

Lucida ~~Sirius~~ and DjNN Tutorial

Speakers: Johann Hauswald, Michael A. Laurenzano, Yunqi Zhang

Organizers: Johann Hauswald, Michael A. Laurenzano, Yunqi Zhang,
Lingjia Tang, Jason Mars



Before We Begin...

- Conference wifi
 - SSID: HPCA_PPoPP_CGO
 - Password: BARCELONA2016
- Suggested prerequisites
 - VirtualBox
 - Chrome
 - VM with demo software
 - lucida-djinn-tutorial.tgz (USB sticks)
 - Hardware resources — ~6GB RAM, 50GB disk

Schedule

9:00-9:15 — Welcome

Section 1 - Lucida

9:15-9:30 — Introduction to Lucida

Section 2 — Lucida Suite

9:30-9:45 — Core Algorithmic Components of IPAs

9:45-10:00 — Hands-on: Lucida Suite

Section 3 - DjNN and Tonic

10:30-10:45 — Deep Learning in Intelligent Web Services

10:45-11:00 — Tonic Suite: Background and Hands-on

Section 4 - LucidaEco

11:00-11:15 — Introduction to LucidaEco

11:15-12:00 — Hands-on: Build Your Own IPA

Lucida and DjNN Tutorial

Topic: Introduction to Lucida

Speakers: Johann Hauswald



Rise of the Wearables

40%



\$80bn

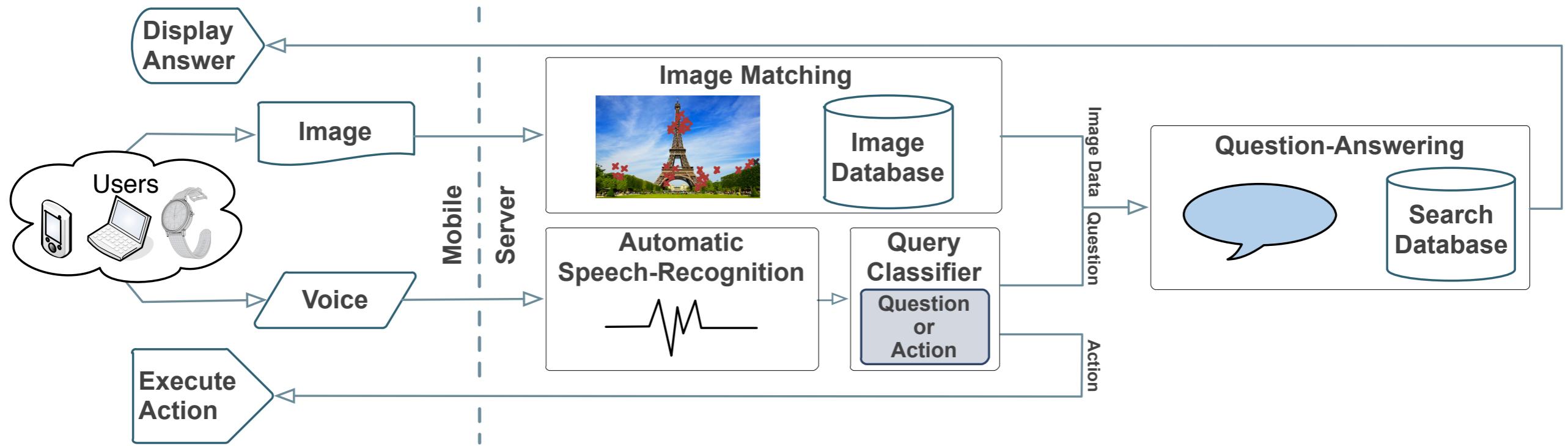
Questions Arise

- What are the impacts on our data centers?
- Can we scale to millions (or billions) of users?
- How do we design the right systems?



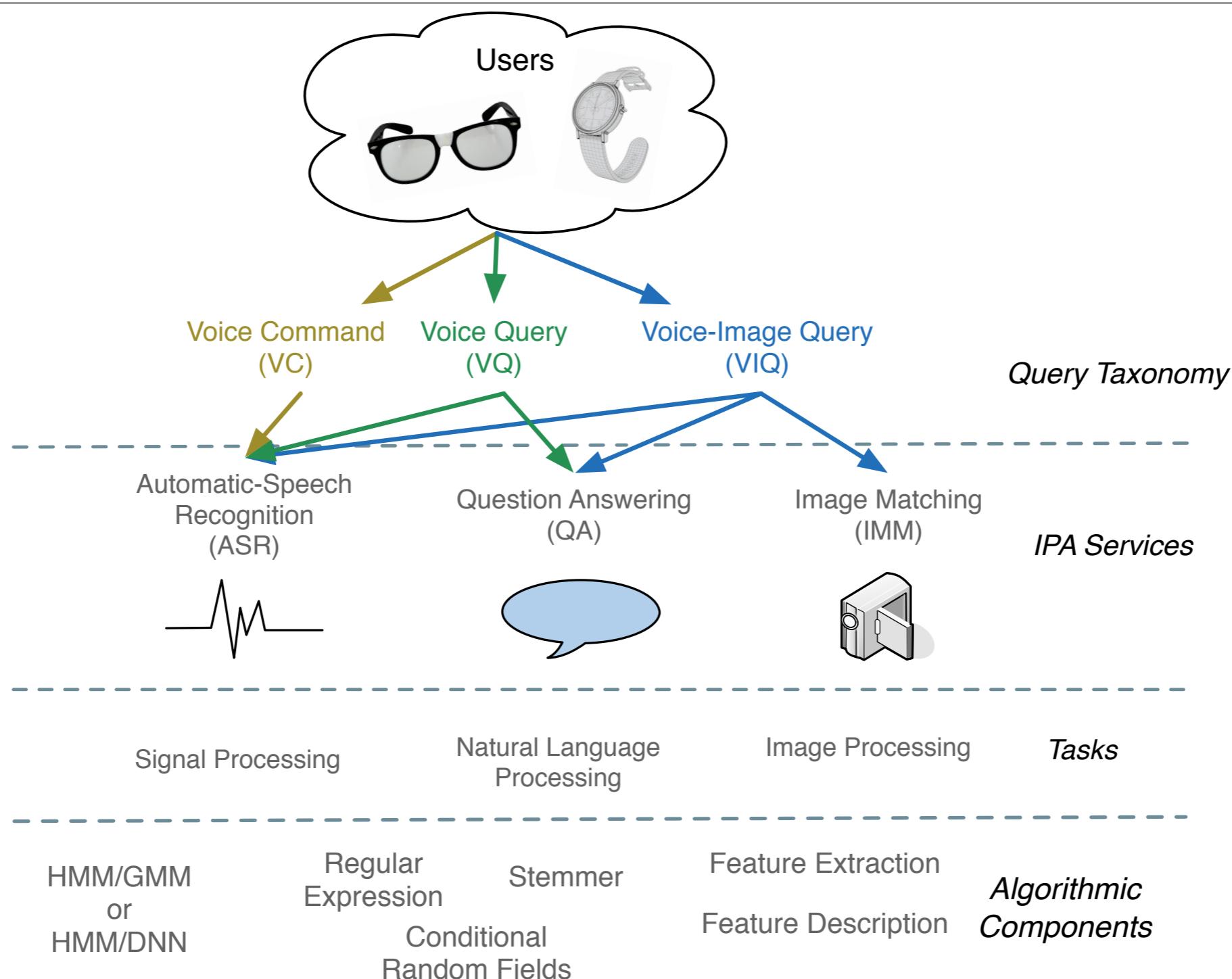
The Hard Place: Build an end to end intelligent personal assistant.
Sirius [ASPLOS' 15]

