Learning from Data

Aaron Clauset
@aaronclauset

Assistant Professor of Computer Science
University of Colorado Boulder
External Faculty, Santa Fe Institute

distribution $p(x)$

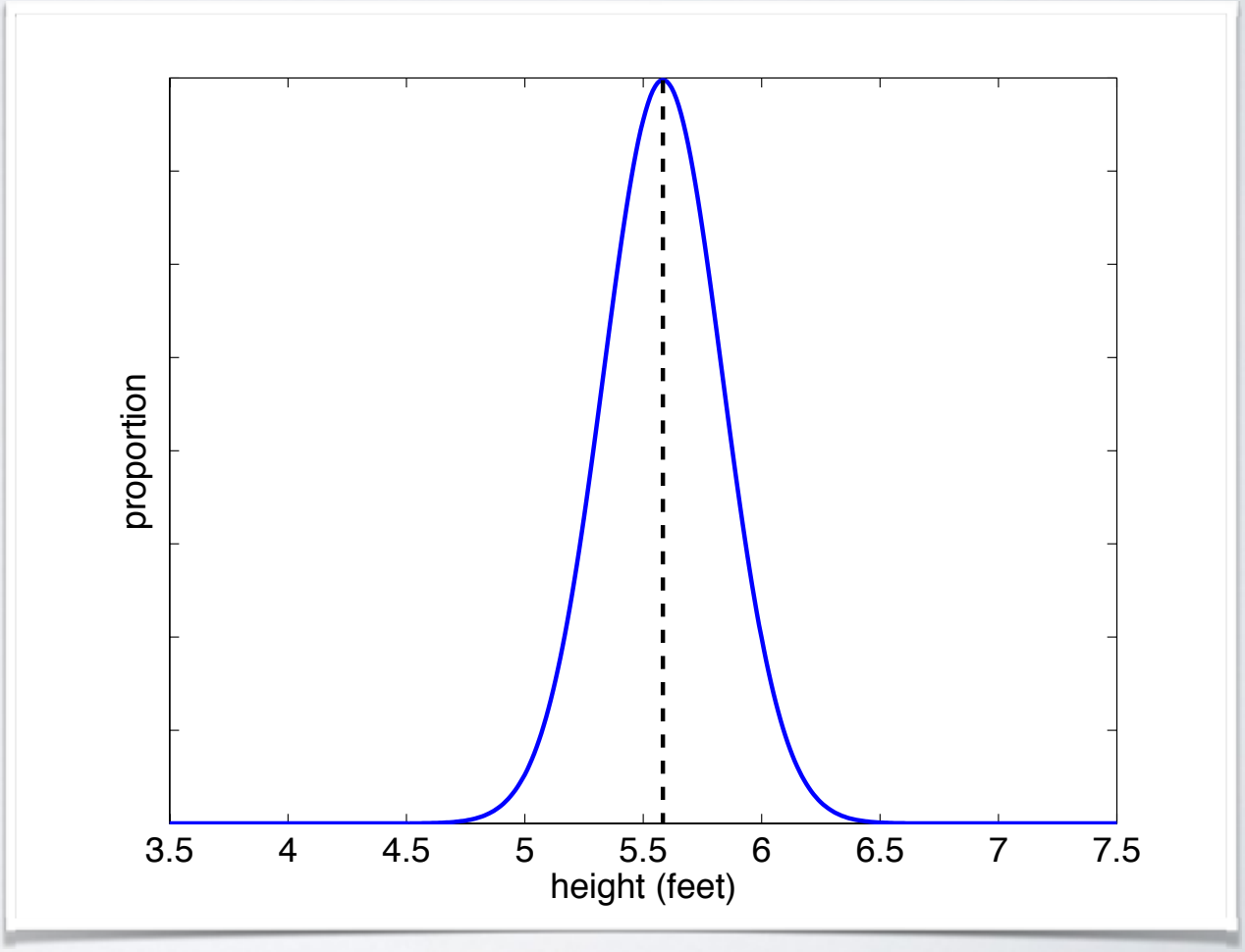
the fraction of times we observe an event of size x

normal distribution

the bell curve

height

average height of an American is 5.583 feet

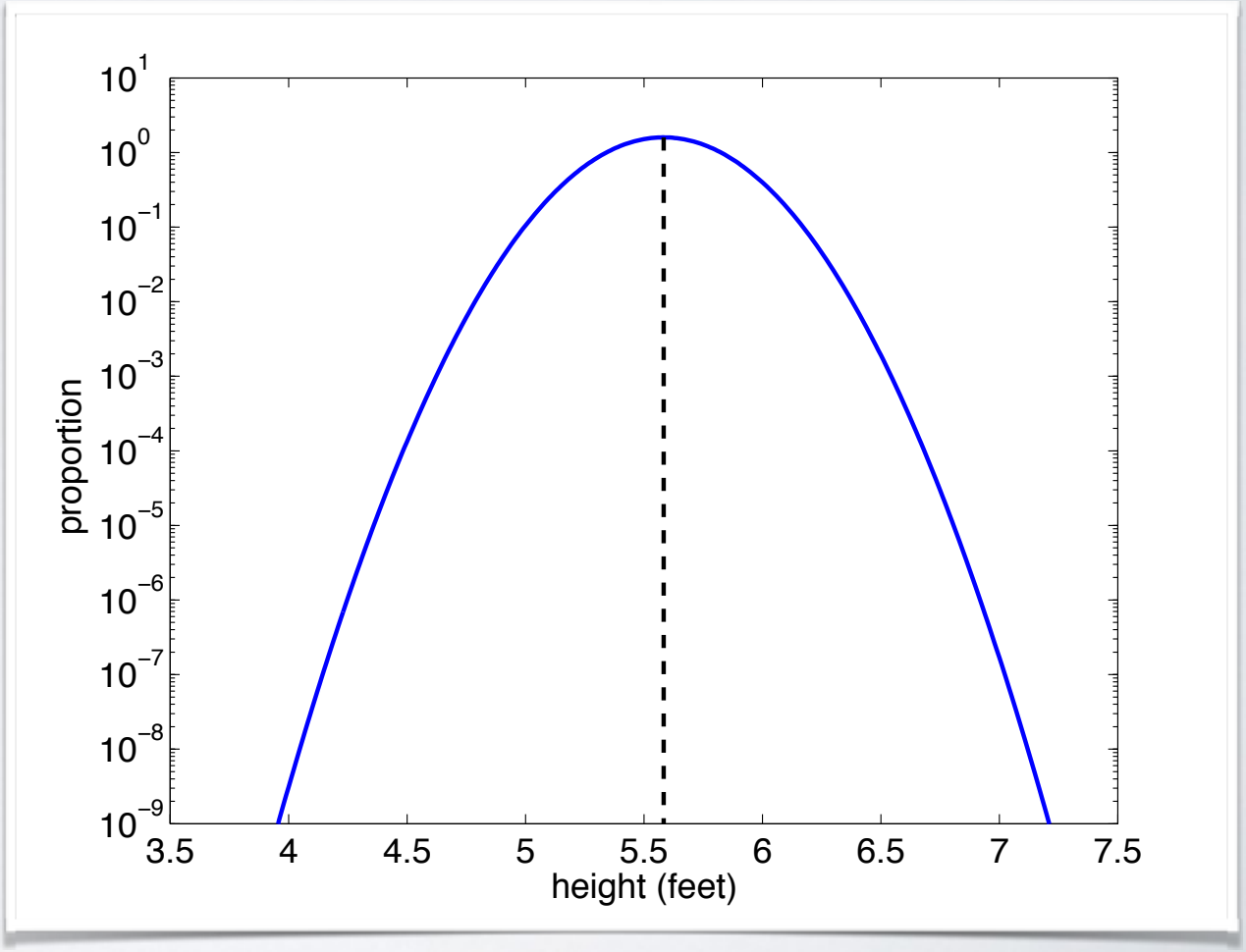


normal distribution

the bell curve

height

average height of an American is 5.583 feet

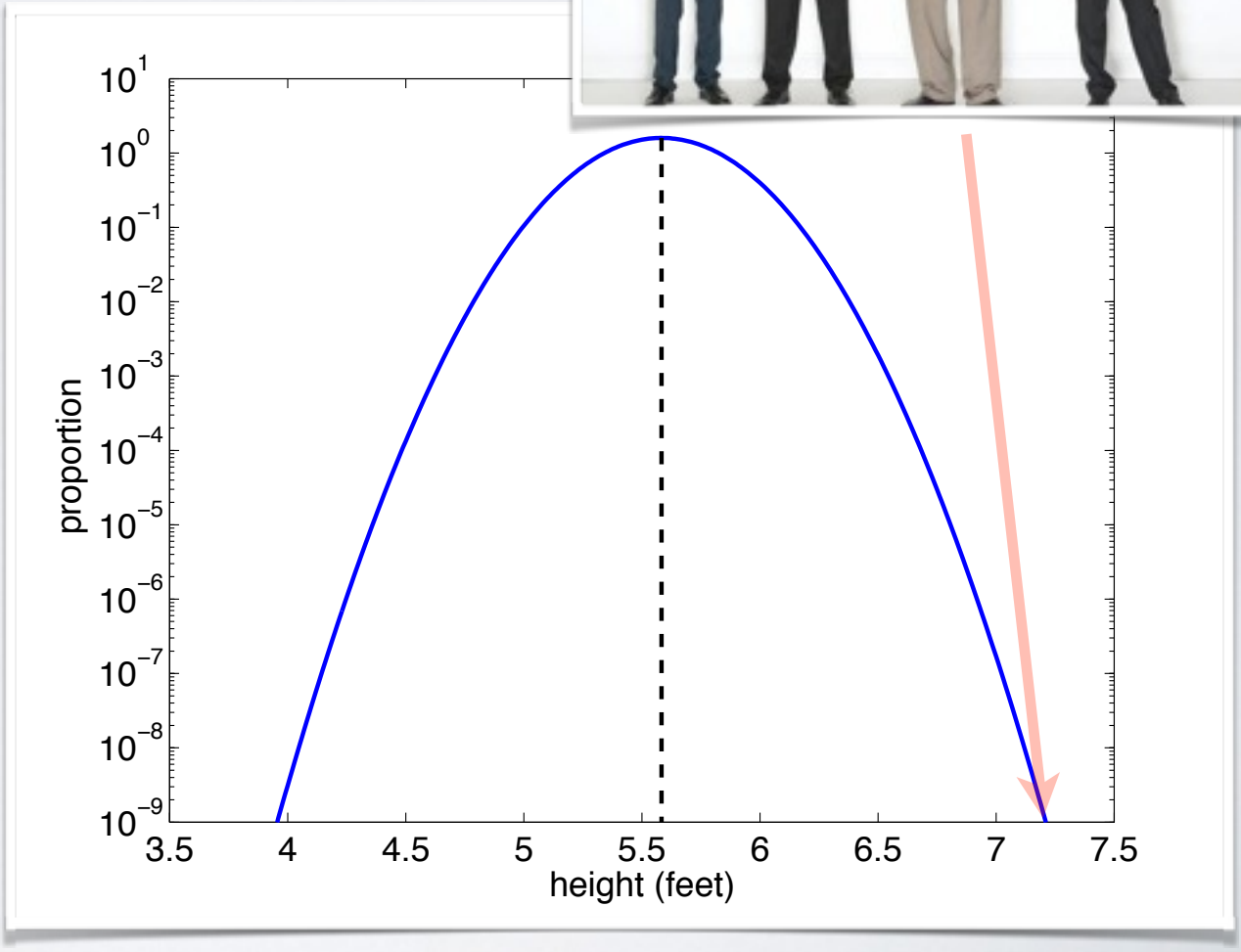


normal distribution

the bell curve

height

average height of an American is 5.583 feet



normal distribution

the bell curve

average is representative

lengths (human height, etc.)
weights (human or otherwise)
speeds (highway, running, etc.)

