

# A DATA-DRIVEN EARLY WARNING SYSTEM FOR MINING ACCIDENT

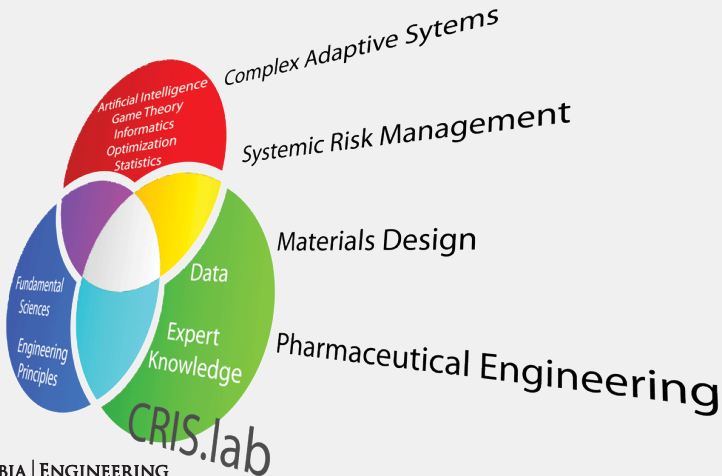
Yu Luo, Ashutosh Nanda, Shiva Rajgopal, Vinay Ramesh,  
Venkat Venkatasubramanian, Zhizun Zhang, and Catherine Zhao

Chemical Engineering, Computer Science, and Business School  
Columbia University

3/27/2017

- 1 INTRODUCTION: A DATA APPROACH TO MINE SAFETY
- 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
- Design complex systems
- Analyze systemic risk

# 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

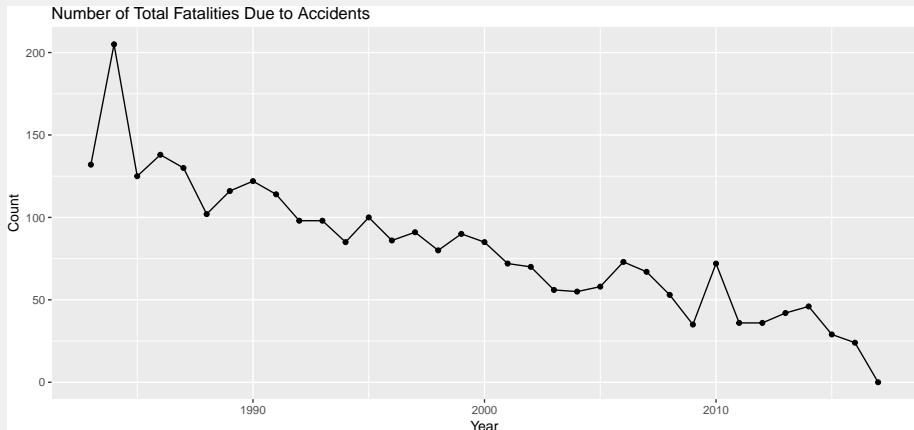
# SAGO MINE DISASTER (2006)

- 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 United States 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

# 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.

# DEPARTMENT OF LABOR ENFORCEMENT DATA

- Link: [https://enforcedata.dol.gov/views/data\\_catalogs.php](https://enforcedata.dol.gov/views/data_catalogs.php)
- Updated daily or weekly
- Enforcement Data Catalog
  - MSHA Enforcement Data
  - OSHA Enforcement Data
  - Wage and Hour Compliance Action Data
  - OFCCP Compliance Evaluation and Complaint Investigation Data
  - EBSA Enforcement Data

# 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](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., selected attributes from the accidents table:

## [1]	"mine_id"	"controller_id"	"ai_dt"
## [4]	"inj_degr_desc"	"ai_class_desc"	"ai_occ_desc"
## [7]	"ai_acty_desc"	"exper_tot_calc"	"exper_mine_calc"
## [10]	"exper_job_calc"	"ai_narr"	"cal_yr"
## [13]	"cal_qtr"	"accident_type_cd"	"accident_type"
## [16]	"no_injuries"	"occupation_cd"	"activity_cd"
## [19]	"injury_source_cd"	"injury_source"	"nature_injury_cd"
## [22]	"nature_injury"	"inj_body_part_cd"	"inj_body_part"
## [25]	"schedule_charge"	"days_restrict"	"days_lost"
## [28]	"trans_term"	"return_to_work_dt"	"immed_notify_cd"
## [31]	"immed_notify"	"invest_begin_dt"	"closed_doc_no"
## [34]	"coal_metal_ind"	"load_dt"	"ai_year"

