HARVARD | BUSINESS | SCHOOL

Open-Source Computing and Research

Ian Gow

HBS Brown Bag

16 September 1995

HARVARD | BUSINESS | SCHOOL

Open-Source Computing and Research

Ian Gow

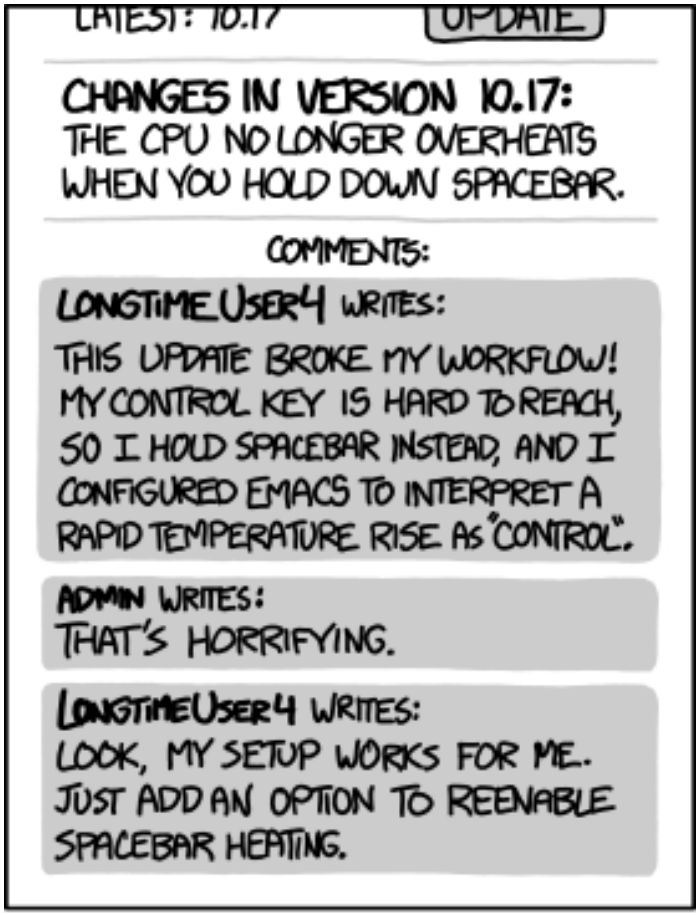
HBS Brown Bag

16 September 2015

Goals of this talk

- Illustrate the power and convenience of having a relational database *as a repository of all of your data*
  - Quality data
  - Separate data from analysis
- Demonstrate the (open-source) relational database system that I happen to use (PostgreSQL)
- Illustrate how open-source computing tools can be used in research
- Argue for the value of version control for collaboration and reproducibility
- Provide some case studies to illustrate above

Caveats



EVERY CHANGE BREAKS SOMEONE'S WORKFLOW.

Research computing at HBS

- **Theory**
  - Data stored in MySQL servers
  - Code run on HBS Research Grid
  - Shared folders used to collaborate with co-authors at HBS and elsewhere
- **Practice**
  - Data stored on laptops and desktops
  - Code run on laptops and desktops
  - Sharing code and data?

Hardware requirements

Component	My setup	Minimum	TSS (DBA)
RAM	16 – 32 GB	8 – 16 GB	4 GB
HDD	512 GB – 3 TB	512 GB	256 – 300 GB
Processors	4 – 6 cores	2 – 4 cores	
Operating system	OS X (Mac)	Anything (Linux best)	OS X or Windows

