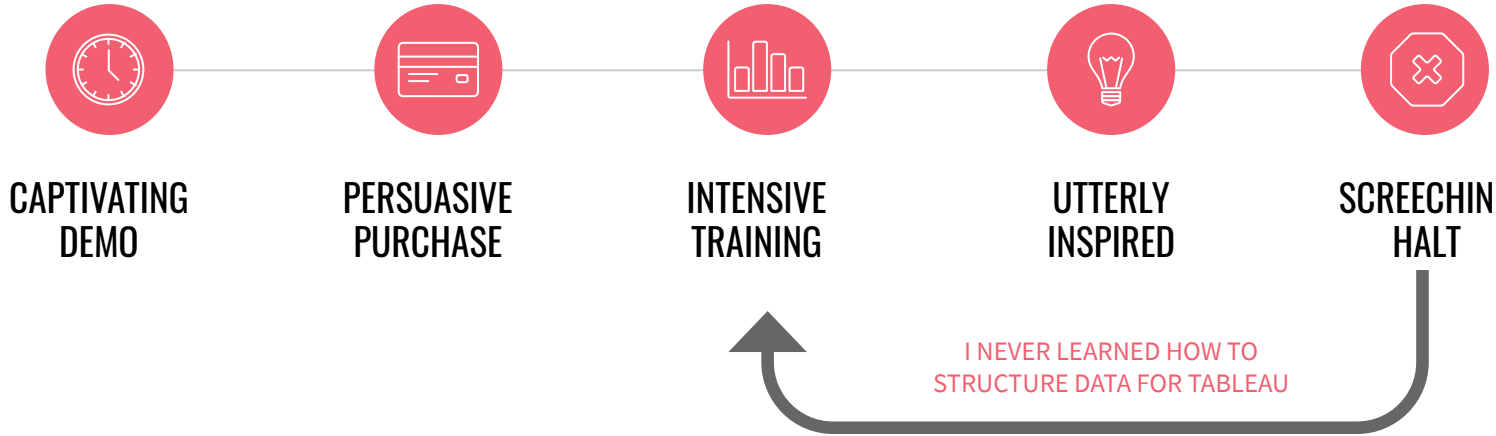
ERIN WALDRON FROM DATA DOZEN • SEPTEMBER 1, 2020 • KANSAS CITY TABLEAU USER GROUP

MAKING TIDY DATA

HOW TO STRUCTURE A DATA SET FOR TABLEAU

ONCE UPON A TIME

Here's how my Tableau fairytale started out.



SCREECHING HALT

Many analysts don't know
how data is structured behind the scenes.

SCREECHING HALT

Tableau training rarely includes any data structure knowledge.

THE DATA STRUCTURE I HAD

STUDENT ID	2014 FALL FUNDING TYPE	2014 FALL FUNDING AMOUNT	2015 SPRING FUNDING TYPE	2015 SPRING FUNDING AMOUNT
1	TEACHING ASSISTANTSHIP	$10,000	TEACHING ASSISTANTSHIP	$10,000
2	TEACHING ASSISTANTSHIP	$10,000	RESEARCH ASSISTANTSHIP	$10,500
3	UNIVERSITY FELLOWSHIP	$10,000	UNIVERSITY FELLOWSHIP	$10,000

THE DATA STRUCTURE I NEEDED
(AND DIDN'T KNOW HOW TO ASK FOR)

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000
2	201405	TEACHING ASSISTANTSHIP	$10,000
2	201502	RESEARCH ASSISTANTSHIP	$10,500
3	201405	UNIVERSITY FELLOWSHIP	$10,000
3	201502	UNIVERSITY FELLOWSHIP	$10,000

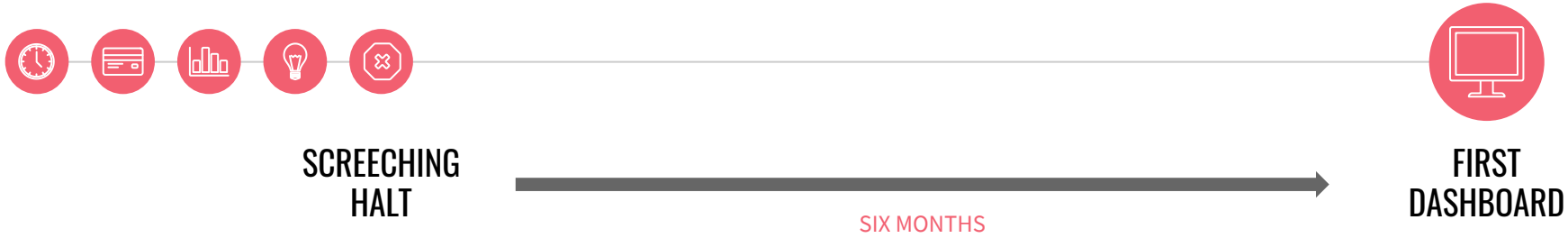
SCREECHING HALT

This is exactly what we had requested.

