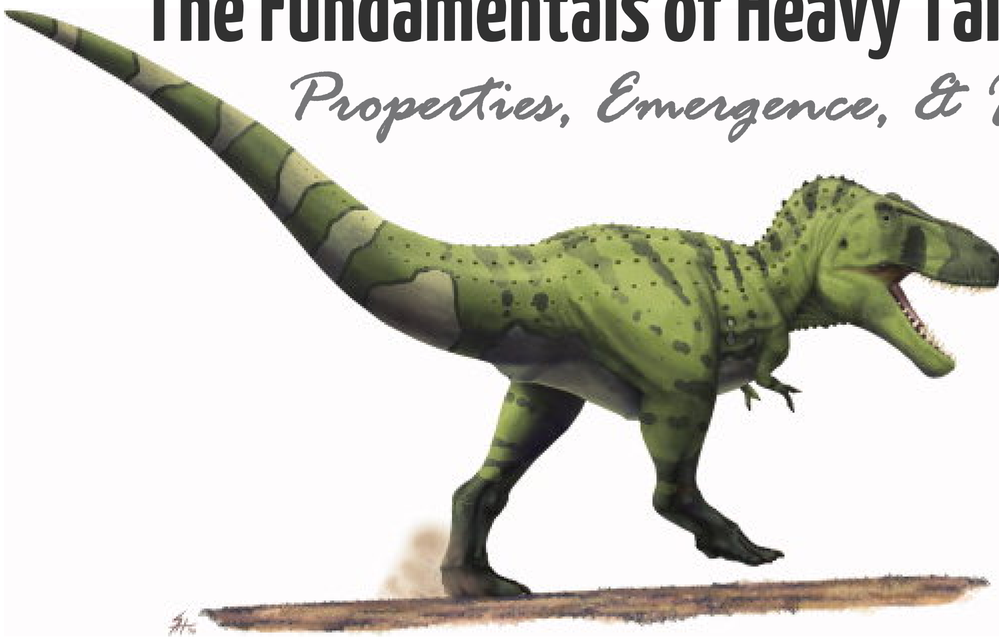
The Fundamentals of Heavy Tails

Properties, Emergence, & Identification



Jayakrishnan Nair, **Adam Wierman**, Bert Zwart

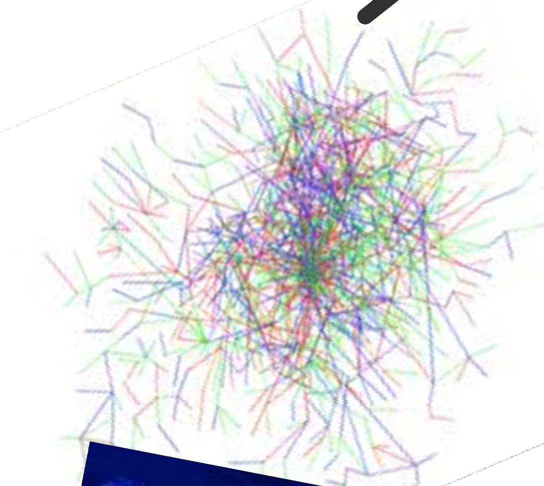
> "The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking […] as if anything but a nice bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world."
>
> – Clay Shirky, as quoted in Newsweek & the Guardian

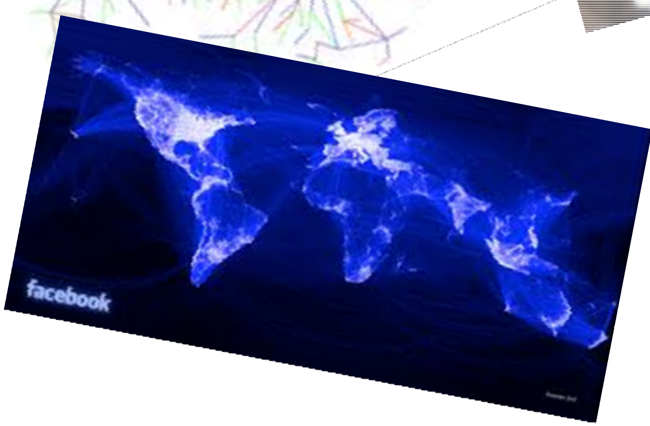
Why am I doing a tutorial on heavy tails?

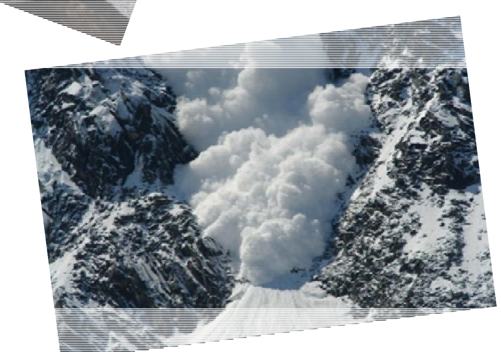
→ Because we're writing a book on the topic...

Why are we writing a book on the topic?

→ Because heavy-tailed phenomena are everywhere!







Why am I doing a tutorial on heavy tails?

Because we're writing a book on the topic...

Why are we writing a book on the topic?

Because heavy-tailed phenomena are everywhere!

BUT, they are extremely misunderstood.

> "The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking […] as if anything but a nice bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world."
>
> – Clay Shirky, as quoted in Newsweek & the Guardian

Heavy-tailed phenomena are treated as something

Mysterious, Surprising, & Controversial

> "The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking […] as if anything but a nice bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world."
>
> – Clay Shirky, as quoted in Newsweek & the Guardian

Heavy-tailed phenomena are treated as something

MYSTERIOUS, Surprising, & Controversial

Simple, appealing statistical approaches have BIG problems

Our intuition is flawed because intro probability classes focus on light-tailed distributions

Heavy-tailed phenomena are treated as something

MYSTERIOUS, Surprising, & Controversial

On Power-Law Relationships of the Internet Topology

Michalis Faloutsos
U.C. Riverside
Dept. of Comp. Science
michalis@cs.ucr.edu

Petros Faloutsos
U. of Toronto
Dept. of Comp. Science
pfal@cs.toronto.edu

Christos Faloutsos *
Carnegie Mellon Univ.
Dept. of Comp. Science
christos@cs.cmu.edu

1999 Sigcomm paper – 4500+ citations!

BUT...

2005, STOC

On the Bias of Traceroute Sampling

or, Power-law Degree Distributions in Regular Graphs

Dimitris Achlioptas
Microsoft Research
Microsoft Corporation
Redmond, WA 98052
optas@microsoft.com

Aaron Clauset
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
aaron@cs.unm.edu

David Kempe
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
dkempe@usc.edu

Cristopher Moore
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
moore@cs.unm.edu

IEEE/ACM TRANSACTIONS ON NET 1205

Understanding Internet Topology: Principles, Models, and Validation

David Alderson, *Member, IEEE*, Lun Li, *Student Member, IEEE*, Walter Willinger, *Fellow, IEEE*, and John C. Doyle, *Member, IEEE*

2005, ToN

Similar stories in electricity nets, citation nets, ...

Heavy-tailed phenomena are treated as something

~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

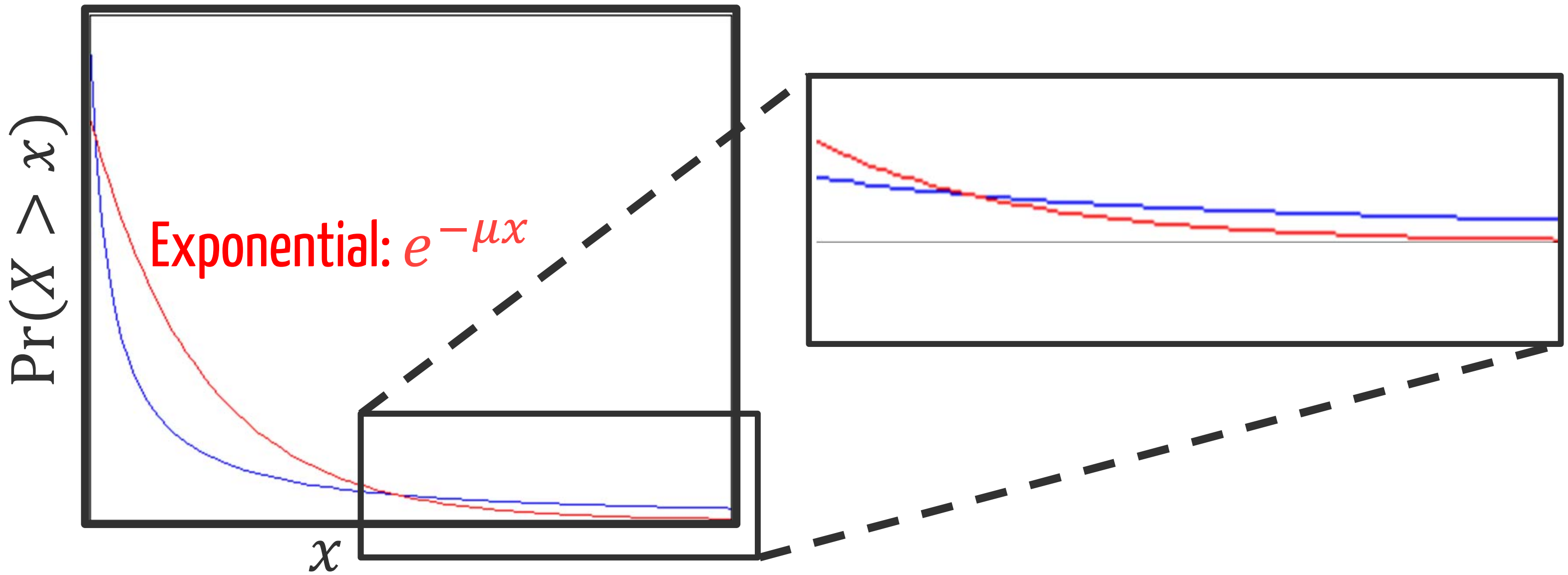
3. Identification

What is a heavy-tailed distribution?

A distribution with a “tail” that is “heavier” than an Exponential

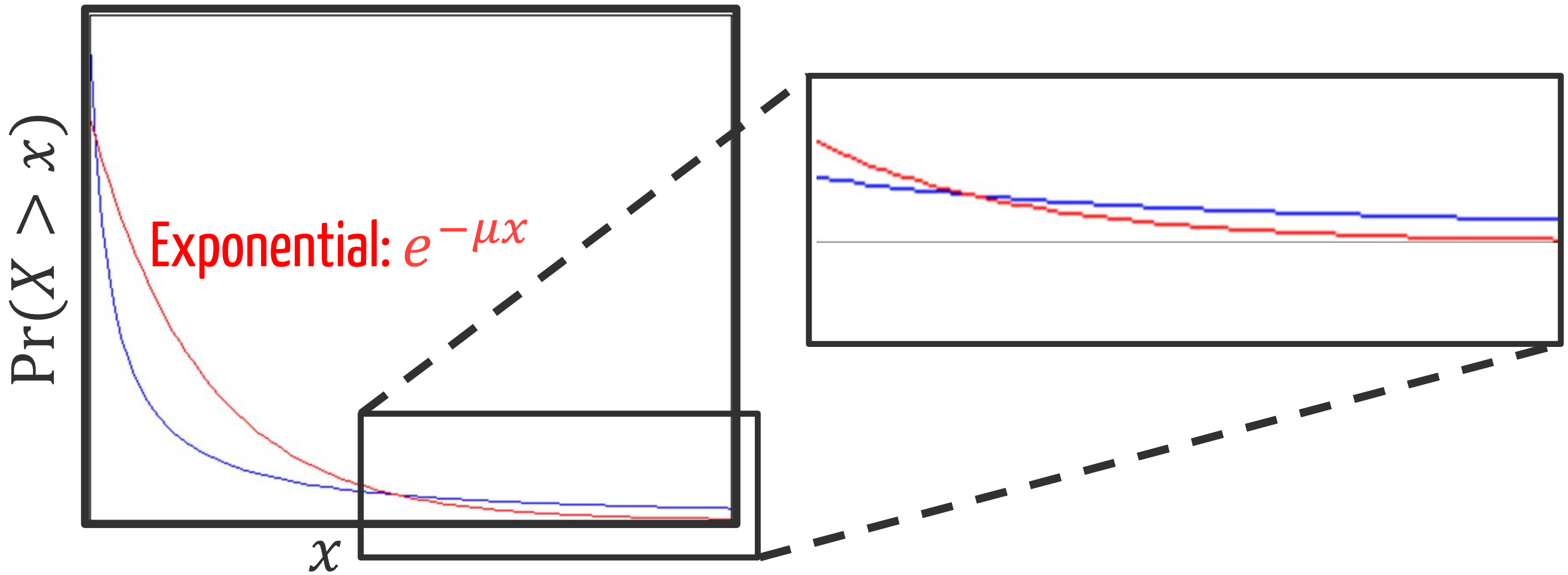
What is a heavy-tailed distribution?

A distribution with a "tail" that is "heavier" than an Exponential



What is a heavy-tailed distribution?

A distribution with a "tail" that is "heavier" than an Exponential



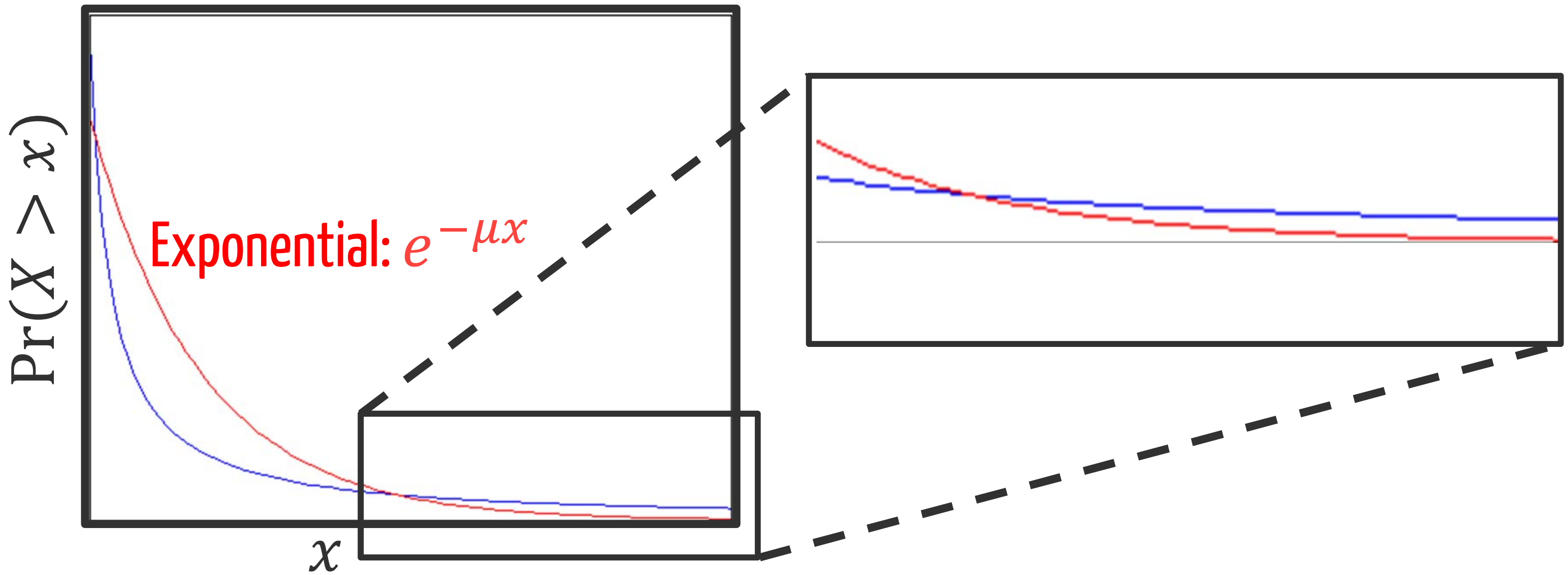
Canonical Example: The Pareto Distribution a.k.a. the "power-law" distribution

$$\Pr(X > x) = \bar{F}(x) = \left(\frac{x_{\min}}{x}\right)^{\alpha} \quad \text{for } x \geq x_{\min}$$

density: $f(x) = \dfrac{\alpha x_{\min}^{\alpha}}{x^{\alpha+1}}$

What is a heavy-tailed distribution?

A distribution with a "tail" that is "heavier" than an Exponential



Canonical Example: The Pareto Distribution a.k.a. the "power-law" distribution

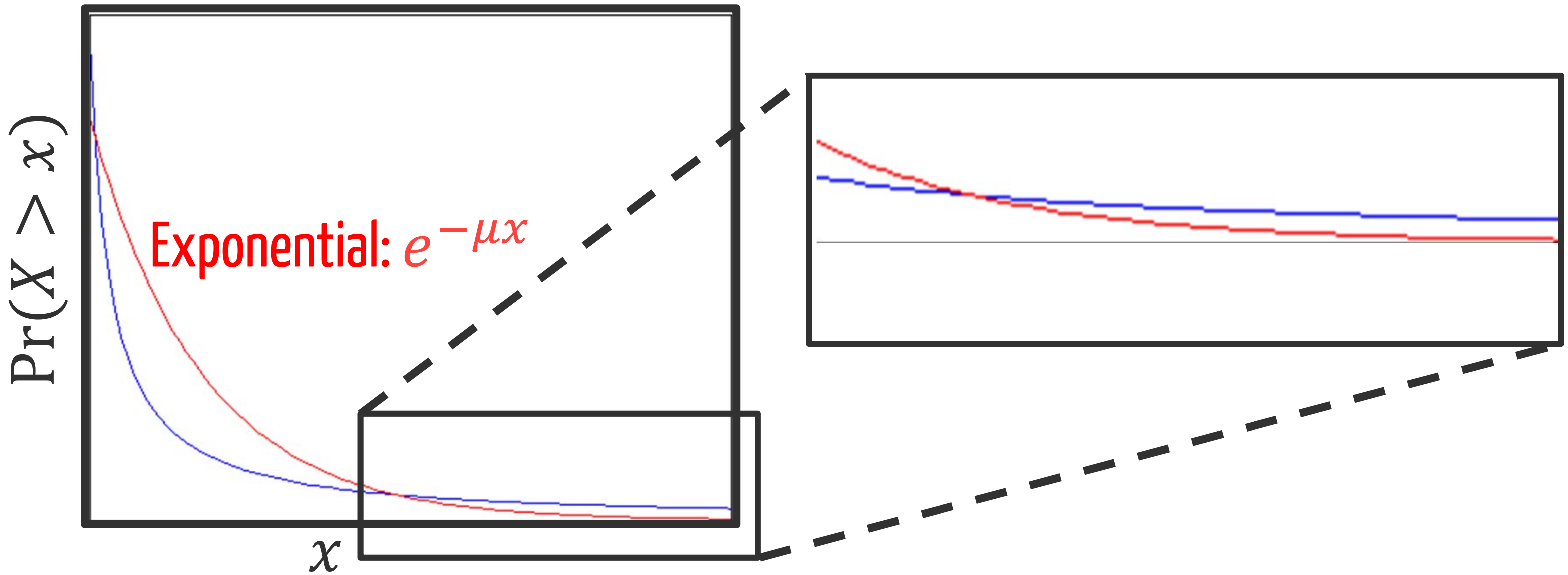
Many other examples: LogNormal, Weibull, Zipf, Cauchy, Student's t, Frechet, …

$X\colon \log X \sim Normal$

$\bar{F}(x) = e^{-(x/\lambda)^k}$

What is a heavy-tailed distribution?

