



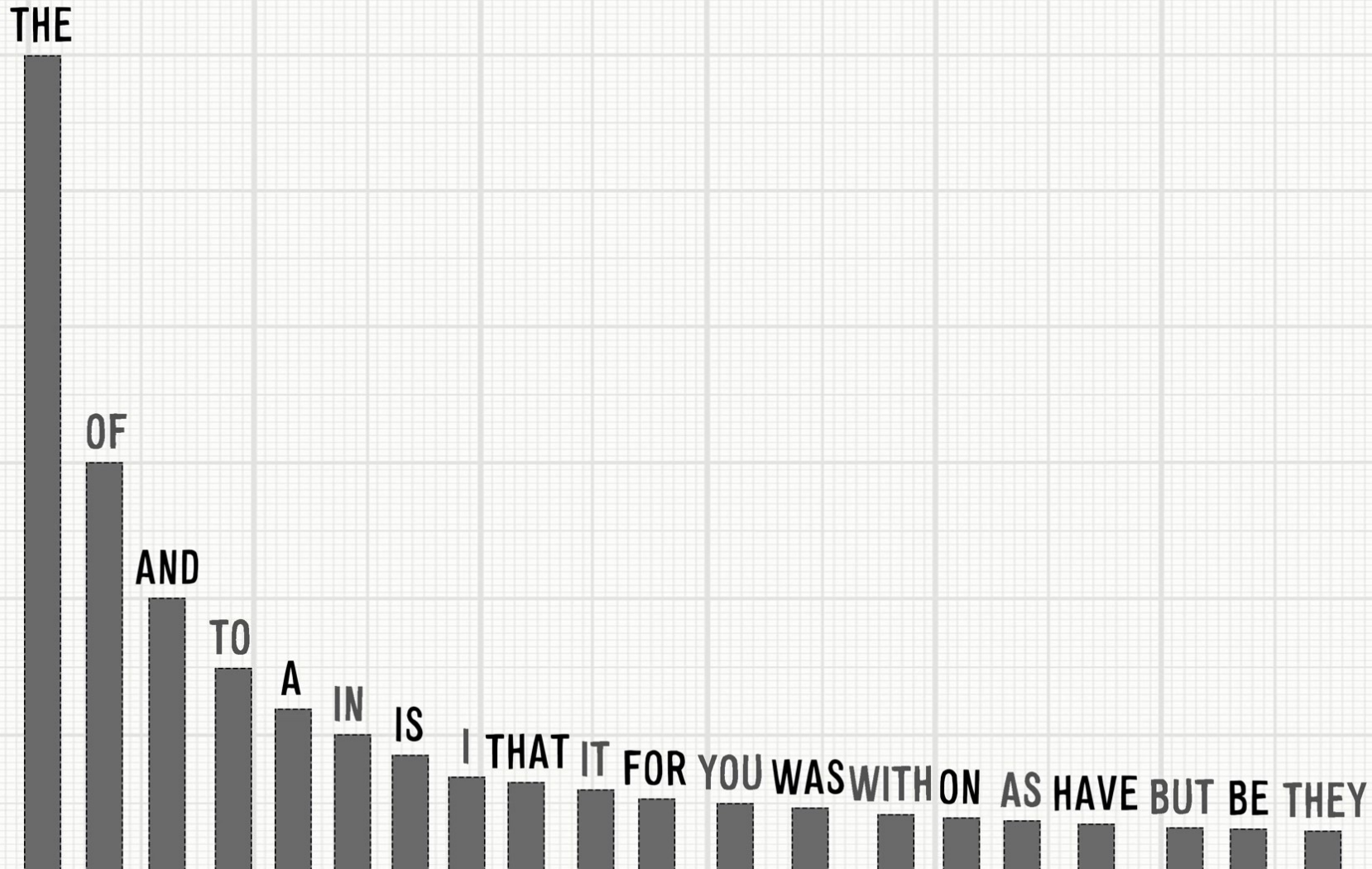
# The Zipf Mystery

Presented By  
SIDDHARTH PANT(1255613)

