Finding Stories in Data

David Tarrant · @davetaz

The data percolator

Validating and cleaning data

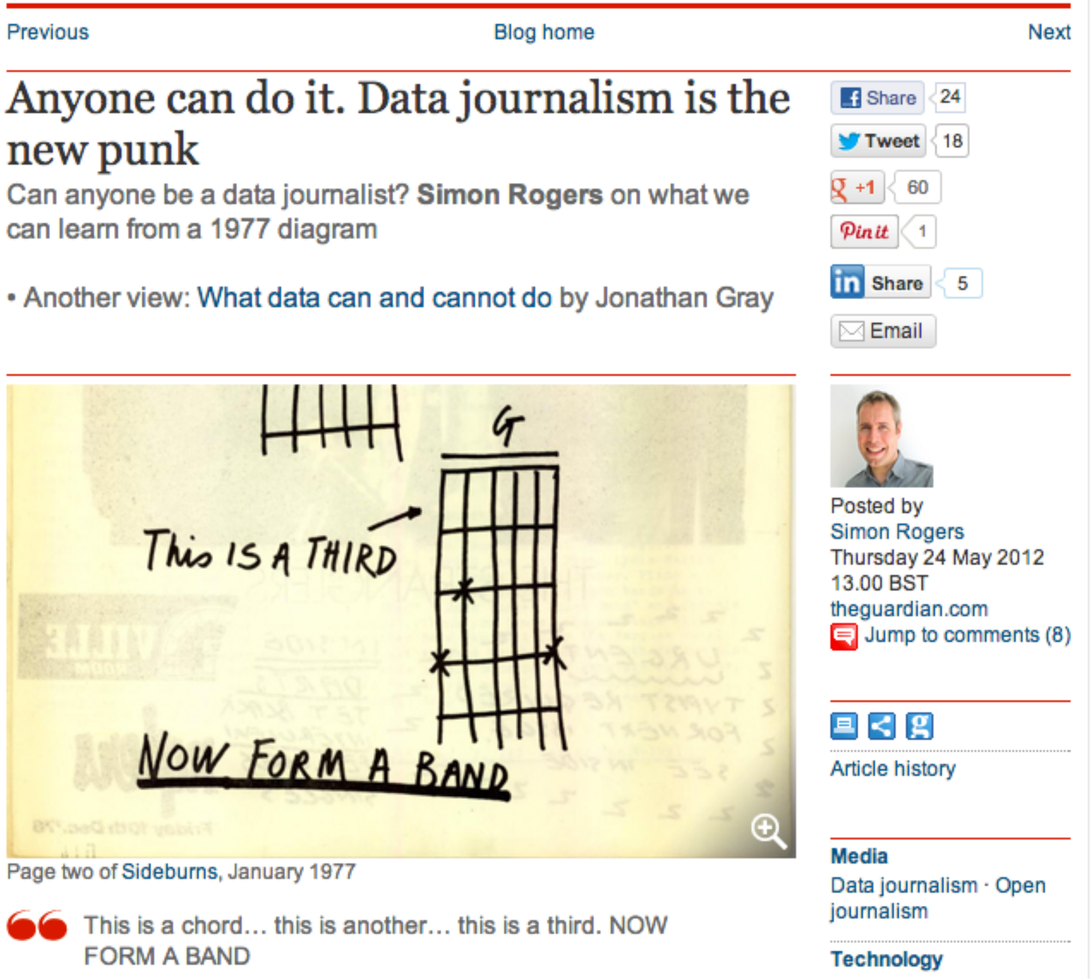
Inception points: cross filters, outliers and comparisons

The data exploration checklist

Session 2

Finding and exploring data

DATABLOG
Facts are sacred



Previous · Blog home · Next

Anyone can do it. Data journalism is the new punk

Can anyone be a data journalist? **Simon Rogers** on what we can learn from a 1977 diagram

• Another view: What data can and cannot do by Jonathan Gray

Page two of Sideburns, January 1977

Posted by
Simon Rogers
Thursday 24 May 2012
13.00 BST
theguardian.com
Jump to comments (8)

