Data Mining
in
Naturalistic Driving

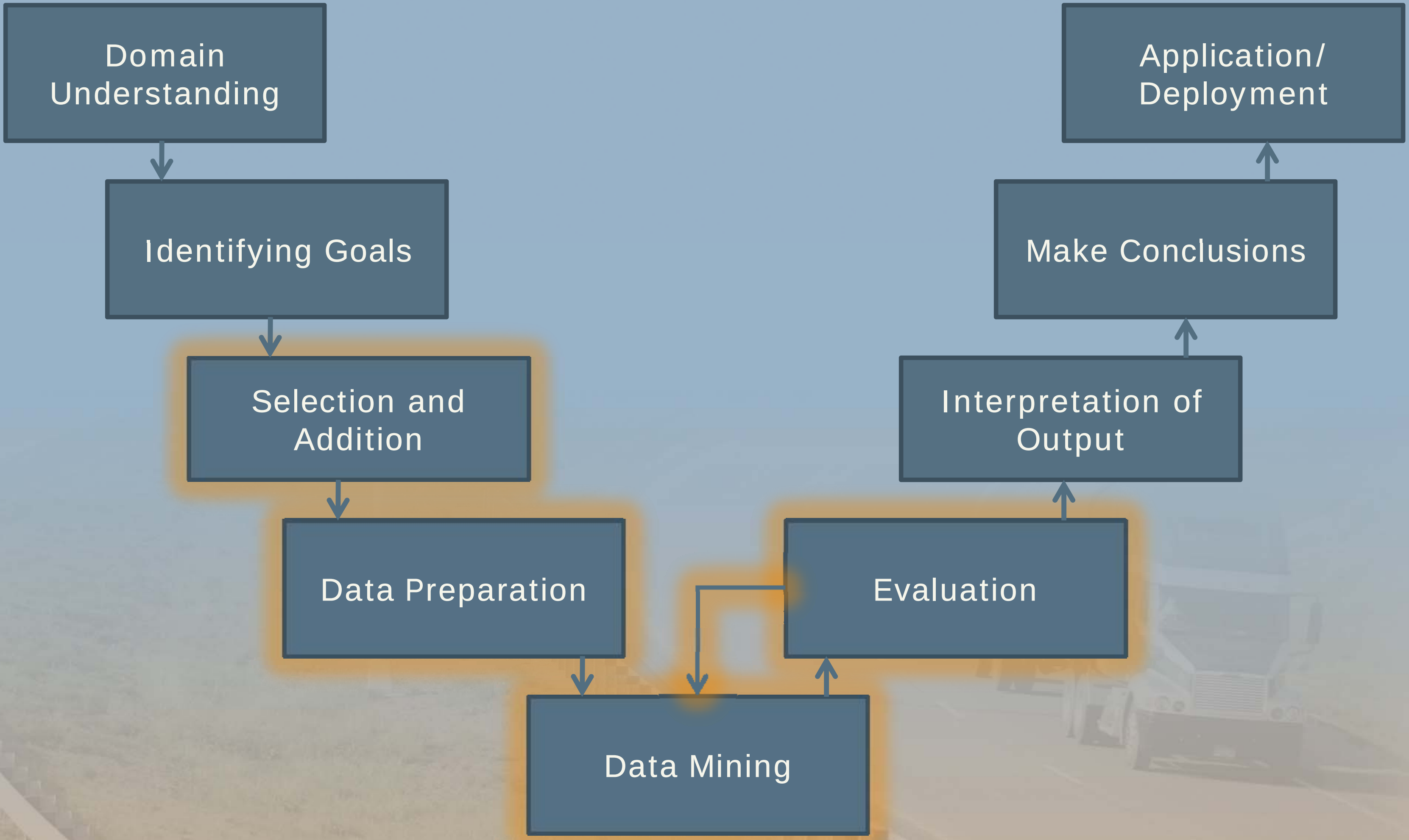
Shane McLaughlin, PhD

Center for Automotive Safety Research

Driving Transportation with Technology

VirginiaTech
TRANSPORTATION
INSTITUTE

Knowledge Discovery in Data (KDD)



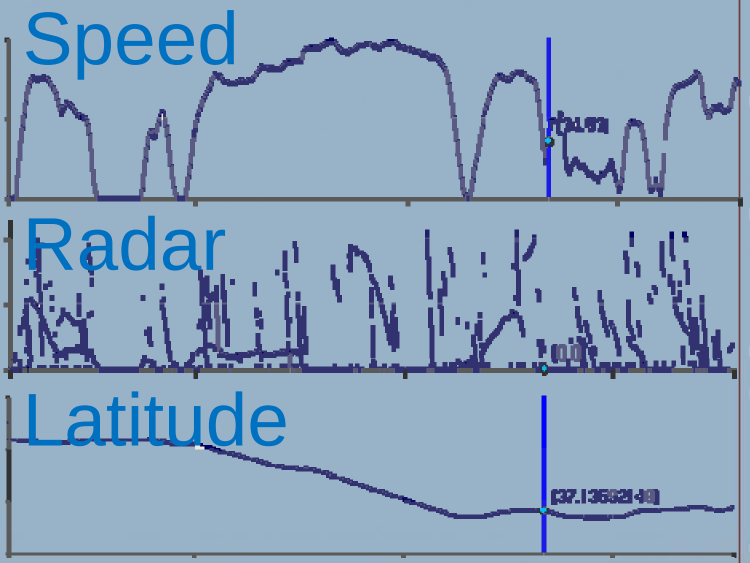
Driving Transportation with Technology

Example Research Goal

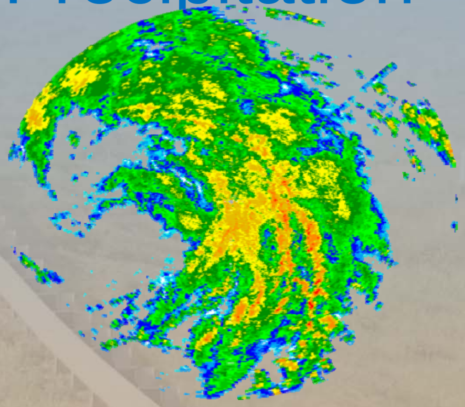
Are there differences in driver following behavior in urban areas during clear weather versus severe rain?