# Most Frequently used words

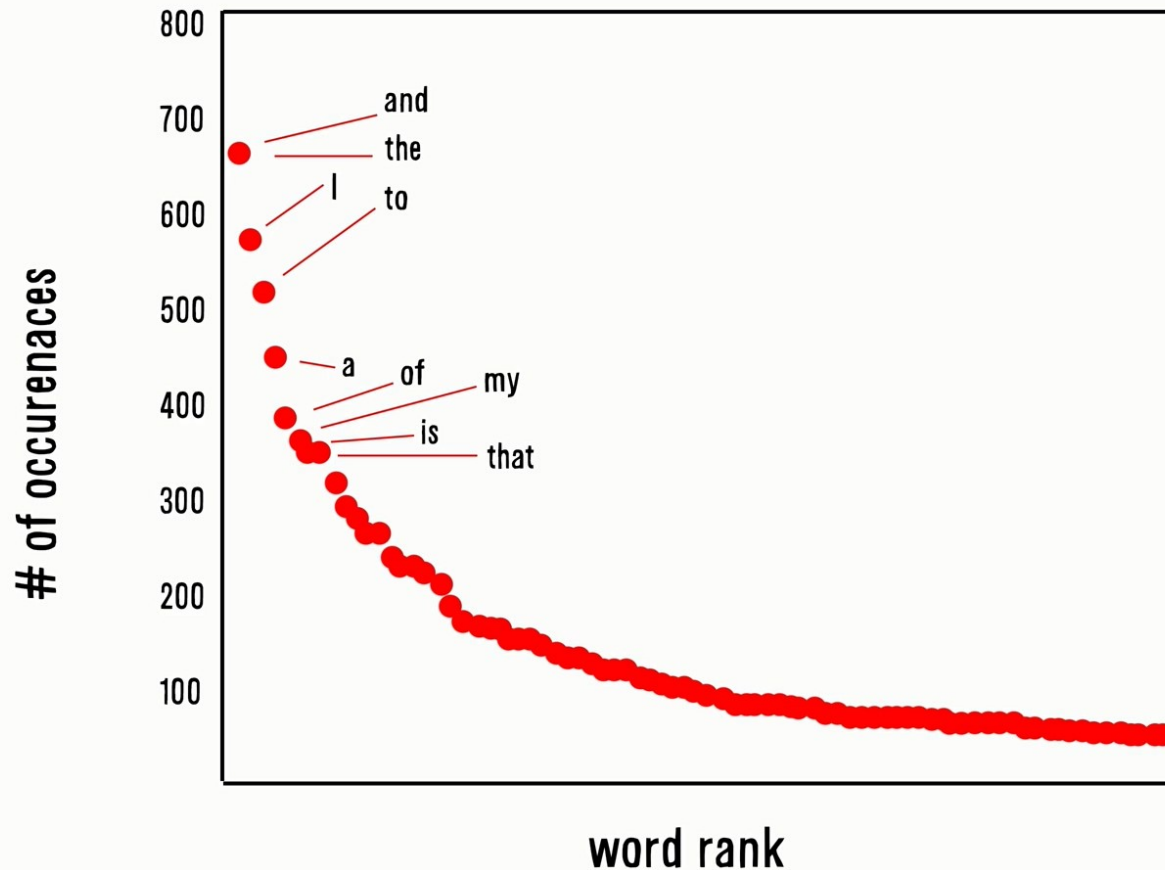
THE OF AND TO A IN IS I THAT IT FOR YOU WAS WITH ON AS HAVE BUT BE THEY