Article history

Media
Data journalism · Open journalism

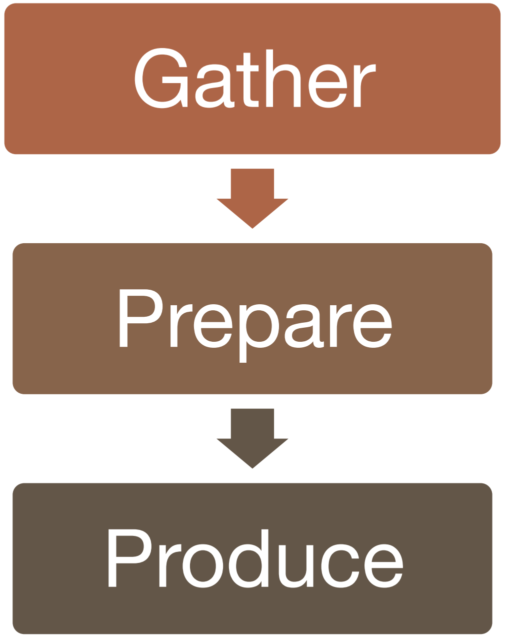
Technology

This is a chord… this is another… this is a third. NOW FORM A BAND

http://www.theguardian.com/news/datablog/2012/may/24/data-journalism-punk

The data percolator

Data percolation: A model of data preparation and analysis

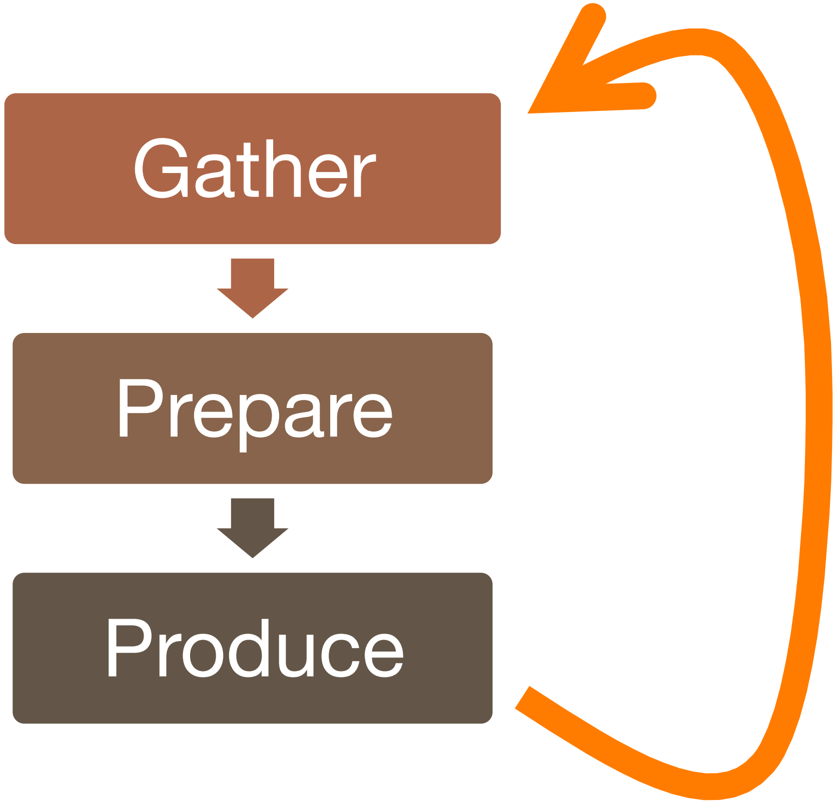

Gather

Prepare

Produce

See also: the Data Journalism Handbook

Data percolation: A model of data preparation and analysis

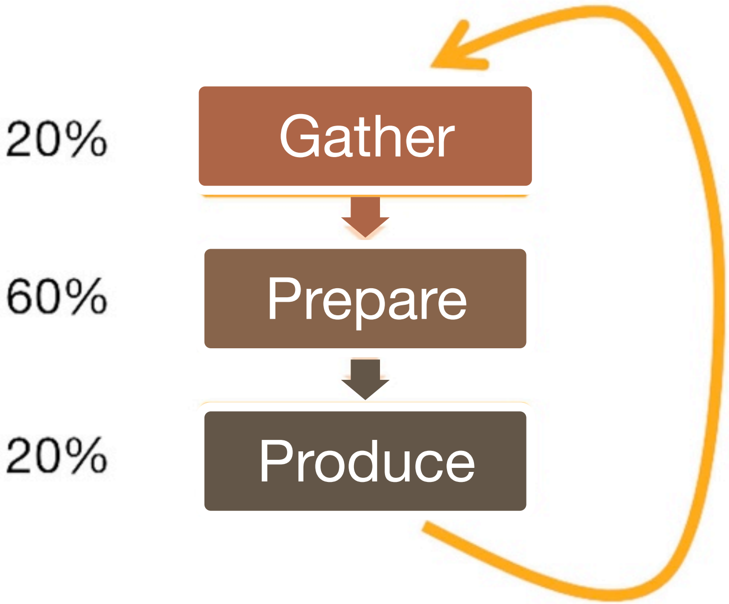


Gather

Prepare

Produce

How should I budget my time?



