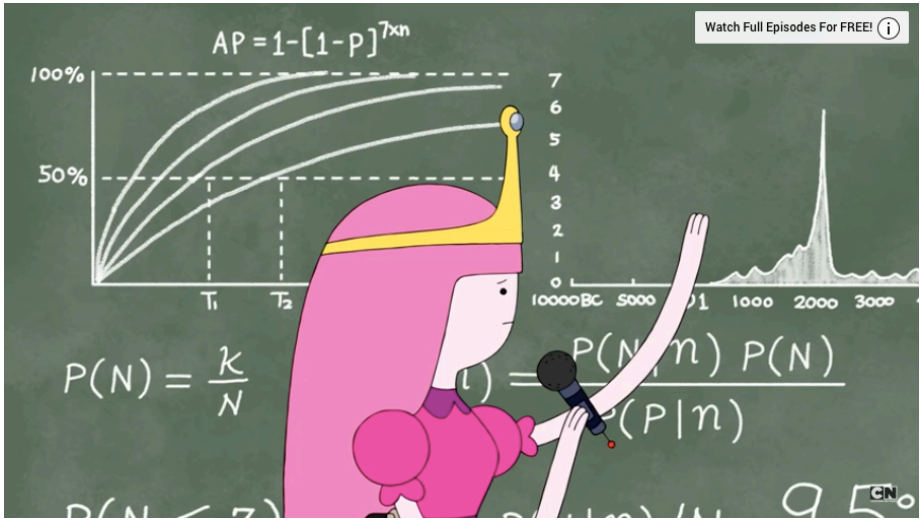
inferencia basada en la probabilidad y la estadística

Germán Chaparro Molano
UNIVERSIDAD ECCI, COLOMBIA



¿Quién soy yo?

- Profesor Titular del Grupo de Simulación, Análisis y Modelado - Universidad ECCI
- Editor en Jefe, Revista Tecciencia de Ingeniería y Ciencias Aplicadas

- PhD en Astronomía Rijksuniversiteit Groningen (Países Bajos)
- MSc en Astronomía Universiteit Leiden (Países Bajos)
- Físico Universidad Nacional de Colombia

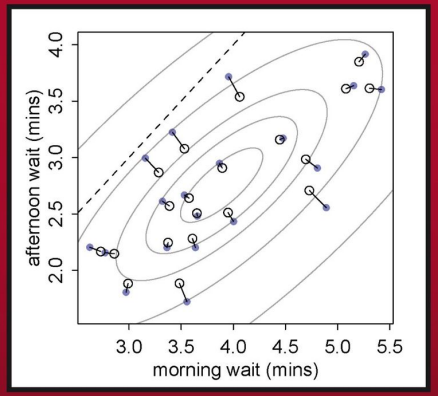
- Coordinador Universidades e Investigación Oficina Andina de Astronomía para el Desarrollo
- Miembro de la Unión Astronómica Internacional
- Miembro de AstroCO, Nodo de Astronomía de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales

- Áreas de Investigación: Radioastronomía, Astroestadística, Inferencia Bayesiana, Machine Learning, Formación Estelar y Planetaria, Astroquímica, Astronomía para el Desarrollo

Texts in Statistical Science

Statistical Rethinking

A Bayesian Course with Examples in R and Stan



Richard McElreath

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

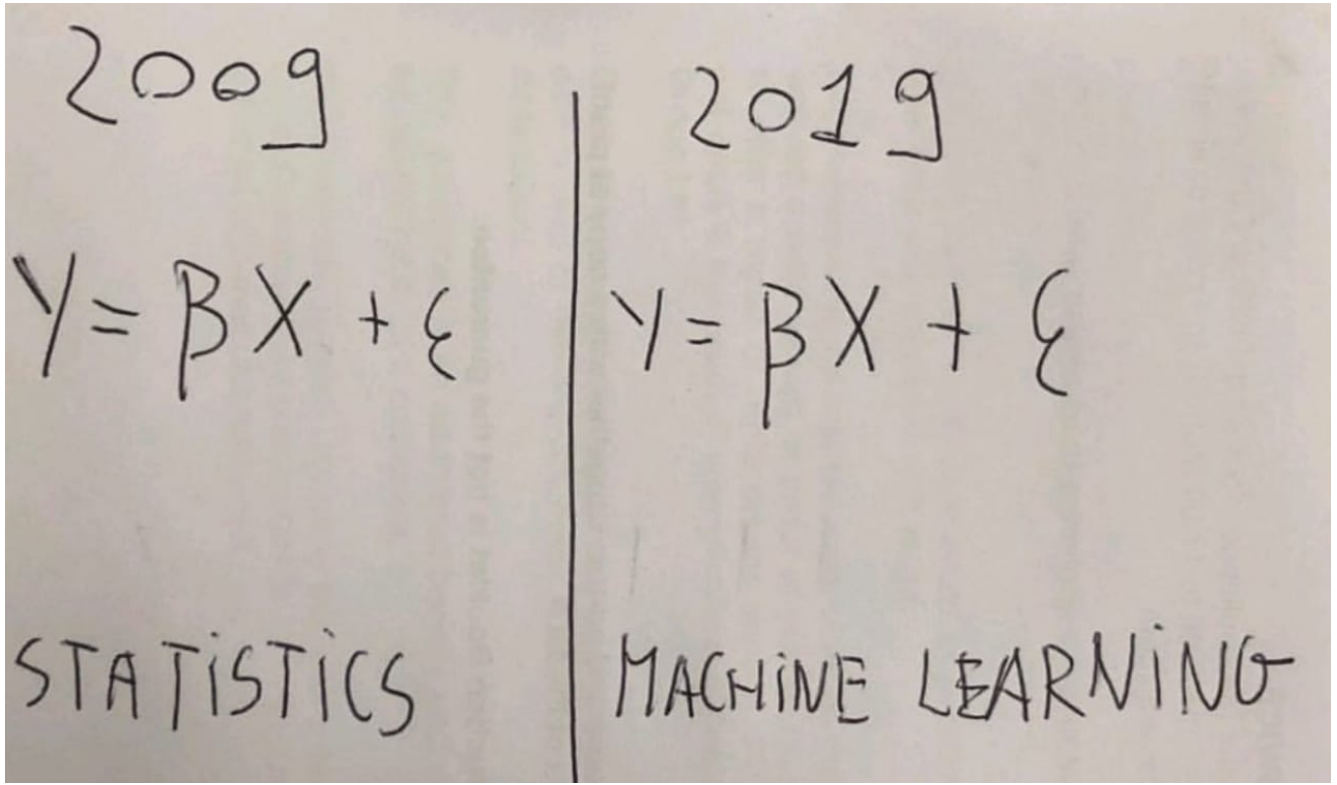
Modern Statistical Methods for Astronomy