A distribution with a "tail" that is "heavier" than an Exponential



Canonical Example: The Pareto Distribution a.k.a. the "power-law" distribution

Many other examples: LogNormal, Weibull, Zipf, Cauchy, Student's t, Frechet, …

Many subclasses: Regularly varying, Subexponential, Long-tailed, Fat-tailed, …

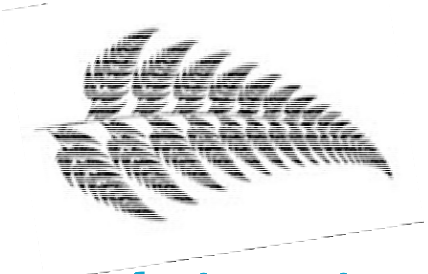
Heavy-tailed distributions have many strange & beautiful properties

- The "Pareto principle": 80% of the wealth owned by 20% of the population, etc.
- Infinite variance or even infinite mean
- Events that are much larger than the mean happen "frequently"

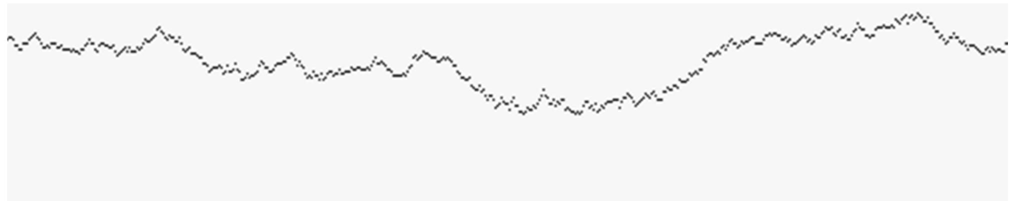
....

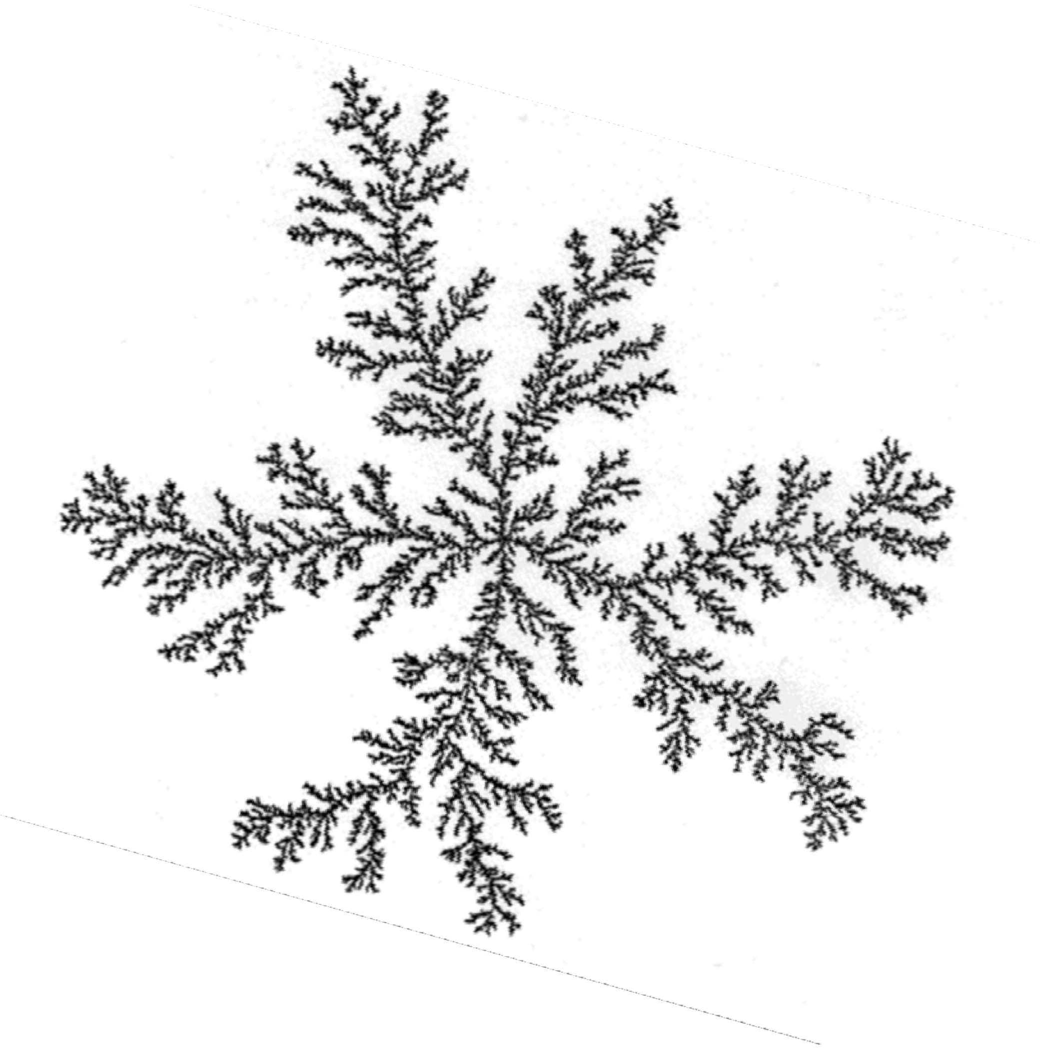
These are driven by 3 "defining" properties

1) Scale invariance
2) The "catastrophe principle"
3) The residual life "blows up"



Scale invariance





Scale invariance

F is scale invariant if there exists an x_0 and a g such that $\boxed{\bar{F}(\lambda x) = g(\lambda)\bar{F}(x)}$ for all λ, x such that $\lambda x \geq x_0$.

"change of scale"

Scale invariance

F is scale invariant if there exists an x_0 and a g such that
$\bar{F}(\lambda x) = g(\lambda)\bar{F}(x)$ for all λ, x such that $\lambda x \geq x_0$.

$\boxed{\text{Theorem: A distribution is scale invariant if and only if it is Pareto.}}$

Example: Pareto distributions

$$\bar{F}(\lambda x) = \left(\frac{x_{\min}}{\lambda x}\right)^{\alpha} = \bar{F}(x)\left(\frac{1}{\lambda}\right)^{\alpha}$$

Scale invariance

F is scale invariant if there exists an x_0 and a g such that
$\bar{F}(\lambda x) = g(\lambda)\bar{F}(x)$ for all λ, x such that $\lambda x \geq x_0$.

$\Downarrow$

Asymptotic scale invariance

F is asymptotically scale invariant if there exists a continuous, finite g such that
$\lim_{x \to \infty} \frac{\bar{F}(\lambda x)}{\bar{F}(x)} = g(\lambda)$ for all λ.

Example: Regularly varying distributions

F is regularly varying if $\bar{F}(x) = x^{-\rho} L(x)$, where $L(x)$ is slowly varying,

i.e., $\lim_{x \to \infty} \frac{L(xy)}{L(x)} = 1$ for all $y > 0$.

Theorem: A distribution is asymptotically scale invariant iff it is regularly varying.

Asymptotic scale invariance

F is asymptotically scale invariant if there exists a continuous, finite g such that

$\lim_{x \to \infty} \frac{\bar{F}(\lambda x)}{\bar{F}(x)} = g(\lambda)$ for all λ.

Example: Regularly varying distributions

F is regularly varying if $\bar{F}(x) = x^{-\rho} L(x)$, where $L(x)$ is slowly varying, i.e., $\lim_{x\to\infty} \frac{L(xy)}{L(x)} = 1$ for all $y > 0$.

Regularly varying distributions are extremely useful. They basically behave like Pareto distributions with respect to the tail:
→ "Karamata" theorems
→ "Tauberian" theorems

Heavy-tailed distributions have many strange & beautiful properties

- The "Pareto principle": 80% of the wealth owned by 20% of the population, etc.
- Infinite variance or even infinite mean
- Events that are much larger than the mean happen "frequently"

....

These are driven by 3 "defining" properties

1) Scale invariance
2) The "catastrophe principle"
3) The residual life "blows up"

A thought experiment

During lecture I polled my 50 students about their heights and the number of twitter followers they have...

The sum of the heights was ~300 feet.
The sum of the number of twitter followers was 1,025,000

What led to these large values?

A thought experiment

During lecture I polled my 50 students about their heights and the number of twitter followers they have...

The sum of the heights was ~300 feet.
The sum of the number of twitter followers was 1,025,000

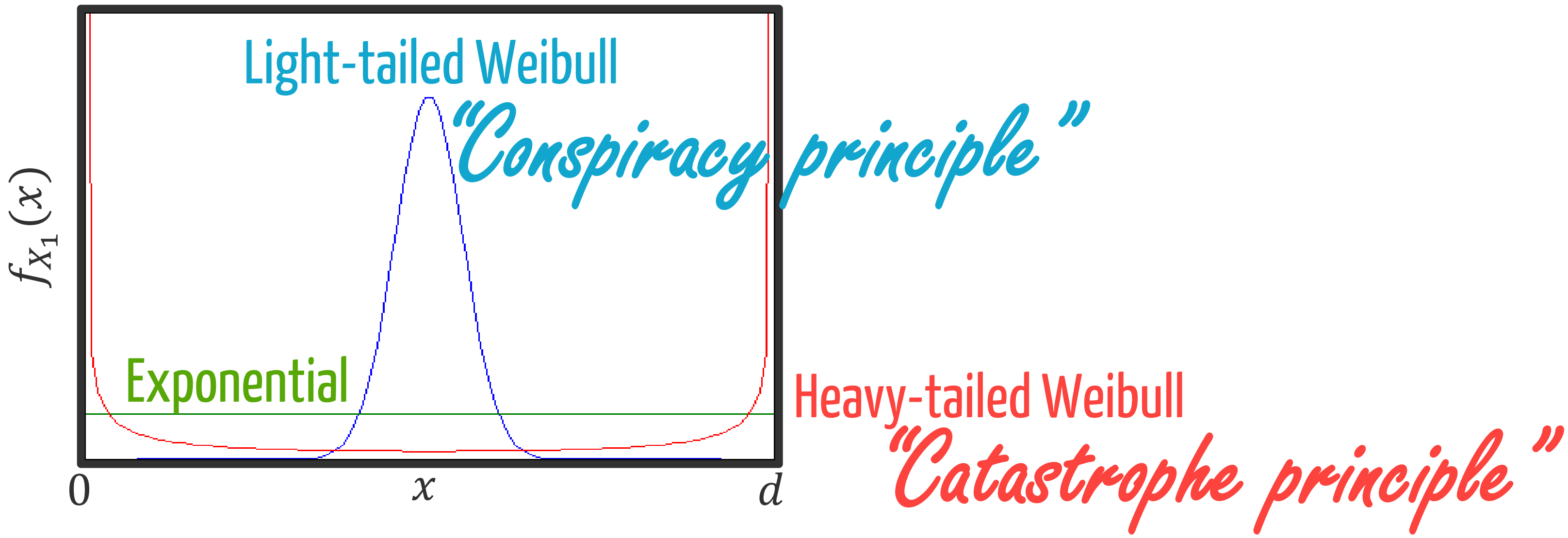
A bunch of people were probably just over 6' tall
(Maybe the basketball teams were in the class.)
"Conspiracy principle"

One person was probably a twitter celebrity and had ~1 million followers.
"Catastrophe principle"

Example

Consider $X_1 + X_2$ i.i.d Weibull.

Given $X_1 + X_2 = d$, what is the marginal density of X_1?



"Catastrophe principle"

$$\Pr(\max(X_1,\dots,X_n) > t) \sim \Pr(X_1 + \cdots + X_n > t)$$

$$\Rightarrow \Pr(\max(X_1,\dots,X_n) > t \mid X_1 + \cdots + X_n > t) \to 1$$

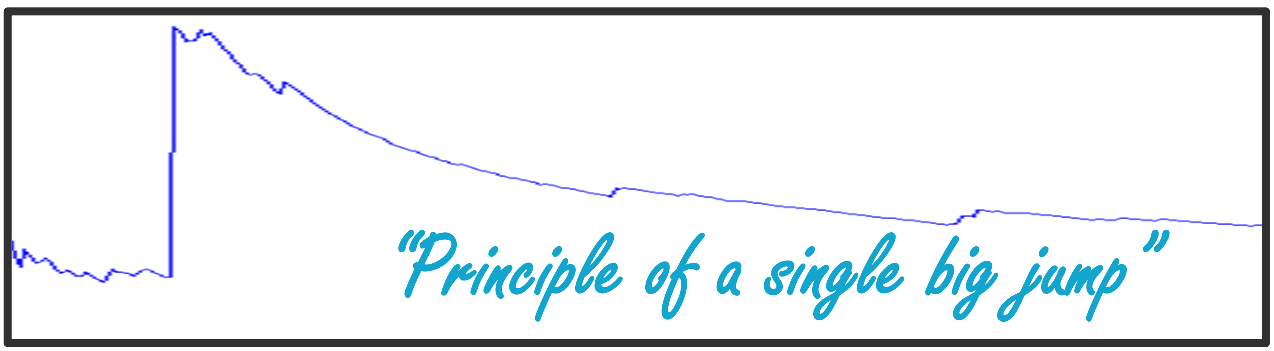
"Conspiracy principle"

$$\Pr(\max(X_1,\dots,X_n) > t) = o(\Pr(X_1 + \cdots + X_n > t))$$

"Catastrophe principle"

$$\Pr(\max(X_1,\ldots,X_n) > t) \sim \Pr(X_1 + \cdots + X_n > t)$$
$$\Rightarrow \Pr(\max(X_1,\ldots,X_n) > t \mid X_1 + \cdots + X_n > t) \to 1$$

Extremely useful for random walks, queues, etc.



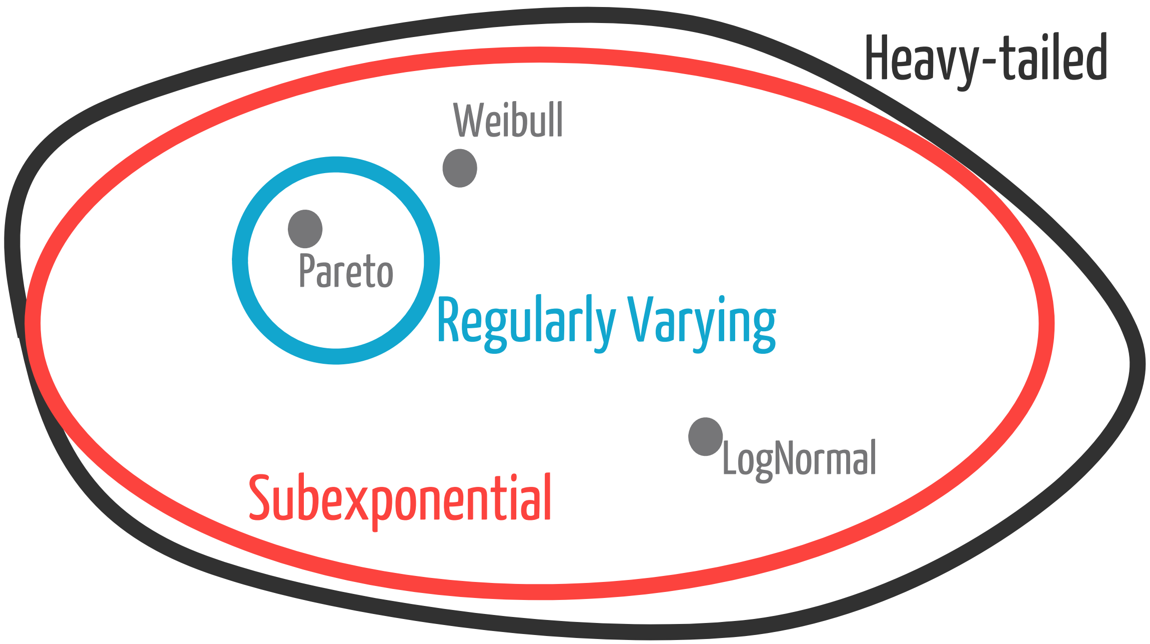
"Catastrophe principle"

$$\Pr(\max(X_1, \ldots, X_n) > t) \sim \Pr(X_1 + \cdots + X_n > t)$$
$$\Rightarrow \Pr(\max(X_1, \ldots, X_n) > t | X_1 + \cdots + X_n > t) \to 1$$

$$\Updownarrow$$

Subexponential distributions

F is subexponential if for i.i.d. X_i, $\Pr(X_1 + \cdots + X_n > t) \sim nPr(X_1 > t)$



Subexponential distributions

F is subexponential if for i.i.d. X_i, $\Pr(X_1 + \cdots + X_n > t) \sim nPr(X_1 > t)$

Heavy-tailed distributions have many strange & beautiful properties

- The "Pareto principle": 80% of the wealth owned by 20% of the population, etc.
- Infinite variance or even infinite mean
- Events that are much larger than the mean happen "frequently"

....

These are driven by 3 "defining" properties

1) Scale invariance
2) The "catastrophe principle"
3) The residual life "blows up"

A thought experiment

What happens to the expected $\underbrace{\text{remaining waiting time}}_{}$ as we wait

residual life

...for a table at a restaurant?

...for a bus?

...for the response to an email?

The remaining wait drops as you wait

If you don't get it quickly, you never will...

The distribution of residual life

The distribution of remaining waiting time given you have already waited x time is $\bar{R}_x(t) = \frac{\bar{F}(x+t)}{\bar{F}(x)}$.

Examples:

Exponential : $\bar{R}_x(t) = \frac{e^{-\mu(x+t)}}{e^{-\mu x}} = e^{-\mu t}$ ⟶ "memoryless"

Pareto : $\bar{R}_x(t) = \frac{\left(\frac{x_{\min}}{x+t}\right)^{\alpha}}{\left(\frac{x_{\min}}{x}\right)^{\alpha}} = \left(1 + \frac{t}{x}\right)^{-\alpha}$ ⟶ Increasing in x

The distribution of residual life

The distribution of remaining waiting time given you have already waited x time is $\bar{R}_x(t) = \frac{\bar{F}(x+t)}{\bar{F}(x)}$.

Mean residual life

$$m(x) = E[X - x | X > x] = \int \bar{R}_x(t)dt$$

Hazard rate

$$q(x) = \frac{f(x)}{\bar{F}(x)} = \bar{R}'_x(0)$$

<u>Heavy-tailed distributions</u> "tend" to have decreasing hazard rates & increasing mean residual lives

<u>Light-tailed distributions</u> "tend" to have increasing hazard rates & decreasing mean residual lives

What happens to the expected remaining waiting time as we wait

...for a table at a restaurant?

...for a bus?

...for the response to an email?

