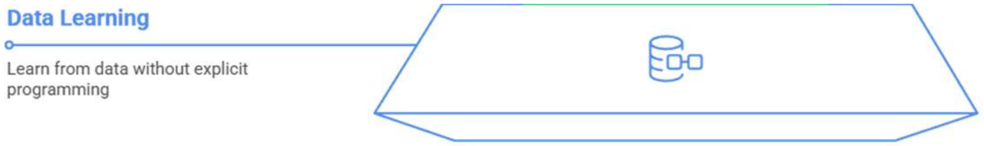
XAI

Visual Analytics for explainable Deep Learning

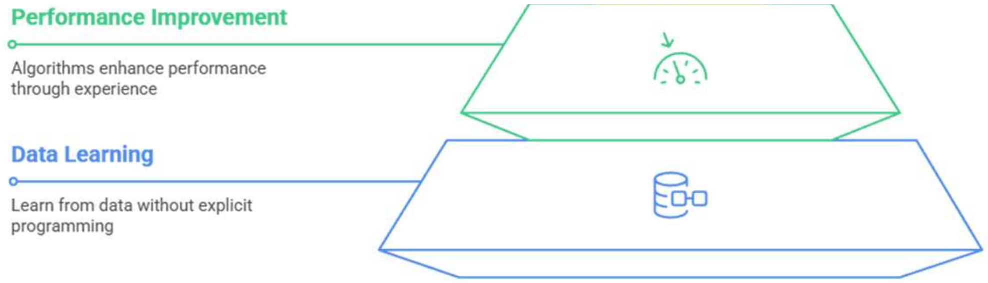
Machine Learning

- Learn from data without explicit programming



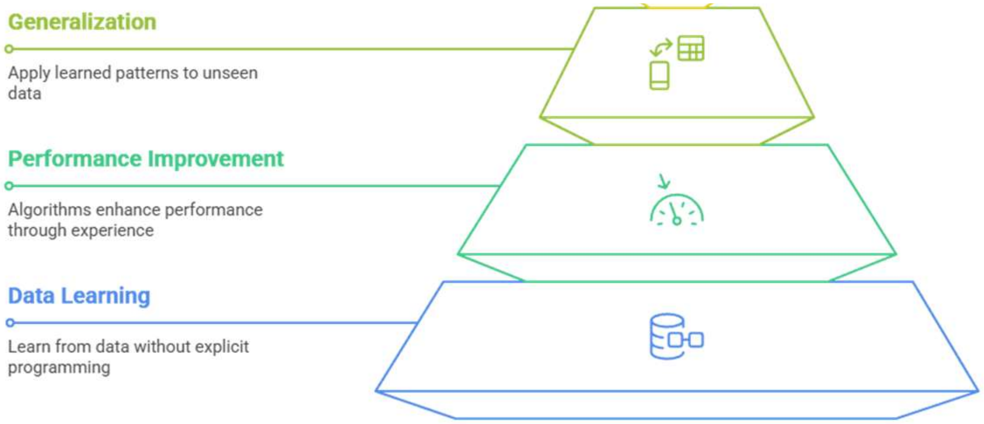
Machine Learning

- Learn from data without explicit programming
- ML algorithms improve performance through experience



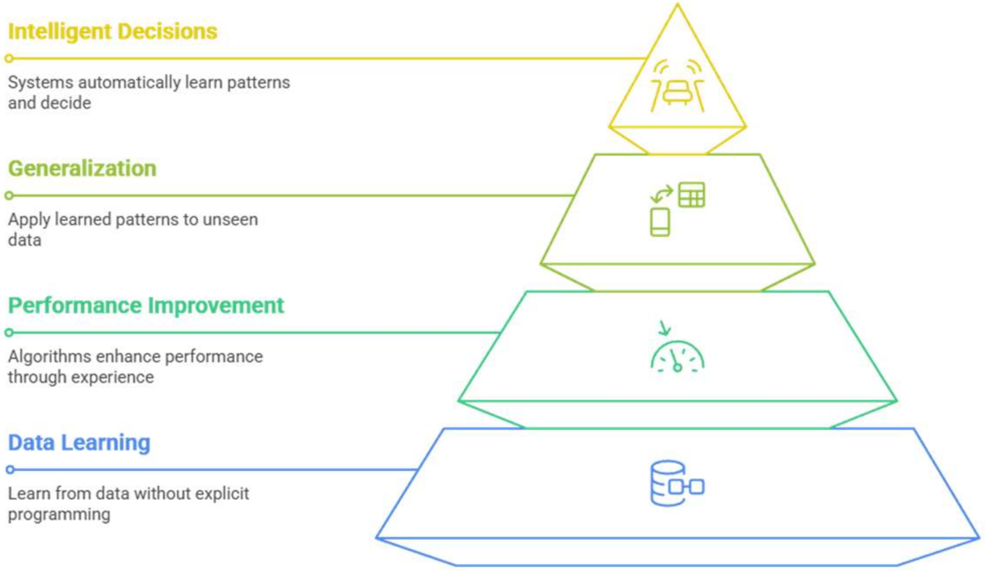
Machine Learning

- Learn from data without explicit programming
- ML algorithms improve performance through experience
  - Learn from data and generalize to unseen data



Machine Learning

- Learn from data without explicit programming
- ML algorithms improve performance through experience
  - Learn from data and generalize to unseen data
- Key goal: Create systems that can automatically learn patterns and make intelligent decisions



Types of Machine Learning

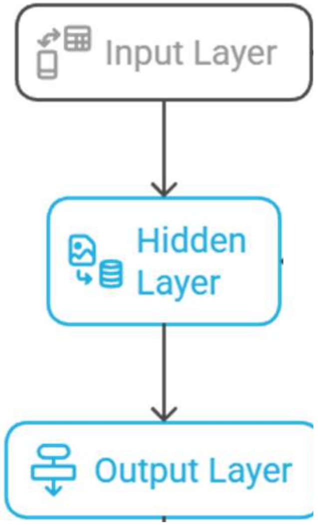
- Supervised Learning: Learns from labelled training data
  - Examples: Classification, Regression
  - Used in spam detection, price prediction
- Unsupervised Learning: Finds patterns in unlabelled data
  - Examples: Clustering, Dimensionality Reduction
  - Used in customer segmentation, anomaly detection
- Reinforcement Learning: Learns through trial and error
  - Used in game AI, robotics, autonomous systems

Deep Learning

- Advanced machine learning technique using multi-layered neural networks
- Mimics human brain's neural structure to process complex data
- Enables automatic feature extraction and learning from raw data

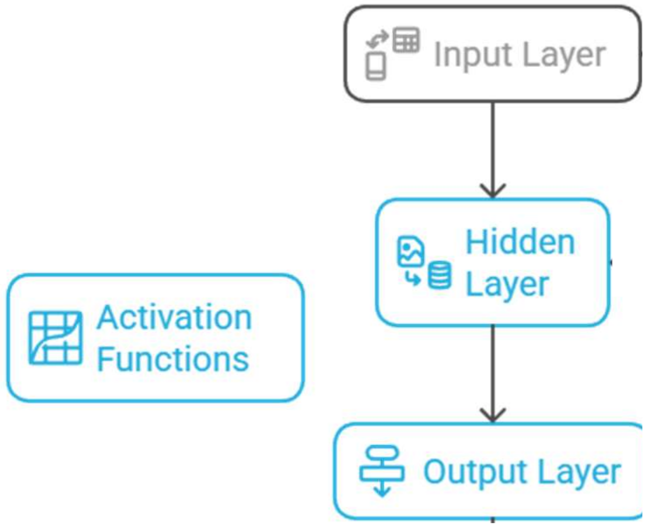
Neural Network Architecture

- Layers: Input, Hidden, Output layers



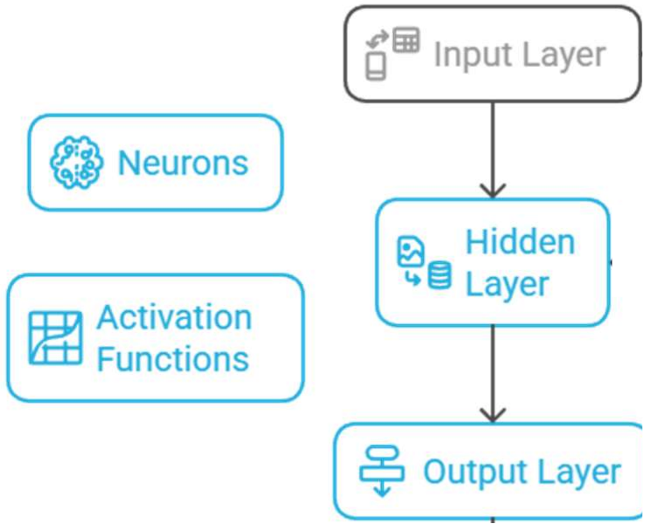
Neural Network Architecture

- Layers: Input, Hidden, Output layers
- Neurons: Process and transmit information



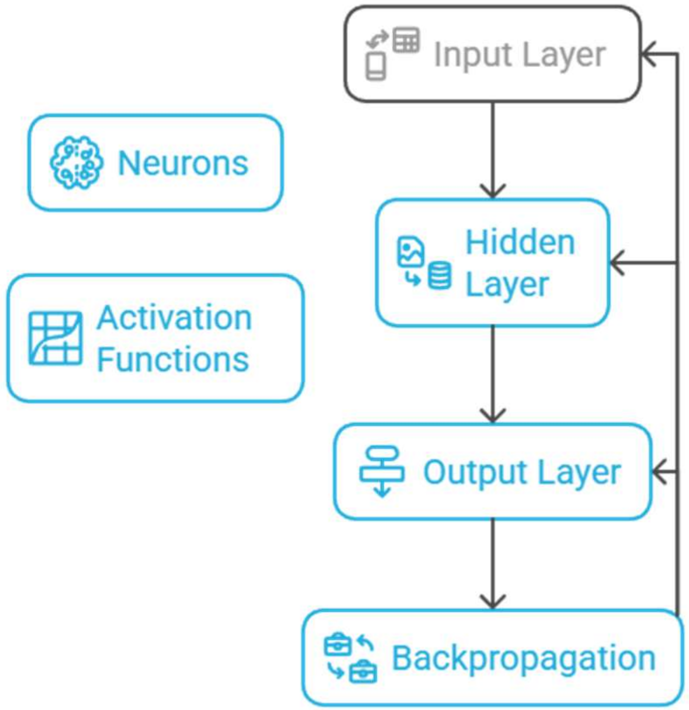
Neural Network Architecture

- Layers: Input, Hidden, Output layers
- Neurons: Process and transmit information
- Activation Functions: Introduce non-linearity



Neural Network Architecture

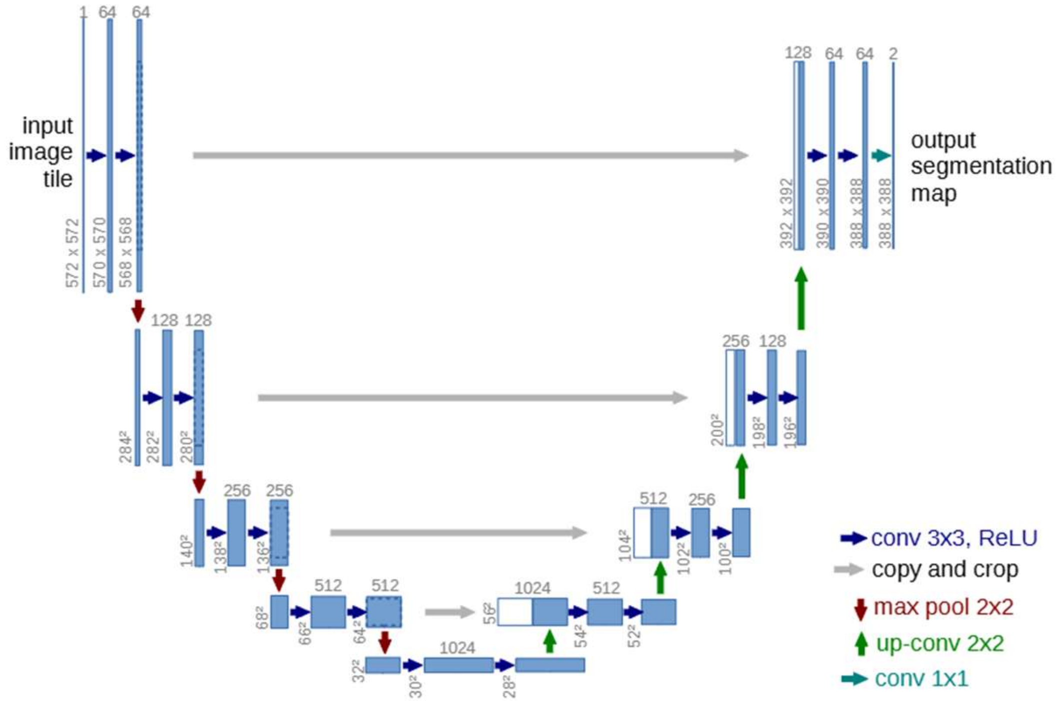
- Layers: Input, Hidden, Output layers
- Neurons: Process and transmit information
- Activation Functions: Introduce non-linearity
- Backpropagation: Learns by adjusting network weights



Various Deep Learning Techniques

Various Deep Learning Techniques

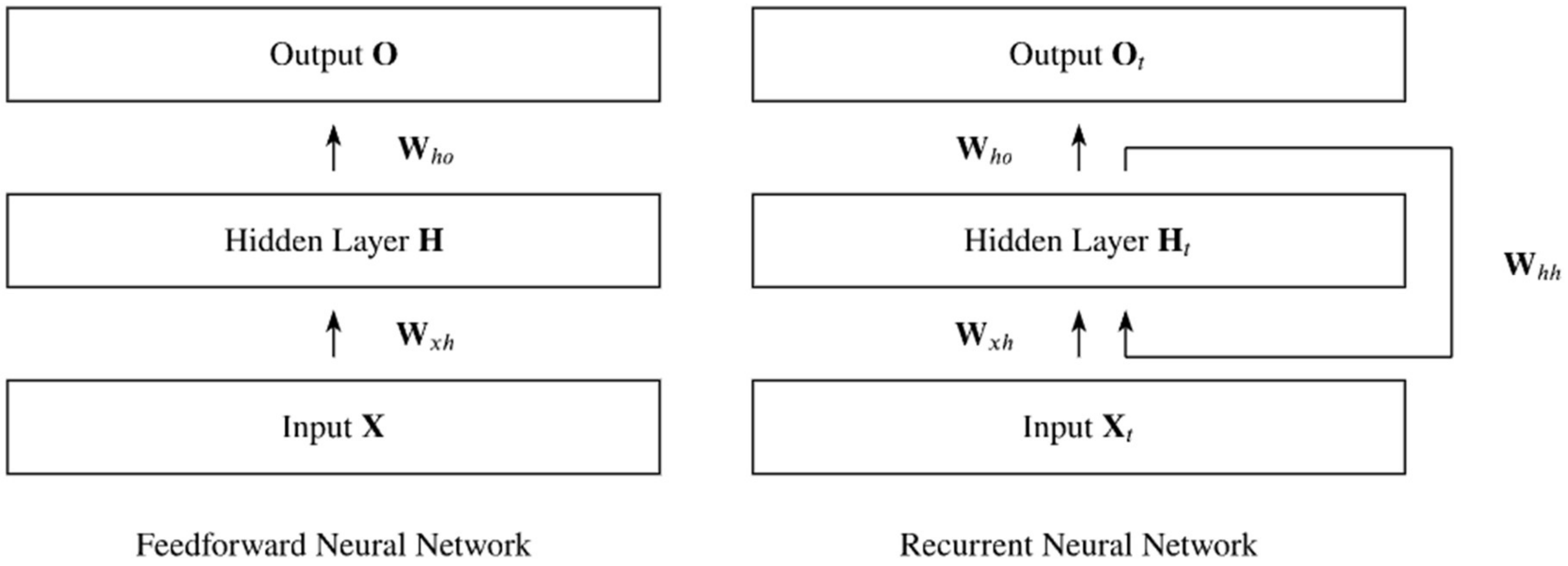
- Convolutional Neural Networks (CNNs)
  - Image and video processing
  - Pattern recognition



U-Net: Convolutional Networks for Biomedical Image Segmentation, https://doi.org/10.48550/arXiv.1505.04597

Various Deep Learning Techniques

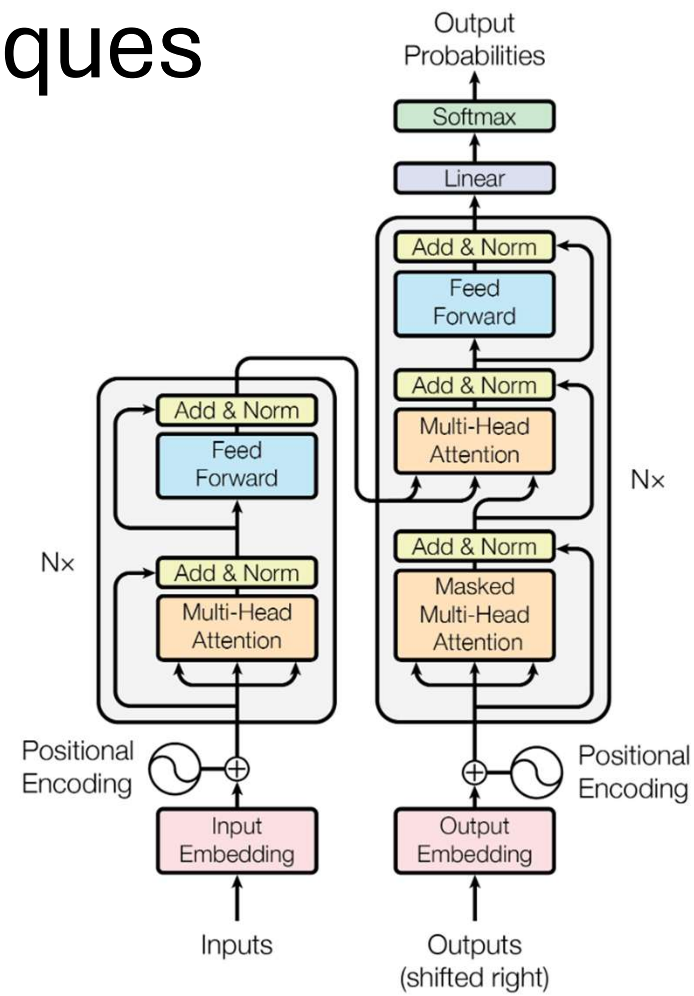
- Recurrent Neural Networks (RNNs)
  - Sequential data analysis
  - Natural language processing



Recurrent Neural Networks (RNNs):A gentle Introduction and Overview, https://doi.org/10.48550/arXiv.1912.05911

Various Deep Learning Techniques

- Transformers
  - Advanced language models
  - Context understanding



Attention Is All You Need, https://doi.org/10.48550/arXiv.1706.03762

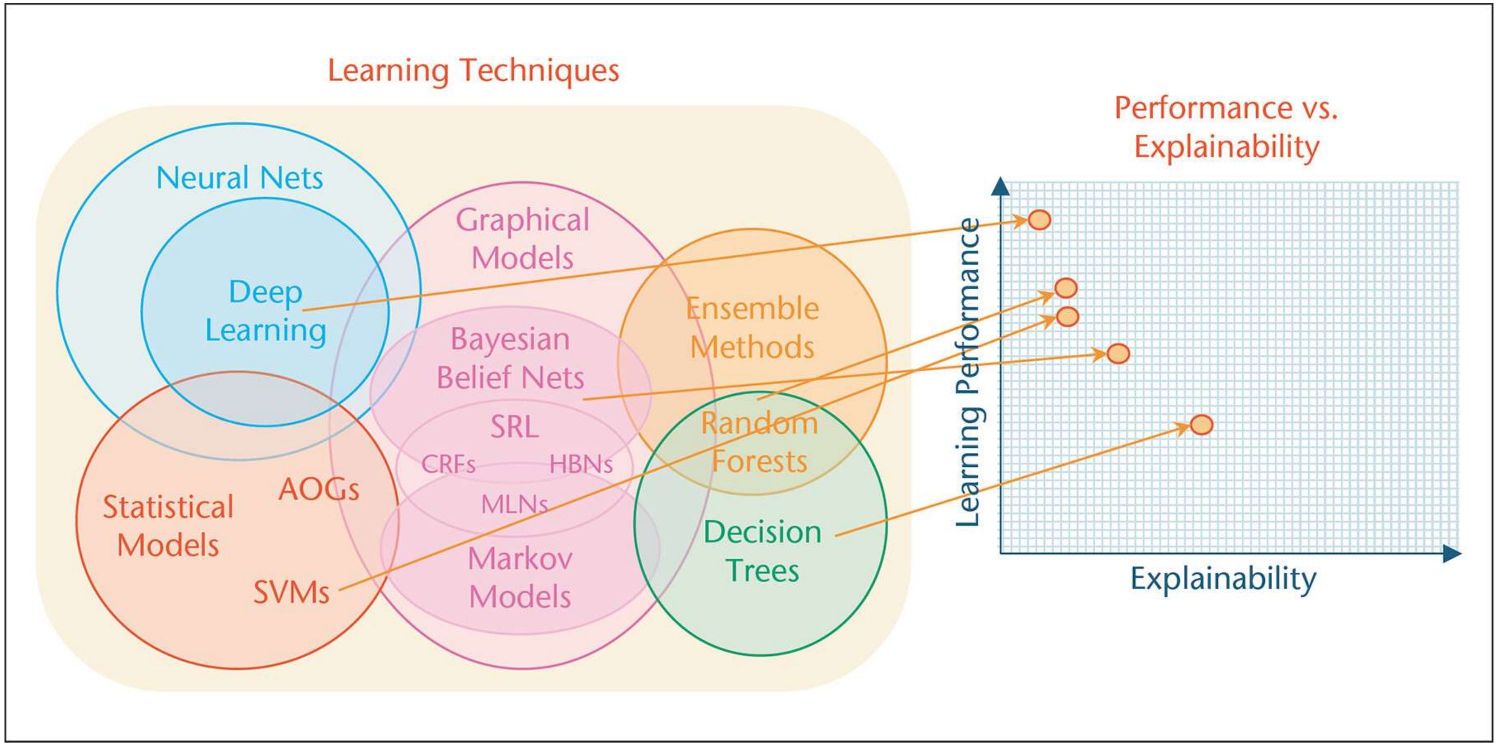
Training Deep Learning Models

- Large datasets critical for performance
- Requires significant computational power
- Techniques: Transfer Learning, Data Augmentation
- Challenges: Overfitting, Computational Complexity

Why We Need Explainable AI

- It is difficult to trust a system whose decisions we don't understand.
- Lack of transparency can lead to bias and unfair outcomes.
- Explainability is essential for accountability and regulation of AI.

Learning Performance vs. Explainability



[Gunning]

Overview

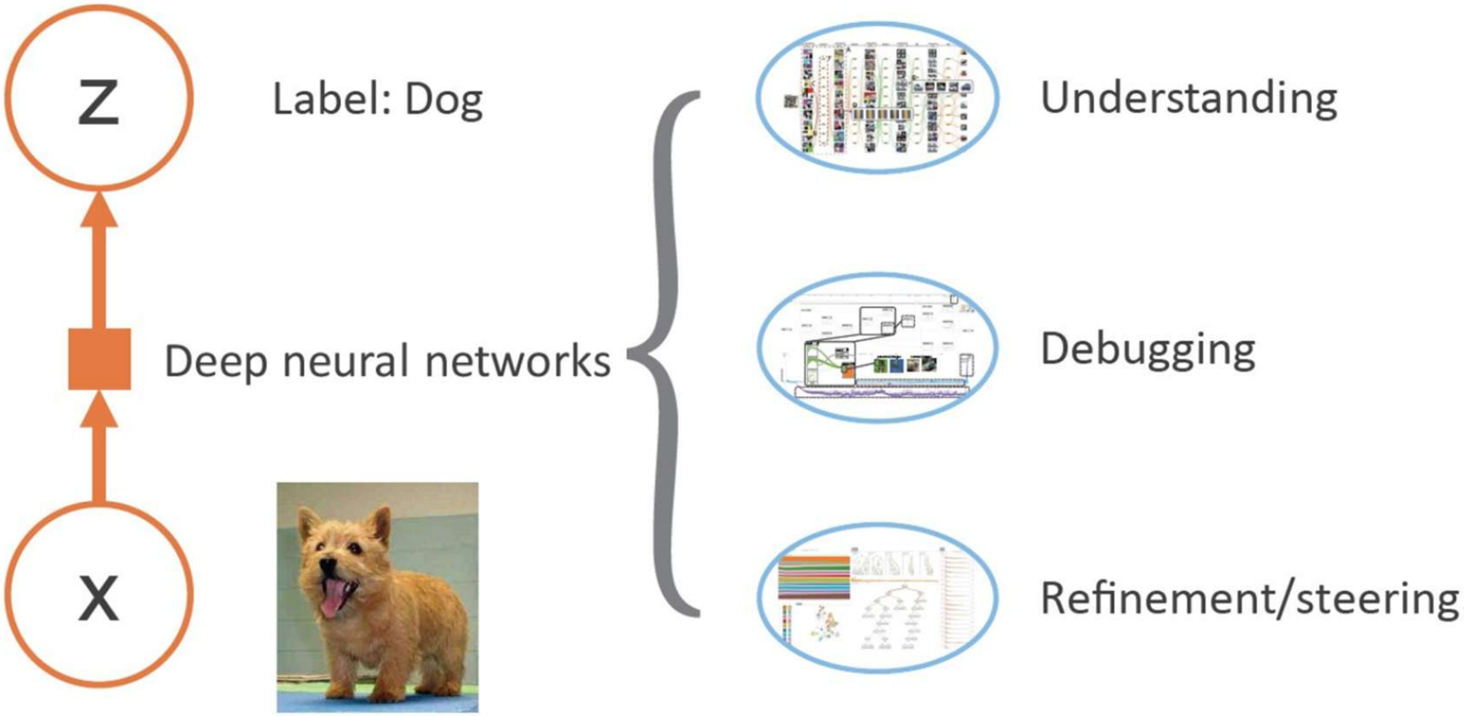


Figure 1. Overview of explainable deep learning

[Choo]

Definitions

- XAI aims to develop machine learning techniques that provide understandable, trustworthy and explainable rationales for decisions made by black-box models. [Adadi, Dragoni, Gunning]
- XAI is the class of systems that provide visibility into how an AI system makes decisions and predictions and executes its actions. [Rai]
- XAI refers to systems that try to explain how a black-box AI model produces its outcomes. [Moradi]

Definitions

- XAI aims to develop machine learning techniques that provide understandable, trustworthy and **explainable** rationales for decisions made by black-box models. [Adadi, Dragoni, Gunning]
- XAI is the class of systems that provide **visibility** into how an AI system makes decisions and predictions and executes its actions. [Rai]
- XAI refers to systems that try to **explain** how a black-box AI model produces its outcomes. [Moradi]

-> **Explain** how Black-Box-Models get to their decisions

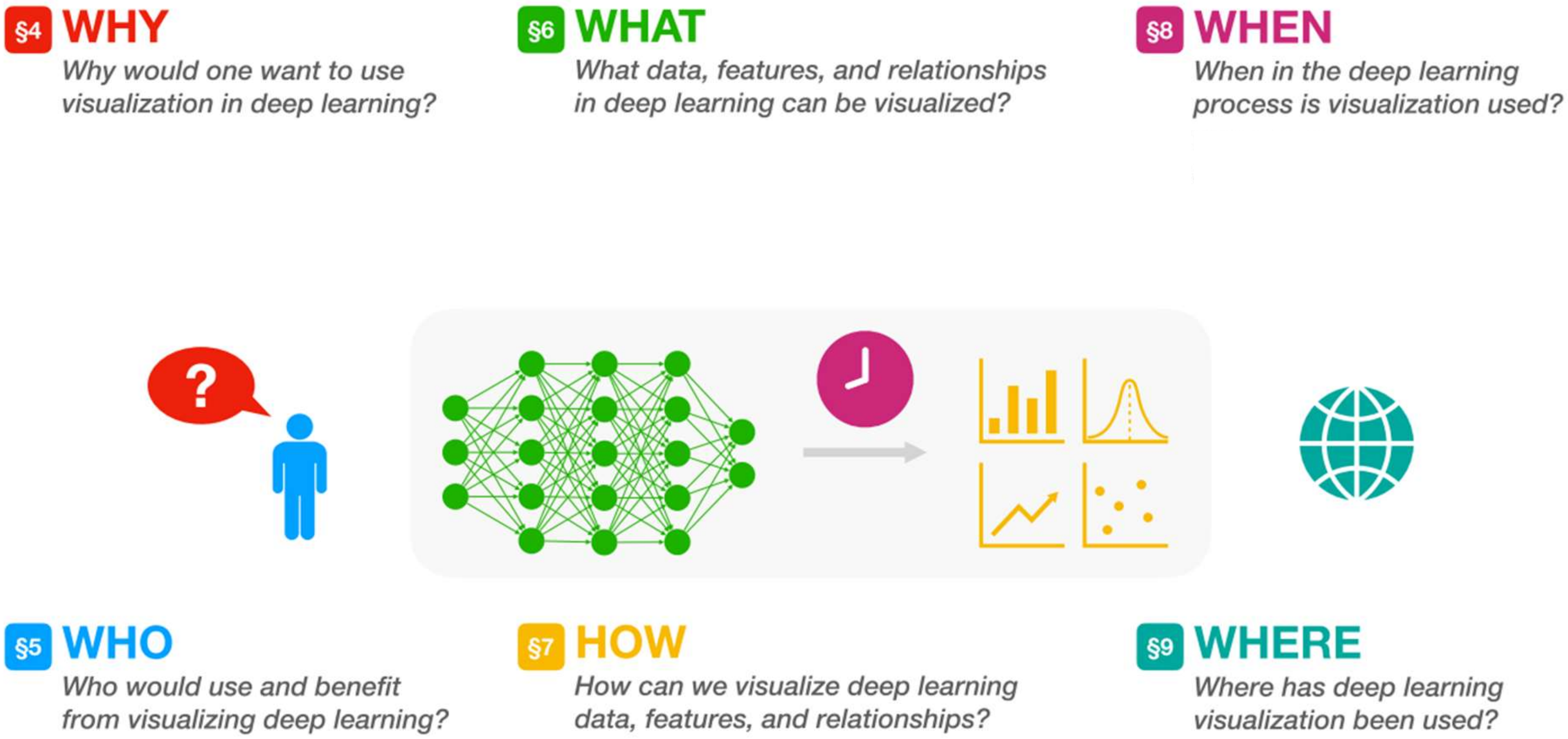
What is XAI

- XAI refers to techniques that make AI systems more understandable to humans.
- It aims to provide insights into how AI systems make decisions and why they reach conclusions.

Stakeholders / Who needs XAI?

- **Data scientists** use XAI to debug and improve models.
- **Domain experts** need to understand the reasoning behind AI's recommendations.
- **Decision-makers** rely on XAI to ensure responsible use of AI.
- **End users** deserve to know how AI affects their lives.

