

Lecture 20: Skewed data

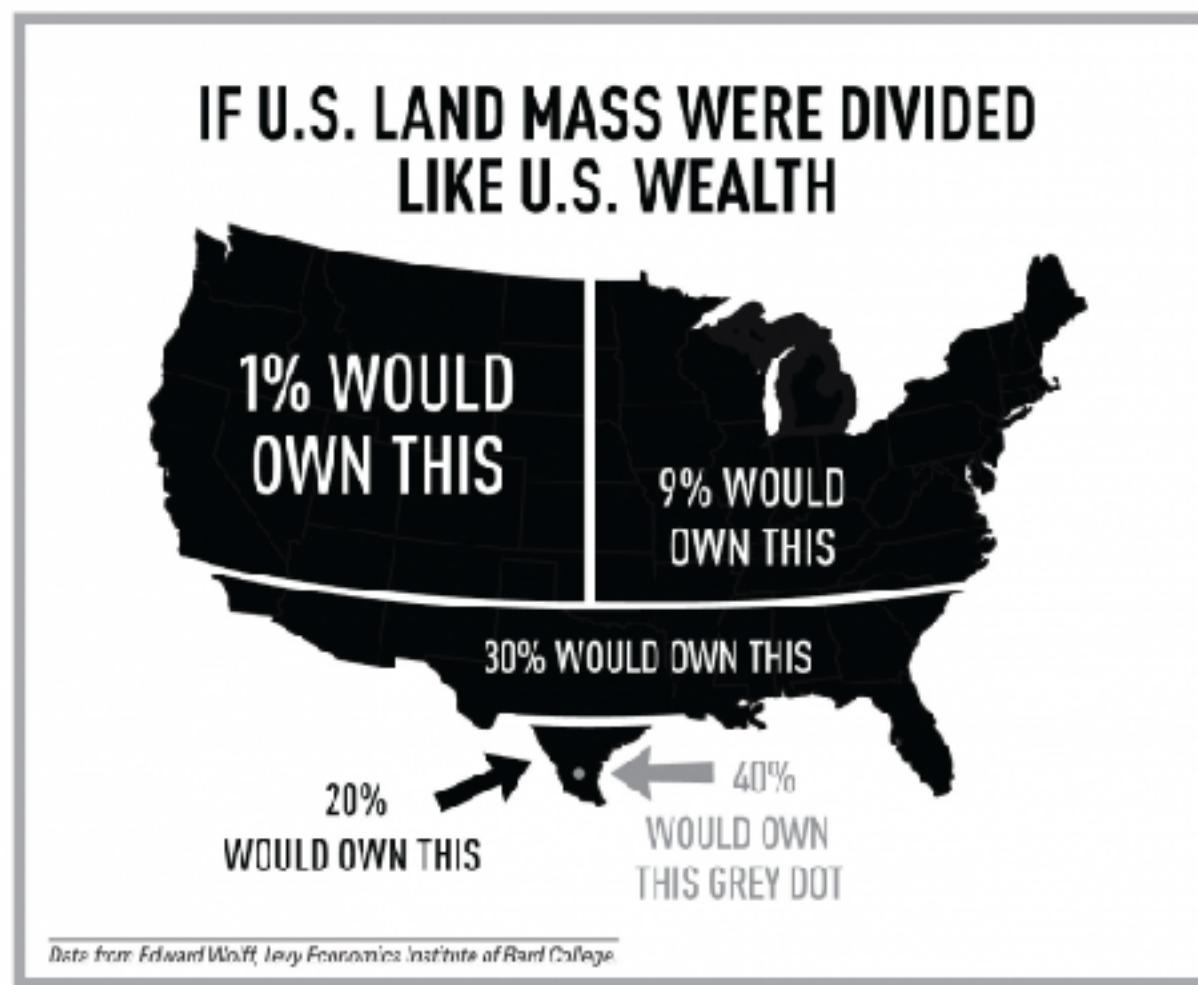
Instructor: Michael Szell

Nov 10, 2023

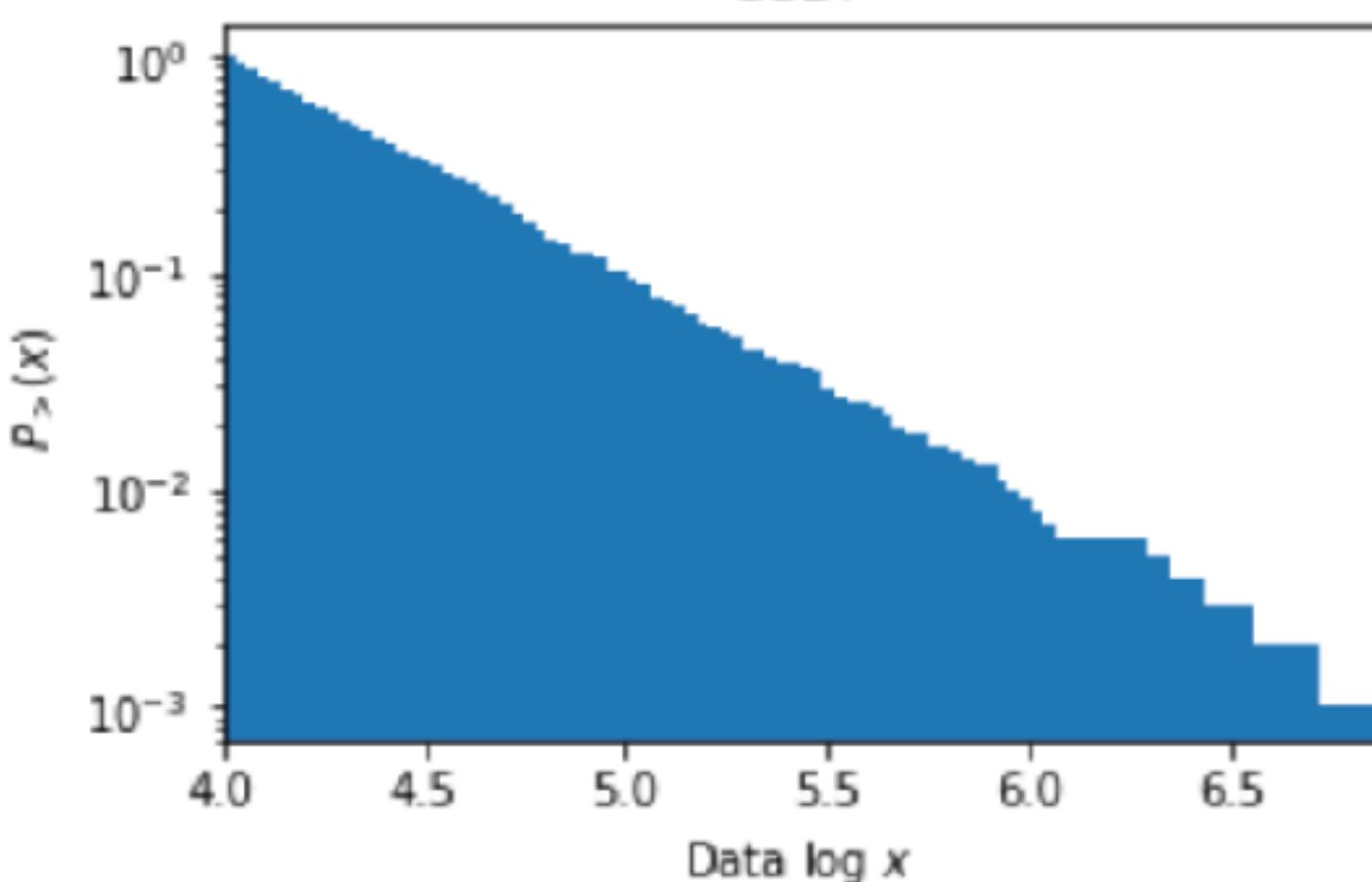


Today you will learn about skewed data distributions

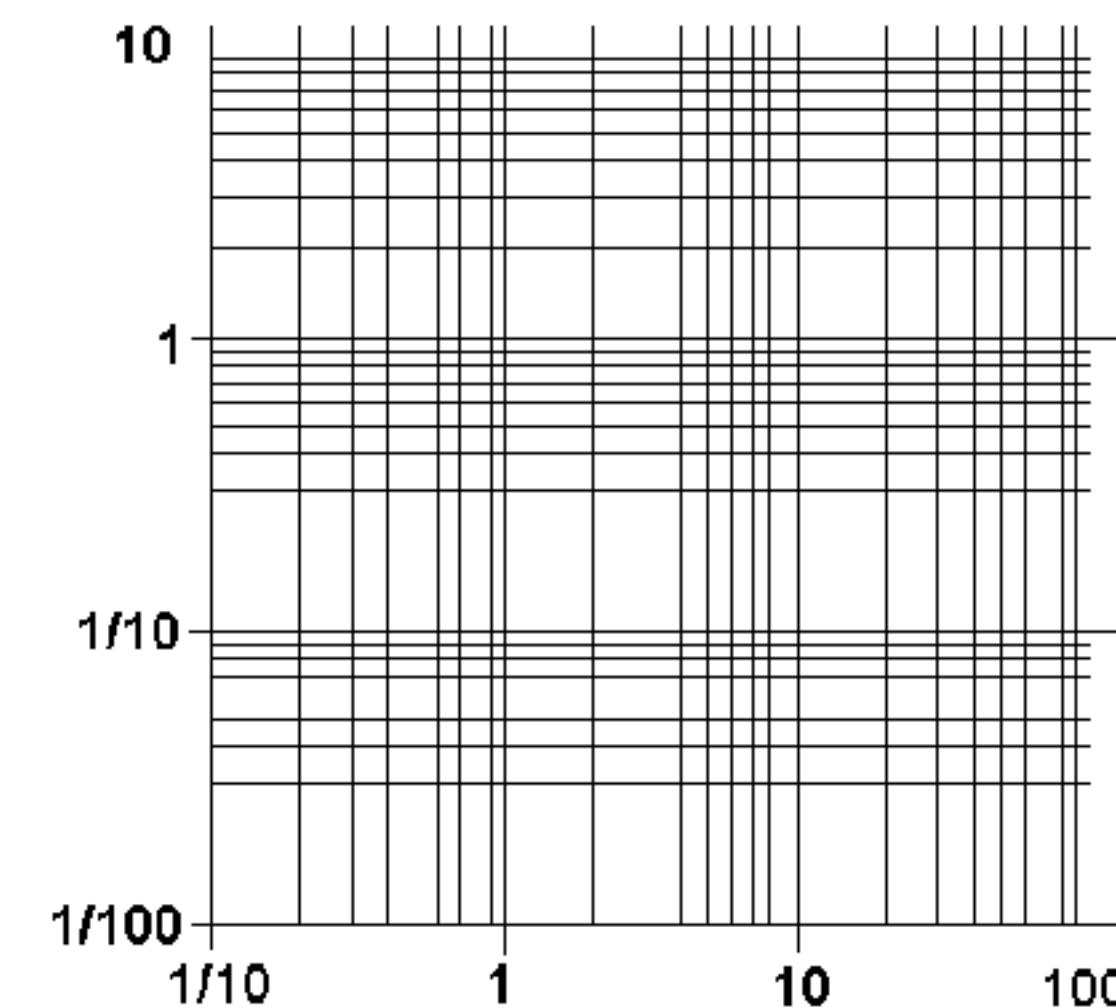
Abundance and importance of skewed data



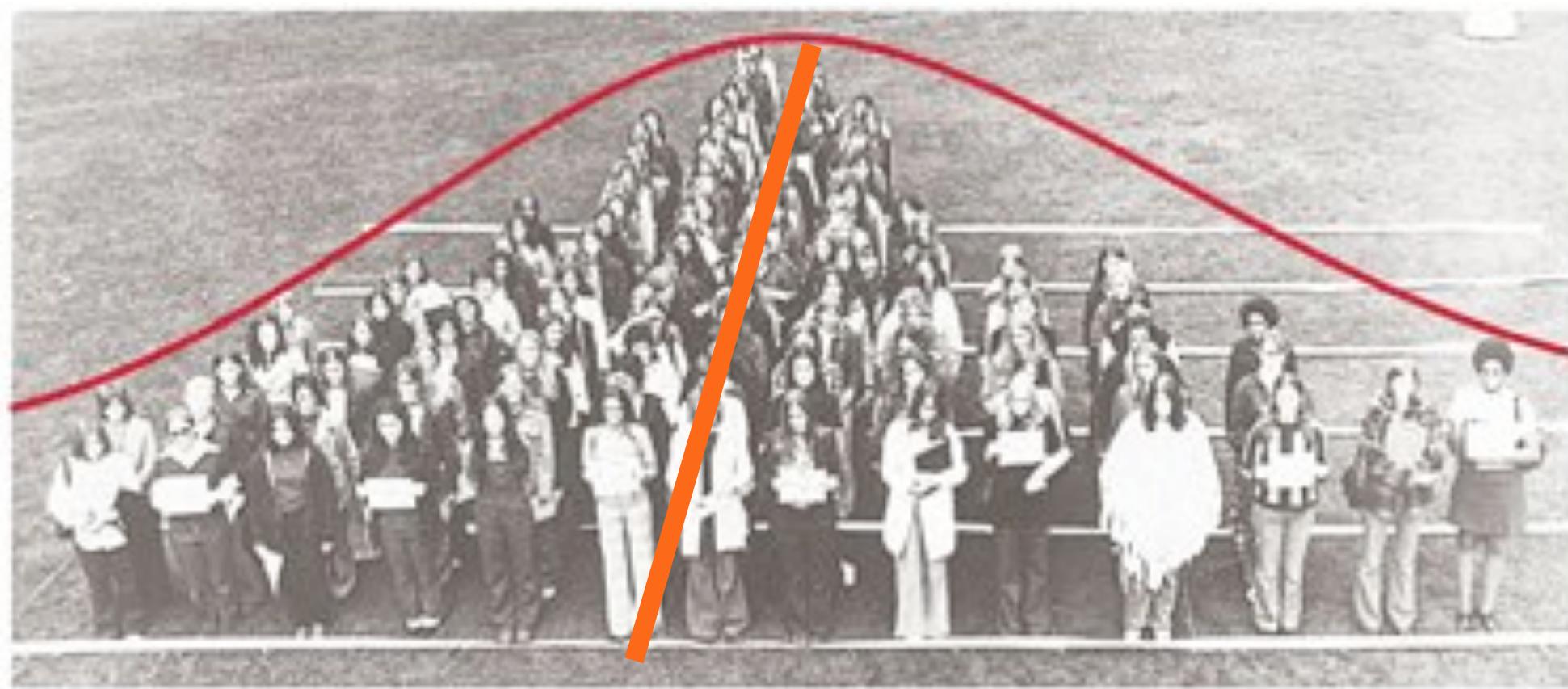
Inspecting distribution tails



Log-transformation



Mediocristan



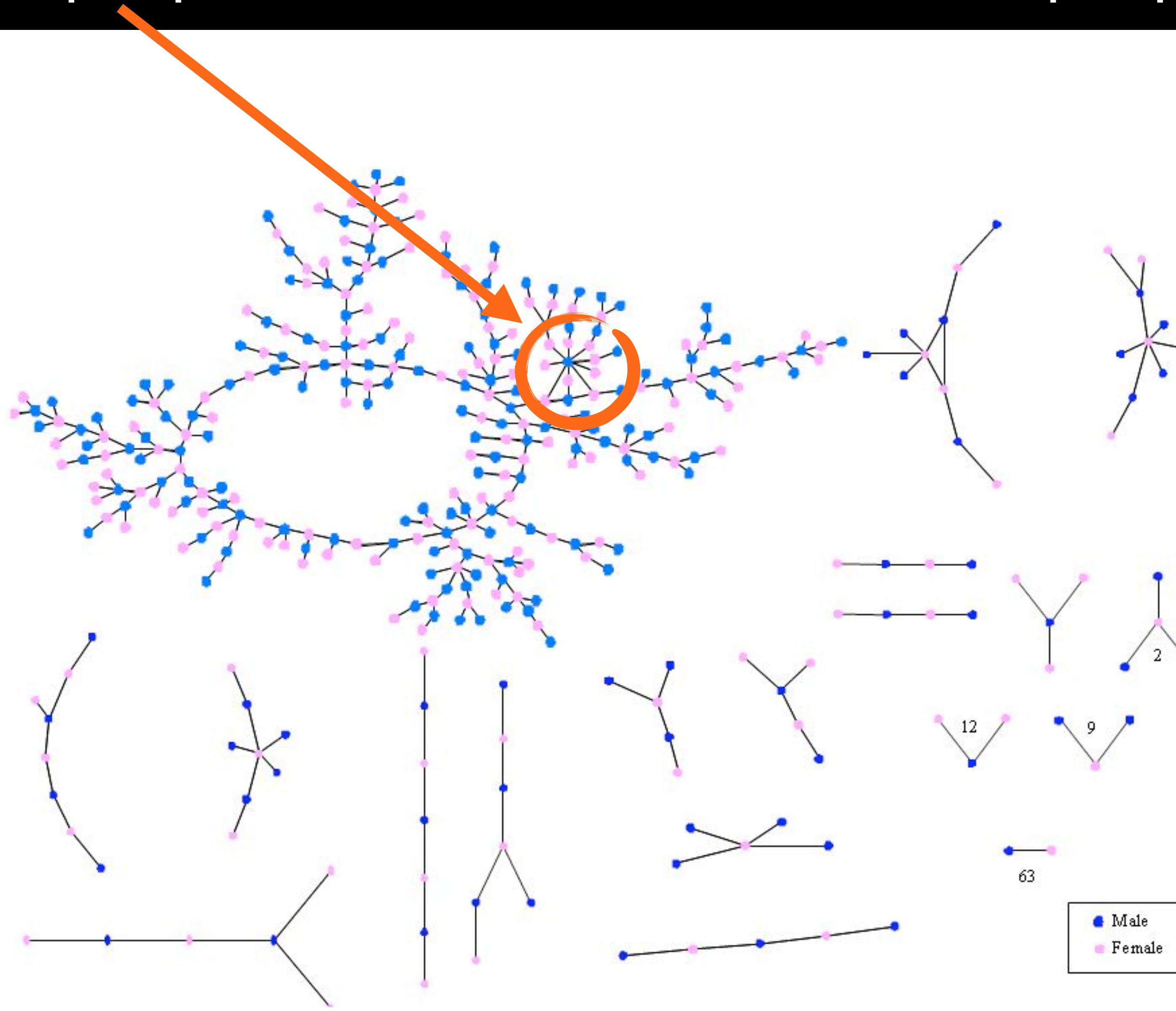
There is a **typical scale**

Extremistan

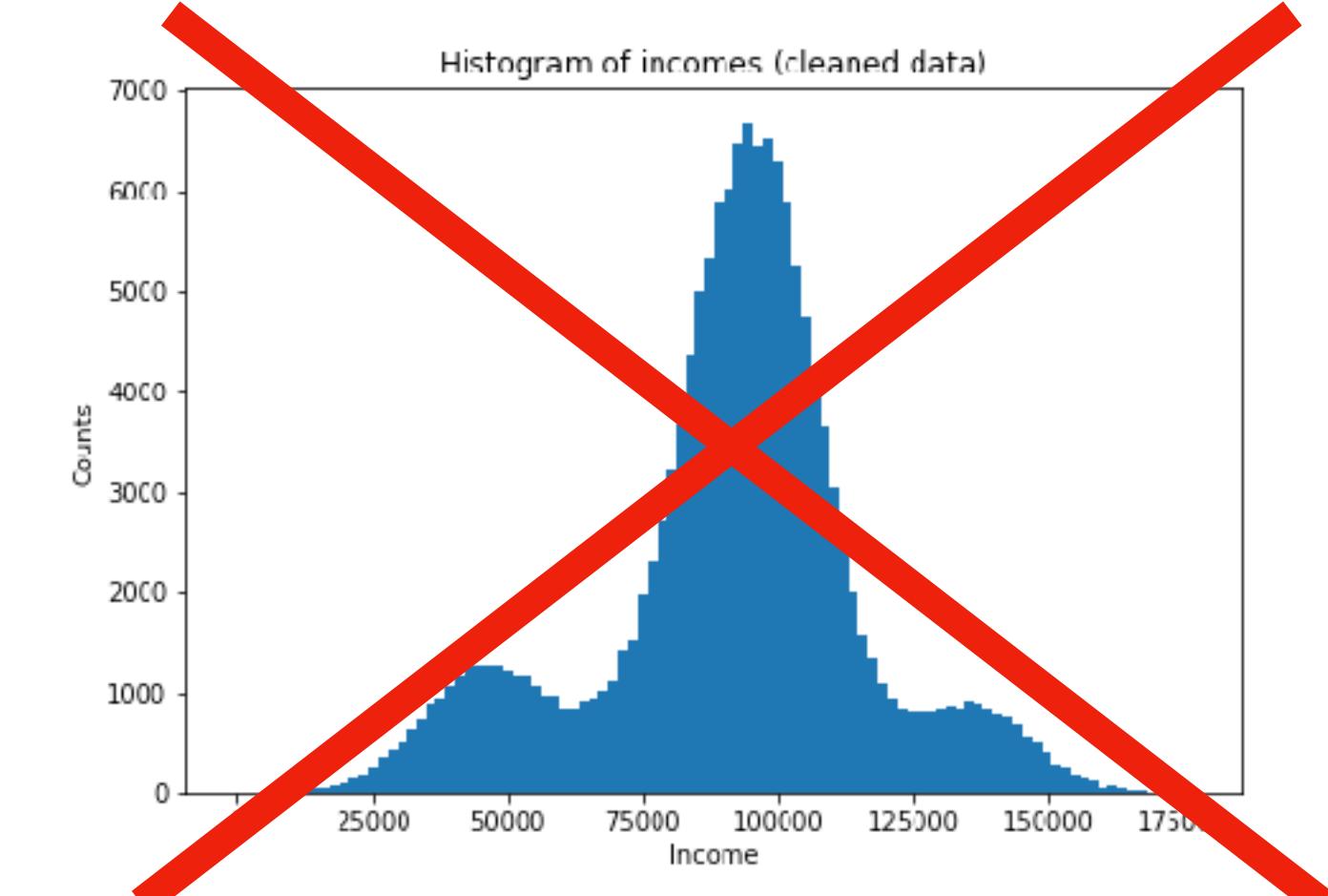
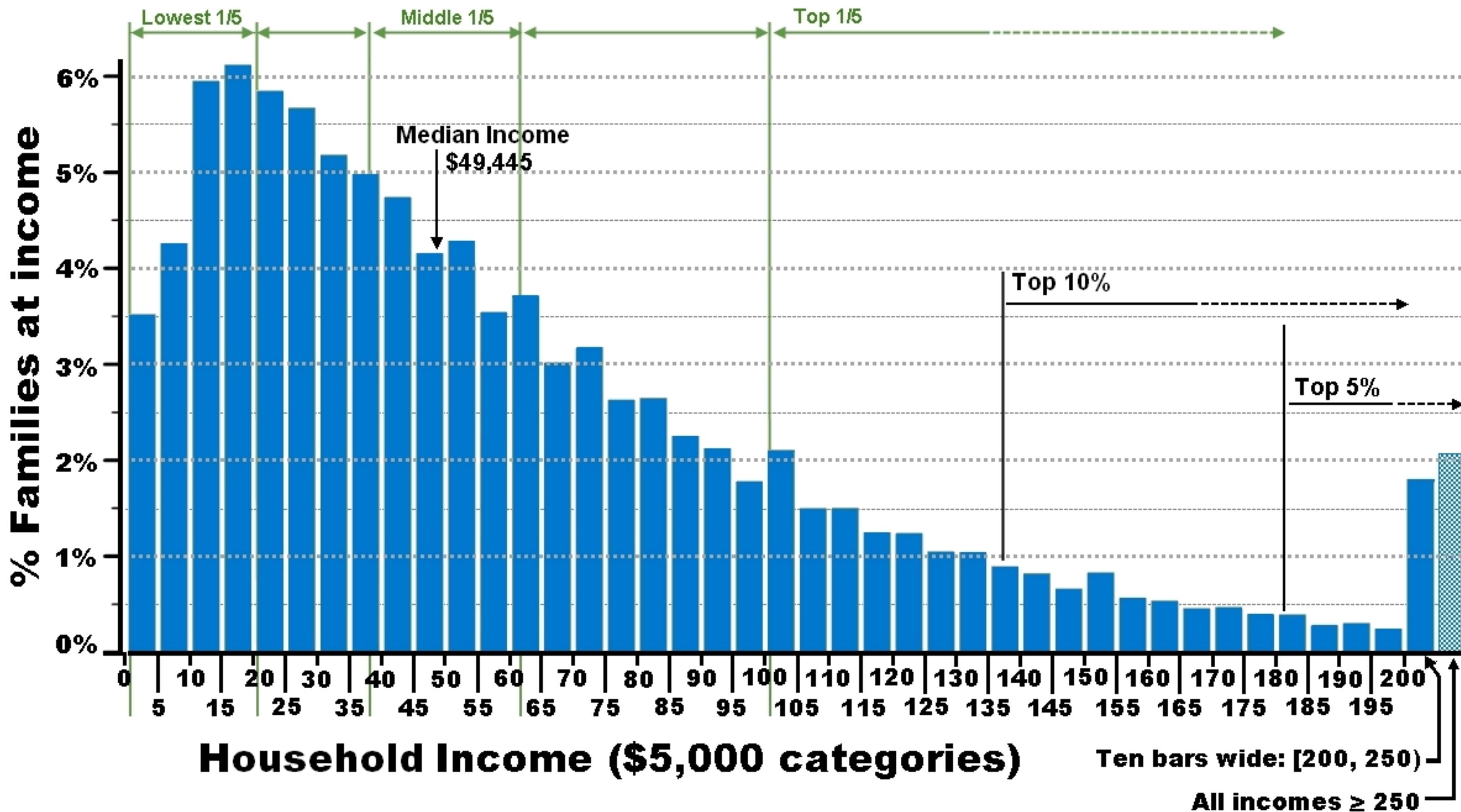


There is **no typical scale**

We live in Extremistan:
Some people have sex with MANY more people than others

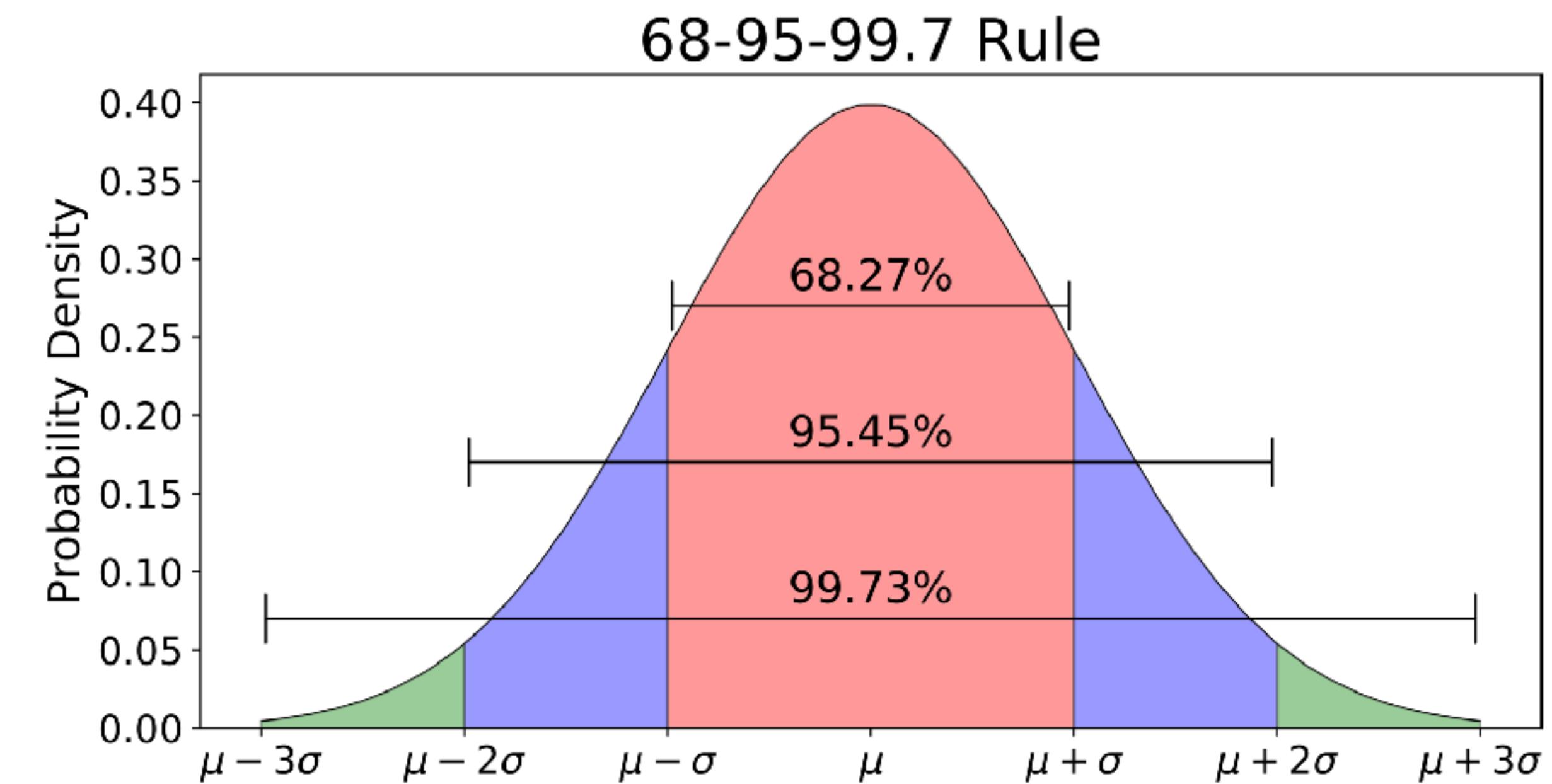


We live in Extremistan: Some people earn MUCH more than others



Data source: http://www.census.gov/hhes/www/cpstables/032011/hhinc/new06_000.htm

We live in Extremistan: Financial collapse is MUCH more likely than expected



" 10σ events" happen regularly

We live in Extremistan: Financial collapse is MUCH more likely than expected



Range	Expected fraction of population inside range	Approximate expected frequency outside range		Approximate frequency for daily event
$\mu \pm 0.5\sigma$	0.382 924 922 548 026	2 in	3	Four or five times a week
$\mu \pm \sigma$	0.682 689 492 137 086	1 in	3	Twice a week
$\mu \pm 1.5\sigma$	0.866 385 597 462 284	1 in	7	Weekly
$\mu \pm 2\sigma$	0.954 499 736 103 642	1 in	22	Every three weeks
$\mu \pm 2.5\sigma$	0.987 580 669 348 448	1 in	81	Quarterly
$\mu \pm 3\sigma$	0.997 300 203 936 740	1 in	370	Yearly
$\mu \pm 3.5\sigma$	0.999 534 741 841 929	1 in	2149	Every 6 years
$\mu \pm 4\sigma$	0.999 936 657 516 334	1 in	15 787	Every 43 years (twice in a lifetime)
$\mu \pm 4.5\sigma$	0.999 993 204 653 751	1 in	147 160	Every 403 years (once in the modern era)
$\mu \pm 5\sigma$	0.999 999 426 696 856	1 in	1 744 278	Every 4776 years (once in recorded history)
$\mu \pm 5.5\sigma$	0.999 999 962 020 875	1 in	26 330 254	Every 72 090 years (thrice in history of modern humankind)
$\mu \pm 6\sigma$	0.999 999 998 026 825	1 in	506 797 346	Every 1.38 million years (twice in history of humankind)
$\mu \pm 6.5\sigma$	0.999 999 999 919 680	1 in	12 450 197 393	Every 34 million years (twice since the extinction of dinosaurs)
$\mu \pm 7\sigma$	0.999 999 999 997 440	1 in	390 682 215 445	Every 1.07 billion years (four times in history of Earth)

