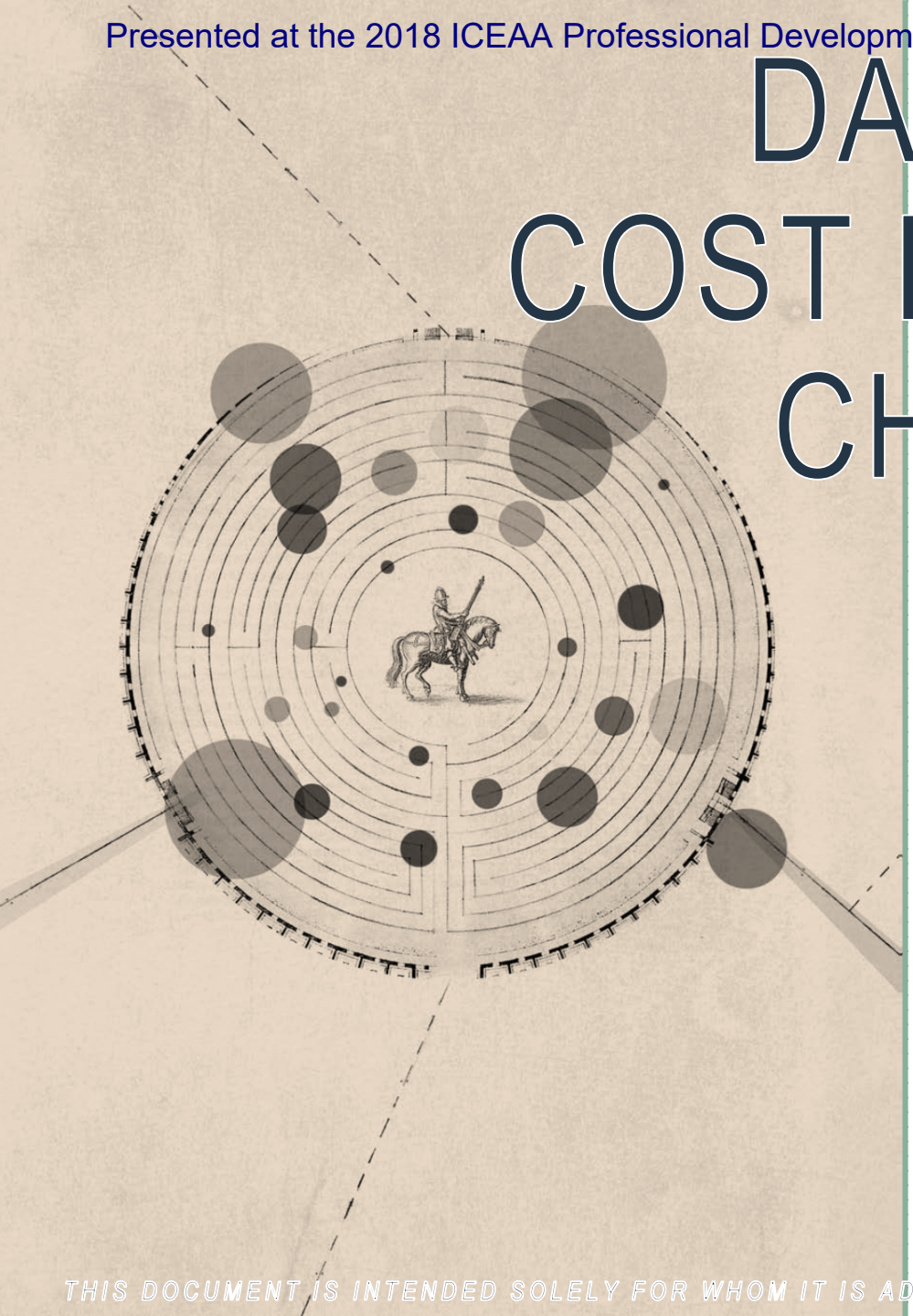
DATA SCIENCE COST ESTIMATING CHALLENGES:

Cut your time in Hack!



INTERNATIONAL COST ESTIMATING AND ANALYSIS ASSOCATION CONFERENCE 2018

JUNE 2018

DATA SCIENCE HACKATHONS...

Hackathons are more about a mindset than a skillset. Despite the name, data science hackathons do not involve physically 'hacking' into any organizational networks or systems. Instead, hackathons are sprint-like events that approach fascinating problem sets with creativity and forward thinking to explore and cultivate impactful ideas and solutions.

