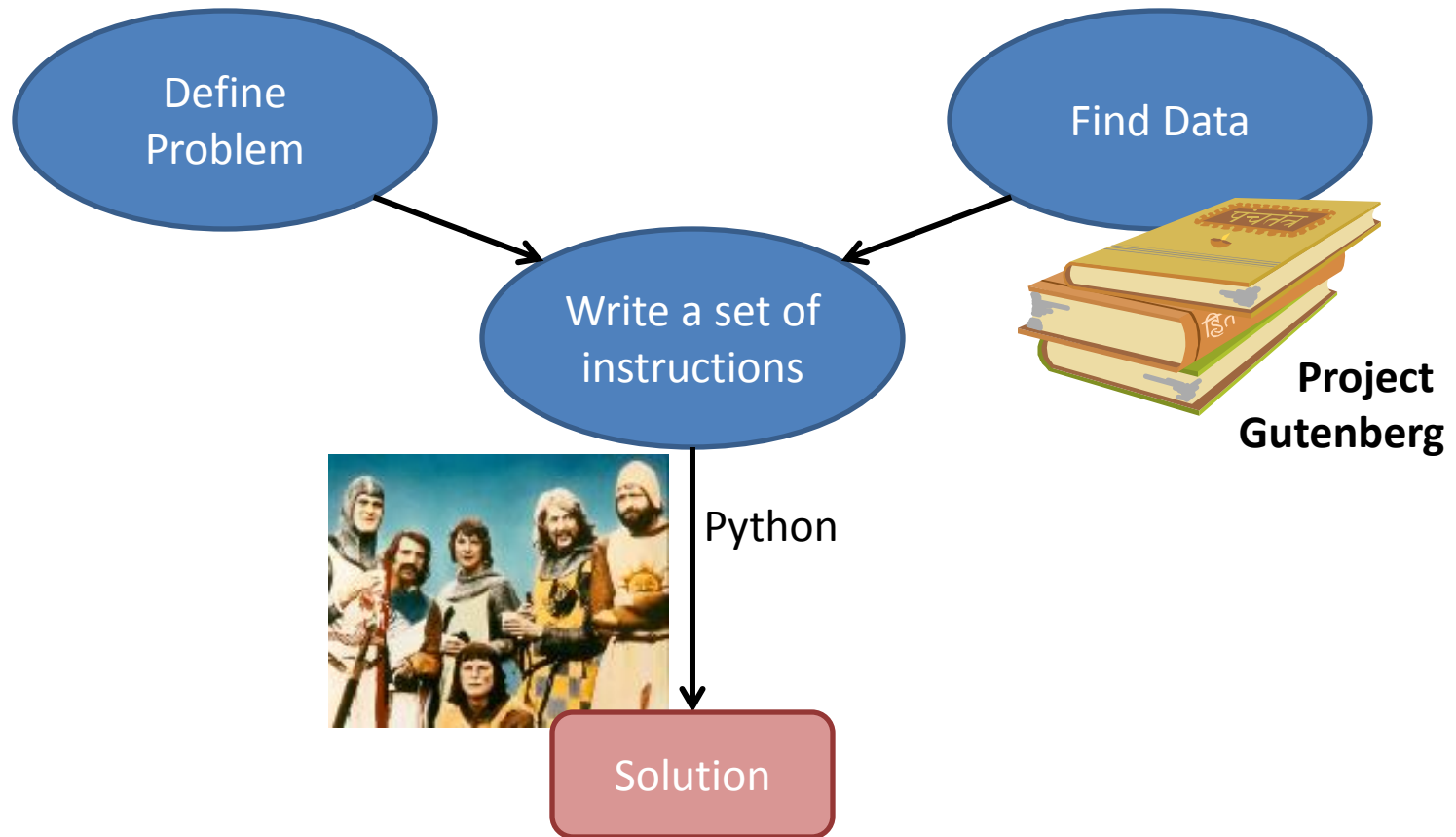


# Determining Authorship

March 21, 2013

# Determining Authorship



# Determining Authorship: Data



Five Books from a Famous  
Children's Series



One Book from a Famous  
Children's Series

# Determining Authorship: Data



Five Books from a Famous  
Children's Series



One Book from a Famous  
Children's Series



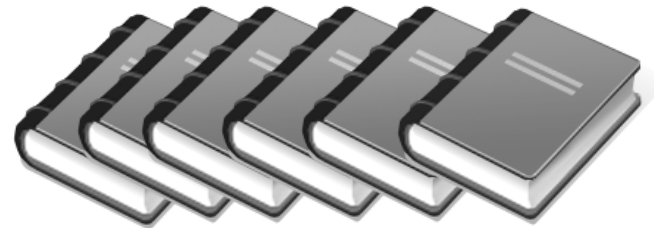
Six Books from Two Famous Children's Series

