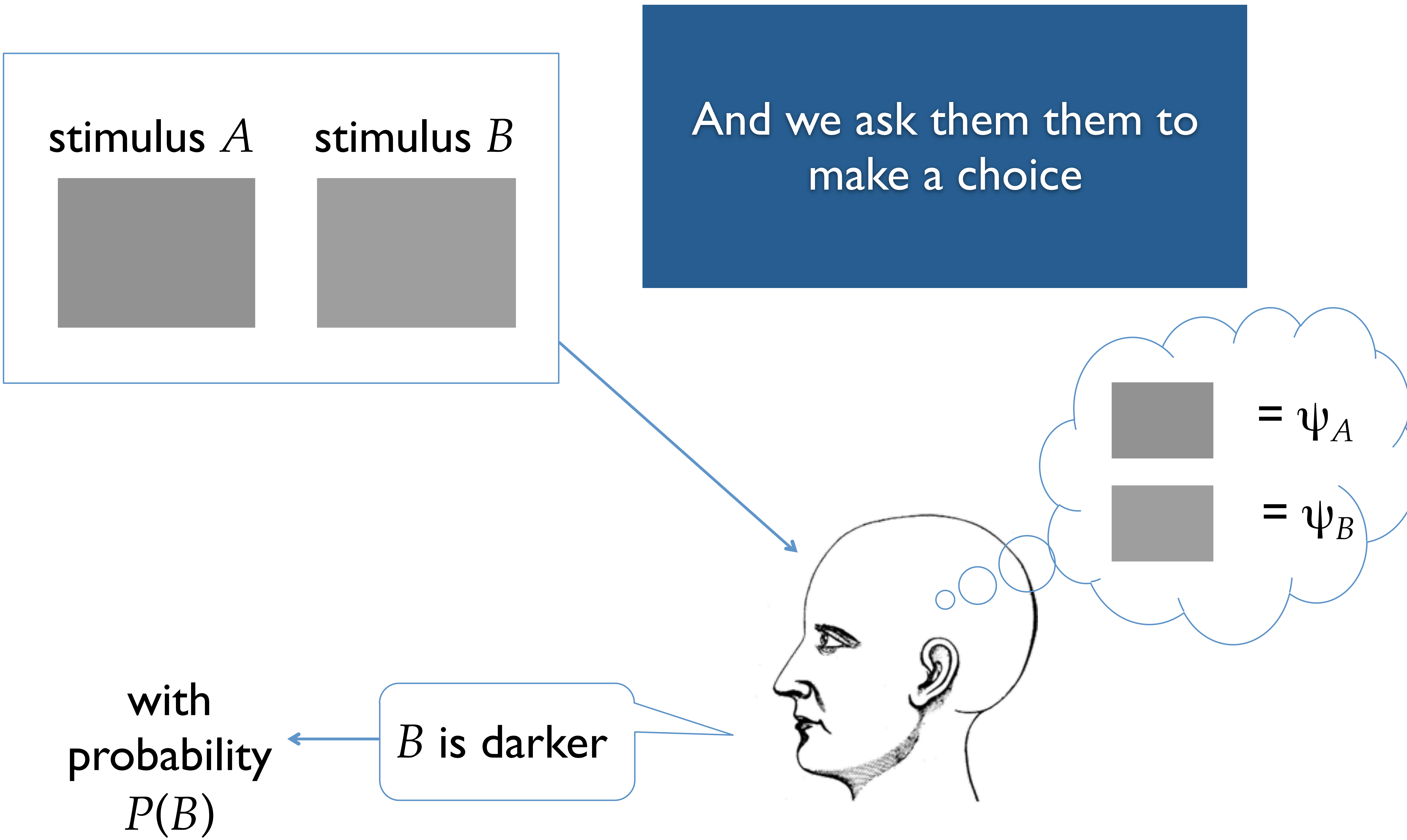


Then we show the stimuli to people



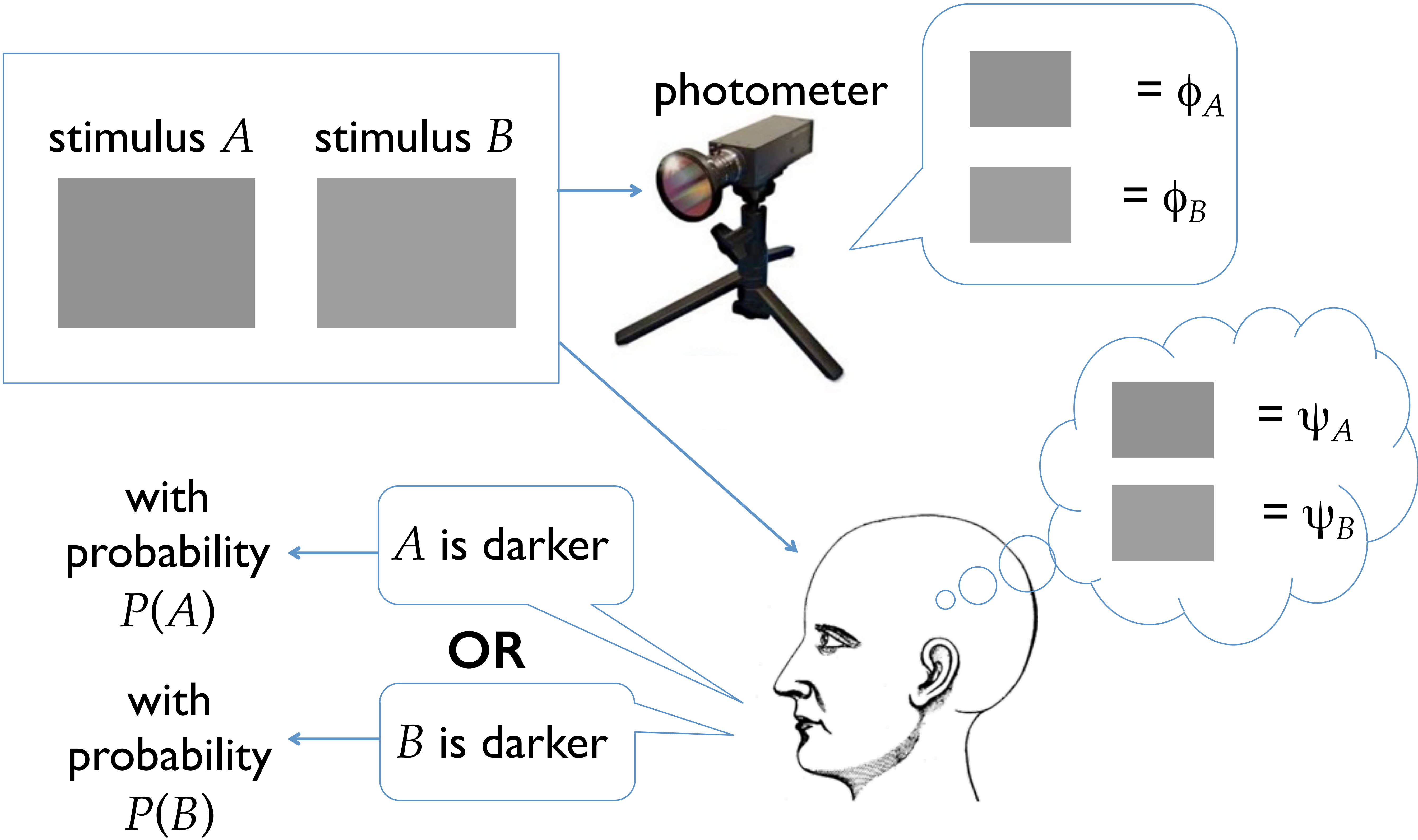
And we ask them them to make a choice





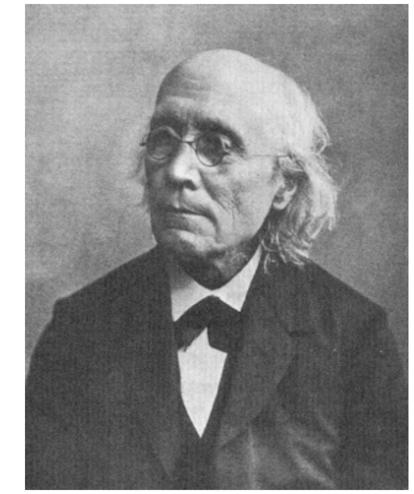
stimulus A
stimulus B
And we ask them them to make a choice
$= \psi_A$
$= \psi_B$
with probability $P(B)$
B is darker

The whole set up in one slide…



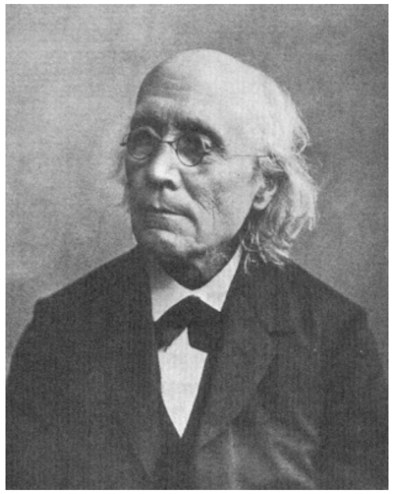
How do you analyse the data?
An introduction to signal detection theory

Fechner's analysis

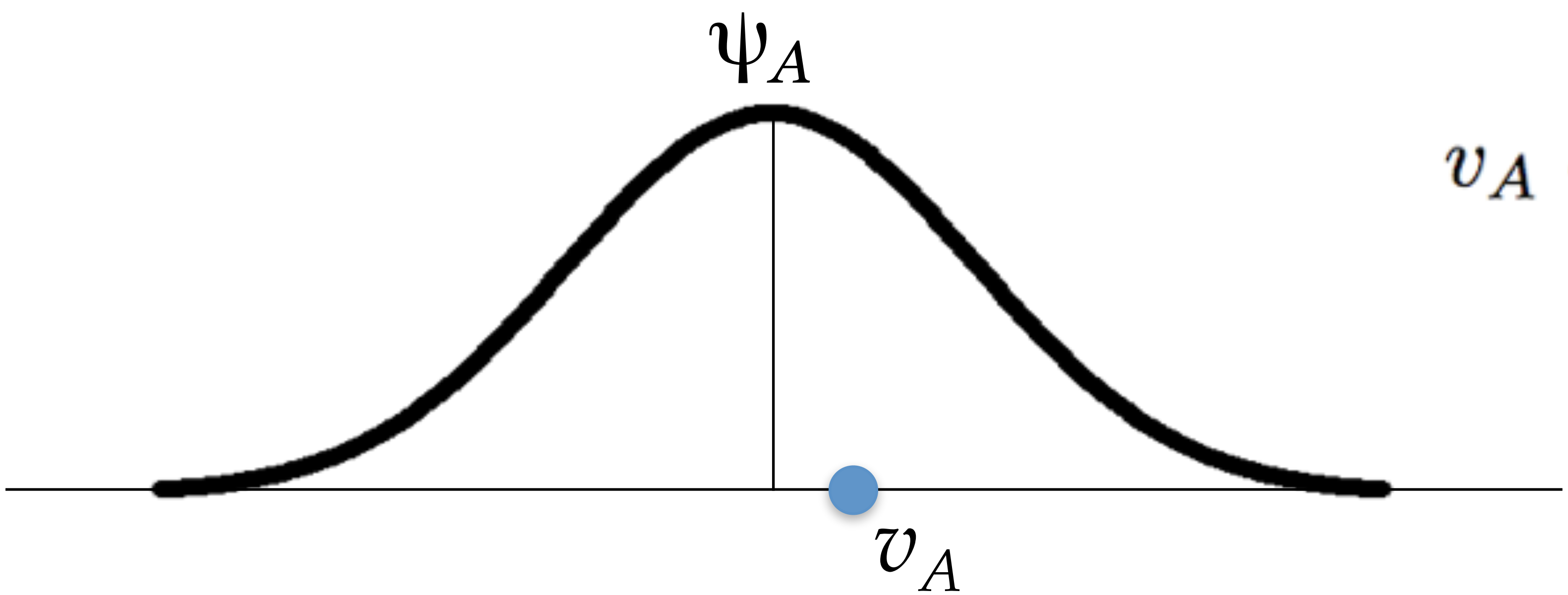


- Visual perception is noisy
  - The "subjective impression" fluctuates from moment to moment, so ψ_A is actually the <u>mean</u> of some distribution over "momentary experiences" v_A

Fechner's analysis

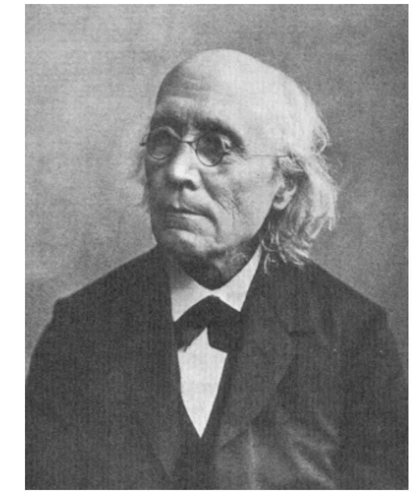


- Visual perception is noisy
  - The "subjective impression" fluctuates from moment to moment, so ψ_A is actually the <u>mean</u> of some distribution over "momentary experiences" v_A



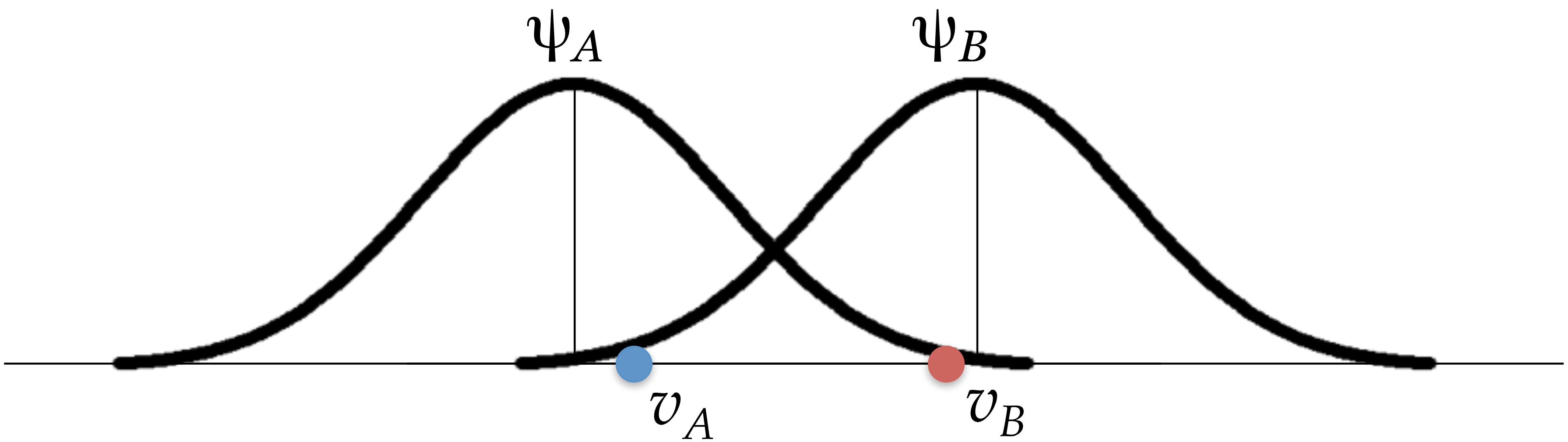
$$v_A \sim \text{Normal}(\psi_A, {\sigma_A}^2)$$

Fechner's analysis



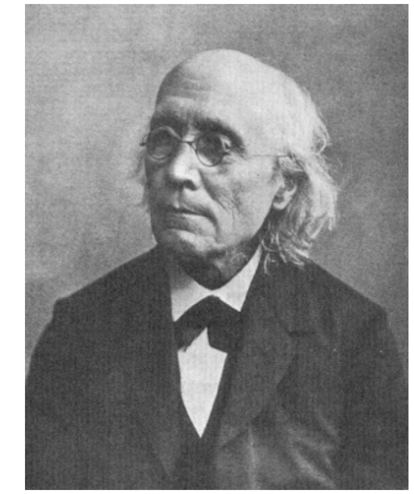
Both stimuli define distributions over subjective experiences