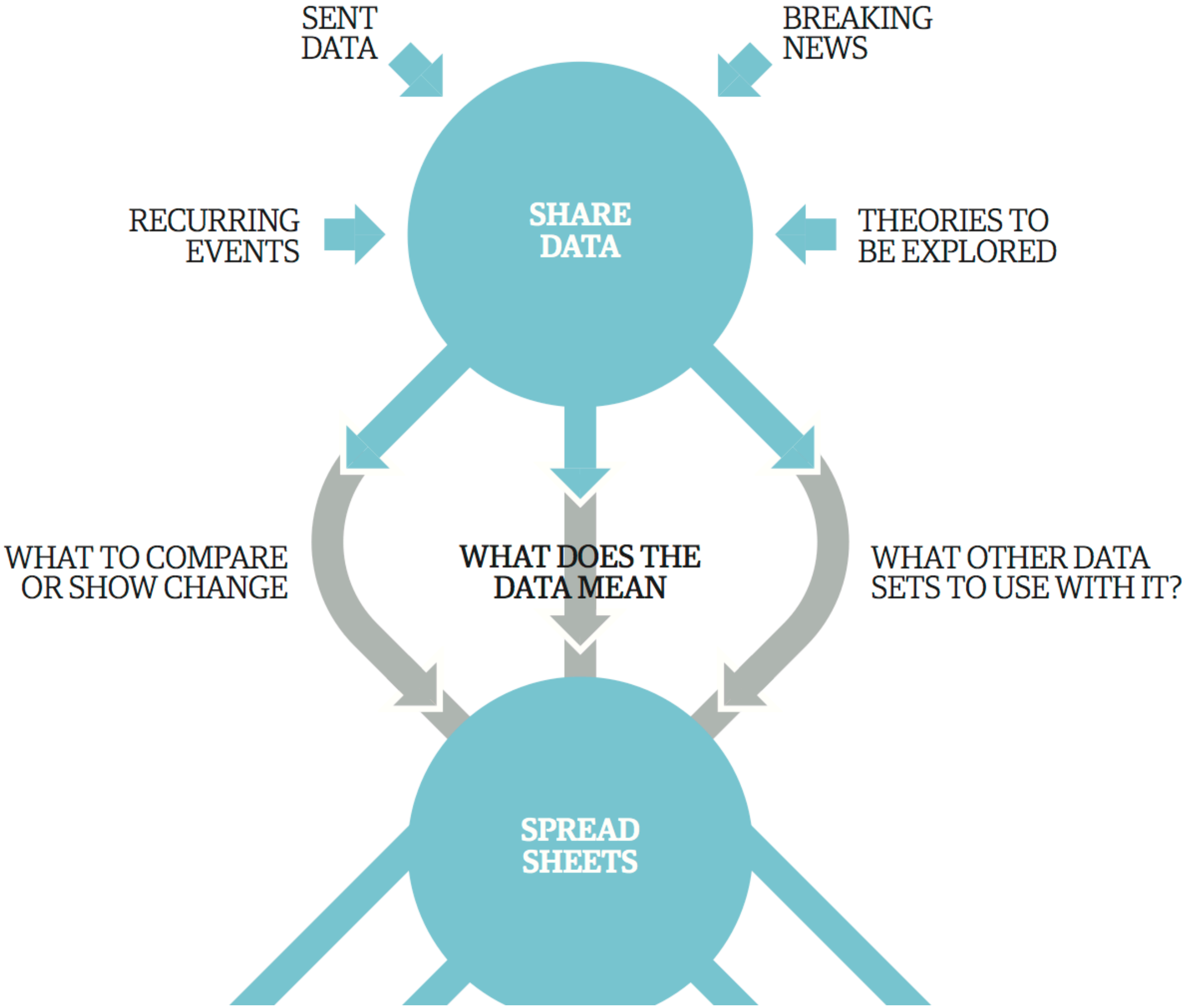
20% Gather

60% Prepare

20% Produce

Finding a story is a creative process.

Let it percolate!



