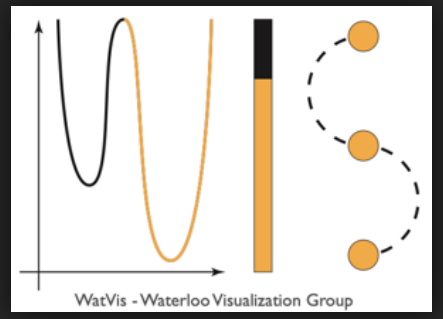
Integrating Predictive Models with Interactive Visualization



Jian Zhao, *Ph.D.*, Assistant Professor
Cheriton School of Computer Science
University of Waterloo
www.jeffjianzhao.com | jianzhao@uwaterloo.ca

Short bio

2009

Ph.D.
@ U Toronto

2015

Researcher
@ Autodesk, Toronto

2016

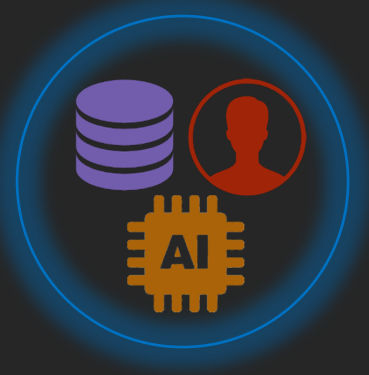
Researcher
@ FXPAL, Palo Alto

2019

Assistant Professor
@ U Waterloo


Data
AI
Machines
Humans
All continuously growing fast!

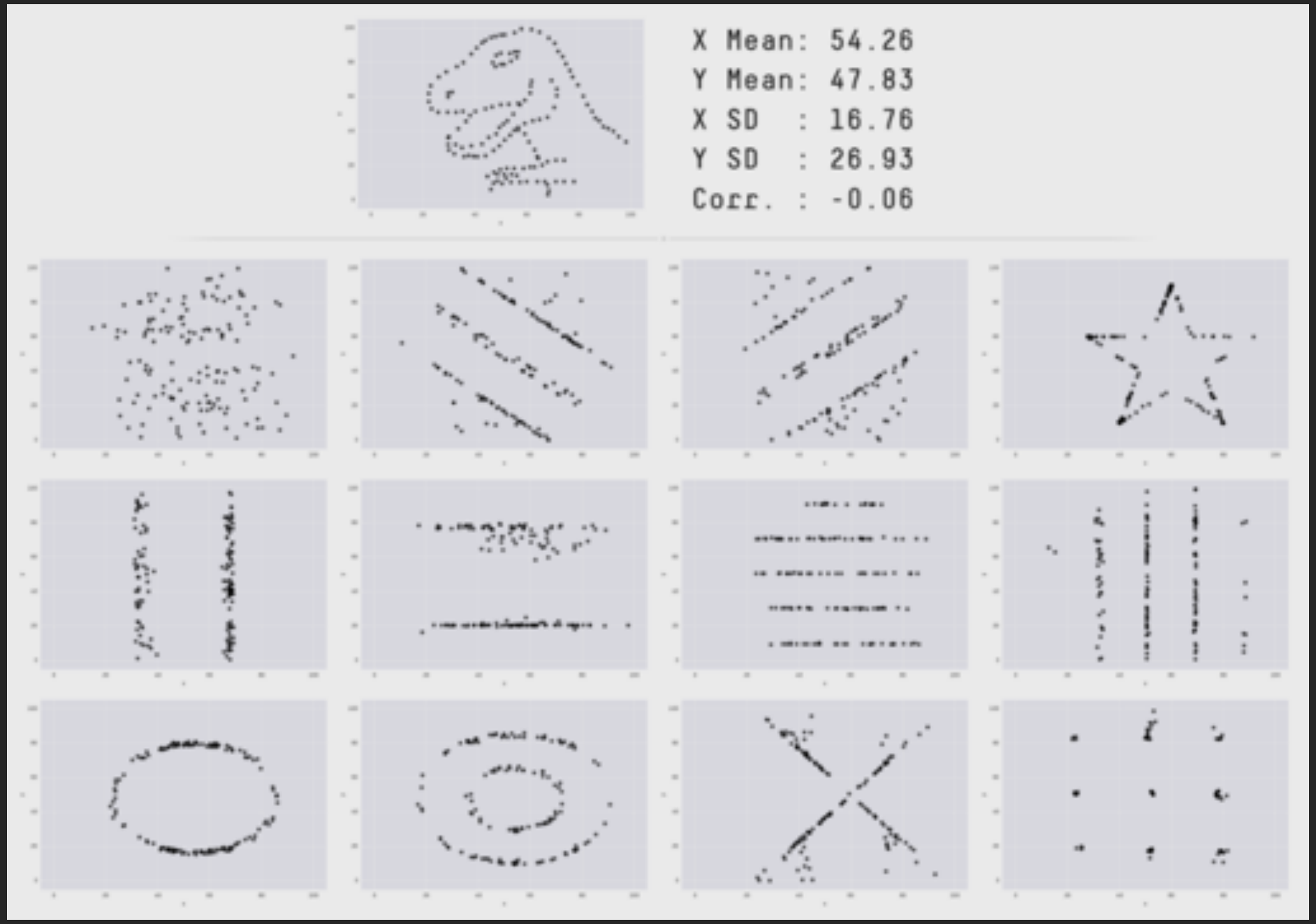
I investigate advanced visualizations (vis) that promote the interplay among data, machines (models), and humans (users) in real-world data science applications.



Bella, Data Scientist

"My input data looks similar, but my classifier performs quite different… Why?"

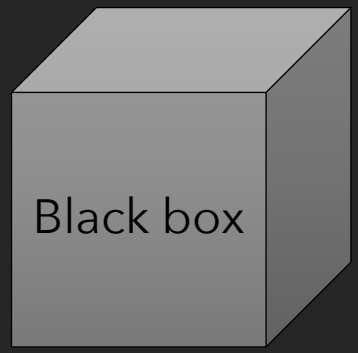
```
X Mean:  54.26
Y Mean:  47.83
X SD   :  16.76
Y SD   :  26.93
Corr.  :  -0.06
```

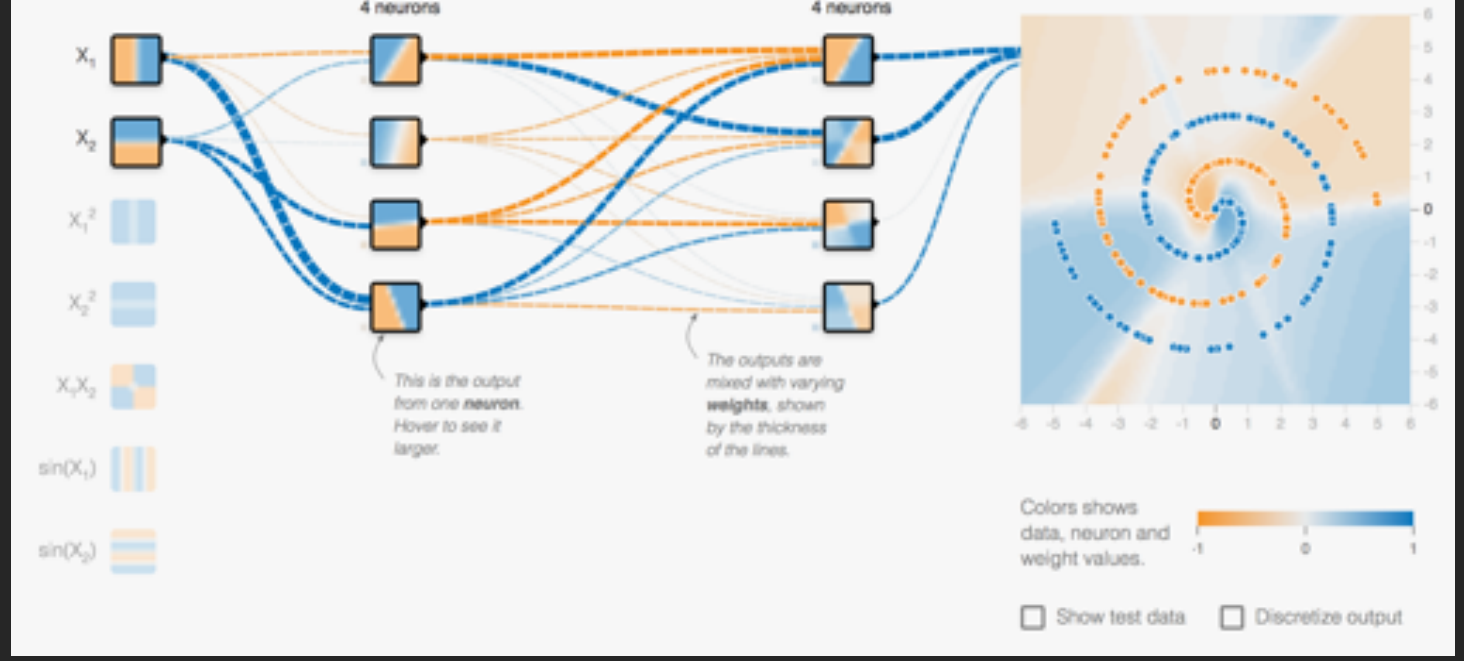


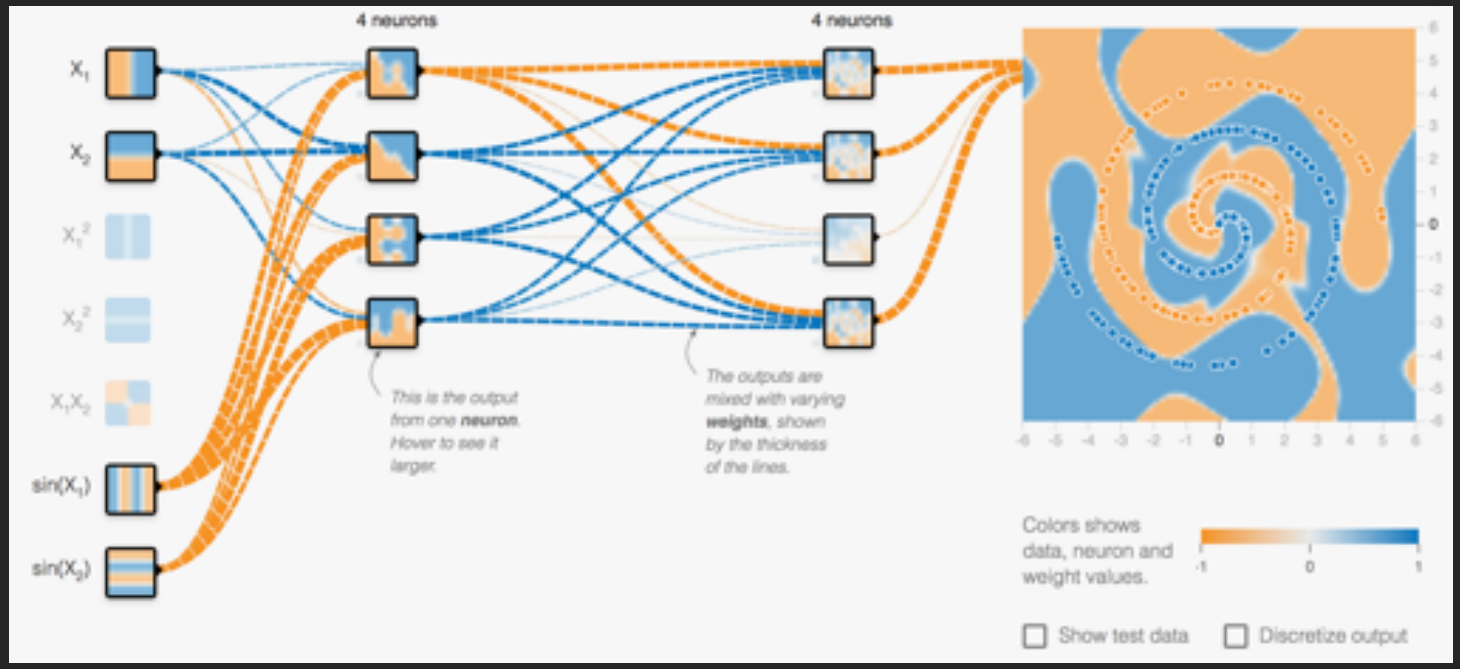
Matejka et al, Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, CHI'17

Bella, Data Scientist

"I'm building a neural network classifier. I tried many ways, but it doesn't work… Why?"





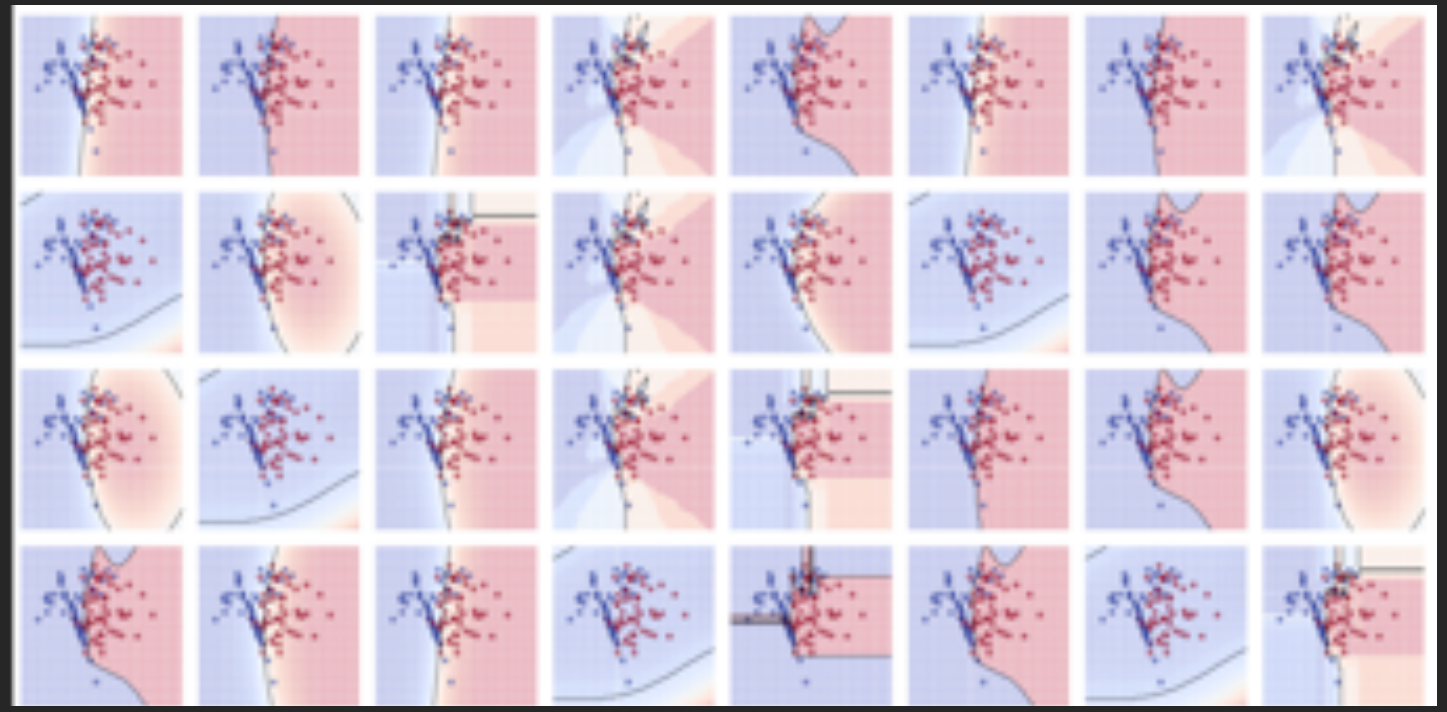


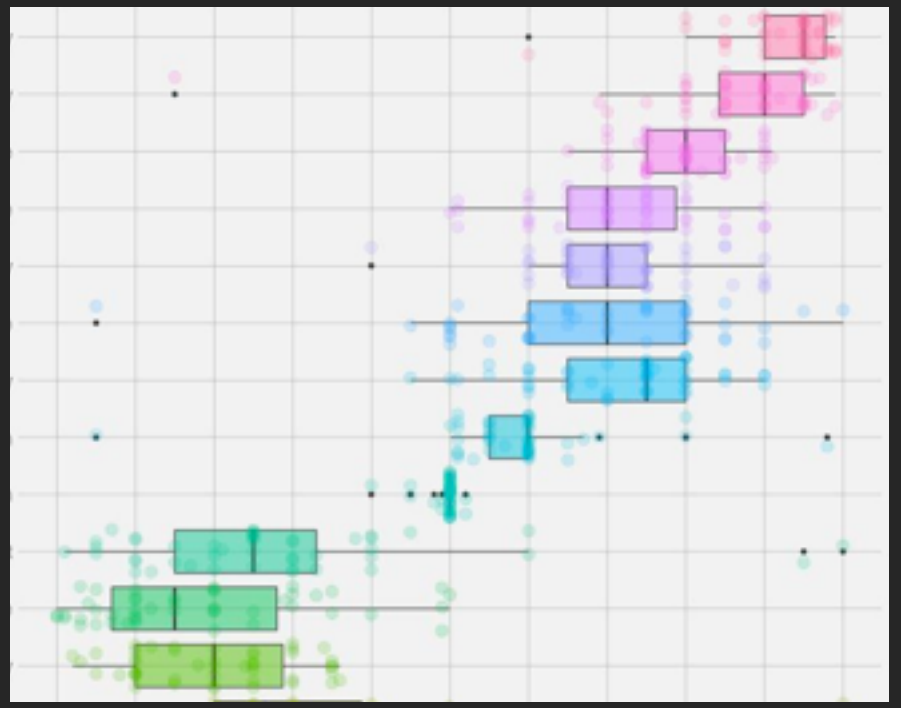
Tensor Flow Playground, http://playground.tensorflow.org/

Bella, Data Scientist

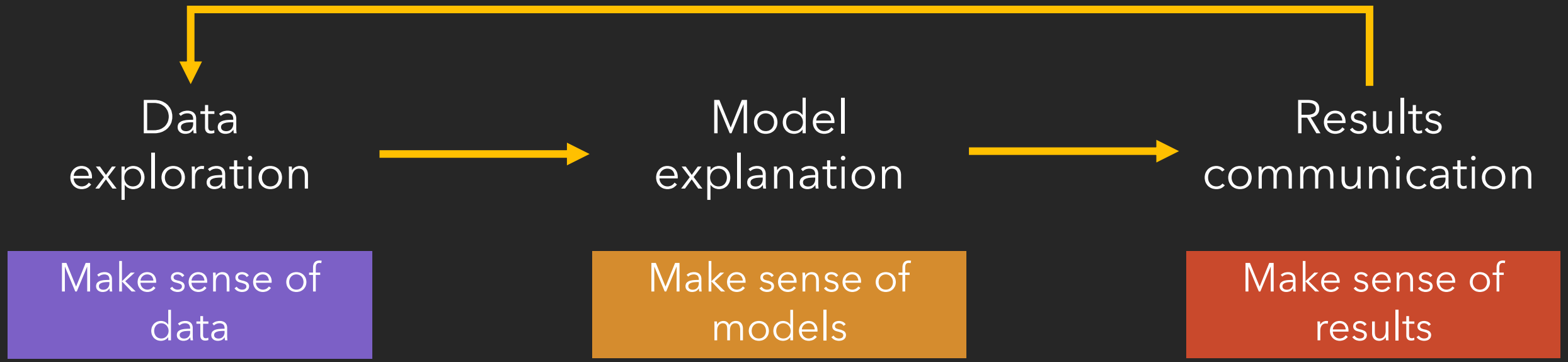
"I finally got some good results, but my boss couldn't understand them..."