# Most Frequently used words



# Every novel every book follows this...

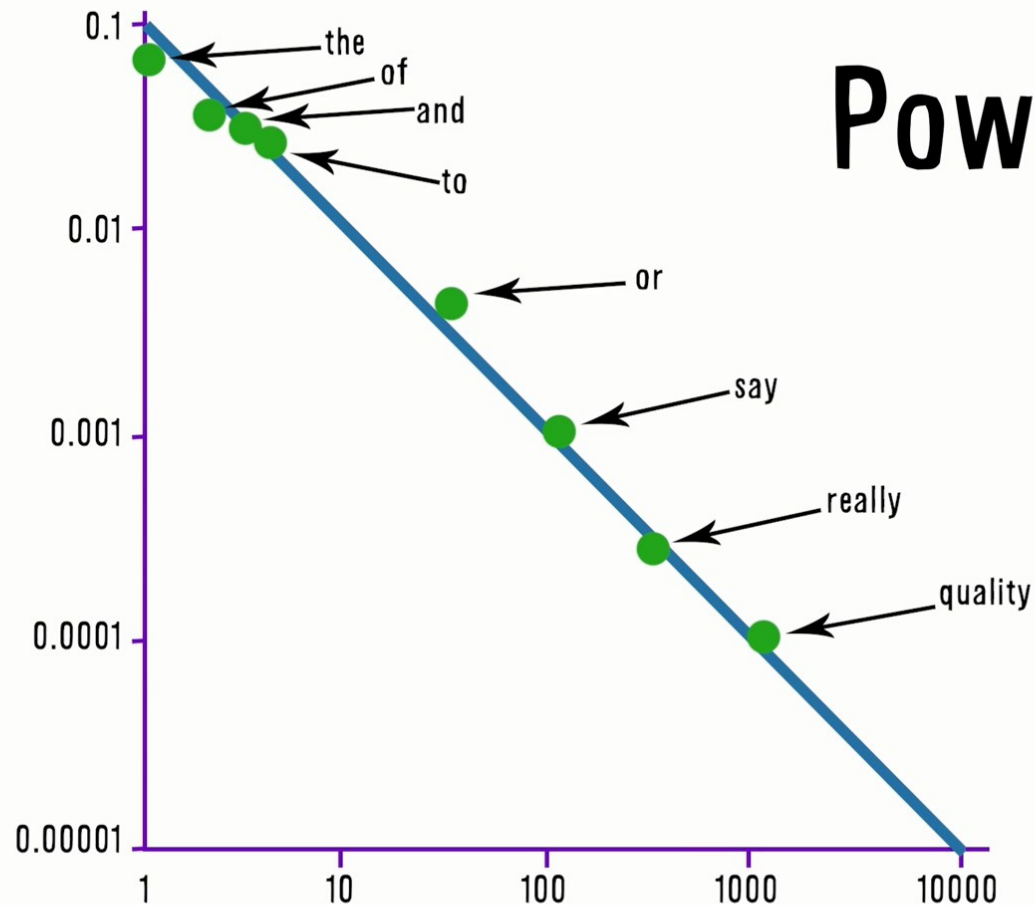
## word frequency and rank in *Romeo and Juliet* (linear-linear)



