HUDK 4051: LEARNING ANALYTICS: PROCESS & THEORY

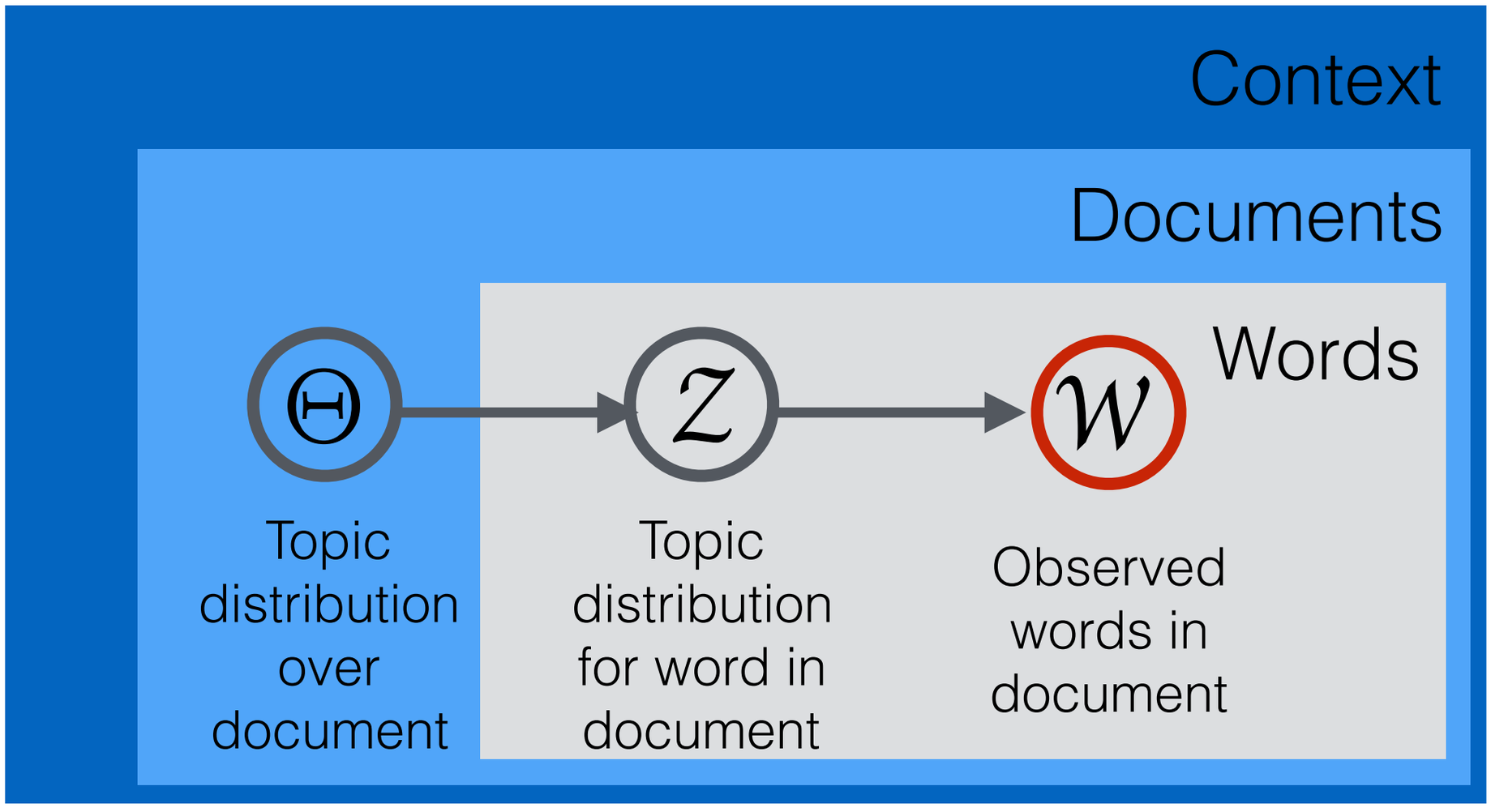
3/26/19 5:03 PM

Topic Modeling with Latent Dirichlet Analysis (LDA)

