

A DATA-DRIVEN EARLY WARNING SYSTEM FOR MINING ACCIDENTS

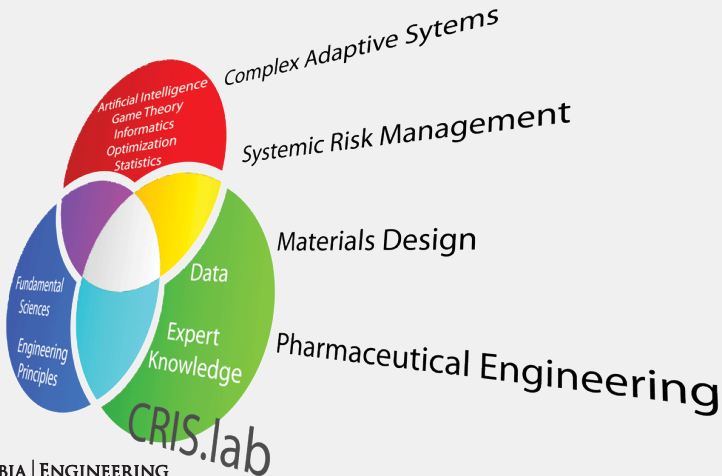
Yu Luo, Ashutosh Nanda, Shivaram Rajgopal, Vinay Ramesh, Zhizun Zhang, Catherine Zhao, and Venkat Venkatasubramanian

Chemical Engineering, Computer Science, and Business School
Columbia University

3/27/2017

- 1 MINE SAFETY: A DATA-DRIVEN APPROACH
- 2 METHODS: DATA SOURCES AND MODEL PRELIMINARIES
- 3 RESULTS AND DISCUSSION
- 4 CONCLUSION

COMPLEX, RESILIENT, INTELLIGENT SYSTEMS (CRIS LAB)



SYSTEMIC RISK

- Systemic disasters
 - SARS (2003)
 - Northeast Blackout (2003)
 - Subprime Crisis (2008)
 - Deepwater Horizon Oil Spill (2010)
- Emerging systemic risks
 - Climate change
 - Income/wealth inequality
 - Cyber-physical security
 - Technological singularity
- Fast-paced and connected
- Prevent systemic disasters
- Analyze systemic risk: go beyond one-off accidents

UPPER BIG BRANCH MINE DISASTER (2010)

- April 5, 2010, Raleigh County, West Virginia, owned by Massey Energy
- 29 deaths, the worst mining in the United States since 1970
- MSHA cites corporate culture as root cause of Upper Big Branch Mine disaster

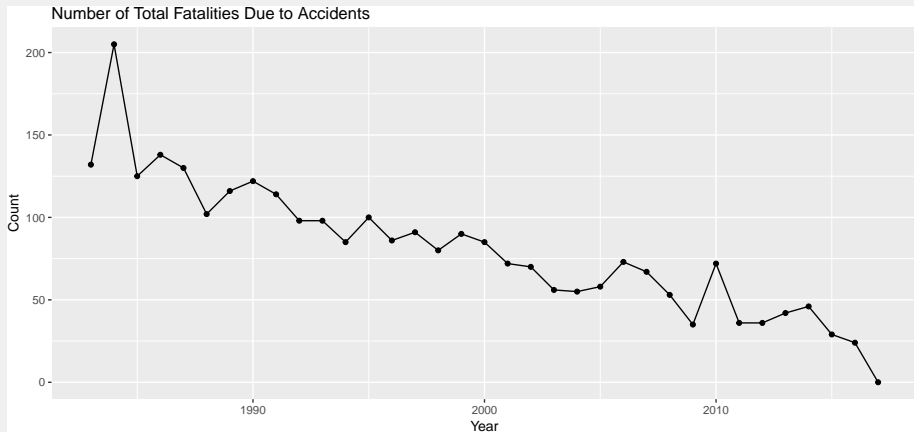
SAGO MINE DISASTER (2006)

