

Computational Husbandry

Seth Russell - seth.russell@cuanschultz.edu

Peter DeWitt - peter.dewitt@cuanschultz.edu

SOM, Dept of Biomedical Informatics

Presentations: <https://github.com/magic-lantern/Computational-Husbandry-2022>

Survey: <https://forms.gle/ZumeuAiWRm83ra9h9>

Computational Husbandry

- Husbandry
- Homesteading
- Permaculture



Computing Resources

- CPU - Central Processing Unit
- RAM - Random Access Memory
- Persistent Storage - Hard Disk, Solid State Drive
- GPU - Graphics Processing Unit
- Network - Communication to other machines (storage, data, programs, etc).
- Software
- Data

Computing Resources

- The Operating System is the main program controlling access to and allocation of these resources. All of these are shared by all users and running programs on a system.
- The slowest/most limited resource involved controls the maximum throughput of the system.
- Networking interfaces allow the sharing of all computing resources among many people and many programs.
 - Network computing increases the complexity of software development

Measure twice, cut once

Avg

Avg[|||||]

Swp[

Mem[|||||]

Tasks: 597; 31 running

Load average: 26.71 25.78 25.44

Uptime: 26 days, 22:07:55

0K/8.00G

581G/2.95T

21.5%

Linux 5.4.0-126-generic (plissken-em-2222)

10/28/2022

_x86_64_

(144 CPU)

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
1971570		20	0	45.3G	18.3G	277M	S	100.	0.6	37h41:10	/usr/local/MATLAB/R2021a/bin/glnxa64/MATLAB -parallelserv
1971913		20	0	45.2G	18.3G	276M	S	100.	0.6	32h	~\$ iostat -cd
1965838		20	0	45.2G	18.3G	276M	S	100.	0.6	33h	
1965253		20	0	45.4G	18.3G	278M	S	100.	0.6	35h	
1967895		20	0	45.4G	18.3G	278M	S	100.	0.6	35h	
1970893		20	0	45.3G	18.4G	276M	S	100.	0.6	33h	
1965547		20	0	45.4G	18.3G	278M	S	99.9	0.6	33h	avg-cpu: %user %nice %system %iowait %steal %idle
1968877		20	0	45.3G	18.3G	278M	S	99.9	0.6	35h	16.64 0.00 0.81 0.00 0.00 82.55
1966420		20	0	45.3G	18.3G	277M	S	99.9	0.6	35h	Device tps kB_read/s kB_wrtn/s kB_dscd/s kB_read kB_wrtn kB_dscd
1969892		20	0	45.2G	18.3G	276M	S	99.9	0.6	35h	dm-0 62.19 105.08 498.81 204.62 246704352 1171119276 480403392
1970220		20	0	45.4G	18.3G	277M	S	99.9	0.6	33h	loop0 0.00 0.00 0.00 0.00 777 0 0
1968534		20	0	45.2G	18.3G	275M	S	99.9	0.6	34h	loop1 0.00 0.00 0.00 0.00 794 0 0
1971229		20	0	45.4G	18.3G	277M	S	99.9	0.6	35h	loop10 0.00 0.00 0.00 0.00 12 0 0
1966711		20	0	45.3G	18.3G	278M	S	99.9	0.6	40h	loop2 0.00 0.00 0.00 0.00 472 0 0
1966129		20	0	45.3G	18.3G	275M	S	99.9	0.6	32h	loop3 0.01 0.01 0.00 0.00 20923 0 0
1972943		20	0	45.4G	18.3G	279M	S	99.9	0.6	35h	loop4 0.00 0.00 0.00 0.00 779 0 0
1967595		20	0	45.4G	18.3G	278M	S	99.9	0.6	38h	loop5 0.00 0.00 0.00 0.00 1337 0 0
1973639		20	0	45.2G	18.3G	276M	S	99.9	0.6	34h	loop6 0.00 0.00 0.00 0.00 1336 0 0
1973984		20	0	45.4G	18.3G	278M	S	99.9	0.6	37h	loop7 0.00 0.00 0.00 0.00 410 0 0
1967003		20	0	45.4G	18.3G	278M	S	99.9	0.6	35h	loop8 0.00 0.00 0.00 0.00 2196 0 0
1968199		20	0	45.4G	18.3G	278M	S	99.9	0.6	34h	loop9 0.01 0.01 0.00 0.00 19691 0 0
1969207		20	0	45.2G	18.3G	277M	S	99.9	0.6	35h	md0 63.12 105.13 498.11 215.77 246821084 1169467220 506587056
1965256		20	0	45.3G	18.4G	276M	S	99.9	0.6	34h	sda 32.01 492.60 500.40 204.62 1156551188 1174859370 480403392
1972579		20	0	45.3G	18.3G	277M	S	99.9	0.6	34h	sdb 26.03 411.23 500.40 204.62 965506694 1174859370 480403392
											sdc 2.90 112.11 37.66 0.00 263209600 88421667 0

F1Help

F2Setup

F3Search

F4Filter

F5Tree

F6SortBy

F7Nice

F8N

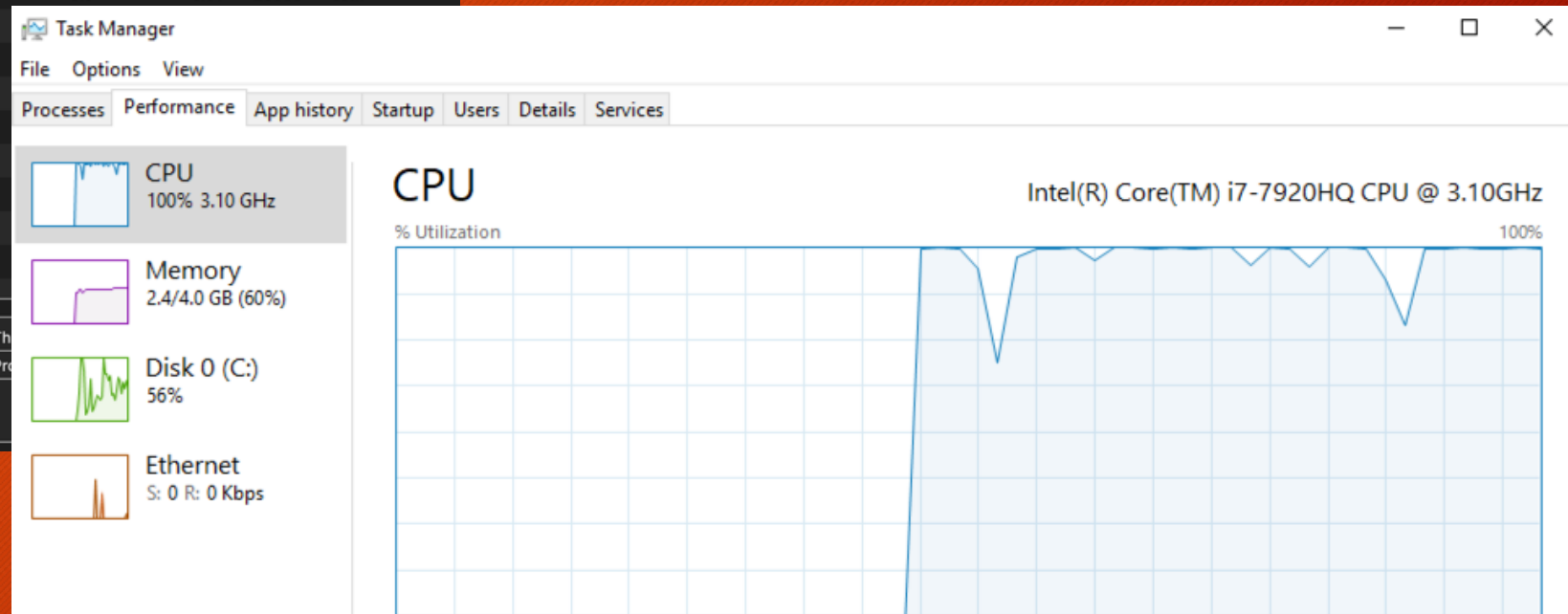
Measure twice, cut once

Activity Monitor
All Processes

Process Name % CPU CPU Time Threads Idle Wake Ups % GPU GPU Time PID User

WindowServer	20.4	35:12:07.57	22	48	1.6	20:46:27.97	152	_windowserver
kernel_task	9.2	20:53:44.25	281	635	0.0	0.00	0	root
Activity Monitor	2.9	6.44	5	2	0.0	0.00	48706	seth
screencapture	2.2	0.36	3	0	0.0	0.00	48719	seth
iTerm2	1.1	1:49.98	6	20	0.0	0.01	29556	seth
sysmond	1.0	14:56.07	3	0				
Egnyte WebEdit	0.9	4:47:25.62	20	110				
loginwindow	0.6	16:42.27	6	66				
Screen Shot	0.3	0.28	4	0				
Control Center	0.3	33:57.50	12	1				
Google Chrome	0.2	55.69	31	1				
R	0.2	50.07	2	1				
FirefoxCP Isolated...	0.2	1:46.51	32	5				
Google Chrome He	0.2	22.97	11	9				

System: 2.90% CPU LOAD
User: 2.90%
Idle: 94.20%



Computer Hardware vs the Human Brain

Line endings - aka “I've got a blank space, baby”

- MS-DOS used the two-character combination of CRLF to denote line endings in files, and modern Windows computers continue to use CRLF as their line ending to this day. Meanwhile, from its very inception, Unix used LF to denote line endings, ditching CRLF for consistency and simplicity. Apple originally used only CR for Mac Classic but eventually switched to LF for OS X, consistent with Unix. <https://www.aleksandrhovhannisyan.com/blog/crlf-vs-lf-normalizing-line-endings-in-git/>



macOS/Linux



Windows

Floating point representations vs Math

✓ $(1/3)+(1/3)+(1/3) == 1$...

... True

✓ $1/10 + 1/10 + 1/10 == 3/10$...

... False

✓ $(0.1 + 0.1 + 0.1) == 0.3$...

... False

✓ $0.1 + 0.1 + 0.1$...

... 0.30000000000000004

✓ *from decimal import ** ...

✓ `Decimal('0.1') + Decimal('0.1') + Decimal('0.1')` ...

... `Decimal('0.3')`

✓ `Decimal('0.1') + Decimal('0.1') + Decimal('0.1') == Decimal('0.3')` ...

... True

```
[$ R
> x <- sqrt(2)
> x ^ 2
[1] 2
> x ^ 2 == 2
[1] FALSE
> all.equal(x ^ 2, 2)
[1] TRUE
```


Floating point representations vs Math

<https://0.30000000000000004.com/>

Java

```
System.out.println(.1 + .2);
```

```
0.30000000000000004
```

AND

AND

```
System.out.println(.1F + .2F);
```

```
0.3
```

Java has built-in support for arbitrary-precision numbers using the [BigDecimal](#) class.

JavaScript

```
console.log(.1 + .2);
```

```
0.30000000000000004
```

The [decimal.js](#) library provides an arbitrary-precision Decimal type for JavaScript.

Julia

```
.1 + .2
```

```
0.30000000000000004
```

Julia has built-in [rational numbers support](#) and also a built-in [arbitrary-precision BigFloat](#) data type. To get the math right, $1/10 + 2/10$ returns $3/10$.

K (Kona)

```
0.1 + 0.2
```

```
0.3
```

Kotlin

```
println(.1 + .2)
```

```
0.30000000000000004
```

AND

AND

```
println(.1F + .2F)
```

```
0.3
```

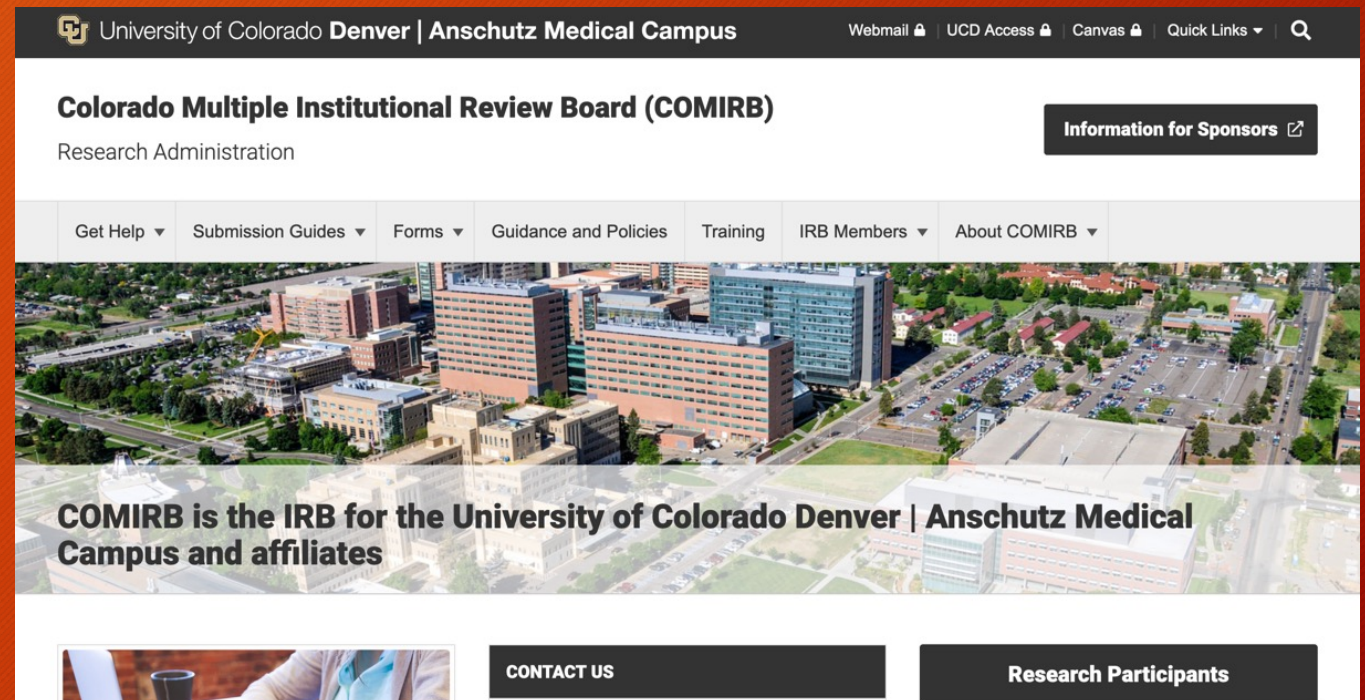
See [Reference documentation](#).

Status check

<https://cups.fast.ai/computationalhusbandry>

Working with health care data

- Data Use Agreements
- Institutional Review Boards



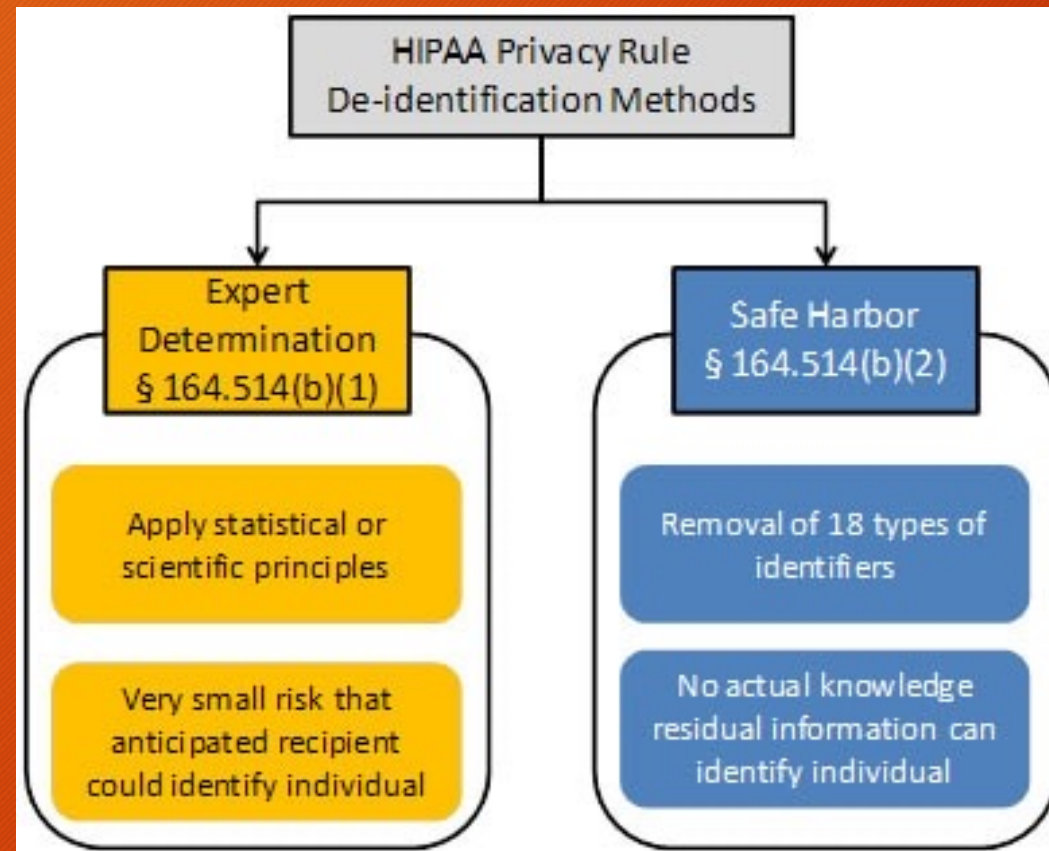
<https://research.cuanschutz.edu/regulatory-compliance/home/hipaa/data-use-agreement>

<https://research.cuanschutz.edu/regulatory-compliance/home/hipaa/data-sharing>

Working with health care data

- Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule
- <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard>
- Protected health information is information, including demographic information, which relates to:
 - the individual's past, present, or future physical or mental health or condition,
 - the provision of health care to the individual, or
 - the past, present, or future payment for the provision of health care to the individual...

Working with health care data



Working with health care data

Safe Harbor method: Remove all of the following

(A) Names

(B) All geographic subdivisions smaller than a state, ... except for the initial three digits of the ZIP code

(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 ...

(D) Telephone numbers

(L) Vehicle identifiers and serial numbers, including license plate numbers

(E) Fax numbers

(M) Device identifiers and serial numbers

(F) Email addresses

(N) Web Universal Resource Locators (URLs)

(G) Social security numbers

(O) Internet Protocol (IP) addresses

(H) Medical record numbers

(P) Biometric identifiers, including finger and voice prints

(I) Health plan beneficiary numbers

(Q) Full-face photographs and any comparable images

(J) Account numbers

(R) Any other unique identifying number, characteristic, or code, except as permitted

(K) Certificate/license numbers

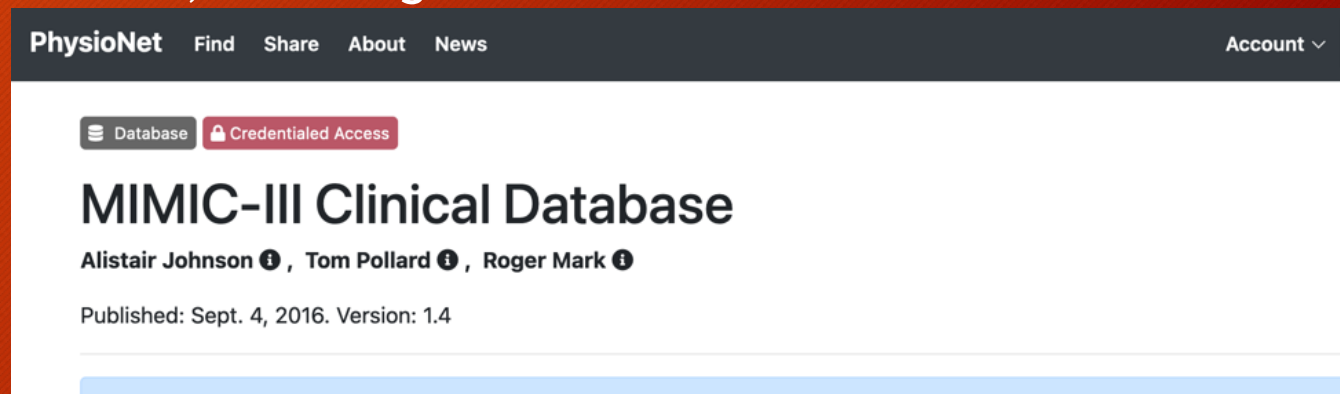
Working with health care data

Example dataset: <https://physionet.org/content/mimiciii/>

MIMIC-III integrates deidentified, comprehensive clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, and makes it widely accessible to researchers internationally under a data use agreement. The open nature of the data allows clinical studies to be reproduced and improved in ways that would not otherwise be possible.

The MIMIC-III database was populated with data that had been acquired during routine hospital care, so there was no associated burden on caregivers and no interference with their workflow. Data was downloaded from several sources, including:

- archives from critical care information systems.
- hospital electronic health record databases.
- Social Security Administration Death Master File.



Working with Data - OS/Human Protections

- OS Protections

- Firewalls
- Access controls
- User permissions
- Group permissions



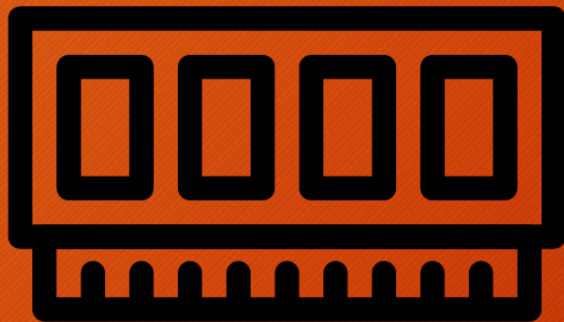
- Human Protections

- Don't reuse passwords
- Don't share passwords
- Report problems ASAP
- Don't move data to other systems
- Keep OS Software current
- Run trusted software
- Specific tasks are performed by specific people

Working with Data - Deep vs Shallow Copies

```
variable A = load_large_dataset()  
B = A[columns 1 to 3]  
C = A[columns 3 to 7]
```

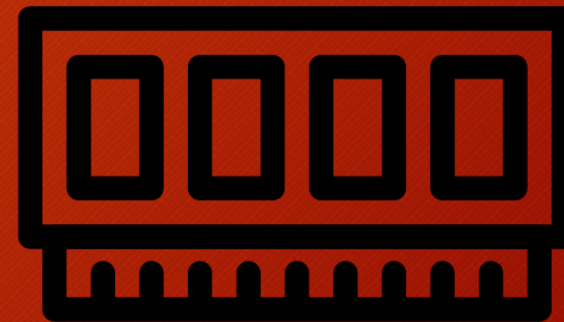
A, B, C



A

B

C



Creating code that others can understand

I'm starting with the man in the mirror
I'm asking him to change his ways
And no message could've been any clearer
If they wanna make the world a better place
Take a look at yourself and then make a change

Glen Ballard and Siedah Garrett; Performed by Michael Jackson, Bad 1987

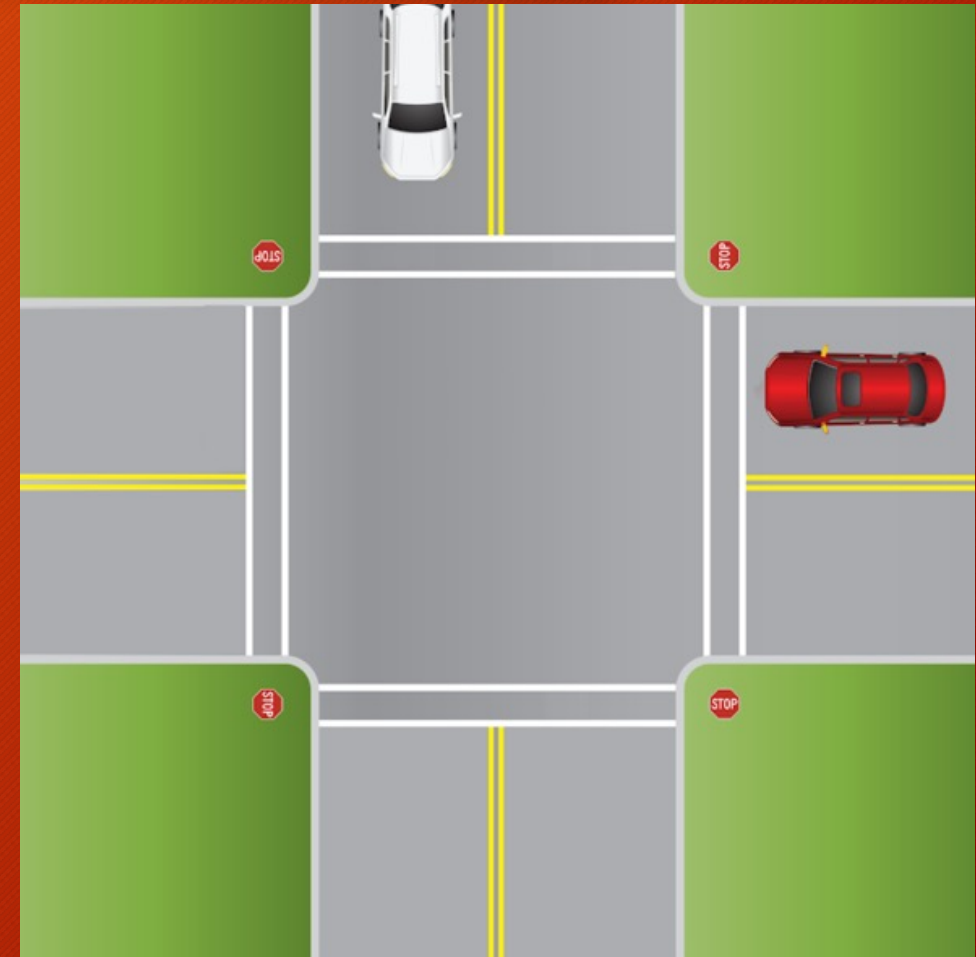
Status check

Close your browser window first if you still have it open

<https://cups.fast.ai/computationalhusbandry>

Project Organization

- Predictable behavior
- Versioning
- Backups
- How to build software/analysis pipeline



Project Naming Conventions

File and folder naming standards

1. Use lower case letters
2. Separate words with “ _ ”

If the language or framework requires uppercase/camelcase or “-” separator, follow those requirements instead.

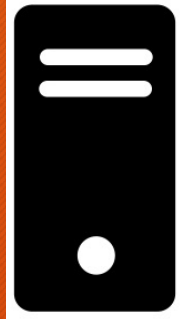
Over all structure and project specific standards should be as follows:

- I. Create teams for all user groups at the Repository Level:
 - a. Ex: Analytics Team, Informatics Team, teams for projects, grants, etc.
 - b. All repositories for a specific team should start with the **team name**:
 - i. EX: **analytics_**”project/grant name”
 - ii. Use all lower case letters
- II. Everyone that creates an account under the CUD2V Organization can set up a repository for their work as well, using the following naming convention:
 - i. EX: name_project/grant
- III. For each repository, follow a standard project structure. At a minimum, there needs to be:
 - i. README.md
 - ii. LICENSE (UC Denver Standard)
 - iii. .gitignore
- IV. If appropriate there should be THREE folders:
 - 1) Documentation
 - If documentation is not stored in the repository, a README.md file should indicate where the documentation is located.
 - 2) data
 - If data not stored in the repository, a README.md file should indicate where the data is located.
 - 3) sourcecode

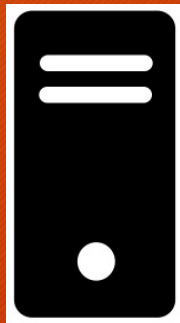
Backups

- If it is important it should be backed up
- If it took a non-trivial amount of work, it should be backed up
- If it changes over time, it should be versioned
- If it is generated by software, it should NOT be backed up
- If you are not sure, back it up

Literate Programming



>



>



Literate Programming

R Markdown

```
1 ---
2 title: "PCCC: An Example Using the Center for Disease Control's Multiple Cause of Death Dataset"
3 author:
4   - James Feinstein
5   - Seth Russell
6   - Tell Bennett
7 date: "`r Sys.Date()`"
8 output: rmarkdown::html_vignette
9 vignette: >
10   %\VignetteIndexEntry{pccc-example}
11   %\VignetteEngine{knitr::rmarkdown}
12   %\VignetteEncoding{UTF-8}
13 ---
14
15 ```{r, include = FALSE}
16 knitr::opts_chunk$set(collapse = TRUE, comment = "#>")
17 ```
18
19 # Introduction
20 This vignette provides an example using publicly available death certificate data to illustrate how the 'pccc' package generates the Complex Chronic
21
22 To evaluate the code chunks in this example you will need to load the
23 following R packages.
24
25 ```{r, message = FALSE}
26 library(pccc)
27 library(dplyr)
28 ```
29
30 # Accessing the Data
31
32 The Center for Disease Control maintains vital statistics including death certificate data. The publicly available death certificate data, known as the
33
34 The data documentation and instructions for direct download are available at: ftp://ftp.cdc.gov/pub/health\_statistics/nchs/datasets/comparability/icd9\_icd10\_ICD09\_ICD10\_comparability\_public\_use\_ASCII
35
36 # Preparing the Data
37
38 For this illustrative example, we have provided just 2 columns of the data for decedents <=21 years old: the ICD-9-CM underlying cause of death diagnosis
39
40 Here's a sample of how the file could be read and processed:
41
42 ```{r, eval = FALSE}
43 # download and unzip file from ftp://ftp.cdc.gov/pub/health_statistics/nchs/datasets/comparability/icd9_icd10_ICD09_ICD10_comparability_public_use_ASCII
44 # columns of interest
45 # start end width description
```

Jupyter

https://github.com/CUD2V/kungfauxpandas/blob/master/sourcecode/python/notebooks/kfp_dork_test.ipynb

Chapter 1: The Problem

Monsters are a tricky lot. They have widely varying statistics and capabilities and many have spent centuries building up their reputations as scary creatures to frighten adventurers from invading their spaces. The smarter varieties of monsters have retained good lawyers through the ages, and it is illegal to publish the "personal" data on any monster.

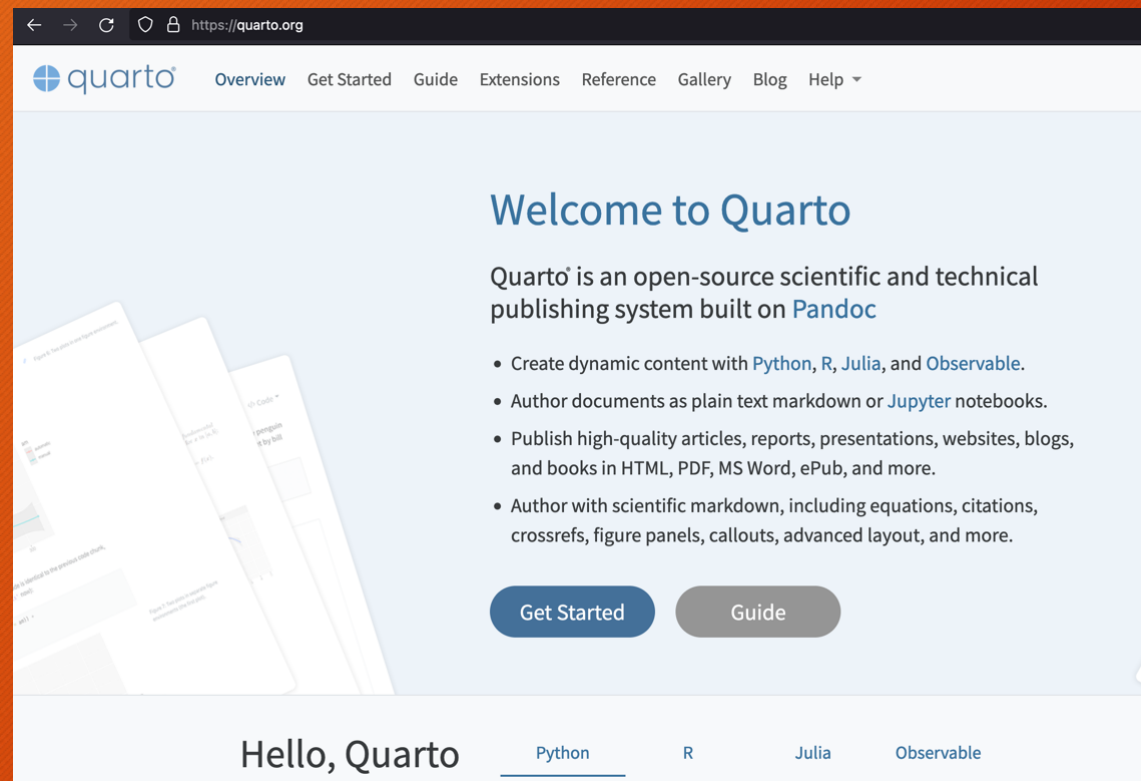
Of course, we would like to study the monsters without violating their privacy rights. KFP can help this problem!

1. Generate a fake data set based on the real stuff
2. Work out all steps necessary to clean the data up
3. Code the data cleaning steps into a function
4. Re-run the data generation using the data cleaner
5. Get a clean, synthetic data set to study
6. Have a trusted agent (DM?) run your study code on the real data and deliver you the results.

```
In [1]: import sqlite3
import pandas as pd
import numpy as np
import scipy as sp
import scipy.stats as stats
import pylab as plt
from collections import Counter
import datetime
import sys
sys.path.append('../')
sys.path.append('../../plugins/DataSynthesizer/DataSynthesizer/')
```


Literate Programming

Quarto



<https://www.rstudio.com/conference/2022/talks/literate-programming-quarto/>

Meta Programming

Programs that write programs

Programs that run programs

Programs that write programs

SQL that writes SQL

```
/* You, a year ago • Updating tables/views to create snapshot in-time ...
Script to create date stamped backup of tables/views that are important to track over time

Could update this to do all tables/views in the covidmab table space or specific named
tables in a table...
*/

DECLARE datestamp STRING;
SET datestamp = (
  SELECT
    substr(table_name, 15)
  FROM
    hdccovidmab.covidmr.INFORMATION_SCHEMA.TABLES
  WHERE table_name like 'table1_patient%'
  ORDER BY table_name DESC
  LIMIT 1
);

EXECUTE IMMEDIATE "CREATE OR REPLACE TABLE hdccovidmab.covidmab.
datestamp || " AS SELECT * FROM hdccovidmab.covidmab.cd4_severit
EXECUTE IMMEDIATE "CREATE OR REPLACE TABLE hdccovidmab.covidmab.
```

Bash that writes SQL

```
#!/bin/bash
# need table created in multiple schemas, so using bash
# to templatzize the script
set -x
query_template=""

for dataset in full harmonized
do
  query_template+="
  CREATE OR REPLACE TABLE \"hdcekapedsepsis1\".${dataset}.tests_checks\" AS

  WITH m AS (
    SELECT DISTINCT
      site,
      test_name_source,
      test_name,
      biennial_admission,
      test_units
```


Programs that run programs

Make

```
TARGETS = /sepsis_data/meds_qa_qc/.split_by_med_name

all: $(TARGETS)

/sepsis_data/meds_qa_qc/meds.csv: ../build_medication_mapping_configuration.sql ../harmonize_medications.sql
→ bq query < $(word 1, $^)
→ bq query < $(word 2, $^)
→ rm -f harmonized_medication/*.csv
→ gsutil -m cp -r "gs://hdcekapedseps1/meds_qa_qc/harmonized_medication" .
→ gsutil rm -f "gs://hdcekapedseps1/meds_qa_qc/harmonized_medication/*.csv"
→ awk 'NR == 1' harmonized_medication/*.csv > /sepsis_data/meds_qa_qc/meds_header.csv
→ awk 'FNR > 1' harmonized_medication/*.csv > $@

/sepsis_data/meds_qa_qc/.split_by_med_name: split_by_med_name.R /sepsis_data/meds_qa_qc/meds.csv
→ R --vanilla -f $<
→ touch $@

/sepsis_data/meds_qa_qc/meds_randomized.csv: /sepsis_data/meds_qa_qc/meds.csv
→ shuf $< > $@

/sepsis_data/meds_qa_qc/meds_25pct.csv: /sepsis_data/meds_qa_qc/meds_randomized.csv
→ split -n 1/1/4 $< > $@

/sepsis_data/meds_qa_qc/meds_50pct.csv: /sepsis_data/meds_qa_qc/meds_randomized.csv
→ split -n 1/1/2 $< > $@
```

<https://www.gnu.org/software/make/>

Programs that run programs

Continuous Integration - Github Actions

```
name: Python package

on: [push]

jobs:
  build:

    runs-on: ubuntu-latest
    strategy:
      matrix:
        python-version: ["3.7", "3.8", "3.9", "3.10"]

    steps:
      - uses: actions/checkout@v3
      - name: Set up Python ${ matrix.python-version }
        uses: actions/setup-python@v4
        with:
          python-version: ${ matrix.python-version }
```

<https://docs.github.com/en/actions>

Status check

Close your browser window first if you still have it open

<https://cups.fast.ai/computationalhusbandry>

Software Testing

It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

A Scandal in Bohemia
Sir Arthur Conan Doyle

Software Testing - What to test

- Identify the most important or unique feature(s) of software being implemented. Software bugs are found to follow a Pareto or Zipfian distribution.
- Test data and software configuration.
- If performance is a key feature, build tests to evaluate performance.

Software Testing - How to test

- Software developer develops unit tests.
- Intended user of software should perform validation/acceptance tests.
- Run all tests regularly.
- Review key algorithms with domain experts.

Software Testing - Antipatterns

- Interdependent tests — When a failure in an early test case breaks a later test, it can cause difficulty in resolution and remediation.
- Testing application performance — Creating an automated test to perform this is difficult and does not carry over well from one machine to another.
- Slow running tests— As much as possible, tests should be automated but still run quickly. If tests are slow they will not be used.
- Only test correct input—A common problem in testing is to only validate expected inputs and desired behavior. Make sure tests cover invalid input, exceptions, and similar items.

Computational Husbandry

questions?

Seth Russell - seth.russell@cuanschultz.edu

Peter DeWitt - peter.dewitt@cuanschultz.edu

SOM, Dept of Biomedical Informatics

Presentations: <https://github.com/magic-lantern/Computational-Husbandry-2022>

Survey: <https://forms.gle/ZumeuAiWRm83ra9h9>