BUT: not all heavy-tailed distributions have DHR / IMRL

some light-tailed distributions are DHR / IMRL

Heavy-tailed distributions "tend" to have decreasing hazard rates & increasing mean residual lives

Light-tailed distributions "tend" to have increasing hazard rates & decreasing mean residual lives

Long-tailed distributions

F is long-tailed if $\lim_{x\to\infty} \bar{R}_x(t) = \lim_{x\to\infty} \frac{\bar{F}(x+t)}{\bar{F}(x)} = 1$ for all t

<u>BUT</u>: not all heavy-tailed distributions have DHR / IMRL

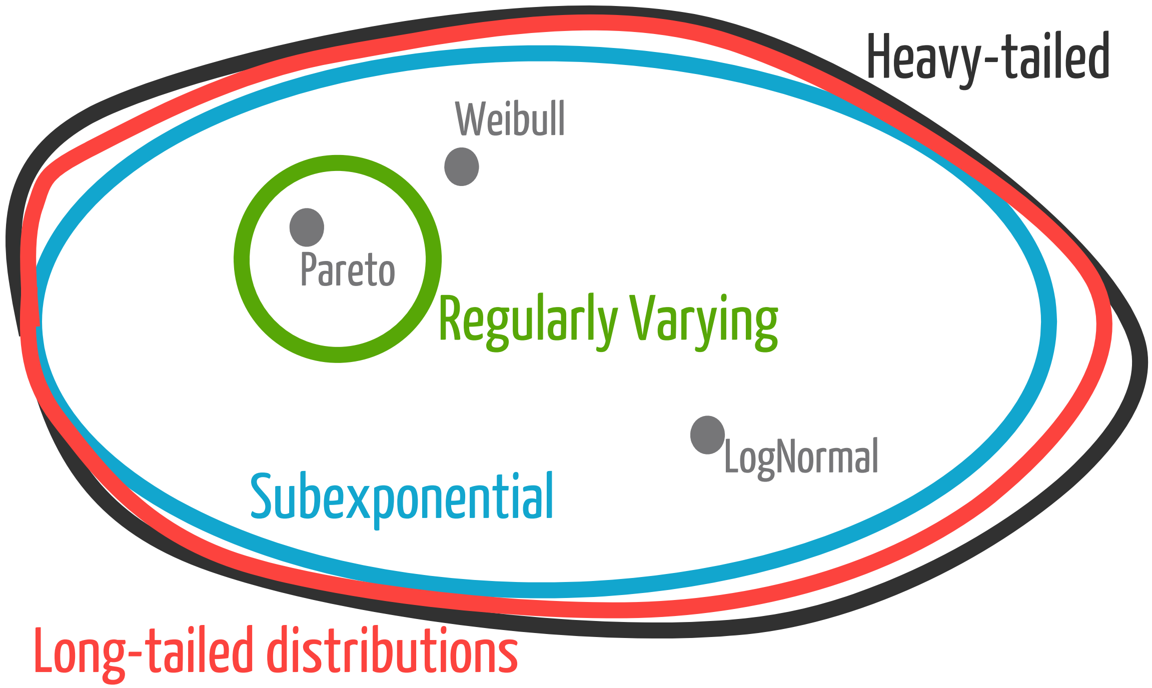
some light-tailed distributions are DHR / IMRL

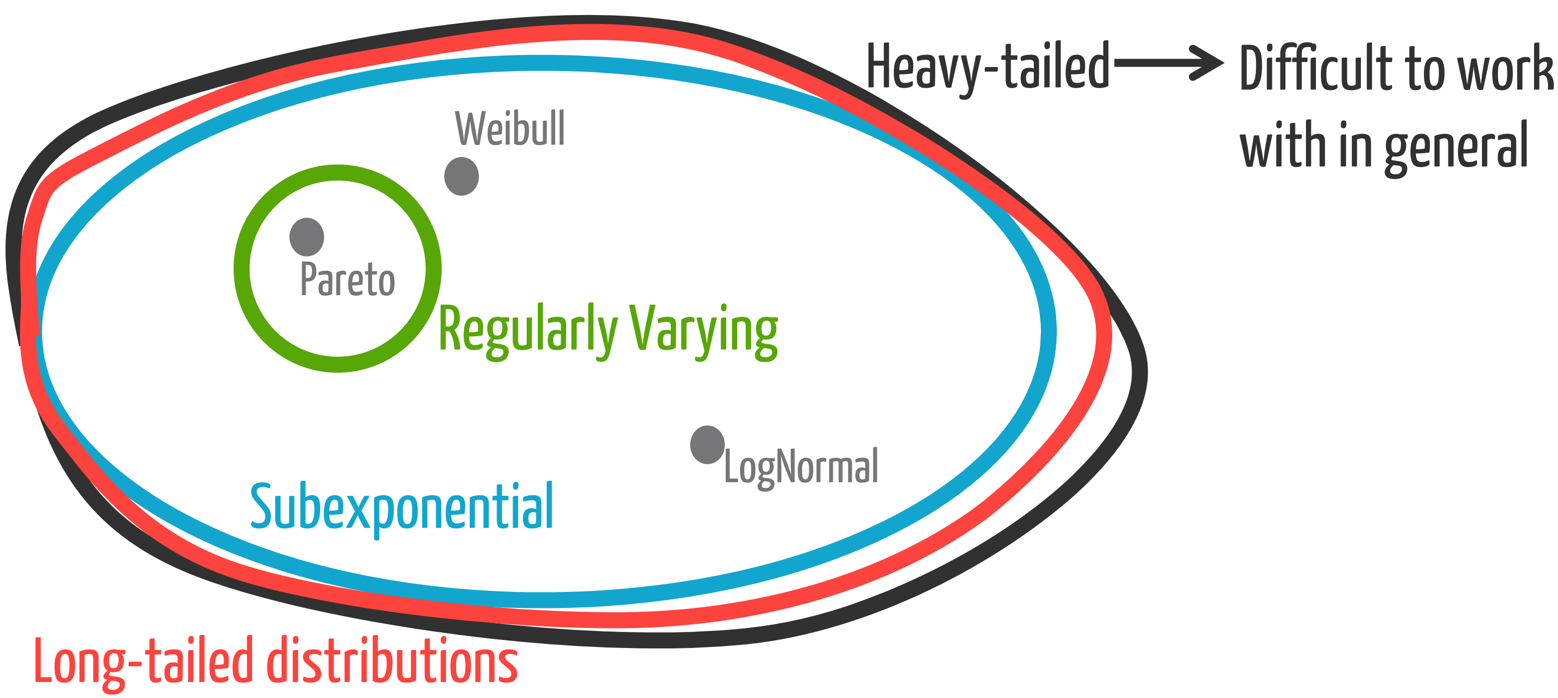
<u>Heavy-tailed distributions</u> “tend” to have decreasing hazard rates & increasing mean residual lives

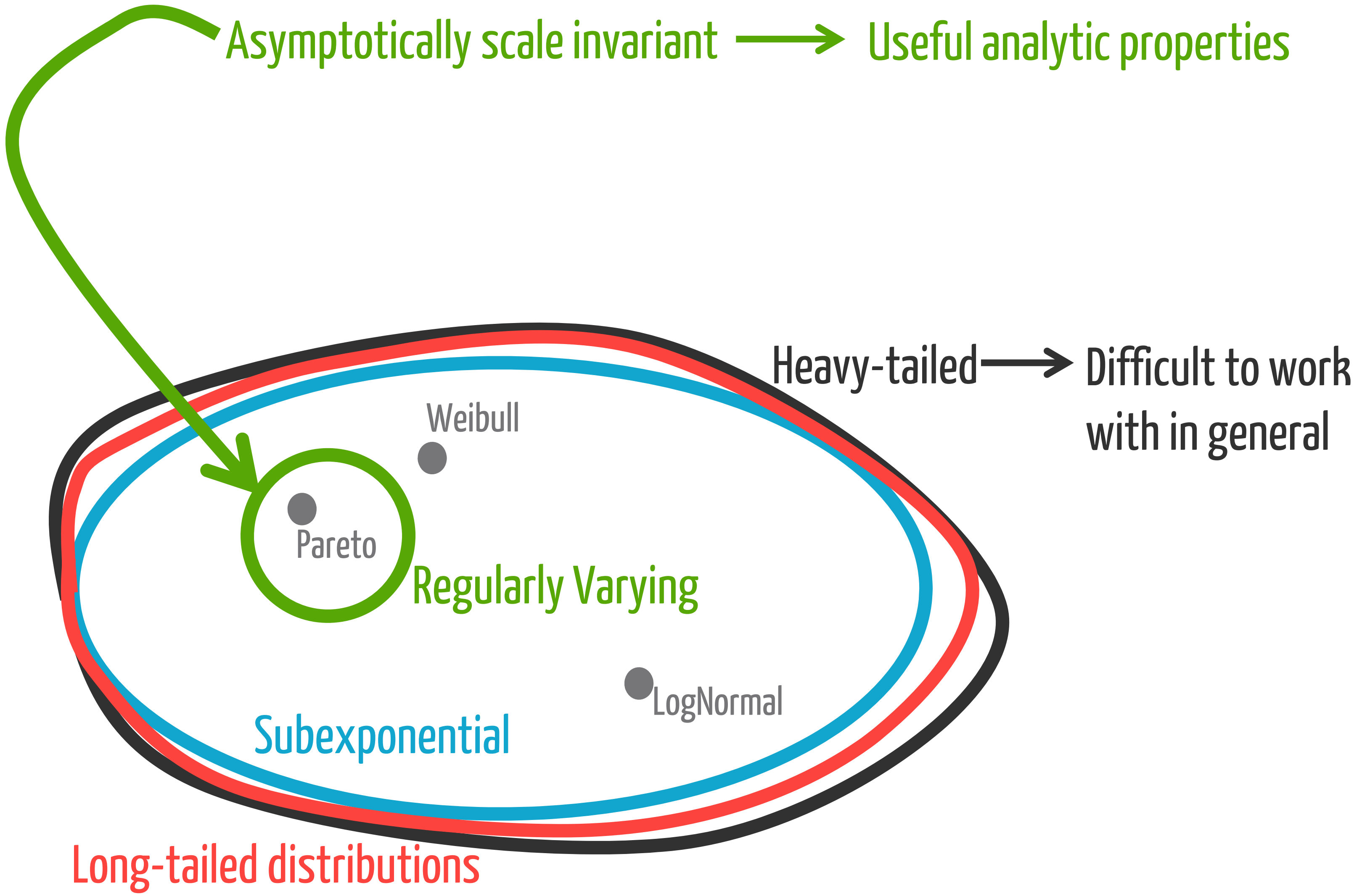
<u>Light-tailed distributions</u> “tend” to have increasing hazard rates & decreasing mean residual lives

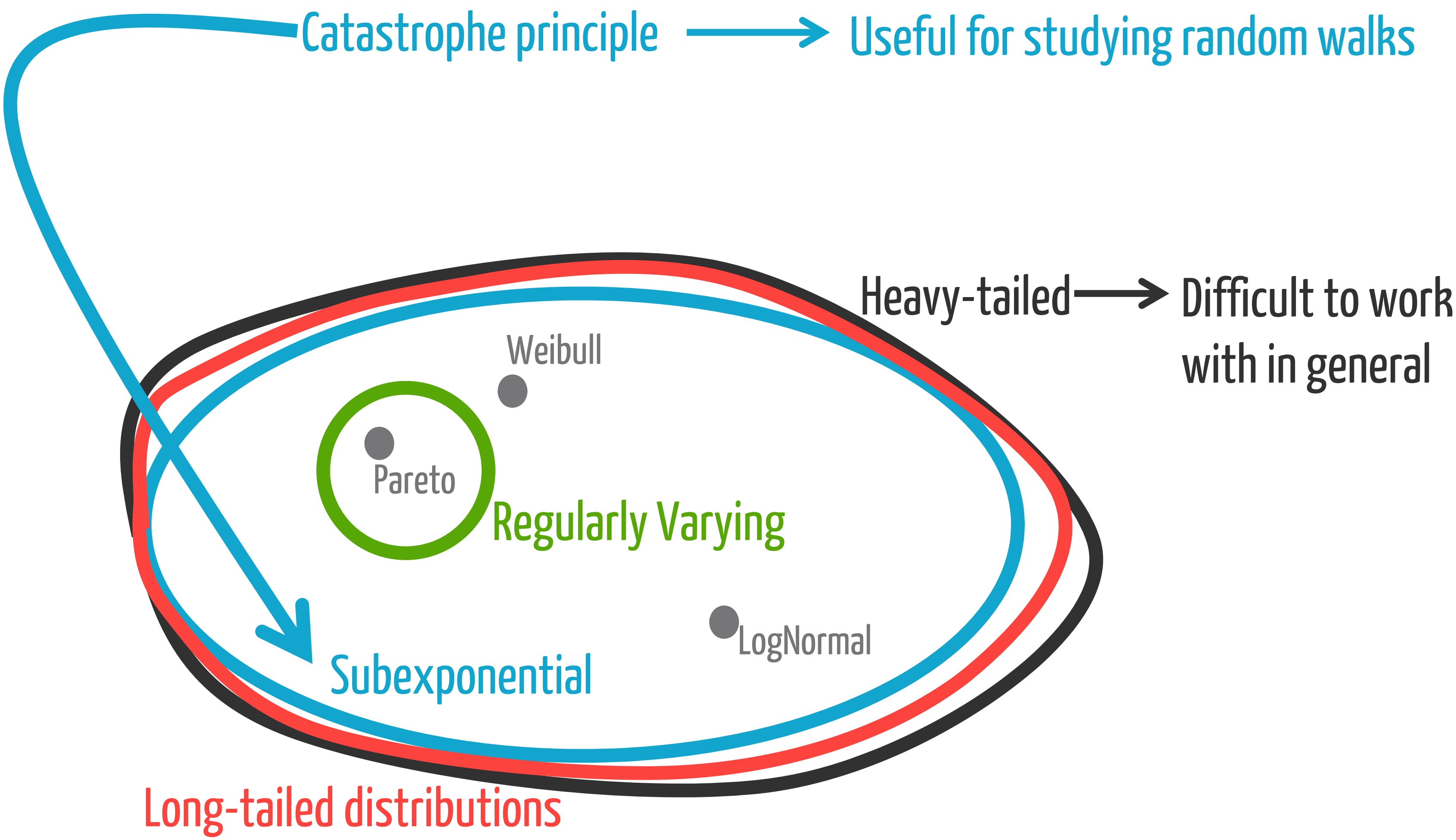
Long-tailed distributions

F is long-tailed if $\lim_{x\to\infty} \bar{R}_x(t) = \lim_{x\to\infty} \frac{\bar{F}(x+t)}{\bar{F}(x)} = 1$ for all t

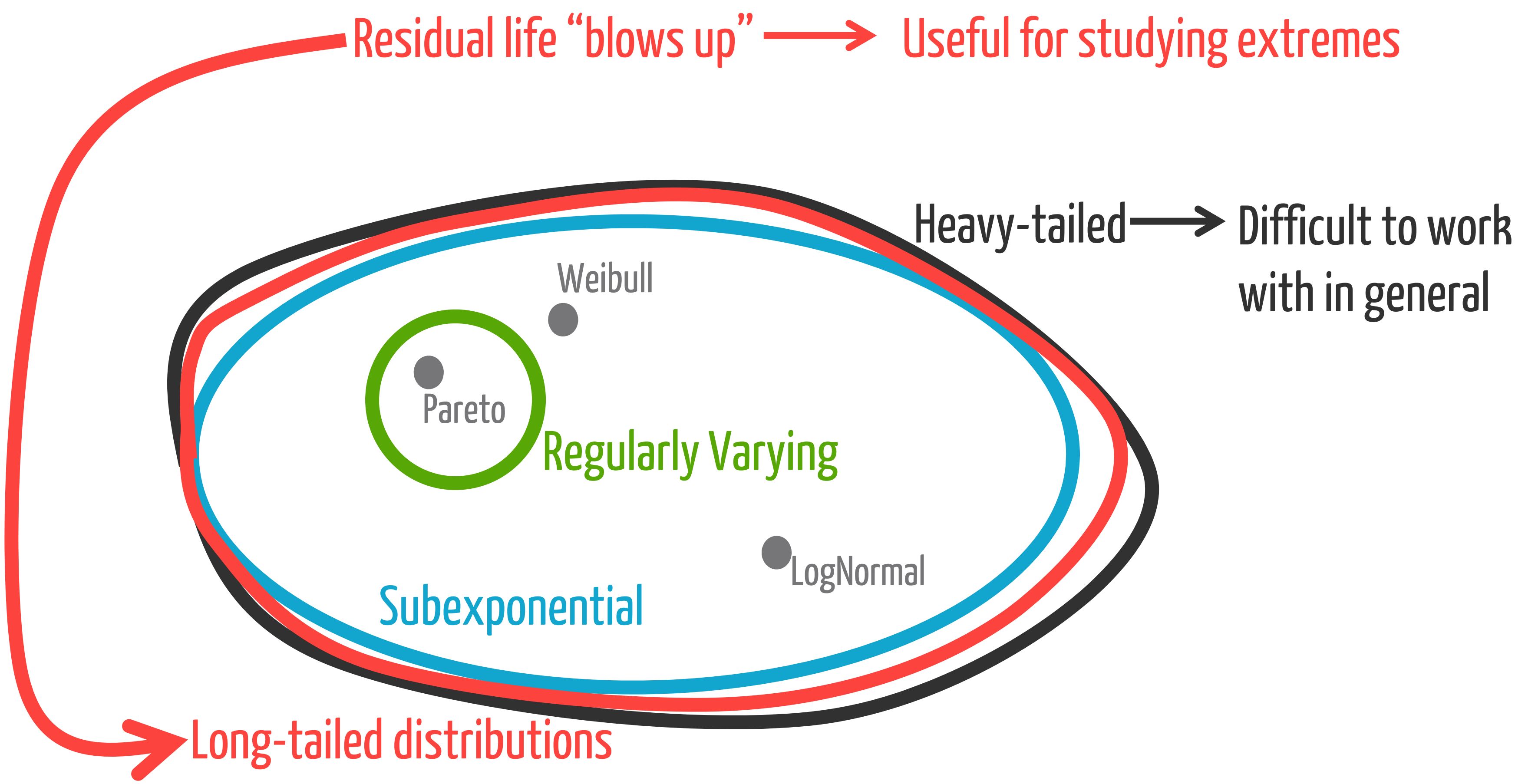



Heavy-tailed
Difficult to work with in general
Weibull
Pareto
Regularly Varying
LogNormal
Subexponential
Long-tailed distributions


Asymptotically scale invariant
Useful analytic properties
Heavy-tailed
Difficult to work with in general
Weibull
Pareto
Regularly Varying
LogNormal
Subexponential
Long-tailed distributions



Catastrophe principle
Useful for studying random walks
Heavy-tailed
Difficult to work with in general
Weibull
Pareto
Regularly Varying
LogNormal
Subexponential
Long-tailed distributions


Residual life "blows up"
Useful for studying extremes
Heavy-tailed
Difficult to work with in general
Weibull
Pareto
Regularly Varying
LogNormal
Subexponential
Long-tailed distributions

Heavy-tailed phenomena are treated as something

~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~

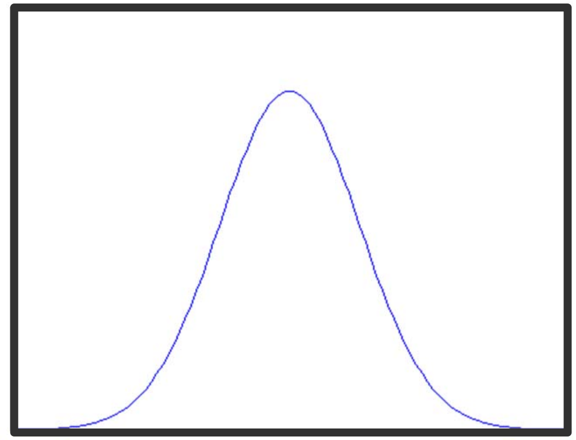
1. Properties

2. Emergence

3. Identification

We've all been taught that the Normal is "normal"

...because of the Central Limit Theorem



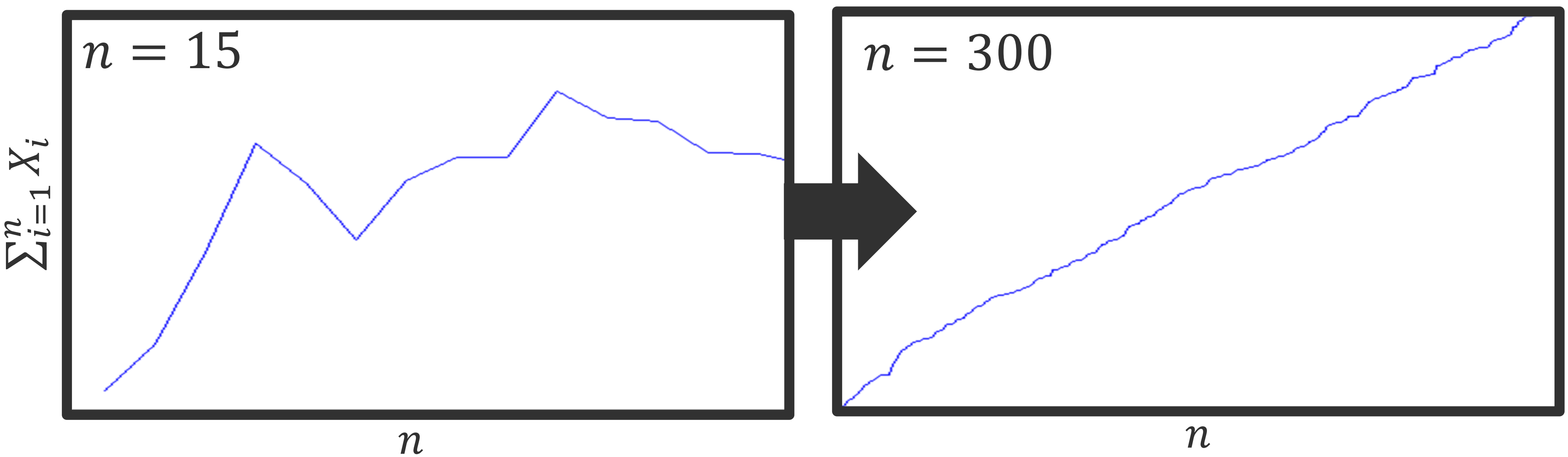
But the Central Limit Theorem we're taught is not complete!

