Liberty Mutual Research Institute for Safety

Reducing the Resource Burden of Narrative Text Classification for Large Administrative Databases

Wellman HM[1], Corns HL[1], Lehto, MR[2]

[1]Liberty Mutual Research Institute for Safety, 71 Frankland Road, Hopkinton, MA 01748, USA

[2]School of Industrial Engineering, Purdue University, 1287 Grissom Hall, West Lafayette, In 47907, USA

Background

- Narrative analysis is used in many different areas of safety research
  - Prevention information (Sorock et al.,1997, Smith 2001).
  - Case Identification (Lombardi et al., 2005).
  - Identifying pre-event circumstances (Stutts et al., 2001, Sorock et al., 1996).
  - Developing hazard scenarios (Smith et al.,2006).

Computerized classification of NHIS injury narratives. (Wellman et al., 2004)

Original injury narratives and e-codes

↓

Calculations of word/word combination probabilities ← Learn on E-code categories assigned by experts

↓

Fuzzy Bayes predictions/ assignment of codes ← Predict e-code categories for new narratives

$$P(A_i|n)=MAX_j \frac{P(n_j|A_i)P(A_i)}{P(n_j)}$$

Objectives

- Develop and evaluate different filtering techniques (i.e. semi-automated strategy using computer classification scheme & strategically assigned manual coders) to improve accuracy
  - Receiver Operating Characteristic (ROC) curves
  - Linear Predictor method
- Prediction and learning datasets unique – eliminate optimistic bias

Research to Reality™ from

Methods:
Data extraction and gold standard codes assigned

- Randomly extracted injury narratives from a large Workers' Compensation insurance database
- Establish 'Gold Standard' injury cause codes
  - 2 expert coders manually classify each narrative
  - Gold standard cases where the 2 coders agreed. (N=10,389)

1 Digit Classification of Injury Events

- BLS Occupational Injury and Illness Classification System (OIICS)[1].

Categories	Description
0* (01, 02…)	Contact With Objects and Equipment
1*	Falls
2*	Bodily Reaction and Exertion
3*	Exposure to Harmful Substances or Environments
4*	Transportation Accidents
5*	Fires and Explosions
6*	Assaults and Violent Acts
9*	Other Events or Exposures
9999	Non-classifiable

[1](ANSI Z16.2-1995, American National Standard for Information Management for Occupational Safety and Health).

