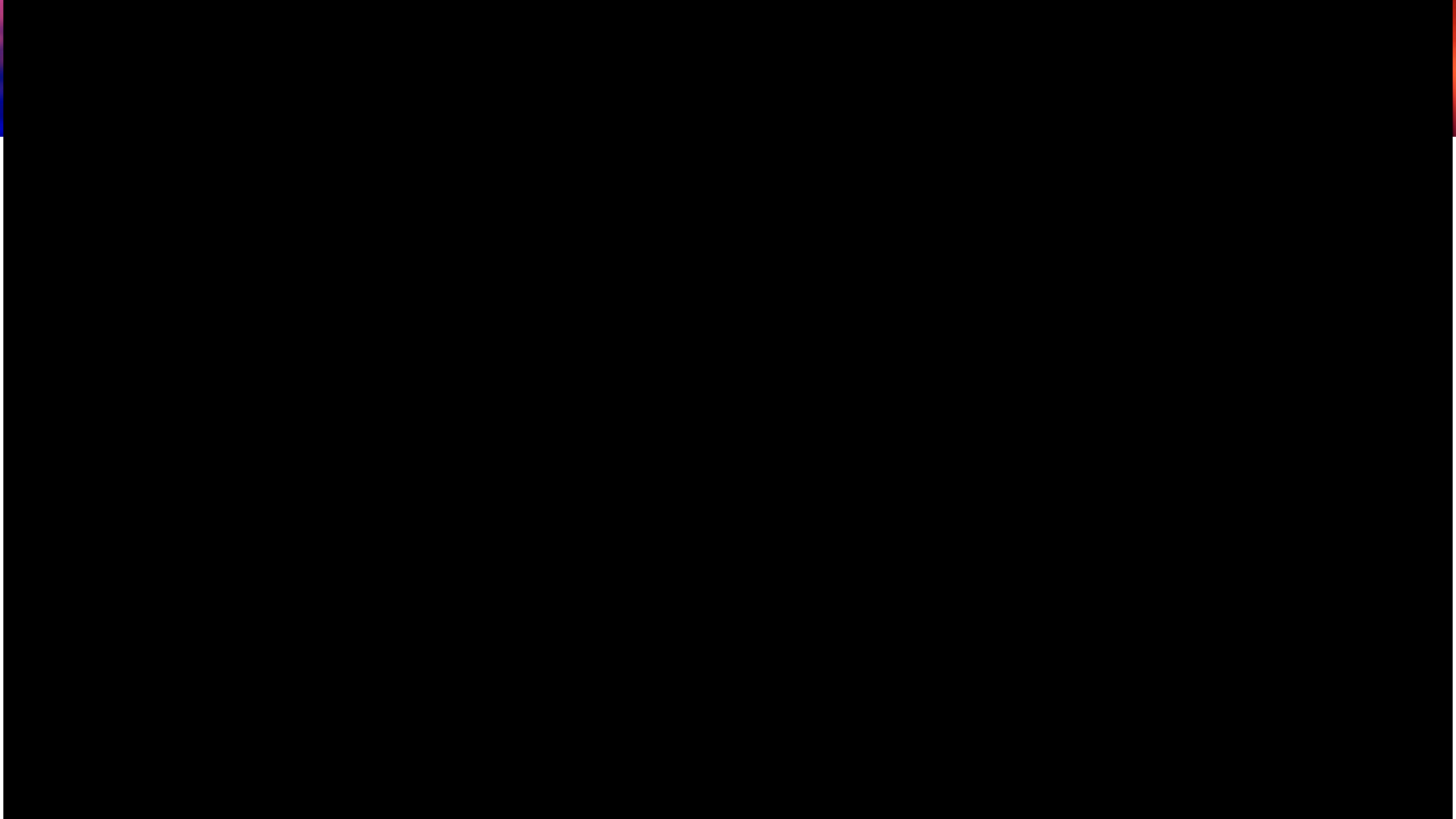


# HPA

## TECH RETREAT 2024

**The Death of Analog:  
Why Right Now Is the Time  
to Digitize Your Archive**

Carin Forman, Amazon  
Andrea Kalas, Paramount  
Heidi Shakespeare, Memnon  
Linda Tadic, Digital Bedrock