Important Questions



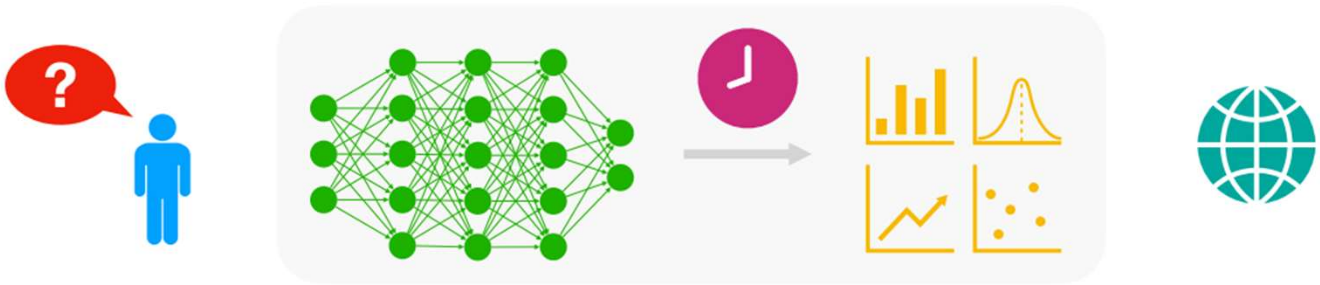
[Homann]

Important Questions

§4 WHY
Why would one want to use visualization in deep learning?
Interpretability & Explainability
Debugging & Improving Models
Comparing & Selecting Models
Teaching Deep Learning Concepts

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

§8 WHEN
When in the deep learning process is visualization used?



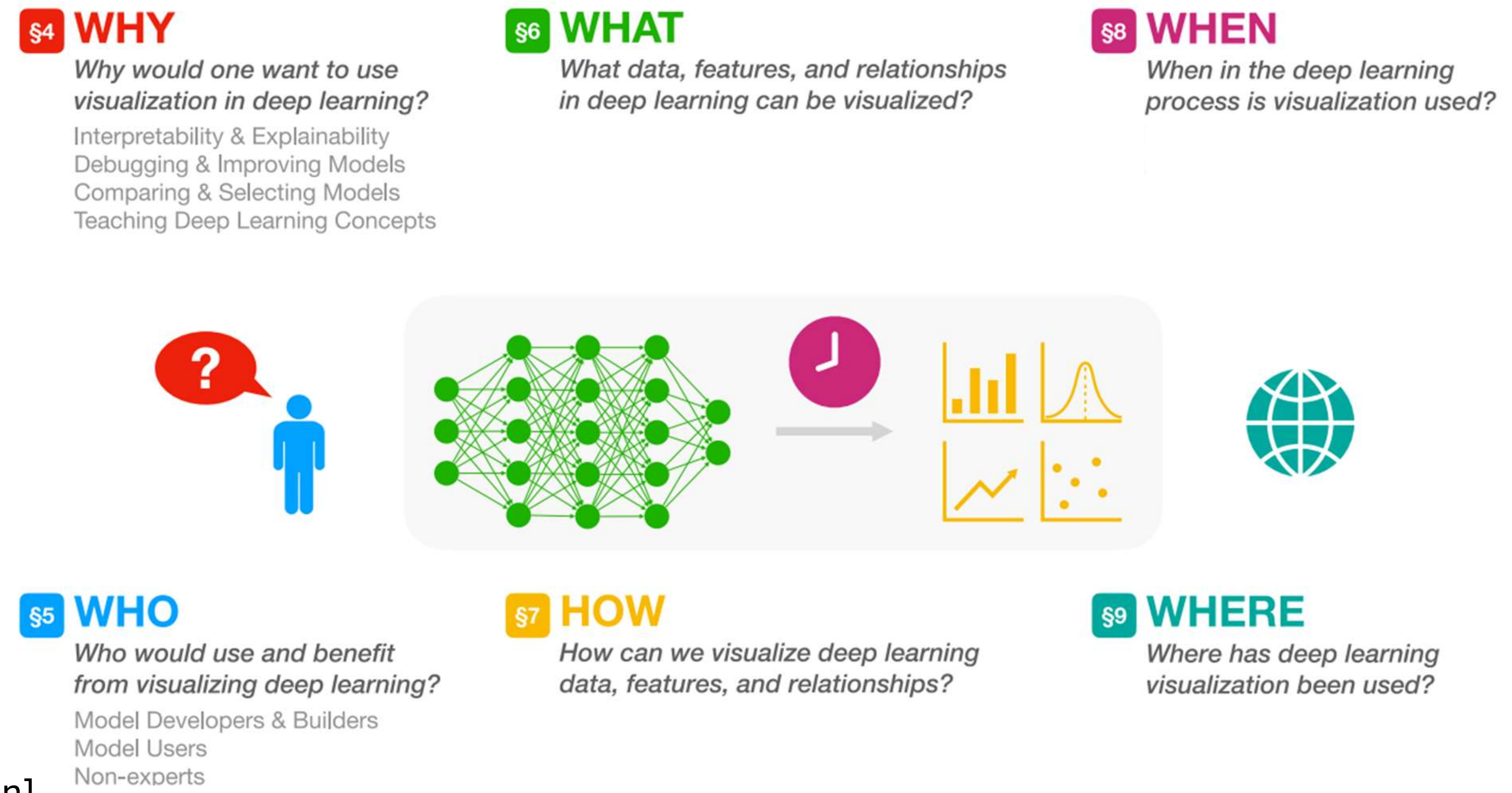
§5 WHO
Who would use and benefit from visualizing deep learning?

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

§9 WHERE
Where has deep learning visualization been used?

[Homann]

Important Questions



[Homann]

Overview – Visual Analytics for Machine Learning

[Yuan]

Overview – Visual Analytics for Machine Learning

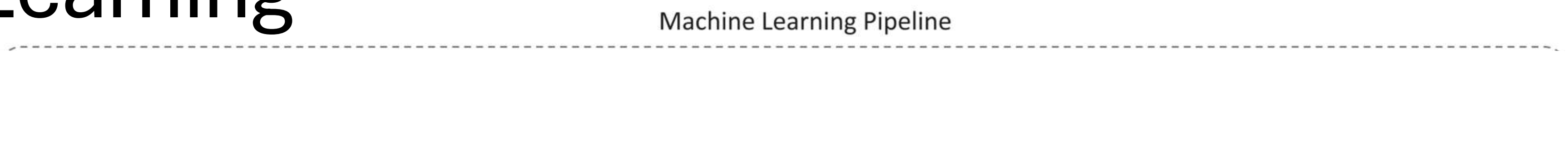


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

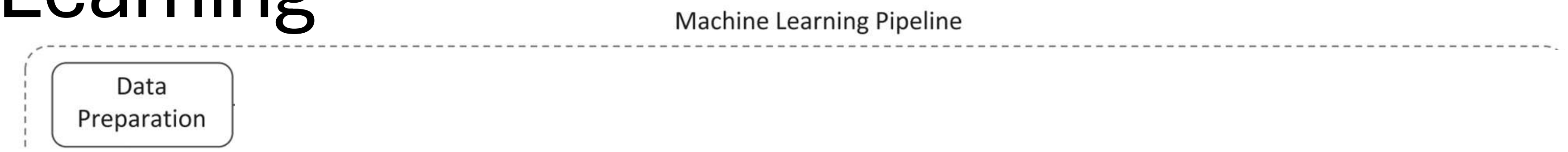


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

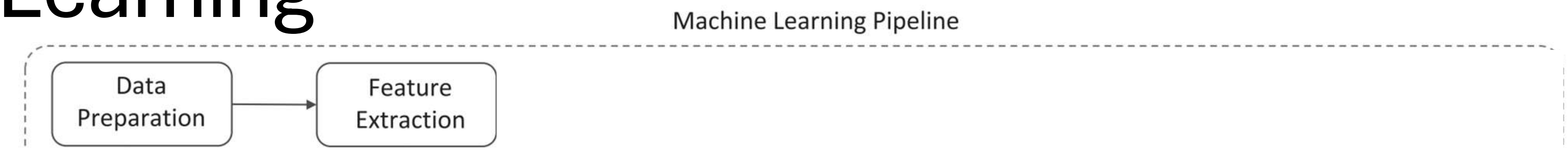


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

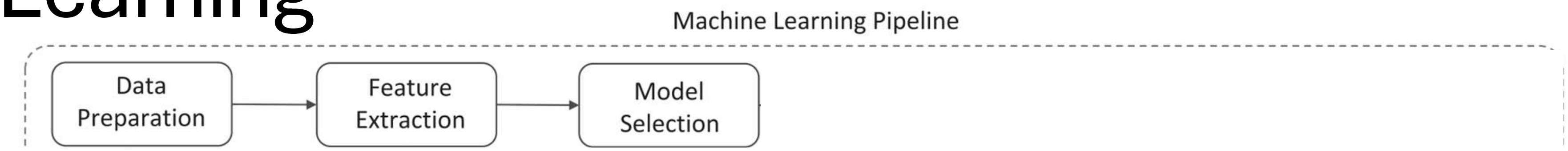


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

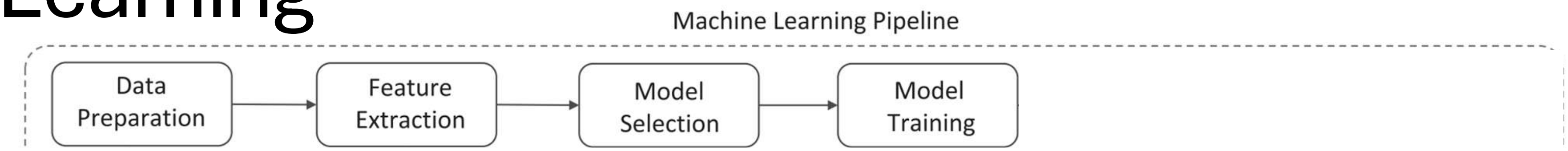


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

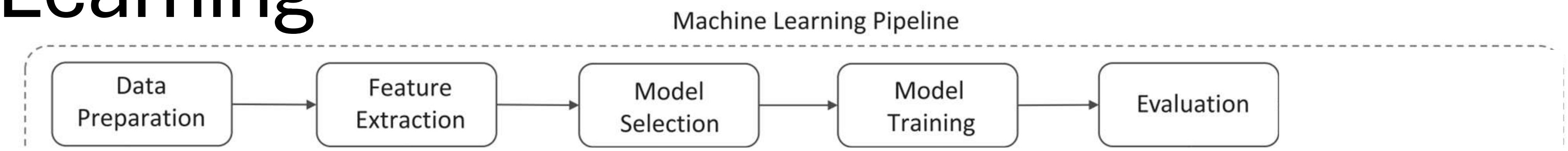


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

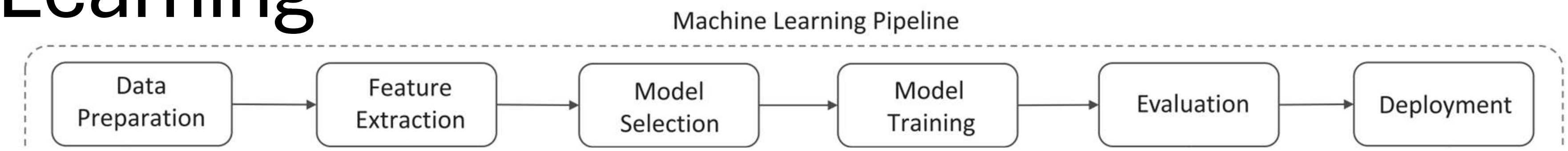


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

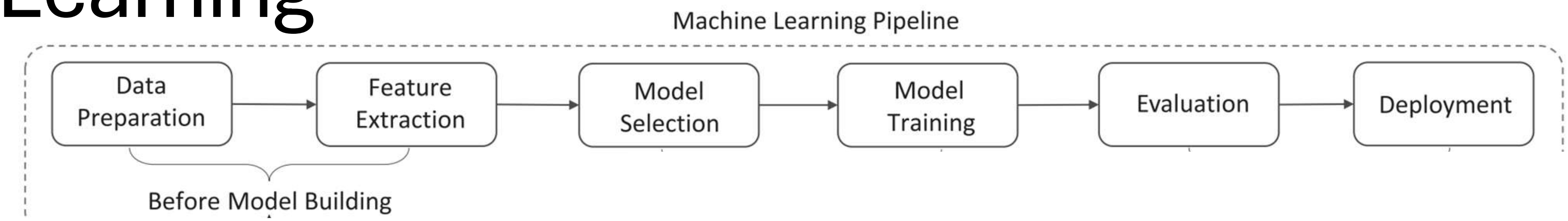


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

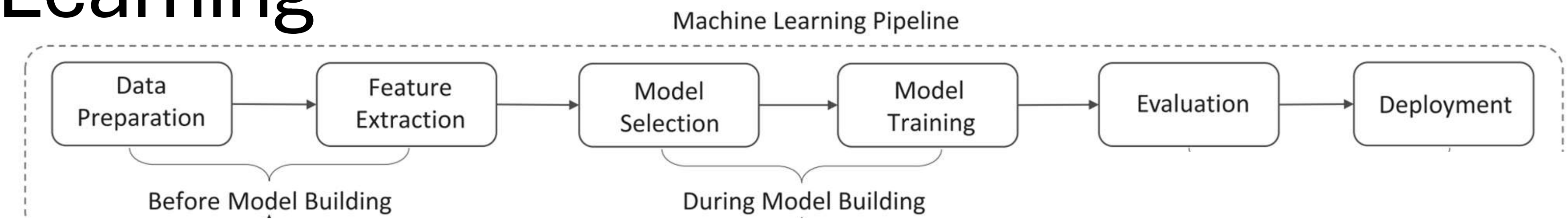


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

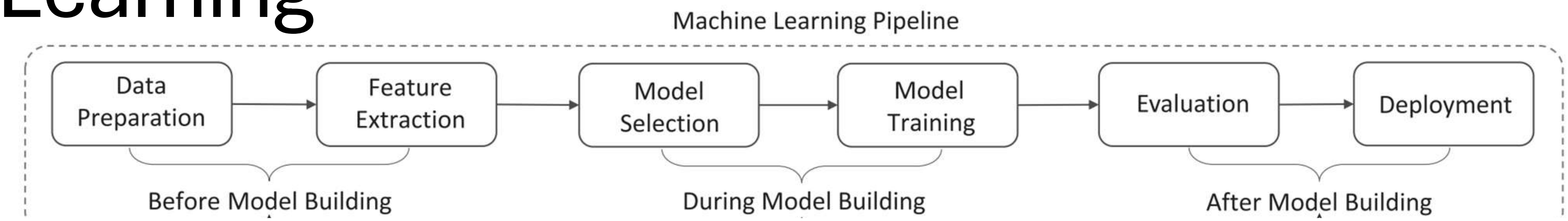


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

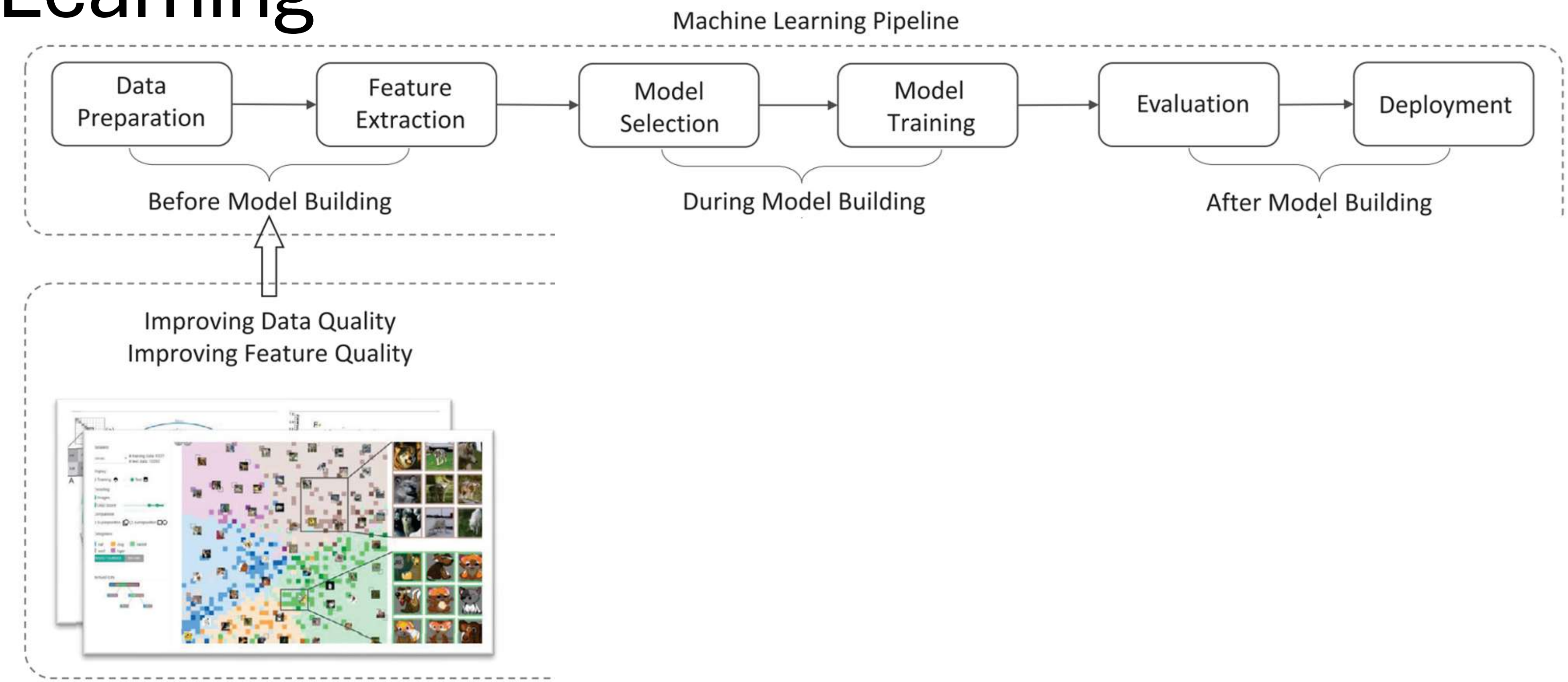


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

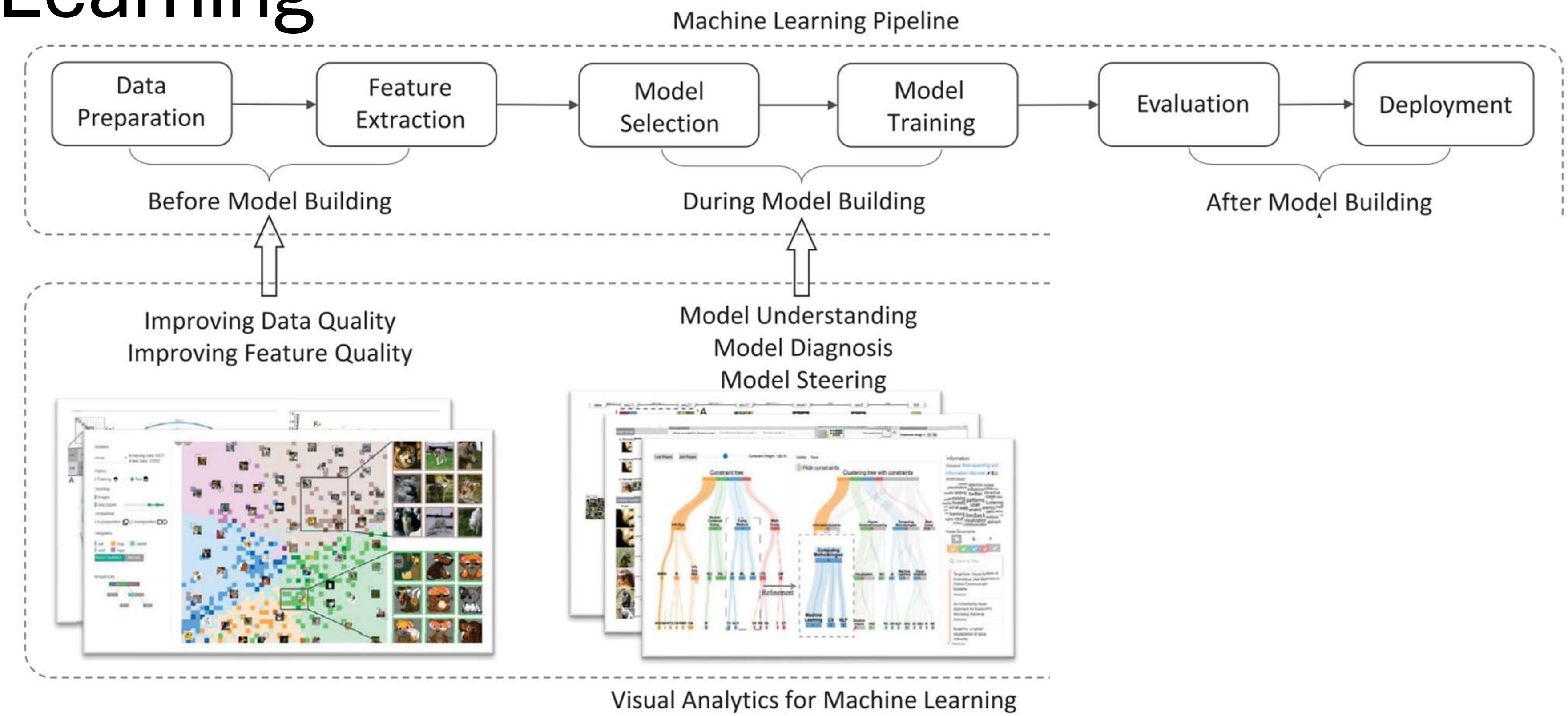


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview – Visual Analytics for Machine Learning

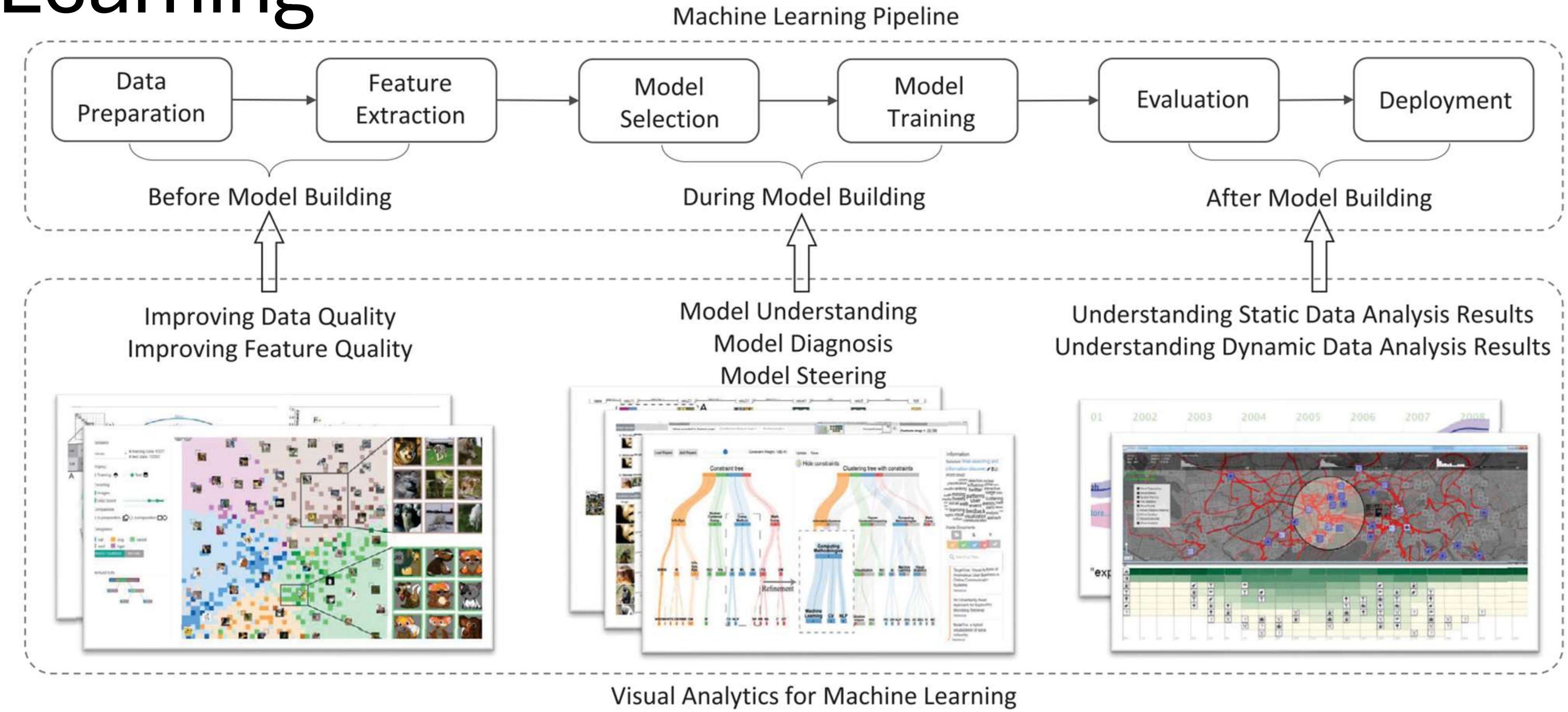


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Overview

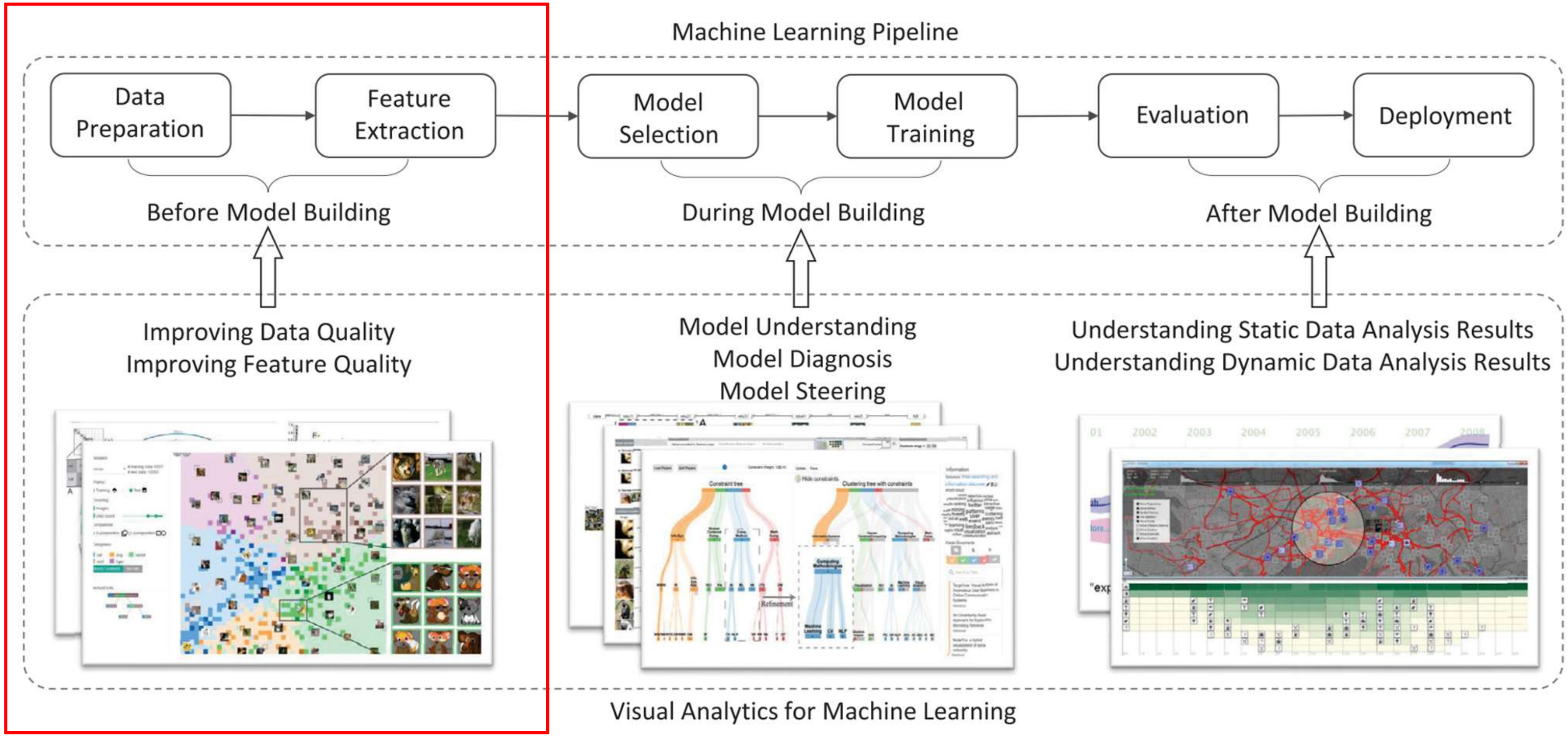


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Techniques **Before** Model Building

Ensure high-quality **data** and **features** to improve model performance and reliability.

