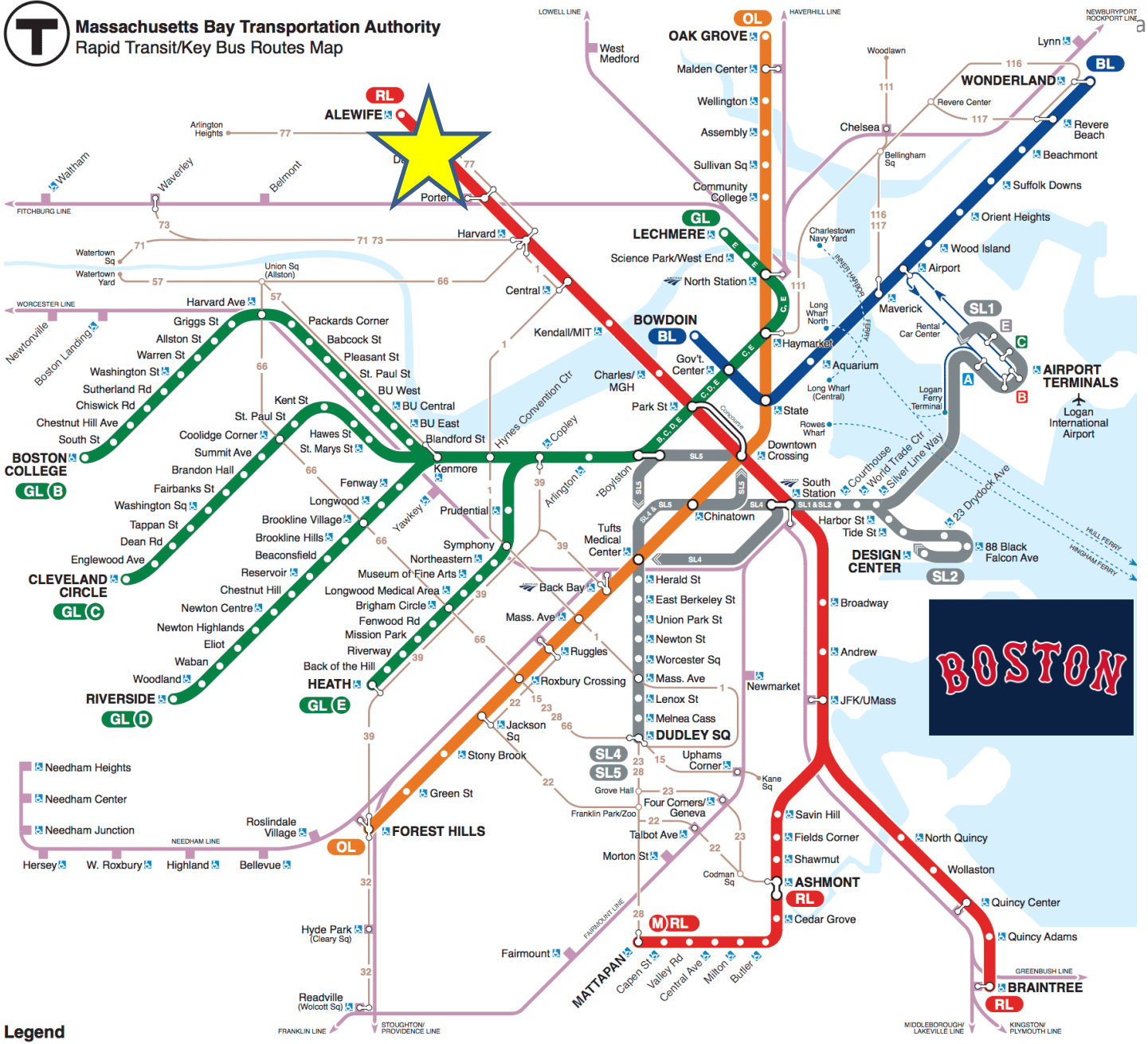
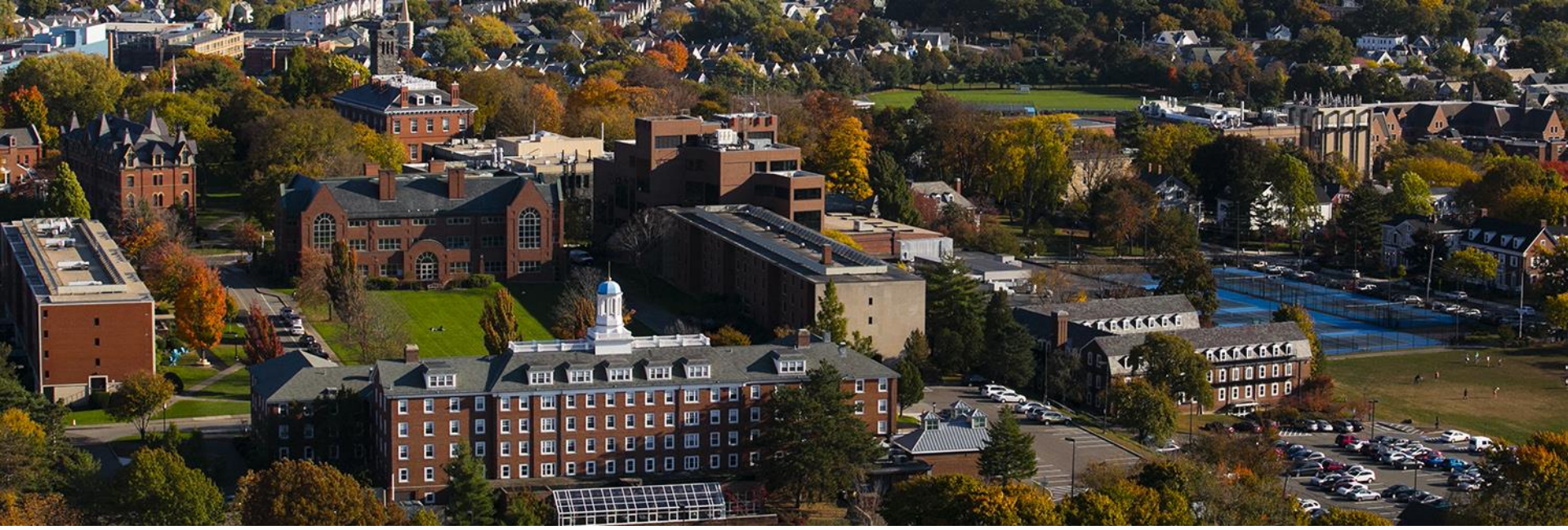
DATA SYSTEMS FOR HUMAN DATA INTERACTION

Remco Chang

Associate Professor
Computer Science, Tufts University

[v]alt


Massachusetts Bay Transportation Authority
Rapid Transit/Key Bus Routes Map
RL
ALEWIFE
OL
OAK GROVE
BL
WONDERLAND
GL
LECHMERE
BOWDOIN
AIRPORT TERMINALS
SL1
SL2
DESIGN CENTER
BOSTON COLLEGE
CLEVELAND CIRCLE
RIVERSIDE
HEATH
FOREST HILLS
DUDLEY SQ
ASHMONT
MATTAPAN
BRAINTREE
BOSTON
Legend
RED LINE
MATTAPAN LINE
ORANGE LINE
BLUE LINE
SILVER LINE and branches
GREEN LINE and branches
COMMUTER RAIL
KEY BUS ROUTE
Frequent service
FERRY
Accessible station
Rapid Transit transfer station
Commuter Rail transfer station
Free Logan Airport shuttle bus
Amtrak service
Back Bay, North & South stations
Customer Communications & Travel Info
617-222-3200, 800-392-6100,
TTY 617-222-5146, www.mbta.com
MBTA Transit Police: 911
TTY 617-222-1200
Elevator/escalator/lift updates: 800-392-6100



Tufts
UNIVERSITY

"The computer is incredibly fast, accurate, and stupid. Man is unbelievably slow, inaccurate, and brilliant. The marriage of the two is a force beyond calculation."

-Leo Cherne, 1977

(often attributed to Albert Einstein)

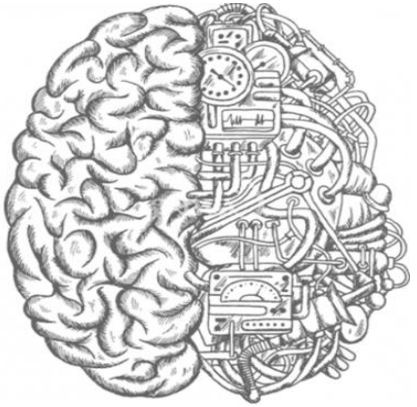
WHICH MARRIAGE?

Which Marriage?

VISUAL ANALYTICS = HUMAN + COMPUTER

- Visual analytics is "*the science of analytical reasoning facilitated by visual interactive interfaces*."[1]

- By definition, it is a collaboration between human and computer to solve data problems.



1. Thomas and Cook, "Illuminating the Path", 2005.

[v]alt

VISUAL ANALYTICS LAB AT TUFTS (VALT)


Jordan Crouser
Prof, Smith


Evan Peck
Prof, Bucknell


Lane Harrison
Prof, WPI


Eli Brown
Prof, DePaul


Beste Yuksel
Prof, USF


Alvitta Ottley
Prof, WashU


Leilani Battle
Prof, Maryland


Shah Humayoun
Prof, SF State


Michael Behrisch
Prof, Utrecht


Dan Afergan
Google Research


Mike Curry
Draper Labs


Dylan Cashman
Novartis Research


Marianne Procopio
MIT Lincoln Lab


Ab Mosca
PhD student


Ashley Suh
PhD student

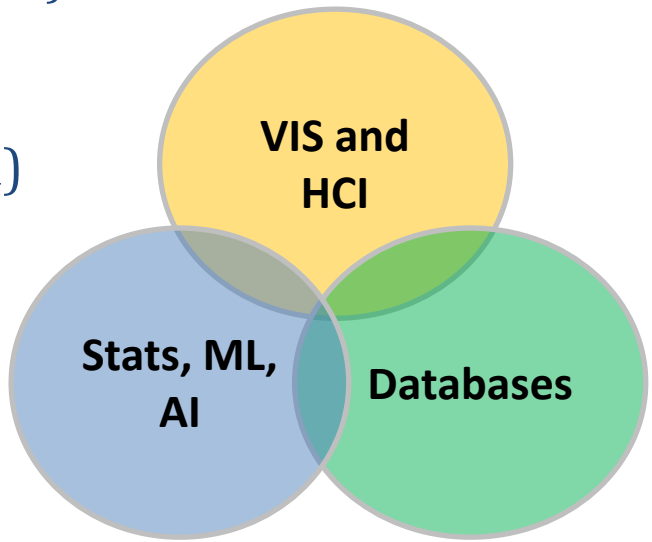

Gabriel Appleby
PhD student


Brian Montambault
PhD student


Wenbo Tao
PhD student
(MIT)

VISUAL ANALYTICS LAB AT TUFTS (VALT)

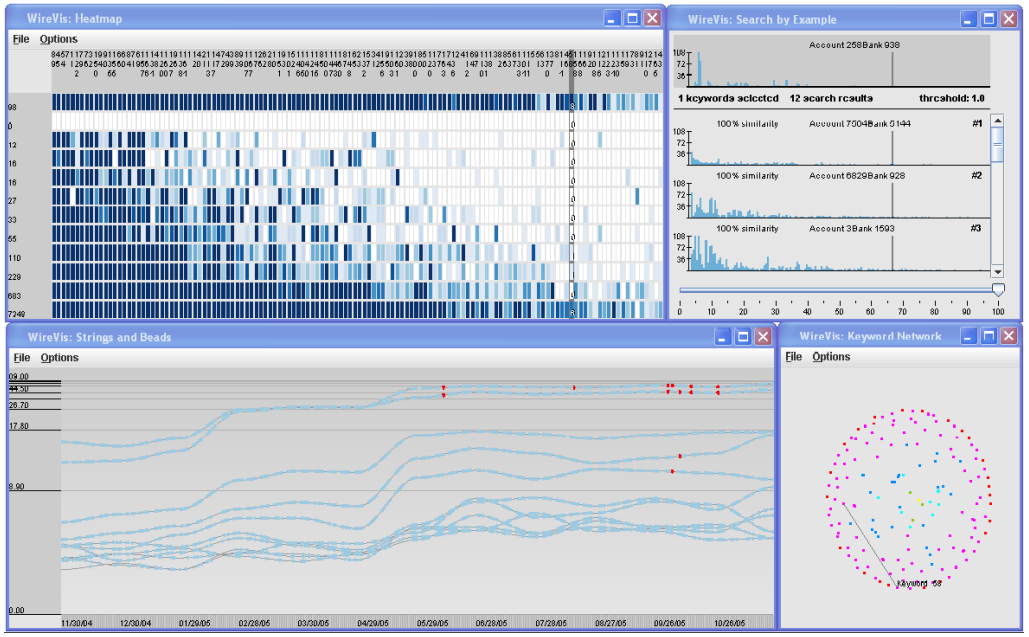
- **User Modeling** (HCI)
  - Complexity of Human Computation (Crouser)
  - Individual Differences (Ottley, Peck)
  - Perceptual Modeling (Harrison, Yang)
  - Adaptive Systems and BCI (Yuksel, Afergan)
  - Decision Making (Mosca, Ottley)
- **Interactive Machine Learning** (ML)
  - Learning from User Interactions (Brown)
  - System for Model Selection (Cashman)
  - Projection Methods (Appleby)
  - Explainable Anomaly Detection (Montambault)
- **Database Systems** (DB)
  - Predictive Prefetching (Battle)
  - Progressive Data Streaming (Procopio)
  - Interactive Pan-Zoom (Tao)