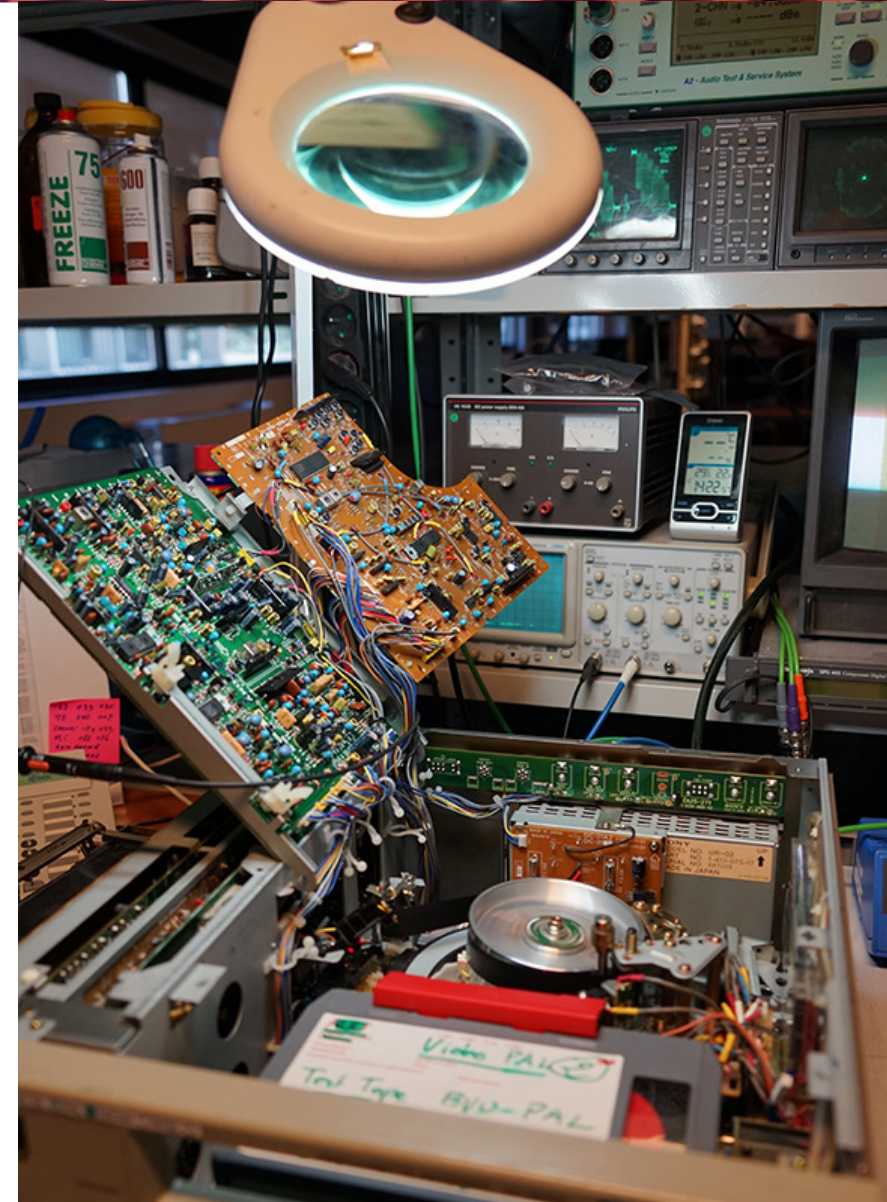
## Supply & Demand

- In 20 years, we have digitized 4 million hours of content
- In the past year alone, we've quoted for over **7 million hours** – close to double what we've achieved in two decades
- At maximum throughput in our largest facility, that would equate to almost **30 years** of 24/7 automated digitization



## Technology Obsolescence

- No playback device = No preserved file
- Legacy equipment no longer manufactured, so the pool of machines is dwindling
- Already most surviving VTRs operate on refurbished heads
- Spares taken from donor machines, reducing the equipment pool further
- Buying and maintaining obsolete technology will become unaffordable



