Understanding LDA in Source Code Analysis

Dave Binkley, Daniel Heinz, Dawn Lawrie, Justin Overfelt

Loyola University Maryland

LDA

models a corpus of documents using probability distributions

- has two parameters
  - ϕ, a **word** by **topic** distribution
  - θ, a **topic** by **document** distribution

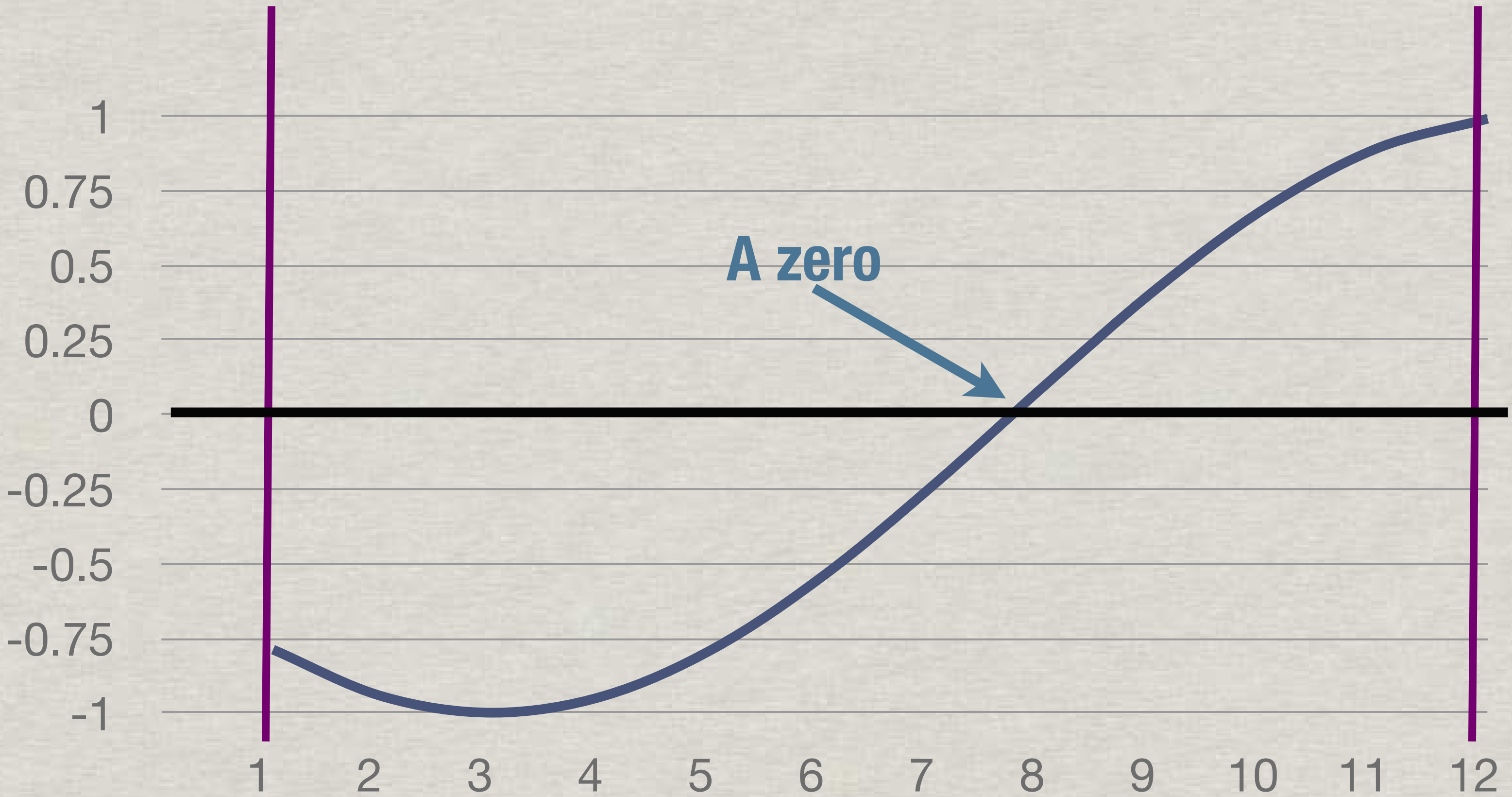
Backwards? Yes

- Two Key Points

  i. *Sampling* is not refinement

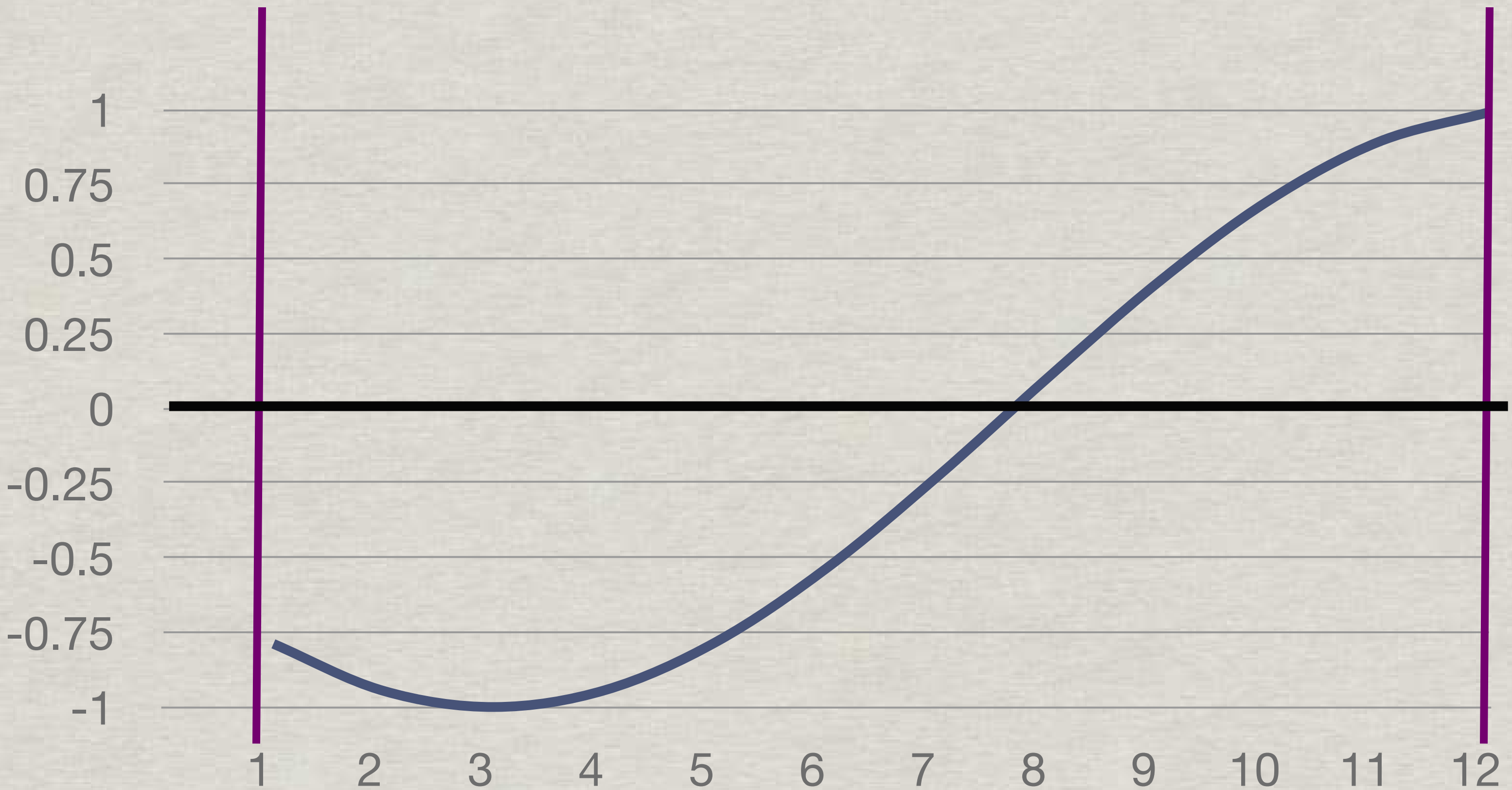
  ii. LDA comprehension starts with *parameter understanding*

Refinement



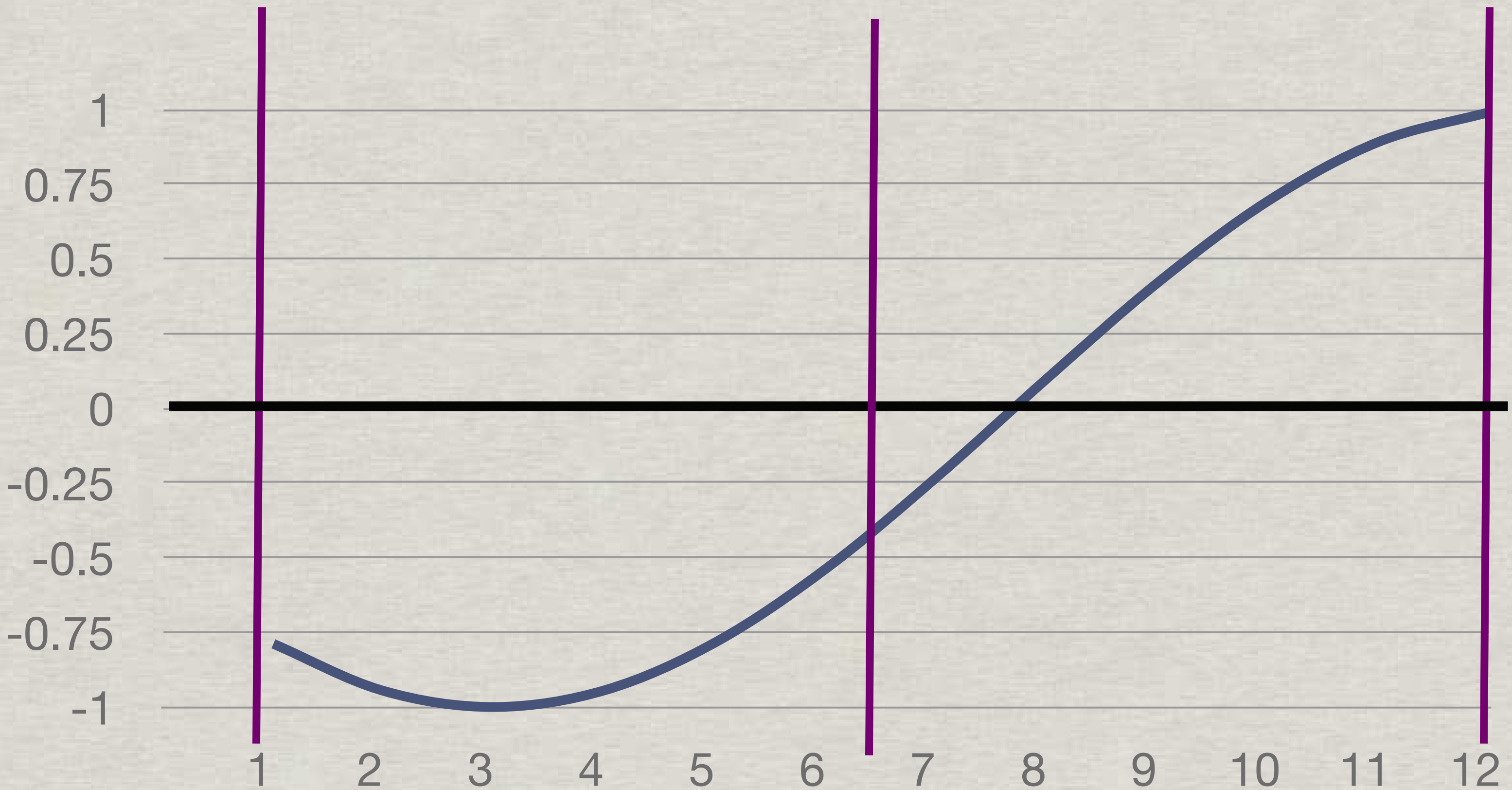
Initially we know a zero exists between 1 and 12

