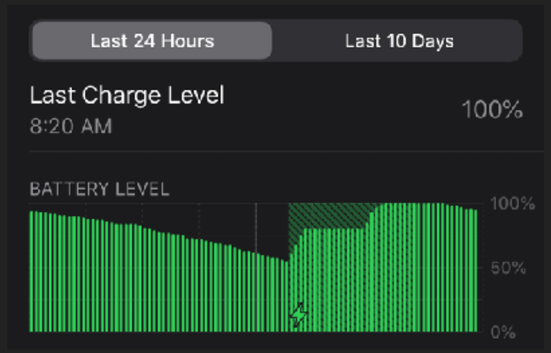
Interactive Scalable Interfaces for Machine Learning Interpretability

Fred Hohman
@fredhohman

Georgia Tech

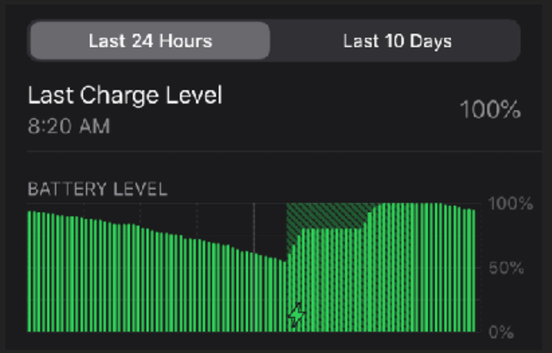
Follow along at fredhohman.com/talk

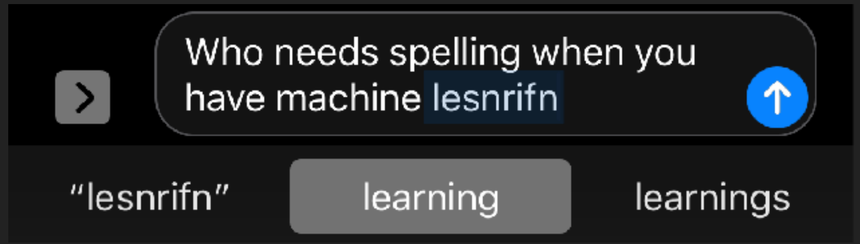
Optimized phone battery charging



Optimized phone battery charging

Spelling & text prediction

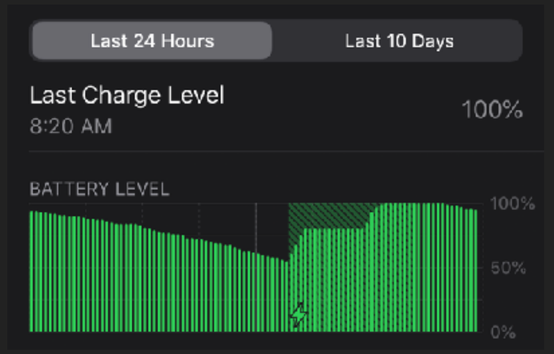


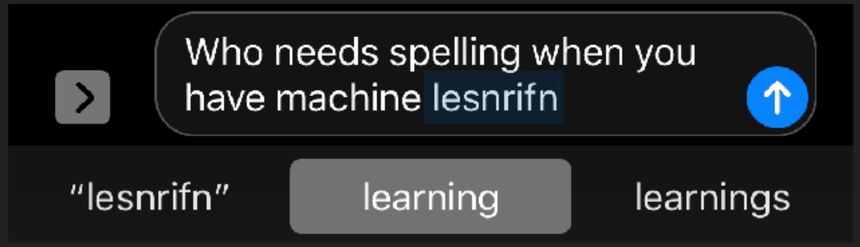


Optimized phone battery charging

Spelling & text prediction

Face unlock



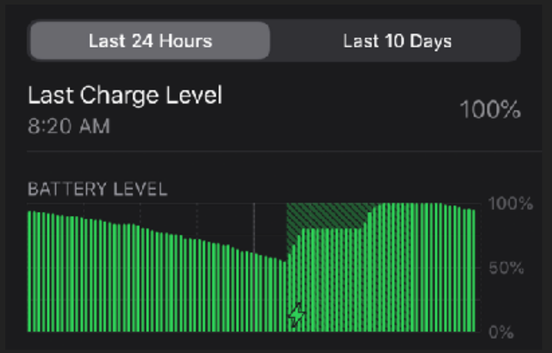


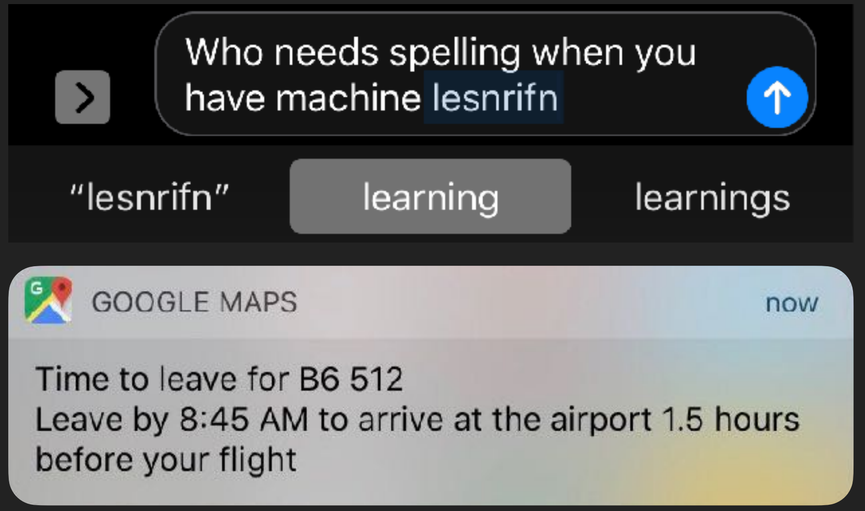
Optimized phone battery charging

Spelling & text prediction

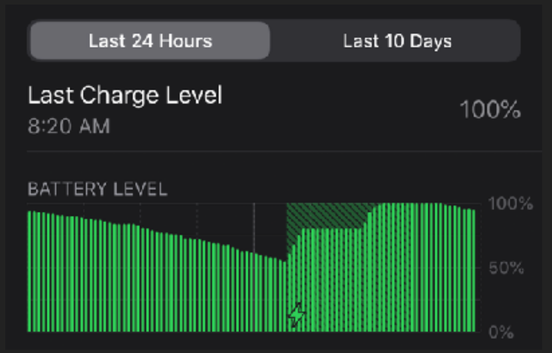
Face unlock

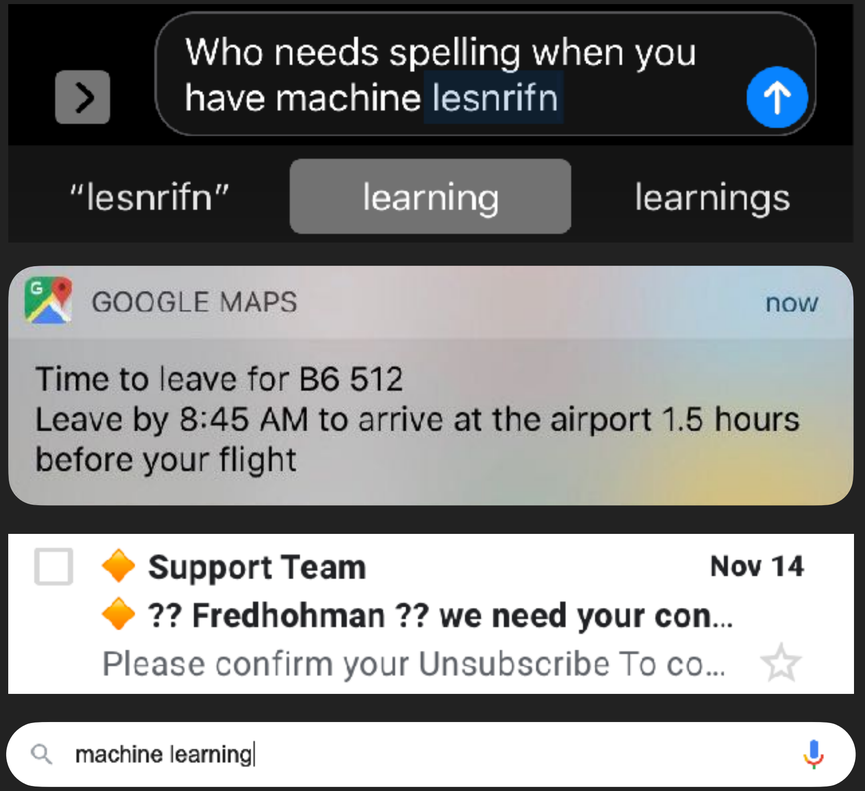
Traffic & commute notifications

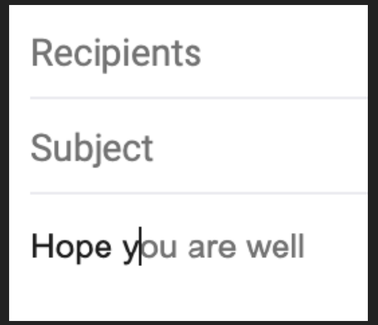




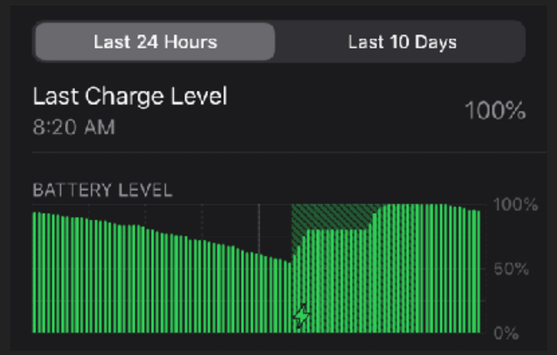
- Optimized phone battery charging
- Spelling & text prediction
- Face unlock
- Traffic & commute notifications
- Email spam filters
- Email “smart” reply
- Search results

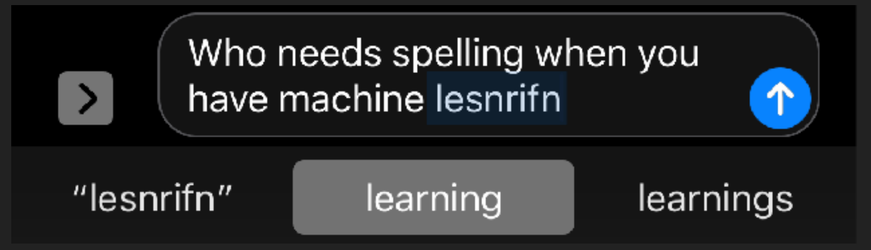


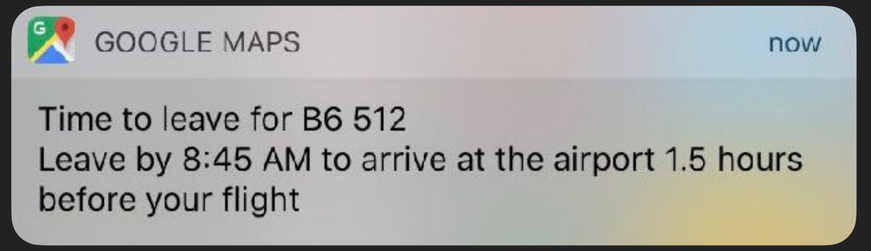


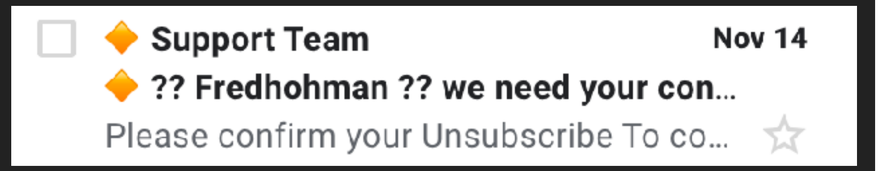


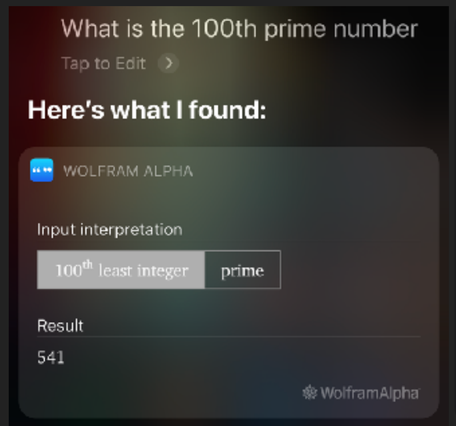
- Optimized phone battery charging
- Spelling & text prediction
- Face unlock
- Traffic & commute notifications
- Email spam filters
- Email “smart” reply
- Search results
- Trending news recommendations
- Social media newsfeeds
- Mobile computational photography
- Automatic voice assistants
- Voice dictation
- Health informatics
- Personalized music & radio
- Entertainment recommendations
- Personalized ads

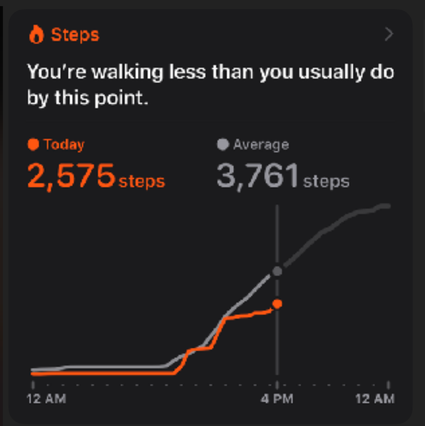


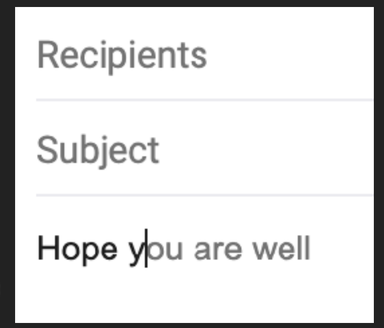


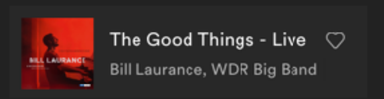












Optimized phone battery charging → Phone not fully charged

Spelling & text prediction

Face unlock

Traffic & commute notifications

Email spam filters

Email “smart” reply

Search results

Trending news recommendations

Social media newsfeeds

Mobile computational photography

Automatic voice assistants

Voice dictation

Health informatics

Personalized music & radio

Entertainment recommendations

Personalized ads

Optimized phone battery charging → Phone not fully charged

Spelling & text prediction → Send wrong message

Face unlock

Traffic & commute notifications

Email spam filters

Email “smart” reply

Search results

Trending news recommendations

Social media newsfeeds

Mobile computational photography

Automatic voice assistants

Voice dictation

Health informatics

Personalized music & radio

Entertainment recommendations

Personalized ads

Optimized phone battery charging → Phone not fully charged

Spelling & text prediction → Send wrong message

Face unlock

Traffic & commute notifications

Email spam filters

Email “smart” reply

Search results → Spread misinformation

Trending news recommendations

Social media newsfeeds

Mobile computational photography

Automatic voice assistants

Voice dictation

Health informatics

Personalized music & radio

Entertainment recommendations

Personalized ads

Optimized phone battery charging → Phone not fully charged

Spelling & text prediction → Send wrong message

Face unlock

Traffic & commute notifications

Email spam filters

Email “smart” reply

Search results → Spread misinformation

Trending news recommendations

Social media newsfeeds → Pigeonhole beliefs

Mobile computational photography

Automatic voice assistants

Voice dictation

Health informatics

Personalized music & radio

Entertainment recommendations

Personalized ads

Optimized phone battery charging → Phone not fully charged

Spelling & text prediction → Send wrong message

Face unlock

Traffic & commute notifications

Email spam filters

Email “smart” reply

Search results → Spread misinformation

Trending news recommendations

Social media newsfeeds → Pigeonhole beliefs

Mobile computational photography

Automatic voice assistants

Voice dictation

Health informatics → Life threatening

Personalized music & radio

Entertainment recommendations

Personalized ads

Optimized phone battery charging → Phone not fully charged

Spelling & text prediction → Send wrong message

Face unlock

Traffic & commute notifications

Email spam filters

Email “smart” reply

Search results → Spread misinformation

Trending news recommendations

Social media newsfeeds → Pigeonhole beliefs

Mobile computational photography

Automatic voice assistants

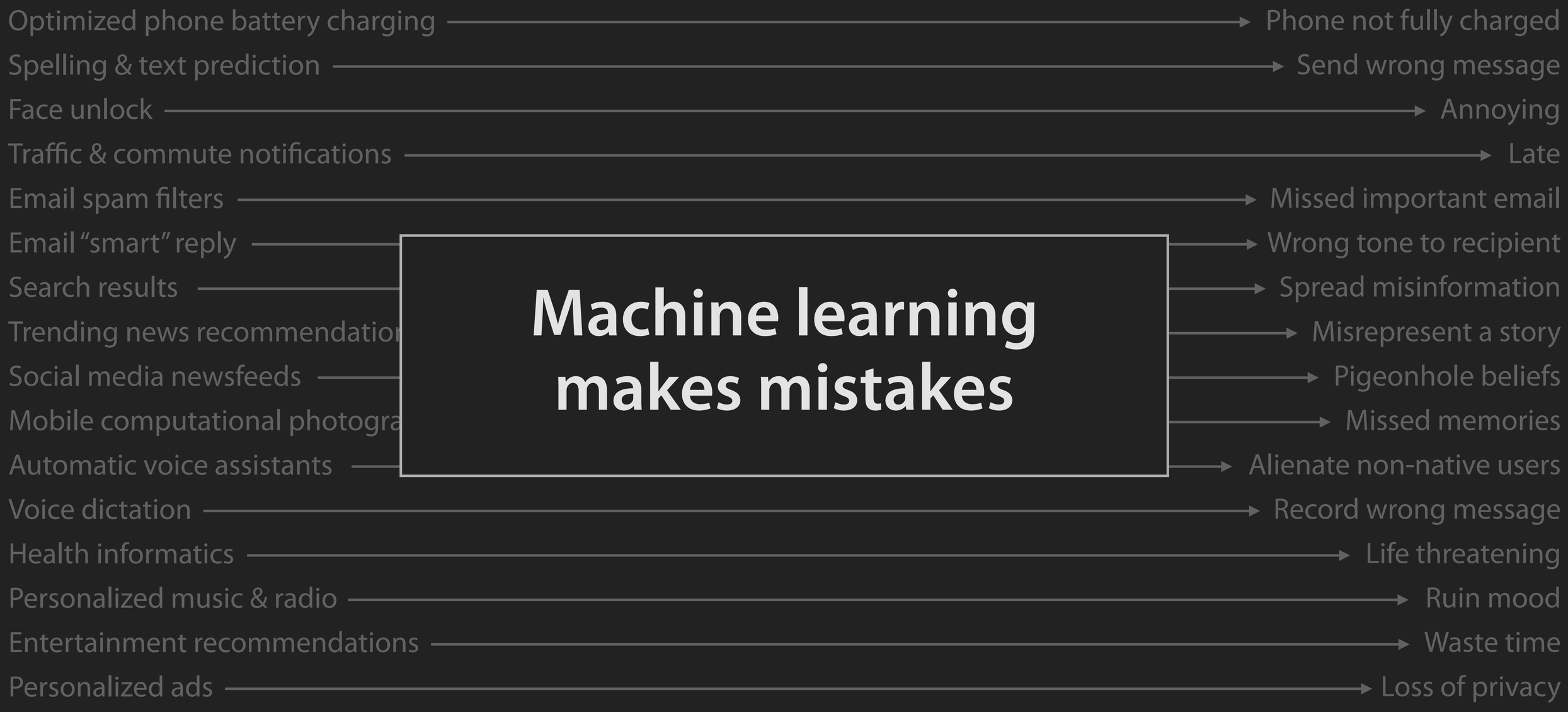
Voice dictation

Health informatics → Life threatening

Personalized music & radio

Entertainment recommendations

Personalized ads → Loss of privacy


Optimized phone battery charging
Phone not fully charged
Spelling & text prediction
Send wrong message
Face unlock
Annoying
Traffic & commute notifications
Late
Email spam filters
Missed important email
Email “smart” reply
Wrong tone to recipient
Search results
Spread misinformation
Trending news recommendatio
Misrepresent a story
Social media newsfeeds
Pigeonhole beliefs
Mobile computational photogra
Missed memories
Automatic voice assistants
Alienate non-native users
Voice dictation
Record wrong message
Health informatics
Life threatening
Personalized music & radio
Ruin mood
Entertainment recommendations
Waste time
Personalized ads
Loss of privacy
Machine learning makes mistakes


Google
ai is
ai is dangerous
ai is the new electricity
ai is taking over
ai is the future
ai is a crapshoot
ai is good
ai is overhyped
ai is not dangerous
ai is eating software
ai is everywhere
Google Search
I'm Feeling Lucky
Report inappropriate predictions

Data

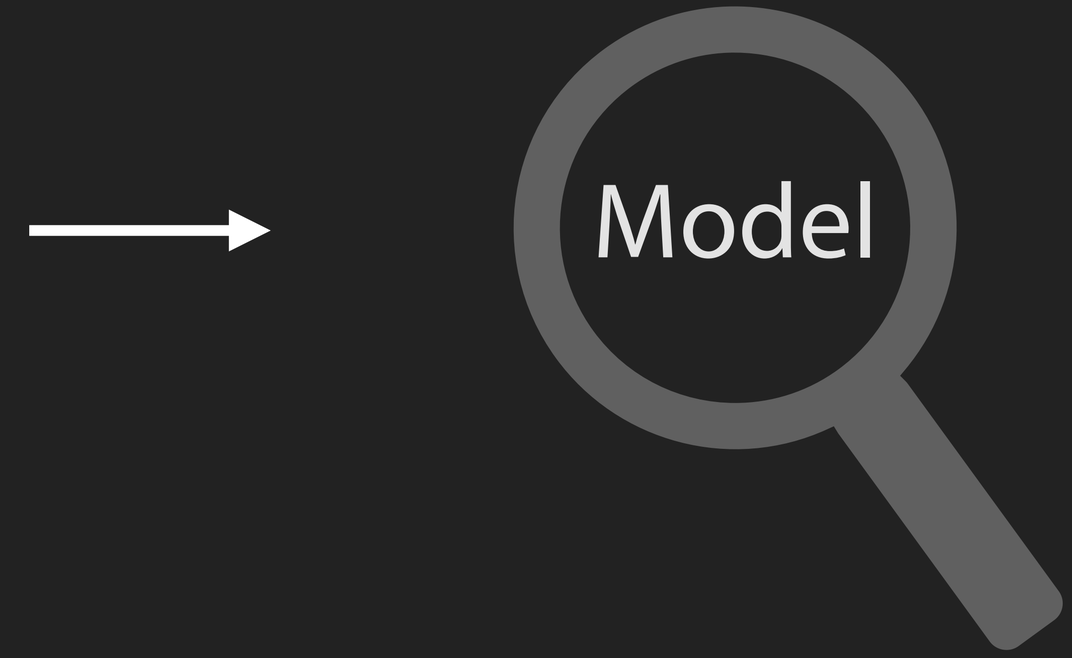
Machine Learning

Model

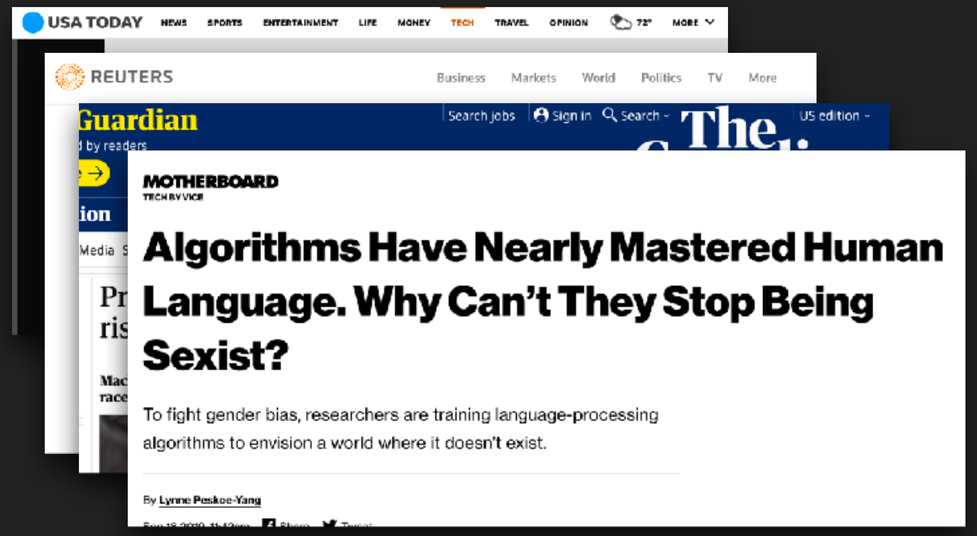
Interpretability

What has a model learned?
Can you explain its predictions?
Do you understand its behavior?

Data



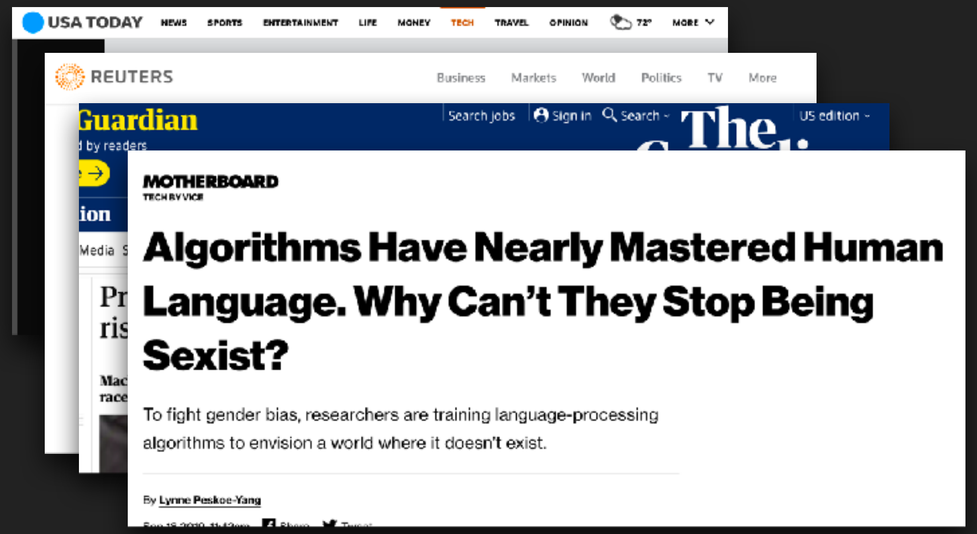
Why Interpretability?



Fairness & GDPR

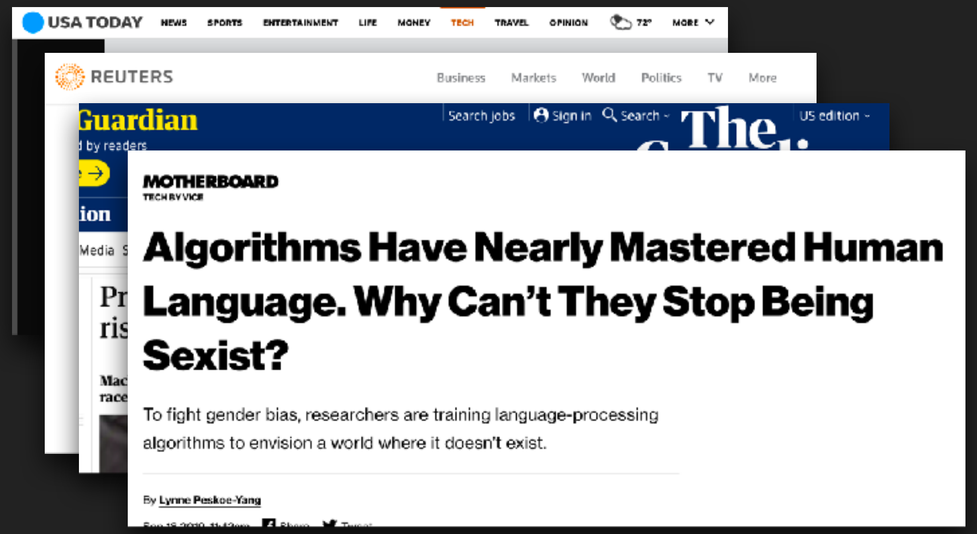
Why Interpretability?

Why Interpretability?



Fairness & GDPR

Science & Discovery



Fairness & GDPR

Science & Discovery

Why Interpretability?

Robustness
To data drift & adversarial attacks

Facilitating User Control
With personalized agents

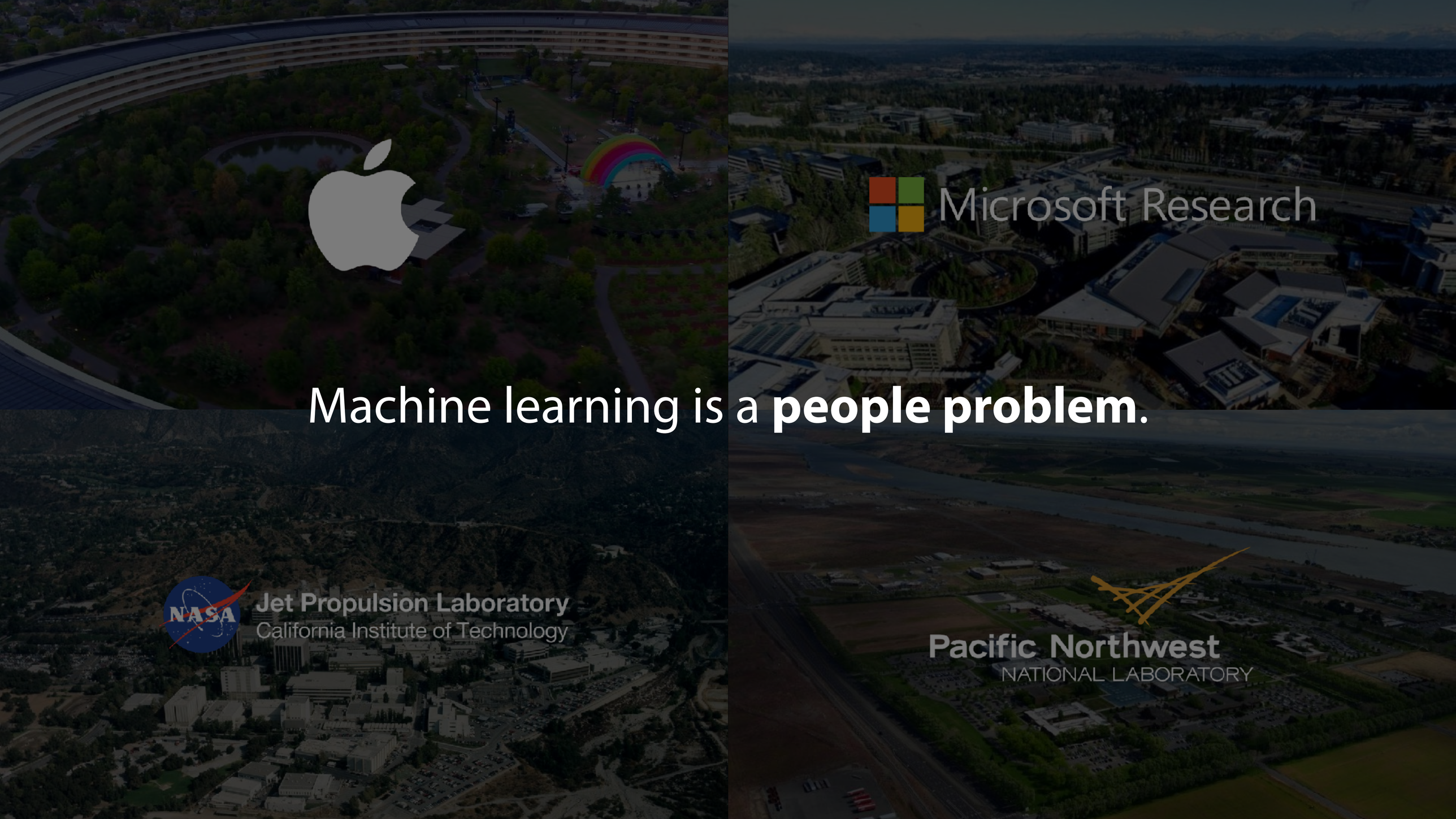
User Acceptance
People happier with explanations

Mismatched Objectives
Is AI right for the right reasons?


Microsoft Research

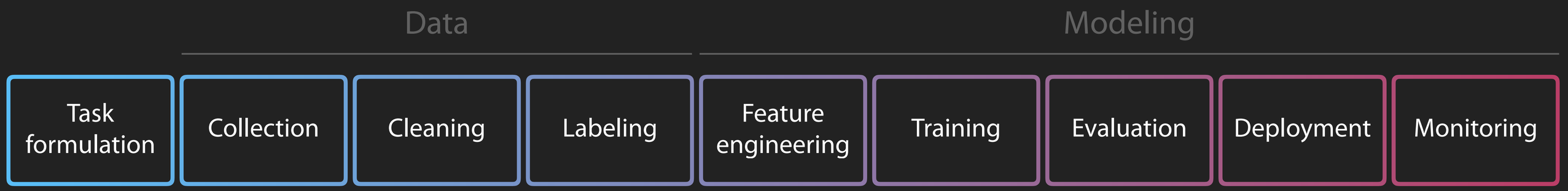

Jet Propulsion Laboratory
California Institute of Technology


Pacific Northwest
NATIONAL LABORATORY



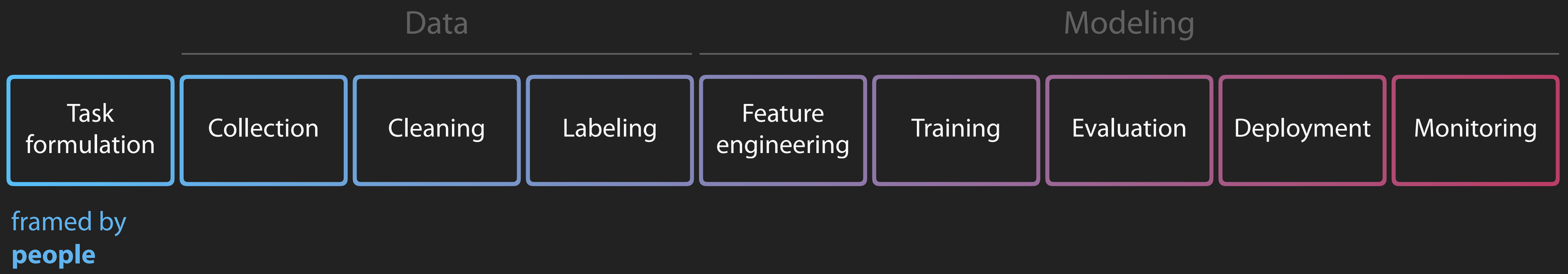
Machine learning is a **people problem**.

Machine learning is a **people problem**.



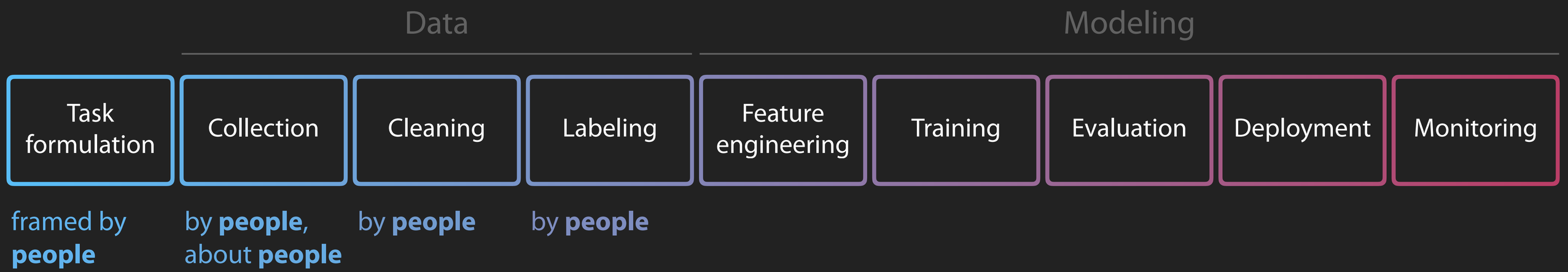
Combined and redesigned from Amershi et al. ICSE 2019; Jeffery Bigham & Joseph Seering, Human-AI course, 2018.

Machine learning is a **people problem**.



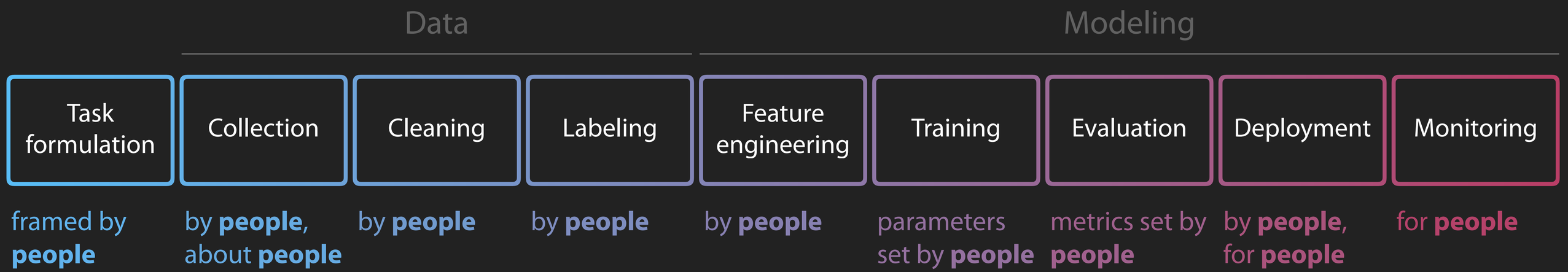
Combined and redesigned from Amershi et al. ICSE 2019; Jeffery Bigham & Joseph Seering, Human-AI course, 2018.

Machine learning is a **people problem**.



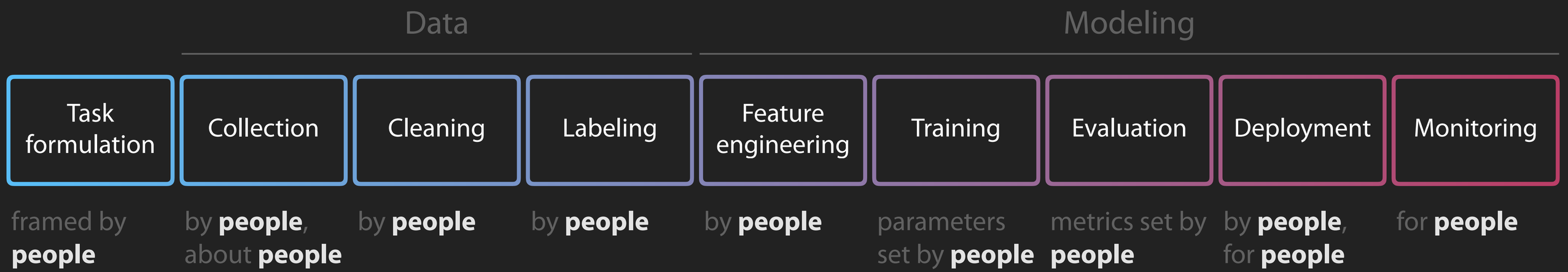
Combined and redesigned from Amershi et al. ICSE 2019; Jeffery Bigham & Joseph Seering, Human-AI course, 2018.

Machine learning is a **people problem**.



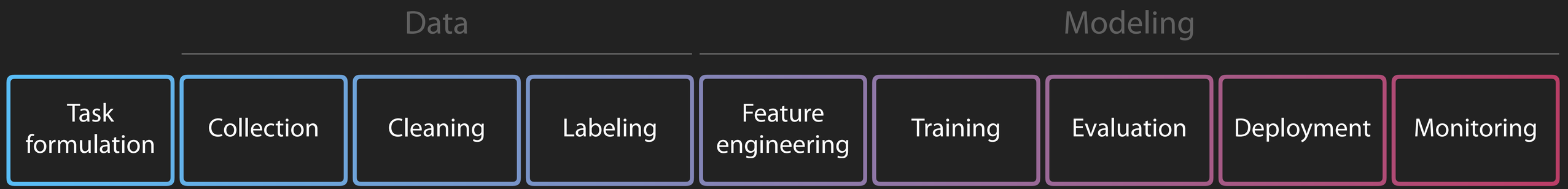
Combined and redesigned from Amershi et al. ICSE 2019; Jeffery Bigham & Joseph Seering, Human-AI course, 2018.

Machine learning is a **people problem**.



Combined and redesigned from Amershi et al. ICSE 2019; Jeffery Bigham & Joseph Seering, Human-AI course, 2018.

Machine learning is a **people problem**.



Human-centered approach

Understand what **people** need to
design and develop the right tools for interpretability.

Combined and redesigned from Amershi et al. ICSE 2019; Jeffery Bigham & Joseph Seering, Human-AI course, 2018.

Research Mission

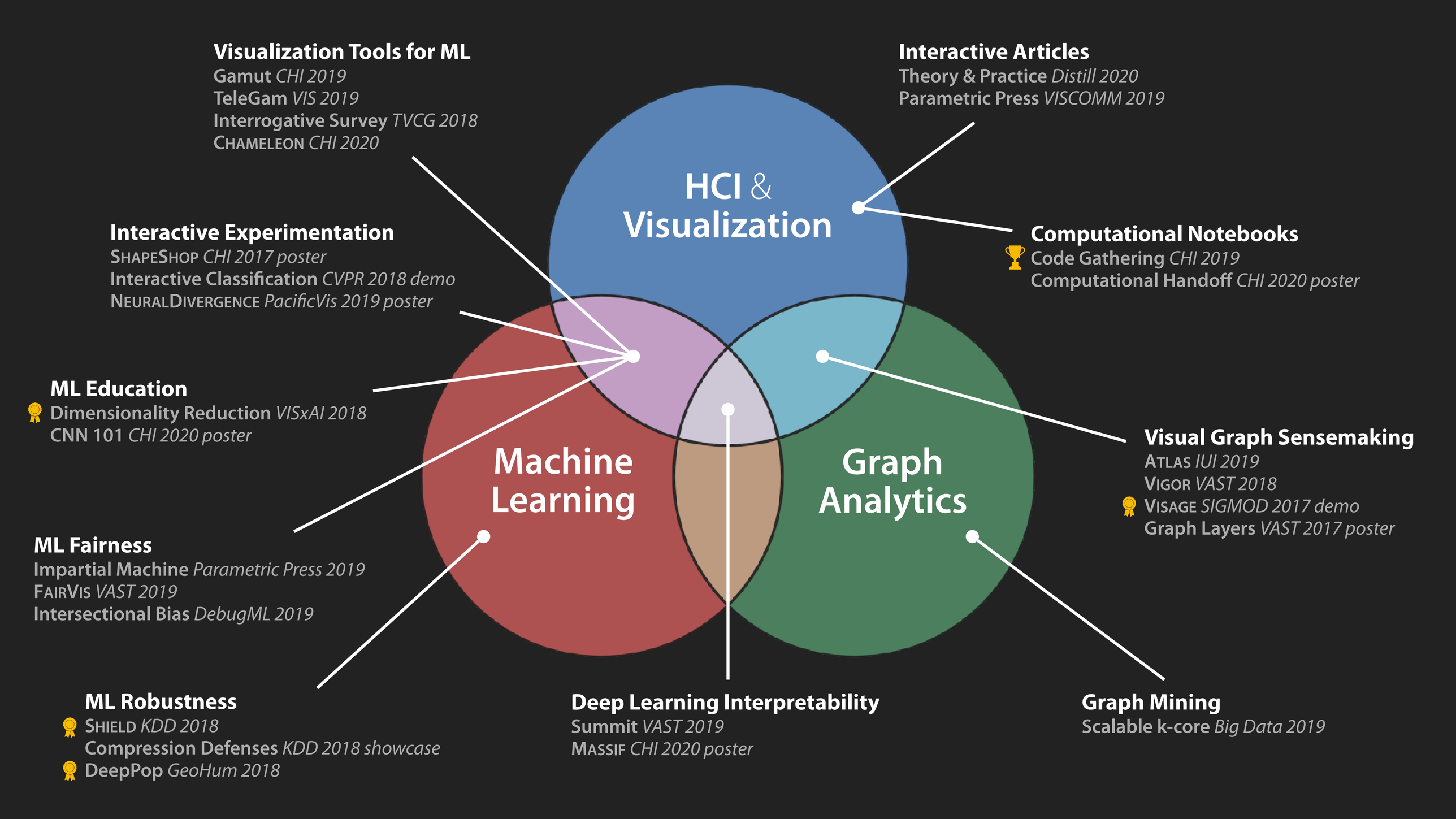
Enable machine learning interpretability at scale and for everyone.

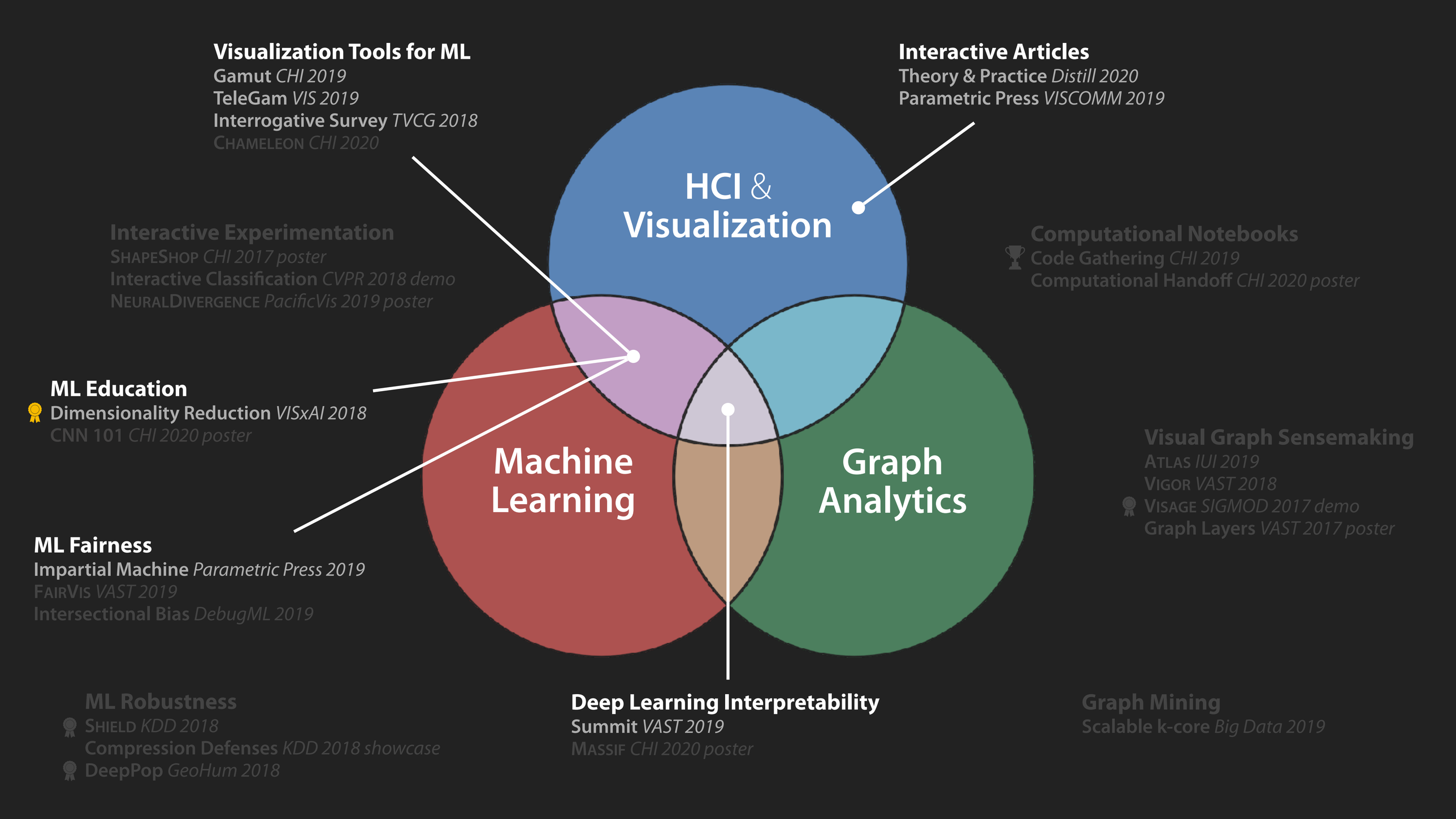
Research Mission

Enable machine learning interpretability at scale and for everyone.

Research Mission

Enable machine learning interpretability at scale and for everyone,
by *designing and developing interactive interfaces* to
help people confidently understand data-driven systems.


Visualization Tools for ML
Gamut CHI 2019
TeleGam VIS 2019
Interrogative Survey TVCG 2018
CHAMELEON CHI 2020
Interactive Articles
Theory & Practice Distill 2020
Parametric Press VISCOMM 2019
HCI & Visualization
Interactive Experimentation
SHAPESHOP CHI 2017 poster
Interactive Classification CVPR 2018 demo
NEURALDIVERGENCE PacificVis 2019 poster
Computational Notebooks
Code Gathering CHI 2019
Computational Handoff CHI 2020 poster
ML Education
Dimensionality Reduction VISxAI 2018
CNN 101 CHI 2020 poster
Machine Learning
Graph Analytics
Visual Graph Sensemaking
ATLAS IUI 2019
VIGOR VAST 2018
VISAGE SIGMOD 2017 demo
Graph Layers VAST 2017 poster
ML Fairness
Impartial Machine Parametric Press 2019
FAIRVIS VAST 2019
Intersectional Bias DebugML 2019
ML Robustness
SHIELD KDD 2018
Compression Defenses KDD 2018 showcase
DeepPop GeoHum 2018
Deep Learning Interpretability
Summit VAST 2019
MASSIF CHI 2020 poster
Graph Mining
Scalable k-core Big Data 2019


Visualization Tools for ML
Gamut CHI 2019
TeleGam VIS 2019
Interrogative Survey TVCG 2018
CHAMELEON CHI 2020
Interactive Articles
Theory & Practice Distill 2020
Parametric Press VISCOMM 2019
HCI & Visualization
Interactive Experimentation
SHAPESHOP CHI 2017 poster
Interactive Classification CVPR 2018 demo
NEURALDIVERGENCE PacificVis 2019 poster
Computational Notebooks
Code Gathering CHI 2019
Computational Handoff CHI 2020 poster
ML Education
Dimensionality Reduction VISxAI 2018
CNN 101 CHI 2020 poster
Machine Learning
Graph Analytics
Visual Graph Sensemaking
ATLAS IUI 2019
VIGOR VAST 2018
VISAGE SIGMOD 2017 demo
Graph Layers VAST 2017 poster
ML Fairness
Impartial Machine Parametric Press 2019
FAIRVIS VAST 2019
Intersectional Bias DebugML 2019
ML Robustness
SHIELD KDD 2018
Compression Defenses KDD 2018 showcase
DeepPop GeoHum 2018
Deep Learning Interpretability
Summit VAST 2019
MASSIF CHI 2020 poster
Graph Mining
Scalable k-core Big Data 2019

Interactive Scalable Interfaces for Machine Learning Interpretability

PART I

Enable interpretability

GAMUT Operationalize interpretability *CHI 2019*

TELEGAM Vis + text for better explanations *VIS 2019*

PART II

Scale interpretability

Interrogative Survey Summarize interpretability vis *TVCG 2018*

SUMMIT Higher-level explanations for neural networks *VAST 2019*

PART III

Communicate interpretability

ML Literacy Interactive mediums & platforms *VISCOMM 2019, VISxAI 2018*

Interactive Articles Formalizing interactive communication *Distill 2020*

Interactive Scalable Interfaces for Machine Learning Interpretability

PART I
Enable interpretability

GAMUT Operationalize interpretability *CHI 2019*
TELEGAM Vis + text for better explanations *VIS 2019*

PART II
Scale interpretability

Interrogative Survey Summarize interpretability vis *TVCG 2018*
SUMMIT Higher-level explanations for neural networks *VAST 2019*

PART III
Communicate interpretability

ML Literacy Interactive mediums & platforms *VISCOMM 2019, VISxAI 2018*
Interactive Articles Formalizing interactive communication *Distill 2020*

What is Interpretability?

