Lucida

~~Sirius~~ and DjiNN 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 - DjiNN 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 DjiNN Tutorial

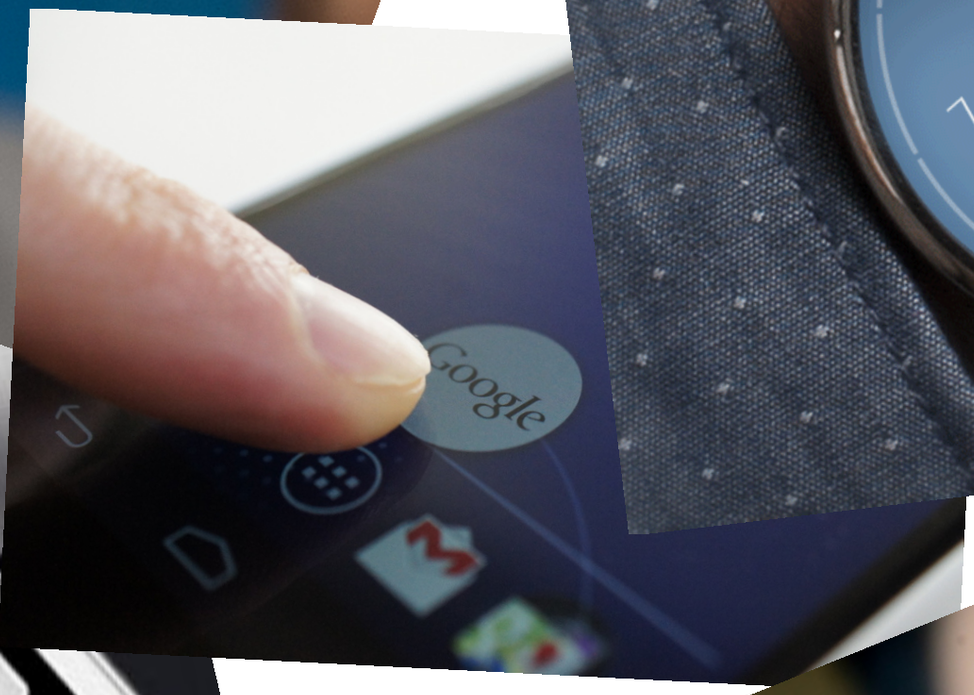
Topic: Introduction to Lucida
Speakers: Johann Hauswald

COMPUTER SCIENCE
& ENGINEERING
UNIVERSITY OF MICHIGAN

ClarityLab


NOKIA
Let me help you get ready for the weekend.
Sunny weekend
Sat & Sun, 0%
Local events
West Park Block Party, IG Alliance Art A
Local restaurants
Argentina Tapas, Coho Vineyard & Win
Reminder
Pack for Frankfurt trip!
Don't show me this
Here's the latest stats for
ask me anything


9:41 AM
"On May 19 remind me it's dad's birthday"
Here's your reminder for May 19, 2012 at 9 am:
Saturday
May 2012


Google


13:37
85%
HUMIDITY
19°
SAN FRANCISCO
feeling awesome
UNIFIER

Rise of the Wearables

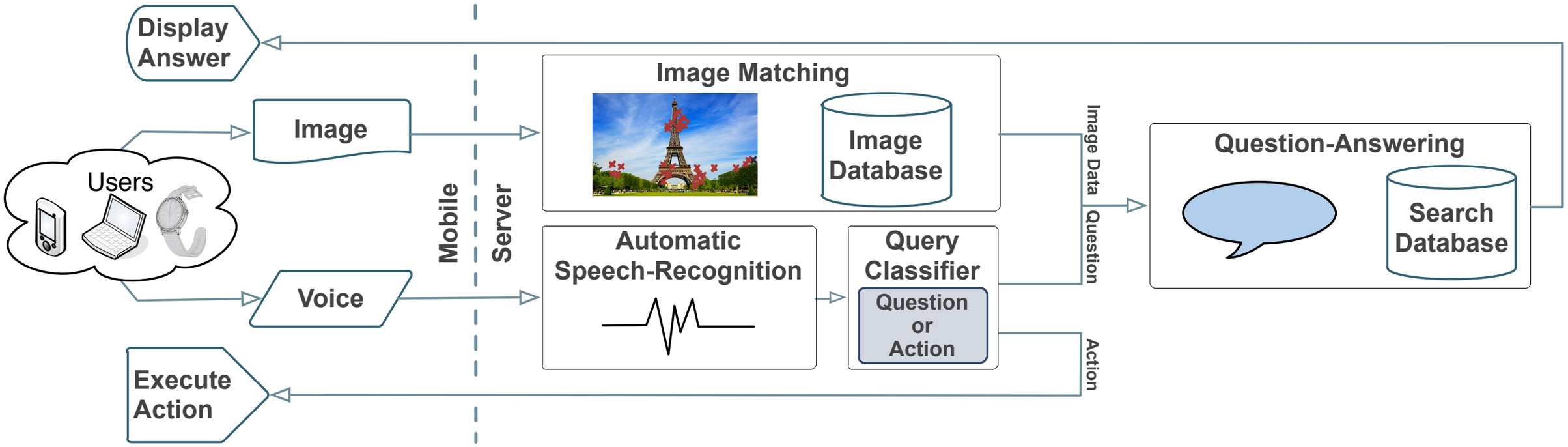
40%

$80bn

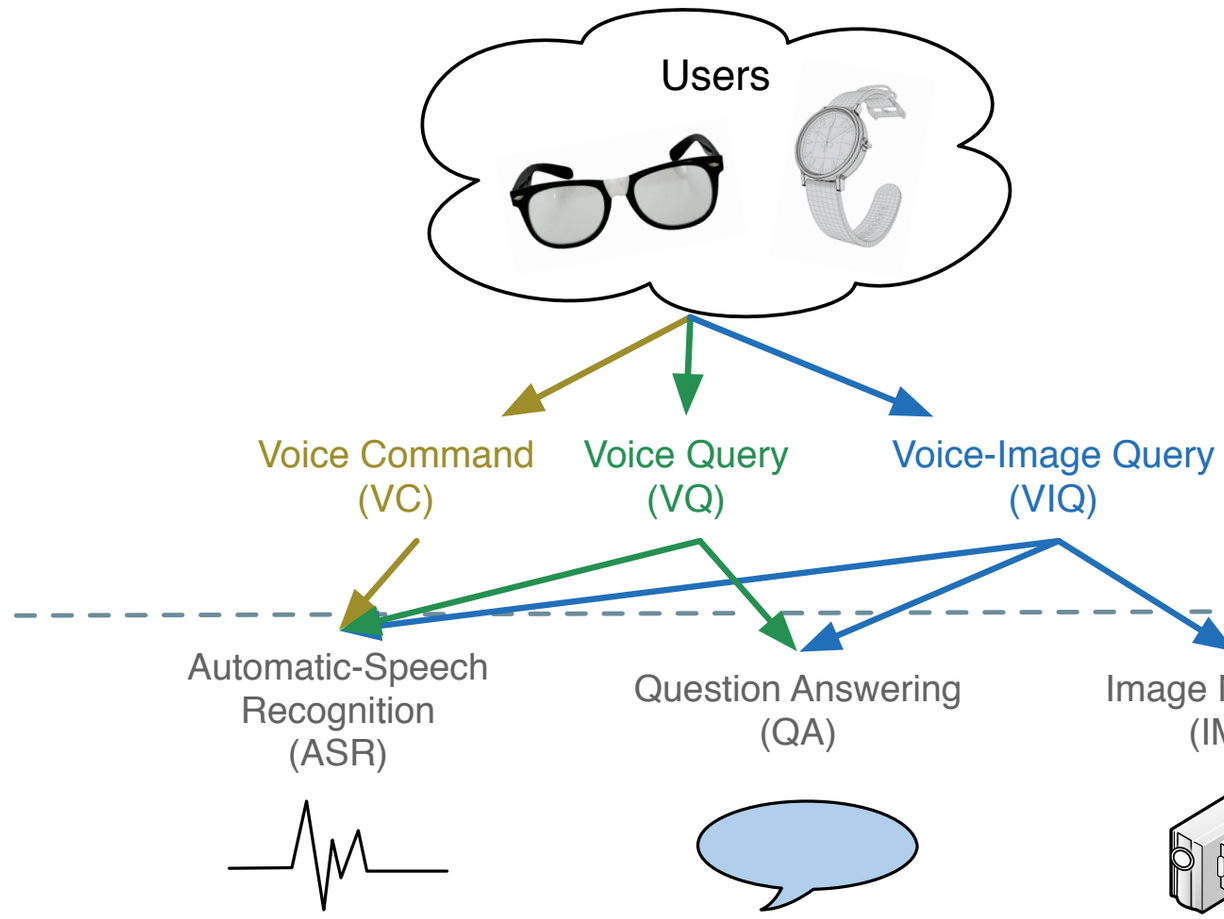
Questions Arise

- 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]


Display Answer
Image
Users
Voice
Mobile
Server
Image Matching
Image Database
Automatic Speech-Recognition
Query Classifier
Question or Action
Image Data
Question
Action
Question-Answering
Search Database
Execute Action

