Artificial Intelligence and Machine Learning in Content Creation

Tom Ohanian
IBM Watson Media and Weather

The Two Core Questions…

Can Artificial Intelligence and Machine Learning produce content that heretofore required humans to produce?

Can creative decisions be codified such that intelligent systems can accomplish those tasks?

Applying AI & ML to Content Creation

Speech recognition providing real-time subtitling and CC in over 80 languages and with 95-99% accuracy.

Automatically creating personalized viewer highlights on a large scale.

Automatic creation of different versions of promos by changing voiceovers by retyping words.

Automated editing of footage from single or multiple cameras to create a coherent narrative of an event.

Automatic creation of frame accurate, lip-synced images from a content library, creating content that is completely fabricated from various source elements.

Evolving Content Creation & Consumption

The Business Models are Rapidly Changing

Broadcast Networks add OTT Services

Broadcast Networks add OTT Services

World’s Population: 7.5B

16.4B Internet Connected Personal Devices

Devices capable of acquiring content at: 4K/30; 1080p/120; 720p/240 fps.

100+ formats

Content Deluge, Lower CPM, Crowdsourcing

6/24/18

YouTube Daily:

- 5B videos
- 500M mobile video views
- 300 hours uploaded per minute

9/7/18

Facebook Daily:

- 8B average daily video views
- 1 in 5 is a live broadcast

All this growth comes with challenges…

Challenges of metadata tagging and categorization

Historical television model, based on CPM, redefined to audiences of thousands and hundreds

Make more content available with fewer resources

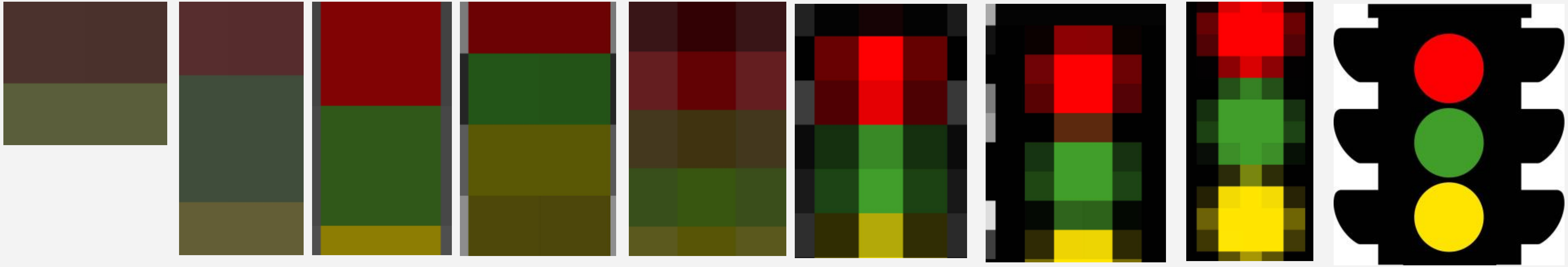
AI & ML

- AI investigates the simulation of human intelligence by computers.
- Machine Learning is a subset of AI. Algorithms capable of “learning” from the data and modifying operations without human intervention.
- AI can assist in image classification, facial recognition, etc. Rules by which systems logically undertake tasks: stock market trading.
- AI focused on imitation of human decision-making, automating the execution of those decisions.

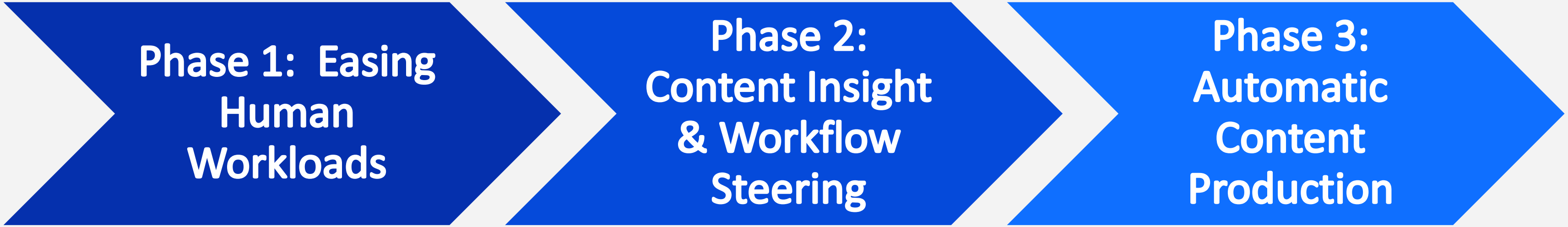
- ML apps analyze large datasets and, based on the learning process, make determinations and predictions.
- Examine written text, determine whether a positive or negative viewpoint is being expressed.
- Speech to text + tonal analysis: text of what is said, indexed to the specific moment in time, and interpret items such as:
  - The sex and approximate age of the speaker and the nature of the communication
  - A resulting indication: the speaker being pleased, displeased, or irate.

