Data sharing and version control

INFO6540

Week 2 - Jan 16, 2018

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Questions?

Week 1

- Data explosion
- Data literacy
- Tables
- Excel

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

What is data sharing?

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Scientific context

Scientific context

The practice of making data used for scholarly research available to other investigators.

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Scientific context

The practice of making data used for scholarly research available to other investigators.

Why is it important?

Coat of arms of the Royal Society
(https://royalsociety.org/)

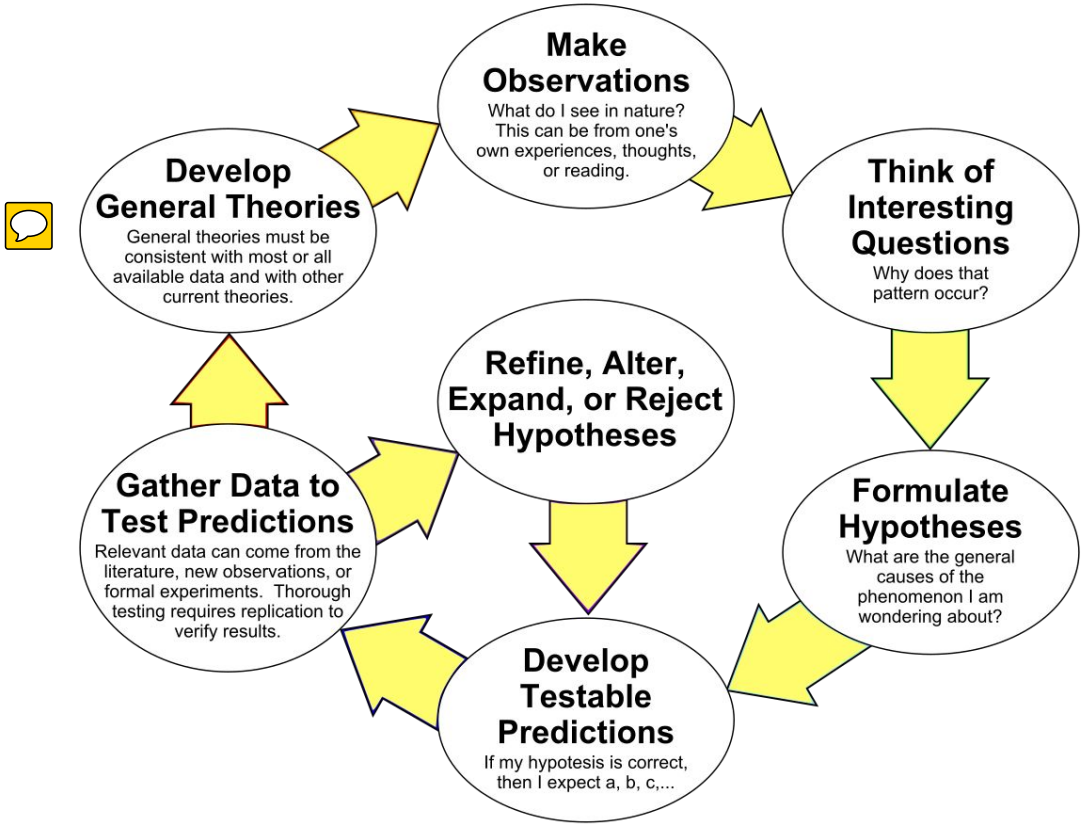
Take nobody's word for it.

Coat of arms of the Royal Society (https://royalsociety.org/)

Why is it important?

Many funding agencies and institutions have policies on data sharing, because openness and transparency are fundamental to the **scientific method**.