- January 2, 2006, Sago, West Virginia, owned by Anker West Virginia Mining
- 13 miners were trapped for nearly two days; only one survived
- Fatality number was exceeded by the Upper Big Branch Mine disaster
- MSHA reports prior history of safety violations and fatalities

MINE SAFETY AND HEALTH ADMINISTRATION (MSHA)

- Formed in 1977, agency of the U.S. Department of Labor
- Mission
 - Prevent death, illness, and injury from mining
 - Promote safe and healthful workplaces for U.S. miners
 - Develop and enforce safety and health rules
 - Provide technical, educational, and other types of assistance
- A constantly improving industry in terms of safety

FATALITY TREND SINCE 1983



CAN WE FURTHER IMPROVE MINE SAFETY?

- Process MSHA safety data
- Understand the underlying causal relationships
- Develop early warning systems based on past behaviors
- Credit rating/score analogy
 - Predict default probability within 18 months
 - Accidents: defaults a month or a year prior to application
 - Violations: missed payments, late payments, etc.
- Can we develop a “credit score” for mine safety?

DEPARTMENT OF LABOR ENFORCEMENT DATA

- Link: https://enforcedata.dol.gov/views/data__catalogs.php
- Updated daily or weekly
- Publicly available
 - Department of Labor: MSHA, OSHA, etc.
 - Other departments: EPA, FDA, DOJ, etc.

MSHA DATA: SOURCES

- Mine accidents table: “msha__accident.csv”
 - 681,386 rows
 - Retrieved 1/26/2017, from
https://enforcedata.dol.gov/views/data_summary.php
- MSHA assessed violations table: “AssessedViolations.csv”
 - 2,169,804 rows
 - Retrieved 12/10/2016, from
<https://arlweb.msha.gov/OpenGovernmentData/OGIMSHA.asp>

MSHA DATA: ADVANTAGES

- Each mine has a unique mine ID, e.g., Upper Big Branch (4608436)
- Rich details: e.g., classification, description, and severity
- Selected attributes from the accidents table (omitting 42 attributes)

## [1]	"mine_id"	"controller_id"	"cal_yr"
## [4]	"cal_qtr"	"ai_dt"	"inj_degr_desc"
## [7]	"ai_class_desc"	"ai_occ_desc"	"ai_acty_desc"
## [10]	"exper_tot_calc"	"exper_mine_calc"	"exper_job_calc"
## [13]	"ai_narr"	"accident_type_cd"	"no_injuries"
## [16]	"days_restrict"	"days_lost"	

MSHA DATA: CHALLENGES

- Missing data, human errors
- No information about inactive/nonoperating mines
- Most data are not numeric
- Lots of zeros, few severe accidents ($\sim 0.5\%$)

CONSOLIDATED DATA

- Group and summarize accidents/violations by mines
- 664,128 rows, 10,377 unique mines
- From 2000 to 2015
- Each row represents data for a unique combination of mine, year, and quarter
 - e.g., Upper Big Branch Mine in the second quarter of 2010
- Each row contains both current and past information
 - i.e., current quarter, past quarter, past year, and past three years

CONSOLIDATED DATA

- All 25 attributes of the consolidated data

## [1]	"mine_id"	"mine.name"
## [3]	"year"	"quarter"
## [5]	"active"	"num.days.lost"
## [7]	"last.quarter.lost"	"last.year.lost"
## [9]	"last.three.years.lost"	"num.days.restrict"
## [11]	"last.quarter.restrict"	"last.year.restrict"
## [13]	"last.three.years.restrict"	"num.death"
## [15]	"last.quarter.death"	"last.year.death"
## [17]	"last.three.years.death"	"num.dis"
## [19]	"last.quarter.dis"	"last.year.dis"
## [21]	"last.three.years.dis"	"viol.quantity"
## [23]	"last.quarter.viol"	"last.year.viol"
## [25]	"last.three.years.viol"	

