Lesson3:
Modeling the Web with Advanced Statistical Descriptive Text Models
Unit1:
A closer look at the word rank frequency plot – Introducing Zipf's law

Rene Pickhardt

Introduction to Web Science Part 2

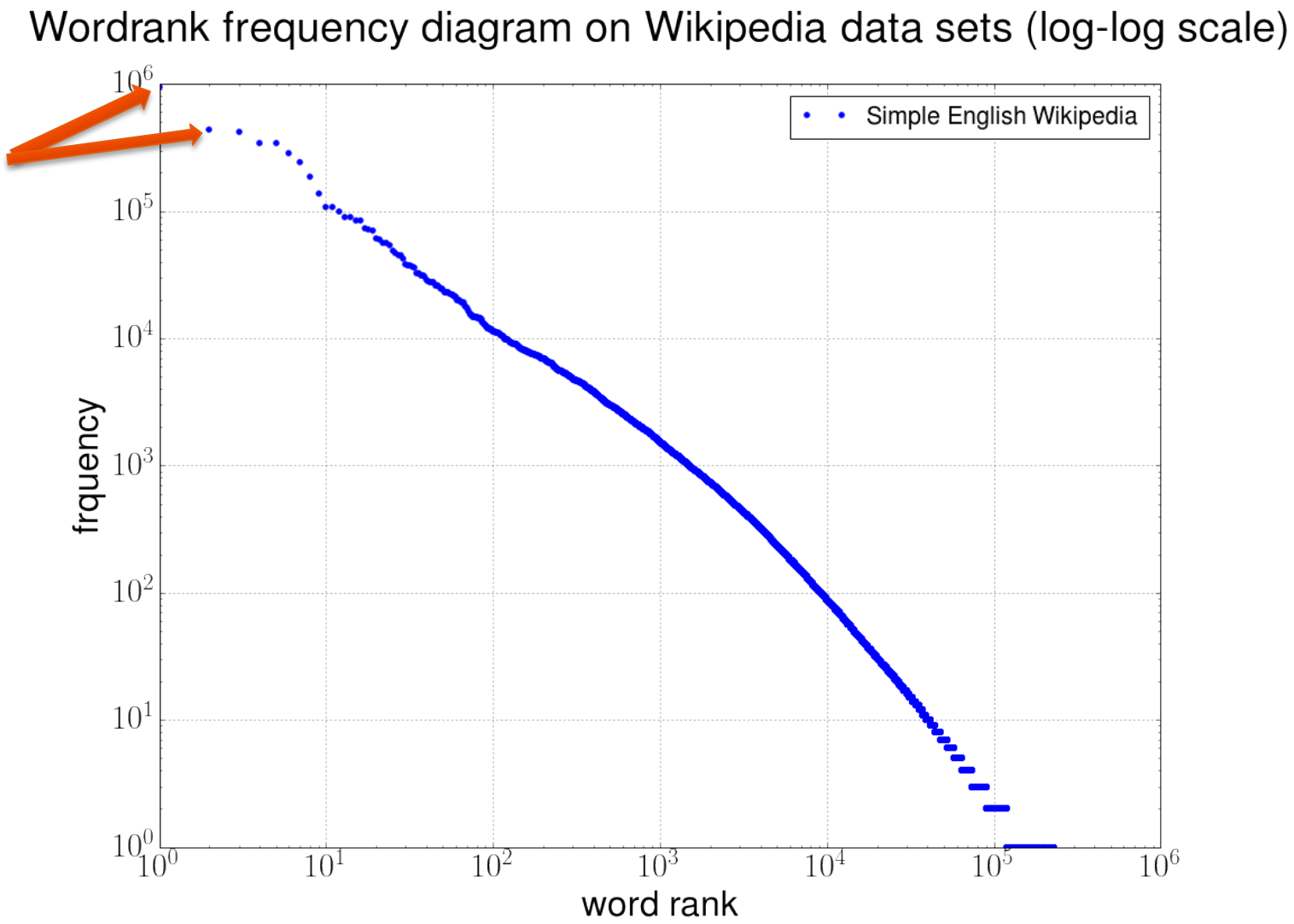
Emerging Web Properties

WeST
People and Knowledge Networks