SENT DATA

BREAKING NEWS

RECURRING EVENTS

SHARE DATA

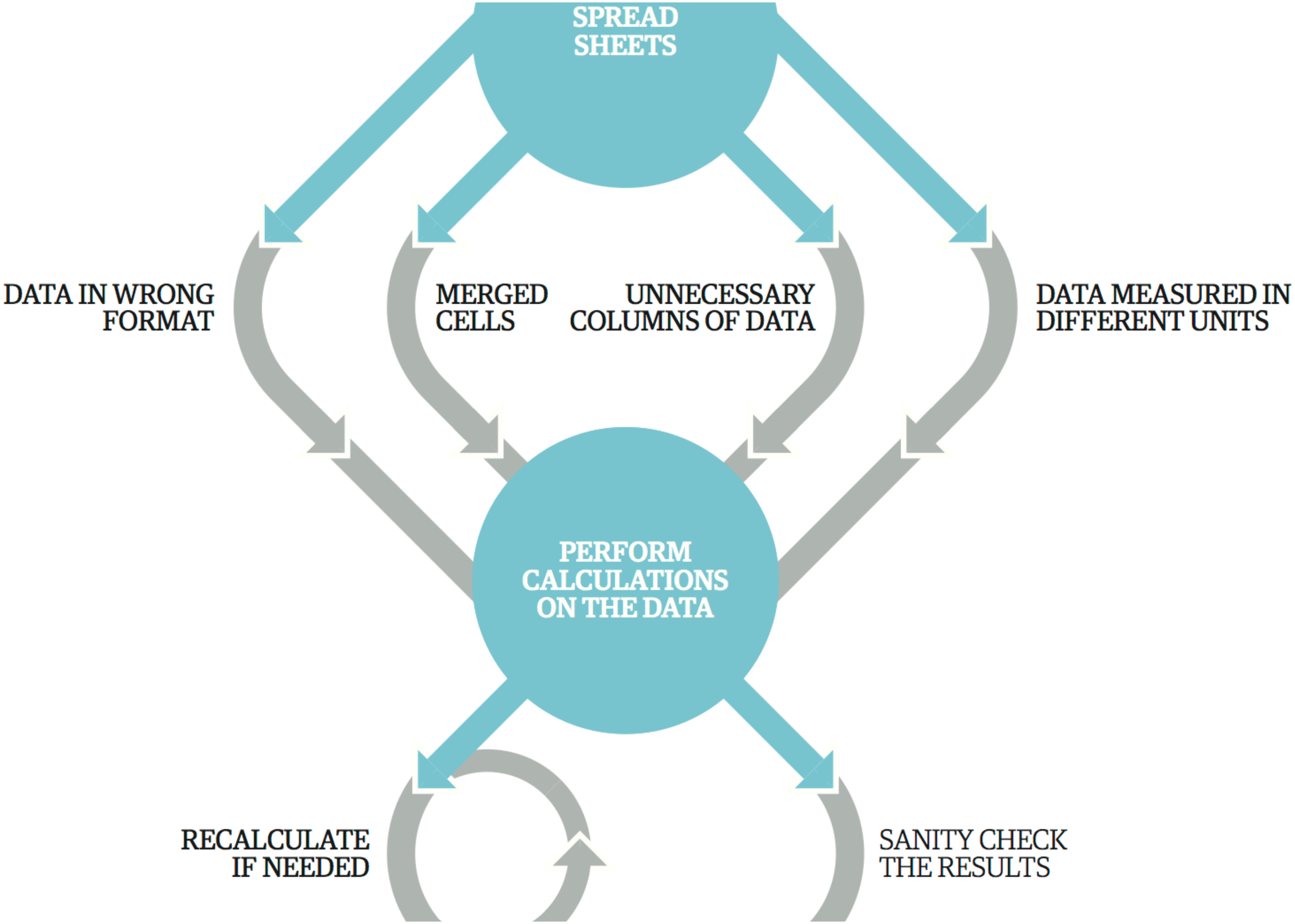
THEORIES TO BE EXPLORED

