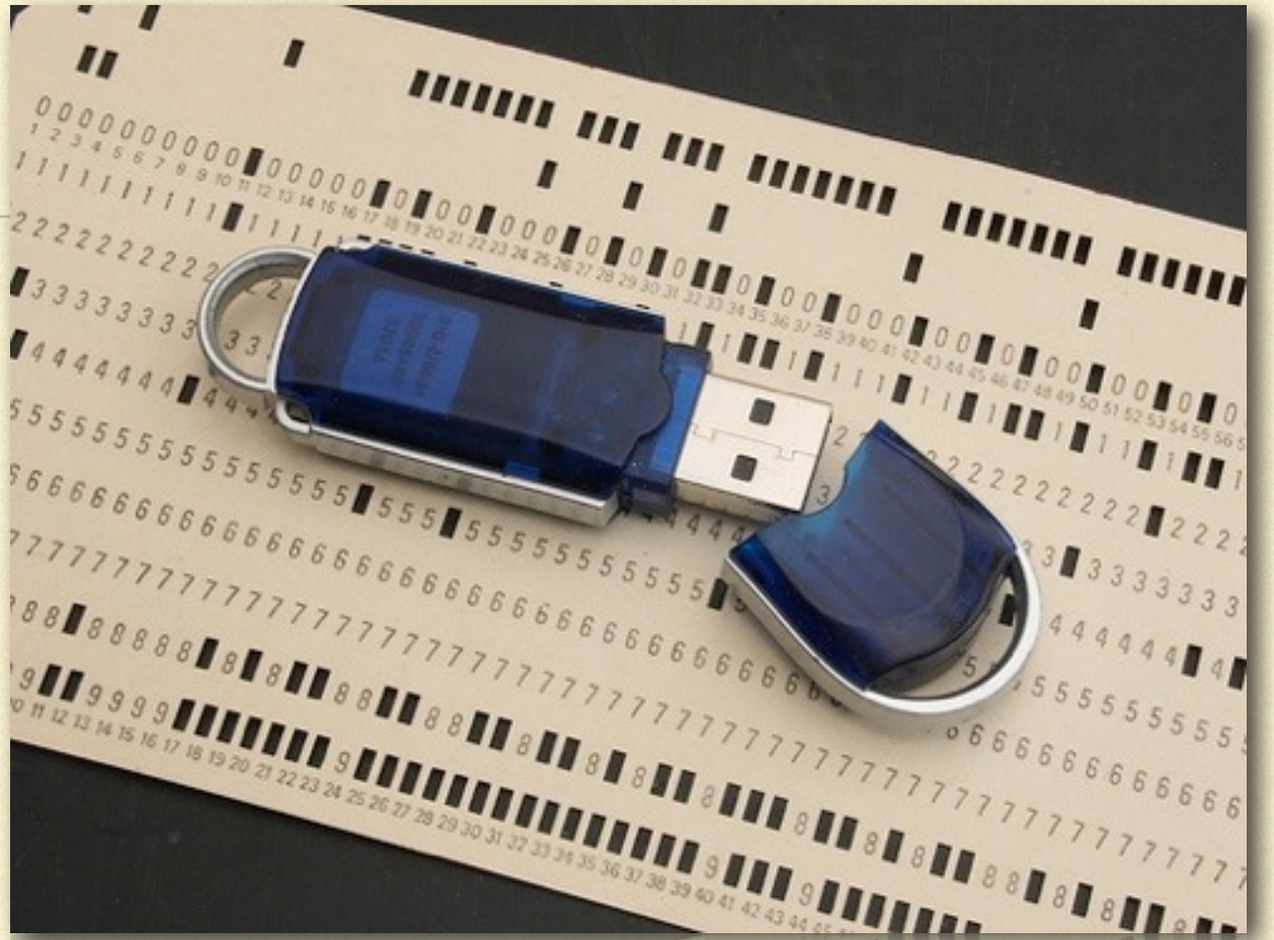
DON’T PANIC!

The Digital Curator’s Guide to the Data Format Methodology



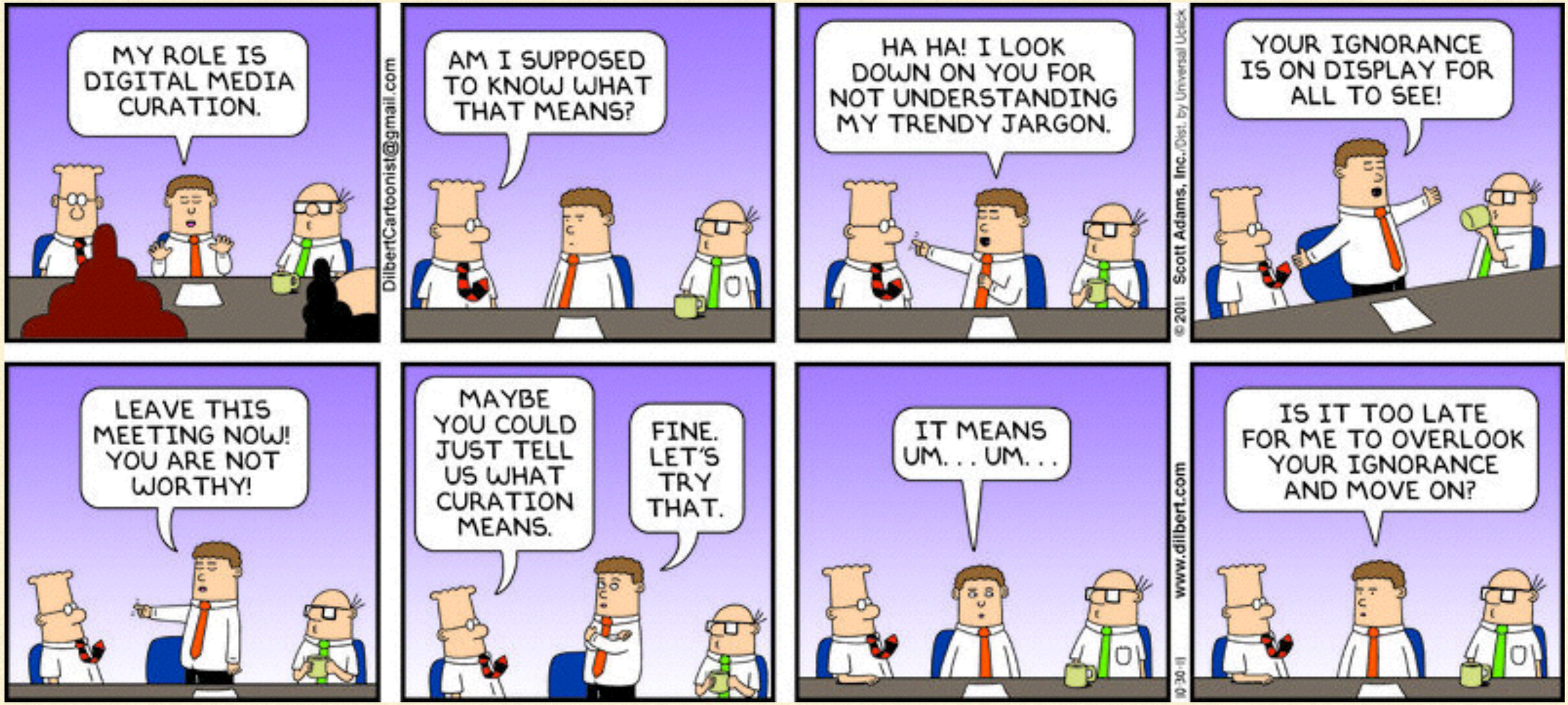
Presented by

Isaiah Beard

Digital Data Curator
Rutgers University Libraries

TOPICS COVERED TODAY:

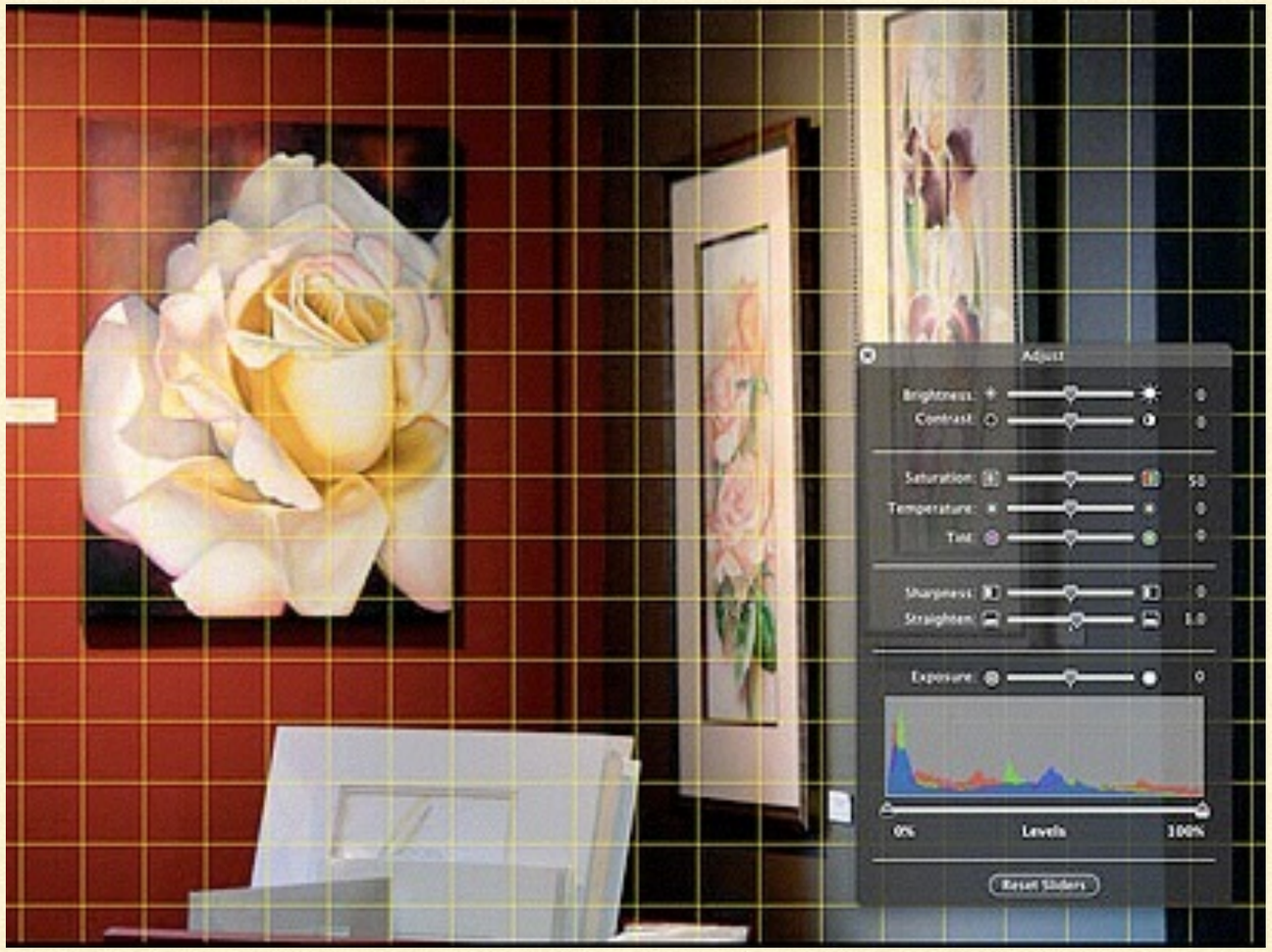
- **Definitions, Concepts**
  - What is Data Curation?
  - Why is it important?
- **The Data Curation Lifecycle**
  - Practical approaches to assessing and handling data
- **Applying these concepts to research data**



WHAT IS DIGITAL CURATION?

*Source: Dilbert, October 30, 2011

"The curation, preservation, maintenance, collection and archiving of digital assets."*



WHAT IS DIGITAL CURATION?

*Source: "What is Digital Curation?" Digital Curation Centre, http://www.dcc.ac.uk/about/what/

WHY SHOULD WE DO THIS?

- Because data is <u>everywhere</u>.

2,500,000,000,000,000,000

2.5 **<u>Quintillion</u>** bytes of data are generated <u>every day</u>.

Conner, M (2012) "Data on Big Data." http://marciaconner.com/blog/data-on-big-data/

WHY SHOULD WE DO THIS?

- Because data is <u>everywhere</u>.

2.5 Petabytes

(2,500 1TB hard drives)

The amount of data currently flowing through Walmart's transaction database.

Cukier, K. (2012) "Data, Data Everywhere." *The Economist*, Feb 27, 2010. http://www.economist.com/node/15557443?story_id=15557443

WHY SHOULD WE DO THIS?

- Because data is <u>everywhere</u>.

10 Petabytes

(5,000 2TB hard drives)

The amount of diagnostic data generated by a jet engine under full engineering monitoring in **<u>1 hour.</u>**

5,000 data samples are taken **<u>per second</u>**.

