Visual Analytics in Deep Learning

An Interrogative Survey for the Next Frontiers

TVCG 2018 Survey

Fred Hohman
@fredhohman

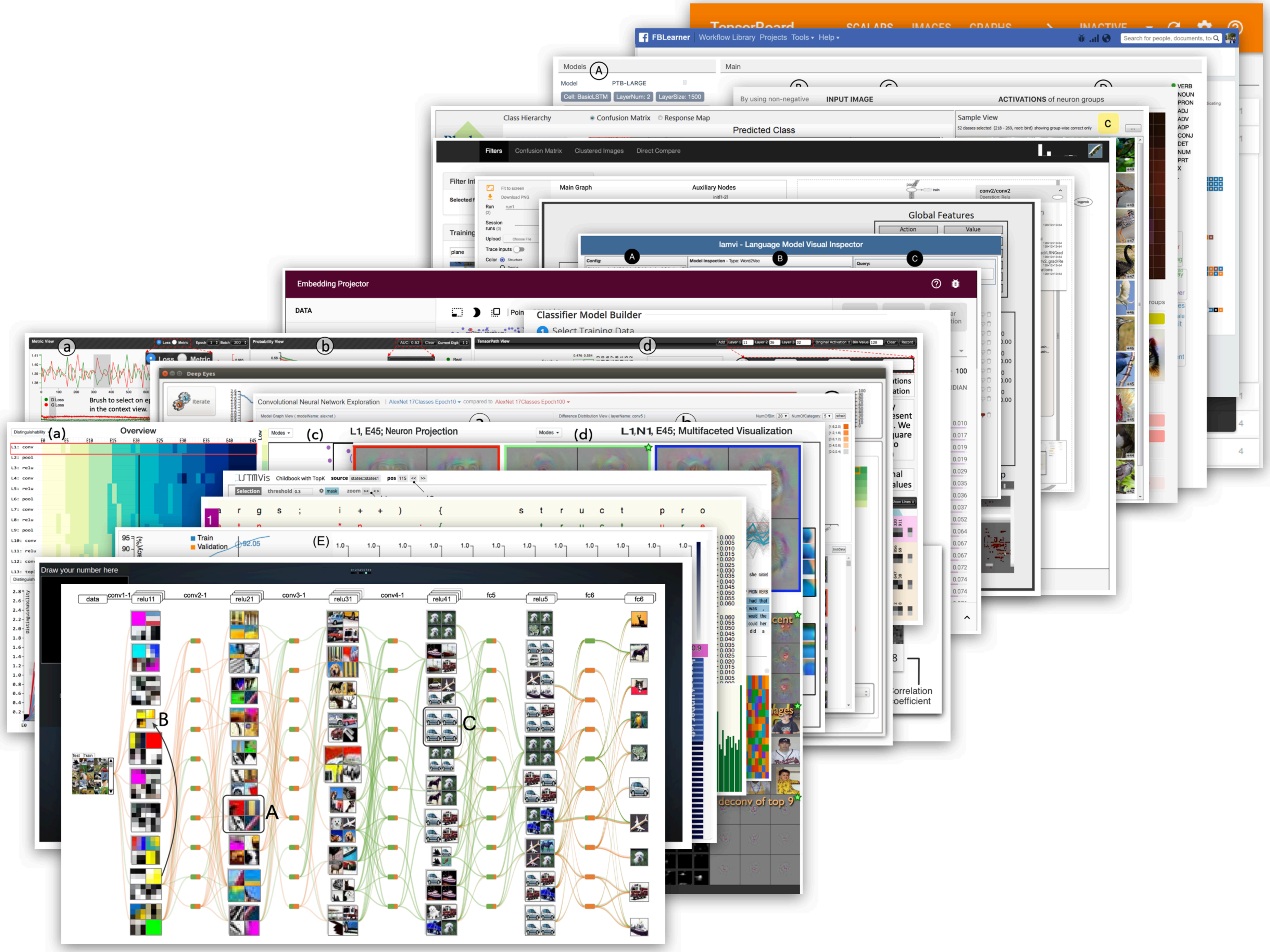
Minsuk Kahng

Robert Pienta

Polo Chau

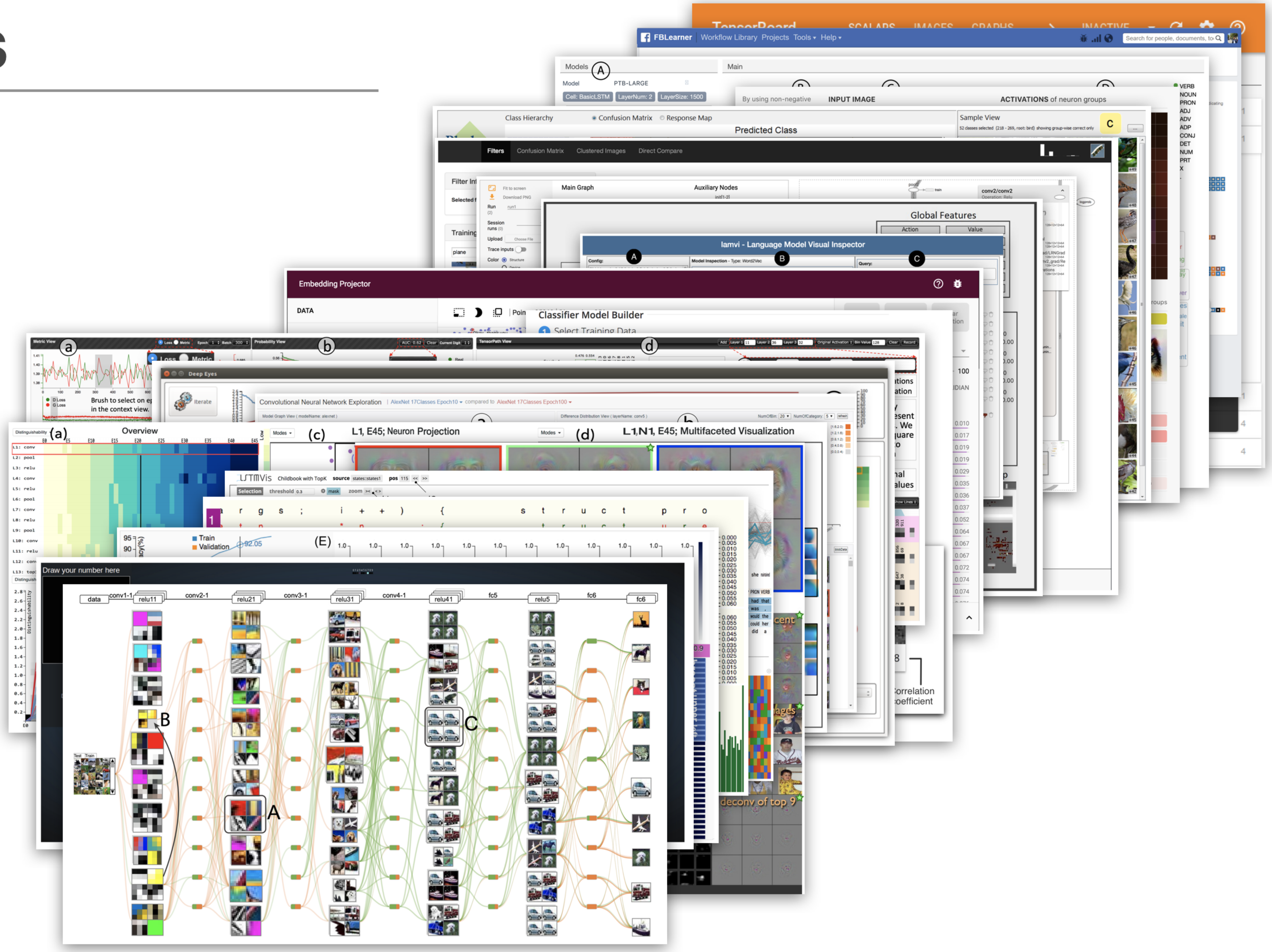
Georgia Tech

Visual Analytics in Deep Learning

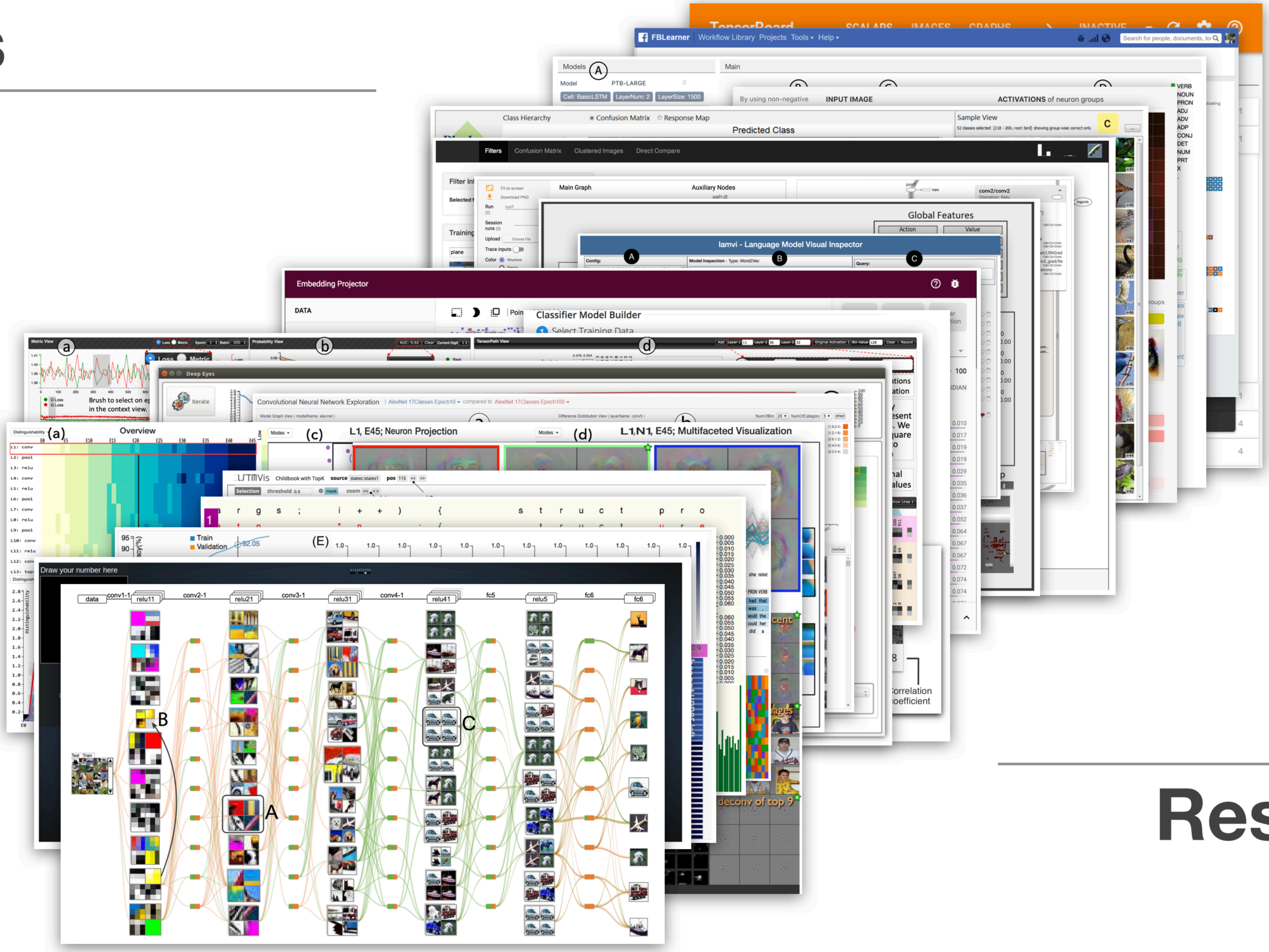


FBLearner
Workflow Library
Projects
Tools
Help
Models
Model
PTB-LARGE
Main
INPUT IMAGE
ACTIVATIONS of neuron groups
Class Hierarchy
Confusion Matrix
Response Map
Predicted Class
Sample View
Filters
Confusion Matrix
Clustered Images
Direct Compare
Main Graph
Auxiliary Nodes
Global Features
Action
Value
lamvi - Language Model Visual Inspector
Embedding Projector
DATA
Classifier Model Builder
Select Training Data
Brush to select on
in the context view.
Deep Eyes
Convolutional Neural Network Exploration
Overview
L1, E45; Neuron Projection
L1,N1, E45; Multifaceted Visualization
Train
Validation
Draw your number here
data
conv1-1
relu11
conv2-1
relu21
conv3-1
relu31
conv4-1
relu41
fc5
relu5
fc6
fc6
deconv of top 9
Correlation
Coefficient

Research Trends



Research Trends



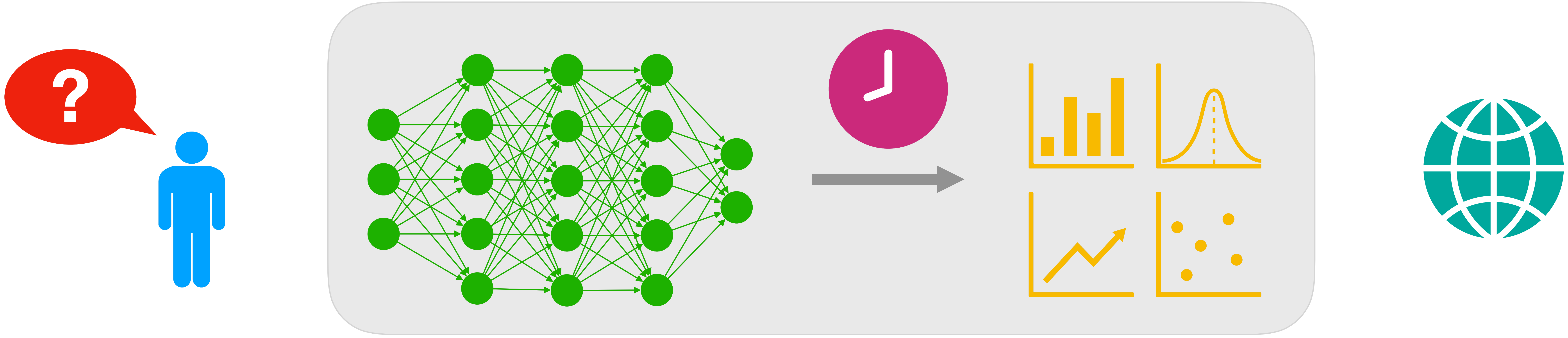
Research Directions

Visual Analytics in Deep Learning | Interrogative Survey Overview

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?