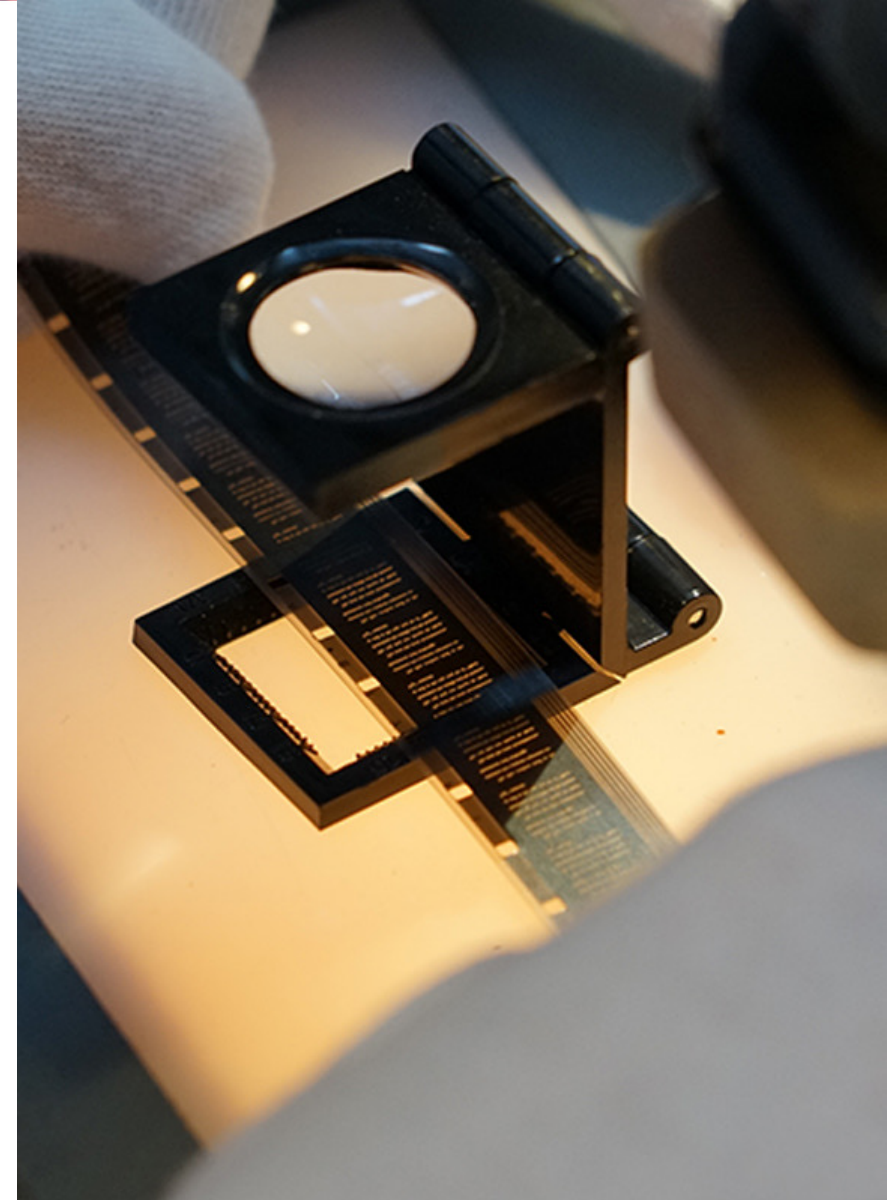
## Format Obsolescence

- All legacy formats have an expiration date for retrieving content
- Physical media degrades with time
- Content can become corrupted, or can be destroyed completely
- Access can be lost without specialized hardware



## Knowledge Obsolescence

- Media migration is a niche service and the skills gap is widening
- Some technology dates back as far as the 1980s
- Engineers with the requisite skills are nearing retirement and attracting younger generations is a challenge
- Urgency to ensure key skills aren't lost for good



## Format Difficulty Ranking

### DIFFICULTY 1 MEDIUM

TYPE	FORMAT	CRITERIA
VIDEO	DVCAM	4 5
VIDEO	DVCPRO	4 5
VIDEO	VHS	5
VIDEO	HDCAM SR	4
VIDEO	XDCAM	5
VIDEO	HDV	4 5
VIDEO	V8/Hi8	5
VIDEO	DBC Gen 1*	4
AUDIO	1/4"	1 5
AUDIO	DA88	1 2 4
AUDIO	Compact Cassettes	1

### DIFFICULTY 2 HIGH

TYPE	FORMAT	CRITERIA
VIDEO	U-Matic	1 5
VIDEO	Betacam SP**	1 3 5
VIDEO	D3	1 3 5
VIDEO	D5	1 2 5
VIDEO	DBC Gen 2*	3
VIDEO	DBC Gen 3*	2 3
VIDEO	1"C	4
AUDIO	DASH	1 2
AUDIO	DAT	1 5

### DIFFICULTY 3 VERY HIGH

TYPE	FORMAT	CRITERIA
VIDEO	D9	1 3
VIDEO	1"B	1 3 5
VIDEO	D1	1 2 3 4
VIDEO	D2	1 2 3 4
VIDEO	Betamax	1
AUDIO	1630	1 2 5
AUDIO	2"	1 2
AUDIO	1"	1 2

### DIFFICULTY 4 CRITICAL

TYPE	FORMAT	CRITERIA
VIDEO	2"	1 2 4 5
VIDEO	EIAJ	1 2 3 5
VIDEO	DCT	1 2 3 4 5
VIDEO	MII	1 2 3 5
VIDEO	CV2000	1 3 5

### CRITERIA

- 1 Lack of Machines  
 2 Cost of Machines  
 3 Lack of Heads  
 4 Cost of Heads  
 5 Playback Difficulty / Format Complexity