Selection and Addition

- Acquiring Samples
- Understanding the data
- Explore
- Evaluate quality
- Select interesting subsets
- Plan integration of datasets
- Selecting fields/attributes
- Sampling design



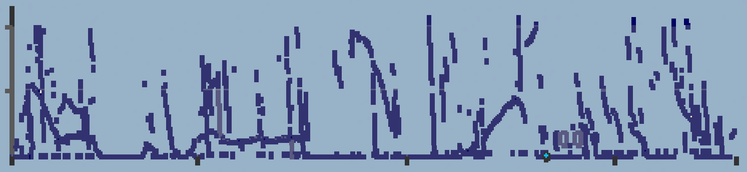
Precipitation



Urban Areas

Time
Date
Demographics
Vehicle Type

Data Preparation



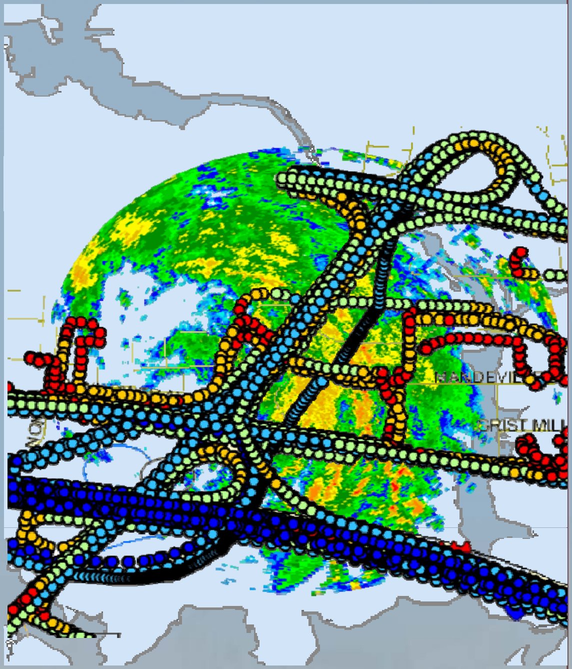
- Organizing
  - Accumulating files
  - Domain specific applications
  - Connections to large datasets
  - Definitions, units, sign, coding
- Storage/processing strategy
  - RAM vs reduced for later
  - Flat table, mixed format, relational
  - Read/write speeds, subsequent analysis
- Transforming
  - Format, creating composite variables, separating
- Cleaning
  - Missing values, noise, outliers, incorrect values
- Prepare data set from raw for use in all subsequent stages

Naturalistic Data Mining (DM)

- Three DM Algorithm Components
- Event Parsing Component
- Crunching

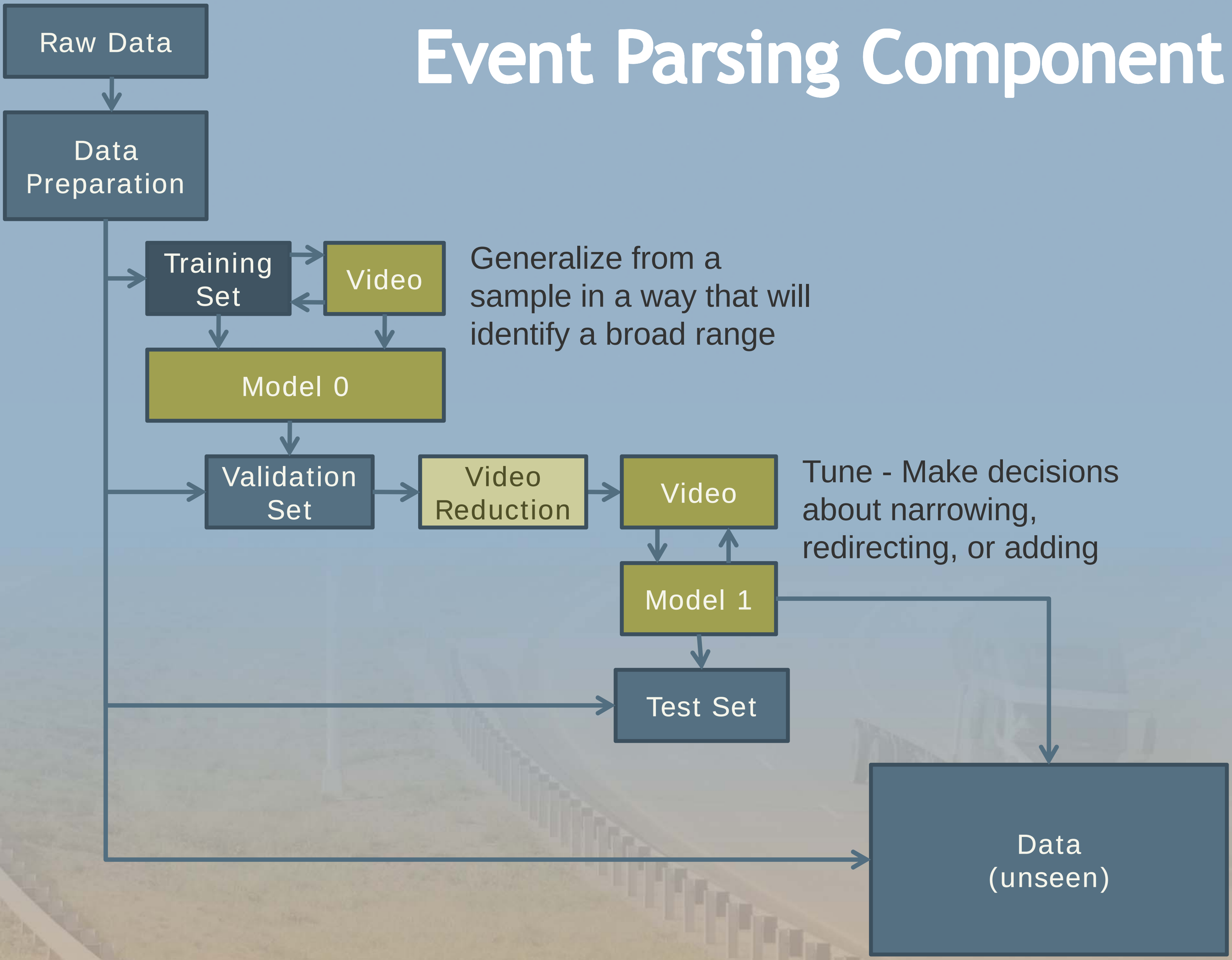
DM Algorithm Components

1. Stream processing
   - Numerical methods
   - Filters
   - Splines
   - FFTs
2. Event parsing
   - Triggers – boolean logic, thresholds and combinations
   - Algorithms
     - Custom scenario recognition code
     - Kinematic models
     - Neural Nets
     - Machine vision



