GDC

Developing and Running Neural Audio in Constrained Environments

Carter Huffman
(CTO / Co-Founder, Modulate)
Brendan Kelly
(Research Engineer, Modulate)

GAME DEVELOPERS CONFERENCE
MARCH 16–20, 2020 | #GDC20

Objectives

1. Practical lessons and tools we have picked up to effectively train speech synthesis models.
2. Running these models real-time in gaming environments.

Agenda

1. Overview of Neural Audio models
2. Challenges for Development
3. Challenges for Runtime
4. Epilogue

Agenda

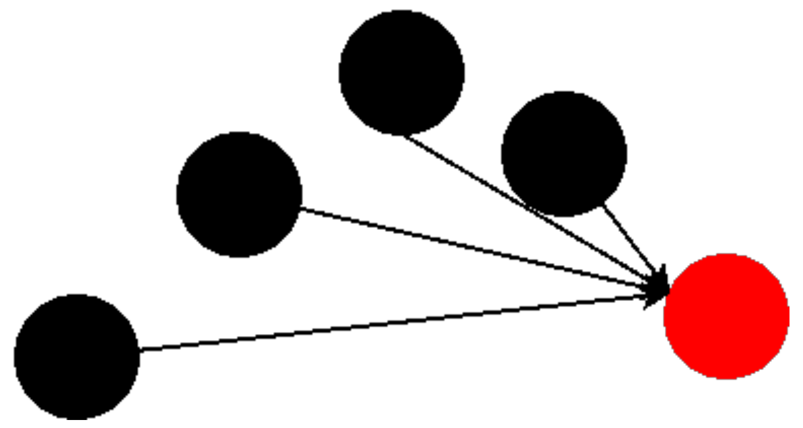
1. **Overview of Neural Audio models**
2. Challenges for Development
3. Challenges for Runtime
4. Epilogue

Autoregressive Approach

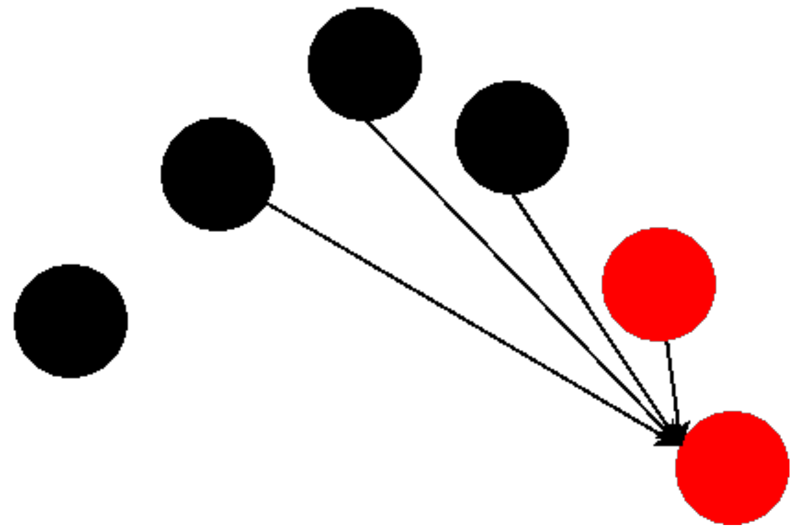
Parametric:

Autoregressive:

Predicting Sample n



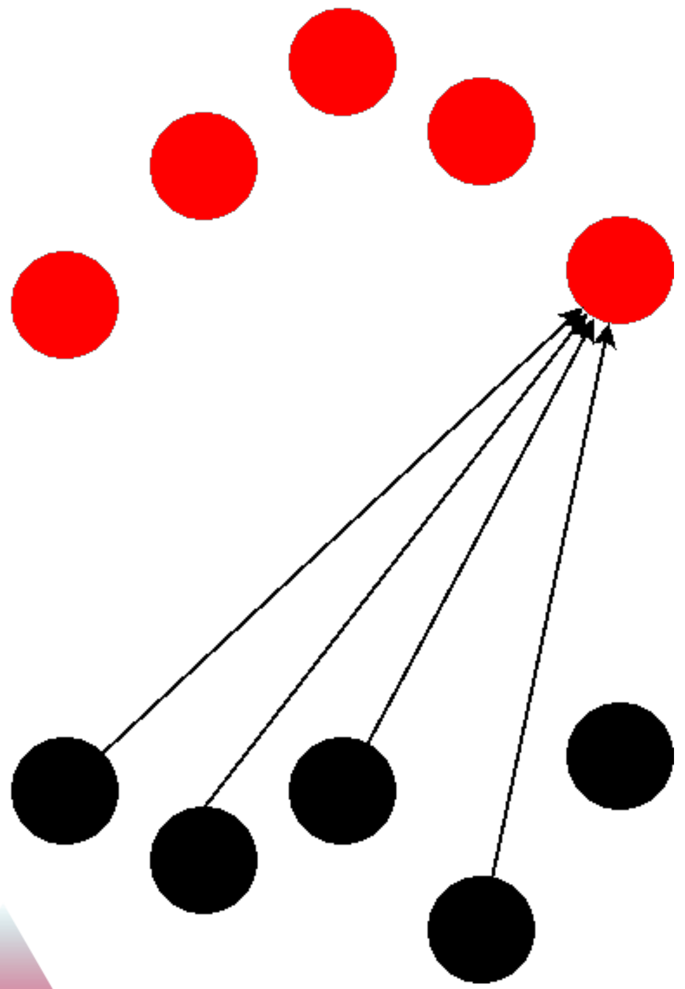
Predicting Sample n+1



Reference: [a]

Parallelized approach

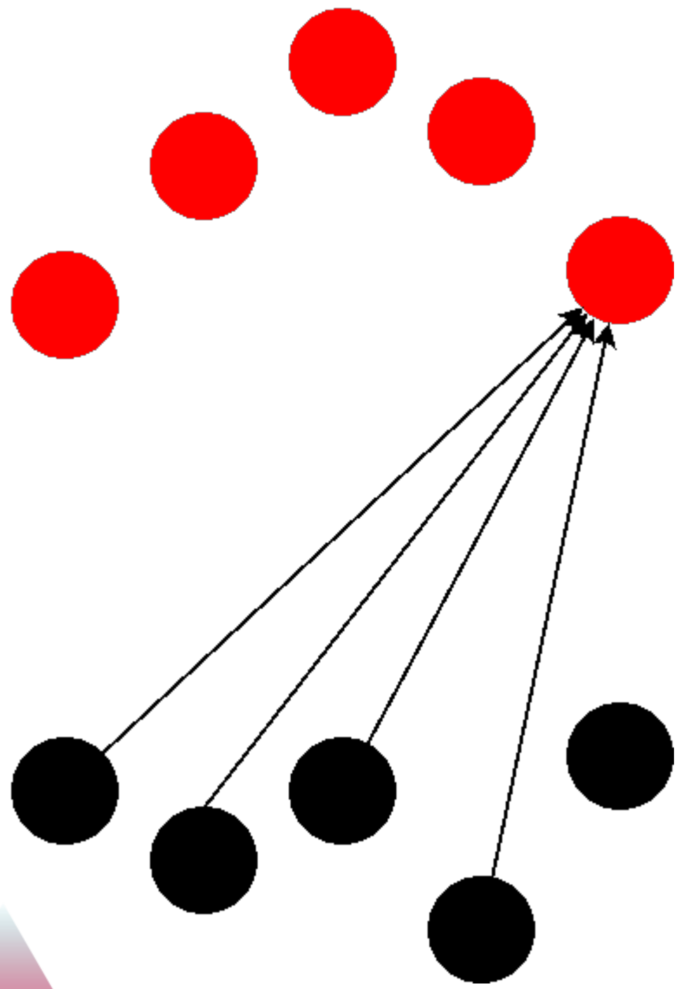
Can predict many samples in parallel



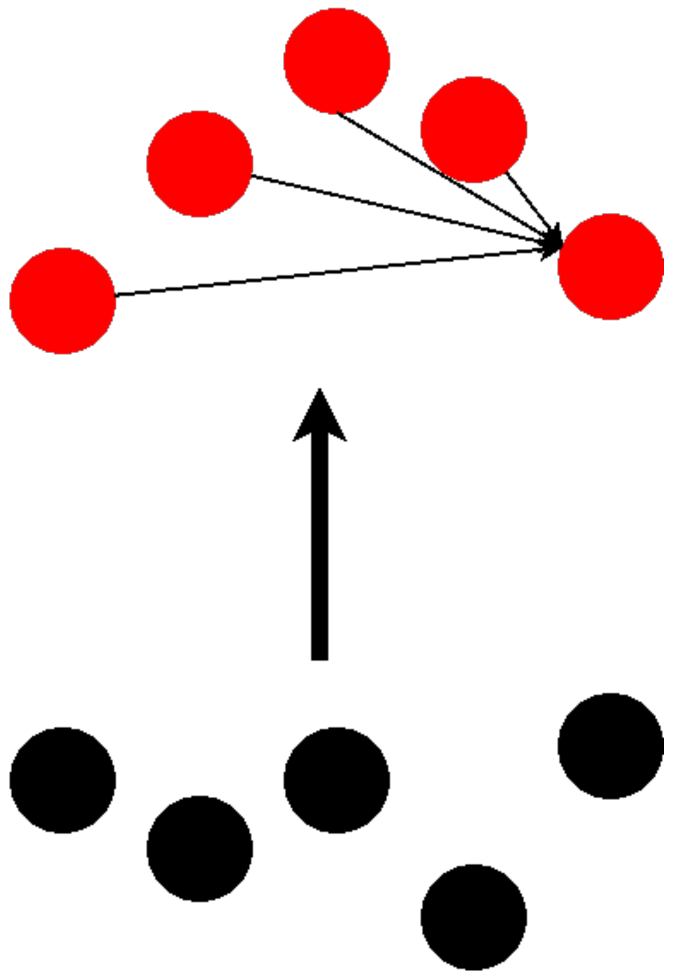
Reference: [b]

