Data Cleaning

February 6, 2020
Data Science CSCI 1951A
Brown University
Instructor: Ellie Pavlick
HTAs: Josh Levin, Diane Mutako, Sol Zitter

Thanks to C. Binning for some stolen slides. :)

Announcements

- Assignment 1: down! Assignment 2: up!
- Projects:
  - Let me know by today at 10:20 if you want to be N != 4
  - Being thinking about your project data…the first deliverable is not just a “ceremonial” checkpoint
  - “Will we know how to do X by in time?” —> maybe/probably/probably not but you should do it regardless!

Today

- 45 minutes—let’s just see how far we get….
- Problems with dirty data
- Cleaning and string matching heuristics
- Monday: bash commands (come with a command line…if you don’t know what that means, ask me)

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL

…

Problems?

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL

…

Problems?

Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL

…

Problems?

Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL

…

Missing Values

Problems?

Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL

…

Typos

Missing Values

Duplicates

Problems?

Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL

…

Typos

Missing Values

Duplicates

Problems?

Inconsistent Representations

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providence	RI	98106	42
2	Natalie Delworth	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	N Dellworth	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Minna Kimura-T	123 University Ave	Providence	Guyana	94305	NULL

…

Typos

Missing Values

Maybe Duplicates?

Dirty Data…

Dirty Data…

- Data is dirty on its own

Dirty Data…

- Data is dirty on its own
- Data sets are clean on their own but combining them introduces errors (e.g. duplicates, different naming conventions)

Dirty Data…

- Data is dirty on its own

- Data sets are clean on their own but combining them introduces errors (e.g. duplicates, different naming conventions)

- Data doesn’t “age well” (inflation, redistricting)

Dirty Data…

- Data is dirty on its own

- Data sets are clean on their own but combining them introduces errors (e.g. duplicates, different naming conventions)

- Data doesn’t “age well” (inflation, redistricting)

- Any combination of the above

Dirty Data…

Dirty Data…

- Parsing input data (e.g., separator issues)

Dirty Data…

- Parsing input data (e.g., separator issues)
- Naming conventions: NYC vs New York

Dirty Data…

- Parsing input data (e.g., separator issues)
- Naming conventions: NYC vs New York
- Formatting issues – esp. dates

Dirty Data…

- Parsing input data (e.g., separator issues)
- Naming conventions: NYC vs New York
- Formatting issues – esp. dates
- Missing values and required fields (e.g., always use 0)

Dirty Data…

- Parsing input data (e.g., separator issues)
- Naming conventions: NYC vs New York
- Formatting issues – esp. dates
- Missing values and required fields (e.g., always use 0)
- Different representations (2 vs Two)

Dirty Data…

- Parsing input data (e.g., separator issues)
- Naming conventions: NYC vs New York
- Formatting issues – esp. dates
- Missing values and required fields (e.g., always use 0)
- Different representations (2 vs Two)
- Fields too long (get truncated)

Dirty Data…

- Parsing input data (e.g., separator issues)
- Naming conventions: NYC vs New York
- Formatting issues – esp. dates
- Missing values and required fields (e.g., always use 0)
- Different representations (2 vs Two)
- Fields too long (get truncated)
- Primary key violations (from data merging)

Dirty Data…

- Parsing input data (e.g., separator issues)
- Naming conventions: NYC vs New York
- Formatting issues – esp. dates
- Missing values and required fields (e.g., always use 0)
- Different representations (2 vs Two)
- Fields too long (get truncated)
- Primary key violations (from data merging)
- Redundant Records (from data merging)

Clicker Questions!

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

How will the dirty data affect the results of this query?

(a) Too high

(b) Too low

(c) Unaffected

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

How will the dirty data affect the results of this query?

(a) Too high

(b) Too low

(c) Unaffected

How many TAs are there?

```
SELECT COUNT(*)
FROM TAS
```

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

How will the dirty data affect the results of this query?

[✓ (a) Too high]

(b) Too low

(c) Unaffected

Duplicates -> Double Counting

How many TAs are there?

```
SELECT COUNT(*)
FROM TAS
```

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

How will the dirty data affect the results of this query?

(a) Too high

(b) Too low

(c) Unaffected

How many TAs have worked zero hours?

```
SELECT COUNT(*)
FROM TAS
WHERE Hours = 0
```

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

How will the dirty data affect the results of this query?

(a) Too high

[✓ (b) Too low]

(c) Unaffected

How many TAs have worked zero hours?

```
SELECT COUNT(*)
FROM TAS
WHERE Hours = 0
```

NULLS arent included in the where clause

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

How will the dirty data affect the results of this query?

(a) Too high

(b) Too low

(c) Unaffected

How many hours do my commuter TAs work?

```
SELECT SUM(Hours)
FROM TAS
WHERE City != “Providence”
```

Clicker Lightening Round!

TAS

ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Natalie Delworth	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	N Dellworth	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	Minna Kimura-T	Warwick	RI	NULL

ID	Name	City	State	Hours
1	Nazem Aldroubi	Providence	Rhode Island	42
2	Natalie Delworth	Pawtucket	Rhode Island	30
3	Nam Do	Providence	Rhode Island	38
7	Minna Kimura-T	Warwick	Rhode Island	0

How will the dirty data affect the results of this query?

[✓ **(a)**] **Too high**

(b) Too low

(c) Unaffected

Inconsistent names, typos, and duplicates...

How many hours do my commuter TAs work?

```
SELECT SUM(Hours)
FROM TAS
WHERE City != "Providence"
```

What’s to be done?

What’s to be done?

- Look at your data!

What’s to be done?

- Look at your data!
- Maybe set (sensible) defaults

What’s to be done?

- Look at your data!
- Maybe set (sensible) defaults
- Maybe remove outliers

What’s to be done?

- Look at your data!
- Maybe set (sensible) defaults
- Maybe remove outliers
- Look at your data

What's to be done?

- Look at your data!
- Maybe set (sensible) defaults
- Maybe remove outliers
- Look at your data
- Maybe machine learn some of the things

What’s to be done?

- Look at your data!
- Maybe set (sensible) defaults
- Maybe remove outliers
- Look at your data
- Maybe machine learn some of the things
- Look at your data

What’s to be done?

- Look at your data!
- Maybe set (sensible) defaults
- Maybe remove outliers
- Look at your data
- Maybe machine learn some of the things
- Look at your data
- When you issue a query, don’t take the answer as gospel.

What’s to be done?

- Look at your data!
- Maybe set (sensible) defaults
- Maybe remove outliers
- Look at your data
- Maybe machine learn some of the things
- Look at your data
- When you issue a query, don’t take the answer as gospel. Instead…

What’s to be done?

- Look at your data!
- Maybe set (sensible) defaults
- Maybe remove outliers
- Look at your data
- Maybe machine learn some of the things
- Look at your data
- When you issue a query, don’t take the answer as gospel. Instead…wait for it…

What’s to be done?

- Look at your data!
- Maybe set (sensible) defaults
- Maybe remove outliers
- Look at your data
- Maybe machine learn some of the things
- Look at your data
- When you issue a query, don’t take the answer as gospel. Instead…wait for it…look at your data!

Look at your data

Look at your data

```
SELECT City, COUNT(*) as pop
FROM PEOPLE
GROUP BY Zip_Code
ORDER BY pop
```

Look at your data

```
SELECT City, COUNT(*) as pop
FROM PEOPLE
GROUP BY Zip_Code
ORDER BY pop
```

City	Count(*)
Schenectady	2,500
New York City	2,200
Los Angeles	1,900
Dallas	1,400