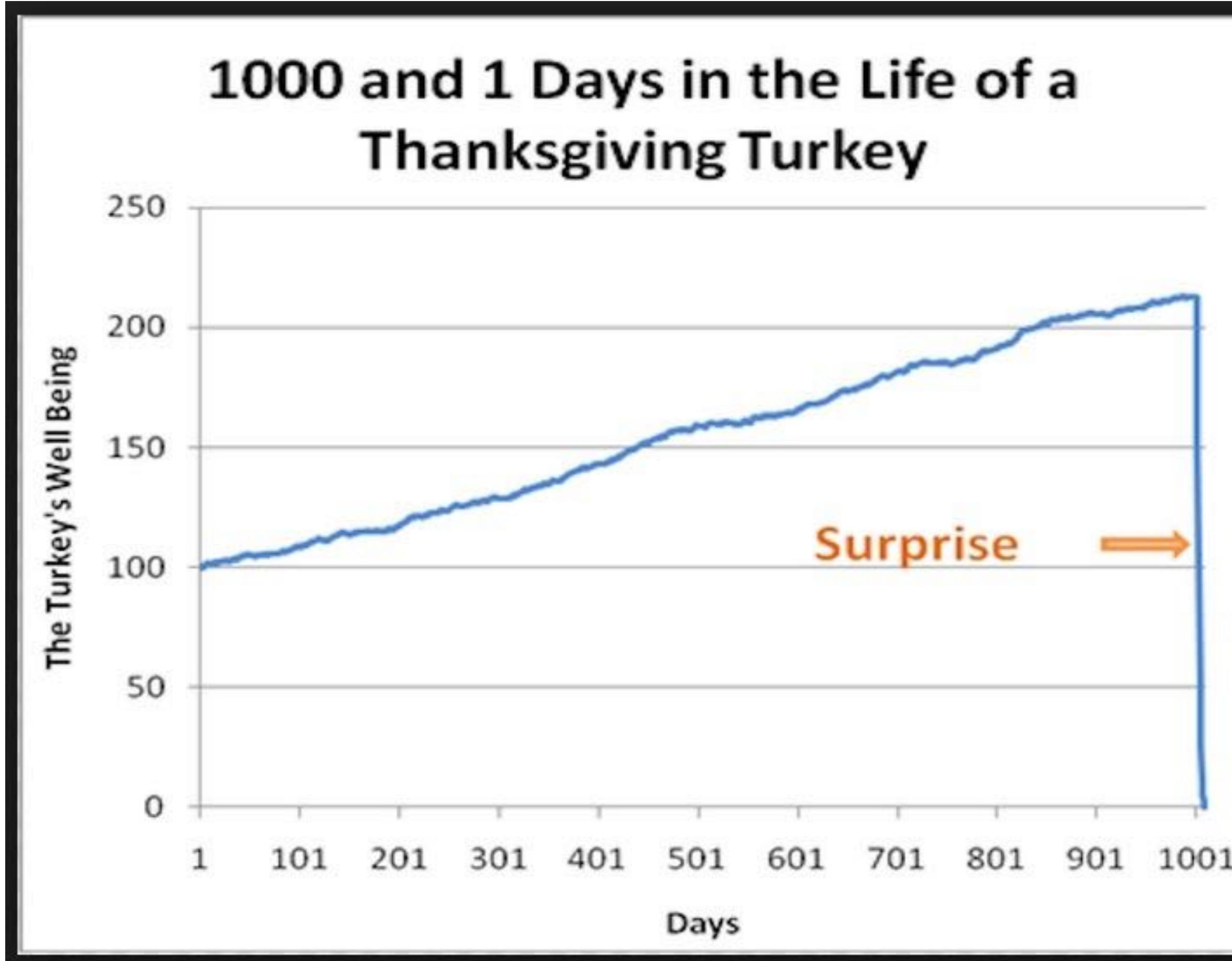
" 10σ events" happen regularly

We live in Extremistan:
Black swans are MUCH more likely than expected

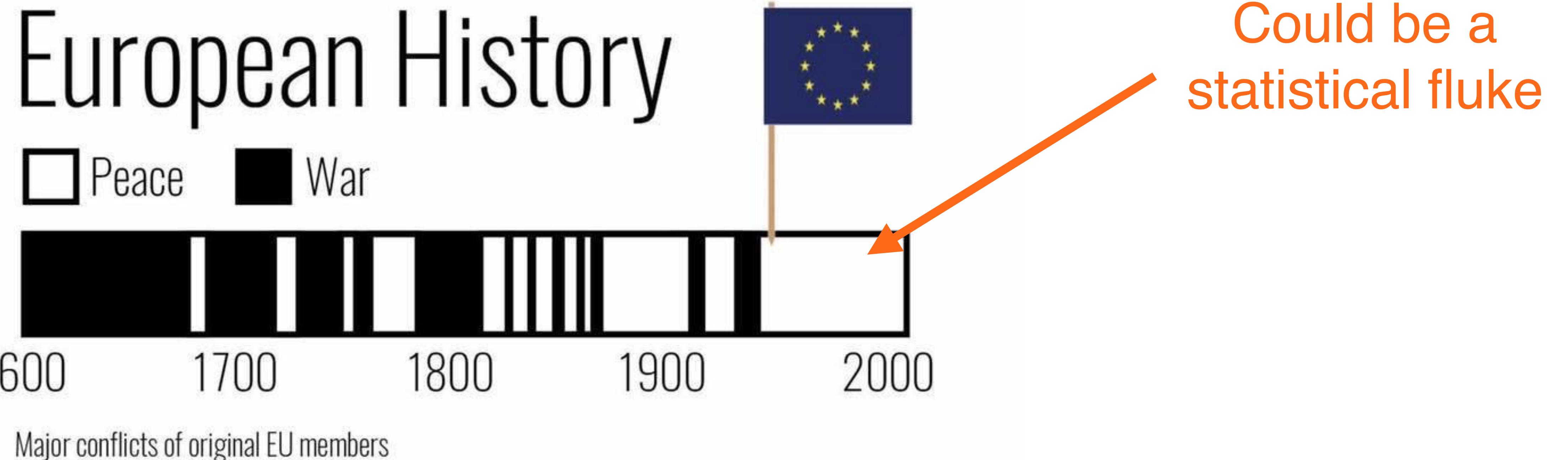


N.N.Taleb: The Black Swan

Black swans are rare, major impacts, rationalized by hindsight. We do not believe they happen until they do.



We live in Extremistan: Wars are more likely than expected



Extremistan produces **skewed data**:
It covers a **VERY** large range of values

Mediocristan

Tallest person: 272 cm

Shortest person: 56 cm

ratio: 4.8

Extremistan

Casanova: 20000

Monk: 1

ratio: 20,000



Mediocristan

Tallest person: 272 cm

Shortest person: 56 cm

ratio: 4.8

Extremistan

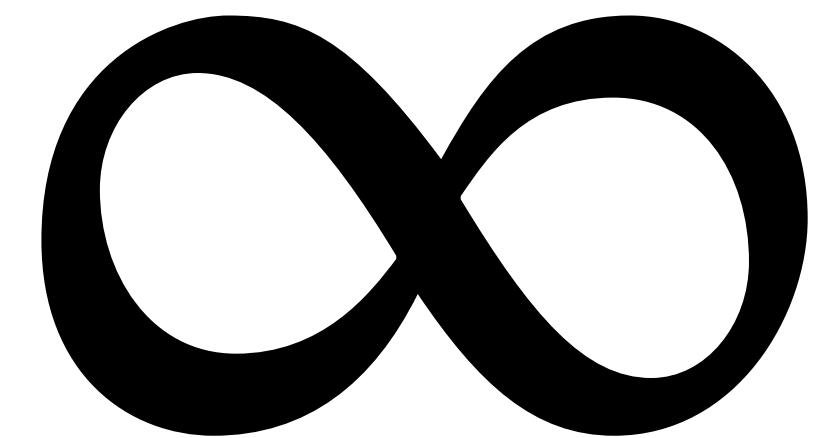
Biggest city: 8,000,000

Smallest city: 52

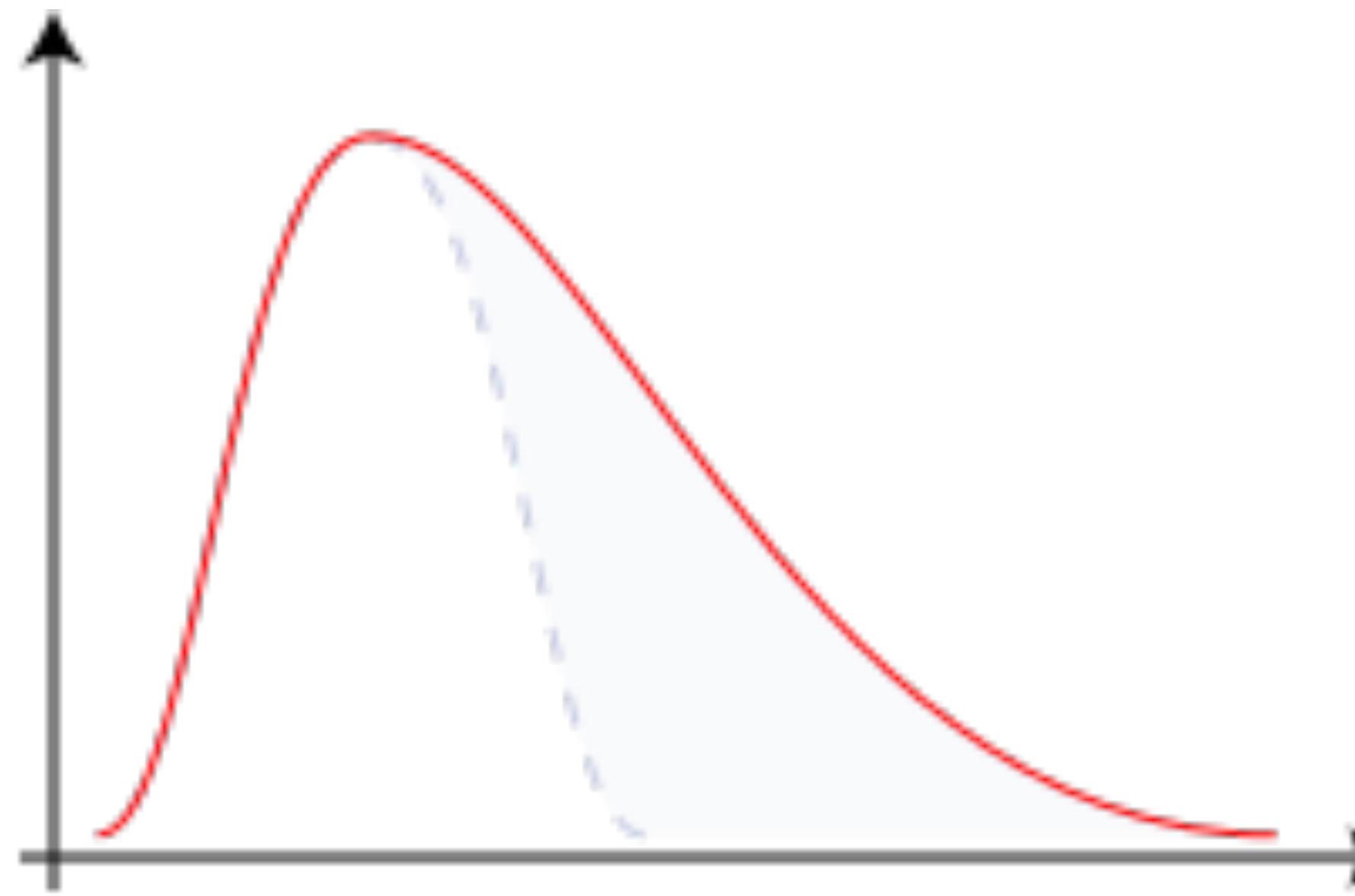
ratio: 150,000



The theoretical spread and mean
in skewed data is often:



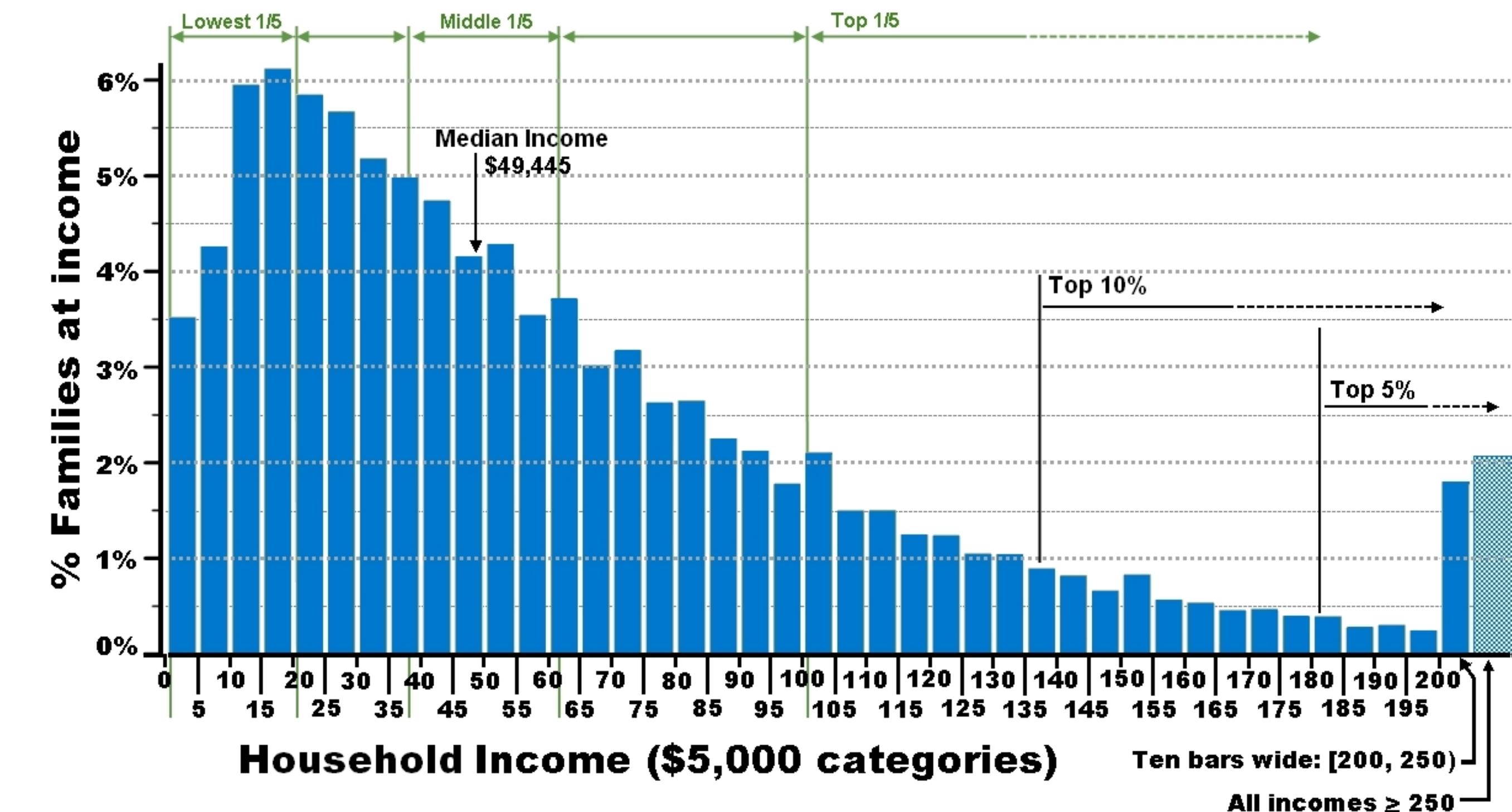
Skewed data means: Mean and median are very different



Positive Skew

Right skew

Long tail to the right



Data source: http://www.census.gov/hhes/www/cpstables/032011/hhinc/new06_000.htm

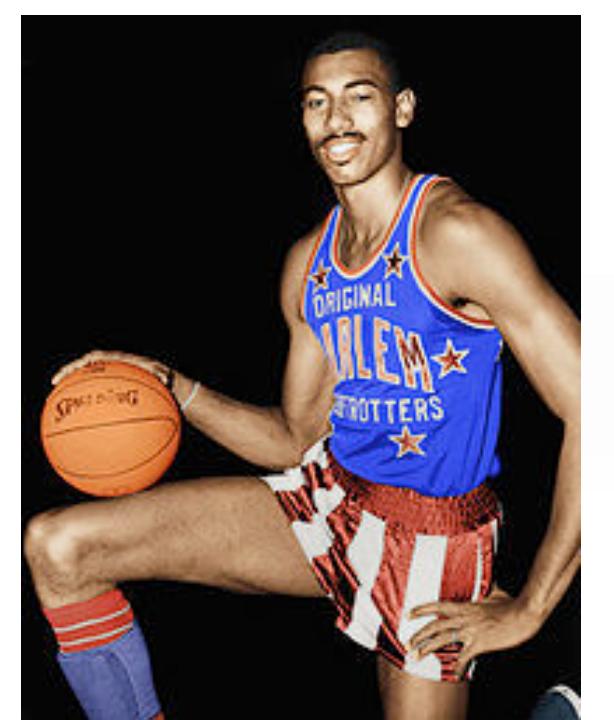
Ten bars wide: [200, 250]

All incomes ≥ 250

Skewed data has a long tail



Skewed data has a long tail



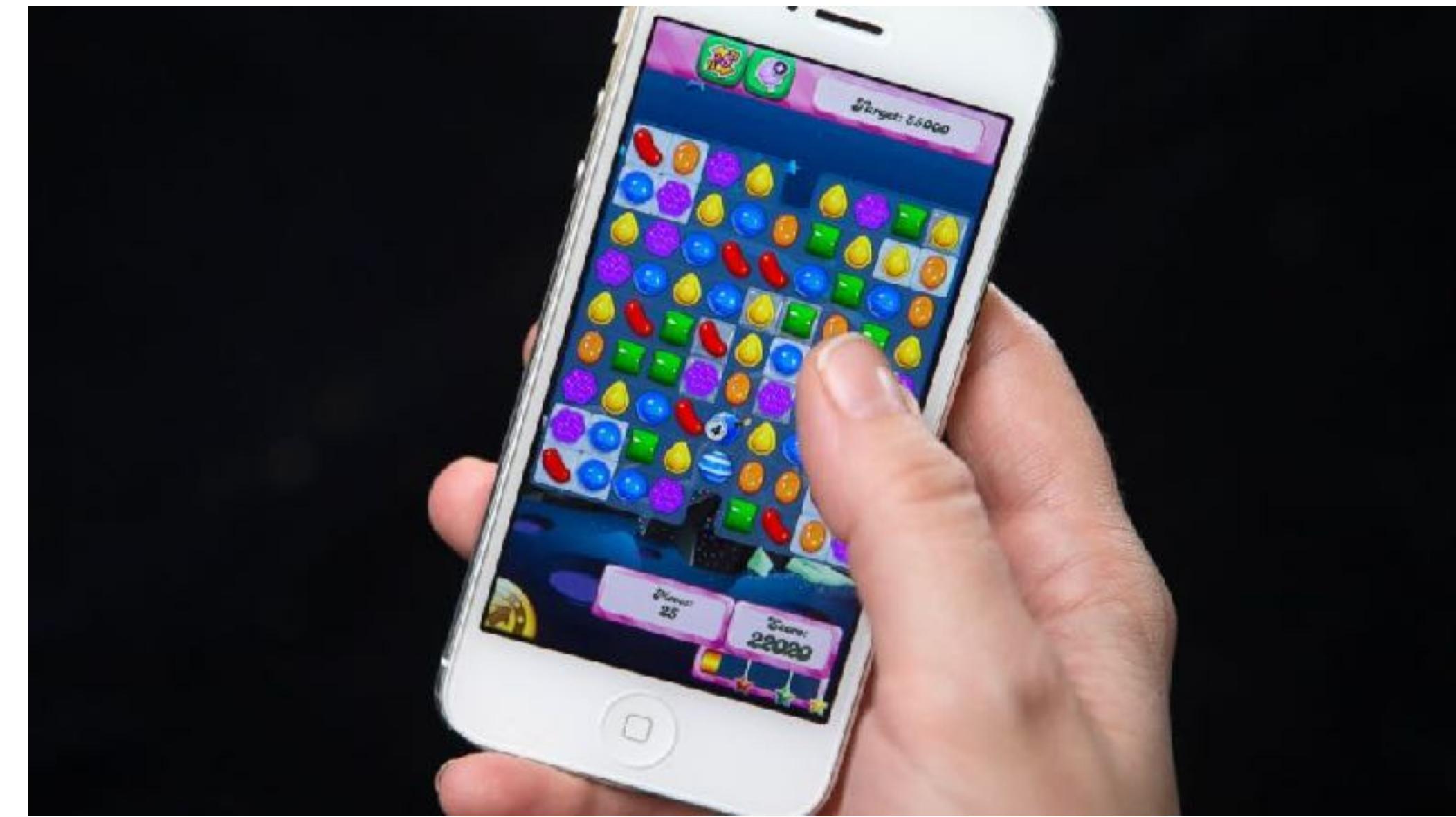
The abundance of long tails has
a fundamental business impact

The pareto principle states:

80% of the effects come from 20% of the causes



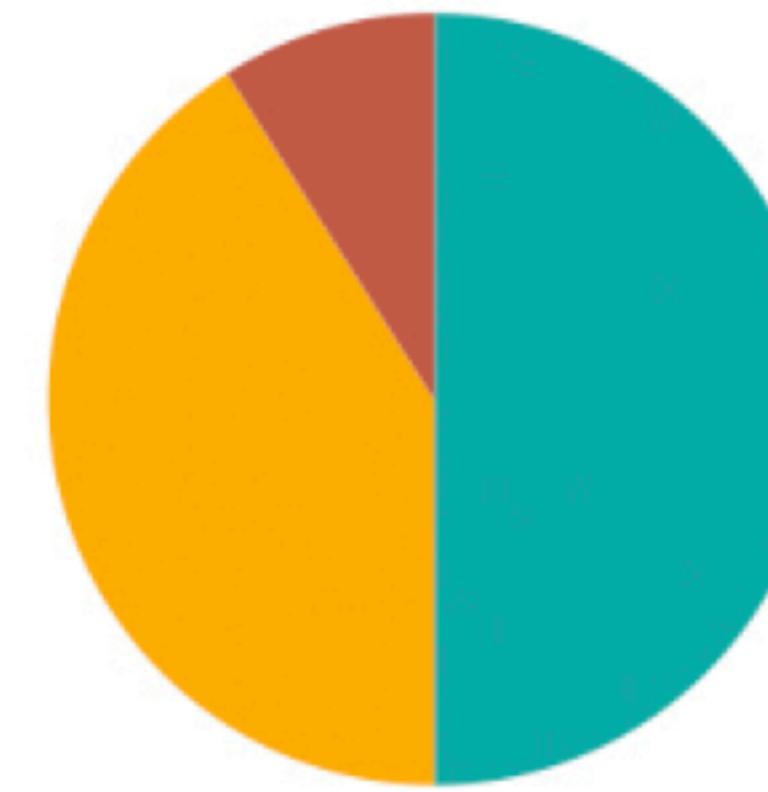
The pareto principle states:
80% of the effects come from 20% of the causes