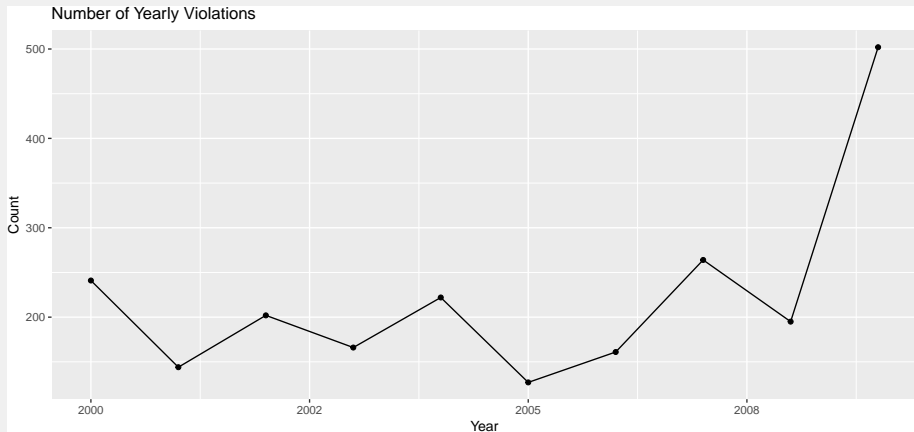
TOP 10 FATAL ACCIDENTS SINCE 2005

- Query the consolidated data on the deadliest accidents

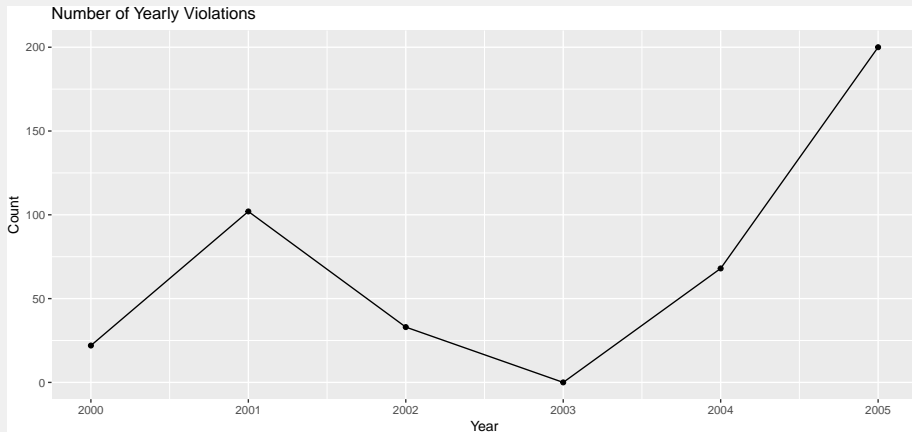
##		mine.name	mine_id	year	quarter	num.death
## 1	Upper Big Branch Mine-South	4608436	2010	2	29	
## 2	Sago Mine	4608791	2006	1	12	
## 3	Crandall Canyon Mine	4201715	2007	3	9	
## 4	Darby Mine No 1	1518185	2006	2	5	
## 5	Gibson Mine	1202215	2007	3	3	
## 6	Affinity Mine	4608878	2013	1	2	
## 7	Aracoma Alma Mine #1	4608801	2006	1	2	
## 8	Black Stallion UG Mine	4609086	2014	2	2	
## 9	Cucumber Mine	4609066	2007	1	2	
## 10	D-14 Stillhouse	1517165	2005	3	2	

- Plot violation trends prior to disasters

VIOLATION TREND: UPPER BIG BRANCH



VIOLATION TREND: SAGO MINE



PREDICTIVE MODEL

- Rising violation trends before disasters
- Develop a disaster classifier based on historical data
- Define a **severe** accident as one with death or permanent disability

```
## # A tibble: 2 × 3
##   severe      n perc
##   <lgl>   <int> <dbl>
## 1  FALSE 477077 99.46
## 2   TRUE  2608  0.54
```

FIXED-MINE EFFECTS

- Biostatisticians and epidemiologists call it “conditional logistic regression” (`survival::clogit`)
- Suitable for **panel data** (e.g., longitudinal data, our consolidated data)
- Model includes mine-specific biases
- Logistic function (for every mine)

$$\Pr(Y = 1|\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \beta\mathbf{X})}}$$

