
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

AIM409

Deploying neural networks at scale on AWS

Leo Simberg
Senior Solutions Architect
AWS

Najeeb Danish
Senior Cloud Architect
Deloitte

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

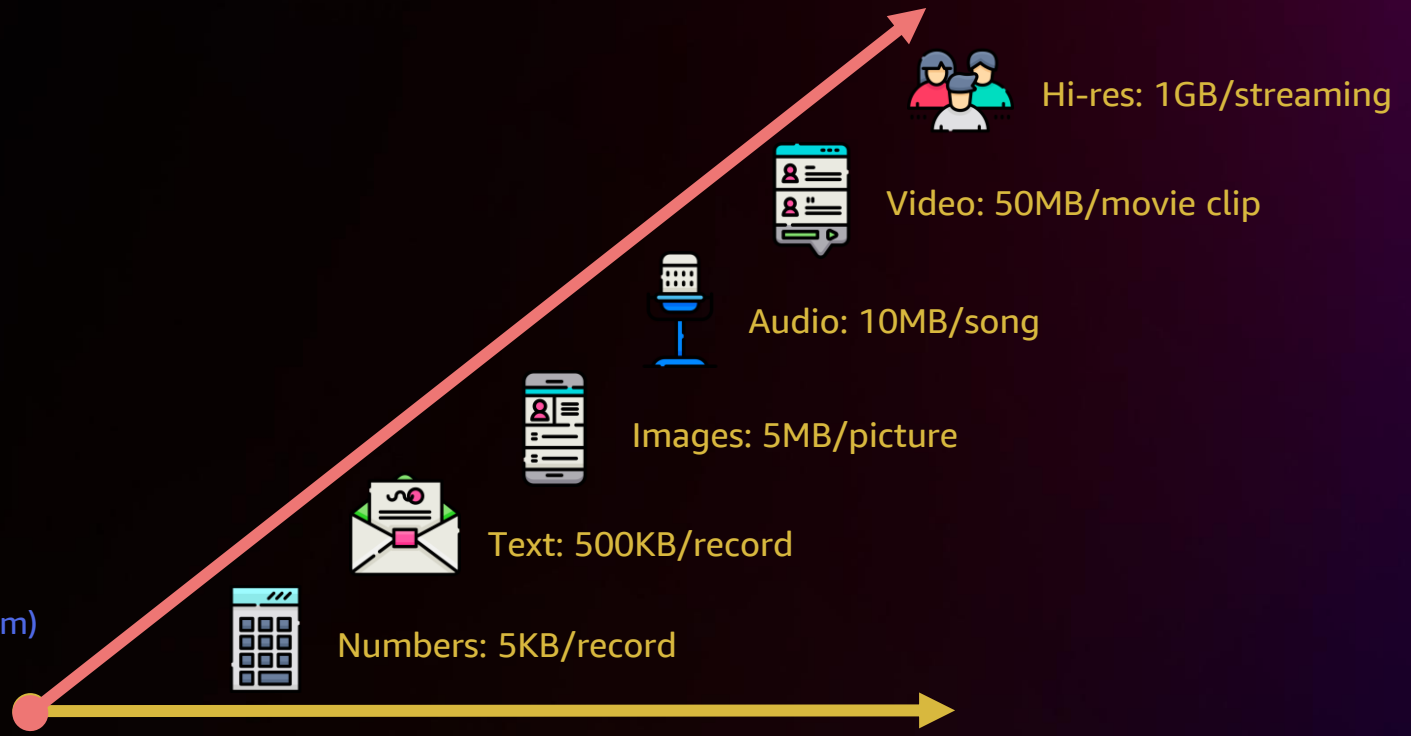
Agenda

- The unstructured data explosion
- The future of neural applications
- Going beyond single-modality neural networks
- Neural search using multimodal data types
- Using open source for neural applications
- Deploying neural networks on the AWS Cloud

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The unstructured data explosion

- Gartner estimates that 80% of enterprise data today is unstructured
- Unstructured data includes emails, files, PDFs, documents, images, and audio
- 80% of enterprises are unable to manage and utilize their unstructured data