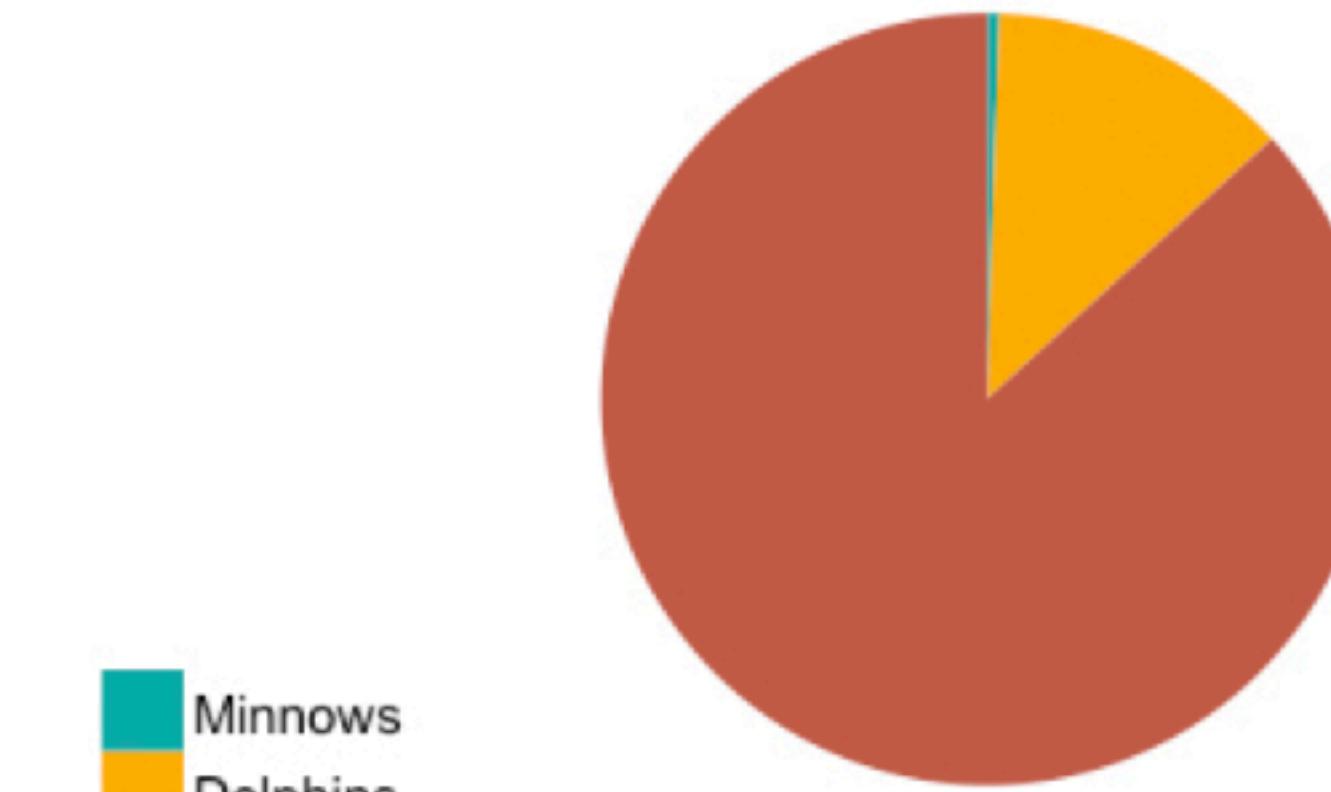
80% of sales come from 20% of clients

The pareto principle states:
80% of the effects come from 20% of the causes

Percentage of players

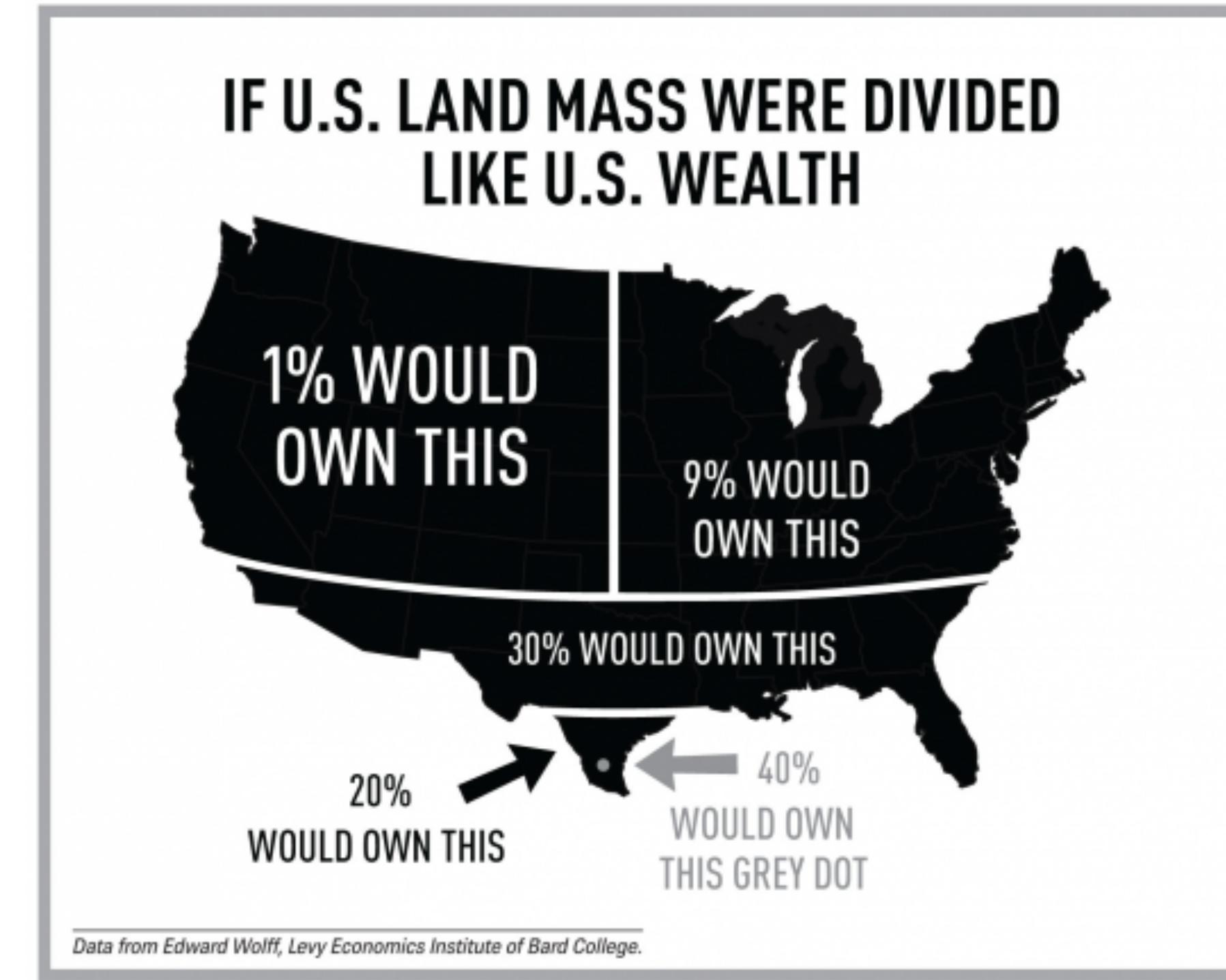


Total revenue income



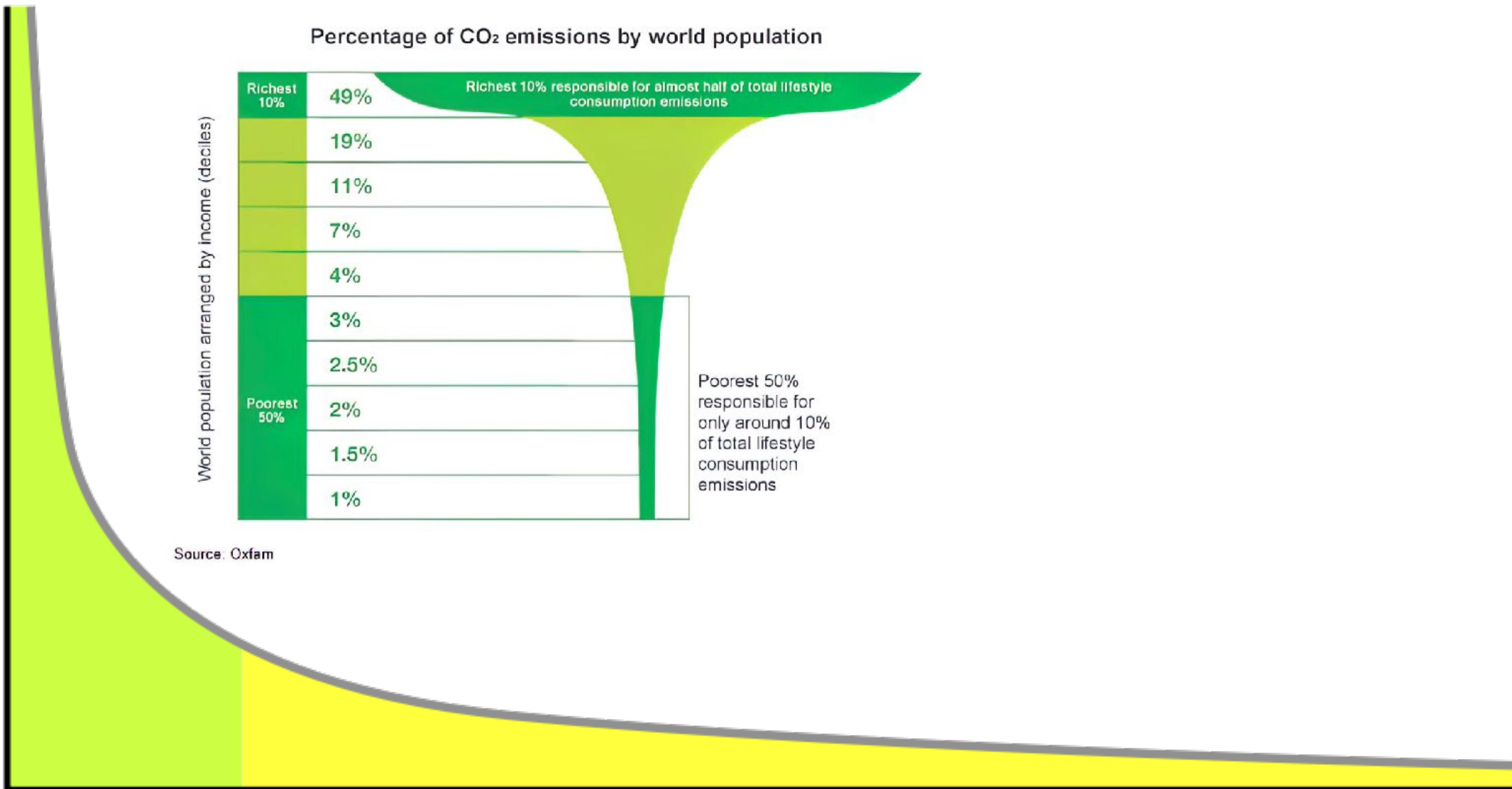
80% of sales come from 20% of clients

The pareto principle states:
80% of the effects come from 20% of the causes



80% of income goes to richest 20%
50% of income goes to richest 1%

The pareto principle states: 80% of the effects come from 20% of the causes



The pareto principle states: 80% of the effects come from 20% of the causes

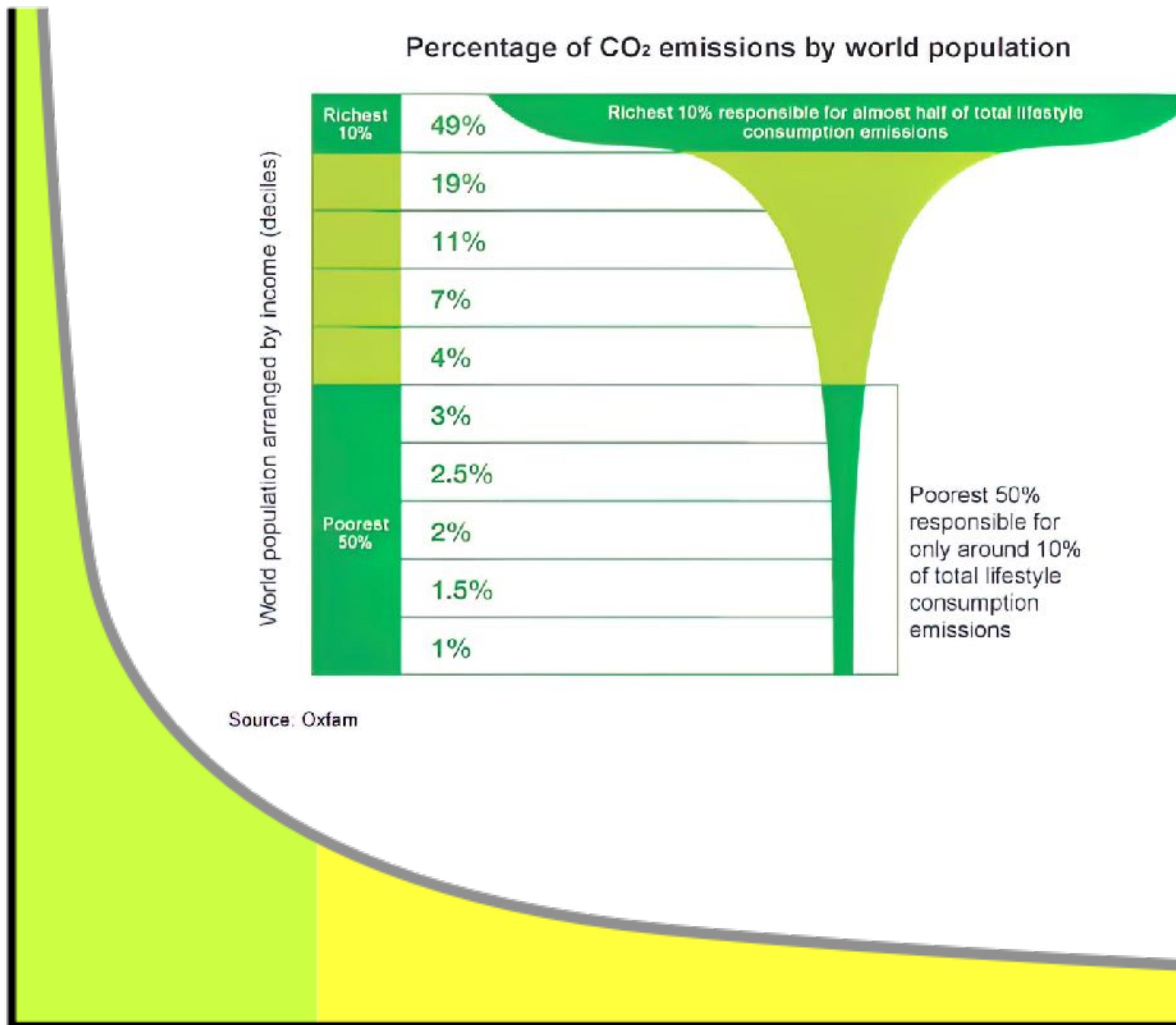
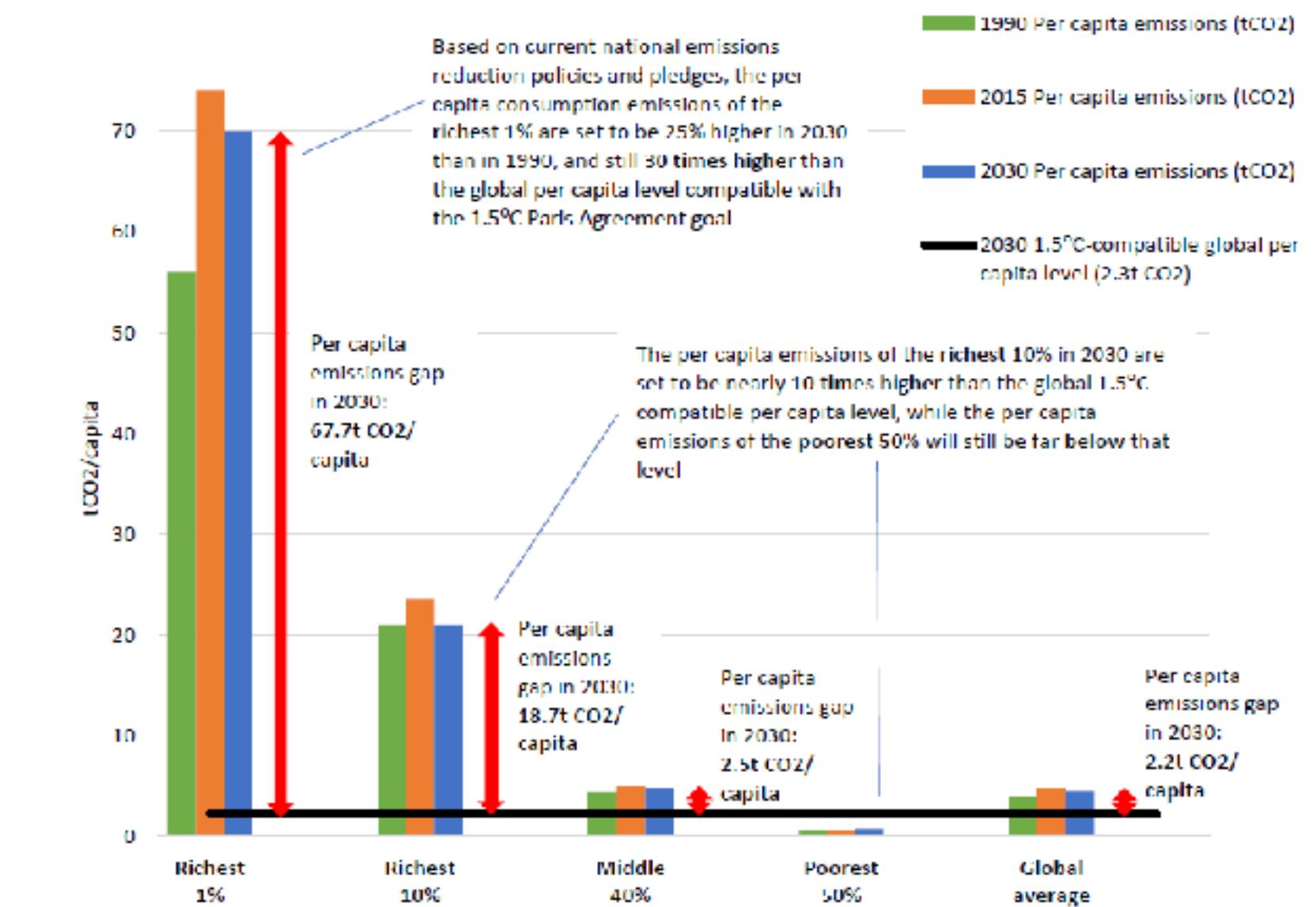


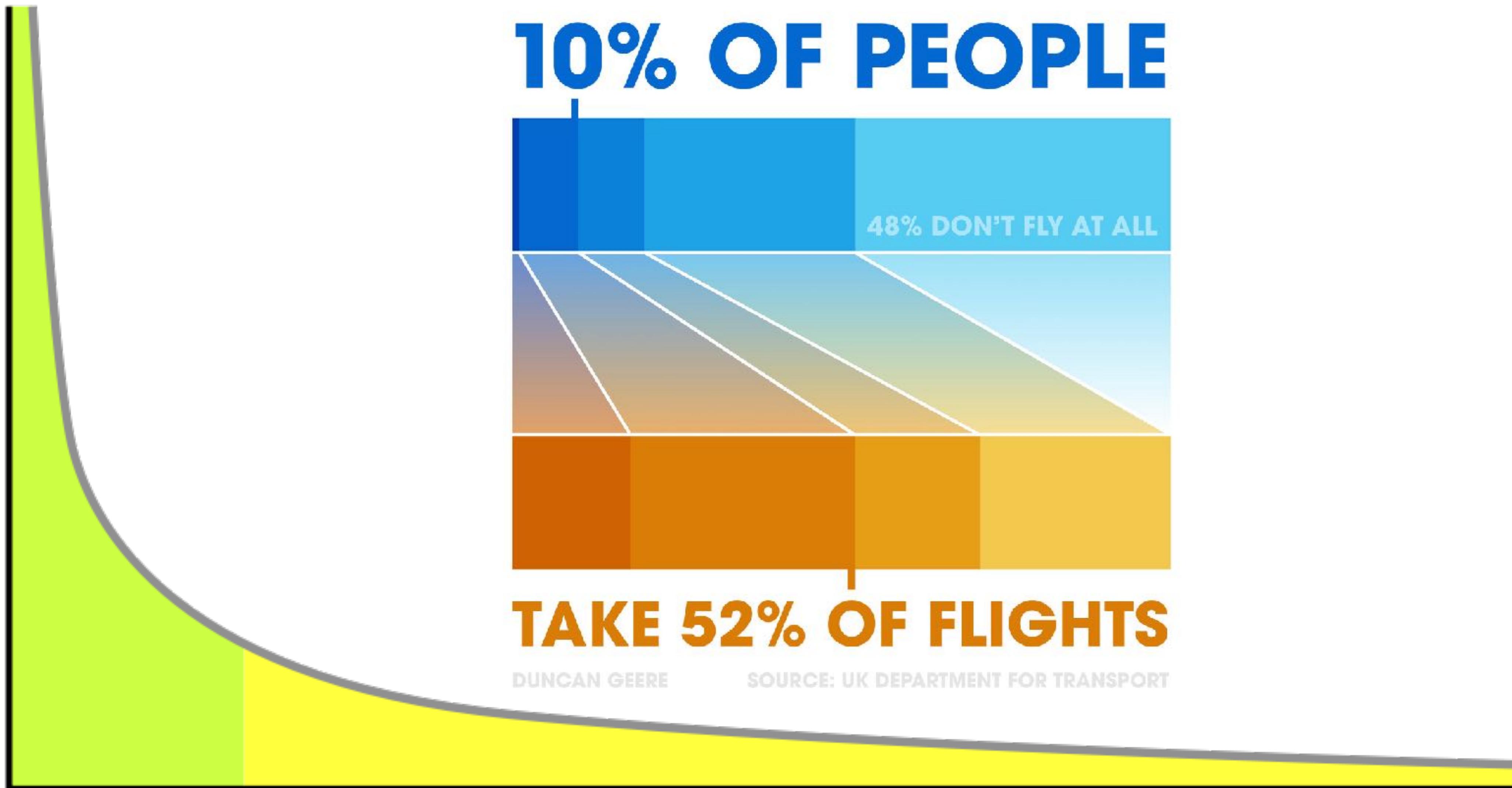
Figure 1: Per capita consumption emissions of global income groups 1990–2030 and the 2030 1.5°C-compatible global per capita goal



Source: IEA and SEI analysis. Annual income in 2030 (\$2011PPP) of richest 1%: >\$172k; richest 10%: >\$55k; middle 40%: \$0.8k; poorest 50%: <\$0.8k. Total population in 2030: c.7.9 billion.

The richest 1% emit 30x more CO₂ than compatible with 1.5°

The pareto principle states:
80% of the effects come from 20% of the causes



The pareto principle states:

80% of the effects come from 20% of the causes



80% of errors come from 20% of bugs

In the past, selection of cars was limited



Hungary 1980s

In the past, media were limited by time or space

1971

	7:30	8:00	8:30	9:00	9:30	10:00	10:30		
SAT	ABC Local	Getting Together	Movie of the Weekend		The Persuaders				
	CBS Local	All in the Family	Funny Face	New Dick Van Dyke Show	Mary Tyler Moore Show	Mission: Impossible			
	NBC Local	The Partners	The Good Life	NBC Saturday Night at the Movies					
SUN	ABC Local	The FBI		The ABC Sunday Night Movie					
	CBS	The CBS Sunday Night Movies		Cade's County		Local			
	NBC	The Wonderful World of Disney	Jimmy Stewart	Bonanza		The Bold Ones			
MON	ABC Local	Nanny and the Professor	Local	NFL Monday Night Football (to be replaced by movies after Jan. 24)					
	CBS Local	Gunsmoke		Here's Lucy	Doris Day Show	My Three Sons	Arnie		
	NBC Local	Rowan and Martin's Laugh-In		NBC Monday Night at the Movies					
TUE	ABC	The Mod Squad		Movie of the Week	Marcus Welby, M.D.				
	CBS	Glen Campbell Goodtime Hour		Hawaii Five-O	Cannon		Local		
	NBC	Ironside	Sarge		The Funny Side		Local		
WED	ABC Local	Bewitched	Eddie's Father	The Smith Family	Shirley's World	The Man and the City			
	CBS Local	Carol Burnett Show		Medical Center		Mannix			
	NBC Local	Adam-12	NBC Mystery Movie		Night Gallery				
THU	ABC Local	Alias Smith and Jones		Longstreet	Owen Marshall: Counselor at Law				
	CBS Local	Bearcats!		The CBS Thursday Night Movies/CBS Reports (once a month)					
	NBC Local	Flip Wilson Show		Nichols	Dean Martin Show				
FRI	ABC Local	The Brady Bunch	The Partridge Family	Room 222	The Odd Couple	Love, American Style			
	CBS Local	The Chicago Teddy Bears	O'Hara, United States Treasury		The New CBS Friday Night Movies				
	NBC Local	The D.A.	NBC World Premiere Movie/Chronolog (once a month)		Local				



In the past, retail of goods was limited by shelf space

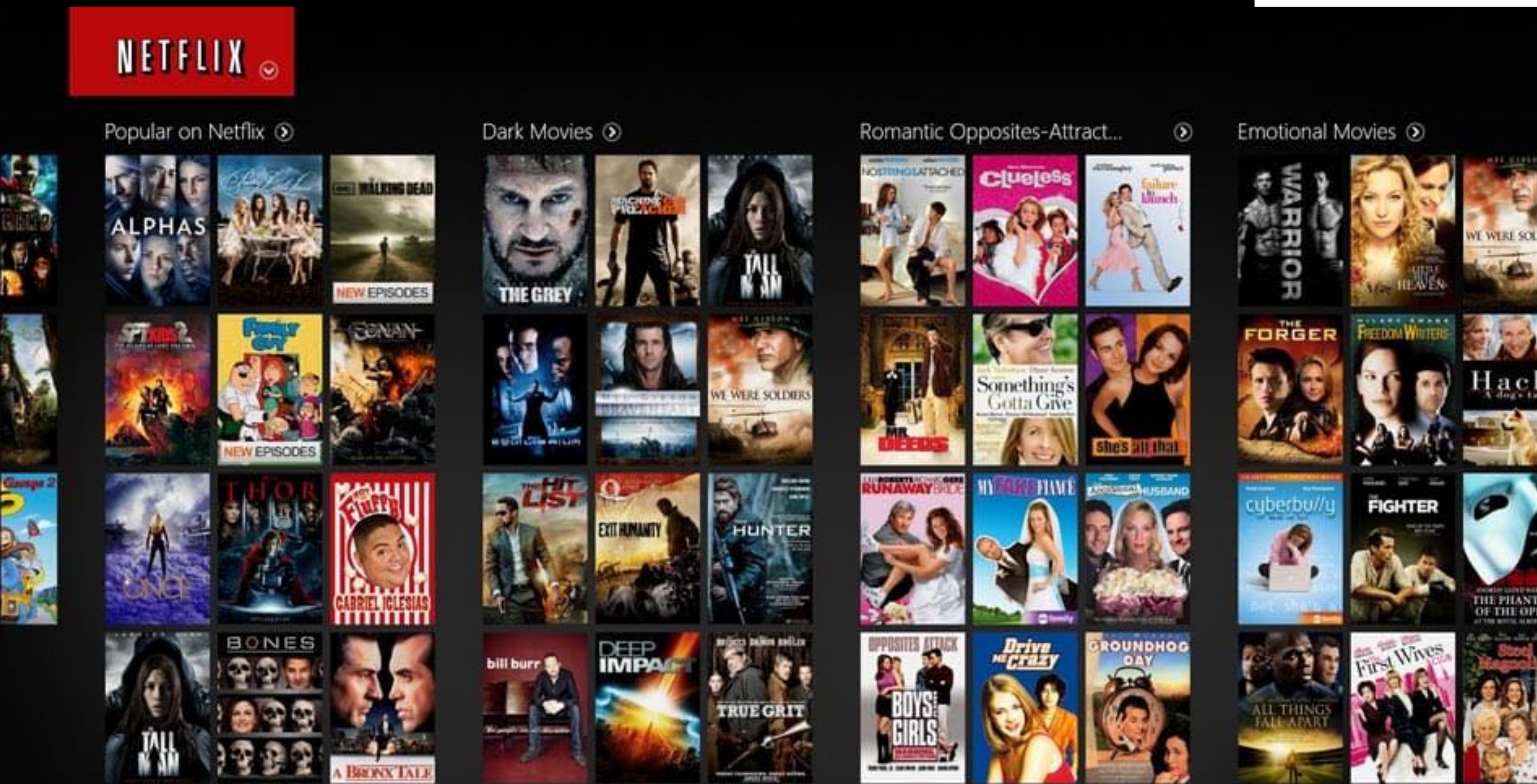


Are people happy with a limited choice?

