

# A DATA-DRIVEN EARLY WARNING SYSTEM FOR MINING ACCIDENT

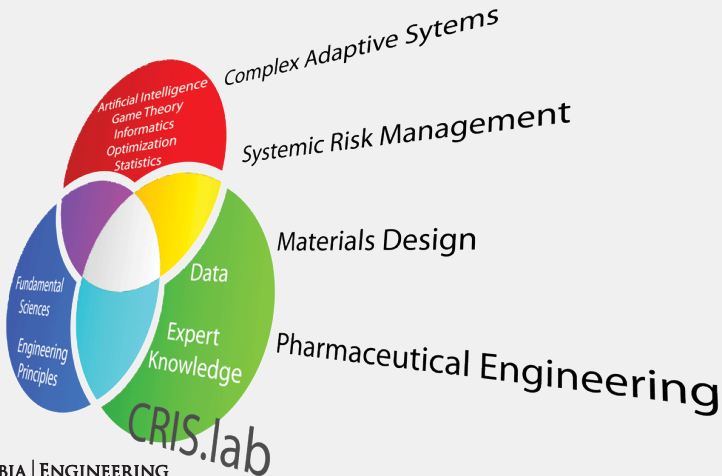
Yu Luo, Ashutosh Nanda, Shiva Rajgopal, Vinay Ramesh,  
Venkat Venkatasubramanian, Zhizun Zhang, and Catherine Zhao

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

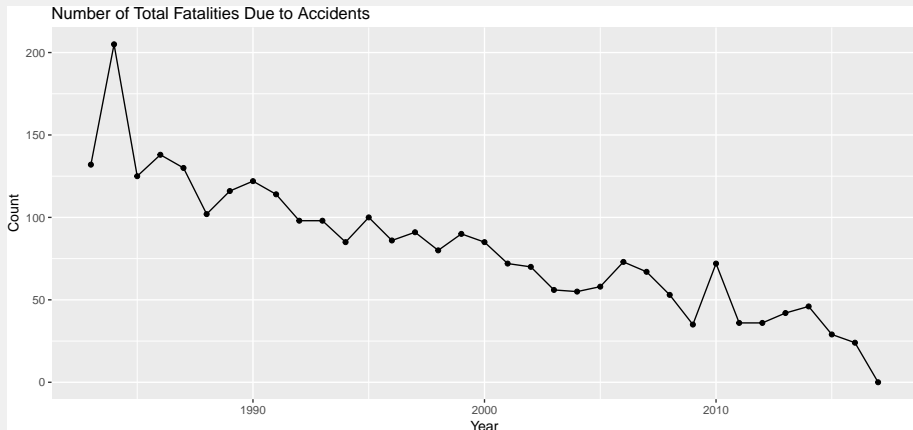
# SAGO MINE DISASTER (2006)

