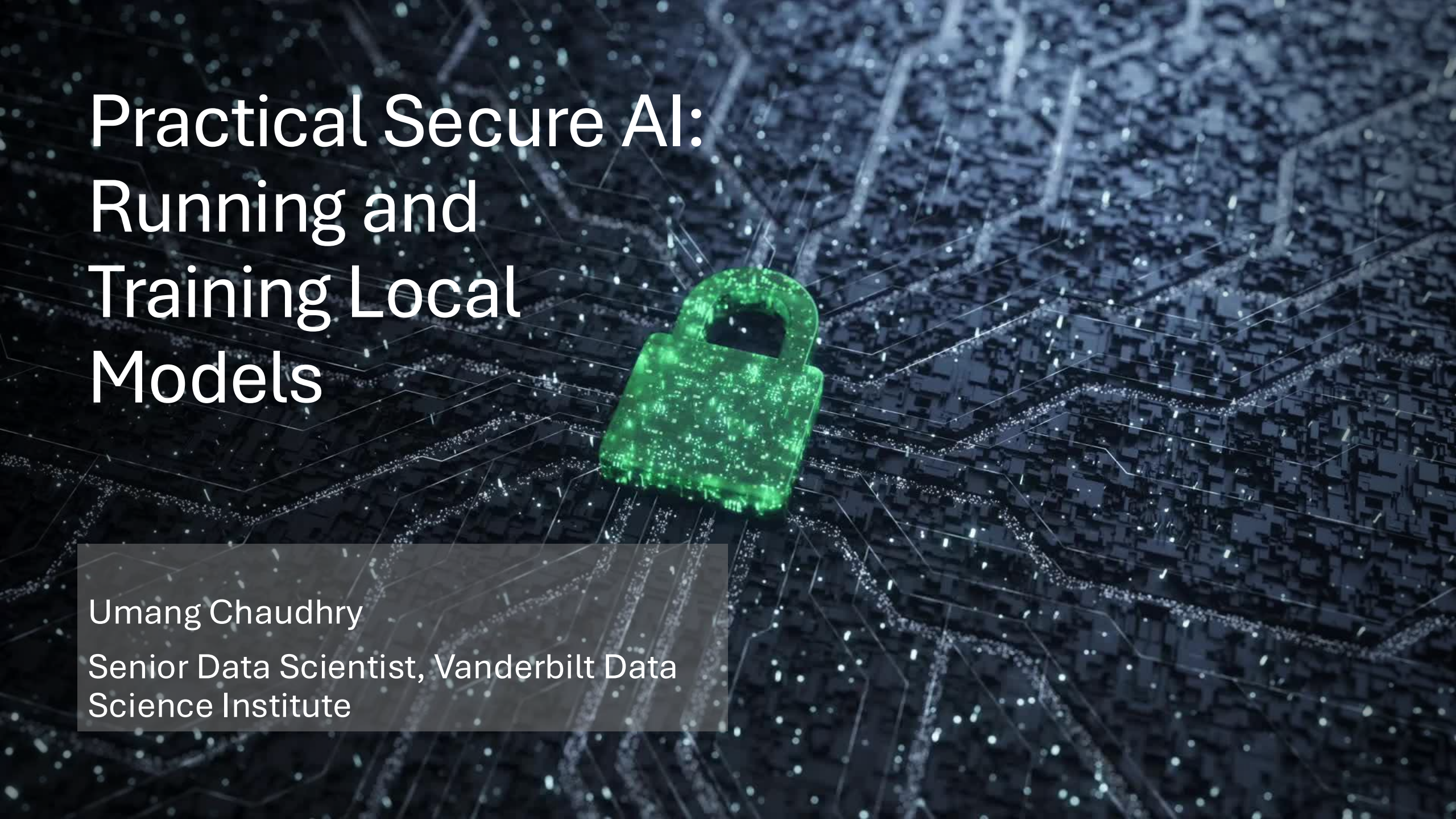
Practical Secure AI: Running and Training Local Models



Umang Chaudhry

Senior Data Scientist, Vanderbilt Data Science Institute

Agenda

- Working with secure data

- Running, Training and Hosting local models
  - HuggingFace
  - LM Studio (Apple Silicon Macs, Windows, Linux)
- Anything LLM (All Machines) – Lightweight UI-Based RAG implementations (not local)

Downloads

- LM Studio - https://lmstudio.ai
- Anything LLM - https://anythingllm.com
- Hugging Face API Key - https://huggingface.co/settings/profile

Secure Solutions

- By default, conversations with ChatGPT may be used for future training
- Any requests through the API to chat models and embedding models may also be used to train new models
- When working with secure data, alternate approaches are required
  - Enterprise contracts
  - Open-source models

Choosing the right solution

Per Vanderbilt Office of Cybersecurity

	Non-Sensitive	Sensitive		
Classification	Level 1 Public	Level 2 Institutional Use Only	Level 3 Restricted	Level 4 Critical
Description	Intended for public release or distribution.	Private to the institution and should not be available to individuals without permission.	Confidential by law or contract or should not be shared with unauthorized persons.	Confidential by law or contract and requires bespoke security requirements.
Examples	Publicly available datasets	Unpublished research data, institutional documents	NDA, Protected Health Information (PHI, HIPPA), FERPA, GDPR	Classified information, law enforcement records

https://www.vanderbilt.edu/cybersecurity/guidelines/data-classification/

Data Classification	AI Solution
Level 1 Public Use	All Chat and Embedding models through API and UI All Training solutions (Local, OpenAI Fine Tuning API)

Level 2 Institutional Use Only	Varies by needs and preferences
	• Opting out of data use for training
	  • ChatGPT UI
	  • OpenAI Team Account
	• Institutional Access/Enterprise Contracts
	  • vanderbilt.ai
	• Open-Source models
	• Data stored on ACCRE or other approved storage solutions
	• Local or hosted Vector Stores
	• Training on ACCRE, DGX, Google Colab

Level 3 Restricted	• Access to closed source models only through enterprise contracts • Open-Source models • Data stored on ACCRE or other private servers • Local or approved Vector Stores • Training on ACCRE, private servers
Level 4 Critical	Open-Source or Custom Models, Custom Data Solutions

Limitations of working locally

- Most open-source LLMs that you can run locally are not as good as state-of-the-art models like O1 and Claude 3.7, and proprietary models are almost never available to run locally (exception – DeepSeek R1)
- You are limited by your hardware – running or training full parameter large LLMs requires multiple GPU nodes, if not multiple racks
  - Solution?
    - Enterprise contracts through AWS or Azure giving you a secure instance for popular proprietary LLMs
    - Use smaller parameter fine-tuned models
    - Use distilled versions of larger models
    - Use larger models with lower quantization
    - PEFT

Running and hosting models locally

- HuggingFace – programmatic approach for inference and training
  - Run locally/Colab

- LM Studio, Anything LLM – UI approach for inferencing and creating RAG infrastructures