When diversity of supply increases, there is a market for almost everything



Digital media allow unlimited shelf space



Digital media allow unlimited shelf space

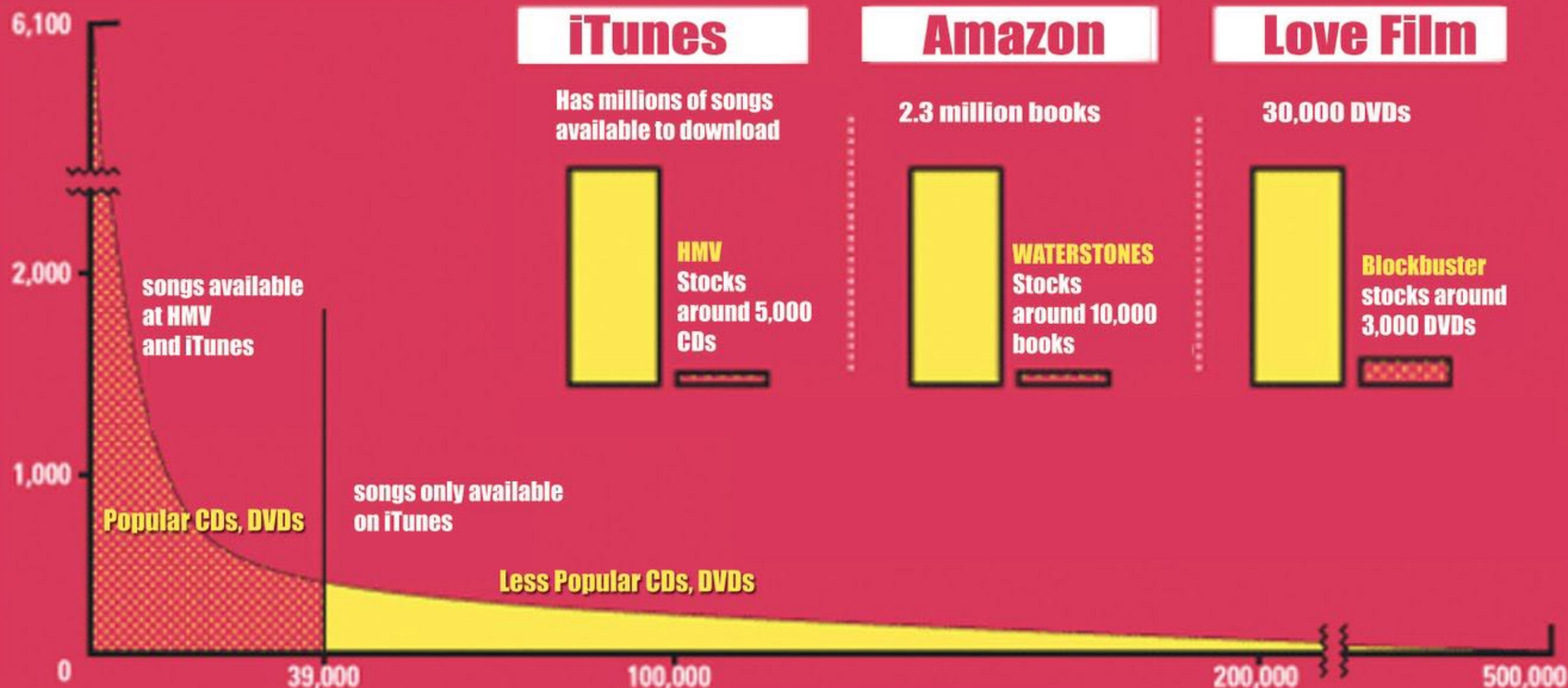


This is why Amazon started and succeeded with books



<https://www.youtube.com/watch?v=rWRbTnE1PEM>

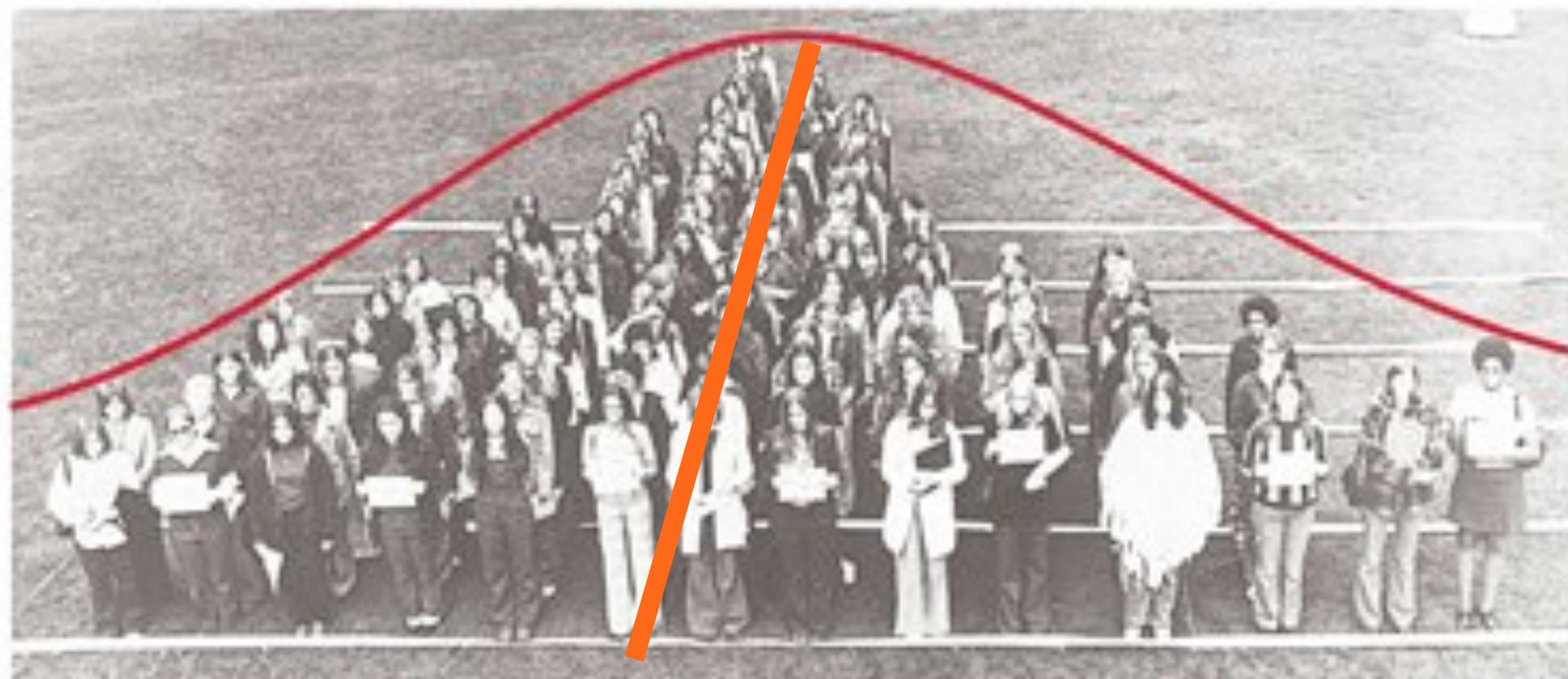
Supply becomes unlimited. Demand becomes long-tailed.



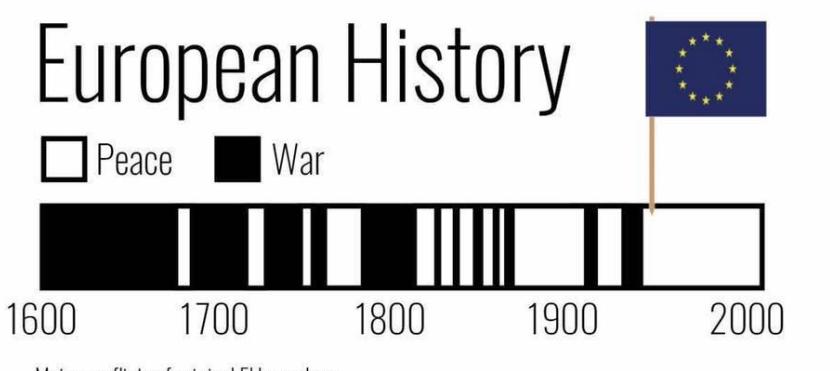
The long tail of demand poses a new challenge:
How to best **recommend** products to people they will like?



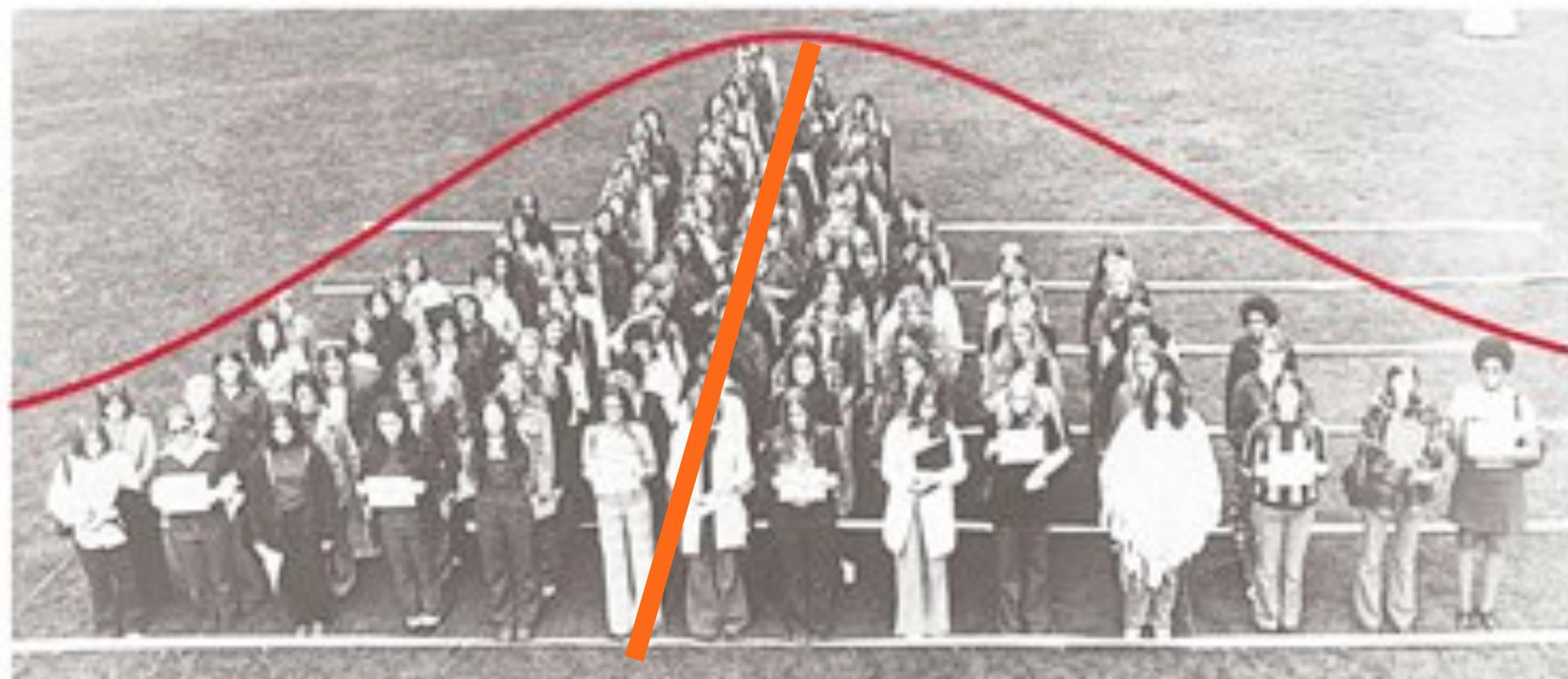
Mediocristan



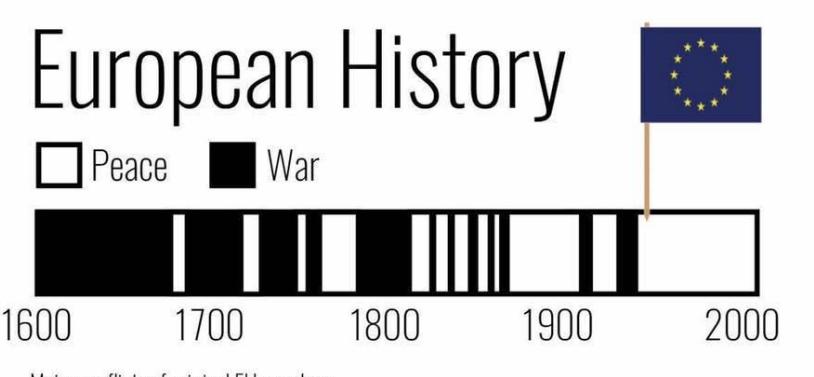
Extremistan



Mediocristan



Extremistan

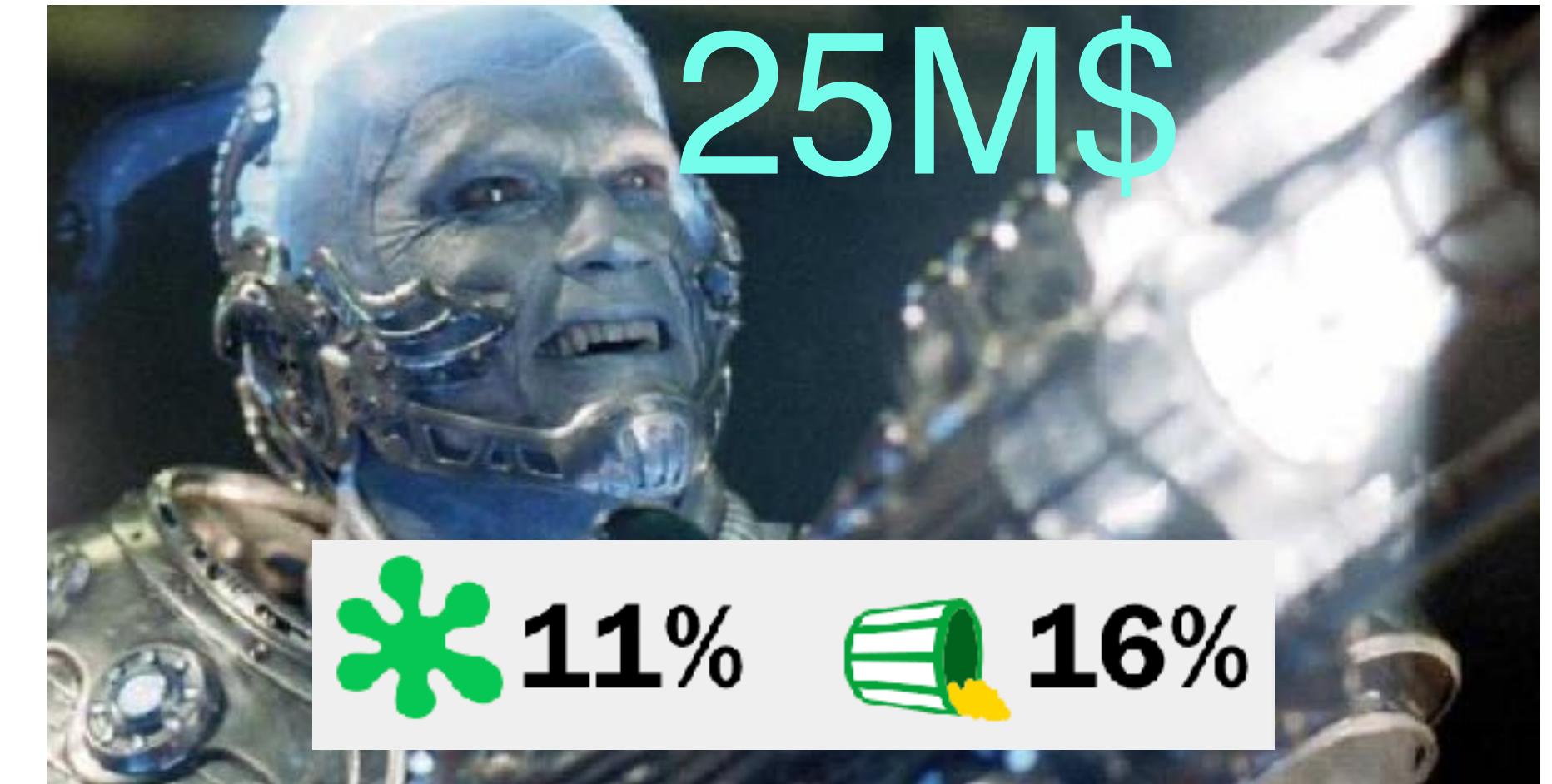


Mediocristan



Low risk, low reward

Extremistan

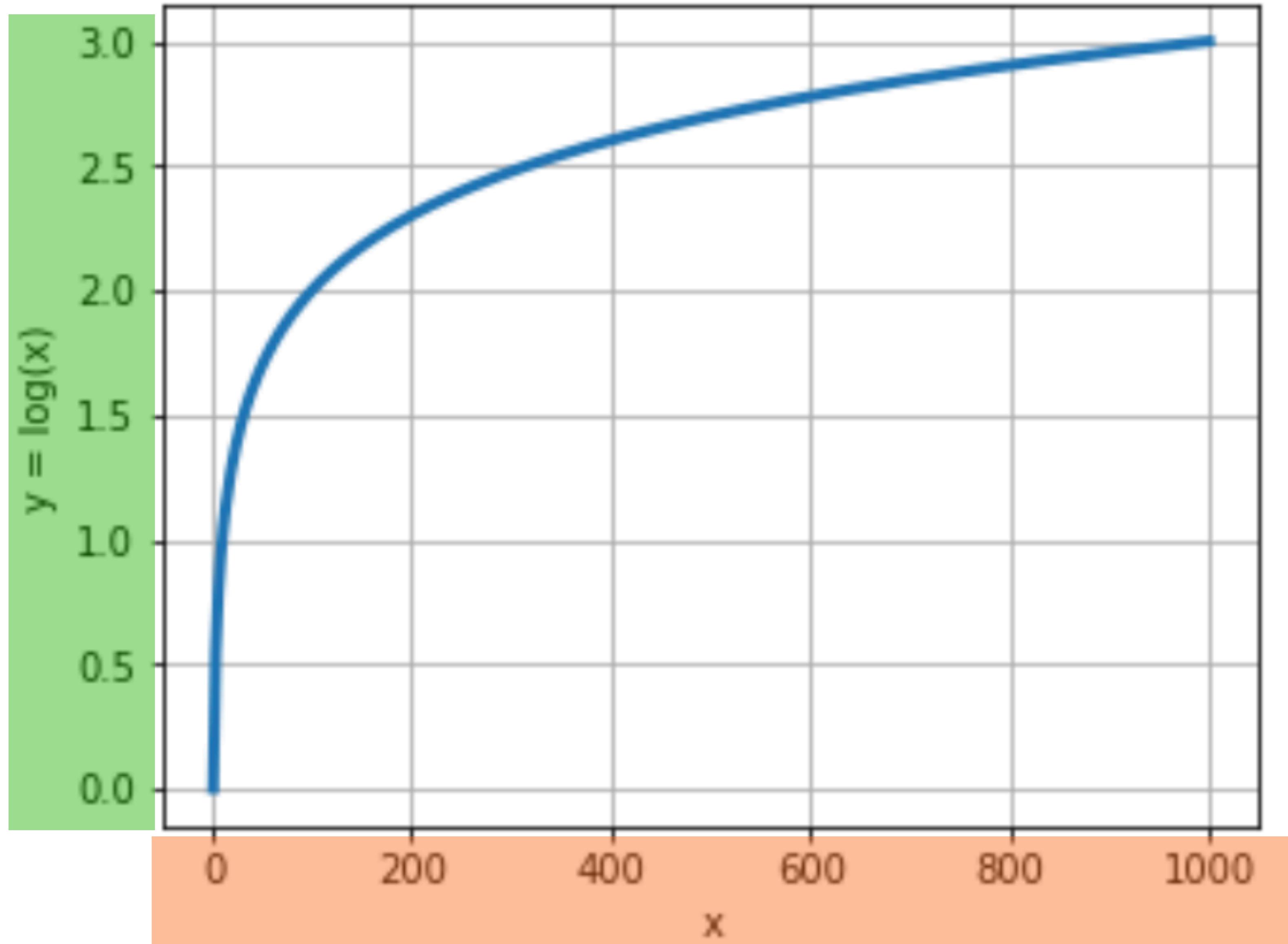


High risk, high reward

Tools to deal with skewed data

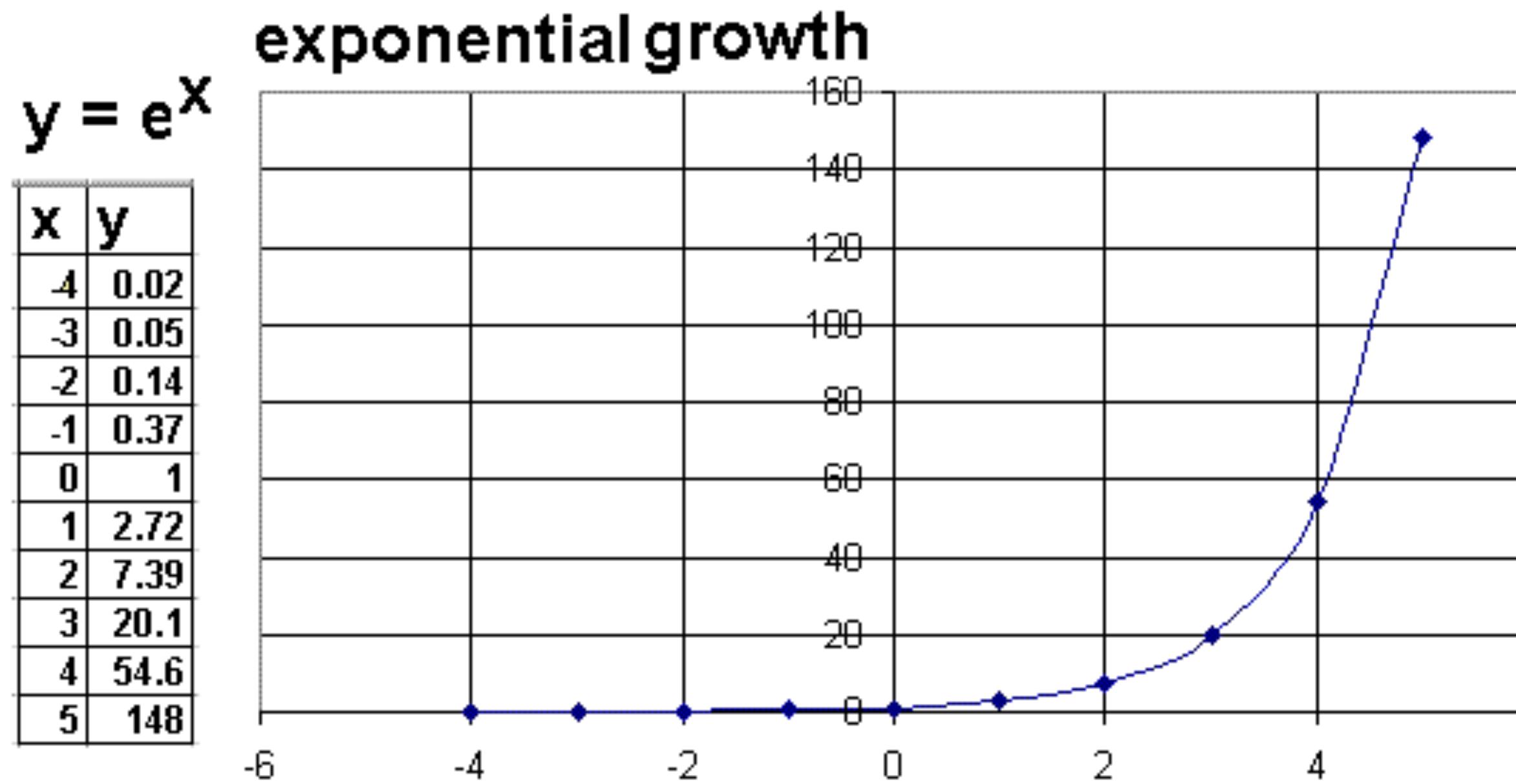
1) Logarithm transformation

The logarithm squashes a **large range of values** onto a **small range**



The log(arithm) transformation turns a linear into a log scale

$$y = ab^x$$

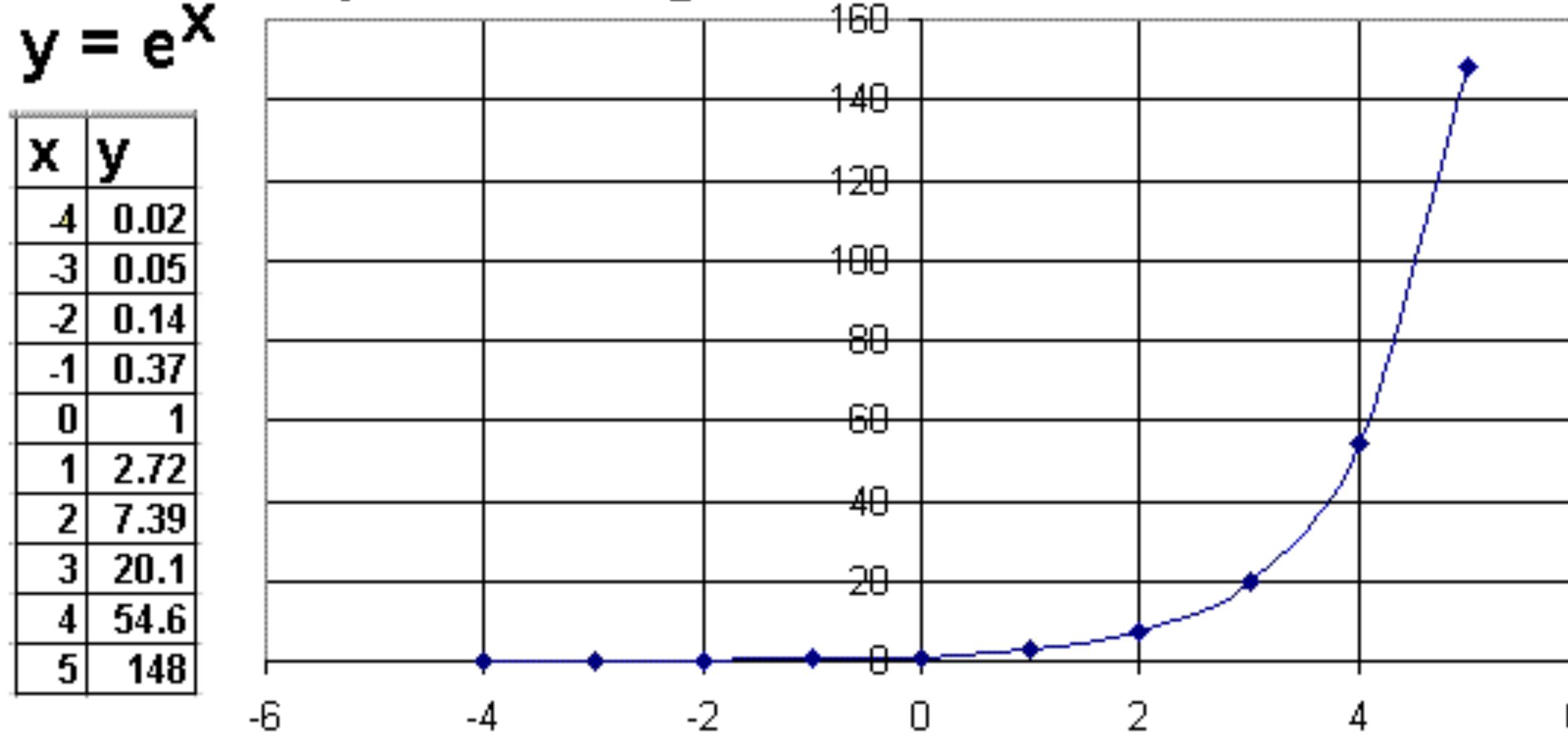


linear scales

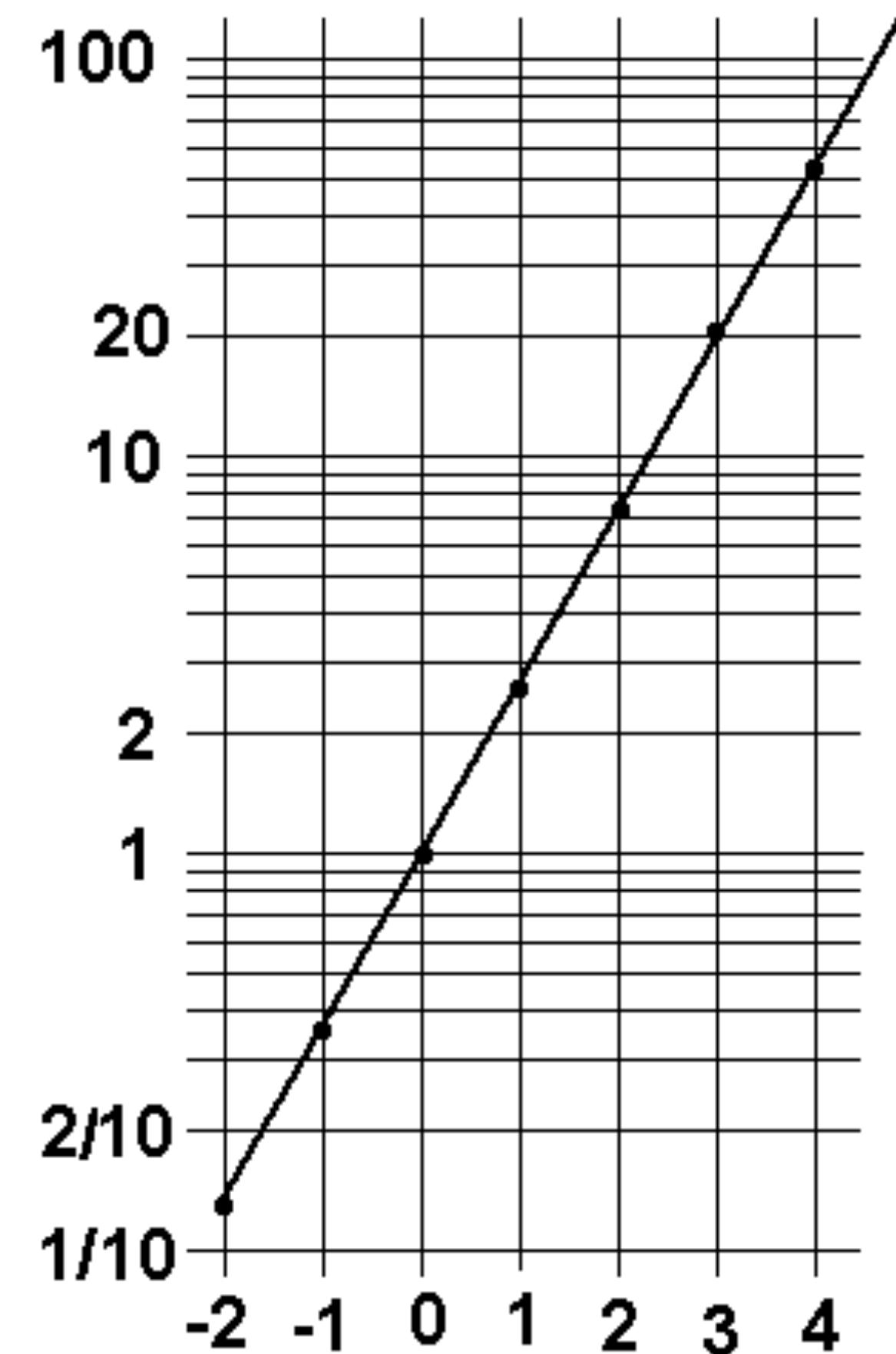
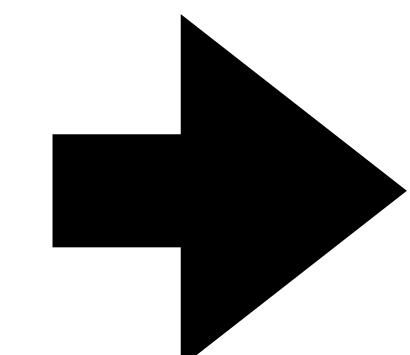
The log(arithm) transformation turns a linear into a log scale

