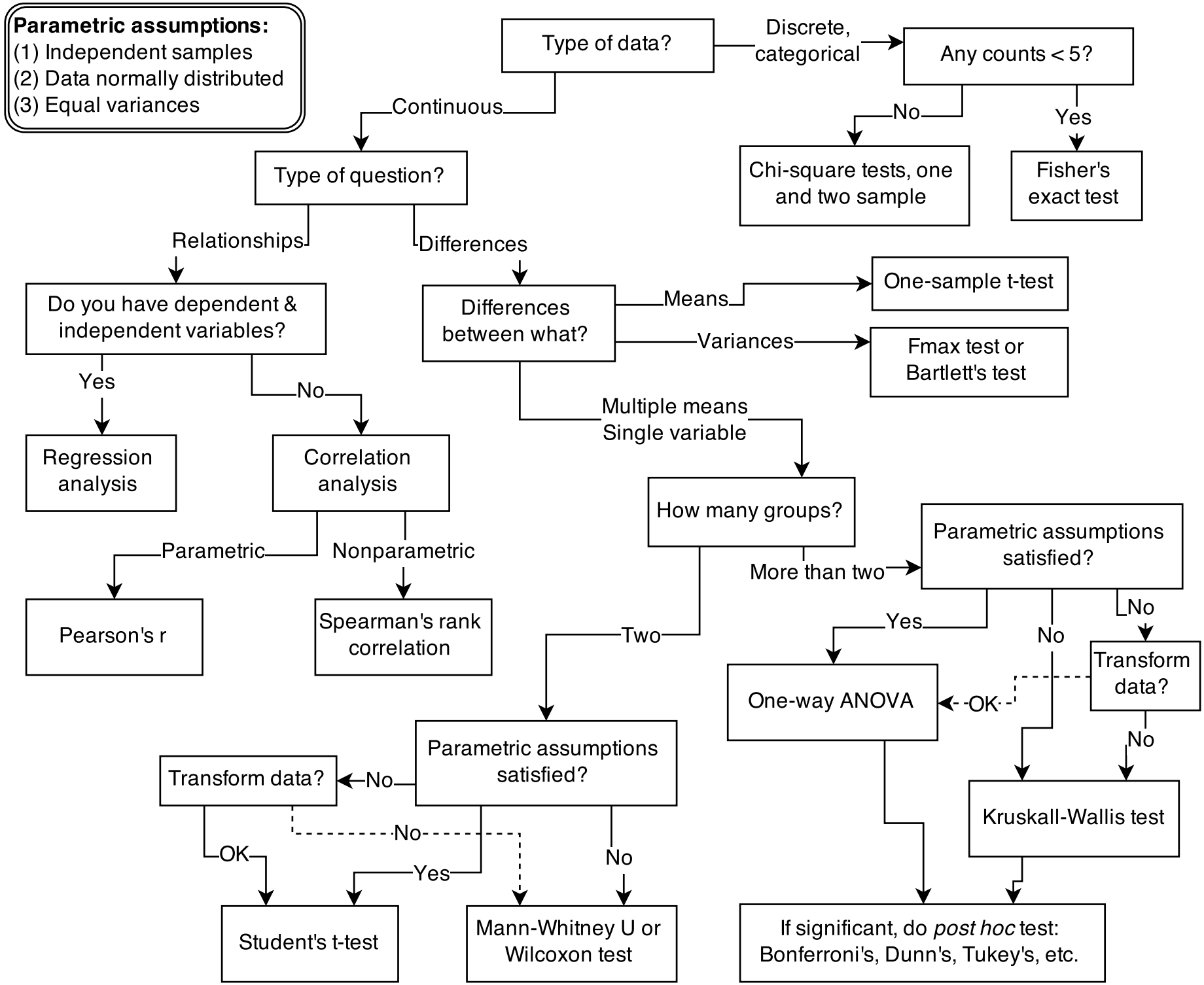
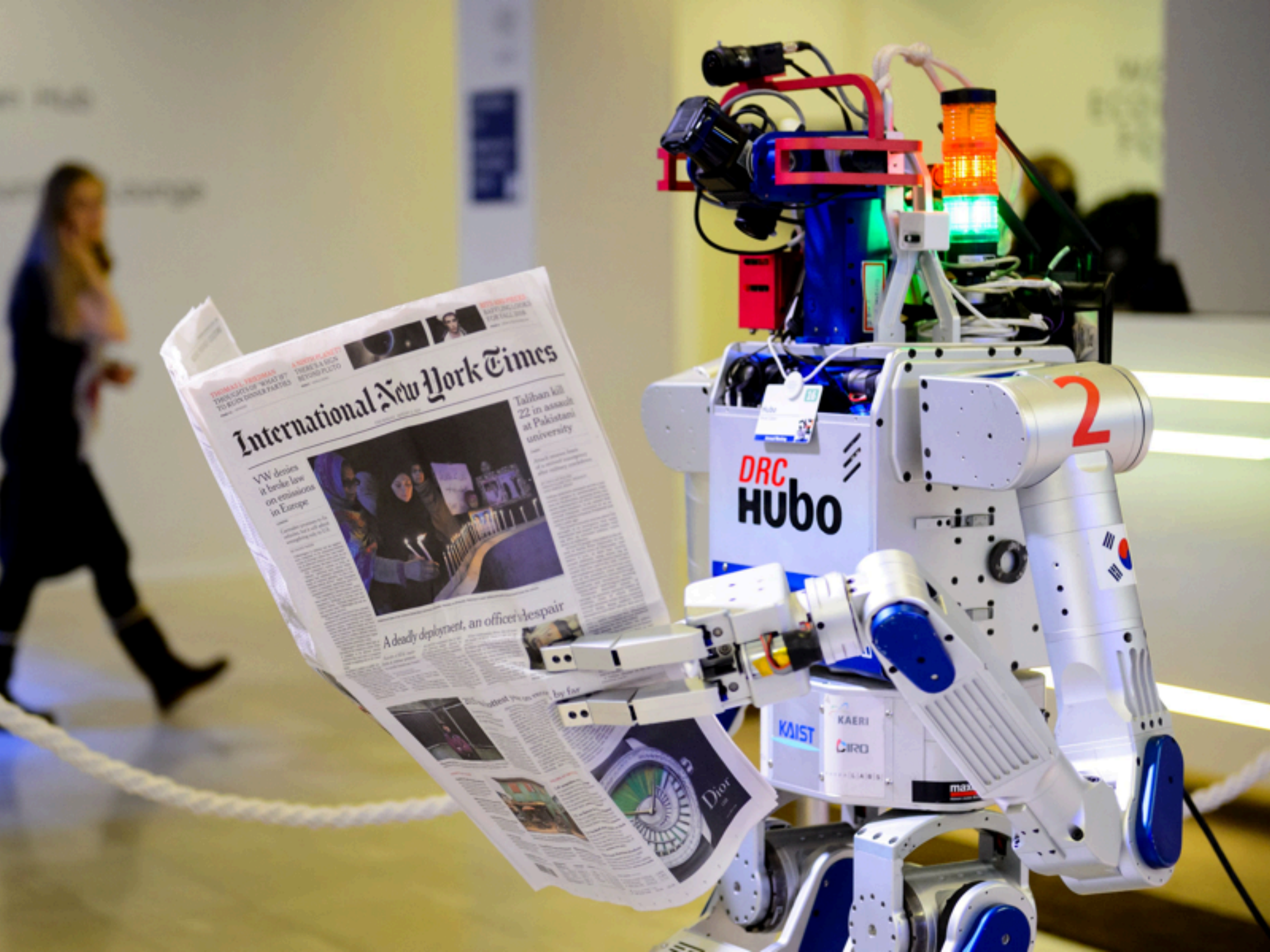
# BAYESIAN INFERENCE IS JUST COUNTING


Richard McElreath
MPI-EVA


Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances
Type of data?
Discrete, categorical
Any counts < 5?
No
Yes
Chi-square tests, one and two sample
Fisher's exact test
Continuous
Type of question?
Relationships
Differences
Do you have dependent & independent variables?
Yes
No
Regression analysis
Correlation analysis
Parametric
Nonparametric
Pearson's r
Spearman's rank correlation
Differences between what?
Means
One-sample t-test
Variances
Fmax test or Bartlett's test
Multiple means Single variable
How many groups?
Two
More than two
Parametric assumptions satisfied?
Yes
No
No
Transform data?
OK
One-way ANOVA
No
Kruskall-Wallis test
Parametric assumptions satisfied?
Transform data?
No
OK
No
Yes
No
Student's t-test
Mann-Whitney U or Wilcoxon test
If significant, do post hoc test: Bonferroni's, Dunn's, Tukey's, etc.


R.A.FISHER
FELLOW 1920-26 1943-62
PRESIDENT 1956-59


International New York Times
DRC
HUBO
2
KAIST
KAERI

# The Golem of Prague

**go•lem** |gōləm|

noun

• (in Jewish legend) a clay figure brought to life by magic.

• an automaton or robot.

ORIGIN late 19th cent.: from Yiddish ***goylem***, from Hebrew ***gōlem***
'shapeless mass.'

# The Golem of Prague

*"Even the most perfect of Golem, risen to life to protect us, can easily change into a destructive force. Therefore let us treat carefully that which is strong, just as we bow kindly and patiently to that which is weak."*


Rabbi Judah Loew ben Bezalel (1512–1609)

From *Breath of Bones: A Tale of the Golem*

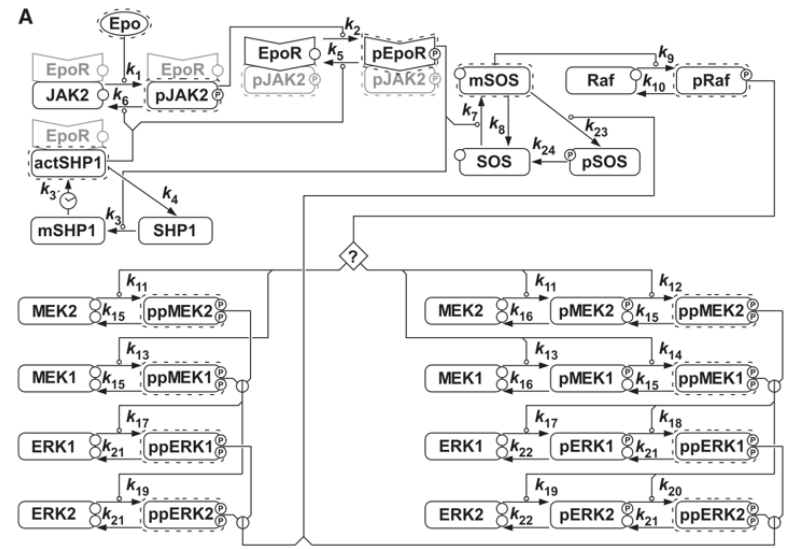
# The Golems of Science

## Golem

- Made of clay
- Animated by “truth”
- Powerful
- Blind to creator’s intent
- Easy to misuse
- Fictional

## Model

- Made of...silicon?
- Animated by “truth”
- Hopefully powerful
- Blind to creator’s intent
- Easy to misuse
- Not even false



# Bayesian data analysis

- Use *probability* to describe uncertainty
  - Extends ordinary logic (true/false) to continuous *plausibility*
- Computationally difficult
  - Markov chain Monte Carlo (MCMC) to the rescue
- Used to be controversial
  - Ronald Fisher: Bayesian analysis "must be wholly rejected."

Pierre-Simon Laplace (1749–1827)

Sir Harold Jeffreys (1891–1989) with Bertha Swirles, aka Lady Jeffreys (1903–1999)

# Bayesian data analysis

*Count all the ways data can happen, according to assumptions.*

*Assumptions with more ways that are consistent with data are more plausible.*

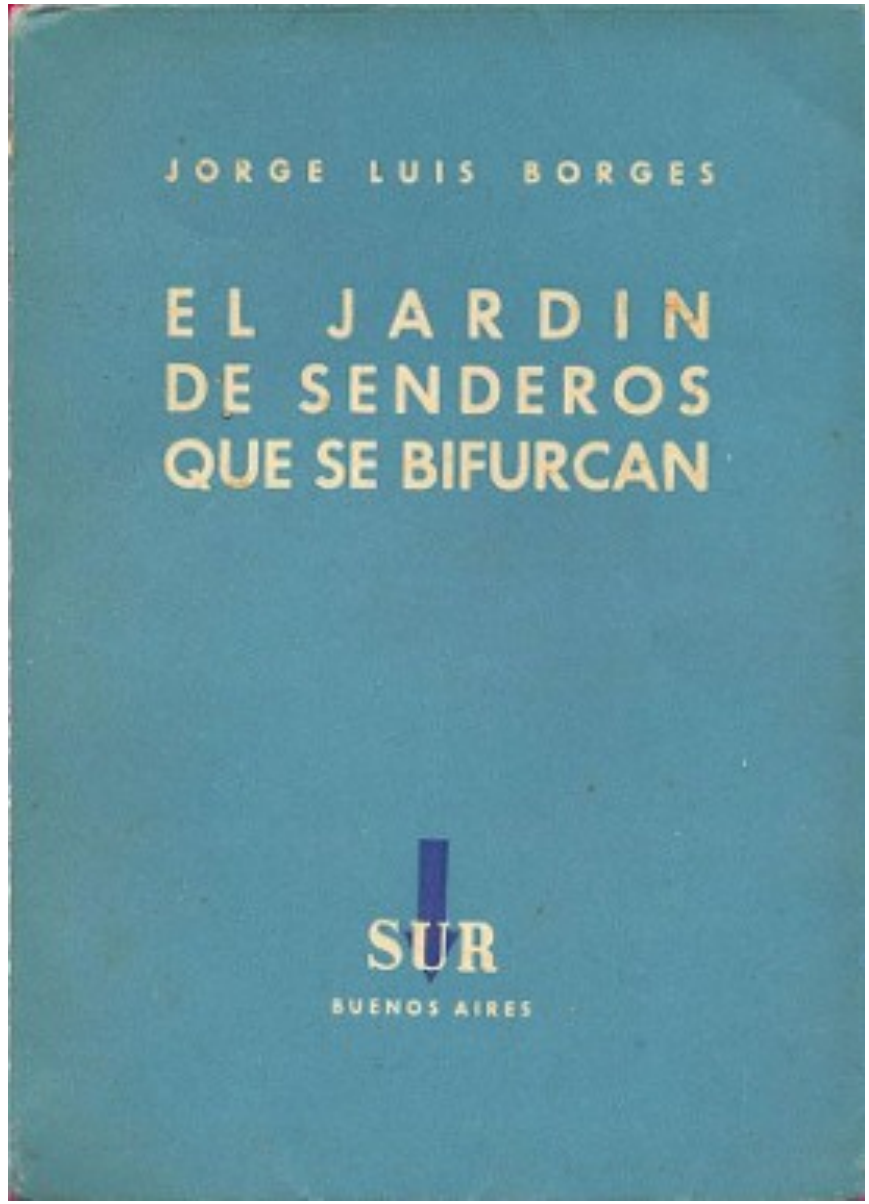
# Bayesian data analysis

- Contrast with *frequentist* view
  - Probability is just limiting frequency
  - Uncertainty arises from *sampling variation*
- Bayesian probability much more general
  - Probability is in the golem, not in the world
  - Coins are not random, but our ignorance makes them so

Saturn as Galileo saw it

# Garden of Forking Data

- The future:
  - Full of branching paths
  - Each choice closes some
- The data:
  - Many possible events
  - Each observation eliminates some



# Garden of Forking Data


Contains 4 marbles

Possible contents:

(1) ○○○○

(2) ●○○○

(3) ●●○○

(4) ●●●○

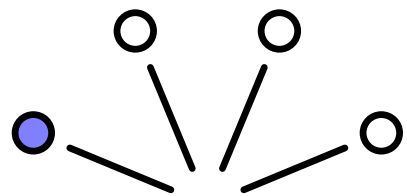
(5) ●●●●

Observe:

●○●

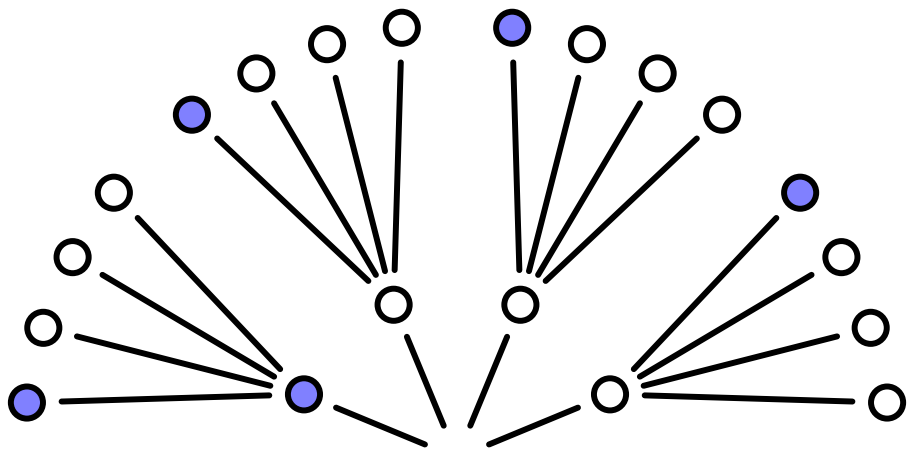
Conjecture: 

Data: 