2 Digit Classification of Injury Events

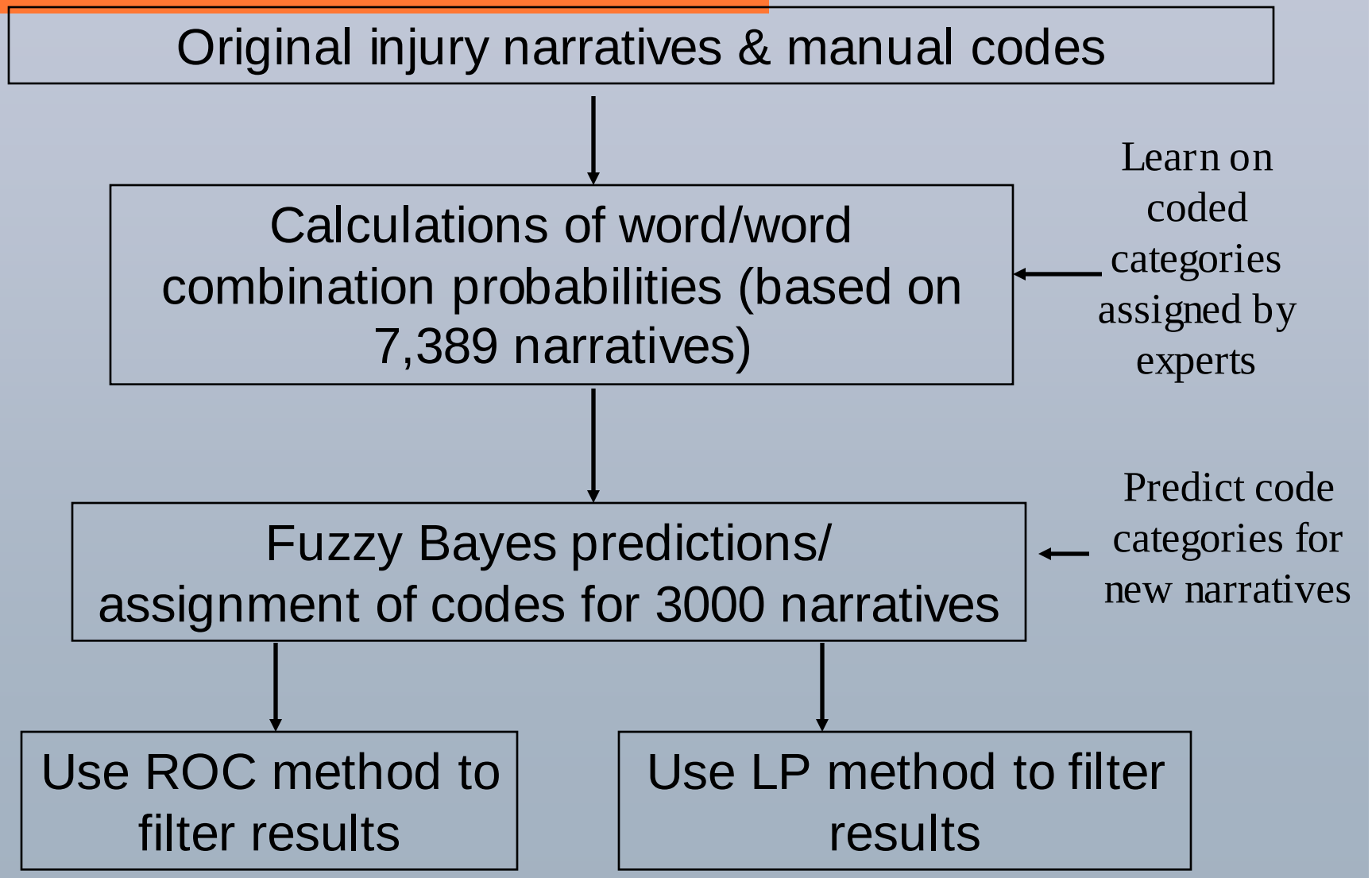
Contact Category ('0*')

2 digit BLS category	Description
00	Contact with Objects
01	Struck against
02	Struck by
03	Caught in
04	Caught in collapsing materials
05	Rubbed/abraded by Friction
06	Rubbed/abraded/ jarred by vibration
09	Contact with objects nec

Fall Category ('1*')

2 digit BLS category	Description
10	Falls, unspecified
11	Fall to lower level
12	Jump to lower level
13	Fall on same level
19	Fall nec