$$y = ab^x$$

exponential growth



linear scales

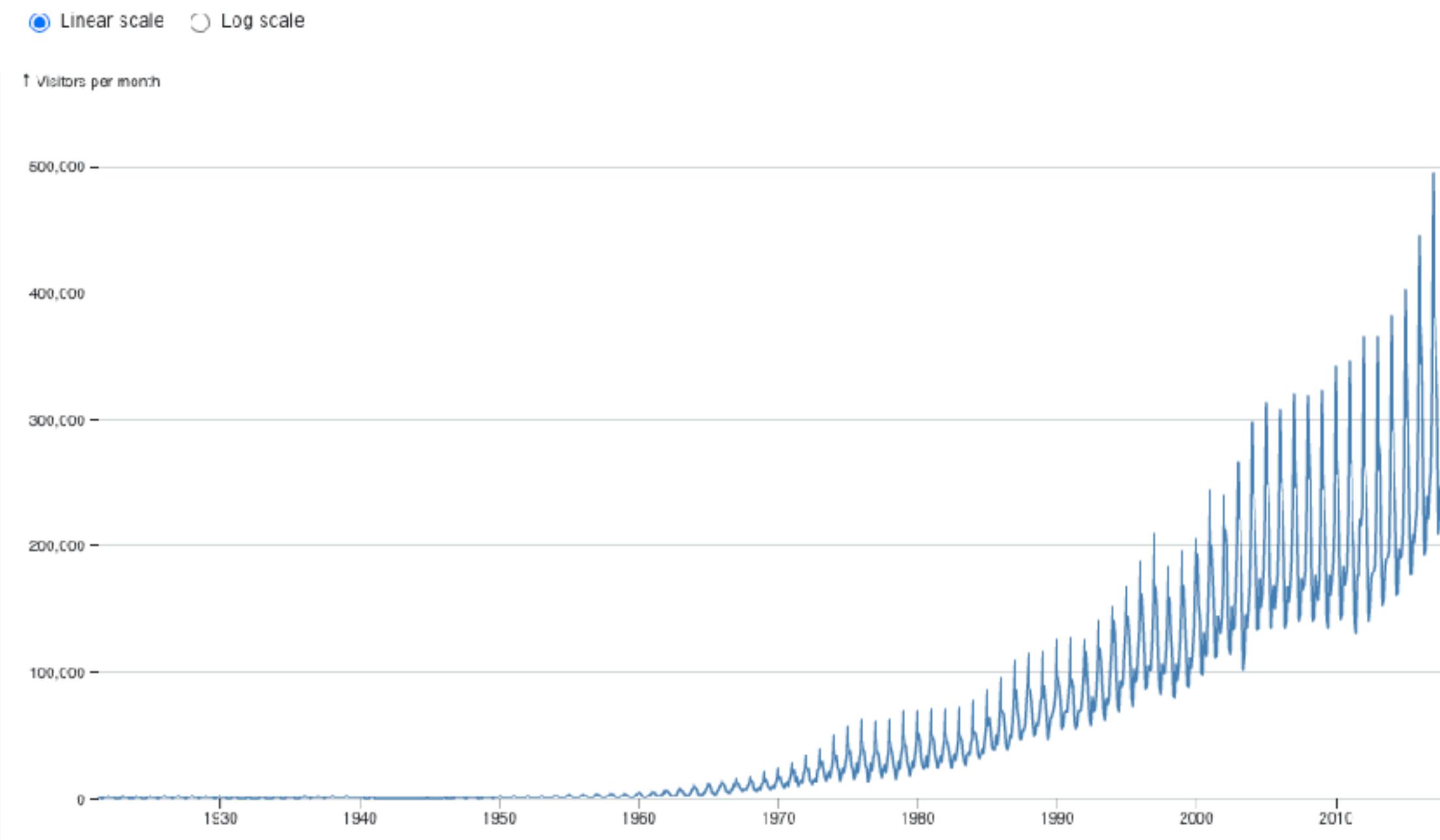


semi-log scale

The log(arithm) transformation turns a linear into a log scale

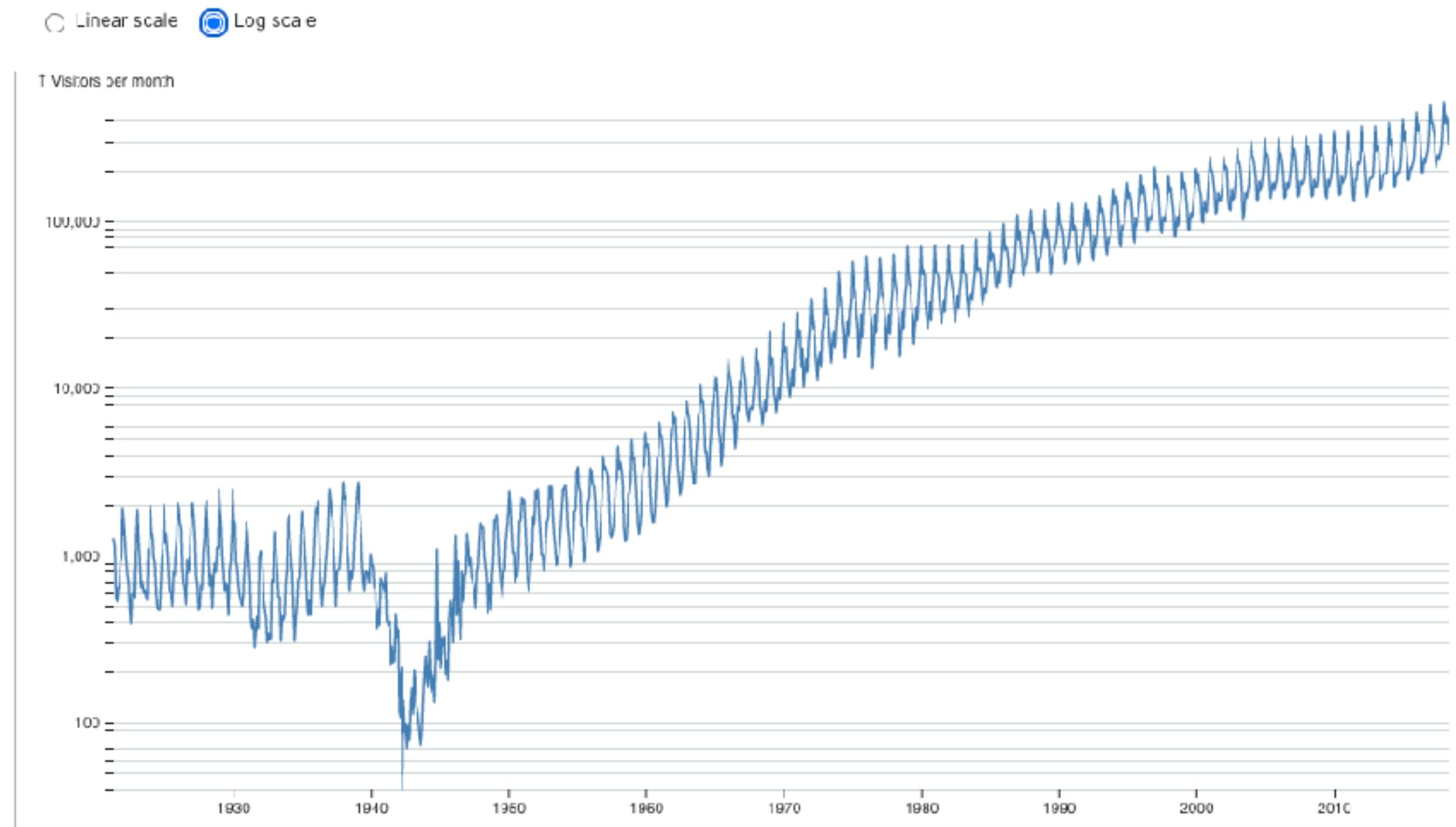
New Zealand Tourists, 1921–2018

Data: [Trading Economics](#) via Lisa Charlotte Rost



New Zealand Tourists, 1921–2018

Data: [Trading Economics](#) via Lisa Charlotte Rost



linear scales

semi-log scale

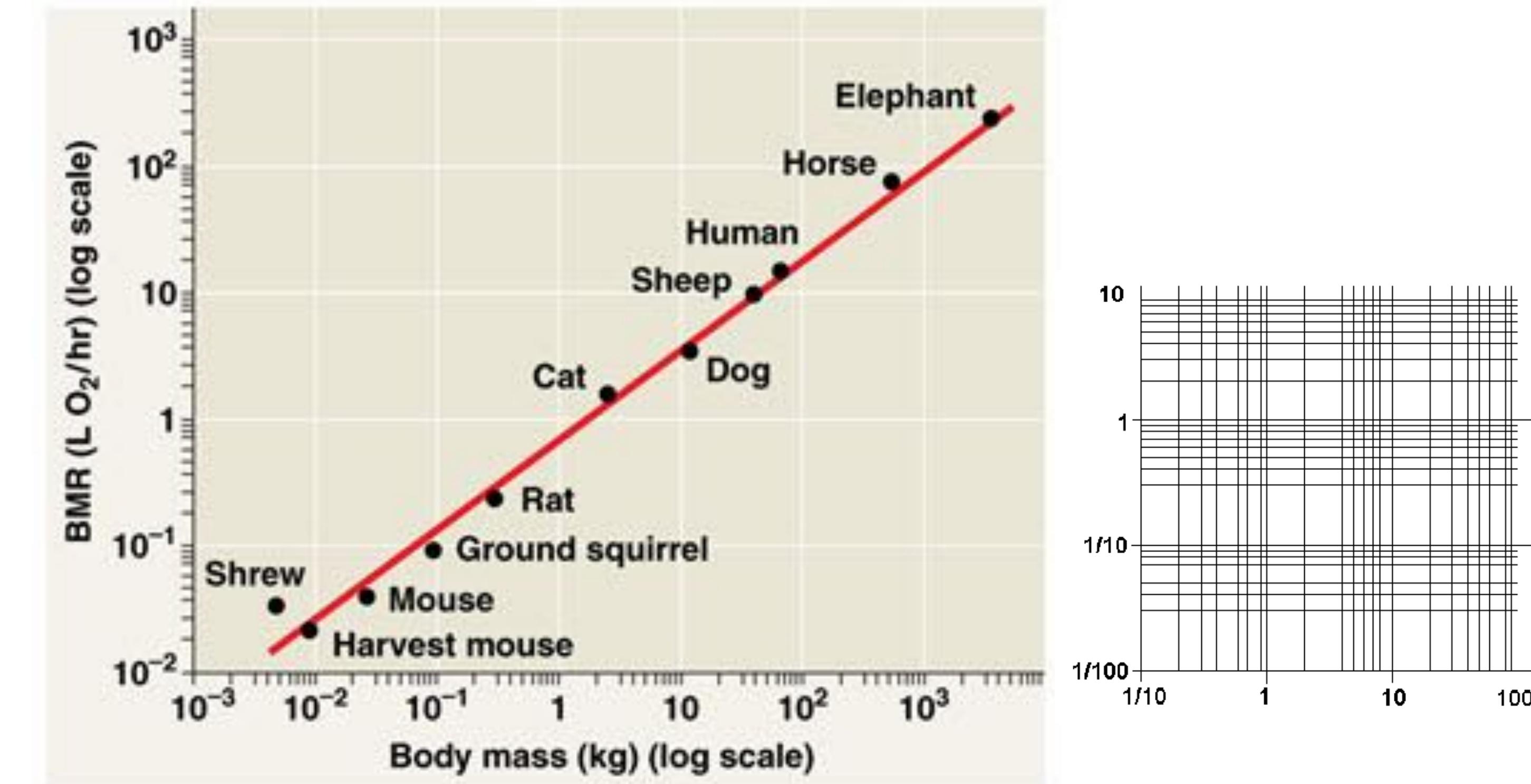
When we have a **power law**, we apply the log transform to both variables

$$y = ax^p$$

When we have a **power law**, we apply the log transform to both variables

$$y = ax^p$$

$$BMR = a \cdot BM^{3/4}$$



(a) Relationship of basal metabolic rate (BMR) to body size for various mammals

© 2011 Pearson Education, Inc.

log-log scale

The log-transform turns bent curves into straight lines,
making them easier to analyze

Exponential $y = ab^x$ log on right side \rightarrow $y = a' + x$

Power law $y = ax^p$ \rightarrow $y' = a' + px'$
log on both sides

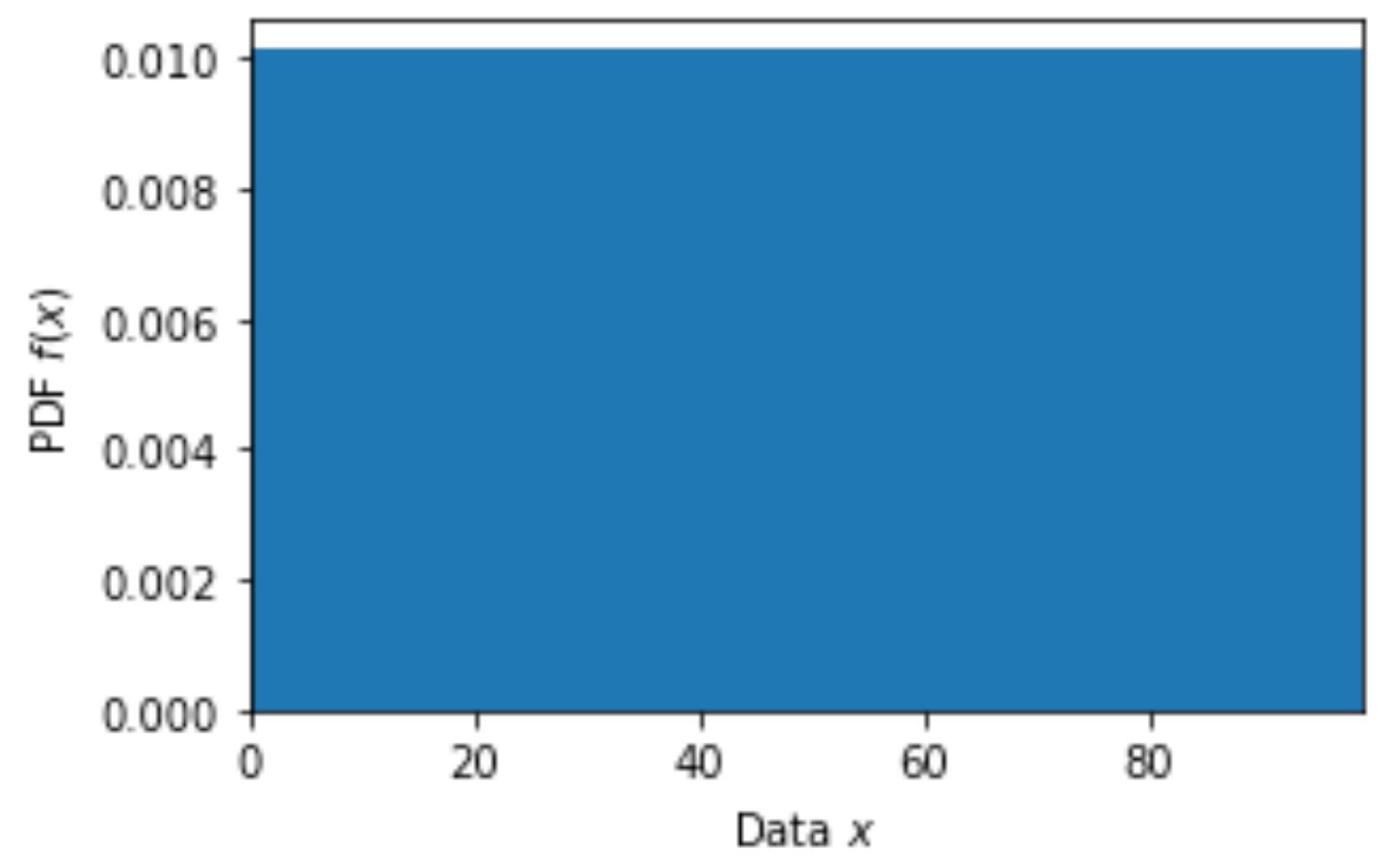
Jupyter

2) Complementary cumulative distribution function (CCDF)