With R Applications

Eric D. Feigelson
G. Jogesh Babu

#10yearchallenge

#10yearchallenge




THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

¿Ciencia + Estadística?

Los pesimistas

"There is no need for these hypotheses to be true, or even to be at all like the truth; rather … they should yield calculations which agree with observations" (Osiander's Preface to Copernicus' De Revolutionibus, quoted by C. R. Rao)

"Essentially, all models are wrong, but some are useful." (Box & Draper 1987)

¿Ciencia + Estadística?

Los optimistas

"The goal of science is to unlock nature's secrets. … Our understanding comes through the development of theoretical models which are capable of explaining the existing observations as well as making testable predictions. … Fortunately, a variety of sophisticated mathematical and computational approaches have been developed to help us through this interface, these go under the general heading of statistical inference."

(P. C. Gregory, Bayesian Logical Data Analysis for the Physical Sciences, 2005)

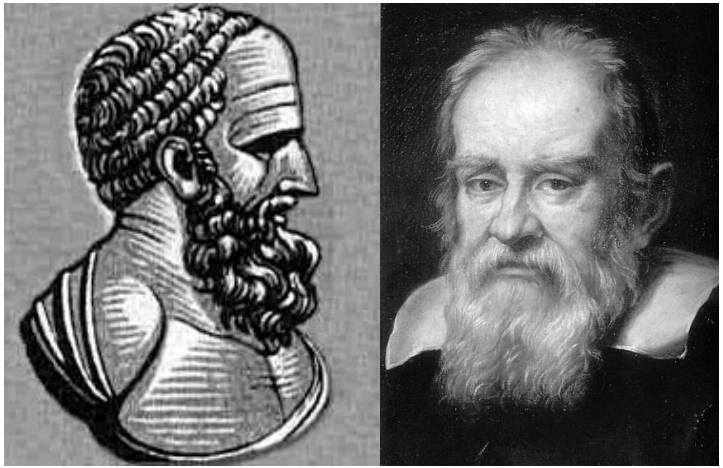
Astroestadística

Durante la mayoría de la historia occidental, los estadísticos eran los astrónomos

Antiguos griegos – Siglo 20

¿Cómo estimar la longitud de un año basándose en datos discrepantes?

- Mitad del Rango: Hiparco (Siglo 4 ACE)
- Observar una vez (medieval)
- Promedio: Brahe (Siglo 16), Galileo (Siglo 17), Simpson (Siglo 18)
- Mediana (Siglo 20)



Astroestadística

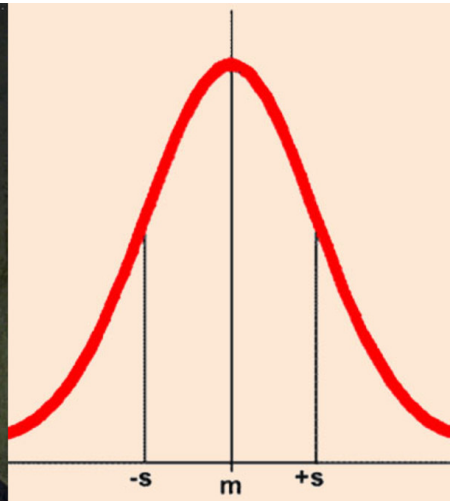
Durante la mayoría de la historia occidental, los estadísticos eran los astrónomos

Siglo 19

Estimación de Parámetros

Observaciones discrepantes de planetas/lunas/ cometas -> parámetros orbitales

Legendre, Laplace & Gauss desarrollan la regresión por mínimos cuadrados y la teoría de errores (c.1800-1820; 1820-1890)