Scientific method



DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Scientific method

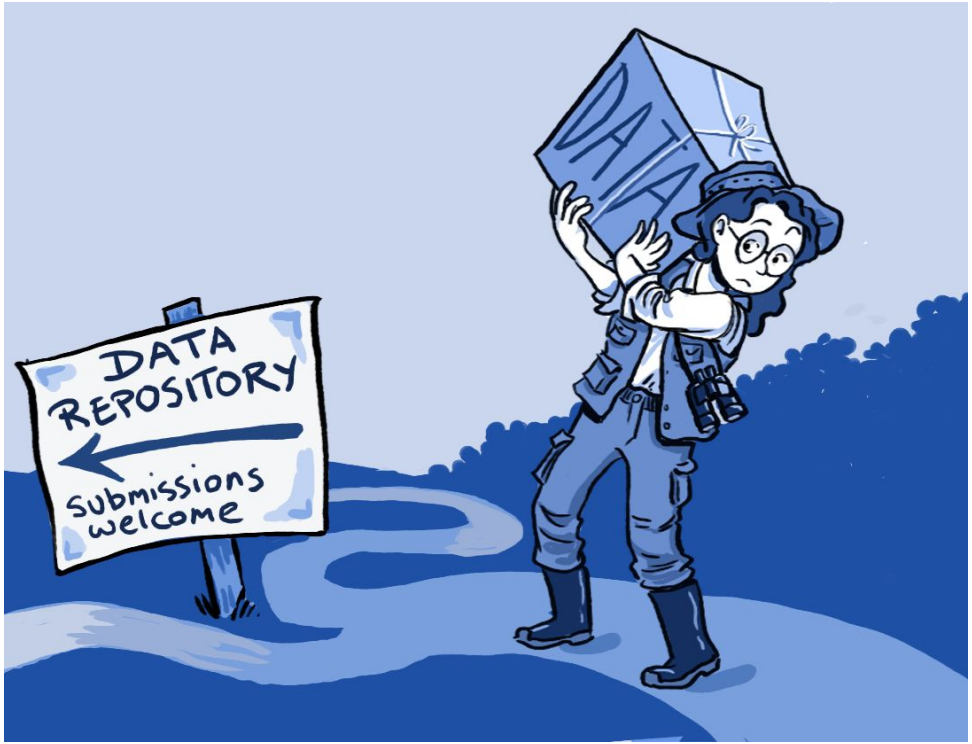
1. Define a question
2. Gather information
3. Form an explanatory hypothesis
4. Test the hypothesis by performing an experiment collect data in a **reproducible** manner
5. Analyze the data
6. Interpret the data and draw conclusions that serve as a starting point for new hypothesis
7. Publish 💬
8. Retest (mostly done by other scientists)

Scientific method

1. Define a question
2. Gather information
3. Form an explanatory hypothesis
4. Test the hypothesis by performing an experiment collect data in a **reproducible** manner
5. Analyze the data
6. Interpret the data and draw conclusions that serve as a starting point for new hypothesis
7. Publish
8. Retest (mostly done by other scientists)

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

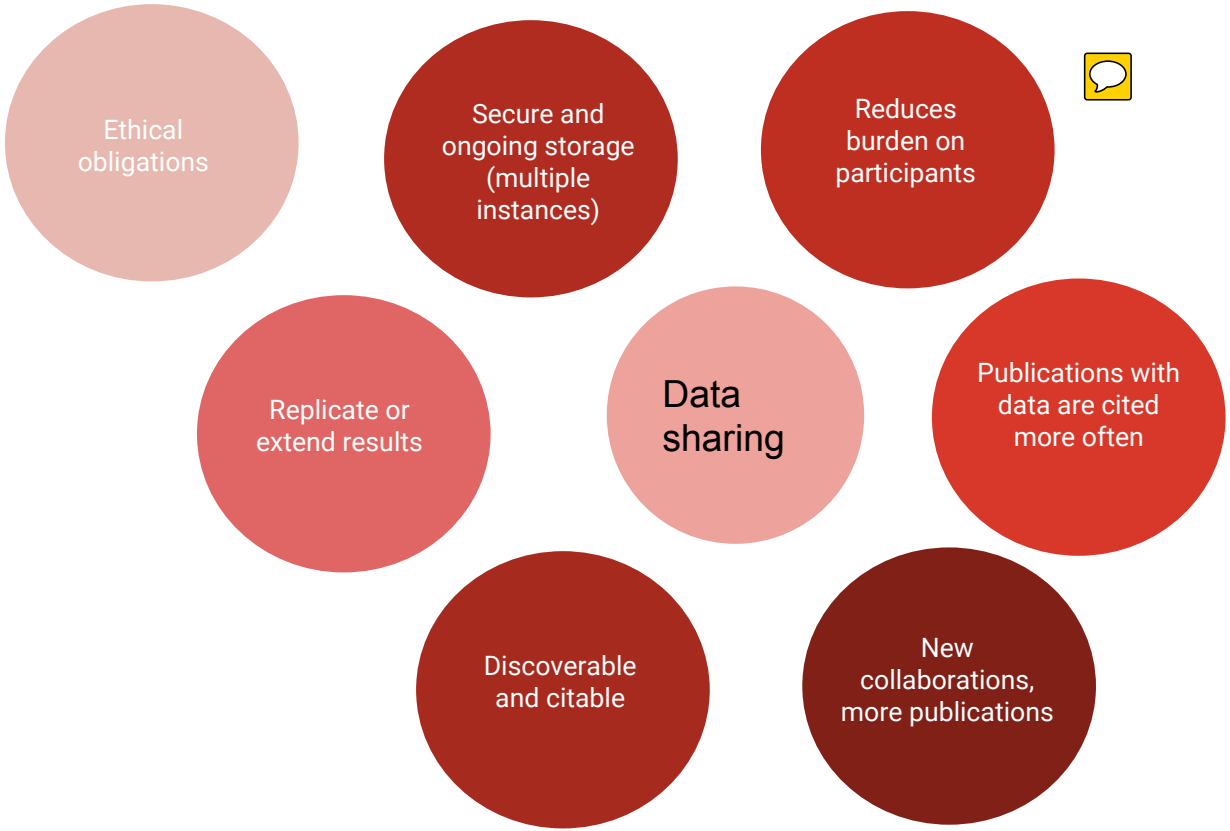
Data sharing!!! 💬



https://dx.doi.org/10.1371/journal.pbio.1001779

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Why should you share data?



DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Non-scientific context

Non-scientific context

The ability to share the same data resource with multiple applications or users. It implies that the data are stored in one or more servers in the network and that there is some software locking mechanism that prevents the same set of data from being changed by two people at the same time. Data sharing is a primary feature of a database management system (DBMS).

Non-scientific context

The ability to share the same data resource with multiple applications or users. It implies that the data are stored in one or more servers in the network and that there is some software locking mechanism that prevents the same set of data from being changed by two people at the same time. Data sharing is a primary feature of a database management system (DBMS).

Why is it important?

Non-scientific context

Shared data:

- Enables collaboration
- Shows what customers value most
- Builds a shared understanding
- Helps anticipate future problems
- Leads to better business practices

Scientific/Non-scientific method?

1. Define a question
2. Gather information
3. Form an explanatory hypothesis
4. Test the hypothesis by performing an experiment collect data in a **reproducible** manner
5. Analyze the data
6. Interpret the data and draw conclusions that serve as a starting point for new hypothesis
7. Publish
8. Retest (mostly done by other scientists)

Scientific/Non-scientific method?

1. Define a question
2. Gather information
3. Form an explanatory hypothesis
4. Test the hypothesis by performing an experiment collect data in a **reproducible** manner
5. Analyze the data
6. Interpret the data and draw conclusions (that serve as a starting point for new hypothesis: not necessarily)
7. Publish
8. Retest ~~(mostly done by other scientists)~~ (not necessarily, re-apply to similar issues or across different businesses)

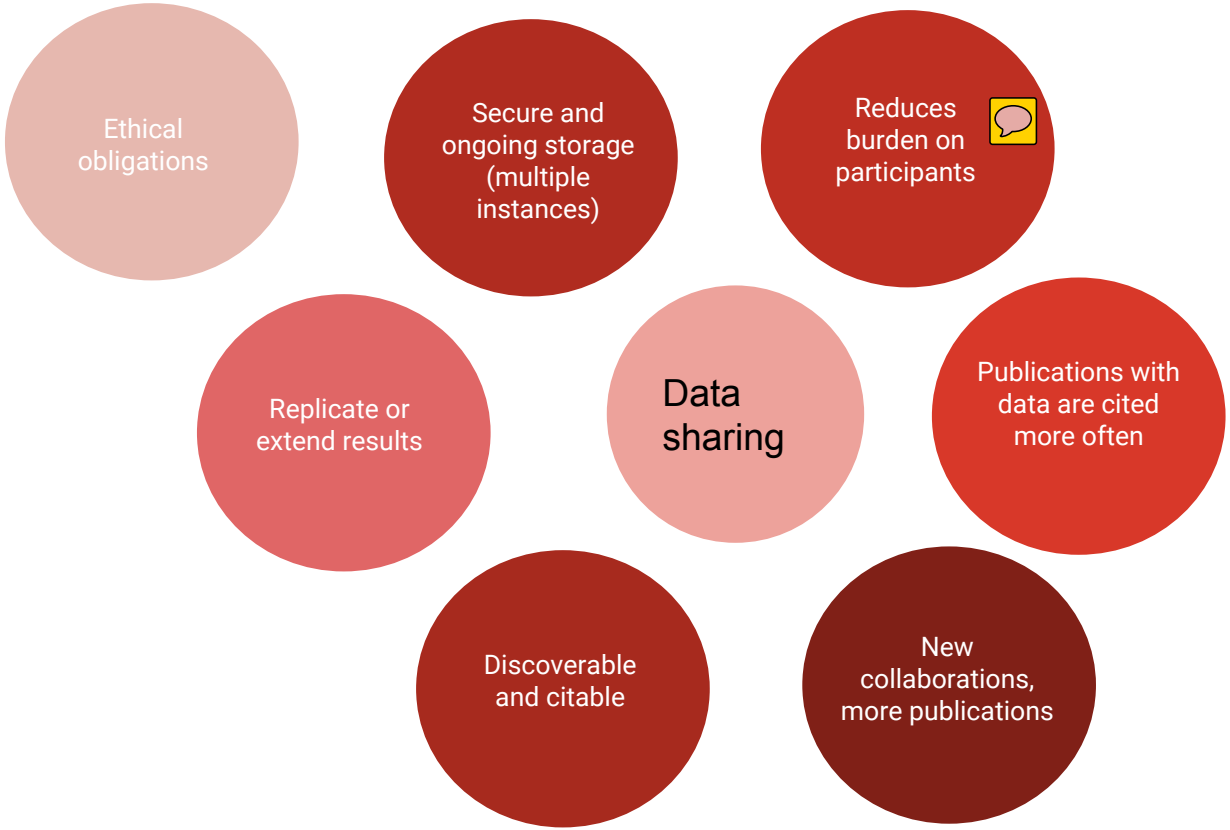
Scientific/Non-scientific method?

