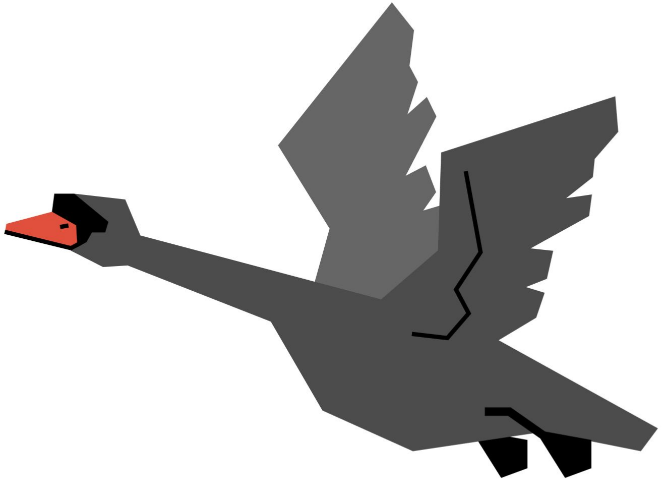
Taming Black Swans

Long-tailed distributions in the natural and engineered world



Allen Downey

slides at

tinyurl.com/longtail23

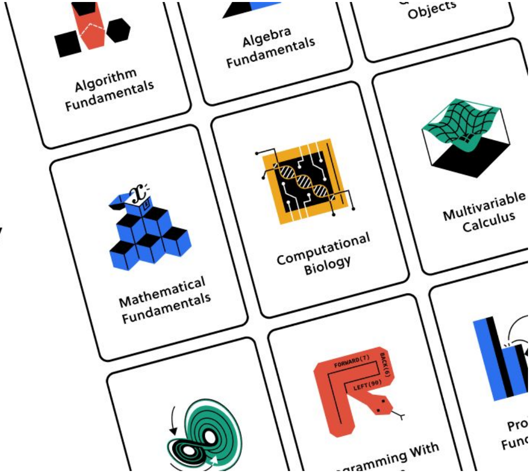
Allen Downey

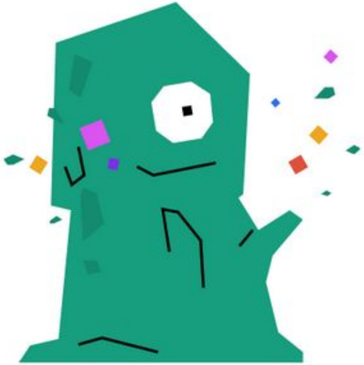
Staff Producer

BRILLIANT

Learn interactively

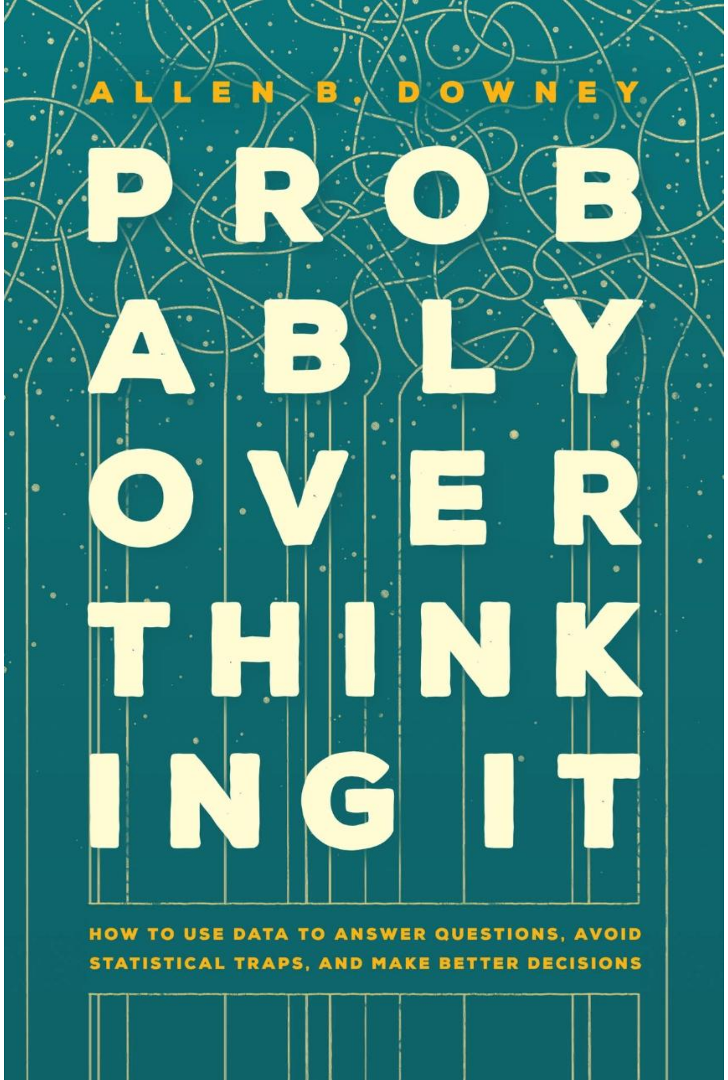
BRILLIANT





tinyurl.com/longtail23

Professor Emeritus at Olin College



Contents

tinyurl.com/longtail23

Long-tailed distributions are common in natural and engineered systems.

Long-tailed distributions

- Violate intuition,
- Defy prediction, and
- Leave us unprepared for disaster.

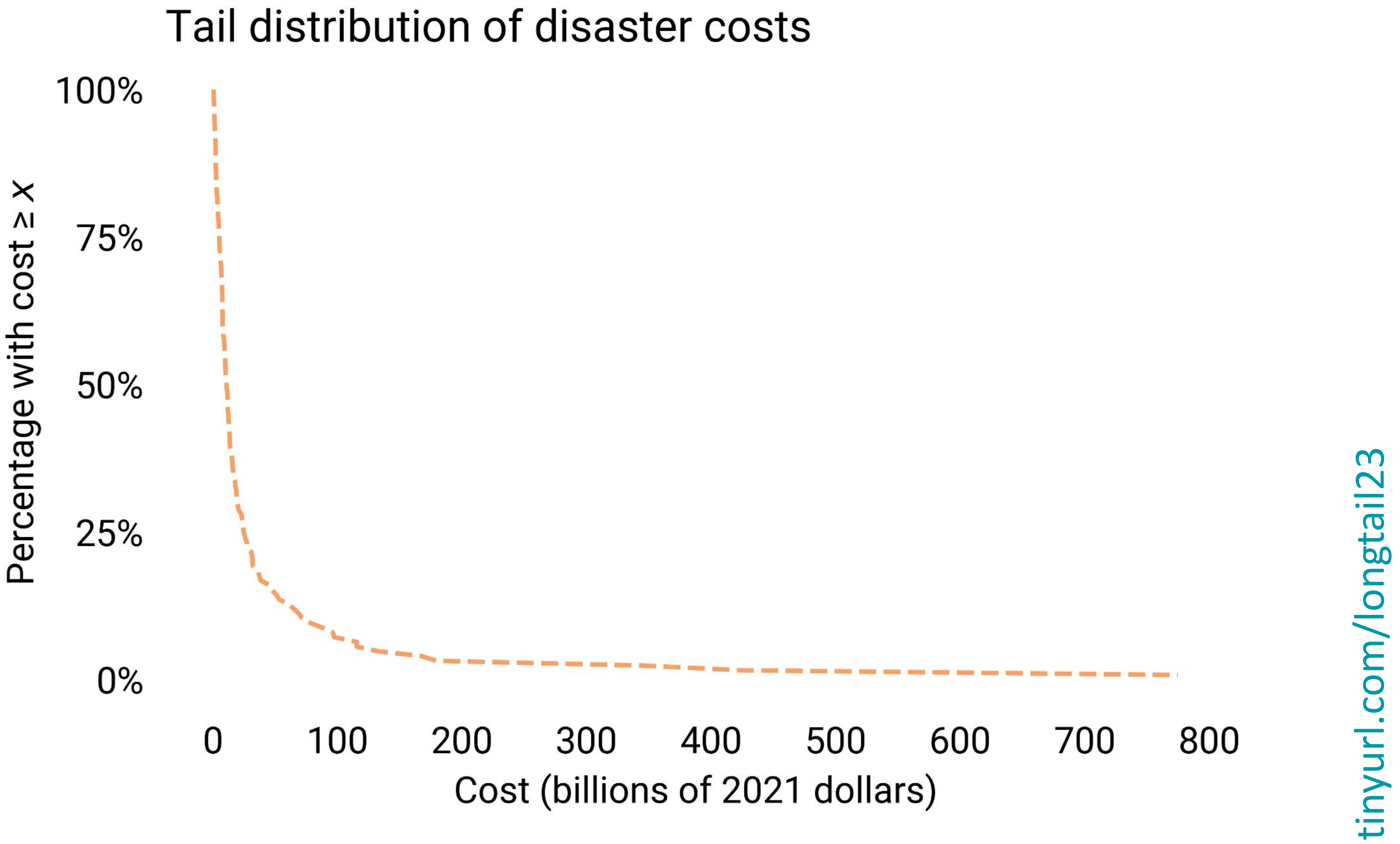
List of disasters by cost

From Wikipedia, the free encyclopedia

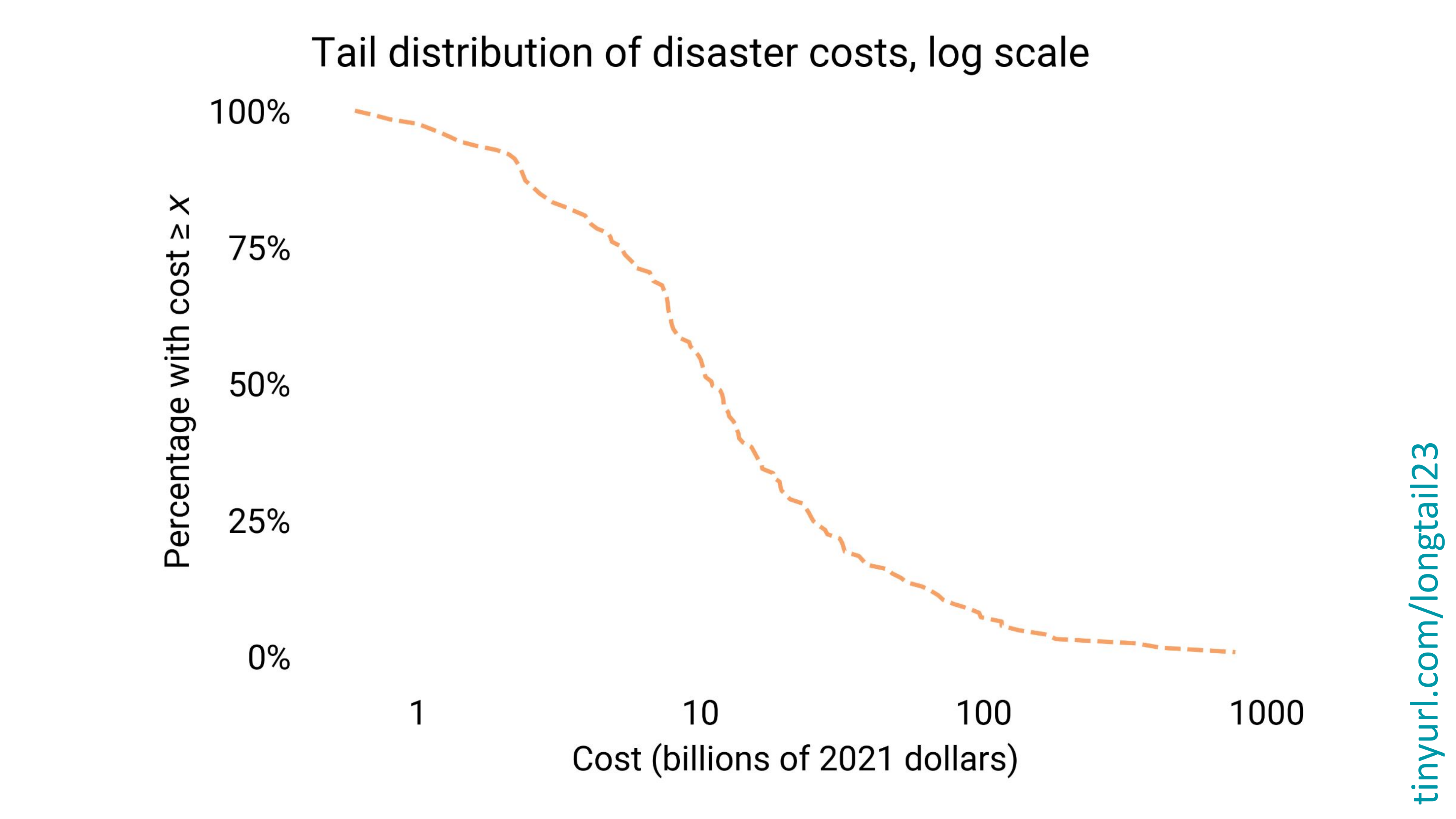
This is a dynamic list and may never be able to satisfy particular standards for completeness. You can help by adding missing items with reliable sources.

Disasters can have high costs associated with responding to and recovering from them. This page lists the estimated economic costs of relatively recent disasters.

The costs of disasters vary considerably depending on a range of factors, such as the geographical location where they occur. When a large disaster occurs in a wealthy country, the financial damage may be large, but when a comparable disaster occurs in a poorer country, the actual financial damage may appear to be relatively small. This is in part due to the difficulty of measuring the financial damage in areas that lack insurance. For example, the 2004 Indian Ocean earthquake and tsunami, with a death toll of around 230,000 people, cost a 'mere' $15 billion,[1] whereas in the Deepwater Horizon oil spill, in which 11 people died, the damage was six times higher.

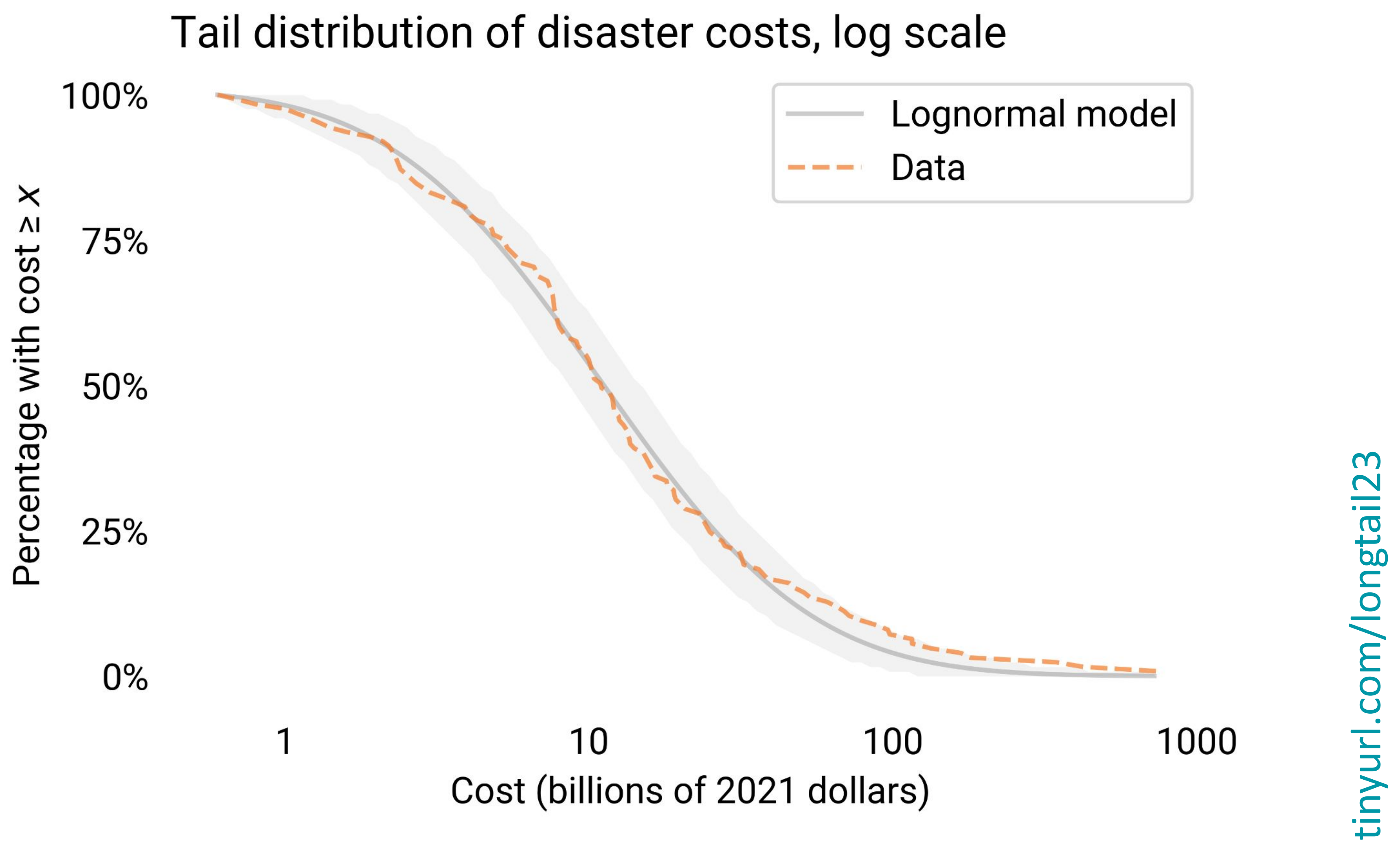

Tail distribution of disaster costs
100%
75%
50%
25%
0%
Percentage with cost ≥ x
0
100
200
300
400
500
600
700
800
Cost (billions of 2021 dollars)
tinyurl.com/longtail23

On a linear scale,
most of the distribution is
mashed against the axes.


Tail distribution of disaster costs, log scale
Percentage with cost ≥ x
100%
75%
50%
25%
0%
1
10
100
1000
Cost (billions of 2021 dollars)
tinyurl.com/longtail23

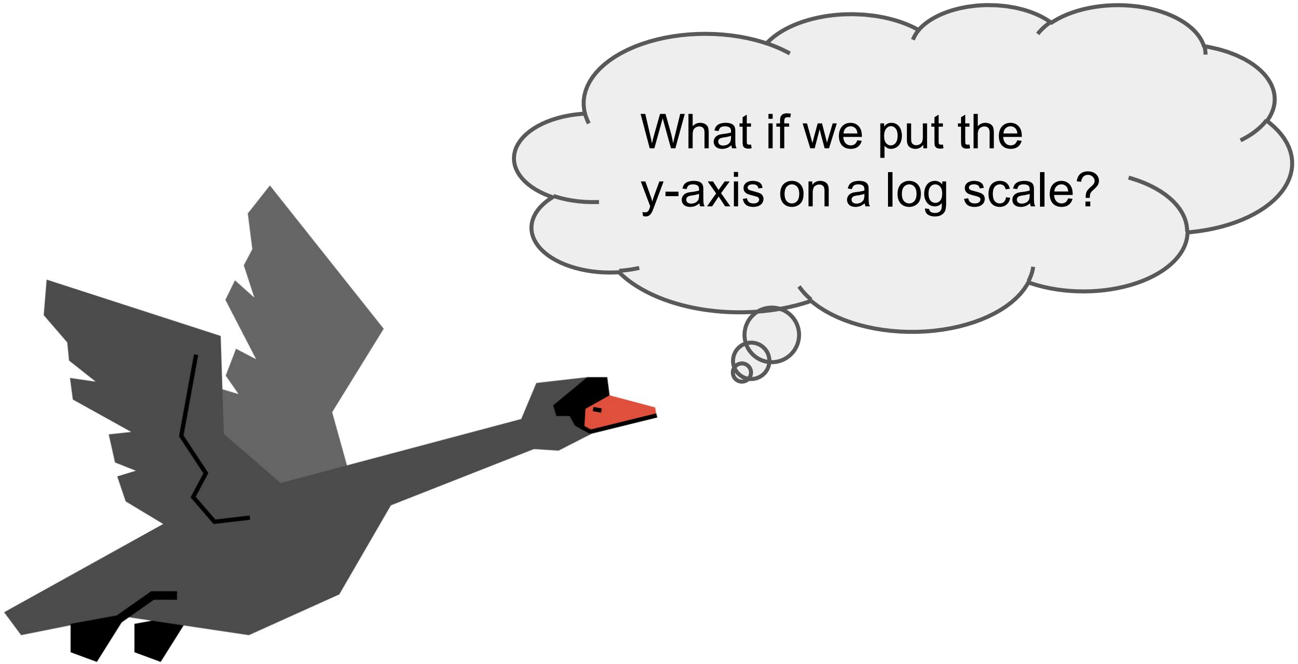
On a log scale,
we can see the middle of the
distribution more clearly.

