



**Lesson5:**  
**Generative Models for Text on the Web**  
**Unit2:**  
**Sample values from a probability  
distribution**

Rene Pickhardt

Introduction to Web Science Part 2  
Emerging Web Properties

**WeST**   
People and Knowledge Networks

## Completing this unit you should

- Understand how to sample values from an arbitrary probability distribution
- Have seen yet another application of the cumulative distribution function
- Understand that sampling from a distribution is just a coordinate transformation of the uniform distribution

## How to generate words with this Model

- Switch to cumulative probabilities

(Simulated) probabilistic model



	' '	'a'	'c'	'b'	'e'	...	'z'
$P(x)$	0.138	0.173	0.002	0.009	0.299	...	0.034
$S(x)$	0.138					...	

## How to generate words with this Model

- Switch to cumulative probabilities

(Simulated) probabilistic model



	' '	'a'	'c'	'b'	'e'	...	'z'
$P(x)$	0.138	0.173	0.002	0.009	0.299	...	0.034
$S(x)$	0.138	0.311				...	

## How to generate words with this Model

- Switch to cumulative probabilities

(Simulated) probabilistic model



	' '	'a'	'c'	'b'	'e'	...	'z'
$P(x)$	0.138	0.173	0.002	0.009	0.299	...	0.034
$S(x)$	0.138	0.311	0.313			...	

## How to generate words with this Model

- Switch to cumulative probabilities

(Simulated) probabilistic model



	' '	'a'	'c'	'b'	'e'	...	'z'
$P(x)$	0.138	0.173	0.002	0.009	0.299	...	0.034
$S(x)$	0.138	0.311	0.313	0.322		...	

## How to generate words with this Model

- Switch to cumulative probabilities

(Simulated) probabilistic model



	' '	'a'	'c'	'b'	'e'	...	'z'
$P(x)$	0.138	0.173	0.002	0.009	0.299	...	0.034
$S(x)$	0.138	0.311	0.313	0.322	0.621	...	

## How to generate words with this Model

- Switch to cumulative probabilities

(Simulated) probabilistic model



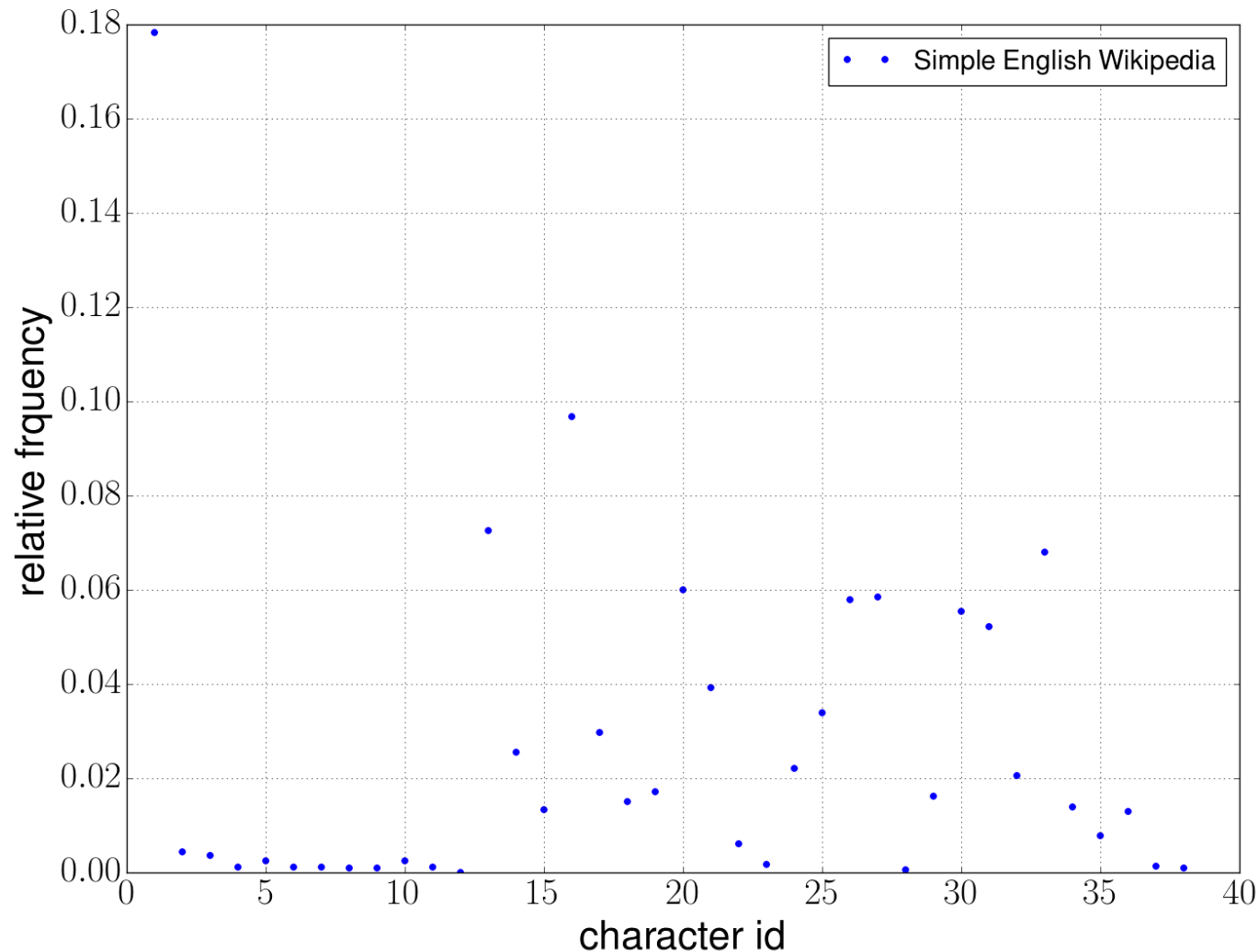
	' '	'a'	'c'	'b'	'e'	...	'z'
$P(x)$	0.138	0.173	0.002	0.009	0.299	...	0.034
$S(x)$	0.138	0.311	0.313	0.322	0.621	...	1