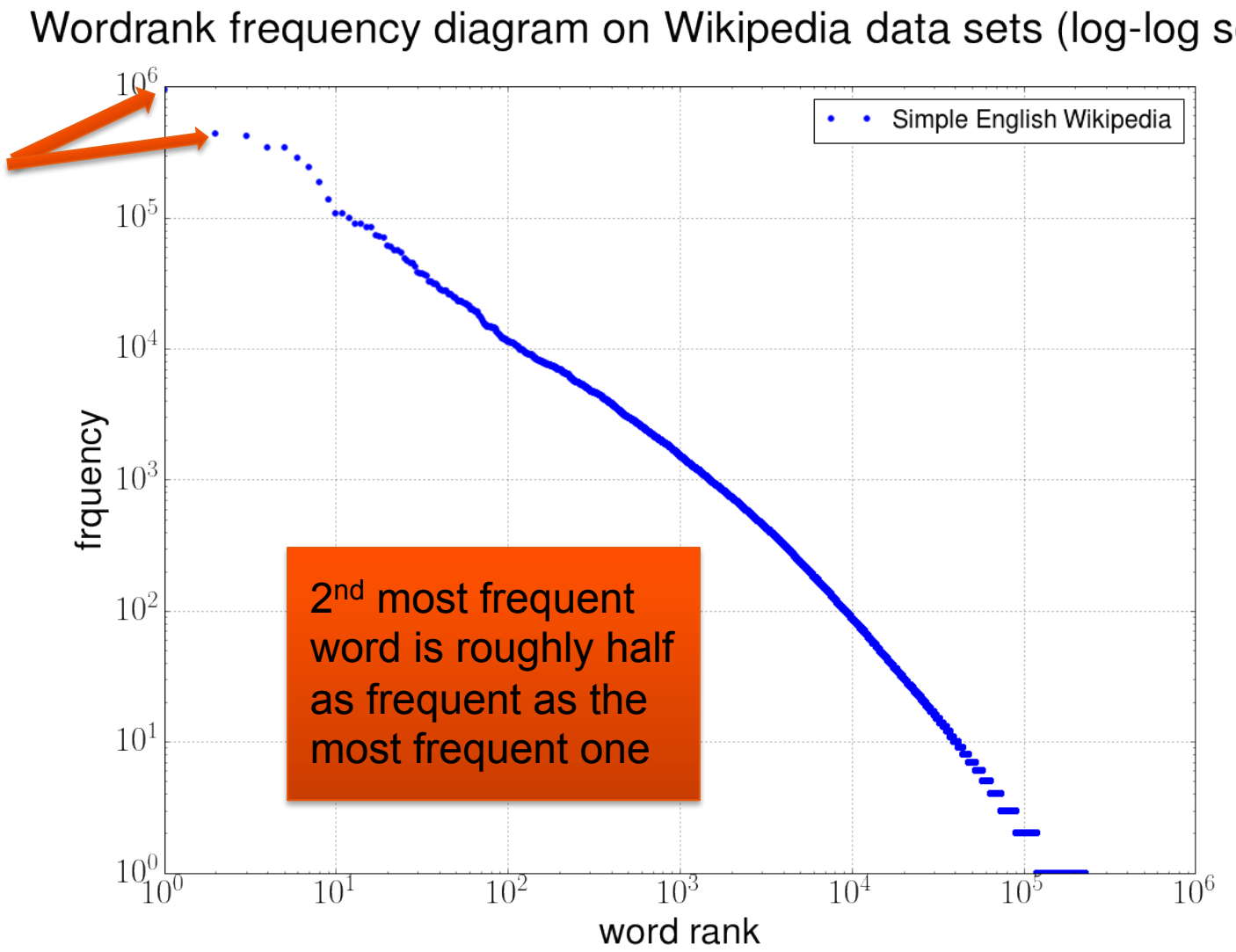
Completing this unit you should

- Be able to name some fundamental properties about how frequencies of words in texts are distributed
- Be a little bit more cautious about visual impressions when looking at log-log plots
- Know both formulations of Zipf's law

