Core Concepts in Data Technologies

An Open Source Tool Chain

Matthew Henderson, PhD, FCACB

Department of Laboratory Medicine

Division of Biochemistry

The Children's Hospital of Eastern Ontario | University of Ottawa

July 7, 2015

Notes

Outline

1 Introduction

2 Fundamental Tools for Data Management

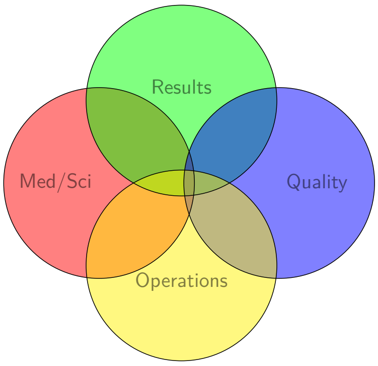
Notes

Information Systems Design

- Single Point of Truth
  - The practice of structuring information models such that every data element is stored exactly once
- Have you looked at your shared network drive lately?

Notes

Laboratory Data Sources



Notes

A Unique Combination of Features

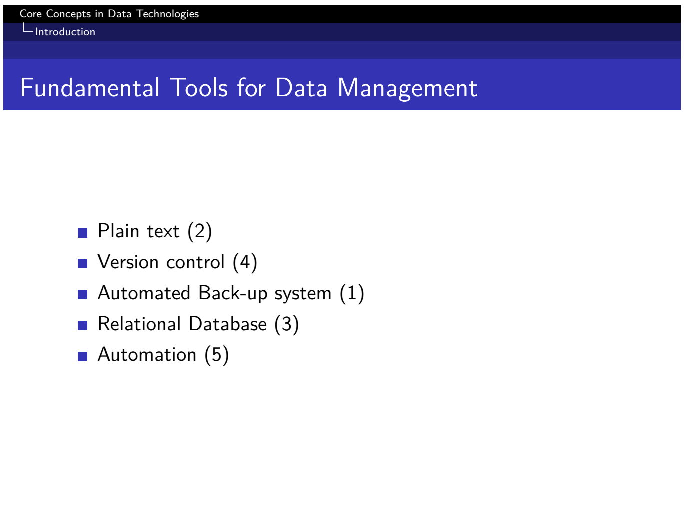
A venomous, electrolocating, egg-laying, duck-billed, beaver-tailed, otter-footed mammal

Notes

Why talk about tools

> *The enjoyment of one's tools is an essential ingredient of successful work.*
>
> *Donald Knuth, Computer Scientist, Turing Award Winner*

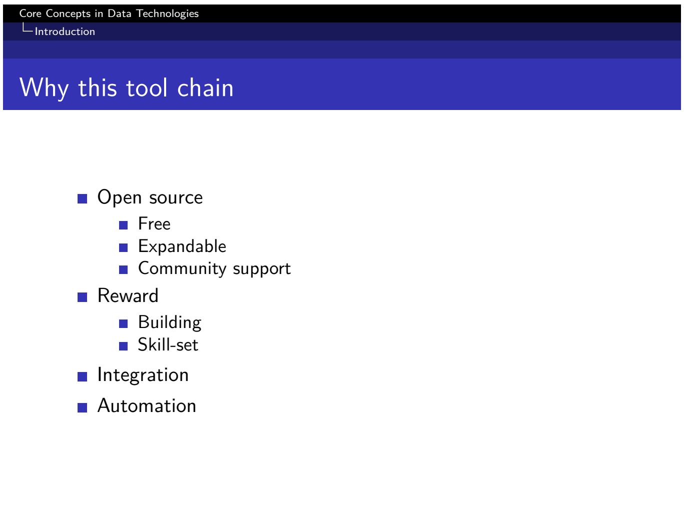
Notes



Fundamental Tools for Data Management

- Plain text (2)
- Version control (4)
- Automated Back-up system (1)
- Relational Database (3)
- Automation (5)

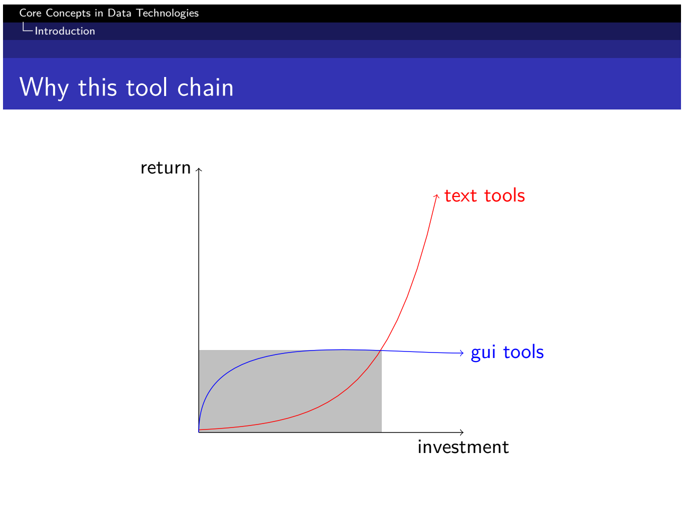
Notes



Why this tool chain

- Open source
  - Free
  - Expandable
  - Community support
- Reward
  - Building
  - Skill-set
- Integration
- Automation

Notes



Why this tool chain

Notes

Plain Text

- Simple data formats: .txt, .csv
  - Read by computers and humans alike.
  - Text editors i.e. Notepad++
    - https://notepad-plus-plus.org/
- Compatibility and Longevity
  - Sophisticated tool chains have been created to manage plain text files
  - 20 year old method validation data - no problem
    - bit rot - .wpd, .doc, .docx, .docxm
    - https://en.wikipedia.org/wiki/List_of_file_formats

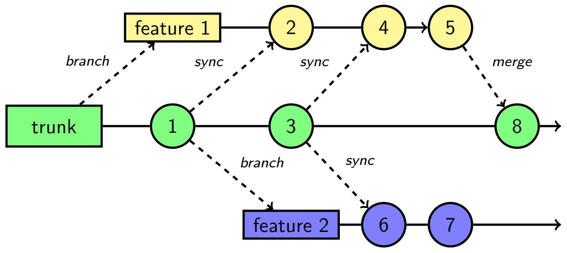
Notes

Plain Text

- Source code
- Markup Languages
  - HTML, XML
- Structured Text
  - HL7, SNOMED CT, LOINC

Notes

Version Control



- Document control software
- Cloud storage services
- Git: https://git-scm.com
  - Github: https://github.com/hendersonmpa/spot-talk

Notes

Automated Back-up

- Automate it!
- Test your system before you need it
- Encrypted cloud storage
  - SpiderOak
    - https://spideroak.com/

Notes

Relational Database

- A collection of data tables
- The tables are part of a **Data Model** called a **Schema**
- The data model defines:
  - The type of data stored in each column
  - The relationship between tables

Notes

Sqlite

SQLite is a software library that implements a self-contained, serverless, zero-configuration, transactional SQL database engine. SQLite is the most widely deployed database engine in the world.

- Sqlite: https://www.sqlite.org/
- Windows installation: https://www.youtube.com/watch?v=VZ20Lh4zbRo
- DB Browser for SQLite: http://sqlitebrowser.org/

Notes

Data Model Concepts

- Entities - samples, physicians, patients, results
- Attributes - names, values, units, reference intervals,
- Relationships
  - Samples come from Patients
  - Results come from Samples

Notes

Database Operations: Structured Query Language

Filter subsetting or removing observations based on some condition

- select, where

Transform adding or modifying variables.

- functions

Aggregate reducing multiple values into a single value

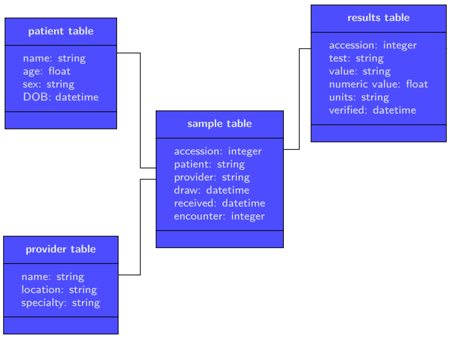
- count, mean, sum with group by

Sort changing the order of observations

- order by

Notes

Schema



patient table

name: string
age: float
sex: string
DOB: datetime

results table

accession: integer
test: string
value: string
numeric value: float
units: string
verified: datetime

sample table

accession: integer
patient: string
provider: string
draw: datetime
received: datetime
encounter: integer

provider table

name: string
location: string
specialty: string

Notes

A Schema for a Single Table

```
-- Make a table
CREATE TABLE "biochemistry" (
‘test‘        TEXT,
‘result‘          NUMERIC,
‘order_date‘      TEXT,
‘patient‘         TEXT,
‘clinic‘          TEXT,
‘physician‘       TEXT);
```

Notes

A month of HbA1c results from the Endo clinic

```
SELECT result, order_date, patient, clinic, physician
FROM biochemistry
WHERE test = ’HbA1C’ AND
clinic = ’clinic_*B7’ AND
order_date BETWEEN ’2014-03-01’ AND ’2014-05-01’
ORDER BY order_date;
```

Notes

Output from the database

```
|--------+---------------------+------------------+------------+-------------|
| result | order_date          | patient          | clinic     | physician   |
|--------+---------------------+------------------+------------+-------------|
|    5.3 | 2014-03-01 09:59:06 | patient_*4C96CD5 | clinic_*B7 | phys_*B13FF |
|    6.0 | 2014-03-01 10:10:09 | patient_*842DEC3 | clinic_*B7 | phys_*B13FF |
|    4.5 | 2014-03-01 10:32:04 | patient_*CD42144 | clinic_*B7 | phys_*B13FF |
|    6.0 | 2014-03-01 11:25:08 | patient_*A85C417 | clinic_*B7 | phys_*8449D |
|    5.5 | 2014-03-01 12:05:05 | patient_*2BC50ED | clinic_*B7 | phys_*B13FF |
|    4.6 | 2014-03-01 14:44:05 | patient_*B3B5C6E | clinic_*B7 | phys_*B13FF |
|    5.6 | 2014-03-01 14:45:02 | patient_*36E9661 | clinic_*B7 | phys_*B13FF |
|    7.8 | 2014-03-01 14:48:04 | patient_*4FE70F0 | clinic_*B7 | phys_*8449D |
|    8.8 | 2014-03-01 18:01:02 | patient_*4C303D5 | clinic_*B7 | phys_*A939A |
|    5.1 | 2014-03-04 10:14:03 | patient_*C7A4177 | clinic_*B7 | phys_*B13FF |
...
```

Notes

Top ten ordering physicians

```
SELECT count(test) AS count, physician FROM biochemistry
WHERE test = ’HbA1C’ AND
order_date BETWEEN ’2014-03-01’ AND ’2014-05-01’
GROUP BY physician
ORDER BY count DESC
LIMIT 10;
```

Notes

Output from the database

```
|-------+-------------|
| count | physician   |
|-------+-------------|
|   168 | phys_*B13FF |
|   167 | phys_*C6301 |
|   161 | phys_*33AC2 |
|   161 | phys_*8449D |
|   140 | phys_*12F17 |
|   123 | phys_*B9396 |
|   110 | phys_*CEC56 |
|   108 | phys_*0698F |
|   107 | phys_*E6DBB |
|    96 | phys_*B0395 |
|-------+-------------|
```

Notes

Number of HbA1c Orders by Day of the Week

```
SELECT STRFTIME(’%w’,order_date) AS day ,
COUNT(STRFTIME(’%w’,order_date)) AS count
FROM biochemistry
WHERE test = "HbA1C"
GROUP BY day;
```

Notes

Output from the database

```
|-----+-------|
| day | count |
|-----+-------|
|   0 |  1027 |
|   1 |   883 |
|   2 |  6358 |
|   3 |  6881 |
|   4 |  7333 |
|   5 |  6578 |
|   6 |  5940 |
|-----+-------|
```

- 0 = Sunday

Notes

Database vs Spreadsheet

Pros

- Data integrity
  - types
  - table level write access
- Automation
  - Pipeline
- Scale
- Relational model

Cons

- Set-up
- Initial Investment

Notes

Why Script?

- A record of your work
- Incremental refinement
  - Forced to think through every step
  - Avoid spending effort recreating
  - Reproducible results
  - Focus on refining and building
  - Plan, Do, Check, Act in minutes
- Gradually gain insight into data and processes

Notes

First steps to automation

```
-- Select all HbA1c results in a date range
SELECT result, order_date, patient, clinic, physician
FROM biochemistry
WHERE test = ’HbA1C’ AND
clinic = ’clinic_*B7’ AND
order_date BETWEEN ’2014-03-01’ AND ’2014-05-01’
ORDER BY order_date;

-- Find the top ten ordering physician for a given test
SELECT count(test) AS count, physician FROM biochemistry
WHERE test = ’HbA1C’ AND
order_date BETWEEN ’2014-03-01’ AND ’2014-05-01’
GROUP BY physician ORDER BY count DESC LIMIT 10;

-- Weekly ordering practices
SELECT STRFTIME(’%w’,order_date) AS day ,
COUNT(STRFTIME(’%w’,order_date)) AS count FROM biochemistry
WHERE test = "HbA1C" GROUP BY day;
```

Notes

References

Introduction to Data Technologies
https://www.stat.auckland.ac.nz/p̃aul/ItDT/

Notes

Notes