Look at your data

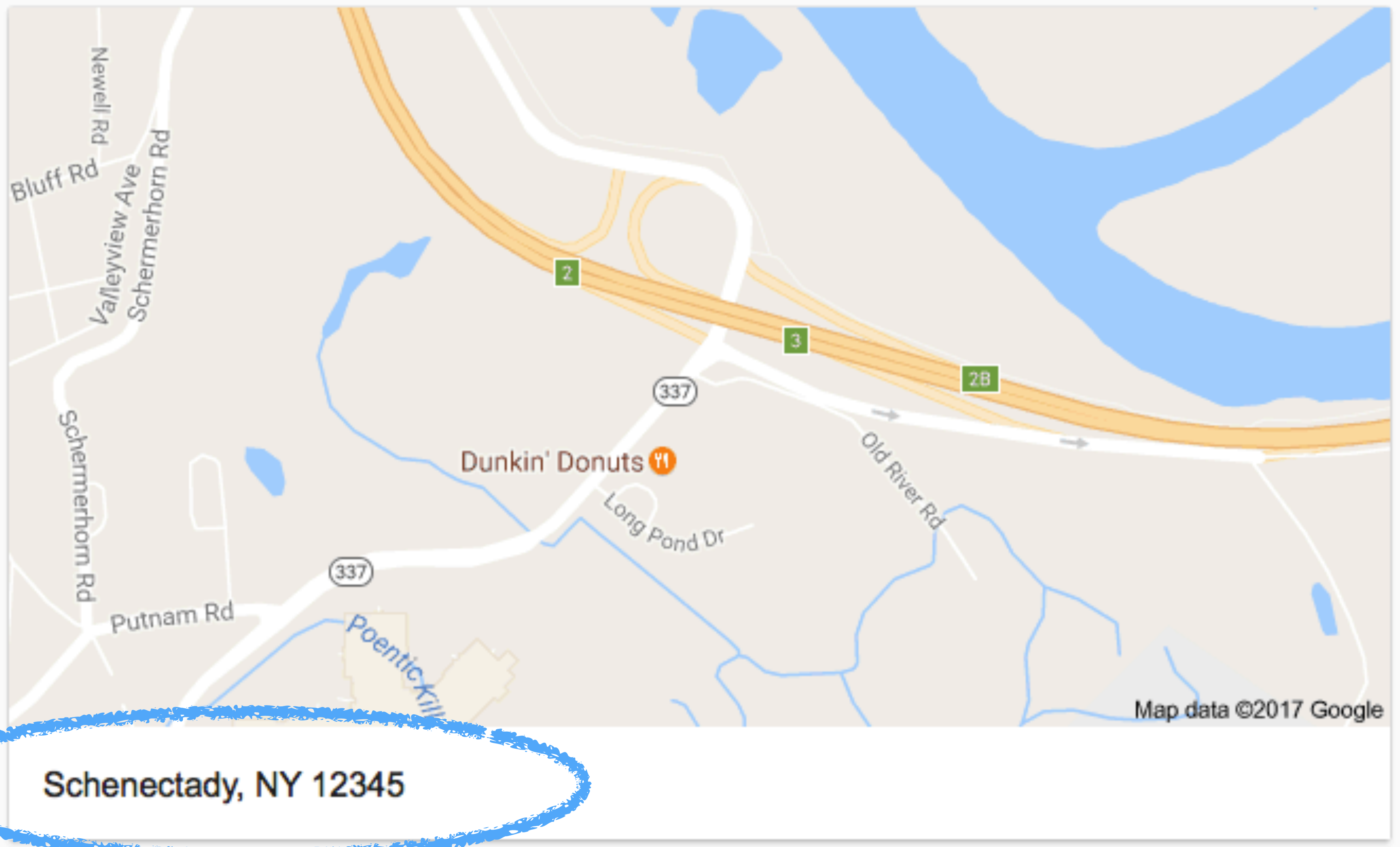
```
SELECT City, COUNT(*) as pop
FROM PEOPLE
GROUP BY Zip_Code
ORDER BY pop
```

City	Count(*)
Schenectady	2,500
New York City	2,200
Los Angeles	1,900
Dallas	1,400

?!?!