- Logistic function with fixed effects (for the i -th mine)

$$\Pr(Y = 1|\mathbf{X}, i) = \frac{1}{1 + e^{-(\alpha_i + \beta\mathbf{x})}}$$

LOGISTIC REGRESSION WITHOUT FIXED EFFECTS

■ In-sample model

Reference

Prediction FALSE TRUE

FALSE 477011 2600

TRUE 66 8

Accuracy Sensitivity Specificity Precision F1

0.9944 0.0031 0.9999 0.1081 0.0060

- $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{P} + \text{N})$
- $\text{Sensitivity/recall} = \text{TP} / \text{P}$
- $\text{Specificity} = \text{TN} / \text{N}$
- $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$
- F1: harmonic mean of sensitivity and precision

LOGISTIC REGRESSION WITHOUT FIXED EFFECTS

- Fails to predict top 10 deadliest disasters

##		mine.name	year	quarter	severe	pred
## 1	Upper Big Branch Mine-South	2010	2	TRUE	FALSE	
## 2	Sago Mine	2006	1	TRUE	FALSE	
## 3	Crandall Canyon Mine	2007	3	TRUE	FALSE	
## 4	Darby Mine No 1	2006	2	TRUE	FALSE	
## 5	Gibson Mine	2007	3	TRUE	FALSE	
## 6	Affinity Mine	2013	1	TRUE	FALSE	
## 7	Aracoma Alma Mine #1	2006	1	TRUE	FALSE	
## 8	Black Stallion UG Mine	2014	2	TRUE	FALSE	
## 9	Cucumber Mine	2007	1	TRUE	FALSE	
## 10	D-14 Stillhouse	2005	3	TRUE	FALSE	

LOGISTIC REGRESSION WITHOUT FIXED EFFECTS

- List of false positive predictions based on predicted probability

##		mine.name	year	quarter	severe	pred
## 1	The American Coal Company	New Era Mine	2008	3	FALSE	TRUE
## 2	The American Coal Company	New Era Mine	2008	2	FALSE	TRUE
## 3	The American Coal Company	New Era Mine	2007	4	FALSE	TRUE
## 4	The American Coal Company	New Era Mine	2008	4	FALSE	TRUE
## 5	The American Coal Company	New Era Mine	2008	1	FALSE	TRUE
## 6	The American Coal Company	New Era Mine	2009	1	TRUE	TRUE
## 7	The American Coal Company	New Era Mine	2007	3	FALSE	TRUE
## 8	The American Coal Company	New Era Mine	2006	1	FALSE	TRUE
## 9	The American Coal Company	New Era Mine	2005	4	FALSE	TRUE
## 10	The American Coal Company	New Era Mine	2006	2	TRUE	TRUE

LOGISTIC REGRESSION WITH FIXED EFFECTS

- Out-of-sample model (randomly select half of the data to train and the other half to test)

Reference

Prediction FALSE TRUE

FALSE 141332 483

TRUE 97167 852

Accuracy Sensitivity Specificity Precision F1

0.5928 0.6382 0.5926 0.0087 0.0172

LOGISTIC REGRESSION WITH FIXED EFFECTS

- Successfully predicts top 10 deadliest disasters

##		mine.name	year	quarter	severe	pred
## 1		Sago Mine	2006	1	TRUE	TRUE
## 2		Crandall Canyon Mine	2007	3	TRUE	TRUE
## 3		Darby Mine No 1	2006	2	TRUE	TRUE
## 4		Cucumber Mine	2007	1	TRUE	TRUE
## 5		Dotiki Mine	2010	2	TRUE	TRUE
## 6		Equality	2011	4	TRUE	TRUE
## 7		Meikle Mine	2010	3	TRUE	TRUE
## 8		Nanuuq Gold Project	2007	3	TRUE	TRUE
## 9	4 J's	Gravel Crushing Plant 2	2011	3	TRUE	TRUE
## 10		Adams	2006	3	TRUE	TRUE