PDF

Probability
density function

$$f(x)$$



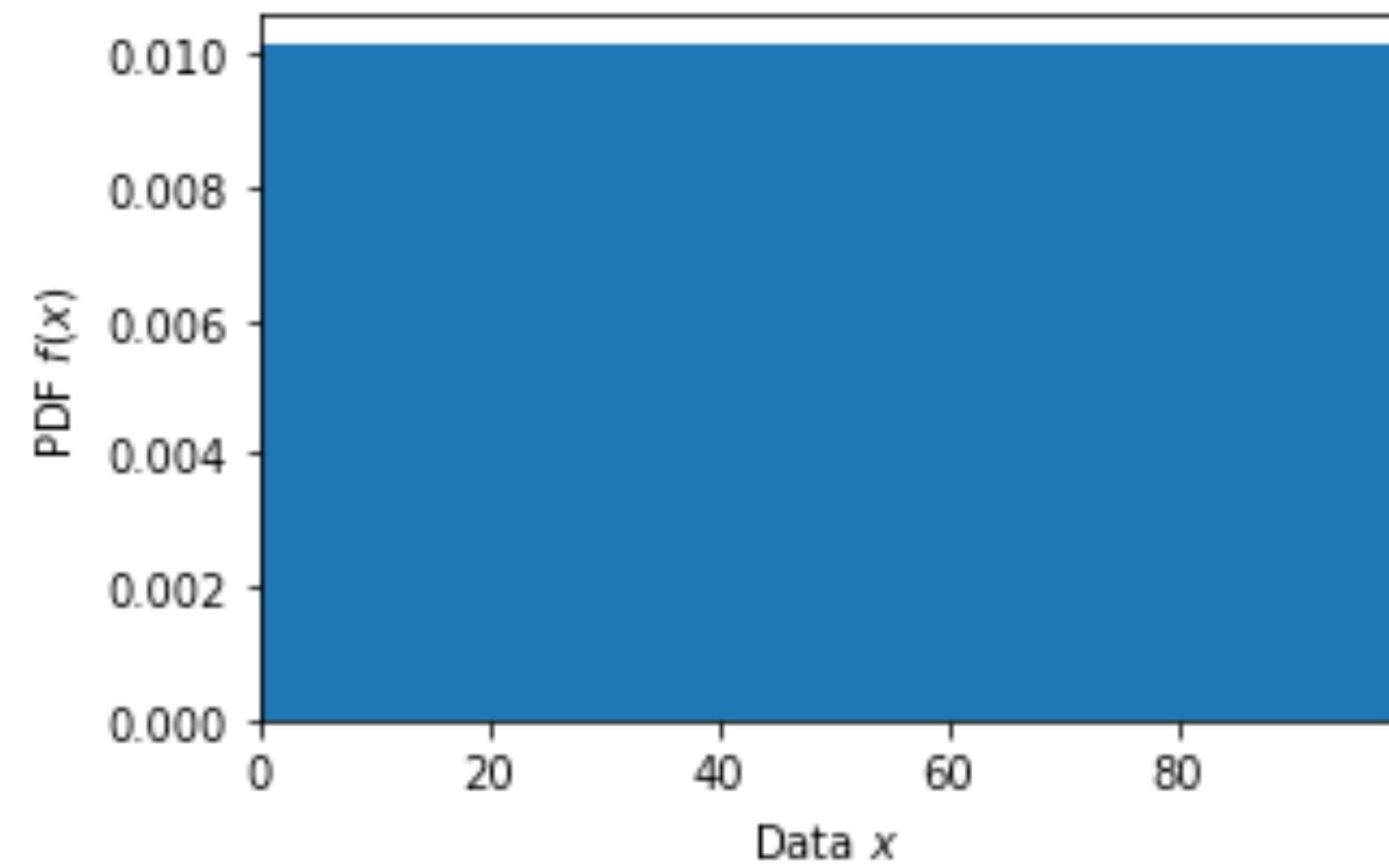
PDF

Probability
density function

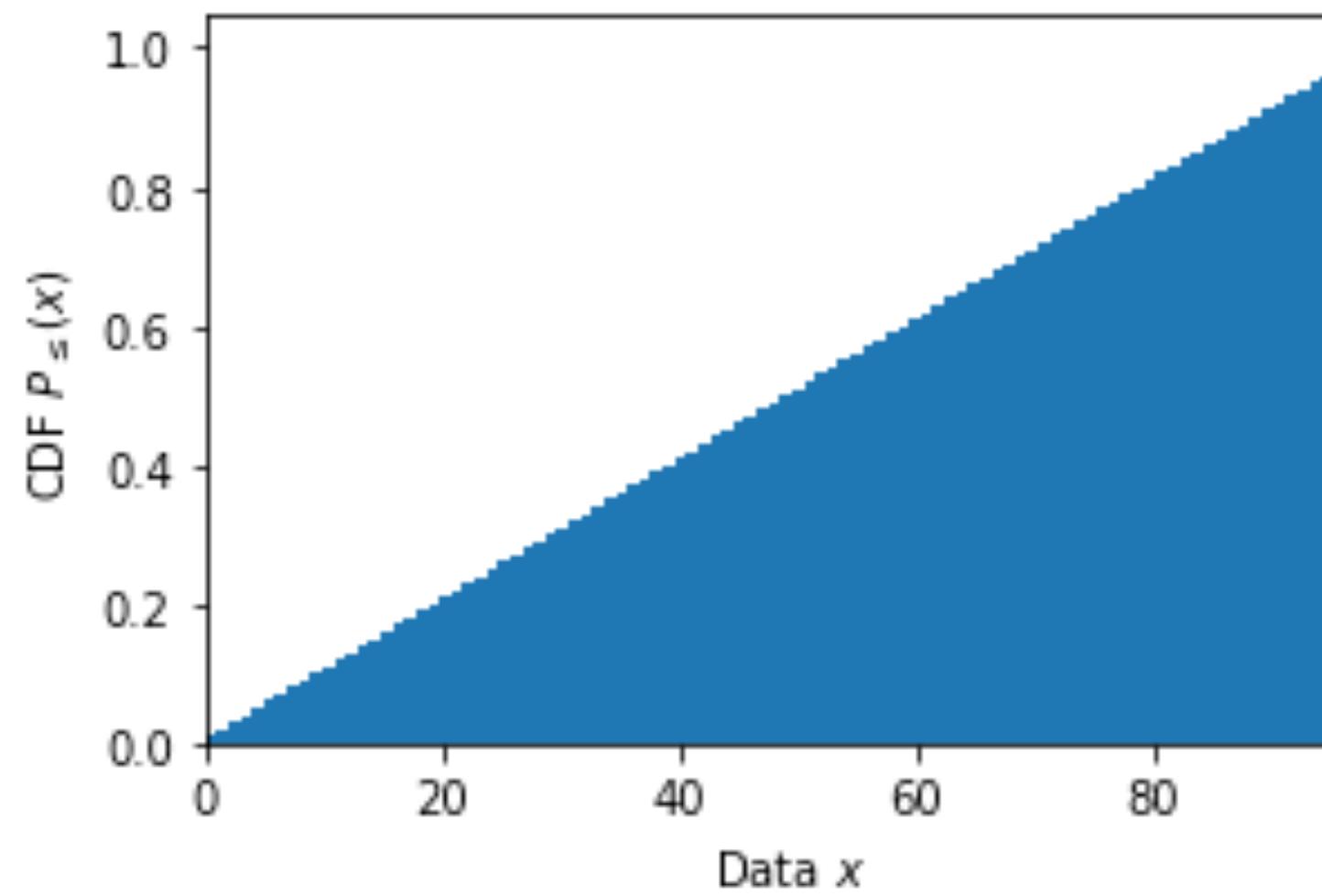
CDF

Cumulative
distribution function

$$f(x)$$



$$P_{\leq}(x)$$



Integrate

$$P(X \leq x)$$

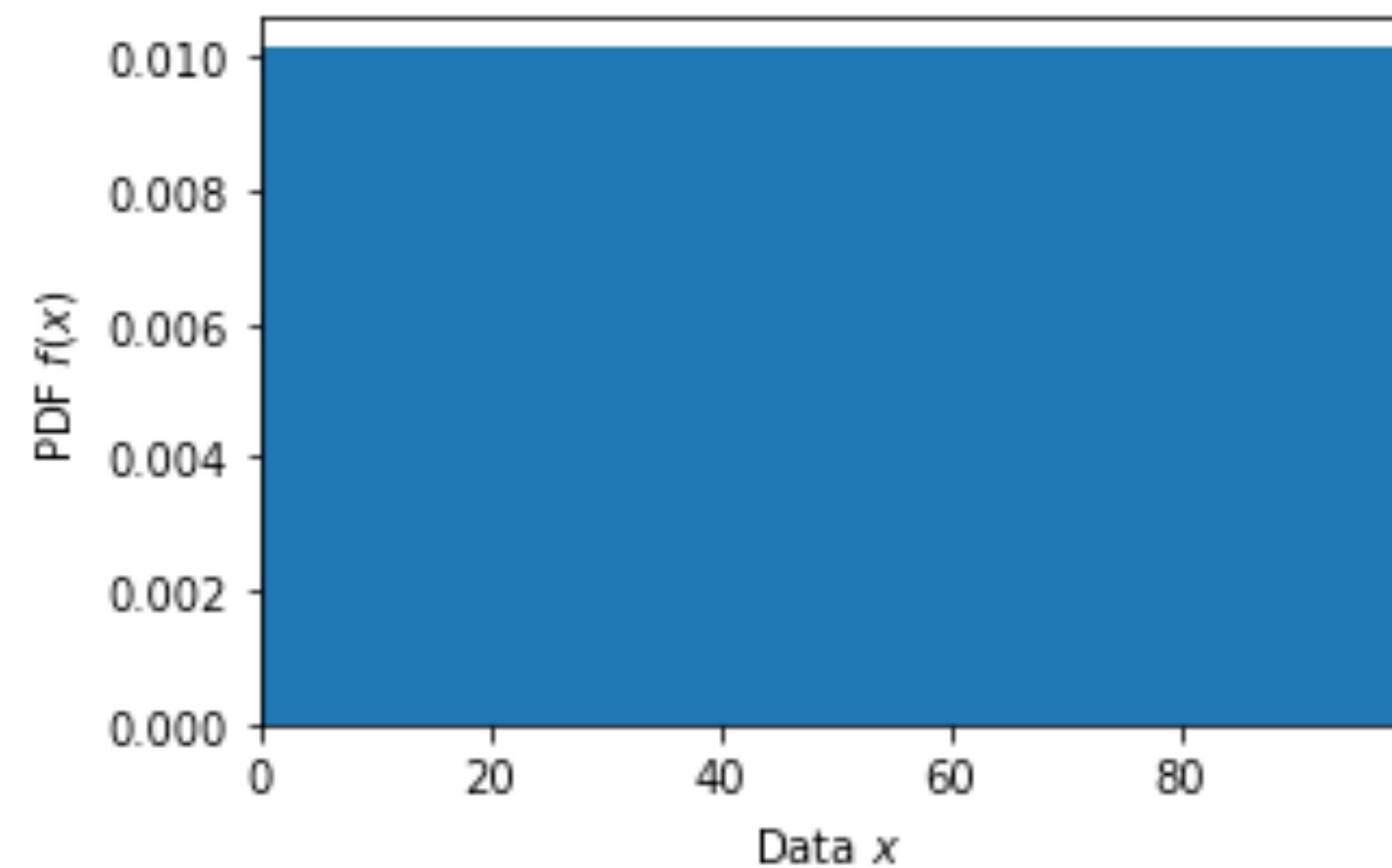
PDF

Probability
density function

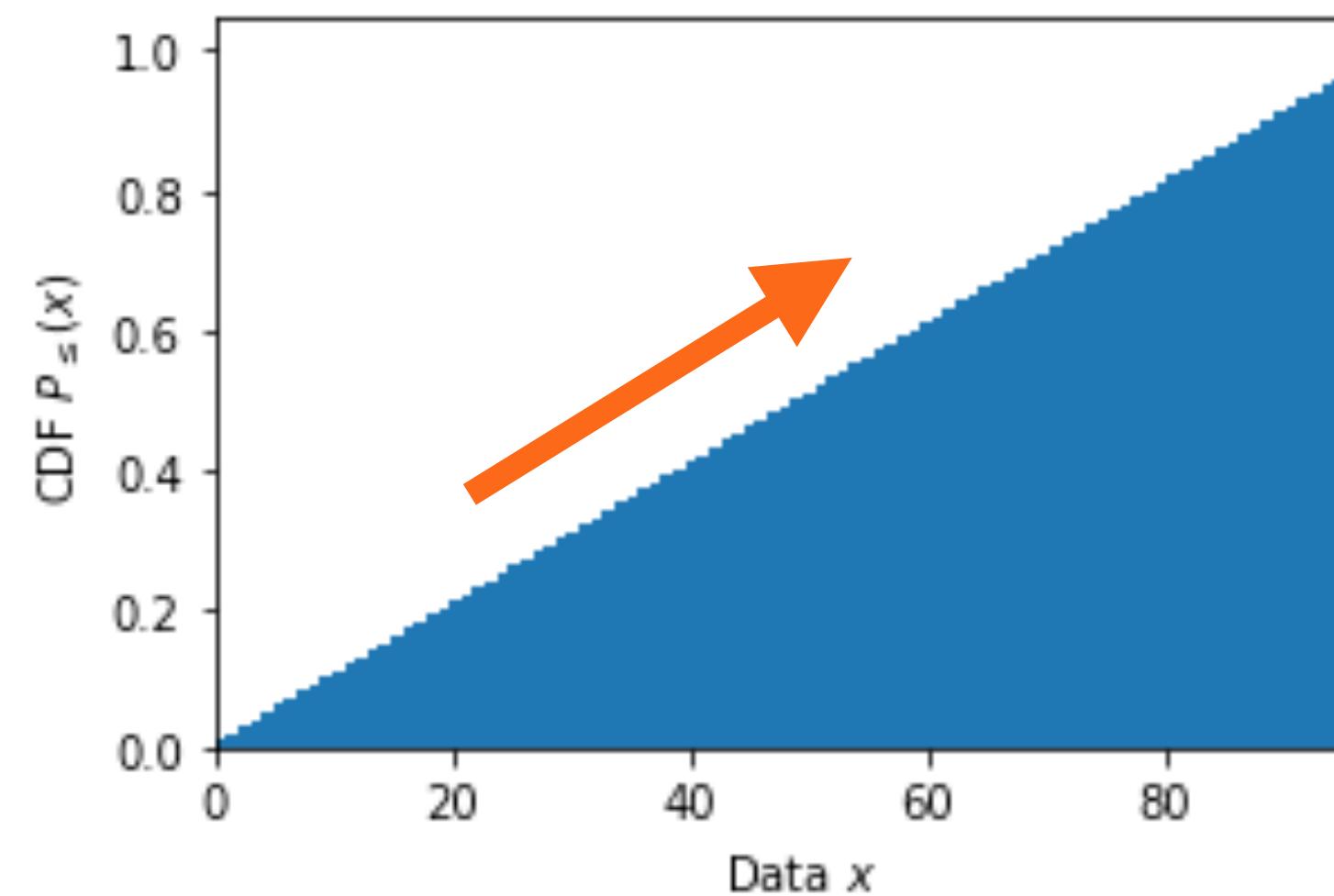
CDF

Cumulative
distribution function

$$f(x)$$



$$P_{\leq}(x)$$



The CDF is
monotonically
non-decreasing
It starts at 0

Integrate

$$P(X \leq x)$$

PDF

Probability
density function

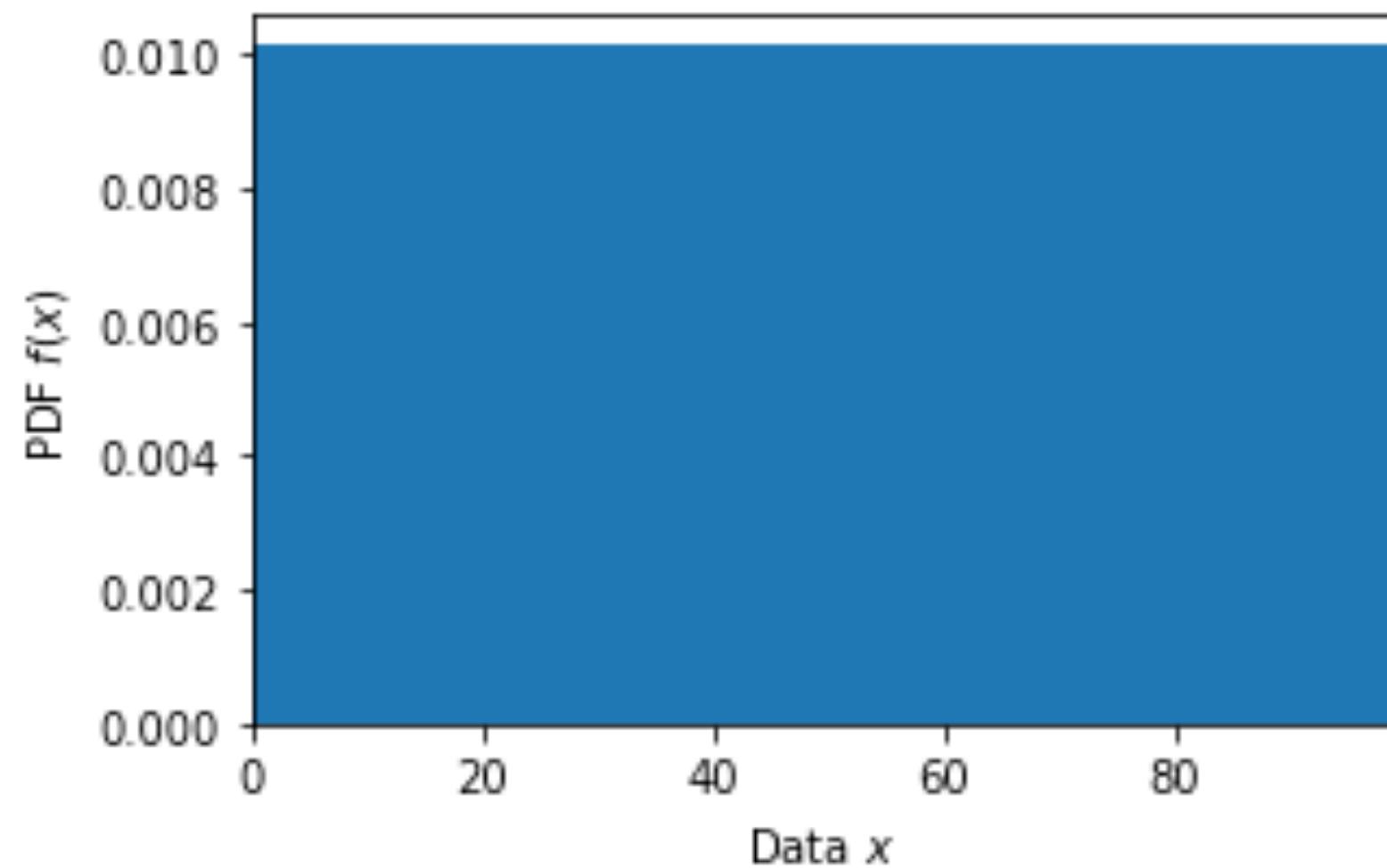
CDF

Cumulative
distribution function

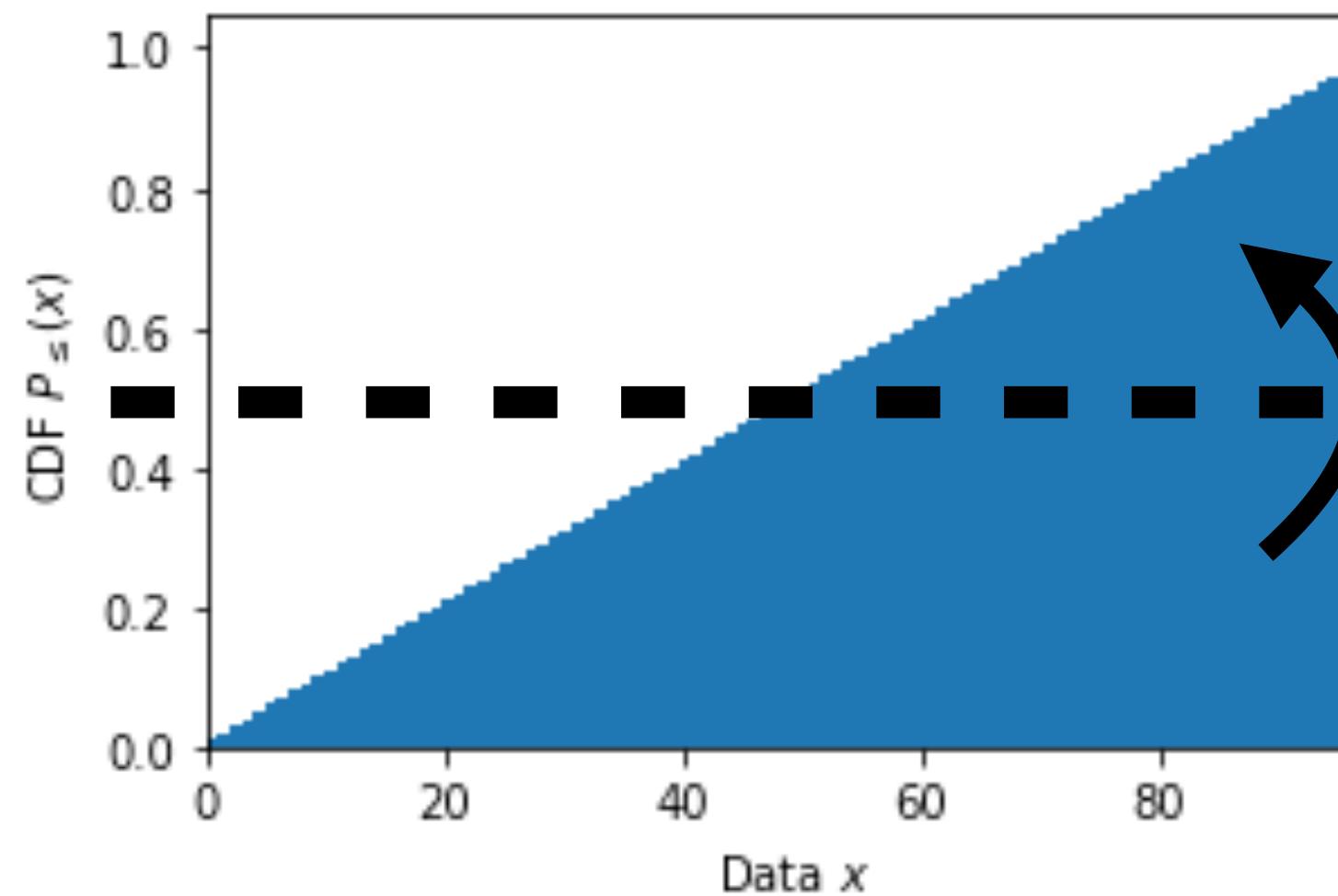
CCDF

Complementary
cumulative
distribution function

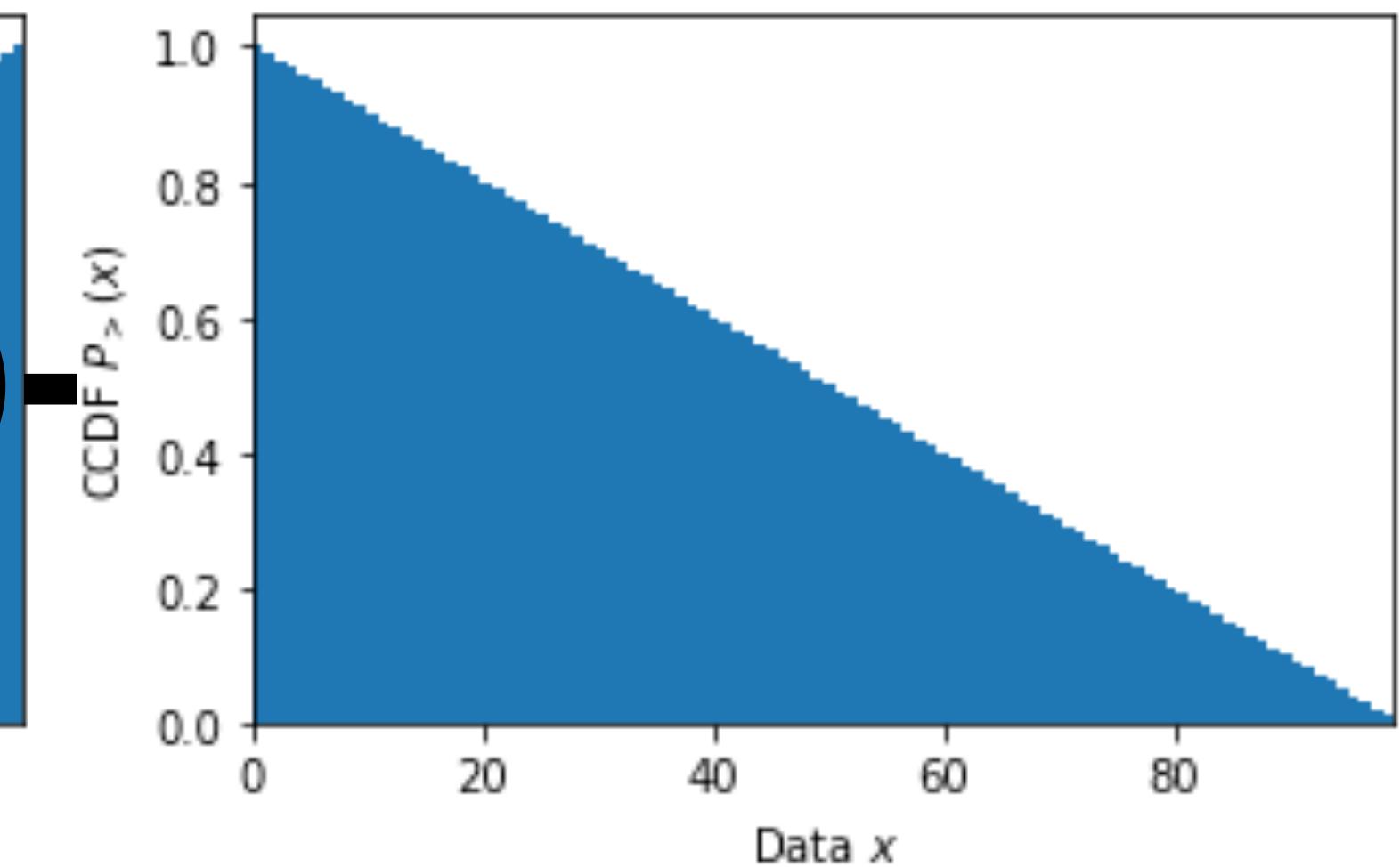
$$f(x)$$



$$P_{\leq}(x)$$



$$P_{>}(x)$$



Integrate

$$P(X \leq x)$$

Flip

$$P(X > x)$$

PDF

Probability
density function

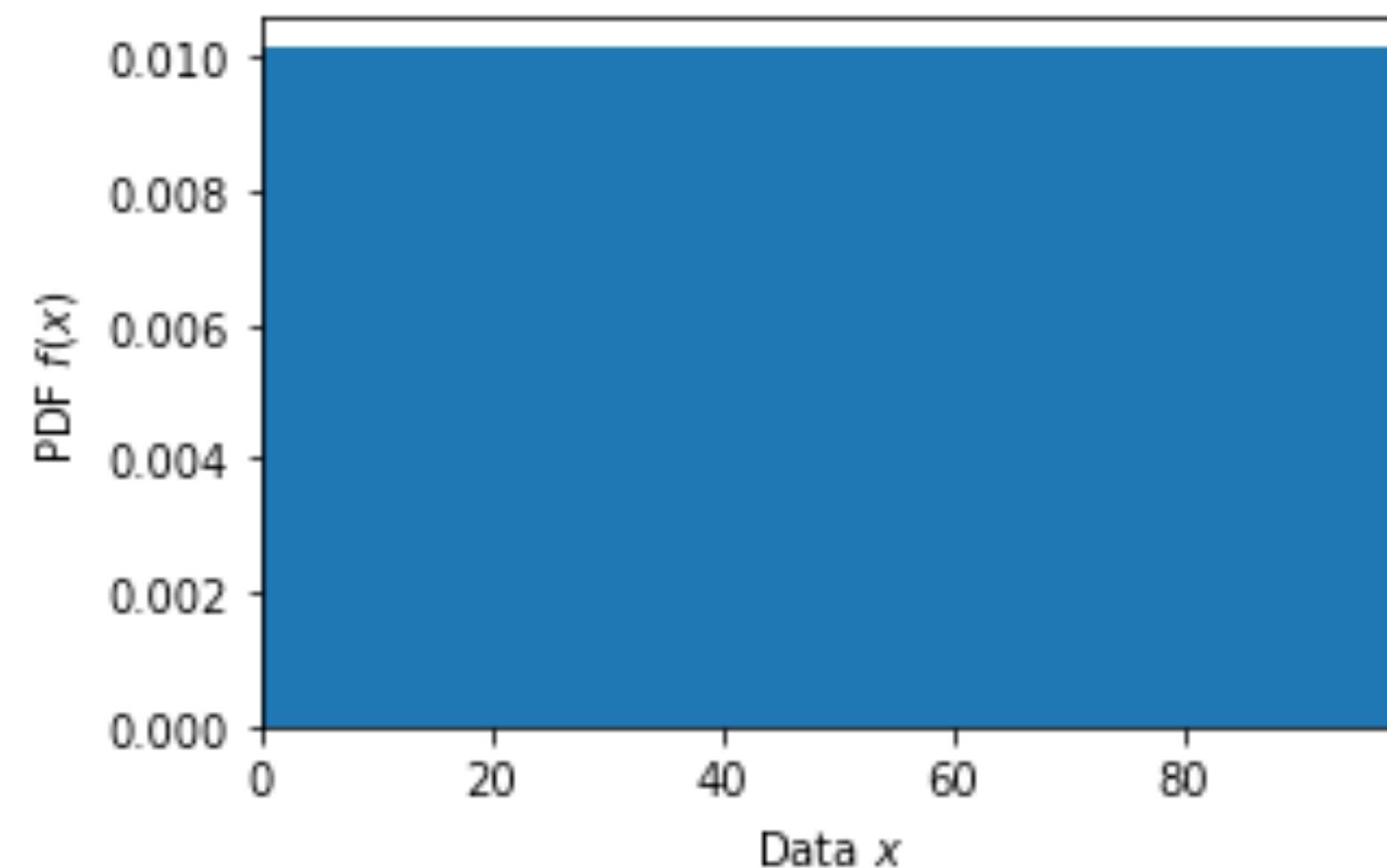
CDF

Cumulative
distribution function

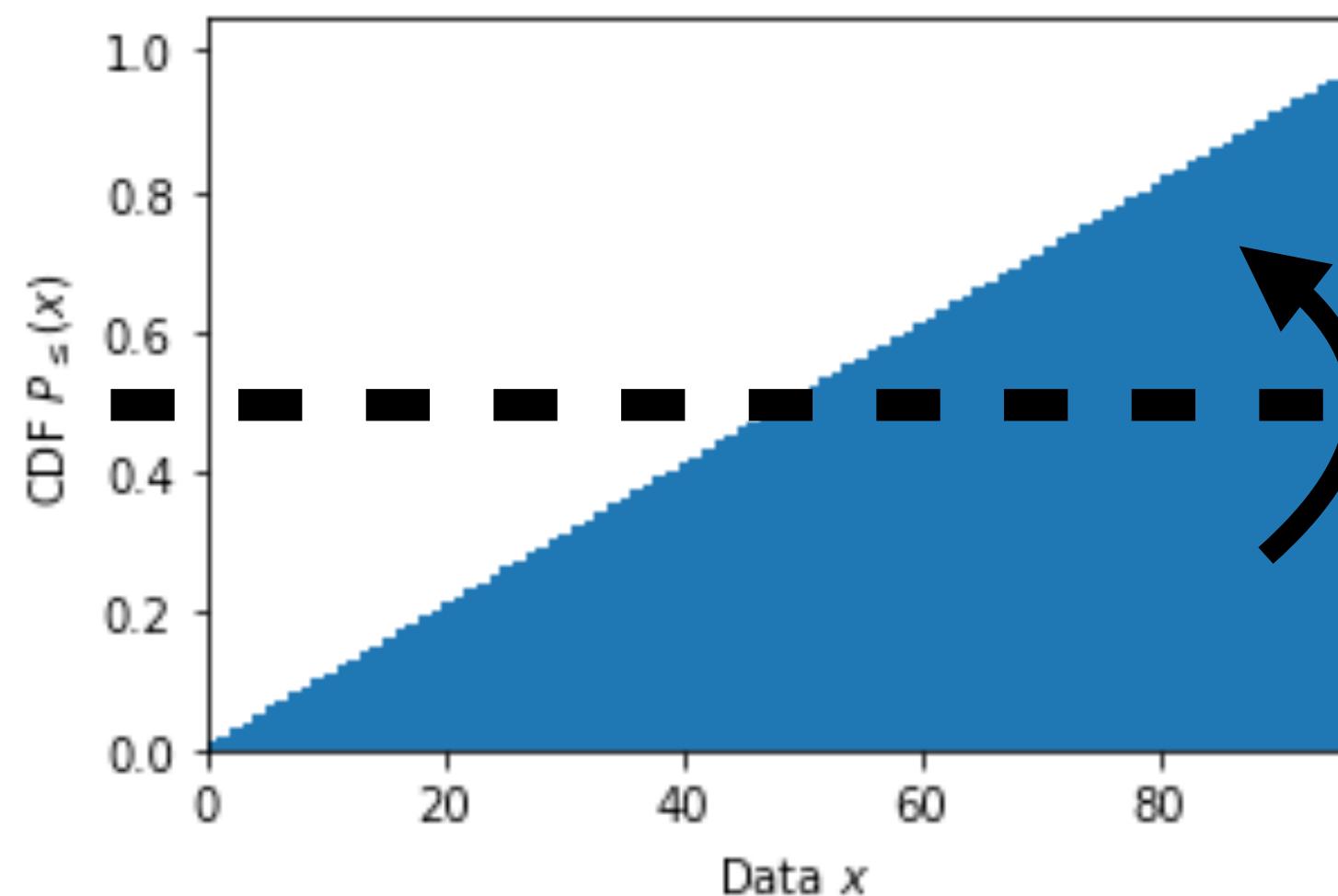
CCDF

Complementary
cumulative
distribution function

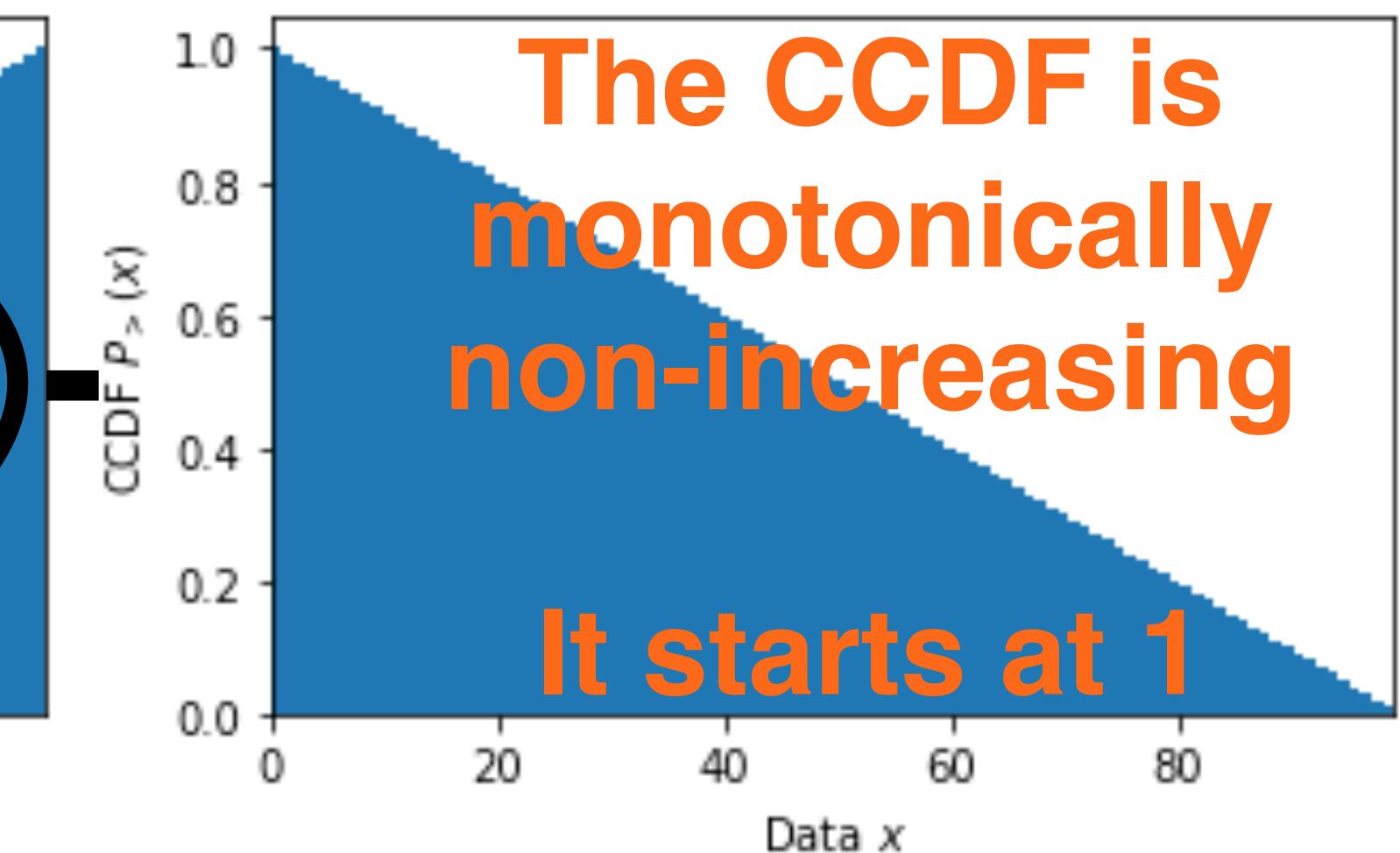
$$f(x)$$



$$P_{\leq}(x)$$



$$P_{>}(x)$$



Integrate

$$P(X \leq x)$$

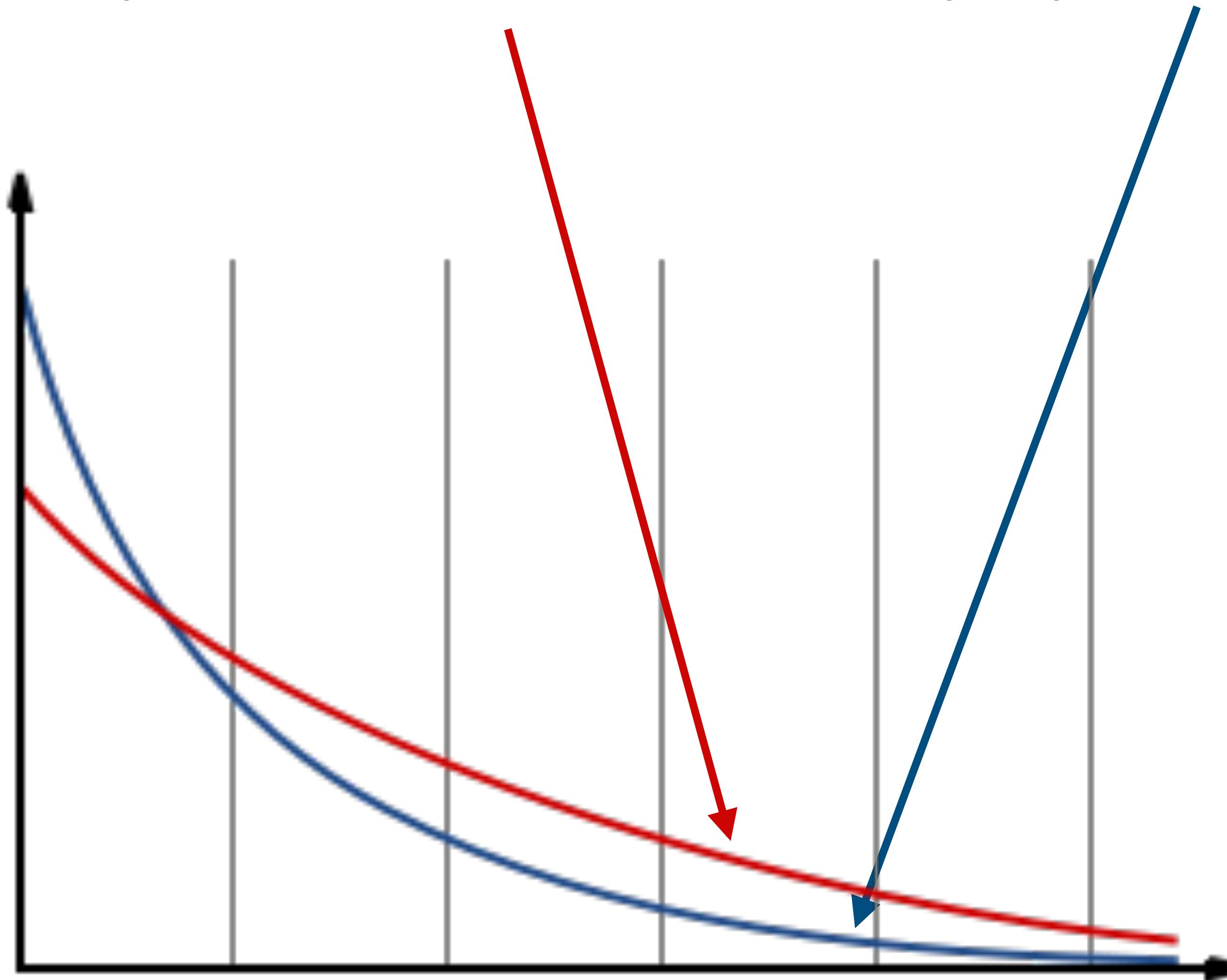
Flip

$$P(X > x)$$

Jupyter

Many real distributions are **heavy-tailed**

A distribution is heavy-tailed if its **tail** is not bounded by any **exponential** (from above)