AI, ML & Neural Networks

- Neural Networks combine AI and ML to process data.
- Assume a rectangle comprised of circles of colored pixels. Based on a library of similar shapes and pixel patterns, the conclusion may be that the rectangle represents:



Three Phases of AI & ML to Content Creation

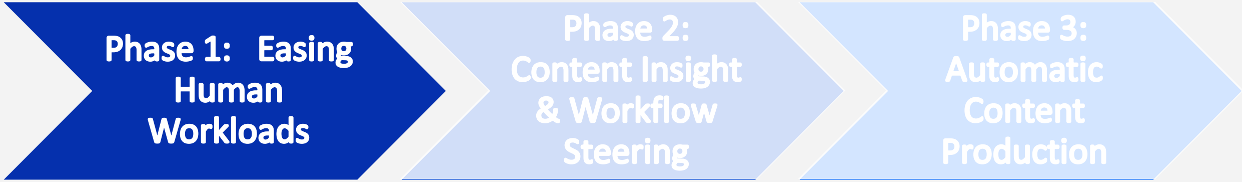


Phase 1: Easing Human Workloads

Phase 2: Content Insight & Workflow Steering

Phase 3: Automatic Content Production

Speech to Text	Language Recognition	Cognitive Metadata Extraction	Speech & Tonal Analysis
Image Recognition	Near Human Voice Quality Dubbing	Real-time Data & Statistical Integration and Analysis	General Automation Routines



Phase 1: Easing Human Workloads

Phase 2: Content Insight & Workflow Steering

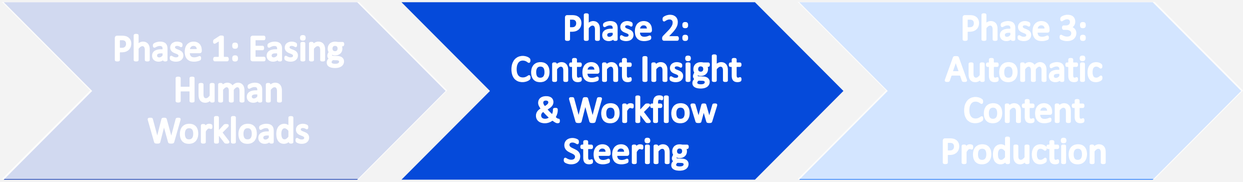
Phase 3: Automatic Content Production

Closed Captioning → Phonetics-to-text (PTT) → Automatic Speech Recognition (ASR)

Personnel would type the spoken words. Text would appear in 2-3 seconds.

Shifted reliance from operators to automated systems

Speech recognition, subtitling and closed-caption creation in over 80 languages and with 95-99% accuracy is a reality.



Phase 1: Easing Human Workloads

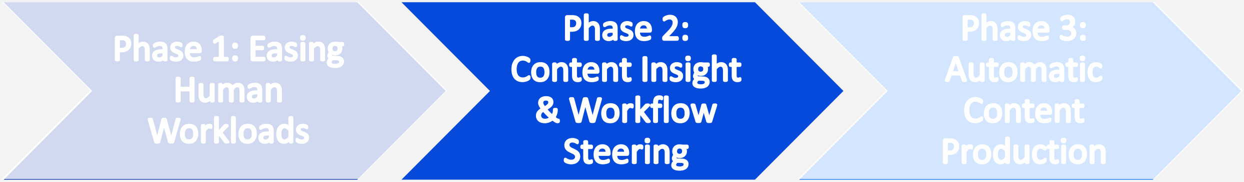
Phase 2: Content Insight & Workflow Steering

Phase 3: Automatic Content Production

Content, contextual metadata and essence extraction that provides content value to the owner and consumer.

Technologies applied to Phase 2 solutions include:

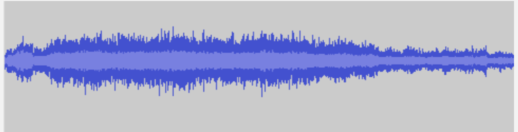
- Image Recognition
- Speech & Tonal Analysis
- Real-time Data & Statistical Integration
- Cognitive Metadata Extraction