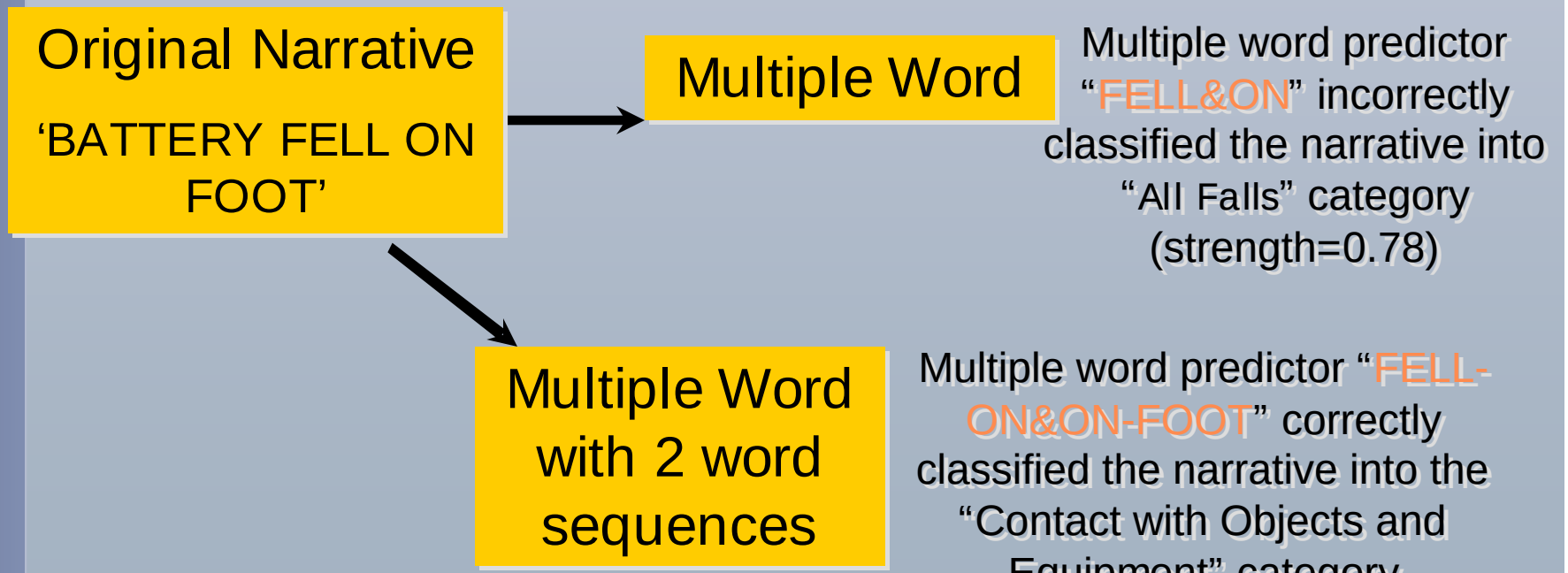
Methods Flowchart



Assigning Category Probabilities – Learning Phase

Word	Classification	Probability
FELL	02:Struck by	0.70
FELL	13:Fall on same level	0.10
FELL-ABOUT	11:Fall to lower level	0.86
FELL-AGAINST	02:Struck by	0.29
FELL-AGAINST	13:Fall on same level	0.57
FELL-BACK	11:Fall to lower level	0.38
FELL-BACK	13:Fall on same level	0.38
FELL-BACKWARD	11:Fall to lower level	0.20
FELL-BACKWARD	13:Fall on same level	0.50
FELL-BACKWARDS	11:Fall to lower level	0.23
FELL-BACKWARDS	13:Fall to same level	0.73

Narrative Coding Prediction Phase Example



Baseline results Category level

BLS Event	n	Baseline 1-digit Category Sensitivity	Baseline 2-digit Category Sensitivity
Contact with objects/equipment	557	84%	64 %
Falls	474	90%	79 %
Bodily Reaction (BR) & Exertion	961	86 %	68 %
Exposure to harmful environment	361	94 %	90 %
Transportation Accident	370	93 %	81 %
Fires & Explosions	21	10 %	10 %
Assaults & Violent Acts	117	74 %	73 %
Non Classifiable	139	25 %	25 %
Total	3000	84 %	71 %