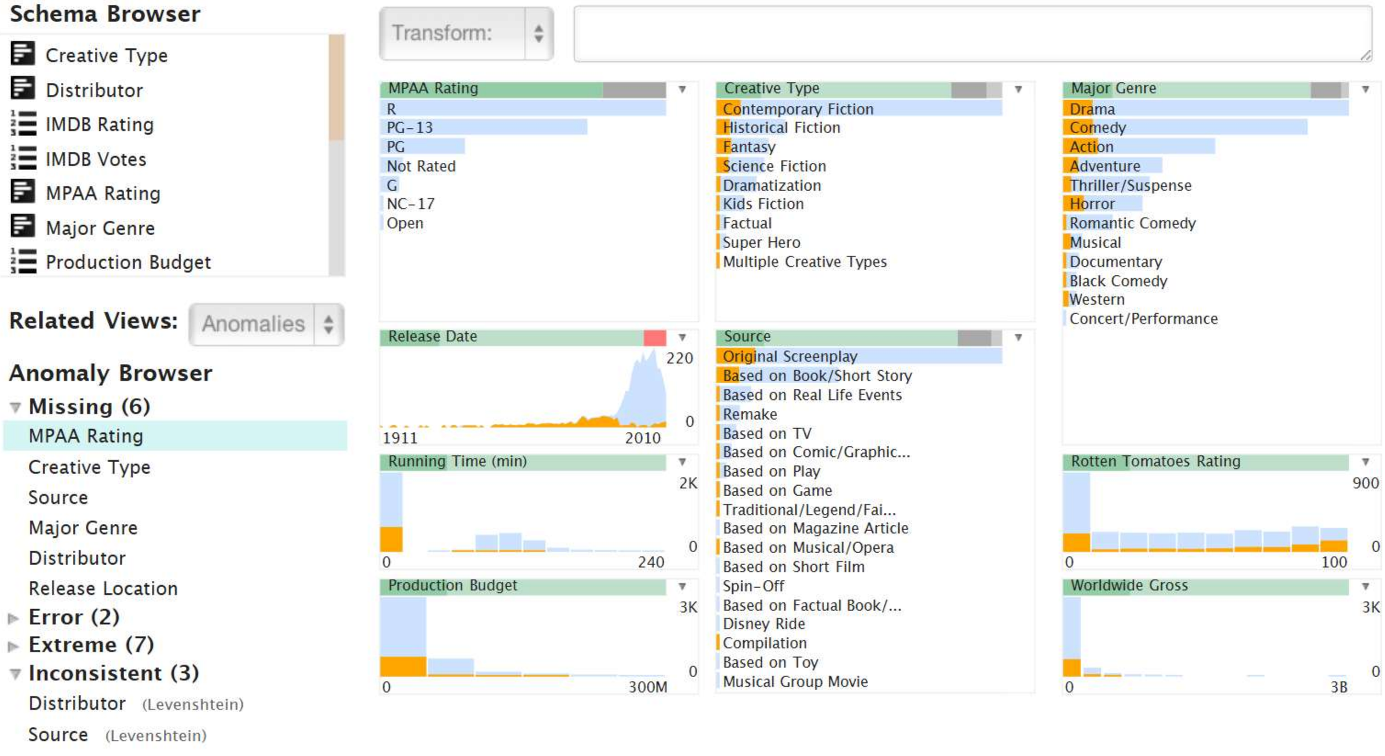
- Improving Data Quality
- Improving Feature Quality

Improving Data Quality - Instance-Level

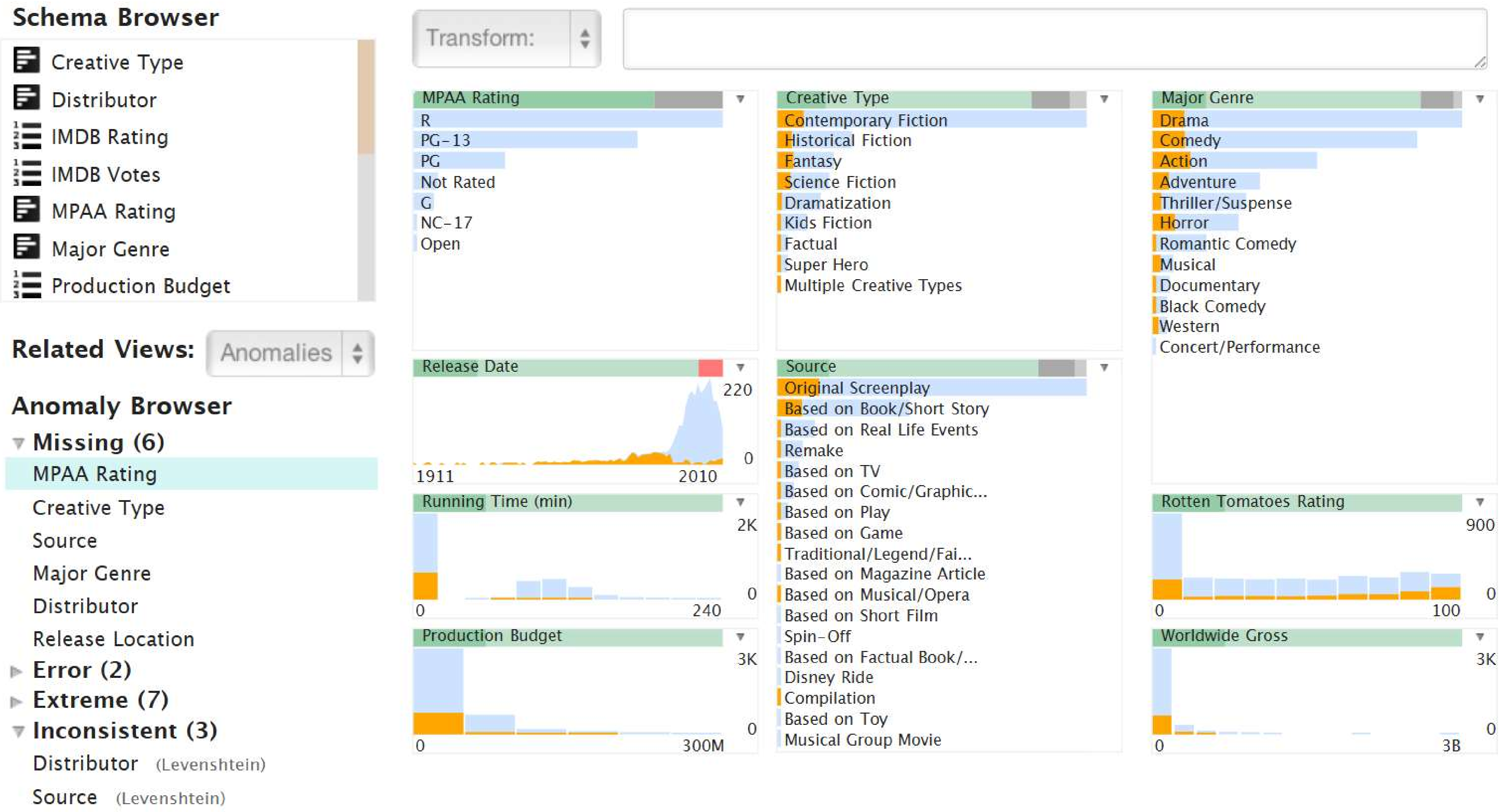
- Anomaly Detection and Correction:
  - Visualising and interacting with data to identify missing values,
  - outliers,
  - duplicates,
  - and out-of-distribution samples.
  - i.e. Profiler and OoDAnalyzer
- Provenance Tracking:
  - Illustrating the impact of data cleaning and preprocessing steps on data quality.
  - i.e. DQProv Explorer
- Privacy Preservation:
  - Balancing data utility with privacy concerns during the cleaning process
  - i.e. Privacy Exposure Risk Tree and GraphProtector.

Improving Data Quality - Instance-Level

- Anomaly Detection and Correction Example: Profiler



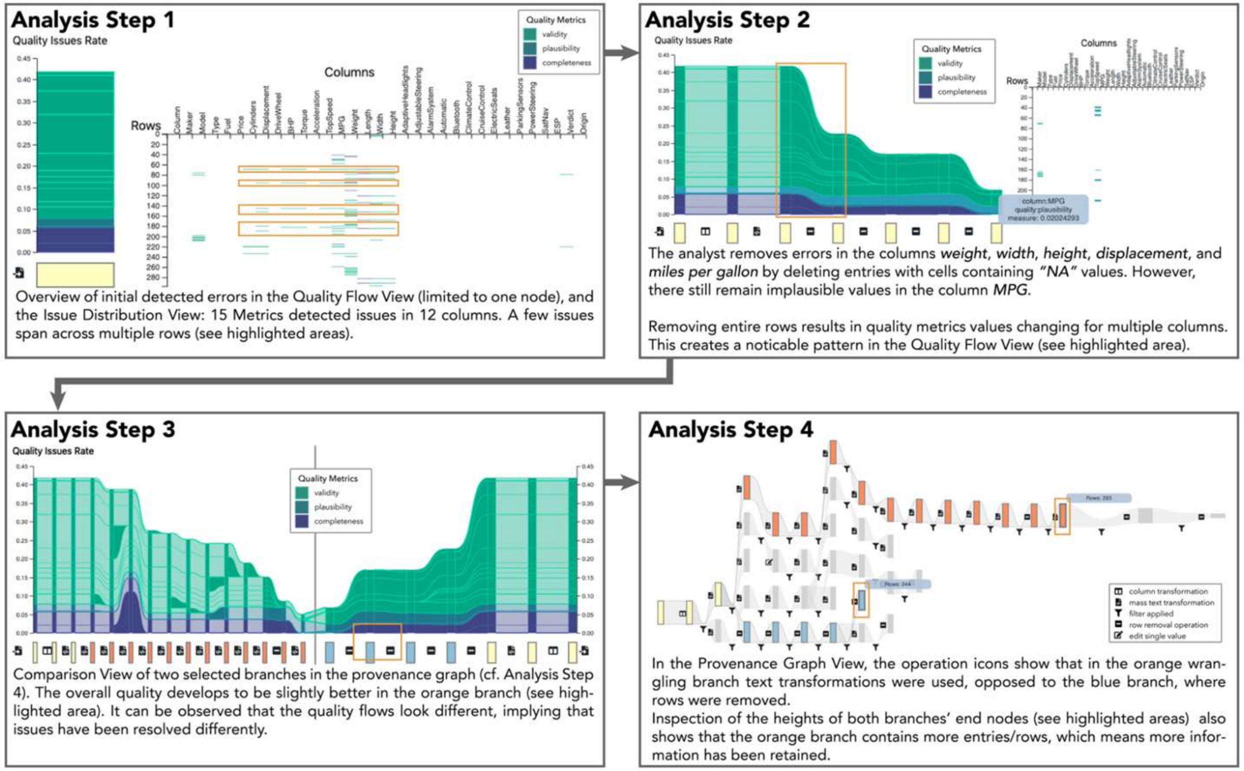
Profiler: integrated statistical analysis and visualization for data quality assessment, https://doi.org/10.1145/2254556.2254659



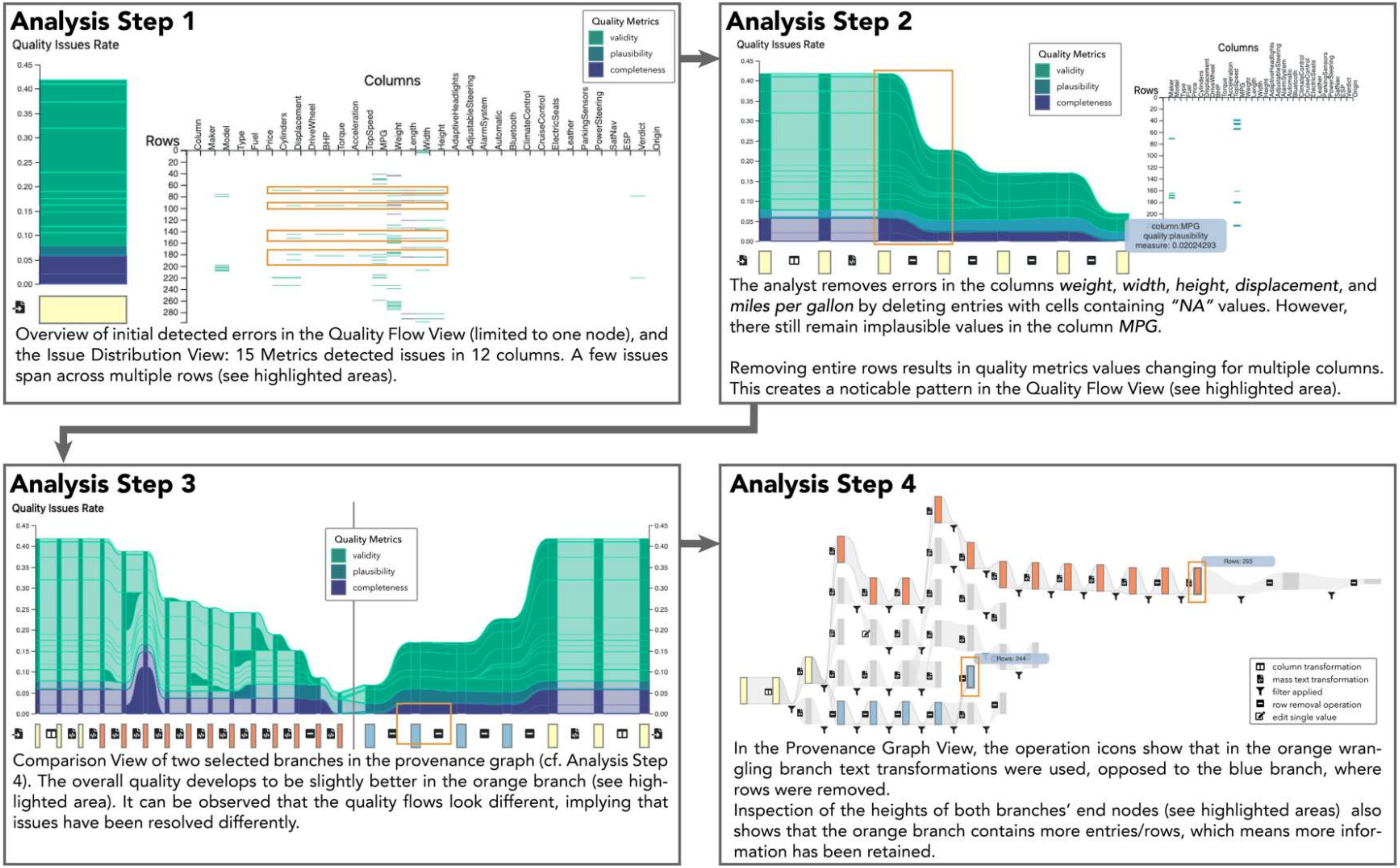
Schema Browser
Creative Type
Distributor
IMDB Rating
IMDB Votes
MPAA Rating
Major Genre
Production Budget
Related Views:
Anomalies
Anomaly Browser
Missing (6)
MPAA Rating
Creative Type
Source
Major Genre
Distributor
Release Location
Error (2)
Extreme (7)
Inconsistent (3)
Distributor (Levenshtein)
Source (Levenshtein)
Transform:
MPAA Rating
R
PG-13
PG
Not Rated
G
NC-17
Open
Creative Type
Contemporary Fiction
Historical Fiction
Fantasy
Science Fiction
Dramatization
Kids Fiction
Factual
Super Hero
Multiple Creative Types
Major Genre
Drama
Comedy
Action
Adventure
Thriller/Suspense
Horror
Romantic Comedy
Musical
Documentary
Black Comedy
Western
Concert/Performance
Release Date
220
0
1911
2010
Source
Original Screenplay
Based on Book/Short Story
Based on Real Life Events
Remake
Based on TV
Based on Comic/Graphic...
Based on Play
Based on Game
Traditional/Legend/Fai...
Based on Magazine Article
Based on Musical/Opera
Based on Short Film
Spin-Off
Based on Factual Book/...
Disney Ride
Compilation
Based on Toy
Musical Group Movie
Running Time (min)
2K
0
0
240
Rotten Tomatoes Rating
900
0
0
100
Production Budget
3K
0
0
300M
Worldwide Gross
3K
0
0
3B

Improving Data Quality - Instance-Level

- Provenance Tracking Example: DQProvExplorer

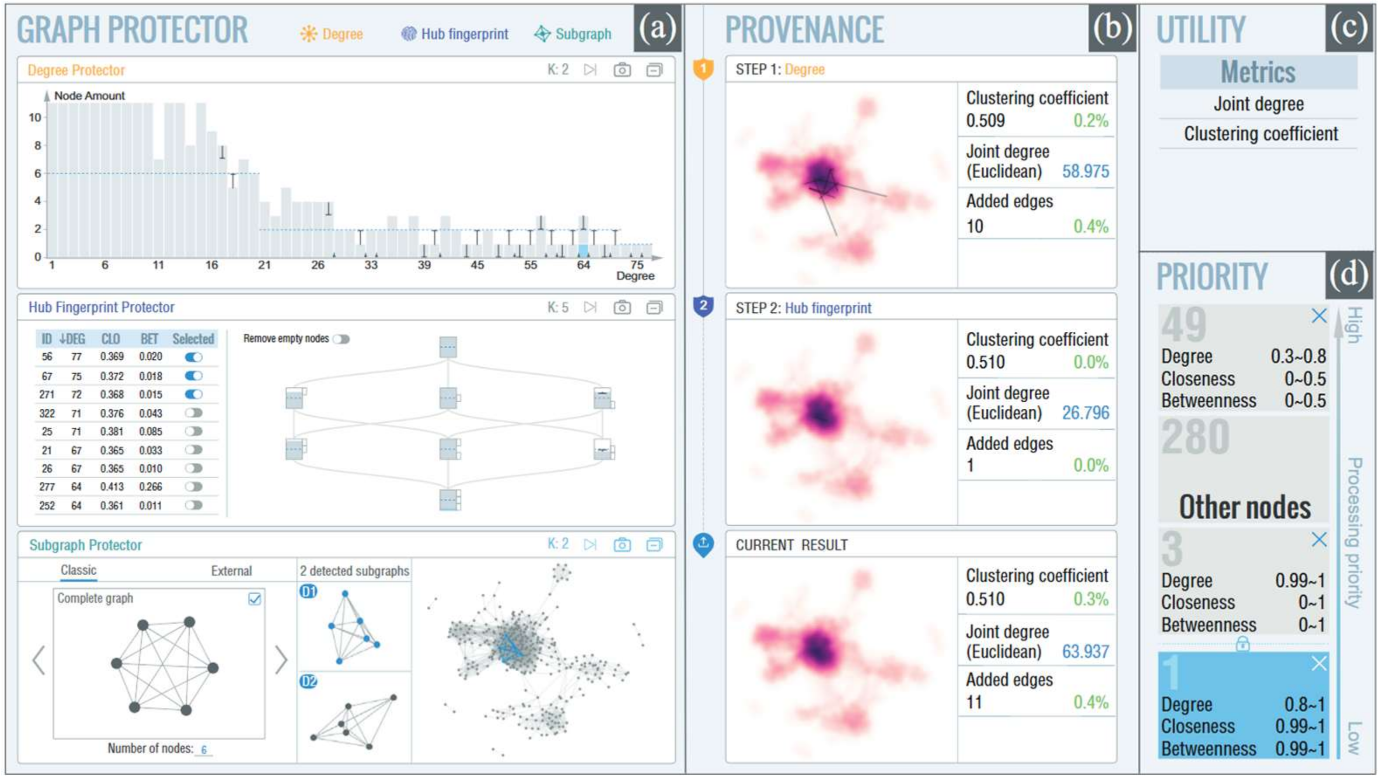


Capturing and Visualizing Provenance From DataWrangling, doi: 10.1109/MCG.2019.2941856.

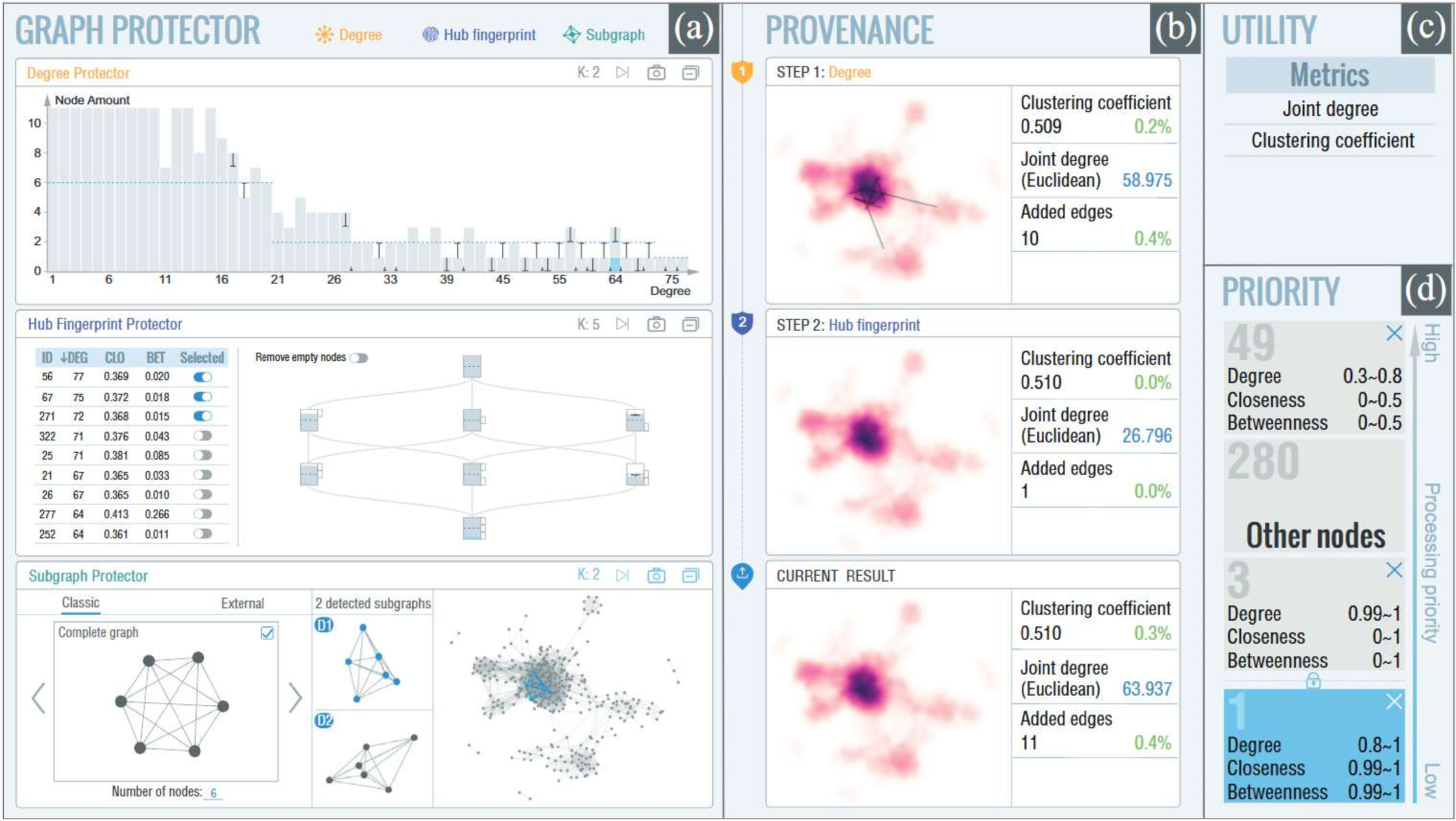

Analysis Step 1
Quality Issues Rate
Quality Metrics
validity
plausibility
completeness
Columns
Rows
Overview of initial detected errors in the Quality Flow View (limited to one node), and the Issue Distribution View: 15 Metrics detected issues in 12 columns. A few issues span across multiple rows (see highlighted areas).
Analysis Step 2
Quality Issues Rate
Quality Metrics
validity
plausibility
completeness
Columns
Rows
column:MPG
quality:plausibility
measure: 0.02024293
The analyst removes errors in the columns weight, width, height, displacement, and miles per gallon by deleting entries with cells containing "NA" values. However, there still remain implausible values in the column MPG.
Removing entire rows results in quality metrics values changing for multiple columns. This creates a noticable pattern in the Quality Flow View (see highlighted area).
Analysis Step 3
Quality Issues Rate
Quality Metrics
validity
plausibility
completeness
Comparison View of two selected branches in the provenance graph (cf. Analysis Step 4). The overall quality develops to be slightly better in the orange branch (see highlighted area). It can be observed that the quality flows look different, implying that issues have been resolved differently.
Analysis Step 4
Rows: 293
Rows: 244
column transformation
mass text transformation
filter applied
row removal operation
edit single value
In the Provenance Graph View, the operation icons show that in the orange wrangling branch text transformations were used, opposed to the blue branch, where rows were removed.
Inspection of the heights of both branches' end nodes (see highlighted areas) also shows that the orange branch contains more entries/rows, which means more information has been retained.

Improving Data Quality - Instance-Level

- Privacy Preservation Example: GraphProtector



GraphProtector: A Visual Interface for Employing and AssessingMultiple Privacy Preserving Graph Algorithms, doi: 10.1109/TVCG.2018.2865021.

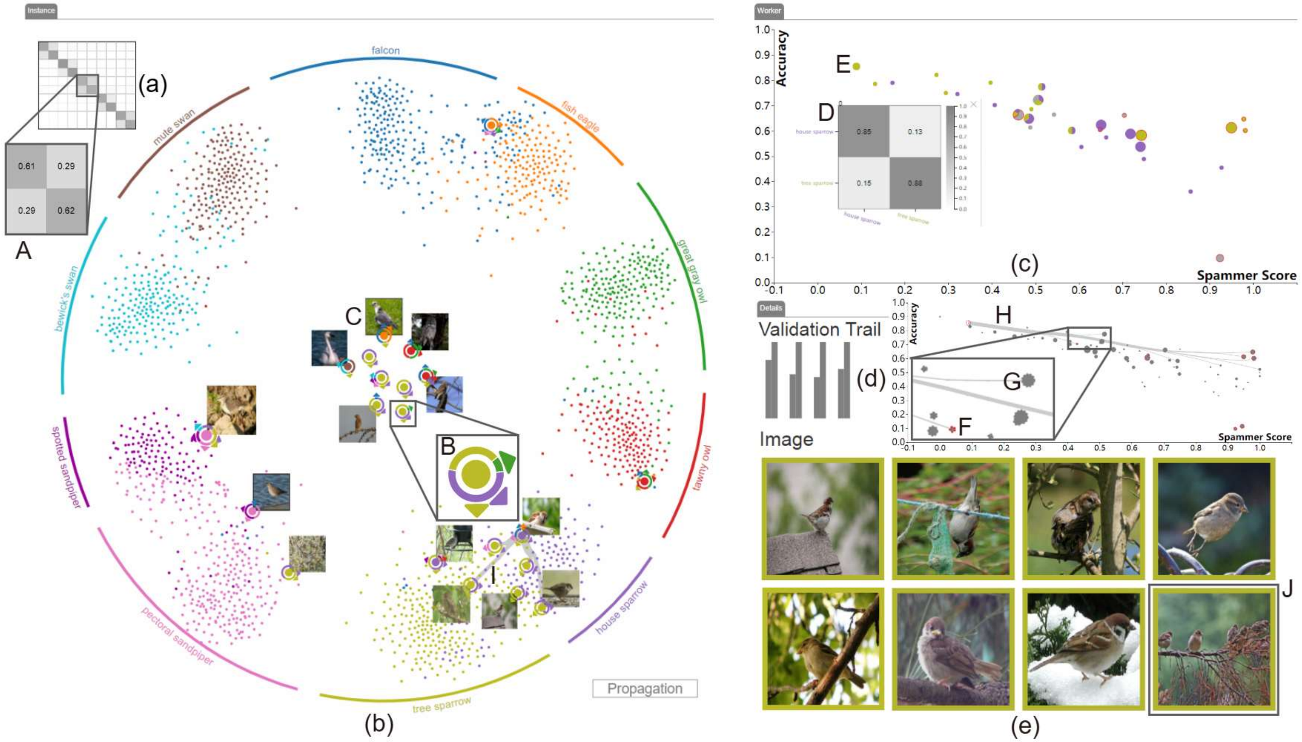

GRAPH PROTECTOR
Degree
Hub fingerprint
Subgraph
(a)
Degree Protector
K: 2
Node Amount
Degree
Hub Fingerprint Protector
K: 5
ID
↓DEG
CLO
BET
Selected
Remove empty nodes
Subgraph Protector
K: 2
Classic
External
Complete graph
Number of nodes: 6
2 detected subgraphs
D1
D2
PROVENANCE
(b)
STEP 1: Degree
Clustering coefficient
0.509 0.2%
Joint degree
(Euclidean) 58.975
Added edges
10 0.4%
STEP 2: Hub fingerprint
Clustering coefficient
0.510 0.0%
Joint degree
(Euclidean) 26.796
Added edges
1 0.0%
CURRENT RESULT
Clustering coefficient
0.510 0.3%
Joint degree
(Euclidean) 63.937
Added edges
11 0.4%
UTILITY
(c)
Metrics
Joint degree
Clustering coefficient
PRIORITY
(d)
49
Degree 0.3~0.8
Closeness 0~0.5
Betweenness 0~0.5
280
Other nodes
3
Degree 0.99~1
Closeness 0~1
Betweenness 0~1
1
Degree 0.8~1
Closeness 0.99~1
Betweenness 0.99~1
High
Processing priority
Low

Improving Data Quality - Label-Level

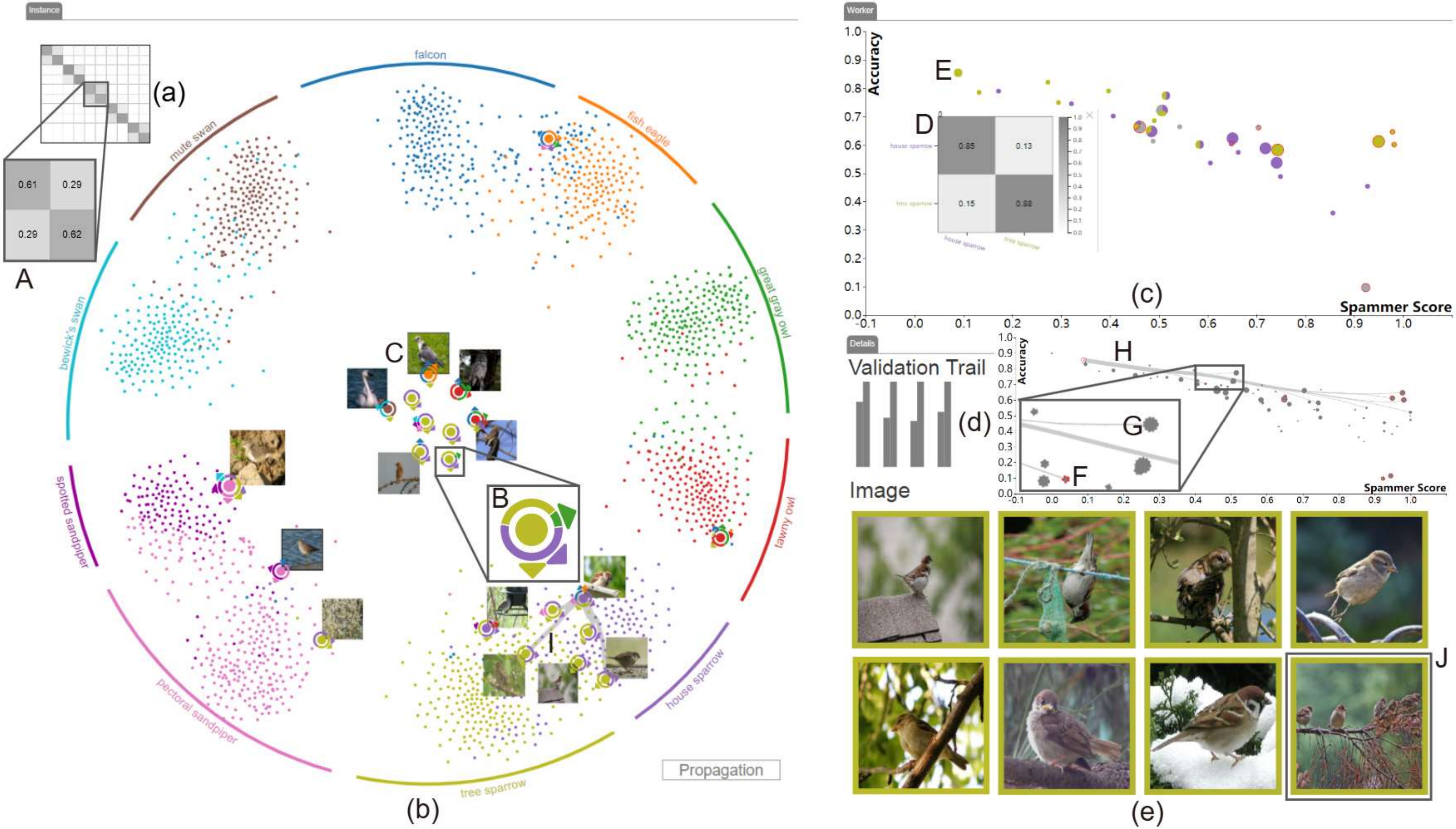
- Noisy Label Handling:
  - Visualising and refining crowdsourced annotations,
  - identifying unreliable workers,
  - correcting mislabelled instances.
  - i.e. LabelInspect and C2A
- Interactive Labelling:
  - Efficiently labelling unlabelled data by leveraging techniques like clustering similar instances and filtering to find items of interest.
  - i.e. MediaTable and the SOM-based visualisation by Moehrmann et al..

Improving Data Quality - Label-Level

- Noisy Label Handling Example: LabelInspect



An Interactive Method to Improve Crowdsourced Annotations, doi: 10.1109/TVCG.2018.2864843.