# The Bizarre Rule of Thumb

1  
—  
rank

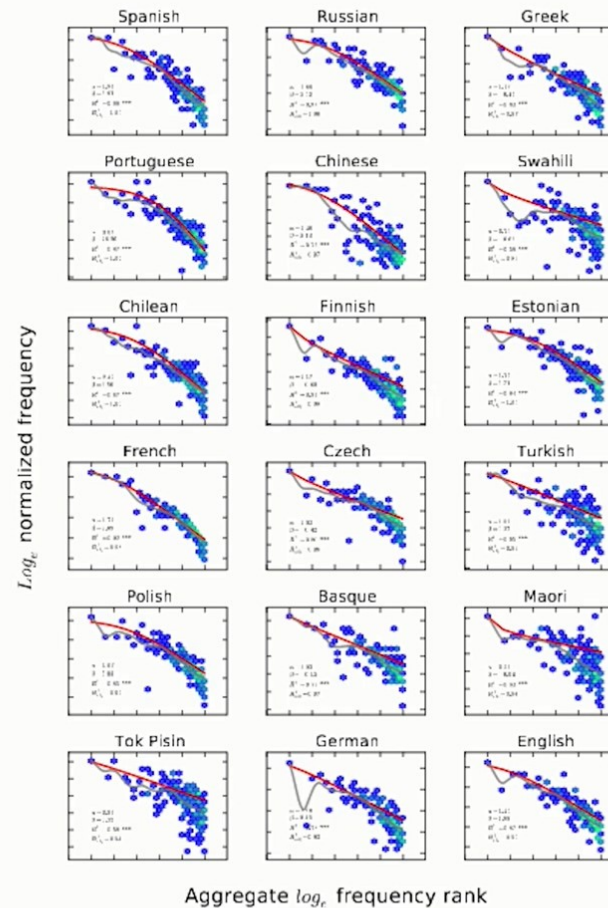
All nicely aligned on a line



The name of the game

**ZIPF'S  
LAW**

# Surprise surprise...Every language has it



Aggregate  $\text{log}_e$  frequency rank



# Let's do some crazy math shall we?

## 5,555

WORDCOUNT

◀ PREVIOUS WORD      NEXT WORD ▶

sauce cheltenham shelf interference

5555      5556      5557      5558

CURRENT WORD

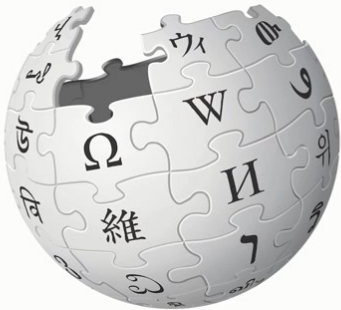
FIND WORD:  ▶ BY RANK:  ▶ REQUESTED WORD: SAUCE RANK: 5555

86800 WORDS IN ARCHIVE  
ABOUT WORDCOUNT

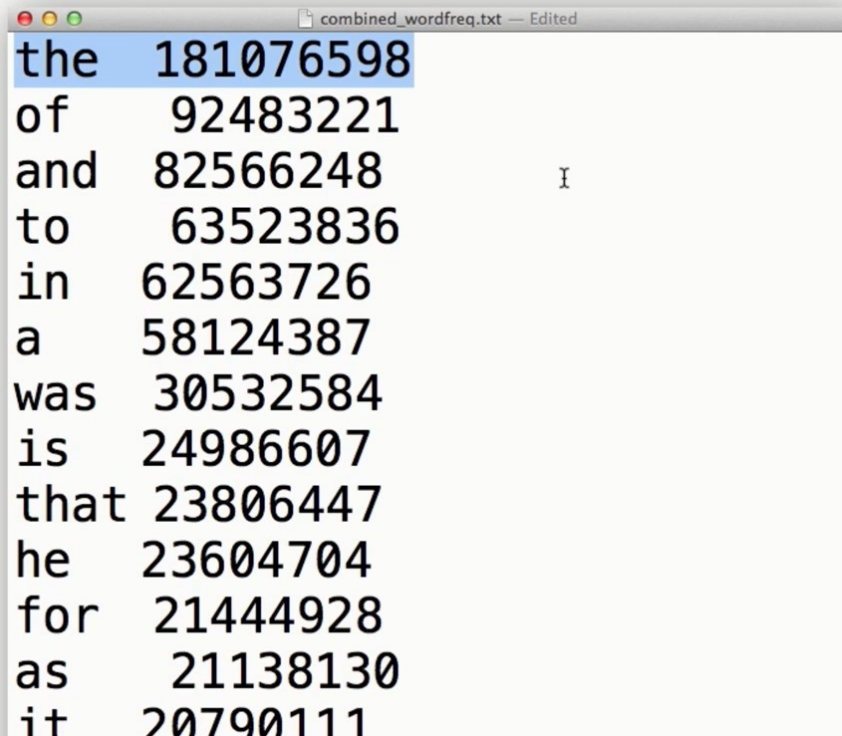
Wordcount ©2003 | Jonathan Harris | Help

Just bear for a second more...

$$181 \text{ million} \times \frac{1}{5,555} = 30,000$$



**WIKIPEDIA**  
The Free Encyclopedia

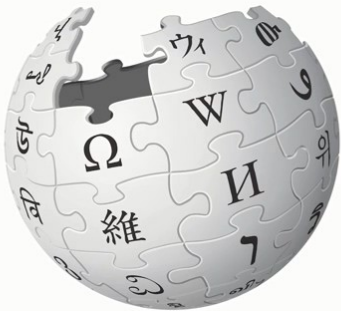


the	181076598
of	92483221
and	82566248
to	63523836
in	62563726
a	58124387
was	30532584
is	24986607
that	23806447
he	23604704
for	21444928
as	21138130
it	20790111