And that sigmoid shape
suggests a lognormal distribution.


Tail distribution of disaster costs, log scale
Lognormal model
Data
Percentage with cost ≥ x
100%
75%
50%
25%
0%
1
10
100
1000
Cost (billions of 2021 dollars)
tinyurl.com/longtail23

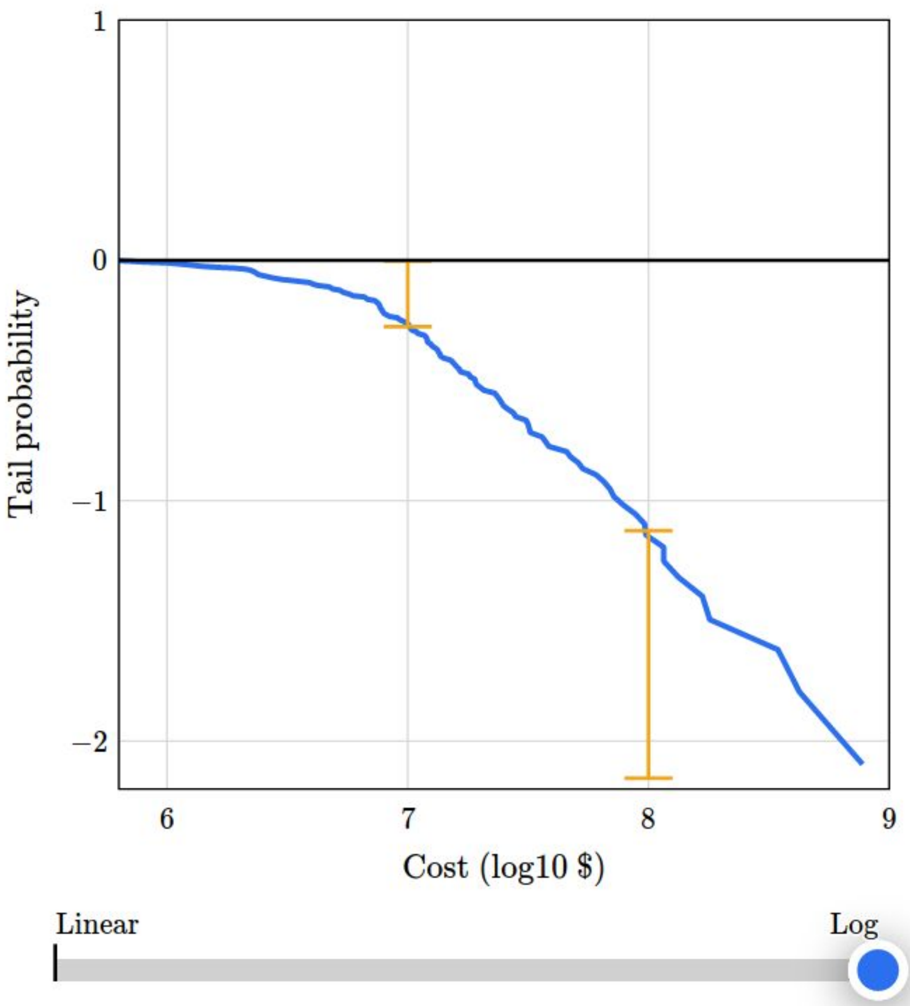
The lognormal model looks good:

- More disasters near $100 billion than expected,
- But within the variation we expect by chance.

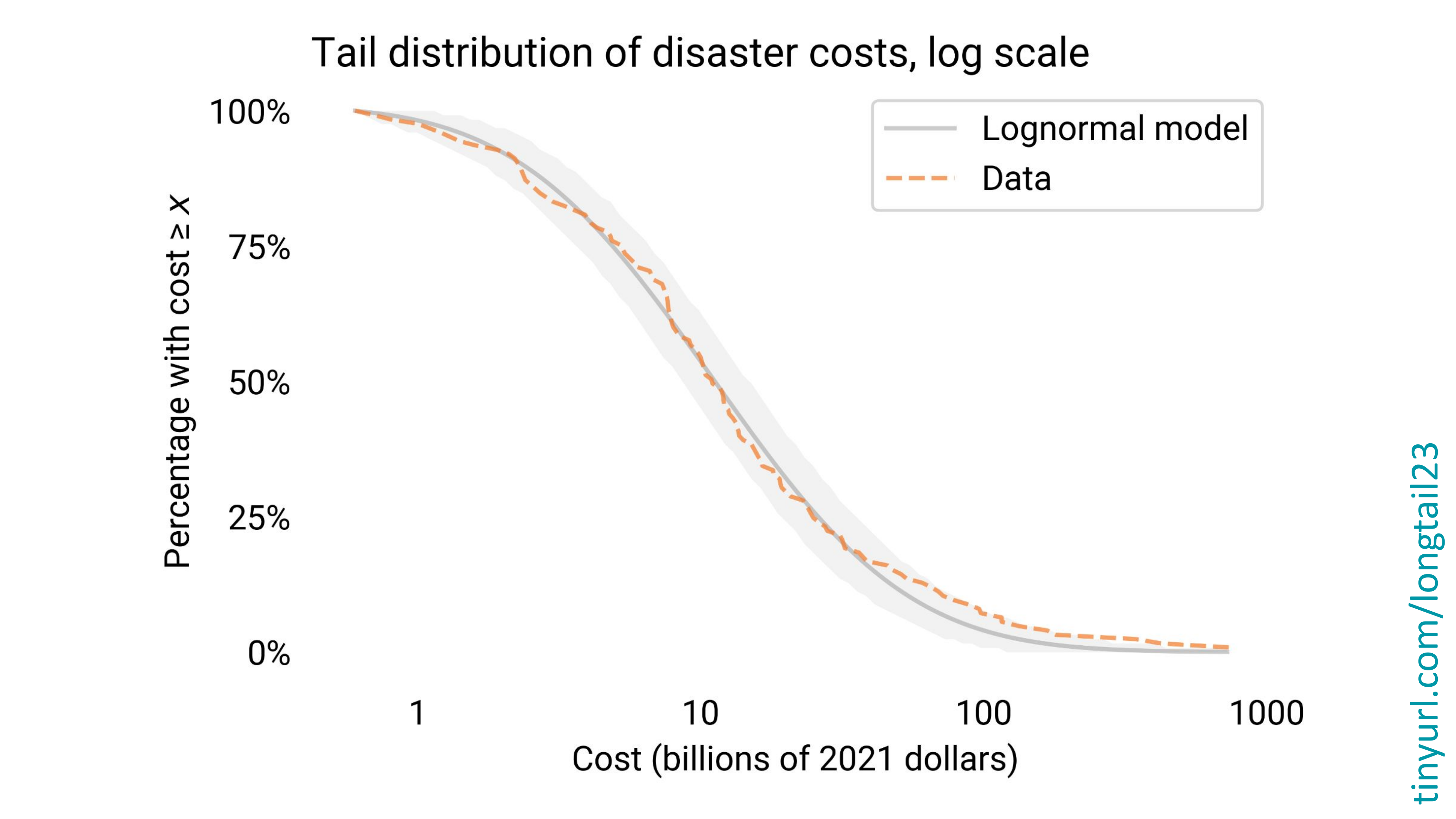


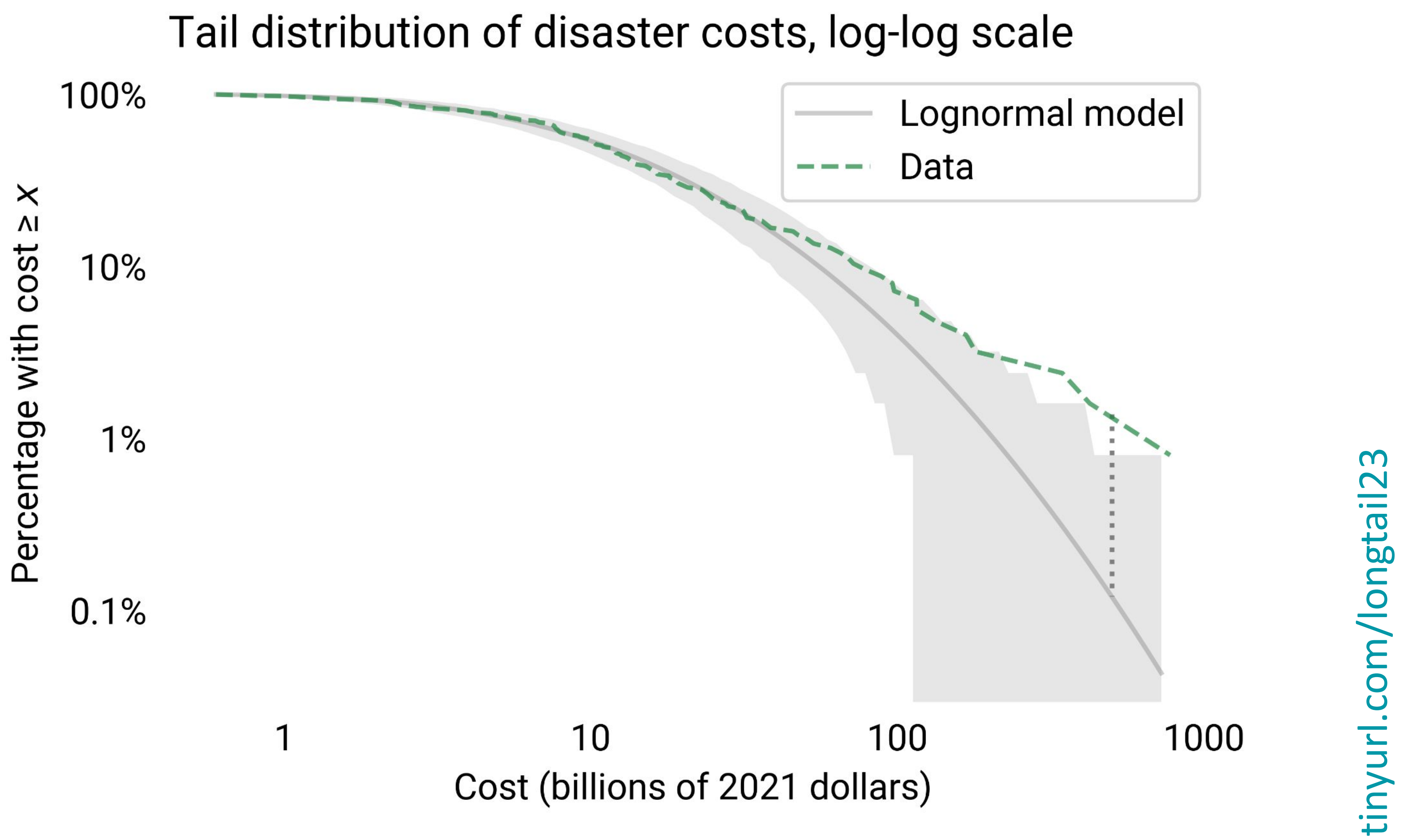
tinyurl.com/longtail23

The log-y scale is like a microscope for inspecting tail behavior.



tinyurl.com/longtail23


Tail distribution of disaster costs, log scale
Lognormal model
Data
Percentage with cost ≥ x
100%
75%
50%
25%
0%
1
10
100
1000
Cost (billions of 2021 dollars)
tinyurl.com/longtail23



Tail distribution of disaster costs, log-log scale
Lognormal model
Data
Percentage with cost ≥ x
100%
10%
1%
0.1%
1
10
100
1000
Cost (billions of 2021 dollars)
tinyurl.com/longtail23

Fraction of disasters that exceed $500 billion

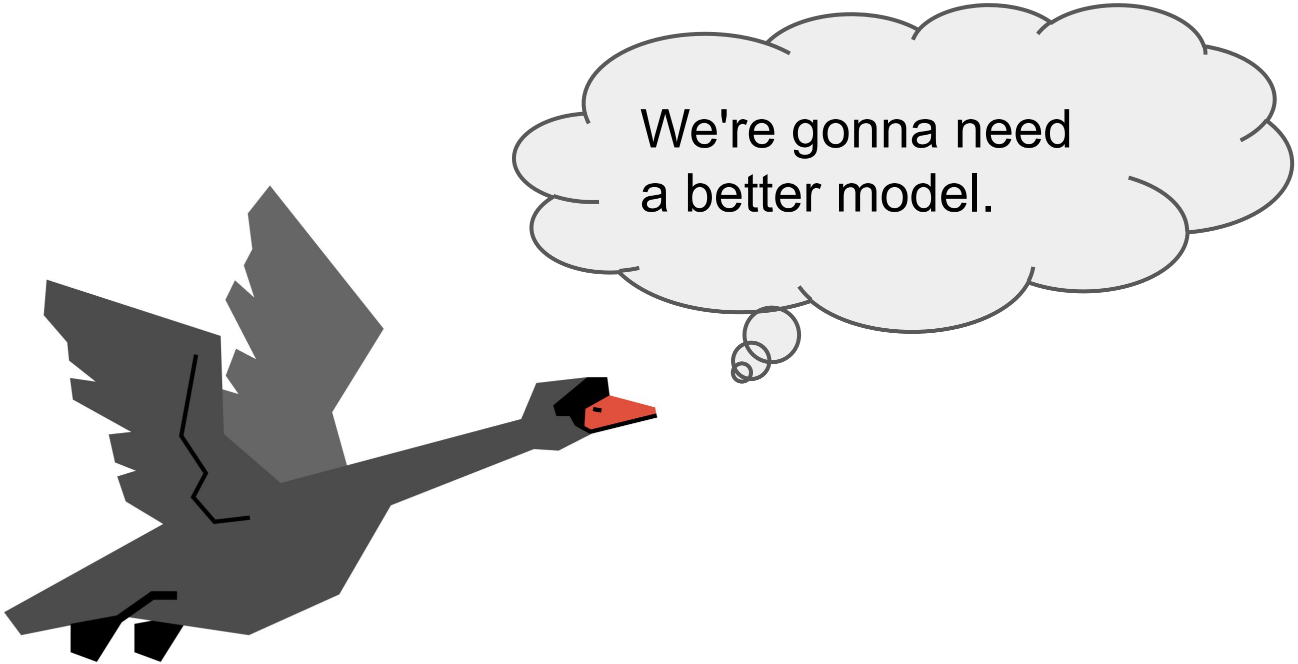
Model: 1 per 1000

Data: 16 per 1000

tinyurl.com/longtail23

The model underestimates the probability of large, rare disasters.

tinyurl.com/longtail23



tinyurl.com/longtail23

Several longer-tailed models to choose from.

	name	distribution $p(x) = Cf(x)$ $f(x)$	 C
continuous	power law	$x^{-\alpha}$	$(\alpha - 1)x_{\min}^{\alpha-1}$
	power law with cutoff	$x^{-\alpha}\mathrm{e}^{-\lambda x}$	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{\min})}$
	exponential	$\mathrm{e}^{-\lambda x}$	$\lambda \mathrm{e}^{\lambda x_{\min}}$
	stretched exponential	$x^{\beta-1}\mathrm{e}^{-\lambda x^{\beta}}$	$\beta\lambda \mathrm{e}^{\lambda x_{\min}^{\beta}}$
	log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$	$\sqrt{\frac{2}{\pi\sigma^2}}\left[\mathrm{erfc}\left(\frac{\ln x_{\min}-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}$
discrete	power law	$x^{-\alpha}$	$1/\zeta(\alpha, x_{\min})$
	Yule distribution	$\frac{\Gamma(x)}{\Gamma(x+\alpha)}$	$(\alpha - 1)\frac{\Gamma(x_{\min}+\alpha-1)}{\Gamma(x_{\min})}$
	exponential	$\mathrm{e}^{-\lambda x}$	$(1 - \mathrm{e}^{-\lambda})\,\mathrm{e}^{\lambda x_{\min}}$
	Poisson	$\mu^x/x!$	$\left[\mathrm{e}^{\mu} - \sum_{k=0}^{x_{\min}-1}\frac{\mu^k}{k!}\right]^{-1}$

TABLE 2.1

Definition of the power-law distribution and several other common statistical distributions. For each distribution we give the basic functional form $f(x)$ and the appropriate normalization constant C such that $\int_{x_{\min}}^{\infty} Cf(x)\,\mathrm{d}x = 1$ for the continuous case or $\sum_{x=x_{\min}}^{\infty} Cf(x) = 1$ for the discrete case.

https://arxiv.org/abs/0706.1062

tinyurl.com/longtail23

Student's *t* distribution

Similar to Gaussian, but longer tail

Three parameters:

- location, μ
- scale, τ
- degrees of freedom, ν

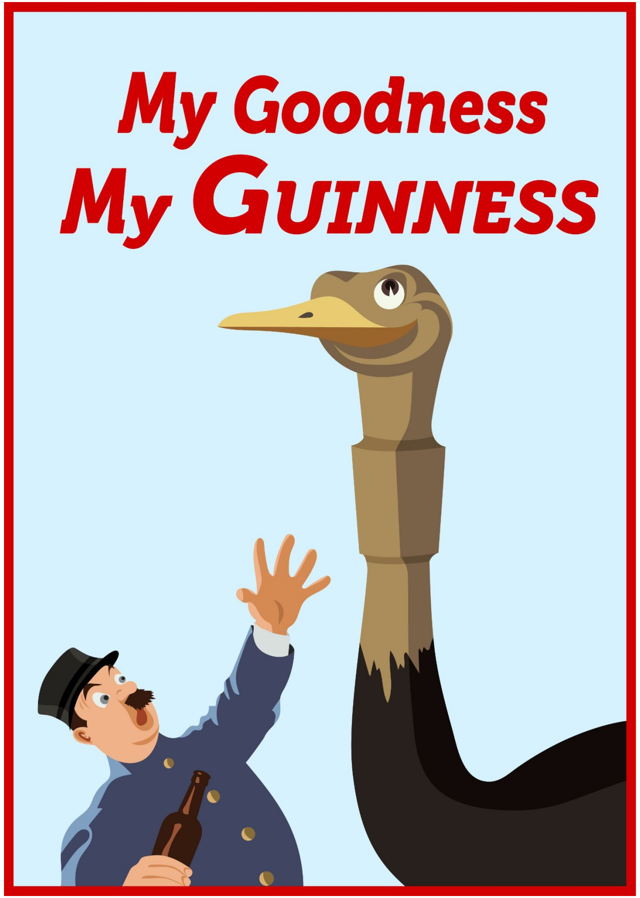
tinyurl.com/longtail23

Student's *t* distribution

Similar to Gaussian, but longer tail

Three parameters:

- location, μ
- scale, τ
- degrees of freedom, ν

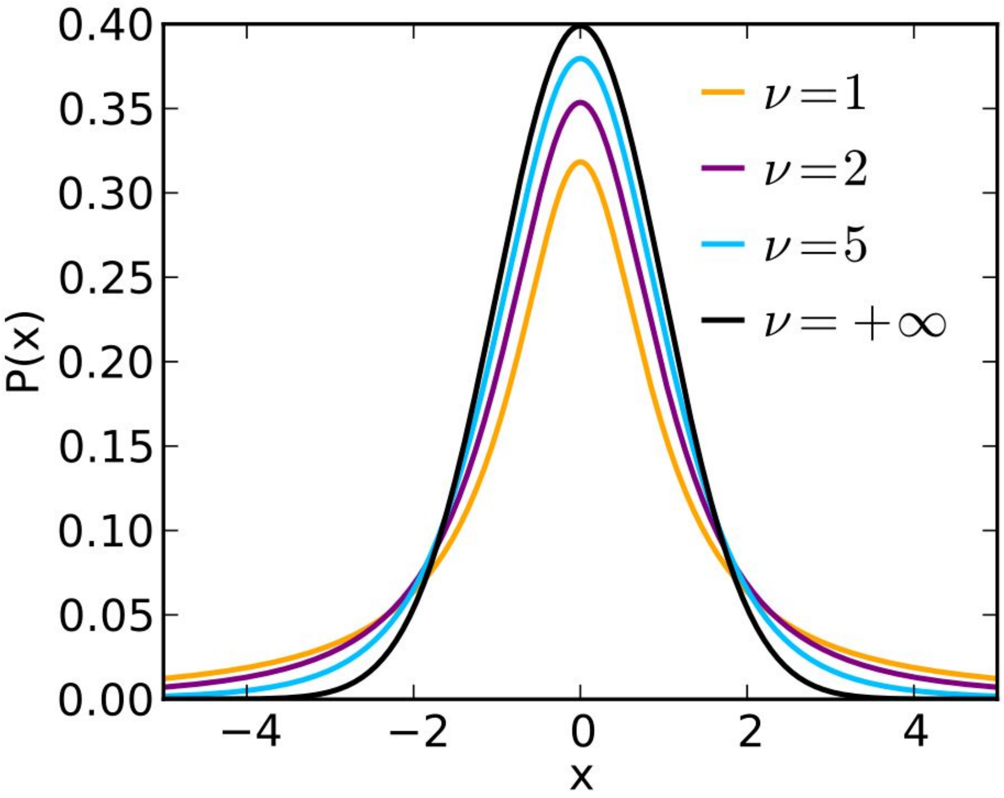


tinyurl.com/longtail23

v = 1 SUPER long tail

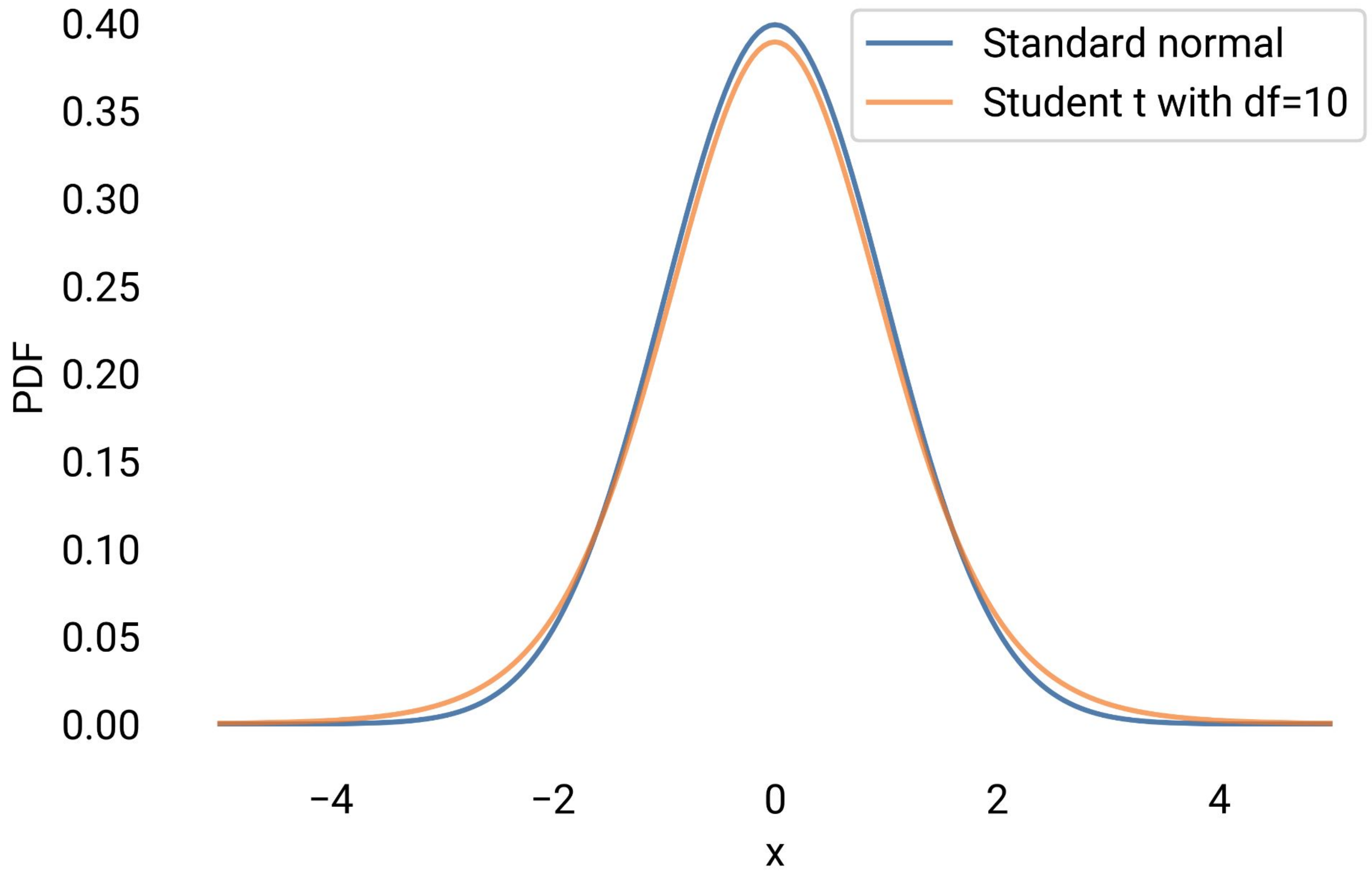
v = 3-10 empirical

v = ∞ same as Gaussian

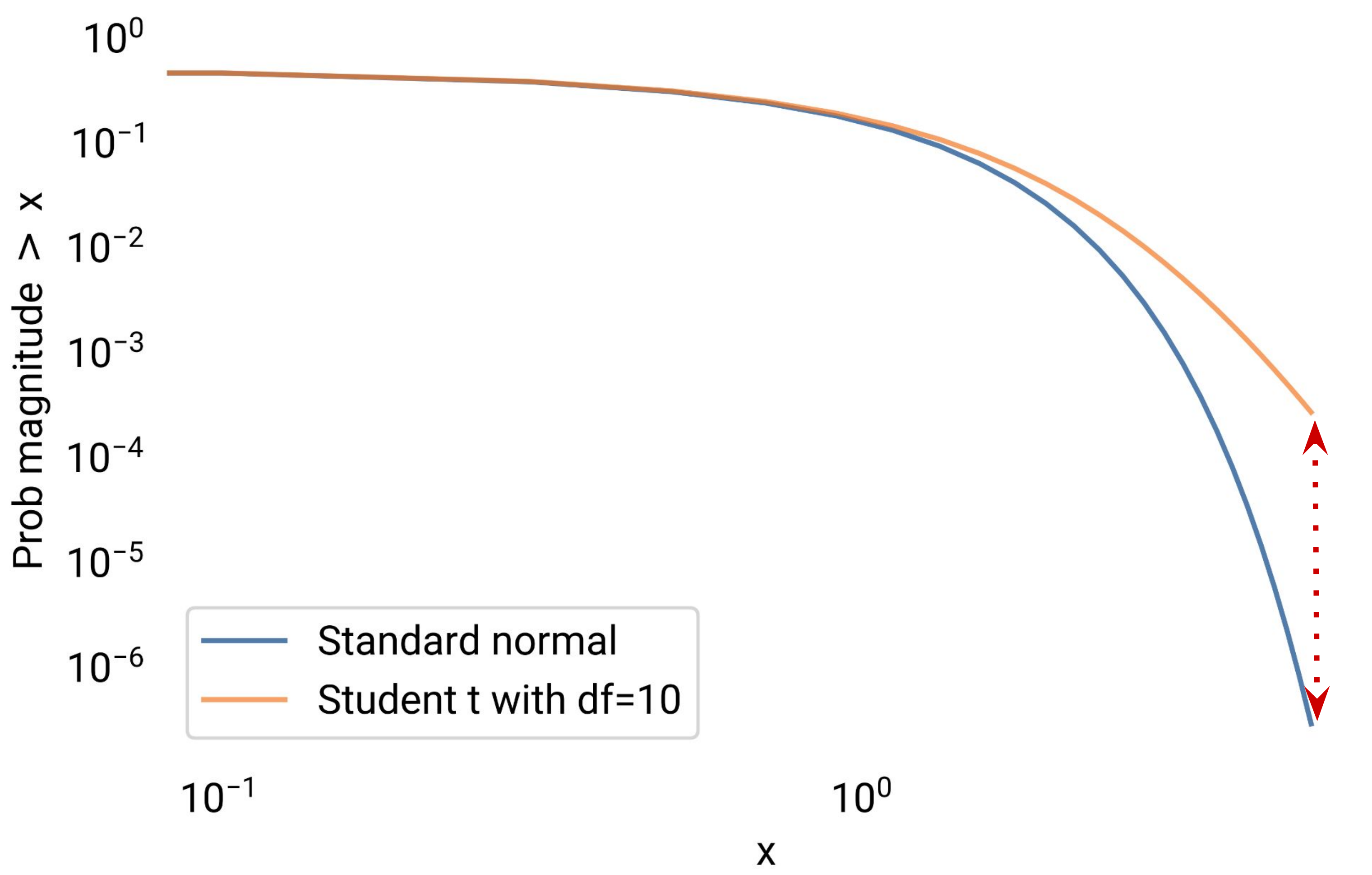


https://upload.wikimedia.org/wikipedia/commons/4/41/Student_t_pdf.svg

tinyurl.com/longtail23



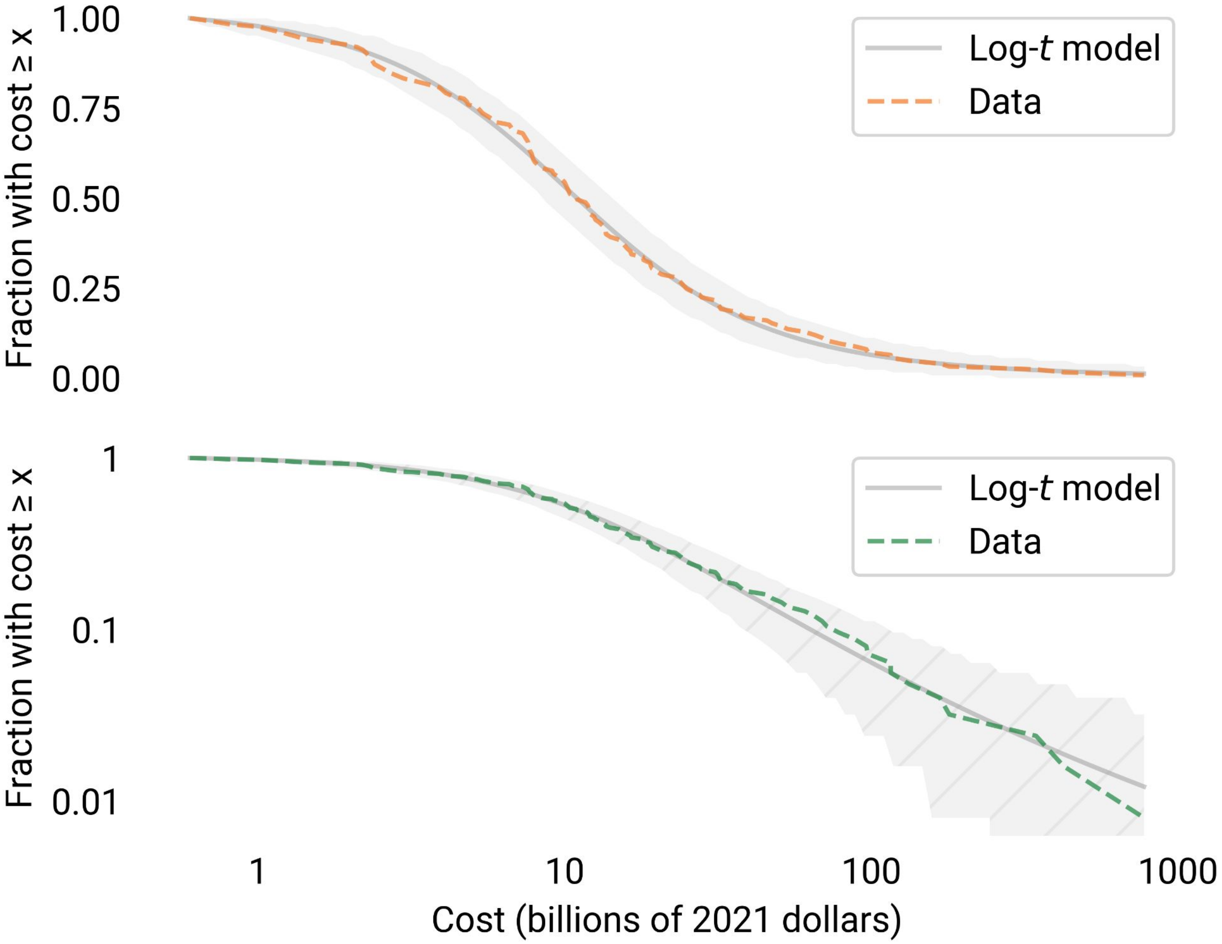
tinyurl.com/longtail23



tinyurl.com/longtail23

With $\nu = 3.5$, the *t* distribution fits the log cost of disasters.

Tail distribution of disaster costs with log-*t* model



tinyurl.com/longtail23

Top

- y-axis on a linear scale
- The model fits the data over the "normal range"

Bottom

- y-axis on a log scale
- The model fits the tail.

Note:

- Normal distribution of logs $\longrightarrow$ lognormal.
- *t* distribution of logs $\longrightarrow$ log-*t*.

That's what I'm calling it.

Let's consider earthquakes.

SCEDC

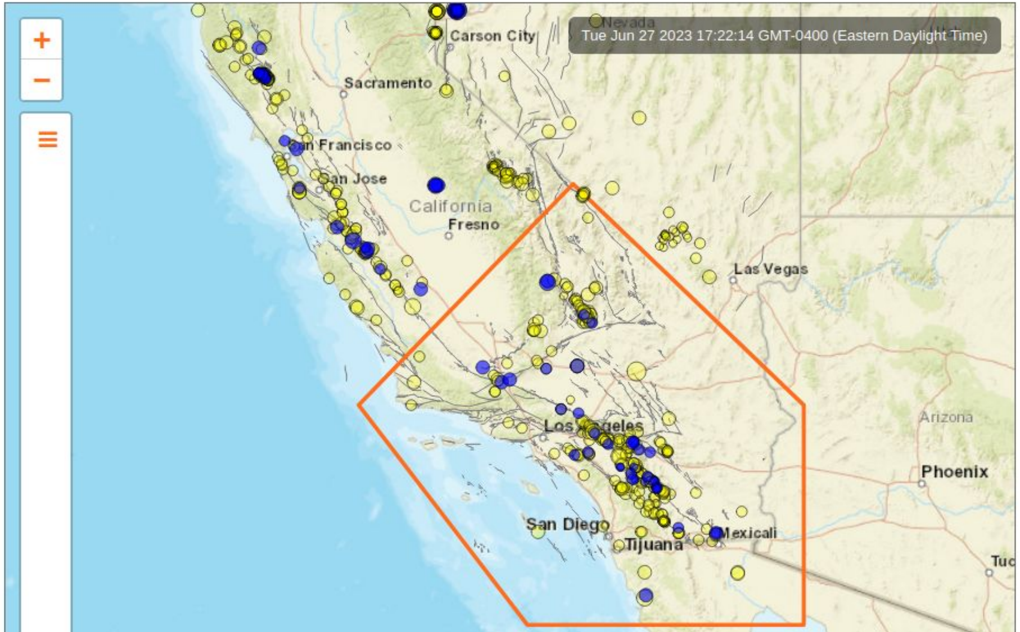
Southern California Earthquake Data Center

For the SCSN Visit

Welcome

The Southern California Earthquake Data Center (SCEDC) is the archive of the Caltech/USGS Southern California Seismic Network (SCSN). It is funded by the U.S. Geological Survey (USGS) and the Southern California Earthquake Center (SCEC). Its primary mission is to distribute data recorded or processed by the SCSN, a component of the California Integrated Seismic Network (CISN).