A quick review

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Law of Large Numbers (LLN): $\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X_i]\ a.s.$ when $E[X_i] < \infty$

$$\sum_{i=1}^{n} X_i = nE[X_i] + o(n)$$



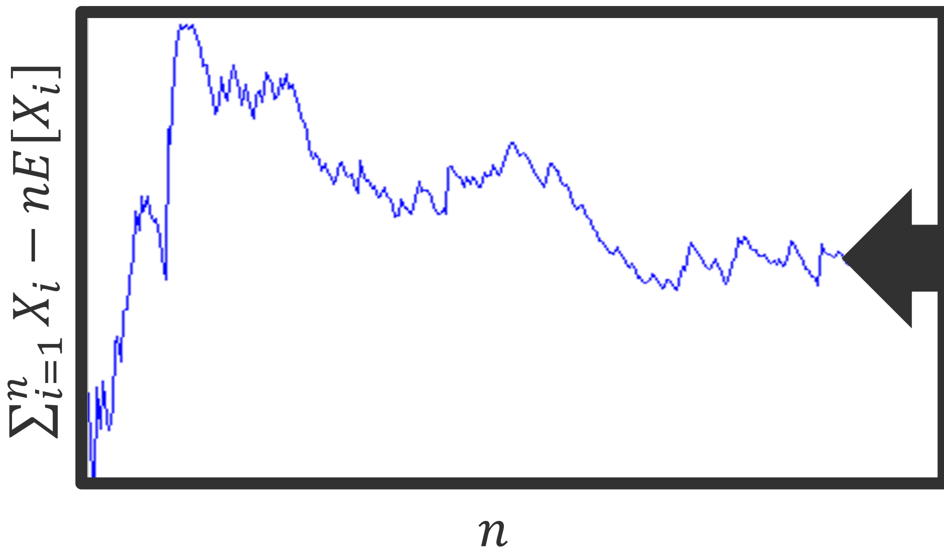
A quick review

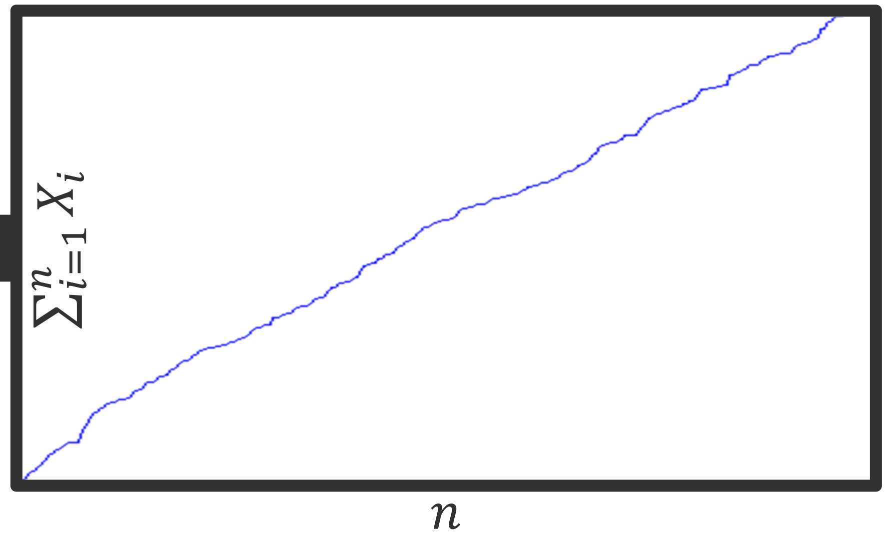
Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT): $\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - nE[X_i]\right) \to Z \sim Normal(0, \sigma^2)$

when $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

$$\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$$





A quick review

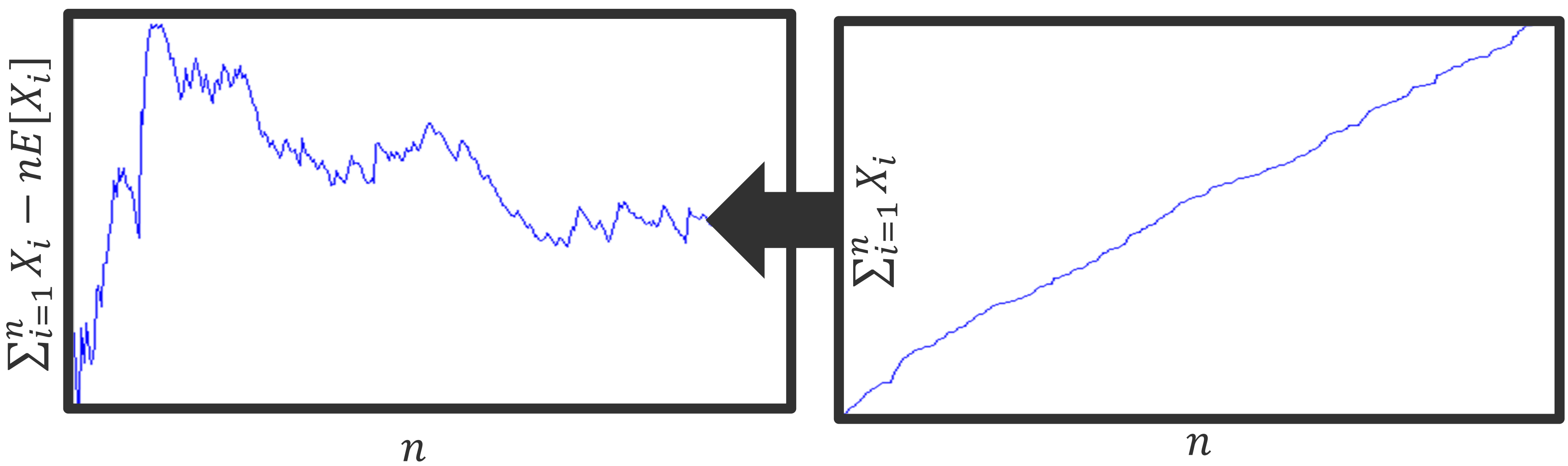
Two key assumptions

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT): $\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - nE[X_i]\right) \to Z \sim Normal(0, \sigma^2)$

when $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

$$\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$$



A quick review

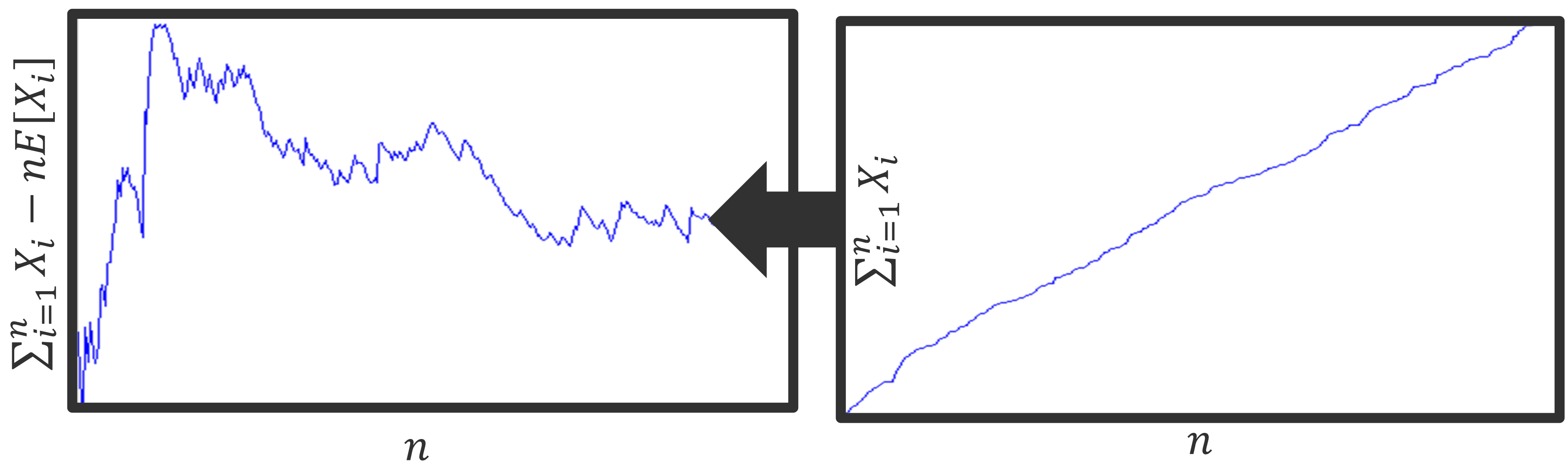
What if $Var[X_i] = \infty$?

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT): $\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - nE[X_i]\right) \to Z \sim Normal(0, \sigma^2)$

when $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

$$\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$$



A quick review

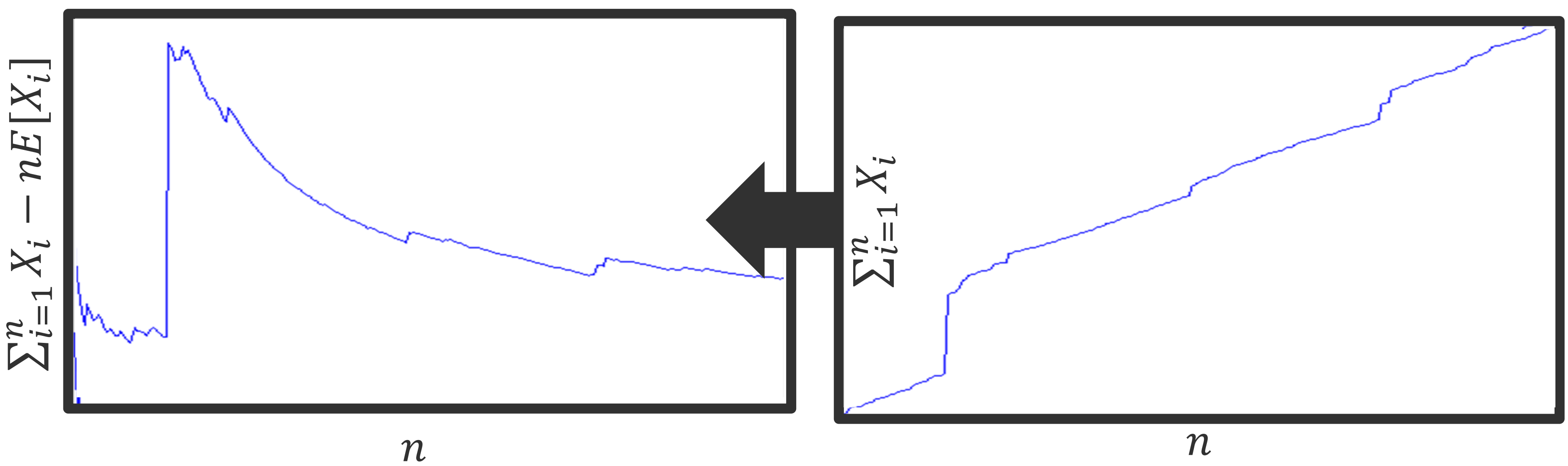
What if $Var[X_i] = \infty$?

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT): $\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - nE[X_i]\right) \to Z \sim Normal(0, \sigma^2)$

when $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

$$\sum_{i=1}^{n} X_i = nE[X_i] + \sqrt{n}Z + o(\sqrt{n})$$



A quick review

What if $Var[X_i] = \infty$?

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT): $\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - nE[X_i]\right) \to Z \sim Normal(0,\sigma^2)$

when $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

The Generalized Central Limit Theorem (GCLT):

$$\frac{1}{a_n}\left(\sum_{i=1}^{n} X_i - b_n\right) \to Z \begin{cases} Normal(0,\sigma^2) \\ Regularly\ varying\ \alpha \in (0,2) \end{cases}$$

$$\sum_{i=1}^{n} X_i = nE[X_i] + n^{1/\alpha} Z + o(n^{1/\alpha})$$

A quick review

What if $Var[X_i] = \infty$?

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Central Limit Theorem (CLT): $\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - nE[X_i]\right) \to Z \sim Normal(0, \sigma^2)$

when $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

The Generalized Central Limit Theorem (GCLT):

$$\frac{1}{a_n}\left(\sum_{i=1}^{n} X_i - b_n\right) \to Z \begin{cases} Normal(0, \sigma^2) \\ Regularly\ varying\ \alpha \in (0,2) \end{cases}$$

Finite variance → Light-tailed (Normal)
Infinite variance → Heavy-tailed (power law)

Returning to our original question...

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

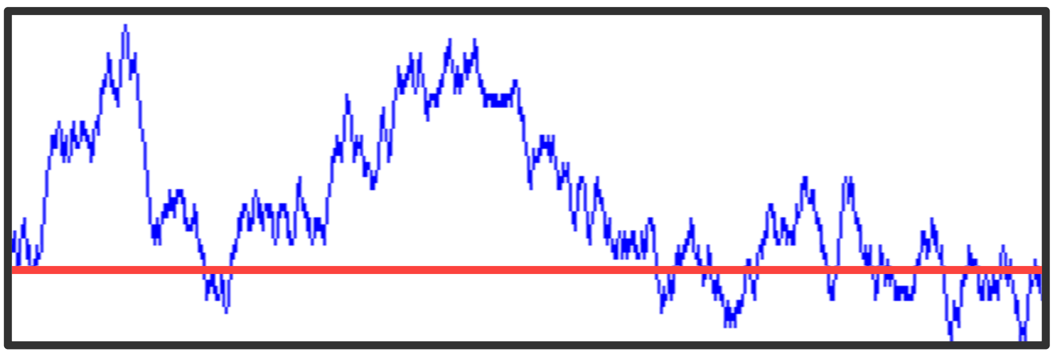
Either the Normal distribution <u>OR</u>
a power-law distribution can emerge!

Returning to our original question...

Consider i.i.d. X_i. How does $\sum_{i=1}^{n} X_i$ grow?

Either the Normal distribution OR
a power-law distribution can emerge!

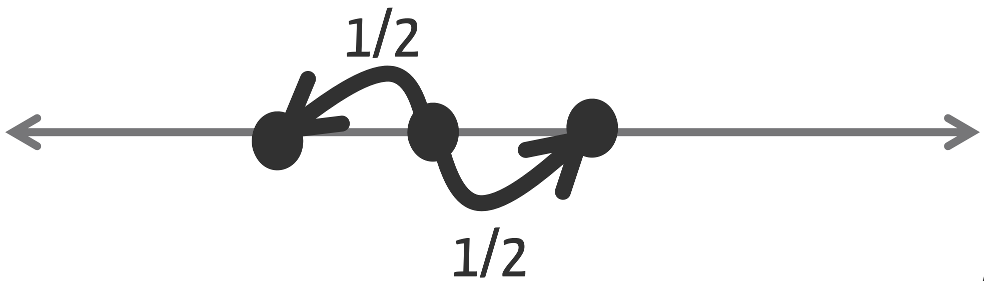
...but this isn't the only question one can ask about $\sum_{i=1}^{n} X_i$.



What is the distribution of the "ruin" time?

The ruin time is always heavy-tailed!

Consider a symmetric 1-D random walk



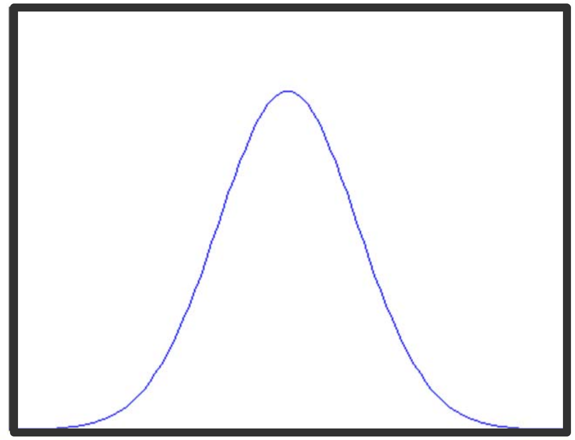
The distribution of ruin time satisfies $\Pr(T > x) \sim \frac{\sqrt{2/\pi}}{\sqrt{x}}$

What is the distribution of the “ruin” time?

The ruin time is always heavy-tailed!

We've all been taught that the <u>Normal is "normal"</u>

...because of the <u>Central Limit Theorem</u>



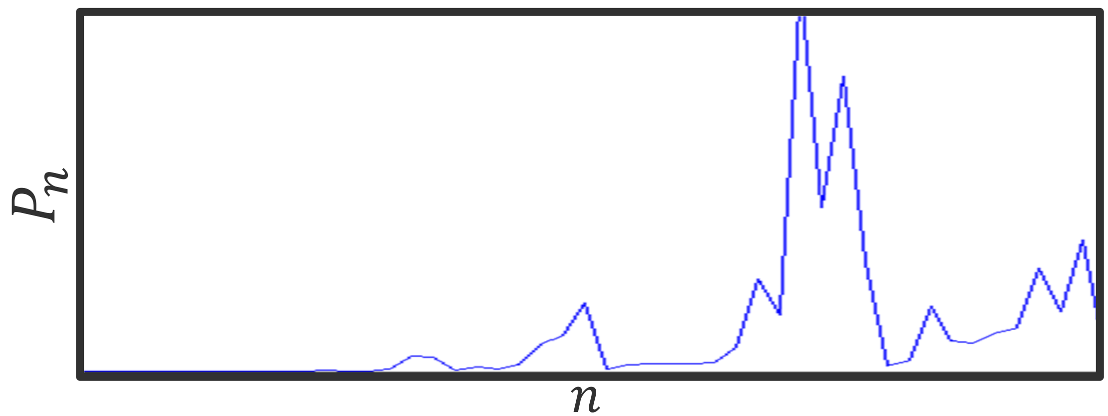
Heavy-tails are more "normal" than the Normal!