- Now draw a uniform random number between 0 and 1 e.g.:  $r = \text{random.random}()$
- Find  $x$  so that  $S(x)$  is the lowest value bigger than  $r$



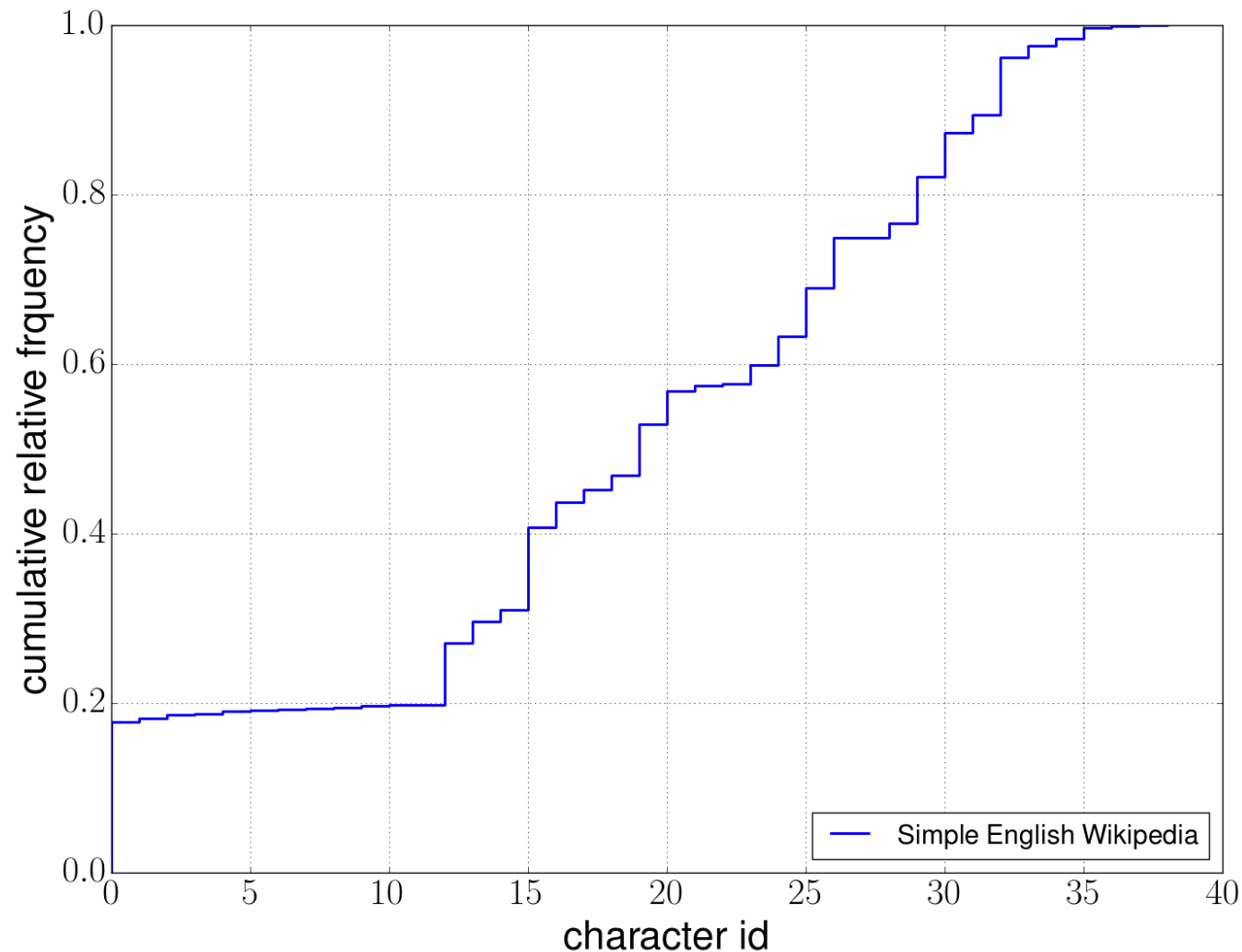
# Shape of the character distribution [a-z0-9]