Algorithm	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Bayesian Linear Regression	3.258579	4.630709	0.508712	0.279148	0.720852
Neural Network Regression	2.738169	3.64928	0.427469	0.173362	0.826638
Boosted Decision Tree Regression	2.091976	2.84549	0.326588	0.105403	0.894597
Linear Regression	3.423023	4.651562	0.534385	0.281668	0.718332
Decision Forest Regression	2.439228	3.497692	0.3808	0.159258	0.840742

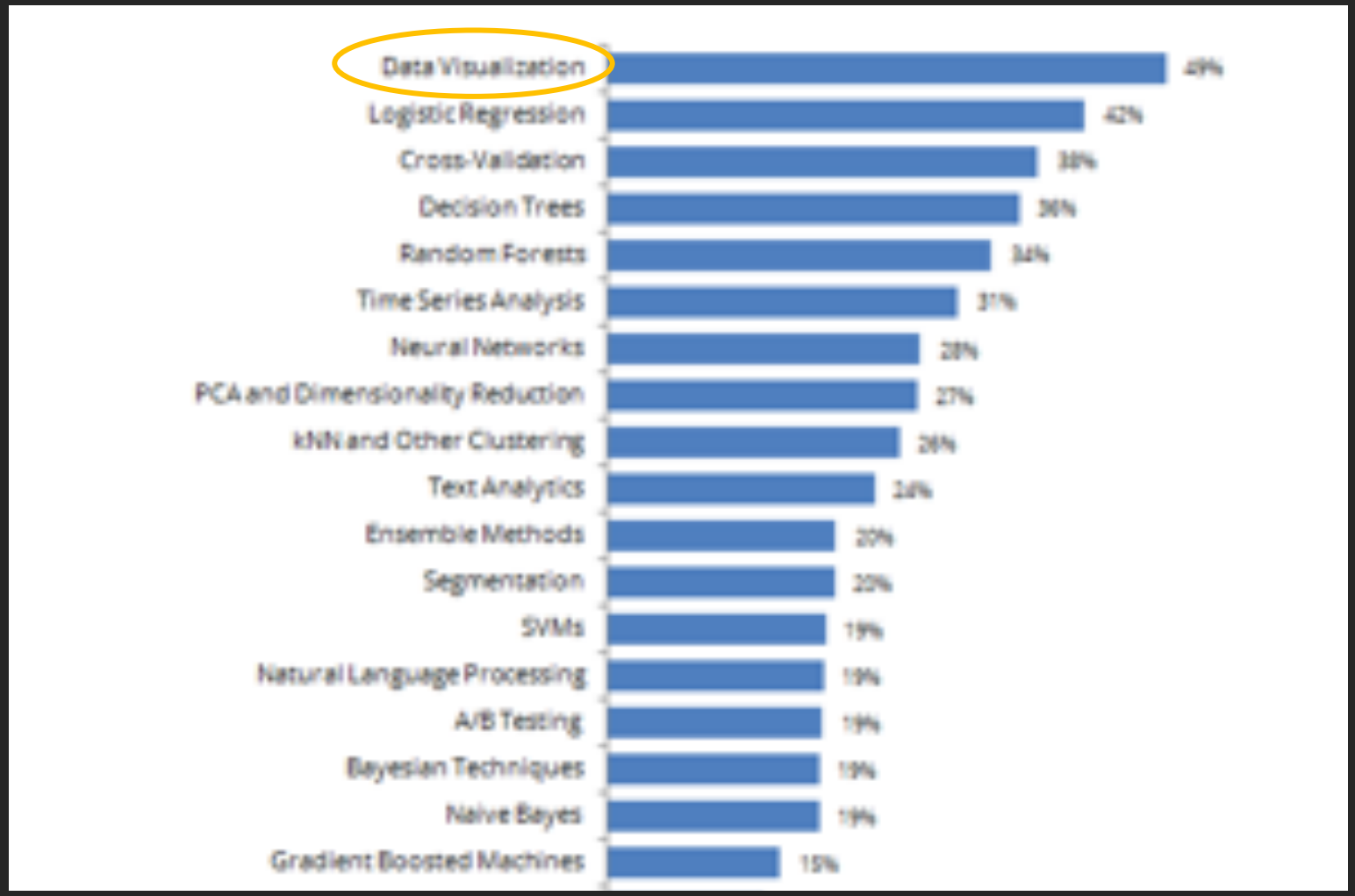




Visualization is critical in data analysis workflow



Top machine learning and data science methods used at work



http://businessoverbroadway.com/top-machine-learning-and-data-science-methods-used-at-work

Creating effective visualizations is hard

Problem/domain specific

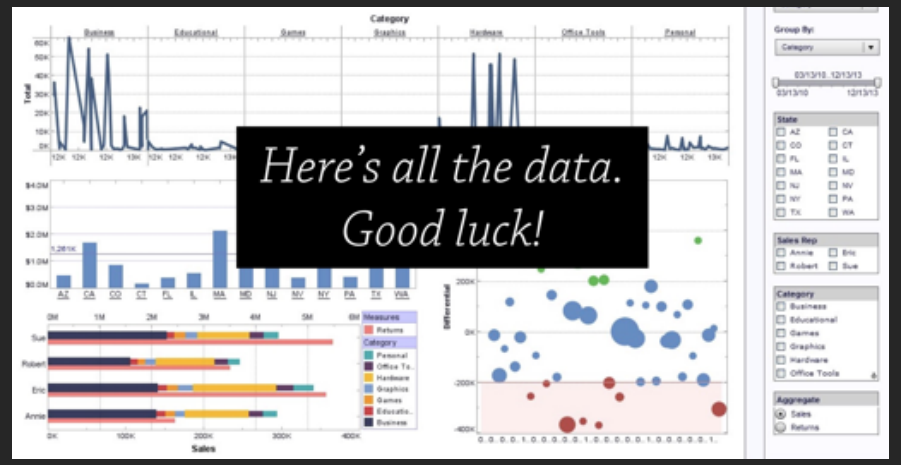
- No easy one-size-fits-all solution

Technical skills

- Matplotlib, D3.js, ggplot2, …

Sense of design

- Huge design space



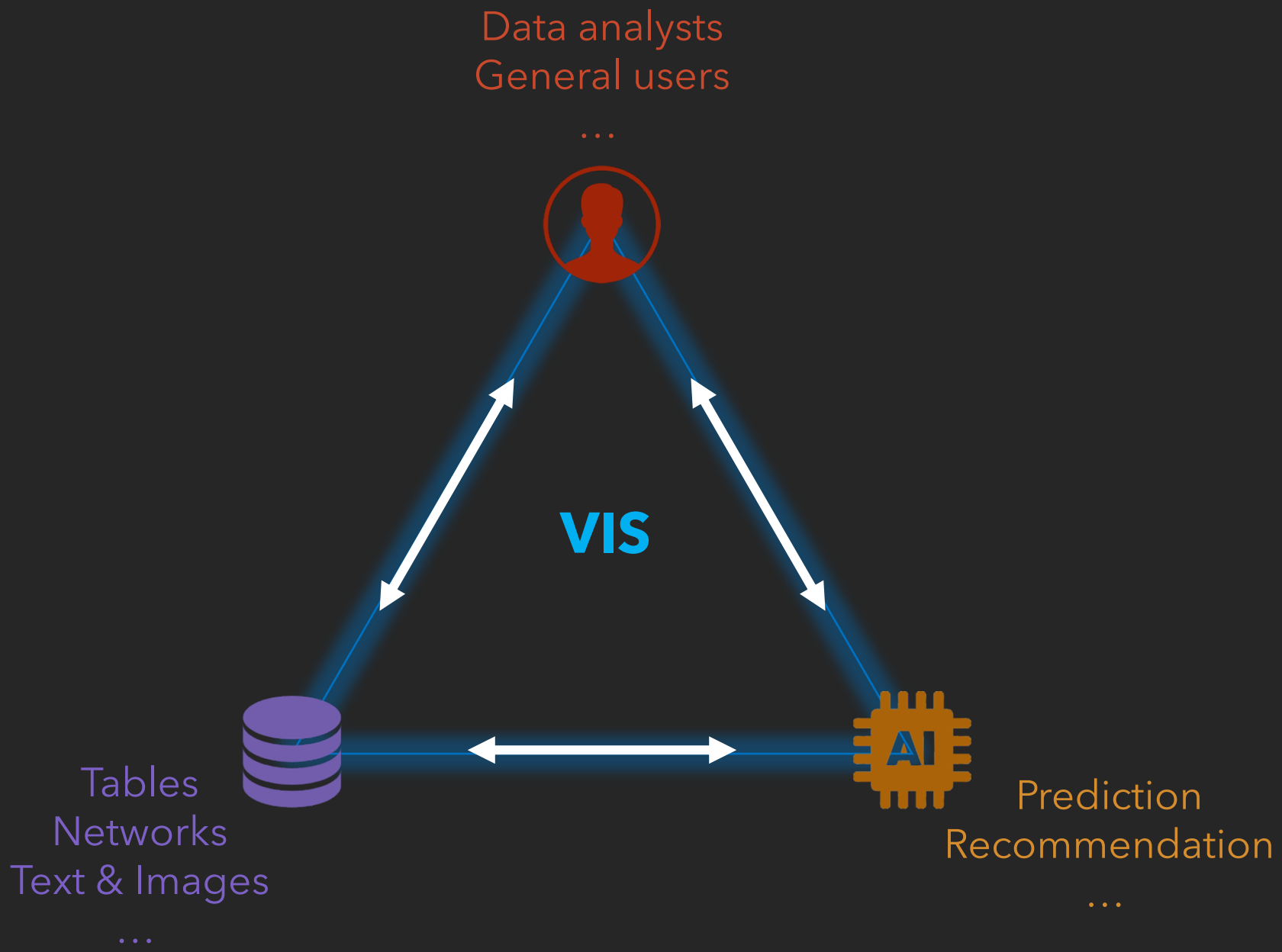
Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
	position, placement	yes	infinite	Good	Good	Good	Good
1, 2, 3; A, B, C	text labels	optional alpha or num	infinite	Good	Good	Good	Good
	length	yes	many	Good	Good		
	size, area	yes	many	Good	Good		
	angle	yes	medium	Good	Good		
	pattern density	yes	few	Good	Good		
	weight, boldness	yes	few		Good		
	saturation, brightness	yes	few		Good		
	color	no	few (<20)			Good	
	shape, icon	no	medium			Good	
	pattern texture	no	medium			Good	
	enclosure, connection	no	infinite			Good	Good
	line pattern	no	few				Good
	line endings	no	few				Good
	line weight	yes	few		Good		

matplotlib

D3

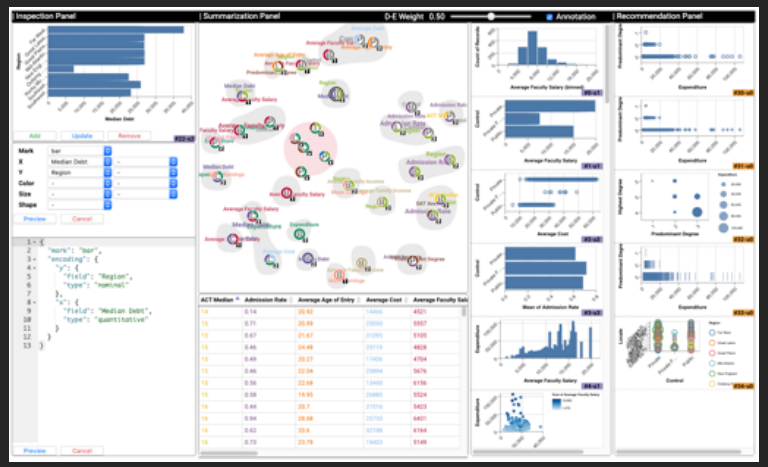
ggplot2

Make sense of data | Make sense of models | Make sense of results



Make sense of data

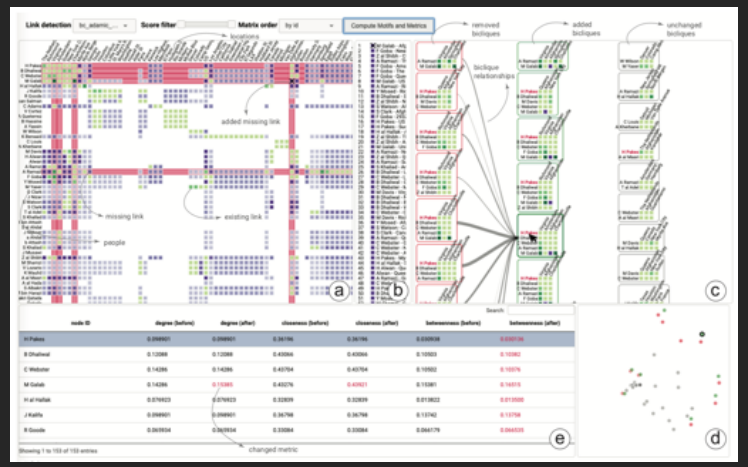
Explore complex data with visualization recommendations



ChartSeer

Make sense of models

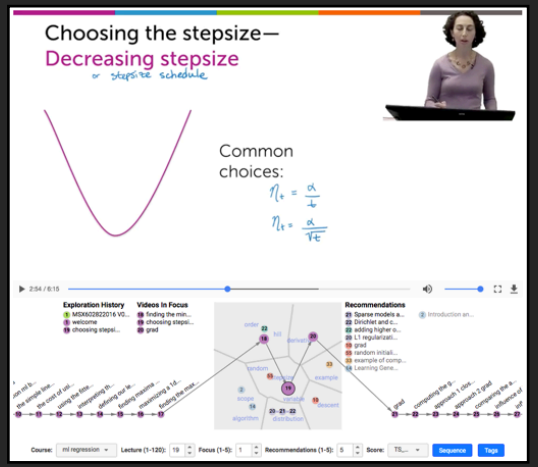
Comprehend missing link prediction in bipartite networks



MissBiN

Make sense of results

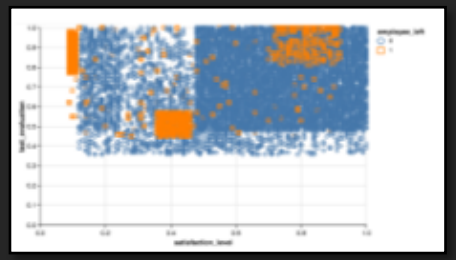
Leverage video recommendations in online learning

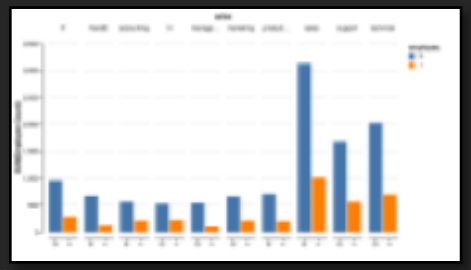


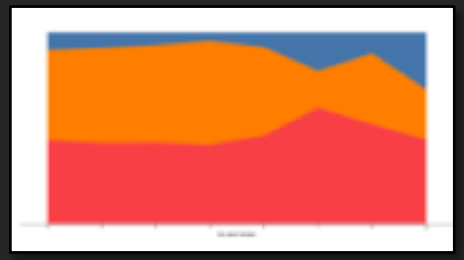
MOOCex

Make sense of data

Exploring large information space







???

Challenges

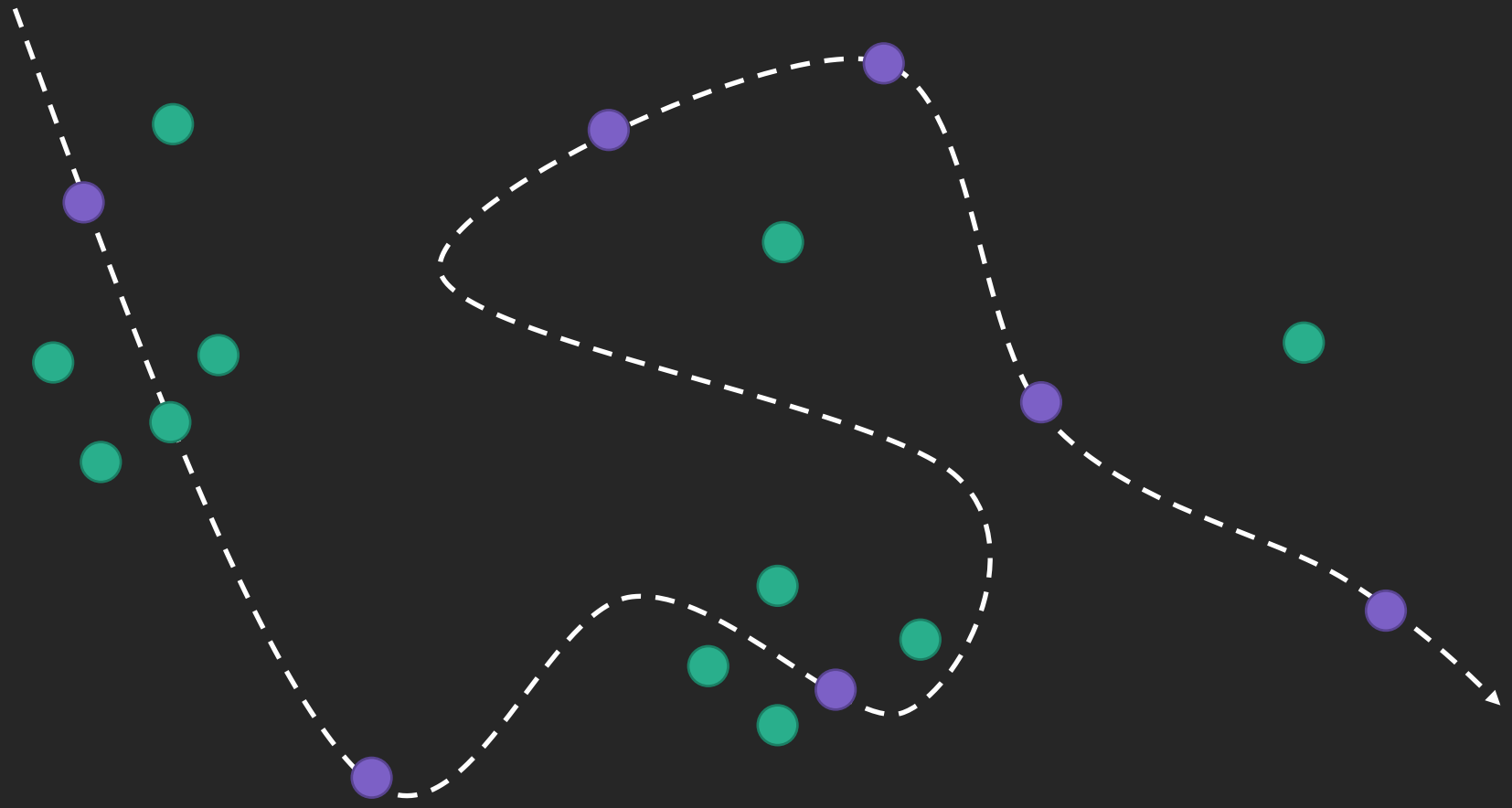
Continuously making decision in a large parameter space

- Which data variables to explore?
- What kind of charts to use?