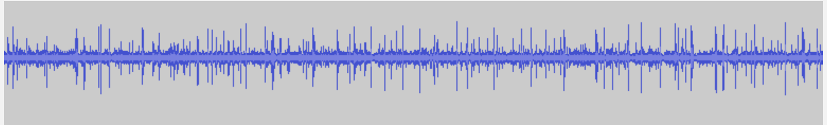
Phase 1: Easing Human Workloads

Phase 2: Content Insight & Workflow Steering

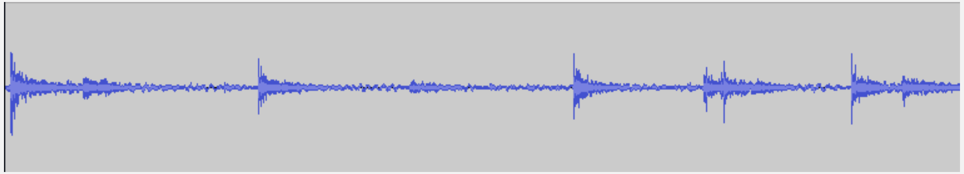
Phase 3: Automatic Content Production



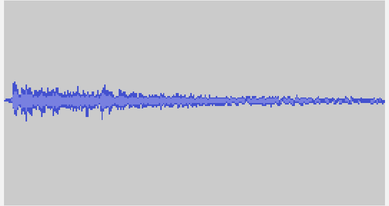
Viewer Catcalls



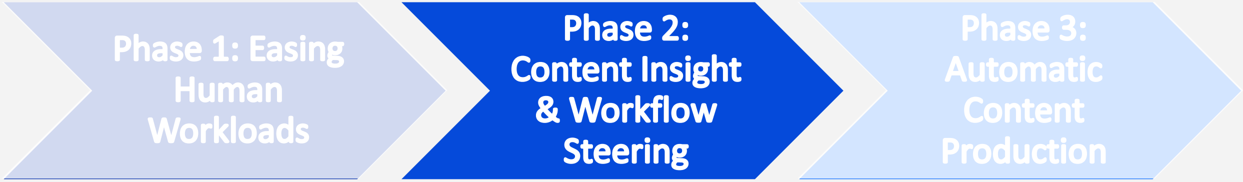
Viewer Cheers



Tennis Volley



Tennis Ace


Phase 1: Easing Human Workloads

Phase 2: Content Insight & Workflow Steering

Phase 3: Automatic Content Production

Automated methods for extracting and delivering added-value clips and content to viewers.

Real-time Data and Statistical Integration and Analysis: Potential clips based on court data and statistics. Breakpoints won, serves, scoring data, historical performances

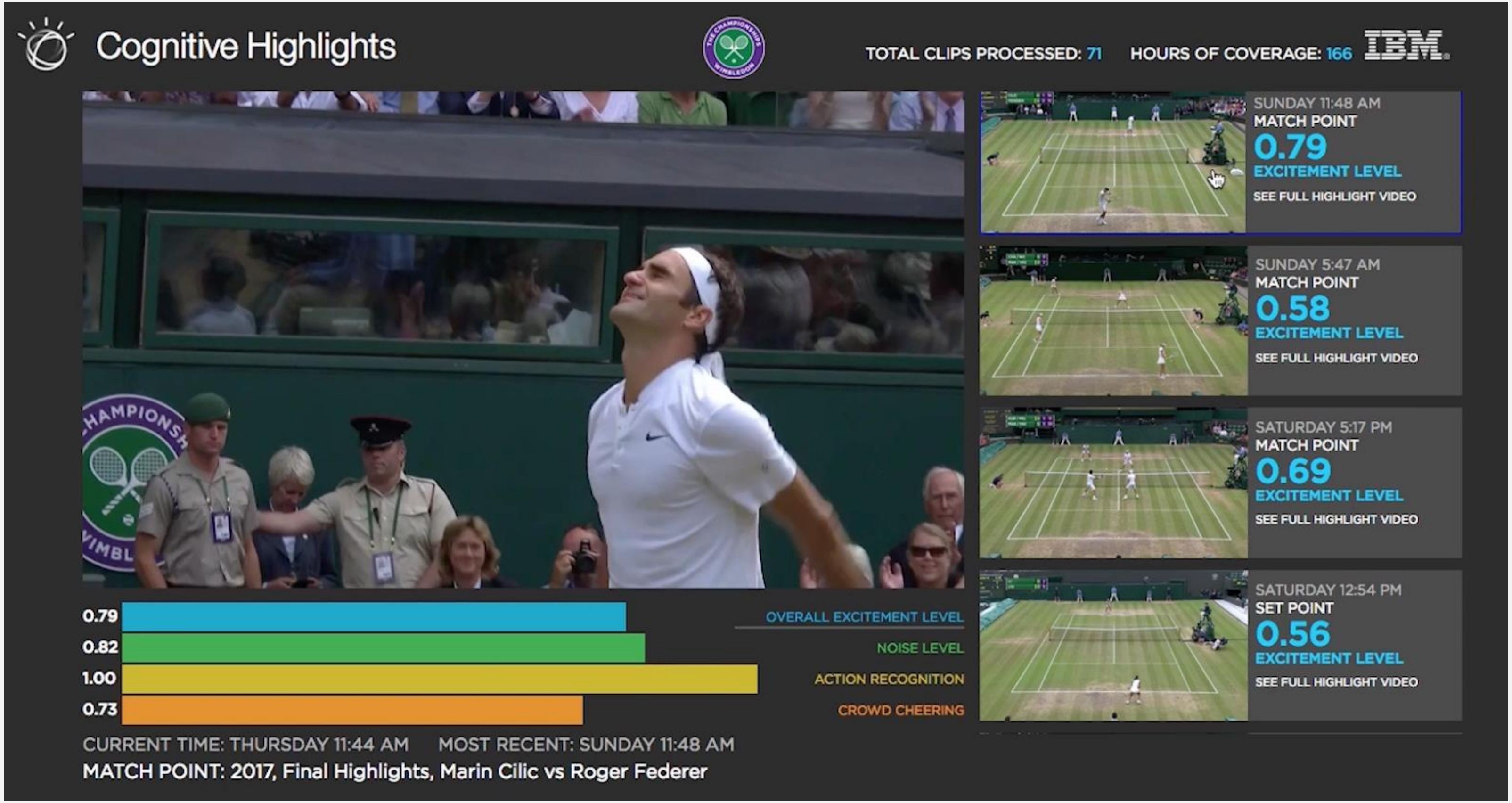
Image Recognition & Speech and Tonal Analysis: Analyze crowd cheering and other noises. Based on library video of players, image recognition identified players and cataloged reactions.

Cognitive Metadata Extraction: Is a player's smile due to a point won? Can it be correlated to a winning volley or serve? Can logical clips be created?

Automatically Creating Highlight Clips: Combining Phase 2 technologies and data feeds of play types correlated to timestamps creates highlight clips.