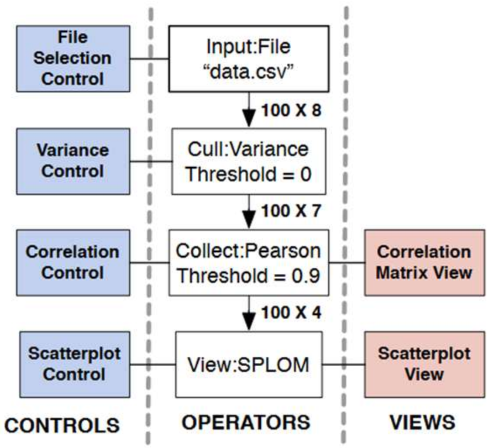
Instance
(a)
0.61
0.29
0.29
0.62
A
falcon
fish eagle
mute swan
bewick's swan
great gray owl
tawny owl
spotted sandpiper
pectoral sandpiper
tree sparrow
house sparrow
C
B
I
Propagation
(b)
Worker
Accuracy
E
D
house sparrow
tree sparrow
0.85
0.13
0.15
0.88
(c)
Spammer Score
Details
Validation Trail
(d)
H
G
F
Image
J
(e)

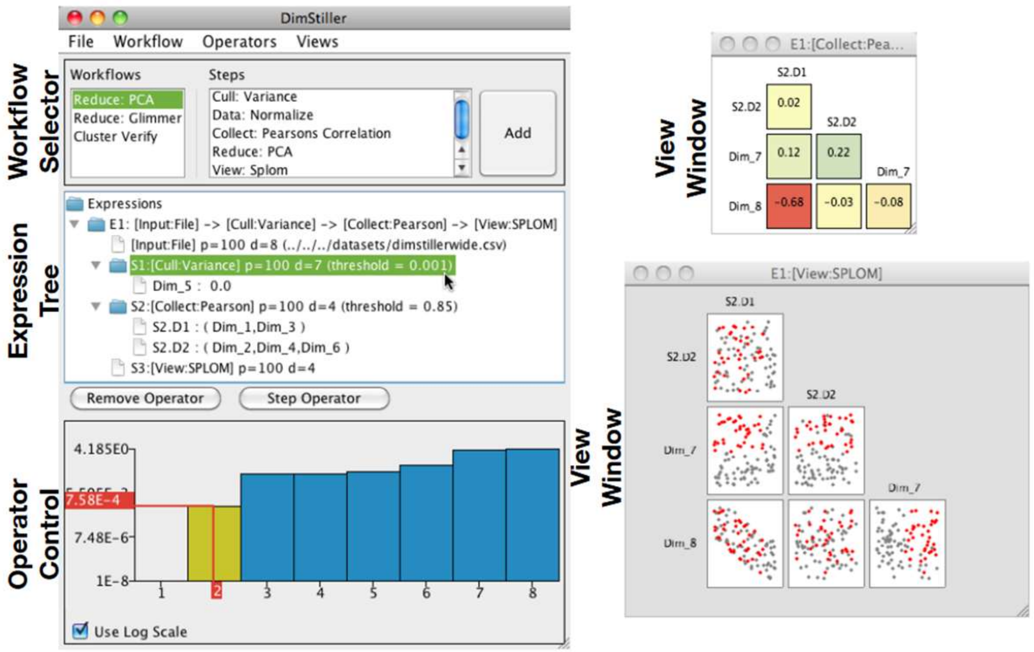
Improving Feature Quality

- Feature Selection:
  - Select useful features that contribute most to the prediction.
  - i.e. DimStiller and SmartStripes
- Feature Construction:
  - Guide the creation of new, more discriminative features.
  - i.e. FeatureInsight

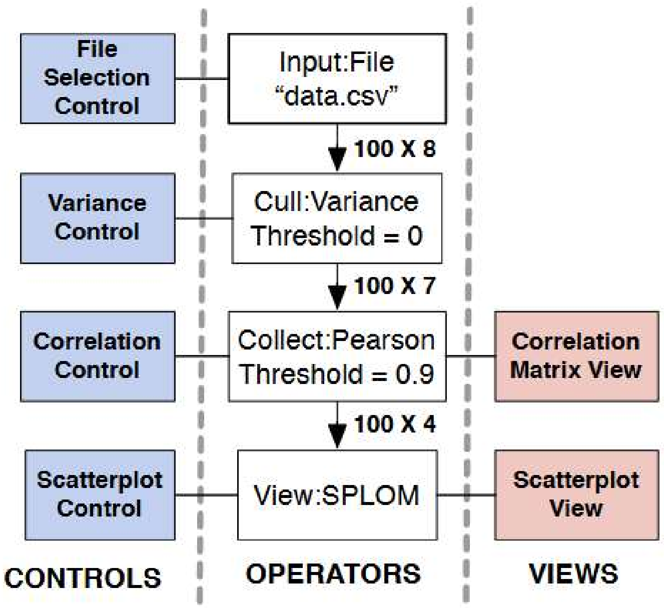
Improving Feature Quality

- Feature Selection Example: DimStiller

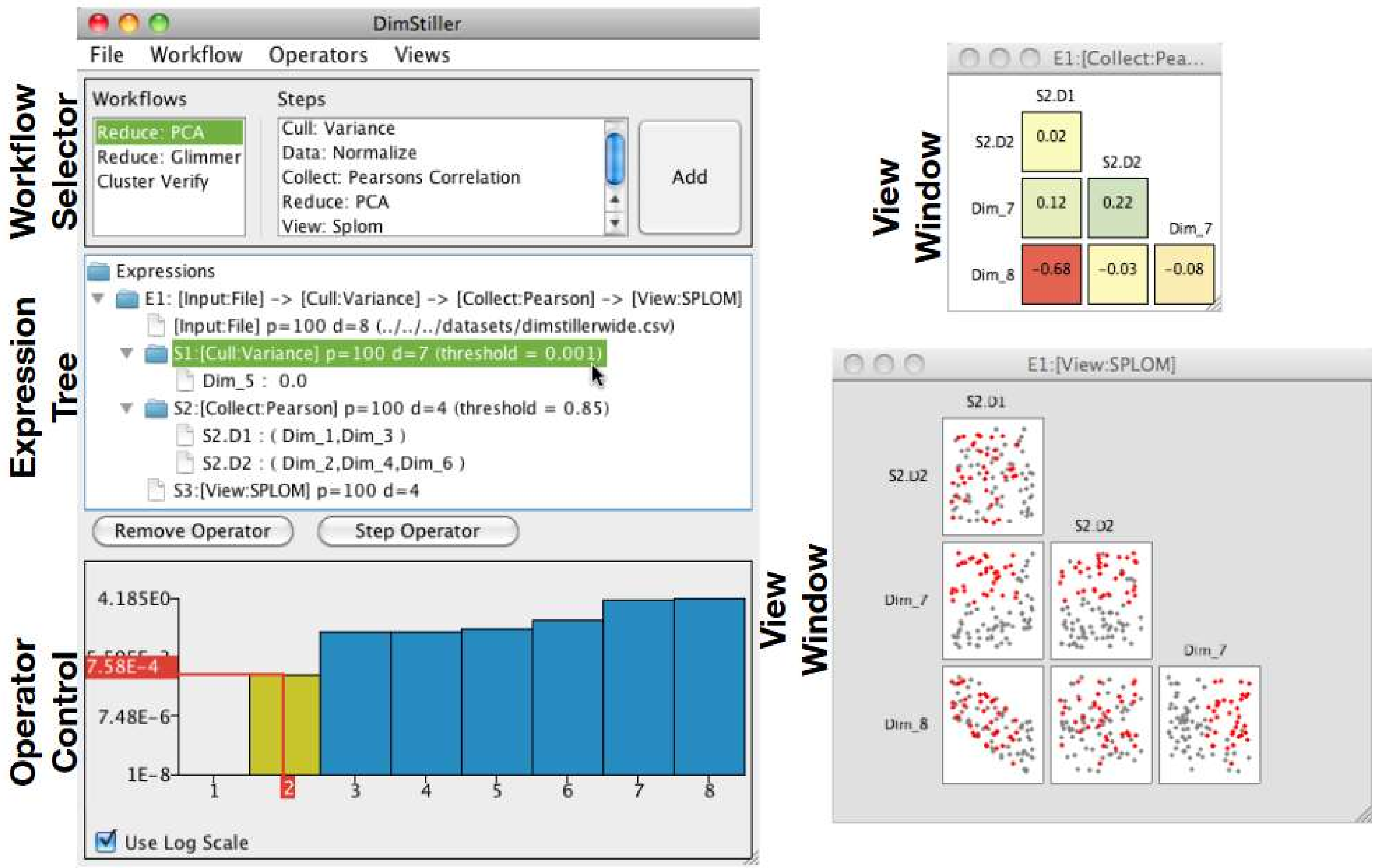




DimStiller: Workflows for Dimensional Analysis and Reduction, doi: 10.1109/VAST.2010.5652392.



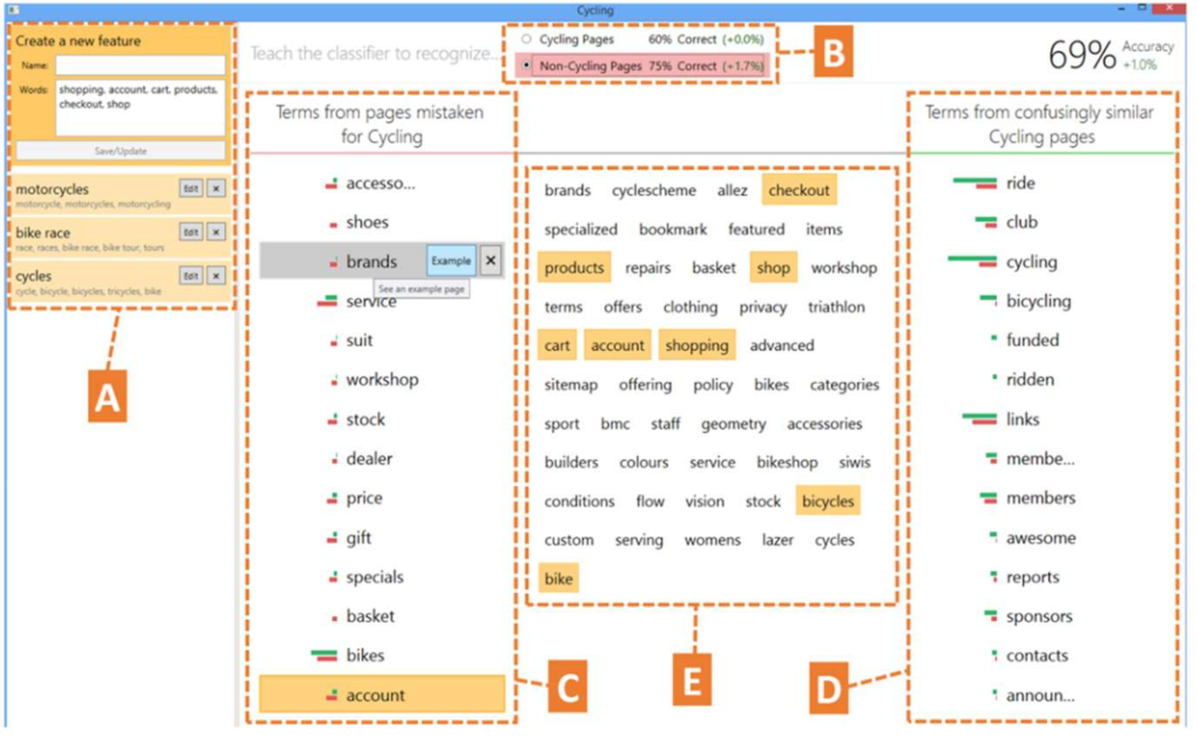
File Selection Control
Input:File "data.csv"
100 X 8
Variance Control
Cull:Variance Threshold = 0
100 X 7
Correlation Control
Collect:Pearson Threshold = 0.9
Correlation Matrix View
100 X 4
Scatterplot Control
View:SPLOM
Scatterplot View
CONTROLS
OPERATORS
VIEWS



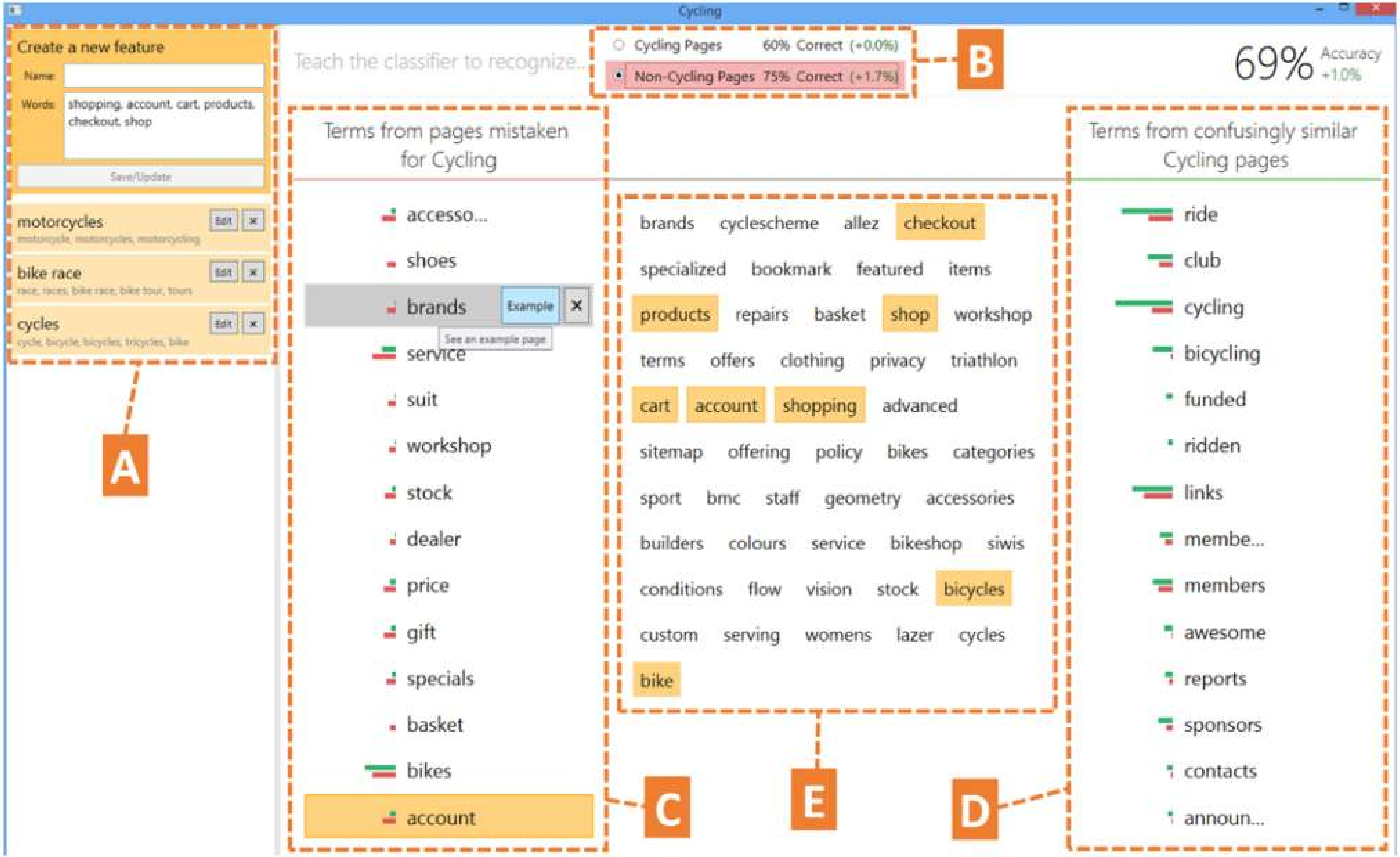
Workflow Selector
DimStiller
File Workflow Operators Views
Workflows
Reduce: PCA
Reduce: Glimmer
Cluster Verify
Steps
Cull: Variance
Data: Normalize
Collect: Pearsons Correlation
Reduce: PCA
View: Splom
Add
Expression Tree
Expressions
E1: [Input:File] -> [Cull:Variance] -> [Collect:Pearson] -> [View:SPLOM]
[Input:File] p=100 d=8 (../../../datasets/dimstillerwide.csv)
S1:[Cull:Variance] p=100 d=7 (threshold = 0.001)
Dim_5 : 0.0
S2:[Collect:Pearson] p=100 d=4 (threshold = 0.85)
S2.D1 : (Dim_1,Dim_3)
S2.D2 : (Dim_2,Dim_4,Dim_6)
S3:[View:SPLOM] p=100 d=4
Remove Operator
Step Operator
Operator Control
4.185E0
7.58E-4
7.48E-6
1E-8
1 2 3 4 5 6 7 8
Use Log Scale
View Window
E1:[Collect:Pea...
S2.D1
S2.D2
Dim_7
Dim_8
0.02
0.12
0.22
-0.68
-0.03
-0.08
View Window
E1:[View:SPLOM]
S2.D1
S2.D2
Dim_7
Dim_8

Improving Feature Quality

- Feature Construction Example: FeatureInsight



FeatureInsight: Visual Support forError-Driven Feature Ideation in Text Classification, doi: 10.1109/VAST.2015.7347637.


Cycling
Create a new feature
Name:
Words: shopping, account, cart, products, checkout, shop
Save/Update
motorcycles
motorcycle, motorcycles, motorcycling
bike race
race, races, bike race, bike tour, tours
cycles
cycle, bicycle, bicycles, tricycles, bike
Edit
A
Teach the classifier to recognize...
Cycling Pages 60% Correct (+0.0%)
Non-Cycling Pages 75% Correct (+1.7%)
B
69% Accuracy +1.0%
Terms from pages mistaken for Cycling
accesso...
shoes
brands
Example
See an example page
service
suit
workshop
stock
dealer
price
gift
specials
basket
bikes
account
C
brands cyclescheme allez checkout
specialized bookmark featured items
products repairs basket shop workshop
terms offers clothing privacy triathlon
cart account shopping advanced
sitemap offering policy bikes categories
sport bmc staff geometry accessories
builders colours service bikeshop siwis
conditions flow vision stock bicycles
custom serving womens lazer cycles
bike
E
Terms from confusingly similar Cycling pages
ride
club
cycling
bicycling
funded
ridden
links
membe...
members
awesome
reports
sponsors
contacts
announ...
D

Overview – Visual Analytics for Machine Learning

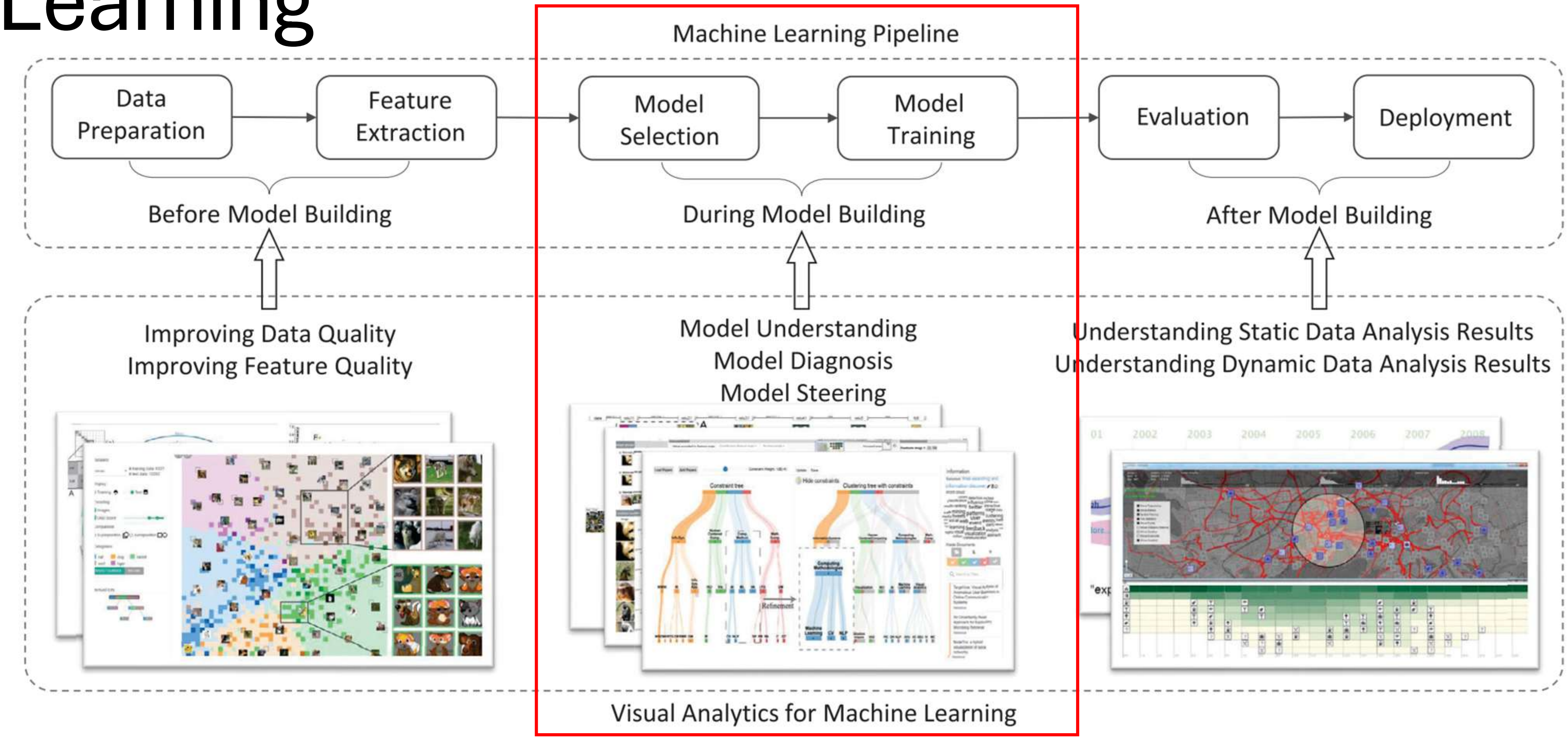


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Techniques **During** Model Building

Gain deeper understanding of model workings, diagnose training issues, and steer model behaviour towards desired outcomes.

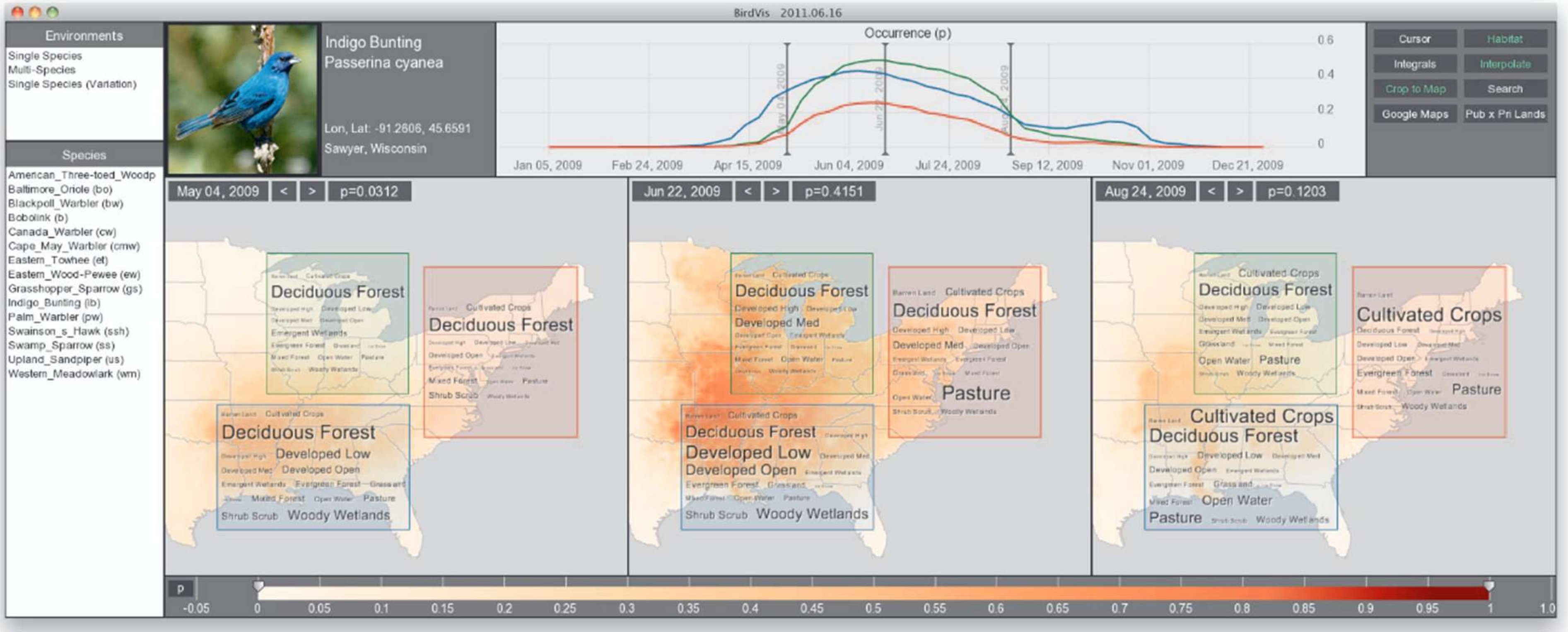
- Model Understanding
- Model Diagnosis
- Model Steering

Model Understanding - Parameter Effects

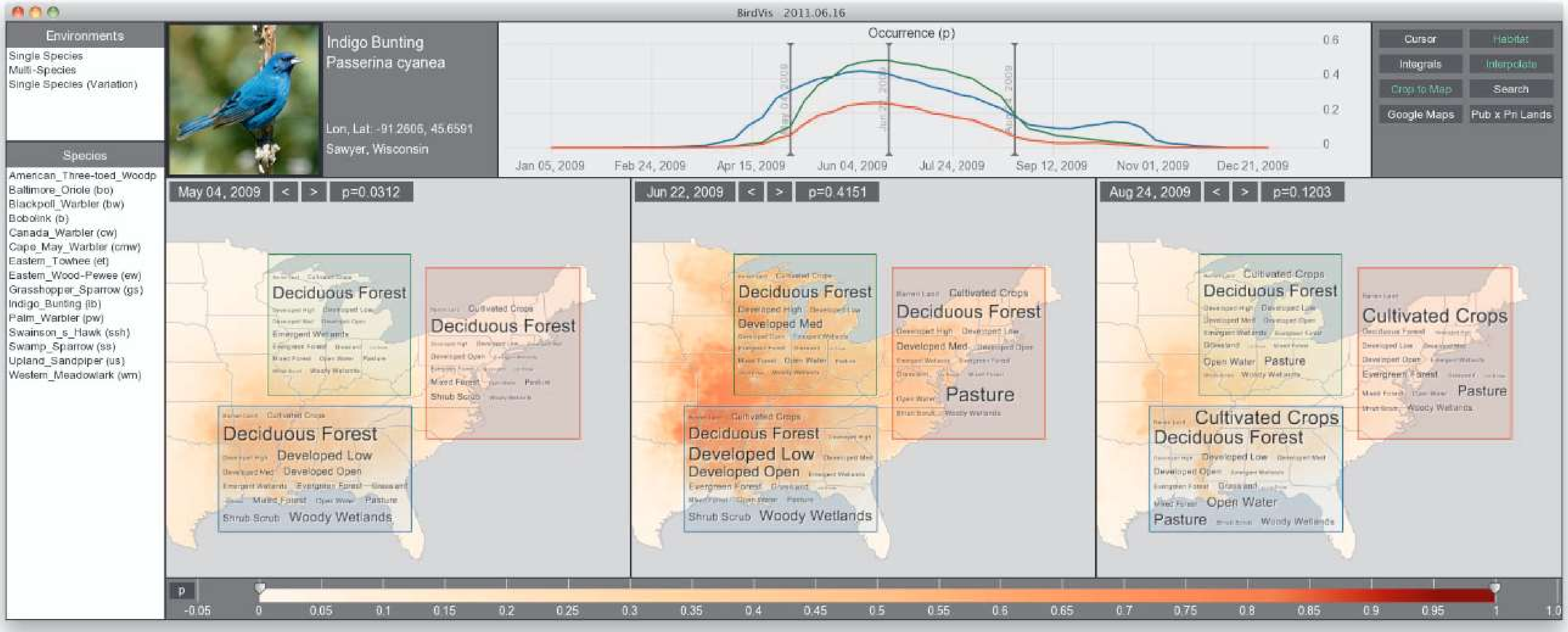
- Understanding Parameter Effects:
  - Visualising how model outputs change with variations in parameter settings.
  - i.e. BirdVis

Model Understanding - Parameter Effects

- Understanding Parameter Effects Example: BirdVis



BirdVis: Visualizing and Understanding Bird Populations, doi: 10.1109/TVCG.2011.176.