El siglo perdido de la astroestadística

Finales del s. 19 al s. 20

Estadística: Nuevas aplicaciones

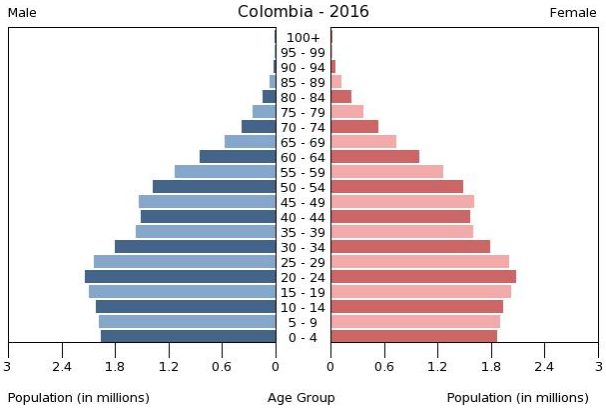
Ciencias Humanas

Demografía, Economía, Psicología,

Medicina, Política

Industria

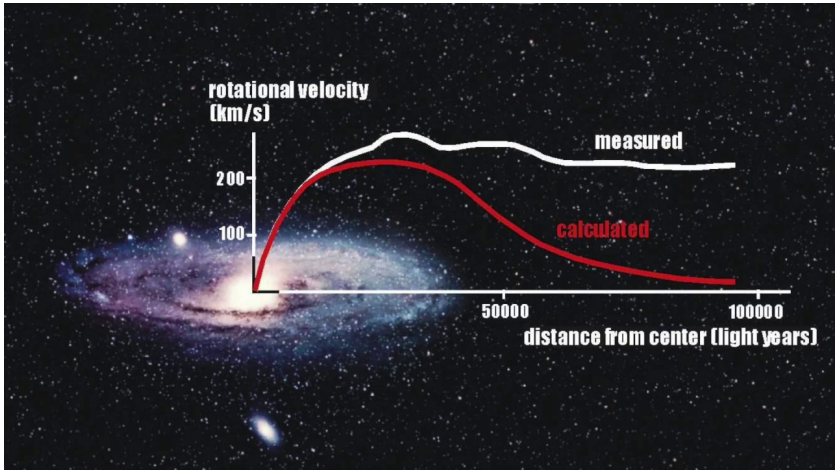
Agricultura, Minería, Manufactura



Astronomía: Alianza con la Física moderna:

Electromagnetismo, Termodinámica,
Mecánica Cuántica, Relatividad

Nace la Astrofísica



Computación

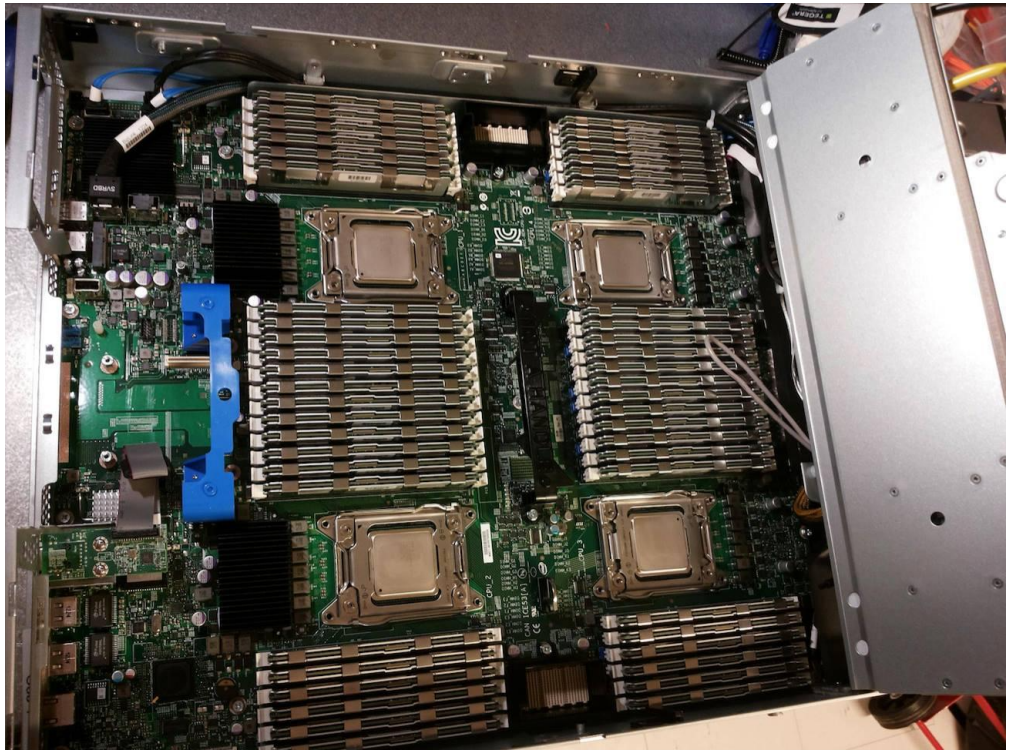
"ENIAC (1940s) contained 17,468 vacuum tubes, 7200 crystal diodes, 1500 relays, 70,000 resistors, 10,000 capacitors and approximately 5,000,000 hand-soldered joints. It weighed 27 t, was roughly 2.4m × 0.9m × 30m in size, occupied 167 m^2 and consumed 150 kW of electricity. This power requirement led to the rumor that whenever the computer was switched on, lights in Philadelphia dimmed."

Computación

Medio acumulador de 1 B

Computación

Medio acumulador de 1 B



1 TB de RAM

Un reencuentro lento

Métodos “clásicos”

Fourier transform for temporal analysis (Fourier 1807)

Least squares regression for model fits (Legendre 1805, Pearson 1901)

Kolmogorov-Smirnov goodness-of-fit test (Kolmogorov, 1933)

Principal components analysis for tables (Hotelling 1936)

Métodos “modernos”

Modeling (MLE, EM Algorithm, BIC, bootstrap, MCMC samplers)

Multivariate classification (GMM, LDA, SVM, CART, RFs)

Time series (autoregressive models, state space models)

Spatial point processes (Ripley’s K, Kriging)

Non-detections (survival analysis)

Image analysis (computer vision methods, False Detection Rate)