Lucida Hierarchy



Query Taxonomy

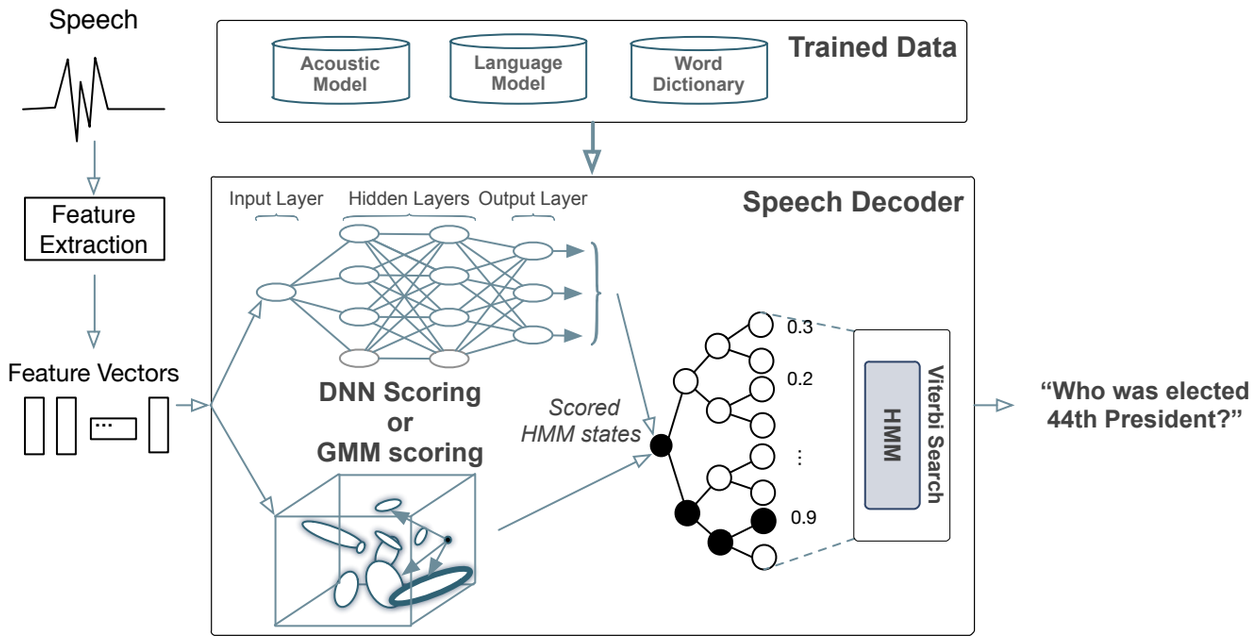
IPA Services

Tasks

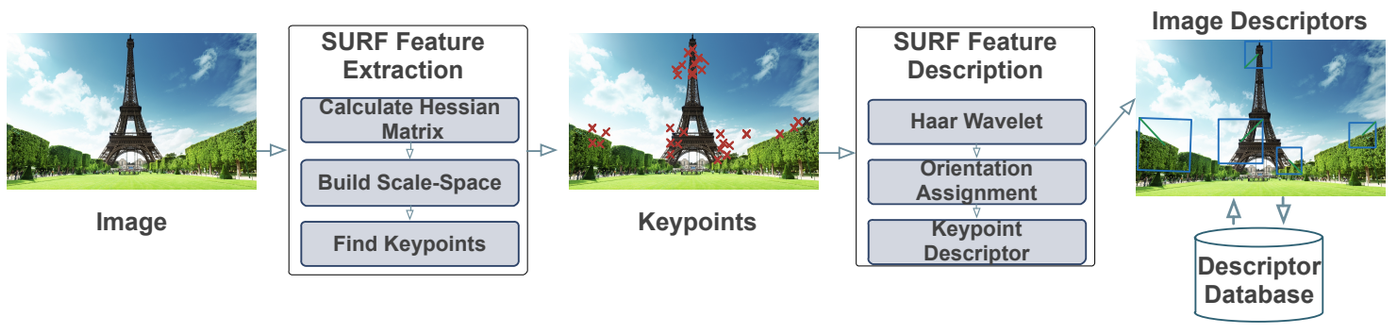
Algorithmic Components

Three Horsemen of Lucida

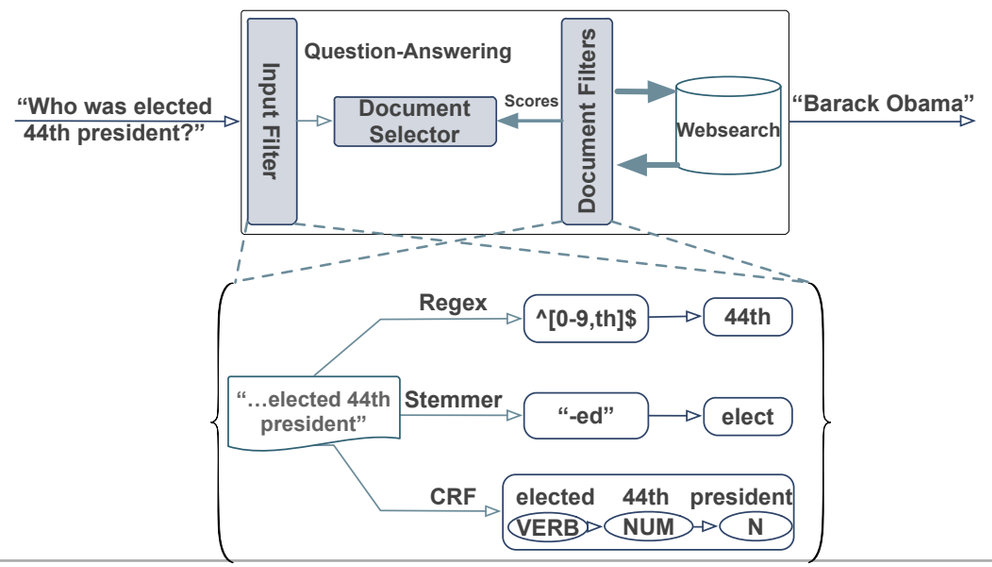
ASR



IMM

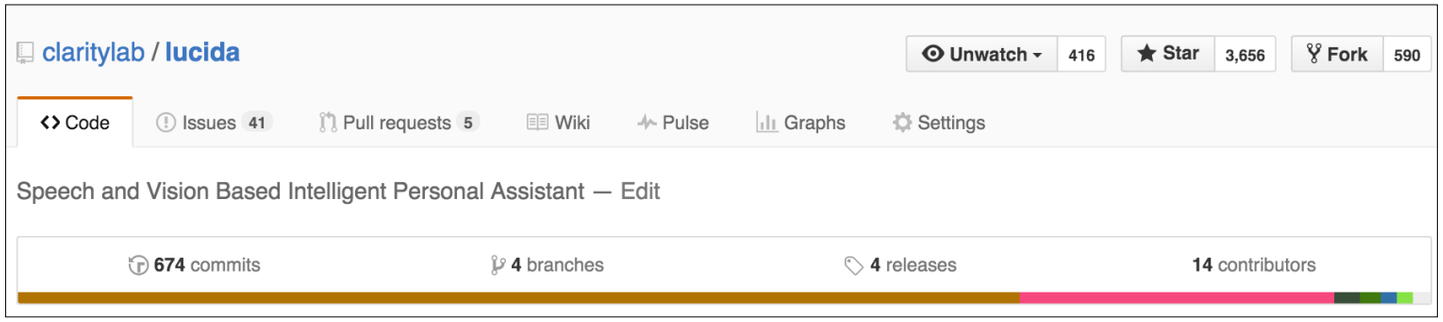


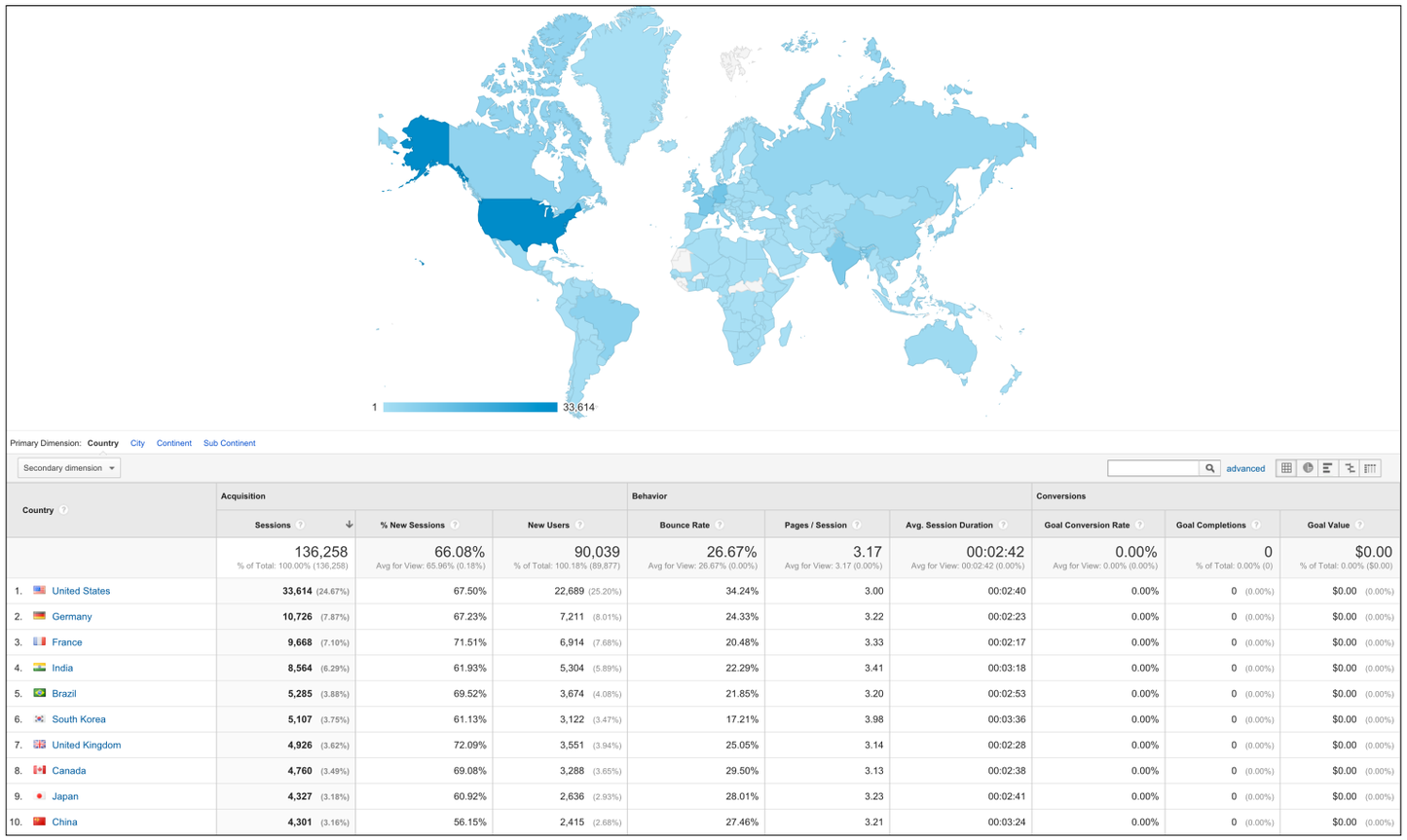
QA



Project Status

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





Country	Acquisition			Behavior			Conversions		
	Sessions	% New Sessions	New Users	Bounce Rate	Pages / Session	Avg. Session Duration	Goal Conversion Rate	Goal Completions	Goal Value
	136,258 % of Total: 100.00% (136,258)	66.08% Avg for View: 65.96% (0.18%)	90,039 % of Total: 100.18% (89,877)	26.67% Avg for View: 26.67% (0.00%)	3.17 Avg for View: 3.17 (0.00%)	00:02:42 Avg for View: 00:02:42 (0.00%)	0.00% Avg for View: 0.00% (0.00%)	0 % of Total: 0.00% (0)	$0.00 % of Total: 0.00% ($0.00)
1. United States	33,614 (24.67%)	67.50%	22,689 (25.20%)	34.24%	3.00	00:02:40	0.00%	0 (0.00%)	$0.00 (0.00%)
2. Germany	10,726 (7.87%)	67.23%	7,211 (8.01%)	24.33%	3.22	00:02:23	0.00%	0 (0.00%)	$0.00 (0.00%)
3. France	9,668 (7.10%)	71.51%	6,914 (7.68%)	20.48%	3.33	00:02:17	0.00%	0 (0.00%)	$0.00 (0.00%)
4. India	8,564 (6.29%)	61.93%	5,304 (5.89%)	22.29%	3.41	00:03:18	0.00%	0 (0.00%)	$0.00 (0.00%)
5. Brazil	5,285 (3.88%)	69.52%	3,674 (4.08%)	21.85%	3.20	00:02:53	0.00%	0 (0.00%)	$0.00 (0.00%)
6. South Korea	5,107 (3.75%)	61.13%	3,122 (3.47%)	17.21%	3.98	00:03:36	0.00%	0 (0.00%)	$0.00 (0.00%)
7. United Kingdom	4,926 (3.62%)	72.09%	3,551 (3.94%)	25.05%	3.14	00:02:28	0.00%	0 (0.00%)	$0.00 (0.00%)
8. Canada	4,760 (3.49%)	69.08%	3,288 (3.65%)	29.50%	3.13	00:02:38	0.00%	0 (0.00%)	$0.00 (0.00%)
9. Japan	4,327 (3.18%)	60.92%	2,636 (2.93%)	28.01%	3.23	00:02:41	0.00%	0 (0.00%)	$0.00 (0.00%)
10. China	4,301 (3.16%)	56.15%	2,415 (2.68%)	27.46%	3.21	00:03:24	0.00%	0 (0.00%)	$0.00 (0.00%)

