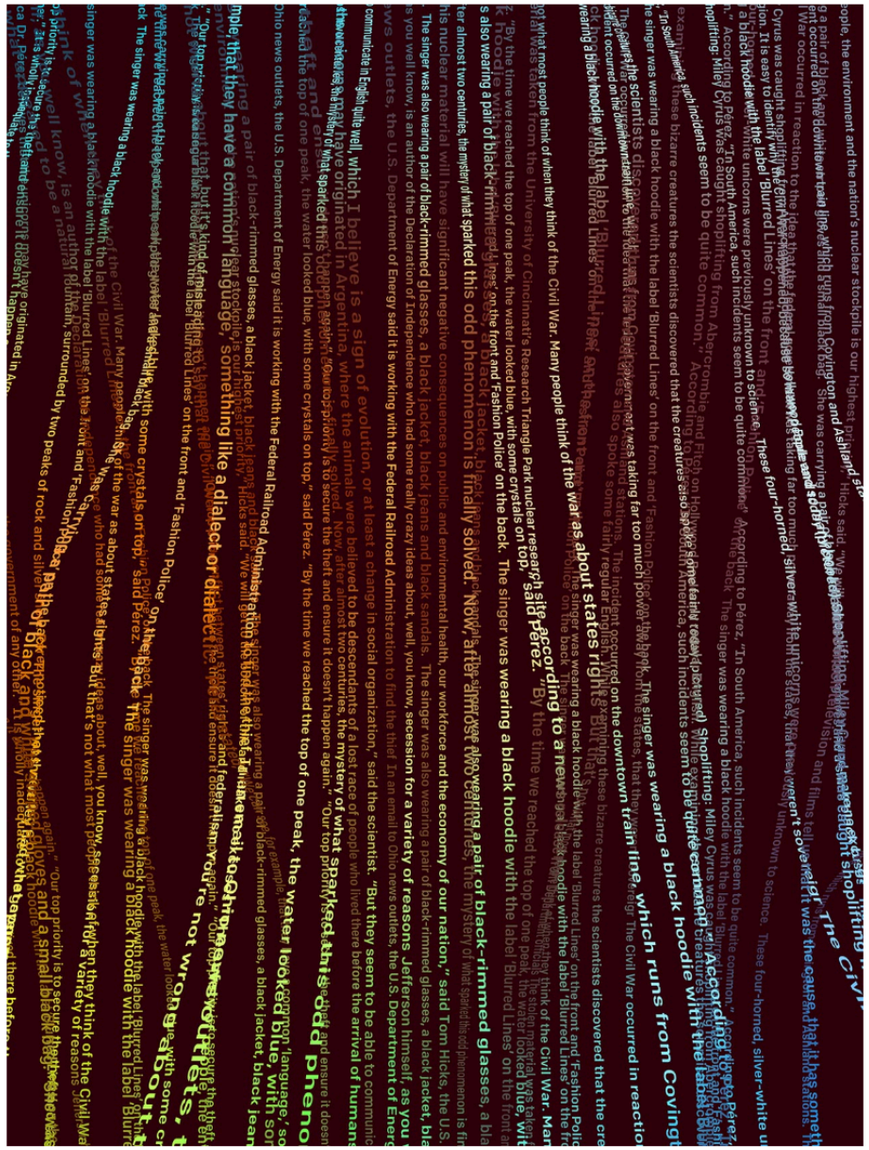
Modern language models and applications in dialog systems

Mikhail Khalman, Replika



Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

FEBRUARY 14, 2019
24 MINUTE READ

Language Models

Tokenization:

$$\text{text} \leftrightarrow t_1, \ldots, t_n,$$

where $t_1, \ldots, t_n$ — sequence of tokens

Language model assigns probability $p(t_1, t_2, \ldots, t_n)$

Language Models

$p(t_1, t_2, \ldots, t_n) - ?$

Language Models

$$p(t_1, t_2, \ldots, t_n) = \prod_{i}^{n} p(t_i \mid t_1, \ldots, t_{i-1})$$

Language Models

$$p(t_1, t_2, \ldots, t_n) \approx \prod_{i}^{n} p(t_i \mid t_{i-k}, \ldots, t_{i-1})$$