Software

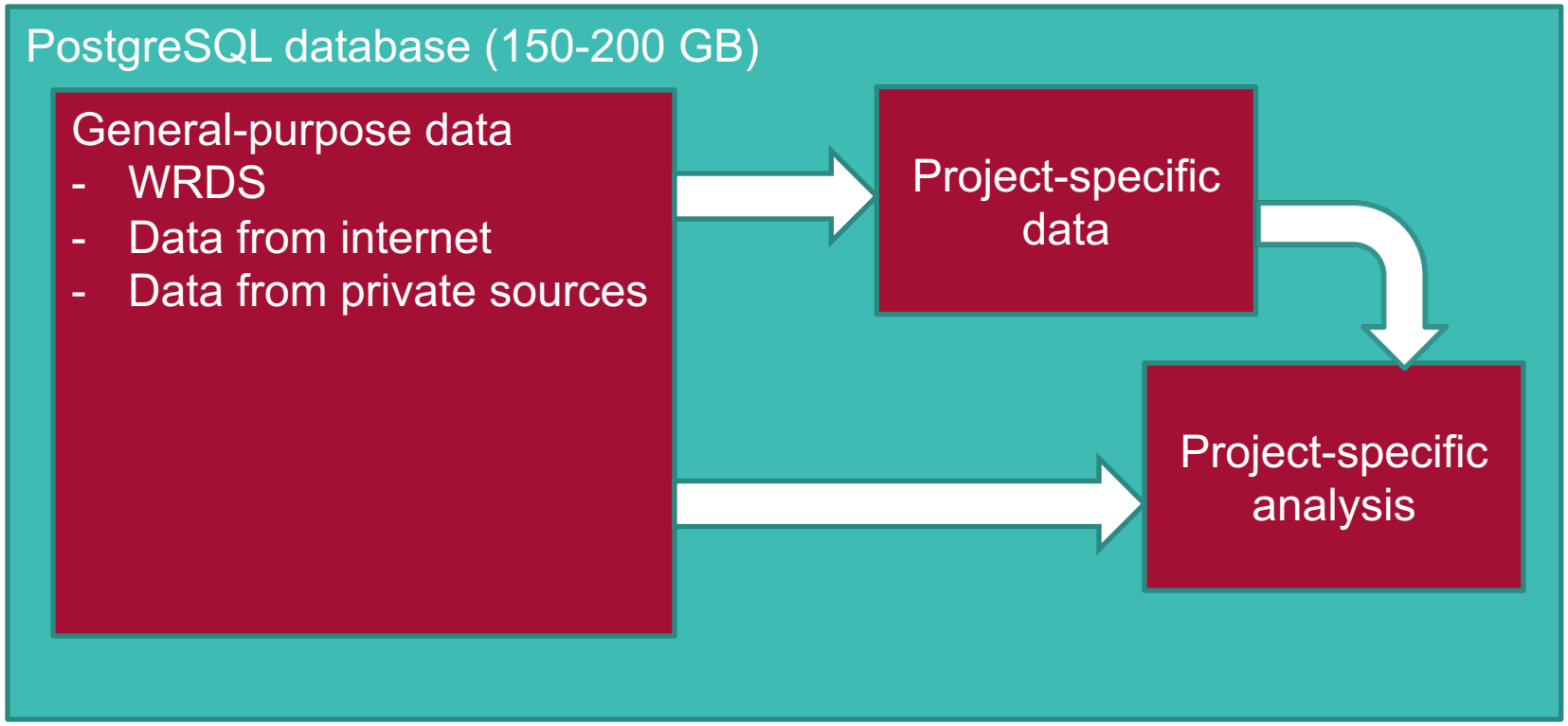
- The software I will describe is open-source, free, and available for all platforms:
  - PostgreSQL
  - R
  - RStudio
  - Python
  - Perl
  - Git
- In the last year or so, I have made the code I use to manage most of my publicly available, so you should be able to replicate what I will describe.