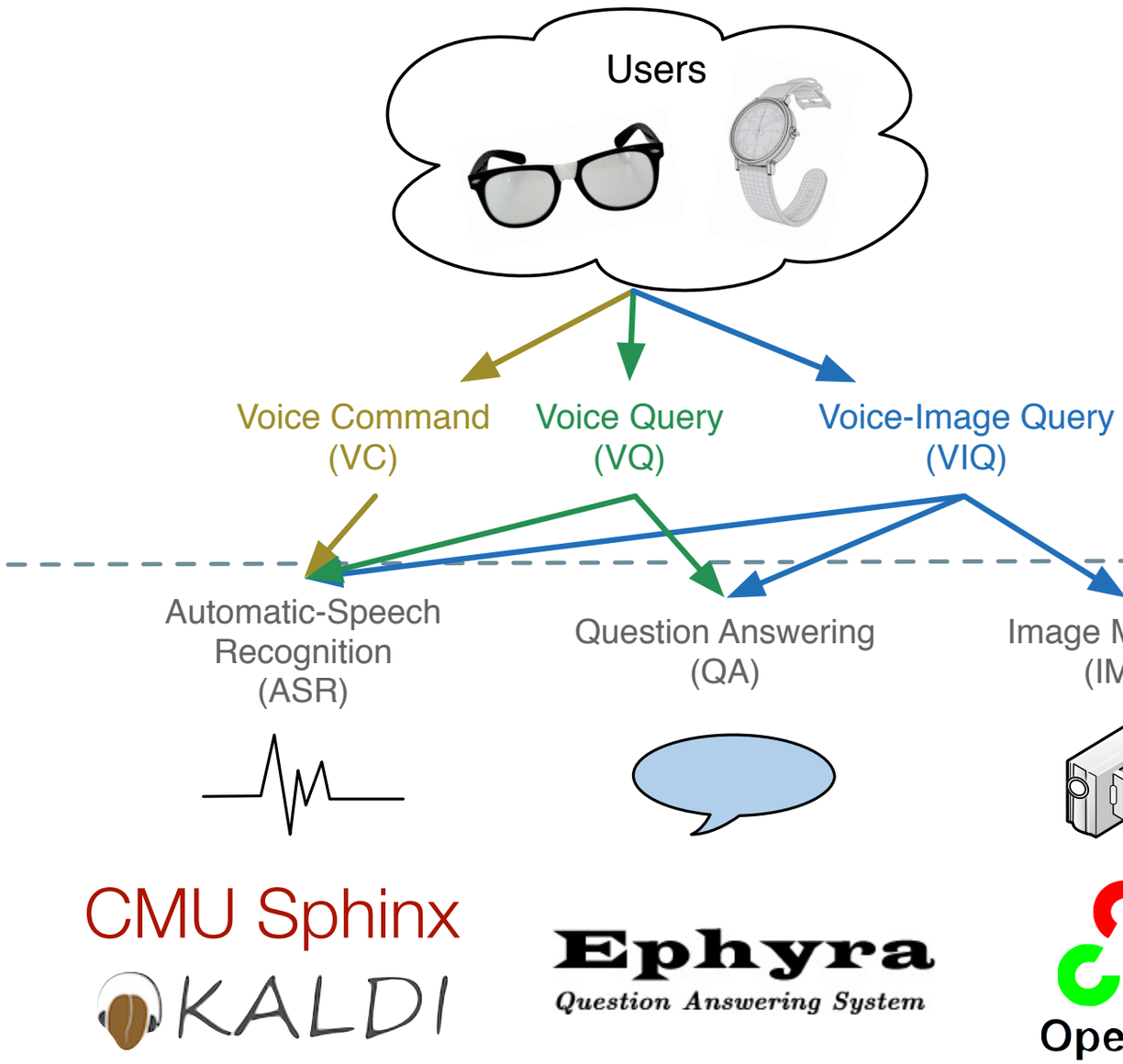
Lucida and DjiNN Tutorial

Topic: Core Algorithmic Components of IPAs
Speaker: Yunqi Zhang

COMPUTER SCIENCE
& ENGINEERING
UNIVERSITY OF MICHIGAN

ClarityLab

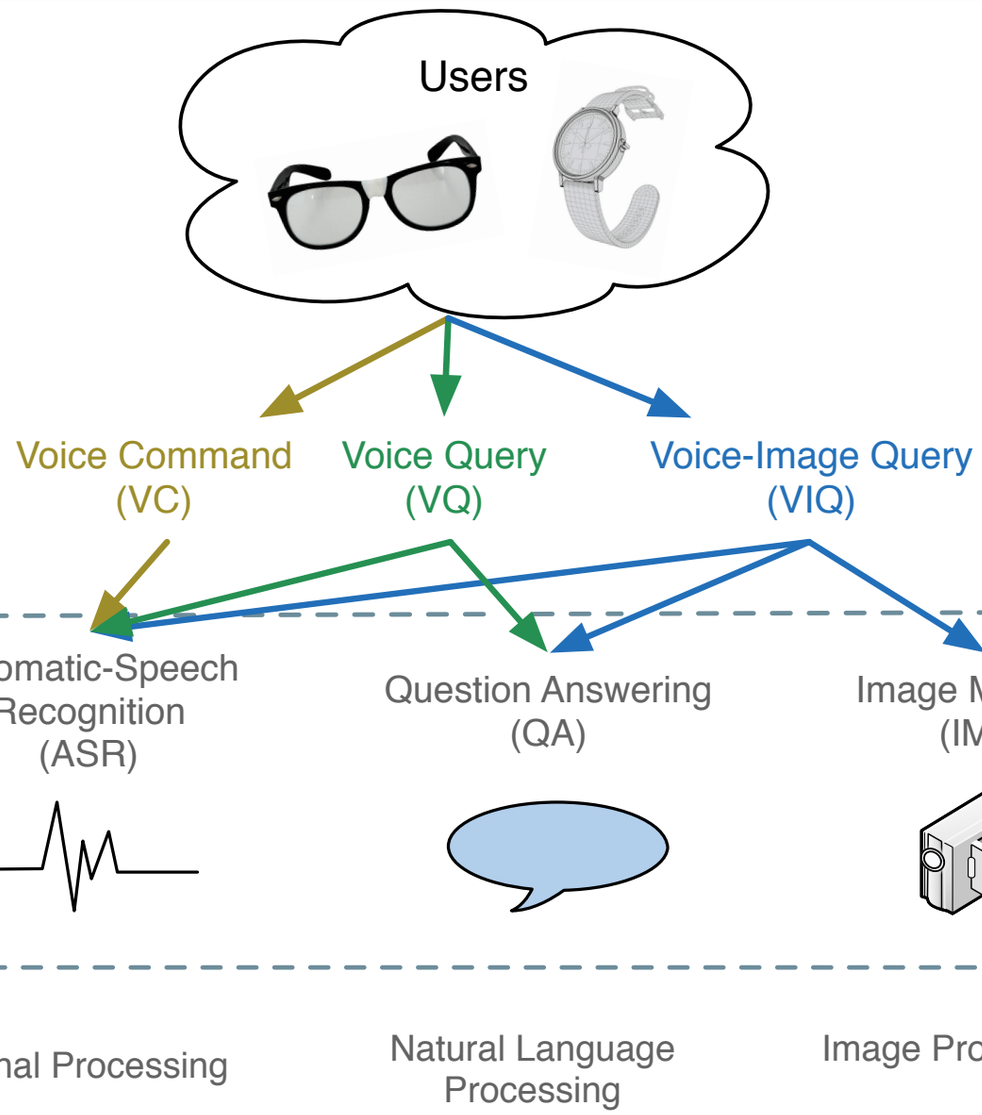
How does Lucida work?



Query Taxonomy

IPA Services

How does Lucida work?



How does Lucida work?

Signal Processing	Natural Language Processing		Image Processing	*Tasks*
Gaussian Mixture Model (GMM) or Deep Neural Network (DNN)	Regular Expression	Stemmer	Feature Extraction	*Algorithmic Components*
		Conditional Random Fields	Feature Description	
85% or 78%	85%		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

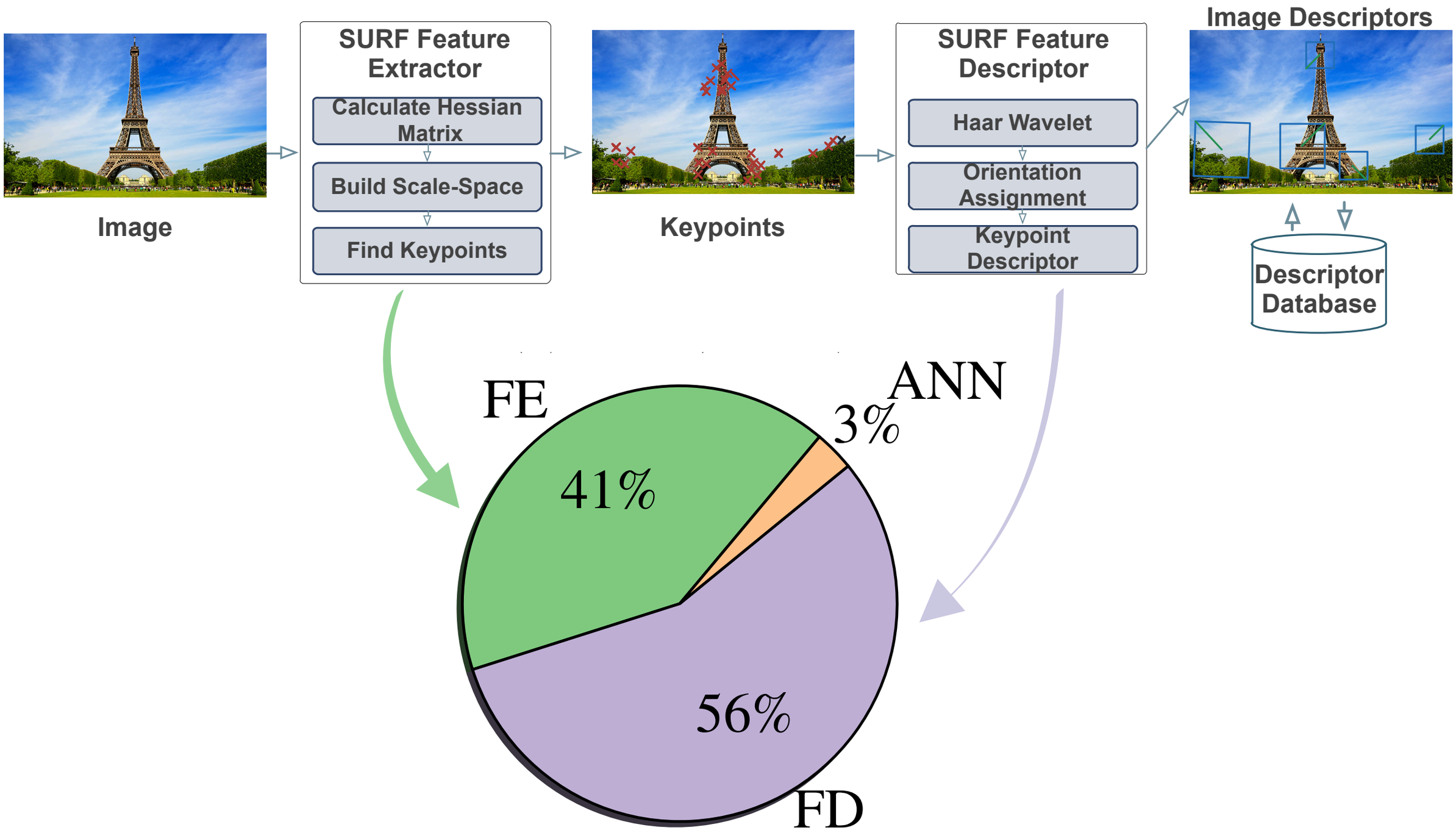
Objectives of Lucida Suite

- 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

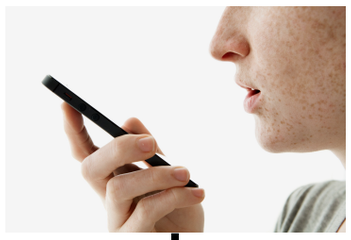
Image Matching



Accelerated workload: 97%

Automatic Speech Recognition

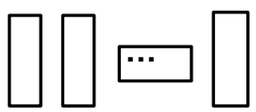
Automatic Speech Recognition (ASR)