# Determining Authorship

Define Problem

Find Data

Write a set of instructions



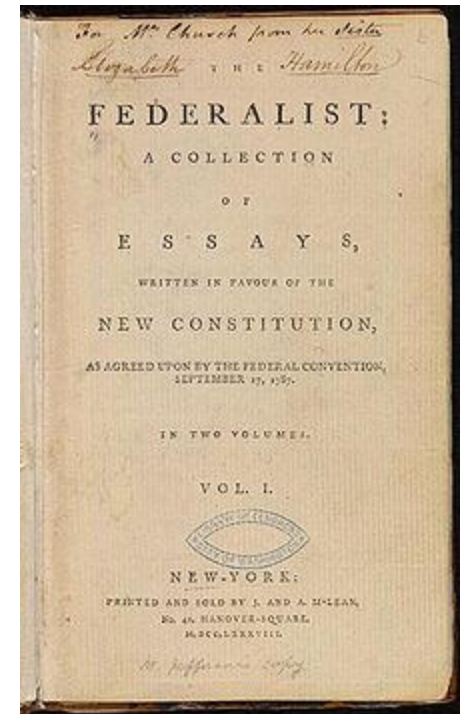
Python

Solution

Discern the **Outlier**:  
The one book that is  
NOT in the series of  
the others.

# Remember the Federalist Papers

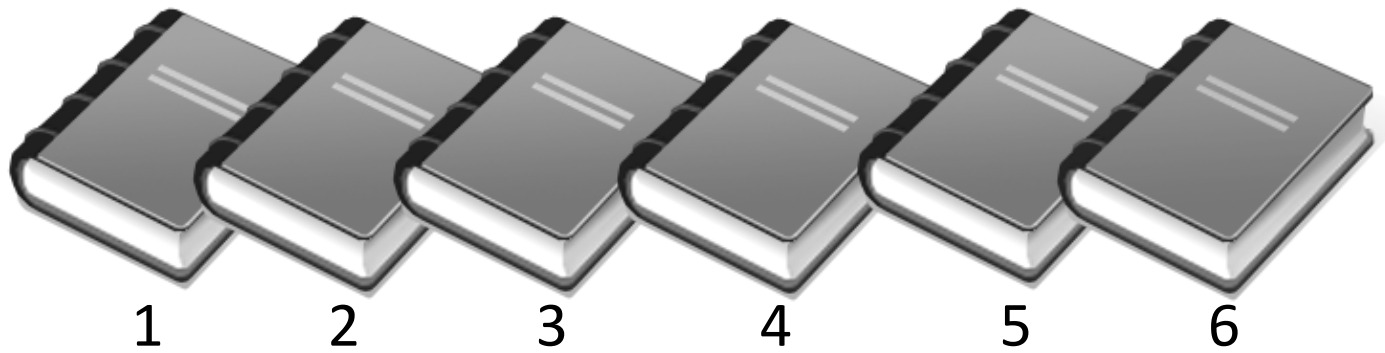
- 85 articles written in 1787 to promote the ratification of the US Constitution
- In 1944, Douglass Adair guessed authorship
  - Alexander Hamilton (51)
  - James Madison (26)
  - John Jay (5)
  - 3 were a collaboration
- Corroborated in 1964 by a computer analysis