LOGISTIC REGRESSION WITH FIXED EFFECTS

- Still has a lot of false positive predictions
- List of false positive predictions based on predicted probability

##		mine.name	year	quarter	severe	pred
## 1	The American Coal Company	New Era Mine	2006	1	FALSE	TRUE
## 2		Upper Big Branch Mine-South	2009	3	FALSE	TRUE
## 3		Upper Big Branch Mine-South	2009	1	FALSE	TRUE
## 4		Upper Big Branch Mine-South	2006	4	FALSE	TRUE
## 5		Upper Big Branch Mine-South	2005	1	FALSE	TRUE
## 6	The American Coal Company	New Era Mine	2005	3	FALSE	TRUE
## 7	The American Coal Company	New Era Mine	2008	1	FALSE	TRUE
## 8	The American Coal Company	New Era Mine	2007	4	FALSE	TRUE
## 9		Upper Big Branch Mine-South	2006	1	FALSE	TRUE
## 10		Upper Big Branch Mine-South	2006	3	FALSE	TRUE

- What happened in the New Era Mine?

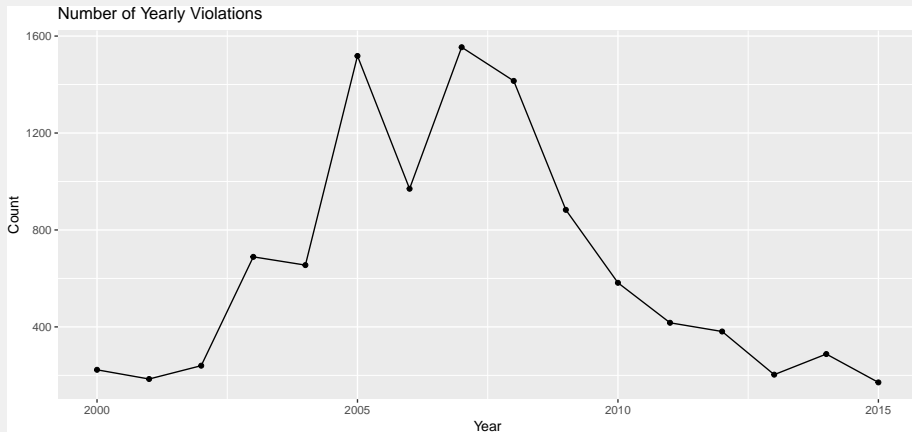
NEW ERA MINE

- Among the worst mines by number of days lost due to accidents

##	mine.name	year	quarter	num.days.lost
## 1	The American Coal Company New Era Mine	2005	2	2940
## 2	The American Coal Company New Era Mine	2003	2	2914
## 3	The American Coal Company New Era Mine	2005	3	2874
## 4	Mathies	2002	1	2840
## 5	The American Coal Company New Era Mine	2004	3	2613
## 6	The American Coal Company New Era Mine	2004	1	2591
## 7	Monongalia County Mine	2013	3	2563
## 8	The American Coal Company New Era Mine	2005	4	2487
## 9	Powhatan No. 6 Mine	2013	1	2409
## 10	Maple Creek	2001	1	2030

- Rising violation trend from 2000 to 2005

NEW ERA MINE



NEW LABELS INCLUDING DAYS LOST

- Updated severe accident label
 - Previously defined criteria plus days lost > 300
- Redo out-of-sample model