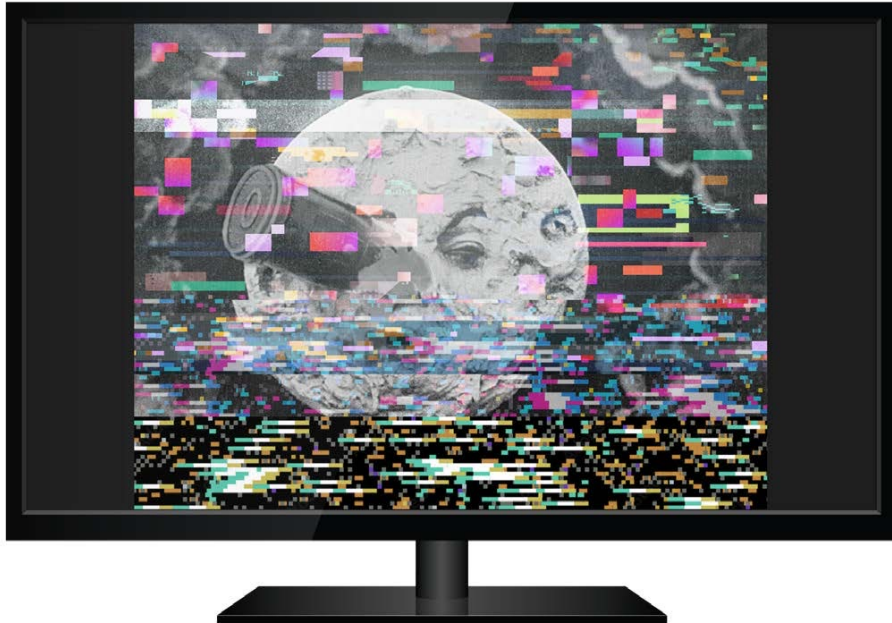
## Your digital archive consists of:



- Digital manifestations created by digitizing analog media
- Born-digital content

**Just because it's digital, doesn't mean  
it's preserved.**

## The Silent Fire



*Image copyright Digital Bedrock 2023*

- Data **storage carriers** can be obsolete or fail
- File **backup (writing) software** can be proprietary and/or obsolete
- **Storage file system** on the carrier can be proprietary or obsolete
- **File formats or codecs** themselves can be obsolete (require obsolete software, OS, hardware to render)
- **Human and machine errors:** the data wasn't written correctly so is corrupted

## Digital preservation requires analysis and appraisal.

- Do you know what is on these older storage carriers? Can you get the files off? Are the files truly what you think they are?
- Are the files OK? (checksums, scheduled fixity checks)
- Do you need to preserve everything? Can you perform selection/appraisal first?
- What content should be stored in the cloud, and what should be stored offline?

## Digital preservation requires ongoing care and management

There is no “store and ignore” medium.

**Migration is necessary to keep your digital content alive.**

**Avoid it becoming a museum object (or, the equivalent of a pet rock).**

Managed digital preservation involves:

- planning
- policies
- processes

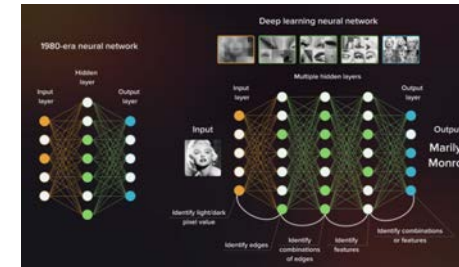
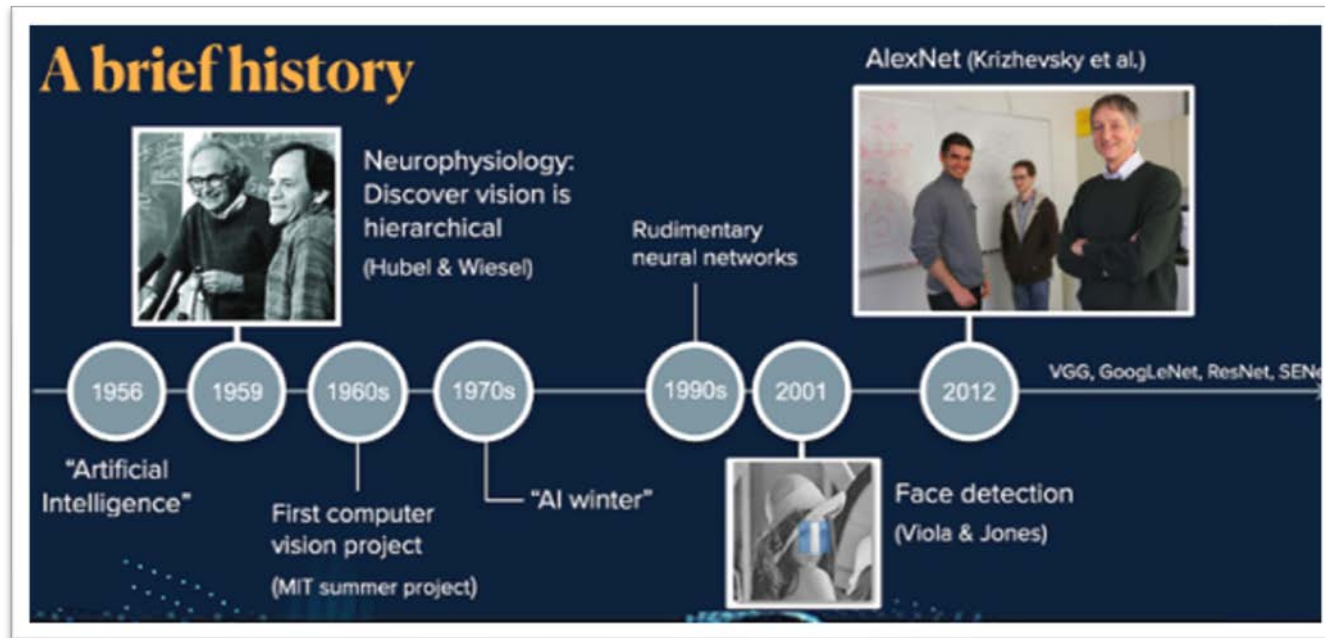