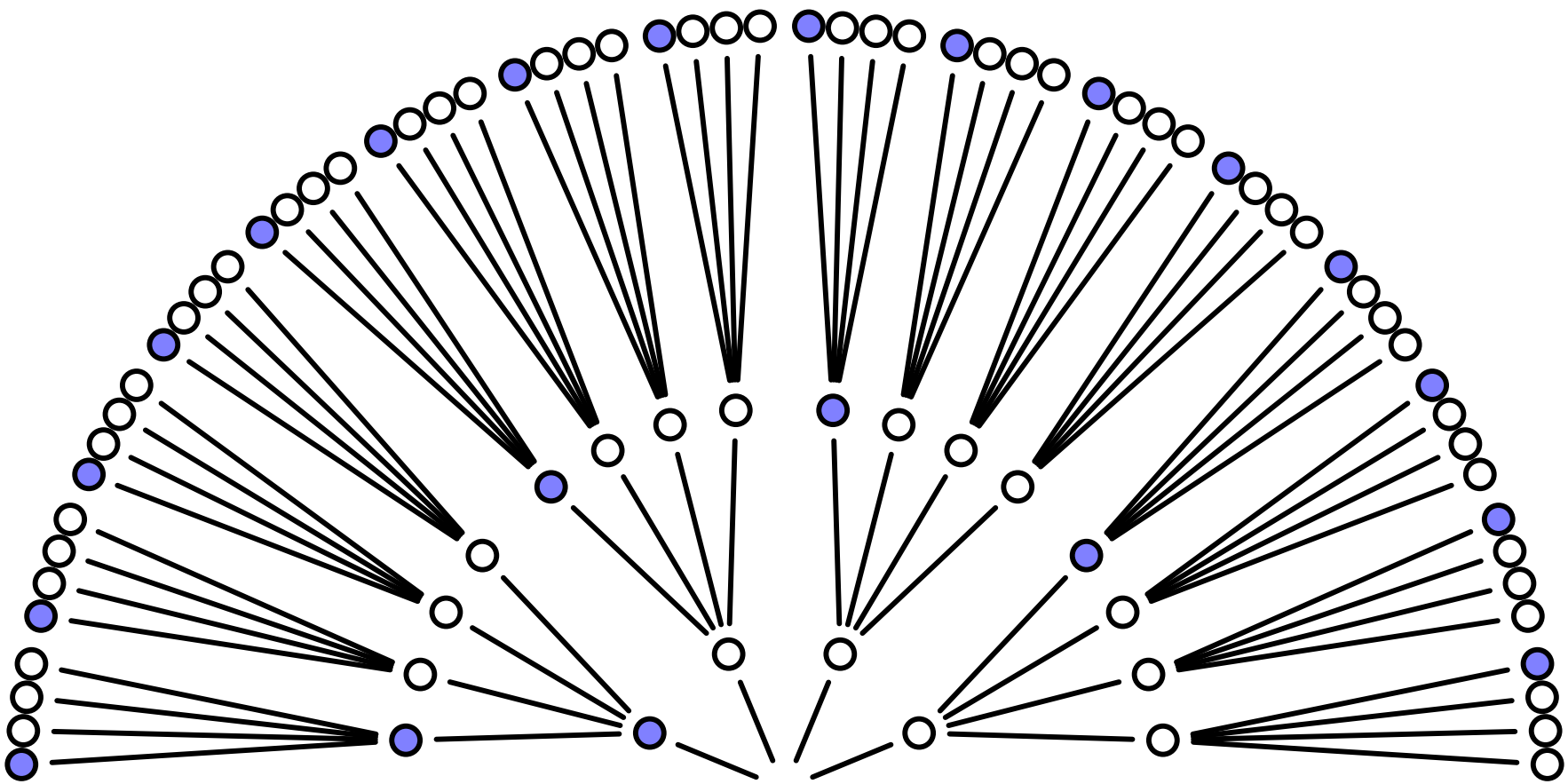


Conjecture:

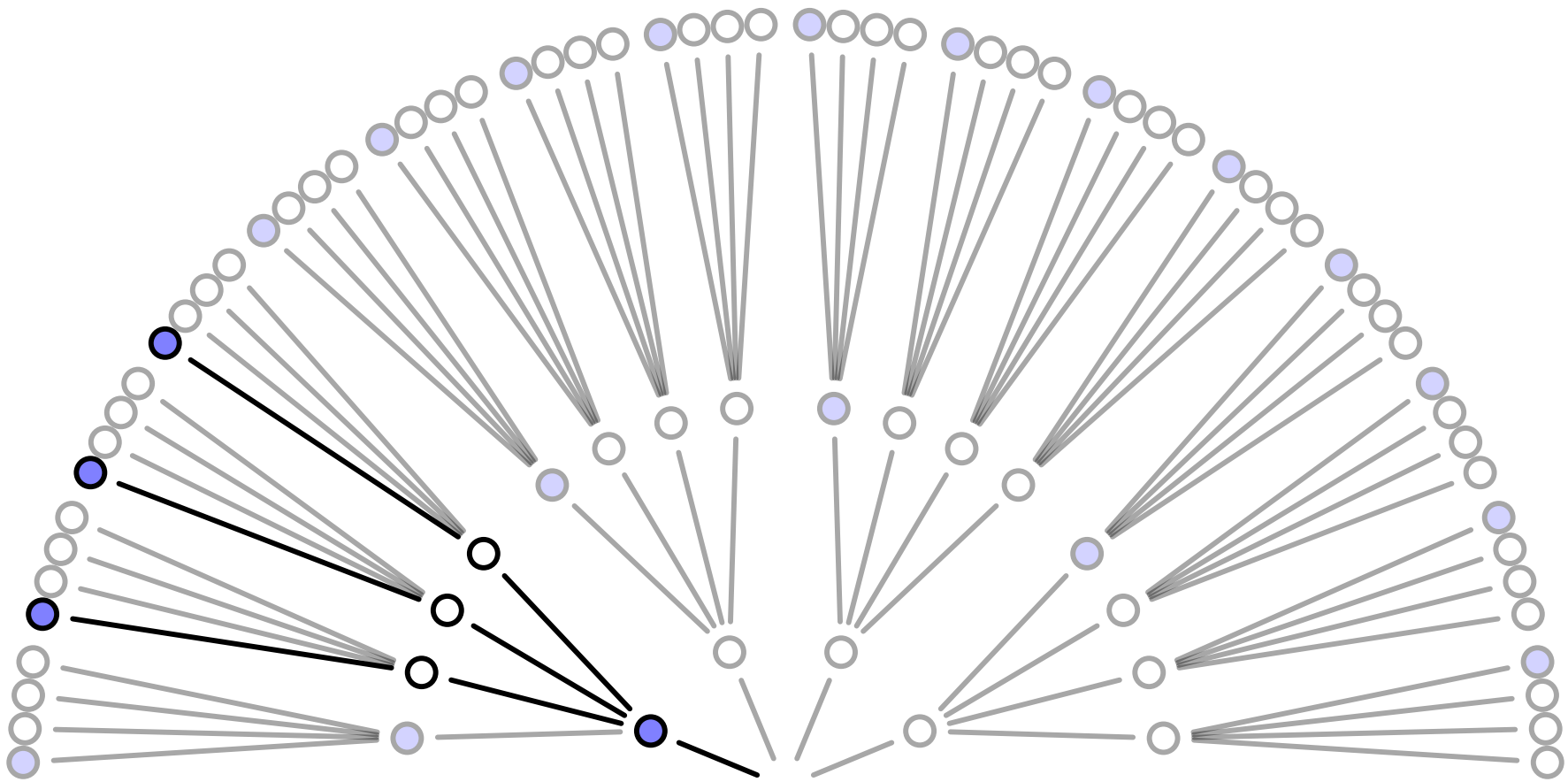
Data:



Conjecture: ●○○○ Data: ●○●



Conjecture: ●○○○ Data: ●○●



3 paths consistent with data

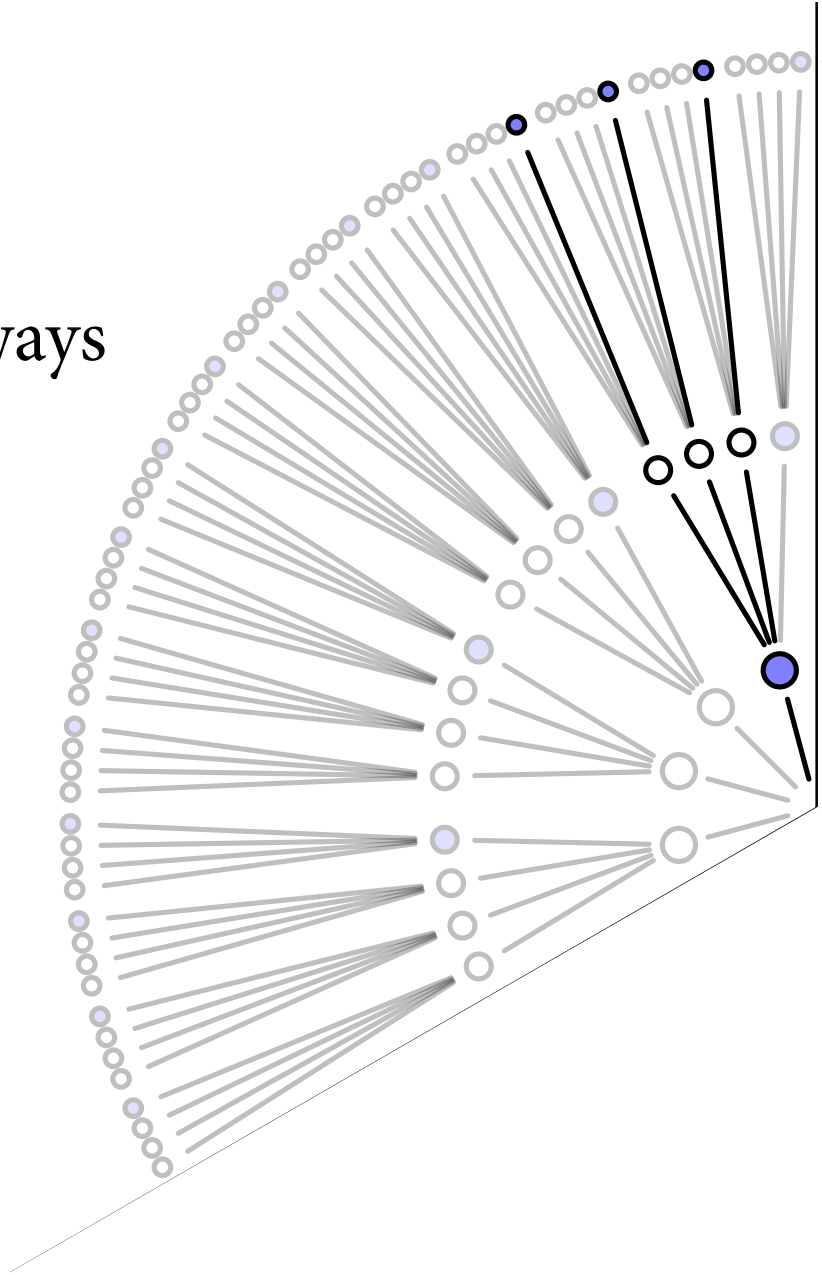
# Garden of Forking Data

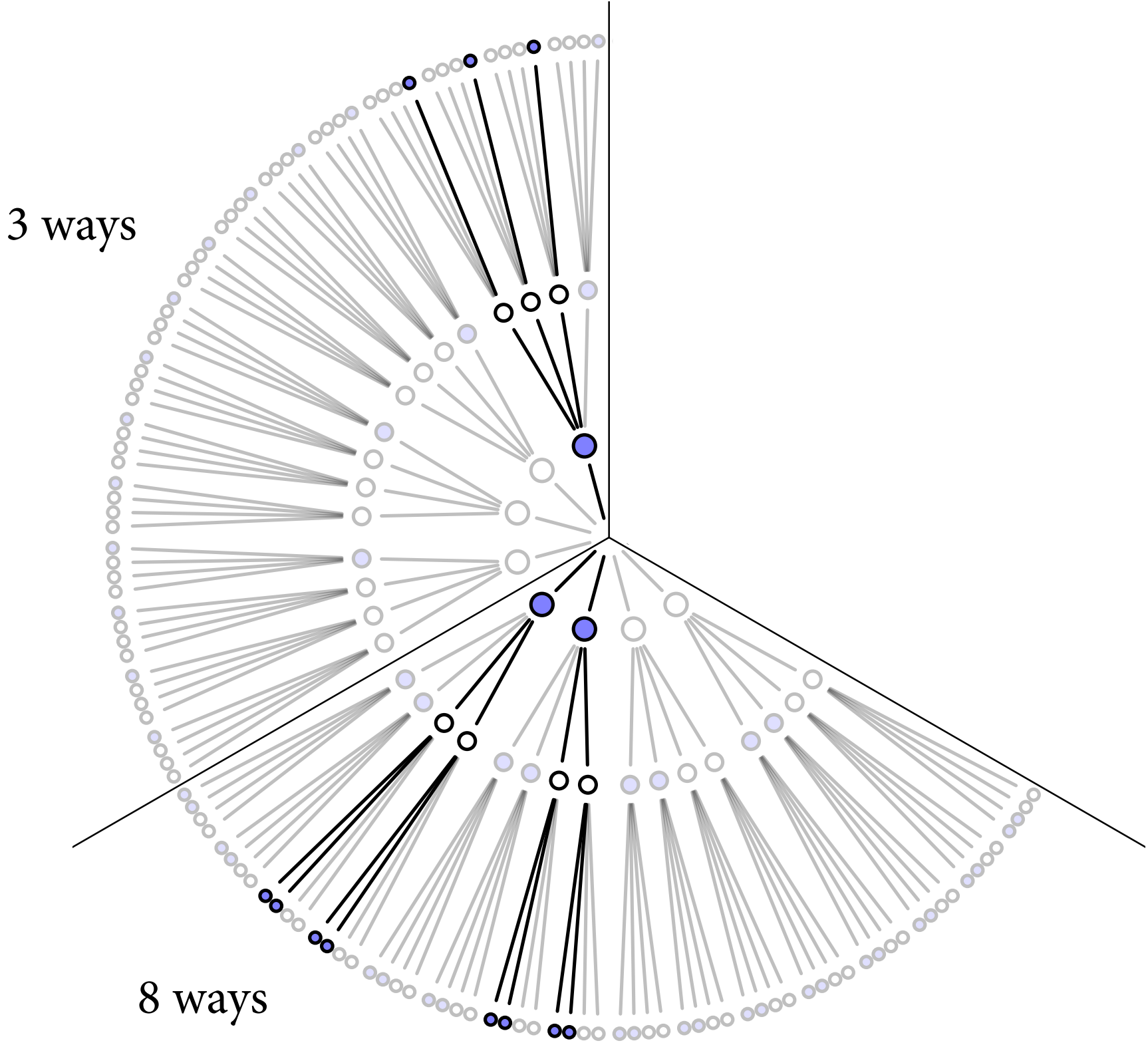
| Possible contents: | Ways to produce ●○● |
| --- | --- |
| (1) ○○○○ | ? |
| (2) ●○○○ | 3 |
| (3) ●●○○ | ? |
| (4) ●●●○ | ? |
| (5) ●●●● | ? |

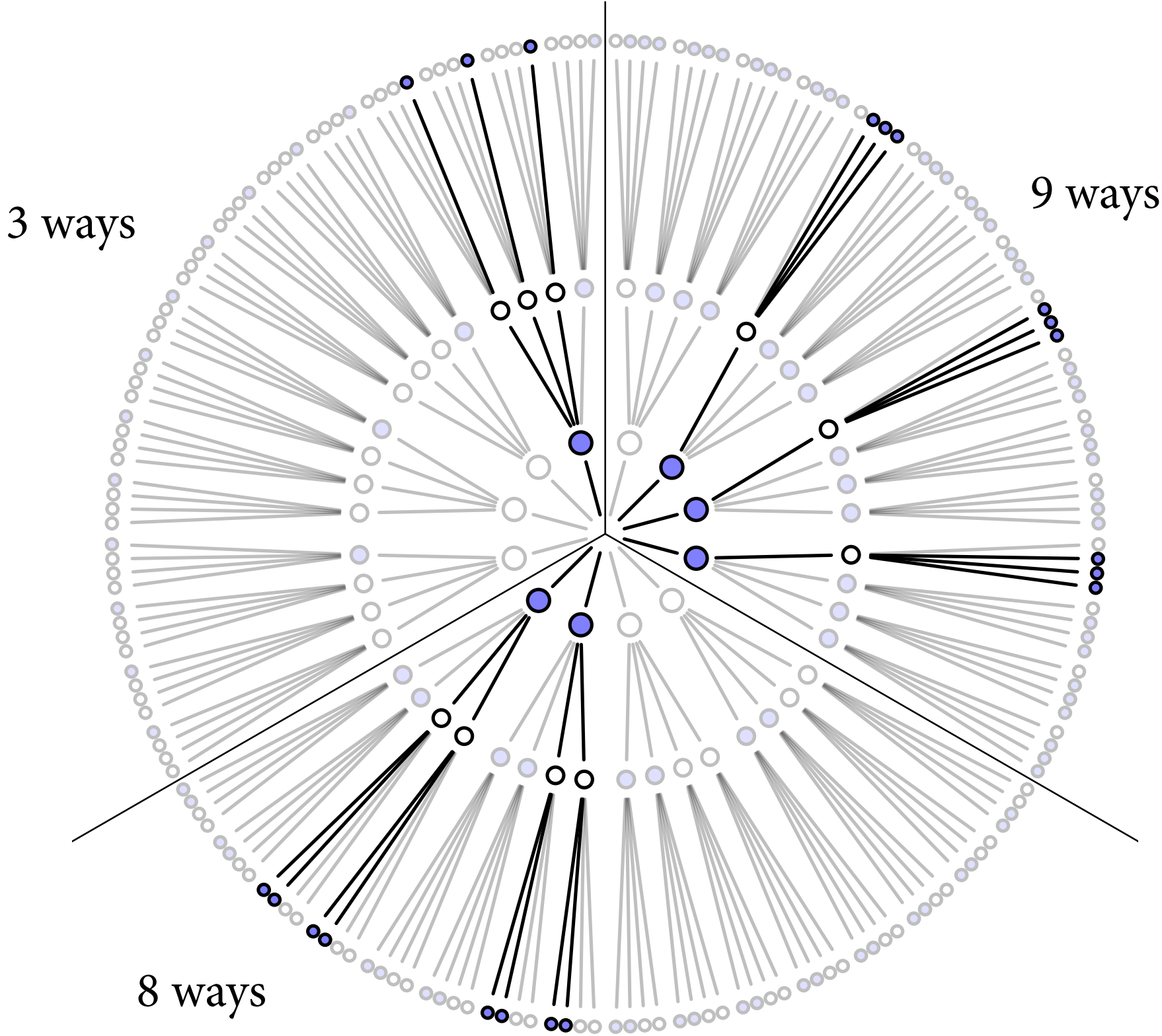
# Garden of Forking Data

| Possible contents: | Ways to produce ●○● |
|---|---|
| (1) ○○○○ | 0 |
| (2) ●○○○ | 3 |
| (3) ●●○○ | ? |
| (4) ●●●○ | ? |
| (5) ●●●● | 0 |