Research to Reality™

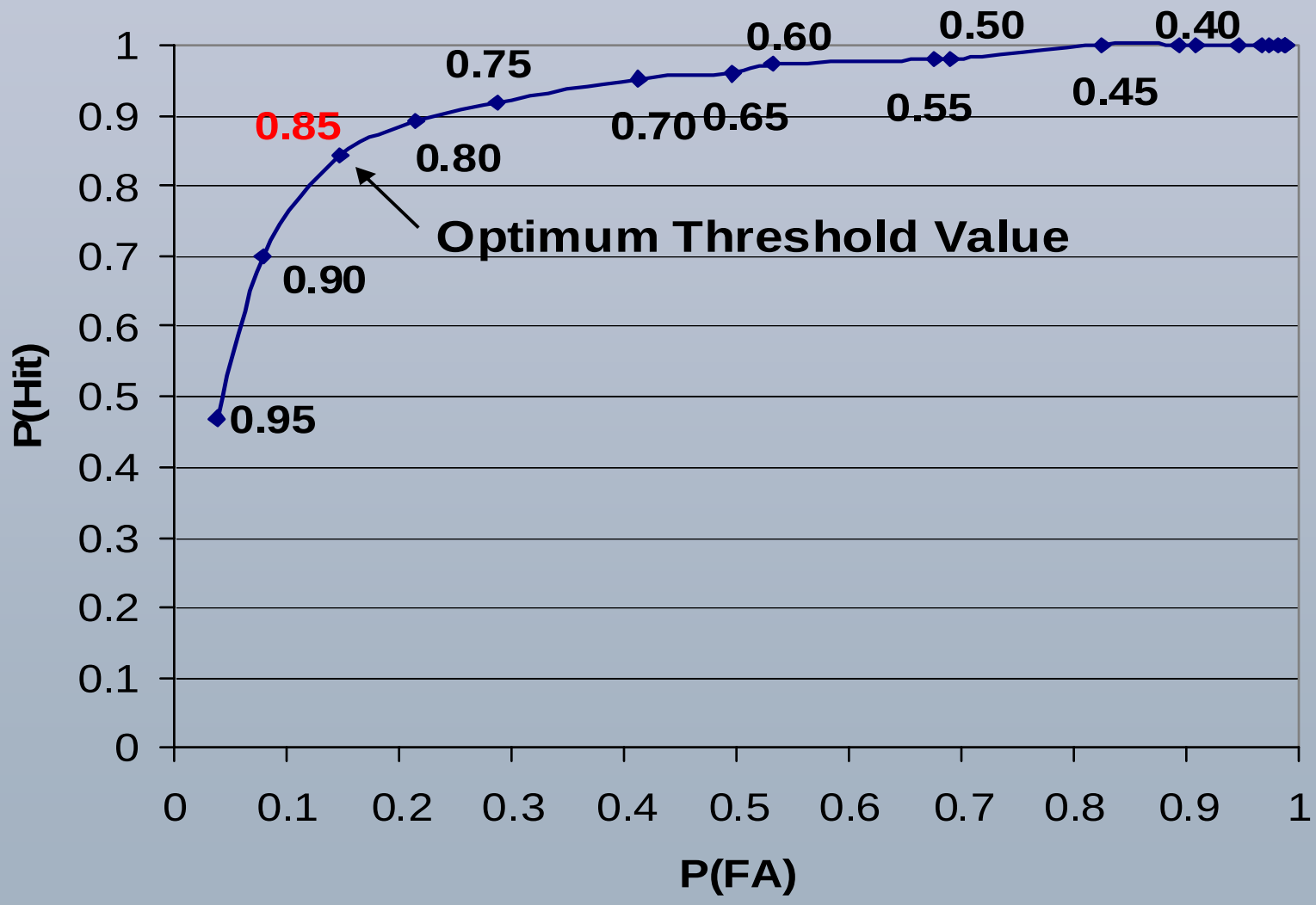
I. Receiver Operating Characteristic (ROC) Filtering

Methods

- ROC curves were plotted for categories (sensitivity vs. 1- specificity).
- Determine the optimal cut off point of prediction strengths
  - Balance the selection of errors and amount of manual coding required at that filter (threshold).

