



NAME _____

DATE _____

CLASS _____

NATIVE AMERICAN HERITAGE MONTH

Michael Running Wolf

Artificial Intelligence, Endangered Languages, and Data Sovereignty

Who Owns a Language's Data?

Michael Running Wolf, a citizen of the Northern Cheyenne Tribe, grew up in Birney, a remote village in southeastern Montana where running water and electricity were intermittent. He was raised in part by a grandmother who spoke Cheyenne. He earned bachelor's and master's degrees in **computer science** from Montana State University, worked as a software engineer on Amazon's Alexa, taught at Northeastern University, and went on to pursue a PhD in computer science at McGill University.

Modern **automatic speech recognition** (ASR) is built on machine learning: a model is trained on large quantities of transcribed audio until it can generalize to speakers it has never heard. Performance depends heavily on the volume and variety of the **training data**. For English and other widely spoken languages, that data is abundant. For many endangered Indigenous languages it is scarce — Michael has noted that for many Montana tribes the available recordings may total around 100 hours, and some tribes have essentially none.

The languages themselves pose a second challenge. Many Indigenous languages are **polysynthetic**: words are built from many meaningful parts, so one word can express what English needs a full sentence to say. Systems designed around vocabularies of shorter, separate words handle this poorly, and Michael has described it as an unsolved problem in AI. Inspired in part by Māori researchers in New Zealand who applied AI to their own language, he co-founded First Languages AI Reality (FLAIR), an initiative developing adaptable AI tools for Indigenous language revitalization.

Running through his work is the principle of **data sovereignty** — the idea that a community should own and control data about itself, including recordings of its language, stories and ceremonies. Michael argues that tribes need their own engineers and computer scientists and must set the terms for how their data is collected and used; otherwise the same technology meant to help a language could end up owned by someone else.

Part 1 • Comprehension & Analysis

1. Explain why the amount and variety of training data determines how well a speech-recognition model performs.

2. Identify the two separate technical obstacles to building ASR for many Indigenous languages described in the passage.

3. Define data sovereignty and explain why it matters more for a language community than for, say, a dataset of product reviews.

4. Michael worked on a major commercial voice assistant before starting FLAIR. How might that experience help, and how might it create tension, in his current work?

Part 2 • Quantitative Challenge

Sizing a Speech Dataset

Uncompressed speech audio for ASR is often stored at a 16,000 Hz sample rate, 16 bits (2 bytes) per sample, one channel. Use $1 \text{ MB} = 10^6$ bytes, $1 \text{ GB} = 10^9$ bytes and $1 \text{ TB} = 10^{12}$ bytes. Show your work and keep units consistent.

1 How many bytes does one hour of this audio take? Give your answer in MB.

2 How much storage do 100 hours of recordings need, in GB?

3 A mainstream system is trained on 50,000 hours. How much storage is that, in TB?

4 Suppose transcribing 1 hour of audio takes 5 hours of work by a fluent speaker. How many 40-hour work weeks would it take one speaker to transcribe the community's 100 hours?

Part 3 • Discussion & Extension

The storage for 100 hours of audio fits on a phone, but the work of recording and transcribing it depends on a small number of fluent speakers. Who should decide how a community's language recordings are collected, stored, shared and used to train AI — the speakers, the tribal government, the researchers who build the tools, or someone else? What rules would you put in place, and what trade-offs do they create?