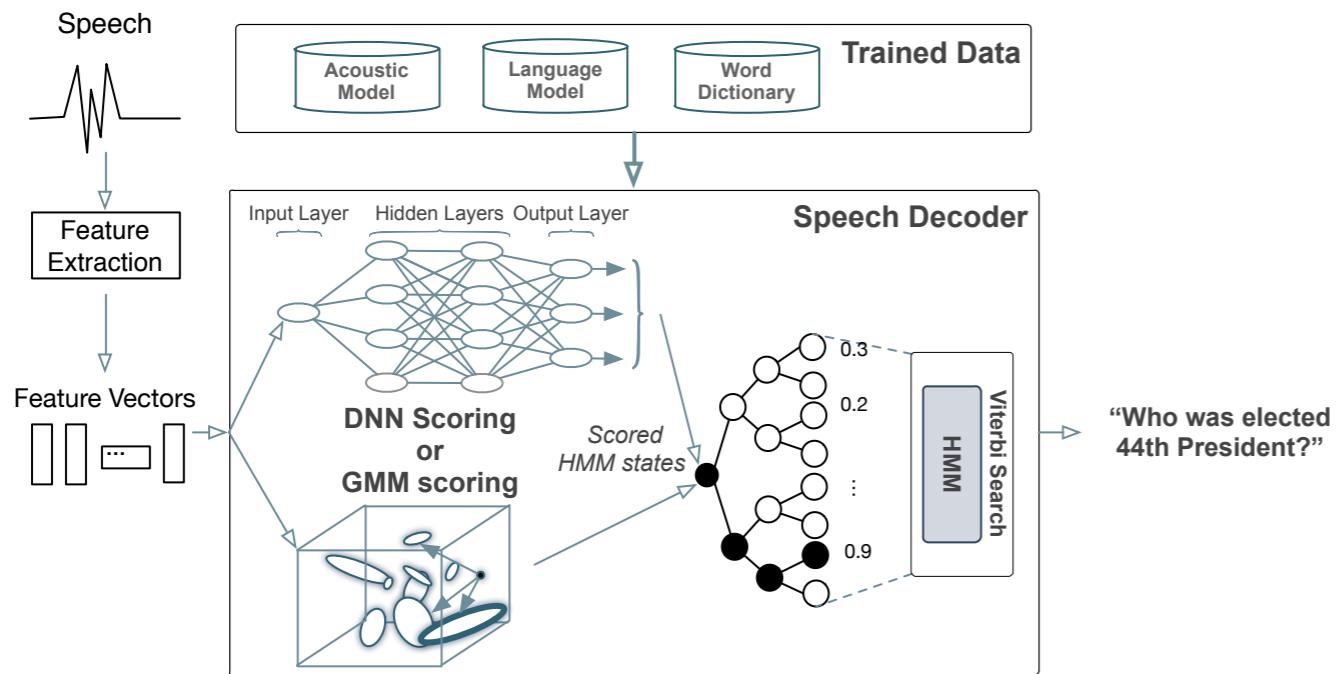
Lucida and Djinn Tutorial



Lucida Hierarchy

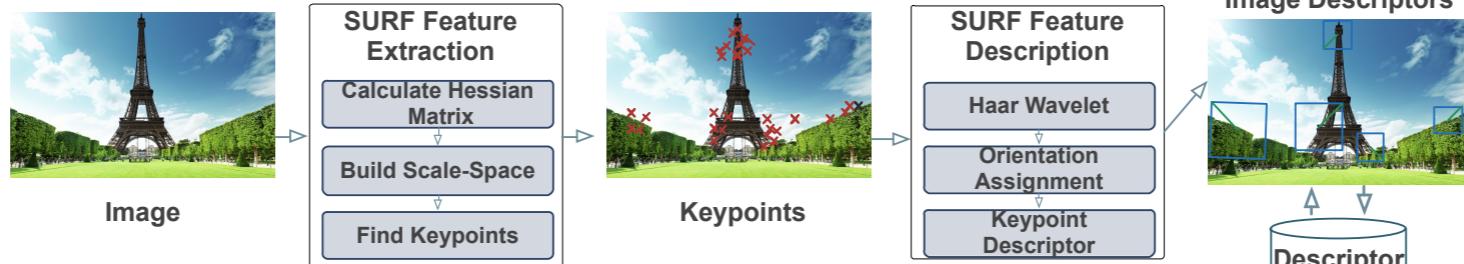
Lucida Hierarchy





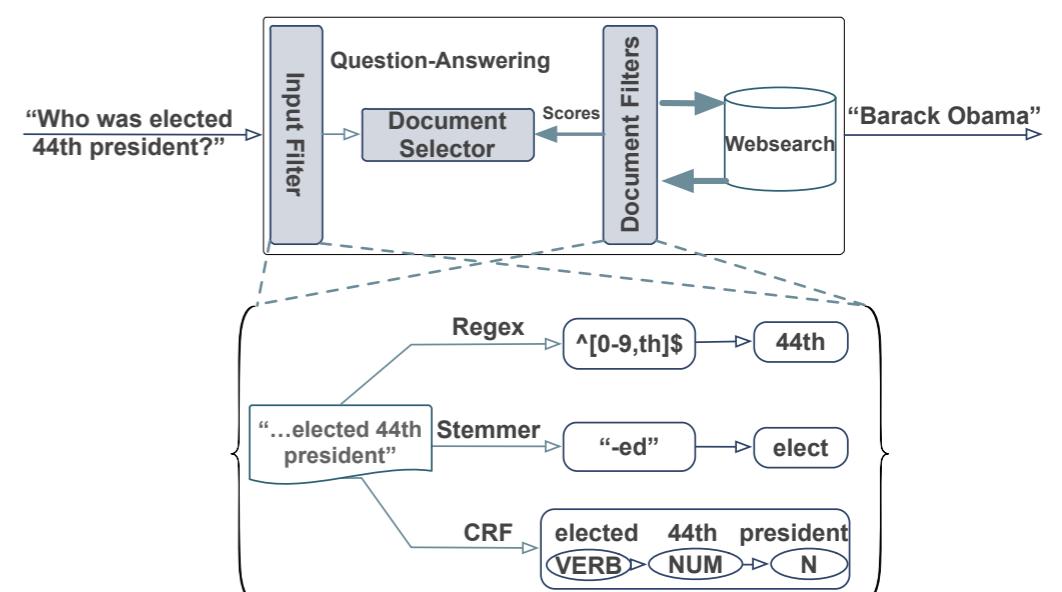
ASR

IMM



Three Horsemen of Lucida

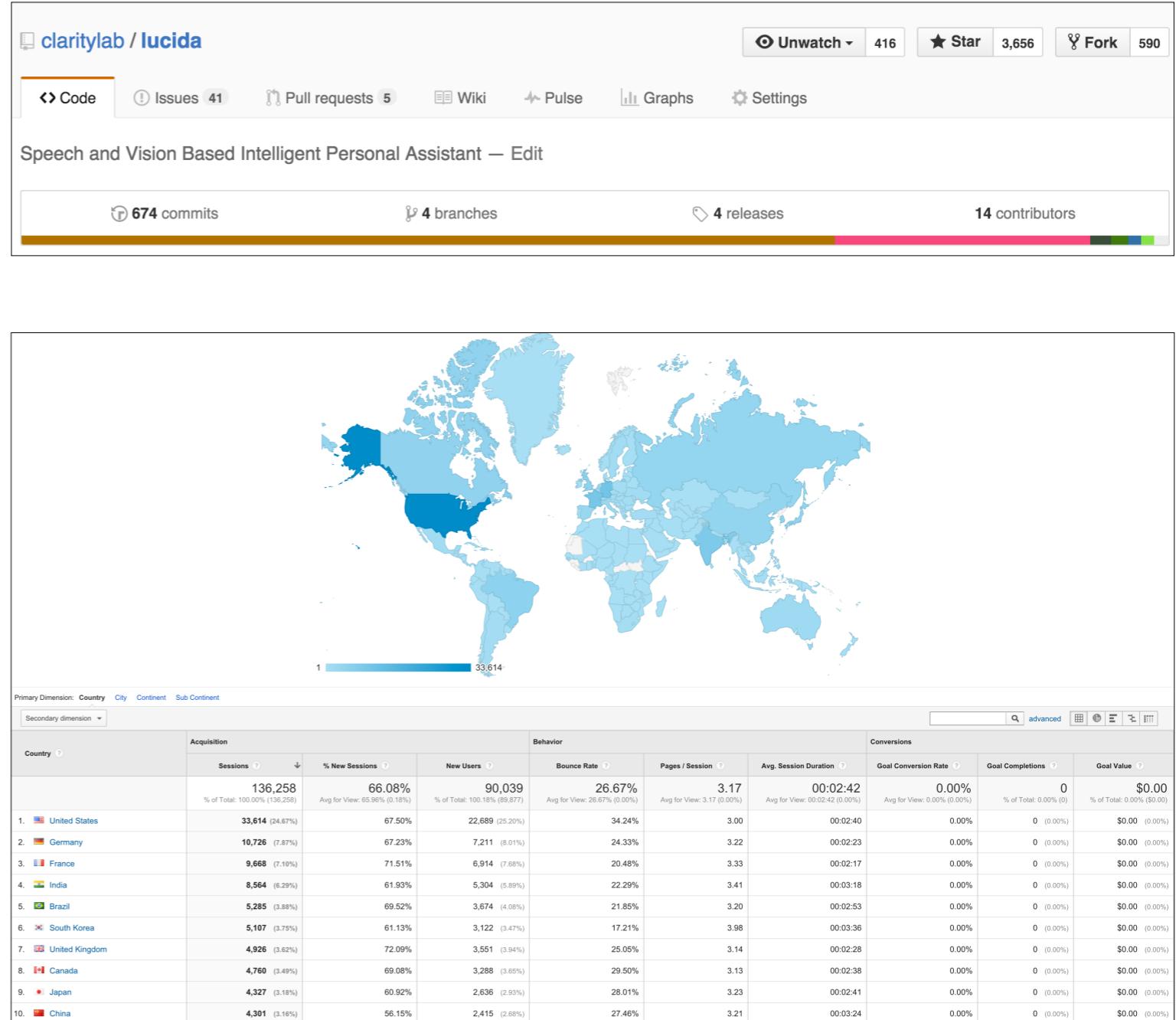
QA



Lucida and DjINN Tutorial

Project Status

- Recent Publication
 - BayMax [ASPLOS '16]
- Open source community
 - lucida-ai.slack.com
- Student projects
 - Independent studies
 - Summer interns

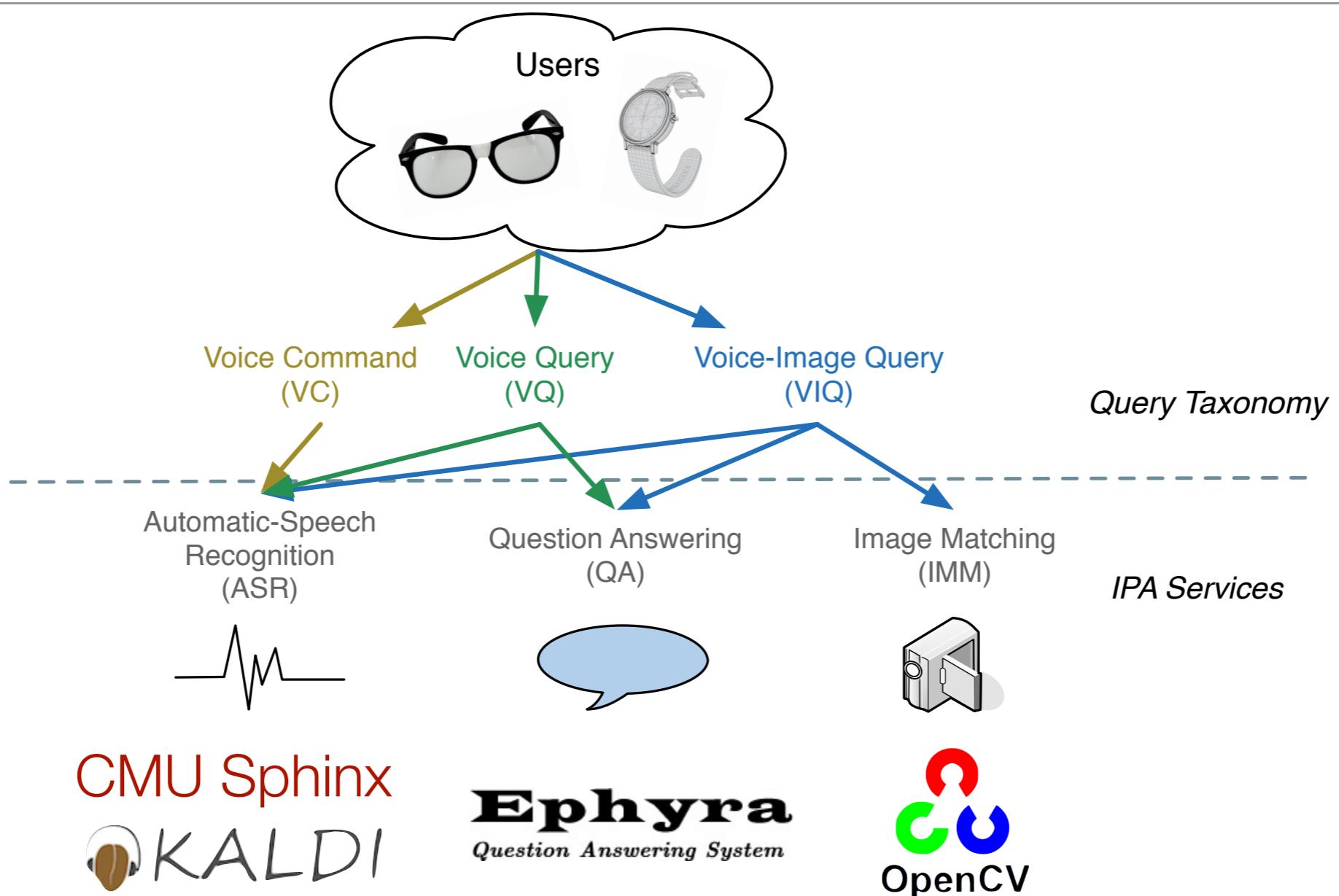


Lucida and Djinn Tutorial

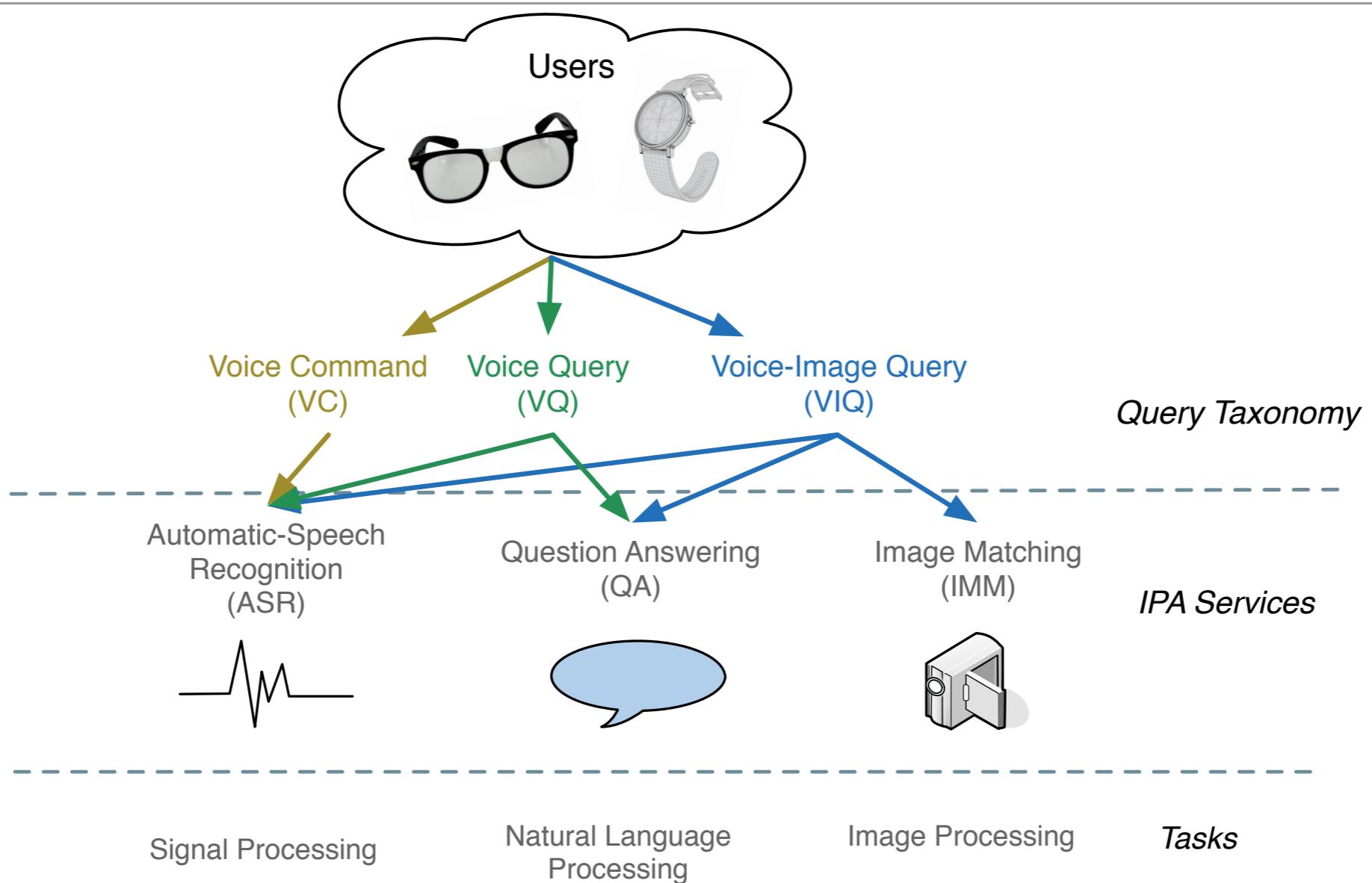
Topic: Core Algorithmic Components of IPAs

Speaker: Yunqi Zhang

How does Lucida work?