Lots of cool stuff I *won't* talk about

- "Big Data"
- Hadoop, Spark, etc.
- Cloud computing (AWS, etc.)
- Docker
- LaTeX, etc.
- Sharing bibliographies and papers (Mendeley, etc.)
- Many R, Python, etc., packages

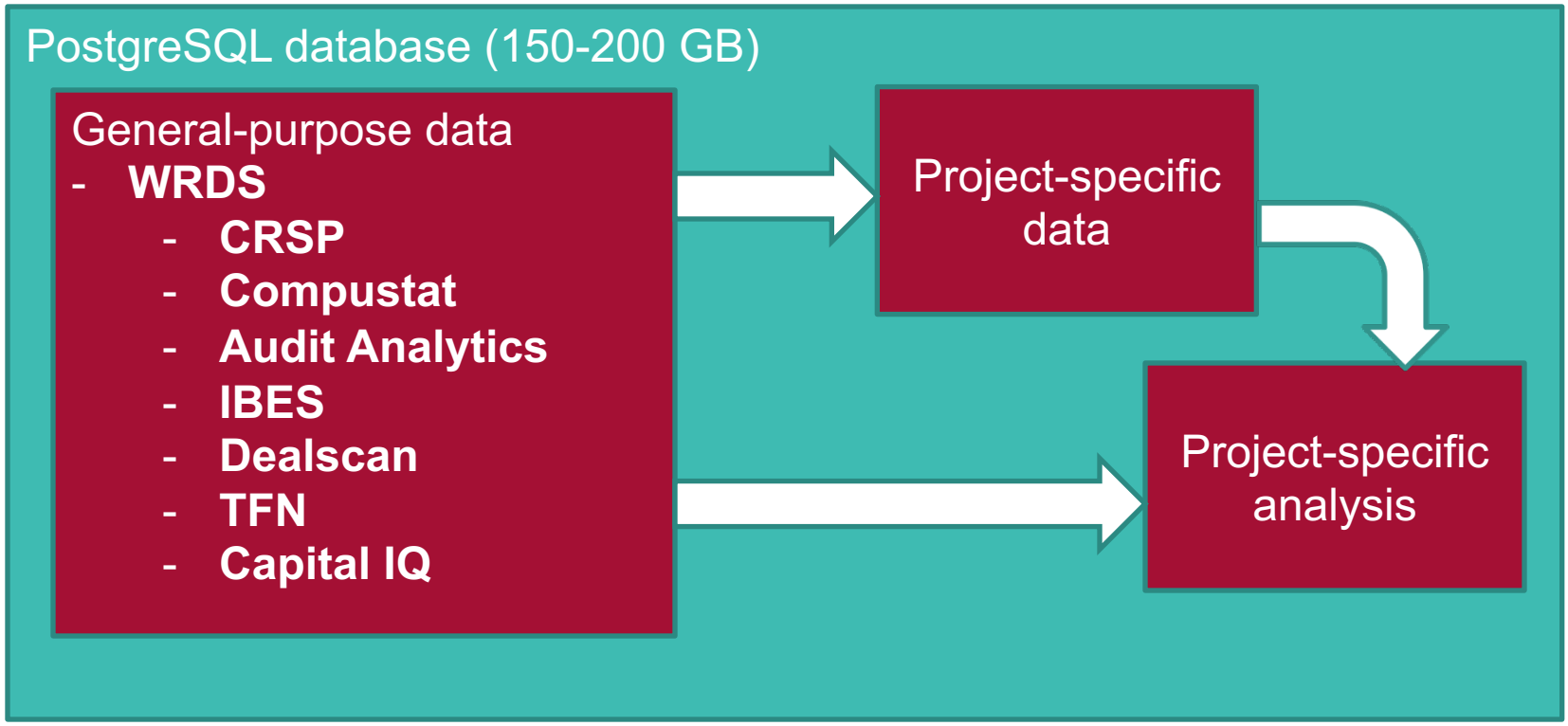
My background

Data



Some data is kept in system files (e.g., raw SEC filings).

Data



Some data is kept in system files (e.g., raw SEC filings).