1. Define a question
2. Gather information
3. Form an explanatory hypothesis
4. Test the hypothesis by performing an experiment collect data in a **reproducible** manner
5. Analyze the data
6. Interpret the data and draw conclusions (that serve as a starting point for new hypothesis: not necessarily)
7. Publish (internally, e.g. reports)
8. Retest (mostly done by other scientists)

Scientific/Non-scientific method?

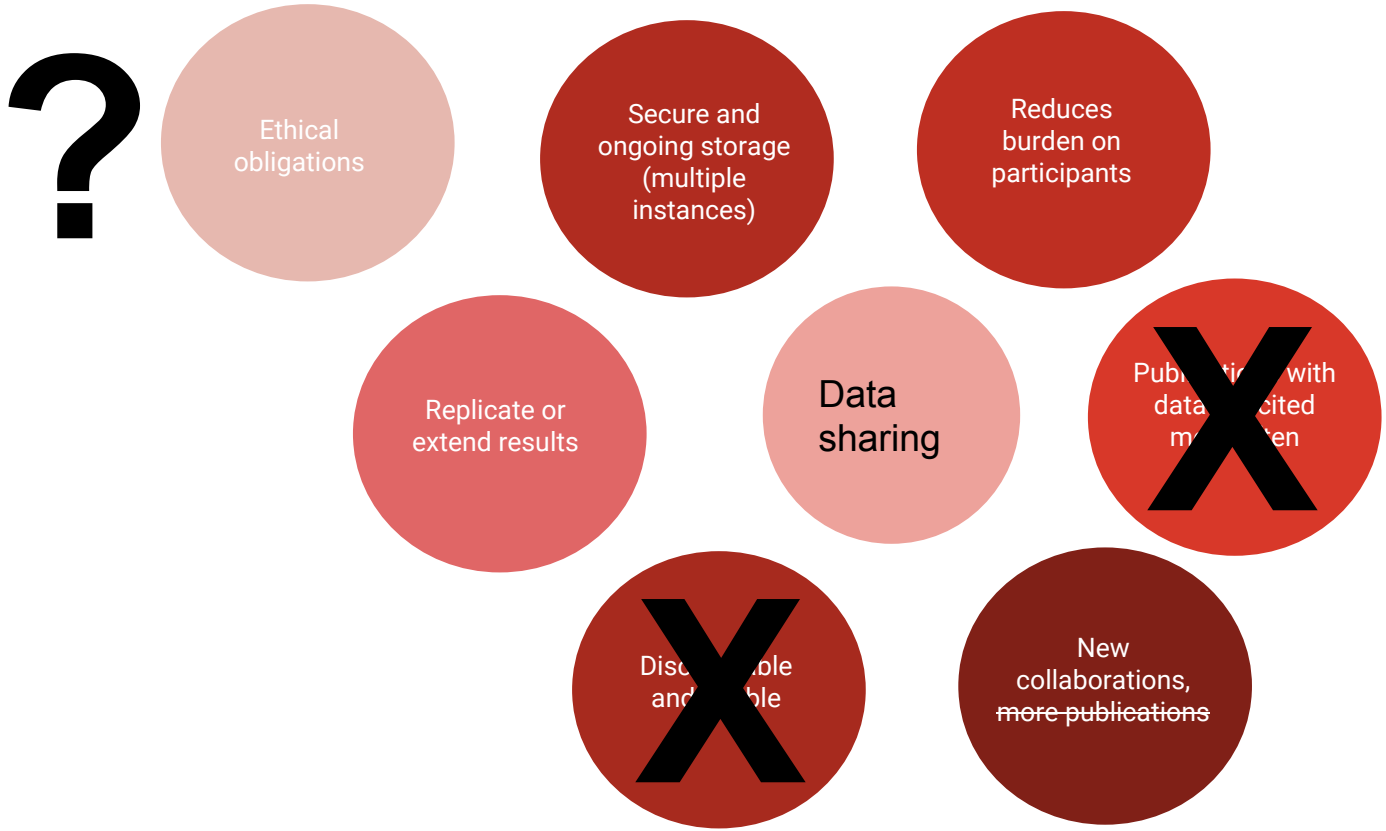
1. Define a question
2. Gather information
3. Form an explanatory hypothesis
4. Test the hypothesis by performing an experiment collect data in a **reproducible** manner
5. Analyze the data
6. Interpret the data and draw conclusions (that serve as a starting point for new hypothesis: not necessarily)
7. Publish (internally, e.g. reports)
8. Retest ~~(mostly done by other scientists)~~ (not necessarily, re-apply to similar issues or across different businesses)

Why should you share data?