How does Lucida work?



How does Lucida work?

Signal Processing	Natural Language Processing	Image Processing	<i>Tasks</i>
Gaussian Mixture Model (GMM) or Deep Neural Network (DNN)	Regular Expression	Stemmer	Feature Extraction
85% or 78%	Conditional Random Fields	Feature Description	97%



7 kernels: 92% of total execution of Lucida

What is Lucida Suite?

- Suite of 7 workloads extracted from the end-to-end Lucida intelligent personal assistant
- Suite includes:
 - Standalone workloads
 - Pretrained models (when applicable)
 - Inputs for all workloads
- Available at:

<https://github.com/claritylab/lucida>

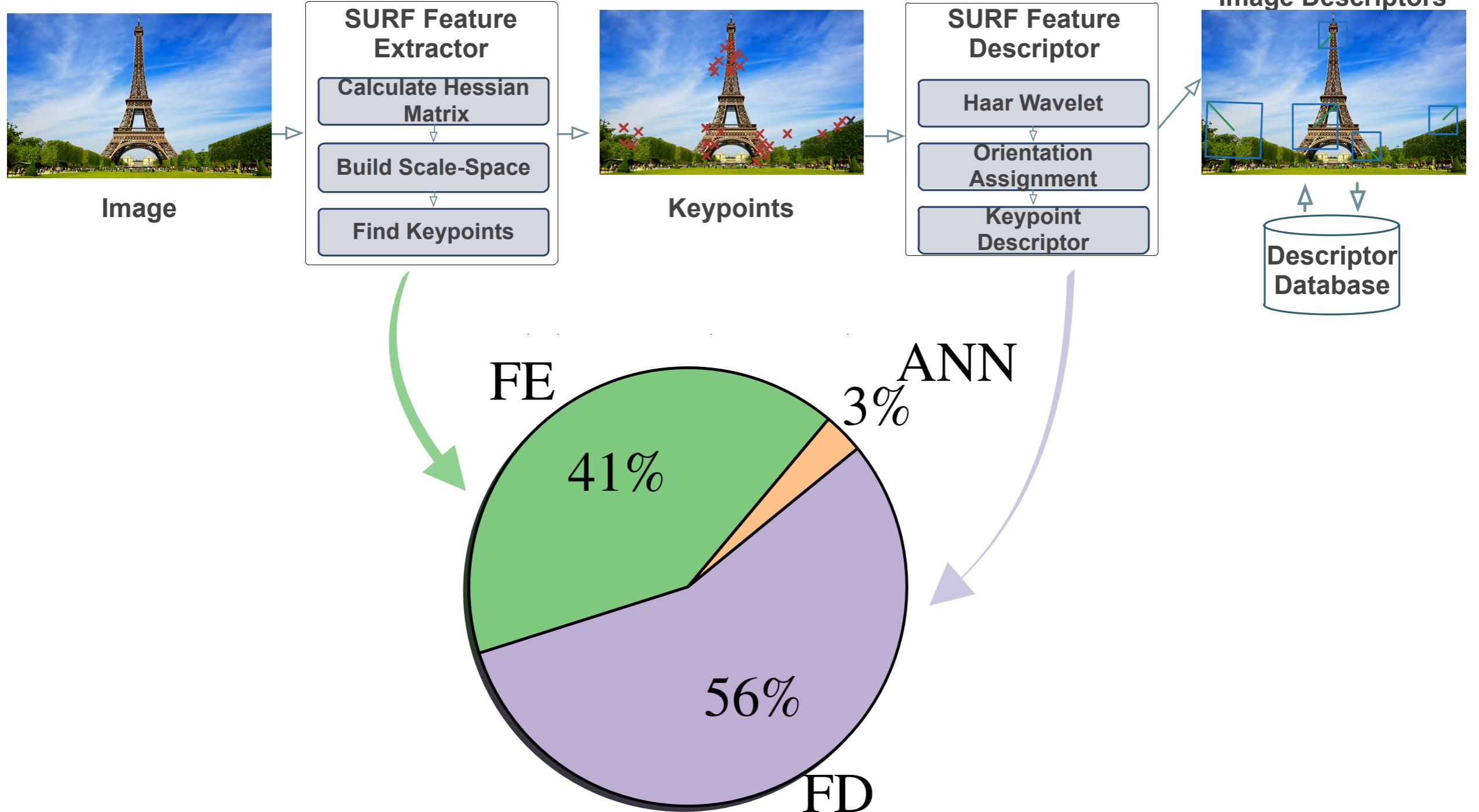
Objectives of Lucida Suite

- Represent emerging datacenter workloads
 - speech recognition, computer vision, natural language processing
- Cover key components in the application
 - Suite represents 92% of the total cycles
- Implementations
 - Single thread CPU
 - Pthread CPU
 - GPU
- Easy to use
 - Implemented in C/C++ and CUDA

Lucida Suite

Service	Workload	Task	Platforms
Image Matching	Feature Extraction	Computer Vision	Pthread, GPU
	Feature Description	Computer Vision	Pthread, GPU
Automatic Speech Recognition	GMM Scoring	Speech Recognition	Pthread, GPU
	DNN Scoring	Speech Recognition	Pthread, GPU
Question Answer	Stemmer	Natural Language Processing	Pthread, GPU
	Regular Expression	Natural Language Processing	Pthread
	Conditional Random Fields	Natural Language Processing	Pthread