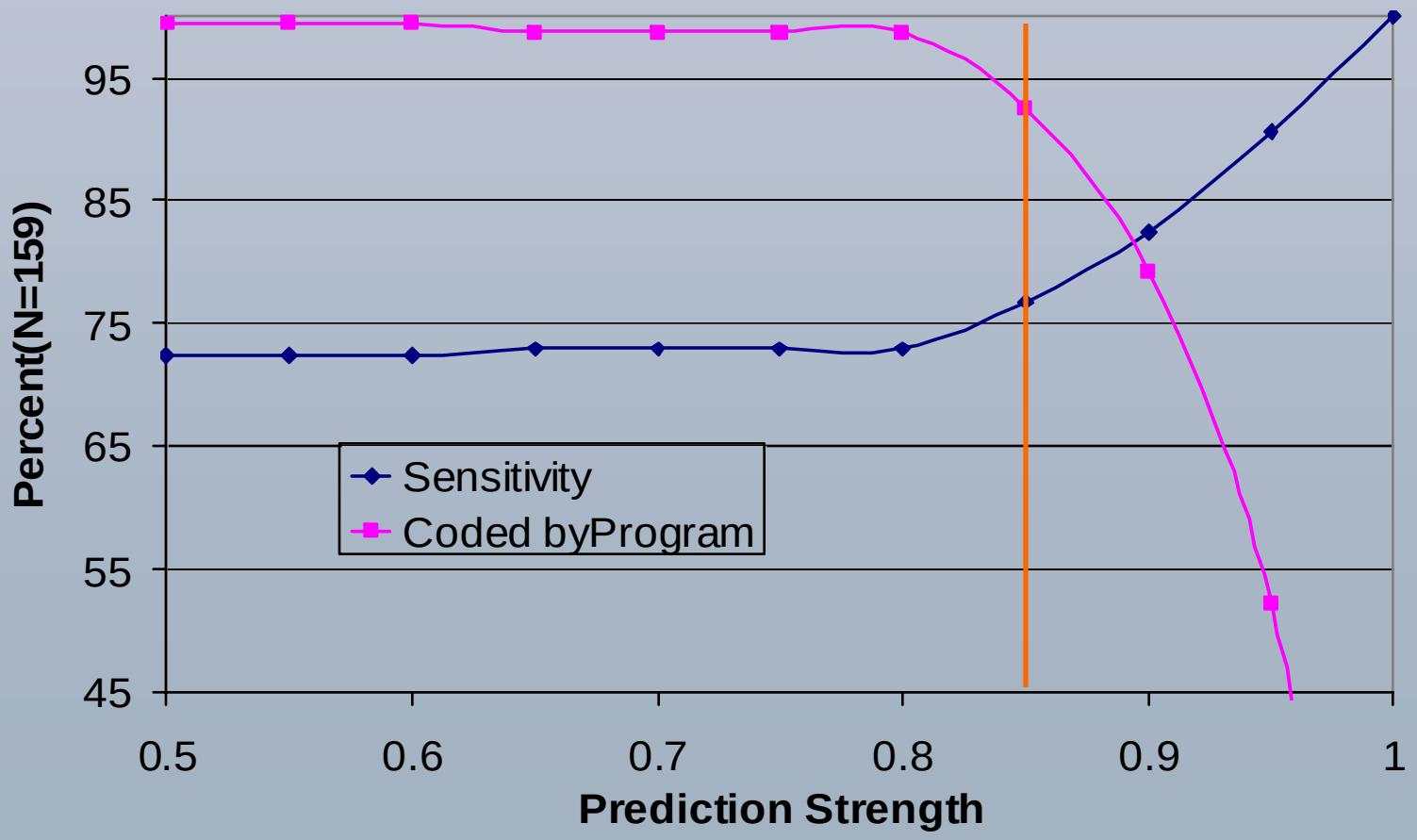
Research to Reality™ from

Example ROC for 'Fall to Lower Level'