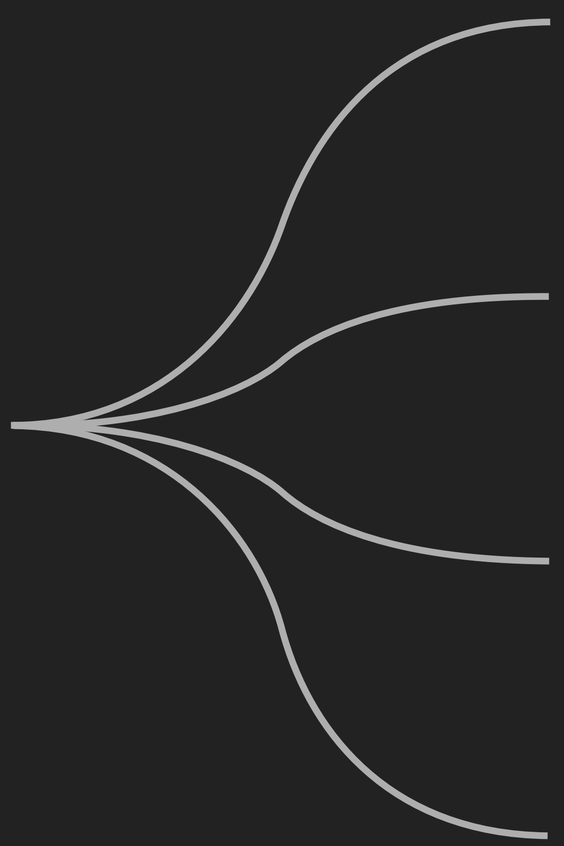
What is Interpretability?

Human understanding

of a system's…

What is Interpretability?

Human understanding
of a system's…



internals
e.g., components
[Gilpin, 2018]

operations
e.g., math
[Biran, 2017]

data mapping
e.g., input to output
[Montavon, 2017]

representation
in an explanation
[Ribeiro, 2016]

What is Interpretability?

Human understanding
of a system's…

No formal, agreed upon definition
[Lipton, 2016]

internals
e.g., components
[Gilpin, 2018]

operations
e.g., math
[Biran, 2017]

data mapping
e.g., input to output
[Montavon, 2017]

representation
in an explanation
[Ribeiro, 2016]

GAMUT

Understanding How Data Scientists Understand Machine Learning Models

CHI 2019

Fred Hohman
Georgia Tech

Andrew Head
UC Berkeley

Rob DeLine
Microsoft Research

Rich Caruana
Microsoft Research

Steven Drucker
Microsoft Research

GAMUT Contributions

Operationalize
interpretability

Design Probe
embodying
operationalization

Evaluation & Investigation
of probe & emerging
practice of interpretability
with real users

How to Operationalize Interpretability?

Formative research with professional data scientists @

- 4 senior ML researchers
- 5 ML practitioners

Prompt: *In a perfect world, given a machine learning model, what questions would you ask it to help you interpret both the model and its predictions?*

Explainable ML Interface Questions

Why this prediction?

What is the difference between these two?

What if I added…

What are similar instances?

Where is it wrong?

What is most important?

Explainable ML Interface Capabilities

C1 Why this prediction?
Local instance explanations

C2 What is the difference between these two?
Instance explanation comparisons

C3 What if I added…
Counterfactuals

C4 What are similar instances?
Nearest neighbors

C5 Where is it wrong?
Regions of error

C6 What is most important?
Feature importance

Explainable ML Interface Capabilities

C1 Why this prediction?
Local instance explanations

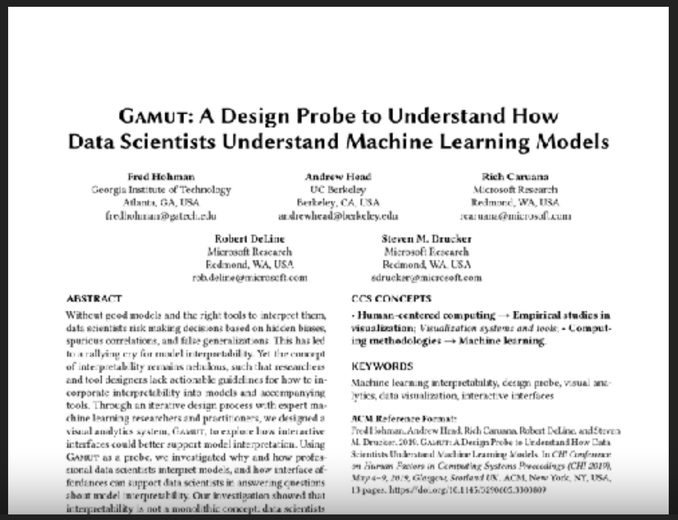
C2 What is the difference between these two?
Instance explanation comparisons

C3 What if I added…
Counterfactuals

C4 What are similar instances?
Nearest neighbors

C5 Where is it wrong?
Regions of error

C6 What is most important?
Feature importance



GAMUT: A Design Probe to Understand How Data Scientists Understand Machine Learning Models

Fred Hohman
Georgia Institute of Technology
Atlanta, GA, USA
fredhohman@gatech.edu

Andrew Head
UC Berkeley
Berkeley, CA, USA
andrewhead@berkeley.edu

Rich Caruana
Microsoft Research
Redmond, WA, USA
rcaruana@microsoft.com

Robert DeLine
Microsoft Research
Redmond, WA, USA
rob.deline@microsoft.com

Steven M. Drucker
Microsoft Research
Redmond, WA, USA
sdrucker@microsoft.com

ABSTRACT

Without good models and the right tools to interpret them, data scientists risk making decisions based on hidden biases, spurious correlations, and false generalizations. This has led to a rallying cry for model interpretability. Yet the concept of interpretability remains nebulous, such that researchers and tool designers lack actionable guidelines for how to incorporate interpretability into models and accompanying tools. Through an iterative design process with expert machine learning researchers and practitioners, we designed a visual analytics system, GAMUT, to explore how interactive interfaces could better support model interpretation. Using GAMUT as a probe, we investigated why and how profes-

CCS CONCEPTS

• Human-centered computing → Empirical studies in visualization; Visualization systems and tools; • Computing methodologies → Machine learning;

KEYWORDS

Machine learning interpretability, design probe, visual analytics, data visualization, interactive interfaces

ACM Reference Format:

Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. GAMUT: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300809

Definitions + examples in the paper!

How to Test Our Capabilities?

[Hutchinson, 2003]

Design probe: *"instrument that is deployed to find out about the unknown—returning with useful or interesting data."*

Balance of:

- ***Design:*** inspire reflection on emerging tech
- ***Social science:*** appreciate needs of users
- ***Engineering:*** field-testing prototypes

House 550

$190,606

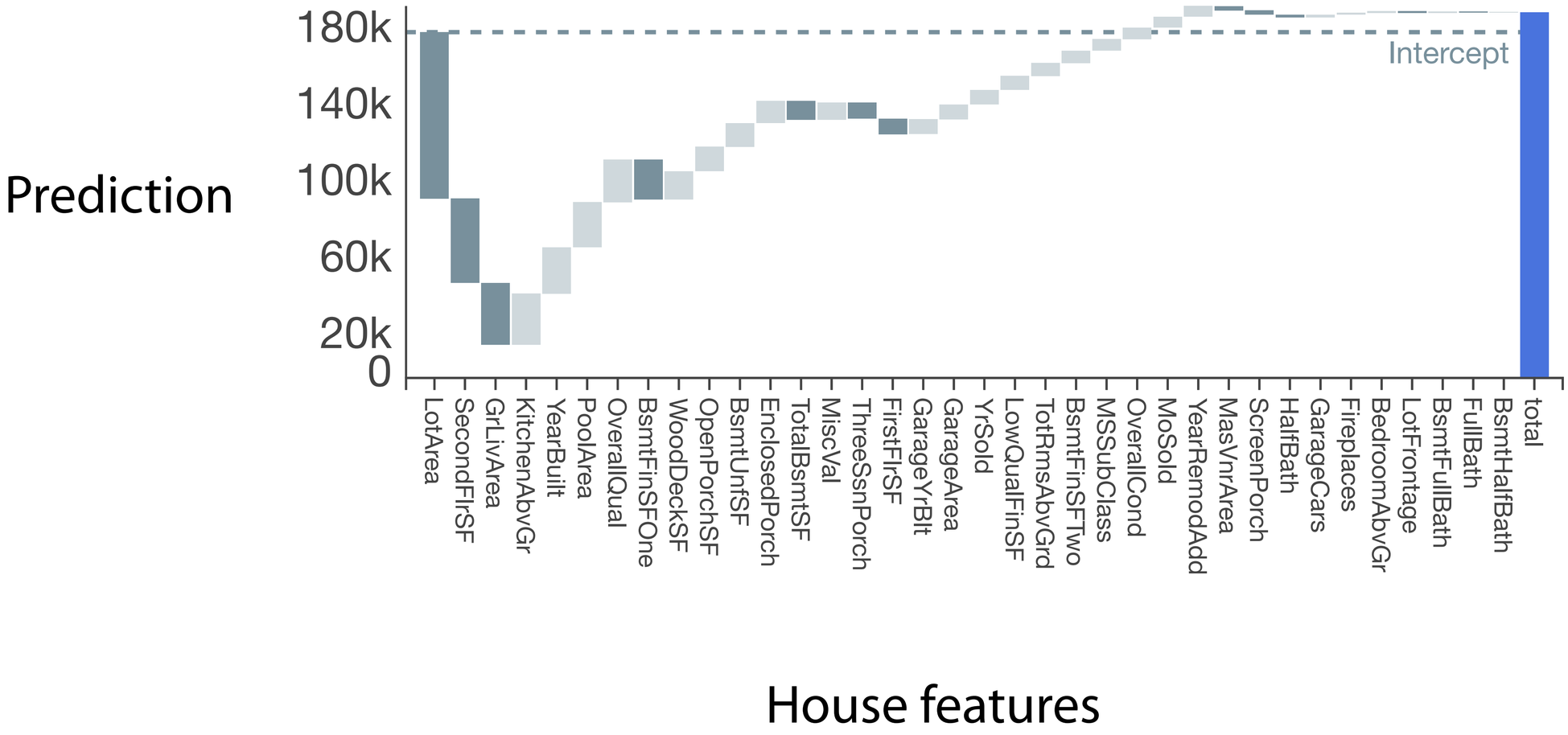
House 550

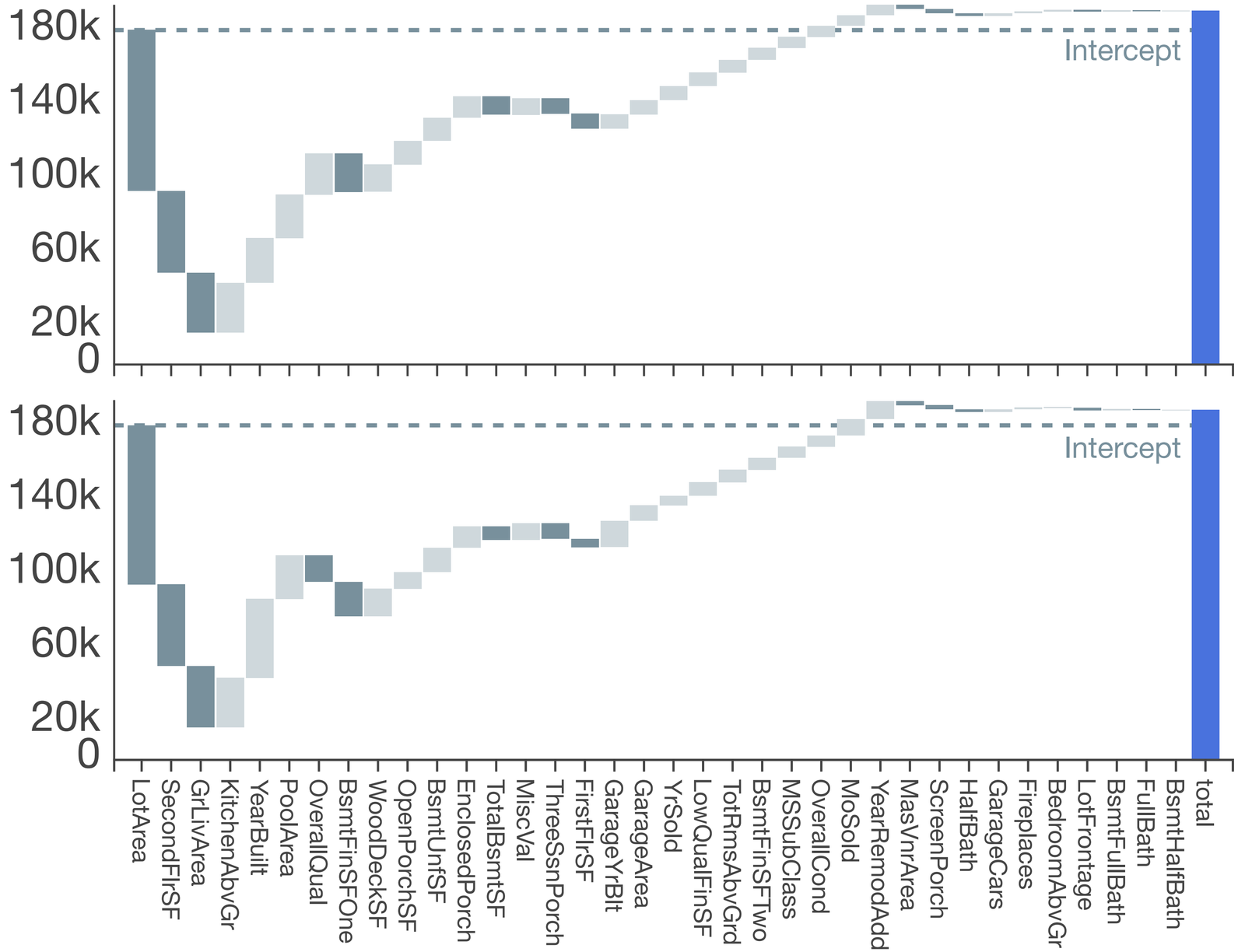
$190,606

BsmtHalfBath
FullBath
BsmtFullBath
LotFrontage
BedroomAbvGr
Fireplaces
GarageCars
HalfBath
ScreenPorch
MasVnrArea
YearRemodAdd
MoSold
OverallCond
MSSubClass
BsmtFinSFTwo
TotRmsAbvGrd
LowQualFinSF
YrSold
GarageArea
GarageYrBlt
FirstFlrSF
ThreeSsnPorch
MiscVal
TotalBsmtSF
EnclosedPorch
BsmtUnfSF
OpenPorchSF
WoodDeckSF
BsmtFinSFOne
OverallQual
PoolArea
YearBuilt
KitchenAbvGr
GrLivArea
SecondFlrSF
LotArea

House 550

$190,606



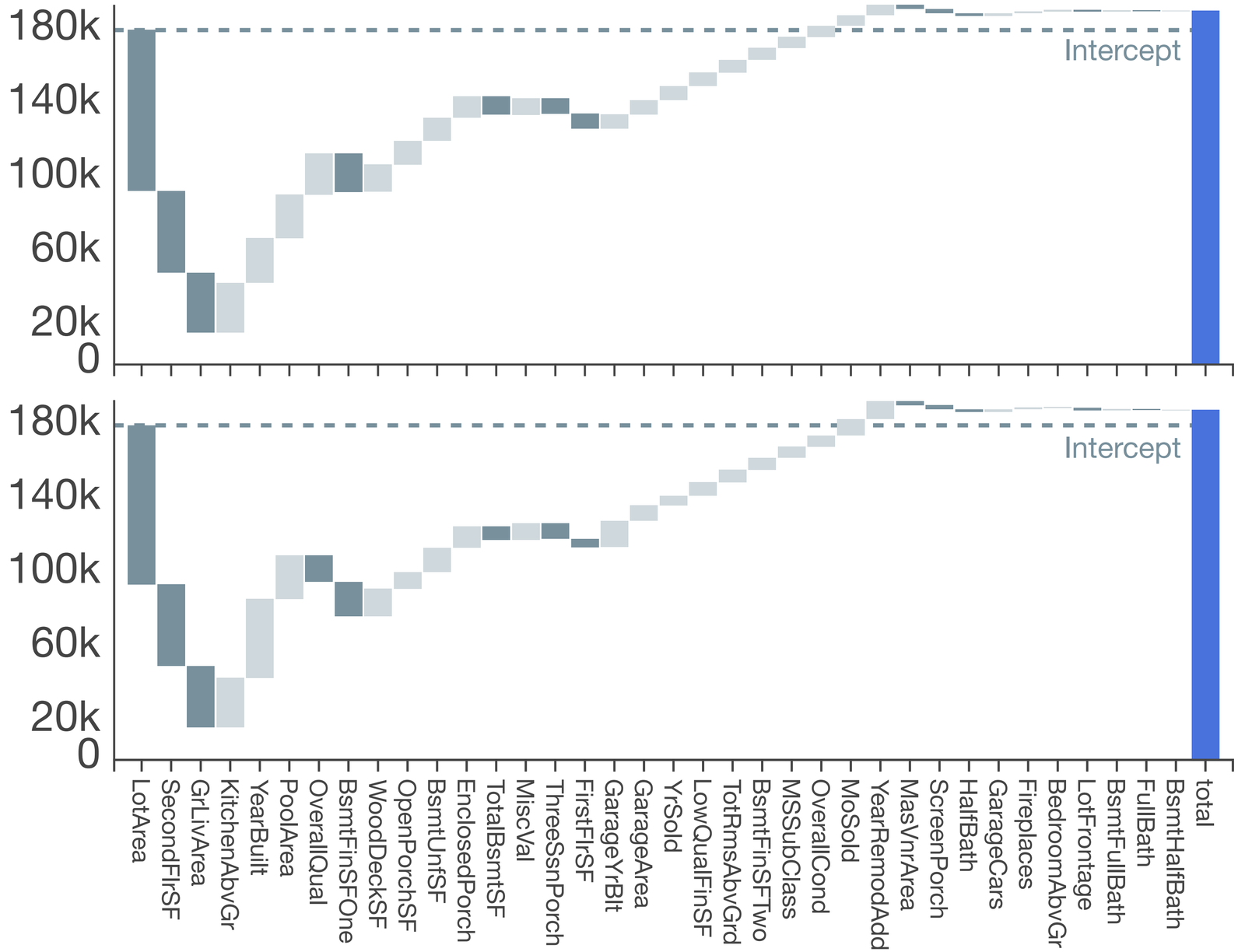


House 550

$190,606

House 798

$188,620

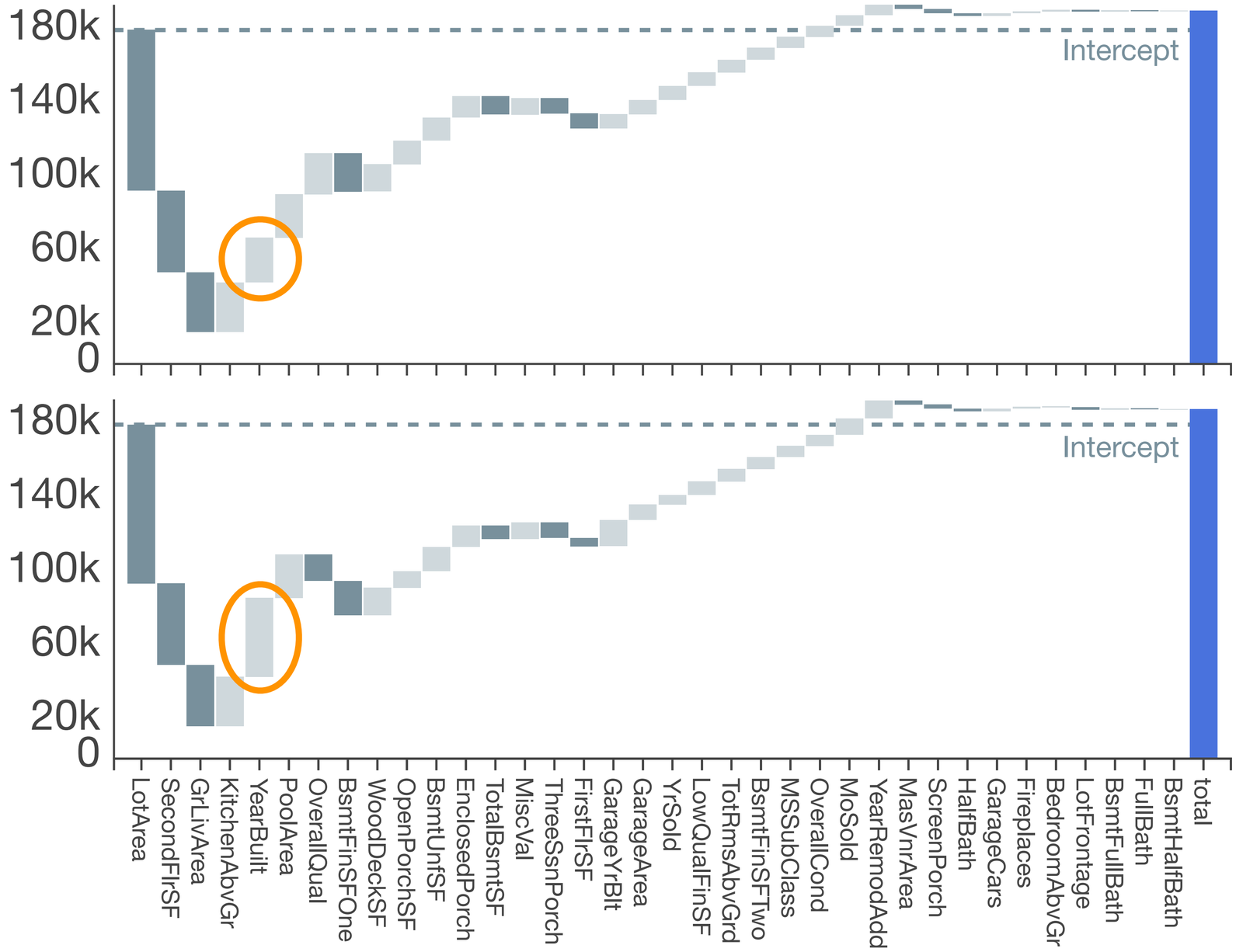


House 550

$190,606

House 798

$188,620

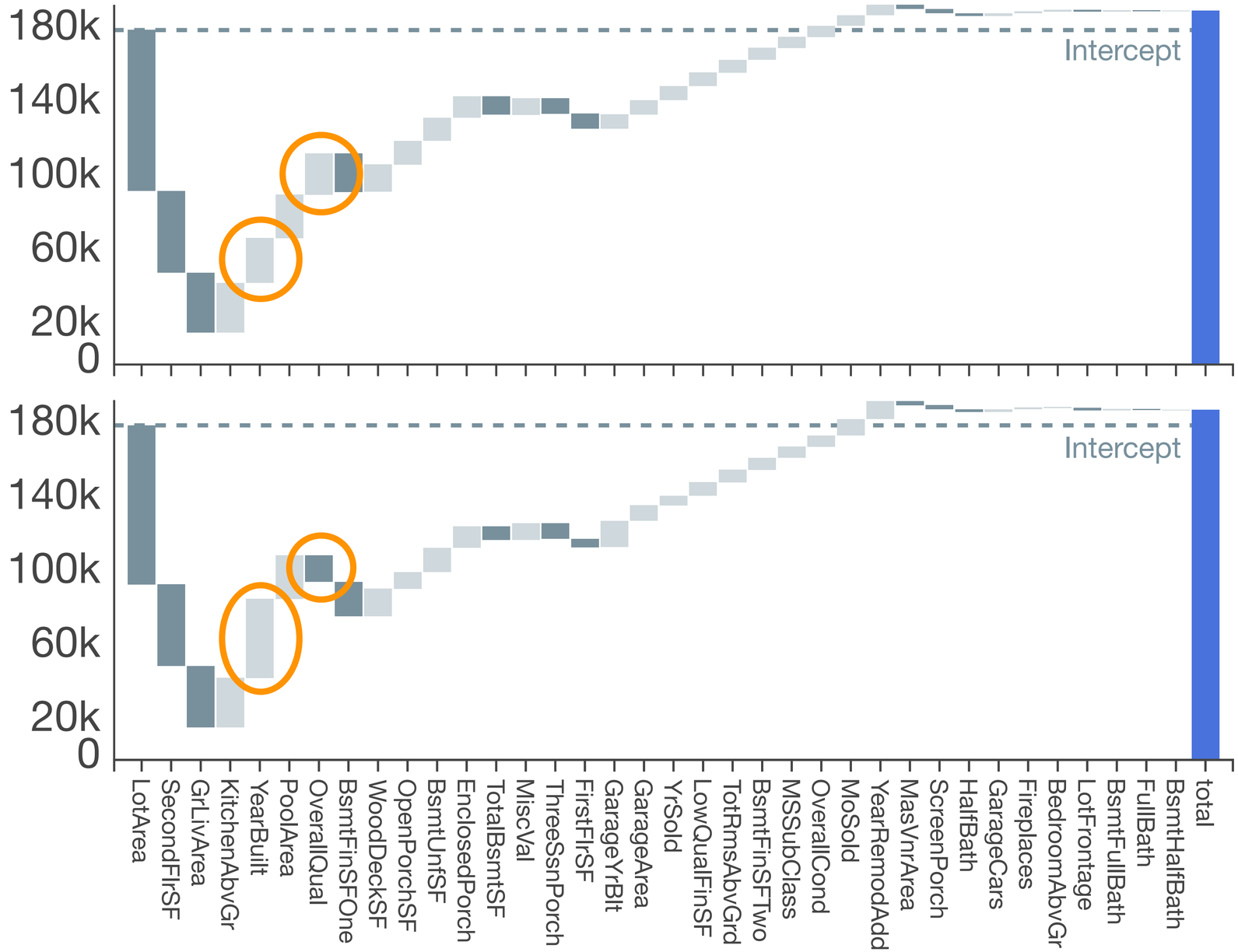


House 550

$190,606

House 798

$188,620

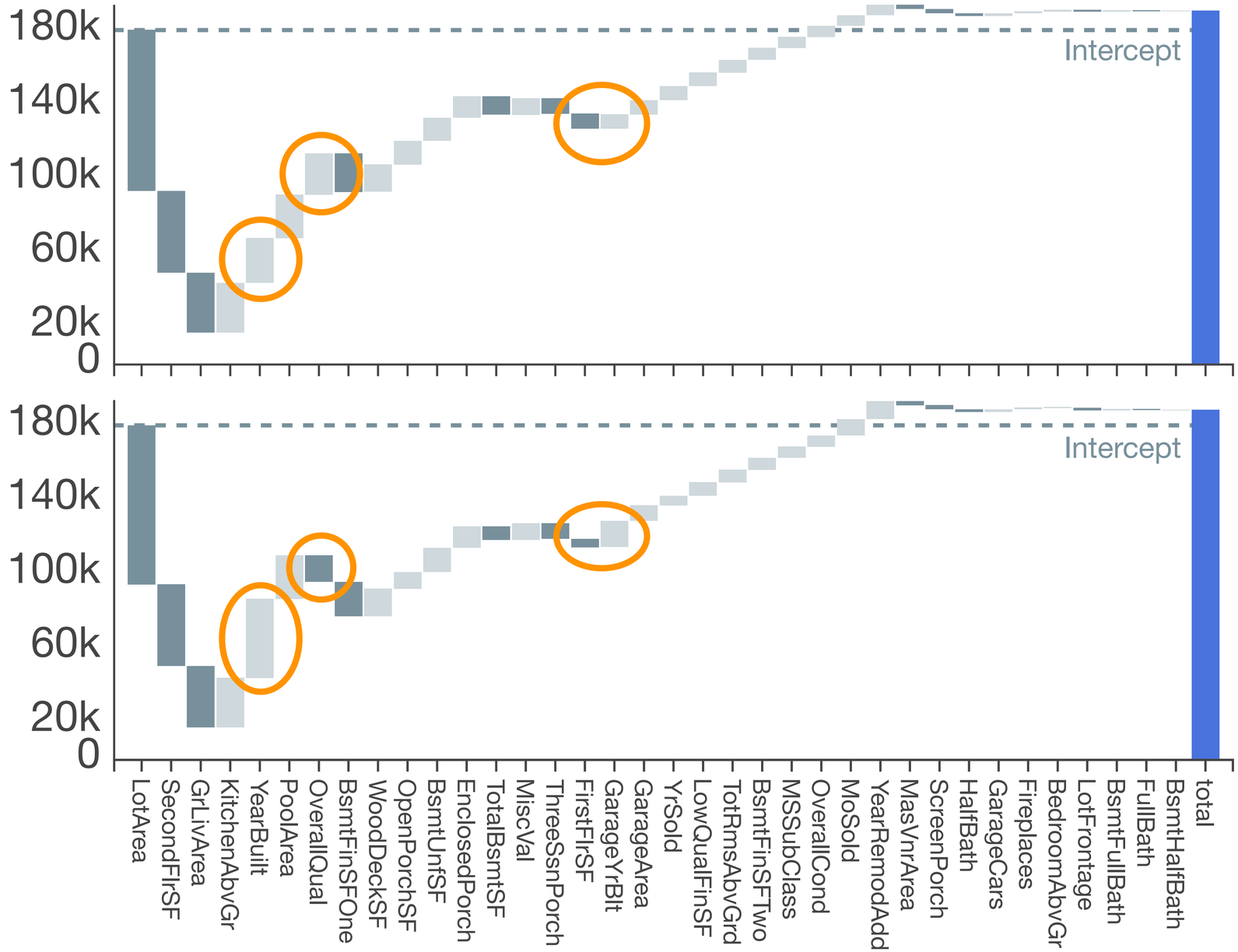


House 550

$190,606

House 798

$188,620

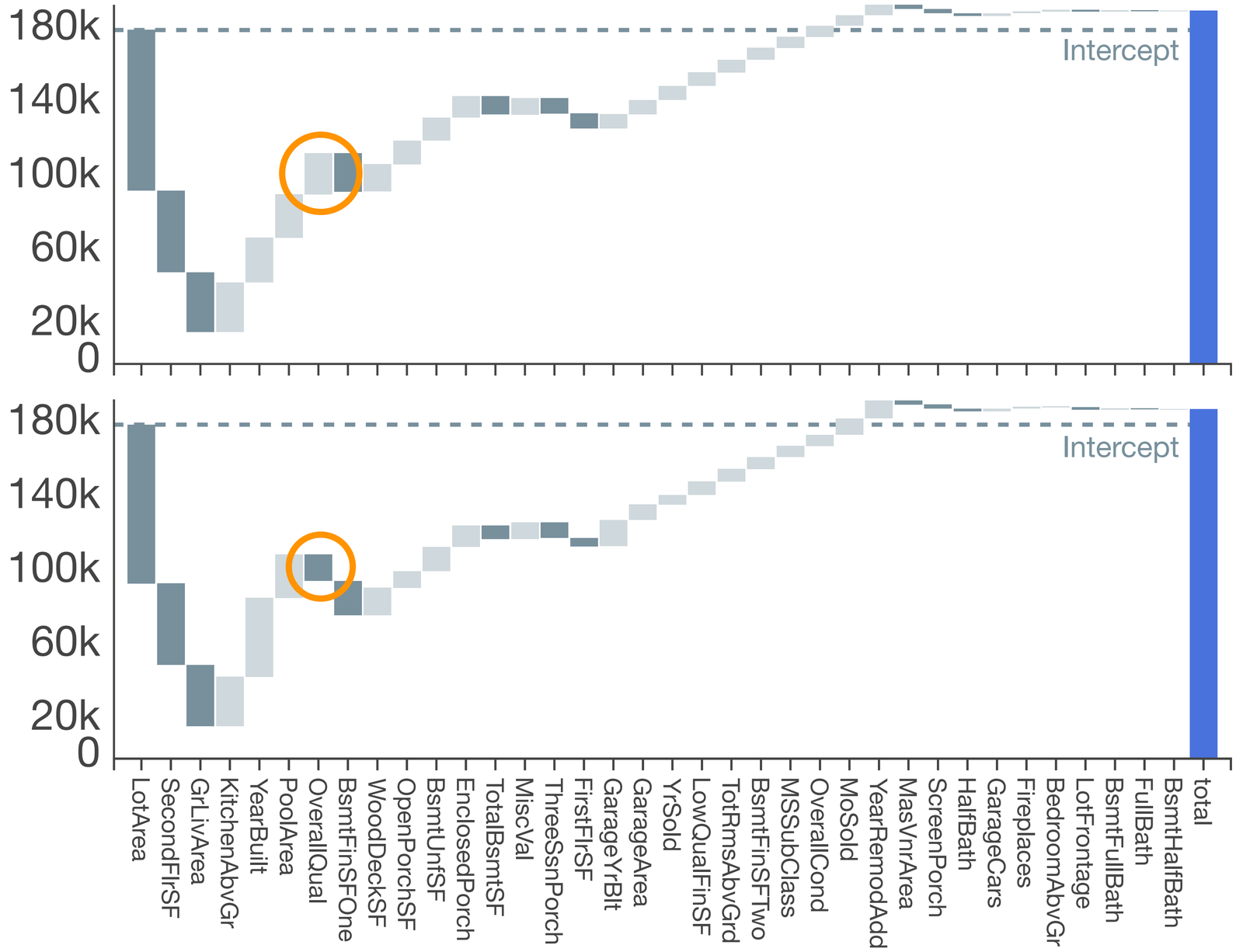


House 550

$190,606

House 798

$188,620

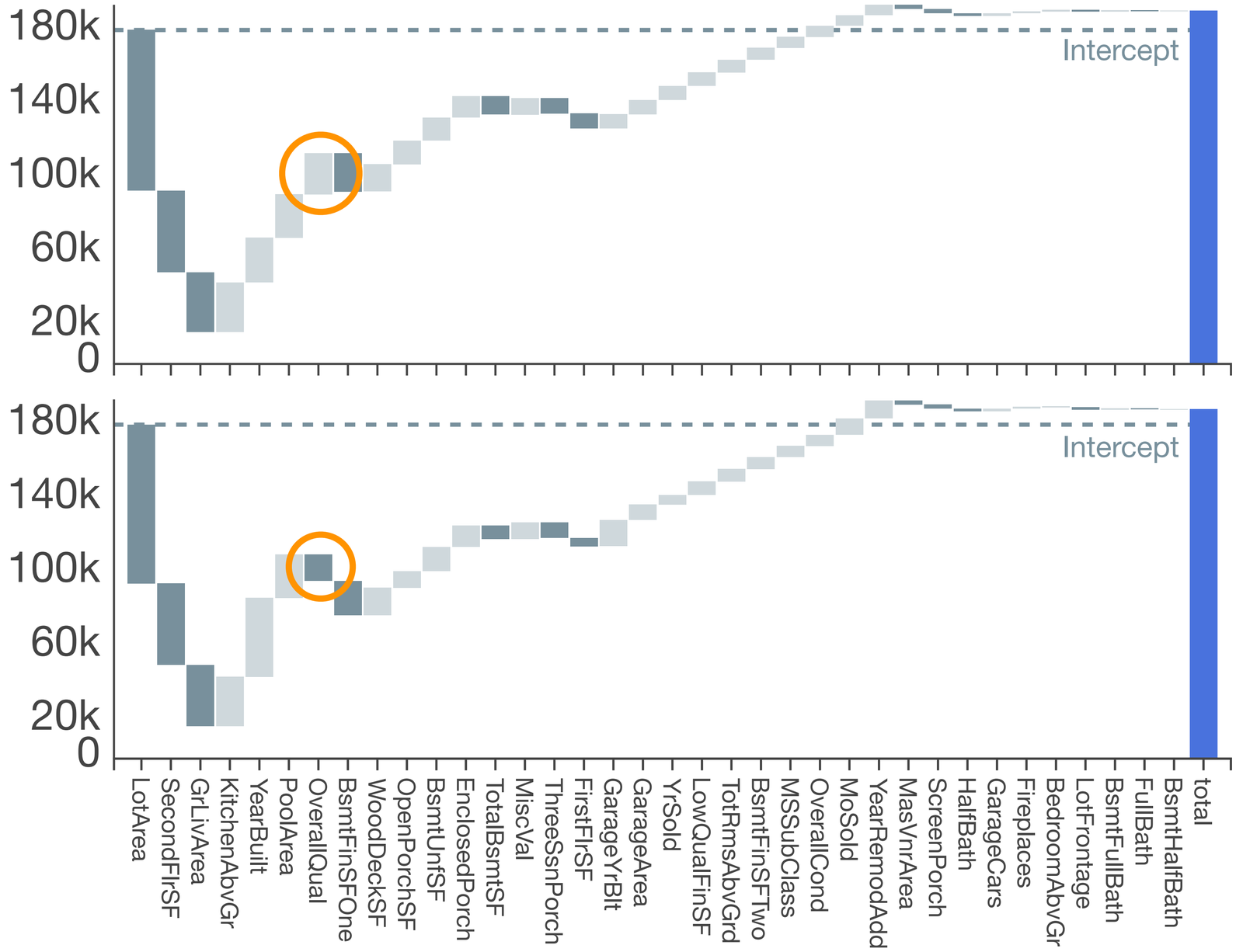


House 550

$190,606

House 798

$188,620

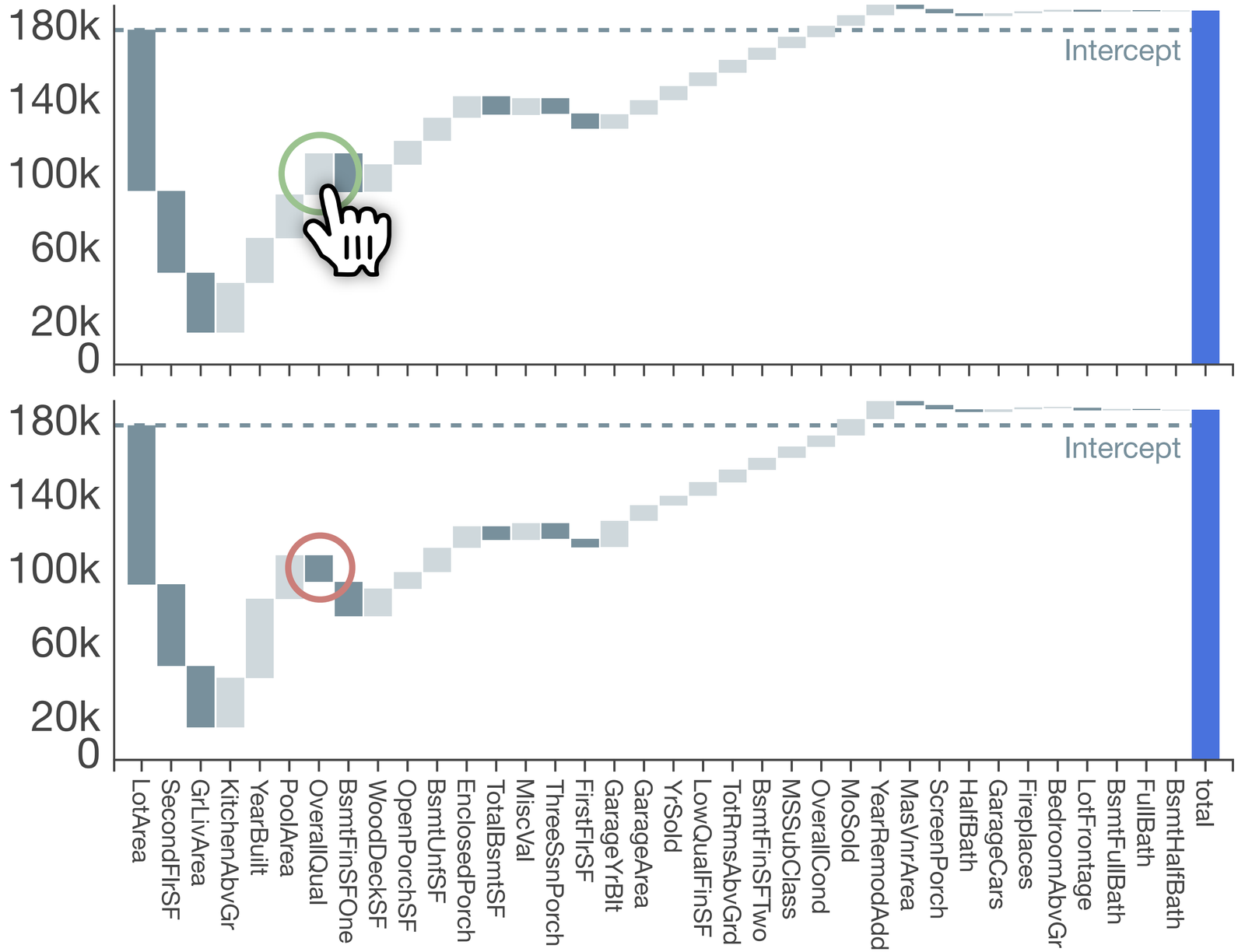


House 550

$190,606

House 798

$188,620

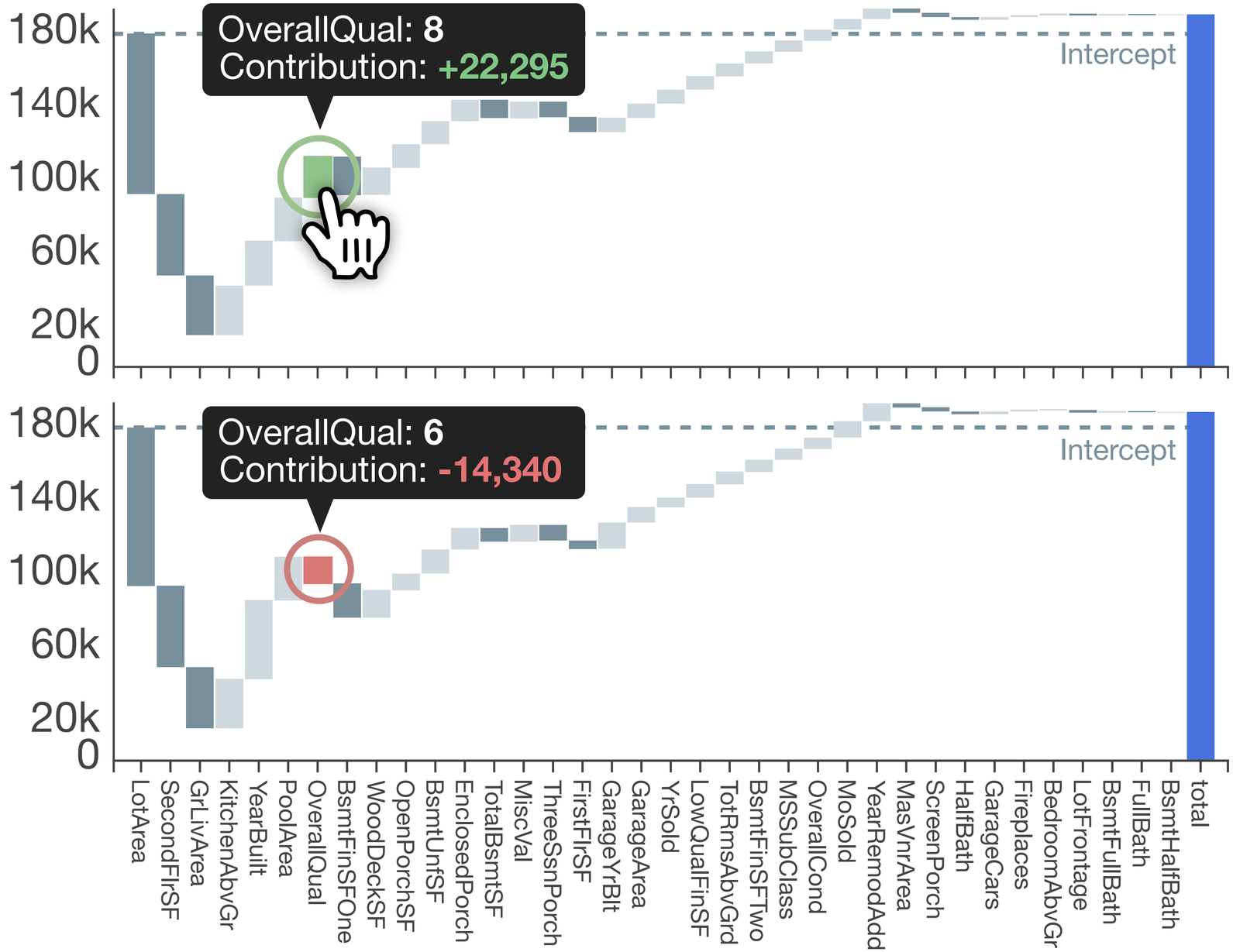


House 550

$190,606

House 798

$188,620

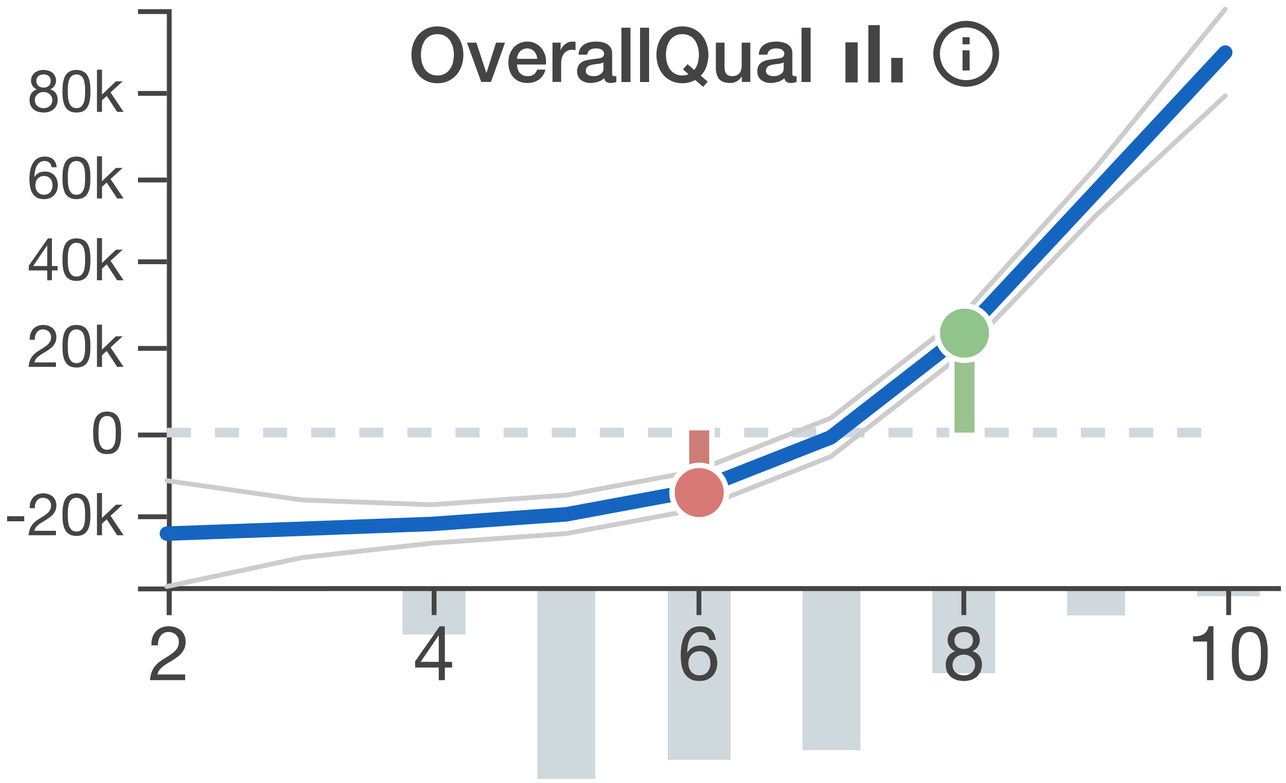


House 550

$190,606

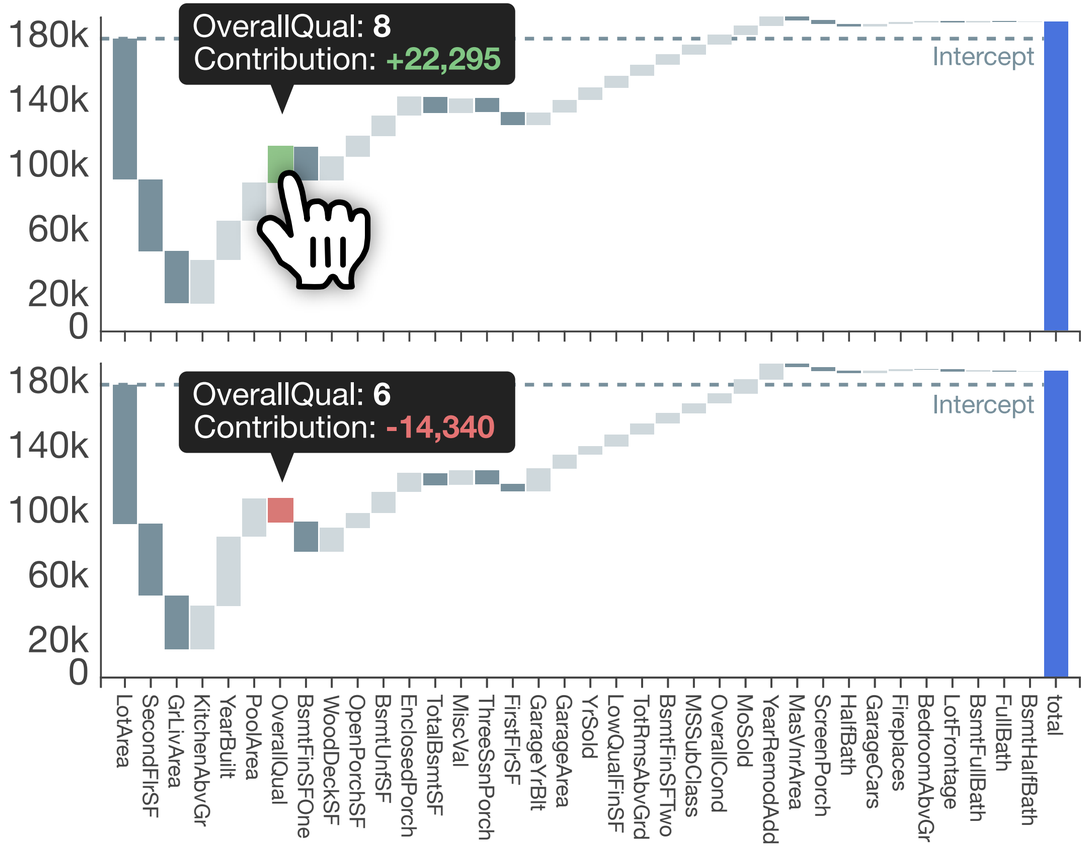
House 798

$188,620



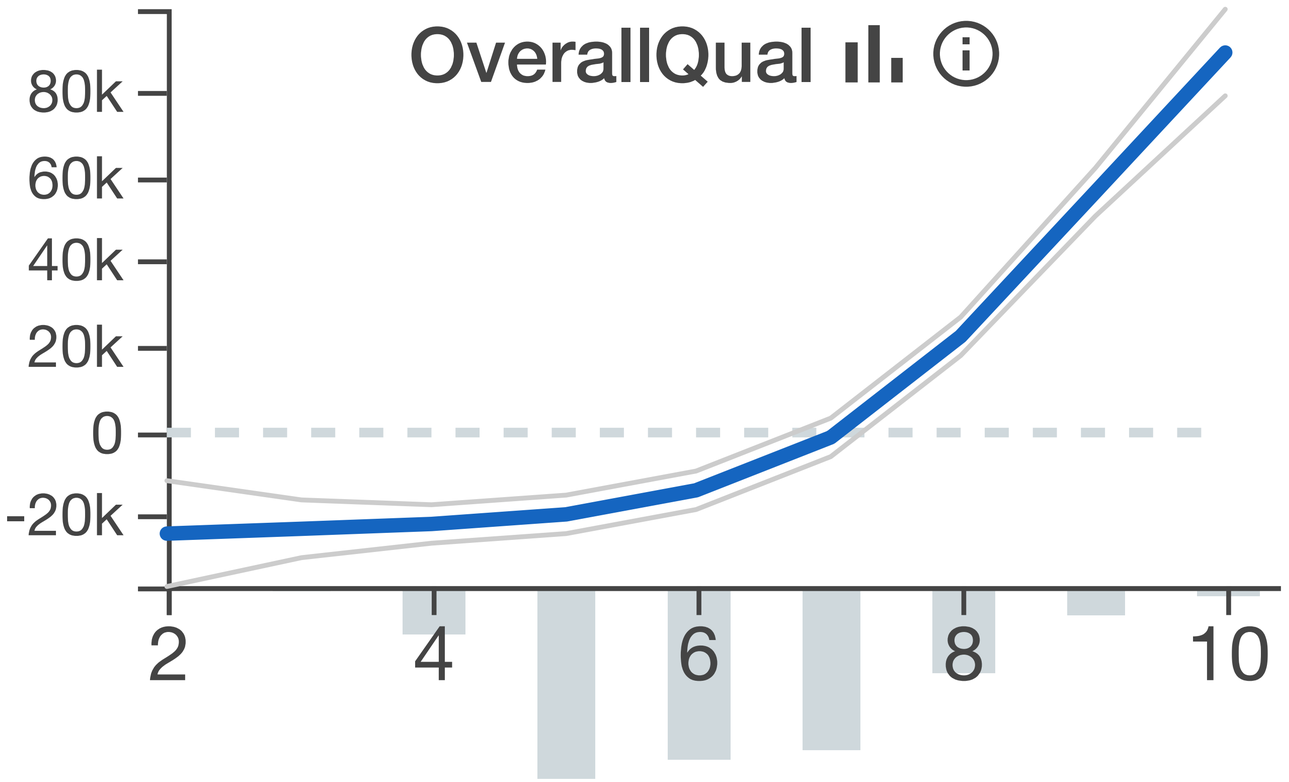
House 550

$190,606



House 798

$188,620



Generalized Additive Model (**GAM**)