3 ways

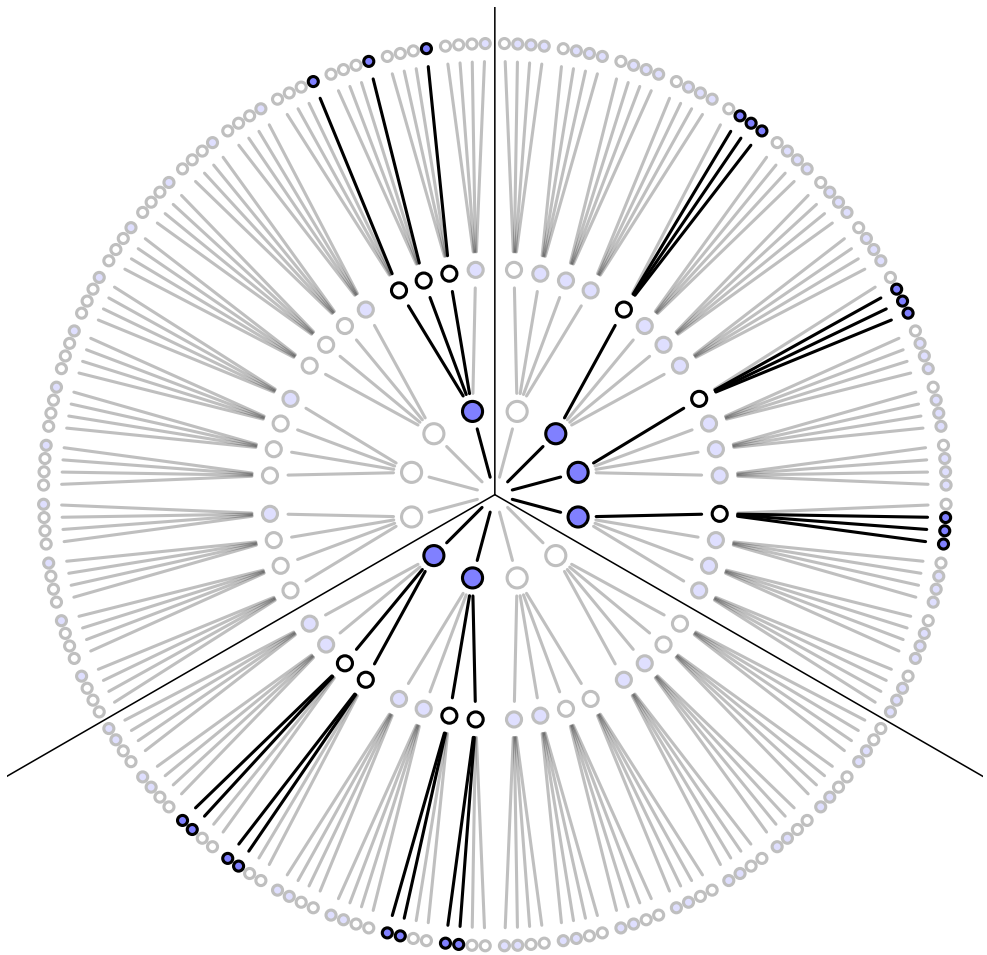



3 ways
8 ways


3 ways
9 ways
8 ways

# Garden of Forking Data

| Conjecture | Ways to produce ●○● |
|---|---|
| [○○○○] | $0 \times 4 \times 0 = 0$ |
| [●○○○] | $1 \times 3 \times 1 = 3$ |
| [●●○○] | $2 \times 2 \times 2 = 8$ |
| [●●●○] | $3 \times 1 \times 3 = 9$ |
| [●●●●] | $4 \times 0 \times 4 = 0$ |



# Updating

Another draw from the bag: 

| Conjecture | Ways to produce ● | Previous counts | New count |
|---|---|---|---|
| [○○○○] | 0 | 0 | $0 \times 0 = 0$ |
| [●○○○] | 1 | 3 | $3 \times 1 = 3$ |
| [●●○○] | 2 | 8 | $8 \times 2 = 16$ |
| [●●●○] | 3 | 9 | $9 \times 3 = 27$ |
| [●●●●] | 4 | 0 | $0 \times 4 = 0$ |

# Using other information

Factory says: ● marbles rare, but every bag contains at least one.

| Conjecture | Factory count |
|---|---|
| [○○○○] | 0 |
| [●○○○] | 3 |
| [●●○○] | 2 |
| [●●●○] | 1 |
| [●●●●] | 0 |

# Using other information

Factory says: ● marbles rare.

| Conjecture | Prior ways | Factory count | New count |
|---|---|---|---|
| [○○○○] | 0 | 0 | $0 \times 0 = 0$ |
| [●○○○] | 3 | 3 | $3 \times 3 = 9$ |
| [●●○○] | 16 | 2 | $16 \times 2 = 32$ |
| [●●●○] | 27 | 1 | $27 \times 1 = 27$ |
| [●●●●] | 0 | 0 | $0 \times 0 = 0$ |

# Counts to plausibility

Unglamorous basis of applied probability:

*Things that can happen more ways are more plausible.*

| Possible composition | $p$ | ways to produce data | plausibility |
|---|---|---|---|
| [○○○○] | 0 | 0 | 0 |
| [●○○○] | 0.25 | 3 | 0.15 |
| [●●○○] | 0.5 | 8 | 0.40 |
| [●●●○] | 0.75 | 9 | 0.45 |
| [●●●●] | 1 | 0 | 0 |

# Counts to plausibility

| Possible composition | $p$ | ways to produce data | plausibility |
|---|---|---|---|
| [○○○○] | 0 | 0 | 0 |
| [●○○○] | 0.25 | 3 | 0.15 |
| [●●○○] | 0.5 | 8 | 0.40 |
| [●●●○] | 0.75 | 9 | 0.45 |
| [●●●●] | 1 | 0 | 0 |

```
ways <- c( 3 , 8 , 9 )
ways/sum(ways)
```

R code 2.1

```
[1] 0.15 0.40 0.45
```

# Counts to plausibility

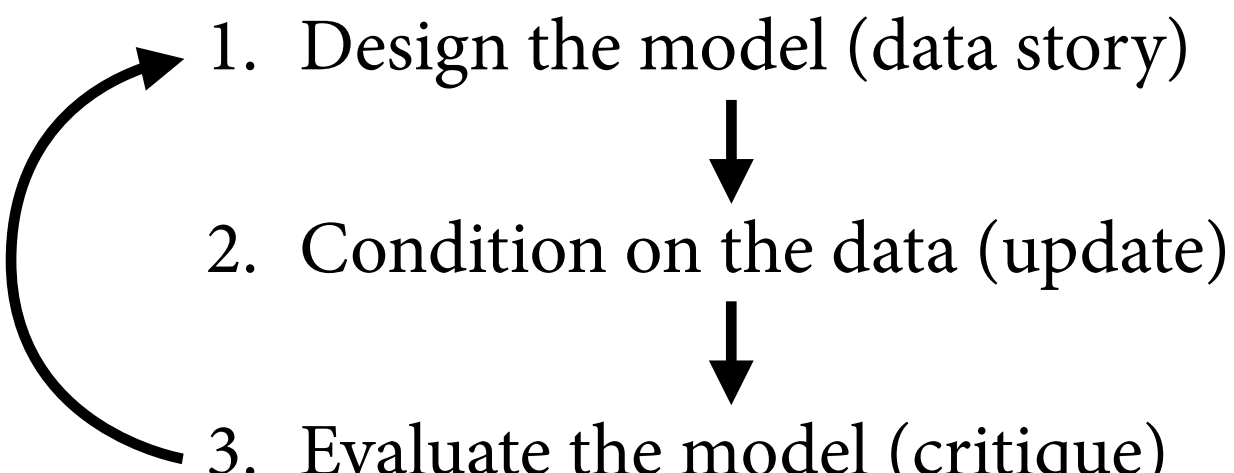
| Possible composition | $p$ | ways to produce data | plausibility |
|---|---|---|---|
| [○○○○] | 0 | 0 | 0 |
| [●○○○] | 0.25 | 3 | 0.15 |
| [●●○○] | 0.5 | 8 | 0.40 |
| [●●●○] | 0.75 | 9 | 0.45 |
| [●●●●] | 1 | 0 | 0 |

Plausibility is *probability*: Set of non-negative real numbers that sum to one.

Probability theory is just a set of shortcuts for counting possibilities.

# Building a model

- How to use probability to do typical statistical modeling?



Nine tosses of the globe:

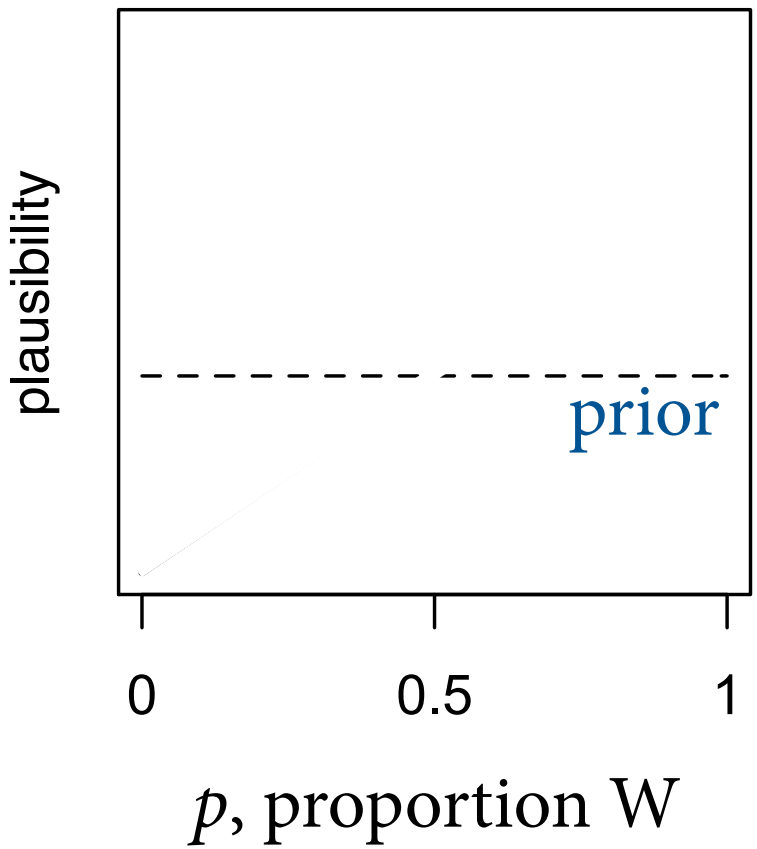
**W L W W W L W L W**

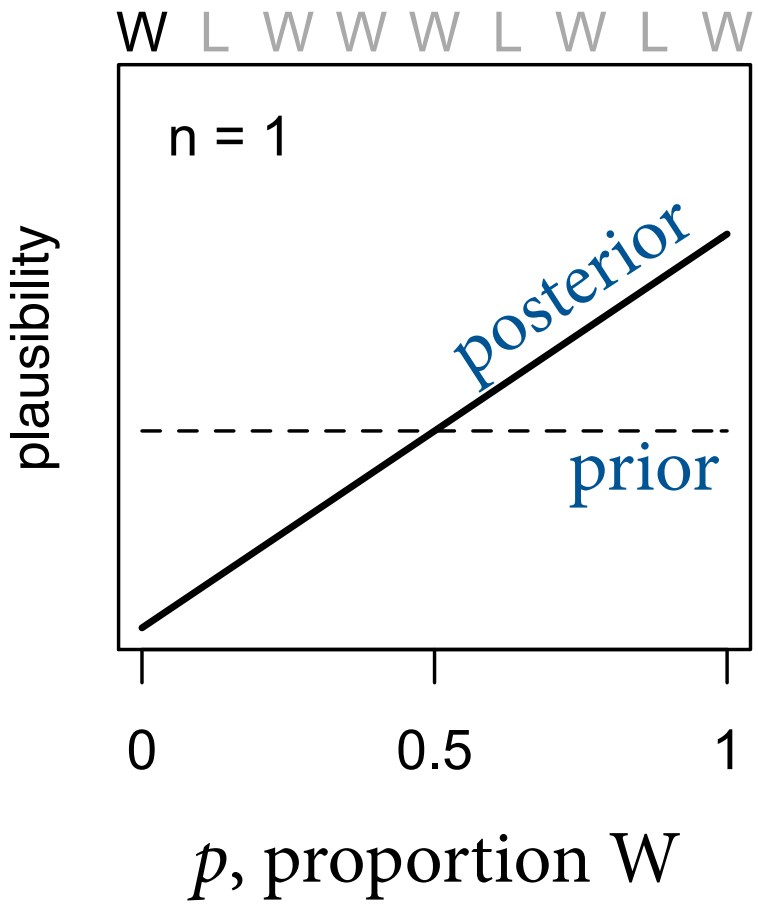
# **Design** > Condition > Evaluate

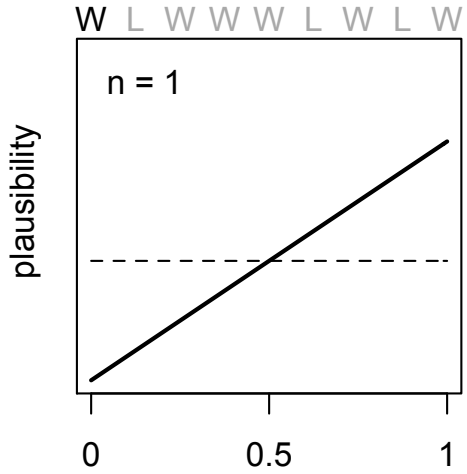
- Data story motivates the model
  - How do the data arise?
- For **W L W W W L W L W**:
  - Some true proportion of water, $p$
  - Toss globe, probability $p$ of observing W, $1-p$ of L
  - Each toss therefore independent of other tosses
- Translate data story into probability statements

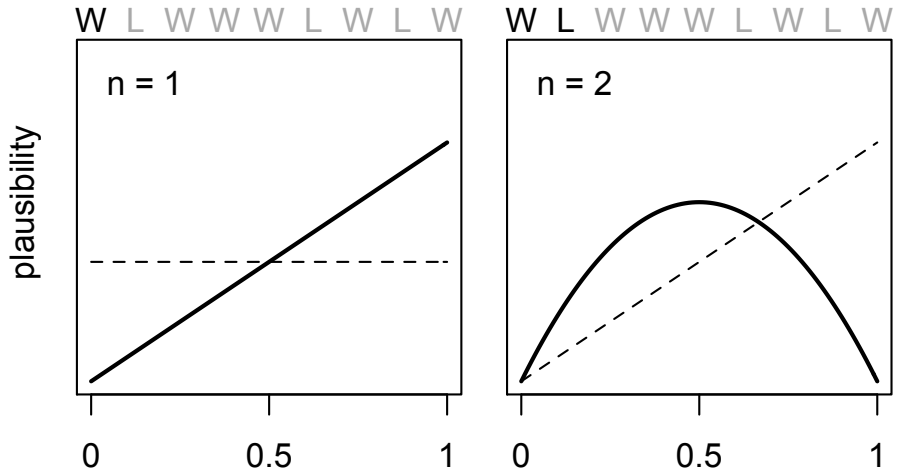
# Design > **Condition** > Evaluate

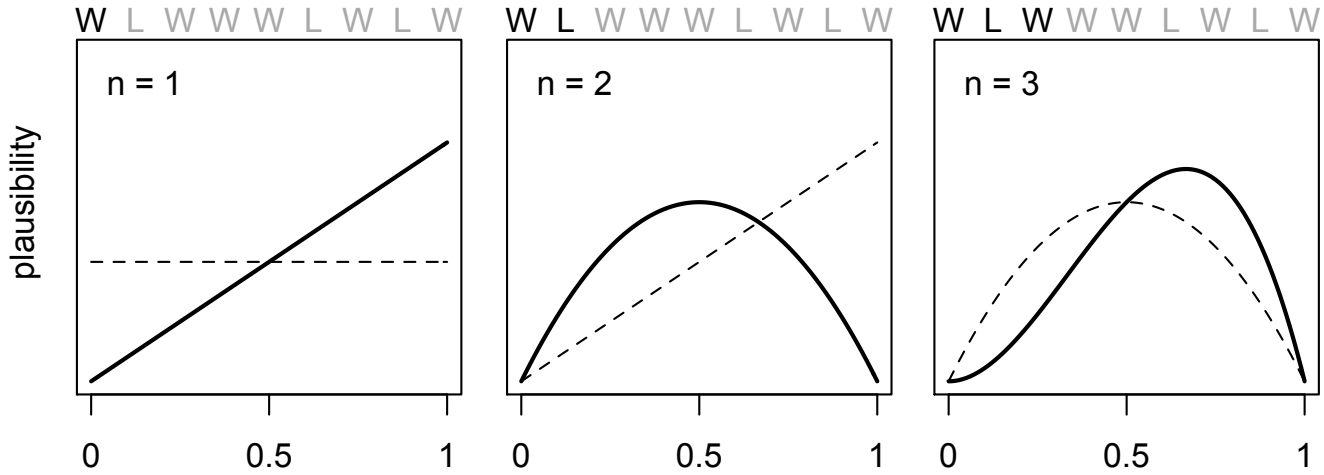
- *Bayesian updating* defines optimal learning in small world, converts *prior* into *posterior*
  - Give your golem an information state, before the data: Here, an initial confidence in each possible value of $p$ between zero and one
  - Condition on data to update information state: New confidence in each value of $p$, conditional on data