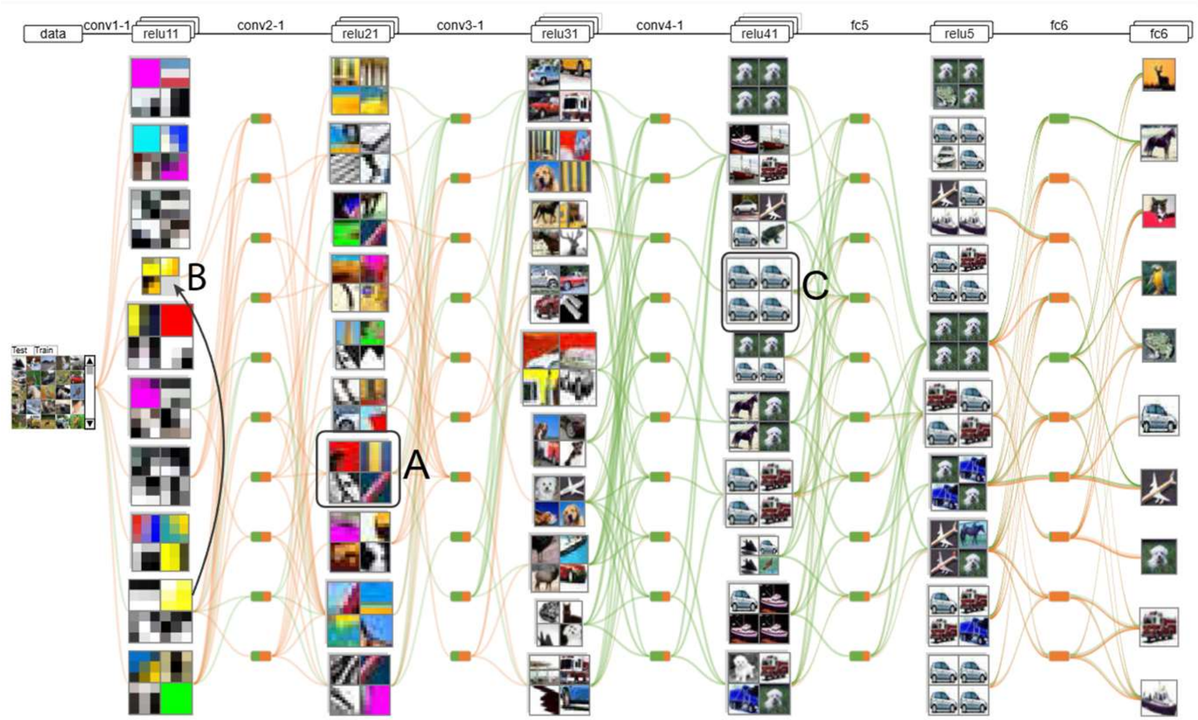

BirdVis 2011.06.16
Environments
Single Species
Multi-Species
Single Species (Variation)
Species
American_Three-toed_Woodp
Baltimore_Oriole (bo)
Blackpoll_Warbler (bw)
Bobolink (b)
Canada_Warbler (cw)
Cape_May_Warbler (cmw)
Eastern_Towhee (et)
Eastern_Wood-Pewee (ew)
Grasshopper_Sparrow (gs)
Indigo_Bunting (ib)
Palm_Warbler (pw)
Swainson_s_Hawk (ssh)
Swamp_Sparrow (ss)
Upland_Sandpiper (us)
Western_Meadowlark (wm)
Indigo Bunting
Passerina cyanea
Lon, Lat: -91.2606, 45.6591
Sawyer, Wisconsin
Occurrence (p)
0.6
0.4
0.2
0
Jan 05, 2009
Feb 24, 2009
Apr 15, 2009
Jun 04, 2009
Jul 24, 2009
Sep 12, 2009
Nov 01, 2009
Dec 21, 2009
Cursor
Habitat
Integrals
Interpolate
Crop to Map
Search
Google Maps
Pub x Pri Lands
May 04, 2009
p=0.0312
Jun 22, 2009
p=0.4151
Aug 24, 2009
p=0.1203
Deciduous Forest
Cultivated Crops
Developed Low
Developed Med
Developed Open
Developed High
Emergent Wetlands
Evergreen Forest
Grassland
Mixed Forest
Open Water
Pasture
Shrub Scrub
Woody Wetlands
P
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
1.0

Model Understanding - Model Behaviours

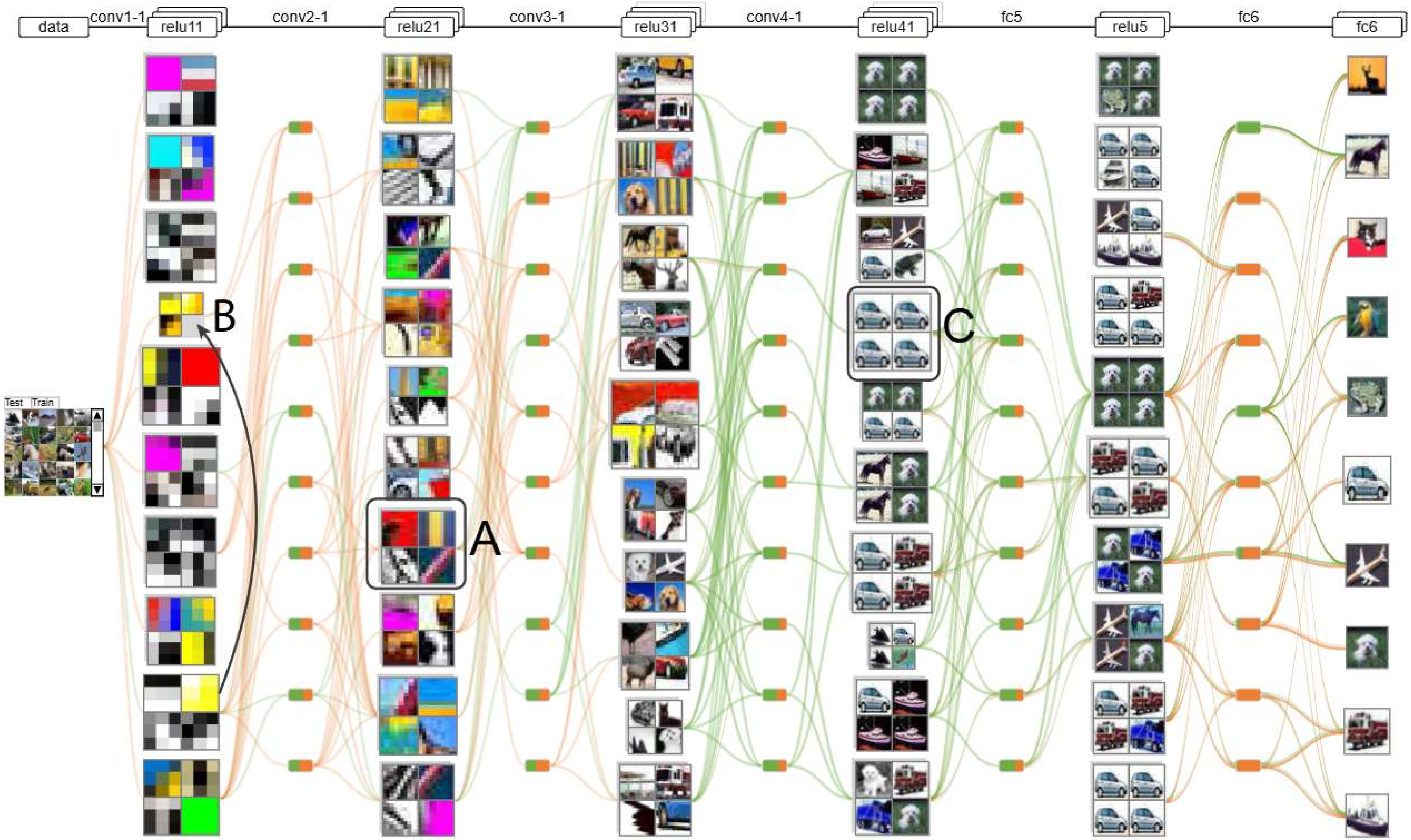
- Network-Centric Methods:
  - Exploring model structure,
  - visualising neuron activations,
  - and interpreting how different parts of the model contribute to the final output.
  - i.e. CNNVis for convolutional neural networks and LSTMVis for recurrent neural networks.
- Instance-Centric Methods:
  - Analysing individual instances and their relationships,
  - visualising the representation space learned by the model.
  - i.e. Rauber et al.
- Hybrid Methods:
  - Combining network-centric and instance-centric approaches
  - i.e. Summit and ActiVis
- Surrogate Model Explanations:
  - Using simpler, interpretable models to explain the behaviour of more complex models.
  - i.e. RuleMatrix and DeepVID

Model Understanding - Model Behaviours

- Network-Centric Methods Example: CNNVis



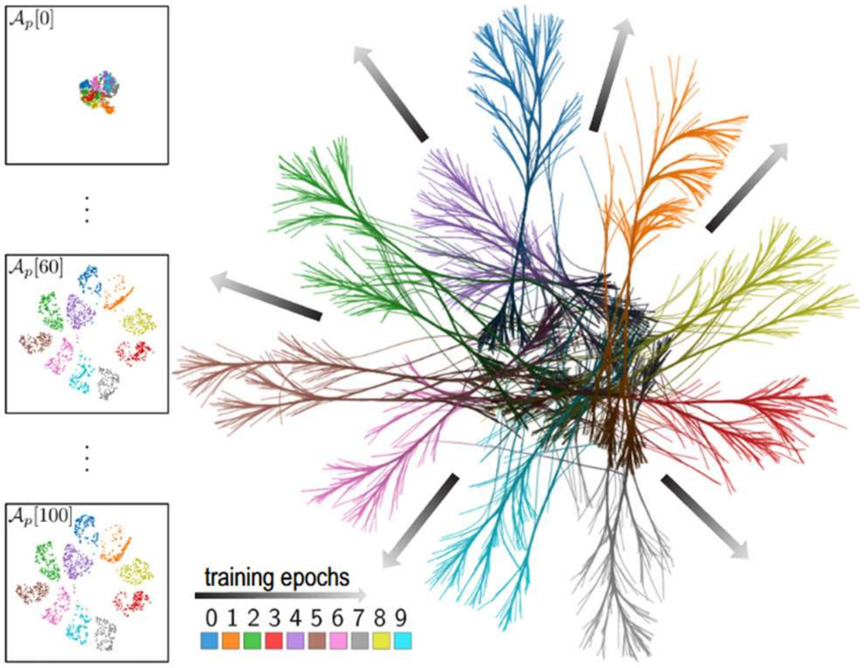
Towards Better Analysis of Deep Convolutional Neural Networks, https://doi.org/10.48550/arXiv.1604.07043



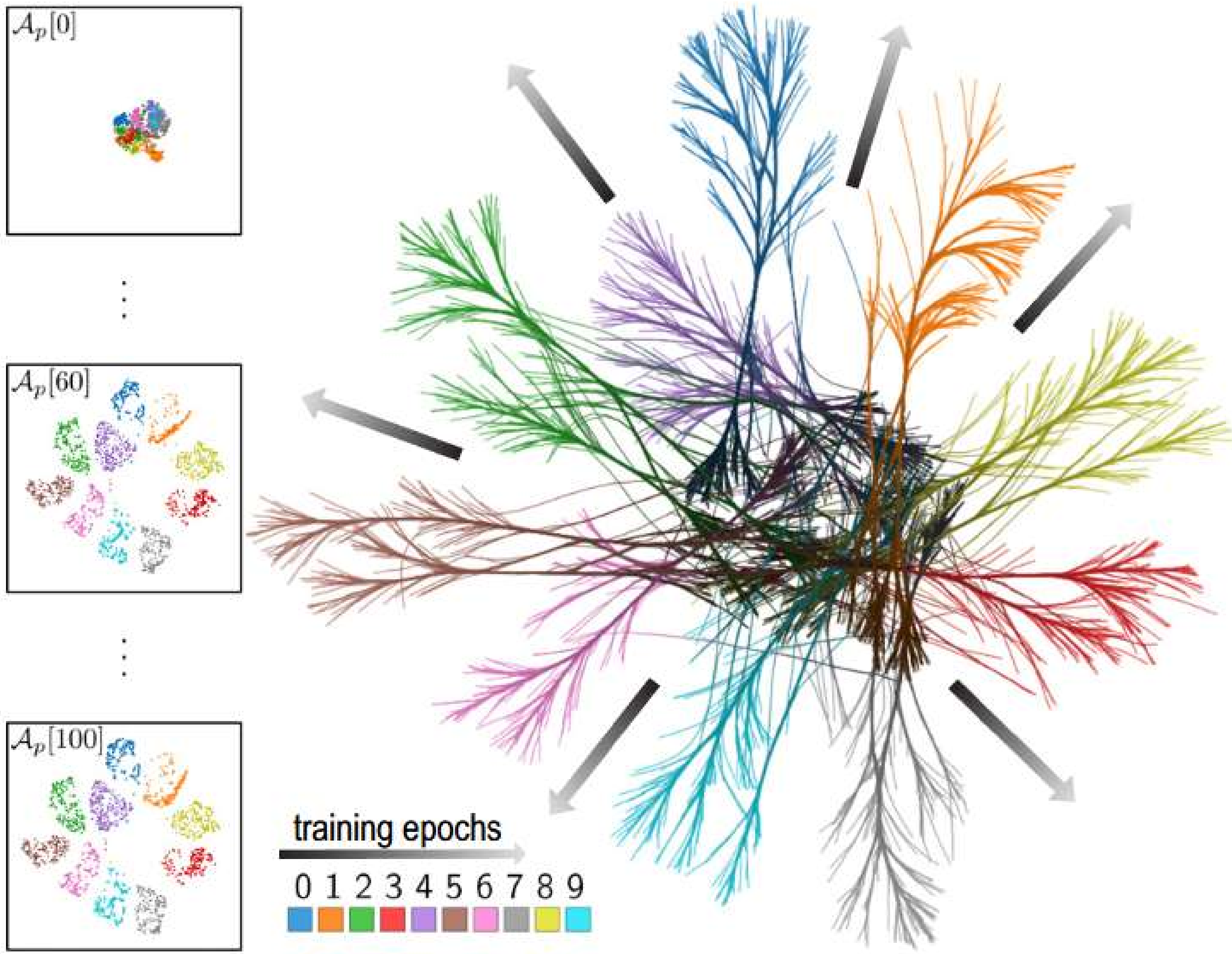
data
conv1-1
relu11
conv2-1
relu21
conv3-1
relu31
conv4-1
relu41
fc5
relu5
fc6
fc6
Test
Train
A
B
C

Model Understanding - Model Behaviours

- Instance-Centric Methods Example: Rauber et al.



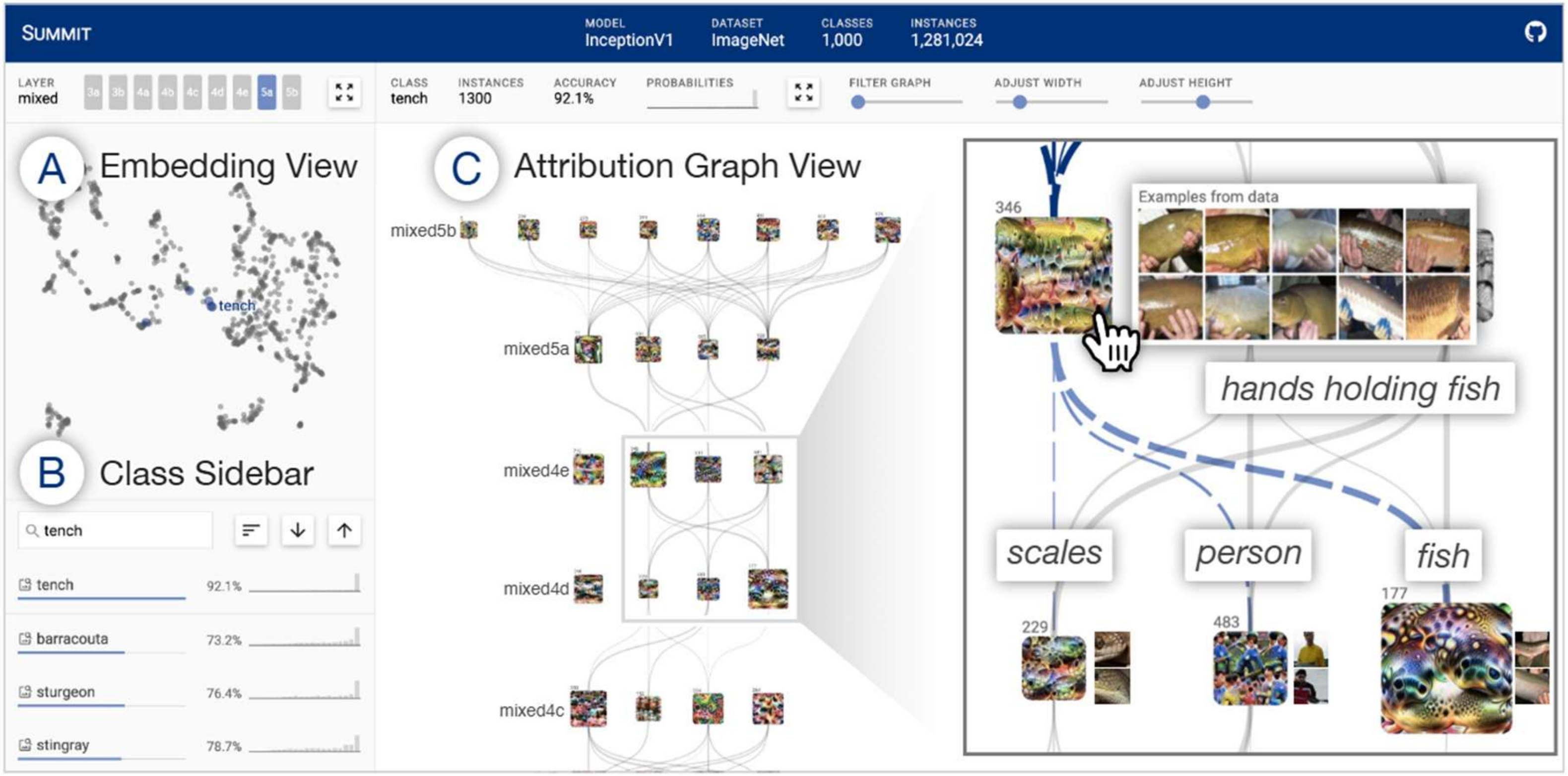
Visualizing the Hidden Activity of Artificial Neural Networks, doi: 10.1109/TVCG.2016.2598838.



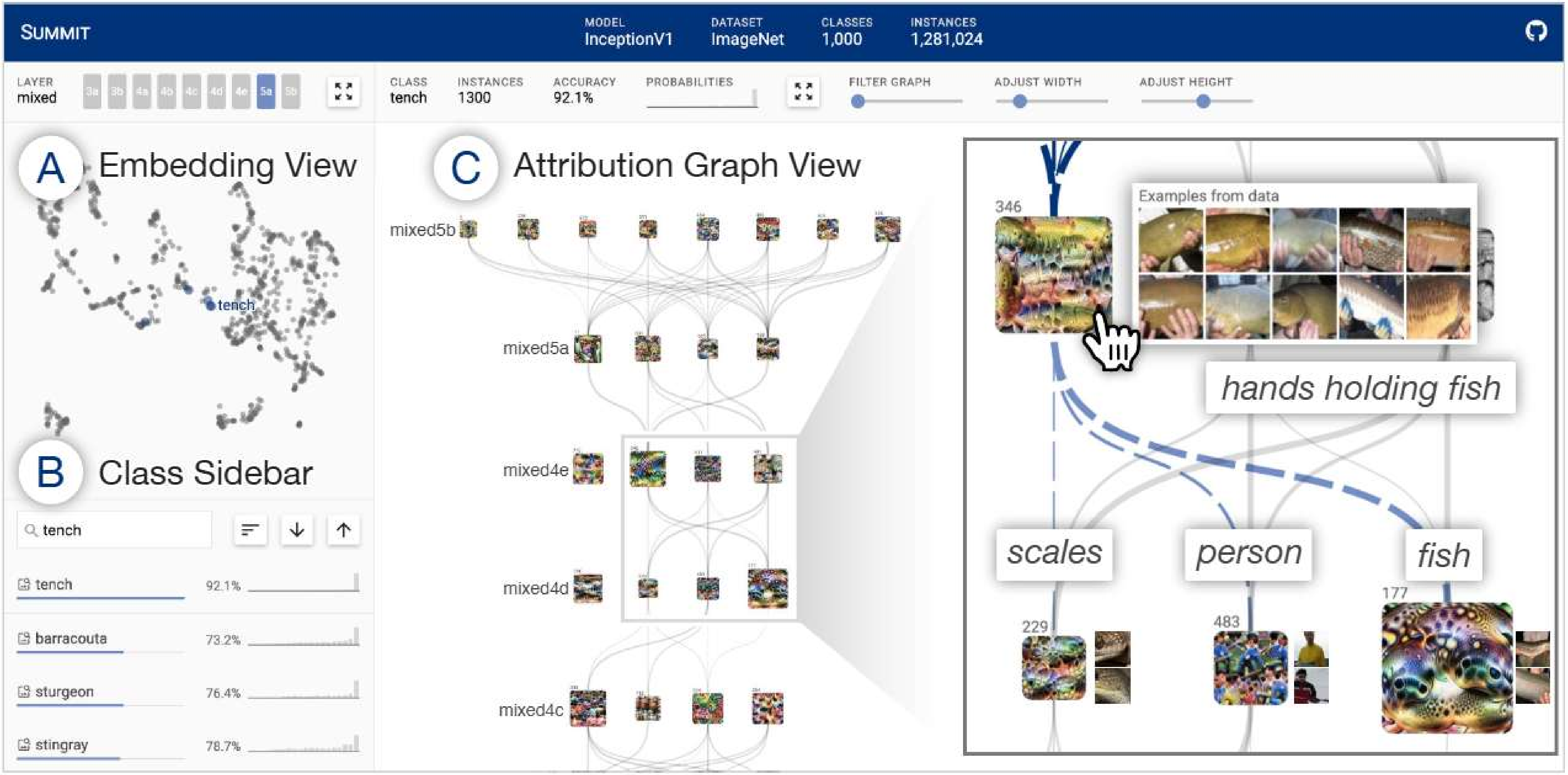
$\mathcal{A}_p[0]$
$\mathcal{A}_p[60]$
$\mathcal{A}_p[100]$
training epochs
0 1 2 3 4 5 6 7 8 9

Model Understanding - Model Behaviours

- Hybrid Methods Example: Summit



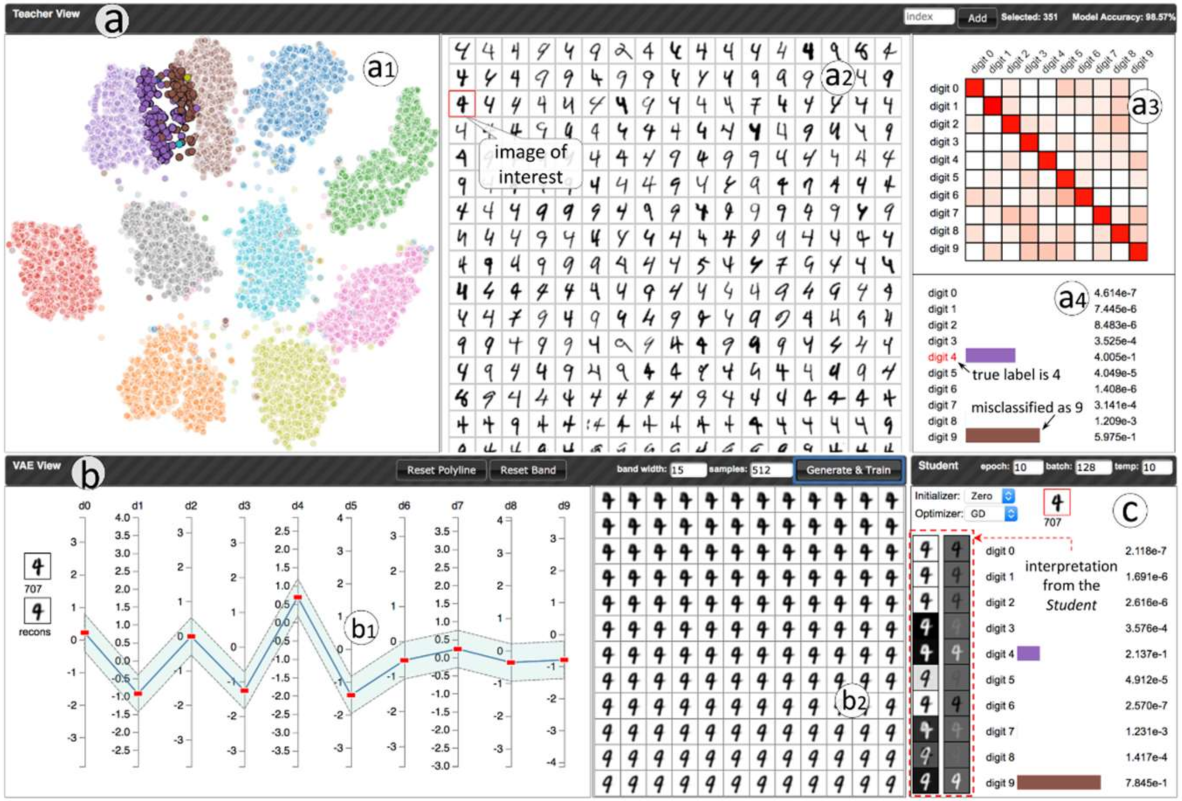
SUMMIT: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations, https://doi.org/10.48550/arXiv.1904.02323.



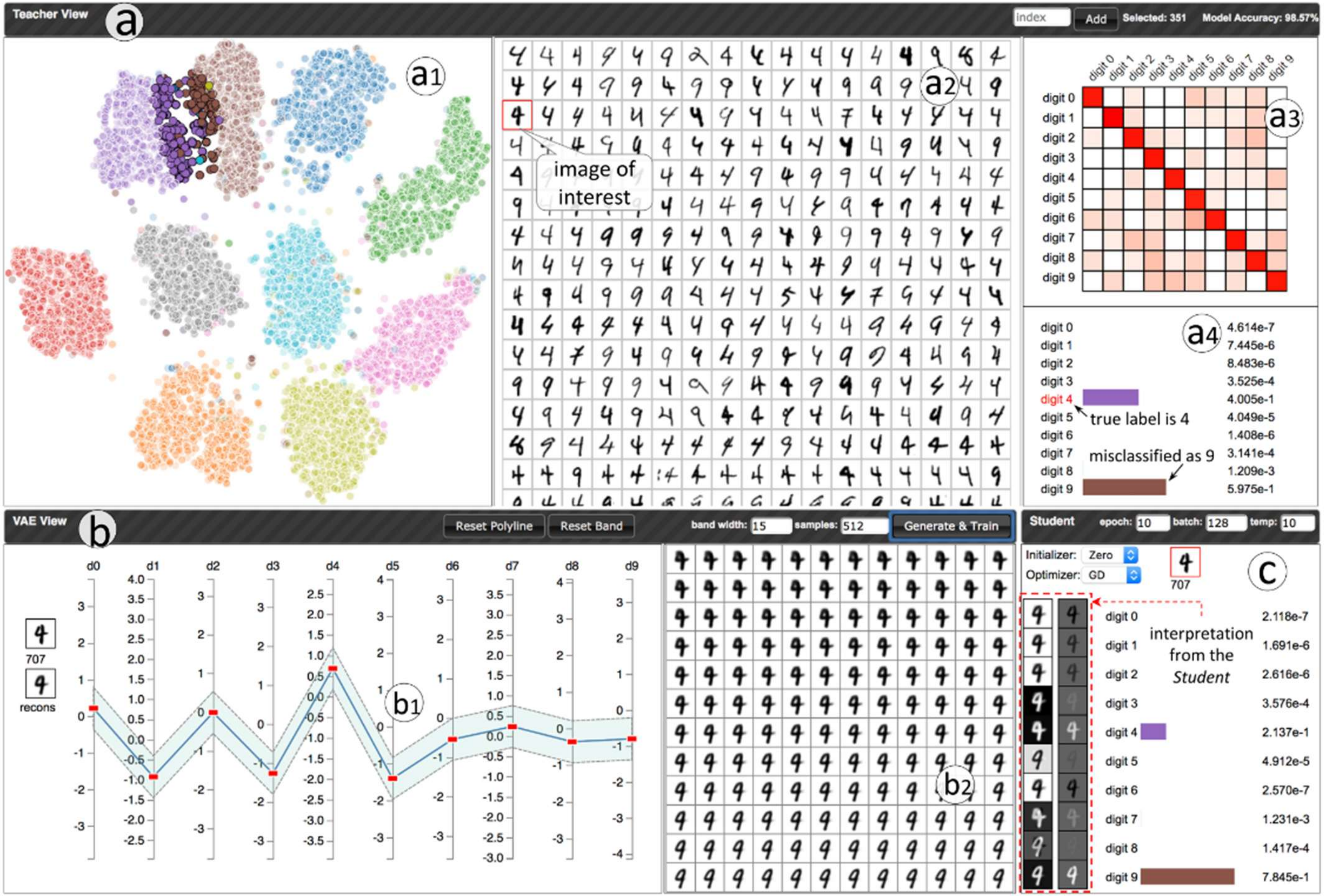
SUMMIT
MODEL
InceptionV1
DATASET
ImageNet
CLASSES
1,000
INSTANCES
1,281,024
LAYER
mixed
3a 3b 4a 4b 4c 4d 4e 5a 5b
CLASS
tench
INSTANCES
1300
ACCURACY
92.1%
PROBABILITIES
FILTER GRAPH
ADJUST WIDTH
ADJUST HEIGHT
A Embedding View
tench
B Class Sidebar
tench
tench 92.1%
barracouta 73.2%
sturgeon 76.4%
stingray 78.7%
C Attribution Graph View
mixed5b
mixed5a
mixed4e
mixed4d
mixed4c
346
Examples from data
hands holding fish
scales
person
fish
229
483
177

Model Understanding - Model Behaviours

- Surrogate Model Explanations Example: DeepVID



DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation, doi: 10.1109/TVCG.2019.2903943.



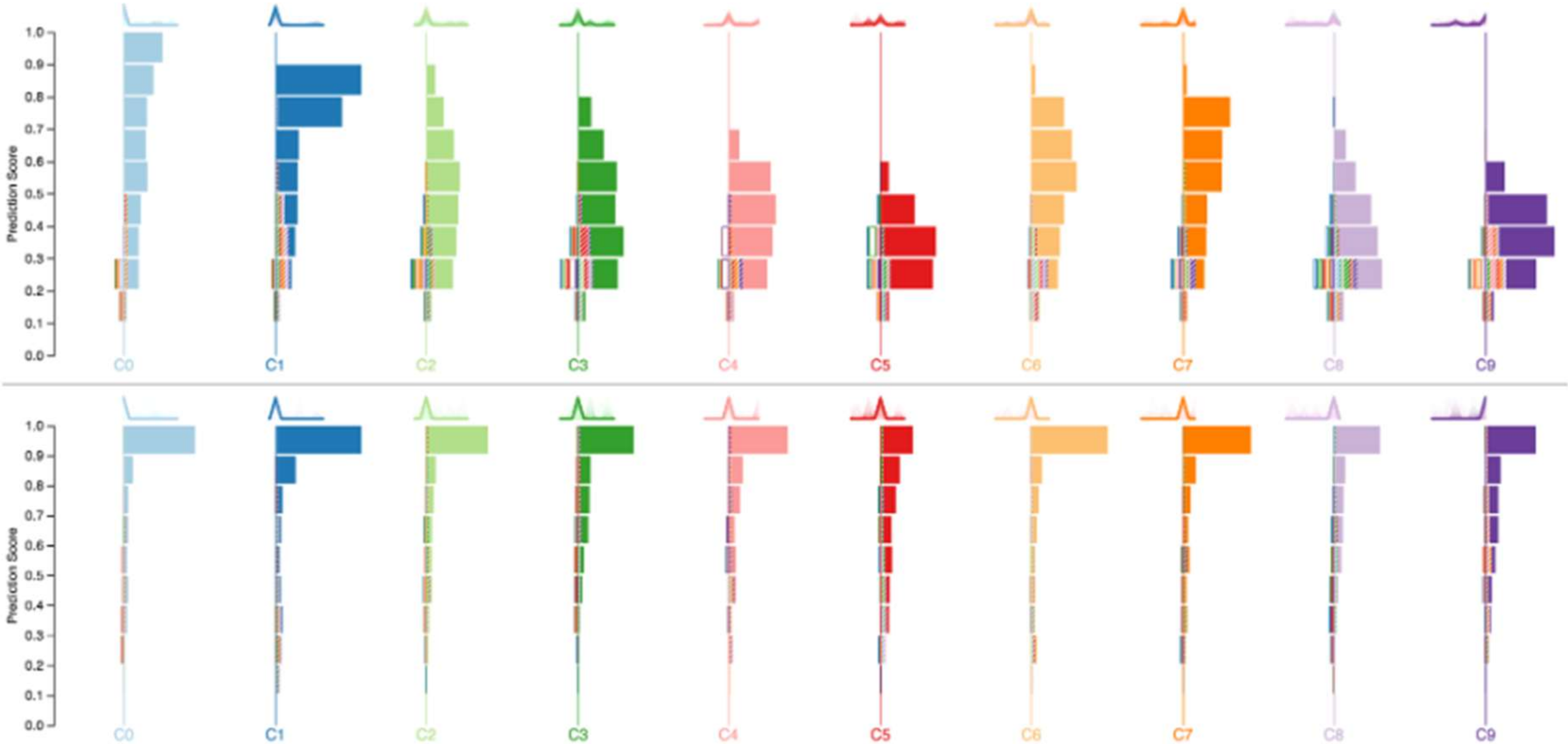
Teacher View
a
index
Add
Selected: 351
Model Accuracy: 98.57%
a1
a2
a3
a4
image of interest
digit 0
digit 1
digit 2
digit 3
digit 4
digit 5
digit 6
digit 7
digit 8
digit 9
4.614e-7
7.445e-6
8.483e-6
3.525e-4
4.005e-1
4.049e-5
1.408e-6
3.141e-4
1.209e-3
5.975e-1
true label is 4
misclassified as 9
VAE View
b
Reset Polyline
Reset Band
band width: 15
samples: 512
Generate & Train
Student
epoch: 10
batch: 128
temp: 10
b1
b2
d0 d1 d2 d3 d4 d5 d6 d7 d8 d9
707
recons
Initializer: Zero
Optimizer: GD
c
interpretation from the Student
2.118e-7
1.691e-6
2.616e-6
3.576e-4
2.137e-1
4.912e-5
2.570e-7
1.231e-3
1.417e-4
7.845e-1