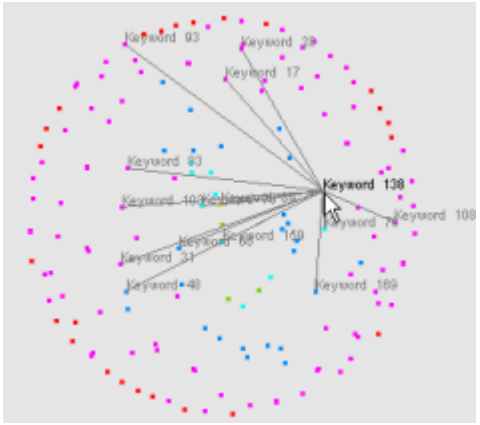


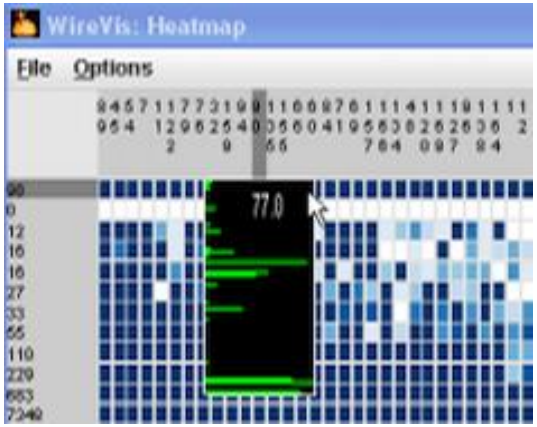
[v]alt

Motivation: Visual Analytics Systems

- Financial Fraud Analysis
  - Bank of America
- Political Simulation
  - Agent-based analysis
- Bridge Maintenance
  - Exploring inspection reports
- Biomechanical Motion
  - Interactive motion comparison
- High-D Data Analysis
  - Dis-Function: high-dimensional metric learning



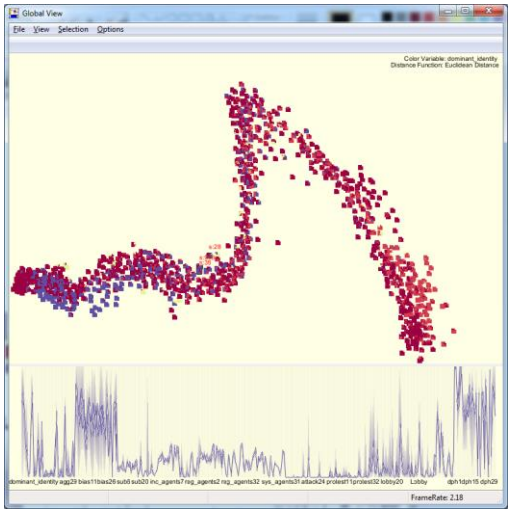


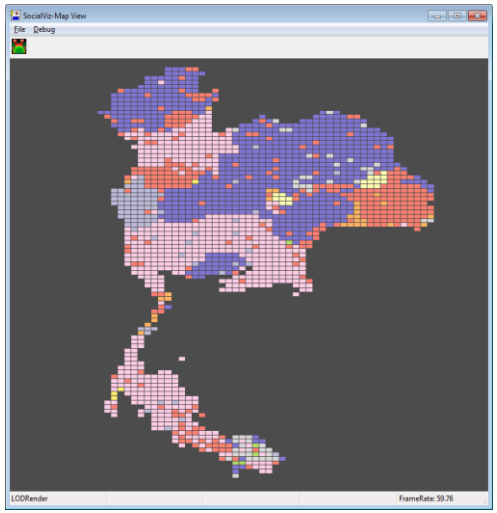


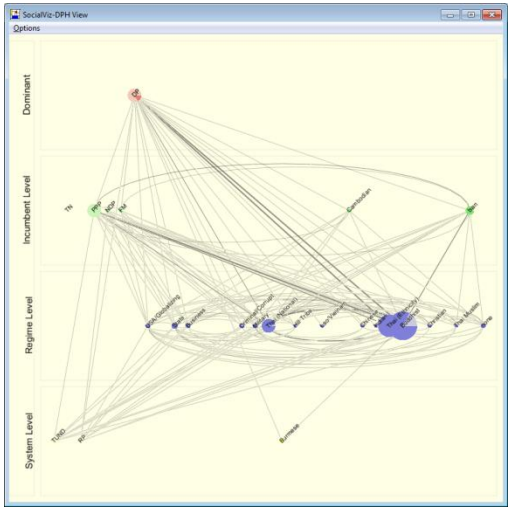
R. Chang et al., WireVis: Visualization of Categorical, Time-Varying Data From Financial Transactions. IEEE VAST, 2007

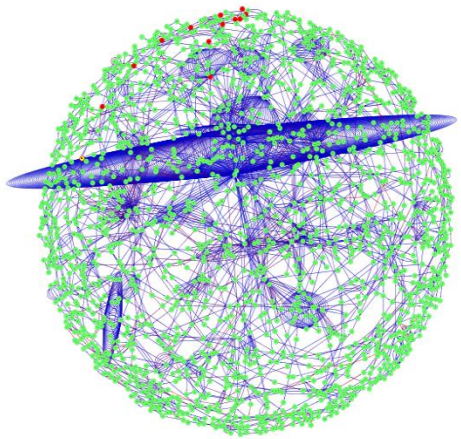
Motivation: Visual Analytics Systems

- Financial Fraud Analysis
  - Bank of America
- Political Simulation
  - Agent-based analysis
- Bridge Maintenance
  - Exploring inspection reports
- Biomechanical Motion
  - Interactive motion comparison
- High-D Data Analysis
  - Dis-Function: high-dimensional metric learning









R. Chang et al., Two Visualization Tools for Analysis of Agent-Based Simulations in Political Science. IEEE CG&A, 2012

[v]alt

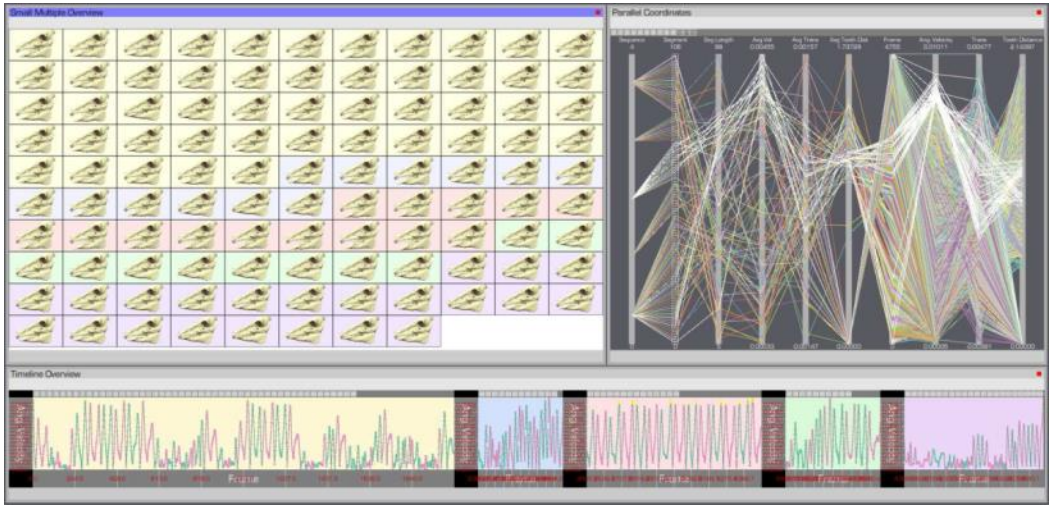
Motivation: Visual Analytics Systems

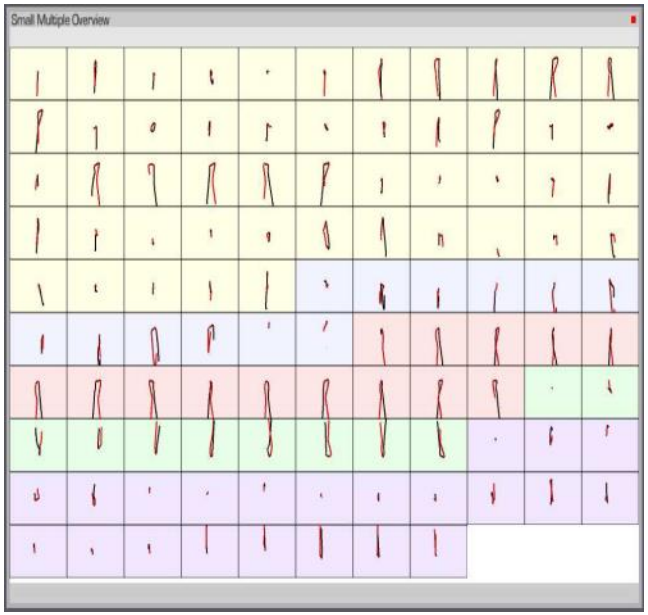
- Financial Fraud Analysis
  - Bank of America
- Political Simulation
  - Agent-based analysis
- **Bridge Maintenance**
  - **Exploring inspection reports**
- Biomechanical Motion
  - Interactive motion comparison
- High-D Data Analysis
  - Dis-Function: high-dimensional metric learning

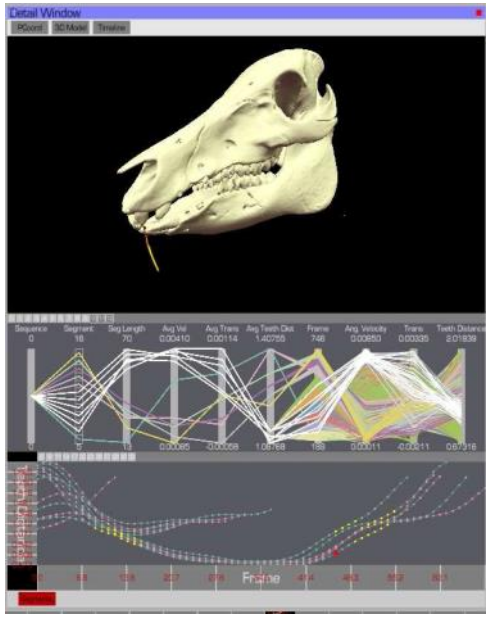
R. Chang et al., An Interactive Visual Analytics System for Bridge Management, EuroVis, 2010.

Motivation: Visual Analytics Systems

- Financial Fraud Analysis
  - Bank of America
- Political Simulation
  - Agent-based analysis
- Bridge Maintenance
  - Exploring inspection reports
- Biomechanical Motion
  - Interactive motion comparison
- High-D Data Analysis
  - Dis-Function: high-dimensional metric learning



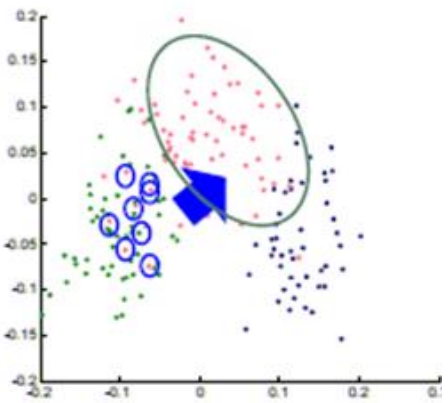


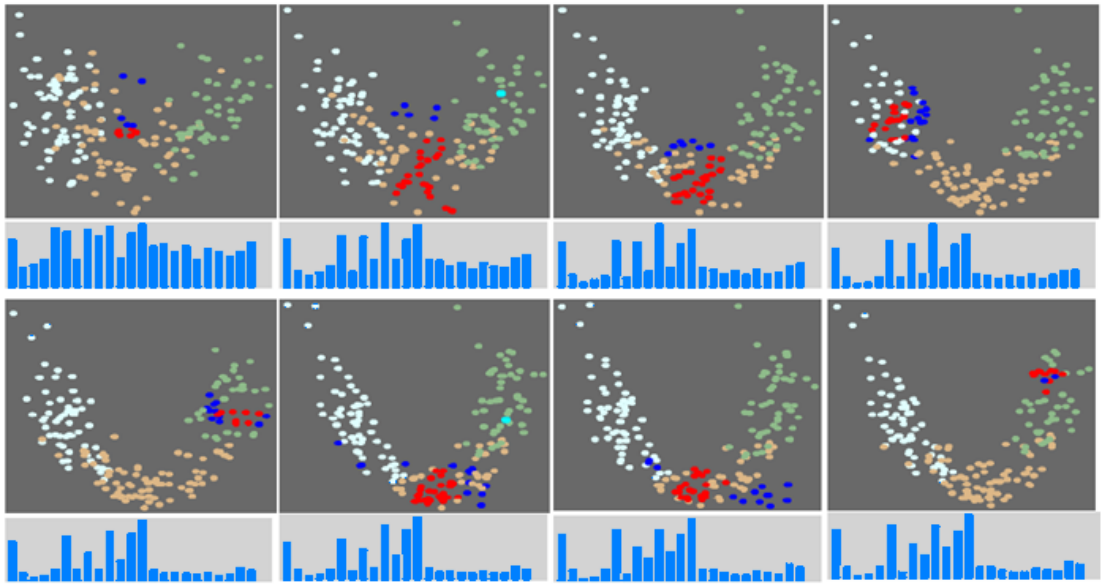


R. Chang et al., Interactive Coordinated Multiple-View Visualization of Biomechanical Motion Data, IEEE Vis (TVCG) 2009.

Motivation: Visual Analytics Systems

- Financial Fraud Analysis
  - Bank of America
- Political Simulation
  - Agent-based analysis
- Bridge Maintenance
  - Exploring inspection reports
- Biomechanical Motion
  - Interactive motion comparison
- High-D Data Analysis
  - Dis-Function: high-dimensional metric learning





R. Chang et al., Dis-function: Learning Distance Functions Interactively, IEEE VAST 2011.

Lessons Learned

- What do all good marriages have in common?