Visual Analytics in Deep Learning | Interrogative Survey Overview

WHY

Why would one want to use visualization in deep learning?

Interpretability & Explainability
Debugging & Improving Models
Comparing & Selecting Models
Teaching Deep Learning Concepts

WHAT

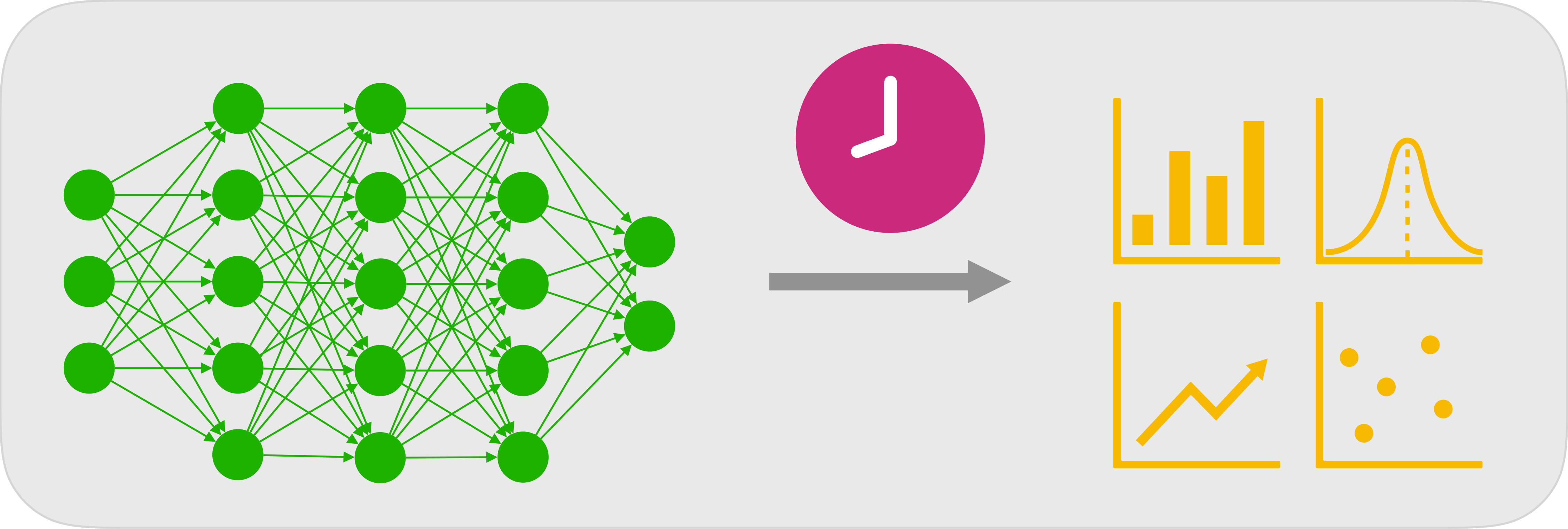
What data, features, and relationships in deep learning can be visualized?

Computational Graph & Network Architecture
Learned Model Parameters
Individual Computational Units
Neurons In High-dimensional Space
Aggregated Information

WHEN

When in the deep learning process is visualization used?

During Training
After Training



WHO

Who would use and benefit from visualizing deep learning?

Model Developers & Builders
Model Users
Non-experts

HOW

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

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

WHERE

Where has deep learning visualization been used?

Application Domains & Models
A Vibrant Research Community

Visual Analytics in Deep Learning | Interrogative Survey Overview

WHY

Interpretability & Explainability
Debugging & Improving Models
Comparing & Selecting Models
Teaching Deep Learning Concepts

WHAT

Computational Graph & Network Architecture
Learned Model Parameters
Individual Computational Units
Neurons In High-dimensional Space
Aggregated Information

WHEN

During Training
After Training

WHO

Model Developers & Builders
Model Users
Non-experts

HOW

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

WHERE

Application Domains & Models
A Vibrant Research Community

		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

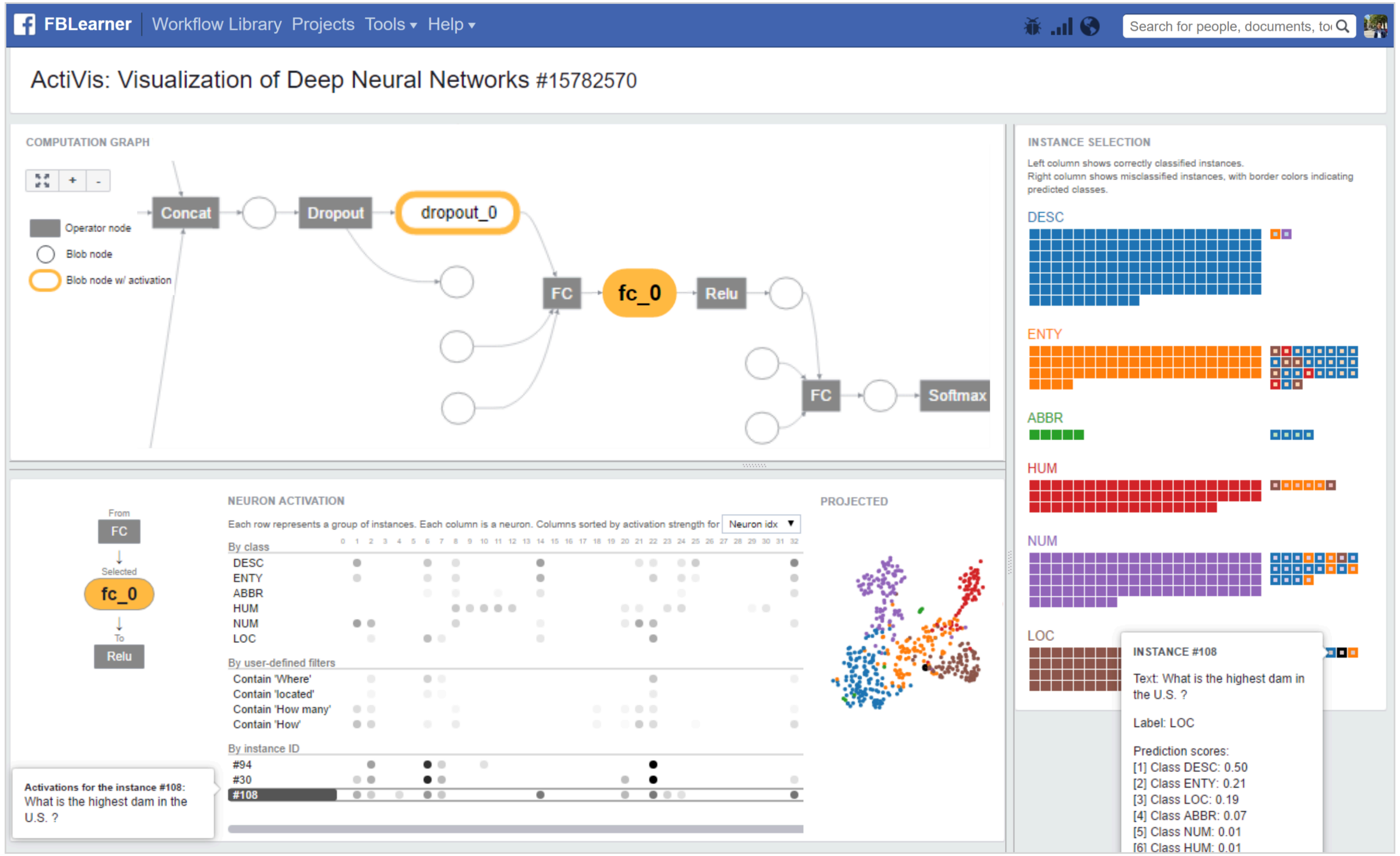
		WHY				WHO			WHAT					HOW						WHEN		WHERE
Author	Year	Interpretability & Explainability	Debugging & Improving Models	Comparing & Selecting Models	Education	Model Developers & Builders	Model Users	Non-experts	Computational Graph & Network Architecture	Learned Model Parameters	Individual Computational 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