WHAT TO COMPARE OR SHOW CHANGE

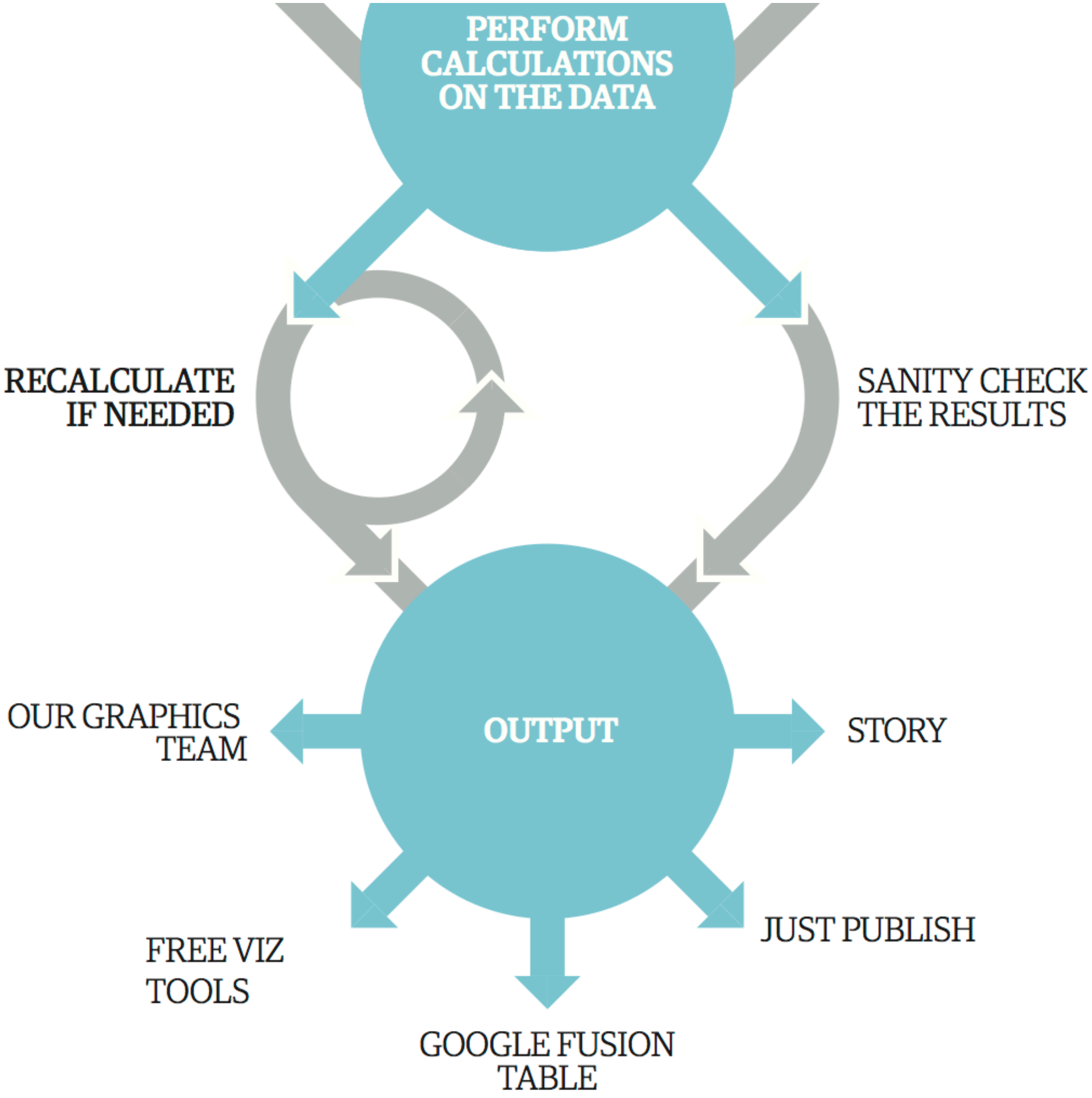
WHAT DOES THE DATA MEAN

WHAT OTHER DATA SETS TO USE WITH IT?