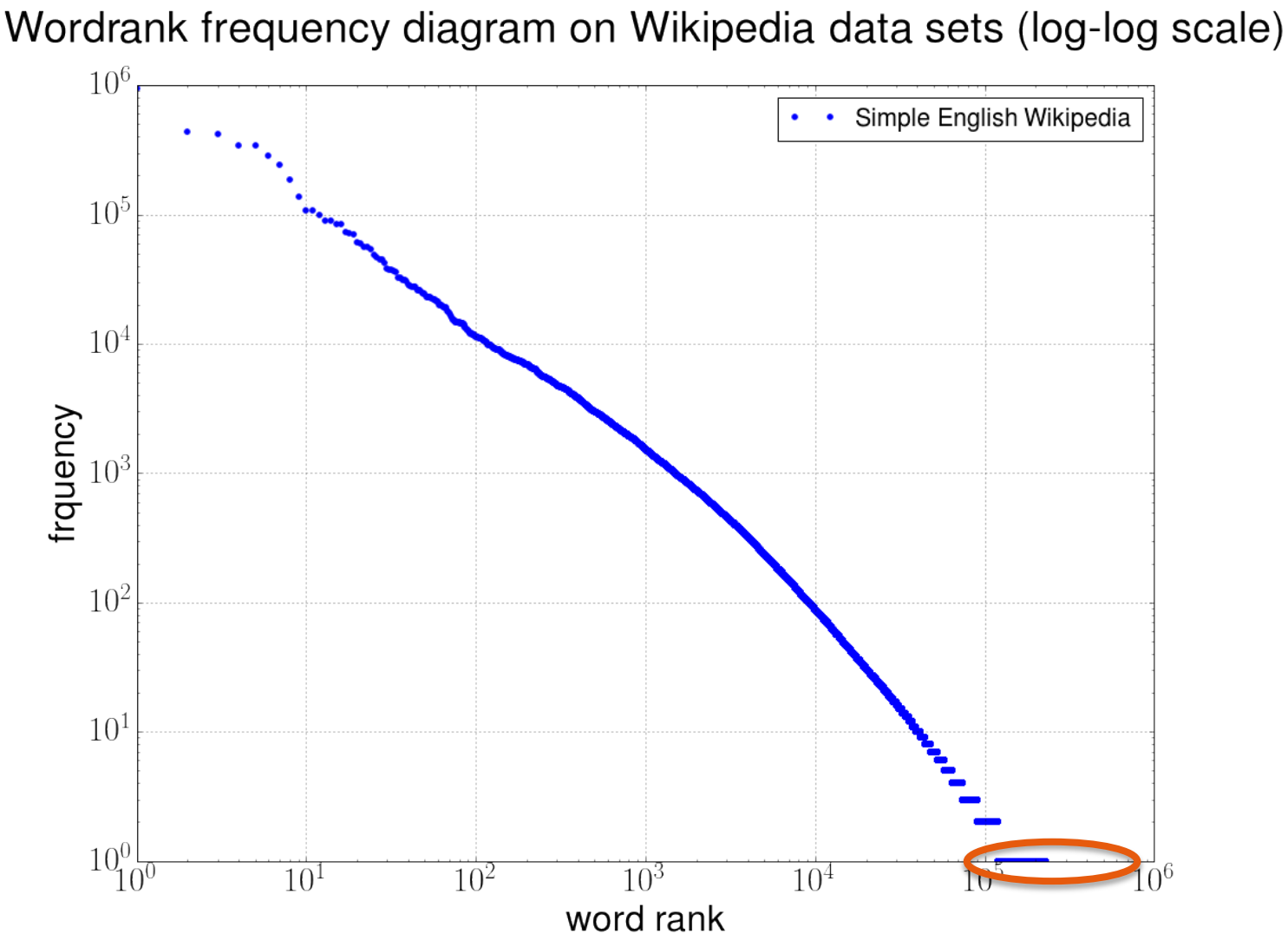
What can we say about these points?