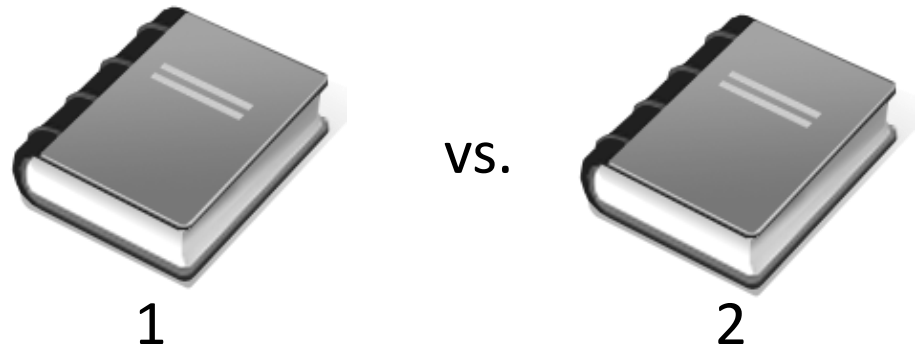
Wikipedia

<http://pages.cs.wisc.edu/~gfung/federalist.pdf>

# Determining Authorship

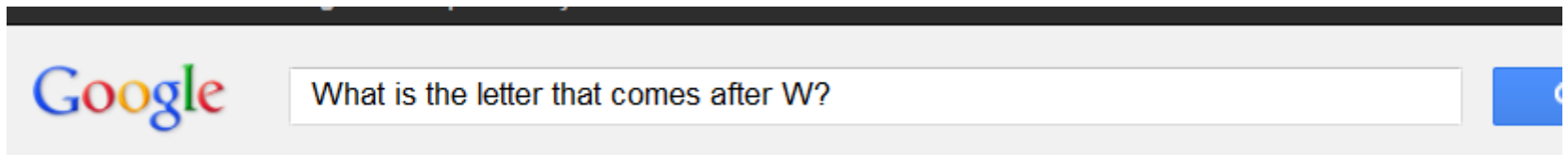


Discern the **Outlier**:  
The one book that is  
NOT in the series of  
the others.



# Stop Words

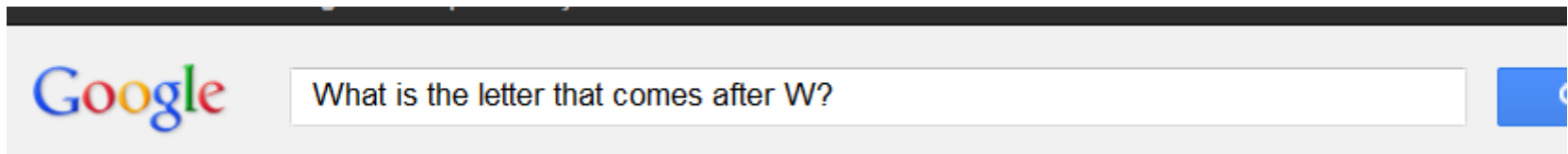
**Stop Words** are words that are filtered out in natural language processing





# Stop Words

**Stop Words** are words that are filtered out in natural language processing



[What letter comes after w](#)

[wiki.answers.com/Q/What\\_letter\\_comes\\_after\\_w](http://wiki.answers.com/Q/What_letter_comes_after_w)

What **letters come after** the **letter** C. The **letters** of the alphabet that follow C are: d e f g h i j k l m n o p q r s t u v w x y z. Does The **Letter A Come After** The **Letter** ...

[What letter comes after A in the alphabet](#)

[wiki.answers.com > ... > Alphabet History > English Alphabet History](#)

Why use alphabets **with** pictures? Answer it! ... What **letter comes after** the twelfth **letter** of the alphabet. L is the ... What **comes after** the **letter** a in the alphabet ...

# Stop Words

**Stop Words** are words that are filtered out in natural language processing

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, i, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

<http://www.textfixer.com/resources/common-english-words.txt>



# Determining Authorship

Discern the **Outlier**:  
The one book that is  
**NOT** in the series of  
the others.



vs.



1. Calculate the word frequencies of the stop words in the two books

	a	able	about	across	after	...
File 1	1000	238	483	12	3	...
File 2	102	93	10	0	15	...

# Determining Authorship

Discern the **Outlier**:  
The one book that is  
**NOT** in the series of  
the others.



vs.



1. Calculate the word frequencies of the stop words in the two books
2. Normalize the word frequencies

	a	able	about	across	after	...
File 1	.3	.01	.003	.0027	0.006	...
File 2	0.238	0.0932	0.0034	0.0021	0.05	...

# Determining Authorship

1. Calculate the word frequencies of the stop words in the two books
2. Normalize the word frequencies

	a	able	about	across	after	...
File 1	.3	.01	.003	.0027	0.006	...
File 2	0.238	0.0932	0.0034	0.0021	0.05	...

3. Design a **metric** to compare the two files
  - A metric is a function that defines a **distance** between two things

# Determining Authorship

1. Calculate the word frequencies of the stop words in the two books
2. Normalize the word frequencies

	a	able	about	across	after	...
File 1	.3	.01	.003	.0027	0.006	...
File 2	0.238	0.0932	0.0034	0.0021	0.05	...

3. Design a **metric** to compare the two files
  - A metric is a function that defines a **distance** between two things

Write a  
`compareTwo(list1, list2)`  
function that returns a float.