And it was the only way I knew how to access the data.

ONCE UPON A TIME

Here's how long my Tableau stalemate lasted.



WHAT WE'LL DO TODAY

1

2

3

4

WHAT WE'LL DO TODAY

1. **Tidy Data: Three Features of a Tableau Ready Data Set**
2.
3.
4.

WHAT WE'LL DO TODAY

1 Tidy Data: Three Features of a Tableau Ready Data Set

2 **Common Culprits: Problem, Telltale Sign, Solution**

3

4

WHAT WE'LL DO TODAY

1. Tidy Data: Three Features of a Tableau Ready Data Set
2. Common Culprits: Problem, Telltale Sign, Solution
3. **Tableau Demo: Restructuring a Data Set for Tableau**
4.

WHAT WE'LL DO TODAY

1. Tidy Data: Three Features of a Tableau Ready Data Set
2. Common Culprits: Problem, Telltale Sign, Solution
3. Tableau Demo: Restructuring a Data Set for Tableau
4. **Q&A**

LISA GITELMAN

Raw data *is an oxymoron.*

MESSY DATA

There are two ways in which data can be messy:

CONTENT

WHAT'S WRITTEN
INSIDE THE CELLS

STRUCTURE

HOW THOSE CELLS
ARE LAID OUT

MESSY DATA

We want to be able to compare and aggregate.

CONTENT

WHAT'S WRITTEN
INSIDE THE CELLS

- We're inconsistent.
  St. Louis, Saint Louis, STL
- We make errors.
  Misspellings, wrong field, added an extra 0
- We don't have all the data.
  Partial or empty cells
- The world is ambiguous.
  World Bank vs United Nations country list
- And things change.
  And this makes consistency hard.

MESSY DATA

We want the computer to be able to see what we see.

STUDENT ID	SEMESTER	FUNDING AMOUNT
1	201405	$10,000
1	201502	$10,000
2	201405	$10,000
2	201502	$10,500

STRUCTURE

HOW THOSE CELLS ARE LAID OUT

TIDY DATA

"Happy families are all alike; every unhappy family is unhappy in its own way." -TOLSTOY

TIDY DATA

"Like families, tidy data sets are all alike, but every messy data set is messy in its own way." -HADLEY WICKHAM

HOW: SQL, TABLEAU PREP,
TABLEAU DESKTOP

TIDY DATA

The path to tidy data will depend on your organization's data, but the resulting data will always have the same structure.

RESULT: THE DATA SOURCE YOU
CONNECT TO TABLEAU

ERIN WALDRON
www.datadozen.com

TIDY DATA

In his article “Tidy Data,” Hadley Wickham identifies three elements that make a successful tabular data structure, which programs like R and Tableau expect.

ONE COLUMN PER VARIABLE

DIMENSIONS, MEASURES, FIELDS

ONE ROW PER OBSERVATION

RECORDS, TRANSACTIONS, EVENTS

ONE TABLE PER OBSERVATION TYPE

CONSISTENT GRANULARITY

TIDY DATA

Imagine a table in Excel.

TIDY DATA

One column per variable.

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT

TIDY DATA

One row per observation.

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000

TIDY DATA

One row per observation.

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000

TIDY DATA

One row per observation.

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000
2	201405	TEACHING ASSISTANTSHIP	$10,000

TIDY DATA

One row per observation.

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000
2	201405	TEACHING ASSISTANTSHIP	$10,000
2	201502	RESEARCH ASSISTANTSHIP	$10,500

TIDY DATA

One row per observation.

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000
2	201405	TEACHING ASSISTANTSHIP	$10,000
2	201502	RESEARCH ASSISTANTSHIP	$10,500
3	201405	UNIVERSITY FELLOWSHIP	$10,000

TIDY DATA

One row per observation.

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000
2	201405	TEACHING ASSISTANTSHIP	$10,000
2	201502	RESEARCH ASSISTANTSHIP	$10,500
3	201405	UNIVERSITY FELLOWSHIP	$10,000
3	201502	UNIVERSITY FELLOWSHIP	$10,000

TIDY DATA

One table per observational unit.