Speech

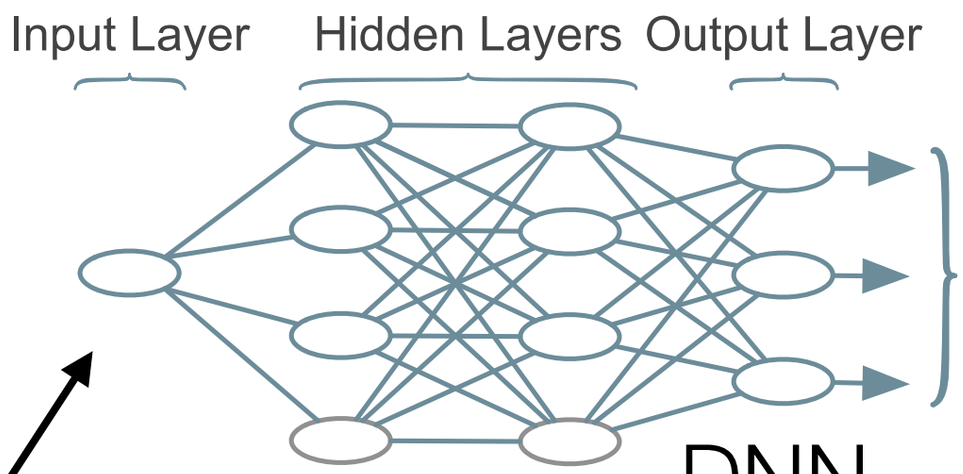
Feature Extraction

Feature Vectors



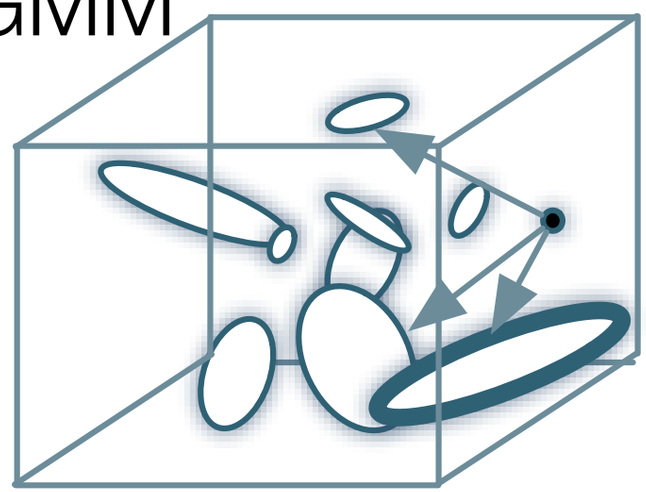
Pre-processing

Recognition Process

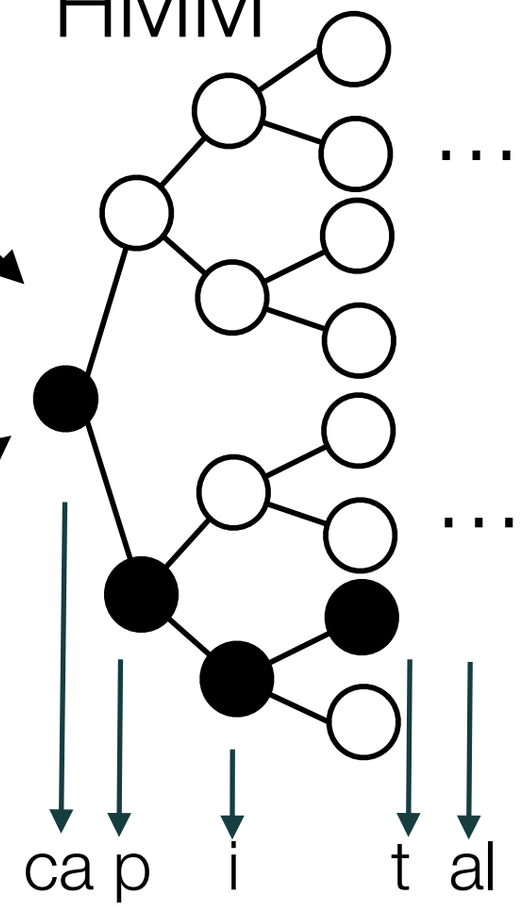


DNN

GMM



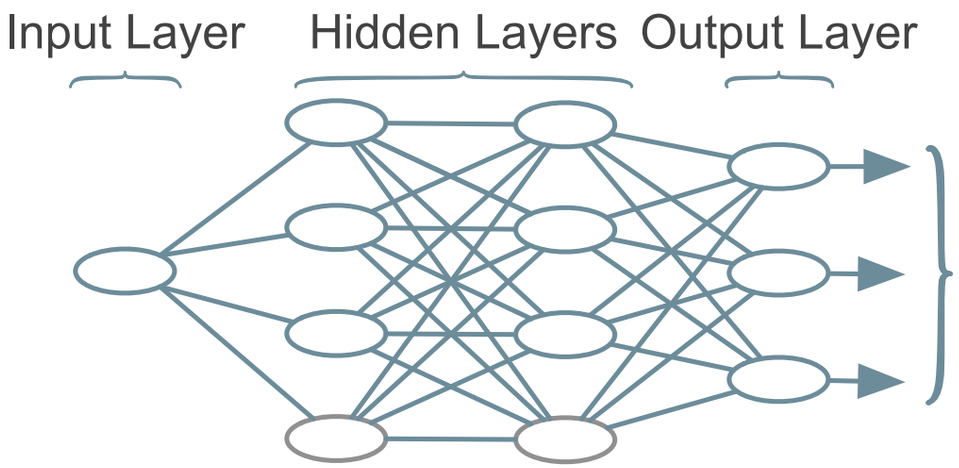
HMM

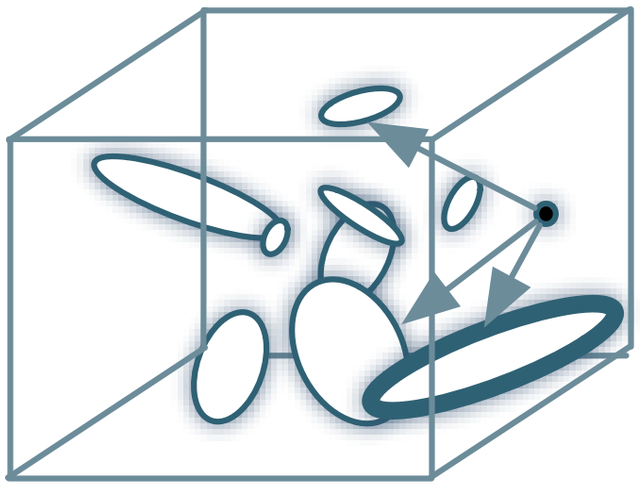


Viterbi Algorithm

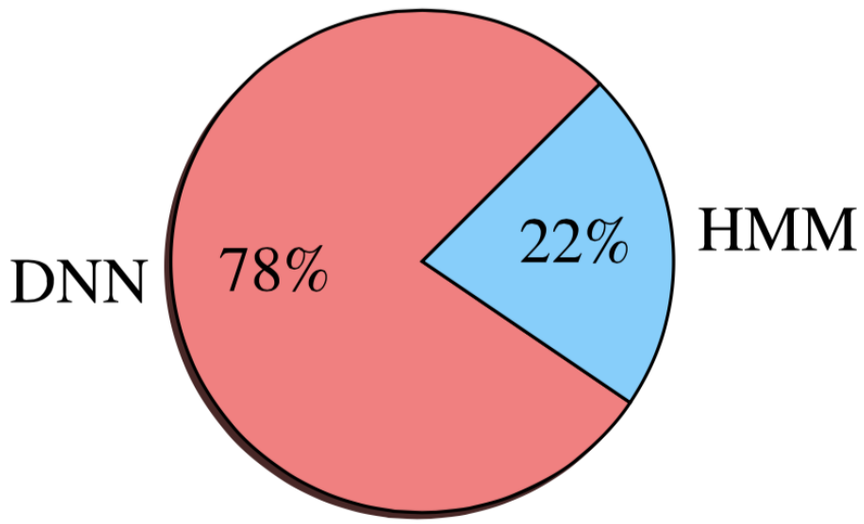
ClarityLab

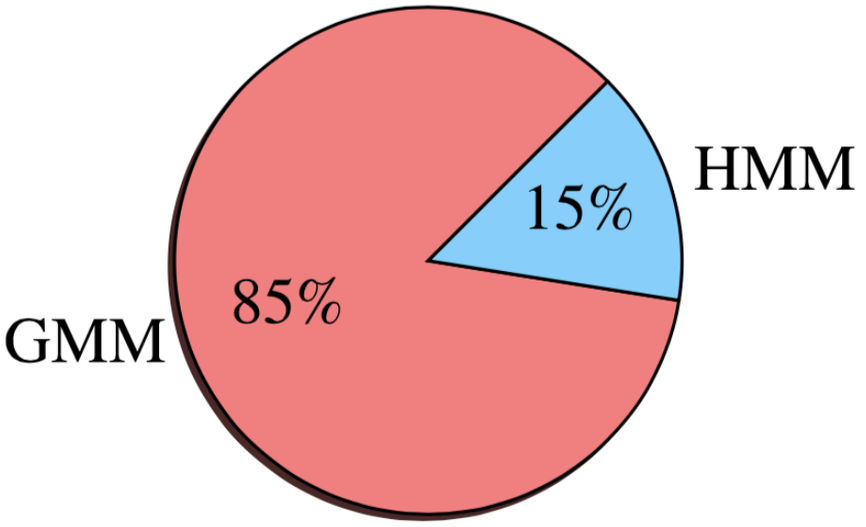
Automatic Speech Recognition (ASR)





Recognition Process





Question Answering

Question Answering

Who was elected 44th president of the United States?

Input Filters

Candidate Phrases:
- (*) was elected …
- (*) was the 44th …

Document Database

Document Filters e.g., wikipedia

Sentences in Candidate Documents:
- wikipedia/Barack_Obama
- wikipedia/Presidents_of_the_US