# MSHA DATA: CHALLENGES

- Missing data
- Inactive mines
- Most data are not numeric
- Relatively few severe accidents ( $\sim 0.5\%$ )

# MODEL PRELIMINARIES

- Create a master data table with selected attributes
- Plot violation trends for Upper Big Branch Mine and Sago Mine
- Design a predictive model

# MASTER DATA TABLE

- Combine and summarize accidents/violations based on mines
- 664,128 rows, 10,377 unique mines
- From 2000 to 2015 in quarters
- 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

# MASTER DATA TABLE

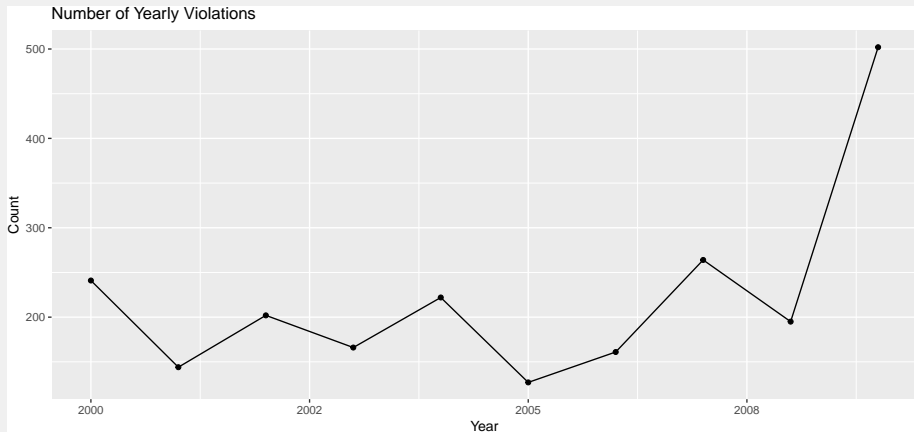
##	[1]	"mine_id"	"mine.name"
##	[3]	"year"	"quarter"
##	[5]	"active"	"num.days.lost"
##	[7]	"last.quarter.lost"	"last.year.lost"
##	[9]	"last.three.years.lost"	"num.days.restrict"
##	[11]	"last.quarter.restrict"	"last.year.restrict"
##	[13]	"last.three.years.restrict"	"num.death"
##	[15]	"last.quarter.death"	"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"	