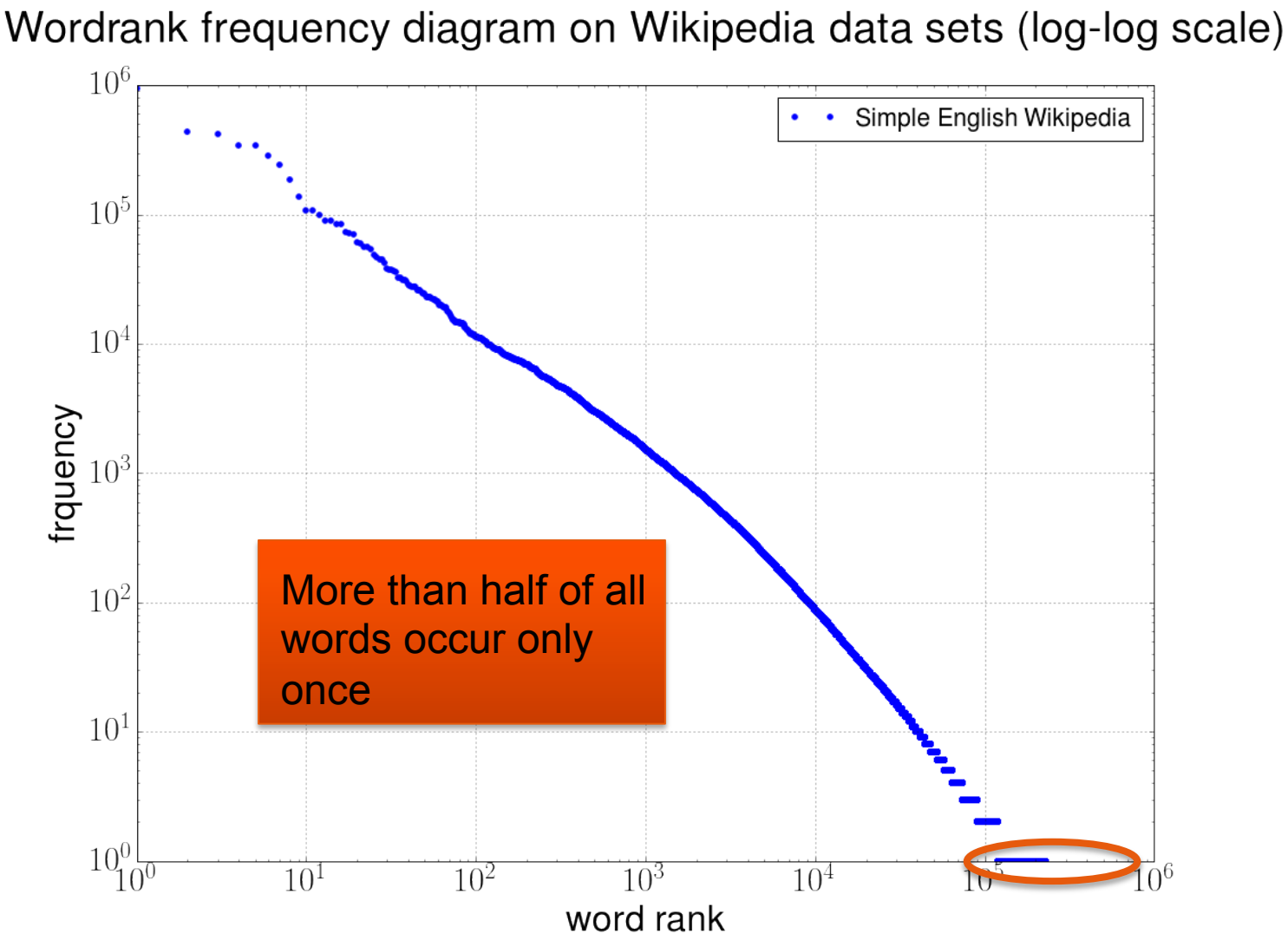
What can we say about these points?



What about those points?



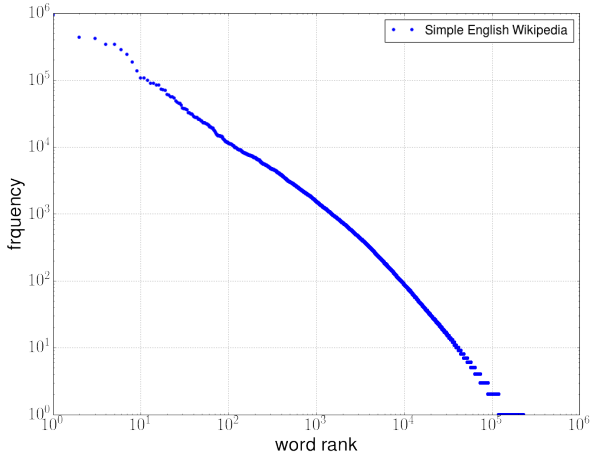
Don’t get tricked by visual appearances



Summary of observations

- Frequencies of words are distributed very “unfair”
- Only few words have a high frequency
- Most words occur only once
- Most frequent word occurs already almost twice as much as second most frequent

Wordrank frequency diagram on Wikipedia data sets (log-log scale)