Image Matching

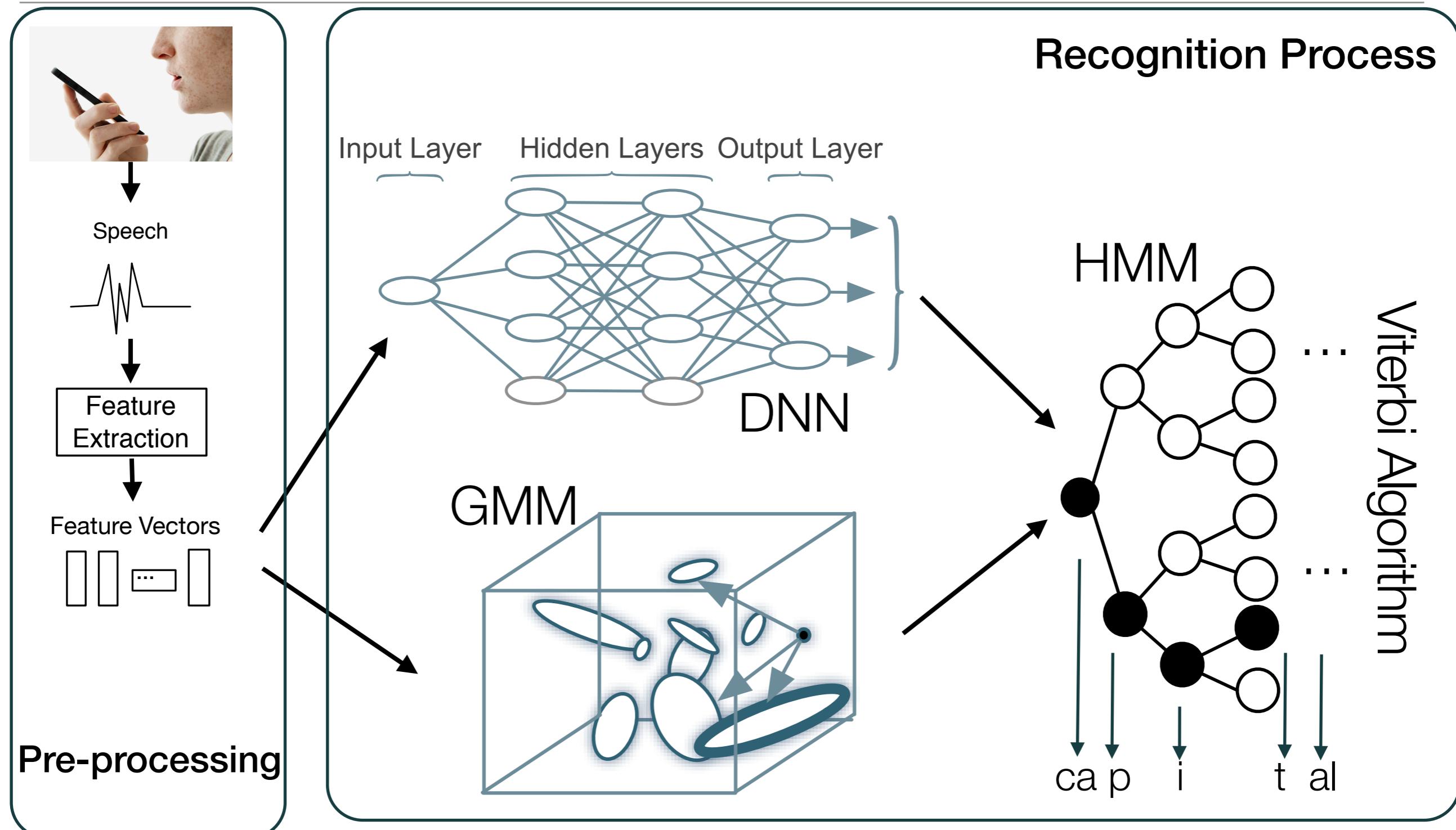


Accelerated workload: 97%

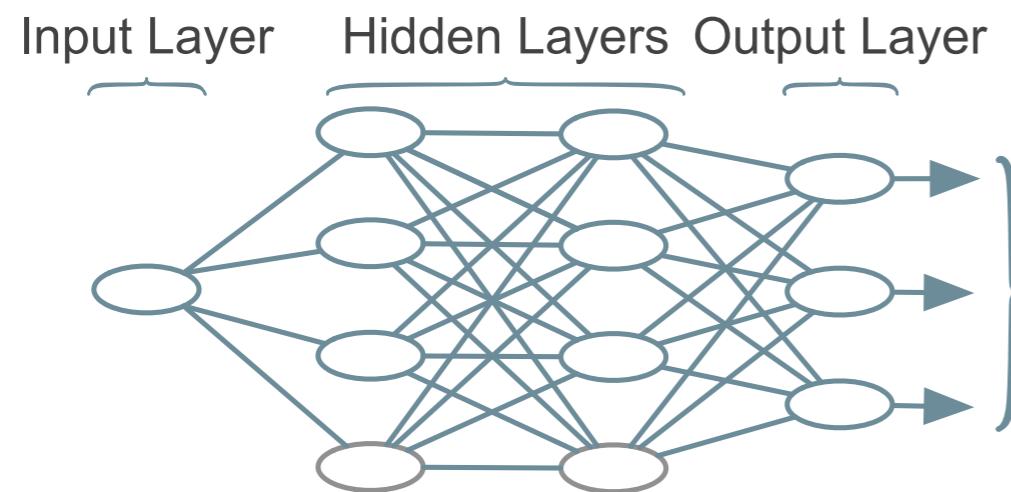
Automatic Speech Recognition



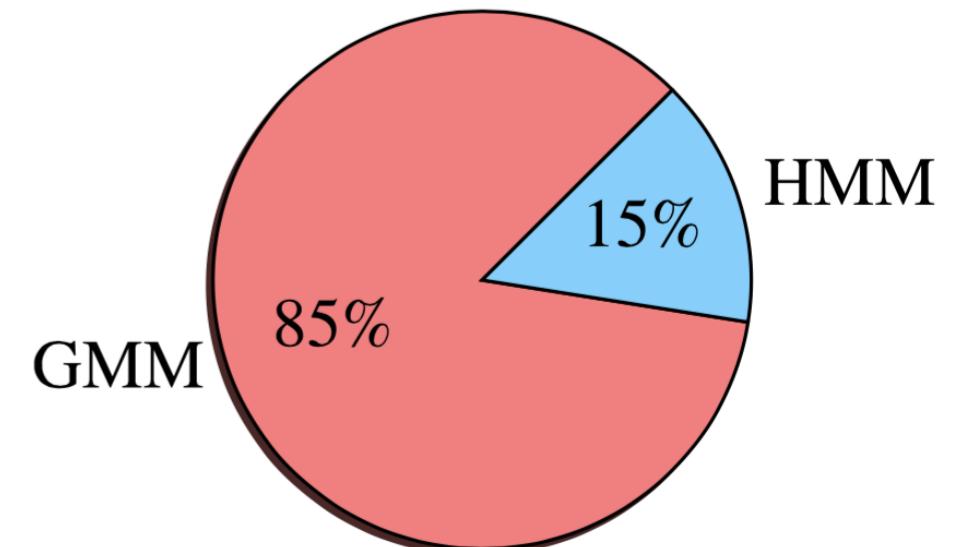
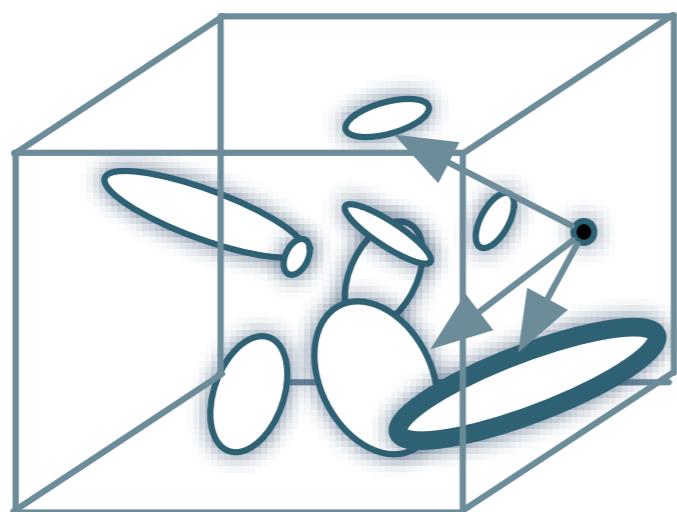
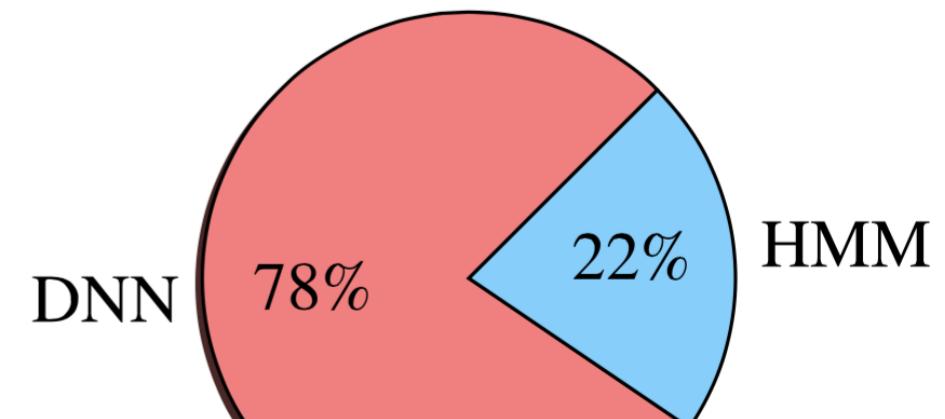
Automatic Speech Recognition (ASR)



Automatic Speech Recognition (ASR)