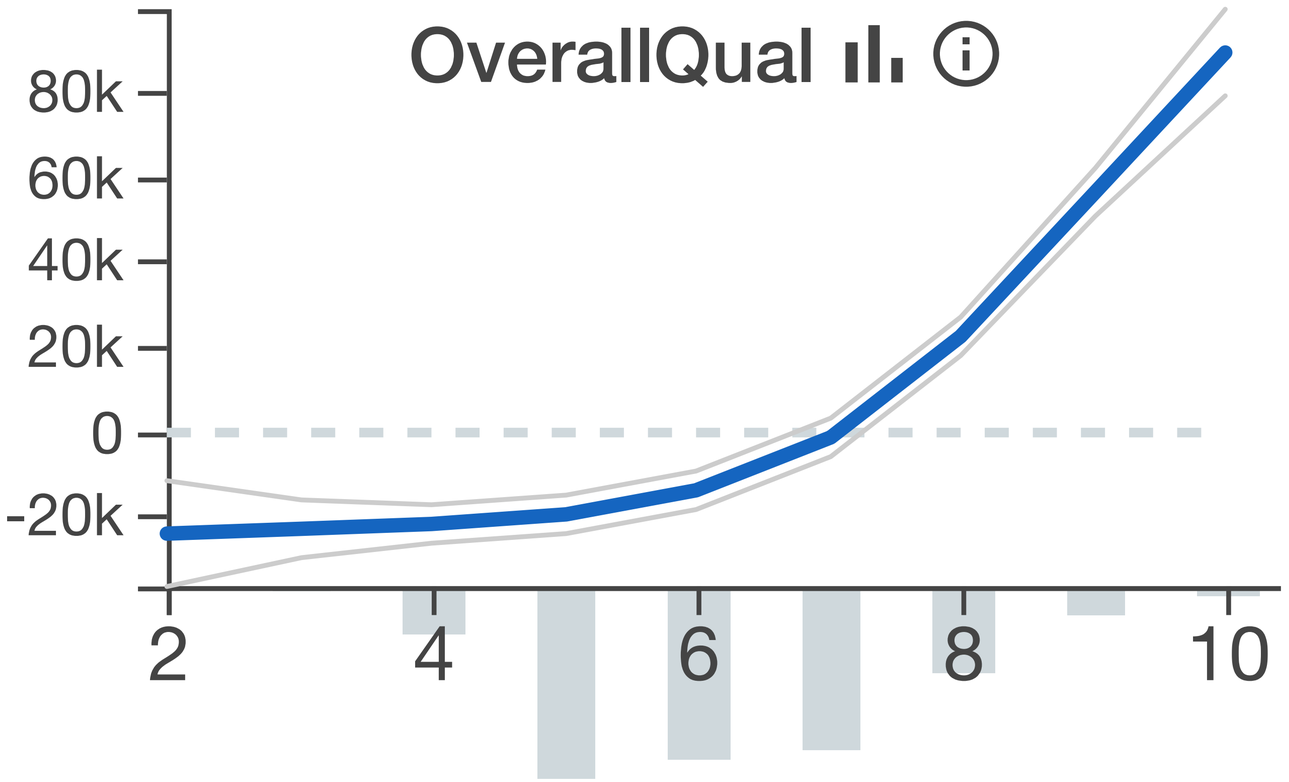
- ☑ Global explanation
- ☑ Easy to understand:
  - ☑ Average math skills
  - ☑ Average graphicacy
- ☑ High accuracy, realistic

Generalized **linear** model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Generalized **additive** model

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$$



Generalized Additive Model (**GAM**)

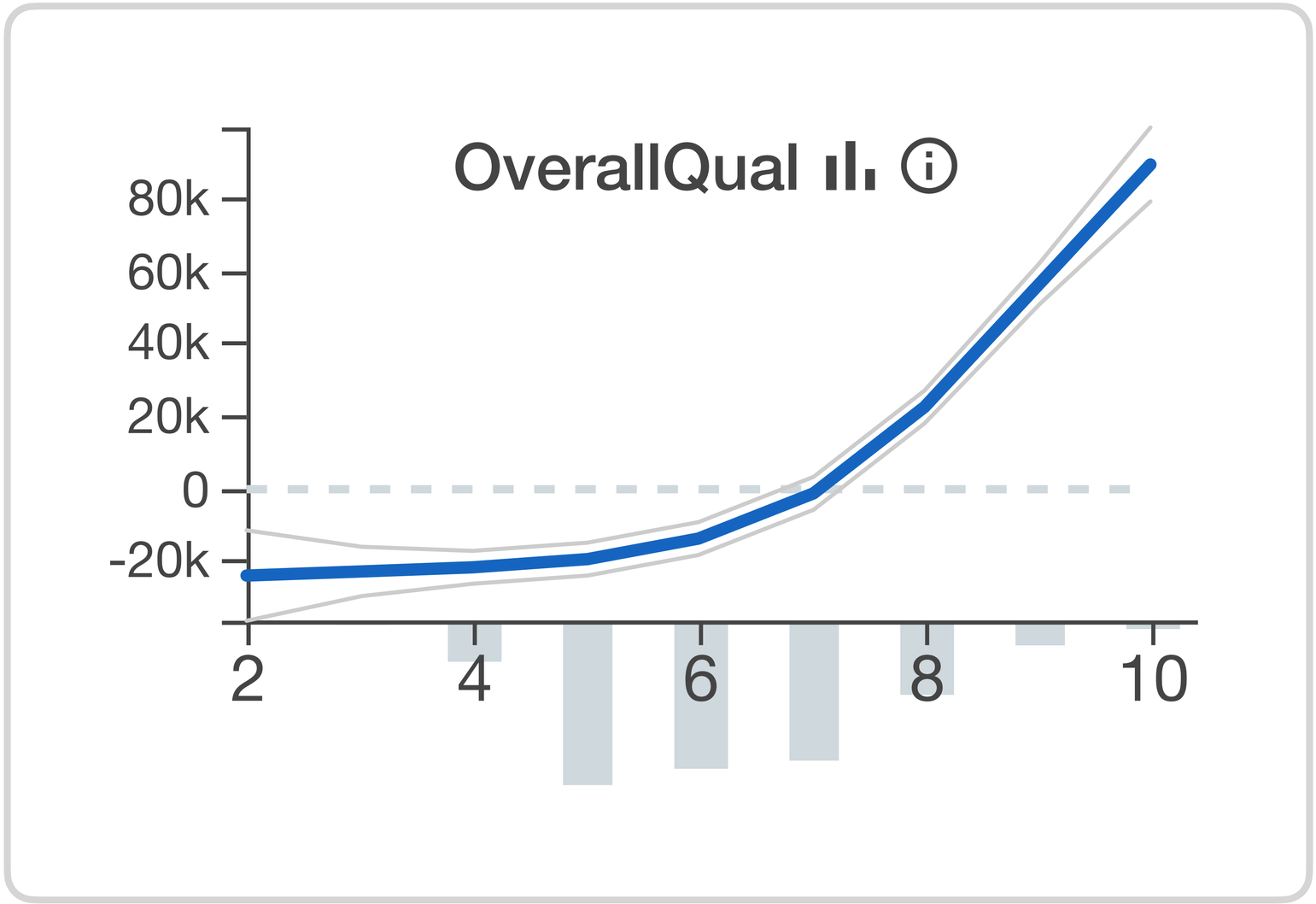
- ☑ Global explanation
- ☑ Easy to understand:
  - ☑ Average math skills
  - ☑ Average graphicacy
- ☑ High accuracy, realistic

Generalized **linear** model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Generalized **additive** model

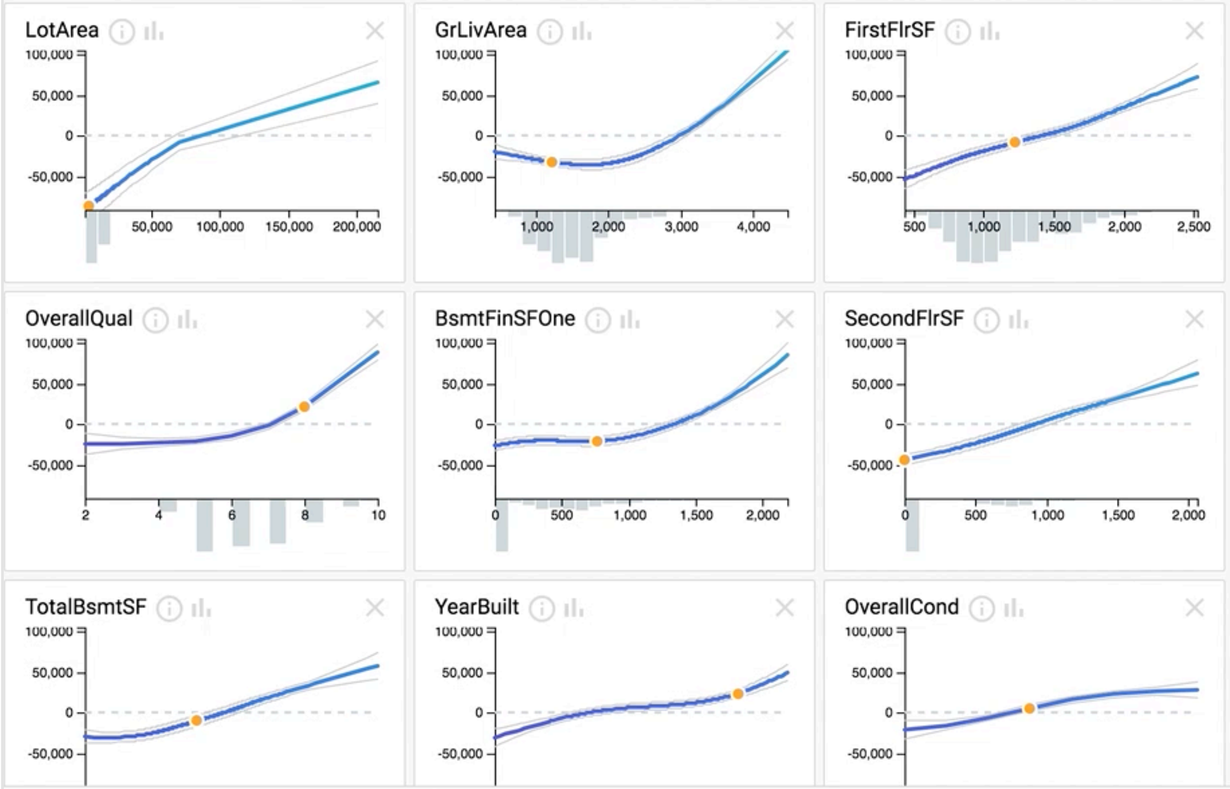
$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$$


OverallQual
80k
60k
40k
20k
0
-20k
2
4
6
8
10

Gamut

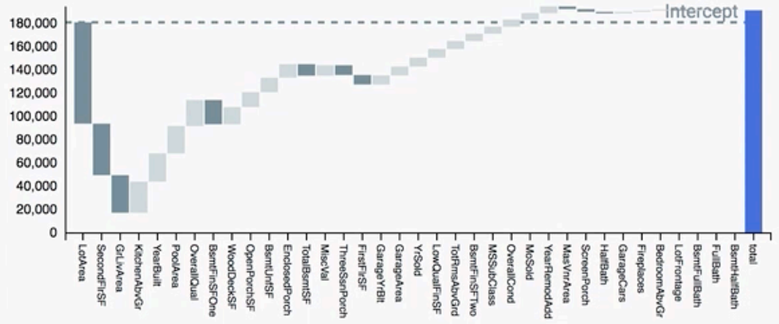
MODEL ames-housing | RSME 20,981.66 | INTERCEPT 180,193.38

Normalize axes | Hide all histograms | Hide zeroline

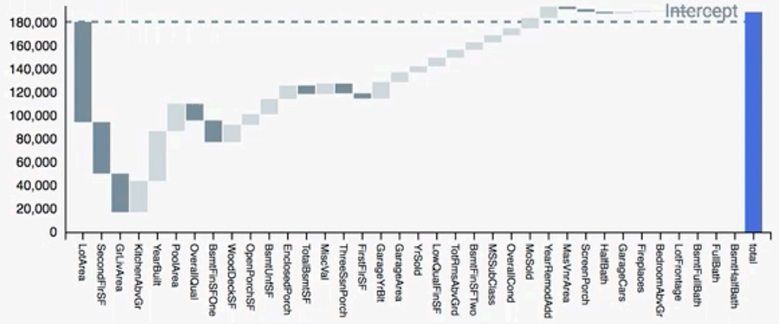


Sort waterfall linear

Instance 550 Actual: 189,000.00 Prediction: 190,606.28



Instance 798 Actual: 196,000.00 Prediction: 188,620.13



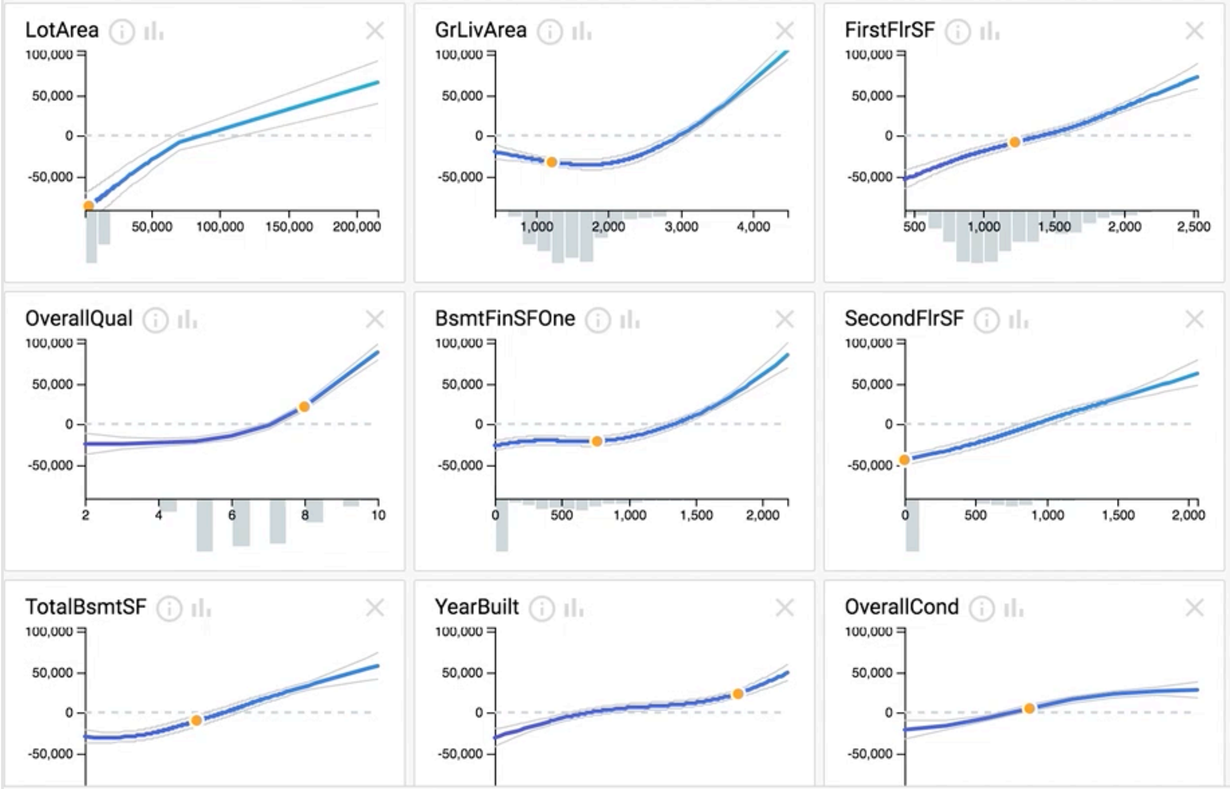
Showing 1119 of 1119 | CLEAR FILTERS | Nearest neighbors in Feature space | SORT BY NEIGHBORS

ID	Actual	Predicted	Differen...	Neighbo...	LotArea	GrLivArea	FirstFlrSF	Overall...	BsmtFin...	Second...	TotalBs...	YearBuilt	Overall...	Enclose...	Kitchen...
795	315500	320700.119...	5200.11951...	0.7800970...	12898	1620	1620	9	1022	0	1620	2007	5	0	1
796	80000	59768.784...	20231.2153...	1.0558532...	1477	630	630	4	509	0	630	1970	4	0	1
797	155000	144391.109...	10608.890...	1.2288659...	13125	1803	1803	5	168	0	1134	1957	4	0	1
798	196000	188620.127...	7379.8722...	0.3635679...	5381	1306	1306	6	900	0	1306	2005	5	0	1
799	262280	263513.121...	1233.12142...	1.0479754...	11839	2329	1532	7	1085	797	1475	1990	5	0	1
800	278000	323164.97...	45164.9701...	1.28911306...	9600	2524	2524	8	1104	0	2524	1981	5	0	1
801	556581	433082.42...	123498.571...	1.45181310...	16056	2868	1992	9	240	876	1992	2005	5	0	1
802	145000	136593.33...	8406.6626...	1.3020229...	9245	990	990	5	686	0	990	1994	5	0	1
803	115000	136060.69...	21060.692...	1.18223071...	21750	1771	1771	5	0	0	0	1960	4	0	1
804	84900	115655.109...	30755.1091...	1.3081934...	11100	930	930	4	0	0	0	1946	7	0	1

Gamut

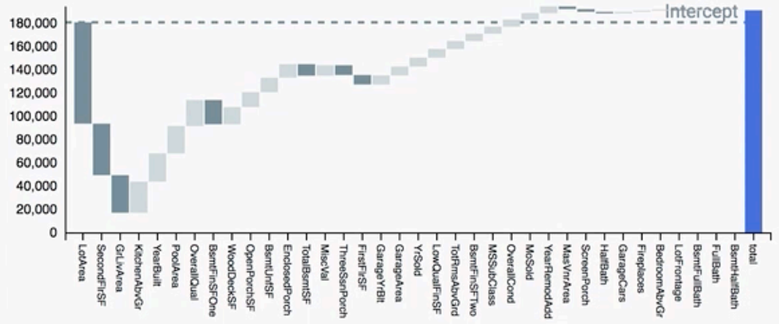
MODEL ames-housing | RSME 20,981.66 | INTERCEPT 180,193.38

Normalize axes | Hide all histograms | Hide zeroline

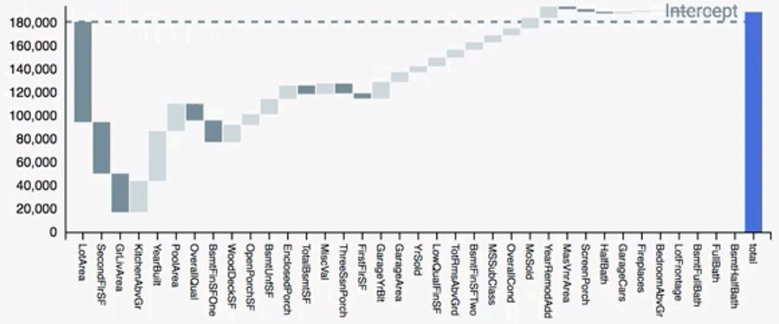


Sort waterfall linear

Instance 550 Actual: 189,000.00 Prediction: 190,606.28



Instance 798 Actual: 196,000.00 Prediction: 188,620.13



Showing 1119 of 1119 | CLEAR FILTERS | Nearest neighbors in Feature space | SORT BY NEIGHBORS

ID	Actual	Predicted	Differen...	Neighbo...	LotArea	GrLivArea	FirstFlrSF	Overall...	BsmtFin...	Second...	TotalBs...	YearBuilt	Overall...	Enclose...	Kitchen...
795	315500	320700.119...	5200.11951...	0.7800970...	12898	1620	1620	9	1022	0	1620	2007	5	0	1
796	80000	59768.784...	20231.2153...	1.0558532...	1477	630	630	4	509	0	630	1970	4	0	1
797	155000	144391.109...	10608.890...	1.2288659...	13125	1803	1803	5	168	0	1134	1957	4	0	1
798	196000	188620.127...	7379.8722...	0.3635679...	5381	1306	1306	6	900	0	1306	2005	5	0	1
799	262280	263513.121...	1233.12142...	1.0479754...	11839	2329	1532	7	1085	797	1475	1990	5	0	1
800	278000	323164.97...	45164.9701...	1.28911306...	9600	2524	2524	8	1104	0	2524	1981	5	0	1
801	556581	433082.42...	123498.571...	1.45181310...	16056	2868	1992	9	240	876	1992	2005	5	0	1
802	145000	136593.33...	8406.6626...	1.3020229...	9245	990	990	5	686	0	990	1994	5	0	1
803	115000	136060.69...	21060.692...	1.18223071...	21750	1771	1771	5	0	0	0	1960	4	0	1
804	84900	115655.109...	30755.1091...	1.3081934...	11100	930	930	4	0	0	0	1946	7	0	1

User Study Findings Summary

Reasons for Interpretability 🔬

Why do data scientists need interpretability and how do they use it in GAMUT?

Global v. Local Explanations 🌎

How do data scientists use different explanation paradigms?

Interactive Explanations ⚡

How does interactivity play a role in explainable machine learning interfaces?

12 data scientists, ~1.5 hours each

Think-aloud + question answering about multiple models

Tutorial → Study → Interview

User Study Findings Summary

Reasons for Interpretability 🔬

Model debugging & improvement

Understanding data

Hypothesis generation for building trust

Communication

Global v. Local Explanations 🌎

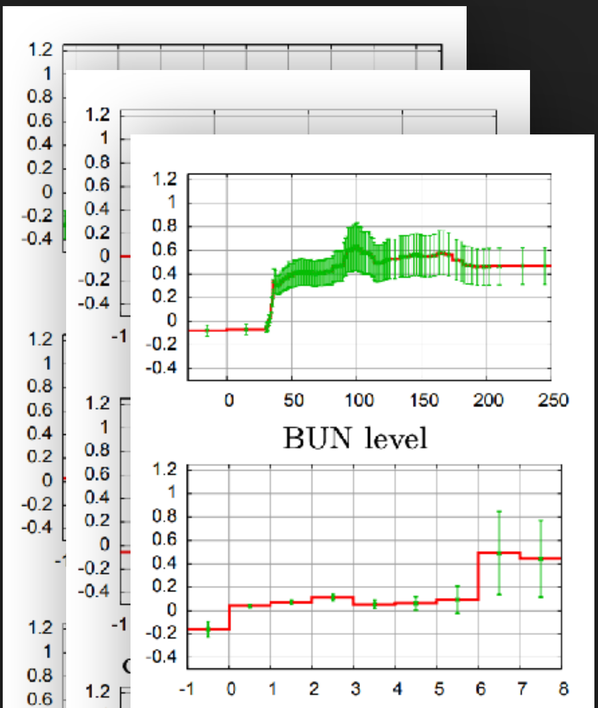
Global
Familiars
[3-5 years]

Both
Experts
[5+ years]

Local
Novices
[1-2 years]

Interactive Explanations ⚡

Primary mechanism for explaining predictions



User Study Findings Summary

Reasons for Interpretability 🔬

Interpretability is not a singular, rigid concept

Operationalization helps people in practice

Global v. Local Explanations 🌎

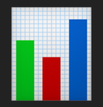
Tailor explanations for specific audiences

Balance simplicity and completeness

Interactive Explanations ⚡

Design and integrate effective interaction

Interaction key to realizing interpretability

Visualization 

Explanations

- Show model context
- Interactive analytics
- Rely on user interpretation

Visualization 📊

Explanations

- Show model context
- Interactive analytics
- Rely on user interpretation

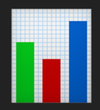
 Verbalization

Explanations

- Direct and concise
- Less cognitive load
- No training needed

Visualization

Explanations



+

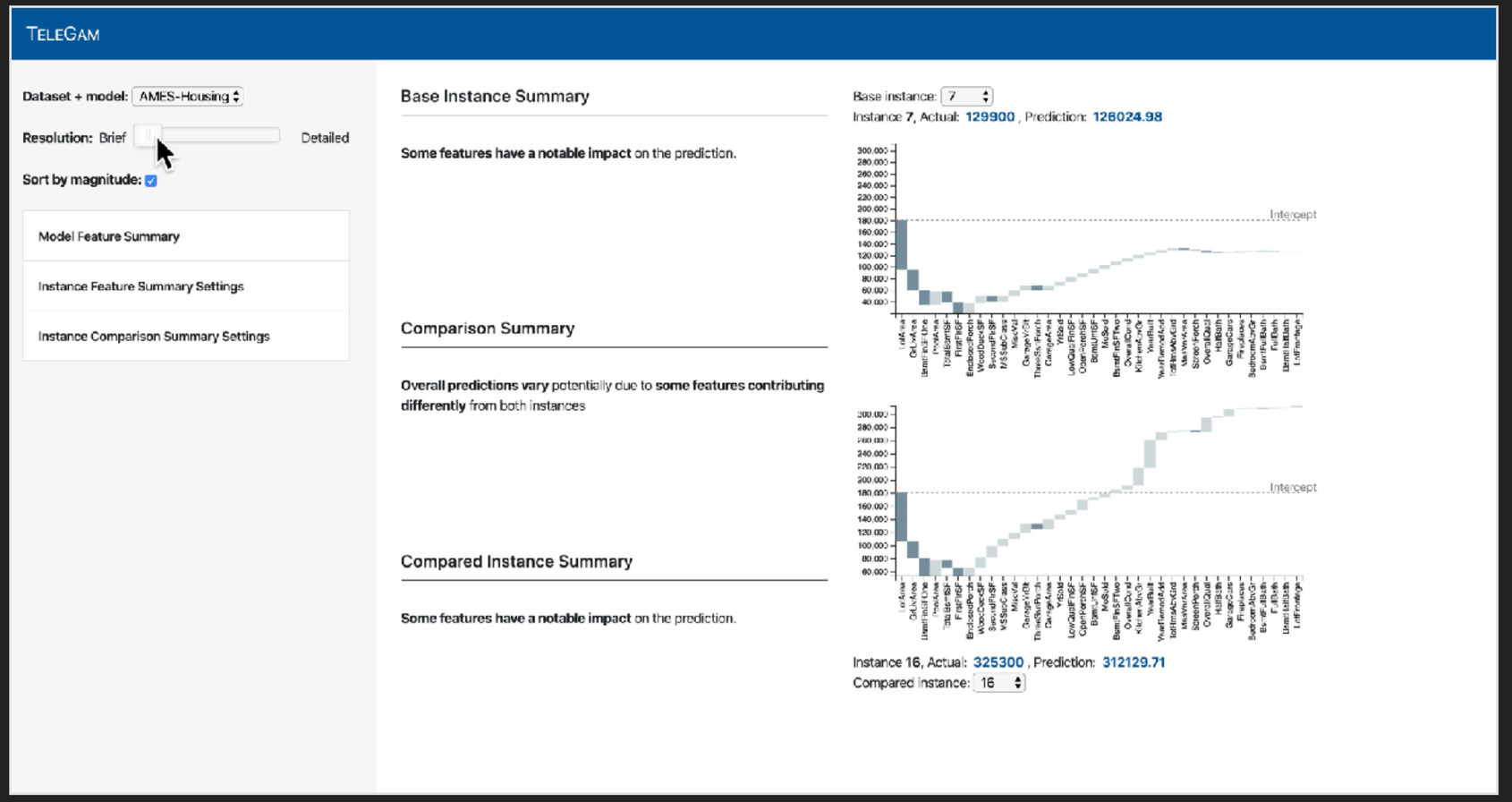
Verbalization

Explanations

TeleGam

VIS 2019

Combining Visualization & Verbalization for Machine Learning Interpretability



Fred Hohman
Georgia Tech

Arjun Srinivasan
Georgia Tech

Steven Drucker
Microsoft Research

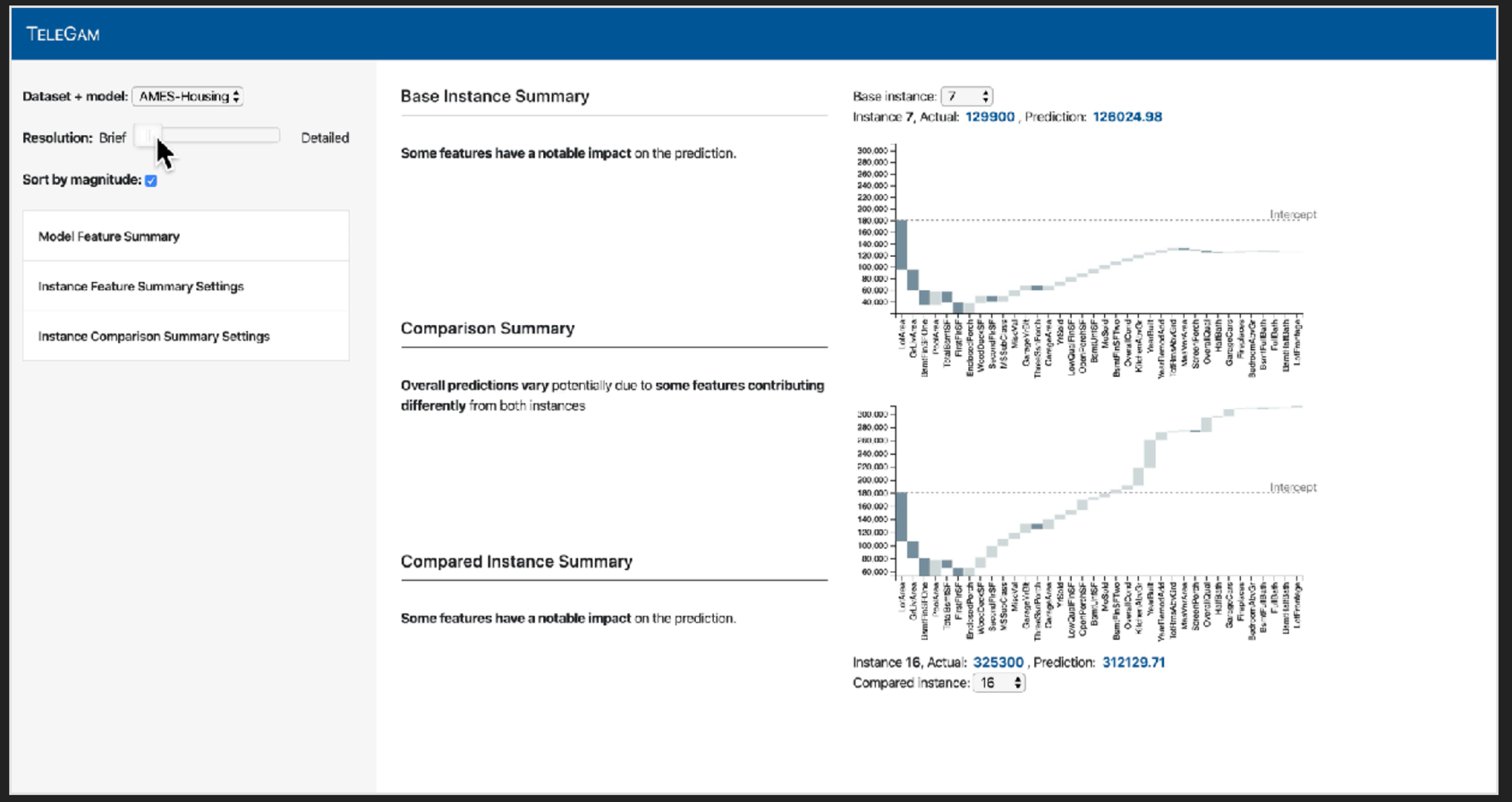
Interactively highlight **verbalization** in context of the **visualization**

Users decide resolution, balance **simplicity** & **completeness**

TeleGam

VIS 2019

Combining Visualization & Verbalization for Machine Learning Interpretability



Fred Hohman
Georgia Tech

Arjun Srinivasan
Georgia Tech

Steven Drucker
Microsoft Research

Interactively highlight **verbalization** in context of the **visualization**

Users decide resolution, balance **simplicity** & **completeness**

Impact

GAMUT

Deployed at Microsoft Research
Demoed at for executive leadership

`msrgamut.microsoft.com`

INTERPRETML 2.4k+ stars

Influenced visualization design for popular interpretability toolkit

`github.com/interpretml/interpret`

SANDDANCE 4.2k+ stars

GAMUT linked to visualization and data exploration tools

`sanddance.js.org`

Interactive Scalable Interfaces for Machine Learning Interpretability

PART I
Enable interpretability

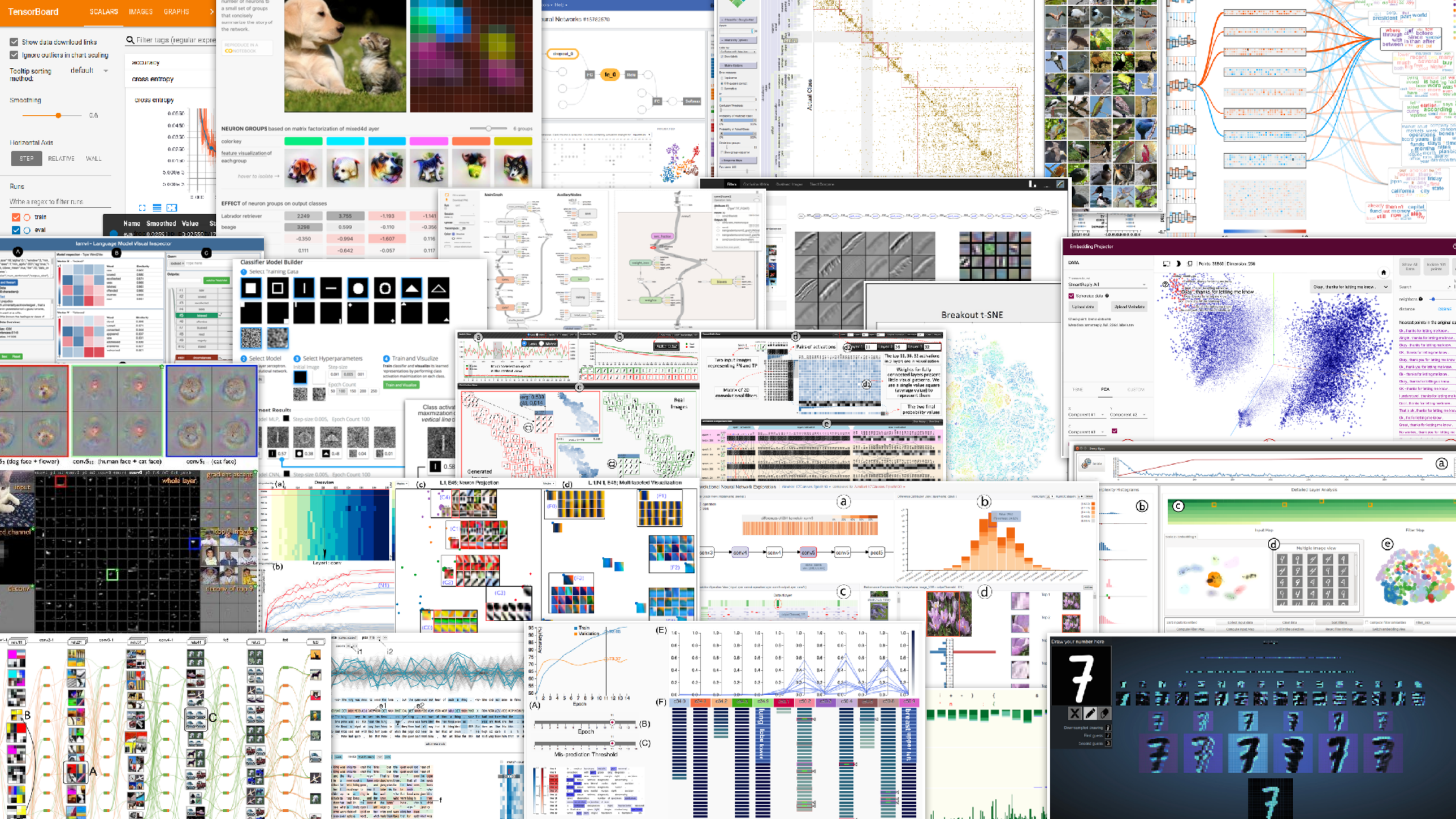
GAMUT Operationalize interpretability *CHI 2019*
TELEGAM Vis + text for better explanations *VIS 2019*

PART II
Scale interpretability

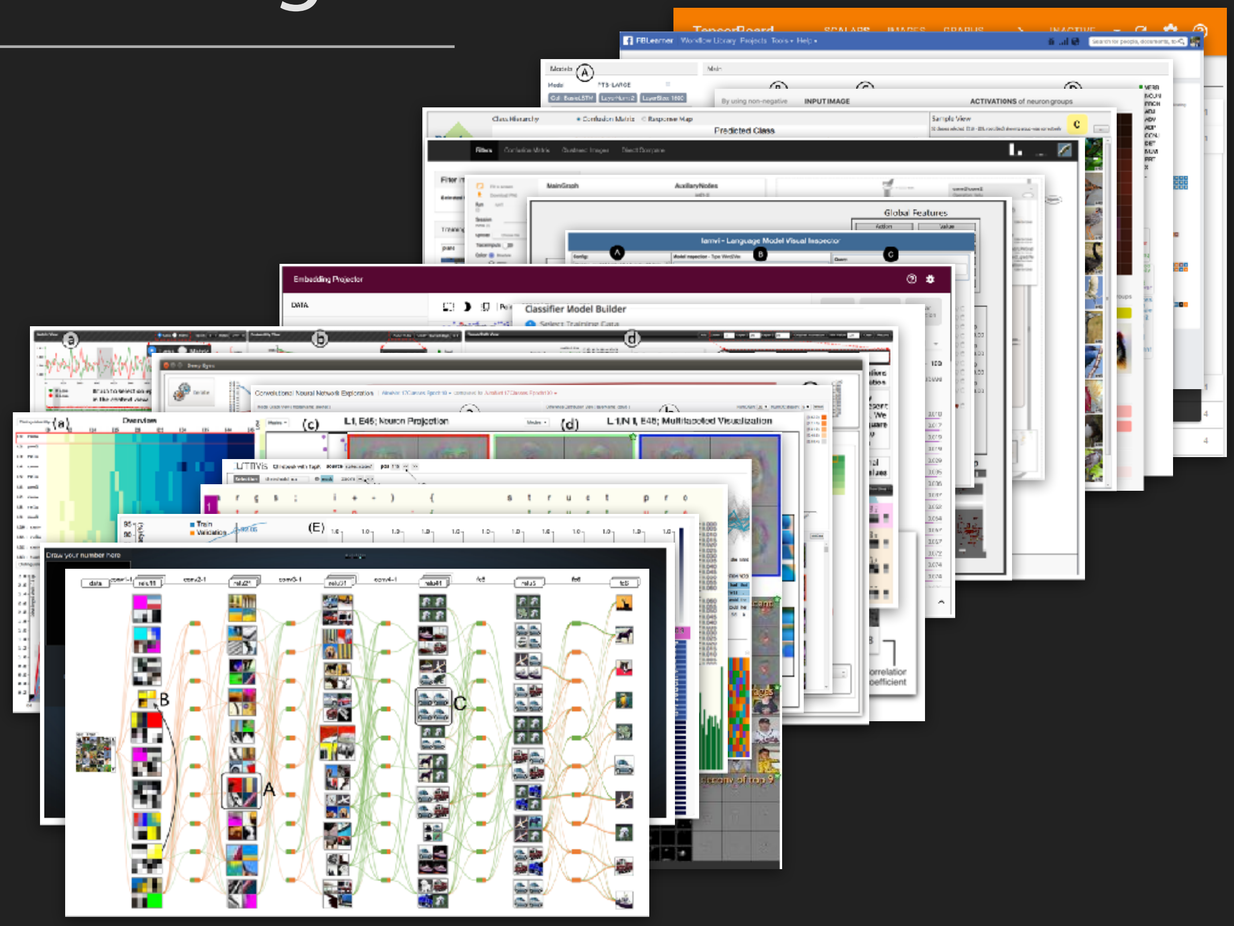
Interrogative Survey Summarize interpretability vis *TVCG 2018*
SUMMIT Higher-level explanations for neural networks *VAST 2019*

PART III
Communicate interpretability

ML Literacy Interactive mediums & platforms *VISCOMM 2019, VISxAI 2018*
Interactive Articles Formalizing interactive communication *Distill 2020*


TensorBoard
SCALARS
IMAGES
GRAPHS
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method:
default
Smoothing
0.6
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
train
eval
accuracy
cross entropy
NEURON GROUPS based on matrix factorization of mixed4d ayer
6 groups
color key
feature visualization of each group
hover to isolate
EFFECT of neuron groups on output classes
Labrador retriever
beagle
lanvi - Language Model Visual Inspector
Classifier Model Builder
Select Training Data
Select Model
Select Hyperparameters
Train and Visualize
Breakout t-SNE
Embedding Projector
DATA
PCA
CUSTOM
input
whole layer
gradients ascent
deconv
Layer: conv
Generated
Real Images
Train
Validation
Epoch
Mis-prediction Threshold
Draw your number here

Organize & Summarize Research for Visual Analytics in Deep Learning



Interrogative Survey

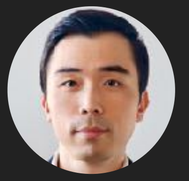
For Visual Analytics in Deep Learning

TVCG 2018

Fred Hohman
Georgia Tech

Minsuk Kahng
Georgia Tech

Robert Pienta
Georgia Tech



Polo Chau
Georgia Tech

Survey Highlights

WHY

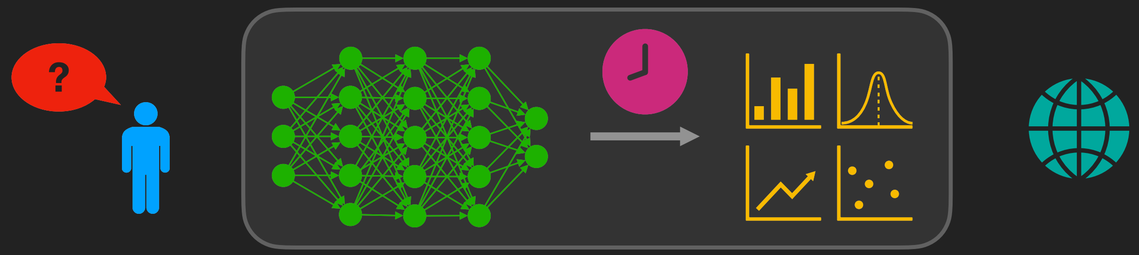
Why would one want to use visualization in deep learning?

WHAT

What data, features, and relationships in deep learning can be visualized?

WHEN

When in the deep learning process is visualization used?



WHO

Who would use and benefit from visualizing deep learning?

HOW

How can we visualize deep learning data, features, and relationships?

WHERE

Where has deep learning visualization been used?