Topic Modeling

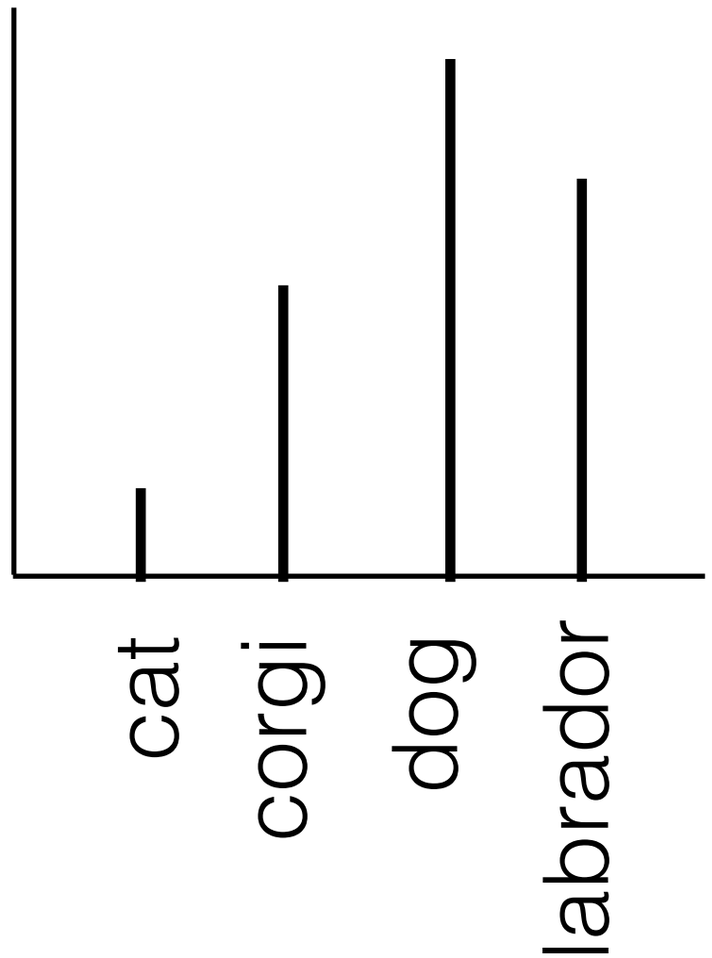
A topic model is a type of statistical model for discovering the abstract topics that occur in a collection of documents

