Lesson2:
Generative Models for Text on the Web
Unit1:
Building (probabilistic) generative models

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

WeST
People and Knowledge Networks

Completing this unit you should

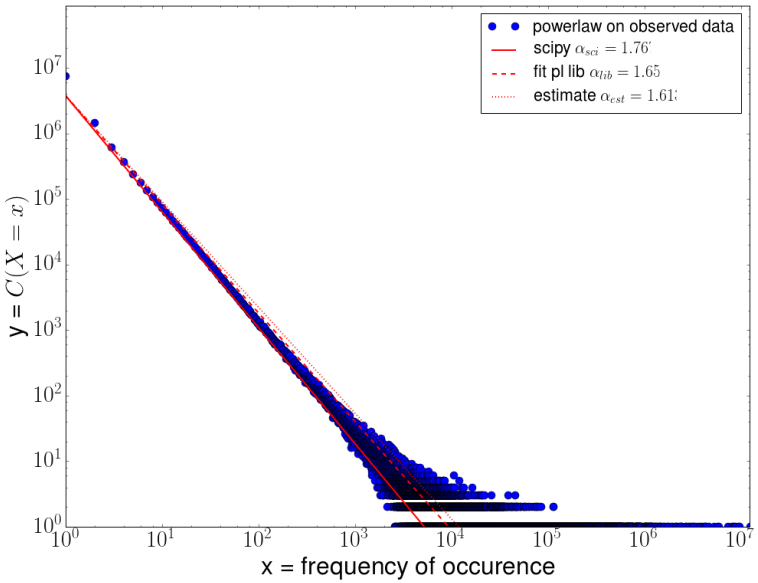
- Understand the principle methodology for building generative models
- Remember why people are interested in generative models
- Know why descriptive models are needed when evaluating a generative model
- Be aware of one way to create a model for text generation

Why are words distributed so unequally?

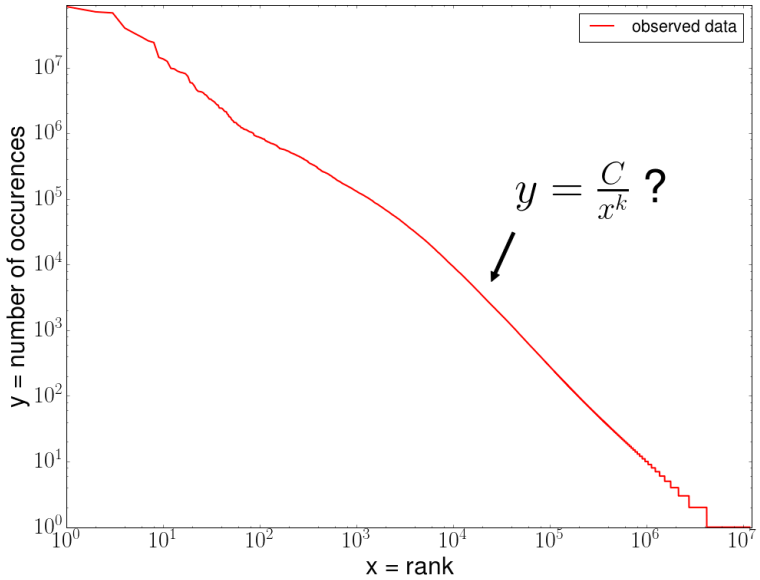
- We have seen that frequencies of word occurrences follow a power law
- Many words occur only once
- A few words occur very frequently
- Can we come up with an explanation?

In the past we just had a descriptive model

Words occuring exactly n times on Simple Englisch Wikipedia



Word frequencies depending on word rank on English wikipedia



Create a generative Model (few parameters!)