Iteratively check the middle



Initially we know a zero exists between 1 and 12

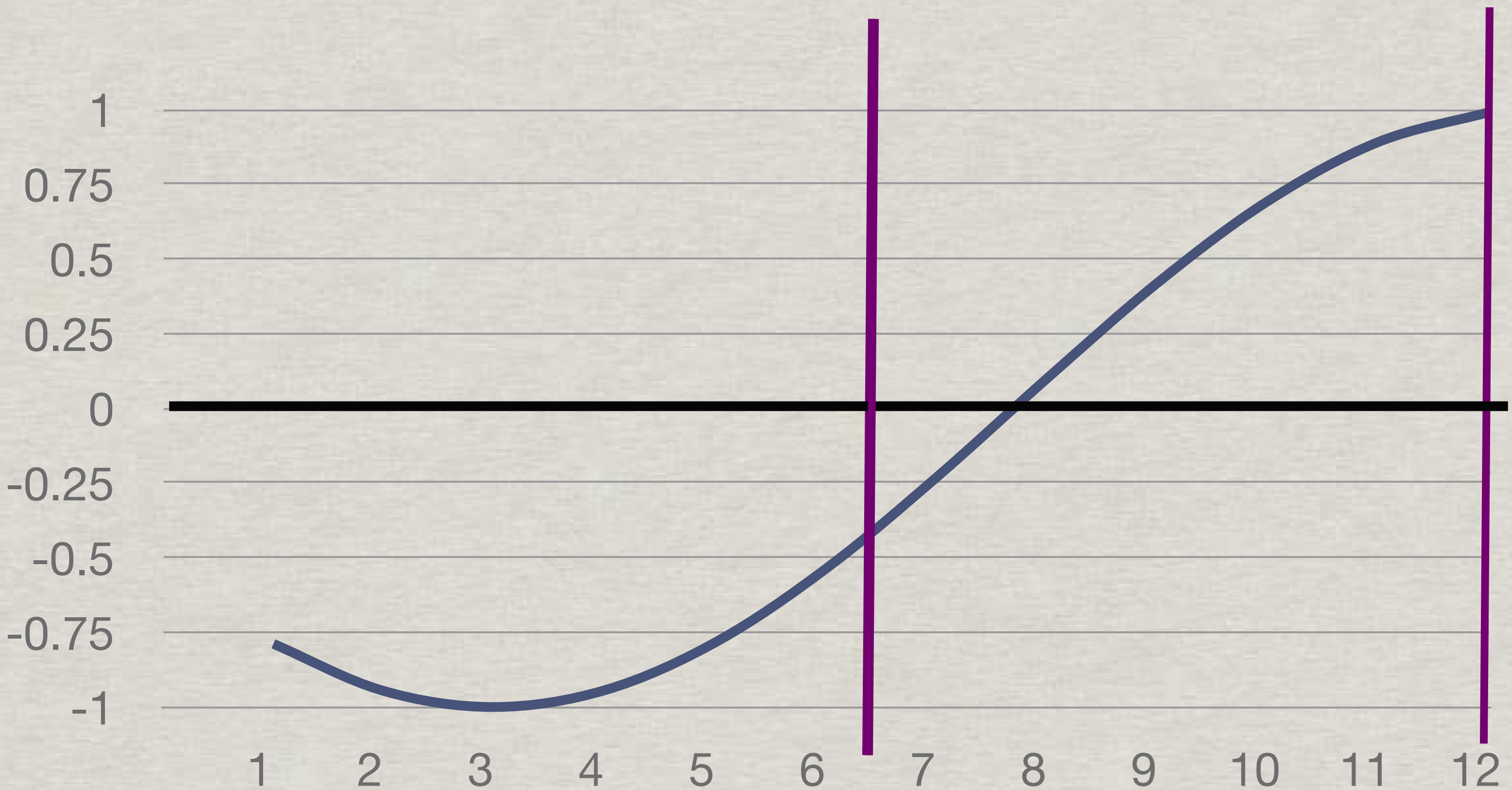
Iteratively check the middle



Initially we know a zero exists between 1 and 12

Move left side in

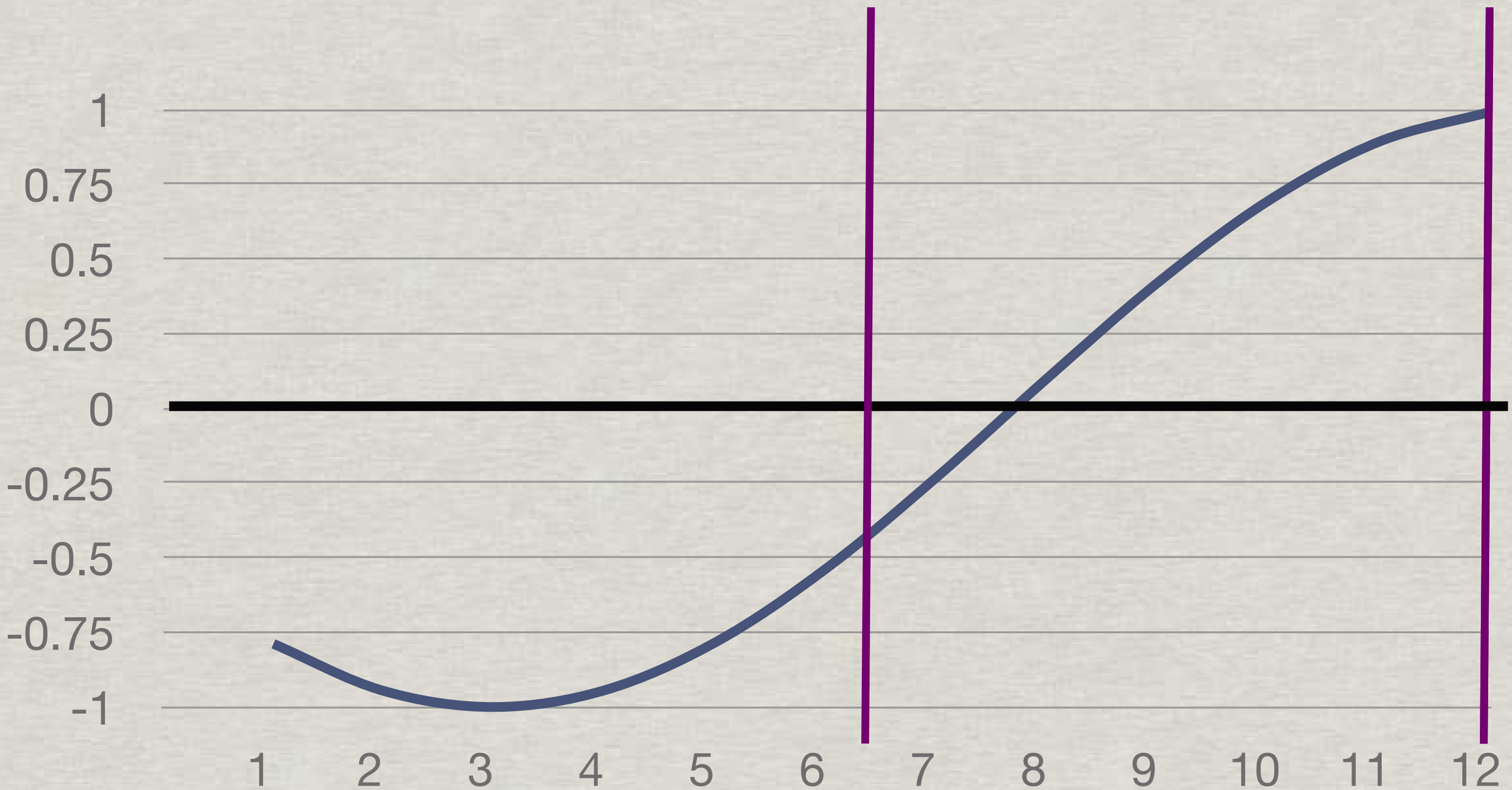
Move left side in



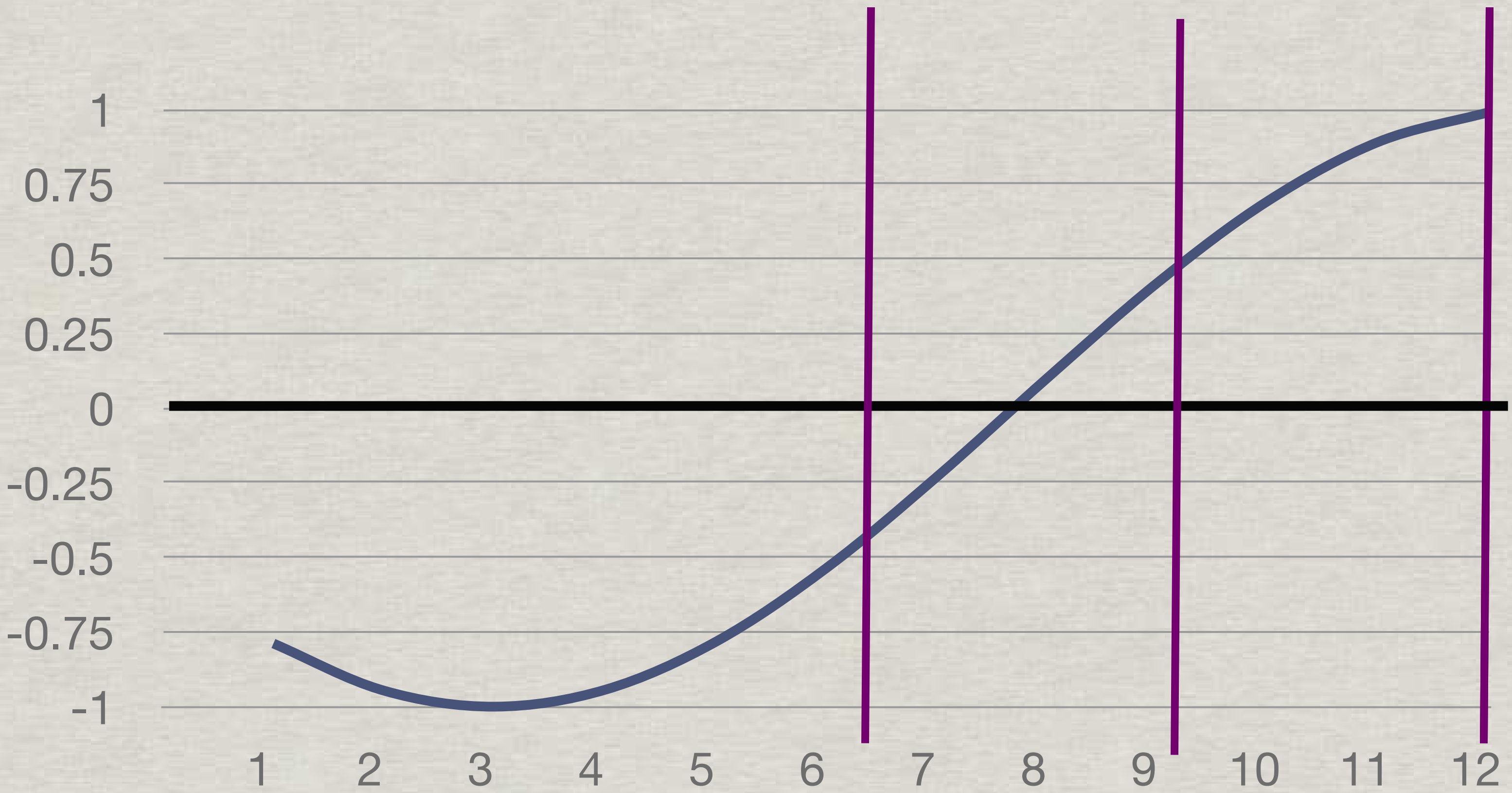