		WHY				WHO			WHAT					HOW								WHEN		WHERE
Author	Year	Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Education	Model Developers & Builders	Model Users	Non-experts	Computational Graph & Network Architecture	Learned Model Parameters	Individual Computational Units	Neurons in High-dimensional Space	Aggregated Information	Node-link Diagrams for Network Architecture	Dimensionality Reduction & Scatter Plots	Line Charts for Temporal Metrics	Instance-based Analysis & Exploration	Interactive Experimentation	Algorithms for Attribution & Feature Visualization	During Training	After Training	Publication Venue
Abadi, et al.	2016	■	■	■		■	■						■			■				■	■	arXiv
Bau, et al.	2017	■		■		■					■		■				■		■		■	CVPR
Bilal, et al.	2017	■	■			■					■		■				■		■	■		TVCG
Bojarski, et al.	2016	■	■			■				■			■				■		■	■	■	arXiv
Bruckner	2014	■	■			■			■	■				■			■		■	■		MS Thesis
Carter, et al.	2016	■			■	■	■	■			■	■	■				■	■			■	Distill
Cashman, et al.	2017	■	■			■	■			■	■						■				■	VADL
Chae, et al.	2017	■	■			■					■		■			■	■			■		VADL
Chung, et al.	2016	■	■			■			■	■	■	■		■	■	■	■			■		FILM
Goyal, et al.	2016	■						■		■							■	■	■		■	arXiv
Harley	2015	■			■			■	■	■	■			■			■	■			■	ISVC
Hohman, et al.	2017	■		■	■			■					■				■	■	■		■	CHI
Kahng, et al.	2018	■	■			■	■		■		■	■	■	■	■		■				■	TVCG
Karpathy, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Li, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Liu, et al.	2017	■	■			■			■	■	■		■	■			■				■	TVCG
Liu, et al.	2018	■	■			■			■	■	■		■	■		■	■			■		TVCG
Ming, et al.	2017	■		■		■					■		■				■				■	VAST
Norton & Qi	2017	■	■		■	■	■	■									■	■			■	VizSec
Olah	2014	■			■			■				■			■		■	■			■	Web
Olah, et al.	2018	■			■	■	■	■	■		■	■	■				■	■	■		■	Distill
Pezzotti, et al.	2017	■	■			■					■	■	■		■	■	■			■		TVCG
Rauber, et al.	2017	■	■	■		■					■	■	■		■		■			■	■	TVCG
Robinson, et al.	2017	■				■	■				■	■	■				■				■	GeoHum.
Rong, et al.	2016	■	■			■	■				■		■				■				■	ICML VIS
Smilkov, et al.	2016	■					■				■	■	■		■		■				■	NIPS Workshop
Smilkov, et al.	2017	■	■		■			■	■	■	■			■		■			■	■	■	ICML VIS
Strobelt, et al.	2017	■	■			■	■				■	■	■		■		■				■	TVCG
Tzeng & Ma	2005	■				■			■	■			■	■		■					■	VIS
Wang, et al.	2018	■	■	■		■				■	■		■		■	■	■			■		TVCG
Webster, et al.	2017			■				■									■	■		■	■	Web
Wongsuphasawat, et al.	2018		■			■			■				■	■								TVCG
Yosinski, et al.	2015	■		■			■	■	■	■	■						■	■	■		■	ICML DL
Zahavy, et al.	2016	■	■			■					■	■	■		■		■				■	ICML
Zeiler, et al.	2014	■	■			■				■	■								■		■	ECCV
Zeng, et al.	2017	■		■		■			■		■						■			■		VADL
Zhong, et al.	2017	■	■			■					■	■	■		■	■	■		■	■		ICML VIS
Zhu, et al.	2016	■				■	■	■				■					■	■	■		■	ECCV

Survey Highlights

		WHY				WHO			WHAT					HOW								WHEN		WHERE
Author	Year	Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Education	Model Developers & Builders	Model Users	Non-experts	Computational Graph & Network Architecture	Learned Model Parameters	Individual Computational Units	Neurons in High-dimensional Space	Aggregated Information	Node-link Diagrams for Network Architecture	Dimensionality Reduction & Scatter Plots	Line Charts for Temporal Metrics	Instance-based Analysis & Exploration	Interactive Experimentation	Algorithms for Attribution & Feature Visualization	During Training	After Training	Publication Venue
Abadi, et al.	2016	■	■	■		■	■						■			■				■	■	arXiv
Bau, et al.	2017	■		■		■					■						■		■		■	CVPR
Bilal, et al.	2017	■	■			■					■		■				■		■	■		TVCG
Bojarski, et al.	2016	■	■			■				■			■				■		■	■	■	arXiv
Bruckner	2014	■	■			■			■	■				■			■		■	■		MS Thesis
Carter, et al.	2016	■			■	■	■	■			■	■	■				■	■			■	Distill
Cashman, et al.	2017	■	■			■	■			■	■						■				■	VADL
Chae, et al.	2017	■	■			■					■		■			■	■			■		VADL
Chung, et al.	2016	■	■			■			■	■	■	■		■	■	■	■			■		FILM
Goyal, et al.	2016	■						■		■							■	■	■		■	arXiv
Harley	2015	■			■			■	■	■	■			■			■	■			■	ISVC
Hohman, et al.	2017	■		■	■			■					■				■	■	■		■	CHI
Kahng, et al.	2018	■	■			■	■		■		■	■	■	■	■		■				■	TVCG
Karpathy, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Li, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Liu, et al.	2017	■	■			■			■	■	■		■	■			■				■	TVCG
Liu, et al.	2018	■	■			■			■	■	■		■	■		■	■			■		TVCG
Ming, et al.	2017	■		■		■					■		■				■				■	VAST
Norton & Qi	2017	■	■		■	■	■	■									■	■			■	VizSec
Olah	2014	■			■			■				■			■		■	■			■	Web
Olah, et al.	2018	■			■	■	■	■	■		■	■	■				■	■	■		■	Distill
Pezzotti, et al.	2017	■	■			■					■	■	■		■	■	■			■		TVCG
Rauber, et al.	2017	■	■	■		■					■	■	■		■		■			■	■	TVCG
Robinson, et al.	2017	■				■	■				■	■	■				■				■	GeoHum.
Rong, et al.	2016	■	■			■	■				■		■				■				■	ICML VIS
Smilkov, et al.	2016	■					■				■	■	■		■		■				■	NIPS Workshop
Smilkov, et al.	2017	■	■		■			■	■	■	■			■		■		■		■	■	ICML VIS
Strobelt, et al.	2017	■	■			■	■				■	■	■		■		■				■	TVCG
Tzeng & Ma	2005	■				■			■	■			■	■		■					■	VIS
Wang, et al.	2018	■	■	■		■				■	■		■		■	■	■			■		TVCG
Webster, et al.	2017				■			■									■	■		■	■	Web
Wongsuphasawat, et al.	2018		■			■			■				■	■								TVCG
Yosinski, et al.	2015	■			■		■	■	■	■	■						■	■	■		■	ICML DL
Zahavy, et al.	2016	■	■			■					■	■	■		■		■				■	ICML
Zeiler, et al.	2014	■	■			■				■	■								■		■	ECCV
Zeng, et al.	2017	■		■		■			■		■						■			■		VADL
Zhong, et al.	2017	■	■			■					■	■	■		■	■	■		■	■		ICML VIS
Zhu, et al.	2016	■				■	■	■					■				■	■	■		■	ECCV

Survey Highlights

30/38 works designed for **model developers**

WHO

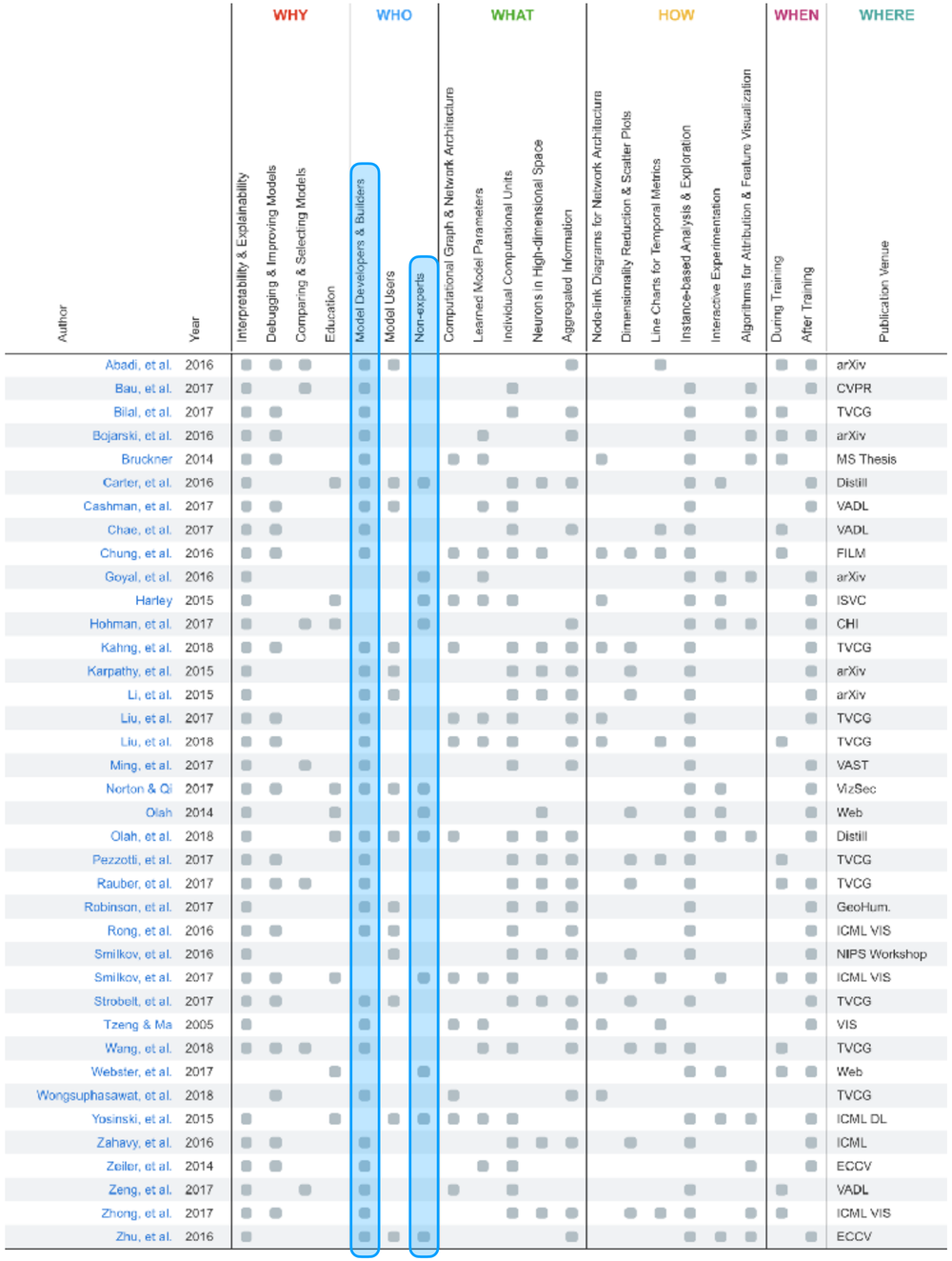
Author	Year	WHY				WHO			WHAT					HOW						WHEN		WHERE
		Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Education	Model Developers & Builders	Model Users	Non-experts	Computational Graph & Network Architecture	Learned Model Parameters	Individual Computational Units	Neurons in High-dimensional Space	Aggregated Information	Node-link Diagrams for Network Architecture	Dimensionality Reduction & Scatter Plots	Line Charts for Temporal Metrics	Instance-based Analysis & Exploration	Interactive Experimentation	Algorithms for Attribution & Feature Visualization	During Training	After Training	Publication Venue
Abadi, et al.	2016	■	■	■		■	■						■			■				■	■	arXiv
Bau, et al.	2017	■		■		■					■		■				■		■		■	CVPR
Bilal, et al.	2017	■	■			■					■		■				■		■	■		TVCG
Bojarski, et al.	2016	■	■			■				■			■				■		■	■	■	arXiv
Bruckner	2014	■	■			■			■	■				■			■		■	■		MS Thesis
Carter, et al.	2016	■			■	■	■	■			■	■	■				■	■			■	Distill
Cashman, et al.	2017	■	■			■	■			■	■						■				■	VADL
Chae, et al.	2017	■	■			■					■		■			■	■			■		VADL
Chung, et al.	2016	■	■			■			■	■	■	■		■	■	■	■			■		FILM
Goyal, et al.	2016	■						■		■							■	■	■		■	arXiv
Harley	2015	■			■			■	■	■	■			■			■	■			■	ISVC
Hohman, et al.	2017	■		■	■			■					■				■	■	■		■	CHI
Kahng, et al.	2018	■	■			■	■		■		■	■	■	■	■		■				■	TVCG
Karpathy, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Li, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Liu, et al.	2017	■	■			■			■	■	■		■	■			■				■	TVCG
Liu, et al.	2018	■	■			■			■	■	■		■	■		■	■			■		TVCG
Ming, et al.	2017	■		■		■					■		■				■				■	VAST
Norton & Qi	2017	■	■		■	■	■	■									■	■			■	VizSec
Olah	2014	■			■			■				■			■		■	■			■	Web
Olah, et al.	2018	■			■	■	■	■	■		■	■	■				■	■	■		■	Distill
Pezzotti, et al.	2017	■	■			■					■	■	■		■	■	■			■		TVCG
Rauber, et al.	2017	■	■	■		■					■	■	■		■		■			■	■	TVCG
Robinson, et al.	2017	■				■	■				■	■	■				■				■	GeoHum.
Rong, et al.	2016	■	■			■	■				■		■				■				■	ICML VIS
Smilkov, et al.	2016	■					■				■	■	■		■		■				■	NIPS Workshop
Smilkov, et al.	2017	■	■		■			■	■	■	■			■		■		■		■	■	ICML VIS
Strobelt, et al.	2017	■	■			■	■				■	■	■		■		■				■	TVCG
Tzeng & Ma	2005	■				■			■	■			■	■		■					■	VIS
Wang, et al.	2018	■	■	■		■				■	■		■		■	■	■			■		TVCG
Webster, et al.	2017				■			■									■	■		■	■	Web
Wongsuphasawat, et al.	2018		■			■			■				■	■								TVCG
Yosinski, et al.	2015	■			■		■	■	■	■	■						■	■	■		■	ICML DL
Zahavy, et al.	2016	■	■			■					■	■	■		■		■				■	ICML
Zeiler, et al.	2014	■	■			■				■	■								■		■	ECCV
Zeng, et al.	2017	■		■		■			■		■						■			■		VADL
Zhong, et al.	2017	■	■			■					■	■	■		■	■	■		■	■		ICML VIS
Zhu, et al.	2016	■				■	■	■					■				■	■	■		■	ECCV

Survey Highlights

WHO

30 / 38 works designed for **model developers**

11 / 38 works designed for **non-experts**



Author	Year	WHY: Interpretability & Explainability	WHY: Debugging & Improving Models	WHY: Comparing & Selecting Models	WHY: Education	WHO: Model Developers & Builders	WHO: Model Users	WHO: Non-experts	WHAT: Computational Graph & Network Architecture	WHAT: Learned Model Parameters	WHAT: Individual Computational Units	WHAT: Neurons in High-dimensional Space	WHAT: Aggregated Information	HOW: Node-link Diagrams for Network Architecture	HOW: Dimensionality Reduction & Scatter Plots	HOW: Line Charts for Temporal Metrics	HOW: Instance-based Analysis & Exploration	HOW: Interactive Experimentation	HOW: Algorithms for Attribution & Feature Visualization	WHEN: During Training	WHEN: After Training	WHERE: Publication Venue
Abadi, et al.	2016	■	■	■		■	■						■			■				■	■	arXiv
Bau, et al.	2017	■		■		■					■						■		■		■	CVPR
Bilal, et al.	2017	■	■			■					■		■				■		■	■		TVCG
Bojarski, et al.	2016	■	■			■				■			■				■		■	■	■	arXiv
Bruckner	2014	■	■			■			■	■				■			■		■	■		MS Thesis
Carter, et al.	2016	■			■	■	■	■			■	■	■				■	■			■	Distill
Cashman, et al.	2017	■	■			■	■			■	■						■				■	VADL
Chae, et al.	2017	■	■			■					■		■			■	■			■		VADL
Chung, et al.	2016	■	■			■			■	■	■	■		■	■	■	■			■		FILM
Goyal, et al.	2016	■						■		■							■	■	■		■	arXiv
Harley	2015	■			■			■	■	■	■			■			■	■			■	ISVC
Hohman, et al.	2017	■		■	■			■					■				■	■	■		■	CHI
Kahng, et al.	2018	■	■			■	■		■		■	■	■	■	■		■				■	TVCG
Karpathy, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Li, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Liu, et al.	2017	■	■			■			■	■	■		■	■			■				■	TVCG
Liu, et al.	2018	■	■			■			■	■	■		■	■		■	■			■		TVCG
Ming, et al.	2017	■		■		■					■		■				■				■	VAST
Norton & Qi	2017	■	■		■	■	■	■									■	■			■	VizSec
Olah	2014	■			■			■				■			■		■	■			■	Web
Olah, et al.	2018	■			■	■	■	■	■		■	■	■				■	■	■		■	Distill
Pezzotti, et al.	2017	■	■			■					■	■	■		■	■	■			■		TVCG
Rauber, et al.	2017	■	■	■		■					■	■	■		■		■			■	■	TVCG
Robinson, et al.	2017	■				■	■				■	■	■				■				■	GeoHum.
Rong, et al.	2016	■	■			■	■				■		■				■				■	ICML VIS
Smilkov, et al.	2016	■					■				■	■	■		■		■				■	NIPS Workshop
Smilkov, et al.	2017	■	■		■			■	■	■	■			■		■		■		■	■	ICML VIS
Strobelt, et al.	2017	■	■			■	■				■	■	■		■		■				■	TVCG
Tzeng & Ma	2005	■				■			■	■			■	■		■					■	VIS
Wang, et al.	2018	■	■	■		■				■	■		■		■	■	■			■		TVCG
Webster, et al.	2017				■			■									■	■		■	■	Web
Wongsuphasawat, et al.	2018		■			■			■				■	■								TVCG
Yosinski, et al.	2015	■			■		■	■	■	■	■						■	■	■		■	ICML DL
Zahavy, et al.	2016	■	■			■					■	■	■		■		■				■	ICML
Zeiler, et al.	2014	■	■			■				■	■								■		■	ECCV
Zeng, et al.	2017	■		■		■			■		■						■			■		VADL
Zhong, et al.	2017	■	■			■					■	■	■		■	■	■		■	■		ICML VIS
Zhu, et al.	2016	■				■	■	■					■				■	■	■		■	ECCV

Survey Highlights

WHO

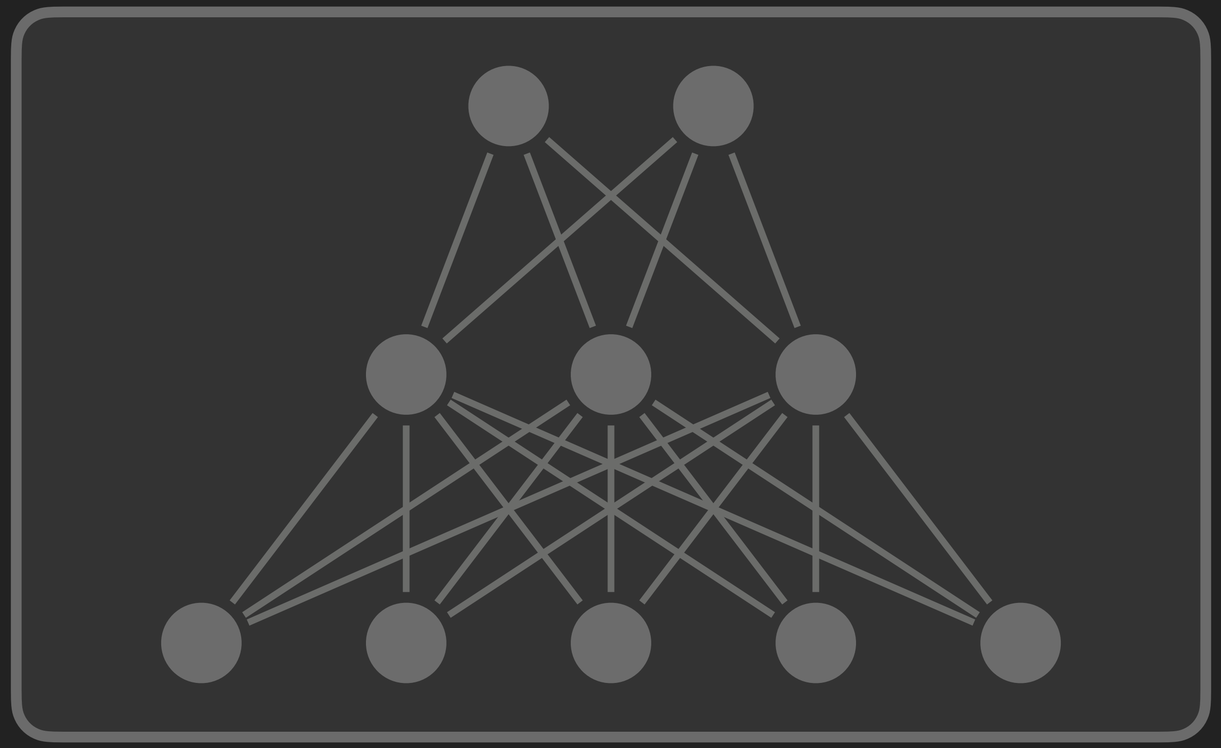
30 / 38 works designed for **model developers**

11 / 38 works designed for **non-experts**

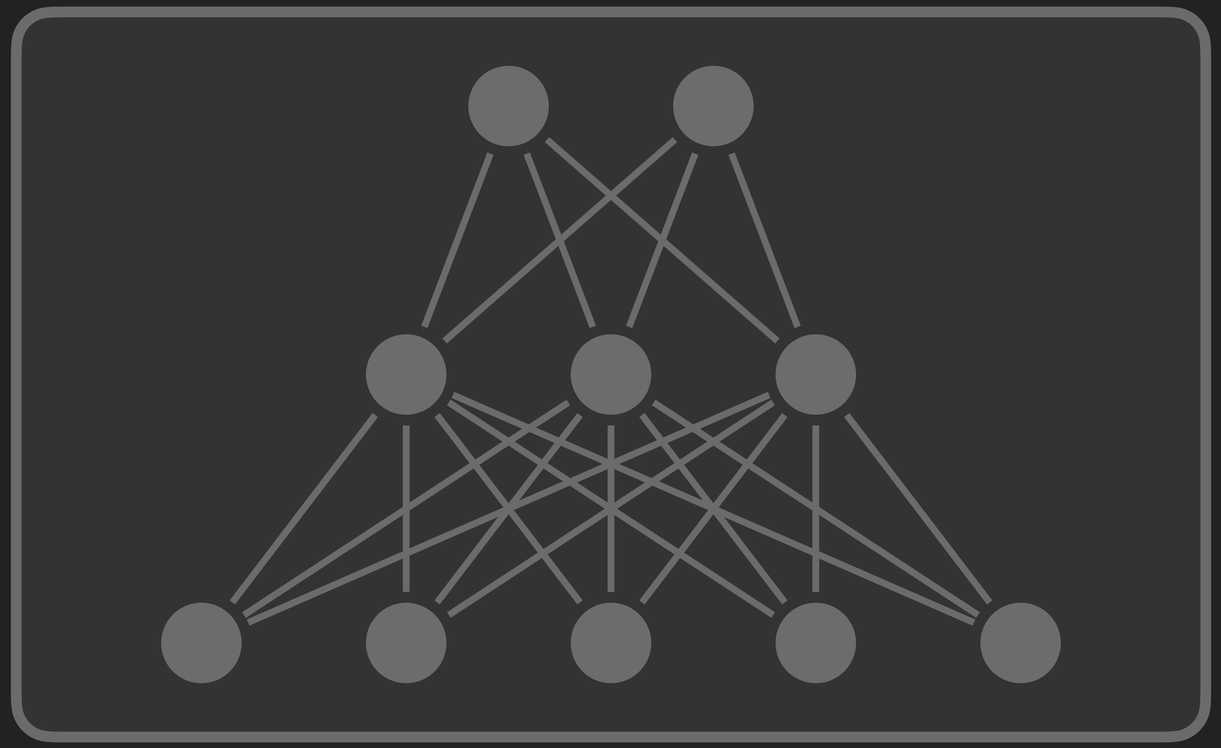
HOW

33 / 38 works use **instance-based analysis**

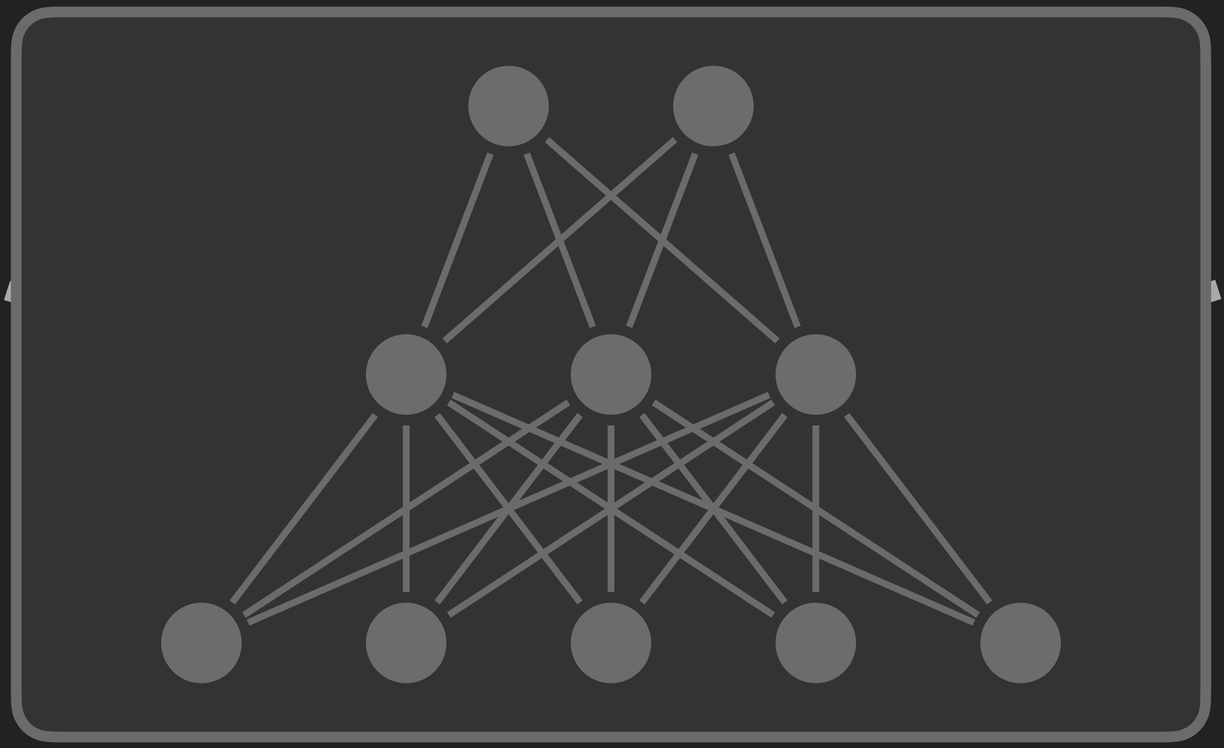
		WHY				WHO			WHAT					HOW							WHEN		WHERE
Author	Year	Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Education	Model Developers & Builders	Model Users	Non-experts	Computational Graph & Network Architecture	Learned Model Parameters	Individual Computational Units	Neurons in High-dimensional Space	Aggregated Information	Node-link Diagrams for Network Architecture	Dimensionality Reduction & Scatter Plots	Line Charts for Temporal Metrics	Instance-based Analysis & Exploration	Interactive Experimentation	Algorithms for Attribution & Feature Visualization	During Training	After Training	Publication Venue
Abadi, et al.	2016	■	■	■		■	■						■			■				■	■	arXiv
Bau, et al.	2017	■		■		■					■		■				■		■		■	CVPR
Bilal, et al.	2017	■	■			■					■		■				■		■	■		TVCG
Bojarski, et al.	2016	■	■			■				■			■				■		■	■	■	arXiv
Bruckner	2014	■	■			■			■	■				■			■		■	■		MS Thesis
Carter, et al.	2016	■			■	■	■	■			■	■	■				■	■			■	Distill
Cashman, et al.	2017	■	■			■	■			■	■						■				■	VADL
Chae, et al.	2017	■	■			■					■		■			■	■			■		VADL
Chung, et al.	2016	■	■			■			■	■	■	■		■	■	■	■			■		FILM
Goyal, et al.	2016	■						■		■							■	■	■		■	arXiv
Harley	2015	■			■			■	■	■	■			■			■	■			■	ISVC
Hohman, et al.	2017	■		■	■			■					■				■	■	■		■	CHI
Kahng, et al.	2018	■	■			■	■		■		■	■	■	■	■		■				■	TVCG
Karpathy, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Li, et al.	2015	■				■	■				■	■	■		■		■				■	arXiv
Liu, et al.	2017	■	■			■			■	■	■		■	■			■				■	TVCG
Liu, et al.	2018	■	■			■			■	■	■		■	■		■	■			■		TVCG
Ming, et al.	2017	■		■		■					■		■				■				■	VAST
Norton & Qi	2017	■	■		■	■	■	■									■	■			■	VizSec
Olah	2014	■			■			■				■			■		■	■			■	Web
Olah, et al.	2018	■			■	■	■	■	■		■	■	■				■	■	■		■	Distill
Pezzotti, et al.	2017	■	■			■					■	■	■		■	■	■			■		TVCG
Rauber, et al.	2017	■	■	■		■					■	■	■		■		■			■	■	TVCG
Robinson, et al.	2017	■				■	■				■	■	■				■				■	GeoHum.
Rong, et al.	2016	■	■			■	■				■		■				■				■	ICML VIS
Smilkov, et al.	2016	■					■				■	■	■		■		■				■	NIPS Workshop
Smilkov, et al.	2017	■	■		■			■	■	■	■			■		■			■	■	■	ICML VIS
Strobelt, et al.	2017	■	■			■	■				■	■	■		■		■				■	TVCG
Tzeng & Ma	2005	■				■			■	■			■	■		■					■	VIS
Wang, et al.	2018	■	■	■		■				■	■		■		■	■	■			■		TVCG
Webster, et al.	2017			■				■									■	■		■	■	Web
Wongsuphasawat, et al.	2018		■			■			■				■	■								TVCG
Yosinski, et al.	2015	■		■		■	■	■	■	■	■						■	■	■		■	ICML DL
Zahavy, et al.	2016	■	■			■					■	■	■		■		■				■	ICML
Zeiler, et al.	2014	■	■			■				■	■						■		■		■	ECCV
Zeng, et al.	2017	■		■		■			■		■						■			■		VADL
Zhong, et al.	2017	■	■			■					■	■	■		■	■	■		■	■		ICML VIS
Zhu, et al.	2016	■				■	■	■					■				■	■	■		■	ECCV



Neural Network

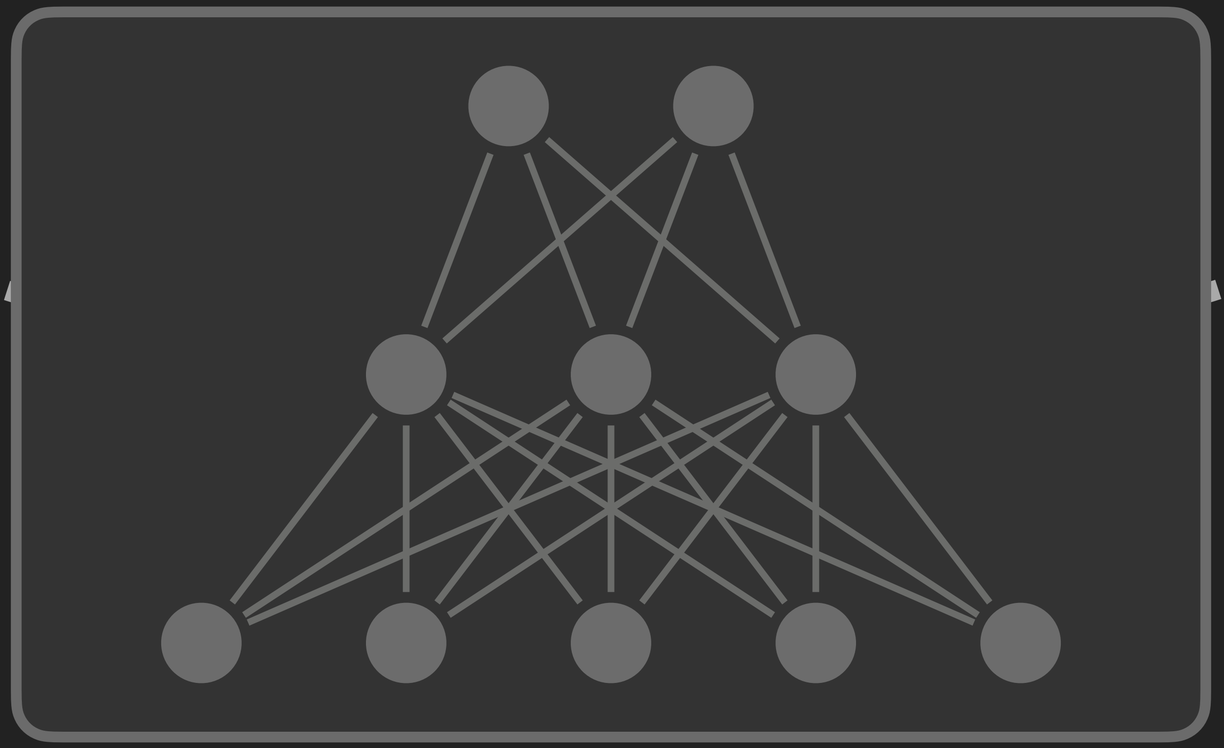


Neural Network



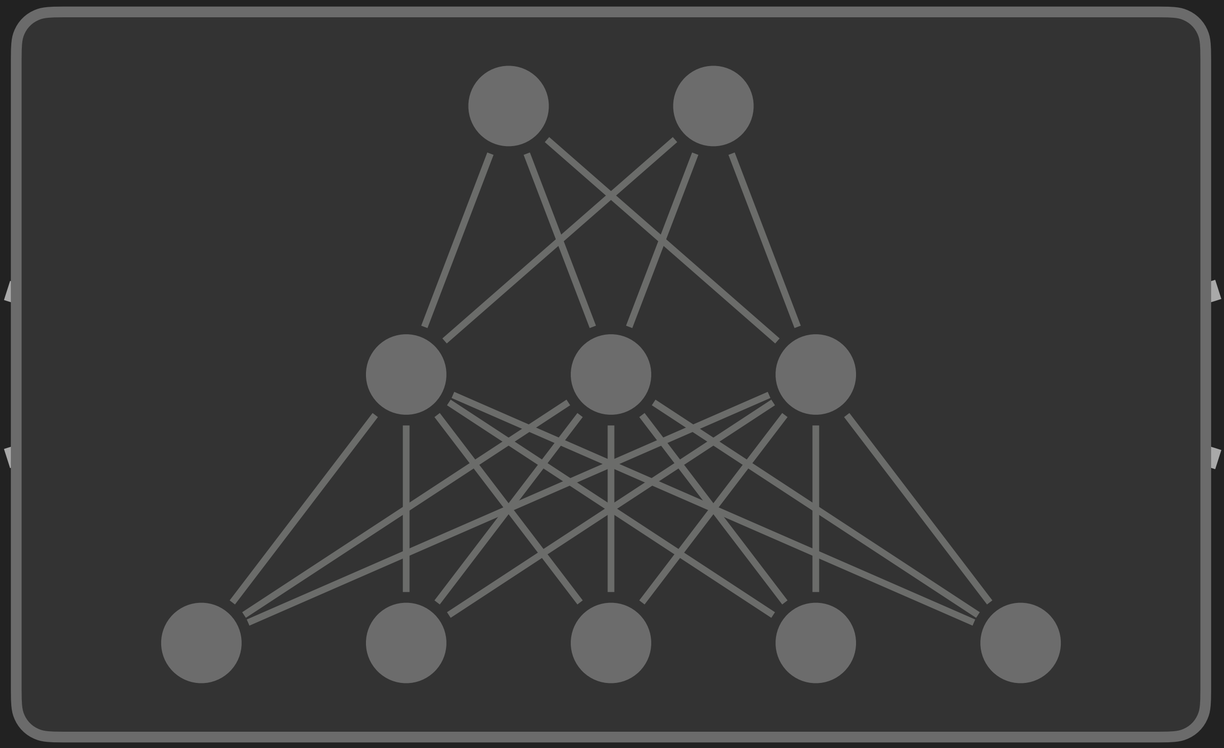
Neural Network

bike ✓



Neural Network

bike ✓



Neural Network

bike ✓

truck ✗

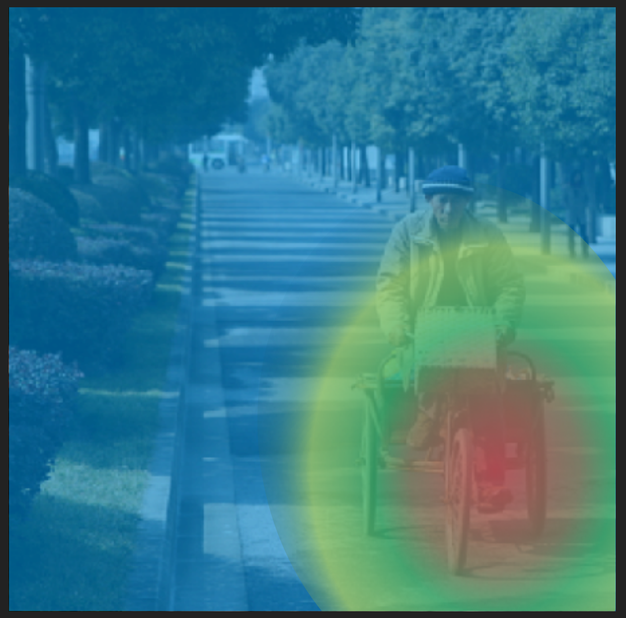
?

truck ✗

?

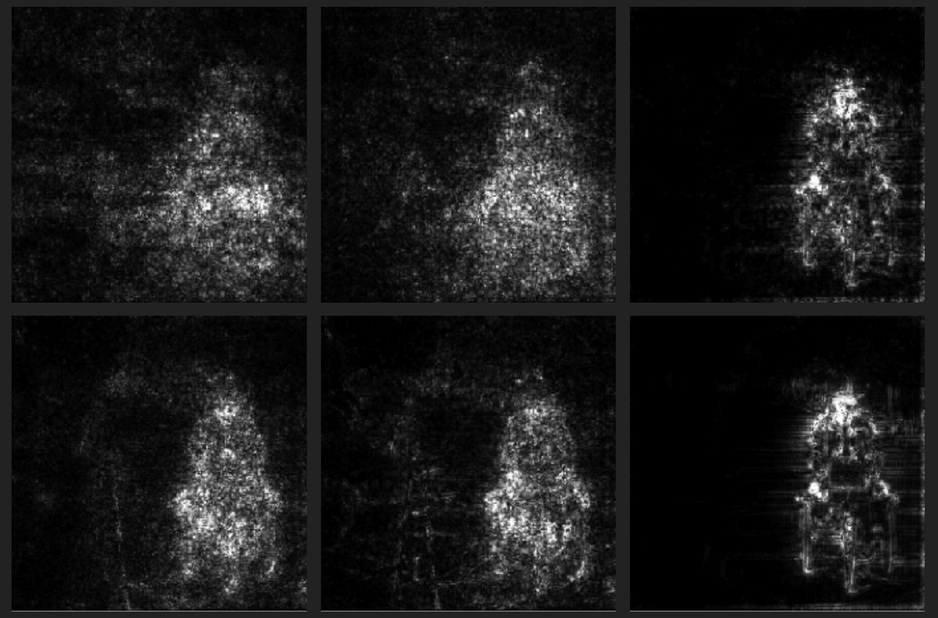
truck ✗

Attention

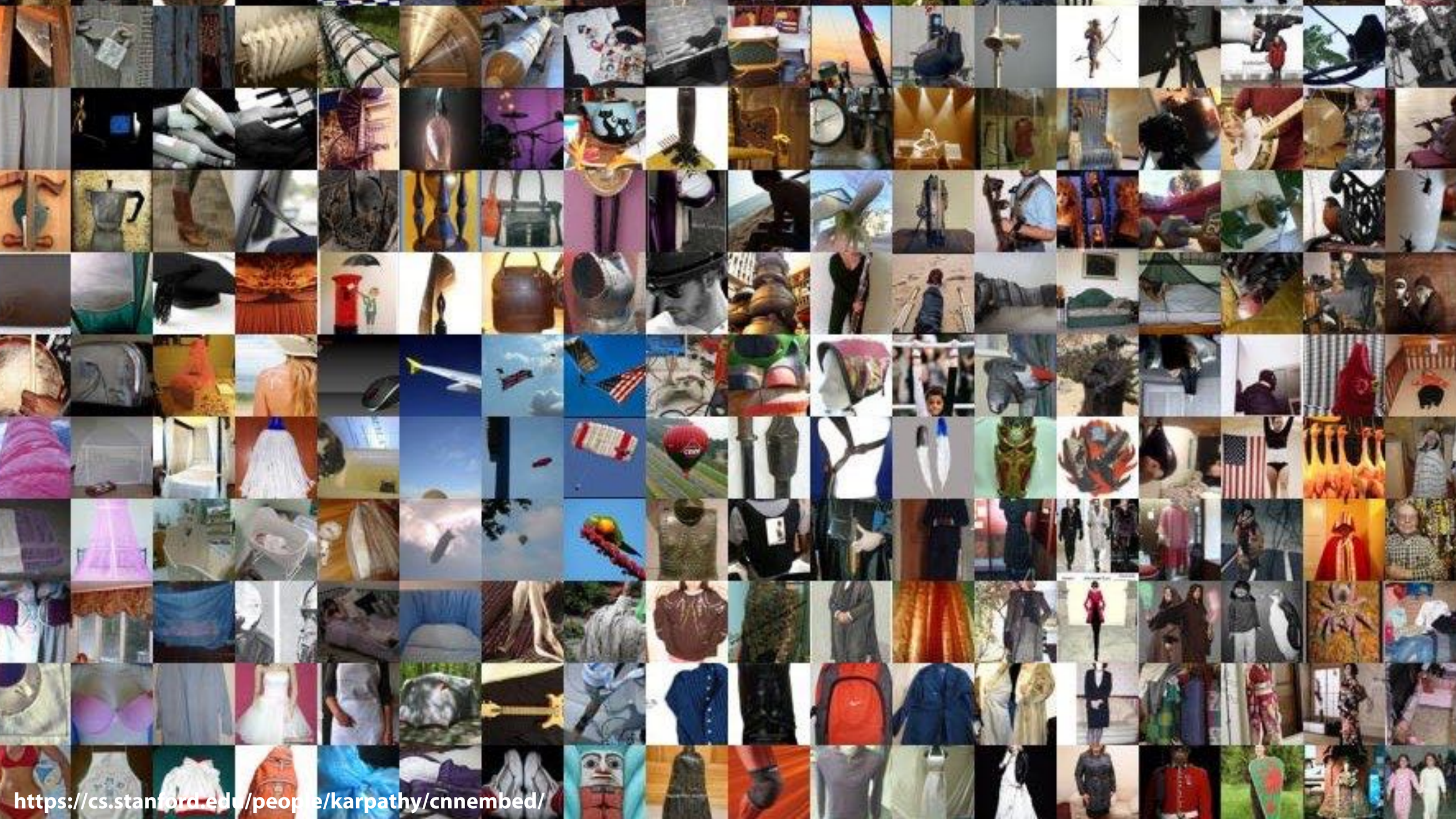


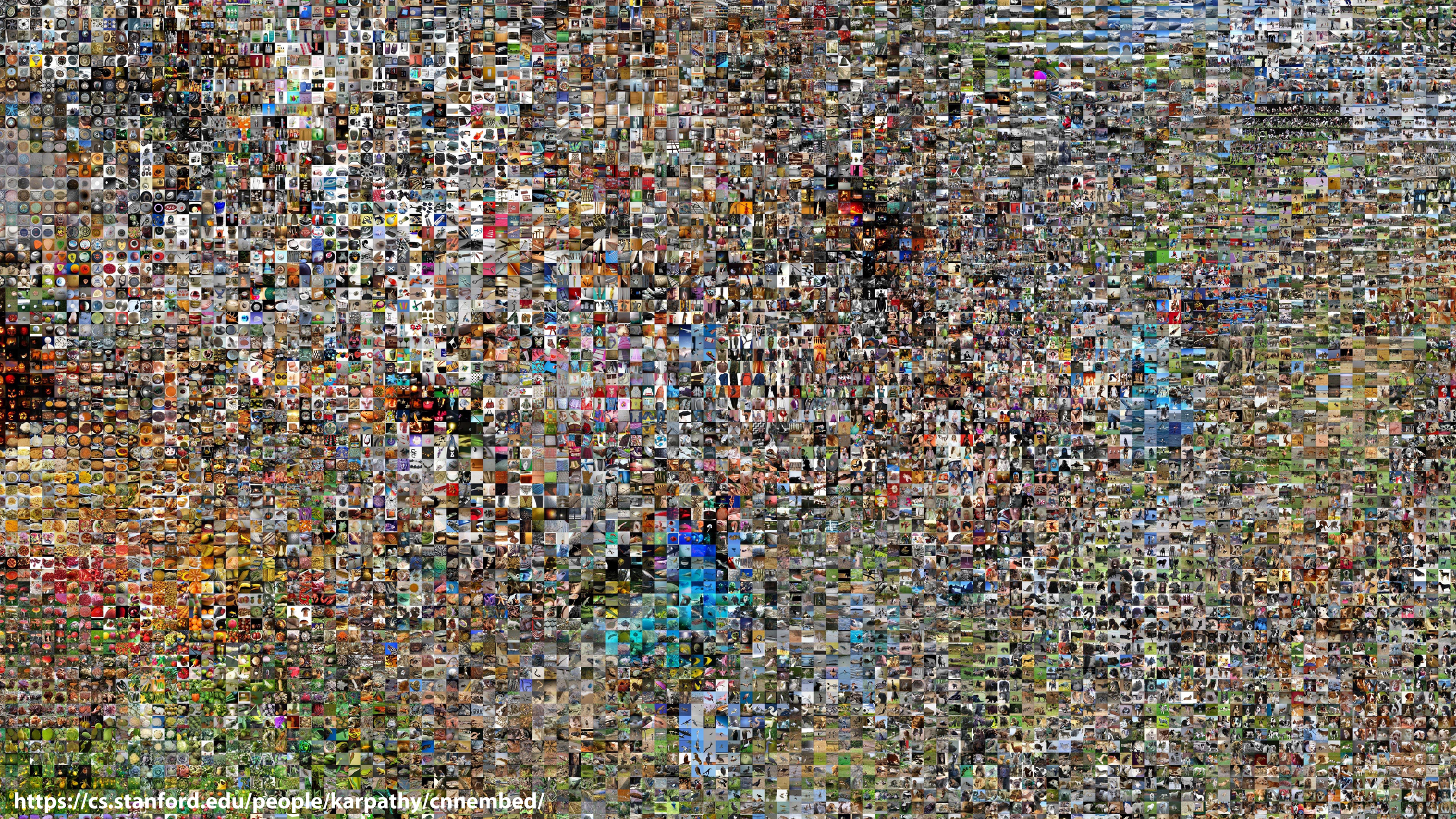
[Selvaraju, et al., ICCV, 2017]

Sensitivity



[Smilkov, et al., arXiv, 2017]


https://cs.stanford.edu/people/karpathy/cnnembed/


https://cs.stanford.edu/people/karpathy/cnnembed/

SUMMIT

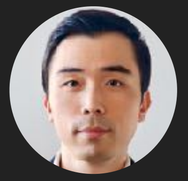
Scaling Deep Learning Interpretability by Visualizing Activation & Attribution Summarizations

VAST 2019

Fred Hohman
Georgia Tech

Haekyu Park
Georgia Tech

Caleb Robinson
Georgia Tech



Polo Chau
Georgia Tech

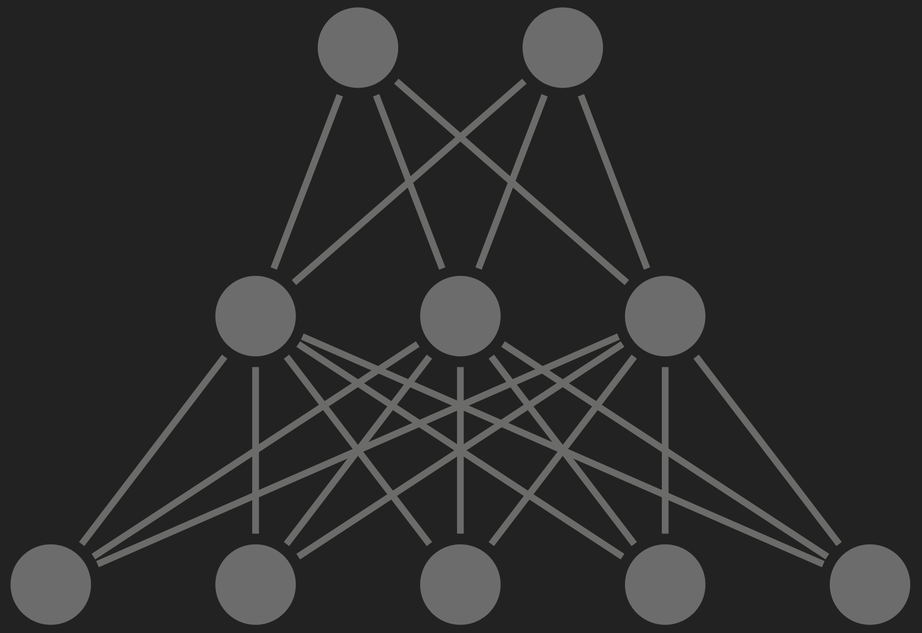
SUMMIT

Scalably summarize and **interactively visualize**
neural network feature representations
for millions of images

SUMMIT

Scalably summarize and **interactively visualize**
neural network feature representations
for millions of images

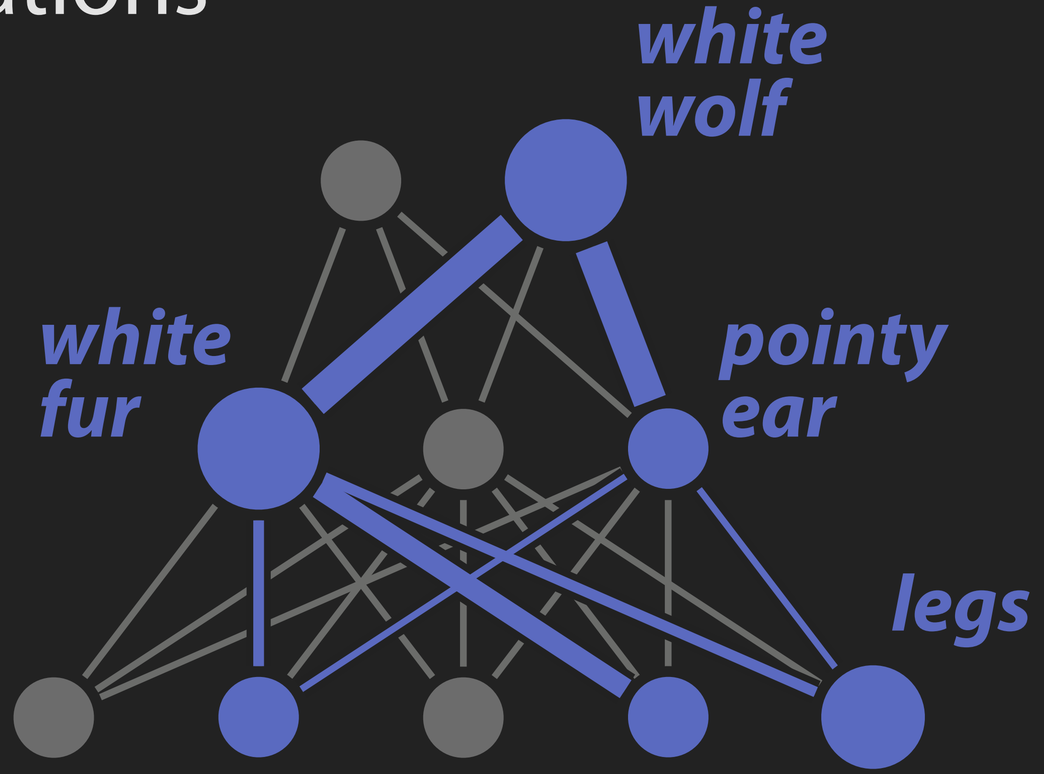
white wolf

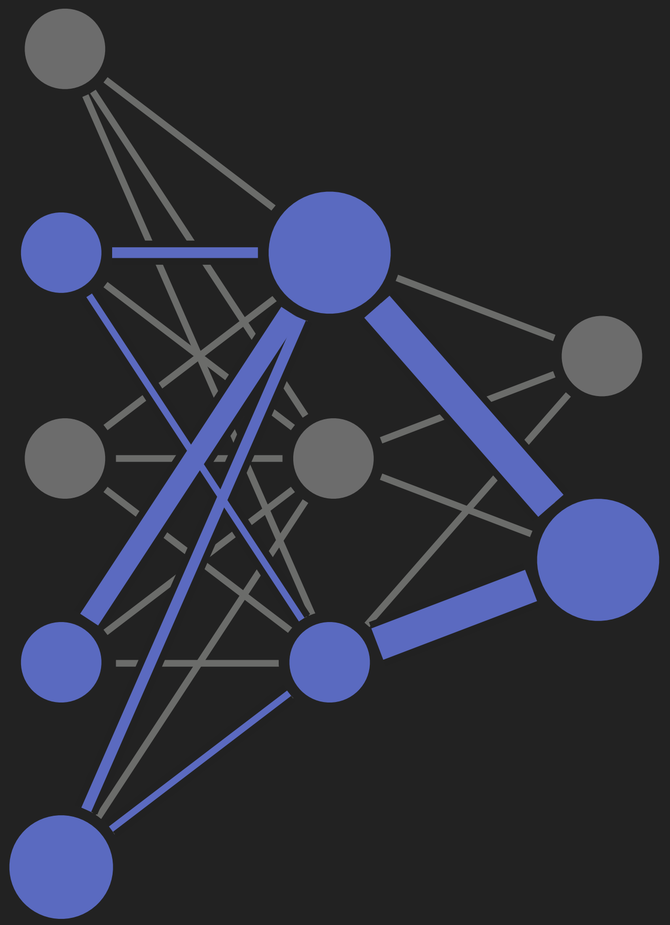


SUMMIT

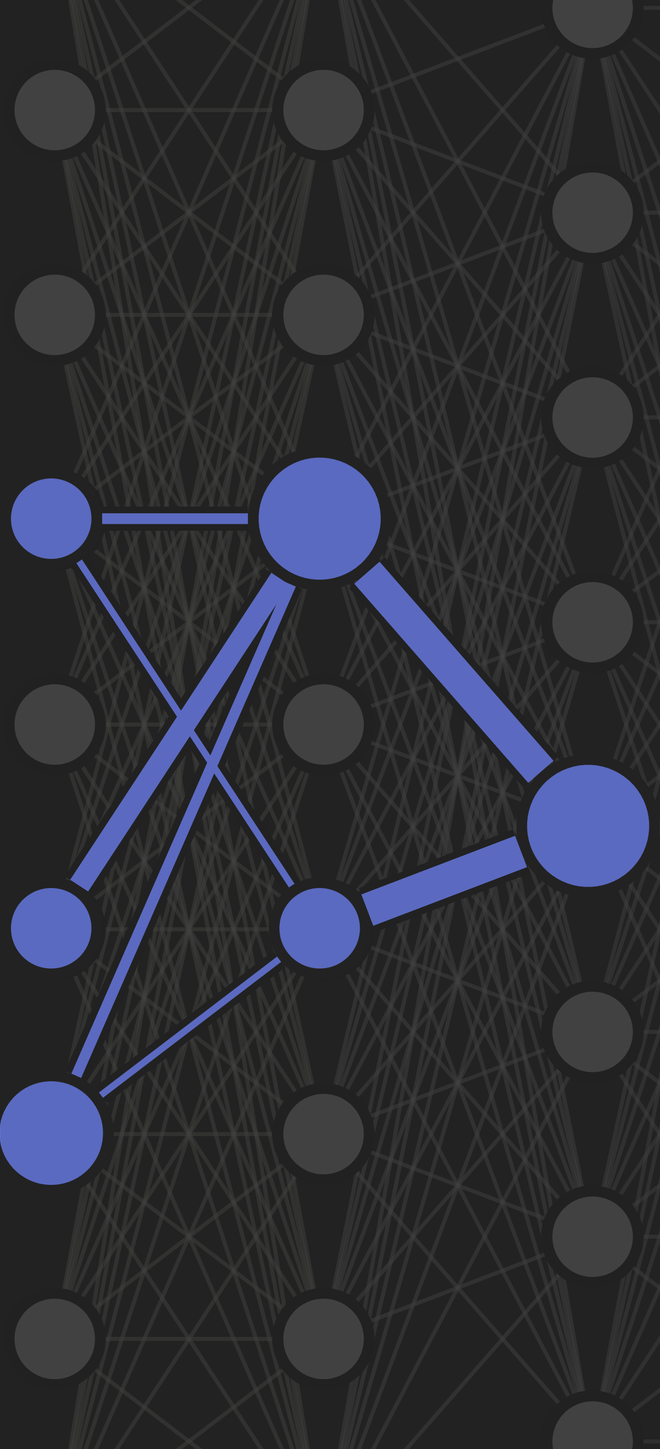
Scalably summarize and **interactively visualize**
neural network feature representations
for millions of images

white wolf



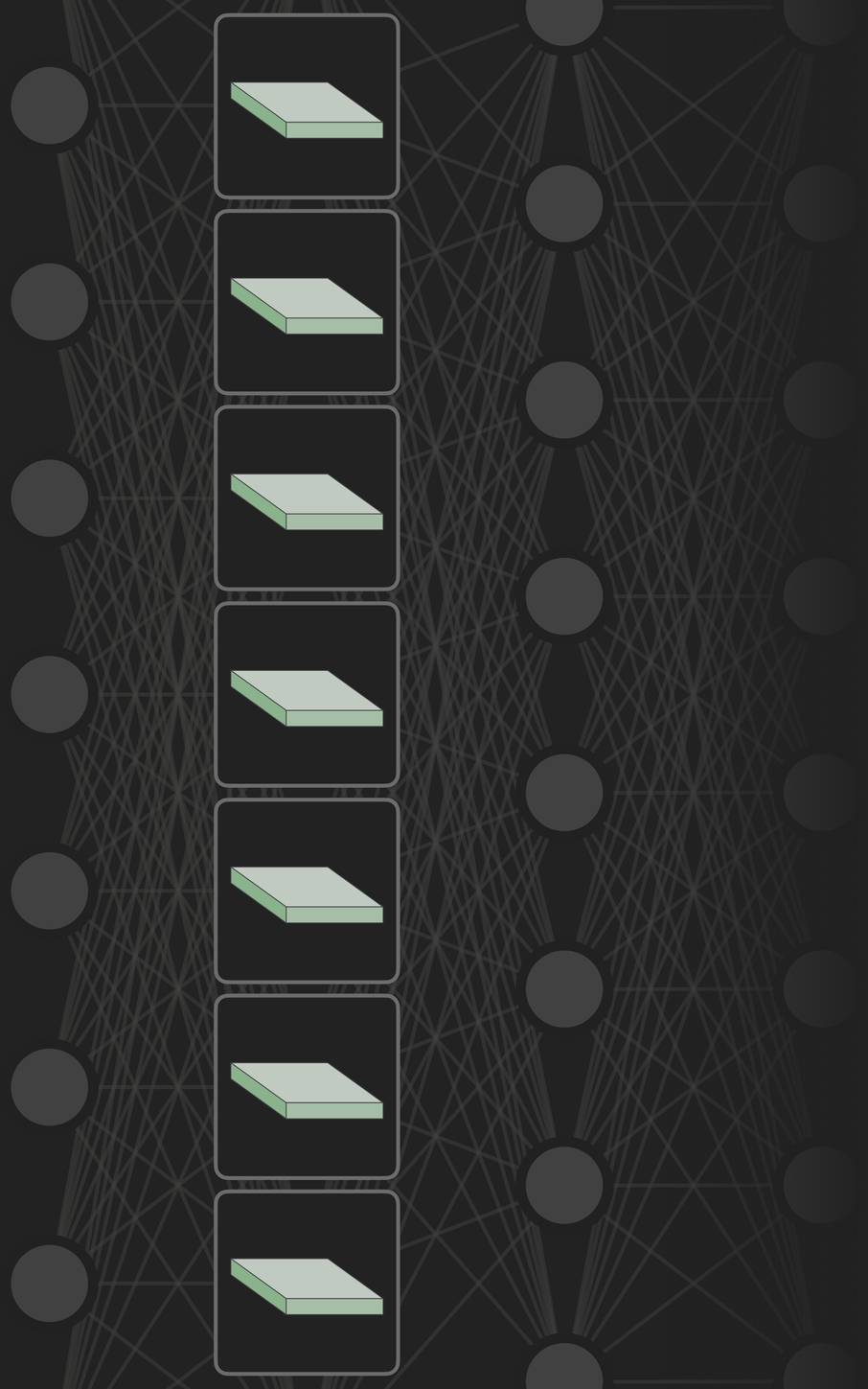


How do we make **attribution graphs**?