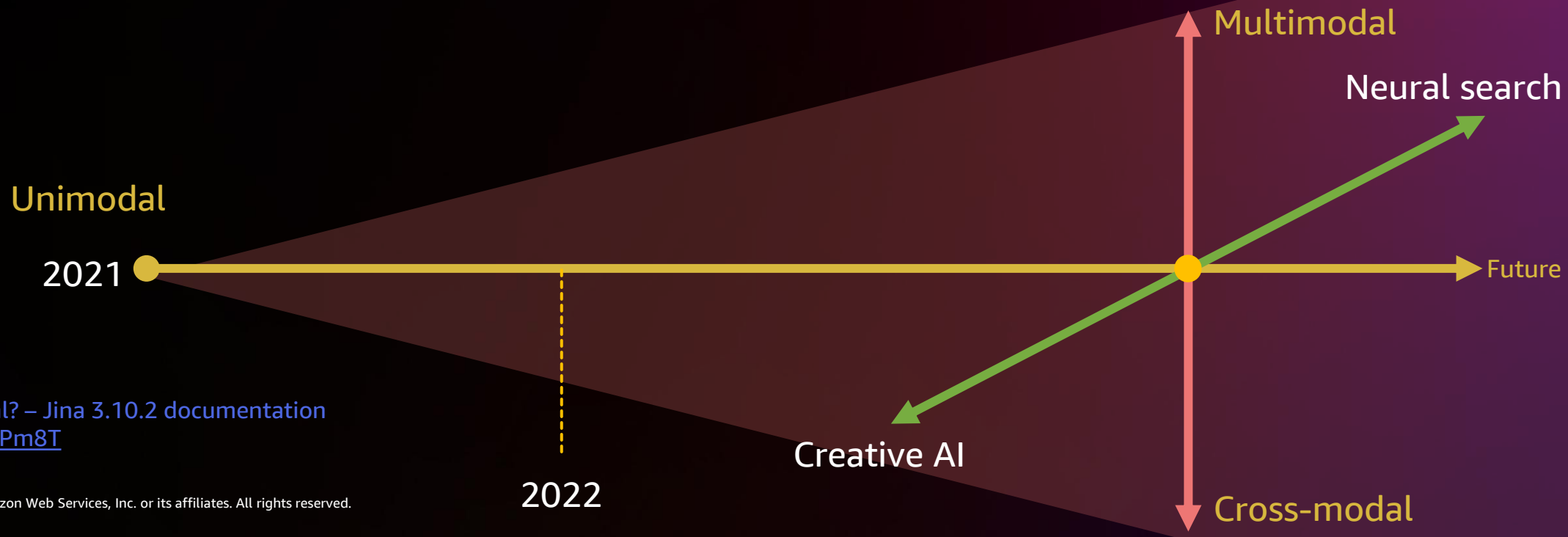


The Big (Unstructured) Data Problem (forbes.com)
https://bit.ly/3WJ9T6I

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The future of neural applications

- Unimodal: algorithms that process a single ”data type” or modality
- Multimodal: algorithms that learn from multiple modalities (image, audio, text)
- Cross-modal: algorithms that learn from neighboring modalities (text to image)



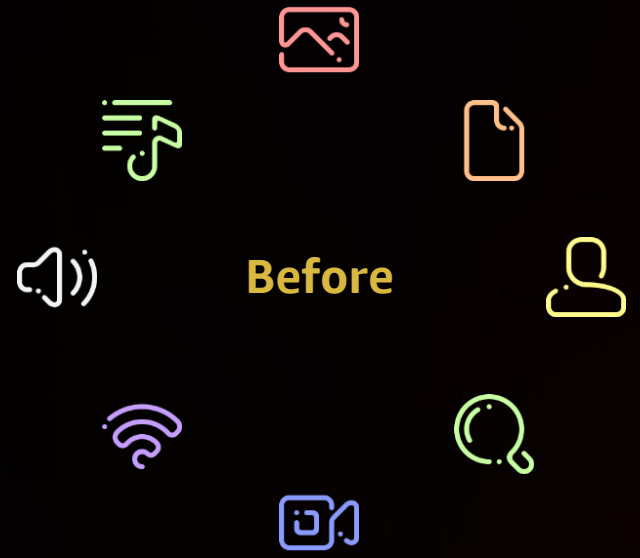
What is Multimodal? – Jina 3.10.2 documentation
https://bit.ly/3DMPm8T

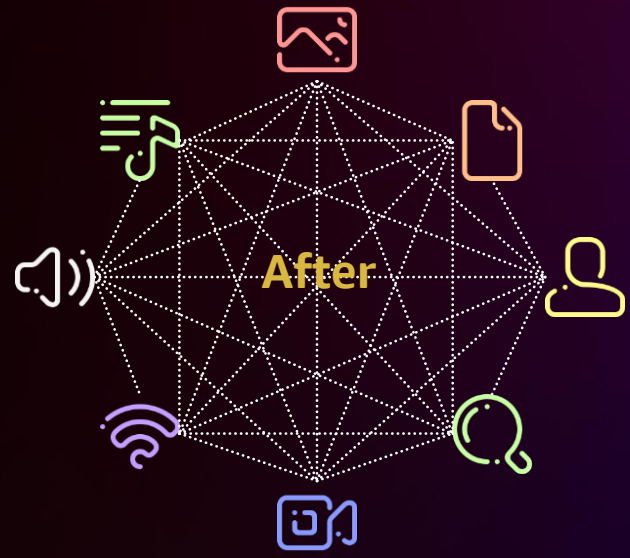
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Going beyond single modality neural networks