Parallelized approach

Can predict many samples in parallel



Autoregressive model predicts likelihood of samples



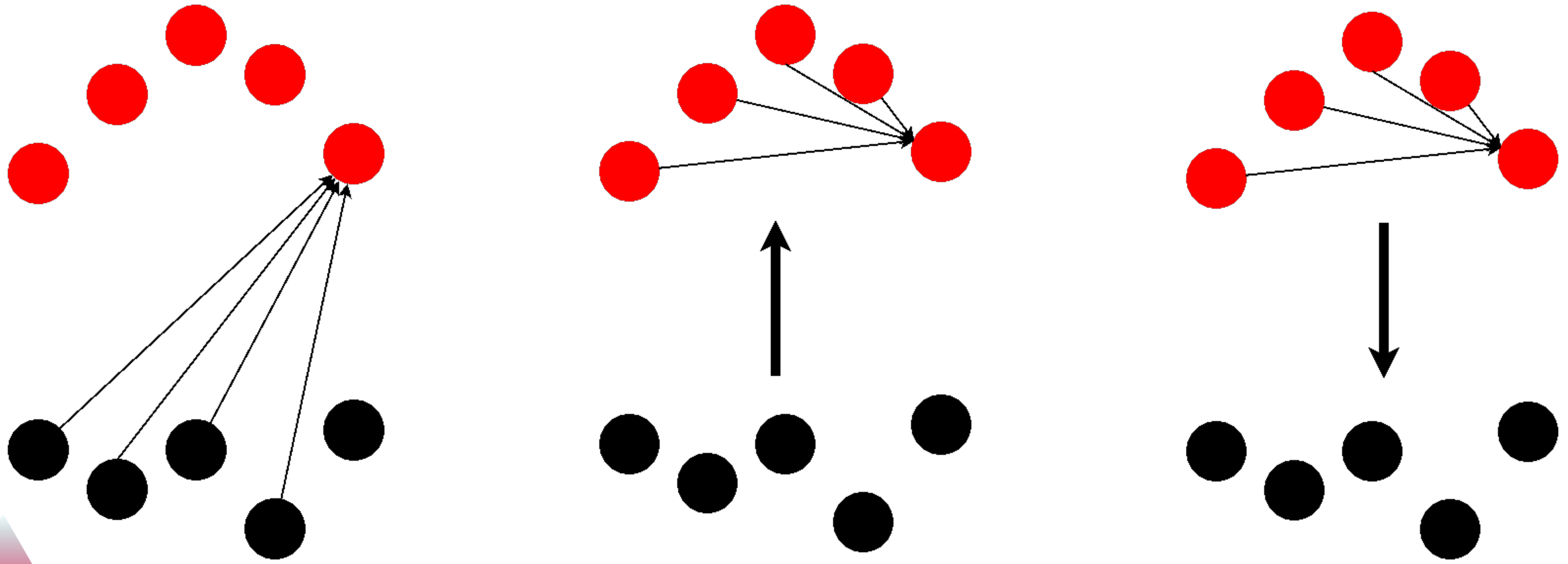
Reference: [b]

Parallelized approach

Can predict many samples in parallel

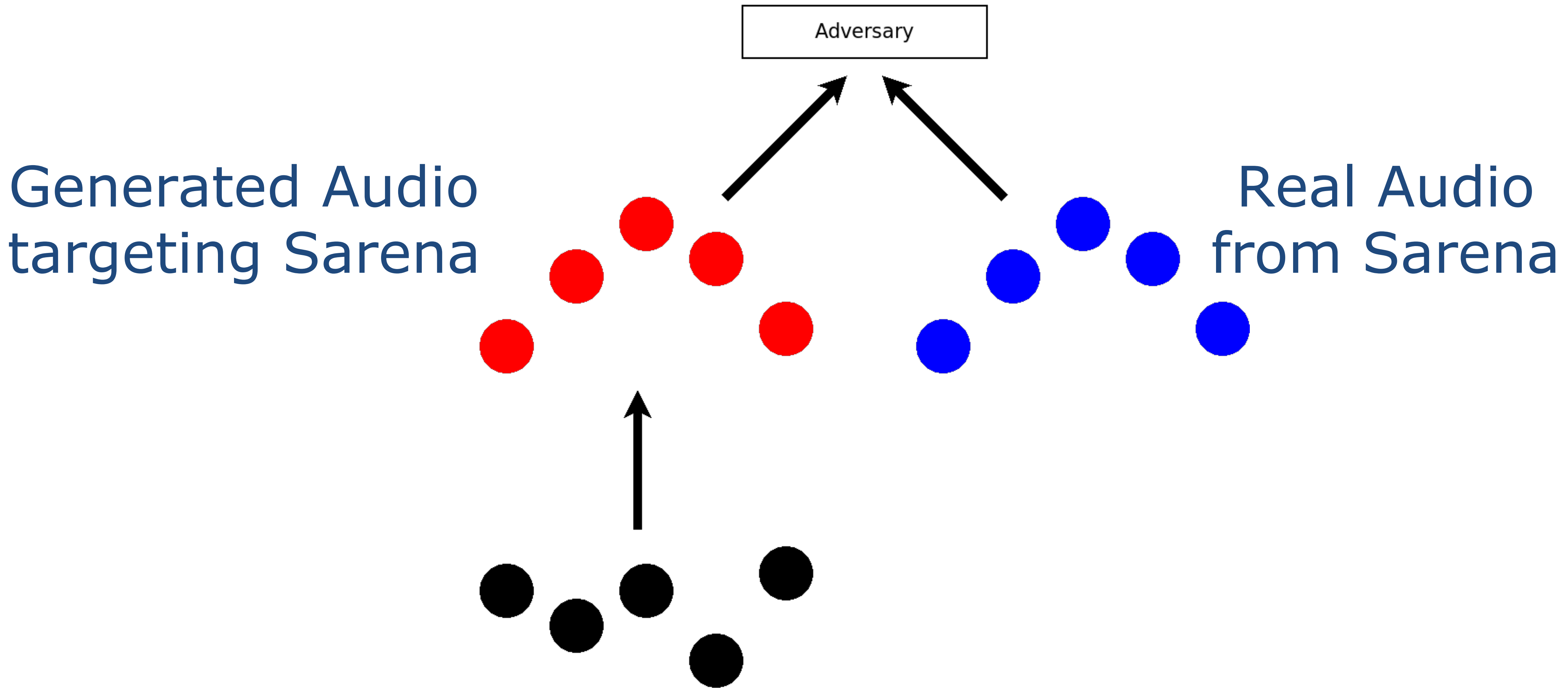
Autoregressive model predicts likelihood of samples

Gradients maximizing this likelihood are passed back

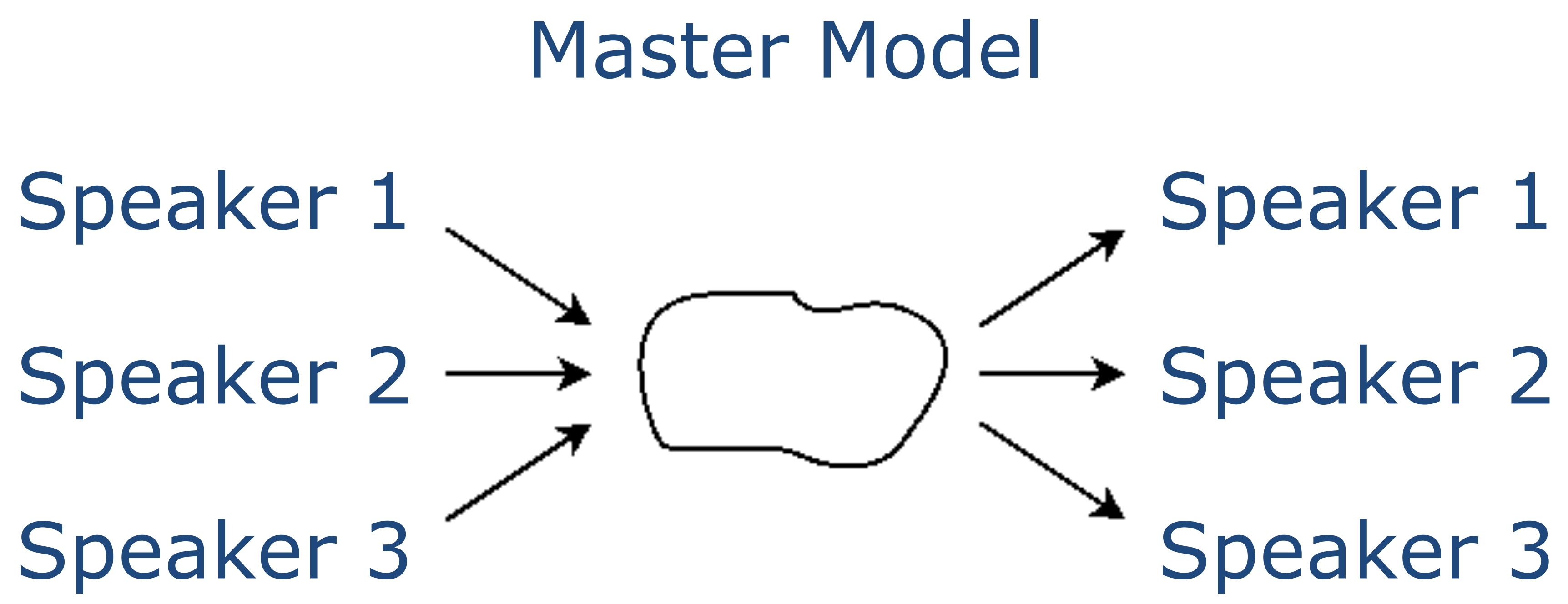


Reference: [b]