1. Additive Processes
2. Multiplicative Processes
3. Extremal Processes

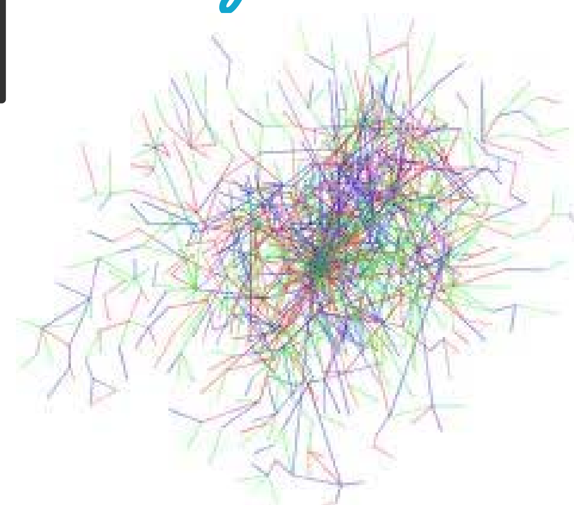
A simple multiplicative process

$P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n$, where Y_i are i.i.d. and positive

Ex: incomes, populations, fragmentation, twitter popularity...



"Rich get richer"



Multiplicative processes almost always lead to heavy tails

An example:

$$Y_1, Y_2 \sim Exponential(\mu)$$

$$\Pr(Y_1 \cdot Y_2 > x) \geq \Pr(Y_1 > \sqrt{x})^2$$

$$= e^{-2\mu\sqrt{x}}$$

$\Rightarrow Y_1 \cdot Y_2$ is heavy-tailed!

Multiplicative processes almost always lead to heavy tails

$$P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n$$

$$\Big\downarrow$$

$$\log P_n = \log Y_1 + \log Y_2 + \cdots + \underbrace{\log Y_n}_{X_n}$$

$$\Big\downarrow \ \text{Central Limit Theorem}$$

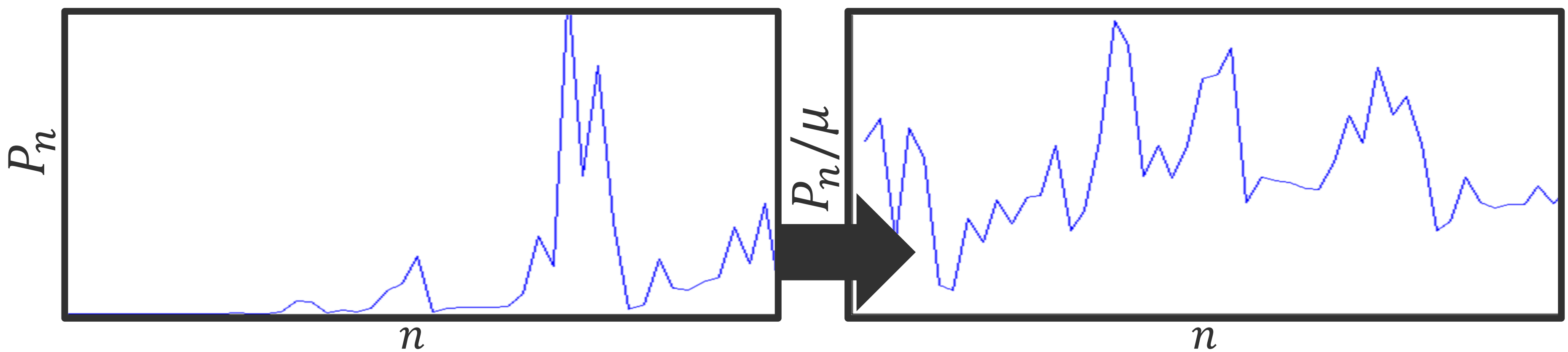
$$\log P_n = n\, E[X_i] + \sqrt{n} Z + o(\sqrt{n}), \text{ where } Z \sim Normal(0, \sigma^2)$$

when $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

$$\Big\Updownarrow$$

$$\left(\frac{Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n}{\mu}\right)^{1/\sqrt{n}} \to H \sim LogNormal(0, \sigma^2)$$

where $\mu = e^{E[\log Y_i]}$
and $\mathrm{Var}[\log Y_i] = \sigma^2 < \infty$.



Multiplicative central limit theorem

$$\left(\frac{Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n}{\mu}\right)^{1/\sqrt{n}} \to H \sim LogNormal(0, \sigma^2)$$

$\boxed{\text{where } \mu = e^{E[\log Y_i]}}$

and $\mathrm{Var}[\log Y_i] = \sigma^2 < \infty$.

Satisfied by all distributions with finite mean and many with infinite mean.

Multiplicative central limit theorem

$$\left(\frac{Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n}{\mu}\right)^{1/\sqrt{n}} \to H \sim LogNormal(0, \sigma^2)$$

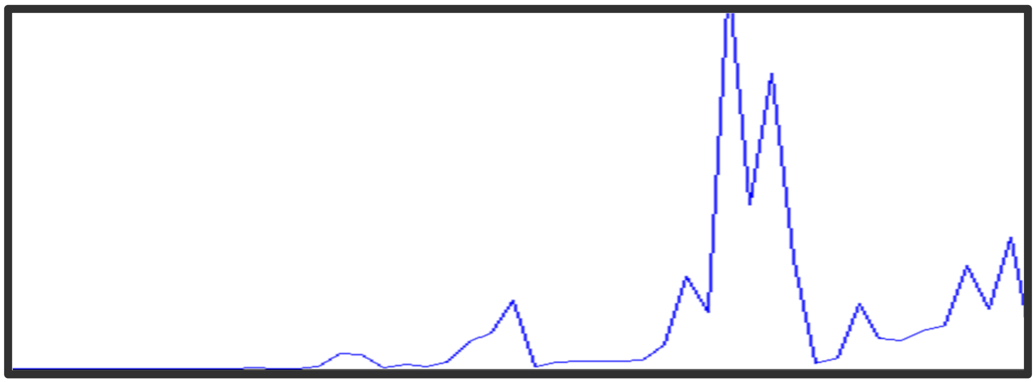
where $\mu = e^{E[\log Y_i]}$

and $\mathrm{Var}[\log Y_i] = \sigma^2 < \infty$.

A simple multiplicative process

$P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n$, where Y_i are i.i.d. and positive

Ex: incomes, populations, fragmentation, twitter popularity...



"Rich get richer"

~~LogNormals~~ emerge

Heavy-tails

A simple multiplicative process

$P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n$, where Y_i are i.i.d. and positive

Ex: incomes, populations, fragmentation, twitter popularity...

Multiplicative process with a lower barrier

$$P_n = \min(P_{n-1} Y_n, \epsilon)$$

Multiplicative process with noise

$$P_n = P_{n-1} Y_n + Q_n$$

Distributions that are approximately power-law emerge

A simple multiplicative process

$P_n = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n$, where Y_i are i.i.d. and positive

Ex: incomes, populations, fragmentation, twitter popularity...

Multiplicative process with a lower barrier

$$P_n = \min(P_{n-1}Y_n, \epsilon)$$

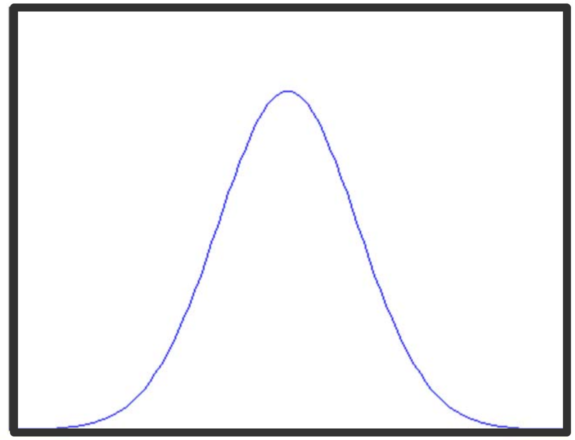
Under minor technical conditions, $P_n \to F$ such that

$\boxed{\lim_{x\to\infty} \frac{\log \bar{F}(x)}{\log x} = s^*}$ where $s^* = \sup(s \geq 0 | E[Y_1^s] \leq 1)$

"Nearly" regularly varying

We've all been taught that the Normal is "normal"

...because of the Central Limit Theorem



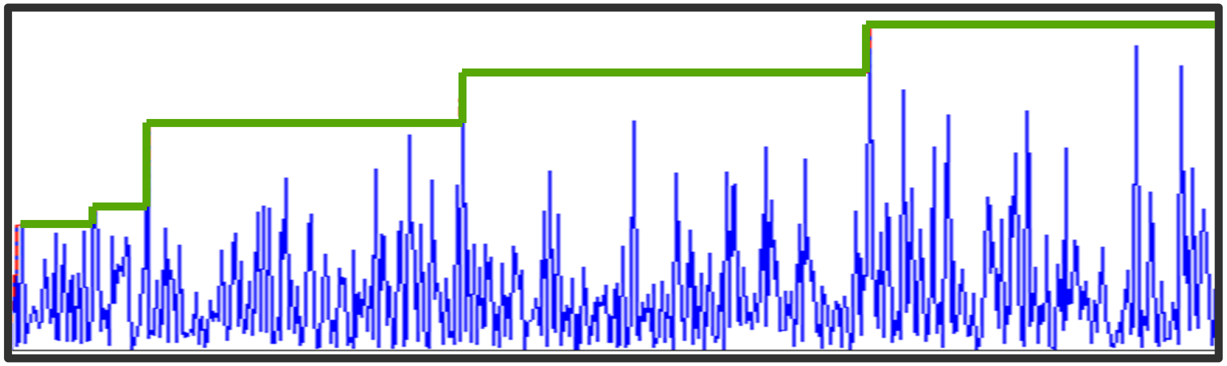
Heavy-tails are more "normal" than the Normal!

1. Additive Processes
2. Multiplicative Processes
3. Extremal Processes

A simple extremal process

$M_n = \max(X_1, X_2, \ldots, X_n)$

Ex: engineering for floods, earthquakes, etc. Progression of world records



"Extreme value theory"

$$M_n = \max(X_1, X_2, \dots, X_n)$$

How does M_n scale? $\frac{M_n - b_n}{a_n}$

A simple example

$X_i \sim Exponential(\mu)$

$$\Pr(\max(X_1, \dots, X_n) > a_n t + b_n) = F(a_n t + b_n)^n$$

$$= \left(1 - \mathrm{e}^{-a_n t - b_n}\right)^n$$

$a_n = 1, b_n = \log n$

$$= \left(1 - \mathrm{e}^{-t - \log n}\right)^n$$

$\rightarrow e^{-e^{-t}}$: **Gumbel distribution**

$M_n = \max(X_1, X_2, \ldots, X_n)$

How does M_n scale? $\frac{M_n - b_n}{a_n}$

"Extremal Central Limit Theorem"

$$\frac{\boldsymbol{M_n - b_n}}{\boldsymbol{a_n}} \to \boldsymbol{Z} \begin{cases} Frechet \longrightarrow \text{Heavy-tailed} \\ Weibull \longrightarrow \text{Heavy or light-tailed} \\ Gumbel \longrightarrow \text{Light-tailed} \end{cases}$$

$M_n = \max(X_1, X_2, \ldots, X_n)$

How does M_n scale? $\dfrac{\boldsymbol{M_n - b_n}}{\boldsymbol{a_n}}$

"Extremal Central Limit Theorem"

$$\frac{\boldsymbol{M_n - b_n}}{\boldsymbol{a_n}} \to \boldsymbol{Z} \begin{cases} Frechet \longrightarrow \text{iff } X_i \text{ are regularly varying} \\ Weibull \longrightarrow \text{e.g. when } X_i \text{ are Uniform} \\ Gumbel \longrightarrow \text{e.g. when } X_i \text{ are LogNormal} \end{cases}$$

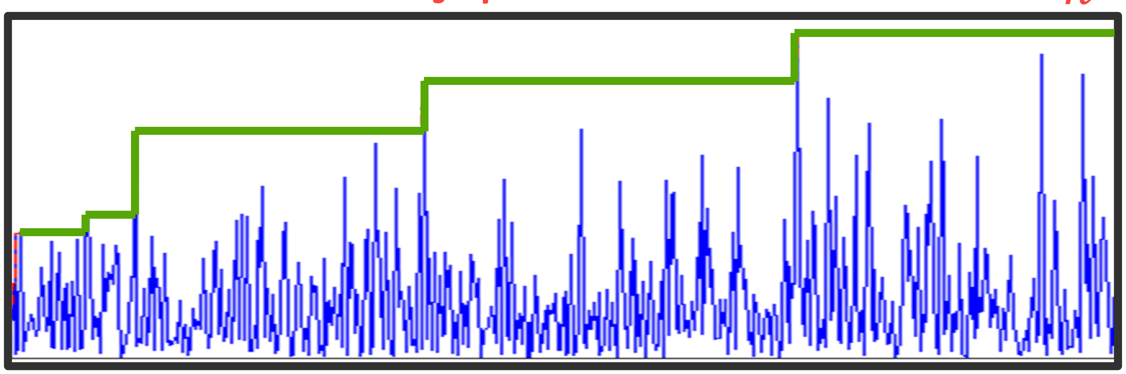
A simple extremal process

$$M_n = \max(X_1, X_2, \ldots, X_n)$$

Ex: engineering for floods, earthquakes, etc. Progression of world records

Either heavy-tailed or light-tailed distributions can emerge as $n \to \infty$

...but this isn't the only question one can ask about M_n.



What is the distribution of the time until a new "record" is set?

The time until a record is always heavy-tailed!

T_k: Time between k & $k+1^{st}$ record

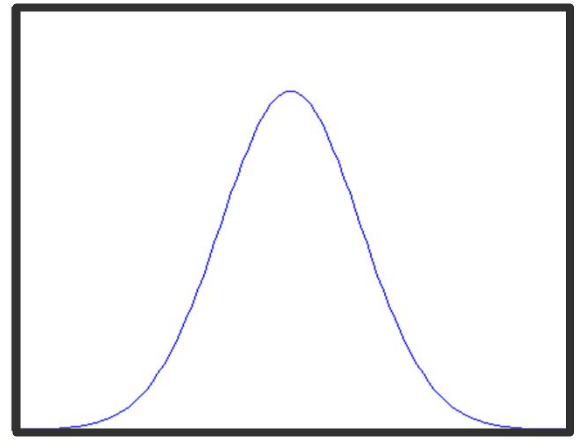
$$\Pr(T_k > n) \sim \frac{2^{k-1}}{n}$$

What is the distribution of the time until a new "record" is set?

The time until a record is always heavy-tailed!

We've all been taught that the Normal is "normal"

...because of the Central Limit Theorem



Heavy-tails are more "normal" than the Normal!

1. Additive Processes
2. Multiplicative Processes
3. Extremal Processes

Heavy-tailed phenomena are treated as something

~~Mysterious~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

3. Identification

Heavy-tailed phenomena are treated as something

MYSTERIOUS, Surprising, & Controversial

On Power-Law Relationships of the Internet Topology

Michalis Faloutsos
U.C. Riverside
Dept. of Comp. Science
michalis@cs.ucr.edu

Petros Faloutsos
U. of Toronto
Dept. of Comp. Science
pfal@cs.toronto.edu

Christos Faloutsos *
Carnegie Mellon Univ.
Dept. of Comp. Science
christos@cs.cmu.edu

1999 Sigcomm paper – 4500+ citations!

BUT...

2005, STOC

On the Bias of Traceroute Sampling

or, Power-law Degree Distributions in Regular Graphs

Dimitris Achlioptas
Microsoft Research
Microsoft Corporation
Redmond, WA 98052
optas@microsoft.com

Aaron Clauset
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
aaron@cs.unm.edu

David Kempe
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
dkempe@usc.edu

Cristopher Moore
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
moore@cs.unm.edu

Similar stories in electricity nets, citation nets, ...

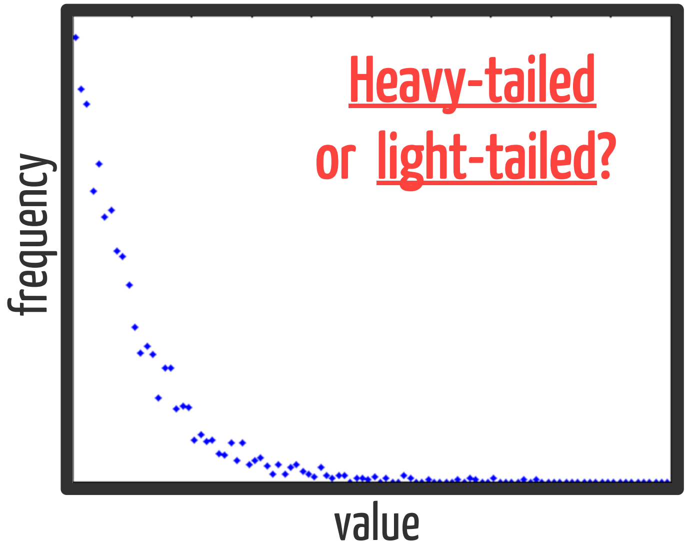
IEEE/ACM TRANSACTIONS ON NET 1205

Understanding Internet Topology: Principles, Models, and Validation

David Alderson, *Member, IEEE*, Lun Li, *Student Member, IEEE*, Walter Willinger, *Fellow, IEEE*, and John C. Doyle, *Member, IEEE*