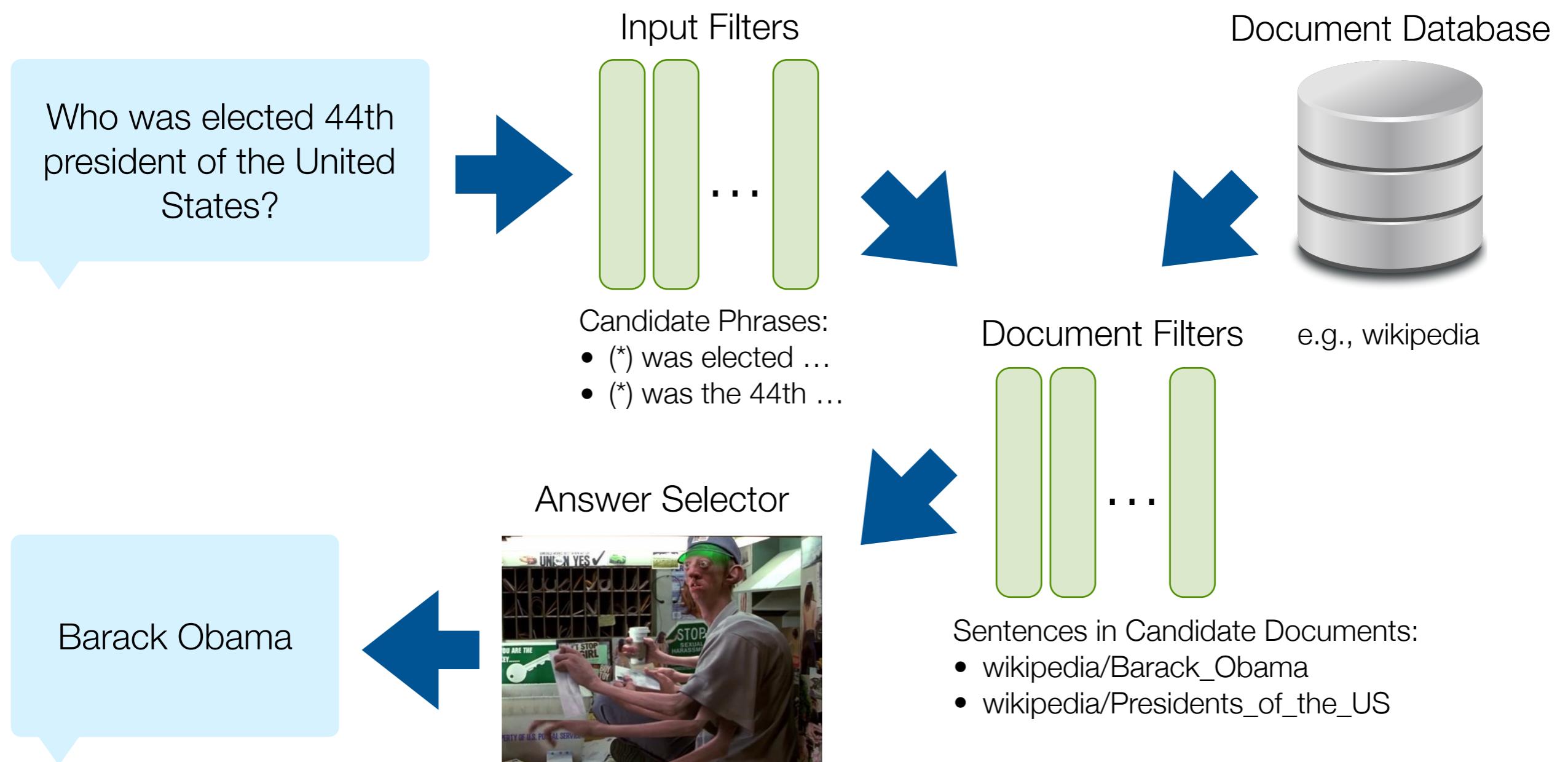
Recognition Process



Question Answering



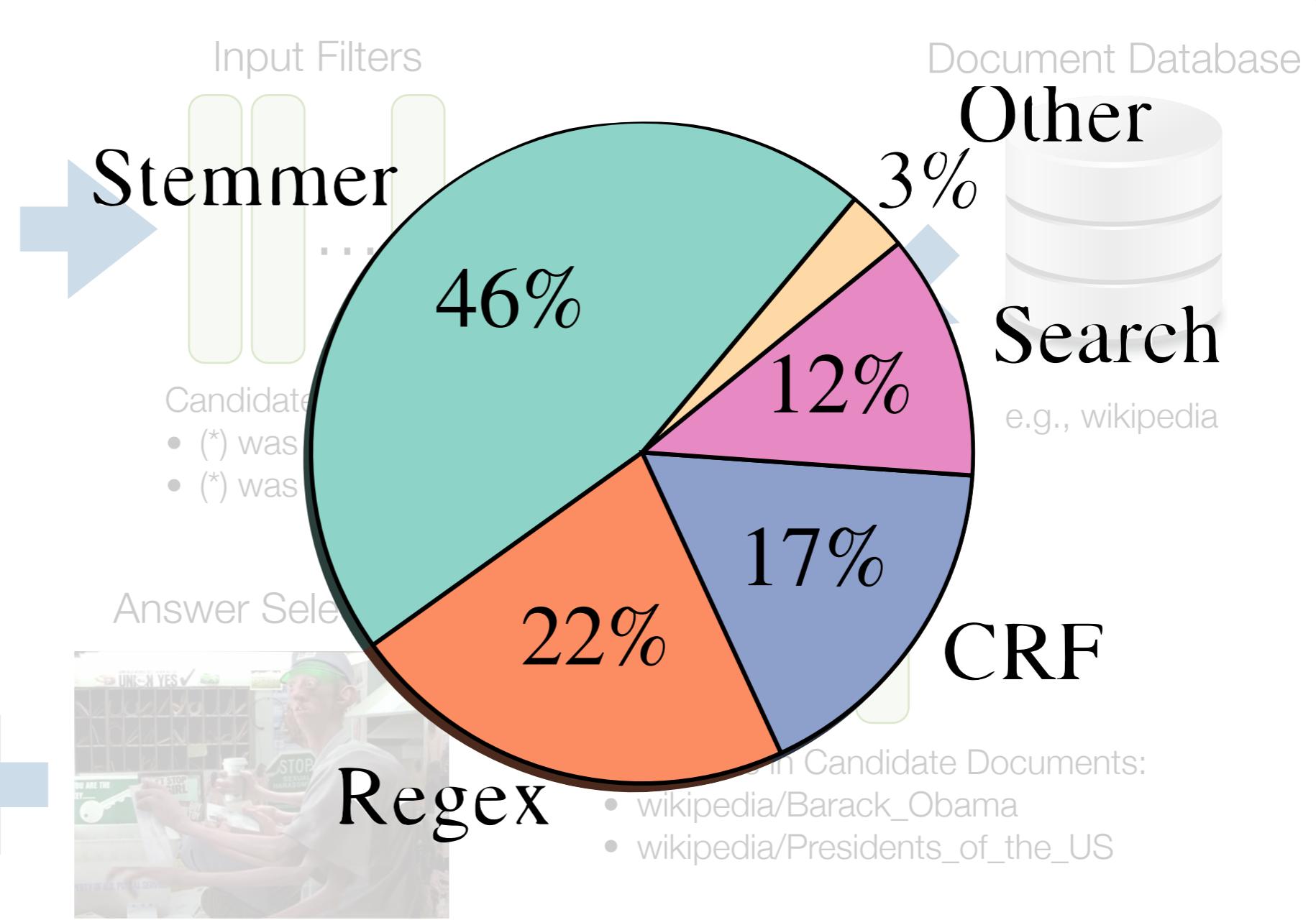
Question Answering



Question Answering

85% of the execution

- Conditional Random Fields
- Stemmer
- Regular Expression



Lucida and DjNN Tutorial

Topic: Hands-on: Lucida Suite

Speaker: Yunqi Zhang

Lucida Suite

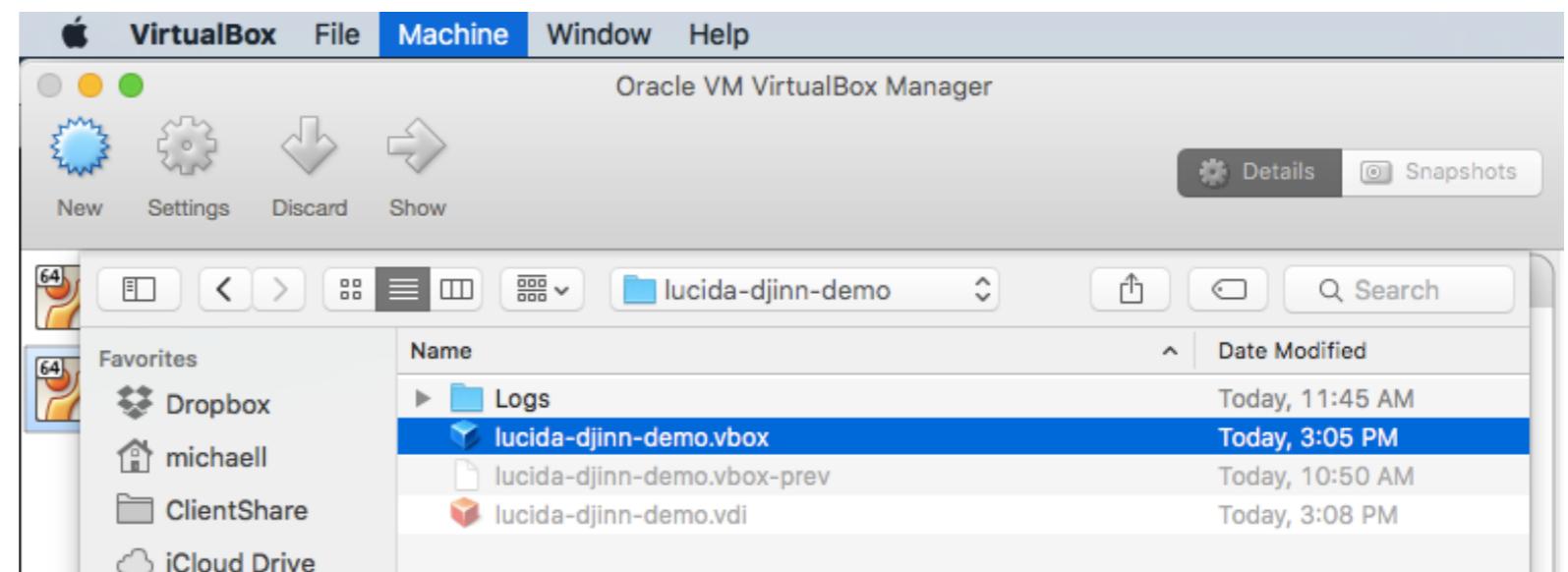


Setup Prerequisites

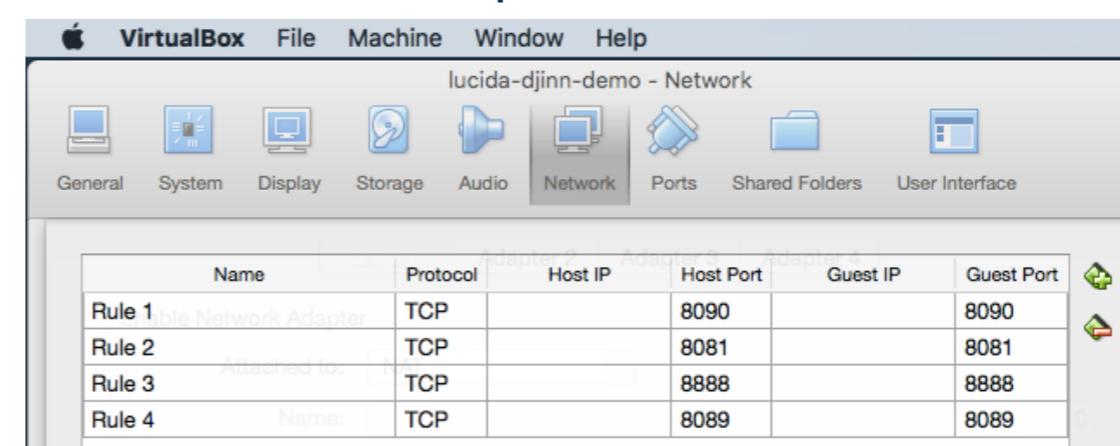
- Install VirtualBox
- Install Chrome
- Copy the tutorial VM – USB thumb drive

Configure VirtualBox

- Unpack lucida-djinn-demo.tgz into ~VirtualBox VMs/
- Add lucida-djinn-demo to VirtualBox
 - Machine -> Add



- Port forwarding
 - Machine -> Settings -> Network -> Adapter 1 -> Advanced -> Port Forwarding



Stand up the VM

- Start!
- username/password: demo/demo
- open Terminal application
- cd ~/h pca-demo/source-code/lucida/lucida-suite
- Follow along with Yunqi!

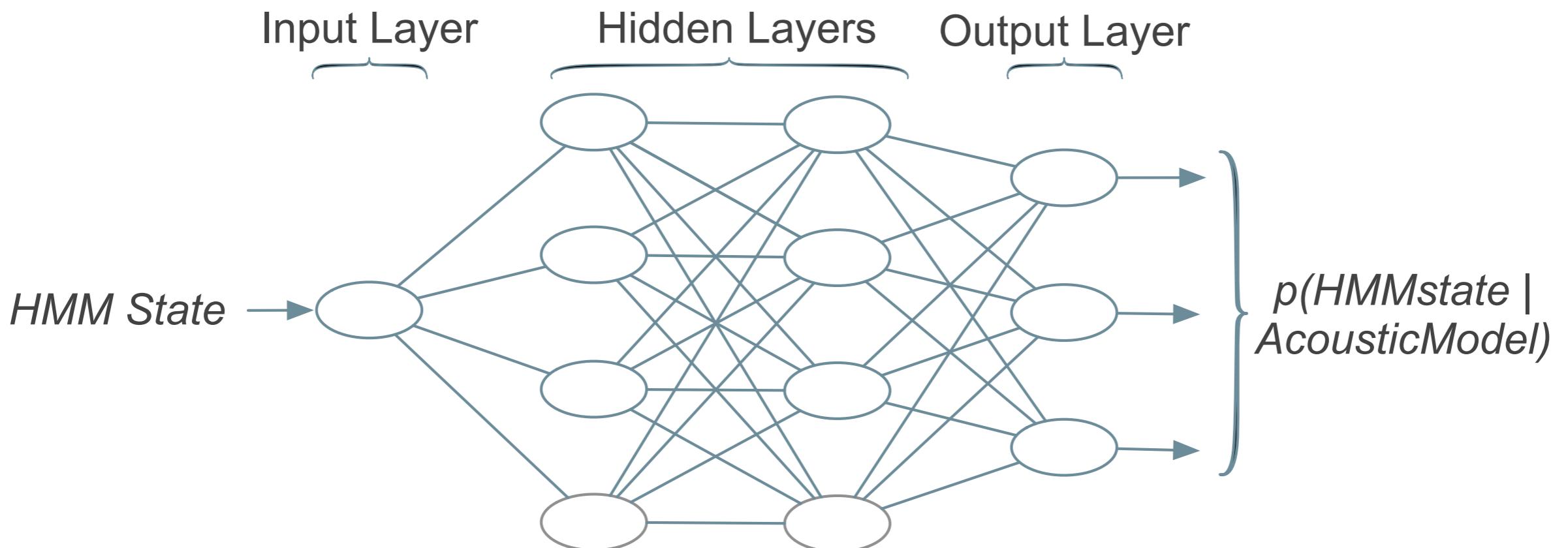
Lucida and Djinn Tutorial

Topic: Deep Learning in Intelligent Web Services

Speaker: Michael A. Laurenzano

10,000 foot view of Deep Neural Networks (DNN)

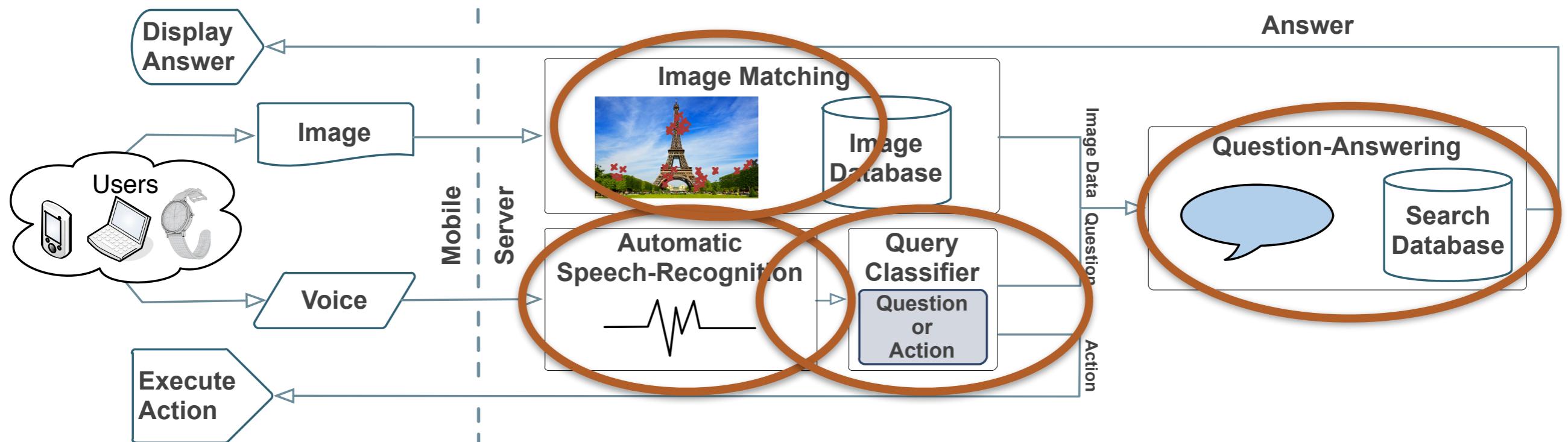
- What is “deep” about a DNN?
 - More layers
 - Often, bigger layers also
- Example — DNN in Lucida ASR



Why DNN?

- Emerging in a number of domains
 - Big data analytics
 - Video/image/audio recognition
 - Language semantics/translation
- Many components of IPAs like Lucida can leverage DNN

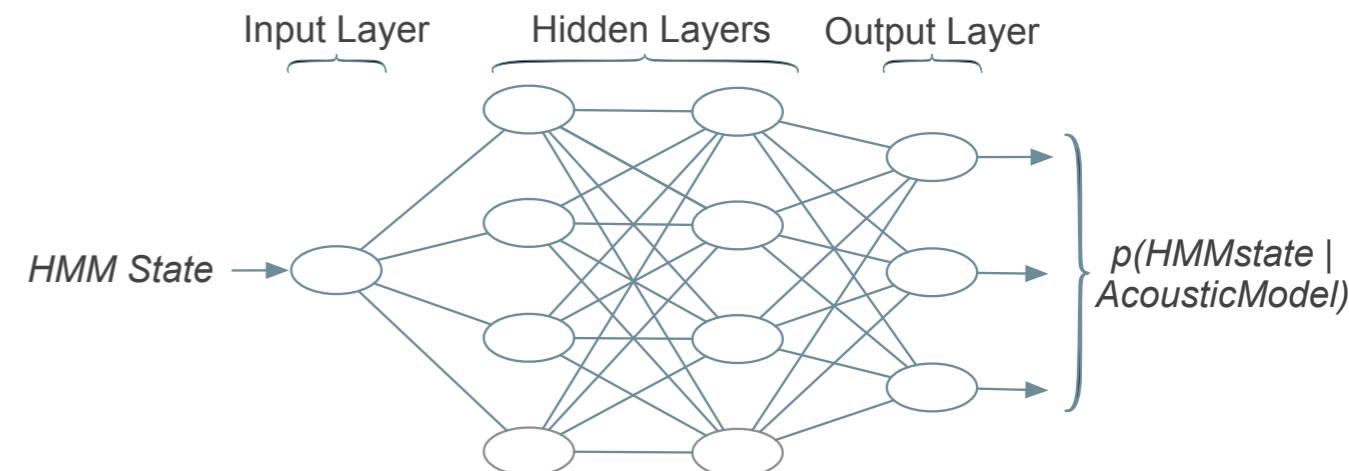
DNNs in the IPA pipeline



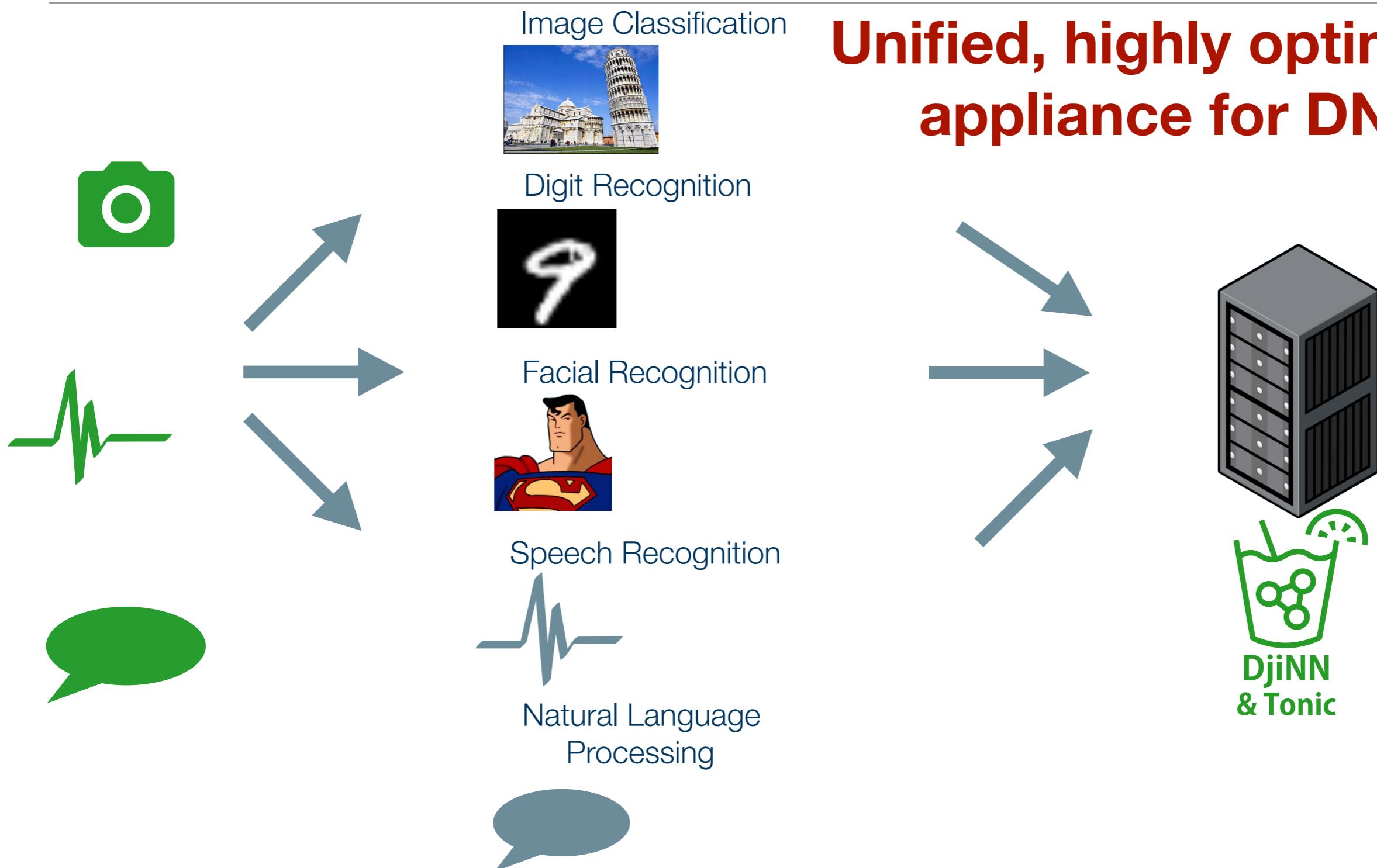
Studying DNNs as an architect is a challenge