- **Collaboration**
  - Computers and humans need to contribute equally to the task at hand.
  - Data systems need to perform complex machine learning tasks with minimal guidance from the user
    - Snowcat
    - "Unprojection"
- **Communication**
  - Computers and humans need to "talk" to each other quickly and effortlessly.
  - Data systems need to respond to a user's interactions within 500ms
    - ForeCache, Kyrix
    - NeuralCubes

[v]alt


Mike Gleicher (Wisconsin)


John Stasko (GT)


Alex Endert (GT)


Dylan Cashman (Tufts)


Shah Rukh Humayoun (Tufts, SF State)


Florian Heimerl (Wisconsin)


Subhajit Das (GT)

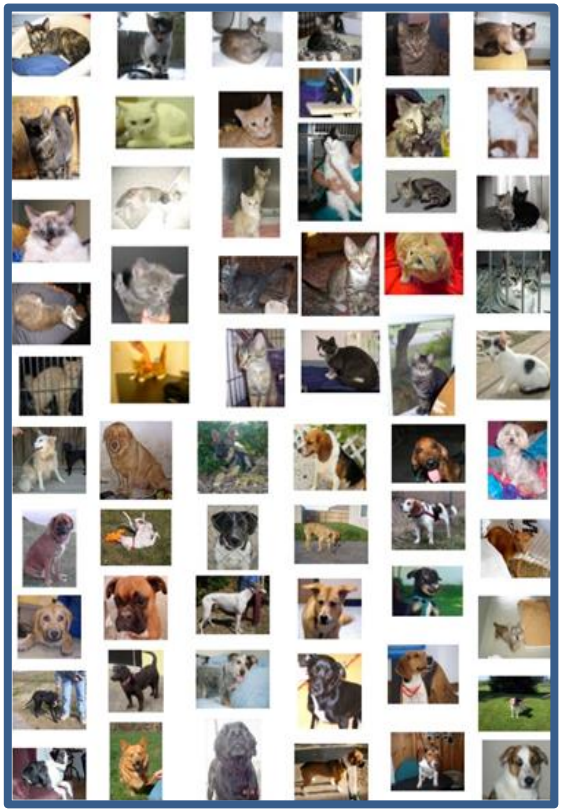

Shenyu Xu (GT)

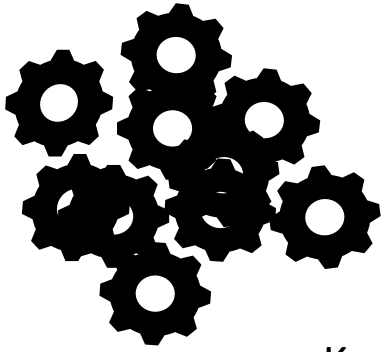
HUMAN-ML COLLABORATION:
VISUAL ANALYTICS WITH AUTOML

R. Chang et al., A User-based Visual Analytics Workflow for Exploratory Model Analysis, EuroVis 2019
R. Chang et al., Defining an Analysis: A Study of Client-Facing Data Scientists, EuroVis 2019 (Short)
R. Chang et al., Ablate, Variate, and Contemplate: Visual Analytics for Discovering Neural Architectures, VAST 2019

[v]alt

Typical ML Process





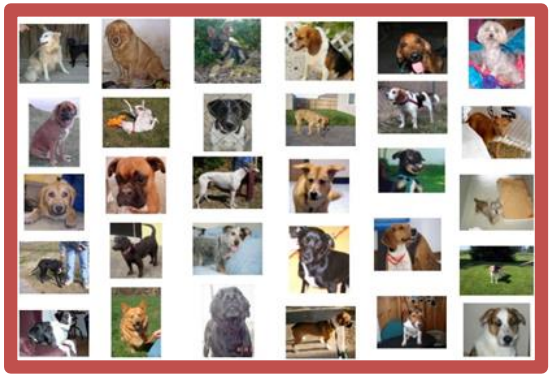
DecisionTree?

SVM?

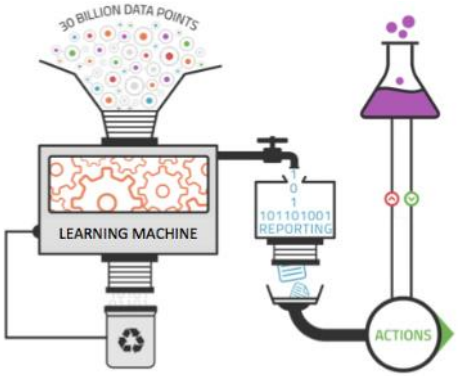
kNN?

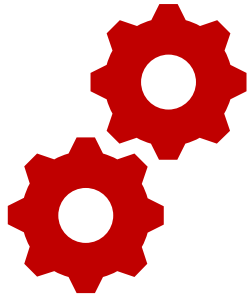
Regression?

K-means?



The AutoML Approach



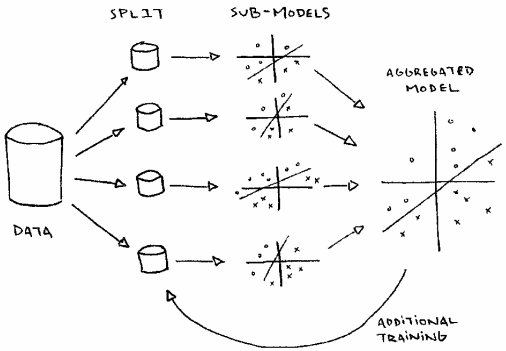


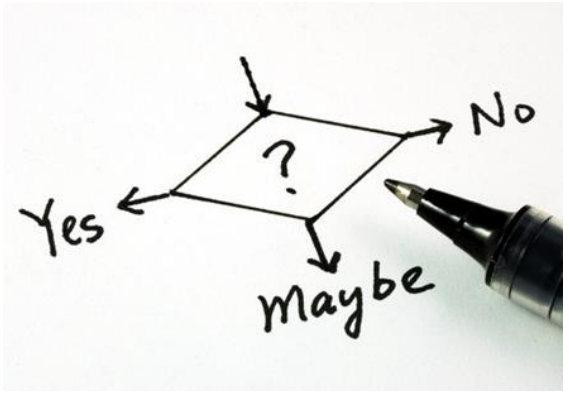
[v]alt

AutoML is Great, but…

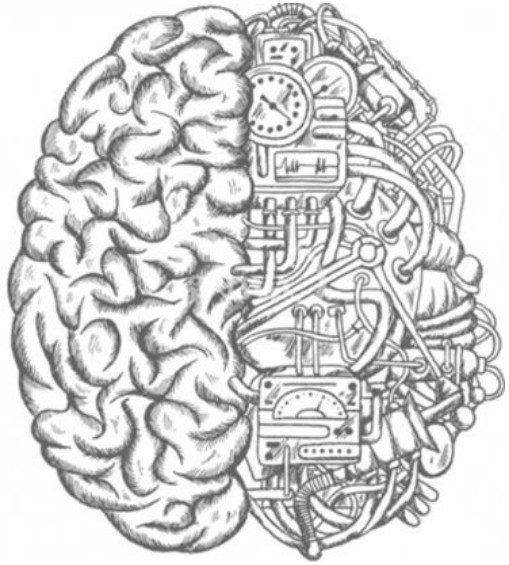
- The use of AutoML significantly reduces the effort of data modeling. However, AutoML still needs help to be useful!

1. Data Preparation
   - AutoML lacks "problem context" and "domain knowledge."
2. Query Specification
   - AutoML needs a carefully specified task. Specification is hard.
3. Model Comparison and Selection
   - AutoML will find models with the highest "score" but don't make sense (e.g. over or under fitted)





Goal: Best of Both Worlds



- Integration of AutoML with VIS:
  - Let AutoML handle all the machine learning
  - Frees the user to do what they do best

- Hypothesis:

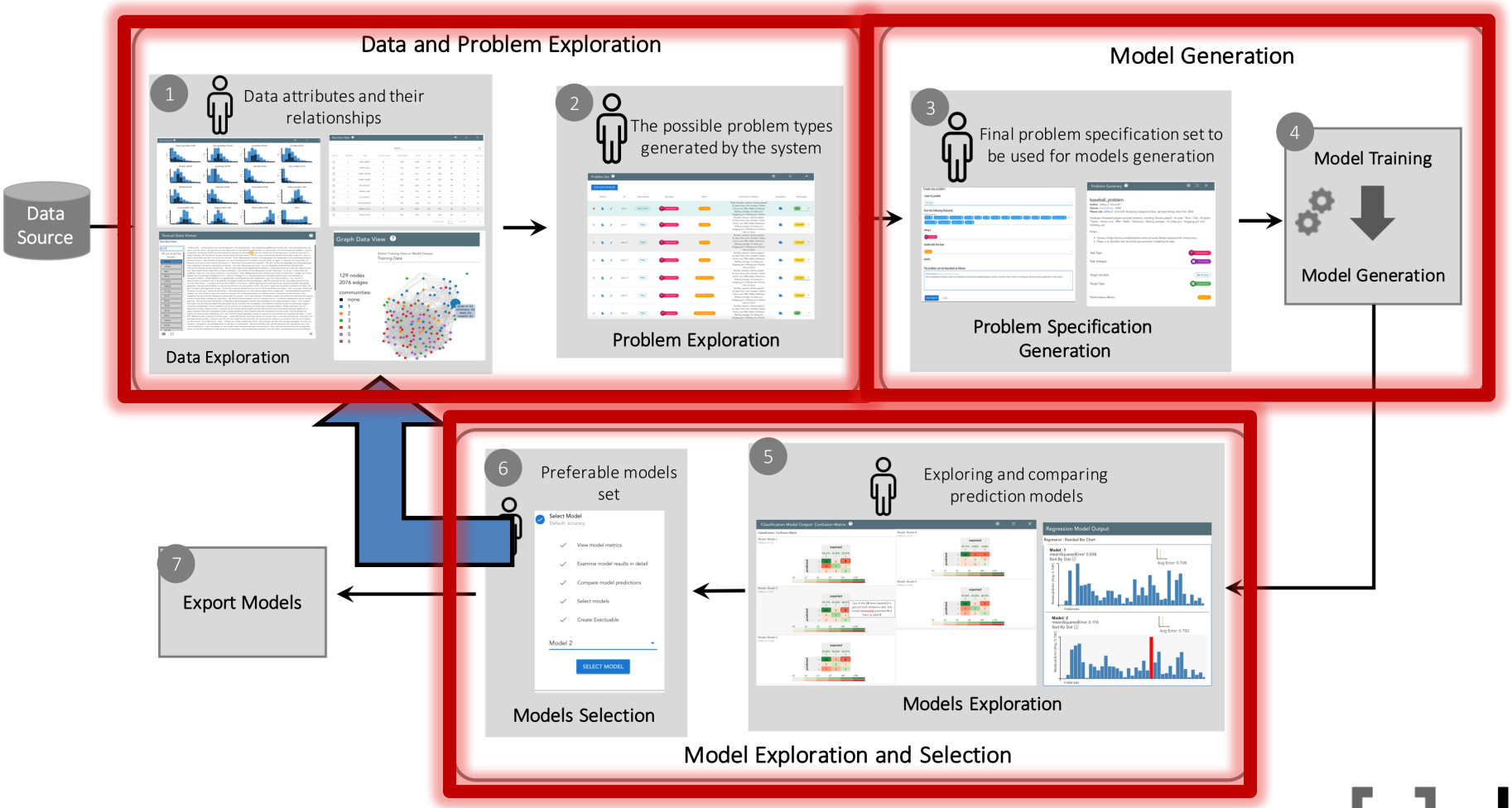
  AutoML+VIS >

  AutoML >

  Human

[v]alt

SNOWCAT: WORKFLOW



SNOWCAT

Snowcat

A visual analytics tool for exploratory model analysis

Tufts, Georgia Tech, Wisconsin-Madison

[v]alt

Evaluation of VIS + AutoML

- Part of the DARPA D3M (Data-Driven Discovery of Models) program
  - Evaluation conducted by NIST
- Evaluation:
  - AutoML > human experts
  - **Snowcat > AutoML**