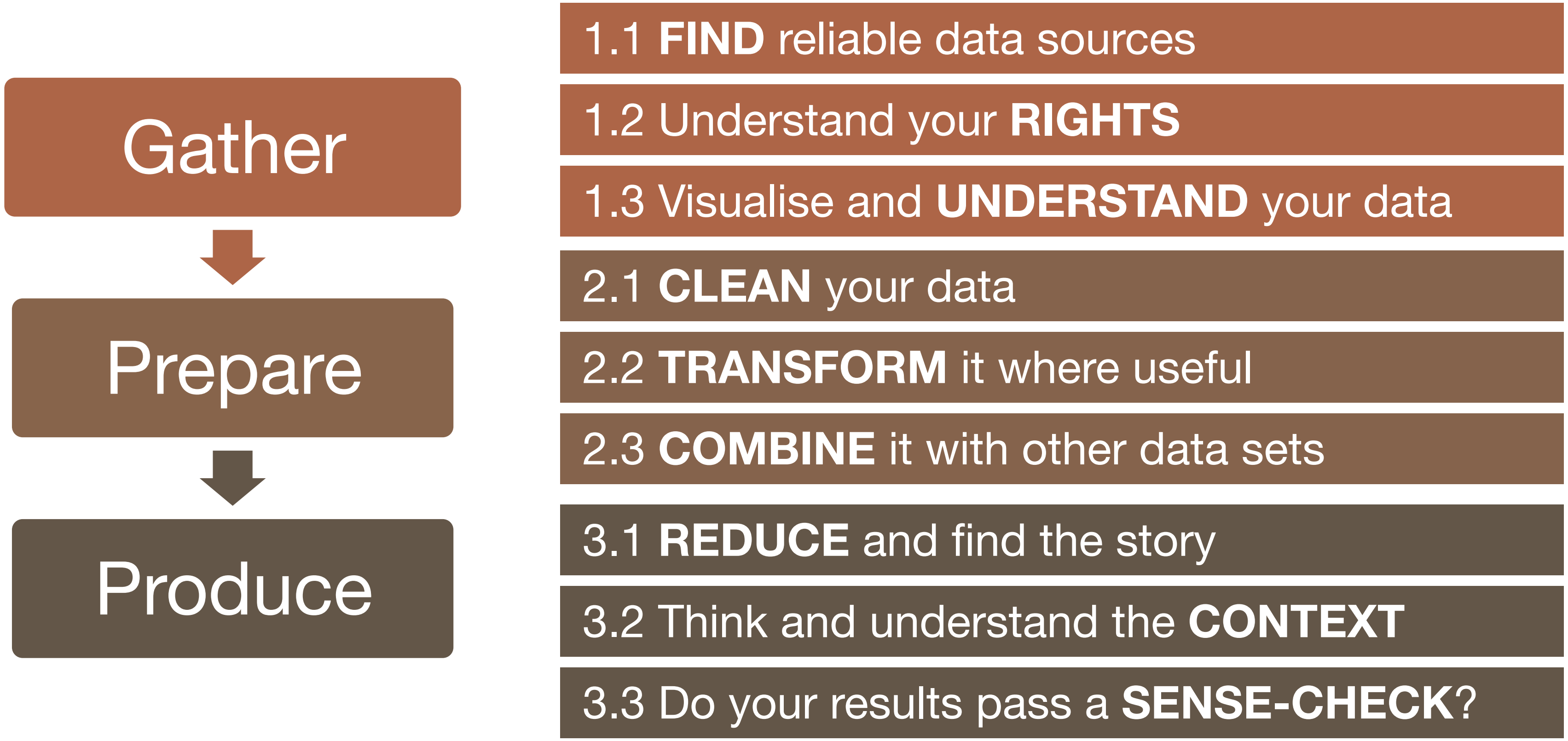
SPREAD SHEETS



SPREAD SHEETS
DATA IN WRONG FORMAT
MERGED CELLS
UNNECESSARY COLUMNS OF DATA
DATA MEASURED IN DIFFERENT UNITS
PERFORM CALCULATIONS ON THE DATA
RECALCULATE IF NEEDED
SANITY CHECK THE RESULTS


PERFORM CALCULATIONS ON THE DATA
RECALCULATE IF NEEDED
SANITY CHECK THE RESULTS
OUTPUT
OUR GRAPHICS TEAM
STORY
FREE VIZ TOOLS
GOOGLE FUSION TABLE
JUST PUBLISH

How should I budget my time?