Recent Earthquakes in the Southern California Region



Access Data

Earthquake Updates

06/24/2023, M3.5 near Barstow

06/18/2023, M3.6 near Lake Isabella

05/30/2023, M3.6 near Port Hueneme

05/20/2023, M4.3 near Big Pine

05/02/2023, M3.6 near Little Lake

Worldwide Earthquakes

M 6.4 - Gulf of California

M 4.8 - 4 km SSE of Courçon, France

M 7.2 - south of the Fiji Islands

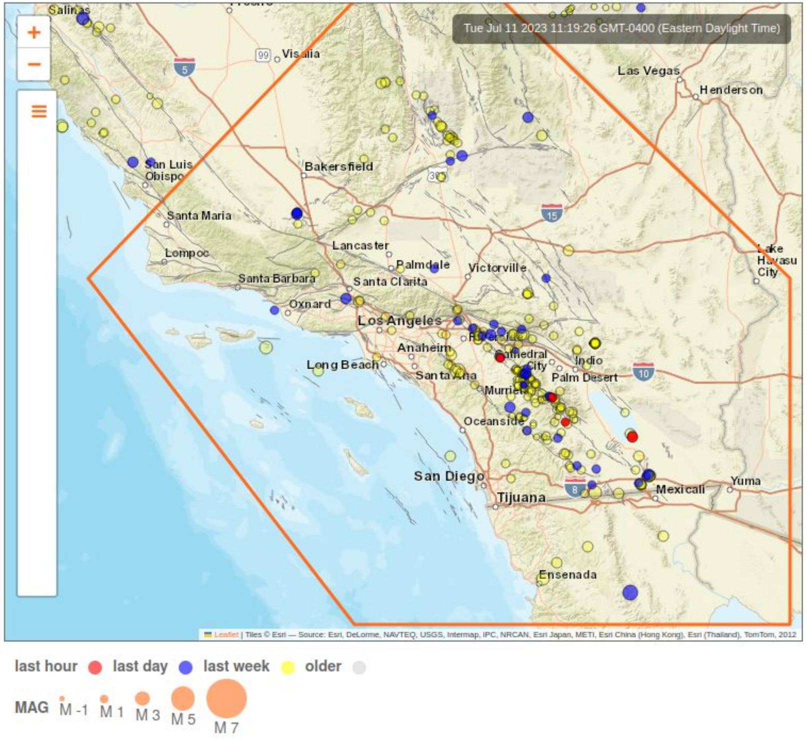
News and Updates

Planned Maintenance Notice 02/14/2023

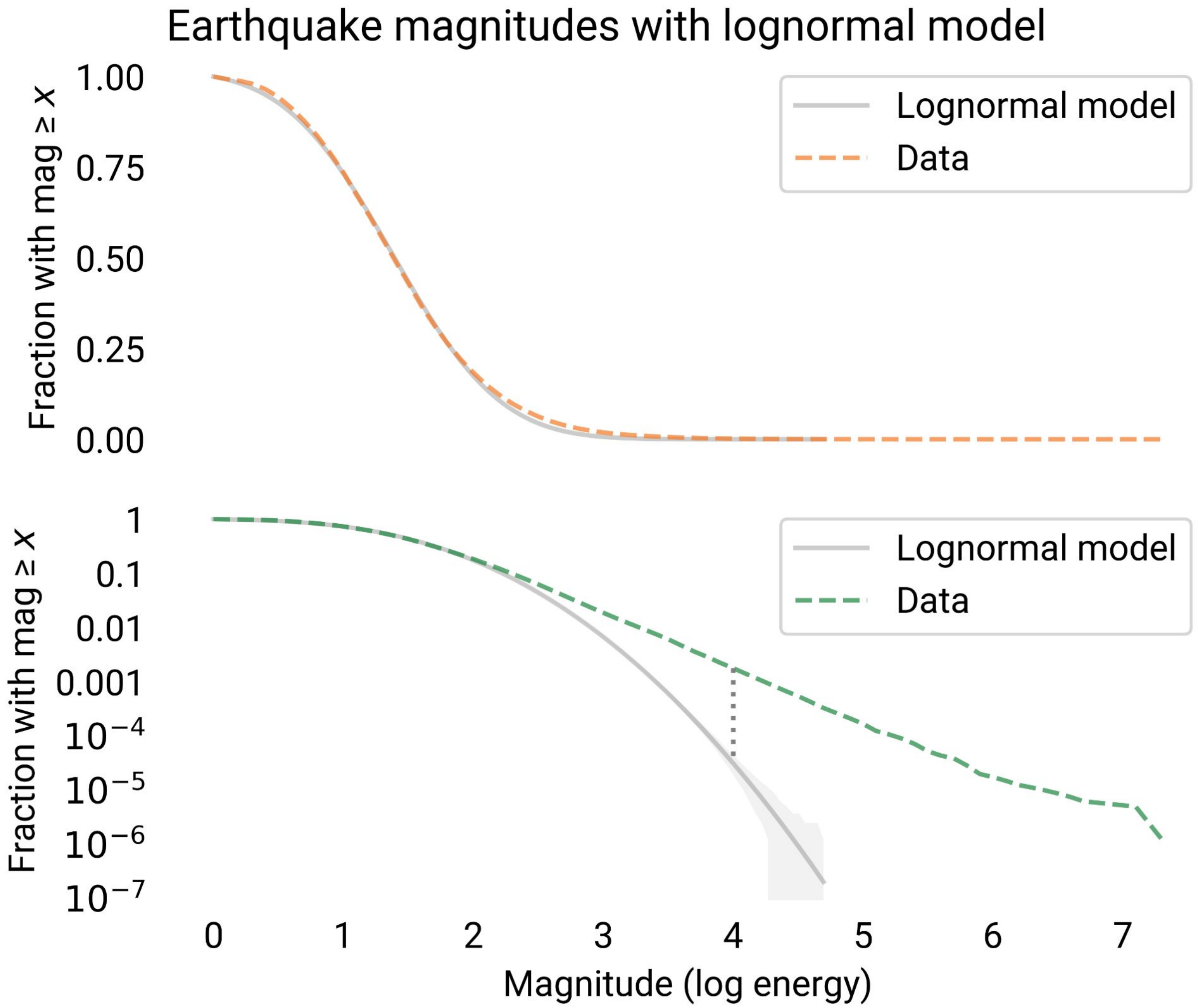
tinyurl.com/longtail23

791,329 earthquakes in Southern California from January 1981 to April 2022.

Recent Earthquakes in the Southern California Region



tinyurl.com/longtail23



tinyurl.com/longtail23

On a linear y-axis, the model seems OK.

It is not.

Fraction of earthquakes with magnitude 4 or more:

Model: 33 per million

Data: 1800 per million

Off by a factor of 55.

Fraction of earthquakes with magnitude 7 or more:

Model: 5 per 10^{18}

Data: 6 per 10^{6}

Off by 12 orders of magnitude.

Wikipedia
The Free Encyclopedia

Search Wikipedia | Search

AllenDowney

Main menu [hide]

- Main page
- Contents
- Current events
- Random article
- About Wikipedia
- Contact us
- Donate

Switch to old look

Contribute

- Help
- Learn to edit
- Community portal
- Recent changes
- Upload file

2019 Ridgecrest earthquakes

14 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

Coordinates: 35.766°N 117.605°W

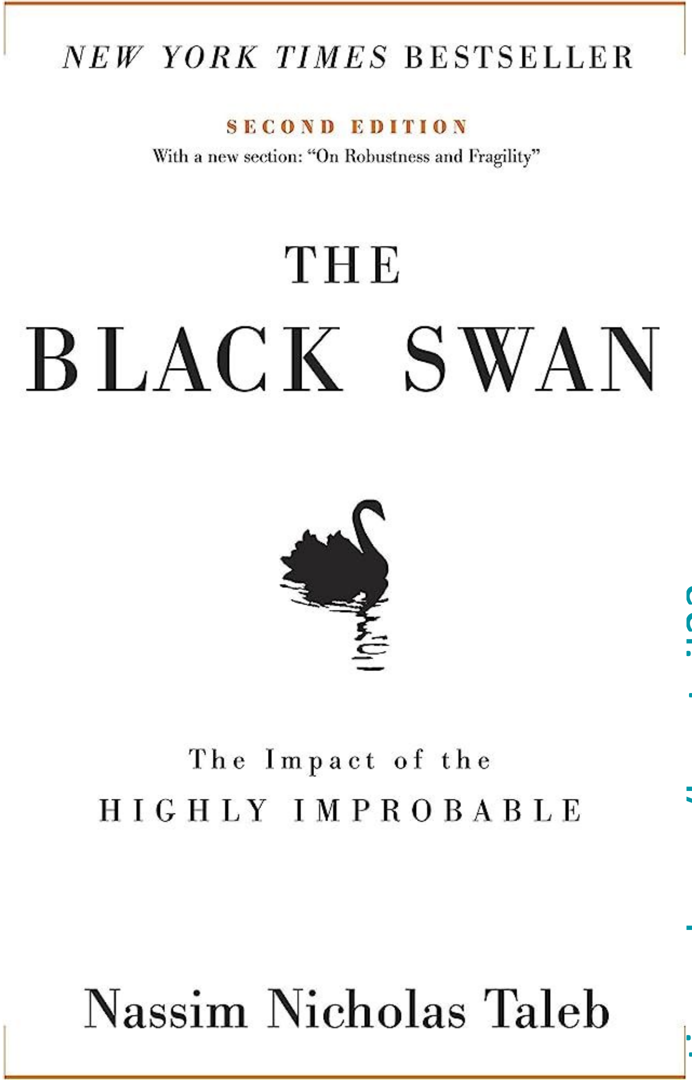
The **2019 Ridgecrest earthquakes** (more commonly referred to in scientific literature as the **2019 Ridgecrest earthquake sequence**) of July 4 and 5 occurred north and northeast of the town of Ridgecrest, California located in Kern County and west of Searles Valley (approximately 200 km [122 mi] north-northeast of Los Angeles). They included three initial main shocks of M_w magnitudes 6.4, 5.4, and 7.1,[8] and many perceptible aftershocks, mainly within the area of the Naval Air Weapons Station China Lake. Eleven months later, a M_w 5.5 aftershock took place (the largest aftershock of the sequence) to the east of Ridgecrest.

2019 Ridgecrest earthquakes

tinyurl.com/longtail23

This is what Taleb calls a Black Swan.

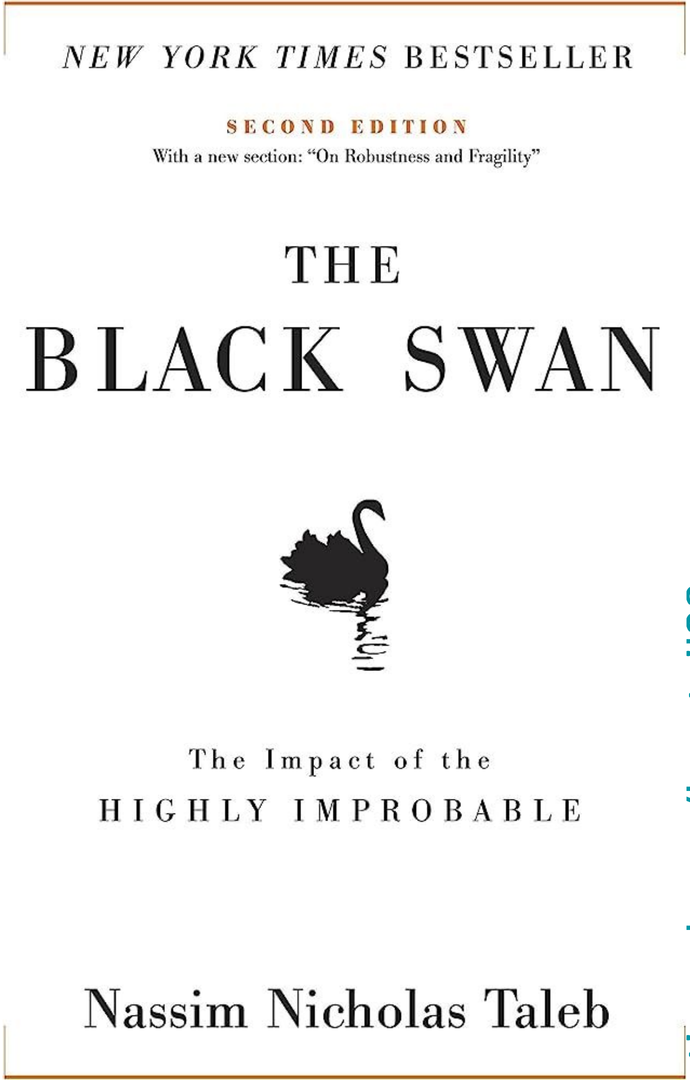
- Large, impactful event
- Considered extremely unlikely
- Based on a model of prior events



tinyurl.com/longtail23

This is what Taleb calls a Black Swan.

- Large, impactful event
- Considered extremely unlikely
- Based on a model of prior events

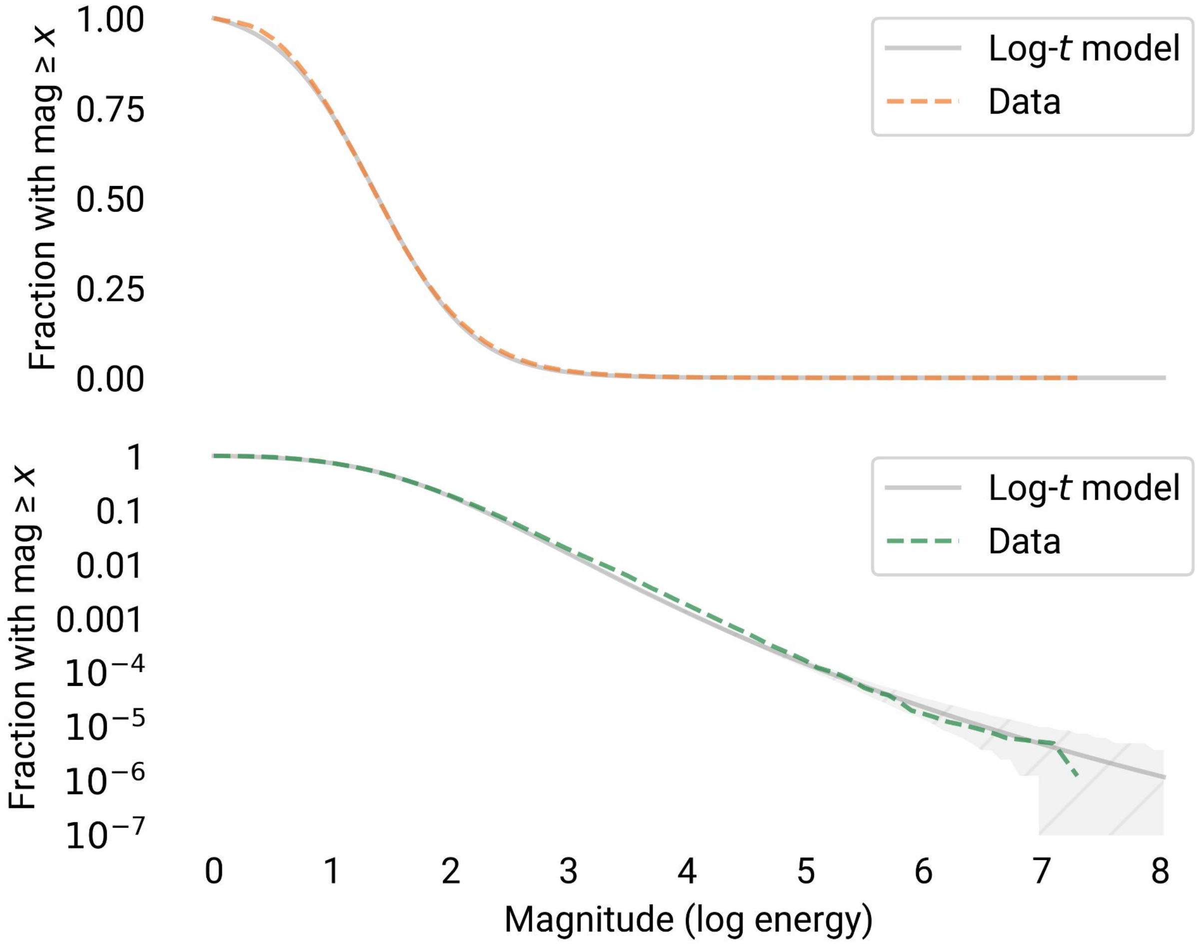


tinyurl.com/longtail23

With the lognormal model,
a magnitude 7.1 earthquake is a
Black Swan.

Let's see if we can do better with a
log-*t* distribution.

Earthquake magnitudes with log-t model



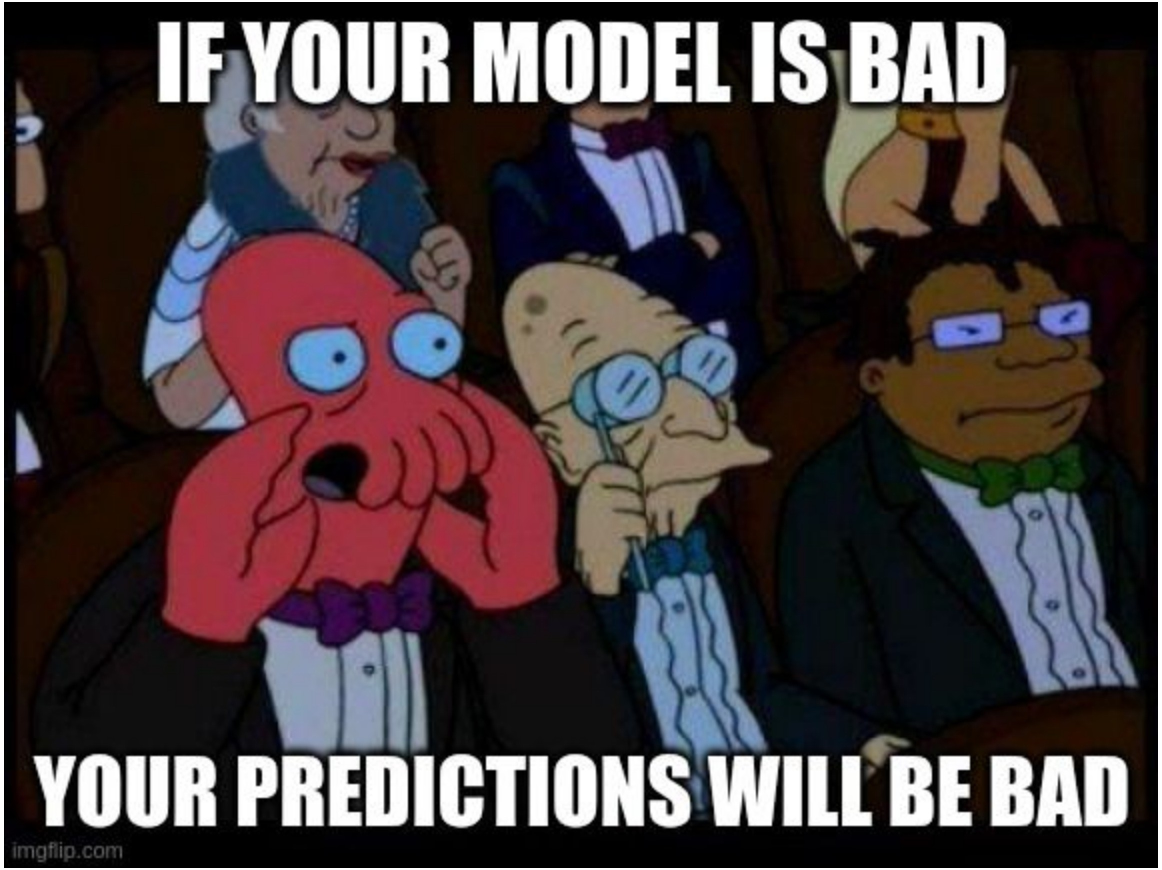
tinyurl.com/longtail23

If you have enough data,
use an appropriate model,
and make reliable predictions,
you have "tamed" the Black Swan.

Now it's a Gray Swan.

Black Swan hypothesis (weak form):

If your model is bad,
your predictions will be bad.



tinyurl.com/longtail23

Black Swan hypothesis (strong form):

Some black swans can't be tamed.