What do you do with this? Voice Skins



What do you do with this? Voice Skins



Input Speech

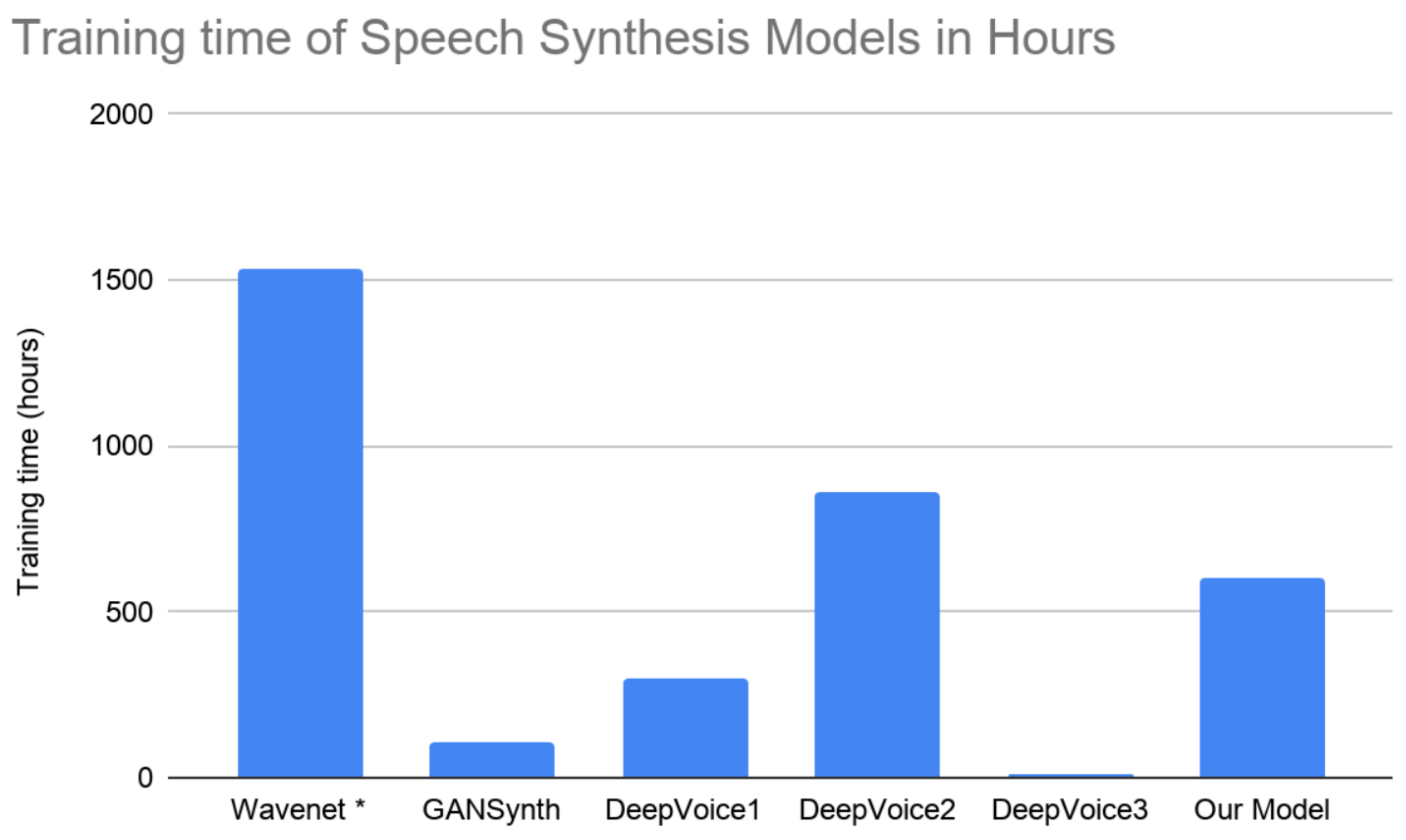
Output Speaker1

Output Speaker2

Agenda

1. Overview of Neural Audio models
2. **Challenges for Development**
3. Challenges for Runtime
4. Epilogue

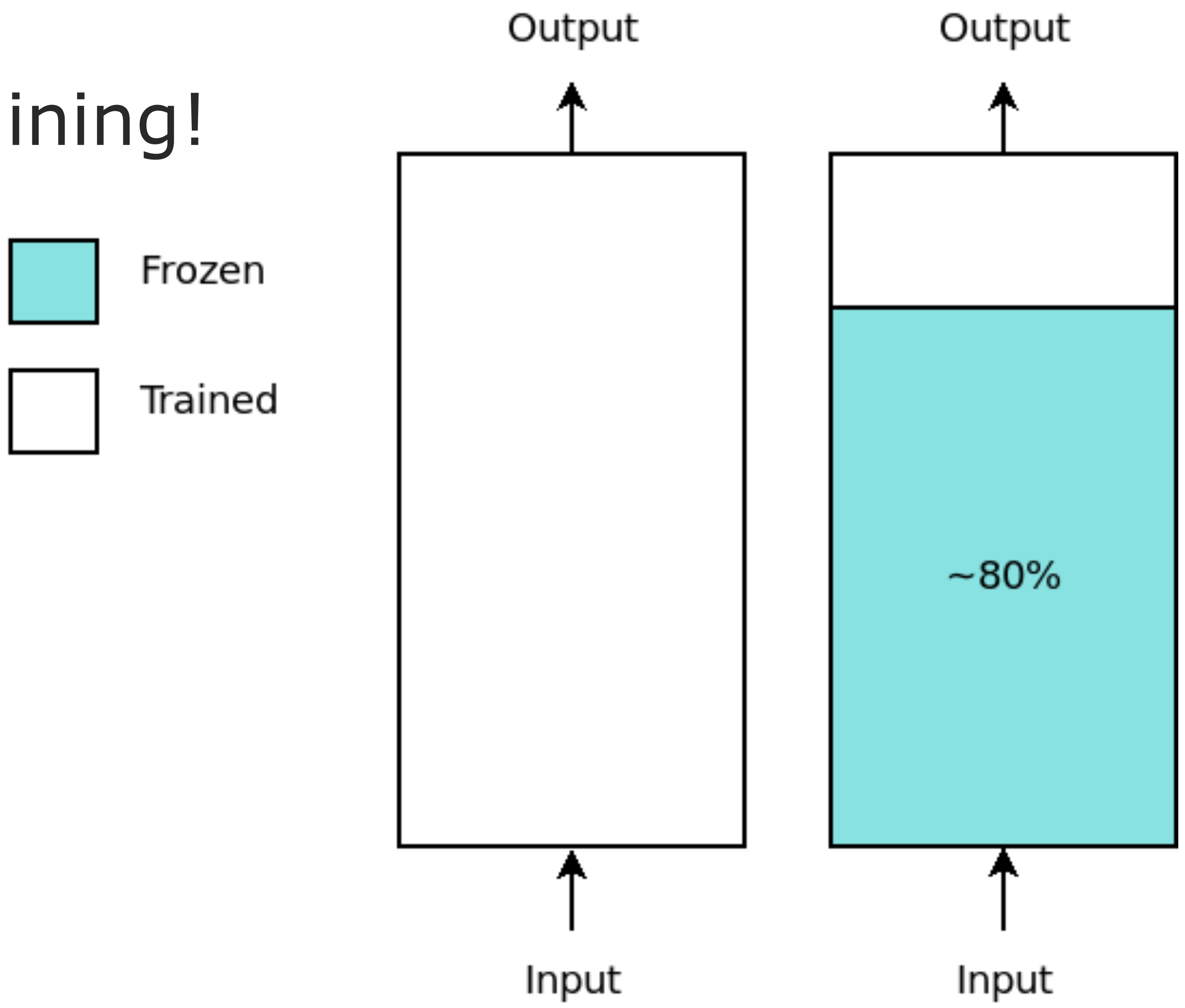
Problem: Model training is time-consuming and costly



Reference: [c], [d], [e], [f], [g], [h]

Solution: Train less and use sponsorship programs

1. Use pretrained models or stages of training!
    - Multistage training on one v100 from 25 days to 5 days.
    - Focuses experimentation on specific aspects of training



FOUNDER FRIENDLY LABS

2. Get Sponsorship!

Azure

Google Cloud

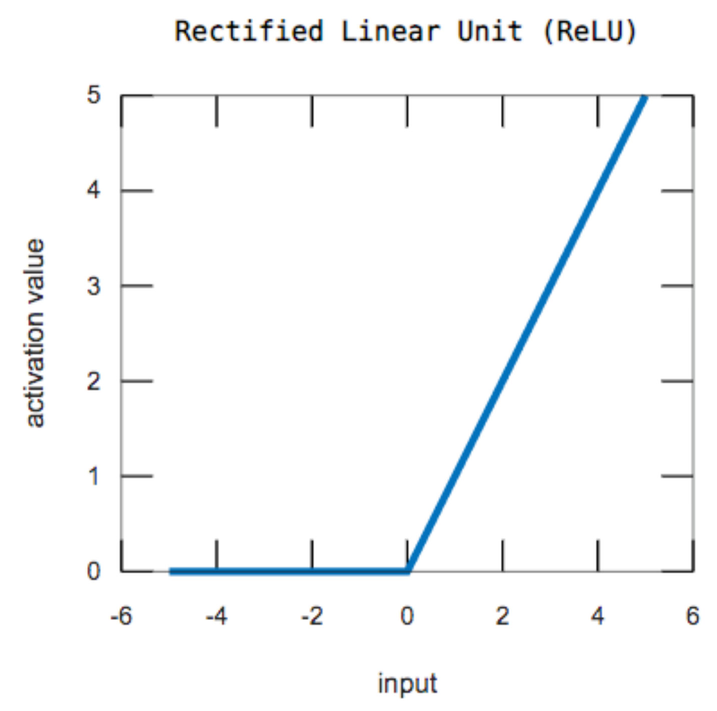
aws

Solution: Kill experiments early based on heuristics

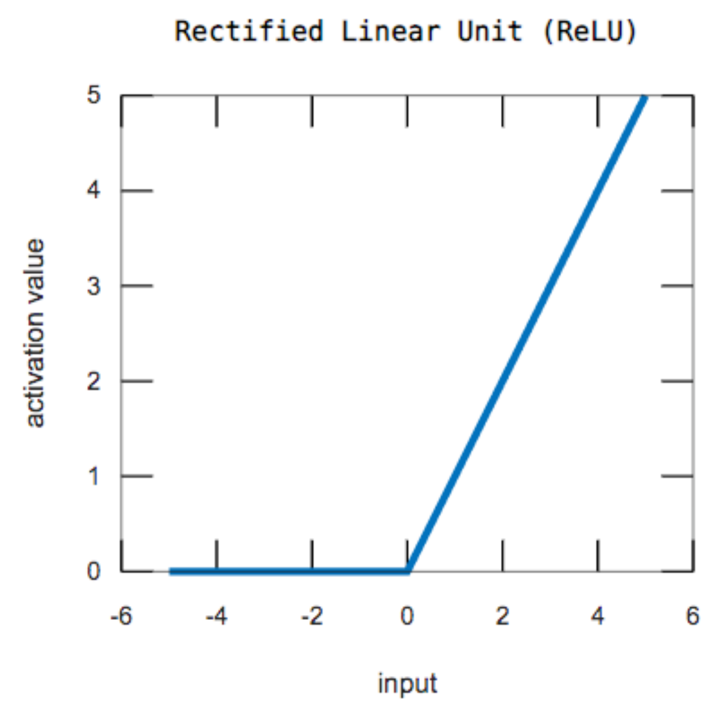
Find correlations between negative results of experiments such as model collapse, poor generalization, poor performance

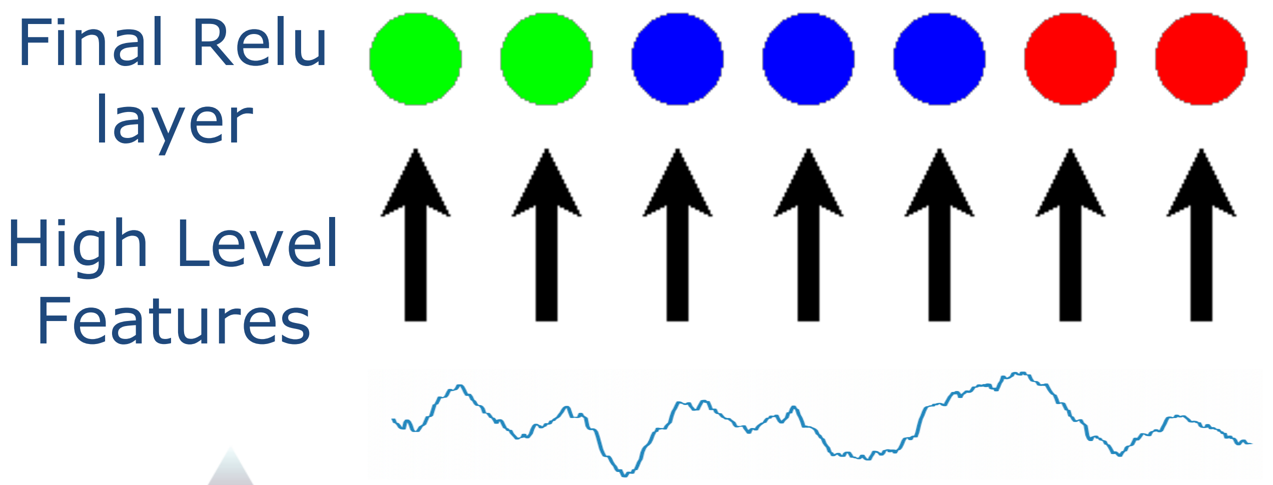
- Dead relu monitoring
- Other heuristics (power loss collapse, healthy average activation spread and more in [1])