End result: Automatic creation of highlight clips. By classifying players via facial pixel makeup, viewers could point a cell phone at a player to receive information unique to that player.

Cognitive Extraction & Highlights



Cognitive Highlights

TOTAL CLIPS PROCESSED: 71 HOURS OF COVERAGE: 166

IBM

0.79 OVERALL EXCITEMENT LEVEL

0.82 NOISE LEVEL

1.00 ACTION RECOGNITION

0.73 CROWD CHEERING

CURRENT TIME: THURSDAY 11:44 AM MOST RECENT: SUNDAY 11:48 AM

MATCH POINT: 2017, Final Highlights, Marin Cilic vs Roger Federer

SUNDAY 11:48 AM
MATCH POINT
0.79
EXCITEMENT LEVEL
SEE FULL HIGHLIGHT VIDEO

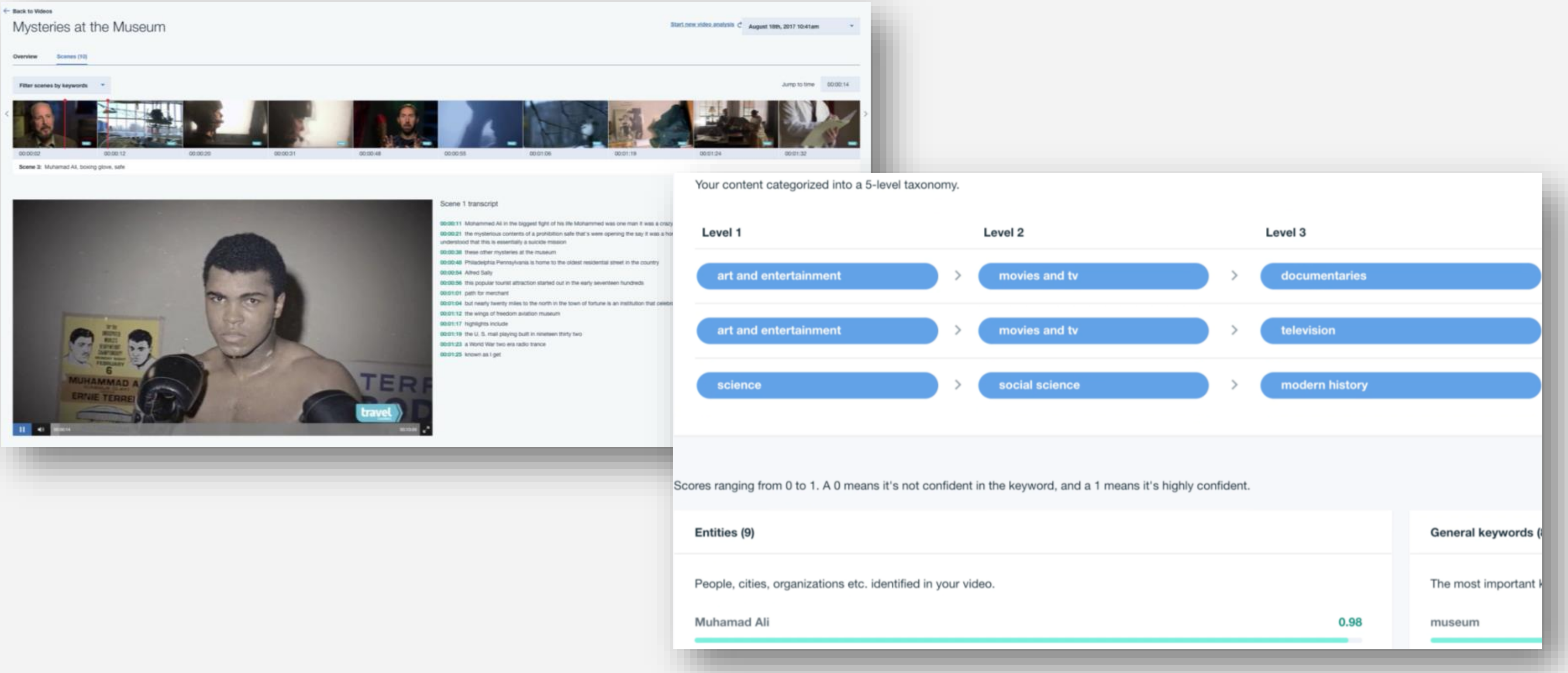
SUNDAY 5:47 AM
MATCH POINT
0.58
EXCITEMENT LEVEL
SEE FULL HIGHLIGHT VIDEO

SATURDAY 5:17 PM
MATCH POINT
0.69
EXCITEMENT LEVEL
SEE FULL HIGHLIGHT VIDEO

SATURDAY 12:54 PM
SET POINT
0.56
EXCITEMENT LEVEL
SEE FULL HIGHLIGHT VIDEO

Content Indexing & Categorization

IBM Watson Media Video Enrichment



Mysteries at the Museum

Scene 3: Muhamad Ali, boxing glove, safe

Scene 1 transcript

00:00:11 Mohammed Ali in the biggest fight of his life Mohammed was one man it was a crazy
00:00:21 the mysterious contents of a prohibition safe that's were opening the say it was a ho
understood that this is essentially a suicide mission
00:00:38 these other mysteries at the museum
00:00:48 Philadelphia Pennsylvania is home to the oldest residential street in the country
00:00:54 Alfred Sally
00:00:56 this popular tourist attraction started out in the early seventeen hundreds
00:01:01 path for merchant
00:01:04 but nearly twenty miles to the north in the town of fortune is an institution that celebr
00:01:12 the wings of freedom aviation museum
00:01:17 highlights include
00:01:19 the U. S. mail playing built in nineteen thirty two
00:01:23 a World War two era radio trance
00:01:25 known as I get

Your content categorized into a 5-level taxonomy.