Data: WRDS

- WRDS data is pulled into my database using pre-existing Perl scripts
  - Scripts determine if updates are available
  - Most data sets are easy, but some require special handling
  - Special handling reflected in custom scripts

WRDS: Compustat Point-in-Time Data

Data Set Name	COMPSNAP.WRDS_CSQ_PIT
Observations	8807954
Type	DATA
Variables	620
Indexes	2
Created	Wednesday, June 10, 2015 05:40:01 PM
Last Modified	Wednesday, June 10, 2015 05:40:01 PM
Label	Compustat Snapshot Quarterly/YTD - Point-In-Time
File Size	44,403,195,904 bytes

WRDS: Compustat Point-in-Time Data

```
2009-Mac-Pro:acct_data igow$ ./wrds_update.pl compsnap.wrds_csq_pit
Getting schema for compsnap.wrds_csq_pit

Beginning file import at 15:27:33

Importing data into compsnap.wrds_csq_pit.

COPY 8807954

Result of system command: 0

Completed file import at 17:10:07

Last modified: 06/10/2015 17:40:02
```

WRDS: Compustat Point-in-Time Data

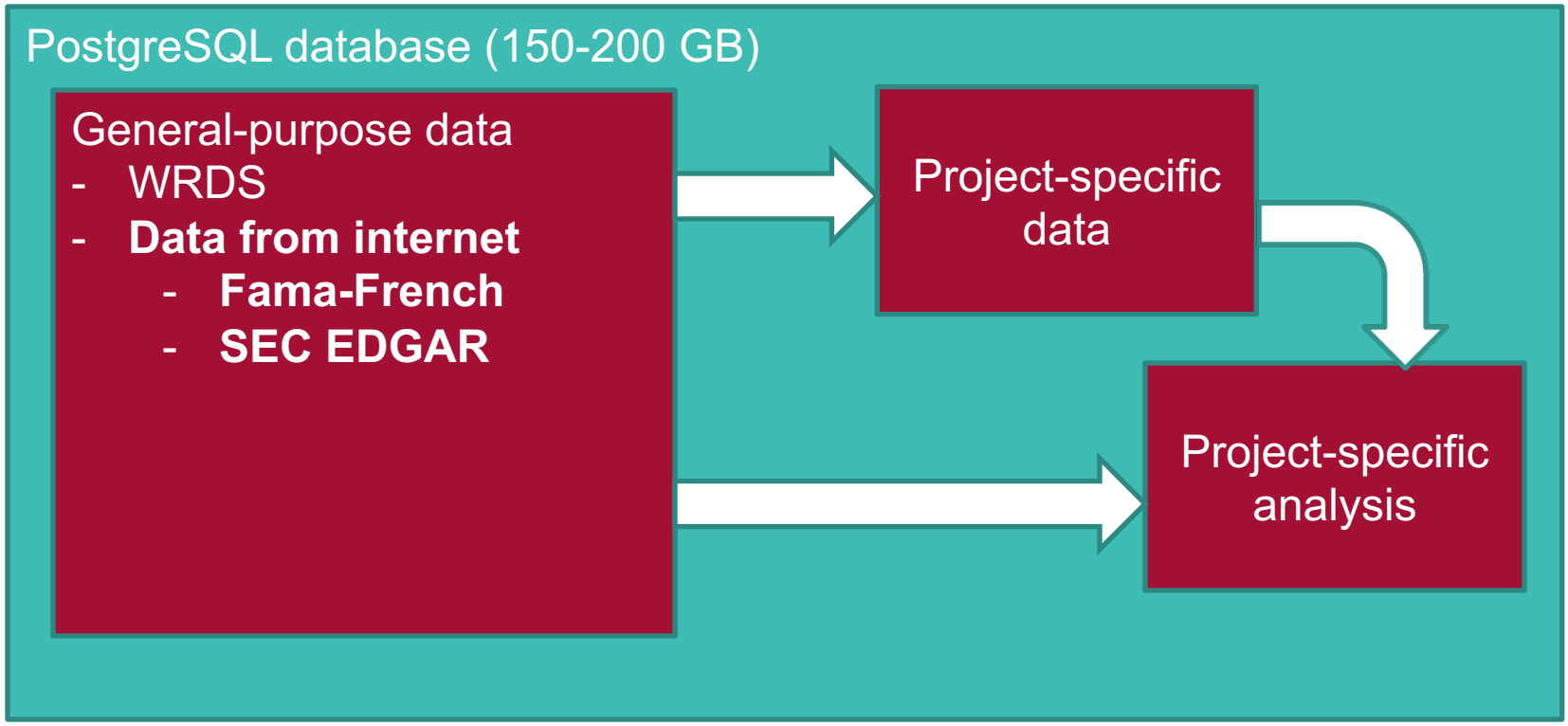
Alphabetic List of Indexes and Attributes