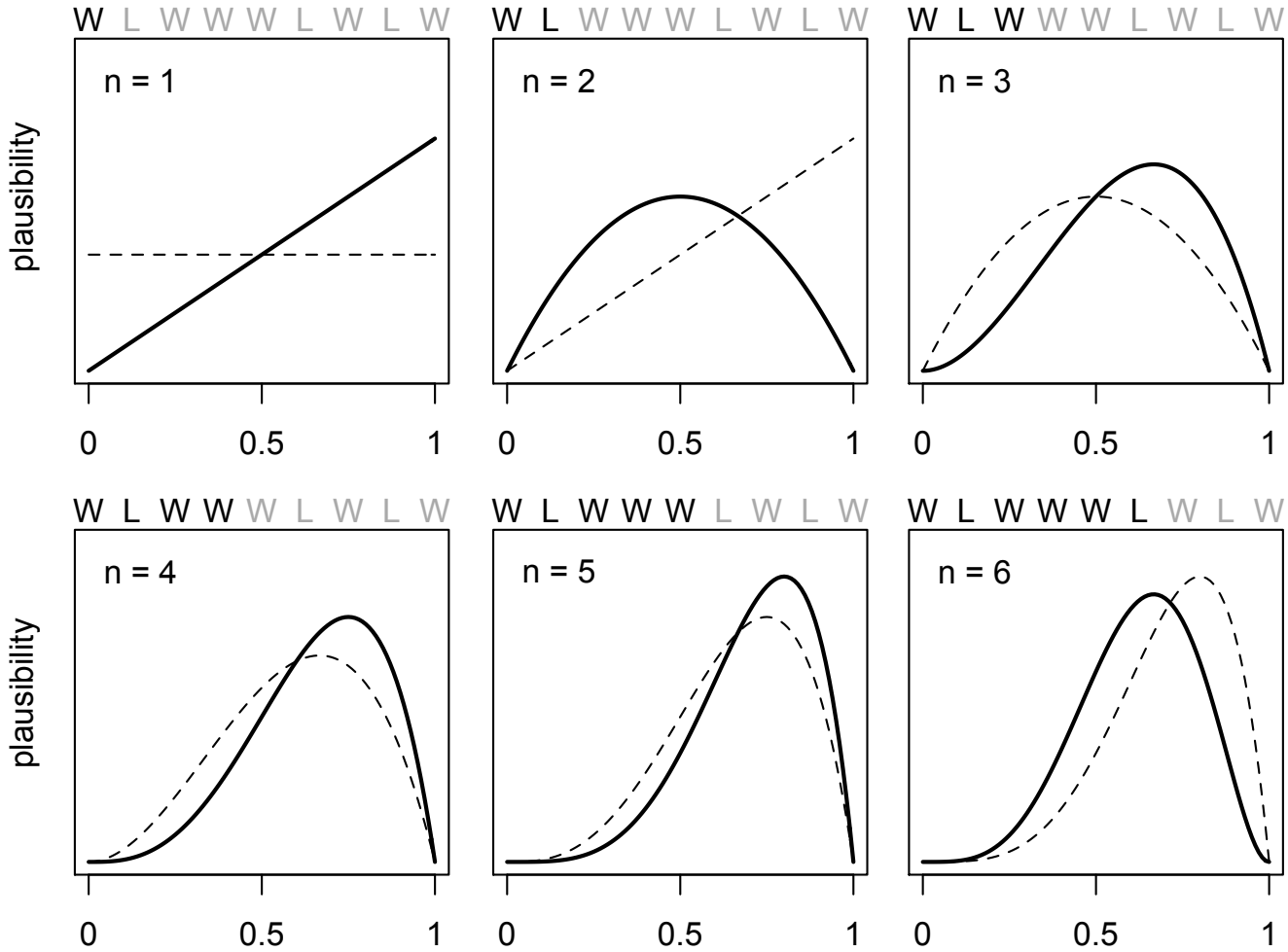

plausibility
prior
0
0.5
1
p, proportion W

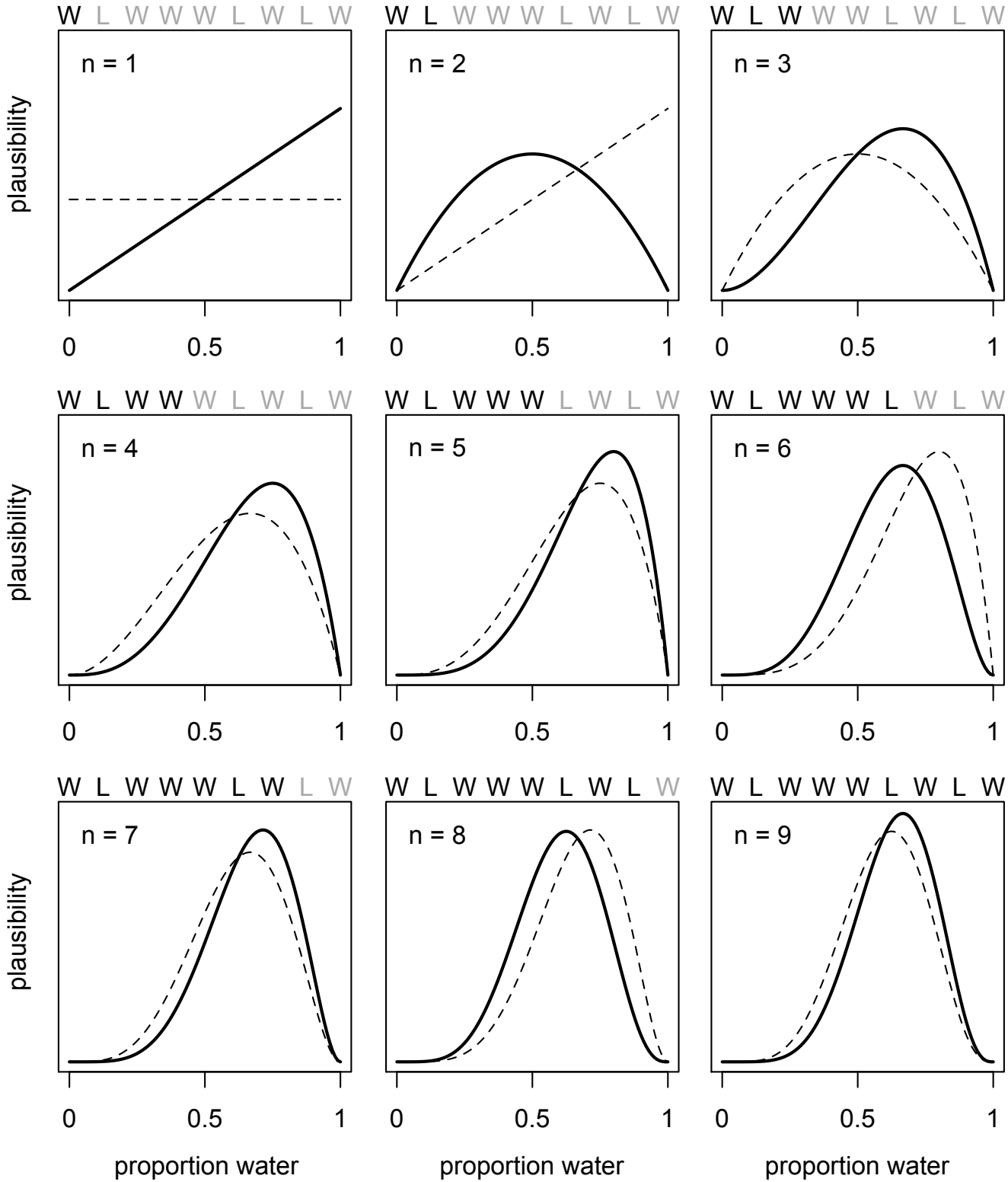

W L W W W L W L W
n = 1
plausibility
posterior
prior
0
0.5
1
p, proportion W


W L W W W L W L W
n = 1
plausibility
0
0.5
1


W L W W W L W L W
n = 1
plausibility
0
0.5
1
W L W W W L W L W
n = 2
0
0.5
1

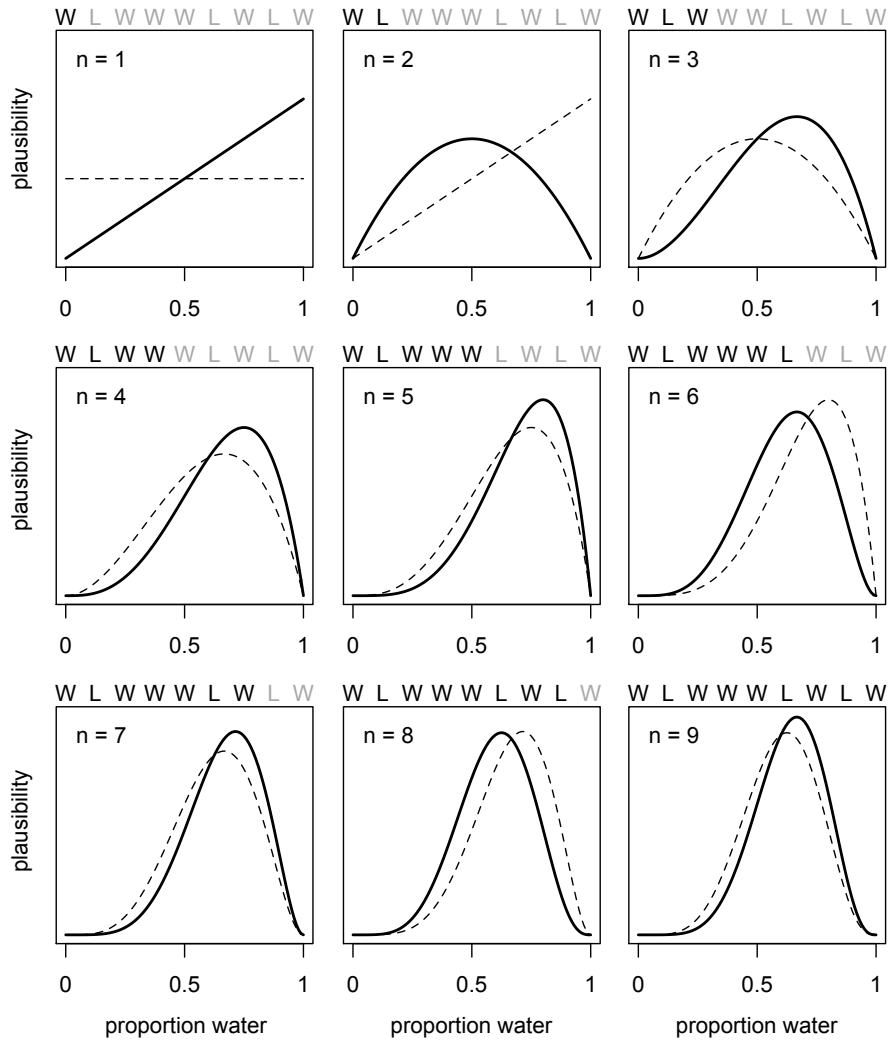

W L W W W L W L W
n = 1
plausibility
0
0.5
1
W L W W W L W L W
n = 2
0
0.5
1
W L W W W L W L W
n = 3
0
0.5
1


W L W W W L W L W
n = 1
plausibility
0
0.5
1
W L W W W L W L W
n = 2
0
0.5
1
W L W W W L W L W
n = 3
0
0.5
1
W L W W W L W L W
n = 4
plausibility
0
0.5
1
W L W W W L W L W
n = 5
0
0.5
1
W L W W W L W L W
n = 6
0
0.5
1


W L W W W L W L W
n = 1
n = 2
n = 3
W L W W W L W L W
n = 4
n = 5
n = 6
W L W W W L W L W
n = 7
n = 8
n = 9
plausibility
0
0.5
1
proportion water

# Design > **Condition** > Evaluate

- Data order irrelevant, because golem assumes order irrelevant
  - All-at-once, one-at-a-time, shuffled order all give same posterior
- Every posterior is a prior for next observation
- Every prior is posterior of some other inference
- Sample size automatically embodied in posterior

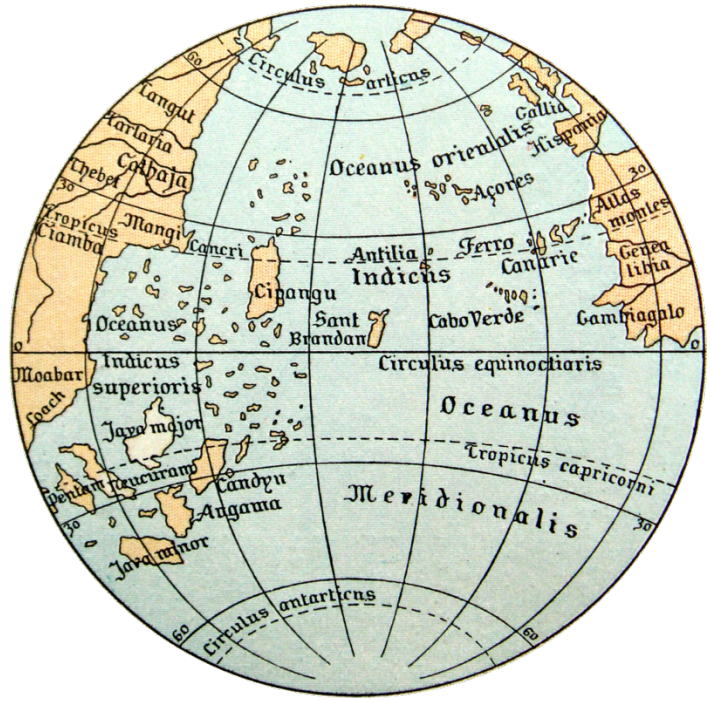


# Design > Condition > **Evaluate**

- Bayesian inference: Logical answer to a question in the form of a model

  *"How plausible is each proportion of water, given these data?"*

- Golem must be supervised
  - Did the golem malfunction?
  - Does the golem's answer make sense?
  - Does the question make sense?
  - Check sensitivity of answer to changes in assumptions

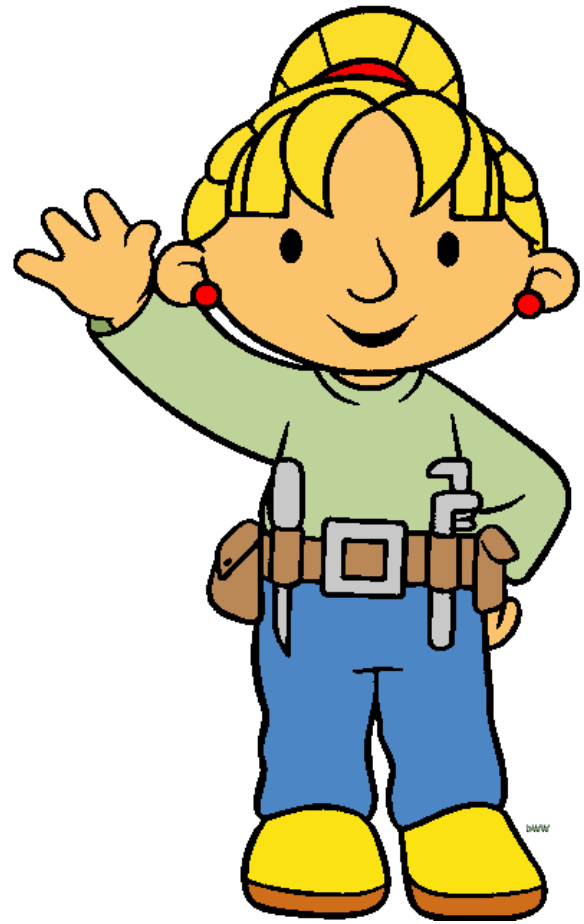


# Construction perspective

- Build joint model:
  (1) List variables
  (2) Define generative relations
  (3) ???
  (4) Profit

- Input: Joint prior

- Deduce: Joint posterior



# The Joint Model

$$W \sim \text{Binomial}(N, p)$$

$$p \sim \text{Uniform}(0, 1)$$

- Bayesian models are **generative**
- Can be run **forward** to generate predictions or simulate date
- Can be run in **reverse** to infer process from data

# The Joint Model

$$W \sim \text{Binomial}(N, p)$$

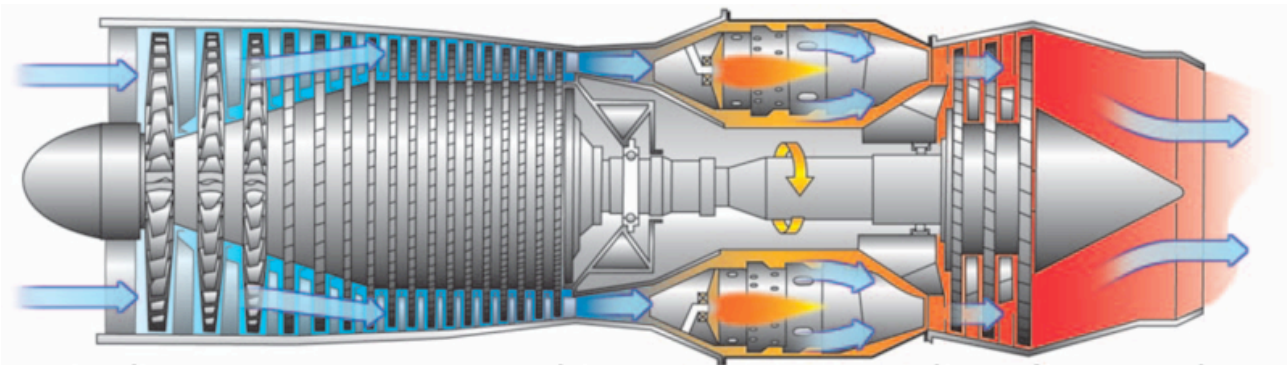
$$p \sim \text{Uniform}(0, 1)$$

- Run **forward**:

```
> N <- 9
> p <- runif( 1e4 , 0 , 1 )
> W <- rbinom( 1e4 , size=N , prob=p )
> table(W)
W
   0    1    2    3    4    5    6    7    8    9
1013 1000  999 1014 1094  999  982  971  980  948
```

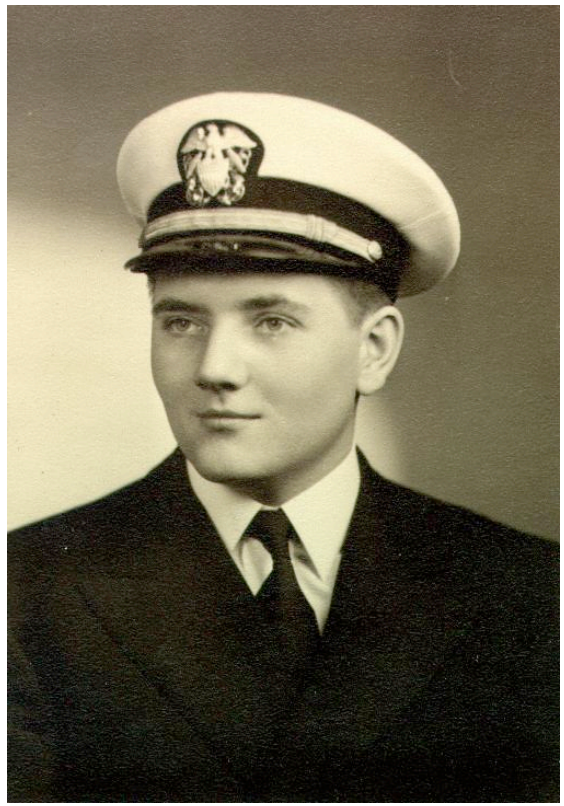
# Run in Reverse: Computing the posterior

1. Analytical approach (often impossible)
2. Grid approximation (very intensive)
3. Quadratic approximation (limited)
4. Markov chain Monte Carlo (intensive)



# Predictive checks

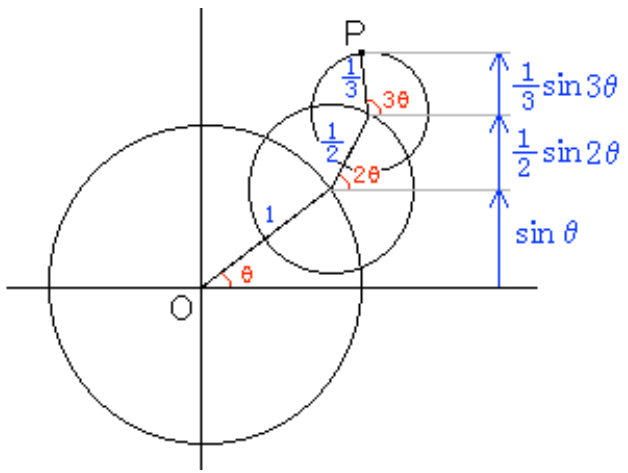
- Something like a *significance test*, but not
- No universally best way to evaluate adequacy of model-based predictions
- No way to justify always using a threshold like 5%
- Good predictive checks always depend upon purpose and imagination