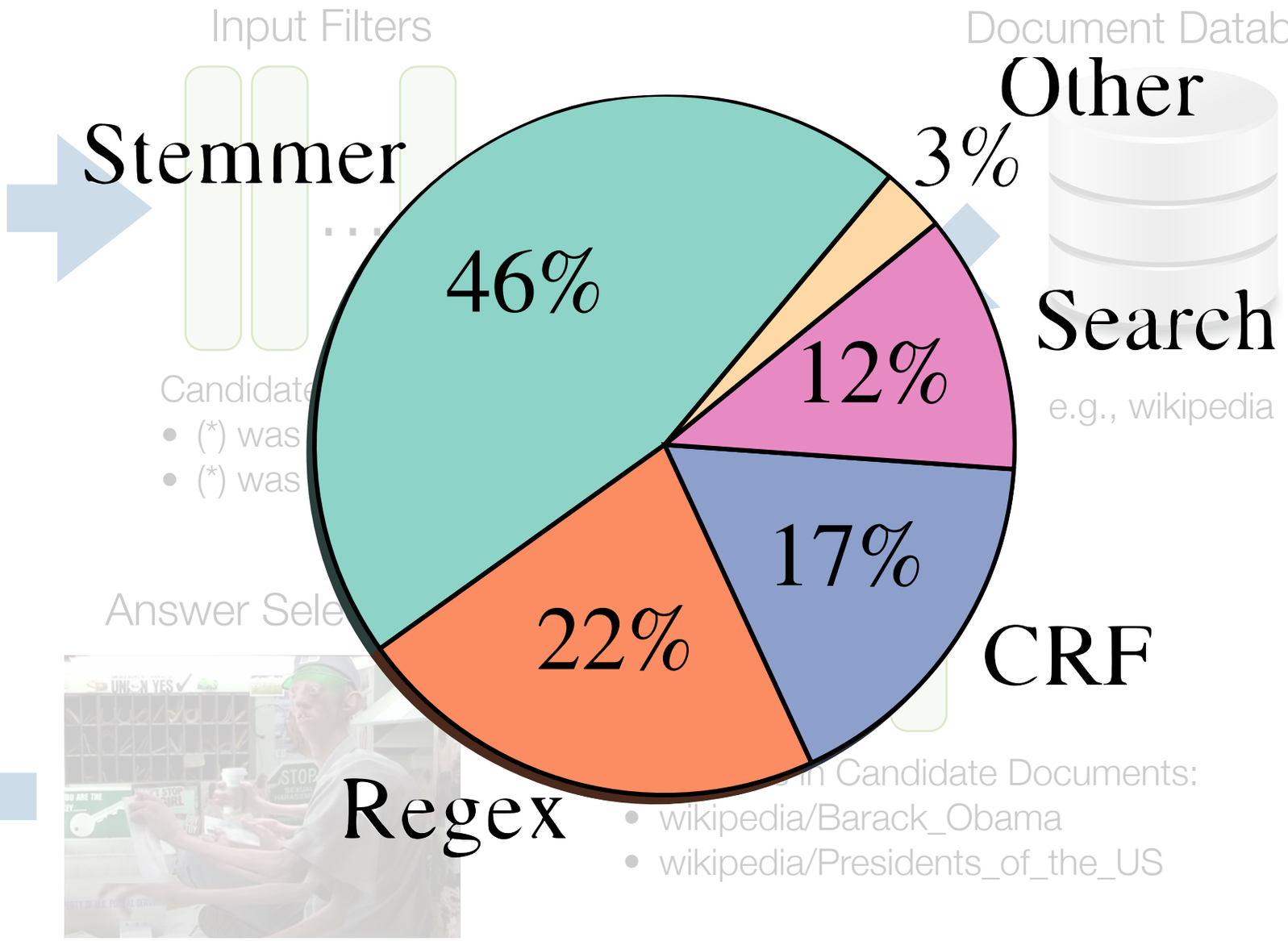
Answer Selector

Barack Obama

Question Answering

85% of the execution

- Conditional Random Fields
- Stemmer
- Regular Expression



Lucida and DjiNN Tutorial

Topic: Hands-on: Lucida Suite
Speaker: Yunqi Zhang

COMPUTER SCIENCE
& ENGINEERING
UNIVERSITY OF MICHIGAN

ClarityLab

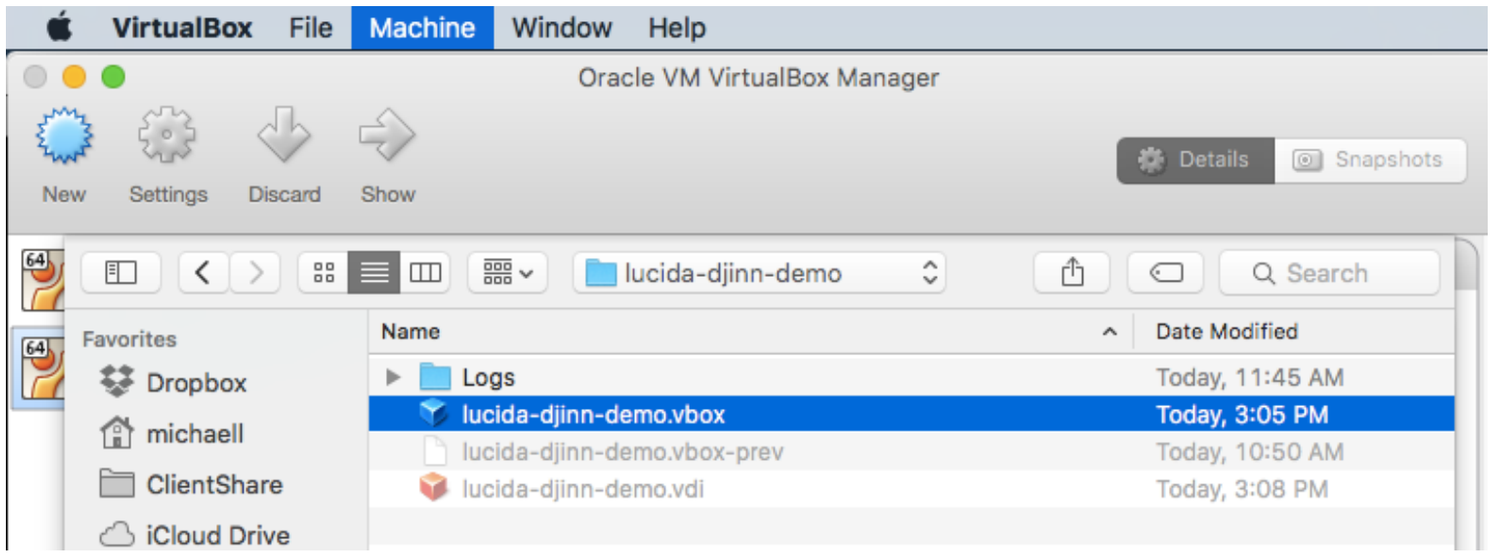
Lucida Suite

Setup Prerequisites

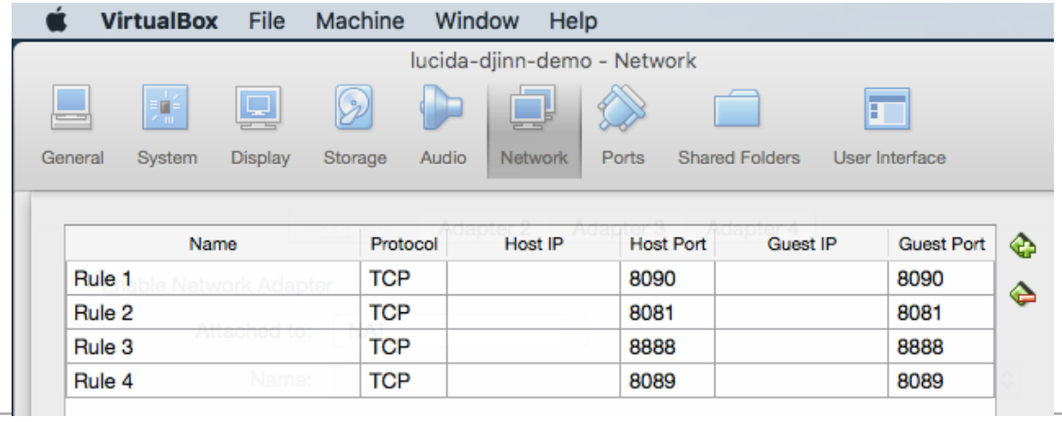
- Install VirtualBox
- Install Chrome
- Copy the tutorial VM — USB thumb drive

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 ~/hpca-demo/source-code/lucida/lucida-suite

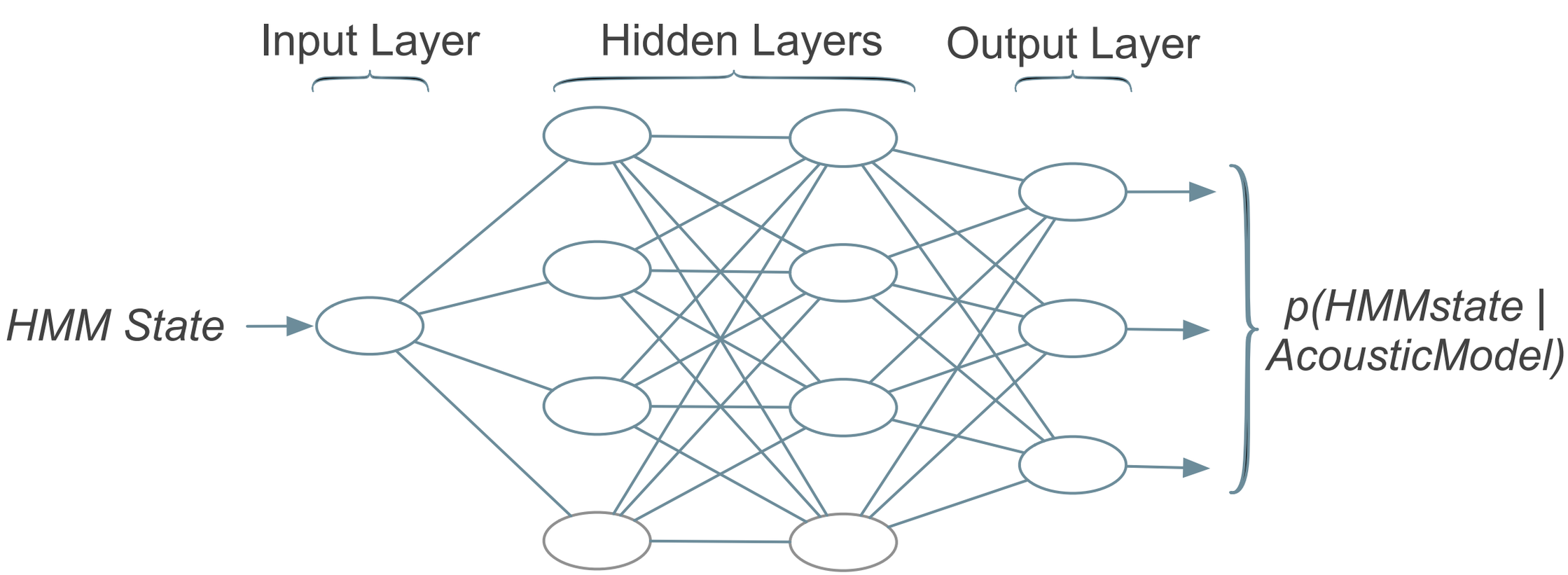
- Follow along with Yunqi!

Lucida and DjiNN Tutorial

Topic: Deep Learning in Intelligent Web Services
Speaker: Michael A. Laurenzano

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

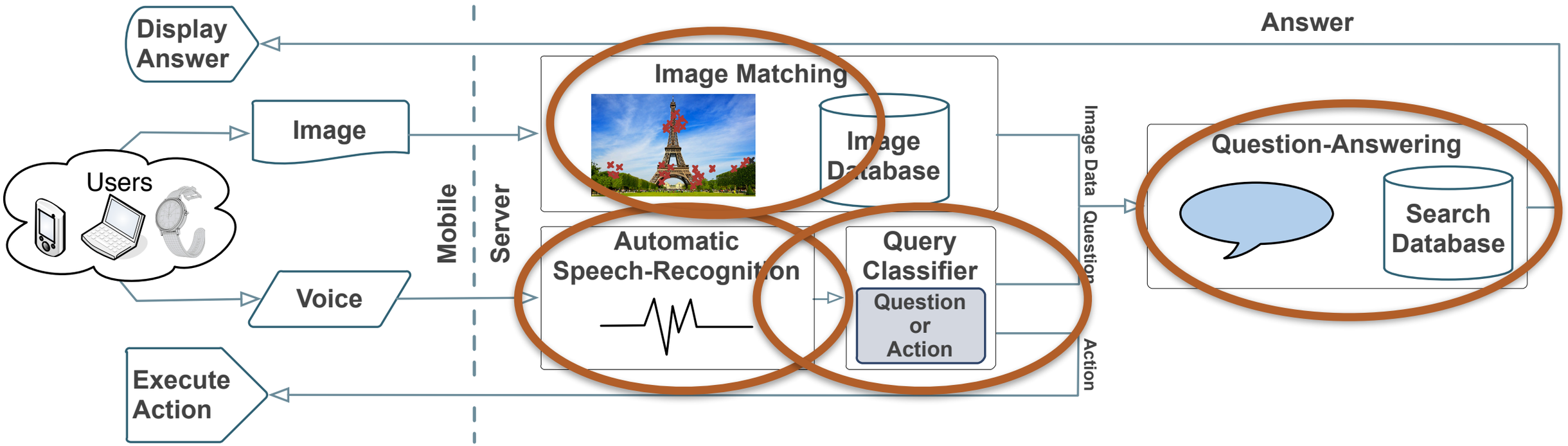
- What is “deep” about a DNN?
  - More layers
  - Often, bigger layers also
- Example — DNN in Lucida ASR