- 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

# PRIMARY DATASETS

- 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>

# MASTER DATA TABLE

- 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

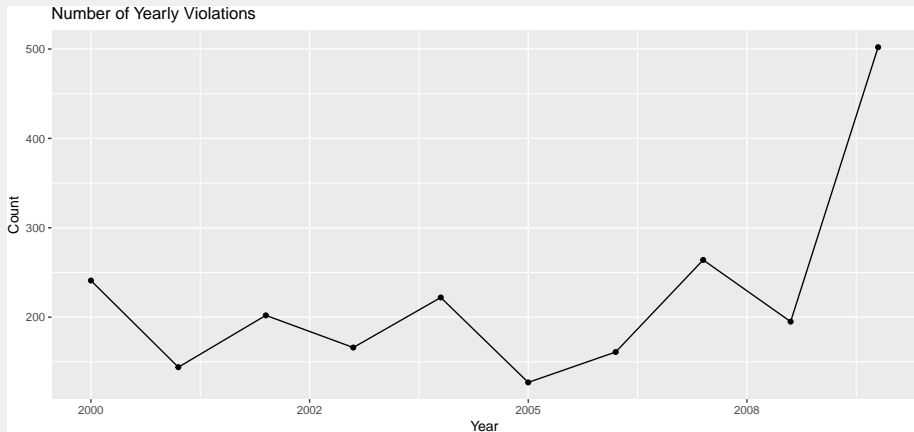
# MASTER DATA TABLE

##	[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"	

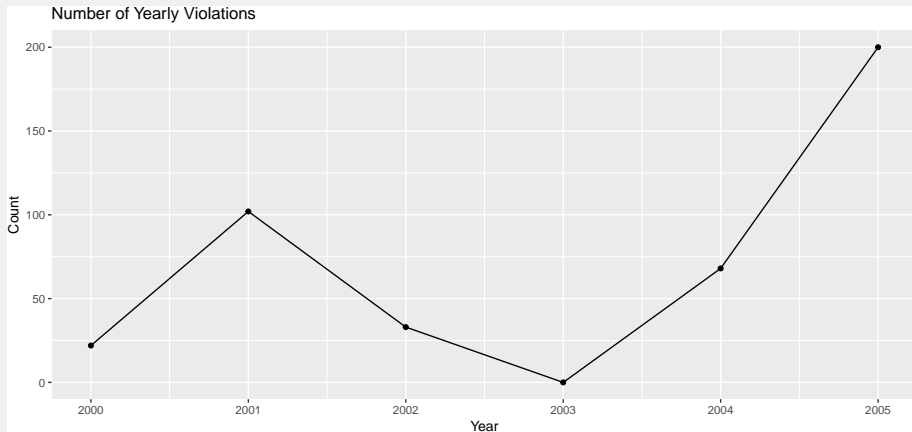
# TOP 10 FATAL ACCIDENTS SINCE 2005

##		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	

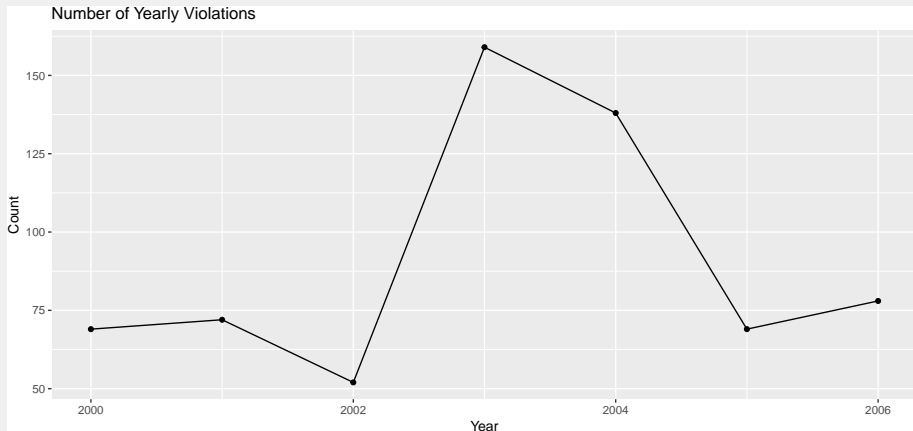
# VIOLATION TREND: UPPER BIG BRANCH



# VIOLATION TREND: SAGO MINE

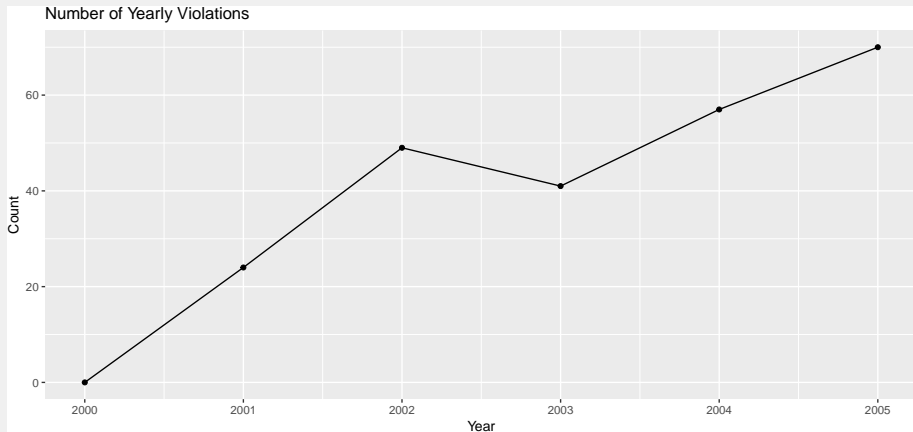


# VIOLATION TREND: CRANDALL CANYON

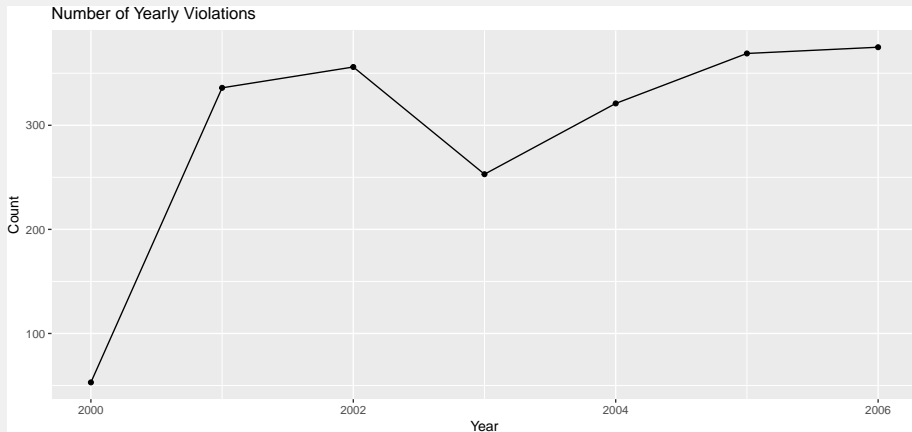




# VIOLATION TREND: DARBY MINE No. 1



# VIOLATION TREND: GIBSON



# 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 661520 99.61
## 2  TRUE  2608  0.39
```

# PROCESSING DATA

## ■ Remove **inactive** quarters

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

# PROCESSING DATA

- Consider fixed-mine effects
  - Biostatisticians and epidemiologists call it “conditional logistic regression” (R, `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 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

- Sort result based on number of deaths

##	mine.name	num.death	severe	prob	pred
## 1	Upper Big Branch Mine-South	29	TRUE	0.0492	FALSE
## 2	Sago Mine	12	TRUE	0.0187	FALSE
## 3	Crandall Canyon Mine	9	TRUE	0.0059	FALSE
## 4	Darby Mine No 1	5	TRUE	0.0055	FALSE
## 5	Gibson Mine	3	TRUE	0.0377	FALSE
## 6	Affinity Mine	2	TRUE	0.0162	FALSE
## 7	Aracoma Alma Mine #1	2	TRUE	0.0104	FALSE
## 8	Black Stallion UG Mine	2	TRUE	0.0786	FALSE
## 9	Cucumber Mine	2	TRUE	0.0094	FALSE
## 10	D-14 Stillhouse	2	TRUE	0.0118	FALSE

# 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

- Sort result based on number of deaths

