Gridspace

SCALED SPEECH AND LANGUAGE TECHNOLOGY IN THE CONTACT CENTER

Anthony Scodary
Wonkyum Lee
Alex Barron

March 18, 2019

[Video on Vimeo Here]()

Interactive

Long-Form

Conversational

Speech &

Language

Application

Roadmap

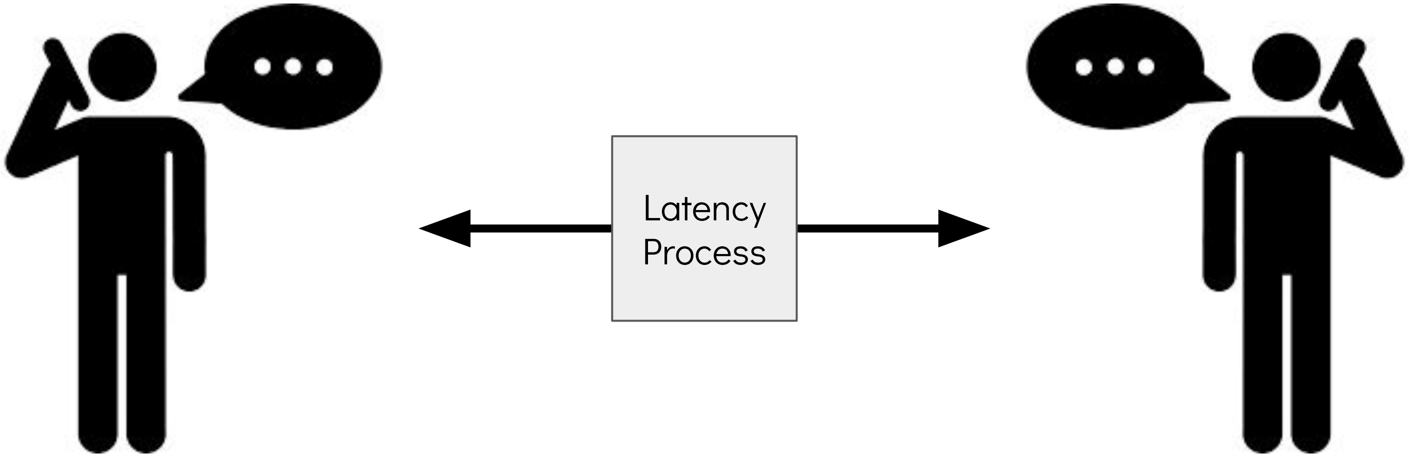
1. Fundamental Challenges
2. Case Study: Closed Loop Feedback in Speech Recognition Training
3. Case Study: Latency and Randomness in Speech Synthesis Inference

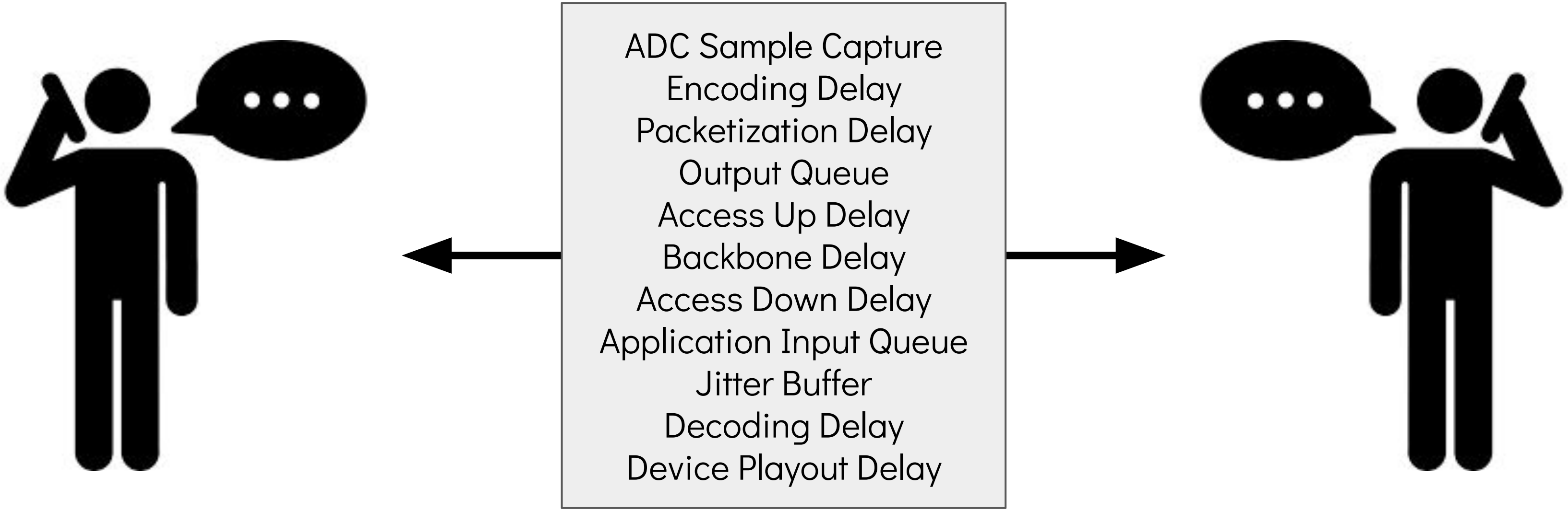
FUNDAMENTAL CHALLENGES

	Time	Predictability	Complexity
Optimize	Latency ⌛	Stochasticity 🎲	Scale ⚖
Manage	Jitter & Aperiodicity 🌊	Nonlinearity 🦋	Interactivity 🤖

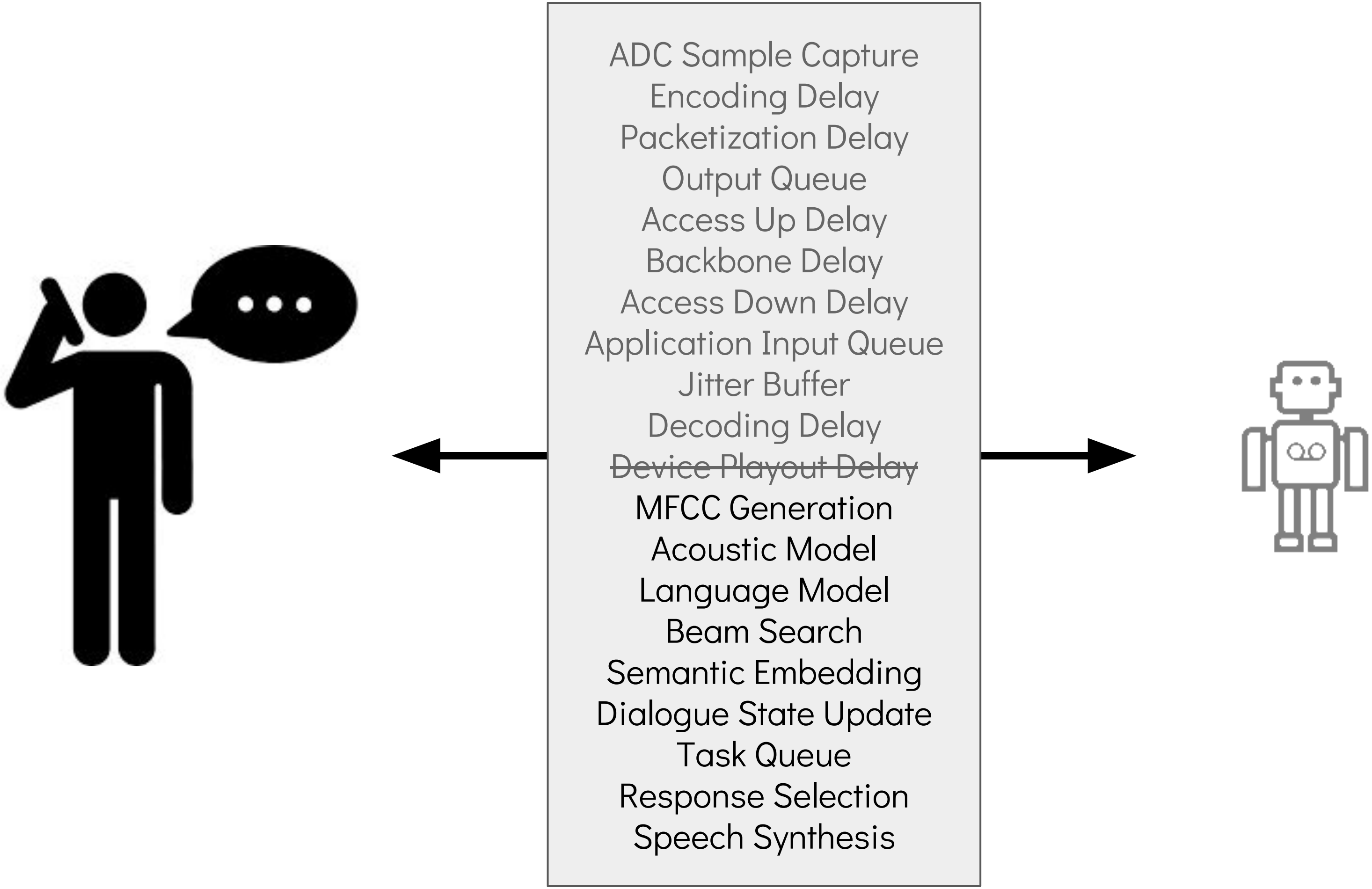
LATENCY

	Time	**Predictability**	**Complexity**
Optimize	Latency ⌛	Stochasticity 🎲	Scale ⚖️
Manage	Jitter & Aperiodicity 🌊	Nonlinearity 🦋	Interactivity 🤖


Latency
Process



ADC Sample Capture
Encoding Delay
Packetization Delay
Output Queue
Access Up Delay
Backbone Delay
Access Down Delay
Application Input Queue
Jitter Buffer
Decoding Delay
Device Playout Delay



ADC Sample Capture
Encoding Delay
Packetization Delay
Output Queue
Access Up Delay
Backbone Delay
Access Down Delay
Application Input Queue
Jitter Buffer
Decoding Delay
~~Device Playout Delay~~
MFCC Generation
Acoustic Model
Language Model
Beam Search
Semantic Embedding
Dialogue State Update
Task Queue
Response Selection
Speech Synthesis

INTERACTIVITY

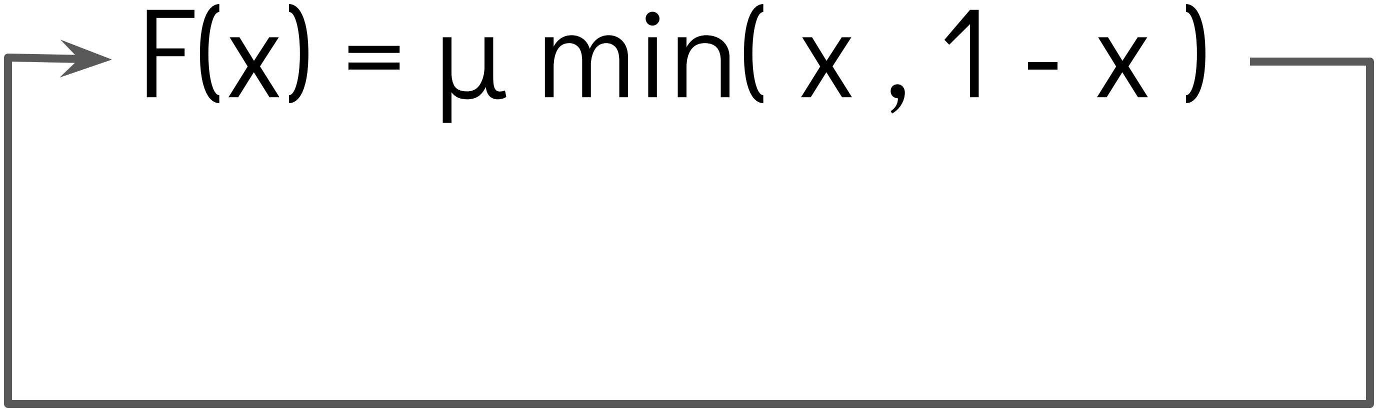
	Time	**Predictability**	**Complexity**
Optimize	Latency ⌛	Stochasticity 🎲	Scale ⚖️
Manage	Jitter & Aperiodicity 🌊	Nonlinearity 🦋	Interactivity

Simple function

The Tent Map

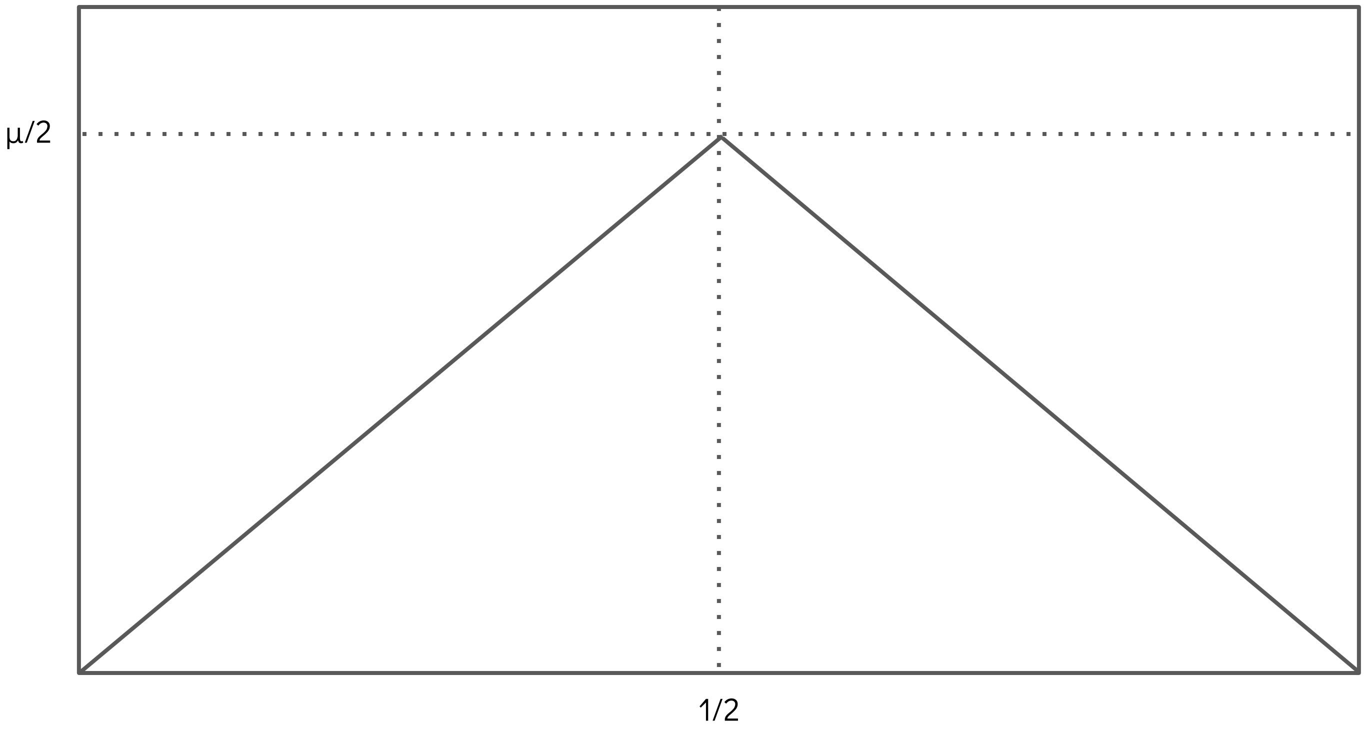
$$F(x) = \mu \min(x , 1 - x)$$

The Tent Map

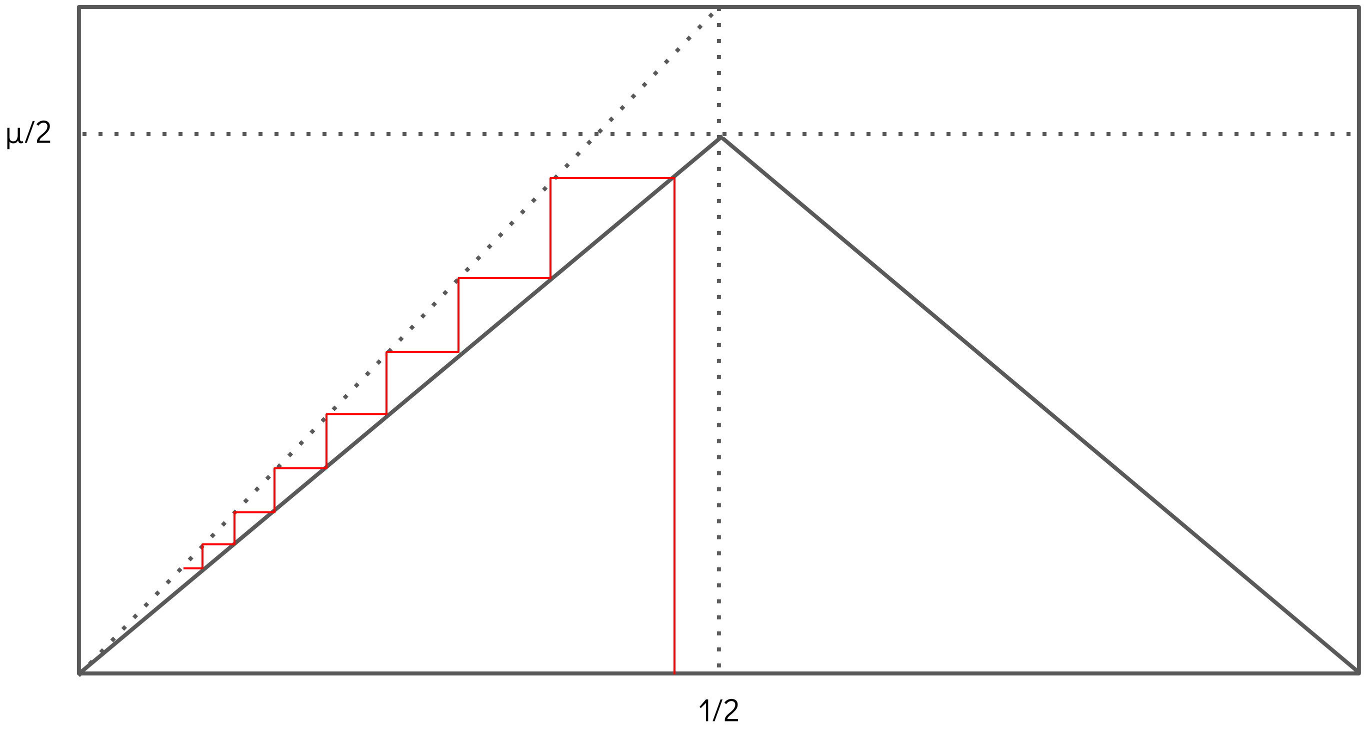


$$F(x) = \mu \min(x , 1 - x)$$

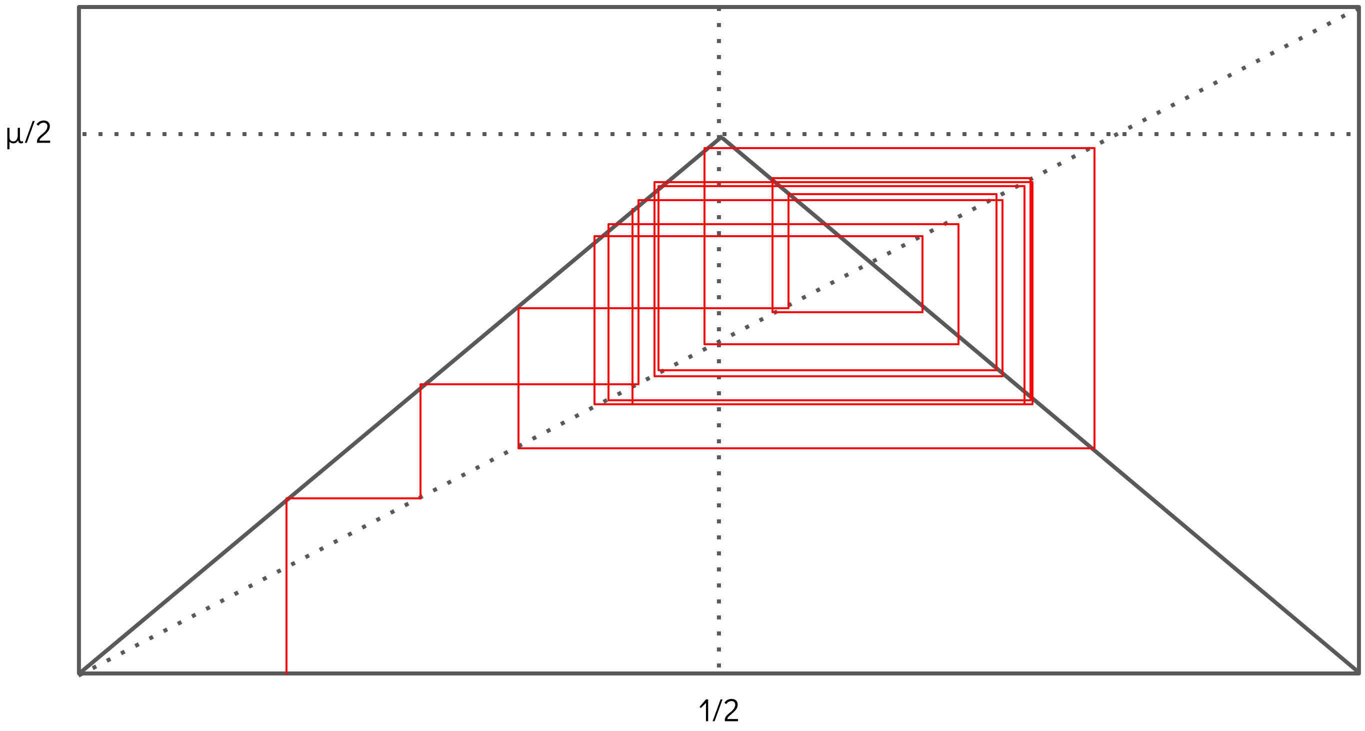
The Tent Map



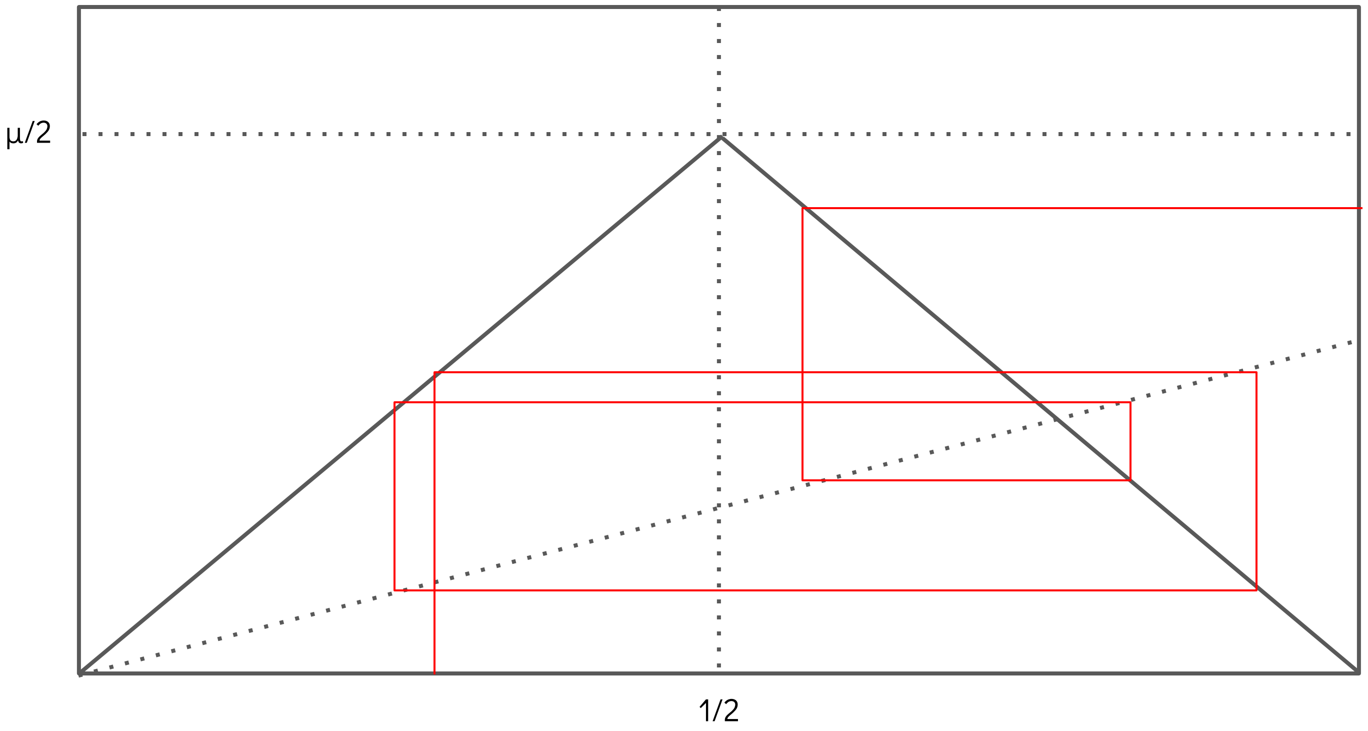
The Tent Map (μ < 1)