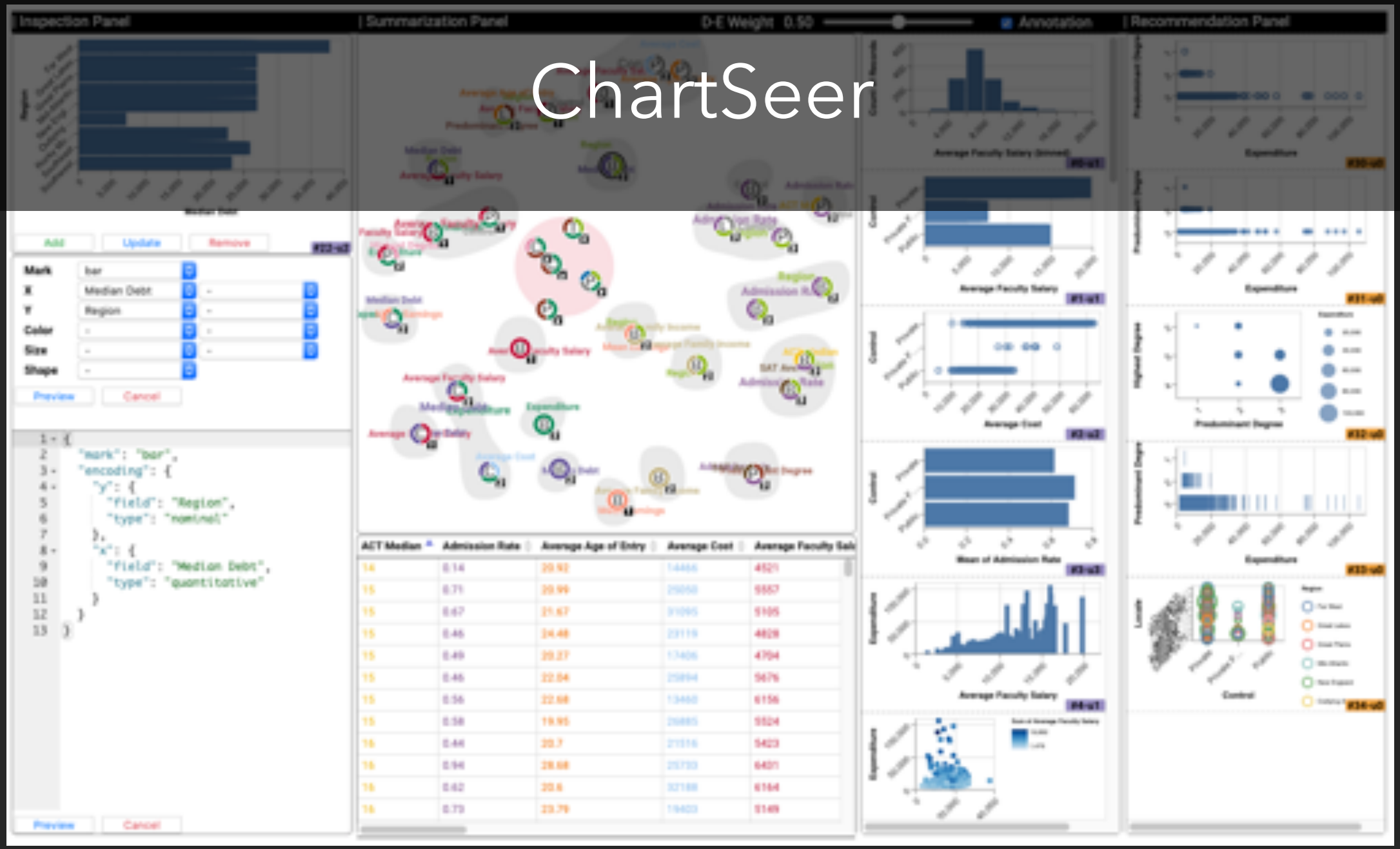
Lacking a holistic view of the analysis space

- How is the current status?
- Where am I?

Exploring large information space with recommendation



ChartSeer



J. Zhao, M. Fan, M. Feng, ChartSeer: Interactive Steering Exploratory Visual Analysis with Machine Intelligence, TVCG


| Inspection Panel
| Summarization Panel
D-E Weight 0.50
Annotation
| Recommendation Panel
Add
Update
Remove
Mark
bar
X
Y
Color
Size
Shape
Preview
Cancel
ACT Median
Admission Rate
Average Age of Entry
Average Cost
Average Faculty Sala
Average Faculty Salary (binned)
Average Faculty Salary
Average Cost
Mean of Admission Rate
Expenditure
Average Faculty Salary
Sum of Average Faculty Salary

System architecture

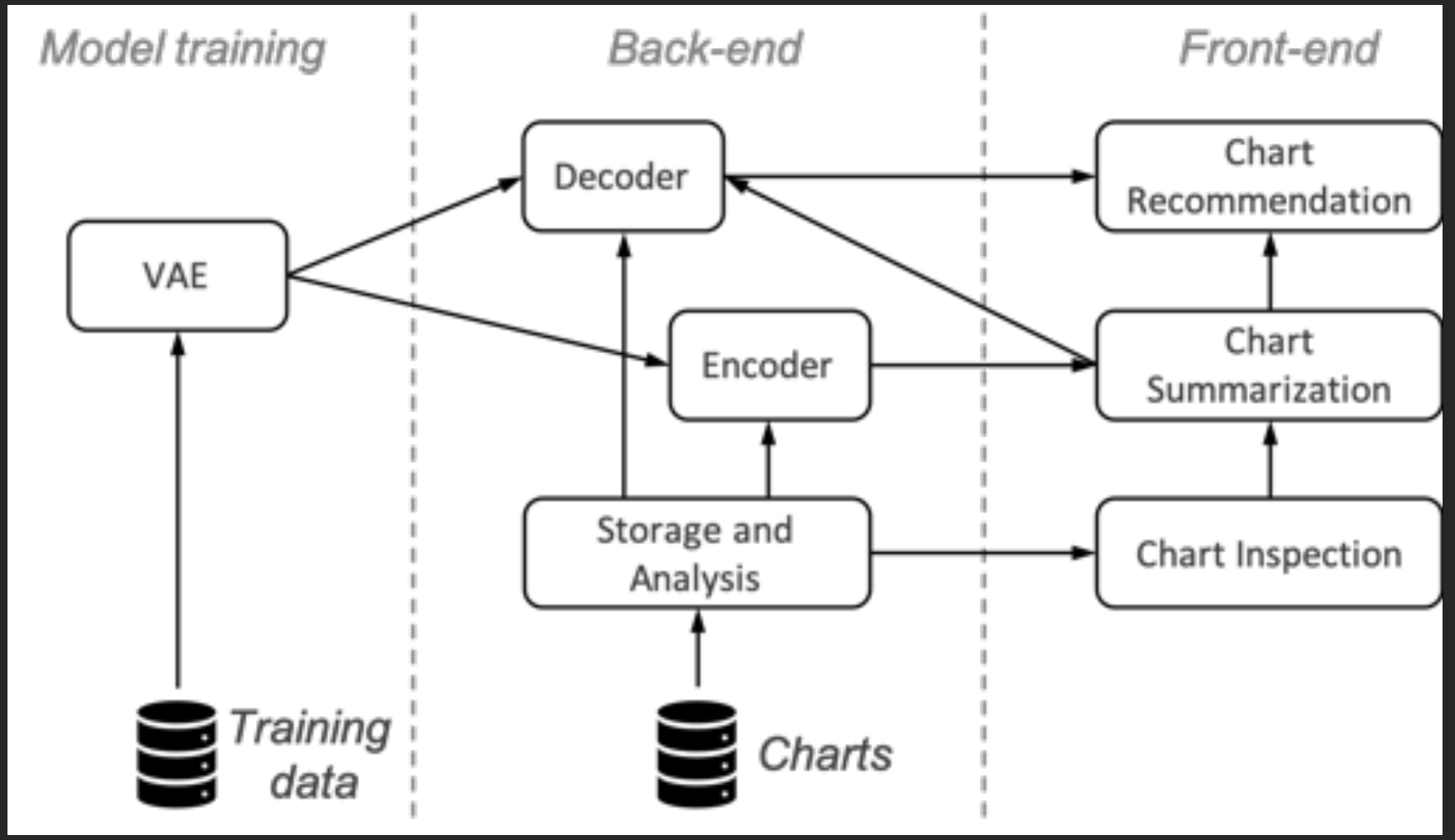


Chart summarization

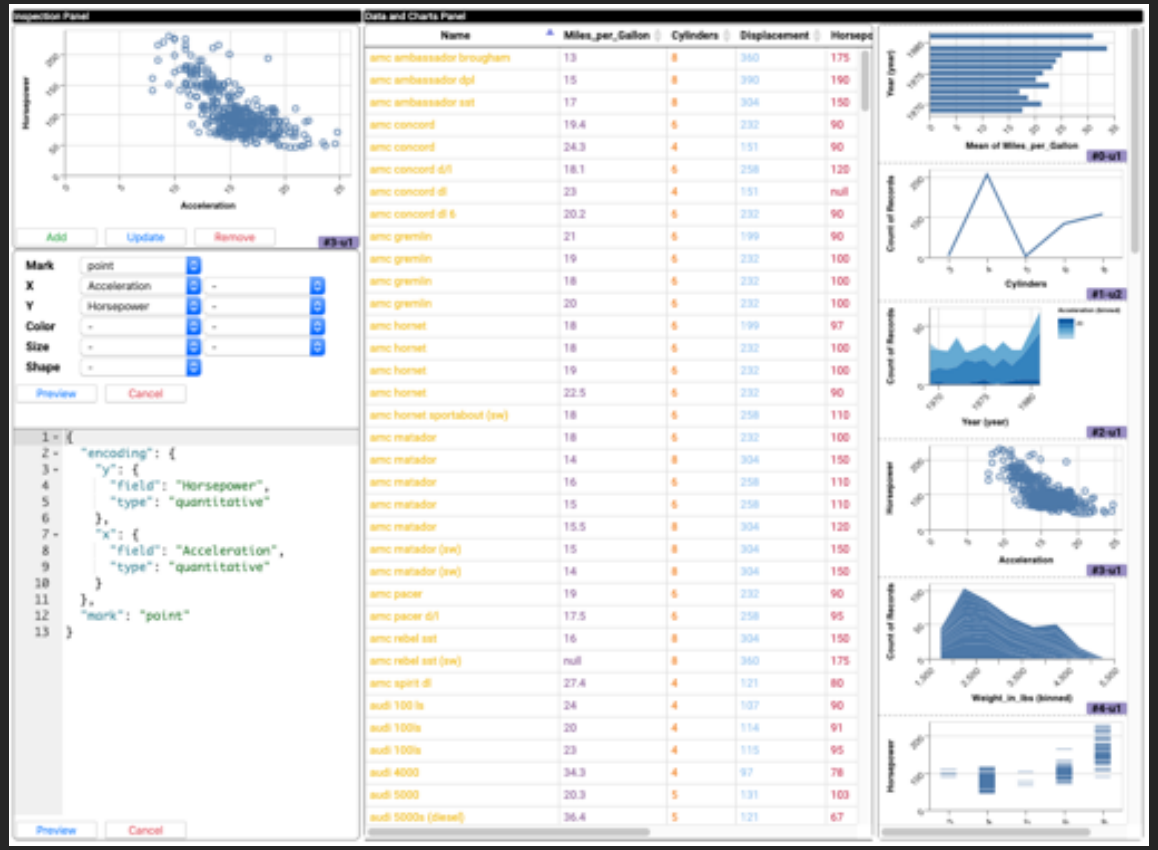
Controlled user study

Between-subjects design
- 24 participants (13 females and 11 males)

Interface conditions
- ChartSeer v.s. Baseline

Dataset
- US college statistics (18 variables)

Tasks
- Summarization task
- Exploration task



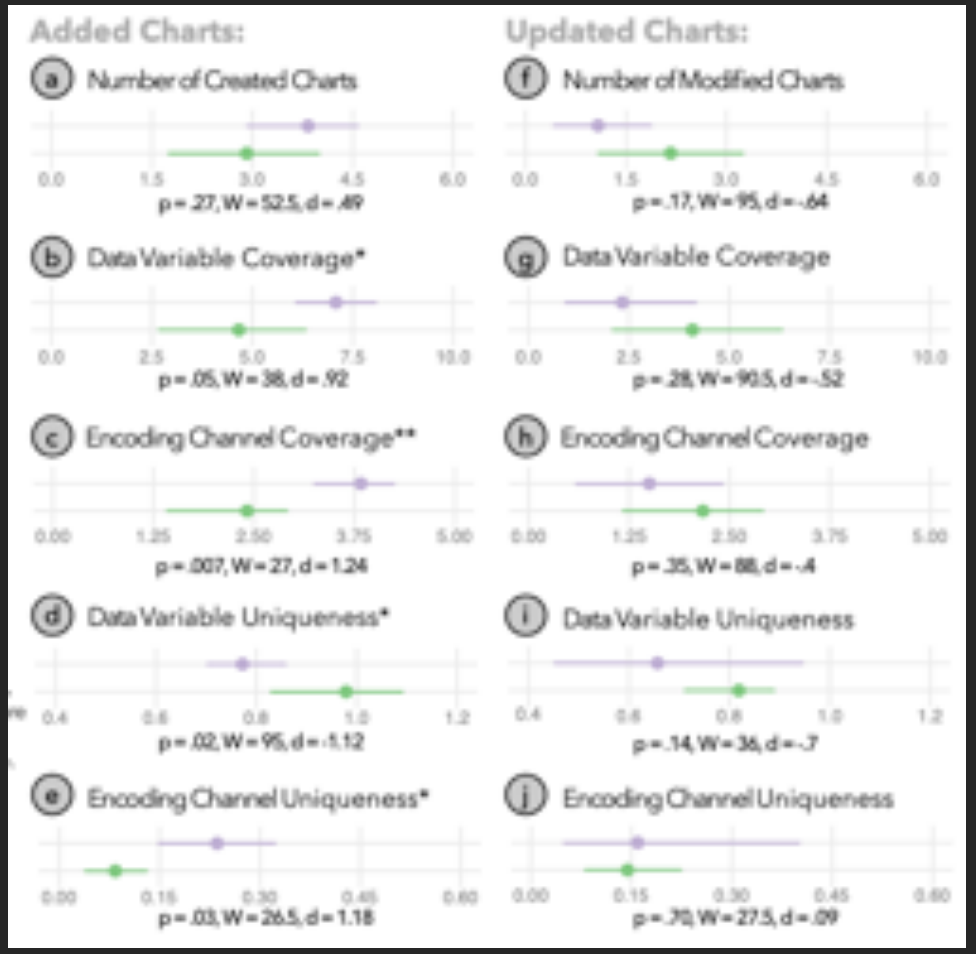
Results of user behaviors

Participants added more charts but updated less charts using ChartSeer

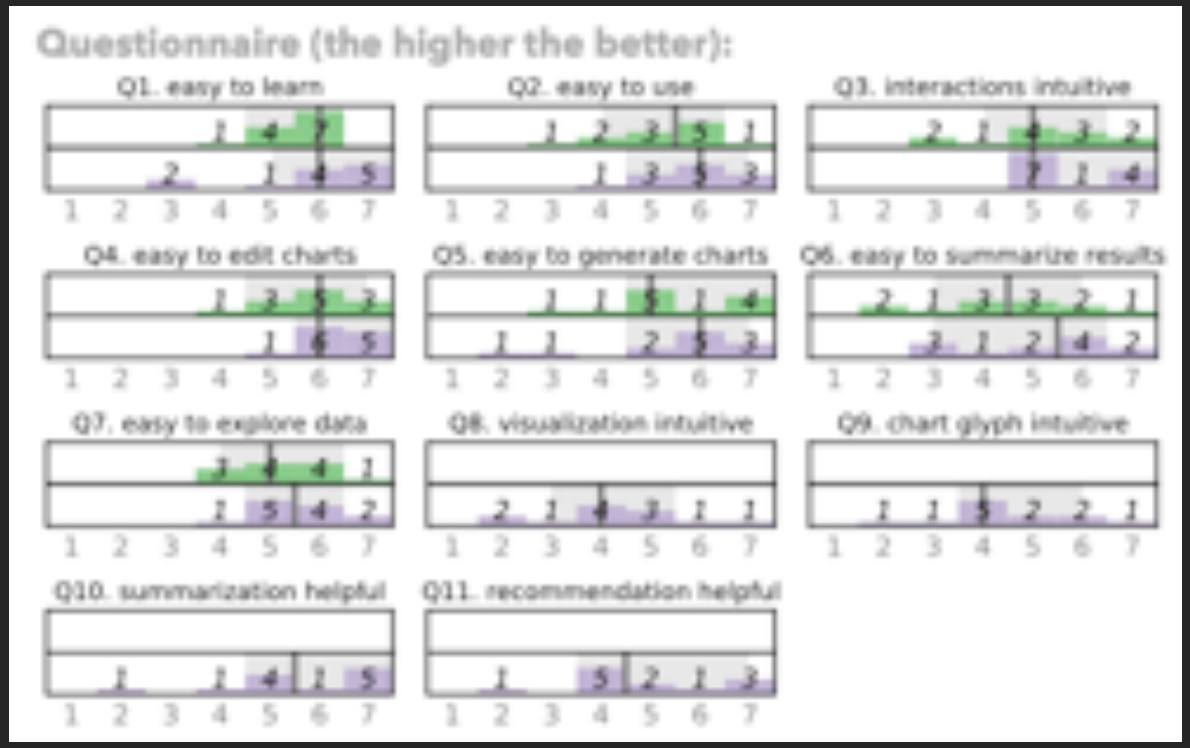
ChartSeer led to a broader range of data variables and visual encodings