Organizing Words

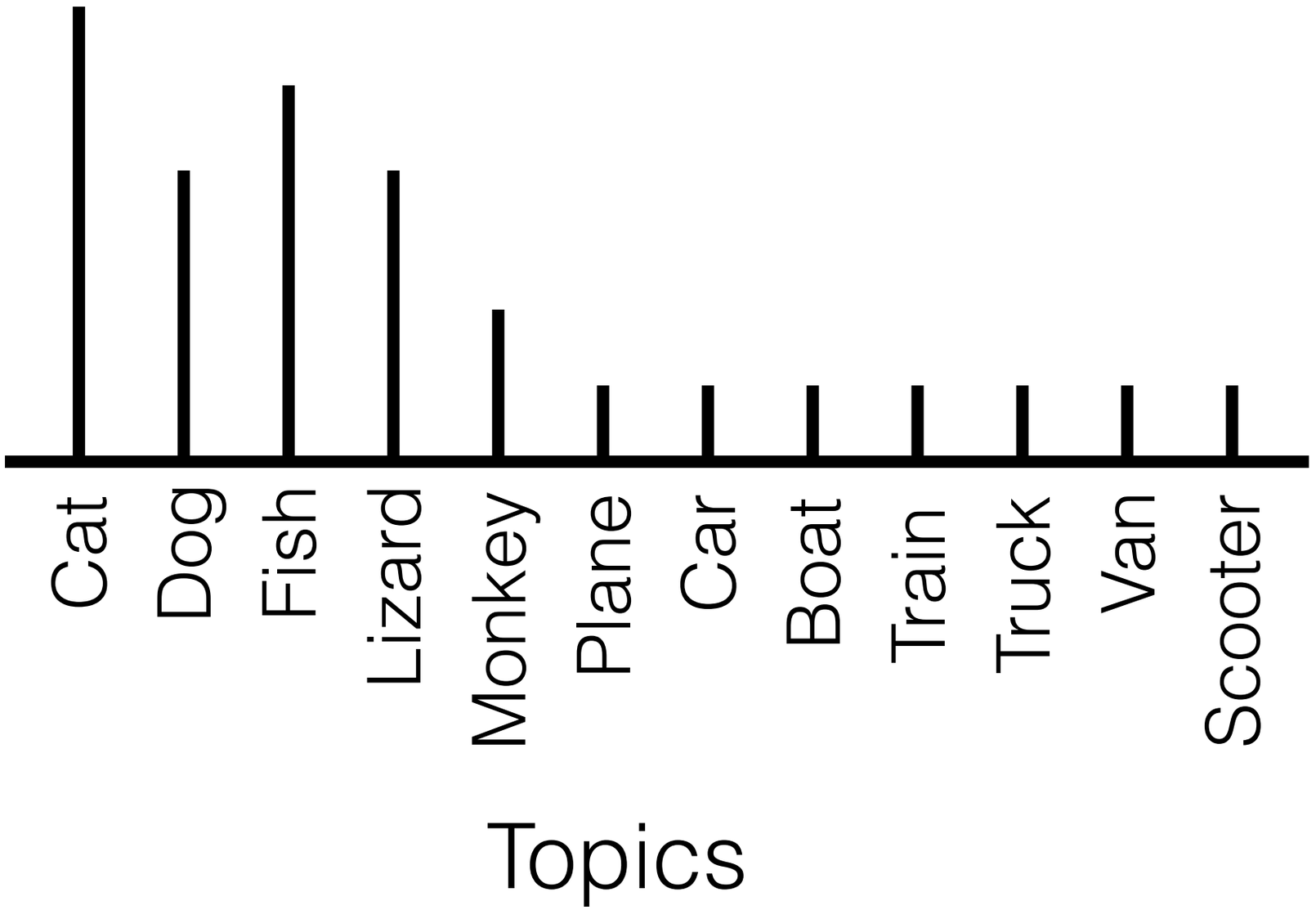


Topics (Z)

- A topic is a probability distribution over words



Topic Distribution for a Document



A document can be described by a recipe of topics and “how much” of each topic it contains