Model Diagnosis

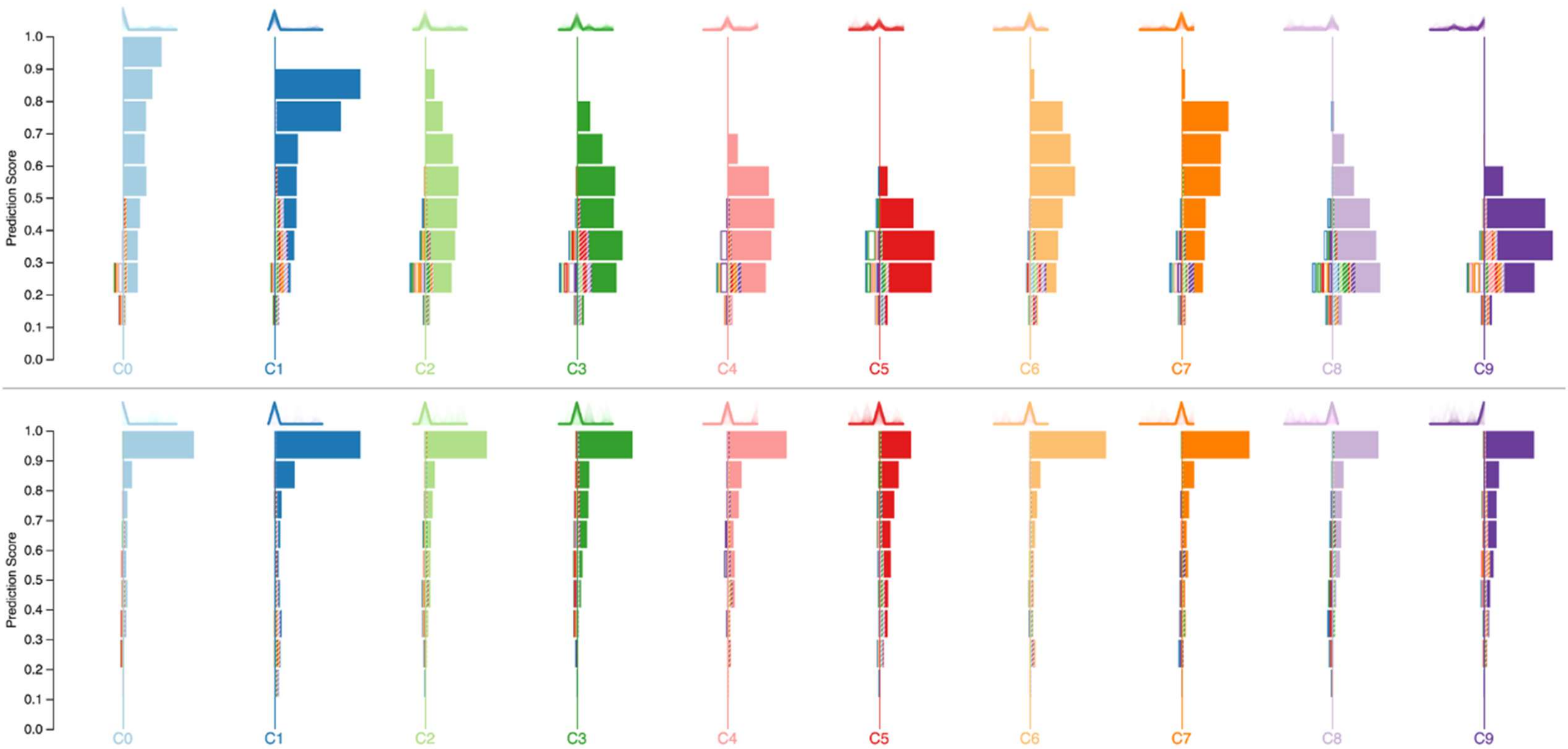
- Analysing Training Results:
  - Diagnose issues by visualising classifier performance,
  - identifying fairness issues,
  - exploring potential model vulnerabilities.
  - i.e. Squares and FairSight
- Analysing Training Dynamics:
  - Monitor the training process,
  - detecting anomalies,
  - understand the evolution of model behaviour over time.
  - i.e. DGMTracker and DQNViz

Model Diagnosis

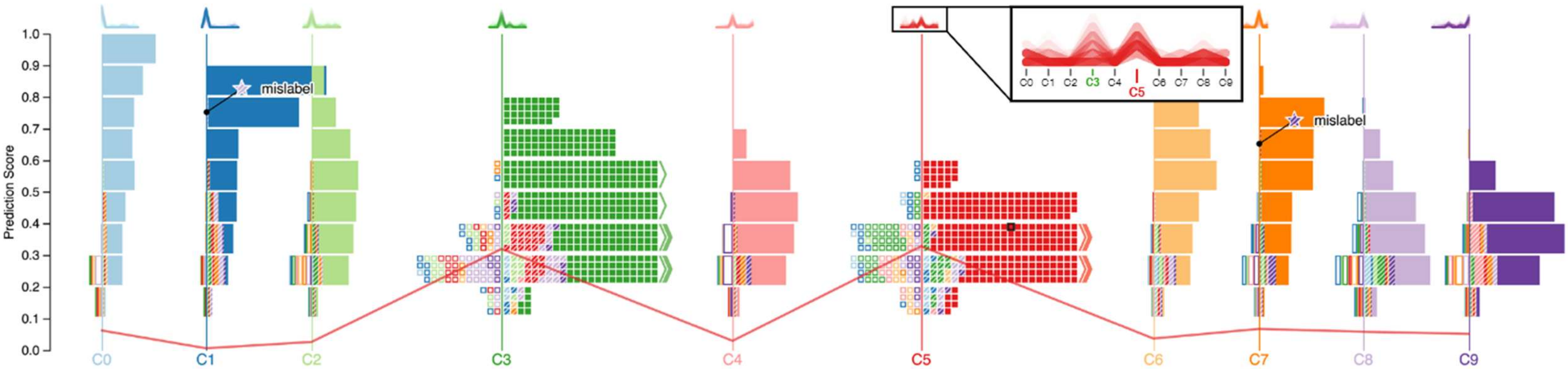
- Analysing Training Results: Squares



Squares: Supporting Interactive Performance Analysis for Multiclass Classifiers, doi: 10.1109/TVCG.2016.2598828



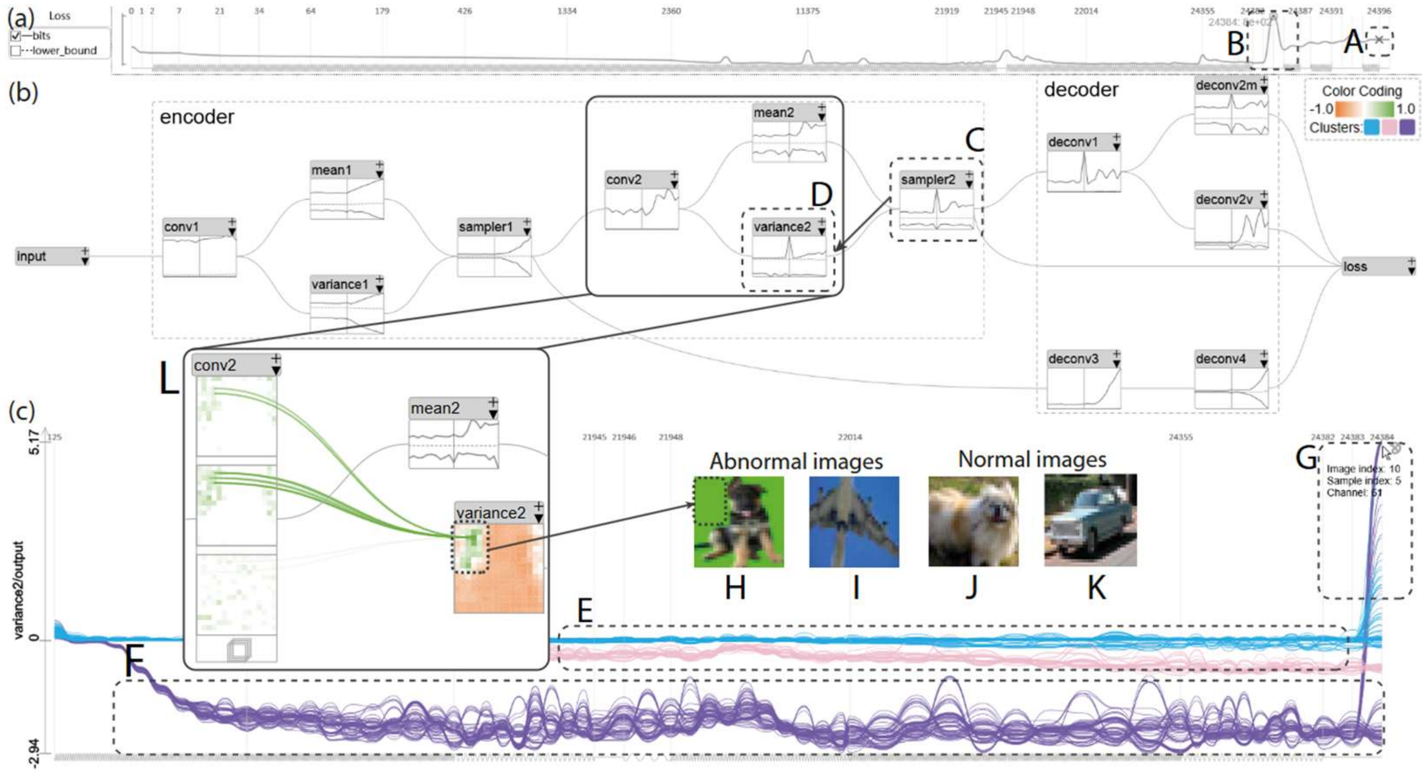
Prediction Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
C0
C1
C2
C3
C4
C5
C6
C7
C8
C9
Prediction Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
C0
C1
C2
C3
C4
C5
C6
C7
C8
C9



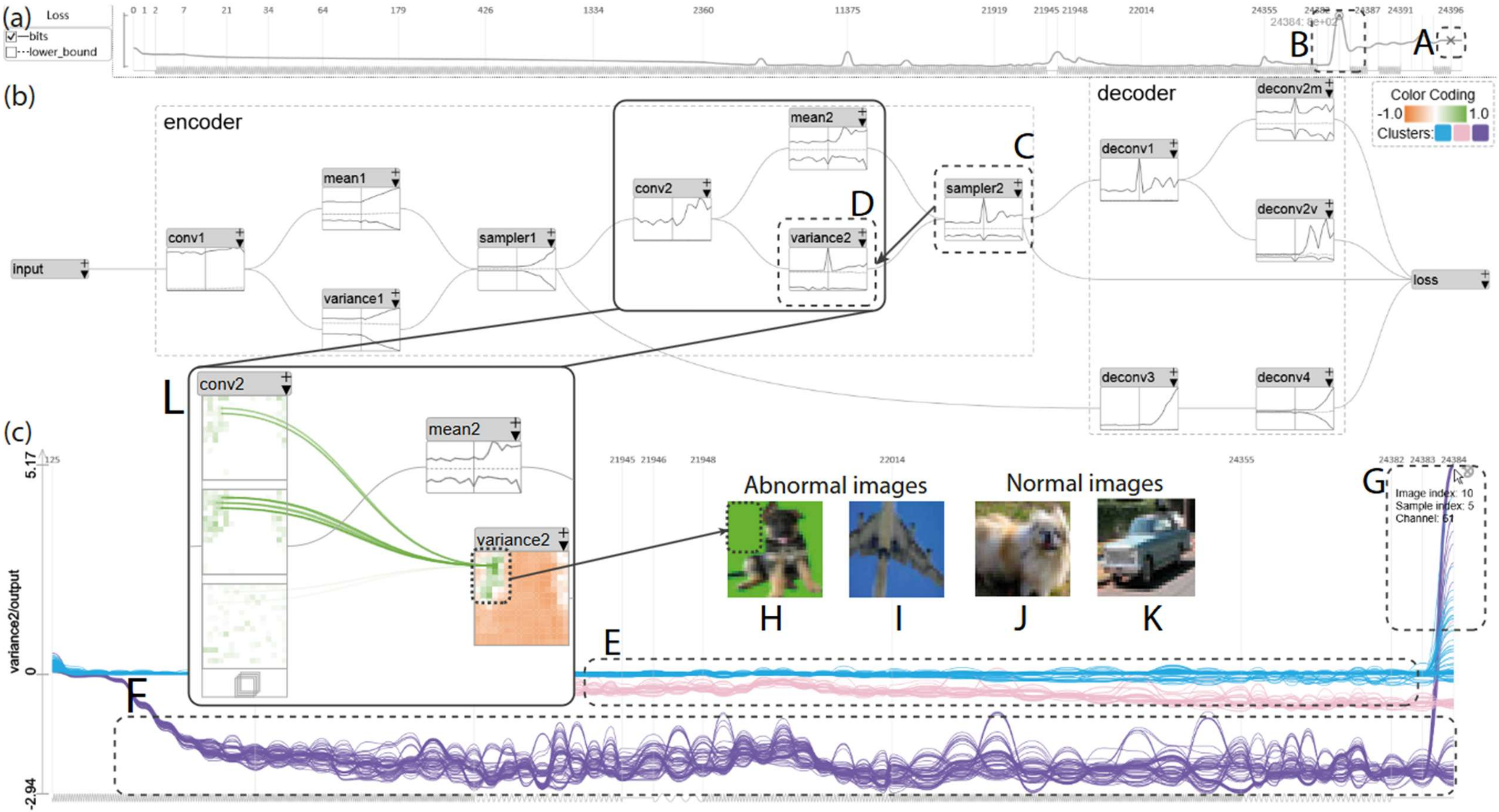
Prediction Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
mislabel
mislabel
C0
C1
C2
C3
C4
C5
C6
C7
C8
C9

Model Diagnosis

- Analysing Training Dynamics Example: DGMTracker



Analyzing the Training Processes of Deep Generative Models, doi: 10.1109/TVCG.2017.2744938.

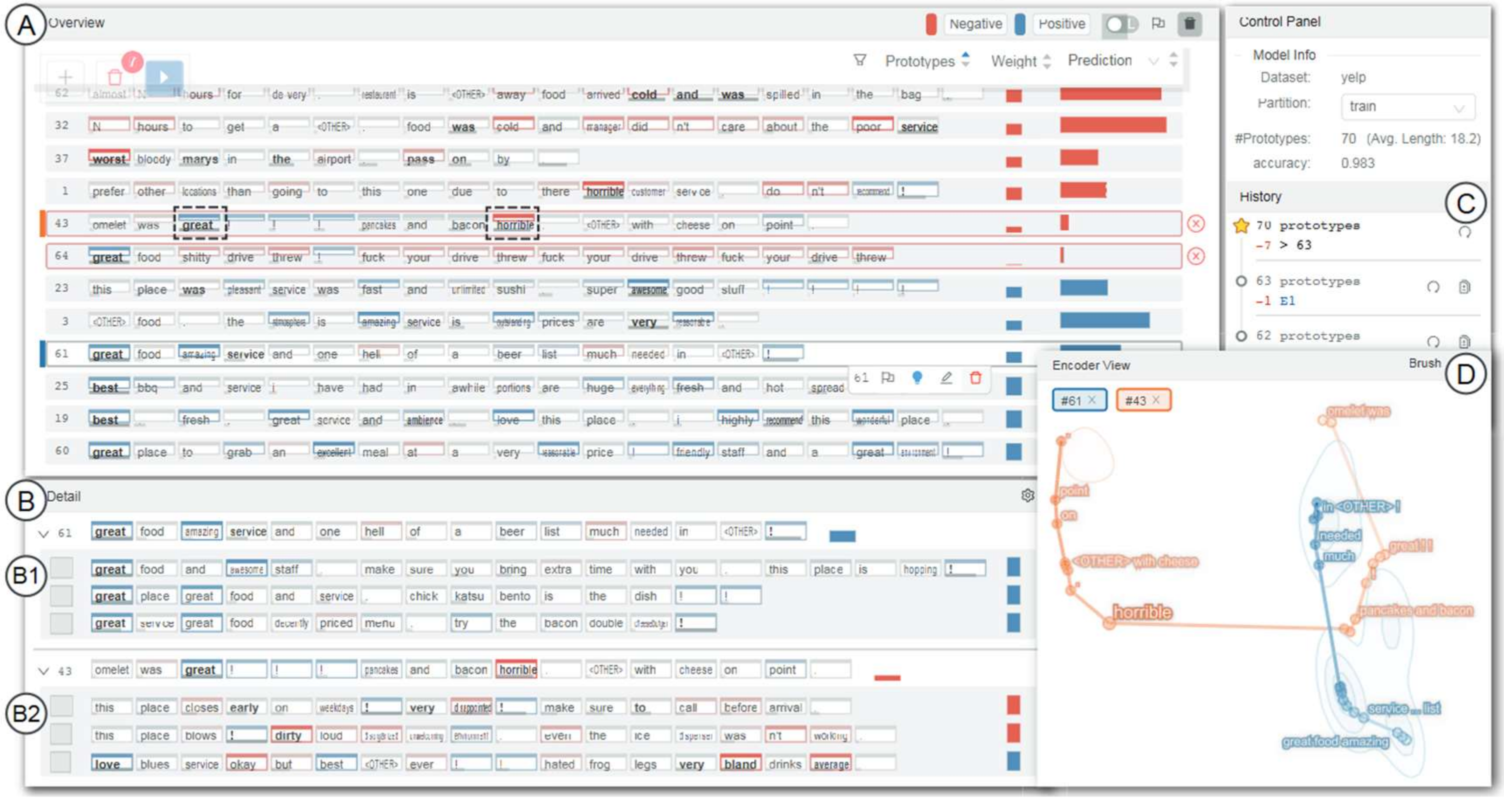

(a)
Loss
bits
lower_bound
A
B
(b)
encoder
conv1
mean1
variance1
sampler1
conv2
mean2
variance2
sampler2
input
C
D
decoder
deconv1
deconv2m
deconv2v
deconv3
deconv4
loss
Color Coding
-1.0
1.0
Clusters:
L
(c)
variance2/output
5.17
0
-2.94
Abnormal images
Normal images
E
F
G
H
I
J
K
Image index: 10
Sample index: 5
Channel: 61

Model Steering

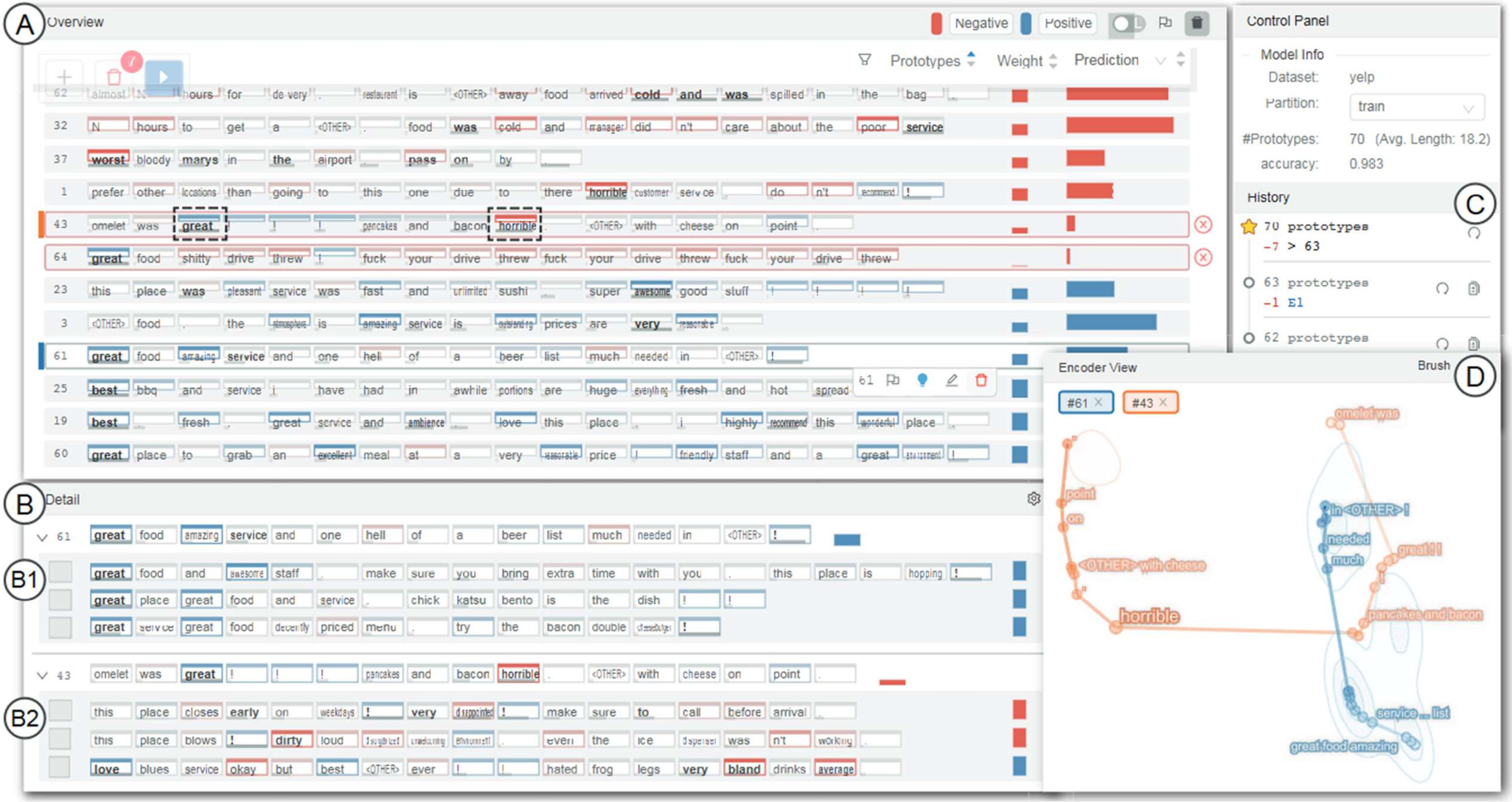
- Model Refinement with Human Knowledge:
  - Allowing users to interactively refine models by editing prototypes,
  - adding constraints,
  - correcting outputs.
  - i.e. ProtoSteer and ReVision
- Model Selection from Ensembles:
  - Comparing and selecting the best model from a set of candidate models.
  - i.e. BEAMES and RegressionExplorer

Model Steering

- Model Refinement with Human Knowledge Example: ProtoSteer



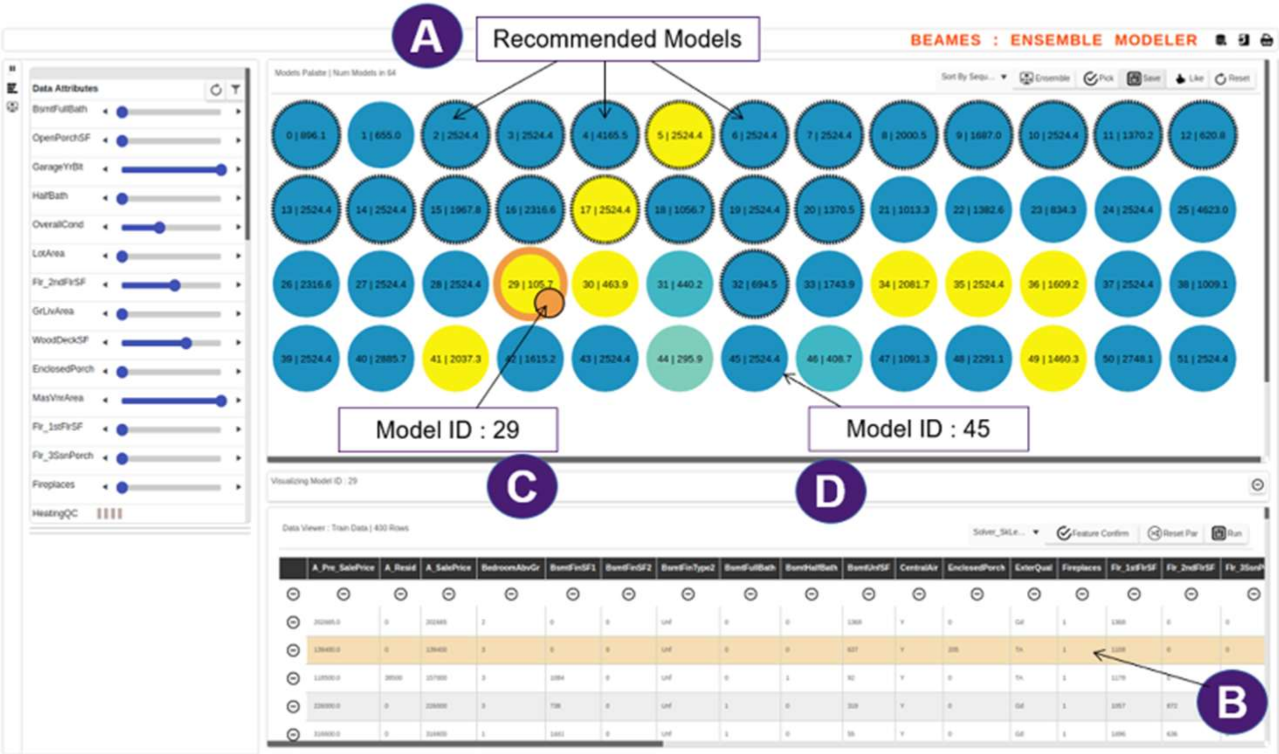
ProtoSteer: Steering Deep Sequence Model with Prototypes, doi: 10.1109/TVCG.2019.2934267.



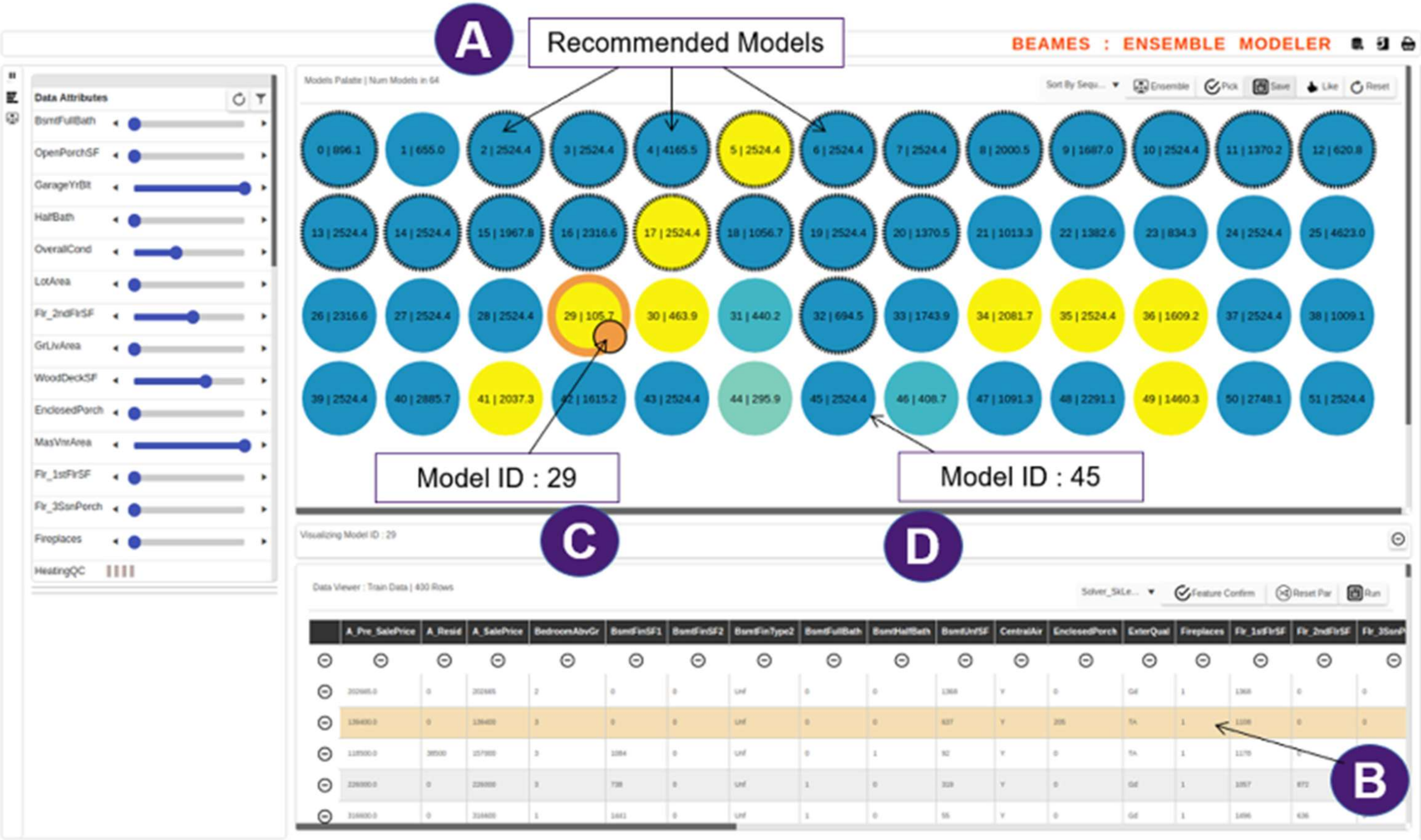
A
Overview
Negative
Positive
Prototypes
Weight
Prediction
B
Detail
B1
B2
C
Control Panel
Model Info
Dataset: yelp
Partition: train
#Prototypes: 70 (Avg. Length: 18.2)
accuracy: 0.983
History
70 prototypes
-7 > 63
63 prototypes
-1 E1
62 prototypes
D
Encoder View
Brush
#61
#43

Model Steering

- Model Selection from Ensembles Example: BEAMES



BEAMES: Interactive Multimodel Steering, Selection, and Inspection for Regression Tasks, doi: 10.1109/MCG.2019.2922592.


A
Recommended Models
BEAMES : ENSEMBLE MODELER
Data Attributes
BsmtFullBath
OpenPorchSF
GarageYrBlt
HalfBath
OverallCond
LotArea
Flr_2ndFlrSF
GrLivArea
WoodDeckSF
EnclosedPorch
MasVnrArea
Flr_1stFlrSF
Flr_3SsnPorch
Fireplaces
HeatingQC
Model ID : 29
Model ID : 45
C
D
B

Overview – Visual Analytics for Machine Learning

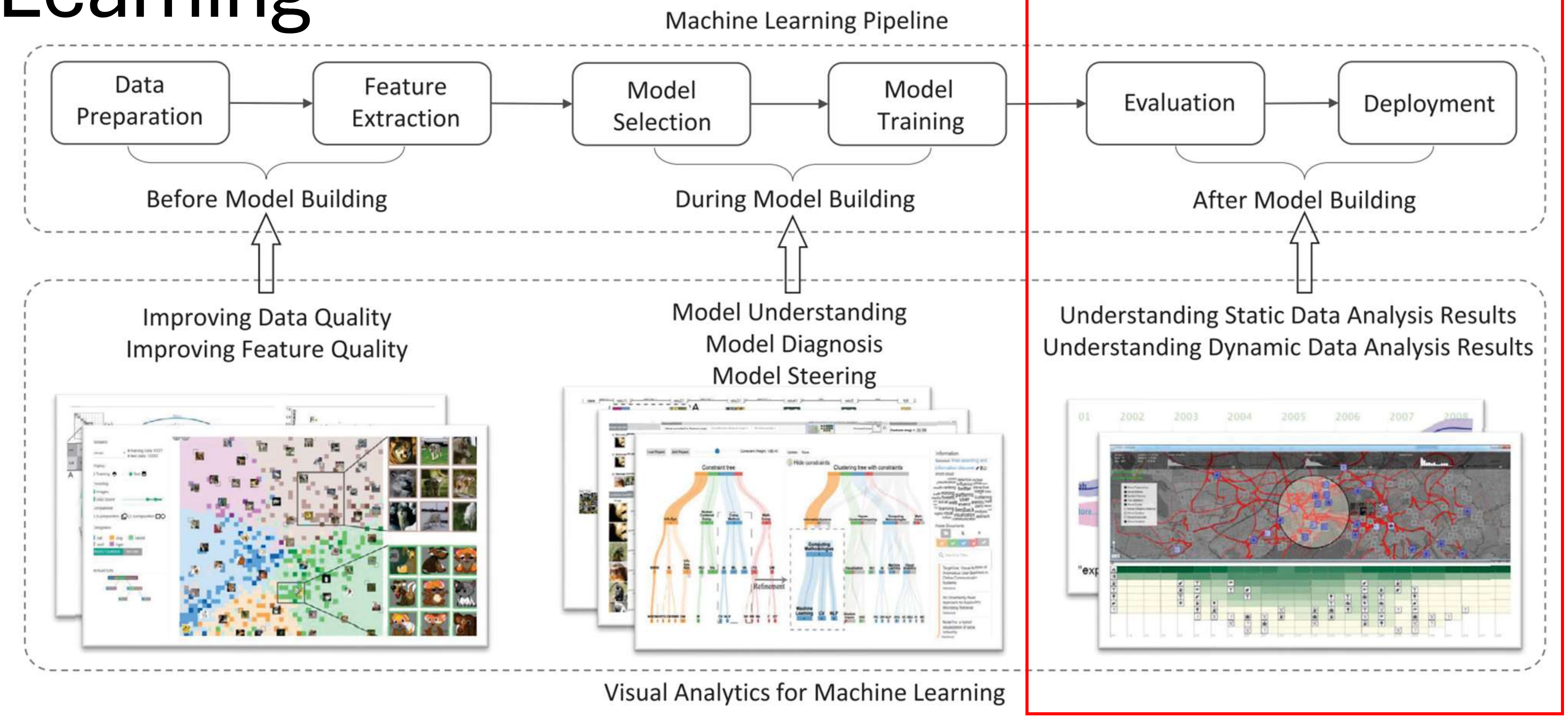


Fig. 1 An overview of visual analytics research for machine learning.