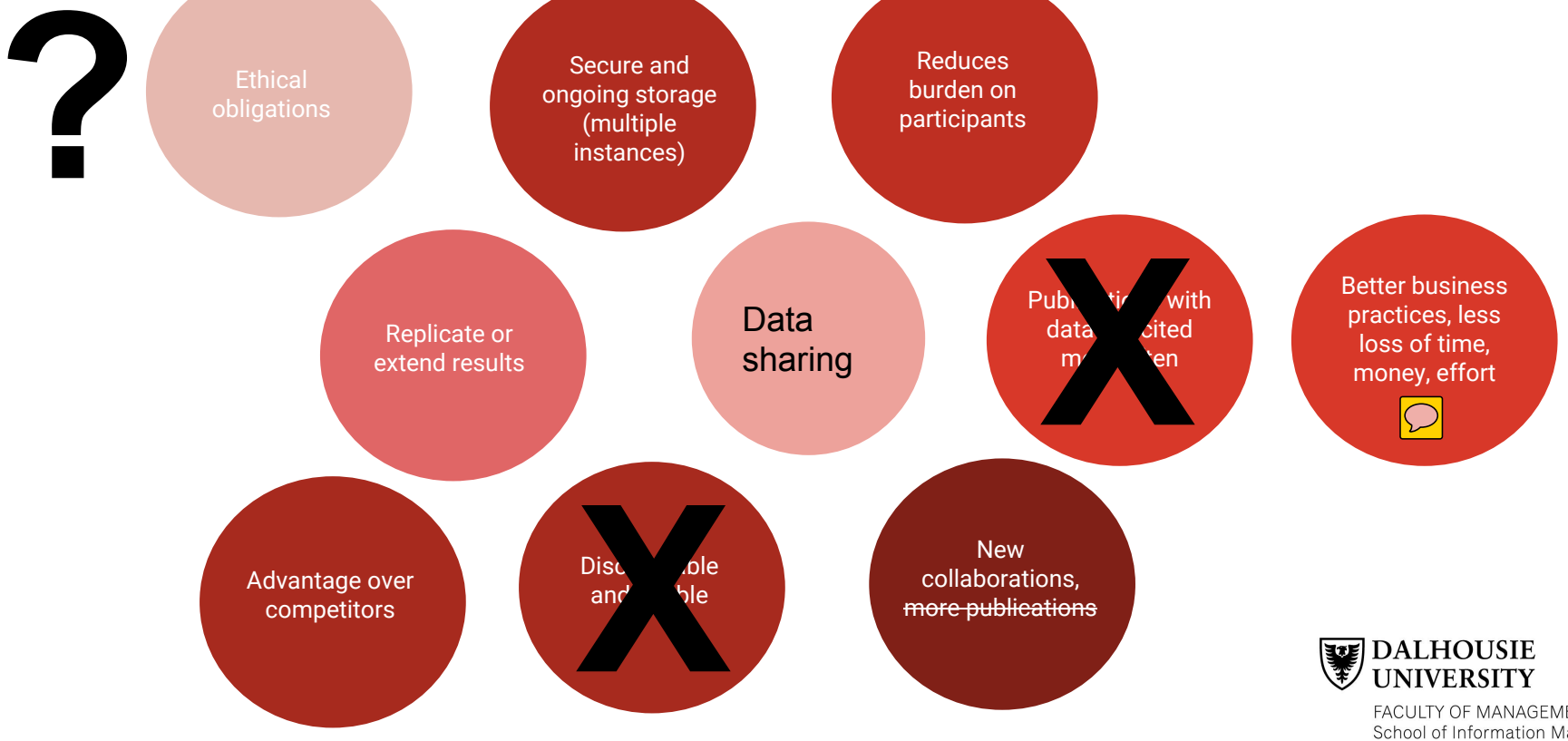
DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Why should you share data?



DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Why should you share data?



DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

How do we share data?

- Journals, articles, in print and online
- Personal communication (e.g. email)
- Hardware (e.g. hard drives, USB sticks)
- Dedicated websites 
- Local databases (servers)
- Online databases (servers)
- The cloud in general

Open Science Framework

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

How do we keep track of changes?