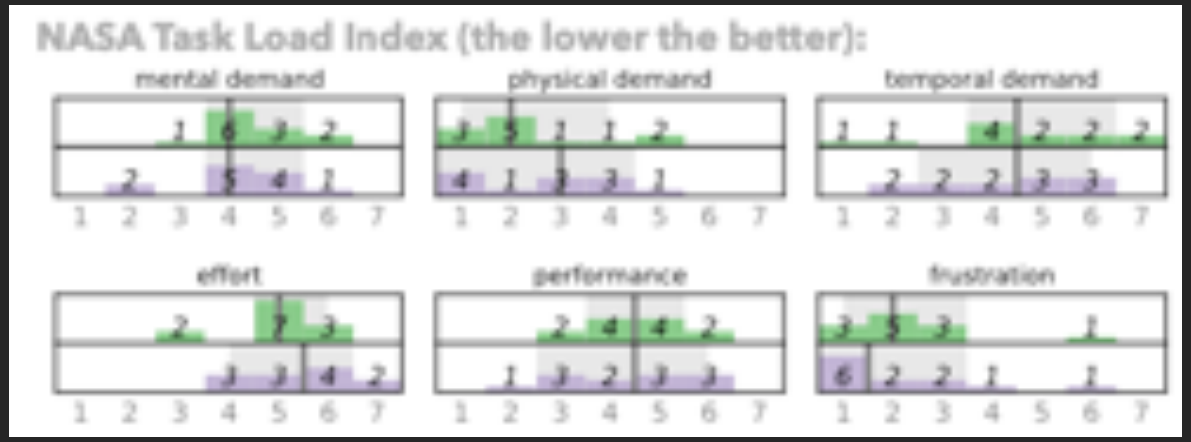
ChartSeer encouraged more focused exploration of data variables

ChartSeer allowed for data exploration from more heterogenous visual perspectives

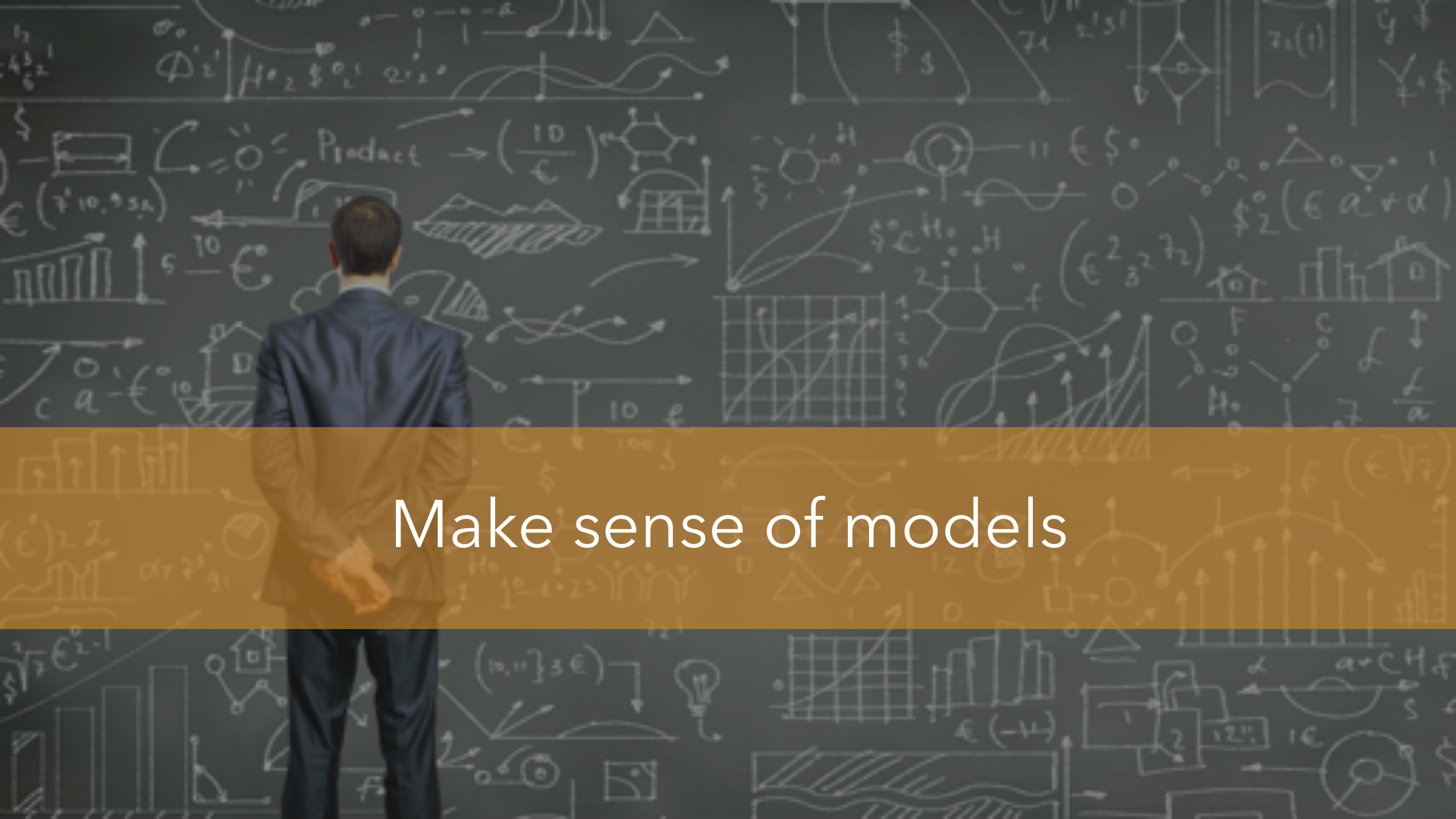


Questionnaire results

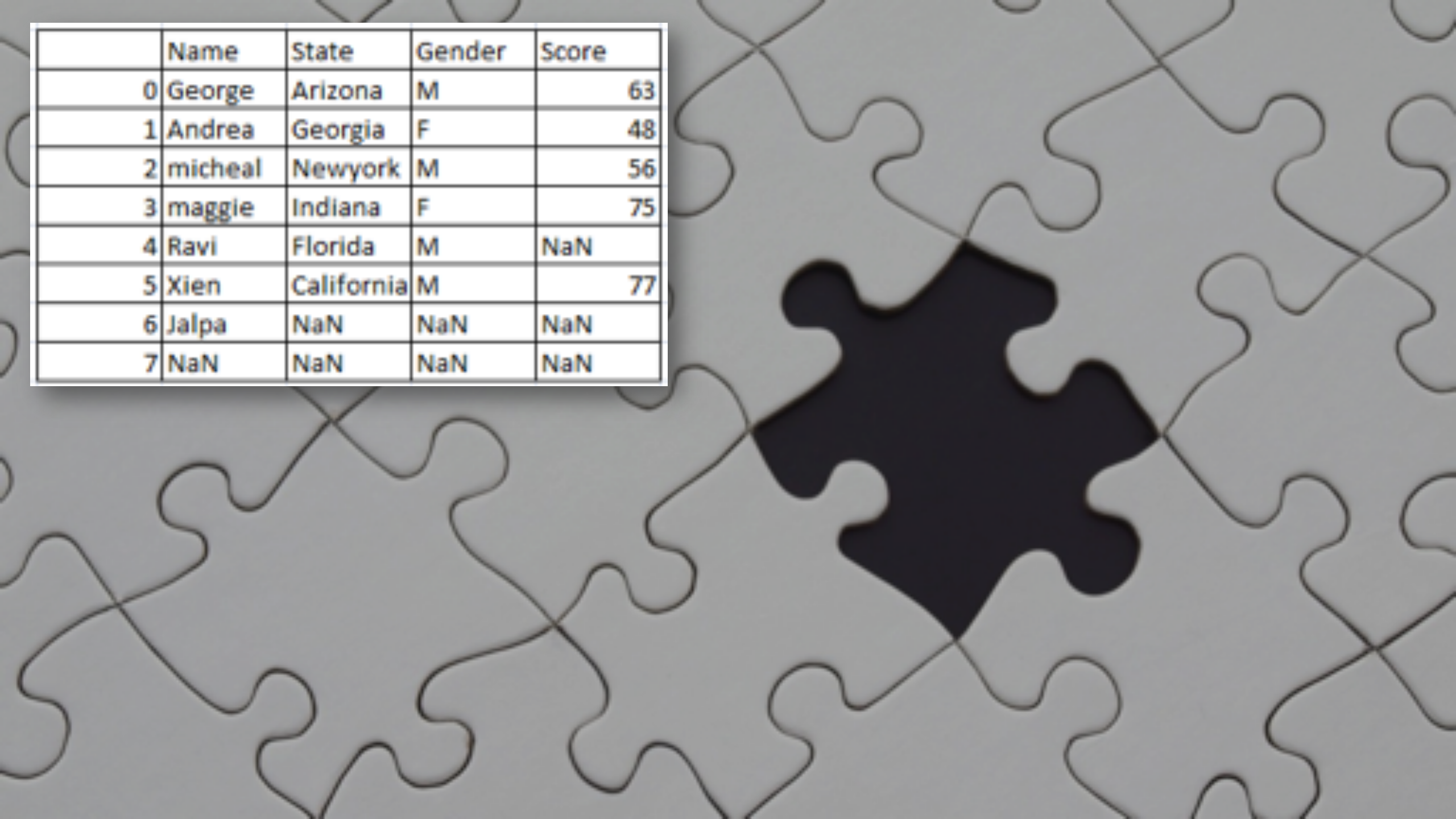




ChartSeer Baseline

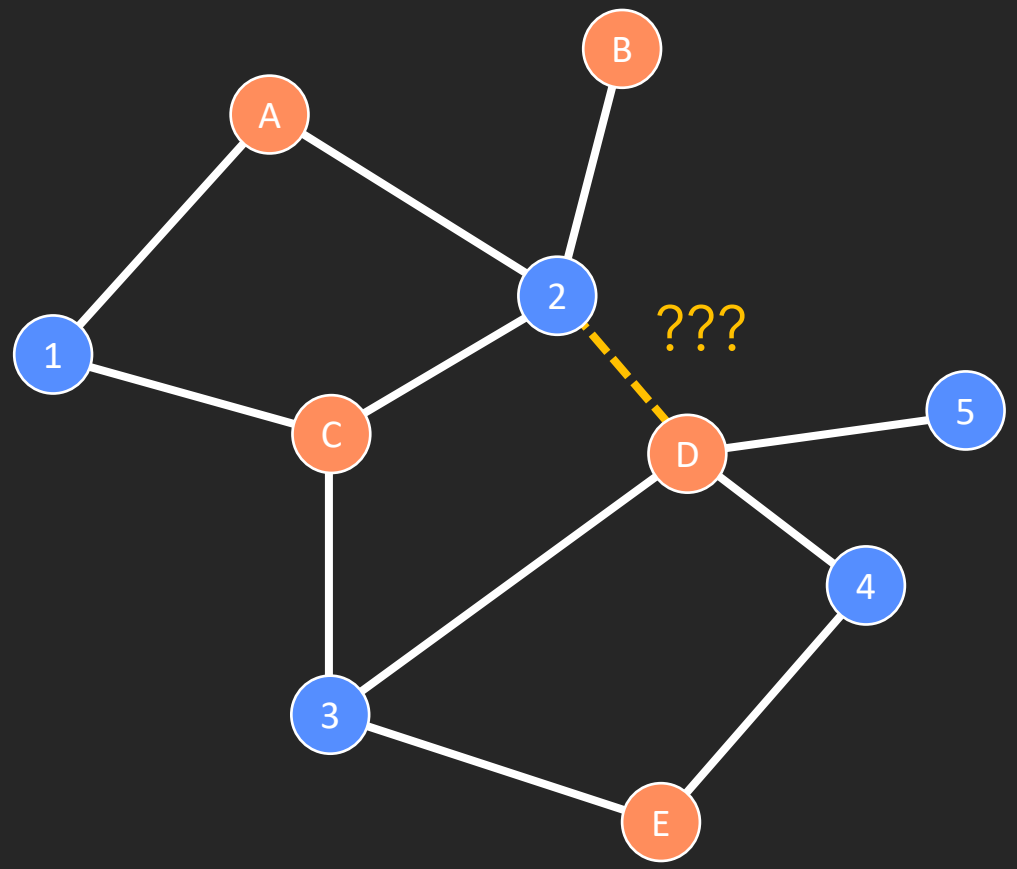


Make sense of models



	Name	State	Gender	Score
0	George	Arizona	M	63
1	Andrea	Georgia	F	48
2	micheal	Newyork	M	56
3	maggie	Indiana	F	75
4	Ravi	Florida	M	NaN
5	Xien	California	M	77
6	Jalpa	NaN	NaN	NaN
7	NaN	NaN	NaN	NaN

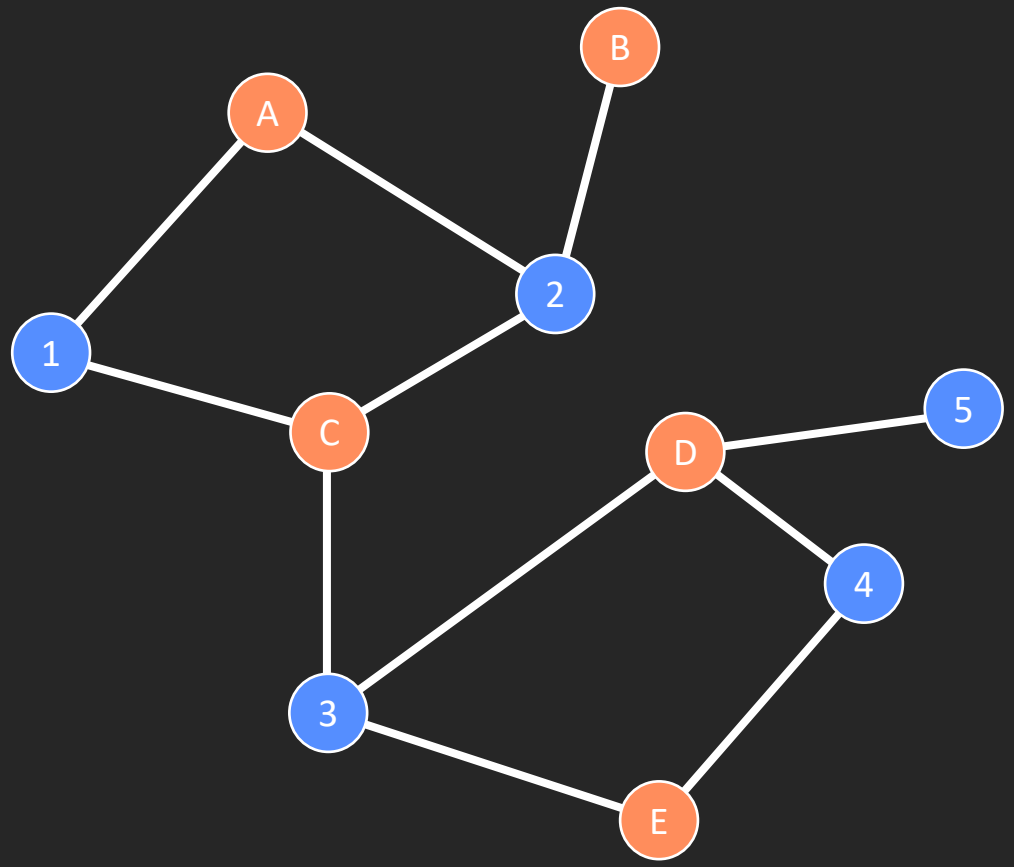
“Missing” links in bipartite networks

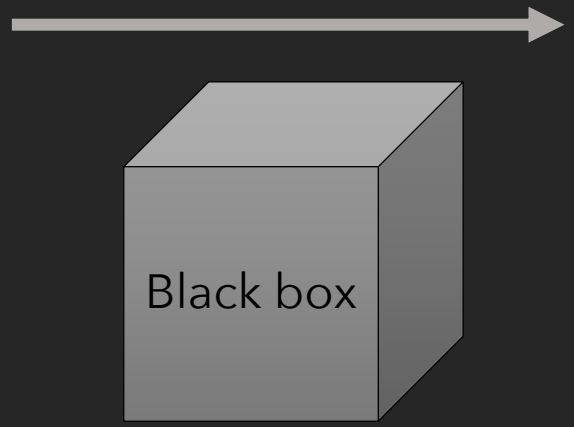


customer

product

Missing link prediction





C – 5: 0.974
D – 2: 0.965
E – 1: 0.873
B – 3: 0.852
…

Analysts' questions

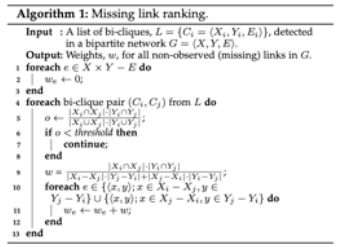
What
are the missing links?

Why
is a link missing?

How
does a missing link impact?

MissBiN

What
are the missing links?



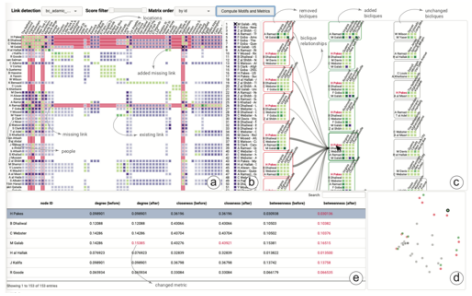
Algorithm 1: Missing link ranking.

