Data and Society

Lecture 5: Data and Entertainment

3/13/15
Announcements

• Section 2 Paper /Mini-proposal due April 3 (Details in syllabus on website)

• Office hours 1-2. Come talk to me if you have any questions about graded work, your paper / mini-proposal, etc.
Today (3/20/15)

• Lecture 5: Data and Entertainment

• L4 Data Roundtable (Kate, Charles, Philip, Alex, Robert)
<table>
<thead>
<tr>
<th>Section 1: The Data Ecosystem -- Fundamentals</th>
<th>January 30</th>
<th>Class introduction; Digital data in the 21st Century (L1)</th>
<th>Data Roundtable / Fran</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>February 6</td>
<td>Data Stewardship and Preservation (L2)</td>
<td>L1 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>February 13</td>
<td>Data and Computing (L3)</td>
<td>L2 Data Roundtable / 6 students</td>
</tr>
<tr>
<td></td>
<td>February 20</td>
<td>Colin Bodel, Time Inc. CTO Guest Lecture and Q&amp;A</td>
<td>L3 Data Roundtable / 5 students</td>
</tr>
<tr>
<td>Section 2: Data and Innovation – How has data transformed science and society?</td>
<td>February 27</td>
<td>Section 1 Exam</td>
<td>Data and the Health Sciences (L4)</td>
</tr>
<tr>
<td></td>
<td>March 6</td>
<td>Paper preparation / no class</td>
<td></td>
</tr>
<tr>
<td></td>
<td>March 13</td>
<td>Data and Entertainment (L5)</td>
<td>L4 Data Roundtable / 6 students</td>
</tr>
<tr>
<td></td>
<td>March 20</td>
<td>Big Data Applications (L6)</td>
<td>L5 Data Roundtable / 5 students</td>
</tr>
<tr>
<td>Section 3: Data and Community – Social infrastructure for a data-driven world</td>
<td>April 3</td>
<td>Data in the Global Landscape (L7) Section 2 paper due</td>
<td>L6 Data Roundtable / 6 students</td>
</tr>
<tr>
<td></td>
<td>April 10</td>
<td>Bulent Yener Guest Lecture, Data Privacy / Bad guys on the Internet (L8)</td>
<td>L7 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>April 17</td>
<td>Data and the Workforce (L9)</td>
<td>L8 Data Roundtable / 6 students</td>
</tr>
<tr>
<td></td>
<td>April 24</td>
<td>Mike Schroepfer, Facebook CTO Guest Lecture and Q&amp;A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>May 1</td>
<td>Data Futures (L10)</td>
<td>L9 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>May 8</td>
<td>Section 3 Exam</td>
<td>L10 Data Roundtable / 5 students</td>
</tr>
</tbody>
</table>
Lecture 5: Data and Entertainment
Data and Entertainment

• Digital Entertainment
• Digitally-enabled film
  – Avatar
    – Stewardship and preservation of digital film assets
• The Grateful Dead Archive Online
  – Stewardship and preservation of cultural assets
Digital Entertainment is Pervasive
Complex digital infrastructure, management needed to support digital entertainment

- **Multi-player games** supported by large communication-intensive systems

- **Digital TV and movies** must support real-time streaming video and audio

- **E-readers and music devices** must incorporate appropriate access and rights for content

- **Smartphone apps** must optimize efficiency of data, compute, networking infrastructure between device, network, cloud, local facilities
Data Stewardship and Preservation especially important as the Arts become more digitally-enabled

- Consumers must move (migrate) downloaded digital music to new media players when old players are too full, sometimes requiring re-registration of Digital Rights Management authorization to insure they do not lose access to favorite songs.

- **Authors** must find applications interoperable with old word processing SW to read manuscripts written with obsolete SW.

- Digital photos recorded on floppies can’t be accessed on modern computers without floppy disk drives.

- Old video games may only run on obsolete game systems.
Digitally-enabled Movies
Digital movies

• Most movies are not shot on film but recorded in bits and bytes through digital media
  – More than 80% of the movie theaters in the U.S. no longer handle film ...

• Many digital technologies used in film-making:
  – Image capture
  – Visual effects
  – Mastering and final color grading
  – Sound capture
  – Sound effects
  – Sound editing and mixing
  – Digital distribution to theaters and other platforms, etc.