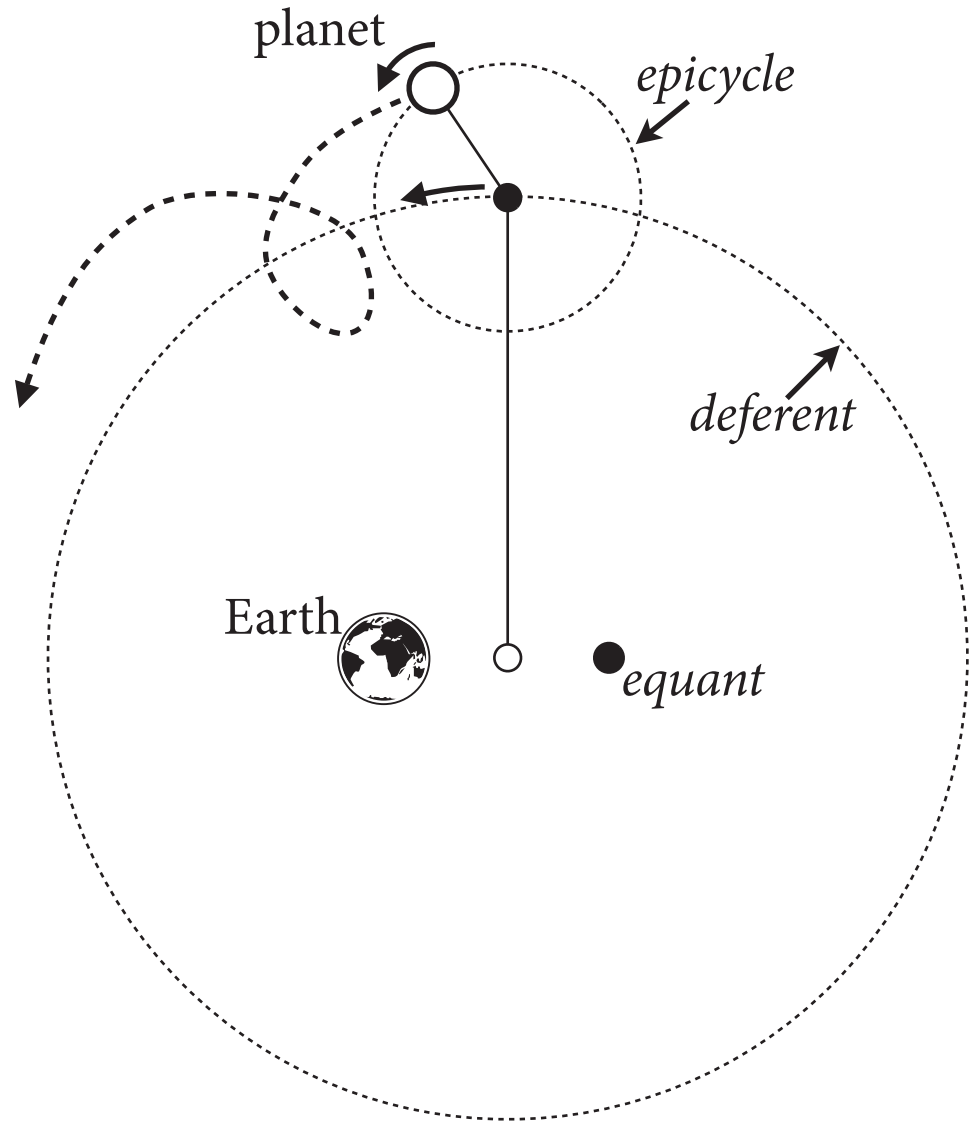


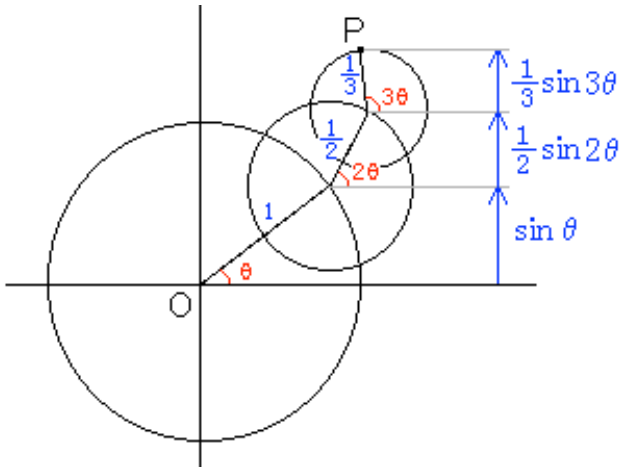
"It would be very nice to have a formal apparatus that gives us some 'optimal' way of recognizing unusual phenomena and inventing new classes of hypotheses [...]; **but this remains an art for the creative human mind.**"
—E.T. Jaynes (1922–1998)

# Triumph of Geocentrism

- Claudius Ptolemy (90–168)
  - Egyptian mathematician
  - Accurate model of planetary motion
  - Epicycles: orbits on orbits
  - Fourier series

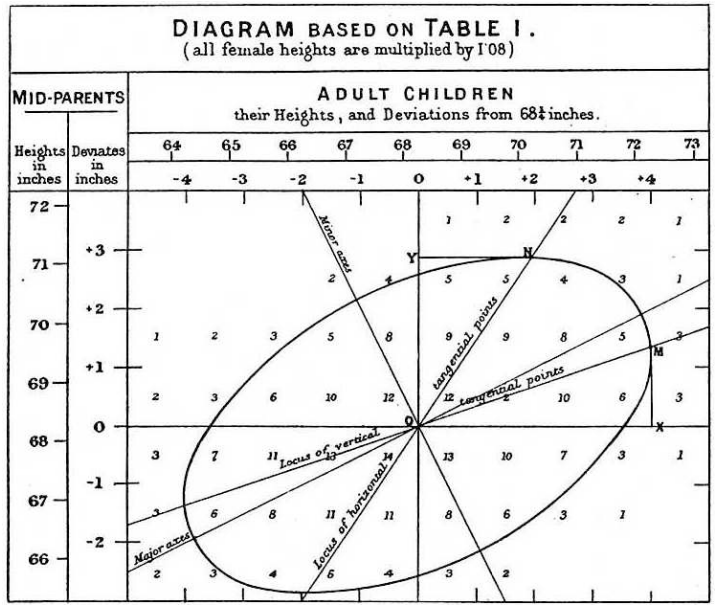





# Geocentrism

- Descriptively accurate
- Mechanistically wrong
- General method of approximation
- Known to be wrong

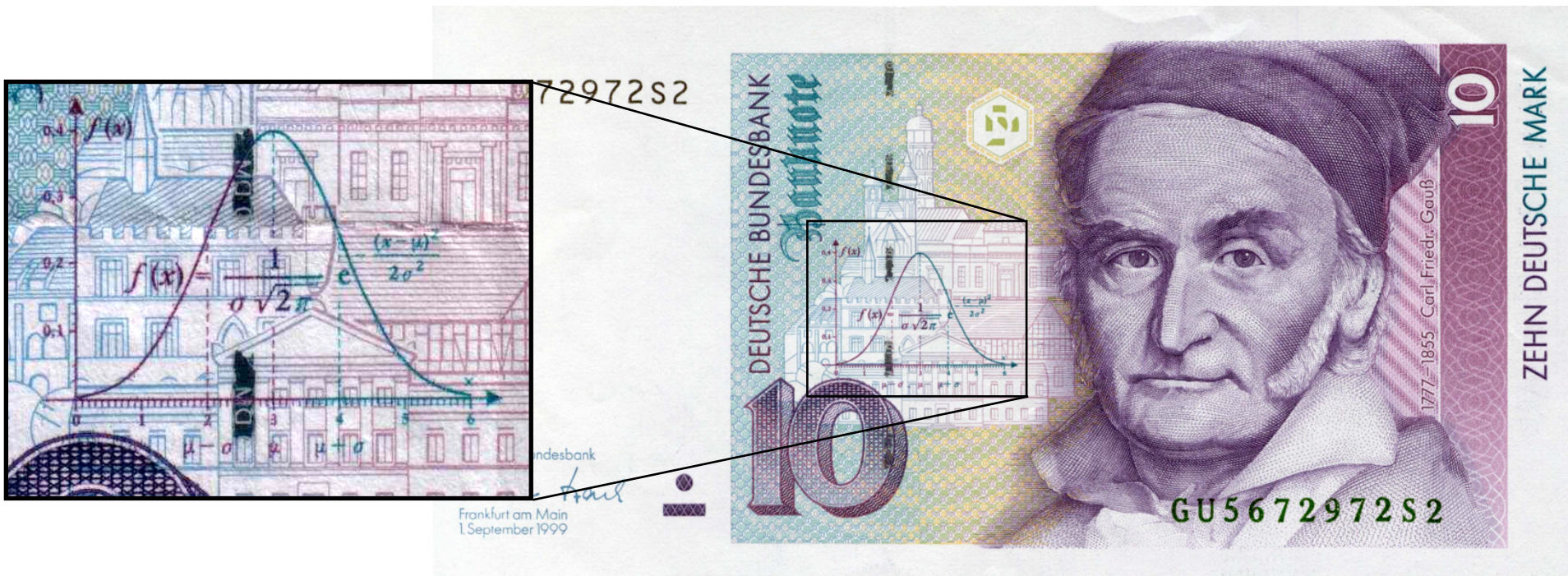



# Regression

- Descriptively accurate
- Mechanistically wrong
- General method of approximation
- Taken too seriously

# Linear regression

- Simple statistical golems
  - Model of mean and variance of normally (Gaussian) distributed measure
  - Mean as *additive* combination of *weighted* variables
  - Constant variance



1809 Bayesian argument for normal error and least-squares estimation

# THEORIA
# MOTVS CORPORVM
# COELESTIVM

IN

SECTIONIBVS CONICIS SOLEM AMBIENTIVM

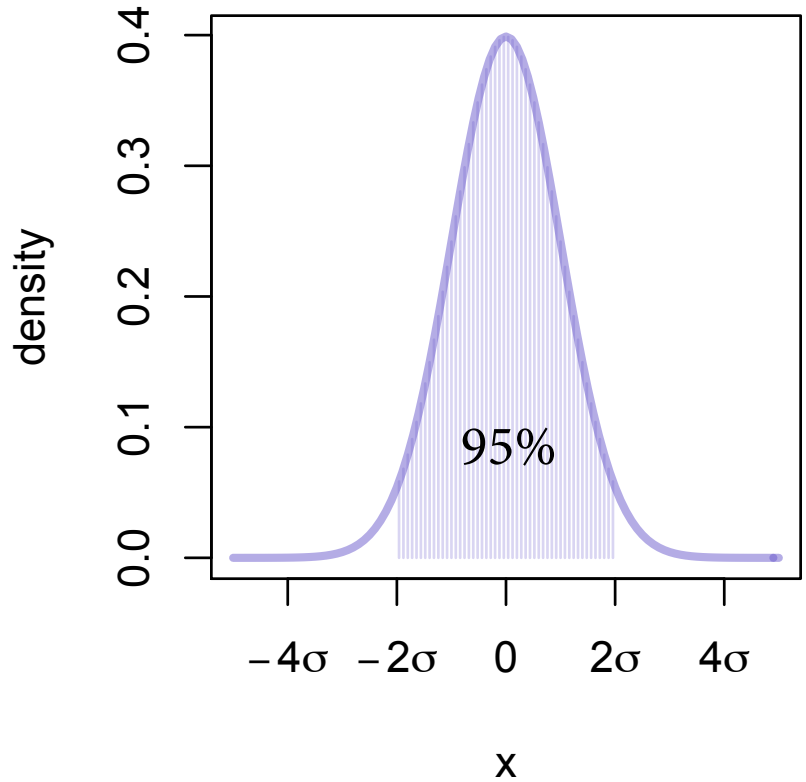
AVCTORE

CAROLO FRIDERICO GAVSS

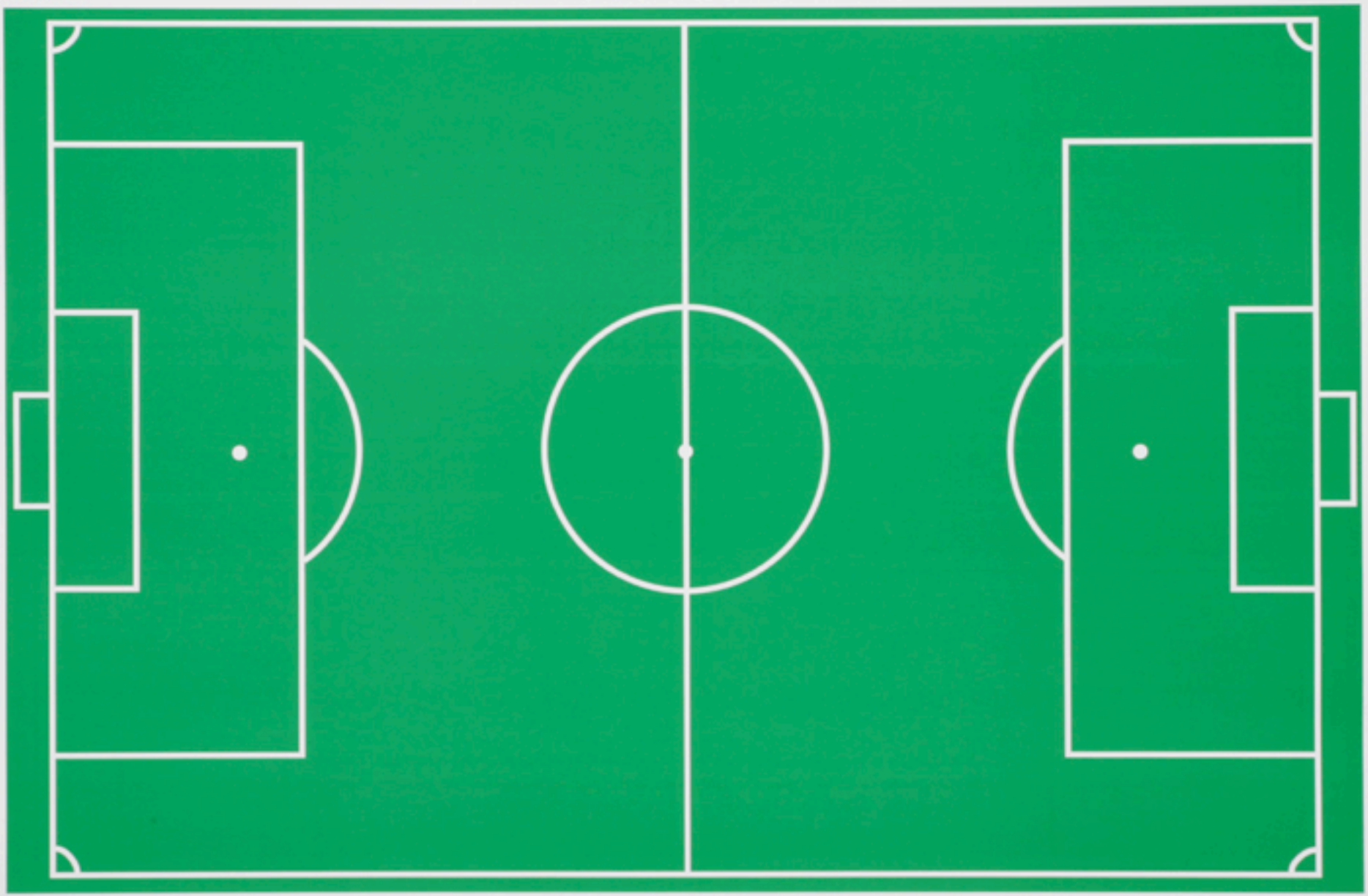
HAMBVRGI SVMTIBVS FRID. PERTHES ET I. H. BESSER
1809.

# Why normal?

- Why are normal (Gaussian) distributions so common in statistics?
  1. Easy to calculate with
  2. Common in nature
  3. Very conservative assumption



Football

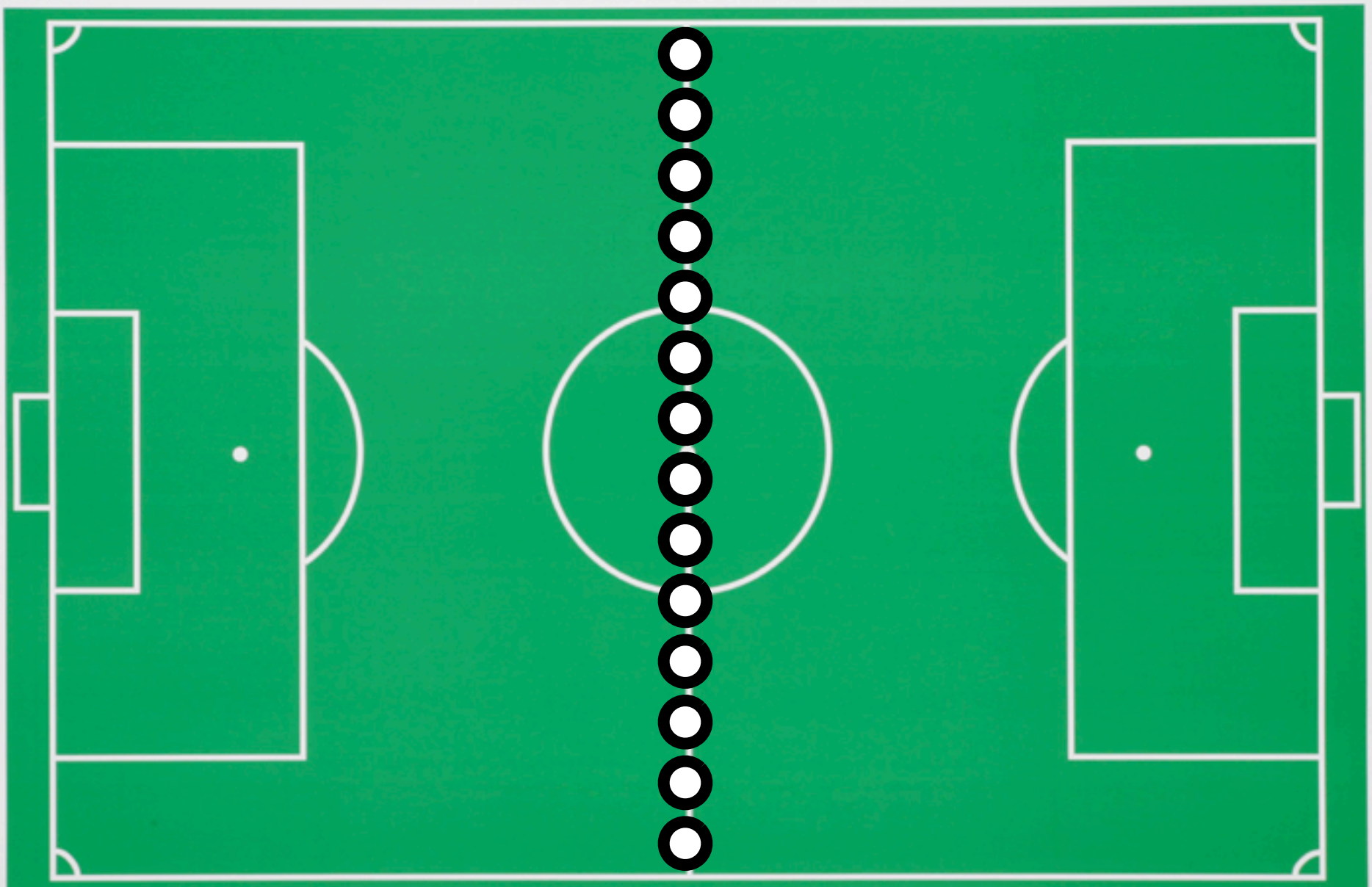


***Game Plan* - Strategy Overlay**

Use in conjunction with a magnetic board such as the A2 Folding Wedge, use only dry-wipe markers, clean with a soft cloth as harsh or abrasive solvents will damage the surface.