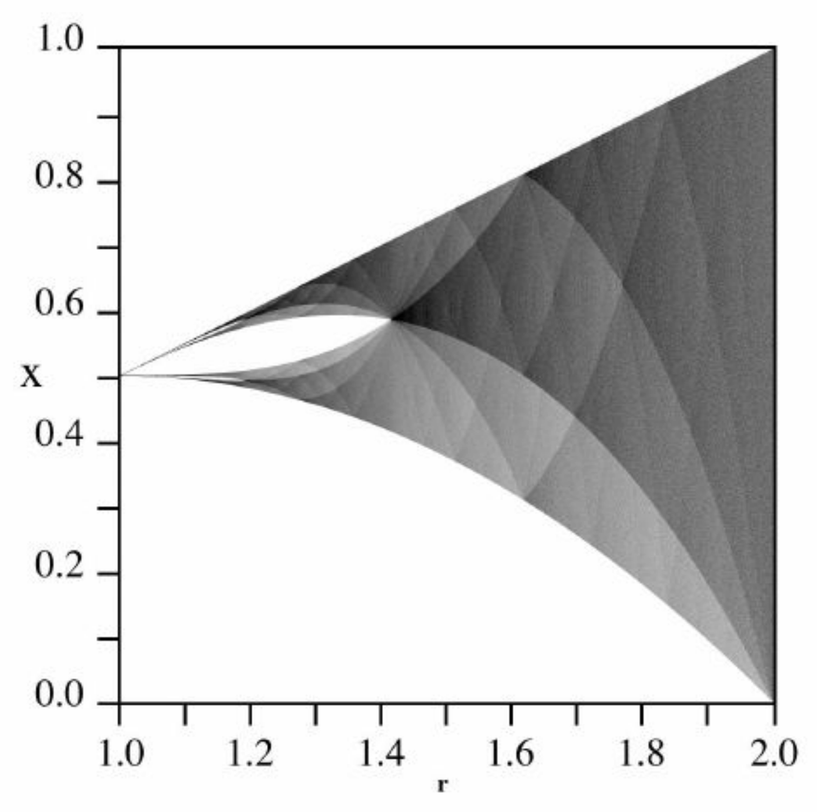
The Tent Map (μ > 1)



The Tent Map (μ >> 1)

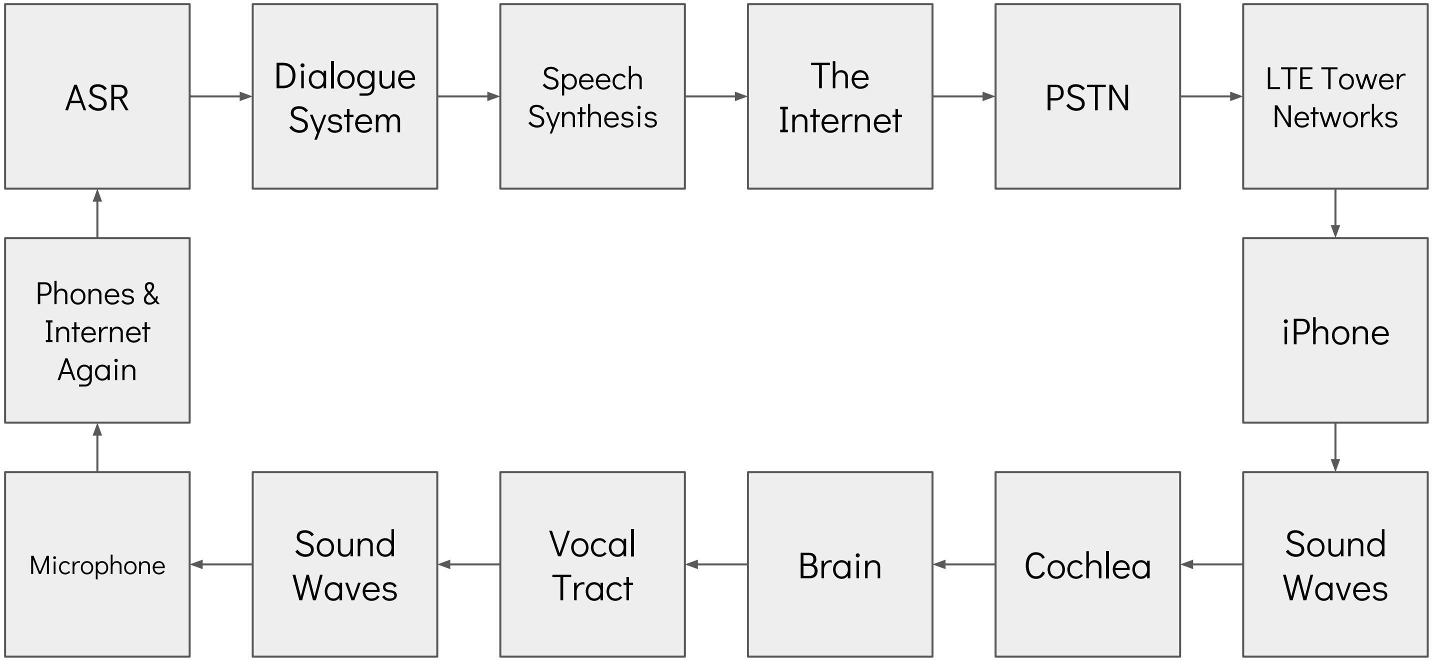


The Tent Map
Bifurcation Diagram



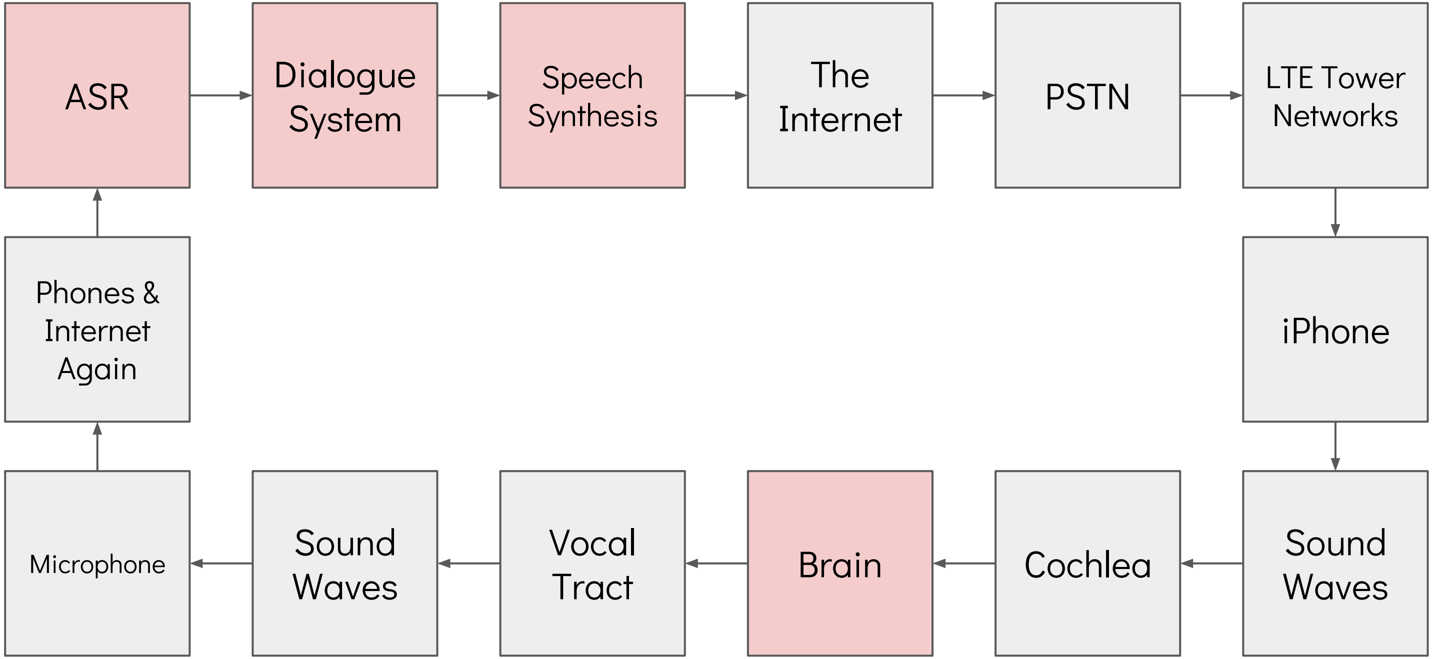
More Complex Function

Full System Feedback Loop

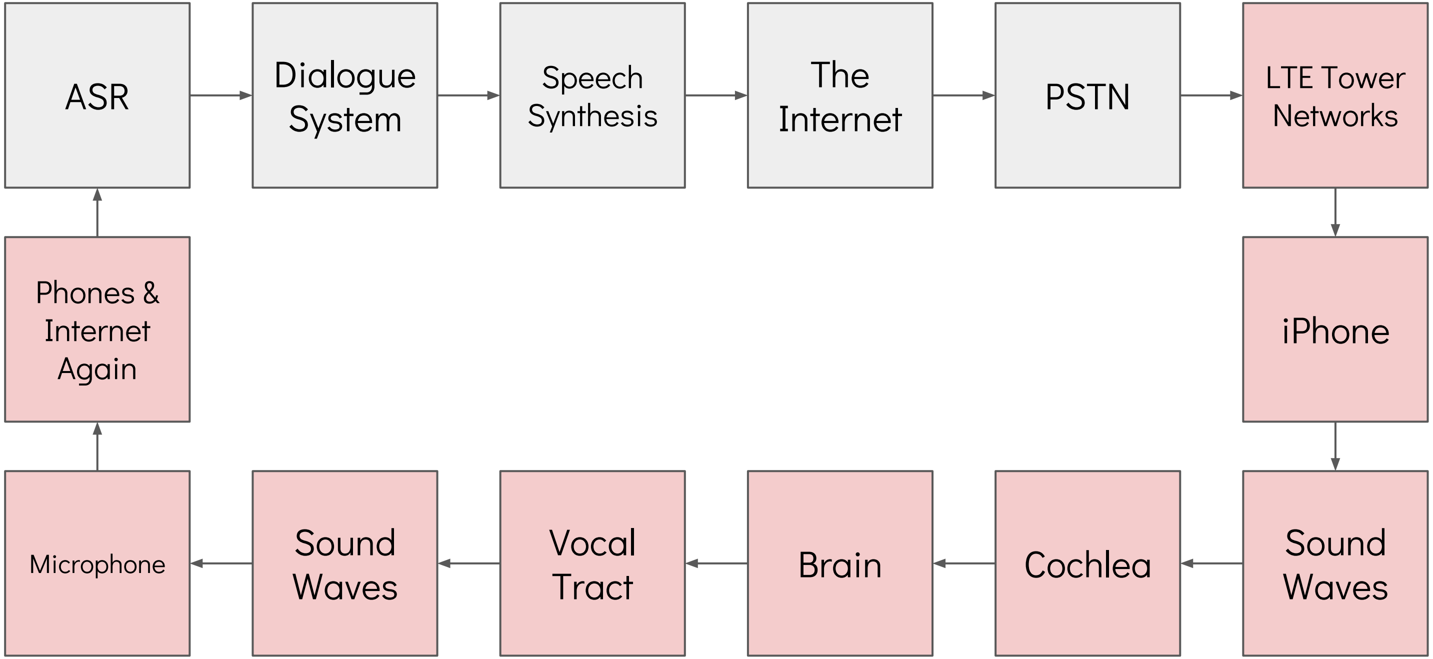


ASR → Dialogue System → Speech Synthesis → The Internet → PSTN → LTE Tower Networks → iPhone → Sound Waves → Cochlea → Brain → Vocal Tract → Sound Waves → Microphone → Phones & Internet Again → ASR

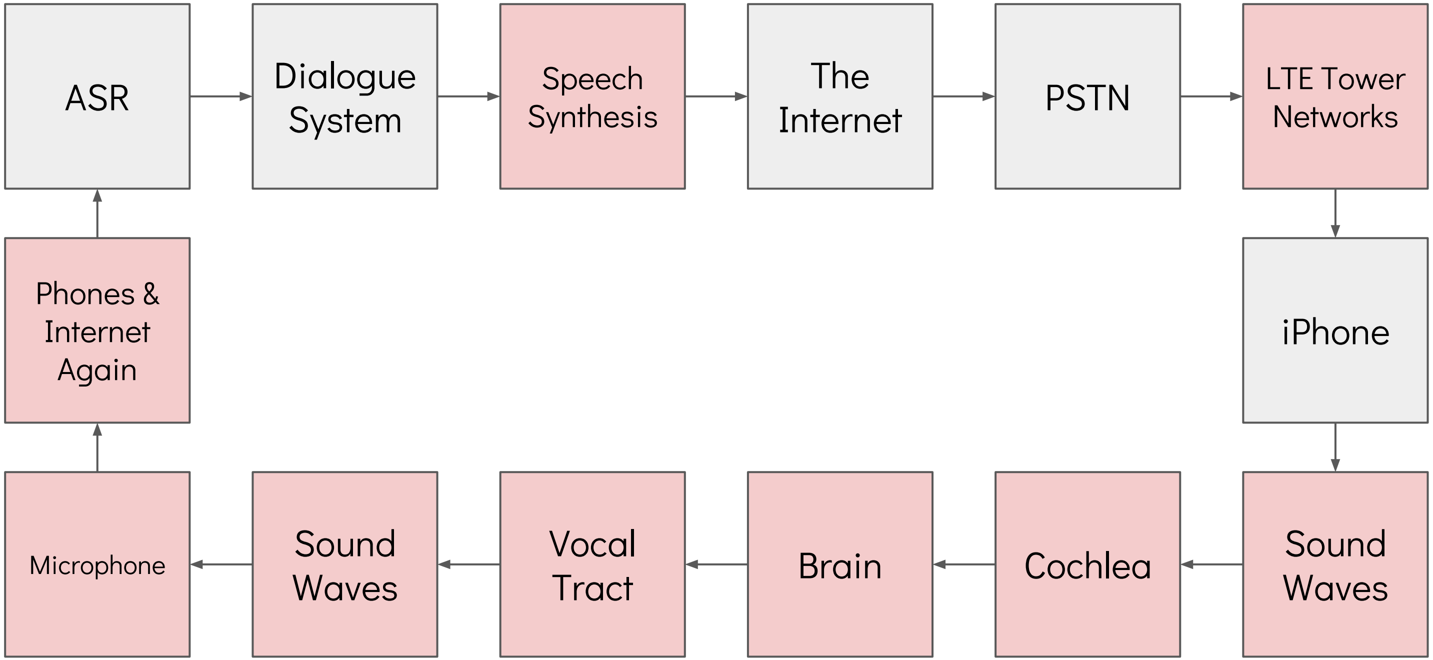
Feature a Neural Network



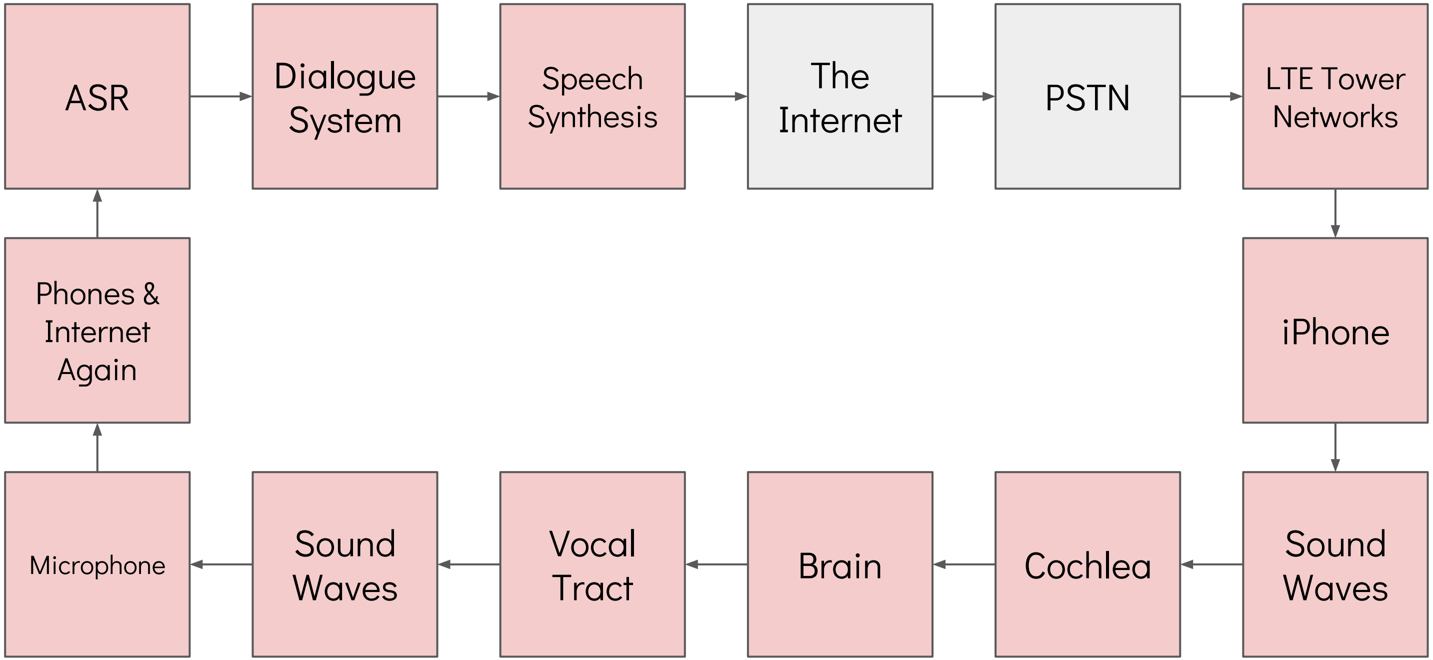
Physical Effects



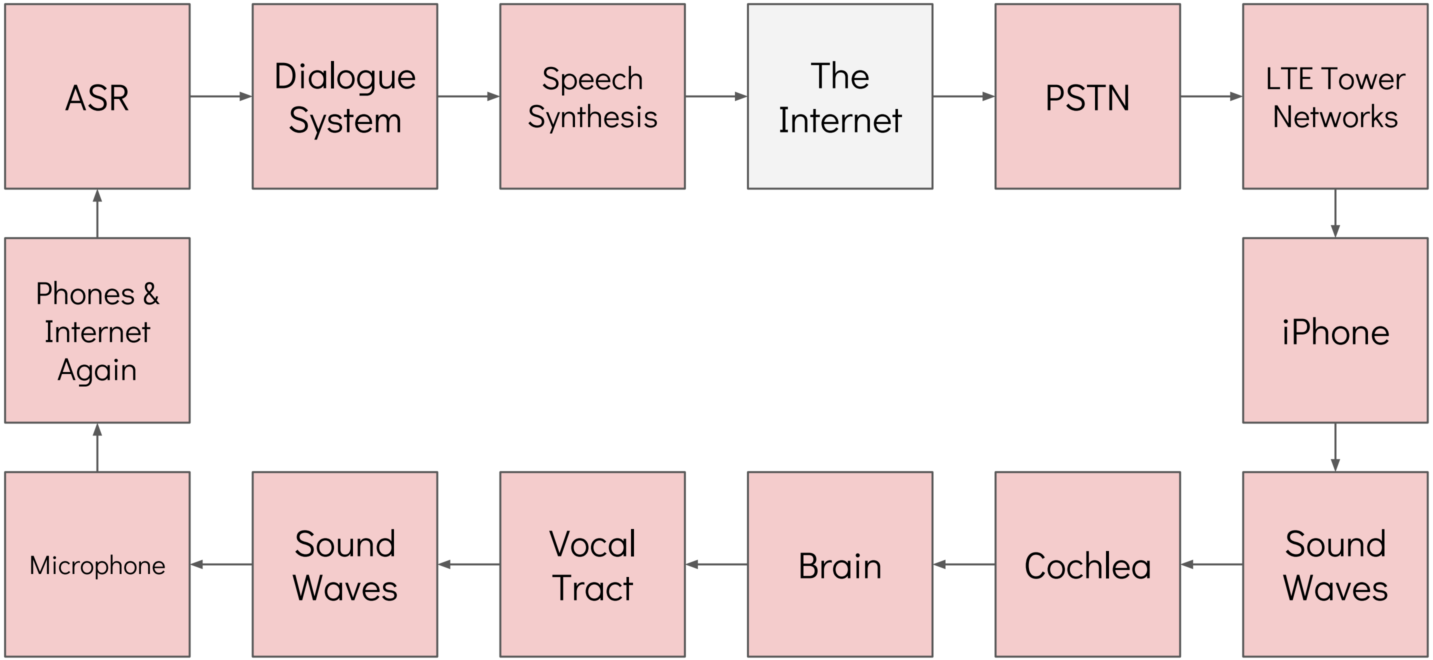
Noisy & Stochastic



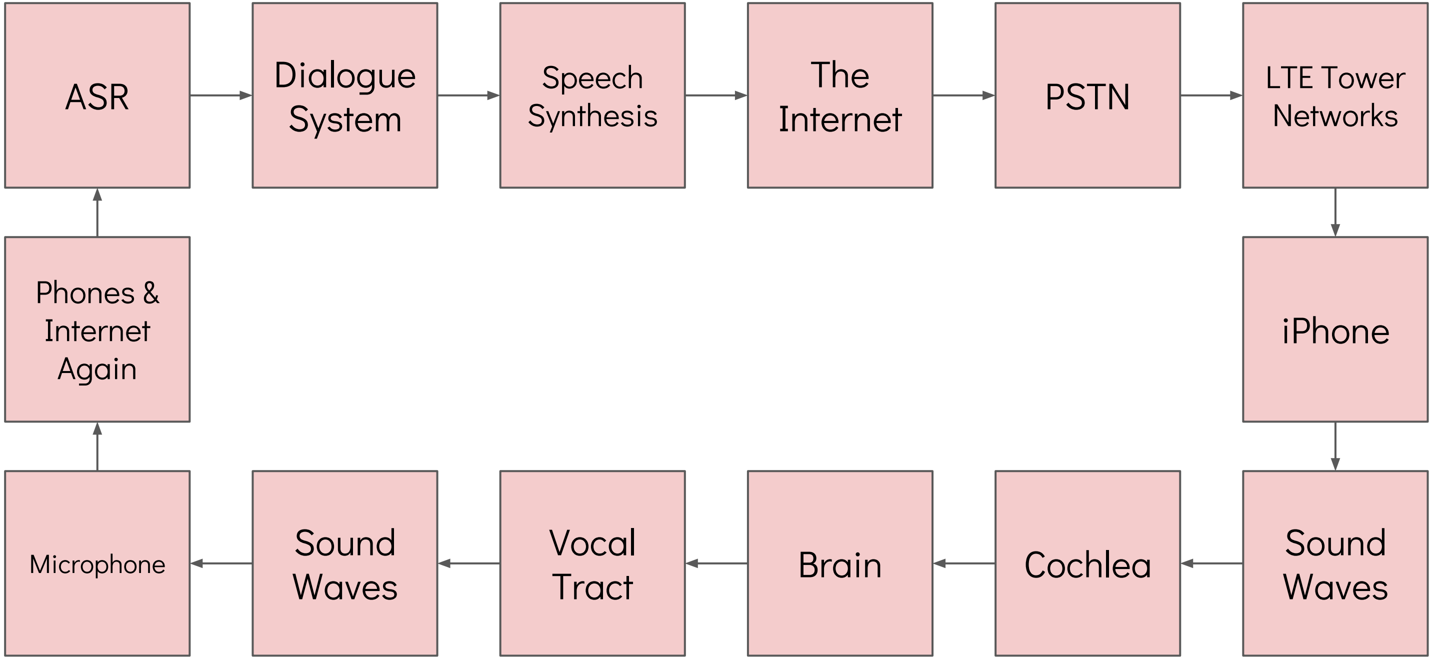
Nonlinearity and Distortion



Information Loss



Consumes Latency Budget



ASR → Dialogue System → Speech Synthesis → The Internet → PSTN → LTE Tower Networks → iPhone → Sound Waves → Cochlea → Brain → Vocal Tract → Sound Waves → Microphone → Phones & Internet Again → ASR

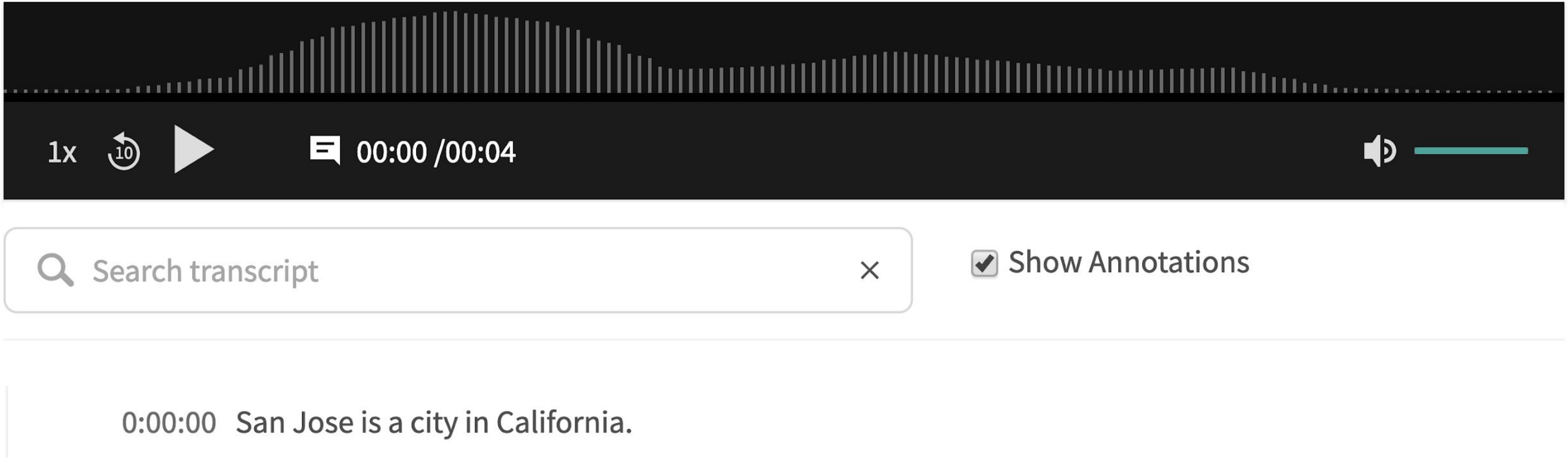
NONLINEARITY

	Time	Predictability	Complexity
Optimize	Latency ⌛	Stochasticity 🎲	Scale ⚖️
Manage	Jitter & Aperiodicity 🌊	Nonlinearity	Interactivity 🤖