And now see the magic

$$181 \text{ million} \times \frac{1}{5,555} = 30,000$$



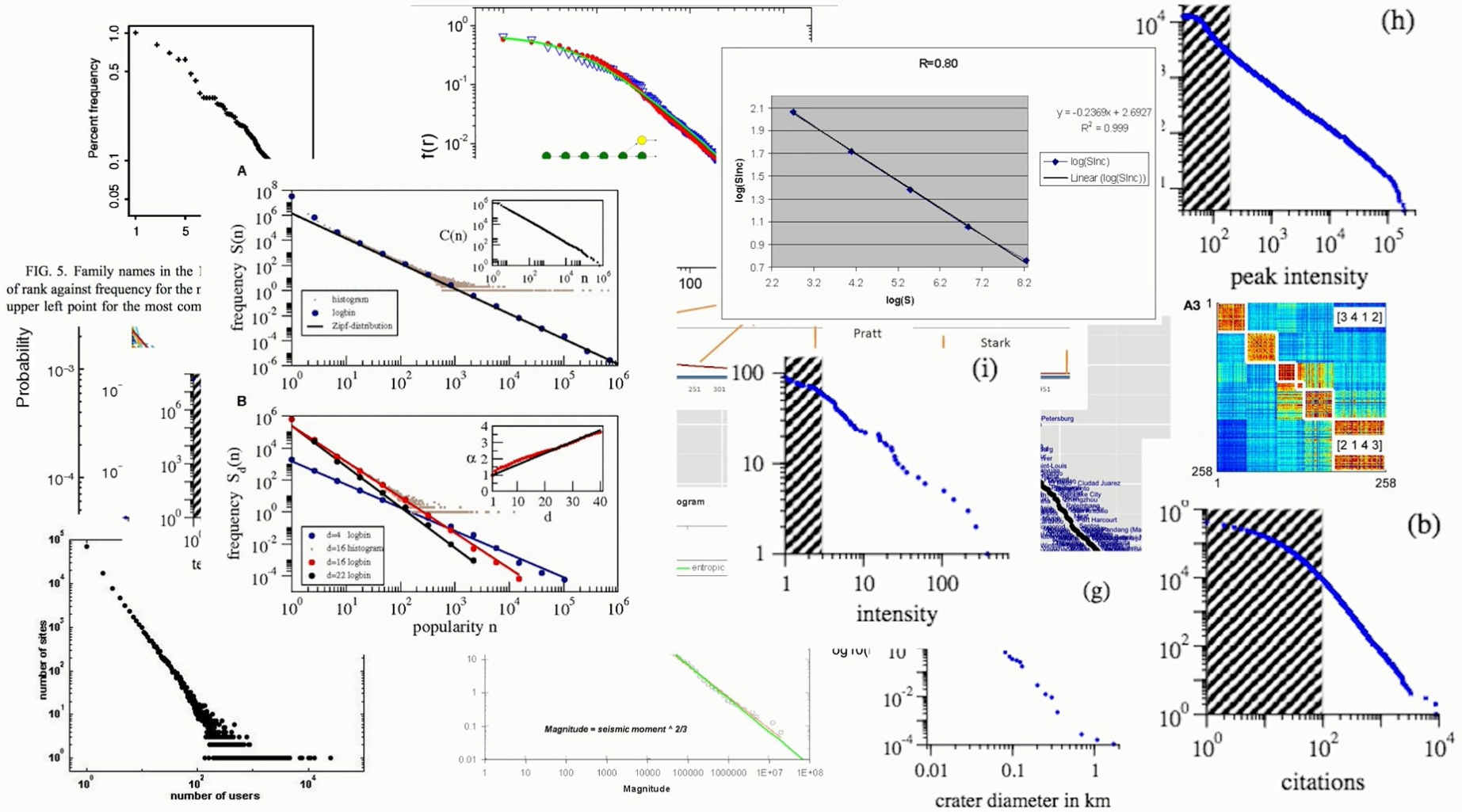
**WIKIPEDIA**  
The Free Encyclopedia

combined\_wordfreq.txt — Edited  
Searching... Done Replace

convoy	29622
parking	29611
gladly	29610
gerald	29608
bending	29604
clause	29595
decisive	29595
assumption	29594
sauce	29594 <sup>1</sup>
jose	29591
shapes	29580
whoever	29569



