Preserving the bits of the Danish Internet

Niels H. Christensen
Royal Library of Denmark
& netarchive.dk

Outline of this talk

- Problem definition: How repository design relates to repository longevity.
- Method: Computer simulations of repositories
- Sample results: What can the simulations tell us?
- Conclusion, future work.

Human memory

Human memory

- In one year, 98% of the atoms in your body will have left you.

Human memory

- In one year, 98% of the atoms in your body will have left you.
- Yet your memories will remain with you!

Human memory

- In one year, 98% of the atoms in your body will have left you.
- Yet your memories will remain with you!
- It is the structures and their interactions that make our memories remain.

Human memory

"So what is this mind of ours: what are these atoms with consciousness? Last week's potatoes! They now can remember what was going on in my mind a year ago — a mind which has long ago been replaced"

R. Feynman

Repository memory

Repository memory

- In 10 years, all disks & tapes in your repository will have been replaced.

Repository memory

- In 10 years, all disks & tapes in your repository will have been replaced.
- Yet the archived data remains.

Repository memory

- In 10 years, all disks & tapes in your repository will have been replaced.
- Yet the archived data remains.
- It is the structures and their interactions that make the archived data remain.

Repository memory

- In 10 years, all disks & tapes in your repository will have been replaced.
- Yet the archived data remains.
- It is the structures and their interactions that make the archived data remain.
- The design determines repository longevity (along with media lifetimes).

Repository memory

“I've seen things you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I watched C-beams glitter in the dark near the Tannhauser gate. All those moments will be lost in time, like tears in rain. Time to die.”

P.K. Dick, H. Fancher, D. Peoples

MTTF: Mean time to failure

- The MTTF of a repository is the expected time span between its launch and its first data loss.

MTTF: Mean time to failure

- The MTTF of a repository is the expected time span between its launch and its first data loss.
- MTTF can be estimated before the repository is built.

MTTF: Mean time to failure

- The MTTF of a repository is the expected time span between its launch and its first data loss.
- MTTF can be estimated before the repository is built.
- MTTF depends on design, media lifetimes, bandwidth

MTTF: Mean time to failure

- The MTTF of a repository is the expected time span between its launch and its first data loss.
- MTTF can be estimated before the repository is built.
- MTTF depends on design, media lifetimes, bandwidth, (ratios of manual errors, software errors, natural disasters…)

Estimating MTTF

Estimating MTTF

- Repository design + assumptions => model

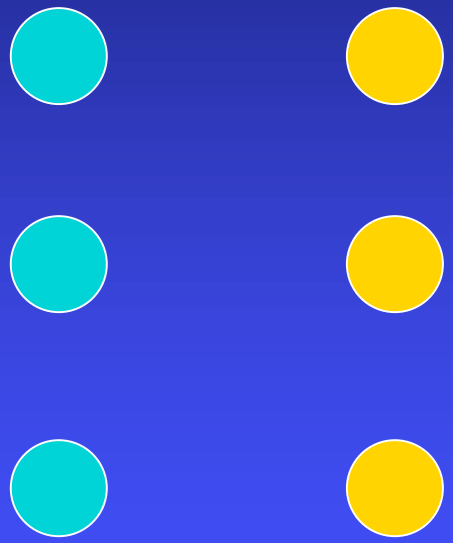
Estimating MTTF

- Repository design + assumptions => model
- Simulate model by computer and compute MTTF.