# Determining Authorship

Download and extract `ACT2-7.zip`

Compile and run `testFiles('output.csv')`



# Determining Authorship

Download and extract `ACT2-7.zip`

Compile and run `testFiles('output.csv')`

We are going to modify two things:

- `compareTwo` function
- Write distance matrix to a file

# Determining Authorship

Download and extract `ACT2-7.zip`

Compile and run `testFiles('output.csv')`

We are going to modify two things:

- `compareTwo` function
- Write distance matrix to a file

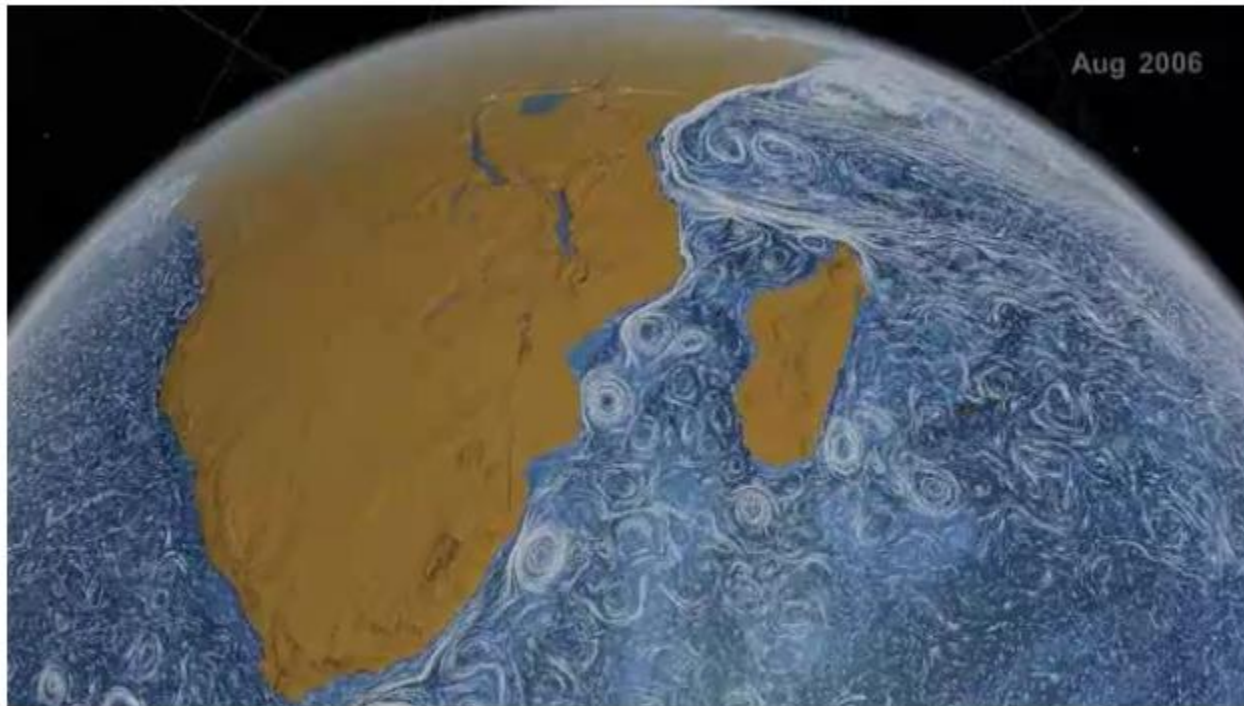
First, what does the current program do?

# Break

## Perpetual Ocean

[PerpetualOcean](#)

March 27, 2012 to Mapping · Add Comment · Share on Twitter



Using a computational model called Estimating the Circulation and Climate of the Ocean, Phase II (ECCO2), the NASA Goddard Space Flight Center Scientific Visualization Studio (I think NASA has a thing for long names.) [visualizes surface currents around the world](#). This is beautiful science here. Make sure you turn on high-def and go full screen.

# Distance Matrix

This matrix looks kind of familiar...

# Distance Matrix

This matrix looks kind of familiar...

Instead of printing to the screen, write it to a file in CSV (comma-separated value) format.

```
myNum = 1
myFile = open('output.csv', 'w')
myFile.write('this is an output file\n')
myFile.write(str(myNum))
myFile.write('\n')
myFile.close()
```

# Distance Matrix

This matrix looks kind of familiar...

Instead of printing to the screen, write it to a file in CSV (comma-separated value) format.

```
myNum = 1
myFile = open('output.csv', 'w')
myFile.write('this is an output file\n')
myFile.write(str(myNum))
myFile.write('\n')
myFile.close()
```

```
this is an output file
1
```

# Distance Matrix

This matrix looks kind of familiar...

