W3C Workshop on Web and Machine Learning

Introduction to the workshop

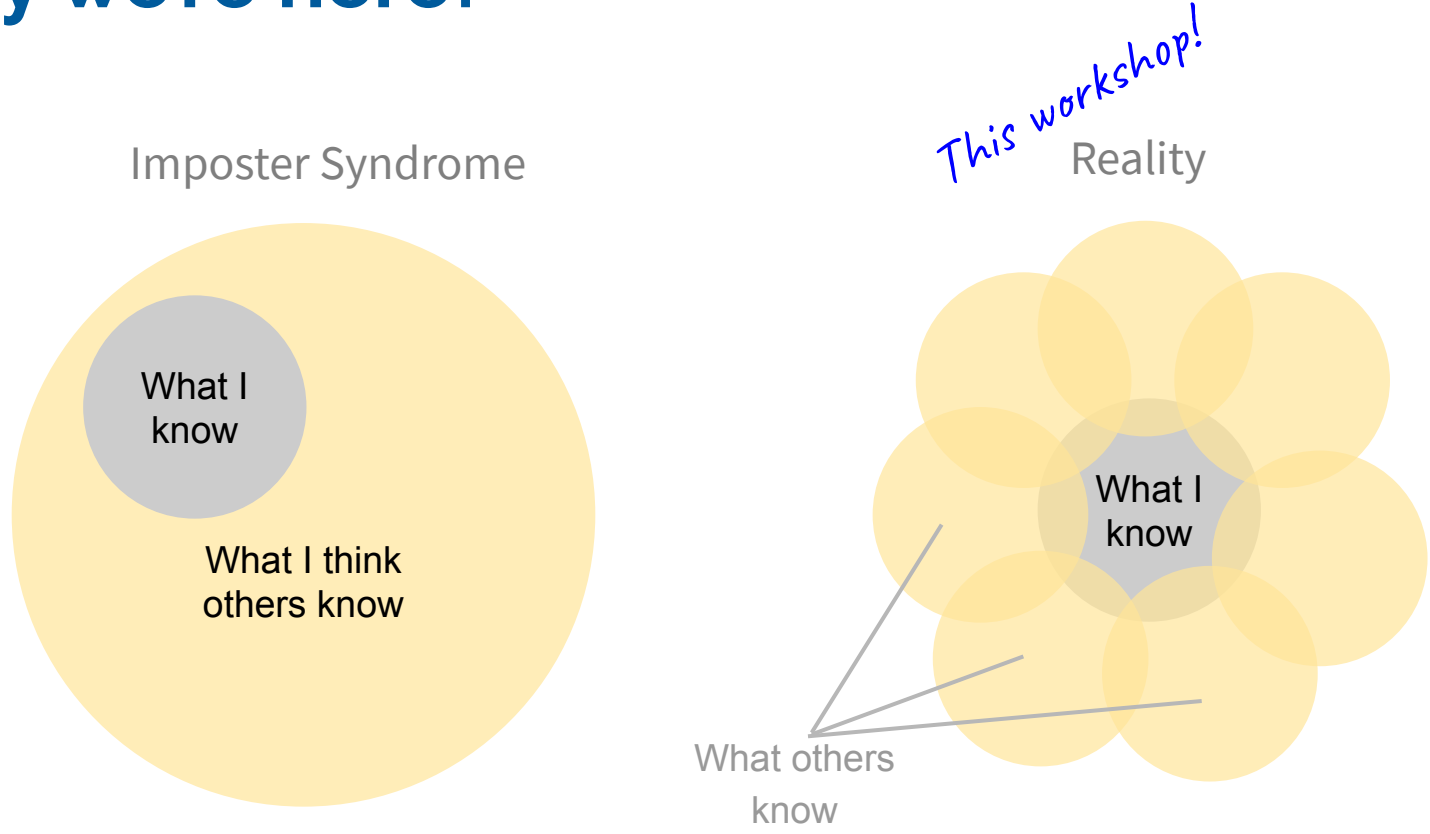
What is the purpose of this workshop?