3. Descriptive Data Capture - IVs and DVs
   - Within event counts, summaries etc (steering reversals)
   - Aggregation, trends descriptive statistics (max, mean, dominant frequencies)
   - Classification (lead vehicle braking, intersection turn)
   - References used for subsequent stages (Target ID, road segment)
   - Temporal landmarks within data (sync of max brake, sync of glance up)

Event Parsing Component



Raw Data

Data Preparation

Training Set

Video

Model 0

Validation Set

Video Reduction

Video

Model 1

Test Set

Data (unseen)

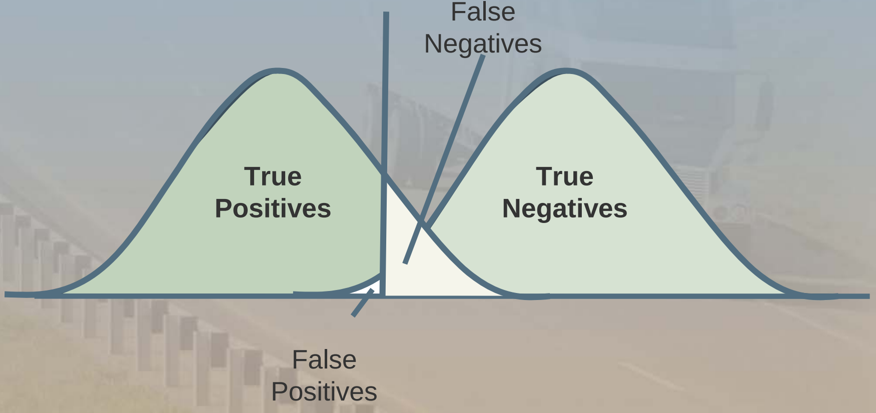
Generalize from a sample in a way that will identify a broad range

Tune - Make decisions about narrowing, redirecting, or adding

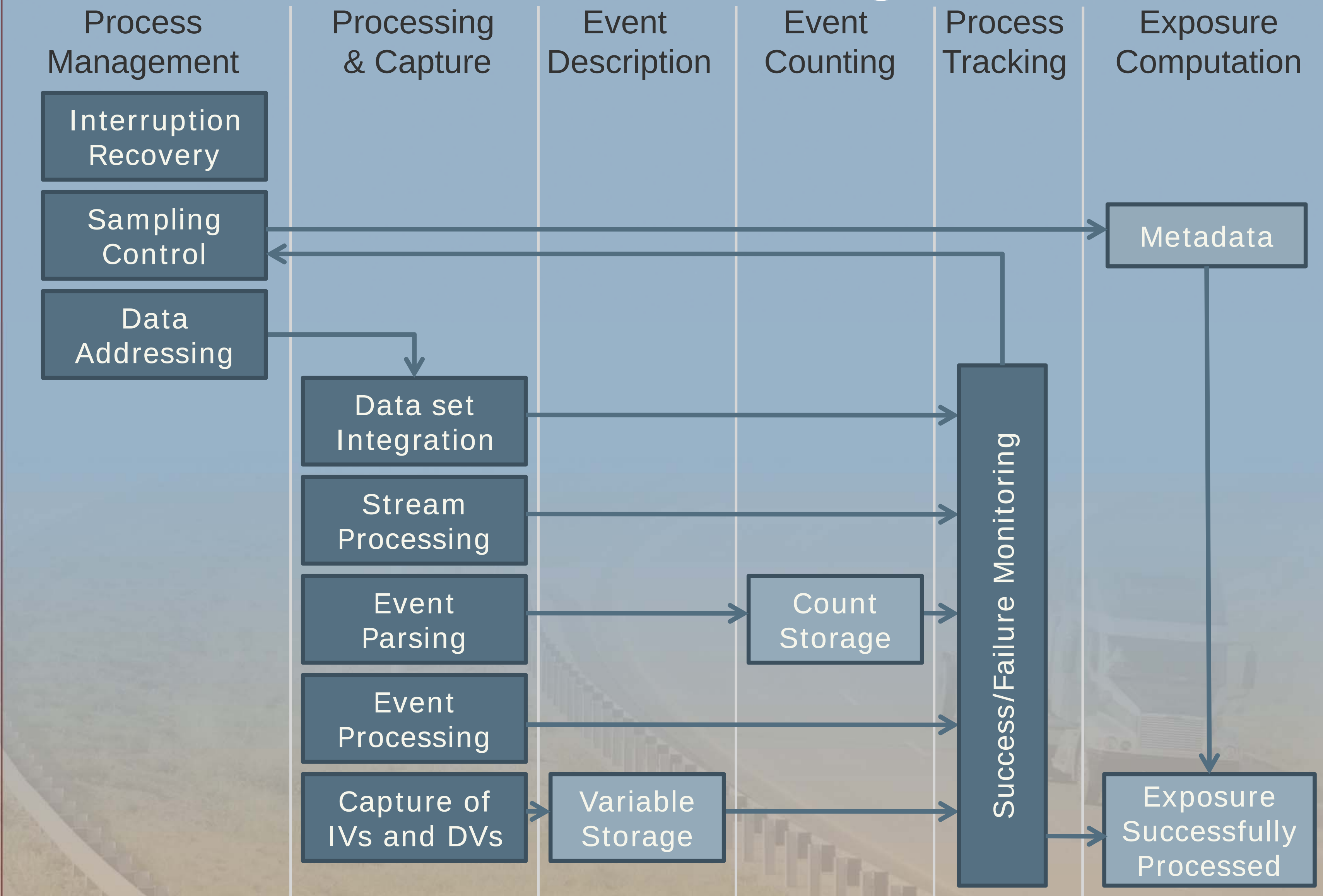
Evaluation

		Predicted				
		Urban Following	Something else			
Actual	Urban Following	**True Positive**	False Negative Type II	Sensitivity	TP/(TP+FN)	Method finds x% of true events
	Something else	False Positive Type I	**True Negative**	Specificity	TN/(TN+FP)	x% correct saying something is not of interest
		Positive Predictive Value	Negative Predictive Value			
		TP/(TP+FP)	TN/(TN+FN)			
		Strength of confirming a true indication	Strength of confirming a false indication			