Linearity

I. $F(x + y) = F(x) + F(y)$

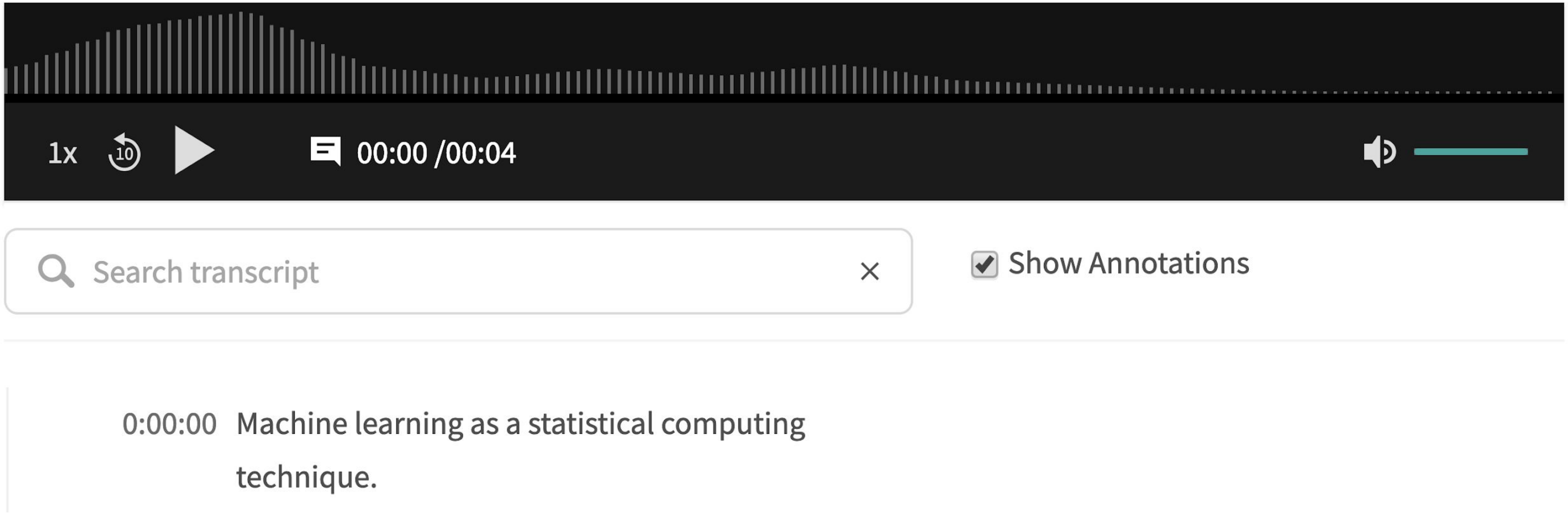
F(x)



☑ Show Annotations

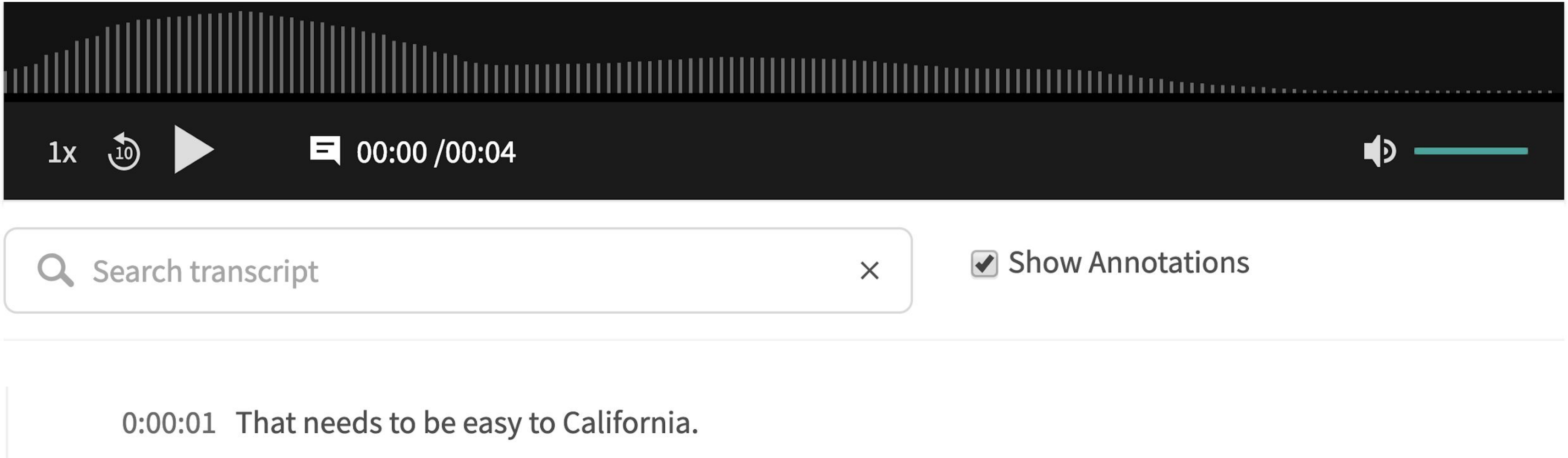
0:00:00 San Jose is a city in California.

F(y)



0:00:00 Machine learning as a statistical computing technique.

F(x+y)

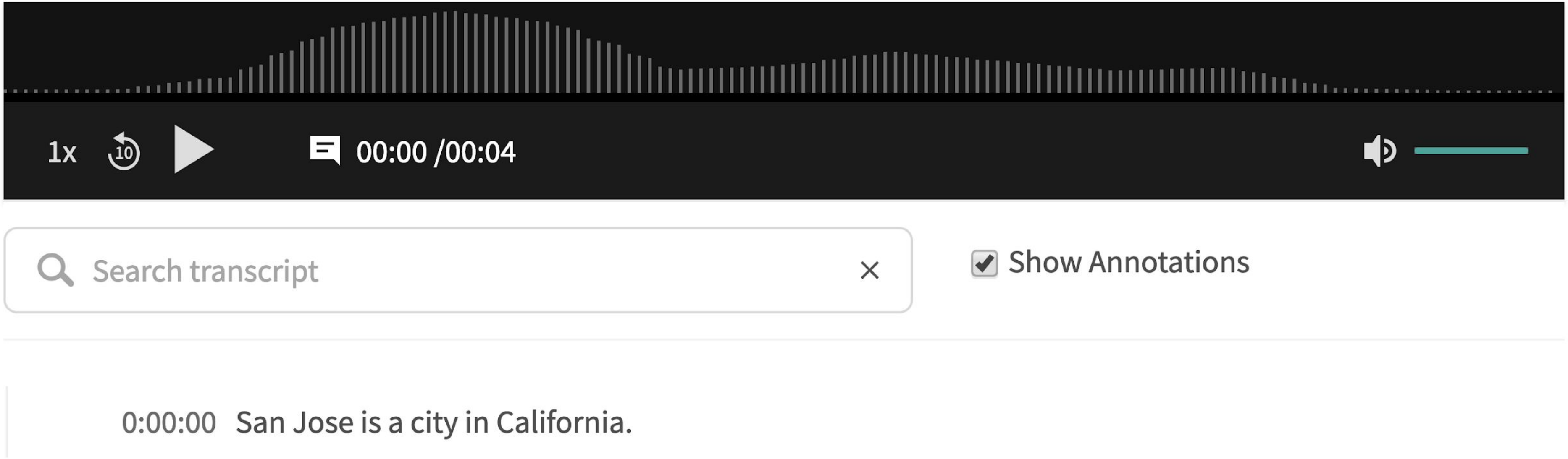


0:00:01 That needs to be easy to California.

Linearity

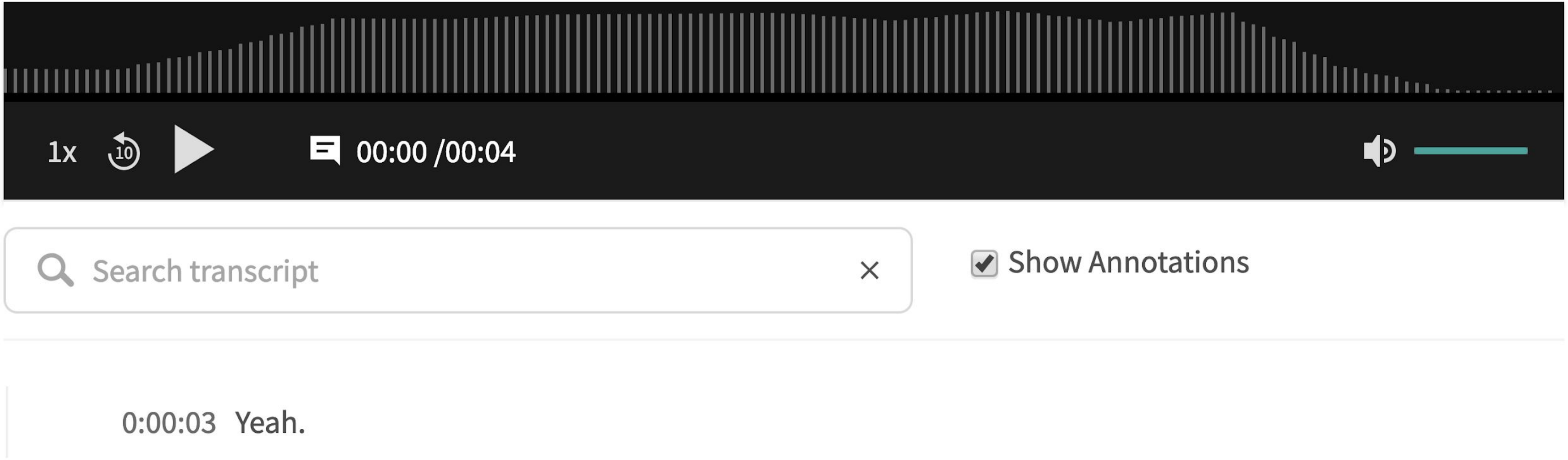
II. $F(ax) = aF(x)$

F(x)



0:00:00 San Jose is a city in California.

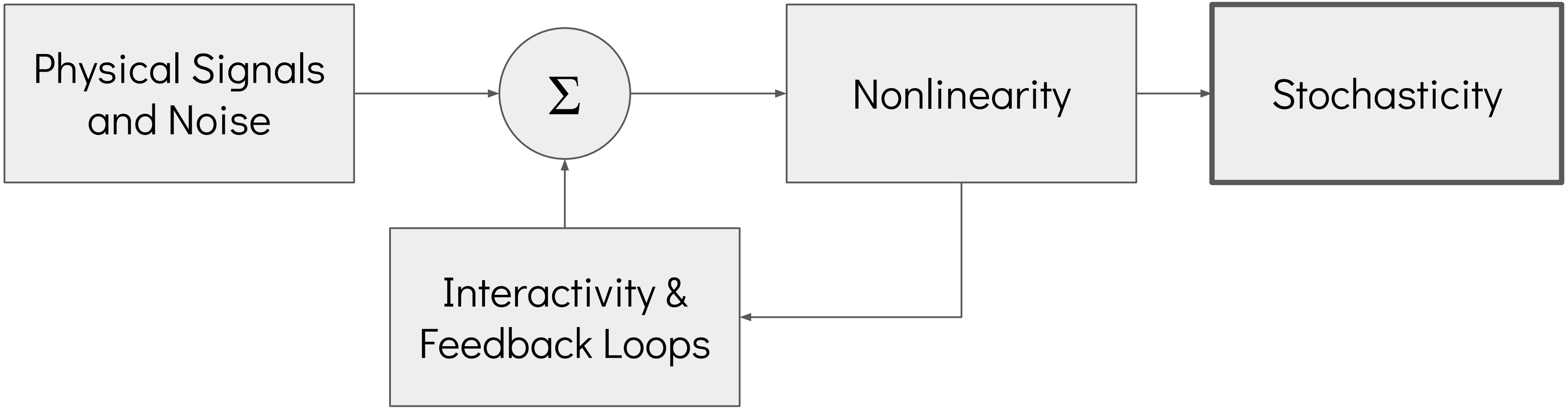
F(ax)



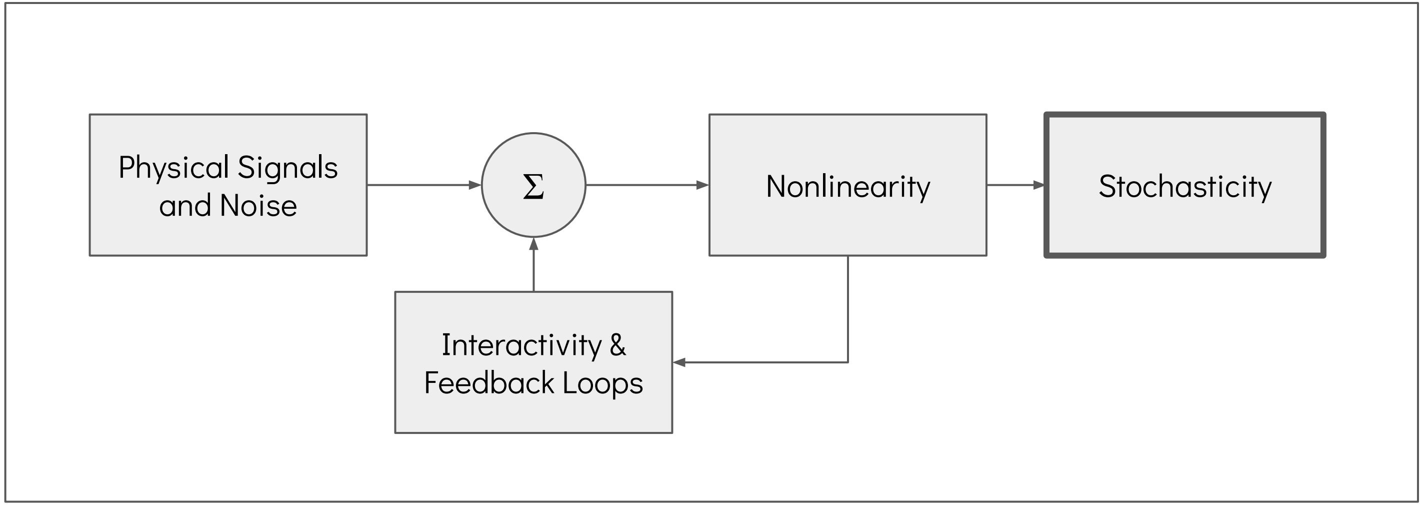
0:00:03 Yeah.

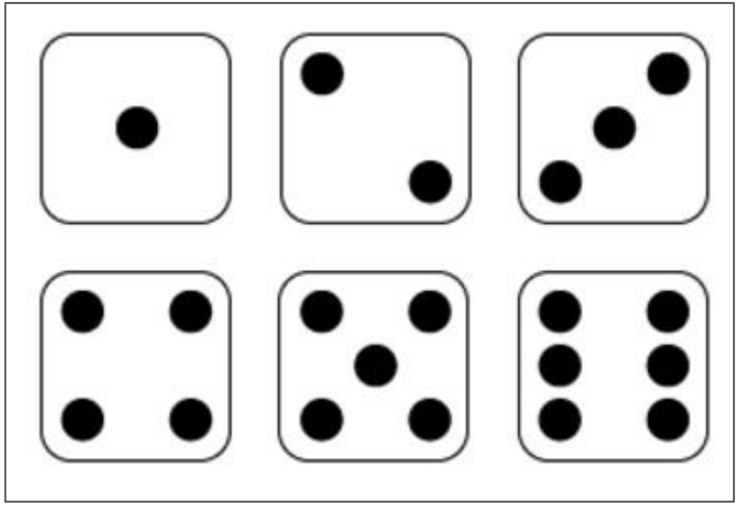
STOCHASTICITY

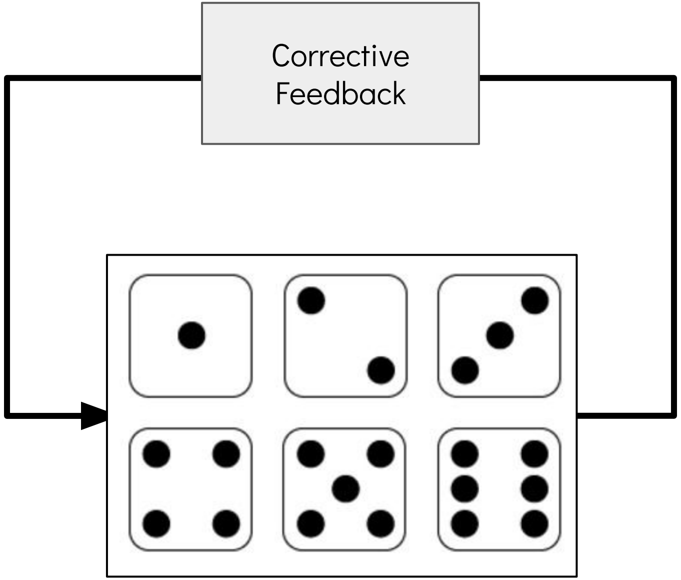
	Time	**Predictability**	**Complexity**
Optimize	Latency ⌛	Stochasticity 🎲	Scale ⚖️
Manage	Jitter & Aperiodicity 🌊	Nonlinearity 🦋	Interactivity 🤖



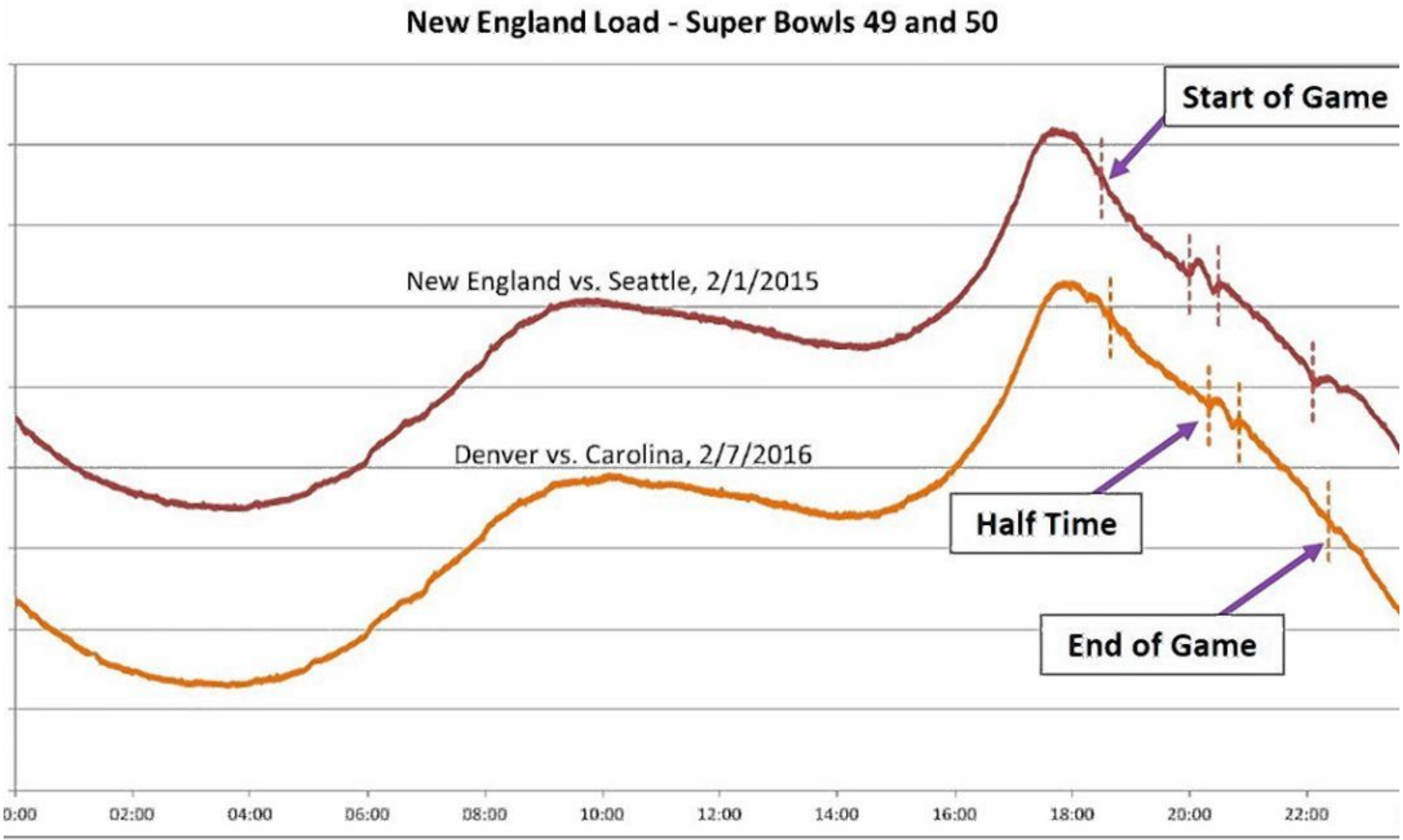
Physical Signals and Noise
Σ
Nonlinearity
Stochasticity
Interactivity & Feedback Loops


Physical Signals and Noise
Σ
Nonlinearity
Stochasticity
Interactivity & Feedback Loops

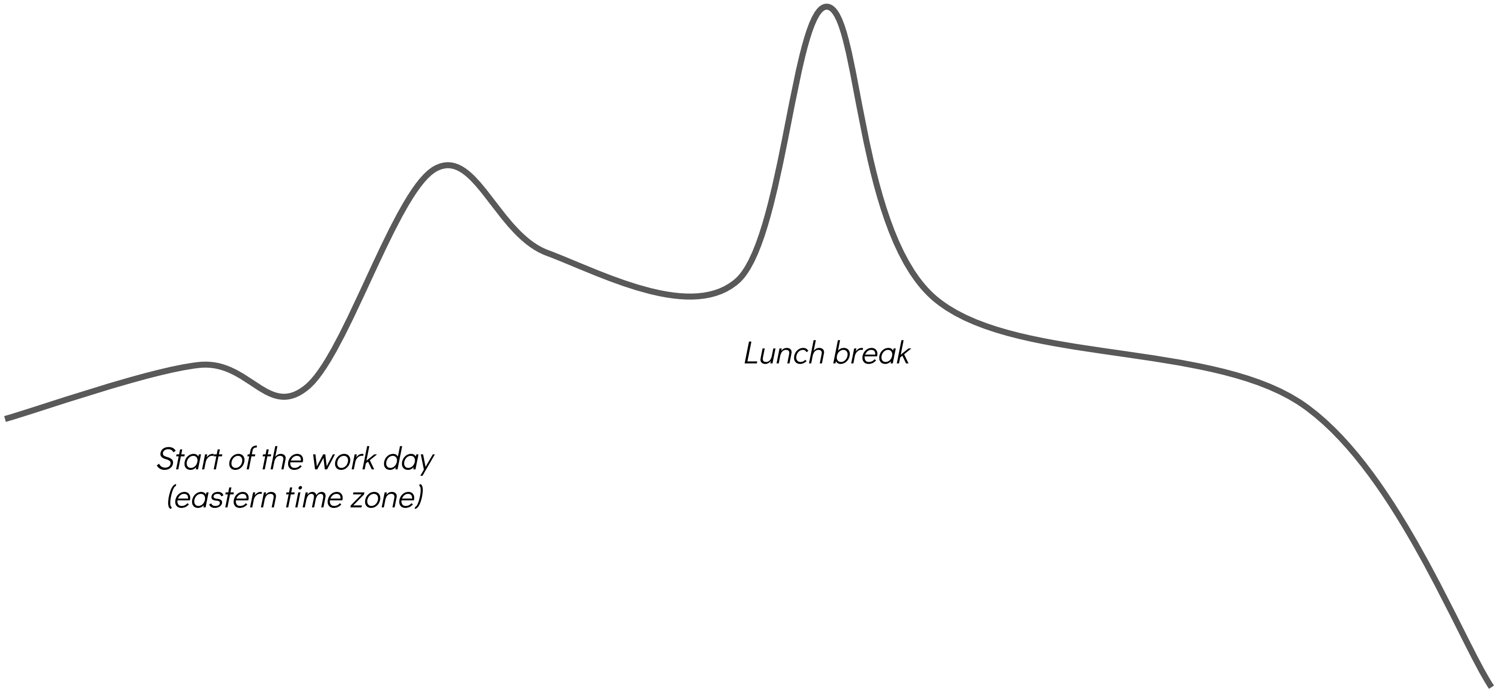



Corrective Feedback

JITTER & APERIODICITY


New England Load - Super Bowls 49 and 50
Start of Game
New England vs. Seattle, 2/1/2015
Denver vs. Carolina, 2/7/2016
Half Time
End of Game
0:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00

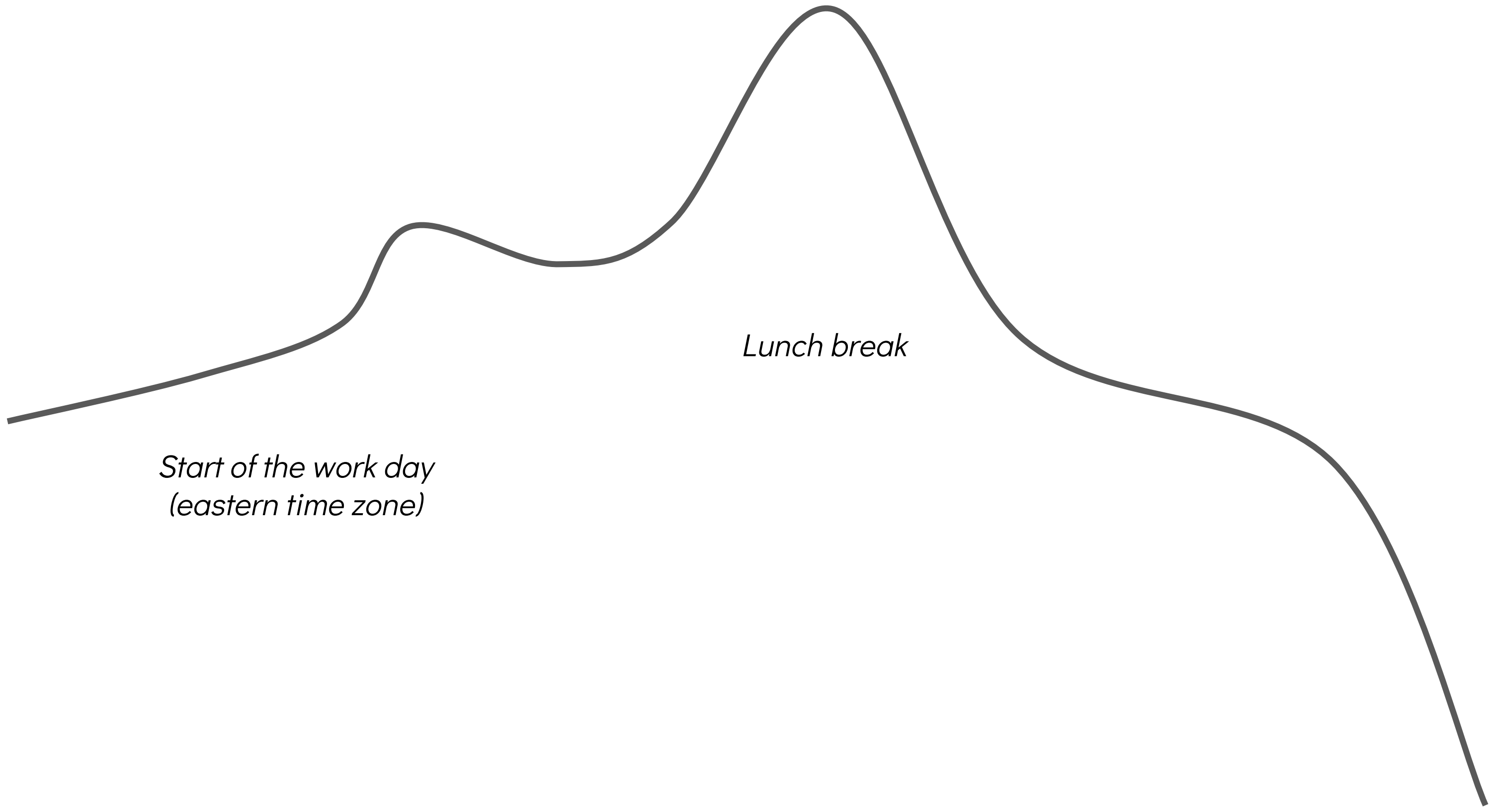
Typical US Call Center Diurnal (Weekday)