power law distributions

"heavy-tailed" patterns

average is not representative

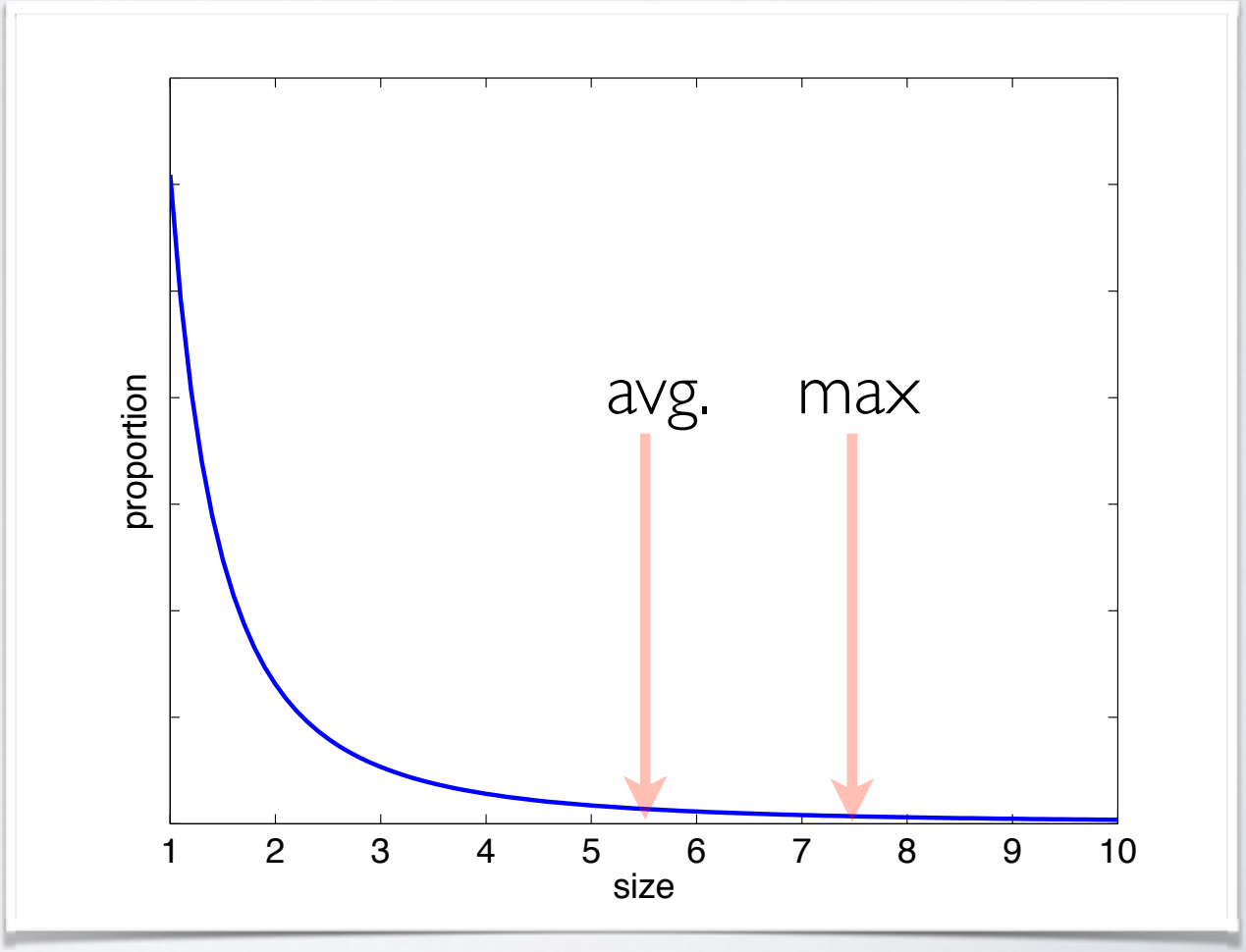
power law distributions

"heavy-tailed" patterns

power law

with average 5.583

most values very small!