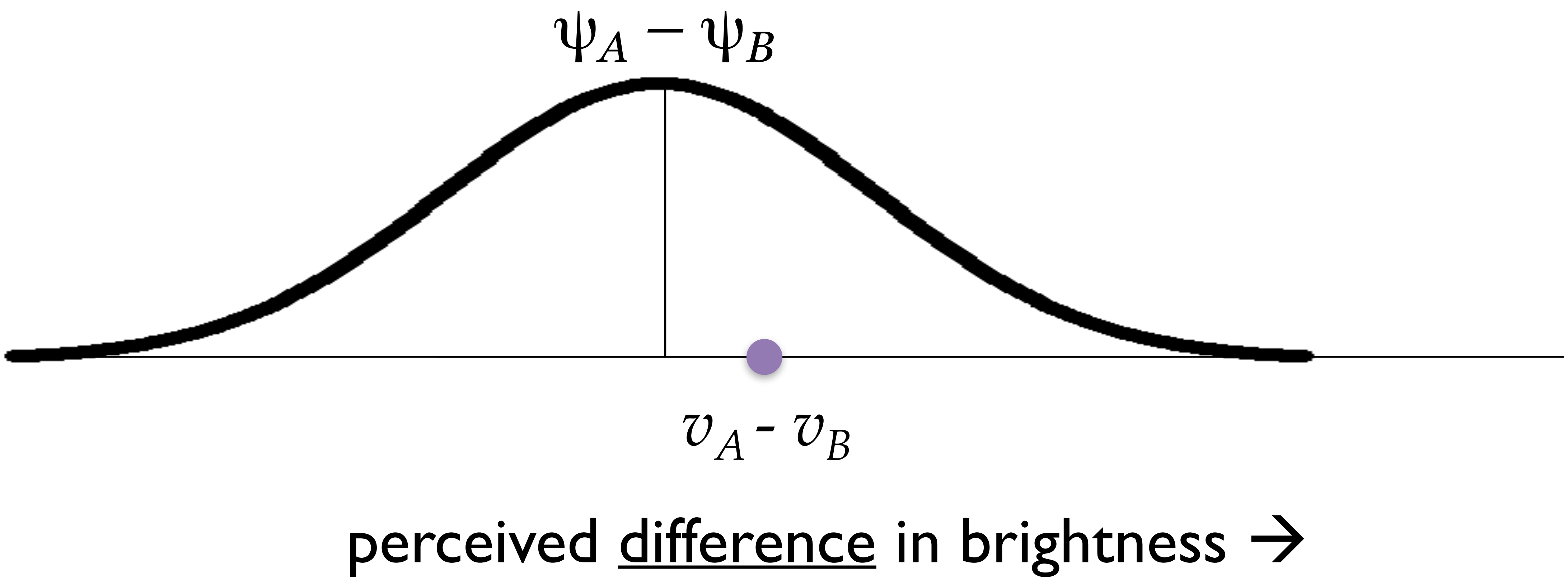
$$v_A \sim \mathrm{Normal}(\psi_A, {\sigma_A}^2)$$
$$v_B \sim \mathrm{Normal}(\psi_B, {\sigma_B}^2)$$



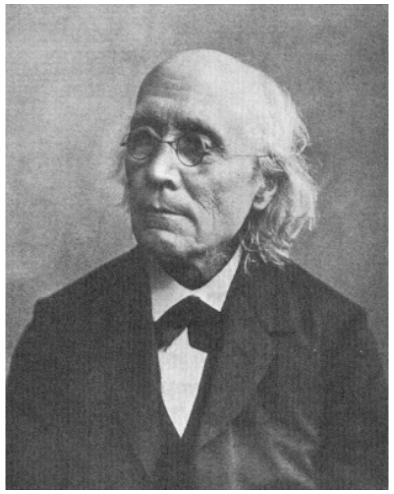
perceived brightness →

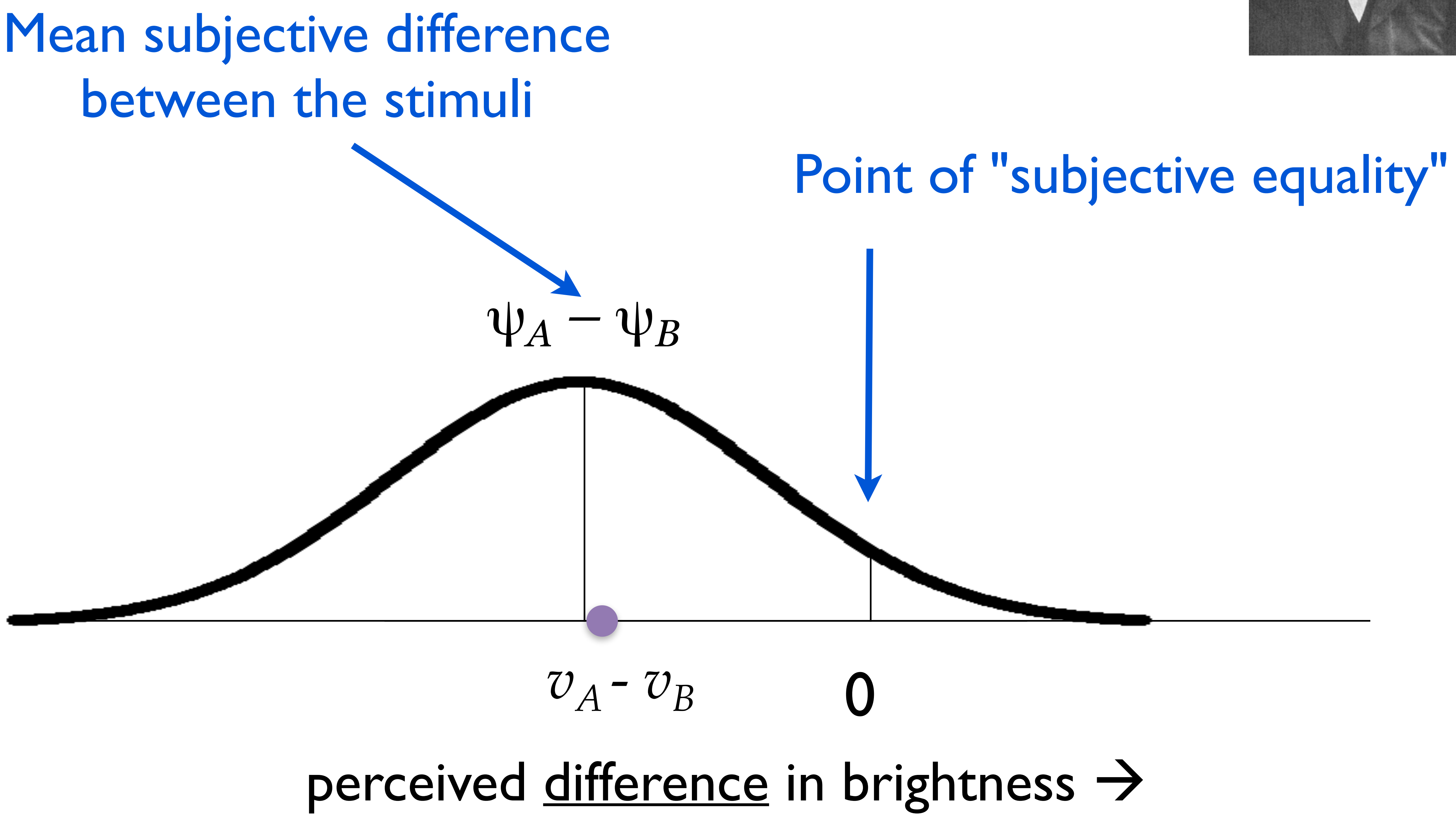
The <u>subjective difference</u> between the two stimuli is v_A - v_B, and is also associated with a distribution



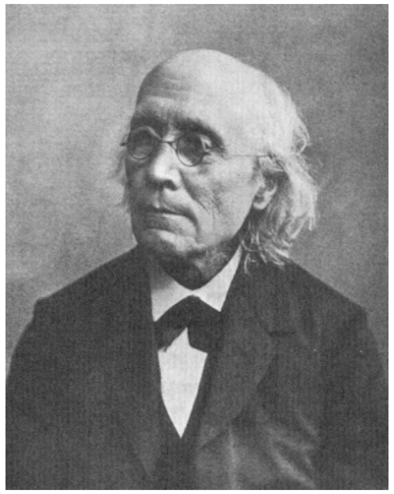


The important point…

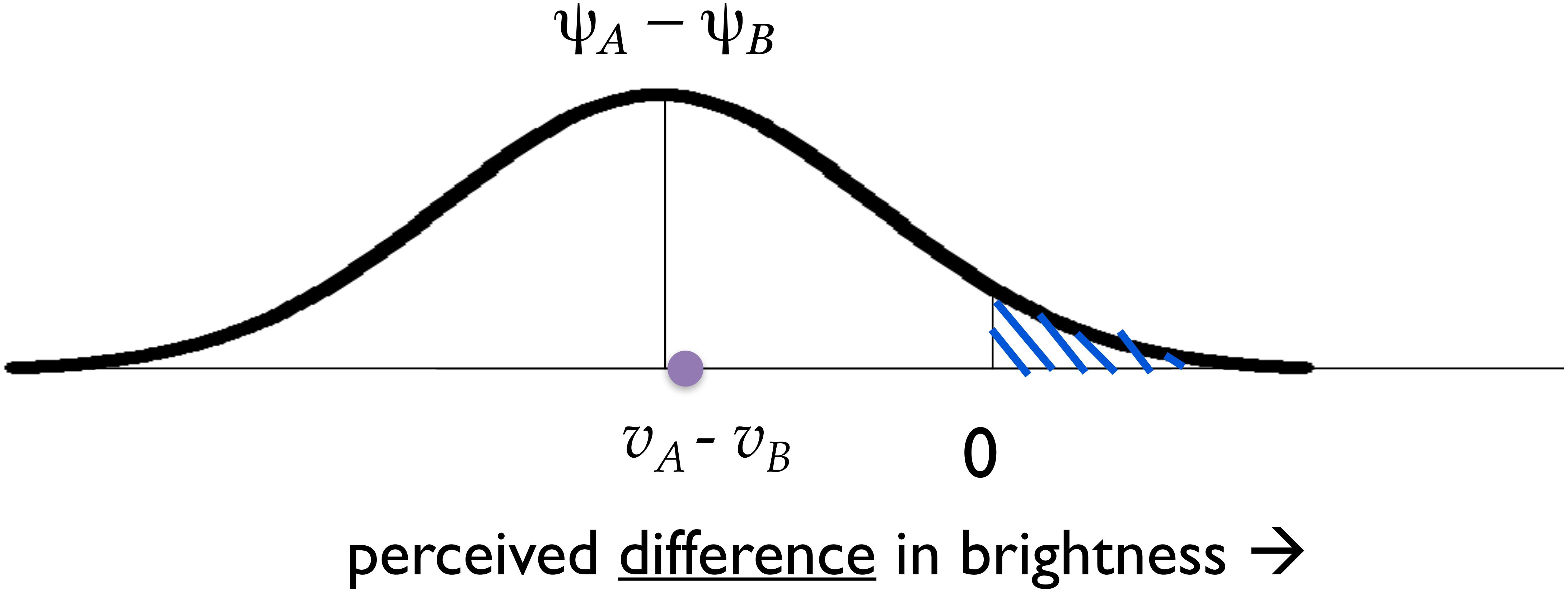




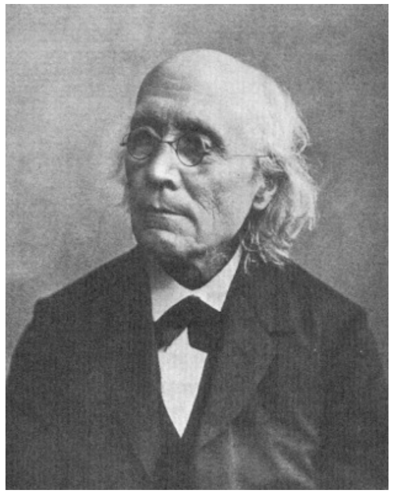
Fechner's analysis



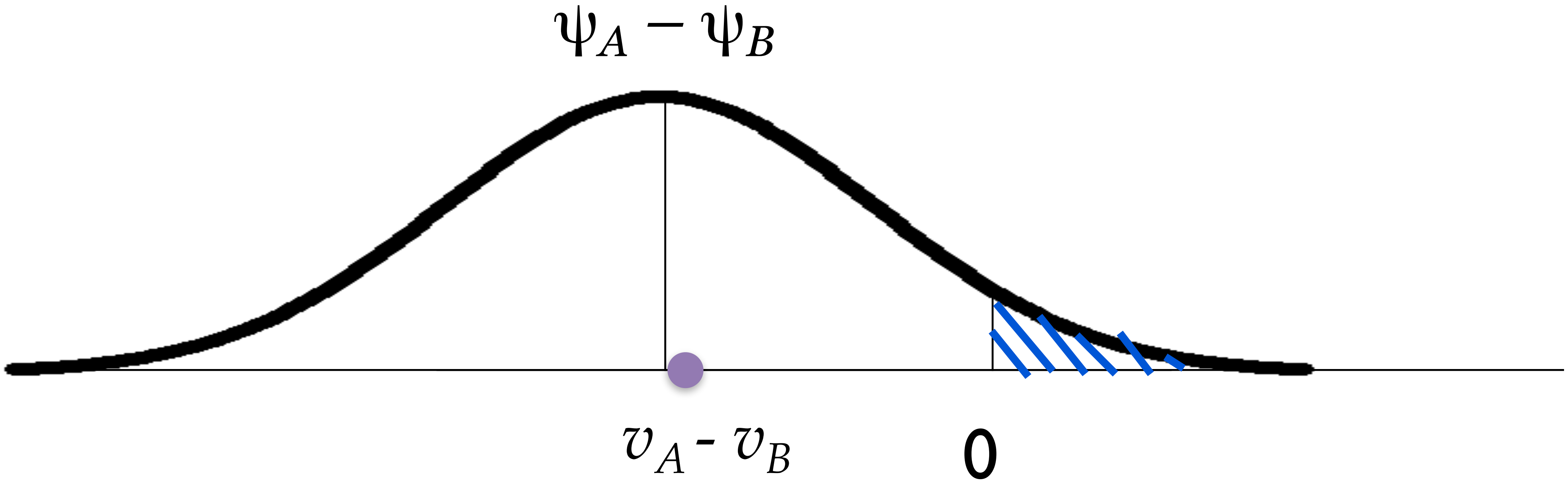
Area under the curve gives the probability that the subjective difference is greater than zero... i.e., the probability of choosing A



Fechner's analysis



This area is given by the cumulative distribution function (CDF) of a normal distribution

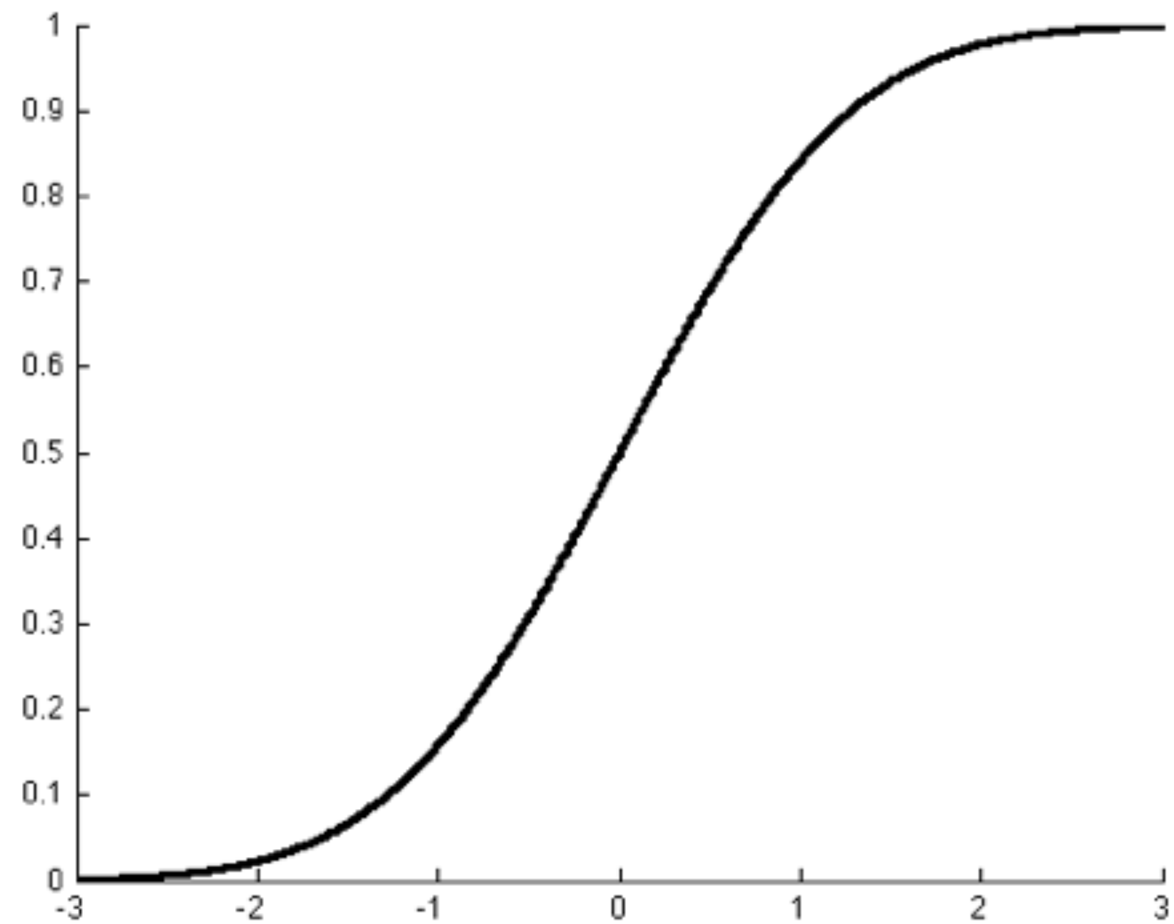


perceived <u>difference</u> in brightness →

The decision model that this implies...