Loc

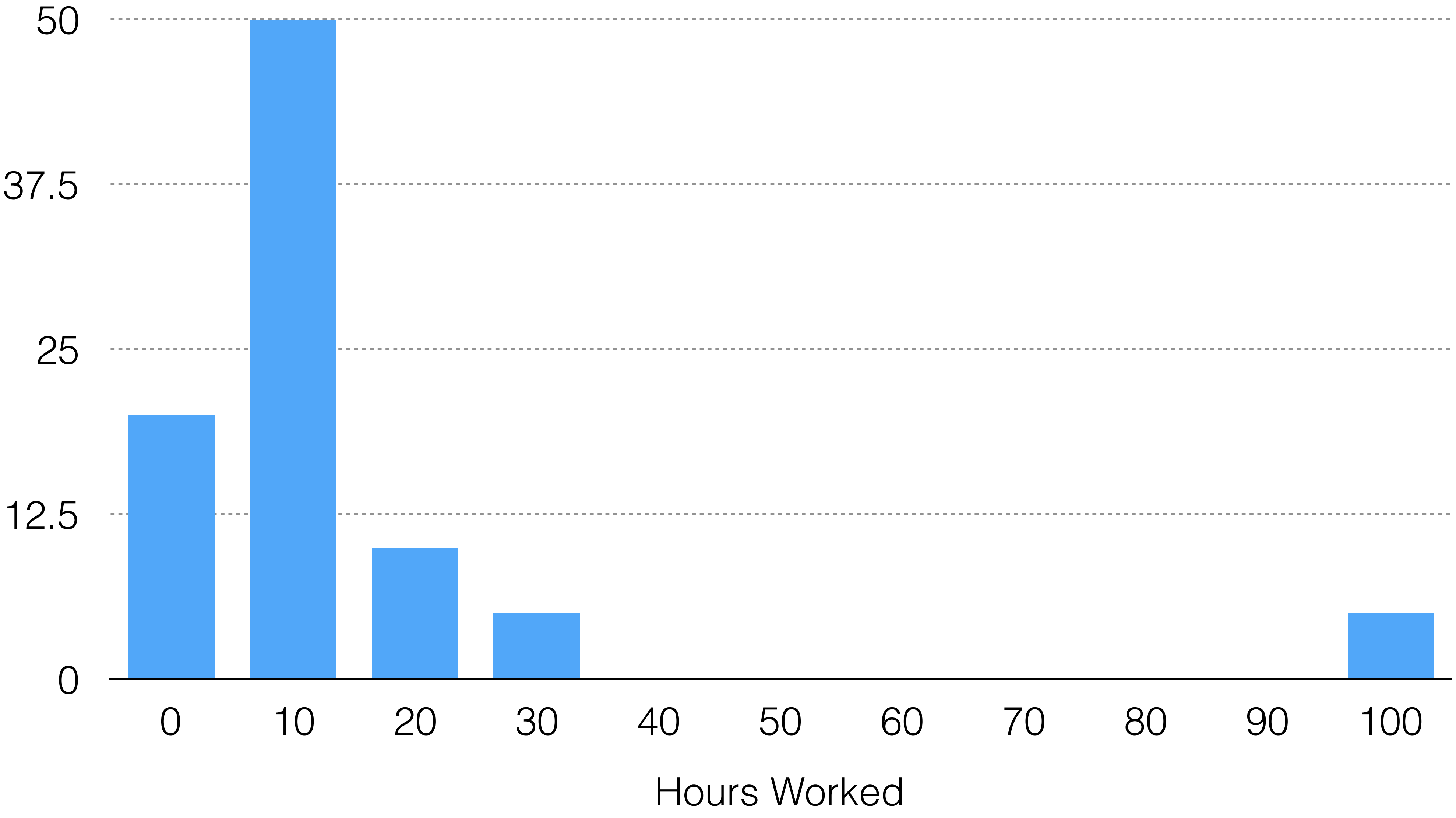
```
SELECT (
FROM PE(
GROUP B
ORDER B
```



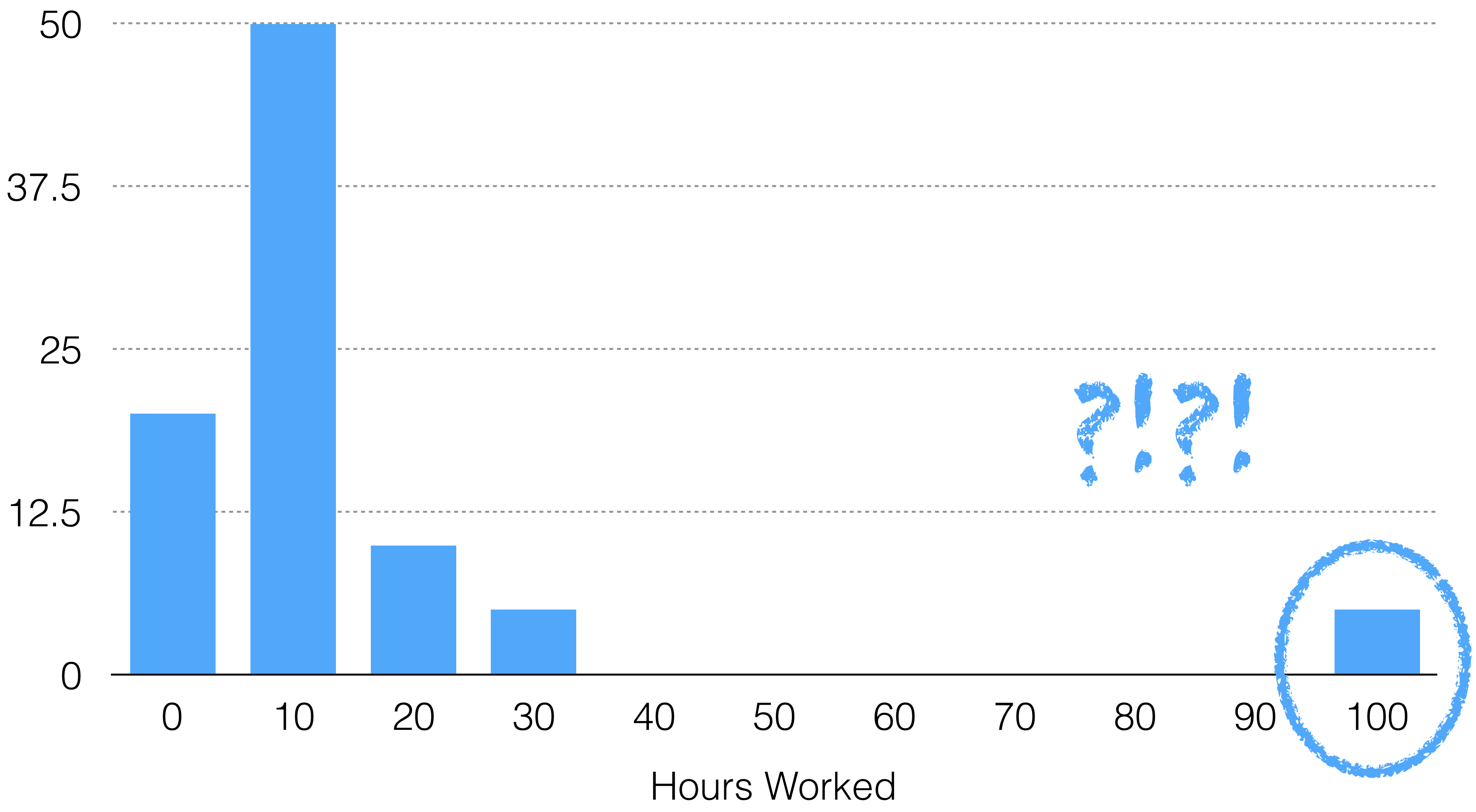
Schenectady, NY 12345

City	Count(*)
12345	2,500
10001	2,2000
90001	1,900
75001	1,400

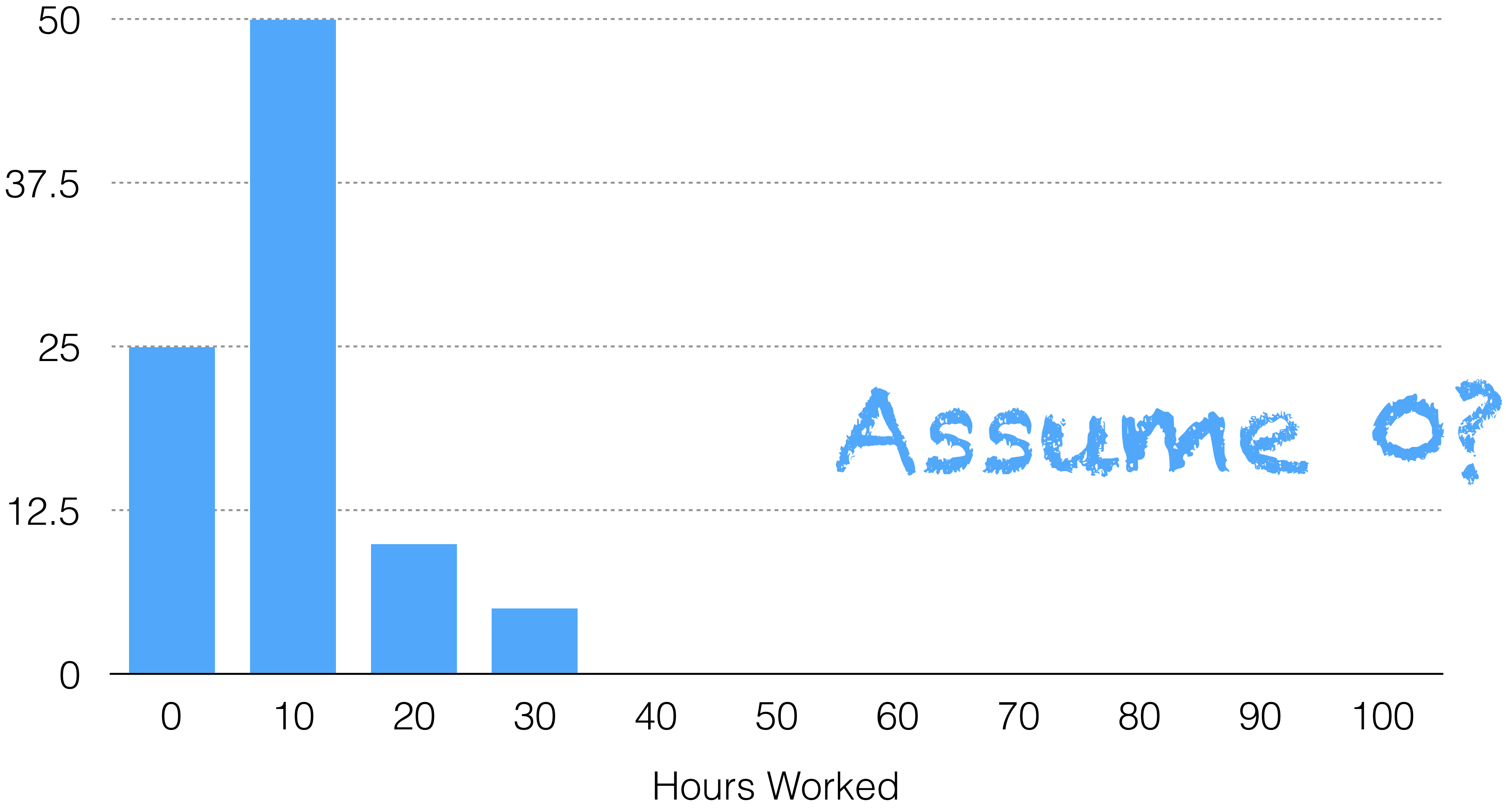
Set Defaults/Remove Outliers



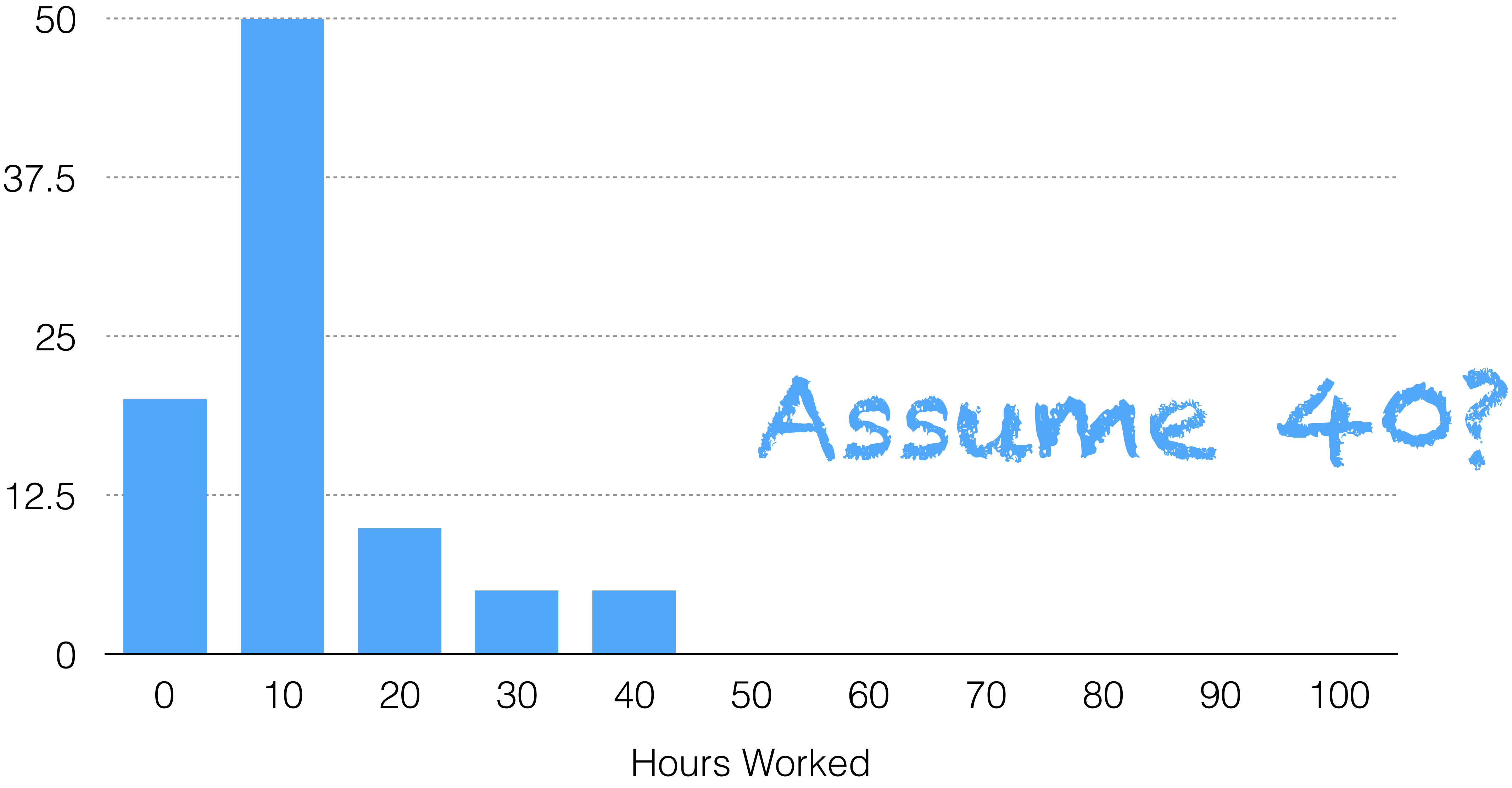
Set Defaults/Remove Outliers



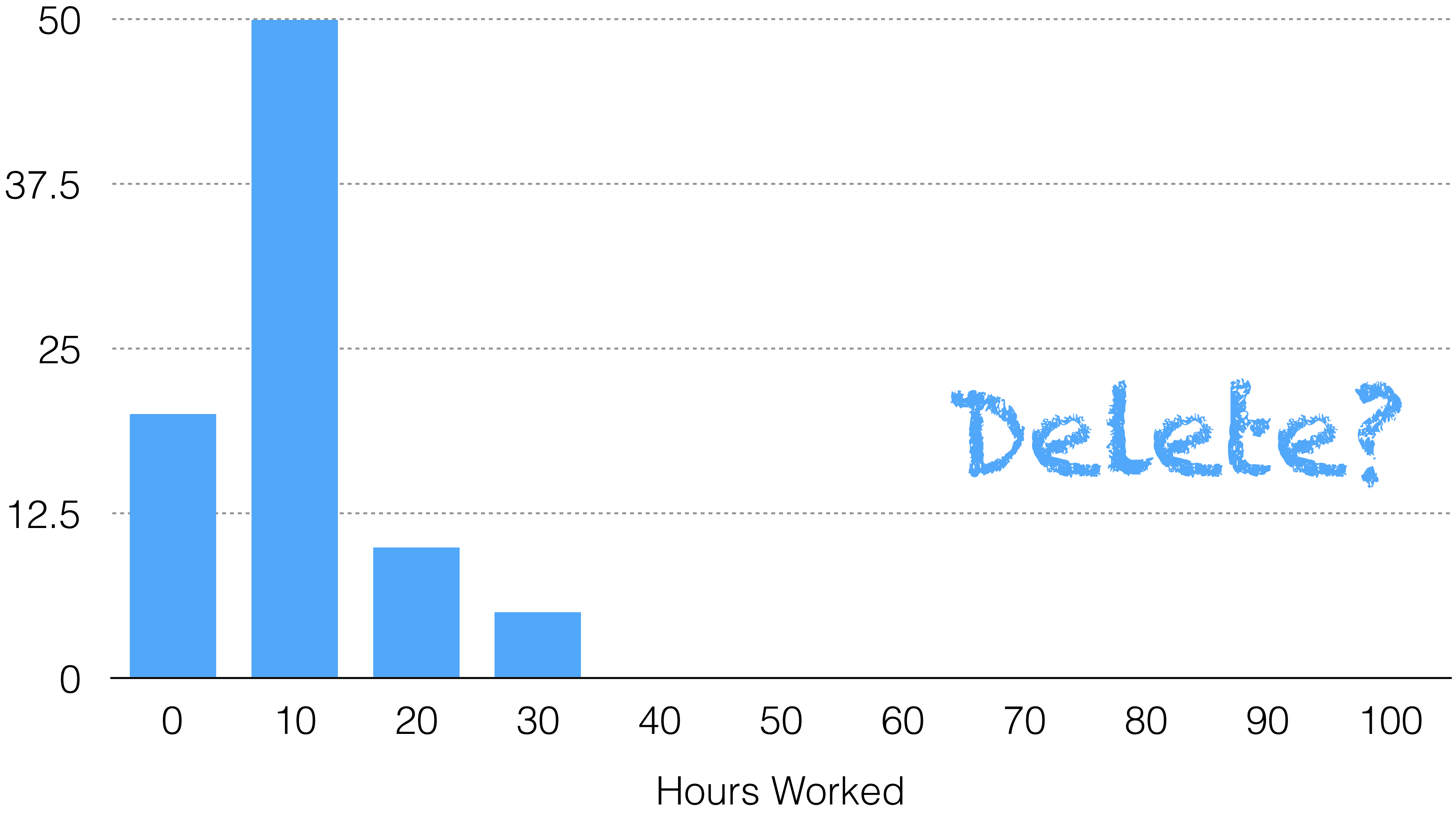
Set Defaults/Remove Outliers



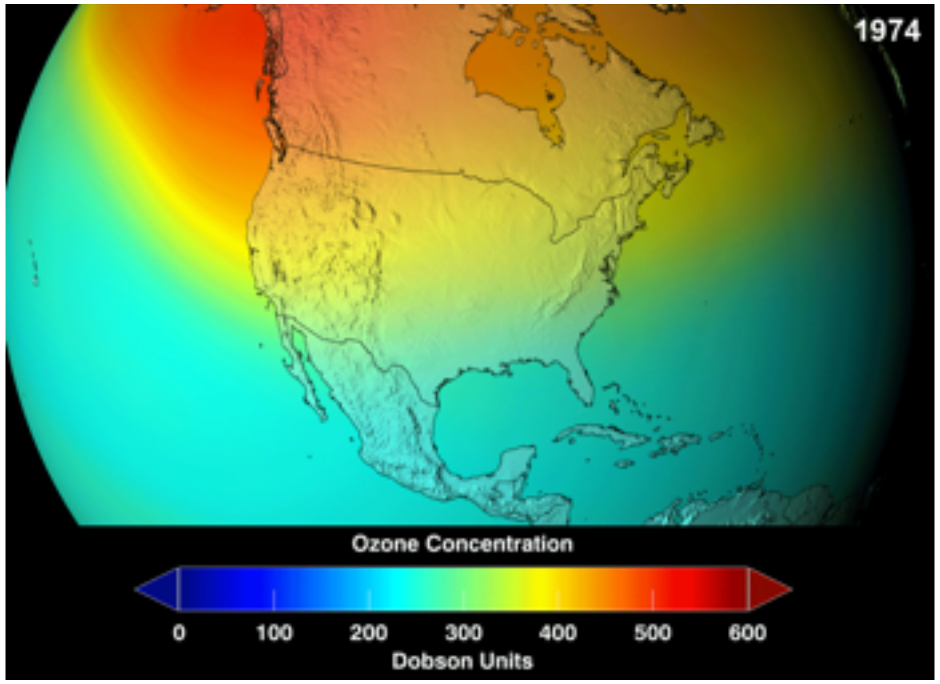
Set Defaults/Remove Outliers



Set Defaults/Remove Outliers



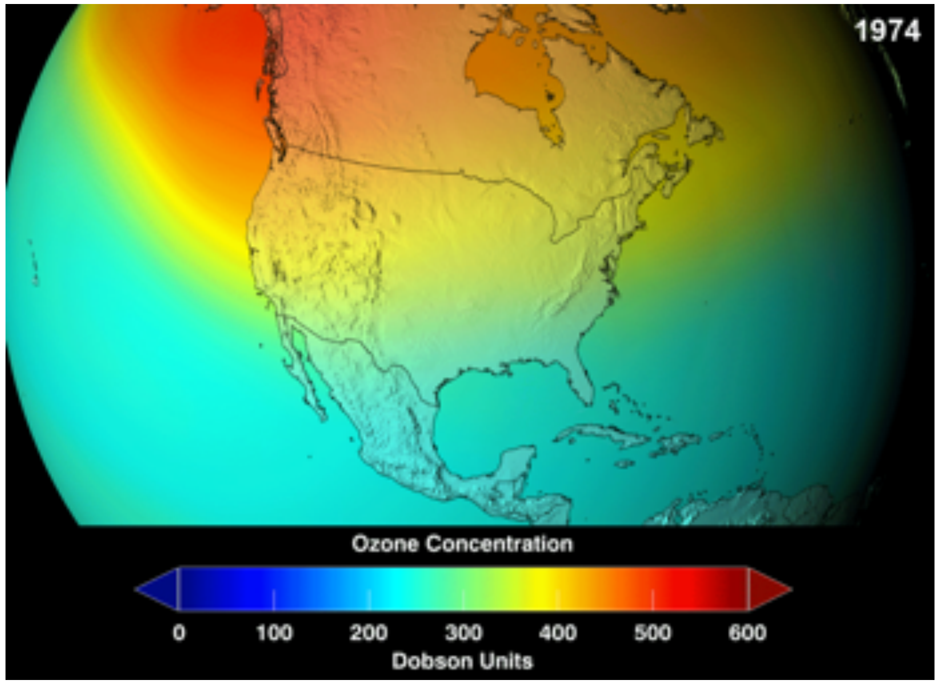
Set Defaults/Remove Outliers



The discovery of the Antarctic "ozone hole" by British Antarctic Survey scientists Farman, Gardiner and Shanklin…came as a shock to the scientific community…[The data] were initially rejected as unreasonable by data quality control algorithms (they were filtered out as errors since the values were unexpectedly low); the ozone hole was detected only in satellite data when the raw data was reprocessed following evidence of ozone depletion in in situ observations. When the software was rerun without the flags, the ozone hole was seen as far back as 1976.

https://en.wikipedia.org/wiki/Ozone_depletion#Antarctic_ozone_hole

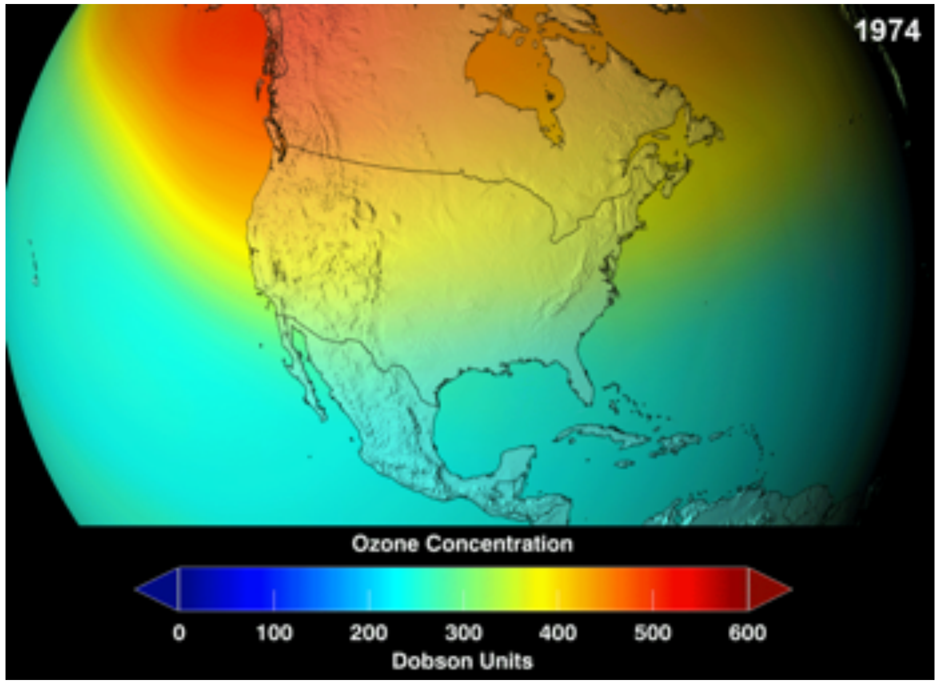
Set Defaults/Remove Outliers



The discovery of the Antarctic "ozone hole" by British Antarctic Survey scientists Farman, Gardiner and Shanklin…came as a shock to the scientific community…[The data] were initially rejected as unreasonable by data quality control algorithms (they were filtered out as errors since the values were unexpectedly low); the ozone hole was detected only in satellite data when the raw data was reprocessed following evidence of ozone depletion in in situ observations. When the software was rerun without the flags, the ozone hole was seen as far back as 1976.

