Database Theory and Applications for Biomedical Research and Practice

BMIN 502 / EPID 635
Week 14: Database administration

John H. Holmes, PhD

Institute for Biomedical Informatics

Agenda for today

- Review of Assignment 12
- Special considerations for databases used in biomedical research
- Monitoring and logging usage
- Database security
- Ensuring data integrity
- Data management plans

Special considerations for databases used in biomedical research

Two critical questions…

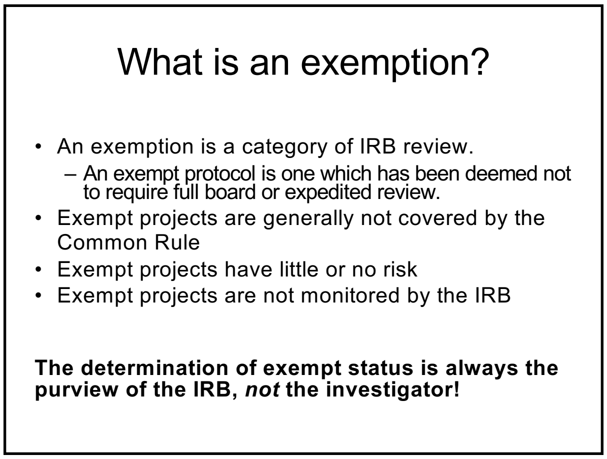
1. Does your proposed research involve human subjects?
2. Is your project exempt from human subjects considerations?

What is a human subject?

- Federal regulations (45 CFR Part 46) define a **human subject** as a living individual about whom an investigator conducting research obtains:
  - Data through intervention or interaction with the individual, or
  - Identifiable private information

What is included in the definition?

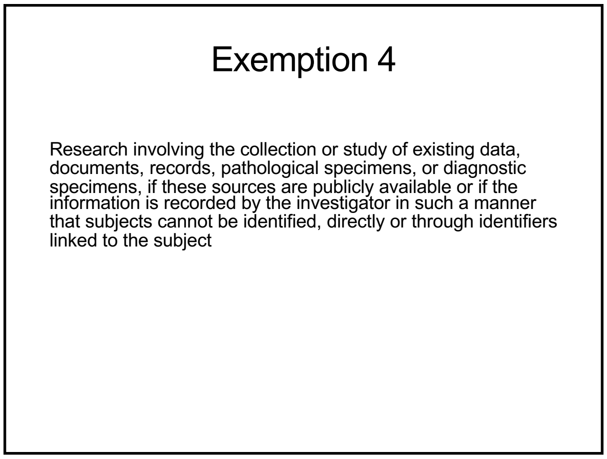
- Use of human organs, tissues, and body fluids, as well as graphic, written, or recorded information, from living individuals if the identity of the subjects can be readily ascertained by the investigator or other members of the research team



What is an exemption?

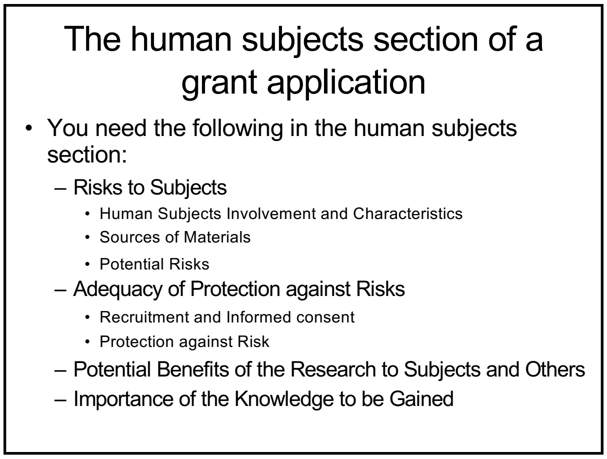
- An exemption is a category of IRB review.
  - – An exempt protocol is one which has been deemed not to require full board or expedited review.
- Exempt projects are generally not covered by the Common Rule
- Exempt projects have little or no risk
- Exempt projects are not monitored by the IRB

The determination of exempt status is always the purview of the IRB, *not* the investigator!



Exemption 4

Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subject



The human subjects section of a grant application

- You need the following in the human subjects section:
  - – Risks to Subjects
    - Human Subjects Involvement and Characteristics
    - Sources of Materials
    - Potential Risks
  - – Adequacy of Protection against Risks
    - Recruitment and Informed consent
    - Protection against Risk
  - – Potential Benefits of the Research to Subjects and Others
  - – Importance of the Knowledge to be Gained

NIH Data Sharing Policy

- Basic premise
  - Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data.
- Data sharing plan
  - Required by NIH for grants >$500K direct costs
  - All research that involves human subjects and laboratory research that does not involve human subjects

What is shared?
Final Research Data

- Recorded factual material commonly accepted in the scientific community as necessary to document and support research findings
- Final research data do not include laboratory notebooks, partial datasets, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects
- NOT summary or aggregated data

Data Sharing Plan
Example 1 (survey data)

This application requests support to collect public-use data from a survey of more than 22,000 Americans over the age of 50 every 2 years. Data products from this study will be made available without cost to researchers and analysts. https://ssl.isr.umich.edu/hrs/

User registration is required in order to access or download files. As part of the registration process, users must agree to the conditions of use governing access to the public release data, including restrictions against attempting to identify study participants, destruction of the data after analyses are completed, reporting responsibilities, restrictions on redistribution of the data to third parties, and proper acknowledgement of the data resource. Registered users will receive user support, as well as information related to errors in the data, future releases, workshops, and publication lists. The information provided to users will not be used for commercial purposes, and will not be redistributed to third parties.

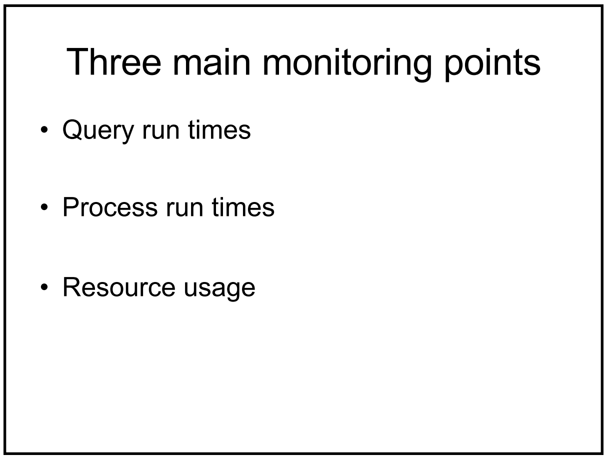
Data Sharing Plan
Example 2 (sensitive data)

The proposed research will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner city STD clinic. The final dataset will include self-reported demographic and behavioral data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

Data Sharing Plan
Example 3 (rationale for not sharing)

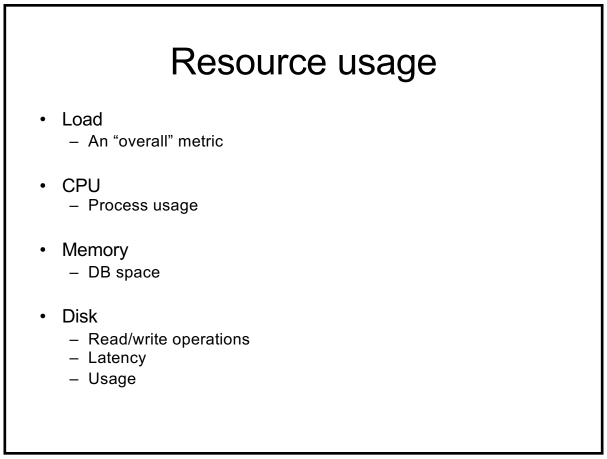
The proposed research will involve a small sample (less than 20 subjects) recruited from clinical facilities in the New York City area with Williams syndrome. This rare craniofacial disorder is associated with distinguishing facial features, as well as mental retardation. Even with the removal of all identifiers, we believe that it would be difficult if not impossible to protect the identities of subjects given the physical characteristics of subjects, the type of clinical data (including imaging) that we will be collecting, and the relatively restricted area from which we are recruiting subjects. Therefore, we are not planning to share the data.

Monitoring and logging usage



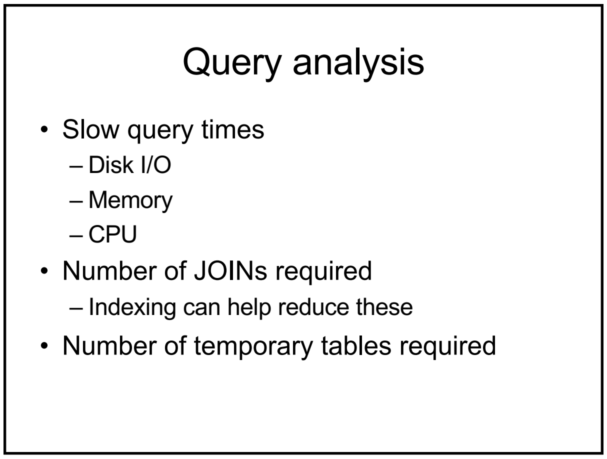
Three main monitoring points

- Query run times
- Process run times
- Resource usage



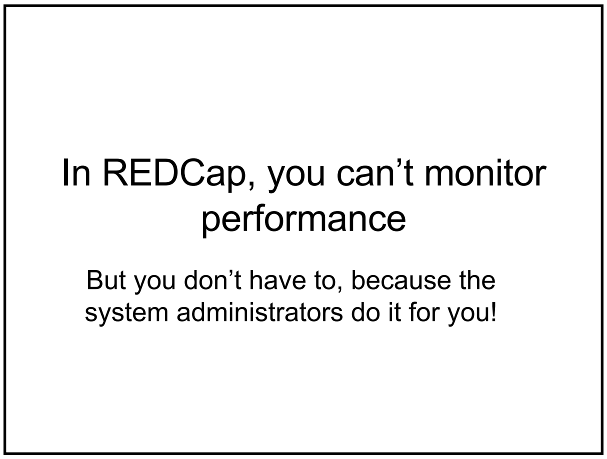
Resource usage

- Load
  - An “overall” metric
- CPU
  - Process usage
- Memory
  - DB space
- Disk
  - Read/write operations
  - Latency
  - Usage



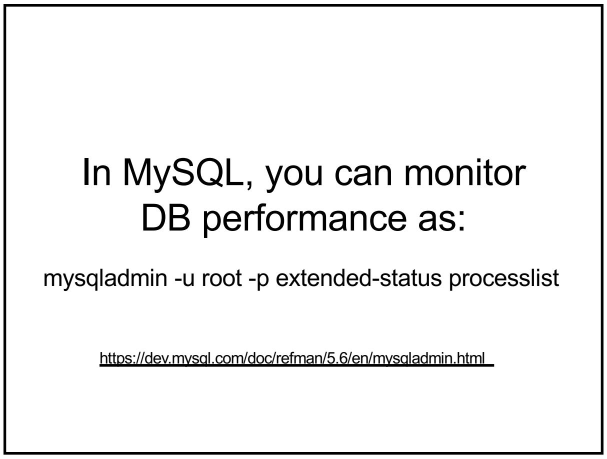
Query analysis

- Slow query times
  - Disk I/O
  - Memory
  - CPU
- Number of JOINs required
  - Indexing can help reduce these
- Number of temporary tables required



In REDCap, you can't monitor performance

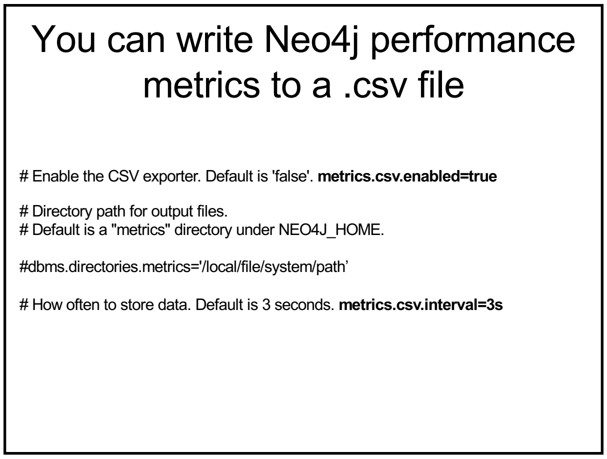
But you don't have to, because the system administrators do it for you!



In MySQL, you can monitor DB performance as:

```
mysqladmin -u root -p extended-status processlist
```

https://dev.mysql.com/doc/refman/5.6/en/mysqladmin.html



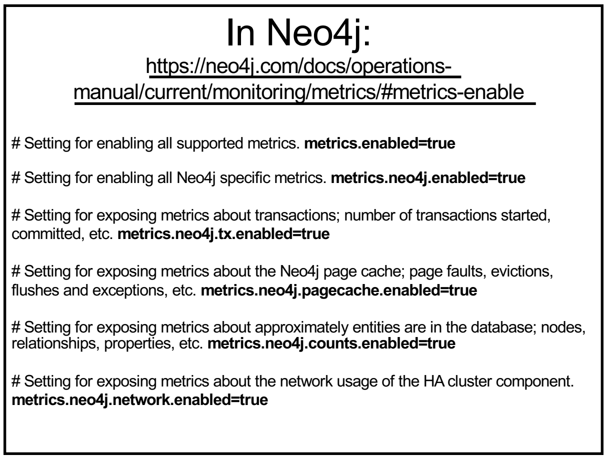
You can write Neo4j performance metrics to a .csv file

```
# Enable the CSV exporter. Default is 'false'. metrics.csv.enabled=true

# Directory path for output files.
# Default is a "metrics" directory under NEO4J_HOME.

#dbms.directories.metrics='/local/file/system/path'

# How often to store data. Default is 3 seconds. metrics.csv.interval=3s
```



In Neo4j:

https://neo4j.com/docs/operations-manual/current/monitoring/metrics/#metrics-enable

# Setting for enabling all supported metrics. **metrics.enabled=true**

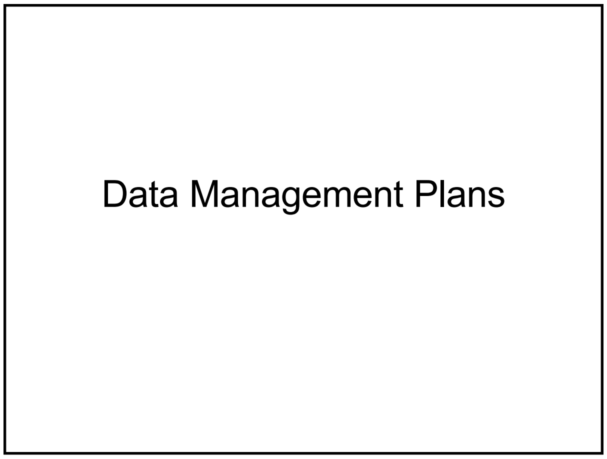
# Setting for enabling all Neo4j specific metrics. **metrics.neo4j.enabled=true**

# Setting for exposing metrics about transactions; number of transactions started, committed, etc. **metrics.neo4j.tx.enabled=true**

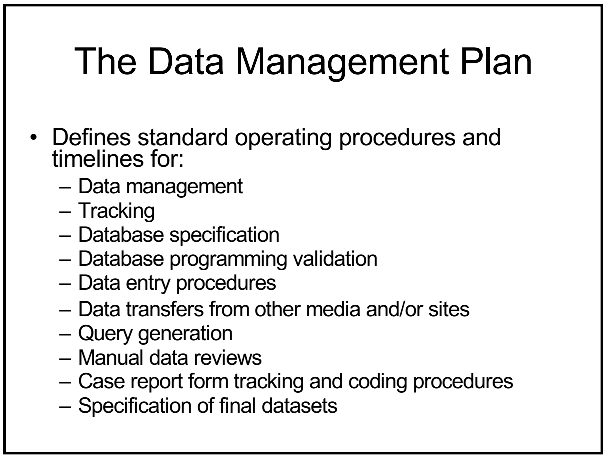
# Setting for exposing metrics about the Neo4j page cache; page faults, evictions, flushes and exceptions, etc. **metrics.neo4j.pagecache.enabled=true**

# Setting for exposing metrics about approximately entities are in the database; nodes, relationships, properties, etc. **metrics.neo4j.counts.enabled=true**

# Setting for exposing metrics about the network usage of the HA cluster component. **metrics.neo4j.network.enabled=true**



Data Management Plans



The Data Management Plan

- Defines standard operating procedures and timelines for:
  - Data management
  - Tracking
  - Database specification
  - Database programming validation
  - Data entry procedures
  - Data transfers from other media and/or sites
  - Query generation
  - Manual data reviews
  - Case report form tracking and coding procedures
  - Specification of final datasets

The DMP Table of Contents

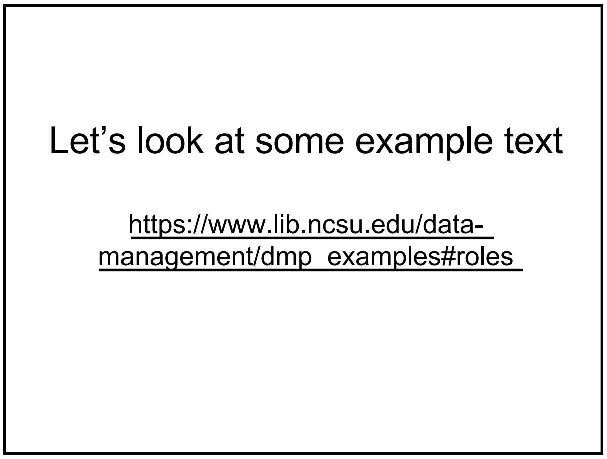
1. Project team
2. Project timeline
3. CRF development and completion instructions
4. Screen designs and database development
5. Application development and validation
6. Backup procedures
7. Disaster recovery
8. Application Approval
9. User manuals and training
10. System maintenance

The DMP Table of Contents, contd.

11. Change control procedures
12. Error handling procedures
13. Data flow model
14. Quality assurance procedures
15. Data monitoring
16. Data entry procedures
17. CRF retrieval and tracking
18. Data validation specifications
19. Data validation programming and testing
20. Query generation, validation, and review

The DMP Table of Contents, contd.

21. Transfer of electronic data
22. Data coding procedures
23. Database validation procedures
24. Data model
25. Data dictionary

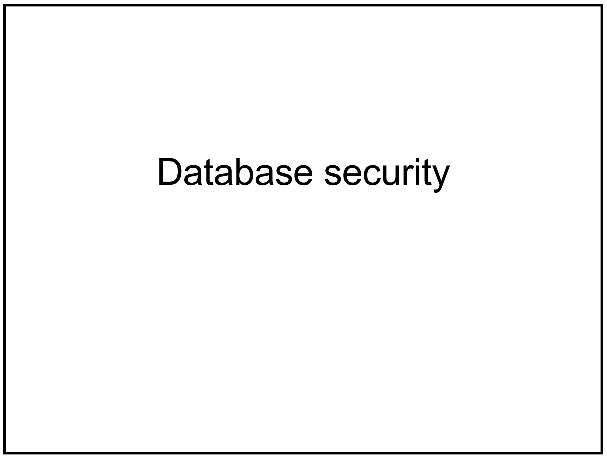


Let’s look at some example text

https://www.lib.ncsu.edu/data-management/dmp_examples#roles

A nice online tool for creating a DMP

https://dmptool.org



Database security

Why do we care about data security?

http://www.healthcareitnews.com/projects/biggest-healthcare-data-breaches-2018-so-far

Physical Environment Security

- Locked files
- Locked offices
- File and office access tracking

Computer hardware security

- Lock-down devices
- Password protection
- Firewalls

An example password policy

- Passwords are changed every 60 days
- Passwords must contain at least eight characters
- Passwords must contain at least three non-alphanumeric characters
- Alphanumeric characters must be mixed case
- The previous 10 passwords cannot be reused

An example “strong” password

Ch^Rl1e_Br0wN

Translates to “Charlie Brown”

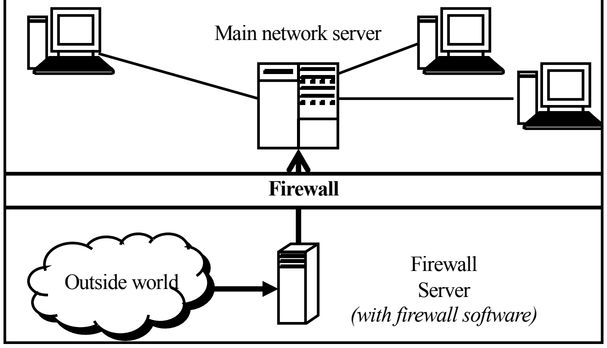
Is this really that strong? Let’s test it…

http://www.passwordmeter.com

Firewalls

- An extra layer of protection between a computer system or network and a larger network, usually the Internet
- Components
  - Firewall software
  - Firewall server

A typical firewall


Main network server

Firewall

Outside world

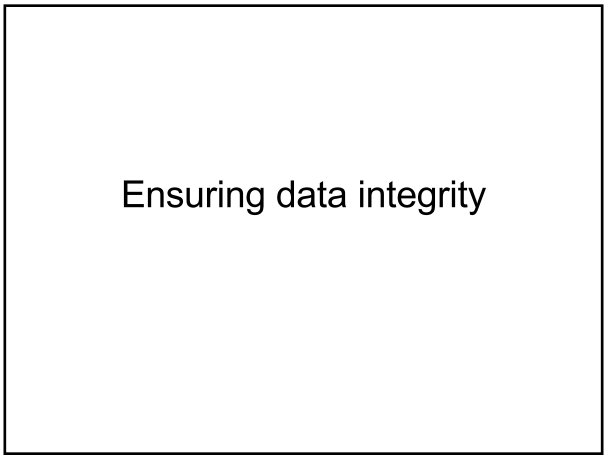
Firewall
Server
(with firewall software)

Personnel Security

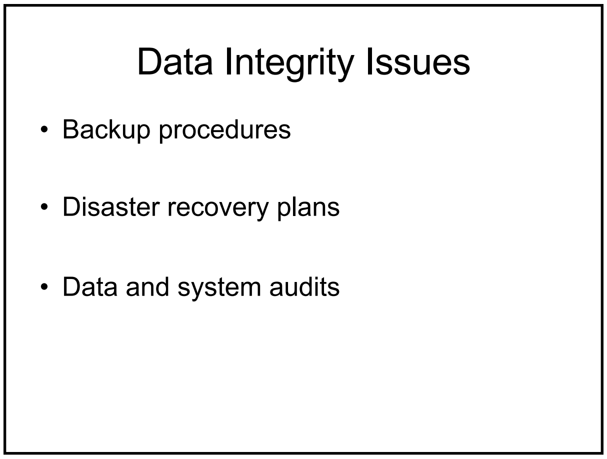
- Background checks
- Inservice education
- Periodic performance evaluations

Research subject security

- Protection of data confidentiality
- Protection of subject privacy

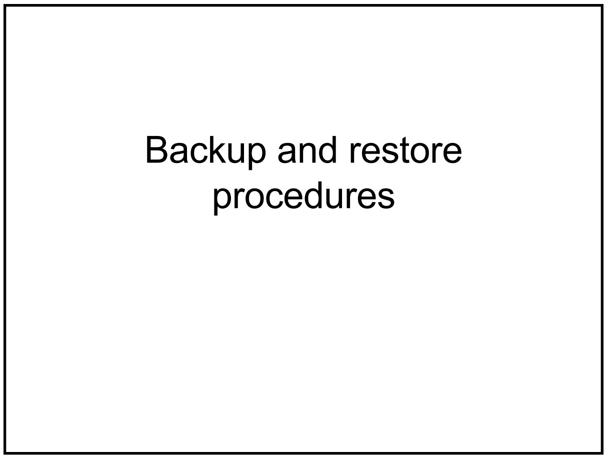


Ensuring data integrity

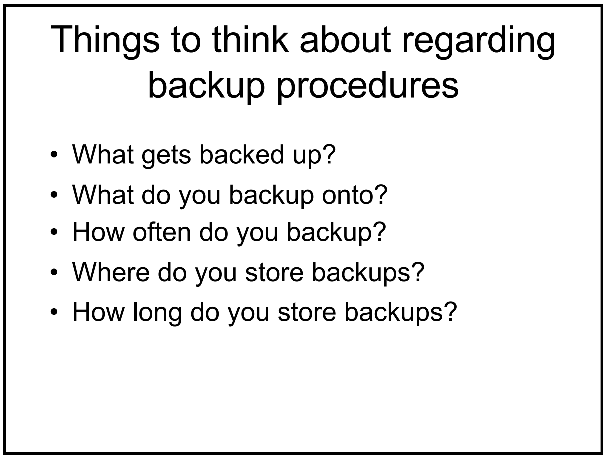


Data Integrity Issues

- Backup procedures
- Disaster recovery plans
- Data and system audits

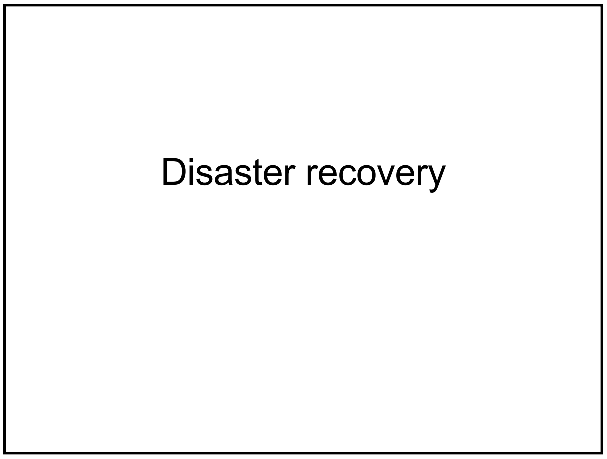


Backup and restore procedures

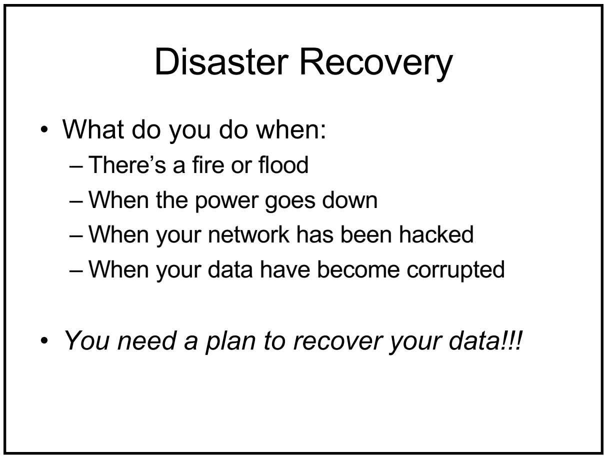


Things to think about regarding backup procedures

- What gets backed up?
- What do you backup onto?
- How often do you backup?
- Where do you store backups?
- How long do you store backups?

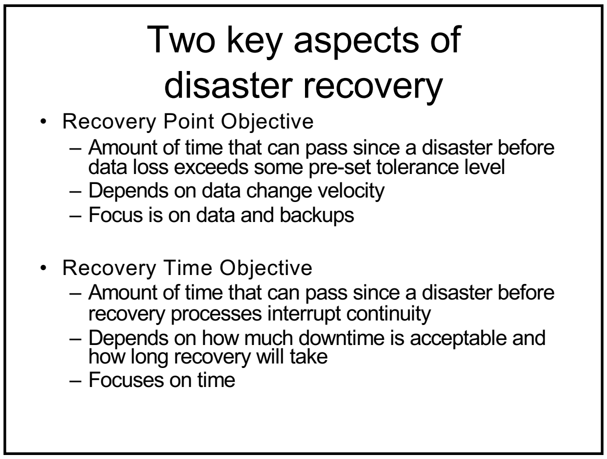


Disaster recovery



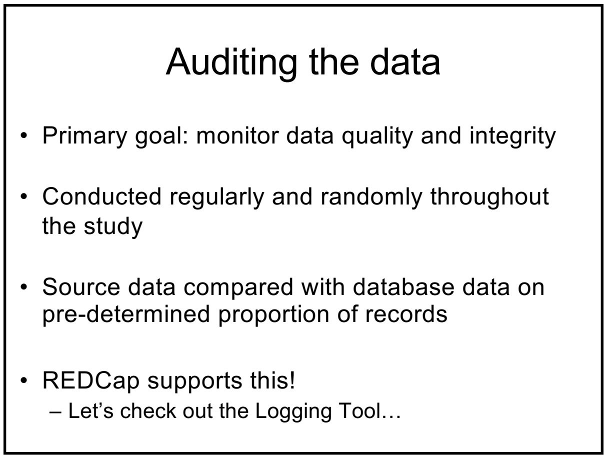
Disaster Recovery

- What do you do when:
  - There’s a fire or flood
  - When the power goes down
  - When your network has been hacked
  - When your data have become corrupted
- *You need a plan to recover your data!!!*



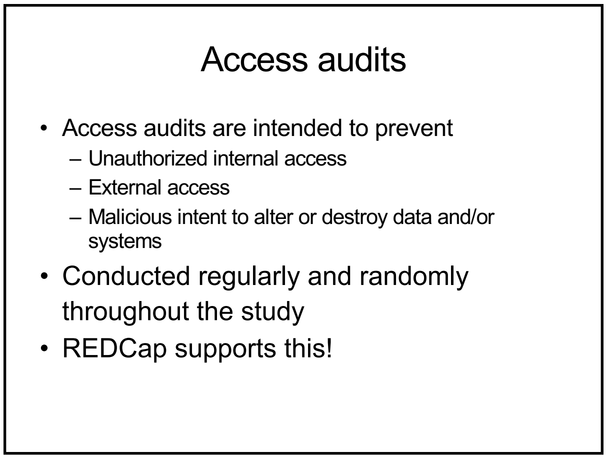
Two key aspects of disaster recovery

- Recovery Point Objective
  - Amount of time that can pass since a disaster before data loss exceeds some pre-set tolerance level
  - Depends on data change velocity
  - Focus is on data and backups
- Recovery Time Objective
  - Amount of time that can pass since a disaster before recovery processes interrupt continuity
  - Depends on how much downtime is acceptable and how long recovery will take
  - Focuses on time



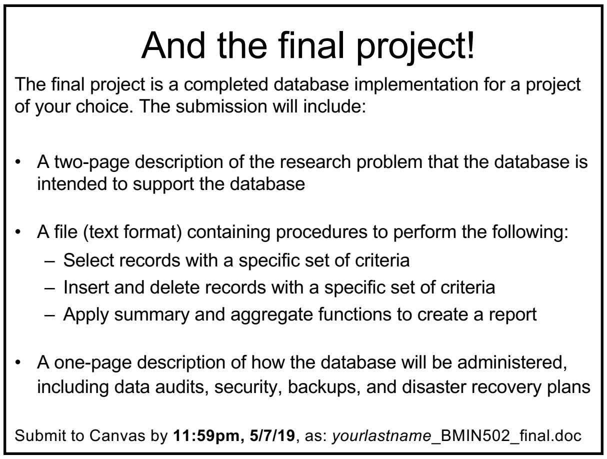
Auditing the data

- Primary goal: monitor data quality and integrity
- Conducted regularly and randomly throughout the study
- Source data compared with database data on pre-determined proportion of records
- REDCap supports this!
  - Let’s check out the Logging Tool…



Access audits

- Access audits are intended to prevent
  - Unauthorized internal access
  - External access
  - Malicious intent to alter or destroy data and/or systems
- Conducted regularly and randomly throughout the study
- REDCap supports this!



And the final project!

The final project is a completed database implementation for a project of your choice. The submission will include:

- A two-page description of the research problem that the database is intended to support the database
- A file (text format) containing procedures to perform the following:
  - Select records with a specific set of criteria
  - Insert and delete records with a specific set of criteria
  - Apply summary and aggregate functions to create a report
- A one-page description of how the database will be administered, including data audits, security, backups, and disaster recovery plans

Submit to Canvas by **11:59pm, 5/7/19**, as: *yourlastname*_BMIN502_final.doc