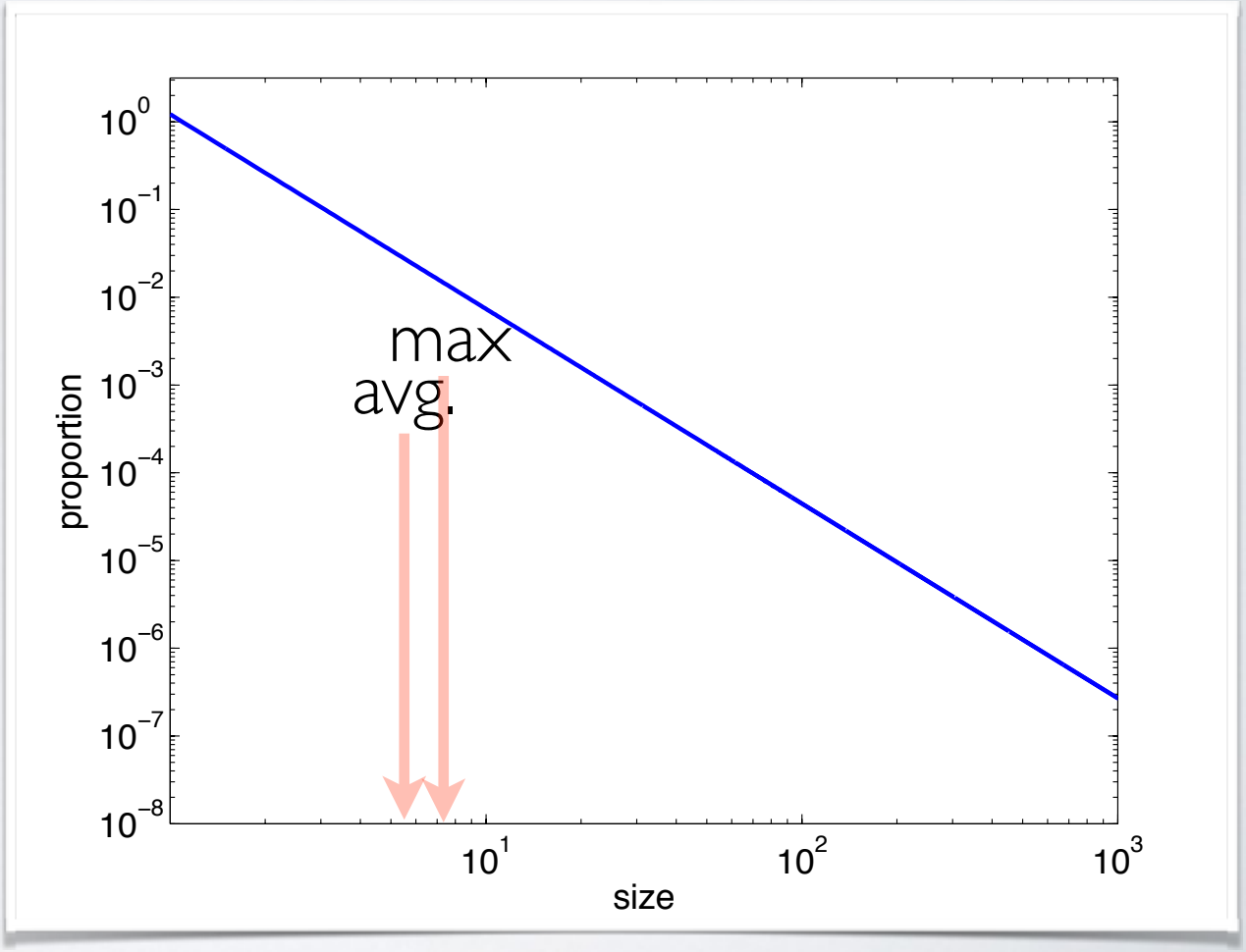
power law distributions

"heavy-tailed" patterns

power law

with average 5.583

most values very small



power law distributions

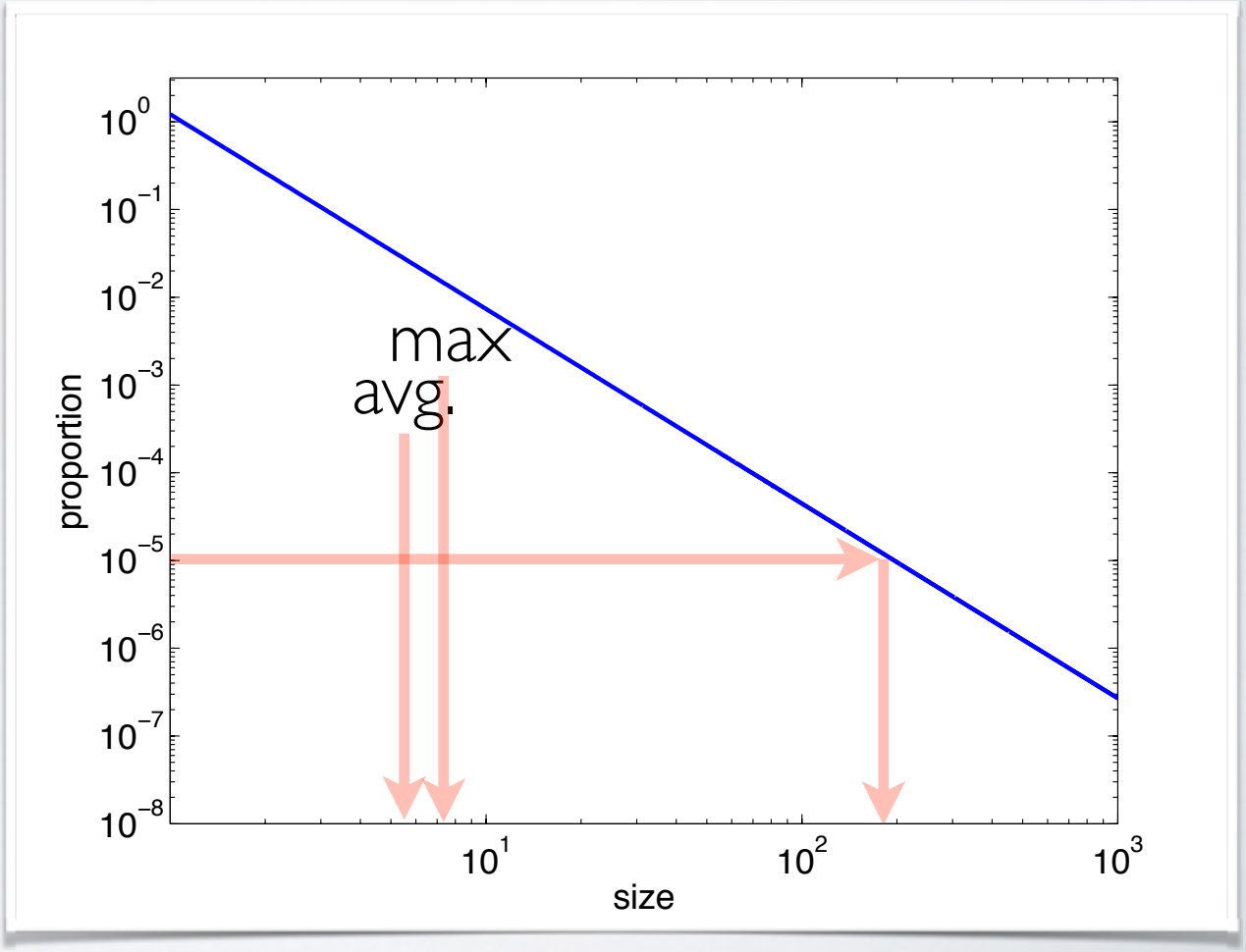
"heavy-tailed" patterns

power law

with average 5.583

most values very small

but a few values VERY big



power law distributions

"heavy-tailed" patterns

power law distributions

average is not representative

earthquake energy
solar flare energy
flood water volume
forest fire size
landslide size
lunar crater size

terrorism events
international wars
book sales
electrical blackouts
financial wealth
city population
financial returns
surname frequency

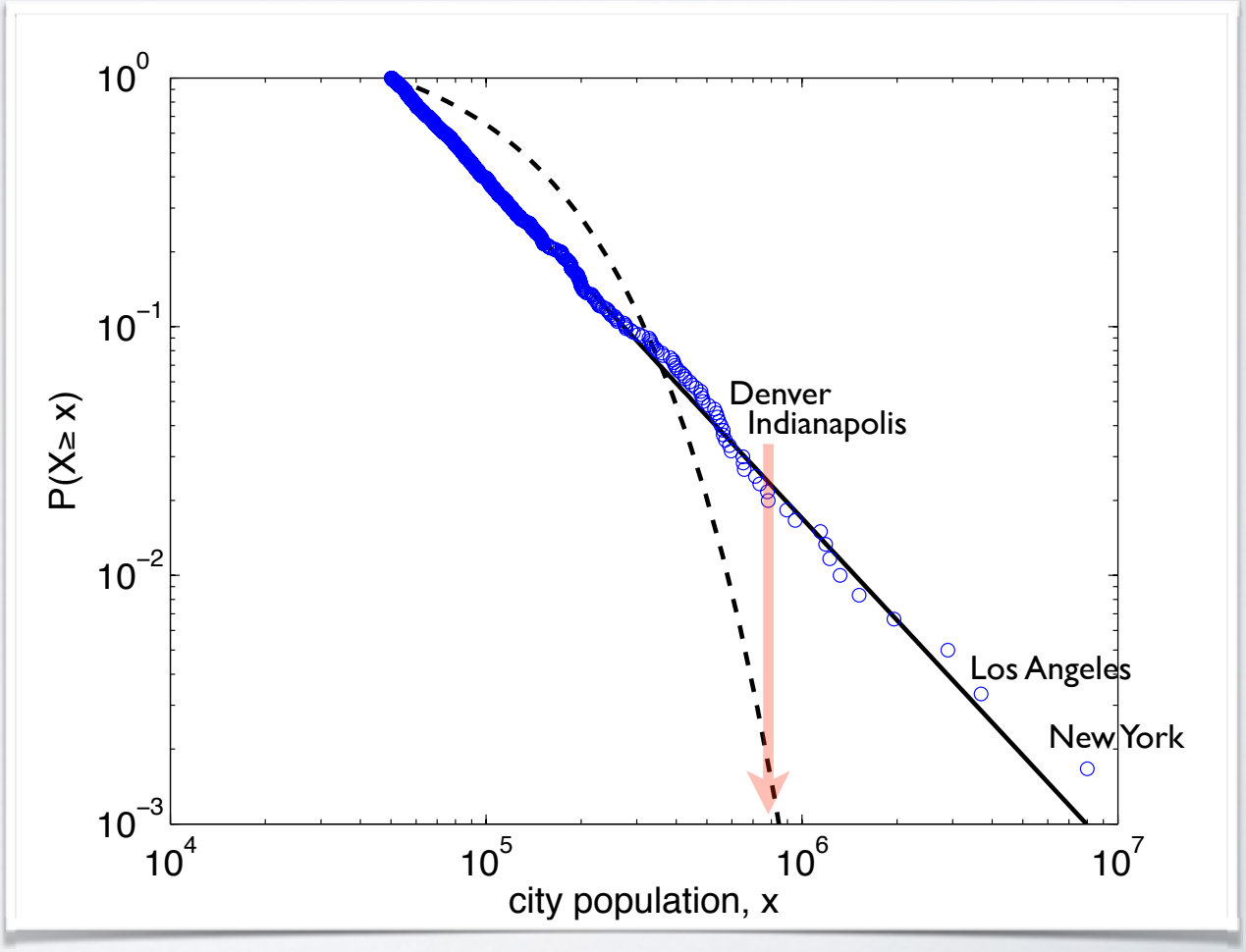
power laws vs. normals

counter-intuitive

power laws vs. normals

counter-intuitive

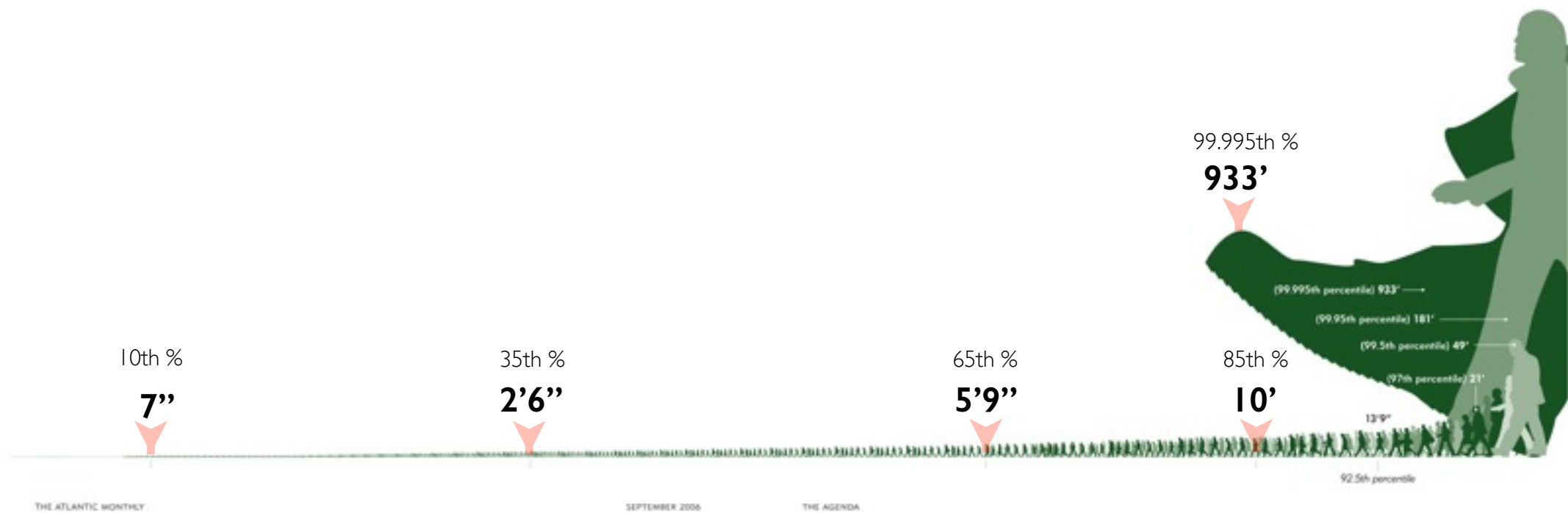
500 largest US cities



power laws vs. normals

counter-intuitive

financial wealth as human heights



"The Height of Inequality" (2006) in The Atlantic

terrorism

terrorism

MIPT TERRORISM
KNOWLEDGE BASE SM

incident profile

ABU HAFS AL-MASRI BRIGADE AND SECRET ORGANIZATION OF AL-QAEDA IN EUROPE ATTACKED TRANSPORTATION TARGET (JULY 7, 2005, UNITED KINGDOM)

Incident Date: July 7, 2005

Terrorist Organization(s): Abu Hafs al-Masri Brigade , Secret Organization of al-Qaeda in Europe