$P(A)$

Probability of choosing A

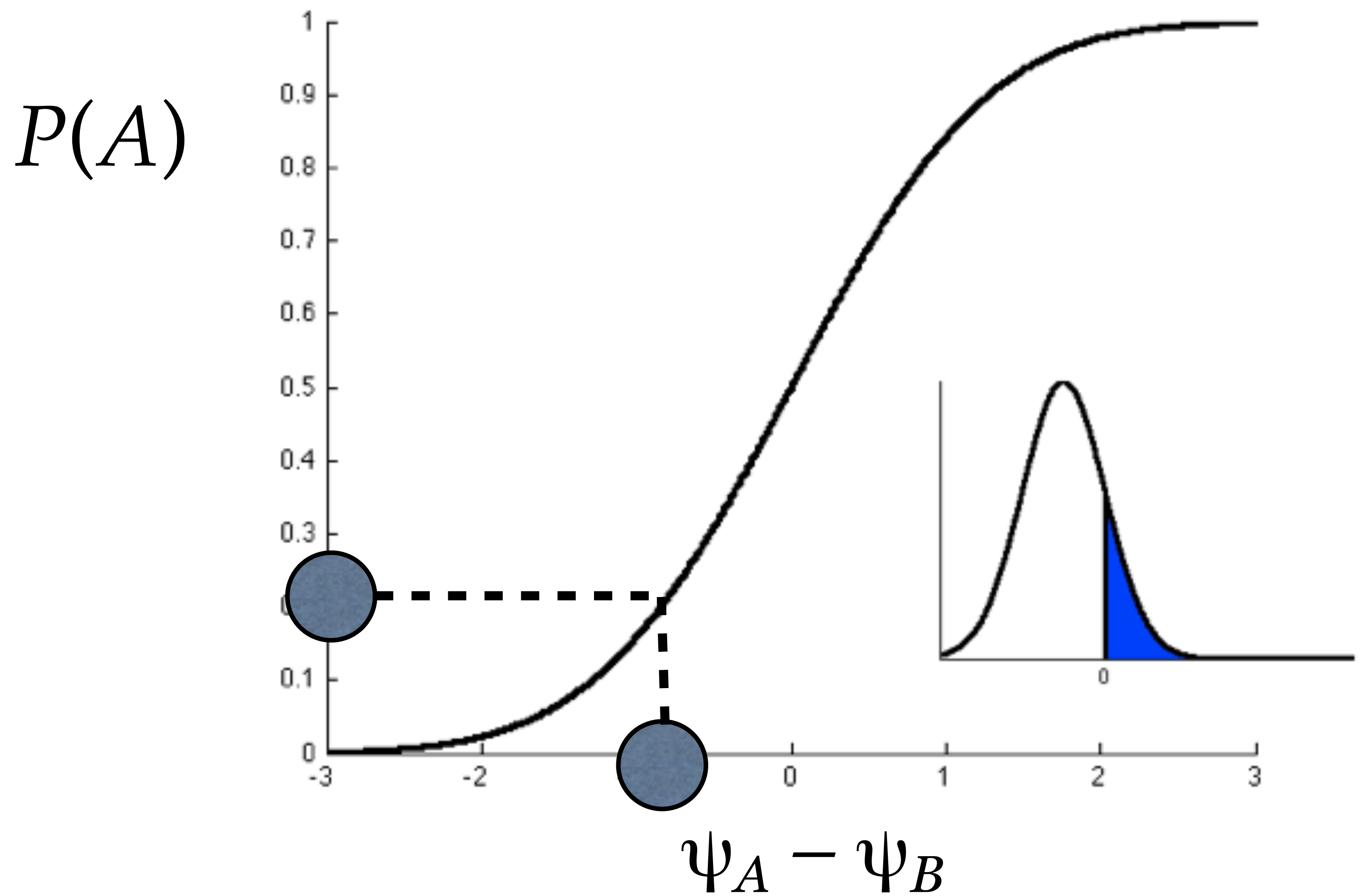


$\psi_A - \psi_B$

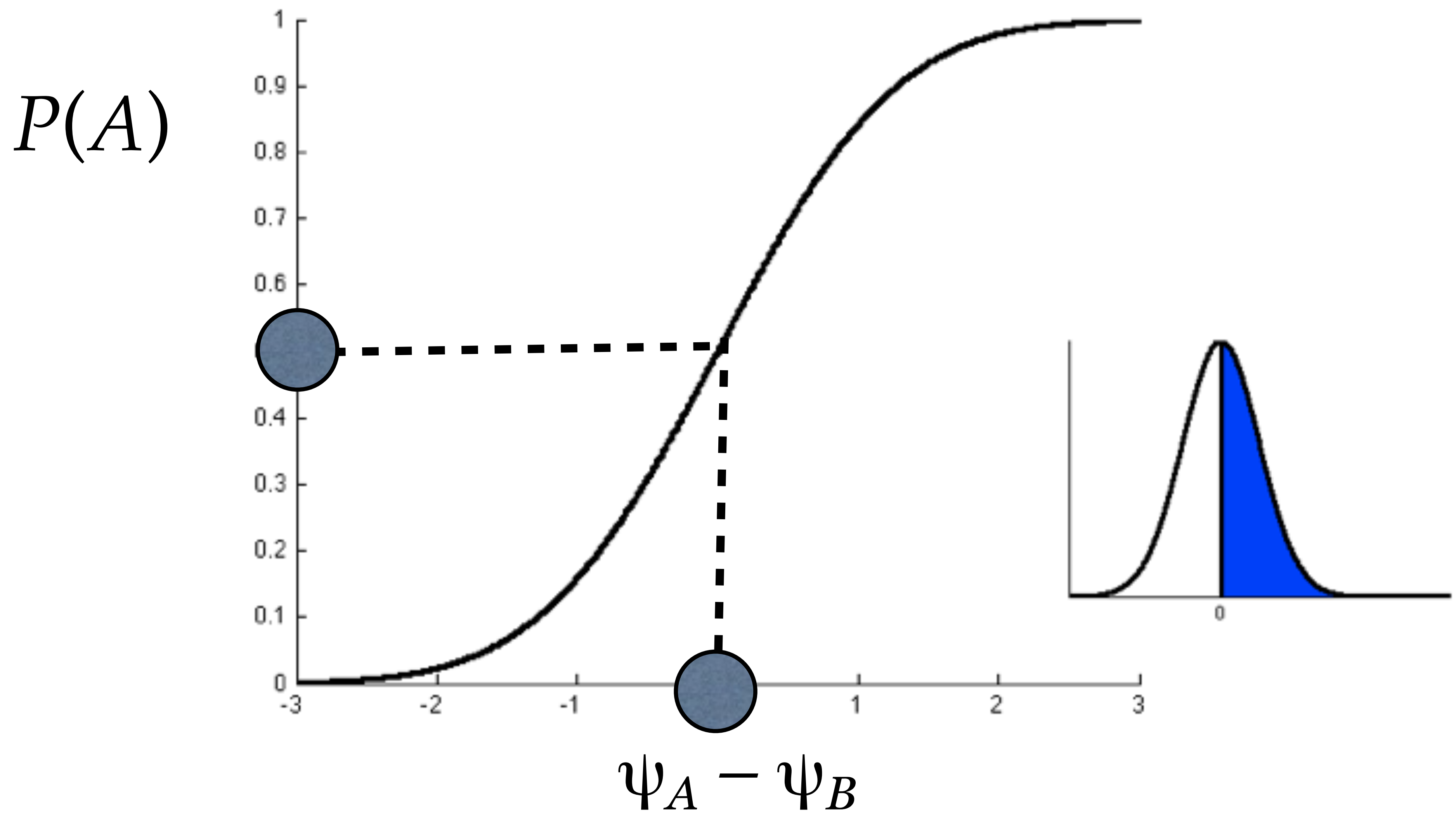
Mean subjective difference

(I'm being a little imprecise here: the slope of this curve depends on how noisy the perceptual system is, but let's ignore that detail for today)

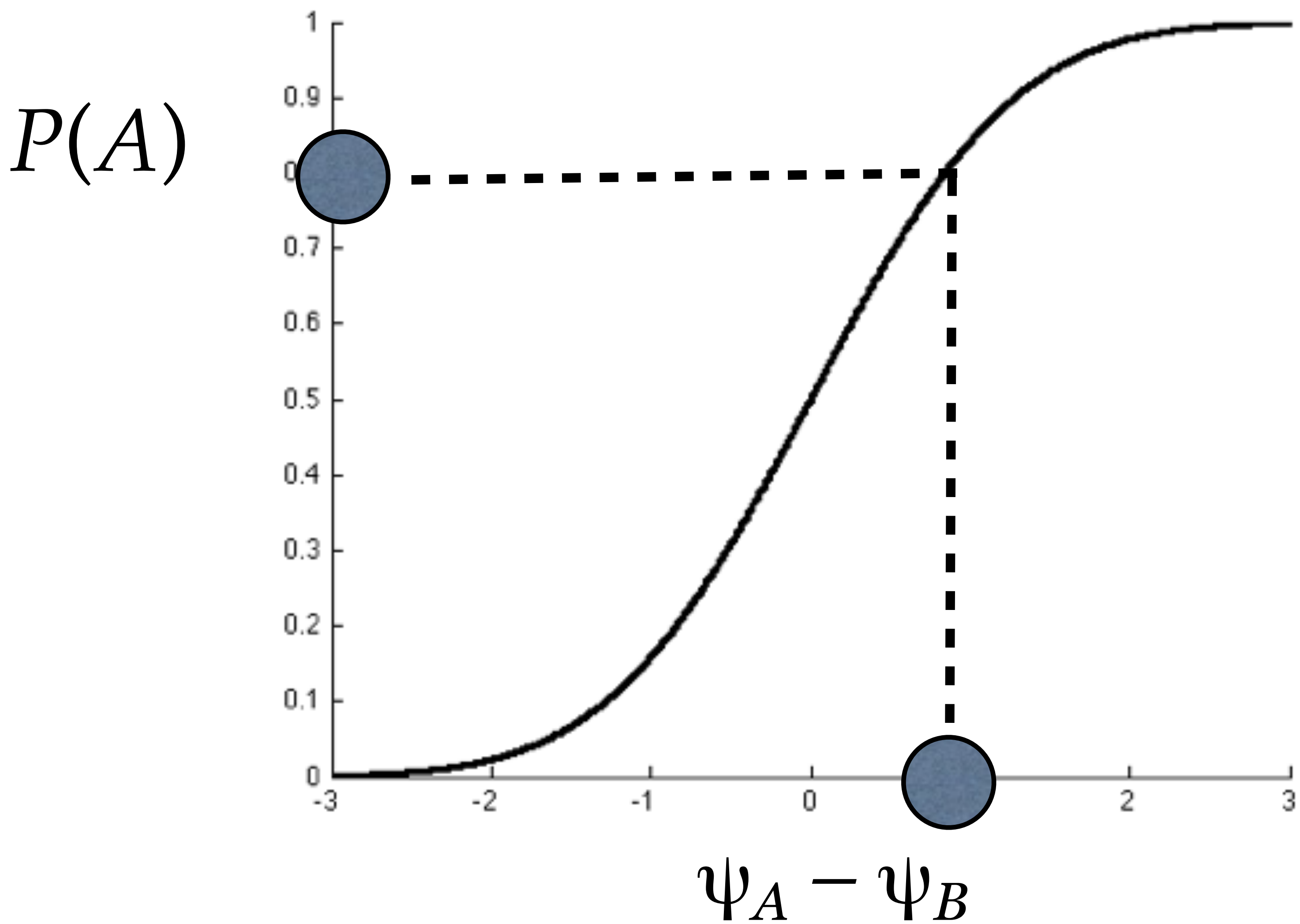
If we know P(A), we can infer $\psi_A - \psi_B$

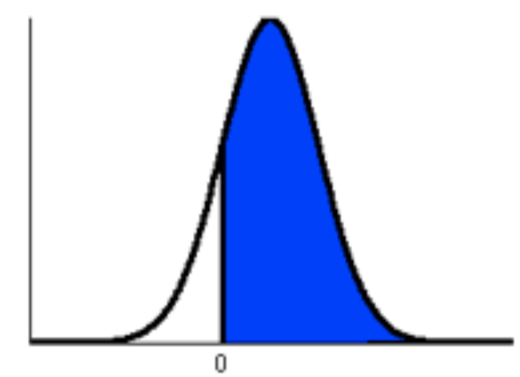


If we know P(A), we can infer $\psi_A - \psi_B$



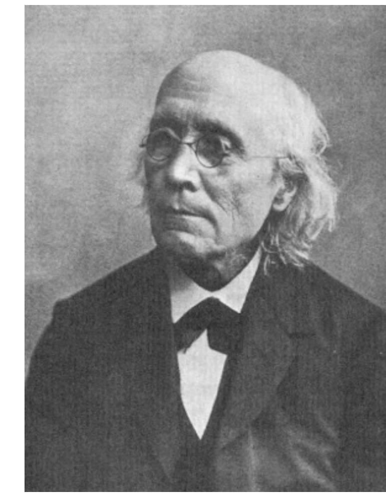
If we know P(A), we can infer $\psi_A - \psi_B$

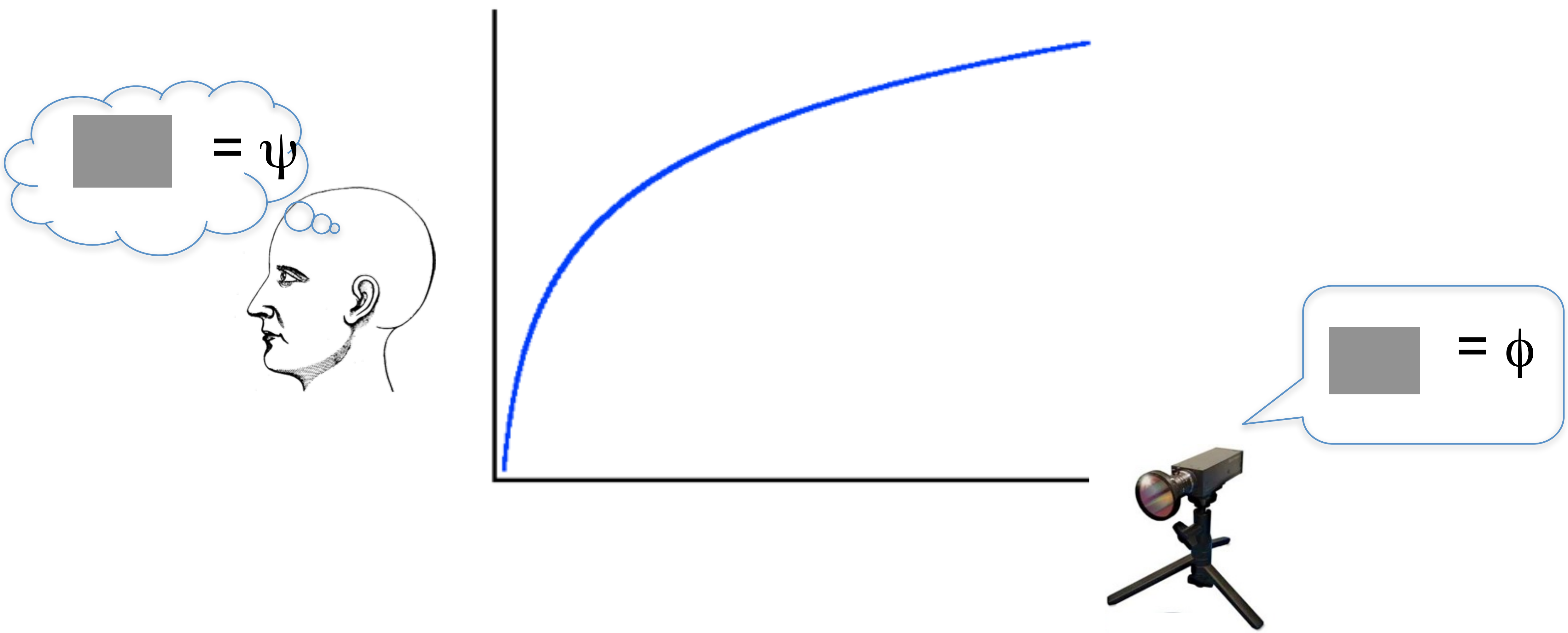




And psychophysics was born

- On the basis of this analysis, Fechner was able to determine that a logarithmic relationship between physical magnitude and subjective experience was best able to explain human choices