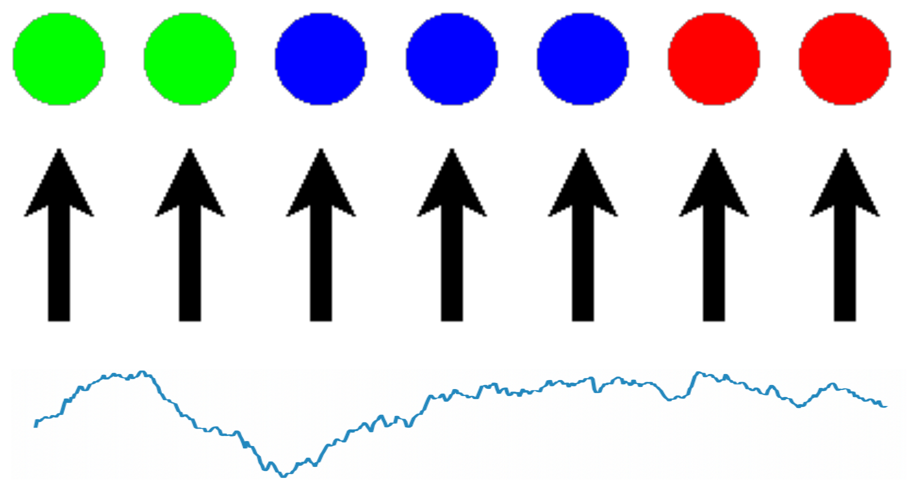
Example Solution: Dead Relus



Example Solution: Dead Relus

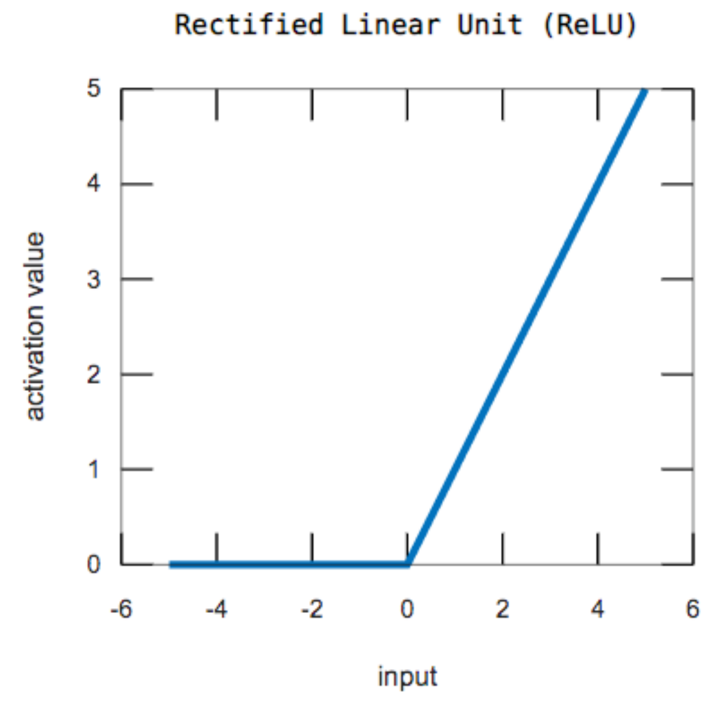


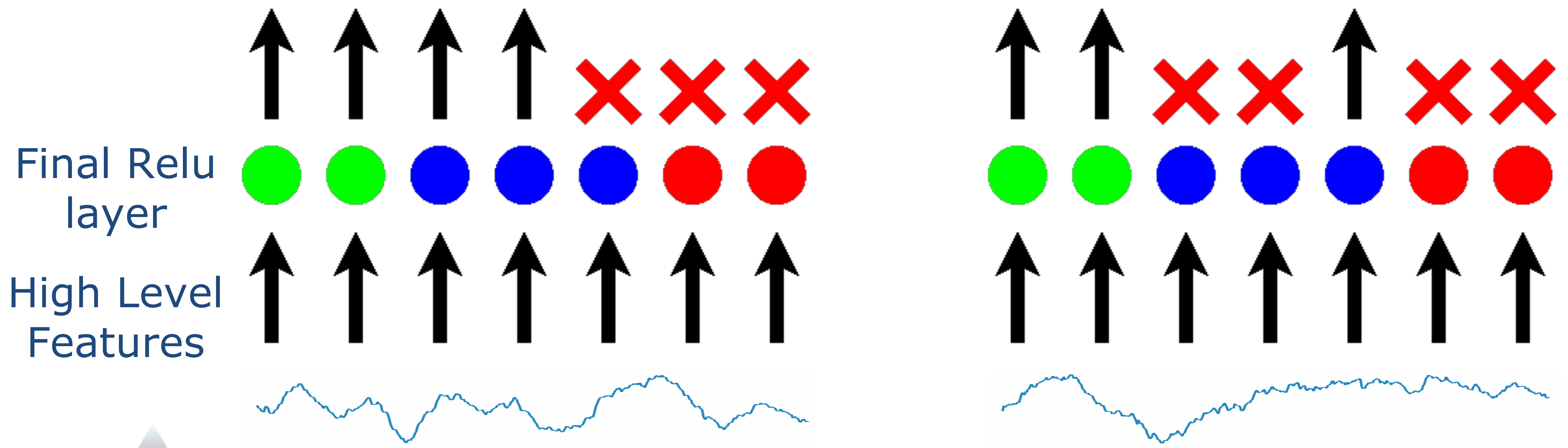




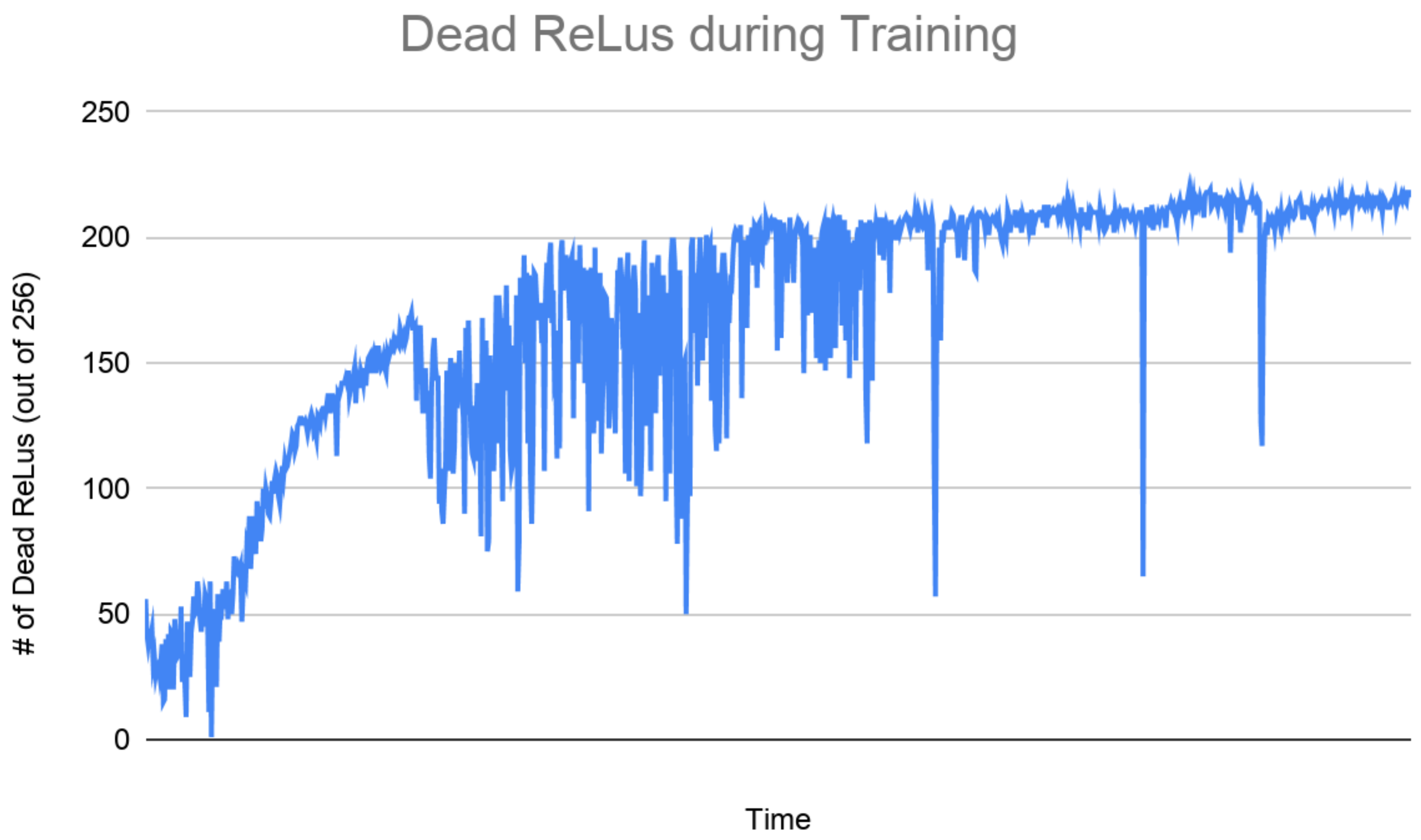
Example Solution: Dead Relus

🟢	Evergreen
🔵	Varies depending on input
🔴	Always 0 – Dead Relus





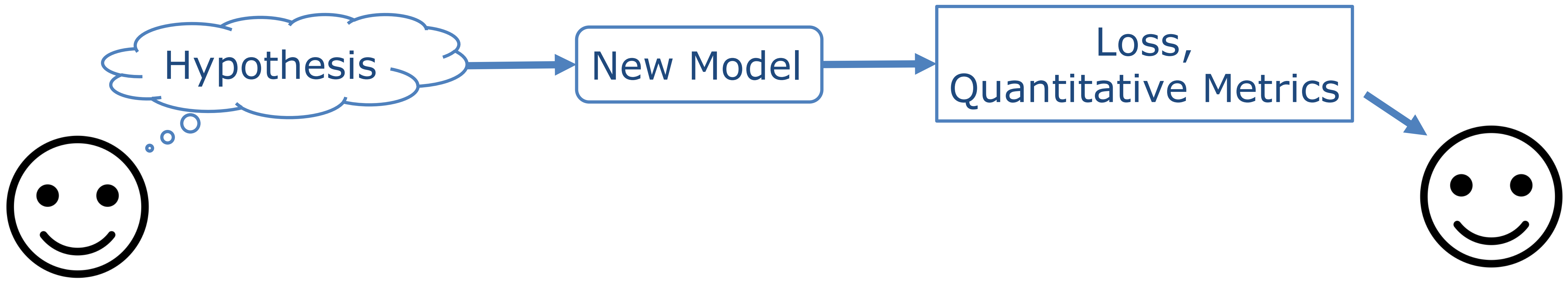
Example Solution: Dead Relus



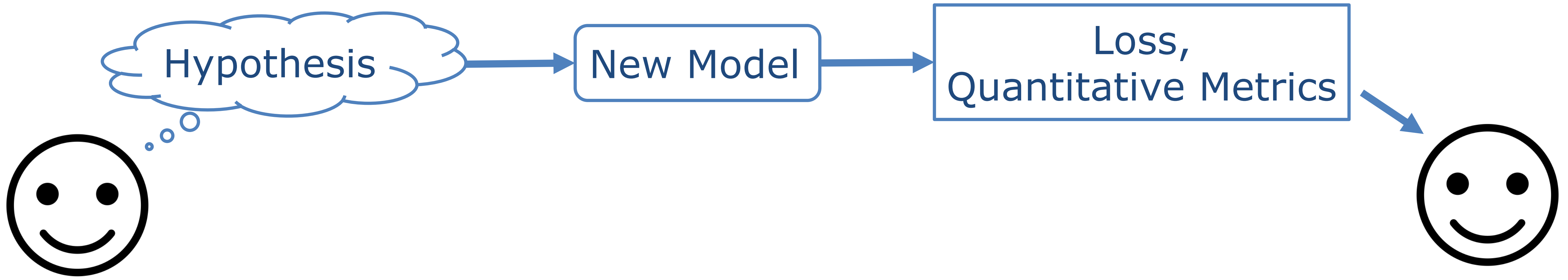
Result of tracking Correlations

- Terminate failing experiments after hours or 1-2 days
- Iterate faster by spending less time heavily analyzing model performance

Problem: How to Evaluate Models

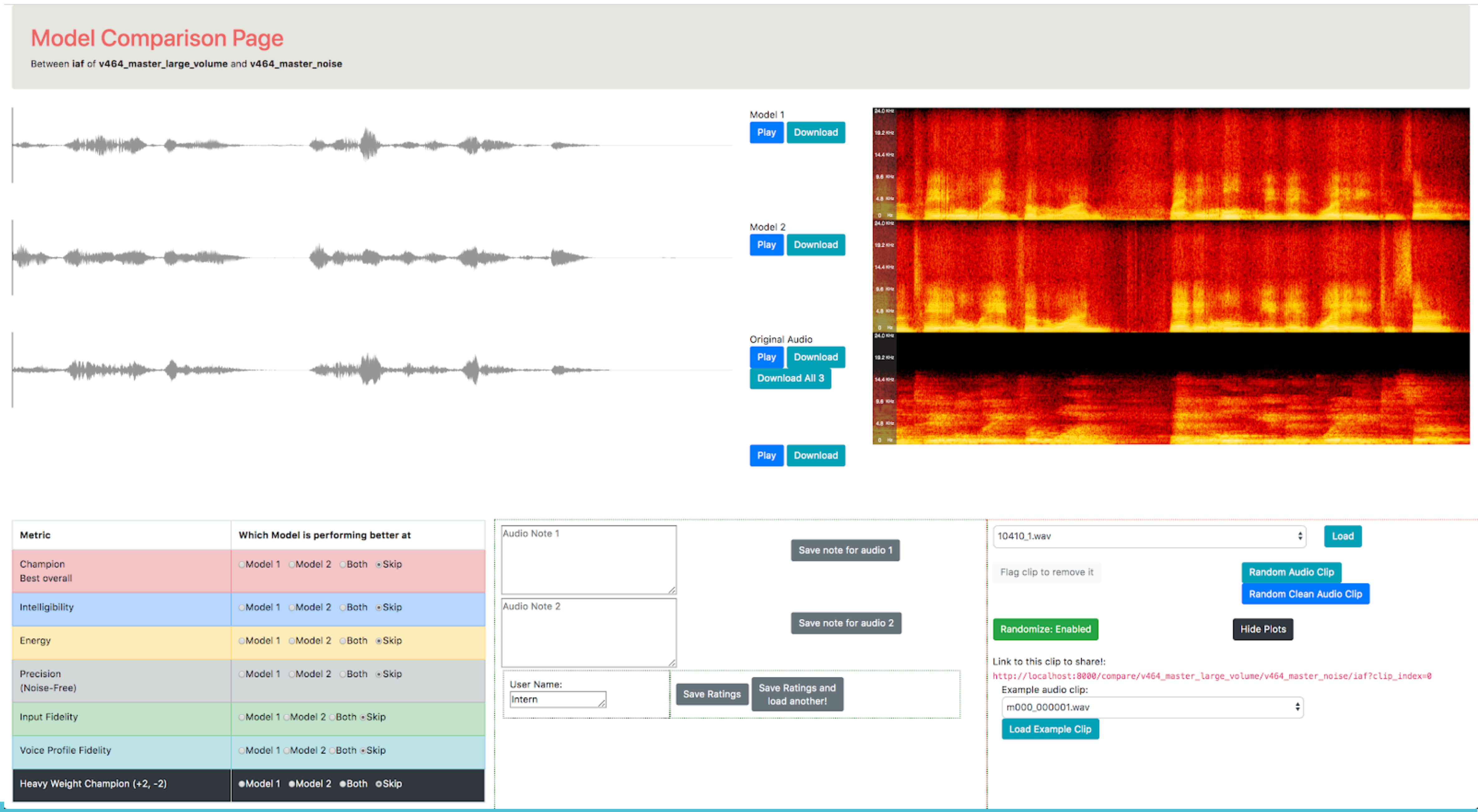


Problem: How to Evaluate Models



Available Metrics		
Training Losses	Other Neural Nets: * Inception Score * Frechet Distance	Human Score: Mean Opinion Score (MOS)
- Difficult to weigh different losses, and doesn't always correlate with performance	- Highly dependent on the dataset and task for evaluation nets	- Slow/costly and difficult to target specific aspects of media

Solution: Design assessment tools for your needs



Model Comparison Page

Between **iaf** of **v464_master_large_volume** and **v464_master_noise**