1. Bring together machine learning library providers with Web platform practitioners to **enrich the Open Web Platform with better foundations for machine learning**

2. The secondary goals of the workshop are as follows:

   - Understand how machine learning fits into the Web stack,
   - Understand how in-browser machine learning fits into the machine learning ecosystem,
   - Explore the impact of machine learning technologies on Web browsers and Web apps,
   - Evaluate the opportunities for Web standardization around machine learning APIs and formats

Why we're here?



Acknowledgements

Program Committee

Kelly Davis, **Anssi Kostiainen**, Göran Eriksson, Dominique Hazaël-Massieux, Ningxin Hu, Dean Jackson, Sangwhan Moon, Roy Ran, Georg Rehm, Amy Siu, Nikhil Thorat

Speakers

Philip Laszkowicz, Bernard Aboba, Cormac Brick, Jason Mayes, Sangwhan Moon, Peter Hoddie, Dominique Hazaël-Massieux, François Daoust, Ningxin Hu, Jonathan Bingham, Mehmet Oguz Derin, Chai Chaoweeraprasit, Miao Wang, Jeff Hammond, Yakun Huang, Xiuquan Qiao, Wolfgang Maß, Mingqiu Sun, Andrew Brown, Ann Yuan, Emma Ning, Ping Wu, Yining Shi, Wenhe Li, Oleksandr Paraska, Piotr Migdal, Bartłomiej Olechno, Josh Meyer, Lindy Rauchenstein, Jutta Treviranus, John Rochford, Lisa Seeman, Joshue O'Connor, Tero Parviainen, Kelly Davis, Ryuichi Tanimoto, Nikolay Bogoychev, Anita Chen, Zelun Chen, Jean-Marc Valin, Louis McCallum

Sponsor

futurice

What topics will be covered?

Opportunities and Challenges

Today's live session focus

Opportunities and Challenges of Browser Based Machine Learning: Improving existing web platform capabilities, Extending beyond the browser, Considerations for creating and deploying models ...

Web Platform Foundations

Developer's Perspective

User's Perspective

Schedule

Opportunities and Challenges

Sep 16, 2020, 2pm UTC

Introduction

Opportunities and Challenges of Browser-Based ML

Developer's Perspective

Sep 23, 2020, 2pm UTC

ML Experiences on the Web: A **Developer**'s Perspective

Web Platform Foundations

Sep 22, 2020, 2pm UTC

Web Platform Foundations for ML

User's Perspective

Sep 29, 2020, 2pm UTC

ML Experiences on the Web: A **User**'s Perspective

Conclusions & Next Steps

What is W3C?