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

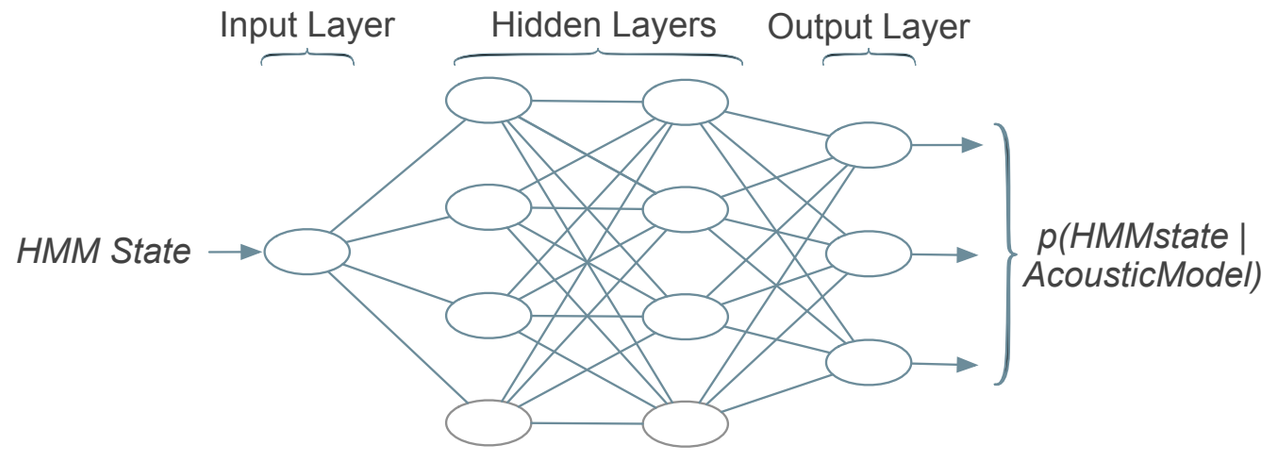
DNNs in the IPA pipeline



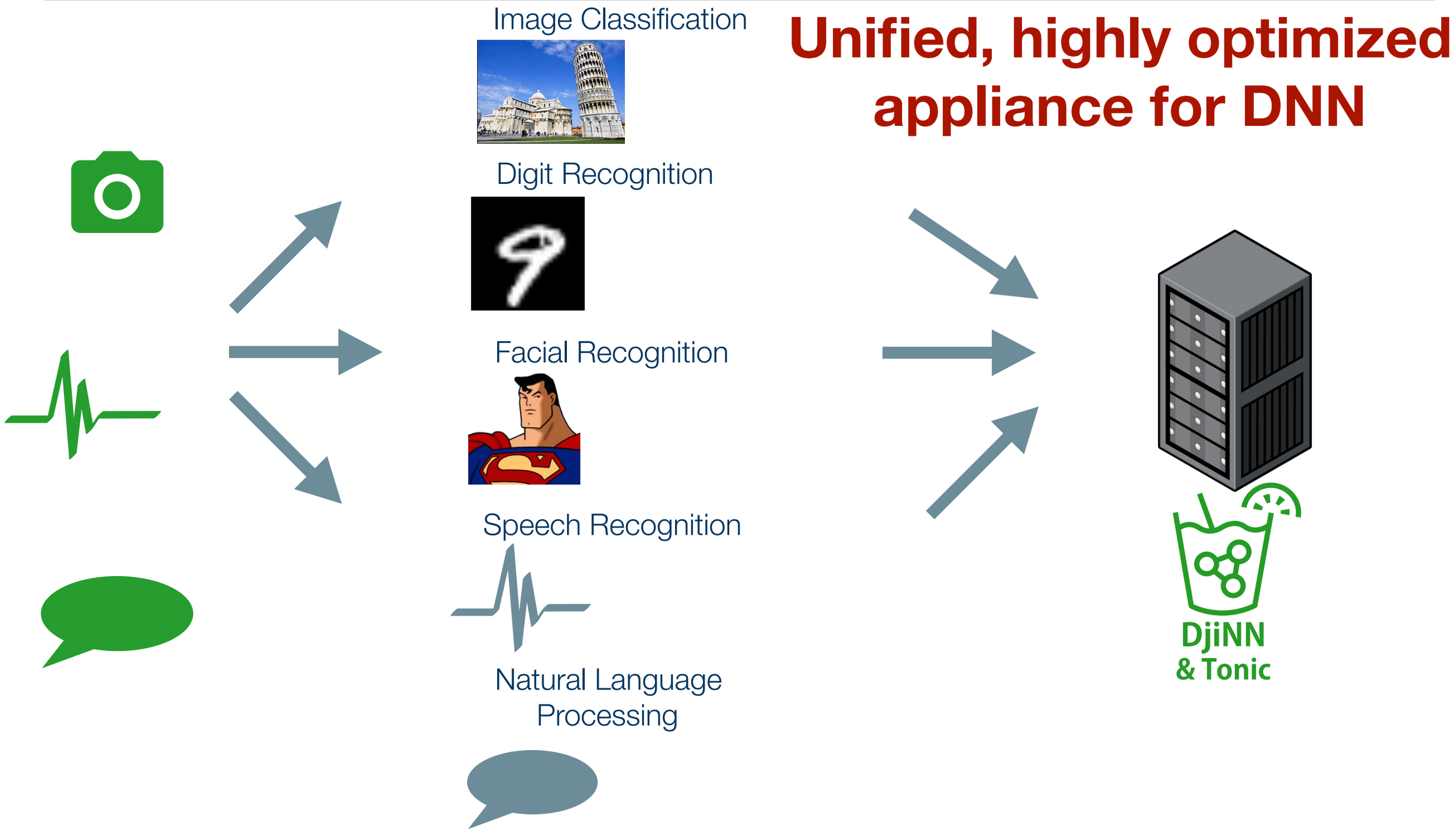
Studying DNNs as an architect is a challenge

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 DjiNN Tutorial

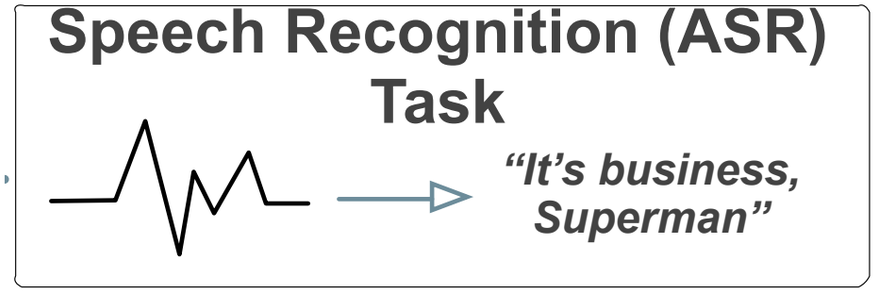
Topic: Tonic Suite Components
Speaker: Johann Hauswald

COMPUTER SCIENCE
& ENGINEERING
UNIVERSITY OF MICHIGAN

ClarityLab

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)

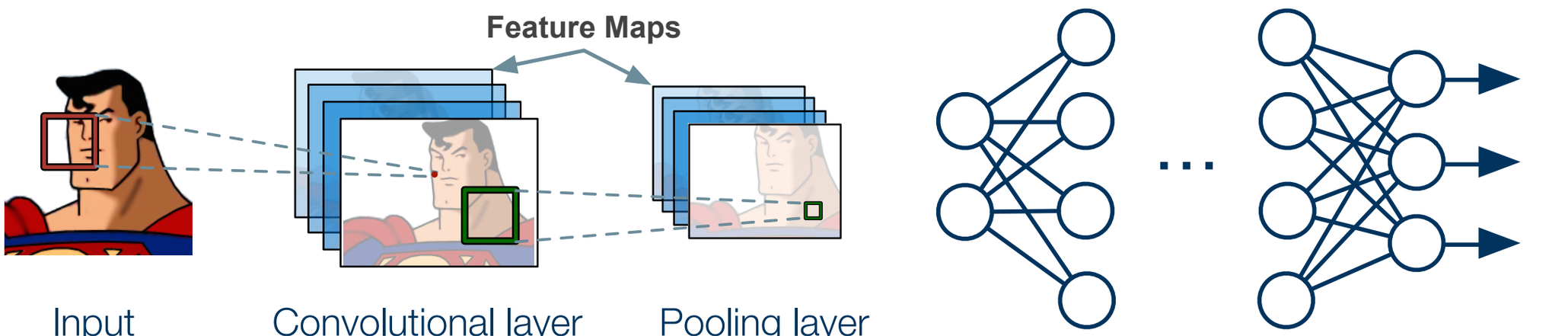


Natural Language Processing Task

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

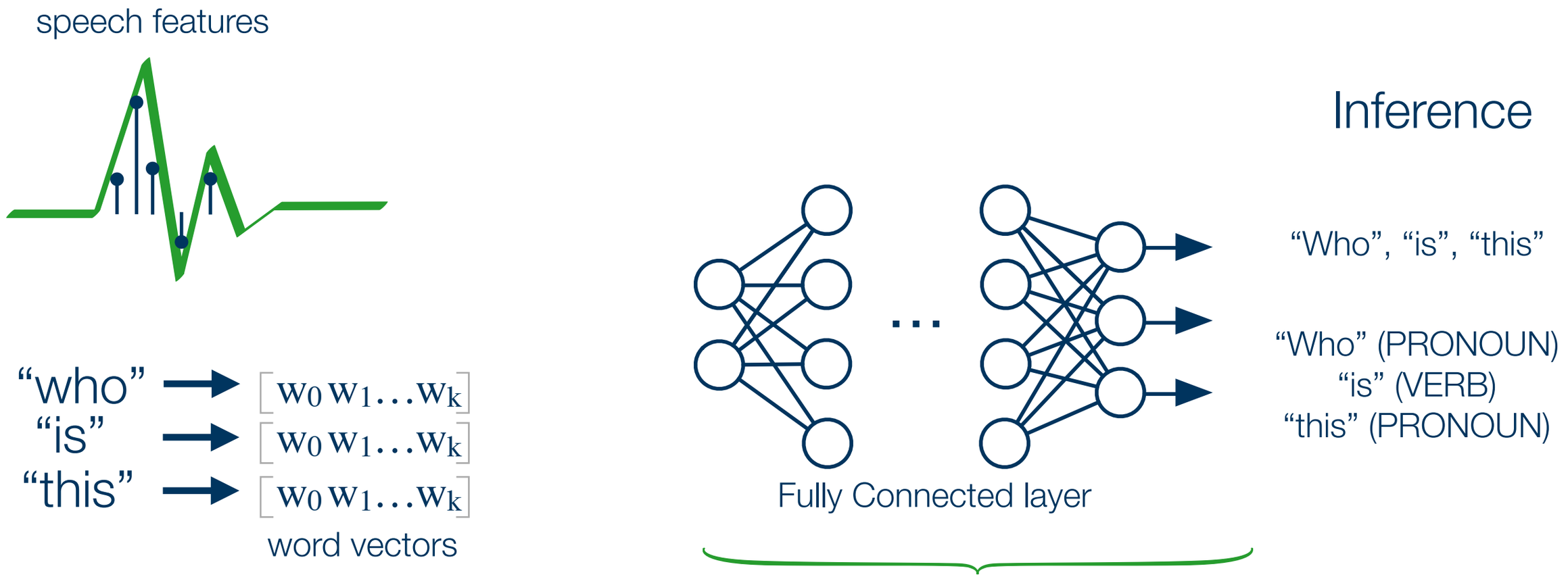
Deep Neural Networks (DNNs)