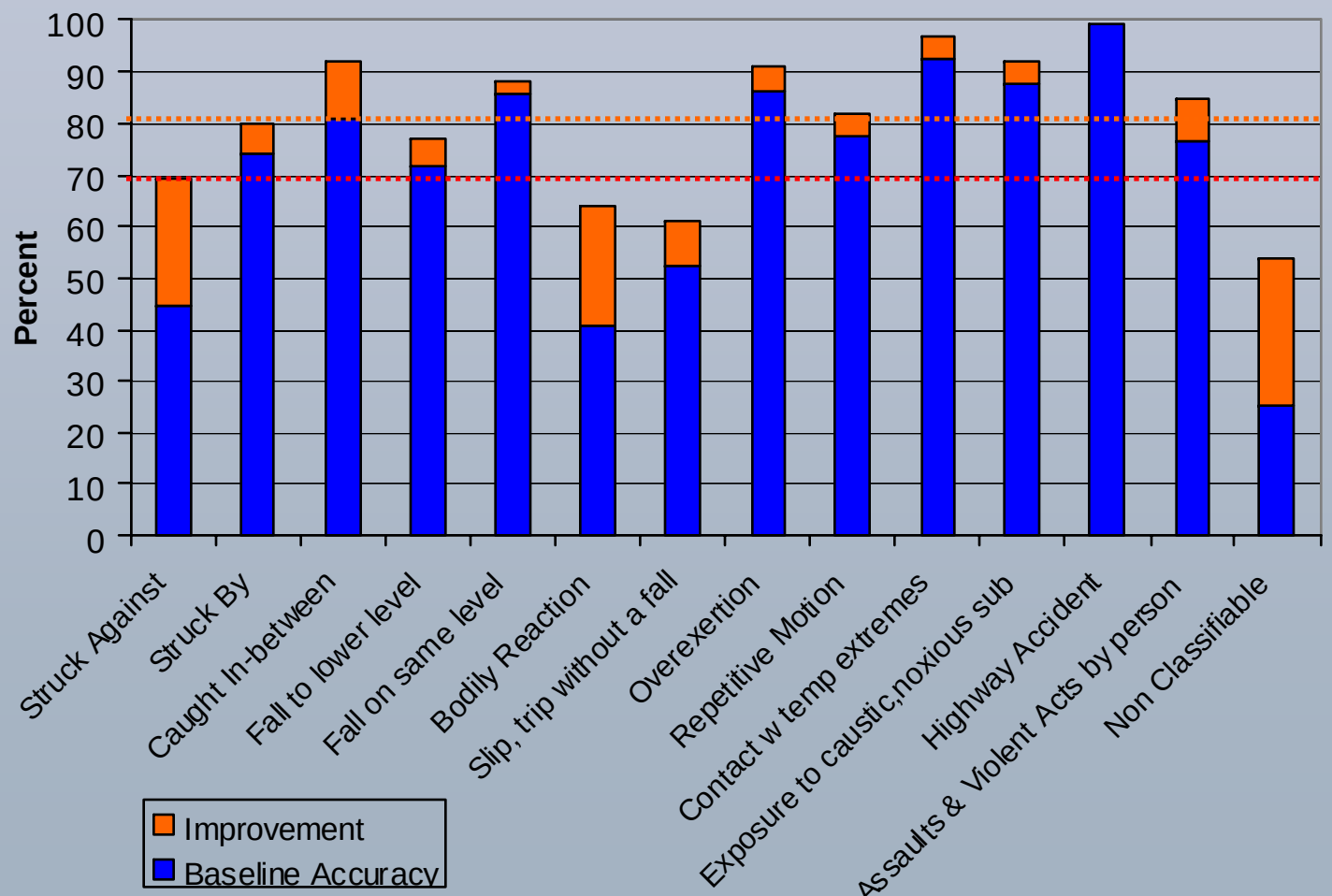


Effect of Threshold on Accuracy: Fall to Lower Level

- A stricter threshold means more manual coding



Comparison with Baseline of ROC Filtering – 2 digit improvement



from Research to Reality™