www.pnpsports.co.uk

Football

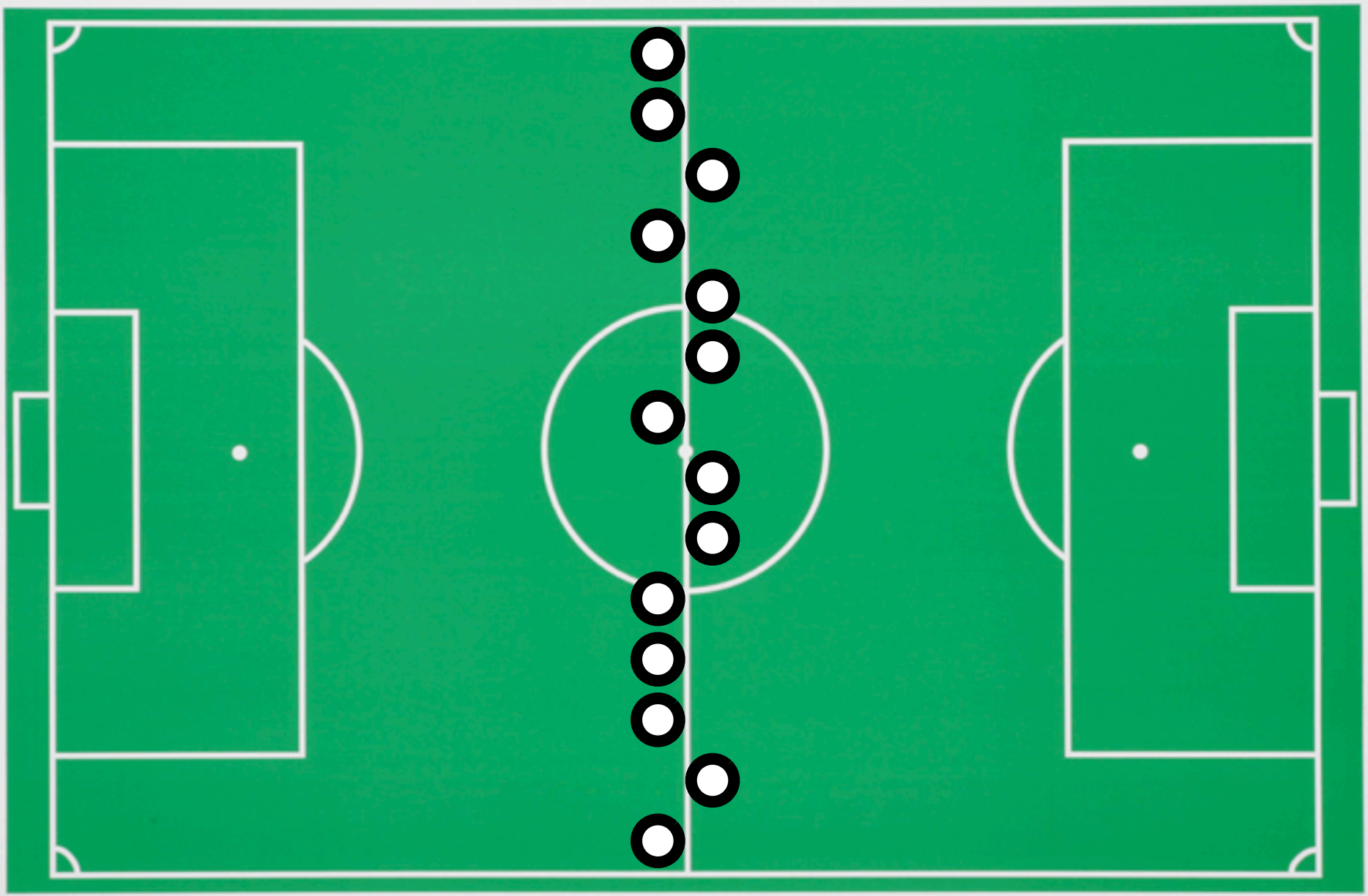


*Game Plan* - Strategy Overlay

Use in conjunction with a magnetic board such as the A2 Folding Wedge, use only dry-wipe markers, clean with a soft cloth as harsh or abrasive solvents will damage the surface.

www.pnpsports.co.uk

Football

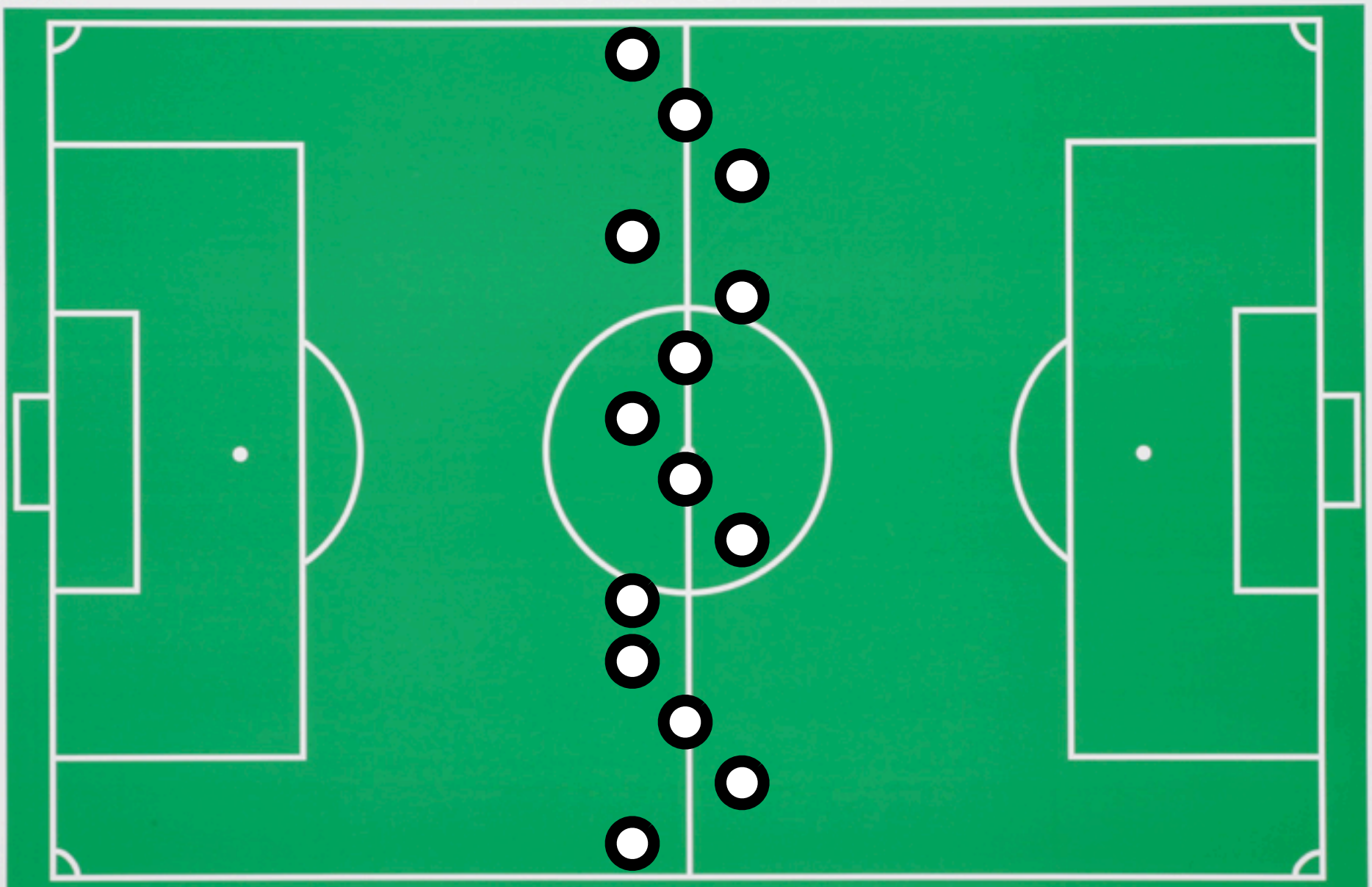


*Game Plan* - Strategy Overlay

Use in conjunction with a magnetic board such as the A2 Folding Wedge, use only dry-wipe markers, clean with a soft cloth as harsh or abrasive solvents will damage the surface.

www.pnpsports.co.uk

Football

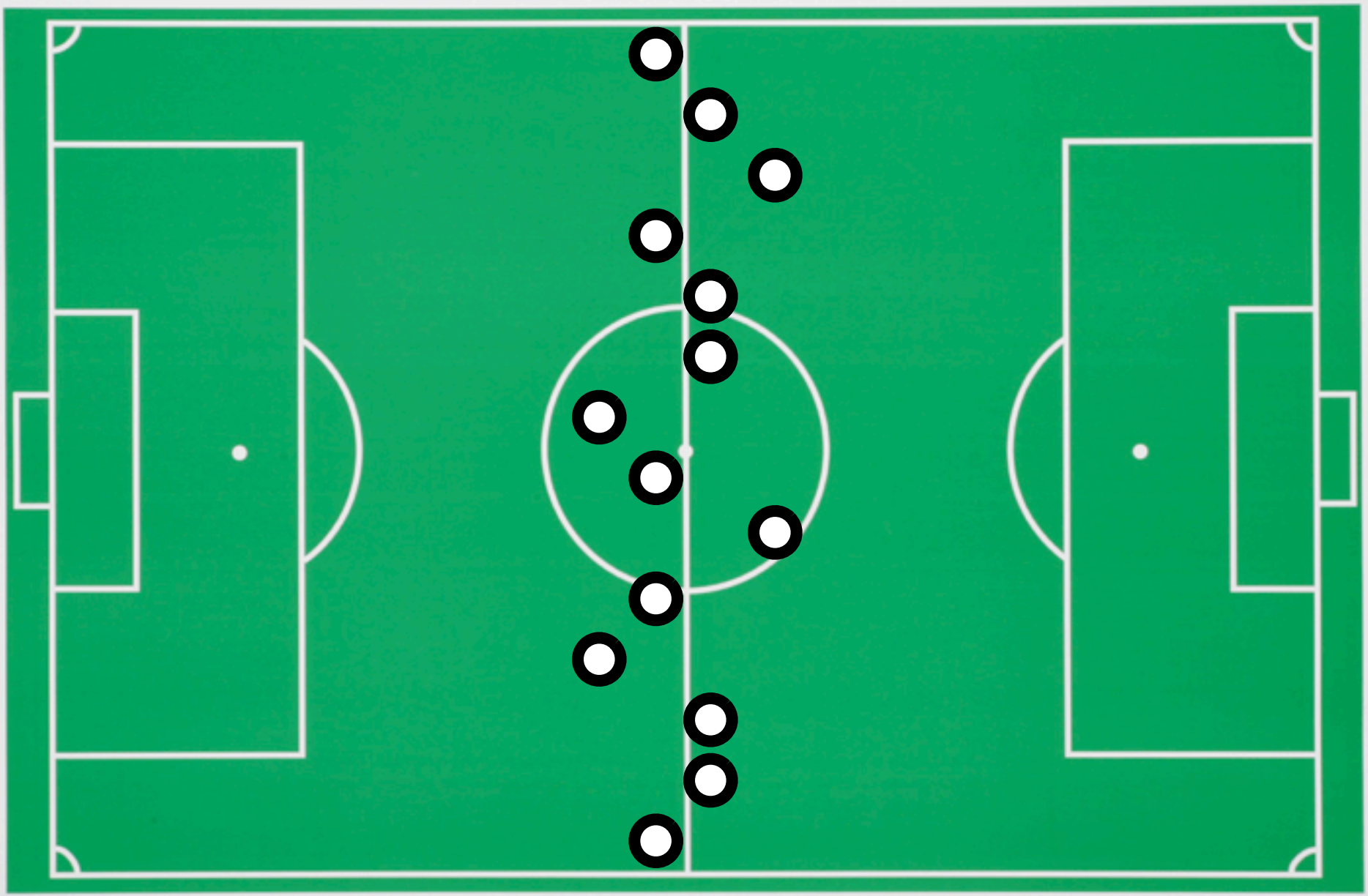


*Game Plan* - Strategy Overlay

Use in conjunction with a magnetic board such as the A2 Folding Wedge, use only dry-wipe markers, clean with a soft cloth as harsh or abrasive solvents will damage the surface.

www.pnpsports.co.uk

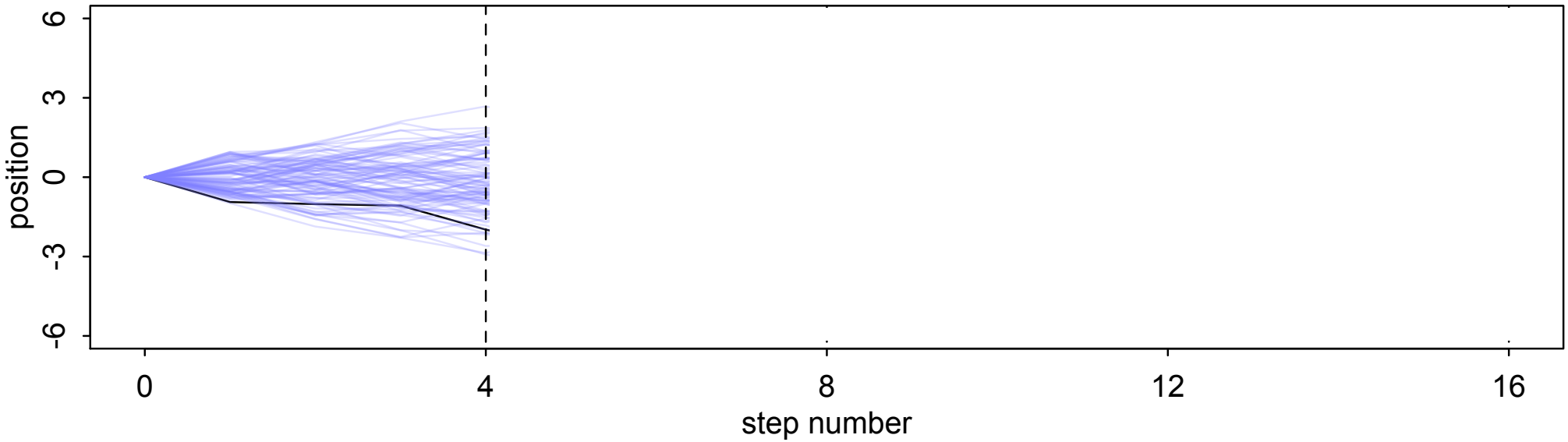
Football



*Game Plan* - Strategy Overlay

Use in conjunction with a magnetic board such as the A2 Folding Wedge, use only dry-wipe markers, clean with a soft cloth as harsh or abrasive solvents will damage the surface.

www.pnpsports.co.uk


position
6
3
0
-3
-6
0
4
8
12
16
step number

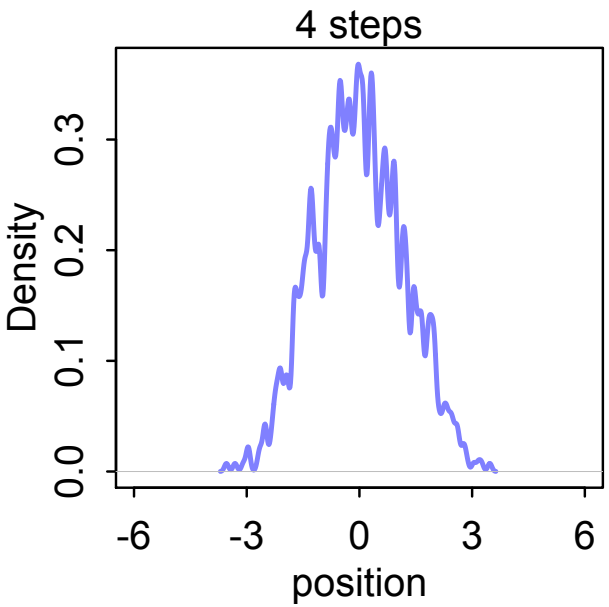

4 steps
Density
0.0
0.1
0.2
0.3
-6
-3
0
3
6
position

Figure 4.2

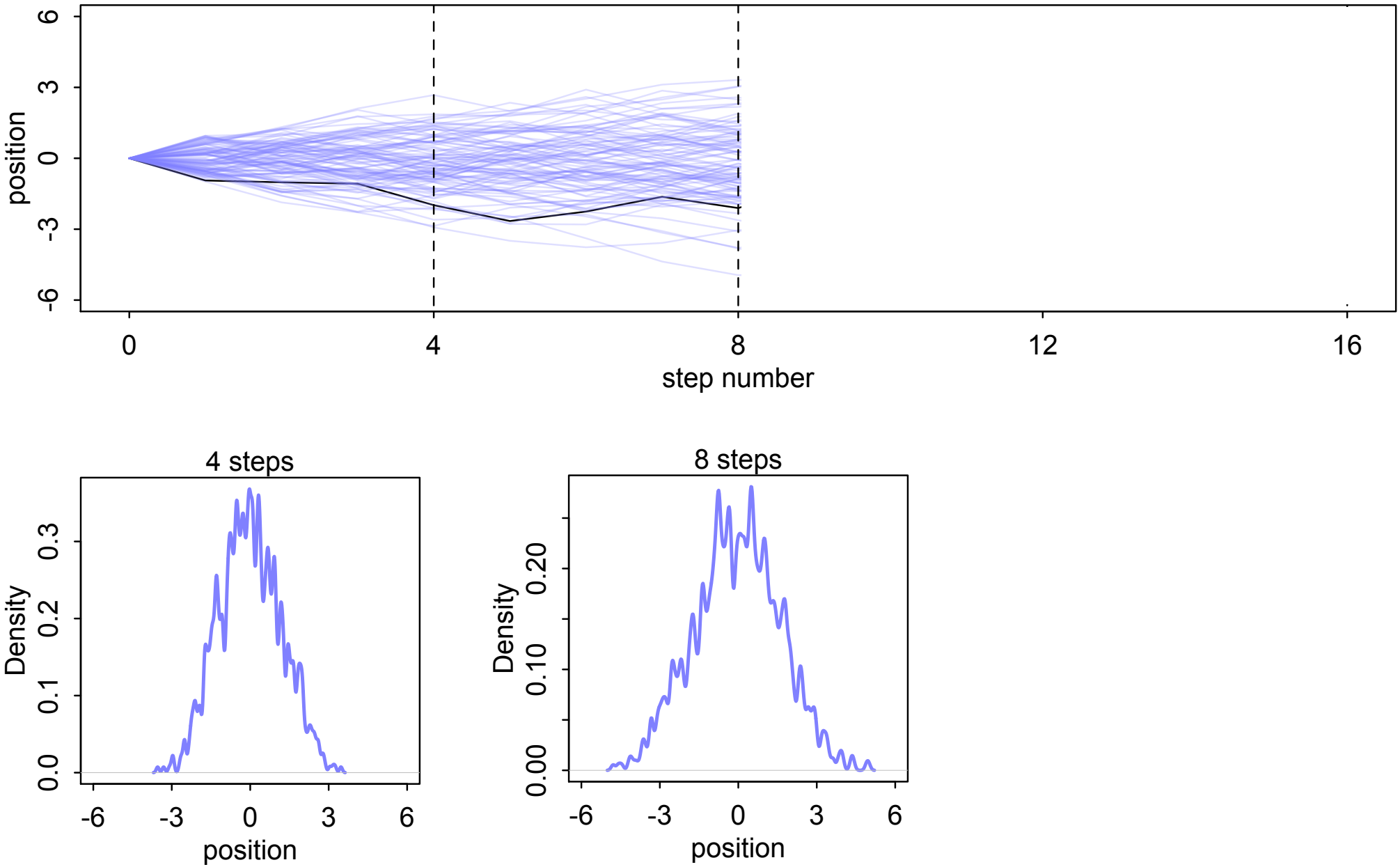

position
6
3
0
-3
-6
0
4
8
12
16
step number
4 steps
Density
0.0
0.1
0.2
0.3
-6
-3
0
3
6
position
8 steps
Density
0.00
0.10
0.20
-6
-3
0
3
6
position