Challenge 1

System	AutoML score	AutoML + VIS score
VA-1	20.2123081	21.20615752
VA-2	19.8853693	19.88951619
VA-3	19.8853693	20.58194965
VA-4	19.8853693	19.88951619
VA-5	19.8853693	19.88951619
VA-6	19.885369	273.1738095
Snowcat	19.8853693	9.440946857

Results are regression error (RMSE), lower is better

Challenge 2

System	AutoML score	AutoML + VIS score
VA-1	0.89707928	0.891516
VA-2	0.90055633	0.90055633
Snowcat	0.89707928	0.90472879

Results are classification accuracy, higher is better

[v]alt

Gabriel Appleby (Tufts)

Erik Anderson (Novartis)

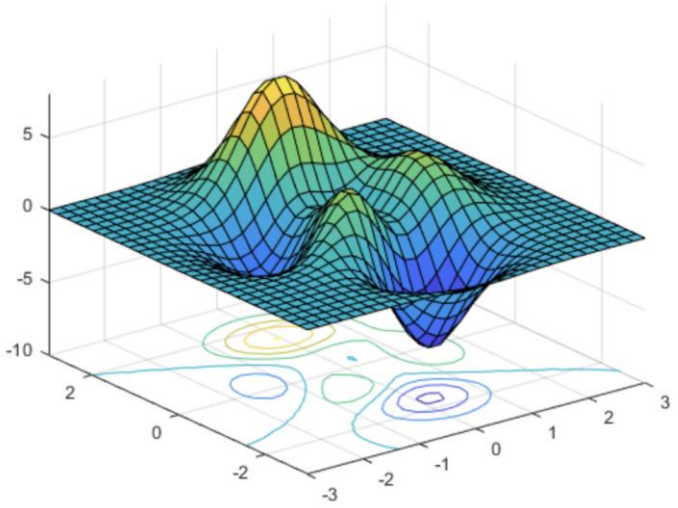
Alex Telea (Utrecht)

Mateus Espadoto (Univ of Sao Paulo)

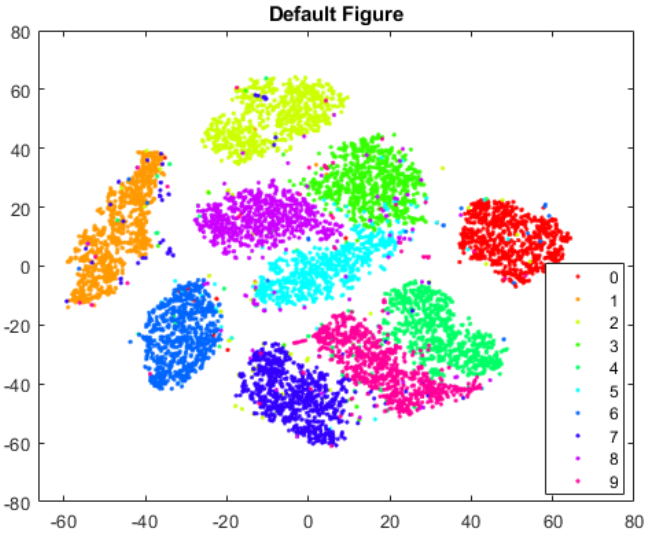
HUMAN-ML COLLABORATION: "UNPROJECTION"

Projection (Dimension Reduction)

- Projection of n-dimensional space to 2D is the *bread and butter* of visual analytics



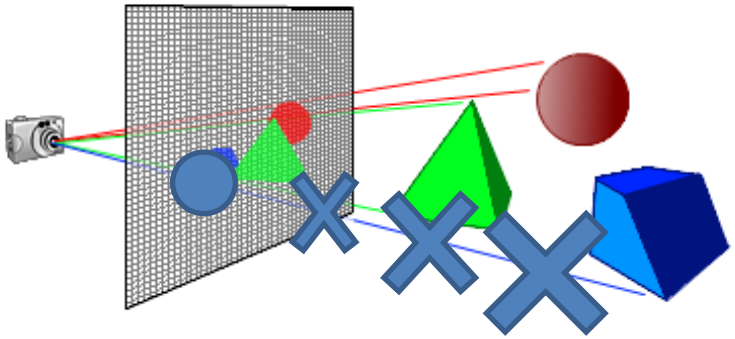
- There exist a number of techniques, both linear (PCA, MDS, Spectral) and non-linear (t-SNE, UMAP, Isomap, SOM).

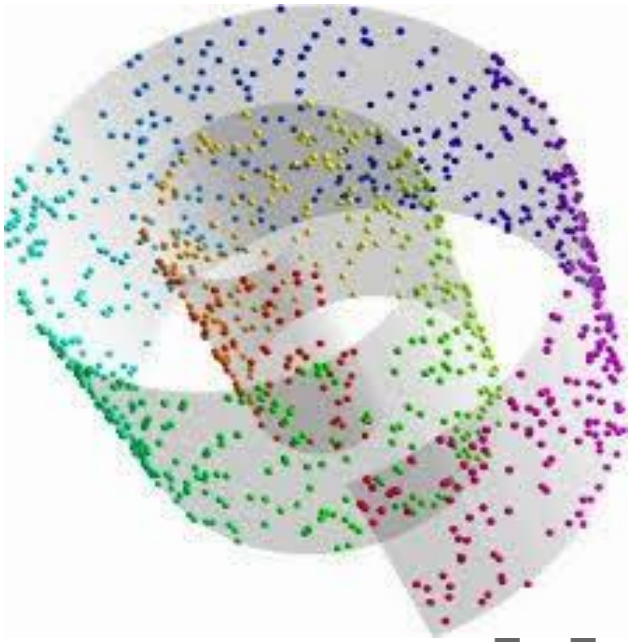


t-SNE of the mnist dataset

Projection as a 1-Way Street

- Typically, projection is considered a 1-way operation.
  - That is, we can apply a projection function to a data point $P \in R^d$ to find its position in 2D (x, y)
  - However, given a 2D position (x, y), it is difficult to recover its high-dimensional position P
- But there might be a solution!
  - Since some projections learn a 2D manifold (e.g. UMAP), there is in fact a finite number of P's given an (x, y)

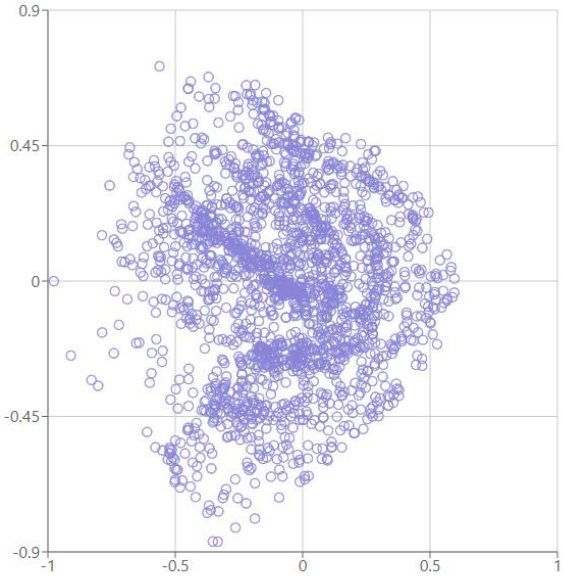


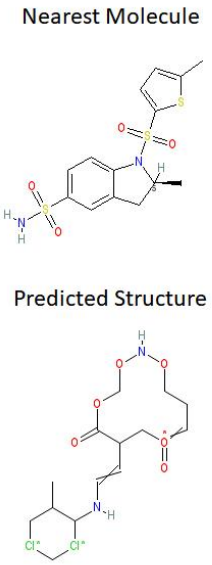


[v]alt

Projection as a 2-way Street (1/3)

- Why this could be really useful:
  - Hypothesis Generation
  - Demo



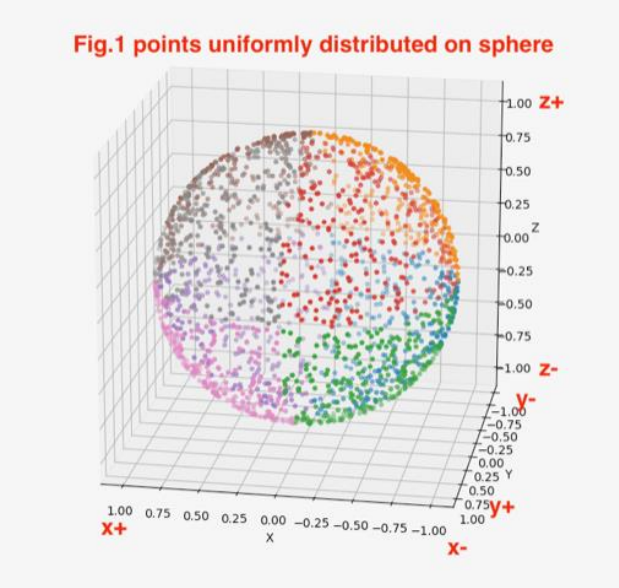


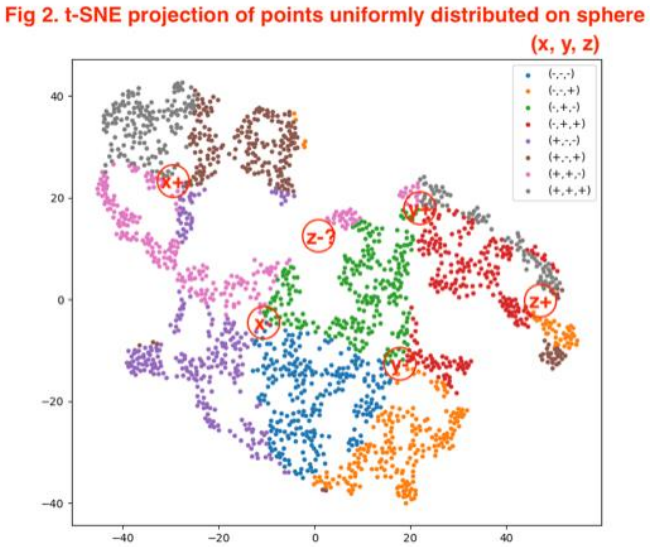
- NSF project on material synthesis
  - Each dot is a known structure
  - Prediction of material structure can save time in unnecessary syntheses (expensive and time consuming)

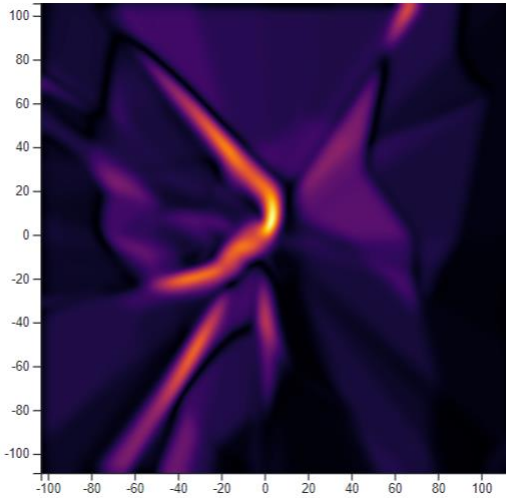
[v]alt

PROJECTION AS A 2-WAY STREET (2/3)

- Visualizing the manifold learned by a projection function (e.g. t-SNE)



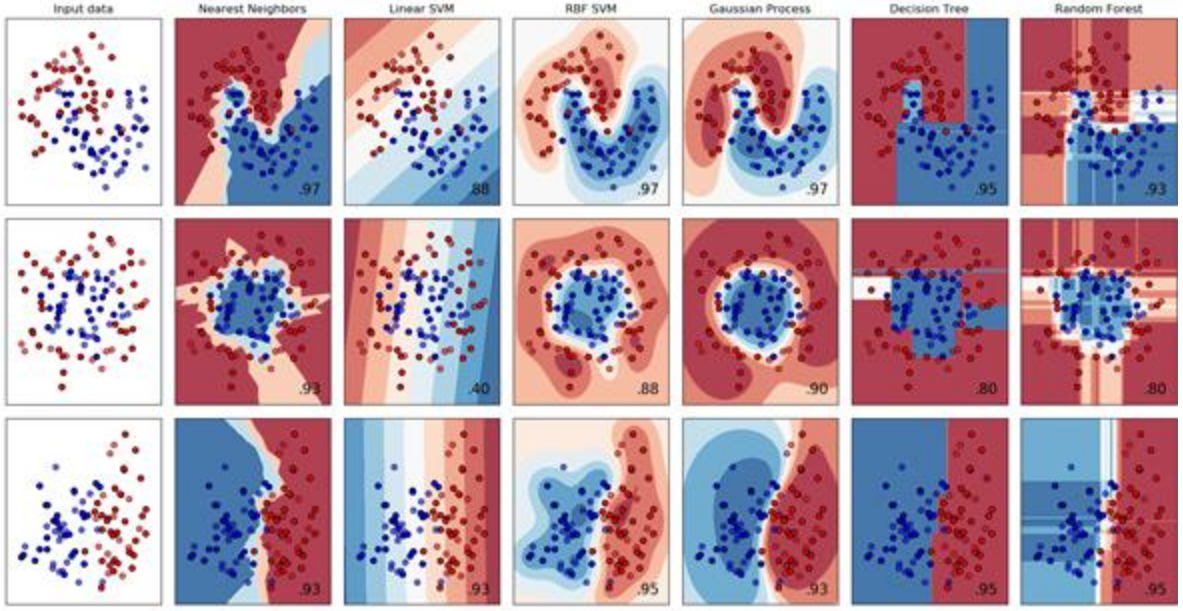


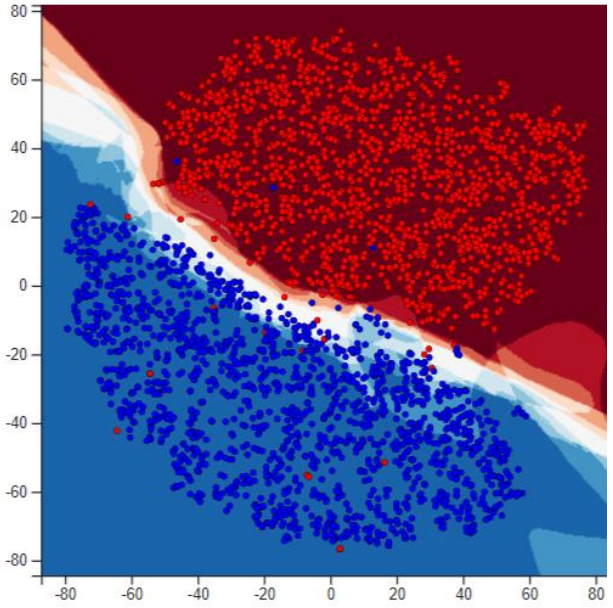


(Middle): t-SNE projection of a 3D sphere is like "peeling an orange", but it's hard to see where the boundaries are