Challenge: Machine learning systems overspecialize on individual tasks, when they could excel at many; today's AI systems are often trained from scratch for each new problem – rather than extending existing models to learn new tasks

Solutions: Tasks are shared and transferred between modalities and knowledge is learned and applied to multiple modalities erasing boundaries of neural networks





© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is Multimodal? – Jina 3.10.2 documentation https://bit.ly/3DMPm8T

Neural search using multimodal data types



GitHub – https://github.com/AndreyGuzhov/AudioCLIP
AudioCLIP: Source code for models described in the paper "AudioCLIP: Extending CLIP to Image, Text and Audio" https://arxiv.org/abs/2106.13043

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Using open source for neural applications

- Jina.AI: open-source advanced MLOps platform for multimodal and creative AI
- Hugging Face: open-source, state-of-the-art, pre-trained deep learning models
- Linkerd: open-source service mesh for Kubernetes security and observability







Build

https://jina.ai/

Train

https://huggingface.co/

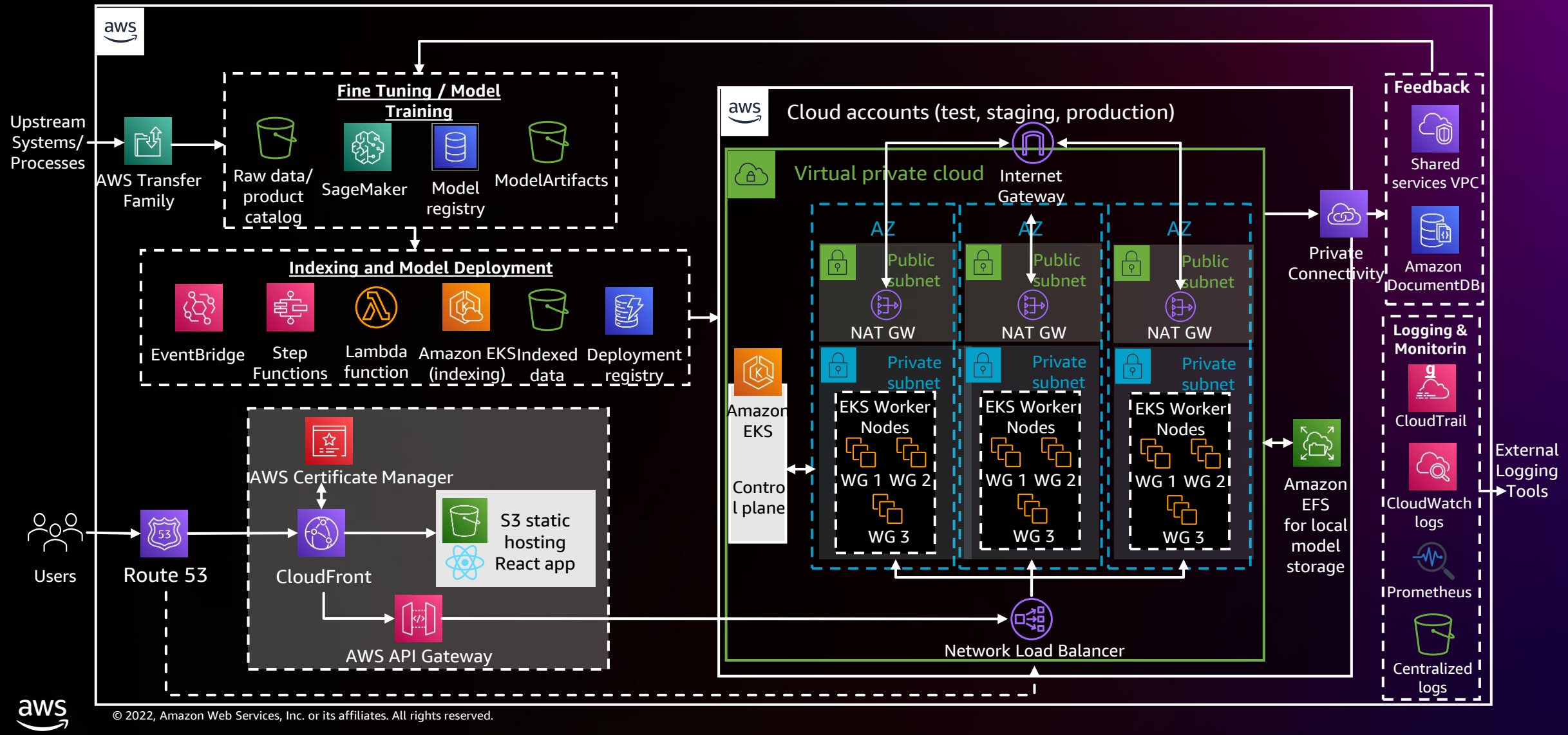
Monitor

https://linkerd.io/

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deploying neural networks on the AWS Cloud



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Leo Simberg
simbel@amazon.com

Najeeb Danish
ndanish@deloitte.com



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.