II. Linear Predictor Threshold Filtering

Methods

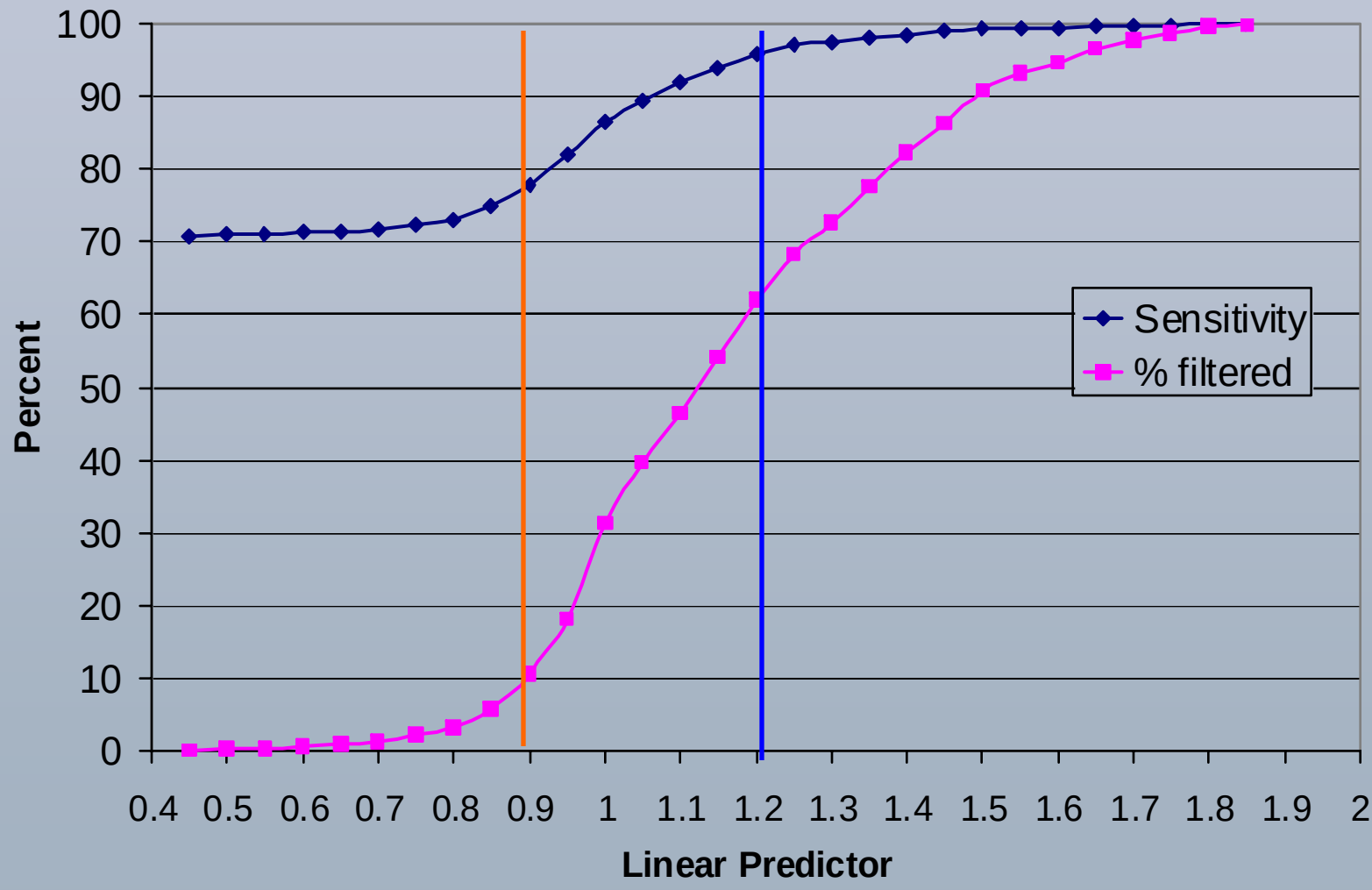
- A linear predictor was created using both
  - The difference in prediction strengths of the top two predictors
  - The actual strength of the highest predictor