Input : A list of bi-cliques, $L = \{C_i = (X_i, Y_i, E_i)\}$, detected in a bipartite network $G = (X, Y, E)$.
Output: Weights, w, for all non-observed (missing) links in G.

```
foreach e ∈ X × Y − E do
    w_e ← 0;
end
foreach bi-clique pair (C_i, C_j) from L do
    o ← (|X_i ∩ X_j| · |Y_i ∩ Y_j|) / (|X_i ∪ X_j| · |Y_i ∪ Y_j|);
    if o < threshold then
        continue;
    end
    w = (|X_i ∩ X_j| · |Y_i ∩ Y_j|) / (|X_i − X_j| · |Y_j − Y_i| + |X_j − X_i| · |Y_i − Y_j|);
    foreach e ∈ {(x, y) : x ∈ X_i − X_j, y ∈ Y_j − Y_i} ∪ {(x, y) : x ∈ X_j − X_i, y ∈ Y_j − Y_i} do
        w_e ← w_e + w;
    end
end
```

A missing link prediction algorithm

Why
is a link missing?



An interactive visualization

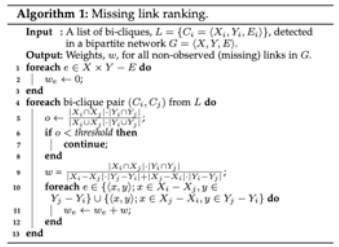
How
does a missing link impact?

A comparative analysis approach

J. Zhao, M. Sun, F. Chen, P, Chiu, MissBiN: Visual Analysis of Missing Links in Bipartite Networks, VIS'19
J. Zhao, M. Sun, F. Chen, P, Chiu, Understanding Missing Links in Bipartite Networks with MissBiN, TVCG

Addressing the questions with MissBiN

What
are the missing links?



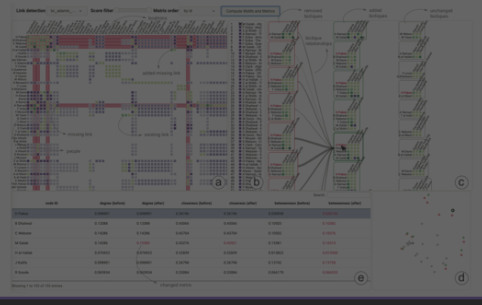
Algorithm 1: Missing link ranking.

Input : A list of bi-cliques, $L = \{C_i = (X_i, Y_i, E_i)\}$, detected in a bipartite network $G = (X, Y, E)$.
Output: Weights, w, for all non-observed (missing) links in G.

foreach $e \in X \times Y - E$ **do**
  $w_e \leftarrow 0$;
end
foreach bi-clique pair (C_i, C_j) from L **do**
  $o \leftarrow \frac{|X_i \cap X_j| \cdot |Y_i \cap Y_j|}{|X_i \cup X_j| \cdot |Y_i \cup Y_j|}$;
  if $o <$ *threshold* **then**
    continue;
  end
  $w = \frac{|X_i \cap X_j| \cdot |Y_i \cap Y_j|}{|X_i - X_j| \cdot |Y_j - Y_i| + |X_j - X_i| \cdot |Y_i - Y_j|}$;
  foreach $e \in \{\langle x, y \rangle; x \in X_i - X_j, y \in Y_j - Y_i\} \cup \{\langle x, y \rangle; x \in X_j - X_i, y \in Y_j - Y_i\}$ **do**
    $w_e \leftarrow w_e + w$;
  end
end

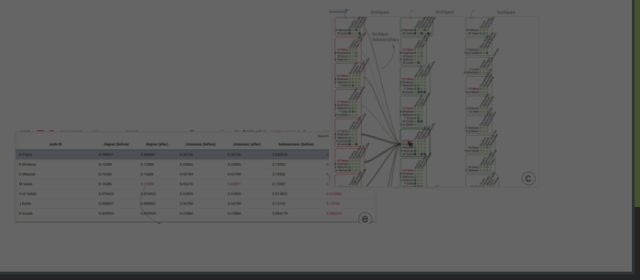
A missing link prediction algorithm

Why
is a link missing?



An interactive visualization

How
does a missing link impact?



A comparative analysis approach

Prediction of missing links

1. Predict the missing links with standard methods (e.g., common neighbors [Chang12])

2. Discover all maximal bicliques, complete subgraphs, of the network (e.g., using MBEA [Zhang14])

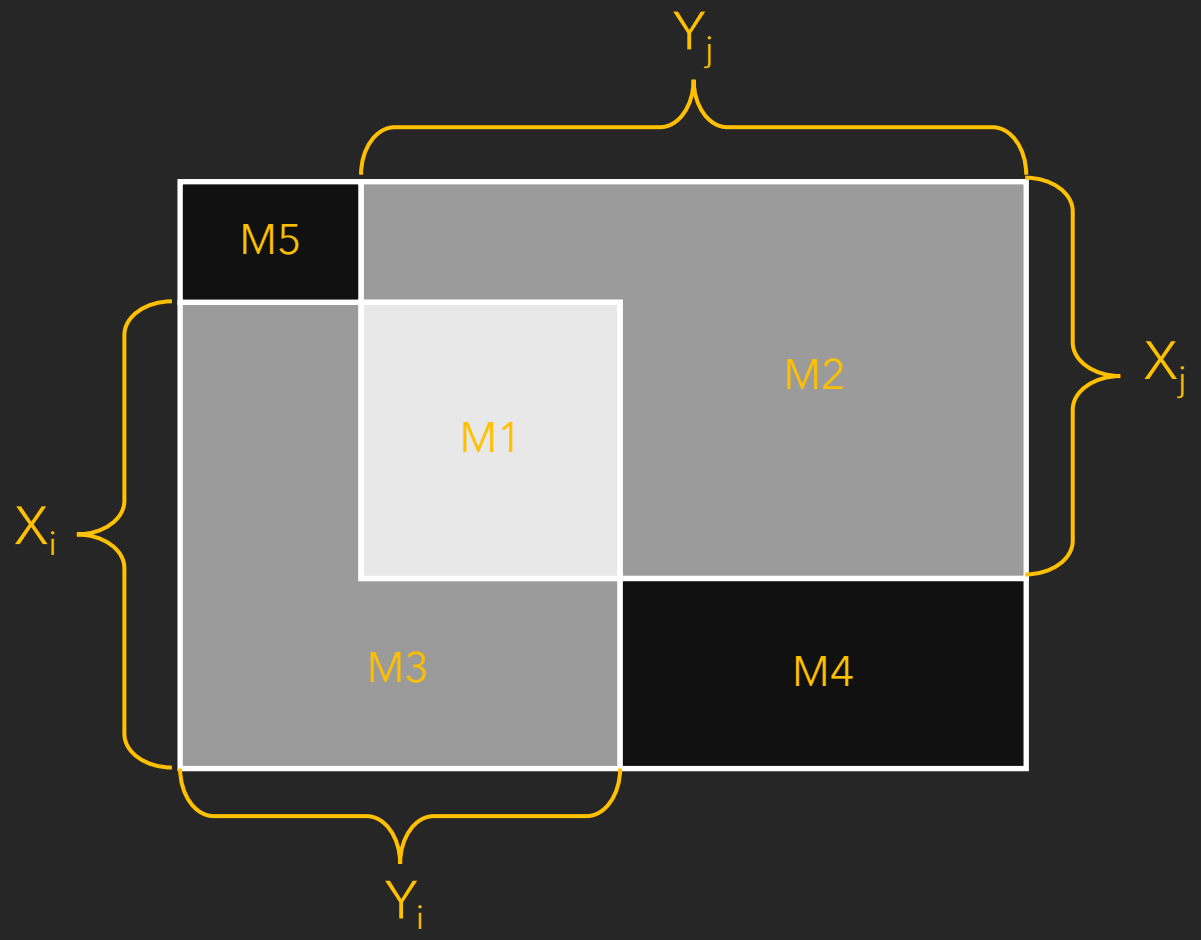
3. Re-rank the missing links based on the overlap of bicliques

Algorithm 1: Missing link ranking.

Input : A list of bi-cliques, $L = \{C_i = (X_i, Y_i, E_i)\}$, detected in a bipartite network $G = (X, Y, E)$.
Output: Weights, w, for all non-observed (missing) links in G.

foreach $e \in X \times Y - E$ **do**
  $w_e \leftarrow 0$;
end
foreach bi-clique pair (C_i, C_j) from L **do**
  $o \leftarrow \frac{|X_i \cap X_j| \cdot |Y_i \cap Y_j|}{|X_i \cup X_j| \cdot |Y_i \cup Y_j|}$;
  if $o <$ *threshold* **then**
    continue;
  end
  $w = \frac{|X_i \cap X_j| \cdot |Y_i \cap Y_j|}{|X_i - X_j| \cdot |Y_j - Y_i| + |X_j - X_i| \cdot |Y_i - Y_j|}$;
  foreach $e \in \{(x, y); x \in X_i - X_j, y \in Y_j - Y_i\} \cup \{(x, y); x \in X_j - X_i, y \in Y_j - Y_i\}$ **do**
    $w_e \leftarrow w_e + w$;
  end
end

In step3, for each pair of bicliques, …



$$w_e \leftarrow w_e + \frac{|X_i \cap X_j| \cdot |Y_i \cap Y_j|}{|X_i - X_j| \cdot |Y_j - Y_i| + |X_j - X_i| \cdot |Y_i - Y_j|}$$

$$\frac{Area(M1)}{Area(M2 + M3 + M4 + M5)}$$

Re-ranking predicted missing links

$$s'_e = w_e \cdot s_e$$

Weights computed in step3, based on bicliques information

Scores computed in step1, based on standard methods

Evaluation of missing link prediction

Test on 3 datasets

- Person-place network from Atlantic Storm corpus [Hughes05]
- User-conversation network from Slack group communication

Compare with 5 base methods

- Jaccard coefficient (JA)
- common neighbors (CN)
- Adamic-Adar coefficient (AA)
- preferential attachment (PA)
- random walk (RW)

Link prediction results

Mostly, PA has the largest performance gain
Secondly, CN performs well

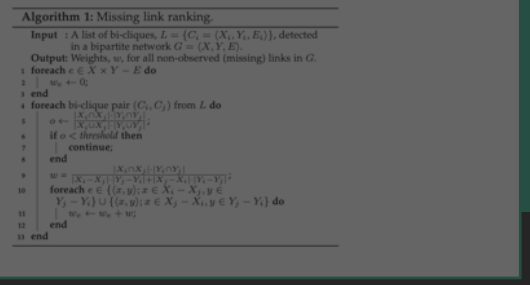
Original method | Our method | Performance gain

		Atlantic Storm			Slack communication			VIS publications			Atlantic Storm*			Slack communication*		
R-Precision	JA	.395	.428	.032	.421	.424	.004	.105	.126	.020	.405	.423	.018	.332	.341	.009
	CN	.359	.580	.221	.179	.424	.245	.066	.129	.063	.342	**.552**	.209	.202	.339	.138
	AA	.440	.455	.016	.202	.219	.016	.107	**.152**	.045	.405	.418	.013	.203	.218	.015
	PA	.021	**.585**	.564	.025	**.590**	.565	.000	.012	.012	.007	.226	.219	.096	**.432**	.336
	RW	.467	.531	.064	.451	.464	.013	.096	.112	.016	.487	**.552**	.065	.398	.407	.009
AUC PR	JA	.398	.451	.053	.301	.325	.024	.039	.053	.014	.393	.444	.051	.225	.249	.024
	CN	.305	.607	.302	.148	.326	.178	.025	.063	.039	.300	**.578**	.277	.133	.285	.152
	AA	.398	.429	.031	.170	.200	.030	.039	**.068**	.029	.379	.406	.027	.153	.177	.025
	PA	.008	**.566**	.557	.021	**.528**	.507	.000	.003	.002	.005	.212	.207	.055	**.406**	.352
	RW	.435	.516	.082	.346	.377	.031	.039	.052	.013	.448	.525	.078	.279	.308	.029

Jaccard coefficient (JA), common neighbors (CN), Adamic-Adar coefficient (AA), preferential attachment (PA), random walk (RW)

Addressing the questions with MissBiN

What
are the missing links?



Algorithm 1: Missing link ranking.

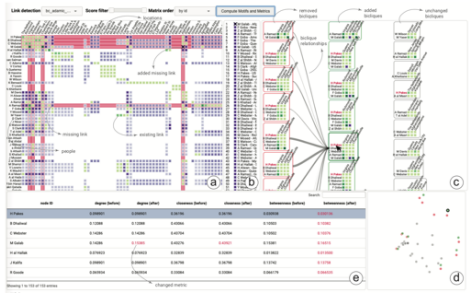
Input : A list of bi-cliques, $L = \{C_i = (X_i, Y_i, E_i)\}$, detected in a bipartite network $G = (X, Y, E)$.

Output: Weights, w_e for all non-observed (missing) links in G.

```
foreach e ∈ X × Y − E do
    w_e ← 0;
end
foreach bi-clique pair (C_i, C_j) from L do
    o ← (|X_i ∩ X_j| · |Y_i ∩ Y_j|) / (|X_i ∪ X_j| · |Y_i ∪ Y_j|);
    if o < threshold then
        continue;
    end
    w = (|X_i ∩ X_j| · |Y_i ∩ Y_j|) / (|X_i − X_j| · |Y_i − Y_j| + |X_j − X_i| · |Y_i − Y_j|);
    foreach e ∈ {(x, y) : x ∈ X_i − X_j, y ∈ Y_j − Y_i} ∪ {(x, y) : x ∈ X_j − X_i, y ∈ Y_j − Y_i} do
        w_e ← w_e + w;
    end
end
```

A missing link prediction algorithm

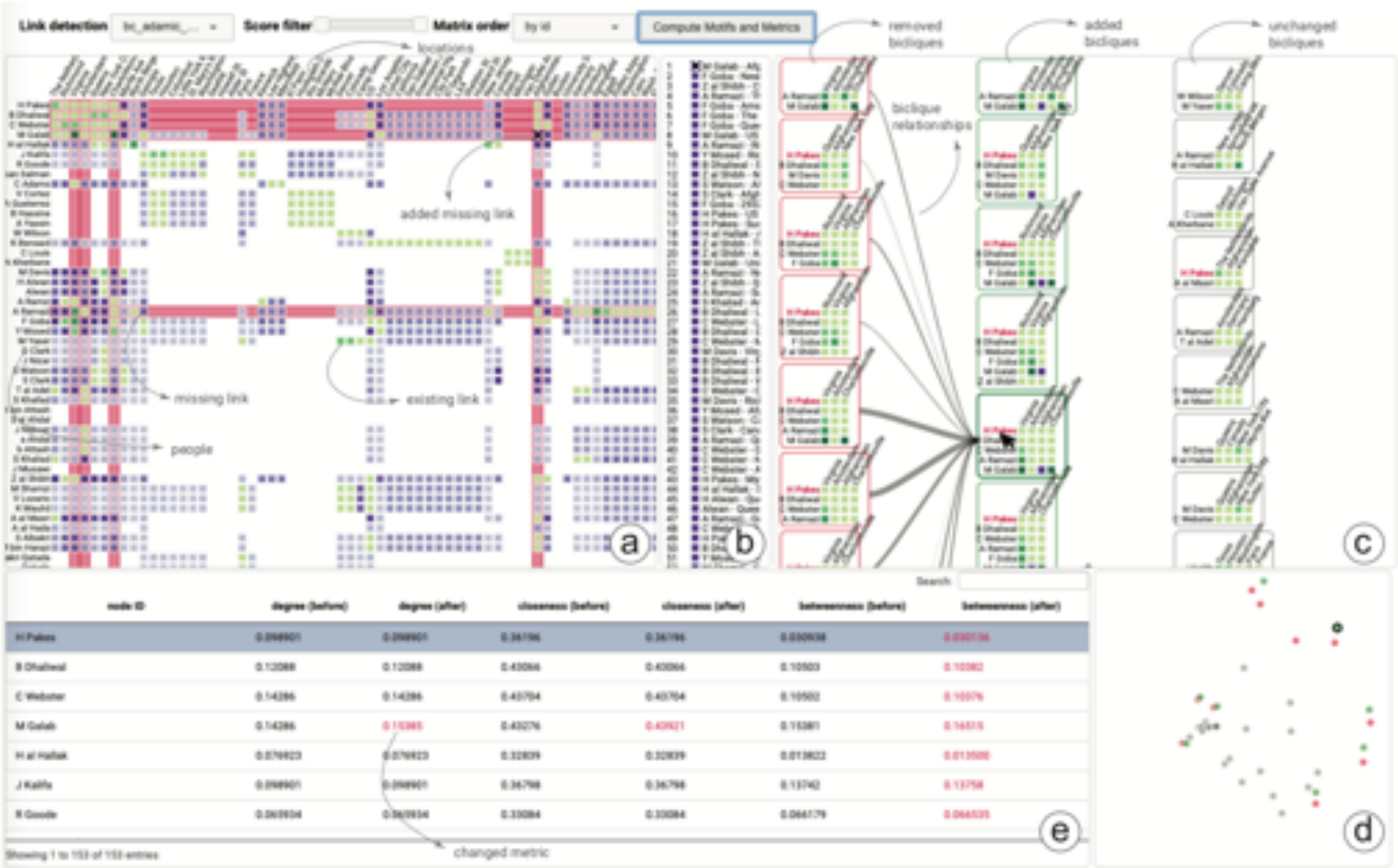
Why
is a link missing?



An interactive visualization

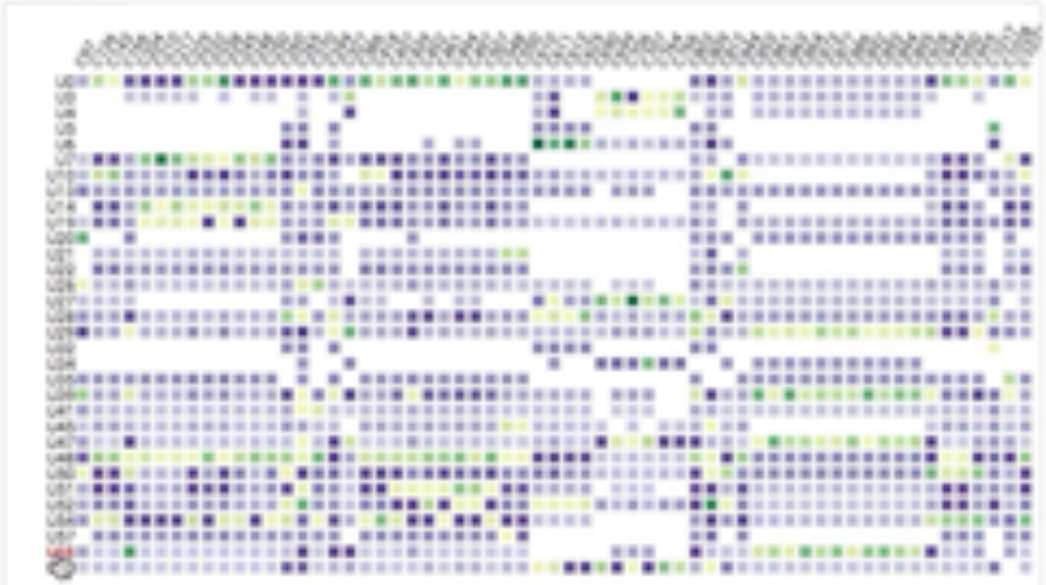
How
does a missing link impact?