Move left side in

Now we know a zero exists between 6.5 and 12

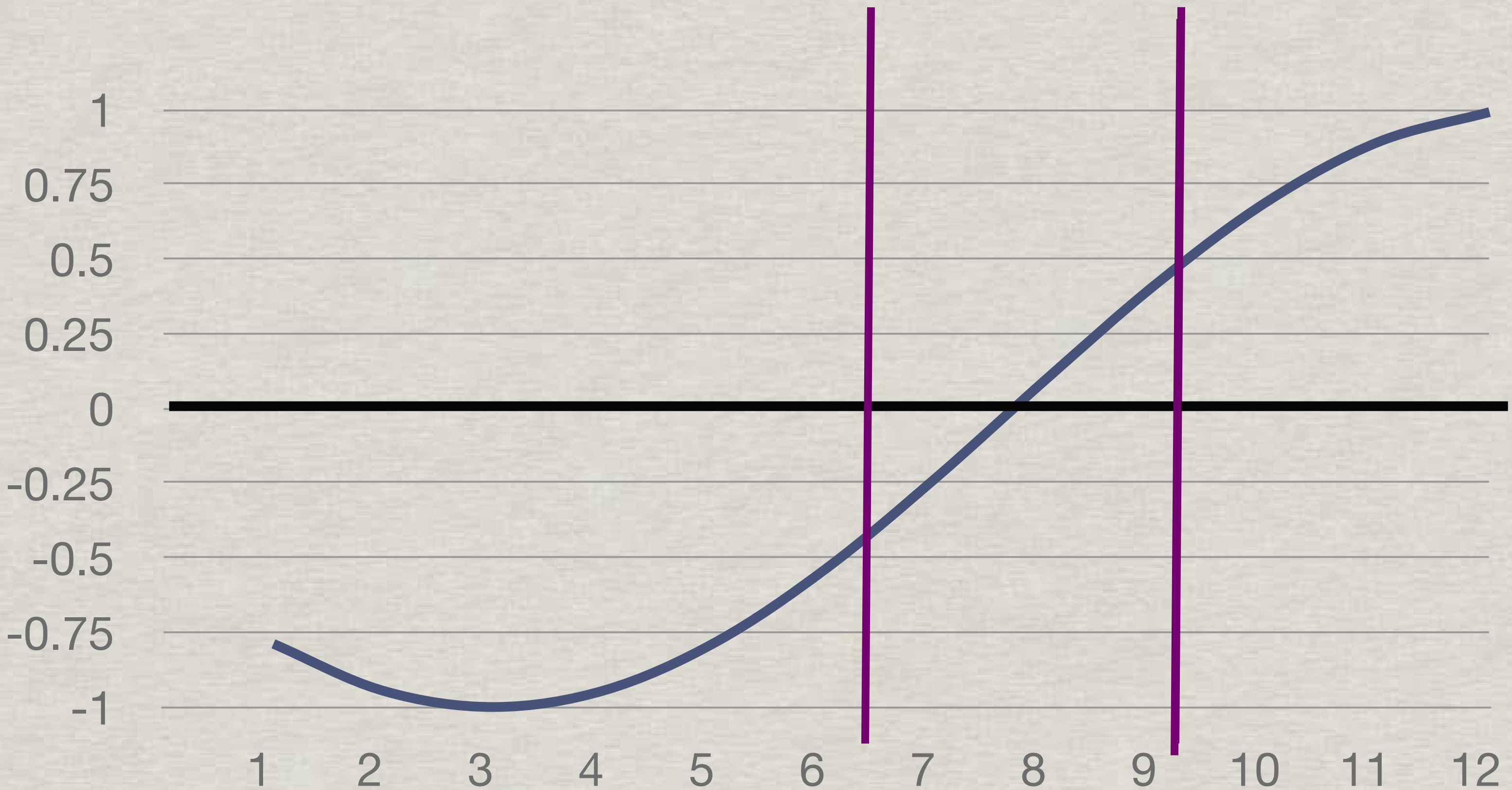
Repeat



Repeat

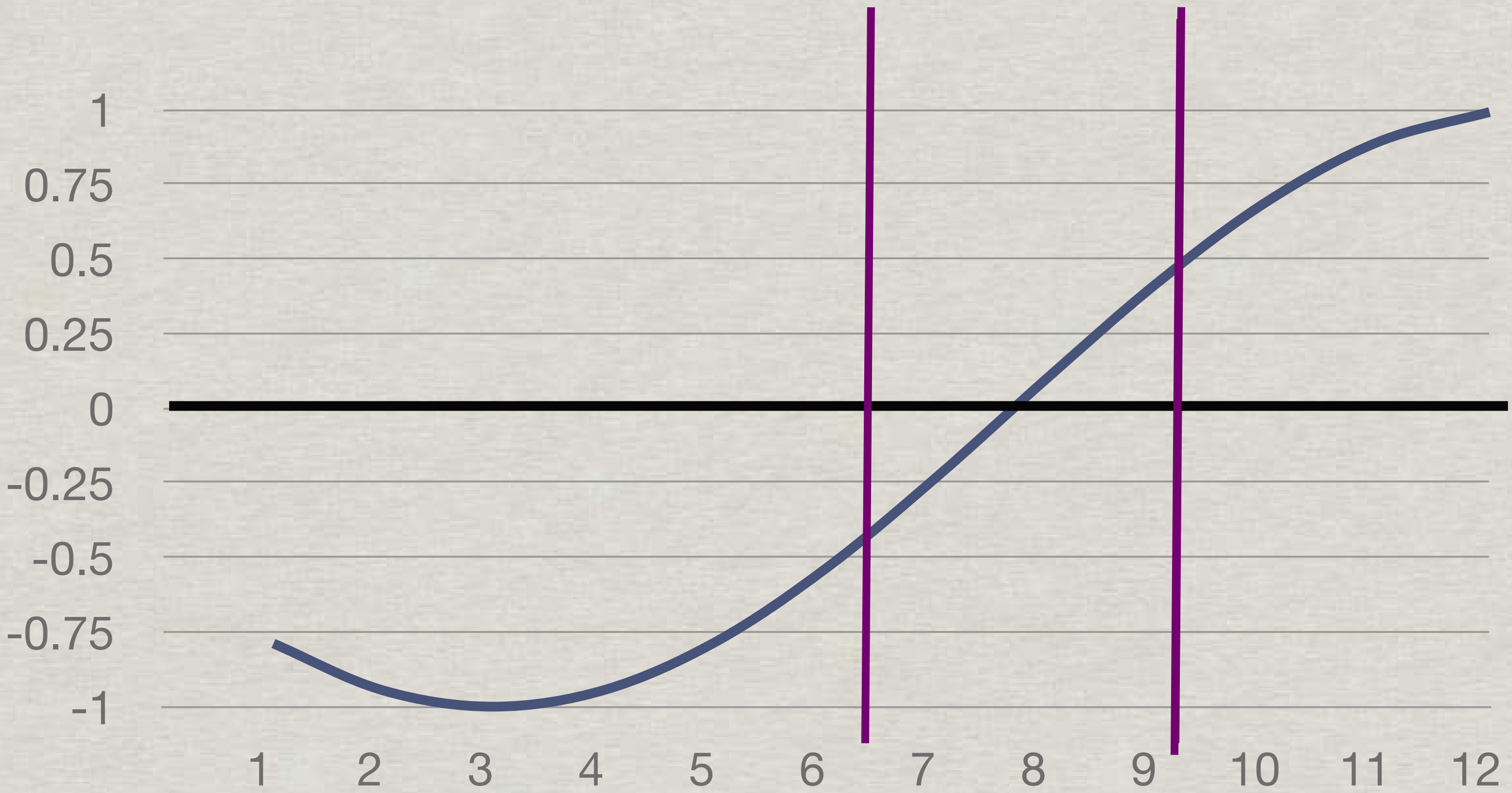


Repeat



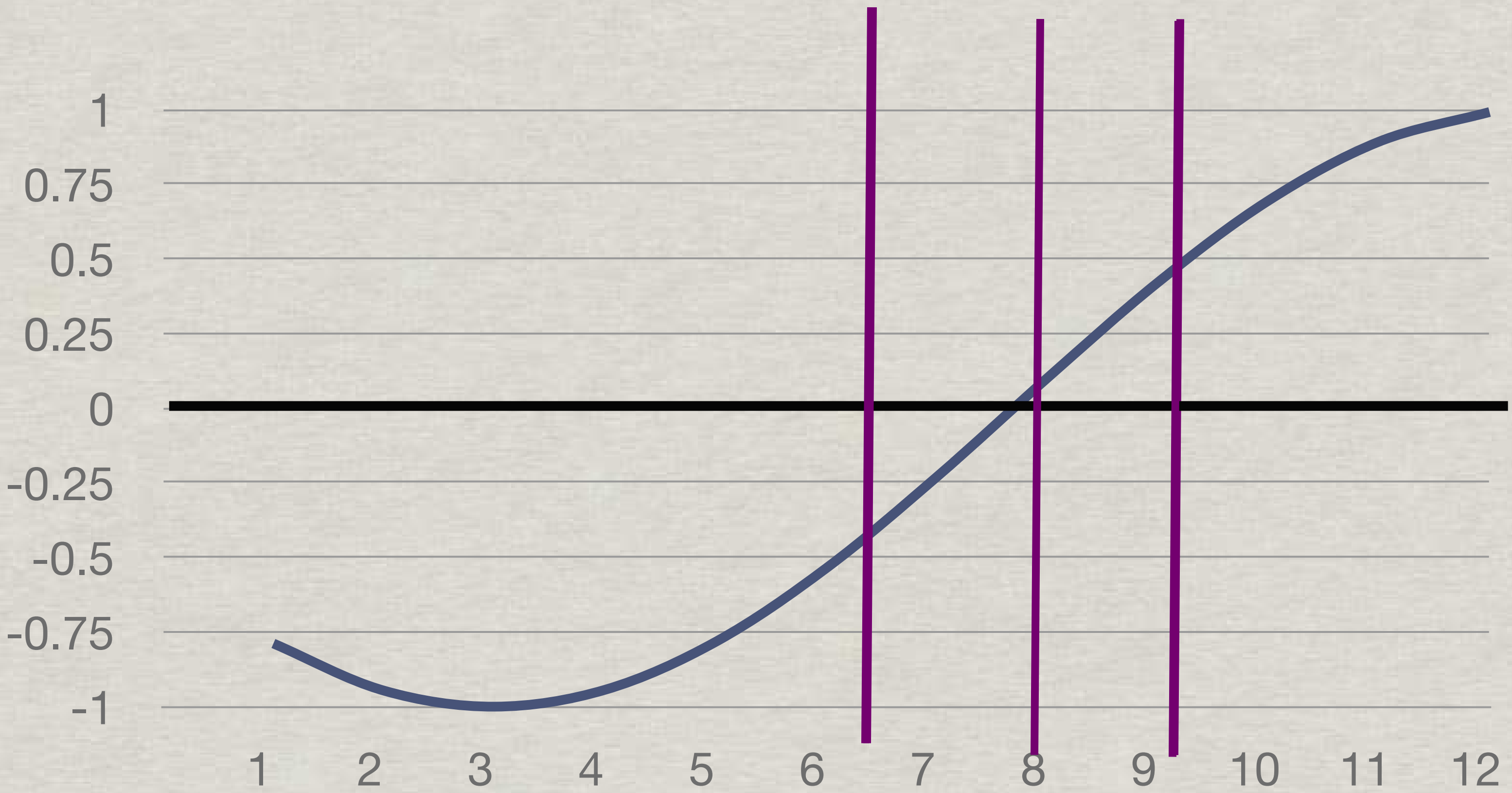
Repeat

Now we know a zero exists between 6.5 and 9.25

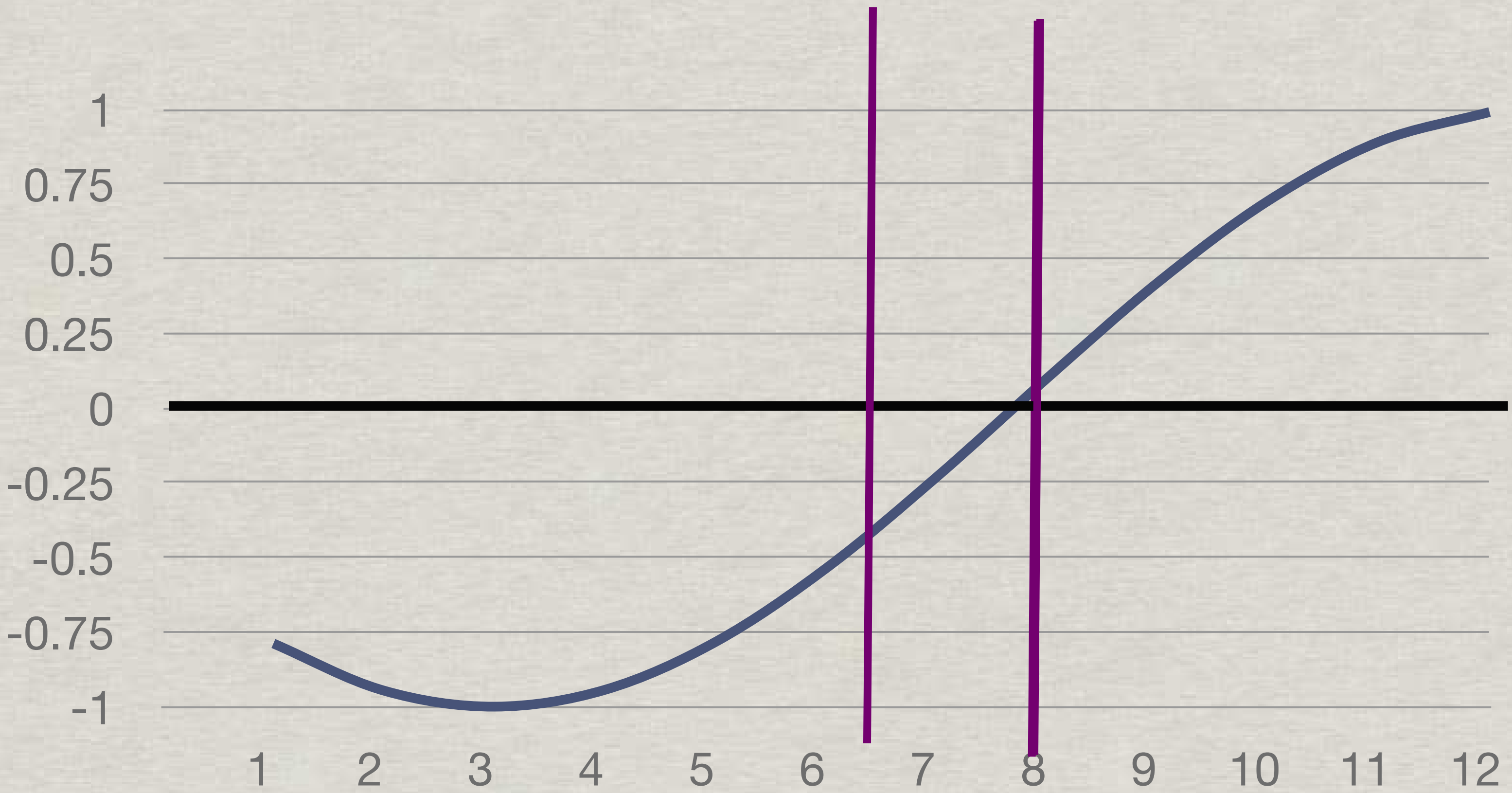