[Yuan]

Techniques **After** Model Building

Help users make sense of model outputs, gain insights from data analysis results, and evaluate model performance in real-world contexts.

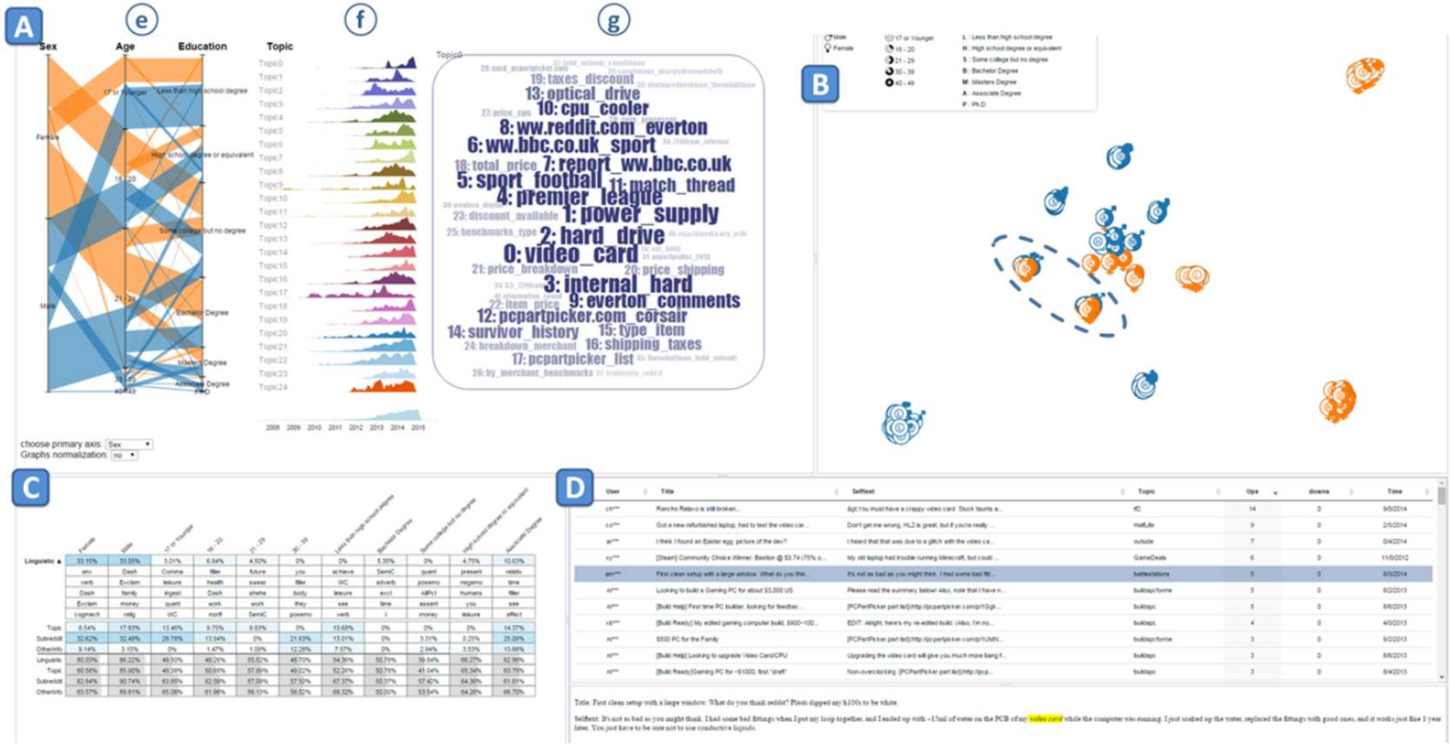
- Understanding Static Data Analysis Results
- Understanding Dynamic Data Analysis Results

Understanding Static Data Analysis Results

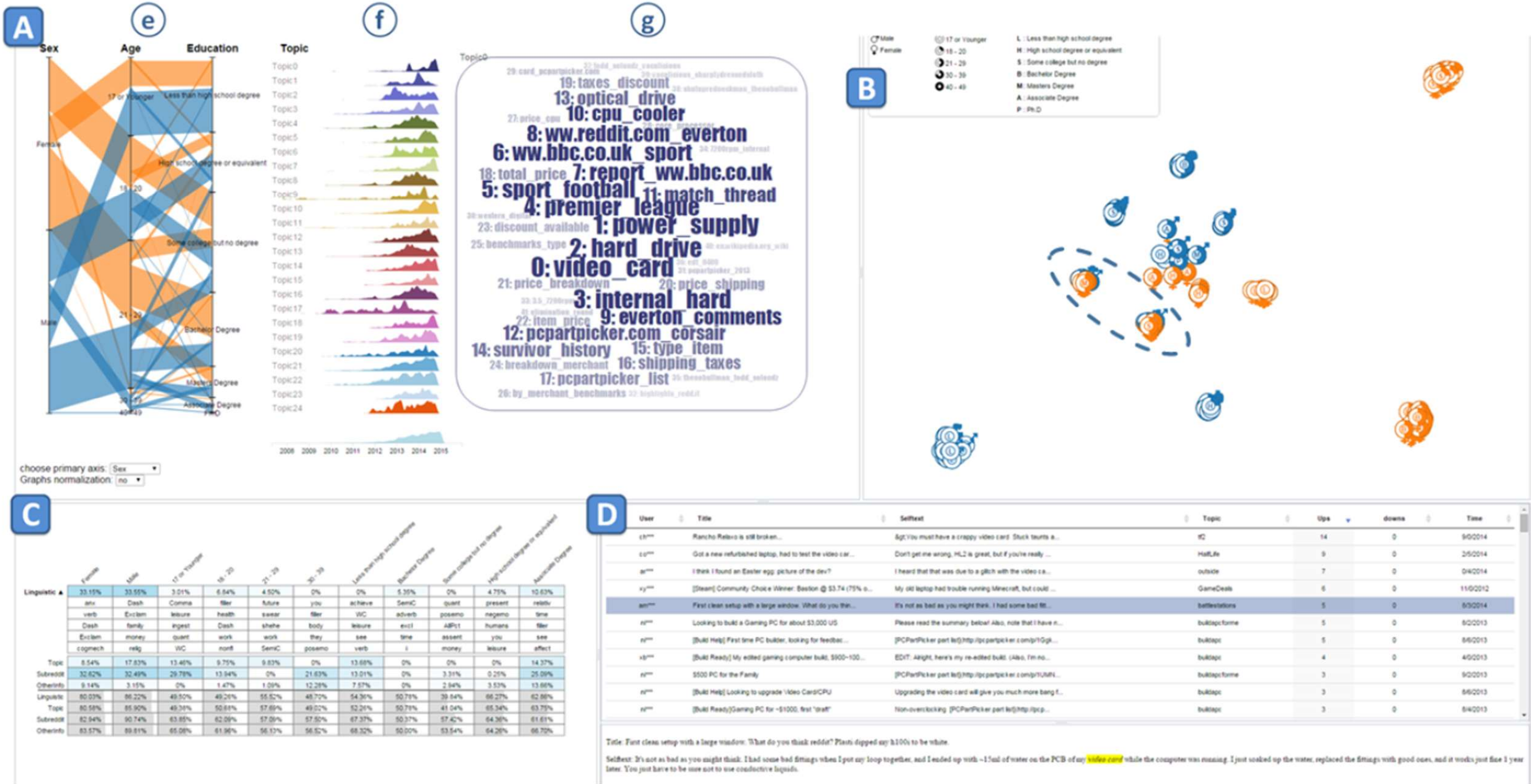
- Textual Data Analysis:
  - Visualizing topics, clusters, and relationships extracted from text data
  - used techniques:
    - topic modelling
    - word embedding
  - i.e. TopicPanorama , DemographicVis , and cite2vec.
- Other Data Analysis:
  - Extending visual analytics to other data types,
  - E.g., flow fields and multi-dimensional data,
  - used techniques:
    - subspace analysis
    - pattern matching
  - i.e. SMARTexplore

Understanding Static Data Analysis Results

- Textual Data Analysis : DemographicVis



DemographicVis: Analyzing demographic information based on user generated content, doi: 10.1109/VAST.2015.7347631.

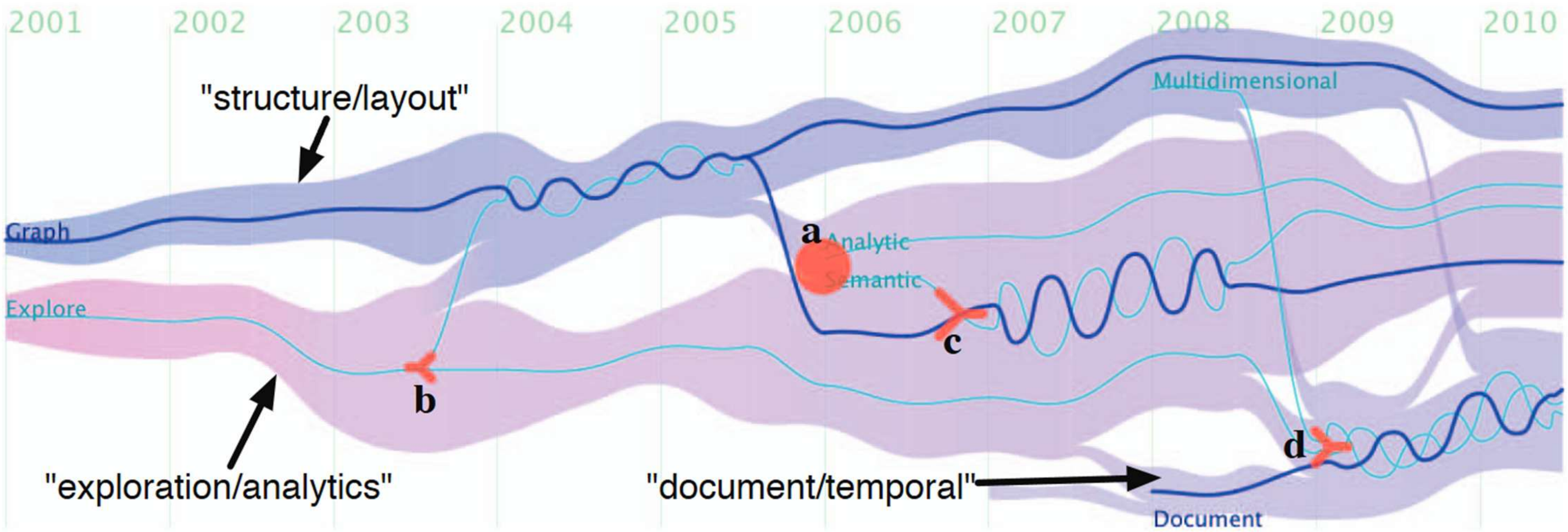

A
e
f
g
B
C
D
Sex
Age
Education
Topic
Female
Male
17 or Younger
Less than high school degree
High school degree or equivalent
Some college but no degree
Bachelor Degree
Masters Degree
Associate Degree
PhD
19: taxes_discount
13: optical_drive
10: cpu_cooler
8: ww.reddit.com_everton
6: ww.bbc.co.uk_sport
7: report_ww.bbc.co.uk
18: total_price
5: sport_football
11: match_thread
4: premier_league
1: power_supply
23: discount_available
25: benchmarks_type
2: hard_drive
0: video_card
21: price_breakdown
20: price_shipping
3: internal_hard
9: everton_comments
22: item_price
12: pcpartpicker.com_corsair
14: survivor_history
15: type_item
24: breakdown_merchant
16: shipping_taxes
17: pcpartpicker_list
26: by_merchant_benchmarks
choose primary axis: Sex
Graphs normalization: no
User
Title
Selftext
Topic
Ups
downs
Time

Understanding Dynamic Data Analysis Results

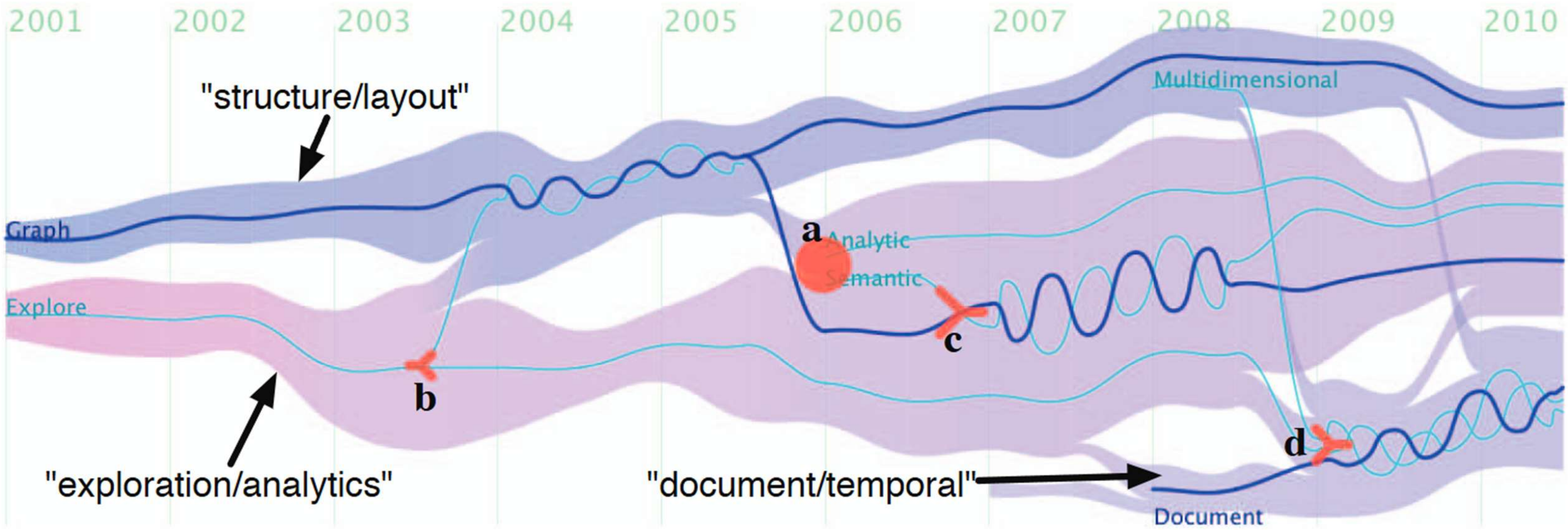
- Offline Analysis: Exploring patterns and trends in data over time, with all data available before analysis.
  - **Topic Analysis**: Visualizing topic evolution using techniques, which often employing a river metaphor to convey changes over time.
    - i.e. ThemeRiver and TextFlow
  - **Event Analysis**: Revealing important sequential patterns in event data, which leverage techniques like tensor decomposition and stage analysis.
    - i.e. EventThread
  - **Trajectory Analysis**: Visualizing and understanding movement patterns, often using techniques like clustering, pattern mining, and semantic enrichment.
    - i.e. Kruger et al. and Chen et al. .

Understanding Dynamic Data Analysis Results

- Offline Analysis Example: TextFlow



TextFlow: Towards Better Understanding of Evolving Topics in Text, doi: 10.1109/TVCG.2011.239.



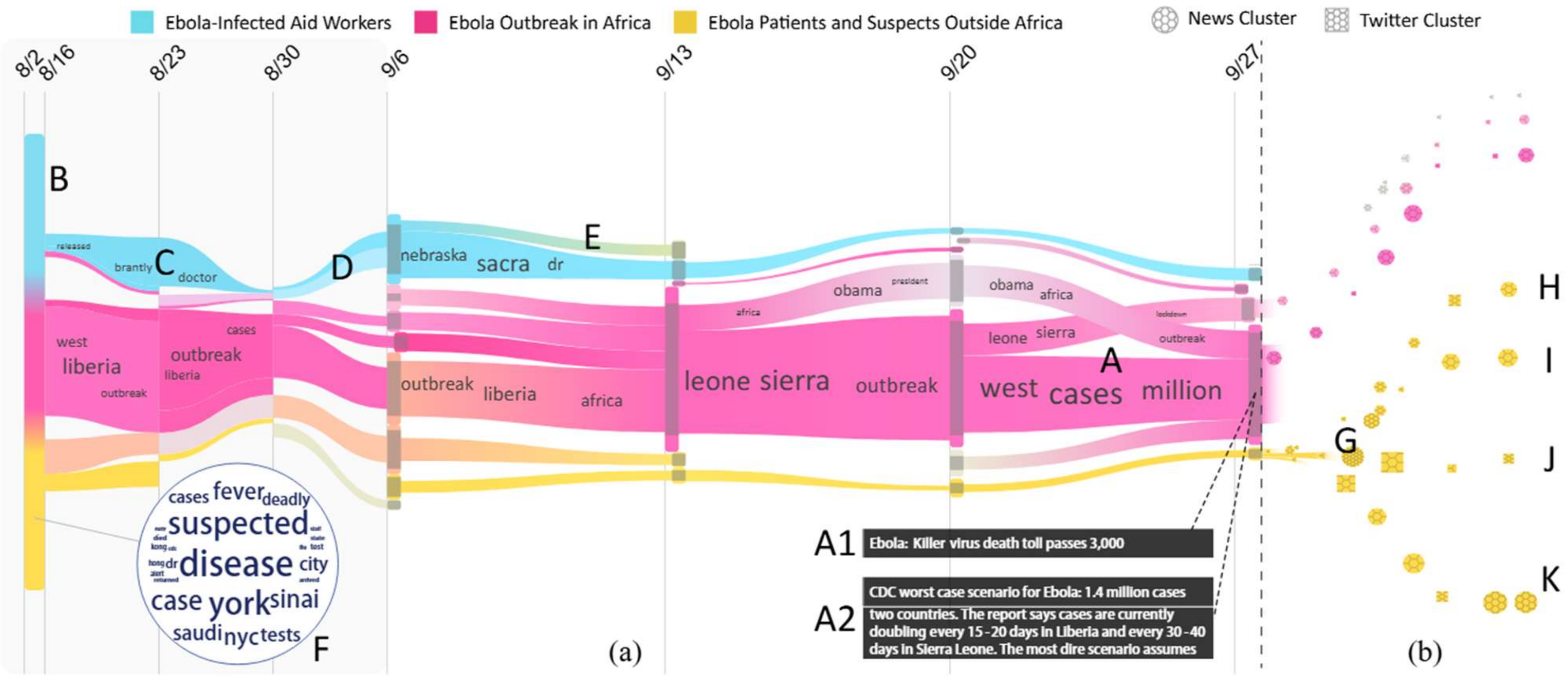
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
"structure/layout"
Multidimensional
Graph
a
Analytic
Semantic
Explore
c
b
d
"exploration/analytics"
"document/temporal"
Document

Understanding Dynamic Data Analysis Results

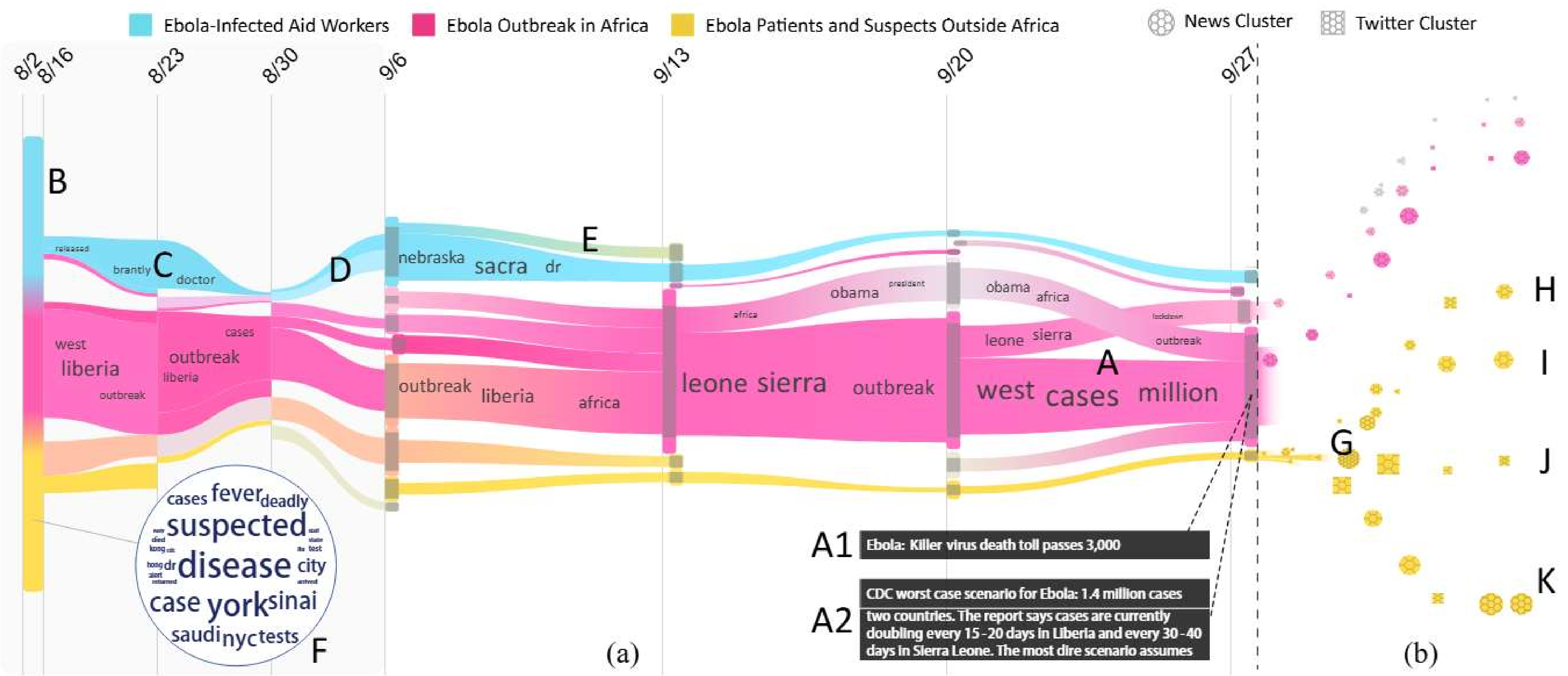
- Online Analysis:
  - Tackling streaming data where new data arrives continuously.
  - Area presents challenges for visualizing evolving patterns and integrating with real-time analysis algorithms.
    - i.e. TopicStream

Understanding Dynamic Data Analysis Results

- Online Analysis Example: TopicStream



Online Visual Analytics of Text Streams, doi: 10.1109/TVCG.2015.2509990.



Ebola-Infected Aid Workers
Ebola Outbreak in Africa
Ebola Patients and Suspects Outside Africa
News Cluster
Twitter Cluster
8/2
8/16
8/23
8/30
9/6
9/13
9/20
9/27
B
C
D
E
F
G
H
I
J
K
A
A1
Ebola: Killer virus death toll passes 3,000
A2
CDC worst case scenario for Ebola: 1.4 million cases
two countries. The report says cases are currently doubling every 15-20 days in Liberia and every 30-40 days in Sierra Leone. The most dire scenario assumes
(a)
(b)

Visual-based XAI (vXAI)

- Model usage
  - Feature-Based Methods
  - Rule-Based Methods
  - Propagation-Based Methods
  - Case-Based Methods
- Visual Approaches
  - Data representation
  - Local Explanations
  - Global Explanations

Feature-Based Methods

Identifying the Key Factors Driving AI's Decisions

- Methods identifying which input features were most influential in the model's decision.
- Methods like LIME and SHAP are commonly used for feature-based explanations.

Rule-Based Methods

Explaining Predictions with IF-THEN Rules

- Techniques expressing the model's logic as a set of human-readable rules.
- Bayesian Rule Lists (BRL) are often used to generate rule-based explanations.

Propagation-Based Methods

Tracing Information Flow in Neural Networks

- Techniques analysing how information flows through a network's layers to identify important features.
- Methods like Saliency Maps, LRP, and Integrated Gradients are used for propagation-based explanations.

Case-Based Methods

Learning from Similar Past Examples

- Methods finding past cases that are like the current input and show how the model treated those cases.
- Case-Based Reasoning (CBR) is a common approach for this type of explanation.

Data representation

- Known visualization techniques for data representation
  - Sankey Diagrams
  - Scatterplots
  - Table based visualizations
  - Etc. -> see Visualization lecture and previous lectures in Visual Analytics

Local vs. Global Explanations

Explaining Individual Predictions vs. Model Behaviour

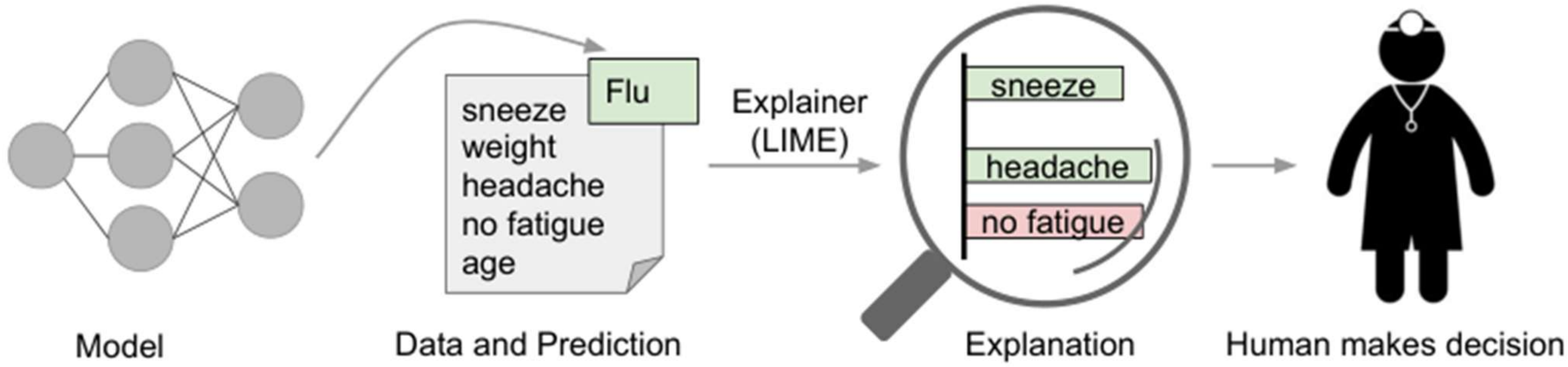
- **Local explanations**: focus on understanding a specific prediction.
- **Global explanations**: aim to provide insights into the overall model logic and behaviour.

XAI for Responsible AI

- XAI is crucial for ensuring that AI is developed and used responsibly.
- It promotes transparency, fairness, accountability, and ethical considerations in AI.

LIME - **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanation

- Explains predictions of any classifier
- Learn interpretable model locally around the prediction
- Model understanding depends on
    1. Trusting a prediction
    2. Trusting a model



[Ribeiro]

LIME – Desired Characteristics for Explainers

- Explanations must be **interpretable**
  - Provide qualitative understanding between the input variables and the response
  - Interpretability depends on the **target audience**
- To be meaningful an explanation must be **locally faithful**
  - Correspond to how the model behaves in the vicinity of the instance being predicted
- An Explainer should be model **agnostic**
- Provide global **perspective**
  - Establishing **trust** in the model

LIME – Goal

The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.