#	Index	# of Unique Values	Variables
1	dt	232765	PITDate1 PITDate2
2	gvkey	34685	

Sort Information

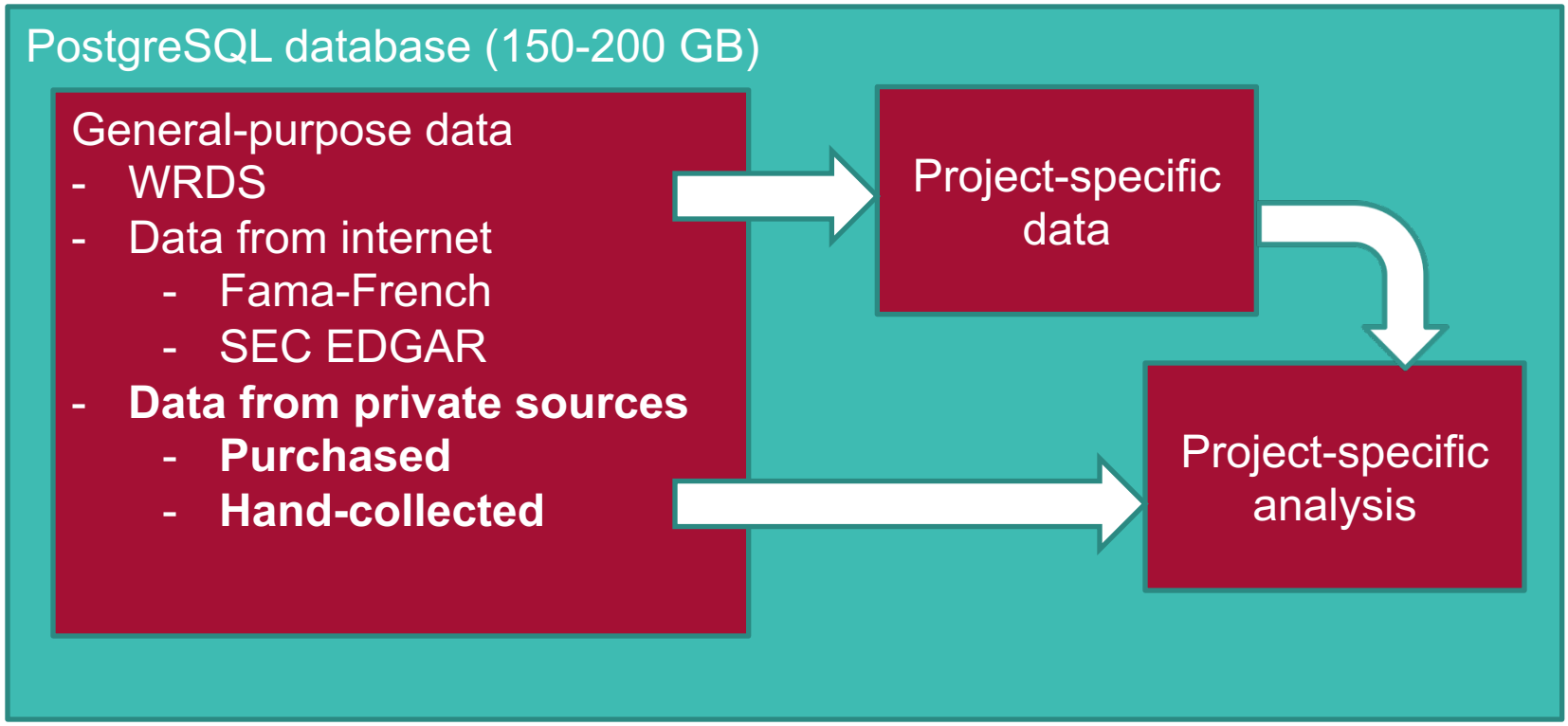
Sortedby	gvkey PITDate1 PITDate2 datadate indfmt datafmt consol popsrc
Sort Option	NODUPKEY

Data



Some data is kept in system files (e.g., raw SEC filings).

Data



Some data is kept in system files (e.g., raw SEC filings).

StreetEvents data

- Transcripts of more than 300,000 conference calls, etc.
- Data updates provided monthly
- Data provided as XML files
- Processed using Perl scripts into “clean” data stored in PostgreSQL
- Then easily subject to analysis

StreetEvents data: Fog

- Fog is a measure of linguistic complexity.
- Used by Feng Li in his dissertation (JAE 2008)
- So I emailed Feng for the code.

StreetEvents data: Fog

Hi Ian,

I used Perl for the calculations. Here is the core part, it's actually very simple. You need to save the file to a .txt file and download and install a Perl module called "`Lingua::EN::Fathom`".

```
use Lingua::EN::Fathom;
my $text = new Lingua::EN::Fathom;
$text->analyse_file("file.txt");

$percent_complex_words = $text->percent_complex_words;
$num_sentences     = $text->num_sentences;
$words_per_sentence = $text->words_per_sentence;

$fog   = $text->fog;
Print "$cik, $filedate, $file, $percent_complex_words, $num_sentences, ";
Print "$words_per_sentence, $fog, \n";
```

Let me know if you need more info about this.

Life is fine, busy with the recruiting stuff lately. Hope all is well with you.

Feng

Note: `sudo cpan -i Lingua::EN::Fathom` installs this module.

StreetEvents data: Fog

Fortunately, PostgreSQL allows you to define functions using other languages, including Perl, Python, R, C, SQL, JavaScript.