Or you can't know whether you have.

tinyurl.com/longtail23

Candidate for an untameable swan: solar flares

tinyurl.com/longtail23

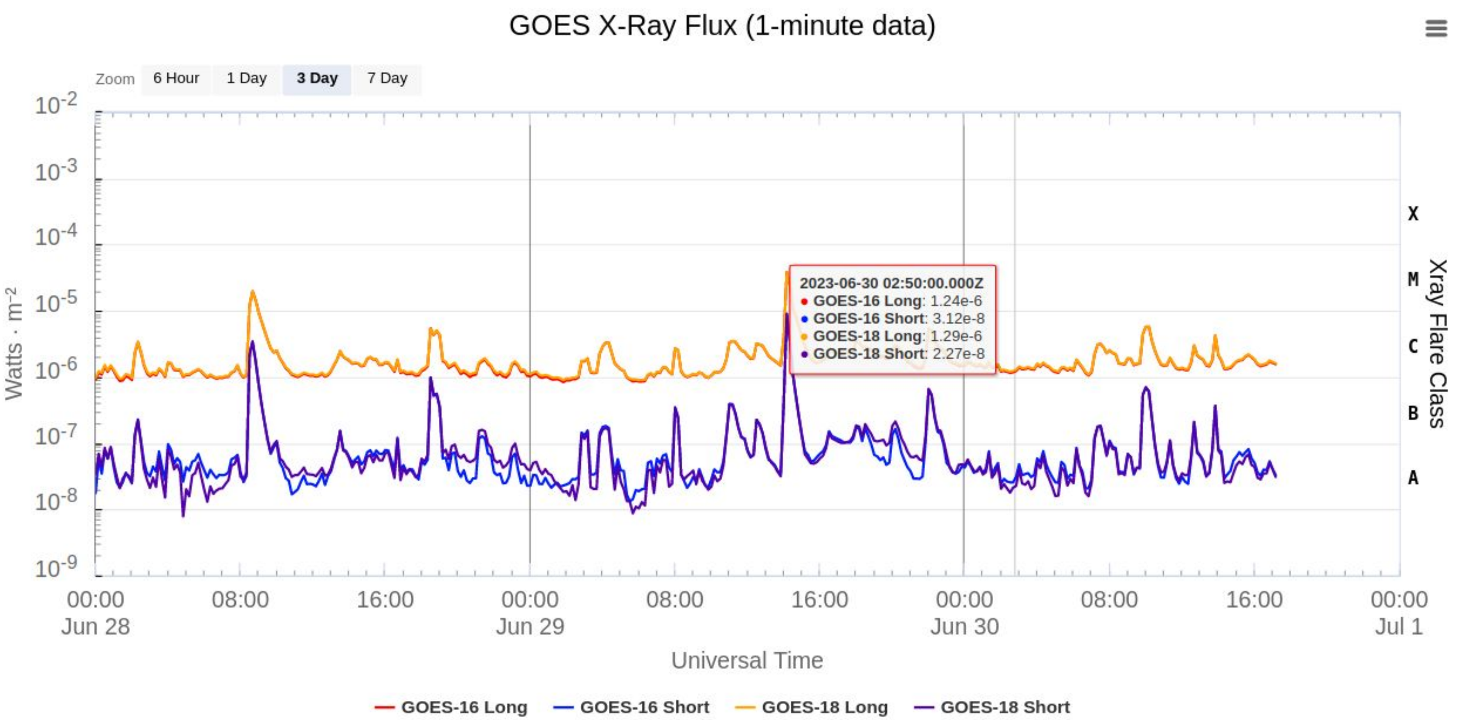
SPACE WEATHER PREDICTION CENTER
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Friday, June 30, 2023 17:23:06 UTC

CURRENT SPACE WEATHER CONDITIONS on NOAA Scales

R none | S none | G none

GOES X-RAY FLUX



Updated 2023-06-30 17:18 UTC

Space Weather Prediction Center

tinyurl.com/longtail23

GOES Satellite Network

GOES Family

- GOES-R
- GOES-N Series

Related Topics

- GOES
- All Topics A-Z

GOES Overview and History

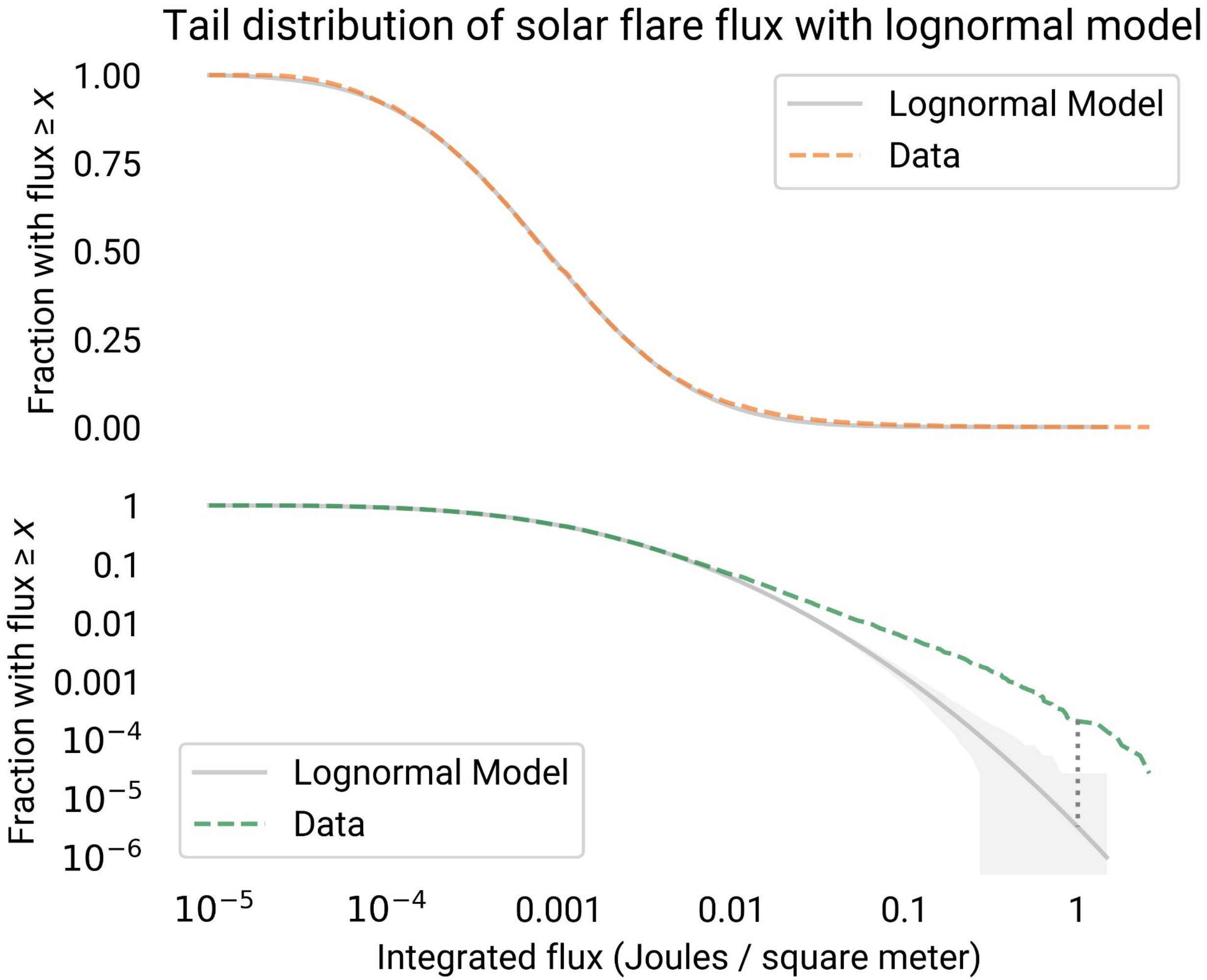
GOES Project Current Status

The Geostationary Operational Environmental Satellite Program (GOES) is a joint effort of NASA and the National Oceanic and Atmospheric Administration (NOAA).

The GOES system currently consists of GOES-13, operating as GOES-East, in the eastern part of the constellation at 75 degrees west longitude and GOES-15, operating as GOES-West, at 135 degrees west longitude. The GOES-R series will maintain the two-satellite system implemented by the current GOES series. However, the locations of the operational GOES-R satellites will be 75 degrees west longitude and 137 degrees west longitude. The latter is a shift in order to eliminate conflicts with other satellite systems. The GOES-R series operational lifetime extends through December 2036.

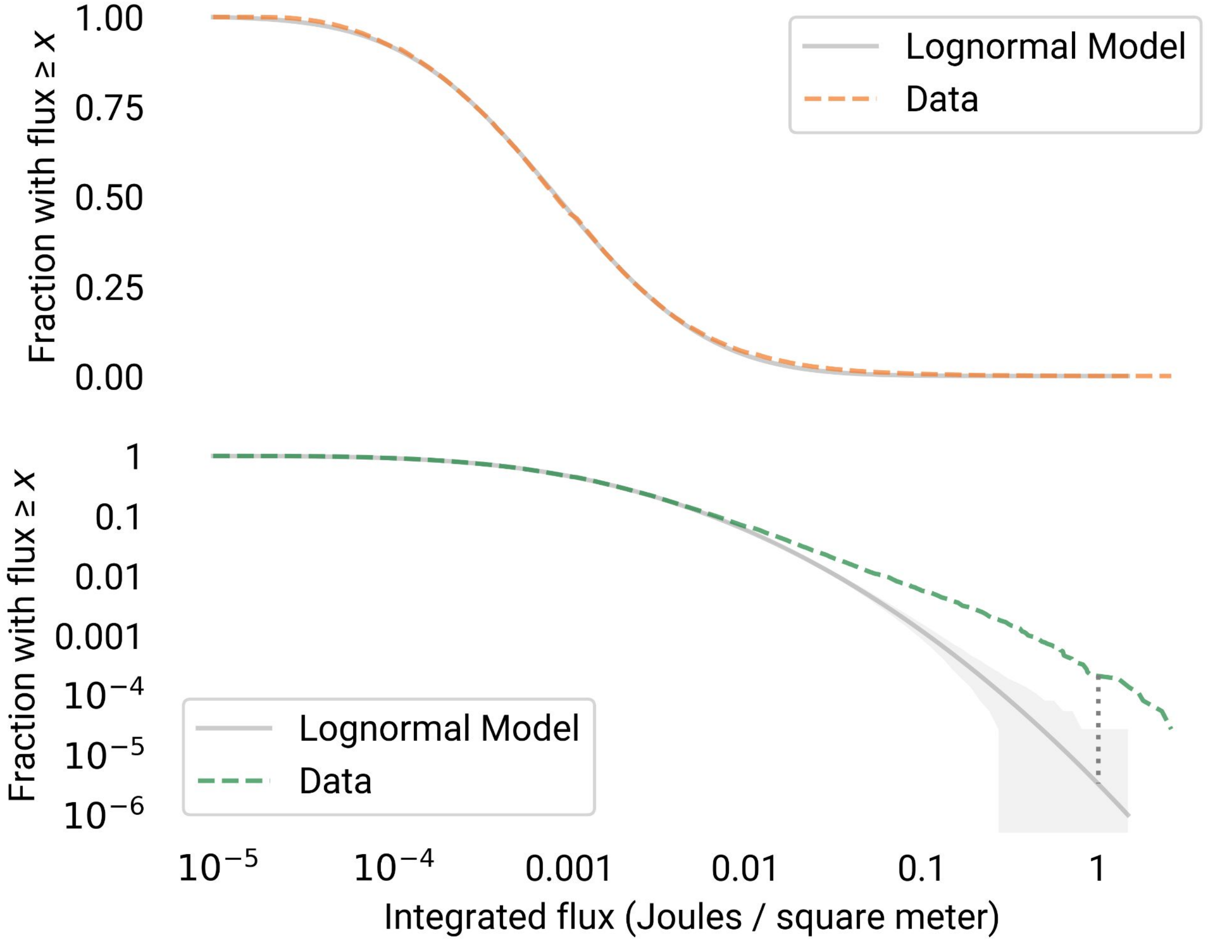
Artist's rendering of GOES-R
Credits: NASA

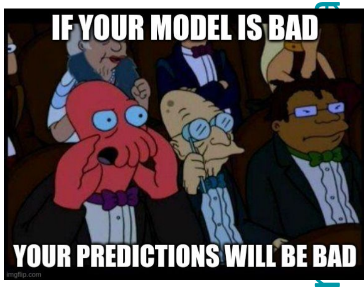
Integrated flux (J / m^2)
from 36,000 solar flares
1997 to 2017.



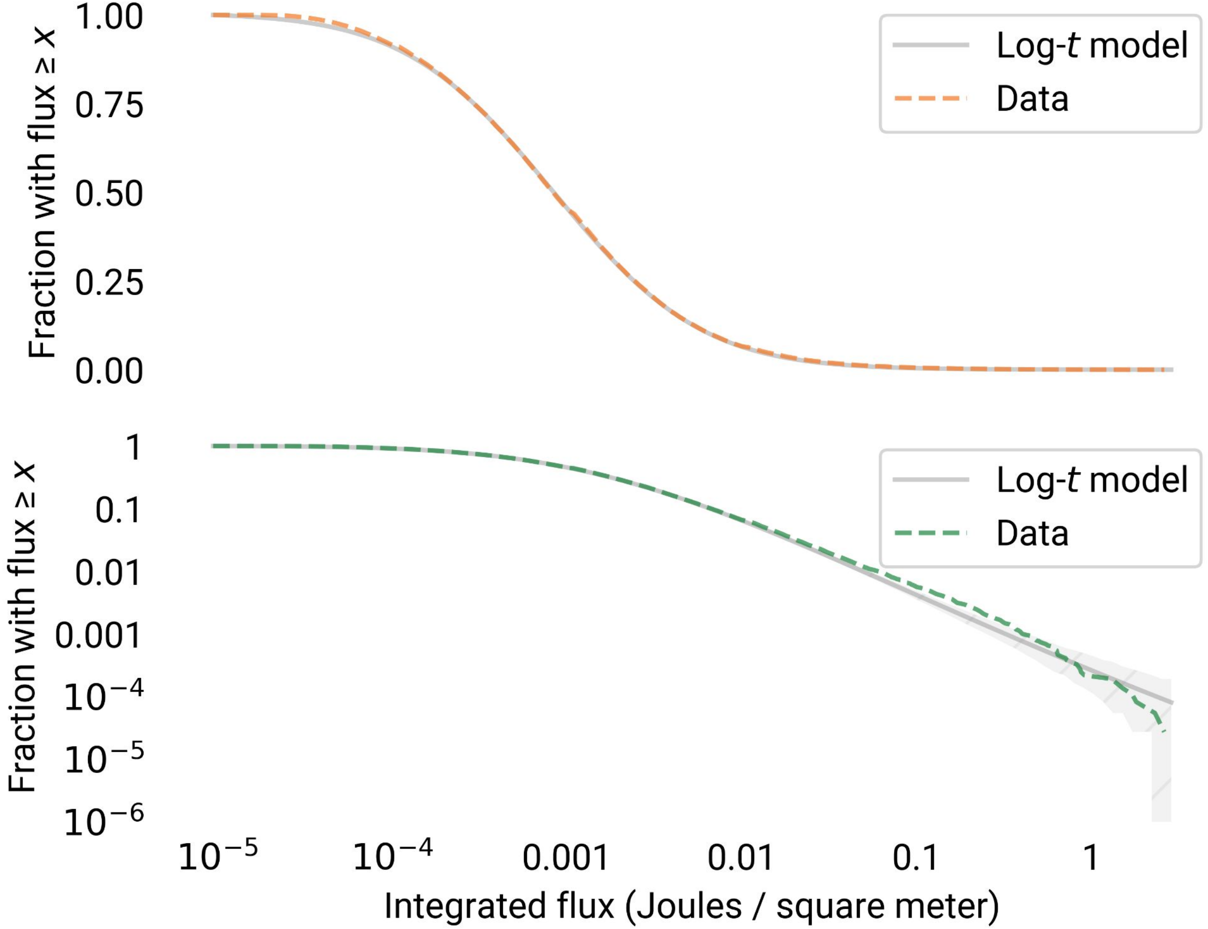
tinyurl.com/longtail23

Tail distribution of solar flare flux with lognormal model





Tail distribution of solar flare flux with log-*t* model



tinyurl.com/longtail23

The log-*t* model is better.

So let's put it to the test…

Main menu [hide]

Main page

Contents

Current events

Random article

About Wikipedia

Contact us

Donate

Switch to old look

Contribute

Help

Learn to edit

Community portal

Recent changes

Superflare

8 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

This article is about the extrasolar phenomenon on solar-type stars. For other uses, see Super flare (disambiguation).

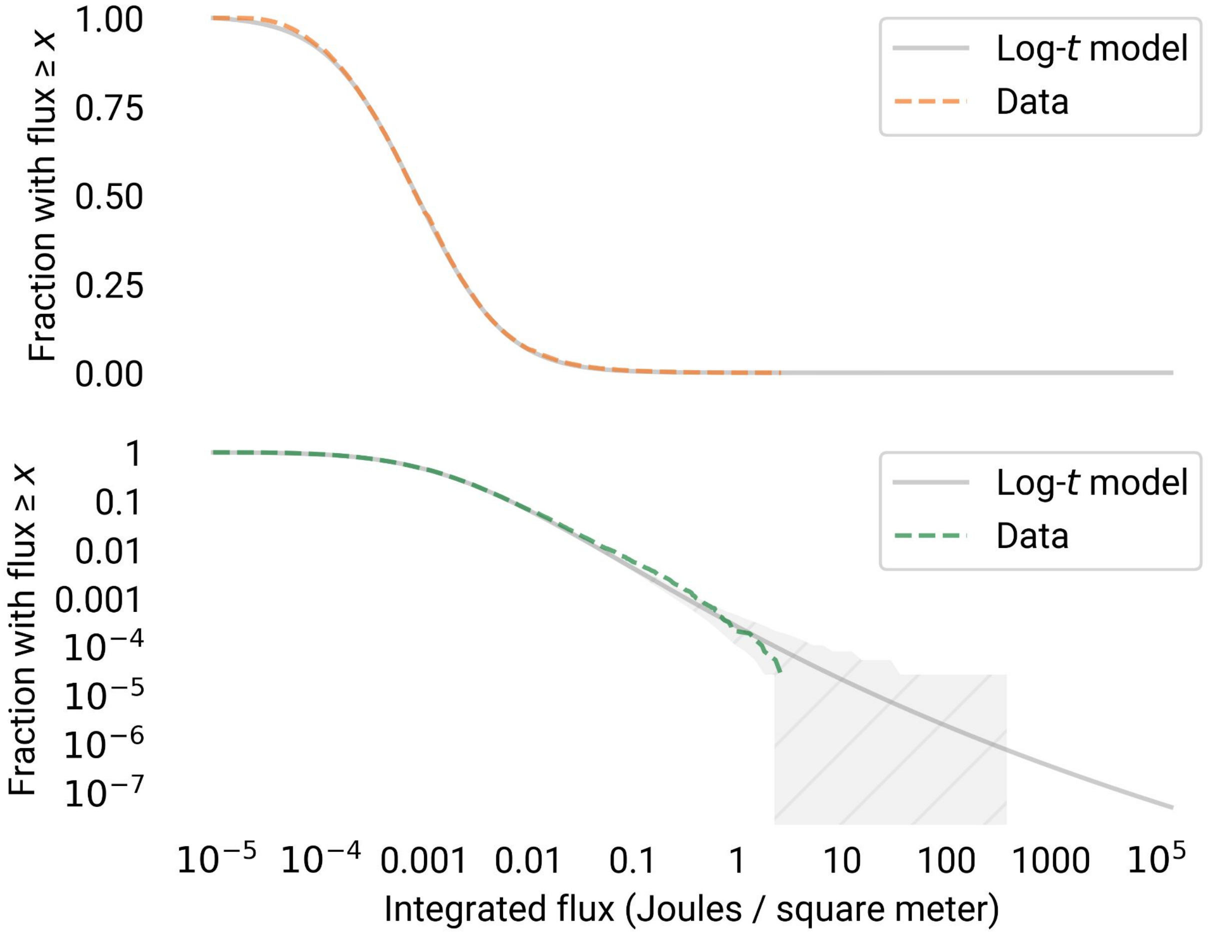
Superflares are very strong explosions observed on stars with energies up to ten thousand times that of typical solar flares. The stars in this class satisfy conditions which should make them solar analogues, and would be expected to be stable over very long time scales. The original nine candidates were detected by a variety of methods. No systematic study was possible until the launch of the Kepler space telescope, which monitored a very large number of solar-type stars with very high accuracy for an extended period. This showed that a small proportion of stars had violent outbursts, up to 10,000 times as powerful as the strongest flares known on the Sun. In many cases there were multiple events on the same star. Younger stars were more likely to flare than old ones, but strong events were seen on stars as old as the Sun.

We would like to know the probability of a superflare from our Sun.

Can we answer that with our model?