Refinement - Each Iteration Leads to a Better Approximation

Refinement - Each Iteration Leads to a Better Approximation



Refinement - Each Iteration Leads to a Better Approximation

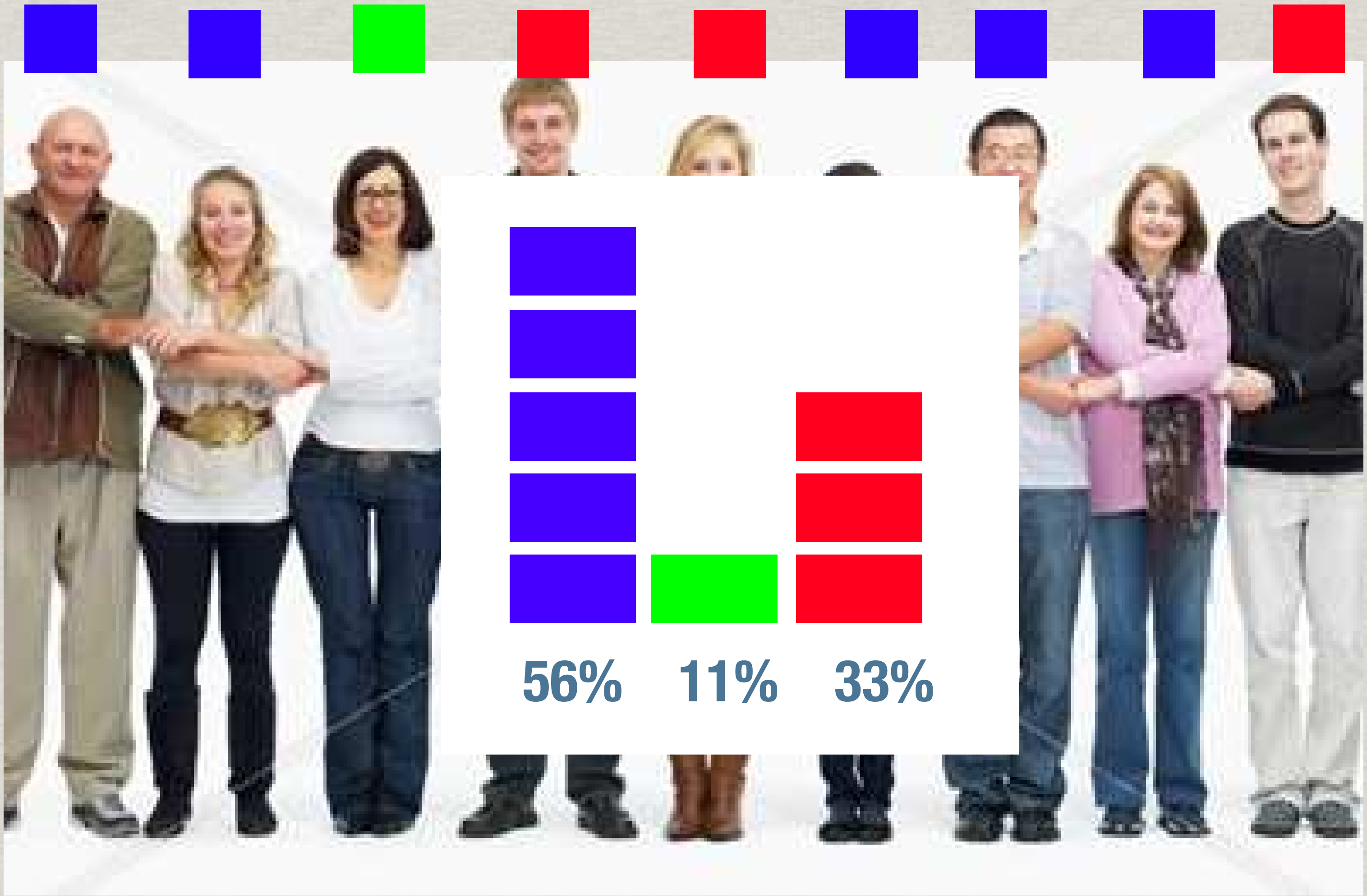


Sampling is **not** Refinement

What is the distribution of favorite colors?