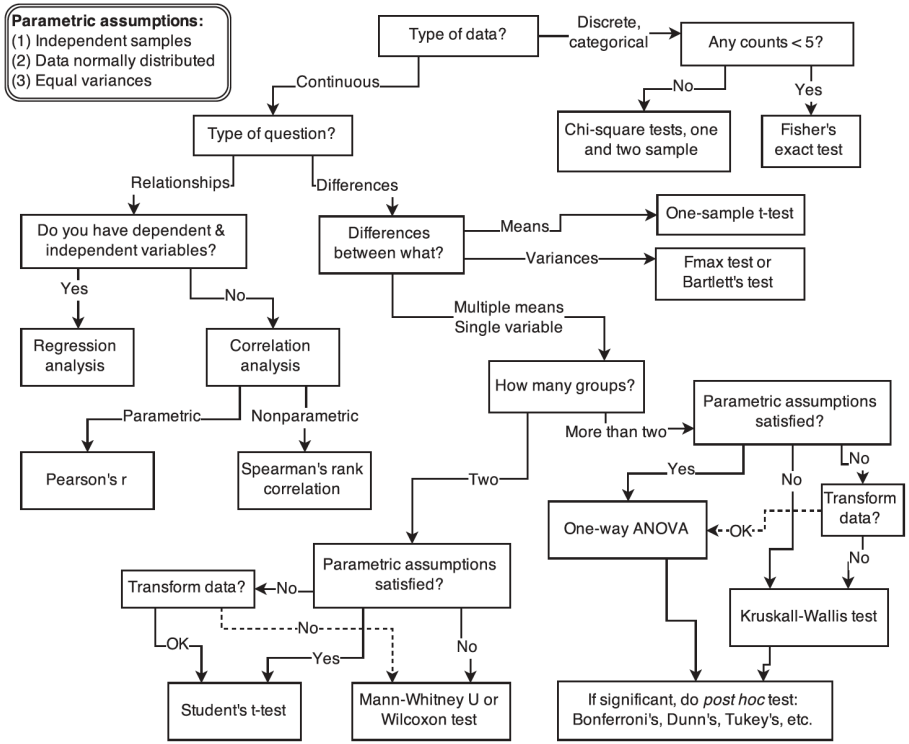
Statistical computing (R, Python)

- Are the available stars/galaxies/sources an unbiased sample of the vast underlying population?
- Sampling, bootstrap methods
- When should these objects be divided into 2/3/… classes?
- Multivariate classification
- What is the intrinsic relationship between two properties of a class (especially with confounding variables)?
- Multivariate regression
- Can we answer such questions in the presence of observations with measurement errors & flux limit?
- Bayesian inference

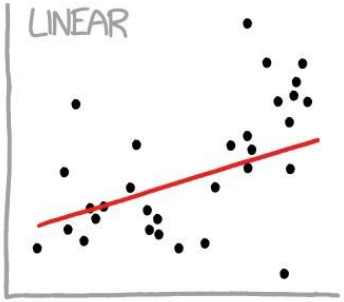
- When is a blip in a spectrum, image or data stream a real signal?
- Model comparison
- How do we model the vast range of variable objects (extrasolar planets, BH accretion, GRBs, …)?
- Time series analysis
- How do we model many-dimensional data points (galaxies in the Universe, photons in a detector)?
- Spatial point processes & image processing
- How do we model continuous structures (cosmic microwave background fluctuations, interstellar medium)?
- Density estimation, regression

Golems estadísticos

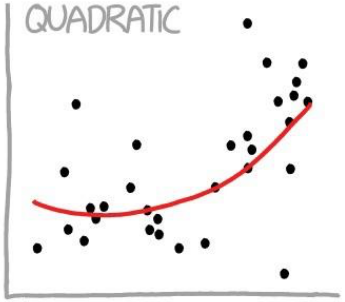
Golems estadísticos



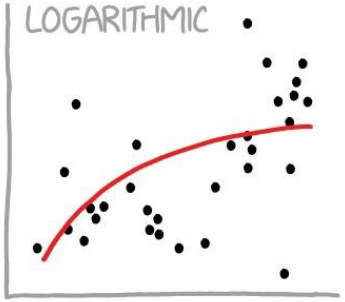
CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND



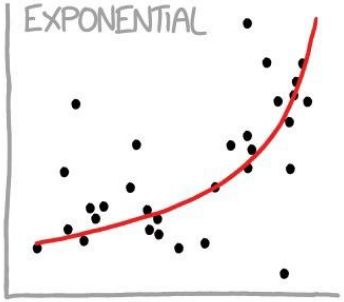
"HEY, I DID A REGRESSION."



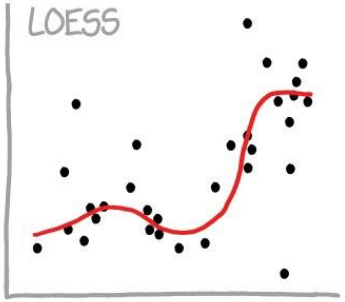
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."



"LOOK, IT'S TAPERING OFF!"

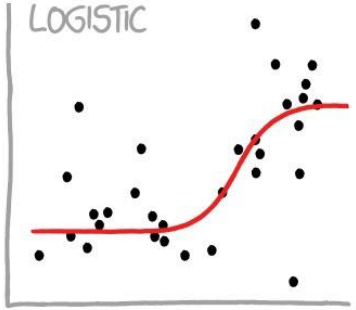


"LOOK, IT'S GROWING UNCONTROLLABLY!"



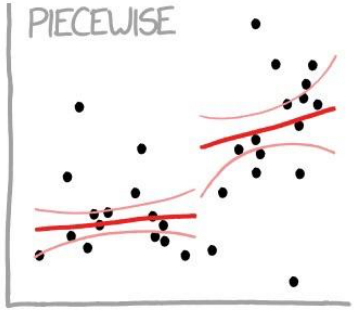
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."

"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."

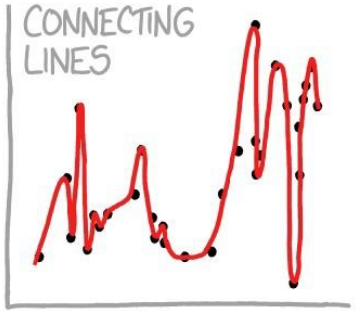


"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."

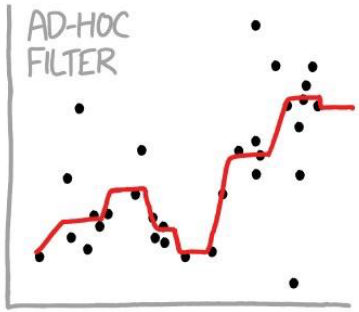
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."