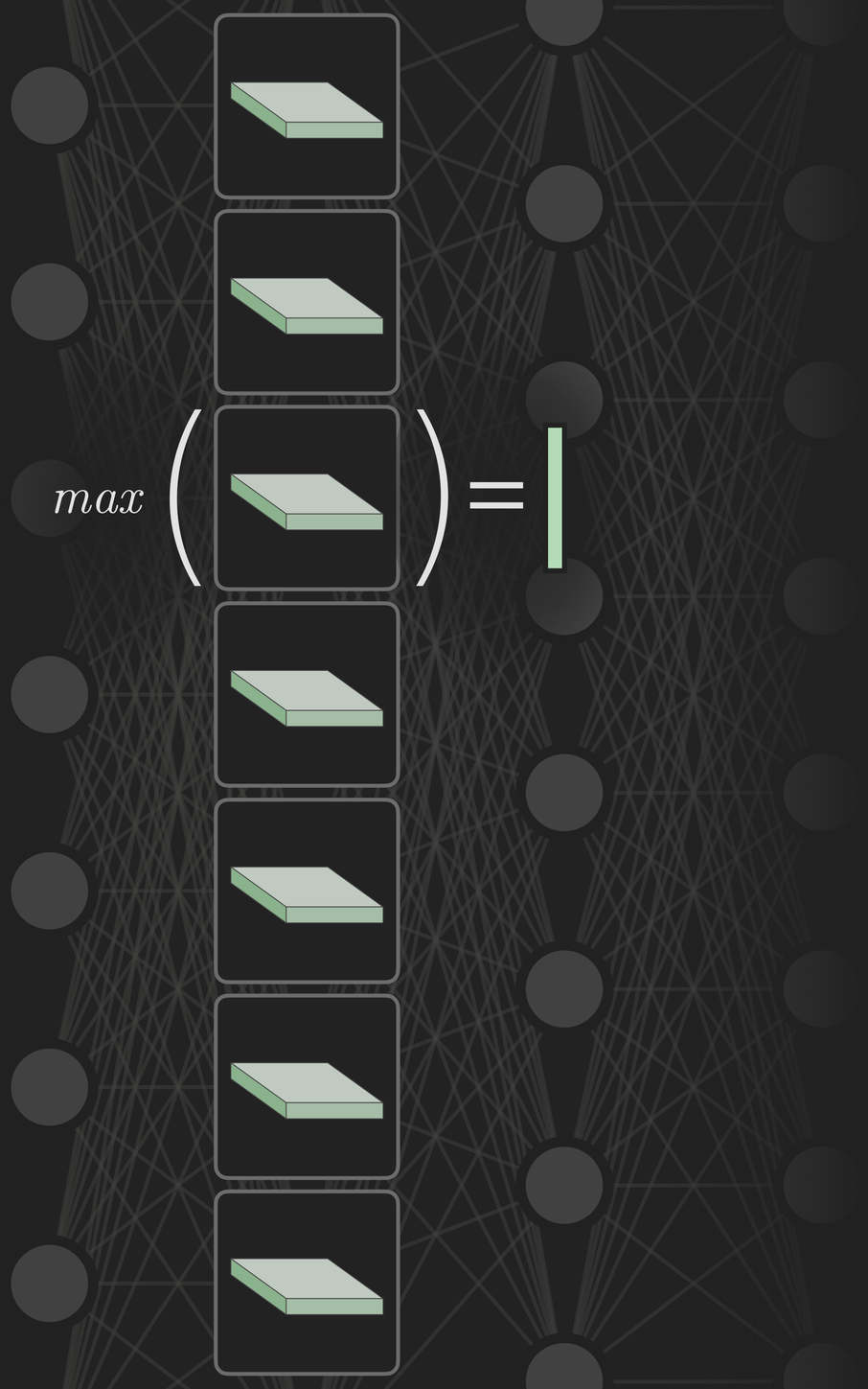


How do we make **attribution graphs**?

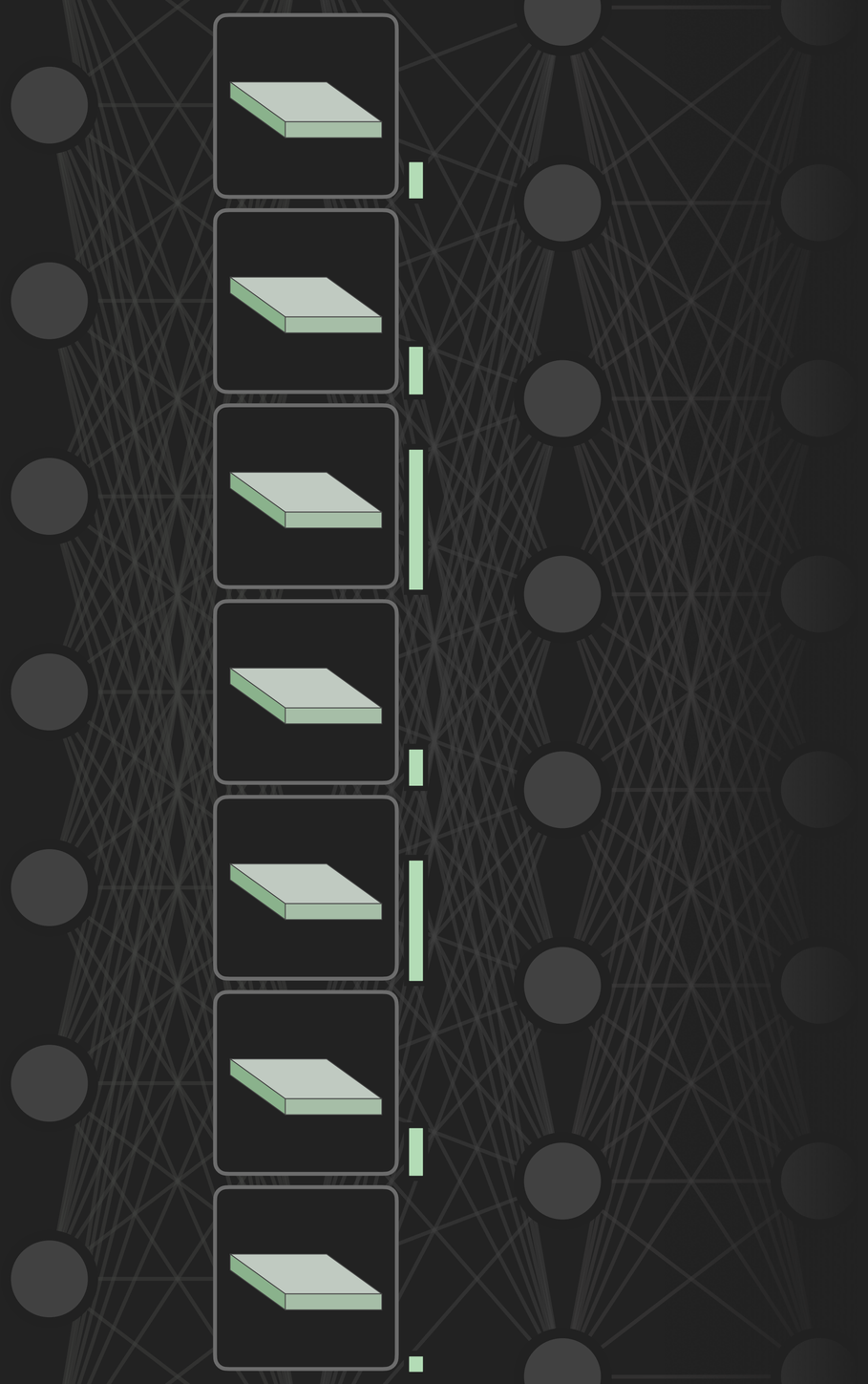
How do we make **attribution graphs**?



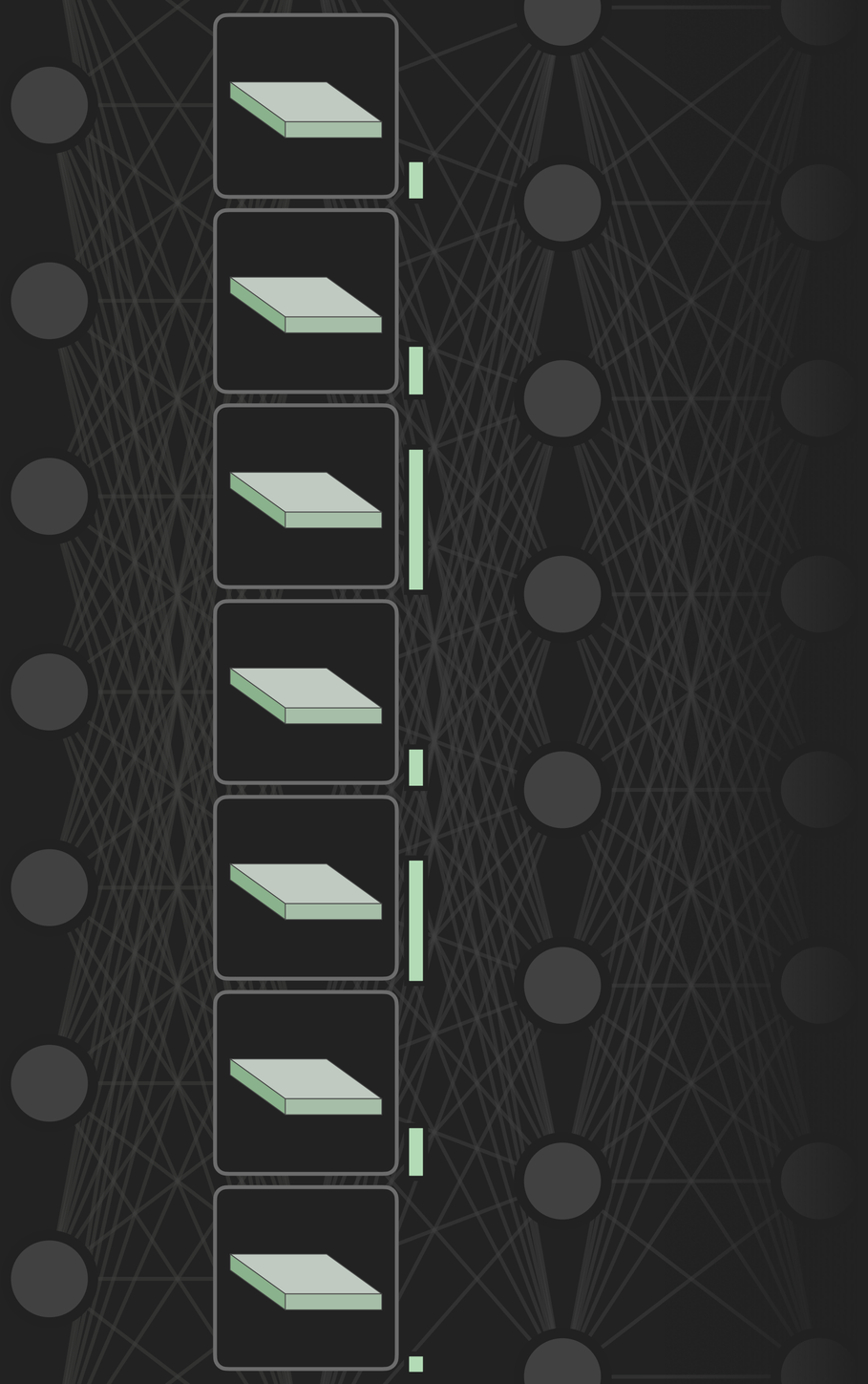
Aggregate network
activations (nodes)

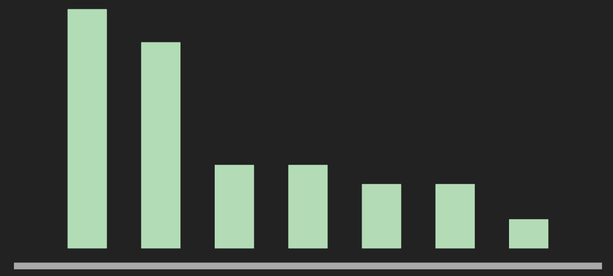


Aggregate network
activations (nodes)

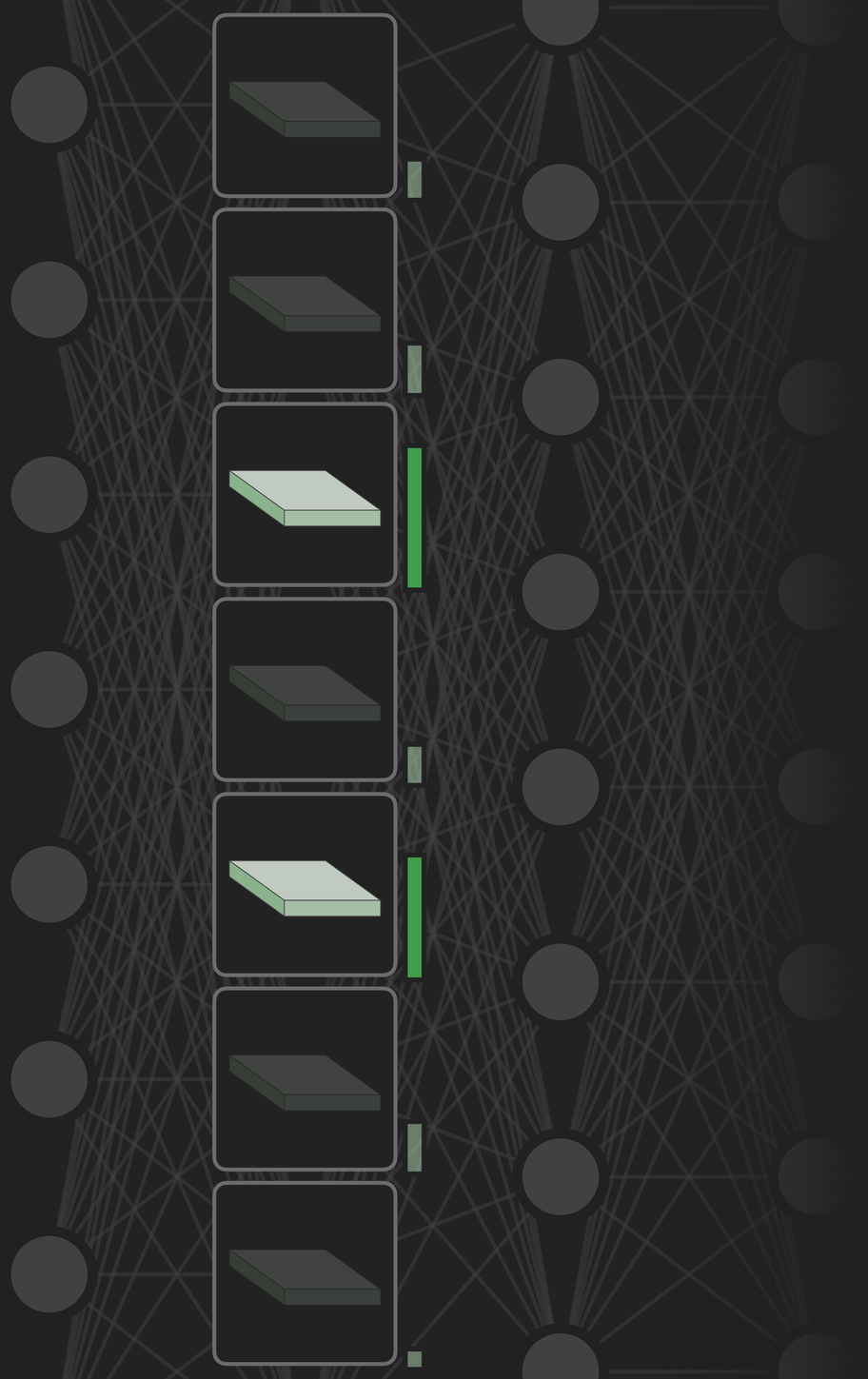


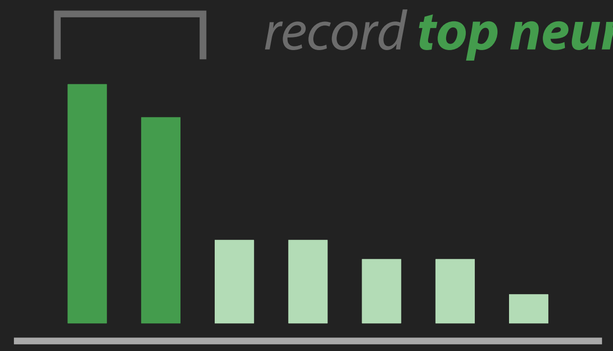
Aggregate network
activations (nodes)



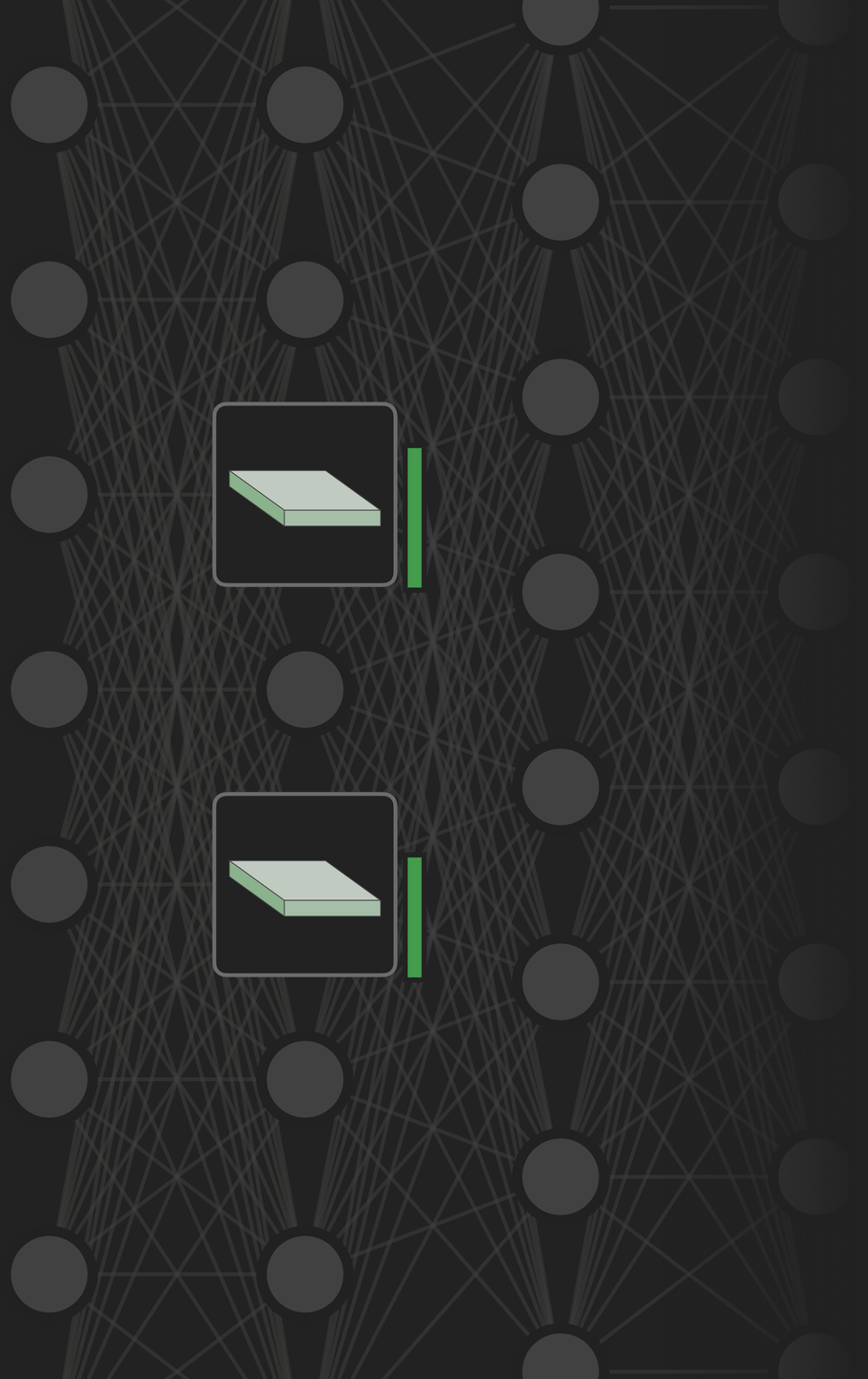


Aggregate network
activations (nodes)

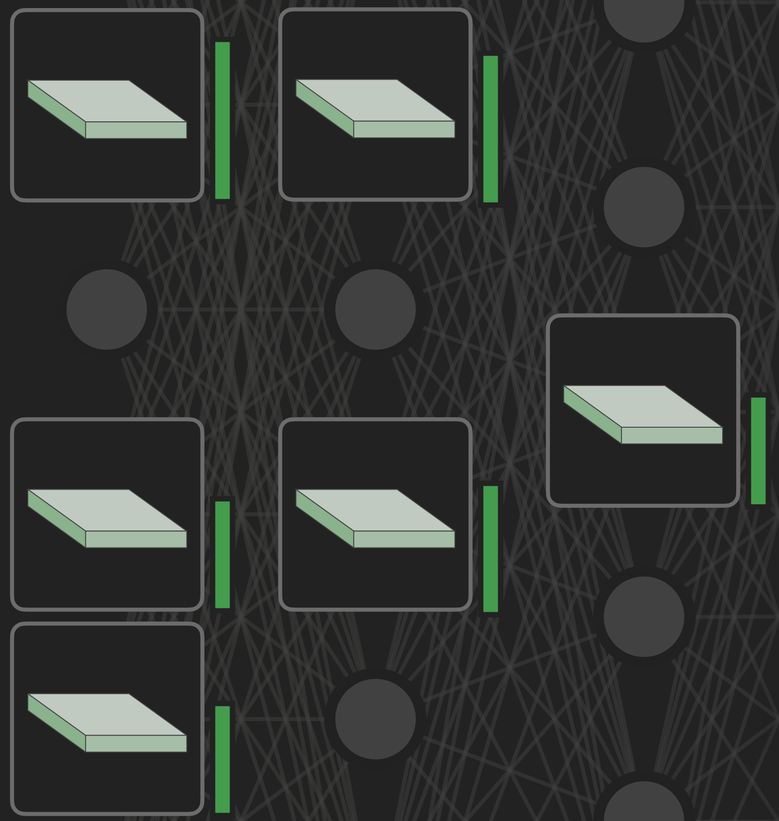




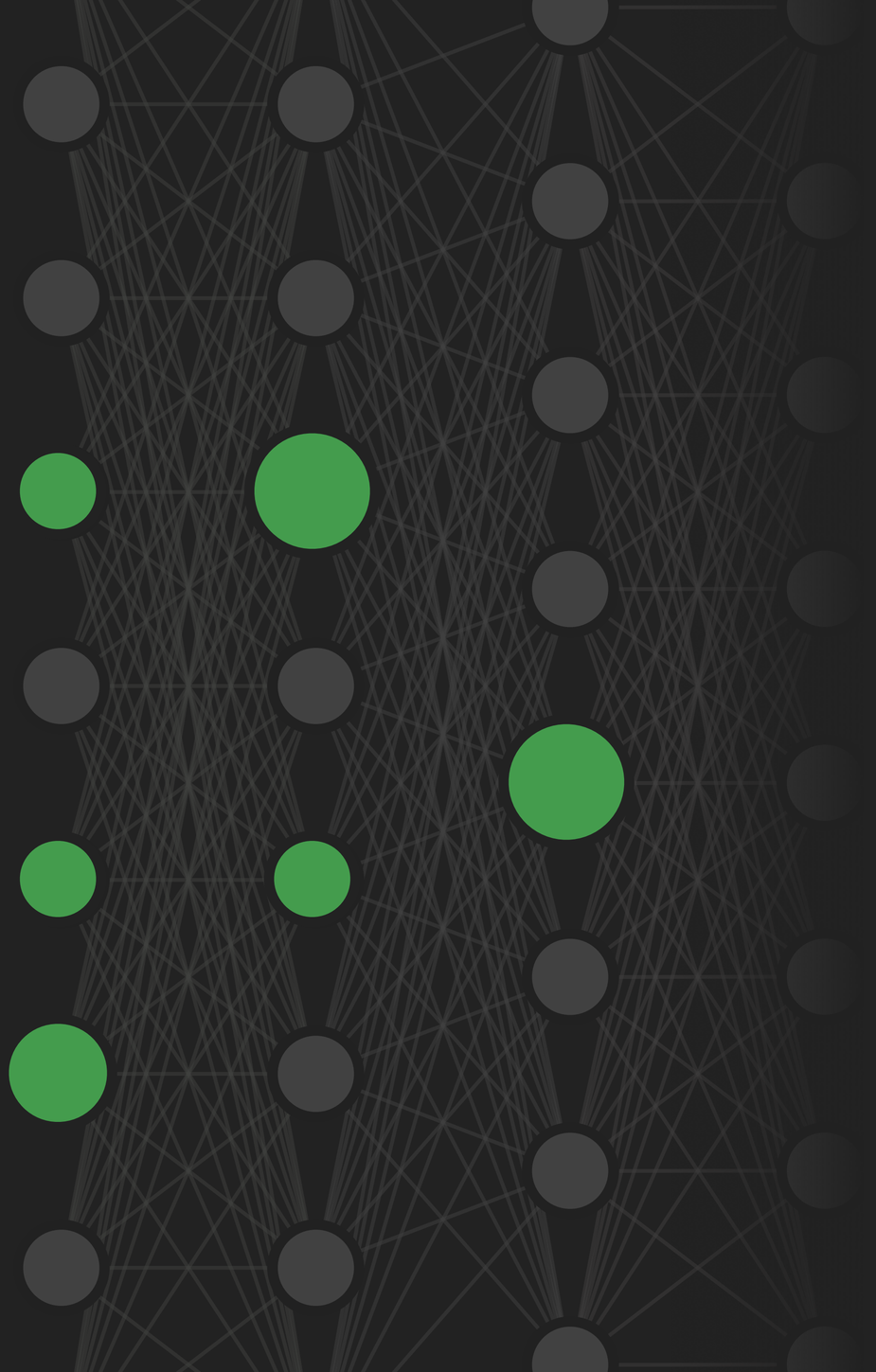
Aggregate network **activations** (nodes)



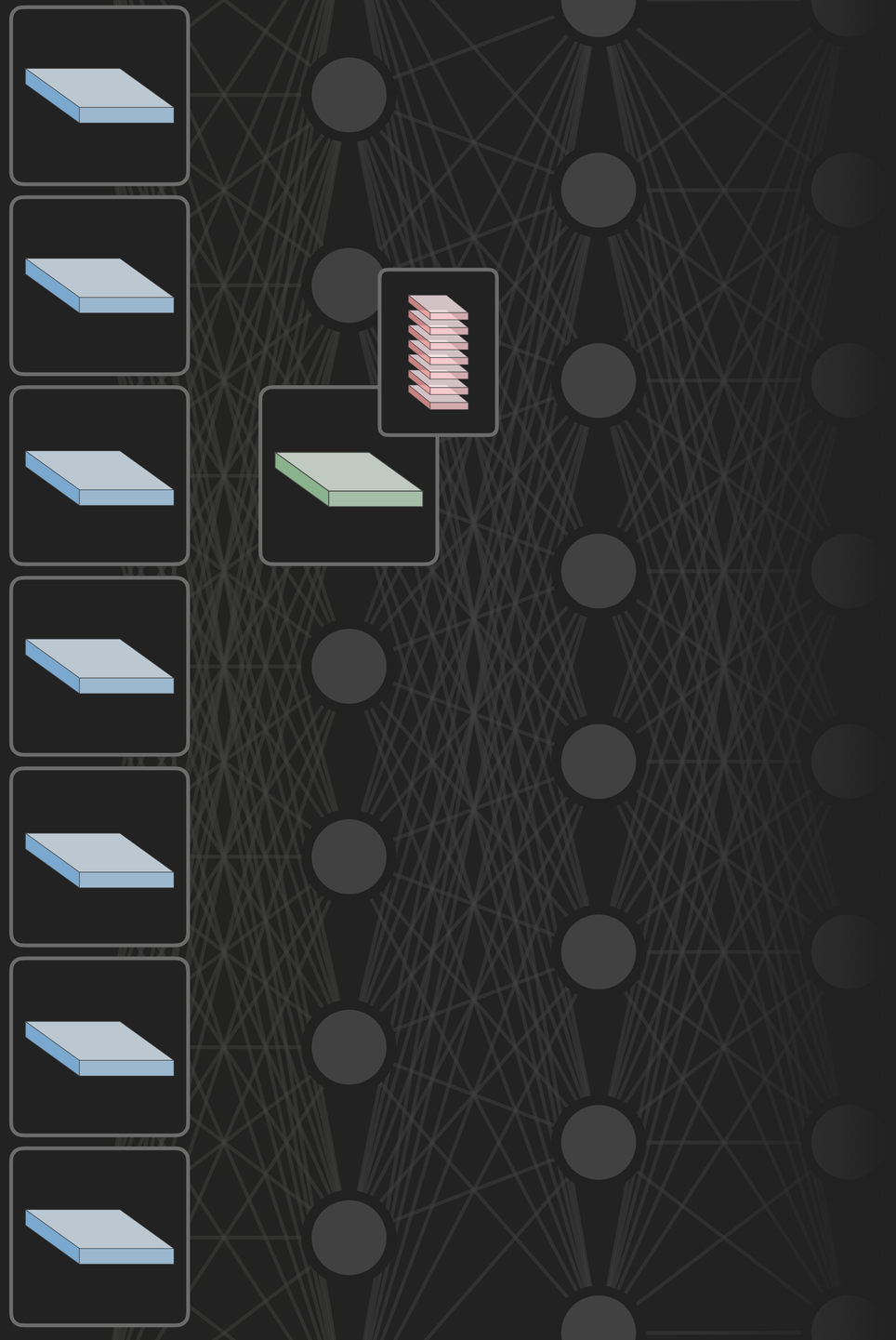
Aggregate network
activations (nodes)



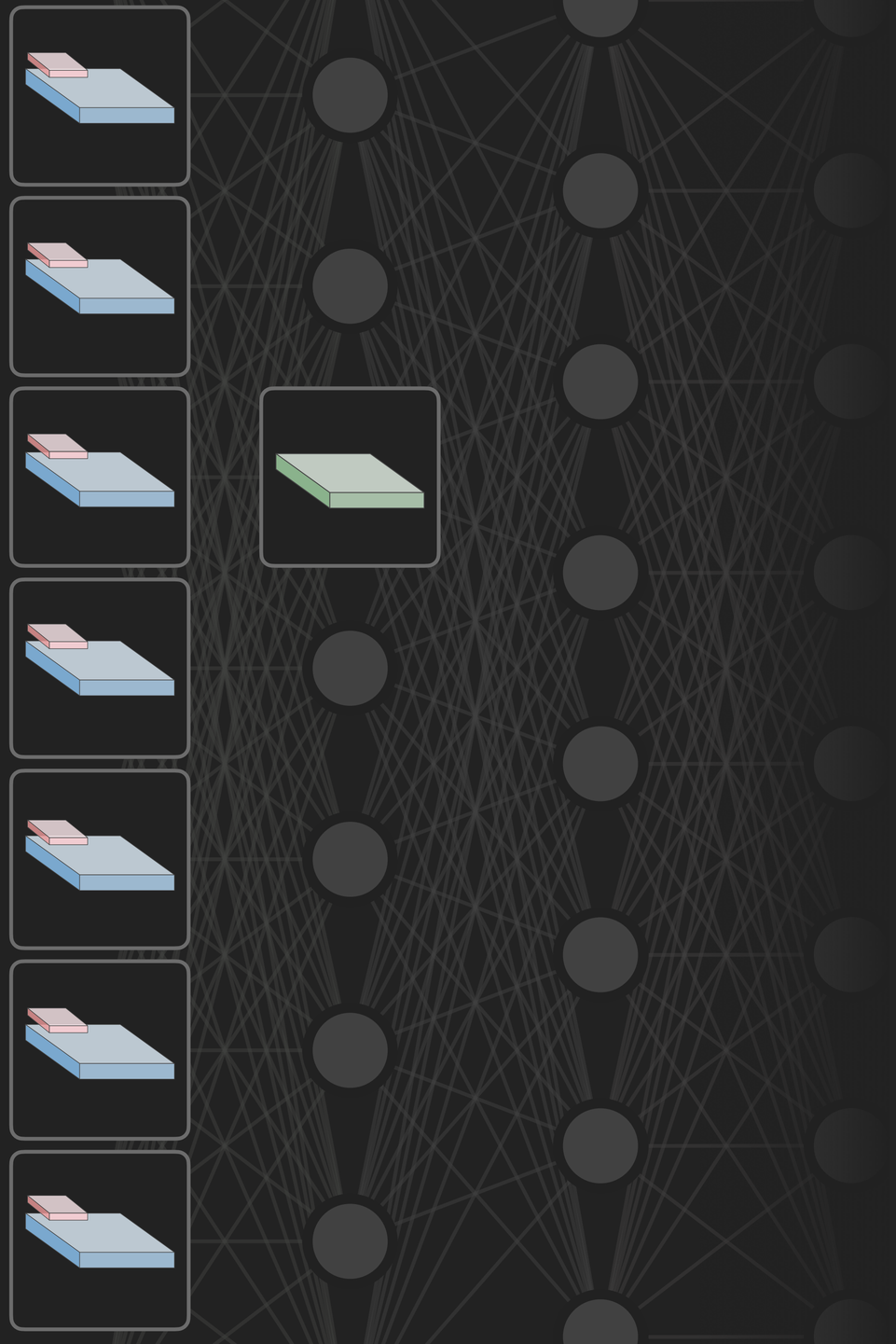
Aggregate network
activations (nodes)



Aggregate network
activations (nodes)



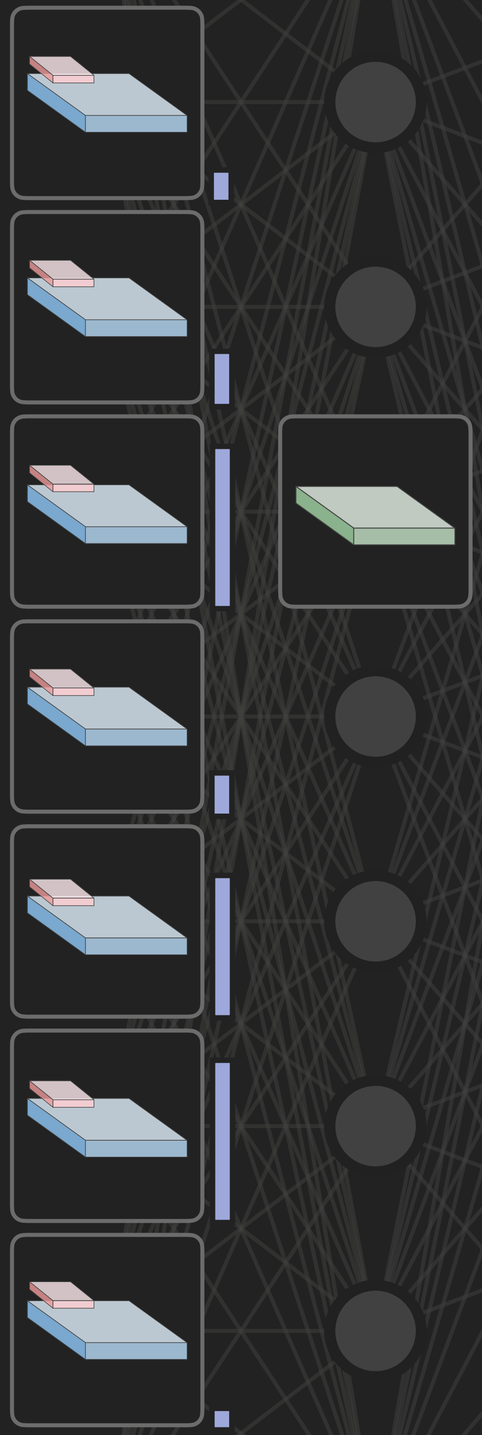
Aggregate network
influences (edges)



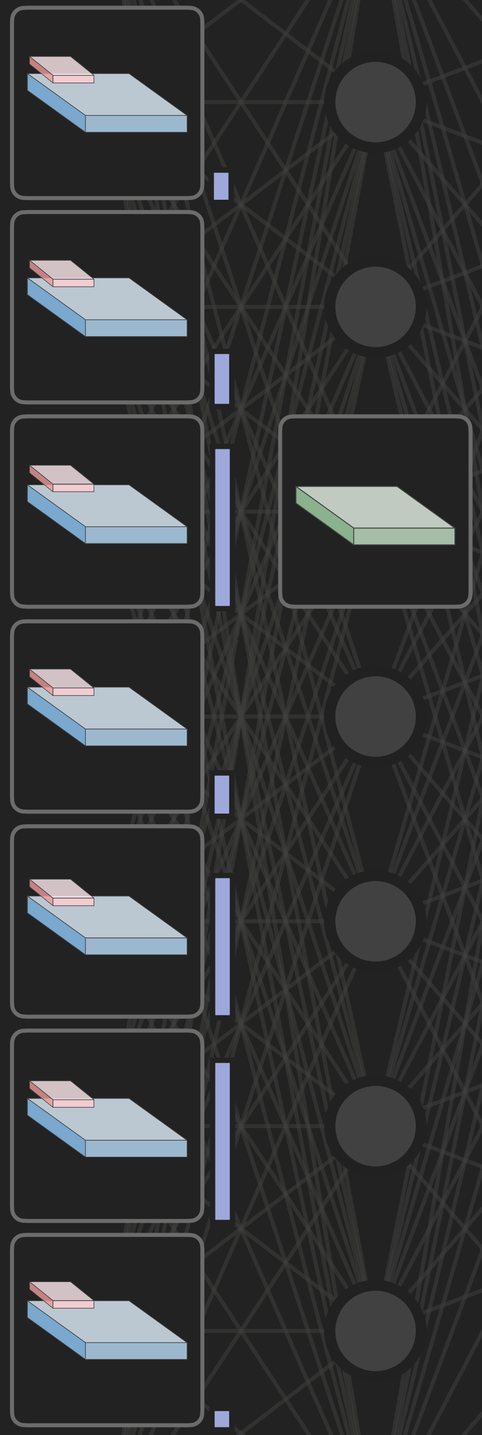
Aggregate network
influences (edges)

max

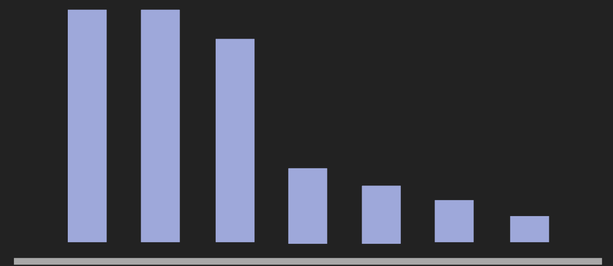
Aggregate network
influences (edges)

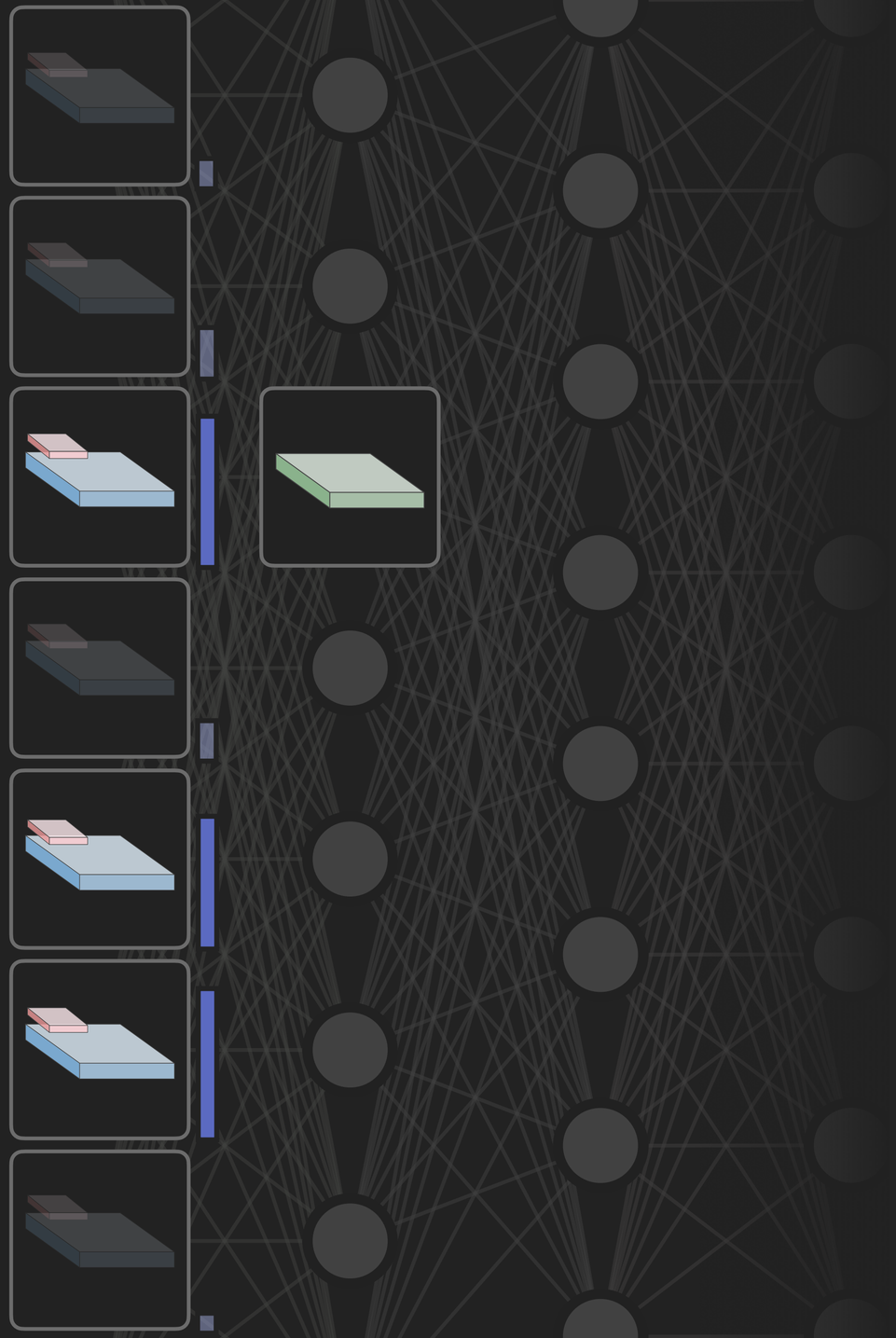


Aggregate network
influences (edges)

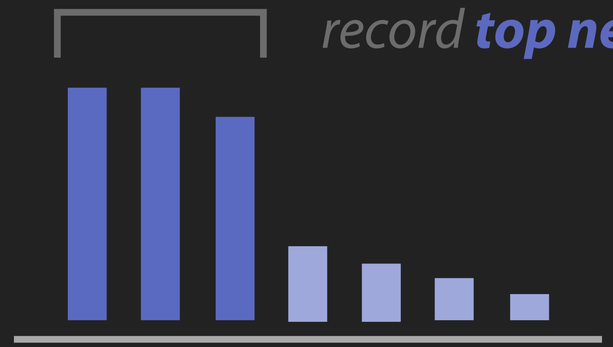


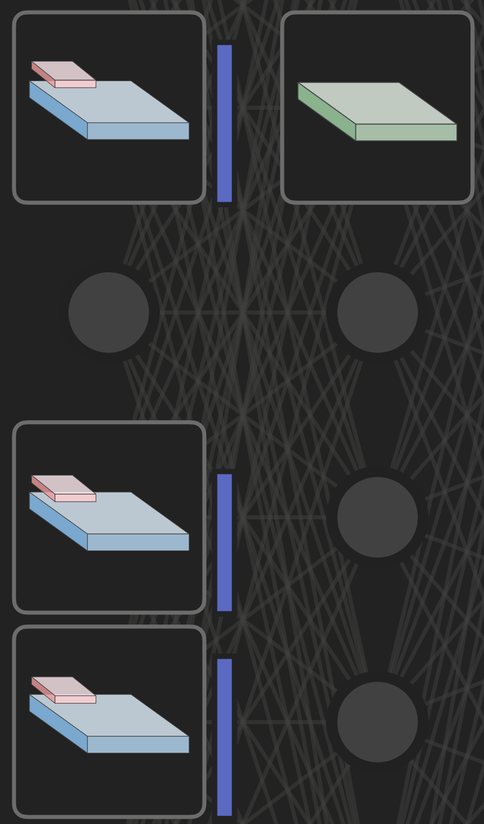
Aggregate network
influences (edges)



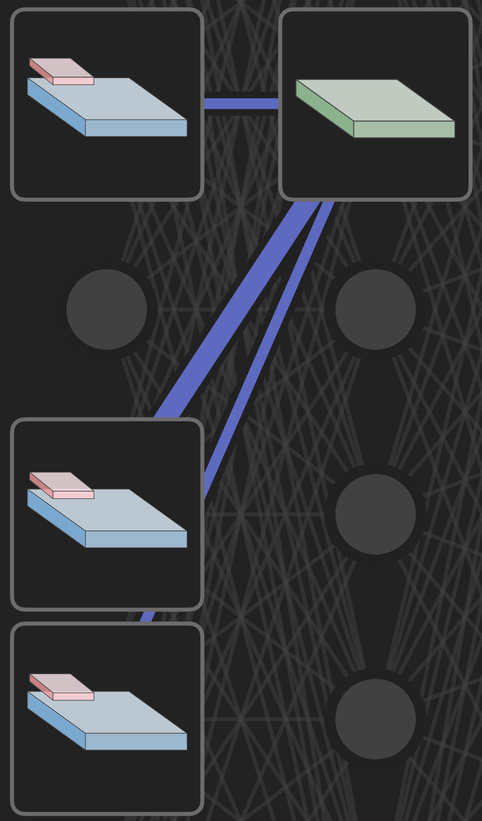


Aggregate network **influences** (edges)

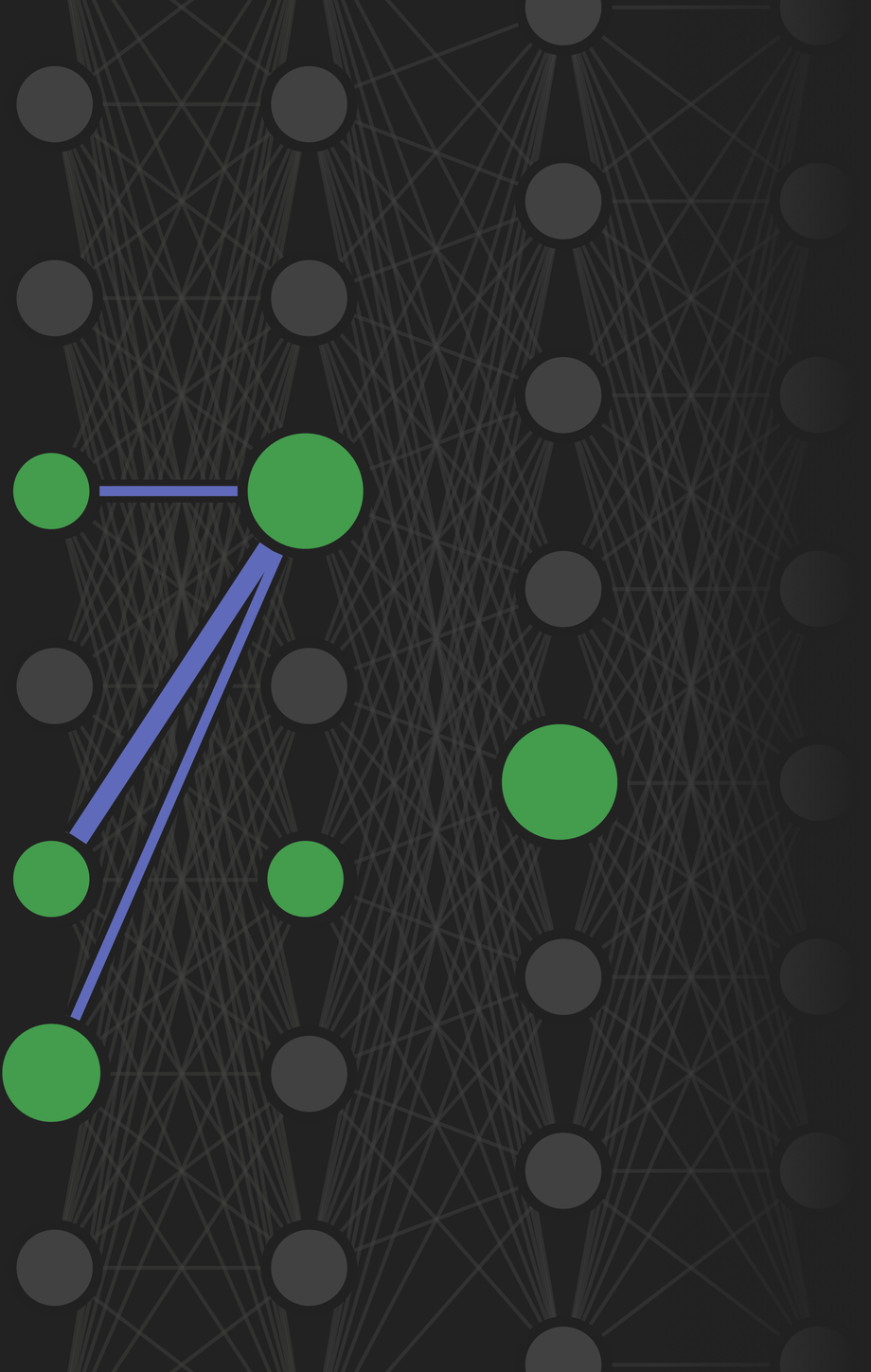




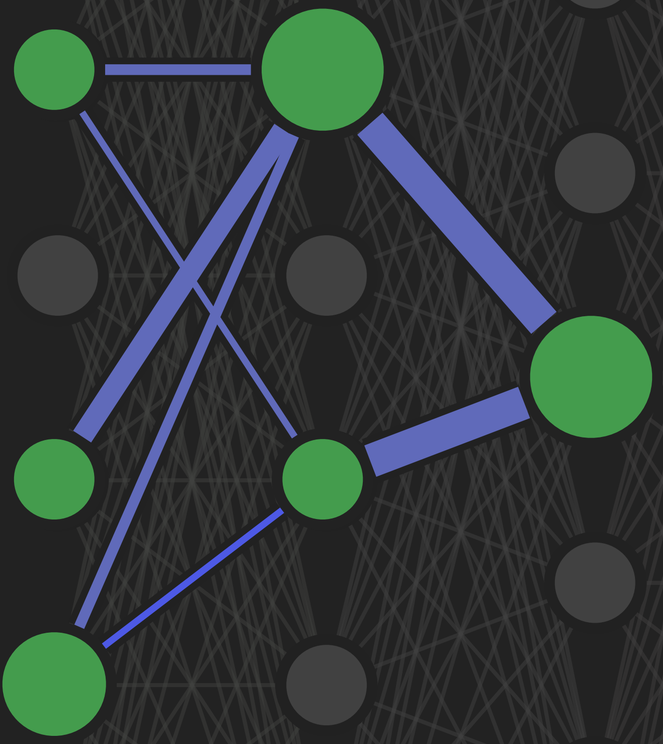
Aggregate network
influences (edges)



Aggregate network
influences (edges)



Combine **activations** and **influences**



Combine **activations** and **influences**

Further summarize graph **personalized PageRank**

Feature Visualization

What kind of input would cause a neuron to maximally activate?