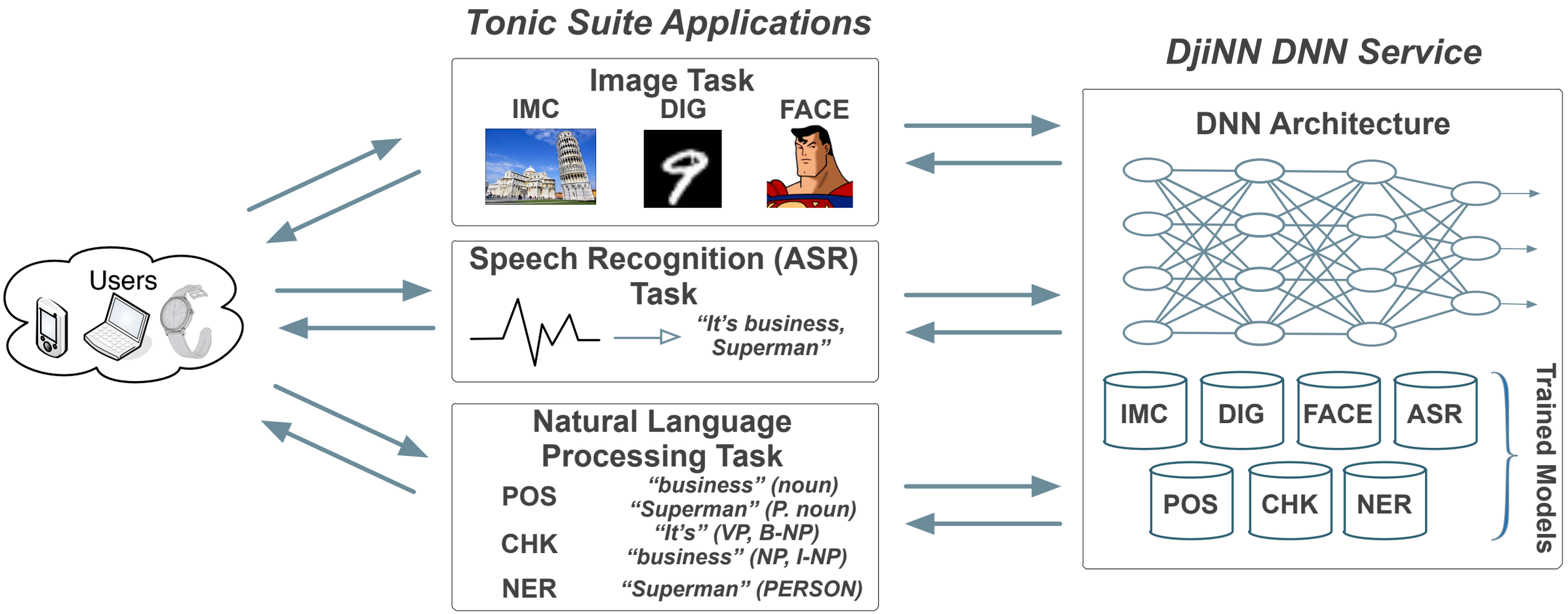
Inference



Network Architecture

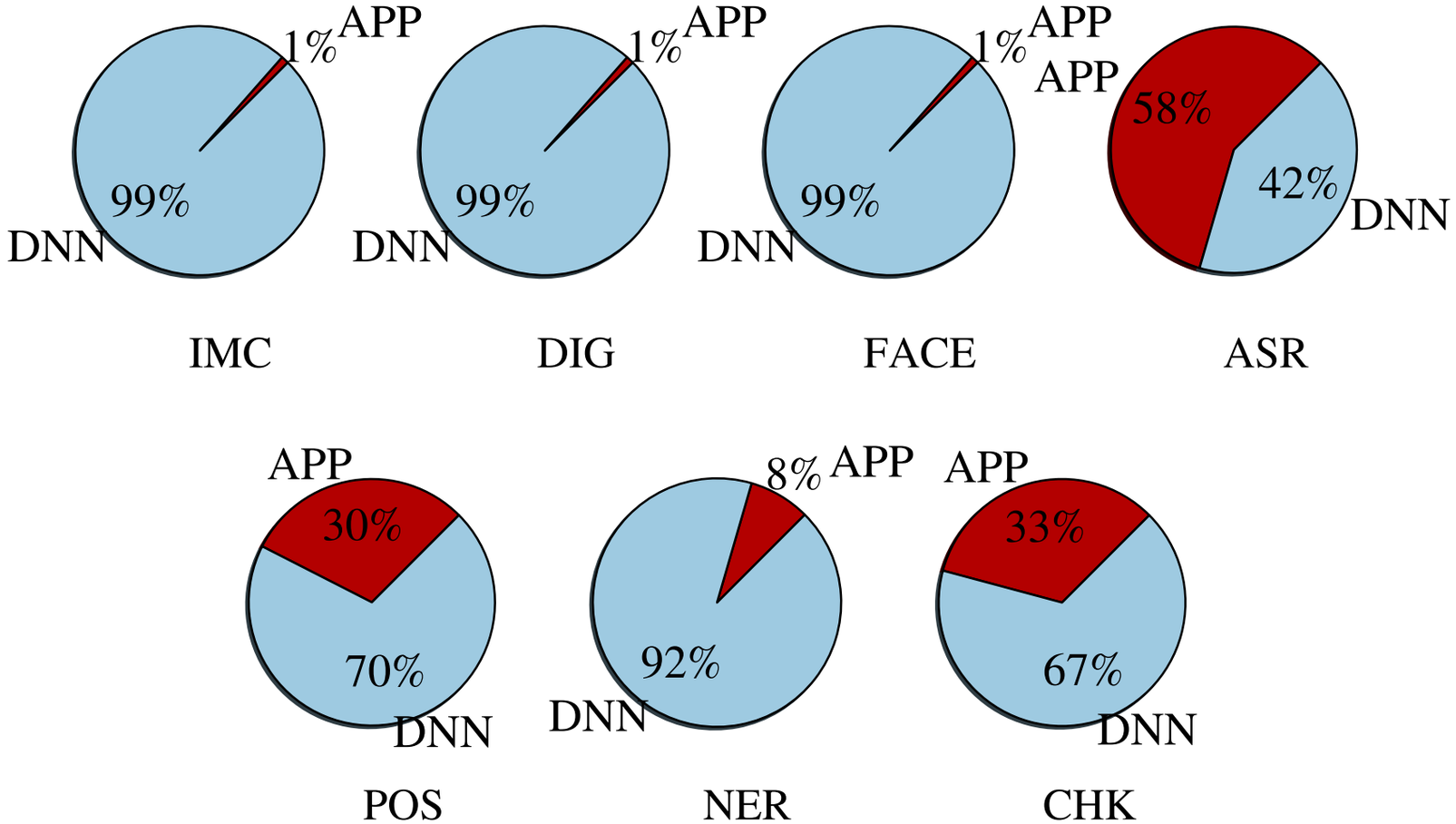
Deep Neural Networks (DNNs)




Tonic Suite Applications
Image Task
IMC
DIG
FACE
Speech Recognition (ASR) Task
"It's business, Superman"
Natural Language Processing Task
POS
"business" (noun)
"Superman" (P. noun)
CHK
"It's" (VP, B-NP)
"business" (NP, I-NP)
NER
"Superman" (PERSON)
Users
DjiNN DNN Service
DNN Architecture
IMC
DIG
FACE
ASR
POS
CHK
NER
Trained Models

Basic characterization

- Single CPU



Not all DNN services are created equal

Accelerating DNN can provide significant speedup

Tonic Suite

Lucida and DjiNN Tutorial

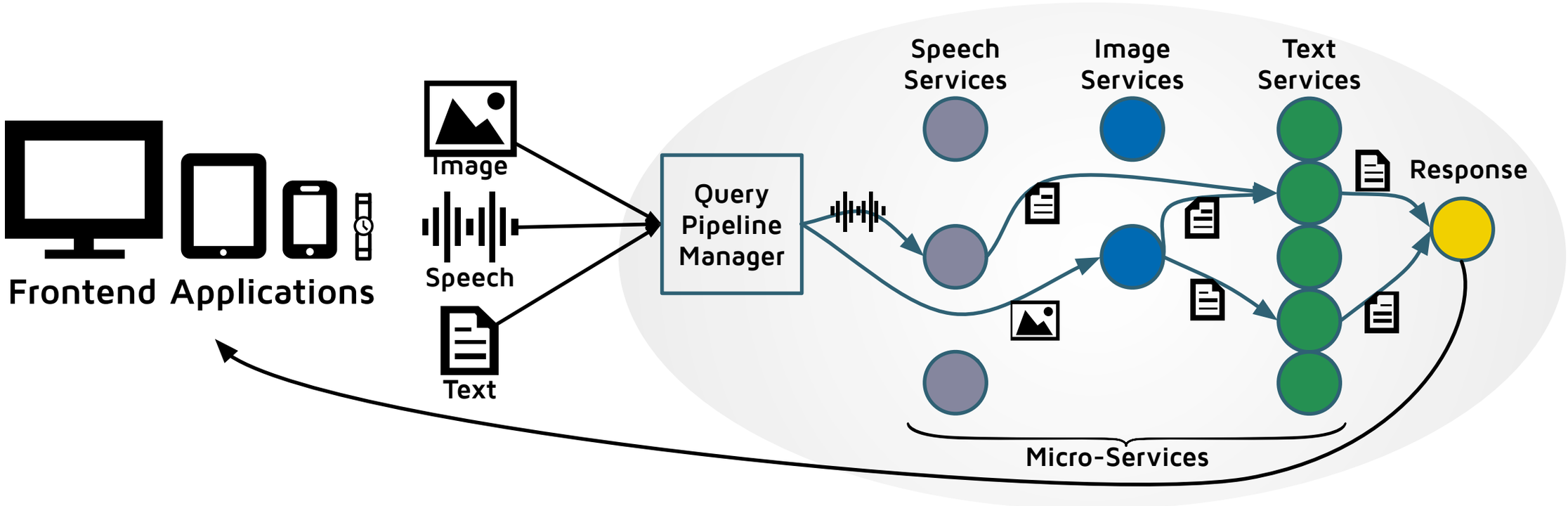
Topic: Introduction to LucidaEco
Speaker: Michael A. Laurenzano

COMPUTER SCIENCE
& ENGINEERING
UNIVERSITY OF MICHIGAN

ClarityLab

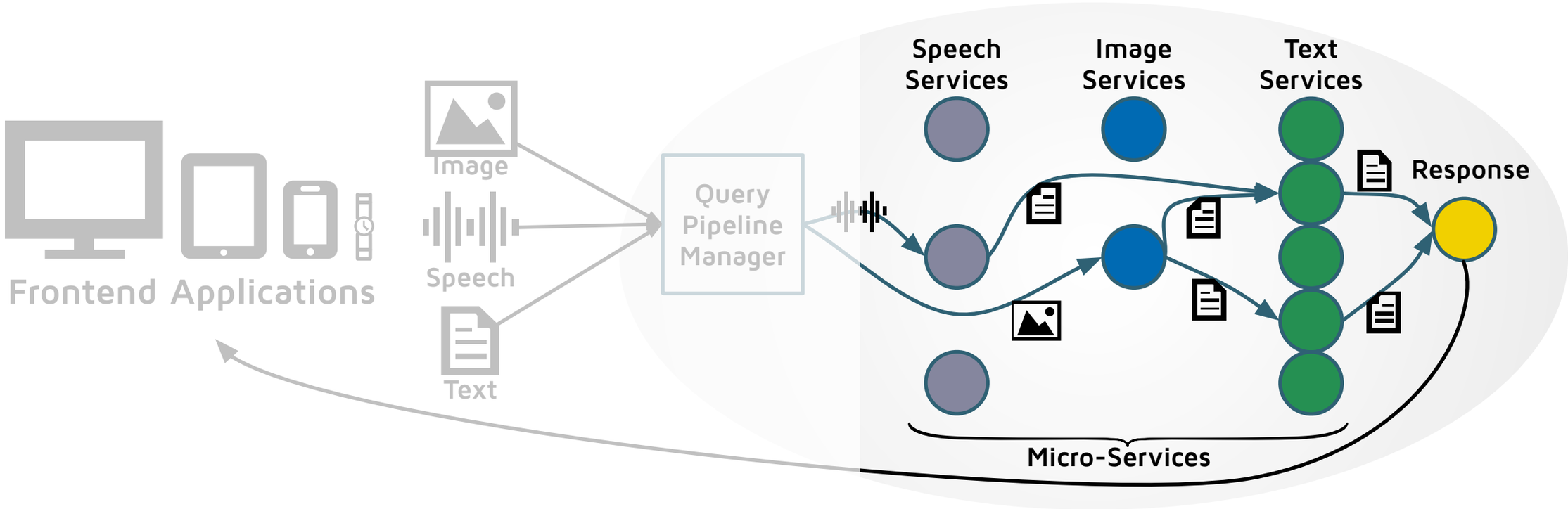
What is LucidaEco?