(Simulated) probabilistic model

Build a (simulated) probabilistic model

- Idea words are generated from characters
  - [a-z] and SPACE
  - Count in Simple English Wiki how often characters occur

- Make probabilities $$P(x) = \frac{c(x)}{\sum_y c(y)}$$

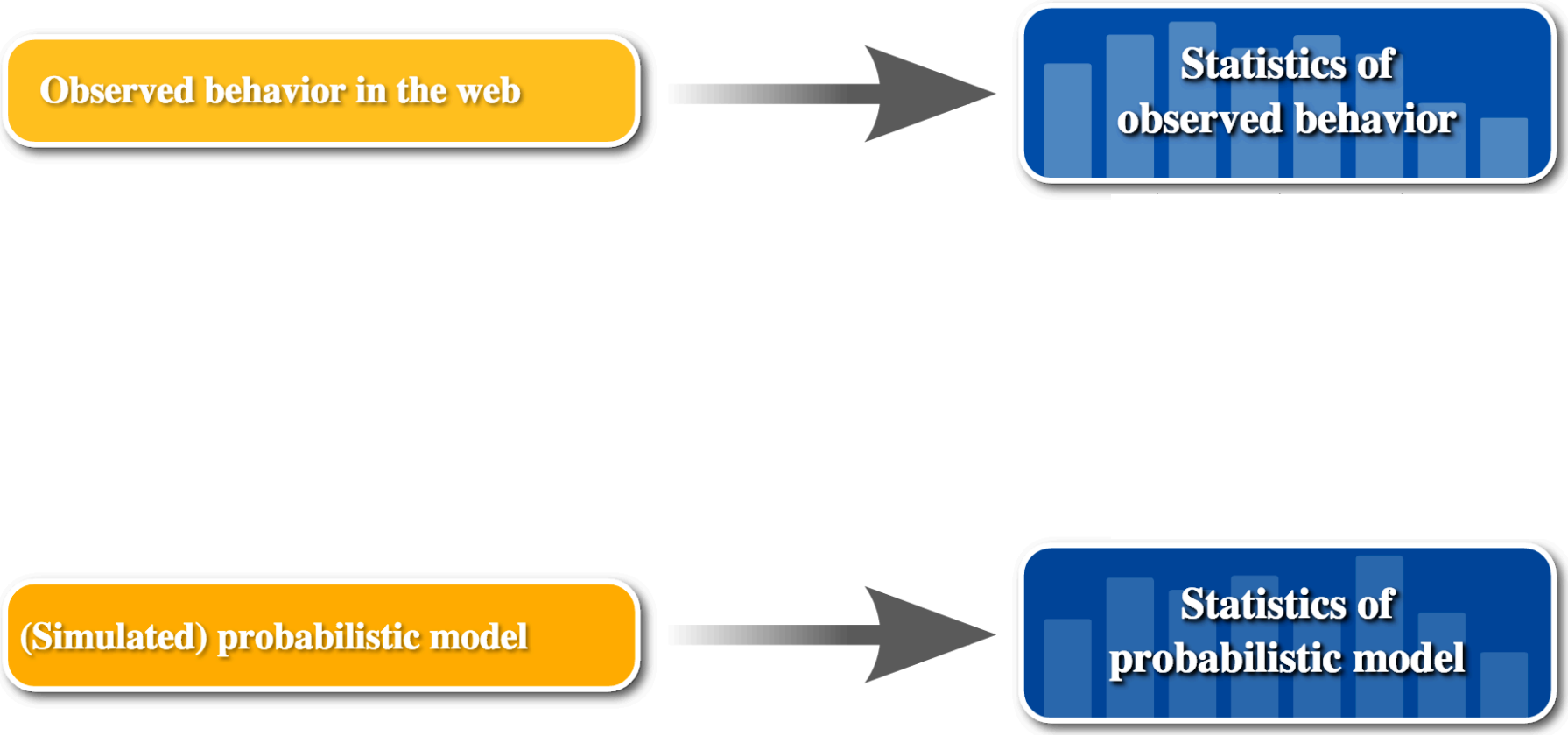
	‘ ‘	‘a‘	‘c‘	‘b‘	‘e‘	...	‘z‘
$c(x)$	30‘298	38‘090	506	1960	65‘818	...	7‘432
$P(x)$	0.138	0.173	0.002	0.009	0.299	...	0.034

What was our model parameter?

