Generative AI models: Go big or go home

Mickey Greaves
Enterprise Sales Executive

Seth Kneeland
Customer Engineer

November 2023

Show of hands:

How many of you work with models today?

SambaNova SYSTEMS

SambaNova delivers
pre-trained foundation models,
on top of a complete and integrated
software/hardware stack,
purpose-built for AI,
delivered as-a-service

What's next:

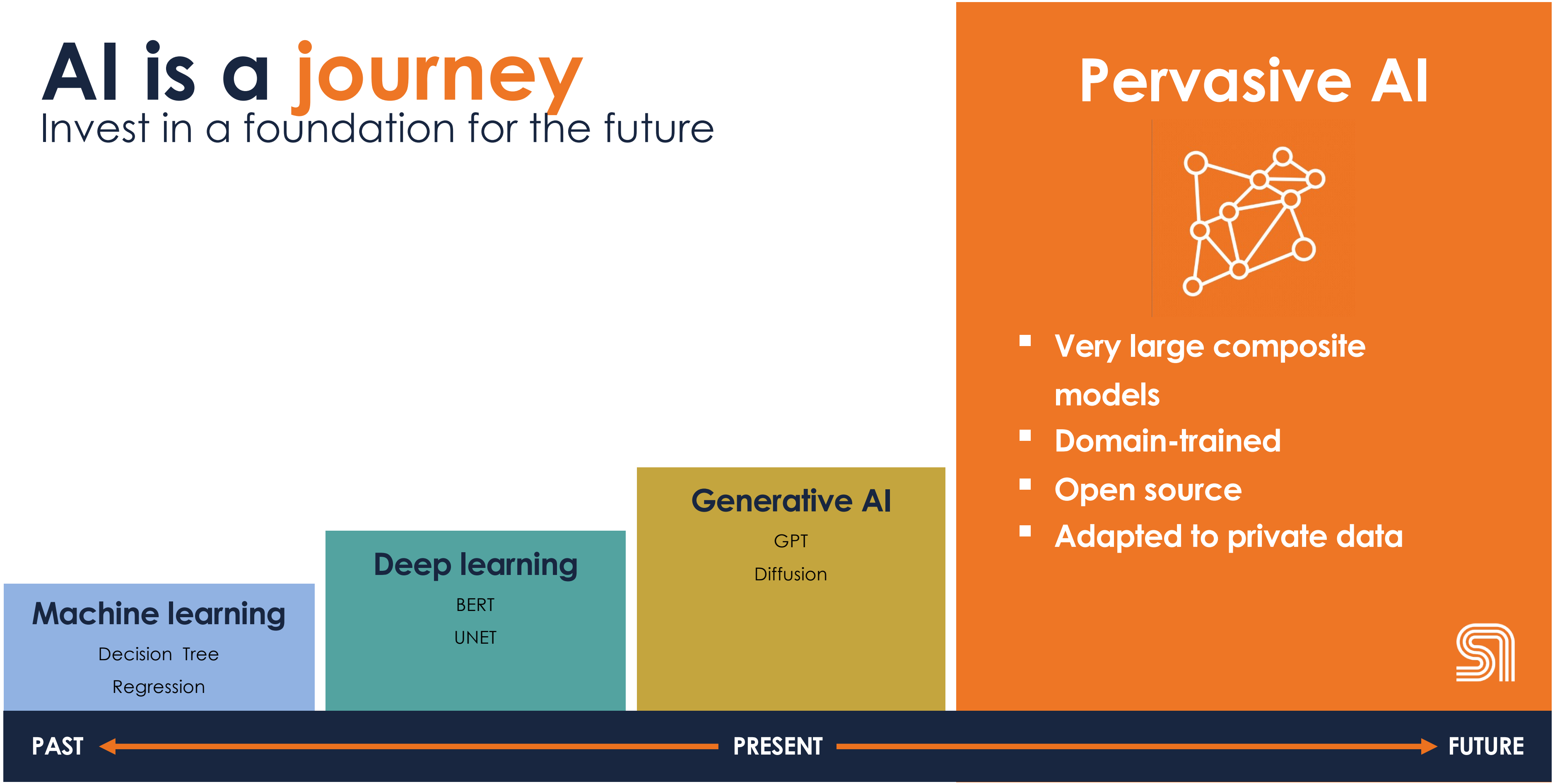
Keeping private data private,

owning your model, enterprise-thinking

And a demo

AI is a journey

Invest in a foundation for the future



Machine learning

Decision Tree

Regression

Deep learning

BERT

UNET

Generative AI

GPT

Diffusion

Pervasive AI

- Very large composite models
- Domain-trained
- Open source
- Adapted to private data

PAST ← PRESENT → FUTURE

Why go big?