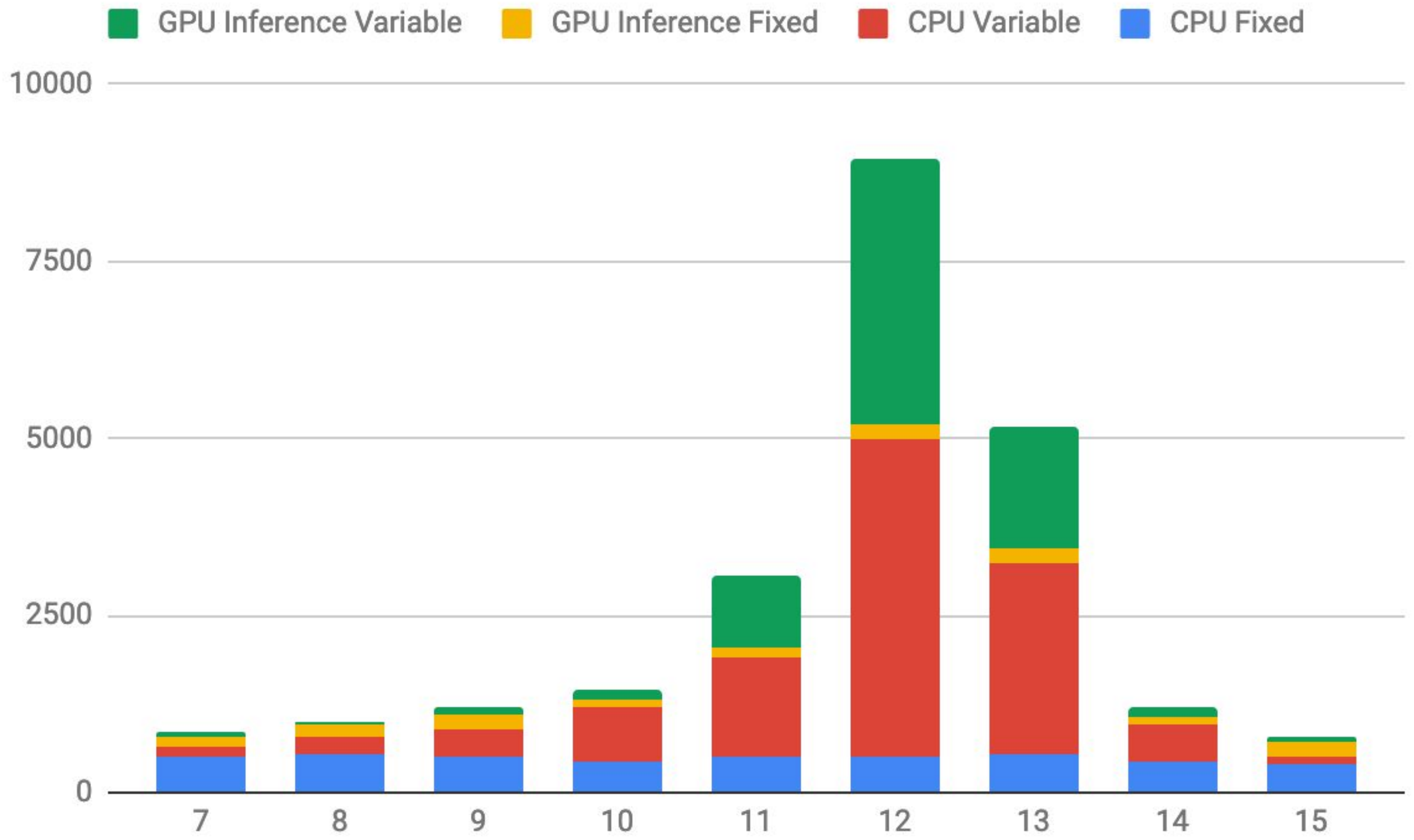


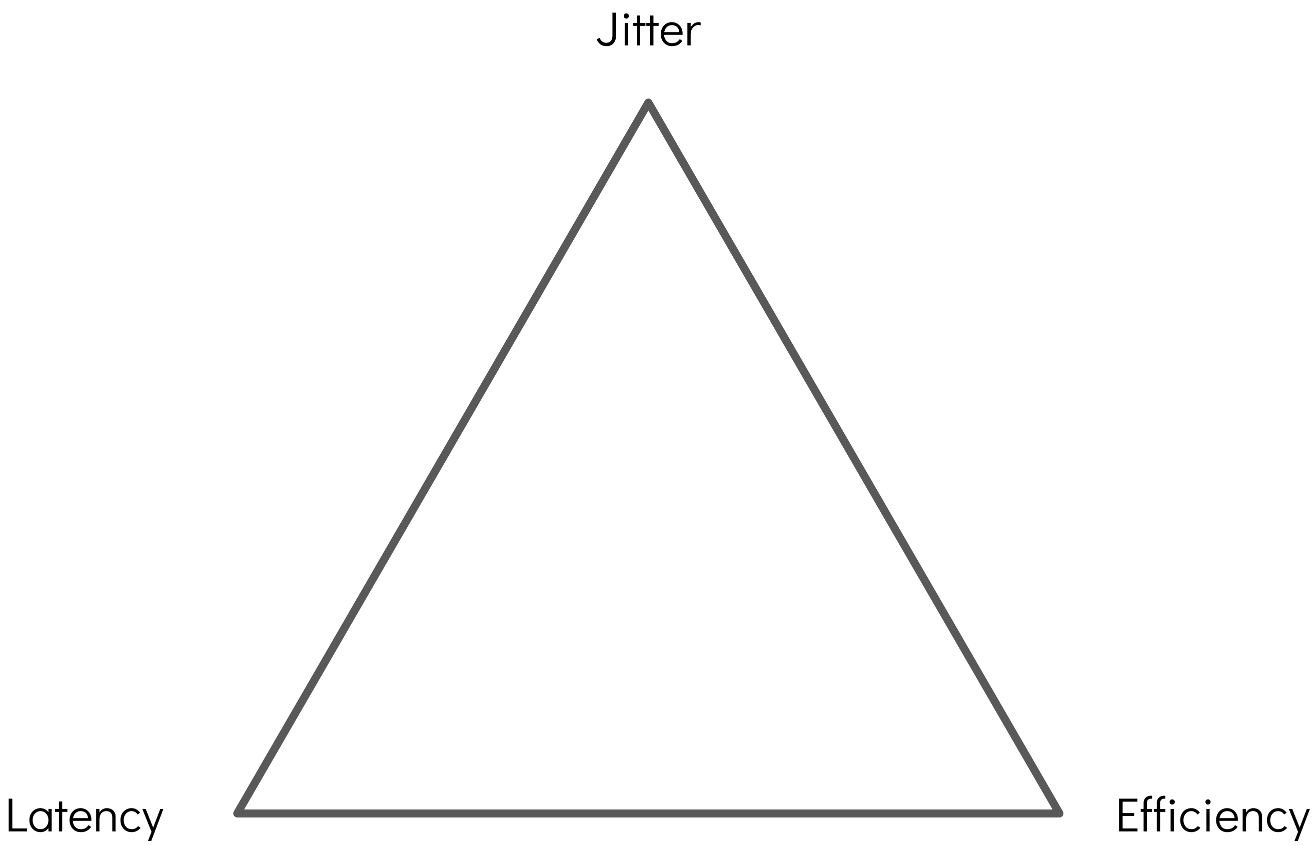
Start of the work day (eastern time zone)

Lunch break

Typical US Call Center Diurnal (First of month)




🟩 GPU Inference Variable
🟨 GPU Inference Fixed
🟥 CPU Variable
🟦 CPU Fixed
10000
7500
5000
2500
0
7
8
9
10
11
12
13
14
15