# The World is “Zipfy”



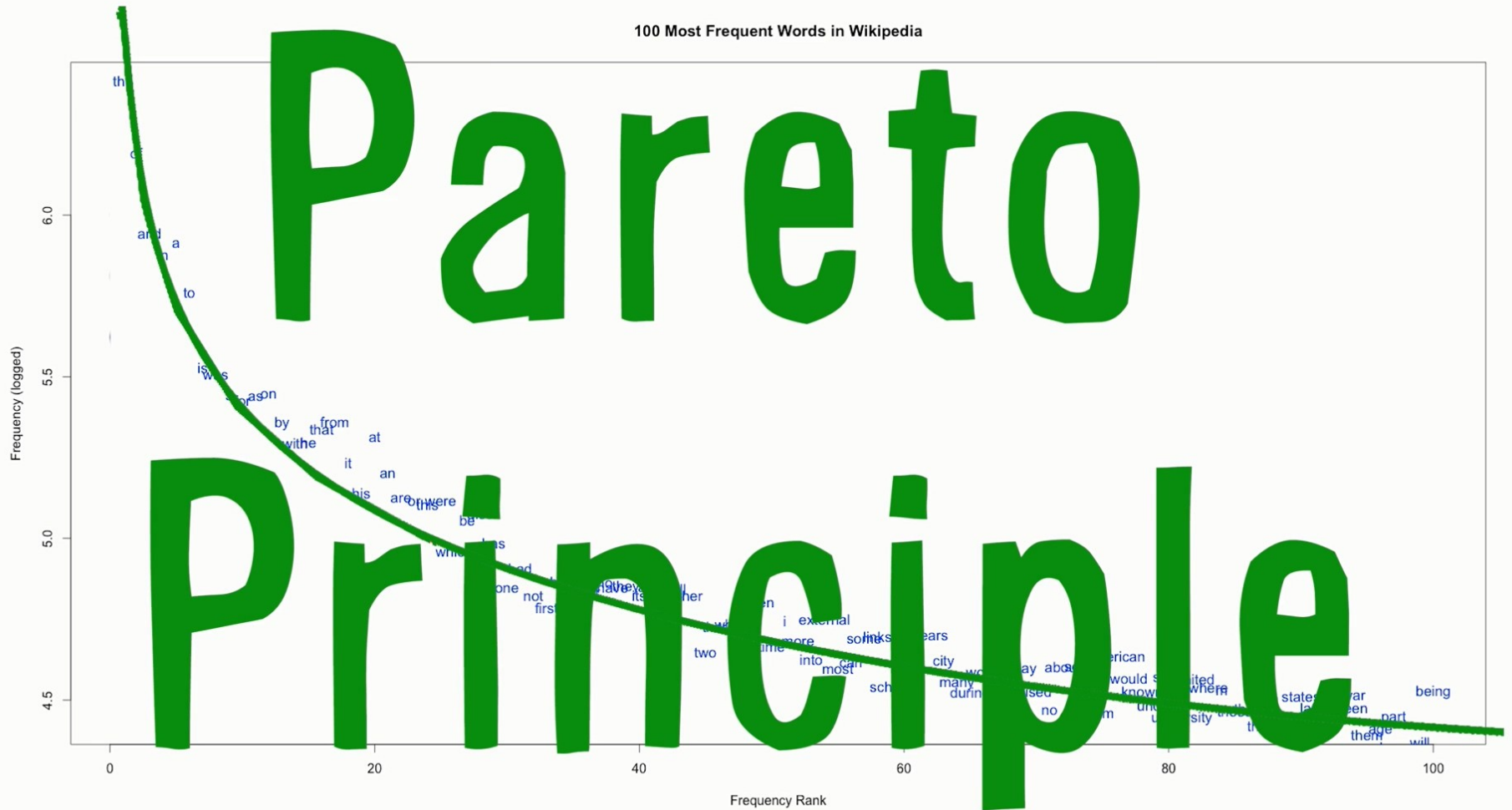
**And the name comes from here**



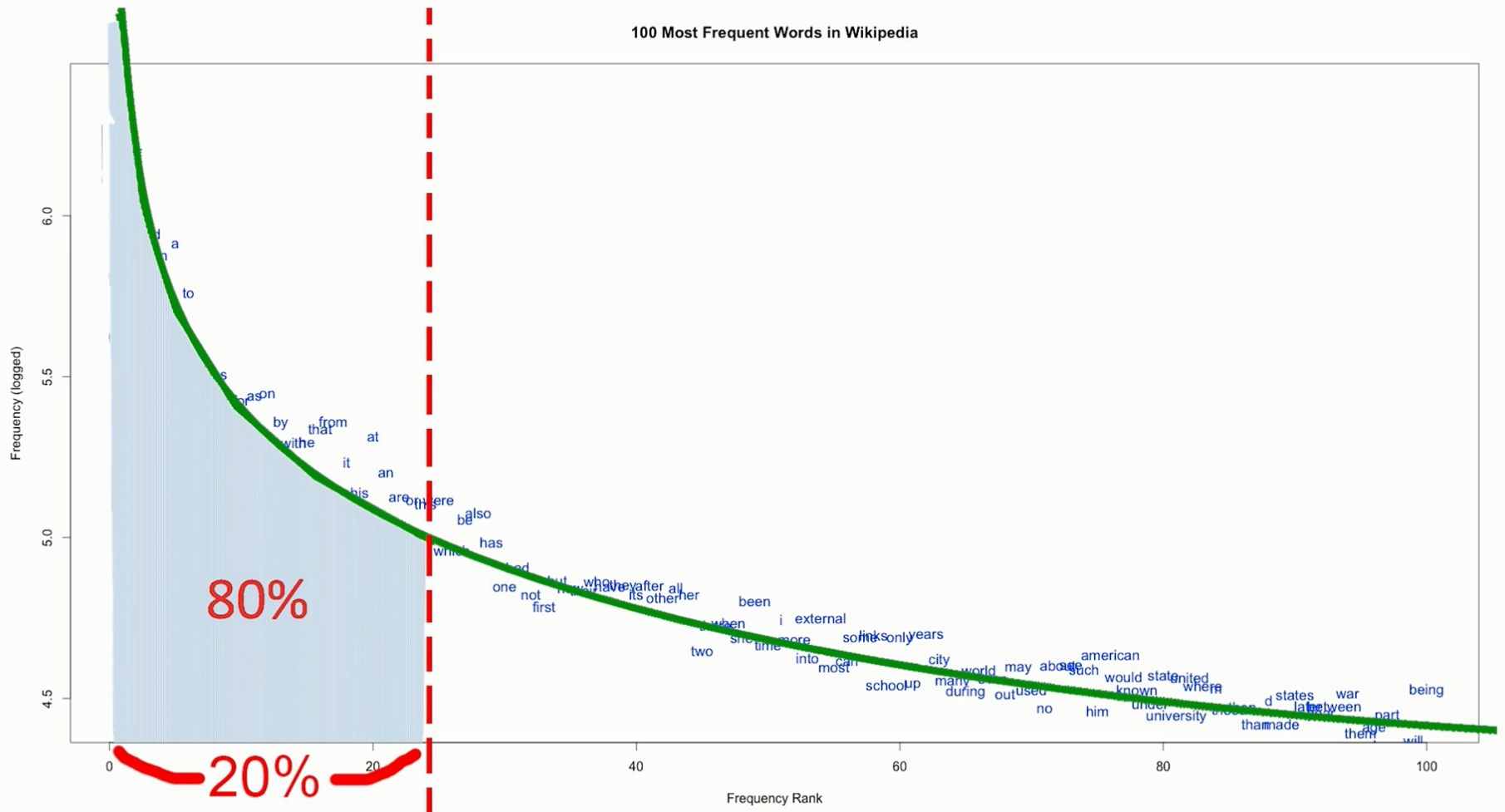
**George Zipf**



It's about to get interesting



# The Bizarre Rule of Thumb



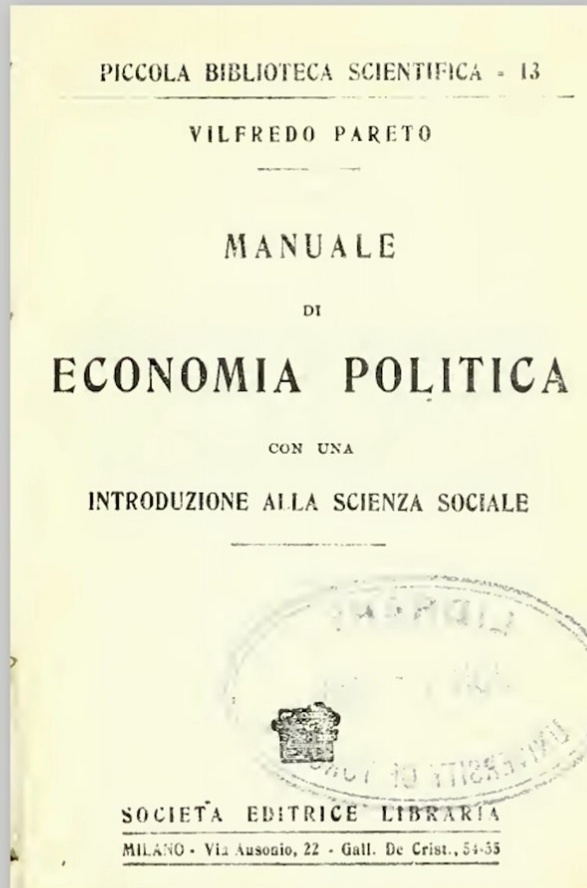
**And this name comes from here**



**Vilfredo Pareto**



# It affects economic ratios



80%

20%

## Even the growing of peas



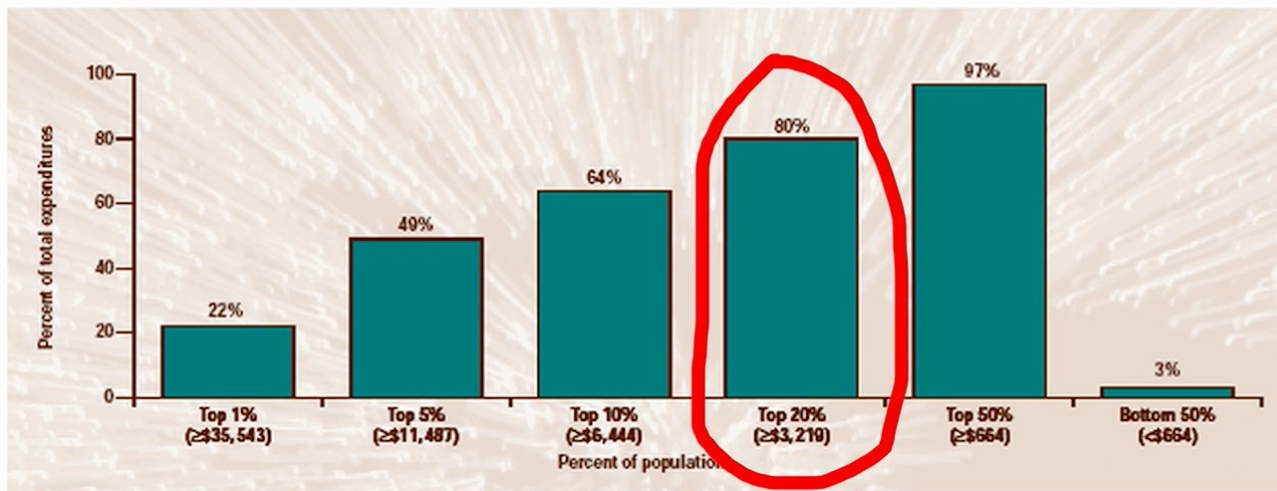