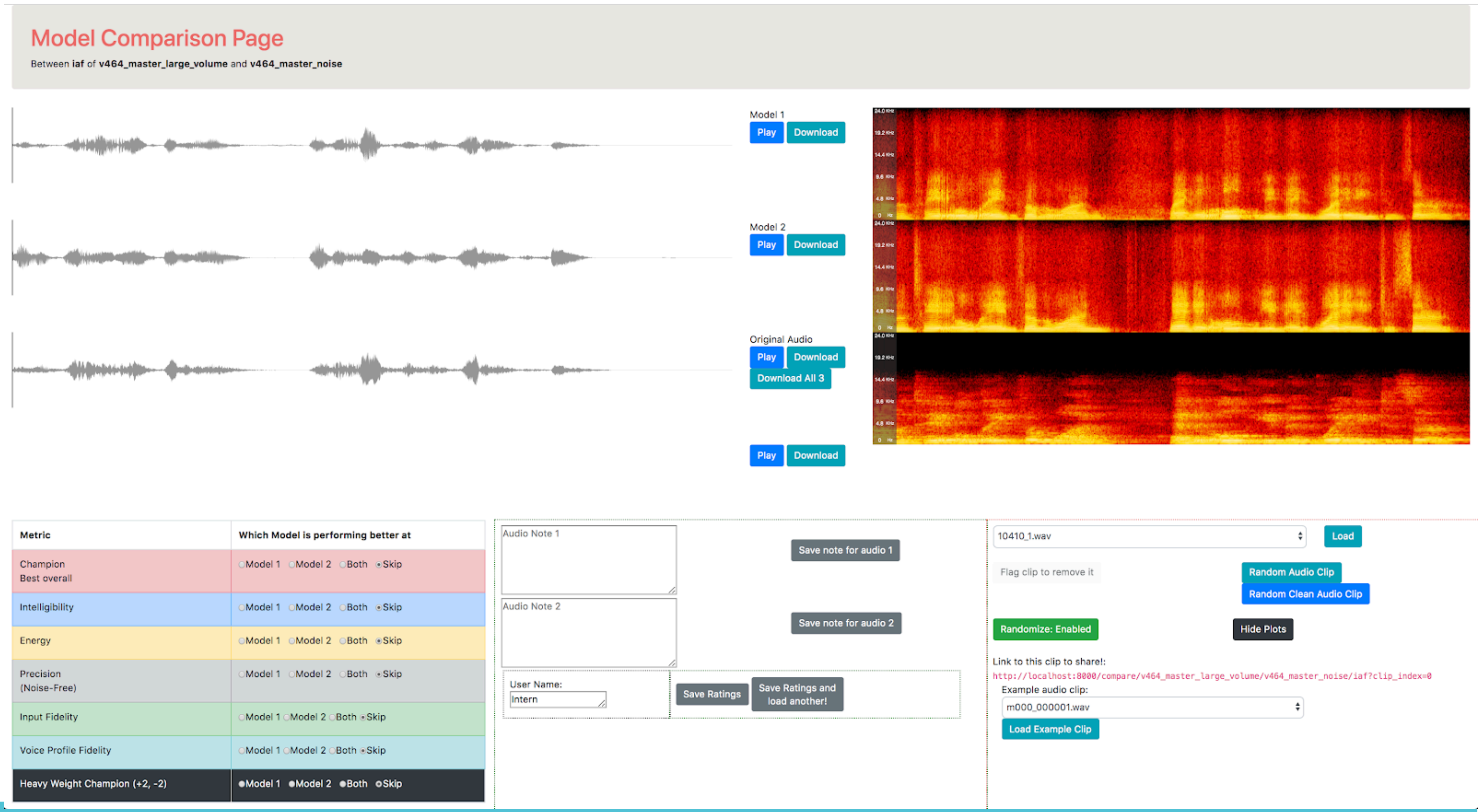
Metric	Which Model is performing better at
Champion Best overall	Model 1 Model 2 Both [✓ Skip]
Intelligibility	Model 1 Model 2 Both [✓ Skip]
Energy	Model 1 Model 2 Both [✓ Skip]
Precision (Noise-Free)	Model 1 Model 2 Both [✓ Skip]
Input Fidelity	Model 1 Model 2 Both [✓ Skip]
Voice Profile Fidelity	Model 1 Model 2 Both [✓ Skip]
Heavy Weight Champion (+2, -2)	Model 1 Model 2 Both Skip

Link to this clip to share!:

`http://localhost:8000/compare/v464_master_large_volume/v464_master_noise/iaf?clip_index=0`

Solution: Design assessment tools for your needs

Comparison of model behaviors along specific dimensions (e.g. input fidelity, intelligibility)



Model Comparison Page

Between **iaf** of **v464_master_large_volume** and **v464_master_noise**

Metric	Which Model is performing better at
Champion Best overall	Model 1 Model 2 Both [✓ Skip]
Intelligibility	Model 1 Model 2 Both [✓ Skip]
Energy	Model 1 Model 2 Both [✓ Skip]
Precision (Noise-Free)	Model 1 Model 2 Both [✓ Skip]
Input Fidelity	Model 1 Model 2 Both [✓ Skip]
Voice Profile Fidelity	Model 1 Model 2 Both [✓ Skip]
Heavy Weight Champion (+2, -2)	Model 1 Model 2 Both Skip