... AND HOW ORGANIZATIONS BENEFITS FROM THEM

PERSPECTIVE

Enables new perspectives on your problem, crowd sources approaches and solutions, and multiplies your workforce for a short period at little cost to the organization

BRANDING

Sets the demand signal across your respective areas for data science and brands your organization as a forward-leaning data science catalyst

EFFICIENCY

Accelerates and improves the military use of machine learning and improves efficiency of intelligence analyst products and workflows

RESOURCES

Provides 'free' talent and a diversification of programmers, mathematicians, and domain experts without the overhead/ administrative costs of having them on payroll

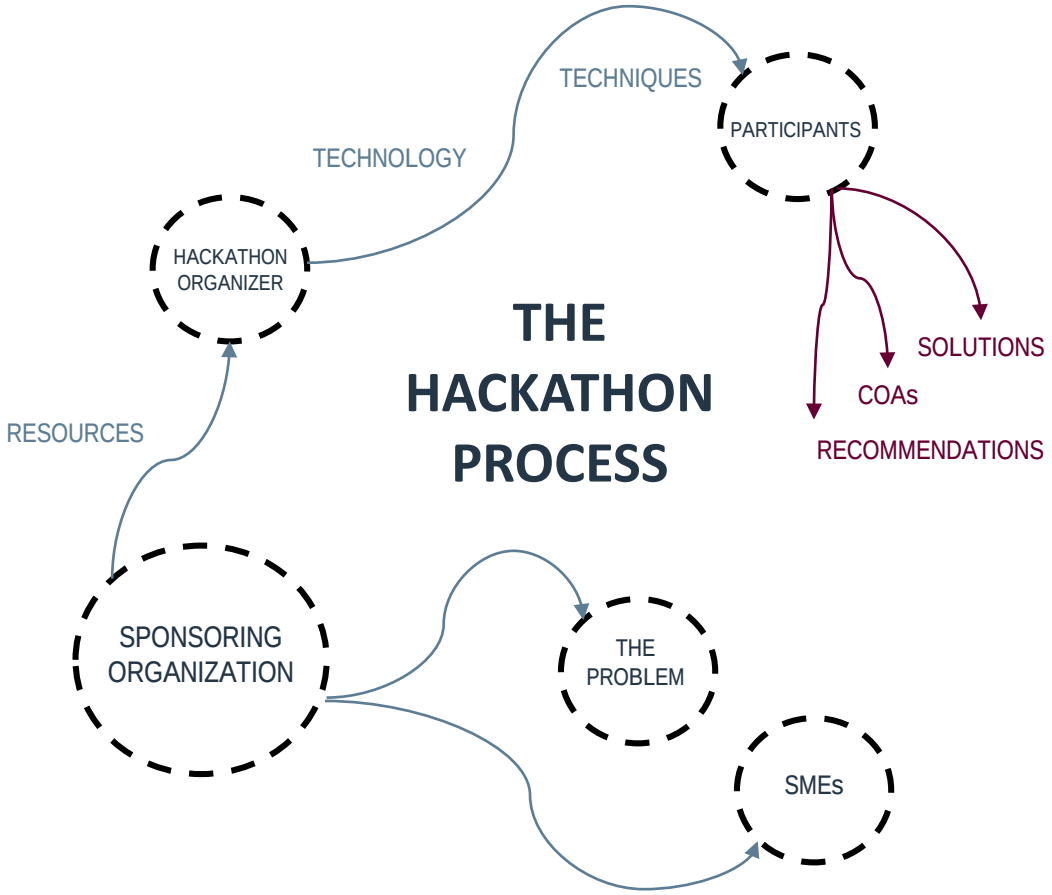
TALENT

Acts as a beacon for talent, revealing that your organization has opportunities to create data solutions, and acts as a recruiting mechanism for the next generation of data scientists