relative frequencies of characters



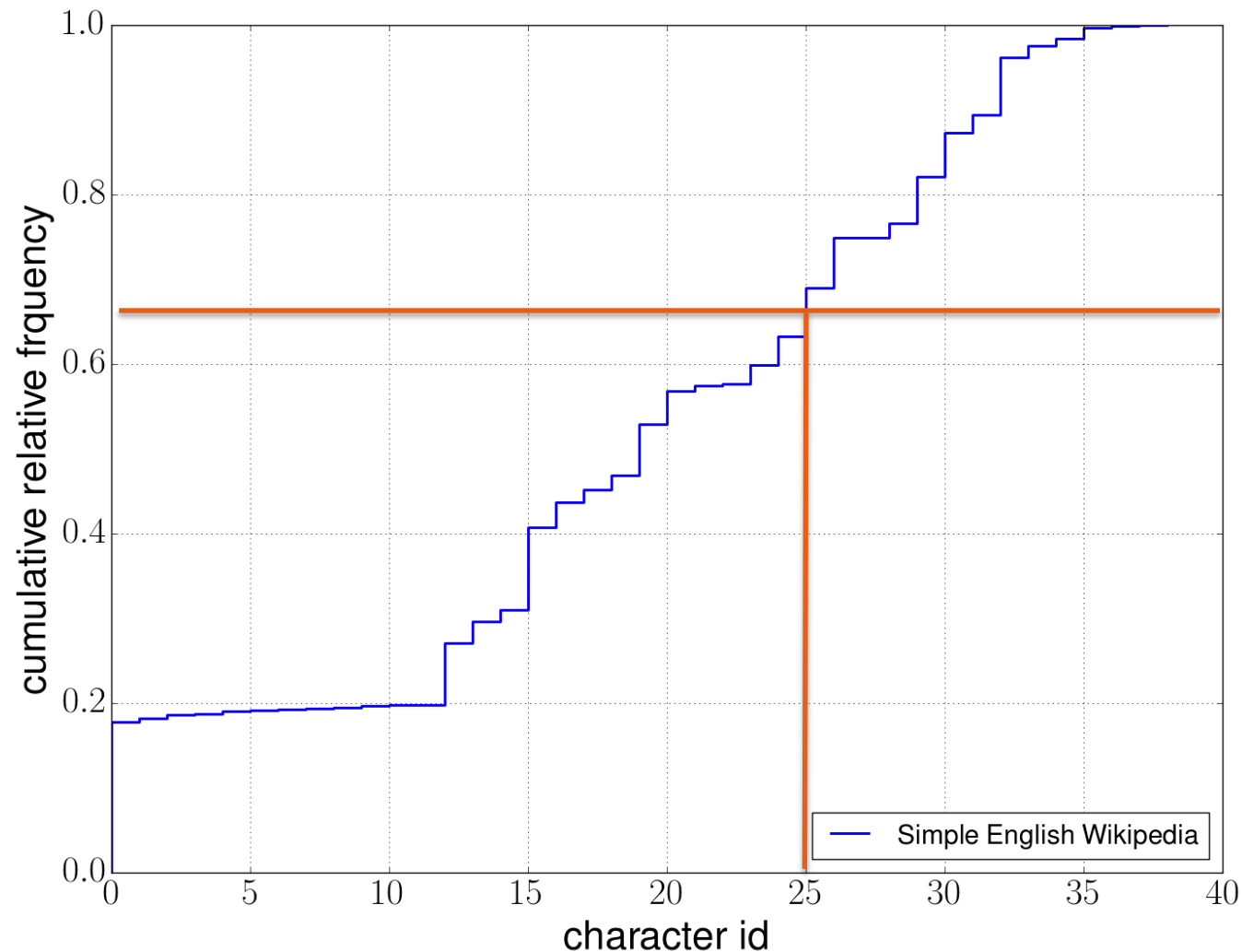
# Rolling a die on the CDF means...

CDF of relative character frequencies



# Rolling a die on the CDF means... char 25

CDF of relative character frequencies





# Thank you for your attention!



Contact:

Rene Pickhardt  
Institute for Web Science and Technologies  
Universität Koblenz-Landau  
[rpickhardt@uni-koblenz.de](mailto:rpickhardt@uni-koblenz.de)

**WeST**   
People and Knowledge Networks

# Copyright:

- **This Slide deck is licensed under creative commons 3.0. share alike attribution license. It was created by Rene Pickhardt. You can use share and modify this slide deck as long as you attribute the author and keep the same license.**
- By ArchonMagnus (Own work) [CC BY-SA 4.0 (<http://creativecommons.org/licenses/by-sa/4.0>)], via Wikimedia Commons