• Film industry has been adopting digital technologies in piecemeal fashion over the last 25 years
Many components of movie process archived beyond the film itself

Motion-picture production has always required many different elements to create the film prints (and now digital files) that audiences see in domestic and international movie theaters. Shown here is a high-level view of these elements, color-coded according to longevity (yellow is for working files, gray for archived elements, and yellow with a gray border for working files that are also archived).

Avatar: Digital tour-de-force

- Film released in 2009 and distributed by 20th Century Fox
- Directed and written by James Cameron, produced by James Cameron and Jon Landau
- Became highest grossing film of all time (>2B)
- Won Academy Awards for Best Art Direction, Best Cinematography and Best Visual Effects
- 3 more Avatar movies forthcoming

Avatar image from Film Education
**Avatar both data-intensive and compute-intensive**

- Avatar technologies developed by Weta Digital Ltd and partners.
  - Weta Digital Ltd. Data Center in New Zealand
  - (Weta Digital also responsible for computer-rendered scenes in *Lord of the Rings* Trilogy, *King Kong*, etc.)

- Avatar IT equal parts of computing power in the data center (creating the visual effects) and data management of artistic processes (driving the film experience)

- Every minute of Avatar represents 17.28 GB of data (~ 3TB in all)

- Avatar used 1 PB of storage space for rendering
Avatar -- Innovative IT

- Technological innovations included:
  - **Performance capture process:** actors wore special gear and cameras that translated live action into realistic animation in real-time
  - **3D Fusion Camera:** 2 high def cameras in a single camera body to create depth perception
  - **Virtual camera system:** shows actors’ virtual counterparts in their digital surroundings in real time
  - **Motion capture stage**, etc.

Avatar image from Wikipedia article with caption “Cameron pioneered a specially designed camera built into a 6-inch boom that allowed the facial expressions of the actors to be captured and digitally recorded for the animators to use later.”

Avatar Technologies

http://www.youtube.com/watch?v=OJ1JzYPjcj0

(8:39)

CGI = Computer-generated imagery
Weta Digital IT Environment

• Computing core included 40K processors and 104TB of RAM
  – 10K square foot server farm with 34 racks of 32 HP Blade servers each
  – Center uses water-cooled racks and leverages chilly climate of New Zealand
  – Interconnected by 10 gigabit network so that storage seems local

• Data storage leveraged partnerships with NetApp and Fujitsu to develop storage system which
  – reduced the amount of manual data management in the process of rendering files
  – balanced the throughput requirements of the renderwall (compute) to maximize access to commonly used files

• Digital Asset Management System “Gaia” developed by Microsoft
Avatar (Weta Digital) computers occupied spots 193-197 on the Top500 List in November 2009

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>System</th>
<th>Cores</th>
<th>Rmax (TFlop/s)</th>
<th>Rpeak (TFlop/s)</th>
<th>Power (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>194</td>
<td>WETA Digital, New Zealand</td>
<td>Cluster Platform 3000 BL2x220, L54xx 2.5 Ghz, GigE Hewlett-Packard</td>
<td>5,936</td>
<td>31.5</td>
<td>59.4</td>
<td></td>
</tr>
<tr>
<td>195</td>
<td>WETA Digital, New Zealand</td>
<td>Cluster Platform 3000 BL2x220, L54xx 2.5 Ghz, GigE Hewlett-Packard</td>
<td>5,936</td>
<td>31.5</td>
<td>59.4</td>
<td></td>
</tr>
<tr>
<td>196</td>
<td>WETA Digital, New Zealand</td>
<td>Cluster Platform 3000 BL2x220, L54xx 2.5 Ghz, GigE Hewlett-Packard</td>
<td>5,936</td>
<td>31.5</td>
<td>59.4</td>
<td></td>
</tr>
<tr>
<td>197</td>
<td>WETA Digital, New Zealand</td>
<td>Cluster Platform 3000 BL2x220, L54xx 2.5 Ghz, GigE Hewlett-Packard</td>
<td>5,936</td>
<td>31.5</td>
<td>59.4</td>
<td></td>
</tr>
<tr>
<td>198</td>
<td>WETA Digital, New Zealand</td>
<td>Cluster Platform 3000 BL2x220, L54xx 2.5 Ghz, GigE Hewlett-Packard</td>
<td>5,936</td>
<td>31.5</td>
<td>59.4</td>
<td></td>
</tr>
<tr>
<td>276</td>
<td>WETA Digital, New Zealand</td>
<td>Cluster Platform 3000 BL2x220, L54xx 2.5 Ghz, GigE Hewlett-Packard</td>
<td>5,120</td>
<td>27.3</td>
<td>51.2</td>
<td></td>
</tr>
<tr>
<td>789</td>
<td>National Institute for Water and Atmospheric Research, New Zealand</td>
<td>Power 575, p6 4.7 GHz, Infiniband IBM</td>
<td>1,792</td>
<td>26.4</td>
<td>33.7</td>
<td>286</td>
</tr>
<tr>
<td>451</td>
<td>WETA Digital, New Zealand</td>
<td>Cluster Platform 3000 BL2x220, L54xx 2.5 Ghz, GigE Hewlett-Packard</td>
<td>4,096</td>
<td>21.9</td>
<td>41.0</td>
<td></td>
</tr>
</tbody>
</table>
Avatar Digital Asset Management (DAM) System