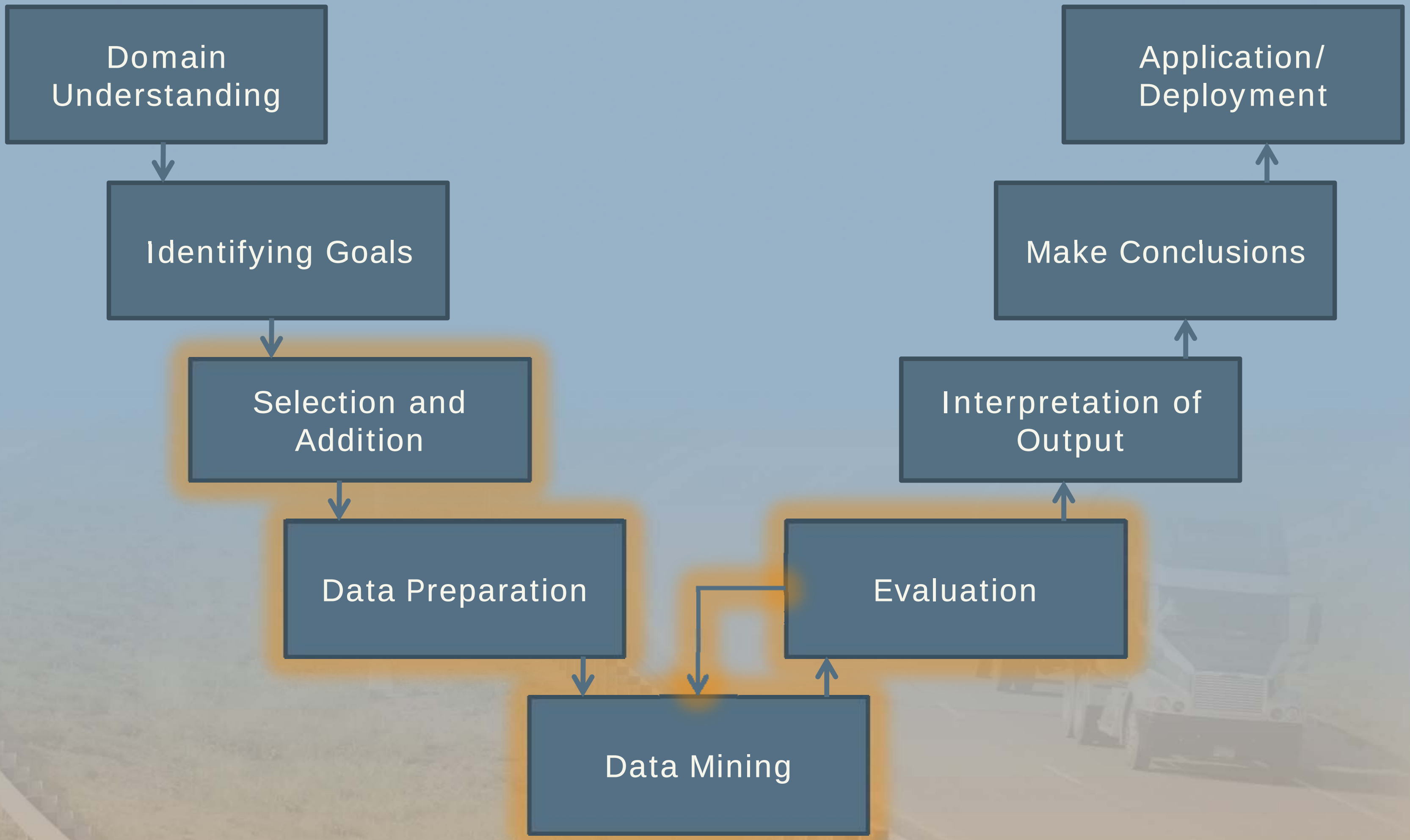
		Predicted	
		Urban Following	Something else
Actual	Urban Following	**Hits**	Misses
	Something else	False Alarms	**Correct Rejections**



DM Crunching



Knowledge Discovery in Data (KDD)