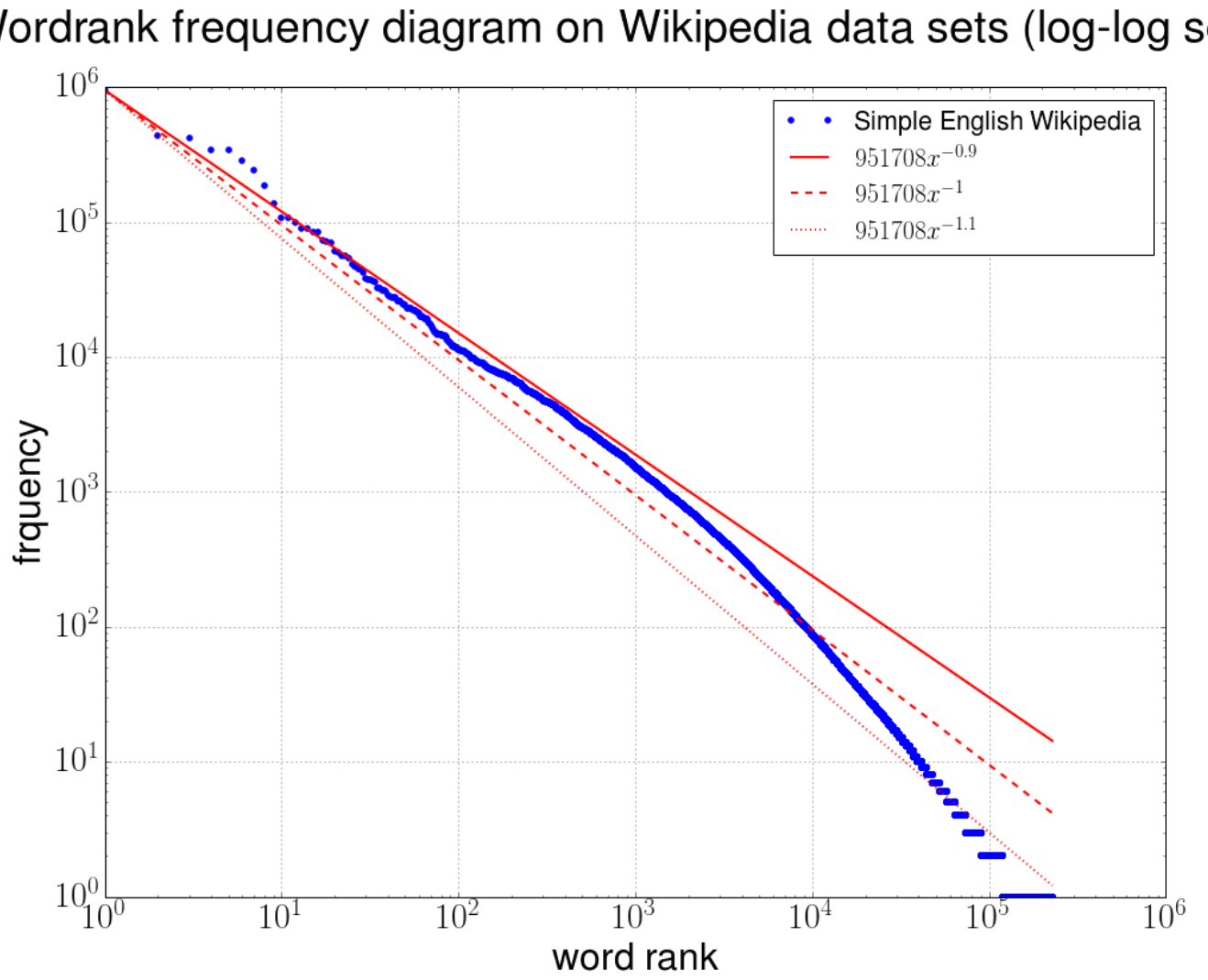
Harvard based Linguist Goerge Kinsley Zipf studied these phenomena