Source: General Electric (2013) "Is Your Jet Engine Linked In?" http://www.linkedin.com/today/post/article/20130619125316-230929989-is-your-jet-engine-linked-in-connecting-minds-and-machines-to-drive-efficiency

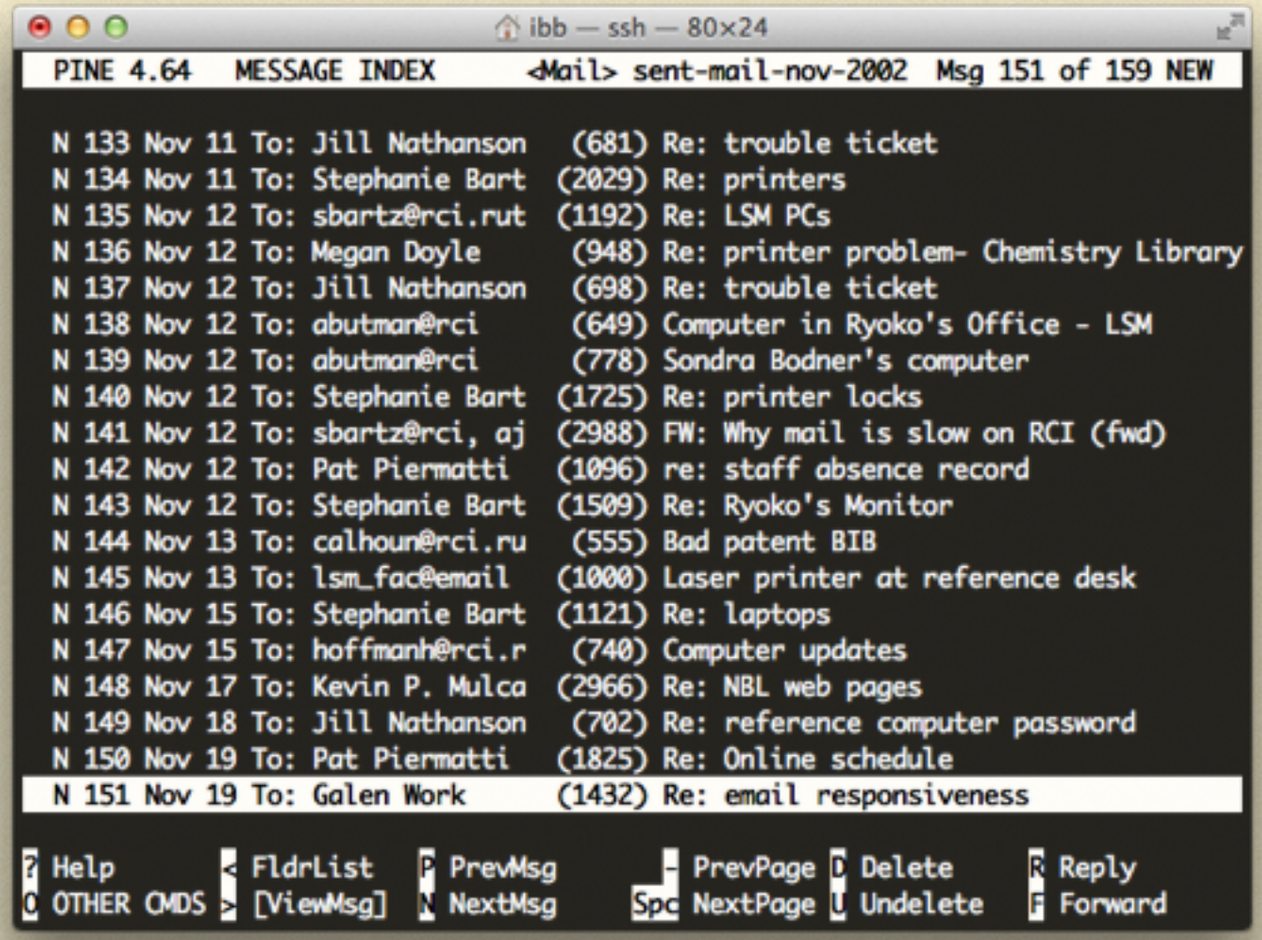
WHY SHOULD WE DO THIS?

- Because data is <u>everywhere</u>.

E-mail

Early 2000s and prior:

- 100MB Quota
- Mostly text-based
- Not very mobile
- Not instant



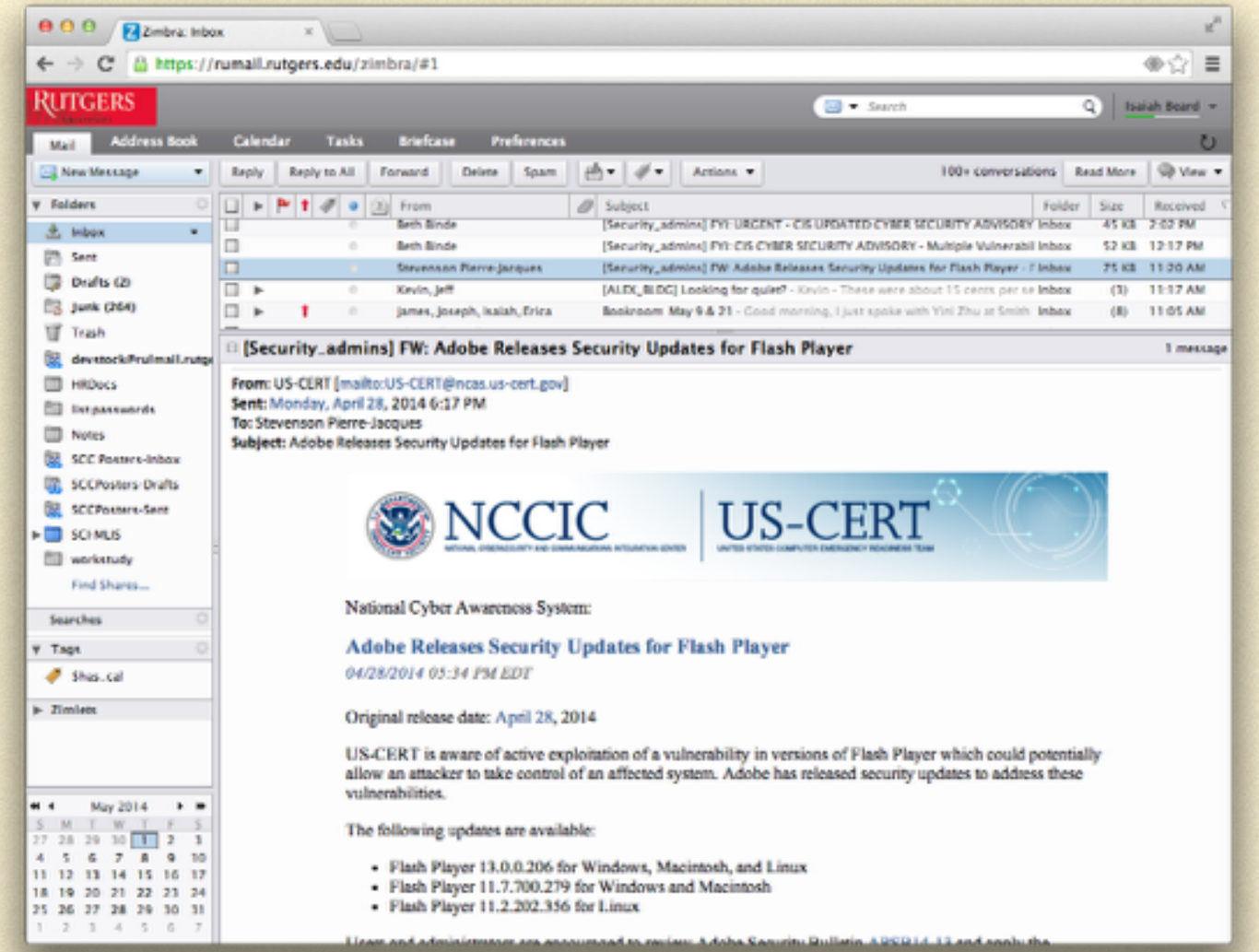
WHY SHOULD WE DO THIS?

- Because data is <u>everywhere</u>.

E-mail

Today:

- GBs of Quota
- Graphical
- Mobile
- Instant
- Web based
- Comes in multiple “faces”



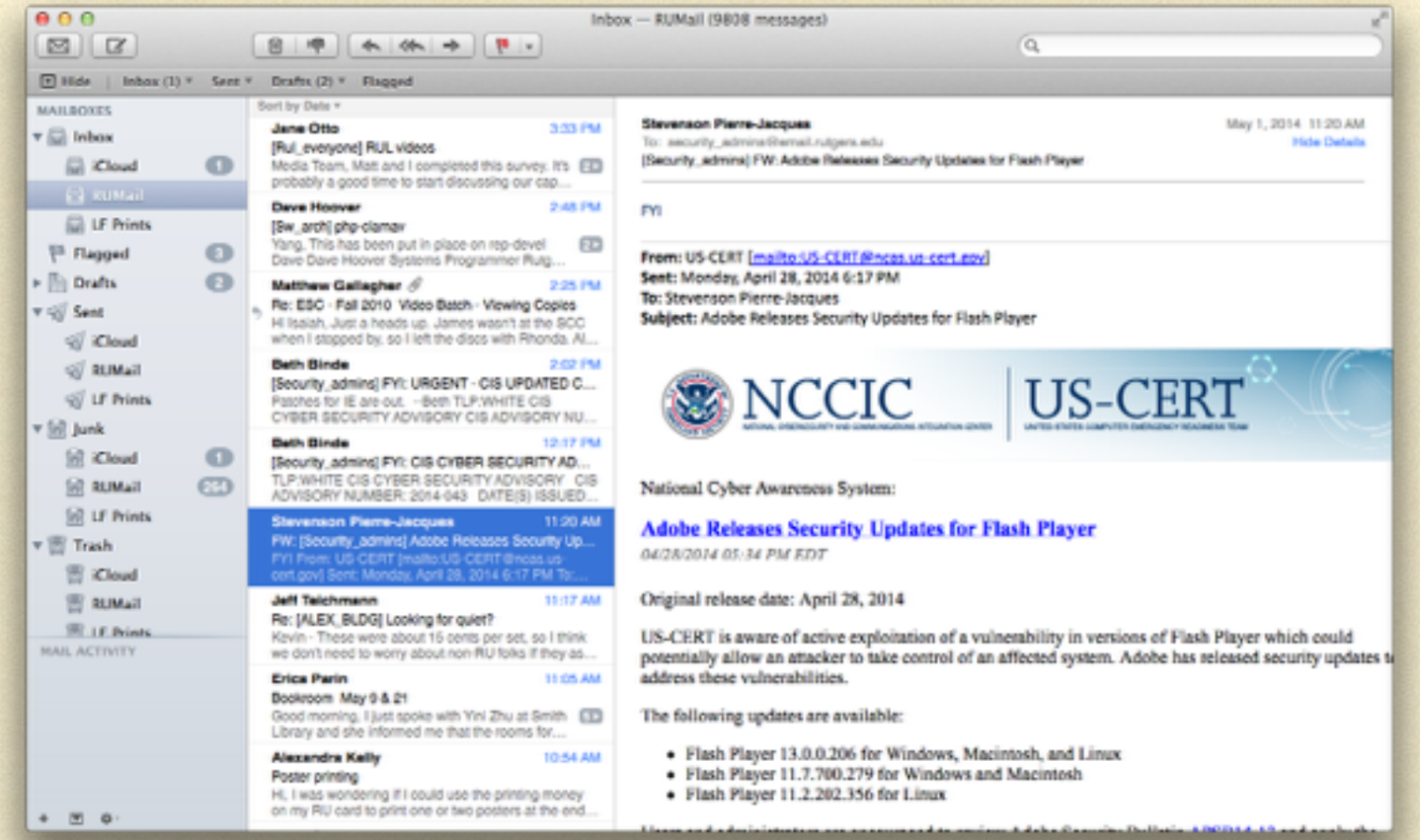
WHY SHOULD WE DO THIS?

- Because data is <u>everywhere</u>.

E-mail

Today:

- GBs of Quota
- Graphical
- Mobile
- Instant
- Web based
- Comes in multiple “faces”



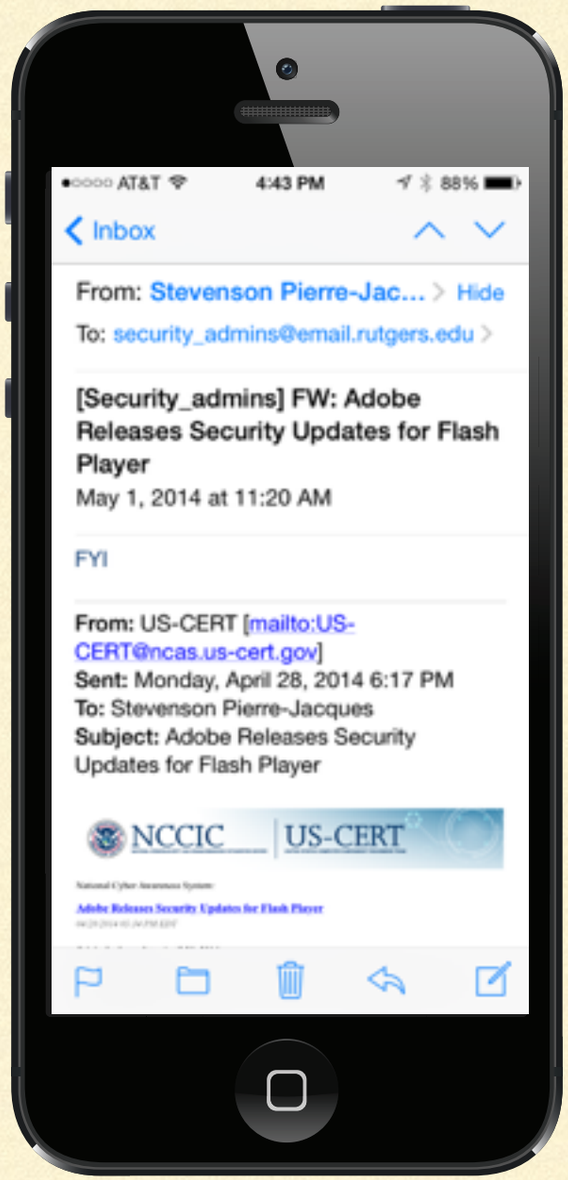
WHY SHOULD WE DO THIS?

- Because data is <u>everywhere</u>.

E-mail

Today:

- GBs of Quota
- Graphical
- Mobile
- Instant
- Web based
- Comes in multiple “faces”



WHY SHOULD WE DO THIS?

- Because data is <u>everywhere</u>.

E-mail

3.37 Billion e-mail accounts worldwide

2.25 Billion
Consumer

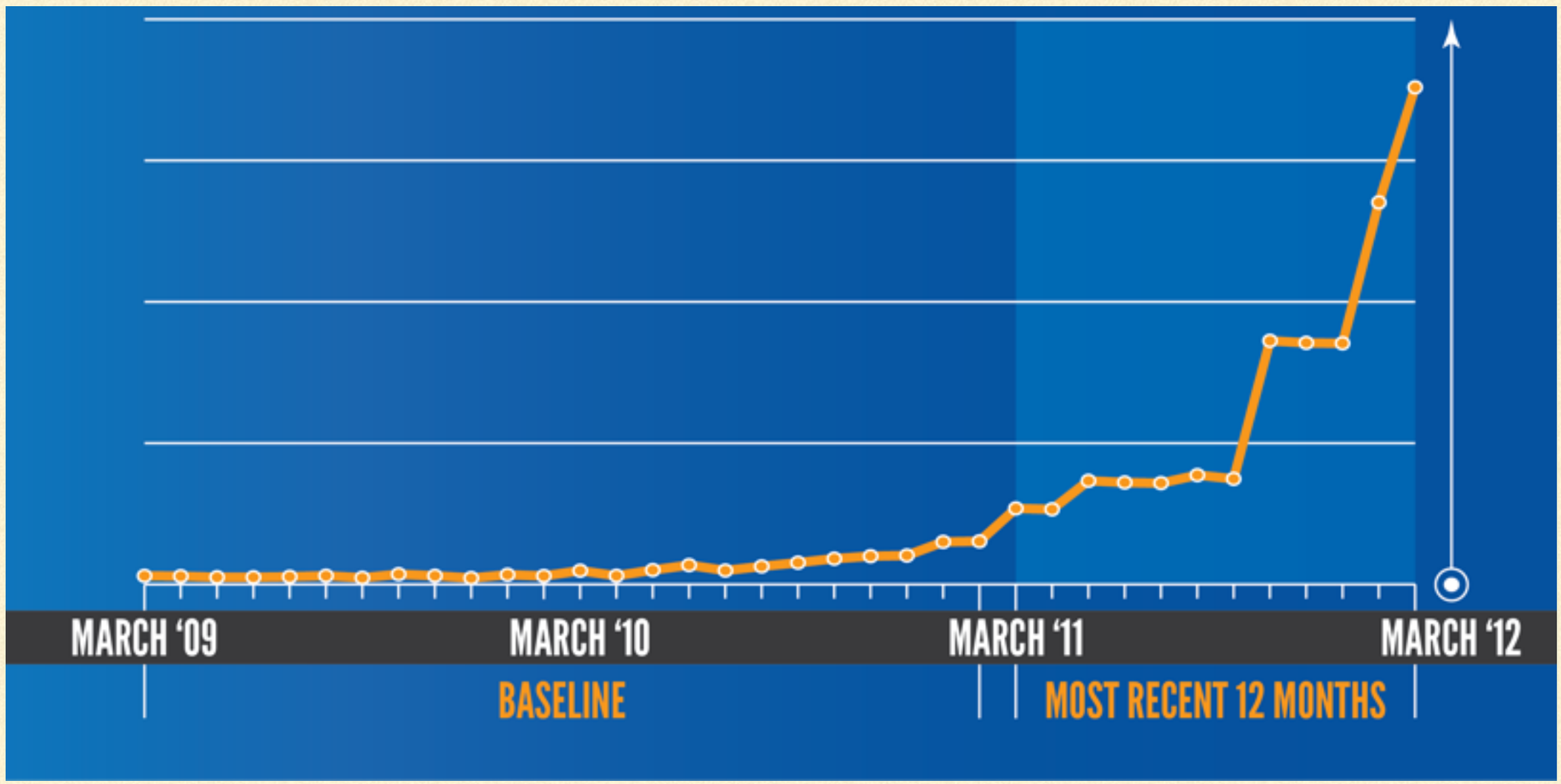
850 Million
Corporate

Radicati Group (2011). "Email Statistics Report (2012-2016)." http://www.radicati.com/wp/wp-content/uploads/2012/04/Email-Statistics-Report-2012-2016-Executive-Summary.pdf

WHY SHOULD WE DO THIS?

- Because data is everywhere.

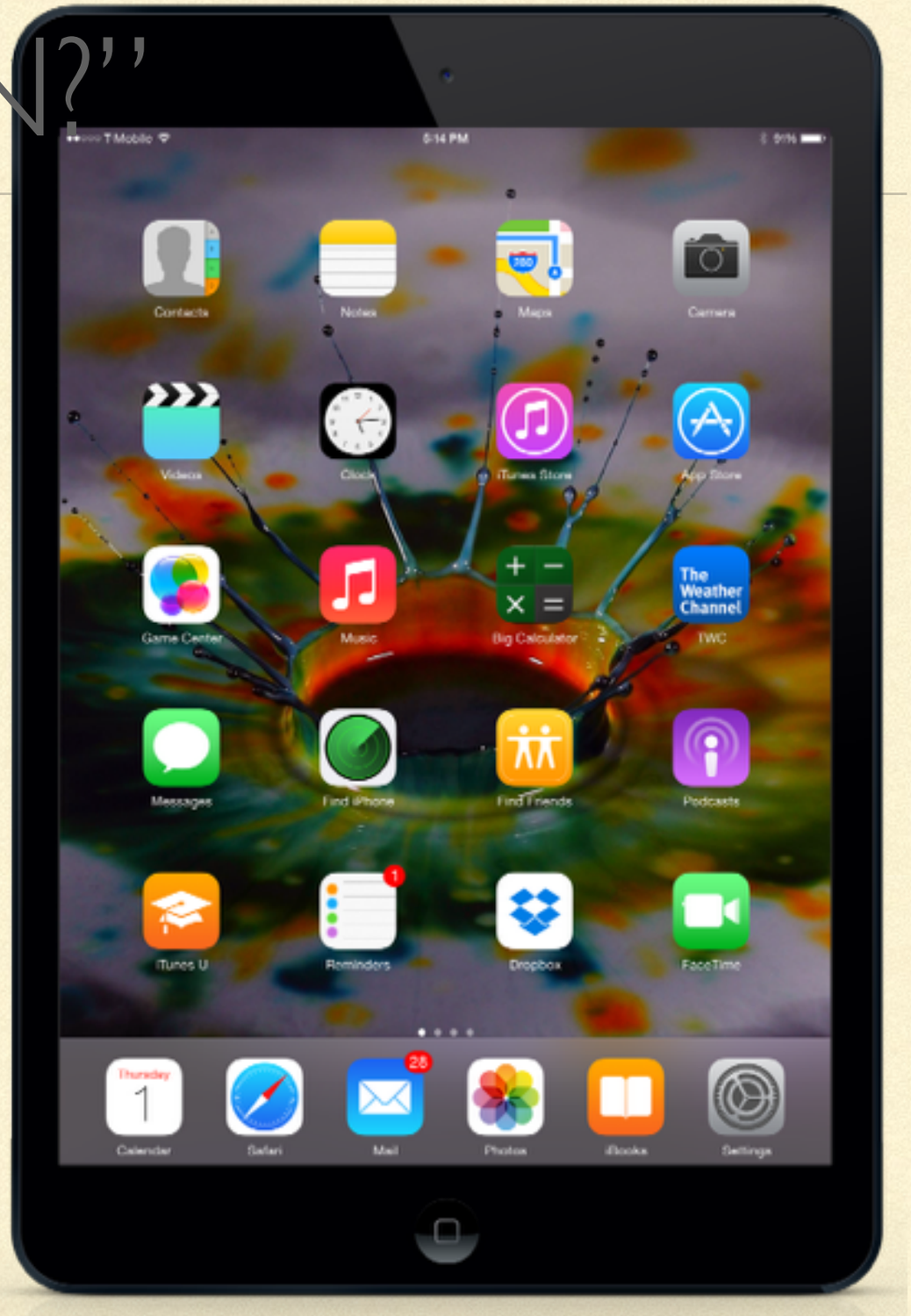
92% of the world's data (by volume) created in the last two years.



IBM (2013, December). "Applying new analytics tools to reveal new opportunities." http://www.ibm.com/smarterplanet/us/en/business_analytics/article/it_business_intelligence.html

WHAT'S CAUSING THIS "DATA EXPLOSION?"

Mobile Devices

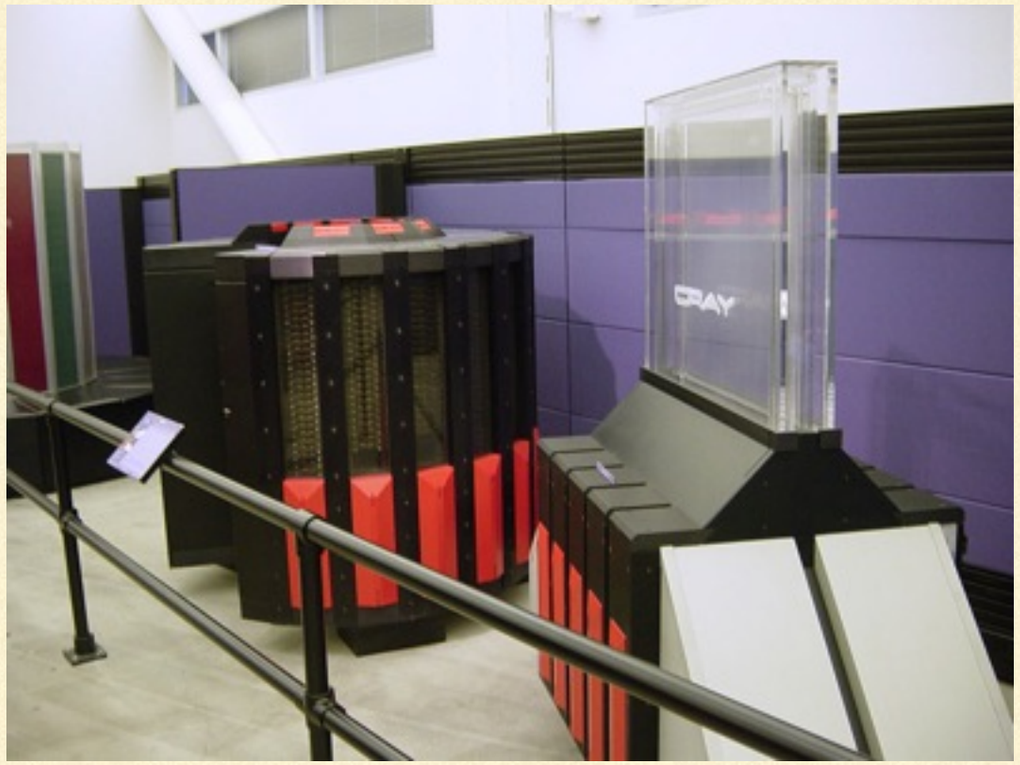


WHAT'S CAUSING THIS "DATA EXPLOSION?"

1984: The Cray XMP Supercomputer

- THE most powerful computer in existence at the time
- 150,000 - 200,00 watts
- 128MB of RAM Max
- Up to 32GB of storage
- 112 square feet of floor space
- Cost: $15 million (disks not included)
- Only a few dozen made

- CPU Processing Power: 4 CPU cores, 400 MFLOPS



WHAT'S CAUSING THIS "DATA EXPLOSION?"

Mobile Devices and Ubiquitous Computing

2007-Present: Smartphones, Tablets

- NOT the most powerful computers in existence for their time
- 5.45 watts **max**
- 1GB of RAM, some higher
- 16 - 64GB of storage, some are expandable
- Fits in your pocket, 1.75 Billion in existence
- Cost: $199-$899, depending on specs (Solid state storage, cloud storage included)

- CPU Processing Power: 4 CPU cores, 984.73 MFLOPS (iPhone 5s example, others may vary)

WHAT'S CAUSING THIS "DATA EXPLOSION?"

Mobile Data Gathering and Processing

- GPS Positioning System
- Compass
  - Magnetometer
- Accelerometer
  - Seismograph
  - Impact Sensor
  - Motion Sensor
  - Gyroscope/Orientation Sensor
- Radios
  - Data/Telemetry
  - Doppler Shift Measuring

WHAT'S CAUSING THIS "DATA EXPLOSION?"

Mobile Data Gathering and Processing

- Microphone
  - Recorder
  - Sound dB meter
  - Pitch meter
  - Frequency counter

WHAT'S CAUSING THIS "DATA EXPLOSION?"

Mobile Data Gathering and Processing

- Camera
  - Recorder
  - Color Densitometer
  - Motion/Pattern detector
  - Barcode Reader

WHAT'S CAUSING THIS "DATA EXPLOSION?"

Mobile Data Gathering and Processing

- Camera
  - Recorder
  - Color Densitometer
  - Motion/Pattern detector
  - Barcode Reader
  - Vital Sign/Health Monitor

A smartphone camera and flash can be used to accurately measure your heart rate and blood oxygen saturation.

UBIQUITOUS DATA

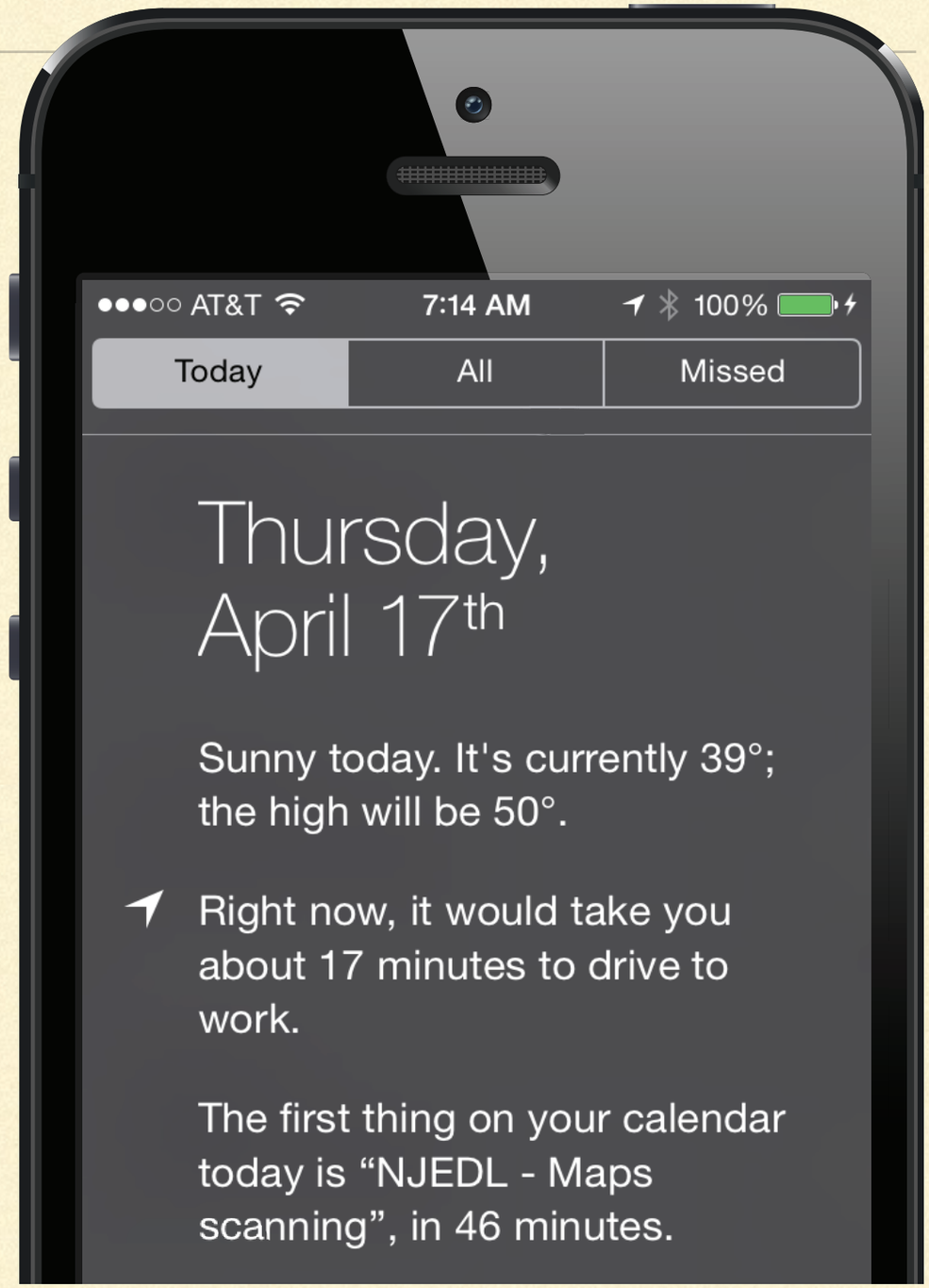
The Smartphone Dilemna:
Data that's maybe a little too smart…

- First Smartphone: Palm Treo 650 (2004)
- Before This: Palm Pilot PDAs
- Also used: Blackberry, Windows Mobile
- Been using iPhones since 2007.
- A LOT of historical data recorded over the past decade.

UBIQUITOUS DATA

"Smart" data, today

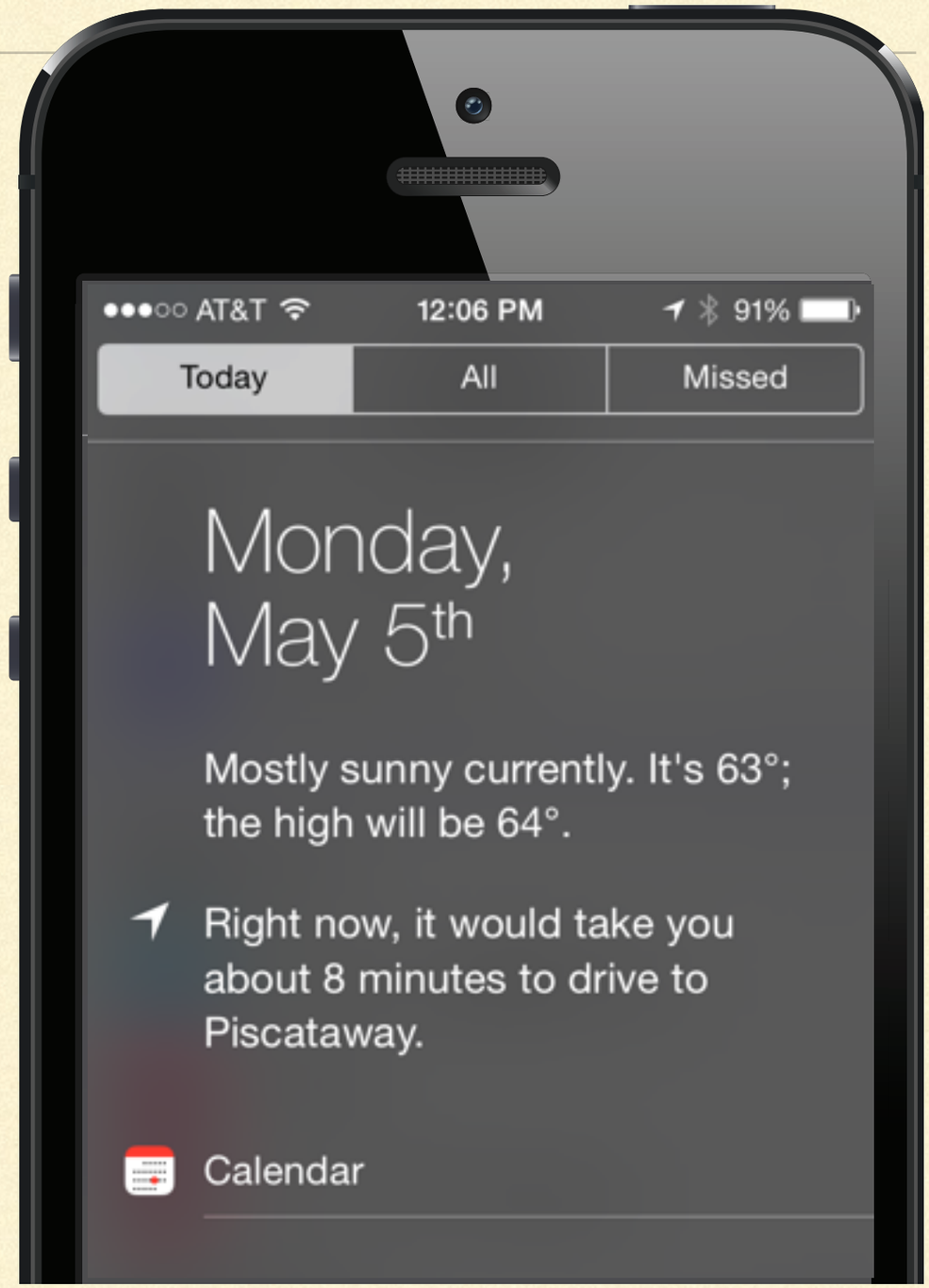
- The Result:
  - My mobile devices are "aware" of my daily routine
    - When and where I go to work
    - When/where I go to lunch
    - When/where I go home
    - Birthdays
    - Anniversaries
    - Meetings



UBIQUITOUS DATA

"Smart" data, today

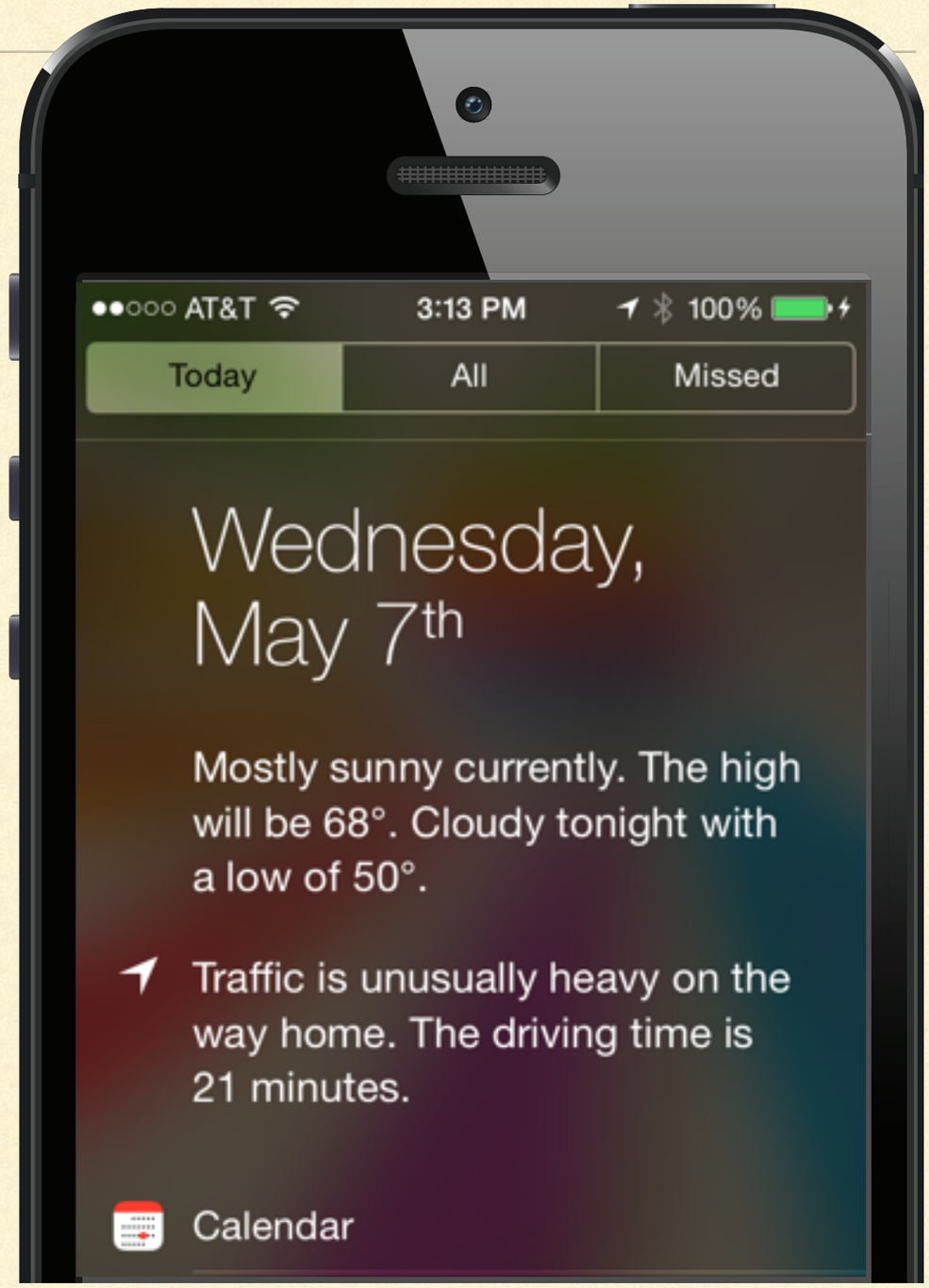
- The Result:
  - My mobile devices are "aware" of my daily routine
    - When and where I go to work
    - When/where I go to lunch
    - When/where I go home
    - Birthdays
    - Anniversaries
    - Meetings



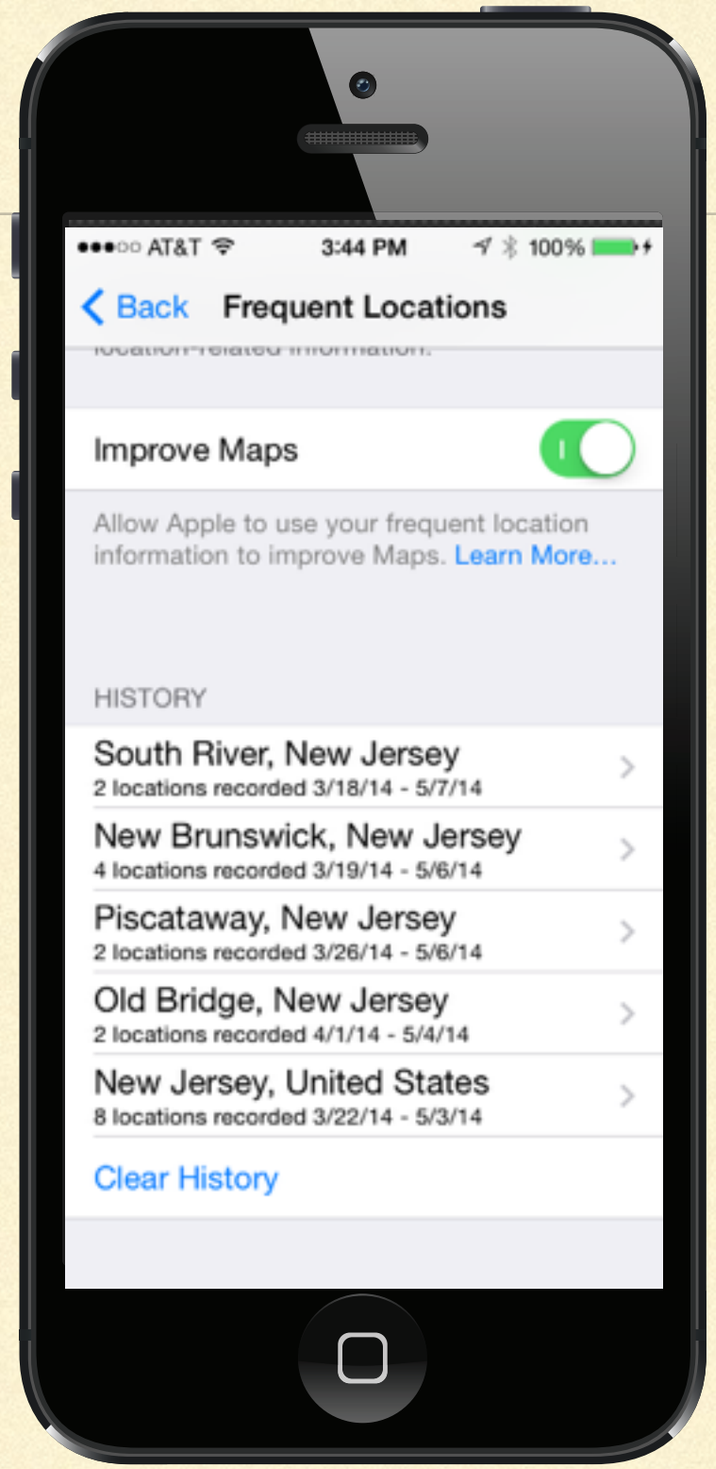
UBIQUITOUS DATA

"Smart" data, today

- The Result:
  - My mobile devices are "aware" of my daily routine
  - When and where I go to work
  - When/where I go to lunch
  - When/where I go home
  - Birthdays
  - Anniversaries
  - Meetings



UBIQUITOUS DATA



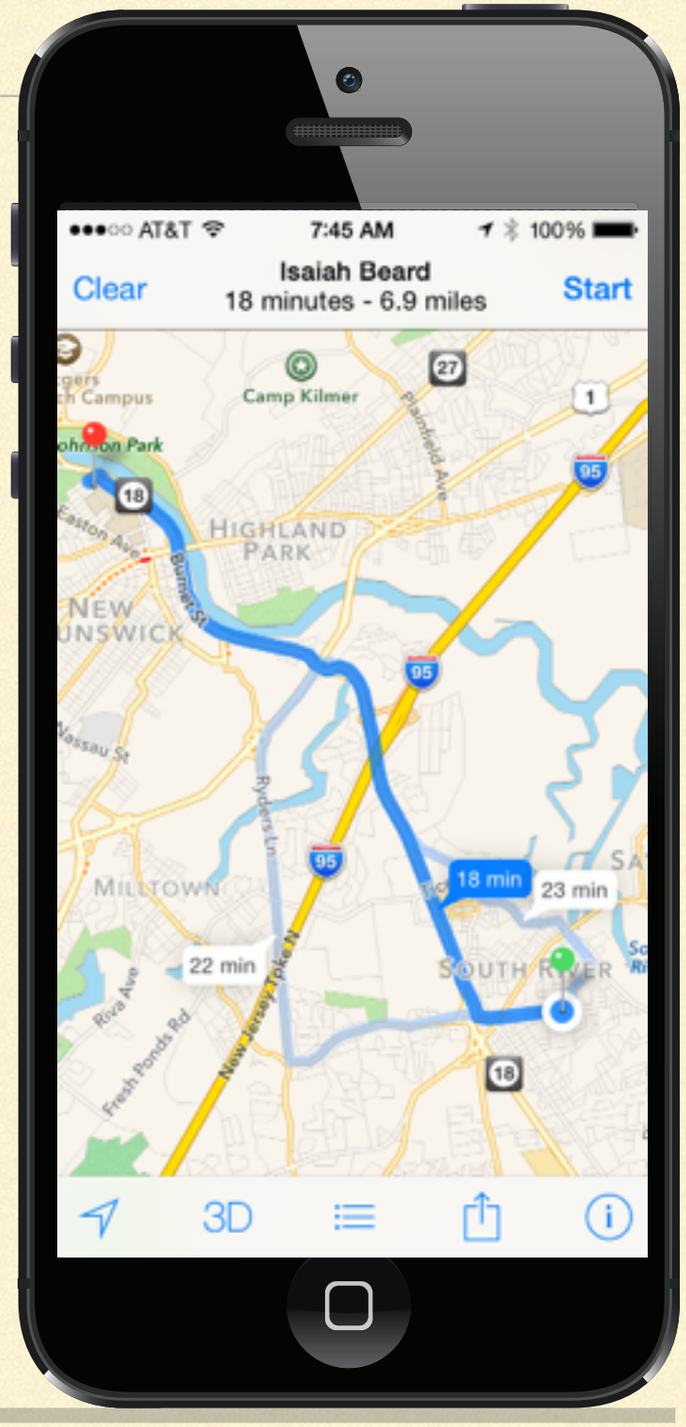
"Smart" data, today

- The Result:
  - My mobile devices are "aware" of my daily routine
    - When and where I go to work
    - When/where I go to lunch
    - When/where I go home
    - Birthdays
    - Anniversaries
    - Meetings

UBIQUITOUS DATA

Traffic data:

- Your smartphone contains a GPS sensor and a motion sensor
- When driving, your location and speed are measured and collected along with other smartphone users that are deemed to be inside moving vehicles on public roads.
- Data is anonymized* and aggregated, and from this, traffic data is obtained.
- Apple, Google, Garmin, OnStar, others.



*This is what we're told, anyway.

IMPLICATIONS

- All of this data collection means there are enormous challenges ahead. For data to be useful, it must be
  - **Sifted**
    - Determine which data is useful; which data is not
    - Determine which data is private; which data is not
    - Long term implications
      - How do we know that we won't in the future, want data we discarded today?

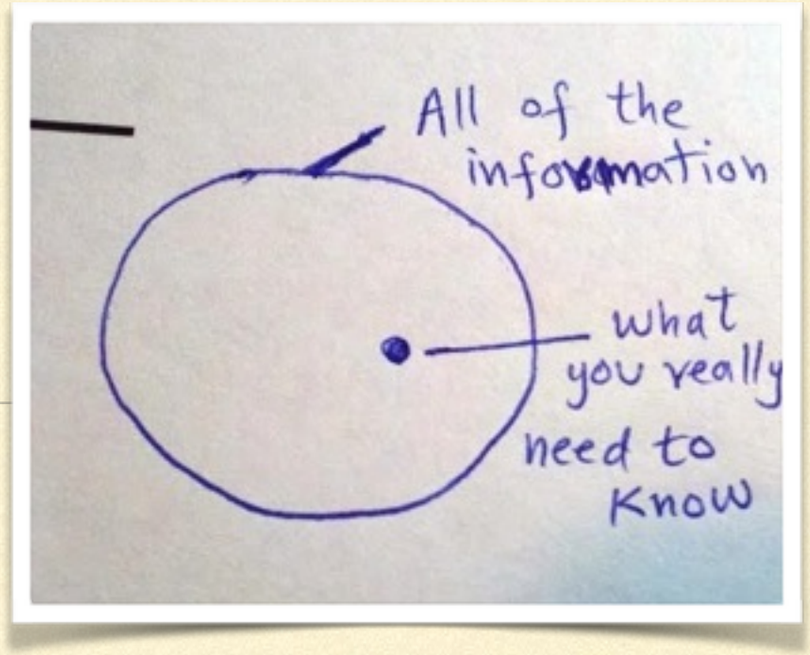
IMPLICATIONS

- All of this data collection means there are enormous challenges ahead. For data to be useful, it must be
  - **Sifted**
  - **Processed**
  - The end result of data collection is that one or more humans, <u>somewhere</u>, will have their question(s) answered.
  - This requires
    - Processing collected data
    - Interpretation of the processing output
    - Visualization / Presentation of the results

IMPLICATIONS

- All of this data collection means there are enormous challenges ahead. For data to be useful, it must be
  - **Sifted**
  - **Processed**
  - **Stored and Preserved**
    - Data has historical value. Other researchers must analyze and validate the data to see if the results can be duplicated.
    - Future researchers may repurpose the data down the road; ask questions the original researcher may not have considered
    - Sensitive/Private data may have access restrictions that must be accommodated.

DON'T PANIC!



- Data should <u>not be feared</u>.
- Yes, it's massive, and yes, it's growing. Yes, the implications are daunting.
- However, the same massive computing power that collects and generates this data can also be harnessed to make it relevant, useful, <u>protected</u>, and durable.
- This gives us a fighting chance to meet the challenge of sifting, processing and preservation.

DIGITAL DATA CURATION IS…

- Understanding what data consists of

- **Digital Data**
  - "That which is collected, observed, or created in a digital form, for purposes of analyzing to produce original research results."*
  - Any related, unique information captured as part of the research process.

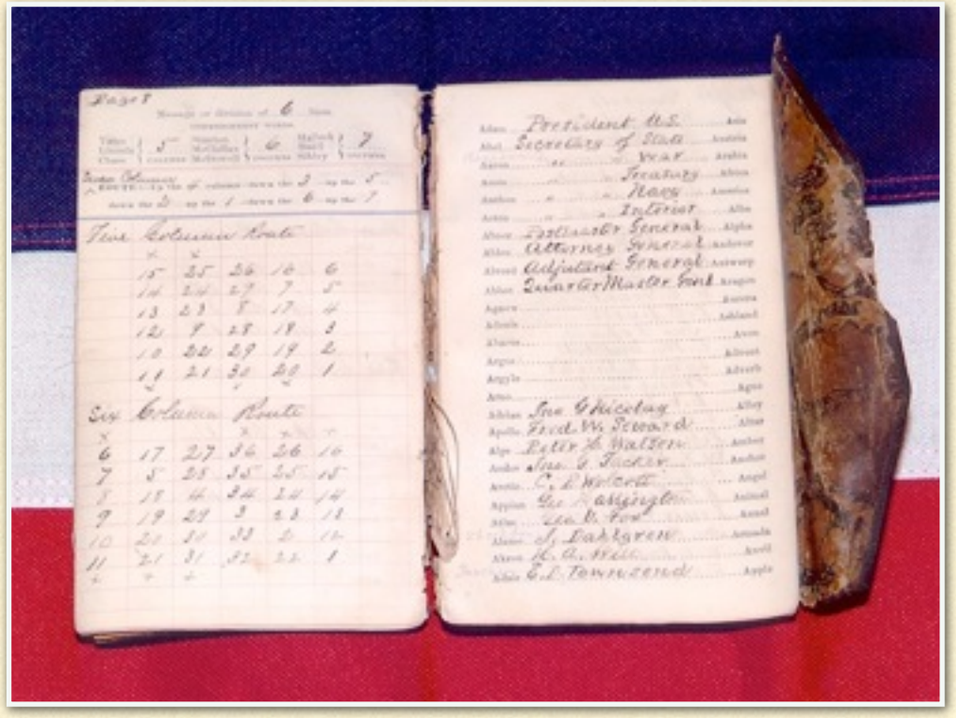
- **Dataset**
  - "A set of files containing both research data - alphanumeric or encoded - and documentation sufficient to make the data re-usable."*

* Source: Information Services, the University of Edinburgh (2012)

DIGITAL DATA CURATION IS…

- Understanding what data consists of

- **Documentation**
  - Any associated assets which explain the research data's
    - production,
    - provenance,
    - processing
    - or interpretation



Such as a codebook, technical or methodology report, or user guide.

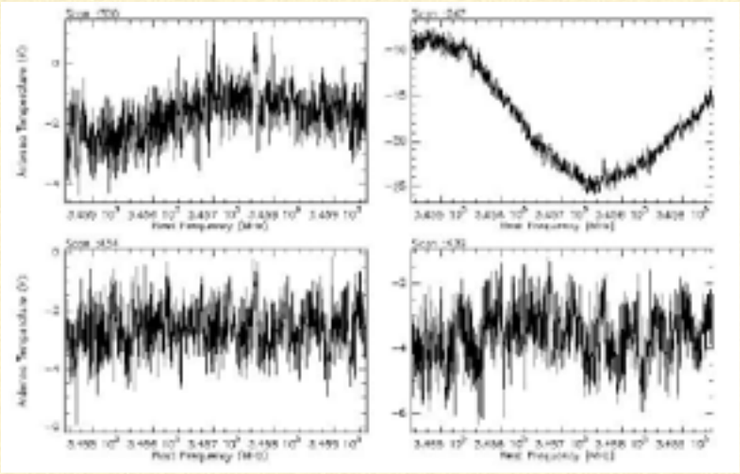
DIGITAL DATA CURATION IS…

- Acquiring verifiable digital data assets
- From analog sources (a "digital surrogate")

DIGITAL DATA CURATION IS…

- Acquiring verifiable digital data assets

- From analog sources (a "digital surrogate")
- Or assets that originated digitally ("born digital")



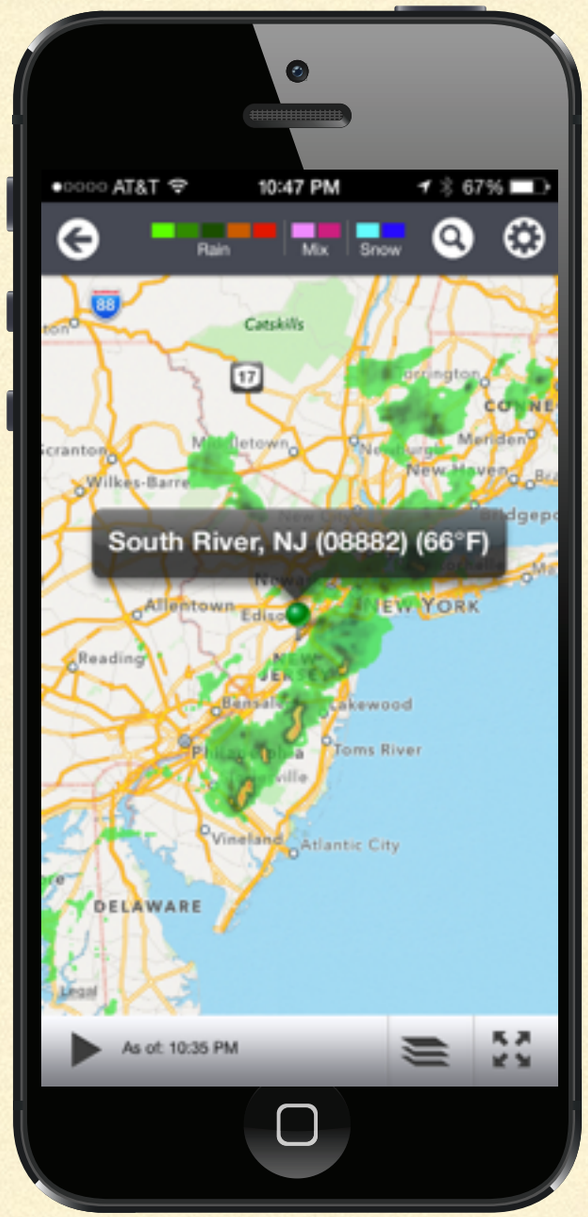
DIGITAL DATA CURATION IS…

- Seeking the best way to present that data in meaningful ways

- Why is this data being collected?
- What questions might users ask that this data can answer?
- How can we answer these questions simply and directly?

DIGITAL DATA CURATION IS…

- Seeking the best way to present that data in meaningful ways



Weather

- Present the user with a map and overlaid data.
- The user must parse that data, and make their own conclusions based on the questions they want answered.
  - What's the temperature?
  - Is it going to rain?
  - If so, for how long?
  - How heavy will the rain be?

How we've done it for 40+ years.

DIGITAL DATA CURATION IS…

- Seeking the best way to present that data in meaningful ways

Weather

- Is there a better way?
  - The processing power now exists for the weather data to be interpreted and related in plain language.
  - So, why not just simply <u>answer the questions</u>?
  - Keep the map, as an <u>option</u>, if the user really wants to see it.

A re-imagined, data-curated approach.*

* This is a real app! http://darkskyapp.com for iOS. All other platforms: http://forecast.io

DIGITAL DATA CURATION IS…

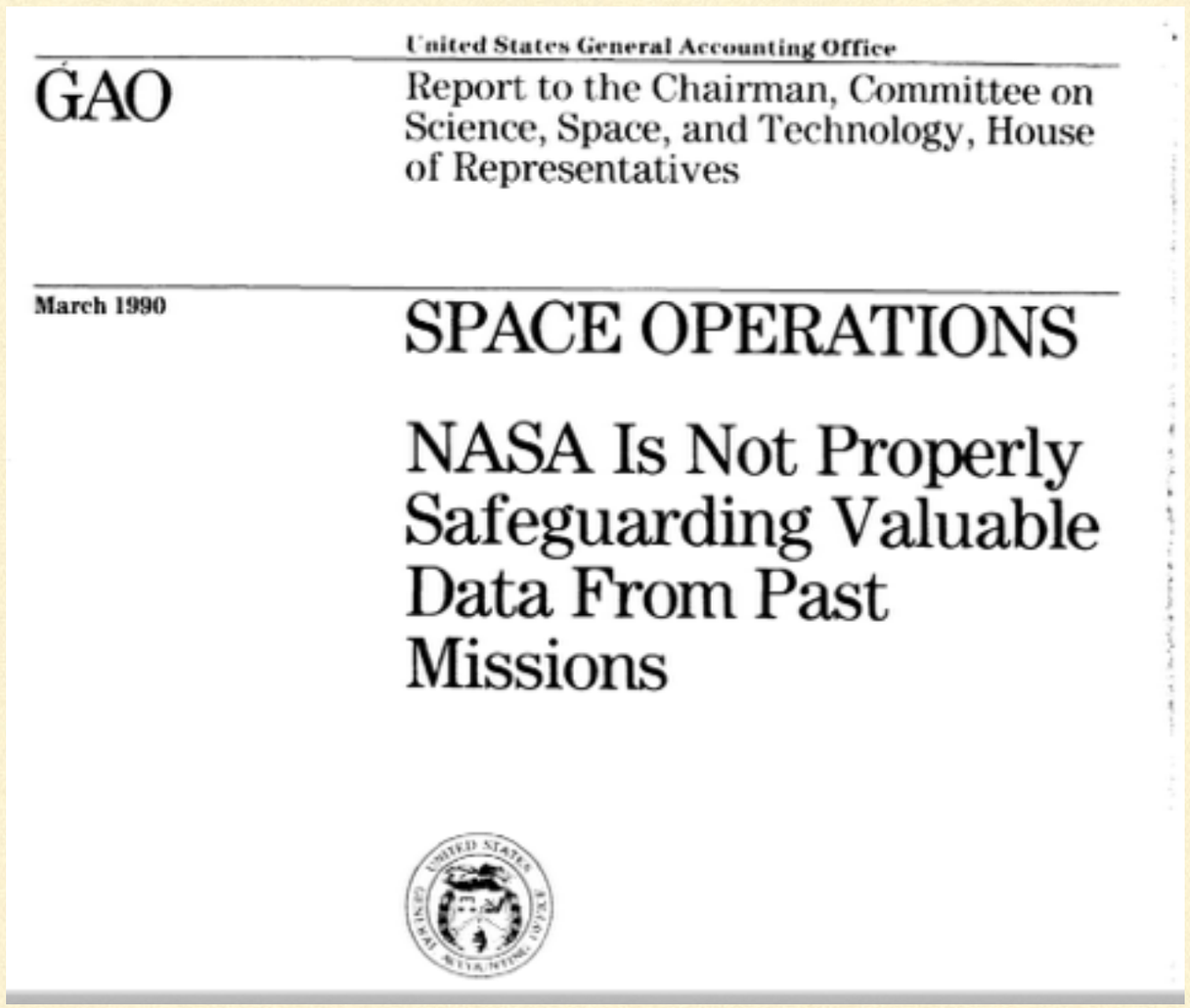
- Making sure data is preserved.

- Certifies data integrity
  - Develops minimum standards and workflow practices
  - Trains staff in handling digital assets and their containers
  - Provides Quality Assurance

- Certifies trustworthiness of the architecture
  - Vets codecs and container formats
  - Plays active role in data storage decisions
  - Implements tools and practices for continued assessment
    - Technical metadata, audit trails, chain of custody

WITHOUT DIGITAL DATA CURATION…

- Important, even historic data, is at serious risk.



United States General Accounting Office

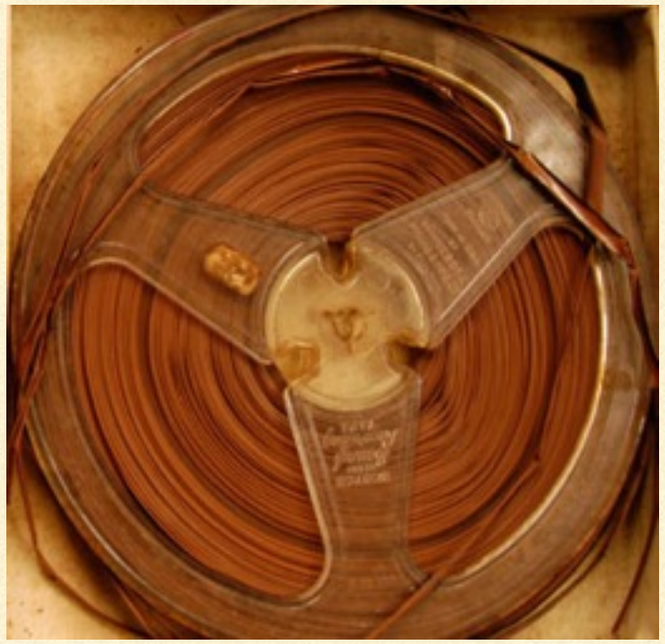
GAO

Report to the Chairman, Committee on Science, Space, and Technology, House of Representatives

March 1990

SPACE OPERATIONS

NASA Is Not Properly Safeguarding Valuable Data From Past Missions



Source: http://www.gao.gov/assets/150/148725.pdf

WITHOUT DIGITAL DATA CURATION…

- Important, even historic data, is at serious risk.

"PICTURE yourself as a historian in 2035, trying to make sense of this year's American election campaign. Many of the websites and blogs now abuzz with news and comment will have long since perished. Data stored electronically decays. Many floppy disks from the early digital age are already unreadable. If you are lucky, copies of campaign material, and of e-mails and other materials (including declassified official documents), will be available in public libraries."

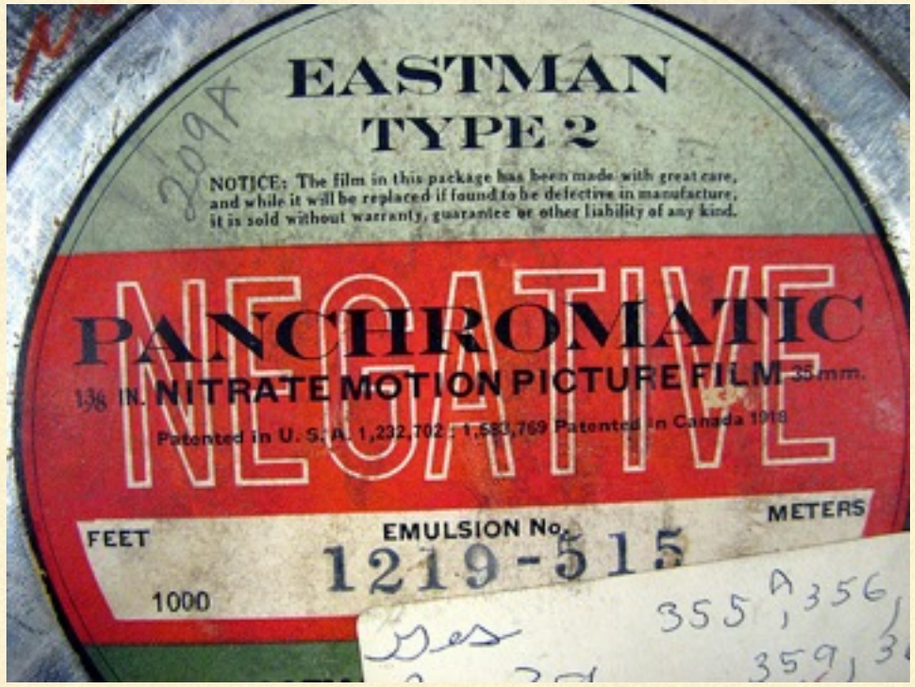
-"Bit Rot," *The Economist*, April 26, 2012

http://www.economist.com/node/21553445

WITHOUT DIGITAL DATA CURATION…

- Important, even historic data, is at serious risk.

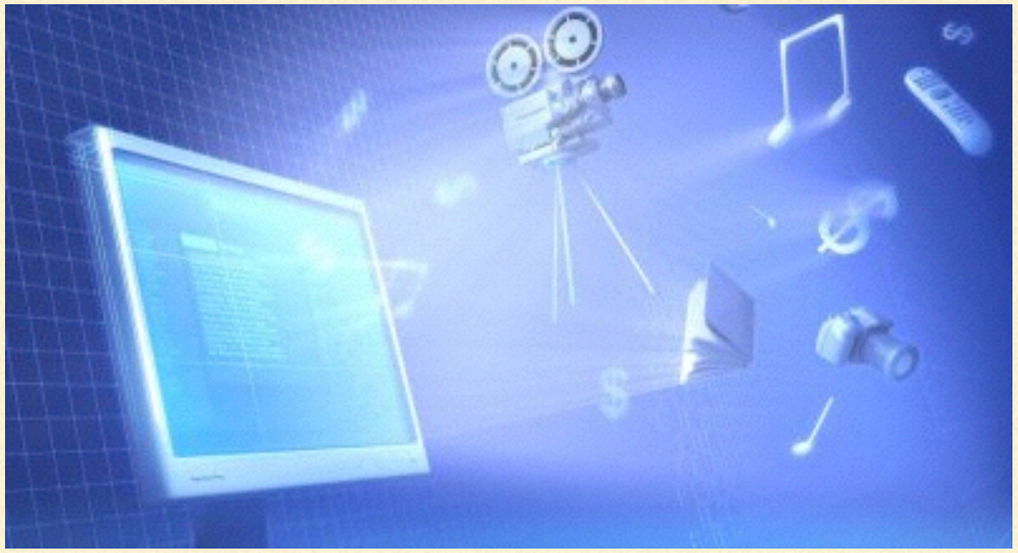
- Digital Assets are easier to destroy, more readily deleted than physical objects
  - Physical objects: typically stored, left behind, forgotten and "rediscovered"



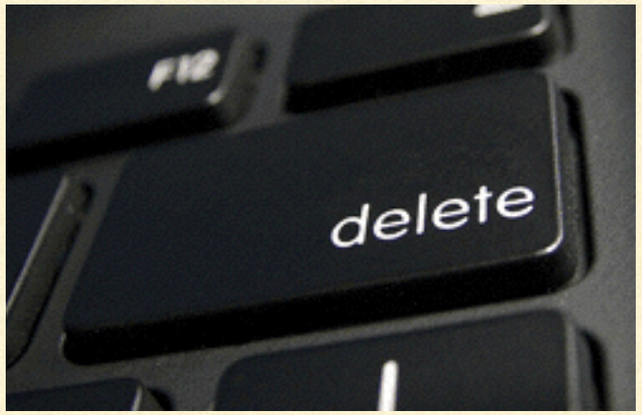
WITHOUT DIGITAL DATA CURATION…

- Important, even historic data, is at serious risk.

- Digital Assets are easier to destroy, more readily deleted than physical objects
  - Digital objects: Casual collectors typically delete what they don't want when they're low on space, or see no immediate need to retain the content.



+



WITHOUT DIGITAL DATA CURATION…

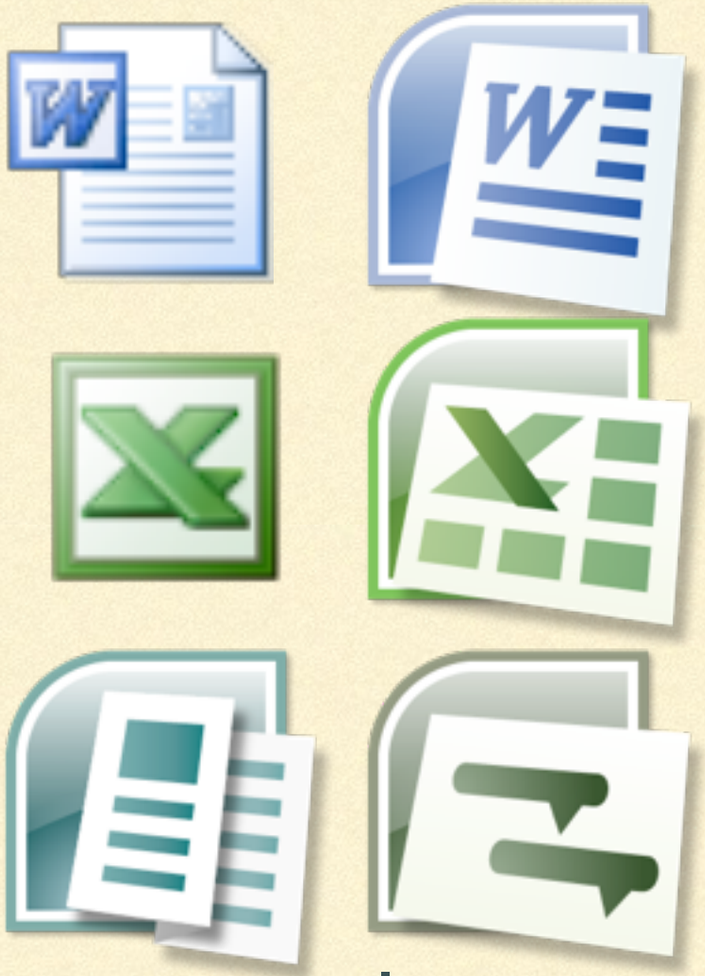
- Important, even historic data, is at serious risk.

- Digital Assets are dependent on file formats and hardware/ software platforms

WITHOUT DIGITAL DATA CURATION…

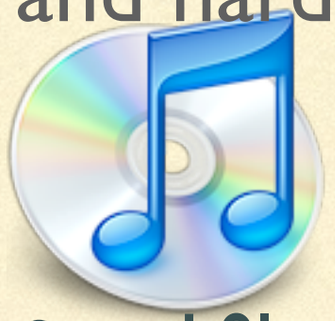
- Important, even historic data, is at serious risk.

- Digital Assets are dependent on file formats and hardware/ software platforms



Born Digital Documents

Still Images



Sound files

- At least 27 common file formats
- At least 90 common codecs

Moving Images

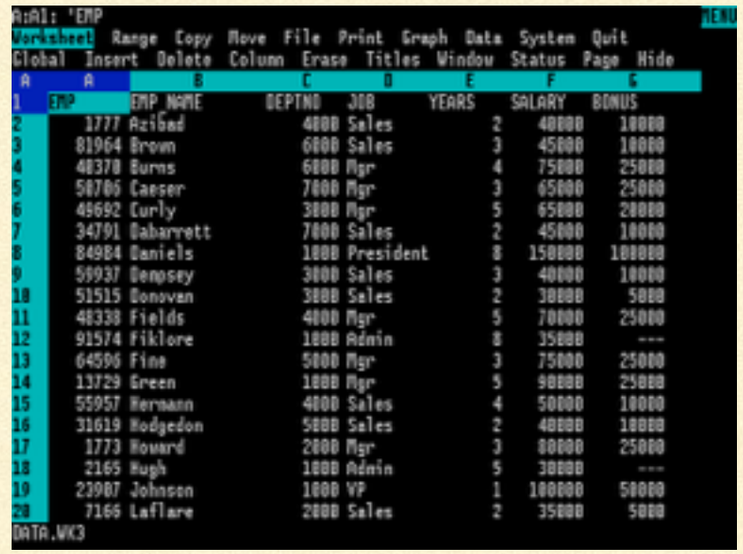
- At least 58 common containers/codecs
- + Audio tracks (27 formats/90 codecs)

WITHOUT DIGITAL DATA CURATION…

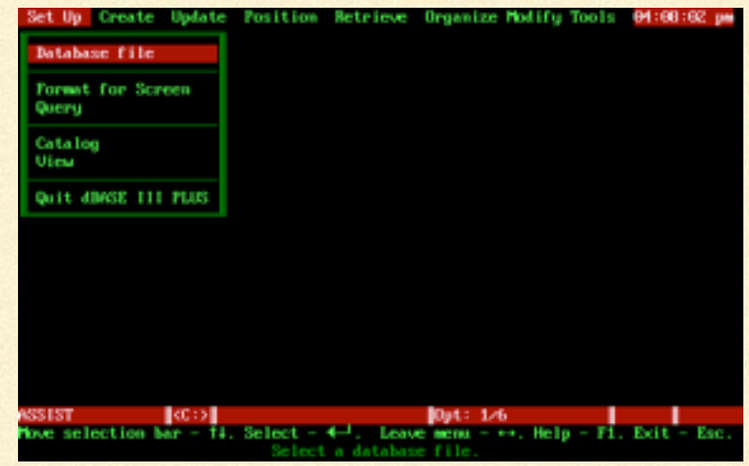
- Important, even historic data, is at serious risk.

- Digital Assets are vulnerable to format obsolescence

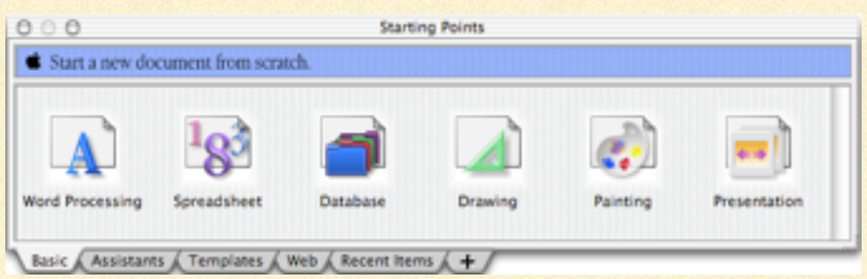
Wordstar
1978-1985



Lotus 1-2-3
1978-2013



dBase
1978 - ?



Appleworks
1984-1991

WITHOUT DIGITAL DATA CURATION…

- Important, even historic data, is at serious risk.

- Digital Assets are vulnerable to format obsolescence

Tape

Floppy Disk

Zip Disk

IBM "Demi-disk"

Video Floppy

WITHOUT DIGITAL DATA CURATION…

- Important, even historic data, is at serious risk.

- Continual data migration is necessary.
  - Storage formats must be evaluated periodically.
    - Are they still commonly used and supported?
    - Are <u>multiple</u> vendors still making the hardware/software to read/write these formats?
    - Is the format still a "standard feature" when purchasing new hardware/software?

If the answer to any of these is "no," it's time to consider migration.

WITHOUT DIGITAL DATA CURATION…

- Important, even historic data, is at serious risk.

- Long-term storage integrity must be considered.
  - One copy is **not** enough.
  - Reliable, redundant storage system for "online access"
  - A "near-line" backup system, preferably using a different storage medium
  - One "off-line, off-site" backup

- Remember: Integrity and support of all storage media <u>must</u> be periodically evaluated. Migration will eventually have to take place as new storage technology emerges.

ADDITIONAL CHALLENGES

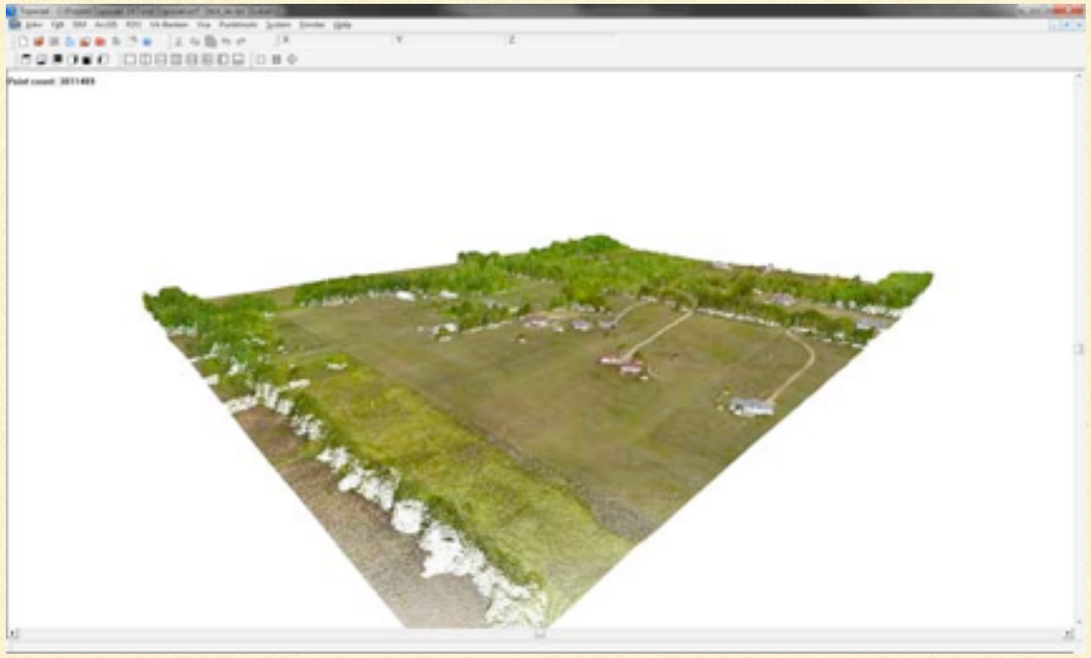
- Datasets and File Formats

- Can be anything.
  - Already known and established formats, OR
  - Totally new formats: SUR, CSFASTA...

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*
```

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```



ADDITIONAL CHALLENGES

- Datasets and File Formats

- Most traditional file types are "known quantities."
  - Predictable use cases
  - Rigid standards
  - Built-in familiarity

ADDITIONAL CHALLENGES

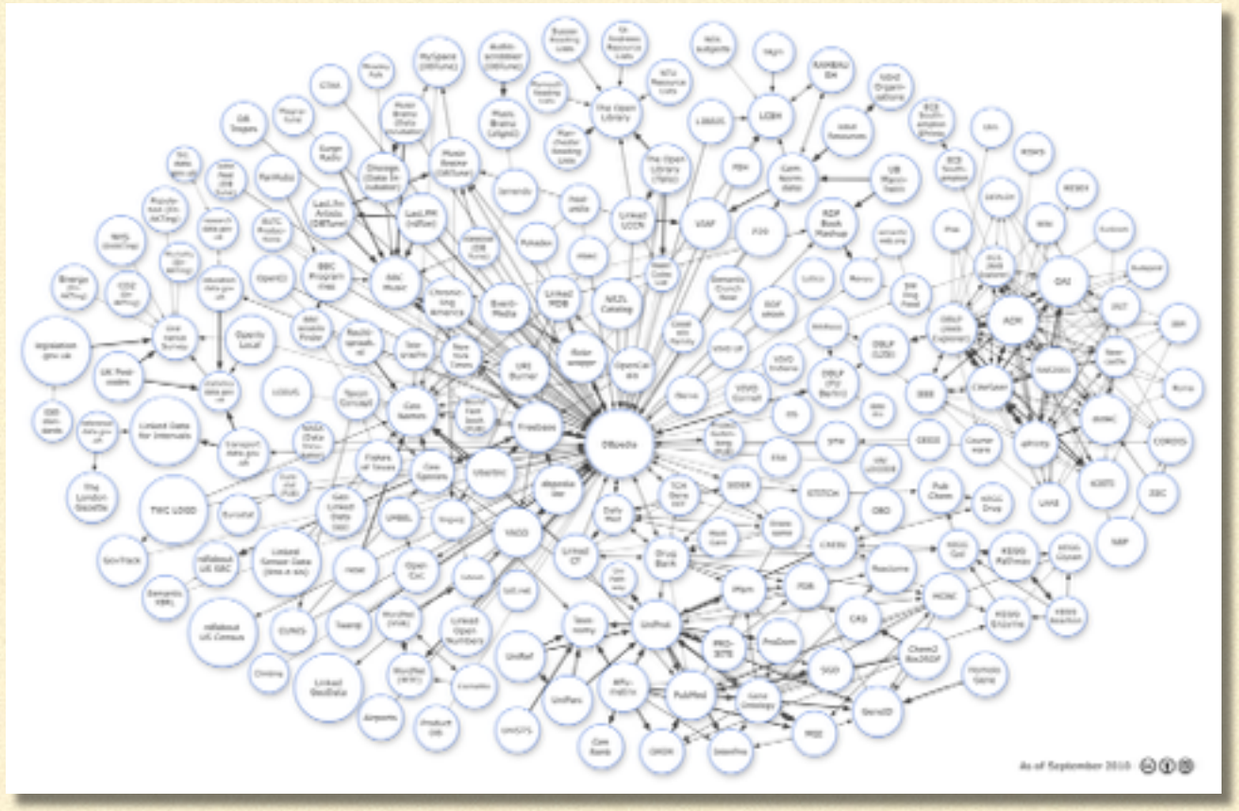
- Datasets and File Formats

- Some datasets are stored merely as common file formats used for the purpose of data gathering, e.g.
  - MS Excel spreadsheets with data points, PDF files with written content.
  - Still images, sound, or moving images captured as part of research.

ADDITIONAL CHALLENGES

- Datasets and File Formats

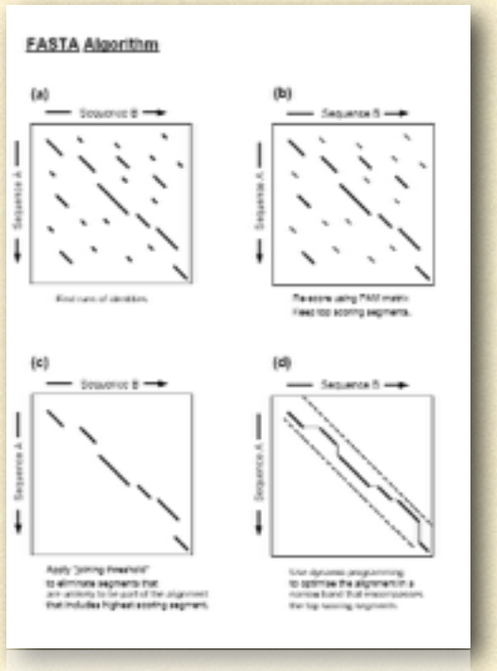
- Research Data will take out us out of our comfort zone
  - Unique, non-traditional, obscure data types
  - Traditional files used in non-traditional ways
  - Usages and implementations that require a learning curve to the uninitiated.



ADDITIONAL CHALLENGES

- Datasets and File Formats

- Other dataset file formats can be extensions of existing file types that are re-purposed. Can be human-read, or interpreted with additional, special-purpose software.
  - e.g. Repurposed UTF-8 (text file) to create a FASTA sequence.

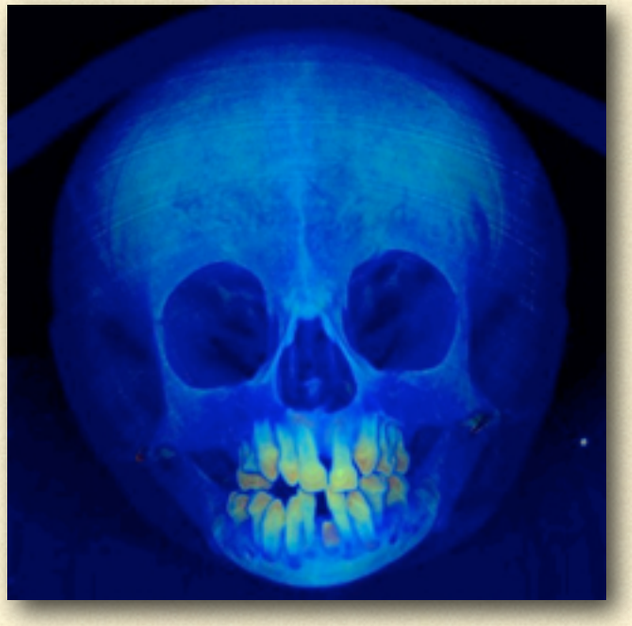


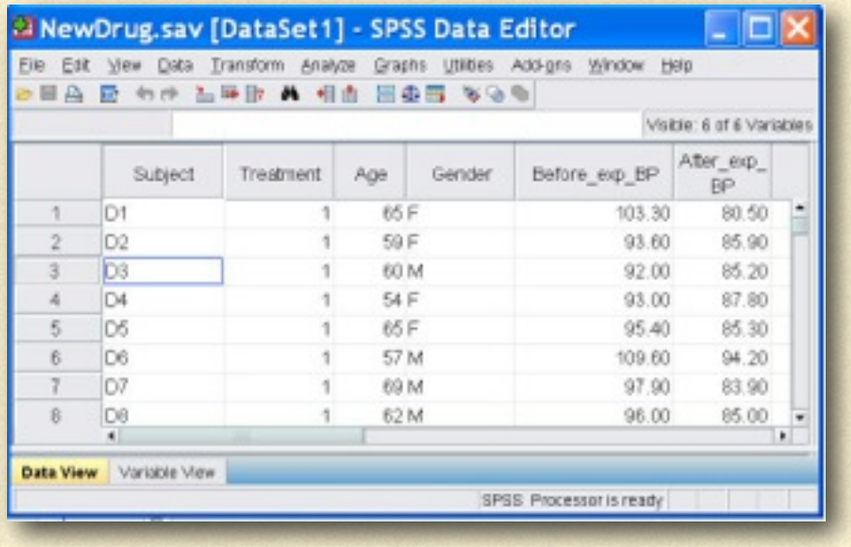

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

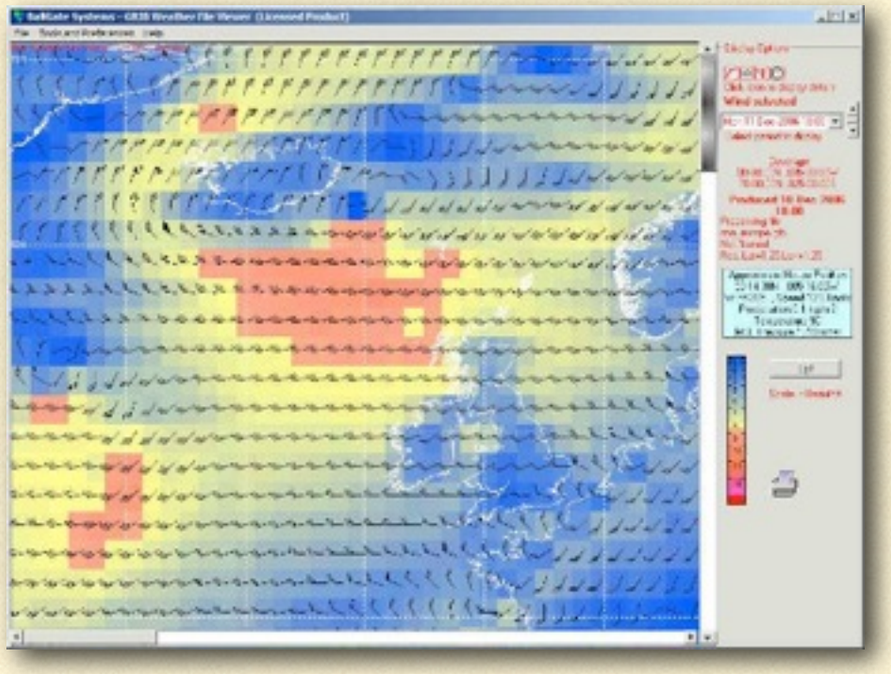
ADDITIONAL CHALLENGES

- Datasets and File Formats

- Finally, some datasets can be completely proprietary, custom and closed, requiring specialized hardware/software to access or interpret.
  - e.g. GRIB, SUR, DWG, SPSS... too many to list!







ADDITIONAL CHALLENGES

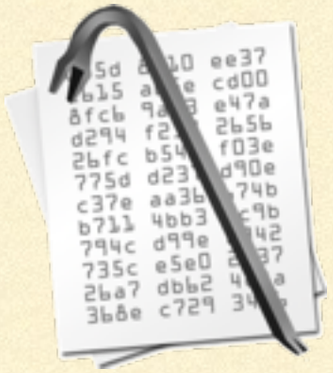
- Datasets and File Formats

- Fortunately, there are some tools to help us out in determining the nature of different files.

<xml>exif</xml>

exiftool

mediainfo



hex/text viewers

command line

EXAMPLE: EXIFTOOL

<xml>exif</xml>

- Used to identify most common file formats, even if renamed.
- Can discover and extract underlaying metadata.

Filename: IMG_2984.JPG
Date Taken: ?
Location: ?
Context: ?

NEXT STEPS

- We've identified and acquired data.
- We're aware that data is big and growing.
- We know the challenges of preserving data.

Now what?

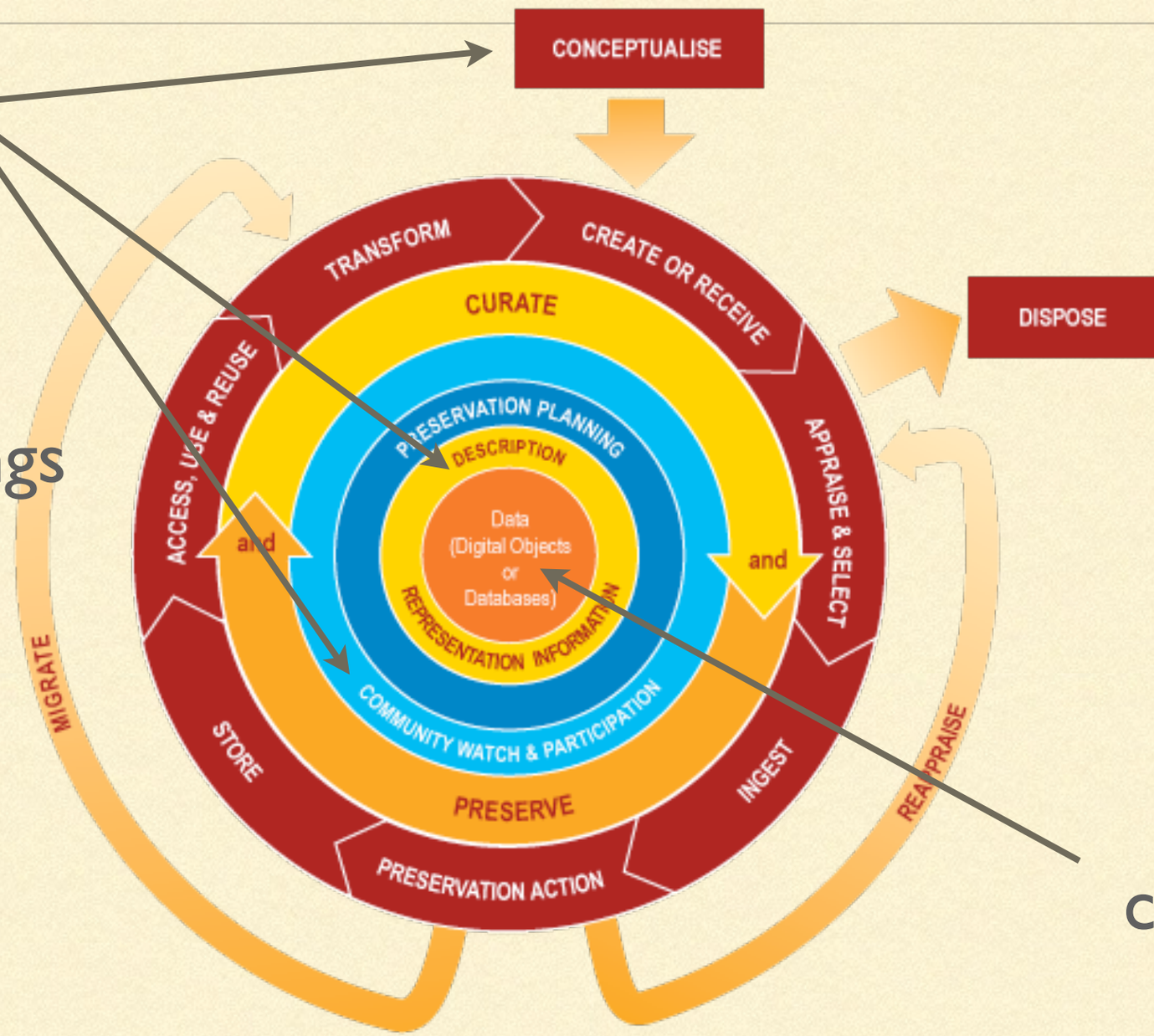
DIGITAL CURATION LIFECYCLE

- A multitiered, continuous process where digital objects of any type are evaluated, preserved, maintained, verified, and re-evaluated.
- Iterative: the cycle doesn't end with one go-round.
- A useful exercise for known and as-yet-unknown file types and formats.

DIGITAL CURATION LIFECYCLE

Collaborate:

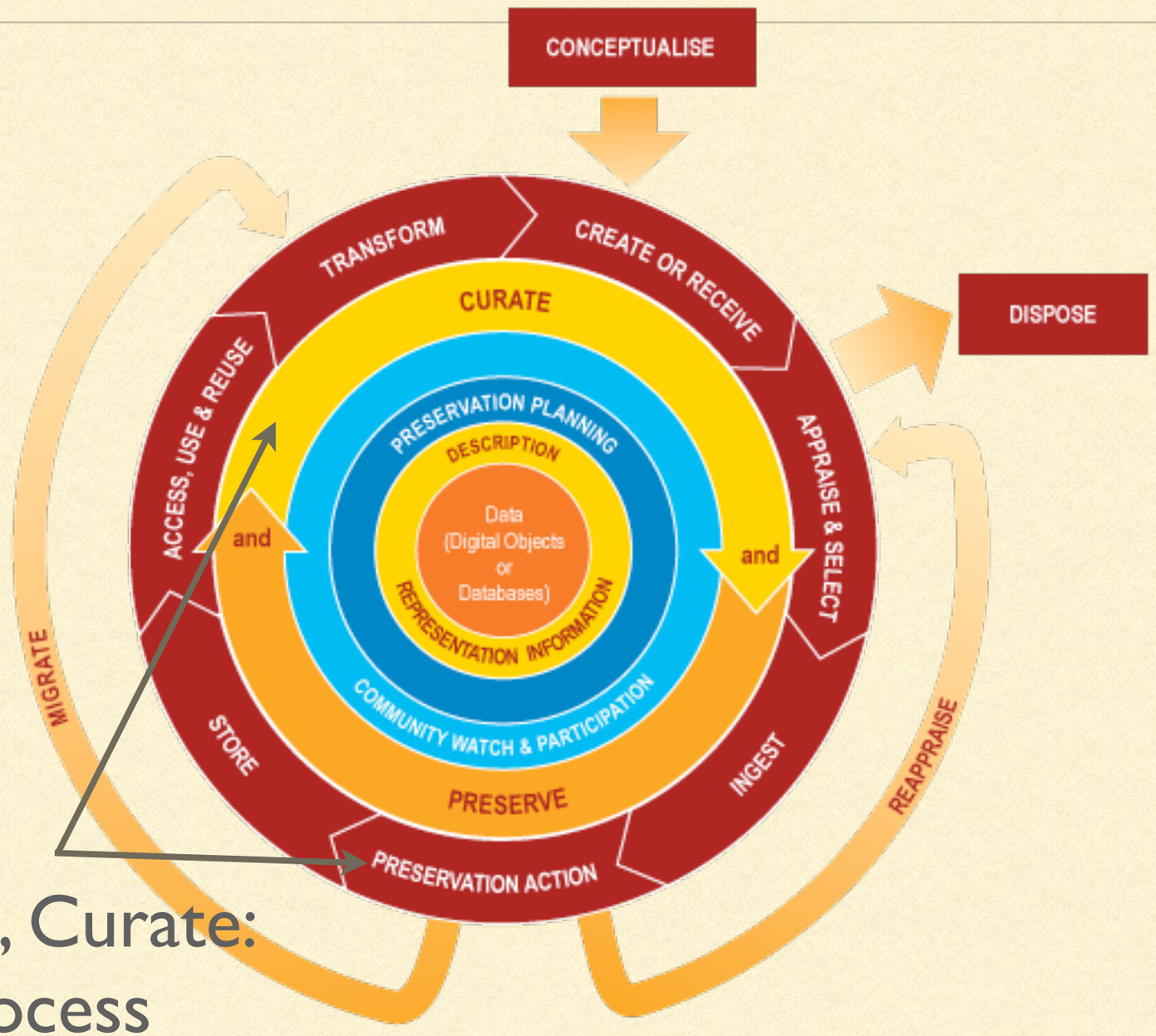
- Plan
- Describe
- Evaluate
- Learn meanings



Data: The center of our "universe."

* Source: Digital Curation Centre, http://www.dcc.ac.uk/resources/curation-lifecycle-model

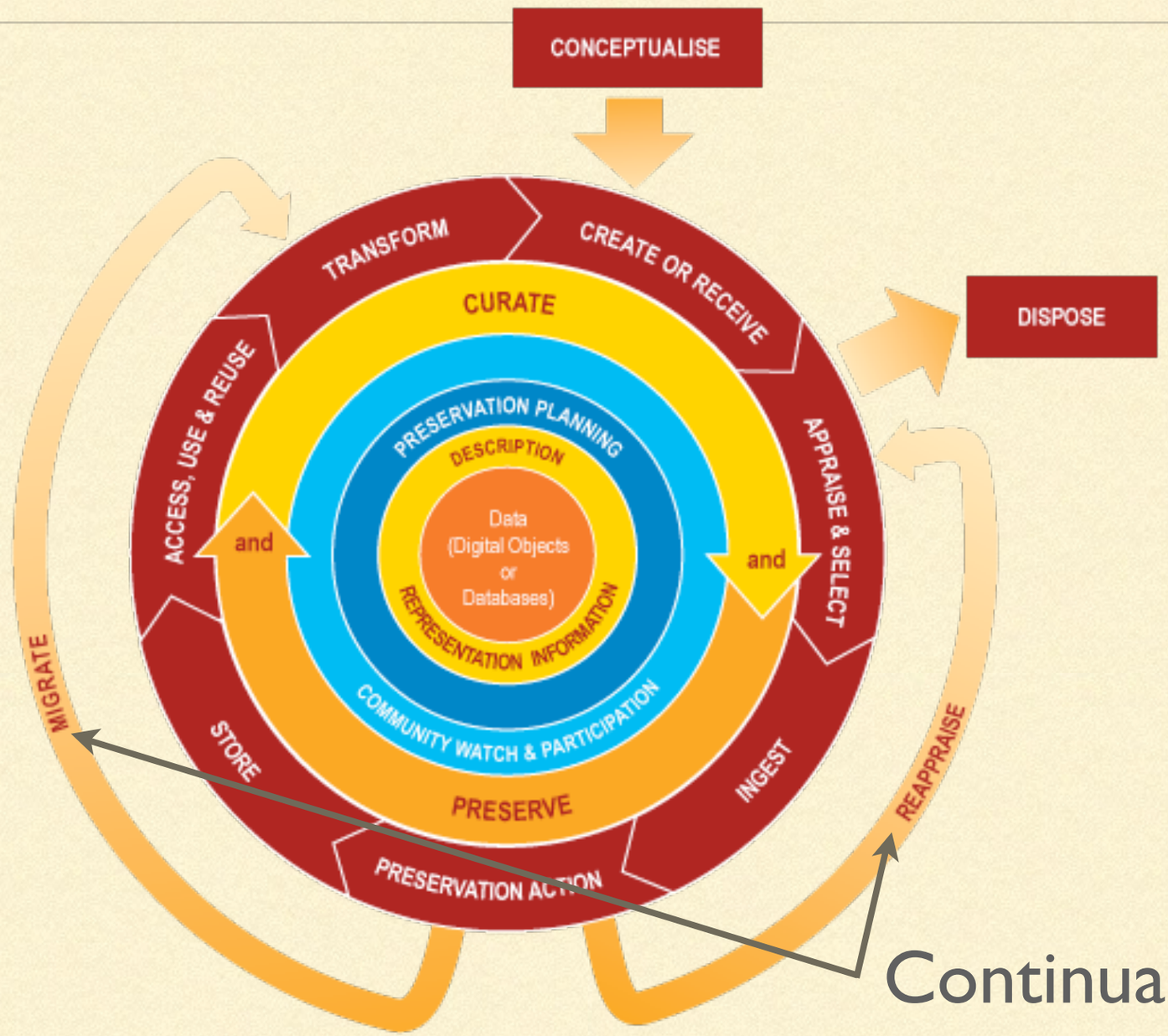
DIGITAL CURATION LIFECYCLE



Ingest, Preserve, Curate:
The RUcore Process

* Source: Digital Curation Centre, http://www.dcc.ac.uk/resources/curation-lifecycle-model

DIGITAL CURATION LIFECYCLE



Continual, iterative post-ingest actions

* Source: Digital Curation Centre, http://www.dcc.ac.uk/resources/curation-lifecycle-model

SOLUTIONS: A "CONTROLLED CHAOS"

- Evaluate the data, the research project, and the researcher's needs. Creation of a descriptive, comprehensive data model for the project is key.
  - Take stock of Software, Systems, measuring/lab equipment, and recording apparatus.
  - Often, we must accept that de facto industry/research standards become de facto preservation standards.

SOLUTIONS: A "CONTROLLED CHAOS"

- **Collaborate and communicate with the researcher**
  - Establish a format guide and handling procedures. Evaluate the veracity and longevity of the data format. Check competitors, alternatives, and potential successor formats. Publish, share and use the findings.
  - Determine methods of access. How are users expected to access and view the data?
    - What are the software and hardware requirements?
    - Do you view the data online? Use a plugin? Download and use separate software?

SOLUTIONS: A "CONTROLLED CHAOS"

- **<u>Do No Harm</u> to digital assets**
  - Preservation masters, derivatives when needed
  - Content modification must be done with extreme care
  - Any changes must be traceable, auditable, reversible

SOLUTIONS: A "CONTROLLED CHAOS"

- **Prepare for the inevitable: format migrations**
  - Periodically re-assess the relevant format
  - Migrate to new formats when the old is obsolete
  - Maintain accessibility while ensuring data integrity

PARTING THOUGHTS

- **Digital Curation is a process where learning is continual.**
  - No single person can know <u>every</u> format, <u>every</u> file type, <u>every</u> technology.
  - It's okay to say "I don't know, but I'll find out."
    - Ask questions. Seek preservationist communities. Share and compare notes.

PARTING THOUGHTS

- **Remember: Don’t Panic!**
  - Getting too anxious over what to do wastes time, and doesn’t get the data any closer to being stabilized or preserved.
  - The internet is your friend! Open Access data is intended to be shared. So, It’s unlikely you’re the only person to have encountered X data format. Someone else online may have the answers you seek.

PARTING THOUGHTS

- **Always pay it forward**
  - Once you've become an "expert" at a new data format, share what you've learned. Document your knowledge and workflow. Share your standards and recommendations.
  - Eventually, you'll be the "other person out there" who someone will come to for help and advice.

QUESTIONS?

Isaiah Beard
Digital Data Curator
+1 848 932 5932
isaiah.beard@rutgers.edu

Additional Resources:

Digital Curation Center:
www.dcc.ac.uk

Australian National Data Service
www.ands.org.au

Blog: From Page2Pixel
page2pixel.org