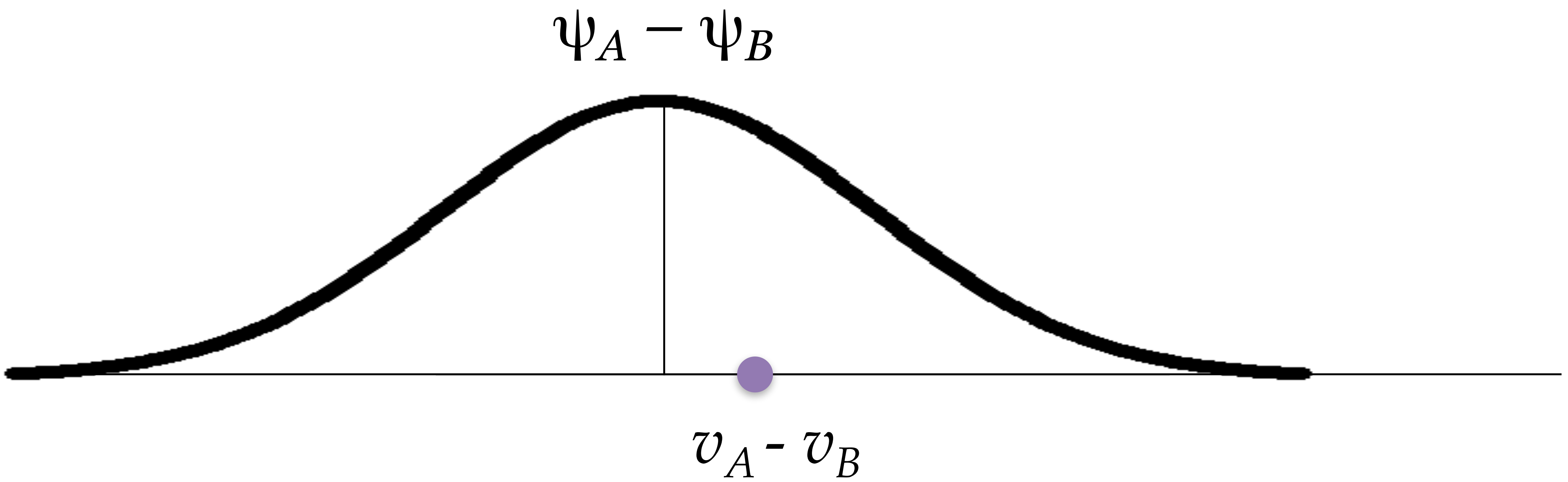
Random utility models, signal detection theory etc

The essential features of Fechner's analysis of human choice behaviour

There is a psychological quantity of interest that guides people's choices

$$\psi_A - \psi_B$$

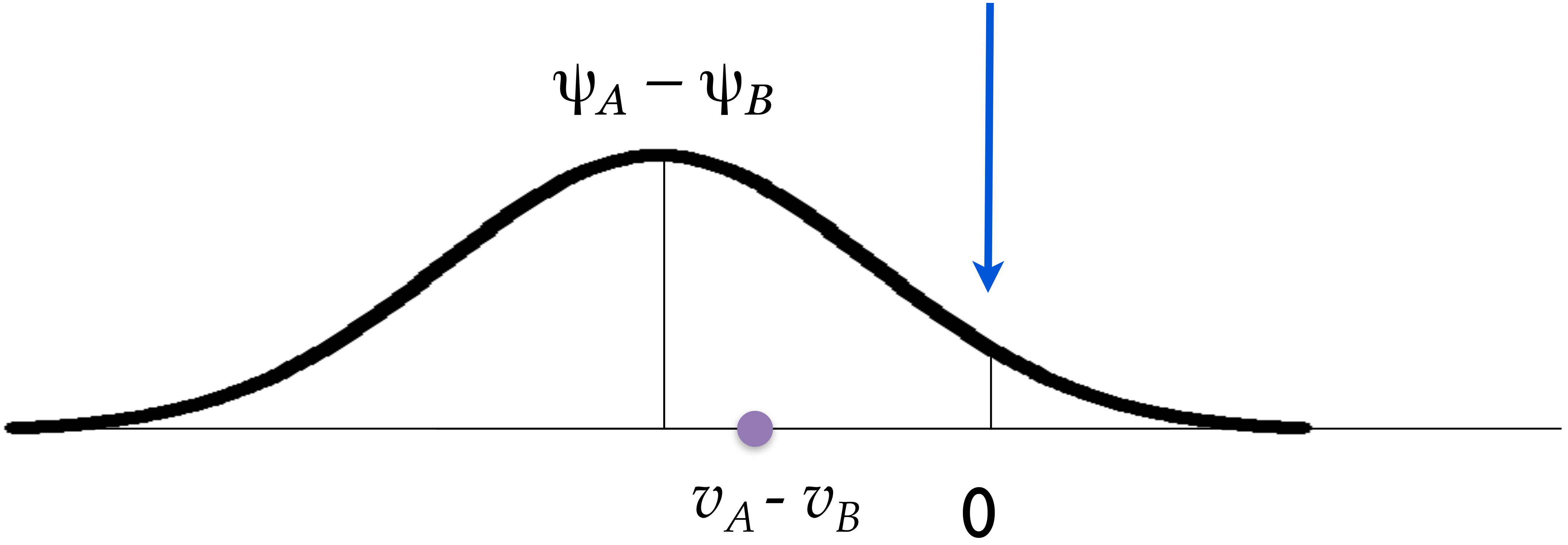
The essential features of Fechner's analysis of human choice behaviour



It defines a probability distribution over subjective experiences

The essential features of Fechner's analysis of human choice behaviour

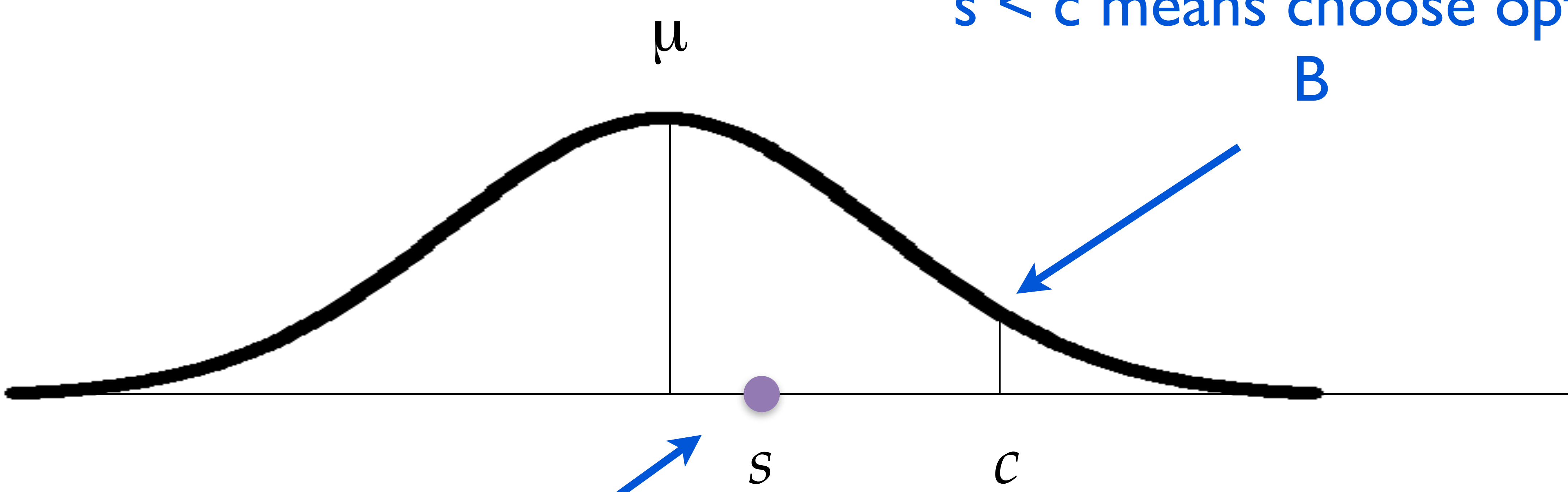
And it is compared to some desired criterion or reference point



Generically...

Quantity of interest that constrains people's choices

The criterion against which s is assessed... s > c means choose option A, whereas s < c means choose option B



The random quantity (sample) that people have access to

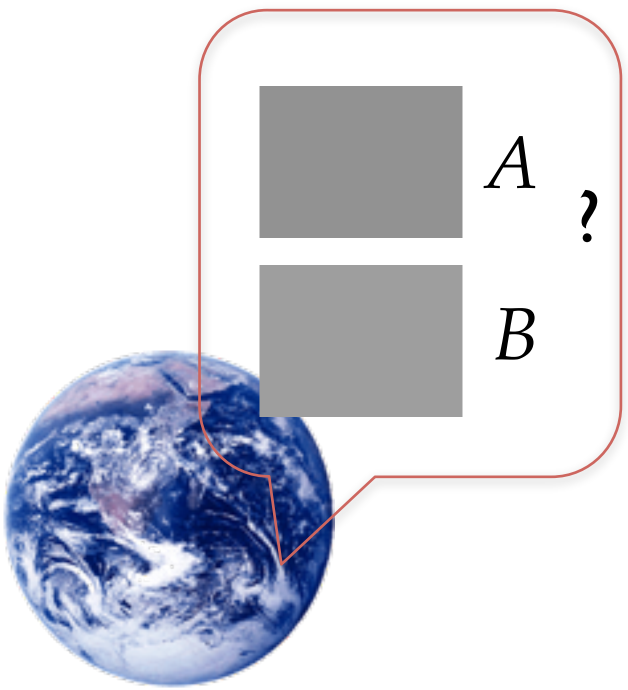
Different names, same thing

- "Signal detection theory"
  - s represents a momentary subjective strength (e.g., feeling of familiarity, feeling of brightness, etc)
  - used a lot throughout cognitive science, especially in memory research

- "Random utility models"
  - s represents the current utility of a particular option (e.g. product you want to buy) that people might want
  - used a lot in economics