##	mine.name	num.death	severe	prob	pred
## 1	Sago Mine	12	TRUE	0.90	TRUE
## 2	Crandall Canyon Mine	9	TRUE	0.76	TRUE
## 3	Darby Mine No 1	5	TRUE	0.82	TRUE
## 4	Cucumber Mine	2	TRUE	0.69	TRUE
## 5	Dotiki Mine	2	TRUE	0.59	TRUE
## 6	Equality	2	TRUE	0.63	TRUE
## 7	Meikle Mine	2	TRUE	0.66	TRUE
## 8	Nanuuq Gold Project	2	TRUE	0.60	TRUE
## 9	4 J's Gravel Crushing Plant 2	1	TRUE	0.62	TRUE
## 10	Adams	1	TRUE	0.51	TRUE

# LOGISTIC REGRESSION WITH FIXED EFFECTS

- Sort result based on model probability

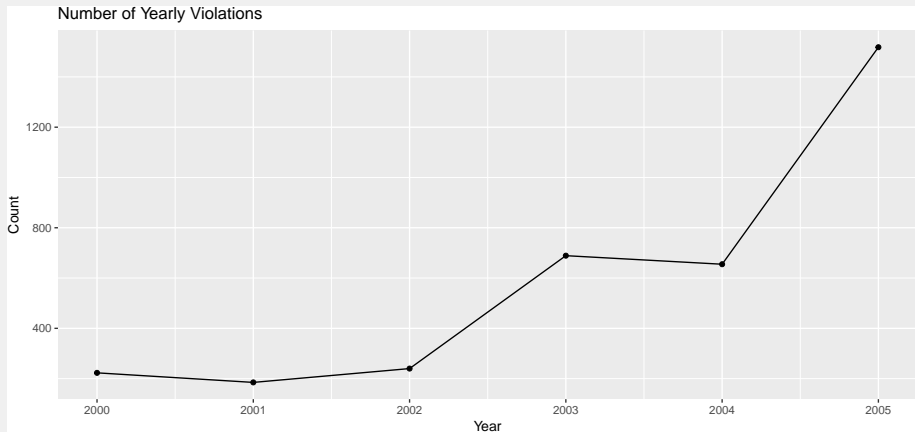
##	mine.name	num.death	severe	prob	pred
## 1	The American Coal Company New Era Mine	0	FALSE	0.99	TRUE
## 2	Upper Big Branch Mine-South	0	FALSE	0.99	TRUE
## 3	Upper Big Branch Mine-South	0	FALSE	0.98	TRUE
## 4	Upper Big Branch Mine-South	0	FALSE	0.98	TRUE
## 5	Upper Big Branch Mine-South	0	FALSE	0.98	TRUE
## 6	The American Coal Company New Era Mine	0	FALSE	0.98	TRUE
## 7	The American Coal Company New Era Mine	0	FALSE	0.98	TRUE
## 8	The American Coal Company New Era Mine	0	FALSE	0.98	TRUE
## 9	Upper Big Branch Mine-South	0	FALSE	0.98	TRUE
## 10	Upper Big Branch Mine-South	0	FALSE	0.98	TRUE

# NEW ERA MINE

- The American Coal Company New Era Mine occupies 6 slots on the top 10 list of most 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

# NEW ERA MINE



# SIMPLE LINEAR MODEL ON NUMBER OF DAYS LOST

■ 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

# CONCLUSION

■ TBD