Gather

Prepare

Produce

1.1 **FIND** reliable data sources

1.2 Understand your **RIGHTS**

1.3 Visualise and **UNDERSTAND** your data

2.1 **CLEAN** your data

2.2 **TRANSFORM** it where useful

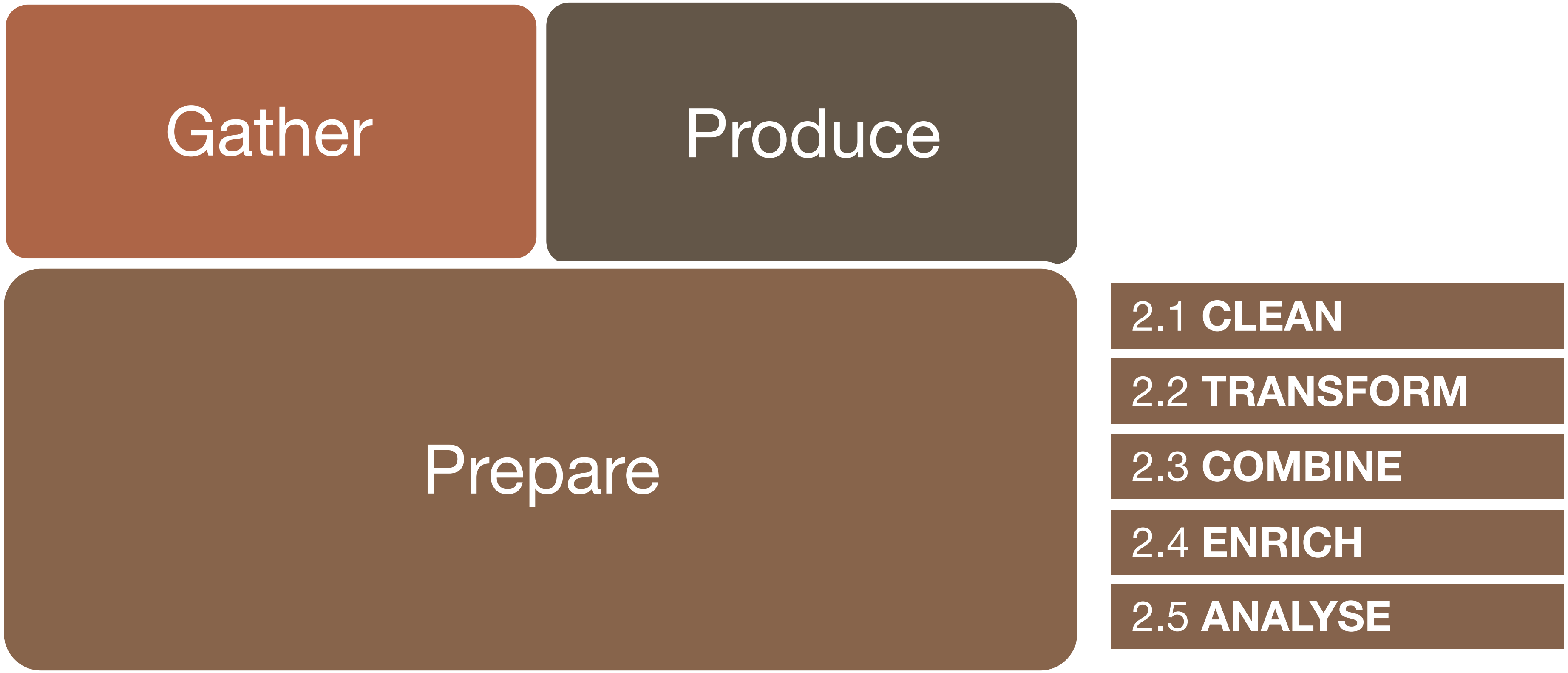
2.3 **COMBINE** it with other data sets

3.1 **REDUCE** and find the story

3.2 Think and understand the **CONTEXT**

3.3 Do your results pass a **SENSE-CHECK**?

Time planning



Gather

Produce

Prepare

2.1 **CLEAN**

2.2 **TRANSFORM**

2.3 **COMBINE**

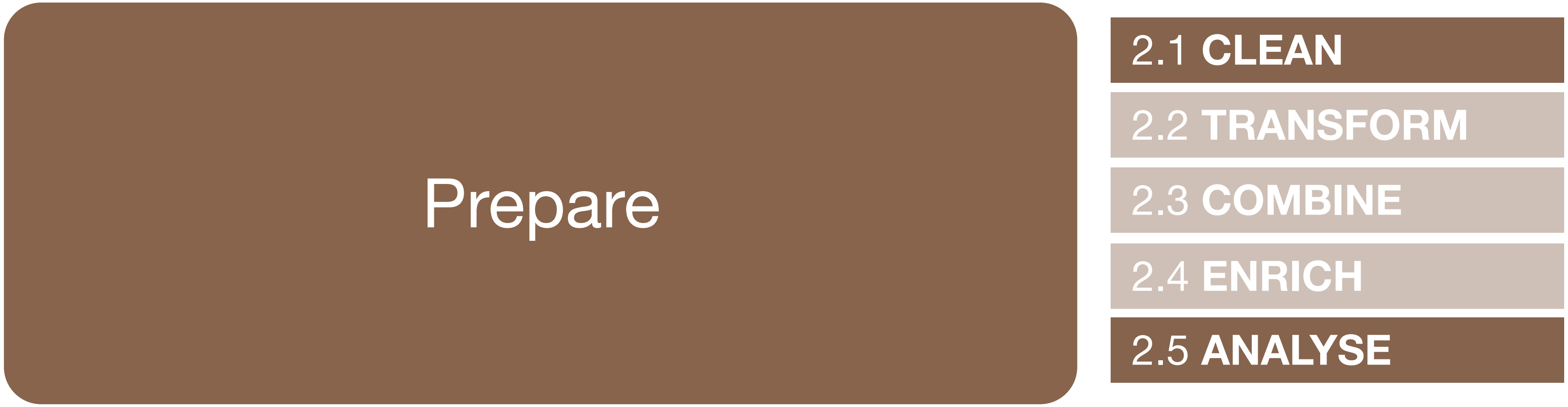
2.4 **ENRICH**

2.5 **ANALYSE**

Validating and cleaning data

Inception points: cross filters, outliers and comparisons

Prepare (Stage 1)