A voluntary standards consortium

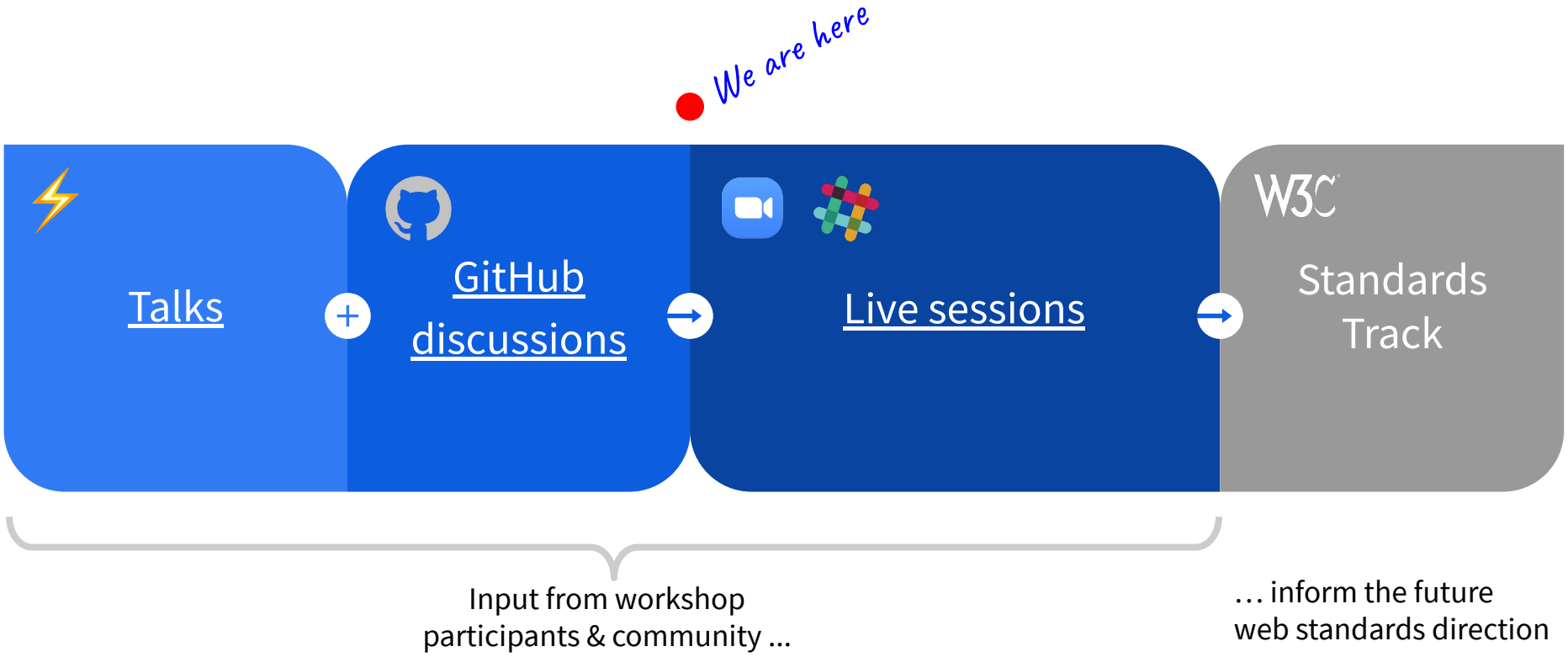
Convenes companies and communities to structure productive discussions around existing and emerging technologies

Royalty-Free patent framework

Focus primarily on client-side browser technologies

Develops work based on the priorities of W3C members and community

Contributing together



Live session practicalities

Zoom *We are here :)*

Mute by default, camera sharing encouraged
Use “raise hand” feature to be added to speaker queue, self-intro when speak up

Slack (https://w3ccommunity.slack.com/messages/machine-learning)

A complementary Slack chat channel for comments, questions

Meeting notes (https://bit.ly/webml-workshop-minutes)

Meeting notes captured in real-time collaboratively

Code of Ethics and Professional Conduct ensures that all voices can be heard

Opportunities and Challenges

✔ **Goal:**

Determine what are the unique opportunities of browser-based ML, what are the obstacles hindering adoption

Discussion topics

Improving existing web platform capabilities

- WebGPU fitness for ML frameworks
- Support for Float16 in JS & Wasm environments
- Memory copies
- Permission model for Machine Learning APIs

Extending beyond the browser

- Applicability to non-browser JS environments
- Targeting WASI-NN and WebNN together

Considerations for creating and deploying models

- Protecting ML models
- ML Model format
- In-browser training
- Training across devices

Improving existing web platform capabilities

Topic: WebGPU fitness for ML frameworks

Does WebGPU expose the right API surface needed to support ML frameworks interactions with GPUs?

Proposal: New WebGPU extensions for subgroups, cooperative matrix multiply.

https://github.com/w3c/machine-learning-workshop/issues/66

Topic: Support for Float16 in JS & Wasm environments

Lack of support for float16 in JS and Wasm environments problematic for quantized models.

Proposal: TBD

https://github.com/w3c/machine-learning-workshop/issues/64

Topic: Memory copies

Machine learning apps within the browser using the media pipeline trigger many more memory copies compared with native applications hindering performance.

Proposal: Introduce a more direct way to feed a video frame, possibly captured from a camera, to a ML model.

https://github.com/w3c/machine-learning-workshop/issues/93

Topic: Permission model for Machine Learning APIs

How to design a forward-looking permission model for ML APIs?

Proposal: TBD

https://github.com/w3c/machine-learning-workshop/issues/72

Extending beyond the browser

Topic: Applicability to non-browser JS environments

Pay attention to the applicability of the browser-targeted work to non-browser JS environments, in particular Node.js.