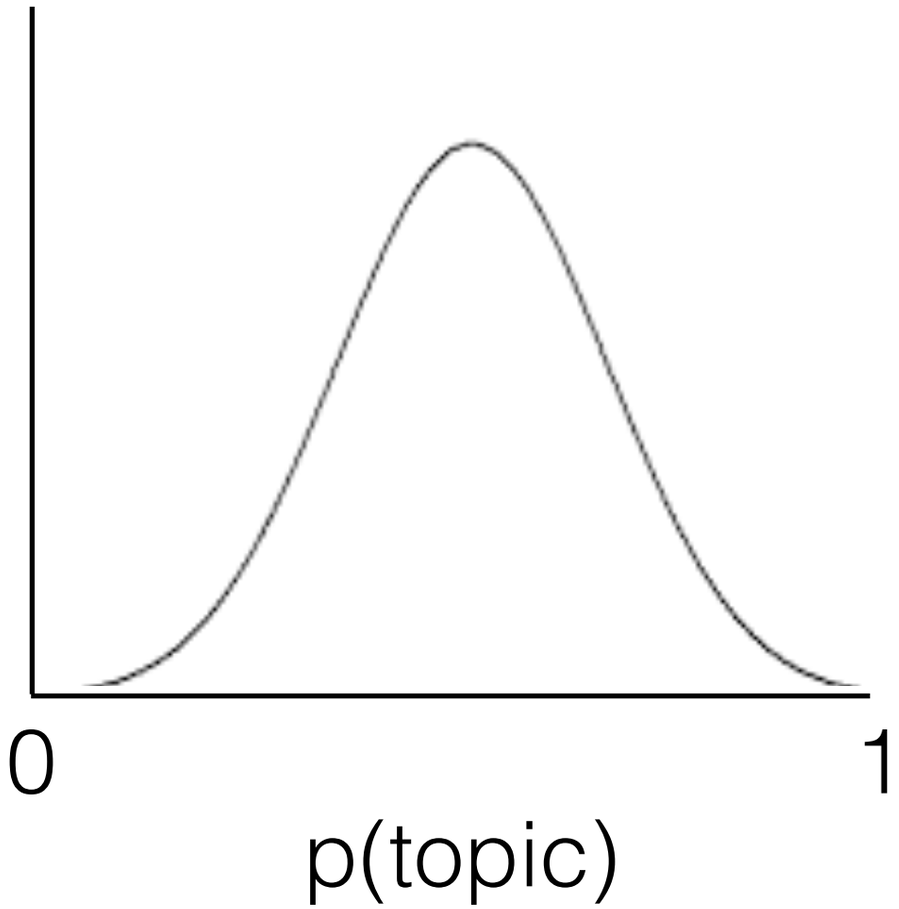
Documents

- A document is a probability distribution over topics

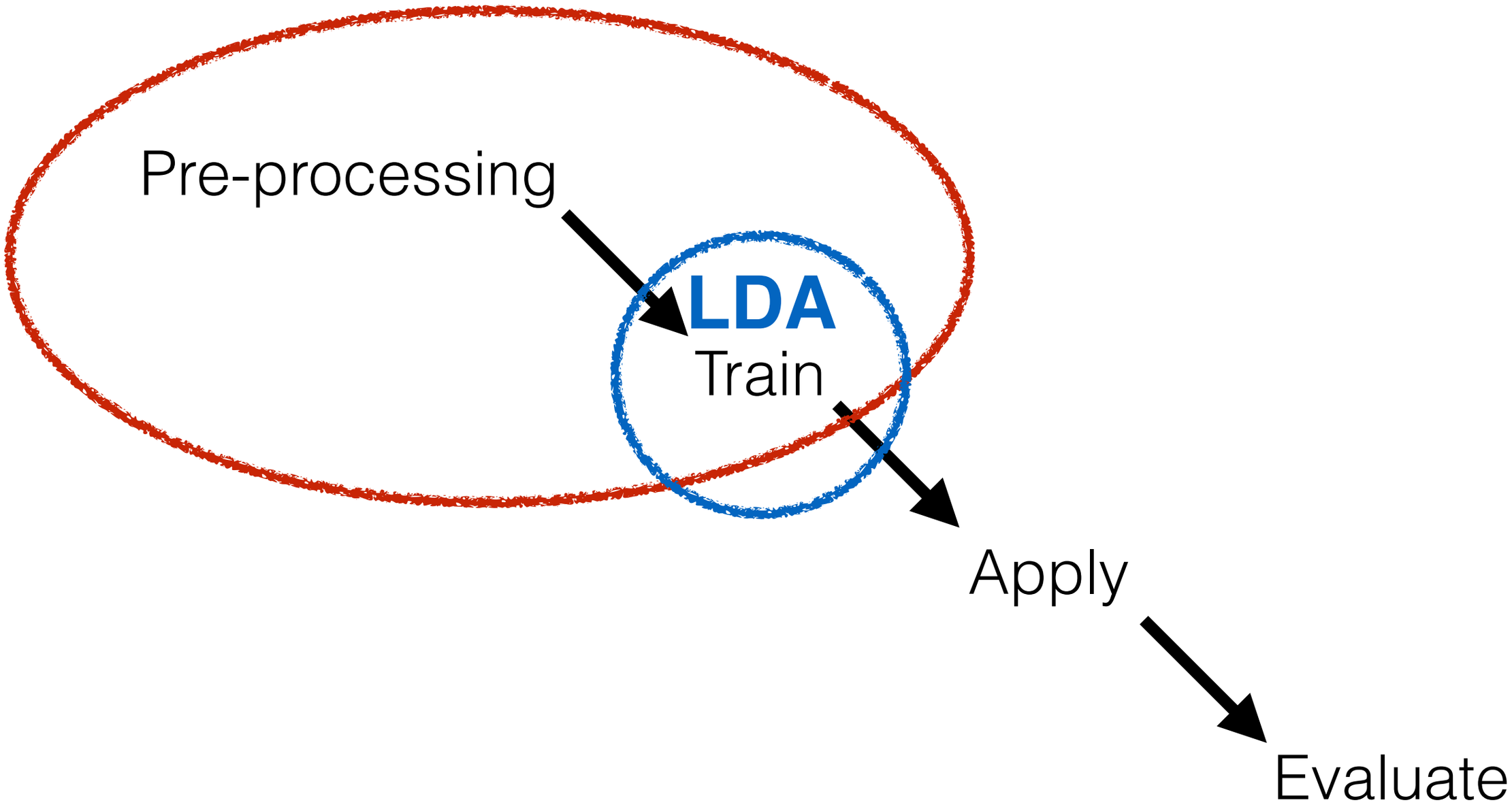
Document

word word word
word word word
word word word
word word word
word word word
word word word
word word word

Topic 1
Topic 2
Topic 3



Process



What does LDA do?

- Assumes that documents cover particular topics and particular topics are covered by particular words

- Therefore, can group similar documents by their word profiles which represent topics

- LDA calculates those distributions

- Like cluster analysis we need to supply the number of topics

Logic of Process

Document

word word word
word word word
word word word
word word word
word word word
word word word
word word word