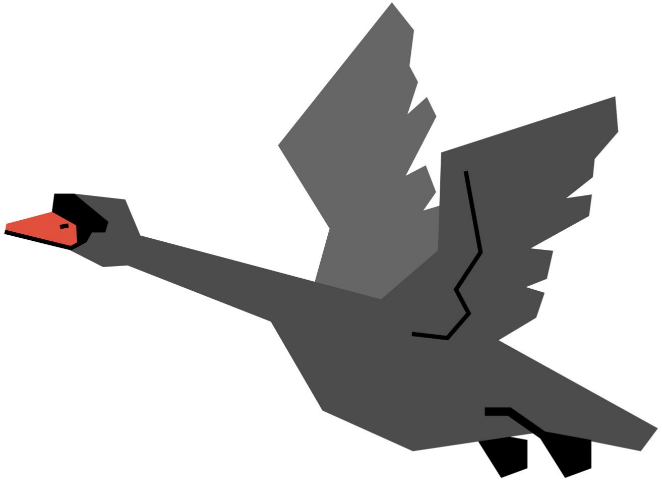
https://earthsky.org/space/what-is-the-suns-name/

Tail distribution of solar flare flux with log-*t* model



tinyurl.com/longtail23

How much confidence should we have in an extrapolation that goes four orders of magnitude beyond the data?



tinyurl.com/longtail23

How much confidence should we have in an extrapolation that goes four orders of magnitude beyond the data?

tinyurl.com/longtail23

Reasons for caution:

- Uncertainty due to random sampling
- Hints that the model is not the right shape
- Extrapolating far beyond the data

To tame this swan,
we need
more data and
more astrophysics.

To tame this swan,
we need
more data and
more astrophysics.

In 20 years we've seen 36,000 flares.

It would take more than 500 years to observe a million.

That would still be short by 2 orders of magnitude.

Anyway, we don't know whether our Sun can produce a superflare at all.

This is where physical models can help (as opposed to purely statistical).

It also helps to know where long-tailed distributions come from.

Main menu [hide]

Main page

Contents

Current events

Random article

About Wikipedia

Contact us

Donate

Switch to old look

Contribute

Help

Learn to edit

Community portal

Recent changes

Upload file

Languages

Language links are at the top

Preferential attachment

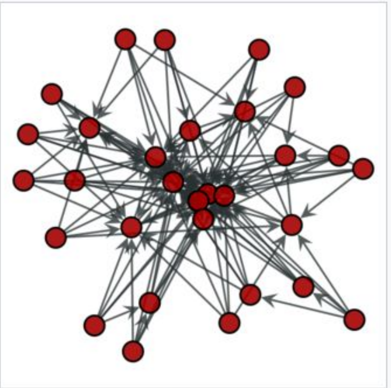
5 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

"Yule process" redirects here. For the type of birth process, see Simple birth process.

A **preferential attachment process** is any of a class of processes in which some quantity, typically some form of wealth or credit, is distributed among a number of individuals or objects according to how much they already have, so that those who are already wealthy receive more than those who are not. "Preferential attachment" is only the most recent of many names that have been given to such processes. They are also referred to under the names **Yule process**, **cumulative advantage**, the rich get richer, and the Matthew effect. They are also related to Gibrat's law. The principal reason for scientific interest in preferential attachment is that it can, under suitable circumstances, generate power law distributions.[1] If preferential attachment is non-linear, measured distributions may deviate from a power law.[2] These mechanisms may generate distributions which are approximately power law over transient periods.[3][4]



Graph generated using preferential attachment. A small number of nodes have a large number of incoming edges, whereas a large number of nodes have a small number of incoming edges.

tinyurl.com/longtail23

Wikipedia
The Free Encyclopedia

Search Wikipedia | Search

AllenDowney

Main menu [hide]

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Switch to old look

Contribute

Help
Learn to edit
Community portal
Recent changes
Upload file

Languages

Language links are at the top

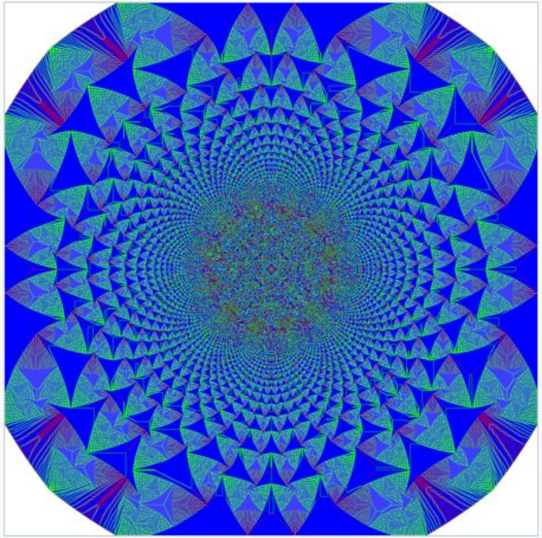
Self-organized criticality

10 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

Self-organized criticality (**SOC**) is a property of dynamical systems that have a critical point as an attractor. Their macroscopic behavior thus displays the spatial or temporal scale-invariance characteristic of the critical point of a phase transition, but without the need to tune control parameters to a precise value, because the system, effectively, tunes itself as it evolves towards criticality.



An image of the 2d Bak-Tang-Wiesenfeld sandpile, the original model of self-organized criticality.

The concept was put forward by Per Bak, Chao Tang and Kurt Wiesenfeld ("BTW") in a paper[1] published in 1987 in *Physical Review Letters*, and is considered to be one of the mechanisms by which complexity[2] arises in nature. Its concepts have been applied across fields as diverse as geophysics,[3][4] physical cosmology, evolutionary biology and ecology, bio-inspired computing and optimization (mathematics),

tinyurl.com/longtail23

Student's *t* distribution:

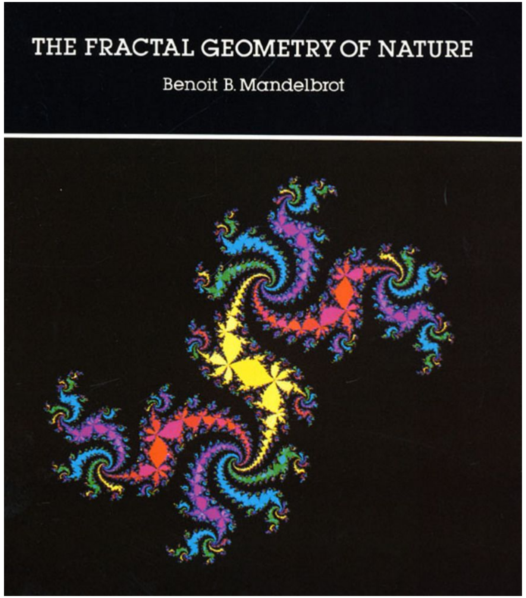
- Mixture of normal distributions with different variance

Log-*t* distribution:

- Mixture of lognormal distributions

And Mandelbrot's "heretical" explanation:

- The data are *the joint effect of a fixed underlying true distribution and a highly variable filter*, which
- *… leaves the asymptotic behavior unchanged*.



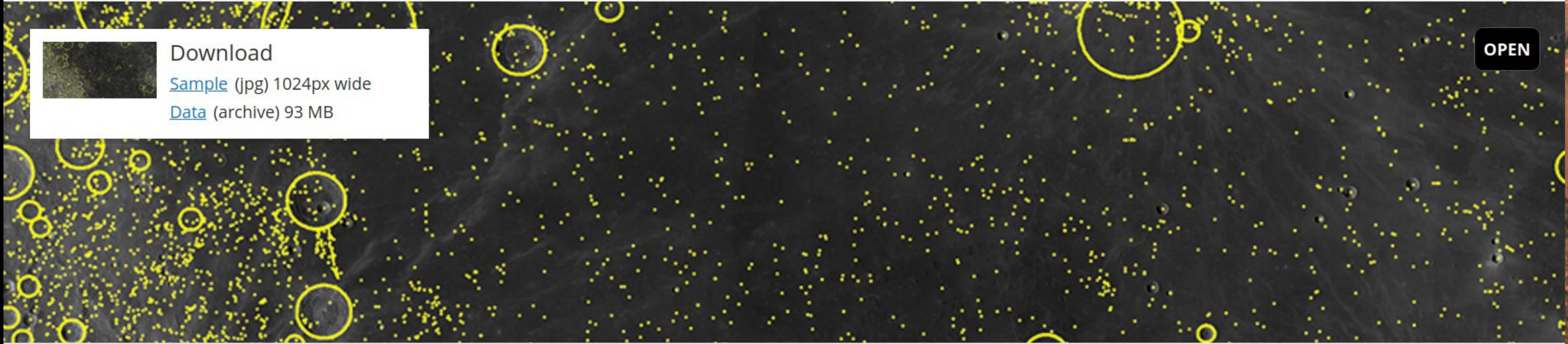
tinyurl.com/longtail23

As an example, lunar craters.

PDS Planetary Data System NASA USGS science for a changing world

Annex NASA PDS and Derived Products

ASTROPEDIA
Lunar and Planetary Cartographic Catalog



Download
Sample (jpg) 1024px wide
Data (archive) 93 MB

OPEN

Moon Crater Database v1 Robbins

Product Information:

The **Lunar Crater Database** contains approximately 1.3 million lunar impact craters and is approximately complete for all craters larger than about 1–2 km in diameter (Robbins, 2018). Craters were manually identified and measured on Lunar Reconnaissance Orbiter (LRO) Camera (LROC) Wide-Angle Camera (WAC) images, in LRO Lunar Orbiter Laser Altimeter (LOLA) topography, SELENE Kaguya Terrain Camera (TC) images, and a merged LOLA+TC DTM (Barker, 2016).

This archive uses PDS4 archiving standards. An overview of PDS4 is provided in the PDS4 Concepts document (2018) and the standards are specified in the PDS4 Standards Reference (2018).

PDS Status: PDS 4 In Review

FGDC: xml metadata

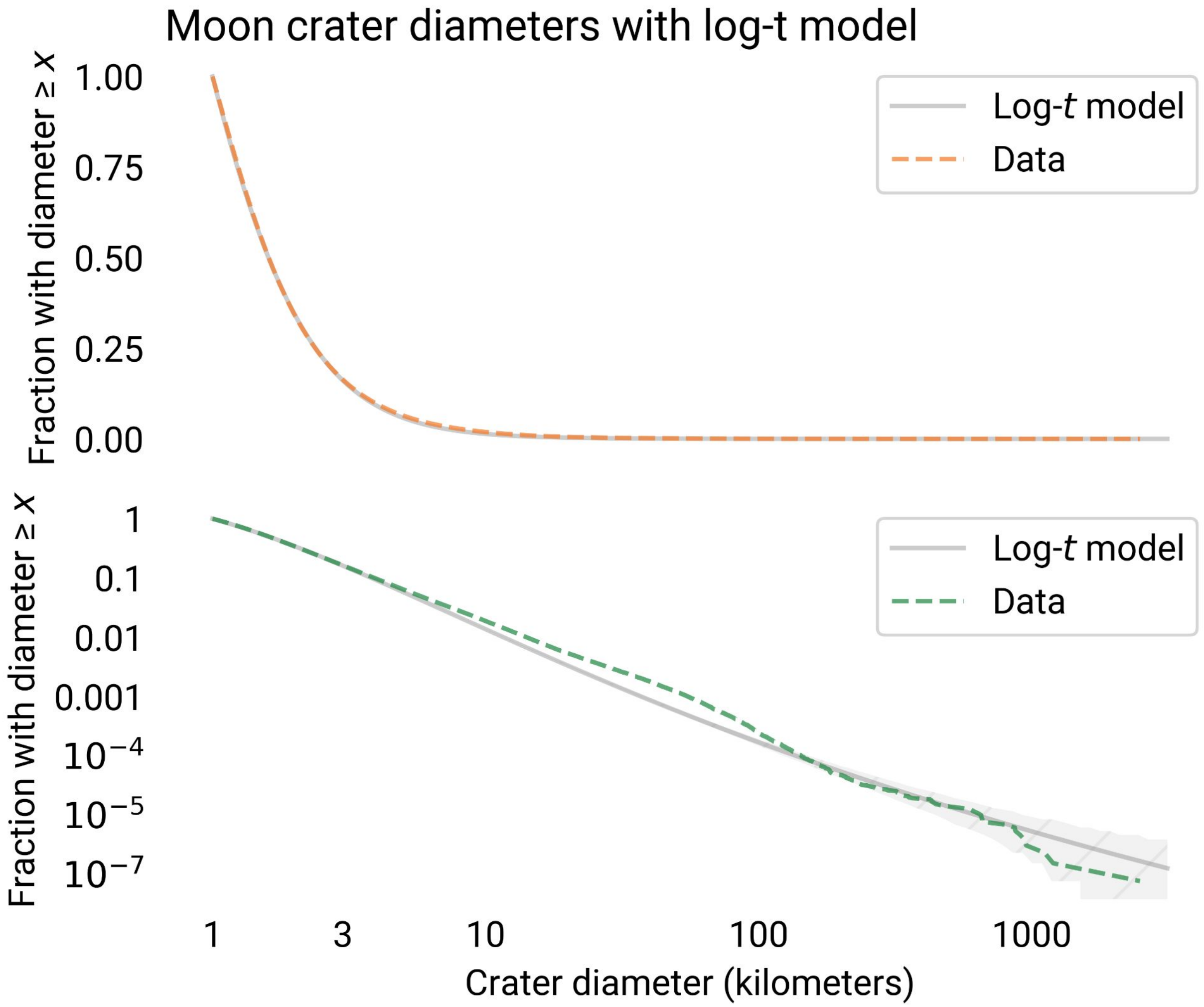
Ancillary Data

Catalog_Moon_Release_20180815_shapefile180.zip (zip) 131 MB
Catalog_Moon_Release_20180815_1kmPlus.vrt (vrt) 2 kB

1.3 million craters

larger than ~1 km diameter.

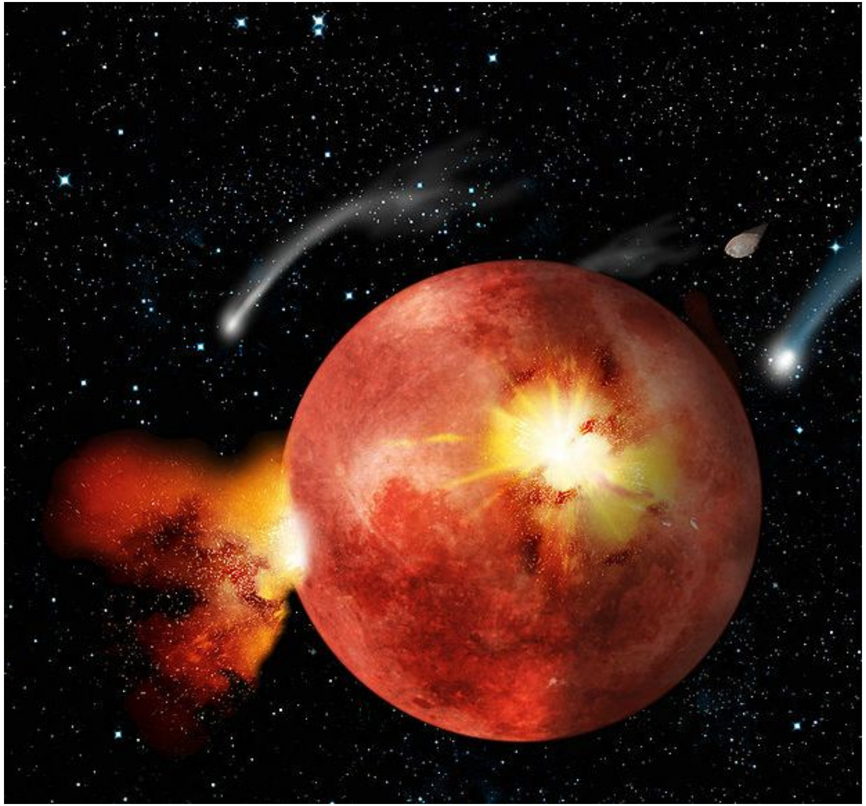
tinyurl.com/longtail23



tinyurl.com/longtail23

Why does the distribution of craters have this shape?

Most formed during the
Late Heavy Bombardment,
by asteroids from the Main Belt.



https://commons.wikimedia.org/wiki/File:Lunar_cataclysm.jpg

tinyurl.com/longtail23

Small-Body Database Query

Use this query tool to generate custom tables of orbital and/or physical parameters for asteroids and/or comets of interest from our small-body database (SBDB). For details on a single specific asteroid or comet, use the SBDB Lookup tool instead.

+ Limit by Object Kind/Group

+ Limit by Orbit Class

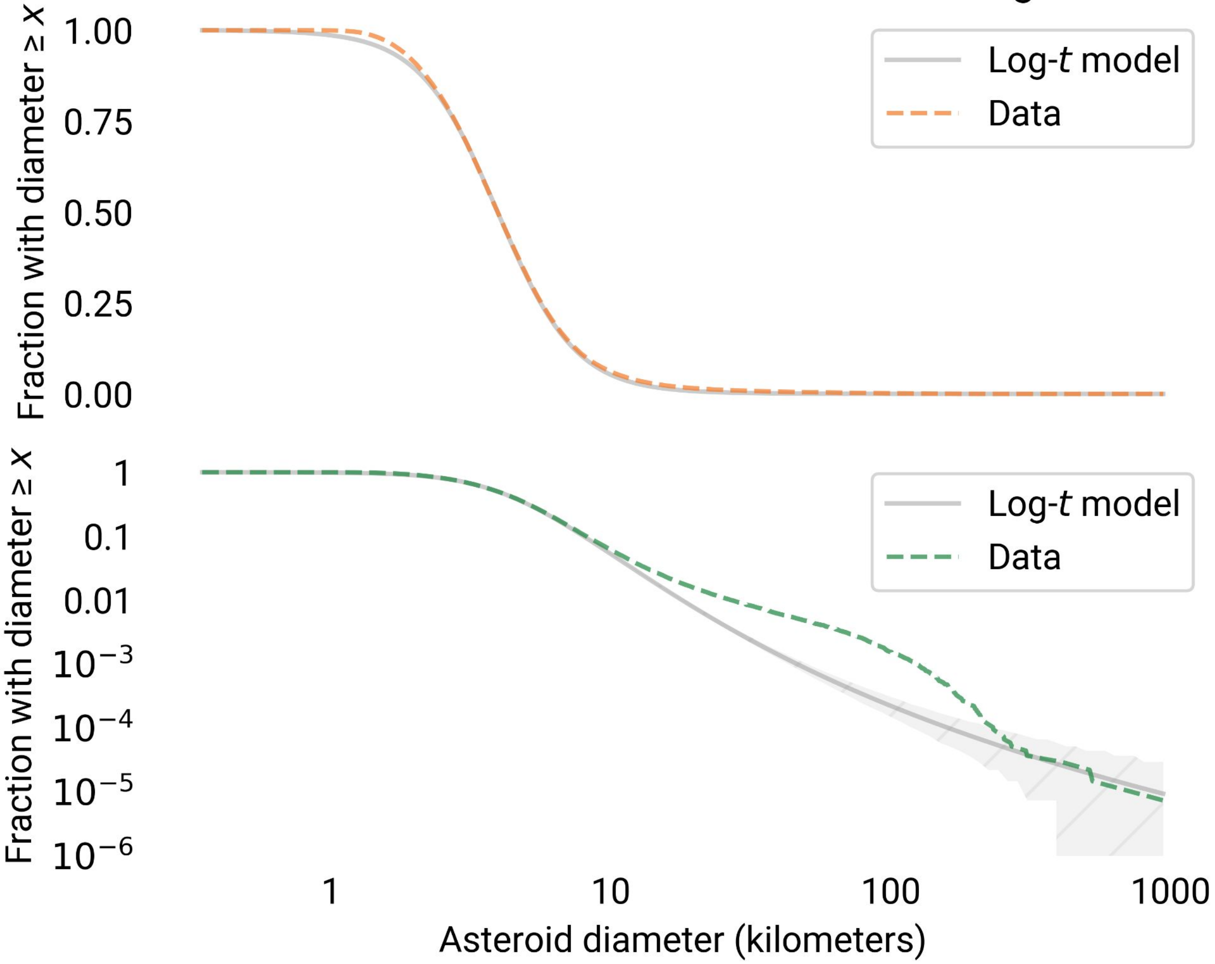
+ Custom Object/Orbit Constraints

+ Output Selection Controls

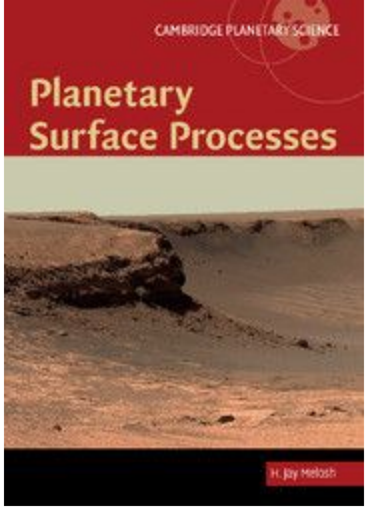
Get Results

135,915 asteroids with known diameter.