Proposal: Extend W3C coordination to TC53 and non-browser projects.

https://github.com/w3c/machine-learning-workshop/issues/62

Thank You!

End of Live Session #1

Next up:

September 22, 2020, 2pm UTC

Web Platform Foundations

Understand how machine learning fits into the Web technology stack

W3C Workshop on Web and Machine Learning

Live Session #2

What topics will be covered?

Opportunities and Challenges

Web Platform Foundations

Today's live session focus

Web Platform Foundations: Considerations for creating and deploying models, Extending the web foundations for ML

Developer's Perspective

User's Perspective

Live session practicalities (recap)

Zoom *We are here :)*

Mute by default, camera sharing encouraged
Use “raise hand” feature to be added to speaker queue, self-intro when speak up

Slack (https://w3ccommunity.slack.com/messages/machine-learning)

A complementary Slack chat channel for comments, questions

Meeting notes (https://bit.ly/webml-workshop-minutes)

Meeting notes captured in real-time collaboratively

Code of Ethics and Professional Conduct ensures that all voices can be heard

Web Platform Foundations

✔ **Goal:**

Understand how machine learning fits into the Web technology stack

Discussion topics

Considerations for creating and deploying models

- ML Model format
- Protecting ML models
- In-browser training
- Training across devices

Extending the web foundations for ML

- Targeting WASI-NN and WebNN together
- Heterogeneous parallel computing for the web

Considerations for creating and deploying models

Topic: ML model format

There is no standard format for packaging and shipping ML models, model formats evolve rapidly.

Proposal: Initially focus on defining a Web API for accelerating established reusable ML operations instead of standardizing a model format.

https://github.com/w3c/machine-learning-workshop/issues/74

Topic: Protecting ML models

Some ML providers need to ensure their ML models cannot be extracted from a browser app.

Proposal: Investigate existing access control mechanisms for video, learnings from 3D assets.

https://github.com/w3c/machine-learning-workshop/issues/67

Topic: In-browser training

The current in-browser efforts are focused on inference rather than training.

Proposal: Understand successful real-world usages (e.g. Teachable Machine) and target transfer learning as the initial training use case for related browser API work.

https://github.com/w3c/machine-learning-workshop/issues/82

Topic: Training across devices

Understand the role of edge computing in training and interactions with the web platform.

Proposal: Work with Web & Networks IG to understand edge computing use cases and ensure input from ML usages is considered.

https://github.com/w3c/machine-learning-workshop/issues/83

Extending the foundations

Topic: Targeting WASI-NN and WebNN together

Should libraries for browsers and/or Wasm execution environments be able to target WebNN and WASI-NN together?

Proposal: TBD

https://github.com/w3c/machine-learning-workshop/issues/96

Topic: Heterogeneous parallel computing for the web

How do the heterogeneous parallel computing abstractions fit in with the web platform?

Proposal: TBD

https://github.com/w3c/machine-learning-workshop/issues/68

Thank You!

End of Live Session #2

Next up:

September 23, 2020, 2pm UTC

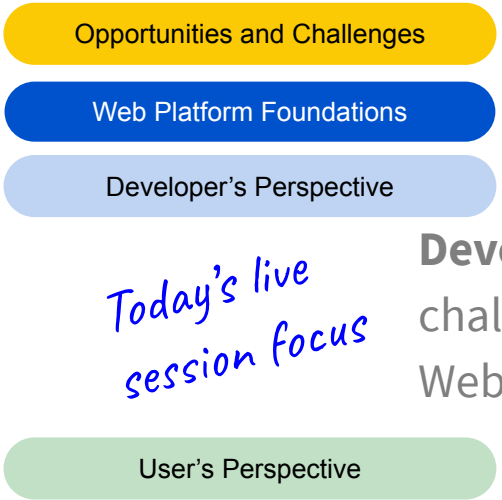
Developer's Perspective

Authoring ML experiences on the Web; challenges and opportunities of reusing existing ML models on the Web; known technical solutions, gaps

W3C Workshop on Web and Machine Learning

Live Session #3

What topics will be covered?



