Learning from Data

Aaron Clauset
@aaronclauset

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

© 2013 Aaron Clauset
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

does not exist

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

terrorism
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, USA 2007

<table>
<thead>
<tr>
<th>Event</th>
<th>Rate (per million)</th>
<th>Multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>terrorism</td>
<td>0.00</td>
<td>--</td>
</tr>
<tr>
<td>lightning</td>
<td>0.15</td>
<td>1</td>
</tr>
<tr>
<td>bee sting</td>
<td>0.18</td>
<td>×1.2</td>
</tr>
<tr>
<td>airplane crash</td>
<td>0.23*</td>
<td>×1.5</td>
</tr>
<tr>
<td>homicide</td>
<td>61.74*</td>
<td>×408.3</td>
</tr>
<tr>
<td>car crash</td>
<td>124.43</td>
<td>×829.5</td>
</tr>
</tbody>
</table>

* Most recent available data, from 2006.*

Sources: US Census, MiPT, NWS, CDC, NTSB and NHTSA.
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

Number of incidents vs. deaths per attack.

- 12280 incidents with 1-9 deaths per attack.
- 957 incidents with 10-99 deaths per attack.
- 36 incidents with 100-999 deaths per attack.
- 1 incident with 1000+ deaths per attack.
deadly terrorist events, 1968-2008

- "normal," 92%
- large, 8%
- very large, 0.3%

number of incidents

12280
957
36
1

1-9
10-99
100-999
1000+

deaths per attack
deadly terrorist events, 1968-2008

large events 8%
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
percent with greater severity

developed nations 5%
developing nations 95%
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
<table>
<thead>
<tr>
<th>No.</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Orly Organization</td>
</tr>
<tr>
<td>2.</td>
<td>People's Liberation Forces (El Salvador)</td>
</tr>
<tr>
<td>3.</td>
<td>Front for the Liberation of the Cabinda Enclave</td>
</tr>
<tr>
<td>4.</td>
<td>Abu al-Rish Brigades</td>
</tr>
<tr>
<td>5.</td>
<td>African National Congress (South Africa)</td>
</tr>
<tr>
<td>6.</td>
<td>Moro National Liberation Front (MNLF)</td>
</tr>
<tr>
<td>7.</td>
<td>Islamic Great Eastern Raiders Front</td>
</tr>
<tr>
<td>8.</td>
<td>Palestinian Revolution Forces General Command</td>
</tr>
<tr>
<td>9.</td>
<td>Chukakuda</td>
</tr>
<tr>
<td>10.</td>
<td>Communist Combatant Cells</td>
</tr>
<tr>
<td>11.</td>
<td>Popular Front for the Liberation of Palestine -- General Command (PFLP-GC)</td>
</tr>
<tr>
<td>12.</td>
<td>Red Brigades</td>
</tr>
<tr>
<td>13.</td>
<td>Japanese Red Army (JRA)</td>
</tr>
<tr>
<td>14.</td>
<td>Animal Liberation Front (ALF)</td>
</tr>
<tr>
<td>15.</td>
<td>Committee of Solidarity with Arab and Middle East Political Prisoners (CSPPA)</td>
</tr>
<tr>
<td>16.</td>
<td>Front for the Liberation of Lebanon from Foreigners (FLLF)</td>
</tr>
<tr>
<td>17.</td>
<td>Jamatul Mujahedin Bangladesh</td>
</tr>
<tr>
<td>18.</td>
<td>Informal Anarchist Federation</td>
</tr>
<tr>
<td>19.</td>
<td>Sudan People's Liberation Army</td>
</tr>
<tr>
<td>20.</td>
<td>Ninth of June Organization</td>
</tr>
<tr>
<td>21.</td>
<td>Guerrilla Army of the Poor</td>
</tr>
<tr>
<td>22.</td>
<td>Loyalist Volunteer Force (LVF)</td>
</tr>
<tr>
<td>23.</td>
<td>Anti-Imperialist International Brigade</td>
</tr>
<tr>
<td>24.</td>
<td>All Tripura Tiger Force (ATTF)</td>
</tr>
<tr>
<td>25.</td>
<td>People's Revolutionary Army (Colombia)</td>
</tr>
<tr>
<td>26.</td>
<td>Social Resistance</td>
</tr>
<tr>
<td>27.</td>
<td>Arab Communist Organization (ACO)</td>
</tr>
<tr>
<td>28.</td>
<td>Anti-Terrorist Liberation Group</td>
</tr>
<tr>
<td>29.</td>
<td>Riyadh us-Salihayn Martyrs' Brigade</td>
</tr>
<tr>
<td>30.</td>
<td>Kosovo Liberation Army (KLA)</td>
</tr>
<tr>
<td>31.</td>
<td>Lashkar-e-Jhangvi (LeJ)</td>
</tr>
<tr>
<td>32.</td>
<td>Revolutionary United Front (RUF)</td>
</tr>
<tr>
<td>33.</td>
<td>Jamiat ul-Mujahedin (JuM)</td>
</tr>
<tr>
<td>34.</td>
<td>Alex Boncayao Brigade (ABB)</td>
</tr>
<tr>
<td>35.</td>
<td>Pattani United Liberation Organization (PULO)</td>
</tr>
<tr>
<td>36.</td>
<td>Group Bakunin Gdansk Paris Guatemala Salvador</td>
</tr>
<tr>
<td>37.</td>
<td>Irish National Liberation Army (INLA)</td>
</tr>
<tr>
<td>38.</td>
<td>Revolutionary Struggle</td>
</tr>
<tr>
<td>39.</td>
<td>Lebanese Armed Revolutionary Faction</td>
</tr>
<tr>
<td>40.</td>
<td>Ananda Marga</td>
</tr>
<tr>
<td>41.</td>
<td>Tupamaros</td>
</tr>
</tbody>
</table>