Topic 1

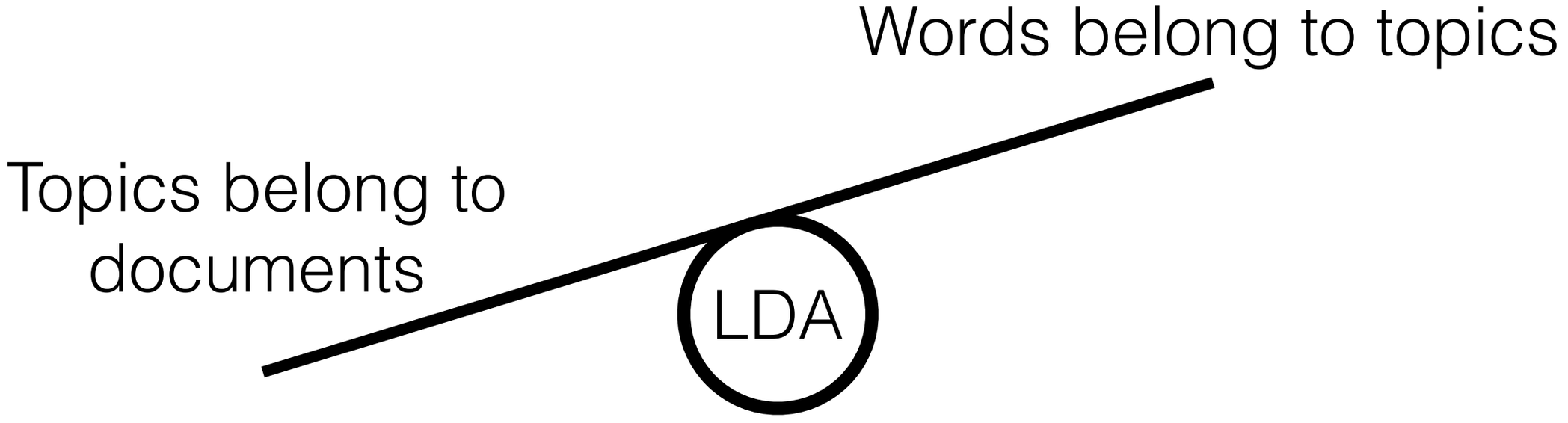
Topic 2

Topic 3

Basic Idea

- Documents are made up of words that belong (with some probability) to topics
- So…We can just reverse engineer these words to learn what a document is about