##		Reference	
##	Prediction	FALSE	TRUE
##	FALSE	148496	1267
##	TRUE	88426	1645

##	Accuracy	Sensitivity	Specificity	Precision	F1
##	0.626	0.565	0.627	0.018	0.035

- Worse true positive rate, improved F1 score

NEW LABELS INCLUDING DAYS LOST

- Successfully predicts 9 out of top 10 deadliest accidents

##	mine.name	year	quarter	severe	pred
## 1	Sago Mine	2006	1	TRUE	TRUE
## 2	Crandall Canyon Mine	2007	3	TRUE	TRUE
## 3	Darby Mine No 1	2006	2	TRUE	TRUE
## 4	Cucumber Mine	2007	1	TRUE	TRUE
## 5	Dotiki Mine	2010	2	TRUE	TRUE
## 6	Equality	2011	4	TRUE	TRUE
## 7	Meikle Mine	2010	3	TRUE	TRUE
## 8	Nanuuq Gold Project	2007	3	TRUE	TRUE
## 9	4 J's Gravel Crushing Plant	2011	3	TRUE	TRUE
## 10	Adams	2006	3	TRUE	FALSE

NEW LABELS INCLUDING DAYS LOST

- Accidents of the New Era mine are now true positives

##		mine.name	year	quarter	severe	pred
## 1	The American Coal Company	New Era Mine	2006	1	TRUE	TRUE
## 2	The American Coal Company	New Era Mine	2005	3	TRUE	TRUE
## 3	The American Coal Company	New Era Mine	2005	1	TRUE	TRUE
## 4		Monongalia County Mine	2014	3	TRUE	TRUE
## 5		Powhatan No. 6 Mine	2013	3	TRUE	TRUE
## 6		Powhatan No. 6 Mine	2013	4	TRUE	TRUE
## 7		Marshall County Mine	2015	4	TRUE	TRUE
## 8	The American Coal Company	New Era Mine	2008	1	TRUE	TRUE
## 9		Willow Lake Portal	2008	2	TRUE	TRUE
## 10		Powhatan No. 6 Mine	2013	1	TRUE	TRUE

CONCLUSION

■ Summary

- Two deadliest mine accidents in the last decade: Upper Big Branch & Sago
- Rich MSHA data that need clean-up
- Supervised predictive model

■ Application

- “Credit score” for mine safety
- Regulators, mines, stakeholders

■ Future

- Improve model performance
- Unsupervised clustering, neural nets, etc.
- Expand data: OSHA, EPA, etc.

APPENDIX: SIMPLE LINEAR MODEL

■ Adjusted $R^2 = 0.36$

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	0.5243	0.06725	7.8	6.4e-15
## last.quarter.lost	0.0566	0.00179	31.6	2.9e-218
## last.year.lost	0.0724	0.00093	77.8	0.0e+00
## last.three.years.lost	0.0338	0.00032	105.6	0.0e+00
## last.quarter.restrict	-0.0173	0.00461	-3.8	1.7e-04
## last.year.restrict	-0.0123	0.00243	-5.1	3.9e-07
## last.three.years.restrict	0.0072	0.00085	8.4	3.8e-17
## last.quarter.viol	0.3083	0.01095	28.1	3.5e-174
## last.year.viol	0.1352	0.00490	27.6	2.1e-167
## last.three.years.viol	-0.0346	0.00141	-24.7	4.2e-134
## last.quarter.death	-5.7149	1.09783	-5.2	1.9e-07
## last.year.death	-3.6943	0.64330	-5.7	9.3e-09
## last.three.years.death	-0.5155	0.33261	-1.5	1.2e-01

APPENDIX: UNSUPERVISED CLUSTERING

- Apply k -means clustering to consolidated data on all 20 features
- 3 clusters: low-risk, mid-risk, and high-risk
- Selected cluster centers (omitting 15 features)

##	num.days.lost	num.days.restrict	num.death	num.dis	viol.quantity
## low	5.3	2.1	0.0013	0.0029	2.6
## mid	100.5	18.6	0.0164	0.0313	34.3
## high	508.4	32.7	0.0431	0.0871	98.9

- Cluster sizes

##	low	mid	high
## size	465203	13299	1183

APPENDIX: MARKOV CHAIN

■ Overall transition matrix

```
##          low   mid   high
## low  0.997 0.003 0.000
## mid  0.087 0.906 0.006
## high 0.000 0.072 0.928
```

■ Steady-state distribution

```
##          low   mid   high
## [1,] 0.97 0.028 0.003
```