HACKATHONS UNDER THE HOOD

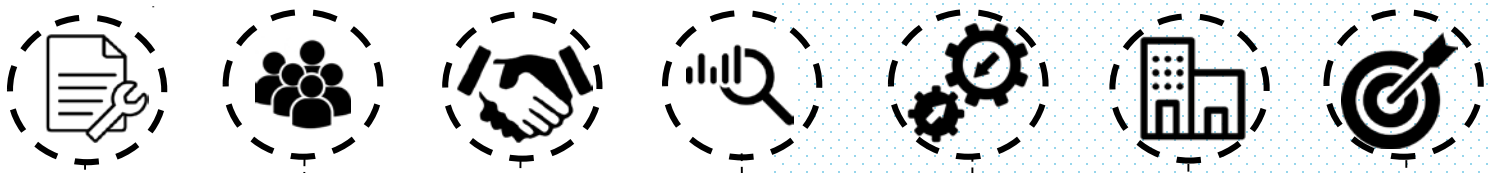
Hackathons scale a data science workforce in a very short period of time to focus on impeding challenges to a mission's success. The acceleration of data science and machine learning capabilities with military use improves efficiency of cost estimation, such as modeling approaches and uncertainty analysis parameters. As a beacon for talent, Data Science Hackathons serve as innovative catalysts to accelerate and transform how your enterprise engages problems and can provide 365 days' worth of work in just a few short days.

+ *How are proposals for new analytic solutions reviewed for investment?*
+ *Who is thinking about how to maximize the potential of your data?*
+ *Do you know who the Data Scientists are in your organization?*
+ *How does your organization engage participants and partners?*
+ *Are data results presented in a way that makes the conclusions obvious?*
+ *Do you know which technologies and technique best suit your problem?*
+ *How do you capture lessons learned from your ongoing analyses?*



WHAT YOU GET FROM EACH PHASE

BEHIND THE SCENES



DEFINE YOUR PROBLEM AND DATA

Get your problem, mission parameters, and valuable datasets defined and identified.

BRING NEW PARTNERS AND PARTICIPANTS TO THE TABLE

Have your problem looked at from every angle using an ecosystem of partners and participants from government, industry, academia, and local incubators; bring additional resources and assets to your fingertips and generating real excitement around your organization.

TEST YOUR DATA TO GUARANTEE RESULTS

Guarantee your data supports the challenge statement and results can be achieved through a pre-event test hack.

FACILITATE THE LOGISTICS, FACILITATING THE EVENT

Ensure you have the right material to generate and capture value intelligence from the event for post-event stakeholder communication.

GET OUTPUTS TO THE RIGHT PEOPLE

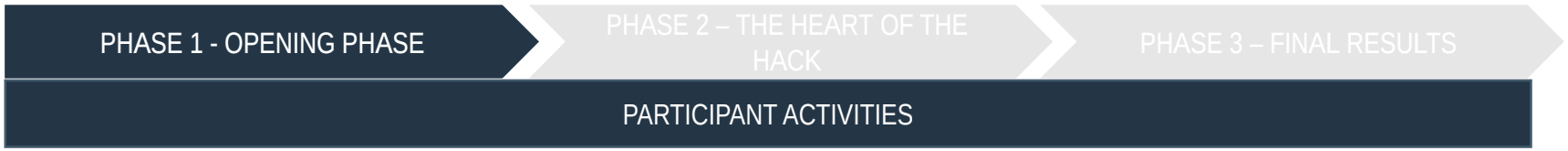
Walk away with actionable insights that inform the decision chain, nominate courses of action, and clearly outline your data science and machine intelligence options. You can also use our consolidated report on all hacker results, code, briefings, and recommendations to organize a follow-on event, announce job opportunities, and summarize event results.

HOW HACKATHONS TRANSFORM YOUR OUTPUTS

CHANGING THE MINDSET FROM SIMPLE ANSWERS TO SINGLE QUESTIONS

NOW	vs.	AFTER A HACKATHON
"We just don't have enough people to think through these problems."	⋯▸	Crowdsourced approaches and solutions
"Not everyone is innovative. You can't solve unique problems with the same stale mindsets."	⋯▸	New perspectives on how to solve specific problems
"We don't have the assets or technologies within our organization."	⋯▸	Exposure to open source and emerging technologies not currently available to your organization
"Proving that we need these capabilities is just too hard without results."	⋯▸	Proven use cases to justify adding new resources and tools to your arsenal
"We don't have buy in from the uppermost leadership."	⋯▸	Curated leadership buy-in at the top of your organization through newly established "internal champions"
"We don't even know where to look or who to leverage to solve some of these problems."	⋯▸	Innovative buzz throughout your organization supported by a fresh ecosystem of industry and academic partners
"It takes too much time from inception to results."	⋯▸	Rapid organizational movement into data science and machine intelligence
"Our hiring process just doesn't cultivate the right type of people."	⋯▸	Fun, interactive recruiting mechanism for the next generation of Data Scientists