Larger models provide better results

- Quality: Highest accuracy; fewer hallucinations
- Versatility: Greater breadth of domains, tasks, etc.
- Capability: Better reasoning

To date, 1T+ models have been out of reach

- Extraordinary compute needs
- Innate restrictions of legacy CPUs and GPUs
- Silicon supply chain constraints
- Necessity of move to smaller models

SambaNova Suite: Highest AI performance

- Breakthrough high-performance, high-efficiency hardware
- Composition of Experts (CoE) model architecture
- Integrated stack to drive highest performance
- Sparse, modular models to optimize training and inference
- Available, reliable, predictable compute costs

SambaNova SYSTEMS

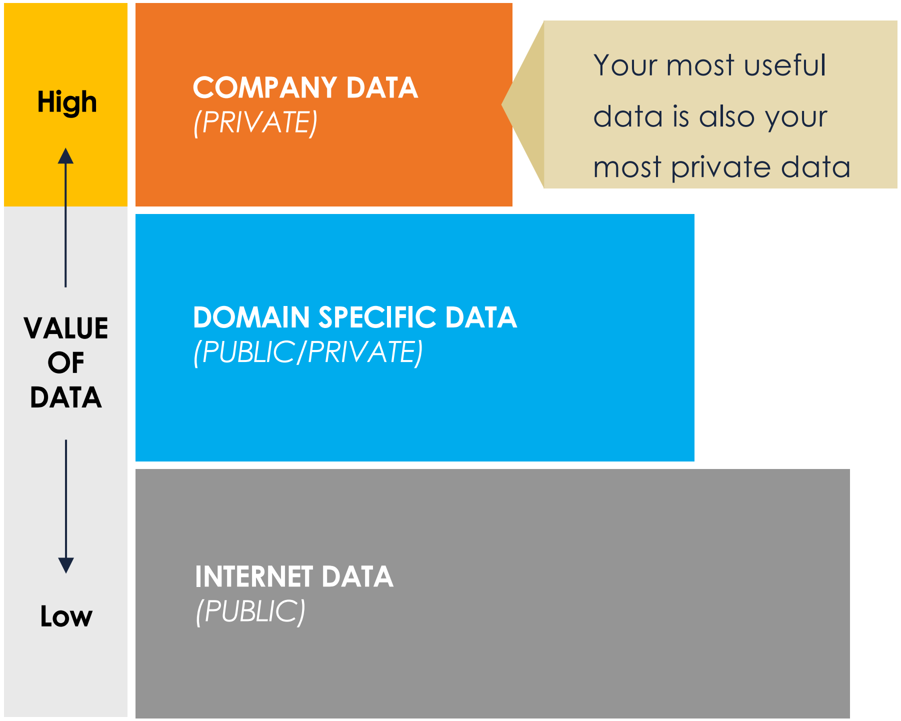
What’s next:

Keeping private data private?

Own your model.