Feature Visualization

What kind of input would cause a neuron to maximally activate?

mixed4b, 409

Generate examples: *starting from random noise, optimize an image to activate a particular neuron*

Feature Visualization

What kind of input would cause a neuron to maximally activate?

mixed4b, 409

Generate examples: *starting from random noise, optimize an image to activate a particular neuron*

[Olah, et al., Distill, 2017]

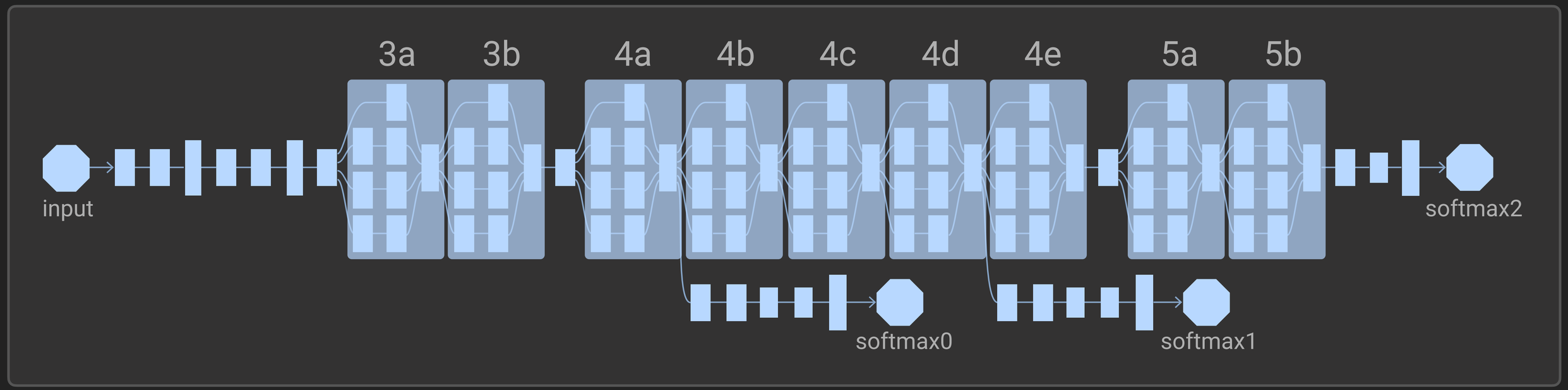
Demo

InceptionV1

Large-scale,
prevalent CNN

ImageNet (ILSVRC)

~1.3M images
1,000 classes



[Olah, et al., Distill, 2017]

SUMMIT

MODEL InceptionV1 | DATASET ImageNet | CLASSES 1,000 | INSTANCES 1,281,024

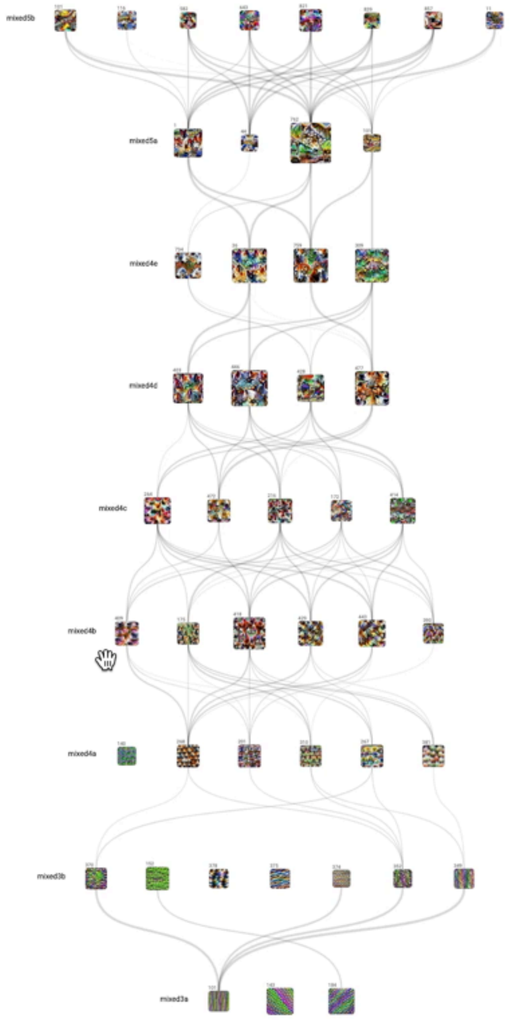
LAYER mixed — 3a 3b 4a 4b 4c 4d 4e 5a 5b

CLASS white_wolf | INSTANCES 1299 | ACCURACY 81.8% | PROBABILITIES

FILTER GRAPH | ADJUST WIDTH | ADJUST HEIGHT

white wolf

- white wolf 81.8%
- red wolf 69.9%
- timber wolf 64.2%
- arctic fox 87.1%
- lion 87.1%
- chow 87.1%
- rottweiler 76.6%
- silky terrier 63.3%



SUMMIT

MODEL InceptionV1 | DATASET ImageNet | CLASSES 1,000 | INSTANCES 1,281,024

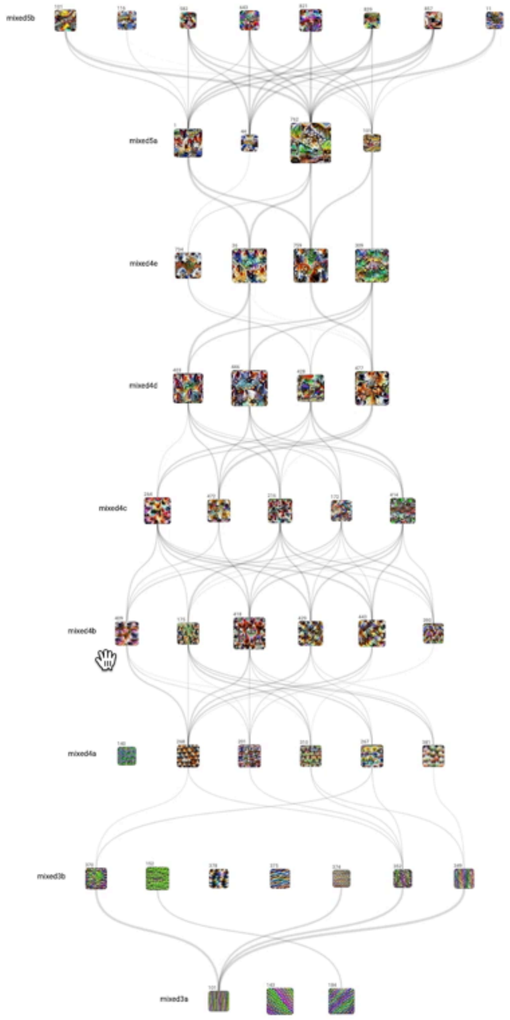
LAYER mixed — 3a 3b 4a 4b 4c 4d 4e 5a 5b

CLASS white_wolf | INSTANCES 1299 | ACCURACY 81.8% | PROBABILITIES

FILTER GRAPH | ADJUST WIDTH | ADJUST HEIGHT

white wolf

- white wolf 81.8%
- red wolf 69.9%
- timber wolf 64.2%
- arctic fox 87.1%
- lion 87.1%
- chow 87.1%
- rottweiler 76.6%
- silky terrier 63.3%



SUMMIT

MODEL InceptionV1 | DATASET ImageNet | CLASSES 1,000 | INSTANCES 1,281,024

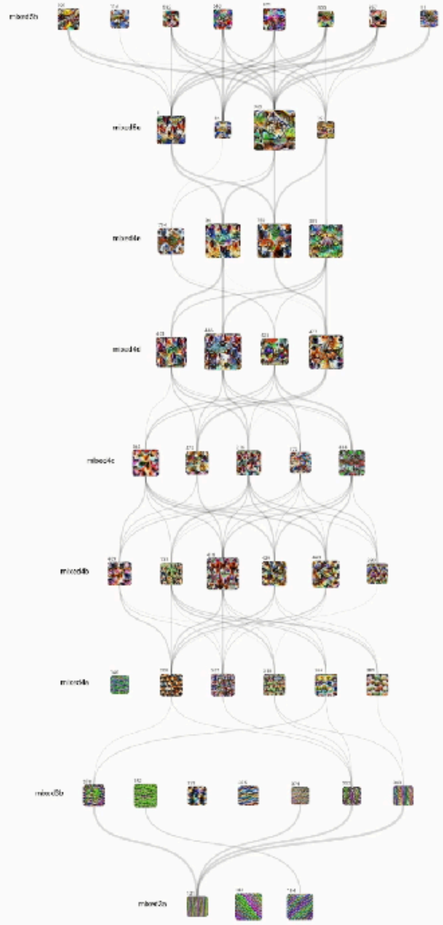
LAYER mixed: 3a 3b 4a 4b 4c 4d 4e 5a 5b

CLASS white_wolf | INSTANCES 1299 | ACCURACY 81.8% | PROBABILITIES | FILTER GRAPH | ADJUST WIDTH | ADJUST HEIGHT

timber wolf
white wolf
malamute
pembroke
samoyed
arctic fox
shetland sheepdog
lesser panda
papillon
keeshond
collie
chow

white wolf

white wolf	81.8%
red wolf	69.9%
timber wolf	64.2%
arctic fox	87.1%
lion	87.1%
chow	87.1%
rottweiler	76.6%
silky terrier	63.3%



SUMMIT

MODEL InceptionV1 | DATASET ImageNet | CLASSES 1,000 | INSTANCES 1,281,024

LAYER mixed: 3a 3b 4a 4b 4c 4d 4e 5a 5b

CLASS white_wolf | INSTANCES 1299 | ACCURACY 81.8% | PROBABILITIES

FILTER GRAPH | ADJUST WIDTH | ADJUST HEIGHT

timber wolf

white wolf

pem

chow

lion 87.1%

chow 87.1%

rottweiler 76.6%

silky terrier 63.3%

What features has a neural network learned for ***tench***?

How are those features related?

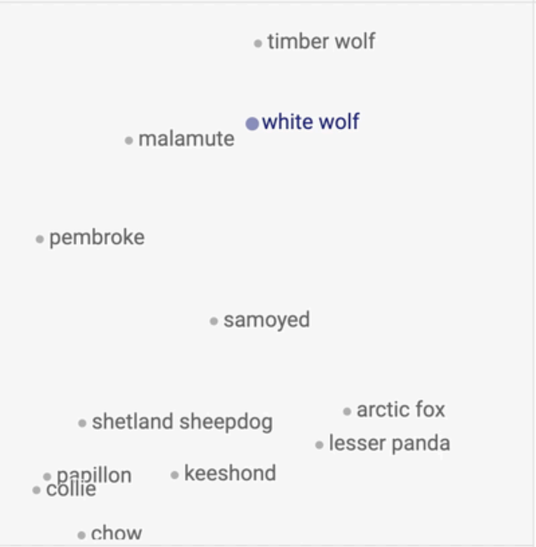
SUMMIT

MODEL InceptionV1 | DATASET ImageNet | CLASSES 1,000 | INSTANCES 1,281,024

LAYER mixed: 3a 3b 4a 4b 4c 4d 4e 5a 5b

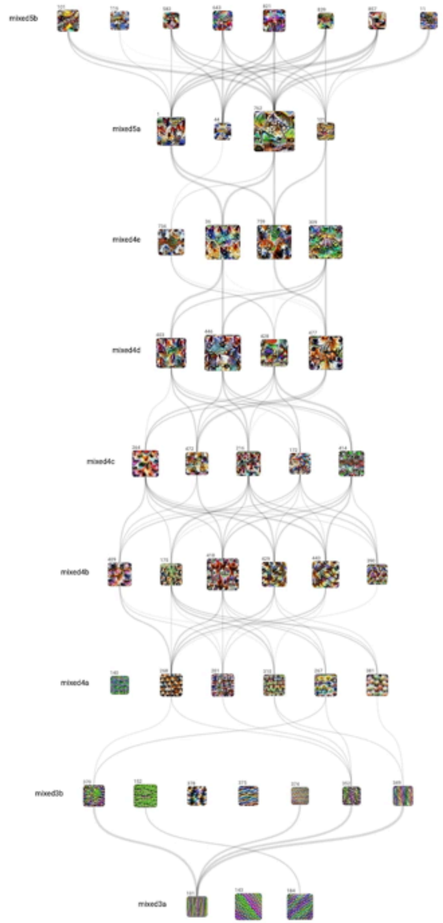
CLASS white_wolf | INSTANCES 1299 | ACCURACY 81.8% | PROBABILITIES

FILTER GRAPH | ADJUST WIDTH | ADJUST HEIGHT



white wolf

white wolf	81.8%
red wolf	69.9%
timber wolf	64.2%
arctic fox	87.1%
lion	87.1%
chow	87.1%
rottweiler	76.6%
silky terrier	63.3%



SUMMIT

MODEL InceptionV1 | DATASET ImageNet | CLASSES 1,000 | INSTANCES 1,281,024

LAYER mixed [3a] [3b] [4a] [4b] [4c] [4d] [4e] [5a] [5b]

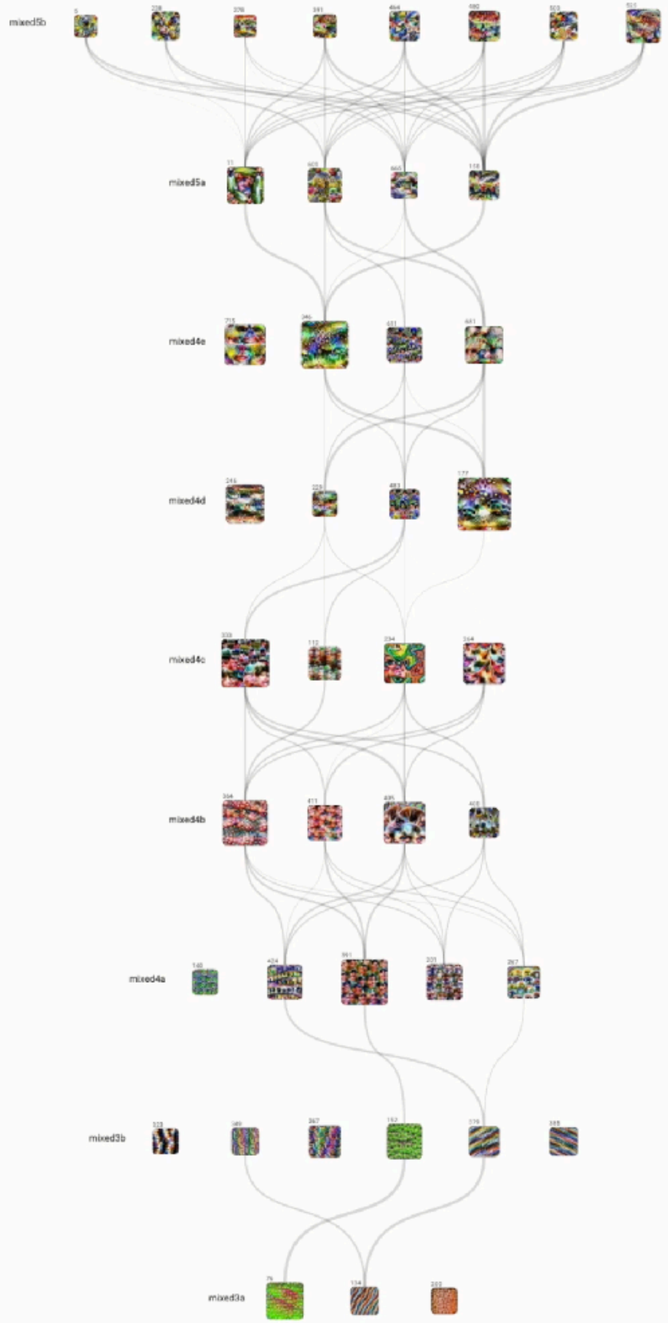
CLASS white_wolf | INSTANCES 1299 | ACCURACY 81.8% | PROBABILITIES

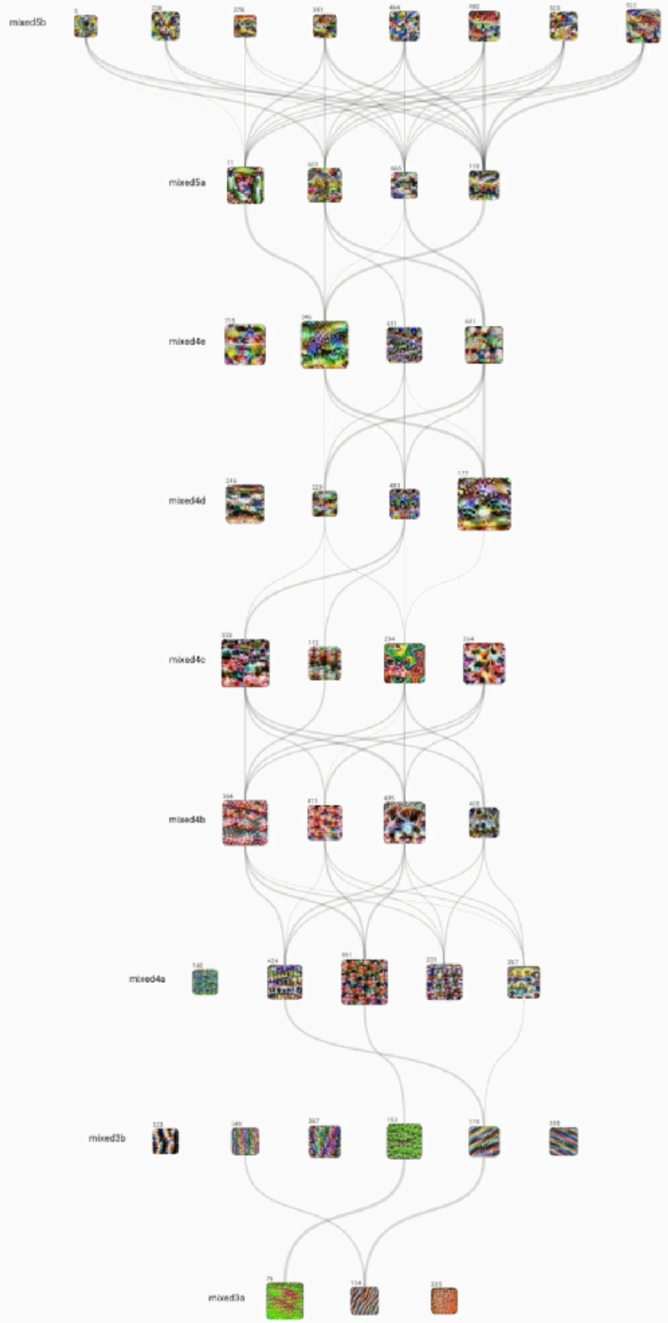
FILTER GRAPH | ADJUST WIDTH | ADJUST HEIGHT

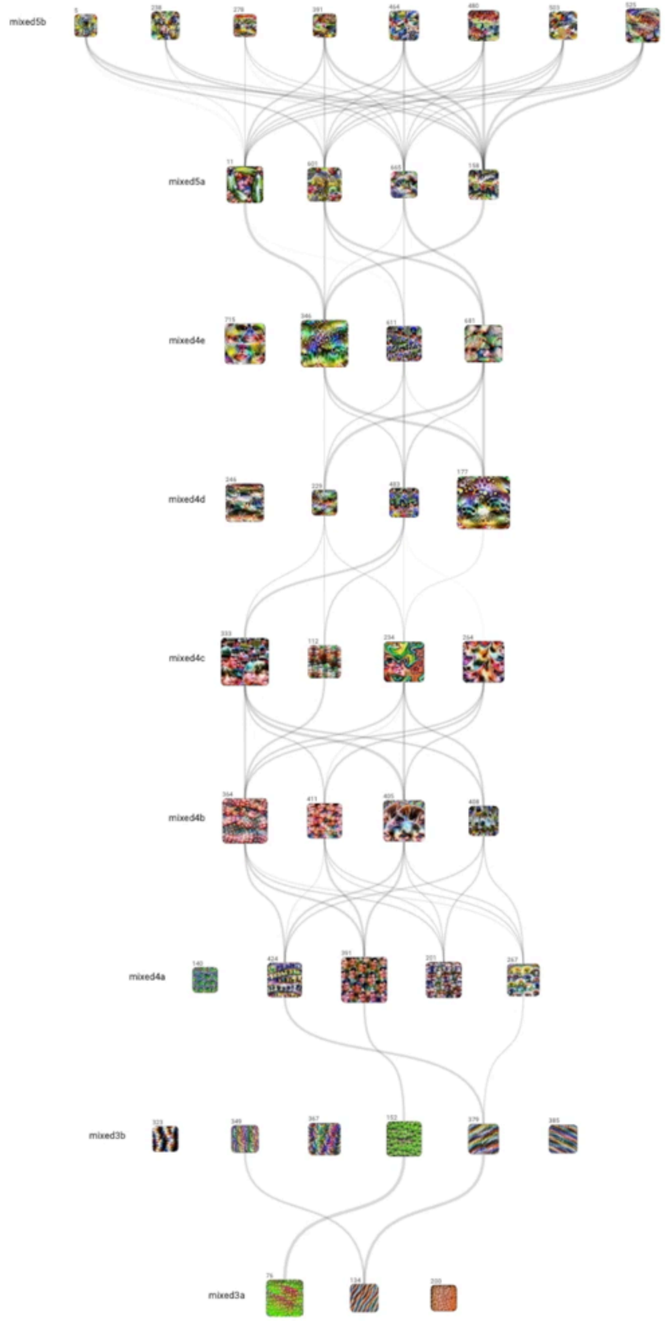
timber wolf
white wolf
malamute
pembroke
samoyed
shetland sheepdog
arctic fox
lesser panda
papillon
keeshond
collie
chow

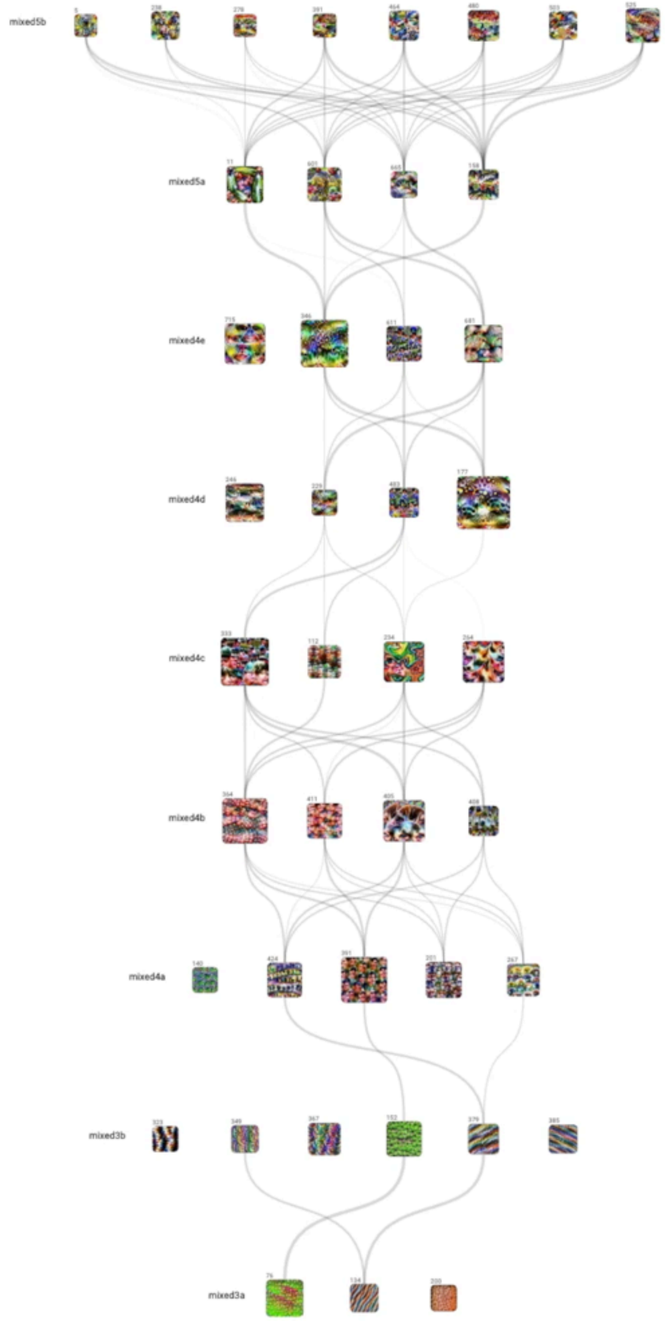
white wolf

white wolf	81.8%
red wolf	69.9%
timber wolf	64.2%
arctic fox	87.1%
lion	87.1%
chow	87.1%
rottweiler	76.6%
silky terrier	63.3%


mixed5b
mixed5a
mixed4e
mixed4d
mixed4c
mixed4b
mixed4a
mixed3b
mixed3a


mixed5b
mixed5a
mixed4e
mixed4d
mixed4c
mixed4b
mixed4a
mixed3b
mixed3a


mixed5b
mixed5a
mixed4e
mixed4d
mixed4c
mixed4b
mixed4a
mixed3b
mixed3a


mixed5b
mixed5a
mixed4e
mixed4d
mixed4c
mixed4b
mixed4a
mixed3b
mixed3a

Google

tench

All | Images | News | Shopping | Videos | More | Settings | Tools

world record | golden | pond | fish | male | float fi

Tench - Wikipedia
en.wikipedia.org

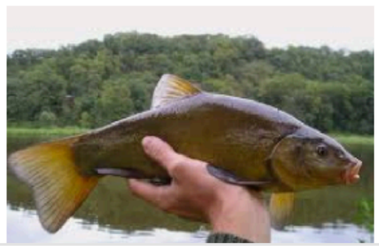
Top Tench Fishing Baits & Tactics...
dynamitebaits.com

Tench Fishing Guide - What Is Tench ...
badangling.com

Early season tench fishing tips ...
dynamitebaits.com

SPRING SPECIMENS Article | Korum ...
korum.co.uk

Boilie Approach For Tench | Drennan ...
drennantackle.com



Google tench

All Images News Shopping Videos More Settings Tools

world record | golden | pond | fish | male | float fi

Tench - Wikipedia
en.wikipedia.org

Top Tench Fishing Baits & Tactics...
dynamitebaits.com

Tench Fishing Guide - What Is Tench ...
badangling.com

Early season tench fishing tips ...
dynamitebaits.com

SPRING SPECIMENS Article | Korum ...
korum.co.uk

Boilie Approach For Tench | Drennan ...
drennantackle.com

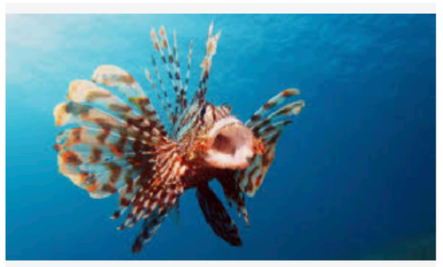
Google lionfish

All Images News Videos Shopping More Settings Tools

mermaid | invasive | aquarium | invasive species | florida

Pterois - Wikipedia
en.wikipedia.org

Invasive lionfish are delicious — but ...
oceana.org



Invasive lionfish are delicious — but ...
oceana.org

Lionfish: The Beautiful and Dange...
livescience.com

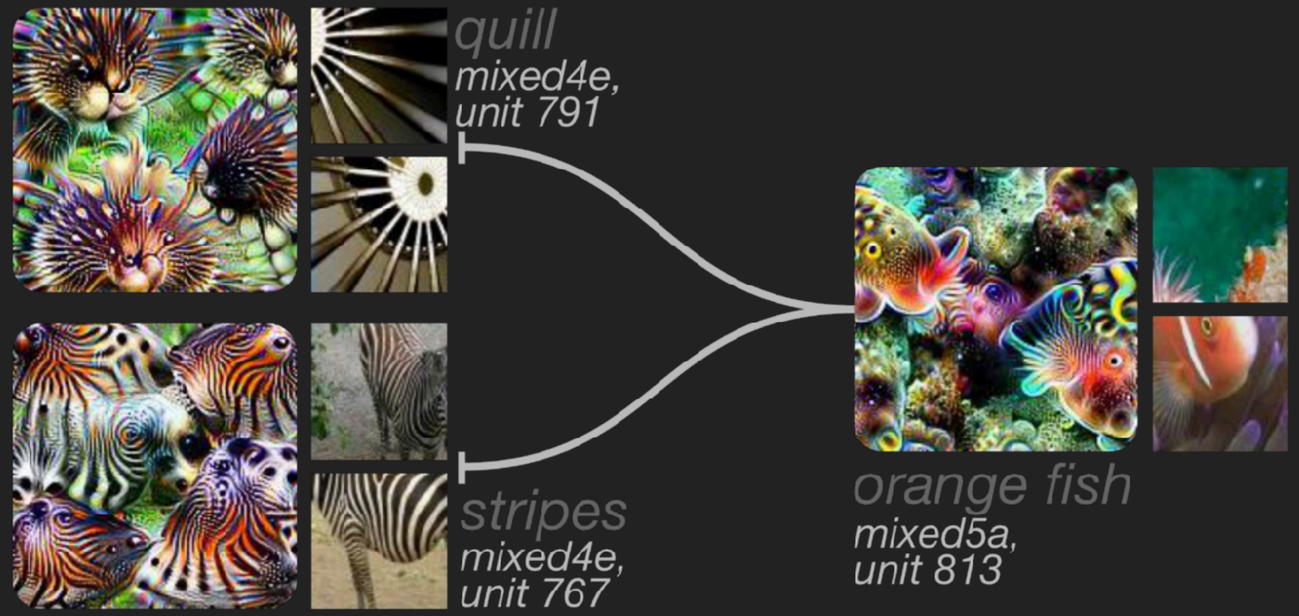
What is a lionfish?
oceanservice.noaa.gov

lionfish | Invasive Species, Sting ...
britannica.com

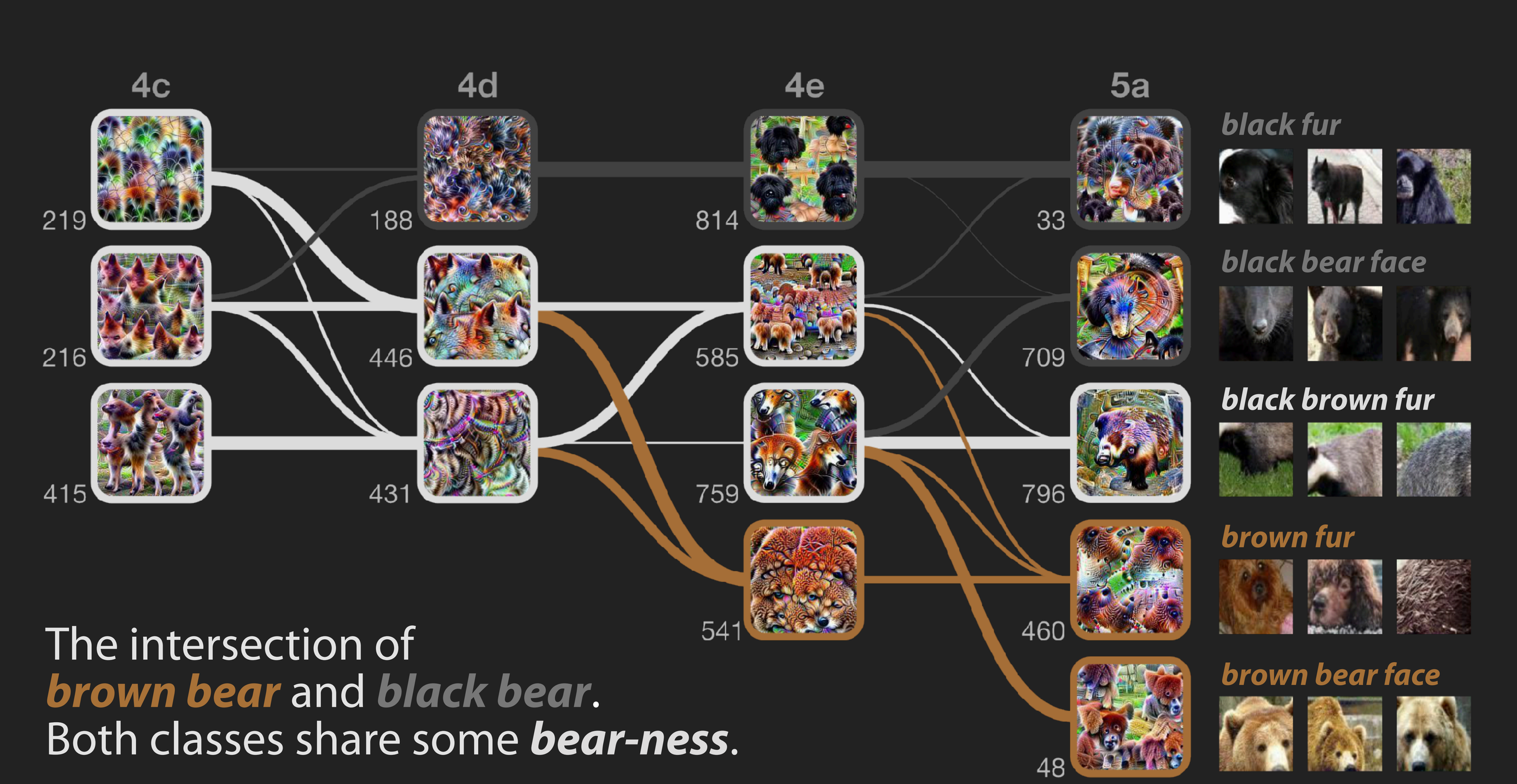
~~Un~~expected Features

lionfish

No more “people” features.
But few “fish” features! Mostly textures.



Attribution graph substructure from ***lionfish*** class.



The intersection of
brown bear and ***black bear***.
Both classes share some ***bear-ness***.

SUMMIT

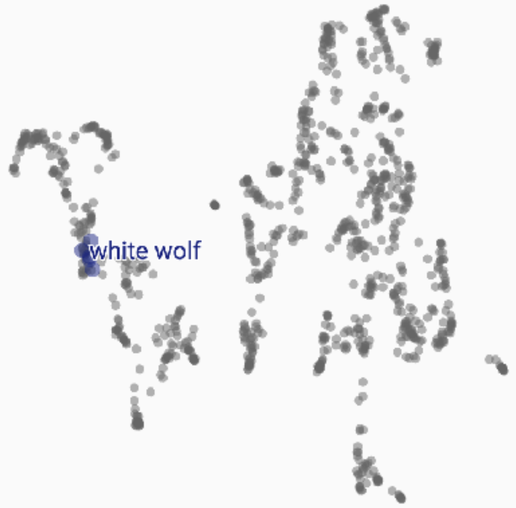
MODEL InceptionV1 | DATASET ImageNet | CLASSES 1,000 | INSTANCES 1,281,024

What is SUMMIT?

LAYER mixed: 3a 3b 4a 4b 4c 4d 4e 5a 5b

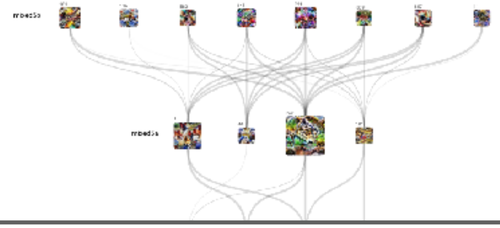
CLASS white_wolf | INSTANCES 1299 | ACCURACY 81.8% | PROBABILITIES

FILTER GRAPH | ADJUST WIDTH | ADJUST HEIGHT



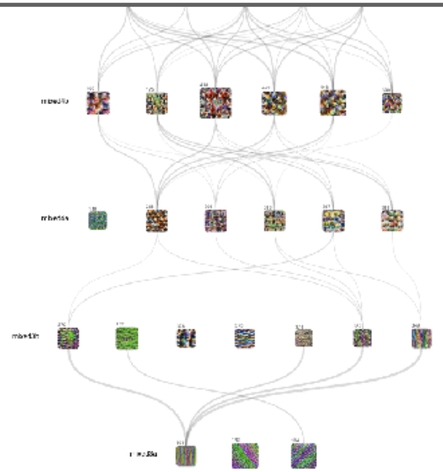
white wolf

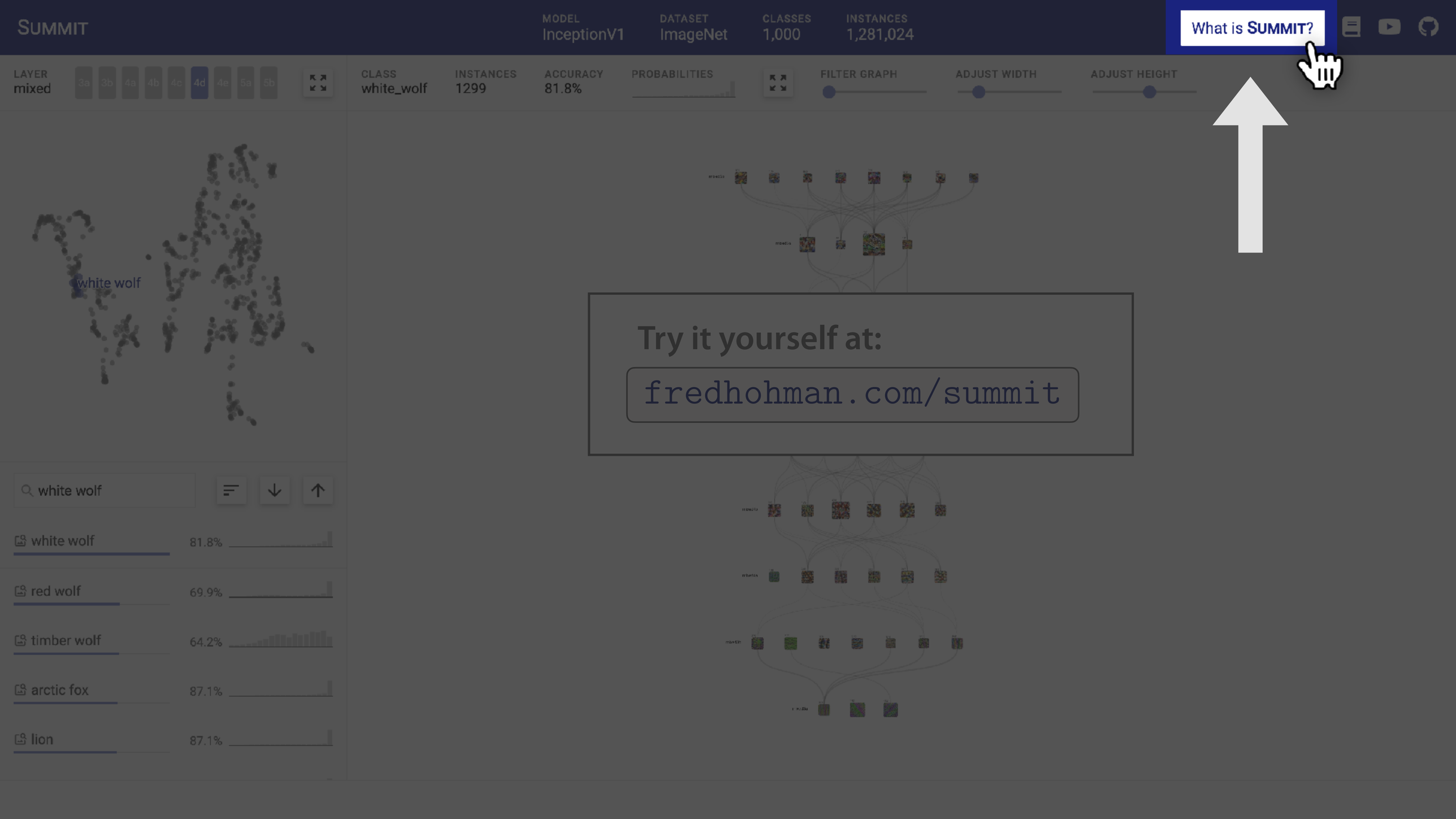
white wolf	81.8%
red wolf	69.9%
timber wolf	64.2%
arctic fox	87.1%
lion	87.1%



Try it yourself at:

fredhohman.com/summit




SUMMIT
MODEL
InceptionV1
DATASET
ImageNet
CLASSES
1,000
INSTANCES
1,281,024
What is SUMMIT?
LAYER
mixed
3a
3b
4a
4b
4c
4d
4e
5a
5b
CLASS
white_wolf
INSTANCES
1299
ACCURACY
81.8%
PROBABILITIES
FILTER GRAPH
ADJUST WIDTH
ADJUST HEIGHT
white wolf
Try it yourself at:
fredhohman.com/summit
white wolf
white wolf
81.8%
red wolf
69.9%
timber wolf
64.2%
arctic fox
87.1%
lion
87.1%

SUMMIT

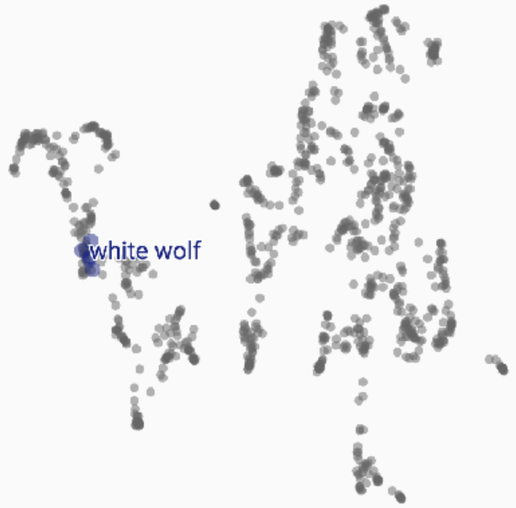
MODEL InceptionV1 | DATASET ImageNet | CLASSES 1,000 | INSTANCES 1,281,024

What is SUMMIT?

LAYER mixed: 3a 3b 4a 4b 4c 4d 4e 5a 5b

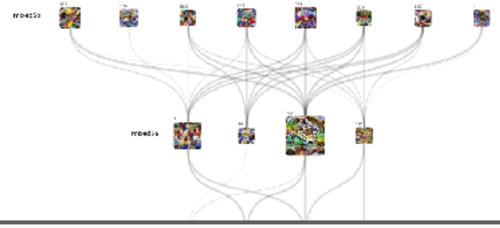
CLASS white_wolf | INSTANCES 1299 | ACCURACY 81.8% | PROBABILITIES

FILTER GRAPH | ADJUST WIDTH | ADJUST HEIGHT



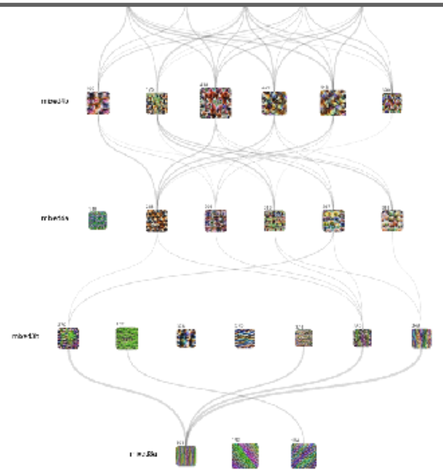
white wolf

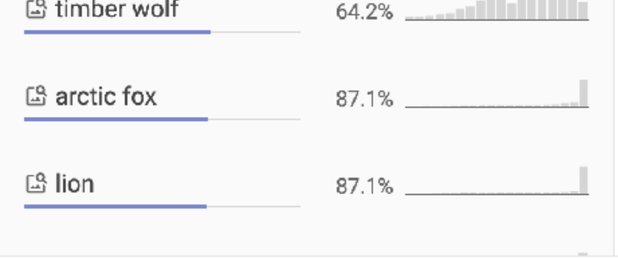
white wolf	81.8%
red wolf	69.9%
timber wolf	64.2%
arctic fox	87.1%
lion	87.1%

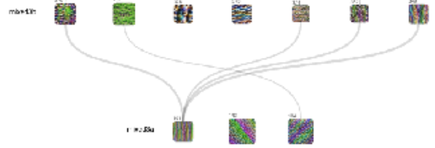


Try it yourself at:

fredhohman.com/summit







What is SUMMIT?

Understanding how neural networks make predictions remains a fundamental challenge. Existing work on interpreting neural network predictions for images often focuses on explaining predictions for single images or neurons, yet predictions are computed from millions of weights optimized over millions of images—such explanations can easily miss a bigger picture.

We present **SUMMIT**, an interactive visualization that scalably summarizes what features a deep learning model has learned and how those features interact to make predictions.

How does it work?

SUMMIT introduces two new scalable summarization techniques that aggregate activations and neuron-influences to create *attribution graphs:* a class-specific visualization that simultaneously highlights *what* features a neural network detects and *how* they are related.

Our work joins a growing body of open-access research that aims to use interactive visualization to explain complex inner workings of modern machine learning techniques. We believe our summarization approach that builds entire class representations is an important step for developing higher-level explanations for neural networks. We hope our work will inspire deeper engagement from both the information visualization and machine learning communities to further develop human-centered tools for artificial intelligence.

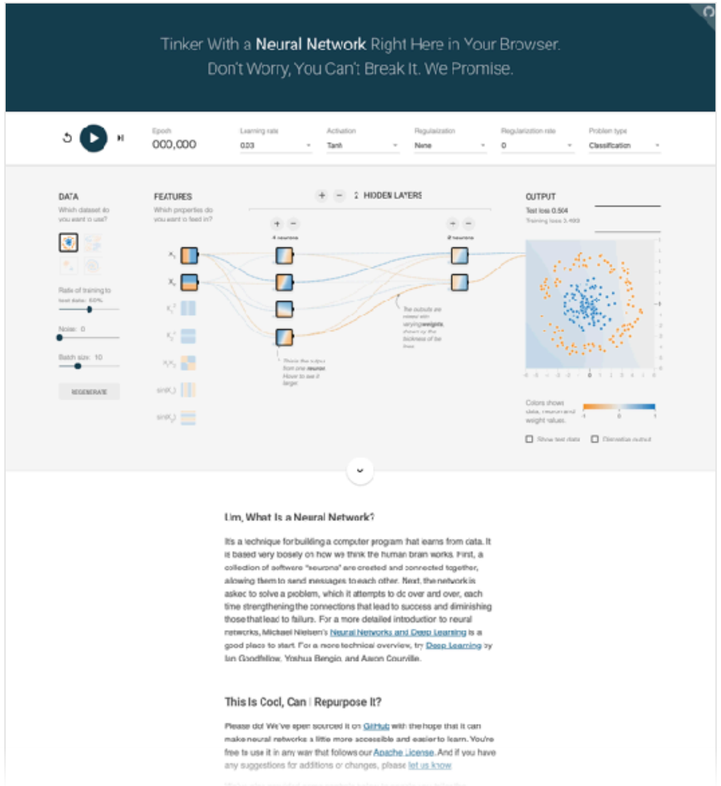
Credits

SUMMIT was created by Fred Hohman, Haekyu Park, Caleb Robinson, and Polo Chau at Georgia Tech. We also thank Nilaksh Das and the Georgia Tech Visualization Lab for their support and constructive feedback. This work is supported by a NASA Space Technology Research Fellowship and NSF grants IIS-1563816, CNS-1704701, and TWC-1526254.

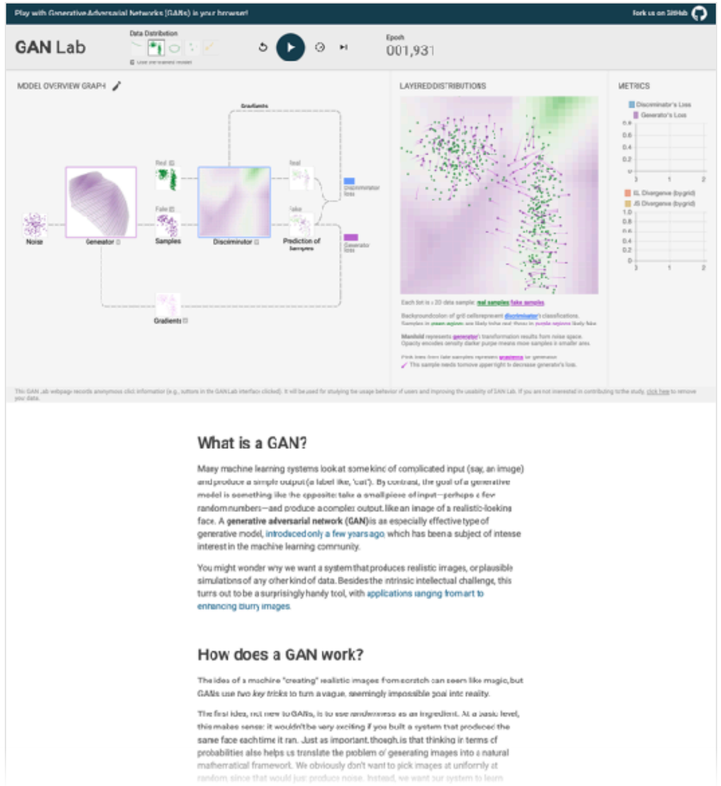
Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations
Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng (Polo) Chau.
IEEE Transactions on Visualization and Computer Graphics (TVCG, Proc. VAST'19). 2020.

- **Live demo:** fredhohman.com/summit
- **Paper:** https://fredhohman.com/papers/19-summit-vast.pdf
- **Video:** https://youtu.be/J4GMLvoH1ZU
- **Code:** https://github.com/fredhohman/summit
- **Slides:** coming October 2019!