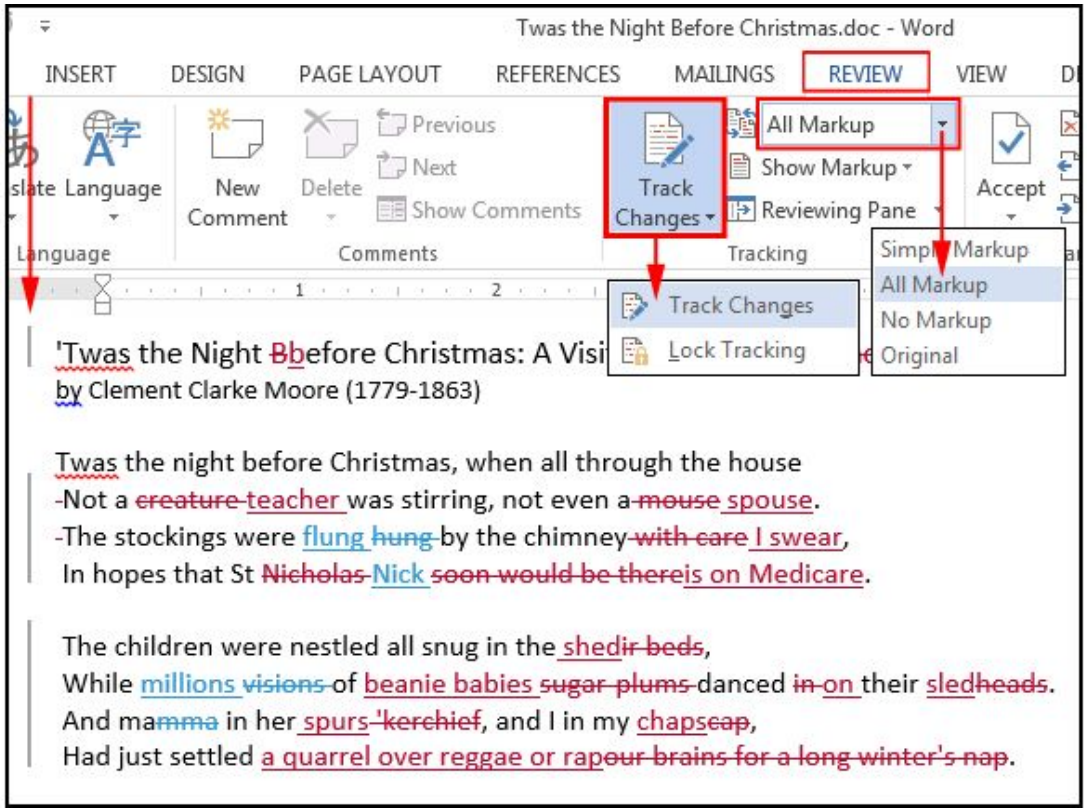
Name	Date modified	Type	Size
idoden_beta0.06.owl	5/28/2013 12:39 PM	OWL File	2,910 KB
idoden_beta0.07.owl	7/9/2013 4:58 AM	OWL File	2,914 KB
idoden_beta0.14.owl	12/22/2013 4:38 PM	OWL File	2,938 KB
idoden_beta0.15.owl	2/3/2014 7:08 PM	OWL File	2,938 KB
idoden_beta0.16.owl	2/18/2014 7:22 AM	OWL File	2,947 KB
idoden_beta0.16b.owl	3/19/2014 5:01 AM	OWL File	2,958 KB
idoden_beta0.17.owl	3/20/2014 12:29 AM	OWL File	2,958 KB
idoden_beta0.18.owl	4/10/2014 6:07 AM	OWL File	2,957 KB
idoden_beta0.19.owl	5/19/2014 2:56 PM	OWL File	2,968 KB
idoden_returned_012813.owl	1/27/2013 7:09 PM	OWL File	2,833 KB
idoden_returned_012913.owl	1/29/2013 7:09 AM	OWL File	2,857 KB
idoden_returned_013013.owl	1/30/2013 9:24 AM	OWL File	2,860 KB
idoden_returned_013113.owl	1/31/2013 2:30 AM	OWL File	2,873 KB
idoden_returned_013113_v002.owl	1/31/2013 11:31 AM	OWL File	2,866 KB
idoden_returned_020113.owl	2/1/2013 9:20 AM	OWL File	2,883 KB
idoden_returned_v0.001	1/10/2013 6:05 AM	001 File	2,773 KB
idoden_rodeimp.owl	12/4/2012 10:39 AM	OWL File	2,773 KB
idoden_synonyms_112112.owl	11/23/2012 7:09 AM	OWL File	2,776 KB
idoden_vicky_returned.obo	12/7/2012 10:06 AM	OBO File	789 KB
idoden_vicky_returned_manolis.obo	12/7/2012 10:06 AM	OBO File	789 KB
idoden_vicky_v3.obo	11/14/2012 3:51 AM	OBO File	763 KB
idoden_vicky_v3.owl	11/20/2012 8:21 AM	OWL File	2,801 KB
idoden091912 - Copy.owl	9/20/2012 2:21 AM	OWL File	1,506 KB
idoden091912 (2).owl	9/20/2012 2:21 AM	OWL File	1,506 KB
idoden091912.owl	10/9/2012 6:20 PM	OWL File	1,507 KB
idoden100812.owl	10/8/2012 6:05 AM	OWL File	1,517 KB

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

How do we keep track of changes?