Studying DNNs as an architect is a challenge

- What should my DNN do?
- Network configuration
 - How many layers?
 - How to configure those layers?
 - Interconnections?
 - Activation functions?
- Model training
 - When are the results accurate enough?
 - How to choose a training set?



DNN as a Service



Introducing Djinn and Tonic [ISCA'15]

- Djinn
 - Infrastructure for DNN as a service
 - Extensible to a number of DNN services
 - CPU-optimized implementation
 - GPU-optimized implementation
- Tonic Suite
 - 7 complete DNN services covering important IPA use cases
 - Configured networks and trained models are provided
 - No machine learning expertise required



<http://djinn.clarity-lab.org>

Lucida and DjNN Tutorial

Topic: Tonic Suite Components

Speaker: Johann Hauswald

Tonic Suite

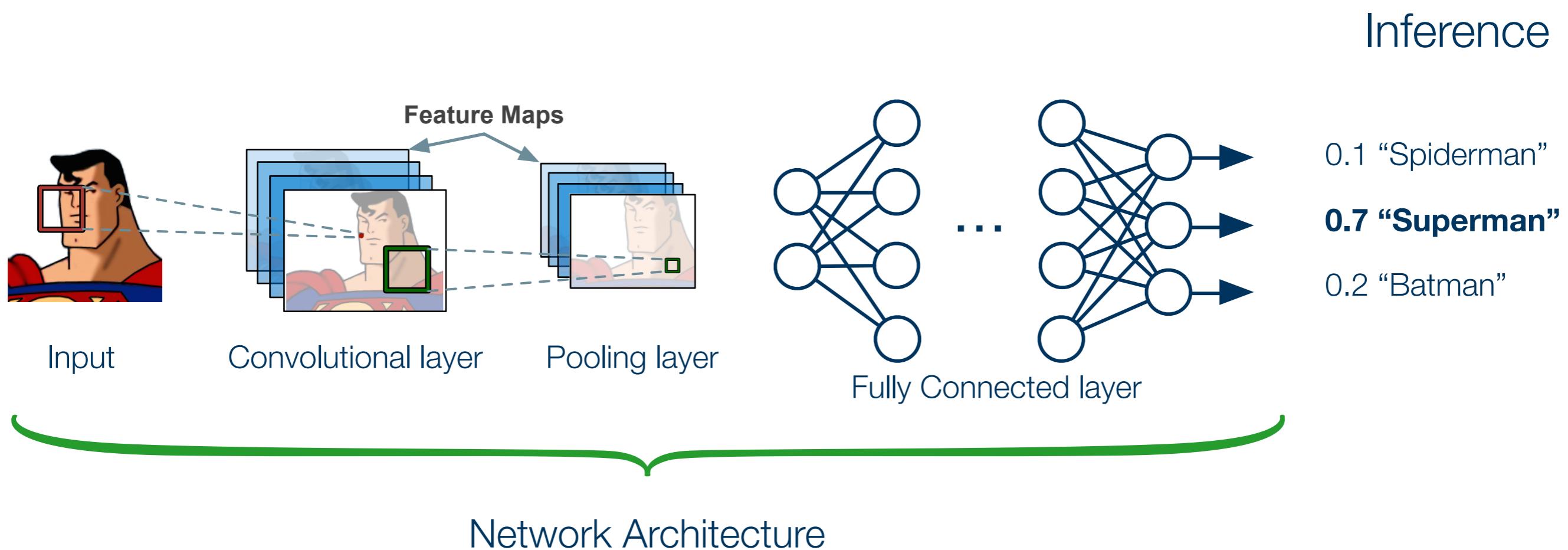
- Image Processing
 - Image classification (IMC)
 - Facial recognition (FACE)
 - Digit recognition (DIG)
- Speech Processing
 - Automatic speech recognition (ASR)
- Natural Language Processing
 - Part-of-speech tagging (POS)
 - Chunking (CHK)
 - Name entity recognition (NER)



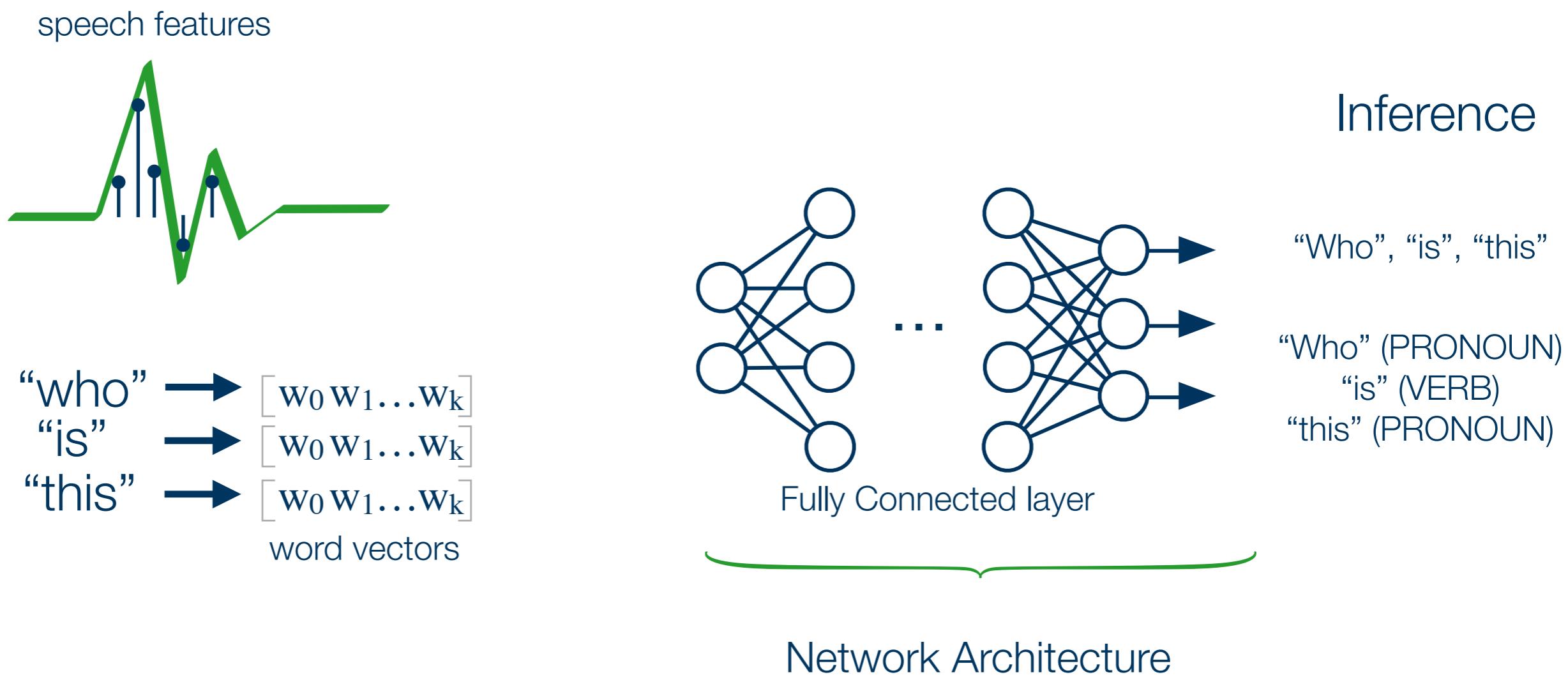
Natural Language Processing Task

POS	"business" (noun) "Superman" (P. noun)
CHK	"It's" (VP, B-NP)
NER	"business" (NP, I-NP) "Superman" (PERSON)

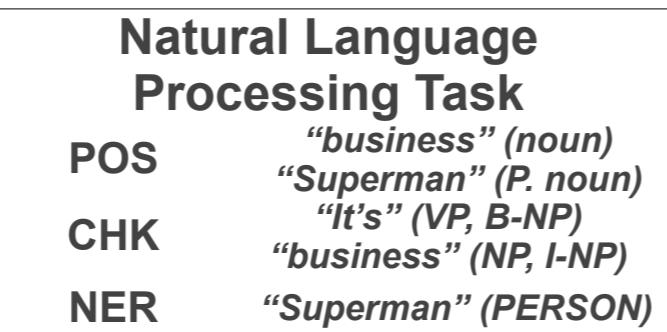
Deep Neural Networks (DNNs)