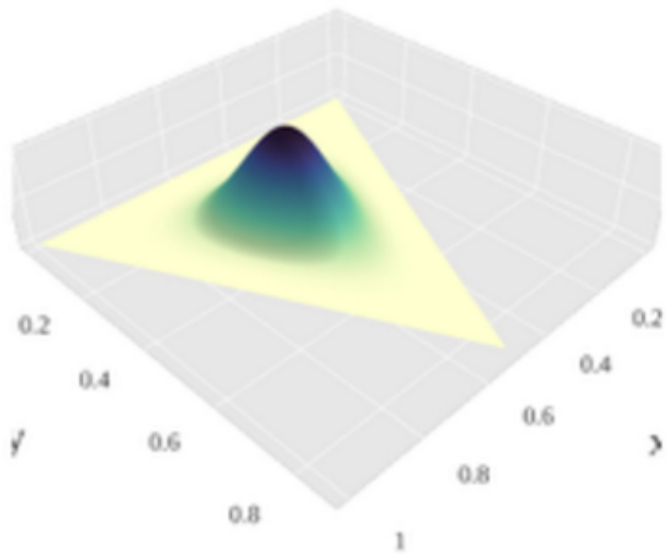
LDA

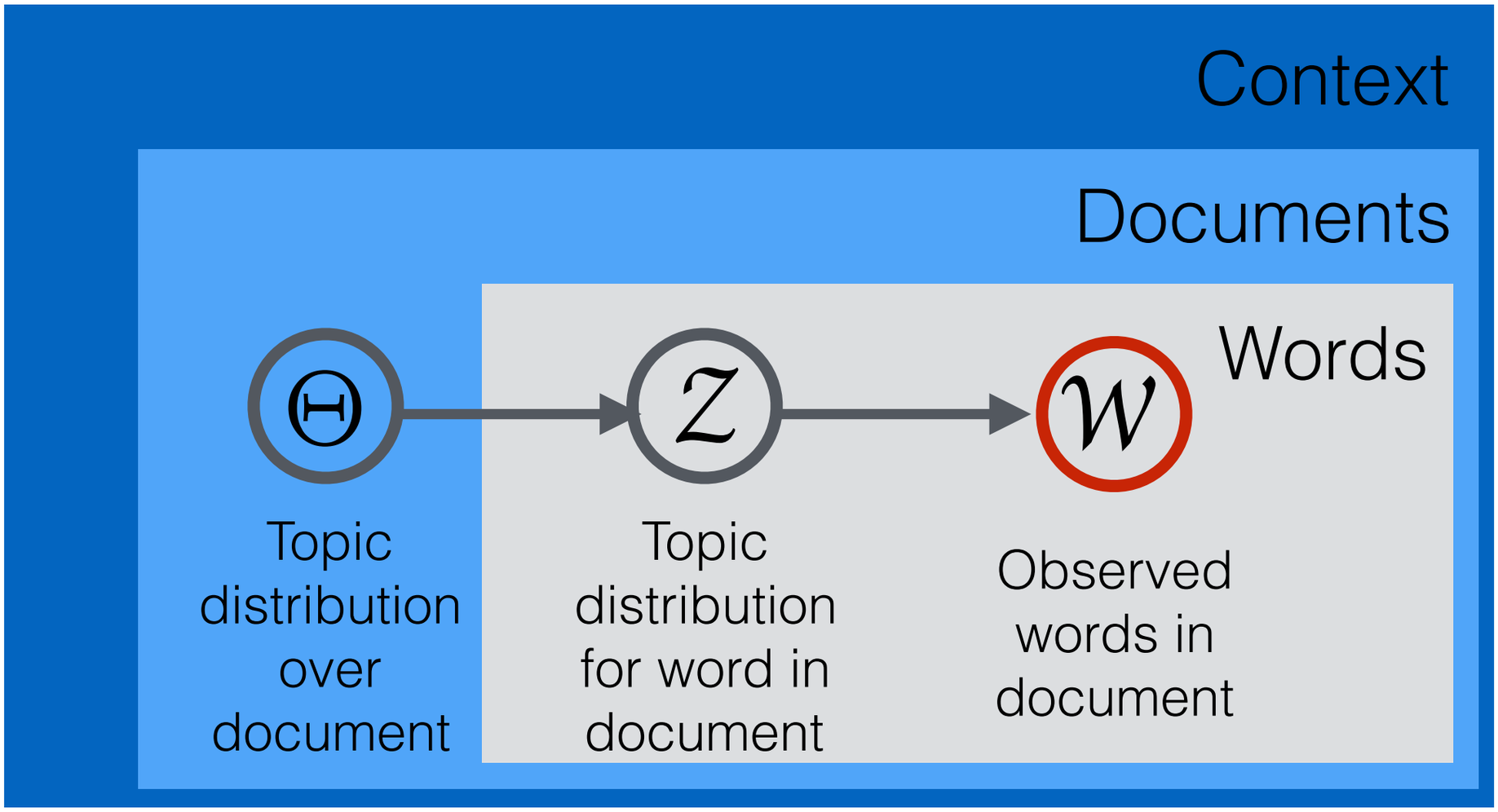


Dirichlet Distribution

- Peter Gustav Lejeune Dirichlet
- 1805 - 1859
- German mathematician
- Helped develop the definition of the word *function*

- Distribution on probability distributions




Context
Documents
Words
Θ
Z
W
Topic distribution over document
Topic distribution for word in document
Observed words in document

Term Document vs. Document Term Matrices

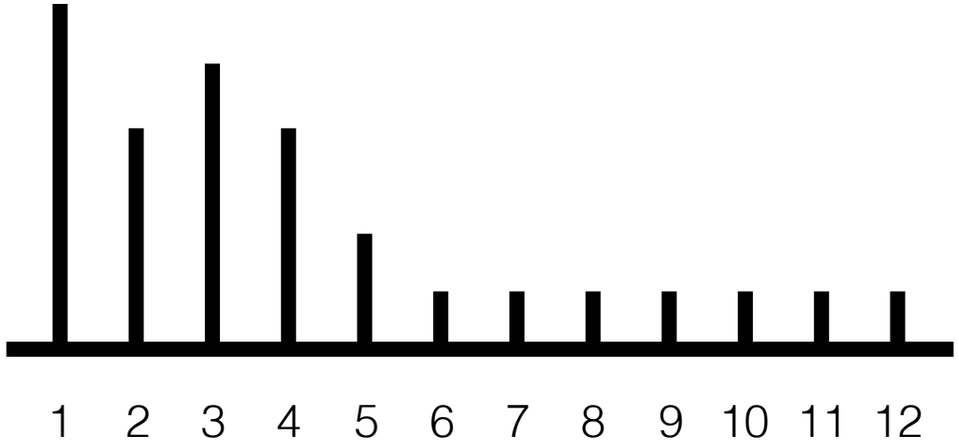
	Term1	Term2	Term3
Doc1			
Doc2			
Doc3			

	Doc1	Doc2	Doc3
Term1			
Term2			
Term3			

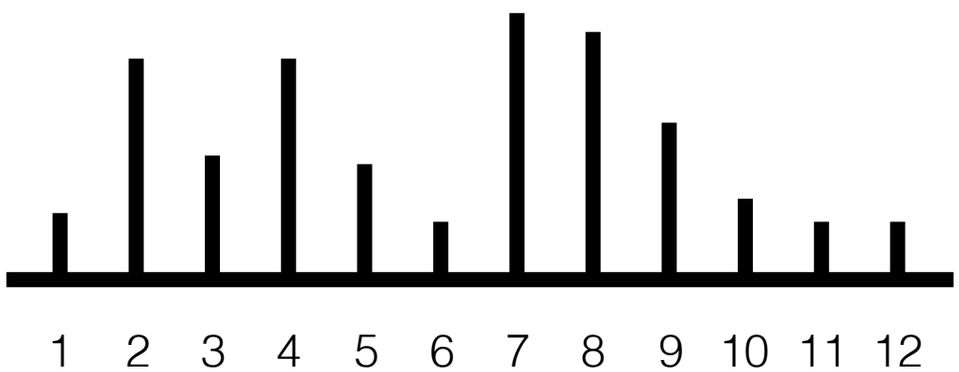
Term Frequency = Number of times a word appears in a document