Figure 4.2

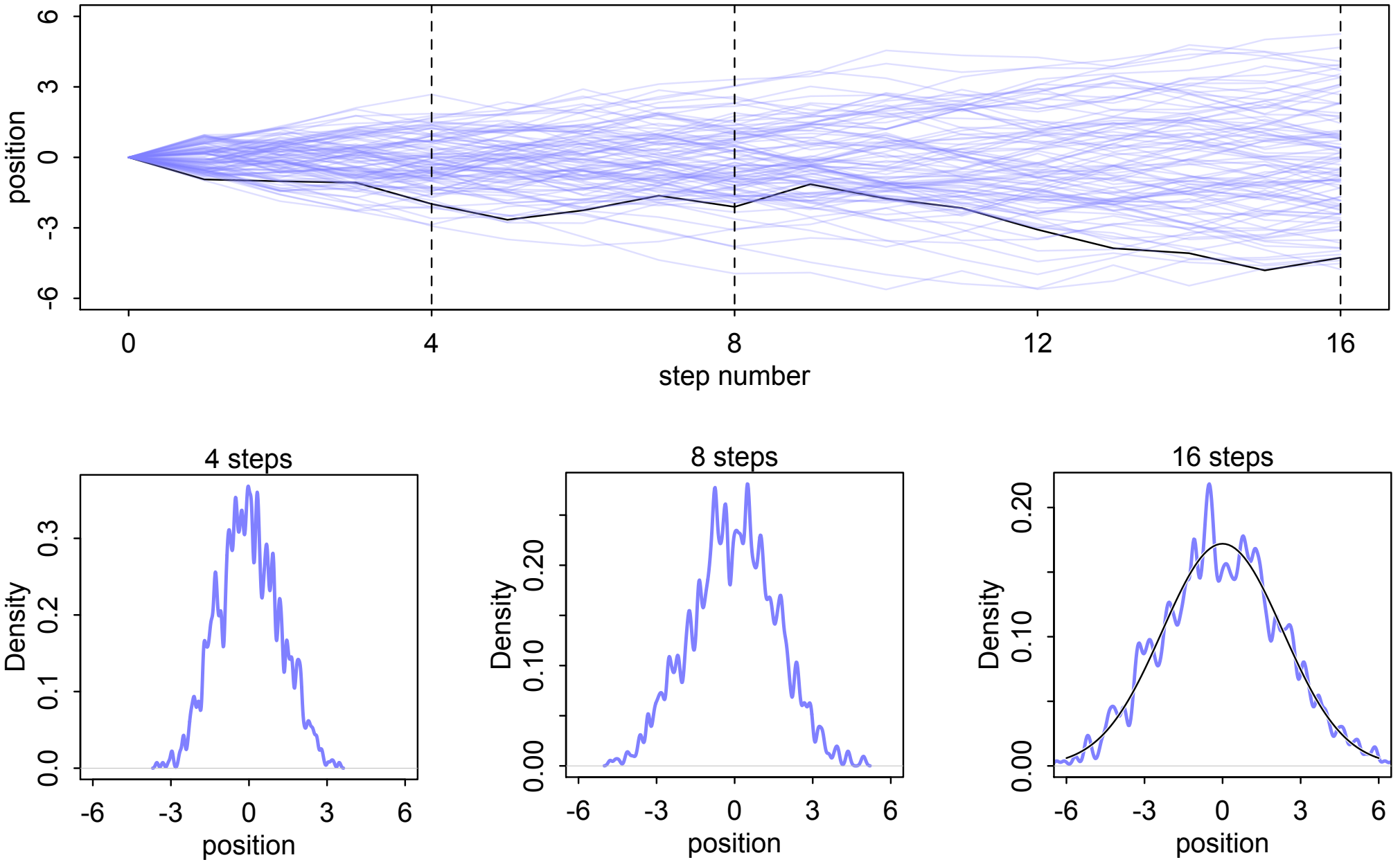
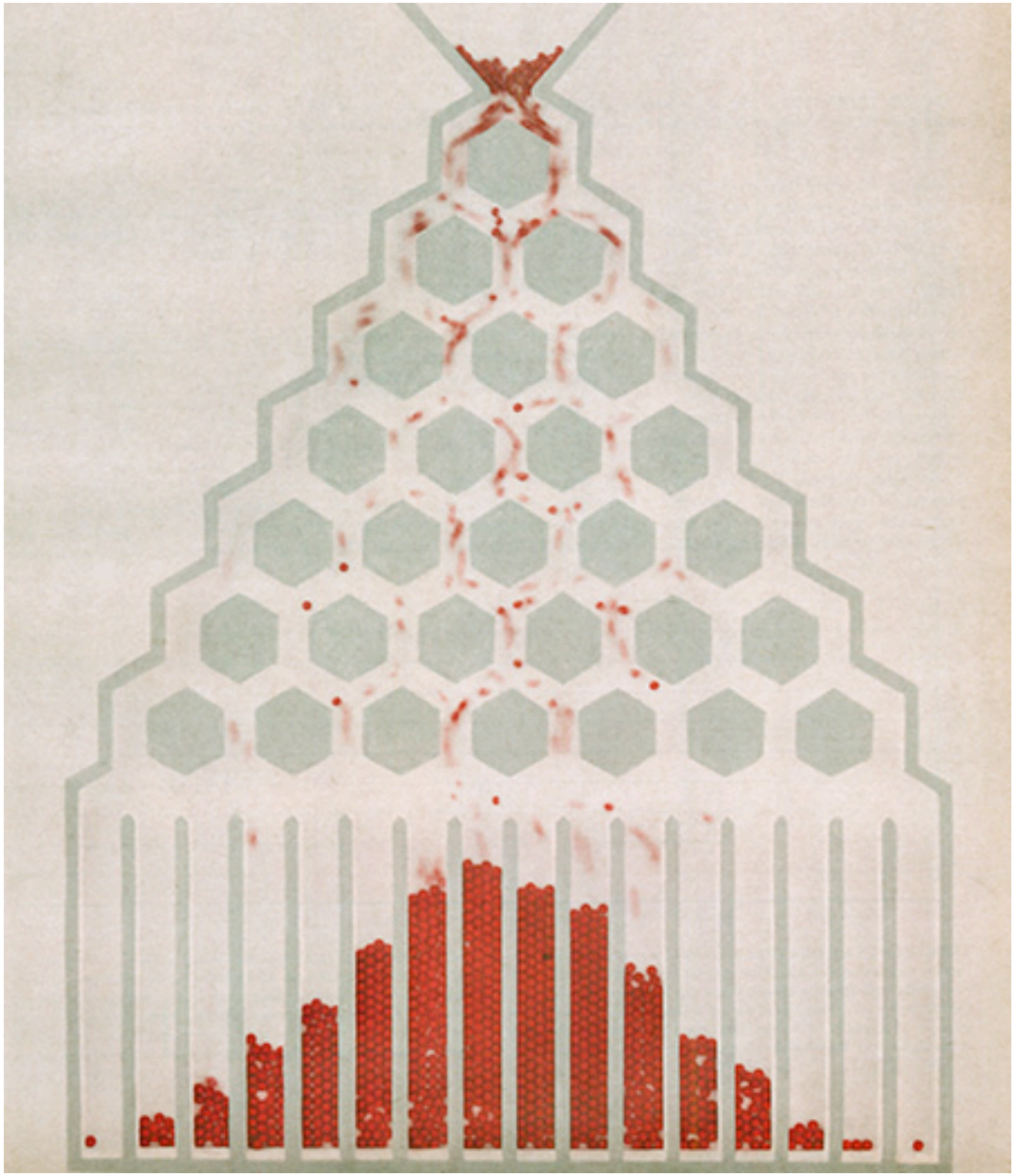

position
6
3
0
-3
-6
0
4
8
12
16
step number
4 steps
8 steps
16 steps
Density
0.0
0.1
0.2
0.3
0.00
0.10
0.20
-6
-3
0
3
6
position

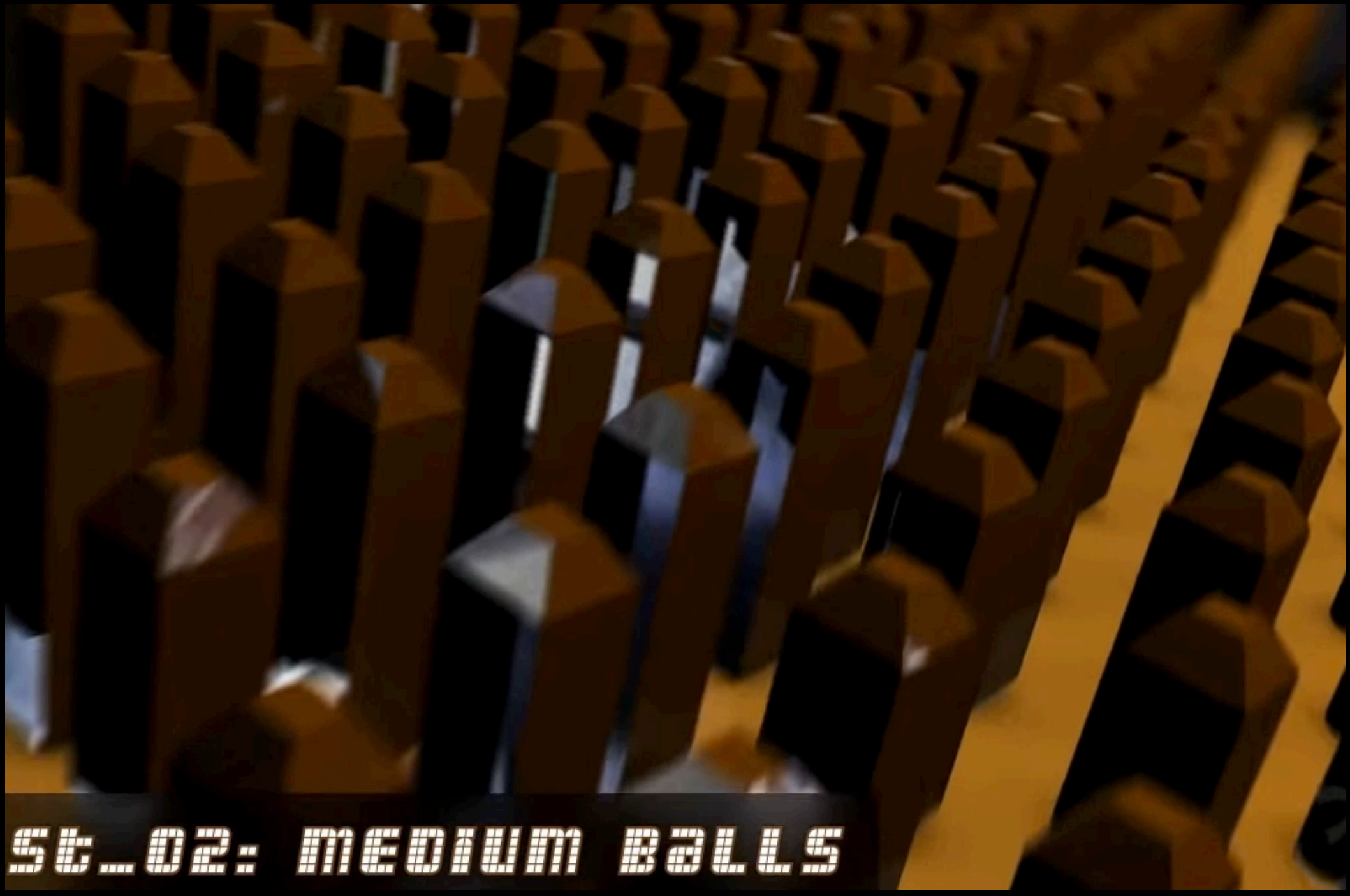
Figure 4.2

# Why normal?

- Processes that produce normal distributions
  - Addition
  - Products of small deviations
  - Logarithms of products

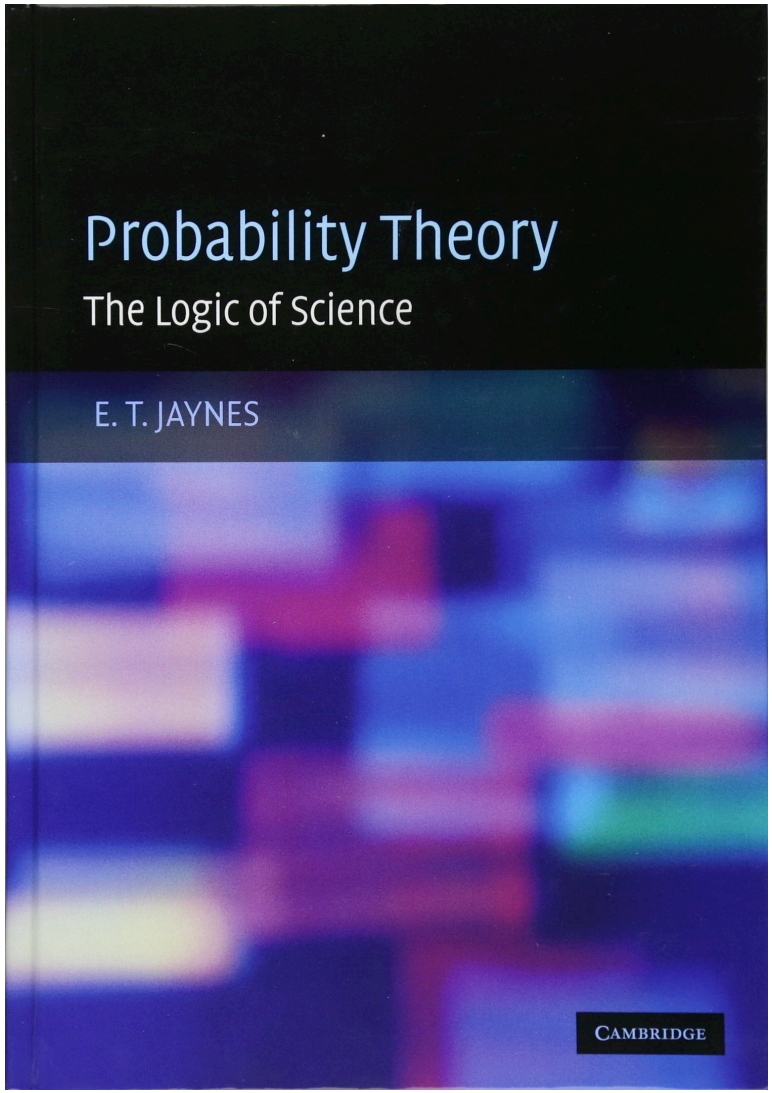


Francis Galton's 1894 "bean machine" for simulating normal distributions


St_02: MEDIUM BALLS

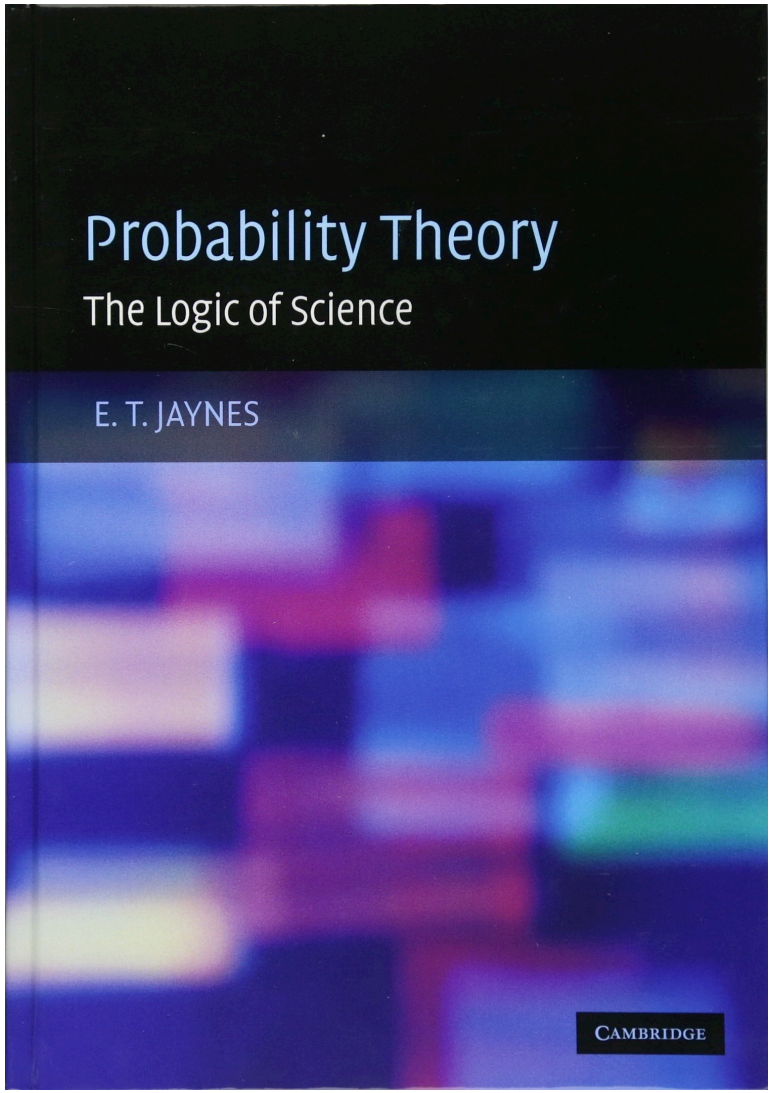
# Why normal?

- Ontological perspective
  - Processes which add fluctuations result in dampening
  - Damped fluctuations end up Gaussian
  - No information left, except mean and variance
  - Can't infer process from distribution!