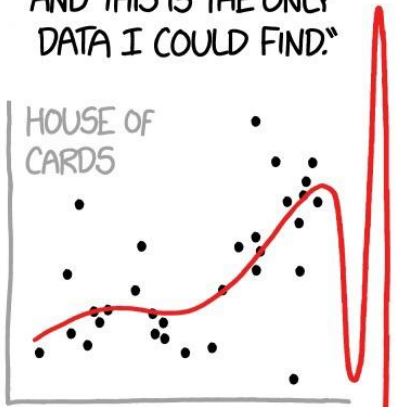
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."



"I CLICKED 'SMOOTH LINES' IN EXCEL."



"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"

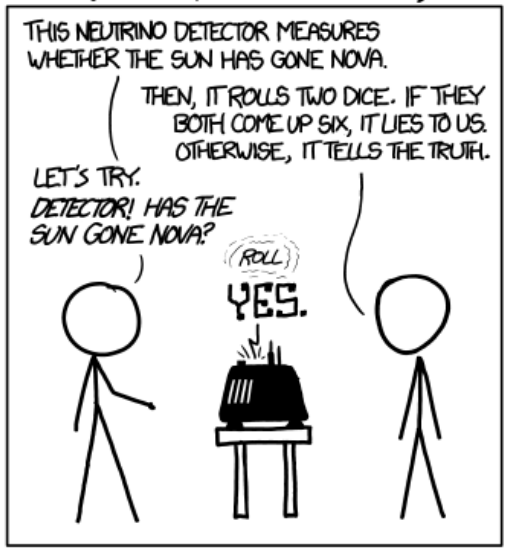


"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– *WAIT NO NO DON'T EXTEND IT AAAAAA!!*"

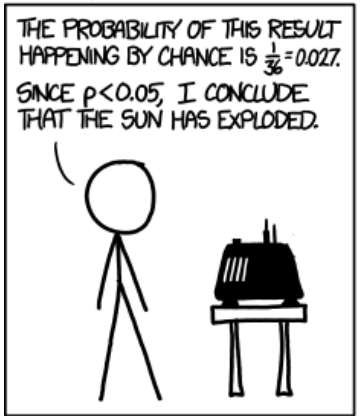
Bayesian Inference

- As evidence accumulates, the degree of belief in a hypothesis ought to change
- Bayesian inference takes prior knowledge into account
- The quality of Bayesian analysis depends on how best one can convert the prior information into mathematical prior probability
- Methods for parameter estimation, model assessment etc.

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)



FREQUENTIST STATISTICIAN:



BAYESIAN STATISTICIAN:

Bayesian Inference

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

Bayesian Inference

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

$P(A|B)$

Dadas las observaciones, ¿cuál es la probabilidad de que cierto modelo las explique?

Posterior

$P(B|A)$

Dado un modelo, ¿cuál es la probabilidad de que se hayan hecho las observaciones?

Verosimilitud

Vampirismo

- Test tiene 95% de efectividad
- El test da falsos positivos el 1% de las veces
- Proporción de vampiros es del 0.1%
- ¿Qué probabilidad hay de que el test de vampirismo diga la verdad si resulta positivo?

Vampirismo

- Test tiene 95% de efectividad

  Pr(positivo|vampiro) = 0.95
- El test da falsos positivos el 1% de las veces

  Pr(positivo|mortal) = 0.01
- Proporción de vampiros es del 0.1%

  Pr(vampiro) = 0.001

- ¿Qué probabilidad hay de que el test de vampirismo diga la verdad si resulta positivo?

Vampirismo

- Aplicamos la regla de Bayes

$$\Pr(\text{vampire}|\text{positive}) = \frac{\Pr(\text{positive}|\text{vampire})\,\Pr(\text{vampire})}{\Pr(\text{positive})}$$

- Con la probabilidad total de dar positivo

$$\begin{aligned}\Pr(\text{positive}) = {} & \Pr(\text{positive}|\text{vampire})\,\Pr(\text{vampire}) \\ & + \Pr(\text{positive}|\text{mortal})\,(1 - \Pr(\text{vampire}))\end{aligned}$$

Vampirismo

- ¿Qué probabilidad hay de que el test de vampirismo diga la verdad si resulta positivo?

  - 99%
  - 80%
  - 50%
  - 10%

Vampirismo

- El resultado es:

  Pr(vampiro|positivo) = 0.087 -> 8%

Vampirismo

- El resultado es:

  Pr(vampiro|positivo) = 0.087 -> 8%

- Contraintuitivo. Una mejor forma de plantearlo es por cuentas:
  - De 100 000 personas, 100 son vampiros
  - De los 100 vampiros, 95 darán positivo en el test
  - De los 99 900 mortales, 999 darán positivo en el test
  - Tan sólo 95 vampiros entre 999+95 tests positivos

Canicas en una bolsa

- 4 canicas en una bolsa oscura
- Las canicas pueden ser blancas o azules

- Tomamos una canica, registramos su color y la devolvemos a la bolsa
- Repetimos 3 veces

- Observación 

Canicas en una bolsa

- Posibilidades

(1) [○○○○], (2) [●○○○], (3) [●●○○], (4) [●●●○], (5) [●●●●]