*Apple “Lisa 2” (1984). On display at the Musée des Arts et Metiers, Paris.*



# Since the invention of AI, brilliant minds have been working on how to identify and find stuff...

Today there are tools on the market that have benefitted from decades of research and learning –

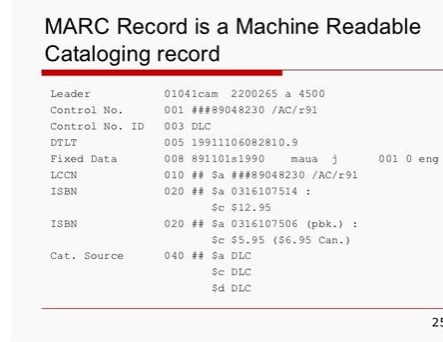


- range vs. kNN queries
    - range query** appropriate when
      - end-user is able to specify  $r$ , i.e., knows the semantics of the model
      - e.g., edit distance on strings, counting the smallest number of character edits to transform  $s_1$  into  $s_2$
      - range query ('drier', 2) = {driver, d\_iver, \_river, drive\_}
      - 100% recall is guaranteed (because of user's confidence on  $r$ )
    - kNN query** appropriate when
      - user cannot specify  $r$ , i.e., does not know the semantics of the model
      - majority of cases
- kNN,  $q =$

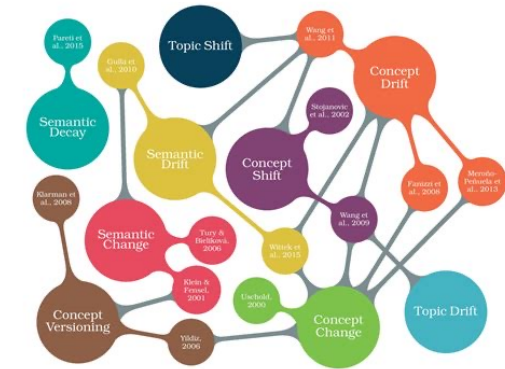
Token String	Token ID	Embedded Token Vector
'<s>'	0	[ 0.1150, -0.1438, 0.0555, ... ]
'<pad>'	1	[ 0.1149, -0.1438, 0.0547, ... ]
'</s>'	2	[ 0.0010, -0.0922, 0.1025, ... ]
'<unk>'	3	[ 0.1149, -0.1439, 0.0548, ... ]
'.'	4	[ -0.0651, -0.0622, -0.0002, ... ]
'the'	5	[ -0.0340, 0.0068, -0.0844, ... ]
','	6	[ 0.0483, -0.0214, -0.0927, ... ]
'to'	7	[ -0.0439, 0.0201, 0.0189, ... ]
'and'	8	[ 0.0523, -0.0208, -0.0254, ... ]
'of'	9	[ -0.0732, 0.0070, -0.0286, ... ]
'a'	10	[ -0.0194, 0.0302, -0.0838, ... ]

- Pattern Recognition
- Convolutional Neural Networks (CNNs)
- Deep Learning
- Multi-Modal Models
- Transformers
- Vector Embeddings

# Library science has accompanied this “rocket science”



<b>Property: Language-based Identifier (Preferred Lexical Label)</b>	
URI:	<a href="http://www.w3.org/2004/02/skos/core#prefLabel">http://www.w3.org/2004/02/skos/core#prefLabel</a>
Label:	language-based identifier (preferred lexical label)
Domain:	rdfs:Resource
Subproperty of:	rdfs:label
Cardinality:	OWL 3.4.2: A resource has exactly one unique language-based identifier (lexical label) per BCP 47 language tag
<p>Note: follow SKOS prefLabel pattern in allowing one (and only one) prefLabel per BCP 47 language tag. Question: How can OWL cardinality rule be adapted to apply to prefLabel plus BCP 47 language tag?</p>	
<b>Property: Variant Language-based Identifier (Alternate Lexical Label)</b>	
URI:	<a href="http://www.w3.org/2004/02/skos/core#altLabel">http://www.w3.org/2004/02/skos/core#altLabel</a>
Label:	variant language-based identifier (alternate lexical label)
Domain:	rdfs:Resource
Subproperty of:	rdfs:label
<p>Note: follow SKOS altLabel pattern in adding BCP 47 language tags to each variant language-based identifier</p>	
<b>Property: Language-based Identifier (Preferred Lexical Label) for</b>	