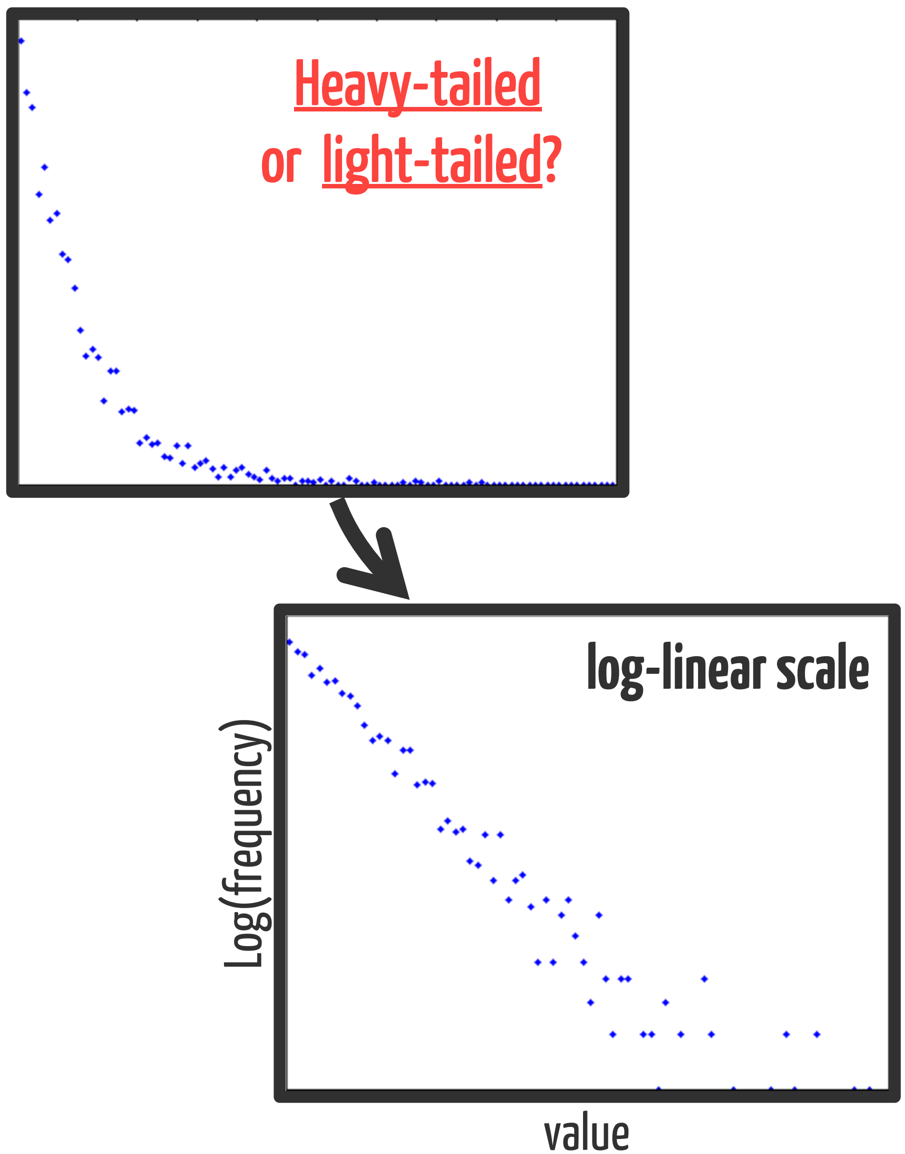
2005, ToN

A "typical" approach for identifying of heavy tails: Linear Regression

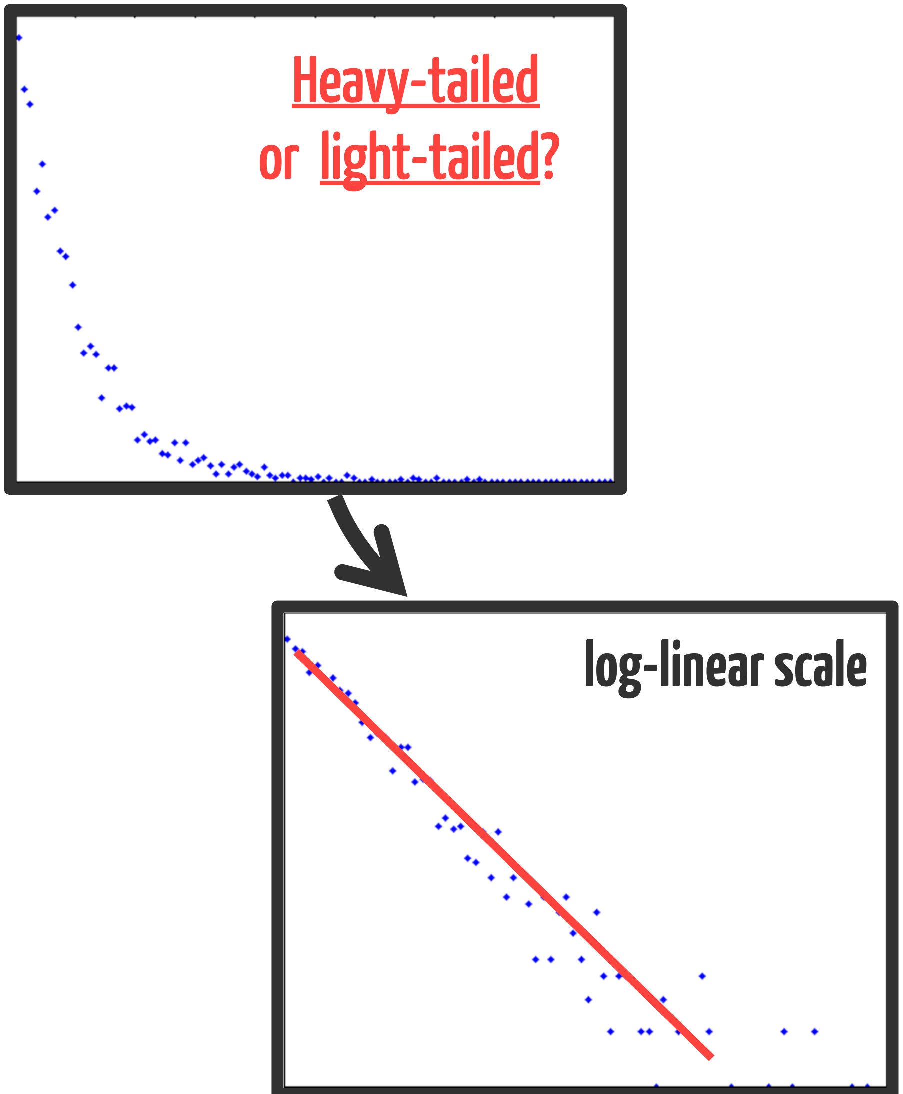


"frequency plot"

A “typical” approach for identifying of heavy tails: Linear Regression



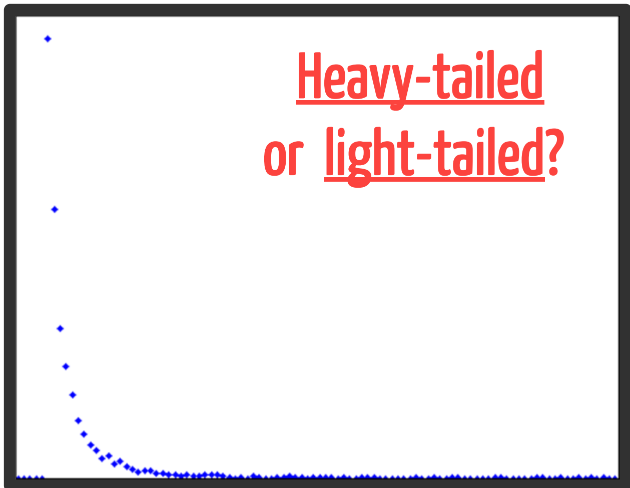
A "typical" approach for identifying of heavy tails: Linear Regression



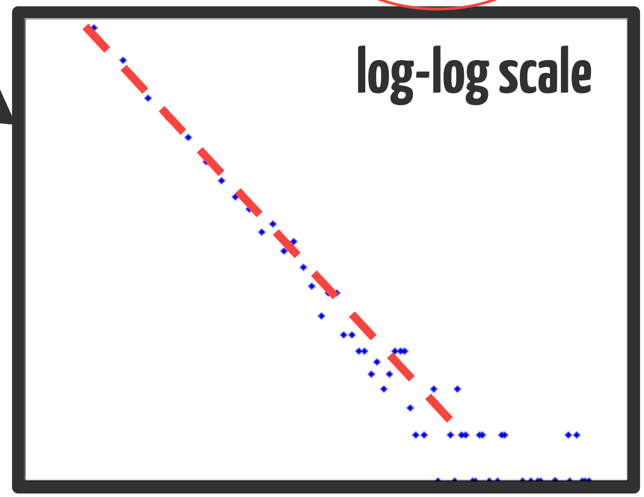
Linear ⇒
Exponential tail

$$f(x) = Ce^{-\mu x} \rightarrow \log f(x) = \log C - \mu x$$

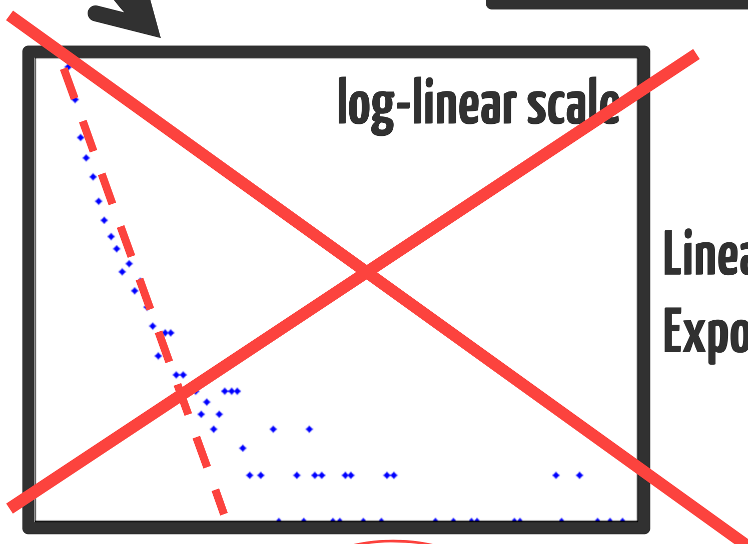
A “typical” approach for identifying of heavy tails: Linear Regression



$$f(x) = Cx^{-\alpha-1} \rightarrow \log f(x) = \log C - (\alpha+1)\log x$$



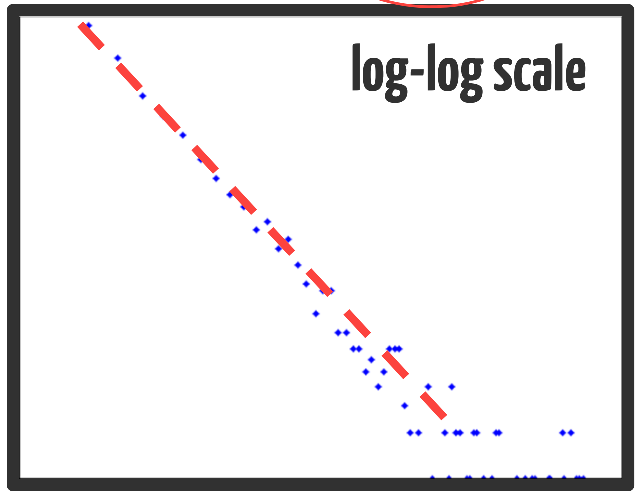
Linear ⇒
Power-law tail



Linear ⇒
Exponential tail

$$f(x) = Ce^{-\mu x} \rightarrow \log f(x) = \log C - \mu x$$

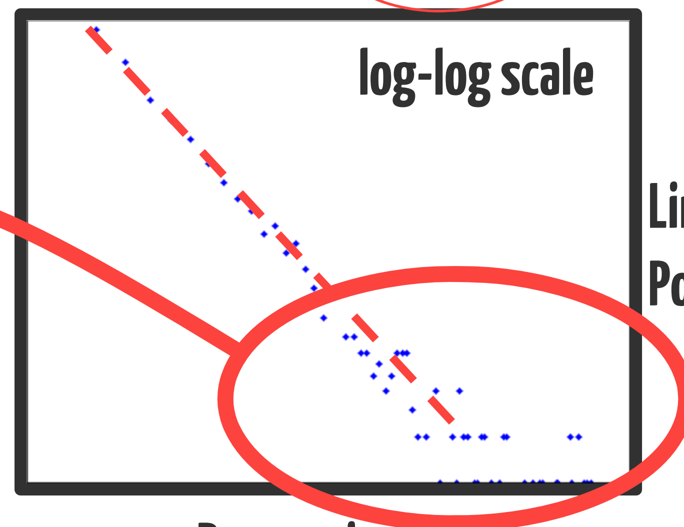
$$f(x) = Cx^{-\alpha-1} \rightarrow \log f(x) = \log C - (\alpha+1)\log x$$



Linear ⇒
Power-law tail

Regression ⇒
Estimate of tail index (α)

$$f(x) = Cx^{-\alpha-1} \rightarrow \log f(x) = \log C - (\alpha + 1) \log x$$

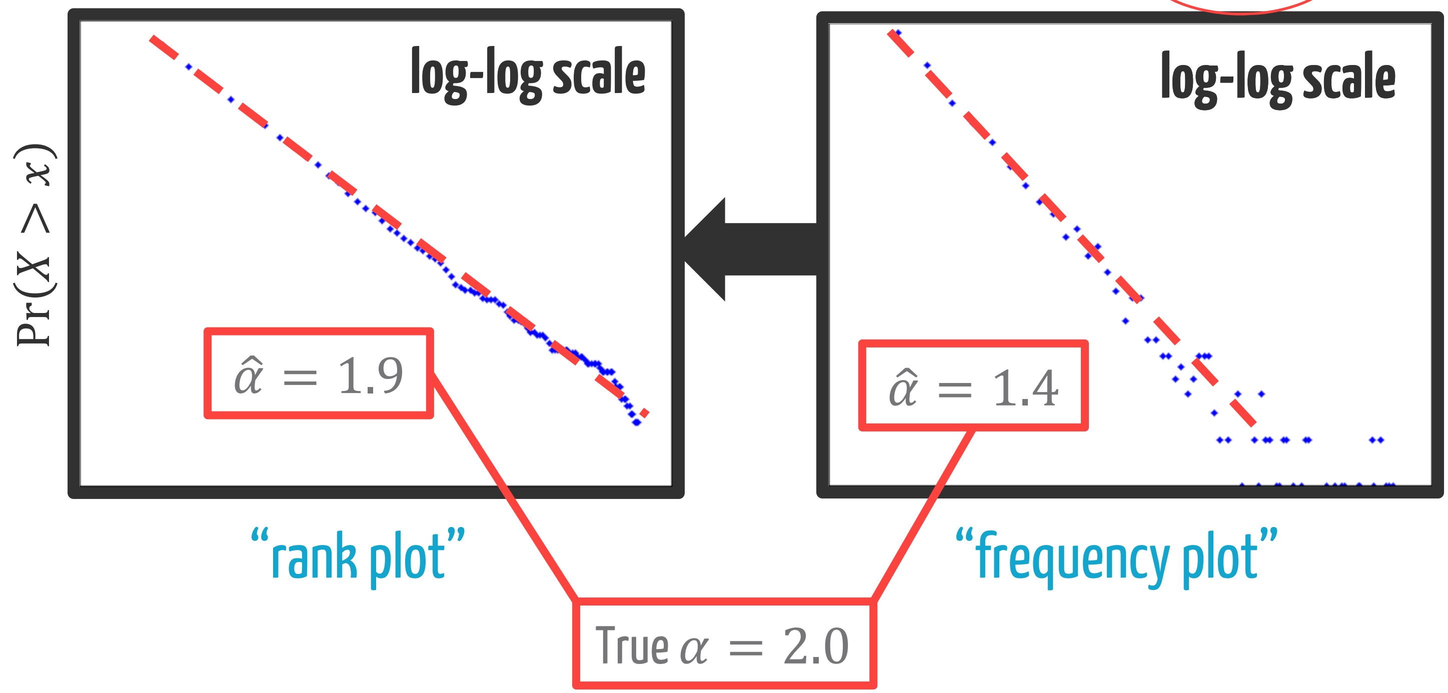


Linear $\Rightarrow$
Power-law tail

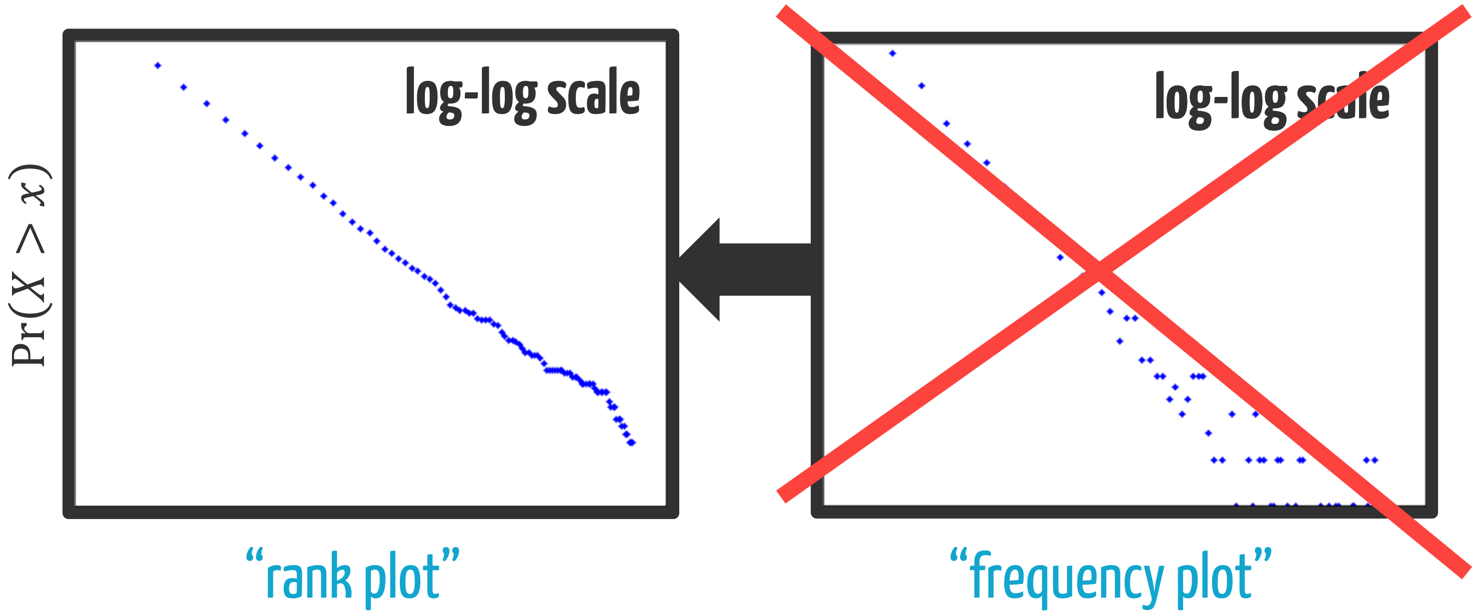
Regression $\Rightarrow$
Estimate of tail index (α)

Is it really linear?
Is the estimate of α accurate?

$$\Pr(X > x) = \bar{F}(x) = C'x^{\alpha} \longleftarrow f(x) = Cx^{-\alpha-1} \longrightarrow \log f(x) = \log C - (\alpha + 1) \log x$$

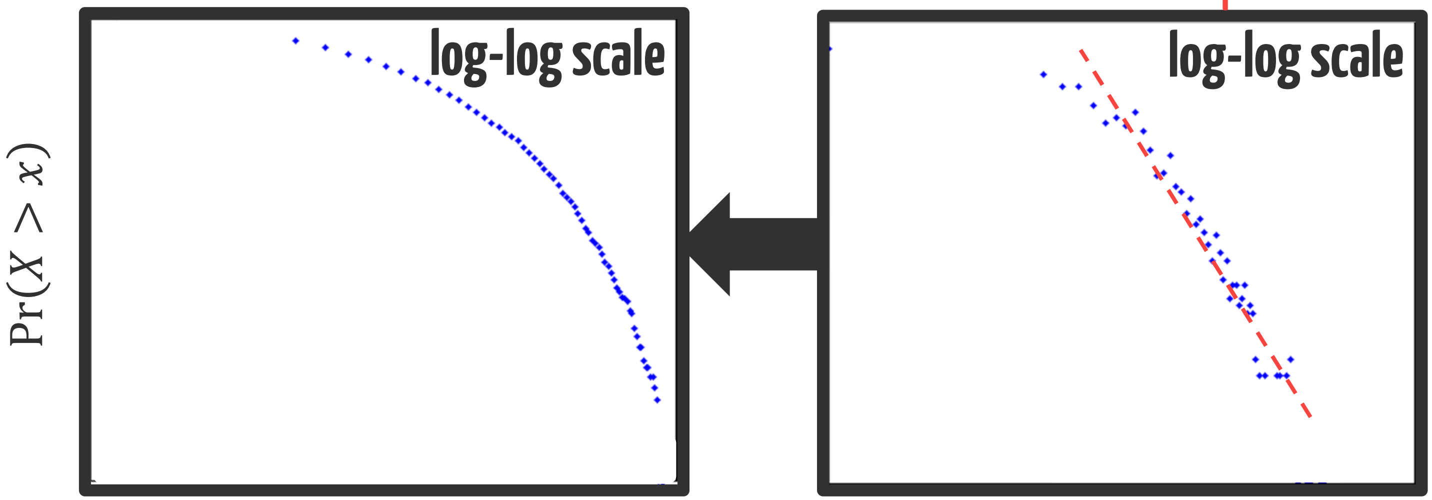


This simple change is extremely important...