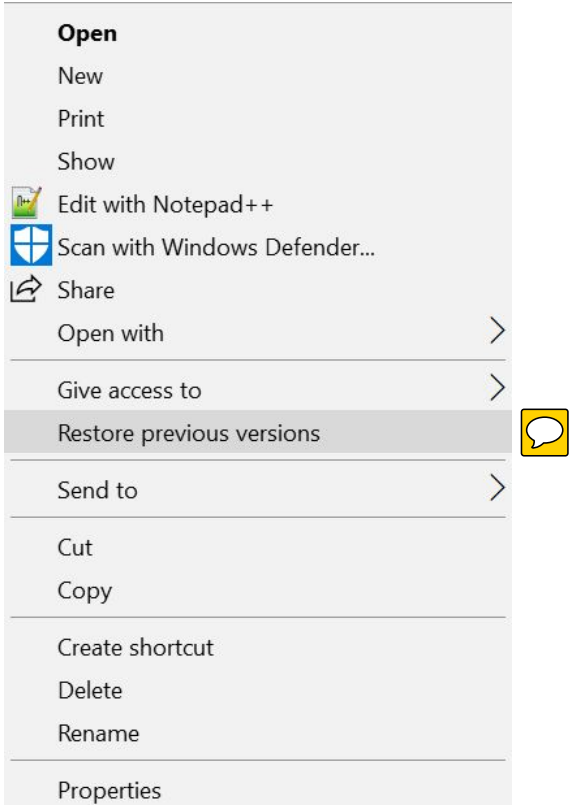
Google docs track history and comments

How do we keep track of changes?



DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

How do we keep track of changes?



DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Version control

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

What is version control?

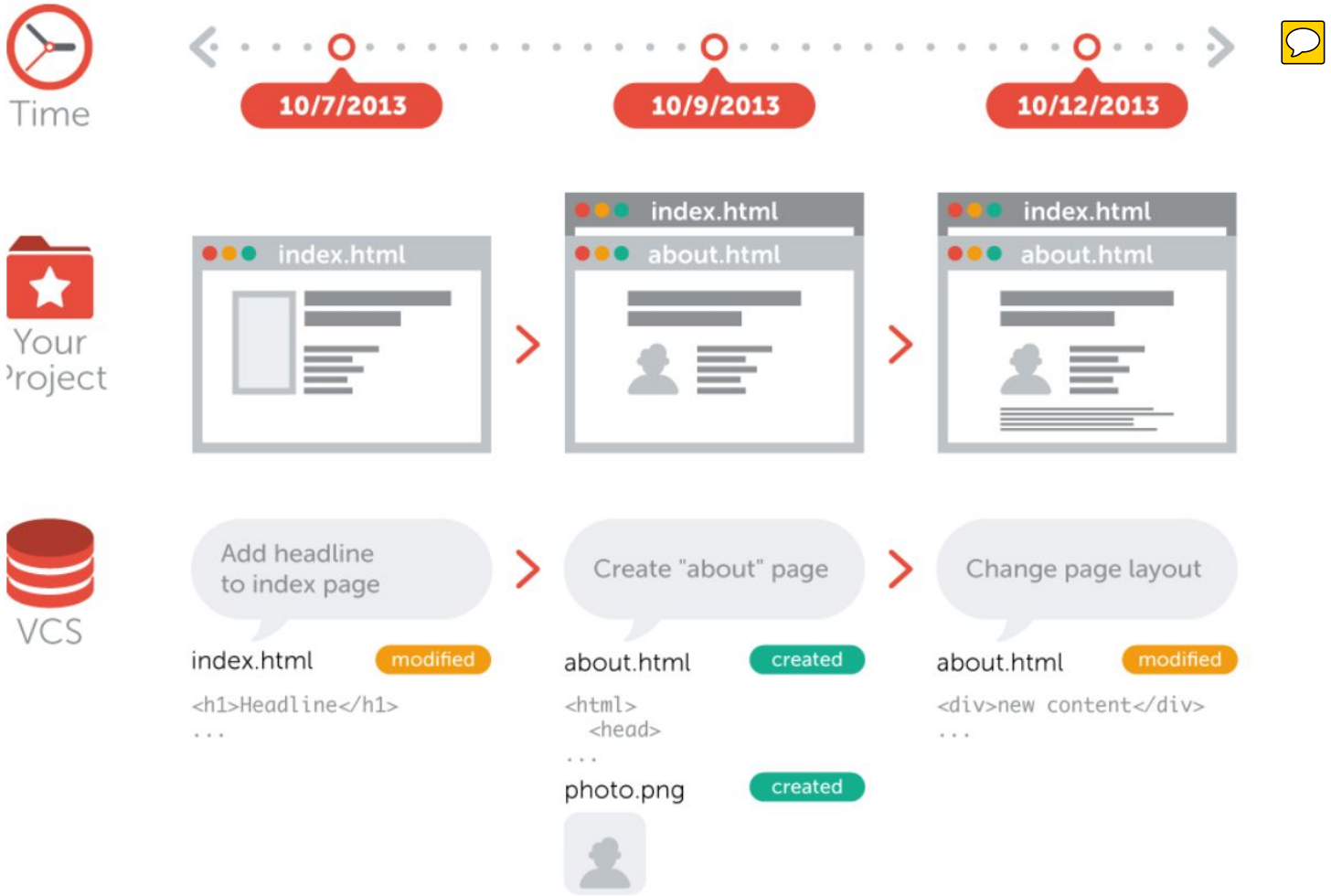
What is version control?