Example

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Kahng, et al. 2018 | TVCG

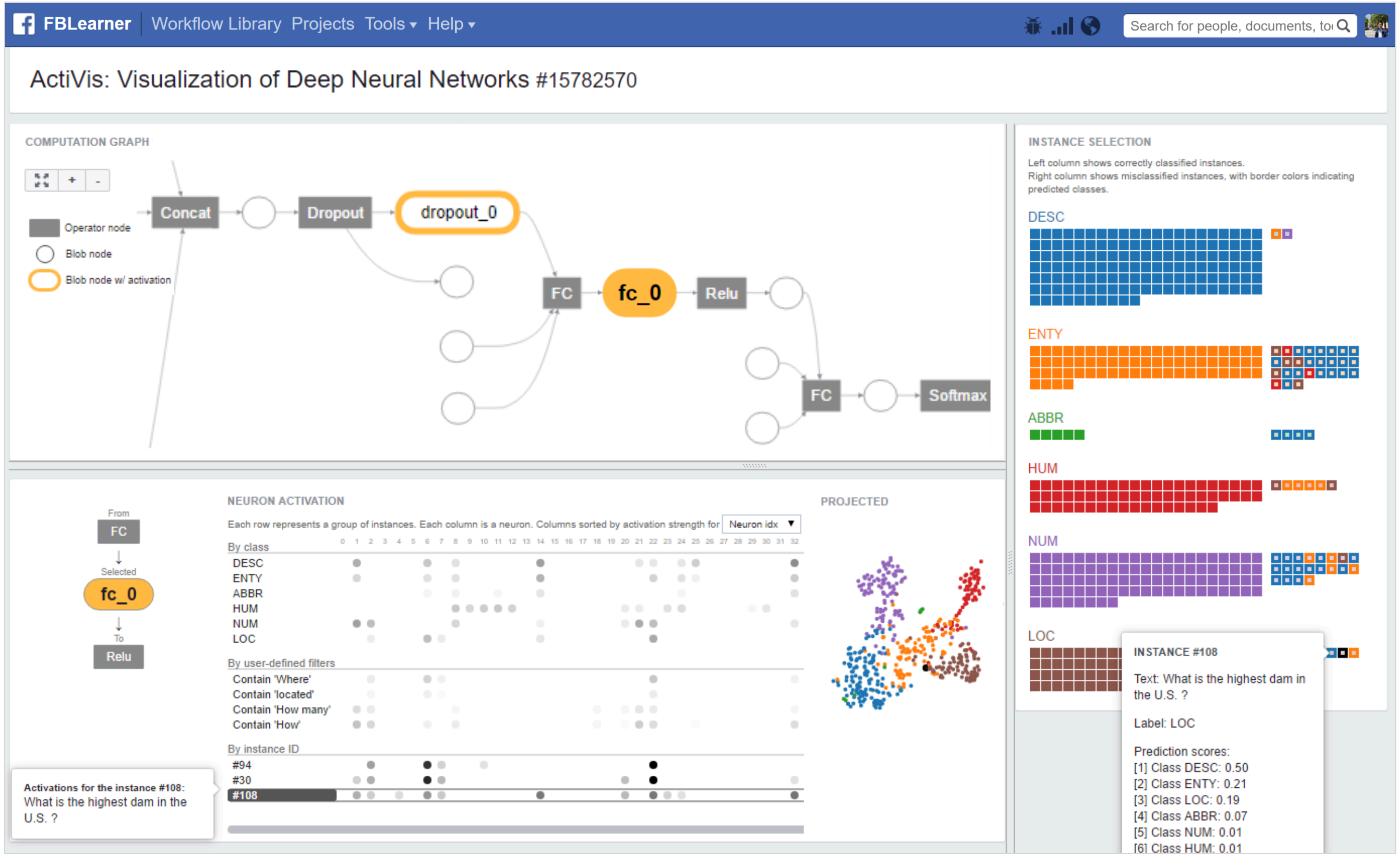


Example

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Kahng, et al. 2018



- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts

- Model Developers
- Model Users
- Non-experts

- Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information

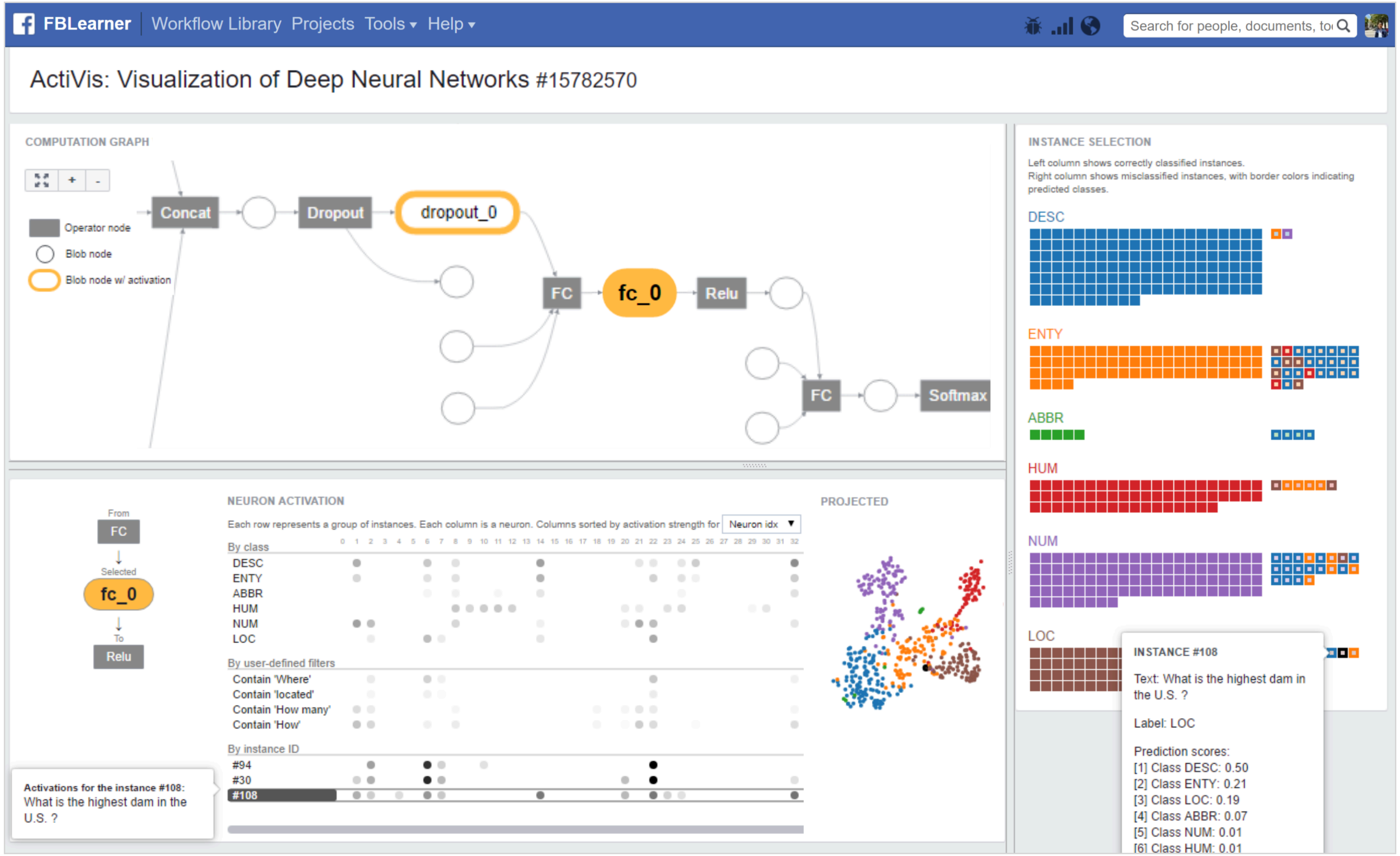
- Node-link Diagrams
- Dimensionality Reduction & Scatter Plots

Example

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau

Kahng, et al. 2018



- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts

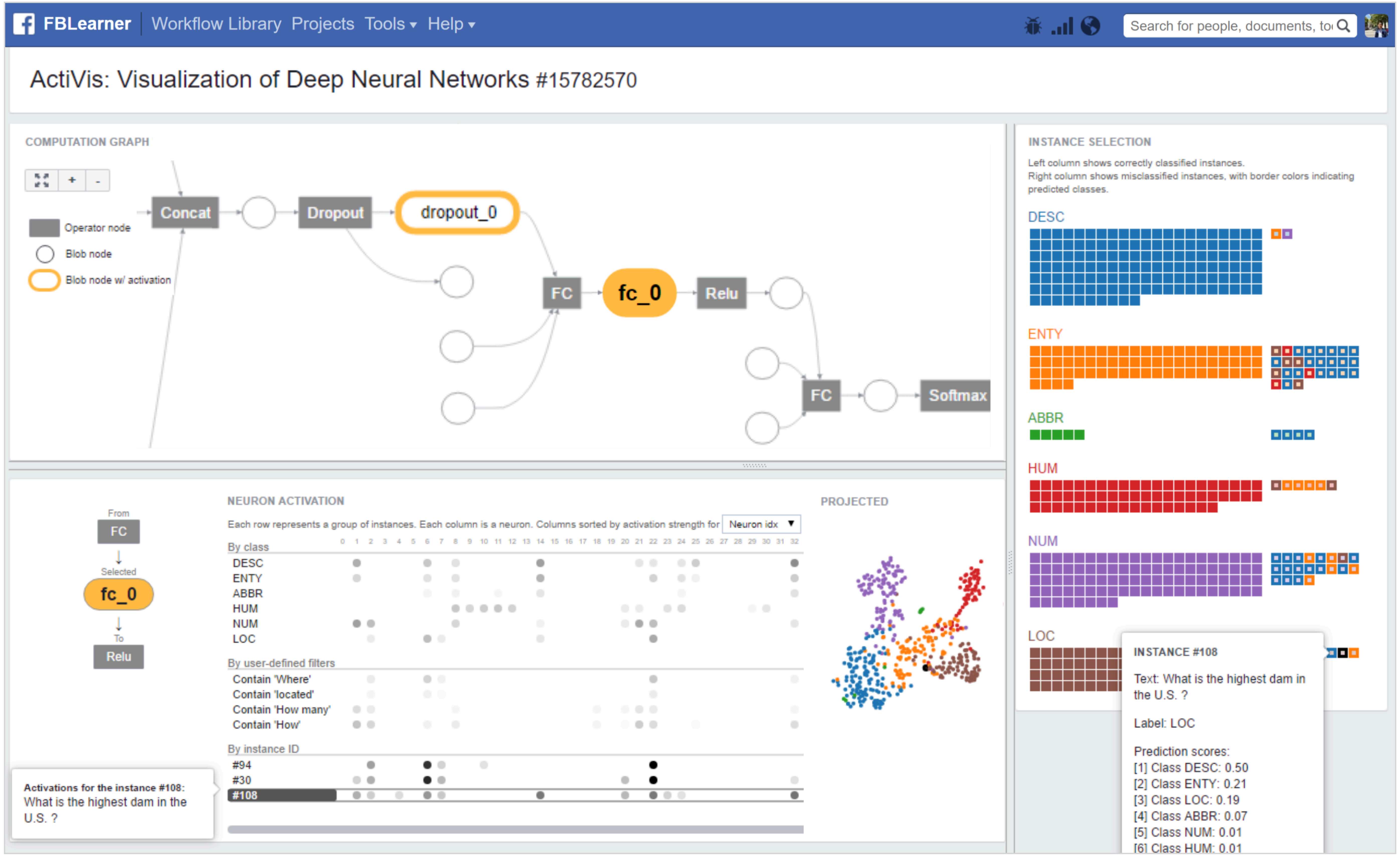
- Model Developers
- Model Users
- Non-experts

- Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information

- Node-link Diagrams
- Dimensionality Reduction & Scatter Plots

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau



Kahng, et al. 2018

WHY

- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts

- Model Developers
- Model Users
- Non-experts

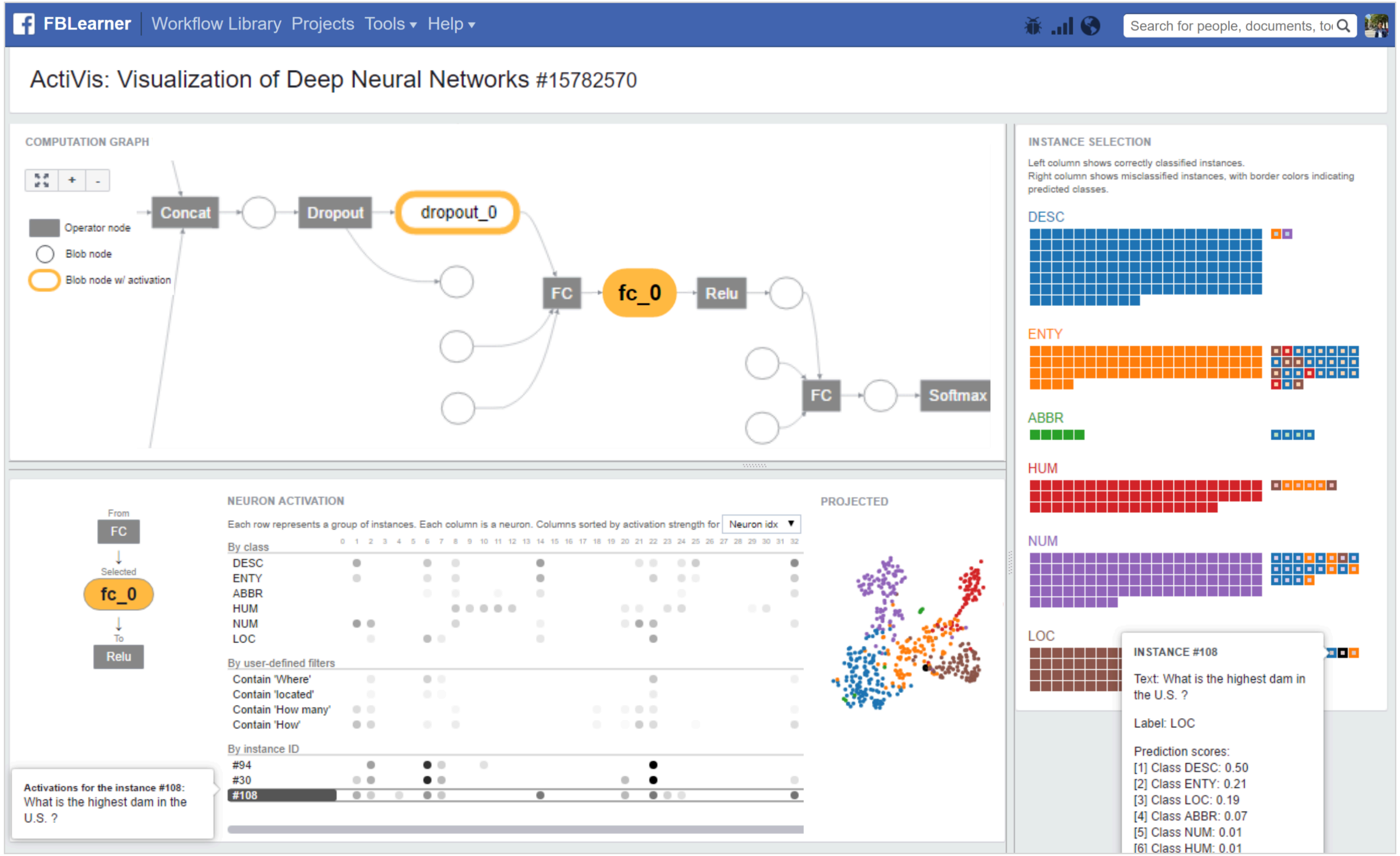
- Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information

