

# How Many Copies Are Enough?

MIT Digital Document Preservation Simulation  
New England NDSA Meeting 2015

**Micah Altman**

**Richard Landau**

- MIT Libraries, Program on  
Information Science

- <http://informatics.mit.edu>

# Problem to Solve

---

- You have a large, valuable, digital document collection
  - How many copies do you need to keep it safe?
  - On what quality level of servers?
  - How often should you audit the servers?
  - Do they still have all the docs?
- Not much hard data on which to base policy decisions
- We are trying to provide some data, admittedly hypothetical

# Assumptions

---

- Everything costs \$
  - Copies
  - Higher quality services to store your docs
  - Data generally not available
  - Bandwidth
  - Bandwidth for auditing
- Trying to provide data you can use for your policies

# Basic Data to Extrapolate

---

- Not keyed to any specific problems
- Many hints on how to extrapolate from our data to your situations
  - Number of docs, doc sizes, storage shelf sizes
  - Server failure rates
  - Audit strategies
- Very preliminary results

# Two Type of Failures

- Common: A copy of a document dies on a particular server
- Less common: A server dies, losing all the documents it contains
  - "Institutional failure:" fire, flood, war, economic downturn, realignment of purpose, failure of credit arrangements, etc.
- All failures are silent (to the client = library)
- Auditing is \*really\* essential

# Form of the Data

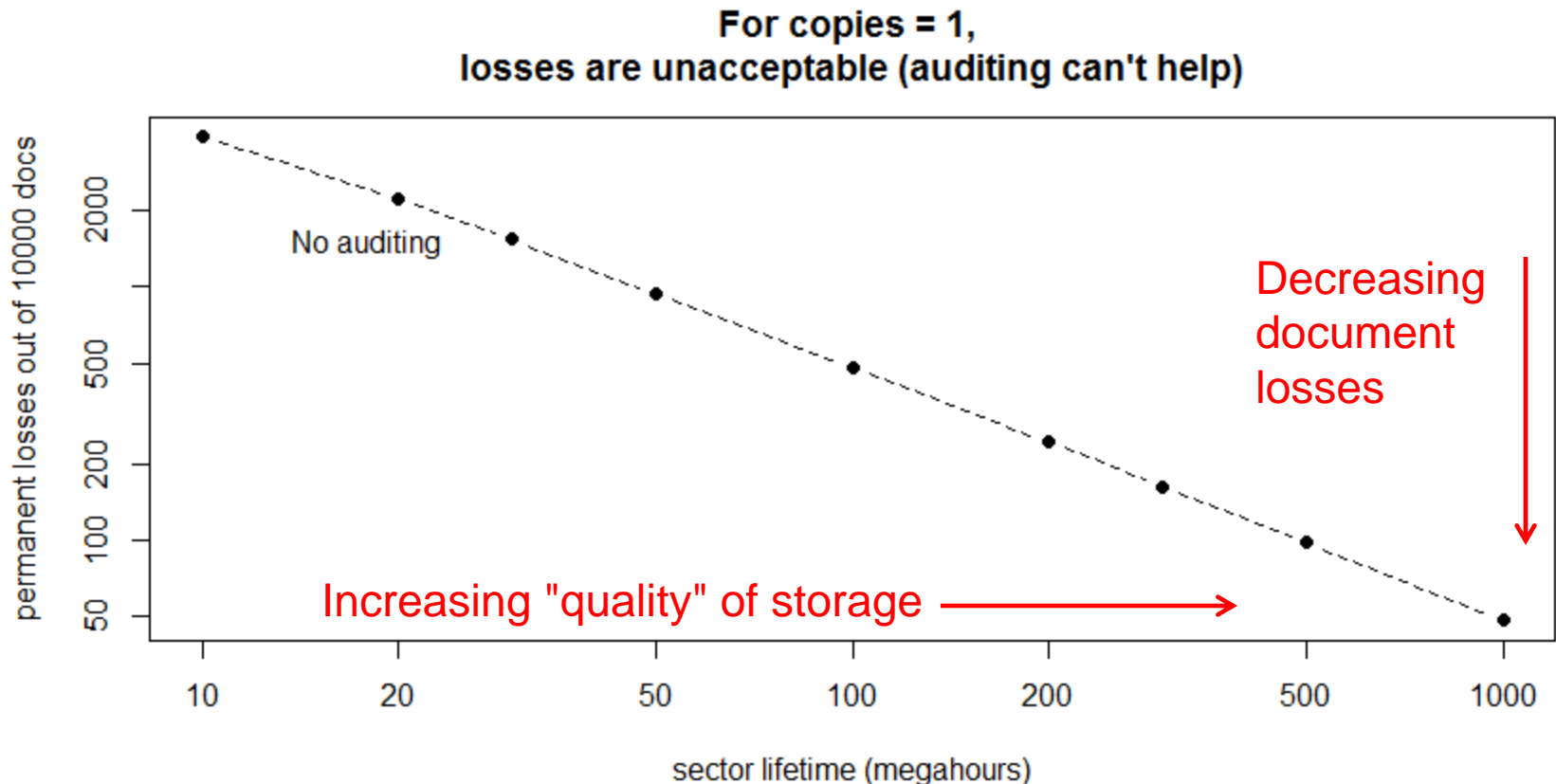
- Fixed number of documents, fixed time
  - Scale to your needs
- Number of copies varies, 1 to 10
- Reliability of storage servers varies
  - Very little real data in this area
- Auditing strategies vary
- Document size varies (but doesn't matter)
- "Glitches:" from minor A/C failures or bad batch of disks, up to institutional failure

# Digital Simulation Programs

---

- Input = error rates, numbers of copies, auditing rate, etc.
- Output = number of documents permanently lost over the life of the test
- Failures happen at random intervals
  - Audits are regularly scheduled
- The programs are "open source:" will be freely available for others to use, test, verify
  - "It's just computer time."

# One Copy? All Losses Are Permanent Losses



Length of time for storage to deteriorate and lose a document

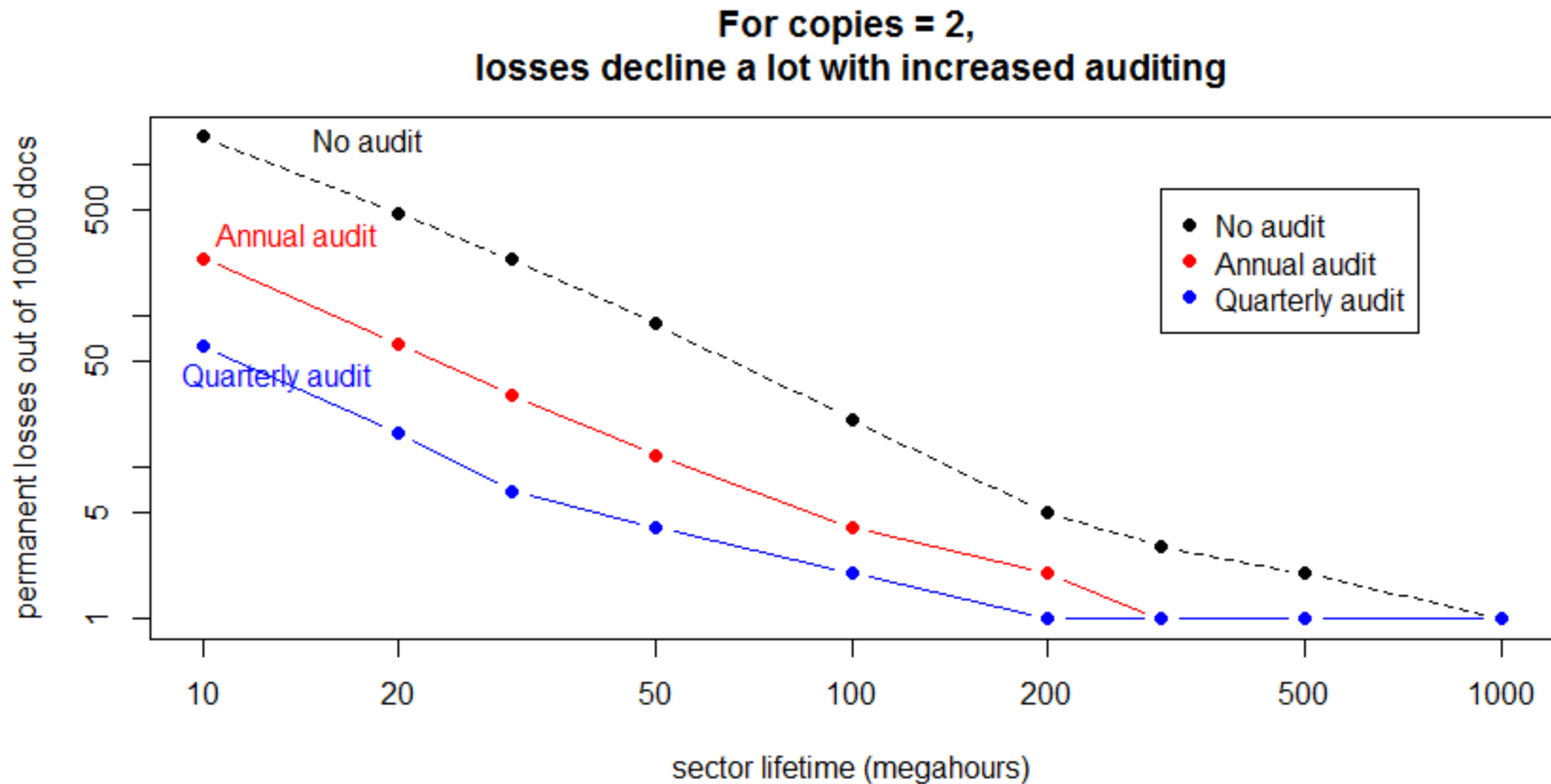


# Good News & Bad News

---

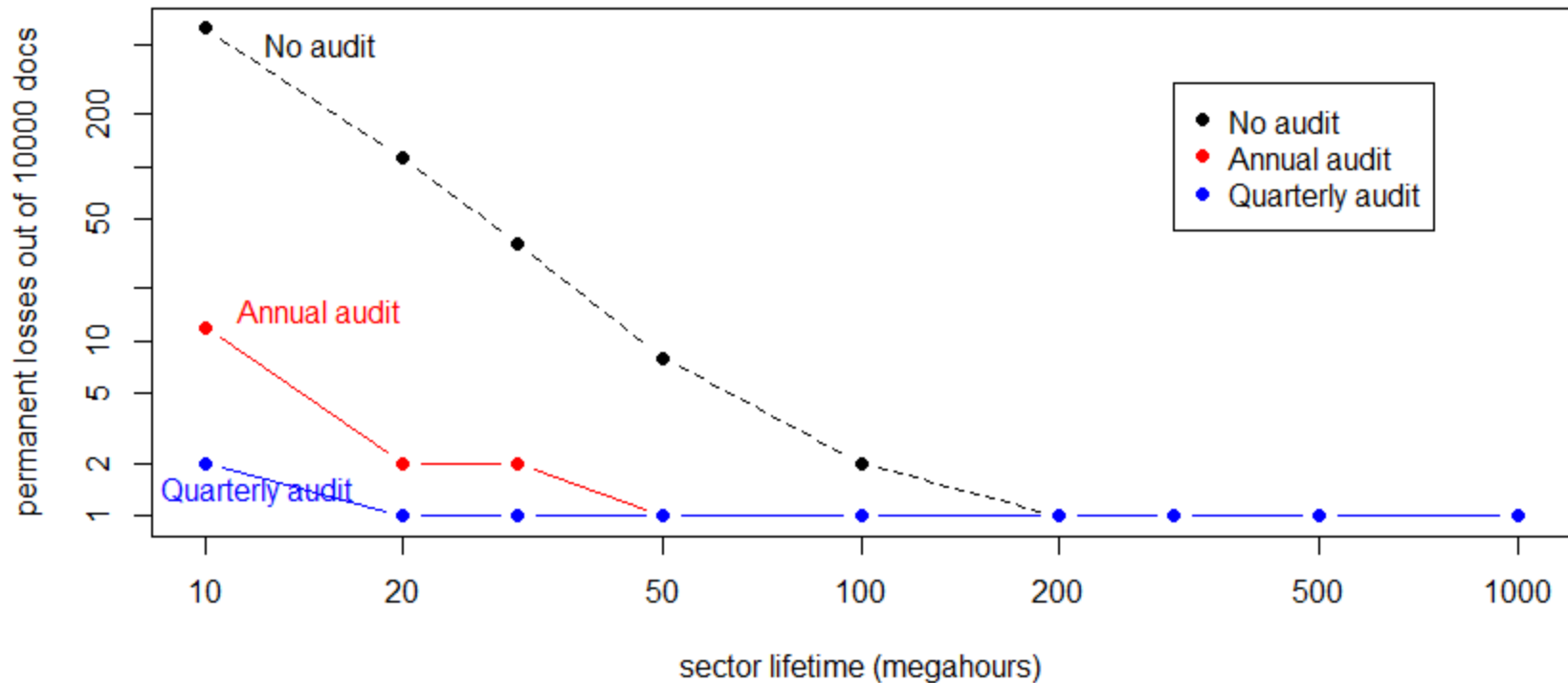
- Bad news: if anyone really understands the "quality" of modern storage, he/she is not talking
- Good news: we \*think\* that modern storage methods, like those used in cloud storage, are very high quality

# Two Copies? Definitely Not Enough



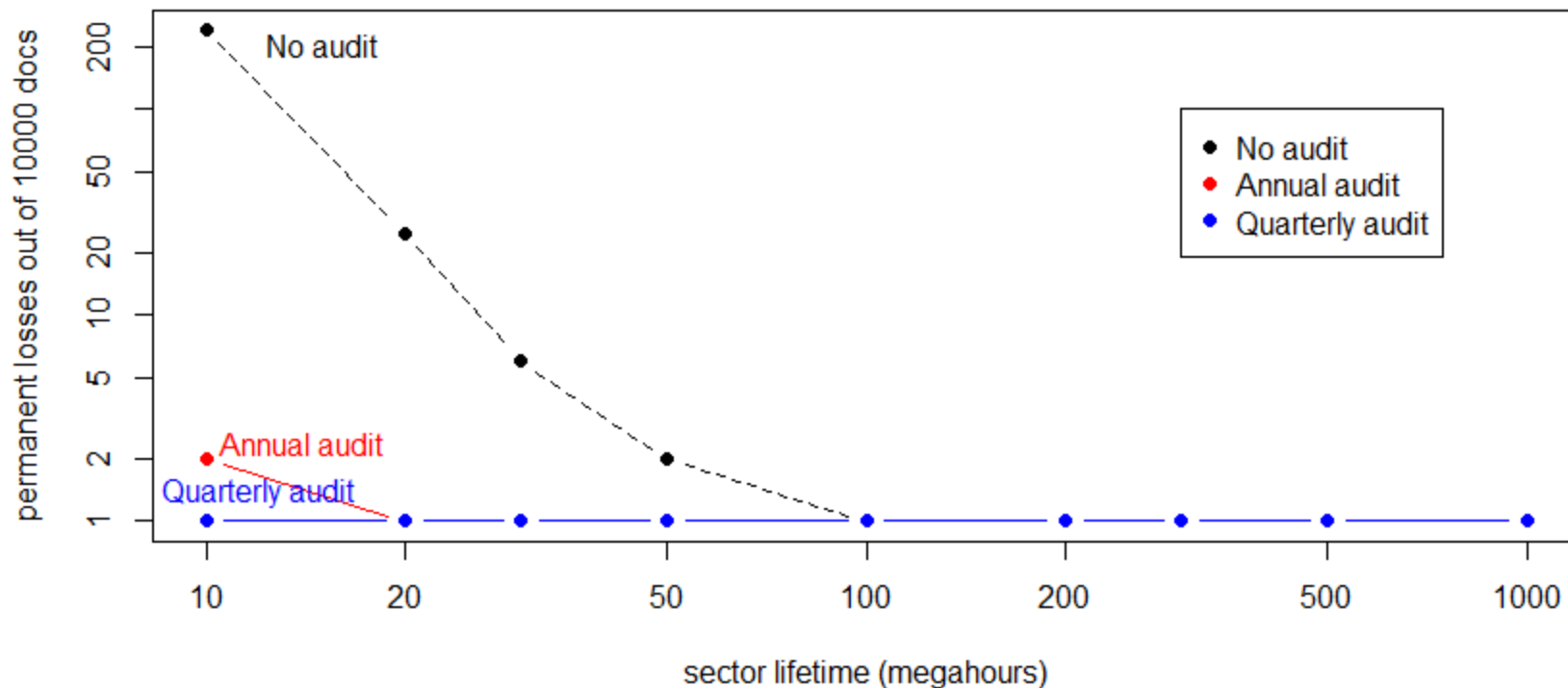
# Three Copies? Marginal

For copies = 3,  
losses decline slightly with increased auditing



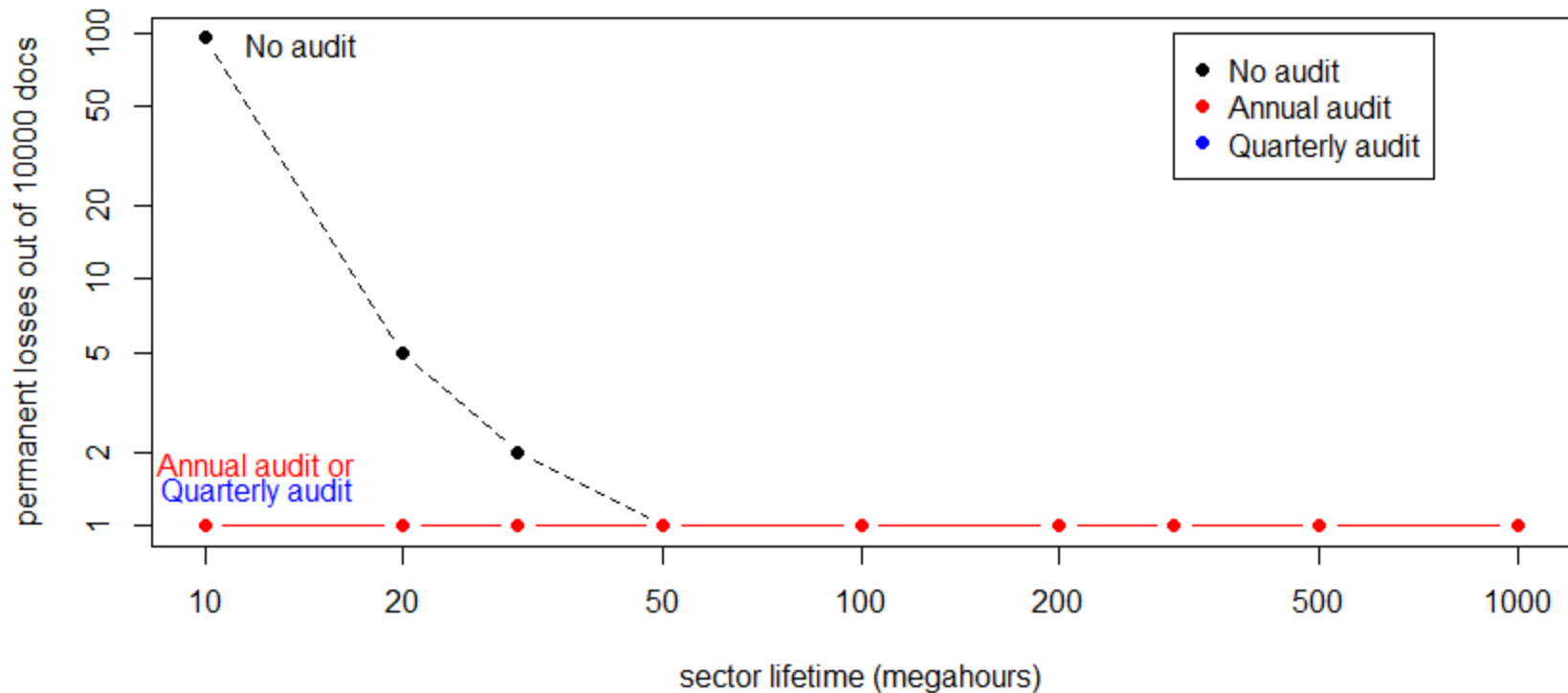
# Four Copies? Looks Good

For copies = 4,  
losses decline only very slightly with increased auditing



# Five Copies? Works

For copies = 5,  
losses are probably negligible with any auditing



# Preliminary Conclusions - 1

---

- More copies are better (duh!)
- Auditing is essential to collection health
  - Very frequent auditing is probably overkill
  - Auditing is expensive (in bandwidth, bytes moved, time)
  - Tricky auditing (subsets, random) are less effective
- We should work toward an efficient cryptographic auditing function
- Simple glitches increase observed error rate

# Preliminary Conclusions - 2

- Institutional failures are pernicious -- but how often do they occur?
  - The problem: a silent institutional failure reduces the number of redundant copies you have stored
    - Thought you had four copies? Well, for a period of time, you actually have only three.
    - And another failure before the audit reduces copies to two
    - Failures may be correlated due to economic conditions
  - Risk is increased until you discover the problem (in auditing) and provision a new server

# Preliminary Conclusions - 3

- How many copies do you need to limit losses?
  - How many to limit permanent losses to some fraction of the collection?
  - Better: How many to keep likelihood of *any* permanent loss under some percentage?
    - 5 per cent, 1 per cent, 0.1 per cent?
  - For institutional failures, particularly correlated, how many copies to keep likelihood of total loss under some percentage?
- What kind of numbers do you need to manage your libraries?



# Backup

---

---

[Ten-ish slides is probably too many for a short session.  
How long is the session?  
How active is the audience likely to be in questioning?  
Relegate some of the slides to "backup." You know the audience; you pick.]