The management of changes to a variety of files.

What is version control?

The management of changes to a variety of files.

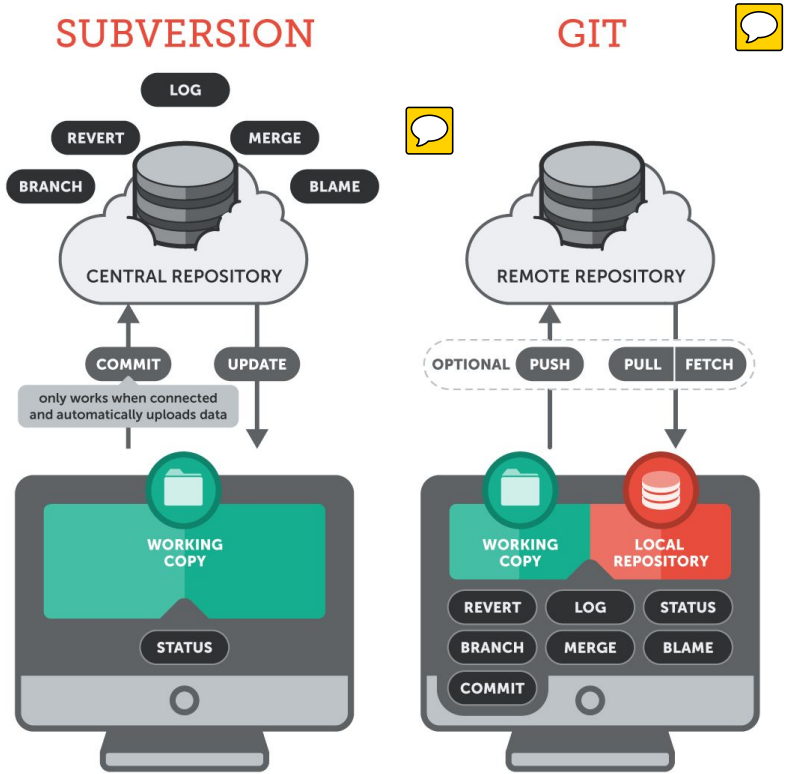
A system that lets you save a specific version (snapshot) of all of your work.


Time
10/7/2013
10/9/2013
10/12/2013
Your Project
index.html
index.html
about.html
index.html
about.html
VCS
Add headline to index page
Create "about" page
Change page layout
index.html
modified
<h1>Headline</h1>
...
about.html
created
<html>
<head>
...
photo.png
created
about.html
modified
new content
...

Version control 💬

- Records all the changes
- Shows you the differences
- Allows you to recover older versions
- Multiple backups
- Allows you to add a description to each commit
- Store versions properly
- You can have a “release” after a milestone
- Allows for multiple collaborators to work on the files simultaneously
- You can see who changed what and when

Centralized vs Distributed/Decentralized model



https://www.git-tower.com/learn/git/ebook/en/desktop-gui/appendix/from-subversion-to-git

DALHOUSIE UNIVERSITY
FACULTY OF MANAGEMENT
School of Information Management

Glossary

https://help.github.com/articles/github-glossary/

Centralized version control - Subversion

https://sourceforge.net/p/diseaseontology/code/HEAD/tree/trunk/

Distributed version control - Git

https://github.com/DiseaseOntology/HumanDiseaseOntology

Lab work