**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. Front di Liberazione Nazionale 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. Poplar 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
46. Democratic Front for the Liberation of Palestine (DFLP)
47. UNITA
48. Revolutionary Nuclei
49. al-Gama'a al-Islamiyya (GAI)
50. Free Aceh Movement (GAM)
51. Kurdistan Freedom Hawks
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)
65. First of October Antifascist Resistance Group (GRAPO)
66. Red Hand Defenders (RHD)
67. Lashkar-e-Taiba (LeT)
68. Hizbul Mujahideen (HM)
69. Mujahideen Youth Movement
70. Bersatu
71. People's Revolutionary Army (Argentina)
72. Farabundo Marti National Liberation Front
73. Jaish-e-Mohammad (JeM)
74. Islamic Jihad Jerusalem
75. Peronist Armed Forces
76. Khormer Rouge
77. Justice Commandos for the Armenian Genocide
78. Continuity Irish Republican Army (CIRA)
79. PKK/KONGRA-GEL
80. National Democratic Front of Bodoland (NDFB)
81. Lautaro Youth Movement
82. Action Directe
83. Polisario Front
84. Mujahedin-e-Khalq (MeK)
85. Maoist Communist Center (MCC)
86. Popular Forces of April 25
87. Third of October Group
88. Baader-Meinhof Group
89. Breton Revolutionary Army (ARB)
"one-hit wonders" 65%

long-lived 35%
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

Gutenberg-Richter law

(frequency) $\propto$ (seismic moment)$^{-\alpha}$
<table>
<thead>
<tr>
<th><strong>earthquakes</strong></th>
<th><strong>terrorism</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gutenberg-Richter law ( F \propto M^{-\alpha} )</td>
<td>Richardson’s law ( F \propto S^{-\alpha} )</td>
</tr>
<tr>
<td>physics largely <em>known</em></td>
<td>processes largely <em>unknown</em></td>
</tr>
<tr>
<td>processes <em>fixed</em></td>
<td>processes <em>dynamic, adaptive</em></td>
</tr>
<tr>
<td>forecasting possible (years of successes)</td>
<td>how do we forecast?</td>
</tr>
<tr>
<td>prediction very hard (years of failures)</td>
<td>what can we predict?</td>
</tr>
<tr>
<td></td>
<td>what can we not predict?</td>
</tr>
</tbody>
</table>
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 c.1400 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\textsuperscript{1,}\textsuperscript{*} and Aaron Clauset\textsuperscript{1,2,3,}\textsuperscript{†}

\textsuperscript{1} Department of Computer Science, University of Colorado, Boulder, CO 80309
\textsuperscript{2} BioFrontiers Institute, University of Colorado, Boulder, CO 80303
\textsuperscript{3} Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501

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

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

Sears Merritt\textsuperscript{1,*} and Aaron Clauset\textsuperscript{1,2,3,†}

\begin{itemize}
  \item Department of Computer Science, University of Colorado, Boulder, CO 80309
  \item BioFrontiers Institute, University of Colorado, Boulder, CO 80303
  \item Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501
\end{itemize}
more

follow me on twitter: @aaronclauset
read my blog : www.structureandstrangeness.com
read my papers : www.santafe.edu/~aaronc/