Instead of printing to the screen, write it to a file in CSV (comma-separated value) format.

Open the CSV file in Excel. Use conditional formatting to look for patterns.

# What's Your Answer?

Discern the **Outlier**:  
The one book that is NOT in the series of the others.

File	Title	Series	Author
file1.txt			
file2.txt			
file3.txt			
file4.txt			
file5.txt			
file6.txt			



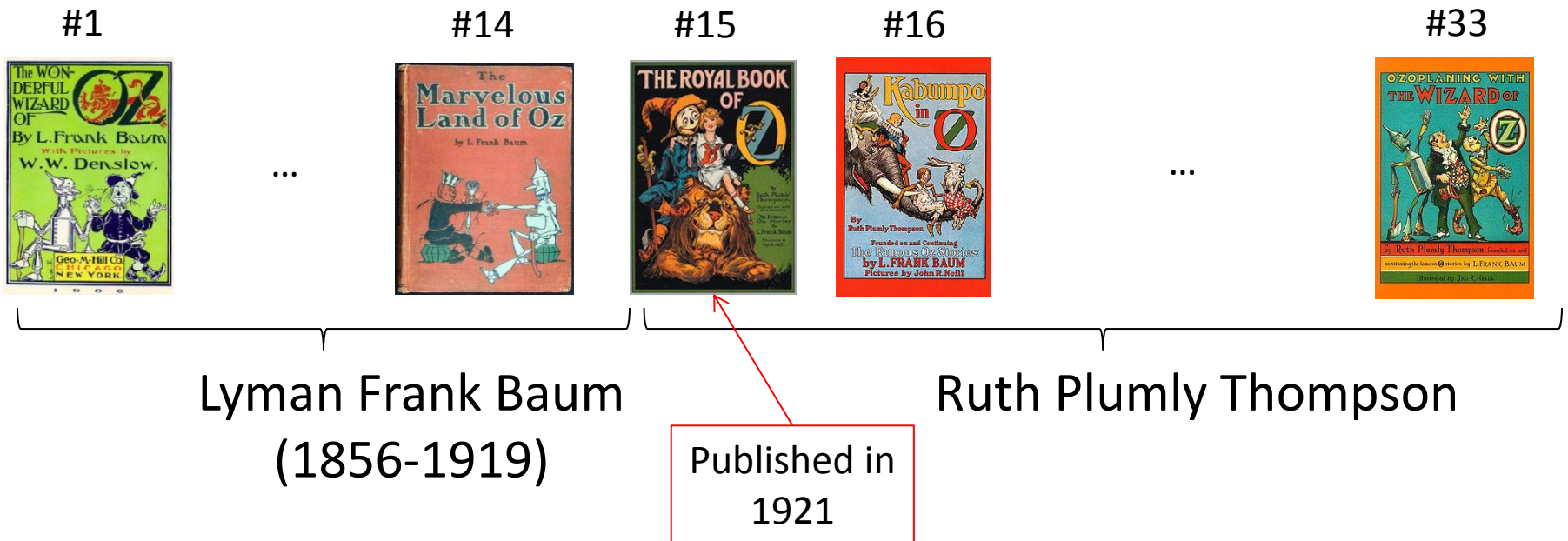
# What's Your Answer?

Discern the **Outlier**:  
The one book that is NOT in the series of the others.

File	Title	Series	Author
file1.txt	Wonder Wizard of Oz	Oz	
file2.txt	Alice's Adventures in Wonderland	Alice in Wonderland	
file3.txt	Dorothy and the Wizard in Oz	Oz	
file4.txt	Emerald City of Oz	Oz	
file5.txt	Royal Book of Oz	Oz	
file6.txt	Glinda of Oz	Oz	

# The Wizard of Oz

- About 40 Books, written by 7 different authors



<http://www.ssc.wisc.edu/~zzeng/soc357/OZ.pdf>

# What's Your Answer?

Discern the **Outlier**:  
The one book that is NOT in the series of the others.

File	Title	Series	Author
file1.txt	Wonder Wizard of Oz	Oz	Lyman Frank Baum
file2.txt	Alice's Adventures in Wonderland	Alice in Wonderland	Lewis Carroll
file3.txt	Dorothy and the Wizard in Oz	Oz	Lyman Frank Baum
file4.txt	Emerald City of Oz	Oz	Lyman Frank Baum
file5.txt	Royal Book of Oz	Oz	Ruth Plumly Thompson
file6.txt	Glinda of Oz	Oz	Lyman Frank Baum