Audio Note 1

Audio Note 2

User Name: Intern

10410_1.wav

Flag clip to remove it

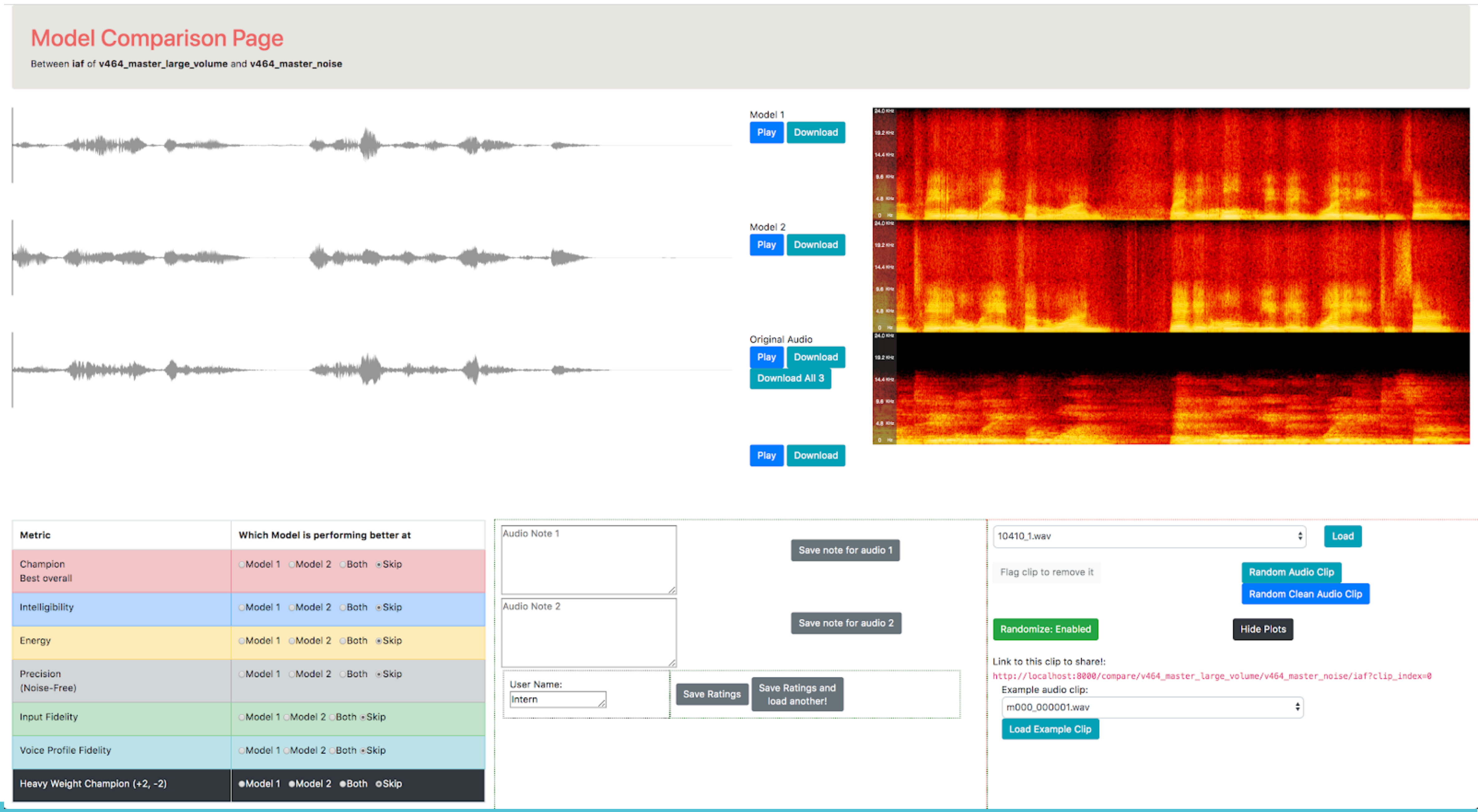
Link to this clip to share!:
`http://localhost:8000/compare/v464_master_large_volume/v464_master_noise/iaf?clip_index=0`

Example audio clip:
m000_000001.wav

Solution: Design assessment tools for your needs

Comparison of model behaviors along specific dimensions (e.g. input fidelity, intelligibility)

Simplifies collaboration



Model Comparison Page

Between **iaf** of **v464_master_large_volume** and **v464_master_noise**

Metric	**Which Model is performing better at**
Champion Best overall	Model 1 Model 2 Both [✓ Skip]
Intelligibility	Model 1 Model 2 Both [✓ Skip]
Energy	Model 1 Model 2 Both [✓ Skip]
Precision (Noise-Free)	Model 1 Model 2 Both [✓ Skip]
Input Fidelity	Model 1 Model 2 Both [✓ Skip]
Voice Profile Fidelity	Model 1 Model 2 Both [✓ Skip]
Heavy Weight Champion (+2, -2)	Model 1 Model 2 Both [✓ Skip]

User Name: Intern

10410_1.wav

Link to this clip to share!:

`http://localhost:8000/compare/v464_master_large_volume/v464_master_noise/iaf?clip_index=0`

Example audio clip: m000_000001.wav

Result: Sets of comparisons

- Takes ~15 minutes of time to compare an experiment with the baseline
- Streamlines team collaboration [2]

Equipped with these tools, training models is possible!

1. Use pretrained models to shorten training time.
2. Use sponsorship if you're in a similar position to us.
3. Use heuristics to kill experiments early.
4. Design assessment tools for specific goals

Agenda

1. Overview of Neural Audio models
2. Challenges for Development
3. **Challenges for Runtime**
4. Epilogue

Why Do You Need to Run on Device

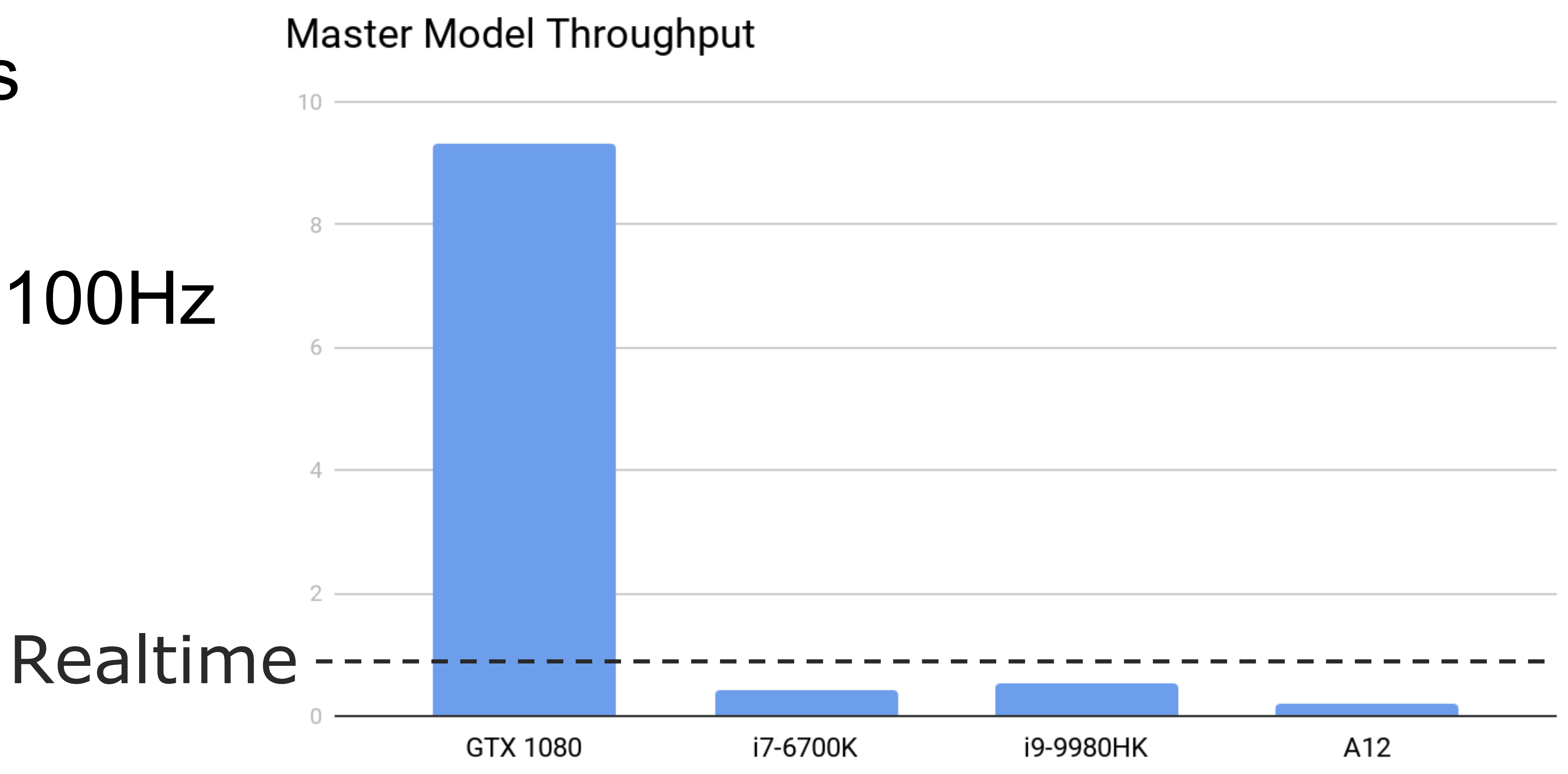
- **Low Latency Player Interaction**
  - **Audio feedback <= 30ms**
- Offline or Peer-to-Peer

Why Do You Need to Run on Device

- Low Latency Player Interaction
  - Audio feedback <= 30ms
- **Offline or Peer-to-Peer**

Constraints Running on Device

- **Reduce Resource Consumption**
- Different Device Capacities
- Optimizing for latency
- Audio Real-time friendly (~100Hz audio framerate)