- Observación 

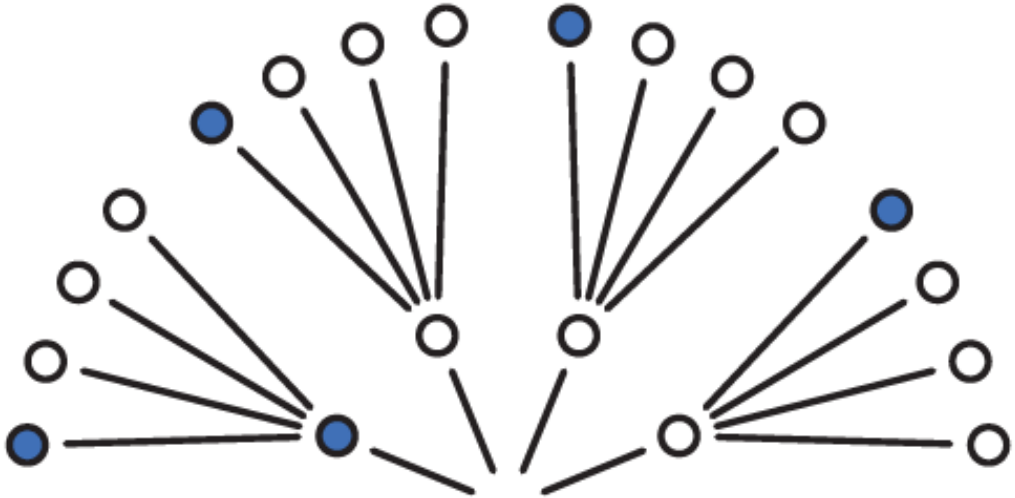
- ¿Conclusiones sobre el sistema real?