https://en.wikipedia.org/wiki/Ozone_depletion#Antarctic_ozone_hole

Set Defaults/Remove Outliers



Always always always! Look at the data!

The discovery of the Antarctic "ozone hole" by British Antarctic Survey scientists Farman, Gardiner and Shanklin…came as a shock to the scientific community…[The data] were initially rejected as unreasonable by data quality control algorithms (they were filtered out as errors since the values were unexpectedly low); the ozone hole was detected only in satellite data when the raw data was reprocessed following evidence of ozone depletion in in situ observations. When the software was rerun without the flags, the ozone hole was seen as far back as 1976.

https://en.wikipedia.org/wiki/Ozone_depletion#Antarctic_ozone_hole

String Similarity

String Similarity: Edit Distance

Minimal number of edits (inserts, deletes, substitutions) needed to transform A into B.

https://en.wikipedia.org/wiki/Levenshtein_distance

String Similarity: Edit Distance

$$
\begin{aligned}
d_{i0} &= \sum_{k=1}^{i} w_{\mathrm{del}}(b_k), & \text{for } 1 \leq i \leq m \\
d_{0j} &= \sum_{k=1}^{j} w_{\mathrm{ins}}(a_k), & \text{for } 1 \leq j \leq n \\
d_{ij} &= \begin{cases} d_{i-1,j-1} & \text{for } a_j = b_i \\ \min \begin{cases} d_{i-1,j} + w_{\mathrm{del}}(b_i) \\ d_{i,j-1} + w_{\mathrm{ins}}(a_j) \\ d_{i-1,j-1} + w_{\mathrm{sub}}(a_j, b_i) \end{cases} & \text{for } a_j \neq b_i \end{cases} & \text{for } 1 \leq i \leq m, 1 \leq j \leq n.
\end{aligned}
$$

https://en.wikipedia.org/wiki/Levenshtein_distance