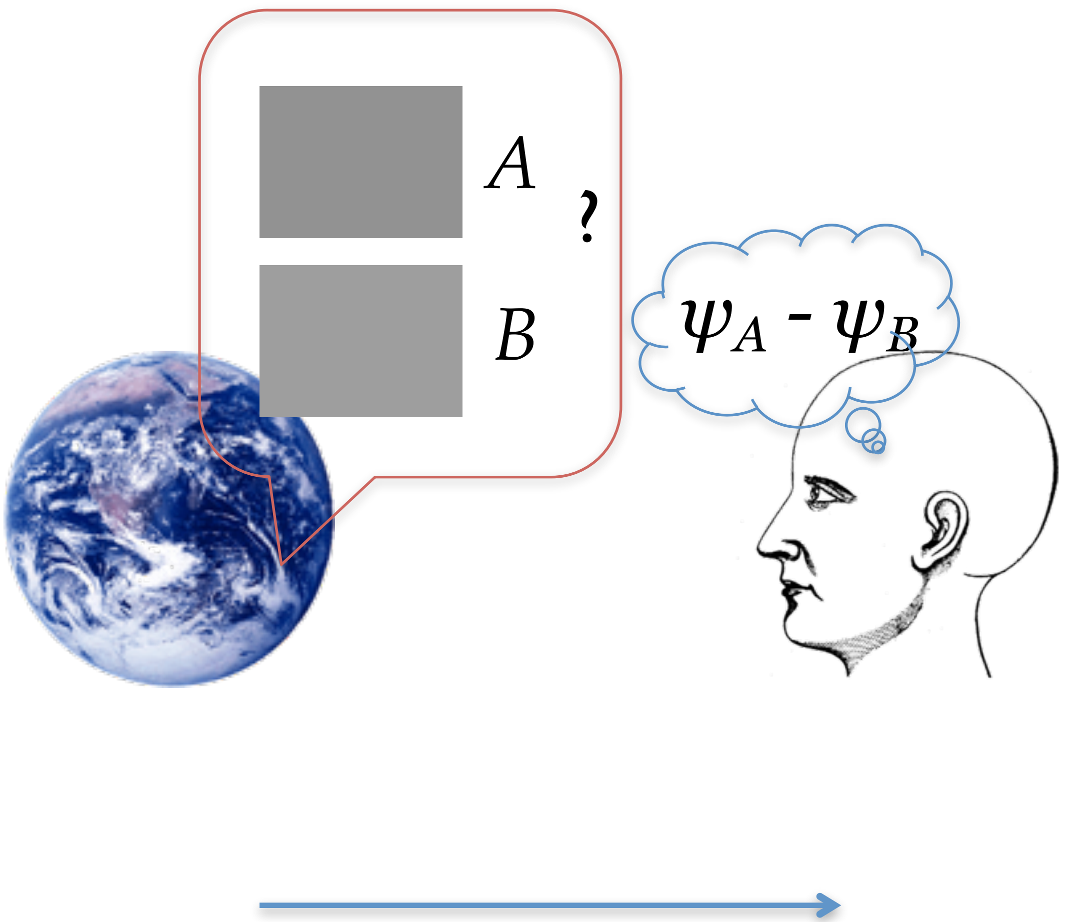
The big picture

The overall decision process



The world presents the **options**

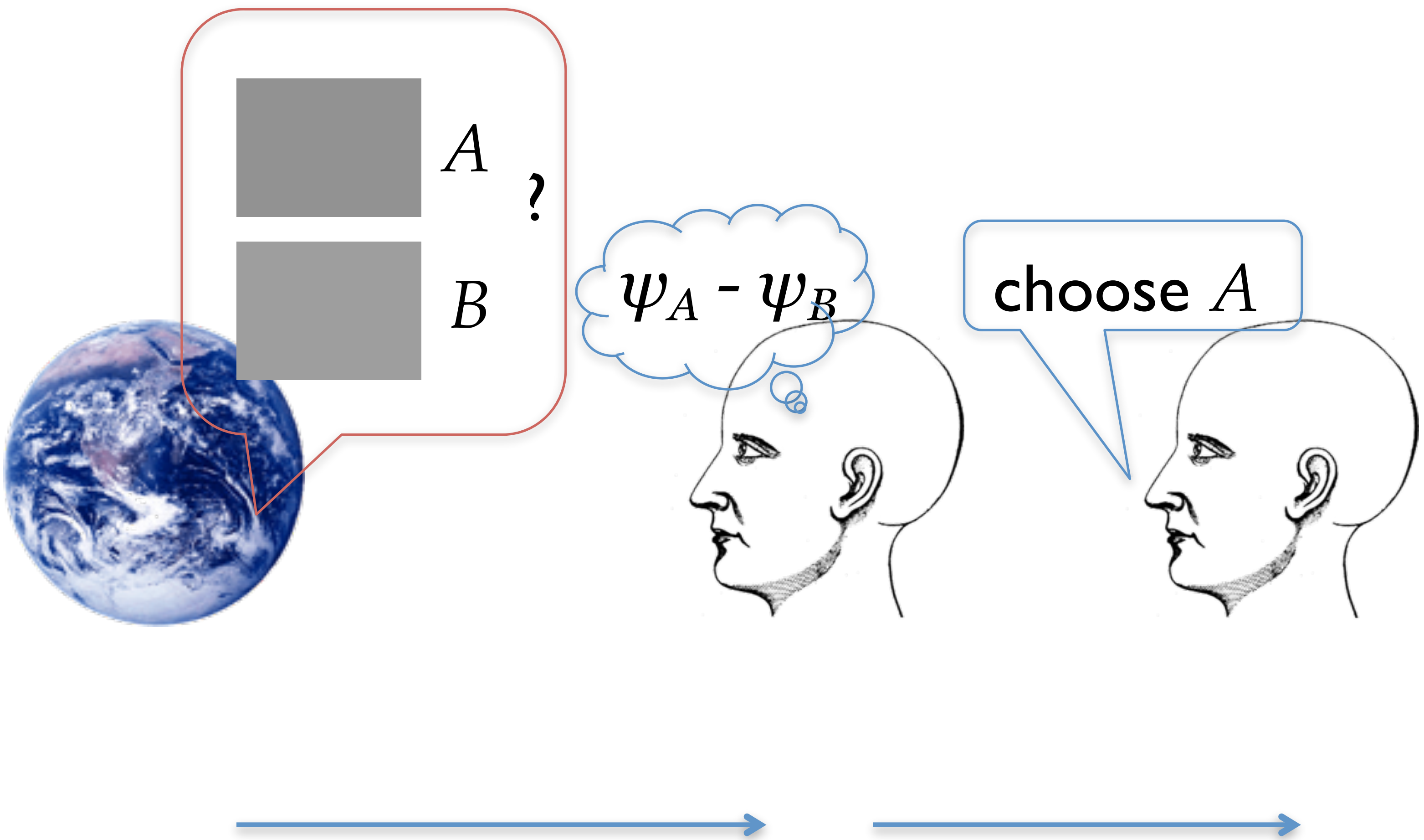
The overall decision process



The world presents the **options**

The decision-maker assesses the **options**

The overall decision process

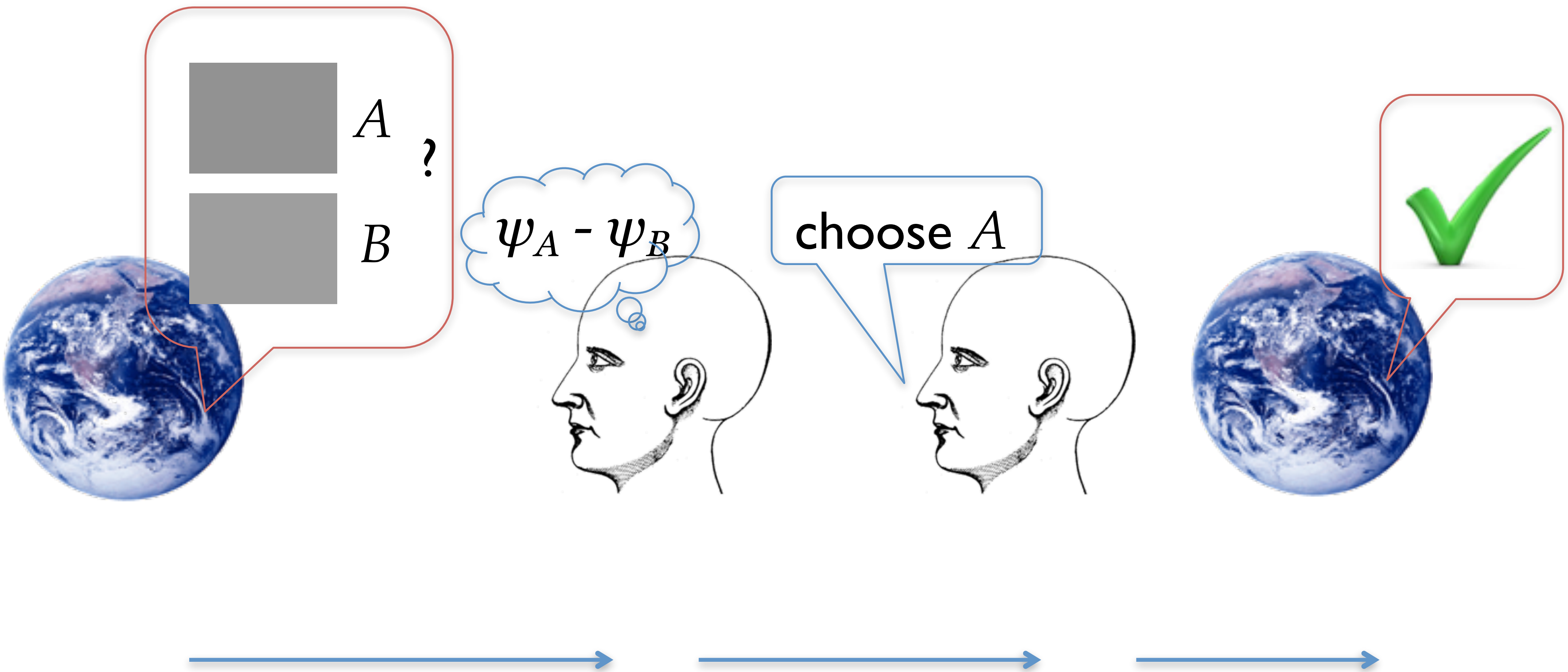


The world presents the **options**

The decision-maker assesses the **options**

An **option** is selected by the decision-maker

The overall decision process



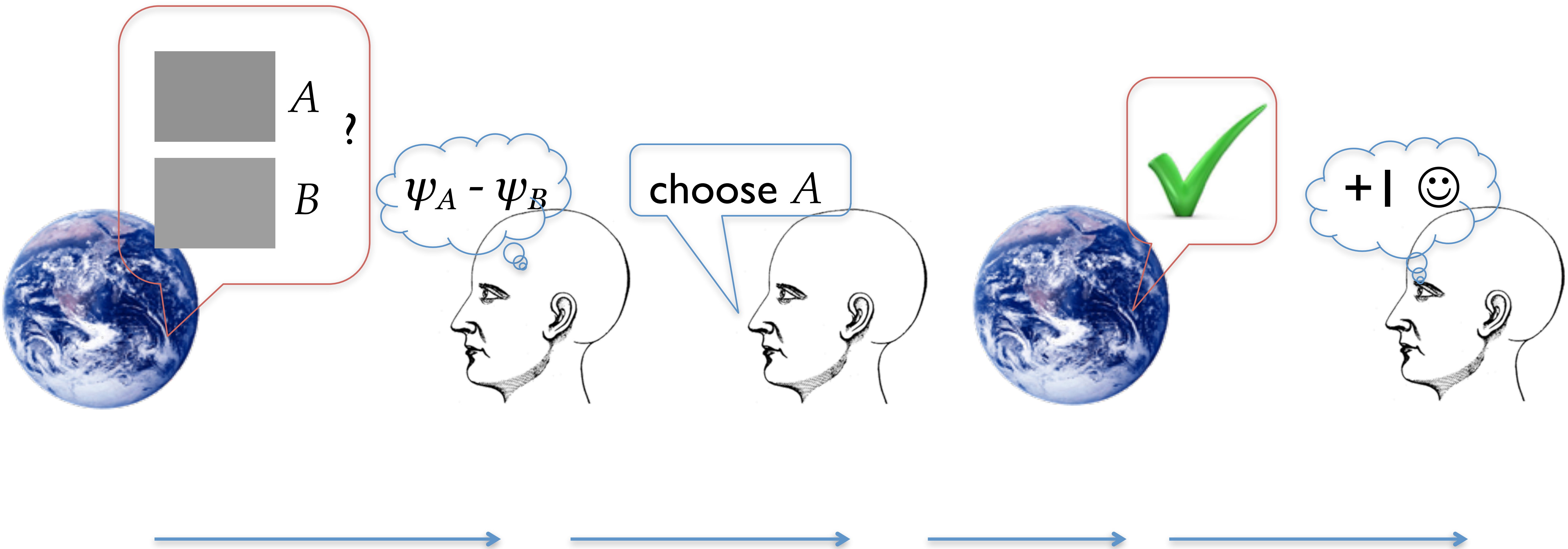
The world presents the **options**

The decision-maker assesses the **options**

An **option** is selected by the decision-maker

The world generates the **outcomes**

The overall decision process



The world presents the **options**

The decision-maker assesses the **options**

An **option** is selected by the decision-maker

The world generates the **outcomes**

The decision-maker gets some **utility**

The overall decision process

The "utilities" are pretty simple here, so EU theory and prospect theory are in agreement

The world presents the **options**

The decision-maker assesses the **options**

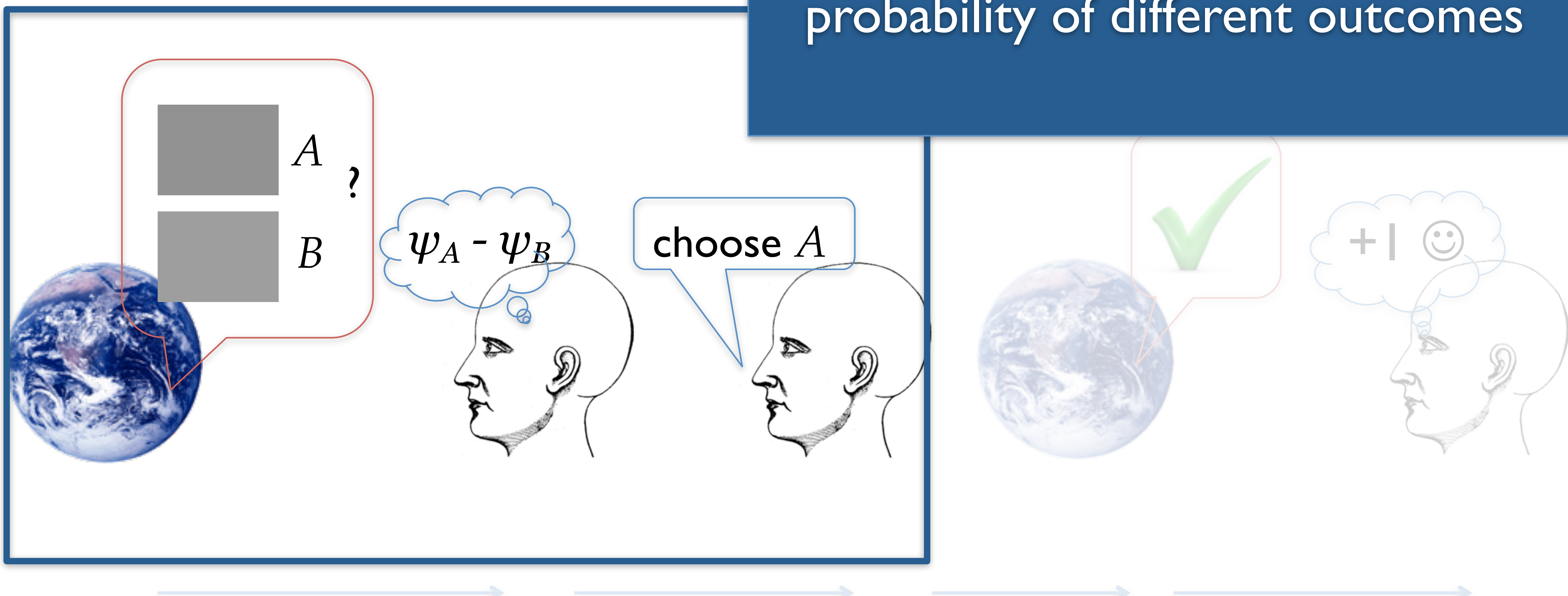
An **option** is selected by the decision-maker

The world generates the **outcomes**

The decision-maker gets some **utility**

The overall c

What we've been doing is developing a theory for how people assess the probability of different outcomes



The world presents the **options**

The decision-maker assesses the **options**

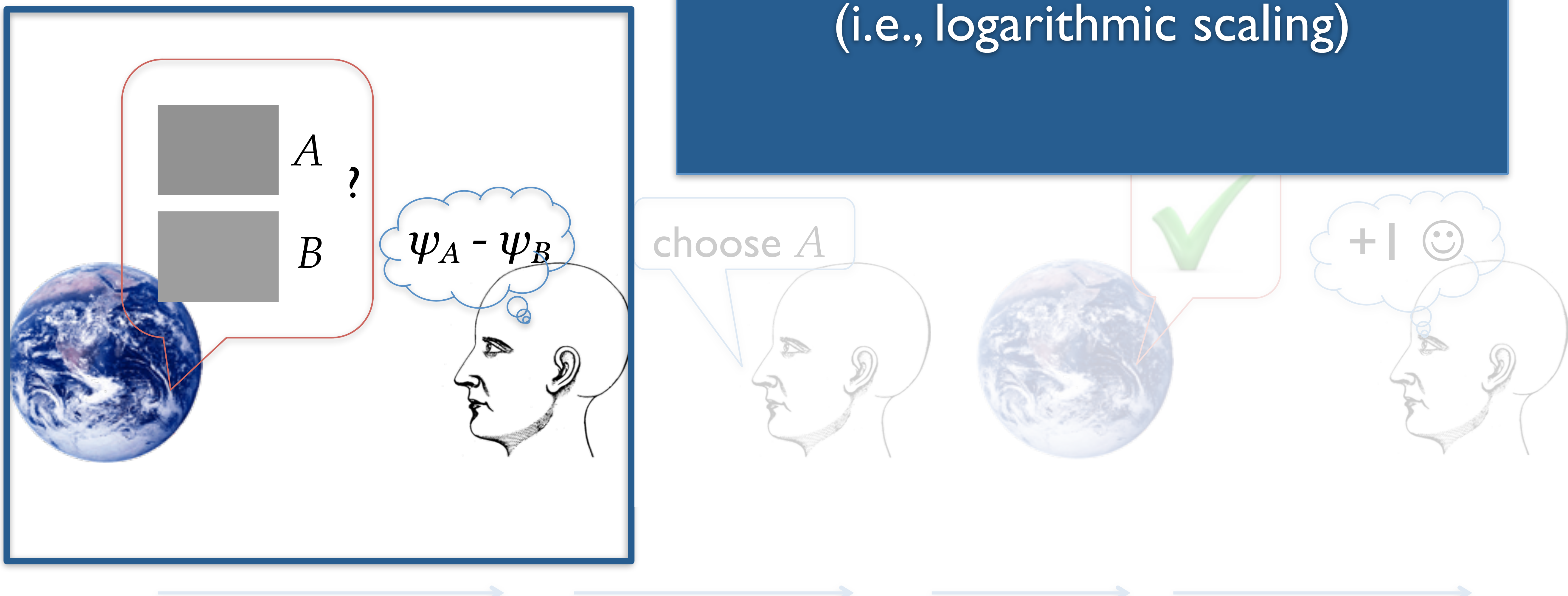
An **option** is selected by the decision-maker

The world generates the **outcomes**

The decision-maker gets some **utility**

The overall decision process

The psychophysical part
(i.e., logarithmic scaling)



The world presents the **options**

The decision-maker assesses the **options**

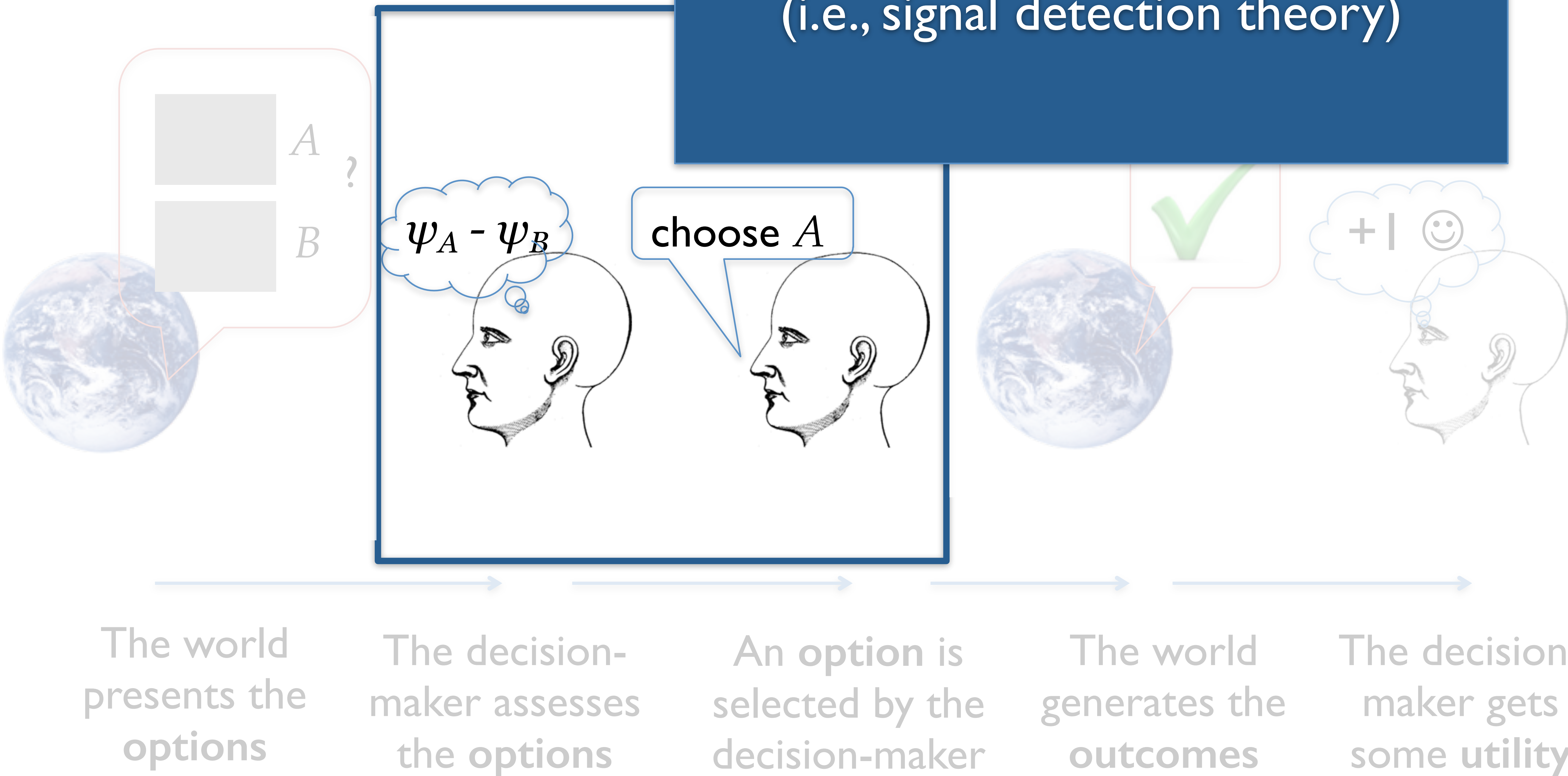
An **option** is selected by the decision-maker