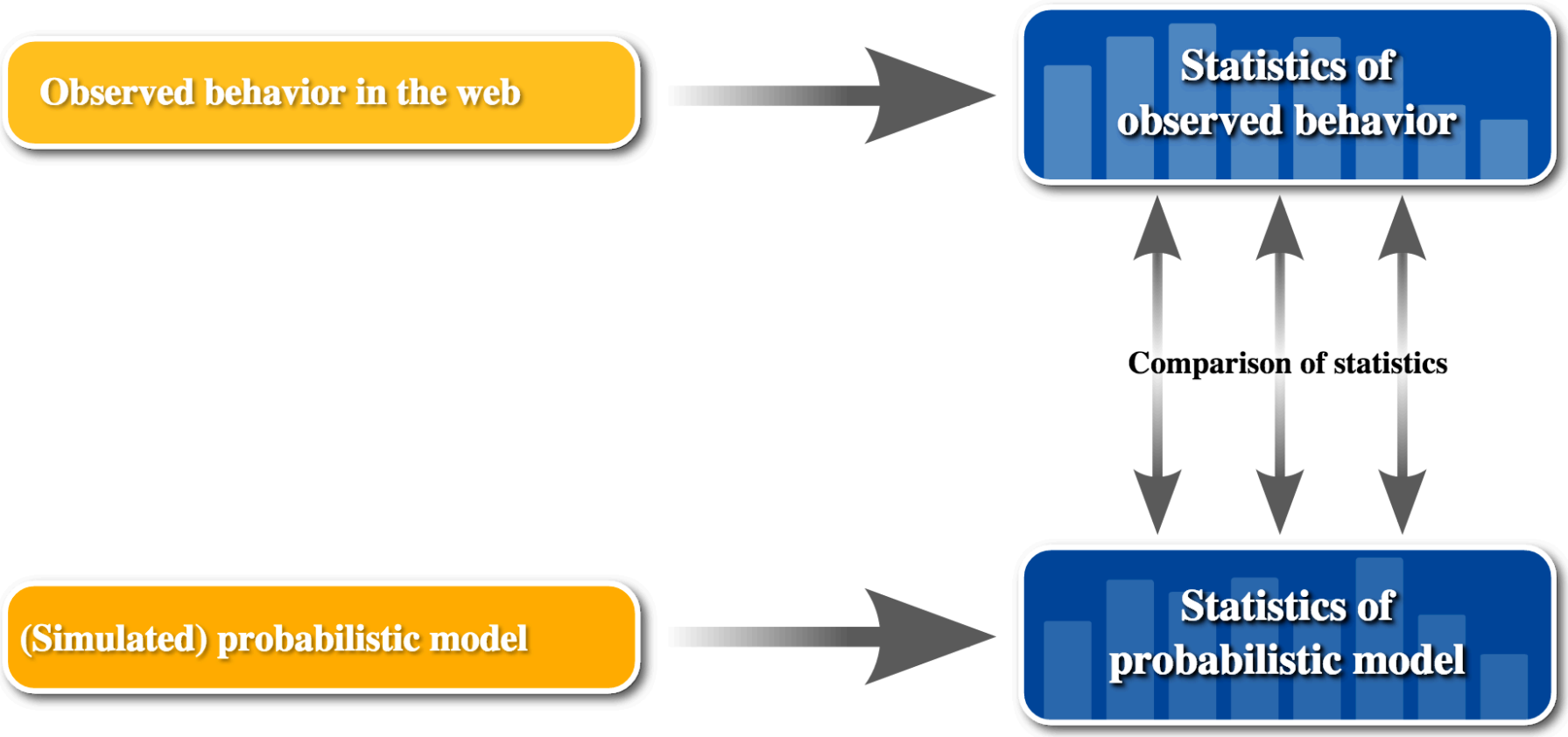
	‘ ‘	‘a‘	‘c‘	‘b‘	‘e‘	...	‘z‘
$c(x)$	30‘298	38‘090	506	1960	65‘818	...	7‘432
$P(x)$	0.138	0.173	0.002	0.009	0.299	...	0.034

- Make probabilities $P(x) = \frac{c(x)}{\sum_y c(y)}$
- Distribution could be a Zipf distribution making the Zipf parameter our one model parameter!

Run the model and calculate statistics again



Finally compare the statistics of both models



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: