

Data Analytics

Professor Ernesto Lee

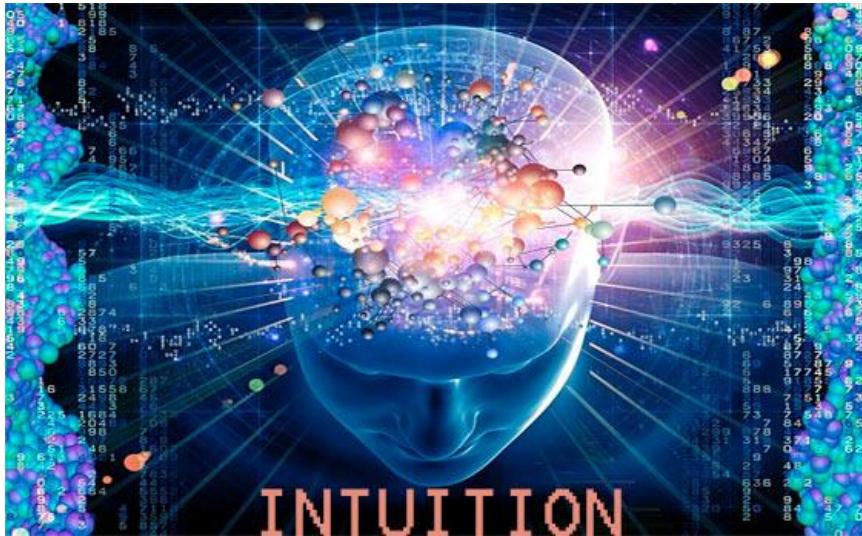


Rocks and Slingshots



Tech and Intuition

- Details and math equations and code and tools can be thought of as the **rocks** that do the actual work, but are cheap & plentiful
- Intuition (the **slingshot**) is the framework that makes the details effective.
 - In pure theory, intuition is optional but in reality - intuition is absolutely not optional.



The Slingshot in your Head



Approaches to Learning Data Analytics



The Dynamite Rock



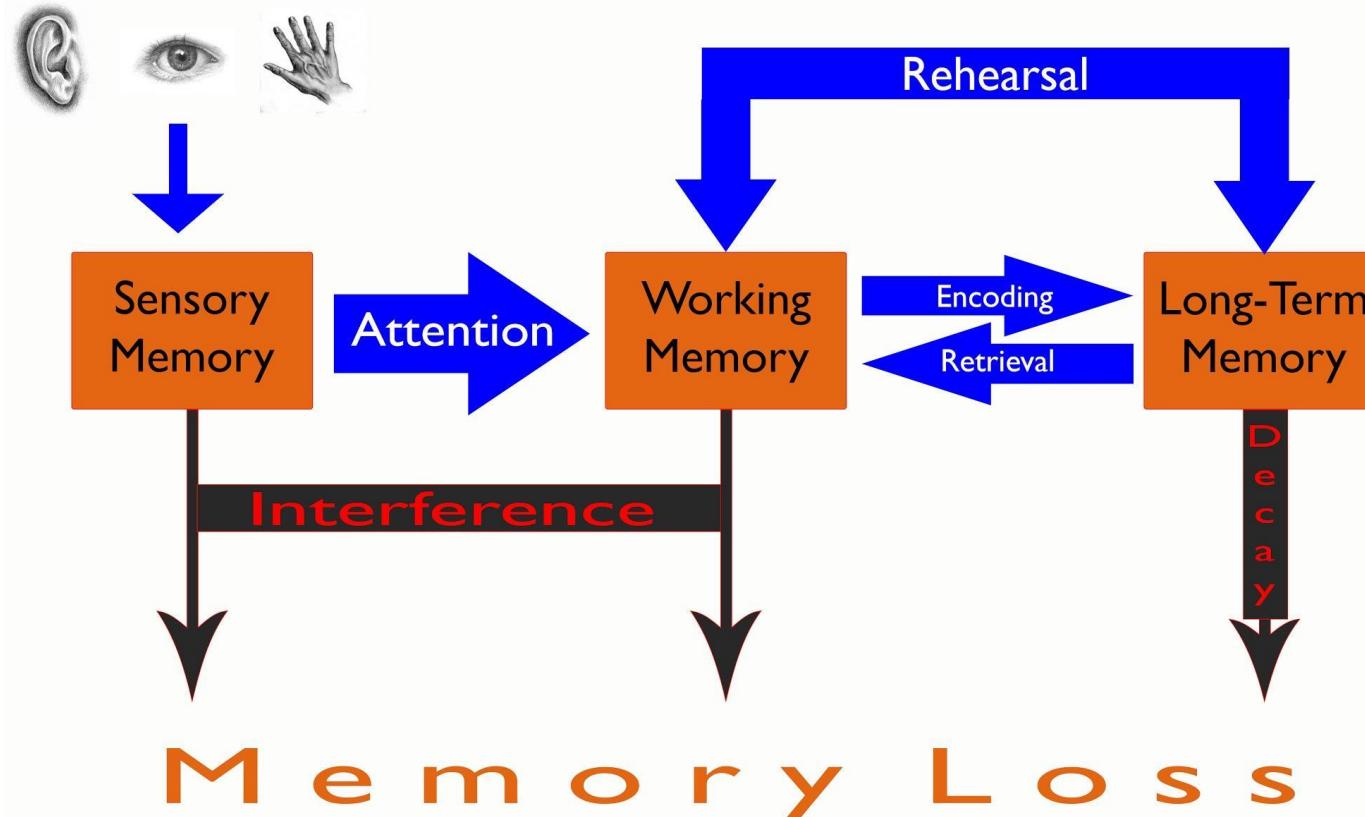
Finding the Essence



Signs of Learning



Cheap Entertainment



Use this class to
upgrade your
slingshot...

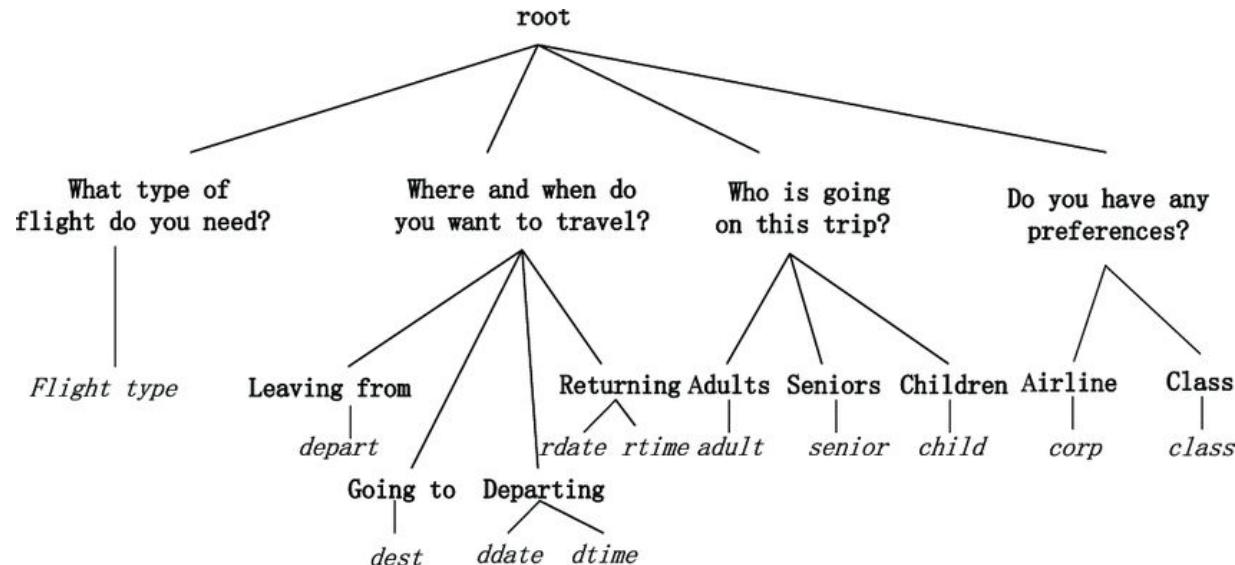


Part II : How to Learn in My Class

ADEPT Method for Learning	
Analogy	Tell me what it's like.
Diagram	Help me visualize it.
Example	Allow me to experience it.
Plain English	Describe it with everyday words.
Technical Definition	Discuss the formal details.

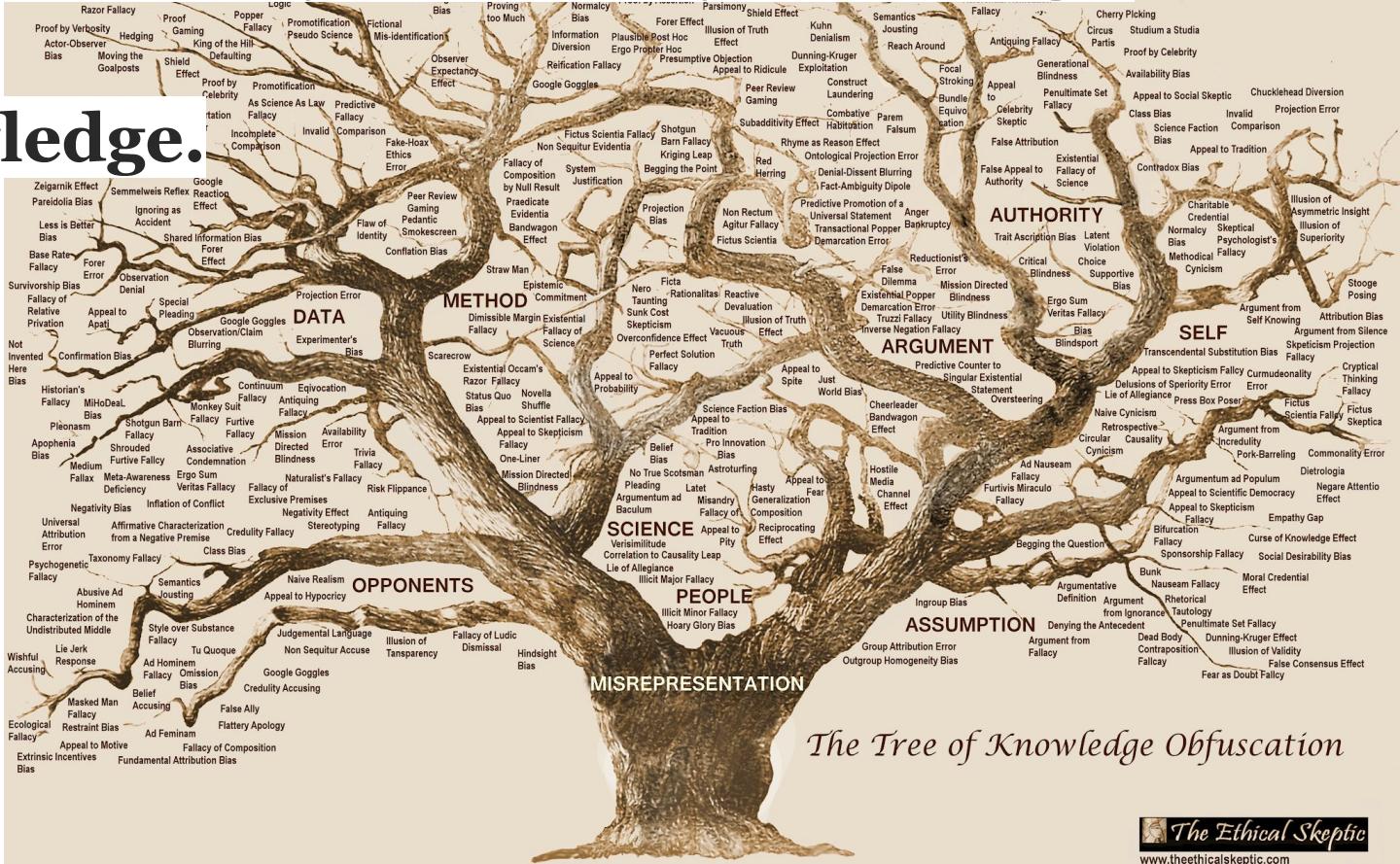
Semantic Tree

Elon Musk believes that learning involves understanding what he called a Semantic Tree. You must learn this concept and identify the different parts of the tree.



“One bit of advice: it is important to view knowledge as sort of a semantic tree — make sure you understand the fundamental principles, i.e. the trunk and big branches, before you get into the leaves/details or there is nothing for them to hang on to.”

Rule #1 — Make sure you're always building a tree of knowledge.

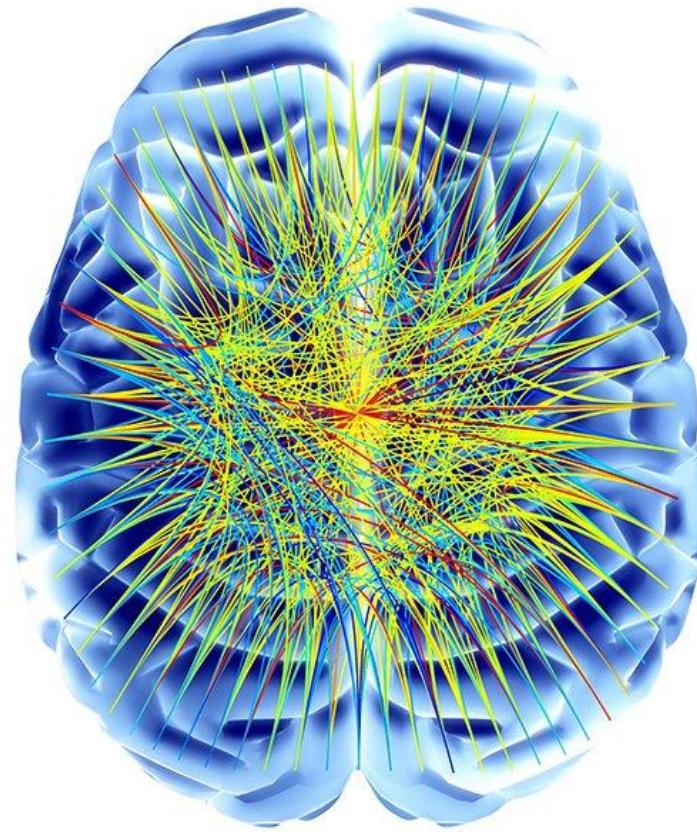


Your Tree

**Your brain is a
network. Therefore
it is the
CONNECTIONS
that are truly the
most powerful part
of your learning.**



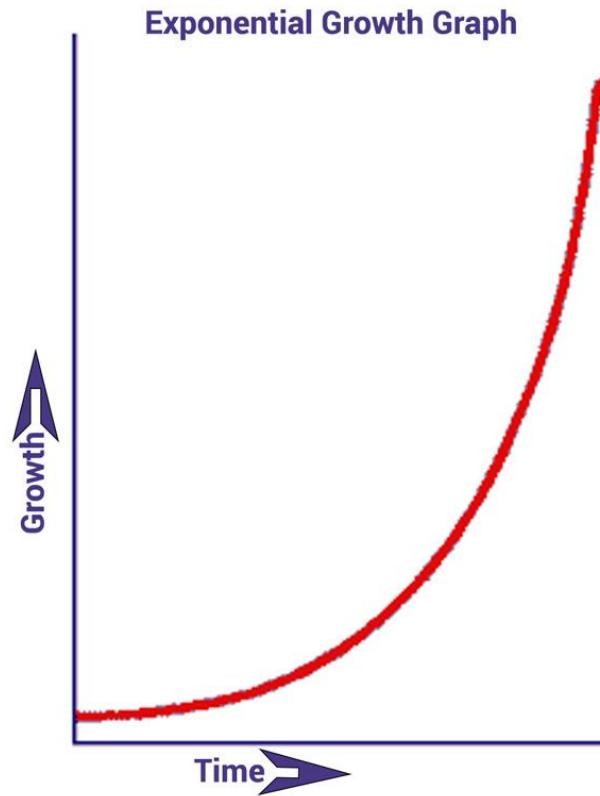
Rule #2 — You can never remember what you cannot connect with.



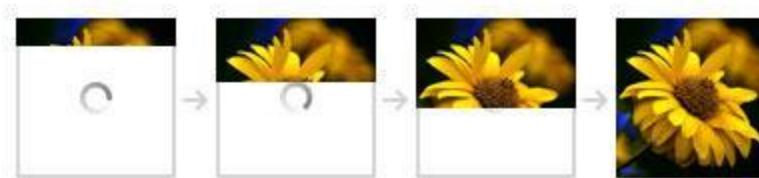
What good are a
bunch of sticks?



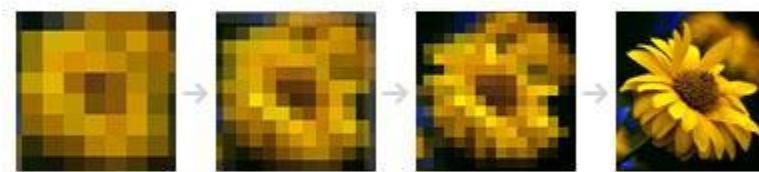
Exponential Growth



Part III : Your Personal Hero Journey



✓ Simple JPEG



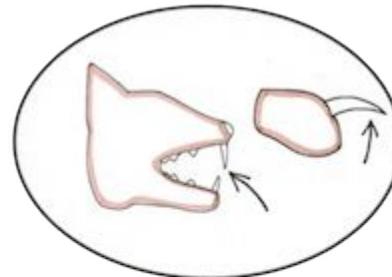
✓ Progressive JPEG



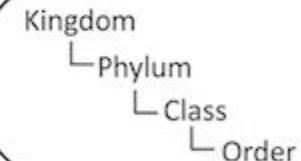
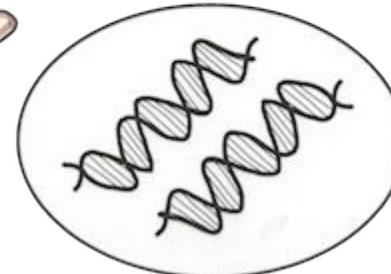
✗ Preload and then show

Develop Insights

- **Caveman definition:** A furry animal with claws, teeth, a tail, 4 legs, that purrs when happy and hisses when angry...
- **Evolutionary definition:** Mammalian descendants of a certain species (*F. catus*), sharing certain characteristics...
- **Modern definition:** You call those *definitions*? Cats are animals sharing the following DNA:
ACATACATACATACAT...



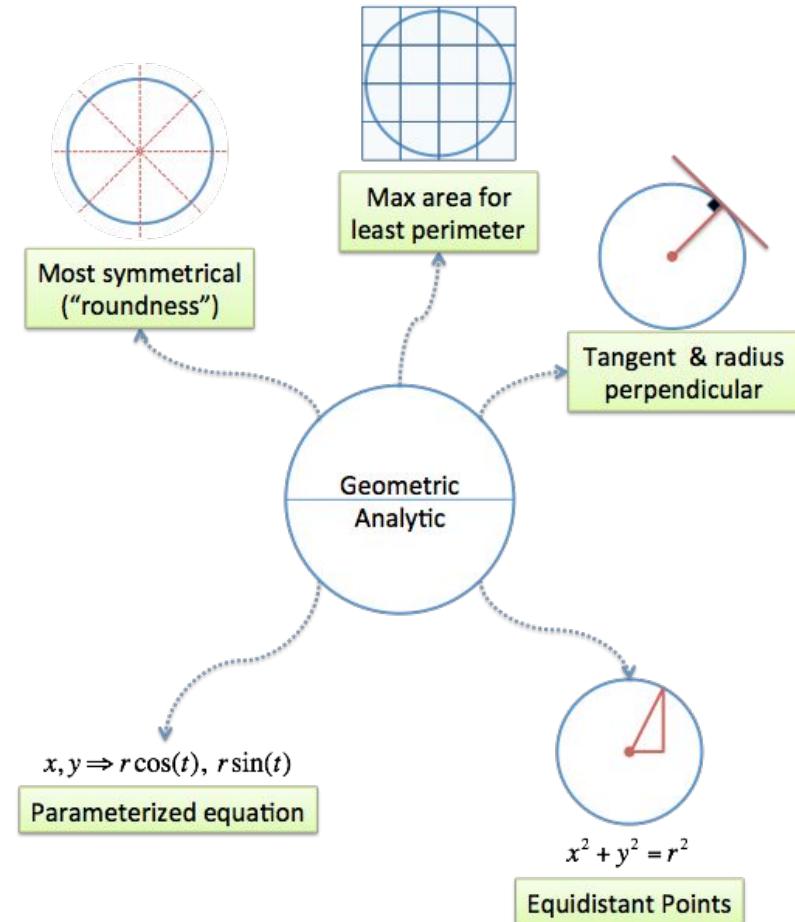
What is a Cat?



Defining a Circle

An Example

Develop Intuitions



- **Step 1: Find the central theme of a math concept.** This can be difficult, but try starting with its history. Where was the idea first used? What was the discoverer doing? This use may be different from our modern interpretation and application.
- **Step 2: Explain a property/fact using the theme.** Use the theme to make an analogy to the formal definition. If you're lucky, you can translate the math equation ($x^2 + y^2 = r^2$) into a plain-english statement ("All points the same distance from the center").
- **Step 3: Explore related properties using the same theme.** Once you have an analogy or interpretation that works, see if it applies to other properties. Sometimes it will, sometimes it won't (and you'll need a new insight), but you'd be surprised what you can discover.

What's the point to all of this?

- **Search for insights and apply them.** That first intuitive insight can help everything else snap into place. Start with a definition that makes sense and “walk around the circle” to find others.
- **Develop mental toughness.** Banging your head against an idea is no fun. If it doesn’t click, come at it from different angles. There’s another book, another article, another person who explains it in a way that makes sense to you.
- **It’s ok to be visual.** We think of math as rigid and analytic — but visual interpretations are ok! Do what develops your understanding. Imaginary numbers were puzzling until their geometric interpretation came to light, decades after their initial discovery. Looking at equations all day didn’t help mathematicians “get” what they were about.

Part IV: Your Assignment

- Build a **lasting intuition** for the key ideas.
- During the course, understand it enough to solve problems.
- After the course, enjoy it enough to revisit.

ADEPT Method for Learning	
Analogy	Tell me what it's like.
Diagram	Help me visualize it.
Example	Allow me to experience it.
Plain English	Describe it with everyday words.
Technical Definition	Discuss the formal details.

Machine Learning Notes

[Machine Learning Notes](#)

[Week 1](#)

[Week 2](#)

[Week 3](#)

1-sentence course summary

- "Create predictive models with Linear Algebra and improve them with Calculus."
 - Linear Algebra efficiently describes complex operations using simple steps.
 - Calculus can optimize any model with derivatives (gradient). Minimize error.

1-sentence core concepts

- [Linear Algebra](#): spreadsheets for your equations. We "pour" data through various operations.
- [Natural log](#): time needed to grow. Helps normalize widely varying numbers.
- [e^x](#): models continuous growth, has a simple derivative.
- [Gradient](#): direction of greatest change, helps optimize.
- [Calculus](#) -Art of breaking a system into steps. With the gradient, we can move in the best direction.

Aha!

Aha!

For the major concepts the course depends on, I keep a 5-second summary in mind. This underlying concept, why does it exist? In plain English, what does it mean?

Aha!

HUH?



Huh's - they can be a question that you have for anything that is unclear. They can be a GOTCHA!

Record them as well!

Summarize this course in PLAIN ENGLISH

WHAT HAVE
YOU LEARNED?

The Results...



Naive Bayes

By: Ernesto Lee

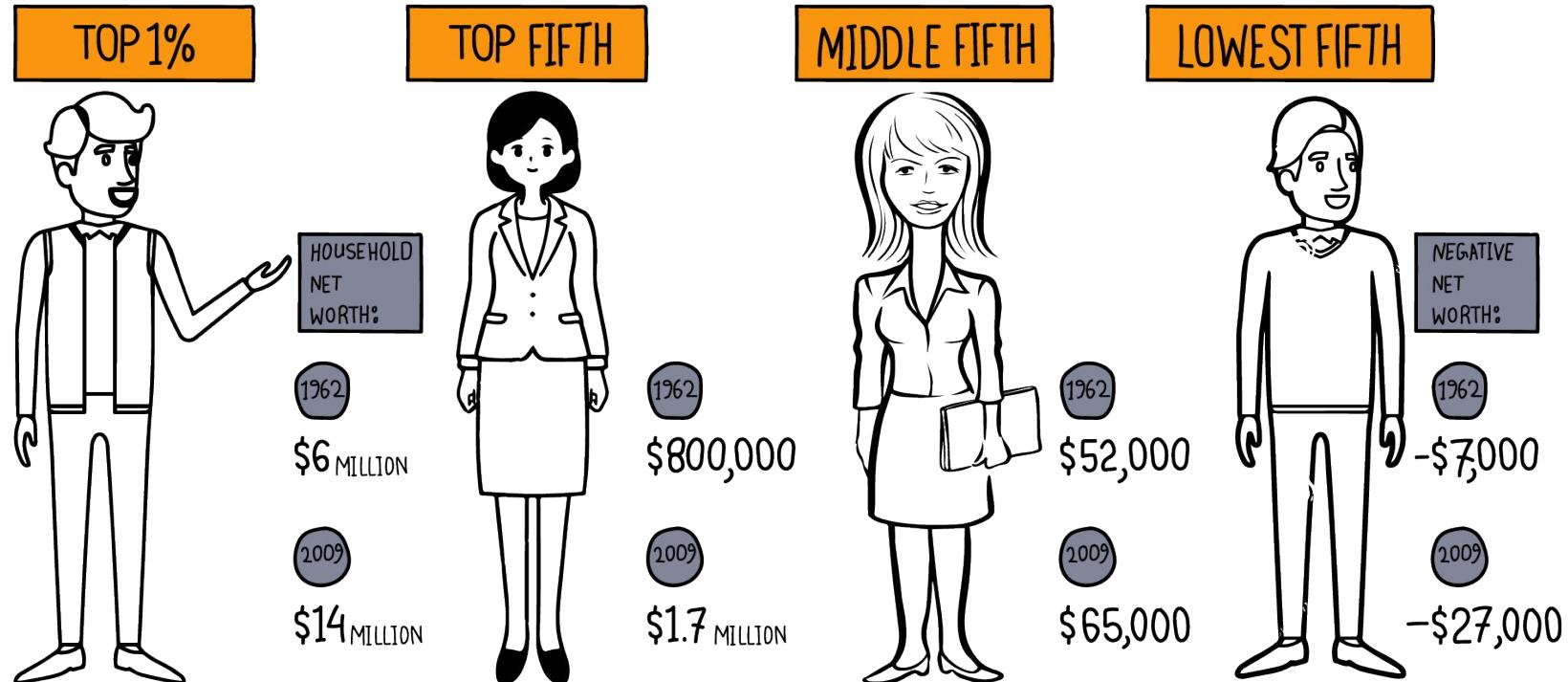
Conditional Probability

WE ARE THE
99%

WE DESERVE
CHANGE !



What are the chances of being part of the 1%?



Research

Research has shown that $\frac{1}{4}$ of the 1% is female

Math Alert:

$\frac{1}{4}$ of 1% is female

or

$.25 * .01 = .0025$ is female

What are the chances of being FEMALE if you are 1%?

CHANCES OF BEING
FEMALE AND IN THE
1%

=

CHANCES OF BEING FEMALE
IF YOU ARE IN THE 1%

×

CHANCES OF BEING IN
THE 1%

What are the chances of being 1% if you are FEMALE?

CHANCES OF BEING
FEMALE AND IN THE
1%

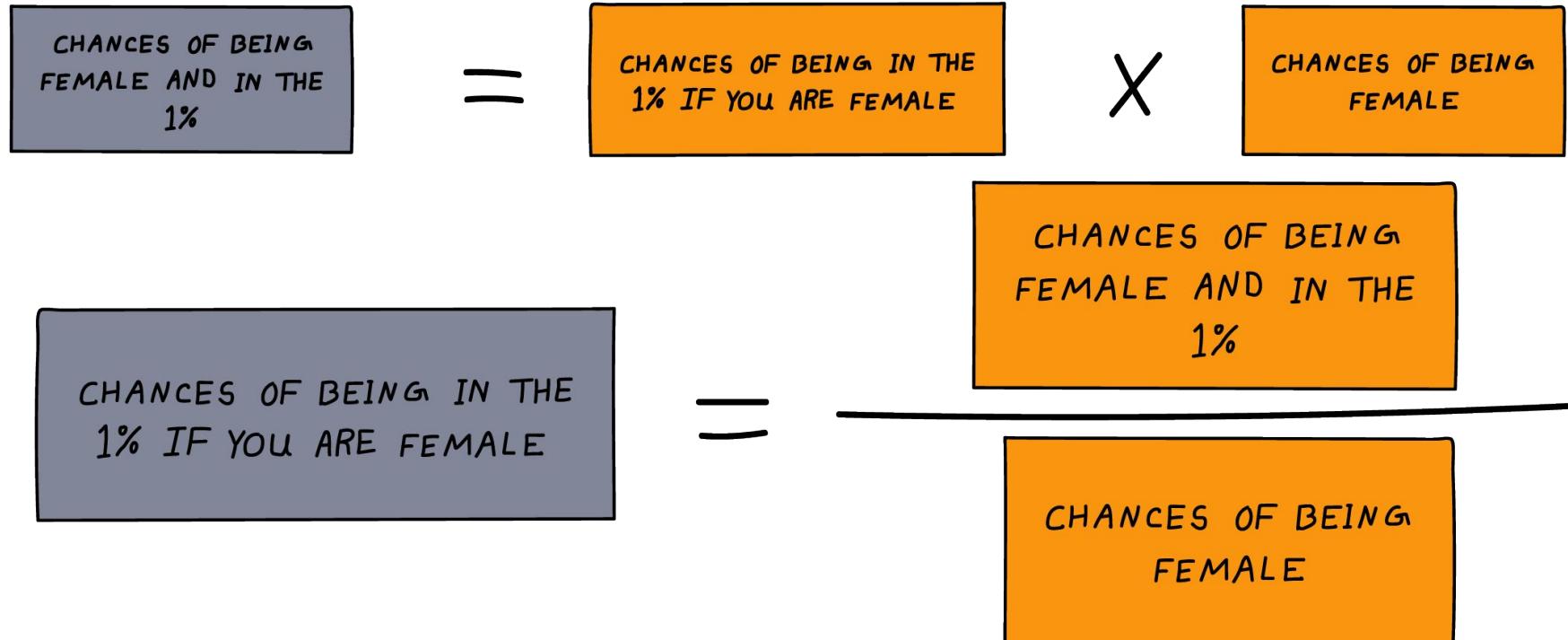
=

CHANCES OF BEING IN THE
1% IF YOU ARE FEMALE

X

CHANCES OF BEING
FEMALE

What are the chances of being 1% if you are FEMALE?



CHANCES OF BEING
FEMALE AND IN THE
1%

=

CHANCES OF BEING IN THE
1% IF YOU ARE FEMALE

X

CHANCES OF BEING
FEMALE

CHANCES OF BEING IN THE
1% IF YOU ARE FEMALE

=

CHANCES OF BEING
FEMALE AND IN THE
1%

CHANCES OF BEING
FEMALE

$$x * 0.5 = .0025$$

$$x = .0025 / 0.5$$

$$x = .005$$

“Bayes Rule” is just a formalization of the logic I just explained.

CHANCES OF BEING IN THE
1% IF YOU ARE FEMALE

=

CHANCES OF BEING
FEMALE AND IN THE
1%

CHANCES OF BEING
FEMALE

$$= x = .0025 / 0.5$$

$$= x = \frac{(.25 * .01)}{0.5}$$

CHANCES OF BEING IN
1% IF YOU ARE FEMALE

=

CHANCE THAT YOU ARE
FEMALE IF YOU'RE IN THE
1%

X

CHANCES OF BEING IN
THE 1%

CHANCES OF
BEING FEMALE

CHANCES A, GIVEN B

=

CHANCES OF B, GIVEN A

X

TOTAL CHANCES
OF A

CHANCES OF
B

Bayes Theorem

Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypothesis before observing the evidence?

$$P(H | e) = \frac{P(e | H) P(H)}{P(e)}$$

Posterior

How probable is our hypothesis given the observed evidence?
(Not directly computable)

Marginal

How probable is the new evidence under all possible hypothesis?
 $P(e) = \sum P(e | H) P(H)$

Use Bayes to discover the chances that you are in the 1% IF you are male

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Probability of A occurring given evidence B has already occurred

Probability of B occurring given evidence A has already occurred

Probability of A occurring

Probability of B occurring

$$P(A|B) = \frac{(0.75) \times (.01)}{0.5}$$

Bayes' Theorem

Bayes' Theorem is a rule (and formula) in probability theory that can help you assess the probability of an event happening given prior knowledge about conditions related to that event.

Mathematically, it looks like this:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Note:
 $P(B)$ must not
be zero

"P" means probability.

$P(A)$ means the probability of event A happening independently \rightarrow whether or not event B happens.

$P(B)$ means the same for event B.

$P(A|B)$ means the probability of event A happening,
given that event B does happen.

$P(B|A)$ is the inverse; it's the probability of event B happening given that event A happens.

By taking the probability of event B into consideration, you can come to a more accurate conclusion about the probability of event A happening.

Software Version Control



So Why Do We Need Version Control?

- **Backup and Restore.**
- **Synchronization.**
- **Short-term undo.**
- **Long-term undo.**
- **Track Changes.**
- **Track Ownership.**
- **Sandboxing**
- **Branching and merging.**

Half of BEING good... is SOUNDING good...

Basic Setup

- **Repository (repo)**: The database storing the files.
- **Server**: The computer storing the repo.
- **Client**: The computer connecting to the repo.
- **Working Set/Working Copy**: Your local directory of files, where you make changes.
- **Trunk/Main**: The primary location for code in the repo. Think of code as a family tree — the trunk is the main line.

You gotta fake it... to make it...

Basic Actions

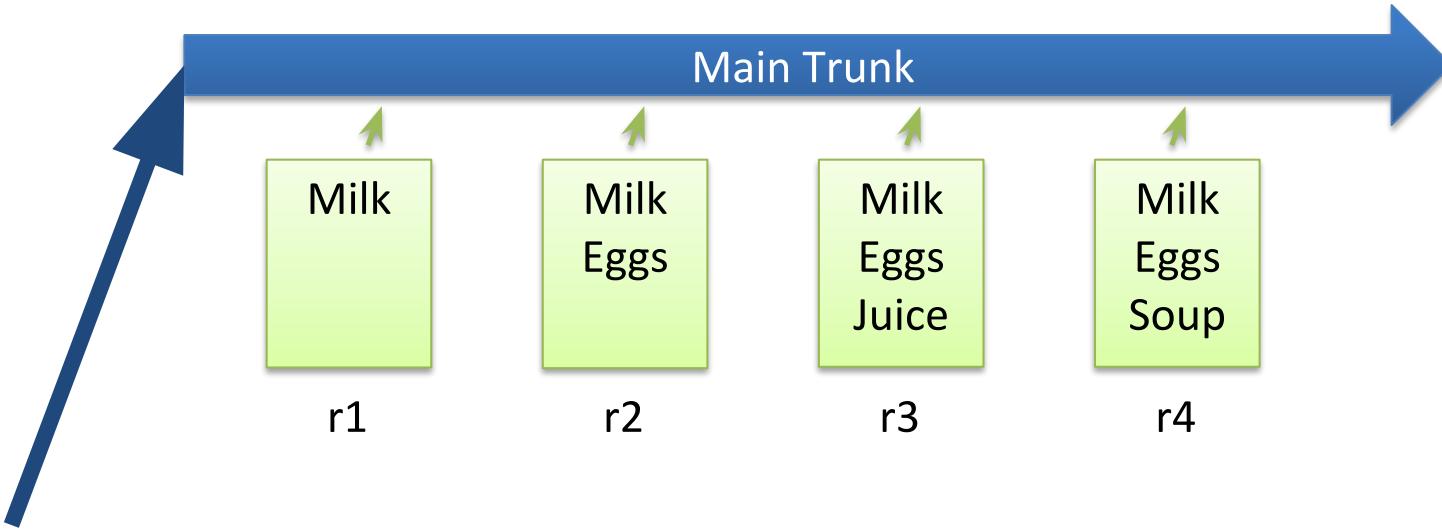
- **Add:** Put a file into the repo for the first time, i.e. begin tracking it with Version Control.
- **Revision:** What version a file is on (v1, v2, v3, etc.).
- **Head:** The latest revision in the repo.
- **Check out:** Download a file from the repo.
- **Check in:** Upload a file to the repository (if it has changed). The file gets a new revision number, and people can “check out” the latest one.
- **Checkin Message:** A short message describing what was changed.
- **Changelog/History:** A list of changes made to a file since it was created.
- **Update/Sync:** Synchronize your files with the latest from the repository. This lets you grab the latest revisions of all files.
- **Revert:** Throw away your local changes and reload the latest version from the repository.

Just Learn the Lingo...

Advanced Actions

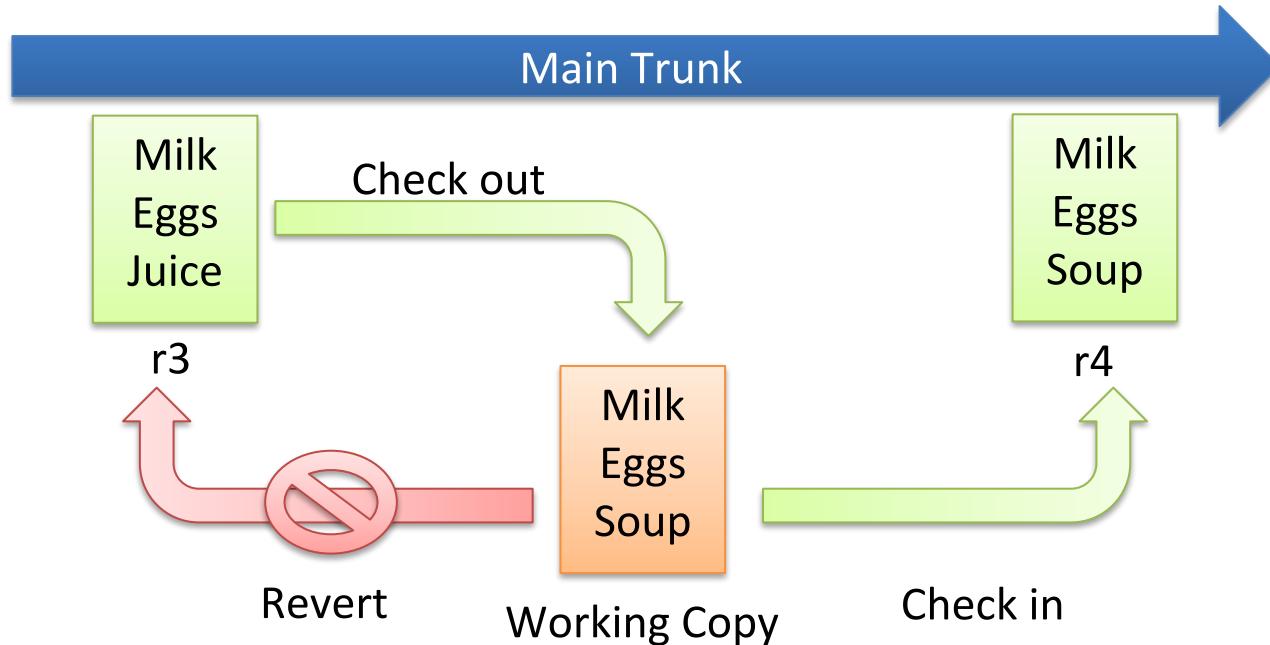
- **Branch:** Create a separate copy of a file/folder for private use (bug fixing, testing, etc). Branch is both a verb (“branch the code”) and a noun (“Which branch is it in?”).
- **Diff/Change/Delta:** Finding the differences between two files. Useful for seeing what changed between revisions.
- **Merge (or patch):** Apply the changes from one file to another, to bring it up-to-date. For example, you can merge features from one branch into another. (At Microsoft this was called [Reverse Integrate and Forward Integrate](#))
- **Conflict:** When pending changes to a file contradict each other (both changes cannot be applied).
- **Resolve:** Fixing the changes that contradict each other and checking in the correct version.
- **Locking:** Taking control of a file so nobody else can edit it until you unlock it. Some version control systems use this to avoid conflicts.
- **Breaking the lock:** Forcibly unlocking a file so you can edit it. It may be needed if someone locks a file and goes on vacation (or “calls in sick” the day Halo 3 comes out).
- **Check out for edit:** Checking out an “editable” version of a file. Some VCSes have editable files by default, others require an explicit command.

Basic Checkins

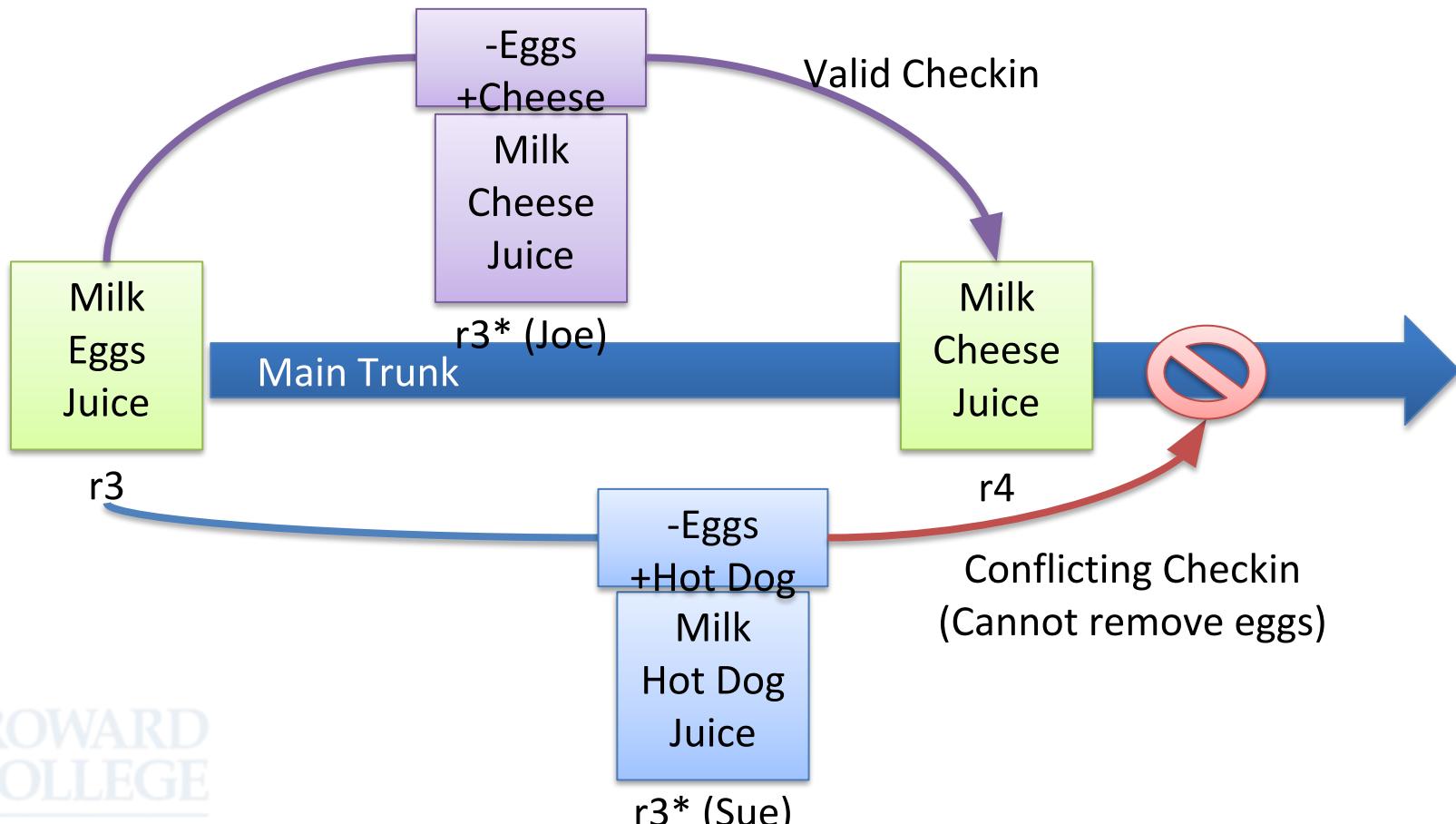


LIST.TXT

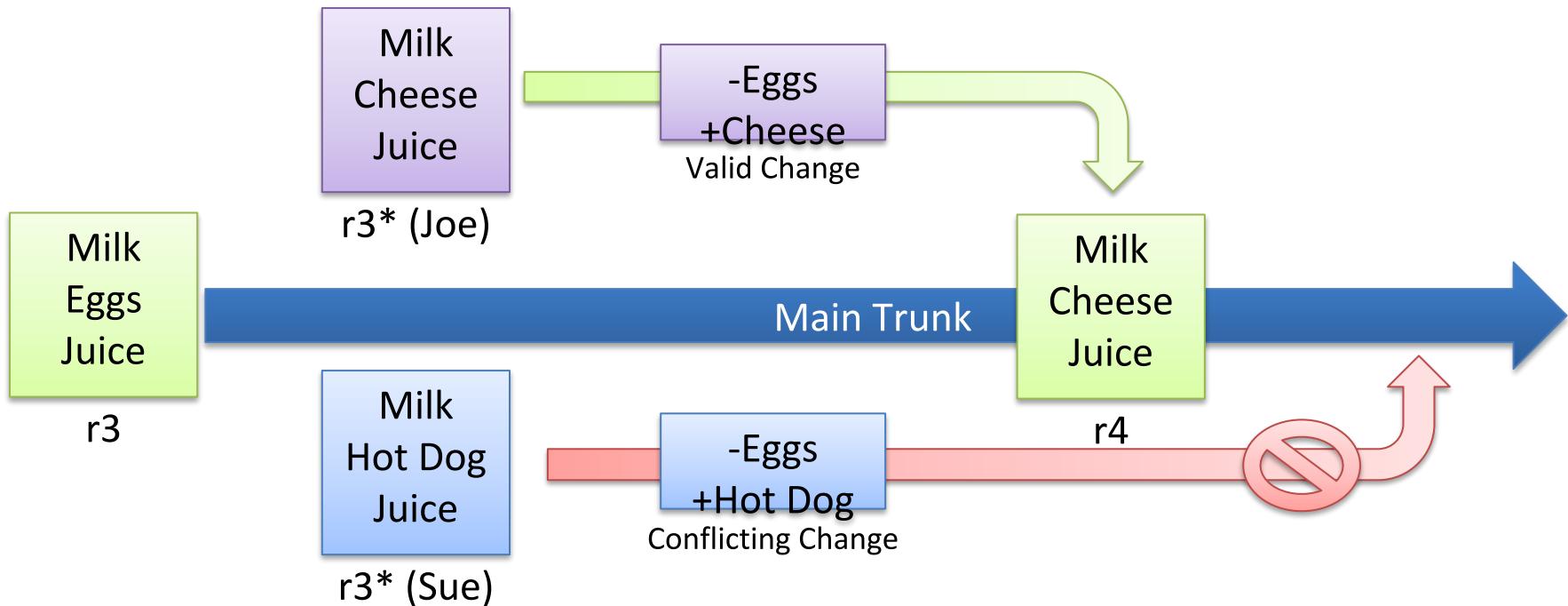
Checkout and Edit



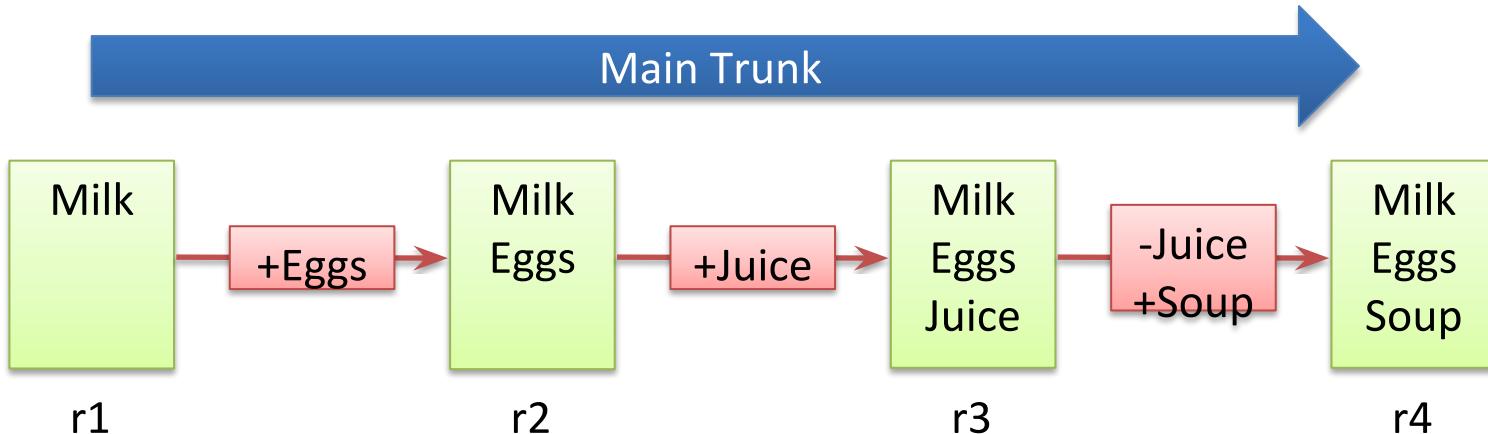
Conflicts



Conflicts (Alt)

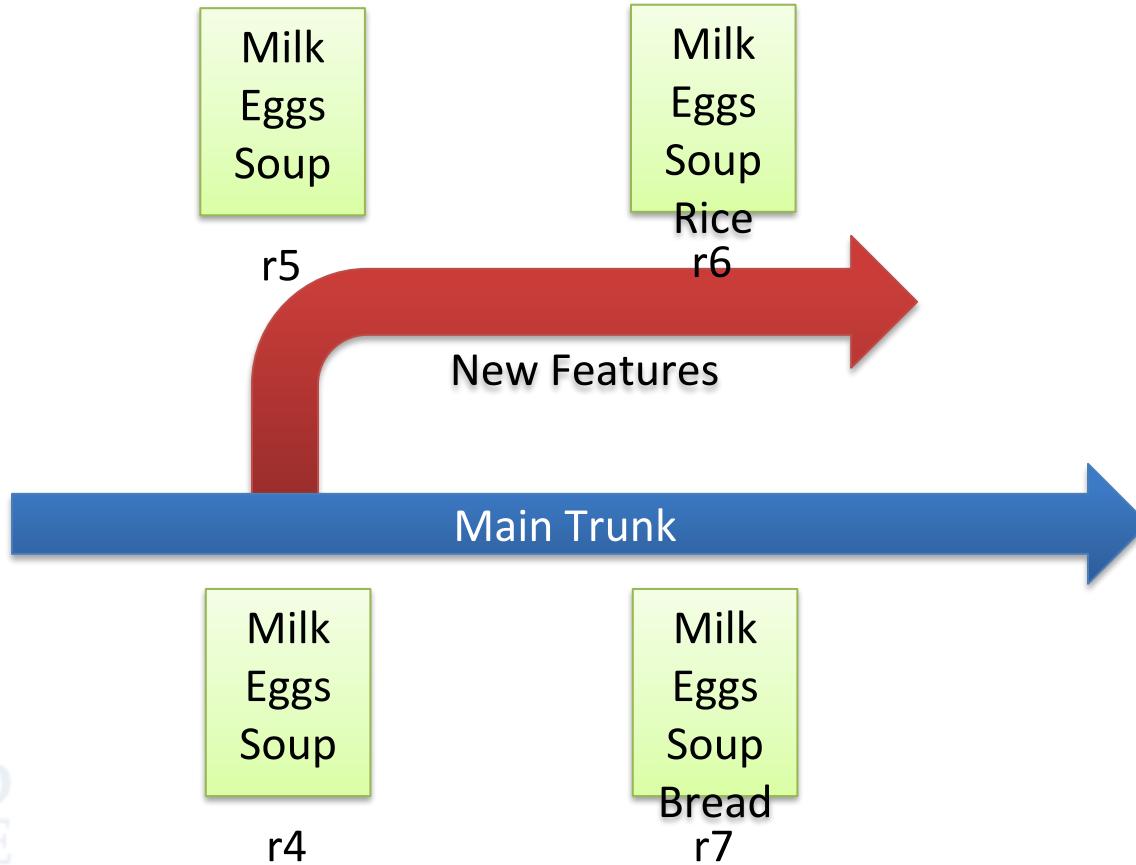


Basic Diffs

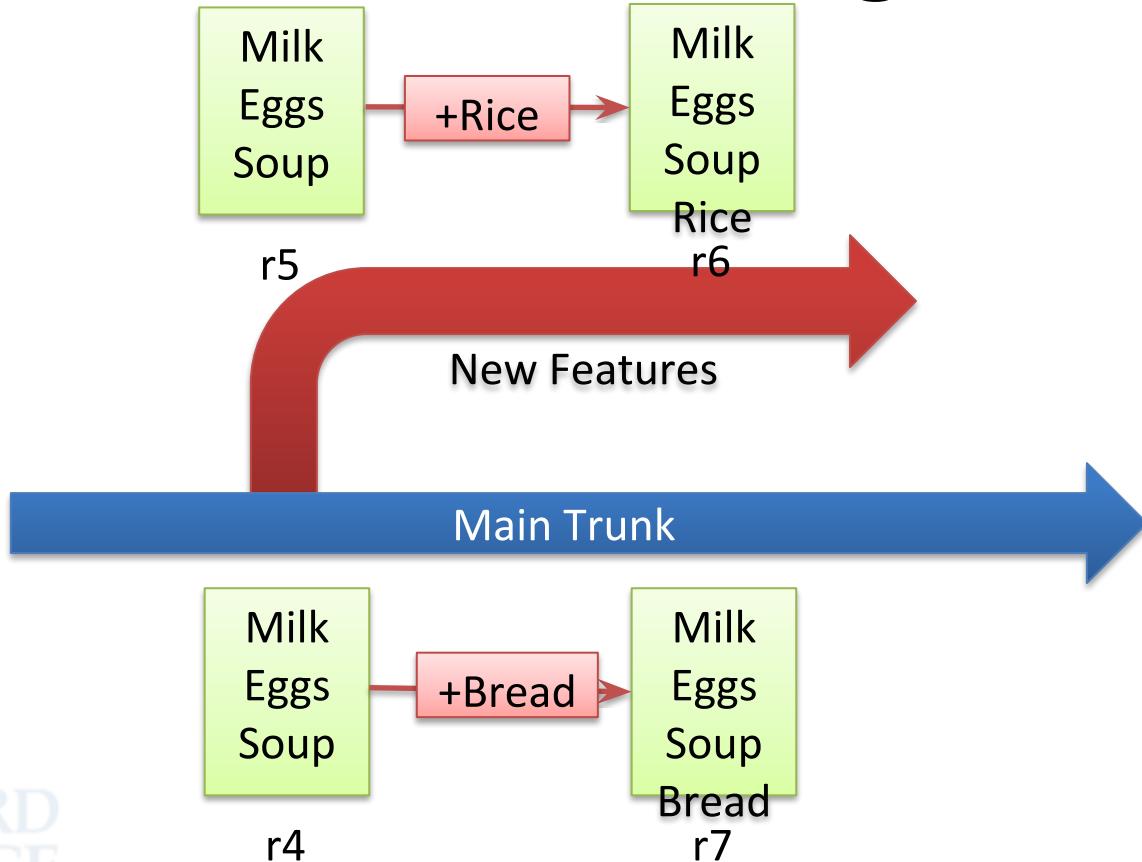


NOTE AVAILABLE IN GIT... ONLY IN SVN

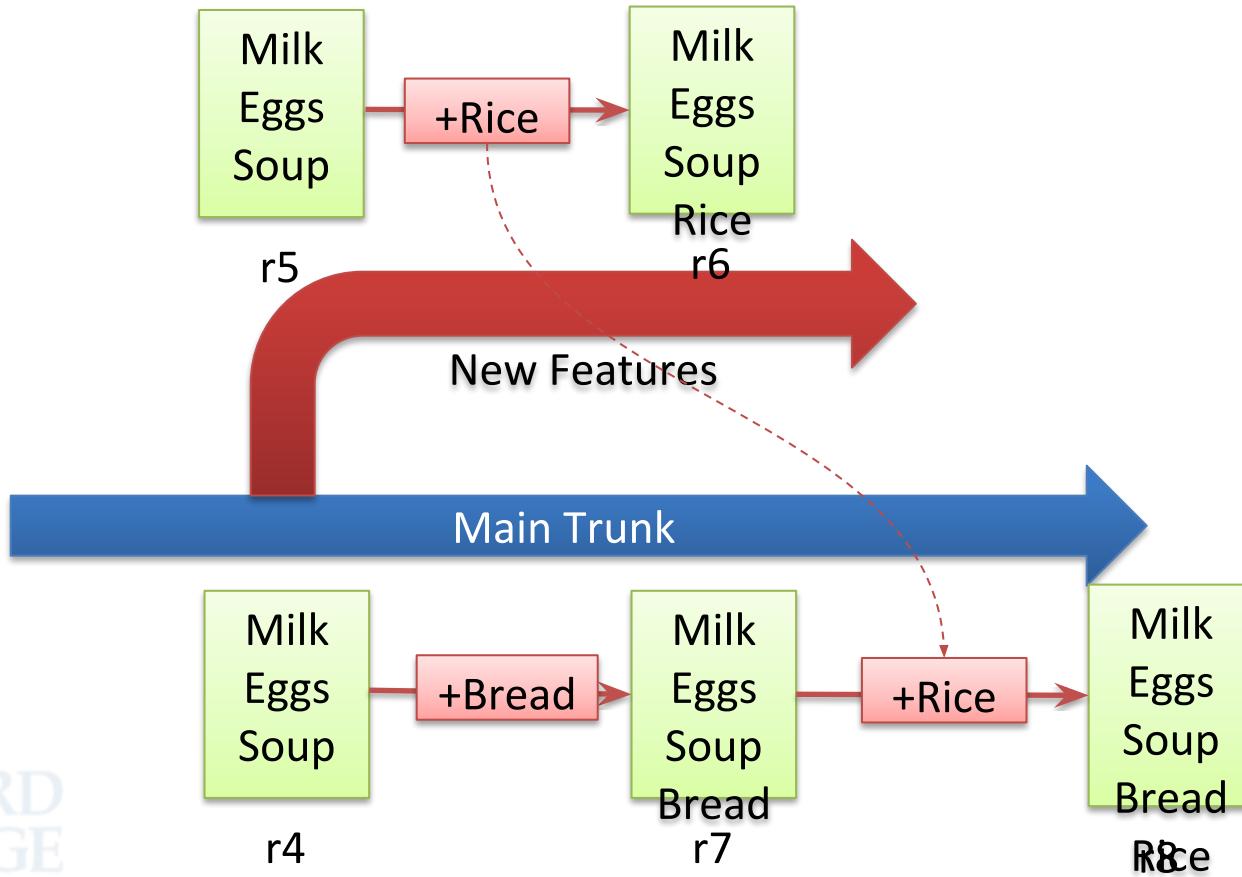
Branching



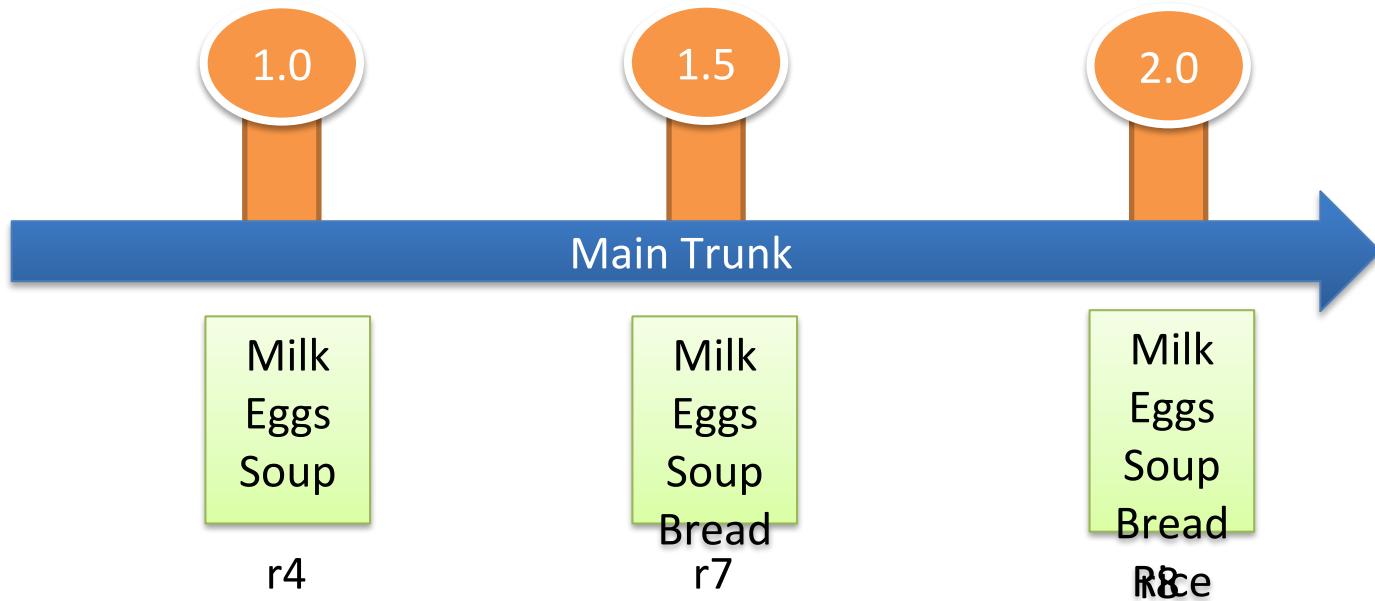
Branch Changes



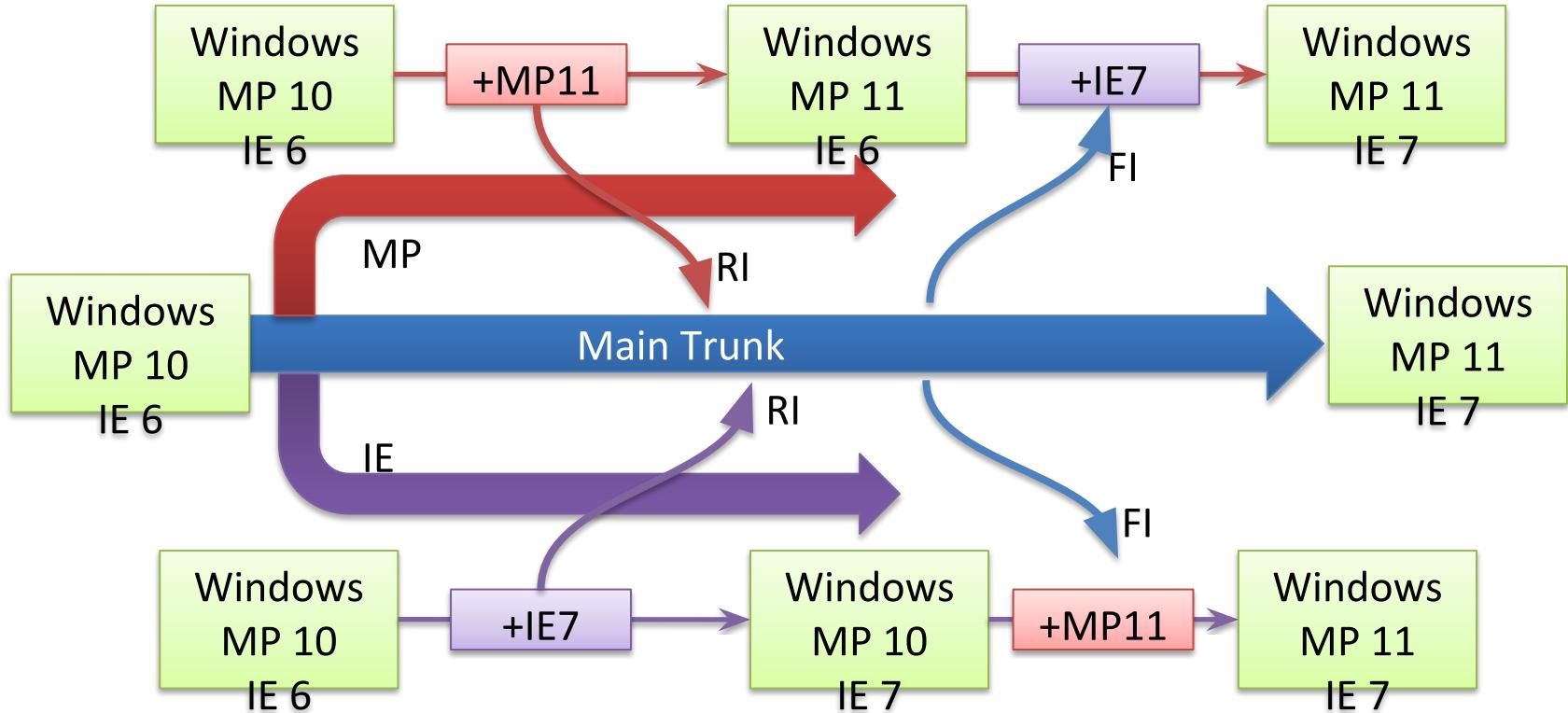
Merging



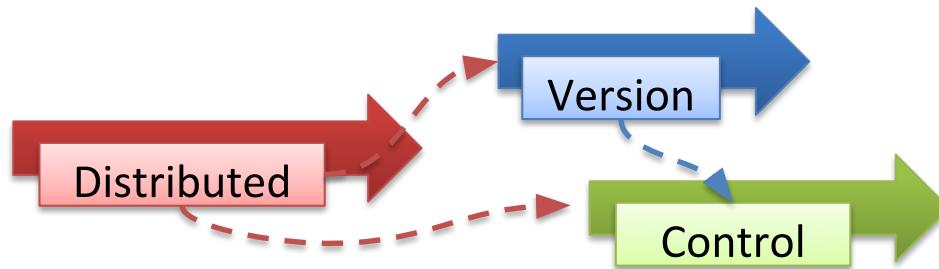
Tagging



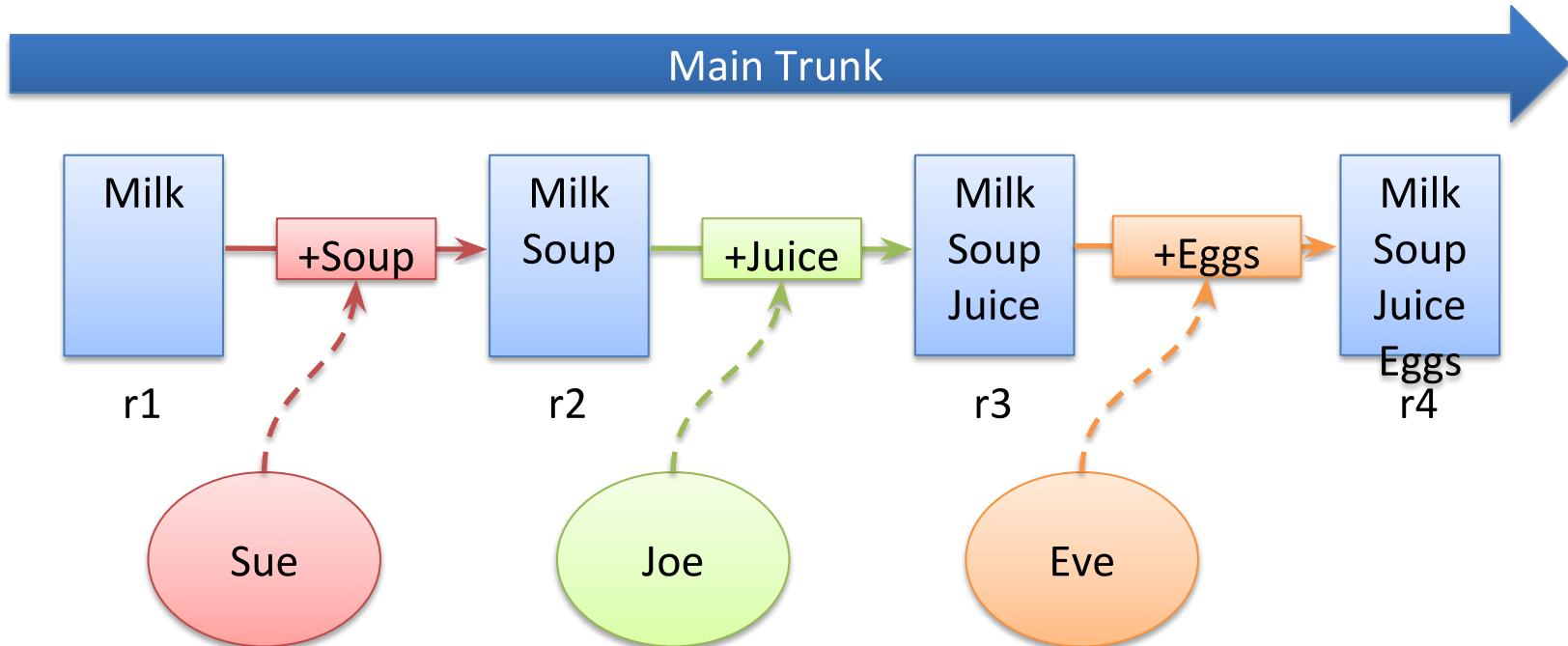
Managing Windows



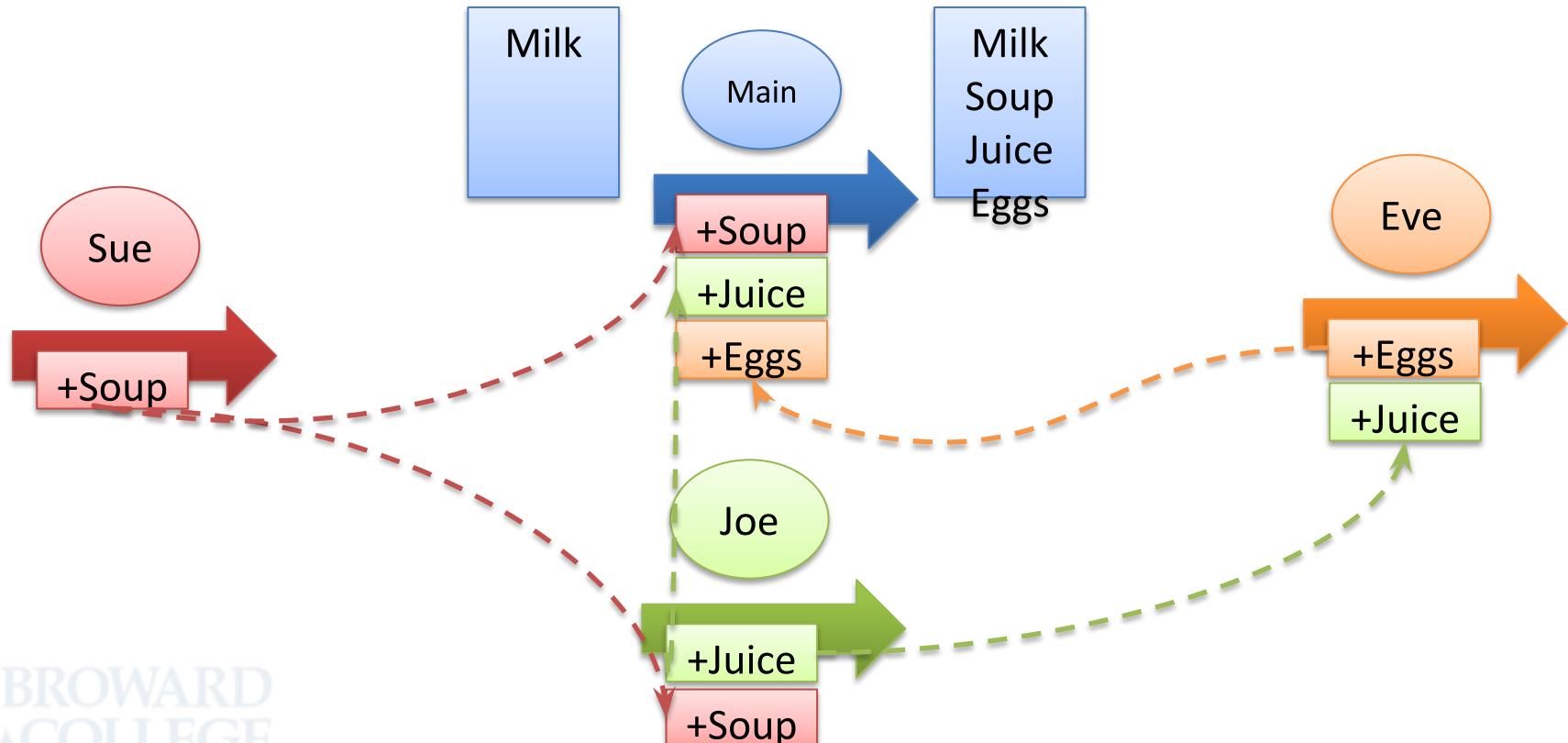
Distributed Version Control



Centralized VCS



Distributed VCS



Core Concepts

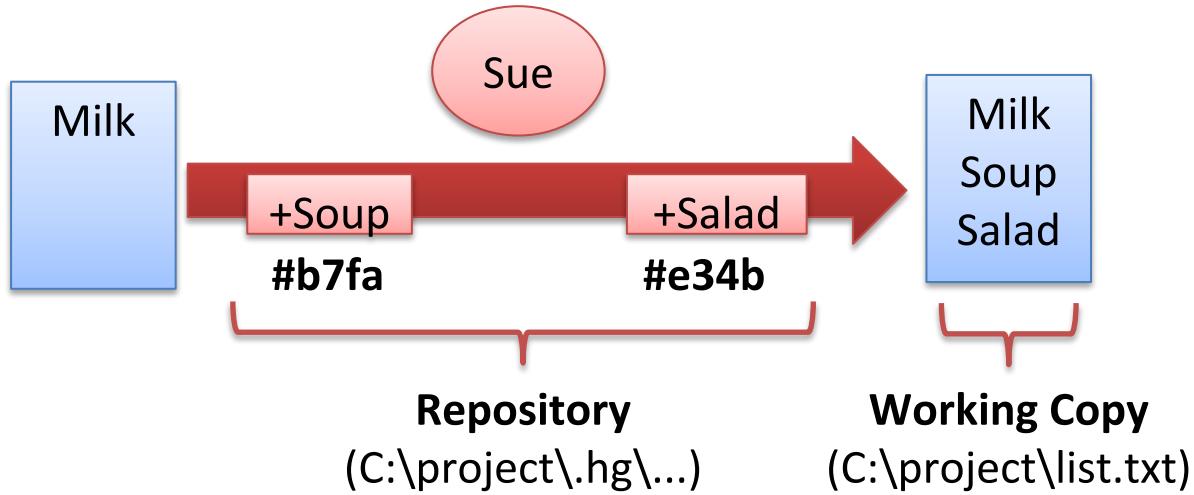
Core Concepts

- Centralized version control focuses on **synchronizing, tracking, and backing up files**.
- Distributed version control focuses on **sharing changes**; every change has a **guid or unique id**.
- **Recording/Downloading** and **applying** a change are separate steps (in a centralized system, they happen together).
- **Distributed systems have no forced structure.** You can create “centrally administered” locations or keep everyone as peers.

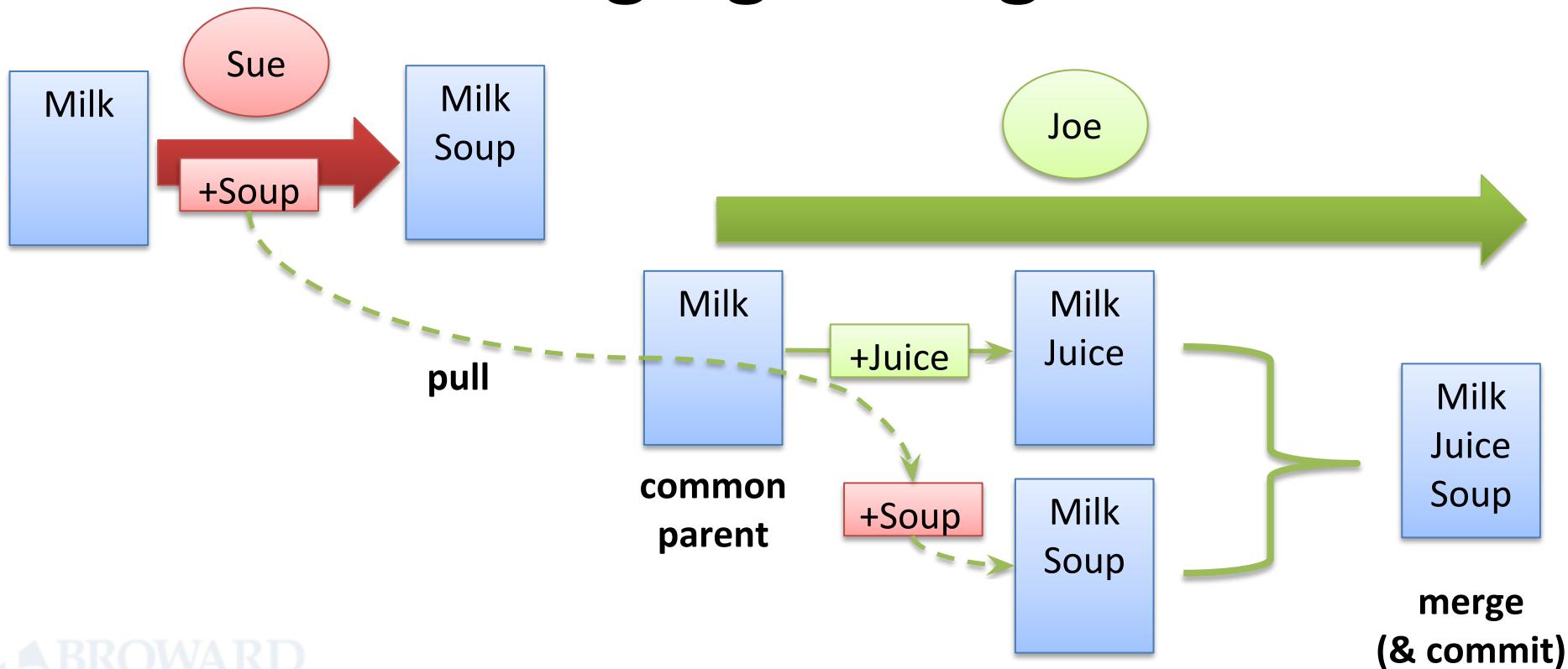
New Terminology

- **push:** send a change to another repository (may require permission)
- **pull:** grab a change from a repository

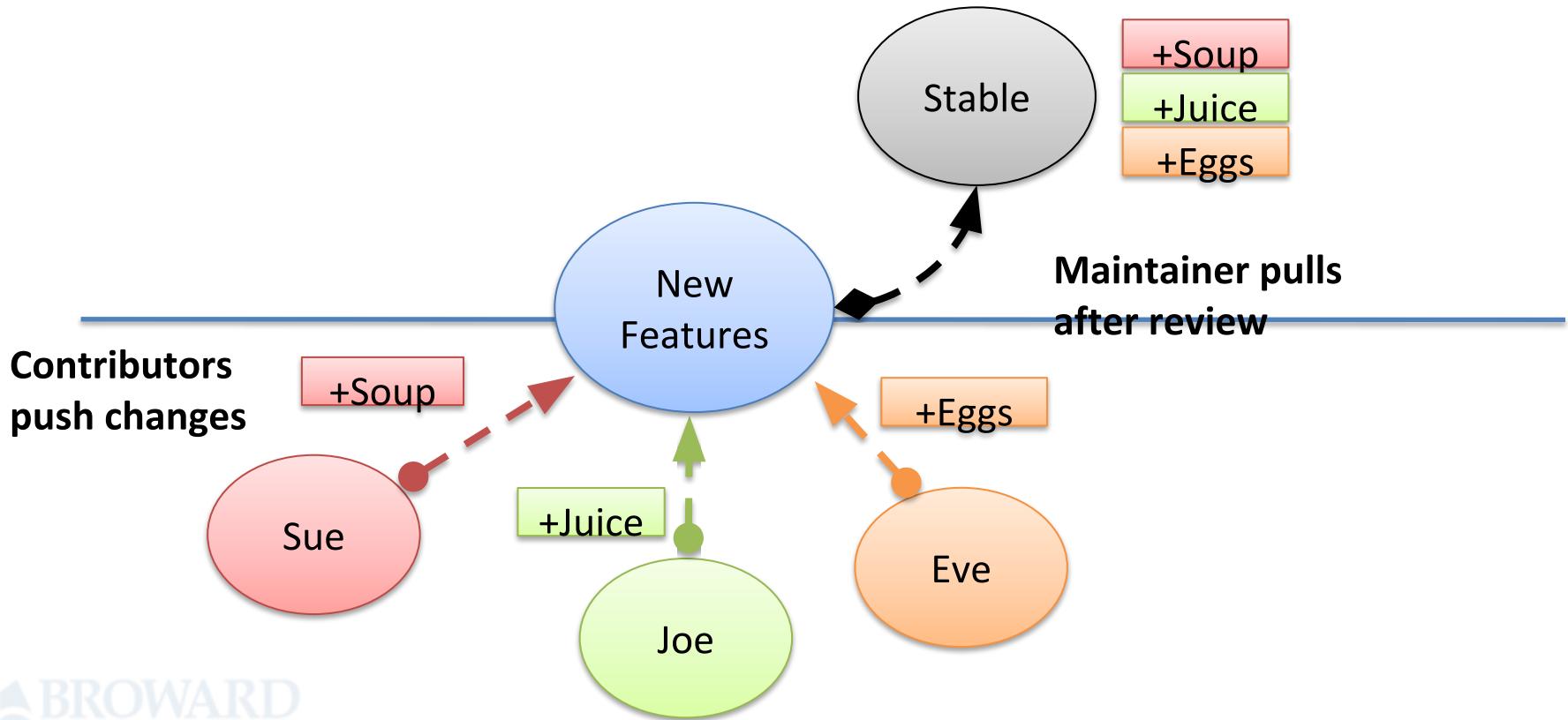
Repository Layout



Merging Changes



Distributed Push/Pull Model



What is the Slingshot?

- **Use version control.**
- **Take it slow.**
- **Keep Learning.**

DEMO

Create your GitHub Account

Create your first Repository

Create your first GitHub WebSite (Bonus)

Git “Aha’s”

Git has a staging area. **Git has a staging area!!!**

Did this ever confuse me. There's both a repo ("object database") and a staging area (called "index"). Checkins have two steps:

- `git add foo.txt`
 - Add `foo.txt` to the index. It's not checked in yet!
- `git commit -m "message"`
 - Put staged files in the repo; they're now tracked
 - You can "`git add --update`" to stage all tracked, modified files

Branching is a “Save As”

Branches are like "Save as..." on a directory. Best of all:

- Easily merge changes with the original (changes tracked and never applied twice)
- No wasted space (common files only stored once)

Imagine Virtual Directories

I see branches as "virtual directories" in the .git folder. While inside a physical directory (c:\project or ~/project), you traverse virtual directories with a checkout.

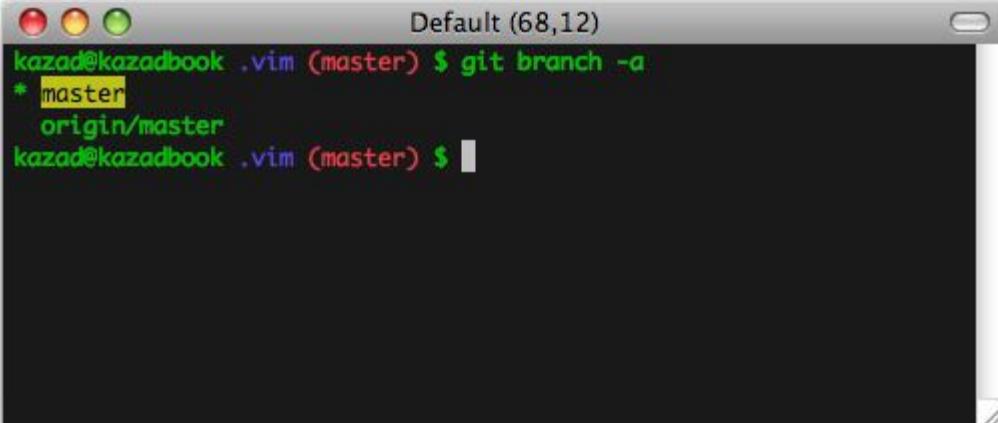
- git checkout master
 - switch to master branch ("cd master")
- git branch dev
 - create new branch from existing ("cp * dev")
 - you still need to "cd" with "git checkout dev"
- git merge dev
 - (when in master) pull in changes from dev ("cp dev/* .")
- git branch
 - list all branches ("ls")

Know the Current Branch

In my .bash_profile:

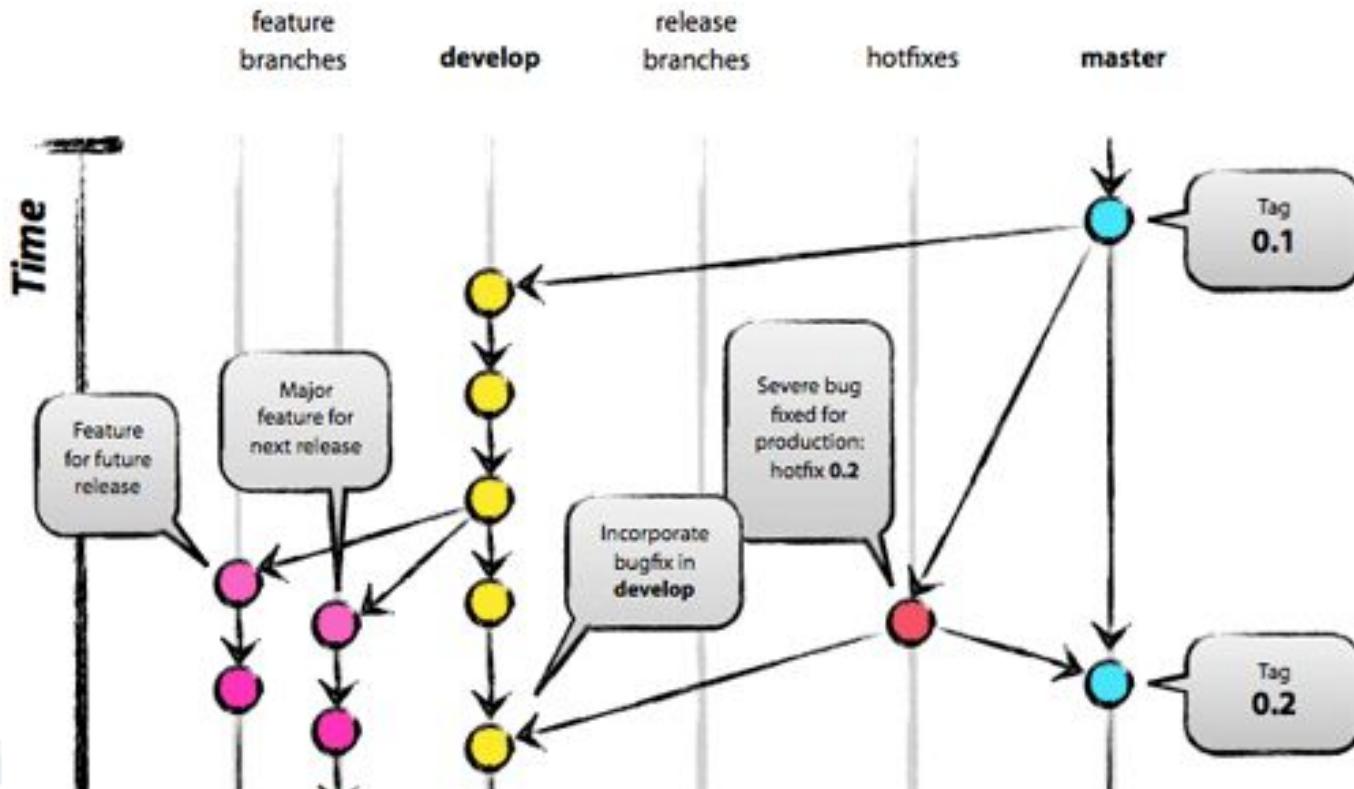
```
parse_git_branch() {  
    git branch 2> /dev/null | sed -e '/^[\^*]/d' -e 's/* (.*)/(1)/*'  
}  
}
```

```
export PS1="[33[00m]u@h[33[01;34m] W [33[31m]$(parse_git_branch)  
[33[00m]$[33[00m] "
```



A screenshot of a terminal window titled "Default (68,12)". The window shows the command `git branch -a` being run. The output indicates that the current branch is "master", which is highlighted with a yellow background. Other branches listed are "origin/master". The terminal prompt is `kazad@kazadbook .vim (master) $`.

Always Visualize your Branches



Git has Local AND Remote

Local data

- `git init`
 - create local repo
 - use `git add/commit/branch` to work locally

Remote data

- `git remote add name path-to-repo`
 - track a remote repo (usually "origin") from an existing repo
 - remote branches are "origin/master", "origin/dev" etc.
- `git branch -a`
 - list all branches (remote and local)
- `git clone path-to-repo`
 - create a new local git repo copied from a remote one
 - local master tracks remote master
- `git pull`
 - merge changes from tracked remote branch (if in dev, pull from origin/dev)
- `git push`
 - send changes to tracked remote branch (if in dev, push to origin/dev)

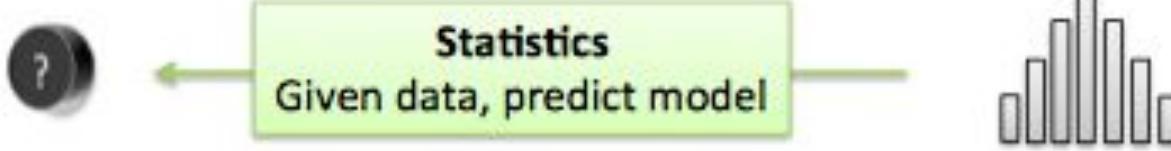
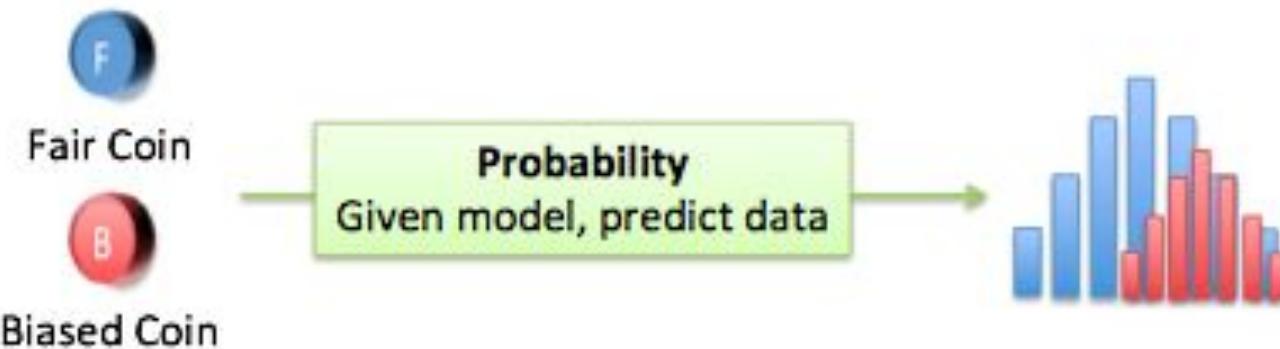
GUIDs

Statistics

Difference Between Probability & Statistics

- **Probability** is starting with an animal, and figuring out what footprints it will make.
- **Statistics** is seeing a footprint, and guessing the animal.

Probability & Statistics



Find the Animal

Get the tracks.

Measure the basic characteristics.

Find the species.

Look up the specific animal.

Make additional predictions.

Slingshot

"Statistics is the study of the collection, organization, analysis, and interpretation of data".

- What are the most common species? ([Common distributions](#))
- Are new ones being discovered?
- Can we predict the next footprint? (Extrapolation)
- Are the tracks following a path? (Regression / trend line)
- Here's two tracks, which animal was faster? Bigger? (Data from two drug trials: which was more effective?)
- Is one animal moving in the same direction as another? (Correlation)
- Are two animals tracking a common source? (Causation: two bears chasing the same rabbit)

Cleaning Data

- Cleansing data
- Visualizing data for preliminary analysis
- Understanding unbalanced datasets



Titanic

<https://github.com/fenago/dataanalysiswithexcel>



File Home Insert Page Layout Formulas Data Review View Add-ins Power Pivot PDF-XChange Team Tell me what you want to do

Get Data ->

- From Text/CSV
- Recent Sources
- Existing Connections
- Refresh All
- Properties
- Edit Links

Sort & Filter

Sort

Filter

Advanced

Text to Columns

What-If Analysis

Forecast Sheet

Group

Ungroup

Subtotal

Outline

Analysis

Connections

From File

From Workbook

From Database

From Text/CSV

From Azure

From XML

From Online Services

From JSON

From Other Sources

From Folder

Combine Queries

From SharePoint Folder

Launch Query Editor...

Data Catalog Search

My Data Catalog Queries

Data Source Settings...

Query Options

Passenger data - Query Editor

File Home Transform Add Column View

Close & Load Refresh Advanced Editor Choose Columns Remove Columns Keep Rows Remove Rows Sort Split Column Group By Data Type: Text Use First Row as Headers Merge Queries Append Queries Combine Files Manage Parameters Data source settings New Source Recent Sources Close Query Manage Columns Reduce Rows Replace Values Parameters Data Sources New Query

Queries

	1 ² 3 pclass	1 ² 3 survived	A ^B C sex	1.2 age	1 ² 3 sibsp	1 ² 3 parch	A ^B C ticket	1.2 fare	A ^B C cabin	A ^B C embark
1		1	1 female		29	0	0 24160	211.3375	B5	S
2		1	1 male	0.9167		1	2 113781	151.55	C22 C26	S
3		1	0 female		2	1	2 113781	151.55	C22 C26	S
4		1	0 male		30	1	2 113781	151.55	C22 C26	S
5		1	0 female		25	1	2 113781	151.55	C22 C26	S
6		1	1 male		48	0	0 19952	26.55	E12	S
7		1	1 female		63	1	0 13502	77.9583	D7	S
8		1	0 male		39	0	0 112050	0	A36	S
9		1	1 female		53	2	0 11769	51.4792	C101	S
10		1	0 male		71	0	0 PC 17609	49.5042	null	C
11		1	0 male		47	1	0 PC 17757	227.525	C62 C64	C
12		1	1 female		18	1	0 PC 17757	227.525	C62 C64	C
13		1	1 female		24	0	0 PC 17477	69.3	B35	C
14		1	1 female		26	0	0 19877	78.85	null	S
15		1	1 male		80	0	0 27042	30	A23	S
16		1	0 male		null	0	0 PC 17318	25.925	null	S
17		1	0 male		24	0	1 PC 17558	247.5208	B58 B60	C
18		1	1 female		50	0	1 PC 17558	247.5208	B58 B60	C
19		1	1 female		32	0	0 11813	76.2917	D15	C
20		1	0 male		36	0	0 13050	75.2417	C6	C
21		1	1 male		37	1	1 11751	52.5542	D35	S
22		1	1 female		47	1	1 11751	52.5542	D35	S
23		1	1 male		26	0	0 111369	30	C148	C
24		1	1 female		42	0	0 PC 17757	227.525	null	C
25										

Query Settings

PROPERTIES

- Name: Passenger data
- All Properties

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Removed Columns

Passenger data - Query Editor

File Home Transform Add Column View

Close & Load Refresh Preview Advanced Editor Choose Columns Remove Columns Keep Rows Remove Rows Split Column Group By Data Type: Text Merge Queries Append Queries Combine Files Manage Parameters Data source settings New Source Recent Sources Close Query Manage Columns Reduce Rows Sort Use First Row as Headers Replace Values Combine Parameters Data Sources New Query

Queries

1 2 3 pclass 1 2 3 survived A^BC sex 1.2 age 1 2 3 sibsp 1 2 3 parch A^BC ticket 1.2 fare A^BC cabin A^BC embac

1		1		1	female		29		0		0	24160		211.3375	B5	S	
2		1		1	male		0.9167		1		2	113781		151.55	C22 C26	S	
3		1		0	female			2		1		2	113781		151.55	C22 C26	S
4		1		0	male		30		1		2	113781		151.55	C22 C26	S	
5		1		0	female		25		1		2	113781		151.55	C22 C26	S	
6		1		1	male		48		0		0	19952		26.55	E12	S	
7		1		1	female		63		1		0	13502		77.9583	D7	S	
8		1		0	male		39		0		0	112050		0	A36	S	
9		1		1	female		53		2		0	11769		51.4792	C101	S	
10		1		0	male		71		0		0	PC 17609		49.5042	unknown	C	
11		1		0	male		47		1		0	PC 17757		227.525	C62 C64	C	
12		1		1	female		18		1		0	PC 17757		227.525	C62 C64	C	
13		1		1	female		24		0		0	PC 17477		69.3	B35	C	
14		1		1	female		26		0		0	19877		78.85	unknown	S	
15		1		1	male		80		0		0	27042		30	A23	S	
16		1		0	male		null		0		0	PC 17318		25.925	unknown	S	
17		1		0	male		24		0		1	PC 17558		247.5208	B58 B60	C	
18		1		1	female		50		0		1	PC 17558		247.5208	B58 B60	C	
19		1		1	female		32		0		0	11813		76.2917	D15	C	
20		1		0	male		36		0		0	13050		75.2417	C6	C	
21		1		1	male		37		1		1	11751		52.5542	D35	S	
22		1		1	female		47		1		1	11751		52.5542	D35	S	
23		1		1	male		26		0		0	111369		30	C148	C	
24		1		1	female		42		0		0	PC 17757		227.525	unknown	C	
25																	

13 COLUMNS, 999+ ROWS

Query Settings

PROPERTIES

Name
Passenger data
All Properties

APPLIED STEPS

Source
Navigation
Promoted Headers
Changed Type
Removed Columns
X Replaced Value

Passenger data - Query Editor

File Home Transform Add Column View

Column From Custom Invoke Custom Function Examples General

Conditional Column Index Column Duplicate Column

Merge Columns ABC Extract 123 Extract Parse

Format Statistics Standard Scientific Information

From Text From Number

Date Time Duration

From Date & Time

Queries [1] Passenger data

Custom Column

New column name: boat_corrected

Custom column formula:

```
= if [survived]=1 and [boat] = null then "unknown" else [boat]
```

Available columns:

- pclass
- survived
- sex
- age
- sibsp
- parch
- ticket

<< Insert

Learn about Power Query formulas

✓ No syntax errors have been detected.

OK Cancel

Query Settings

PROPERTIES

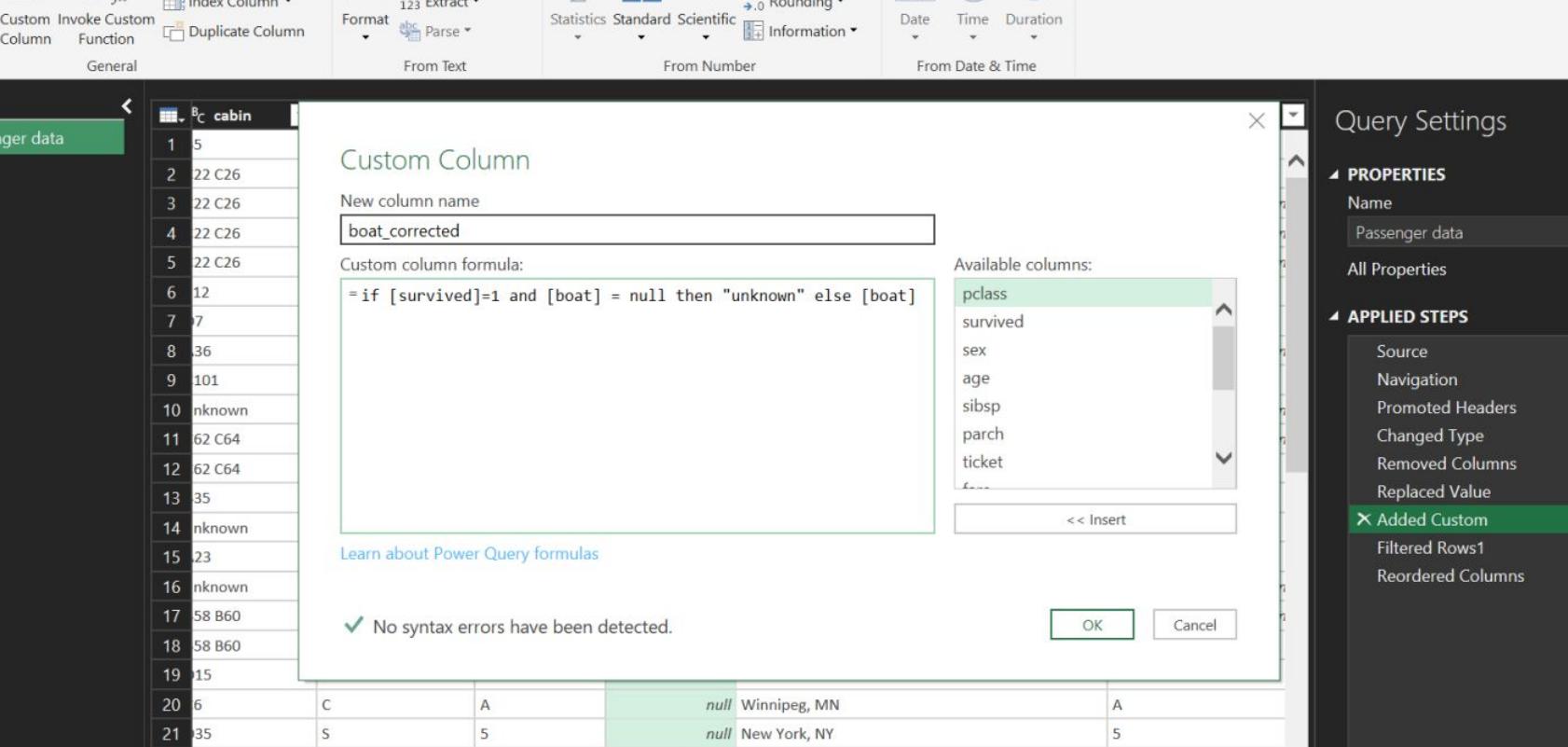
Name: Passenger data

All Properties

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Removed Columns
- Replaced Value
- Added Custom**

Filtered Rows1
Reordered Columns



Passenger data - Query Editor

File Home Transform Add Column View

Column From Custom Invoke Custom Examples Column Function General

Merge Columns Statistics Standard Scientific Trigonometry Rounding Information Date Time Duration

Conditional Column Index Column Duplicate Column Format ABC 123 Extract ABC Parse From Text From Number From Date & Time

Queries [1]

Passenger data

ranked	ABC boat	ABC 123 boat_corrected	123 body	ABC 123 body_corrected	ABC home.dest
1	2	2		null	St Louis, MO
2	11	11		null	Montreal, PQ / Chesterville, ON
3		null		null	Montreal, PQ / Chesterville, ON
4		null		135	Montreal, PQ / Chesterville, ON
5		null		null	Montreal, PQ / Chesterville, ON
6	3	3		null	New York, NY
7	10	10		null	Hudson, NY
8		null		null	Belfast, NI
9	D	D		null	Bayside, Queens, NY
10		null		22	Montevideo, Uruguay
11		null		124	New York, NY
12	4	4		null	New York, NY
13	9	9		null	Paris, France
14	6	6		null	
15	B	B		null	Hessle, Yorks
16		null		null	New York, NY
17		null		null	Montreal, PQ
18	6	6		null	Montreal, PQ
19	8	8		null	
20	A	A		null	Winnipeg, MN
21	5	5		null	New York, NY
22	5	5		null	New York, NY
23	5	5		null	New York, NY
24	4	4		null	
25					

Query Settings

PROPERTIES

- Name: Passenger data
- All Properties

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Removed Columns
- Replaced Value
- Added Custom
- Filtered Rows1
- Reordered Columns
- Added Custom1
- Reordered Columns1

15 COLUMNS, 999+ ROWS PREVIEW DOWNLOADED AT 16:03

Passenger data - Query Editor

File **Home** **Transform** **Add Column** **View**

Close & Load **Refresh Preview** **Advanced Editor** **Properties** **Choose Columns** **Remove Columns** **Keep Rows** **Remove Rows** **Sort** **Split Column** **Group By** **Data Type: Any** **Merge Queries** **Append Queries** **Combine Files** **Manage Parameters** **Parameters** **New Source** **Recent Sources**

Close **Query** **Manage Columns** **Reduce Rows** **Transform** **Combine** **Data source settings** **New Query**

Queries [1]

Passenger data

ranked	ABC boat	ABC 123 boat_corrected	123 body	ABC 123 body_corrected	ABC home.dest
1	2	2		null	N/A
2	11	11		null	N/A
3		null	N/A	null	not recovered
4		null	N/A	135	
5		null	N/A	null	not recovered
6	3	3		null	N/A
7	10	10		null	N/A
8		null	N/A	null	not recovered
9	D	D		null	N/A
10		null	N/A	22	
11		null	N/A	124	
12	4	4		null	N/A
13	9	9		null	N/A
14	6	6		null	N/A
15	B	B		null	N/A
16		null	N/A	null	not recovered
17		null	N/A	null	not recovered
18	6	6		null	N/A
19	8	8		null	N/A
20	A	A		null	not recovered
21	5	5		null	N/A
22	5	5		null	N/A
23	5	5		null	N/A
24	4	4		null	N/A
25					

Query Settings

▲ PROPERTIES

Name

Passenger data

All Properties

▲ APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Removed Columns
- Replaced Value
- Added Custom
- Filtered Rows1
- Reordered Columns
- Added Custom1
- Reordered Columns1
- Replaced Value1



Add Conditional Column

Add a conditional column that is computed from the other columns or values.

New column name

Age group

	Column Name	Operator	Value ⓘ		Output ⓘ	
If	age	equals	ABC 123	-1	Then	ABC 123
Else If	age	is less than	ABC 123	1	Then	ABC 123
Else If	age	is less than	ABC 123	12	Then	ABC 123
Else If	age	is less than	ABC 123	18	Then	ABC 123
Else If	age	is less than	ABC 123	65	Then	ABC 123
Else If	age	is greater than or...	ABC 123	65	Then	ABC 123

Add rule

Otherwise ⓘ

ABC
123

unknown

OK

Cancel

Passenger data - Query Editor

File Home Transform Add Column View

Close & Load Refresh Preview Advanced Editor Choose Columns Remove Columns Keep Rows Remove Rows Split Column Group By Data Type: Decimal Number Merge Queries Append Queries Combine Files Manage Parameters Data source settings New Source Recent Sources Close Query Manage Columns Reduce Rows Sort Use First Row as Headers Replace Values Transform Combine Parameters Data Sources New Query

Queries

	1 ² 3	pclass	1 ² 3	survived	A ^B C	sex	1.2	age	A ^B C	123	Age group	1 ² 3	sibsp	1 ² 3	parch	A ^B C	ticket	1.2	fare	A ^B C	ca
1				1		female			29	adult			0		0	24160		211.3375	B5		
2				1		male			0.9167	infant			1		2	113781		151.55	C2		
3				0		female			2	child			1		2	113781		151.55	C2		
4				0		male			30	adult			1		2	113781		151.55	C2		
5				0		female			25	adult			1		2	113781		151.55	C2		
6				1		male			48	adult			0		0	19952		26.55	E1		
7				1		female			63	adult			1		0	13502		77.9583	D7		
8				0		male			39	adult			0		0	112050		0	A3		
9				1		female			53	adult			2		0	11769		51.4792	C10		
10				0		male			71	elderly			0		0	PC 17609		49.5042	unl		
11				1		male			47	adult			1		0	PC 17757		227.525	C6		
12				1		female			18	adult			1		0	PC 17757		227.525	C6		
13				1		female			24	adult			0		0	PC 17477		69.3	B3		
14				1		female			26	adult			0		0	19877		78.85	unl		
15				1		male			80	elderly			0		0	27042		30	A2		
16				0		male			-1	unknown			0		0	PC 17318		25.925	unl		
17				1		male			24	adult			0		1	PC 17558		247.5208	B5		
18				1		female			50	adult			0		1	PC 17558		247.5208	B5		
19				1		female			32	adult			0		0	11813		76.2917	D1		
20				0		male			36	adult			0		0	13050		75.2417	C6		
21				1		male			37	adult			1		1	11751		52.5542	D3		
22				1		female			47	adult			1		1	11751		52.5542	D3		
23				1		male			26	adult			0		0	111369		30	C1		
24				1		female			42	adult			0		0	PC 17757		227.525	unl		
25																					

16 COLUMNS, 999+ ROWS

PREVIEW DOWNLOADED AT 18:00

Query Settings

PROPERTIES

Name

Passenger data

All Properties

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Removed Columns
- Replaced Value
- Added Custom
- Filtered Rows1
- Reordered Columns
- Added Custom1
- Reordered Columns1
- Replaced Value1
- Replaced Value2
- Added Custom2
- Reordered Columns2

Visualize Data!



B2																				
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
1	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat_corrected	boat	body_corrected	boat_corrected2					
2	1	1	1 female	29 adult		0	0	24160	211.3375	B5	S	2	2	0	2					
3	1					1	2	113781	151.55	C22 C26	S	11	11	0	11					
4	1					1	2	113781	151.55	C22 C26	S		0	unknown	N/A					
5	1					1	2	113781	151.55	C22 C26	S		0		135	N/A				
6	1					1	2	113781	151.55	C22 C26	S		0	unknown	N/A					
7	1					0	0	19952	26.55	E12	S	3	3	0	3					
8	1					1	0	13502	77.9583	D7	S	10	10	0	10					
9	1					0	0	112050	0 A36	S		0	unknown	N/A						
10	1					2	0	11769	51.4792	C101	S	D	D	0	D					
11	1					0	0	PC 17609	49.5042	unknown	C		0		22	N/A				
12	1					1	0	PC 17757	227.525	C62 C64	C		0		124	N/A				
13	1					1	0	PC 17757	227.525	C62 C64	C	4	4		0	4				
14	1					0	0	PC 17477	69.3	B35	C	9	9		0	9				
15	1					0	0	19877	78.85	unknown	S	6	6		0	6				
16	1					0	0	27042	30 A23	S	B	B	B	0	B					
17	1					0	0	PC 17318	25.925	unknown	S		0	unknown	N/A					
18	1					0	1	PC 17558	247.5208	B58 B60	C		0	unknown	N/A					
19	1					0	1	PC 17558	247.5208	B58 B60	C	6	6		0	6				
20	1					0	0	11813	76.2917	D15	C	8	8		0	8				
21	1					0	0	13050	75.2417	C6	C	A	A	unknown	A					
22	1					1	1	11751	52.5542	D35	S	5	5		0	5				
23	1					1	1	11751	52.5542	D35	S	5	5		0	5				
24	1					0	0	111369	30 C148	C	5	5	5		0	5				
25	1					0	0	PC 17757	227.525	unknown	C	4	4		0	4				
26	1	1	1 female	29 adult		0	0	PC 17483	221.7792	C97	S	8	8		0	8				
27	1	0	0 male	25 adult		0	0	13905	26	unknown	C		0		148	N/A				
28	1	1	1 male	25 adult		1	0	11967	91.0792	B49	C	7	7		0	7				

Formatting | **Charts** | Totals | Tables | Sparklines

Clustered Column Clustered Column Clustered Column Clustered Column Clustered Column More Charts

Recom Use this chart type to:
• Compare values across a few categories.
Use it when:
• The order of categories is not important.

titanic_book.xlsx - Excel

PivotCh... Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

A B C D E F G H I J K L M

1

2

3 Age group Sum of age

4 adult 28943.5

5 child 417.5

6 elderly 910.5

7 infant 8.1667

8 teenager 976

9 unknown -263

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

Sum of age by Age group

35000
30000
25000
20000
15000
10000
5000
0
-5000

adult child elderly infant teenager unknown

Age group ▾

PivotChart Fields

Choose fields to add to report:

Search

pclass
 survived
 sex
 age
 Age group
 sibsp
 parch
 ticket
 fare

Move Up
Move Down
Move to Beginning
Move to End
Move to Report Filter
Move to Axis Fields (Categories)
Move to Legend Fields (Series)
Move to Values
Hide Value Field Buttons on Chart
Hide All Field Buttons on Chart
Remove Field

Value Field Settings...

Axis (Categories)

Age group

Sum of age

Defer Layout Update

Update

Sheet1 Sheet2 +

Ready Calculate

100%

Value Field Settings

?