Level 1		Level 2		Level 3
art and entertainment	>	movies and tv	>	documentaries
art and entertainment	>	movies and tv	>	television
science	>	social science	>	modern history

Scores ranging from 0 to 1. A 0 means it's not confident in the keyword, and a 1 means it's highly confident.

Entities (9)

People, cities, organizations etc. identified in your video.

Muhamad Ali 0.98

General keywords (

The most important k

museum

Content Indexing & Categorization

IBM Watson Media Video Enrichment

Watson Captioning

AI-powered captioning technology automates and captions for live broadcasts and on-demand video

THE SOLUTION

Optimize and automate the process of closed captioning using artificial intelligence capabilities of IBM Watson.

USE CASES

Live captioning for broadcast communications

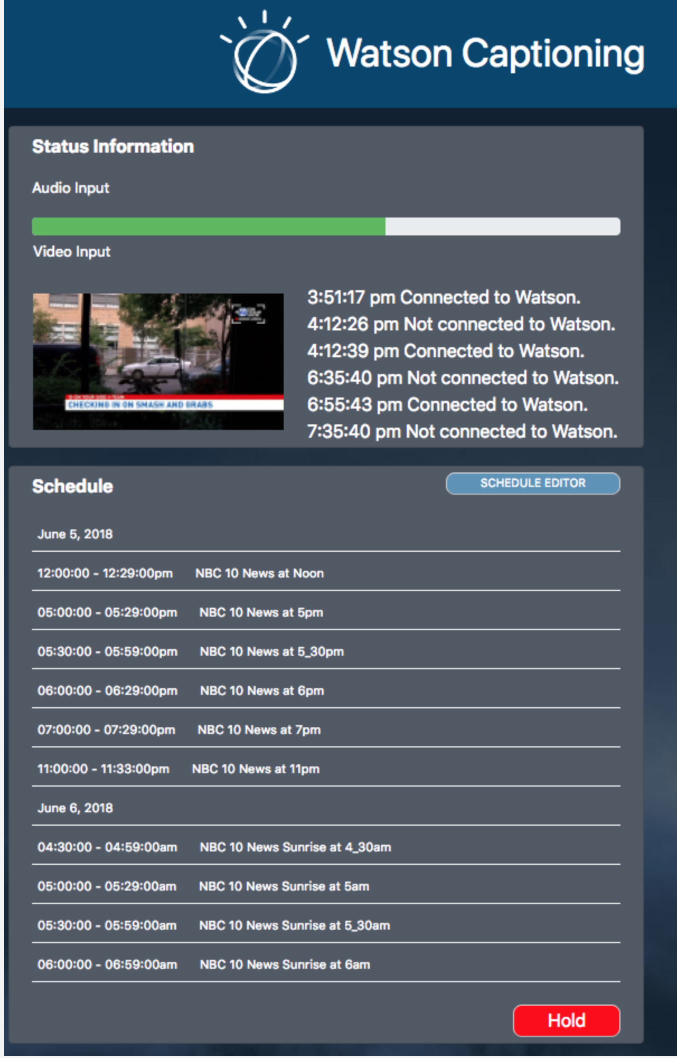
Captioning for video content libraries

BUSINESS BENEFITS

Speeds up delivery time – cost savings

Training model and custom corpora produces initial accuracy rate of 92-96% and will continue to increase with time

Make content more accessible to the deaf community



Status Information

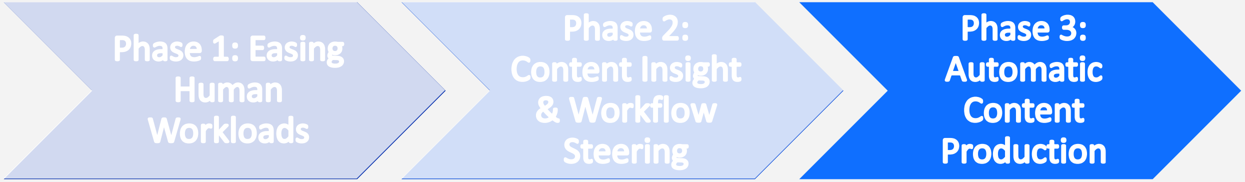
Audio Input

Video Input

3:51:17 pm Connected to Watson.
4:12:26 pm Not connected to Watson.
4:12:39 pm Connected to Watson.
6:35:40 pm Not connected to Watson.
6:55:43 pm Connected to Watson.
7:35:40 pm Not connected to Watson.