Sampling is **not** Refinement

What is the distribution of favorite colors?



Sampling is **not** Refinement

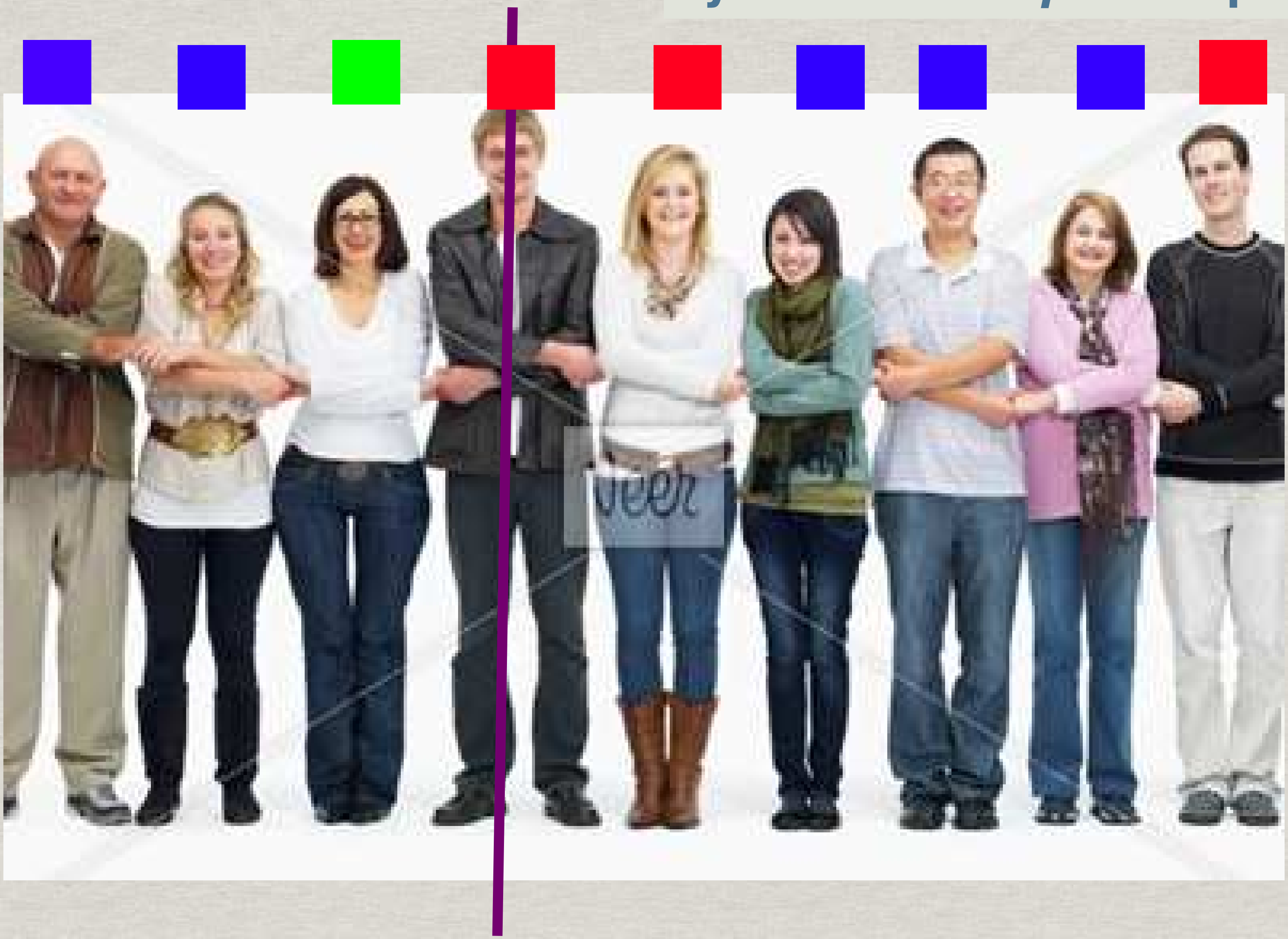
What is the distribution of favorite colors?

Sampling is **not** Refinement

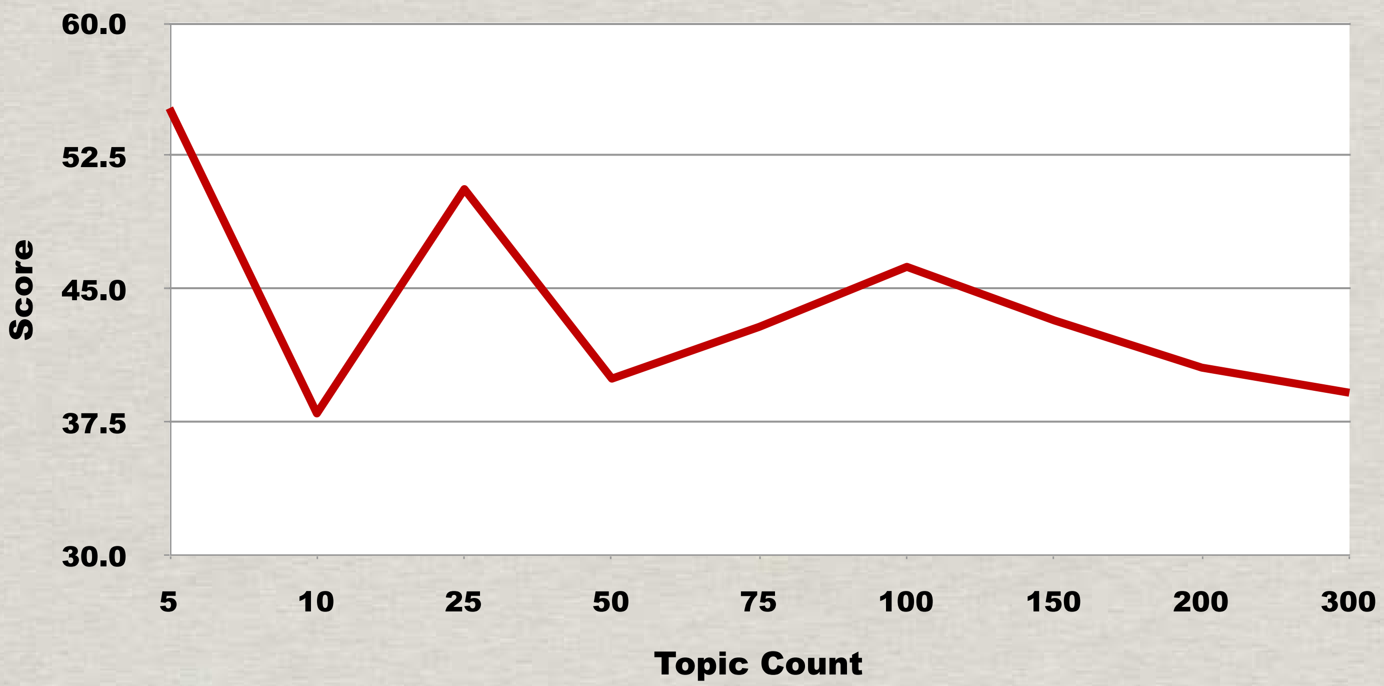
What is the distribution of favorite colors?

No *one* is *the* answer

you need *multiple* samples

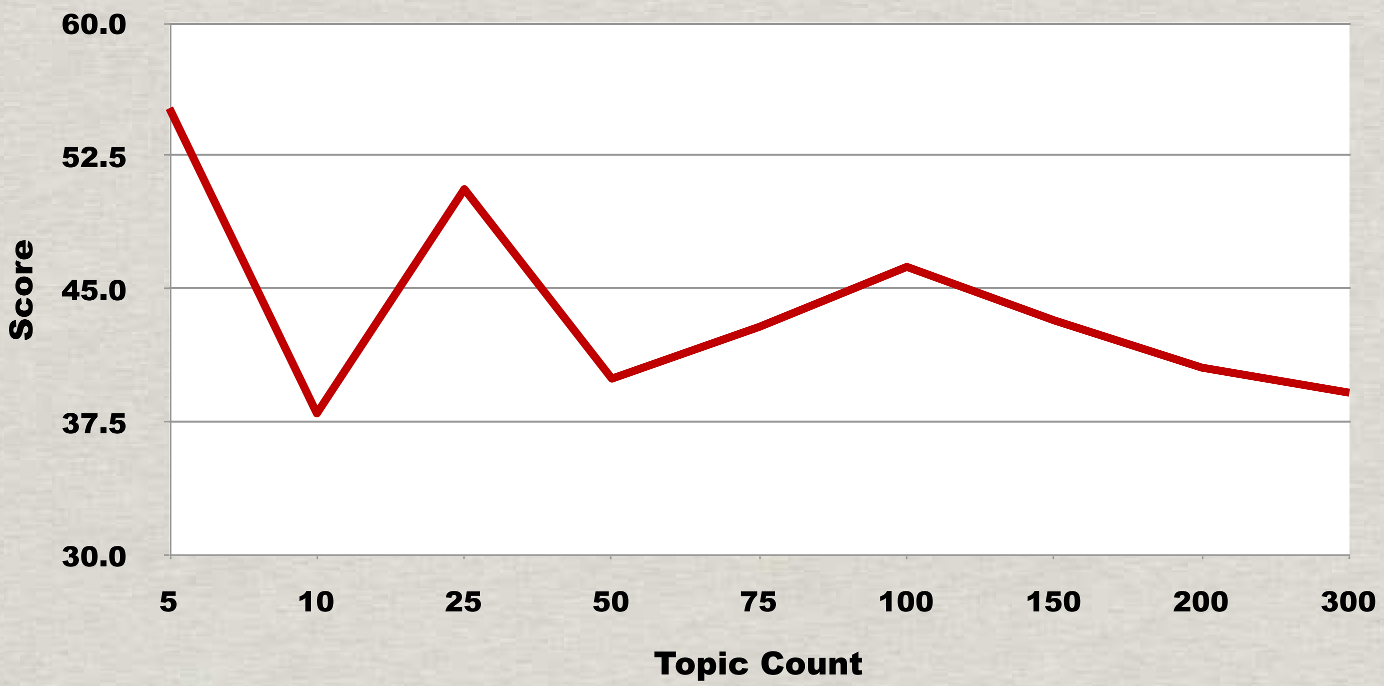


What about LDA?