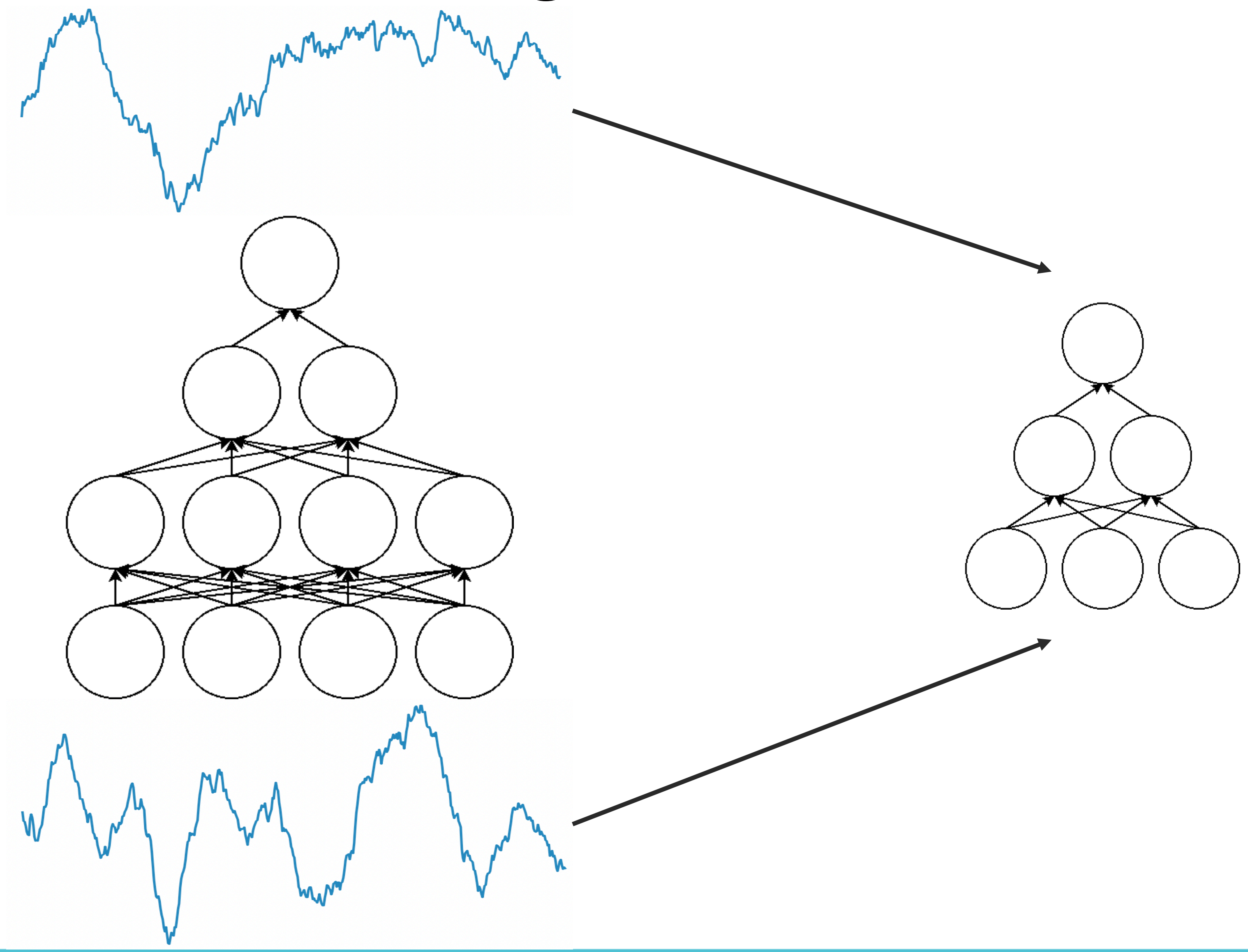


Constraints Running on Device

- Reduce Resource Consumption
- **Different Device Capacities**
- Optimizing for latency
- Audio Real-time friendly (~100Hz audio framerate)

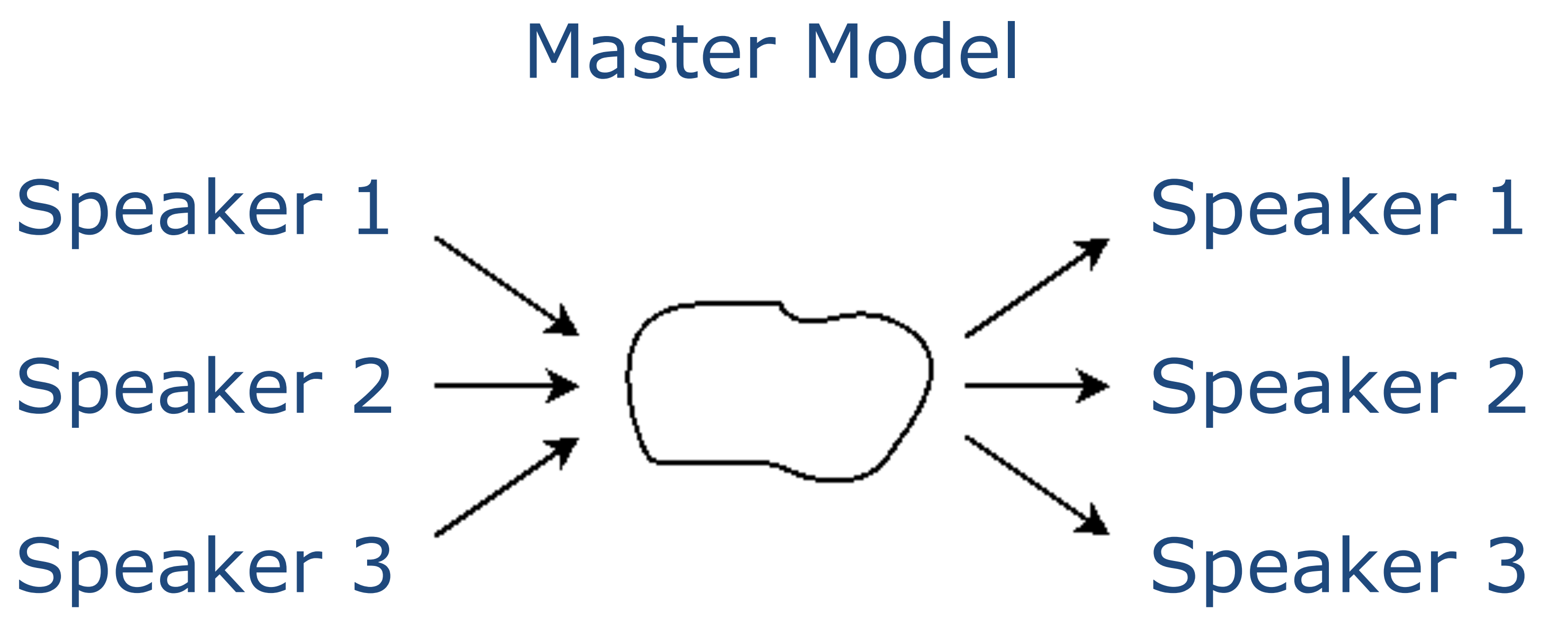
Solution - Distillation

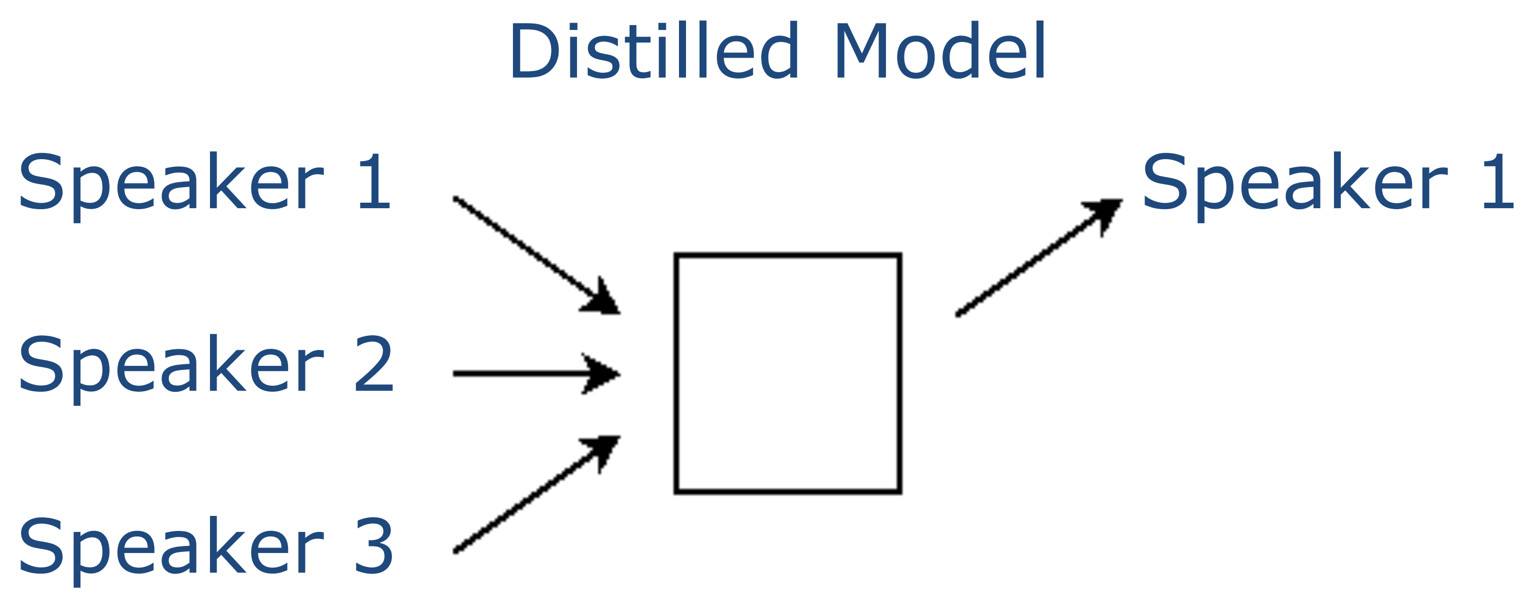
- Supervised training of smaller "distilled" model