Developer's Perspective: Authoring ML experiences on the Web; challenges and opportunities of reusing existing ML models on the Web; known technical solutions, gaps

Live session practicalities (recap)

Zoom *We are here :)*

Mute by default, camera sharing encouraged
Use “raise hand” feature to be added to speaker queue, self-intro when speak up

Slack (https://w3ccommunity.slack.com/messages/machine-learning)

A complementary Slack chat channel for comments, questions

Meeting notes (https://bit.ly/webml-workshop-minutes)

Meeting notes captured in real-time collaboratively

Code of Ethics and Professional Conduct ensures that all voices can be heard

Developer's Perspective

✔ **Goal:**

Authoring ML experiences on the Web; challenges and opportunities of reusing existing ML models on the Web; known technical solutions, gaps

Developer's Perspective

Discussion topics

Applying web design principles to ML

- Progressive Enhancement / Graceful degradation
- Conformance testing of ML APIs for the Web

Improving web developer ergonomics

- JS Operator overloading for Machine Learning
- WebGL garbage collection
- Neural network-oriented graph database

Developing interactive web experiences with ML

- Action-Response Cycle bottlenecks in interactive music apps
- Noise suppression with DSP+DNN, WebNN and Web Audio API feature gaps

Applying web principles to ML

Topic: Progressive Enhancement / Graceful degradation

How to bring more ML features as optional improvements on more powerful devices and browsers without breaking web compatibility?

Proposal: TBD

https://github.com/w3c/machine-learning-workshop/issues/68

Topic: Conformance testing of ML APIs for the Web

Robust conformance testing is a cornerstone of the interoperable web platform, how to scale that to the ML APIs and formats?

Proposal: TBD

https://github.com/w3c/machine-learning-workshop/issues/80

Improving web developer ergonomics

Topic: JS Operator overloading for Machine Learning

Limitations in ECMAScript expressiveness impose ergonomics limitations for JS APIs on the web platform e.g. in vector matrix or tensor operations.

Proposal: TBD

https://github.com/w3c/machine-learning-workshop/issues/73

Topic: WebGL garbage collection

Garbage collection in the WebGL API affects multiple ML libraries through side effects.

Proposal: Identify any improvements in graphics APIs to alleviate the GC issue, ensure purpose-built APIs designed around computational graph abstraction (e.g. WebNN) optimize GC from library usage perspective.

https://github.com/w3c/machine-learning-workshop/issues/63

Topic: Neural network-oriented graph database

Understand model storage issues on the client, research the feasibility of a neural network-oriented graph database for the web.

Proposal: TBD

https://github.com/w3c/machine-learning-workshop/issues/102

Developing interactive web experiences with ML

Developer's Perspective

Topic: Action-Response Cycle bottlenecks in interactive music apps

Action-Response Cycle in interactive (music) apps must execute within 20 ms. Today, web developers need to do some API gymnastics to meet the requirement.

Proposal: Investigate inference in AudioWorklet context and media integration e.g. fast streaming inputs from MediaStream.

https://github.com/w3c/machine-learning-workshop/issues/97

Developer's Perspective

Topic: Noise suppression with DSP+DNN, WebNN and Web Audio API feature gaps

What areas needs work on the web platform to ensure noise suppression models perform? The need for primitives like Basic Linear Algebra Subprograms, Web Audio API enhancements to allow better analysis of waveforms?

Proposal: TBD

https://github.com/w3c/machine-learning-workshop/issues/100

Thank You!

End of Live Session #3

Next up:

September 29, 2020, 2pm UTC

User's Perspective

Goal: Web & ML for all: education, learning, accessibility, cross-industry experiences, cross-disciplinary ML: music, art, and media meet ML; Share learnings and best practices across industries

W3C Workshop on Web and Machine Learning

Live Session #4

Live session practicalities (recap)

Zoom *We are here :)*

Mute by default, camera sharing encouraged
Use “raise hand” feature to be added to speaker queue, self-intro when speak up

Slack (https://w3ccommunity.slack.com/messages/machine-learning)

A complementary Slack chat channel for comments, questions

Meeting notes (https://bit.ly/webml-workshop-minutes)

Meeting notes captured in real-time collaboratively

Code of Ethics and Professional Conduct ensures that all voices can be heard

What topics will be covered?