- **Gaia (developed by Microsoft) provided sophisticated digital management resources and workflow**
  - Central repository used to capture, store, retrieve digital information (TBs) from multiple locations (LA, New Zealand, etc.)
  - Highly customized SQL queries used to balance workload between client and server
  - Special attention paid to security of assets to prevent leaks
  - “Stage App” managed digital workflow for all phases of the shoot:
    - Collected new content from relevant workstations and processed it in the background
    - Ingested into Gaia system
    - Tracked universal time code and attached appropriate metadata
    - Made available for rendering and editing

- **Gaia took 5 years to develop by 6 people**

Fran Berman, Data and Society, CSCI 4967/6963
What happens next?

• What happens to digital films such as Avatar after their release?

• How are they and their associate materials preserved?
  – How long?
  – What is preserved?
  – Who makes the decision?
  – Is it important to keep them?
  – What is the business model?
Goals for Film archiving

• 100+ long-term access (and thus preservation) with no loss of quality
• Ability to create duplicate masters to fulfill future (and unknown) distribution needs and opportunities
• Picture and sound quality that meets or exceeds that of original camera negative and production sound recording
• Independence from shifting technological platforms
• Interoperability
• Immunity from escalating financial investment
• Current practice for many studios in Hollywood is “Save Everything” ...
• Contrast to ESPN (TV) which only saves that which cannot be reconstructed

Term usage:
• Library = collection of films
• Archive = preserved films
Pre-digital archiving

• Earliest movies lost because long-term preservation of motion pictures was not considered important – culturally or commercially

• Many titles in early film libraries on nitrate stock destroyed by fire or trashed

• Some film masters turned to vinegar in hot, humid warehouses until climate control requirements for preservation well-understood

• Fewer than half of the feature films made before 1950 have survived; less than 20% survive from the 1920’s
Digital film preservation economics

• Most movies now shot with digital technologies.
• Digital movies costly to preserve.
  – 8.3 TB digital master costs $12K/year to preserve.
  – Currently studios tend to save everything – finished movies, original camera negative film, original audio recordings, still photographs, etc. Single movie can generate 2+ PB of data.
  – ~7000 new feature films released each year.

• Economic incentives for studios to preserve valuable assets
• Major Hollywood studios investing heavily in archiving facilities
Digital Motion Picture Archive Economics (circa 2007) from The Digital Dilemma

Annual Storage Costs of Motion Picture Materials

<table>
<thead>
<tr>
<th>Source Material</th>
<th>Archival Master</th>
<th>Annual Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film Capture, 4K Master</td>
<td>$1,830</td>
<td>$12,514</td>
</tr>
<tr>
<td>Digital Capture to HDCAM SR Tape, 1920 x 1080 Master</td>
<td>$1,955</td>
<td>$12,514</td>
</tr>
<tr>
<td>Digital Capture to 2K Data, 2K Master</td>
<td>$486</td>
<td>$486</td>
</tr>
<tr>
<td>Digital Capture to 4K Data, 4K Master</td>
<td>$180</td>
<td>$208,569</td>
</tr>
</tbody>
</table>

Diagram is not to scale.
Archiving can be profitable

- Can broaden viewership for archived films through TV, home video distribution, foreign language versions, airplane versions, etc.
- Can get extra value by packing archived films with scenes, bloopers, out-takes, etc.
- Studios have begun “asset protection” programs to ensure the long-term shelf-life (and basis for potential revenue) of film collections