String Similarity: Edit Distance

115th Waterman St., Providence, RI

110th Waterman St., Providence, RI

EditDistance = 1

String Similarity: Edit Distance

Waterman Street, Providence, RI

Waterman St, Providence, RI

EditDistance = 4

String Similarity: Edit Distance

Problems?

String Similarity:
Edit Distance

148th Ave NE, Redmond, WA

148th Ave NE, Redmond, WA

String Similarity: Edit Distance

Edit Distance = 0

148th Ave NE, Redmond, WA

148th Ave NE, Redmond, WA

String Similarity: Edit Distance

Edit Distance = 0

148th Ave NE, Redmond, WA

148th Ave NE, Redmond, WA

148th Ave NE, Redmond, WA

NE 148th Ave, Redmond, WA

String Similarity: Edit Distance

Edit Distance = 0

148th Ave NE, Redmond, WA

148th Ave NE, Redmond, WA

148th Ave NE, Redmond, WA

NE 148th Ave, Redmond, WA

Edit Distance = 4

String Similarity:
Jaccard Similarity

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

https://en.wikipedia.org/wiki/Jaccard_index

String Similarity: Jaccard Similarity

148th Ave NE, Redmond, WA

140th Ave NE, Redmond, WA

https://en.wikipedia.org/wiki/Jaccard_index

String Similarity: Jaccard Similarity

148th Ave NE, Redmond, WA

140th Ave NE, Redmond, WA

https://en.wikipedia.org/wiki/Jaccard_index

String Similarity: Jaccard Similarity

148th Ave NE, Redmond, WA

140th Ave NE, Redmond, WA

Jaccard = 4 / 6 = .67

https://en.wikipedia.org/wiki/Jaccard_index

String Similarity: Jaccard Similarity

148th Ave NE, Redmond, WA

NE 148th Ave, Redmond, WA

Jaccard = ???

https://en.wikipedia.org/wiki/Jaccard_index

String Similarity: Jaccard Similarity

148th Ave NE, Redmond, WA

NE 148th Ave, Redmond, WA

Jaccard = 1

https://en.wikipedia.org/wiki/Jaccard_index

Clicker Question!