The world generates the **outcomes**

The decision-maker gets some **utility**

The overall decision process

The "choice model"
(i.e., signal detection theory)



Sequential sampling models

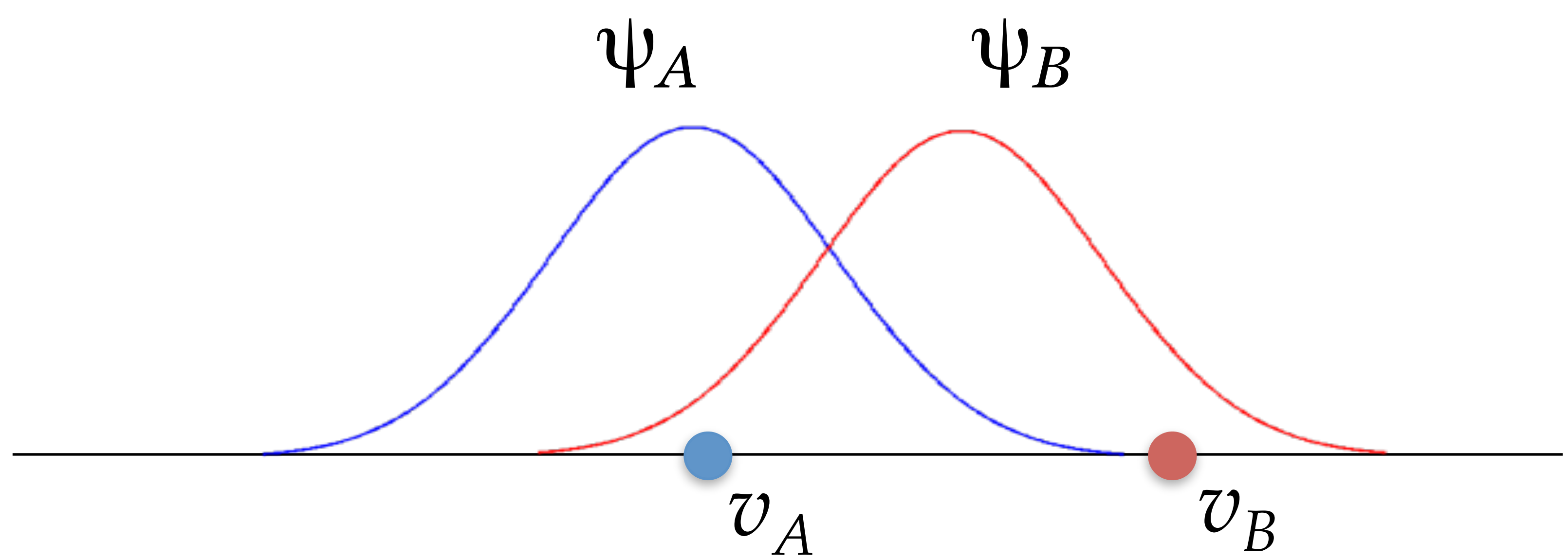
Time is money, and computation isn't free

- Suppose we were to try to account for human decision making using a combination of signal detection theory and expected utility theory.
- This will not work (not without modification)

- There are two big flaws here:
  - Decision making processes take time
  - Decision making processes require computation
  - Neither one is free.

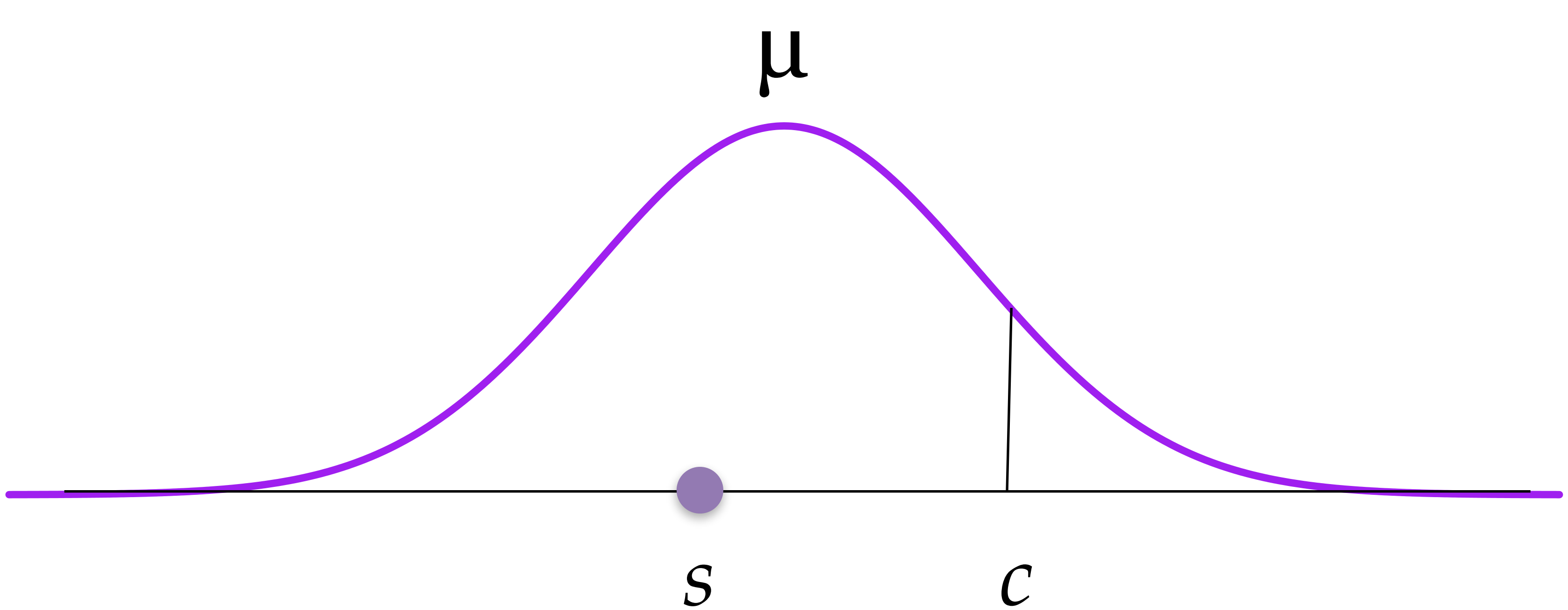
Let's be a little more precise now

- The decision process:
  - Draw one sample v_A from the blue distribution
  - Draw one sample v_B from the red distribution
  - If $v_A > v_B$, choose A
  - If $v_A < v_B$, choose B



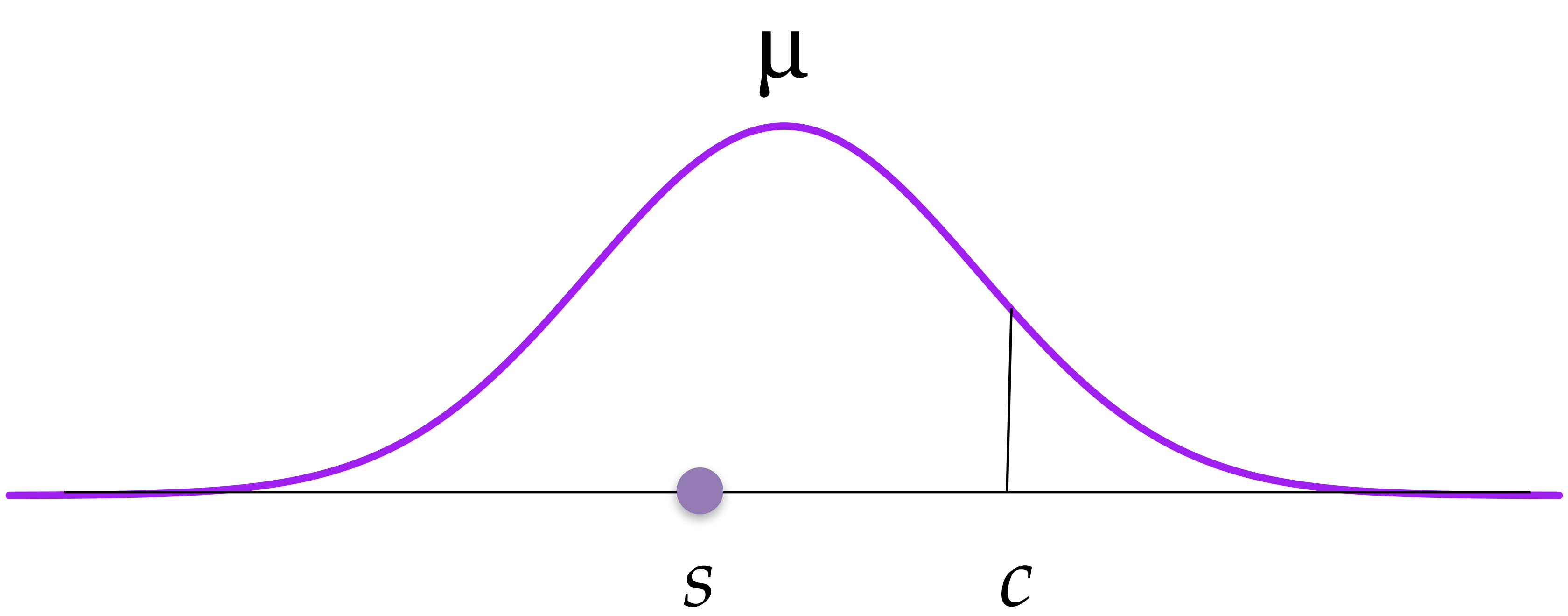
Or, equivalently

- The decision process:
  - Set a criterion c
  - Draw one sample s from the purple distribution
  - Choose option A if s > c
  - Choose option B if s < c



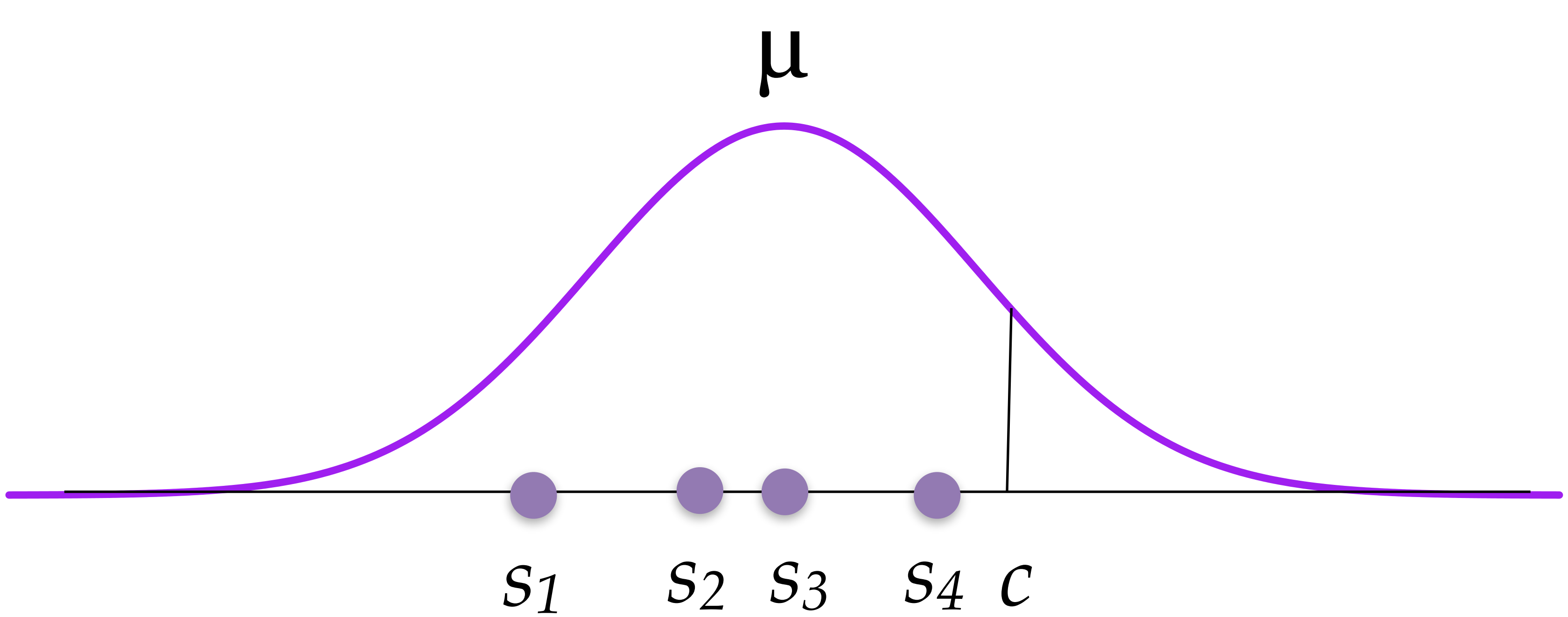
The big question

Why only a single sample?

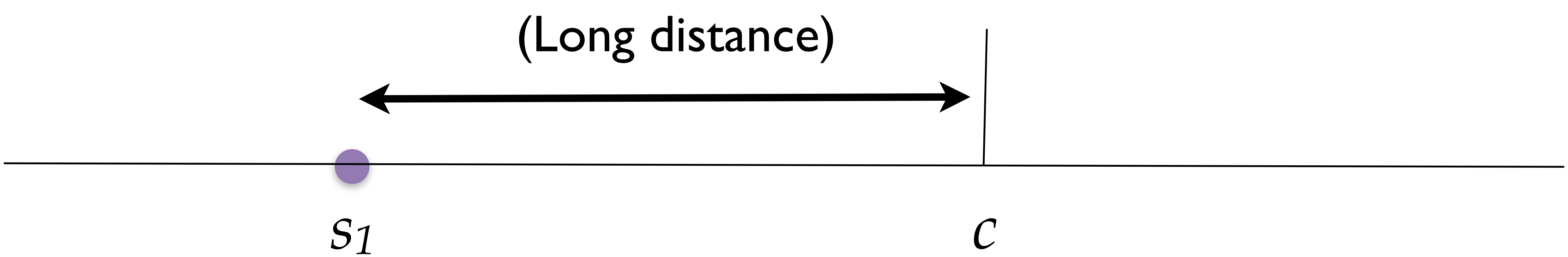


The big question

If the goal is to infer whether μ < c, then multiple samples will provide more evidence, since the decision maker will have much more accurate knowledge of μ

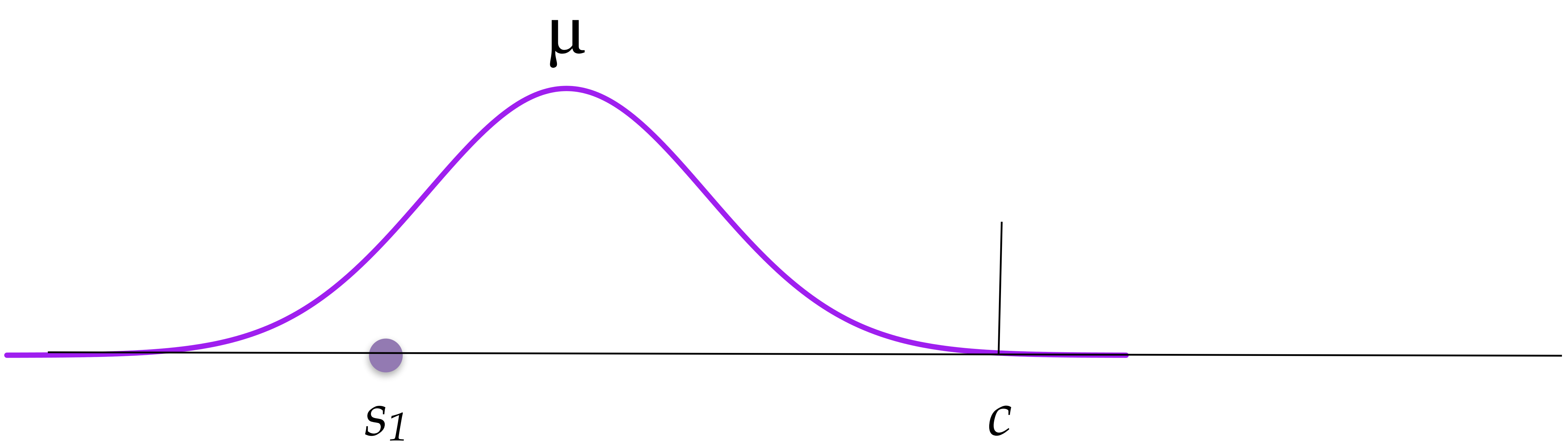


How many samples? The decision-maker's perspective



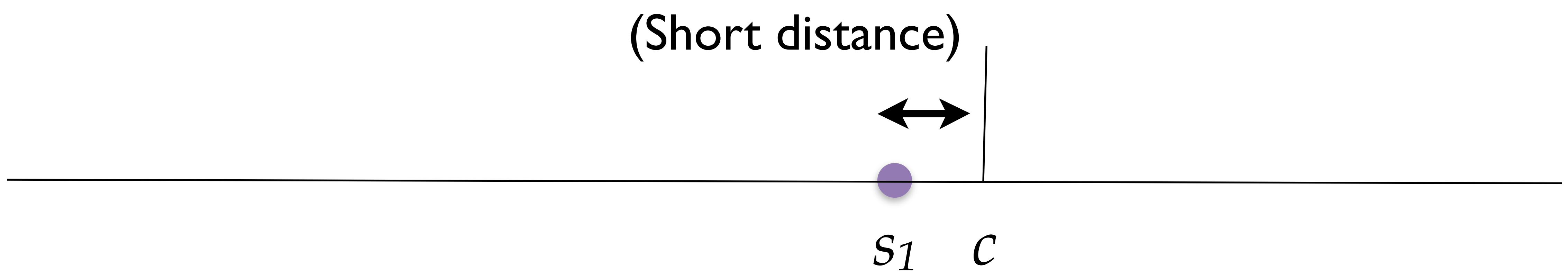
In this situation, it feels like s_1 provides very strong evidence that $\mu < c$, so we only NEED one sample

How many samples? The decision-maker's perspective



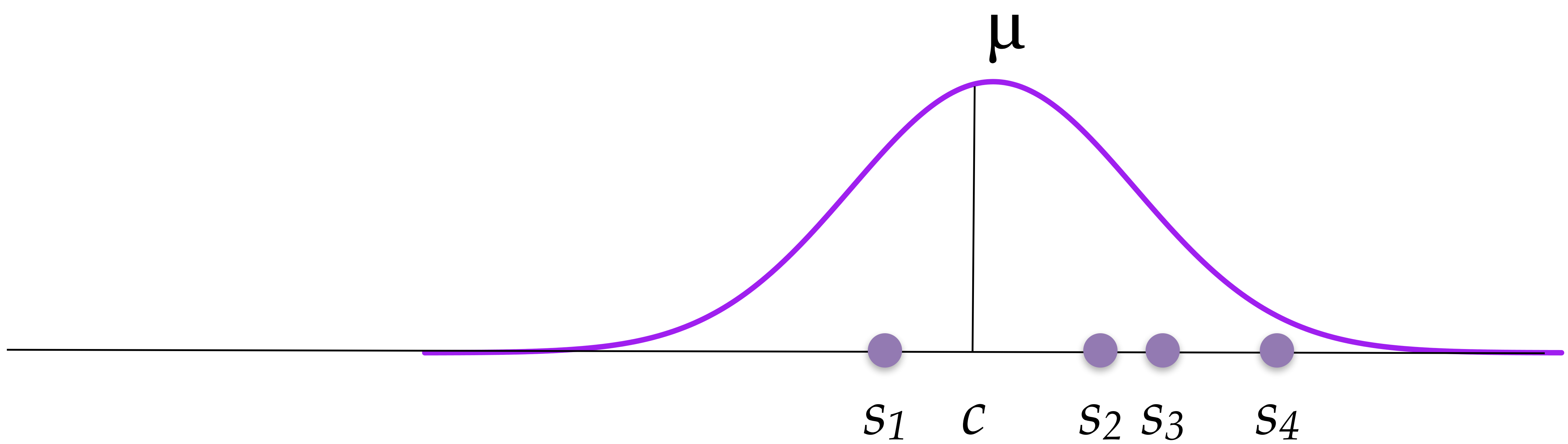
And that would be correct

How many samples? The decision-maker's perspective



But in this situation it feels like you might need more than one data point to justify making your decision

How many samples? The decision-maker's perspective



And that's also true

A computational analysis

- Utility function is just +1☺ for a correct decision
- So maximising utility means making the correct decision, so that you have the greatest possible chance of getting the +1☺.

$$
\begin{aligned}
N^* &= \arg\max_N P(\text{correct decision} \mid N \text{ samples}) \\
&= \infty
\end{aligned}
$$

- So the decision-maker should collect an infinite number of samples.

A computational analysis

- Utility function is just +1☺ for a correct decision?
- So m … rrect decis … sible chan…

Something is very wrong here

$$
\begin{aligned}
N^* &= \arg\max_N P(\text{correct decision} \mid N \text{ samples}) \\
&= \infty
\end{aligned}
$$

- So the decision-maker should collect an infinite number of samples.

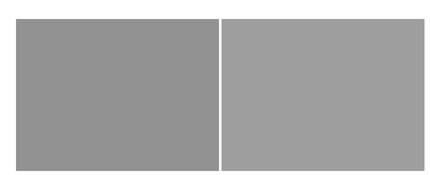
A <u>better</u> computational analysis

- I lied... The utility function is not "just +1☺ for a correct decision"

- Information is not free
  - The brain absorbs a huge proportion of the body's energy budget: each datum costs <u>energy</u>
  - Neurons can only fire at a finite rate, so each datum cost <u>time</u>. Given that we're all going to die, time is expensive
  - Time costs and energy costs are part of the human utility function

A better computational analysis

- The reward for being right is only +1☺
- There must be some tolerable error probability ε for which you would be willing to give up +1☺, in order to save yourself time and effort
- In order to save time, the learner's goal is to achieve a particular success rate, 1-ε
- This is called the <u>speed-accuracy tradeoff</u>. Only an idiot would spend the rest of their life on this problem...



Bayesian analysis

- If we have n samples, **s** = (s_1, s_2, ... s_n)
- Posterior probability that option A is correct

$$P(A|\mathbf{s}) = \frac{P(\mathbf{s}|A)P(A)}{P(\mathbf{s})}$$

- Posterior odds ratio for A versus B:

$$\frac{P(A|\mathbf{s})}{P(B|\mathbf{s})} = \frac{P(\mathbf{s}|A)}{P(\mathbf{s}|B)} \times \frac{P(A)}{P(B)}$$

Bayesian analysis

- If the samples s_1, s_2 ... s_n are conditionally independent, the likelihood function factorises

$$P(\mathbf{s}|A) = \prod_{i=1}^{n} P(s_i|A)$$

- So the posterior odds ratio looks like this

$$\frac{P(A|\mathbf{s})}{P(B|\mathbf{s})} = \prod_{i=1}^{n} \frac{P(s_i|A)}{P(s_i|B)} \times \frac{P(A)}{P(B)}$$

Bayesian analysis

- Taking logarithms makes everything additive

$$\ln \frac{P(A|\mathbf{s})}{P(B|\mathbf{s})} = \sum_{i=1}^{n} \ln \frac{P(s_i|A)}{P(s_i|B)} + \ln \frac{P(A)}{P(B)}$$

Bayesian analysis

- Taking logarithms makes everything additive

$$\ln \frac{P(A|\mathbf{s})}{P(B|\mathbf{s})} = \sum_{i=1}^{n} \ln \frac{P(s_i|A)}{P(s_i|B)} + \ln \frac{P(A)}{P(B)}$$

Define this as x_n, the total (log) evidence for option A after n samples

Bayesian analysis

- Taking logarithms makes everything additive

$$\ln \frac{P(A|\mathbf{s})}{P(B|\mathbf{s})} = \sum_{i=1}^{n} \ln \frac{P(s_i|A)}{P(s_i|B)} + \ln \frac{P(A)}{P(B)}$$

x_n

Define this as y_i, the relative probability of observing sample s_i under the two alternative hypotheses

Bayesian analysis

- Taking logarithms makes everything additive

$$\ln \frac{P(A|\mathbf{s})}{P(B|\mathbf{s})} = \sum_{i=1}^{n} \ln \frac{P(s_i|A)}{P(s_i|B)} + \ln \frac{P(A)}{P(B)}$$

(x_n) (y_i)

Call this y_0, the (log) prior odds favouring A over B

Bayesian analysis

- Taking logarithms makes everything additive

$$\ln \frac{P(A|\mathbf{s})}{P(B|\mathbf{s})} = \sum_{i=1}^{n} \ln \frac{P(s_i|A)}{P(s_i|B)} + \ln \frac{P(A)}{P(B)}$$

(x_n) (y_i) (y_0)

- So we can rewrite our analysis like this

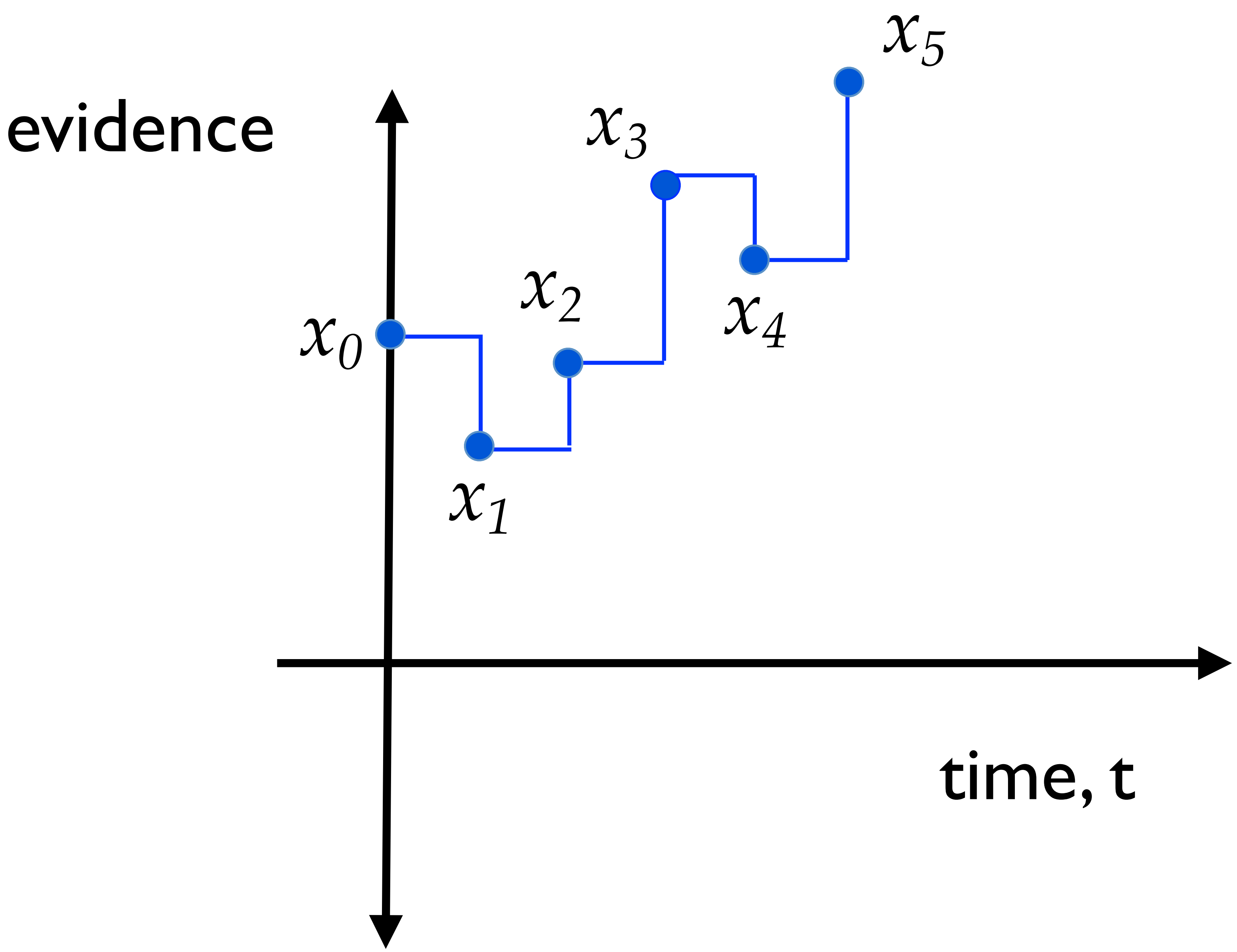
$$x_n = \sum_{i=0}^{n} y_i$$

Bayesian analysis

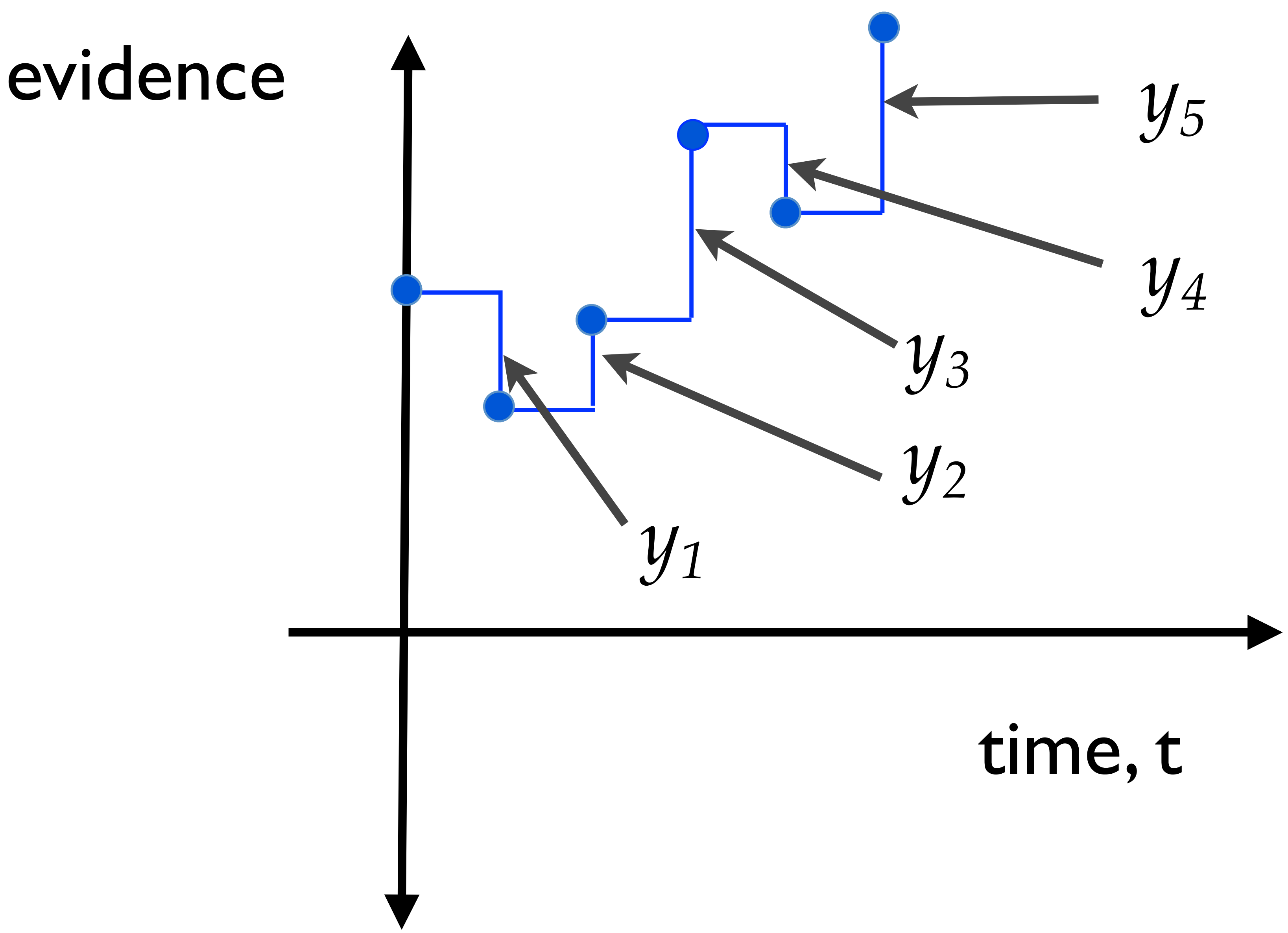
- Next, let's be explicit about the fact that this process unfolds over time. Assume that the samples arrive one at a time.
- At time t:

$$
\begin{aligned}
x_t &= \sum_{i=0}^{t} y_i \\
&= y_t + \sum_{i=0}^{t-1} y_i \\
&= y_t + x_{t-1}
\end{aligned}
$$

A random walk over "evidence space"



The size of each "step" corresponds to the evidence provided by a sample



Okay, when do we stop?

- Recall, our primary goal was to limit the probability of an incorrect decision to some level ε
- Therefore, the sampling must continue so long as

$$\epsilon < P(A|\mathbf{s}) < 1 - \epsilon$$

- Rewriting P(A|s) in terms of x...

$$\epsilon < \frac{1}{1 + \exp(-x_t)} < 1 - \epsilon$$

Okay, when do we stop?

- Recall, our primary goal was to limit the probability of an incorrect decision to some level ε
- Therefore, the sampling must continue so long as

$$\epsilon < P(A|\mathbf{s}) < 1 - \epsilon$$

- Rewriting P(A|s) in terms of x...

$$\epsilon < \frac{1}{1 + \exp(-x_t)} < 1 - \epsilon$$

Okay, when do we stop?

- A little algebra shows that this is equivalent to a decision algorithm that continues to sample new information while

$$|x_t| < \ln \frac{\epsilon}{1-\epsilon}$$

- More simply, $|x_t| < \gamma$ where $\gamma = \ln \frac{\epsilon}{1-\epsilon}$
- This is Wald's (1947) "sequential probability ratio test" (SPRT)

The random walk model for simple decisions

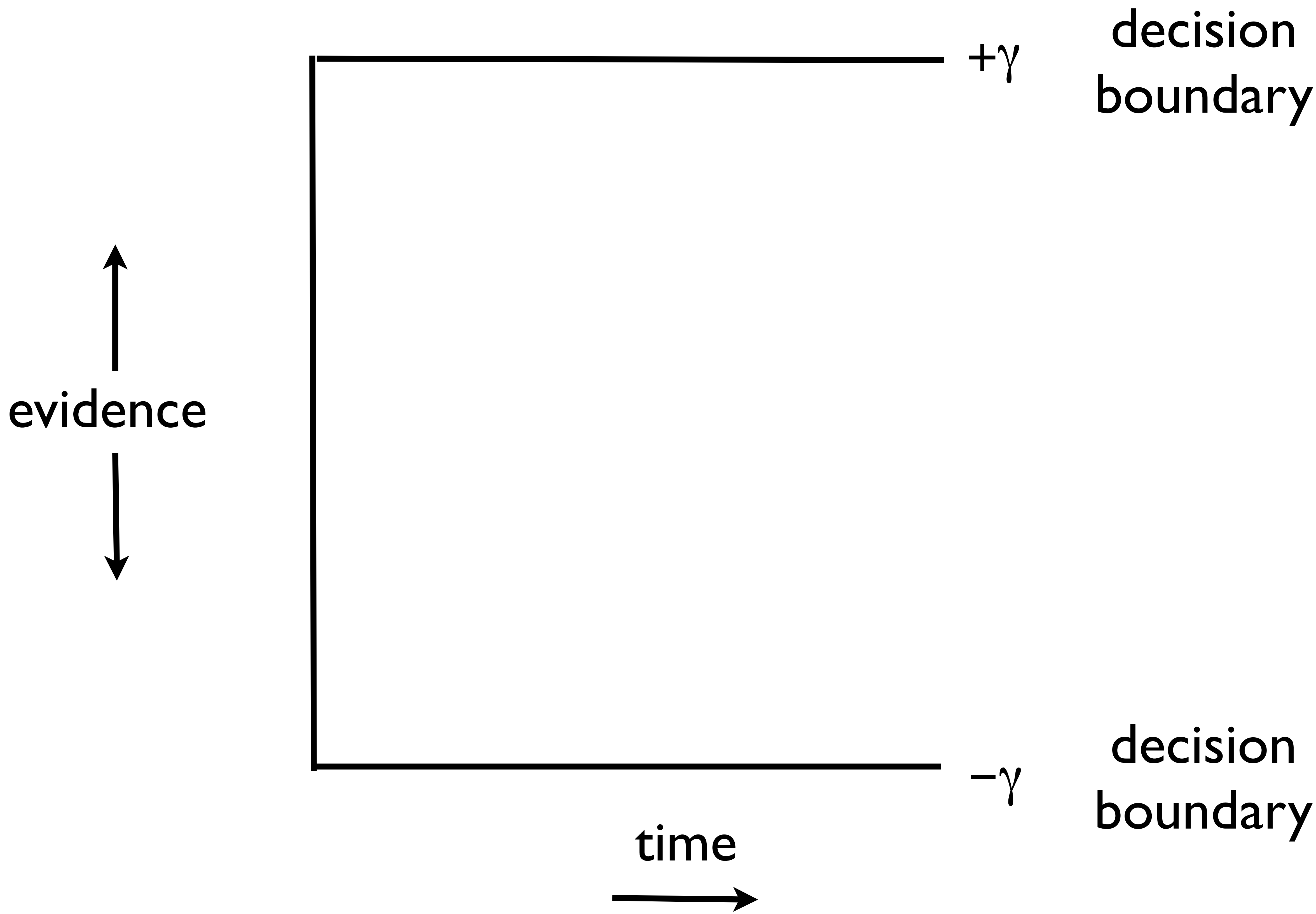
1. Time $t = 0$
2. Set x_0, based on your prior biases
3. Do while $|x_t| < \gamma$
   i. Time increments, $t = t+1$
   ii. Collect sensory sample s_t
   iii. Evaluate the log-odds for that sample, y_t
   iv. Increment evidence tally, $x_t = x_{t-1} + y_t$
4. If $x_t \geq \gamma$, choose option A
5. If $x_t \leq -\gamma$, choose option B

The random walk

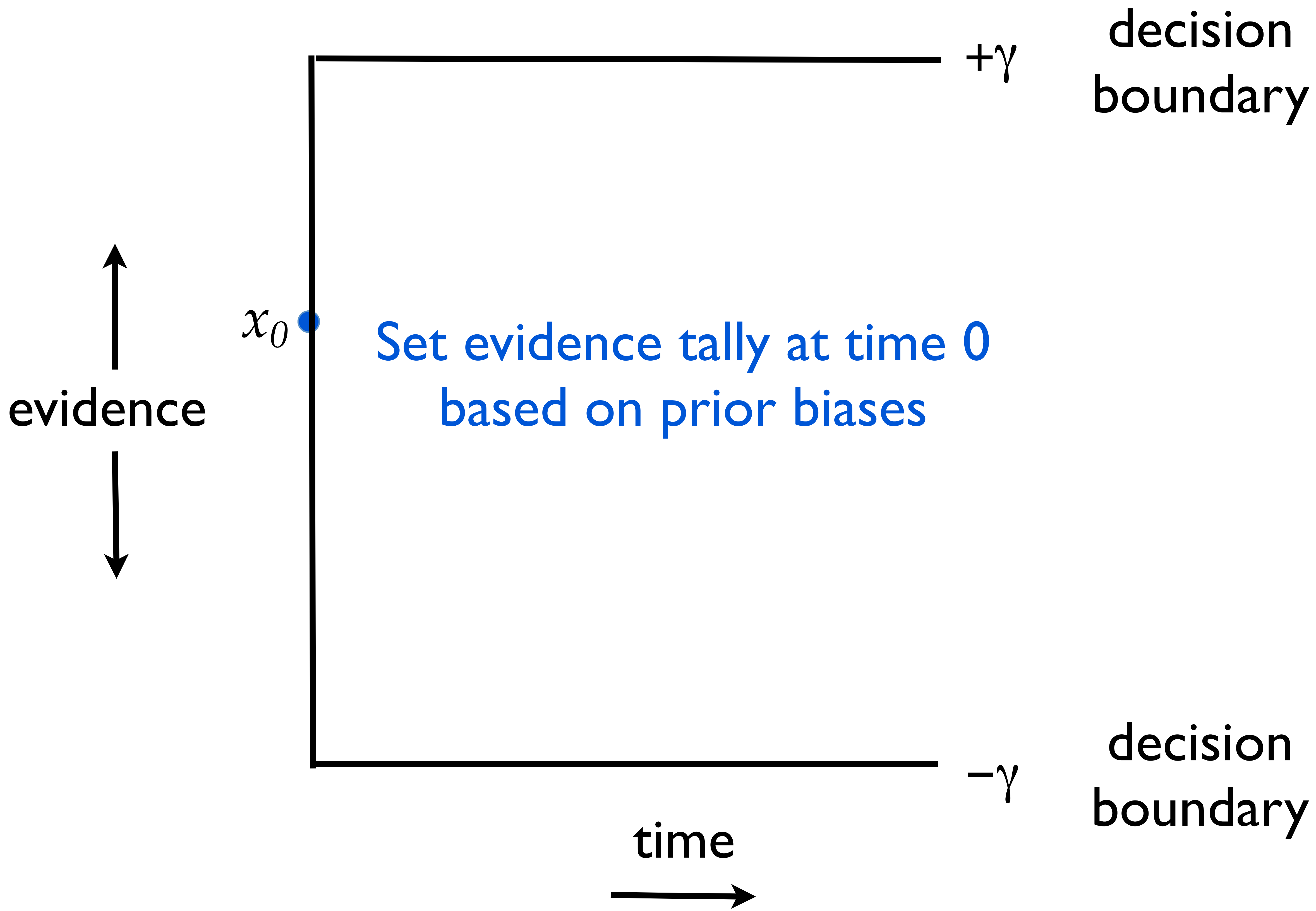
1. Time $t = 0$
2. Set x_0, based on
3. Do while $|x_t| <$
   i. Time increment
   ii. Collect sensory
   iii. Evaluate the log-odds for that sample, y_t
   iv. Increment evidence tally, $x_t = x_{t-1} + y_t$
4. If $x_t \geq \gamma$, choose option A
5. If $x_t \leq -\gamma$, choose option B

This random walk model is one of the simplest examples of a class of "sequential sampling" models that have dominated the theory of perceptual choice since the 1960s

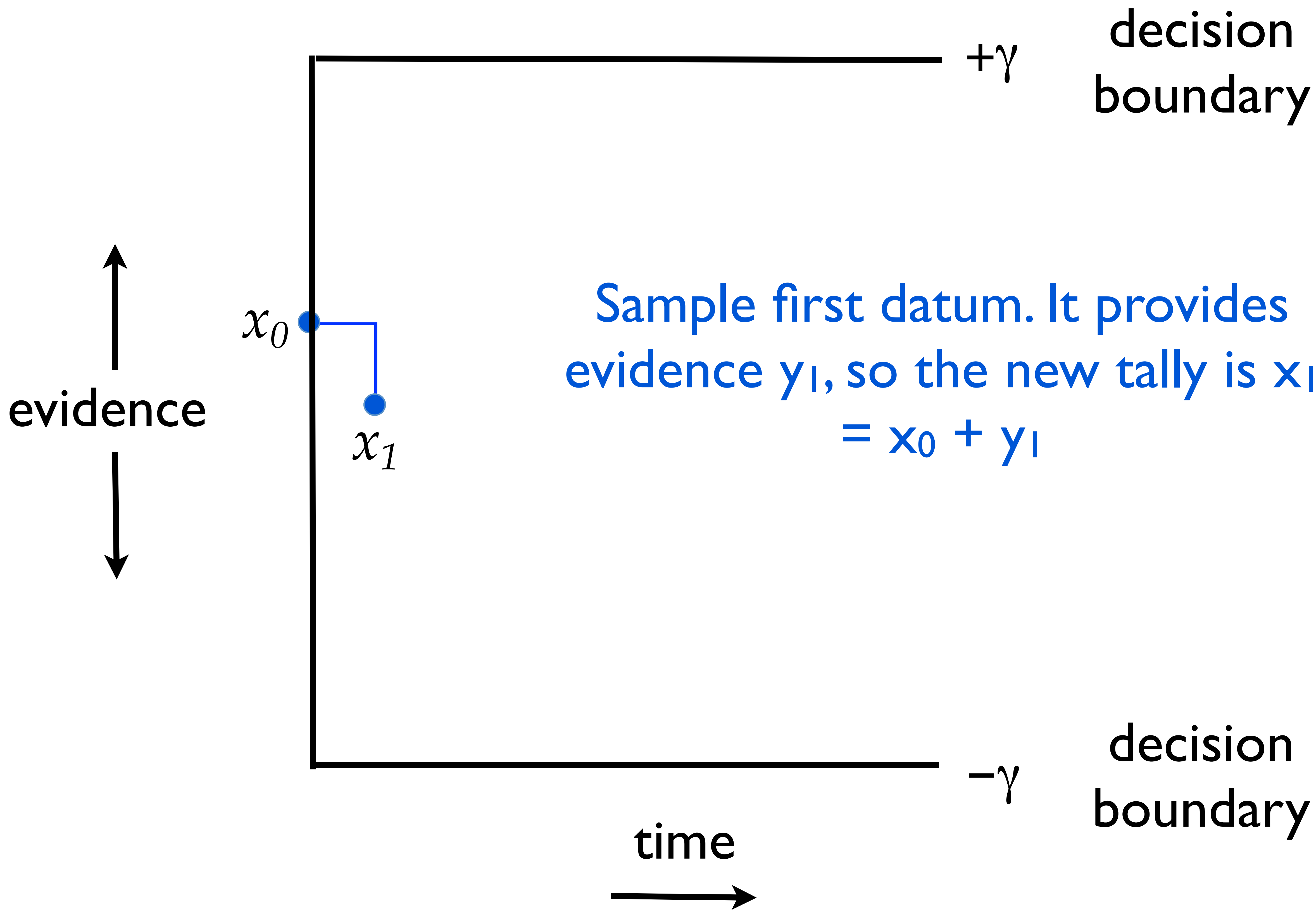
The random walk model



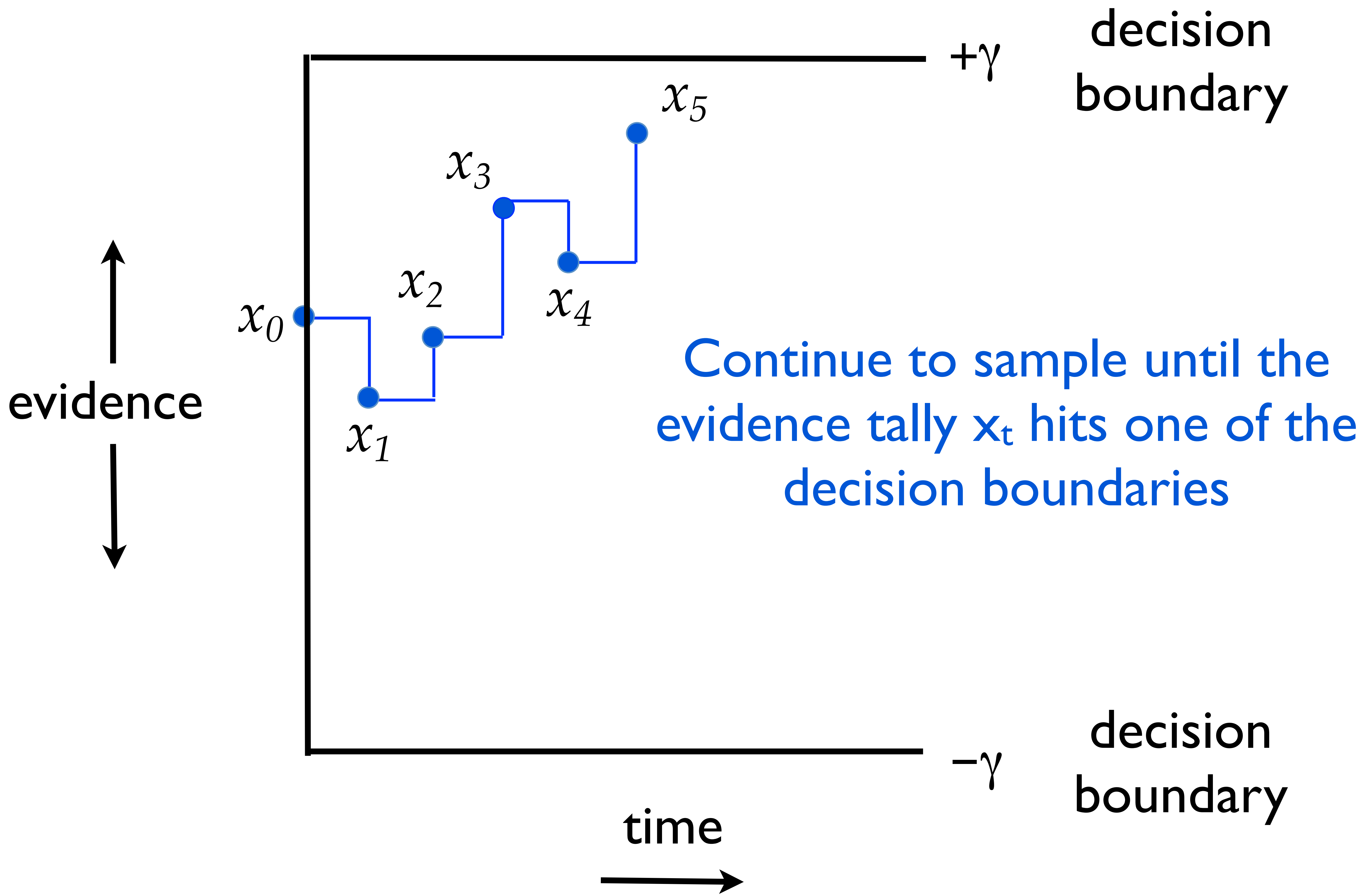
The random walk model



The random walk model



The random walk model



Terminology

- The time taken to reach the decision boundary is called the "first passage time"
- The step sizes (y values) are generated probabilistically from an "information function"
  - In some cases we know the actual information function, and we can calculate this directly
  - Most of the time we tend to assume that the information function generates y values from a nice tractable distribution (e.g., normal distribution, Bernoulli distribution)

Demonstration code: ssm.R