This simple change is extremely important…

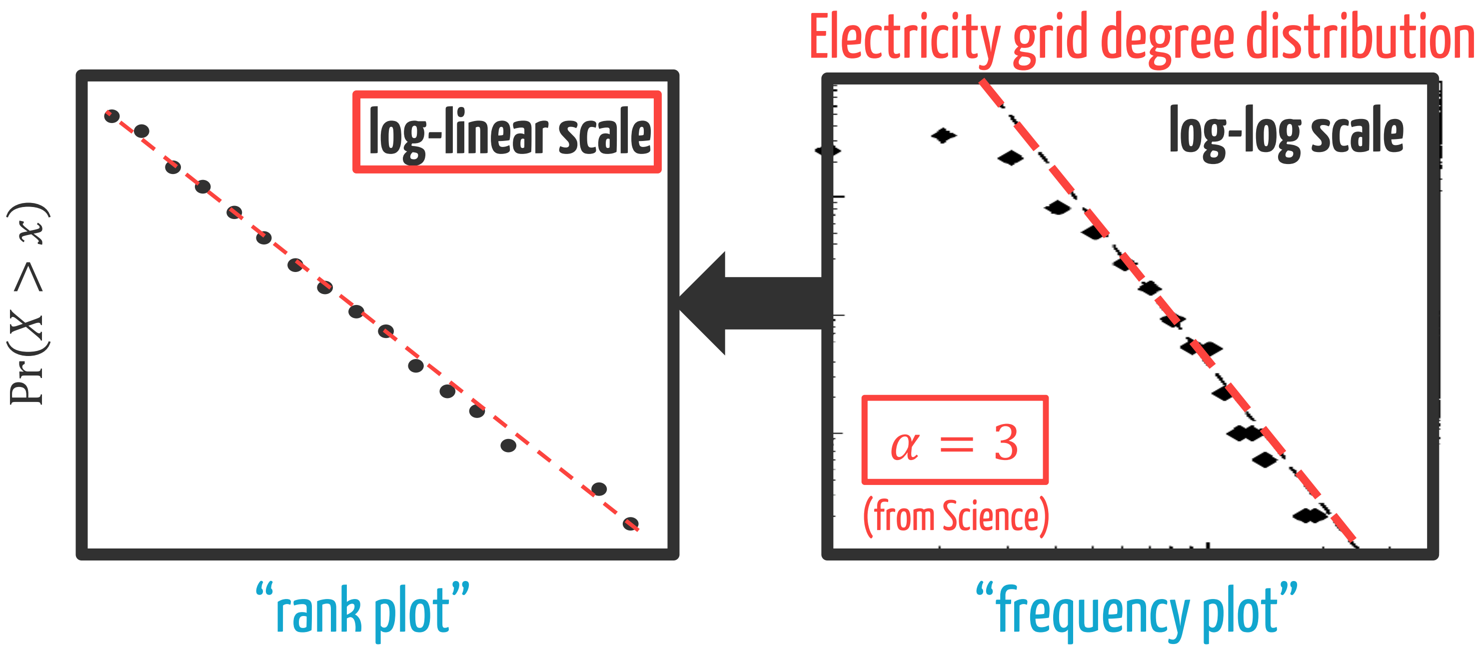
Does this look like a power law?



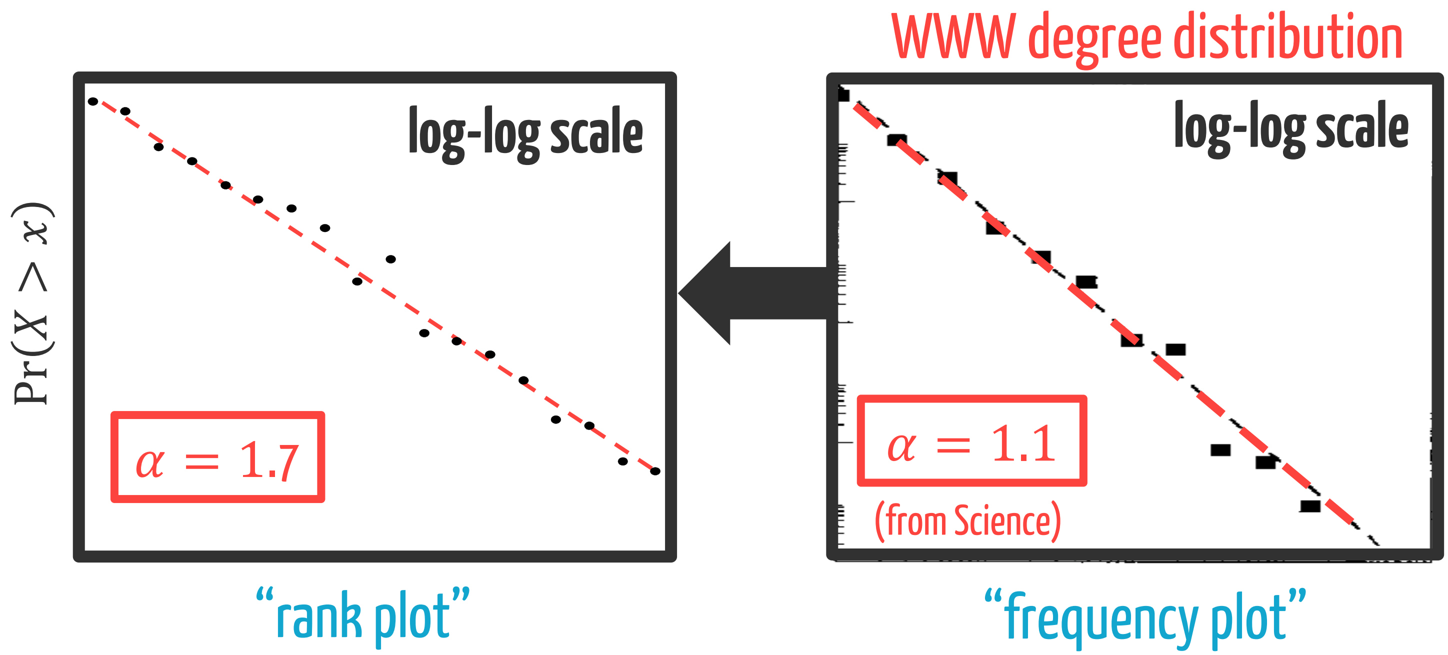
“rank plot” “frequency plot”

The data is from an Exponential!

This mistake has happened A LOT!

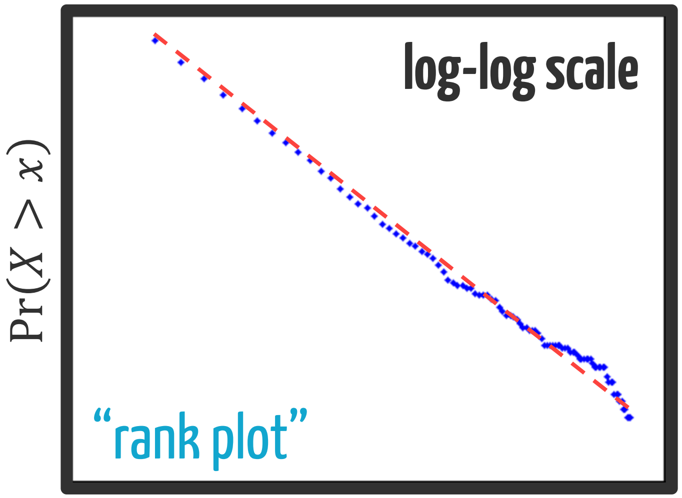


This mistake has happened A LOT!



This simple change is extremely important...

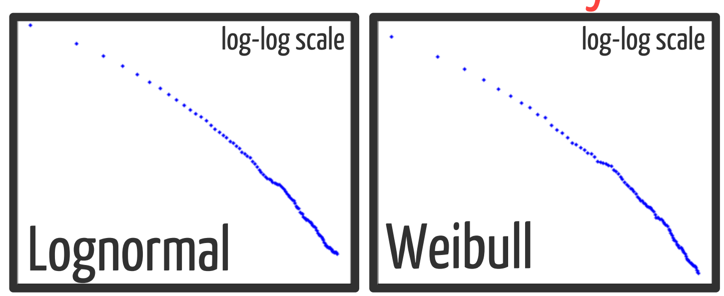
But, this is still an error-prone approach



Regression ⇒
Estimate of tail index (α)

Linear ⇒
Power-law tail

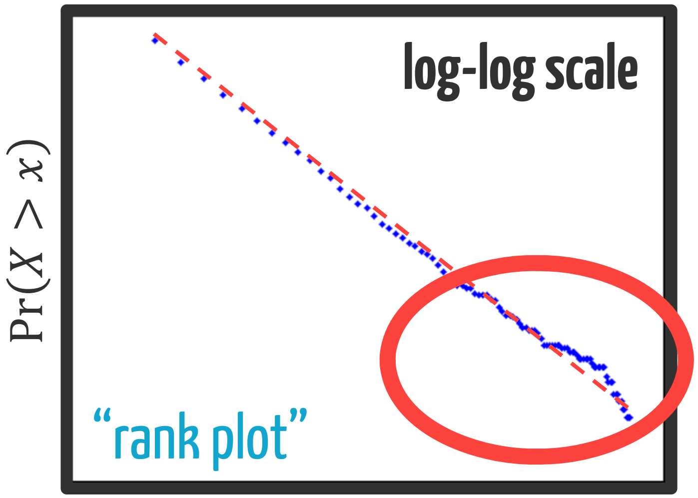
...other distributions can be nearly linear too



•••

This simple change is extremely important...

But, this is still an error-prone approach



Linear $\Rightarrow$

Power-law tail

...other distributions can be nearly linear too

Regression $\Rightarrow$

Estimate of tail index (α)

...assumptions of regression are not met

...tail is much noisier than the body

A completely different approach: Maximum Likelihood Estimation (MLE)

What is the α for which the data is most "likely"?

$$L(x;\alpha) = \prod_{i=1}^{n} \frac{\alpha x_{\min}^{\alpha}}{x_i^{\alpha+1}}$$

$$\log L(x;\alpha) = \sum_{i=1}^{n} \log(\alpha x_{\min}^{\alpha}) - \log x_i^{\alpha+1}$$

Maximizing gives $\hat{\alpha}_{MLE} = \dfrac{n}{\sum_{i=1}^{n} \log(x_i/x_{\min})}$

This has many nice properties:

→ $\hat{\alpha}_{MLE}$ is the minimal variance, unbiased estimator.

→ $\hat{\alpha}_{MLE}$ is asymptotically efficient.

A ~~completely~~ not so different approach: Maximum Likelihood Estimation (MLE)

Weighted Least Squares Regression (WLS)

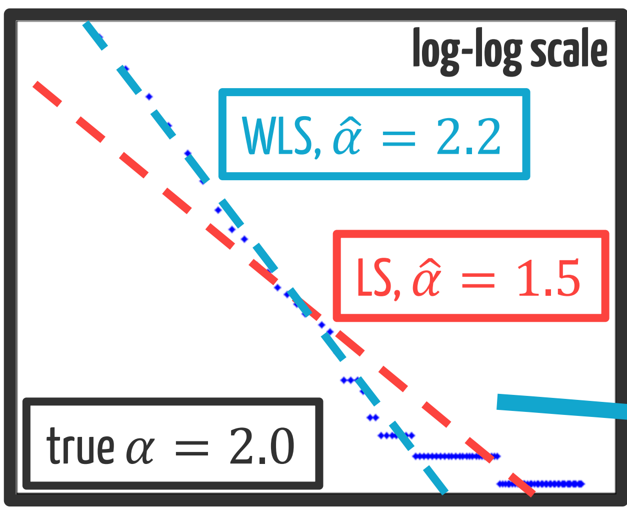
asymptotically for large data sets, when weights are chosen as $w_i = 1/\left(\log x_i - \log x_0\right)$.

$$\hat{\alpha}_{WLS} = \frac{-\sum_{i=1}^{n} \log(\hat{r}_i/n)}{\sum_{i=1}^{n} \log(x_i/x_0)}$$
$$\sim \frac{n}{\sum_{i=1}^{n} \log(x_i/x_0)}$$
$$= \hat{\alpha}_{MLE}$$

A ~~completely~~ not so different approach: Maximum Likelihood Estimation (MLE)

Weighted Least Squares Regression (WLS)

asymptotically for large data sets, when weights are chosen as $w_i = 1/\left(\log x_i - \log x_0\right)$.



"Listen to your body"

A quick summary of where we are:

Suppose data comes from a power-law (Pareto) distribution $\bar{F}(x) = \left(\frac{x_0}{x}\right)^{\alpha}$.

Then, we can identify this visually with a log-log plot,
and we can estimate α using either MLE or WLS.

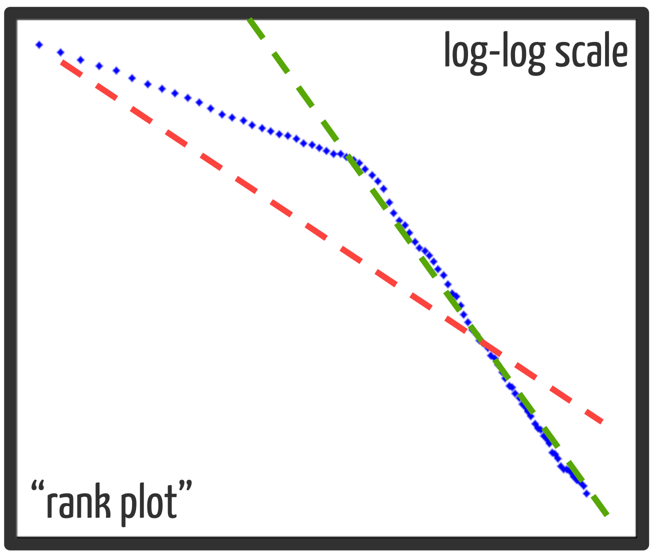
What if the data is not exactly a power-law?

What if only the tail is power-law?

Suppose data comes from a ~~power-law (Pareto) distribution~~ $\bar{F}(x) = \left(\frac{x_0}{x}\right)^{\alpha}$.

Then, we can identify this visually with a log-log plot,

and we can estimate α using either MLE or WLS.



Can we just use MLE/WLS on the "tail"?

But, where does the tail start?

Impossible to answer...

An example

Suppose we have a mixture of power laws:

$$\bar{F}(x) = q\bar{F}_1(x) + (1-q)\bar{F}_2(x)$$

$$\alpha_1 < \alpha_2$$

We want $\hat{\alpha}_{MLE} \to \alpha_1$ as $n \to \infty$.

...but, suppose we use $x_{\min}$ as our cutoff:

$$\frac{1}{\hat{\alpha}_{MLE}} \to \frac{q\bar{F}_1(x_{\min})}{\alpha_1\bar{F}(x_{\min})} + \frac{(1-q)\bar{F}_2(x_{\min})}{\alpha_2\bar{F}(x_{\min})} \neq \alpha_1$$

Identifying power-law <u>distributions</u> → MLE/WLS

"Listen to your body"

v.s.

Identifying power-law <u>tails</u> → Extreme value theory

"Let the tail do the talking"

Returning to our example

Suppose we have a mixture of power laws:

$$\bar{F}(x) = q\bar{F}_1(x) + (1-q)\bar{F}_2(x)$$

$$\alpha_1 < \alpha_2$$

We want $\hat{\alpha}_{MLE} \to \alpha_1$ as $n \to \infty$.

...but, suppose we use $x_{\min}$ as our cutoff:

$$\frac{1}{\hat{\alpha}_{MLE}} \to \boxed{\frac{q\bar{F}_1(x_{\min})}{\alpha_1\bar{F}(x_{\min})} + \frac{(1-q)\bar{F}_2(x_{\min})}{\alpha_2\bar{F}(x_{\min})}}$$

The bias disappears as $x_{\min} \to \infty$!

The idea: **Improve robustness by throwing away nearly all the data!**

$$x_{\min} \longrightarrow x_{\min}(n)$$, where $x_{\min}(n) \to \infty$ as $n \to \infty$.

$$\left[\begin{array}{l} \text{+ Larger } x_{\min}(n) \Rightarrow \text{Small bias} \\ \text{- Larger } x_{\min}(n) \Rightarrow \text{Larger variance} \end{array}\right]$$

The idea: Improve robustness by throwing away nearly all the data!

$x_{\min} \rightarrow x_{\min}(n)$, where $x_{\min}(n) \to \infty$ as $n \to \infty$.

The Hill Estimator

$$\hat{\alpha}(k,n) = \frac{k}{\boxed{\sum_{i=1}^{k} \log\left(\frac{x_{(i)}}{x_{(k)}}\right)}}$$

where $x_{(k)}$ is the kth largest data point

Looks almost like the MLE, but uses order kth order statistic

The idea: **Improve robustness by throwing away nearly all the data!**

$x_{\min} \longrightarrow x_{\min}(n)$, where $x_{\min}(n) \to \infty$ as $n \to \infty$.

The Hill Estimator

$$\hat{\alpha}(k,n) = \frac{k}{\boxed{\sum_{i=1}^{k} \log\left(\frac{x_{(i)}}{x_{(k)}}\right)}}$$

where $x_{(k)}$ is the kth largest data point

Looks almost like the MLE, but uses order kth order statistic

...how do we choose k?

$\hat{\alpha}(k,n) \to \alpha$ as $n \to \infty$ if

$\underbrace{k(n)/n \to 0}_{\text{throw away nearly all the data,}}$ & $\underbrace{k(n) \to \infty}_{\text{but keep enough data for consistency}}$

The idea: Improve robustness by throwing away nearly all the data!

$x_{\min} \longrightarrow x_{\min}(n)$, where $x_{\min}(n) \to \infty$ as $n \to \infty$.

The Hill Estimator

$$\hat{\alpha}(k,n) = \frac{k}{\boxed{\sum_{i=1}^{k} \log\left(\frac{x_{(i)}}{x_{(k)}}\right)}}$$

where $x_{(k)}$ is the kth largest data point

Looks almost like the MLE, but uses order kth order statistic

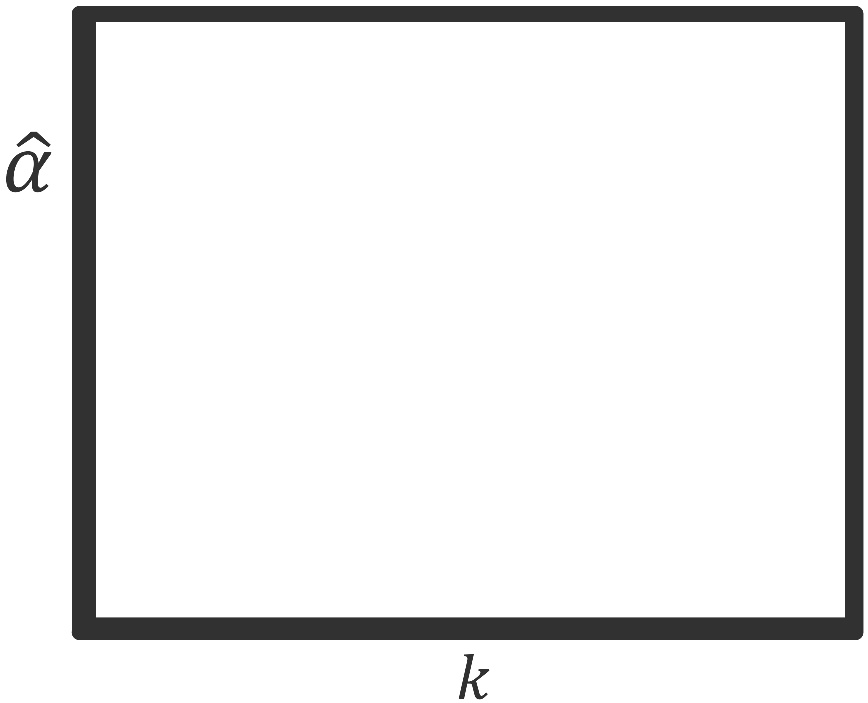
...how do we choose k?