Clicker Question!

iPad Two 16GB WiFi White

iPad 2nd generation 16GB WiFi White

What's the Jaccard Similarity?

(a) 3/8

(b) 4/11

(c) 4/7

Clicker Question!

iPad Two 16GB WiFi White

iPad 2nd generation 16GB WiFi White

What's the Jaccard Similarity?

(a) 3/8

(b) 4/11

[✓ (c)] 4/7

$$\frac{\#(\text{iPad, 16GB, Wifi, White})}{\#(\text{iPad, Two, 2nd, generation, 16GB, Wifi, White})}$$

String Similarity: Jaccard Similarity

Michigan State University

Michigan State Univ.

Michigan State University

Ohio State University

https://en.wikipedia.org/wiki/Jaccard_index

String Similarity:
Jaccard Similarity

Jaccard = 0.5

Michigan State University

Michigan State Univ.

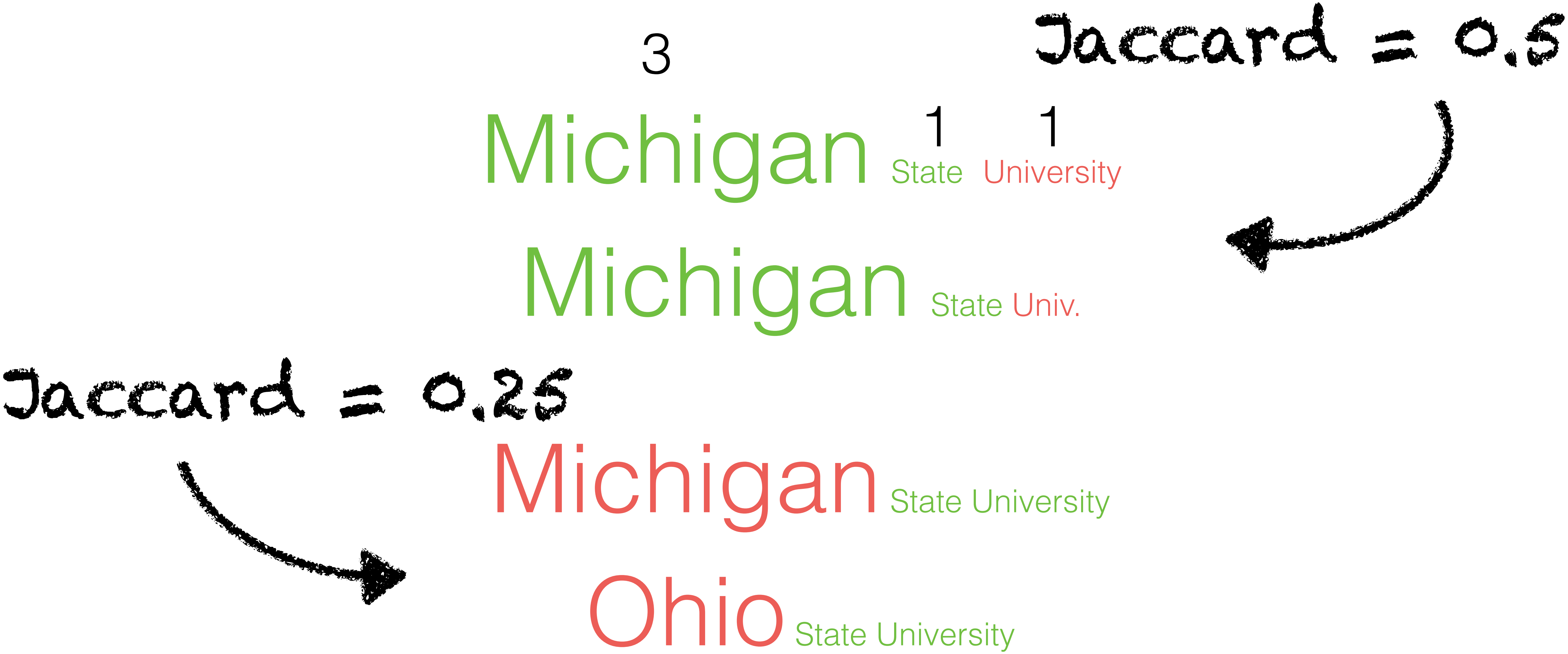
Jaccard = 0.5

Michigan State University

Ohio State University

https://en.wikipedia.org/wiki/Jaccard_index

String Similarity:
(Weighted) Jaccard Similarity



Jaccard = 0.5

3 1 1

Michigan State University

Michigan State Univ.

Jaccard = 0.25

Michigan State University

Ohio State University

https://en.wikipedia.org/wiki/Jaccard_index

String Similarity: (Weighted) Jaccard Similarity

3 1 1

Jaccard = 0.5

Michigan State University

Michigan State Univ.