Tail distribution of asteroid diameters with log-t model



tinyurl.com/longtail23



6.5.1 *Crater diameter scaling*

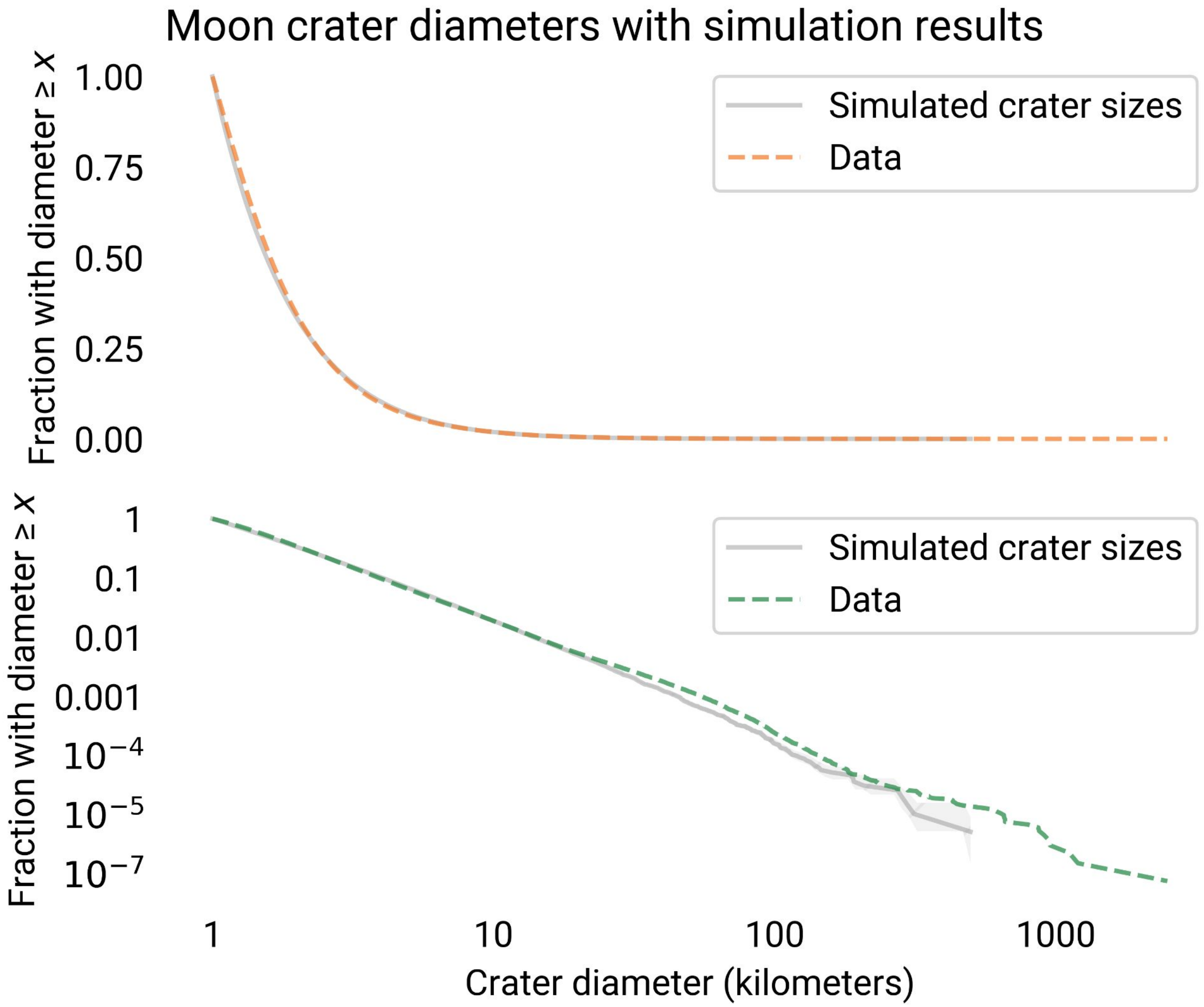
This situation has changed rapidly in the last few decades, however, thanks to more impact cratering experiments specifically designed to test scaling laws. It has been shown that the great expansion of the crater during excavation tends to decouple the parameters describing the final crater from the parameters describing the projectile. If these sets of parameters are related by a single, dimensional "coupling parameter" (as seems to be the case), then it can be shown that crater parameters and projectile parameters are related by power-law scaling expressions with constant coefficients and exponents. Although this is a somewhat complex and rapidly changing subject, the best current scaling relation for impact craters forming in competent rock (low-porosity) targets whose growth is limited by gravity rather than target strength (i.e. all craters larger than a few kilometers in diameter) is given by:

$$D_{tc} = 1.161 \left(\frac{\rho_p}{\rho_t} \right)^{1/3} L^{0.78} \, v_i^{0.44} \, g^{-0.22} \sin^{1/3} \theta \qquad (6.12)$$

where D_{tc} is the diameter of the transient crater at the level of the original ground surface, ρ_p and ρ_t are densities of the projectile and target, respectively, g is surface gravity, L is projectile diameter, v_i is impact velocity and θ is the angle of impact from the horizontal. All quantities are in SI units.

Raising these factors to powers multiplying them are *filters* that *leave their asymptotic behavior unchanged*.

If we simulate actual crater sizes and lognormal other factors…



tinyurl.com/longtail23

Plausibly:

Craters are long-tailed

because asteroids are long-tailed

because of preferential attachment (accretion).

Plausibly:

Craters are long-tailed

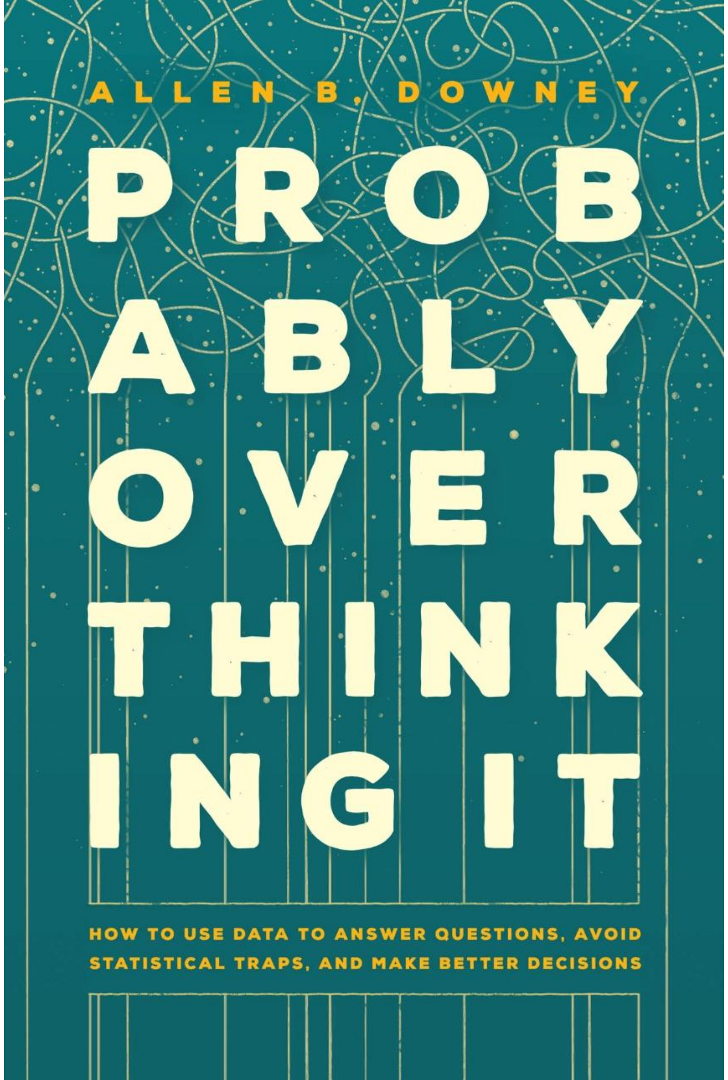
because asteroids are long-tailed — highly variable filter

because of preferential attachment. — underlying process

Summary

- Long-tailed distributions appear in many fields
- Bad models might seem OK on a linear scale
- Look at tail distributions on a log-log scale
- Use long-tailed models
- Some Black Swans may be untameable

Sources and further reading



Contents

tinyurl.com/longtail23

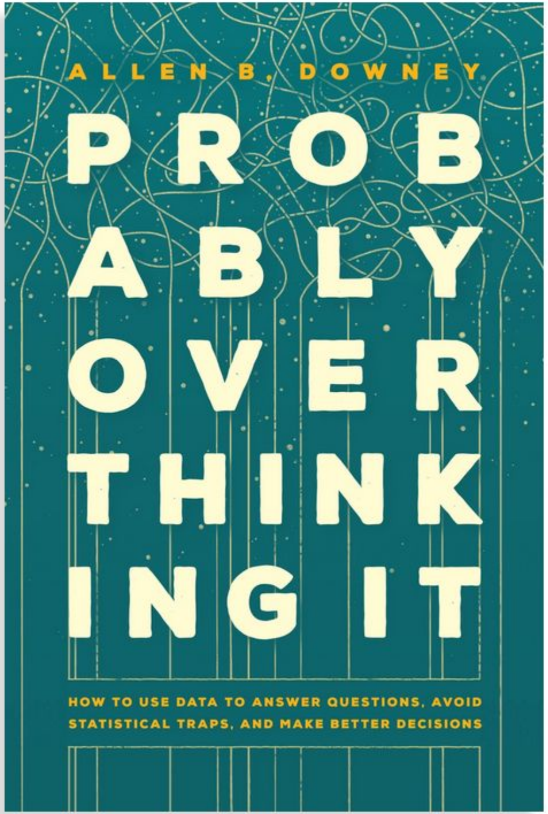
CHICAGO

The University of Chicago Press

Search Site

BOOKS JOURNALS RESOURCES DISTRIBUTION WHO WE ARE

AddThis



Probably Overthinking It

How to Use Data to Answer Questions, Avoid Statistical Traps, and Make Better Decisions

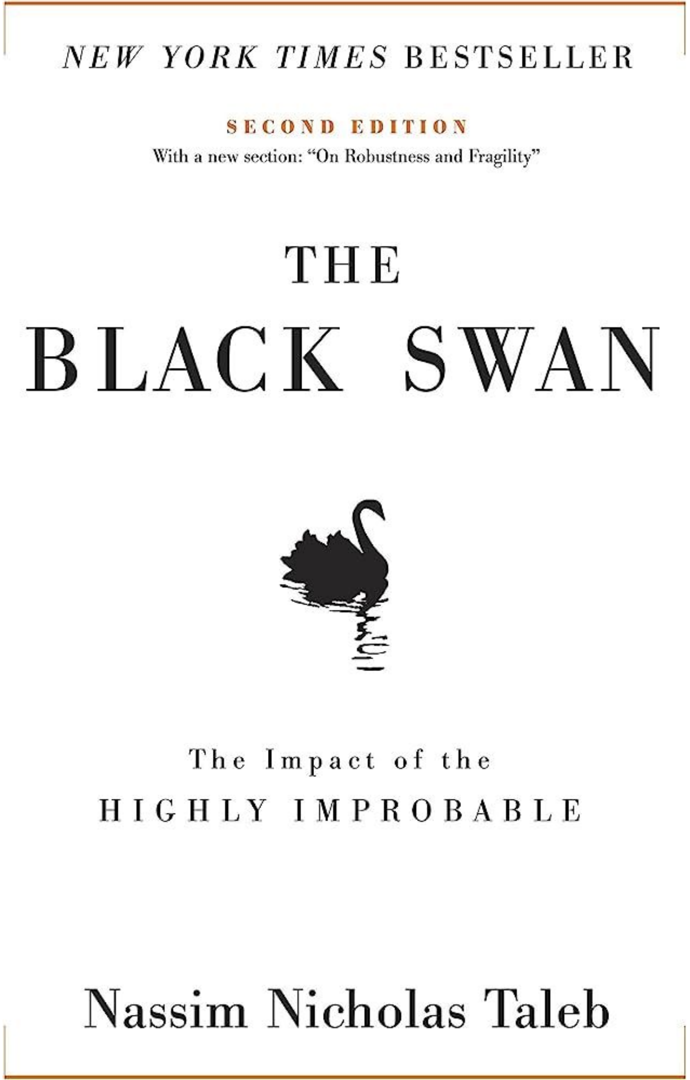
Allen B. Downey

An essential guide to the ways data can improve decision making.

Statistics are everywhere: in news reports, at the doctor's office, and in every sort of forecast, from the stock market to the weather report. Blogger, teacher, and computer scientist Allen B. Downey knows well that we have both an innate ability to understand statistics and to be fooled by them. As he makes clear in this accessible introduction to statistical thinking, the stakes are big. Simple misunderstandings have led to incorrect patient prognoses,

READ MORE +

tinyurl.com/longtail23



tinyurl.com/longtail23

[Submitted on 24 Jan 2020 (v1), last revised 14 Nov 2022 (this version, v3)]

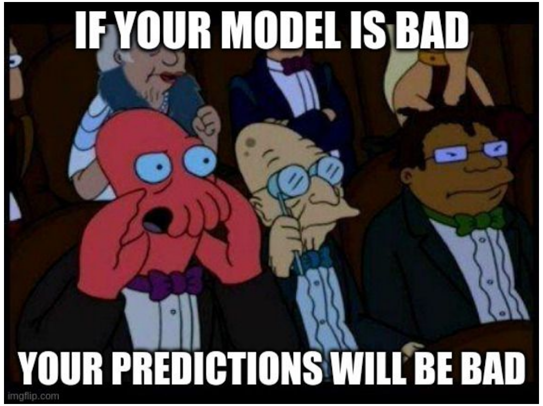
Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications

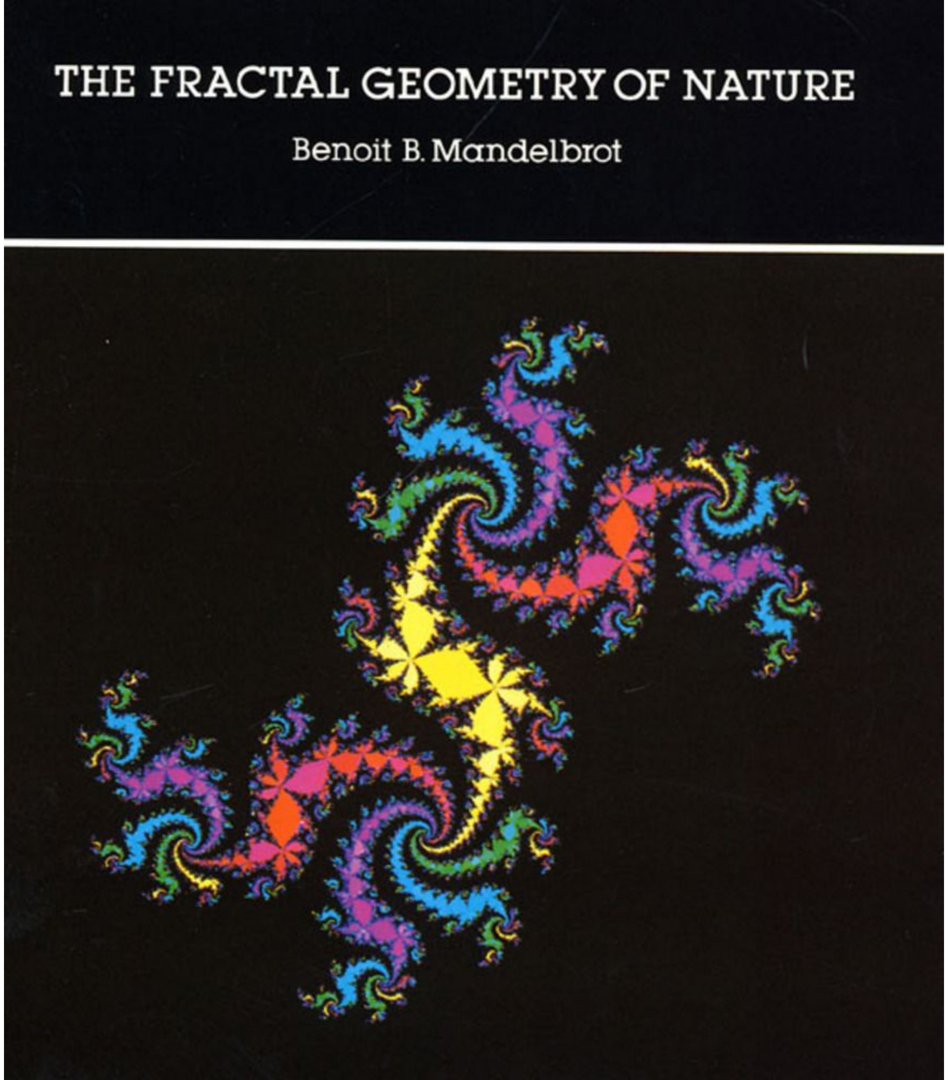
Nassim Nicholas Taleb

https://arxiv.org/abs/2001.10488

The monograph investigates the misapplication of conventional statistical techniques to fat tailed distributions and looks for remedies, when possible.
Switching from thin tailed to fat tailed distributions requires more than "changing the color of the dress". Traditional asymptotics deal mainly with either n=1 or $n = \infty$, and the real world is in between, under of the "laws of the medium numbers" --which vary widely across specific distributions. Both the law of large numbers and the generalized central limit mechanisms operate in highly idiosyncratic ways outside the standard Gaussian or Levy-Stable basins of convergence.
A few examples:
+ The sample mean is rarely in line with the population mean, with effect on "naive empiricism", but can be sometimes be estimated via parametric methods.
+ The "empirical distribution" is rarely empirical.
+ Parameter uncertainty has compounding effects on statistical metrics.
+ Dimension reduction (principal components) fails.
+ Inequality estimators (GINI or quantile contributions) are not additive and produce wrong results.
+ Many "biases" found in psychology become entirely rational under more sophisticated probability distributions
+ Most of the failures of financial economics, econometrics, and behavioral economics can be attributed to using the wrong distributions.
This book, the first volume of the Technical Incerto, weaves a narrative around published journal articles.