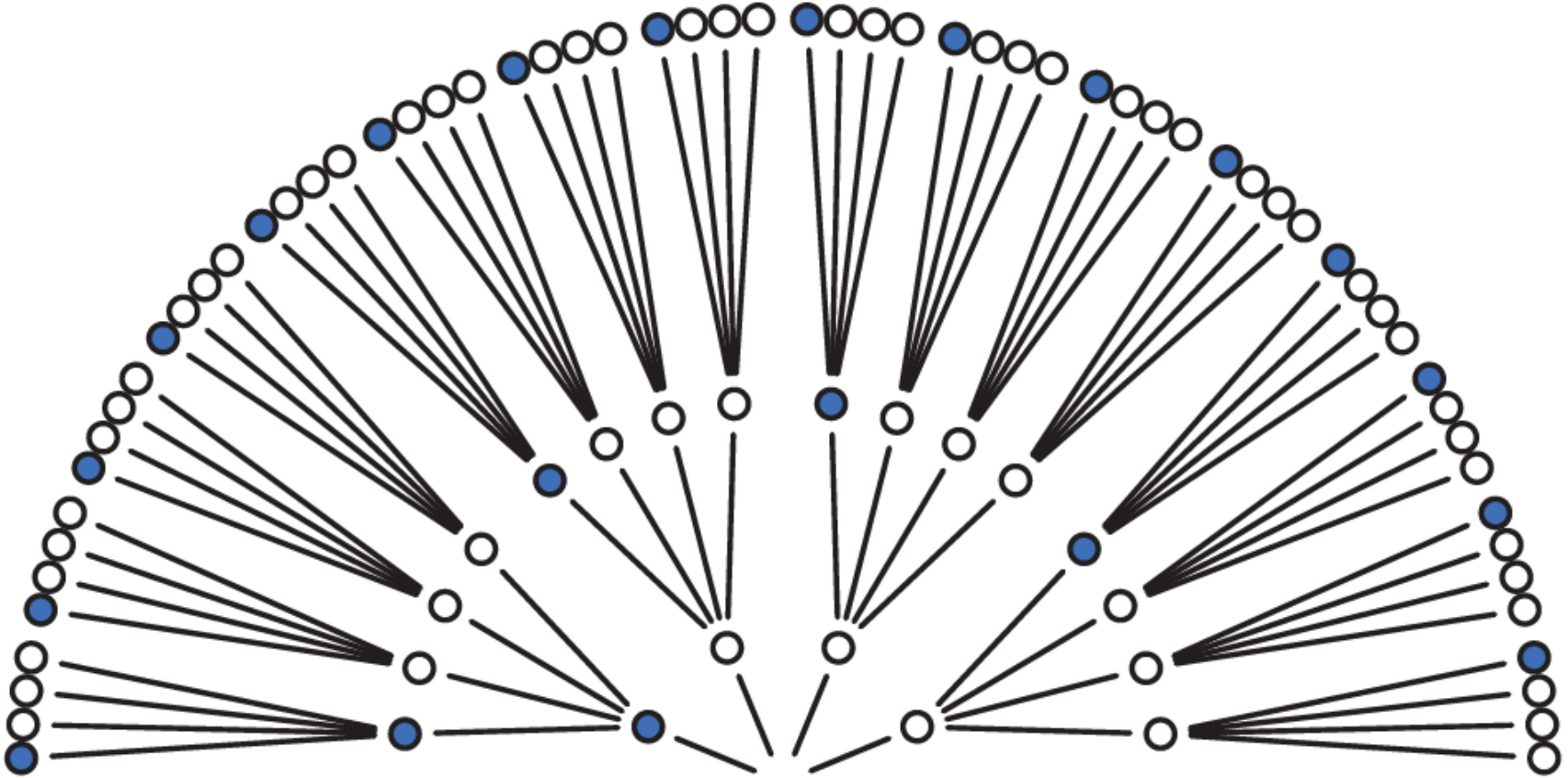
El Jardín de Senderos que se Bifurcan

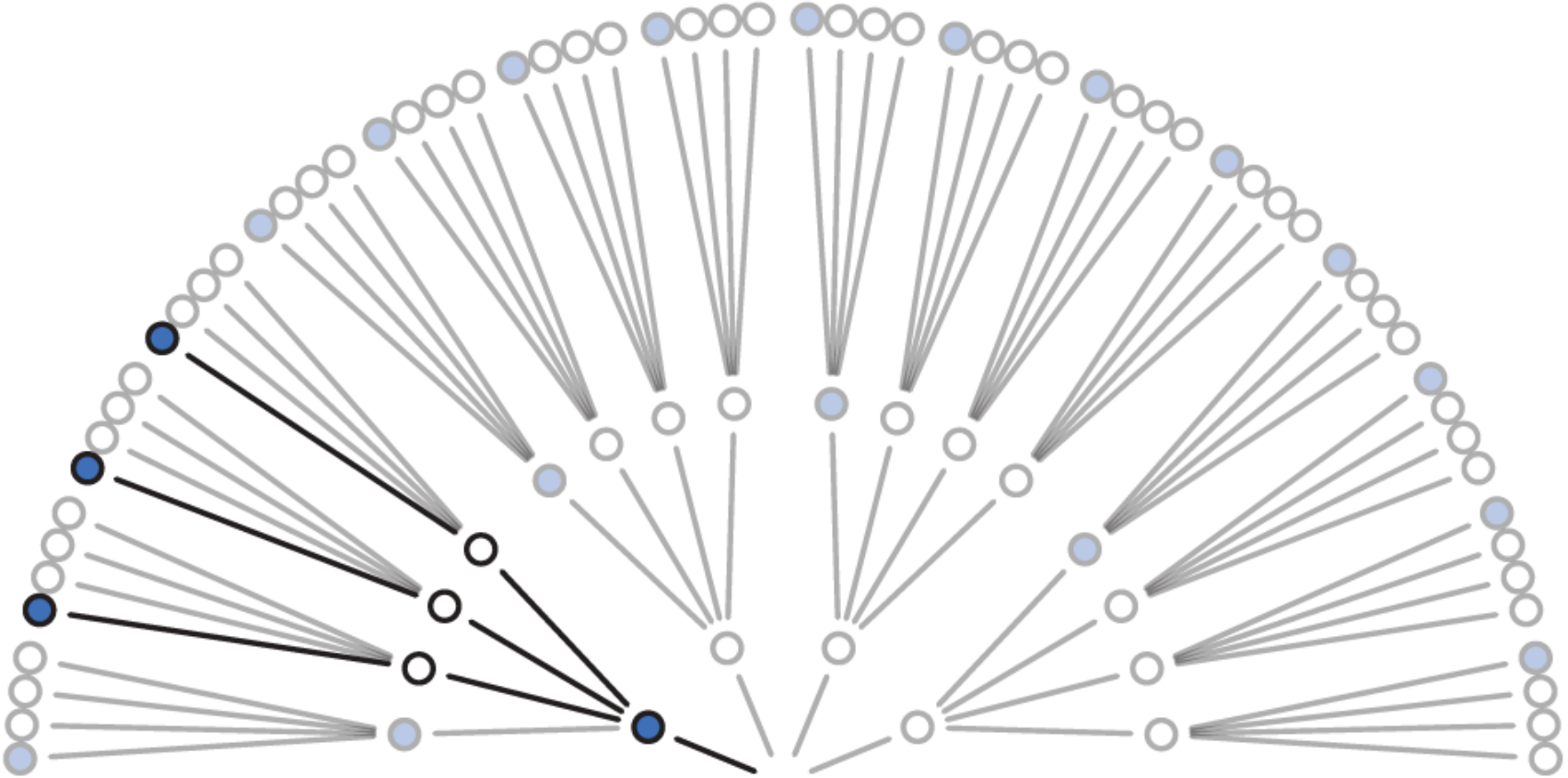
- Conjetura 
- Combinaciones posibles (1era toma)

El Jardín de Senderos que se Bifurcan

- Combinaciones posibles (2da toma)