Prepare

2.1 **CLEAN**

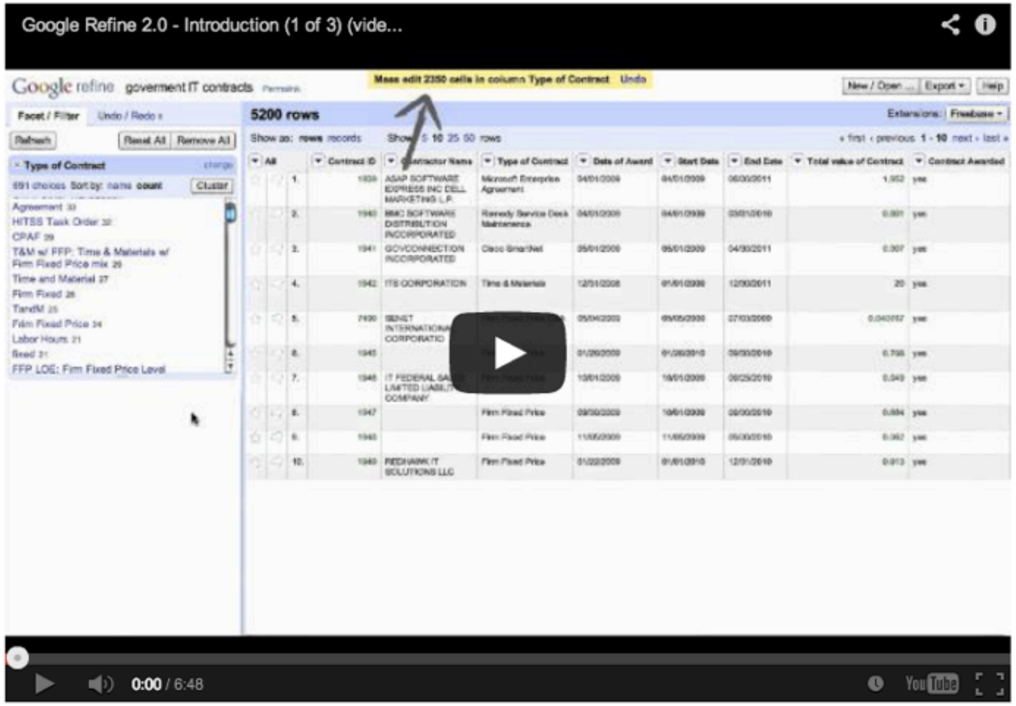
2.2 **TRANSFORM**

2.3 **COMBINE**

2.4 **ENRICH**

2.5 **ANALYSE**

Introducing Open Refine



Mass edit 2350 cells in column Type of Contract Undo

× Type of Contract change

691 choices Sort by: name count

- Agreement 33
- HITSS Task Order 32
- CPAF 29
- T&M w/ FFP: Time & Materials w/ Firm Fixed Price mix 28
- Time and Material 27
- Firm Fixed 26
- TandM 25
- Firm Fixed Price 24
- Labor Hours 21
- fixed 21
- FFP LOE: Firm Fixed Price Level

5200 rows

Show as: rows records   Show: 5 10 25 50 rows   « first ‹ previous 1 - 10 next › last »

All		Contract ID	Contractor Name	Type of Contract	Date of Award	Start Date	End Date	Total value of Contract	Contract Awarded
	1.	1939	ASAP SOFTWARE EXPRESS INC DELL MARKETING L.P.	Microsoft Enterprise Agreement	04/01/2009	04/01/2009	06/30/2011	1.952	yes
	2.	1940	BMC SOFTWARE DISTRIBUTION INCORPORATED	Remedy Service Desk Maintenance	04/01/2009	04/01/2009	03/31/2010	0.001	yes
	3.	1941	GOVCONNECTION INCORPORATED	Cisco SmartNet	05/01/2009	05/01/2009	04/30/2011	0.307	yes
	4.	1942	ITS CORPORATION	Time & Materials	12/01/2008	01/01/2009	12/30/2011	20	yes
	5.	7430	SENET INTERNATIONA CORPORATIO	[illegible]	05/04/2009	05/05/2009	07/03/2009	0.040057	yes
	6.	1945		[illegible]	01/28/2009	01/28/2010	09/30/2010	0.708	yes
	7.	1946	IT FEDERAL SA LIMITED LIABILI COMPANY	Firm Fixed Price	10/01/2009	10/01/2009	09/25/2010	0.049	yes
	8.	1947		Firm Fixed Price	09/30/2009	10/01/2009	09/30/2010	0.884	yes
	9.	1948		Firm Fixed Price	11/05/2009	11/05/2009	05/30/2010	0.362	yes
	10.	1949	REDHAWK IT SOLUTIONS LLC	Firm Fixed Price	01/22/2009	01/01/2010	12/31/2010	0.913	yes

http://openrefine.org

Exercise

Validating and cleaning data exercise

Focus on sections:

- 2) Multiple representations
- 4) Summation records
- 5) Mixed use of numerical scales

Thank-you