- In 1930 far before we had todays computing power or the internet

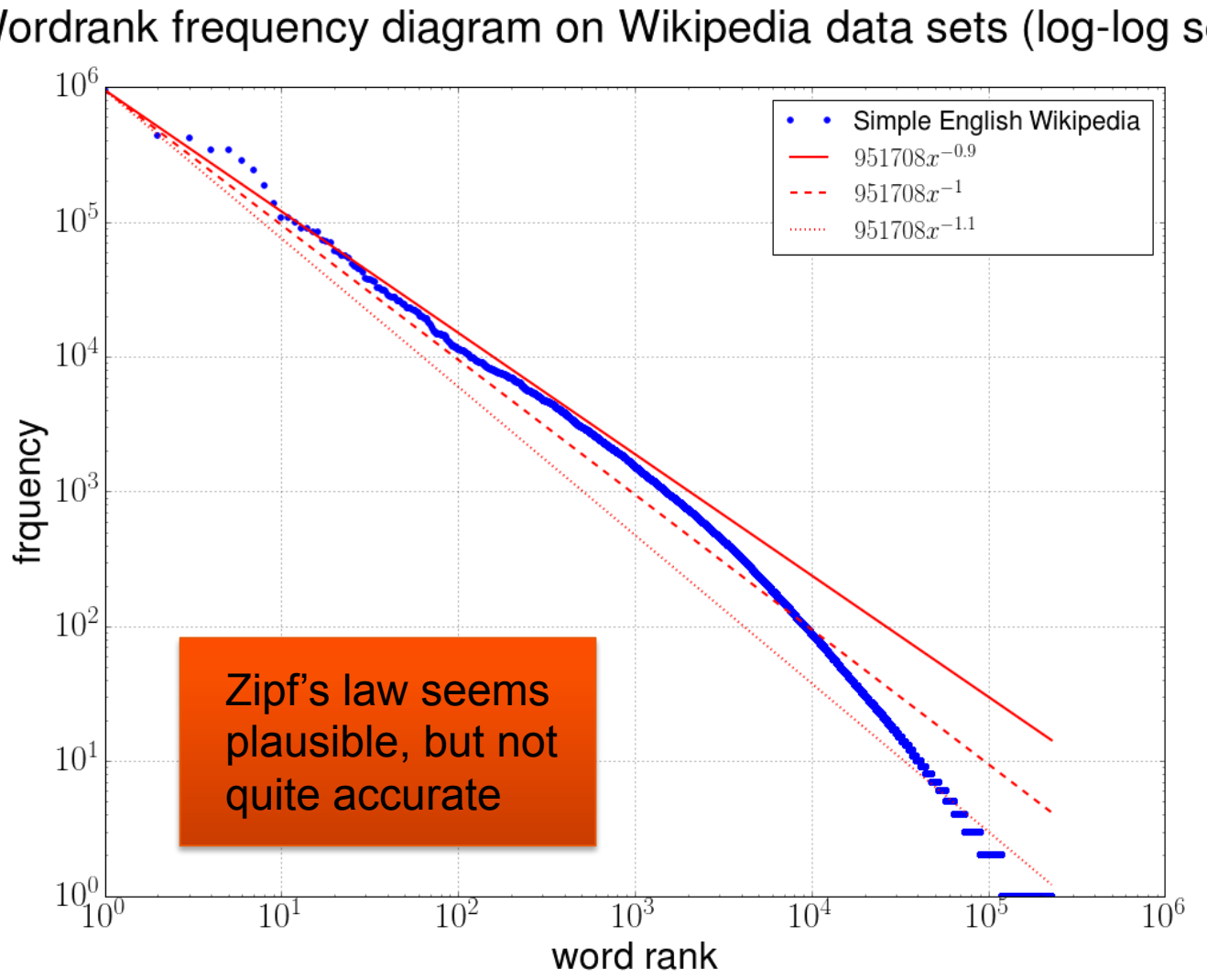
- He proposed a law saying:
  - Word rank multiplied with the frequency is constant

- It is often formulated as $f \sim 1/r^k$ with k having roughly a value of 1.