# Why normal?

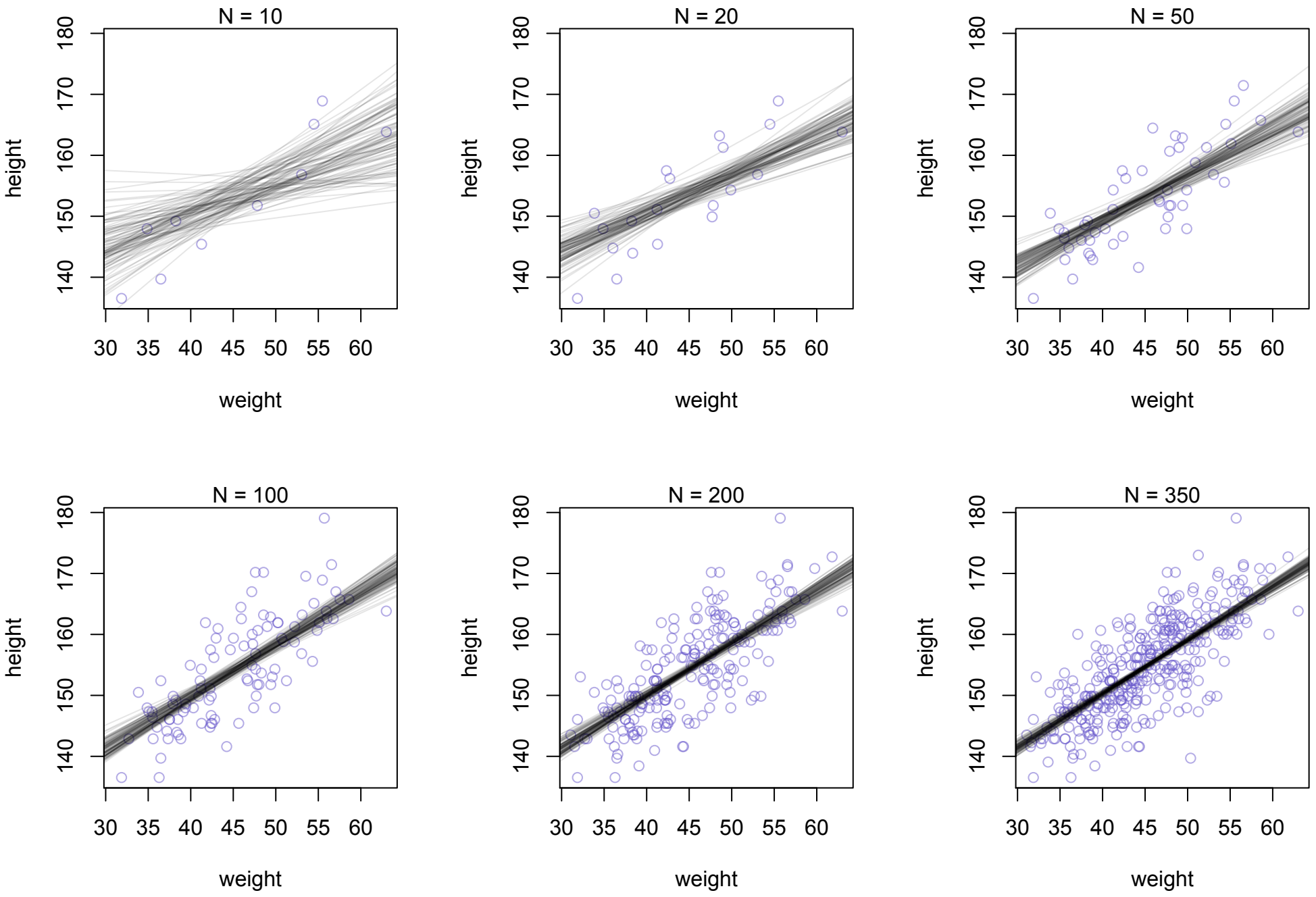
- Ontological perspective
  - Processes which add fluctuations result in dampening
  - Damped fluctuations end up Gaussian
  - No information left, except mean and variance
  - Can't infer process from distribution!
- Epistemological perspective
  - Know only *mean* and *variance*
  - Then least surprising and most conservative (*maximum entropy*) distribution is Gaussian
  - Nature likes maximum entropy distributions

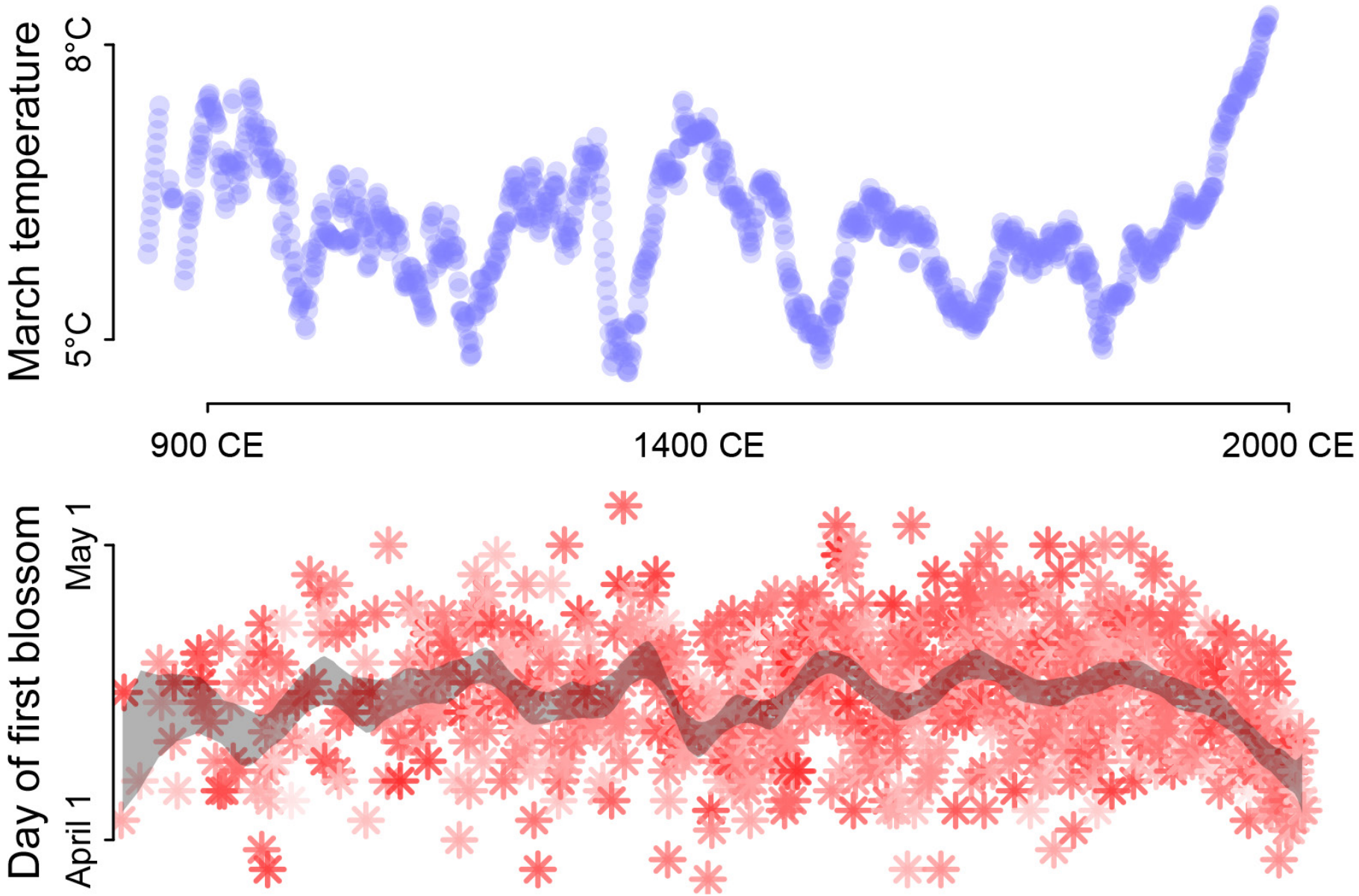


# Linear models

- Models of normally distributed data common
  - "General Linear Model": *t*-test, single regression, multiple regression, ANOVA, ANCOVA, MANOVA, MANCOVA, yadda yadda yadda
  - All the same thing
- Learn strategy, not procedure

Willard Boepple


N = 10
N = 20
N = 50
N = 100
N = 200
N = 350
height
weight
140
150
160
170
180
30
35
40
45
50
55
60

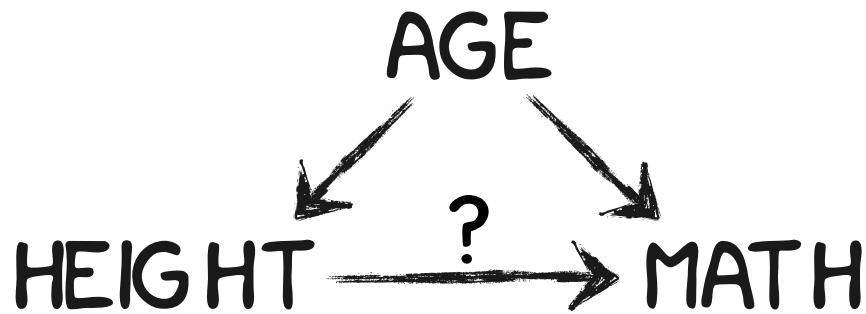

March temperature
8°C
5°C
900 CE
1400 CE
2000 CE
Day of first blossom
May 1
April 1

# Regression as a wicked oracle

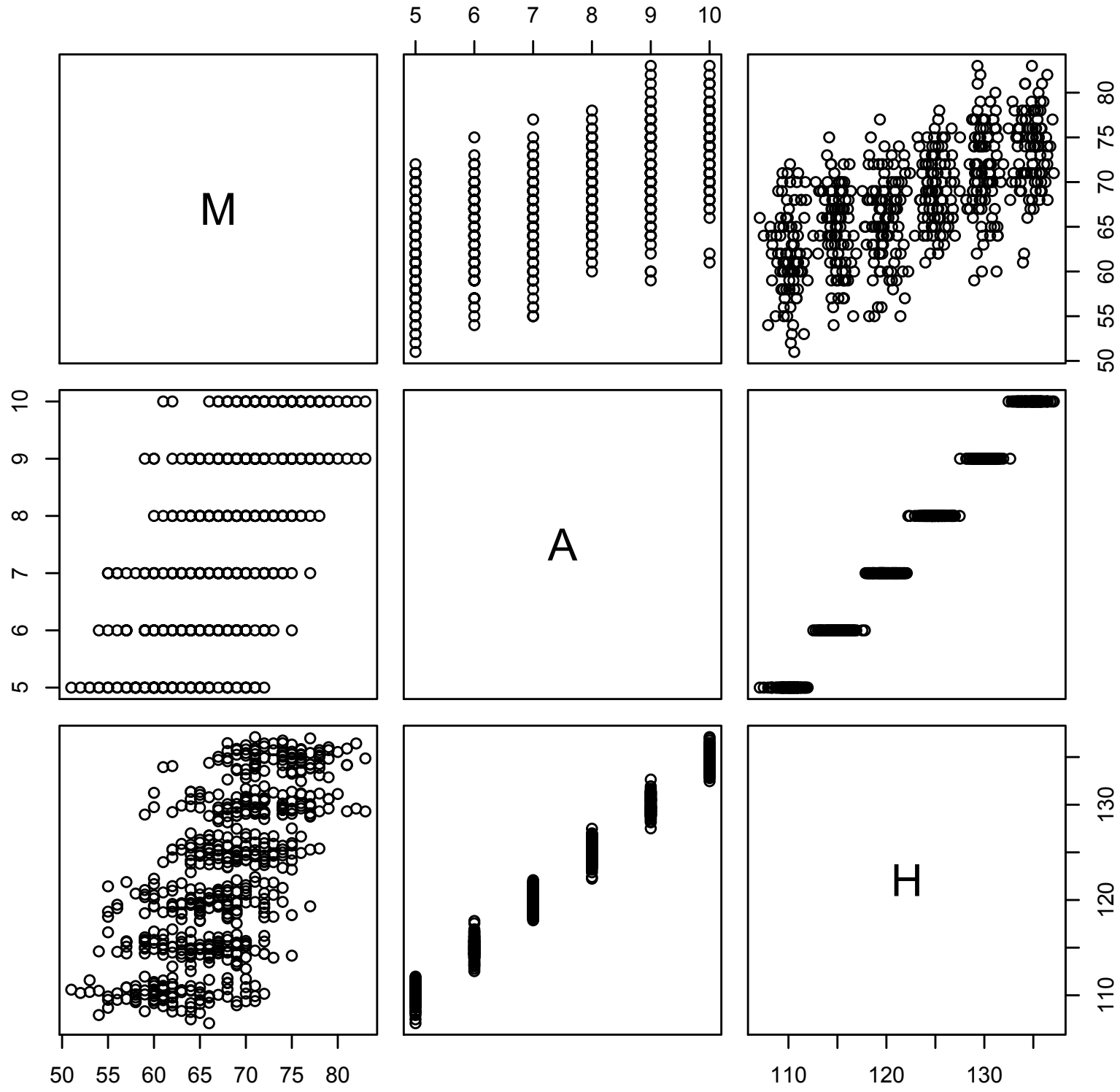
- Regression automatically focuses on the most informative cases
- Cases that don't help are automatically ignored
- But not kind — ask carefully

# Why not just add everything?

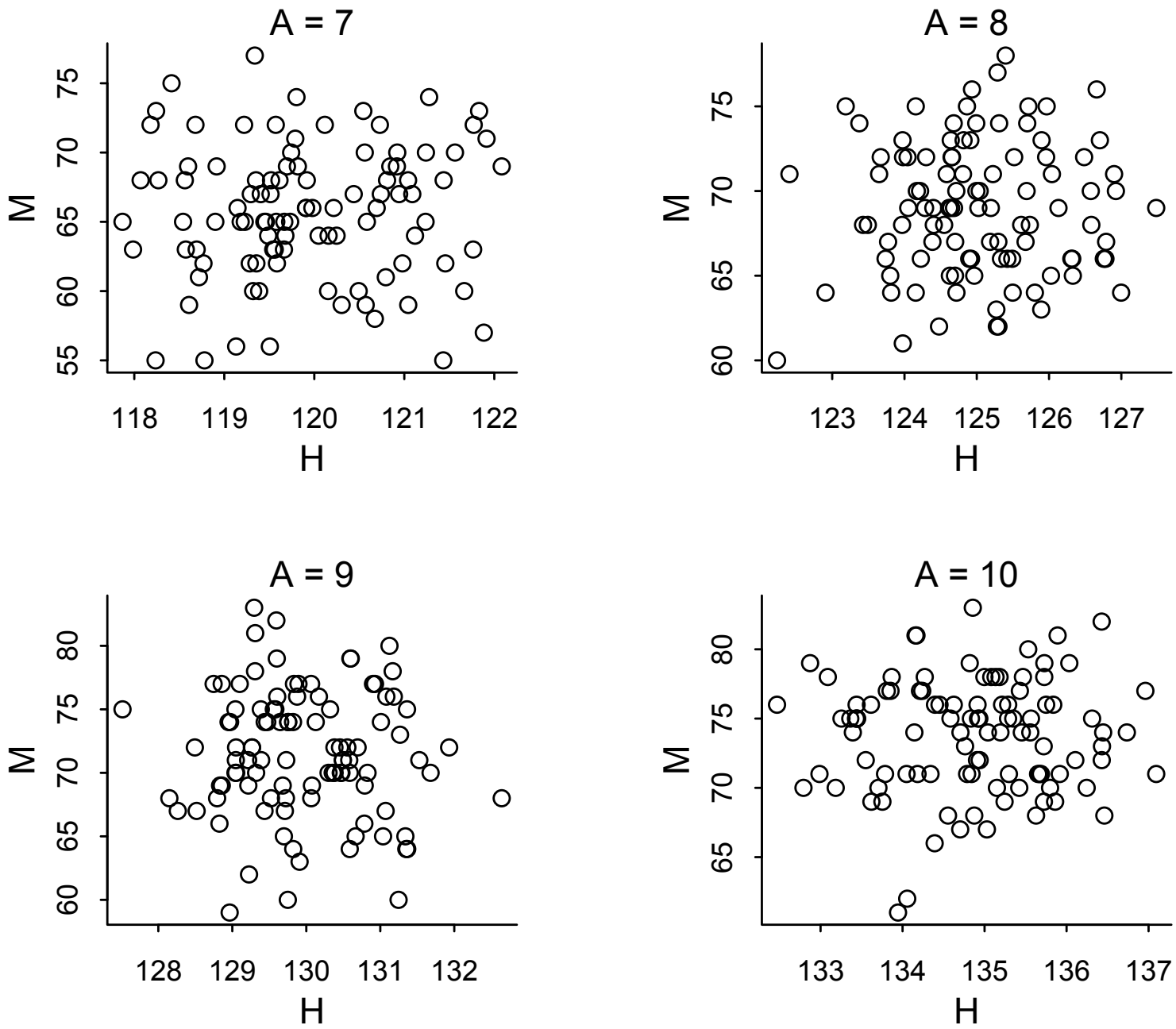
- Could just add all available predictors to model
  - "We controlled for..."
- Almost always a bad idea
  - Adding variables *creates* confounds
  - Residual confounding
  - Overfitting



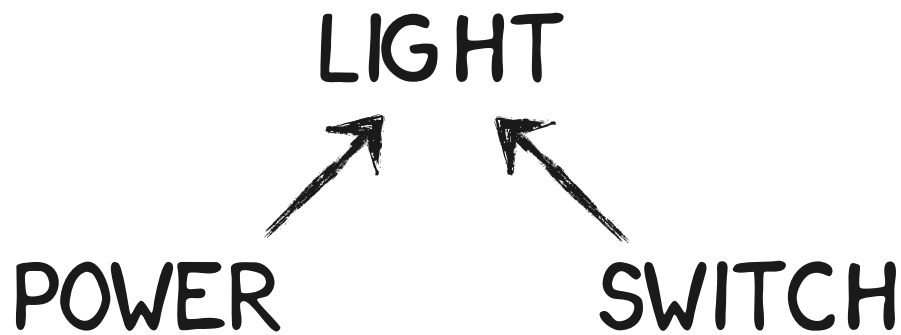

MATH independent of HEIGHT, conditional on AGE


M
A
H
5
6
7
8
9
10
50
55
60
65
70
75
80
110
120
130

# MATH independent of HEIGHT, conditional on AGE



SWITCH independent of POWER



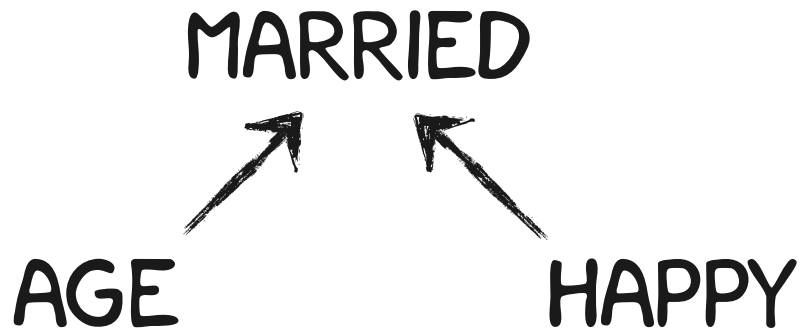

SWITCH dependent on POWER, conditional on LIGHT

| LIGHT | POWER | SWITCH |
|---|---|---|
| ON | ON | ? |
| OFF | ON | ? |

SWITCH dependent on POWER, conditional on LIGHT

**This effect known as "collider bias"**

HAPPY independent of AGE




HAPPY dependent on AGE, conditional on MARRIED

# Why not just add everything?

- Matters for experiments as well
  - Conditioning on post-treatment variables can be very bad
  - Conditioning on pre-treatment can also be bad (colliders)
- Good news!
  - Causal inference possible in observational settings
  - But requires good theory

# JUST COUNTING IMPLICATIONS OF ASSUMPTIONS