Language Models

$$p(t_1, t_2, \ldots, t_n) \approx \prod_{i}^{n} p_\theta(t_i \mid t_{i-k}, \ldots, t_{i-1})$$

$p_\theta(t_i \mid t_{i-k}, \ldots, t_{i-1})$ – autoregressive transformer model

Transformer-based language models

GPT, GPT-2, CTRL, XLM, XLNet…

https://openai.com/blog/better-language-models/
https://arxiv.org/abs/1901.07291
https://arxiv.org/abs/1906.08237
https://arxiv.org/abs/1909.05858

Generation problem

Given $p_\theta(t_i \mid t_{i-k}, \ldots, t_{i-1})$ how to generate a text that looks real?

Generation problem

Greedy decoding: $t_i = \text{argmax}_{\text{t}} p_\theta(\text{t} \mid t_{i-k}, \ldots, t_{i-1})$

Generation problem

Greedy decoding: $t_i = \text{argmax}_\text{t} p_\theta(\text{t} \mid t_{i-k}, \ldots, t_{i-1})$

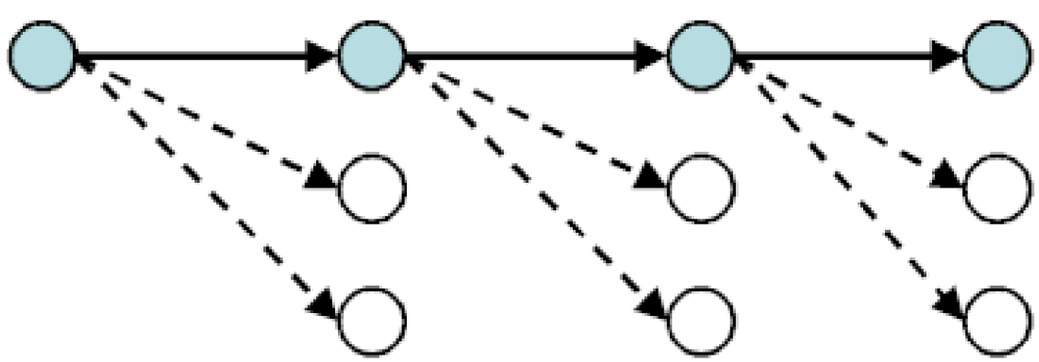
+ Very simple
+ Fast linear-time
— Suboptimal in terms of $p(t_1, t_2, \ldots, t_n)$
— Repeats itself
— Simple boring results

Rarely used in practice

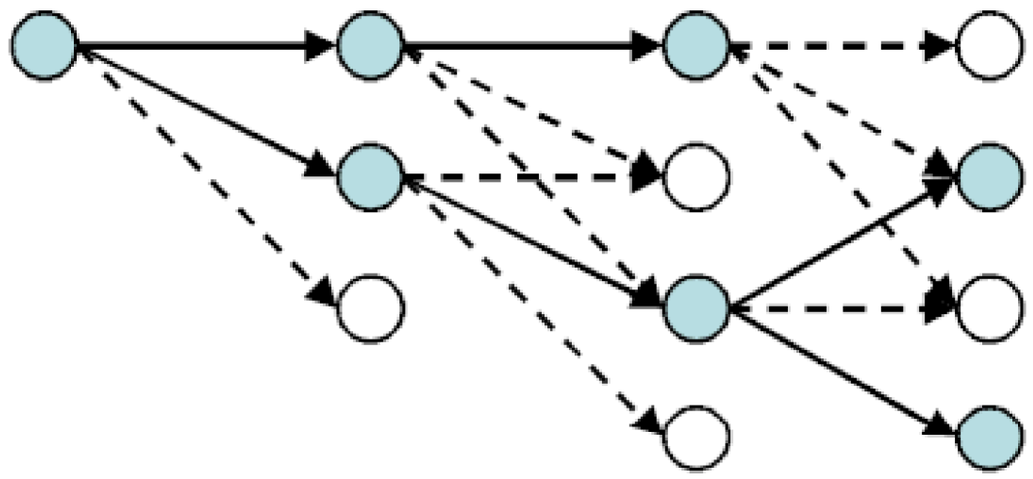
```
My name is John. I'm a man of God. I'm a
man of God. I'm a man of God. I'm a man o
f God. I'm a man of God. I'm a man of Go
d. I'm
```