WHAT A HACKATHON LOOKS LIKE



Orientation to the event, host, data, and objective

Teams and resources distributed into collaborative groups

Tutorials are given to the teams and a brief Q&A fills gaps and initiates the kick off

GUIDED BY YOUR ORGANIZATION TO

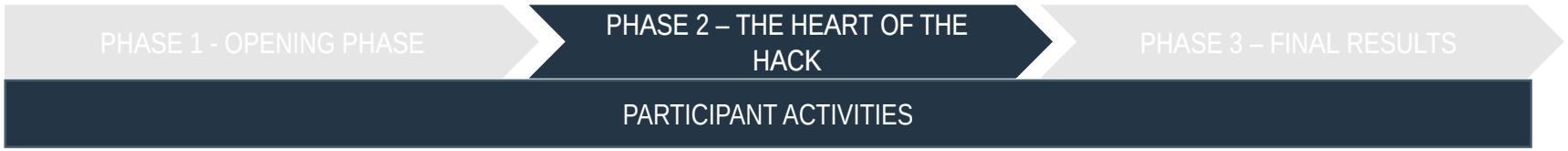
Contextualize the operating environment, mission, problem set, and message

Tailor and maximize assets based on dynamics and expectations envisioned by your organization

Identify the techniques and technologies most appropriate for the problem set and operating environment

PERSISTENT CAPTURE OF PROCESSES, TECHNIQUES, LANGUAGES, TOOLKITS, LIBRARIES, RESULTS, CODE, SOLUTIONS, RECOMMENDATIONS, AND AARS

WHAT A HACKATHON LOOKS LIKE



In-stride collaboration between participants and the organization

Real time feedback and status reports

Constant iteration

Down selection of teams

GUIDED BY YOUR ORGANIZATION TO

Provide domain subject matter expertise to ensure results can be operationalized

Enhance mentoring sessions

Funnel ideas toward the finish line

Curate the most Suitable or successful approaches

PERSISTENT CAPTURE OF PROCESSES, TECHNIQUES, LANGUAGES, TOOLKITS, LIBRARIES, RESULTS, CODE, SOLUTIONS, RECOMMENDATIONS, AND AARS

WHAT A HACKATHON LOOKS LIKE

PHASE 1 - OPENING PHASE | PHASE 2 – THE HEART OF THE HACK | PHASE 3 – FINAL RESULTS

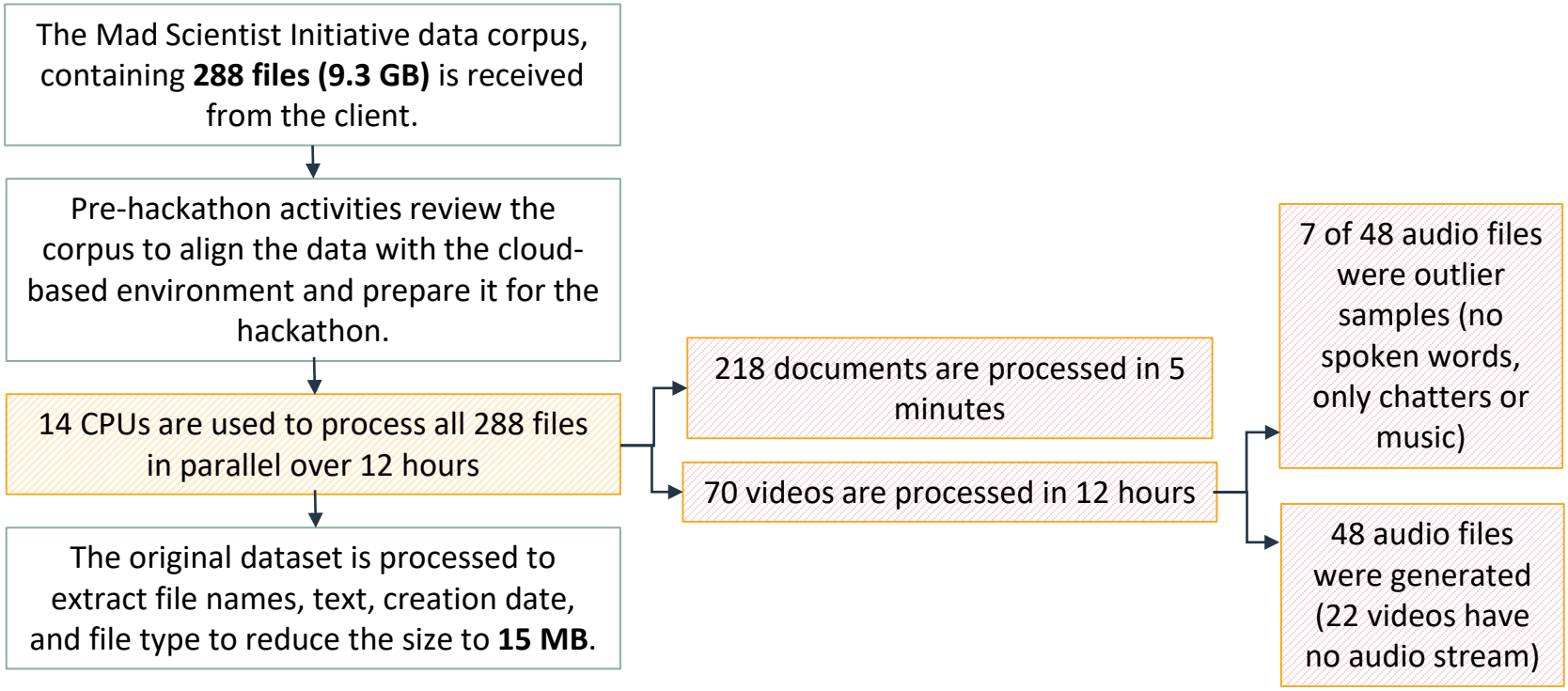
PARTICIPANT ACTIVITIES

Participant presentations	Tie results and lessons learned to mission set	Judges and criteria organization	Final results delivered
Backbrief the organization and provide AARs	Align hackathon results with the organization's message and commander's intent	Effectively adjudicate COAs and nominate recommendations	Walk away with use cases, proven results, and next steps

GUIDED BY YOUR ORGANIZATION TO

PERSISTENT CAPTURE OF PROCESSES, TECHNIQUES, LANGUAGES, TOOLKITS, LIBRARIES, RESULTS, CODE, SOLUTIONS, RECOMMENDATIONS, AND AARS

DATASET STATISTICS



EVENT STATISTICS

HUMAN CAPITAL

54 PARTICIPANTS | 48 HOURS

↓

11 FINALISTS | 3 TEAMS

↓

3 FINAL BRIEFS

MACHINE CAPITAL

12 MICROSOFT AZURE GOVERNMENT CLOUD-BASED LINUX/WINDOWS VM INSTANCES

- $15,000 Donated Azure Gov't Credit
- EACH VM = 16 CPUs, 56 GB of System Memory, 12 GB of SSD Storage

10 OPEN DATA PLATFORM VM

1 SAILFISH VM

PRELOADED SOFTWARE

← Cloudera Manager, NiFi, ElasticSearch, Kibana, Python 2.7, Apache Tika, Sphinx Speech Recognition, FFMPEG

HACKATHON TUTORIALS

Tutorials provide domain subject matter expertise to fill gaps, initiate the event, and ensure results can be operationalized

Text Extraction from Document and Video – Donovan Lo, Data Scientist, Booz Allen Hamilton

- **Overview**: detailed end-to-end tutorial on how to load, parse, and extract text information from the entire unstructured TRADOC Mad Scientist Initiative dataset, including digital documents (PDF, Microsoft Office products) and videos (.avi, .mov, .wmv, .mp4, .m4v)
- **End result**: a structured table format with the following fields – file name, extracted text, creation date, file type
- **Value**: derive structured information from unstructured data

Topic Modeling of Unstructured Data – Phi Vu Tran, Data Scientist, Booz Allen Hamilton

•**Overview**: detailed end-to-end tutorial on how to leverage unsupervised machine learning to automatically discover and extract underlying themes or "topics" from a collection of text data and correlate topics to available metadata such as date and location

•**End result**: an algorithm that can be used to organize massive amounts of text data into categories or topics of interest for visualization, search and retrieval, among other applications

•**Value**: discover and extract semantic meaning from text to aid analysis and decision-making

HACKATHON RESULTS | TEAM ZERO WINNER HIGHLIGHT

Members: Muntaser Syed, Jim Cannaliato, Sally Tan, Kent Huang

TOOLS AND PACKAGES

AWS REKOGNITION

GOOGLE CLOUD API
TEXTRAZOR

GDELT (DATES TO TOPIC QUERY ENGINE)