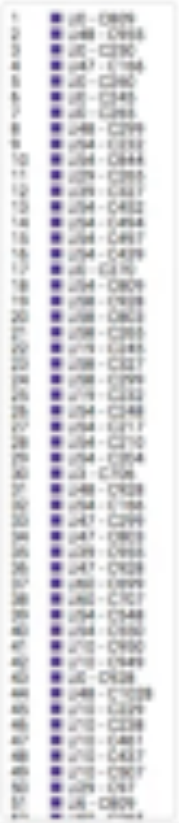
A comparative analysis approach


Link detection
Score filter
Matrix order
by id
Compute Motifs and Metrics
locations
added missing link
missing link
existing link
people
a
b
removed bicliques
added bicliques
unchanged bicliques
biclique relationships
c
Search
node ID
degree (before)
degree (after)
closeness (before)
closeness (after)
betweenness (before)
betweenness (after)
H Pakes 0.098901 0.098901 0.36196 0.36196 0.030938 0.030134
B Dhaliwal 0.12088 0.12088 0.40066 0.40066 0.10503 0.10382
C Webster 0.14286 0.14286 0.40704 0.40704 0.10502 0.10376
M Golab 0.14286 0.15385 0.40276 0.40921 0.15381 0.16515
H al Hallak 0.076923 0.076923 0.32839 0.32839 0.013822 0.013500
J Kalifa 0.098901 0.098901 0.36798 0.36798 0.13742 0.13758
R Goode 0.065934 0.065934 0.33084 0.33084 0.064179 0.064535
changed metric
e
d
Showing 1 to 153 of 153 entries

Link detection bc_adamic_... Score filter Matrix order by id Compute Motifs and Metrics

Link detection bc_atomic_... Score filter Matrix order by id Compute Motifs and Metrics



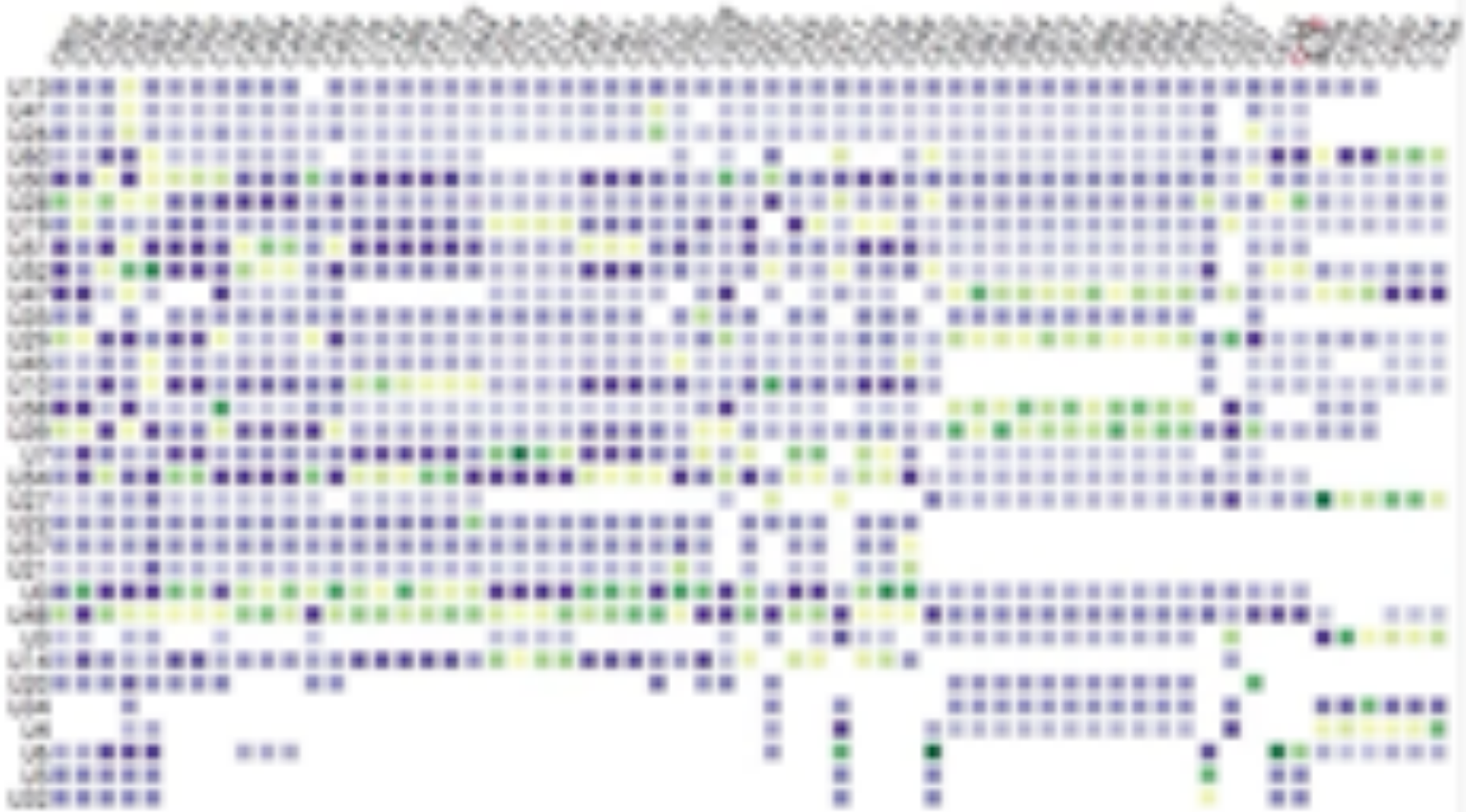


Search:

node ID	degree (before)	degree (after)	closeness (before)	closeness (after)	betweenness (before)	betweenness (after)
U57	0.016393	0.0000	0.33699	0.0000	0.0000	0.0000
U58	0.19672	0.0000	0.41695	0.0000	0.011734	0.0000
U60	0.11475	0.0000	0.44727	0.0000	0.003516	0.0000
C97	0.12500	0.0000	0.31007	0.0000	0.027018	0.0000
C136	0.12500	0.0000	0.56296	0.0000	0.0036435	0.0000
C143	0.12500	0.0000	0.56296	0.0000	0.0036435	0.0000
C166	0.15625	0.0000	0.63866	0.0000	0.040093	0.0000

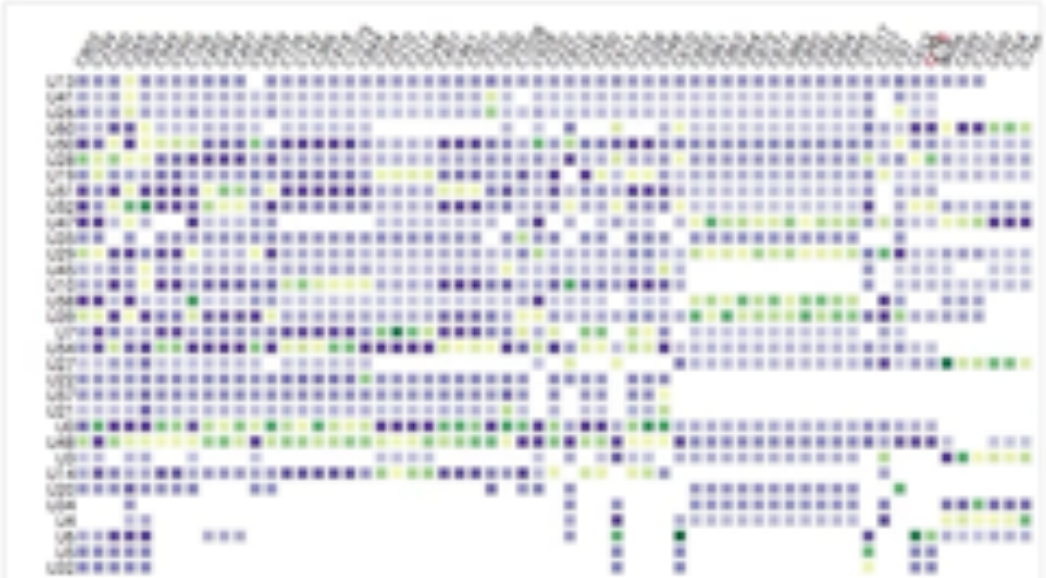
Showing 1 to 90 of 90 entries

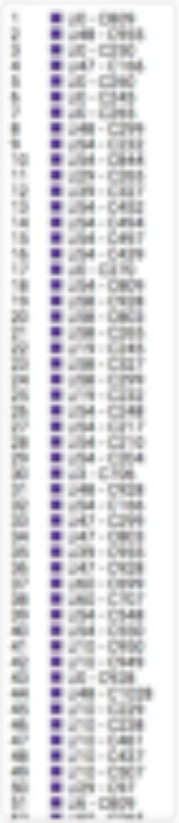
Link detection bc_atomic_... Score filter Matrix order by total-num... Compute Motifs and Metrics



Search:

Link detection bc_atomic_... Score filter Matrix order by total-num... Compute Motifs and Metrics





Search:

node ID	degree (before)	degree (after)	closeness (before)	closeness (after)	betweenness (before)	betweenness (after)
C613	0.093750	0.0000	0.46626	0.0000	0.0049100	0.0000
C625	0.093750	0.0000	0.46626	0.0000	0.0049100	0.0000
C690	0.12500	0.0000	0.43931	0.0000	0.0049347	0.0000
C699	0.12500	0.0000	0.48718	0.0000	0.012570	0.0000
C706	0.12500	0.0000	0.49673	0.0000	0.016961	0.0000
C757	0.15625	0.0000	0.49032	0.0000	0.034775	0.0000
C764	0.12500	0.0000	0.43931	0.0000	0.0049347	0.0000

Showing 1 to 90 of 90 entries

Evaluation of MissBiN

Interview study

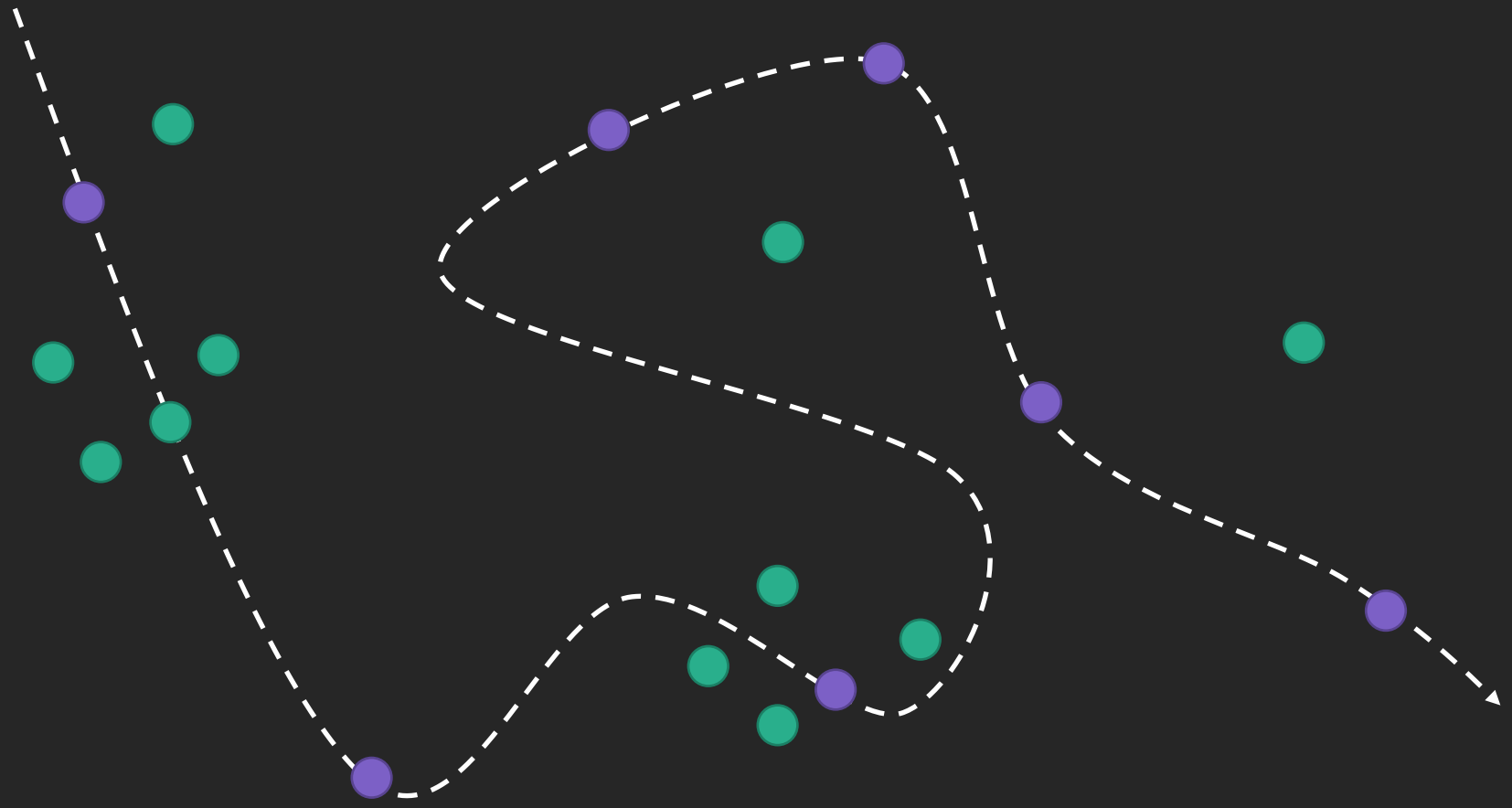
- A management school professor on exploring organizational communication networks
- A computer scientist on investigating relationships of crimes and locations in Washington DC

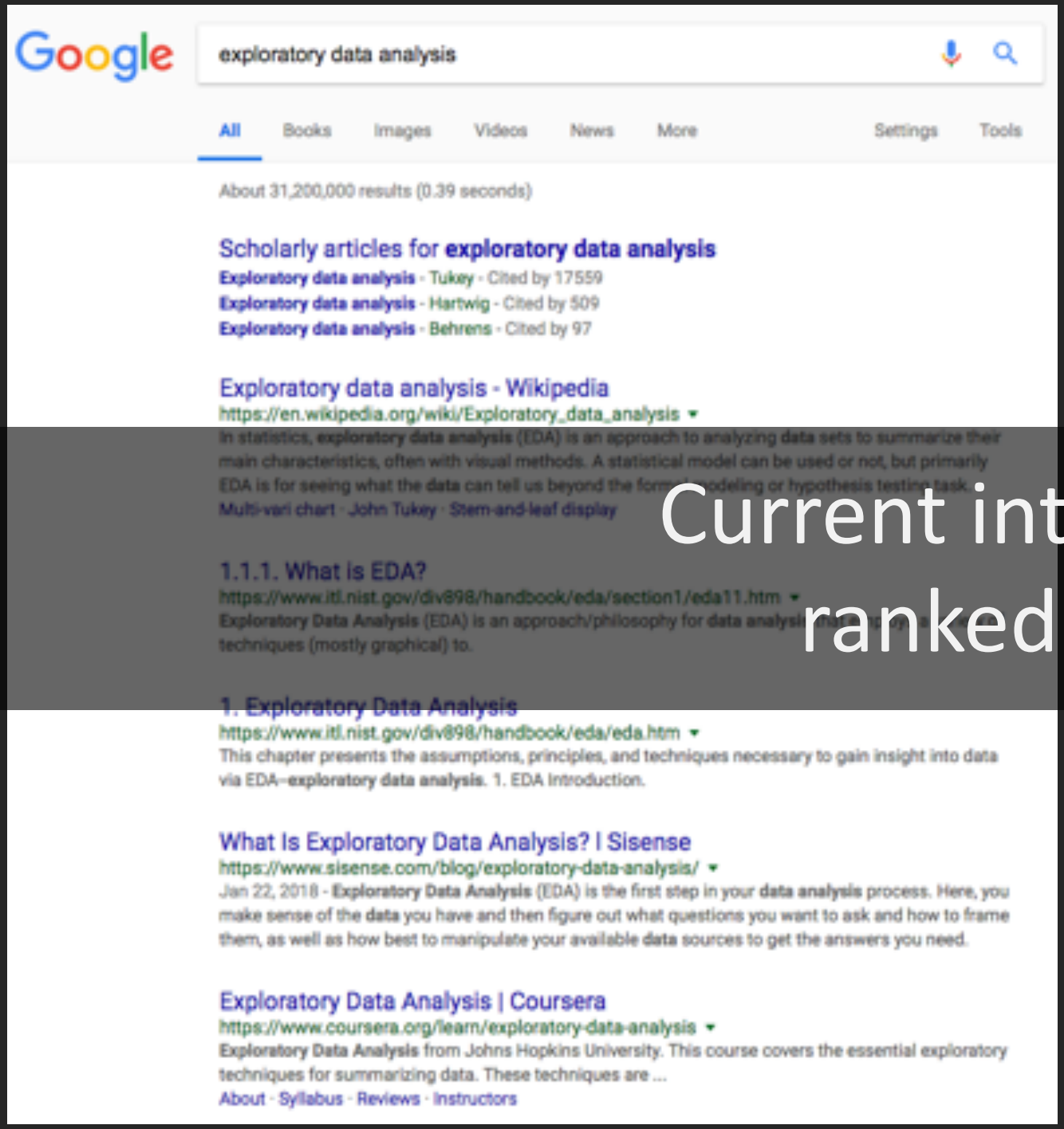
Case study

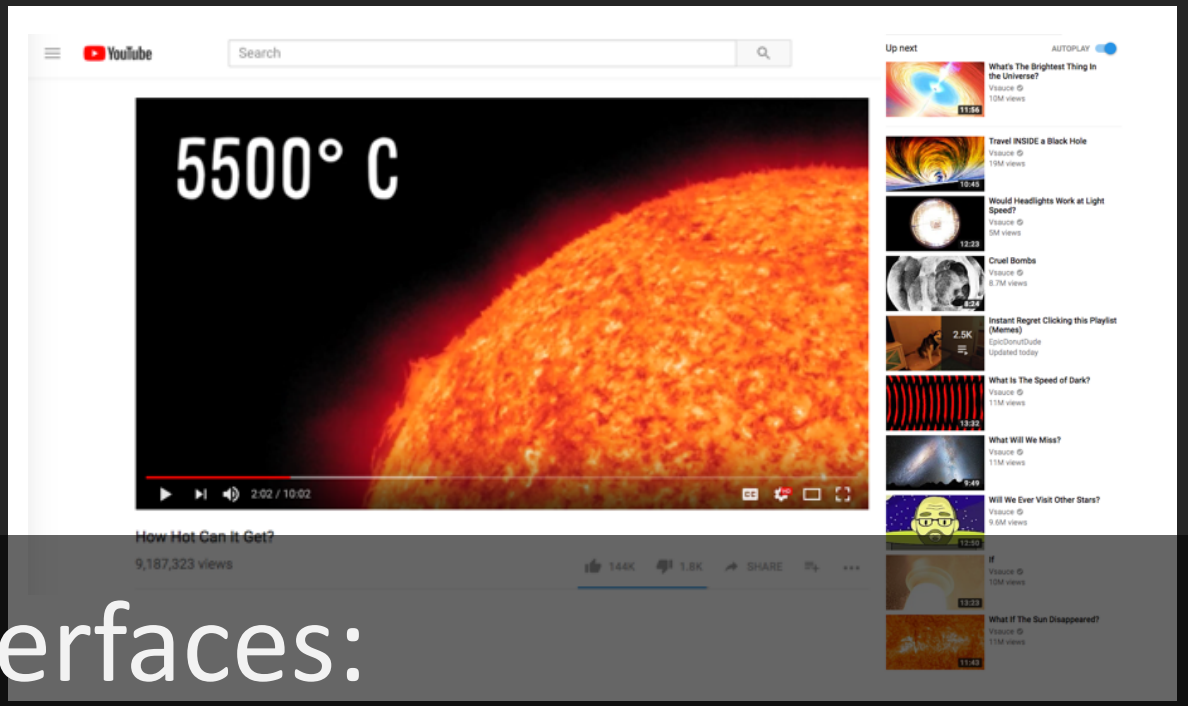
- The Sign of the Crescent [Hughes03]
  - 41 fictional intelligence reports
- Extracted person-location network
  - 49 persons and 104 locations, with 328 links
- Analysis task
  - Identify suspicious persons and activities from the reports