Generation problem

Beam search decoding: $\{t_1, \ldots, t_n\} \approx \text{argmax}_t p_\theta(t_1, \ldots, t_n)$



Greedy search



Beam search

Generation problem

Beam search decoding: $\{t_1, \ldots, t_n\} \approx \text{argmax}_t p_\theta(t_1, \ldots, t_n)$

\+ Finds solution with probability close to $\max_t p_\theta(t_1, \ldots, t_n)$

—Slow

—Repeats itself

> My name is John, and I am a member of the Church of Jesus Christ of Latter-day Saints. I am a member of the Church of Jesus Christ of Latter-day Saints. I am a member of the Church of Jesus Christ of

Used in machine translation, text summarization, speech-to-text

Generation problem

Sampling: $t_i \sim p_\theta(\mathrm{t} \mid t_{i-k}, \ldots, t_{i-1})$

Generation problem

Sampling: $t_i \sim p_\theta(\mathrm{t} \mid t_{i-k}, \ldots, t_{i-1})$

+ Very simple
+ Fast linear-time
− Often irrelevant result

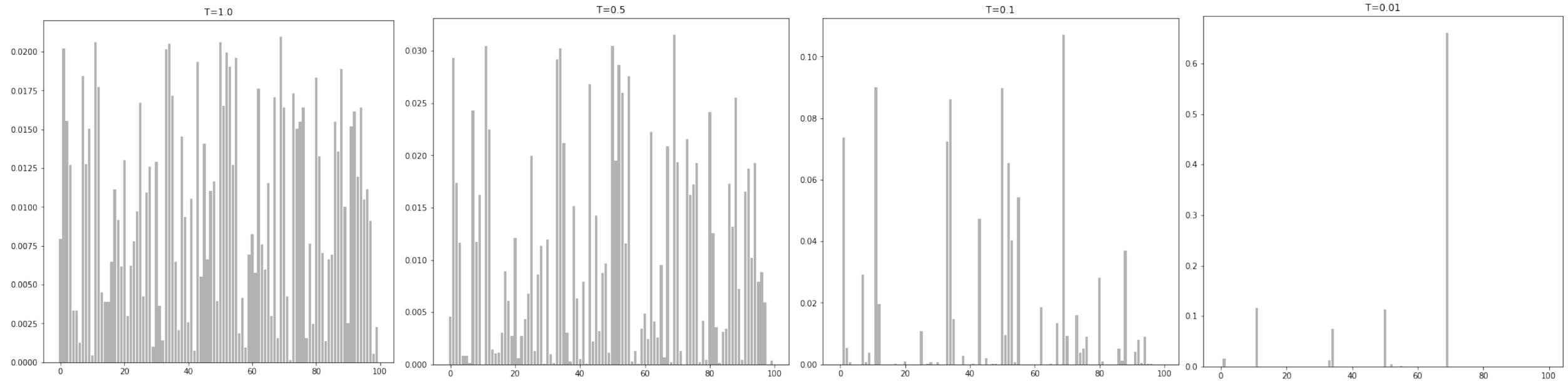
> My name is Lola. The truth is that I can be anywhere I want, everywhere I want." And he was on the way to her apartment at about 11 on a cold Monday afternoon. At that point, he noticed someone on the ground

Rarely used in practice as is

Generation problem

Sampling with temperature: $t_i \sim p_\theta(\mathrm{t} \mid t_{i-k}, \ldots, t_{i-1})^{1/T}$