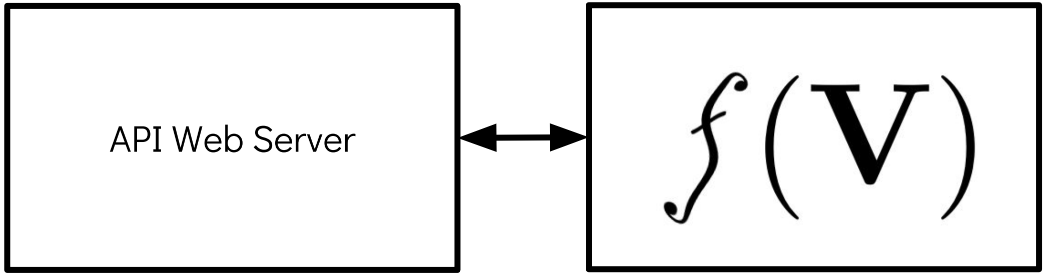


Jitter
Latency
Efficiency

SCALE

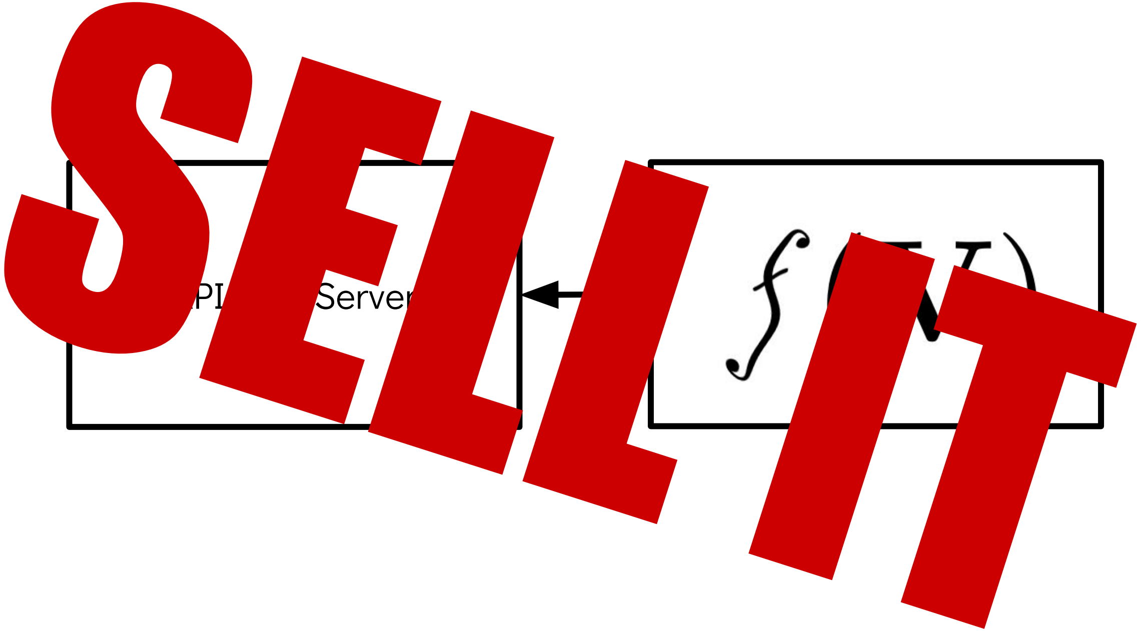
$$f(\mathbf{V})$$

$$f(\mathbf{V})$$



API Web Server

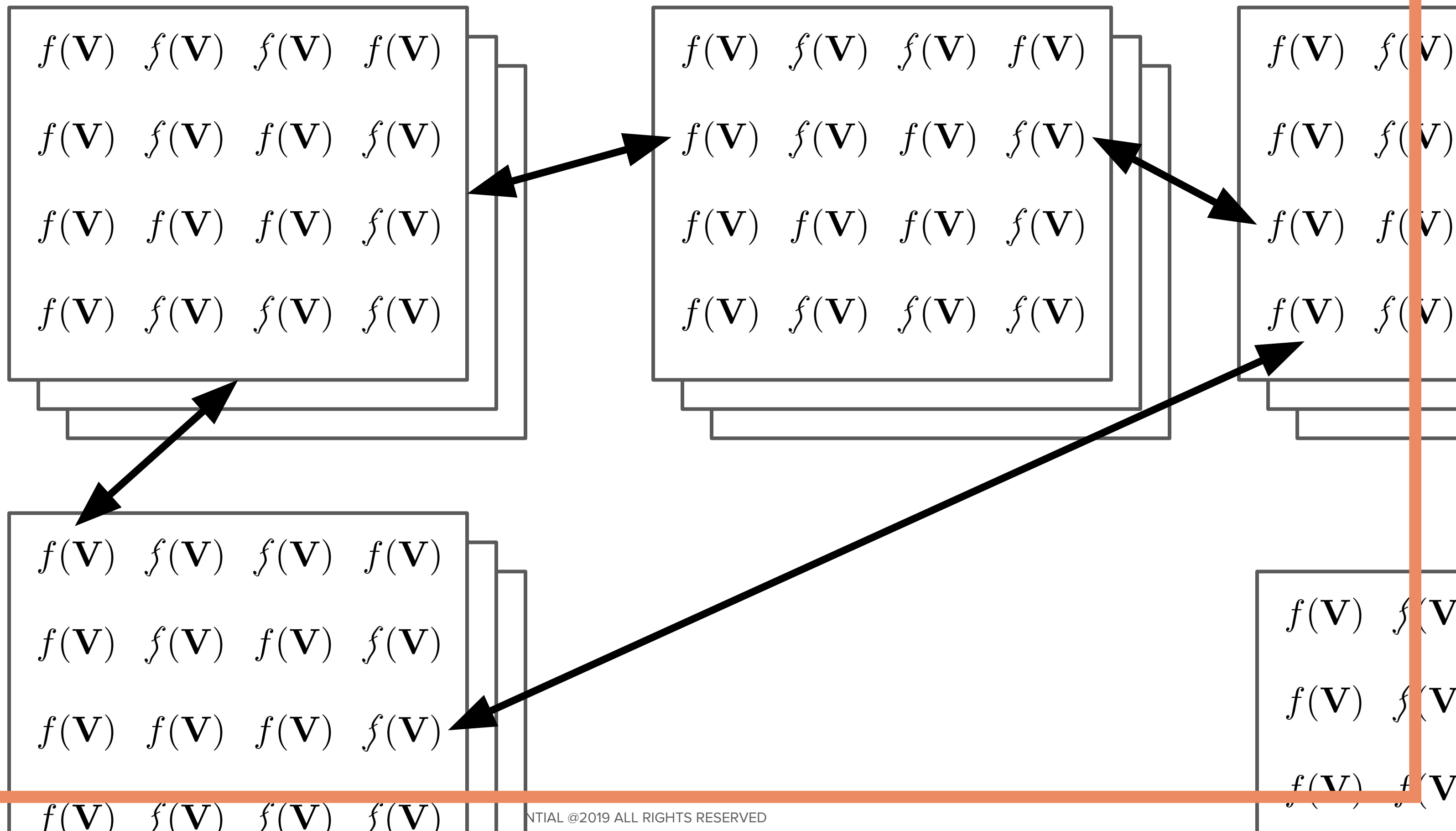
$f(\mathbf{V})$

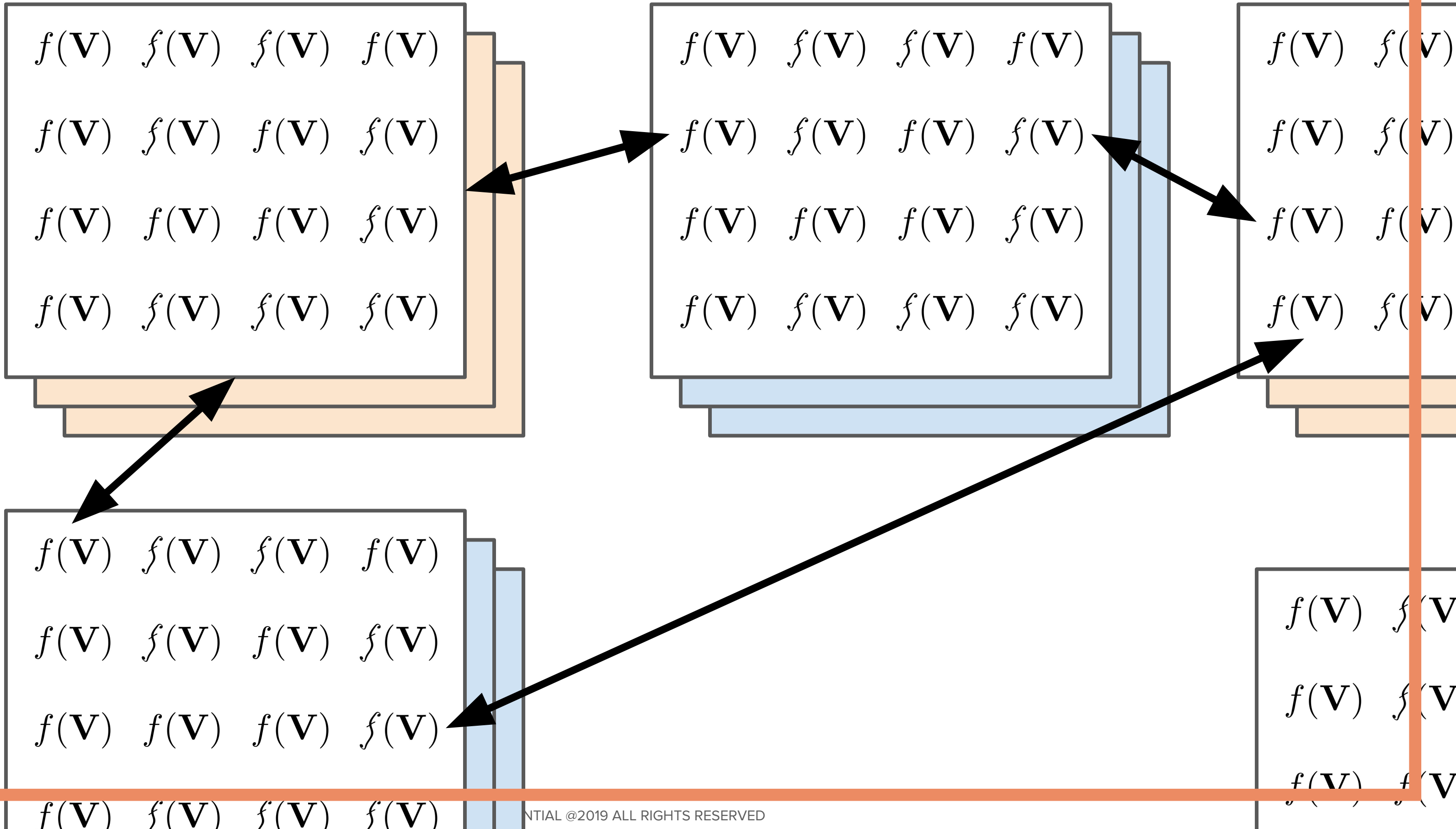


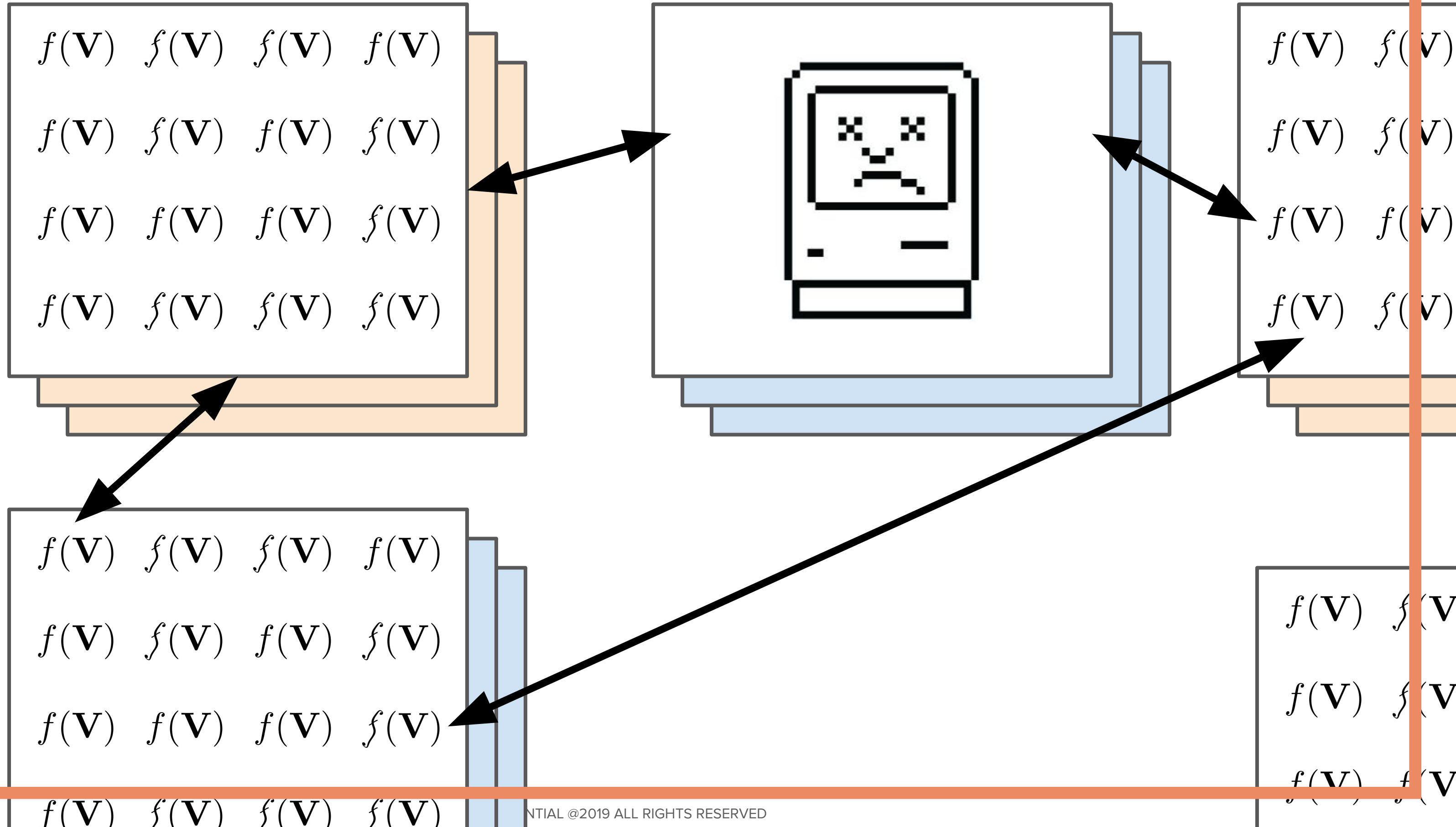
API Server

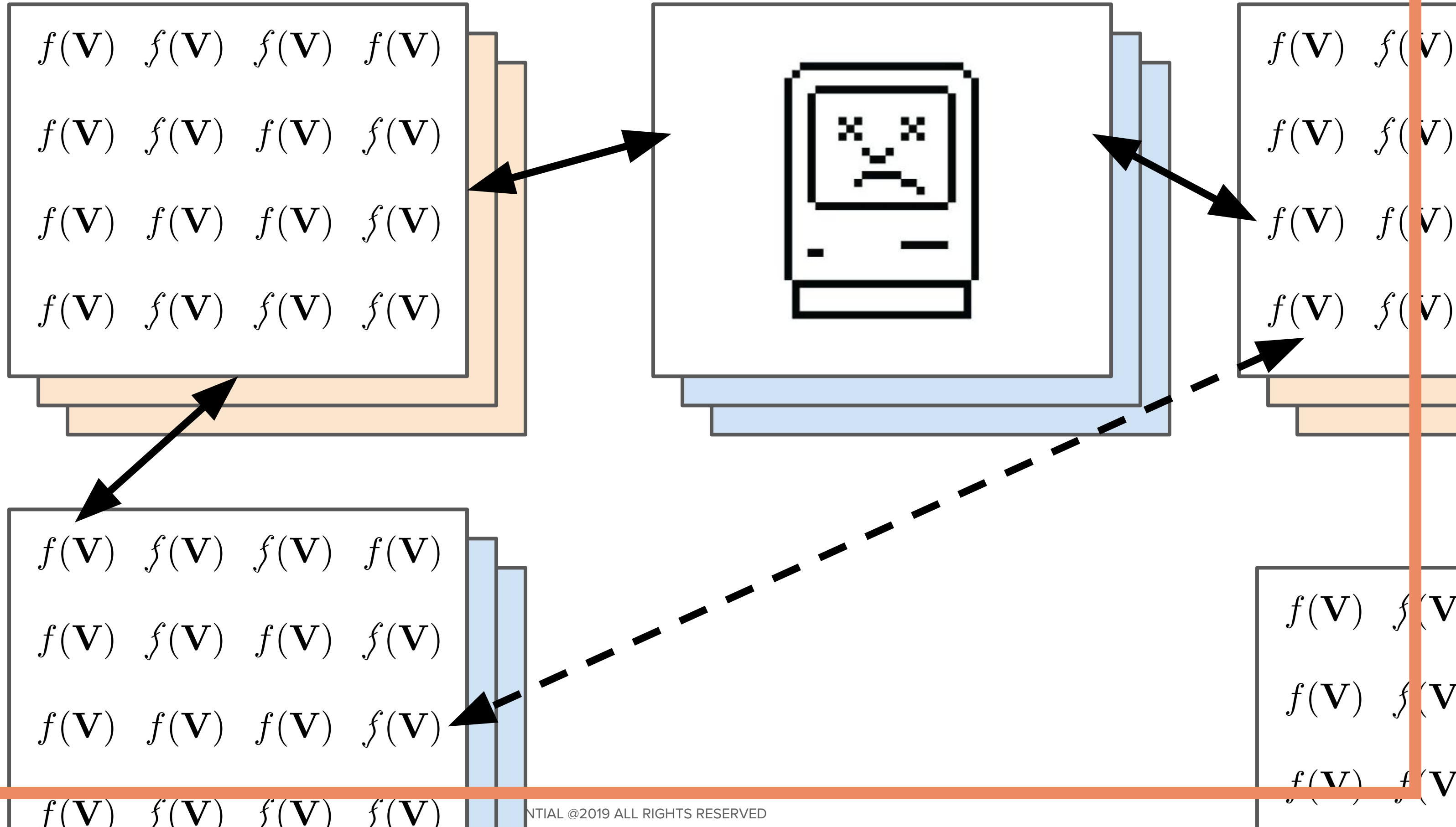
$f(N)$

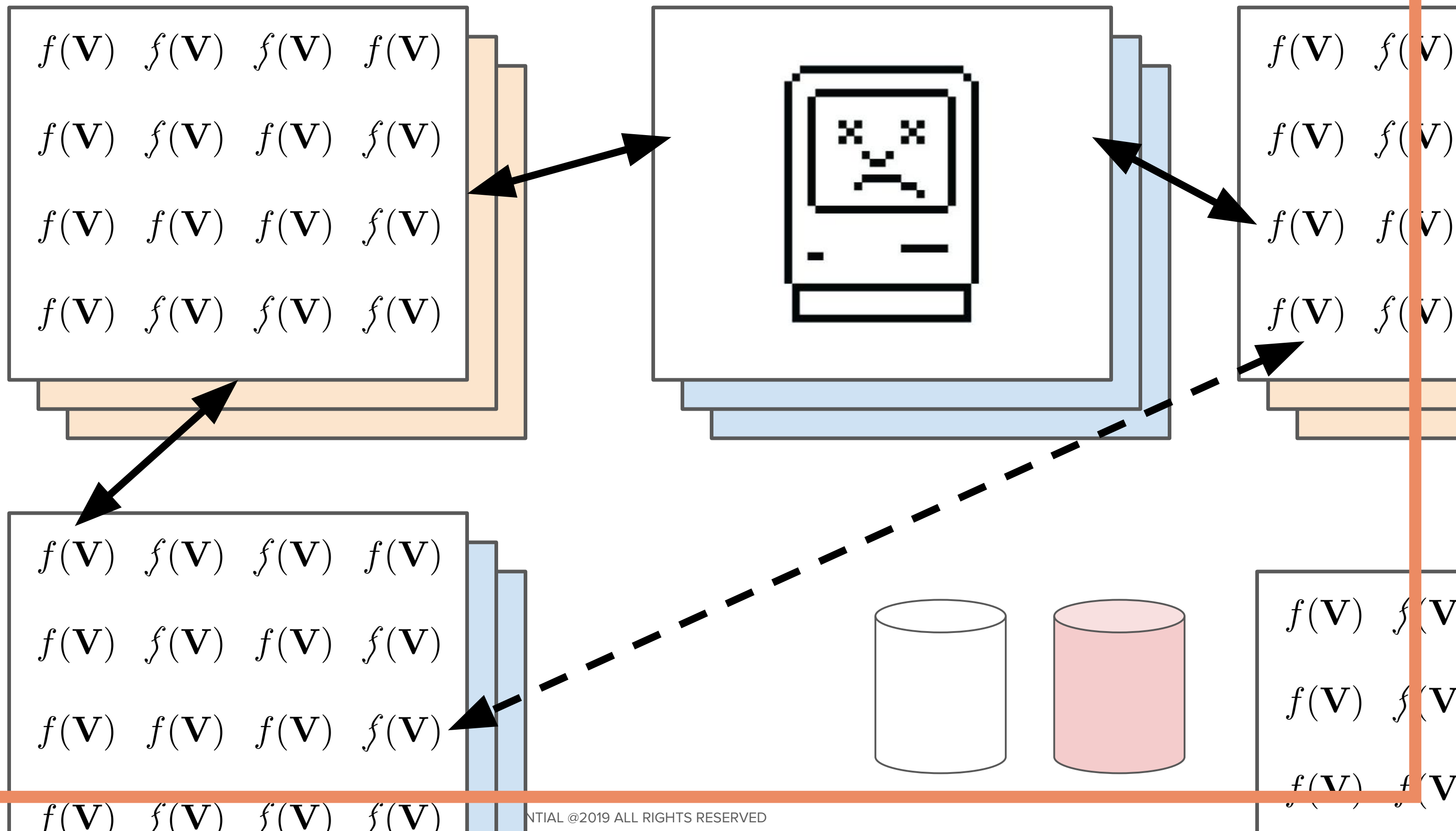
SELL IT


$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$


$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$


$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$

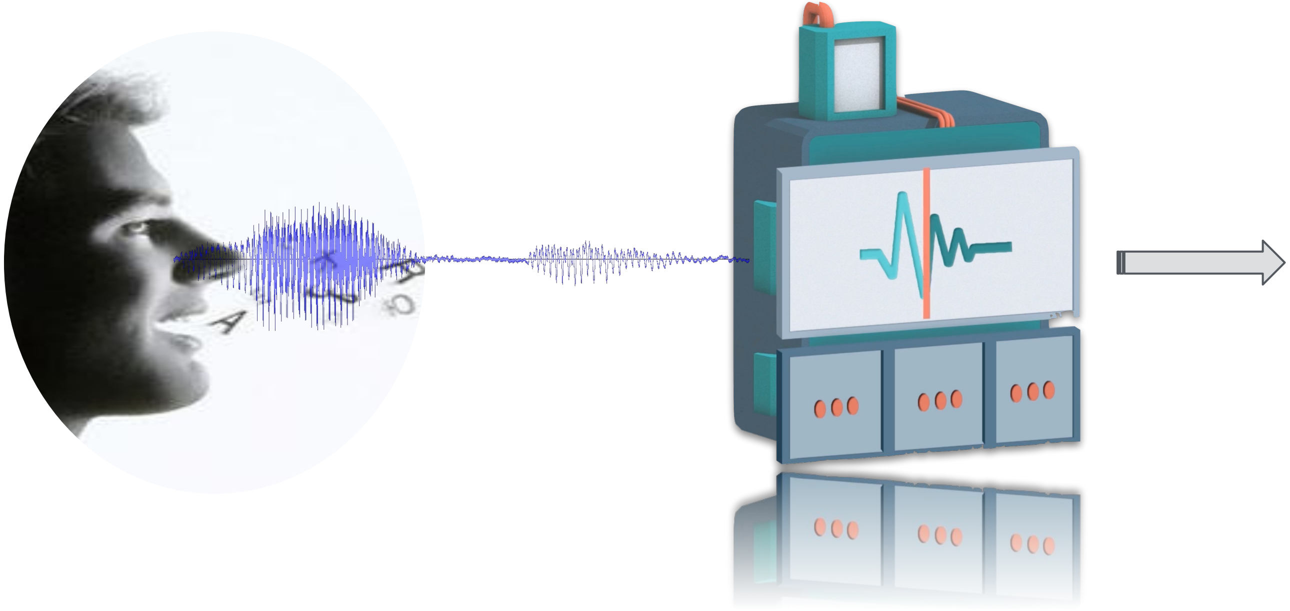

$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V}$
$f(\mathbf{V})$ $f(\mathbf{V}$
$f(\mathbf{V})$ $f(\mathbf{V}$


$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$
$f(\mathbf{V})$ $f(\mathbf{V})$

CASE STUDY:
CLOSED LOOP FEEDBACK IN SPEECH RECOGNITION TRAINING

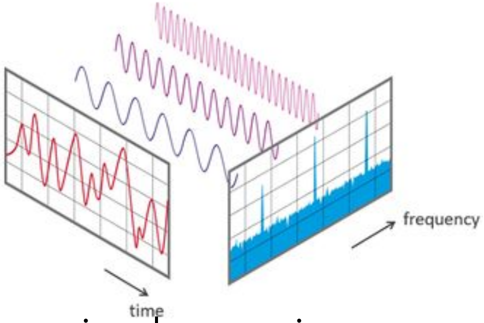
	Time	**Predictability**	**Complexity**
Optimize	Latency ⌛	Stochasticity 🎲	Scale ⚖️
Manage	Jitter & Aperiodicity 🌊	Nonlinearity 🦋	Interactivity 🤖

Speech Recognition



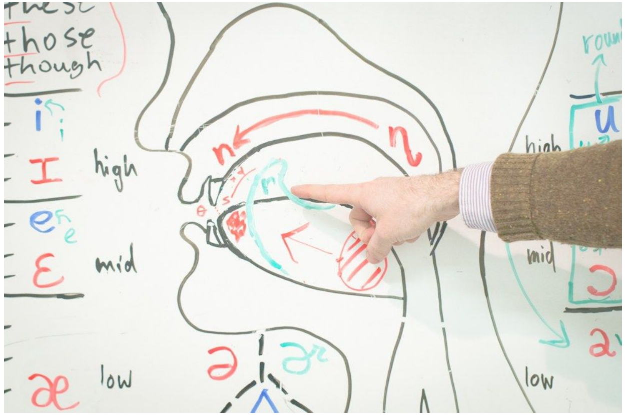
“Welcome to Gridspace”

Speech Recognition

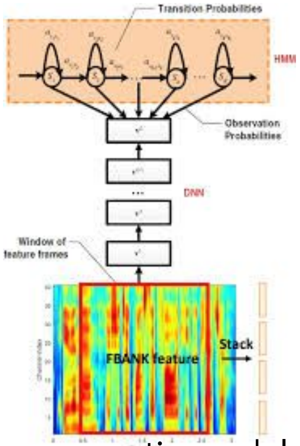


< signal processing >

< Language>

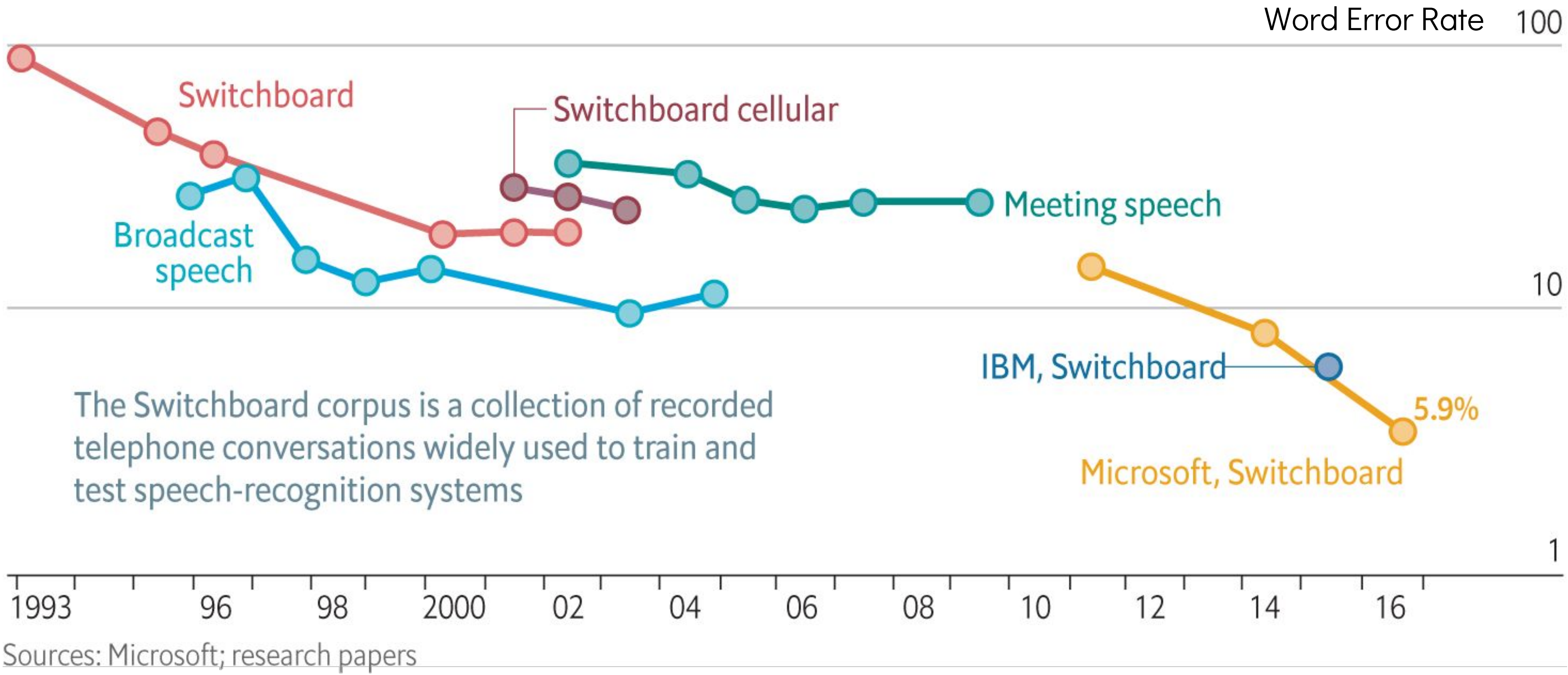


< pronunciation model>



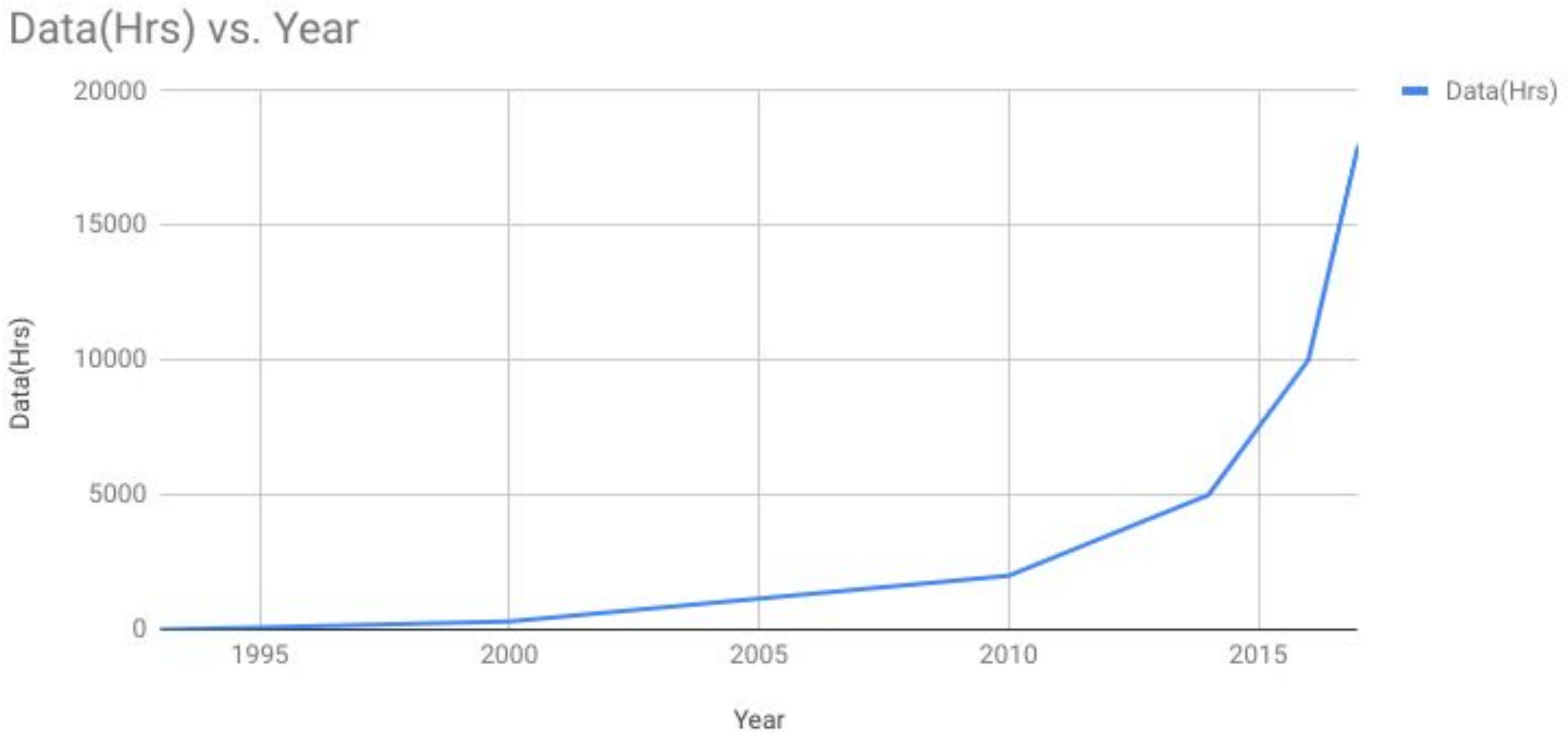
< acoustic model>

Speech Recognition - History



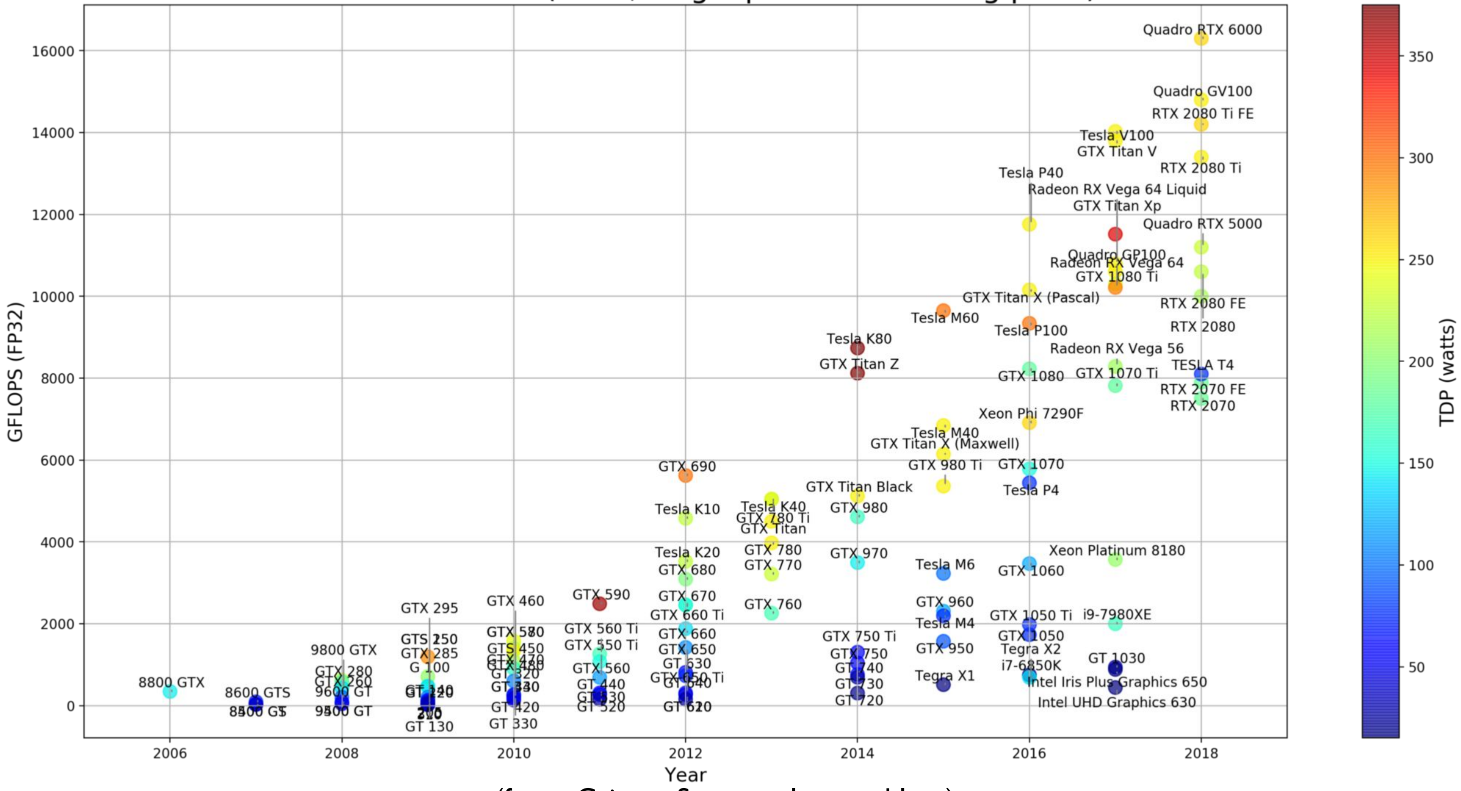
Sources: Microsoft; research papers

Data Boost


Data(Hrs) vs. Year
Data(Hrs)
20000
15000
10000
5000
0
1995
2000
2005
2010
2015
Year
Data(Hrs)

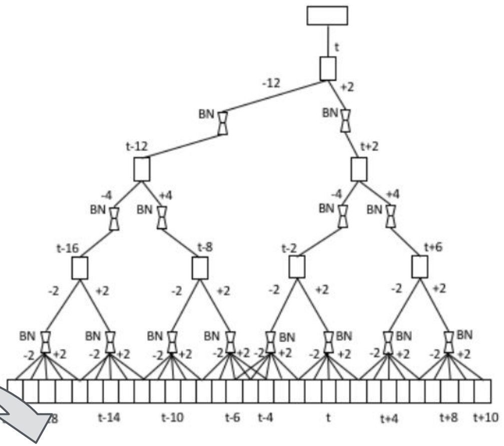
Computation Boost

GPU Performance (FP32, single precision floating point)

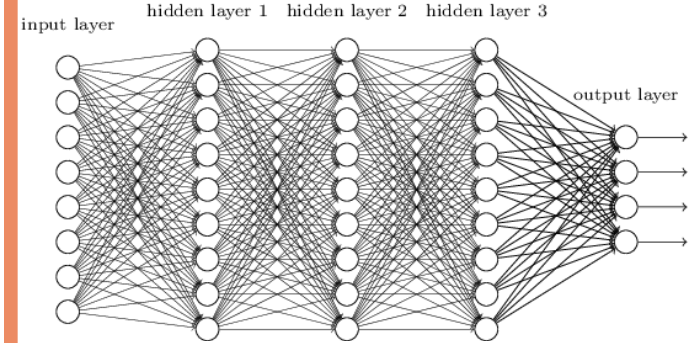


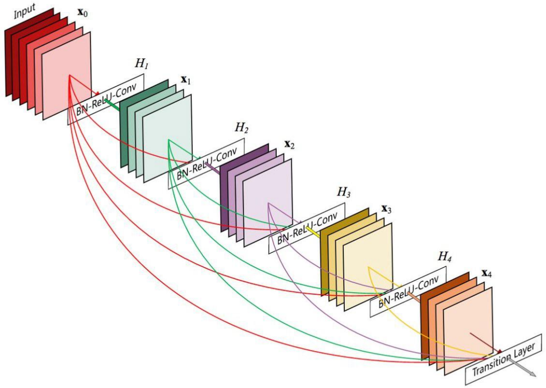
(from: Grigory Sapunov, Intento blogs)

Model Boost



(from: Ming Sun et el., IS 2017 paper)





(from: Kyu Han et el., IS 2018 paper)

Speech recognition in the Wild - Still Challenging

"When my oldest child was a baby..."

"It didn't send me any kind of verification code. It did pop up a message saying that the account was locked"

Switchboard	DATA	Gridspace Call Center Data
normal vocab, normal emotion	Speech Type	domain specific vocab, emotional modulated speech
7.5%	Word Error Rate	18% (without optimization)

Speech recognition in the Wild - Still Challenging

Each domain has each specific language.

- command query
- financial
- media
- customer support

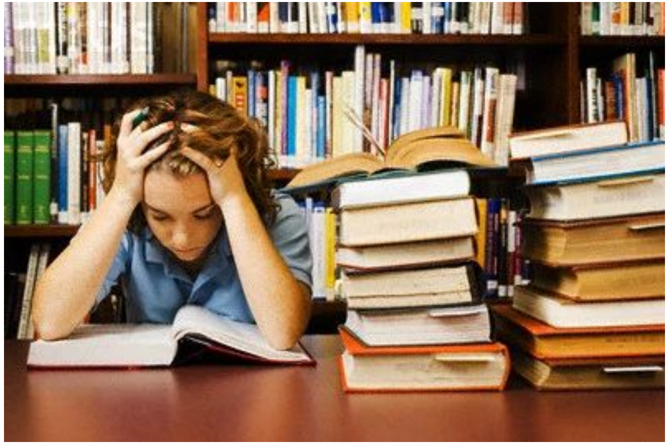
Speech recognition in the Wild - Still Challenging

youse?
yuns?
you?
y'all?
yiz?

