



NAME _____

DATE _____

CLASS _____

NATIVE AMERICAN HERITAGE MONTH

Michael Running Wolf

Training Data and the Fight for Endangered Languages

Why AI Needs So Many Examples

Michael Running Wolf is a citizen of the Northern Cheyenne Tribe who grew up in Birney, Montana, a remote village where running water and electricity were not always reliable. He earned bachelor's and master's degrees in **computer science** at Montana State University, then worked as a software engineer on Amazon's Alexa voice assistant. He later taught at Northeastern University and went on to pursue a PhD in computer science at McGill University in Canada.

Voice assistants rely on **automatic speech recognition** — AI that turns spoken sound into words. These systems are trained on **training data**: enormous collections of recorded speech, often with written transcripts. The more varied examples a model hears — different voices, accents, speeds and background noise — the better it becomes at understanding someone it has never heard before.

That is exactly the problem for endangered Native languages. Michael has pointed out that for many Montana tribes, the recorded audio that exists adds up to perhaps around 100 hours, and some tribes have almost none. Mainstream speech systems learn from vastly more. Many Indigenous languages are also **polysynthetic**, meaning a single long word can carry the meaning of a whole English sentence. Handling that kind of word-building is still an unsolved challenge for AI.

To tackle it, Michael co-founded First Languages AI Reality (FLAIR), which develops AI tools that communities can adapt to revitalize their languages. He argues that communities need their own engineers and computer scientists, and control over their own data, so the technology serves the people whose language it is.

Part 1 • Reading Comprehension

1. What is training data, and why does an AI speech system need a lot of it?

2. About how much recorded audio exists for many Montana tribal languages, according to Michael?

3. What does it mean that a language is polysynthetic, and why does that make AI harder?

4. Name two jobs or roles Michael has had, in order.

Part 2 • Critical Thinking

Michael says communities should have their own engineers and control over their own language data. What could go wrong if an outside company collected a tribe's recordings and owned them instead?

Part 3 • STEM Challenge

How Much Data Is Enough?

Suppose a community has 100 hours of recorded speech in its language. For comparison, imagine a mainstream speech AI trained on 50,000 hours of recordings. Show your work.

1 What percent of 50,000 hours is 100 hours?

2 Write the ratio of the mainstream data to the community's data in simplest form.

3 Elders and speakers record 2 hours of new audio every week. How many weeks would it take to grow the collection from 100 hours to 1,000 hours?

4 About how many years is that? (Use 52 weeks in a year and round to one decimal place.)