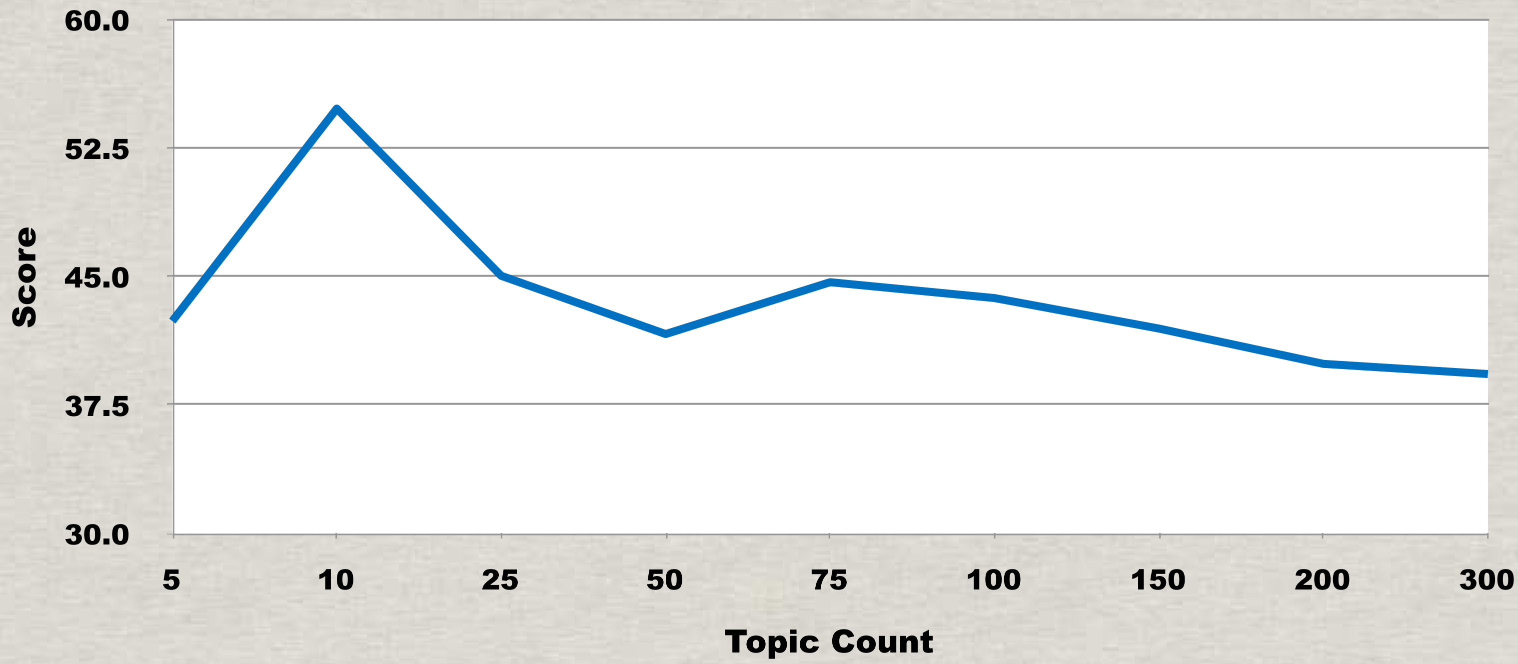


What about LDA?