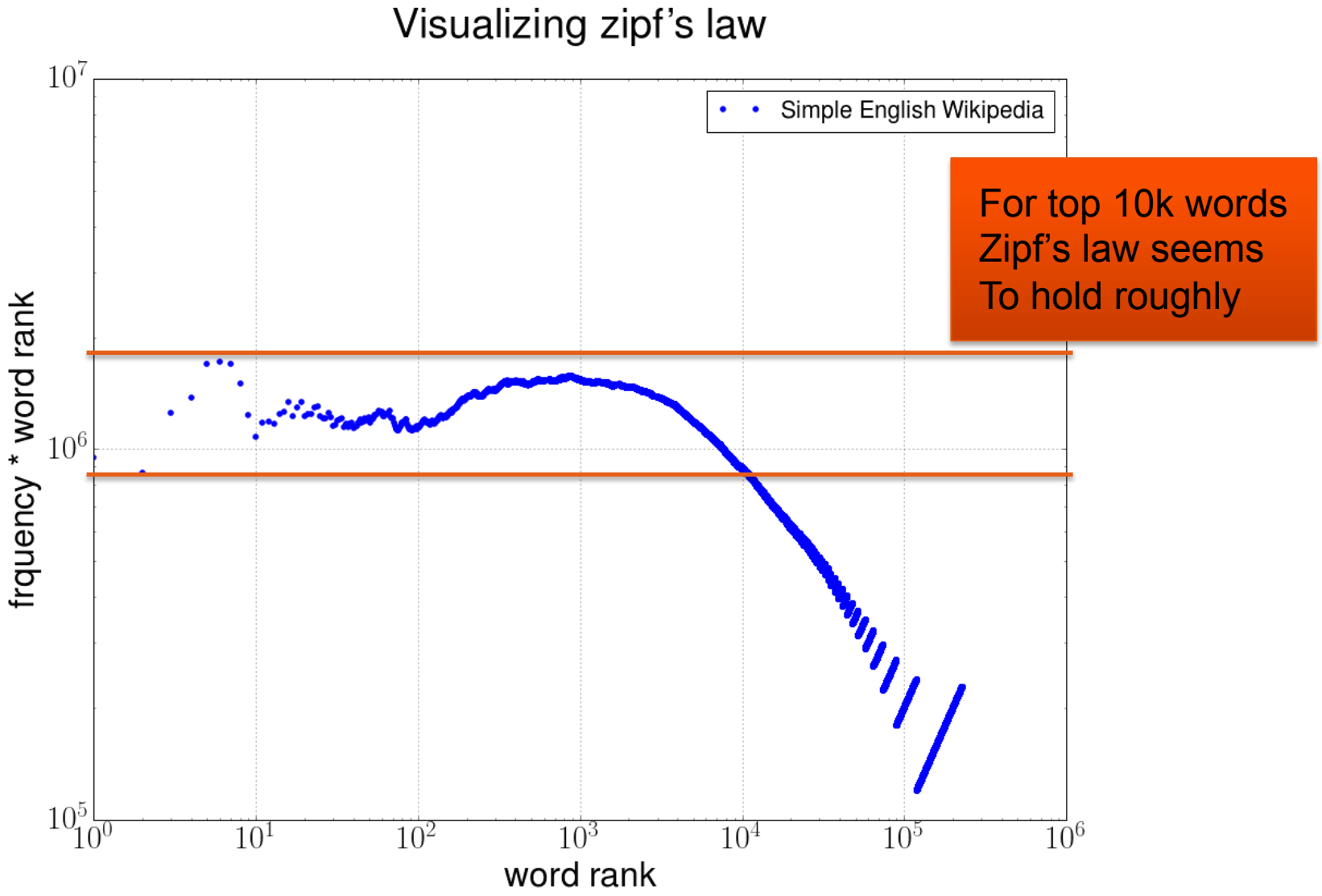
Lets look at the reformulation of Zipf's law



Lets look at the reformulation of Zipf's law



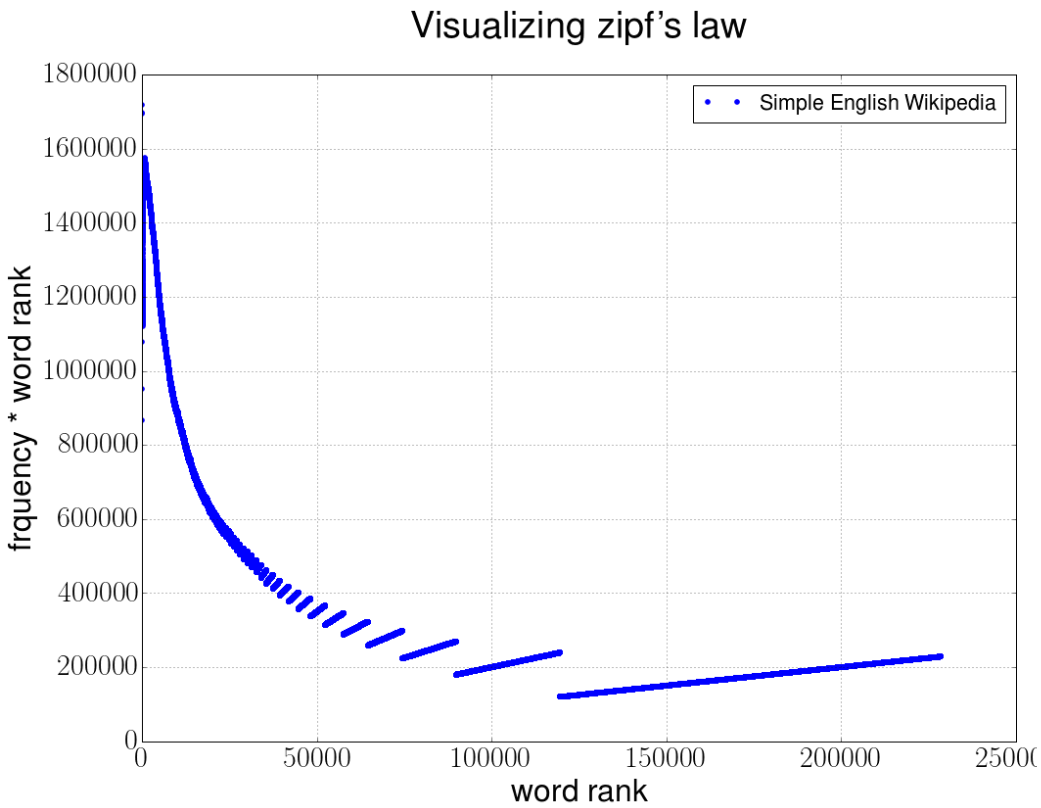
What about the original version of the law?