PROJECTION AS A 2-WAY STREET (3/3)

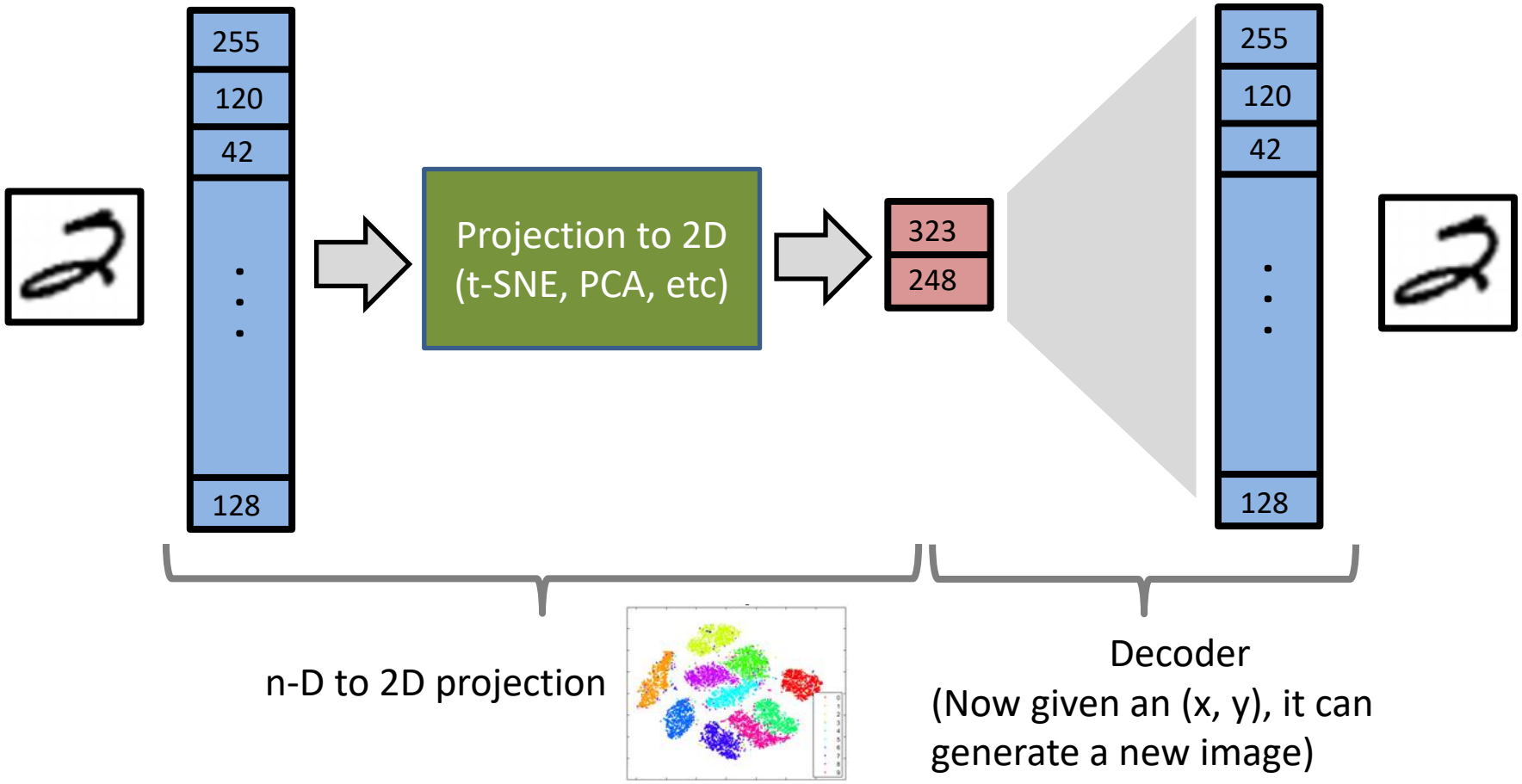
- Visualizing decision boundaries (of high-dimensional data)





*(right): shows the agreement of an ensemble of 9 classification models. Dark blue and dark red backgrounds indicate agreement between all 9 models. Lighter hues represent the amount of disagreements

How it Works: AutoEncoder

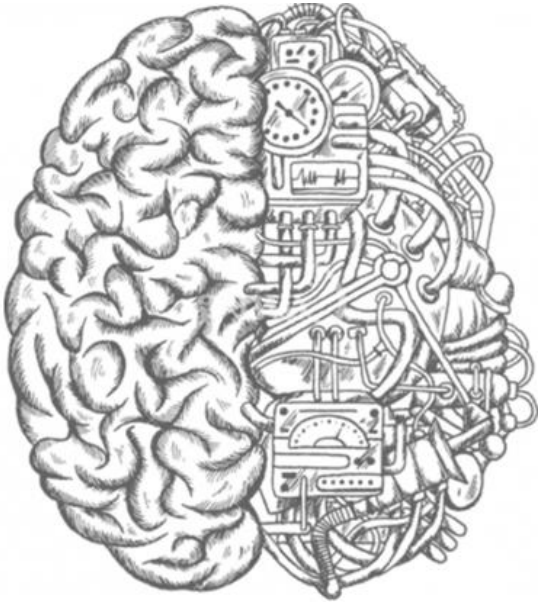


Espadoto et al. “UnProjection: Leveraging Inverse-Projections for Visual Analytics of High-Dimensional Data.” In submission to TVCG.

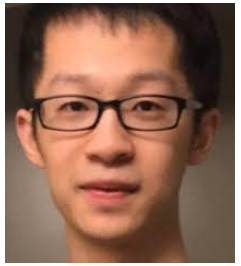
HUMAN+DATA SYSTEMS

- What do all good marriages have in common?

- **Collaboration (Human+AI)**
  - AutoML + VIS
    - Snowcat
  - Deep learning of projections
    - "UnProjection"
- **Communication (Human+DB)**
  - Computers and humans need to "talk" to each other quickly and effortlessly.
  - Data systems need to respond to a user's interactions within 500ms
    - ForeCache, Kyrix
    - NeuralCubes



[v]alt



Wenbo Tao
(MIT)

Mike
Stonebraker
(MIT)

Leilani Battle
(Maryland)

DATA SYSTEM FOR SUPPORTING INTERACTIVE DATA EXPLORATION: **DATABASE OPTIMIZATION**

R. Chang et al., Dynamic Prefetching of Data Tiles for Interactive Visualization. SIGMOD 2016

R. Chang et al., Kyrix: Interactive Pan/Zoom Visualization at Scale. EuroVis 2019

R. Chang et al., Smile: A System to Support Machine Learning on EEG Data at Scale. VLDB 2019

R. Chang et al., Kyrix-S: Authoring Scalable Scatterplot Visualizations of Big Data. InfoVis 2020

[v]alt

Interactive Exploration of Big Data

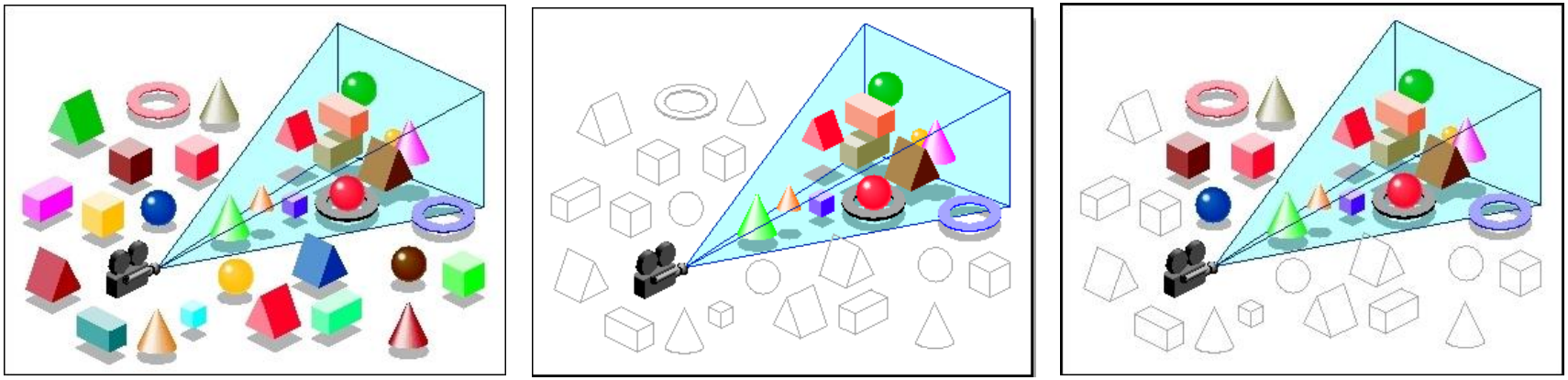
Visualization on a Commodity Hardware

Large Data in a Data Warehouse

[v]alt

FORECACHE: PREDICTIVE PRE-FETCHING

- If we know what the user might do next, we can pre-compute the task and pre-fetch the result
  - Reduces user wait time (e.g. after they click a button to perform a task)



- ForeCache: Three-tiered architecture
  - Thin client (visualization)
  - Backend (array-based database)
  - Fat middleware