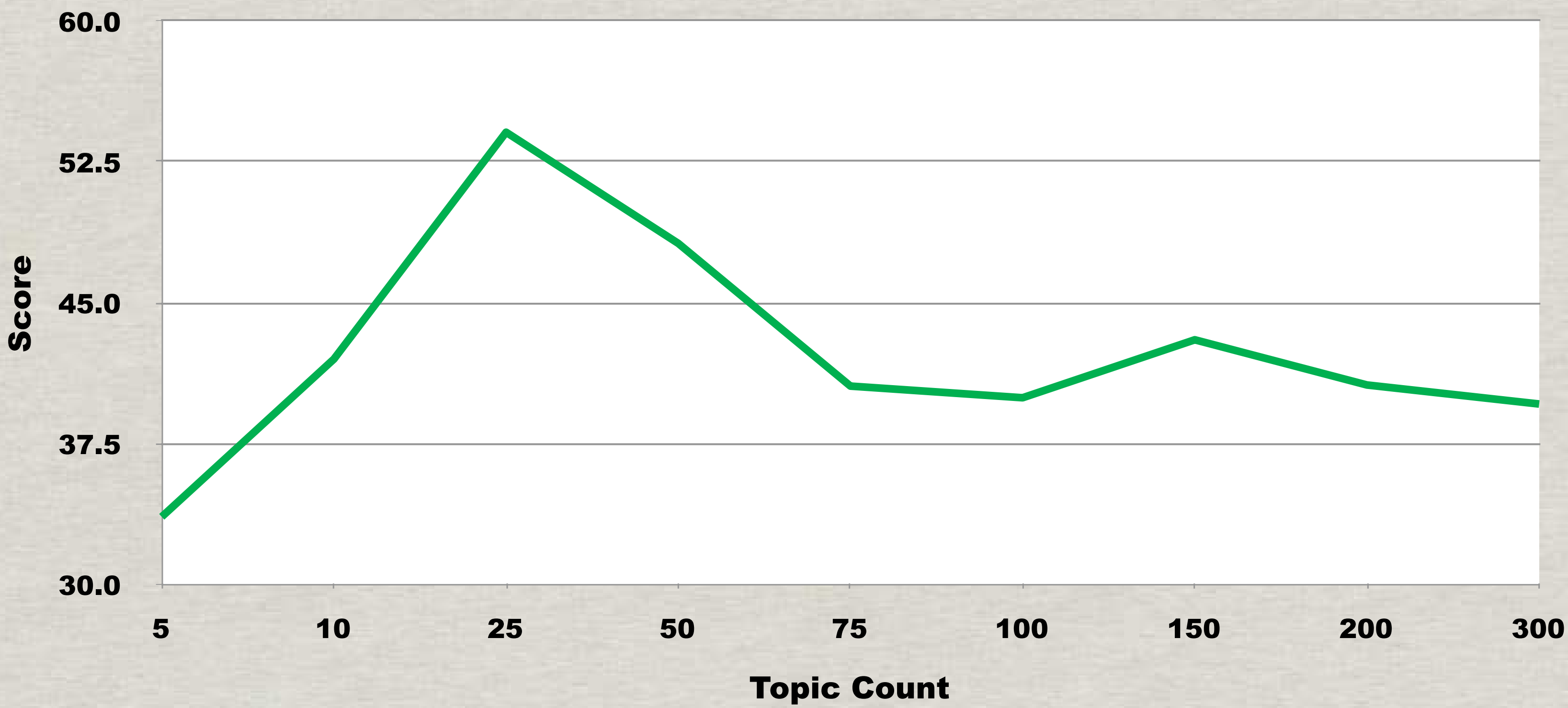
Five Topics is Best



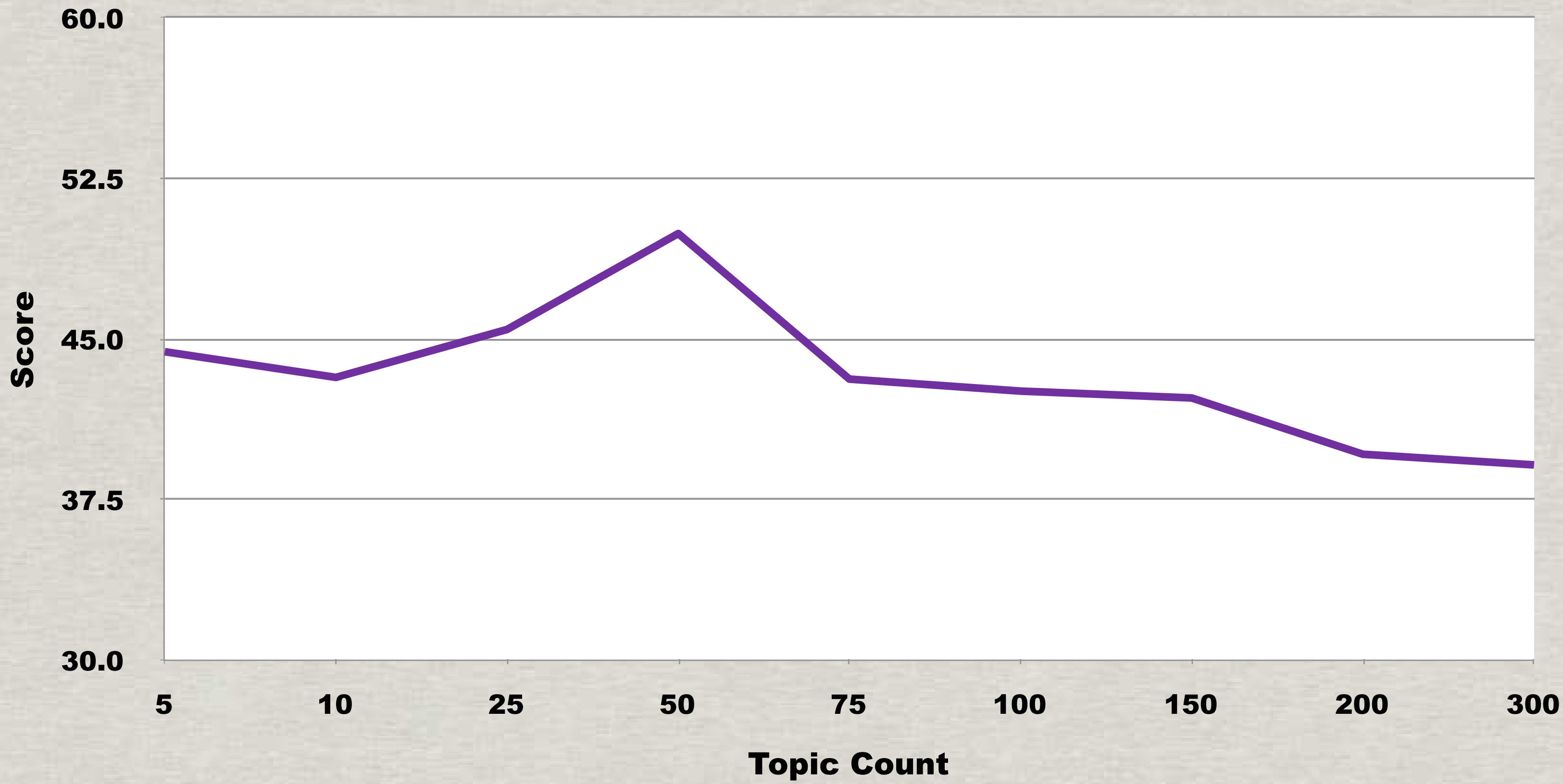
Ten Topics is Best



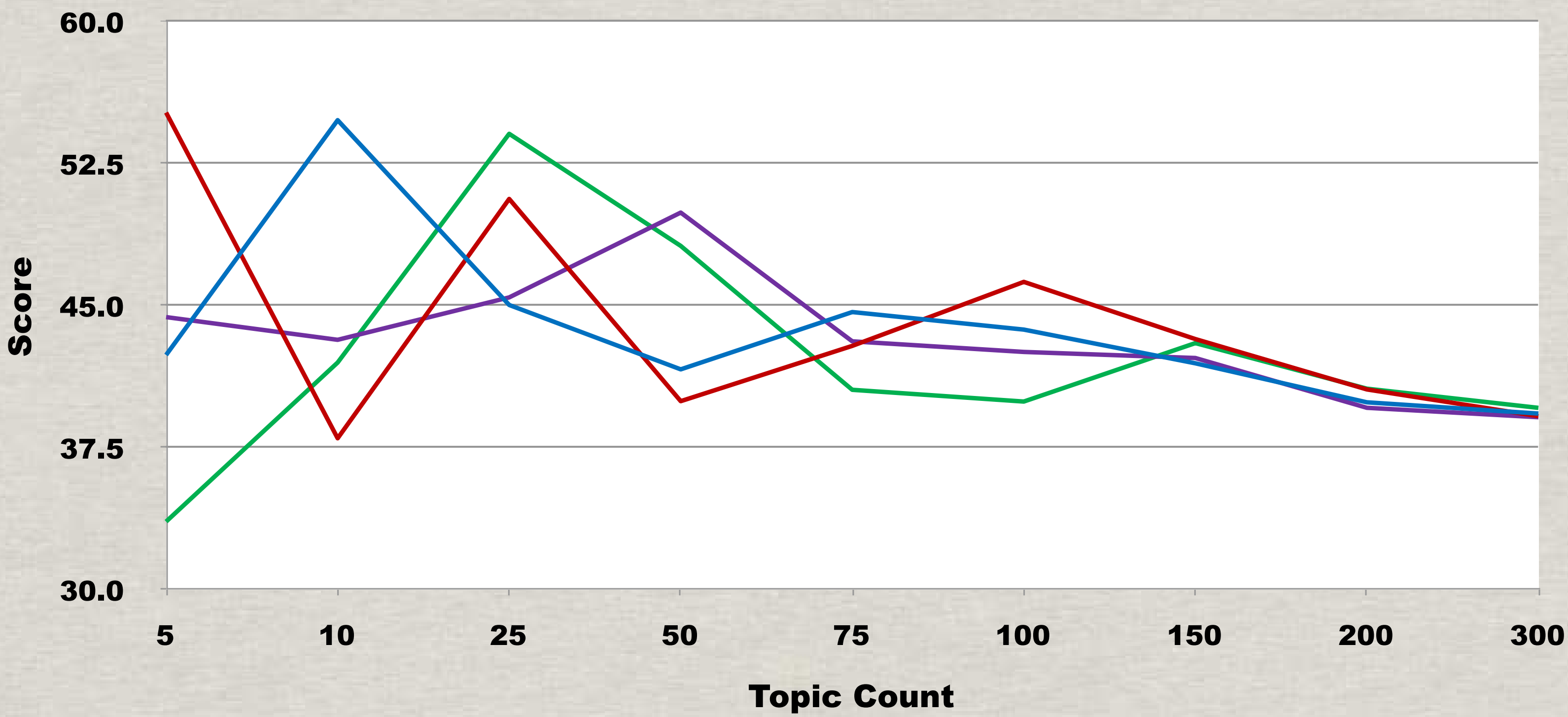
Twenty-Five Topics is Best



Fifty Topics is Best



No One Sample Is **THE** answer



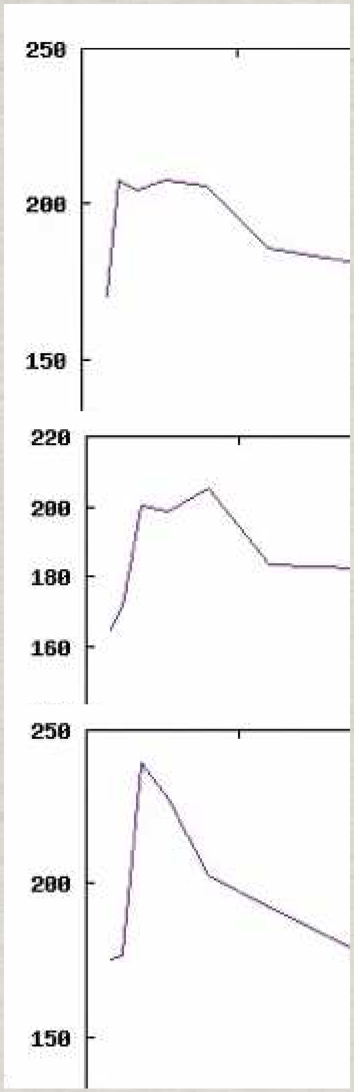
One vs. All Samples

One Sample

Run 1

Run 2

Run 3



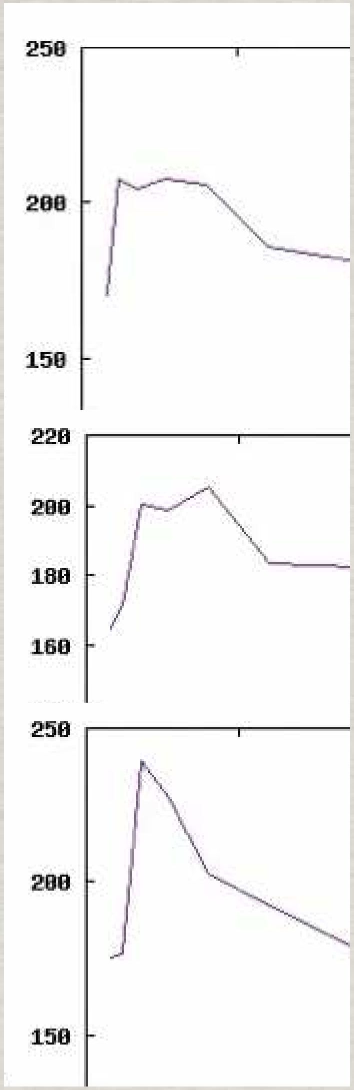
One vs. All Samples

One Sample

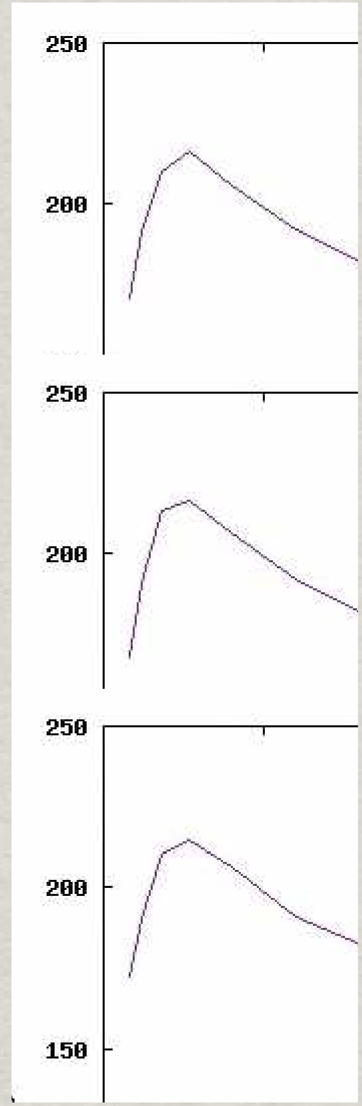
Run 1

Run 2

Run 3



50 Samples



Key Points

- Two Key points
  - LDA requires multiple ***samples***
  - The (sampling and hyper) parameters aren't just "***set and go***"

Parameters Abound

- LDA involves
  - Parameters
  - Hyper-Parameters
  - Gibbs' Sampling Parameters

LDA Reminder

- The **parameters** θ and ϕ specify a statistical model for generating random documents (sequences of words)

Finding θ and ϕ

- *Fitting* an LDA model to a corpus of documents means inferring θ and ϕ
  - parameterized by ***hyper-parameters***
  - done using **Gibbs' sampling**

Gibbs and Hyper Parameters

- Gibbs' Sampling
  - n – samples (random variates)
  - b – burn-in iterations
  - si – sampling interval
- Hyper-Parameters
  - tc – topic count (number of topics) in the model
  - α – Dirichlet prior on the per-document topic distribution, θ
  - β – Dirichlet prior on the per-topic word distribution, ϕ