- Node-link Diagrams
- Dimensionality Reduction & Scatter Plots

Example

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau



Interpretability & Explainability

Debugging & Improving Models

Comparing & Selecting Models

Teaching Deep Learning Concepts

WHO

- Model Developers
- Model Users
- Non-experts

Network Architecture

Learned Model Parameters

Individual Computational Units

Neurons In High-dimensional Space

Aggregated Information

Node-link Diagrams

Dimensionality Reduction & Scatter Plots

Line Charts for Temporal Metrics

Instance-based Analysis & Exploration

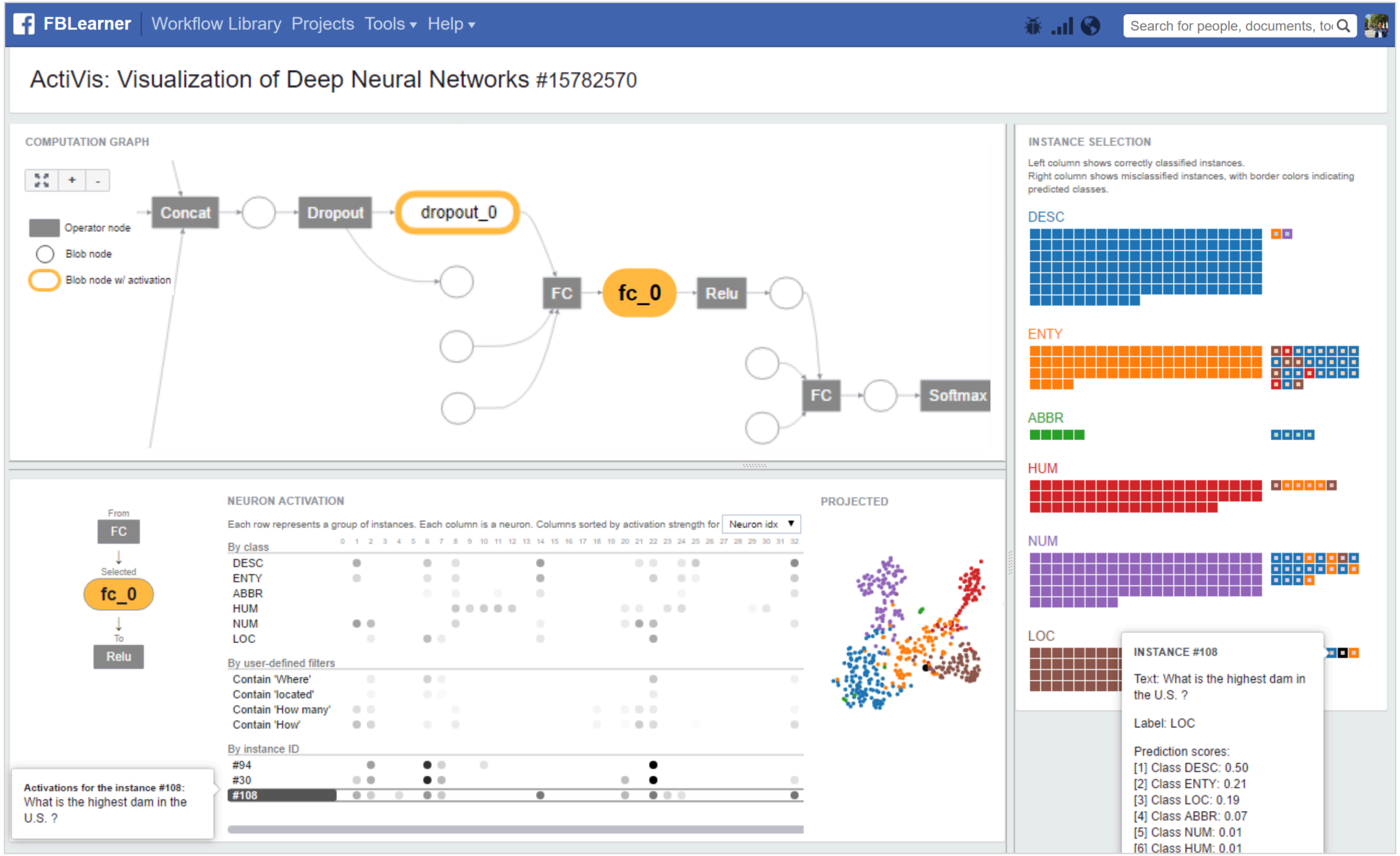
Attribution & Feature Visualization

Interactive Experimentation

Example

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau



Comparing & Selecting Models

Teaching Deep Learning Concepts

Model Developers

Model Users

WHAT

Network Architecture

Learned Model Parameters

Individual Computational Units

Neurons In High Dimensions

Aggregated Information

Node-link Diagrams

Dimensionality Reduction & Scatter Plots

Line Charts for Temporal Metrics

Instance-based Analysis & Exploration

Attribution & Feature Visualization

Interactive Experimentation

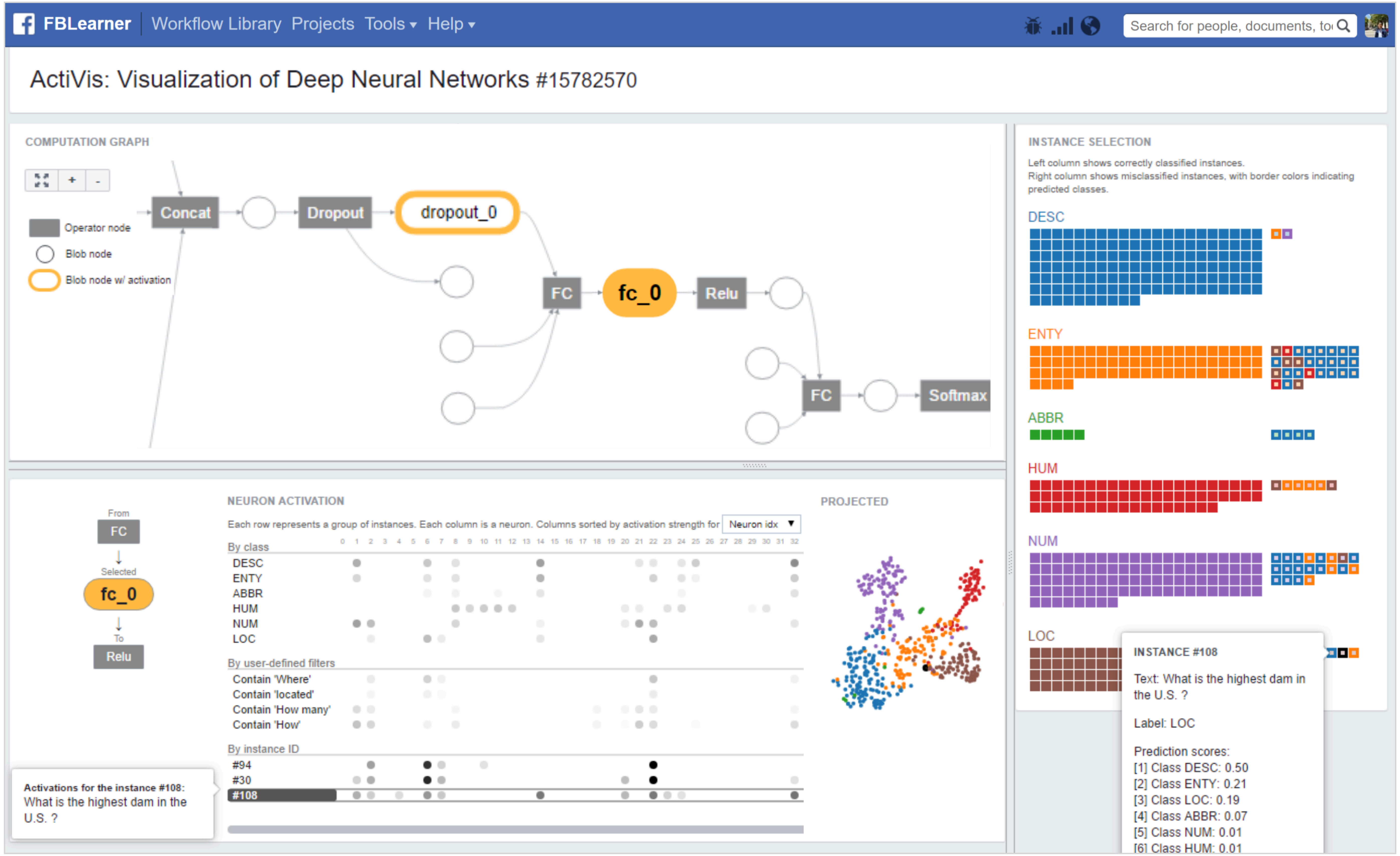
During Training

After Training

Example

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau



Network Architecture

Learned Model Parameters

Individual Computational Units

Neurons In High Dimensions

Aggregated Information

HOW

Node-link Diagrams

Dimensionality Reduction & Scatter Plots

Line Charts for Temporal Metrics

Instance-based Analysis & Exploration

Attribution & Feature Visualization

Interactive Experimentation

During Training

After Training

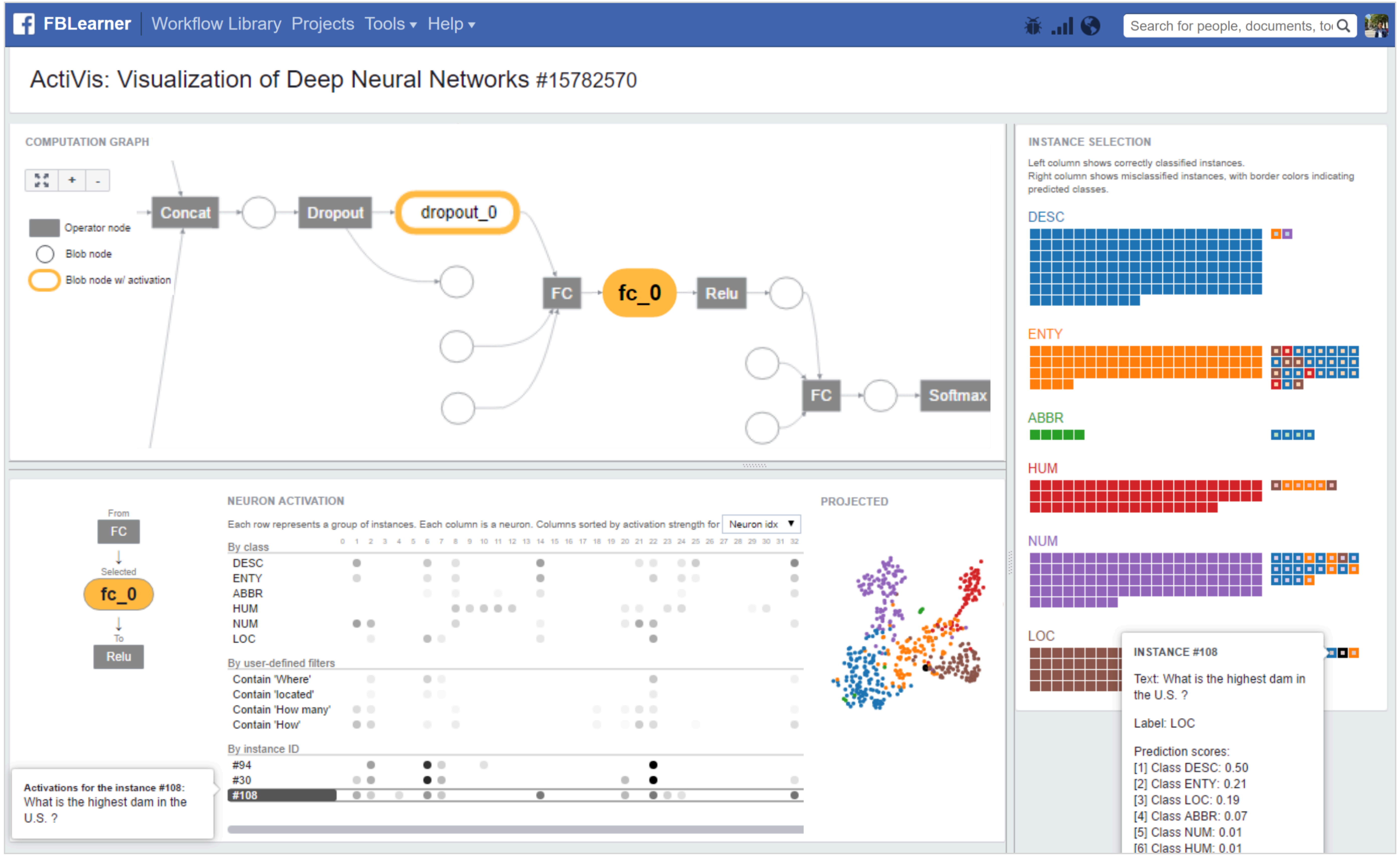
TVCG

Publication Venue

Example

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau



Dimensionality Reduction & Scatter Plots

Line Charts for Temporal Metrics

Instance-based Analysis & Exploration

Attribution & Feature Visualization

Interactive Experimentation

WHEN

During Training

After Training

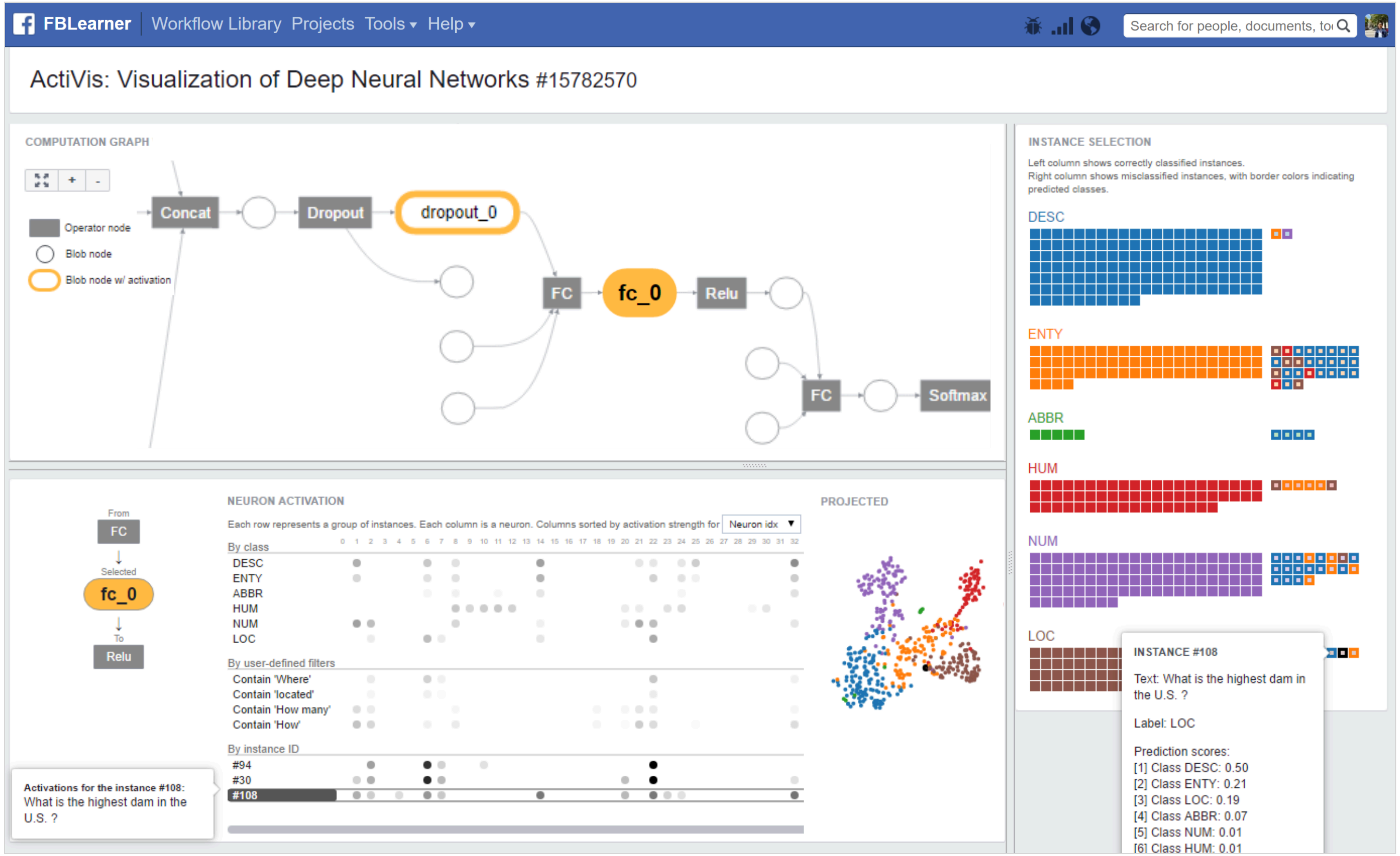
Publication Venue

TVCG

Example

ActiVis *Visual Exploration of Industry-Scale Deep Neural Network Models*

Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, Polo Chau



Instance-based Analysis & Exploration

Attribution & Feature Visualization

Interactive Experimentation

During Training

After Training

WHERE

TVCG

Publication Venue

8 Survey Highlights

8 Survey Highlights