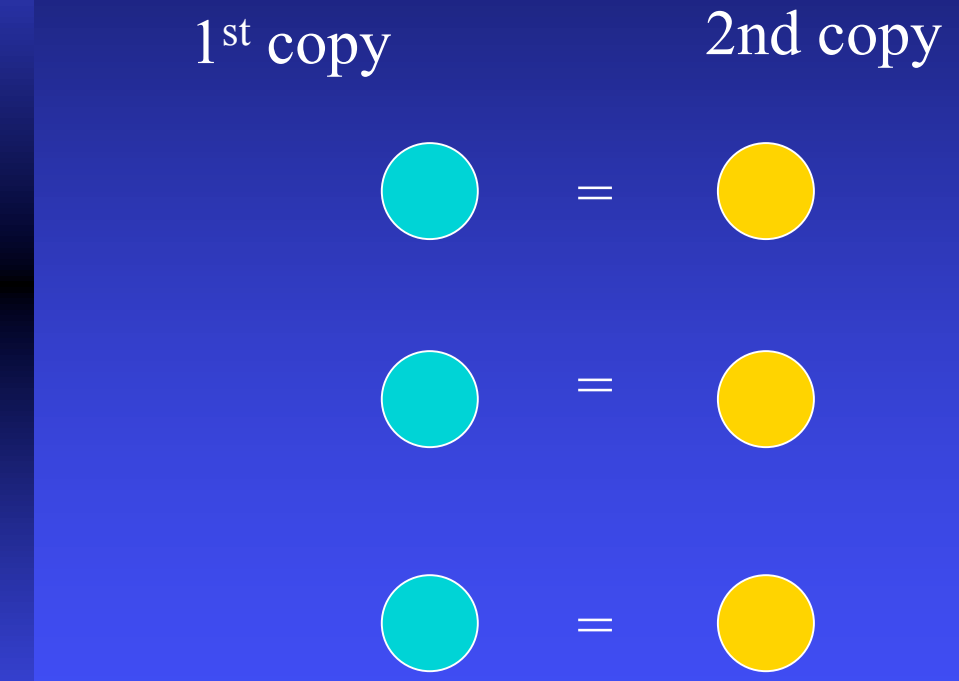
Estimating MTTF

- Repository design + assumptions => model
- Simulate model by computer and compute MTTF.
- Redesign repository or change assumptions and go to step 1.

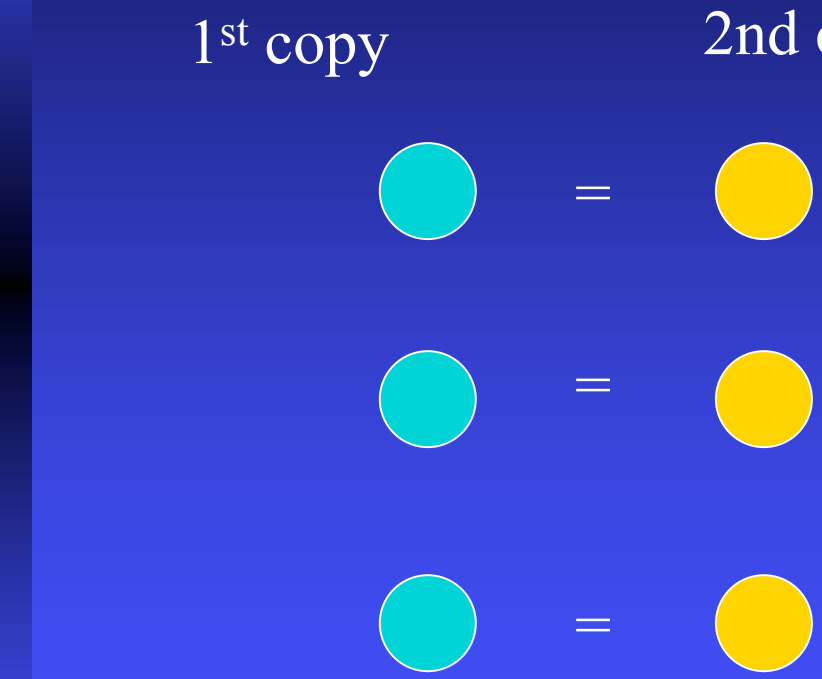
Example simulation



Example simulation



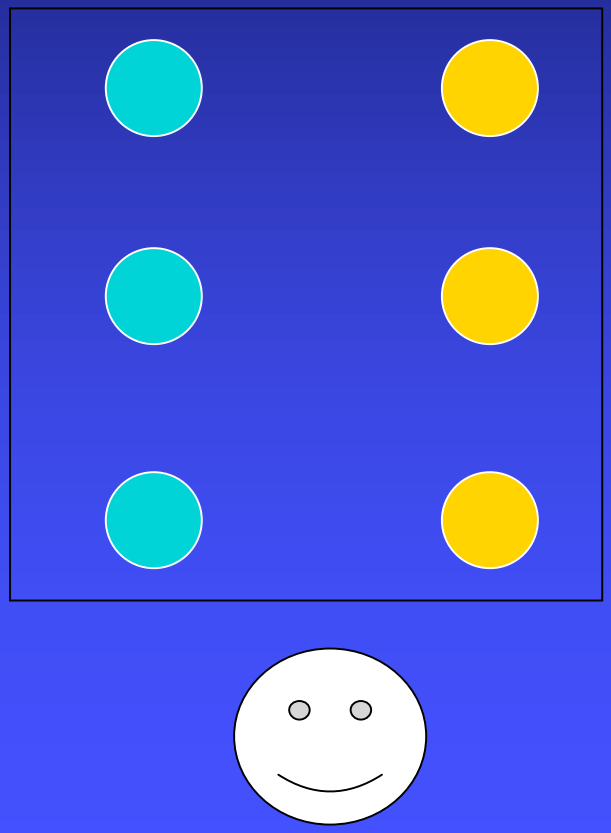
Example simulation



All 6 CDs verfied every Monday

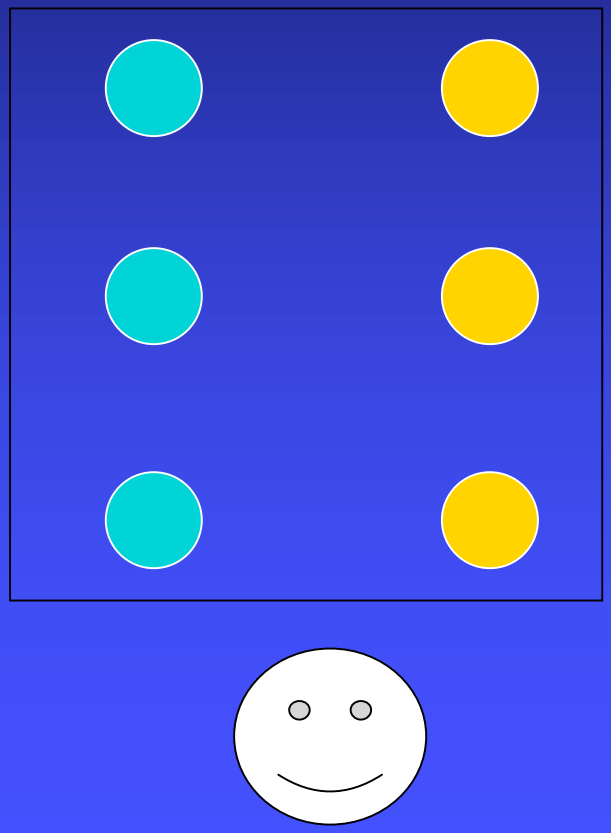
Example simulation

Day 1



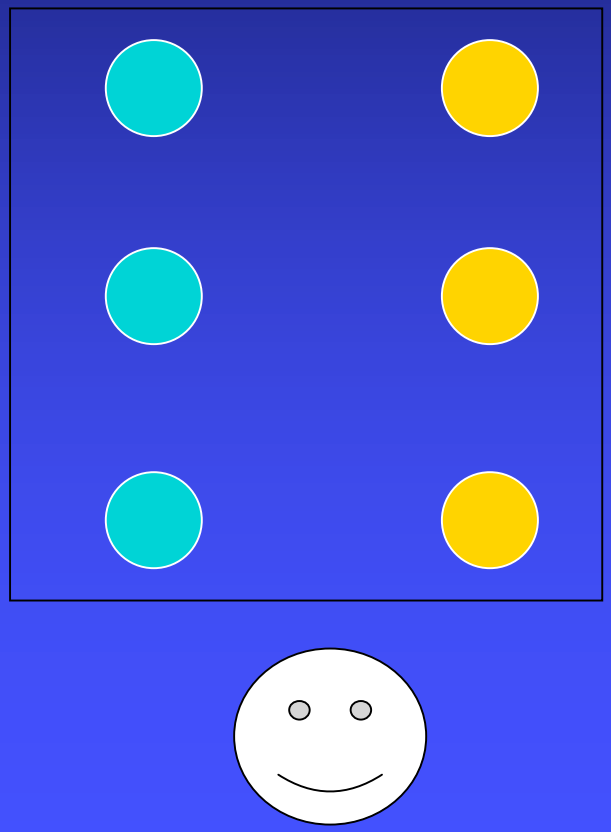