Domain Understanding

Identifying Goals

Selection and Addition

Data Preparation

Data Mining

Evaluation

Interpretation of Output

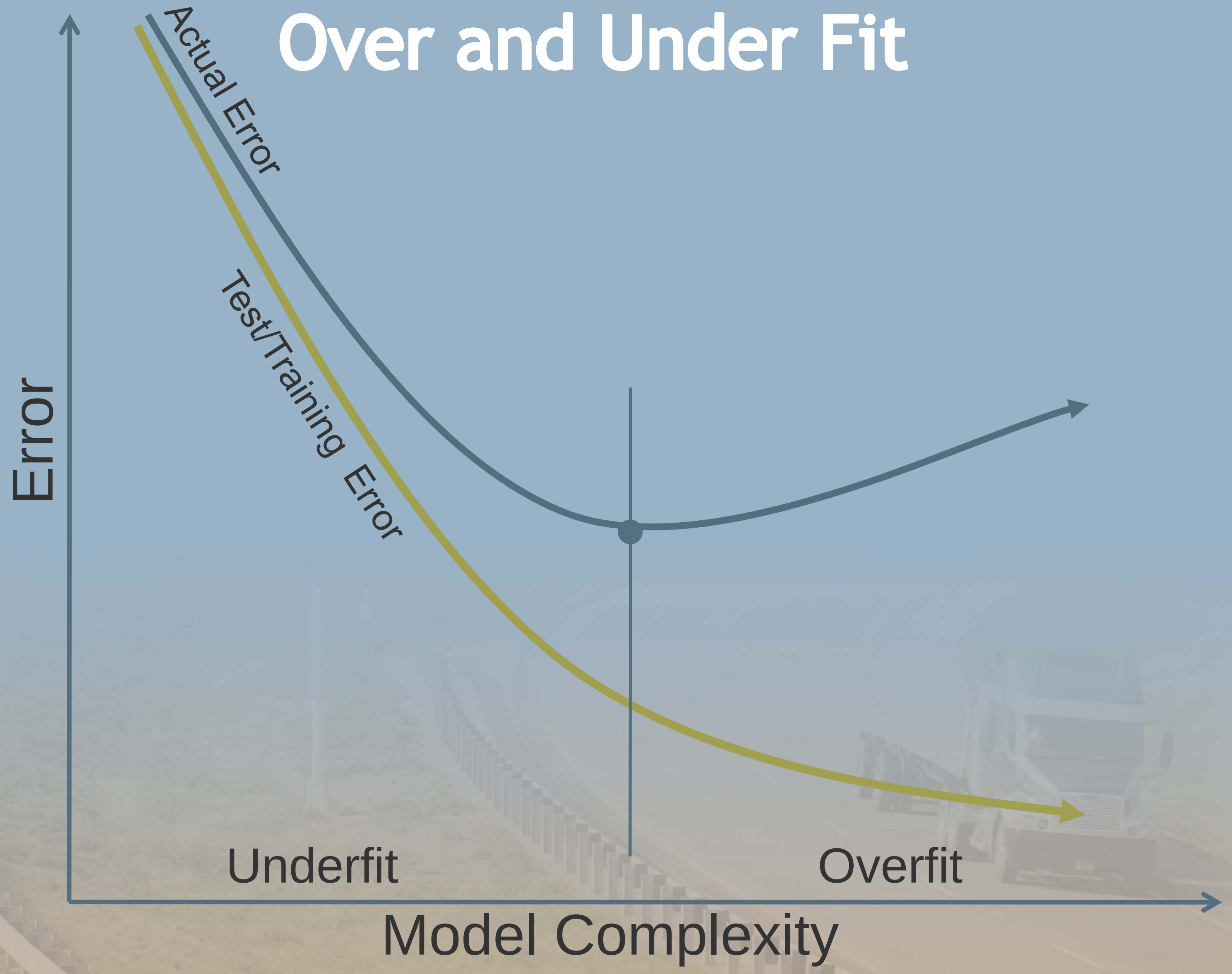
Make Conclusions

Application/ Deployment

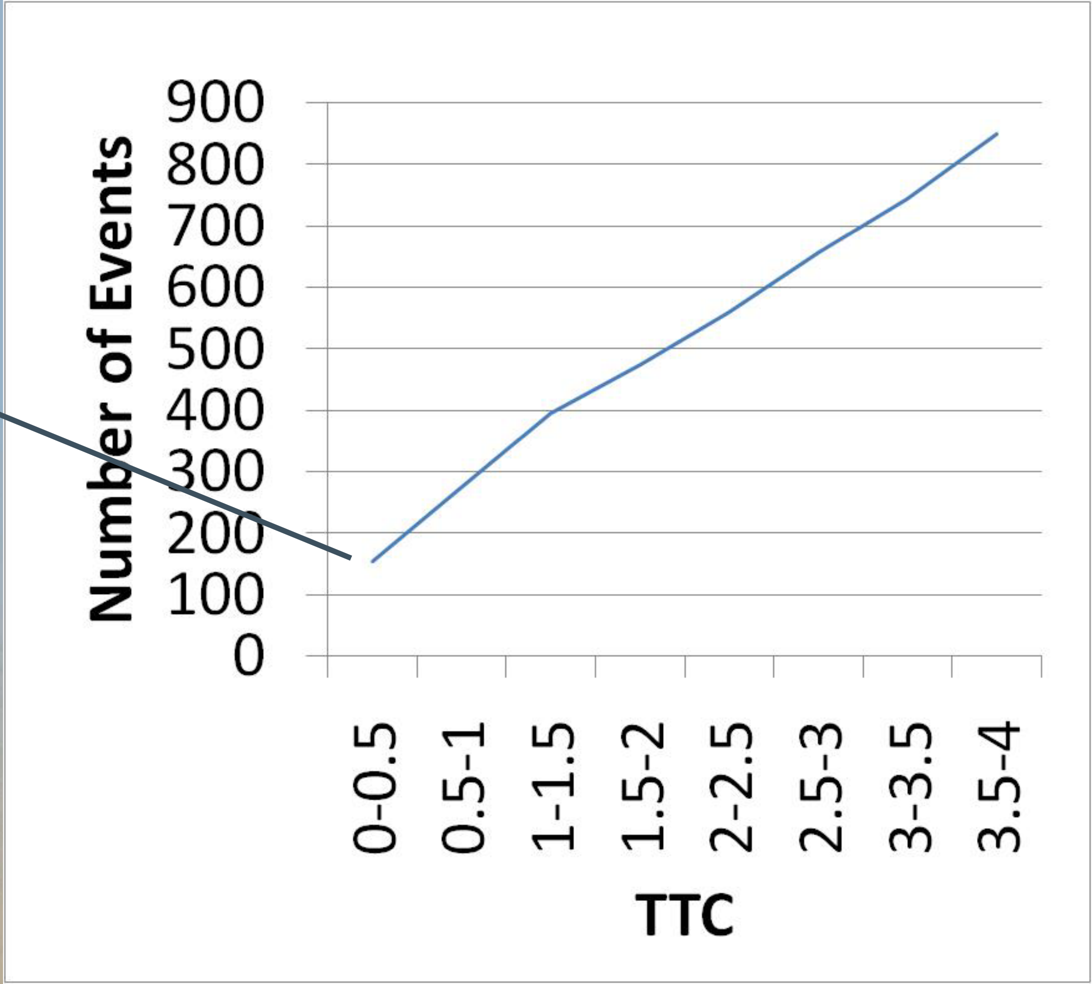
Pitfalls

- Not familiarizing with domain and details of data
  - Faulty from start
  - Imbedding assumptions early - too narrow
- Starting analysis before the data is clean
  - If detected, rework
  - If not detected, faulty conclusions
  - Data versioning difficulty
- Not designing a DM sampling strategy and monitoring successes.
  - Sampling bias
  - Incorrect exposure estimates
  - Insufficient data
- Evaluating on the same data used for developing a model
  - Optimistic estimates of performance