There is lots of variabilities in speech

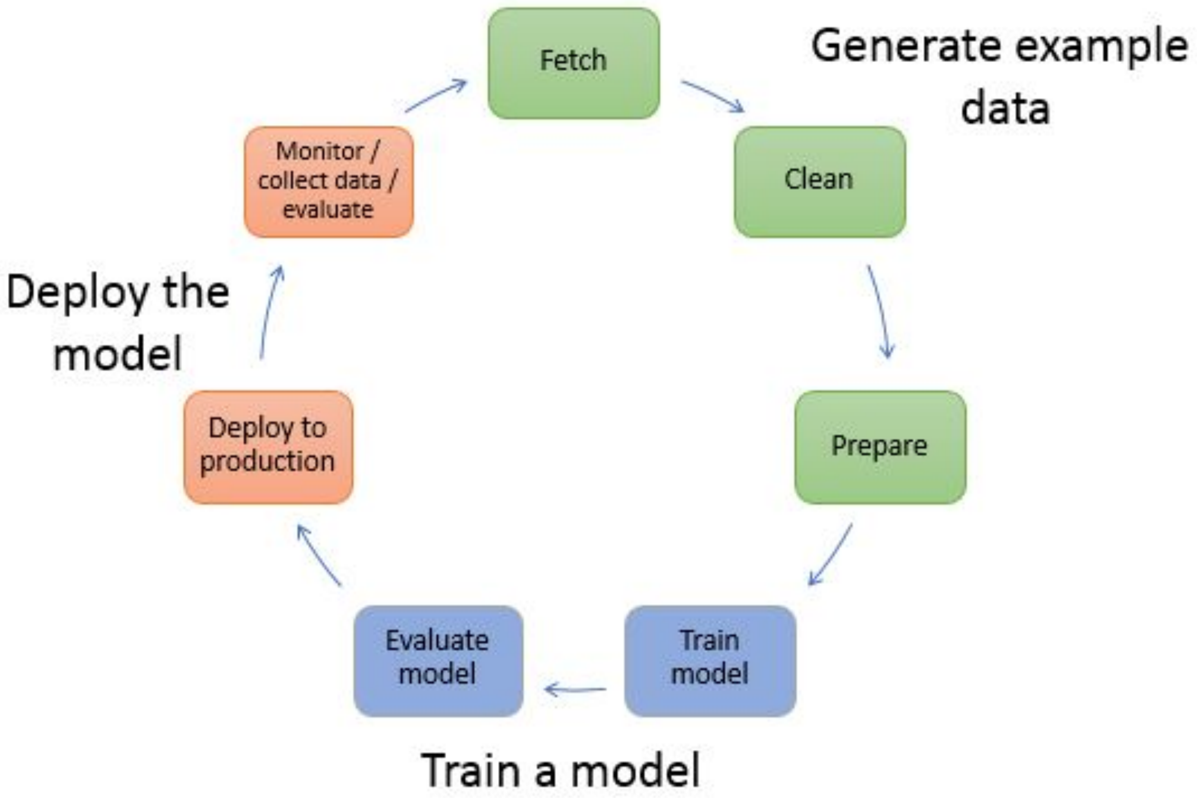
- accent
- noise
- emotion modulated speech
- mis-pronunciation

Speech recognition in the Wild - Still Challenging

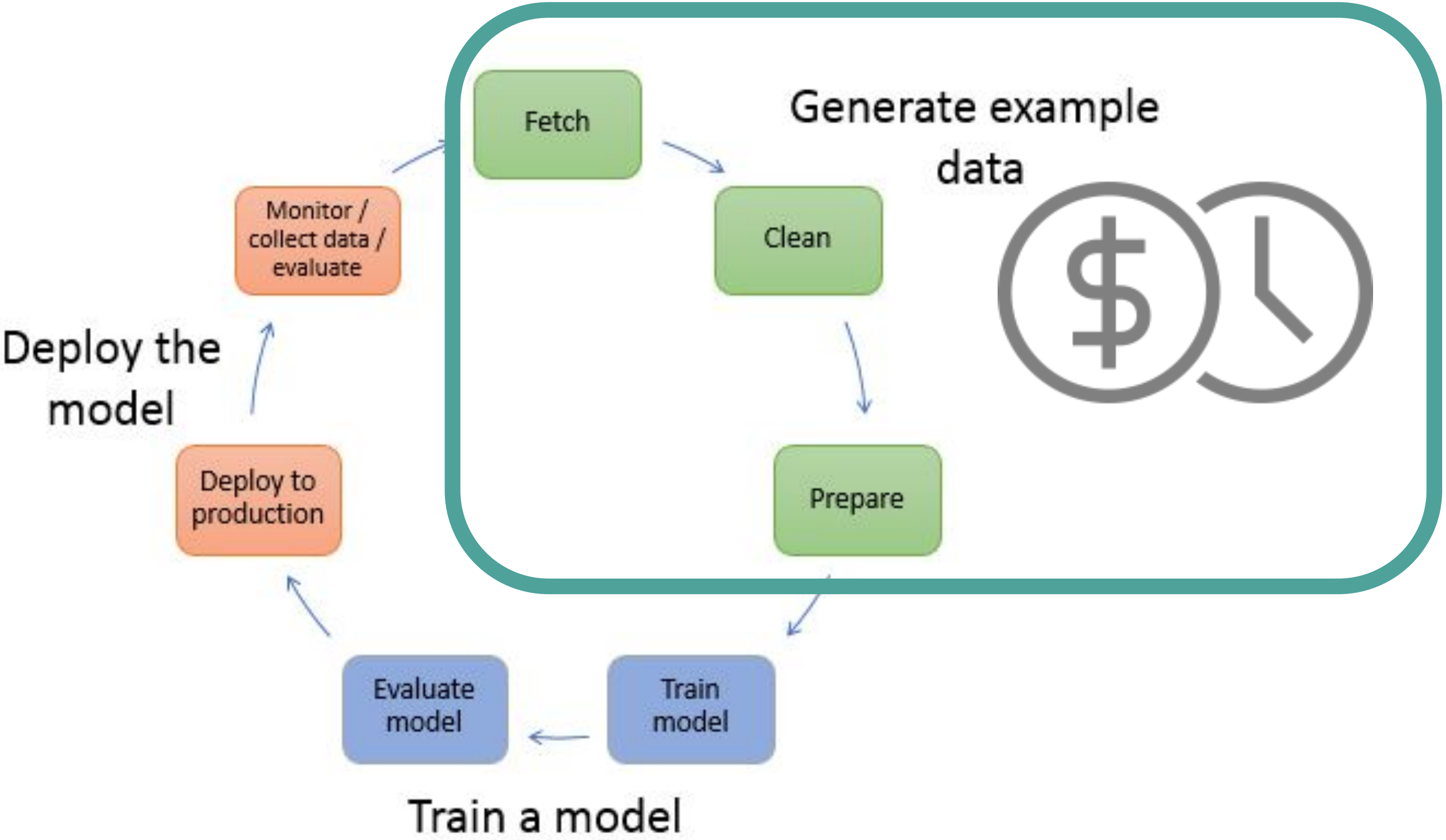


It's hard to learn all beforehand!

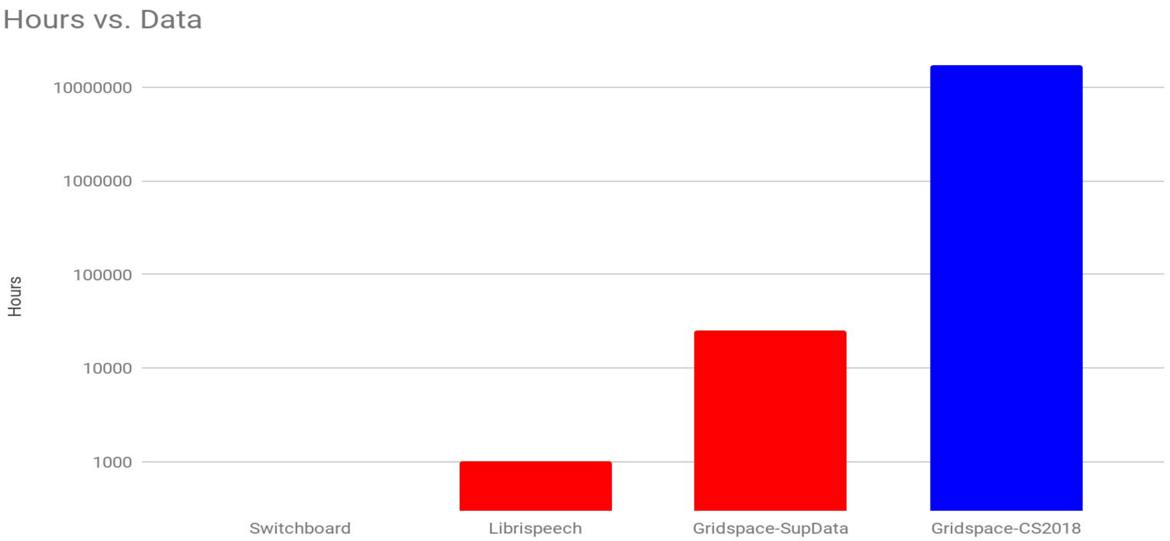
Speech recognition in the Wild - in Practice



Speech recognition in the Wild - in Practice



Use of Unsupervised Data



	SWBD	Librispeech	GS-SupData	GS-CS2018
Hours	0.3k	1k	30k	17M (2000 years)
Data	Supervised	Supervised	Supervised	Unsupervised

Use of Unsupervised Data

Semi-supervised training(SST)

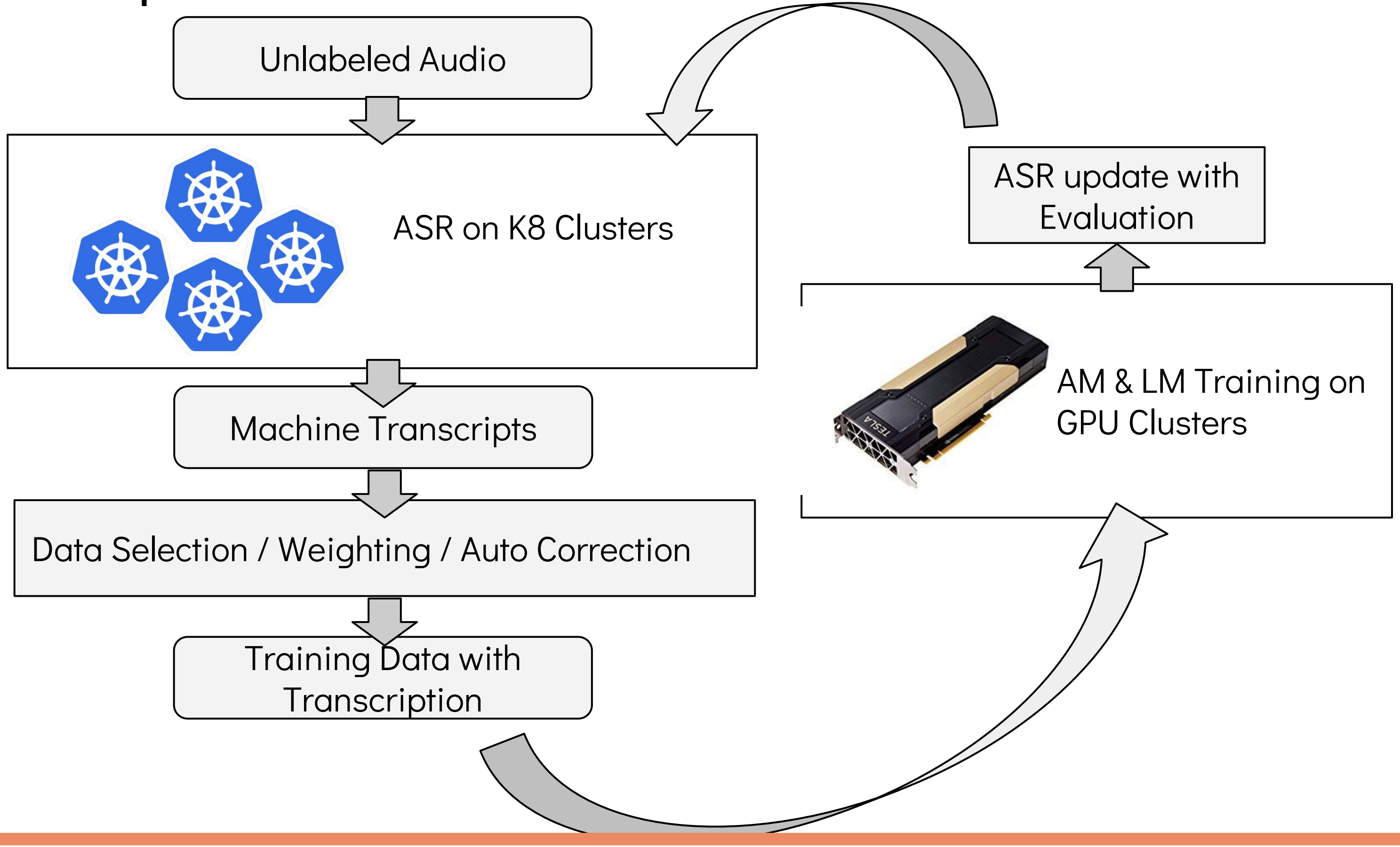
- a good way to use unsupervised data for supervised tasks
- It has to deal with uncertainties
- We can update AM and LM iteratively

Use of Unsupervised Data

Data selection for SST

- 100% use(no filtering) of Unsupervised data would cause model's degradation on accuracy
- knowledge based selection helps
  - confidence score, length, topic, speaker info

Use of Unsupervised Data



Unlabeled Audio

ASR on K8 Clusters

Machine Transcripts

Data Selection / Weighting / Auto Correction

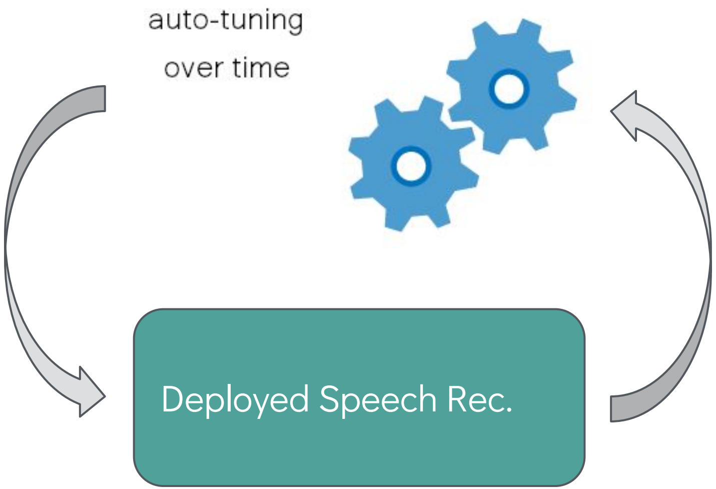
Training Data with Transcription

AM & LM Training on GPU Clusters

ASR update with Evaluation

Continuous Learner

- accent learning
- noise learning
- language/grammar learning



auto-tuning over time

Deployed Speech Rec.

… and become better learner over time

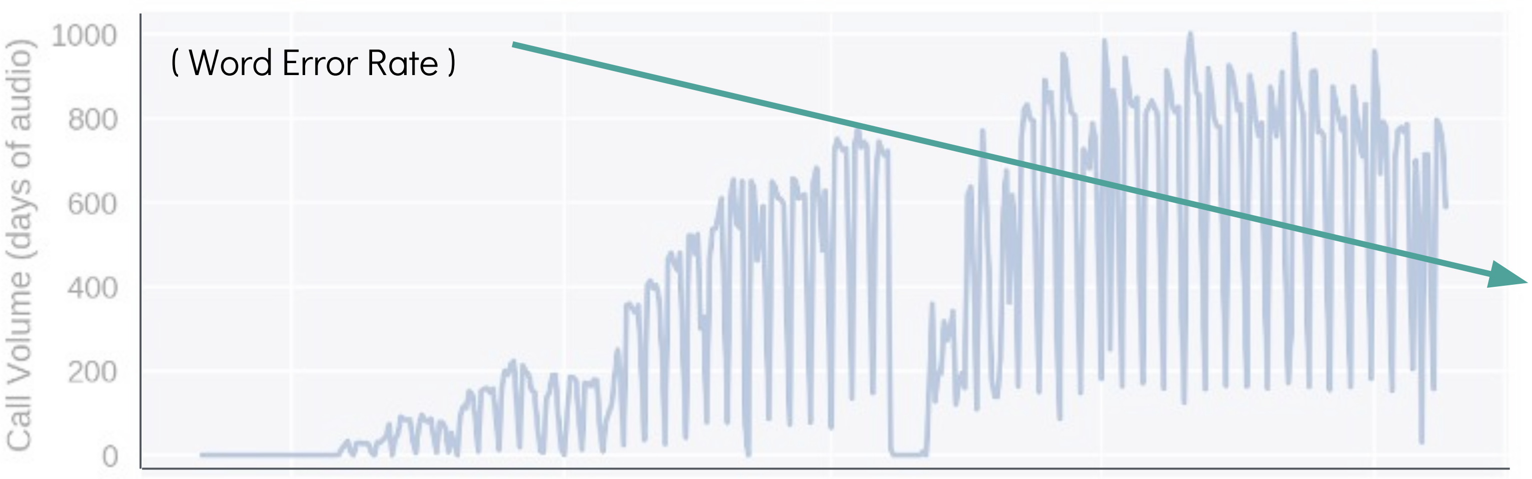
Continuous Learner

Throughput

	P-100	V-100
SST training / 1 sec / 1 gpu	450 sec	580 sec
SST training / 24 hours / 1 gpu	10800 hours	13920 hours

* Training for Acoustic Model(Resent TDNN), 150 frames per example, 64 example per minibatch

Continuous Learner



Time

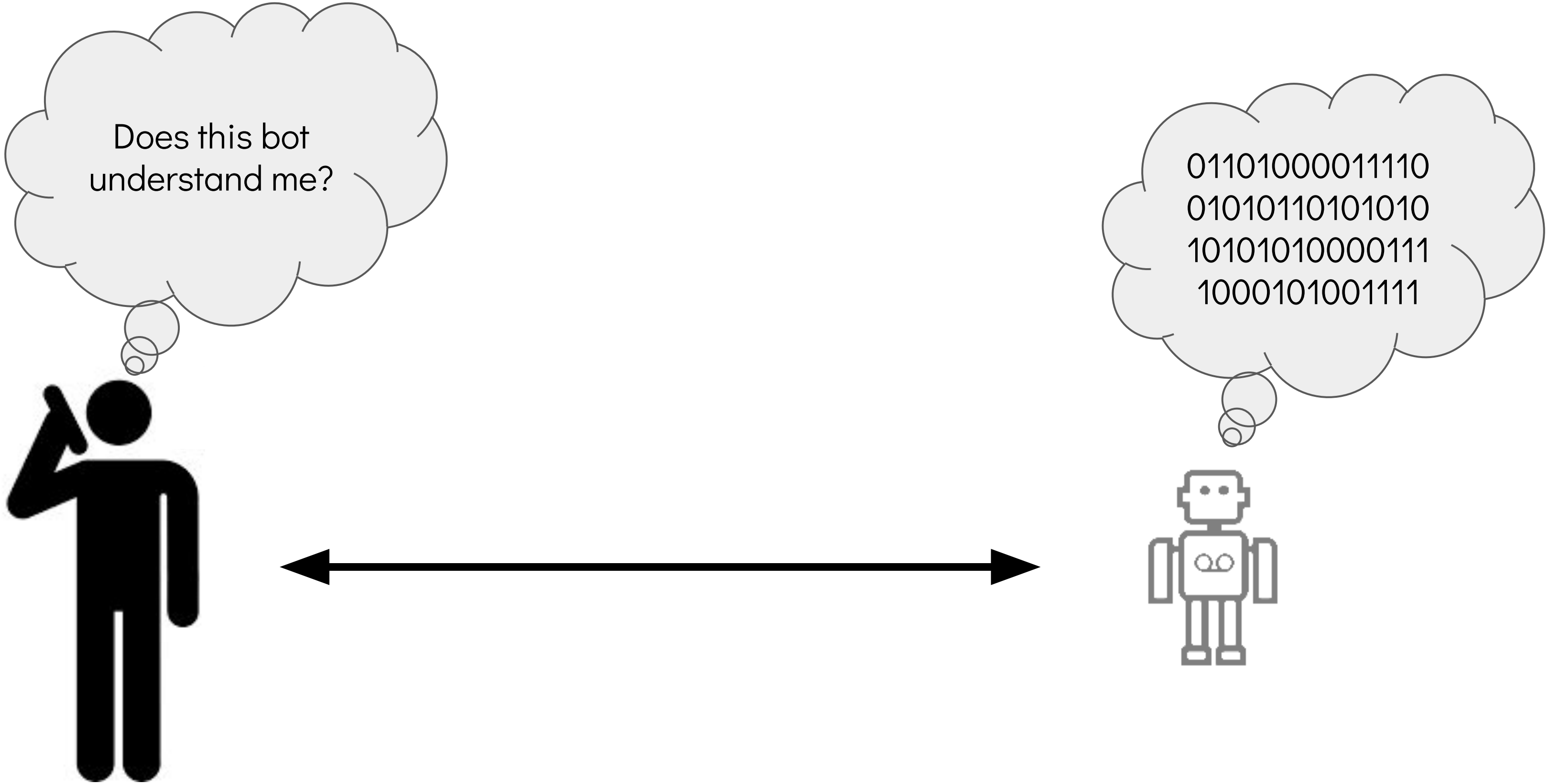
Ideal Error Graph?

Continuous Learner

Experimental Results

	Unmatched Domain	Supervised Training	Semi-Sup Training
Word Error Rate	14.29%	9.52%	7.83%
System Building Hours		3 months	< 1 DAY

CASE STUDY: STOCHASTICITY AND LATENCY IN SPEECH SYNTHESIS INFERENCE


Does this bot understand me?
01101000011110
01010110101010
10101010000111
1000101001111