TABLE A

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000
2	201405	TEACHING ASSISTANTSHIP	$10,000
2	201502	RESEARCH ASSISTANTSHIP	$10,500
3	201405	UNIVERSITY FELLOWSHIP	$10,000
3	201502	UNIVERSITY FELLOWSHIP	$10,000

TABLE B

STUDENT ID	LAST NAME	DEPARTMENT
1	O'MALLEY	ANTHROPOLOGY
2	PERÉZ	ANTHROPOLOGY
3	SMITH	ENGLISH

COMMON CULPRITS

PROBLEM	TELL TALE SIGN	SOLUTION
The three most common data structure problems analysts often encounter.	How to identify the characteristics those structures have, and how Tableau reacts.	How we can restructure the data for analysis (or so that Tableau can see what we see.)

1) DATA IN THE HEADERS

PROBLEM

There are actually pieces of the data in the column headers instead of variable names in the column headers. This is the most common problem with handmade or archival data. (Think data assembled by hand in Excel.)

TELL TALE SIGN

The number of columns doesn't equal the number of variables in the data set. In Tableau, the data pane has all of the individual items of a variable, like year, listed out as individual data pills. (You want to make a line chart, but can't.)

SOLUTION

To start, simply identify and list the actual variables in the data set. Pivot the columns with data in the header. The number of columns will shrink to match the number of variables in the data set, and the number of rows will grow.

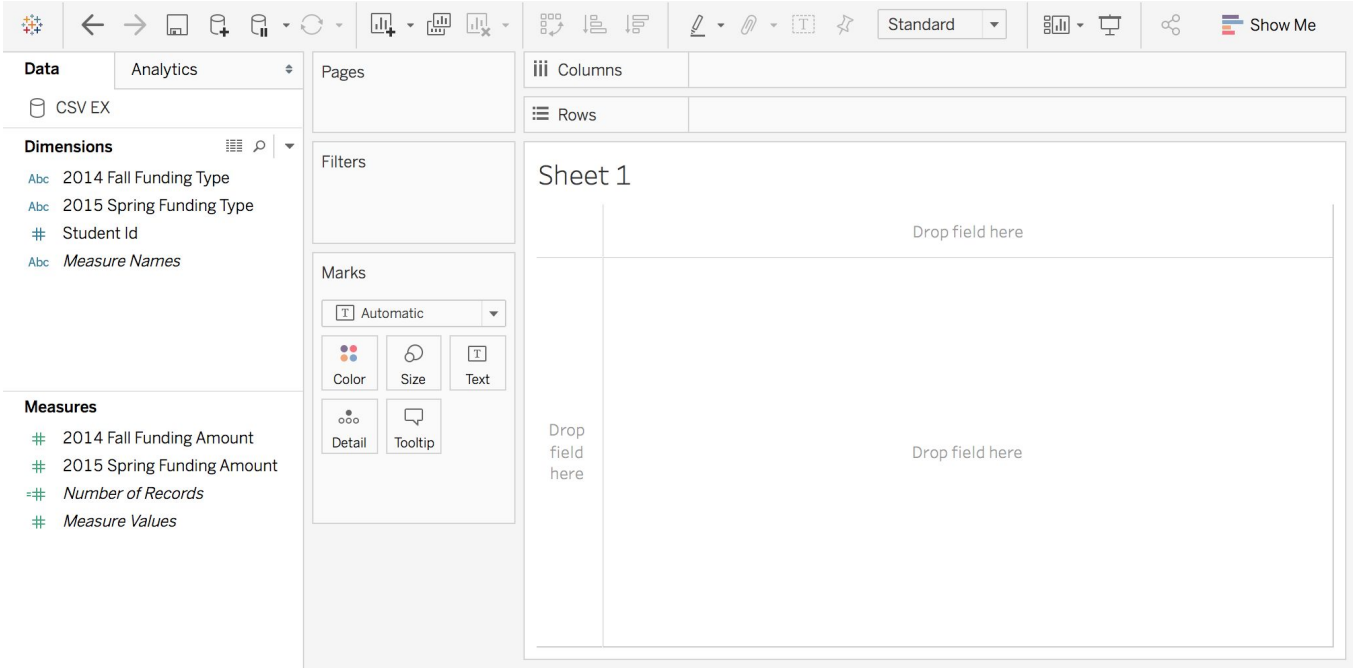
BEFORE PIVOT

1) DATA IN THE HEADERS

STUDENT ID	2014 FALL FUNDING TYPE	2014 FALL FUNDING AMOUNT	2015 SPRING FUNDING TYPE	2015 SPRING FUNDING AMOUNT
1	TEACHING ASSISTANTSHIP	$10,000	TEACHING ASSISTANTSHIP	$10,000
2	TEACHING ASSISTANTSHIP	$10,000	RESEARCH ASSISTANTSHIP	$10,500
3	UNIVERSITY FELLOWSHIP	$10,000	UNIVERSITY FELLOWSHIP	$10,000