Solution - Distillation

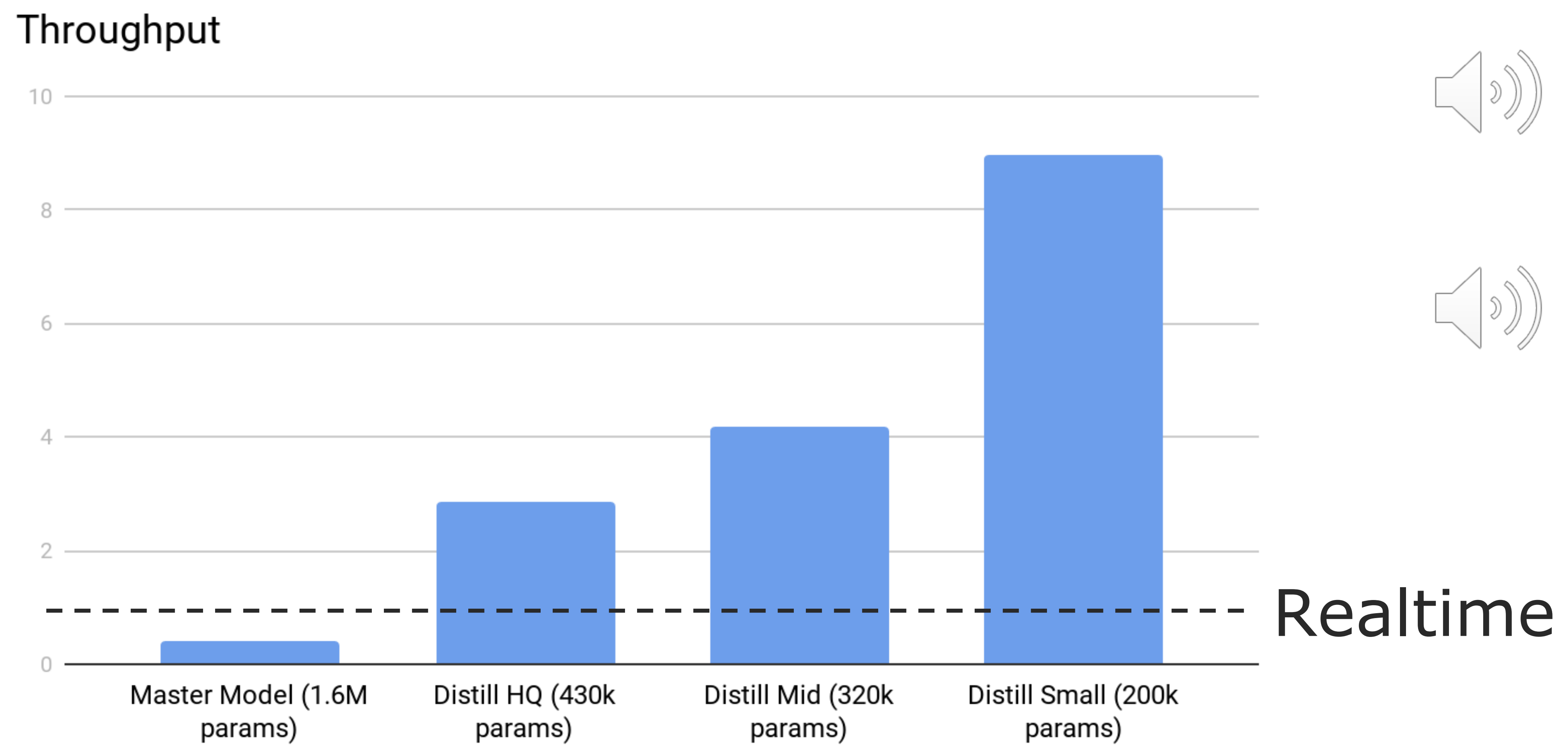
- Reduce unneeded capacity





Solution - Distillation

- Ship default & HQ versions of models



(i7-6700K)

Master Model

Distill Mid

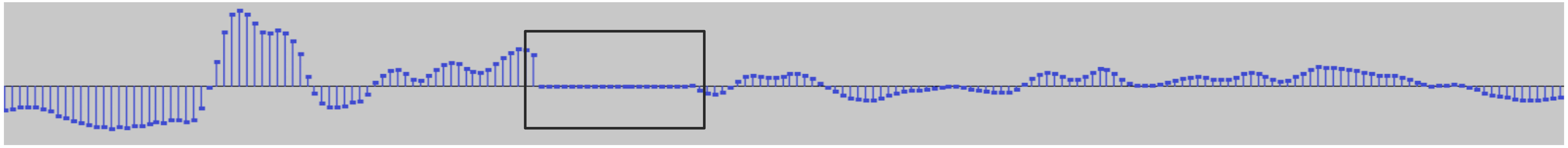
Constraints Running on Device

- Reduce Resource Consumption
- Different Device Capacities
- **Optimizing for latency**
- Audio Real-time friendly (~100Hz audio framerate)

Constraints Running on Device

- Reduce Resource Consumption
- Different Device Capacities
- Optimizing for latency
- **Audio Real-time friendly (~100Hz audio framerate)**

With Mallocs



Solution - Custom Inference Code

- Custom C++ avoids memory allocations, blocking [3]
- Use vectorization and fused-multiply-add

Solution - Custom Inference Code

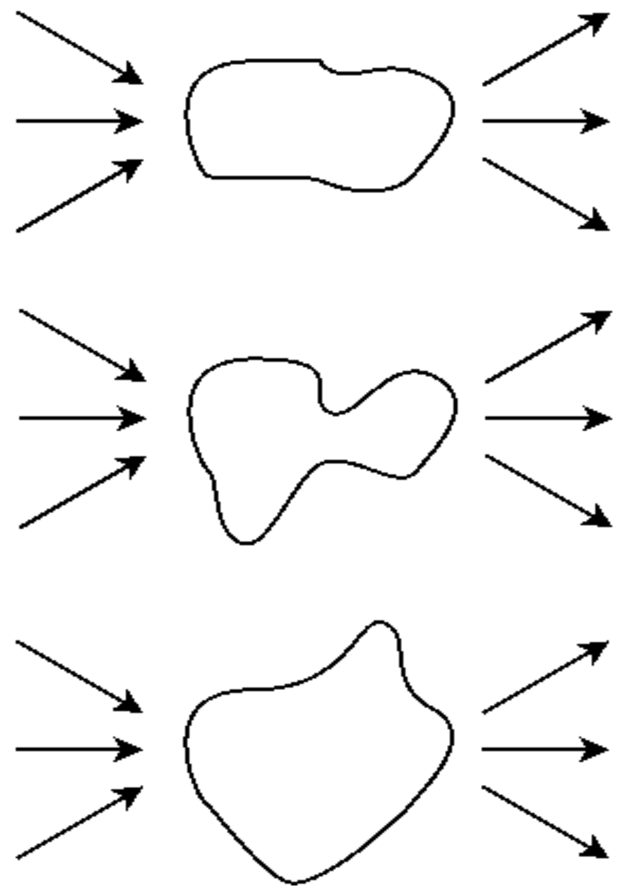
- Portable - cross-compiles to everything
- Easy integration, easy pipelines

GAME DEVELOPERS CONFERENCE
MARCH 16–20, 2020 | #GDC20

Custom Code Time Investment

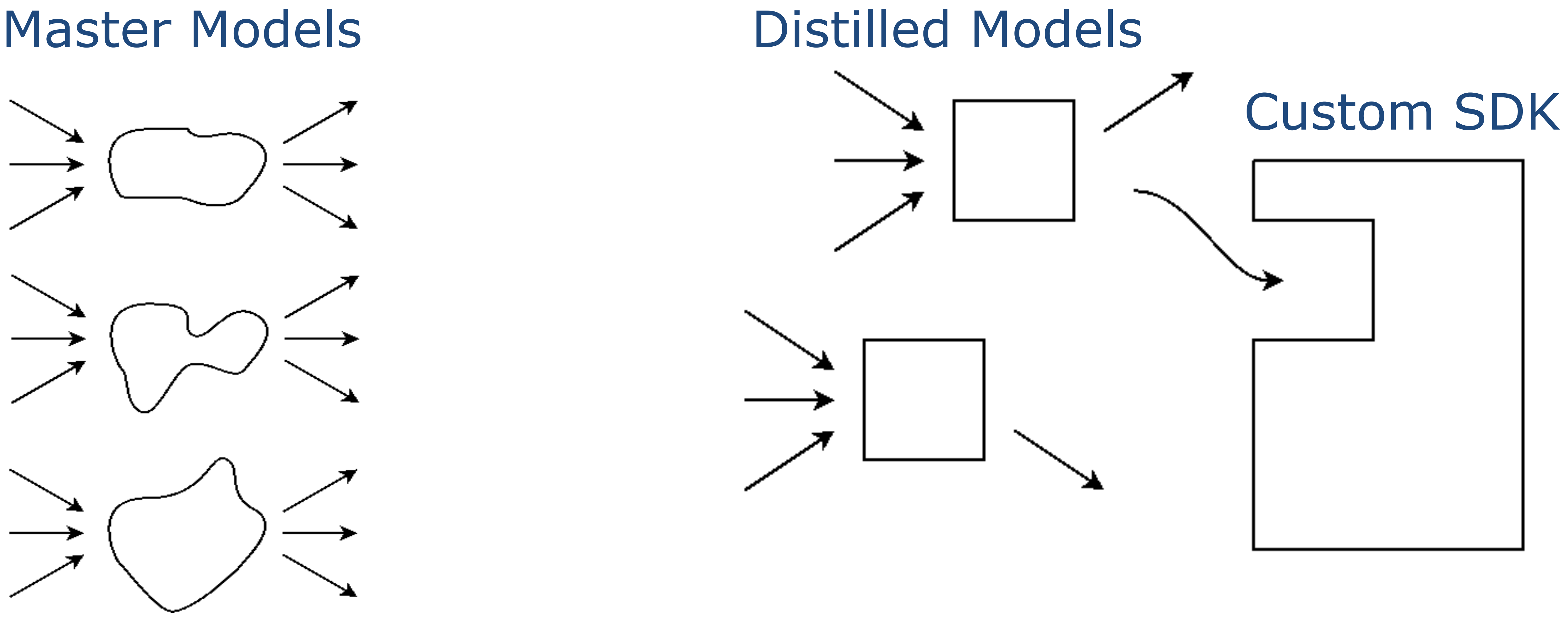
- Distillation avoids rewrites [4]

Master Models



Custom Code Time Investment

- Distillation avoids rewrites [4]



Custom Audio Code Tips

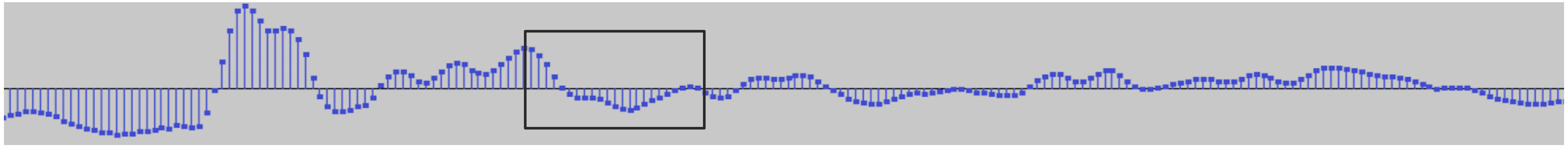
- Bundle in SRC
- Adapt to various frame sizes with flexible pre-allocation
- Include real-time audio logging [5]

Secret Rabbit Code

GDC
GAME DEVELOPERS CONFERENCE
MARCH 16–20, 2020 | #GDC20

Custom Code Results

No Mallocs



Agenda

1. Overview of Neural Audio models
2. Challenges for Development
3. Challenges for Runtime
4. **Epilogue**

Summary

- Even a small team can train and deploy ML models
- Use Pretraining, Pruning, and QA tools
- Optimize with Distillation and Custom Inference Code

Acknowledgements

- Also come talk to us for live demos!

Referenced Posts

1. https://modulate.ai/blog-collection/2020/03/trainingheuristics
2. https://modulate.ai/blog-collection/2020/03/qatools
3. https://modulate.ai/blog-collection/2020/03/nonblockingconv
4. https://modulate.ai/blog-collection/2020/03/distillation
5. https://modulate.ai/blog-collection/2020/03/realtimeaudiologging

Online References

a. https://www.deepmind.com/blog/article/wavenet-generative-model-raw-audio
b. https://arxiv.org/abs/1711.10433
c. https://cloud.google.com/compute/gpus-pricing
d. https://openreview.net/pdf?id=H1xQVn09FX
e. https://arxiv.org/pdf/1702.07825.pdf
f. https://aws.amazon.com/ec2/instance-types/p2/
g. https://arxiv.org/pdf/1705.08947.pdf
h. https://arxiv.org/abs/1710.07654