There are many names for certain heavy-tailed distributions

Fat tail



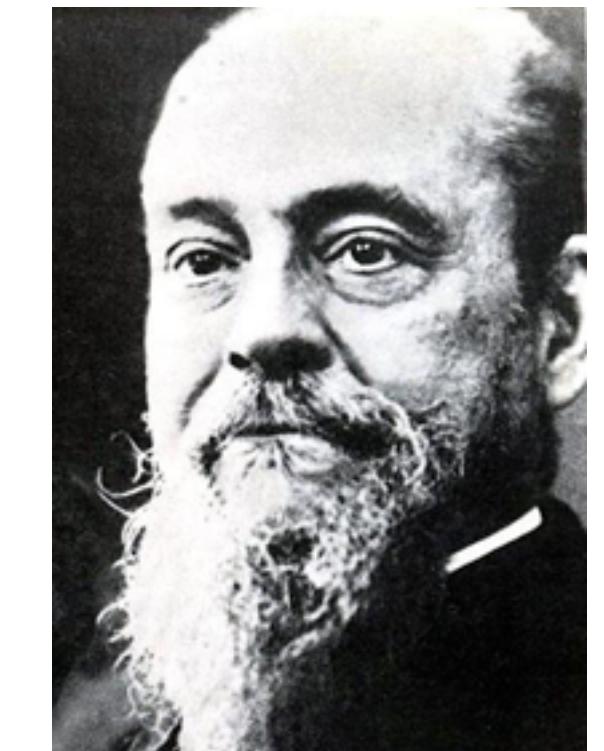
Power law



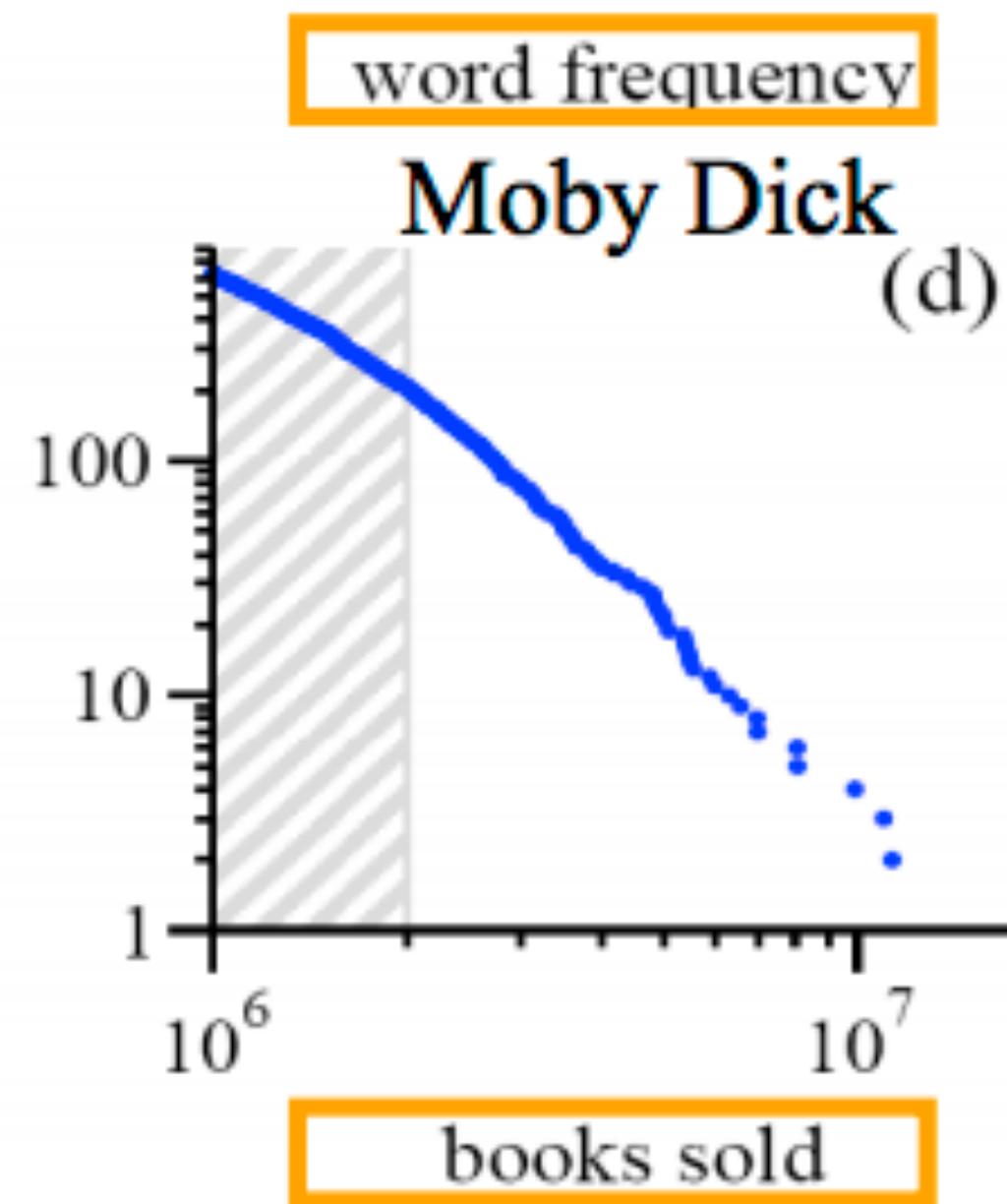
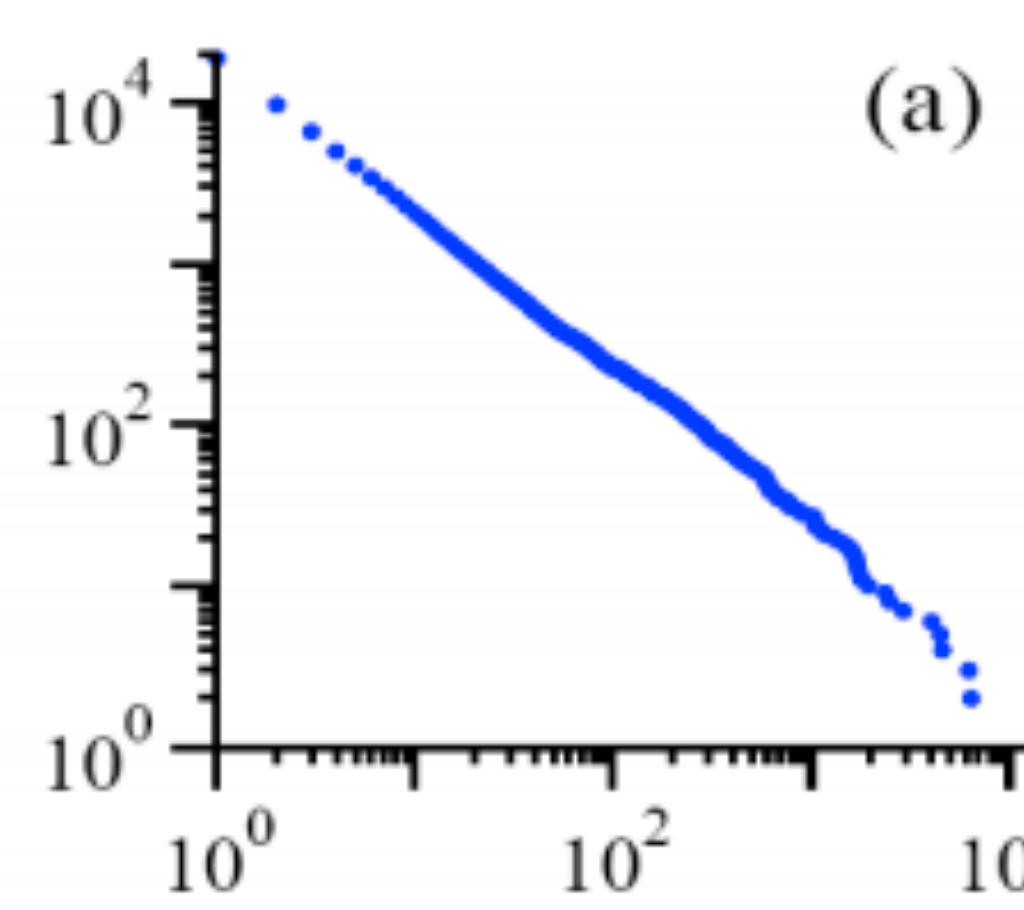
Long tail



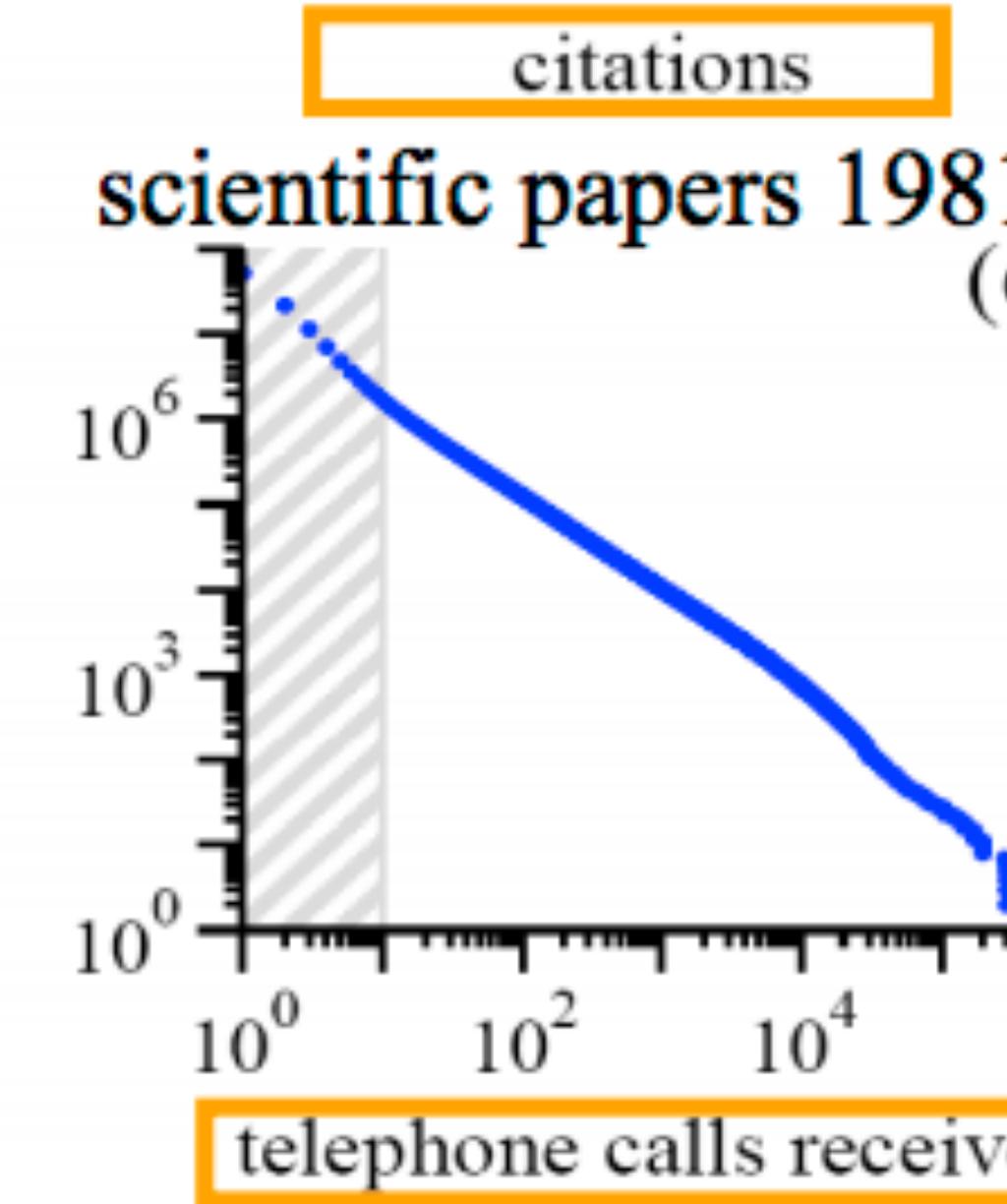
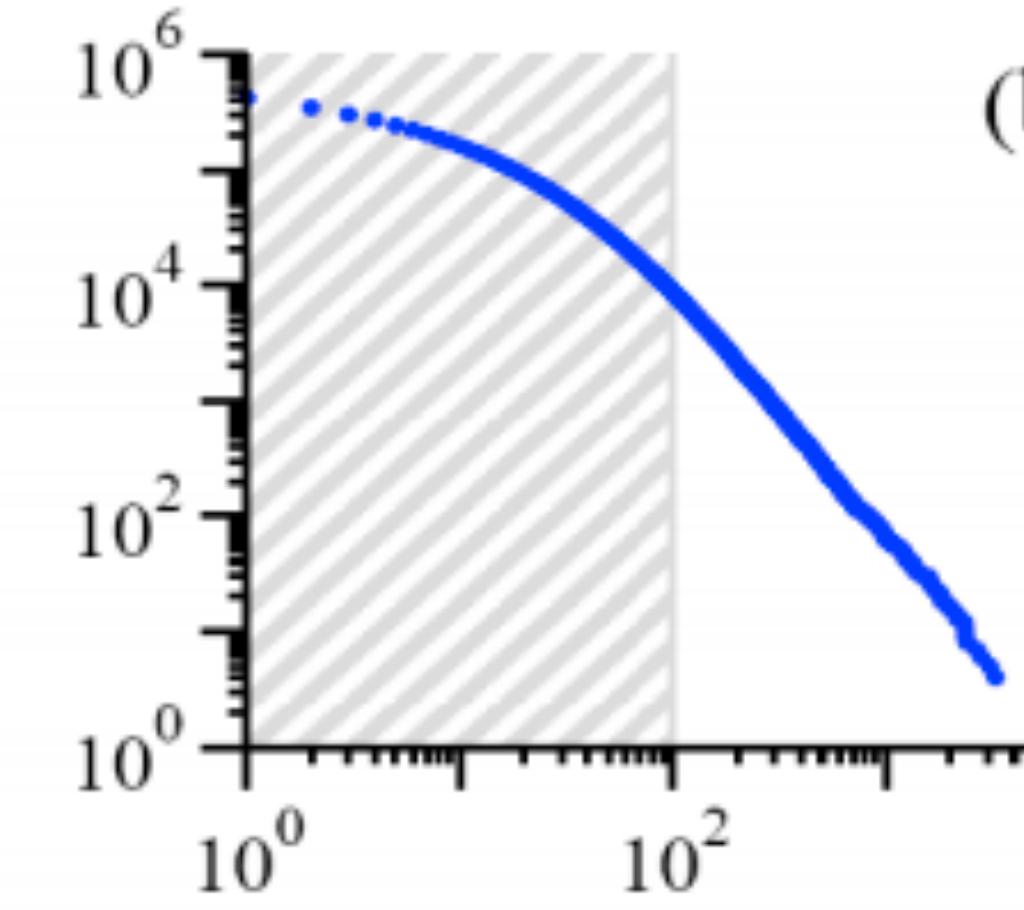
Pareto



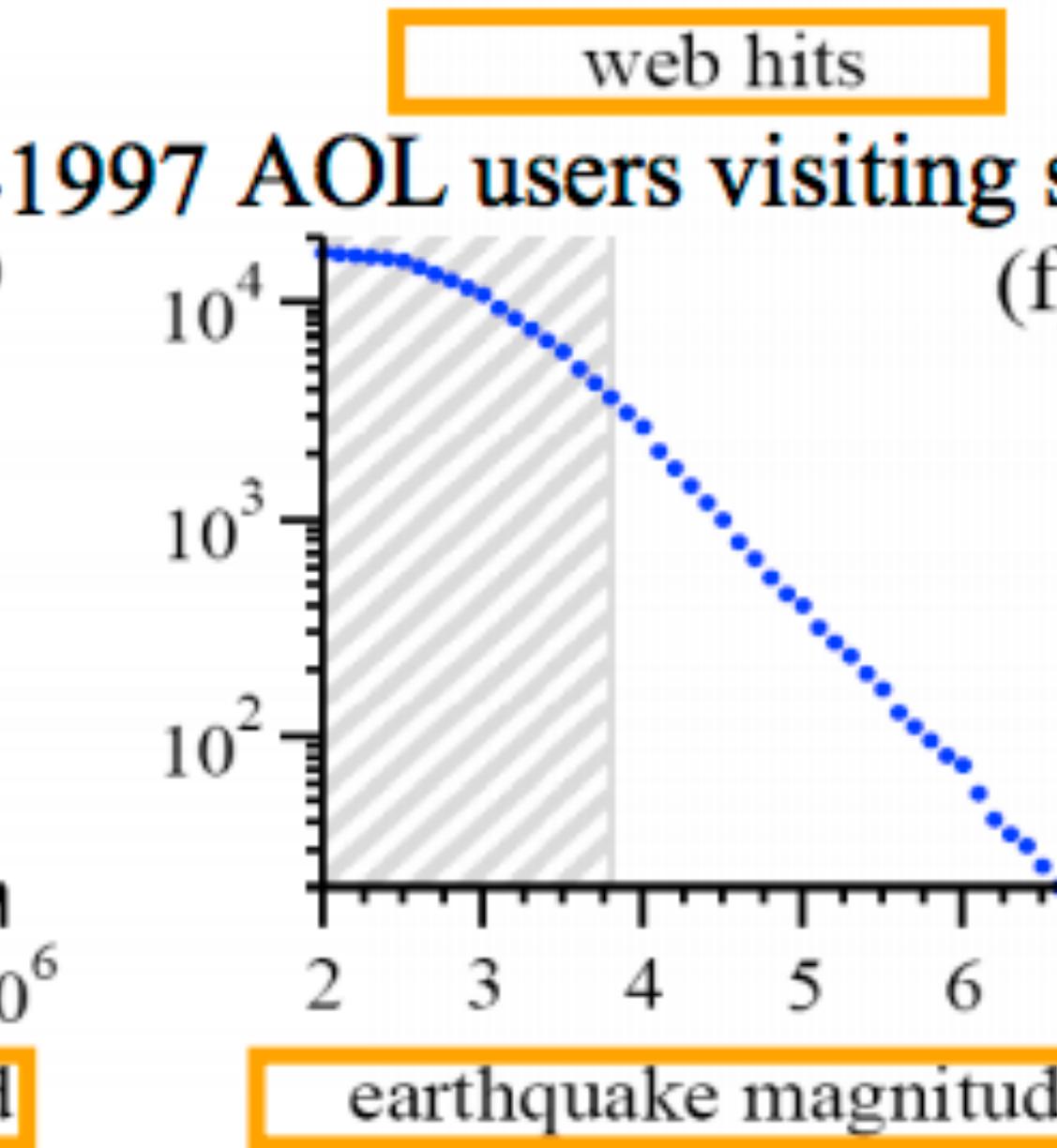
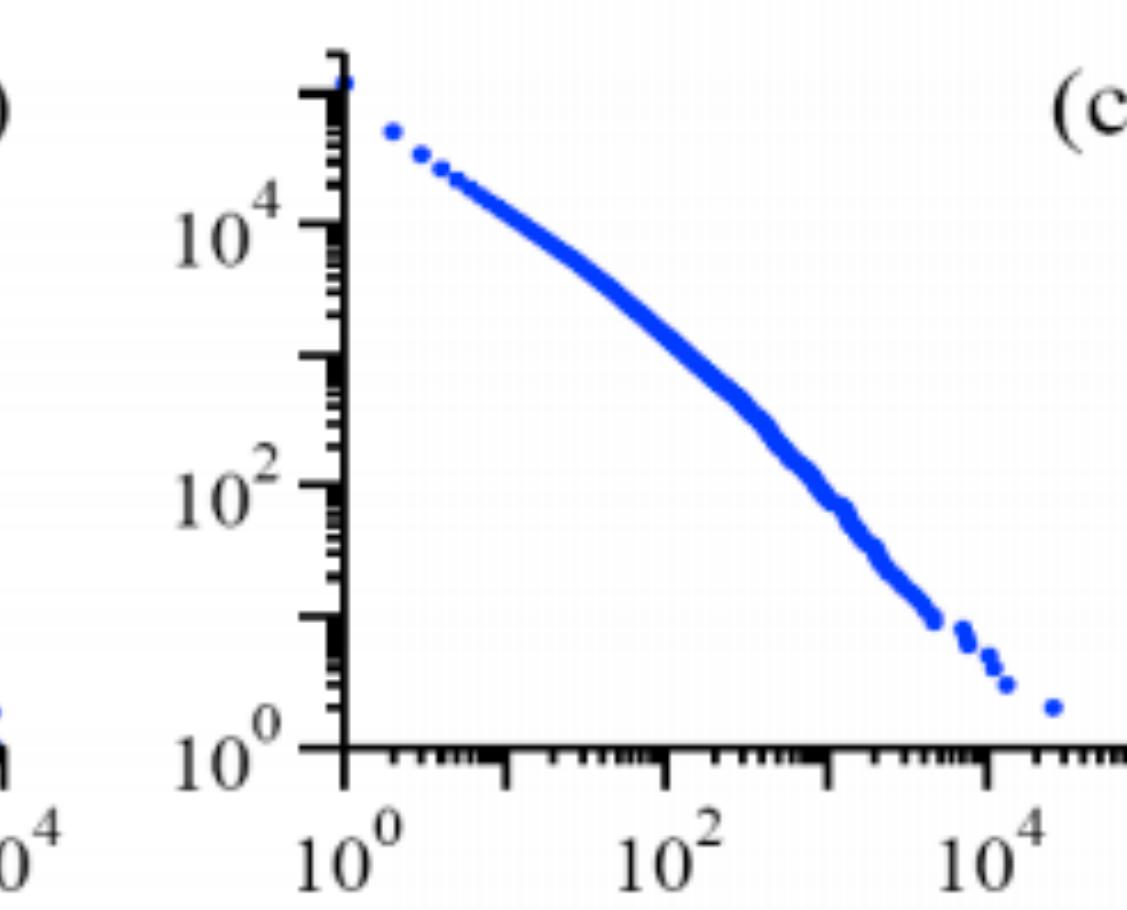
Heavy tails appear in all social systems, also in natural systems