1. Model Interpretation
2. Expert Tool Focus
3. Instance-based Analysis
4. Expanding Audience

Research trends

5. Furthering Interpretability
6. Human-in-the-loop
7. Evaluating Explanations
8. Protecting Against Attacks

Research directions

Research Trend

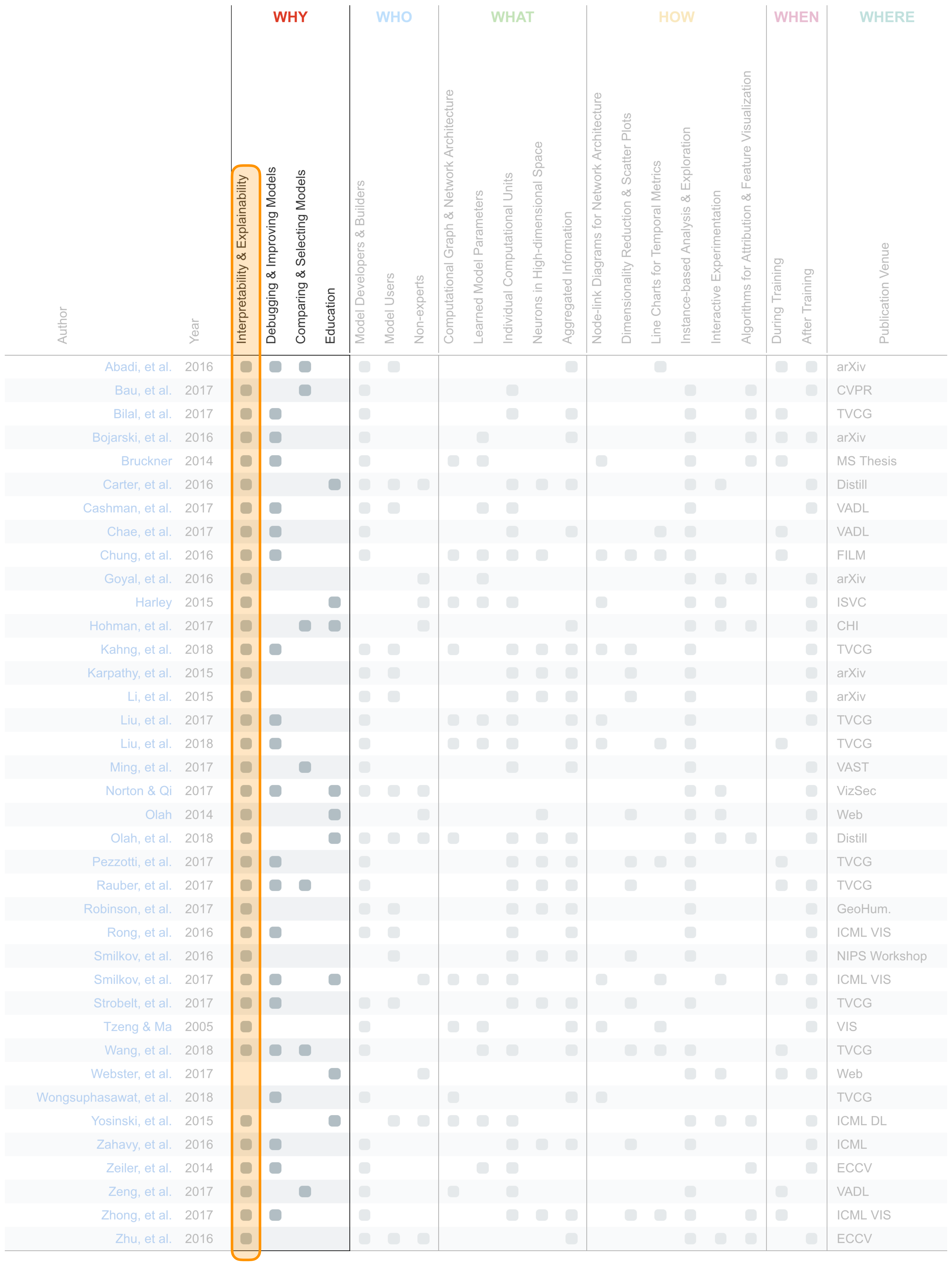
1. Model Interpretation

36/38 works support **model interpretation**

But…
formal, agreed def. remains open

Human understanding of…

internals, operations, mapping of data, or representation



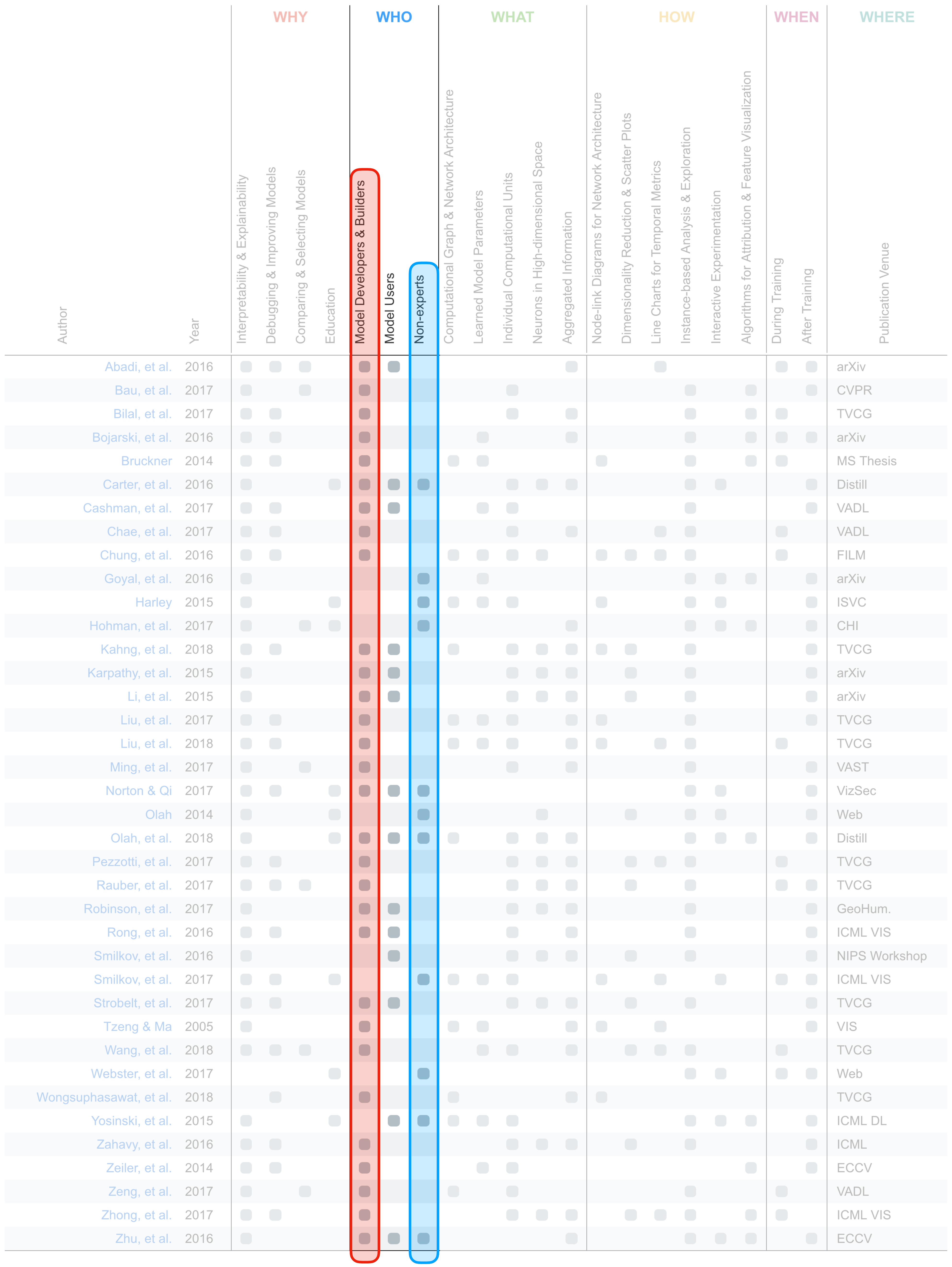
Research Trend

2. Expert Tool Focus

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

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

3x more work for developers



Research Trend

3. Instance-based Analysis

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

Neural networks
lack **global explanations**

Instance-based analysis
enables **local explanations**

		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

Research Trend

4. Expanding Audience

Note: list current as of early 2018.

VIS, HCI Conferences

Venue	
TVCG	IEEE Transactions on Visualization and Computer Graphics
VAST	IEEE Conference on Visual Analytics Science and Technology
InfoVis	IEEE Information Visualization
CHI	ACM Conference on Human Factors in Computing Systems

Research Trend

4. Expanding Audience

Note: list current as of early 2018.

ML, DL Conferences

Venue	
NeurIPS	Conference on Neural Information Processing Systems
ICML	International Conference on Machine Learning
CVPR	Conference on Computer Vision and Pattern Recognition
ICLR	International Conference on Learning Representations

Research Trend

4. Expanding Audience

Note: list current as of early 2018.

		Venue
Workshops	VADL	IEEE VIS Workshop on Visual Analytics for Deep Learning
	HCML	CHI Workshop on Human Centered Machine Learning
	IDEA	KDD Workshop on Interactive Data Exploration & Analytics
		ICML Workshop on Visualization for Deep Learning
	WHI	ICML Workshop on Human Interpretability in ML
		NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning
		NIPS Interpretable ML Symposium
	FILM	NIPS Workshop on Future of Interactive Learning Machines
		ACCV Workshop on Interpretation and Visualization of Deep Neural Nets
		ICANN Workshop on Machine Learning and Interpretability
Online	Distill	Distill: Journal for Supporting Clarity in Machine Learning
	arXiv	arXiv.org e-Print Archive

Research Trend

4. Expanding Audience

Note: list current as of early 2018.

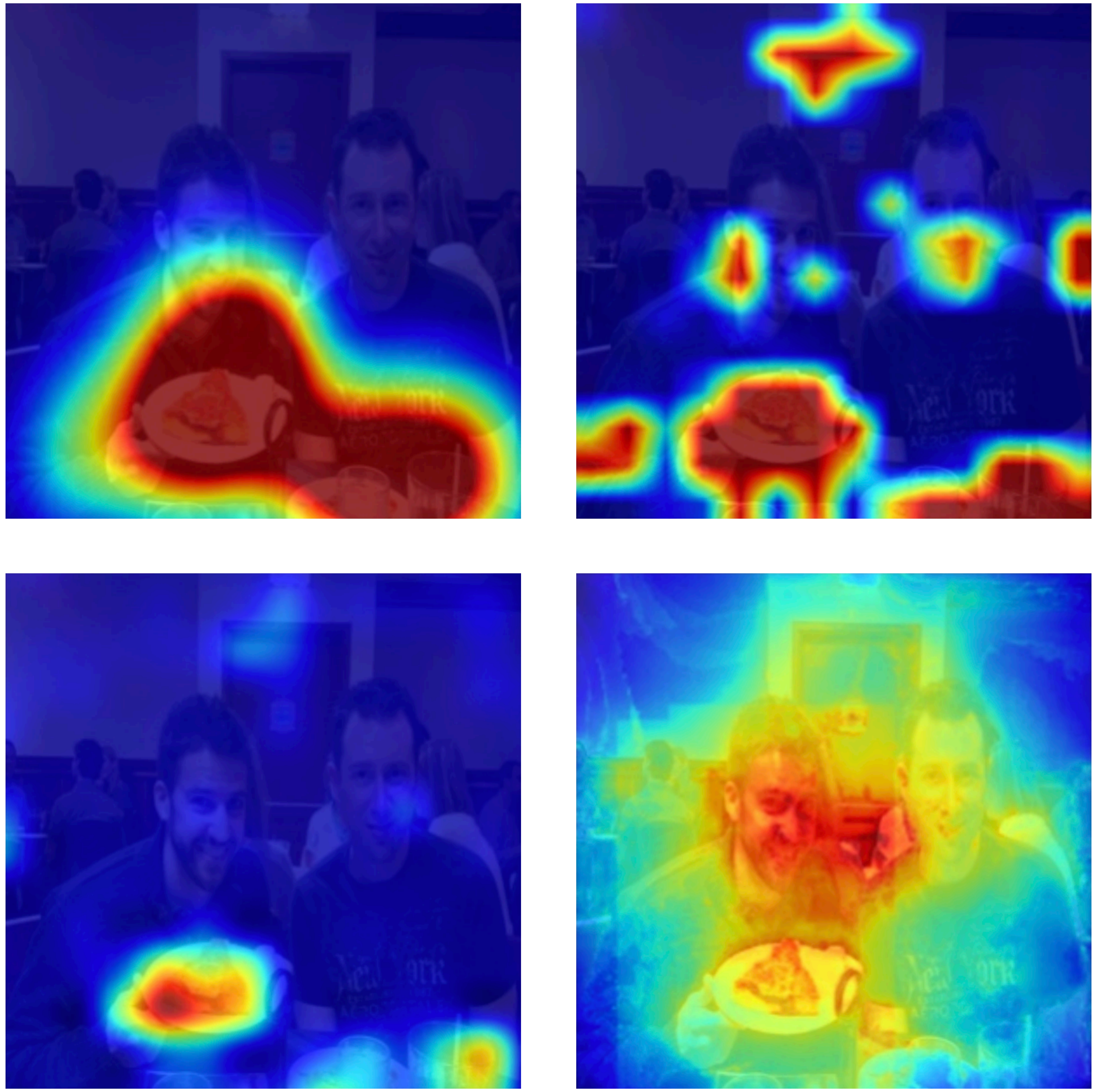
Top venues highly value **open source**

Research Direction

5. Furthering Interpretability

Q: *What are they doing?*

A: *eating*



Attention

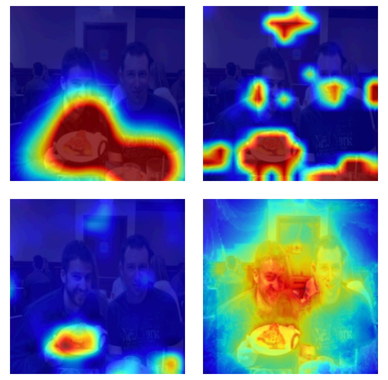
Das, Agrawal, et al. 2016

Research Direction

5. Furthering Interpretability

Attention

Das, Agrawal, et al. 2016



Saliency

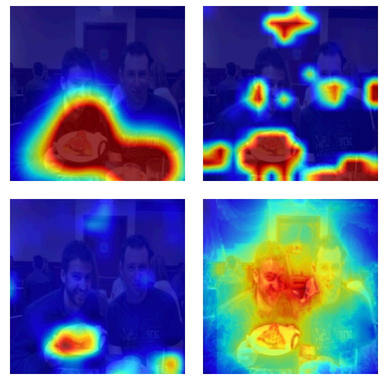
Smilkov, et al. 2017

Research Direction

5. Furthering Interpretability

Attention

Das, Agrawal, et al. 2016



Saliency

Smilkov, et al. 2017

Feature visualization

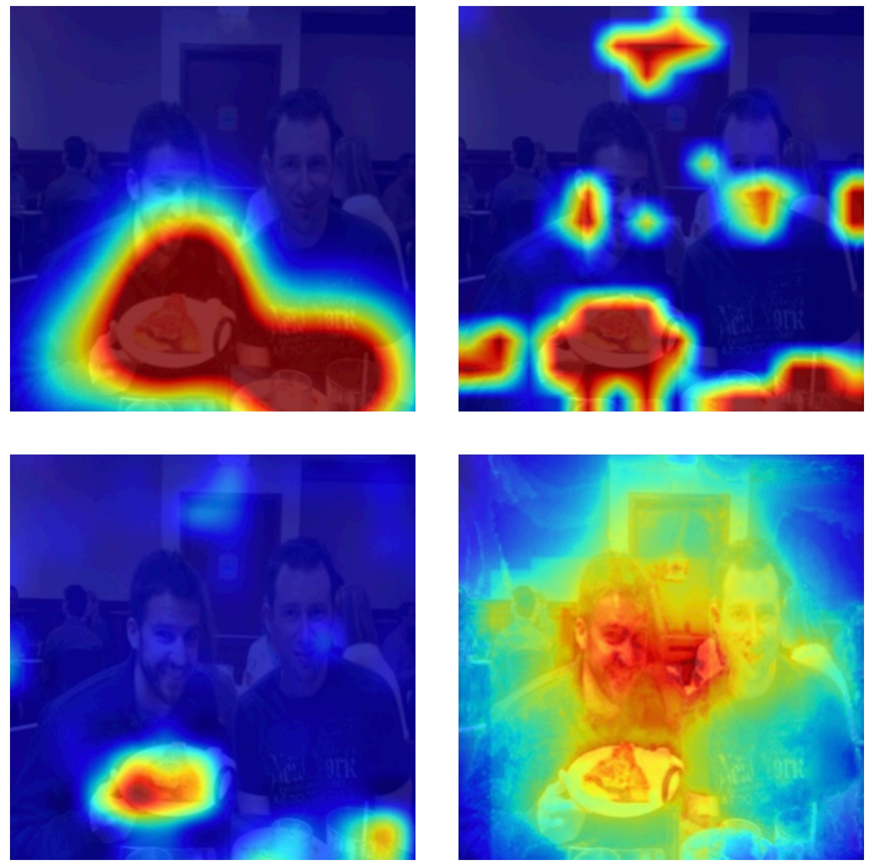
Olah, et al. 2017

Research Direction

5. Furthering Interpretability

Attention

Das, Agrawal, et al. 2016



Saliency

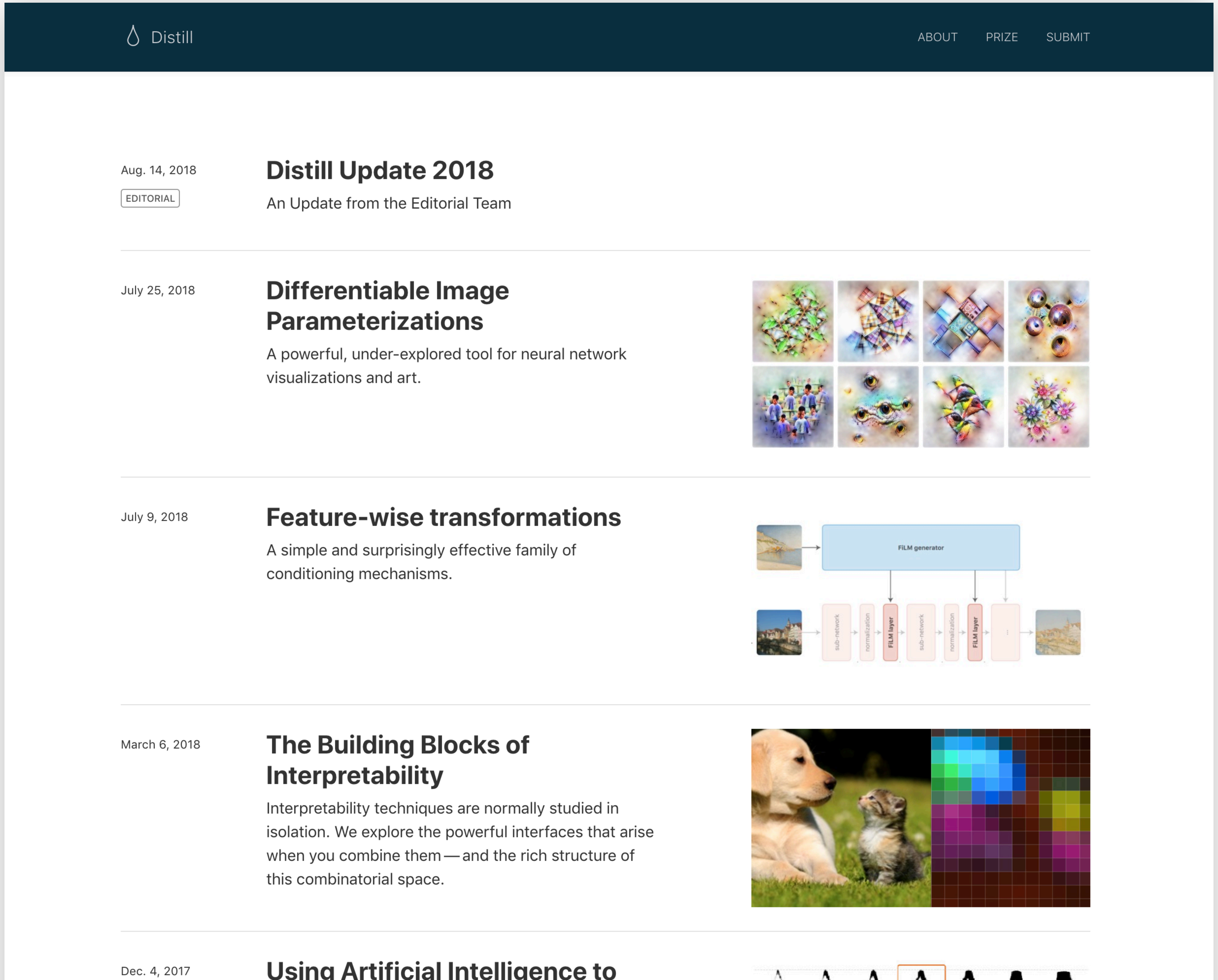
Smilkov, et al. 2017

Feature visualization

Olah, et al. 2017

Research Direction

5. Furthering Interpretability



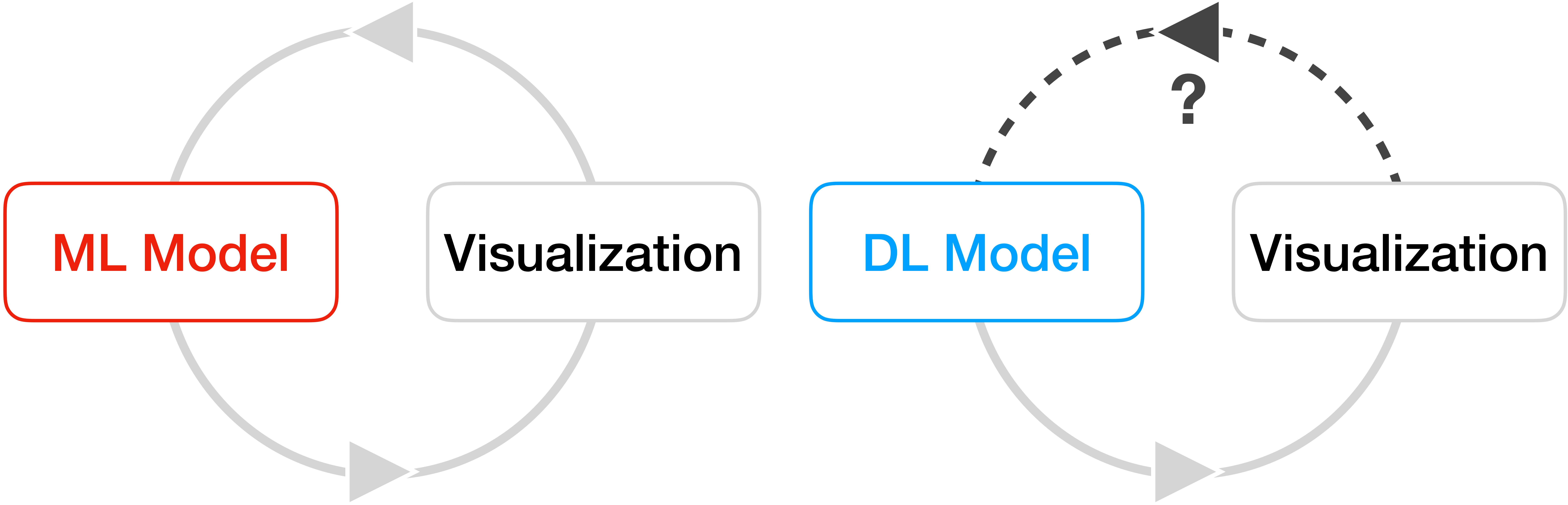
Distill

Journal for Supporting Clarity in Machine Learning

6. Human-in-the-loop

Interactive **Machine Learning**

Interactive **Deep Learning**



Research Direction

7. Evaluating Explanations

Towards A Rigorous Science of Interpretable Machine Learning

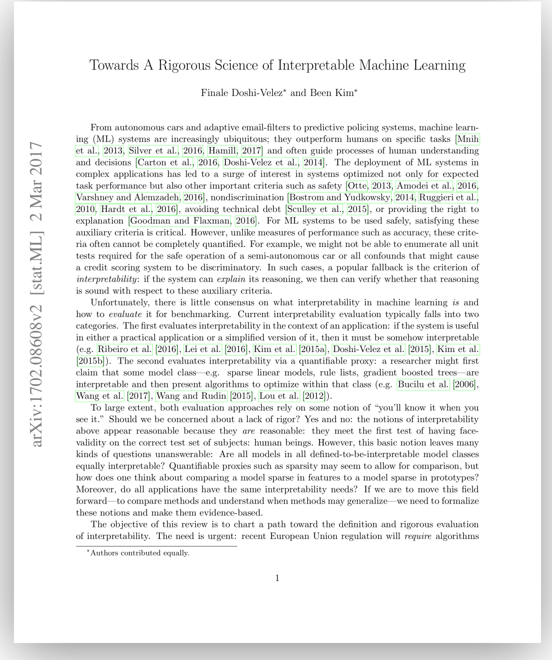
Finale Doshi-Velez* and Been Kim*