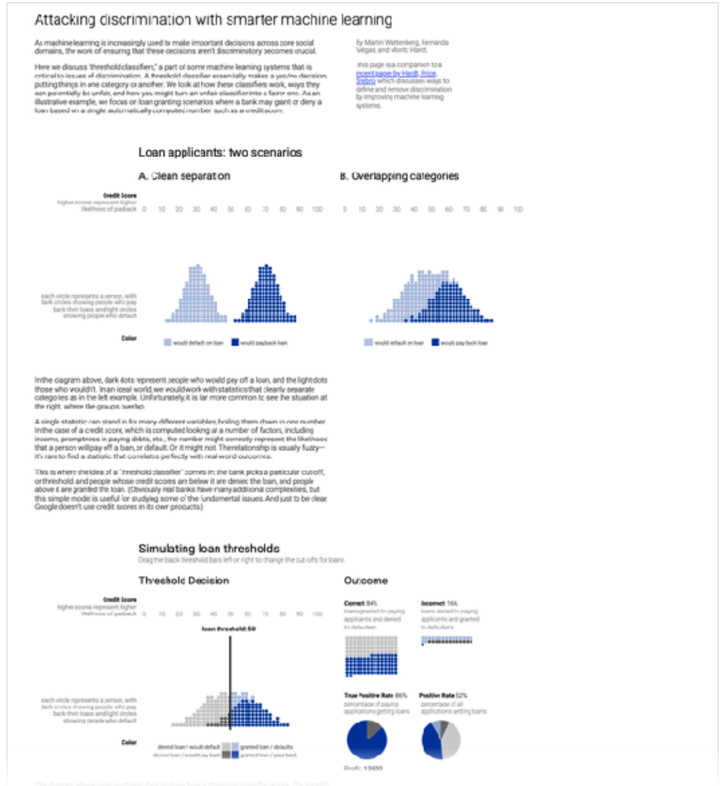
TensorFlow Playground
2016



GAN Lab
2018



Attacking Discrimination with Smarter Machine Learning
2016



Interactive Scalable Interfaces for Machine Learning Interpretability

PART I
Enable interpretability

GAMUT Operationalize interpretability *CHI 2019*
TELEGAM Vis + text for better explanations *VIS 2019*

PART II
Scale interpretability

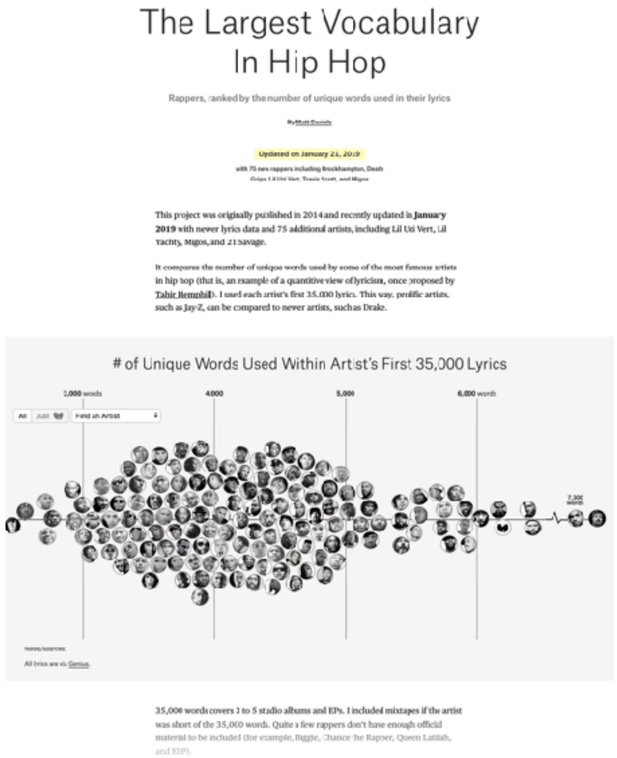
Interrogative Survey Summarize interpretability vis *TVCG 2018*
SUMMIT Higher-level explanations for neural networks *VAST 2019*

PART III
Communicate interpretability

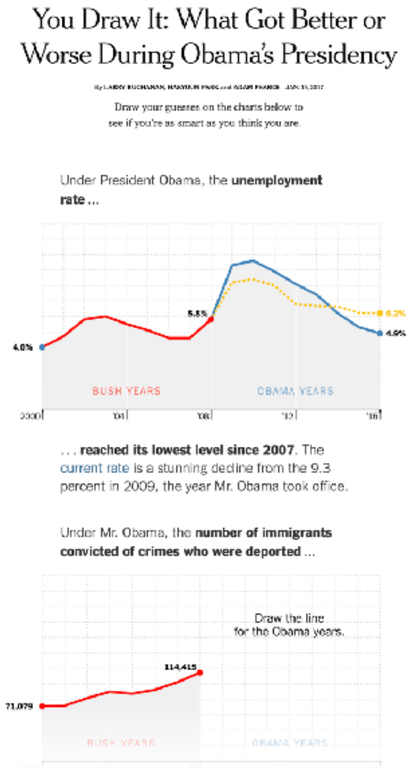
ML Literacy Interactive mediums & platforms *VISCOMM 2019, VISxAI 2018*
Interactive Articles Formalizing interactive communication *Distill 2020*

Interactive Articles for Politics, Pop Culture, …

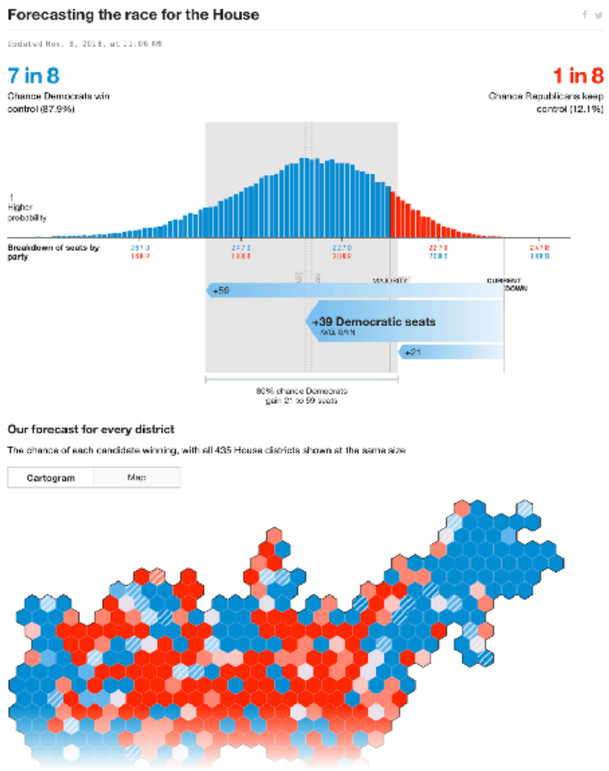
The Pudding



The New York Times



FiveThirtyEight



Interactive Articles now for Machine Learning!

Parametric Press

Issue 01
Science + Society

Table of Contents

Impartial Machine

The Myth of the Impartial Machine

May 1, 2019

Created by
Alice Fung
Shuyan Wu

Metadata
Source Code
Offline Archive
DOI

Wide-ranging applications of data science bring utopian proposals of a world free from bias, but in reality, machine learning models reproduce the inequalities that shape the data they're fed. Can programmers free their models from prejudice?

From voice assistants to image recognition, fraud detection to social media feeds, machine learning (ML) and artificial intelligence (AI) are becoming an increasingly important part of society. The two fields have made enormous strides in recent years thanks to gains in computing power and the so-called "information explosion." Such algorithms are being used in fields as varied as medicine, agriculture, insurance, transportation and art, and the number of companies rushing to embrace what ML and AI can offer has increased rapidly in recent years.

According to a survey conducted by Teradata in July 2017, 80% of enterprises have already begun investing in AI technologies and 30% plan to increase their spending over the next 36 months. Investment in such models is also forecasted to grow from $12 billion in 2017 to over $50 billion by 2021. Billed as being more

The Myth of the Impartial Machine

May 1, 2019

Created by
Alice Feng
Shuyan Wu

Metadata
Source Code
Offline Archive
DOI

Wide-ranging applications of data science bring utopian proposals of a world free from bias, but in reality, machine learning models reproduce the inequalities that shape the data they're fed. Can programmers free their models from prejudice?

From voice assistants to image recognition, fraud detection to social media feeds, machine learning (ML) and artificial intelligence (AI) are becoming an increasingly important part of society. The two fields have made enormous strides in recent years thanks to gains in computing power and the so-called "information explosion." Such algorithms are being used in fields as varied as medicine, agriculture, insurance, transportation and art, and the number of companies rushing to embrace what ML and AI can offer has increased rapidly in recent years.

According to a survey conducted by Teradata in July 2017, 80% of enterprises have already begun investing in AI technologies and 30% plan to increase their spending over the next 36 months. Investment in such models is also forecasted to grow from $12 billion in 2017 to over $50 billion by 2021. Billed as being more accurate, consistent and objective than human judgment, the promises and expectations of what ML and AI can achieve have never been greater.

However, for every neural network that can defeat Jeopardy champions and outplay Go masters, there are other well-documented instances where these algorithms have produced highly disturbing results. Facial-analysis programs were found to have an error rate of 20 to 34 percent when trying to determine the gender of African-American women compared to an error rate of less than one percent for white men. ML algorithms used to predict which criminals are most likely to reoffend tended to incorrectly flag black defendants as being high risk at twice the rate of white defendants. A word embedding model used to help machines determine the meaning of words based on their similarity likewise associated men with being computer programmers and women with homemakers.

If data-trained models are supposed to be objective and impartial, how did these algorithms get things so wrong? Can such bias be fixed?

What's the difference between Artificial Intelligence and Machine Learning?
Artificial intelligence and machine learning are often used interchangeably but there are in fact differences between the two.

Artificial intelligence refers to the broader science of getting computers to act intelligently without being explicitly programmed.

Machine learning is the use of statistical algorithms to detect patterns in large datasets. It is one way in which computers can become better at a task and thus considered to be a subset of artificial intelligence.

The Machine Learning Pipeline

Being able to use data to meaningfully answer questions via machine learning requires several steps. Before getting into the details of bias, it is important to understand them.

1. **Data gathering.** All machine learning models require data as inputs. In today's increasingly digitized world, data can be derived from various sources including user interactions on a website, collections of photo images and sensor recordings.
2. **Data preparation.** Data collected are rarely in a usable state as-is. Data often need to be cleaned, transformed and checked for errors before they are ready to be fed into a model.
3. **Split dataset into training and testing sets.** The training dataset is used to build and train the model while the testing dataset, which is kept separate, is used to evaluate how well the model performs. It is important to assess the model on data it has not seen before in order to ensure that it has indeed learned something about the underlying structure of the data rather than simply "memorized" the training data.
4. **Fit and train models.** This is the step where various types of ML models such as regression models, random forests and neural networks are built and applied to the training data. Models are iterated on by making small adjustments to their parameters in order to improve their performance with the goal of generating the most accurate predictions possible.
5. **Evaluate model on the test dataset.** The top performing model is used on the testing data to get a sense of how the model will perform on real world data it's never seen before. Based on the results, further refinement and tuning of the model may be needed.
6. **Make predictions!** Once the model is finalized, it can begin to be used to answer the question it was designed for.

Sources of bias

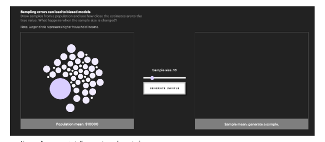
There are two key ways bias can be introduced and amplified during the machine learning process: by using non-representative data and while fitting and training models.

Biased data

Data that are non-representative and biased.

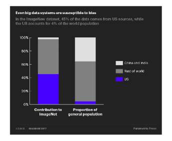
When one examines a data sample, it is imperative to check whether the sample is representative of the population of interest. A non-representative sample where some groups are over- or under-represented inevitably introduces bias in the statistical analysis. A dataset may be non-representative due to sampling error and non-sampling errors.

Sampling errors refer to the difference between a population value and a sample estimate that exists only because of the sample that happened to be selected. Sampling errors are especially problematic when the sample size is small relative to the size of the population. For example, suppose we sample 100 residents to estimate the average US household income. A sample that happened to include Jeff Bezos would result in an overestimate, while a sample that happened to include predominantly low-income households would result in an underestimate.



Non-sampling errors are typically more serious and may arise from many different sources such as errors in data collection, non-response, and selection bias. Typical examples include poorly phrased data-collection questions, web-only data collection that leave out people who don't have easy access to the internet, over-representation of people that feel particularly strongly about a subject, and responses that may not reflect one's true opinion.

Even big data are susceptible to non-sampling errors. A study by researchers at Google found that the United States (which accounts for 4% of the world population) contributed over 45% of the data for ImageNet, a database of more than 14 million labelled images. Meanwhile, China and India combined contribute just 3% of images, despite accounting for over 36% of the world population. As a result of this skewed data distribution, image classification algorithms that use the ImageNet database would often correctly label an image of a traditional US bride with words like "bride" and "wedding" but label an image of an Indian bride with words like "costume".



Data that are representative but still biased.

What if a dataset is representative of its target population? No more worries about bias coming through the data, right? Not so fast—a representative dataset may still be biased to the extent that it reflects any historical social bias. A recruiting algorithm once used by Amazon was found to have disproportionately favored male candidates. The data used to train the algorithm were based on resumes collected over a 10-year period.

Because the tech industry was historically dominated by men, a majority of the resumes were submitted by men. Therefore, even though the dataset was "representative" of the historical applicant pool, it captured the inherent gender bias and passed it on to the recruiting algorithm. Representative but biased data is in some sense a more problematic issue than non-representative data since fixing the former would require addressing biases that are inherent in society.

Biases that are amplified through algorithms

Machine learning algorithms themselves may amplify bias if they make predictions that are more skewed than the training data. Such amplification often occurs through two mechanisms: 1) incentives to predict observations as belonging to the majority group and 2) runaway feedback loops.

More from the Parametric Press

Data Science for Fair Housing
Cities across America covertly exclude racial minorities from majority-white residential neighborhoods, while gentrification drives people of color out of their homes. In Atlanta, a new nonprofit seeks to resist displacement by supporting the city's most vulnerable residents—but how effective is their project?

Flatland Follies: An Adjunct Simulator
This college used to be one of the best in the country. Fifty years later the campus is destitute, they can't pay professors, and it's filled with decaying art.

On Particle Physics
A CERN particle physicist walks through the history and science of particle physics, and why you should care about it—even outside of the laboratory.

Algorithms incentivized to predict the majority group.

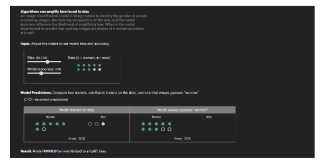
In order to maximize predictive accuracy when faced with an imbalanced dataset, machine learning algorithms are incentivized to put more learning weight on the majority group, thus disproportionately predicting observations to belong to that majority group. The next interactive example illustrates this tendency.

Suppose we have an image-classification algorithm tasked with identifying the gender of the person in a given image. The training dataset contains 10 cooking images; each image shows either a man or a woman in the kitchen. This dataset may range from a balanced, unbiased sample (five images showing men, five images showing women) to a completely biased sample (all 10 images showing either women or men).

Having unbiased training data is the ideal scenario, but in a more realistic setting, women are likely to be over-represented in a sample of cooking images. Suppose the training dataset contains eight images of women cooking and two images of men and that the model has a 70% predictive accuracy (i.e., a bias of 80/20 and model accuracy of 70% in the interactive below). When this data is fed through the model, the model correctly labels the gender of six of the photos of women and one of the photos of a man while mistaking the gender of the remaining three photos (30% error rate).

If the model were to instead simply predict that every cooking image is of a woman, it would lower its error rate to 20% because it has correctly labeled all of the photos with a woman while only mistaking the gender of the two photos of men. In pursuit of higher predictive accuracy, the model has amplified the bias from 80% women in the training data to 100% women in the model prediction.

Play with the bias and model accuracy settings below to see how different mixes affect the tendency for the model to amplify bias. For a given level of predictive accuracy, how does the tendency for the model to amplify bias change with the amount of bias in the data, and vice versa? When is the model incentivized to amplify bias, and when it is not?



This graphic is based on a scenario described in this article by Laura Douglas.

Runaway feedback loops.

In the previous image-classification example, bias amplification stops at the model prediction stage. However, in machine learning models where the prediction is fed back into the model as inputs for the next round of predictions, bias can be amplified further in the form of a feedback loop.

Consider a predictive policing algorithm used to determine the optimal allocation of police force across a city. Suppose the city has two precincts (A and B) and one policeman, John. The city wishes to use historical data on past crime occurrences to determine to which precinct to dispatch John. The goal is to send John to each precinct in propotion to the frequency of crimes that occur in each precinct. If 40% of crimes occur in precinct A, the algorithm should send John to patrol A 40% of the time.

In our example, let's say the city experienced a total of 22 crimes in the past year, with 12 of those occurring in precinct A and 10 in precinct B, as shown below:

The predictive algorithm uses this historical data to determine where to send John. The likelihood that the algorithm will dispatch John to precinct A is in proportion to the share of all crimes that occurred in A. In this case, this probability is 12/22 = 54.5%. The algorithm will choose to send John to precinct B with a probability of 10/22 = 45.5%.

Let's say the algorithm decides to dispatch John to patrol precinct A. While there, John encounters three crimes and logs them in the system. The updated data now show a total of 15 past crimes in A and 10 in B. However, because no one was sent to precinct B, crimes that occured in B on the same day were not captured in the data.

The next day, when the algorithm decides where to send John to, it will have a 15/25 = 60% probability of sending John to A and 10/25 = 40% probability of sending him to B. The algorithm now has an even higher likelihood of sending John to precinct A as a result of its decision to assign John to A yesterday. When the model chooses to send John to precinct A, more instances of crime will be logged for precinct A, while crime that occur in B are ignored and remain uncaptured in the data.

Over time, the data will become increasingly distorted, with the gap between the observed crime ratio and the actual crime ratio of the two precincts widening due to the decisions the model makes.

Officer sent to A: 0	Officer sent to B: 0
Observed crimes in A: 4 (67%)	Observed crimes in B: 2 (33%)
Total actual crimes in A: 4 (67%)	Total actual crimes in B: 2 (33%)

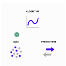
Feedback loops are especially problematic when sub-groups in the training data exhibit large statistical differences (e.g. one precinct has a much higher crime rate than others); a model trained on such data will quickly "run away" and make predictions that fall into the majority group only, thereby generating ever-more lopsided data that are fed back into the model.

Even when sub-groups are statistically similar, feedback loops can still lead to noisy and less accurate predictions. Algorithms where the predictive outcome determines what feedback the algorithm receives—e.g. recidivism prediction, language translation, and social media news feeds—should always be diligently monitored for the presence of feedback loops bias.

Bias in data and in algorithms are interrelated

It should be clear by this point that bias in data and algorithms are interrelated. When an algorithm is fed training data where one group dominates the sample, it is incentivized to prioritize learning about the dominant group and over-predict the number of observations that belong to the dominant group. This tendency is exacerbated when the model's predictive accuracy is relatively low. Conversely, if the data were balanced relative to the predictive accuracy, the model would have nothing to gain by over-predicting the dominant group.

Bias can also be perpetuated through a feedback loop if the model's own biased predictions are repeatedly fed back into it, becoming its own biased source data for the next round of predictions. In the machine learning context, we no longer just face the risk of garbage in, garbage out—when there's garbage in, more and more garbage may be generated through the ML pipeline if one does not monitor and address potential sources of bias.



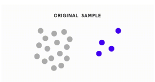
How people are tackling bias

How does one actually tackle bias in the ML pipeline? While a suitable fix depends on each specific circumstance, here are some ways that companies and researchers are trying to reduce bias in machine learning.

1. De-biasing data

One key to de-biasing data is to ensure that a representative sample is collected in the first place. Bias from sampling errors can be mitigated by collecting larger samples and adopting data collection techniques such as stratified random sampling. While sampling errors won't go away entirely, rapid data growth—2.5 quintillion bytes per day and counting—and growing data collection capability have made it easier than ever to mitigate sampling errors compared to the past.

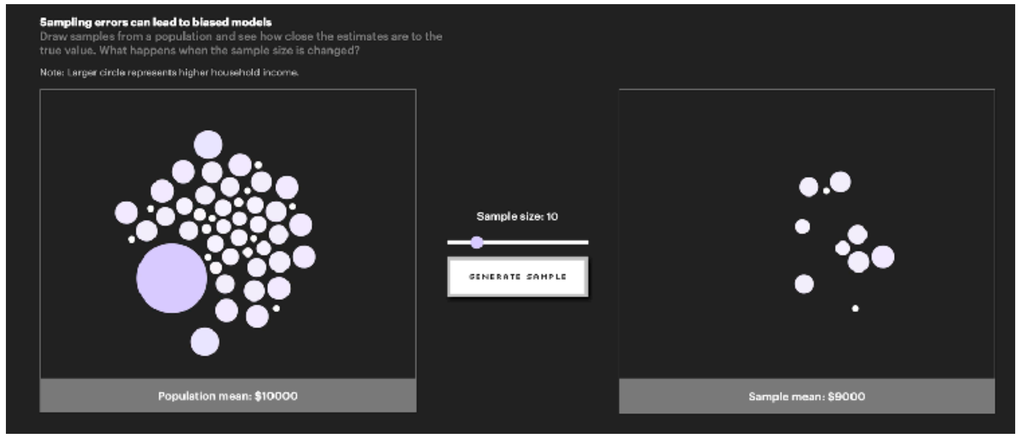
Bias from non-sampling errors are much more varied and harder to tackle, but one should still strive to minimize these kinds of errors through means such as proper training, establishing a clear purpose and procedure for data collection, and conducting careful data validation. For example, in response to the image-classification database that contained disproportionately few wedding images from India, Google deliberately sought out contributions from India to make the database more representative.



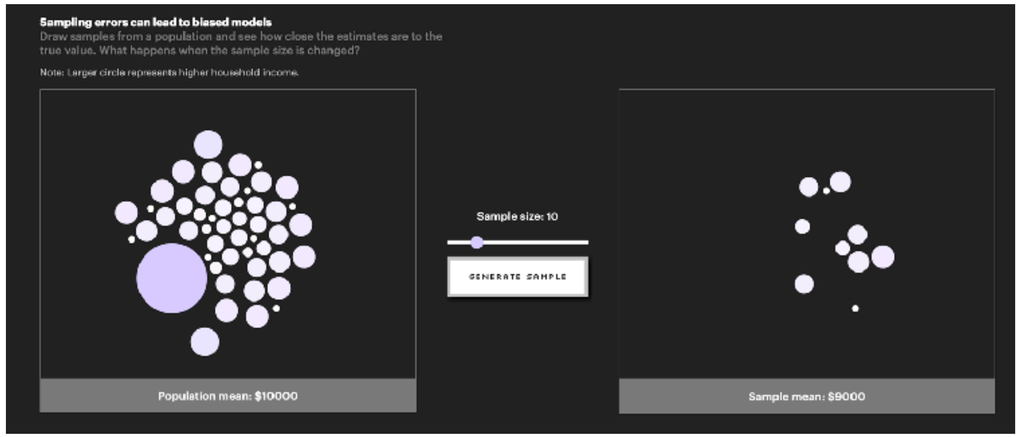
What about data that are representative but reflect bias inherent in the population? For datasets that contain minority groups, one can oversample those minority groups to create a more balanced training dataset. For datasets that may contain biased associations, one can first quantify and remove any biased associations from the dataset before proceeding to the model training and prediction stages.

Word2Vec is a model that can be used to quantify relationships between words. De-biasing Word2Vec data is an example of this latter approach: researchers first measured how closely two words relate to each other along the gender dimension and assessed whether different strengths of association reflect gender bias or appropriate relationships.

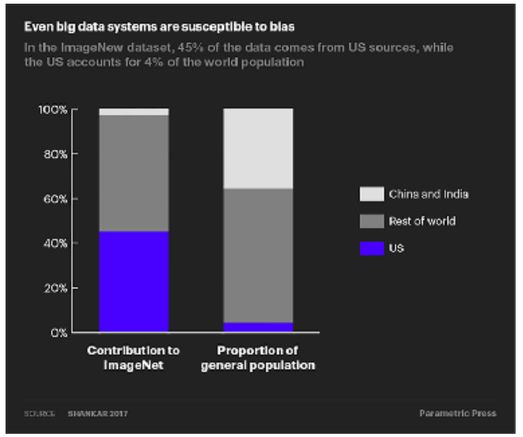
Interactive Graphics



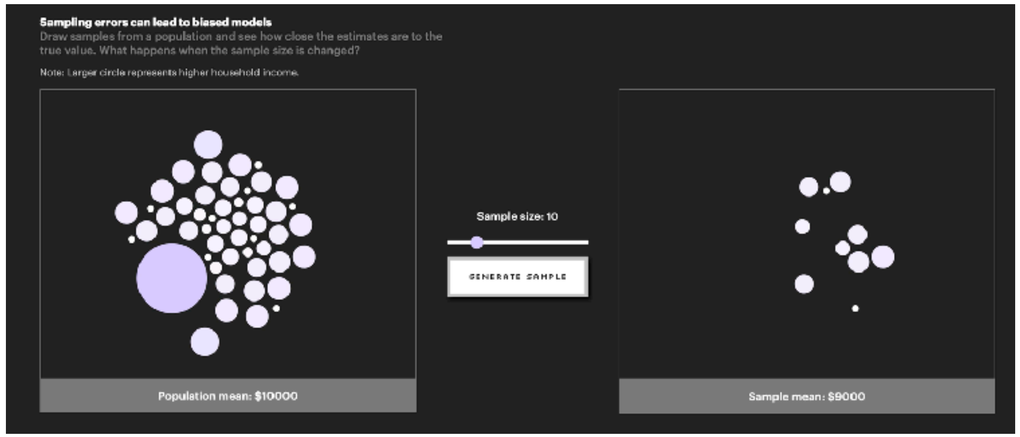
Interactive Graphics



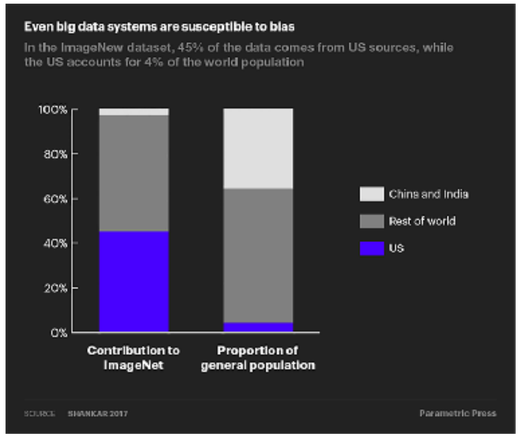
Visualization



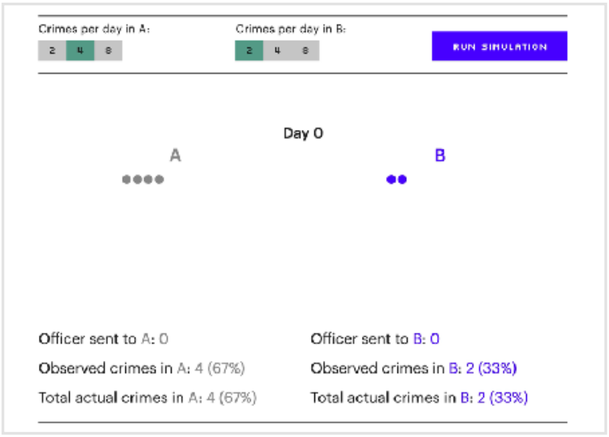
Interactive Graphics



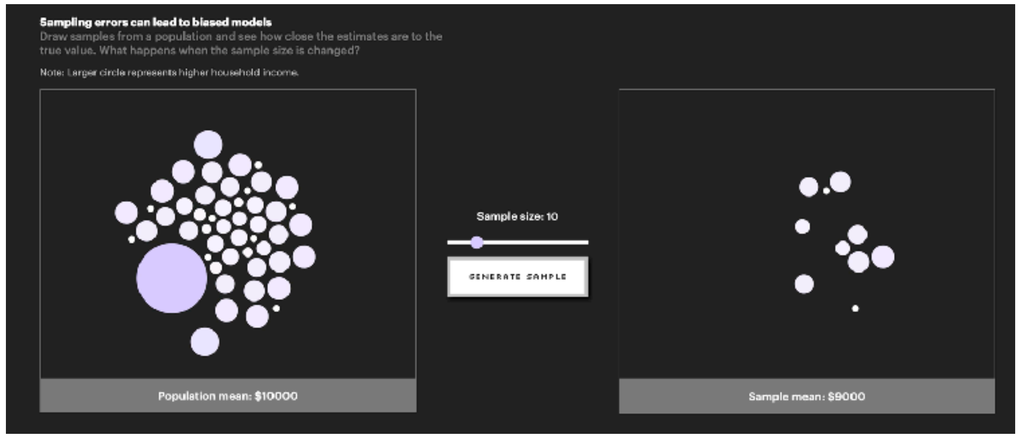
Visualization



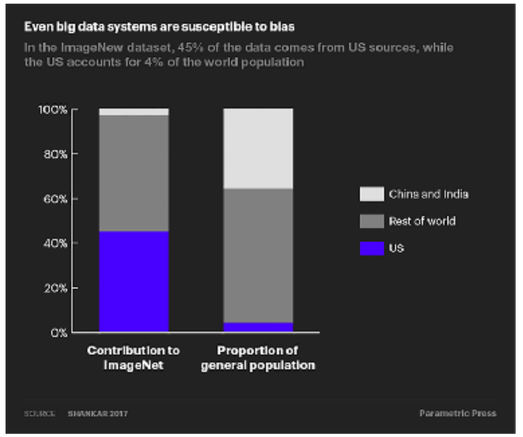
Simulation



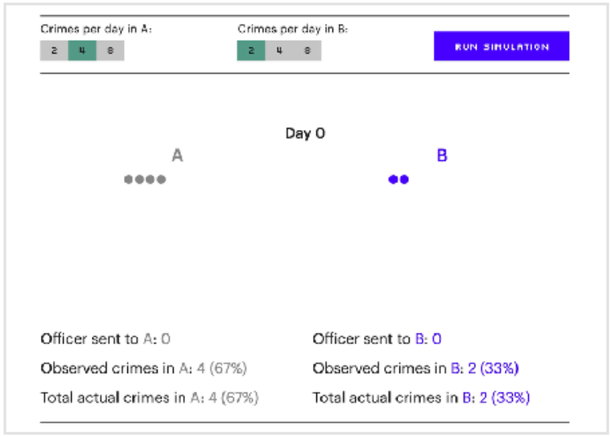
Interactive Graphics



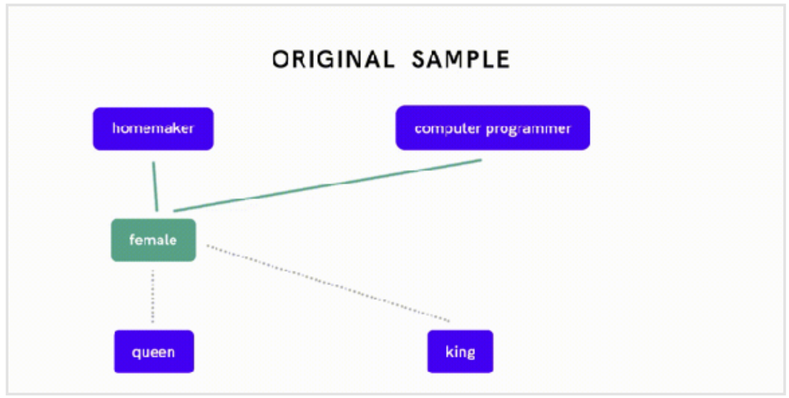
Visualization

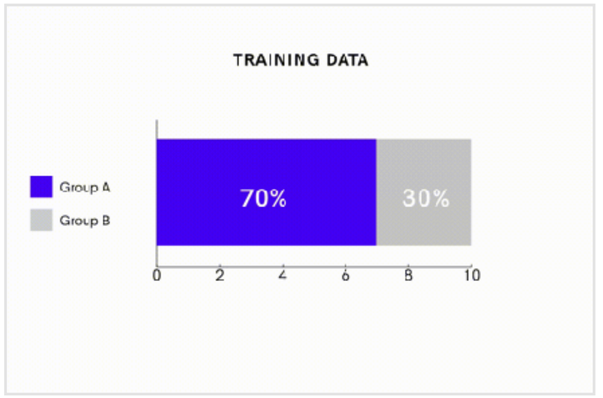


Simulation

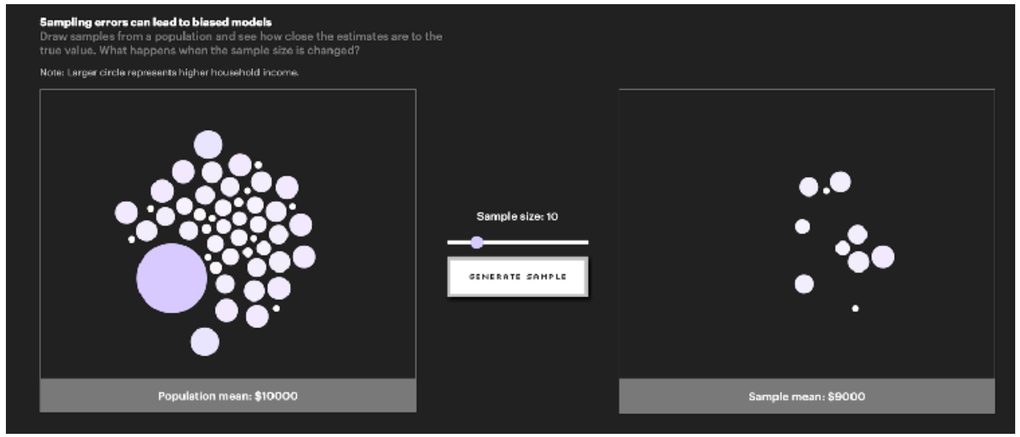


Animation

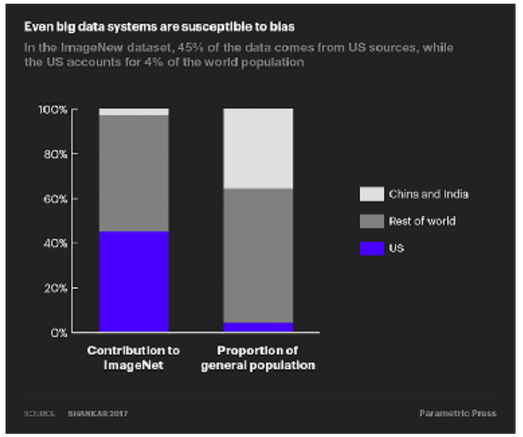




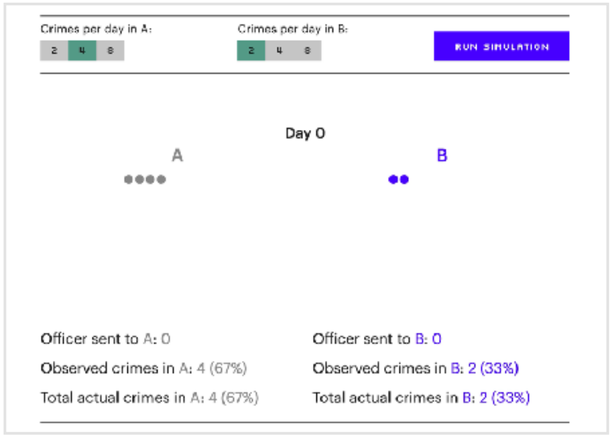
Interactive Graphics



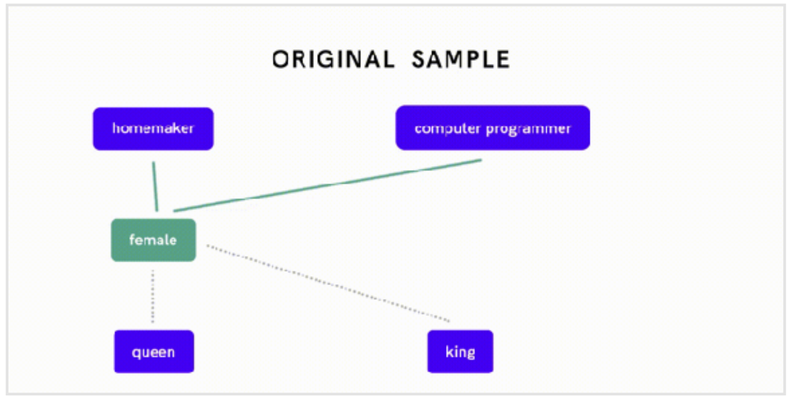
Visualization

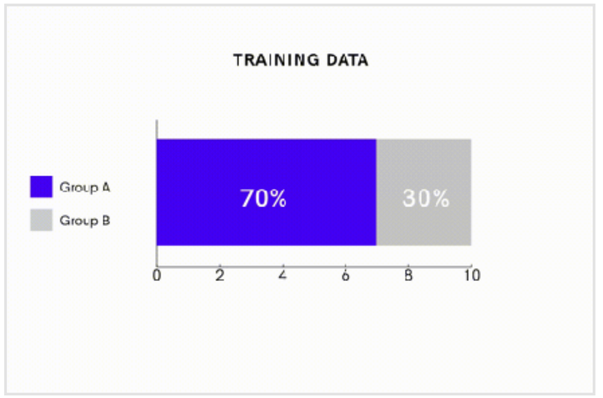


Simulation

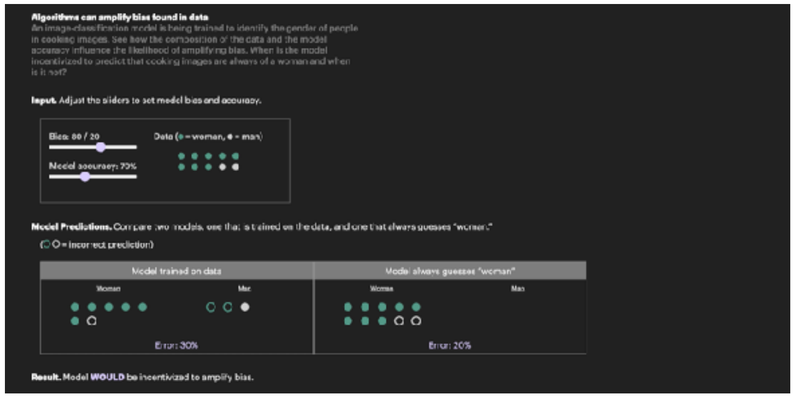


Animation





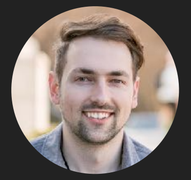
Modeling



Co-launched Parametric Press

Interactive publishing platform & zine to break down complex topics.

Matthew Conlen
University of Washington



Fred Hohman
Georgia Tech

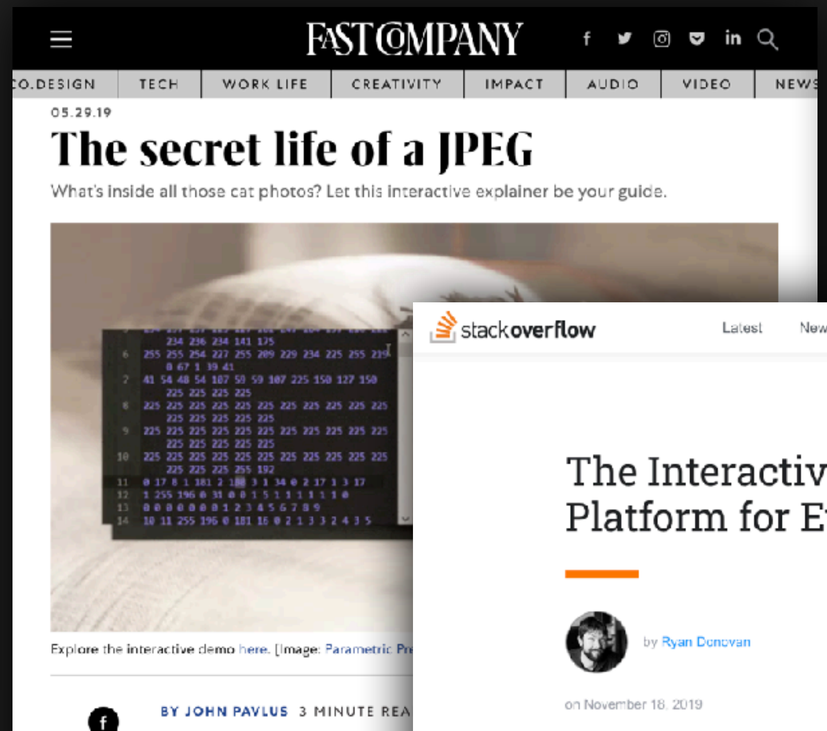
`parametric.press`

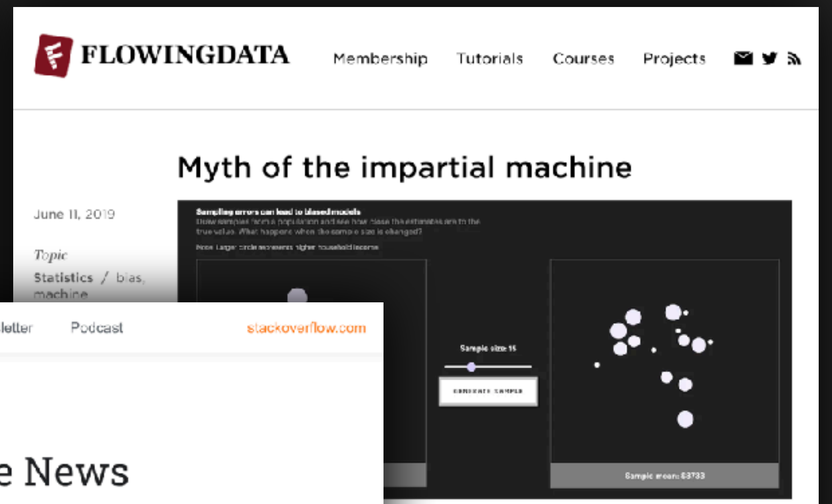
VISCOMM 2019

200,000+ views over 1st year

Top of Hacker News

Featured on r/programming, r/coding, r/FrontEnd





People loved it!

Helped Teach People About ML's Impact

jonathan.ai
@jonathandinu

First issue of @ParametricPress is 🔥🔥🔥

"The Myth of the Impartial Machine" is one of the pithiest and accessible treatments of bias in machine learning... and the interactivity puts it in a league of its own

Jon Fisher
@sciencejon

This is a fascinating and really important read about how AI and machine learning can amplify bias if not corrected:

Patrick Cleary
@PatJCleary

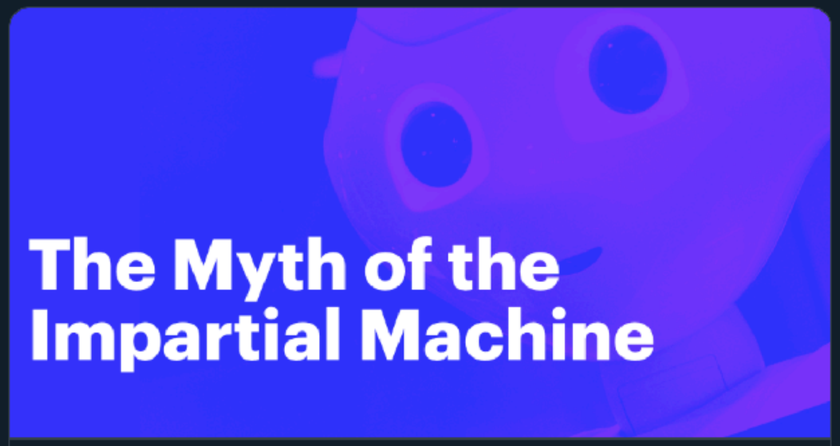
An excellent overview of how many biases can factor into "impartial" #ai systems, even despite the best of intentions.

Brett2point0
@Brett2point0

Sharing my Top 5 favorite #AI reads of 2019, crucial education for 2020 and beyond: "The Myth of the Impartial Machine"

Zak Rogoff
@Zakkai

Interactive article with sliders and infographics that teaches how #AI bias works.



The Myth of the Impartial Machine
Wide-ranging applications of data science bring utopian proposals of a world free from bias, but in reality, machine learning models reproduce the inequalities that ...
parametric.press

Frank Ostermann
@f_ostermann

Great interactive introduction to AI/ML and what can possibly go wrong (and how to fix it):

Dheeraj Tommandru
@realtlsdheeraj

This is great article explaining biases from data gathering to model productionization, with a lot of interesting visual and interactive graphs.

Brent Roraback
@doublepower

Really good piece - with great interactive visuals - from @ParametricPress on biases in machine learning models.

Callum Flack
@callumflack

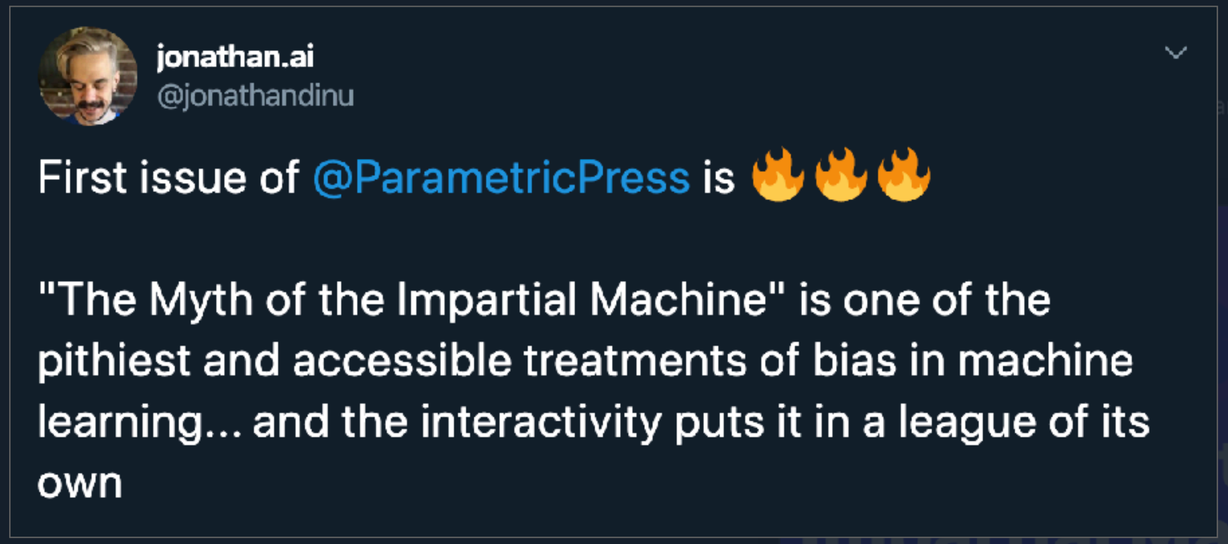
Yum, @ParametricPress. A pertinent article, visualised with D3. And then they have all their articles on Github: open publishing, file an issue! This is smart and it looks great.

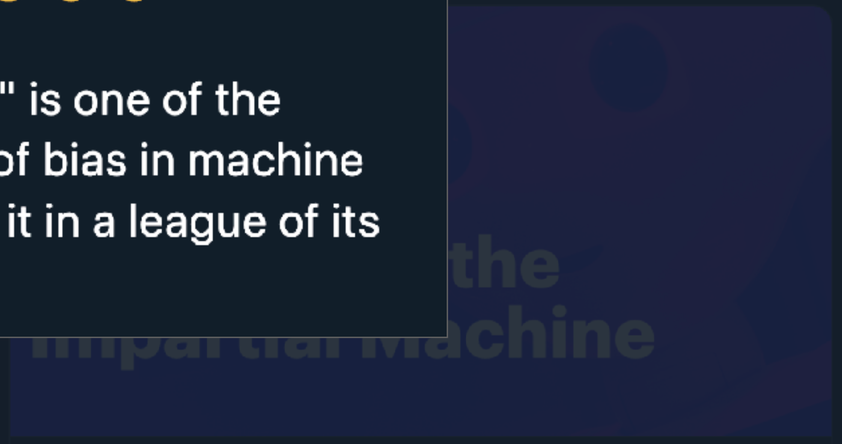
Chelsea Waite
@chelseawaite

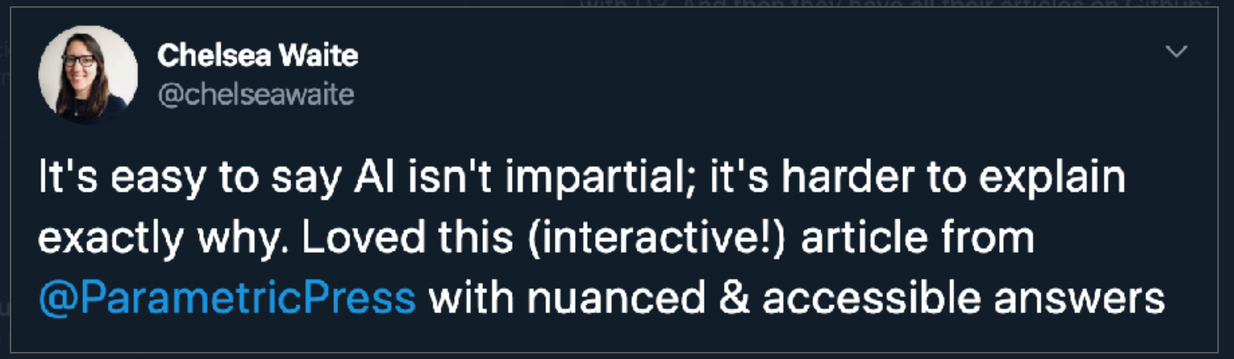
It's easy to say AI isn't impartial; it's harder to explain exactly why. Loved this (interactive!) article from @ParametricPress with nuanced & accessible answers

People loved it!

Helped Teach People About ML's Impact

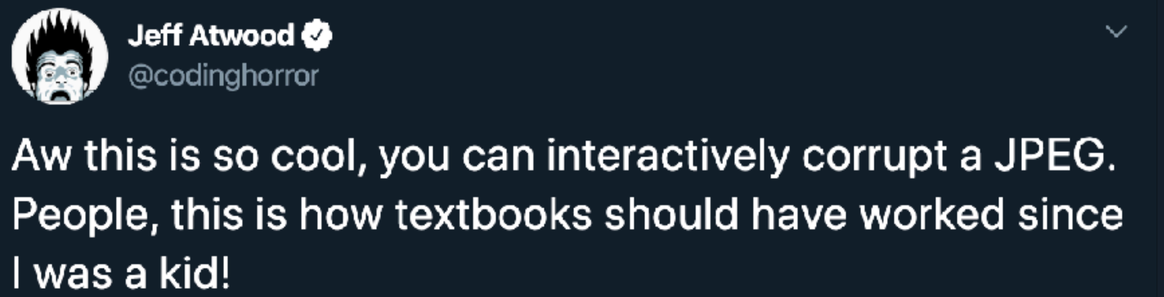






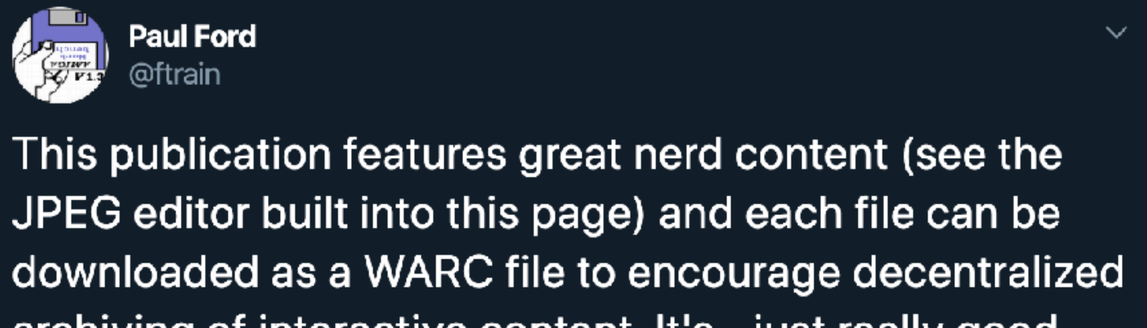
People loved it!

Glimpse at an Interactive Reading Future



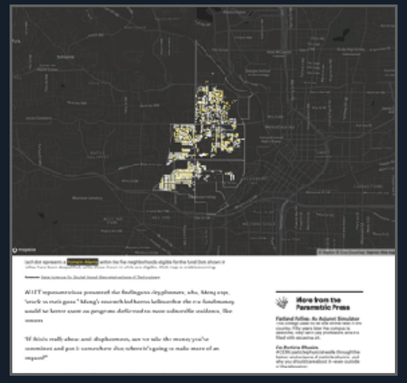
Jeff Atwood ✓
@codinghorror

Aw this is so cool, you can interactively corrupt a JPEG. People, this is how textbooks should have worked since I was a kid!



Paul Ford
@ftrain

This publication features great nerd content (see the JPEG editor built into this page) and each file can be downloaded as a WARC file to encourage decentralized archiving of interactive content. It's...just really good and thoughtful.



Interact with systems—no setup required. 100% open-source. Article archival.