Make sense of results

Exploring large information space with recommendation







Current interfaces: ranked lists

Linear ranked list is not enough

Semantic map significantly improves users' comprehension capability compared to a ranked list [Peltonen 2017]

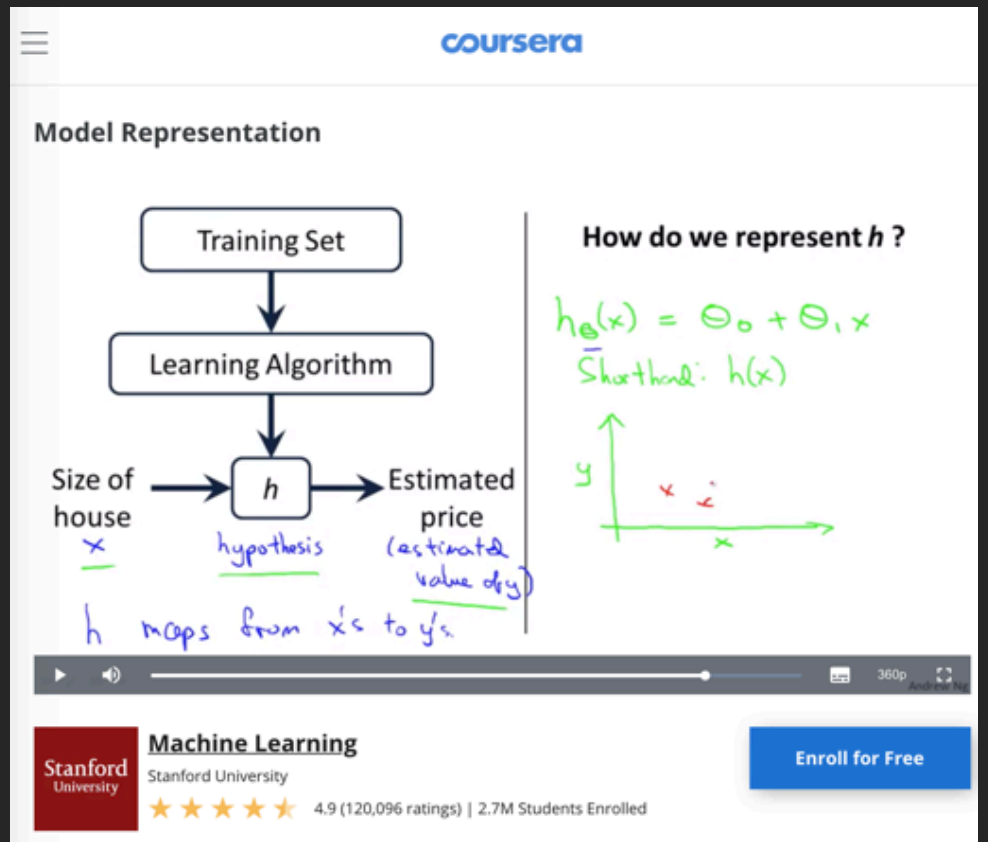
Orienteering helps understand and trust the answers using both prior and contextual information [Teevan 2004]

Support stepping behavior by clustering the information or suggesting query refinements [Teevan 2004]

Mike, the confused

Want to solve an optimization problem in his work

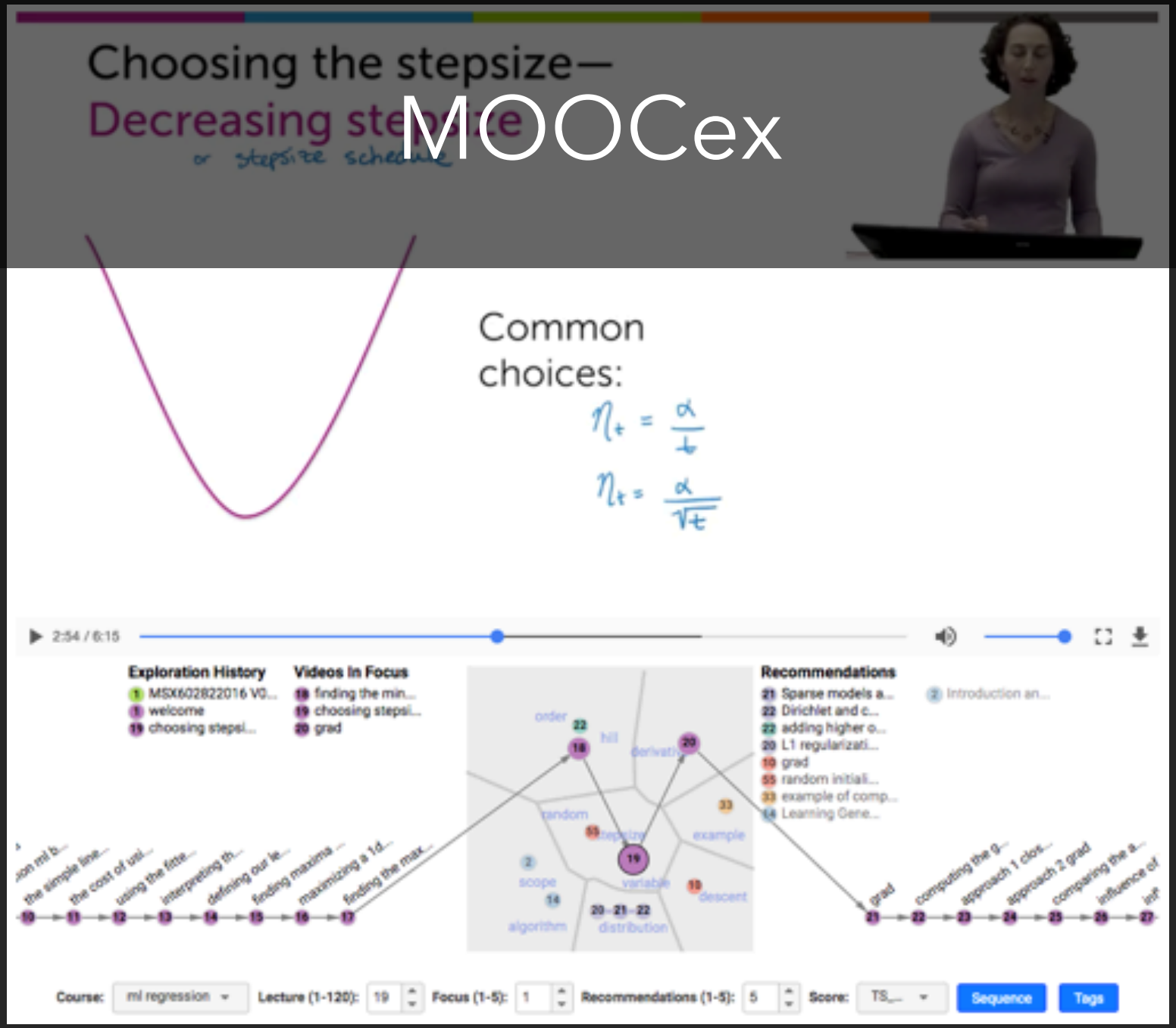
Just watched *#19 – choosing stepsize and convergence criteria*



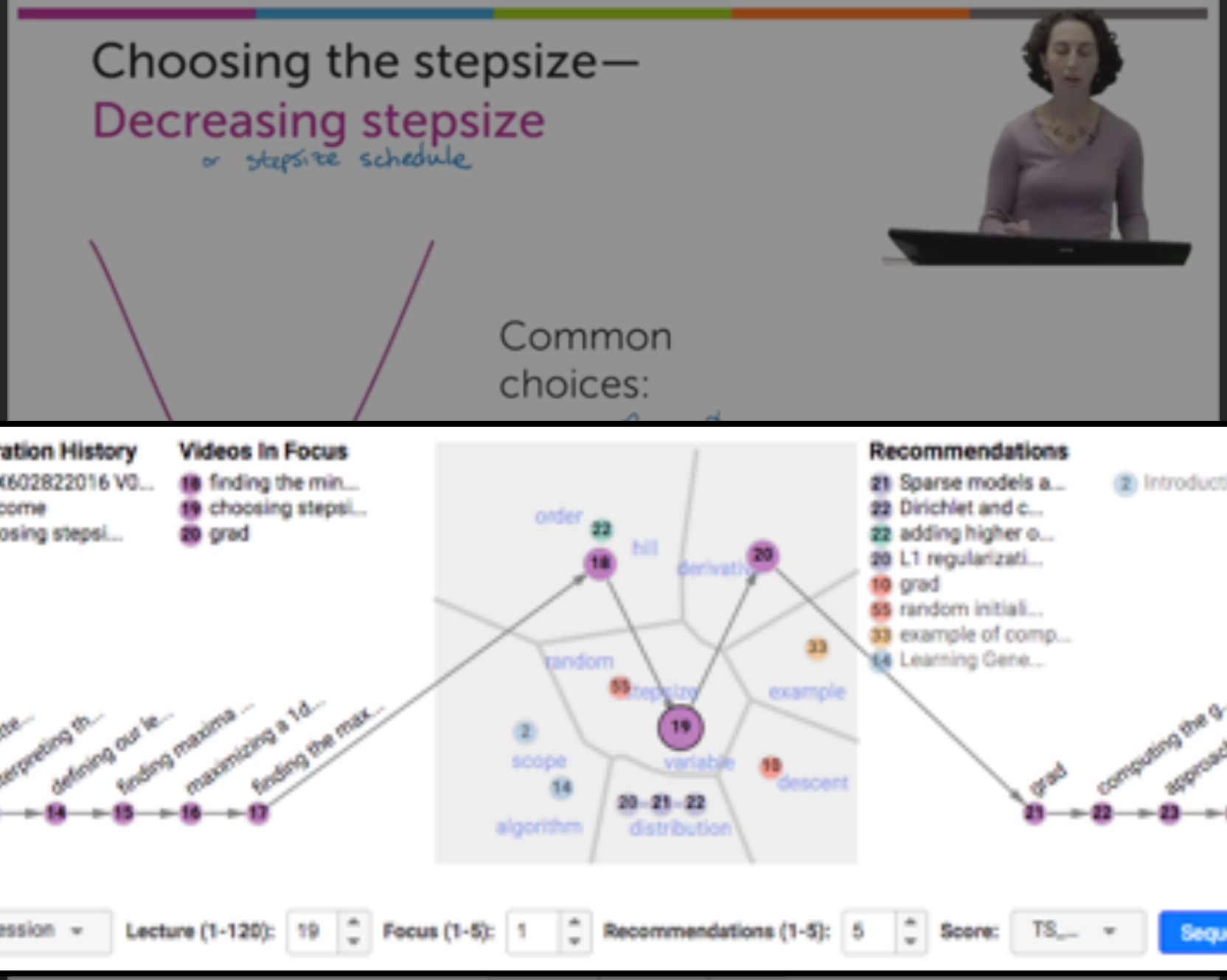
Recommendations:

1. *Sparse models selection*
2. *Dirichlet distribution*
3. *Gradient descent intuition*
4. *Hill climbing*
5. *…*

MOOCex



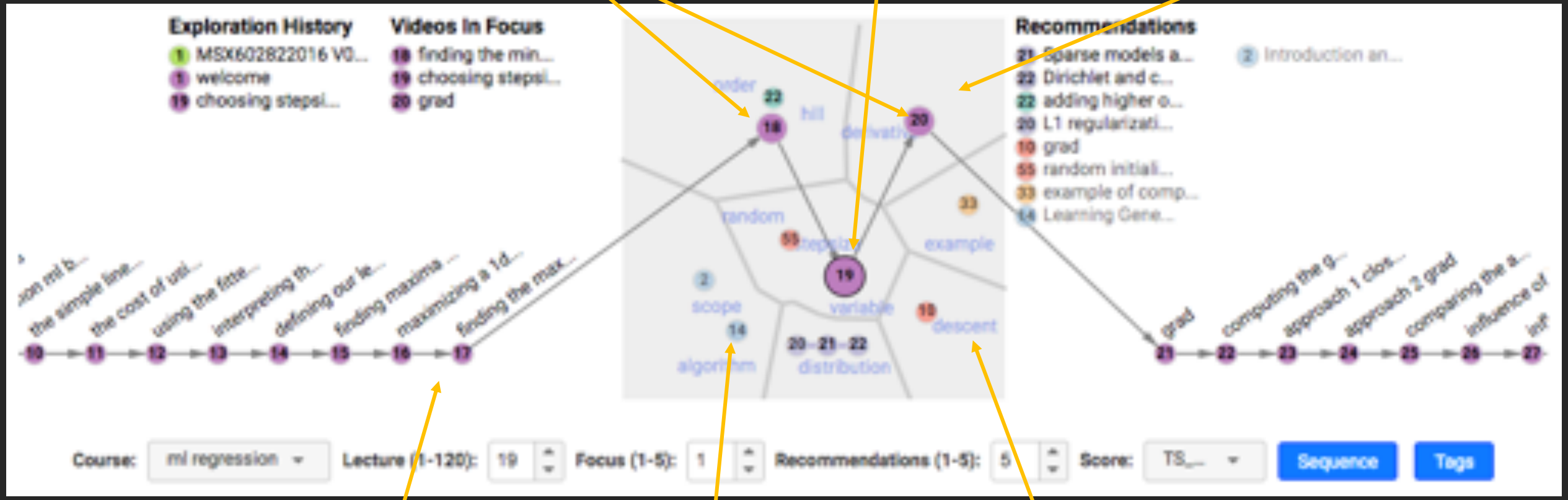
J. Zhao, C. Bhatt, M. Cooper, D. Shamma, Flexible Learning with Semantic Visual Exploration and Sequence-Based Recommendation of MOOC Videos, CHI'18


Choosing the stepsize—
Decreasing stepsize
or stepsize schedule
Common
choices:
Exploration History
1 MSX6028220160...
1 welcome
19 choosing stepsi...
Videos In Focus
18 finding the min...
19 choosing stepsi...
20 grad
Recommendations
21 Sparse models a...
22 Dirichlet and c...
22 adding higher o...
20 L1 regularizati...
10 grad
55 random initiali...
33 example of comp...
14 Learning Gene...
2 Introduction an...
order
22
hill
18
derivative
20
random
55
stepsize
19
33
example
2
scope
14
variable
13
descent
20 21 22
algorithm
distribution
the simple line...
the cost of usi...
using the fitte...
interpreting th...
defining our le...
finding maxima ...
maximizing a 1d...
finding the max...
grad
computing the g...
approach 1 clos...
approach 2 grad
comparing the a...
influence of
Course:
ml regression
Lecture (1-120):
19
Focus (1-5):
1
Recommendations (1-5):
5
Score:
TS...
Sequence
Tags

Neighboring videos (learning context)

Current video

Projection based on semantics and context



Current course

Recommendation

Topics & keywords (sub-region)

Convergence criteria

For convex functions,
optimum occurs when

$$\frac{dg(w)}{dw} = 0$$

In practice, stop when

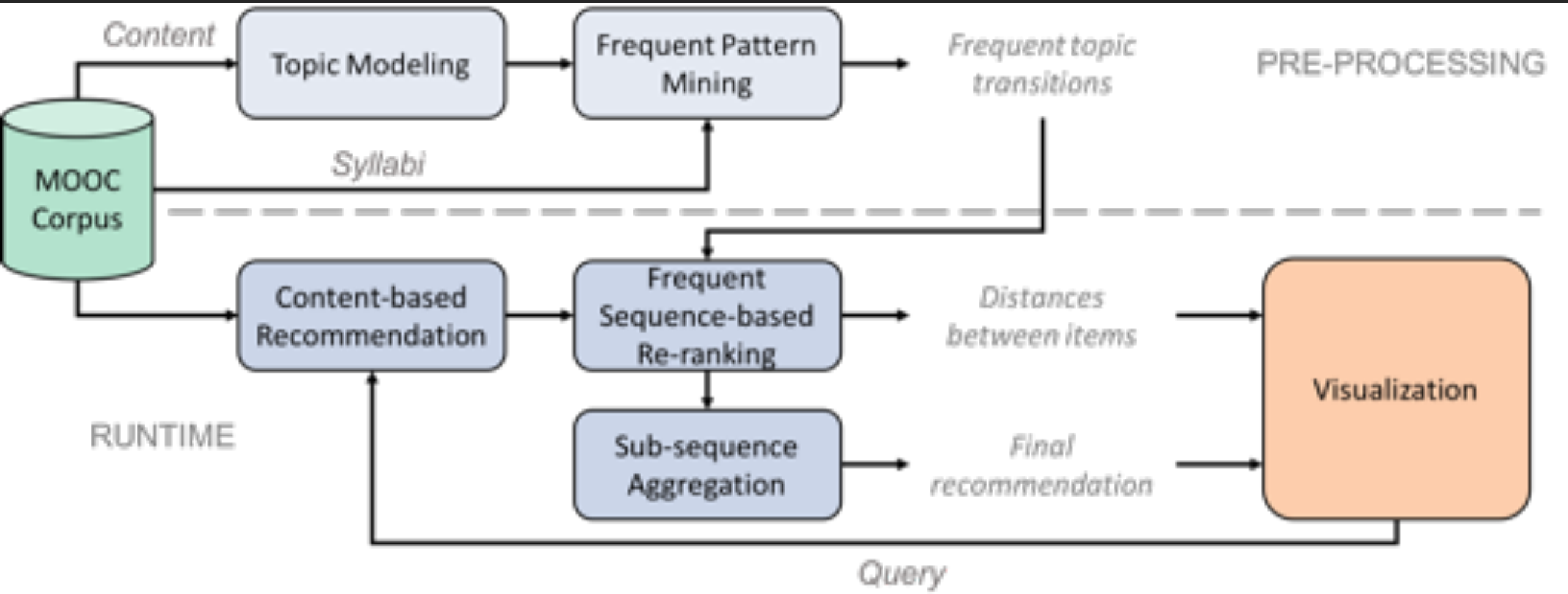
$$\left|\frac{dg(w)}{dw}\right| \text{ small}$$