*Team Zero creatively **utilized metadata, visual, and audio sources** with various software tools to extrapolate TRADOC related topics and categories to:*

- Extract topics from video, audio using AWS Rekognition and Google Cloud API
- Extract categories from metadata using TextRazor and GDELT

STRENGTHS	WEAKNESSES	JUDGES FEEDBACK
• **Technical Performance** Successfully processed all files in 4 hours . • **Creativity** - Utilized both audio and video to extract categories and topics. - Tested multiple technologies to validate assumption (IBM Watson, Google Cloud API, DeepDive, etc.)	• **User Focus** - Need to show the reasons for correlations and predictive analysis. - Need to show cited recommendations for further analysis	• Liked that the team improved data visualization for showing results • Impressed that the team was able to pull out theme from video • Would like to see potential cost associated with research (licenses for analytic tool) • Enjoyed the real world GDELT query engine

HACKATHON RESULTS | TEAM SYNERGY RUNNER UP HIGHLIGHT

Members: Justin, Angela, Tom, Hellen, AmaR

TOOLS AND PACKAGES

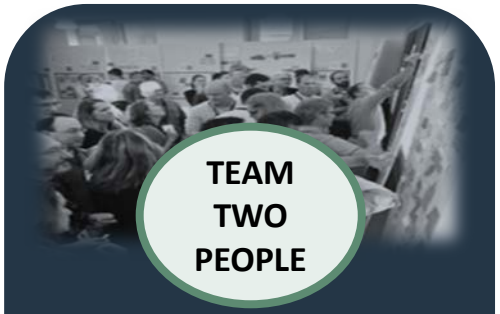
APACHE TIKA **FFMPEG**
SHINX (SPEECH RECOGNITION)

Team Synergy addressed the specific problem within the process flow and innovatively proposed the concept of worker thread into the system to:

- Parallel processes at file level
- Maximize number of CPU and memory resource being utilized

STRENGTHS	WEAKNESSES	JUDGES FEEDBACK
• **Technical Performance** - Successfully processed all files. • **Creativity** - Implemented Multi-thread video parsing.	• **Technical Performance** - Need to explore other applications or technical solution. • **User Focus** - Need more graphics / flowcharts • **General Technical Achievement** - Need to show text analysis	• Excited to see significant reduction in processing time • Would like to know what other tools were used • Like to see what information was gleaned from the work

HACKATHON RESULTS | TEAM TWO PEOPLE RUNNER UP HIGHLIGHT



Members: Randy Brou, Hillary Fleenor

TOOLS AND PACKAGES

APACHE TIKA
SHINX (SPEECH RECOGNITION)
PYTHON GENISM
PYTHON NLTK

Team Two-People ingeniously mimic the human cognitive process by grouping documents into year-group corpora and applying a two-pass topic analysis with LDA algorithm and ConceptNet approach to:

- Separated documents into year-group corpora
- 1st Pass: LDA algorithm to return 9 topics for each corpus
- 2nd Pass: output of 1st Pass & ConceptNet for topic analysis

STRENGTHS	WEAKNESSES	JUDGES FEEDBACK
• **Technical Performance** - Succinct on extraction strategy for topics over time. • **Creativity** - Implemented a 2 pass topic analysis system by LDA algorithm and ConceptNet • **User Focus** - Considered the user's request needs associated with the task.	• **Technical Performance** - Need to evaluate entire dataset including more videos. (Team started on second day)	• Thought that the team explained very well on how data is parsed by humans • Would like to see if solution can be scaled • Interested in seeing process include video analysis. • Liked that the team rationally targeted analysis to seek meaningful categories rather than rely on uncategorized semantic analysis

HOW CAN YOU LEARN MORE ABOUT HOSTING A DATA SCIENCE HACKATHON?

HOW YOUR SUCCESS IS OUR SUCCESS

1. Sit down with a Hackathon Organizer and begin to brainstorm on how hosting a data science hackathon can help solve some of your most challenging problems
2. Hear about data science hackathon success stories and learn how you can leverage previous results
3. Scope out next steps including a sample schedule, potential partners, and data sources to explore

JOE ROHNER
CHIEF TECHNOLOGIST
DATA SCIENCE AND MACHINE INTELLIGENCE
ROHNER_JOSEPH@BAH.COM

LORRAINE FEURY
STAFF SCIENTIST
DATA SCIENCE AND MACHINE INTELLIGENCE
FEURY_LORRAINE@BAH.COM