LIME – Interpretable Representation

Text classification:

- Feature:
  - Word Embedding
- Interpretable representation:
  - Binary vector indicating presence or absence of a word

Image classification:

- Feature:
  - Tensor with three color channels
- Interpretable representation:
  - Binary vector indicating the “presence” or “absence” of a super-pixel

LIME – Fidelity-Interpretablility Trade-off

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}\,(f, g, \pi_x) + \Omega(g)$$

LIME – Fidelity-Interpretablility Trade-off

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}\,(f, g, \pi_x) + \Omega(g)$$

$g \in G$, where G is a class of potentially **interpretable** models

- Linear models
- Decision trees

LIME – Fidelity-Interpretablility Trade-off

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}\left(f, g, \pi_x\right) + \Omega(g)$$

$\Omega(g)$ is a measure of **complexity**

- Linear model: number of non-zero weights
- Decision trees: depth of the tree

LIME – Fidelity-Interpretablility Trade-off

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}\,(f, g, \pi_x) + \Omega(g)$$

$f: R^d \to R$, **model** to be explained

- In classification $f(x)$ is the probability that x belongs to a certain class

LIME – Fidelity-Interpretablility Trade-off

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}\,(f, g, \pi_x) + \Omega(g)$$

$\pi_x(z)$, **proximity** measure between an instance z to x

- To define locality around x

LIME – Fidelity-Interpretablility Trade-off

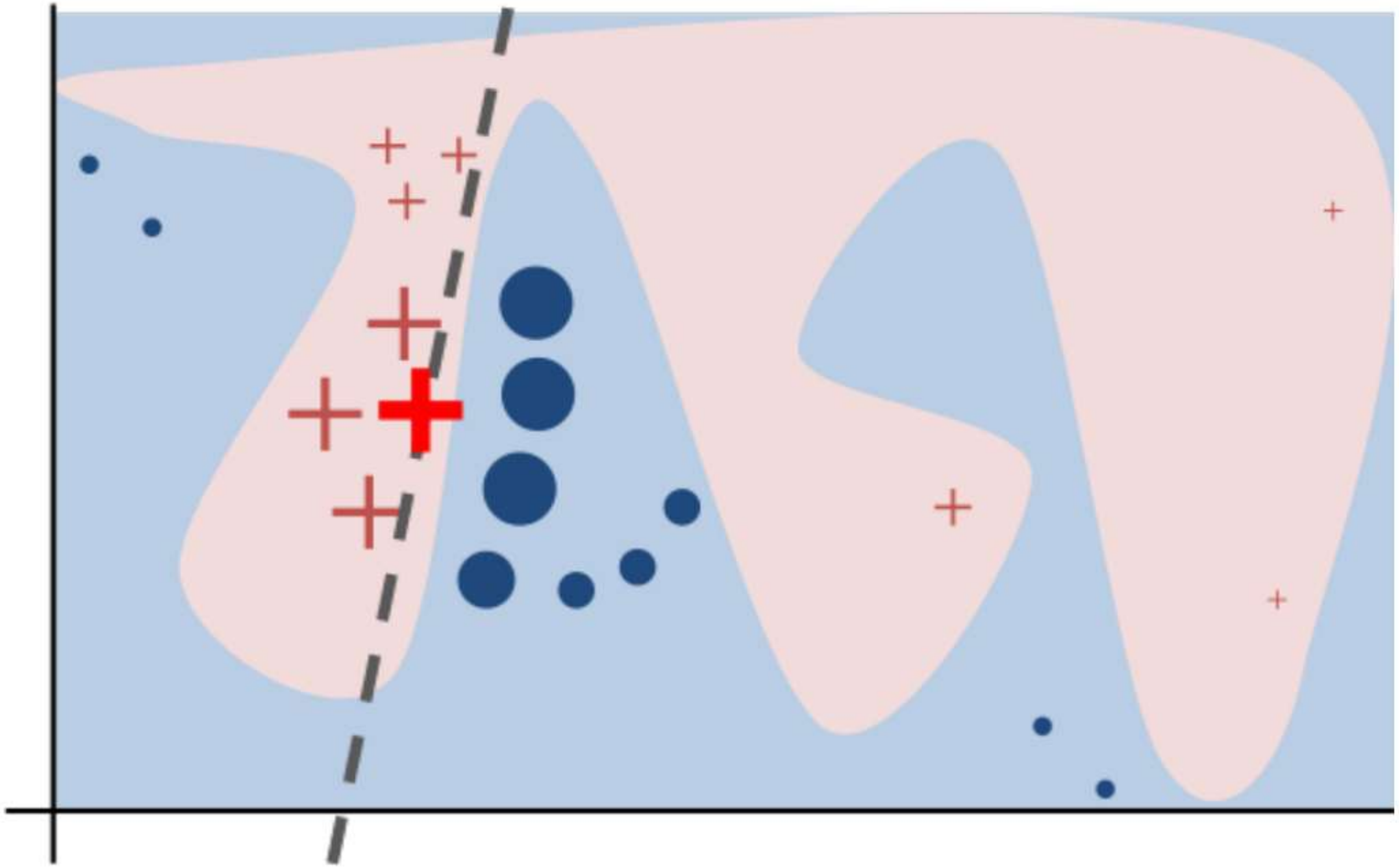
$$\xi(x) = \arg\min_{g \in G} \mathcal{L}\,(f, g, \pi_x) + \Omega(g)$$

$\mathcal{L}(f, g, \pi_x)$, measure of **unfaithfulness** of g in approximating f in locality defined by π_x

- **Minimize** $\mathcal{L}(f, g, \pi_x)$ while keeping $\Omega(g)$ low to ensure interpretability by humans

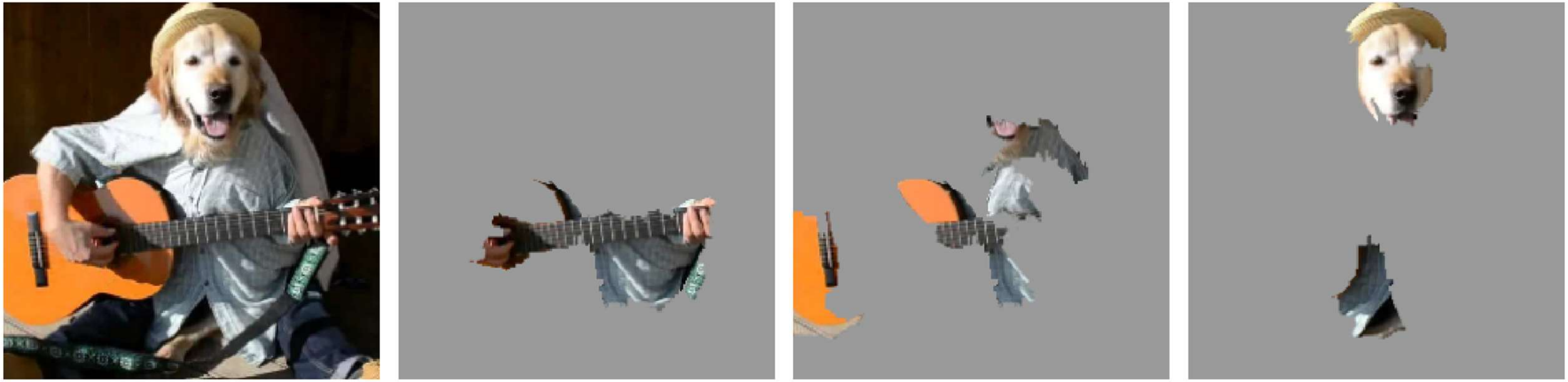
LIME - Sampling for Local Exploration

- Black box model f (pink/blue)
- Explain single instance (bold red plus) with linear model
- Dashed line is learned locally faithful explanation
- **NOT GLOBALLY FAITHFUL**



[Riberio]

LIME - Superpixels



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

[Riberio]

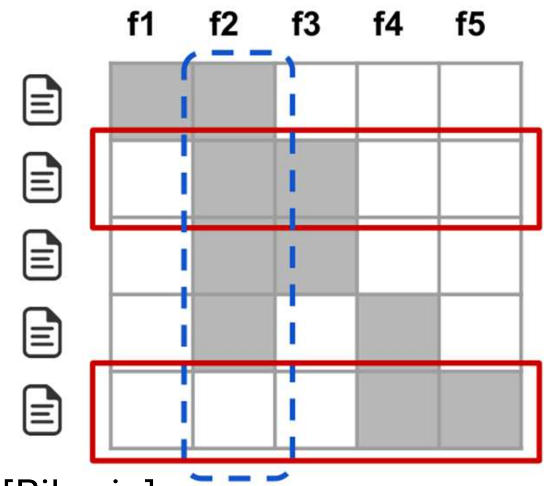
- Explanation of a single instance using superpixels

SP-LIME - Explaining Models

Explanation of a single prediction is not sufficient to evaluate and assess trust in the model as a whole

- give global understanding of the model by explaining **set** of individual instances
- Use **pick step**
  - task of selecting B instances from X for the user to inspect
  - B is the budged of the user (time/patience)
  - pick a diverse, representative set of explanations to show the user
  - construct **explanation matrix**

SP-LIME – Explanation Matrix



[Riberio]

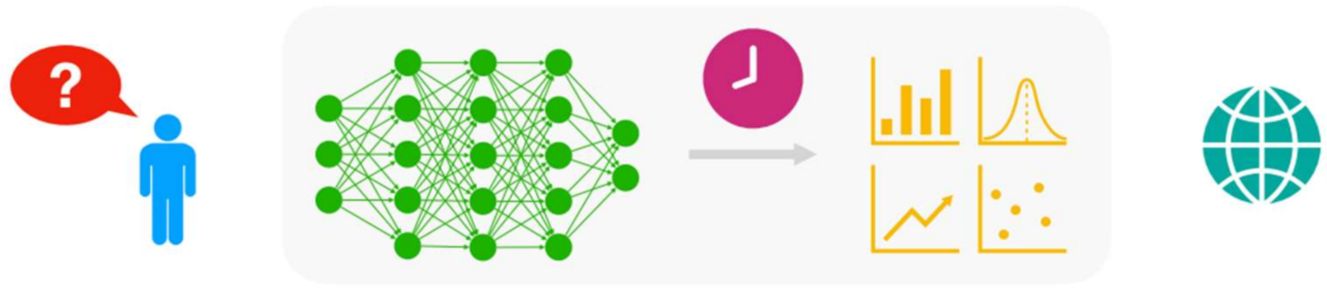
- Columns: features
- Rows: instances
- Cells contain local importance
- Global importance I
  - choose I such that features that explain many different instances have higher importance scores
- Pick problem consists of finding the set of Instances V that achieves the **highest coverage**

Important Questions

§4 WHY
Why would one want to use visualization in deep learning?

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

§8 WHEN
When in the deep learning process is visualization used?



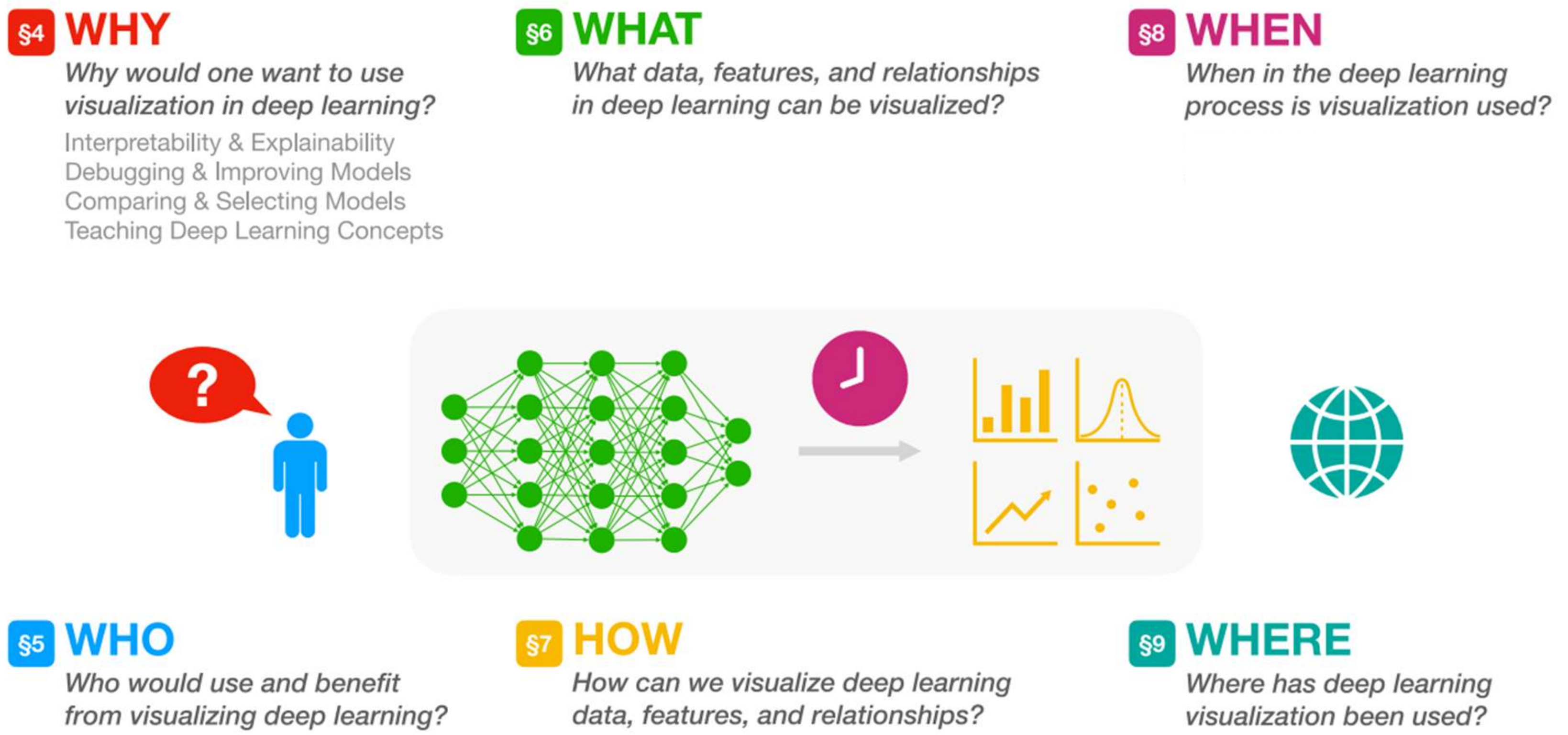
§5 WHO
Who would use and benefit from visualizing deep learning?

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

§9 WHERE
Where has deep learning visualization been used?

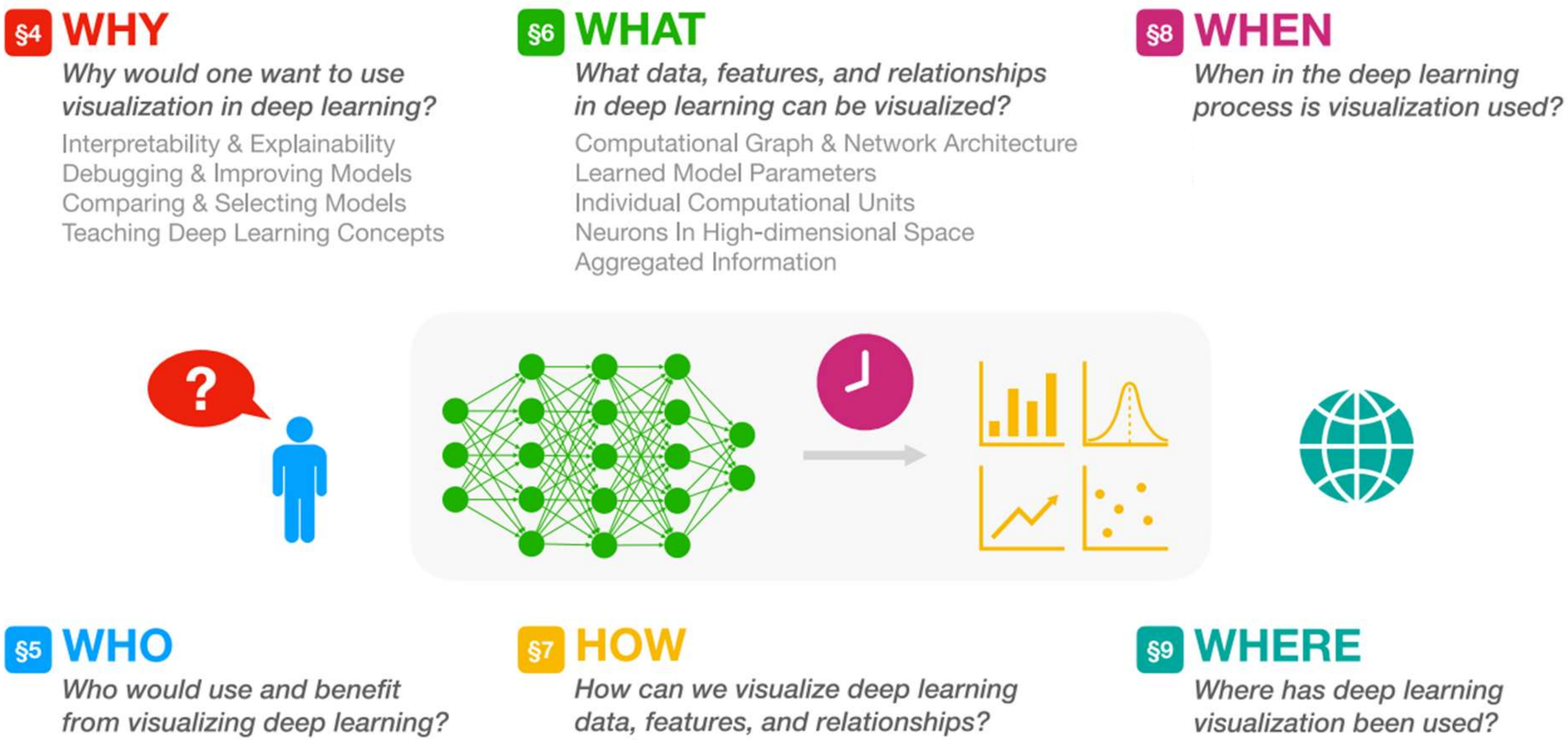
[Homann]

Important Questions



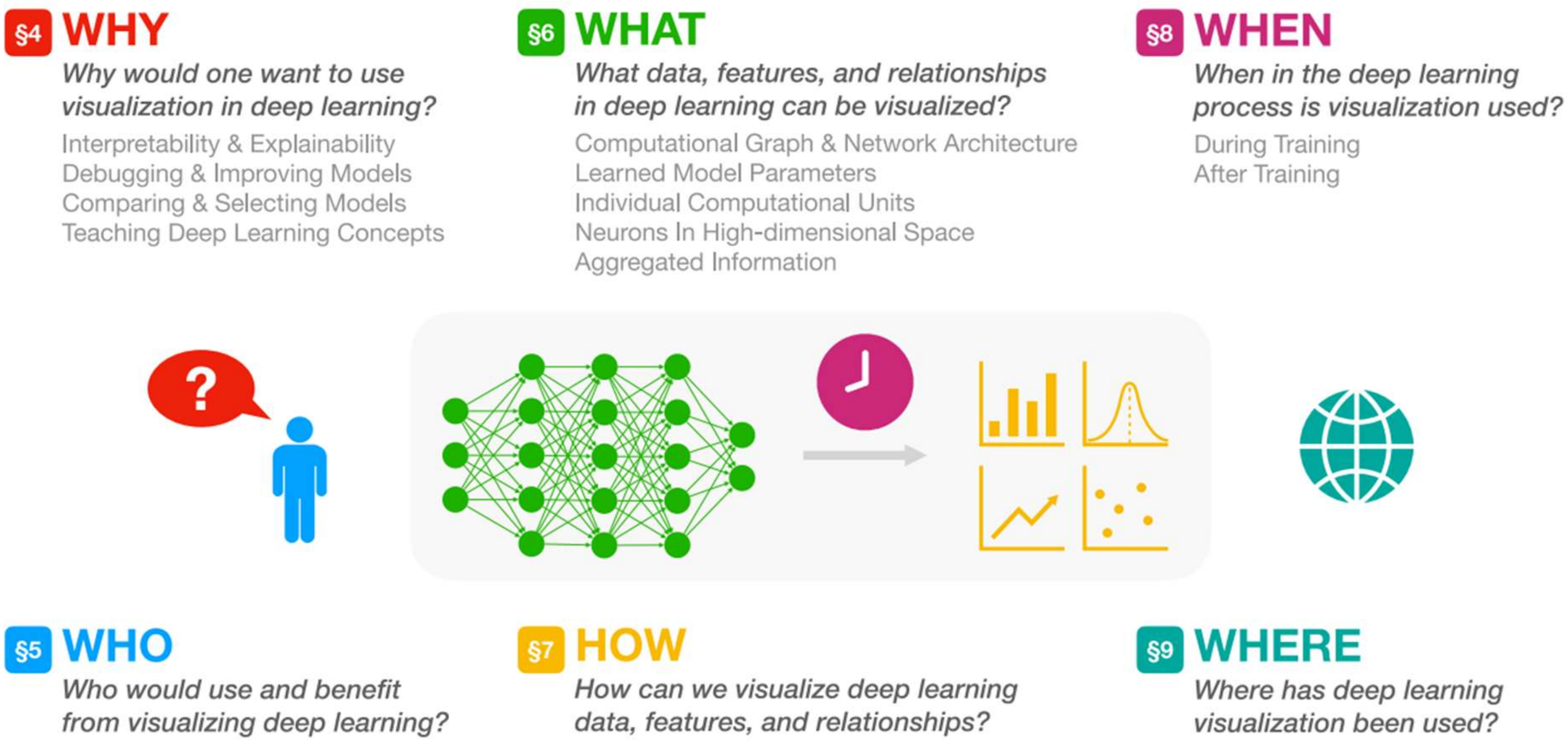
[Homann]

Important Questions



[Homann]

Important Questions



[Homann]

Important Questions

§4 WHY

Why would one want to use visualization in deep learning?

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

§6 WHAT

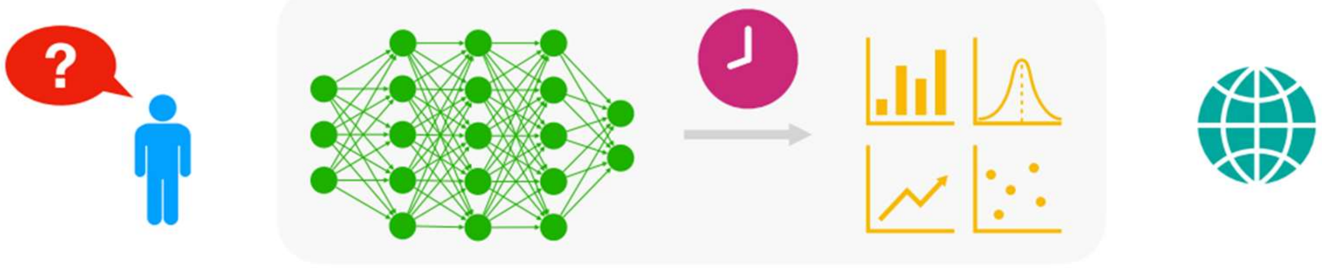
What data, features, and relationships in deep learning can be visualized?

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

§8 WHEN

When in the deep learning process is visualization used?

During Training
After Training



§5 WHO

Who would use and benefit from visualizing deep learning?

Model Developers & Builders
Model Users
Non-experts

§7 HOW

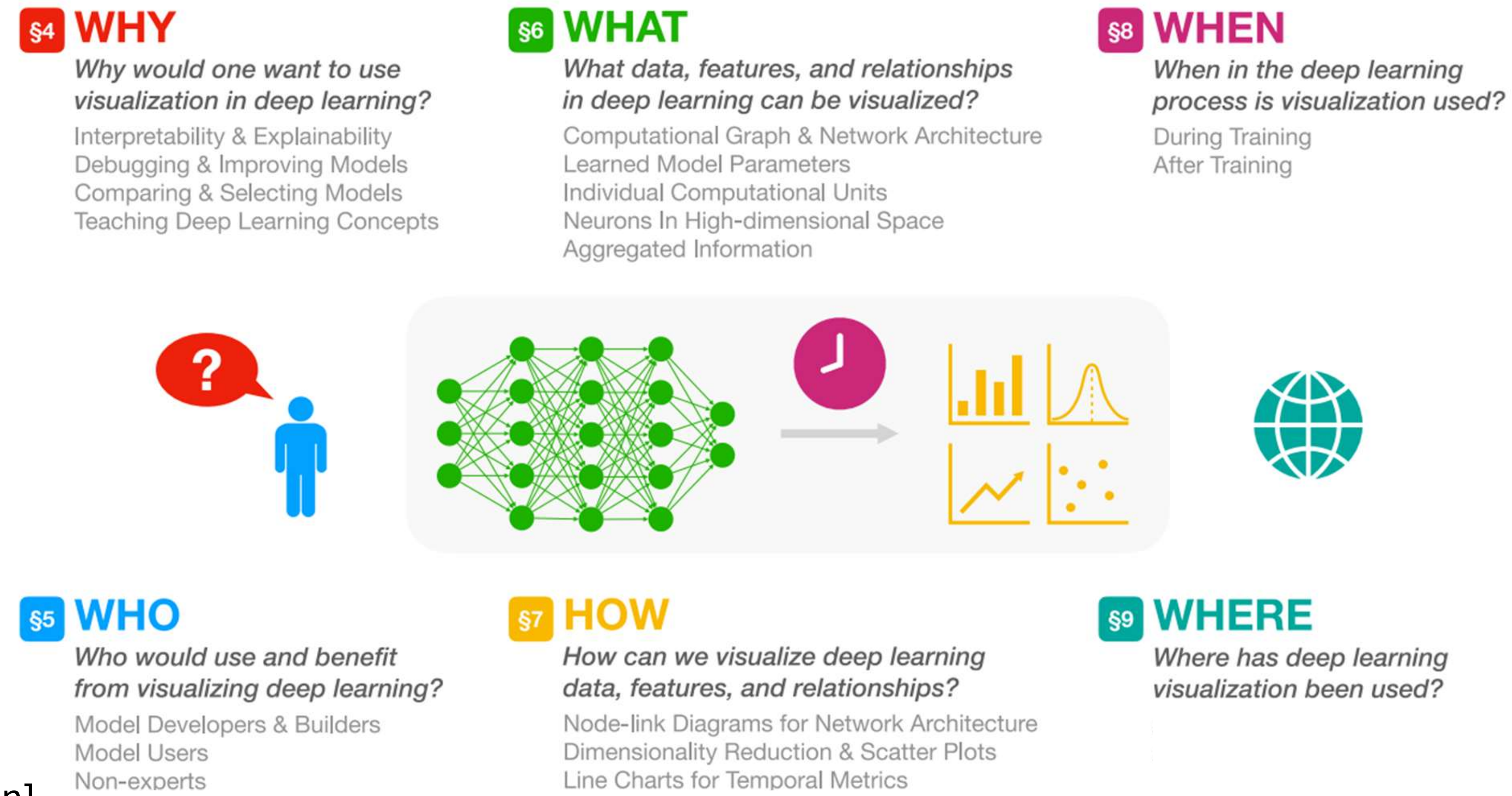
How can we visualize deep learning data, features, and relationships?

§9 WHERE

Where has deep learning visualization been used?

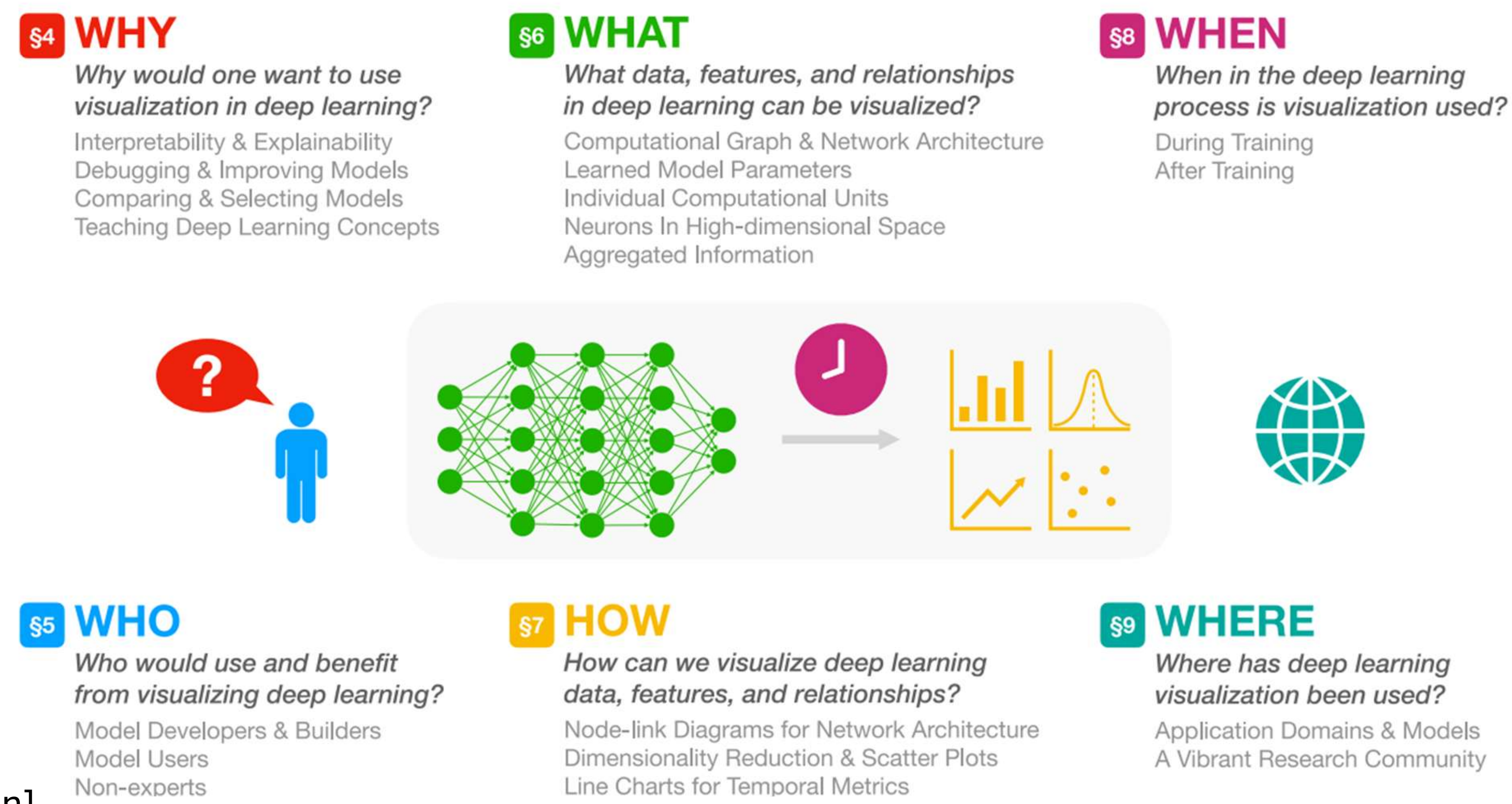
[Homann]

Important Questions



[Homann]

Important Questions



[Homann]

Future Development - Challenges

- Scalability
- Performance analysis
- Bias in representative examples
- Consensus on common visual approach

Future Development - Opportunities

- Using expert knowledge
- Progressive visual analytics
- Advanced deep learning architectures
- Protection against adversarial attacks

Technical Term	Synonyms	Meaning
Neural Network	Artificial neural net, net	Biologically-inspired models that form the basis of deep learning; approximate functions dependent upon a large and unknown amount of inputs consisting of *layers* of *neurons*
Neuron	Computational unit, node	Building blocks of *neural networks*, entities that can apply *activation functions*
Weights	Edges	The trained and updated parameters in the *neural network* model that connect *neurons* to one another
Layer	Hidden layer	Stacked collection of *neurons* that attempt to extract features from data; a *layer's* input is connected to a previous *layer's* output
Computational Graph	Dataflow graph	Directed graph where nodes represent operations and edges represent data paths; when implementing *neural network* models, often times they are represented as these
Activation Functions	Transform function	Functions embedded into each *layer* of a *neural network* that enable the network represent complex non-linear decisions boundaries
Activations	Internal representation	Given a trained network one can pass in data and recover the *activations* at any *layer* of the network to obtain its current representation inside the network
Convolutional Neural Network	CNN, convnet	Type of *neural network* composed of convolutional *layers* that typically assume image data as input; these *layers* have depth unlike typical *layers* that only have width (number of *neurons* in a *layer*); they make use of filters (feature & pattern detectors) to extract spatially invariant representations
Long Short-Term Memory	LSTM	Type of *neural network*, often used in text analysis, that addresses the vanishing gradient problem by using memory gates to propagate gradients through the network to learn long-range dependencies
Loss Function	Objective function, cost function, error	Also seen in general ML contexts, defines what success looks like when learning a representation, i.e., a measure of difference between a *neural network's* prediction and ground truth
Embedding	Encoding	Representation of input data (e.g., images, text, audio, time series) as vectors of numbers in a high-dimensional space; oftentimes reduced so data points (i.e., their vectors) can be more easily analyzed (e.g., compute similarity)
Recurrent Neural Network	RNN	Type of *neural network* where recurrent connections allow the persistence (or "memory") of previous inputs in the network's internal state which are used to influence the network output
Generative Adversarial Networks	GAN	Method to conduct unsupervised learning by pitting a generative network against a discriminative network; the first network mimics the probability distribution of a training dataset in order to fool the discriminative network into judging that the generated data instance belongs to the training set
Epoch	Data pass	A complete pass through a given dataset; by the end of one *epoch*, a *neural network* will have seen every datum within the dataset once

Homann

Sources

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138–52160. doi:10.1109/access.2018.2870052

Alicioglu, G., & Sun, B. (2022, February). A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics, 102*, 502–520. doi:10.1016/j.cag.2021.09.002

Choo, J., & Liu, S. (2018). Visual Analytics for Explainable Deep Learning. *Visual Analytics for Explainable Deep Learning*. arXiv. doi:10.48550/ARXIV.1804.02527

Dragoni, M., Donadello, I., & Eccher, C. (2020, May). Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice. *Artificial Intelligence in Medicine, 105*, 101840. doi:10.1016/j.artmed.2020.101840

Gunning, D., & Aha, D. W. (2019, June). DARPA’s Explainable Artificial Intelligence Program. *AI Magazine, 40*, 44–58. doi:10.1609/aimag.v40i2.2850

Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2019, August). Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics, 25*, 2674–2693. doi:10.1109/tvcg.2018.2843369

La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., . . . Angelini, M. (2023, February). State of the Art of Visual Analytics for eXplainable Deep Learning. *Computer Graphics Forum, 42*, 319–355. doi:10.1111/cgf.14733

Moradi, M., & Samwald, M. (2021, March). Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications, 165*, 113941. doi:10.1016/j.eswa.2020.113941

Rai, A. (2019, December). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science, 48*, 137–141. doi:10.1007/s11747-019-00710-5

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. arXiv. doi:10.48550/ARXIV.1602.04938

Yuan, J., Chen, C., Yang, W., Liu, M., Xia, J., & Liu, S. (2021, March). A survey of visual analytics techniques for machine learning. *Computational Visual Media, 7*, 3–36. doi:10.1007/s41095-020-0191-7