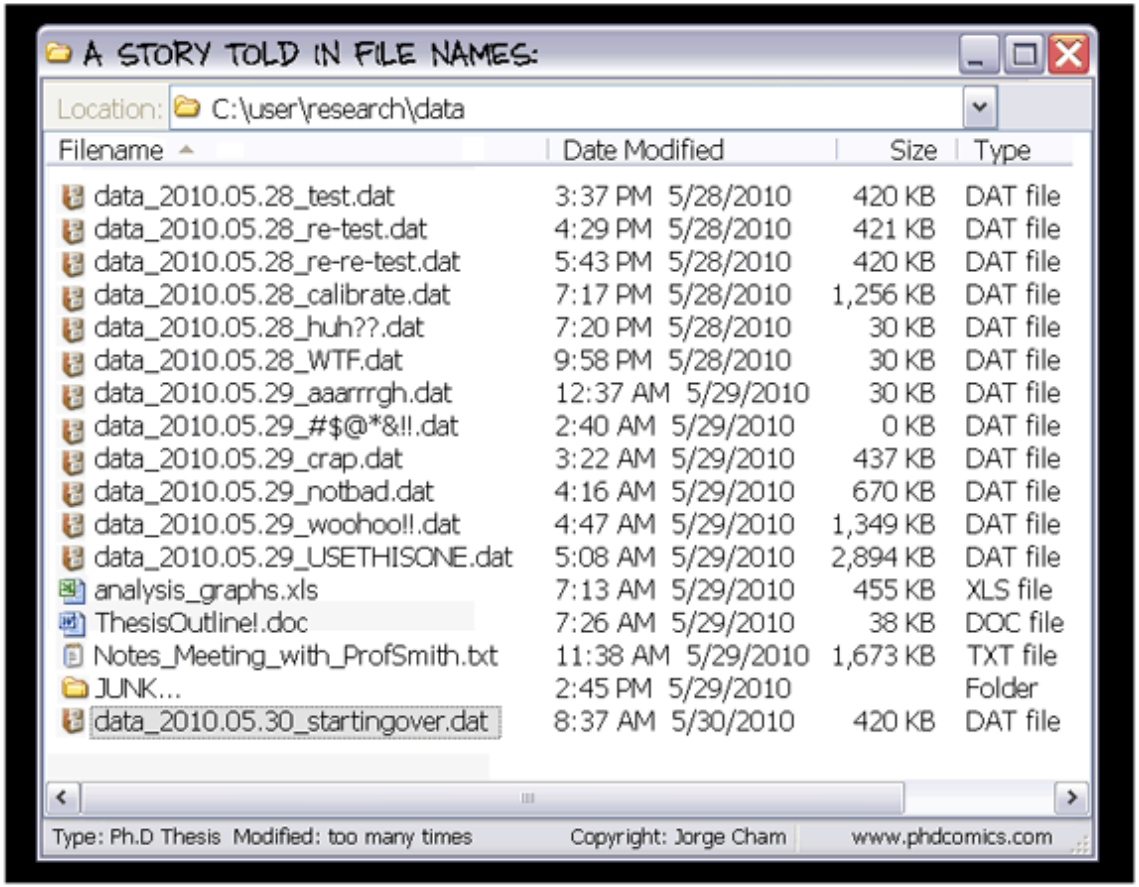
```
CREATE OR REPLACE FUNCTION fog_original(text)
  RETURNS double precision AS $BODY$
  # Load Perl modules that calculate fog, etc.
  use Lingua::EN::Fathom;
  my $text = new Lingua::EN::Fathom;
  if (defined($_[0])) {
    $text->analyse_block($_[0]);
    return($text->fog);
  }
$BODY$ LANGUAGE plperlu;
```

Some other examples

- Hand-collecting data
- Annotating director biographies

Version control?

More version control?



Version control, collaboration & reproducibility

- Dropbox
- Google Docs
- Google Drive
- Git (Bitbucket, GitHub)
- ShareLaTeX

Why PostgreSQL?

- Rich, modern SQL (cf. MySQL)
  - WINDOW functions
  - Common table expressions
  - USING, etc.
- Robust (cf. SQLite)
- Support for rich data types
  - Arrays (map to lists in Python)
  - JSON (easy to map to and from Python dictionaries, etc.)
- Procedural languages (cf. MySQL)
- Excellent documentation, licensing, etc.

Open-source is awesome

- PL/R example
- StackOverflow examples
- R packages for pretty much everything

Why not WRDS?

- WRDS is fine, but …
  - Need to use SAS predominantly
  - Some WRDS data is a mess (e.g., Audit Analytics)
  - What about all my *other* data? (StreetEvents, EDGAR, etc.)
  - Ultimately much more practicable to get WRDS data locally than trying to manage data on WRDS servers … see survey results

Why not HBS Research Grid?

- No PostgreSQL, just MySQL (see above)
- No RStudio Server
- Collaboration is difficult
- Even with MySQL, still (seem to) need to manage data
- HBS Research Grid seems walled off from the internet (none of my scripts work there, even though the servers run Linux).
- I seem not to be alone in *not* using it.

Challenges

- Collaboration is less of a pain
  - I don't want to be a DBA or systems administrator
  - Learning curve for co-authors
- Challenging to find resources to manage this
  - Research Computing and Baker Research Services are not in the business of providing comprehensive data services
  - I want to do *less* data-cleaning, not more
- Sharing setup is not easy
  - Value would come from scale
  - Need better documentation
  - Hardware limitations (no resources to leverage)