Sampling Interval

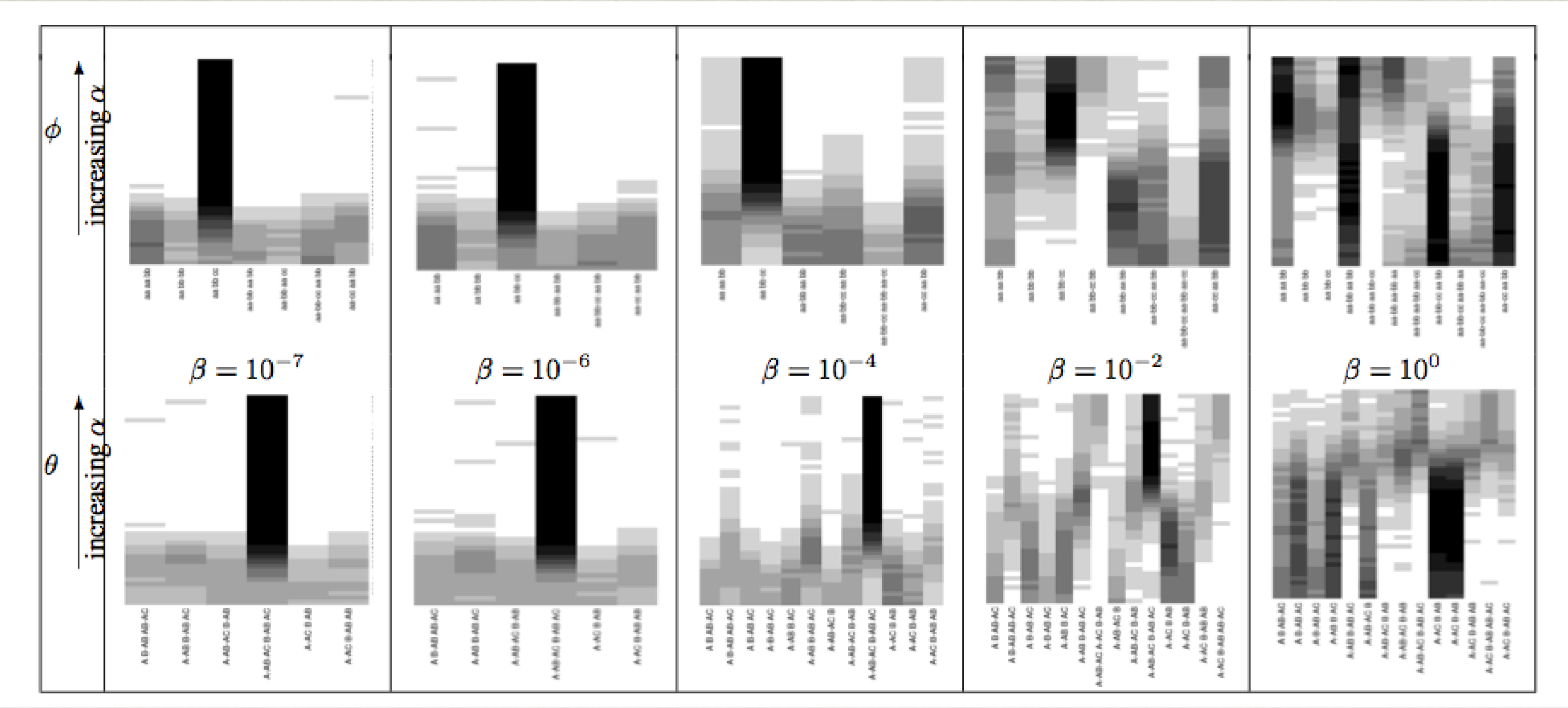
Katrina

Shah Rukh

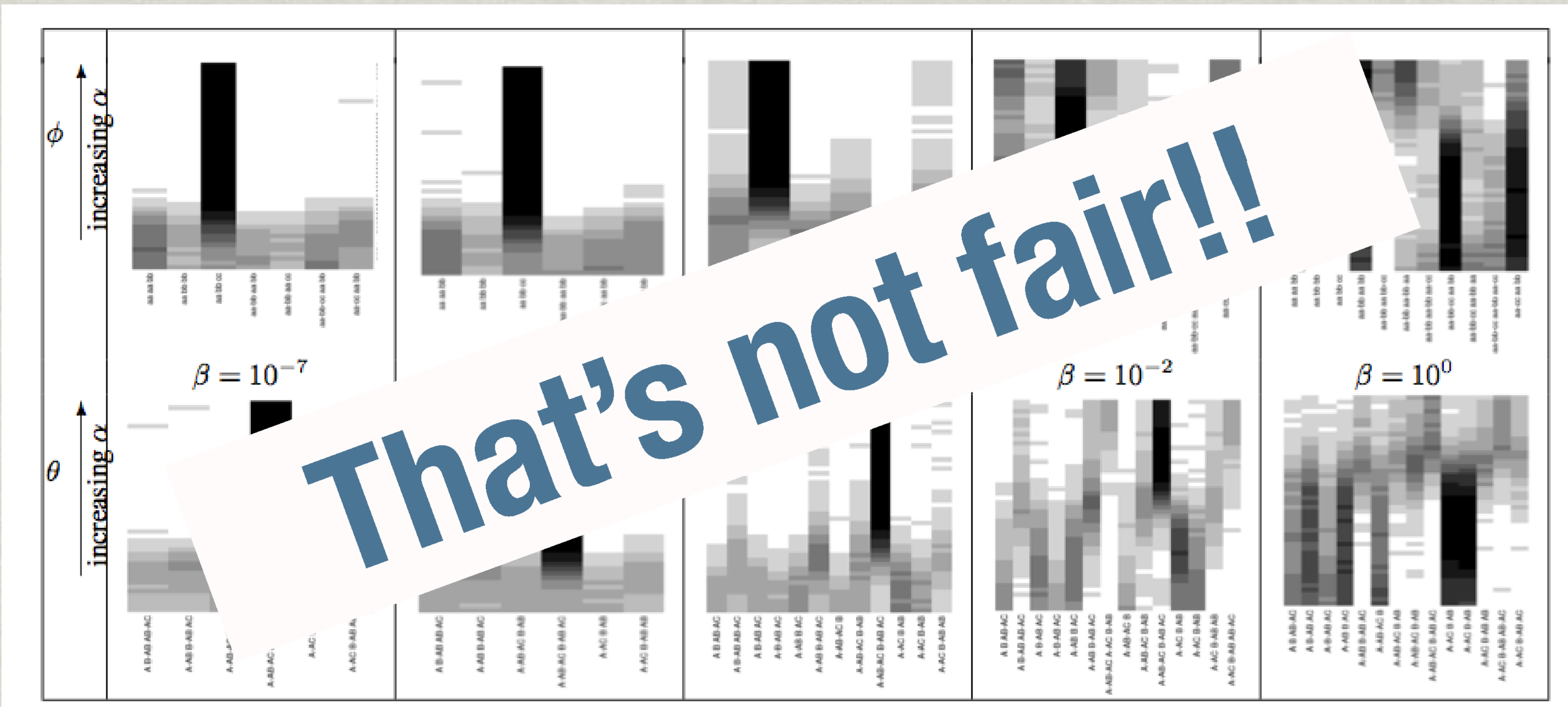
Hyper-Parameters
Topic Count

- Getting it wrong
  - **too many** topics yields diluted topics
    - t1 = {boot}, t2 = {shoe}, t3 = {sneaker}
  - **too few** topics yields non-discriminating topics
    - {bank, money, mud, river, robbery}

Finally α and β



Finally α and β



Impact of β
(the paper expands on this visually)

β determines the strength of prior belief that each topic is a uniform mixture of the words

- larger β
  - means a more uniform word per topic distribution (the prior drowns out the empirical information of the corpus)
  - favors topics with more words
- smaller β
  - means more impact from the corpus
  - favors fewer words per topic

Impact of α
(the paper expands on this visually)

α determines the strength of prior belief that each document is a uniform mixture of the topics

- larger α
  - means a more uniform topic per document distribution (the prior drowns out the empirical information of the corpus)
  - favors documents that include more topics
- smaller α
  - means more impact from the corpus
  - favors fewer topics per document

Hey I'm a Software Engineer

- When **refactoring** code where each topic captures a concept from the code a **smaller** α will encourage documents to receive a dominant topic and consequently suggest a refactoring
- When summarizing methods a smaller β encourages each topic to be dominated by a small number of key words.
- When generating concept labels a smaller β encourages topics that included fewer words that provide more focused concept labels.
- For feature location a larger α encourages more topics per document and thus increases recall as a topic of interest is more likely lead to a document of interest.

Hey I'm a Software Engineer

- When refactoring code where each topic captures a concept from the code a smaller α will encourage documents to receive a dominant topic and consequently suggest a refactoring
- When s**ummarizing methods** a **smaller** β encourages each topic to be dominated by a small number of key words.
- When generating concept labels a smaller β encourages topics that included fewer words that provide more focused concept labels.
- For feature location a larger α encourages more topics per document and thus increases recall as a topic of interest is more likely lead to a document of interest.

Hey I'm a Software Engineer

- When refactoring code where each topic captures a concept from the code a smaller α will encourage documents to receive a dominant topic and consequently suggest a refactoring
- When summarizing methods a smaller β encourages each topic to be dominated by a small number of key words.
- When generating **concept labels** a **smaller** β encourages topics that included fewer words that provide more focused concept labels.
- For feature location a larger α encourages more topics per document and thus increases recall as a topic of interest is more likely lead to a document of interest.

Hey I’m a Software Engineer

- When refactoring code where each topic captures a concept from the code a smaller α will encourage documents to receive a dominant topic and consequently suggest a refactoring
- When summarizing methods a smaller β encourages each topic to be dominated by a small number of key words.
- When generating concept labels a smaller β encourages topics that included fewer words that provide more focused concept labels.
- For **feature location** a **larger** α encourages more topics per document and thus increases recall as a topic of interest is more likely lead to a document of interest.

Hey I'm a Software Engineer

- When refactoring code where each topic captures a concept from the code a smaller α will encourage documents to receive a dominant topic and consequently suggest a refactoring
- When summarizing methods a smaller β encourages each topic to be dominated by a small number of key words.
- When generating concept labels a smaller β encourages topics that included fewer words that provide more focused concept labels.
- For **feature location** a **larger** α encourages more topics per document and thus increases recall as a topic of interest is more likely lead to a document of interest.

This is also supported by a **large** β value which leads to more inclusive topics

In Summary

1. Sampling is required! Don't leave home without it!!
2. The **parameters aren't just "*set and go*"**
   - Both the **problem** being solved and the **objectives** of the software engineer impact the choice ... especially for α and β

- On your train/flight home check out how the paper visually illustrates each parameter's impact
- www.cs.loyola.edu/~binkley/topic_models/additional-images

In Summary

Thanks!

1. Sampling is required! Don't leave home without it!!

2. The **parameters aren't just "*set and go*"**
   - Both the **problem** being solved and the **objectives** of the software engineer impact the choice ... especially for α and β

- On your train/flight home check out how the paper visually illustrates each parameter's impact

- www.cs.loyola.edu/~binkley/topic_models/additional-images