Target: Transportation

City: London

Country: United Kingdom

Region: Western Europe

Tactic: Bombing

Weapon: Explosives

Fatalities: 27

Injuries: 0

RAND-MIPT data

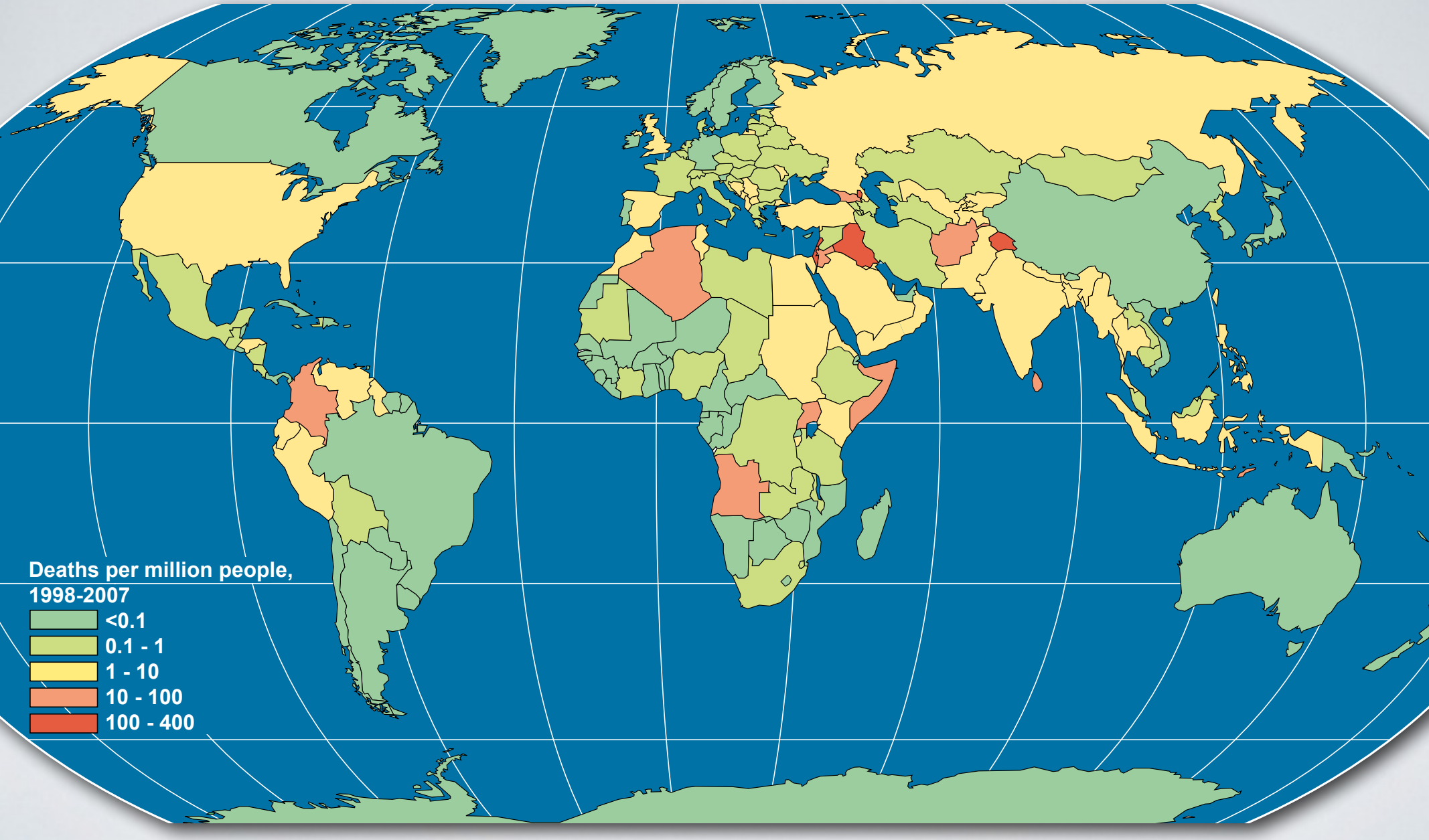
- 40 years (1968-2008)
- domestic + international
- 5000+ cities, 187 countries
- 36,018 events (37% deadly)

terrorism

- where does terrorism occur?
- what is risk of dying from terrorism?

data analysis:

1. *take all events 1998-2007*
2. *count deaths in each country*
3. *divide total by country's population*
4. *yields per capita risk of death*
5. *visualize on a world map*


Deaths per million people,
1998-2007
<0.1
0.1 - 1
1 - 10
10 - 100
100 - 400

deaths per million people, USA 2007

terrorism	0	--
lightning	0.15	1
bee sting	0.18	x1.2
airplane crash	0.23*	x1.5
homicide	61.74*	x408.3
car crash	124.43	x829.5

sources: , US Census, MIPT, NWS, CDC, NTSB and NHTSA

* most recent available data, from 2006

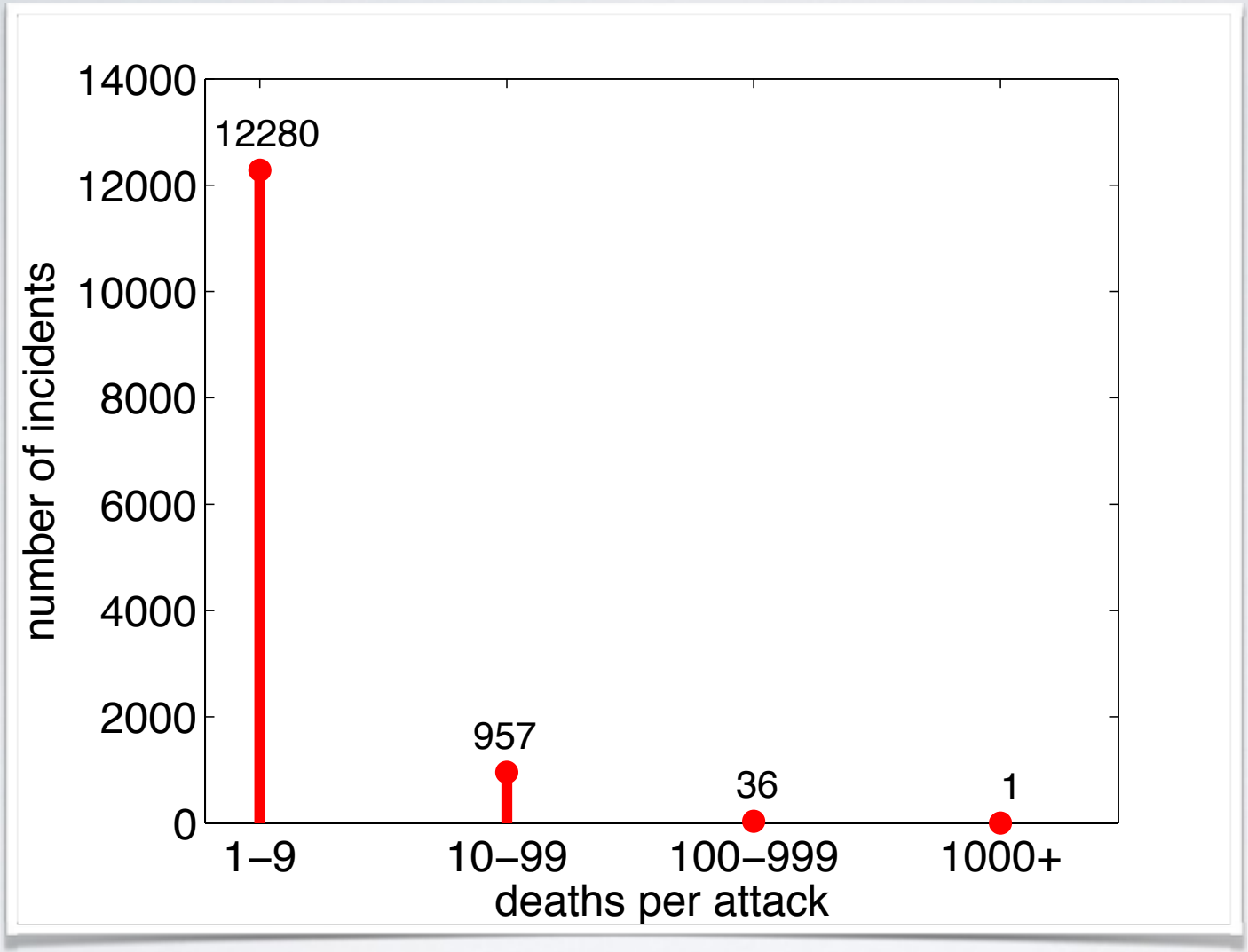
terrorism

- **how many people die in a terrorist event?**

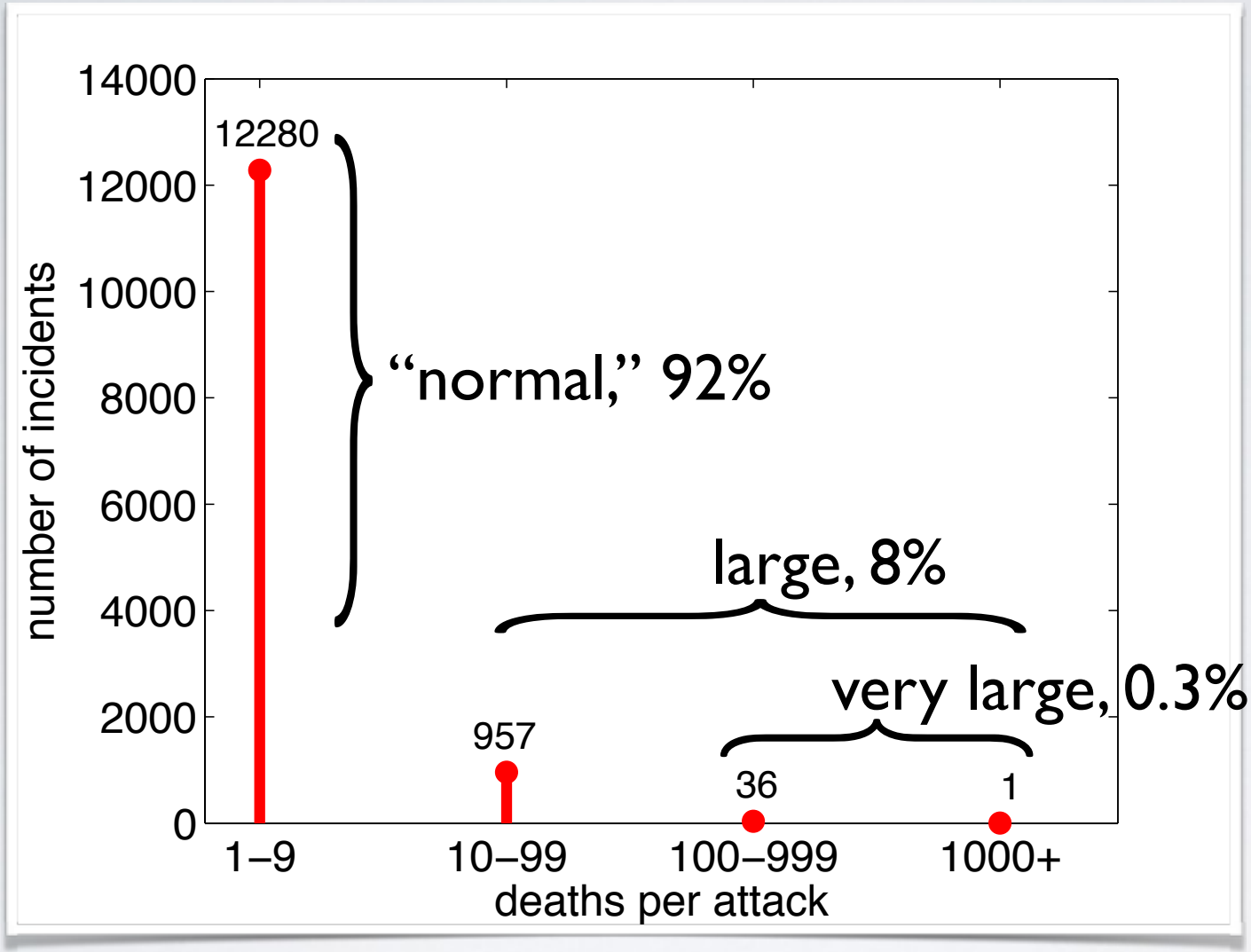
data analysis:

1. *take all events*
2. *count # times 1 death, 2 deaths, etc.*
3. *visualize as distribution*

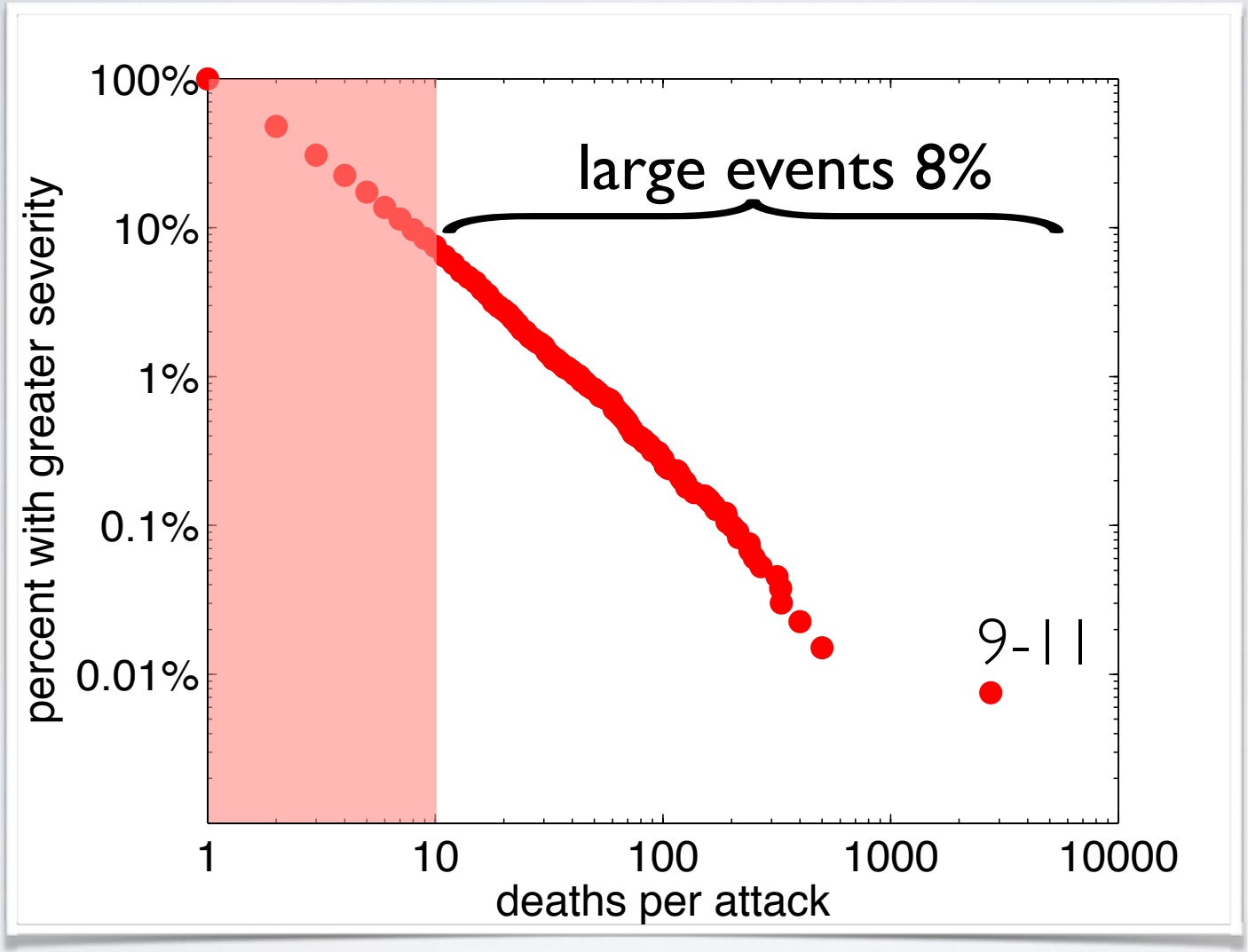
deadly terrorist events, 1968-2008



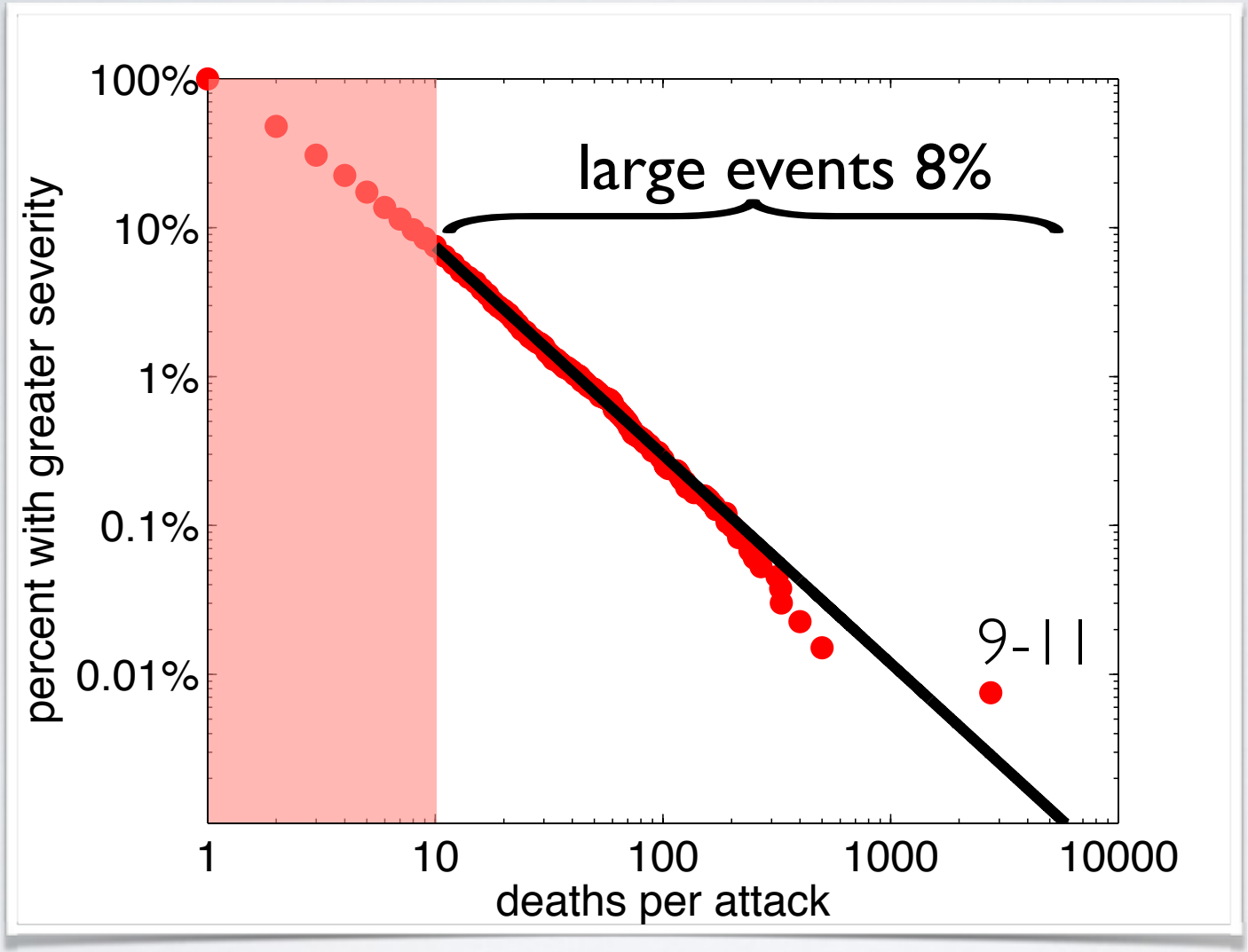
deadly terrorist events, 1968-2008



deadly terrorist events, 1968-2008



deadly terrorist events, 1968-2008



it follows a power-law distribution

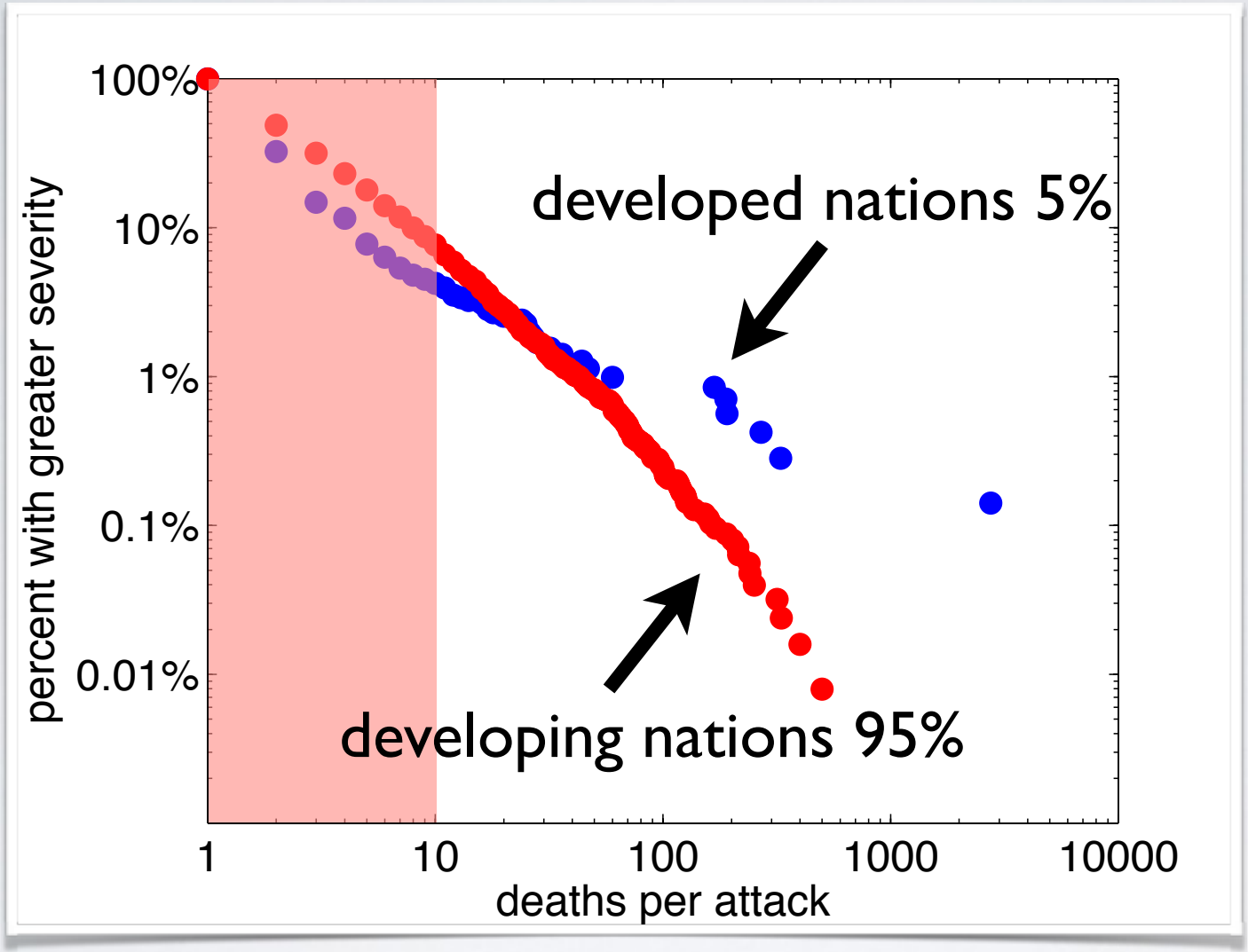
up by 10x in severity = down by 250x in frequency

terrorism

- **do big events happen everywhere equally?**

data analysis:

1. *divide events by (i) developed nation or (ii) developing nation*
2. *for each type, count # times 1 death, 2 deaths, etc.*
3. *visualize the two distributions*


100%
10%
1%
0.1%
0.01%
percent with greater severity
developed nations 5%
developing nations 95%
1
10
100
1000
10000
deaths per attack

terrorism

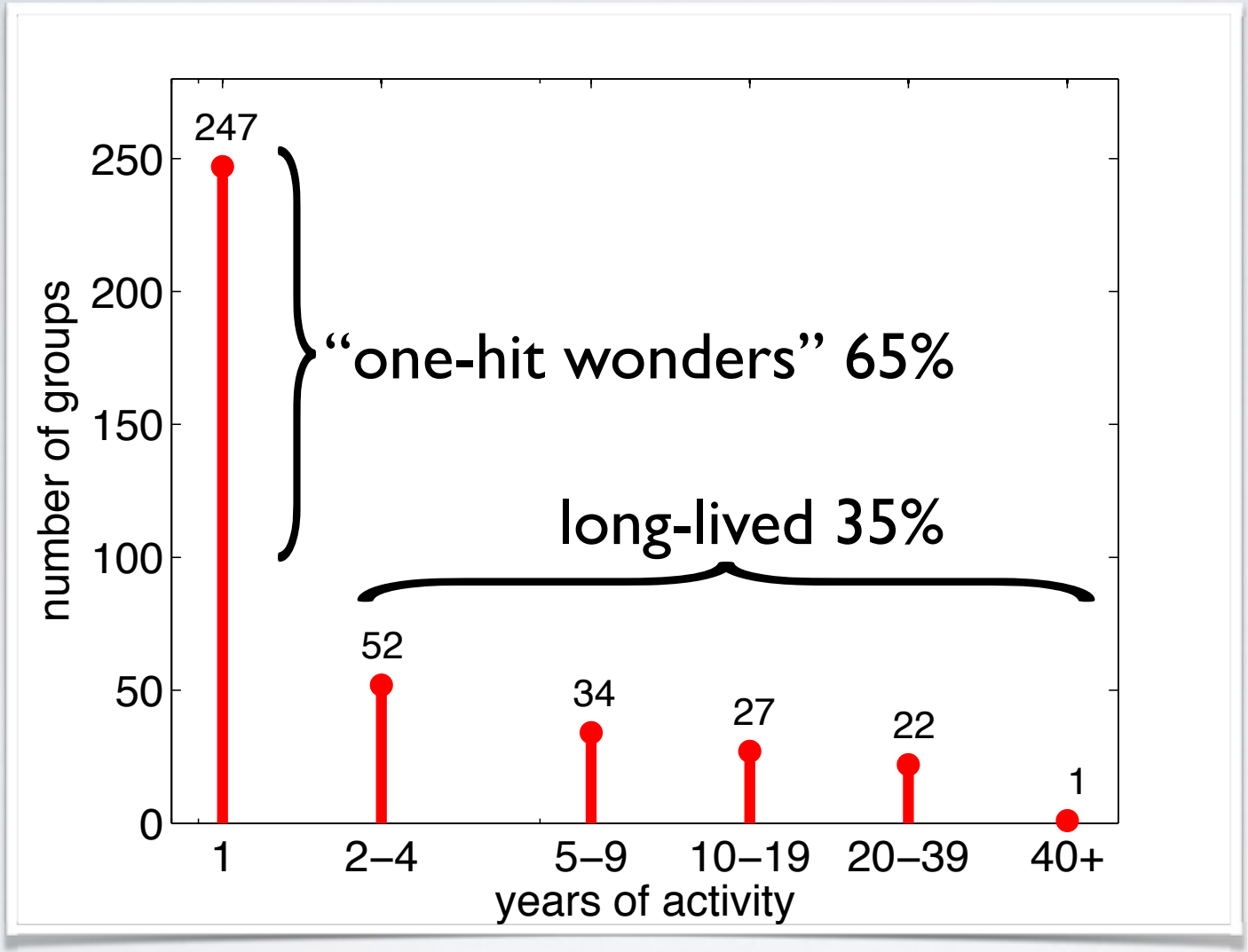
- **how long do terrorist groups last?**

data analysis:

1. *for each unique terrorist group*
2. *count # years between oldest and newest event*
3. *then count # groups with 1 year, 2 years, etc. of activity*
4. *visualize the distribution*

major terrorist organizations 1968-2008

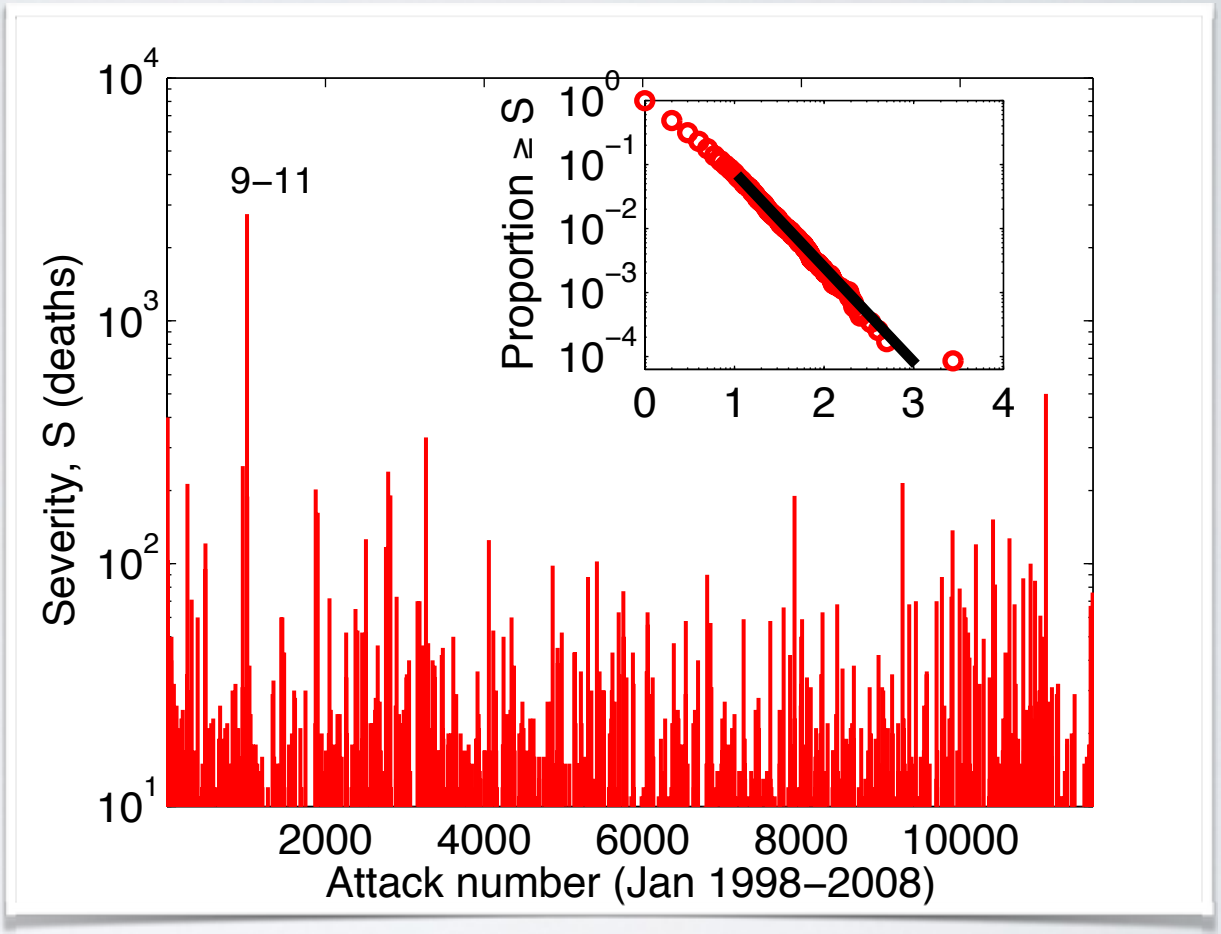
1. Revolutionary Armed Forces of Colombia (FARC)
2. Hamas
3. Taliban
4. Basque Fatherland and Freedom (ETA)
5. Communist Party of Nepal-Maoist (CPN-M)
6. National Liberation Army (Colombia)
7. Palestinian Islamic Jihad (PIJ)
8. Liberation Tigers of Tamil Eelam (LTTE)
9. al-Fatah
10. Communist Party of India-Maoist
11. al-Qaeda Organization in the Land of the Two Rivers
12. Anti-Castro Cubans
13. Hezbollah
14. Fronte di Liberazione Naziunale di a Corsica (FLNC)
15. Shining Path
16. Islamic State of Iraq
17. Popular Front for the Liberation of Palestine (PFLP)
18. United Liberation Front of Assam (ULFA)
19. al-Aqsa Martyrs Brigades
20. Kurdistan Workers' Party (PKK)
21. Tupac Amaru Revolutionary Movement
22. Ansar al-Sunnah Army
23. Black September
24. New People's Army (NPA)
25. Abu Nidal Organization (ANO)
26. Mujahideen Shura Council
27. Armenian Secret Army for the Liberation of Armenia
28. Irish Republican Army (IRA)
29. Revolutionary People's Liberation Party/Front (DHKP/C)
30. People's War Group (PWG)
31. United Self-Defense Forces of Colombia (AUC)
32. Jewish Defense League (JDL)
33. Amal
34. Armed Islamic Group
35. Palestine Liberation Organization (PLO)
36. Earth Liberation Front (ELF)
37. Abu Sayyaf Group (ASG)
38. Popular Resistance Committees
39. Manuel Rodriguez Patriotic Front
40. Revolutionary Organization 17 November (RO-N17)
41. al-Qaeda Organization in the Islamic Maghreb
42. Baloch Liberation Army (BLA)
43. Revolutionary People's Struggle
44. Red Army Faction
45. Islamic Army in Iraq
45. Democratic Front for the Liberation of Palestine (DFLP)
46. UNITA
47. Revolutionary Nuclei
48. al-Gama'a al-Islamiyya (GAI)
49. Free Aceh Movement (GAM)
50. Kurdistan Freedom Hawks
51. April 19 Movement
52. Lord's Resistance Army (LRA)
53. Moro Islamic Liberation Front (MILF)
54. Real Irish Republican Army (RIRA)
55. al-Qaeda
56. Tawhid and Jihad
57. Popular Liberation Army
58. Eritrean Liberation Front (ELF)
59. Montoneros
60. Turkish Communist Party Marxist-Leninist (TKP/ML-TIKKO)
61. Mozambique National Resistance Movement
62. Ulster Defence Association/Ulster Freedom Fighters
63. Purbo Banglar Communist Party (PBCP)
64. National Liberation Front of Tripura (NLFT)
45. First of October Antifascist Resistance Group (GRAPO)
46. Red Hand Defenders (RHD)
47. Lashkar-e-Taiba (LeT)
48. Hizbul Mujahideen (HM)
49. Mujahideen Youth Movement
50. Bersatu
51. People's Revolutionary Army (Argentina)
52. Farabundo Marti National Liberation Front
53. Jaish-e-Mohammad (JeM)
54. Islamic Jihad Jerusalem
55. Peronist Armed Forces
56. Khmer Rouge
57. Justice Commandos for the Armenian Genocide
58. Continuity Irish Republican Army (CIRA)
59. PKK/KONGRA-GEL
60. National Democratic Front of Bodoland (NDFB)
61. Lautaro Youth Movement
62. Action Directe
63. Polisario Front
64. Mujahedin-e-Khalq (MeK)
65. Maoist Communist Center (MCC)
66. Popular Forces of April 25
67. Third of October Group
68. Baader-Meinhof Group
69. Breton Revolutionary Army (ARB)
1. Orly Organization
2. People's Liberation Forces (El Salvador)
3. Front for the Liberation of the Cabinda Enclave
4. Abu al-Rish Brigades
5. African National Congress (South Africa)
6. Moro National Liberation Front (MNLF)
7. Islamic Great Eastern Raiders Front
8. Palestinian Revolution Forces General Command
9. Chukakuha
10. Communist Combatant Cells
11. Popular Front for the Liberation of Palestine -- General Command (PFLP-GC)
12. Red Brigades
13. Japanese Red Army (JRA)
14. Animal Liberation Front (ALF)
15. Committee of Solidarity with Arab and Middle East Political Prisoners (CSPPA)
16. Front for the Liberation of Lebanon from Foreigners (FLLF)
17. Jamatul Mujahedin Bangladesh
18. Informal Anarchist Federation
19. Sudan People's Liberation Army
20. Ninth of June Organization
21. Guerrilla Army of the Poor
22. Loyalist Volunteer Force (LVF)
23. Anti-Imperialist International Brigade
24. All Tripura Tiger Force (ATTF)
25. People's Revolutionary Army (Colombia)
26. Social Resistance
27. Arab Communist Organization (ACO)
28. Anti-Terrorist Liberation Group
29. Riyad us-Saliheyn Martyrs' Brigade
30. Kosovo Liberation Army (KLA)
31. Lashkar-e-Jhangvi (LeJ)
32. Revolutionary United Front (RUF)
33. Jamiat ul-Mujahedin (JuM)
34. Alex Boncayao Brigade (ABB)
35. Pattani United Liberation Organization (PULO)
36. Group Bakunin Gdansk Paris Guatemala Salvador
37. Irish National Liberation Army (INLA)
38. Revolutionary Struggle
39. Lebanese Armed Revolutionary Faction
40. Ananda Marga
41. Tupamaros
42. ....