Martha Yee:  
Computational  
Cataloging  
Pioneer

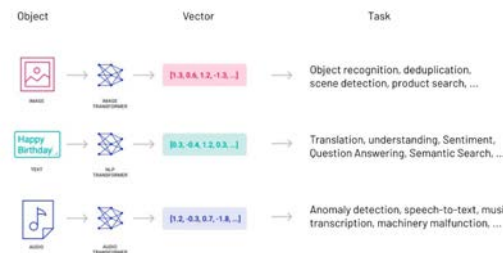
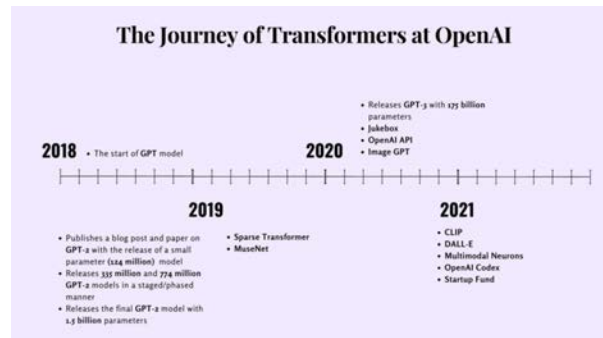


Since the mid-1980s, Library Science as well have worked towards a goal of automating the technology of discovery. From converting card catalogs to online searches; to being the first to use structured data for discovery; to mapping cataloging rules to semantic search technologies; the major use case: Accessibility

## Content Discovery has forever changed

★ OpenAI CLIP - is a multimodal AI model that combines knowledge of English-language concepts with semantic knowledge of images.

★ Open AI Clip uses vectors to power Video Semantic Search, Deciphering Corrupted Images, Image Captioning, Image Classification, Image Similarity, Image Ranking



**After half a century of research and innovation:**

**VIDEO CAN CATALOG ITSELF**




# MTV Music awards needs historic clips for current awards show

Example search understands logos, music stars, awards:


**MTV moon man**  
Min. Score: 0.27  
Found 10 results

#1




Time: 111.50 Frame: 3347

#2



Time: 745.92 Frame: 22377


#3



Time: 260.99 Frame: 7029

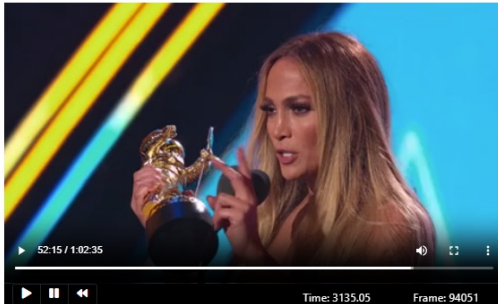
**jennifer lopez accepting an award**  
Min. Score: 0.27  
Found 10 results