From autonomous cars and adaptive email-filters to predictive policing systems, machine learning (ML) systems are increasingly ubiquitous; they outperform humans on specific tasks [Mnih et al., 2013, Silver et al., 2016, Hamill, 2017] and often guide processes of human understanding and decisions [Carton et al., 2016, Doshi-Velez et al., 2014]. The deployment of ML systems in complex applications has led to a surge of interest in systems optimized not only for expected task performance but also other important criteria such as safety [Otte, 2013, Amodei et al., 2016, Varshney and Alemzadeh, 2016], nondiscrimination [Bostrom and Yudkowsky, 2014, Ruggieri et al., 2010, Hardt et al., 2016], avoiding technical debt [Sculley et al., 2015], or providing the right to explanation [Goodman and Flaxman, 2016]. For ML systems to be used safely, satisfying these auxiliary criteria is critical. However, unlike measures of performance such as accuracy, these crite-

Research Direction

7. Evaluating Explanations

Doshi-Velez, Kim. 2017


Towards A Rigorous Science of Interpretable Machine Learning

Finale Doshi-Velez* and Been Kim*

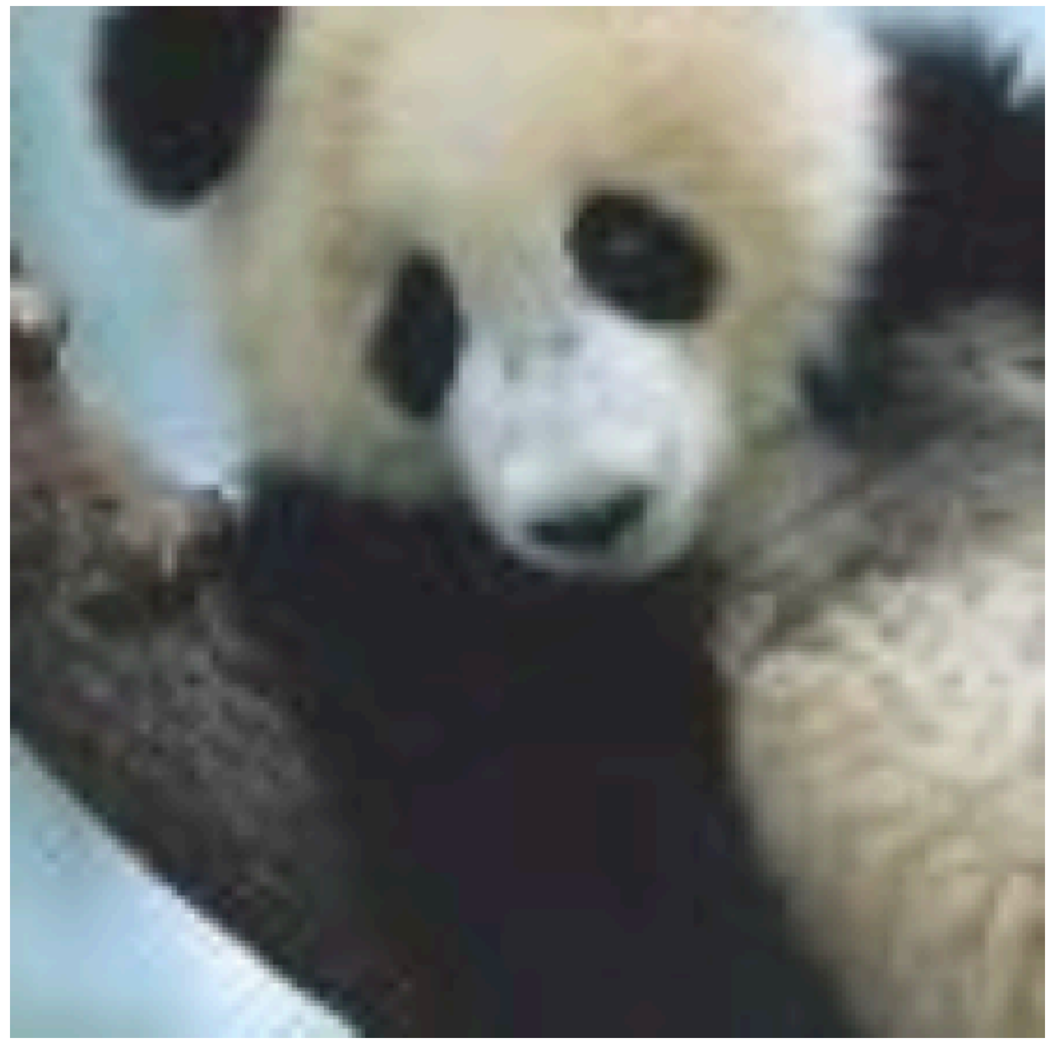
More specific and costly

Evaluation	Humans	Tasks
Application-grounded	Yes	Real
Human-grounded	Yes	Simple
Functionally-grounded	No	Proxy