Dialog System Evaluators

Content

Naturalness

Cadence

Influenced by TTS

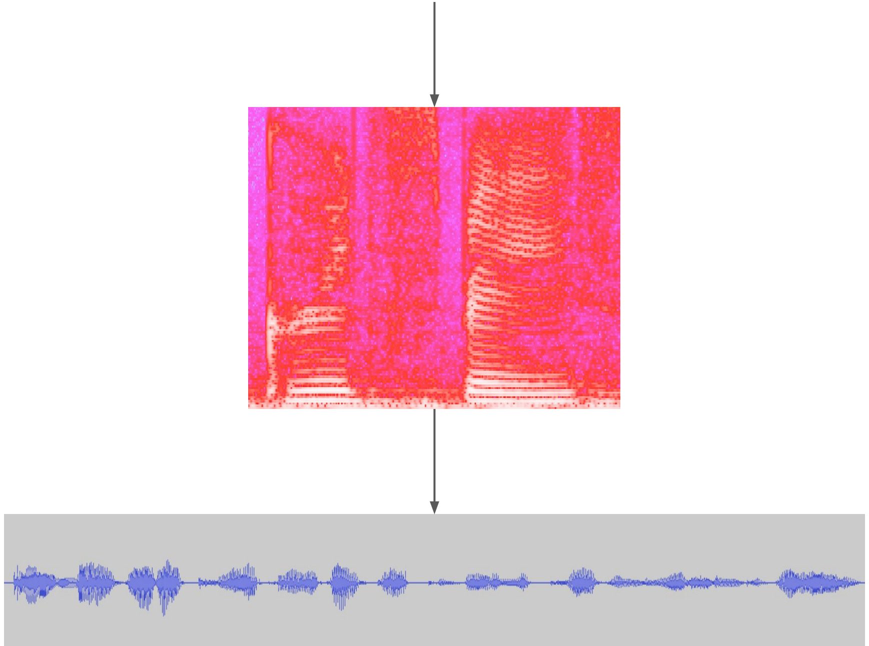
Content

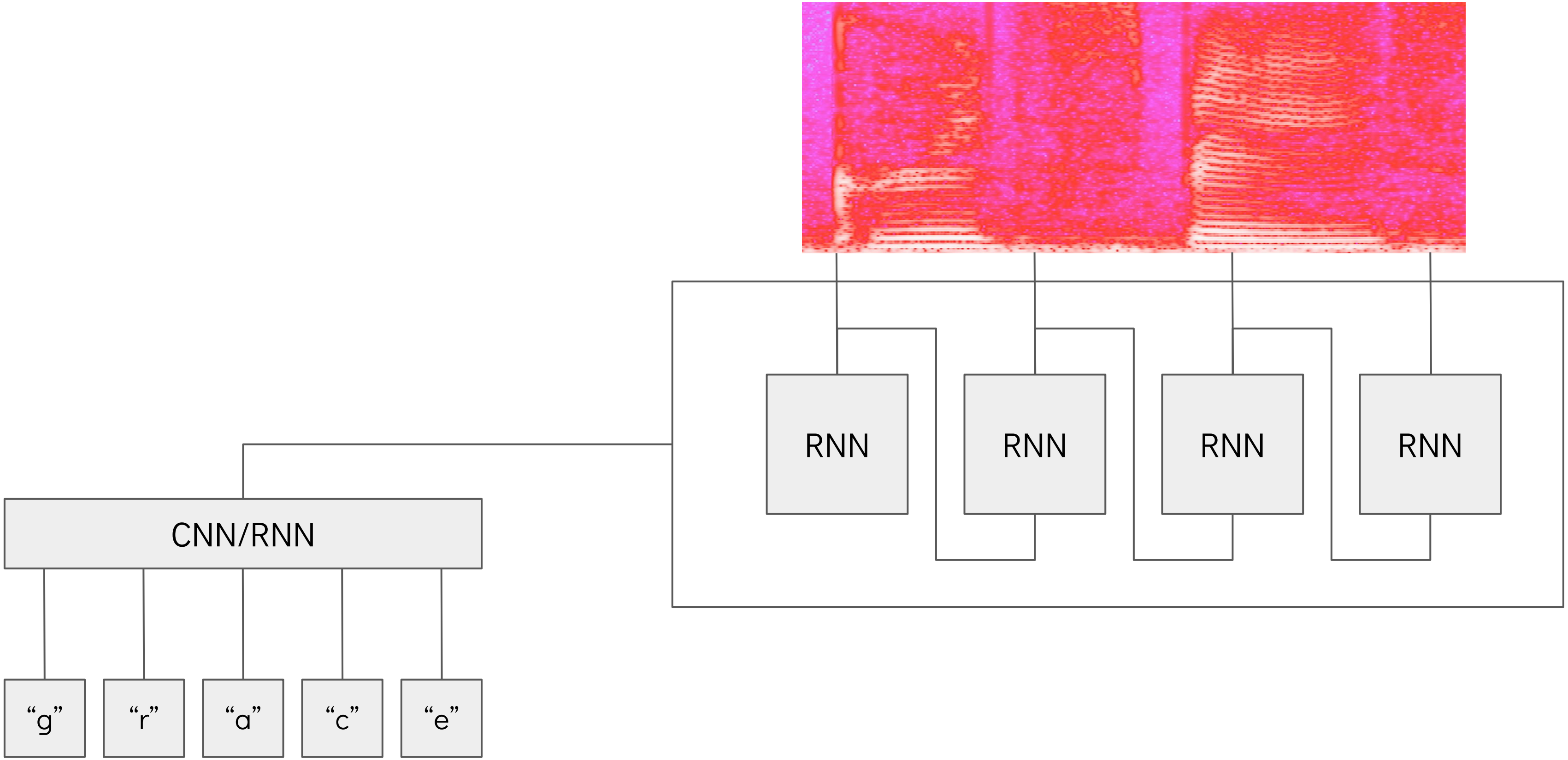
Naturalness

Cadence

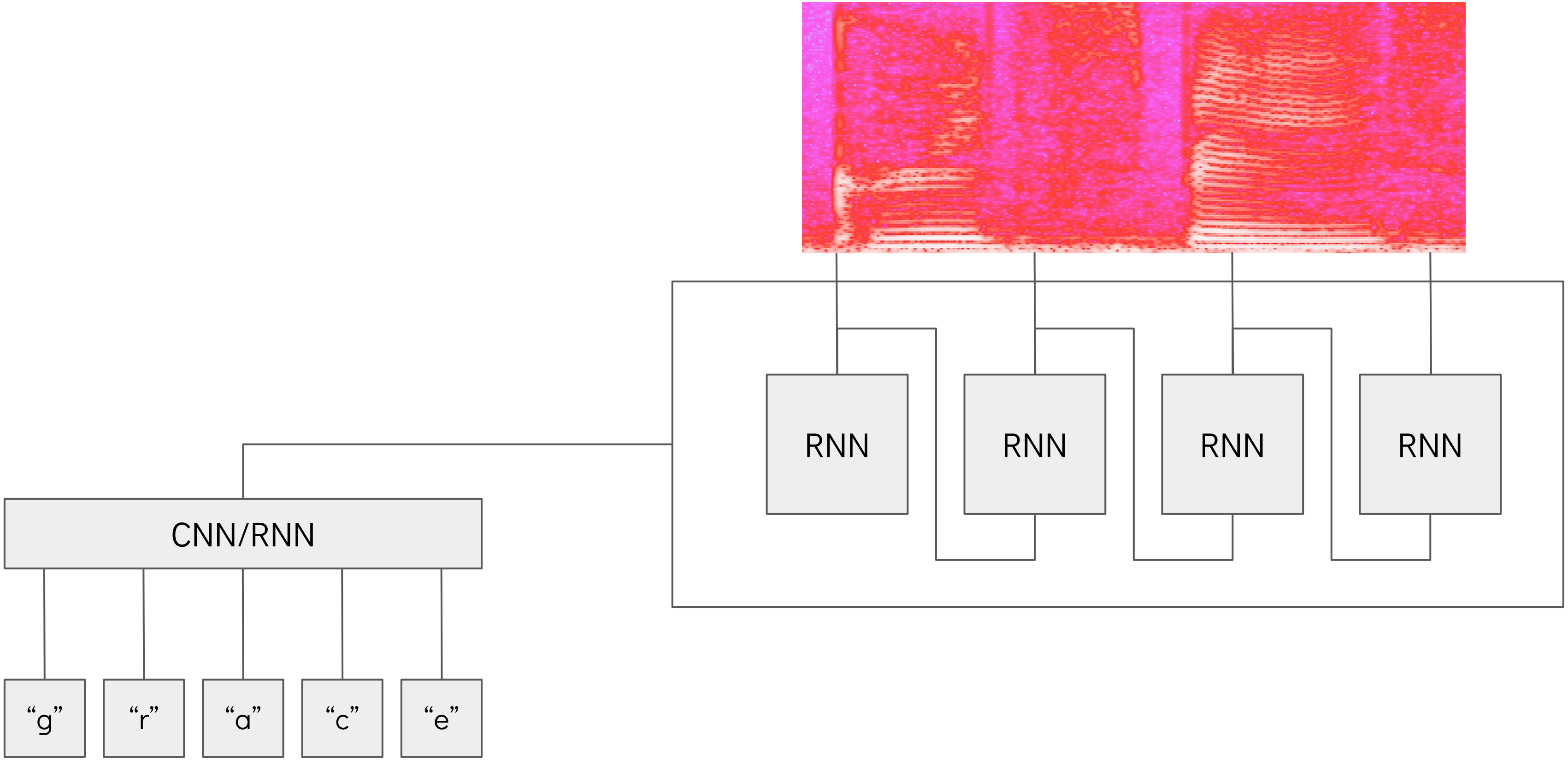
Modern Neural TTS

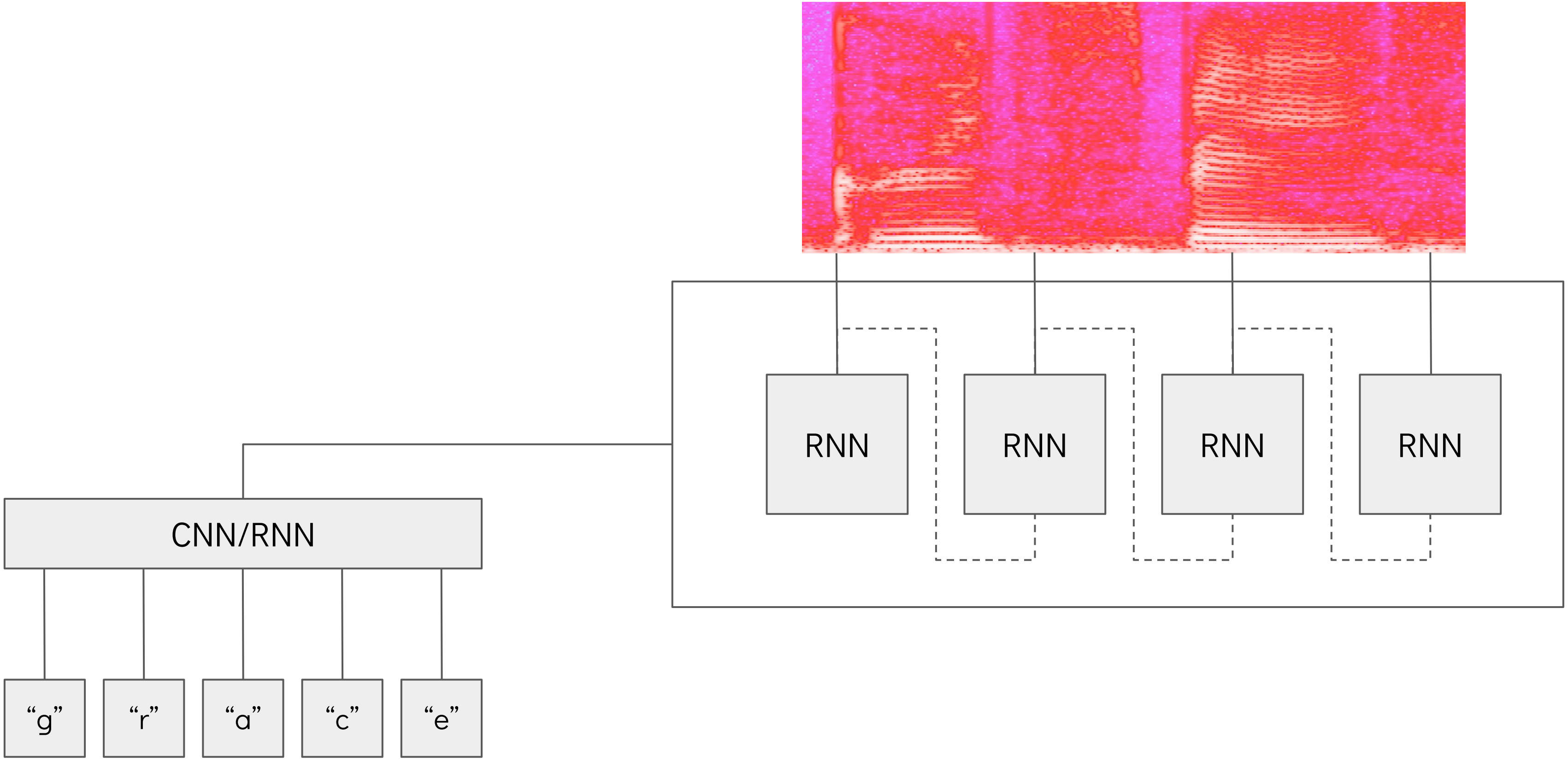
“I’m speech that came from a big neural network”

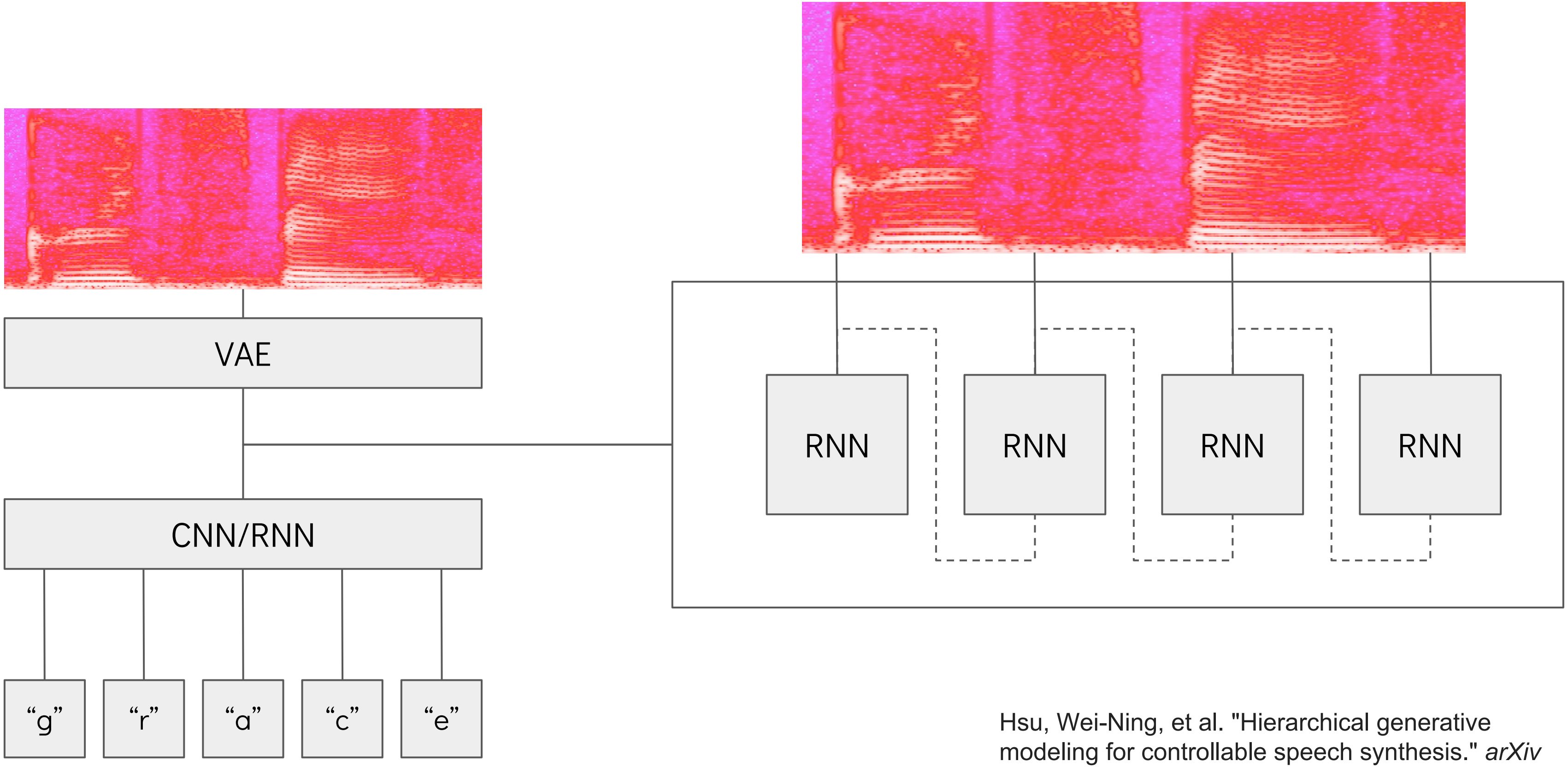



RNN
RNN
RNN
RNN
CNN/RNN
“g”
“r”
“a”
“c”
“e”

	Time	Predictability	Complexity
Optimize	Latency ⌛	Stochasticity 🎲	Scale ⚖
Manage	Jitter & Aperiodicity 🌊	Nonlinearity	Interactivity


RNN
RNN
RNN
RNN
CNN/RNN
“g”
“r”
“a”
“c”
“e”


RNN
RNN
RNN
RNN
CNN/RNN
“g”
“r”
“a”
“c”
“e”



Hsu, Wei-Ning, et al. "Hierarchical generative modeling for controllable speech synthesis." *arXiv preprint arXiv:1810.07217* (2018).