Schedule

June 5, 2018	
12:00:00 - 12:29:00pm	NBC 10 News at Noon
05:00:00 - 05:29:00pm	NBC 10 News at 5pm
05:30:00 - 05:59:00pm	NBC 10 News at 5_30pm
06:00:00 - 06:29:00pm	NBC 10 News at 6pm
07:00:00 - 07:29:00pm	NBC 10 News at 7pm
11:00:00 - 11:33:00pm	NBC 10 News at 11pm
June 6, 2018	
04:30:00 - 04:59:00am	NBC 10 News Sunrise at 4_30am
05:00:00 - 05:29:00am	NBC 10 News Sunrise at 5am
05:30:00 - 05:59:00am	NBC 10 News Sunrise at 5_30am
06:00:00 - 06:59:00am	NBC 10 News Sunrise at 6am



Phase 1: Easing Human Workloads

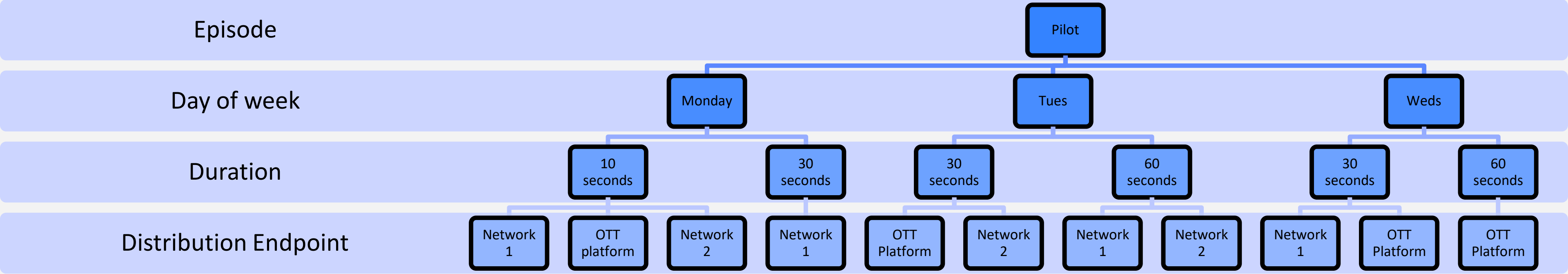
Phase 2: Content Insight & Workflow Steering

Phase 3: Automatic Content Production

- Produce content using rulesets and creative conventions in the form of idiomatic expressions.
- Speech Synthesis: Artificial production of human speech by concatenating pieces of recorded speech. Text to Speech (TTS) uses a database of recorded speech to create new combinations of speech. Database must be very large and emphasis of the spoken phrase may be difficult to shape.
- Heiga, Tokuda, and Black: Parametric TTS (P-TTS) where model parameters are adjusted to shape both content and characteristics of the speech. Output of the model is processed by algorithms in vocoders (voice encoders) and audio signals are generated.

Changing Words in Voiceovers by Retyping Words

- Generating synthetic speech provides countless possibilities.
- Instead of creating different versions by requiring new voiceovers be recorded, type the text of the new versions and have the audio changes automatically conformed.