BEFORE PIVOT

1) DATA IN THE HEADERS



AFTER PIVOT

1) DATA IN THE HEADERS

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000
2	201405	TEACHING ASSISTANTSHIP	$10,000
2	201502	RESEARCH ASSISTANTSHIP	$10,500
3	201405	UNIVERSITY FELLOWSHIP	$10,000
3	201502	UNIVERSITY FELLOWSHIP	$10,000

2) A COLUMN OF VARIABLES

PROBLEM

The analyst can see all of the variables they need when they look inside the "Column Name" or "Measures" data pill, but has trouble actually using them. This is common with exceedingly large or illogically combined data.

TELL TALE SIGN

In Tableau, we're constantly using the filter card to narrow down to just one variable, or we're creating a calculated field for every measure we want to bring into our view.

SOLUTION

Unpivot, or cast, these variables into their own individual columns. The number of columns should equal the number of variables. Now each variable can easily be used independently.

BEFORE UNPIVOT

2) A COLUMN OF VARIABLES

PROGRAM	MEASURE	VALUE
Anthropology	Minimum	4 years
Anthropology	Maximum	7.3 years
Art History	Minimum	3.5 years
Art History	Maximum	9 years

BEFORE UNPIVOT

2) A COLUMN OF VARIABLES

PROGRAM	MEASURE	VALUE	CALCULATED FIELD: MIN TIME TO DEGREE	CALCULATED FIELD: MAX TIME TO DEGREE
Anthropology	Minimum	4 years	4 years	
Anthropology	Maximum	7.3 years		7.3 years
Art History	Minimum	3.5 years	3.5 years	
Art History	Maximum	9 years		9 years

AFTER UNPIVOT

2) A COLUMN OF VARIABLES

PROGRAM	MIN TIME TO DEGREE	MAX TIME TO DEGREE
Anthropology	4 years	7.3 years
Art History	3.5 years	9 years

3) ROWS WITH DIFFERENT GRANULARITIES

PROBLEM

Subtotals and grand totals have been included as rows along with the granular data.

TELL TALE SIGN

When we sum up any of our data pills, we get multiples of what we should, and therefore we have to filter out the total rows each time. Often there is implied data in the totals variables that is unusable in Tableau.

SOLUTION

Remove the subtotals and grand totals. Add columns for any implied information that might be hidden in the subtotal labels.

BEFORE REMOVING

3) ROWS WITH DIFFERENT GRANULARITIES

DEPARTMENT	STUDENTS ENROLLED
ART HISTORY	15
ENGLISH	85
HUMANITIES DISCIPLINE TOTAL	100
POLITICAL SCIENCE	60
PSYCHOLOGY	90
SOCIAL SCIENCES DISCIPLINE TOTAL	150

BEFORE REMOVING

3) ROWS WITH DIFFERENT GRANULARITIES

DEPARTMENT	STUDENTS ENROLLED
ART HISTORY	15
ENGLISH	85
HUMANITIES DISCIPLINE TOTAL	100
POLITICAL SCIENCE	60
PSYCHOLOGY	90
SOCIAL SCIENCES DISCIPLINE TOTAL	150
sum(Students Enrolled)	500

AFTER REMOVING

3) ROWS WITH DIFFERENT GRANULARITIES

DISCIPLINE	DEPARTMENT	STUDENTS ENROLLED
HUMANITIES	ART HISTORY	15
HUMANITIES	ENGLISH	85
SOCIAL SCIENCE	POLITICAL SCIENCE	60
SOCIAL SCIENCE	PSYCHOLOGY	90

ENOUGH THEORY.

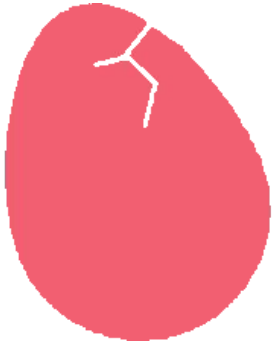
Let's build.

THANK YOU!

ERIN WALDRON

Owner of Data Dozen &
Data Visualization Specialist

www.datadozen.com
linkedin.com/in/egwaldron/
erin@datadozen.com
913-303-1125



DATA DOZEN

DATA VISUALIZATION EDUCATION