Jaccard = 0.5

Michigan State University

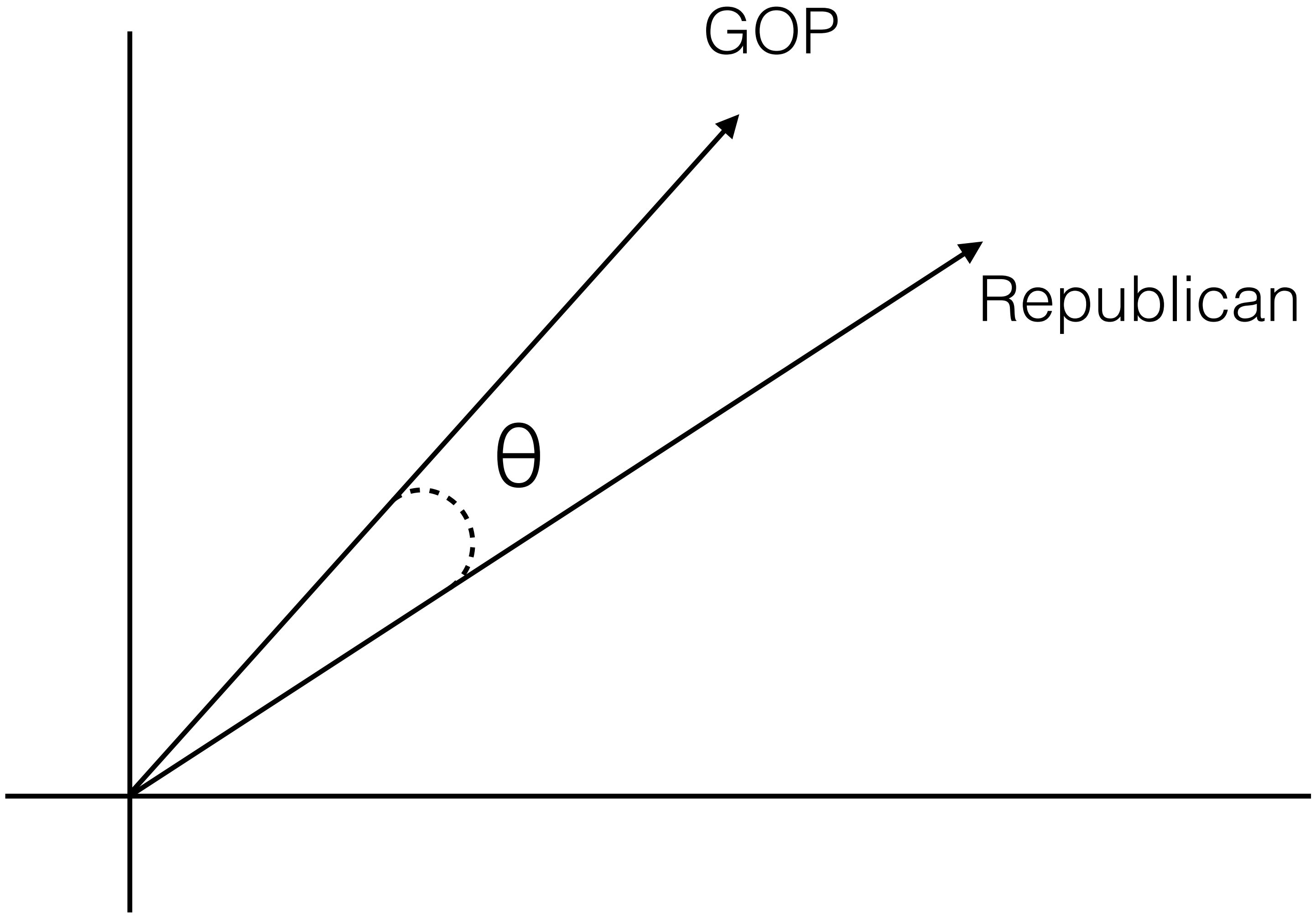
University of Michigan

https://en.wikipedia.org/wiki/Jaccard_index

String Similarity: Cosine Similarity

	Senator	Washington	announced	party	primary	chairman
GOP	1002	41	502	700	400	3
Republican	800	35	521	698	423	10

String Similarity: Cosine Similarity



Clicker Question!

Clicker Question!

Brown
Brown Uni.

Which metric would (likely) consider the above words more similar?

(a) Jaccard
(b) Cosine

Clicker Question!

Brown
Brown Uni.

Which metric would (likely) consider the above words more similar?

[✓ (a)] Jaccard
(b) Cosine

Clicker Question!

Motown
Detroit

Which metric would (likely) consider the above words more similar?

(a) Jaccard
(b) Cosine

Clicker Question!

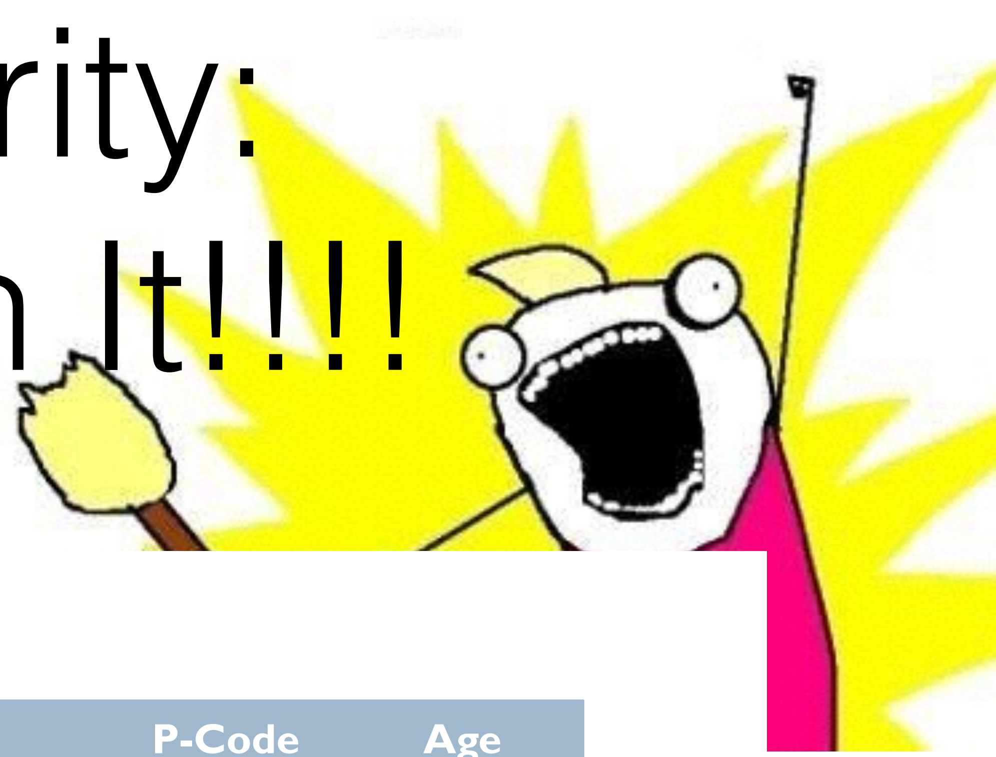
Motown
Detroit

Which metric would (likely) consider the above words more similar?

(a) Jaccard
[✓ (b)] Cosine

String Similarity: Machine Learn It!!!!

String Similarity: Machine Learn It!!!!

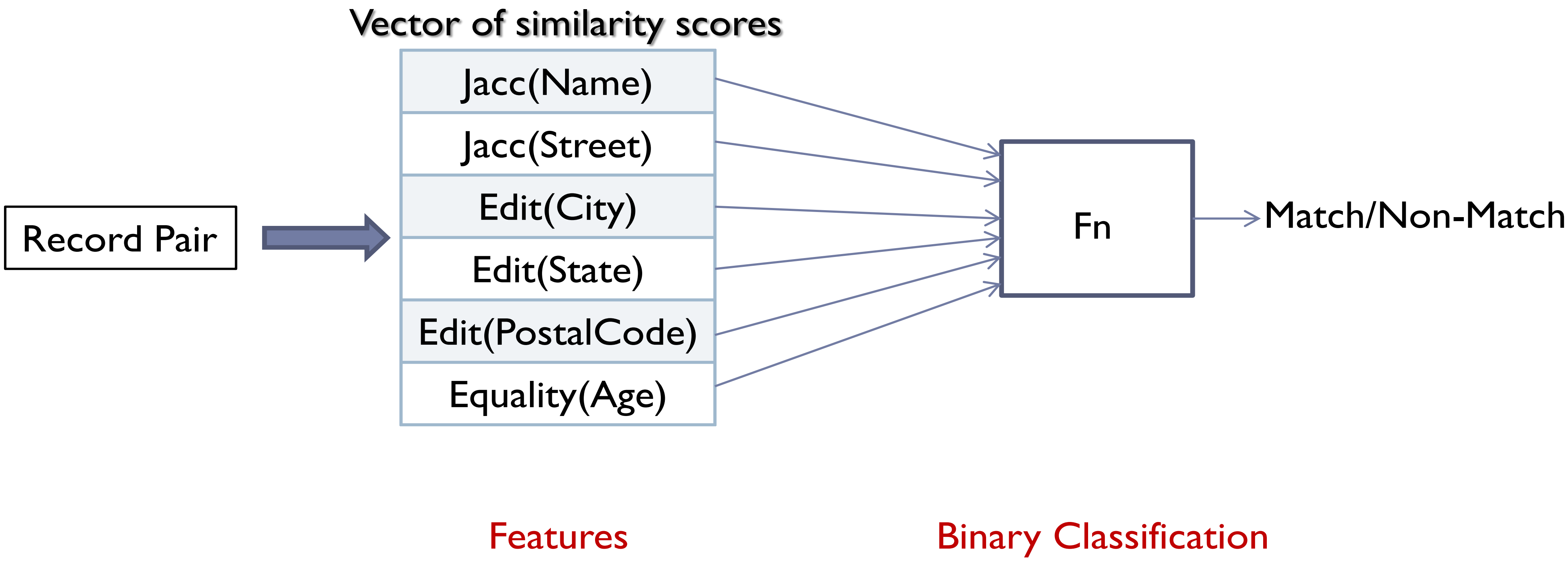


Customer

Id	Name	Street	City	State	P-Code	Age
1	J Smith	123 University Ave	Seattle	Washington	98106	42
2	Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
3	Bob Wilson	345 Broadway	Seattle	Washington	98101	19
4	M Jones	245 Third Street	Redmond	NULL	98052	299
5	Robert Wilson	345 Broadway St	Seattle	WA	98101	19
6	James Smith	123 Univ Ave	Seatle	WA	NULL	41
7	J Widom	123 University Ave	Palo Alto	CA	94305	NULL
…	…	…	…	…	…	…

$WtJaccard =$ 0.57 0.91 1.0 0.0 1.0 1.0

String Similarity: Machine Learn It!!!!



String Similarity: Machine Learn It!!!!