Episode

Pilot

Day of week

Monday

Tues

Weds

Duration

10 seconds

30 seconds

30 seconds

60 seconds

30 seconds

60 seconds

Distribution Endpoint

Network 1

OTT platform

Network 2

Network 1

OTT Platform

Network 2

Network 1

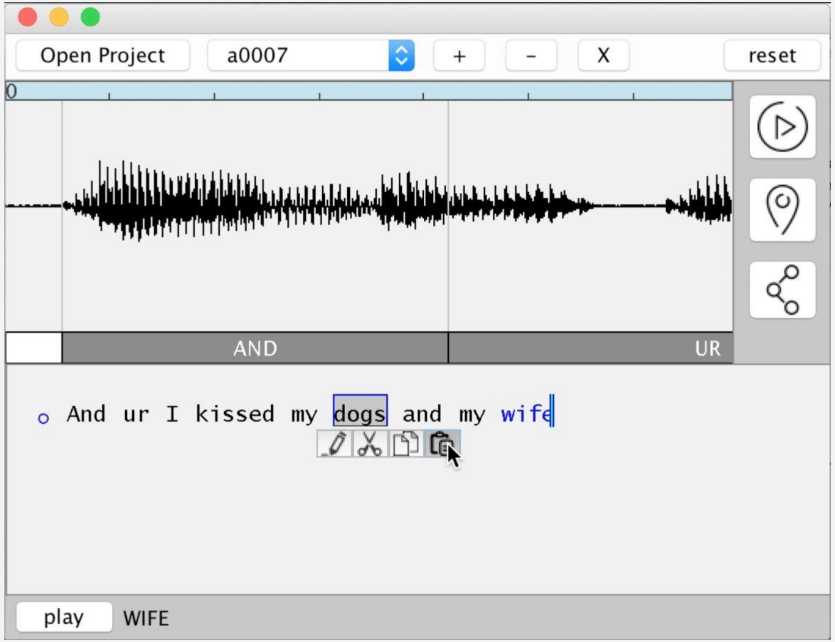
Network 2

Network 1

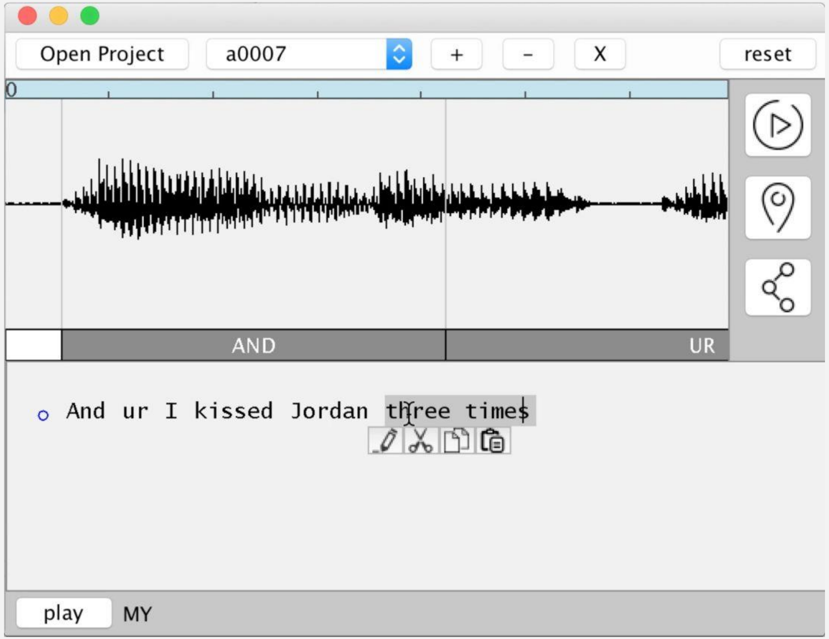
OTT Platform

OTT Platform

Project VoCo, Edit Speech in Text



○ And ur I kissed my dogs and my wife



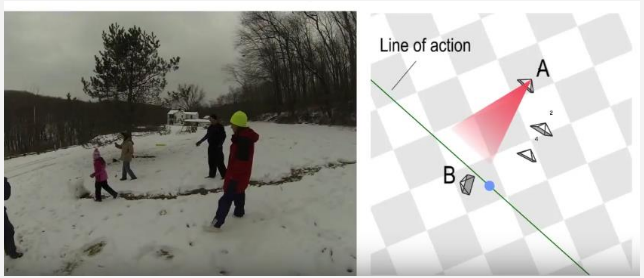
○ And ur I kissed Jordan three times

Zeyu Jin, "VoCo for Adobe Creative Cloud", Adobe Max

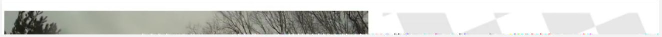
Auto Editing of Multicam Footage

Disney Research

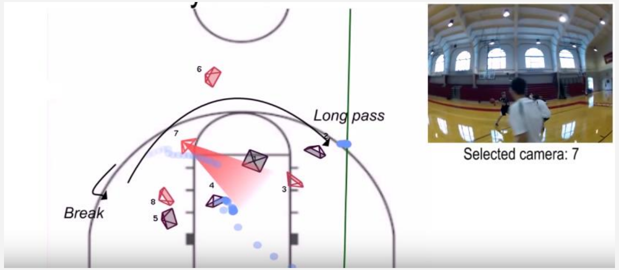
Four consumer / mobile cameras



Observing the 180-degree Line of Action Rule



Avoiding jump cuts between nearby cameras



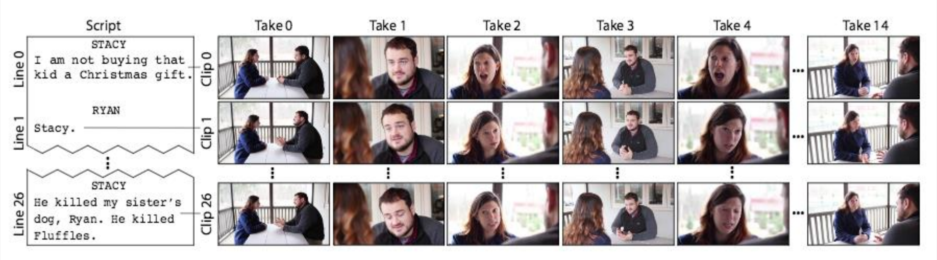
Selected camera: 7

3D camera motion estimation to identify prime interest areas to decide when to cut to a different angle.

Auto Editing of Multicam Footage

Content: 3D Joint Attention

Auto Editing of Single Camera Footage

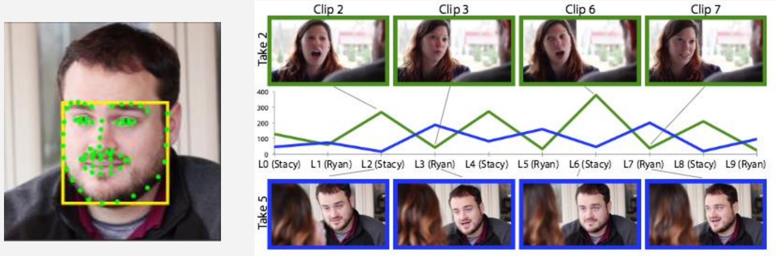


STACY
I am not buying that kid a Christmas gift.

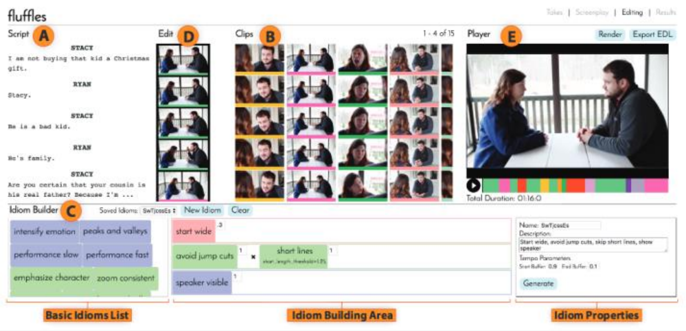
RYAN
Stacy.