Inverse Document Frequency = number of documents in the corpus which contain a term

Topic Distribution for a Document



Topic A



Topic B

New Document

If we have both of those pieces of information & the model…

We can predict the topic of a document

Sign Up for HUDK 5053

http://bit.ly/HUDK5053APP

Open laptop, install

https://github.com/rstudio/shiny

Shiny

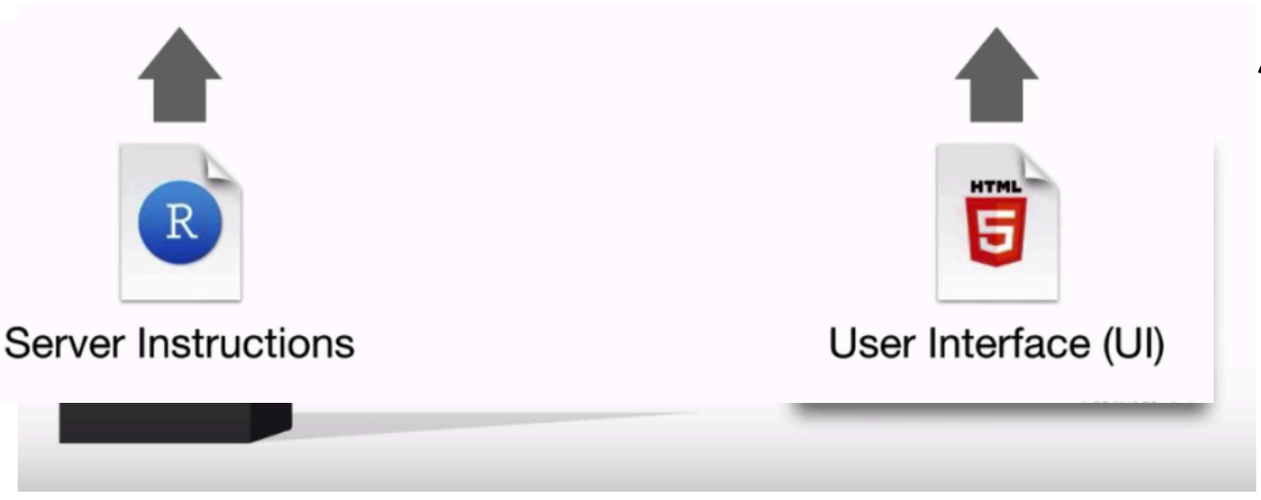
- Web Application Framework
- Allows you to make html applications from within R
- For us that means interactive data visualizations
- Example A*
- Example B*

Shiny

- Architecture
- Template
- Adding elements
- Reactive inputs
- Reactive results

Shiny Architecture

- Two components:
  - Computer running R
  - Webpage running html (user interface)



*

Shiny Template

```
library(shiny)

ui <- fluidPage()

server <- function(input, output) {}

shinyApp(ui = ui, server = server)
```

Example
HTML

Shiny Template

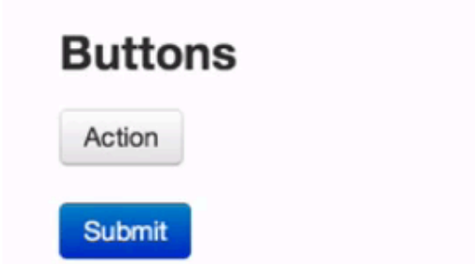
```
library(shiny)

ui <- fluidPage()

server <- function(input, output) {}

shinyApp(ui = ui, server = server)
```

Example
Stop Sign

Input Functions

Things that your user will see and manipulate.

Input Functions



`actionButton()`
`submitButton()`

`checkboxInput()`

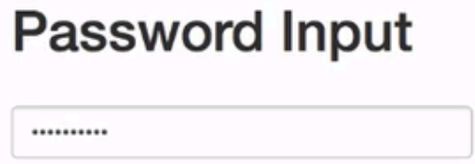
`checkboxGroupInput()`

`dateInput()`

`dateRangeInput()`

`fileInput()`

`numericInput()`



`passwordInput()`

`radioButtons()`

`selectInput()`

`sliderInput()`

`textInput()`

Input Function Syntax

```
xxxInput(inputId = “”, label = “”…)
```

Internal use

External use

Example

Output Function

Things that your user will see when they manipulate something in your web app.