Interactive Articles: Contribution 1

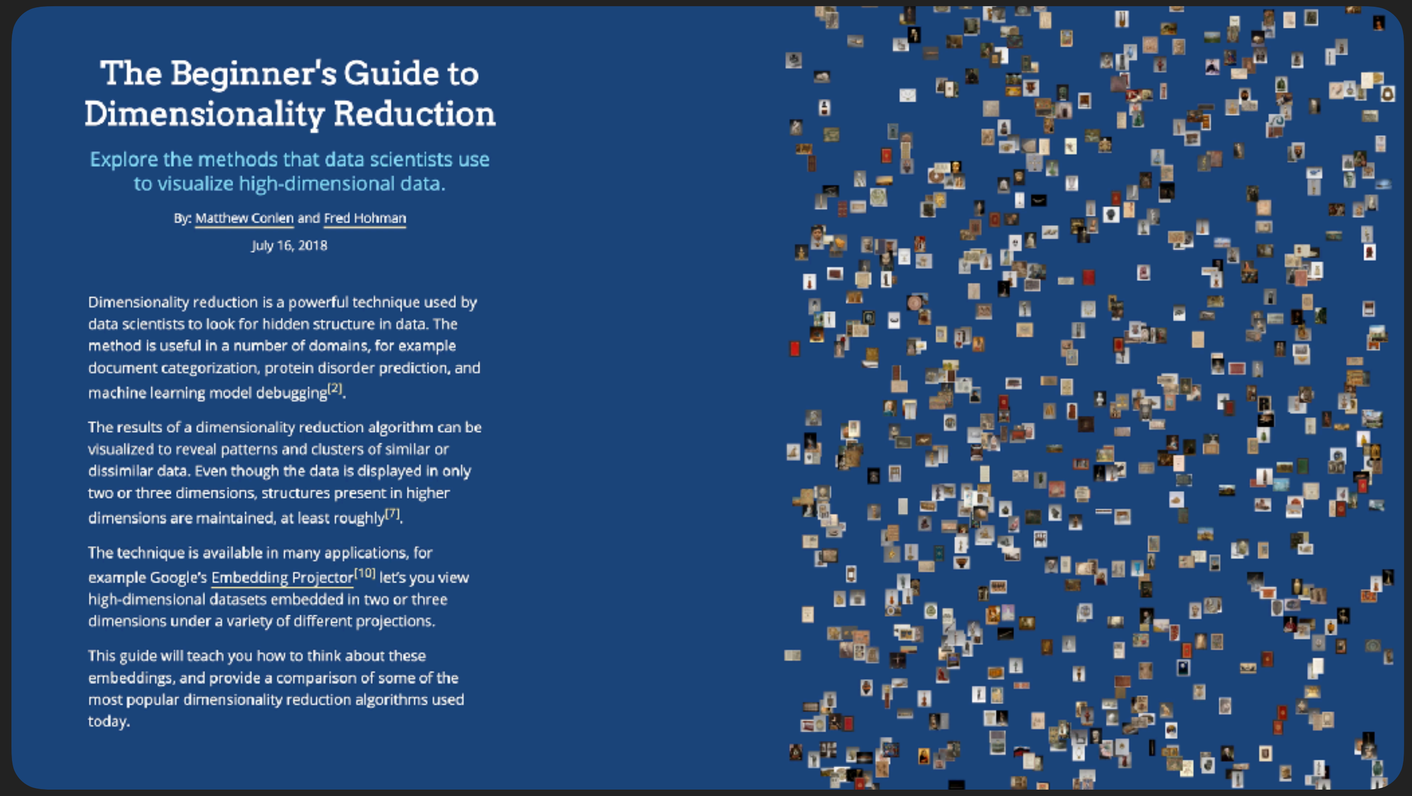
Help broad audiences learn about machine learning's impact on their lives.

Understanding (Interactive) Reading Behavior

The Beginner's Guide to Dimensionality Reduction

Matthew Conlen
University of Washington

Fred Hohman
Georgia Tech



The Beginner's Guide to Dimensionality Reduction

Explore the methods that data scientists use to visualize high-dimensional data.

By: Matthew Conlen and Fred Hohman

July 16, 2018

Dimensionality reduction is a powerful technique used by data scientists to look for hidden structure in data. The method is useful in a number of domains, for example document categorization, protein disorder prediction, and machine learning model debugging[2].

The results of a dimensionality reduction algorithm can be visualized to reveal patterns and clusters of similar or dissimilar data. Even though the data is displayed in only two or three dimensions, structures present in higher dimensions are maintained, at least roughly[7].

The technique is available in many applications, for example Google's Embedding Projector[10] let's you view high-dimensional datasets embedded in two or three dimensions under a variety of different projections.

This guide will teach you how to think about these embeddings, and provide a comparison of some of the most popular dimensionality reduction algorithms used today.

VISXAI 2018 🏆

Recorded reader interactions

7,000+ views over 1st month

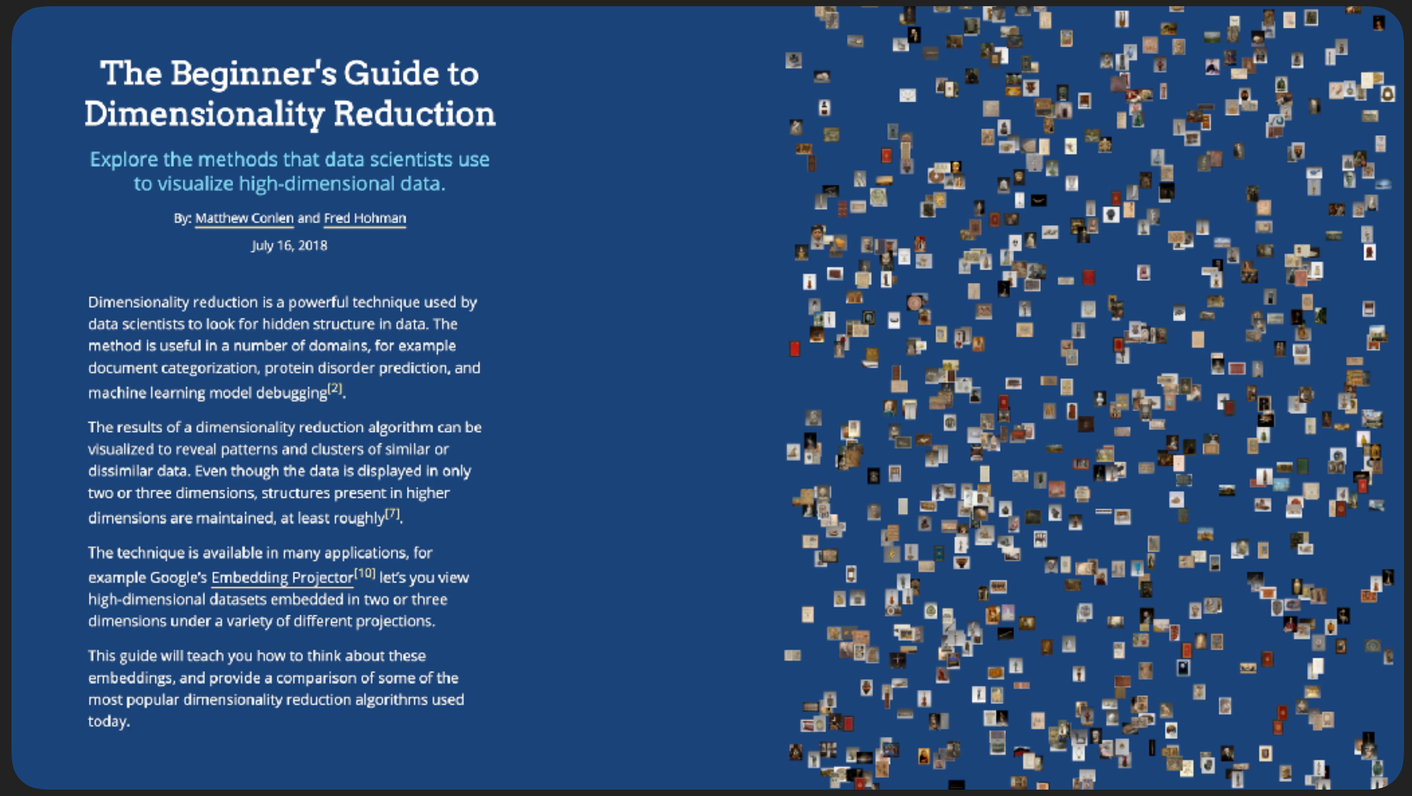
#1 of r/datascience subreddit

Understanding (Interactive) Reading Behavior

The Beginner's Guide to Dimensionality Reduction

Matthew Conlen
University of Washington

Fred Hohman
Georgia Tech



The Beginner's Guide to Dimensionality Reduction

Explore the methods that data scientists use to visualize high-dimensional data.

By: Matthew Conlen and Fred Hohman

July 16, 2018

Dimensionality reduction is a powerful technique used by data scientists to look for hidden structure in data. The method is useful in a number of domains, for example document categorization, protein disorder prediction, and machine learning model debugging[2].

The results of a dimensionality reduction algorithm can be visualized to reveal patterns and clusters of similar or dissimilar data. Even though the data is displayed in only two or three dimensions, structures present in higher dimensions are maintained, at least roughly[7].

The technique is available in many applications, for example Google's Embedding Projector[10] let's you view high-dimensional datasets embedded in two or three dimensions under a variety of different projections.

This guide will teach you how to think about these embeddings, and provide a comparison of some of the most popular dimensionality reduction algorithms used today.

VISXAI 2018 🏆

Recorded reader interactions

7,000+ views over 1st month

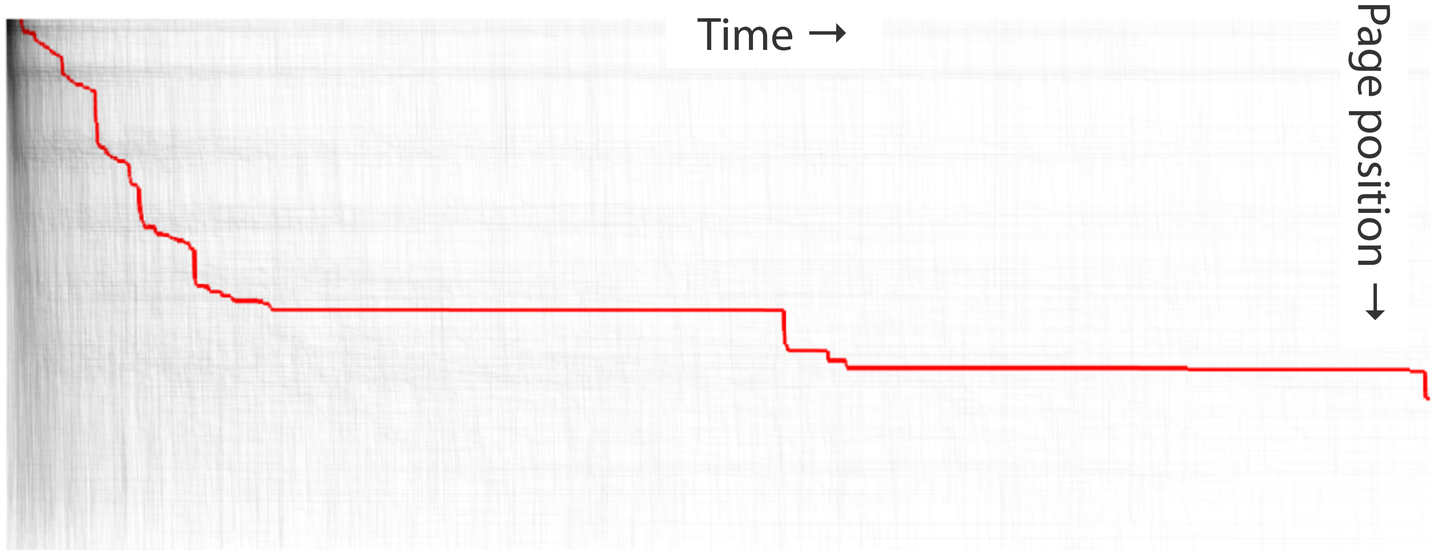
#1 of r/datascience subreddit

Discovery #1

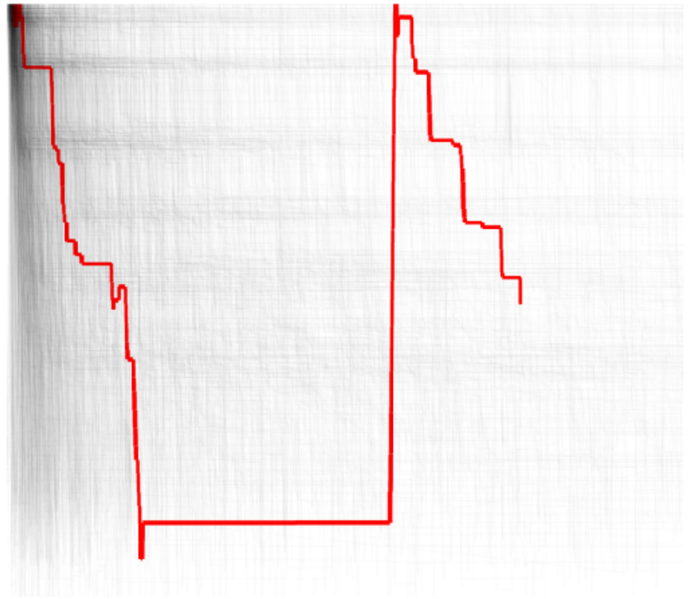
Revealing Different Reader Patterns

Critical, yet understudied: How do people read interactive content?

A. “Scroll & stop”



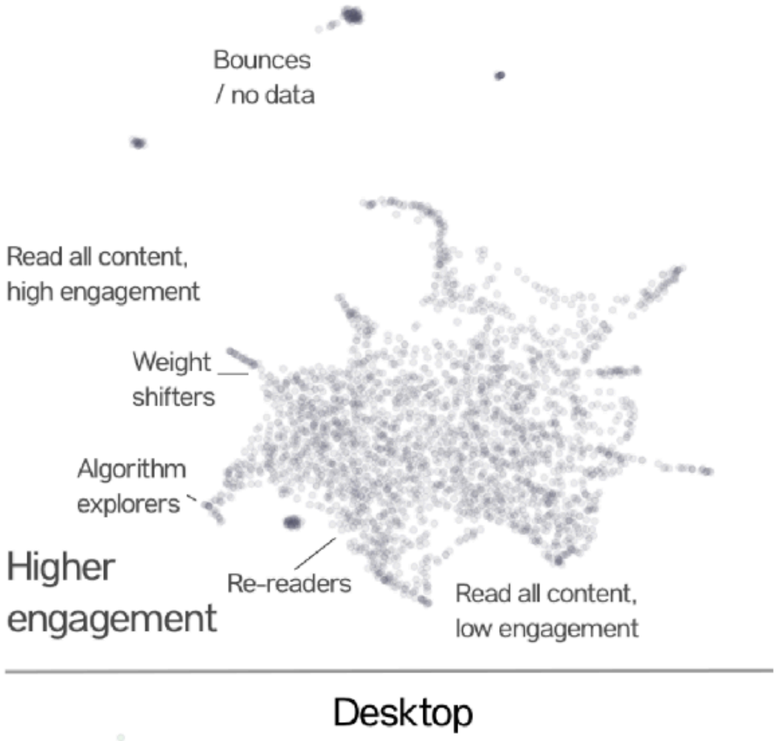
B. “Preview & read”

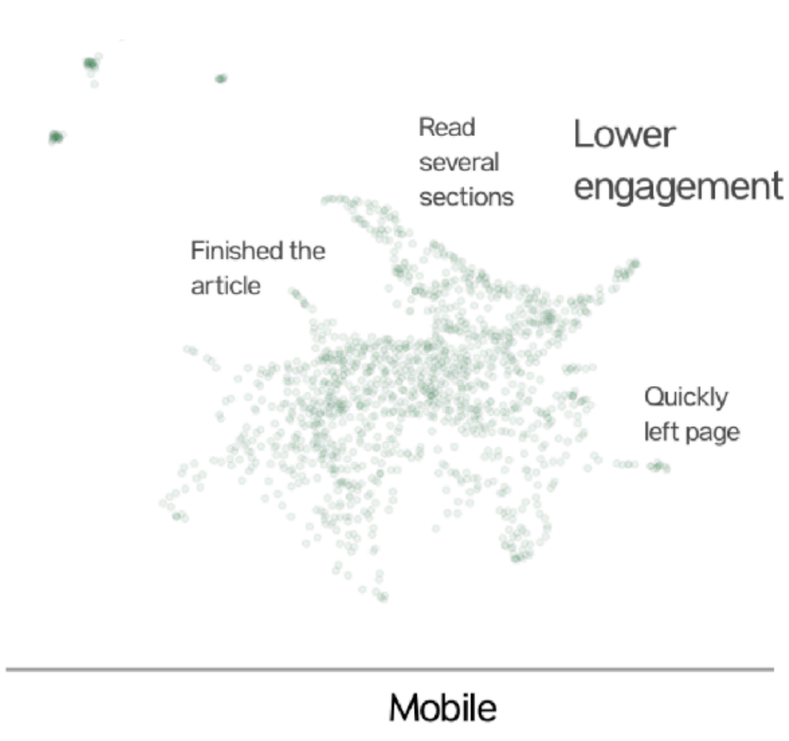


Discovery #2

Enable Macro-analysis of Readership

Reveals broad clusters of reading behavior across device formats.

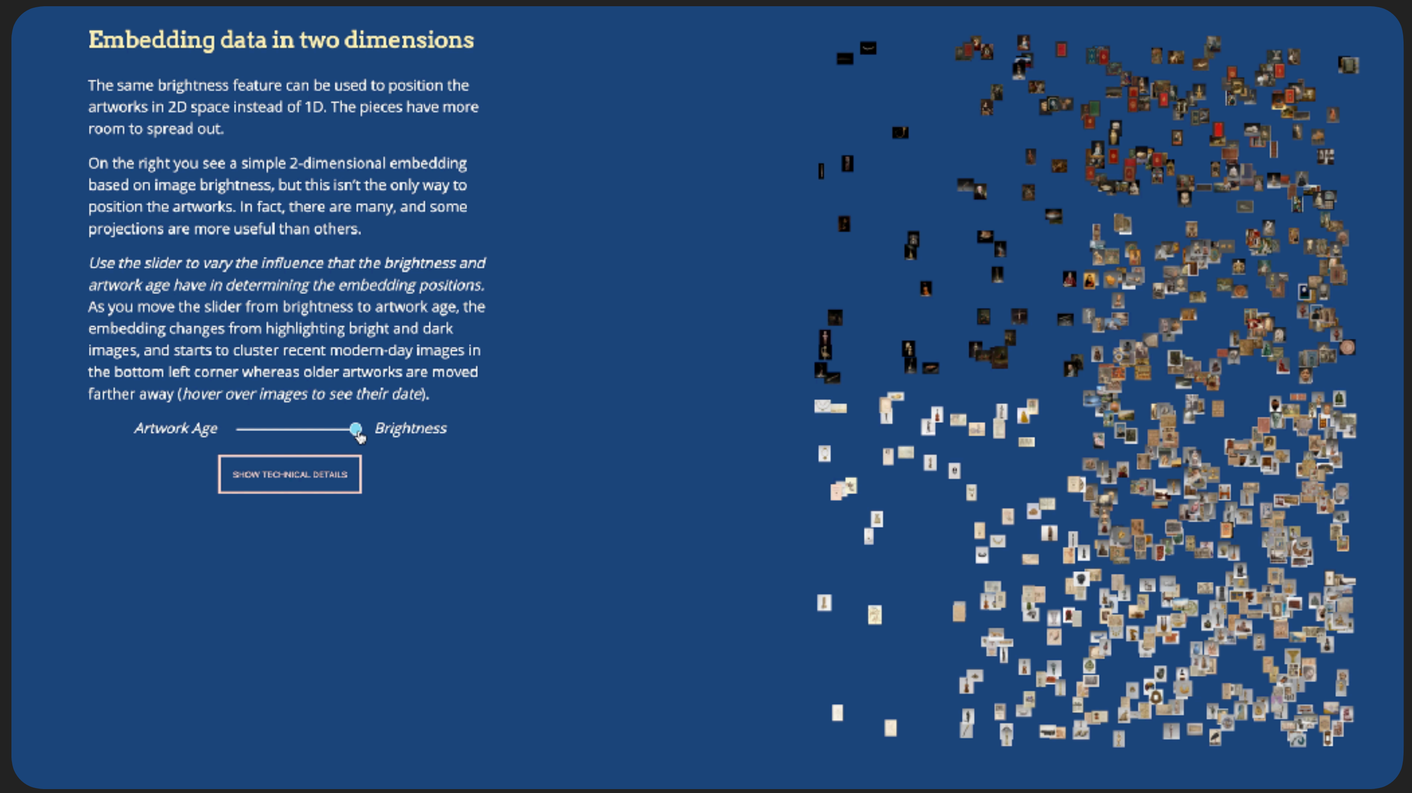


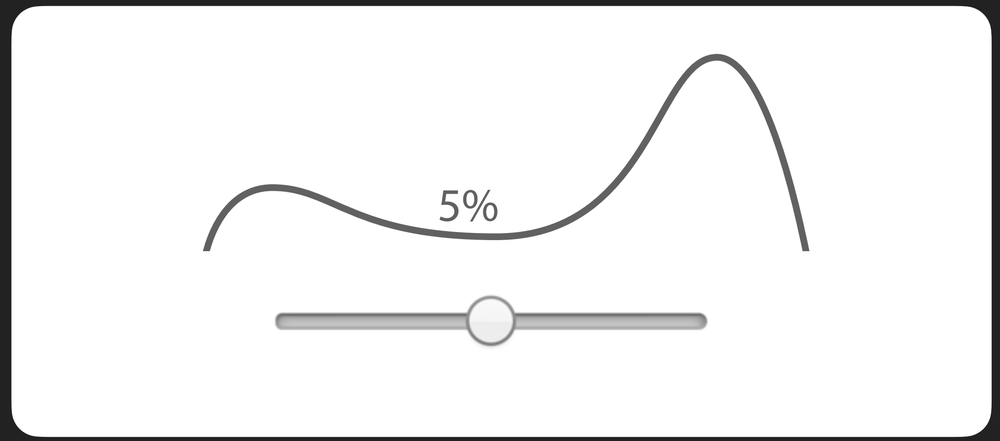


Discovery #3

Content-specific Design Evaluation

People can read in ways that don't match original design intention.





Minimum value distribution

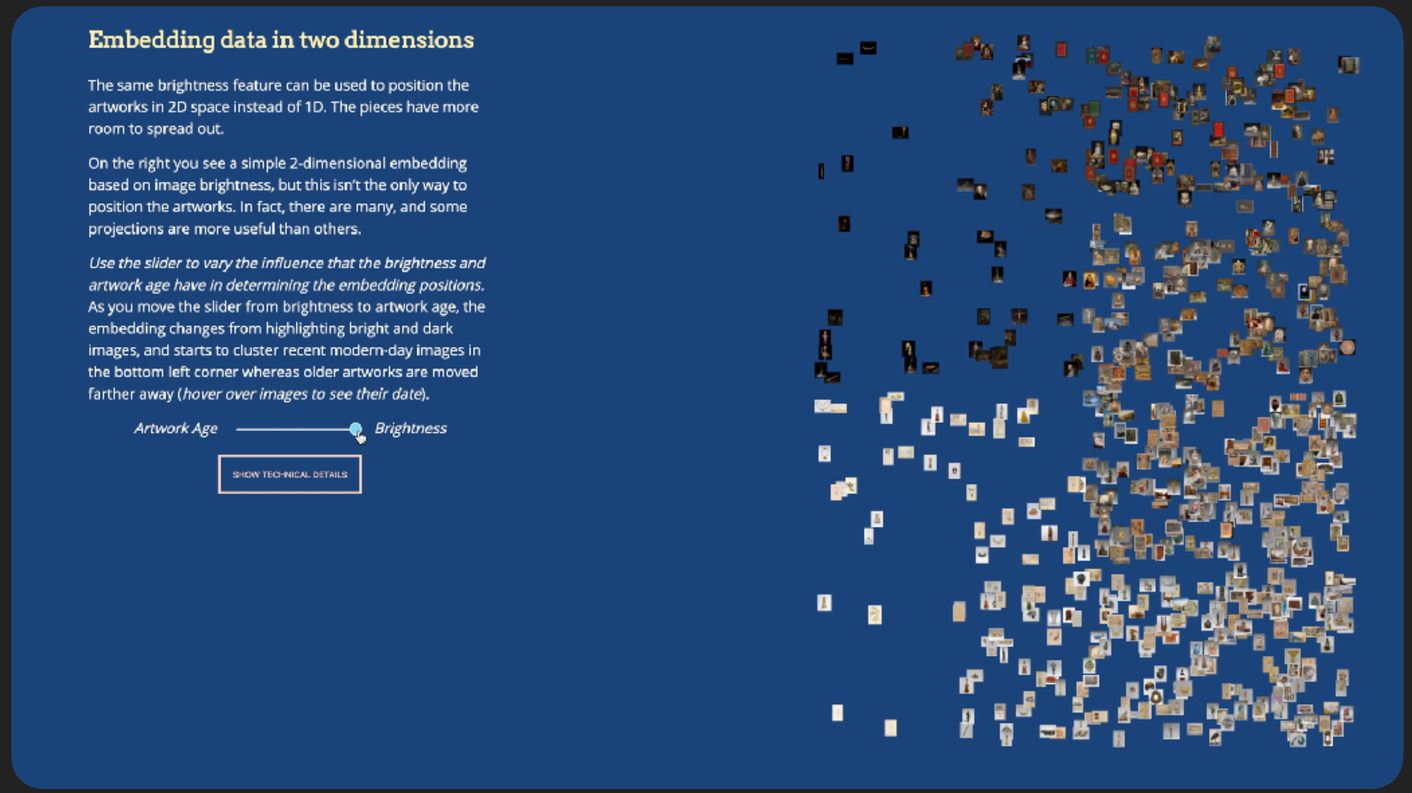
→ *This could've been a toggle?*

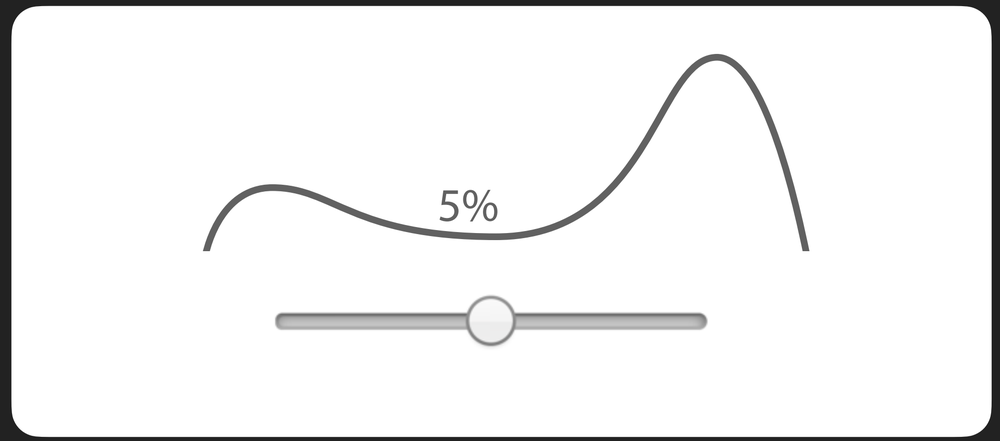
→ *Nudge readers towards the middle?*

Discovery #3

Content-specific Design Evaluation

People can read in ways that don't match original design intention.





Minimum value distribution

→ *This could've been a toggle?*

→ *Nudge readers towards the middle?*

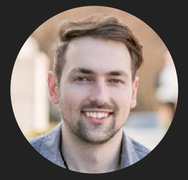
Interactive Articles: Contribution 1

Help broad audiences learn about machine learning's impact on their lives.

Interactive Articles: Contribution 2

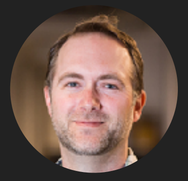
Encourage active reading and engage people with learning material.

Theory + Practice for Interactive Articles

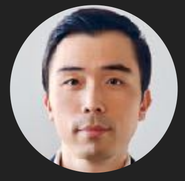


Fred Hohman
Georgia Tech

Matthew Conlen
Univ. of Washington



Jeffrey Heer
Univ. of Washington



Polo Chau
Georgia Tech

Why care about interactives?

- Better learning & engagement
- Highlight important topics
- Attract broad readership & acclaim

Distill 2020, under review

Communicating with Interactive Articles

Examining the design of interactive articles by synthesizing theory from disciplines such as education, journalism, and visualization.

Contents

Introduction
Interactive Articles: Theory and Practice
Engagement and Emotion
Interactive Visualization
Improving Recall
Personalization
Context in Abstract
Critical Reflections
Challenges of Interactive Writing
Looking Forward

Computing has changed how people communicate. The transmission of news, messages, and ideas is instant. Anyone's voice can be heard. In fact, access to digital communication technologies such as the internet is so fundamental to daily life that their disruption by government is condemned by the United Nations Human Rights Council [1]. But while the technology to distribute our ideas has grown in scope and bounds, the interfaces have remained largely the same.

Parallel to the development of the internet, researchers like Alan Kay and Douglas Engelbart worked to build technology that would empower individuals and enhance cognition. Kay imagined the Dynabook [2] in the hands of children across the world. Engelbart, while best remembered for his "mother of all demos," was more interested in the ability of computation to augment human intellect [3]. Neal Stephenson wrote speculative fiction that imagined interactive paper that could display videos and interfaces, and books that could teach and respond to their readers [4].

More recent (yet still historical by personal computing standards) designs point to a future where computers are connected and assist people in decision-making and communicating using rich graphics and interactive user interfaces [5]. While some technologies have seen mainstream adoption, such as Hypertext [6], unfortunately, many others have not. The most popular publishing platforms (for example, WordPress and Medium) choose to prioritize social features and ease-of-use while limiting the ability for authors to communicate using the dynamic features of the web.

In the spirit of previous computer-assisted cognition technologies, a new type of computational communication medium has emerged that leverages active reading techniques to make ideas more accessible to a broader range of people. These interactive articles build on a long history, from Parkes [7] to PHOT [8] to explorable explanations [9]. They have been shown to be more engaging, can help improve recall and learning, and attract broad readership and acclaim [10], yet we do not know that much about them.

Because diverse communities create interactive content, this medium goes by many different names and has not yet settled on a standardized format nor definition [11]. For newsrooms, data journalists, developers, and designers work together to make complex news and investigative reporting clear and engaging using interactive stories [12]. Researchers have proposed artifacts such as explorable multiverse analyses [13], explanadoes [14], and explanations [15] to more effectively disseminate their work, communicate their results to the public, and remove research debt [16]. Educators use interactive textbooks as an alternative learning format to give students hands-on experience with learning material [17].

Besides these groups, others such as academics, game developers, web developers, and designers blend editorial, design, and programming skills to create and publish explorable explanations [9], interactive fiction [18], interactive non-fiction [19], active essays [20], and interactive games [21]. While these all slightly differ in their technical approach and target audience, they all largely leverage the interactivity of the modern web.

Research Dissemination | Journalism | Education | Policy and Decision Making

Research Dissemination

Conducting novel research requires deep understanding and expertise in a specific area. Once achieved, researchers continue contributing new knowledge for future researchers to use and build upon. Over time, this consistent addition of new knowledge can build up, contributing to what some have called research debt. Not everyone is an expert in every field, and it can be easy to lose perspective and forget the bigger picture. Yet research should be understood by many. Interactive articles can be used to distill the latest progress in various research fields and make their methods and results accessible and understandable to a broader audience.

OPPORTUNITIES

- Engage and excite broader audience with latest research progress
- Remove research debt, onboard new researchers
- Make faster and clearer research progress

CHALLENGES

- No clear incentive structure for researchers
- Little funding for bespoke research dissemination and communication
- Not seen as a legitimate research contribution (e.g., to the field, or one's career)

Attacking Discrimination with Smarter Machine Learning [24]

A companion piece to a traditional research paper that uses interactive visualizations to let readers adjust a machine learning model's behavior and explore alternative fairness strategies for a loan granting scenario.

Coeffects: Context-aware Programming Languages [25]

A PhD thesis that contributes a programming language sketcpad for experimenting new programs across the context or environment in which they execute, and walks readers through the web using two simple context-aware languages with live in-browser demos.

What is Complexity Science? [23]

A mock course in complex systems science, created by leading experts, practitioners, and students in the field, with accompanying interactive sandboxes to simulate, control, and visualize different complex systems.

Today there is a growing excitement around the use of interactive articles for communication

Interactive Articles: Contribution 1

Help broad audiences learn about machine learning's impact on their lives.

Interactive Articles: Contribution 2

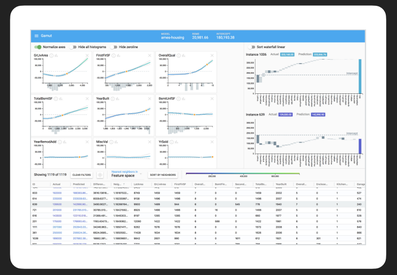
Encourage active reading and engage people with learning material.

Interactive Articles: Contribution 3

Synthesize unique capabilities and formalize research challenges in diverse domains.

Impact

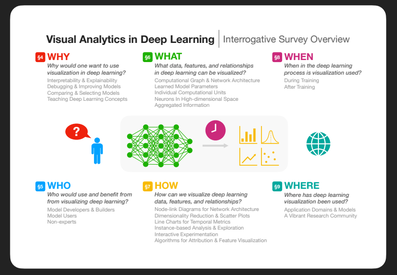
Enable interpretability

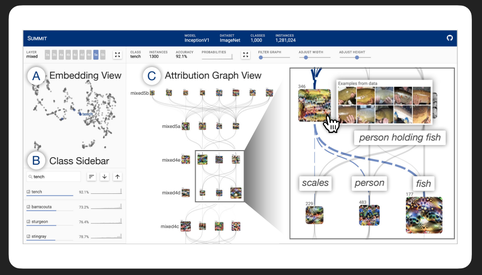


GAMUT + TELEGAM

Deployed and **demoed** at Microsoft Research
Integrated in popular toolkits: **InterpretML**, **SandDance**

Scale interpretability





Interrogative Survey + SUMMIT

Recognized by **NASA PhD Fellowship**
First-of-its-kind survey for deep learning visualization

Communicate interpretability

PARAMETRIC PRESS + Interactive Articles

Helped **250,000+ people** learn about machine learning
Synthesis & formalization of interactive articles

Future Research Vision

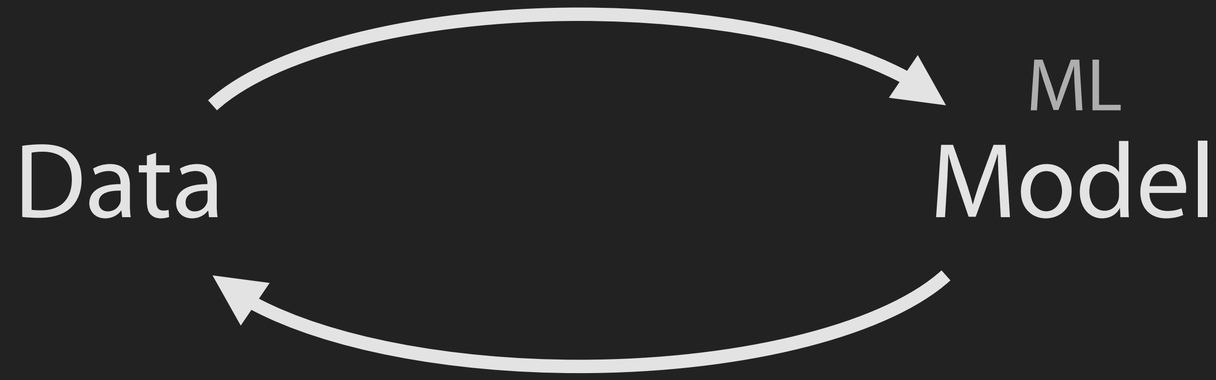
HCI + AI

Research Thrust #1

Mixed-initiative Model Development

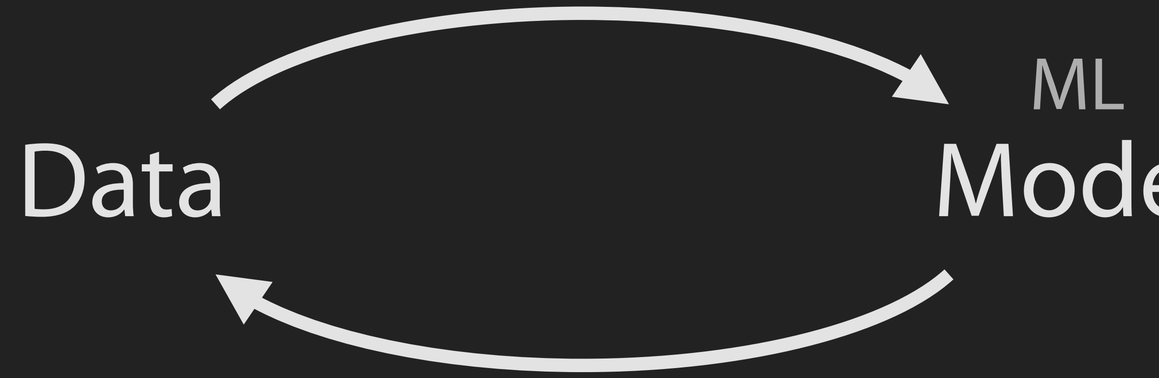
Research Thrust #1

Mixed-initiative Model Development



Research Thrust #1

Mixed-initiative Model Development



CHAMELEON *CHI 2020*

Understanding and Visualizing Data Iteration in Machine Learning

Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, Kayur Patel. *CHI, 2020.*

FAIRVIS *VAST 2019*

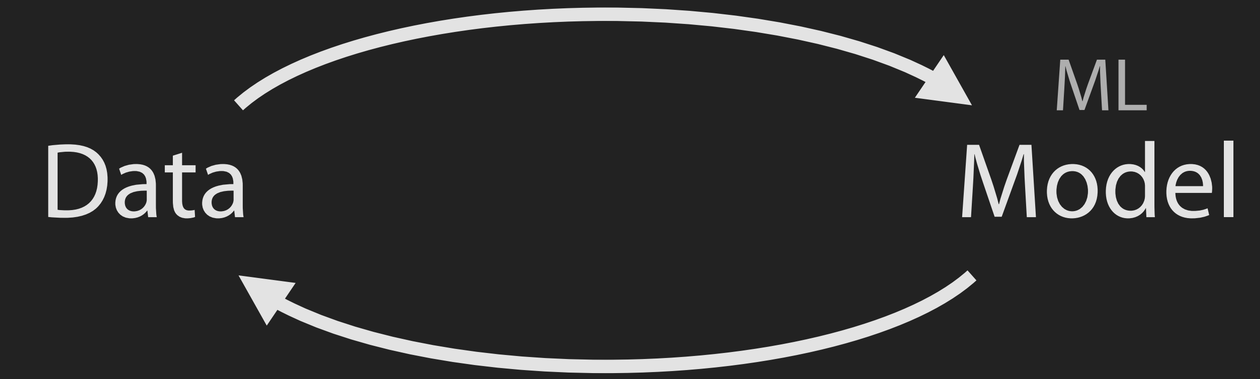
Visual Analytics for Discovering Intersectional Bias in Machine Learning

Angel Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, Duen Horng (Polo) Chau. *VAST, 2019.*

Research Thrust #1

Mixed-initiative Model Development

CHAMELEON *CHI 2020*



FAIRVIS *VAST 2019*

RESEARCH DIRECTIONS

Interactive Data Programming: Visualize and track updating labels.

Augmented Models: New models for underperforming data subgroups.

Unit Tests for Interpretability: Specification for learned rules.

Research Thrust #2

Making Interpretability Common Practice

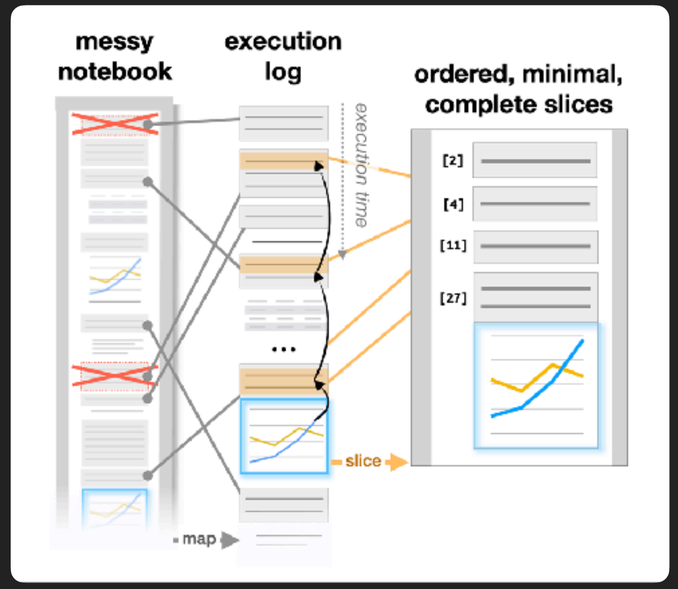
RESEARCH DIRECTIONS

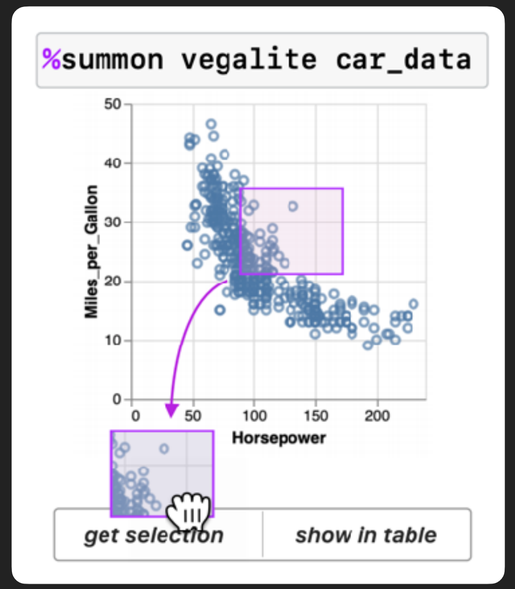
Integrate interpretability into existing code environments.

Generalize to create best practices for interpretability.

Preserve healthy analysis skepticism with explainable AI.

PRELIMINARY WORK





Managing Messes in Computational Notebooks
Andrew Head, Fred Hohman, Titus Barik, Steven Drucker, Robert DeLine. *CHI, 2019.*

The Future of Notebook Programming Is Fluid
Mary Beth Kery, Donghao Ren, Kanit Wongsuphasawat, Fred Hohman, Kayur Patel. *CHI, 2020.*

Research Dissemination

	Paper	Video	Code	Slides	Demo + Article
GAMUT CHI 2019	✓	✓		✓	✓
TELEGAM VAST 2019	✓	✓	✓	✓	✓
Interrogative Survey TVCG 2018	✓	✓	✓	✓	✓
SUMMIT TVCG 2020	✓	✓	✓	✓	✓
Parametric Press 2019	✓	✓	✓	✓	✓
Dimensionality Reduction VISxAI 2018	✓		✓	✓	✓
Interactive Articles Distill 2020	✓		✓		✓
CHAMELEON CHI 2020	✓	✓		○	○
FAIRVIS VAST 2019	✓	✓	✓	✓	✓
Code Gathering CHI 2019	✓	✓	✓	✓	✓
Notebook Handoff CHI 2020 poster	✓	✓	✓	○	○

Research Thrust #3

Accessible Research Distillation

Challenges

Technical: diverse devices, accessible archival

Design: authoring requires multiple skillsets

Legitimization: few incentives, time consuming

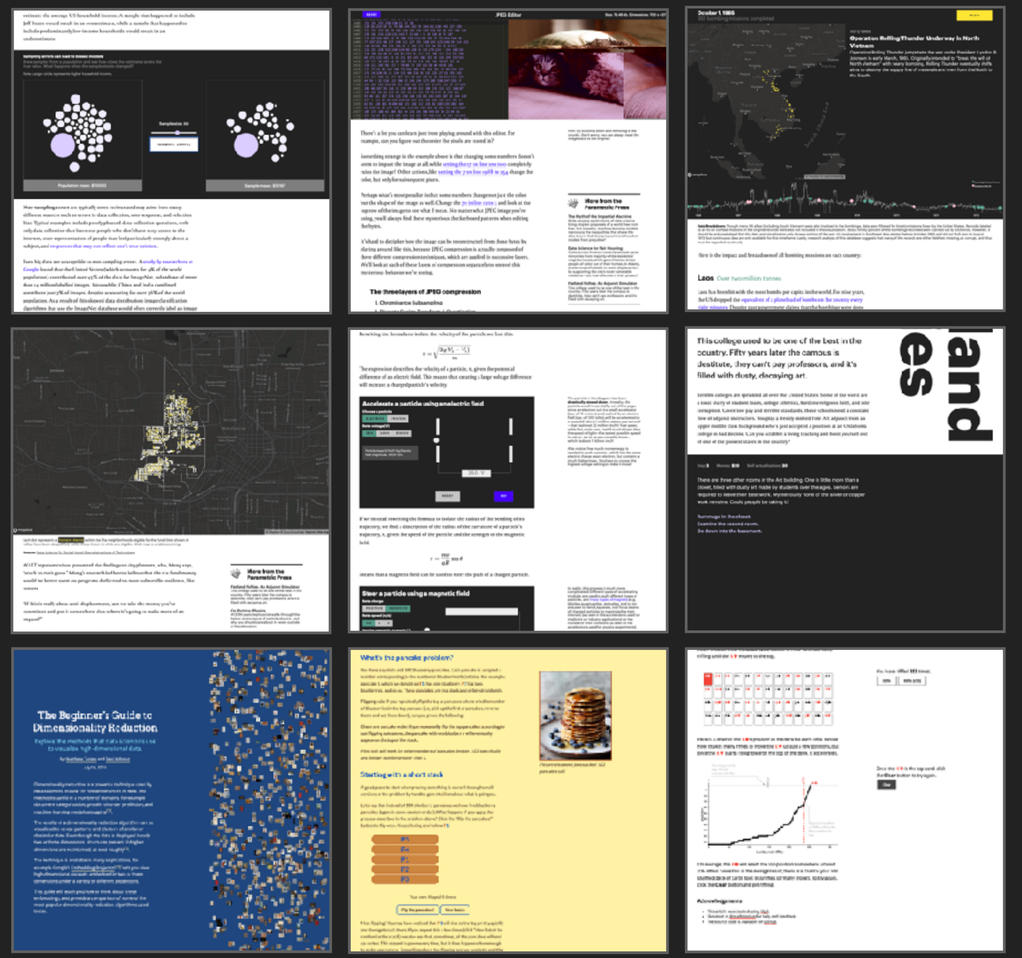
RESEARCH DIRECTIONS

Lower cost of creation with authoring tools and guidelines.

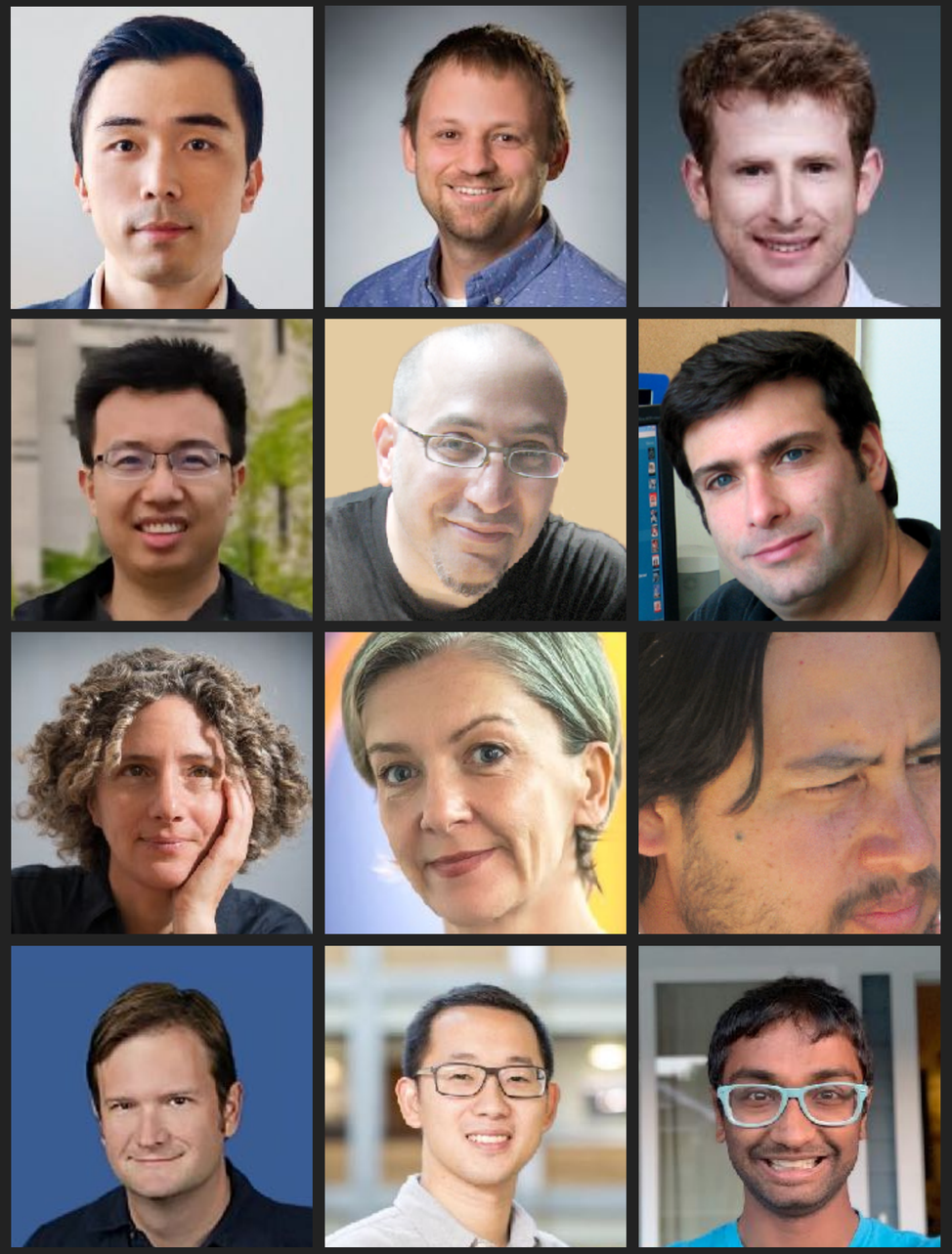
Transform ML education at scale.

Cultivate research communication culture.

BUILD ON MY EXPERIENCE



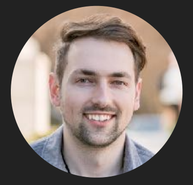
Thank you to collaborators from academia, industry, and the government!



NASA PhD Fellowship

Georgia Tech Visualization Lab

Interactive Scalable Interfaces for Machine Learning Interpretability



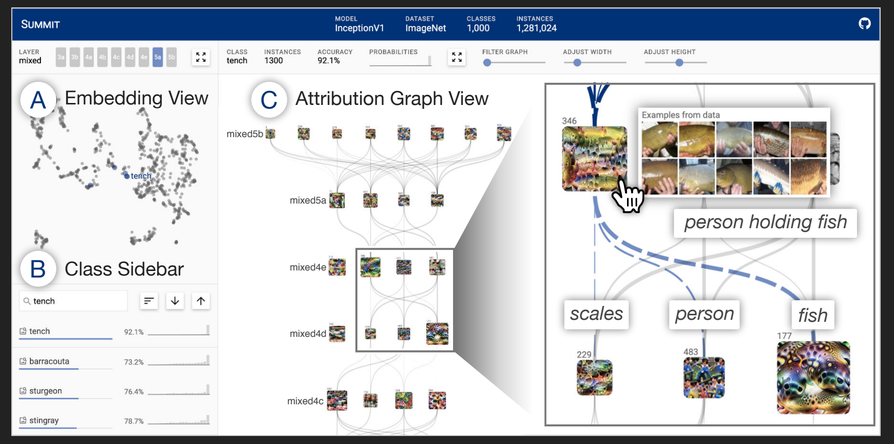
Fred Hohman @fredhohman
fredhohman.com

Georgia Tech

NASA

Enabling interpretability at scale and for everyone.

GAMUT & TELEGAM



Interrogative Survey & SUMMIT

PARAMETRIC PRESS & Interactive Articles