Archiving can be expensive

Paramount Pictures

- spent $35+M in 1987-1993 inspecting negatives, audio tracks and black and white separations, doing film repairs, printing new preservation materials
- created a new $11M archives building with environmentally controlled vaults for preprints and color materials.
- stores some master elements at an underground facility in PA
- tracks millions of items in a computerized inventory and tracking systems
Digital film preservation technical challenges

• Digital information stored in master copy of movie represents the movie only indirectly
  – Need to be able to simulate / emulate **HW and SW environment** (including OS, digital image and sound file formats, imaging and sound applications SW) as well

• **Archiving by data migration time-intensive**: Moving a PB of data over GB ethernet connection takes 3+ months.

• **Digital movies need active preservation** (error checking, alignment of multiple copies, sound economic model, etc.) like other valuable digital content
Digital image standards efforts ongoing

• **ACES (Academy Color Encoding System)** created by Motion Picture Academy as standard for digital image files.
  – Standard replacing DPX (Digital Picture Exchange) file format that leaves many technical details up to the end user, leading to many different interpretations and implementations

• ACES bringing increased clarity of interpretation of digital files during later steps in film-making process including visual effects, postproduction, mastering.
Content Lifetime
(from the Digital Dilemma, circa 2007)
Storage and Distribution: Film Archives and Libraries

• “Dark” long-term archive – cold-temperature underground vaults
  - Accessed only when good quality film print master elements cannot be found locally
  - Function as “insurance policy” for valuable digital assets
  - Typically full pixel count, full bit-depth, uncompressed and unencrypted
  - Changing culture: traditional film archiving can no longer preserve all forms of outputs of increasingly use of digital technologies in film making

• Short-term film distribution libraries
  - Contain release prints and other finish elements needed to meet commercial distribution requirements
  - Accessed frequently, actively managed during primary commercial window (3-5 years typically)
  - Typically formatted at lower pixel count, lower bit-depth and compressed
Challenges for long-term data preservation – many things can go wrong ...

Challenges for preserving movies 100+ years

- Lifetime of physical media (<30 years)
- Vulnerability to heat, humidity, static electricity, electromagnetic fields
- Degradation by natural errors, corruption by processing or communication errors
- Malicious viruses
- Incomplete labeling and metadata
- Evolutionary divergence between data and relevant SW applications
- Obsolete standards
- Human error
- Etc.

### Digital Archive Layers

<table>
<thead>
<tr>
<th>Lifespan</th>
<th>Hardware</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-5 years</td>
<td>Host Computer</td>
<td>• Application SW</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Operating System</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Device Drivers</td>
</tr>
<tr>
<td>5-10+ years</td>
<td>Physical Interface</td>
<td>• Interface Firmware</td>
</tr>
<tr>
<td>3-5 years</td>
<td>Media Drive</td>
<td>• Drive Control Firmware</td>
</tr>
<tr>
<td>5-10 years</td>
<td>Media</td>
<td>• File System</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data File Format</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Physical Recording Format</td>
</tr>
<tr>
<td>Varies</td>
<td>Trained personnel</td>
<td></td>
</tr>
<tr>
<td>Varies</td>
<td>Funding</td>
<td></td>
</tr>
</tbody>
</table>

Graph from The Digital Dilemma, AMPAS
Great reads …

**IEEE SPECTRUM**

*Will Today’s Digital Movies Exist in 100 Years?*

We need new storage technologies, standards, and practices to preserve modern cinema

By Andy Maltz

Posted 21 Feb 2014 | 21:15 GMT

---

Fran Berman, Data and Society, CSCI 4967/6963
The Grateful Dead Archive Online
The Grateful Dead Archive Online (GDAO)
http://www.gdao.org/

- The Grateful Dead Archive Online documents the band’s activities and influence on contemporary music, Deadhead community and fans, and the cultural milieu that arose around the Grateful Dead Community.

- Donation of 600 linear feet of boxed materials including paper documents, books, photographs, graphics, film, video, recordings and textiles, original archives by fans, etc. began the archive.

- GDAO represents one of the most significant popular culture archives in the 20th century.
Who studies the Dead?