Opportunities and Challenges

Web Platform Foundations

Developer's Perspective

User's Perspective

User's Perspective: Web & ML for all: education, learning, accessibility, cross-industry experiences, cross-disciplinary ML: music, art, and media meet ML; Share learnings and best practices across industries

+

Conclusions & Next Steps

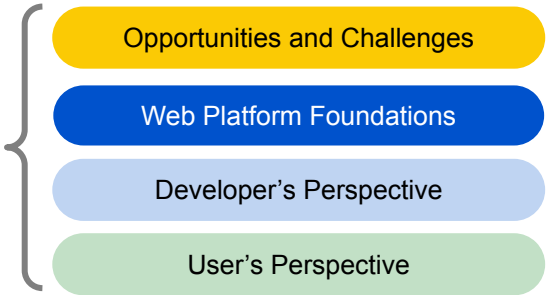
Discussion topics

Web & ML for all User's Perspective

- Bias and model transparency
- Speech recognition privacy issues and solutions
- Designing privacy-preserving ML APIs
- Building an extensible web platform for ML, one abstraction at a time

Conclusions & Next Steps

- What we have learned
- What we still need to figure out
- Next steps in incubation
- Next steps in standardization



Web & ML for all

User's Perspective

Topic: Bias and model transparency

Model bias and lack of ML model transparency impact minorities and underrepresented groups, could the Web help mitigate this issue by providing a browser-assisted mechanism to detail a ML model's limitations and performance characteristics?

Proposal: Explore role of machine-readable Model Cards to bring more transparency.

https://github.com/w3c/machine-learning-workshop/issues/108

User's Perspective

Topic: Speech recognition privacy issues and solutions

Standardization of the Web Speech API and its speech recognition part has been challenging due to privacy issues. What obstacles could be lifted to help make speech recognition an ubiquitous interoperable web capability?

Proposal: Find champion to bring Web Speech API to standardization.

https://github.com/w3c/machine-learning-workshop/issues/99

User's Perspective

Topic: Designing privacy-preserving ML APIs

We build the web platform with responsibility to our global user base, how to ensure the tight feedback loop and productive joint effort between ML ecosystem and privacy experts?

Proposal: Organize early review of WebNN API by Privacy Interest Group.