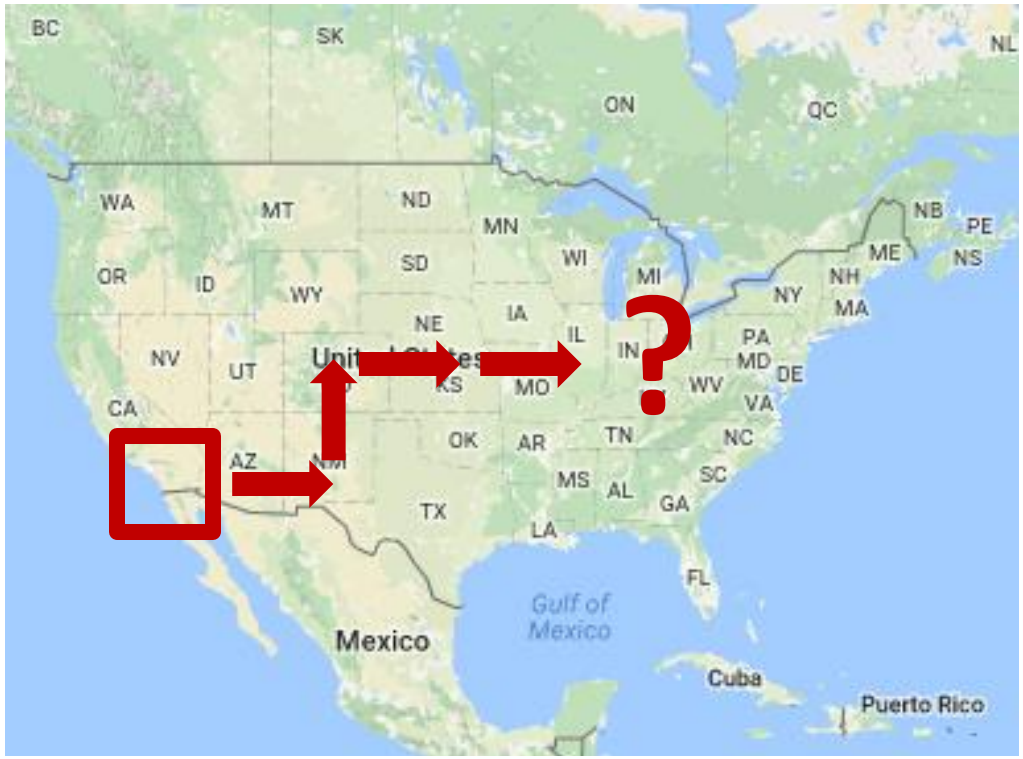
MIT CSAIL

[v]alt

R. Chang et al., Dynamic Prefetching of Data Tiles for Interactive Visualization. SIGMOD 2016

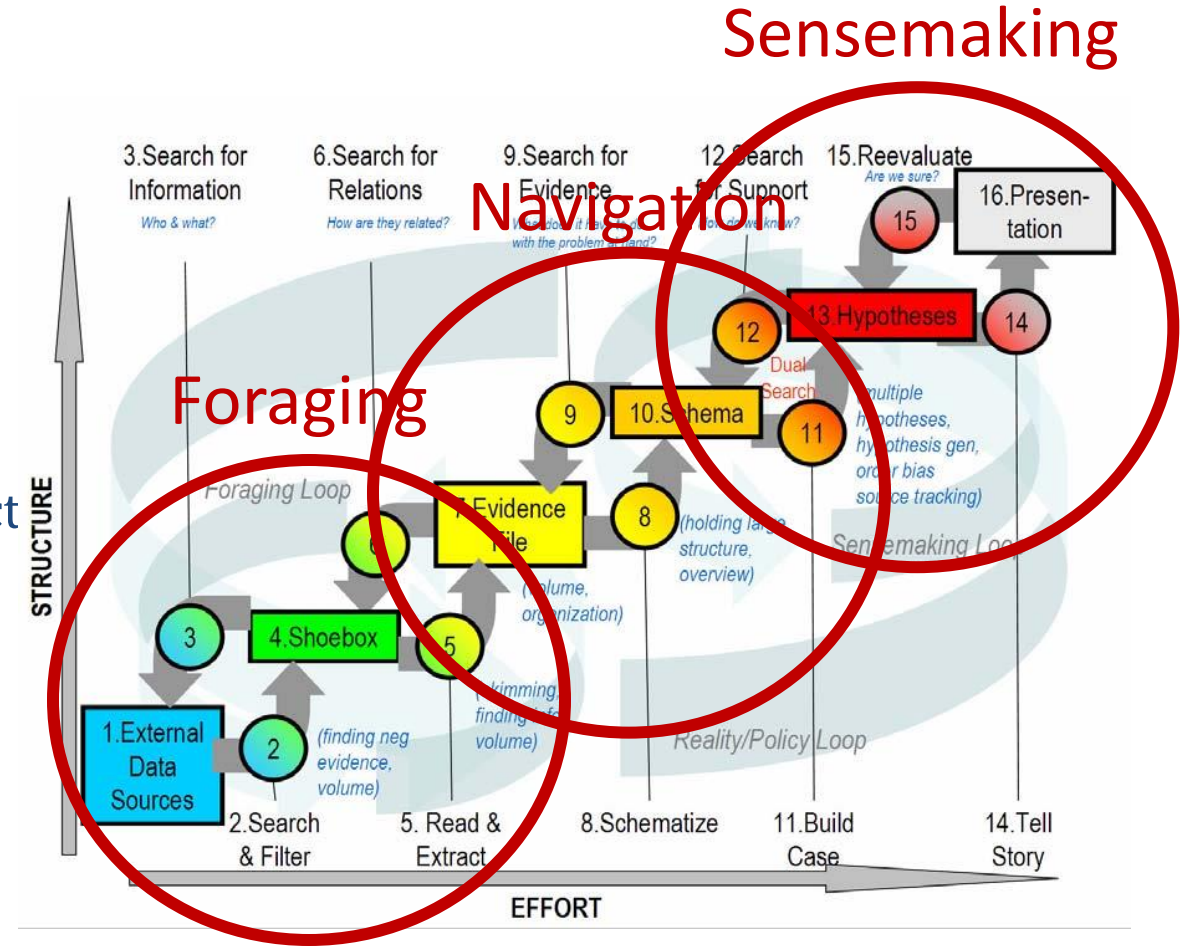
Predictions, an Example

- What is the user going to do next?
  - Likely "right" or "up"
- Prediction of behavior is pretty easy IF we know that the user is "navigating"
- But what if the user is doing something else?



Predicting User Actions

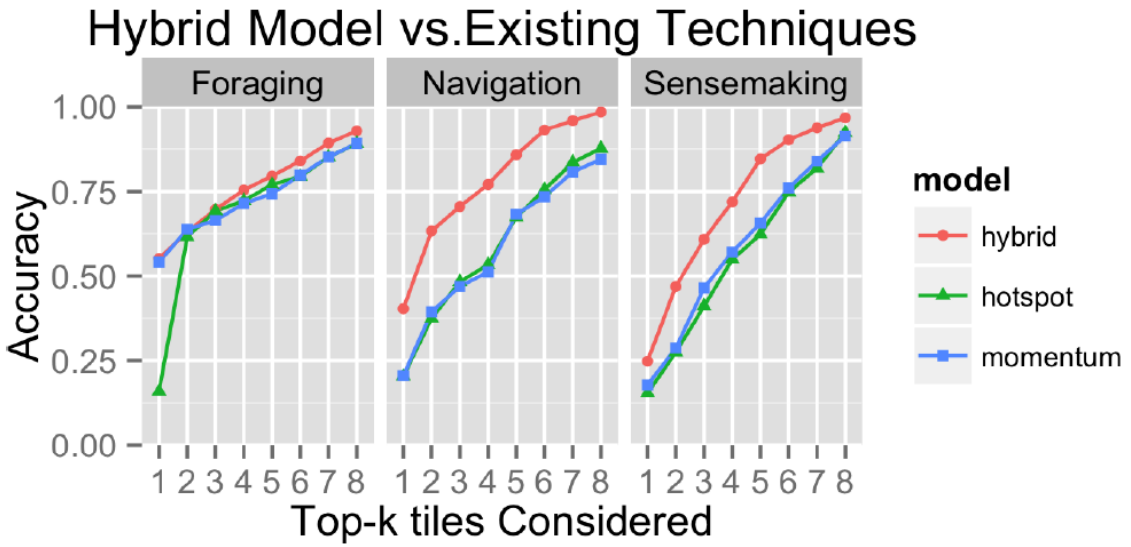
- Two-tiered approach using Markov
- First tier: predict what "phase" of analysis the user is in.
- Second tier: given a "phase", use phase-specific models to predict user's next actions.
  - For example, for "navigation", use momentum-based prediction
  - For others, use a combination of access-frequency, statistical distribution, SIFT (image-based), etc.



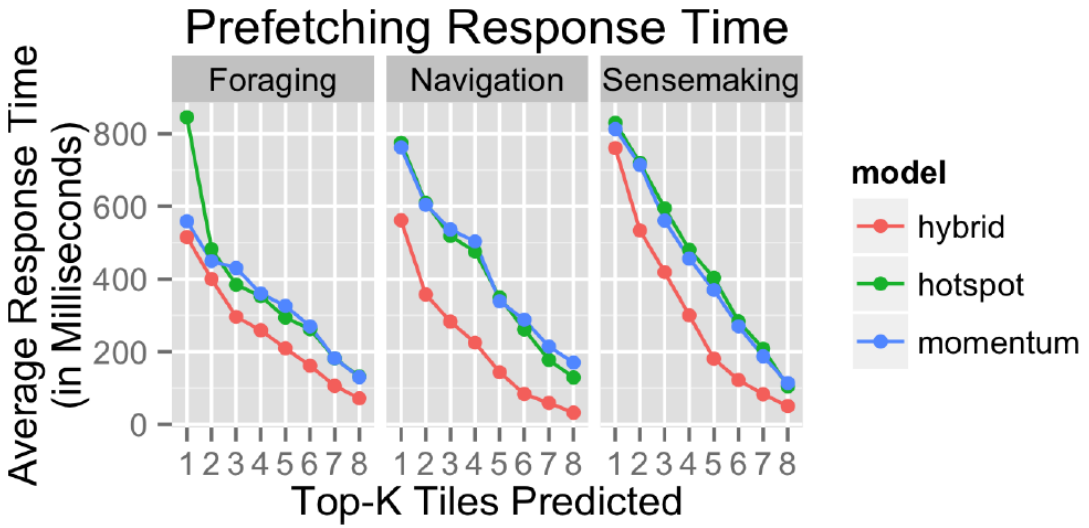
Card-Pirolli Sensemaking Loop

Evaluation: Prediction Accuracy

- Comparison against existing techniques



- "Random guessing" accuracy is: k/n
  - n: number of possible user actions
  - k: number of allowed "guesses"

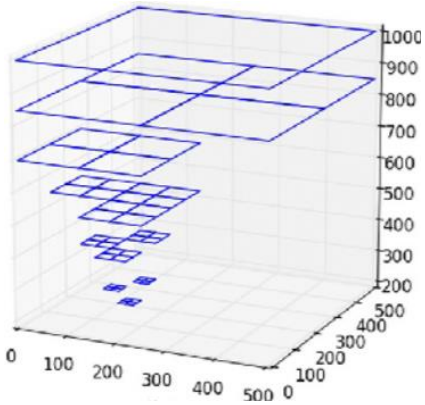


KYRIX: GENERALIZED PAN-ZOOM

- Kyrix is an extension to ForeCache with two improvements:

1. It does not rely on pre-computed tiles
   - Data is stored in a modified R-tree
   - Update to an R-tree is much cheaper than re-computation of tiles

2. Kyrix includes a specification language
   - Authoring of a Kyrix "app" is fast and easy
   - The language, based on D3, allows for a wide range of visualization designs and layouts



	⟨SSV⟩	::= ⟨Marks⟩⟨Layout⟩⟨Data⟩[Config]	(1)
; marks			
	⟨Marks⟩	::= ⟨Cluster⟩[Hover]	(2)
	⟨Cluster⟩	::= ⟨Mode⟩⟨Aggregate⟩[Config]	(3)
	⟨Hover⟩	::= ⟨Ranklist⟩⟨Boundary⟩[Config]	(4)
	⟨Mode⟩	::= *Circle* | *Contour* | *heatmap* | *Radar* | *Pie* | ⟨Custom⟩	(5)
	⟨Aggregate⟩	::= ⟨Dimension⟩ ∗ ⟨Measure⟩ +	(6)
	⟨Ranklist⟩	::= ⟨Topk⟩(*Tabular* | ⟨Custom⟩)	(7)
	⟨Boundary⟩	::= *Convex Hull* | *BBox*	(8)
	⟨Custom⟩	::= Custom JS mark renderer	(9)
	⟨Dimension⟩	::= ⟨Field⟩[Domain]	(10)
	⟨Measure⟩	::= ⟨Field⟩⟨Function⟩[Extent]	(11)
	⟨Topk⟩	::= A positive integer	(12)
	⟨Domain⟩	::= A list of string values	(13)
	⟨Function⟩	::= *Count* | *Sum* | *Avg* | *Min* | *Max* | *Sqrsum*	(14)
; layout			
	⟨Layout⟩	::= ⟨X⟩⟨Y⟩⟨Z⟩[Theta]	(15)
	⟨X⟩	::= ⟨Field⟩[Extent]	(16)
	⟨Y⟩	::= ⟨Field⟩[Extent]	(17)
	⟨Z⟩	::= ⟨Field⟩⟨Order⟩	(18)
	⟨Theta⟩	::= A number between 0 and 1	(19)
	⟨Field⟩	::= A database column name	(20)
	⟨Extent⟩	::= A pair of float numbers	(21)
	⟨Order⟩	::= *Ascending* | *Descending*	(22)
; data			
	⟨Data⟩	::= a database query	(23)
; config			
	⟨Config⟩	::= Key value pairs	(24)

[v]alt

KYRIX: EXAMPLE APPS (NBA)

Data: All plays in the 2017-2018 NBA season with 560k records

[v]alt

KYRIX: SPECIFICATION LANGUAGE

```
var logoTransform = new Transform("logoTransform",
    "select * from teams;",
    "nba",
    function (row){
        var id = parseInt(row[0]);
        var y = Math.floor(id / 6);
        var x = id - y * 6;
        var ret = [];
        ret.push((x * 2 + 1) * 80);          // x coordinate of logos
        ret.push((y * 2 + 1) * 80 + 100);    // y coordinate of logos
        for (var i = 1; i <= 4; i ++)
            ret.push(row[i]);                // raw data attributes
        return ret;
    },
    ["x", "y", "team_id", "city", "name", "abbr"]);
```