- Grateful Dead a focus of academic research from broad spectrum of disciplines
  - Ethnomusicology
  - Audio and sound
  - Drugs and pharmaceuticals (recreational drugs)
  - Social science (“lifestyle enclave” of Deadhead community)
  - History
  - Philosophy and Religion
  - Psychology
  - Art
  - Management and Business (wrt “customer service”, taping, etc.)
  - Law (discrimination against Deadheads by the system)
  - Popular culture (Tom Wolfe’s “The Electric Kool Aid Acid Test” ...)

Photo from the Grateful Dead Archive Online
Theses using materials from the Grateful Dead Archive

- *Deadheads as a Moral Community*, Ph.D. Dissertation, Deborah Baiano Berman, Northeastern University
- *Media Framing as Brand Positioning: Analysis of coverage linking Phish to the Grateful Dead*, Ph.D. Dissertation, Jordan McClain, Temple University
- “Never Could Read No Road Map”: *Geographic Perspectives on the Grateful Dead*, MS Thesis, Daniel Culli, Louisiana State University
Brief History of GDAO

- Archive gifted by the band with the stipulation that contents be digitized to broaden access
  - Library would seek grant funding for digitization (IMLS stepped up)
  - University would provide archivist

- Donated to UC Santa Cruz (UCSC). Archive contains 40 years worth of commercial recordings and videotapes, press clippings, stage sets, business records and correspondence

- Over 115K unique users since 2012, around 1K unique users per month

- Unusually engaged user community.
  - Fans donate funding and materials
  - Curation is crowd-sourced:
    - UCSC library post materials to community to help curate material and build up record
    - Users review and flag new content for appropriateness
Partnership

• UCSC received support in the form of “knowledge, collections and services” to create Archive
  – California Digital Library
  – Institute of Museum and Library Sciences
  – Cosmic Design Group
  – Internet Archive
  – Omeka
  – Donors

• Major costs are staffing. UCSC department copies and Internet Archive provide backup during heavy traffic
GDAO Cost Model

• Startup costs were $615K from IMLS grant for digitization.

• UCSC provided 1.6M in cost sharing for
  – Scanning equipment
  – Full-time archivist
  – Programmer to develop the website
  – Project management staff
  – Consultants on copyright,
  – Metadata specialists, etc.

• Post-start-up:
  – Endowment being raised for archivist’s position
  – Additional funding needed for web development, project management, rights management materials expertise, digitization of new materials
Sustainability

- Currently UCSC Library supporting GDAO
- Sustainability plan is to
  - Fundraise to create a $1M endowment to support archivist position and archive
  - Continue to apply for grants
  - Solicit on-line donations from Grateful Dead fan base and through the GDAO website
  - Solicit large donations from the community
  - Leverage UCSC fund-raising efforts
  - Crowd-fund for specific GDA projects
Lecture 5 Materials (not already on slides)

- Lecture Materials (links on website)
  - Data Plays a Supporting Role in Avatar, ComputerWorld [http://www.computerworld.com/s/article/346361/Data_center_plays_supporting_role_in_i_Avatar_i](http://www.computerworld.com/s/article/346361/Data_center_plays_supporting_role_in_i_Avatar_i)
  - Theses from the Grateful Dead Archive [http://lgdata.s3_website-us-east-1.amazonaws.com/docs/2740/592738/diss_checklist.pdf](http://lgdata.s3_website-us-east-1.amazonaws.com/docs/2740/592738/diss_checklist.pdf)
Data Roundtable
April 3: Big Data (L6) Data Roundtable


- “How pro teams are using data analytics to draft better players,” Financial Post, http://business.financialpost.com/2013/09/03/pro-sports-teams-turning-to-data-analytics-to-fill-seats/?__lsa=88c9-3dab (Dennis Fogerty)

Next week: L5 (Data and Entertainment) Data Roundtable

• “Management Secrets of the Grateful Dead”, The Atlantic

• “The Shazam Effect”, The Atlantic,
  http://www.theatlantic.com/magazine/archive/2014/12/the-shazam-effect/382237/?single_page=true (Sumit Munshi)

• “At Disney Parks, a Bracelet Meant to Build Loyalty (and Sales)”, The New York Times,

• “Dancing Data”, re/code,
  http://recode.net/2014/01/28/dancing-data/ (Miguel Inoa-Lantigua)

• “Here’s How Piracy Hurts Indie Film”, Indiewire,
Today: L4 Data Roundtable