#1



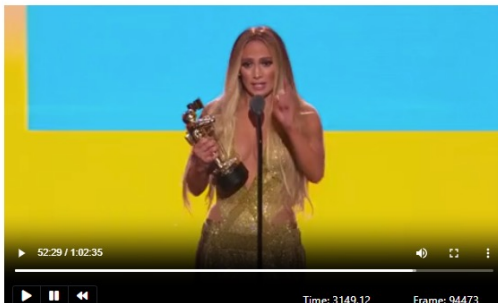
Time: 1582.22 Frame: 47466

#2



Time: 3135.05 Frame: 94051


#3



Time: 3149.12 Frame: 94473


**Foo Fighters performing on stage**  
Min. Score: 0.27  
Found 10 results

#1




Time: 4798.92 Frame: 143967

#2



Time: 4717.75 Frame: 141532


#3



Time: 4880.55 Frame: 146416


**Madonna**  
Min. Score: 0.27  
Found 10 results

#1




Time: 3344.39 Frame: 100331

#2



Time: 119.42 Frame: 3582

#3

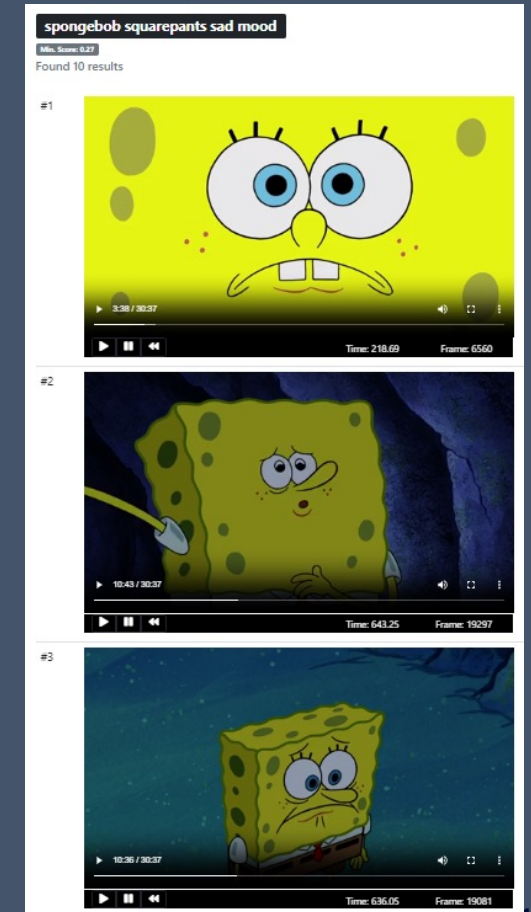
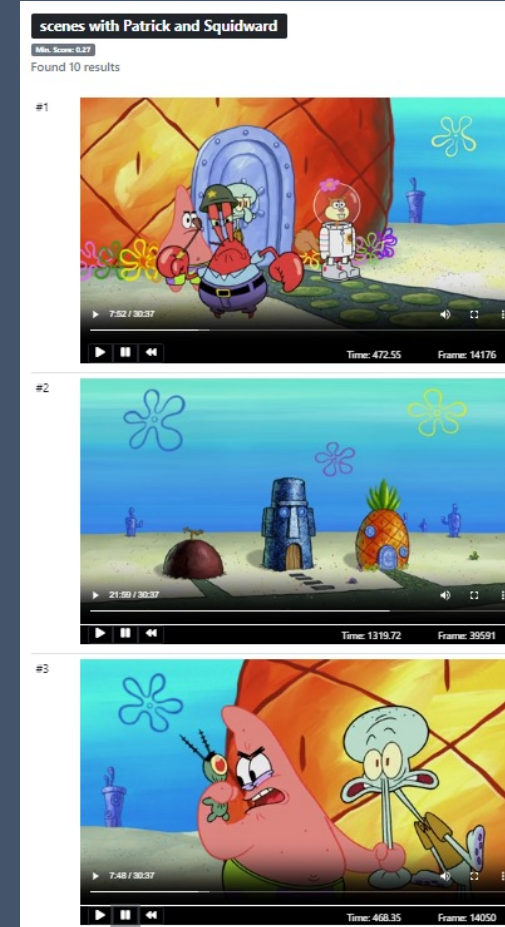
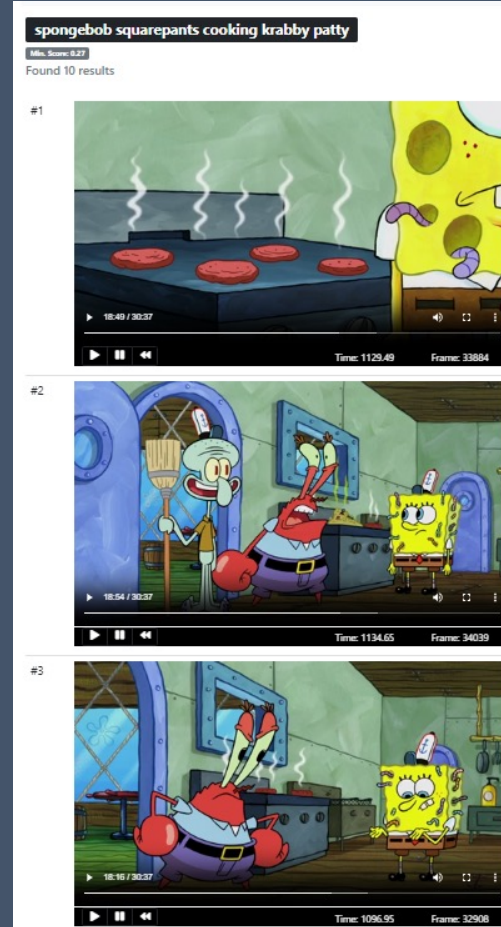
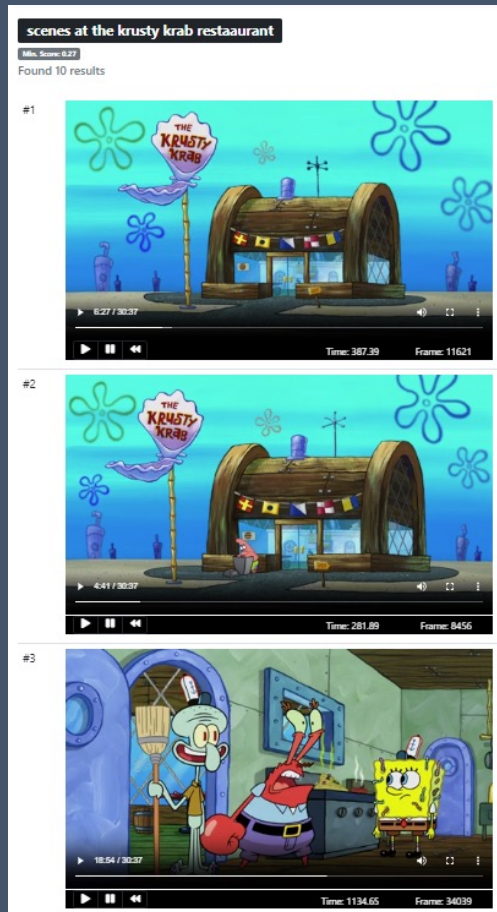