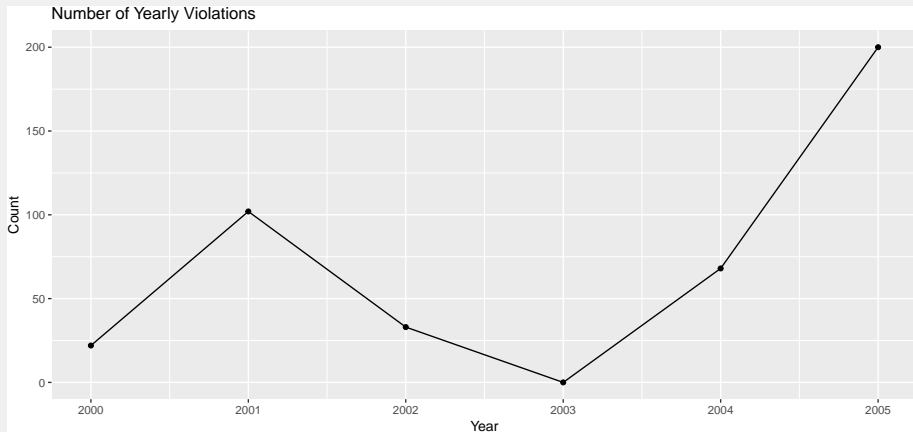
# TOP 10 FATAL ACCIDENTS SINCE 2005

##		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	

# VIOLATION TREND: UPPER BIG BRANCH



# VIOLATION TREND: SAGO MINE



# PREDICTIVE MODEL

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

```
## # 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., our master data table)
- Model includes mine-specific but time-invariant variables (e.g., same slope but different intercepts for different mines)
- Logistic regression (for every mine)

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

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

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

# LOGISTIC REGRESSION WITHOUT FIXED EFFECTS

## ■ Train and test on all data

##		Reference	
##	Prediction	FALSE	TRUE
##	FALSE	477011	2600
##	TRUE	66	8

##	Sensitivity	Specificity	Pos Pred Value
##	0.003067	0.999862	0.108108
##	Neg Pred Value	Precision	Recall
##	0.994579	0.108108	0.003067
##	F1	Prevalence	Detection Rate
##	0.005966	0.005437	0.000017
##	Detection Prevalence	Balanced Accuracy	
##	0.000154	0.501465	

# LOGISTIC REGRESSION WITHOUT FIXED EFFECTS

## ■ True positives

##		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

## ■ Top predicted positives (possibly false positives)

##		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

- Randomly select half of the data to train and the other half to test

##	Reference		
##	Prediction	FALSE	TRUE
##	FALSE	141332	483
##	TRUE	97167	852

##	Sensitivity	Specificity	Pos Pred Value
##	0.6382	0.5926	0.0087
##	Neg Pred Value	Precision	Recall
##	0.9966	0.0087	0.6382
##	F1	Prevalence	Detection Rate
##	0.0172	0.0056	0.0036
##	Detection Prevalence	Balanced Accuracy	
##	0.4087	0.6154	

# LOGISTIC REGRESSION WITH FIXED EFFECTS

## ■ Top true positives (deadliest accidents)

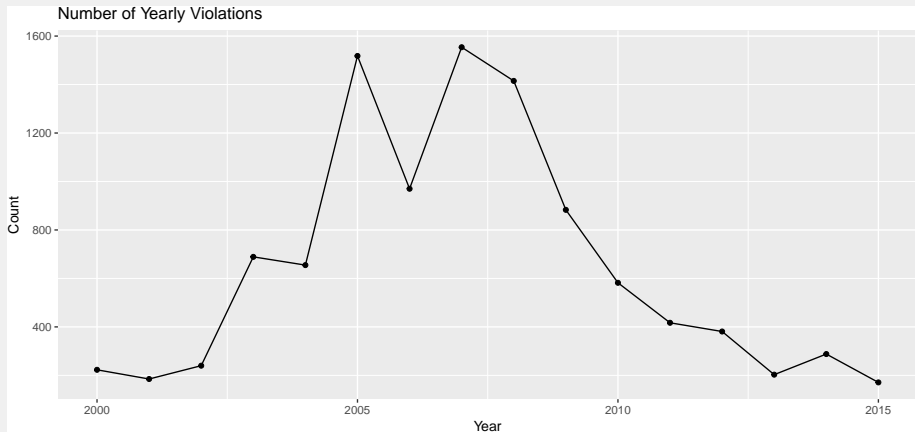
##		mine.name	year	quarter	severe	pred
## 1		Sago Mine	2006	1	TRUE	TRUE
## 2		Crandall Canyon Mine	2007	3	TRUE	TRUE
## 3		Darby Mine No 1	2006	2	TRUE	TRUE
## 4		Cucumber Mine	2007	1	TRUE	TRUE
## 5		Dotiki Mine	2010	2	TRUE	TRUE
## 6		Equality	2011	4	TRUE	TRUE
## 7		Meikle Mine	2010	3	TRUE	TRUE
## 8		Nanuuq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

# LOGISTIC REGRESSION WITH FIXED EFFECTS

## ■ Top predicted positives (possibly false positives)

##	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

# NEW ERA MINE



# NEW LABELS INCLUDING DAYS LOST

- Previously defined severe accidents and days lost  $> 300$

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

##	Sensitivity	Specificity	Pos Pred Value
##	0.5649	0.6268	0.0183
##	Neg Pred Value	Precision	Recall
##	0.9915	0.0183	0.5649
##	F1	Prevalence	Detection Rate
##	0.0354	0.0121	0.0069
##	Detection Prevalence	Balanced Accuracy	
##	0.3756	0.5958	

# NEW LABELS INCLUDING DAYS LOST

## ■ Top true positives (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 2	2011	3	TRUE	TRUE
## 10	Adams	2006	3	TRUE	FALSE

# NEW LABELS INCLUDING DAYS LOST

## ■ Top predicted positives (possibly false 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
- MSHA



# 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