Over and Under Fit



Hidden Bias



Mined events

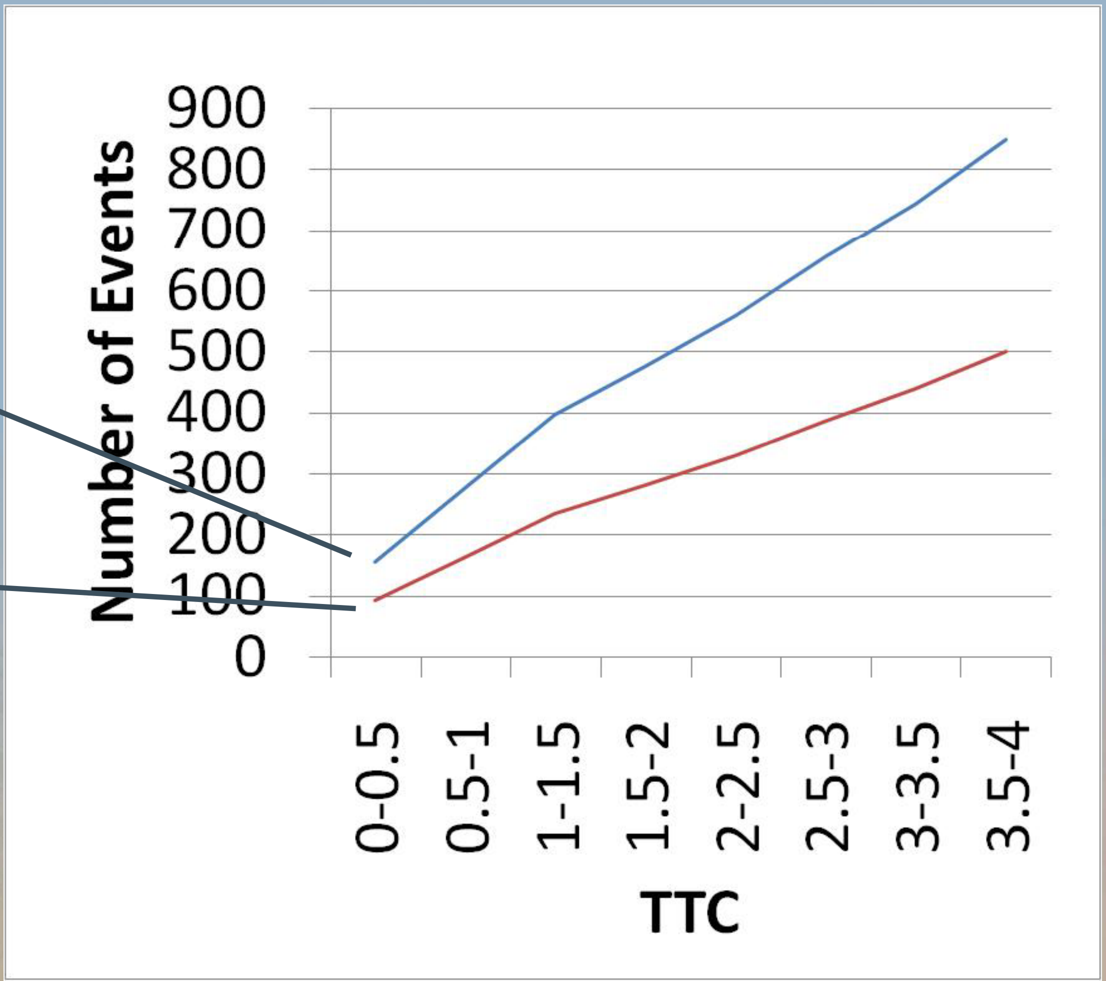
Driving Transportation with Technology

Virginia Tech Transportation Institute

Hidden Bias

Mined events

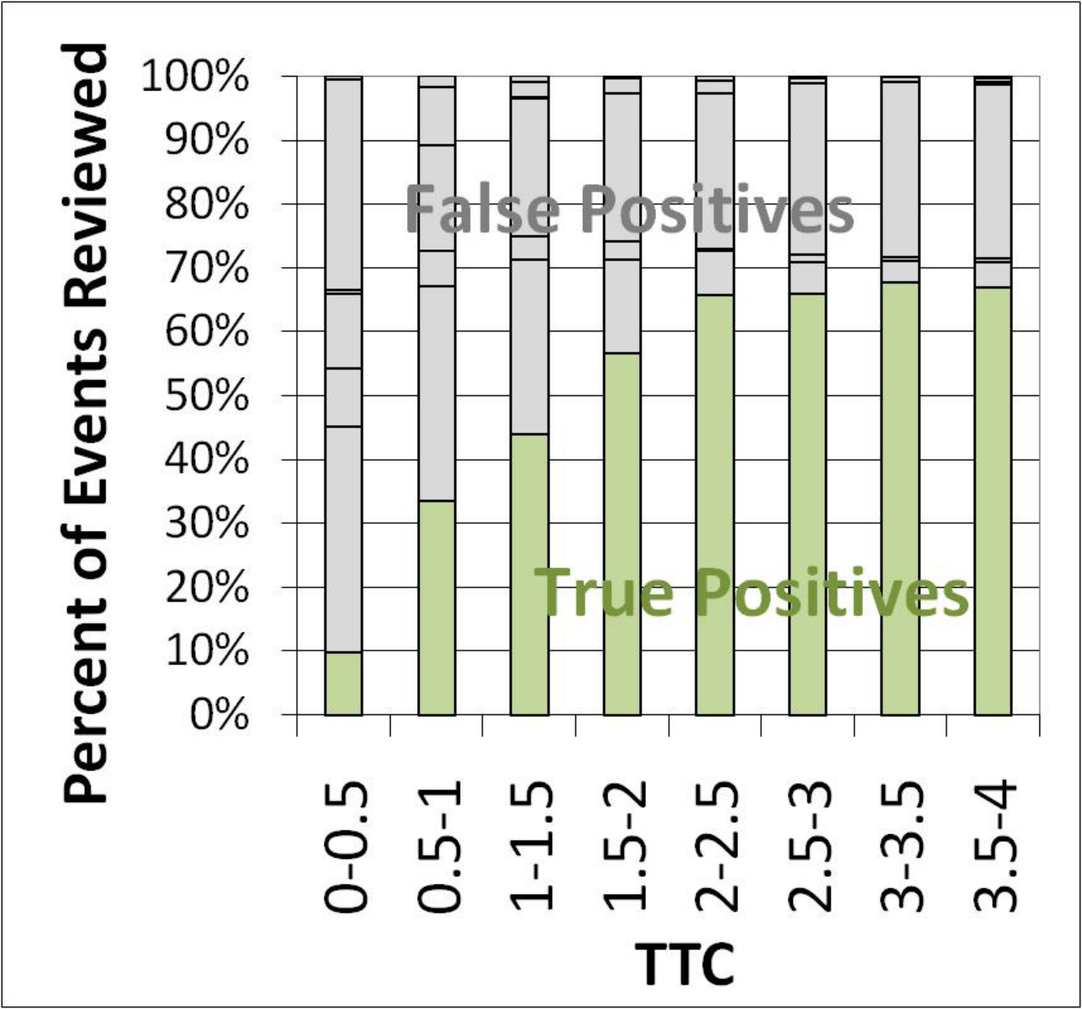
Adjustment - random sample. 31% found to be false positives.