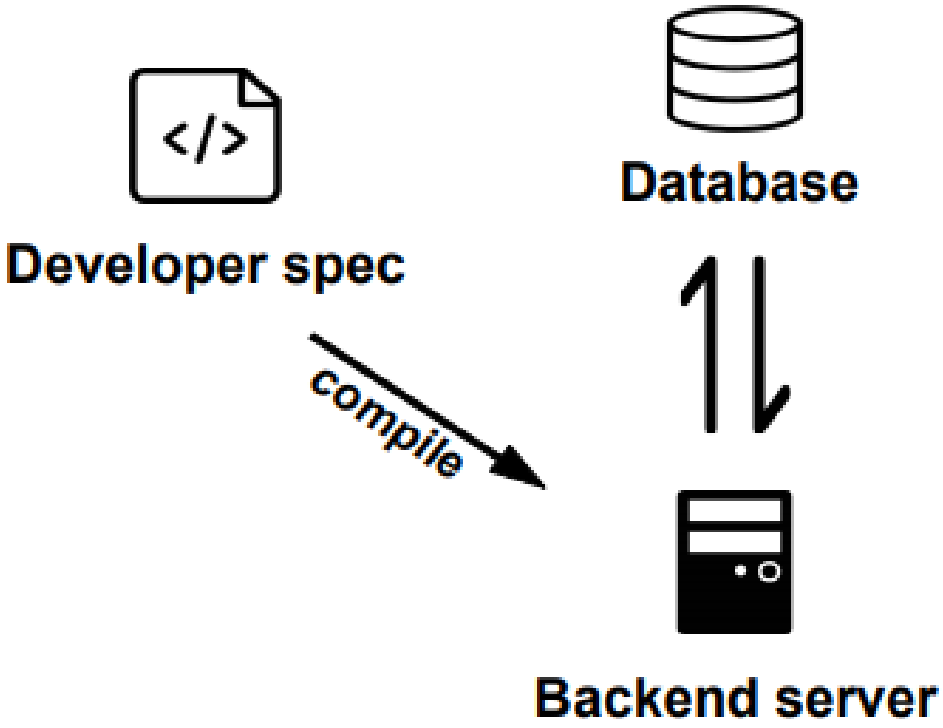
Data Transform

```
// ======== teamlogo -> teamtimeline ========
var selector = function (row) {
    return true;
};

var viewport = function (row) {
    return [0, 0];
};

var predicate = function (row) {
    return {
        "layer 0" : "home_team=" + row.team_id + " and "
            + "away_team=" + row.team_id,
        "layer 1" : "id=" + row.team_id
    };
};

p.addZoom(new Zoom(logoCanvas,
    timelineCanvas,
    selector,
    viewport,
    predicate);
```

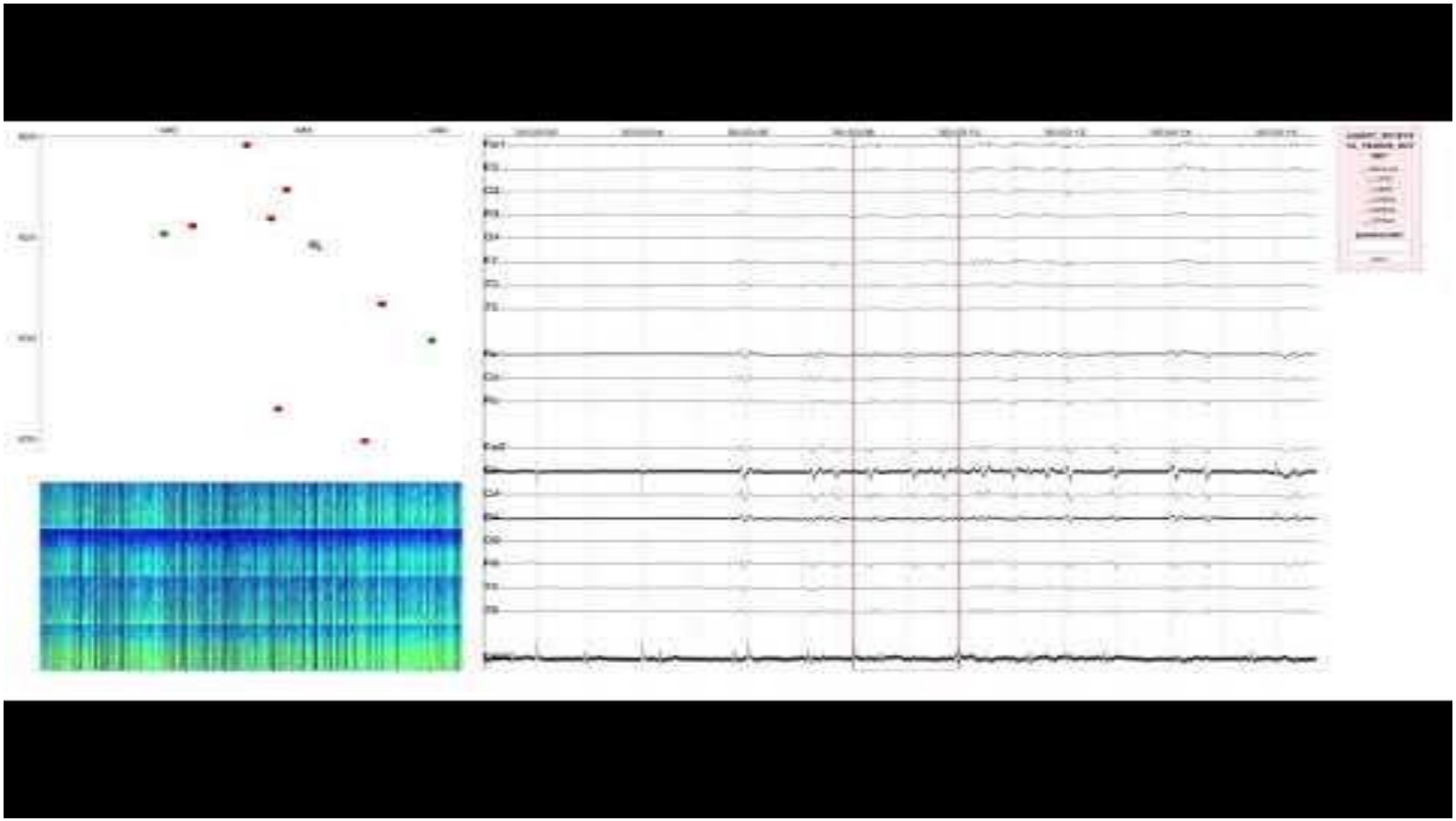
Transition Functions

[v]alt

KYRIX: SYSTEM ARCHITECTURE (REALTIME)



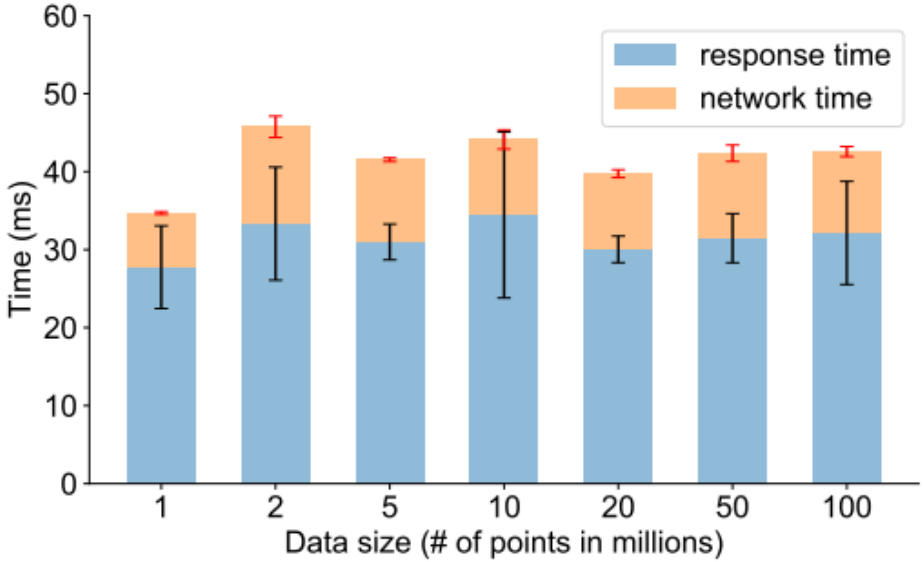
KYRIX: EXAMPLE APP (MGH EEG)



Dataset: 17 million patients (shown as dots in the scatterplot) along with the 100+ million data points for all of their EEG data (right).

EVALUATION

- Benchmark Performance
  - Beats target of 500ms by 10x



- User study
  - 8 participants (4 with visualization programming experience, 4 without)
  - Participants asked to recreate (parts of) the NBA example*
  - Completion time of avg: 26.25 minutes, stdev: 7.07 minutes

*Participants wrote the placement functions and the transition functions of the top two layers, but did not have to write the transform or the rendering functions.

[v]alt


Zhe Wang (Arizona)


Dylan Cashman (Tufts)


Mingwei Li (Arizona)


Jixian Li (Arizona)


Matt Berger (Vanderbilt)


Josh Levine (Arizona)


Carlos Scheidegger (Arizona)