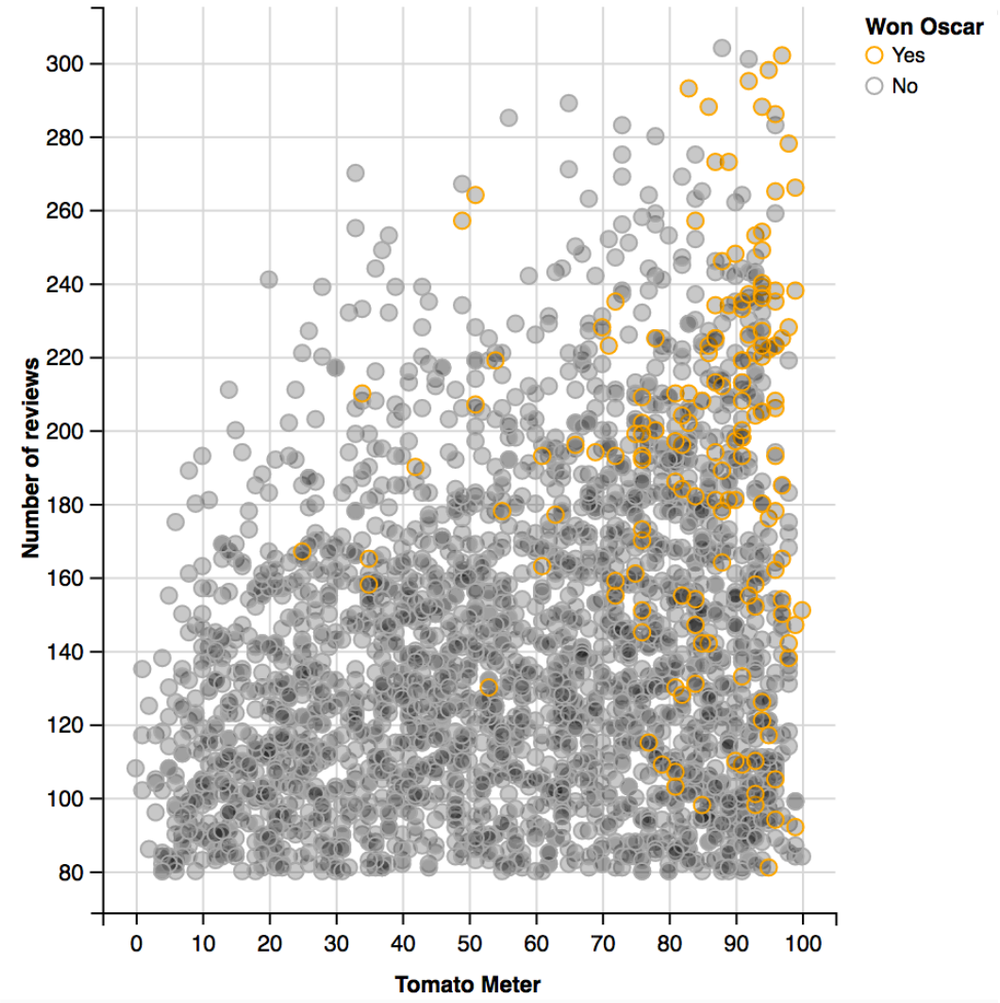
Output Function

Manufacturer: ford

Transmission: All

Show 10 **entries**

	manufacturer	model	displ	year	cyl	trans
1	ford	expedition 2wd	4.6	1999	8	auto(l4)
2	ford	expedition 2wd	5.4	1999	8	auto(l4)
3	ford	expedition 2wd	5.4	2008	8	auto(l6)
4	ford	explorer 4wd	4	1999	6	auto(l5)
5	ford	explorer 4wd	4	1999	6	manual(m5)
6	ford	explorer 4wd	4	1999	6	auto(l5)
7	ford	explorer 4wd	4	2008	6	auto(l5)
8	ford	explorer 4wd	4.6	2008	8	auto(l6)
9	ford	explorer 4wd	5	1999	8	auto(l4)
10	ford	f150 pickup 4wd	4.2	1999	6	auto(l4)



Output Function

Function	Inserts
`dataTableOutput()`	an interactive table
`htmlOutput()`	raw HTML
`imageOutput()`	image
`plotOutput()`	plot
`tableOutput()`	table
`textOutput()`	text
`uiOutput()`	a Shiny UI element
`verbatimTextOutput()`	text

Output Function Syntax

```
plotOutput(outputId = “name”)
```

Example

Shiny Template

```
library(shiny)

ui <- fluidPage()

server <- function(input, output) {}

shinyApp(ui = ui, server = server)
```