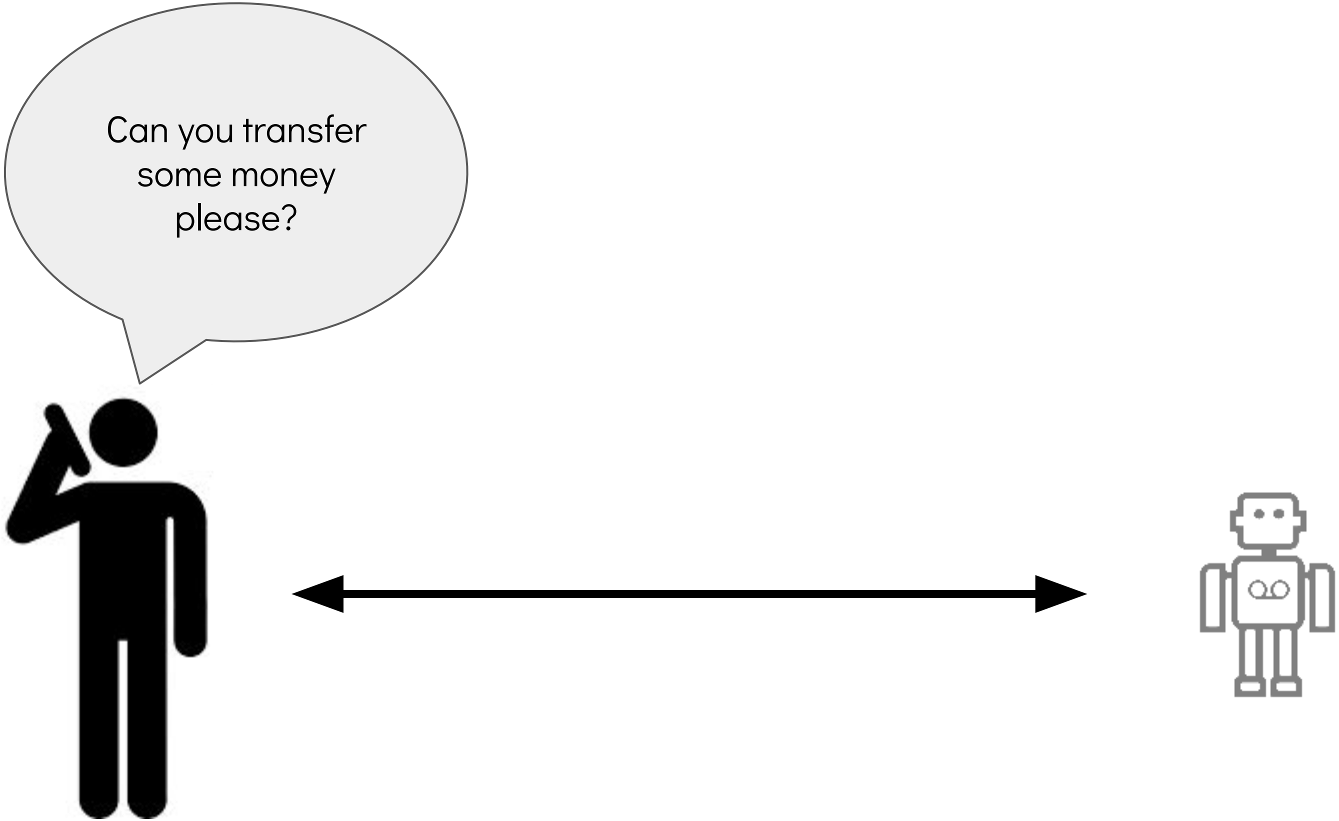
Influenced by TTS

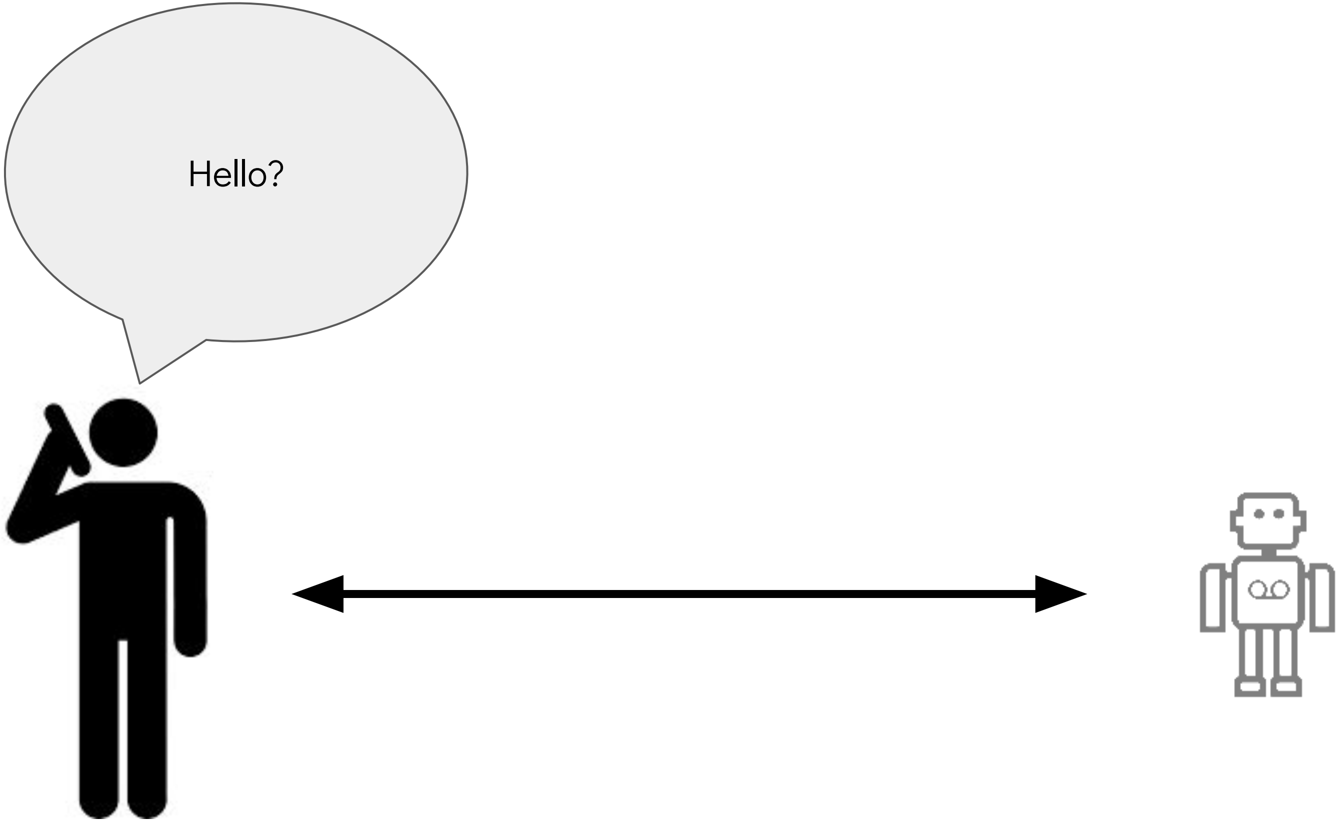
Content

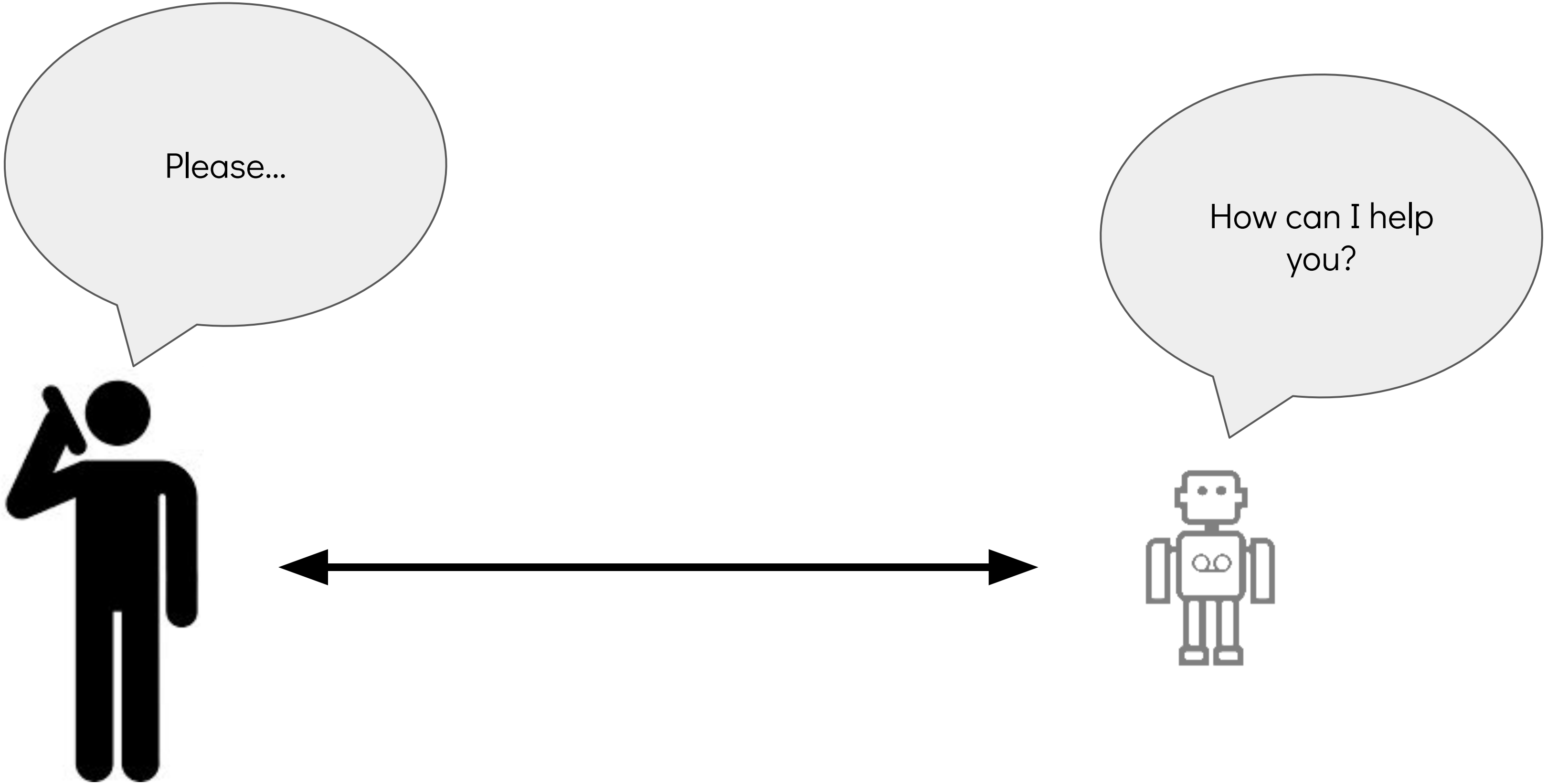
Naturalness

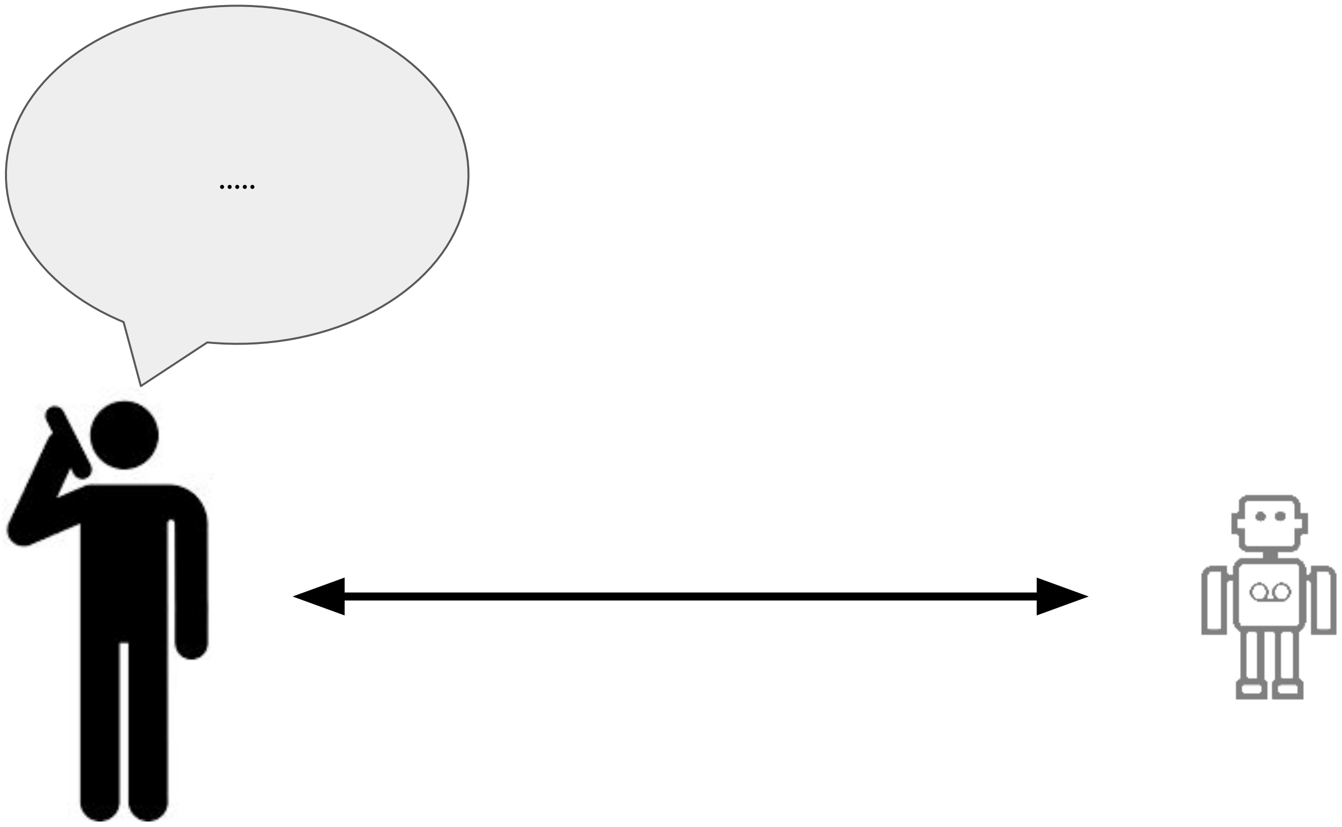
Cadence

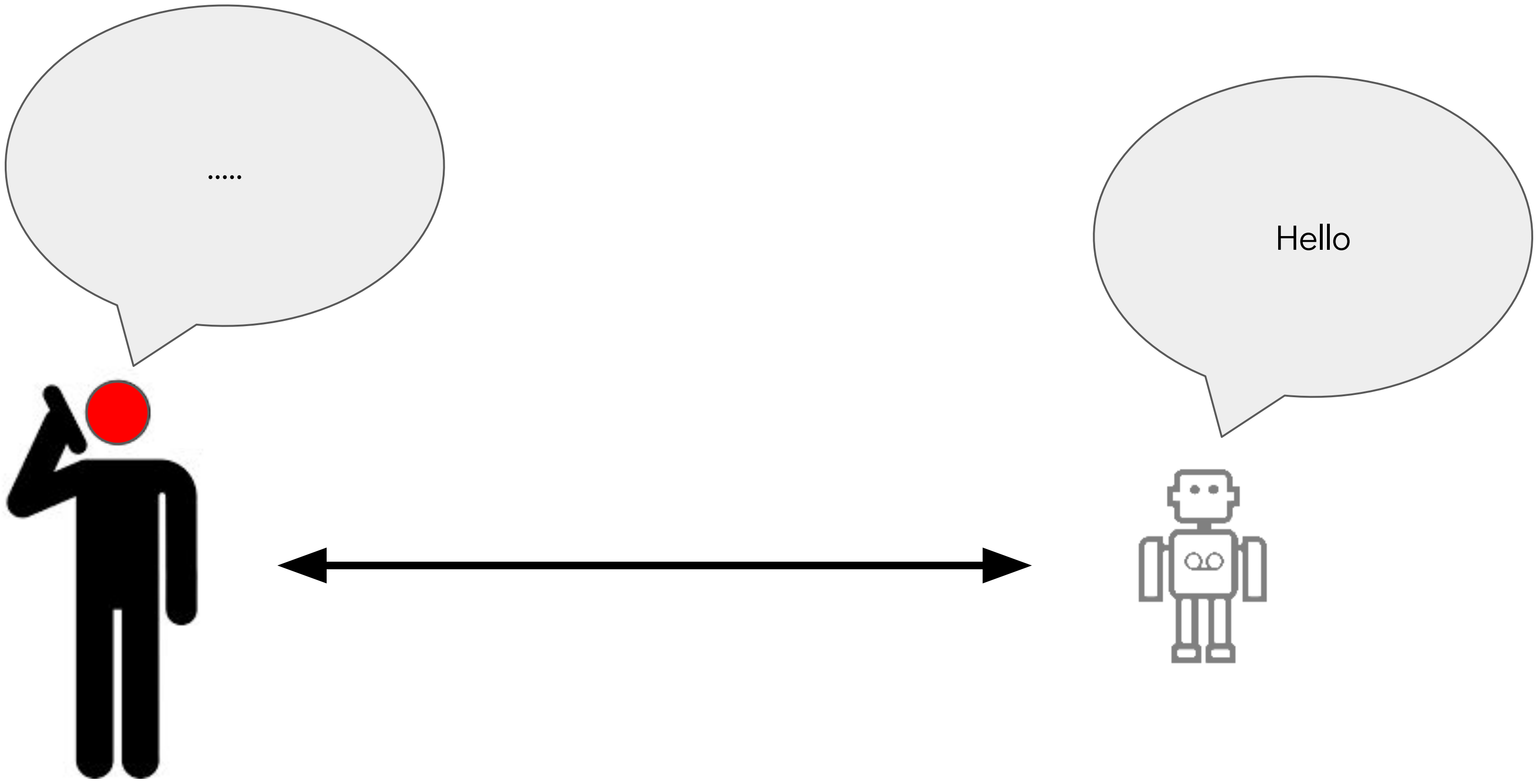
	Time	**Predictability**	**Complexity**
Optimize	Latency ⌛	Stochasticity 🎲	Scale ⚖
Manage	Jitter & Aperiodicity 🌊	Nonlinearity □	Interactivity □


Can you transfer some money please?

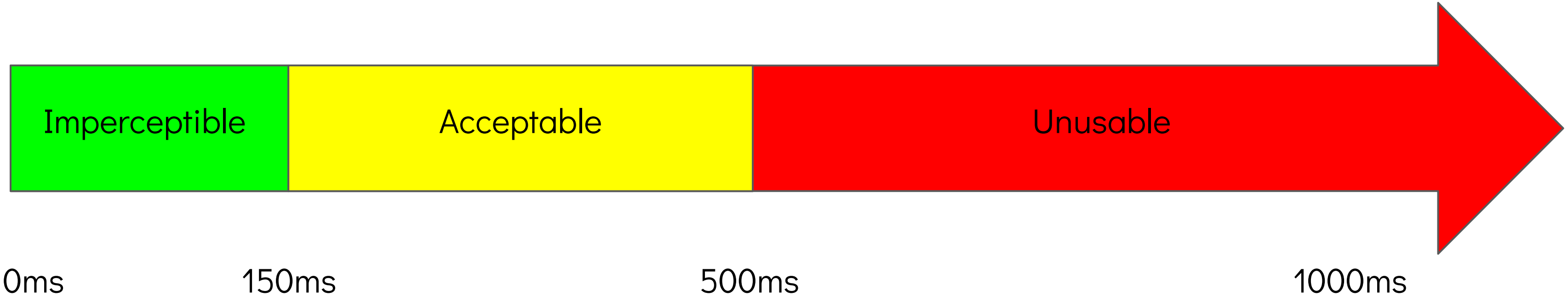

Hello?


Please...
How can I help you?

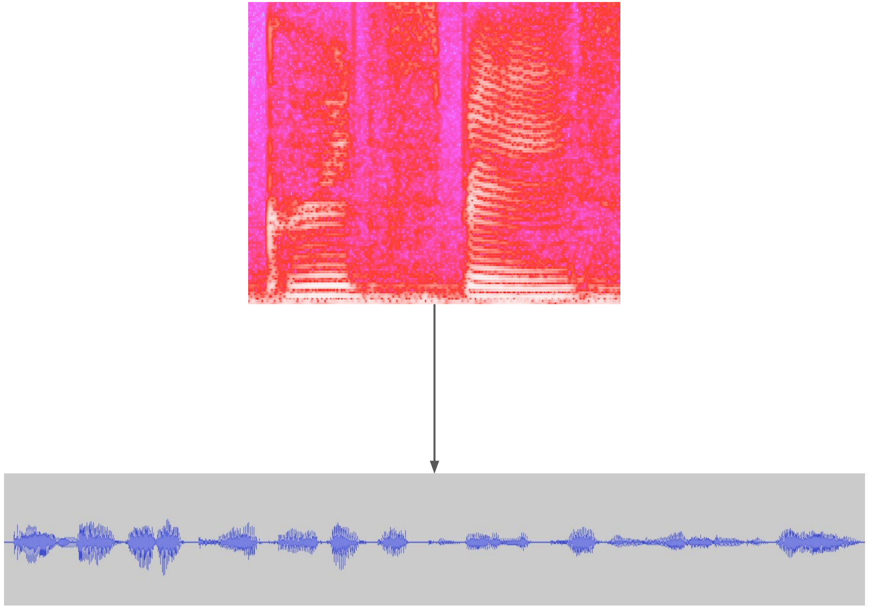



.....
Hello

Latency


Imperceptible
Acceptable
Unusable
0ms
150ms
500ms
1000ms

Where does the latency come from?



WaveNet Vocoding

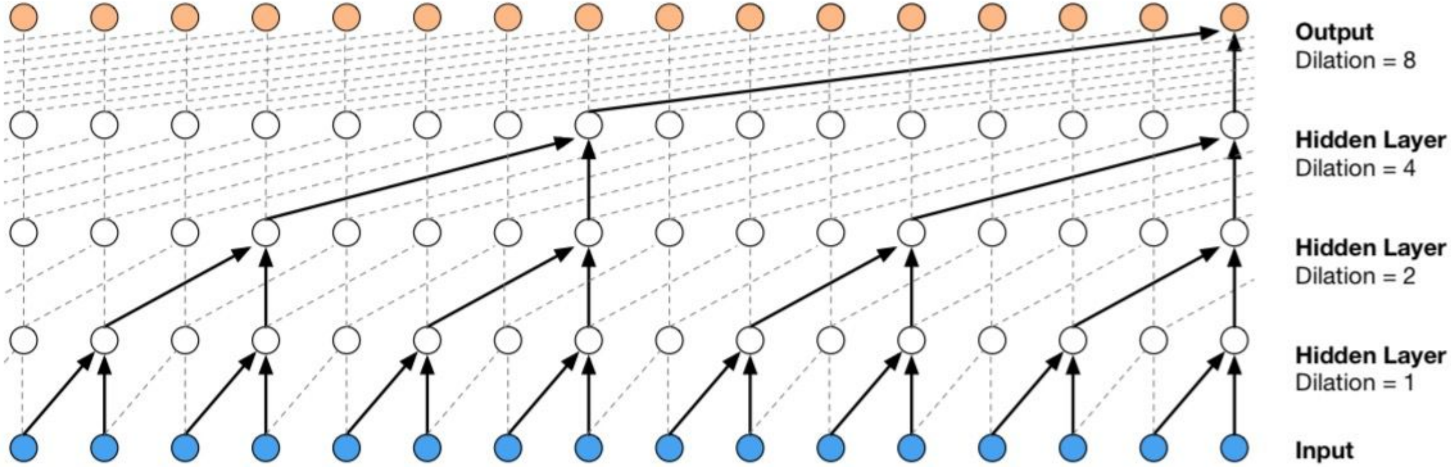
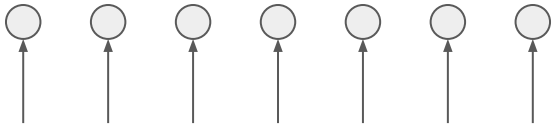


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

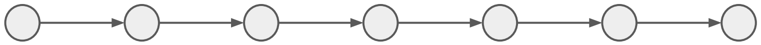
Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).

WaveNet

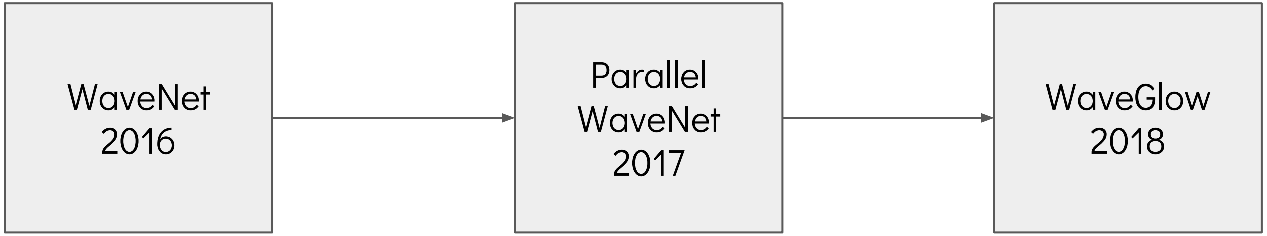
Training



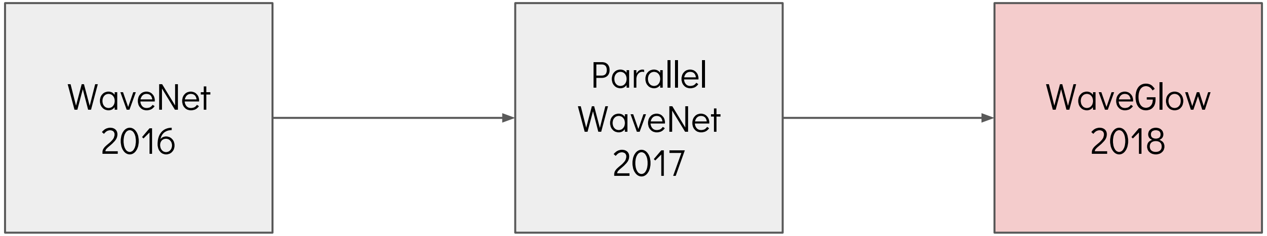
Inference



Vocoder Evolution



Vocoder Evolution



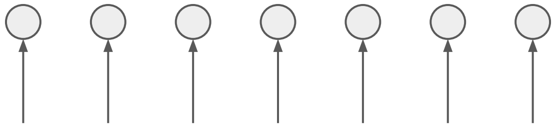
WaveNet 2016

Parallel WaveNet 2017

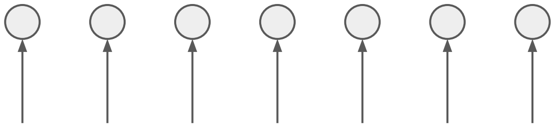
WaveGlow 2018

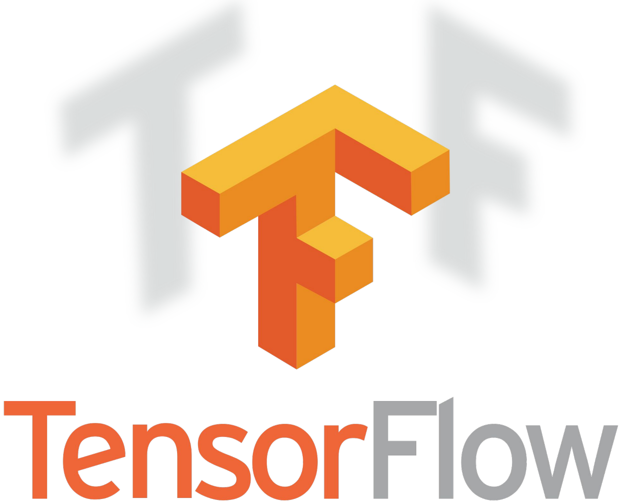
WaveGlow

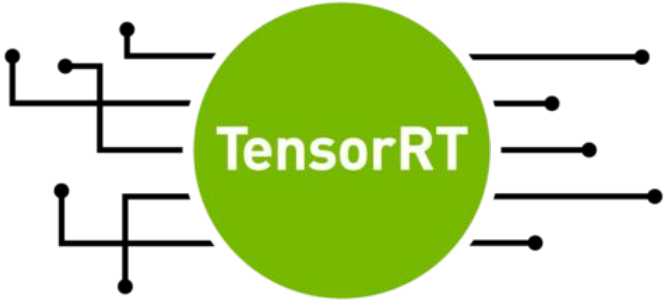
Training

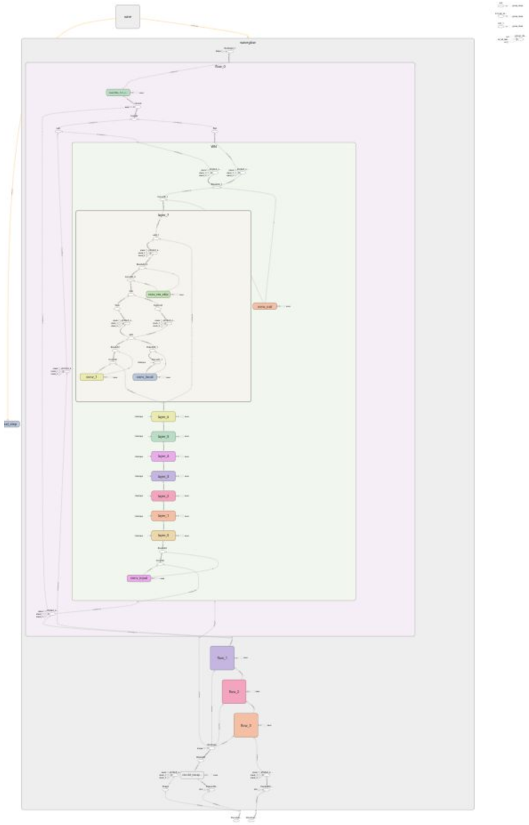


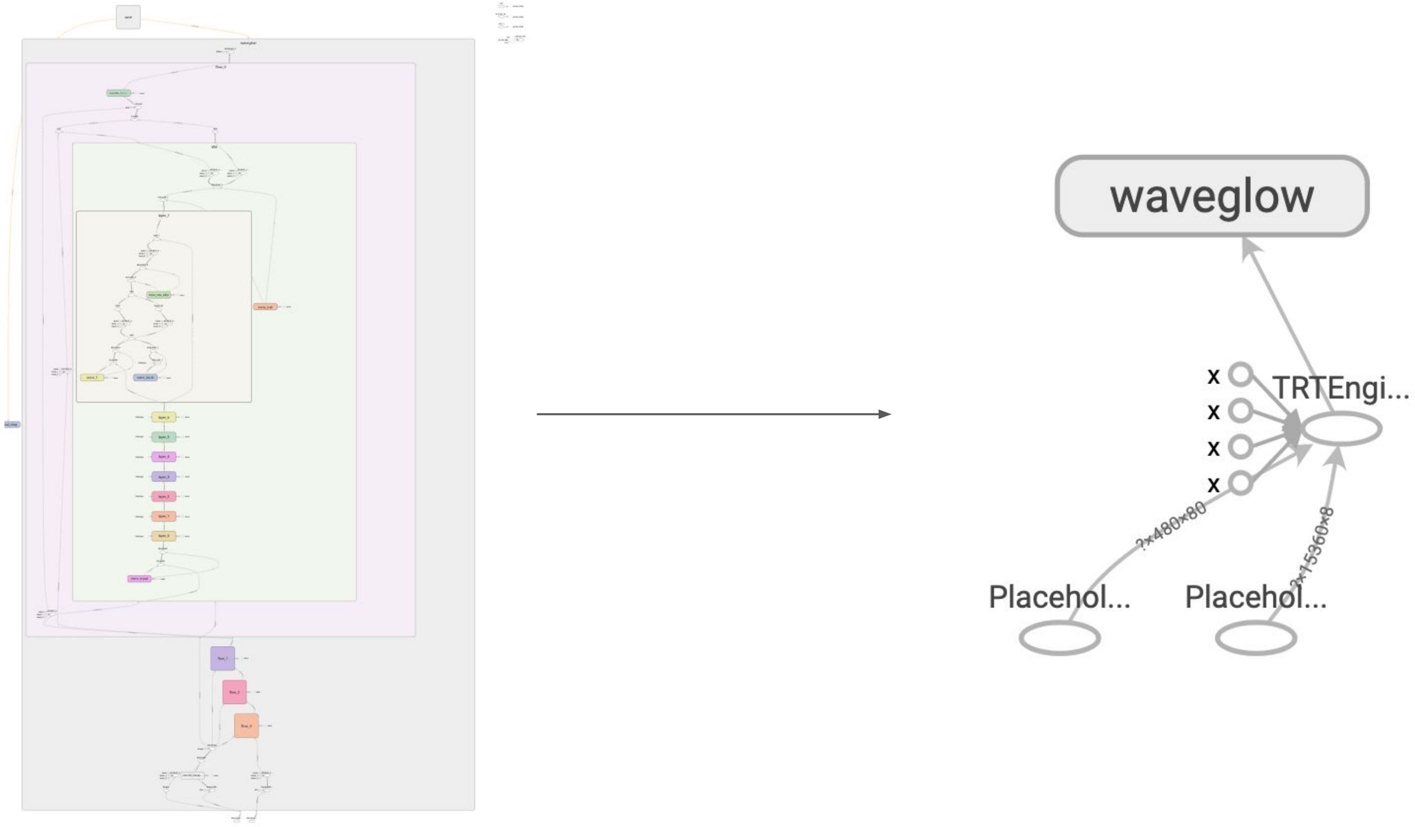
Inference




TensorFlow


TensorRT


saver
waveglow
flow_0
WN
layer_7
layer_6
layer_5
layer_4
layer_3
layer_2
layer_1
layer_0
flow_1
flow_2
flow_3


waveglow
x
x
x
x
TRTEngi...
?×480×80
?×15360×8
Placehol...
Placehol...

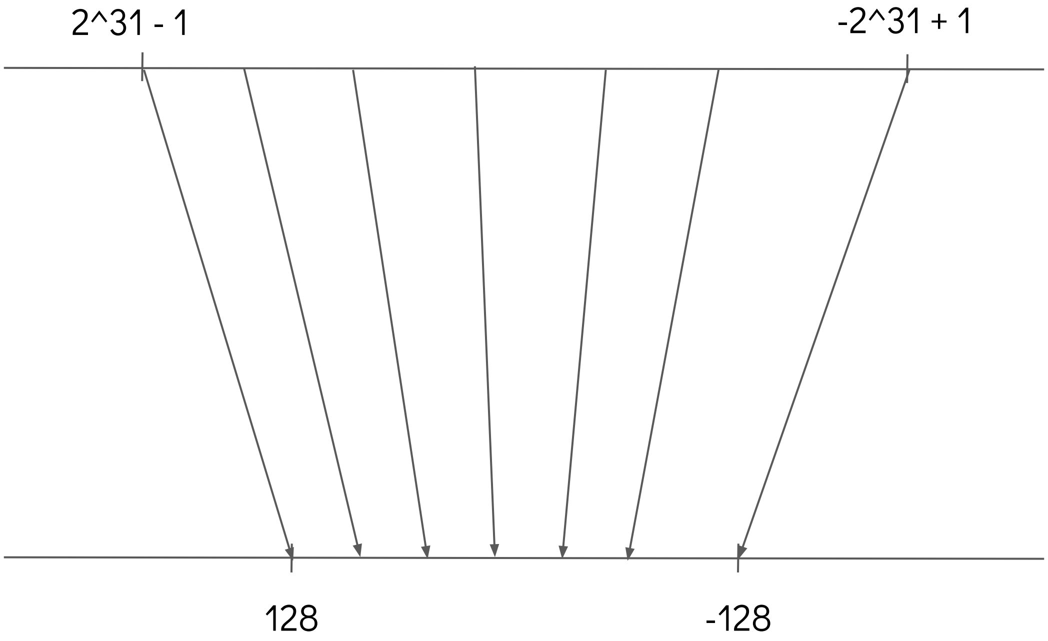
```python
from tensorflow.python.compiler.tensorrt.trt_convert import TrtGraphConverter

converter = TrtGraphConverter(
    input_saved_model_dir='my_saved_model',
    precision_mode=FP16
)
converter.convert()
converter.save('trt_saved_model')
```

Tesla V100 Latency (batch size=1)

Precision	Latency /ms	Samples per second /Hz	Speed Up
FP32	277	520	1x
FP16 (TRT)	196	735	1.4x

Can we go faster?

INT8 Calibration

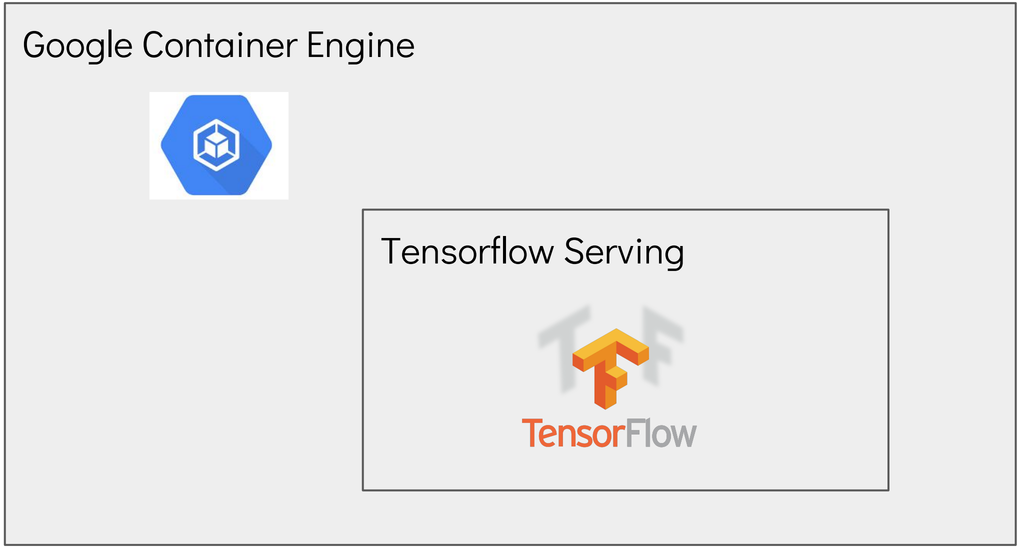


```python
from tensorflow.python.compiler.tensorrt.trt_convert import TrtGraphConverter

converter = TrtGraphConverter(
    input_saved_model_dir='my_saved_model',
    precision_mode='INT8'
)
converter.convert()
converter.calibrate(
    fetch_names=['output:0'],
    num_runs=1000,
    input_map_fn=get_examples
)
converter.save('trt_saved_model')
```

Tesla V100 Latency (batch size=1)

Precision	Latency /ms	Samples per second /Hz	Speed Up
FP32	277	520	1x
FP16 (TRT)	196	735	1.4x
INT8 (TRT)	164	878	1.7x

Deployment



Google Container Engine

Tensorflow Serving

	Time	Predictability	Complexity
Optimize	✓Latency ⌛	✓Stochasticity 🎲	✓Scale ⚖️
Manage	✓Jitter & Aperiodicity 🌊	✓Nonlinearity🦋	✓Interactivity 🤖

Enables...

THE FOLLOWING IS A REAL INTERACTION

Access the demo video:

Video on Vimeo Here

Want to demo Grace at GTC?
Do you operate a call center?
Do you need speech processing or automation?
gtc@gridspace.com

Do you want to work at Gridspace?
hiring@gridspace.com

Thank you everyone!