247
250
200
150
100
50
0
number of groups
"one-hit wonders" 65%
long-lived 35%
52
34
27
22
1
1
2–4
5–9
10–19
20–39
40+
years of activity

terrorism

terrorism

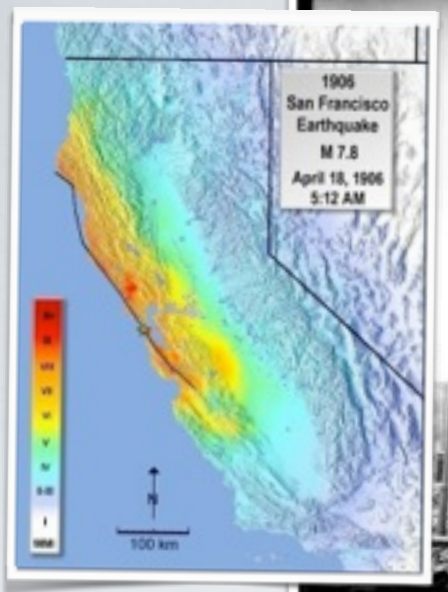
- terrorism occurs mostly in global "hot spots"
- deaths follow a power law
- big events often in developed countries
- most terrorist groups short-lived



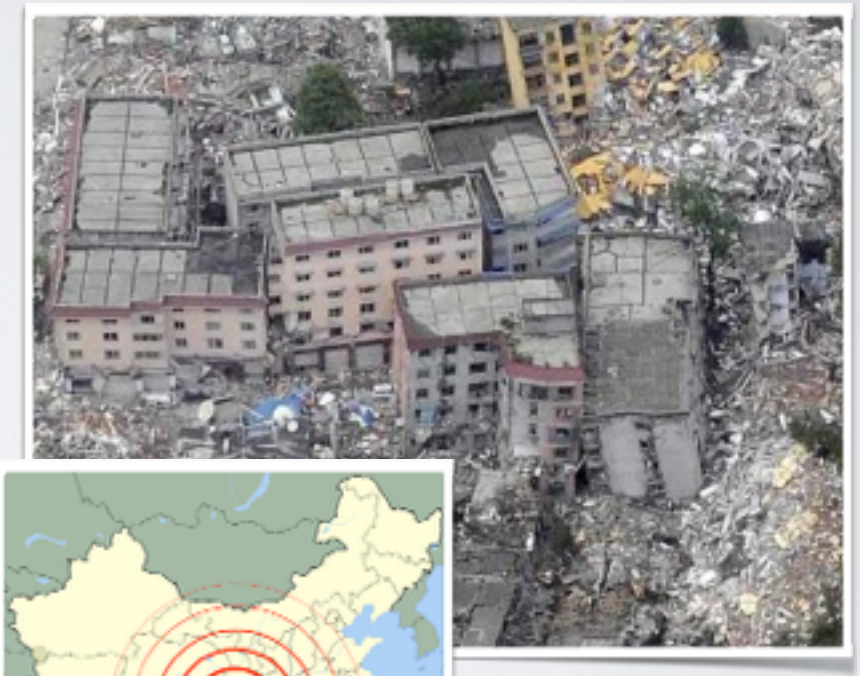
$$(\text{frequency}) \propto (\text{deaths})^{-\alpha}$$

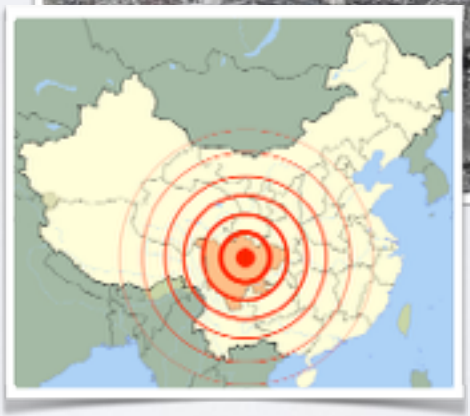
what else follows power laws?

