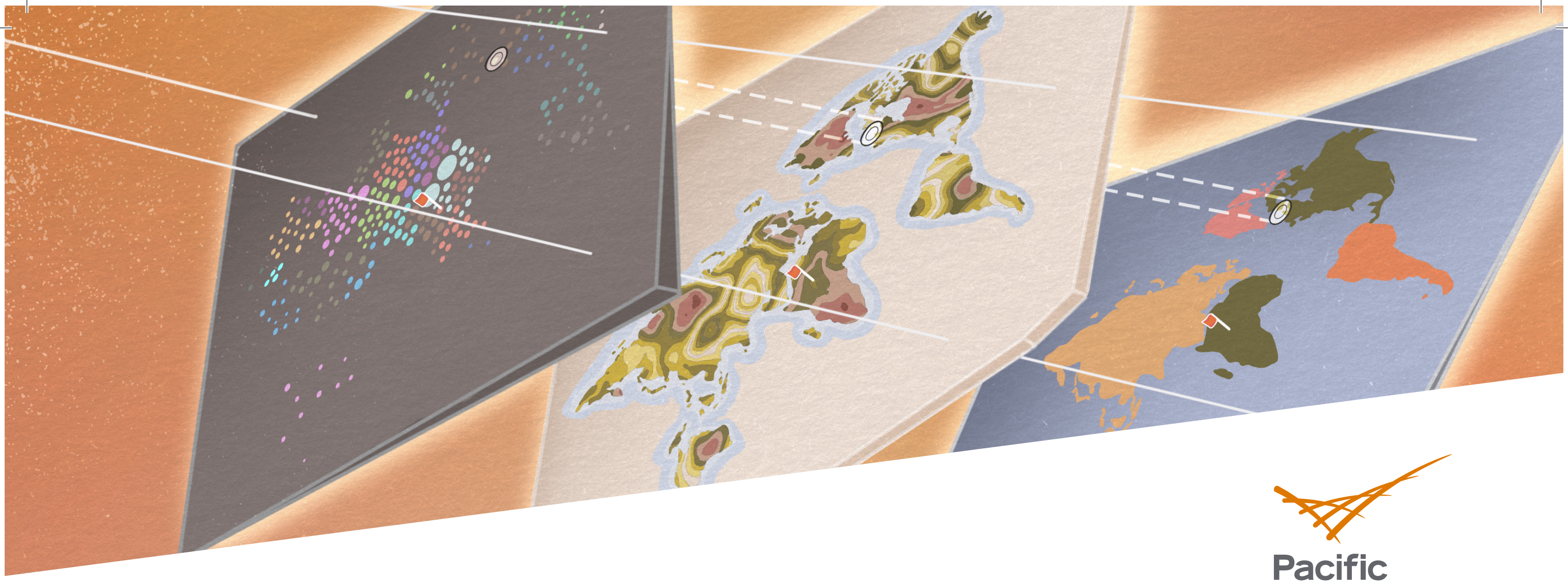


Text and Media

Into the Dataverse

THE CHALLENGE

Information overload makes the need to find connections in the data more vital than it was 20 years ago. The things we need to know from analyses of text and media sources today are more complex. Early software tools typically focused on aggregating keywords from various sources to show the strength of particular topics in news or conversations. Today there are billions of records to parse, and they come from sources as varied as social media, online traffic data, sensors, and photos. Analysts need to merge these records, geolocate the data, and show overarching topics in time and space.

Machine learning is making this easier, but it presents its own issues. With advanced computational processes running in the background, it is more challenging for humans to understand how the computer arrived at its conclusions and recommendations. Thus, users may be less comfortable with the data.

APPROACH

Pacific Northwest National Laboratory (PNNL) develops exciting new technologies that use text, image, audio, video and numerical data to help users discover patterns, trends, relationships and events. Designed for end-users, we deliver software products that put useful analytic capabilities to work combining text and images with analysis and context, which allows people to have confidence in the resulting data.

At PNNL we work in a variety of mission areas including energy, national security, and scientific discovery. We build custom algorithms to meet our sponsors' needs and, because we also know their mission areas on a technical level, the algorithms we provide just work better.

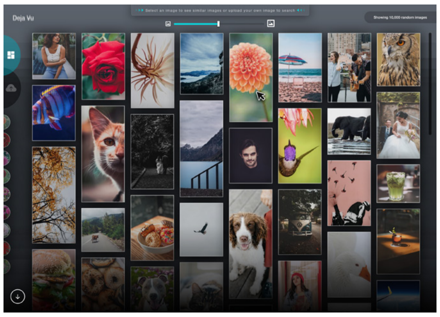
Researchers develop software that focuses on 'Concept drift.' In other words, how do conversations and topics drift from one topic to another? Instead of having to look at daily data, you can see emerging trends and detect anomalies. It's like giving the computer a bin of words and it will tell the analyst when they start to change.

Geoinference models also help analysts know where discussions are happening. Out of 700–800 million tweets per day, only approximately two percent are geotagged. PNNL software, developed by machine-learning techniques, helps to identify location by analysis of images and more.

EXAMPLE PROJECTS

Déjà vu

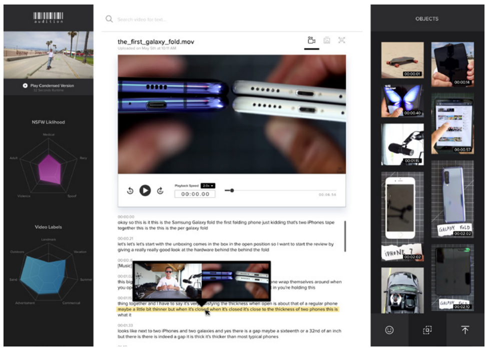
Déjà vu is an image feature similarity tool for finding images with similar features. It is a compelling combination of novel PNNL research ideas


Deja Vu
Select an image to see similar images or upload your own image to search
Showing 10,000 random images

and PNNL engineering savvy. Industry best-of-breed approaches are used to extract features from images, index them, and make them searchable. The innovative techniques used by PNNL has seen the features of 50 million images extracted and indexed in a matter of a few days with the final 50-million feature index returning search results in under a second. The result is a simple but powerful user interface that can help a user find similarities in a sea of pictures.

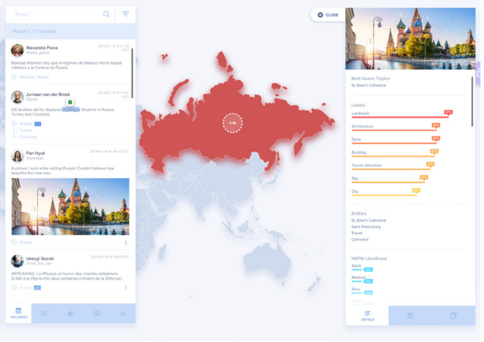
Audition

Audition is a video analysis tool that excels at picking out objects from video, including speech-to-text translation for on-screen transcriptions. Audition automatically condenses video to a series of key frames for easier analysis. Through topic modeling it is able to describe the general theme of the video, so issues of interest can be reviewed. Researchers are honing in on ways to identify speakers in the videos and being able to identify objects in each frame for faster searching. Audition helps reduce workloads and effort for analysts involved in identifying critical intelligence for decision makers.



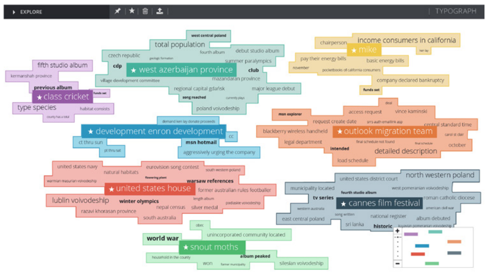
Clairvoyant

Clairvoyant is a tool that can search social media for text and images that are similar to a topic of interest. If there is a user of interest, Clairvoyant can not only search for tweets similar to a specific tweet, but it can take the users 'body of work' or all their tweets and identify users who have similar text histories. Clairvoyant also incorporates Déjà vu's image similarity, giving analysts the ability to get context around an image based off of tweets. It can also provide geoinference modeling. Only about two percent of posts on social media are geotagged. So, geoinferencing is needed but is difficult to do with text. For instance, if someone is posting about a place, it's difficult to know if they are just writing about it or actually at that place. Clairvoyant is taking steps to better help analysts distinguish these differences through continuing research.



Typograph

Statistical models exist for extracting keywords and generating topic models, but there were no effective tools to visually explore these topic spaces — until Typograph. Organized something like a geospatial word cloud, Typograph reveals topical areas within large datasets and is flexible for visualizing many kinds of topic models.



About PNNL

PNNL advances the frontiers of knowledge, taking on some of the world's greatest science and technology challenges. Distinctive strengths in chemistry, earth sciences, and data analytics are the heart of our science mission, laying a foundation for innovations that improve America's energy resiliency and enhance our national security. PNNL's computing research encompasses data and computational engineering, high-performance computing, applied mathematics, and semantic and human language technologies.

Contacts

Collaborate with us | Tap into our capabilities to meet your needs | Explore technology transfer opportunities | Join our team to grow your career

Aaron Harper
User Experience Designer
Pacific Northwest National Laboratory
(509) 371-6463
Aaron.Harper@pnnl.gov

David Gillen
Software Engineer
Pacific Northwest National Laboratory
(509) 375-5935
David.Gillen@pnnl.gov

Russ Burtner
Technical Group Manager, Visual Analytics
Pacific Northwest National Laboratory
(509) 371-6736
Russ.Burtner@pnnl.gov

U.S. DEPARTMENT OF ENERGY
BATTELLE

October 2019 | PNNL-SA-148324