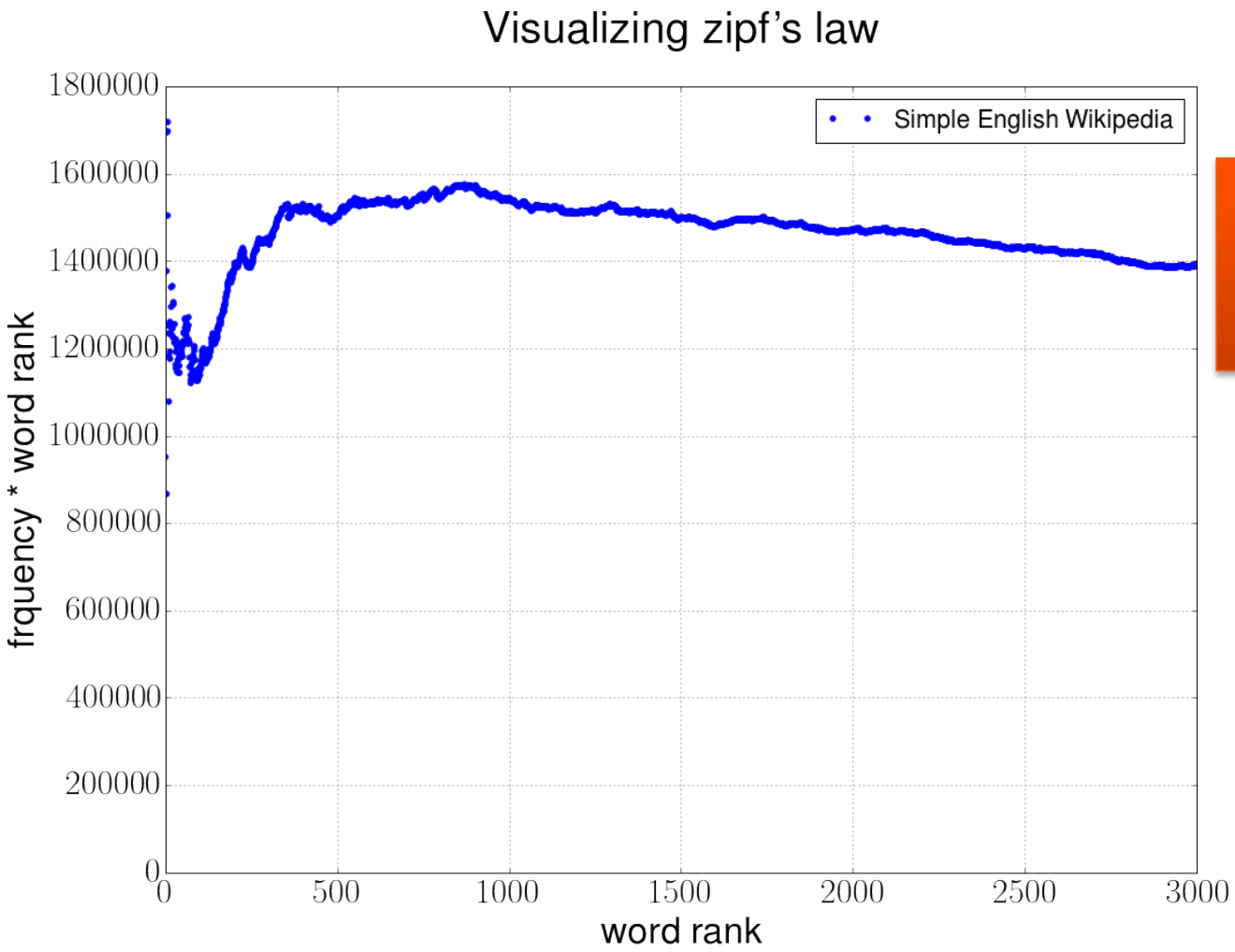
For top 10k words
Zipf's law seems
To hold roughly

Same (!) plot with linear scales

- How could one ever deduce a law like Zipf's law from this curve?
- Remember Zipf did not have nearly as much data as we
- Nor computational power and means of visualization



Looking at top ranked words (linear scale)



Zipf probably did not have the chance to see more

Thank you for your attention!

Contact:

Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST
People and Knowledge Networks

Copyright: