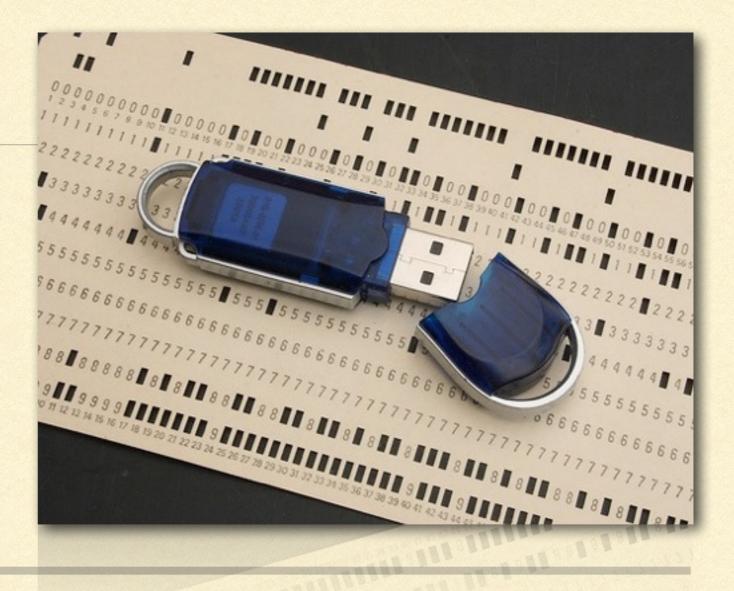


#### The Digital Curator's Guide to the Data Format Methodology

#### Presented by

#### **Isaiah Beard**

Digital Data Curator Rutgers University Libraries



### TOPICS COVERED TODAY:

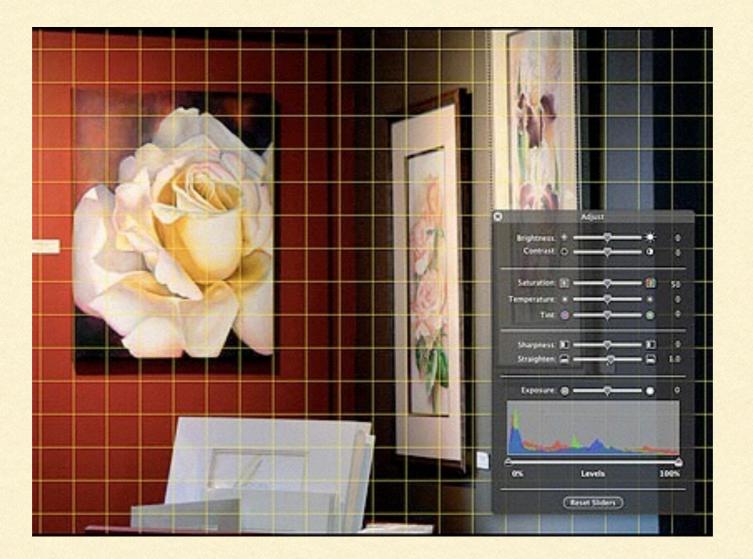
- Definitions, Concepts
  - What is Data Curation?
  - Why is it important?
- The Data Curation Lifecycle
  - Practical approaches to assessing and handling data
- Applying these concepts to research data



### 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/

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



# 2.5 Quintillion bytes of data are generated every day.

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

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

2.5 Petabytes (2,500 ITB 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

#### 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>I hour.</u>



#### 5,000 data samples are taken per second.

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

E-mail

Early 2000s and prior:

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

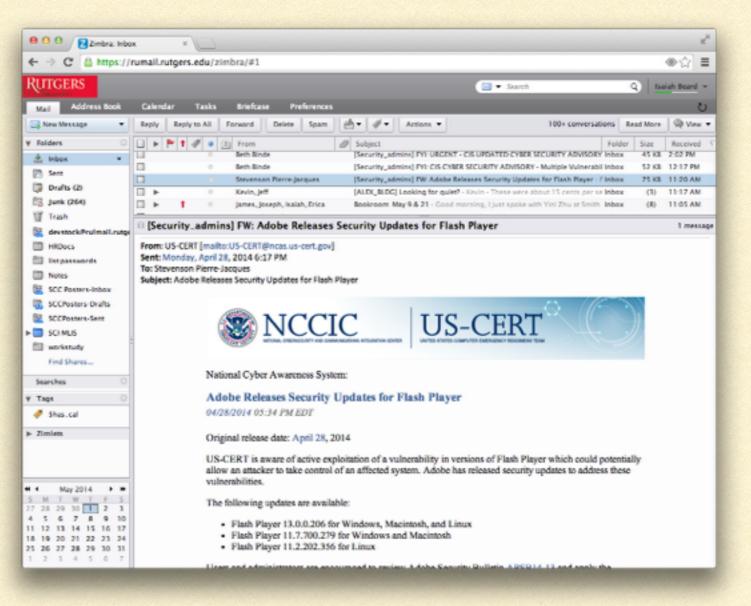
| ● ● ● ●      |                      |  |  |  |  |
|--------------|----------------------|--|--|--|--|
| PINE 4.64    | MESSAGE INDEX <      | Mail> sent-mail-nov-2002 Msg 151 of 159 NEW  |  |  |  |
|              | T                    |  |  |  |  |
|              | 1 To: Jill Nathanson | (681) Re: trouble ticket                     |  |  |  |
|              | 1 To: Stephanie Bart |  |  |  |  |
|              | 2 To: sbartz@rci.rut |  |  |  |  |
|              | 2 To: Megan Doyle    | (948) Re: printer problem- Chemistry Library |  |  |  |
|              | 2 To: Jill Nathanson |  |  |  |  |
| N 138 Nov 1  | 2 To: abutman@rci    | (649) Computer in Ryoko's Office - LSM       |  |  |  |
| N 139 Nov 1  | 2 To: abutman@rci    | (778) Sondra Bodner's computer               |  |  |  |
| N 140 Nov 1  | 2 To: Stephanie Bart | (1725) Re: printer locks                     |  |  |  |
| N 141 Nov 1  | 2 To: sbartz@rci, aj | (2988) FW: Why mail is slow on RCI (fwd)     |  |  |  |
| N 142 Nov 1  | 2 To: Pat Piermatti  | (1096) re: staff absence record              |  |  |  |
| N 143 Nov 12 | 2 To: Stephanie Bart | (1509) Re: Ryoko's Monitor                   |  |  |  |
| N 144 Nov 1  | 3 To: calhoun@rci.ru | (555) Bad patent BIB                         |  |  |  |
| N 145 Nov 1  | 3 To: lsm_fac@email  | (1000) Laser printer at reference desk       |  |  |  |
| N 146 Nov 1  | 5 To: Stephanie Bart | (1121) Re: laptops                           |  |  |  |
|              | 5 To: hoffmanh@rci.r | (740) Computer updates                       |  |  |  |
| N 148 Nov 1  | 7 To: Kevin P. Mulca |  |  |  |  |
|              | 8 To: Jill Nathanson | (702) Re: reference computer password        |  |  |  |
|              | 9 To: Pat Piermatti  |  |  |  |  |
|              | 9 To: Galen Work     |  |  |  |  |
|              |                      |  |  |  |  |
| 7 Help       | FldrList P PrevMs    | g – PrevPage 🖸 Delete 🖪 Reply                |  |  |  |
|              | [ViewMsg] N NextMs   |  |  |  |  |

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

E-mail

Today:

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



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

### E-mail

Today:

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

| and the second  |  |   |  |
|---|--|---|--|
| 000   |  | ox — RUMail (9808 messages) y   |  |
|   | • • • • •  | Q   |  |
| Filds   Inbox (1) * Sent *  | Drafts (2) * Flagged   |   |  |
| MAILBOXES   | Bort by Date *         Jane Otto         3.03 FM           [Ru[_everyond] RUL videos         Media Toam, Matt and I completed this survey. It's probably a good time to start discussing our cap           Dave Heaver         2.48 FM           [Bw_arch] php-clamav         2.48 FM           [Bw_arch] php-clamav         2.48 FM           Dave Boever         2.48 FM           [Bw_arch] php-clamav         100 probably a good time to place on rep-devel           Dave Dave Hoover Systems Programmer Rug         Image: Dave Dave Action of the place on rep-devel           Dave Dave Action of the place on rep-devel         Image: Dave Dave Action of the place on rep-devel           Dave Dave Action of the place on rep-devel         Image: Dave Dave Action of the place on rep-devel           Dave Dave Action of the place on rep-devel         Image: Dave Dave Action of the place on rep-devel           Dave Dave Action of the place on rep-devel         Image: Dave Dave Action of the place on rep-devel           Dave Dave Action of the place on rep-devel         Image: Dave Dave Action of the place on rep-devel           Dave Dave Action of the place on rep-devel         Image: Dave Dave Action of the place on rep-devel           Dave Dave Action of the place on rep-devel         Image: Dave Dave Action of the place on rep-devel           Dave Dave Action of the place on rep-devel         Image: Dave Action of the place on rep-devel <t< td=""><td>Stevenson Pierre-Jacques         May 1, 2014         11:20:AM           To:::ecurity_admins/Ferrel.rutgers.edu         Fide Datable         Fide Datable           [Security_admins/Ferrel.rutgers.edu         Fide Datable         Fide Datable           FVI         From: US-CERT [mailto:US-CERT/Rincol.us-cert.mov]         Sent: Monday, April 28, 2014 6:17 PM           To: Stevenson Pierre-Jacques         Fide Datable         Fide Datable</td></t<> | Stevenson Pierre-Jacques         May 1, 2014         11:20:AM           To:::ecurity_admins/Ferrel.rutgers.edu         Fide Datable         Fide Datable           [Security_admins/Ferrel.rutgers.edu         Fide Datable         Fide Datable           FVI         From: US-CERT [mailto:US-CERT/Rincol.us-cert.mov]         Sent: Monday, April 28, 2014 6:17 PM           To: Stevenson Pierre-Jacques         Fide Datable         Fide Datable  |  |
| <ul> <li>v ≪ Sent</li> <li>√ Kloud</li> <li>√ RUMail</li> <li>√ Lf Prints</li> <li>√ MUMail</li> <li>☆ Kloud</li> <li>⊕ RUMail</li> <li>⊕ Lf Prints</li> <li>♥ 🗇 Trash</li> <li>☆ Kloud</li> <li>☆ RUMail</li> <li>☆ LF Prints</li> <li>MAL ACTIVITY</li> </ul> | <ul> <li><sup>9</sup> Hi Isalah, Just a heads up. James wasn't at the BCC<br/>when I stopped by, so I left the discs with Rhonda, AL.,</li> <li>Bath Binde 2:02 PM<br/>(Security, admins) FYL URGENT - CIS UPDATED C,<br/>Patches to IE are out Ben TLP:WHITE CIS<br/>CYBER SECURITY ADVISORY CIS ADVISORY NU</li> <li>Bath Binde 12:17 PM<br/>(Security, admins) FYL CIS CYBER SECURITY AD.,<br/>TLP:WHITE CIS CYBER SECURITY ADVISORY CIS<br/>ADVISORY NUMBER: 2014-043 DATE(S) ISSUED</li> <li>Stevenson Pierre-Jacques 11:20 AM</li> </ul>   | Subject: Adobe Releases Security Updates for Flash Player           With the second sec |  |
|   | FW: [Security_admins] Addde Releases Security Up<br>Print Prove: US CERT (Invate-US CERT Braces us<br>cont pow) Sent: Monday, April 28, 2014 617 PM Tor<br>Jeff Teichmann 15:17 AM<br>Re: [ALEX, BLDG] Looking for quiet?<br>Kevin - These were about 15 cents per set, so I think<br>we don't need to worry about non RU toks if they as<br>Erica Parin 11:05 AM<br>Docksoom May 9:4, 21  | Original release date: April 28, 2014<br>US-CERT is aware of active exploitation of a vulnerability in versions of Flash Player which co<br>potentially allow an attacker to take control of an affected system. Adobe has released security u<br>address these vulnerabilities.  |  |
|   | Good moming, I just spoke with Yini Zhu at Smith.  | The following updates are available:<br>• Flash Player 13.0.0.206 for Windows, Macintosh, and Linux<br>• Flash Player 11.7.700.279 for Windows and Macintosh<br>• Flash Player 11.2.202.356 for Linux   |  |
|   |  |   |  |

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

E-mail

Today:

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

|   | >>>> AT&T ♥   | 4:43 PM     | 7 %  | 88%        |  |  |
|---|---|-------------|------|------------|--|--|
|   | From: Steve   |             |      |            |  |  |
|   | [Security_ad<br>Releases Se<br>Player<br>May 1, 2014 a  | ecurity Upd |      | Flash      |  |  |
| i | FYI   |             |      |            |  |  |
|   | From: US-CERT [mailto:US-<br>CERT@ncas.us-cert.gov]<br>Sent: Monday, April 28, 2014 6:17 PM<br>To: Stevenson Pierre-Jacques<br>Subject: Adobe Releases Security<br>Updates for Flash Player |             |      |            |  |  |
|   | 8 NCCI  | C US-       | CERT | $\bigcirc$ |  |  |
|   | National Cyber Assurances System<br>Adobe: Releases Security, Cj<br>10 (19 (19 10) Art 20 (19 10)   |             |      |            |  |  |
| 1 | P 🗖   | Û           | \$   |            |  |  |
|   |   |             |      |            |  |  |

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

E-mail

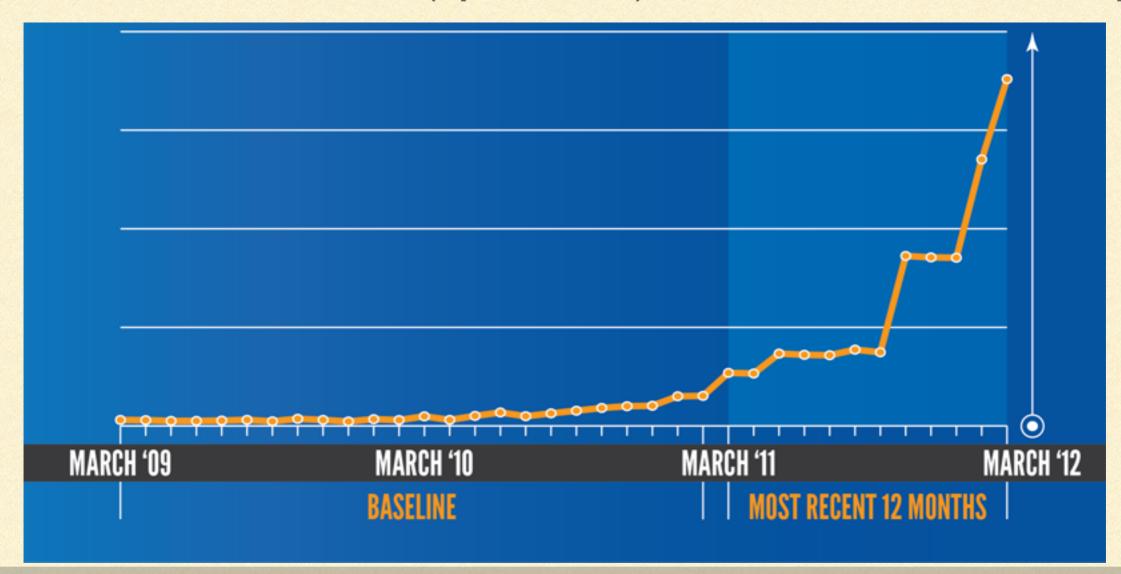
3.37 Billion e-mail accounts worldwide

2.25 Billion Consumer

850 Million Corporate

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

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

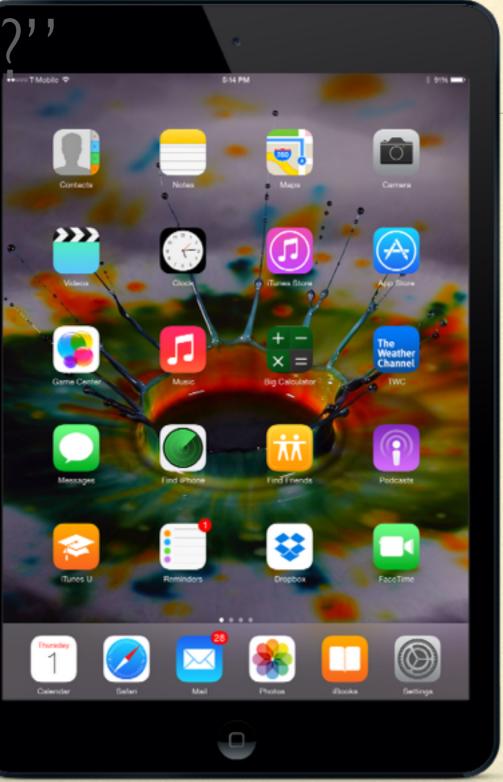


IBM (2013, December). "Applying new analytics tools to reveal new opportunities." <u>http://www.ibm.com/smarterplanet/us/en/business\_analytics/article/it\_business\_intelligence.html</u>

#### **Mobile Devices**







#### **1984: The Cray XMP Supercomputer**

- THE most powerful computer in existence at the time
- I 50,000 200,00 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

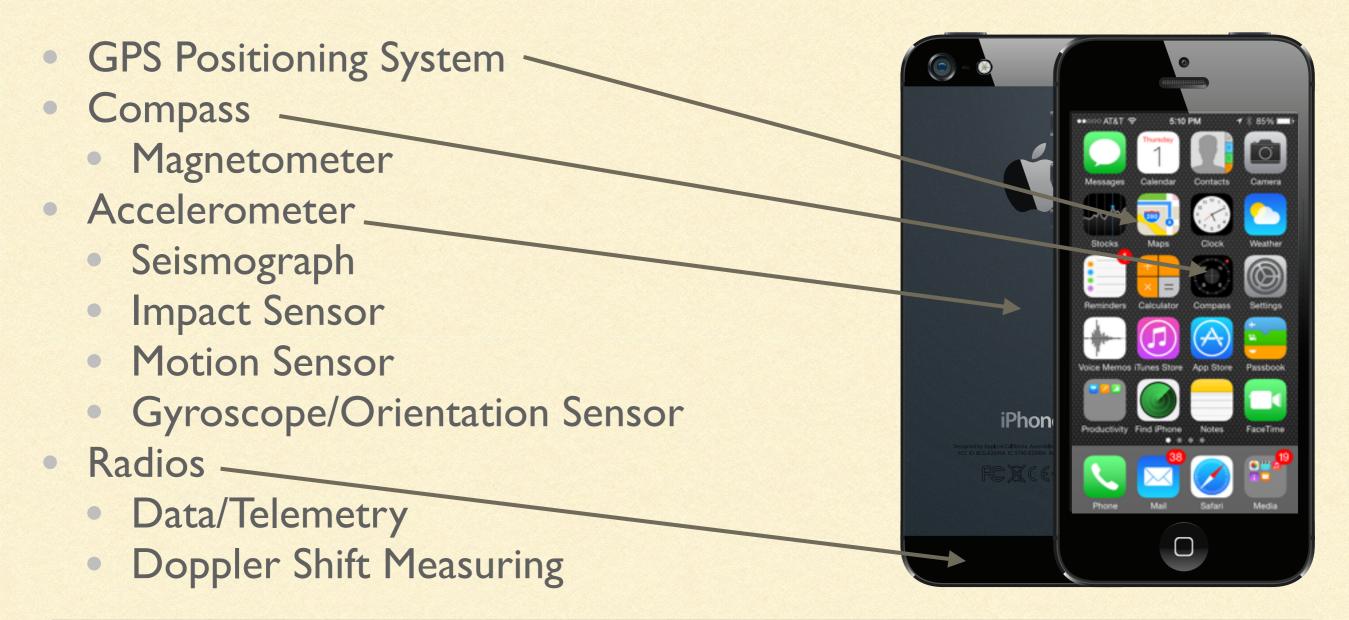


# Mobile Devices and Ubiquitous Computing 2007-Present: Smartphones, Tablets

- NOT the most powerful computers in existence for their time
- 5.45 watts max
- IGB of RAM, some higher
- I6 64GB of storage, some are expandable
- Fits in your pocket, 1.75 <u>Billion</u> 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)



#### **Mobile Data Gathering and Processing**



#### **Mobile Data Gathering and Processing**

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



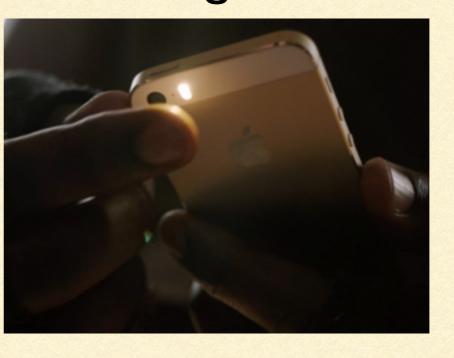
#### **Mobile Data Gathering and Processing**

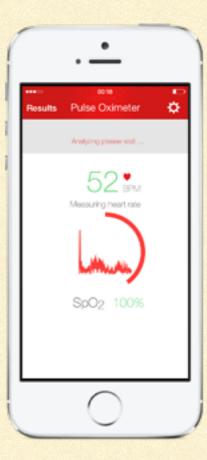
- Camera
  - Recorder
  - Color Densitometer
  - Motion/Pattern detector
  - Barcode Reader



#### **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.

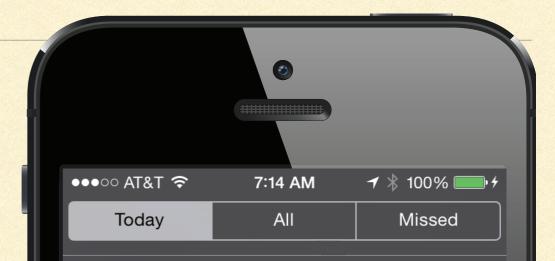
#### The Smartphone Dilemna: Data that's <u>maybe</u> a little too smart...

- 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.



#### "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



#### Thursday, April 17<sup>th</sup>

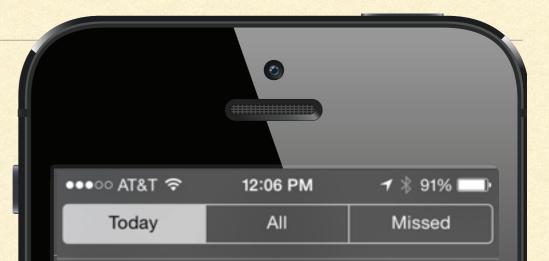
Sunny today. It's currently 39°; the high will be 50°.

 Right now, it would take you about 17 minutes to drive to work.

The first thing on your calendar today is "NJEDL - Maps scanning", in 46 minutes.

#### "Smart" data, today

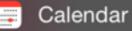
- The Result:
  - My mobile devices are "aware" of my daily routine
    - When and where I go to work
    - When/where I go to lunch
    - When/where I go home
    - Birthdays
    - Anniversaries
    - Meetings



Monday, May 5<sup>th</sup>

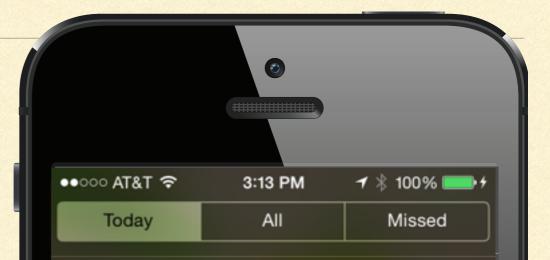
Mostly sunny currently. It's 63°; the high will be 64°.

 Right now, it would take you about 8 minutes to drive to Piscataway.



#### "Smart" data, today

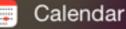
- The Result:
  - My mobile devices are "aware" of my daily routine
    - When and where I go to work
    - When/where I go to lunch
    - When/where I go home
    - Birthdays
    - Anniversaries
    - Meetings



#### Wednesday, May 7<sup>th</sup>

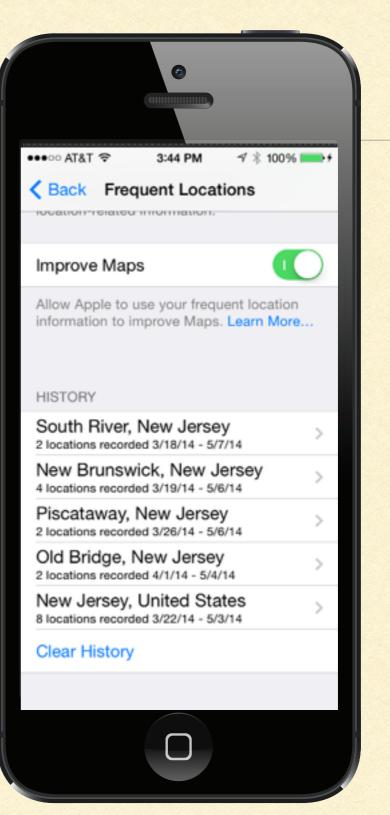
Mostly sunny currently. The high will be 68°. Cloudy tonight with a low of 50°.

Traffic is unusually heavy on the way home. The driving time is 21 minutes.



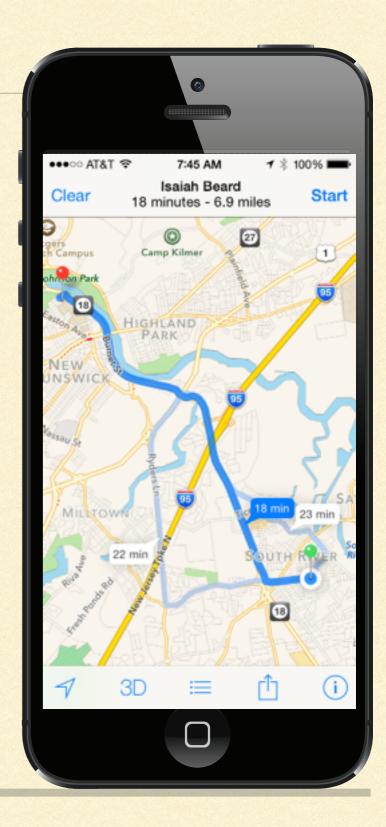
#### "Smart" data, today

- 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



#### 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.



### 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

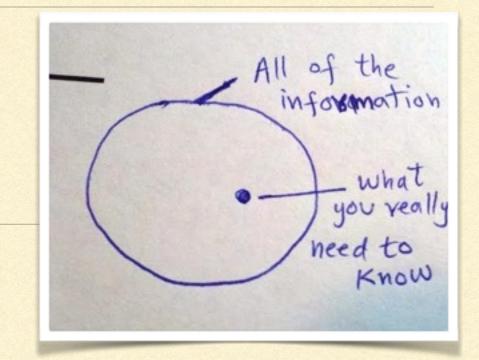
- All of this data collection means there are enormous challenges ahead. For data to be useful, it must be
  - Sifted
  - Processed
    - The end result of data collection is that one or more humans, <u>somewhere</u>, will have their question(s) answered.
    - This requires
      - Processing collected 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, protected, and durable.
- This gives us a fighting chance to meet the challenge of sifting, processing and preservation.

#### 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

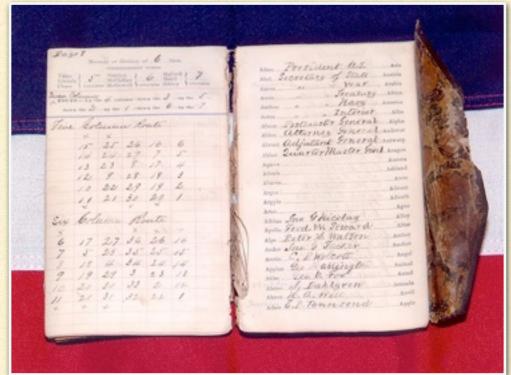
 "A set of files containing both research data - alphanumeric or encoded - and <u>documentation</u> sufficient to make the data re-usable."\*

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

#### Understanding what data consists of

#### Documentation

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

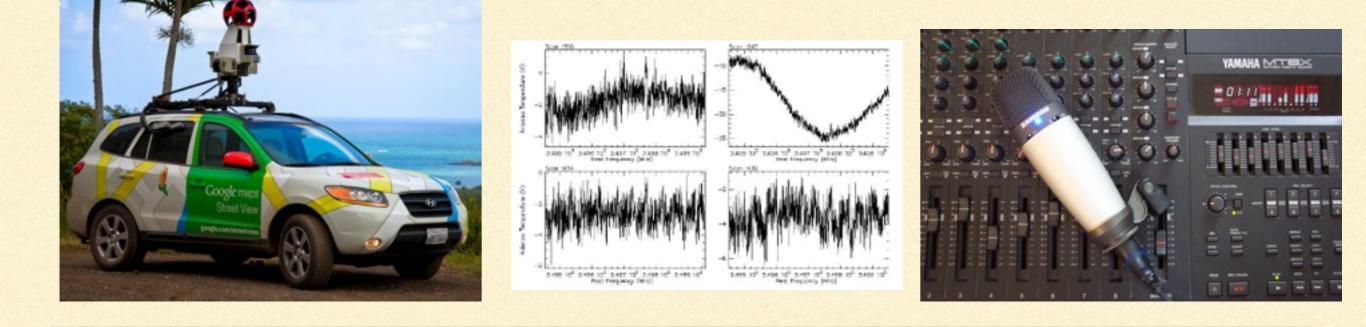


Such as a codebook, technical or methodology report, or user guide.

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



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



- 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 <u>simply</u> and <u>directly</u>?

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</u> <u>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! <u>http://darkskyapp.com</u> for iOS. All other platforms: <u>http://forecast.io</u>

# 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

Important, even historic data, is at serious risk.

#### ĠAO

Control States General Accounting Office 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

Important, even historic data, is at serious risk.



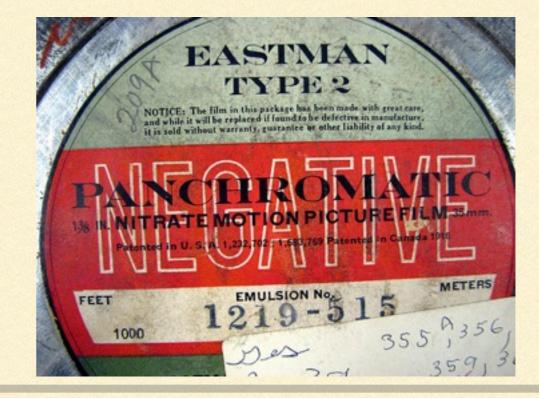
"PICTURE yourself as a historian in 2035, trying to make sense of this year's American election campaign. Many of the websites and blogs now abuzz with news and comment will have long since perished. Data

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

- 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"





- 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.



- Important, even historic data, is at serious risk.
- Digital Assets are dependent on file formats and hardware/ software platforms



Important, even historic data, is at serious risk.

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





At least 58 common containers / or

• At least 58 common containers/codecs

+ Audio tracks (27 formats/90 codecs)

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

Digital Assets are vulnerable to format obsolescence

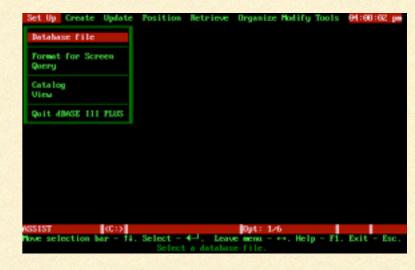
| CITEST  | P01 L07 C17 In   | sert Align   |                               |
|---|--|--|-------------------------------|
| ^E up ^¥ u<br>^X doun ^Z do<br>^S left ^R u<br>^D right ^C do | CROLL EPSTE<br>p C char<br>son T word<br>p screen Y line<br>son Del char<br>creen U uneras | T NENU<br>OTHER<br>J help<br>1 tab<br>2 turn insert off<br>2 B align paragraph<br>2 N split the line Es<br>2 N split the line Sa<br>2 L find/replace again | c shorthand                   |
| L!!!!   | ****-  | !!!!   | R                             |
|   | A centred ti   | tle  | 5                             |
| The quick brown for   | x jumped over the 1  | azy dog  | Ś                             |
| "Bone hold text"B   | ~~   |  |                               |
| Display Centre C.<br>Melp Blado Bl                            | hkäest Chkword Bel<br>sdrlin Beld Dell   | Elk HideElk NoveBlk Cop<br>ine DelVordrAliyn - Mul   | An under a state from a state |

Wordstar 1978-1985

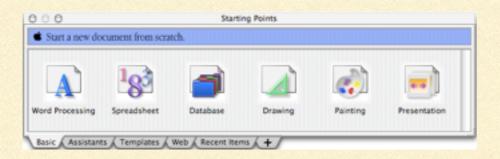
|      | l: 'EMP<br>sheet R | ange Copy          | Nove File    | Print G          | raph Data | System  | Quit      |
|------|--------------------|--------------------|--------------|------------------|-----------|---|-----------|
| Clob | al Inse            | rt Delete          | Column Eras  | e Title          | s Window  | Status  | Page Hide |
| Ĥ    | A                  |                    | C            |                  | E         | F   | 6         |
| 1    | EMP                | EMP_NAME           | DEPTND       | 30B              | YEARS     | SALARY  | BDNUS     |
| 2    | 1777               | Azibad             | 4808         | Sales            | 2         | 48888   | 10000     |
| 3    |                    | Brown              |              | Sales            |           | 45888   | 10000     |
| 4    |                    | Burns              | 6888         | fige<br>fige     | 4         | 75888   | 25080     |
| 5    | 58785              | Caeser             | 7808         | fign             | 3         | 75088<br>65090<br>65898<br>45000<br>158888<br>40000<br>38888<br>70000<br>35888<br>75000 | 25888     |
| 6    | 49692              | Curly<br>Dabarrett | 3808         | lige-            | 5         | 65888   | 28888     |
| 7    | 34791              | Dabarrett          | 7808         | Sales<br>Preside | 2         | 45888   | 10000     |
| 8    | 84984              | Daniels            | 1000         | Preside          | nt 8      | 150800  | 188888    |
| 9    | 59937              | Denosey            | 3808         | Sales            | 3         | 40800   | 10000     |
| 18   | 51515              | Donovan            | 3888         | Sales            | 2         | 38888   | 5888      |
| 11   |                    | Fields             |              | fign             | 5         | 70800   | 25080     |
| 12   |                    | Fiklore            |              | Admin            | 8         | 35888   |           |
| 13   | 64596              | Fine               | 5808         | fign             | 3         | 75000   | 25000     |
| 14   | 13729              | Green              | 1000         | Fign             |           | 98888   | 25888     |
| 15   | 55957              | Ternann            | 4808         | Sales            | 4         | 50000   | 10000     |
| 16   | 31619              | Hodgedon           | 5888         | Sales            | 2         | 4888B   | 10000     |
| 17   | 1773               | Foward             | 2888         | <b>Ngr</b>       | 3         | 88888   | 25888     |
| 18   | 2165               | Eugh               | 1888         | Admin            | 5         | 38888   |           |
| 19   | 23987              | Johnson            | 100B<br>100B | VP               | 1         | 100000  | 58888     |
| 28   |                    | Laflare            |              | Sales            |           | 35888   | 5080      |
| DATA | a.WK3              |                    |              |                  |           |   |           |

Lotus 1-2-3

1978-2013



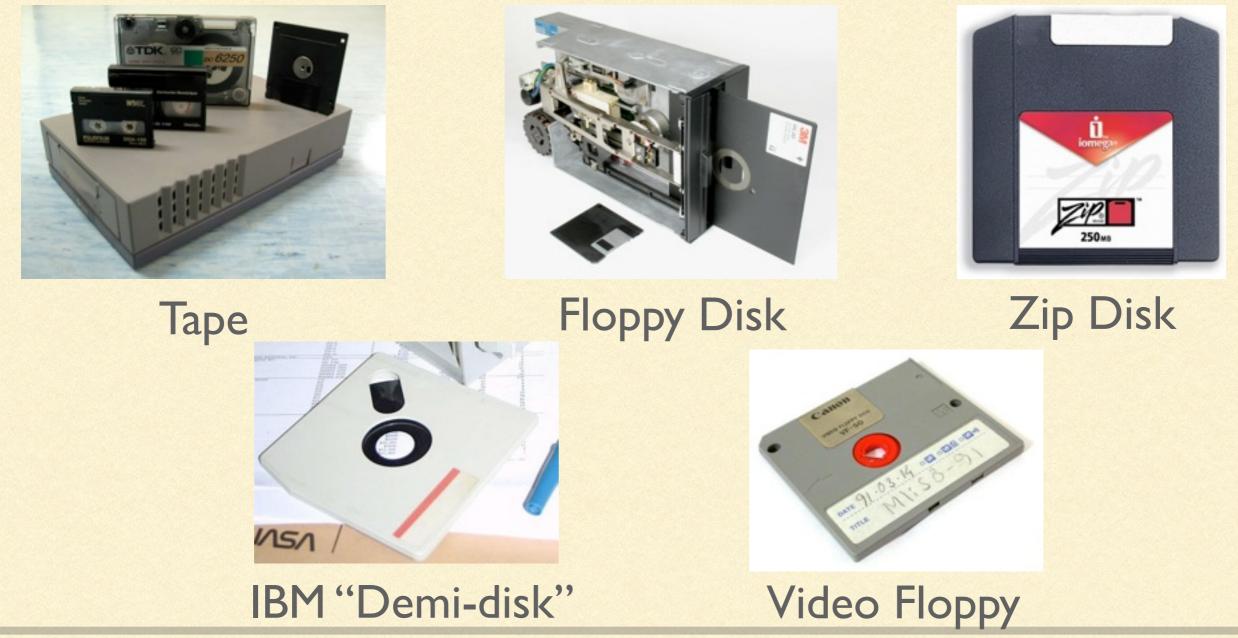
dBase 1978 - ?



Appleworks 1984-1991

Important, even historic data, is at serious risk.

Digital Assets are vulnerable to format obsolescence



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.

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

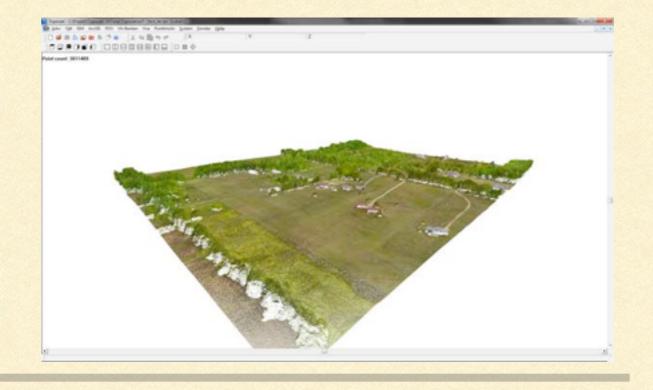
#### • Can be anything.

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

;LCB0 - 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] LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWCQMSFWGATVITNLFSAIPYIGTNLV EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG LLILILLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLCGVLALFLSIVIL GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX IENY



# ADDITIONAL CHALLENGES Datasets and File Formats

#### Most traditional file types are "known quantities."

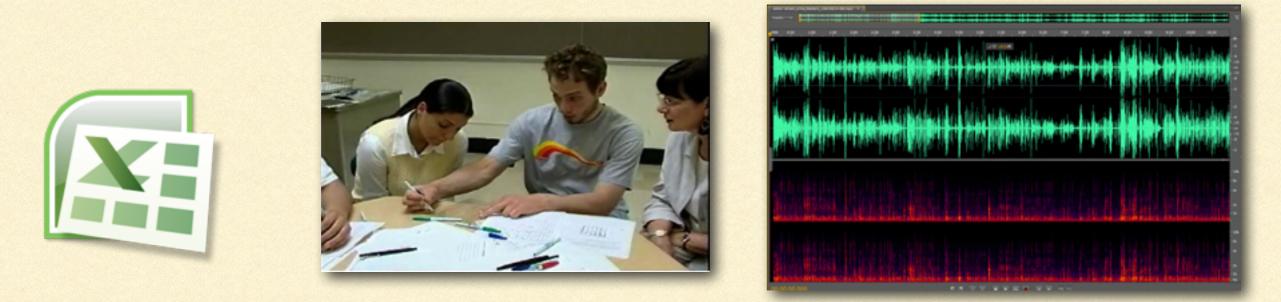
- Predictable use cases
- Rigid standards
- Built-in familiarity



# ADDITIONAL CHALLENGES

Datasets and File Formats

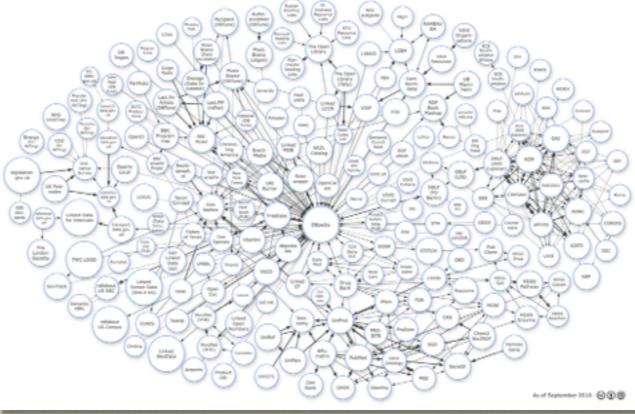
- Some datasets are stored merely as common file formats used for the purpose of data gathering, e.g.
  - MS Excel spreadsheets with data points, PDF files with written content.
  - Still images, sound, or moving images captured as part of research.



# ADDITIONAL CHALLENGES

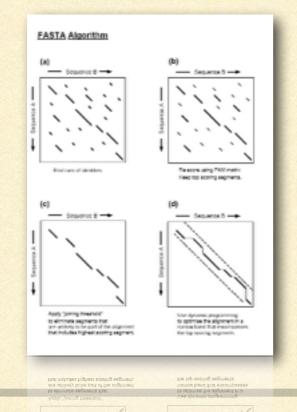
Datasets and File Formats

- 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.

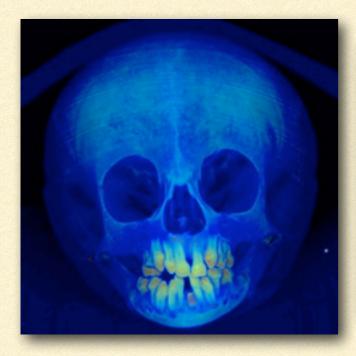




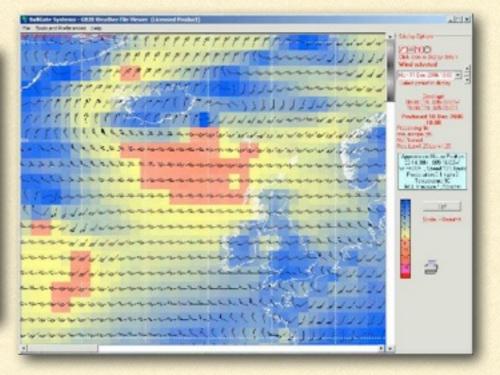
# ADDITIONAL CHALLENGES

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!



|    | E etter F | IF [] A 4] | 10 10 9 | 0.00   |               | ble: 6 of 6 Vari | abia |
|----|-----------|------------|---------|--------|---------------|------------------|------|
|    | Subject   | Treatment  | Age     | Gender | Before_exp_BP | Atter_exp_<br>BP |      |
| 1  | D1        | 1          | 65 F    |        | 103.30        | 80.50            | ŀ    |
| 2  | D2        | 1          | 59 F    |        | 93.60         | 85.90            | 1    |
| 3  | D3        | 1          | 601     | N      | 92.00         | 85.20            | 1    |
| -4 | D4        | 1          | 54 1    | F.     | 93.00         | 87.80            | 1    |
| 5  | 05        | 1          | 65 8    | E.     | 95.40         | 85.30            | 1    |
| 6  | D6        | 1          | 571     | M      | 109.60        | 94.20            | 1    |
| 7  | 07        | 1          | 691     | N      | 97.90         | 83.90            | 1    |
| 8  | D8        | 1          | 621     | N      | 96.00         | 85.00            |      |
|    | 4         |            |         |        |               |                  | •    |



# 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.



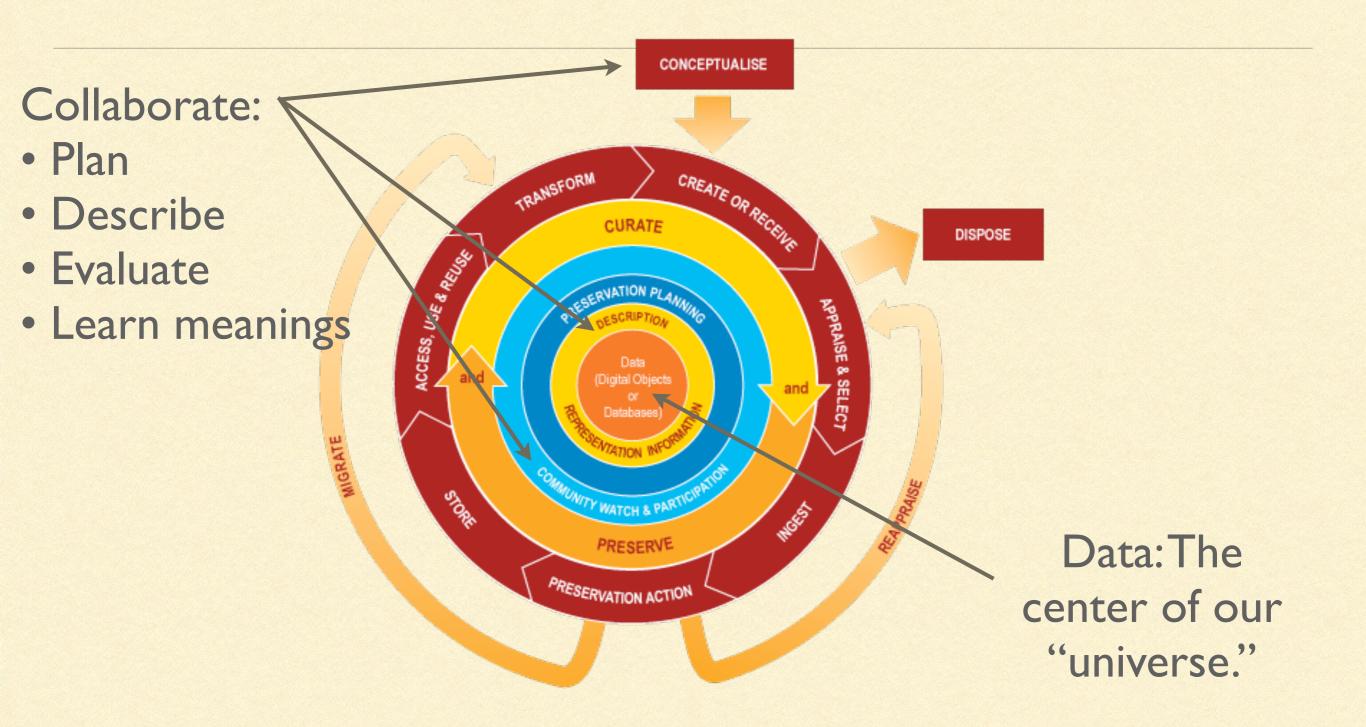
Filename: IMG\_2984.JPG Date Taken: ? Location: ? Context: ?

# NEXT STEPS

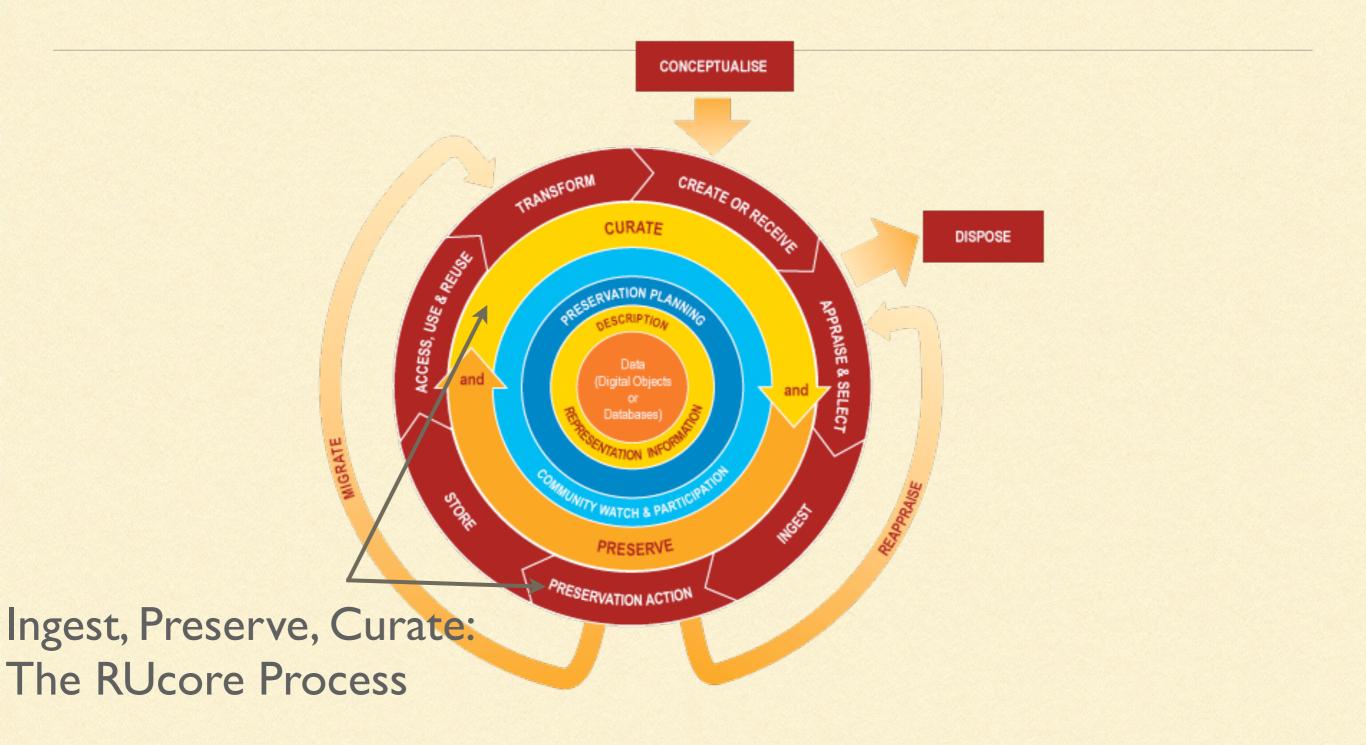
- We've identified and acquired data.
- We're aware that data is big and growing.
- We know the challenges of preserving data.

### Now what?

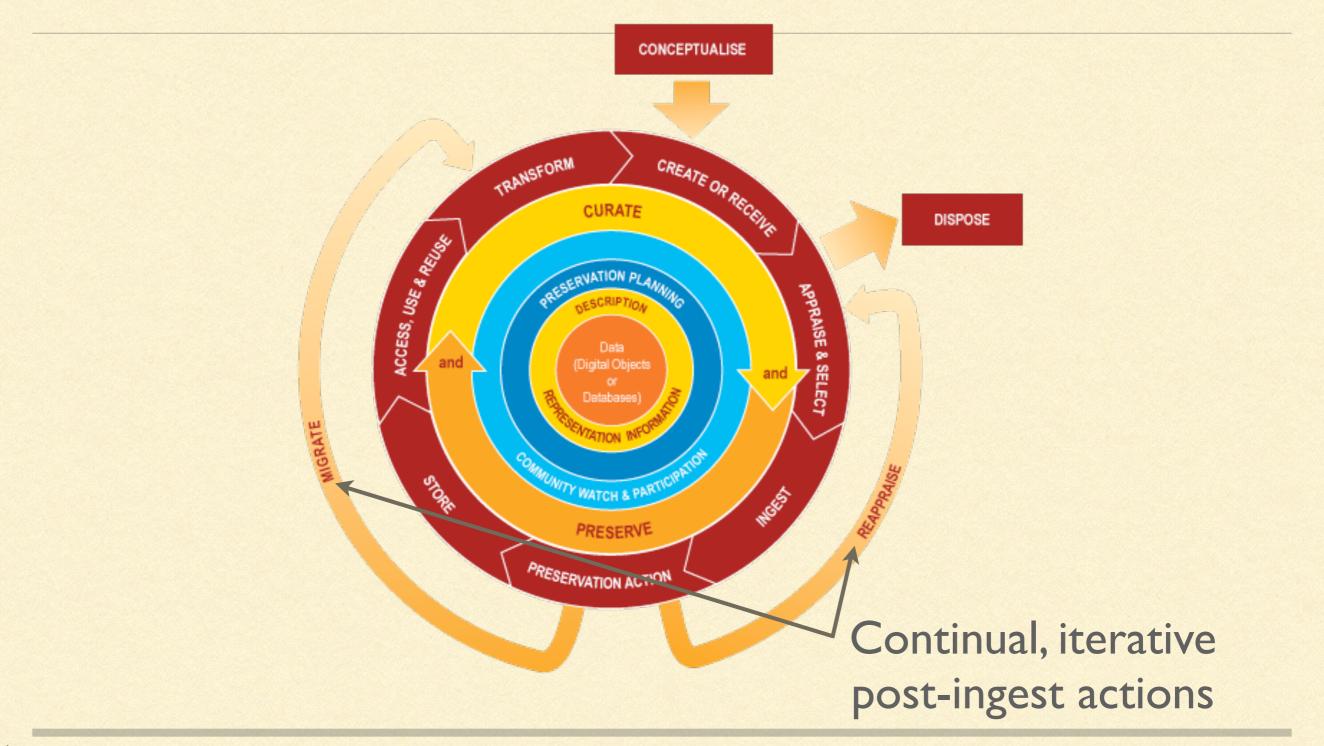
- A multitiered, continuous process where digital objects of any type are evaluated, preserved, maintained, verified, and reevaluated.
- Iterative: the cycle doesn't end with one go-round.
- A useful exercise for known and as-yet-unknown file types and formats.



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



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



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

- 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.

#### 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?

#### Do No Harm to digital assets

- Preservation masters, derivatives when needed
- Content modification must be done with extreme care
- Any changes must be traceable, auditable, reversible

- 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

# PARTINGTHOUGHTS

- 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.

# PARTINGTHOUGHTS

#### 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 <u>shared</u>. So, It's unlikely you're the only person to have encountered X data format. Someone else online may have the answers you seek.

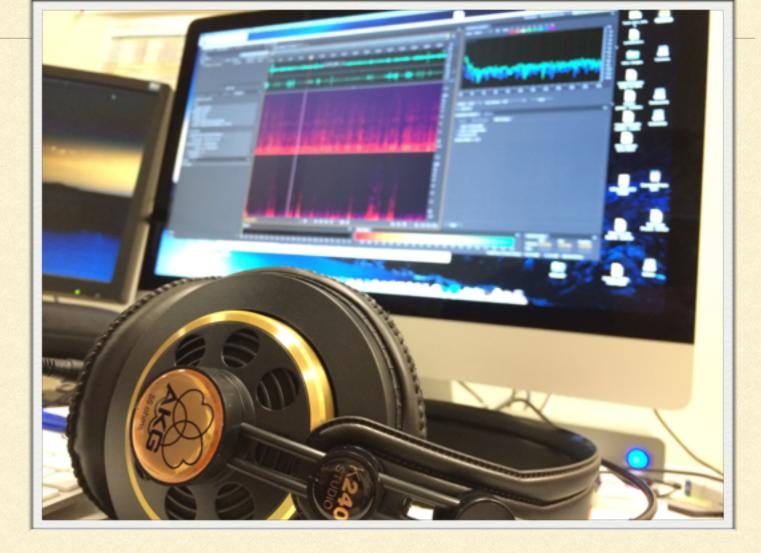
# PARTINGTHOUGHTS

#### 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:**

www.dcc.ac.uk

Digital Curation Center: Australian National Data Service www.ands.org.au

**Blog: From Page2Pixel** page2pixel.org