https://github.com/w3c/machine-learning-workshop/issues/90

User's Perspective

Topic: Building an extensible web platform for ML, one abstraction at a time

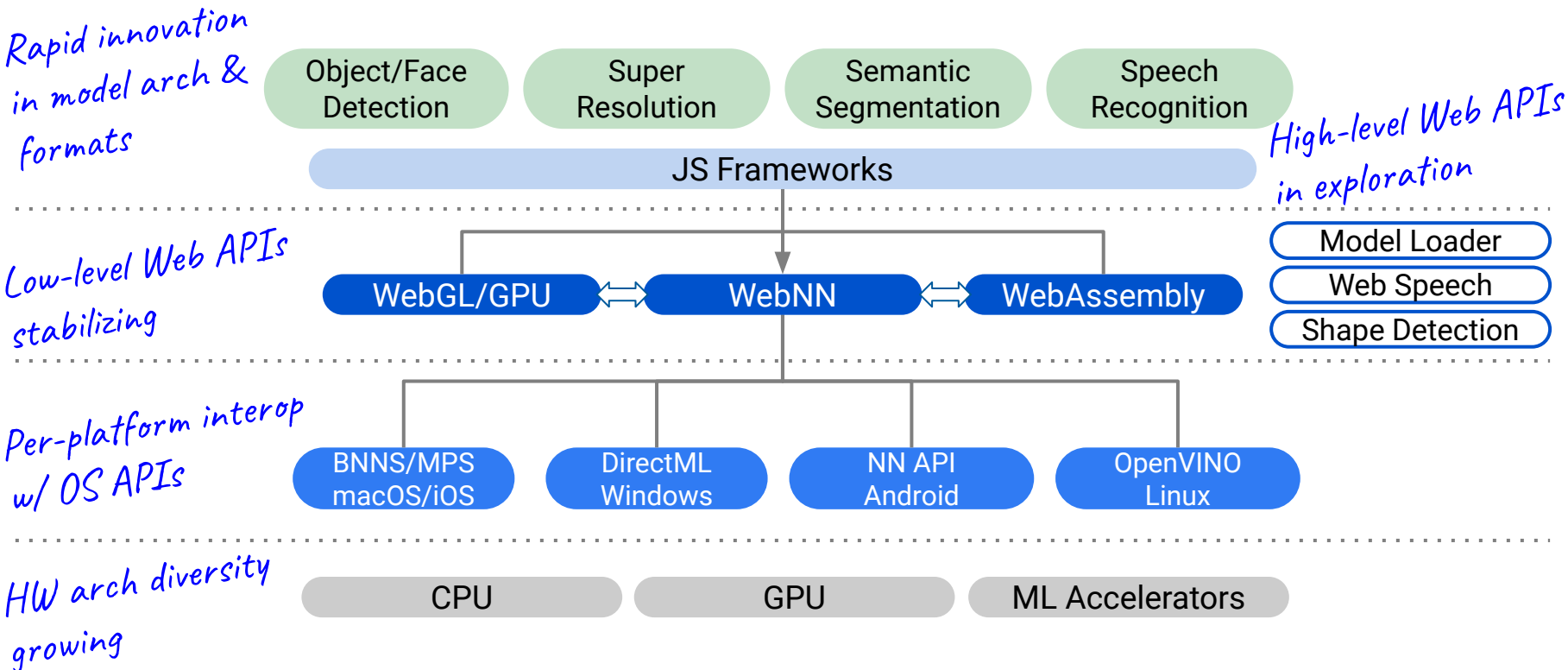
Are we in agreement that advancing with standardization of low-level capabilities e.g. WebNN API is the pragmatic first step?

Proposal: Propose a charter for standardizing WebNN API.

https://github.com/w3c/machine-learning-workshop/issues/109

Conclusions & Next Steps

A look at the Web & ML opportunity space



What we have learned

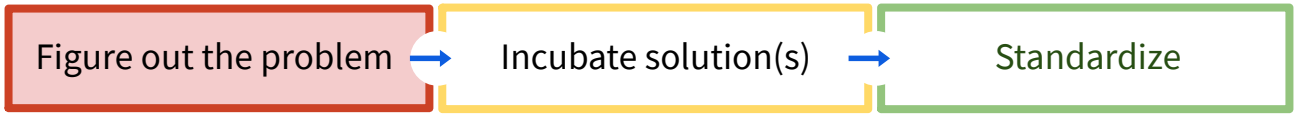
- Browser-based ML inference has a key role to play to help **privacy-friendly and real-time compatible** usage of Machine Learning models
- A **graph-based API layered over OS APIs** (à la WebNN) is a key primitive for efficient ML inference in Web browsers
- ML model formats are evolving fast and are not ready for standardization; there may be room for a **format-agnostic model loader API** to cater for the lack of standard format
- Efficient ML processing of media requires improvements throughout the processing pipeline (incl e.g. in memory management)
- Layering conformance testing of ML Web APIs on top of existing tests of OS APIs

What we have learned (2)

- **JS and WebAssembly** needs some upgrades to cater optimally for ML in browsers
- While ML inference needs to be the priority, ML **training** needs to be on the radar (esp. in the context of federated learning)
- Scaling up ML via browsers creates **risks of scaling up bias issues** linked to ML training

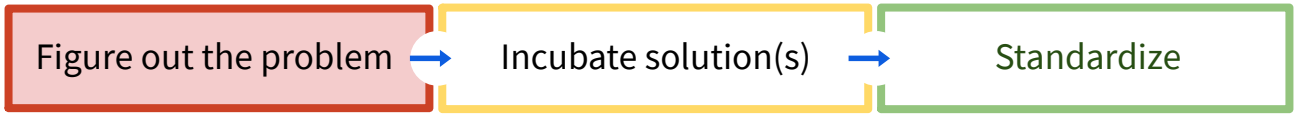
What we still need to figure out

- How to address the need to protect ML models while preserving user's privacy and security?
- How to manage access to compute-intensive APIs (incl ML APIs)?
- How much the in-browser APIs can and should be adapted to non-browser environments?
- How to integrate existing or emerging high-level ML-based APIs (speech & object recognition) with the lower level ones (WebNN, Model Loader)?
- How to scale up awareness of bias review and bias correction for ML model adopters?
- How important is it for ML that WebGPU provides ML-useful optimizations?