X

Source Name: age

Custom Name: Count of age

Summarize Values By

Show Values As

Summarize value field by

Choose the type of calculation that you want to use to summarize data from the selected field

Sum

Count

Average

Max

Min

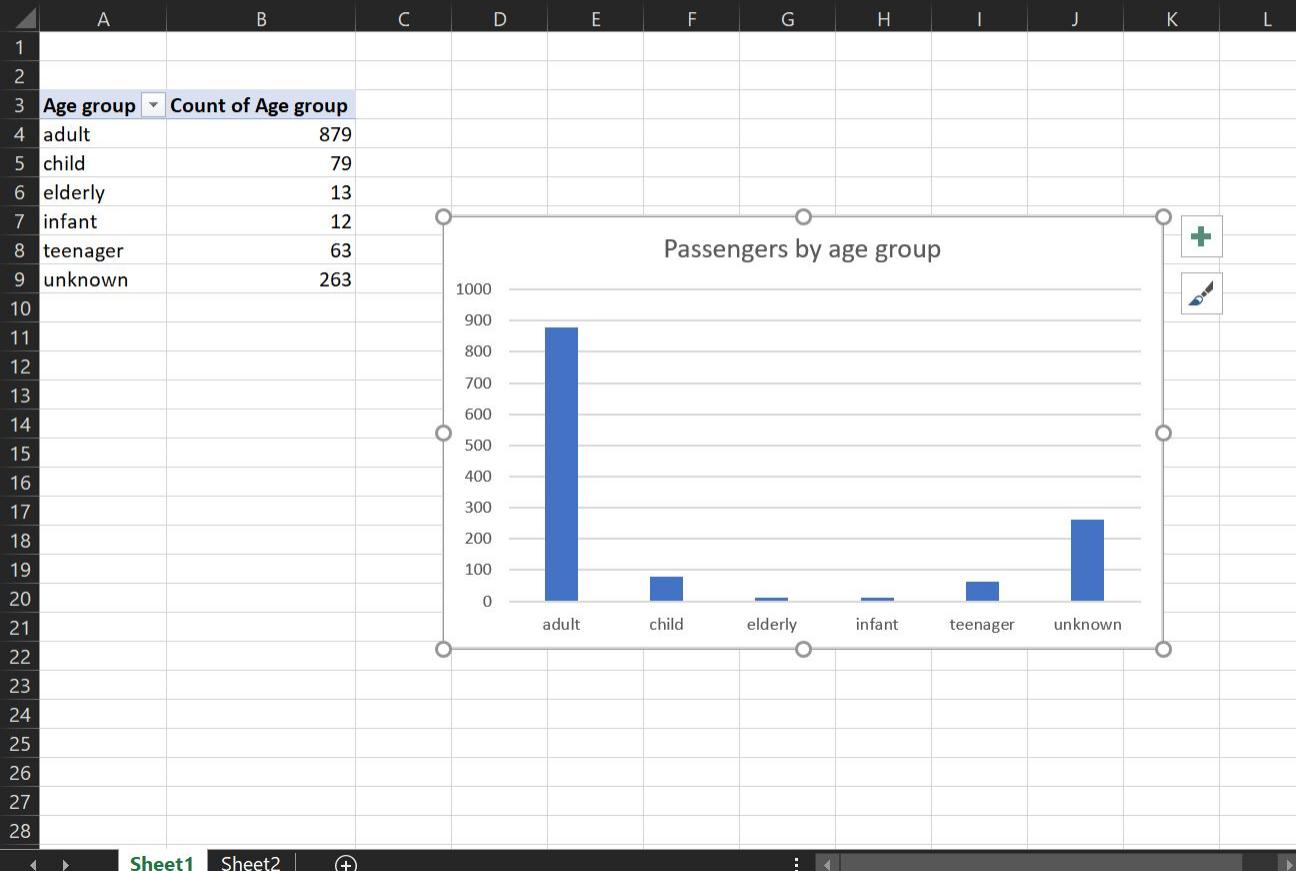
Product

Number Format

OK

Cancel

Chart 1

**PivotChart Fields**

Choose fields to add to report:

Search

- pclass
- survived
- sex
- age
- Age group
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

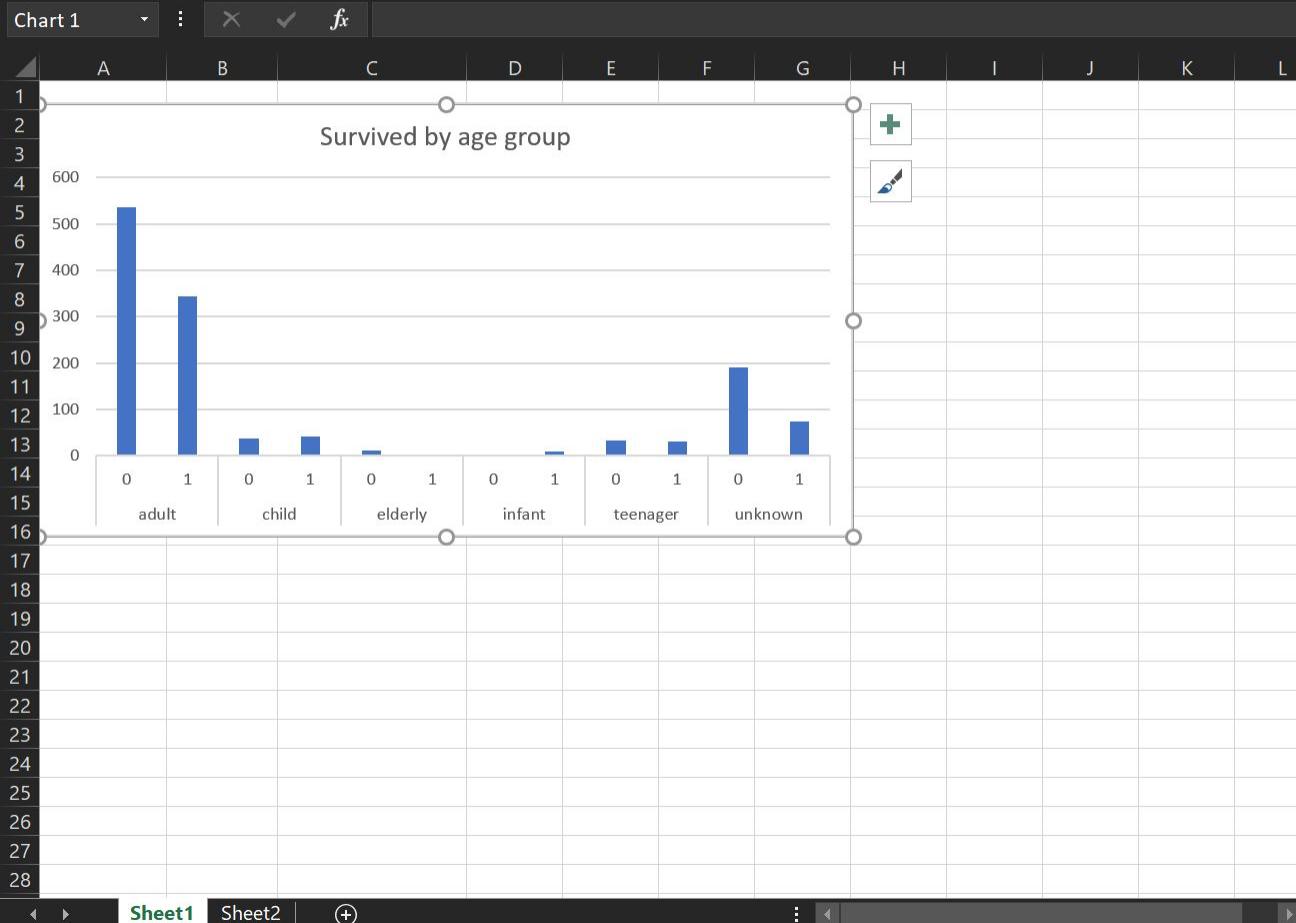
Age group

Values

Count of Age group

Defer Layout Update

Update



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- survived
- sex
- age
- Age group
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters

Legend (Series)
Axis (Categories)
Age group

Values

Count of Age group

Defer Layout Update

Update

Value Field Settings

?

X

Source Name: Age group

Custom Name: Count of Age group

Summarize Values By

Show Values As

Show values as

% of Parent Total



Base field:

pclass
survived
sex
age
Age group
sibsp

Base item:



Number Format

OK

Cancel

titanic_book.xlsx - Excel

PivotCh... Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

	A	B	C	D	E	F	G	H	I	J	K	L
3	Age group	survived	Count of Age group									
4	adult	0	60.86%									
5	adult	1	39.14%									
6	child	0	48.10%									
7	child	1	51.90%									
8	elderly	0	84.62%									
9	elderly	1	15.38%									
10	infant	0	16.67%									
11	infant	1	83.33%									
12	teenager	0	52.38%									
13	teenager	1	47.62%									
14	unknown	0	72.24%									
15	unknown	1	27.76%									

PivotChart Fields

Choose fields to add to report:

Search

pclass
 survived
 sex
 age
 Age group
 sibsp
 parch
 ticket
 fare

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories) Values

Age group Count of Age group

Defer Layout Update

Update

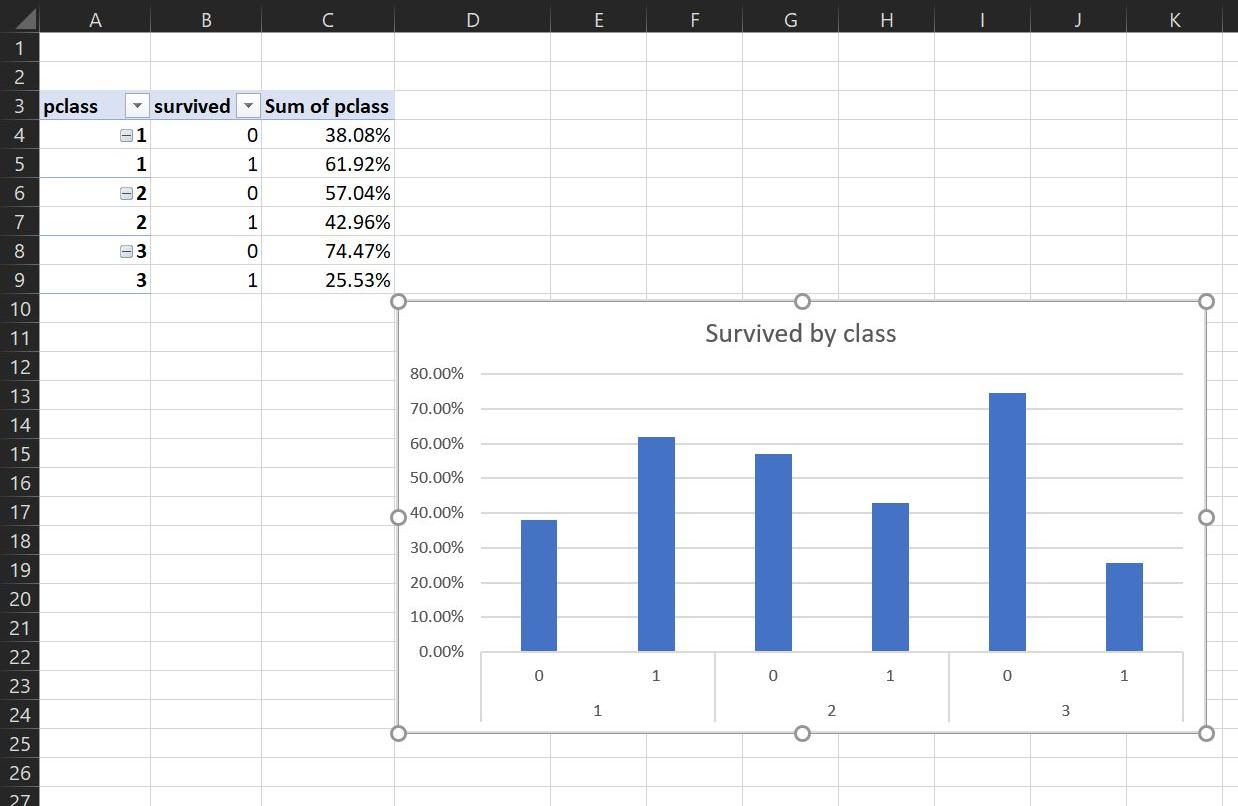
Survived by age group

Age group	0	1	survived (%)
adult	0	1	~39.14%
child	0	1	~51.90%
elderly	0	1	~84.62%
infant	0	1	~16.67%
teenager	0	1	~47.62%
unknown	0	1	~27.76%

Sheet1 Sheet2

Ready

Chart 1



PivotChart Fields

Choose fields to add to report:

Search

 survived sex age Age group sibsp parch ticket fare cabin

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

pclass

survived

Σ Values

Sum of pclass

Defer Layout Update

Update

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

titanic_book.xlsx - Excel

PivotCh... Julio C Rodriguez Martino

Chart 1

	A	B	C	D	E	F	G	H	I	J	K	L
3	sex	survived	Count of sex									
4	fema	0	27.25%									
5	fema	1	72.75%									
6	male	0	80.90%									
7	male	1	19.10%									
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												
25												
26												
27												
28												
29												
30												

Survived by sex

The chart displays the percentage of survivors ('survived') for two categories: 'female' and 'male'. The y-axis represents the percentage from 0.00% to 90.00%. For females ('female'), the survival rates are 27.25% (labeled 0) and 72.75% (labeled 1). For males ('male'), the survival rates are 80.90% (labeled 0) and 19.10% (labeled 1).

Category	Survived	Percentage (%)
female	0	27.25%
female	1	72.75%
male	0	80.90%
male	1	19.10%

PivotChart Fields

Choose fields to add to report:

Search

- boat_corrected
- boat
- body_corrected
- boat_corrected2
- body
- body_corrected3
- home.dest
- % Total