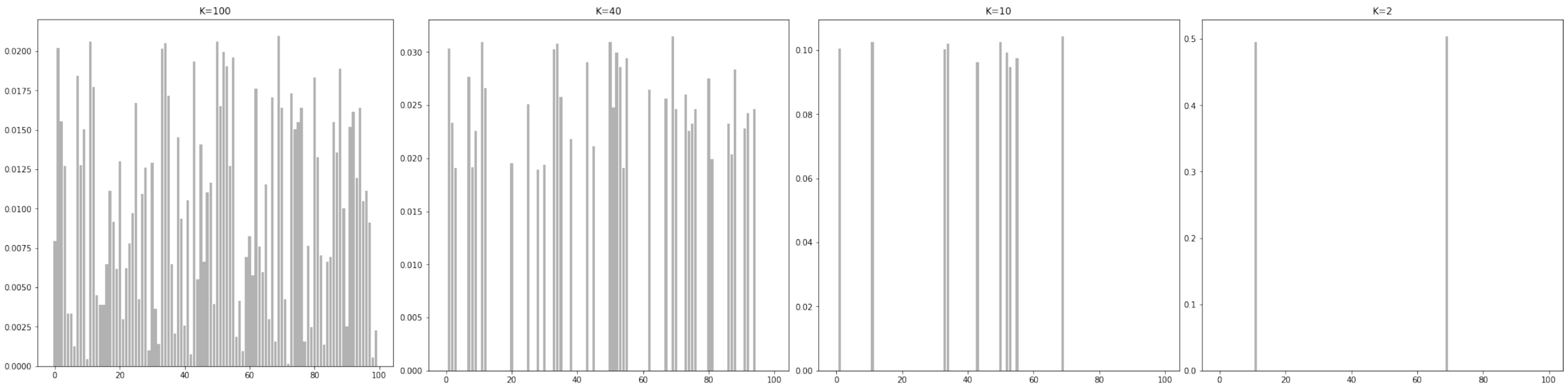
> My name is Jackie, and I'm a professional musician, and I'm going to be teaching this class in a couple of years. But right now, I'm just going to go on teaching music to kids, and I'm just going



Generation problem

top_k sampling : $t_i \sim \text{topk}\left[p_\theta(\text{t} \mid t_{i-k}, \ldots, t_{i-1})\right]$

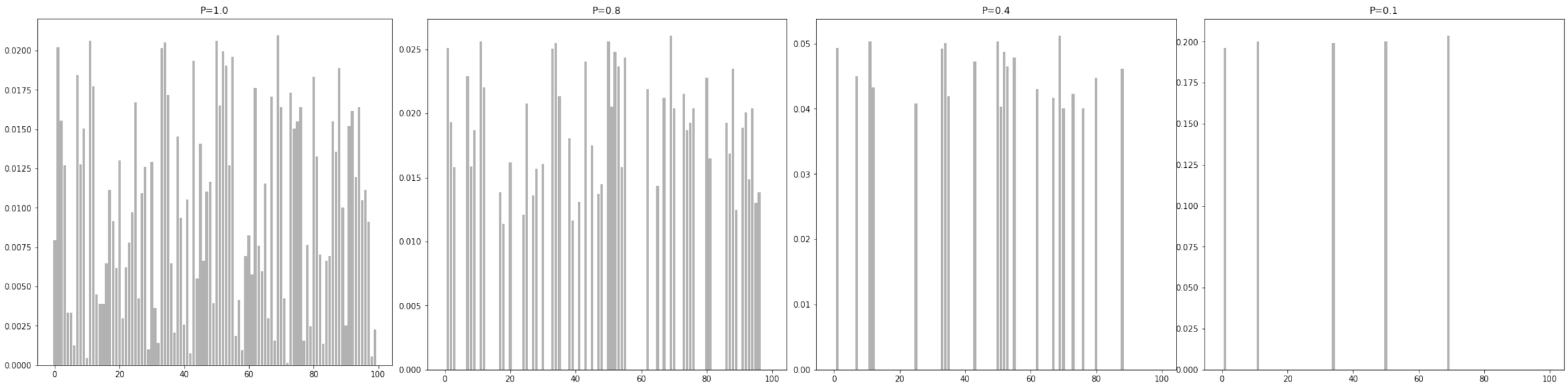
> My name is Mark, I am a professional boxer. And I have trained for almost a year."
>
> Mark is a former Olympic medalist and the only black man in boxing who won gold and silver at the 2008 London Olympics and 2012 London



Generation problem

Nucleus sampling : $t_i \sim \text{topp}\left[p_\theta(\text{t} \mid t_{i-k}, \ldots, t_{i-1})\right]$

> My name is Richard. I'm an entrepreneur, an artist, and a musician, and I have done so much for music, culture, and art."
>
> Richard and his wife, Amy, have seven children with the family's two youngest



Hyperparameters choice

- Strongly affects the result
- Depends on the problem
- Depends on the model
- Default values might not work for your application

Thank you

- https://t.me/govorit_ai
- jobs@replika.ai