    $LP=[P(Ai/n)_1-P(Ai/n)_2] + P(Ai/n)_1$

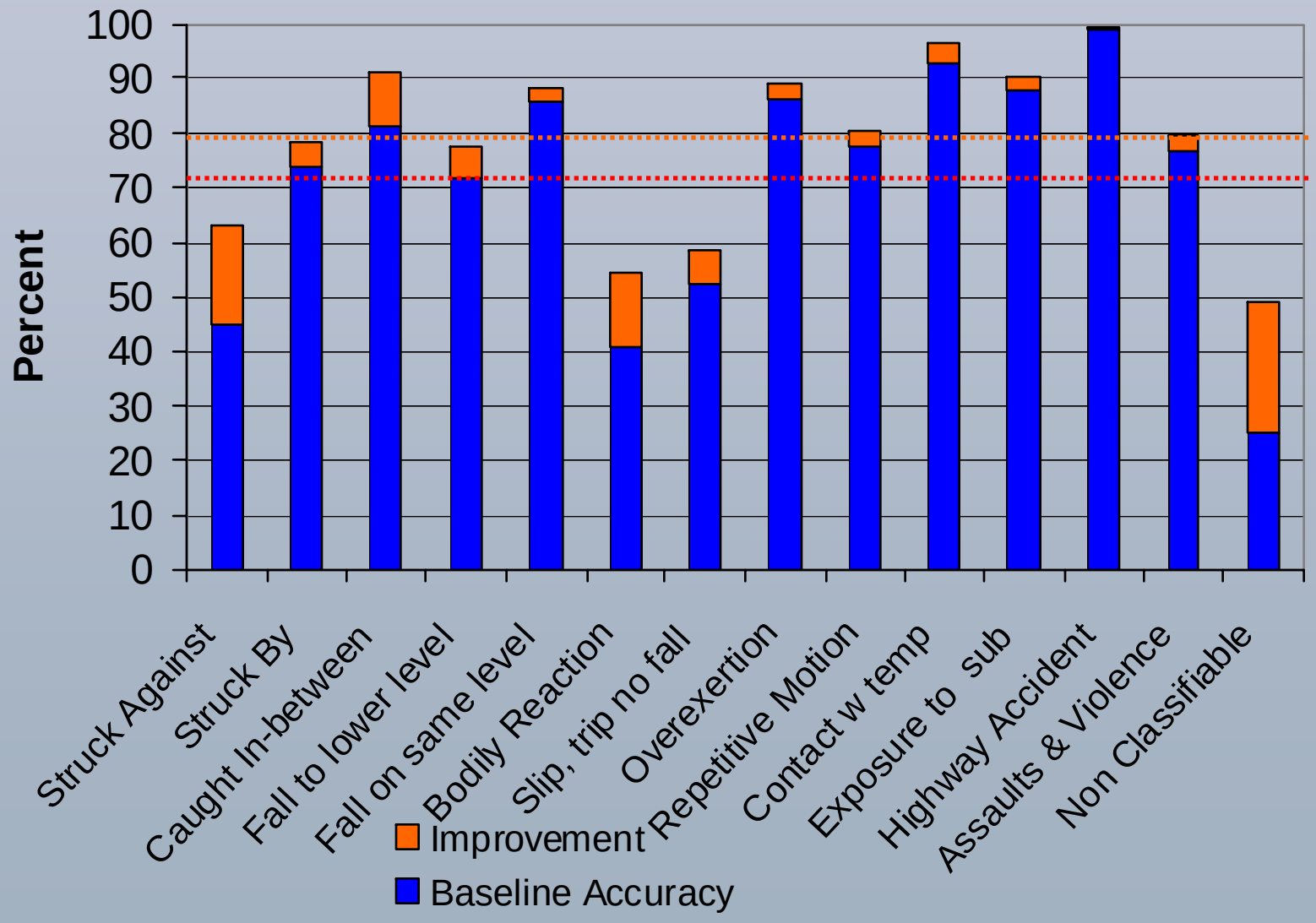
    An example might be: $P(Ai/n)_1=0.76$, $P(Ai/n)_2=0.74$

    So LP = [0.76 – 0.74] + 0.76 = 0.78
- Determine the optimal cut off point of prediction strengths
  - Balance the selection of errors and amount of manual coding required at that linear predictor level.

Balancing Resources & Sensitivity



Comparison with Baseline of LP Filtering – 2 digit improvement



Conclusions

- Filtering improves accuracy substantially from baseline
- This approach opens up the opportunity to code narrative data from large administrative databases
- Other classification protocols can be used because the program learns from manually coded narratives

www.libertymutualresearch.com

Liberty Mutual®

Generating knowledge to help people live safer, more secure lives.