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

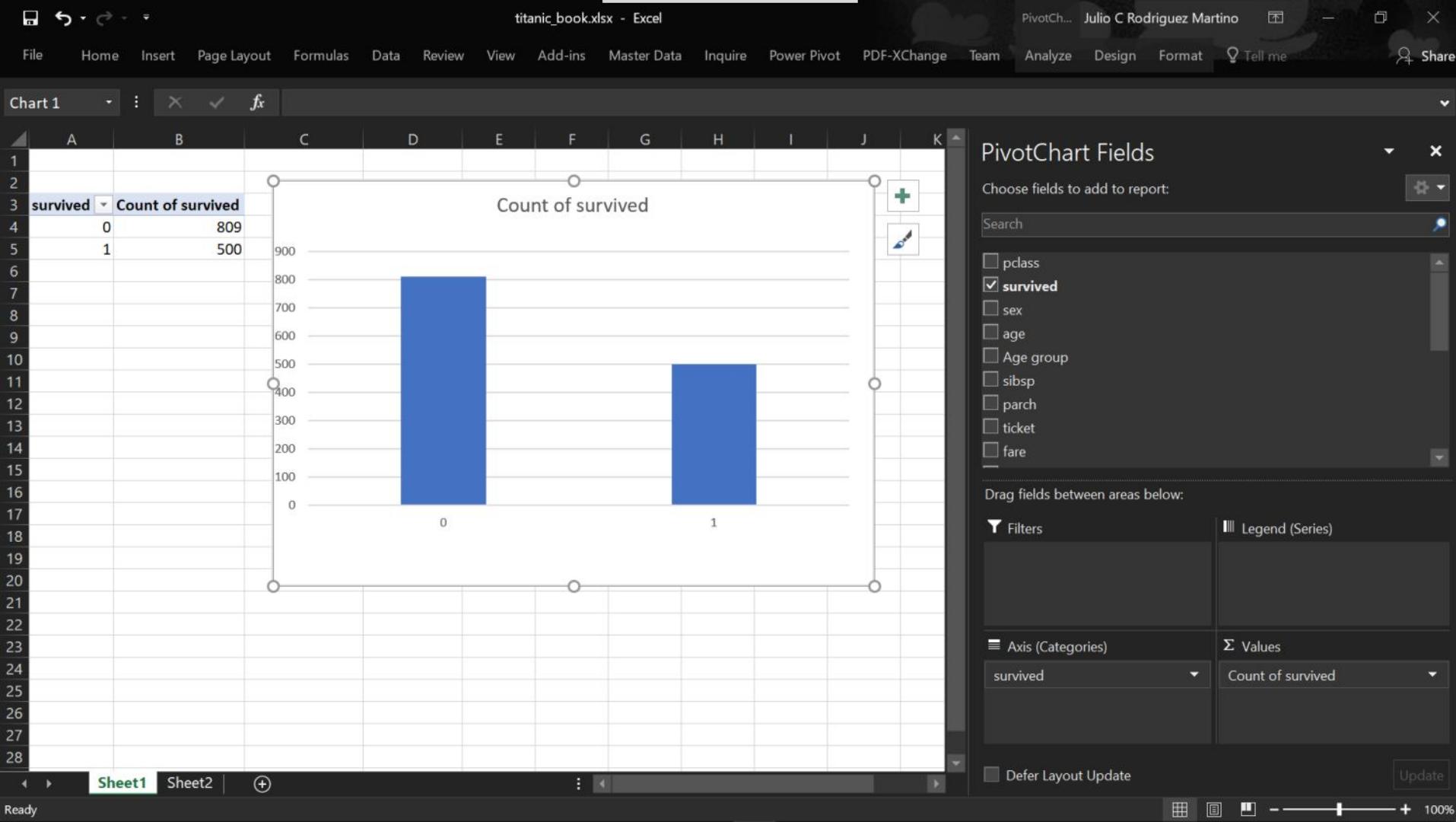
Values

Defer Layout Update

Update

Bias and Balance



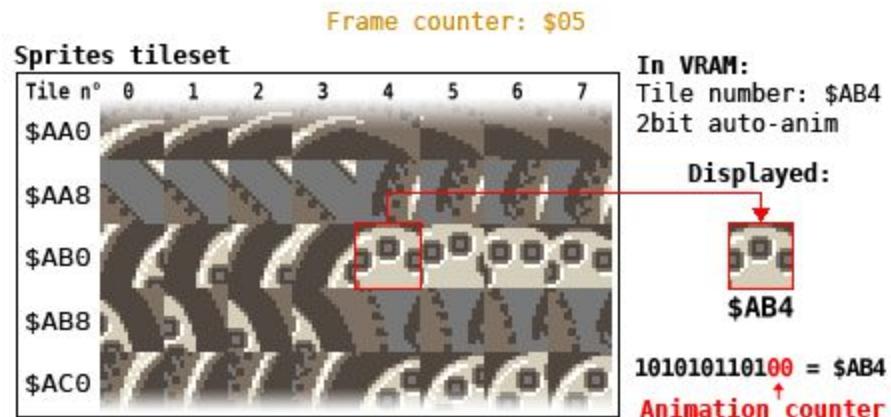


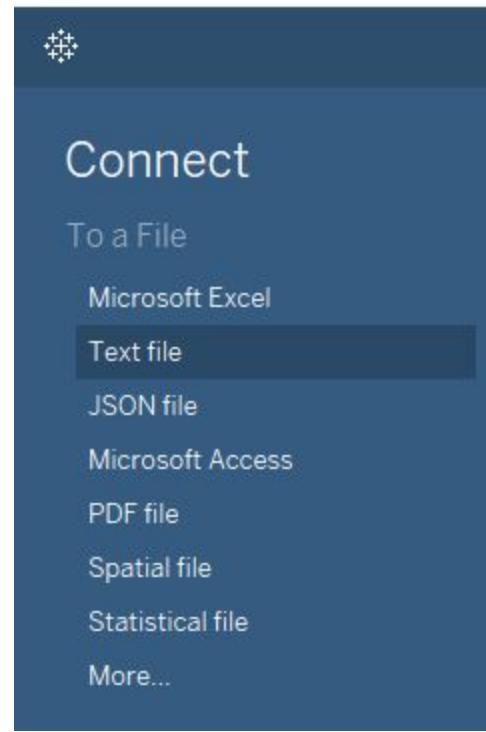
J41 B35

	pclass	survived	sex	age	Age group	sibsp	parch	ticket	fare	cabin	embarked	boat_corrected	boat	body_corrected	boat_corrected2
Sort Smallest to Largest				adult		1	0	17474		57 B20	S	3	3		0 3
Sort Largest to Smallest				teenager		1	0	17474		57 B20	S	3	3		0 3
Sort by Color				elderly		1	1	WE/P 5735		71 B22	S	0		269 N/A	
Clear Filter From "survived"				adult		0	2	WE/P 5735		71 B22	S	7	7		0 7
Filter by Color				adult		0	0	12749		93.5 B24	S	0	unknown	N/A	
Number Filters				adult		1	1	112901		26.55 B26	S	7	7		0 7
Search				adult		0	0	113572		80 B28		6	6		0 6
				adult		0	0	113572		80 B28		6	6		0 6
				adult		0	1	24160		211.3375 B3	S	2	2		0 2
				elderly		0	1	113509		61.9792 B30	C	0		234 N/A	
				adult		0	0	PC 17477		69.3 B35	C	9	9		0 9
				adult		0	0	PC 17477		69.3 B35	C	9	9		0 9
				adult		0	1	113509		61.9792 B36	C	5	5		0 5
				adult		0	0	11771		29.7 B37	C	0		258 N/A	
				adult		0	0	113050		26.55 B38	S	0	unknown	N/A	
				adult		0	2	13568		49.5 B39	C	5	5		0 5
47	1	1 female	44	adult		0	0	PC 17610		27.7208 B4	C	6	6		0 6
48	1	1 male	60	adult		1	1	13567		79.2 B41	C	5	5		0 5
49	1	1 female	48	adult		1	1	13567		79.2 B41	C	5	5		0 5
50	1	1 female	19	adult		0	0	112053		30 B42	S	3	3		0 3
51	1	1 male	24	adult		1	0	21228		82.2667 B45	S	7	7		0 7
52	1	1 female	23	adult		1	0	21228		82.2667 B45	S	7	7		0 7
53	1	1 male	25	adult		1	0	11967		91.0792 B49	C	7	7		0 7
54	1	1 female	19	adult		1	0	11967		91.0792 B49	C	7	7		0 7
55	1	1 female	29	adult		0	0	24160		211.3375 B5	S	2	2		0 2
56	1	1 female	15	teenager		0	1	24160		211.3375 B5	S	2	2		0 2
57	1	1 male	32	adult		0	0	13214		30.5 B50	C	3	3		0 3
58	1	1 male	36	adult		0	1	PC 17755		512.3292 B51 B53 B55	C	3	3		0 3

Data Analysis Report										
Demographic Data		Health Metrics			Performance Indicators			Financial Summary		
Index	ID	Gender	Age Group	Height (cm)	Weight (kg)	BMI	Score	Rank	Total Sales	Avg Profit
1	0.754856	2	0 male	32 adult	0	0 237216	13.5 unknown S	0	0	0
2	0.678131	3	0 male	32 adult	0	0 STON/O 2	7.925 unknown S	0	0	0
3	0.352361	3	0 male	18 adult	0	0 349912	7.775 unknown S	0	0	0
4	0.512693	3	0 male	-1 unknown	0	0 A/5 2817	8.05 unknown S	0	0	0
5	0.549755	3	0 male	31 adult	0	0 21332	7.7333 unknown Q	0	0	0
6	0.267527	3	0 male	22 adult	0	0 350045	7.7958 unknown S	0	0	0
7	0.691454	3	0 female	37 adult	0	0 368364	7.75 unknown Q	0	0	0
8	0.531422	3	0 male	-1 unknown	0	0 2681	6.4375 unknown C	0	0	0
9	0.500349	3	0 male	28 adult	0	0 363611	8.05 unknown S	0	0	0
10	0.294431	3	0 female	21 adult	0	0 315087	8.6625 unknown S	0	0	0
11	0.508635	3	0 male	9 child	4	2 347077	31.3875 unknown S	0	0	0
12	0.928319	3	0 male	39 adult	1	5 347082	31.275 unknown S	0	0	0
13	0.886834	3	0 male	17 teenager	0	0 315095	8.6625 unknown S	0	0	0
14	0.978843	2	0 male	19 adult	1	1 C.A. 33112	36.75 unknown S	0	0	101 N/A
15	0.202766	1	0 male	58 adult	0	2 35273	113.275 D48 C	0	0	122 N/A
16	0.461524	3	0 male	-1 unknown	1	0 2689	14.4583 unknown C	0	0	unknown N/A
17	0.373138	1	0 male	47 adult	0	0 111320	38.5 E63 S	0	0	275 N/A
18	0.289079	1	0 male	55 adult	1	0 PC 17603	59.4 unknown C	0	0	unknown N/A
19	0.397568	2	0 female	18 adult	1	1 250650	13 unknown S	0	0	unknown N/A
20	0.167581	3	0 female	-1 unknown	0	0 364859	7.75 unknown Q	0	0	unknown N/A
21	0.526137	3	0 male	23 adult	1	0 347072	13.9 unknown S	0	0	unknown N/A
22	0.118658	3	0 male	28 adult	0	0 347464	7.8542 unknown S	0	0	unknown N/A
23	0.495476	3	0 male	51 adult	0	0 347064	7.75 unknown S	0	0	unknown N/A
24	0.851977	3	0 male	17 teenager	0	0 315086	8.6625 unknown S	0	0	unknown N/A
25	0.702529	3	0 male	31 adult	3	0 345763	18 unknown S	0	0	unknown N/A
26	0.291475	2	0 male	23 adult	0	0 29751	13 unknown S	0	0	unknown N/A
27	0.953934	3	0 female	-1 unknown	0	0 65305	8.1125 unknown S	0	0	unknown N/A

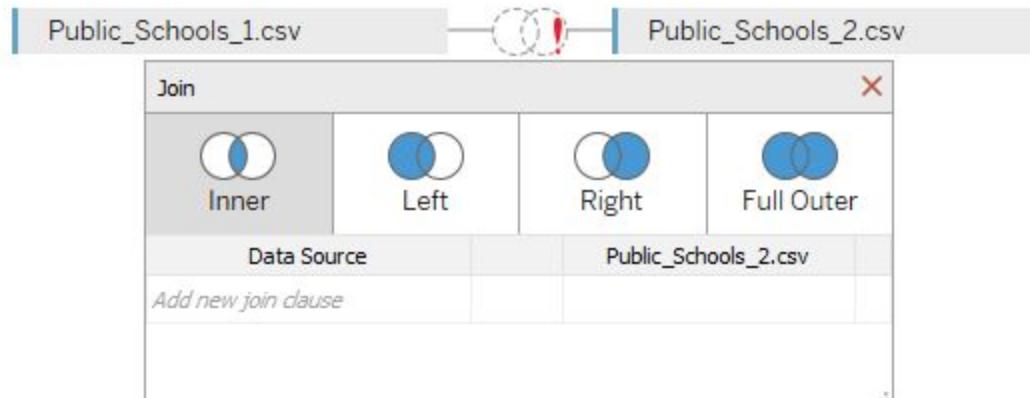
Data Manipulation



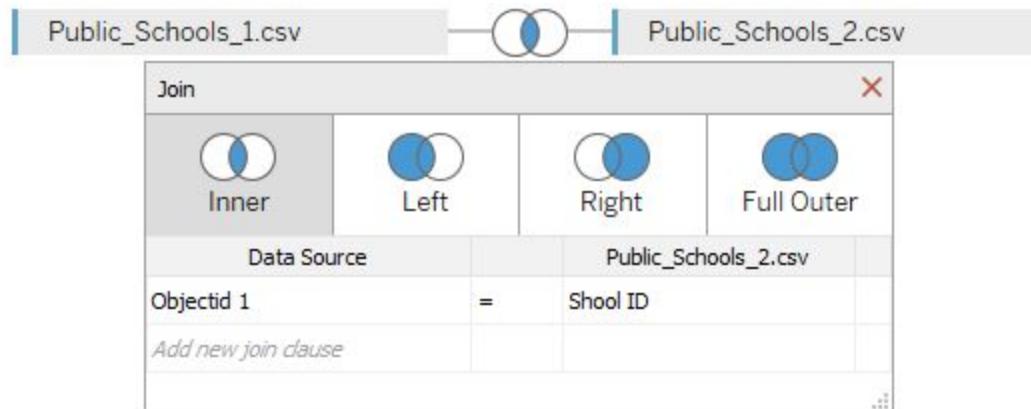


The screenshot shows a software interface for managing data connections and files. On the left, there's a sidebar with a 'Connections' section containing a single entry: 'Public_Schools_1' (Text File). Below this is a 'Files' section with a checkbox for 'Use Data Interpreter'. A note below the checkbox states: 'Data Interpreter might be able to clean your Text File workbook.' Under the 'Files' section, three files are listed: 'Internet_sati...by_region.csv', 'Public_Schools_1.csv', and 'Public_Schools_2.csv'. The 'Public_Schools_1.csv' file is highlighted with a blue selection bar. The main workspace on the right displays the title 'Public_Schools_1'.

↳ Public_Schools_1+



Public_Schools_1+





Sort fields

Data source order

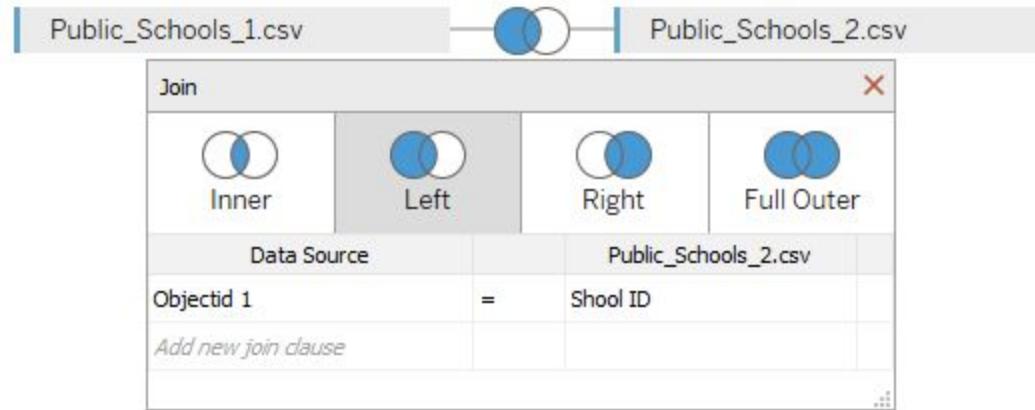
 Show aliases Show hidden fields

131

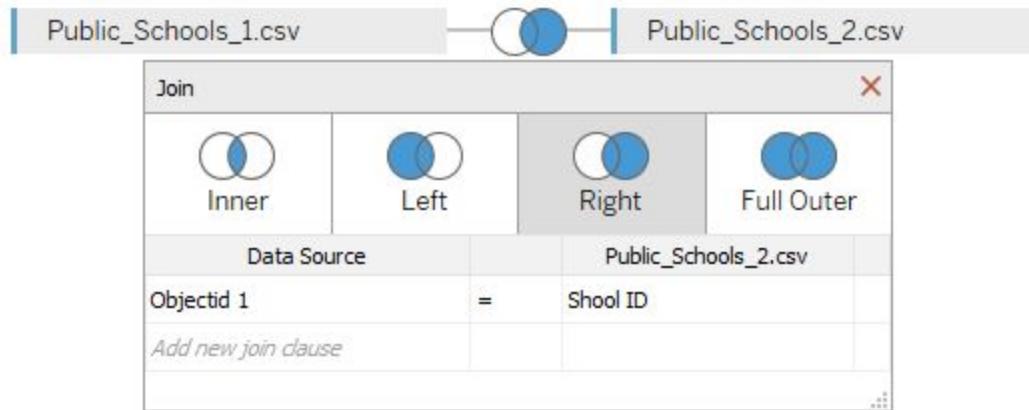


# Public_Schools_1.csv Objectid 1	# Public_Schools_1.... Objectid	# Public_Schools... Bldg Id	Abc Public_Schools_1.csv Bldg Name	Abc Public_Schools_1.csv Address	@ Public_Schools_1... City	@ Public_Schools_1... Zipcode	# Public_Schools_1.csv Csp Sch Id	# Public_Schools_1.... Sch Id
1	1	1	Guild Bldg	195 Leyden Street	East Boston	02128	4061	4061
2	2	3	Kennedy, P Bldg	343 Saratoga Street	East Boston	02128	4541	4541
3	3	4	Otis Bldg	218 Marion Street	East Boston	02128	4322	4322
4	4	6	Odonnell Bldg	33 Trenton Street	East Boston	02128	4543	4543
5	5	7	East Boston High Bldg	86 White Street	East Boston	02128	1070	1070
6	6	8	Umana / Barnes Bldg	312 Border Street	East Boston	02128	4323	4323
7	7	10	East Boston Eec Bldg	135 Gove Street	East Boston	02128	4450	4450
8	8	11	Mckay Bldg	122 Cottage Street	East Boston	02128	4360	4360
9	9	12	Adams Bldg	165 Webster Street	East Boston	02128	4361	4361
10	10	13	Harvard-Kent	50 Bunker Hill Street	Charlestown	02129	4280	4280
11	11	14	Charlestown High Bld...	240 Medford Street	Charlestown	02129	1050	1050
12	12	15	Edwards Bldg	28 Walker Street	Charlestown	02129	2010	2010
13	13	16	Warren-Prescott Bldg	50 School Street	Charlestown	02129	4283	4283

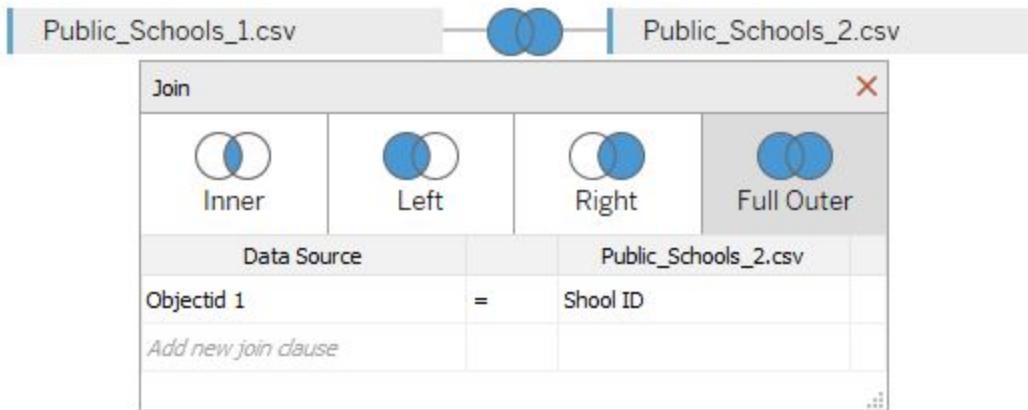
- Public_Schools_1+



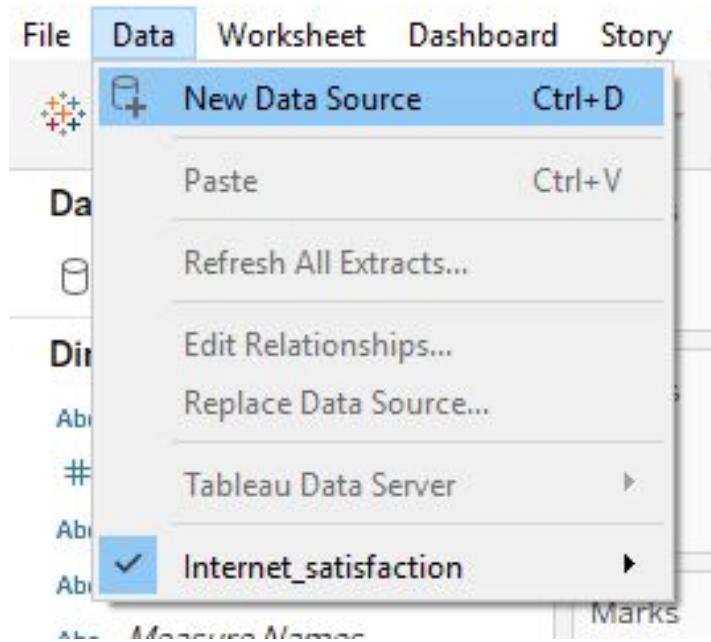
Public_Schools_1+



Public_Schools_1+



Adding Secondary Sources



Data	Analytics
<input type="checkbox"/> Internet_satisfaction	
<input type="checkbox"/> Internet_usage	

Marks

Automatic

Color Size Text

Detail Tooltip

SUM(Satisf..)

Filter...

Show Filter

Format...

✓ Include in Tooltip

Dimension

Attribute

● Measure (Sum) ▾

- Sum
- Average
- Median
- Count
- Count (Distinct)

ADSL Cable internet

ADSL	Cable	internet
1,507	1,233	92

The screenshot shows the Tableau Data Prep interface. At the top, there are three buttons: 'Pages' (disabled), 'Columns' (disabled), and 'Rows' (disabled). Below these are two sections: 'Filters' and 'Marks'. The 'Marks' section includes a dropdown for 'Automatic' and several buttons for 'Color', 'Size', 'Text', 'Detail', 'Tooltip', and a button for 'AVG(Satisfacti..)' which is highlighted with a green rounded rectangle.

iii Columns

Rows

Pages

Filters

Marks

Automatic

Color

Size

Text

Detail

Tooltip

AVG(Satisfacti..)

Sheet 1

HH internet type

	Wireless		
	ADSL	Cable	internet
3.9554	4.0693	4.0000	

The screenshot shows the Tableau Data Prep interface. On the left, there are three panels: 'Pages' (empty), 'iii Columns' (with 'Settlement type' highlighted in blue), and 'Rows' (empty). Below these is a 'Filters' panel (empty) and a 'Marks' panel. The 'Marks' panel contains a dropdown menu set to 'Automatic' and several buttons: 'Color', 'Size', 'Text', 'Detail', and 'Tooltip'. At the bottom of the marks shelf is a button labeled 'AVG(Internet ..)'.

iii Columns

Settlement type

Rows

Pages

Filters

Marks

AVG(Internet ..)

Sheet 2

Settlement type	Rural	Urban
	47.53	69.77