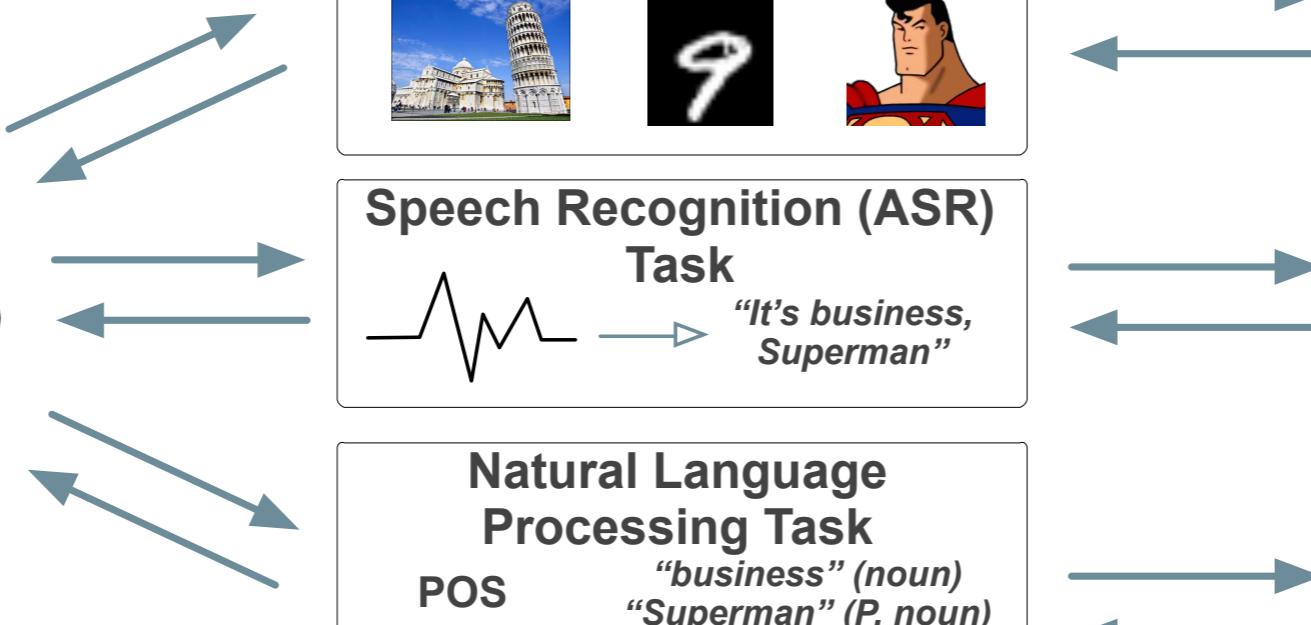
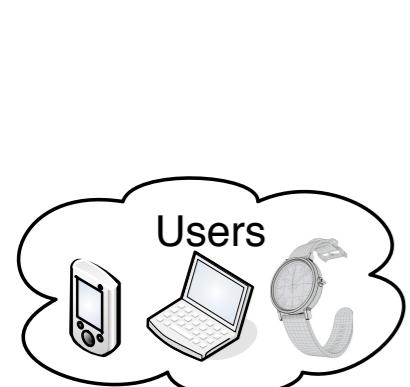
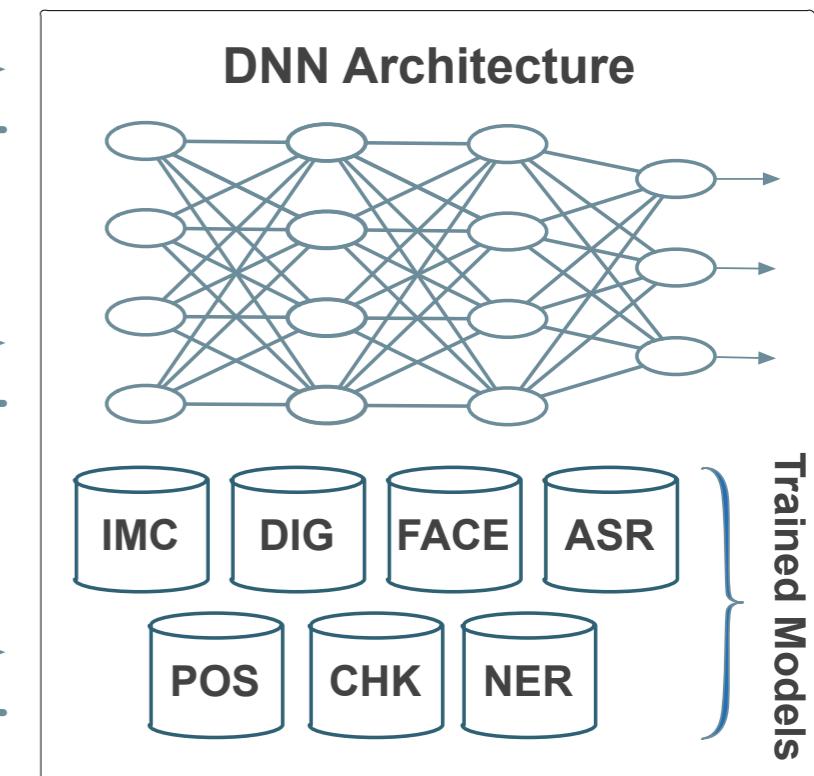
Deep Neural Networks (DNNs)



Tonic Suite Applications

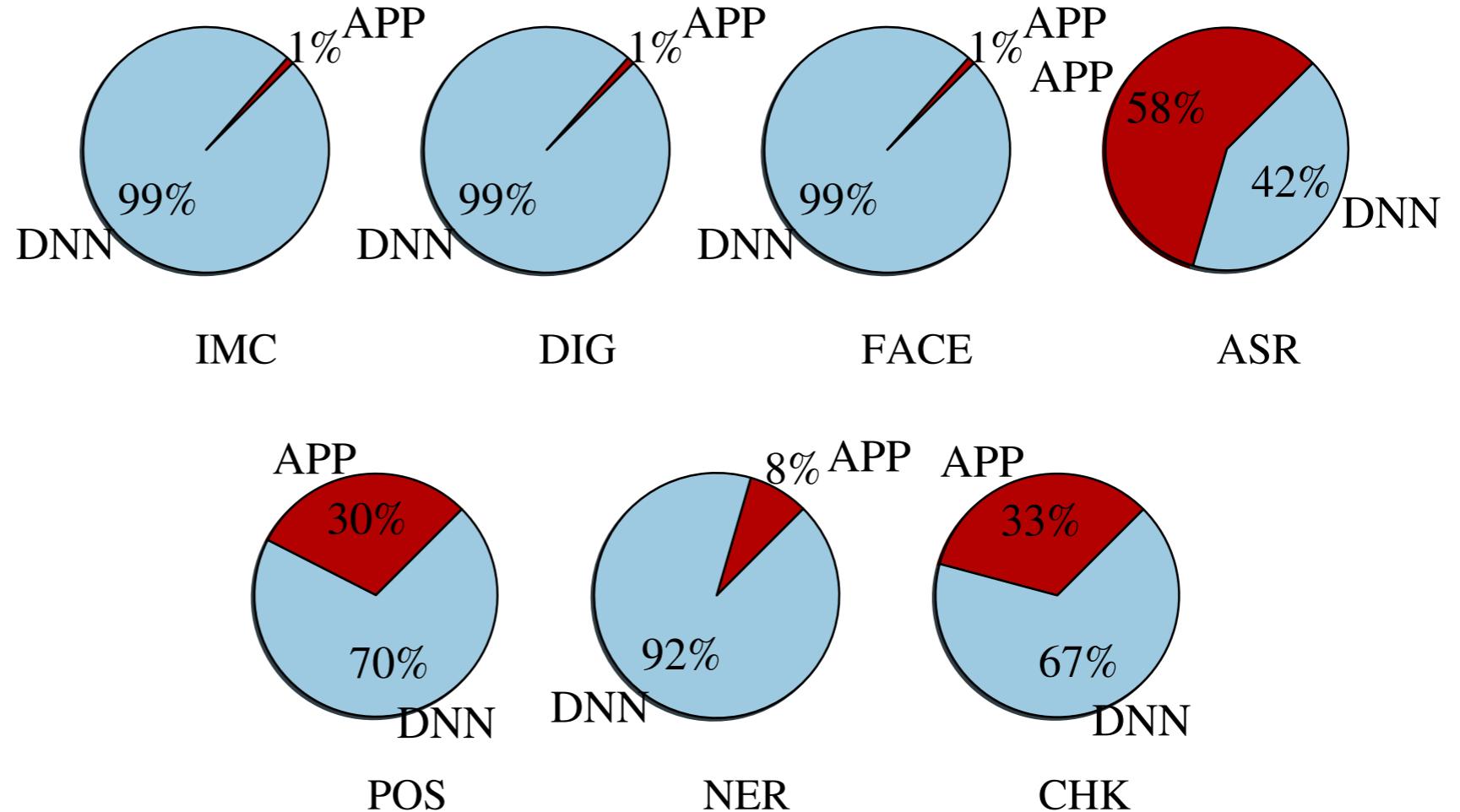


DjiNN DNN Service



Basic characterization

- Single CPU



Not all DNN services are created equal

Accelerating DNN can provide significant speedup

Tonic Suite



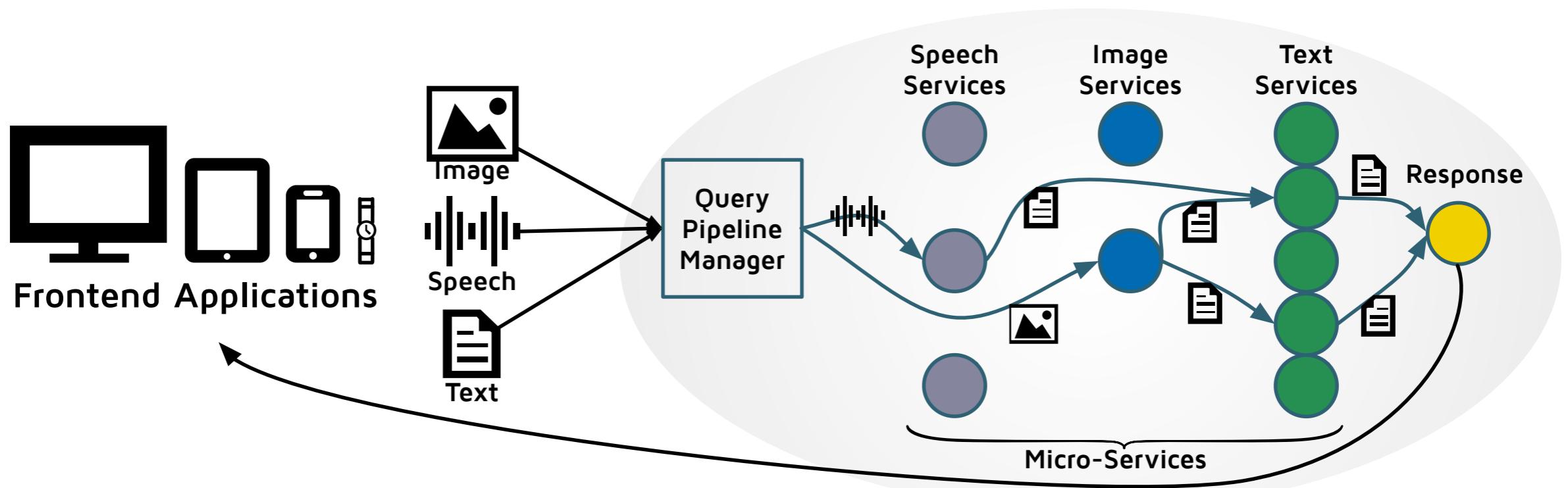
Lucida and Djinn Tutorial

Topic: Introduction to LucidaEco

Speaker: Michael A. Laurenzano

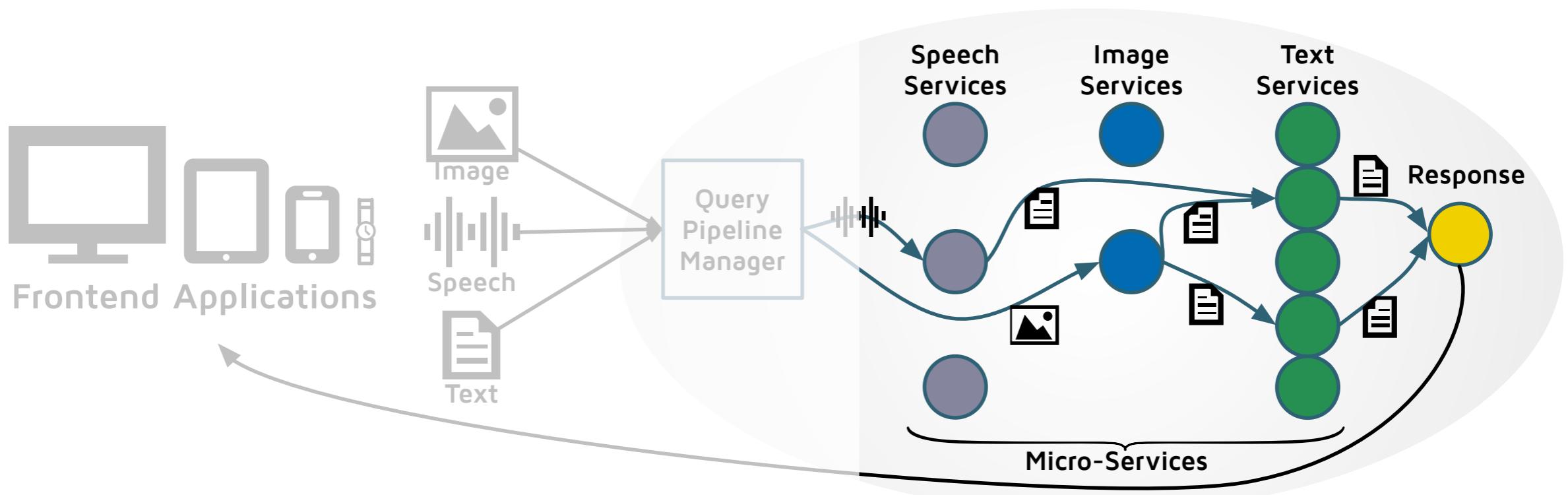
What is LucidaEco?