"Most of the failures of financial economics, econometrics, and behavioral economics can be attributed to using the wrong distributions."





tinyurl.com/longtail23

blog

website

github

downey@allendowney.com/blog

twitter

email

tinyurl.com/longtail23

Let me tell you about Long-Tailed World

tinyurl.com/longtail23

Distribution of height in Long-Tailed World

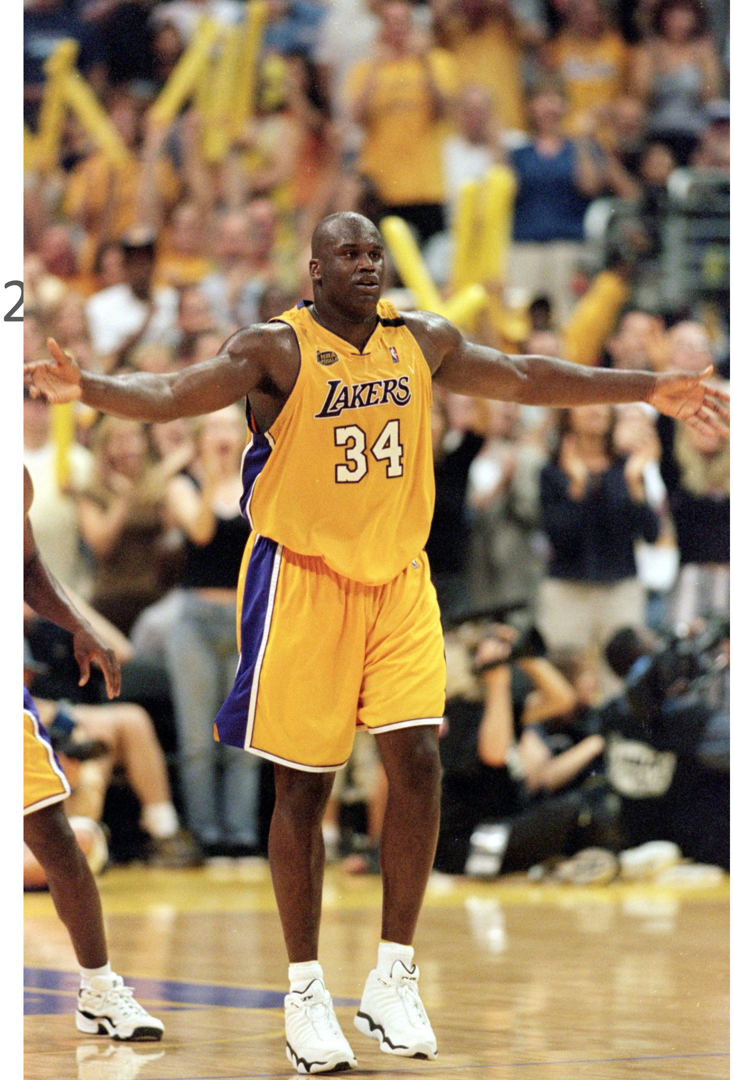
- 25th percentile 163 cm.
- 75th percentile 178 cm.

Like the distribution of height on Earth.

But with the tail of the distribution of disaster.

Of 10 people, the tallest might be 186 cm -- taller than a gorilla, shorter than me.

Of 100 people, the tallest might be 2



tinyurl.com/longtail23

Out of 10,000 people, the tallest w

Like the statues at
Stadio dei Marmi, Rome.

tinyurl.

Out of a million people in Long-ta

59 meters.

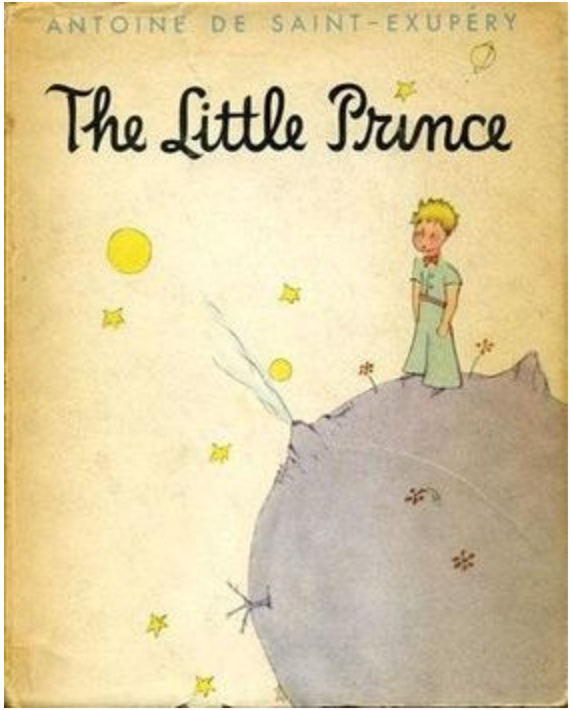
Like this statue of Guan Yu,

Jingzhou, China.

tinyurl.com/longtail23

In a country the size of the United States, the tallest person would be 160,000 kilometers tall.

About a third of the way to the moon.

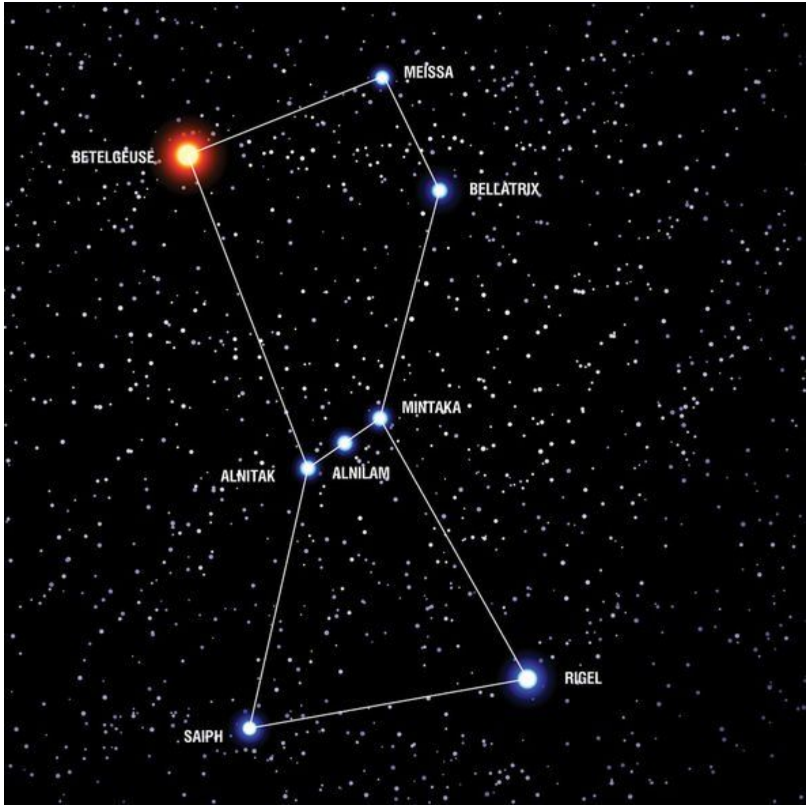


tinyurl.com/longtail23

And the tallest person in the world would be 14 quintillion kilometers.

About 1500 light years

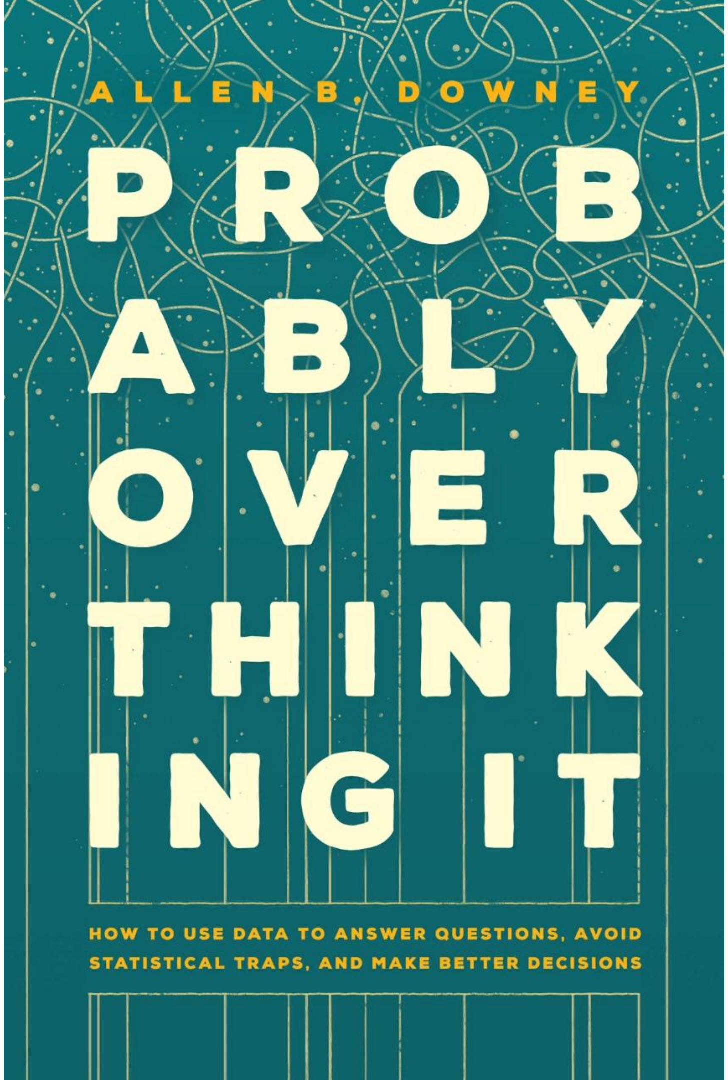
Three times the distance to Betelgeuse.



tinyurl.com/longtail23

This is the fourth stop on my stealth book tour.

tinyurl.com/longtail23

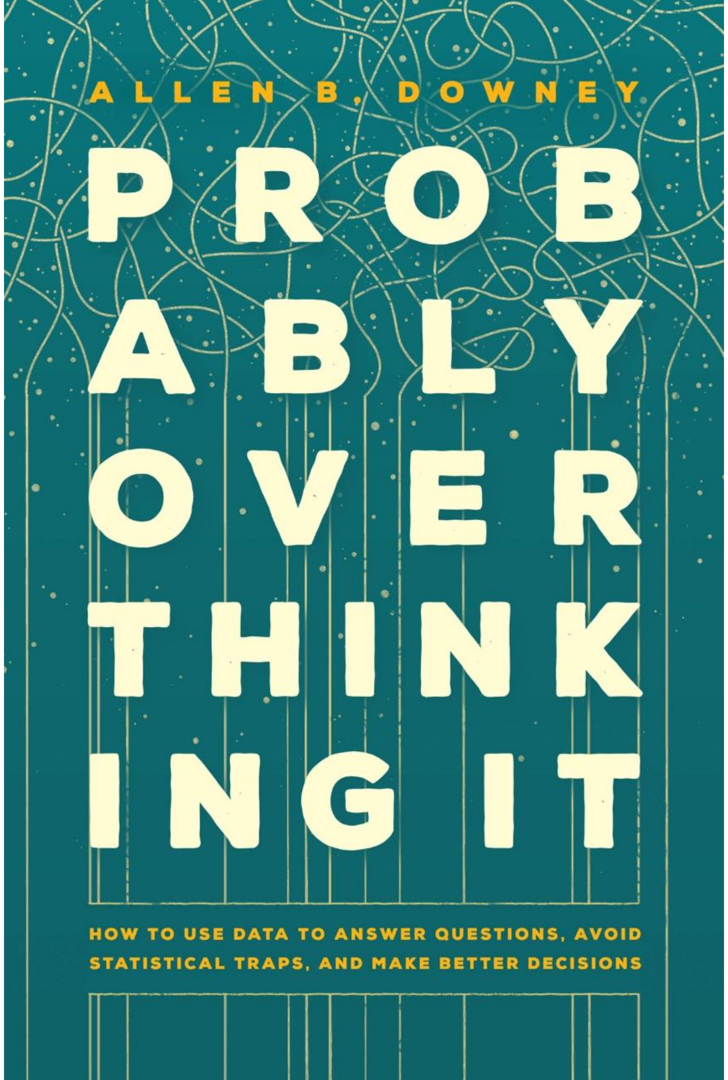


Contents

tinyurl.com/longtail23

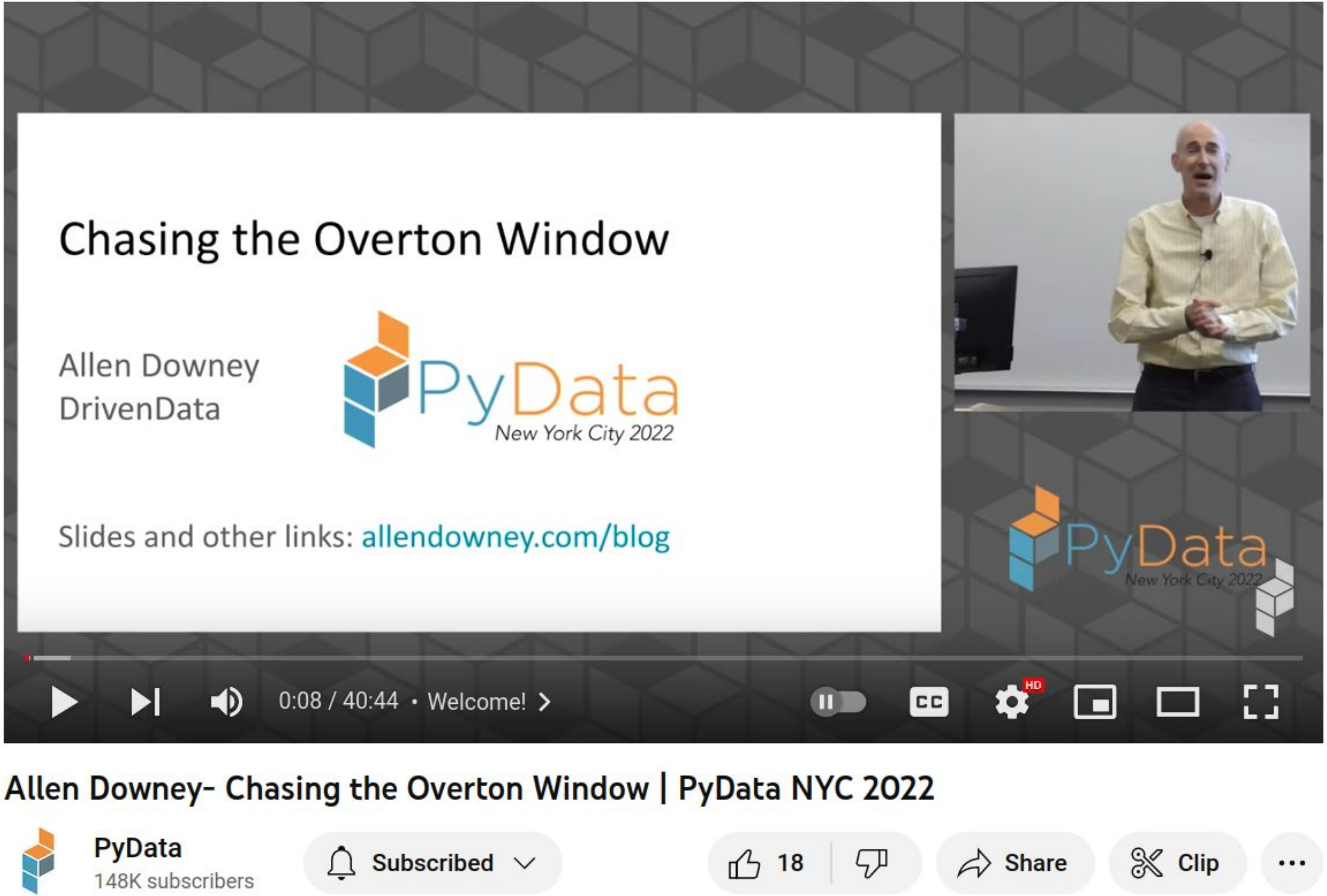
https://www.youtube.com/watch?v=cXWTHfvycyM

tinyurl.com/longtail23



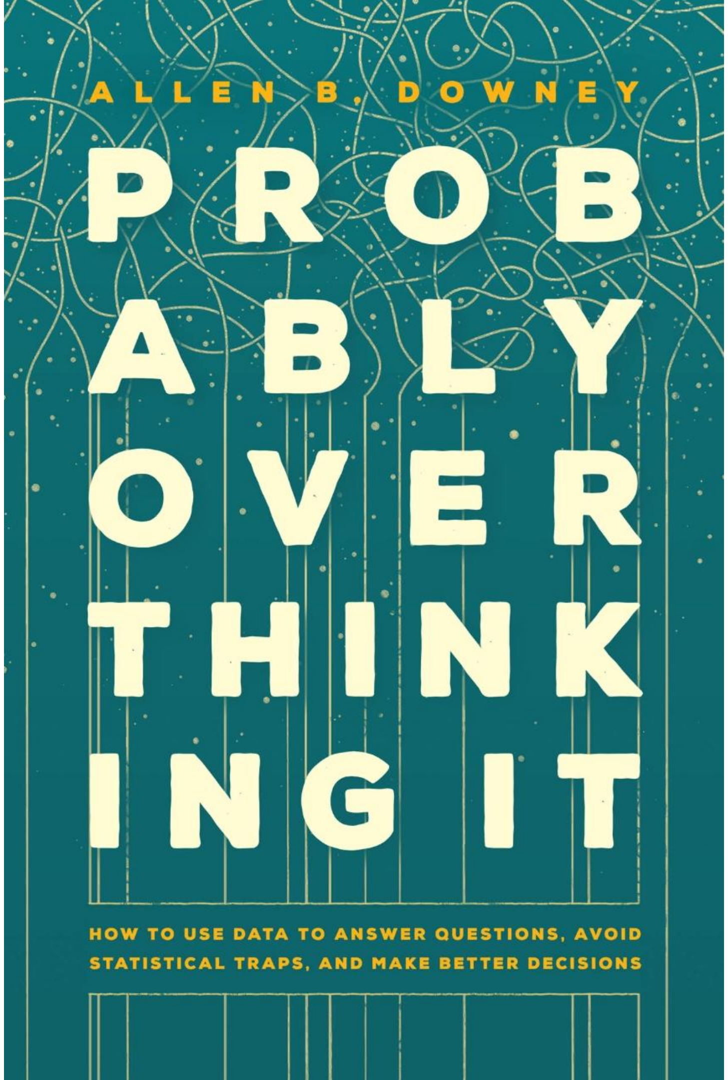
Contents

tinyurl.com/longtail23



https://www.youtube.com/watch?v=VpuWECpTxmM

tinyurl.com/longtail23



Contents

tinyurl.com/longtail23

Open Data Science Conference

BOOTCAMP | EAST | EUROPE | APAC | WEST | Ai+ | Blog | Partners | Jobs | Newsletter | Swag Store | About |

TRACKS BOOTCAMP ATTEND SCHEDULE ODSC REGISTER SPEAKERS AI EXPO VENUE INFO

PROBABLY OVERTHINKING IT

Data Science, Bayesian Statistics, And Other Ideas

CAUSATION, COLLISION, AND CONFUSION

May 10, 2023 AllenDowney

Today I presented a talk about Berkson's paradox at ODSC East 2023. If you missed it, the slides are here. When the video is available, I'll post it here.

> *Abstract: Collision bias is the most treacherous error in statistics: it can be subtle, it is easy to induce it by accident, and the error it causes can be bigger than the effect you are trying to measure. It is the cause of Berkson's paradox, the low birthweight paradox, and the obesity paradox, among other famous historical errors. And it might be the cause of your next blunder! Although it is best known in epidemiology, it*

ABOUT ME

Allen Downey is a curriculum designer at Brilliant, professor emeritus at Olin College, and author of *Think Python*, *Think Bayes*, and other books available from Green Tea Press.

I am working on a book, also called *Probably Overthinking It*, which is about using evidence and reason to answer questions and guide decision making. If you would like to get an occasional update about the

tinyurl.com/longtail23