STACY
He killed my sister's dog, Ryan. He killed Fluffles.

Correlating the Input Script into Lines of Dialogue Spoken by Each Character



Facial Analysis & Tracking and Computing Speakers Visible by Changes in Mouth Area



STACY
I am not buying that kid a Christmas gift.

RYAN
Stacy.

STACY
He is a bad kid.

RYAN
He's family.

STACY
Are you certain that your cousin is his real father? Because I'm ...

Total Duration: 01:16:0

Name: SwTjcssEs
Description: Start wide, avoid jump cuts, skip short lines, show speaker
Tempo Parameters
Start Buffer: 0.9 End Buffer: 0.1

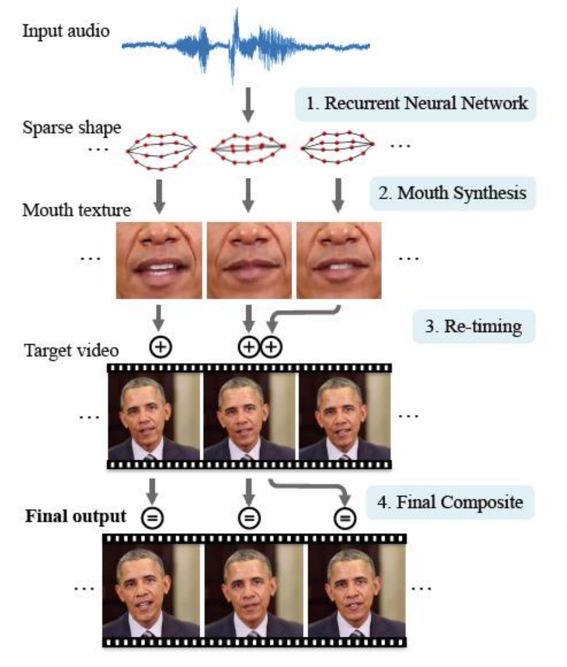
Choosing from the Idioms List and Placing into the Building Area results in Automatic Scene Construction.

Agrawala, Davis, Leake, Truong: Stanford & Adobe Research.

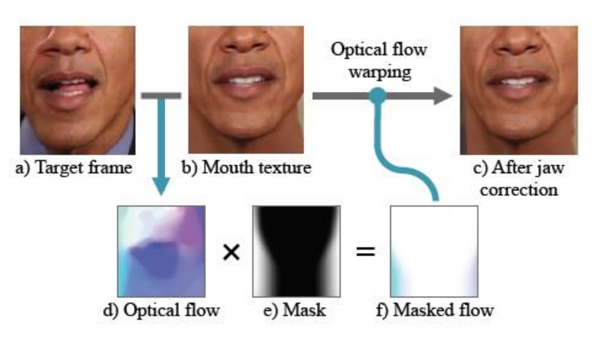
Auto Editing of Single Camera Footage

Agrawala, Davis, Leake, Truong: Stanford & Adobe Research.

Creating a Shot that Never Existed

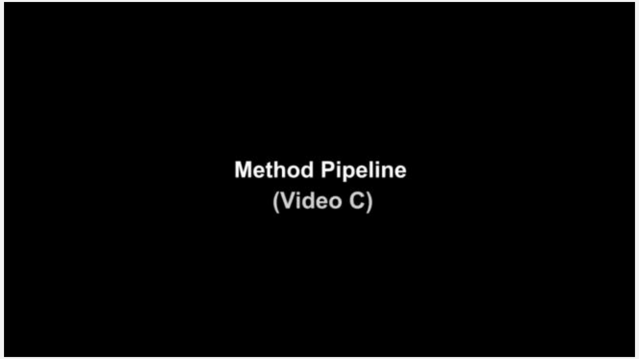


With a database of mouth shapes associated with time instances, mouth textures were synthesized and then composited with 3D matching to change what he appears to be saying. The result is that synthetic, photorealistic shots can be created.



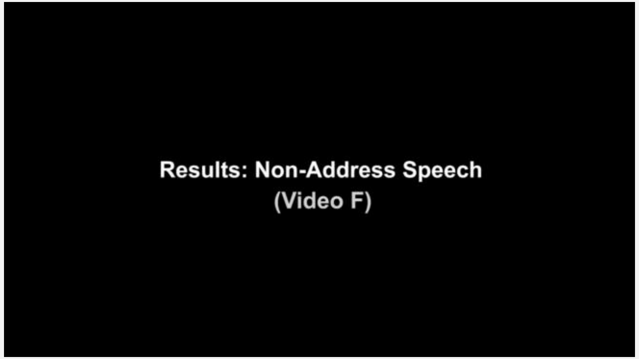
Audio Converted to Time Varying Mouth Shapes and Fixing Jawline Discrepancies.

Creating a Shot That Never Existed



Method Pipeline
(Video C)

Results: Weekly Address Speech
(Video E)



Results: Non-Address Speech
(Video F)

Conclusion

Artificial Intelligence, Machine Learning, and Neural Networks will make (and in some cases already are) significant contributions to the content creation-to-consumption process.