- The next step in the evolution of Lucida
- Open source end-to-end IPA ecosystem
 - Rich set of infrastructure and AI tools
 - Easily compose customized, scalable, production-quality IPAs



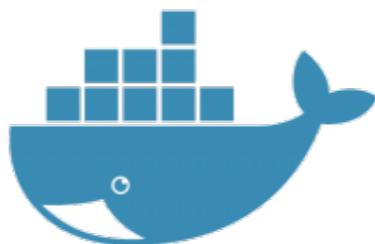
Micro-service Architecture

- Designed to achieve
 - Scalability
 - Modularity
 - Extensibility
- Independently deployable intelligent services
- Common interface - create, learn, infer



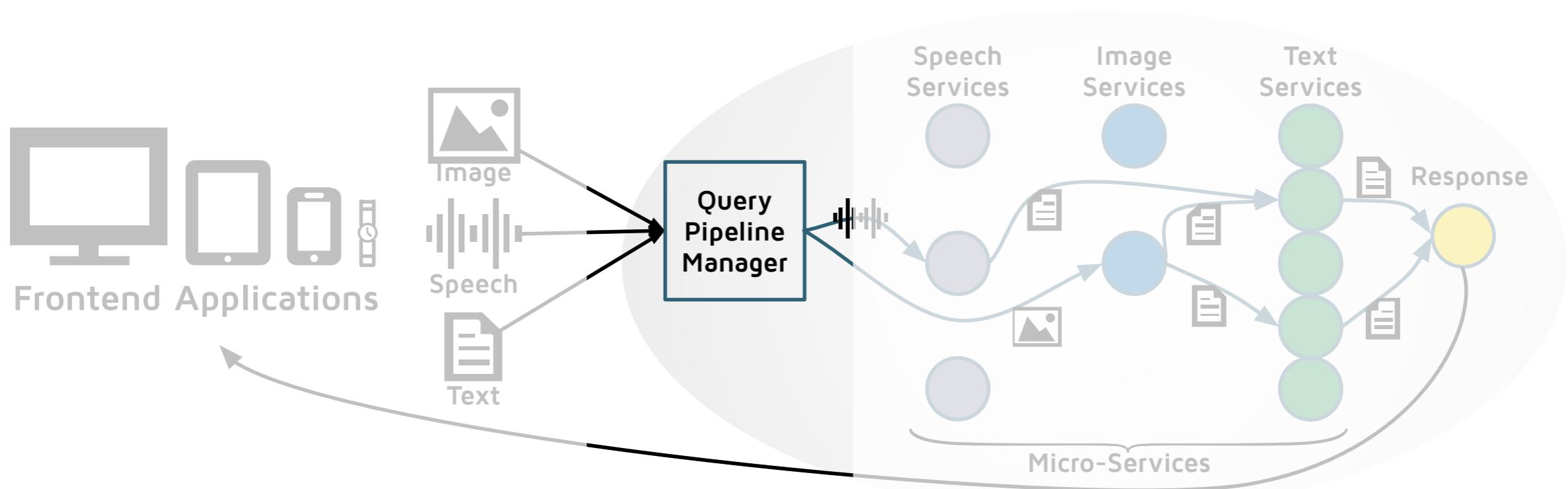
Micro-service Design

- Common interface to each micro-service
 - Create - new knowledge base/model
 - Learn - add information to KB/model
 - Infer - retrieve knowledge
- Inter-service communication
 - RPC abstraction (Facebook's Thrift) for most
 - websockets for streaming audio
 - Language/platform issues hidden from micro-service
- “Virtualized”, platform-independent deployment
 - docker (container-based)



Query Pipeline Manager

- Orchestrate the pipeline of IPA components
- Goals
 - Optimize computational footprint
 - Hardware — accelerators and heterogeneity
 - Micro-service co-location



LucidaEco is Growing

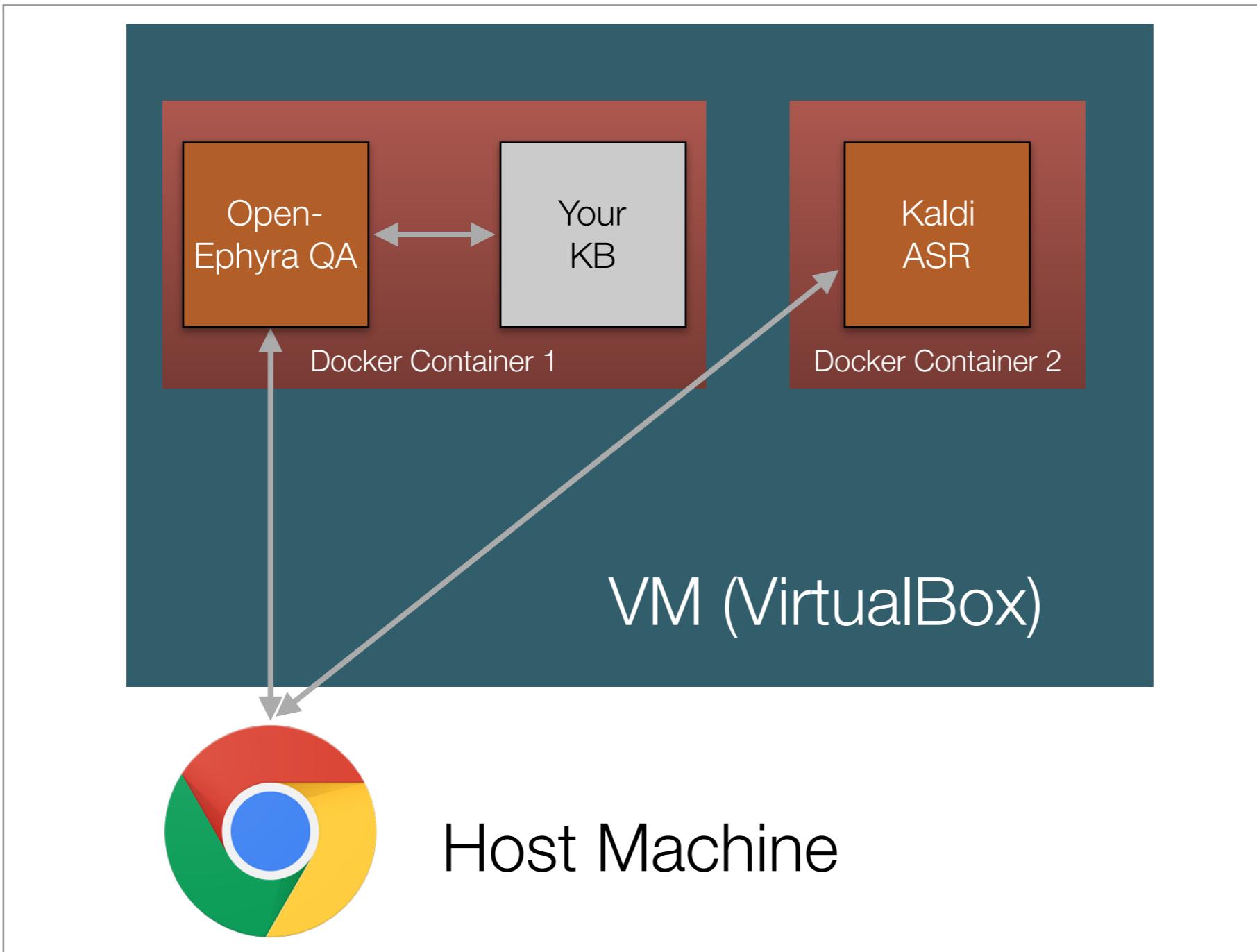
- Micro-service ecosystem
 - ASR (Kaldi, Sphinx, different languages)
 - QA (OpenEphyra, Qanta, YodaQA)
 - Machine Translation, Sentiment, Summarization
 - Image matching, object recognition
- Frontend apps — web, mobile, tablet; others coming
- LucidaEco as a research platform
 - AI algorithms/implementations in an end-to-end pipeline
 - Scalability - acceleration, virtualization, high-BW communication substrates
- Burgeoning community of developers/researchers
 - Join us!

Lucida and Djinn Tutorial

Topic: Build Your Own IPA

Speakers: Michael A. Laurenzano and Johann Hauswald

The Architecture of Your IPA

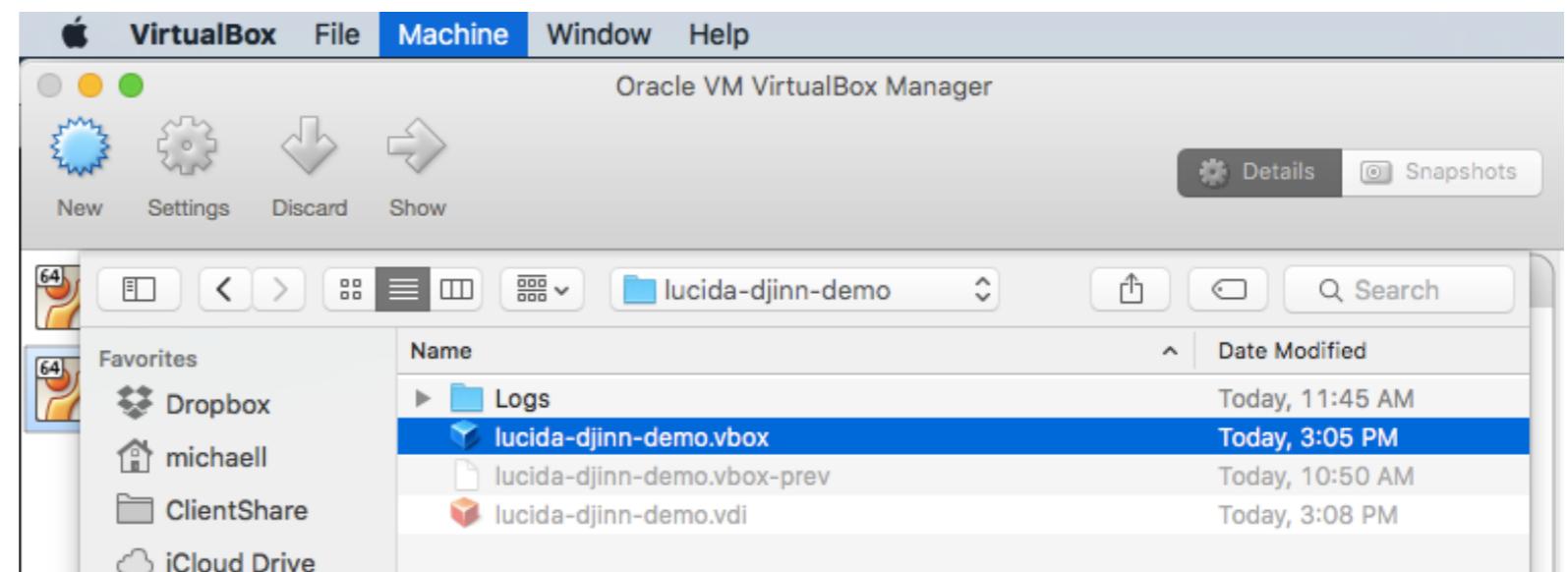


Setup Prerequisites

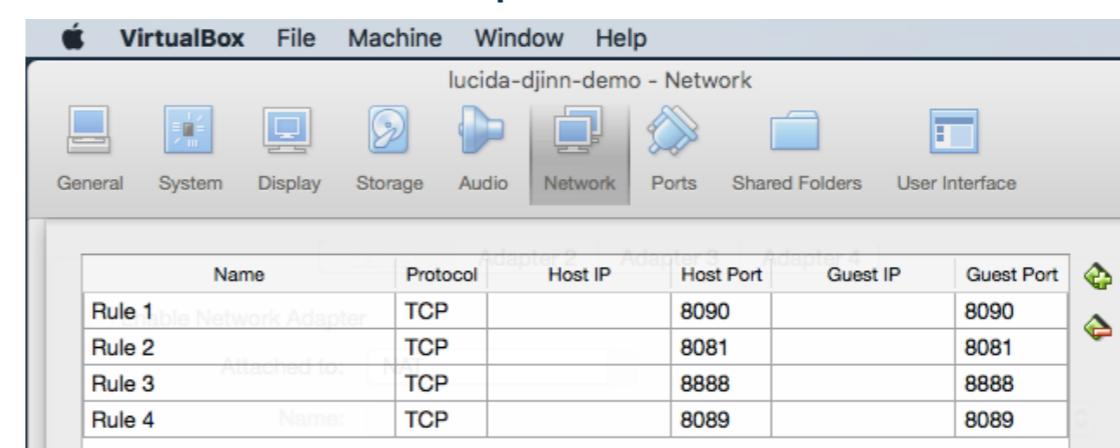
- Install VirtualBox
- Install Chrome
- Copy the tutorial VM – USB thumb drive

Configure VirtualBox

- Unpack lucida-djinn-demo.tgz into ~VirtualBox VMs/
- Add lucida-djinn-demo to VirtualBox
 - Machine -> Add



- Port forwarding
 - Machine -> Settings -> Network -> Adapter 1 -> Advanced -> Port Forwarding



Run Your IPA

- Start!
- username/password: demo/demo
- open Terminal application
- cd ~/h pca-demo/source-code/lucida/
- Running the demo
 - \$ docker-compose up
 - ON THE HOST: <http://localhost:8081/>

Build Your Own IPA



Thank You!!!

- **Tutorial Website** — lucida.ai
- **Lucida Suite** — <http://sirius.clarity-lab.org/downloads/>
- **Tonic Suite** — <http://djinn.clarity-lab.org/downloads/>
- **LucidaEco** — <http://github.com/claritylab/lucida/>