Time: 1560.79 Frame: 46823



# Nickelodeon creates themed social campaigns

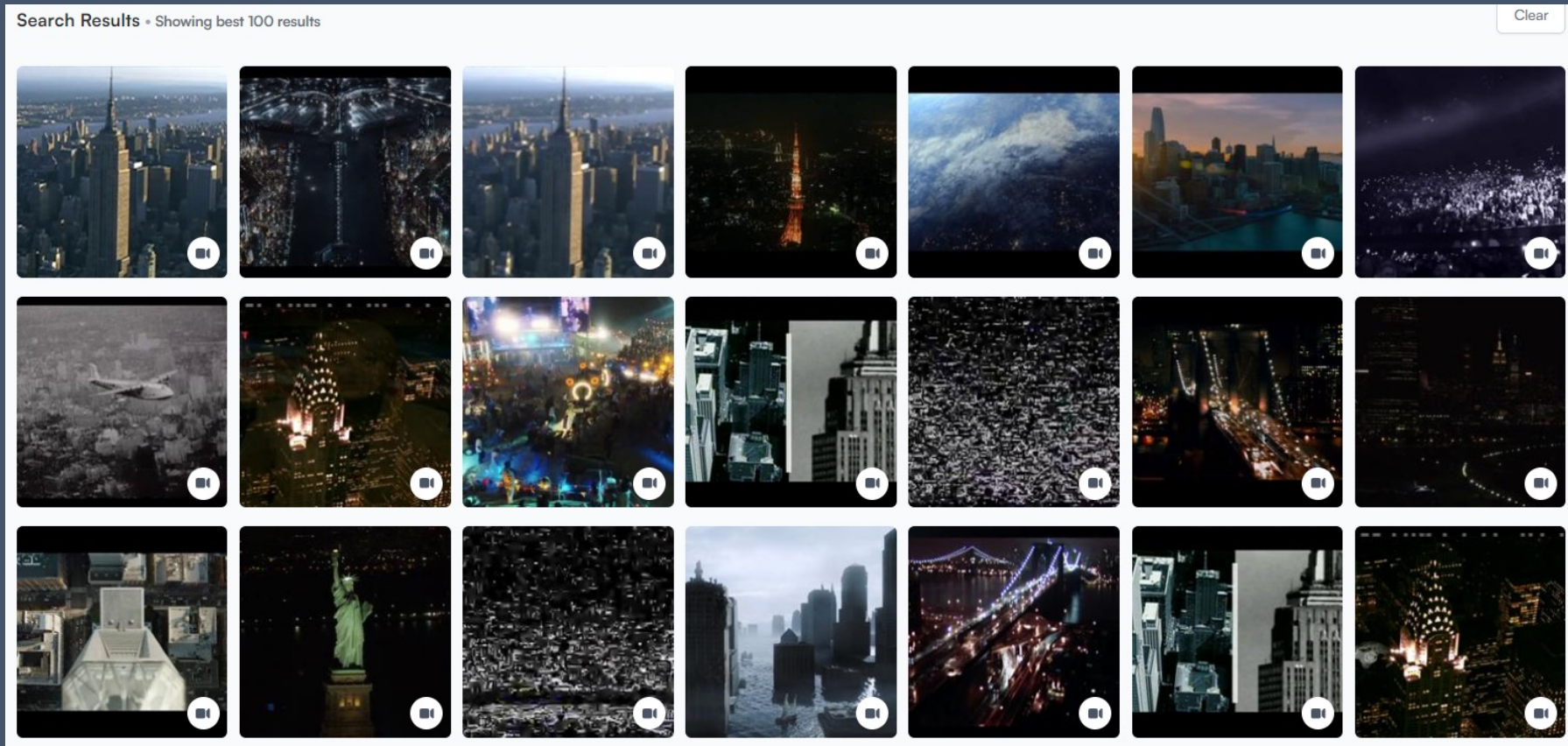
Example search shows these models understand text, animated characters, food and mood





# Stock footage need: NYC aerals

Example search understand landmarks, locations and shot angles





## TECH RETREAT 2024

Carin Forman – AWS - [forcarin@amazon.com](mailto:forcarin@amazon.com)

Heidi Shakespeare, Memnon - [heidi.shakespeare@memnon.com](mailto:heidi.shakespeare@memnon.com)

Linda Tadic, Digital Bedrock - [ltadic@digitalbedrock.com](mailto:ltadic@digitalbedrock.com)

Andrea Kalas - [Andrea\\_Kalas@Paramount.com](mailto:Andrea_Kalas@Paramount.com)