Example simulation

Day 2



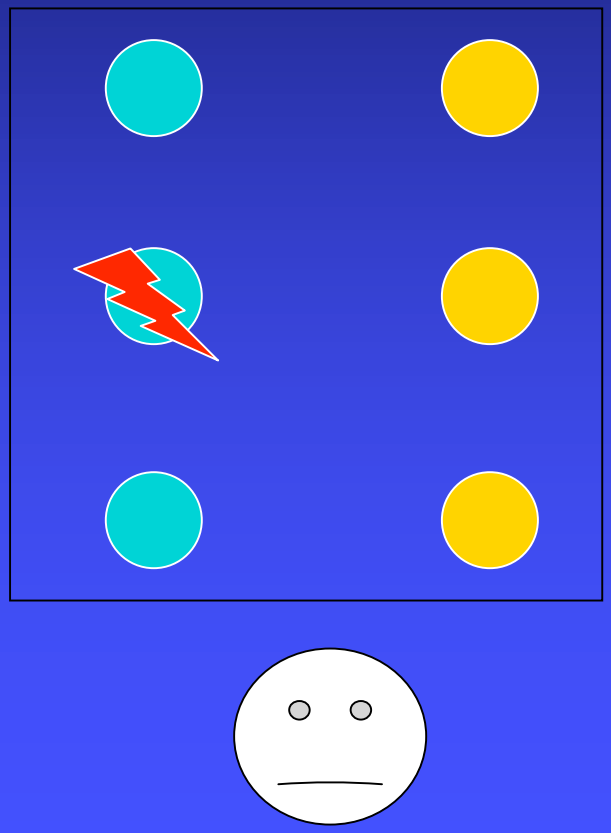
Example simulation

Day 3



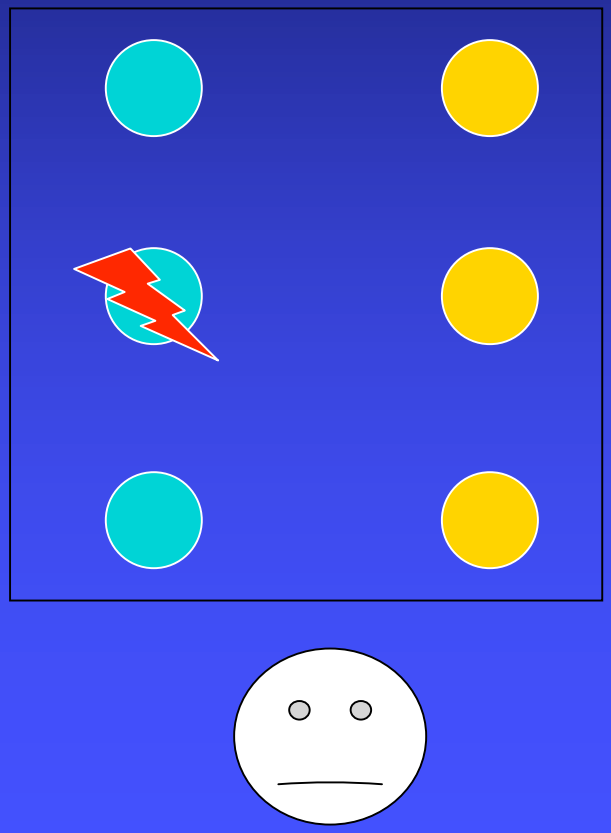
Example simulation

Day 698



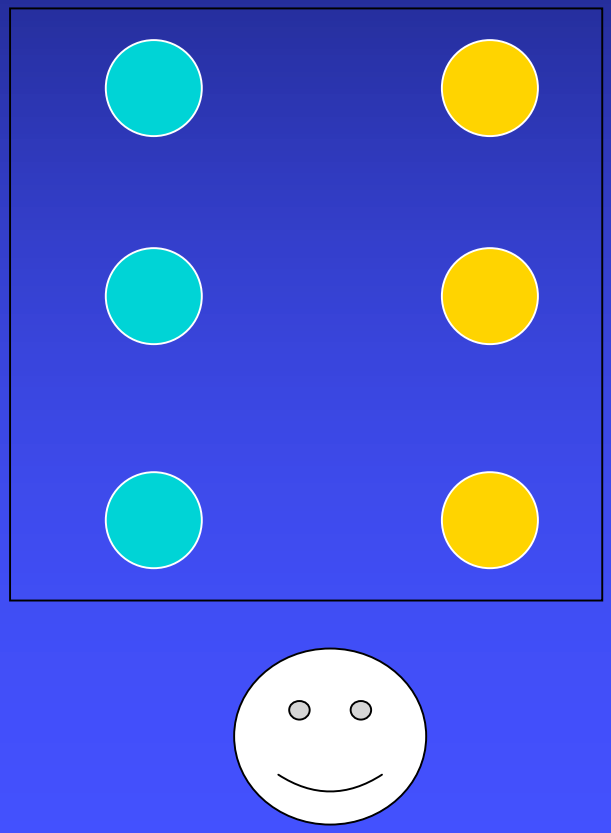
Example simulation

Day 699



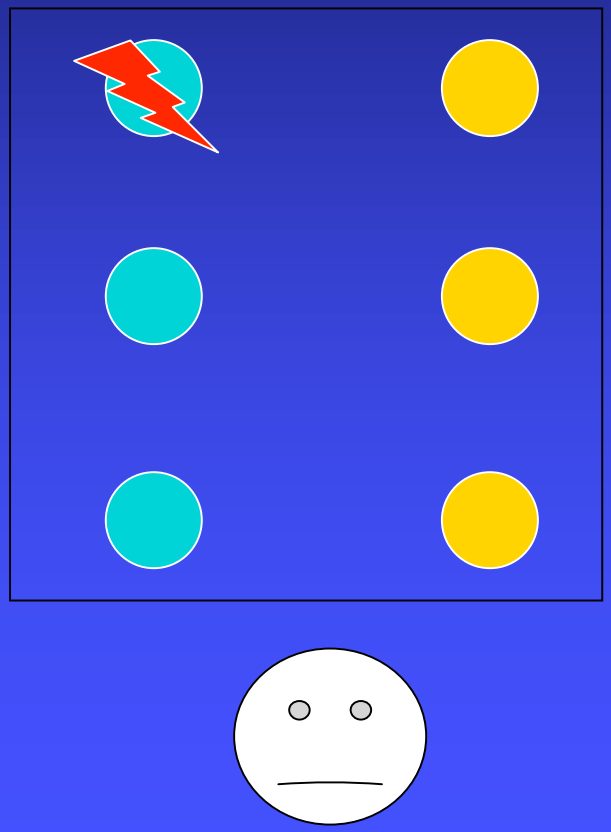
Example simulation

Day 700



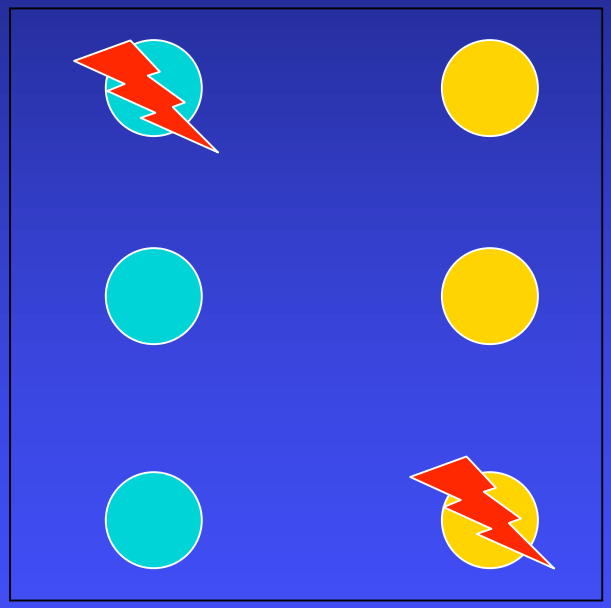
Example simulation

Day 1055



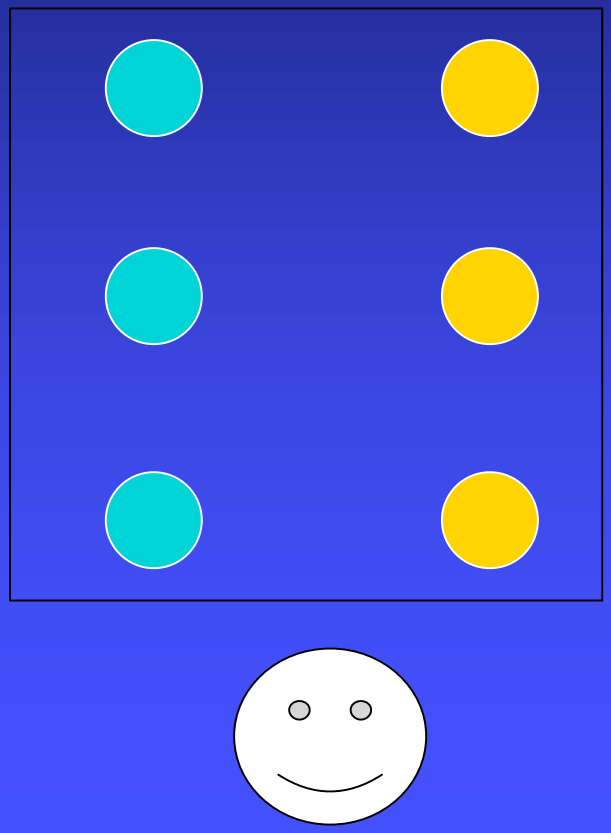
Example simulation

Day 1056



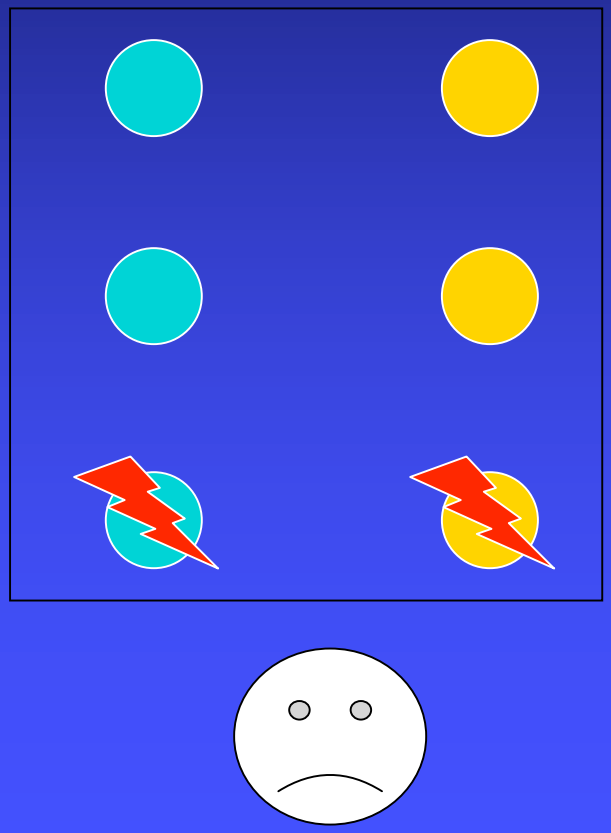
Example simulation

Day 1057



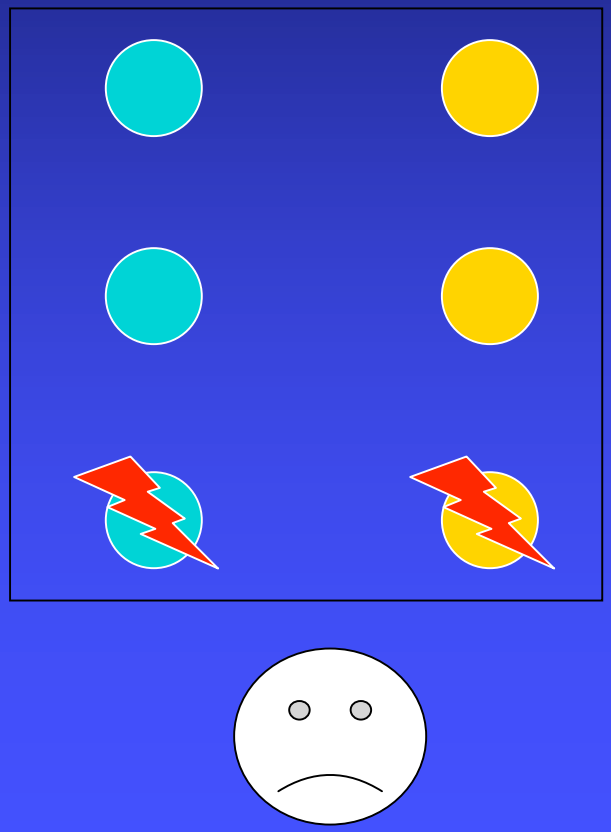
Example simulation

Day 6666



Example simulation

TTF = 6666



Sample results

Sample results

- The MTTF of netarchive.dk's repository is 144 years.

Sample results

- The MTTF of netarchive.dk's repository is 144 years…not counting software errors, operator errors, natural disasters etc.

Sample results

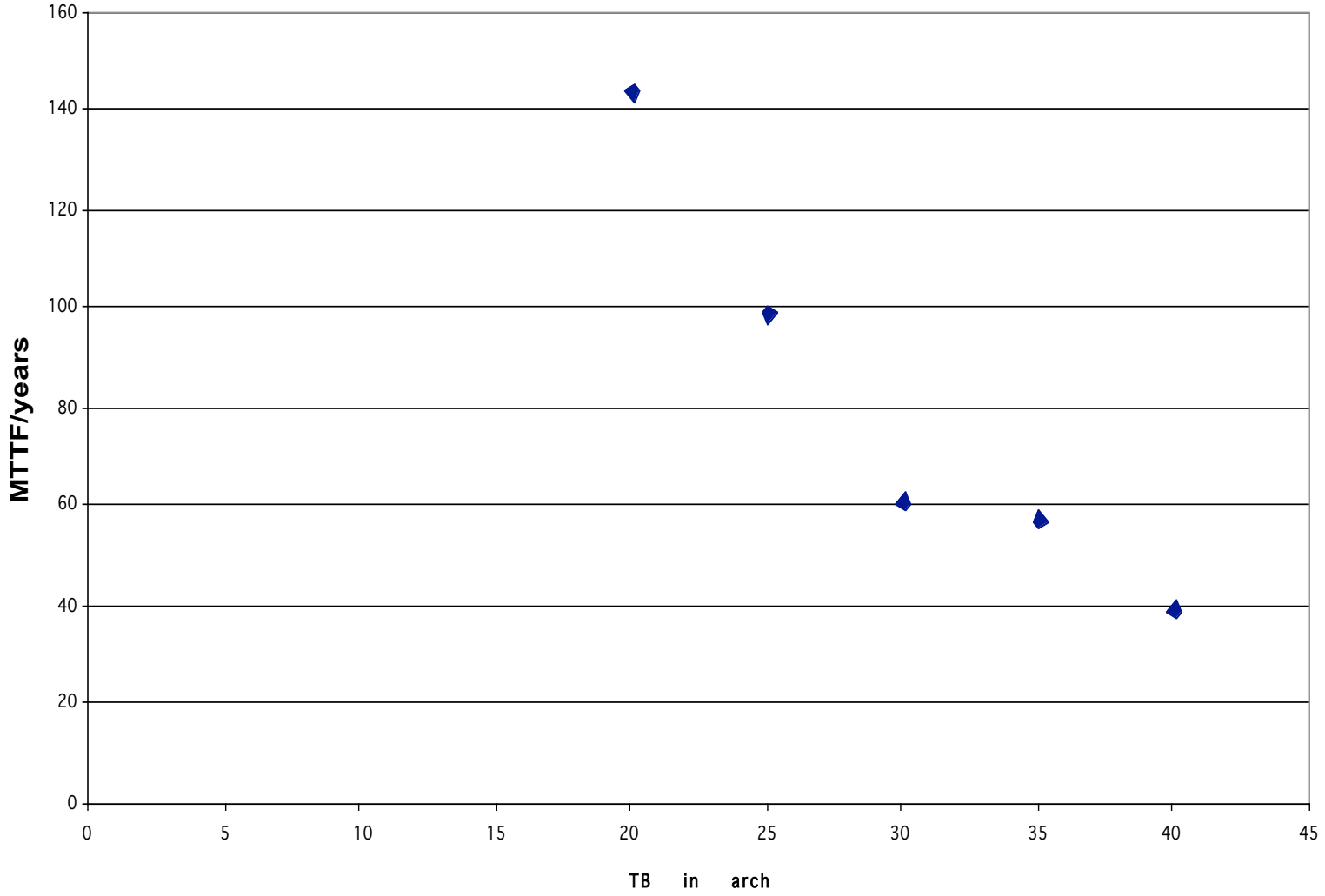
- The MTTF of netarchive.dk's repository is 144 years…not counting software errors, operator errors, natural disasters etc.
- What use is this information?

Sample results

- The MTTF of netarchive.dk's repository is 144 years…not counting software errors, operator errors, natural disasters etc.
- What use is this information? Focus on other risks!

Sample results

- The MTTF of netarchive.dk's repository is 144 years…not counting software errors, operator errors, natural disasters etc.
- What use is this information? Focus on other risks!
- Scalability: doubling the amount of ingested data reduces MTTF to 40 years.


160
140
120
100
80
60
40
20
0
MTTF/years
0
5
10
15
20
25
30
35
40
45
TB in arch

Sample results: static vs explosive

- The simulation example was static (no ingest after day 1).

Sample results: static vs explosive

- The simulation example was static (no ingest after day 1).
- W.A.s face exponential growth in ingest.

Sample results: static vs explosive

- The simulation example was static (no ingest after day 1).
- W.A.s face exponential growth in ingest.
- MTTF(explosive) ~ MTTF(static)

Sample results: static vs explosive

- The simulation example was static (no ingest after day 1).
- W.A.s face exponential growth in ingest.
- MTTF(explosive) ~ MTTF(static)
- - if storage capacities increase at same rate

Sample results: static vs explosive

- The simulation example was static (no ingest after day 1).
- W.A.s face exponential growth in ingest.
- MTTF(explosive) ~ MTTF(static)
- - if storage capacities increase at same rate
- - and bandwidth does too.

Sample results: static vs explosive

- The simulation example was static (no ingest after day 1).
- W.A.s face exponential growth in ingest.
- MTTF(explosive) ~ MTTF(static)
- - if storage capacities increase at same rate
- - and bandwidth does too.
- Failure to keep up => decreasing MTTF

Summary

- Digital repositories do not last forever.
- Their longevity depends on their structure.
- The MTTF of a repository can be estimated through simulation.
- Results can point out the risks that have/have not been addressed appropriately.

Conclusion

- Compared to the total effort of designing and implementing a repository, simulation is fast & cheap.
- Do it!
- (I'd like to see your designs/models).

Future work

- Make it simpler to define models.
- Make the model more comprehensive, removing assumptions.
- Estimate other figures, e.g. costs.

Questions?