$\hat{\alpha}(k,n) \to \alpha$ as $n \to \infty$ if

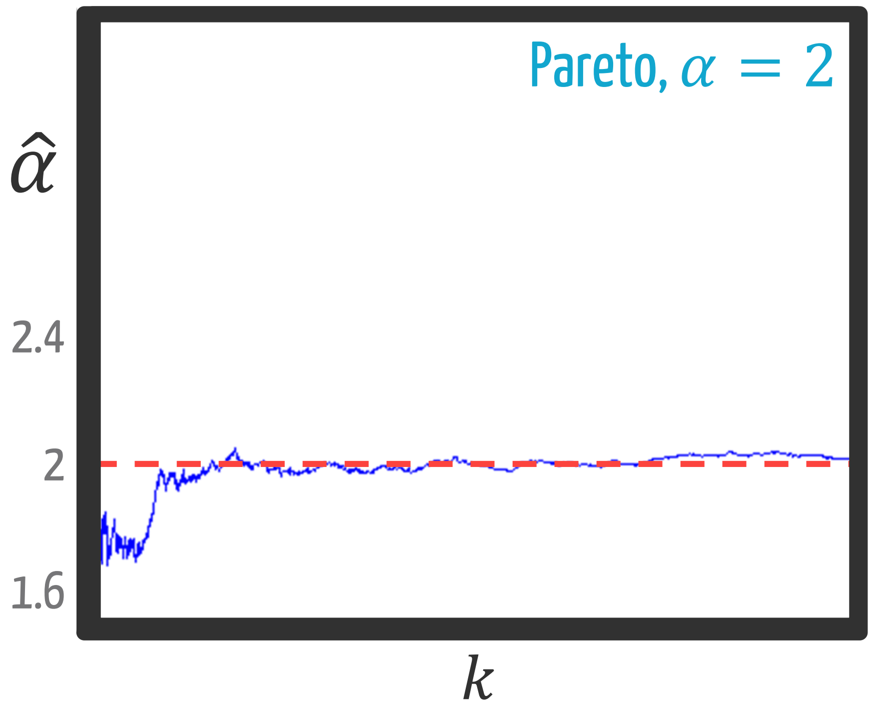
$\underline{k(n)/n \to 0}$ & $\underline{k(n) \to \infty}$

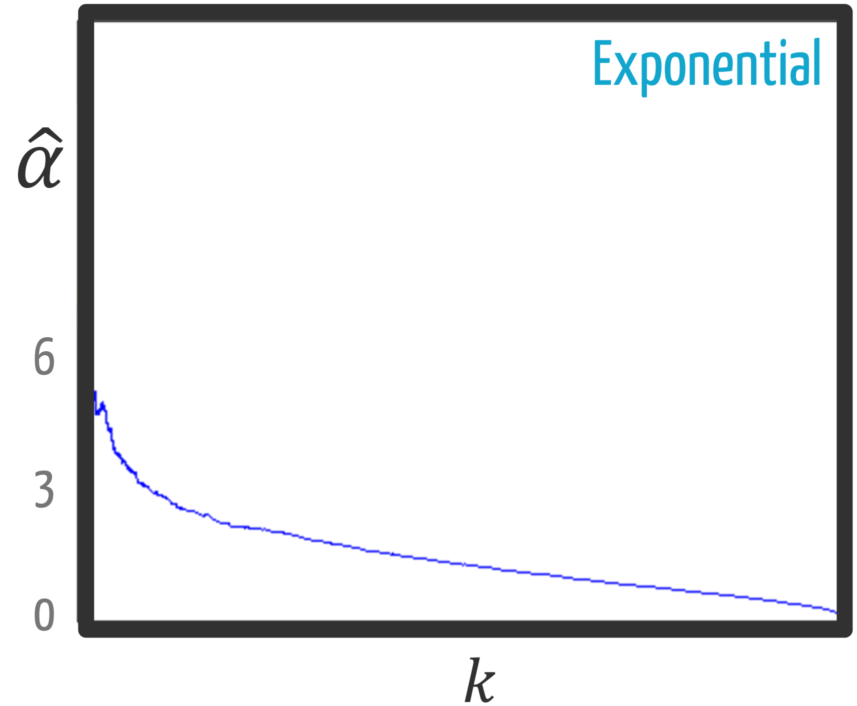
Throw away everything except the outliers!

Choosing k in practice: The Hill plot

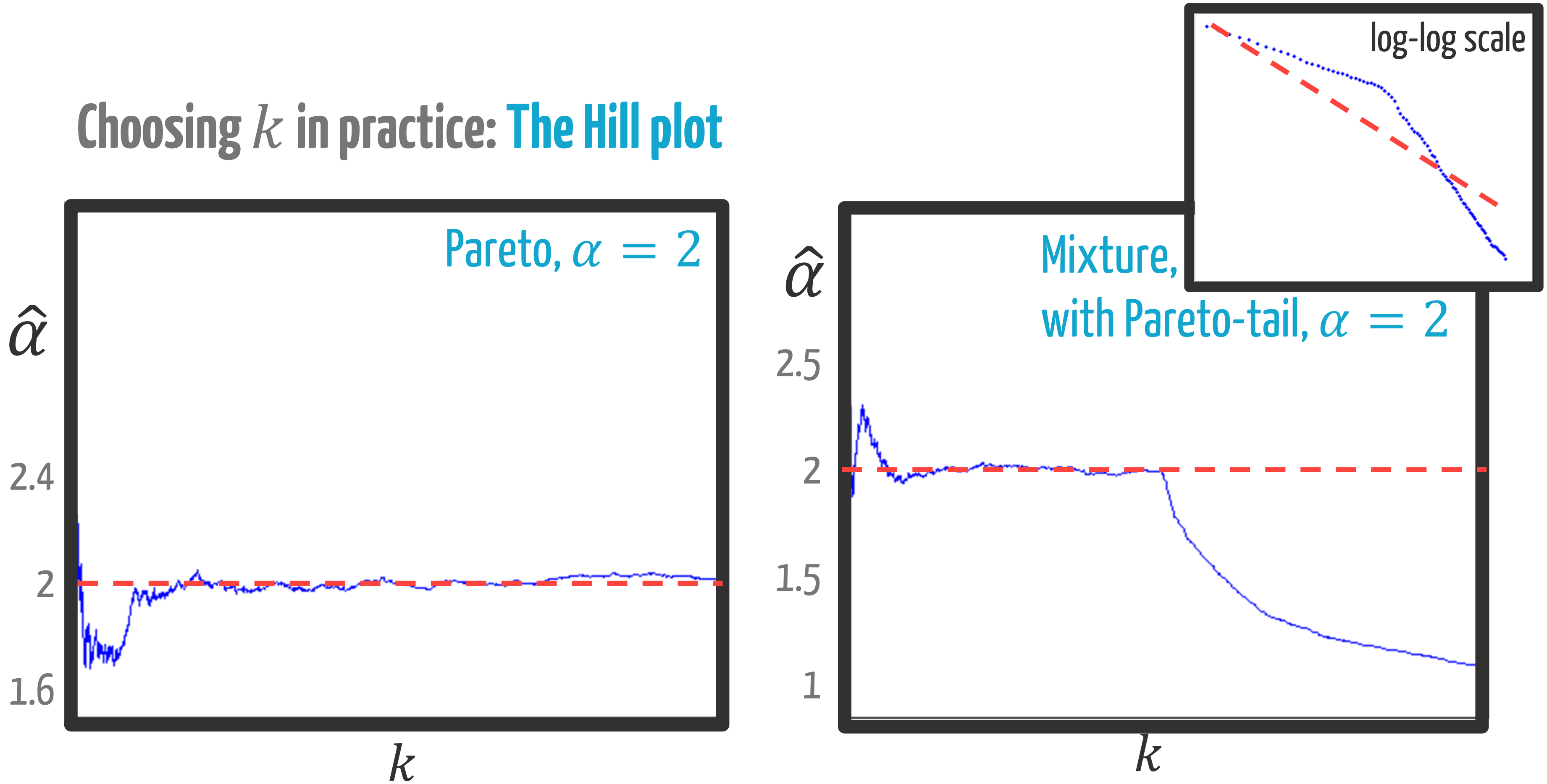


Choosing k in practice: The Hill plot

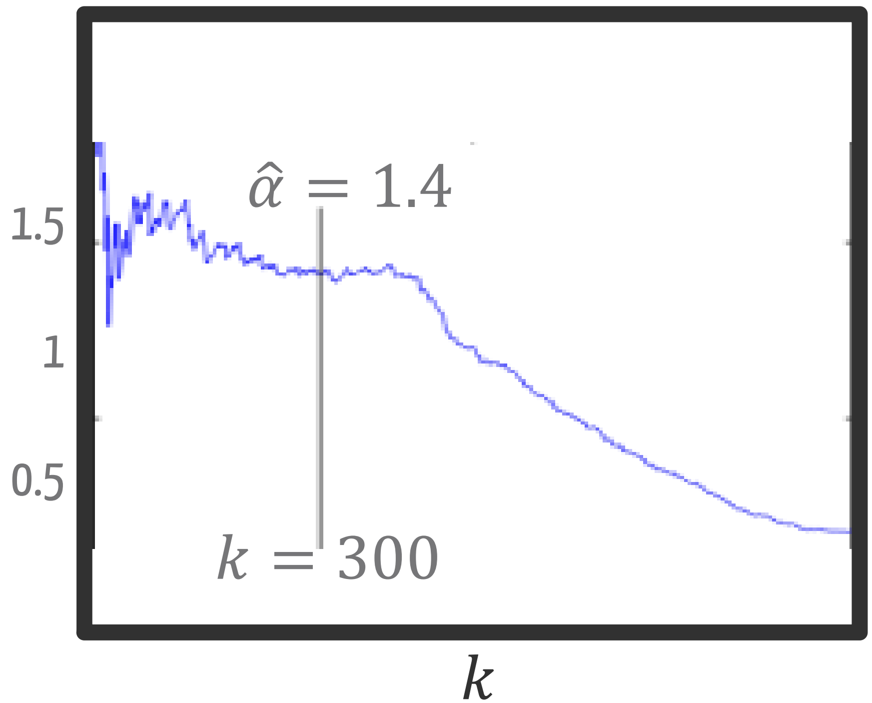


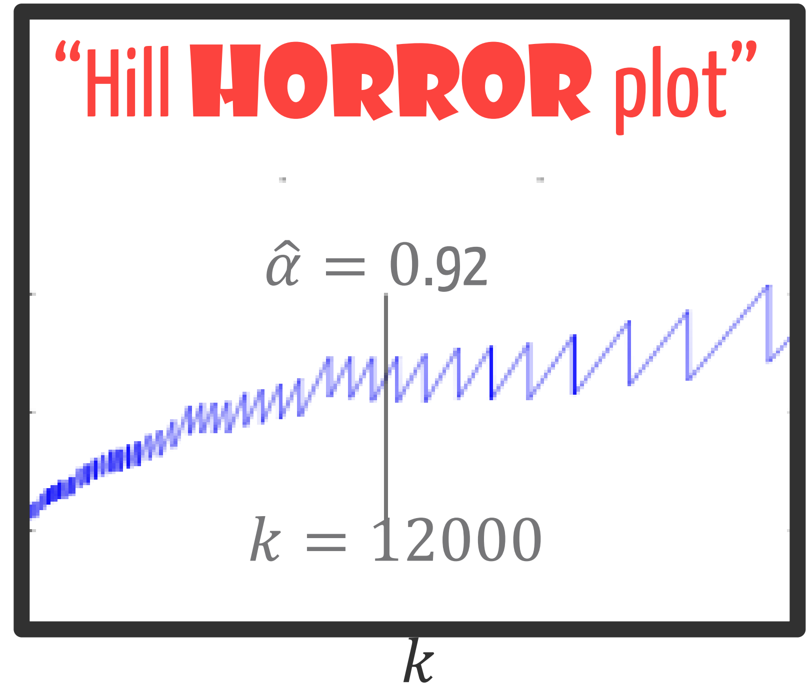


Choosing k in practice: The Hill plot



...but the hill estimator has problems too





This data is from TCP flow sizes!

Identifying power-law distributions → MLE/WLS

"Listen to your body"

Identifying power-law tails → Hill estimator

"Let the tail do the talking"

☆ It's dangerous to rely on any one technique!

(see our forthcoming book for other approaches)

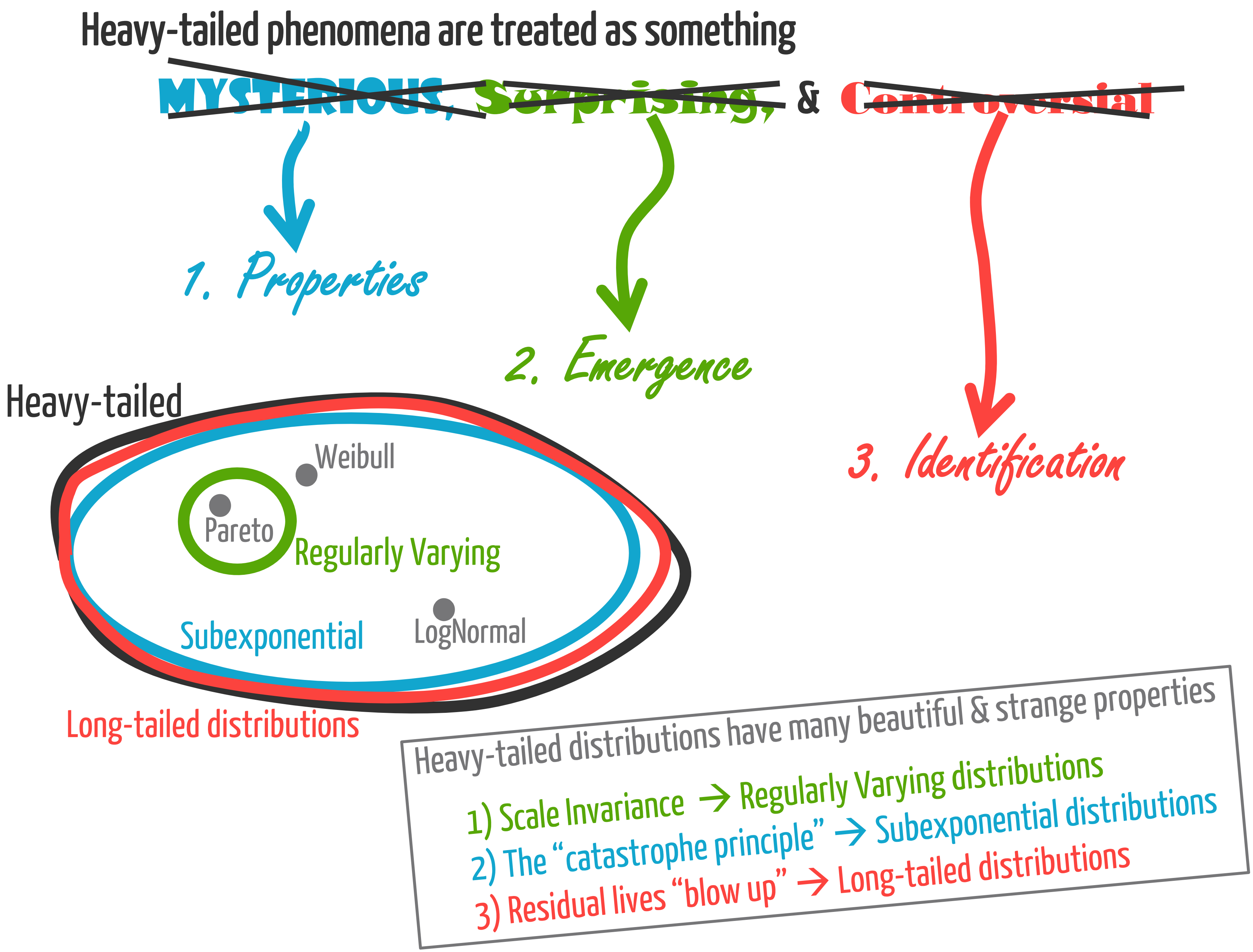
Heavy-tailed phenomena are treated as something

~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

3. Identification


Heavy-tailed phenomena are treated as something
~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~
1. Properties
2. Emergence
3. Identification
Heavy-tailed
Weibull
Pareto
Regularly Varying
Subexponential
LogNormal
Long-tailed distributions
Heavy-tailed distributions have many beautiful & strange properties
1) Scale Invariance → Regularly Varying distributions
2) The "catastrophe principle" → Subexponential distributions
3) Residual lives "blow up" → Long-tailed distributions

Heavy-tailed phenomena are treated as something

~~MYSTERIOUS~~, ~~Surprising~~, & ~~Controversial~~

1. Properties

2. Emergence

3. Identification

We've all been taught that the Normal is "normal"
because of the Central Limit Theorem, **BUT**
Heavy-tails are more "normal" than the Normal!

Heavy-tailed phenomena are treated as something

~~Mysterious~~, ~~Surprising~~, & ~~Controversial~~

1. *Properties*

2. *Emergence*

3. *Identification*

Identifying power-law distributions → MLE/WLS

"Listen to your body"

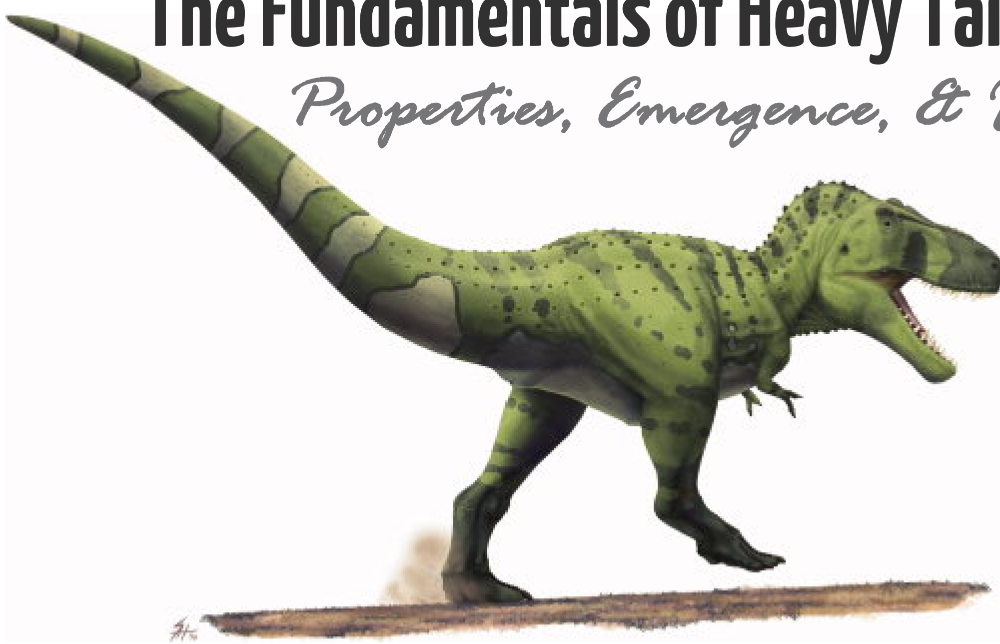
Identifying power-law tails → Hill estimator

"Let the tail do the talking"

...and others we didn't talk about

The Fundamentals of Heavy Tails

Properties, Emergence, & Identification



Jayakrishnan Nair, Adam Wierman, Bert Zwart

"The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of the tweets. These figures are always reported as shocking […] as if anything but a nice bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world."
– Clay Shirky, as quoted in Newsweek & the Guardian