80%

20%

# Or the health of a nation's people

## Chart 1. Percent of Total Health Care Expenses Incurred by Different Percentiles of U.S. Population: 2002

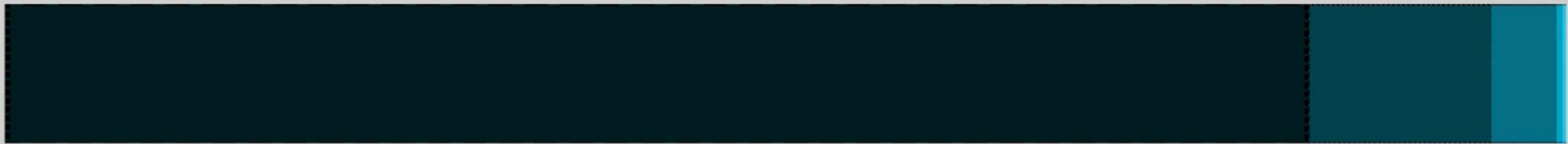
Research in Action, Issue 19



**Source:** Conwell LJ, Cohen JW. Characteristics of people with high medical expenses in the U.S. civilian noninstitutionalized population, 2002. *Statistical Brief #73*. March 2005. Agency for Healthcare Research and Quality, Rockville, MD. Web site: [http://meps.ahrq.gov/mepsweb/data\\_files/publications/st73/stat73.pdf](http://meps.ahrq.gov/mepsweb/data_files/publications/st73/stat73.pdf). Accessed April 7, 2006.

# And of course their wealth

● Top 20%   ● 2nd 20%   ● middle 20%   ● 4th 20%   ● bottom 20%





# Thank You

IT Services  
Business Solutions  
Consulting