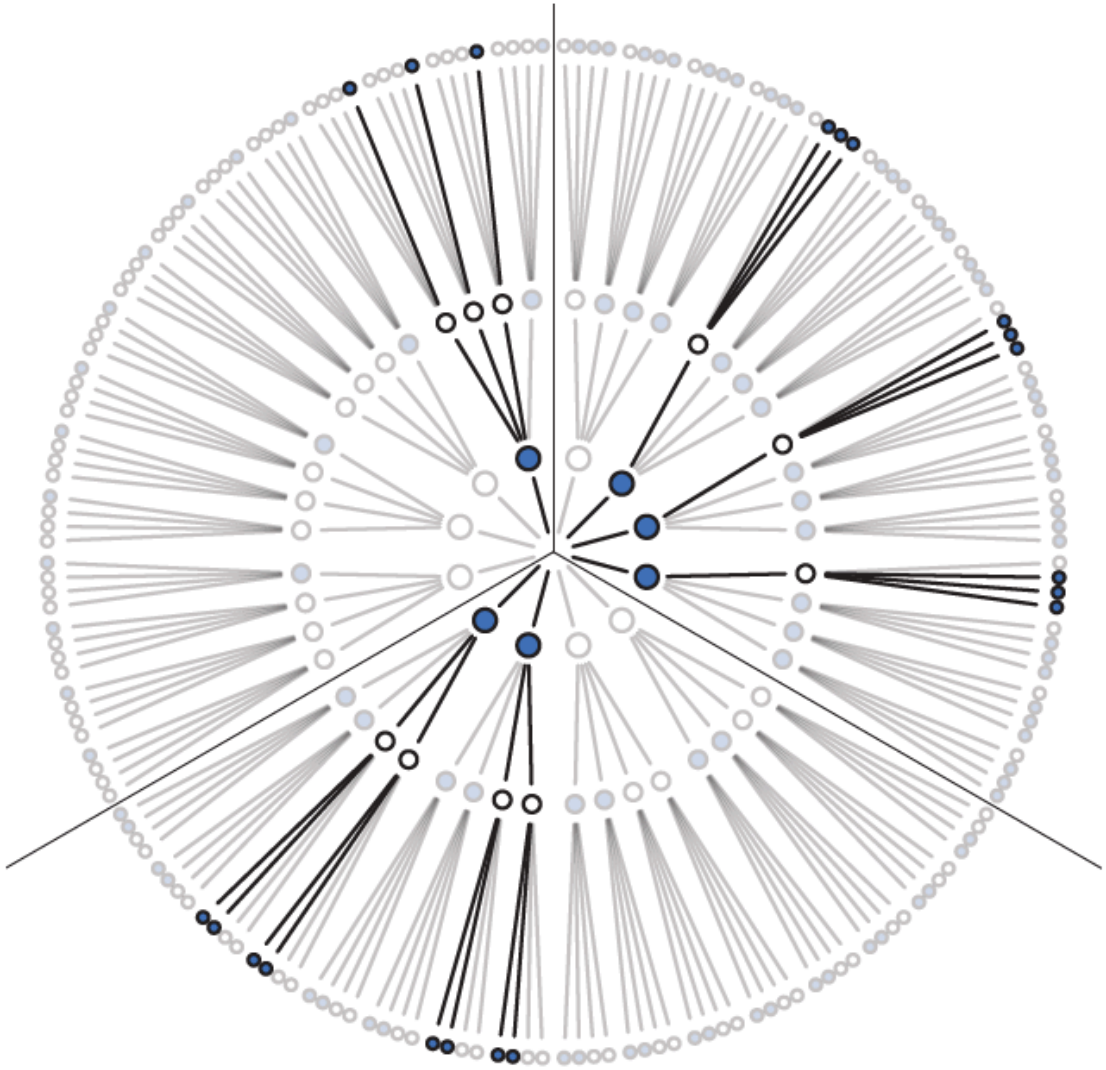






Conteos según cada conjetura

Conjetura	Maneras de producir ●○●
[○○○○]	$0 \times 4 \times 0 = 0$
[●○○○]	$1 \times 3 \times 1 = 3$
[●●○○]	$2 \times 2 \times 2 = 8$
[●●●○]	$3 \times 1 \times 3 = 9$
[●●●●]	$4 \times 0 \times 4 = 0$



Nueva observación:

- Datos actualizados

- Opciones:
  - Contar de nuevo
  - Actualizar conteos

Conteo actualizado

Conjetura	Maneras de producir ●	Conteo anterior	Conteo nuevo
[○○○○]	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	$0 \times 4 = 0$

Actualizando conteos

- Si para cada conjetura:
  - (1) Hay M_{prior} maneras de producir observación D_{prior}
  - (2) Una observación nueva D_{nuevo} tiene M_{nuevo} maneras de ser producida
- Entonces:
  - (3) Las maneras de producir D_{nuevo} habiendo observado D_{prior} anteriormente es $M_{\text{prior}} \times M_{\text{nuevo}}$

Información (más) prior

- Puede venir antes de las observaciones
- Ejemplo: algunas combinaciones pueden ser más probables de fábrica

Conjetura	Conteo de fábrica
[○○○○]	0
[●○○○]	3
[●●○○]	2
[●●●○]	1
[●●●●]	0

Actualizando conteos con la información (prior) de fábrica

Conjetura	Conteo prior	Conteo de fábrica	Conteo nuevo
[○○○○]	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	$0 \times 0 = 0$

De posibilidades a probabilidades

- Los valores de los conteos son importantes sólo cuando se comparan entre sí

plausibilidad de [●○○○] después de observar ●○●

$\propto$

maneras de que [●○○○] produzca ●○●

$\times$

plausibilidad prior de[●○○○]

De posibilidades a probabilidades

- Cierta proporción de canicas azules: p
- Cierta observación: D_{nuevo}

$$\text{plausibilidad de } p \text{ después de observar } D_{\text{nuevo}}$$
$$\propto$$
$$\text{maneras de que } p \text{ produzca } D_{\text{nuevo}}$$
$$\times$$
$$\text{plausibilidad prior de } p$$

De posibilidades a probabilidades

- Normalizando respecto a todas las posibilidades, llegamos a la probabilidad.

$$\text{plausibilidad de } p \text{ después de observar } D_{\text{nuevo}} = \frac{\text{maneras de que } p \text{ produzca } D_{\text{nuevo}} \times \text{plausibilidad prior de } p}{\text{suma de productos}}$$

De posibilidades a probabilidades

- Normalizando respecto a todas las posibilidades, llegamos a la probabilidad.

Composición posible	p	Maneras de producir los datos	Plausibilidad
[○○○○]	0	0	0
[●○○○]	0.25	3	0.15
[●●○○]	0.5	8	0.40
[●●●○]	0.75	9	0.45
[●●●●]	1	0	0

Formalizando

- Parámetro: p (posible explicación de los datos)
- Verosimilitud: Número relativo de maneras en las que algún valor de p puede explicar los datos (modelo)
- Plausibilidad prior: probabilidad prior
- Plausibilidad actualizada: probabilidad posterior

Using Bayes' rule != Bayesian Inference

Construyendo un Modelo Bayesiano

- (1) Historia para los datos: Explicar la motivación detrás de modelo narrando cómo éste genera las observaciones
- (2) Actualizar: Educar el modelo alimentándolo con los datos
- (3) Evaluar: Todos los modelos estadísticos requieren supervisión, lo que puede llevar a una revisión del modelo

The Bayesian Mindset

- Posterior distributions are probability distributions
- Samples from the posterior = model parameter values*
- The Bayesian formalism treats parameter distributions as relative plausibilities, not as any physical random process.
- Randomness is a property of information, never of the real world.
- Inside the computer, parameters are empirical.
- The posterior defines the expected frequency that different parameter values will appear, once we start plucking parameters out of it.

Andr(é|ew) MacDonald
@polesasunder

Follow

build an understanding of statistics and you too can cripple you publication record while also alienating all your colleagues

7:43 PM - 3 Apr 2019

193 Retweets **1,419** Likes

26 193 1.4K