- 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



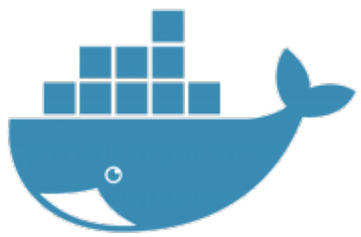
Micro-service Architecture

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



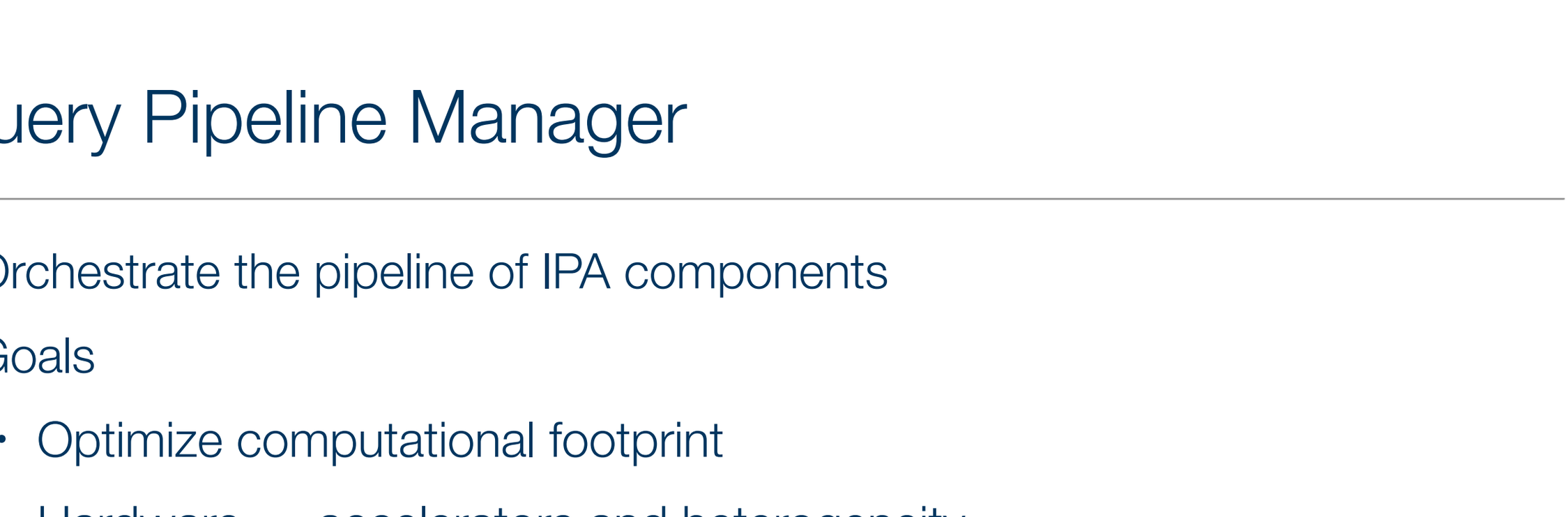
Micro-service Design

- 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

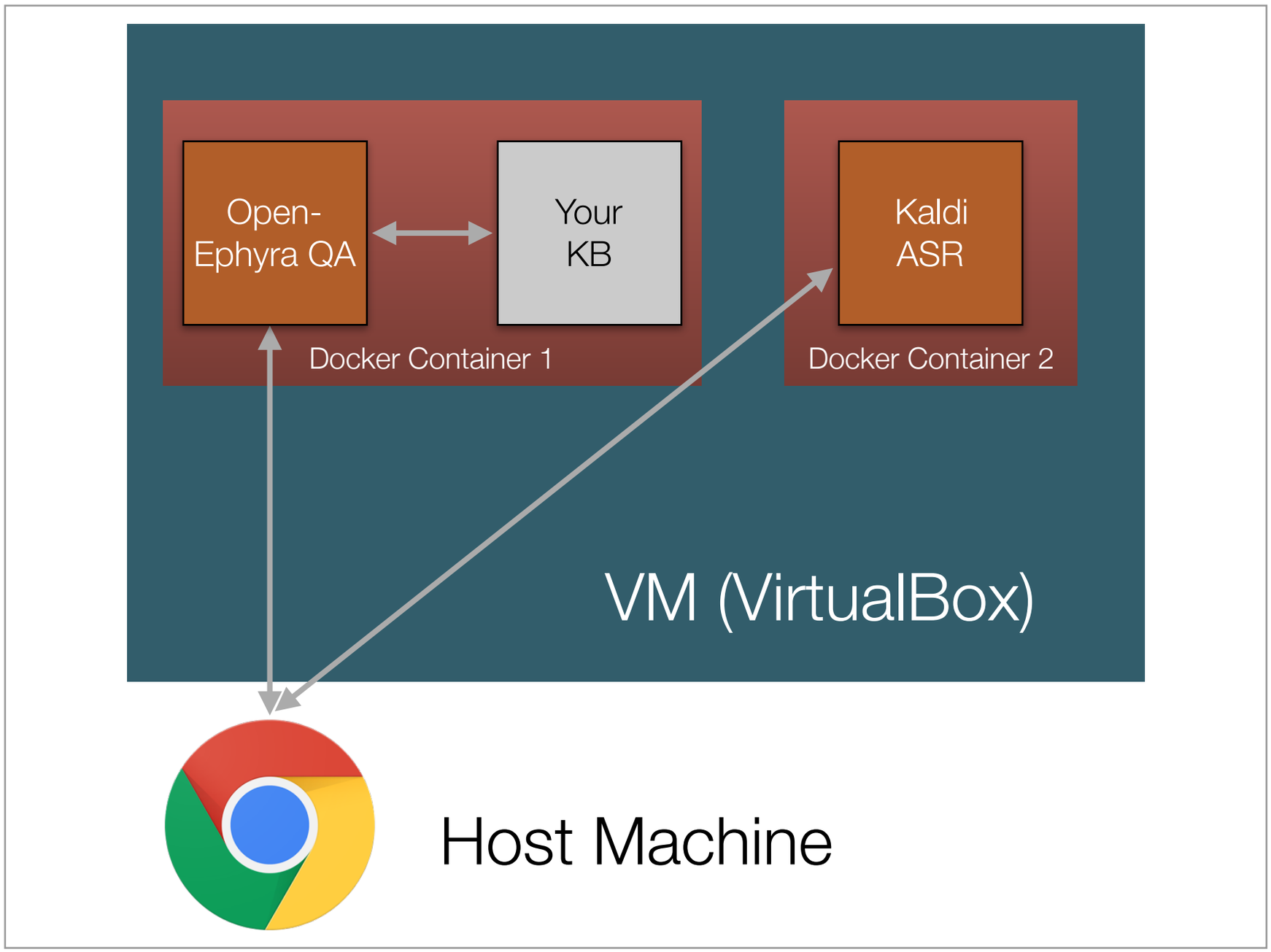
- Micro-service ecosystem
  - ASR (Kaldi, Sphinx, different languages)
  - QA (OpenEphyra, Qanta, YodaQA)
  - Machine Translation, Sentiment, Summarization
  - Image matching, object recognition
- Frontend apps — web, mobile, tablet; others coming
- LucidaEco as a research platform
  - AI algorithms/implementations 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

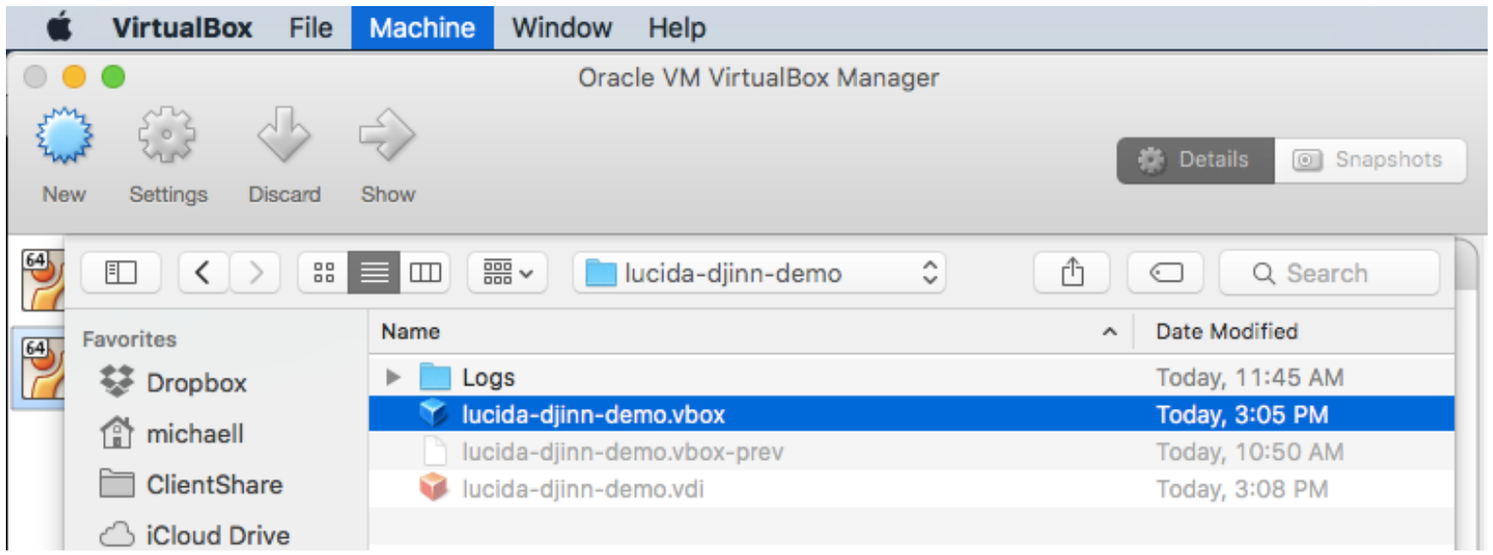


Setup Prerequisites

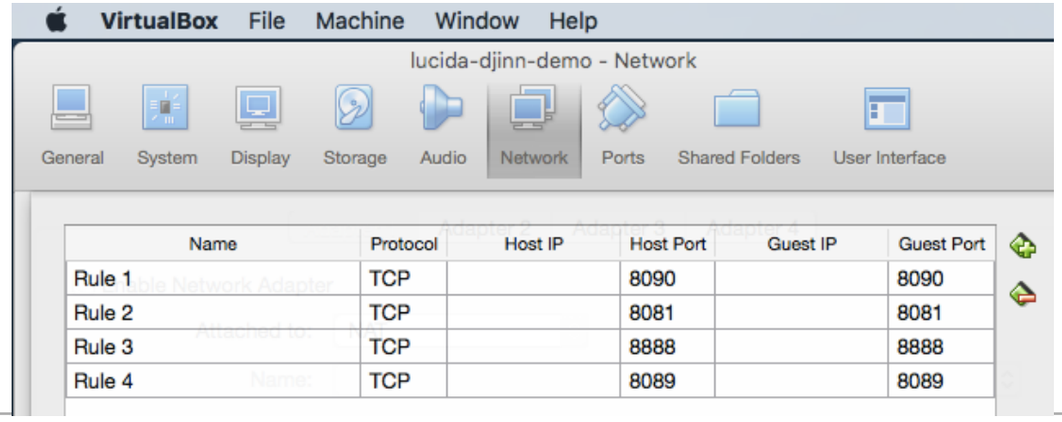
- Install VirtualBox
- Install Chrome
- Copy the tutorial VM — USB thumb drive

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 ~/hpca-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/