Research Direction

8. Protecting Against Attacks

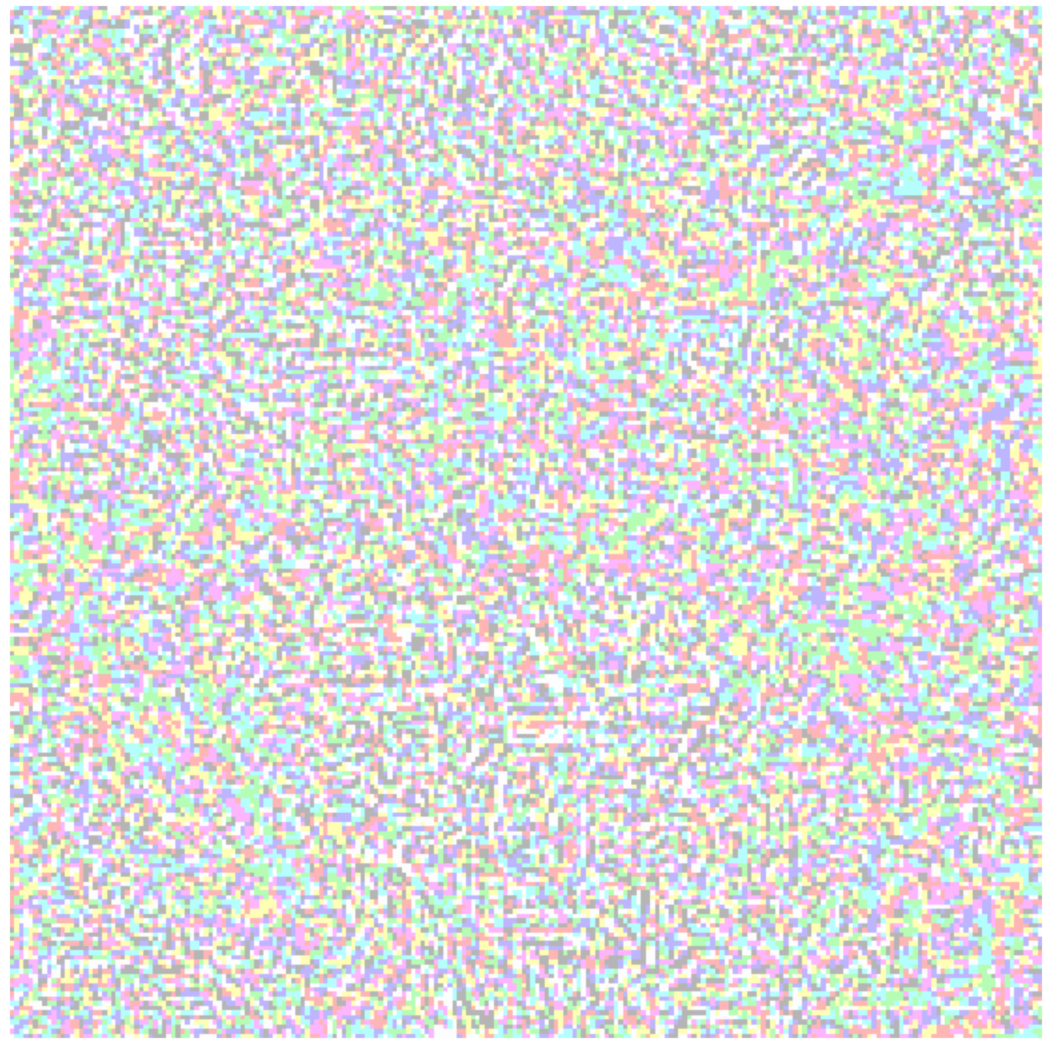
Benign



"panda" ✓

+

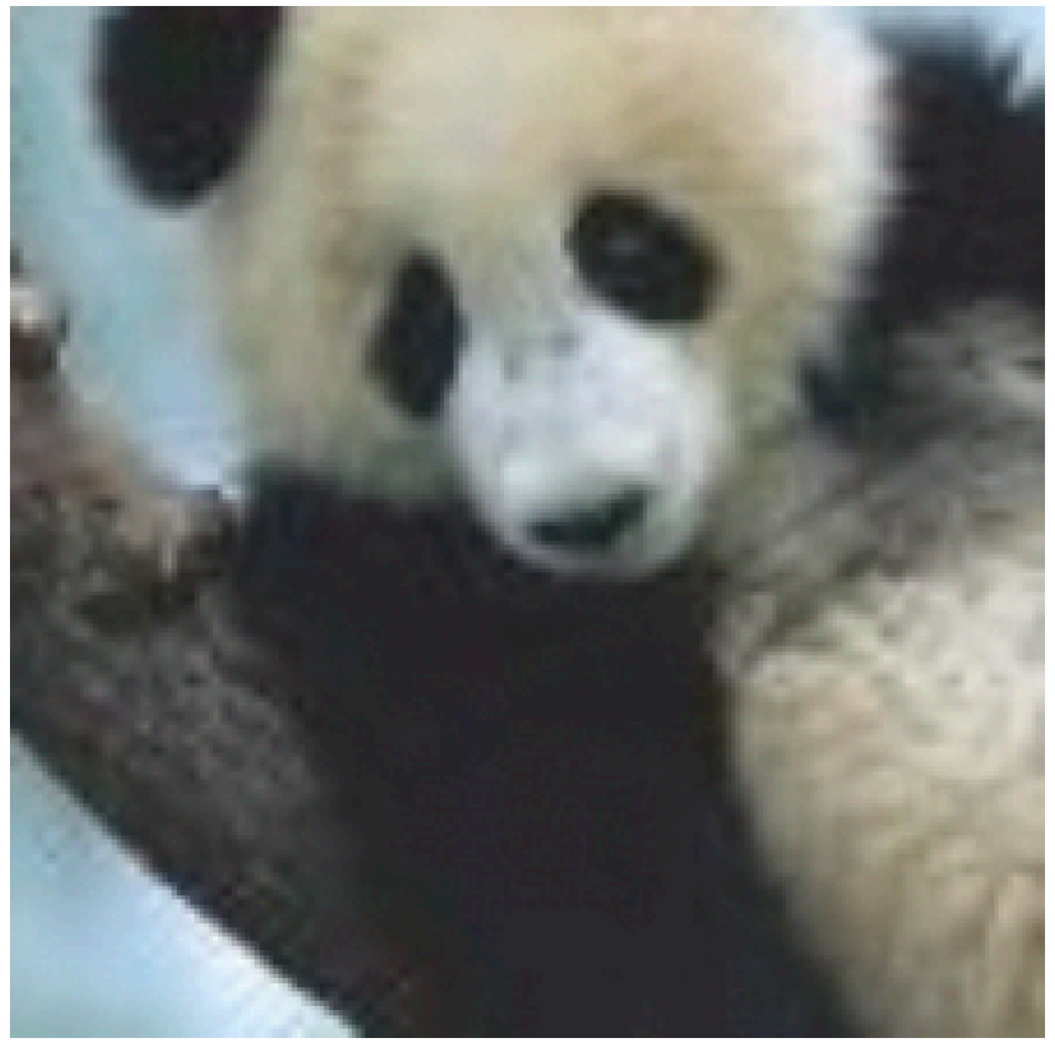
Perturbation



attack

=

Attacked



"gibbon" ✗

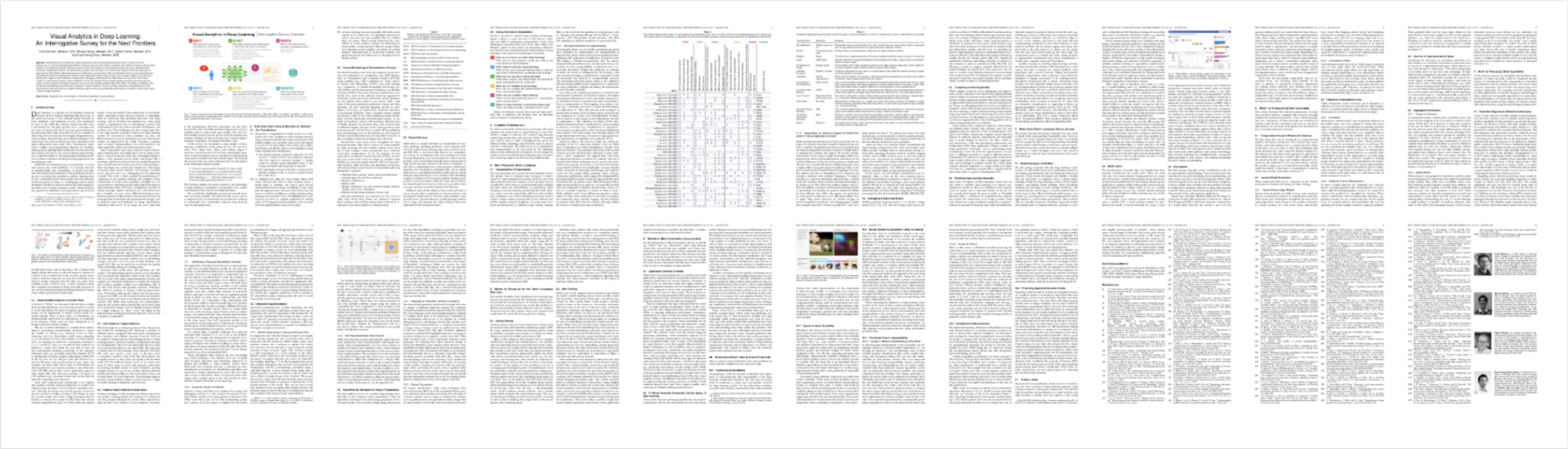
Visual Analytics in Deep Learning

An Interrogative Survey for the Next Frontiers

Fred Hohman, Minsuk Kahng, Robert Pienta, Duen Horng Chau

Deep learning has recently seen rapid development and significant attention due to its state-of-the-art performance on previously-thought hard problems. However, because of the innate complexity and nonlinear structure of deep neural networks, the underlying decision making processes for why these models are achieving such high performance are challenging and sometimes mystifying to interpret.

As deep learning spreads across domains, it is of paramount importance that we equip users of deep learning with tools for understanding when a model works correctly, when it fails, and ultimately how to improve its performance. Standardized toolkits for building neural networks have helped democratize deep learning; visual analytics systems have now been developed to support model explanation, interpretation, debugging, and improvement.



Read the paper.

We present a survey of the role of visual analytics in deep

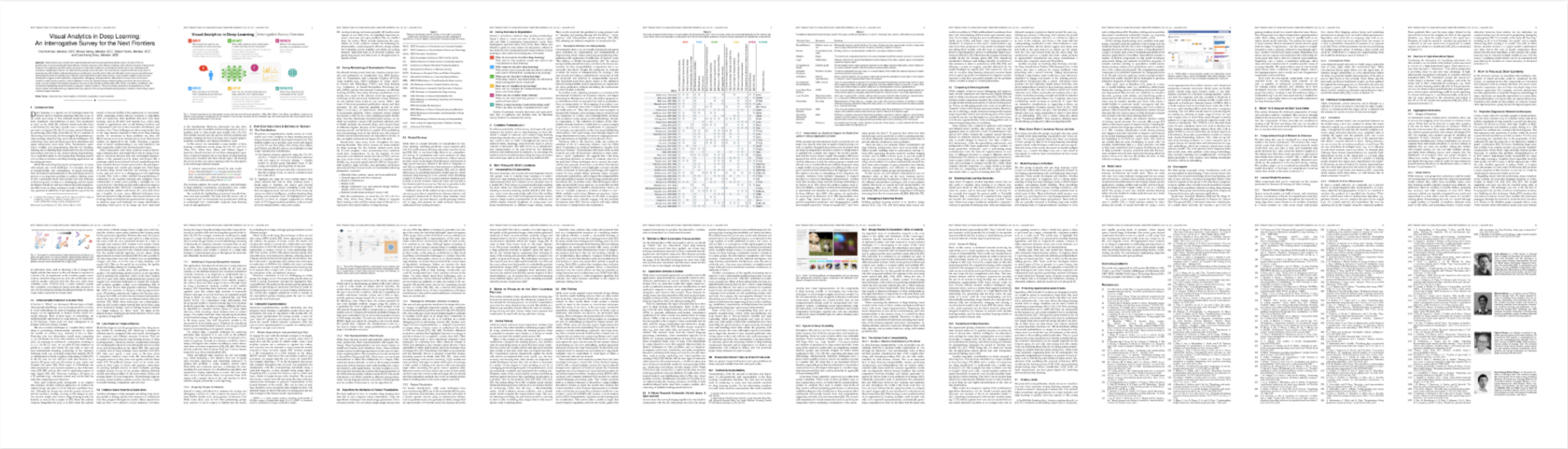
Visual Analytics in Deep Learning

An Interrogative Survey for the Next Frontiers

Fred Hohman, Minsuk Kahng, Robert Pienta, Duen Horng Chau

Deep learning has recently seen rapid development and significant attention due to its state-of-the-art performance on previously-thought hard problems. However, because of the innate complexity and nonlinear structure of deep neural networks, the underlying decision making processes for why these models are achieving such high performance are challenging and sometimes mystifying to interpret.

As deep learning spreads across domains, it is of paramount importance that we equip users of deep learning with tools for understanding when a model works correctly, when it fails, and ultimately how to improve its performance. Standardized toolkits for building neural networks have helped democratize deep learning; visual analytics systems have now been developed to support model explanation, interpretation, debugging, and improvement.



Read the paper.

We present a survey of the role of visual analytics in deep

bit.ly/
va-dl-survey

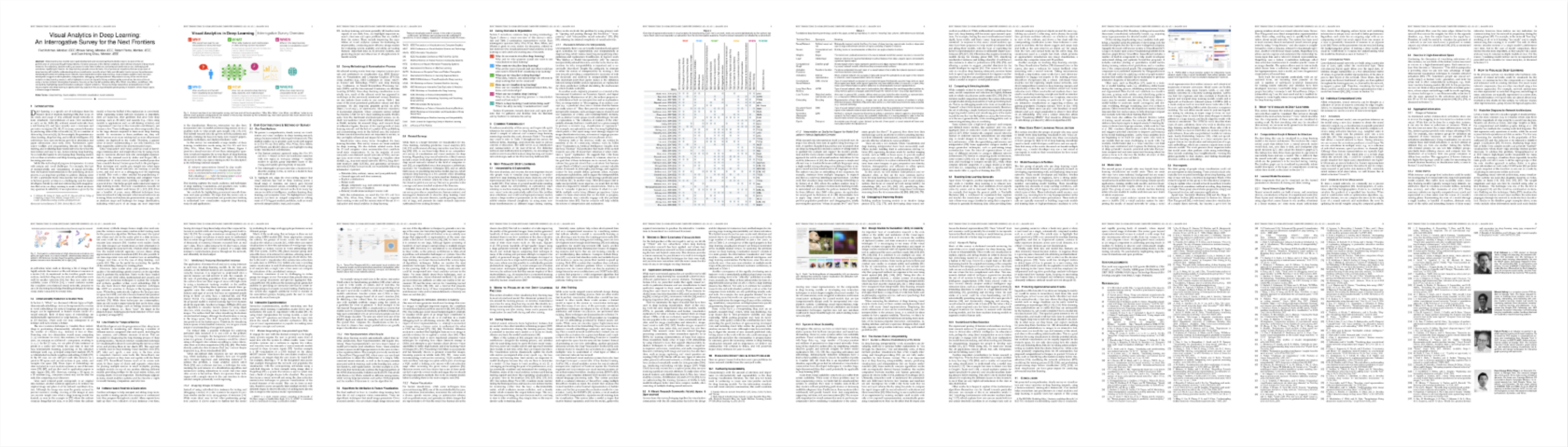
Deep Learning

An Interrogative Survey for the Next Frontiers

Fred Hohman, Minsuk Kahng, Robert Pienta, Duen Horng Chau

Deep learning has recently seen rapid development and significant attention due to its state-of-the-art performance on previously-thought hard problems. However, because of the innate complexity and nonlinear structure of deep neural networks, the underlying decision making processes for why these models are achieving such high performance are challenging and sometimes mystifying to interpret.

As deep learning spreads across domains, it is of paramount importance that we equip users of deep learning with tools for understanding when a model works correctly, when it fails, and ultimately how to improve its performance. Standardized toolkits for building neural networks have helped democratize deep learning; visual analytics systems have now been developed to support model explanation, interpretation, debugging, and improvement.



Read the paper.

We present a survey of the role of visual analytics in deep learning research, noting its short yet impactful history and summarize the state-of-the-art using a human-centered interrogative framework, focusing on the Five W's and How (**WHY**, **WHO**, **WHAT**, **HOW**, **WHEN**, and **WHERE**), to thoroughly summarize deep learning visual analytics research. We conclude by highlighting research directions and open research problems.

bit.ly/
va-dl-survey

Read the paper

		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

bit.ly/
va-dl-survey

Paper table,
with links

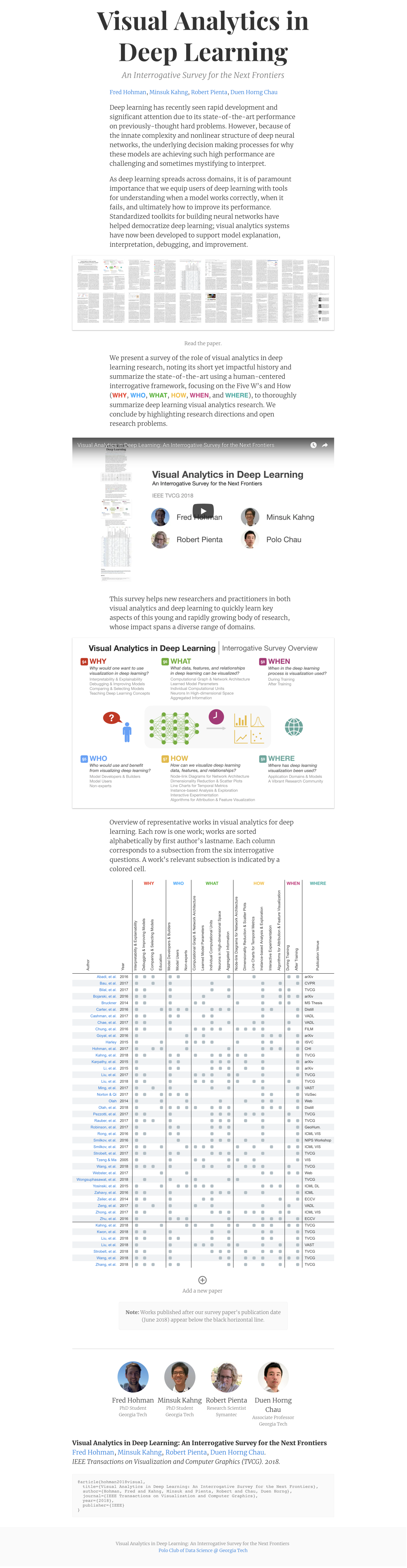
Abadi, et al.	2016	●	●	●		●	●						●			●				●	●	arXiv
Bau, et al.	2017	●		●		●					●						●		●		●	CVPR
Bilal, et al.	2017	●	●			●					●		●				●		●	●		TVCG
Bojarski, et al.	2016	●	●			●				●			●				●		●	●	●	arXiv
Bruckner	2014	●	●			●			●	●				●			●		●	●		MS Thesis
Carter, et al.	2016	●			●	●	●	●			●	●	●				●	●			●	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
Kahng, et al.	2018	●			●			●	●		●		●	●		●	●	●		●	●	TVCG
Kwon, et al.	2018	●	●			●					●		●				●	●			●	TVCG
Liu, et al.	2018	●	●			●	●				●		●				●	●			●	TVCG
Liu, et al.	2018	●				●			●	●	●		●	●			●		●		●	VAST
Strobelt, et al.	2018	●	●			●					●	●	●		●		●	●	●		●	TVCG
Wang, et al.	2018	●	●			●						●	●		●	●	●		●	●	●	TVCG
Zhang, et al.	2018	●	●	●		●	●						●		●		●		●		●	TVCG

Add a new paper

Note: Works published after our survey paper's publication date (June 2018) appear below the black horizontal line.

bit.ly/
va-dl-survey

VIS 2018 Papers



Visual Analytics in Deep Learning

An Interrogative Survey for the Next Frontiers

IEEE TVCG 2018

Thanks!

bit.ly/
va-dl-survey

Fred Hohman
@fredhohman

Minsuk Kahng

Robert Pienta

Polo Chau

Georgia Tech

NASA

Google AI