1906 San Francisco, M7.8



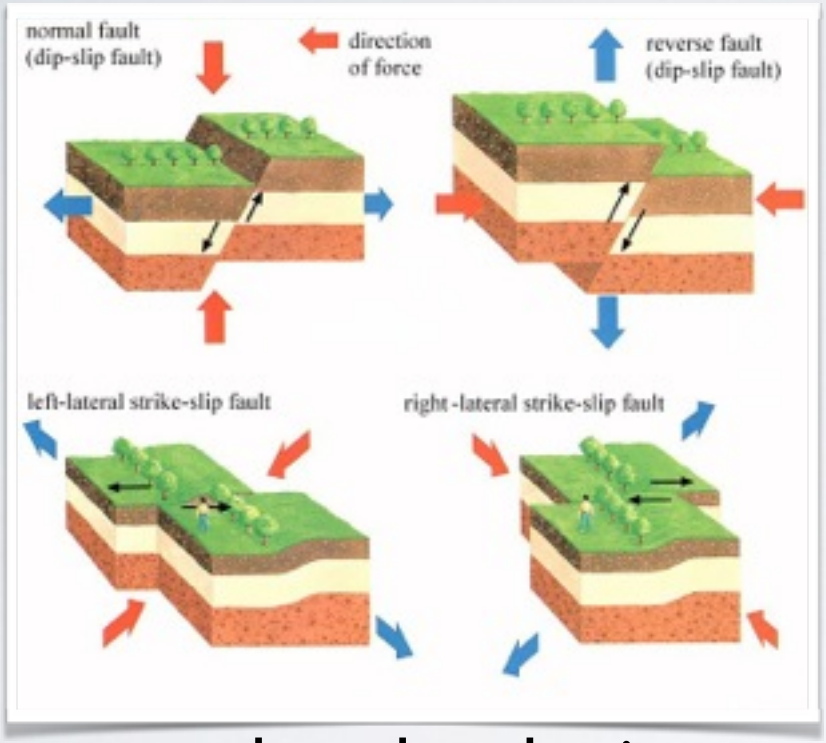
2008 Sichuan, M7.9





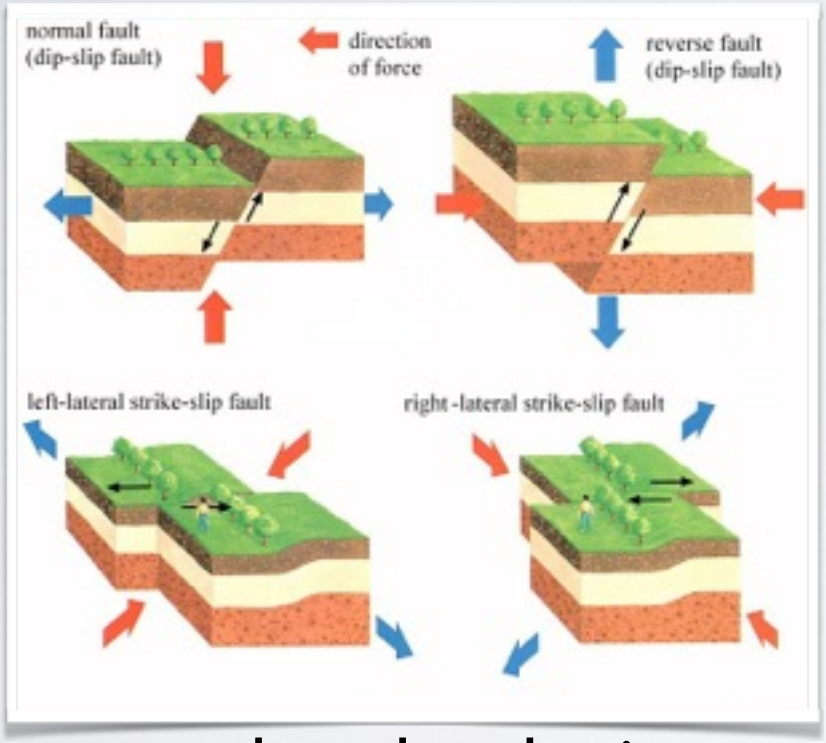
2011 Japan, M8.9

earthquakes

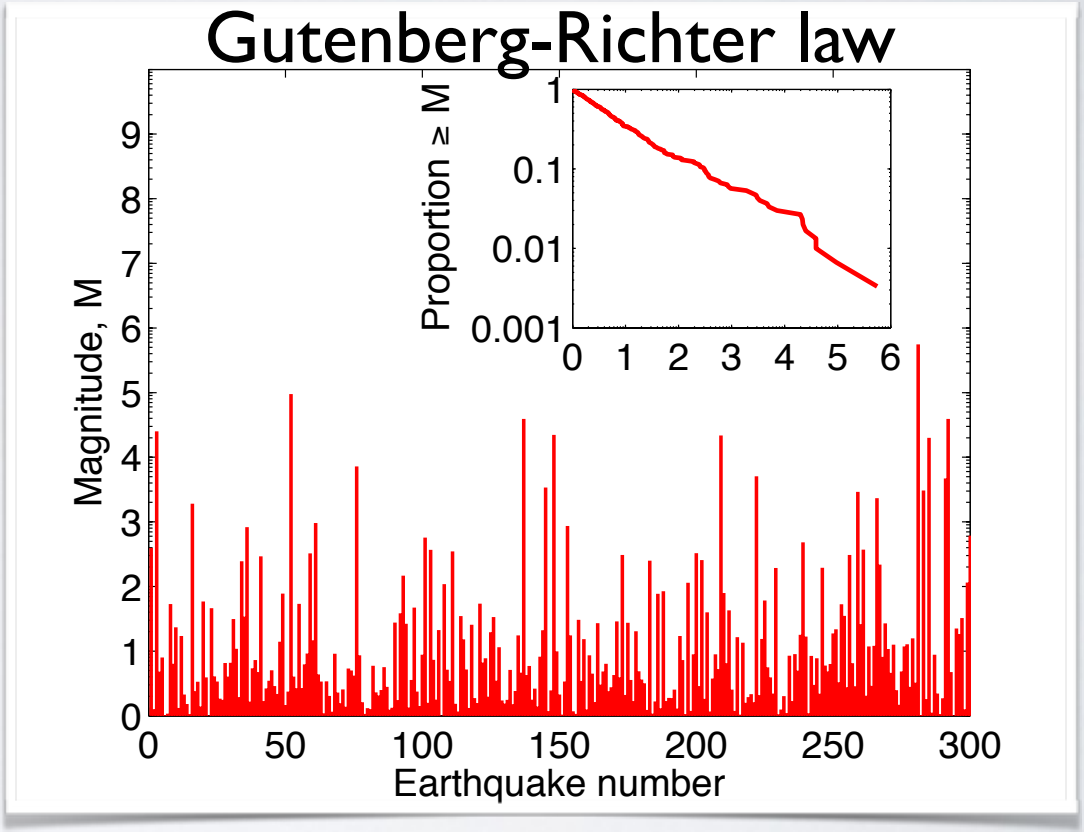


earthquake physics

earthquakes



earthquake physics



$$(\text{frequency}) \propto (\text{seismic moment})^{-\alpha}$$

earthquakes	**terrorism**
Gutenberg-Richter law $F \propto M^{-\alpha}$	Richardson's law $F \propto S^{-\alpha}$
physics largely *known*	processes largely *unknown*
processes *fixed*	processes *dynamic, adaptive*
forecasting possible (years of successes)	how do we forecast?
prediction very hard (years of failures)	what can we predict? what can we not predict?

data, data, data

data, data, data

some of my projects published in 2013

- scoring dynamics in professional team sports
- scoring dynamics in the video game *Halo*
- identifying patterns in malaria gene networks
- detecting friendships in online social networks
- body size evolution of horses over the past 55 million years
- social networks in c1400 American Southwest
- forecasting large events in terrorism
- how large should whales be?
- ...

data, data, data

Scoring dynamics across professional team sports: tempo, balance and predictability

Sears Merritt[1,*] and Aaron Clauset[1,2,3,†]

[1] *Department of Computer Science, University of Colorado, Boulder, CO 80309*
[2] *BioFrontiers Institute, University of Colorado, Boulder, CO 80303*
[3] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501*

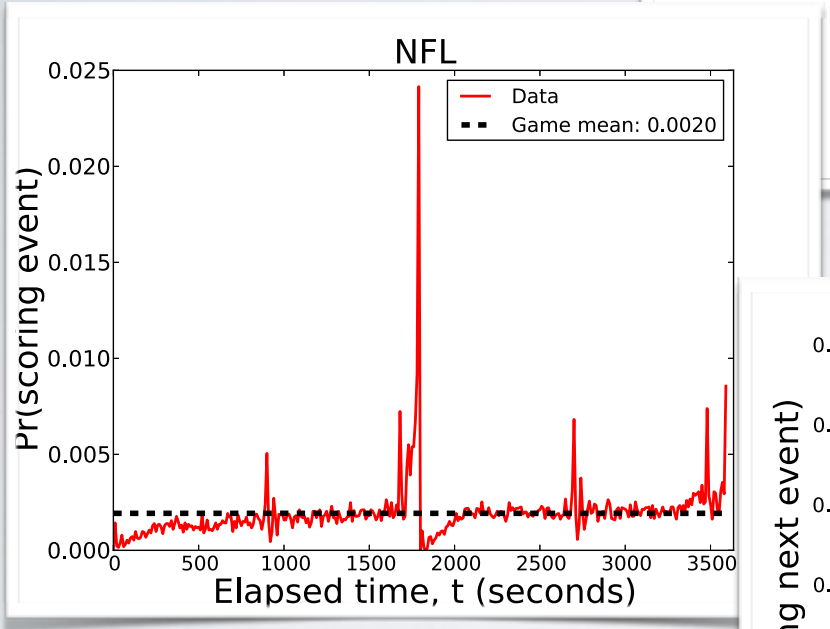
sport	abbrv.	seasons	teams	competitions	scoring events
Football (college)	CFB	10, 2000–2009	486	14,588	120,827
Football (pro)	NFL	10, 2000–2009	31	2,654	19,476
Hockey (pro)	NHL	10, 2000–2009	29	11,813	44,989
Basketball (pro)	NBA	9, 2002–2010	31	11,744	1,080,285

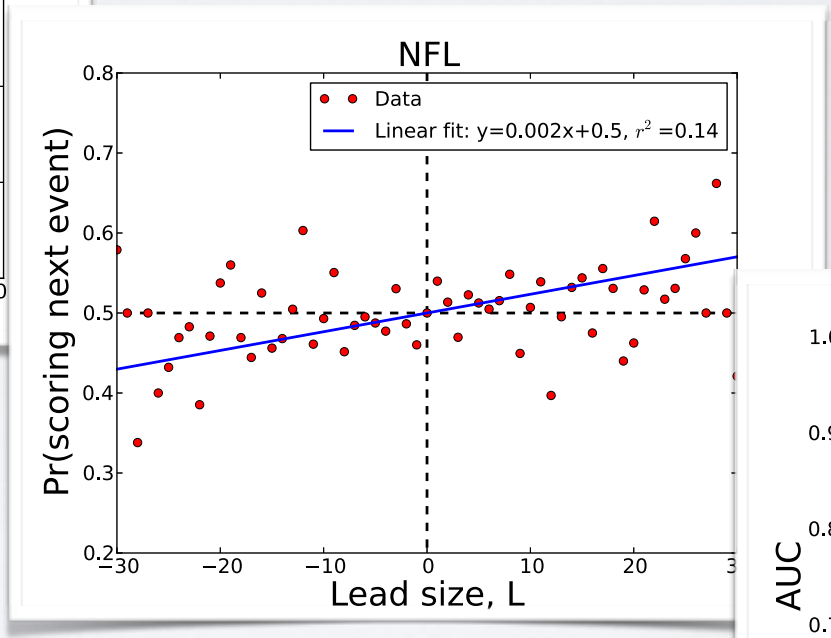
data, data, data

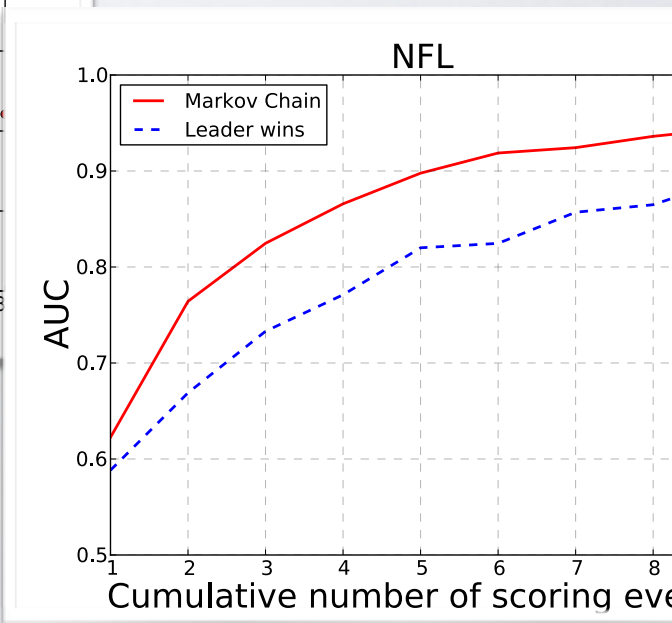
Scoring dynamics across professional team sports: tempo, balance and predictability

Sears Merritt[1,*] and Aaron Clauset[1,2,3,†]

[1] *Department of Computer Science, University of Colorado, Boulder, CO 80309*
[2] *BioFrontiers Institute, University of Colorado, Boulder, CO 80303*
[3] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501*







fin

more

follow me on twitter: @aaronclauset

read my blog : www.structureandstrangeness.com

read my papers : www.santafe.edu/~aaronc/