Bob Wilson	345 Broadway	Seattle	Washington	98101	19
Robert Wilson	345 Broadway St	Seattle	WA	98101	19

Match

B Wilson	123 Broadway	Boise	Idaho	83712	19
Robert Wilson	345 Broadway St	Seattle	WA	98101	19

Non-Match

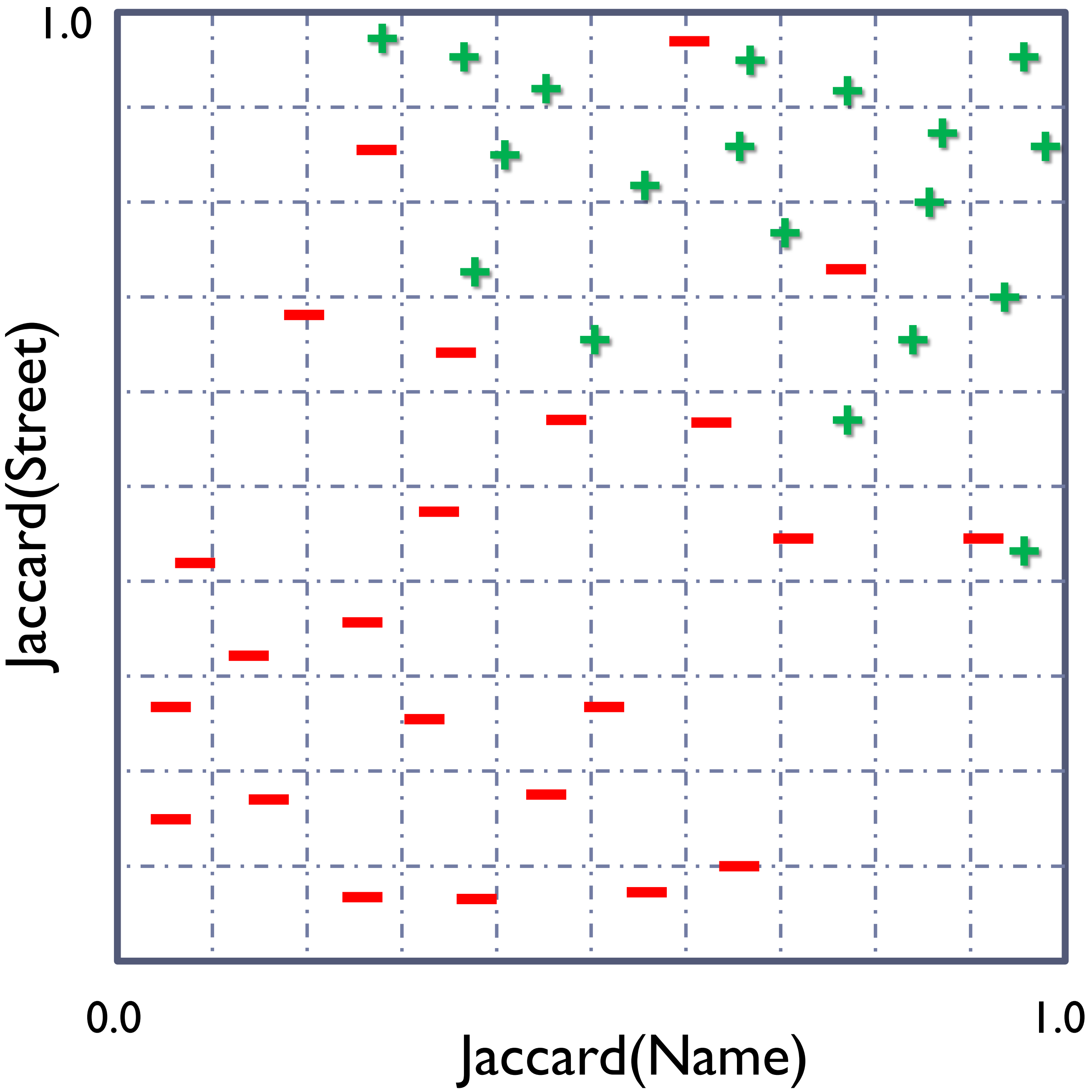
Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
M Jones	245 Third Street	Redmond	NULL	98052	299

Match

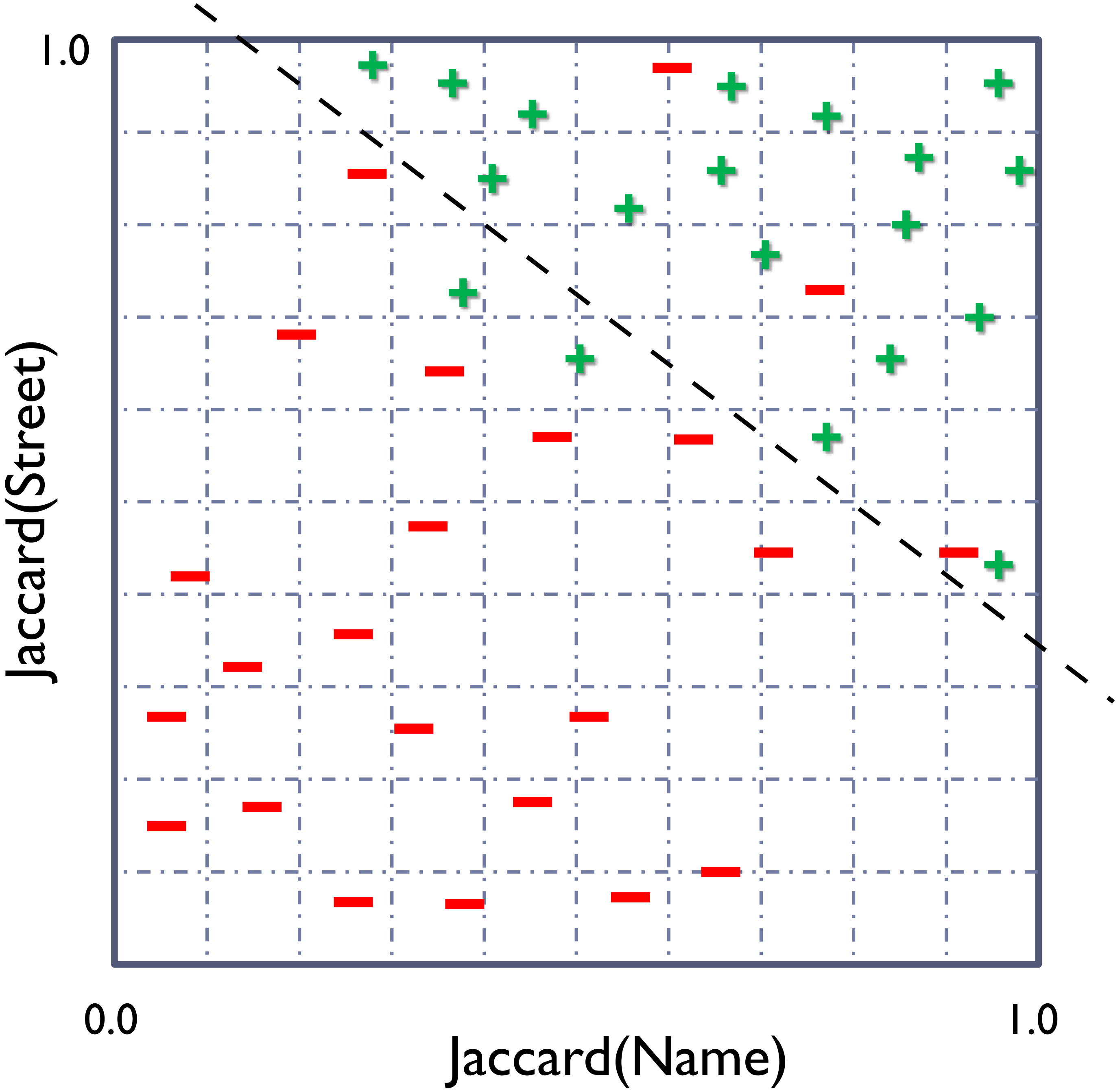
Mary Jones	245 3rd St	Redmond	WA	98052-1234	30
Robert Wilson	345 Broadway St	Seattle	WA	98101	19

Non-Match

String Similarity: Machine Learn It!!!!



String Similarity: Machine Learn It!!!!



And now….a word from your HTAs

(Meanwhile: I HAVE TO GO I’M GONNA MISS MY TRAIN EMAIL ME YOUR QUESTIONS HAVE A GOOD WEEKEND BYEEEEE)