SambaNova SYSTEMS

Own your model: Build an asset



High

VALUE OF DATA

Low

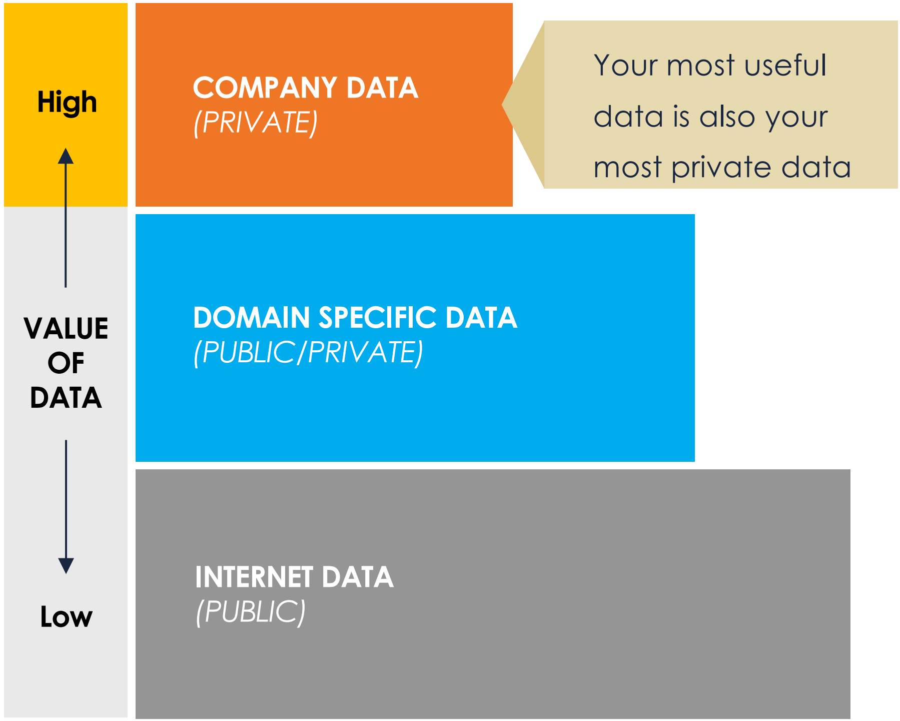
COMPANY DATA
(PRIVATE)

Your most useful data is also your most private data

DOMAIN SPECIFIC DATA
(PUBLIC/PRIVATE)

INTERNET DATA
(PUBLIC)

Own your model: Build an asset



High

VALUE OF DATA

Low

COMPANY DATA
(PRIVATE)

Your most useful data is also your most private data

DOMAIN SPECIFIC DATA
(PUBLIC/PRIVATE)

INTERNET DATA
(PUBLIC)

Accuracy

Securely fine-tune models so they are up to date and safe to use

Ownership

Retain ownership of any model adapted with private data

Privacy/Security

Complete control over what data is used, how the model is built, and how results are delivered

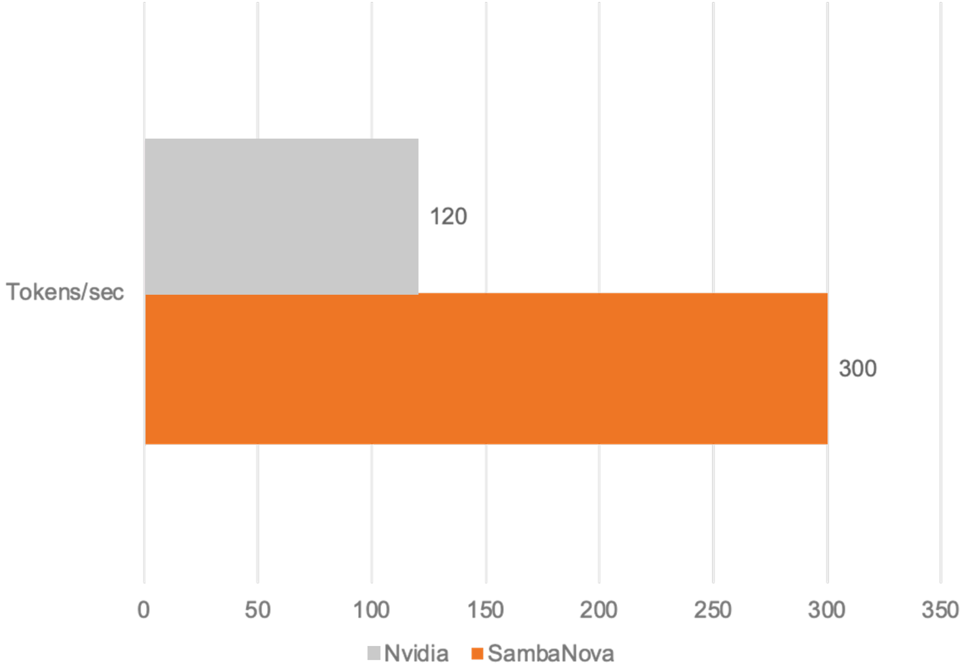
Regulation/ Compliance

Full explainability into weights and training methods

What’s next:

Enterprise thinking

Best-in-class 300 tokens/sec inference at 1T parameter scale

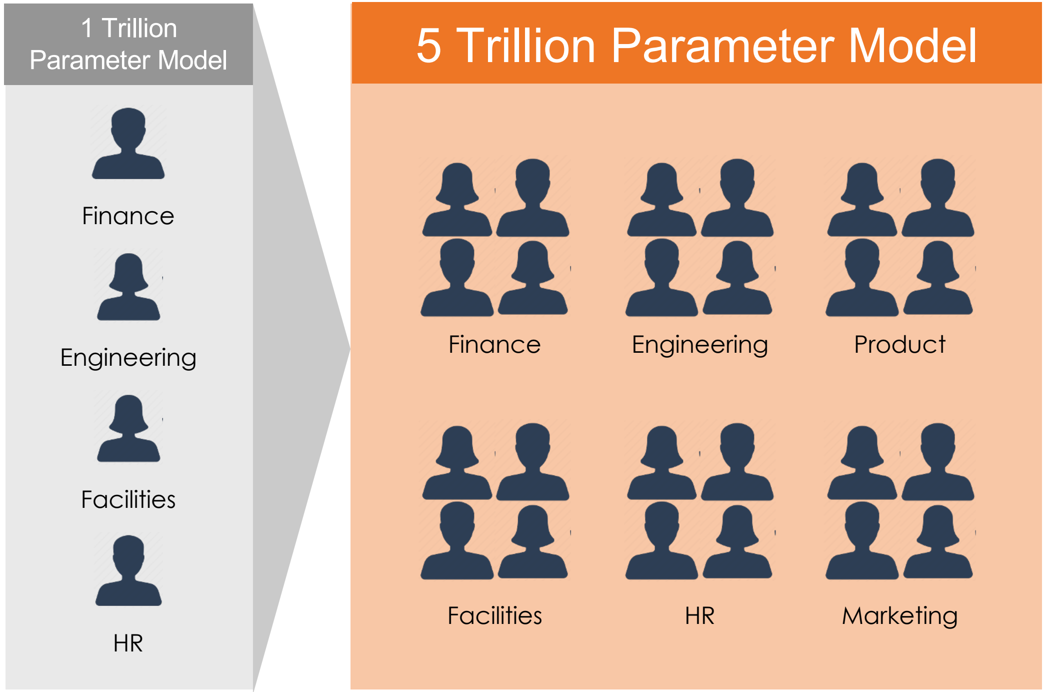


2.5x

higher inference throughput than NVIDIA DGX H100

Composition of Experts (CoE)
The new model architecture for enterprises

- Power: Built on state-of-the-art open-source foundation models
- Accuracy & relevance: Deep domain expertise from independently-trained expert models
- Modular: Models can be added, swapped, retrained
- Cumulative: Can be trained on new data, without sacrificing previous learning
- Scalable: Saves compute & latency
- Secure: Role-based access control & security isolation



Value of Composition of Experts

Best TCO for inference

Best TCO for running many expert models (up to 5T parameters worth) and up to 256k sequence length on a single node at 300 tokens/second throughput (with 7B parameter experts)

Modular training

Train expert-by-expert, meaning quicker time to value and ability to incrementally scale.

Avoids alignment tax

Train on new domains, new tasks, new languages, new modalities, without becoming worse at existing capabilities.

Granular access control

Ensure access control of information within the model, resulting in higher security. Eg. Accounting expertise should only be made available to Finance.

SambaNova® SYSTEMS

Why SambaNova?

	Off-the-shelf AI application	SambaNova Suite	Build-your-own (NVIDIA+)
Ownership	Vendor-owned model	Customer-owned model	Buy or build your model
Privacy/Security	Shared, general model	Securely and privately adapted	Model-dependent
IP	IP ambiguity	Own your IP	Model-dependent
Model Visibility	Opaque: Training data and methodology is hidden	Full explainabiity, visibility into training data and methodology	Model-dependent
Model Portability	Vendor lock-in; no model portability	Full model portability; no vendor lock-in	Model-dependent
Version Management	Version-dependent	Continuous fine-tuning & dynamic management	Model-dependent
Cost	Pay-as-you-go	28x lower TCO than BYO	High effort: Tech + people + models
Availability	Cloud-based offering	Available now!	Supply chain constraints

SambaNova delivers
pre-trained foundation models,
on top of a complete and integrated
software/hardware stack,
purpose-built for AI,
delivered as-a-service

Demo

Q&A

SambaNova.ai

Thank you

We look forward to hearing from you:

Mickey Greaves
mickey.greaves@sambanova.ai
347 995 5153

Seth Kneeland
seth.kneeland@sambanova.ai

Go big or go home: https://sambanova.ai/resources