DATA SYSTEM FOR SUPPORTING INTERACTIVE DATA EXPLORATION: **NEURALCUBES**

R. Chang et al., Neuralcubes: Deep representations for visual data exploration. ArXiv, 2018

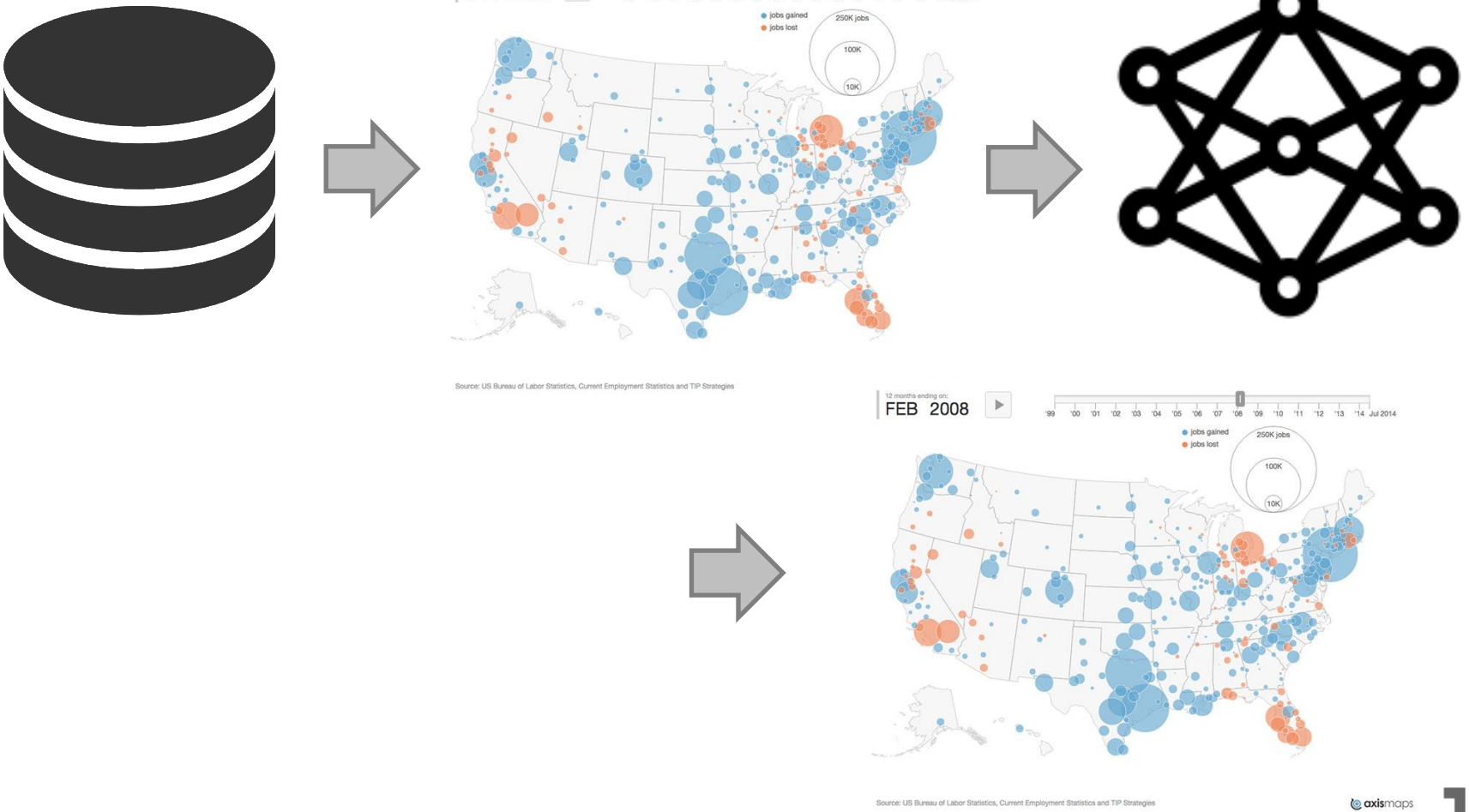
[v]alt

Interactive Exploration of Big Data

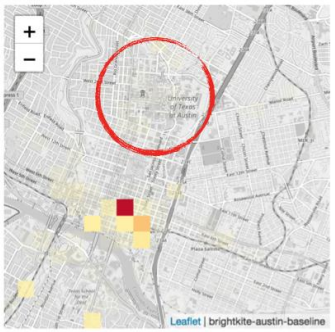
Visualization on a Commodity Hardware

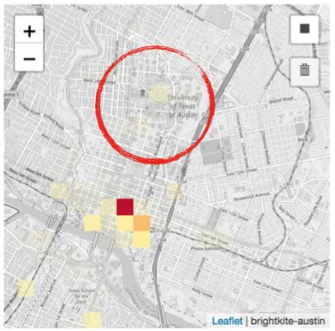
Large Data in a Data Warehouse

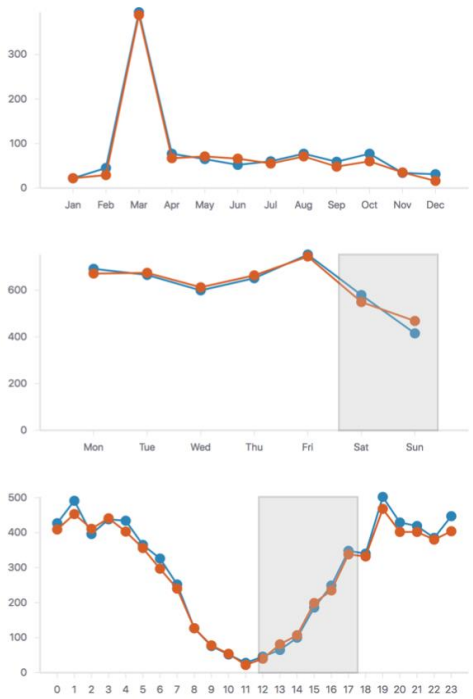
NeuralCubes -- Big Picture Idea

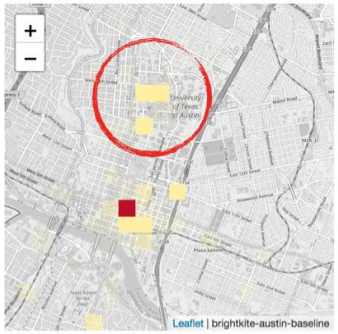


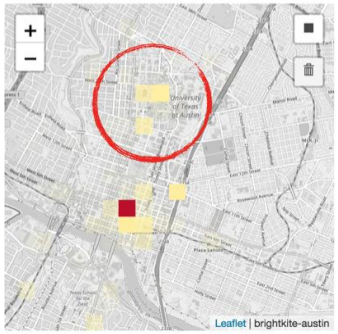
SAMPLE RESULTS

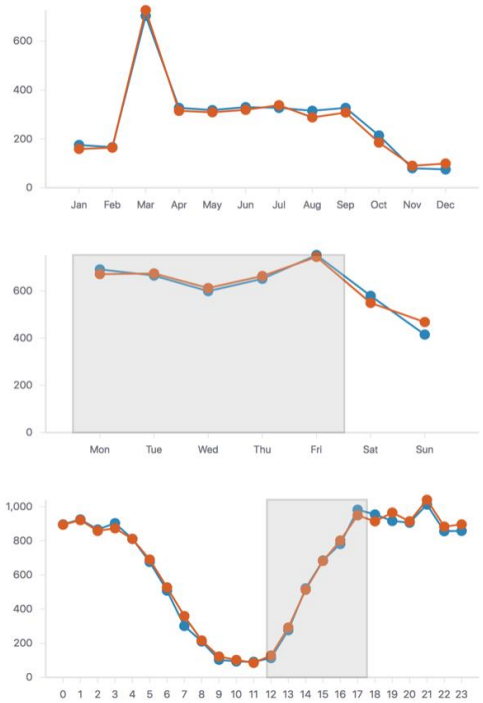












brightkite-austin-baseline brightkite-austin-nn

Storage Saving

DataSet	Raw Data Size	# queries	# net params	network size	# iteration	training time	testing error
SPLOM	100,000 (3.9MB)	1000k	50,653	232KB	1000	3 h 5 mins	0.022
	100,000 (3.9MB)	2000k	53,903	245KB	1000	3 h 8 mins	0.028
	100,000 (3.9MB)	3000k	57,153	258KB	1000	8 h 44 mins	0.038
	100,000 (3.9MB)	4000k	60,403	271KB	1000	10 h 32 mins	0.048
	100,000 (3.9MB)	5000k	63,653	284KB	1000	10 h 17 mins	0.057
BrightKite NYC	79,357 (3.1MB)	8,860k	169,032	688KB	500	4 h 29 mins	0.124
BrightKite Austin	22,171 (868KB)	8,860k	169,032	688KB	300	5 h 55 mins	0.094
Flights (count)	5,092,321 (204MB)	856,689	24,267	152KB	250	1 h 16 mins	0.042
Flights (delay)	5,013,088 (175MB)	670,000	123,330	830KB	500	2 h 14 mins	0.025

- When original data is small, not much saving
- When original data is larger, the saving can start to become more significant

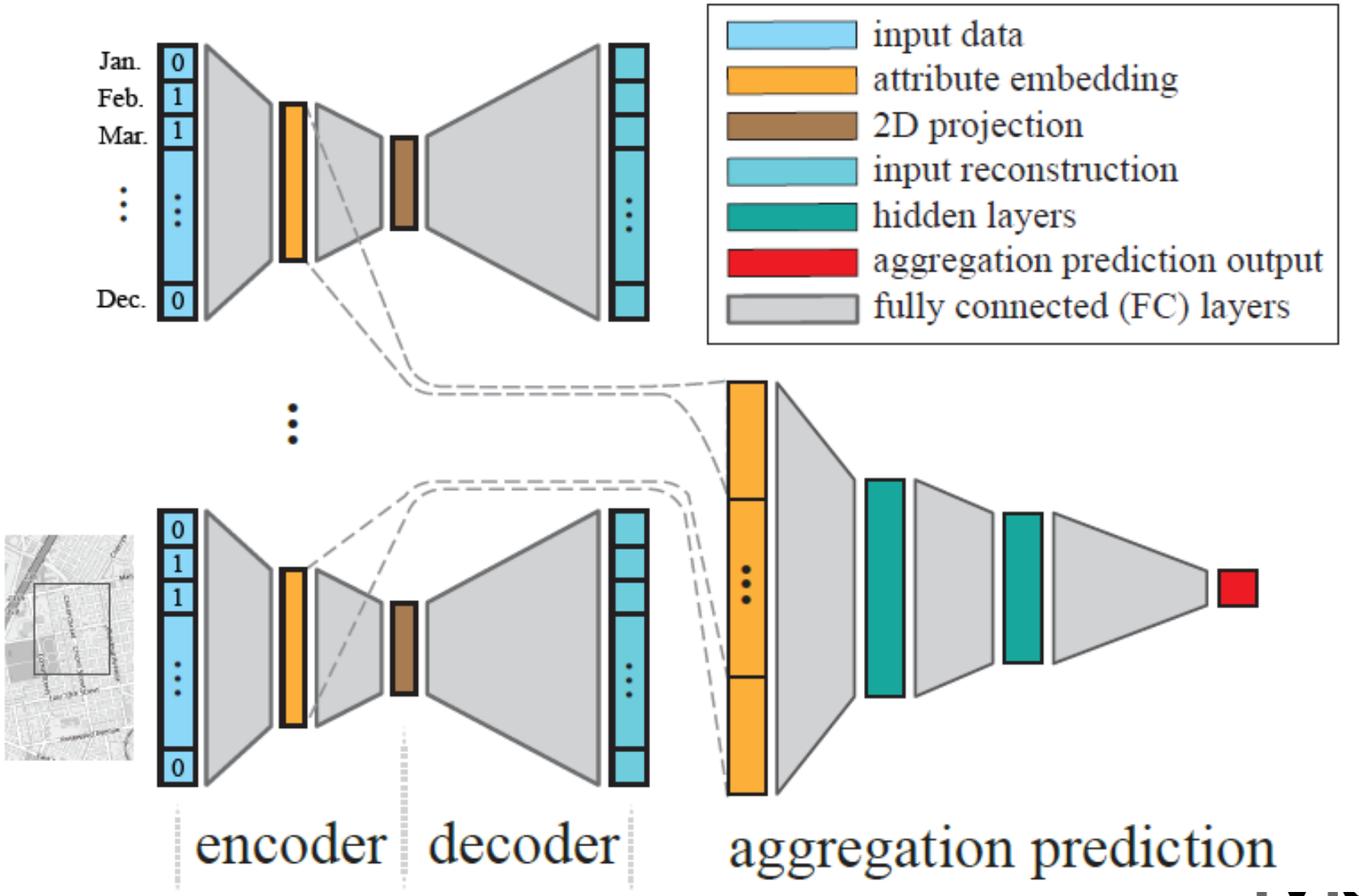
[v]alt

But Wait, There's More!!

Raw Data Size	# States	Model Size	Testing RAE
12k (1.8MB)	30k	798KB	3.70%
120k (18MB)	30k	798KB	2.04%
1.2m (180MB)	30k	798KB	1.35%
12m (1.8GB)	30k	798KB	0.97%

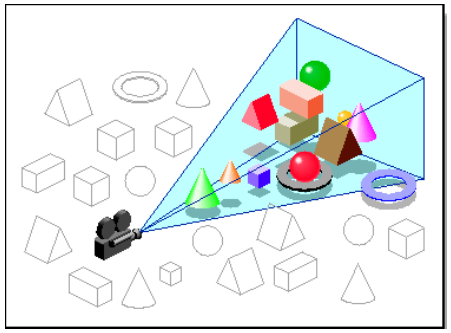
- As the raw data size goes up, so does the amount of training data
  - As a result, the amount of error actually decreases (while the model size stays constant)!!
  - In this case, the space saving is 2000X
    - (2 GB -> 1MB)

2D AUTOENCODER

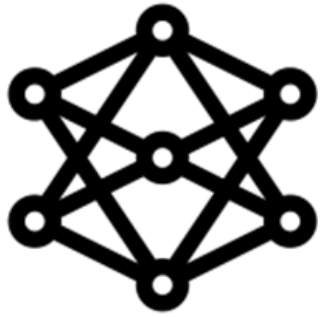


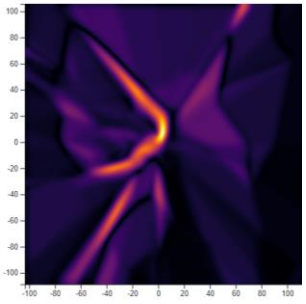
Recap: Data Systems

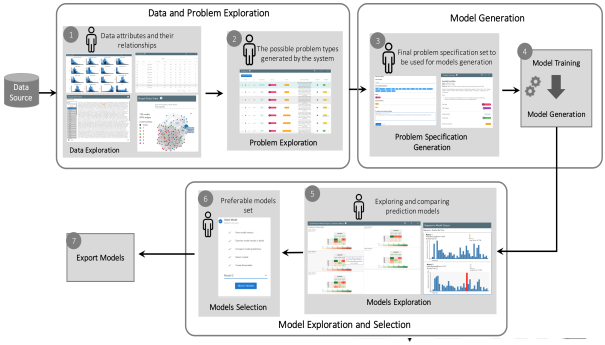
- **Human + ML:**
  - SnowCat
    - AutoML + VIS system that is better than AutoML and human expert
  - "Unproject"
    - Learns the inverse of a projection function



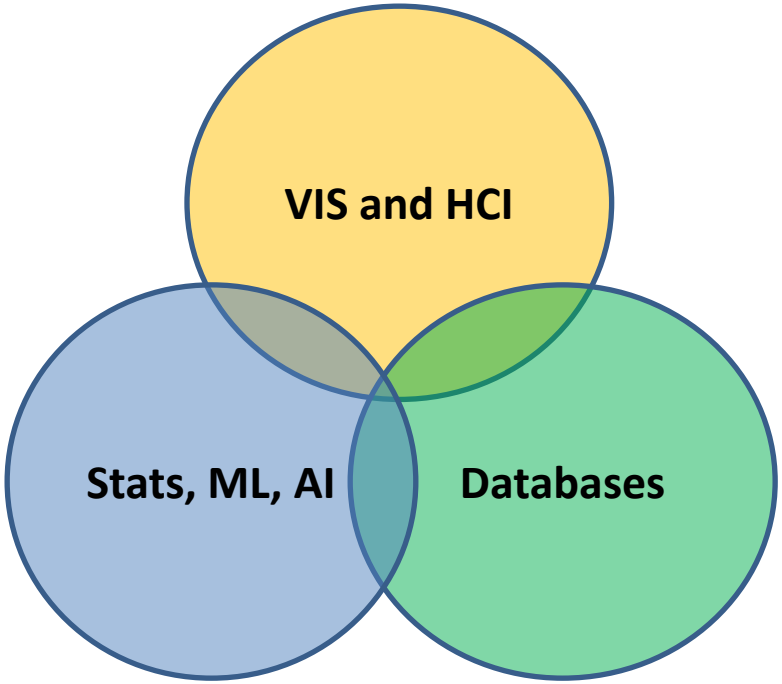
- **Human + DB:**
  - ForeCach & Kyrix
    - Optimized database systems for pan-zoom visualization systems
  - NeuralCubes
    - Use of deep learning to compress the data







The Future of Data Systems for VIS



- Future of data analysis (and data science) will be at the intersection of **VIS/HCI**, **Stats/ML/AI**, and **DB/Cloud**
- Each community is becoming more aware of this need
  - KDD: IDEA workshop
  - SIGMOD: HILDA workshop
  - VIS: DSIA workshop
- Work I presented today is just scratching the surface.
  - How to integrate all three is hard
  - Space is wide open. Join us!

[v]alt

[v]alt
Visual Analytics Laboratory @ Tufts

QUESTIONS?

REMCO
@CS.TUFTS.EDU