bestsellers 1895-1965

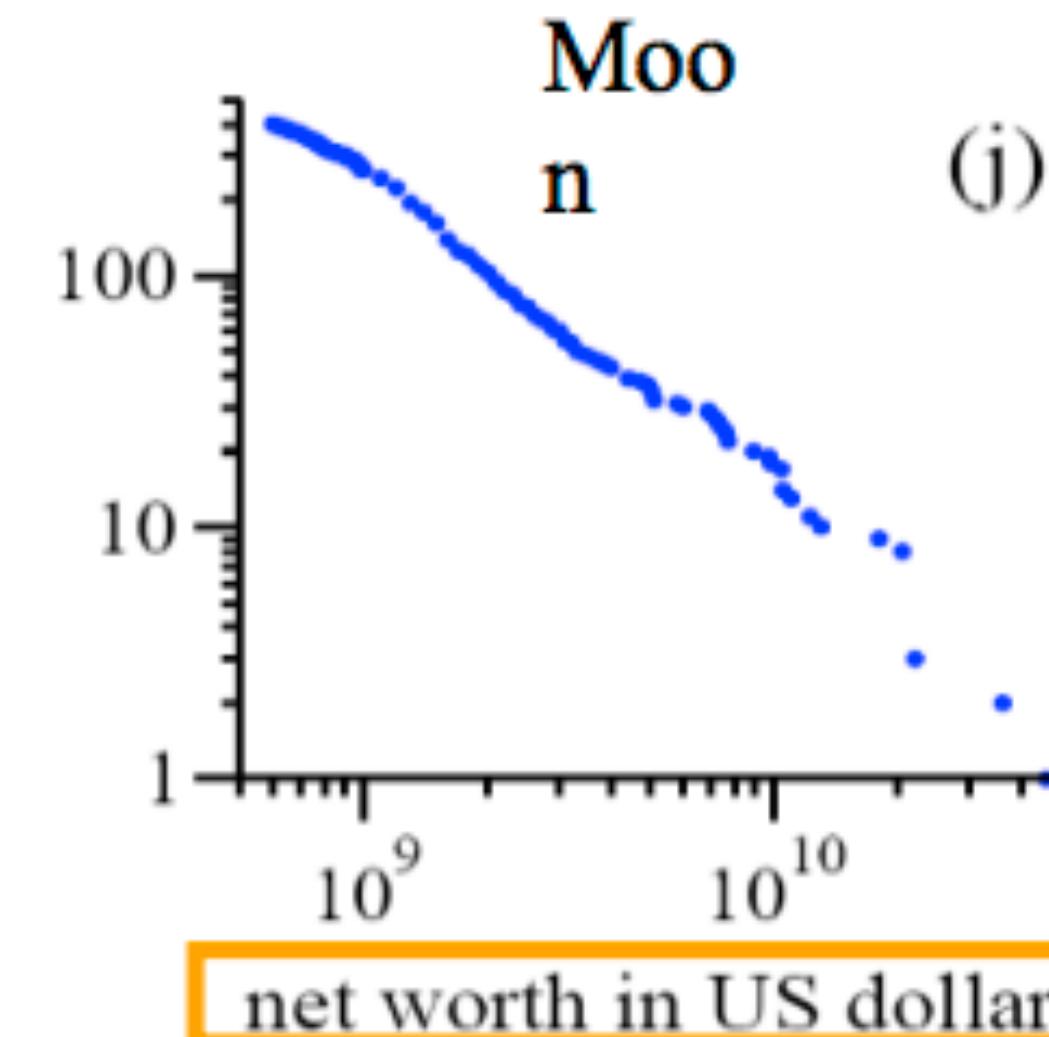
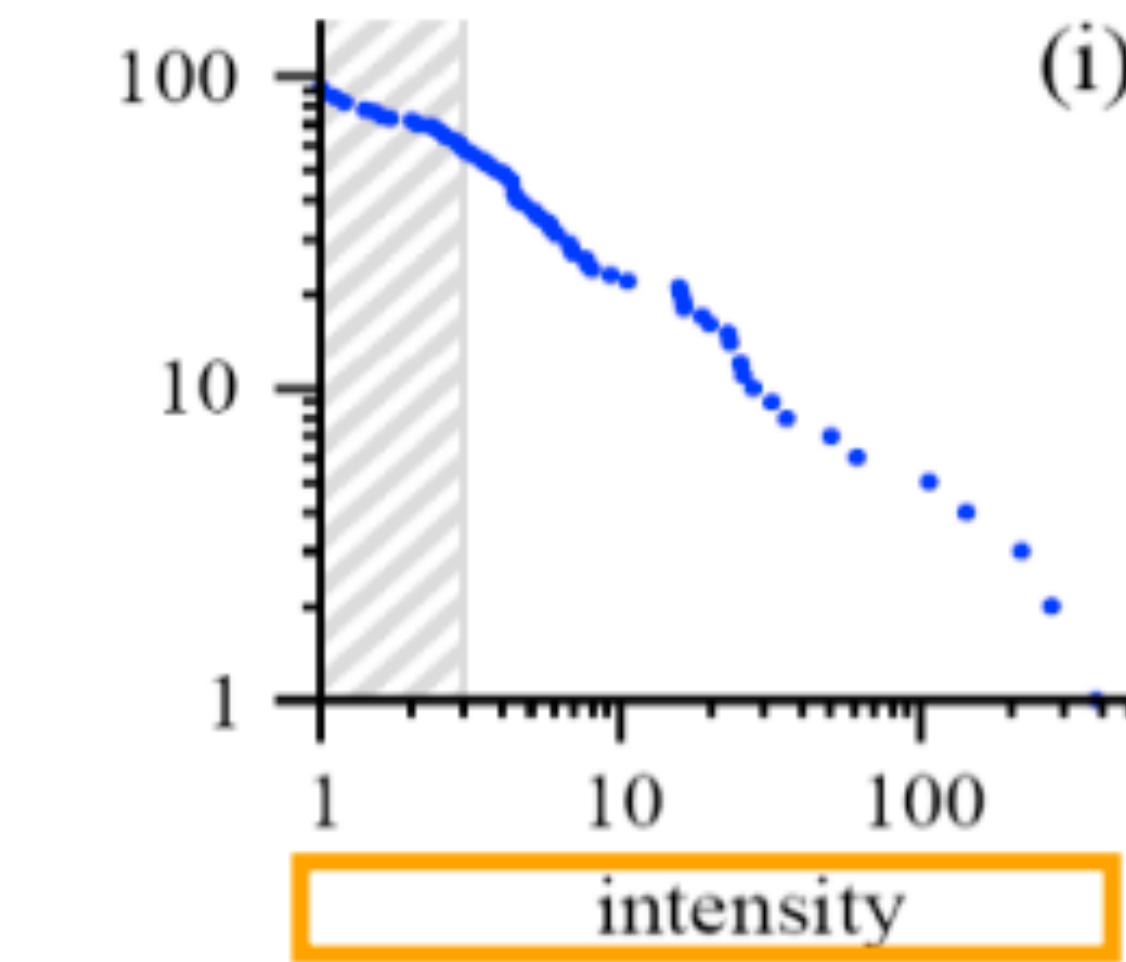
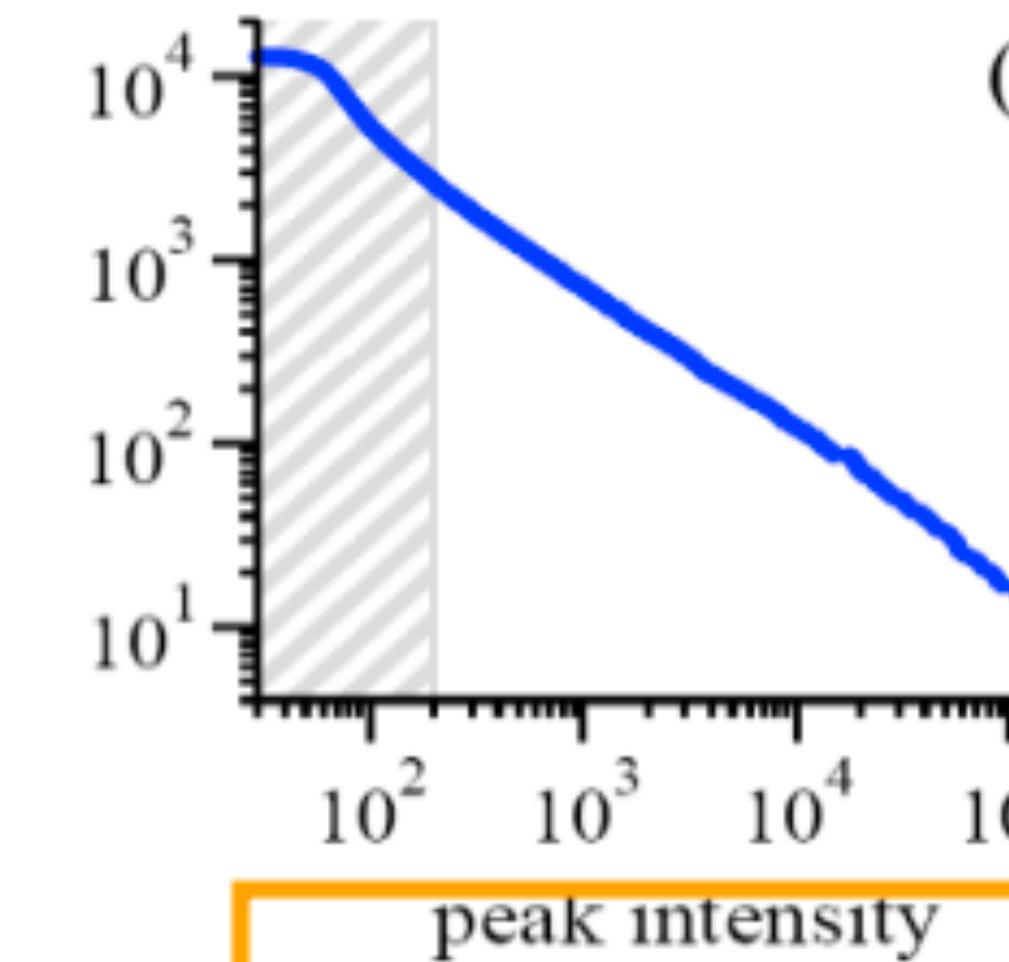
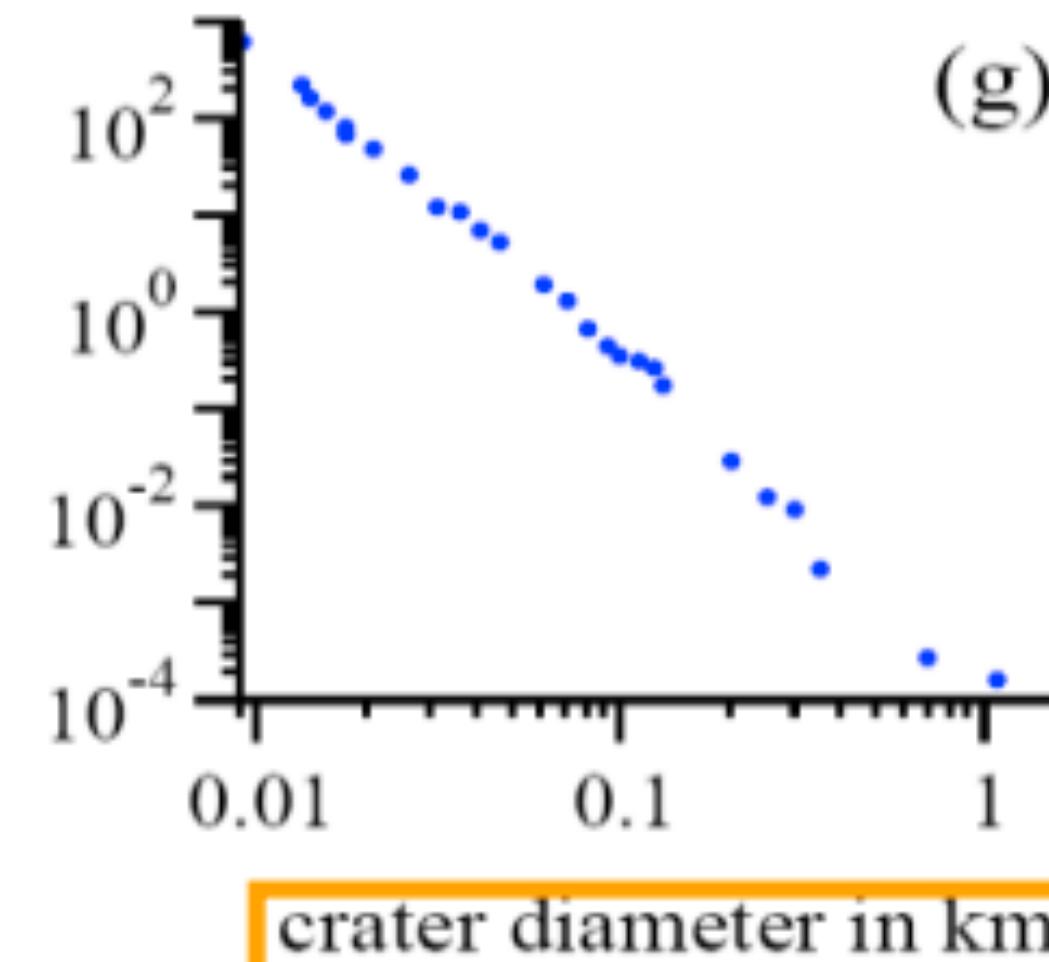


AT&T customers on 1 day

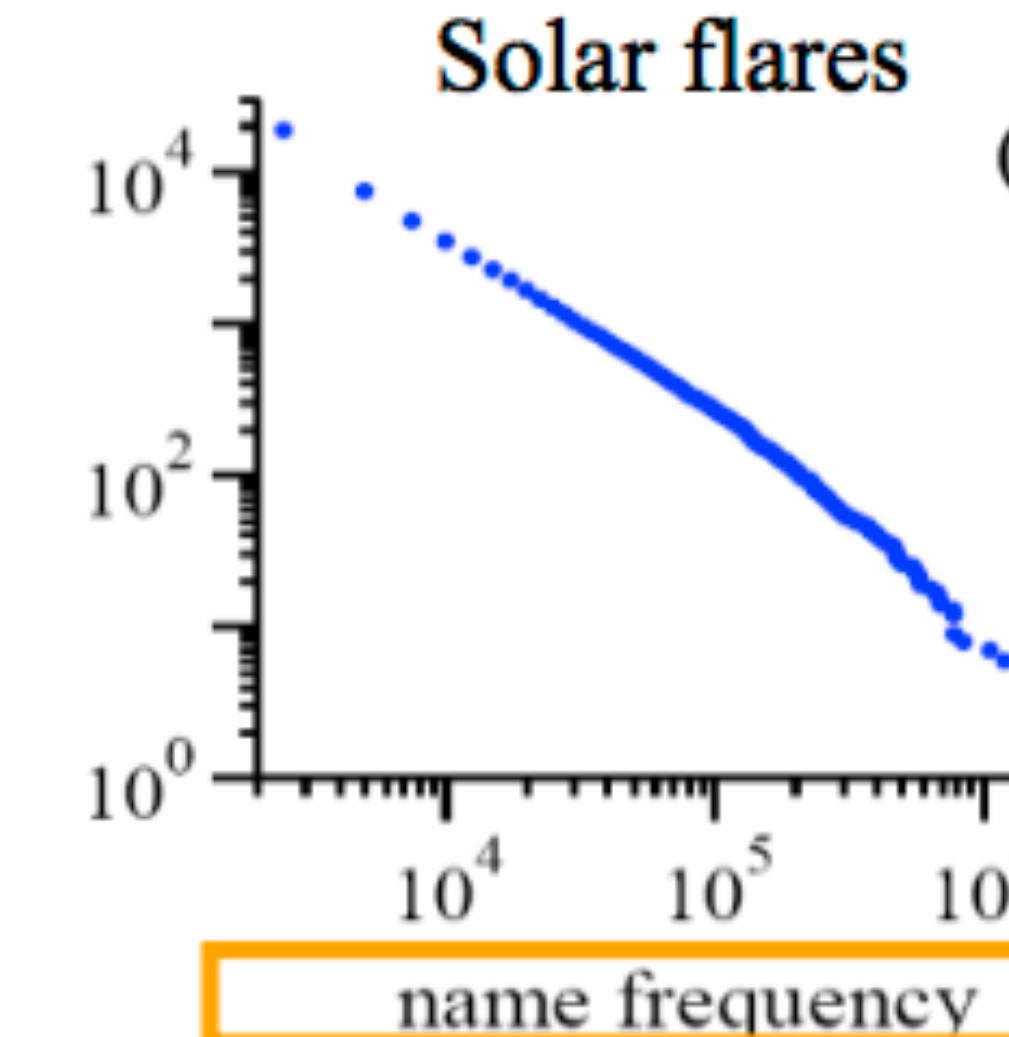


California 1910-1992

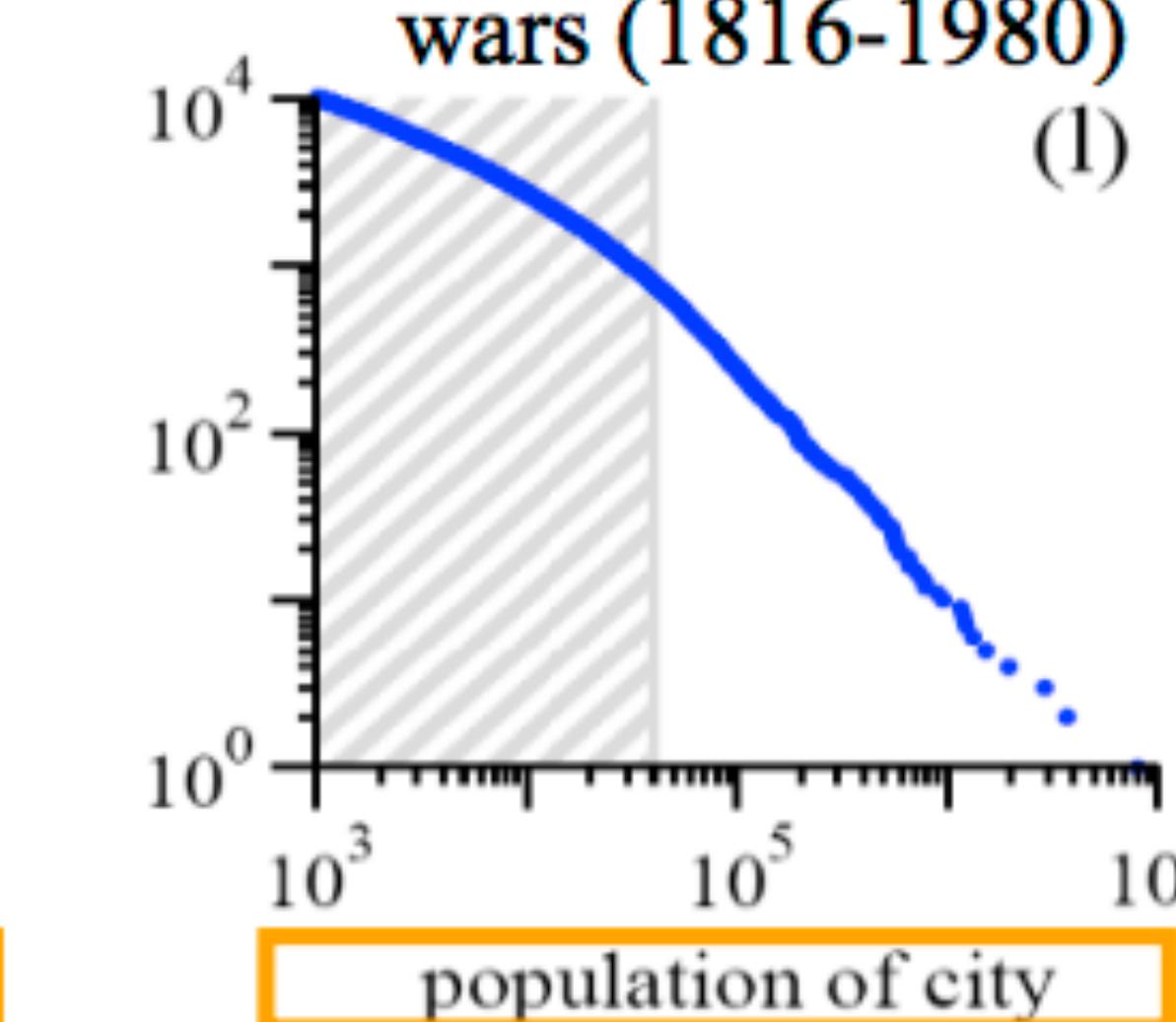
Heavy tails appear in all social systems, also in natural systems



richest individuals
2003

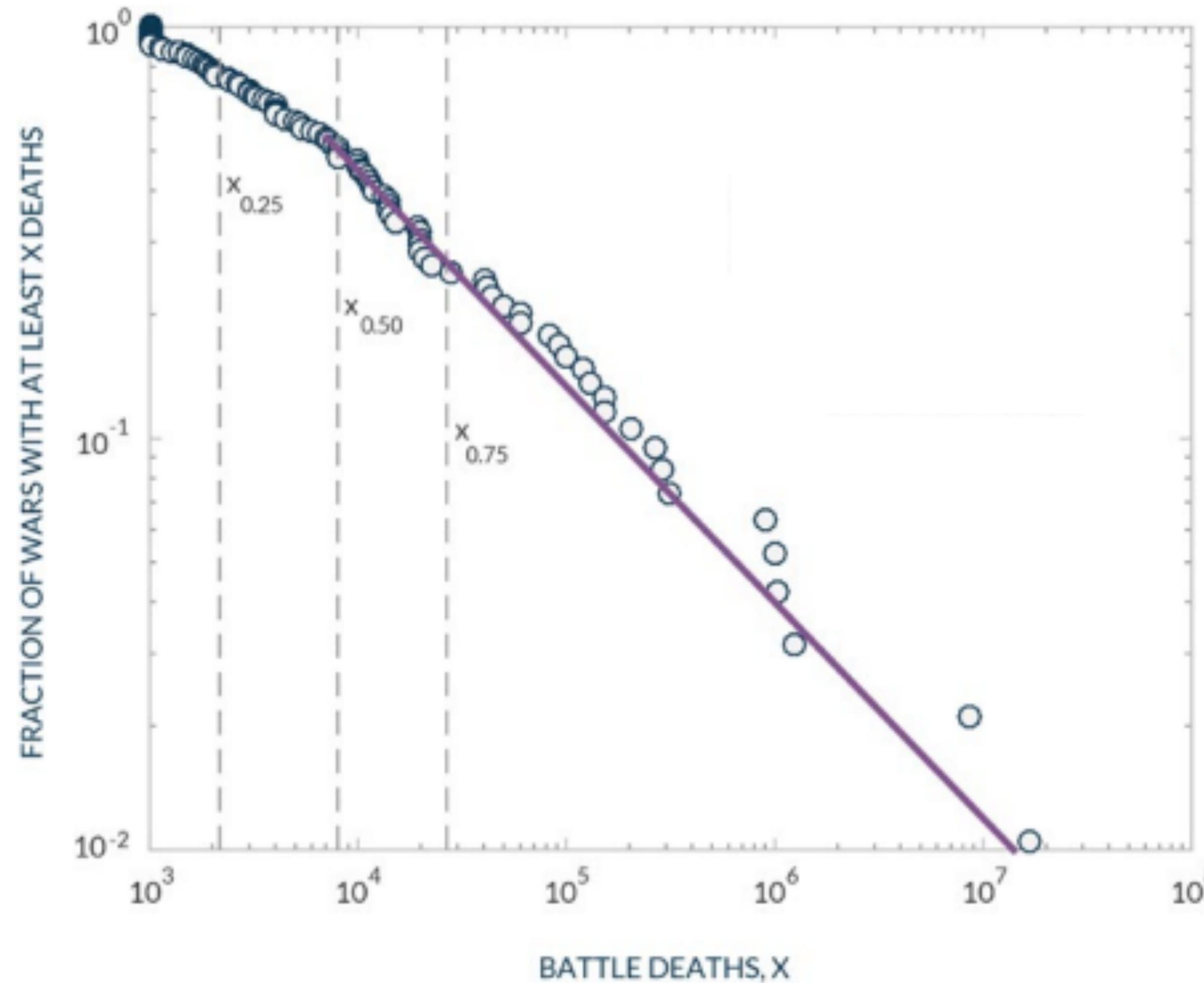


US family names
1990



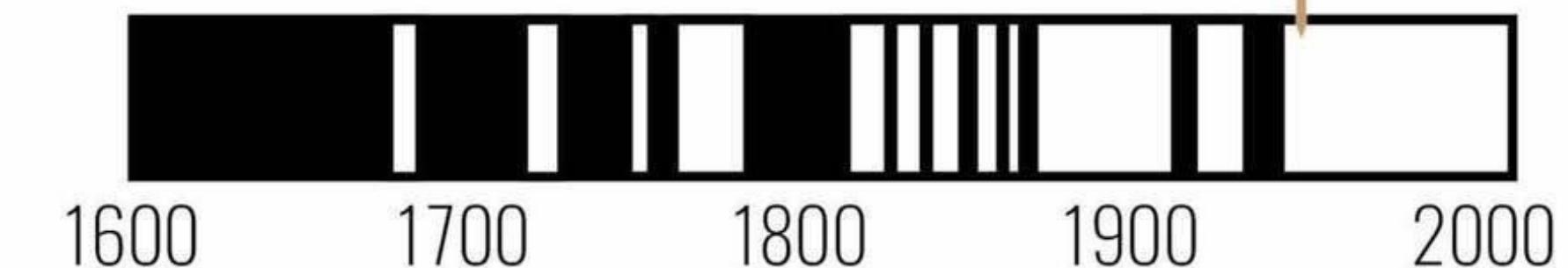
US cities 2003

Heavy tails appear in all social systems, also in natural systems

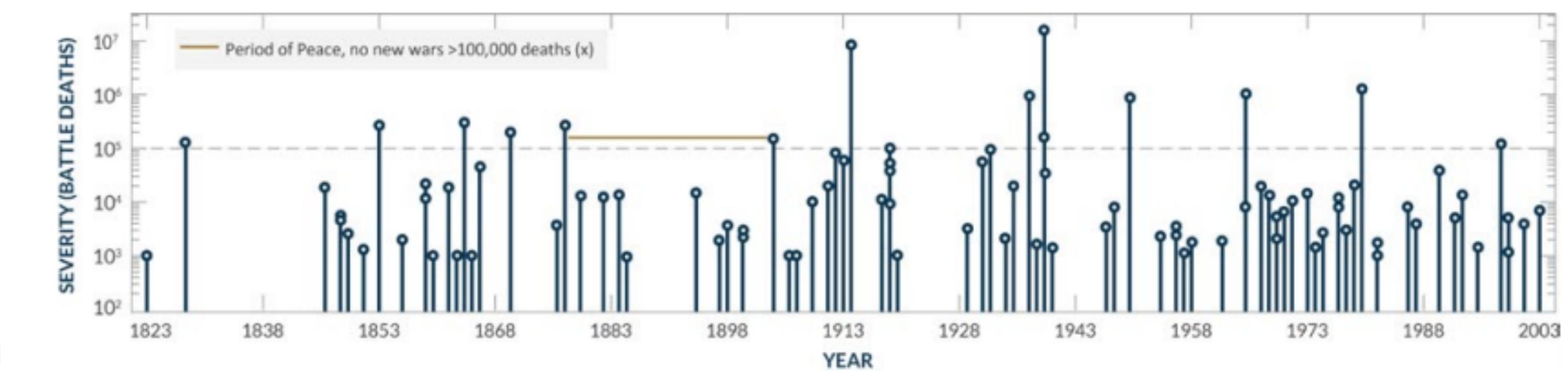


European History

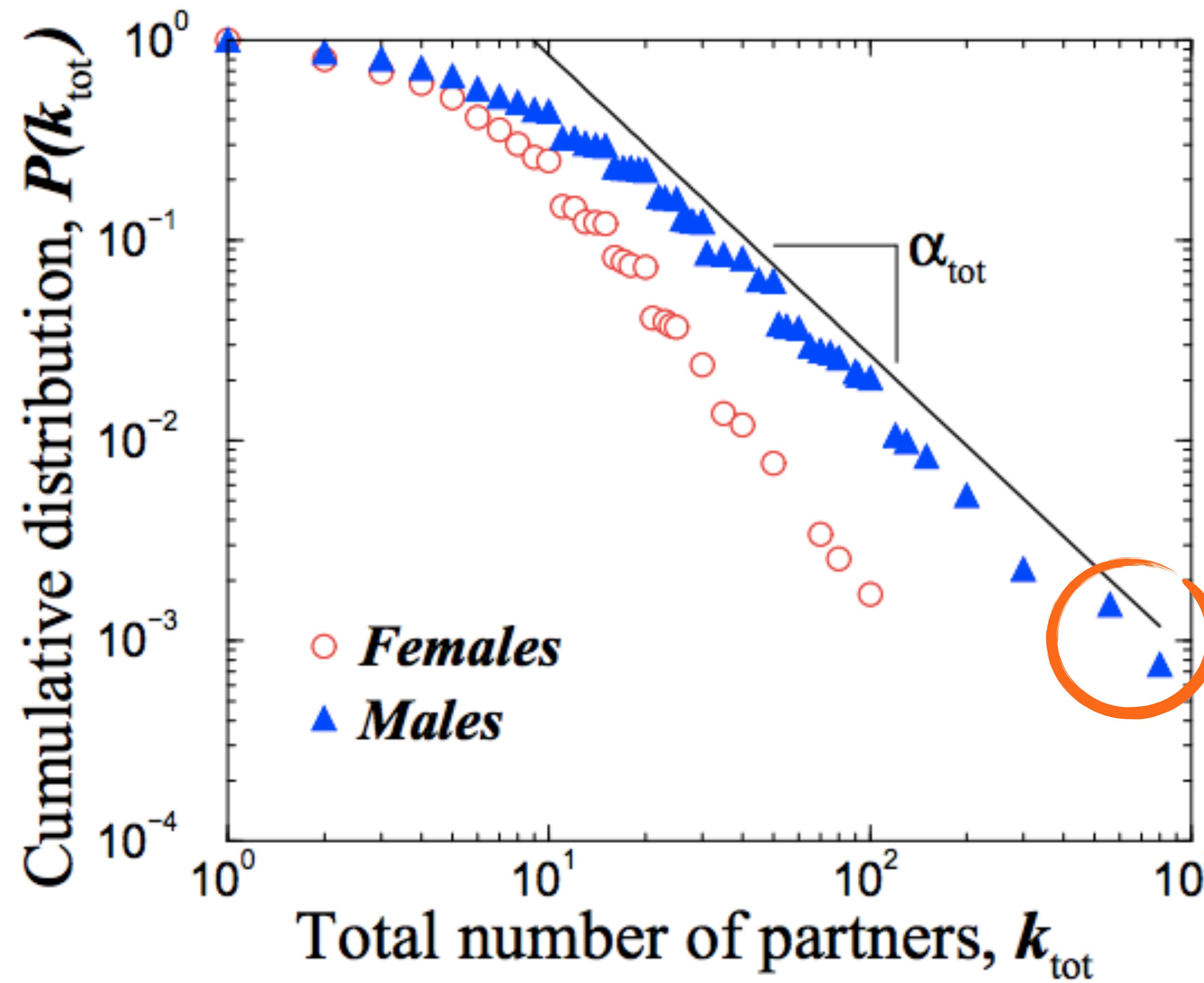
□ Peace ■ War



Major conflicts of original EU members



Extreme events are usually not outliers, but part of the system



Extreme events are usually not outliers, but part of the system

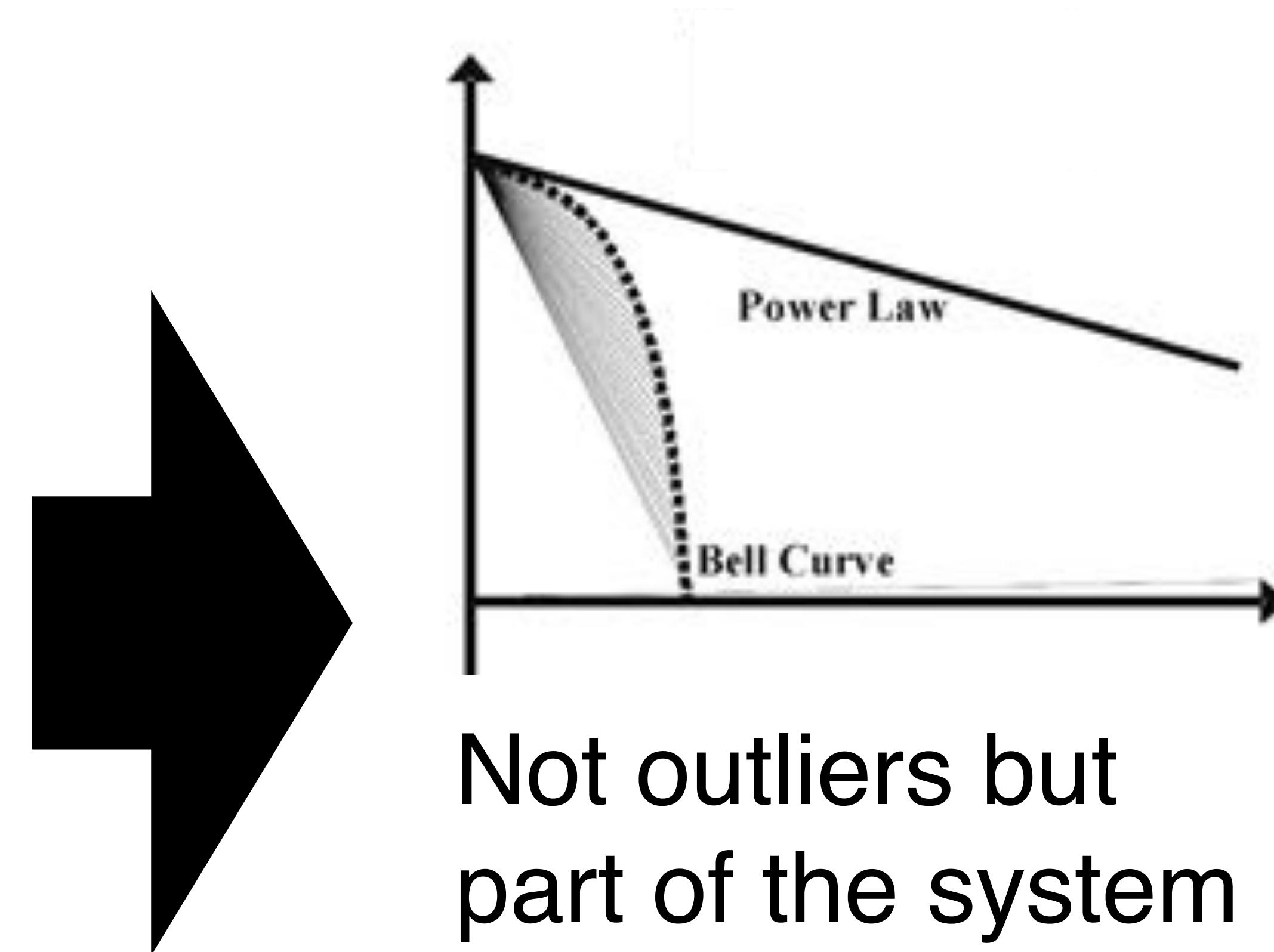
20th century statistics



Strange outliers

Focus on the
head/center

21th century statistics add:



Not outliers but
part of the system

Focus on the tail

Today we learned about the abundance of skewed data

Skewed data are inherent
to social systems



The log transform helps
to analyze skewed data

The CCDF allows to
investigate heavy tails