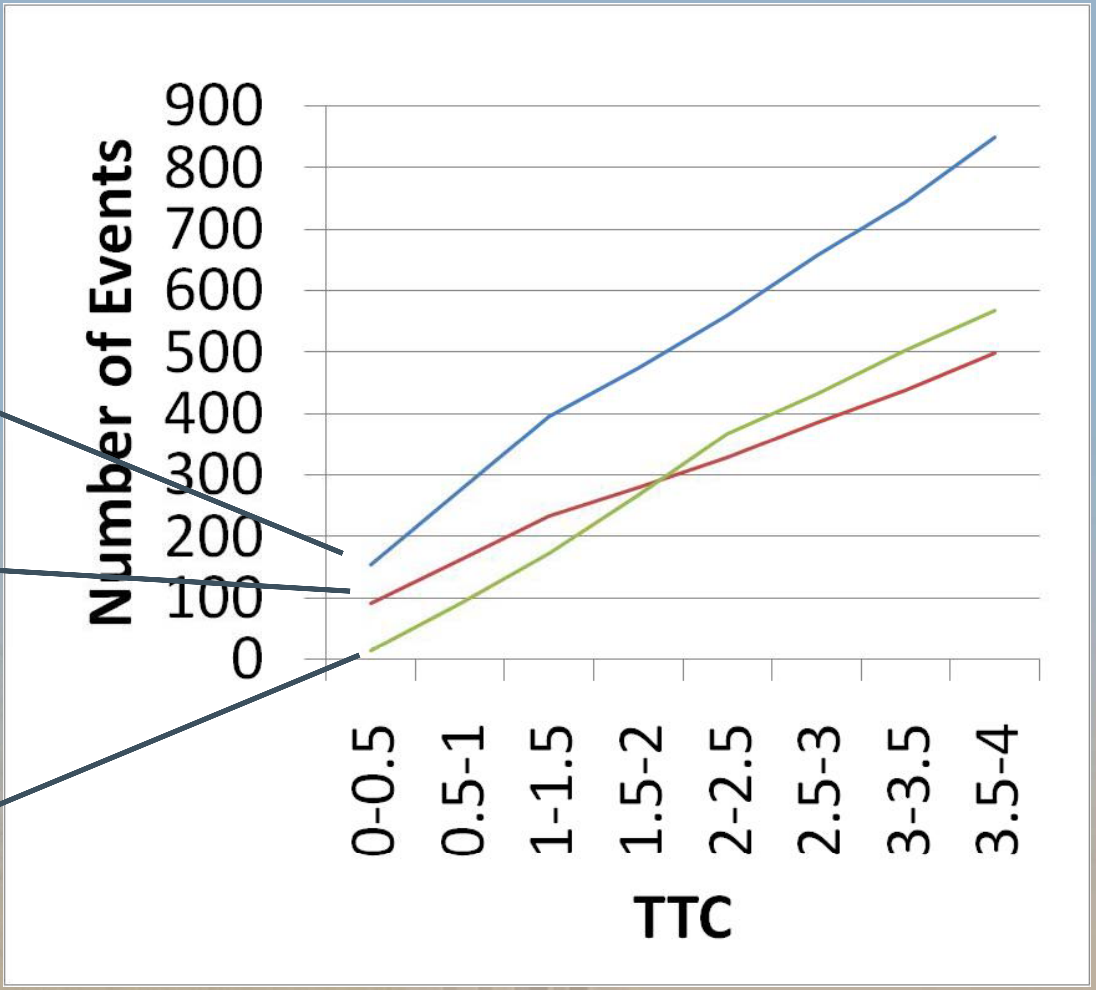
Hidden Bias

Stratified Evaluation Approach

Bias present in proportion of valid events across variable of interest



Hidden Bias

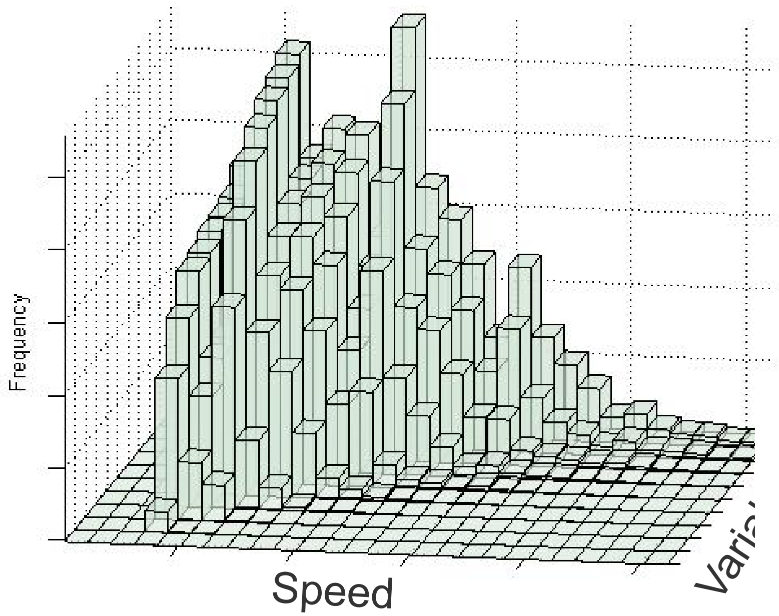


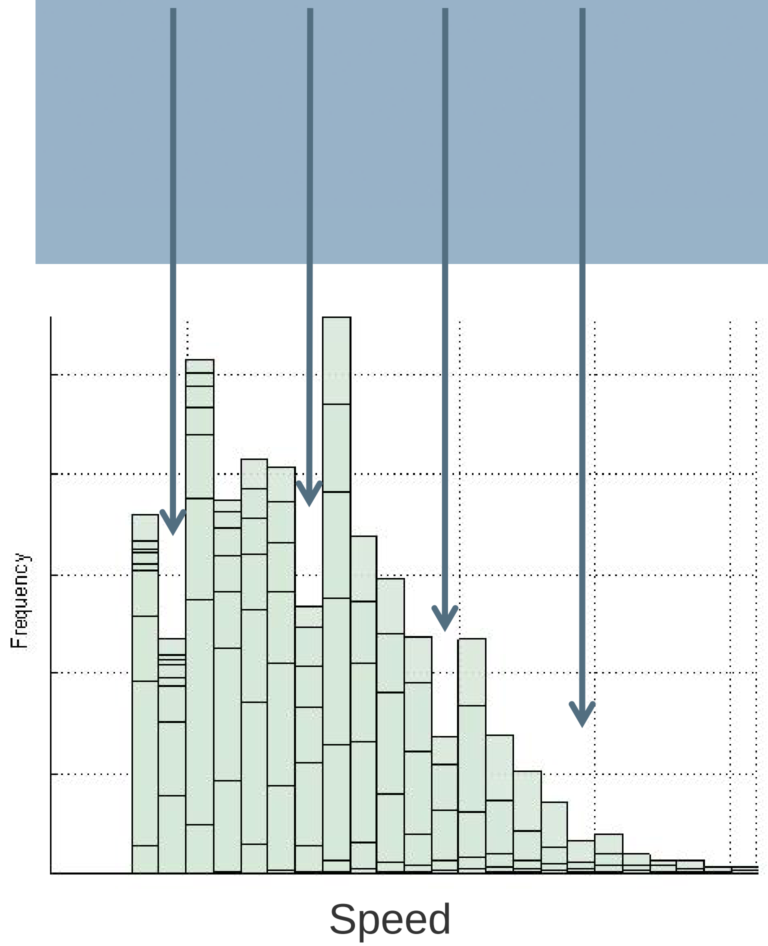
Mined events

Adjustment - random sample. 31% found to be false positives.

Adjustment correcting for bias in data mining code.

Aliasing or Real Effect?

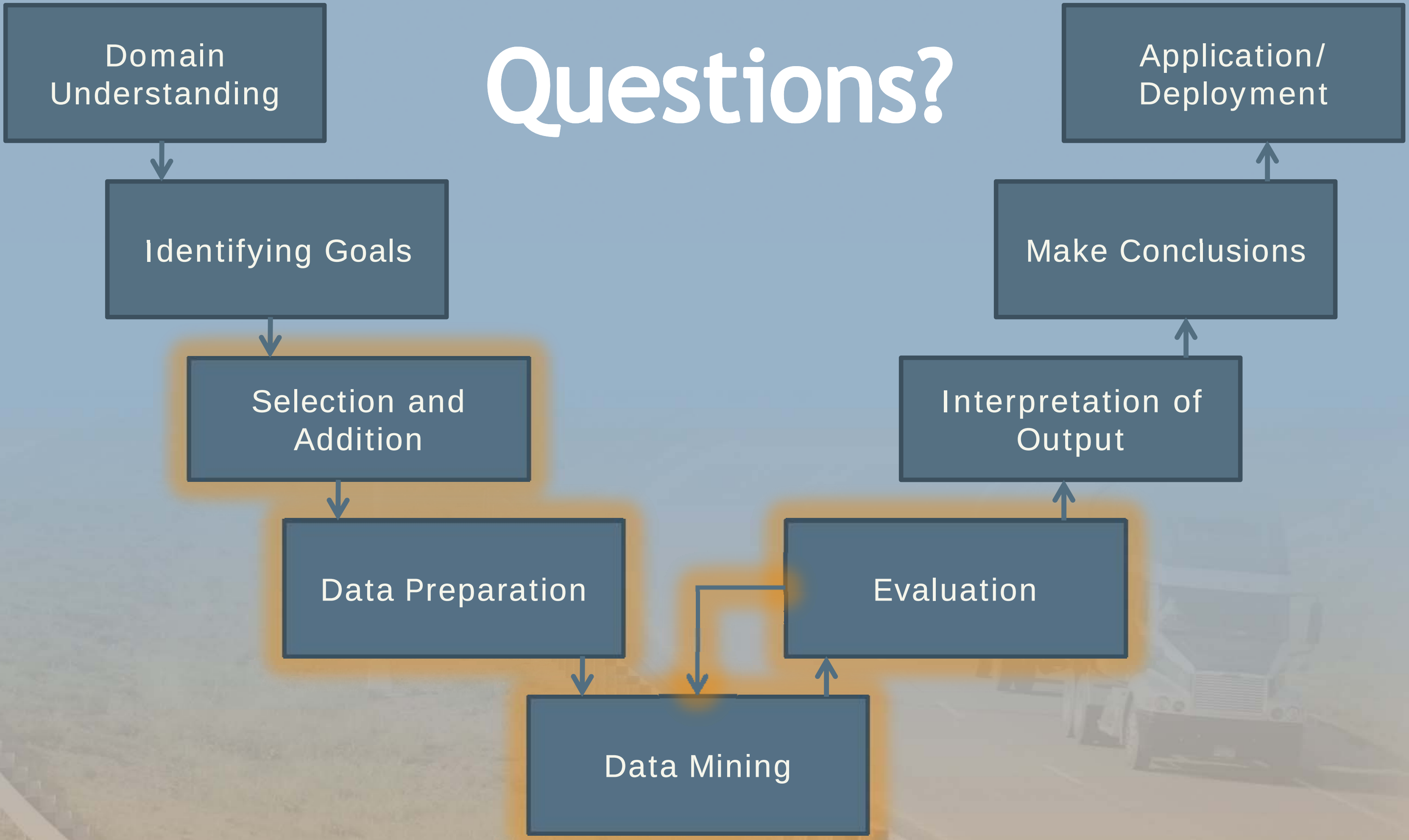




Driving Transportation with Technology

Knowledge Discovery in Data (KDD)

Questions?



Driving Transportation with Technology

References and Links

- Larose, D. T. (2005). Discovering knowledge in data: an introduction to data mining. John Wiley & Sons. Hoboken, NJ.
- Maimon, O., Rokach, L. Eds. (2005). Data mining and knowledge discovery handbook. Springer. New York, NY.
- Witten, I., Frank, E. (2005). Data mining: practical machine learning tools and techniques 2nd ed. Elsevier. San Fransico, CA.
- http://en.wikipedia.org/wiki/Sensitivity_(tests)
- http://www.sigkdd.org/
- http://www.kdnuggets.com