Algorithm:

while not converged

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \left.\frac{dg}{dw}\right|_{w^{(t)}}$$

System architecture



Recommendation engine

Content-based recommendation
- Based on TF-IDF

Sequence-based re-ranking
- Topic similarity score (TS)
- Global sequence score (GS)
- Local sequence score (LS)

Sub-sequence aggregation
- Greedy search down the ranked list

$$S_{TS}(P^{(q)}, P^{(r)}) = \mathrm{J}(V_q, V_r) + \frac{1}{Z}\sum_{z=1}^{Z} \delta\left(|P^{(q)}(z) - P^{(r)}(z)| \leq t\right)$$

$$S_{GS}(V_q, V_r) = \frac{1}{N}\sum_{i=0}^{N} c_i \frac{y_i}{|V_r|} + \frac{s_i}{D_G}$$

$$S_{LS}(V_q, V_r) = \frac{1}{M}\sum_{i=0}^{M} c_i \frac{y_i}{|V_r|} + \frac{s_i}{D_q}$$

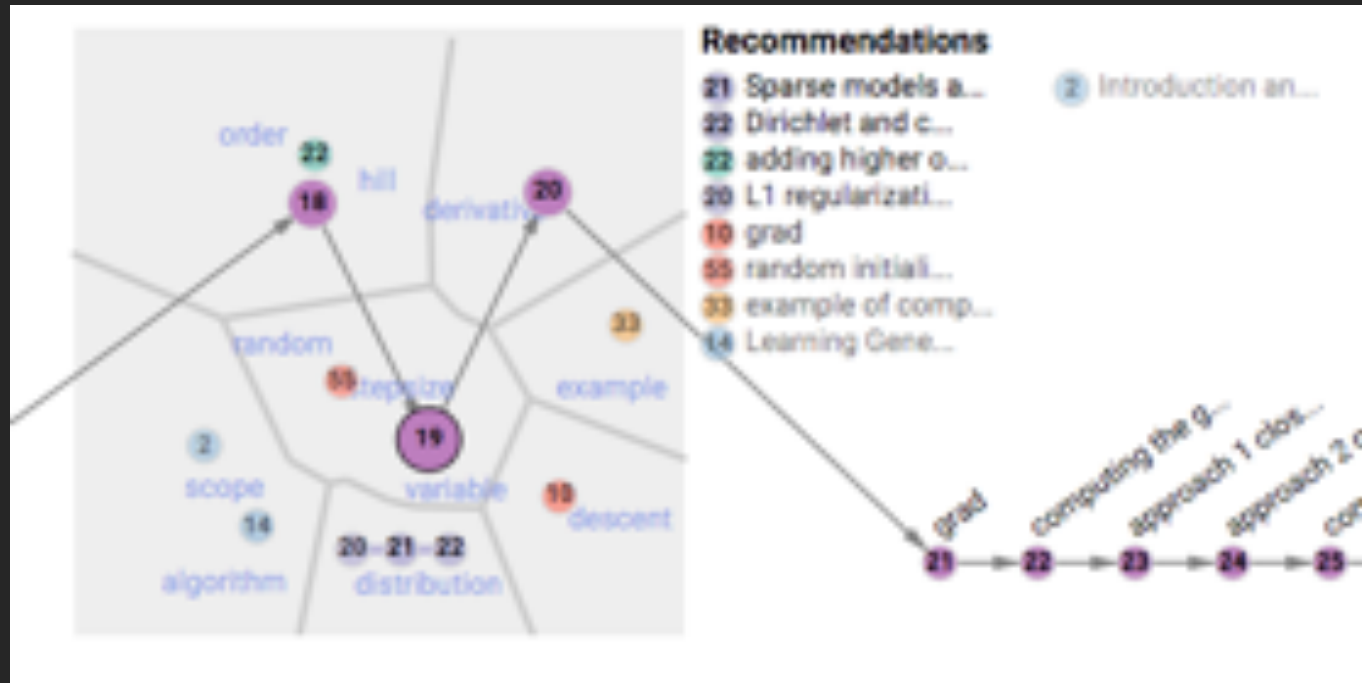
Dataset
- ~4000 videos, ~350 hours running time, from Coursera, EdX, and Udacity

Visualization generation

Multidimensional scaling (MDS) in feature space
- Rotate to comply with left-right browsing flow
- Tune positions to avoid overlap
- Merge consecutive videos

Hierarchical clustering
- Context-based region division
- Voronoi tessellation

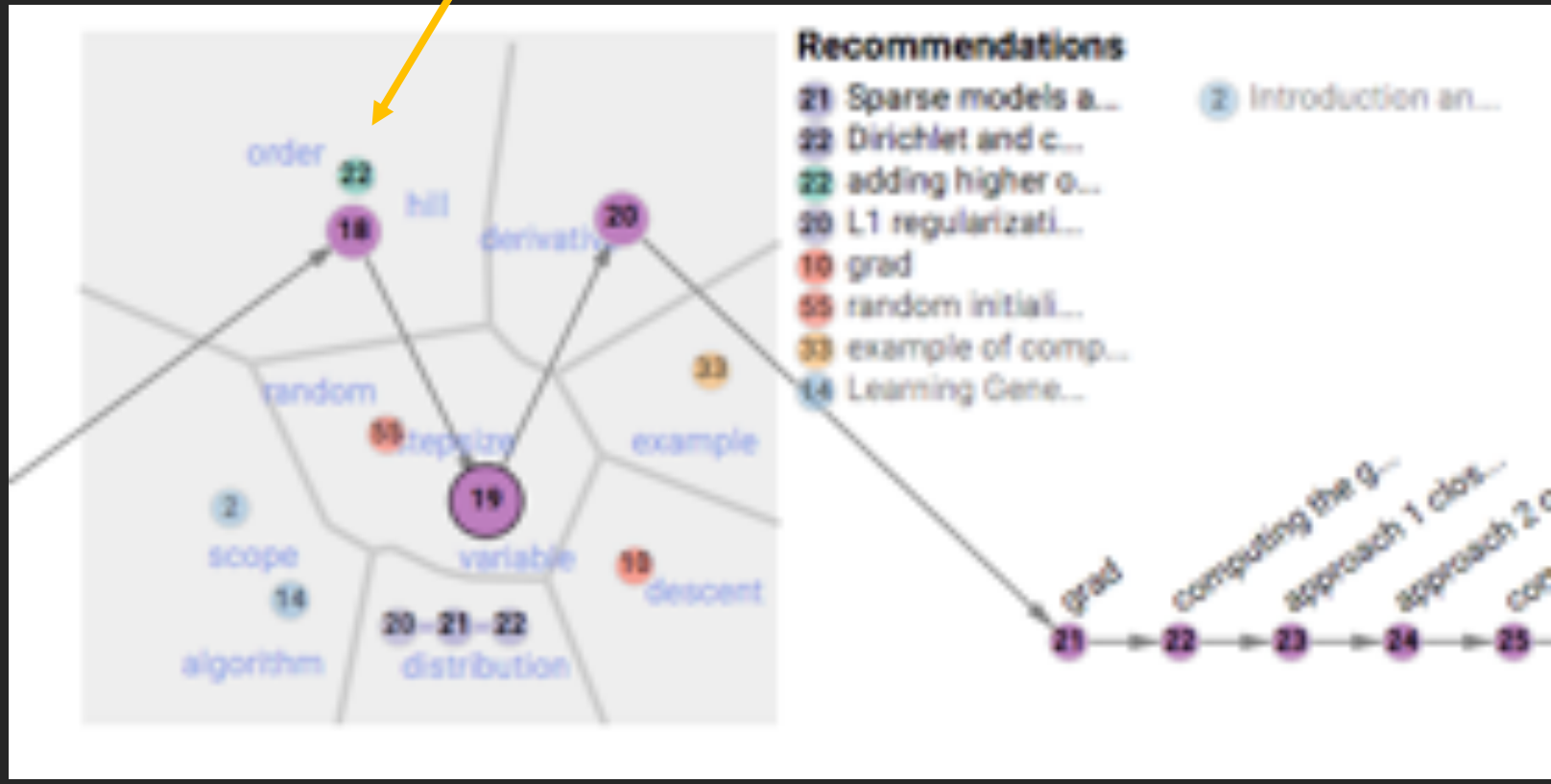
Topical keywords extraction
- Force-directed placement



Scenario I: “I missed anything?”

Mike

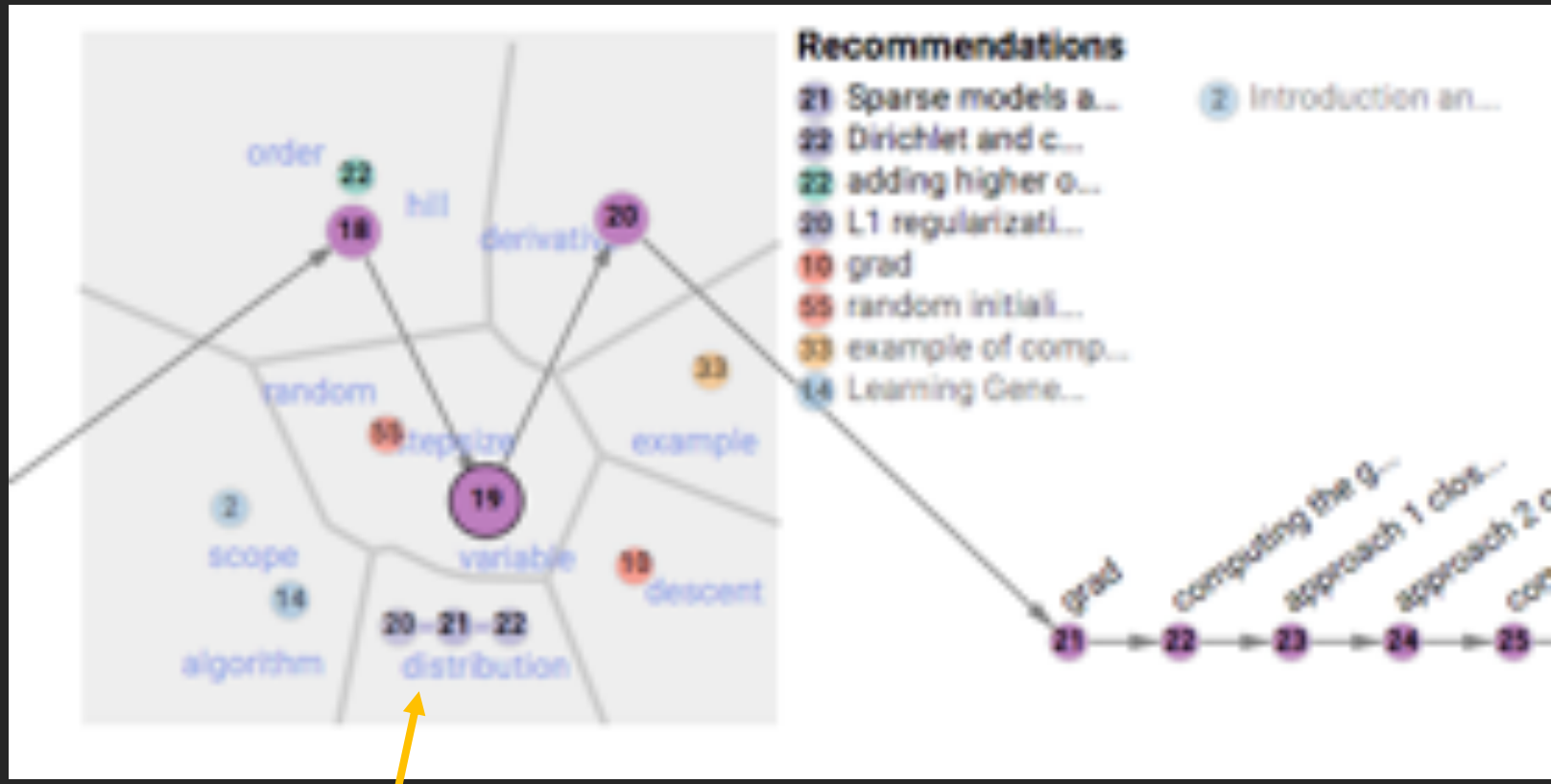
Confused about this lecture. Wants to check if missed anything.



Scenario II: “I want to know more.”

Lisa

Already knows about this.
Wants to extend her horizon.



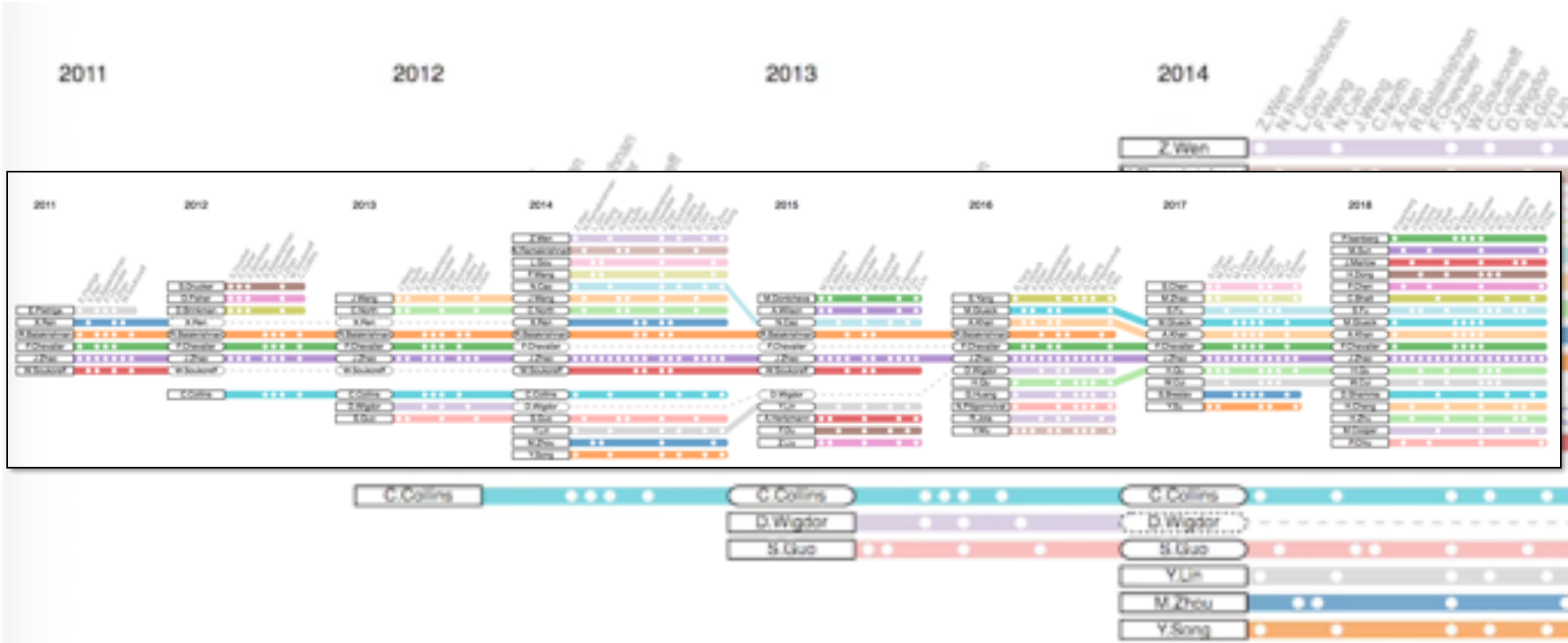
Used by MOOC instructors

Semi-structured interviews with two university instructors

> *"I normally don't look at what others teach, but the tool provides the awareness of related lectures, so I could borrow some materials to enhance my lecture, and avoid unnecessary duplication."*
>
> *"If you see one lecture is here [on the Exploration Canvas], then you go very far for the second lecture, and back here again for the third lecture, you should really think about reordering the content presented in the videos."*

One more thing…

Thank all my collaborators!



Available on https://www.jeffjianzhao.com/webapp/EgoLines/egolines.html

Another thing…

Welcome to apply to Waterloo HCI

WATERLOO HCI
CHERITON SCHOOL OF COMPUTER SCIENCE

http://hci.cs.uwaterloo.ca/

Jeff Avery
Lecturer

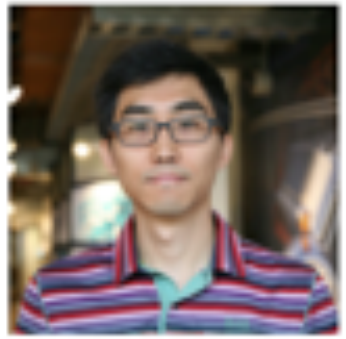
Géry Casiez
Adjunct Professor

Keiko Katsuragawa
Research Assistant Professor and Research Officer at the NRC

Edward Lank
Professor

Edith Law
Assistant Professor

Daniel Vogel
Associate Professor
interaction techniques



Jian Zhao
Assistant Professor

Integrating Predictive Models with Interactive Visualization

Jian Zhao, *Ph.D.*, Assistant Professor
Cheriton School of Computer Science
University of Waterloo
www.jeffjianzhao.com
jianzhao@uwaterloo.ca