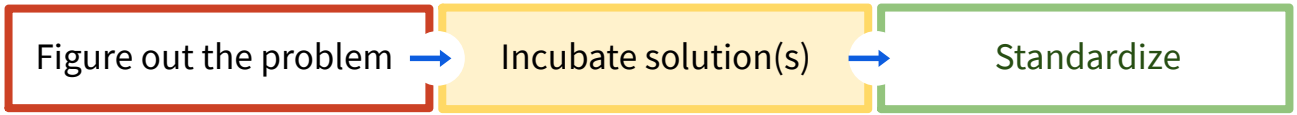
What we still need to figure out (2)

- What introspection data on models is needed to cater for progressive enhancements approaches?
- What architecture do we need to distribute ML tasks (inference, training) across multiple devices (incl edge computing)?
- Does ML model storage require any specific browser adaptation or would the File System Access API cover all that is needed?



Next steps in incubation

- Validate that Model Loader API can support interoperable deployment of ML
- Explore gating access to compute intensive APIs (via the W3C TAG?)
- Explore optimizing memory copy across media pipeline (Media & Entertainment IG?)
- Explore machine-readable Model Cards (Web ML CG proposal?)



Next steps in standardization

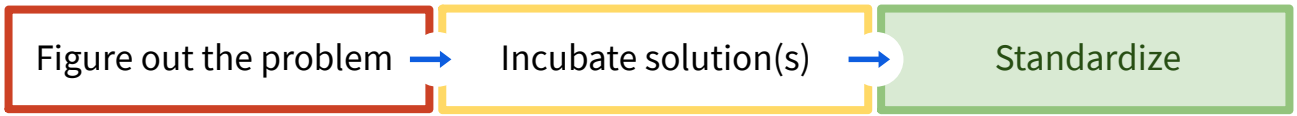
Bring Web Neural Network API to W3C standardization

➔ Early draft of a charter for a new W3C Working Group

Reinvigorate efforts for JS operators overloading

➔ Liaise with ECMA TC39 to raise priority

Reinvigorate efforts around Float16 support in JS



Where to follow up

- The GitHub repo of the workshop is open and active for the foreseeable future
- So will our Slack channel
- Input to proposed Working Group charter
- W3C TPAC Breakout on Memory Copies
- Web Machine Learning Community Group (open to anyone to join) - proposals repo

Upcoming feedback survey - let dom@w3.org know if/when we should do this again!

Acknowledgements

Program Committee **Kelly Davis**, **Anssi Kostiainen**, Göran Eriksson, Dominique Hazaël-Massieux, Ningxin Hu, Dean Jackson, Sangwhan Moon, Roy Ran, Georg Rehm, Amy Siu, Nikhil Thorat

Speakers Philip Laszkowicz, Bernard Aboba, Cormac Brick, Jason Mayes, Sangwhan Moon, Peter Hoddie, Dominique Hazaël-Massieux, François Daoust, Ningxin Hu, Jonathan Bingham, Mehmet Oguz Derin, Chai Chaoweeraprasit, Miao Wang, Jeff Hammond, Yakun Huang, Xiuquan Qiao, Wolfgang Maß, Mingqiu Sun, Andrew Brown, Ann Yuan, Emma Ning, Ping Wu, Yining Shi, Wenhe Li, Oleksandr Paraska, Piotr Migdal, Bartłomiej Olechno, Josh Meyer, Lindy Rauchenstein, Jutta Treviranus, John Rochford, Lisa Seeman, Joshue O'Connor, Tero Parviainen, Kelly Davis, Ryuichi Tanimoto, Nikolay Bogoychev, Anita Chen, Zelun Chen, Jean-Marc Valin, Louis McCallum

Sponsor futurice