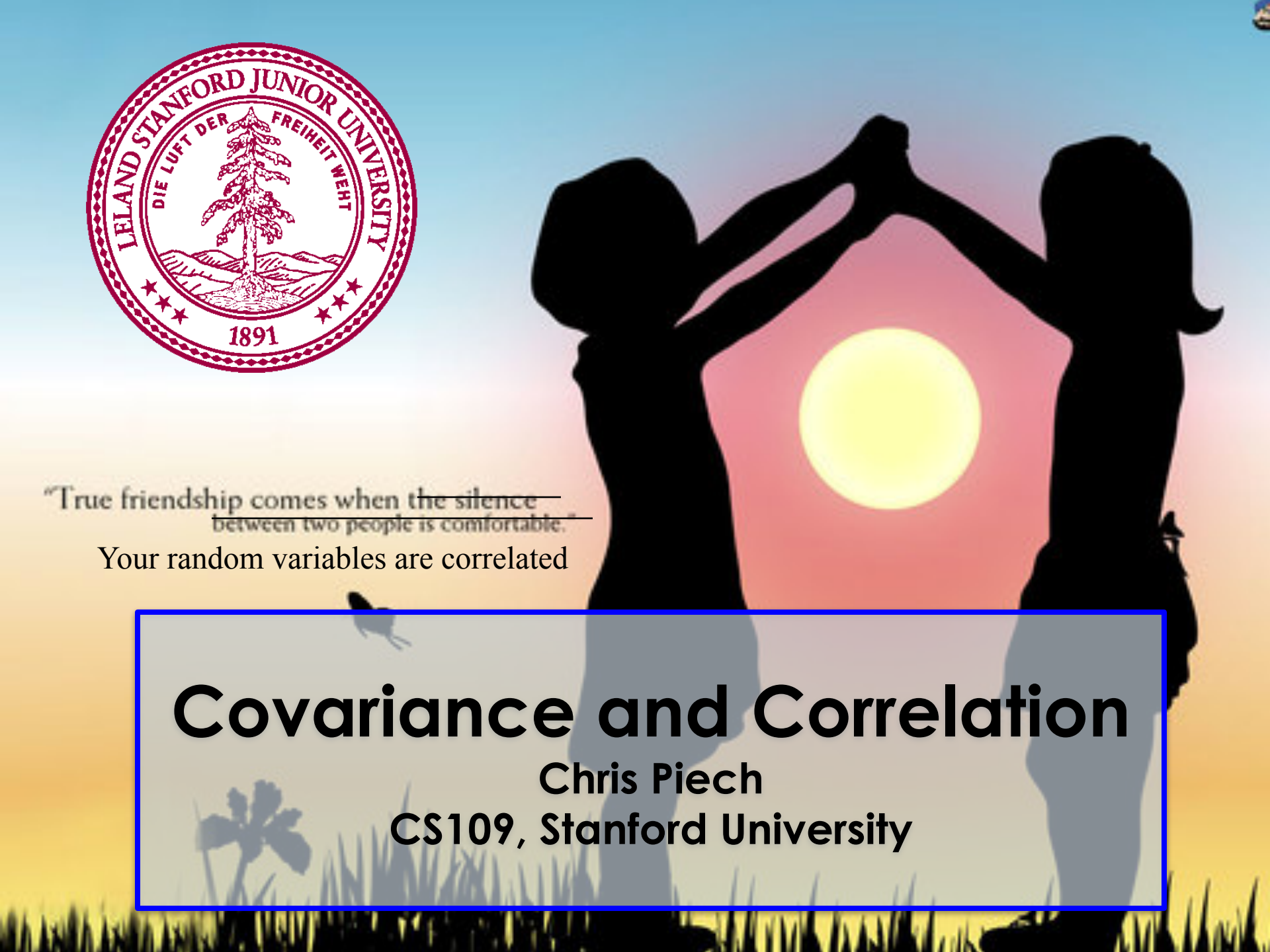


Covariance and Correlation

Chris Piech
CS109, Stanford University

Revisión

Expected Values of Functions

$$E[g(X)] = \sum_x g(x)p(x)$$

$$E[g(X,Y)] = \sum_{x,y} g(x,y)p(x,y)$$

For example: X, Y are **independent** random variables:

$$E[X \cdot Y] = E[g(X,Y)] \qquad \text{Let } g(X,Y) = X \cdot Y$$

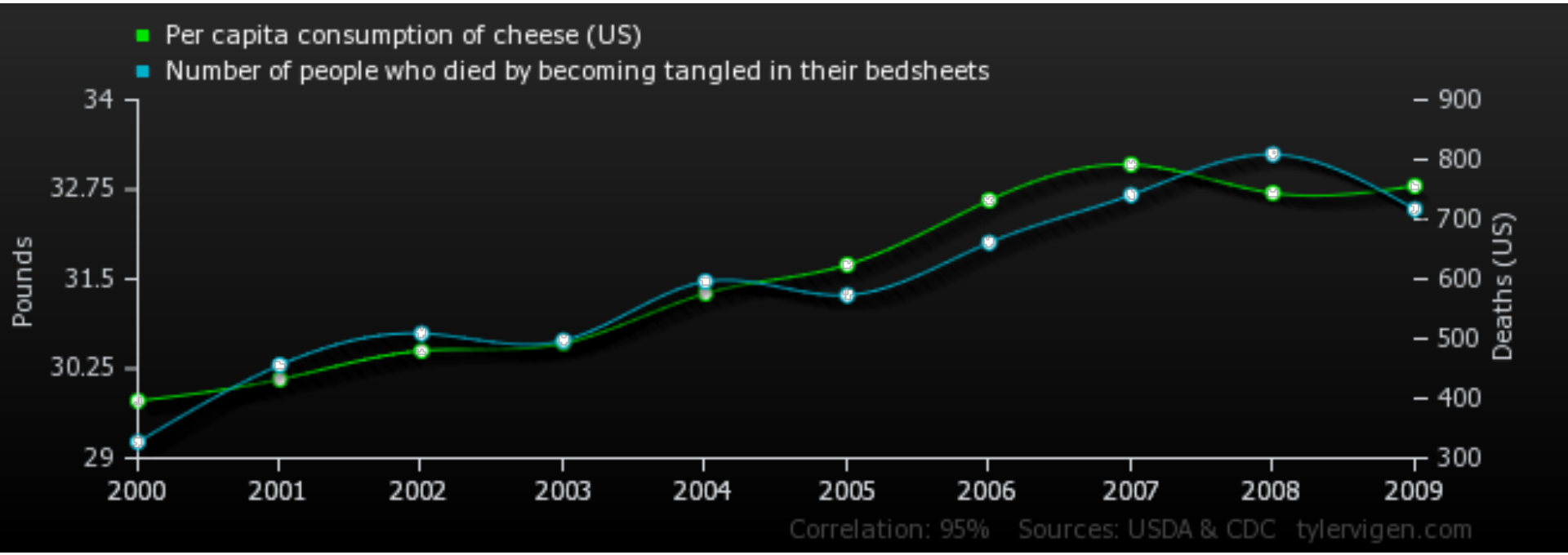
$$= \sum_{x,y} g(x,y) \cdot p(x,y)$$

$$= \sum_{x,y} xy \cdot p(x)p(y)$$

$$= \sum_x xp(x) \cdot \sum_y yp(y) = E[X] \cdot E[Y]$$

Fin de la revisión

Tell your friends!



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) *Pounds (USDA)*	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717

Correlation: 0.947091

Recall our Ebola Bats

Bat Data

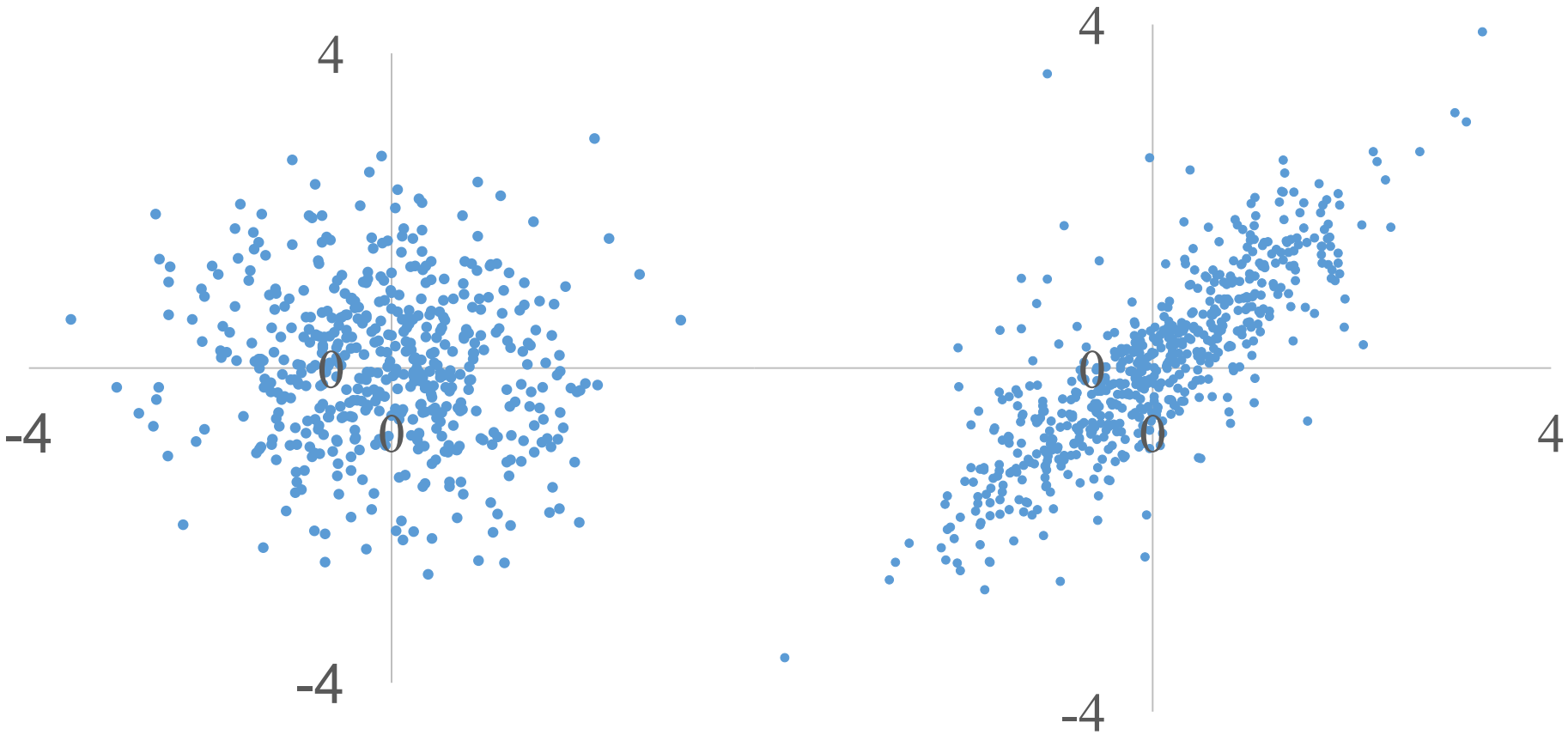
Gene1	Gene2	Gene3	Gene4	Gene5	Trait
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
			...		
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE

Expression Amount

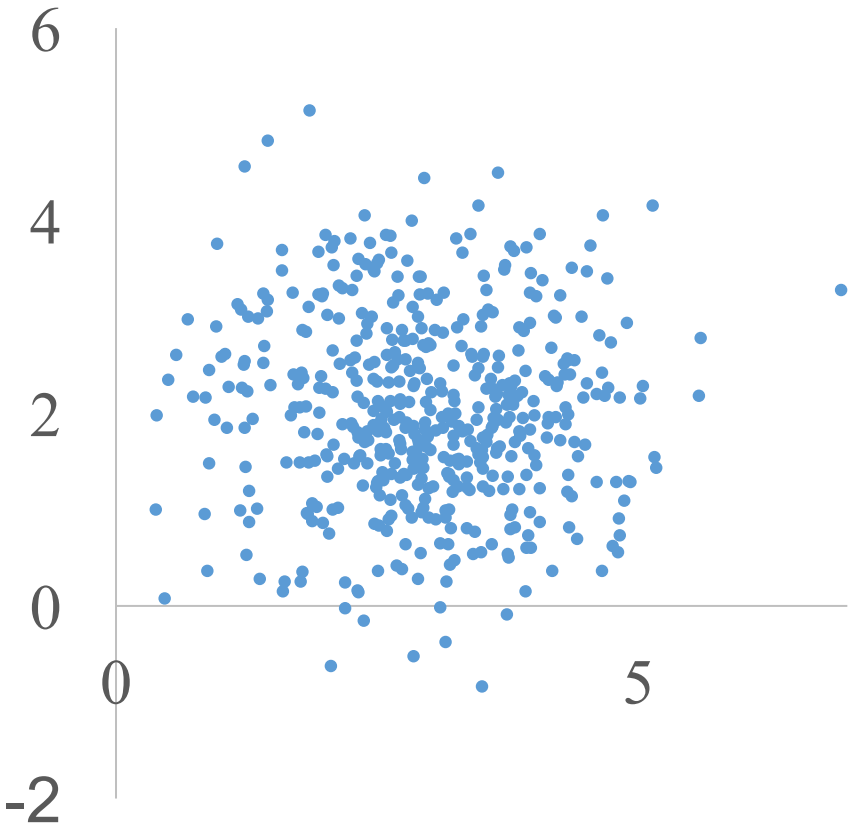
Gene1	Gene2	Gene3	Gene4	Gene5	Trait
0.71	0.29	0.89	0.82	0.76	0.83
0.17	0.02	0.89	0.02	0.94	0.85
0.01	0.63	0.76	0.38	0.82	0.03
0.19	0.95	0.63	0.89	0.94	0.32
0.46	0.96	0.36	0.12	0.50	0.10
0.48	0.51	0.45	0.16	0.40	0.53
0.20	0.77	0.27	0.23	0.90	0.67
0.49	0.24	0.77	0.37	0.29	0.71
0.59	0.95	0.38	0.42	0.72	0.25
0.43	0.66	0.57	0.03	0.15	0.24
0.32	0.42	0.25	0.12	0.79	0.98
0.77	0.31	0.66	0.78	0.68	0.77
0.46	0.59	0.38	0.99	0.71	0.37
0.97	0.66	0.05	0.99	0.36	0.18
0.50	0.66	0.35	0.41	0.62	0.08
0.70	0.85	0.98	0.29	0.59	0.38
			...		
0.78	0.09	0.69	0.41	0.82	0.76

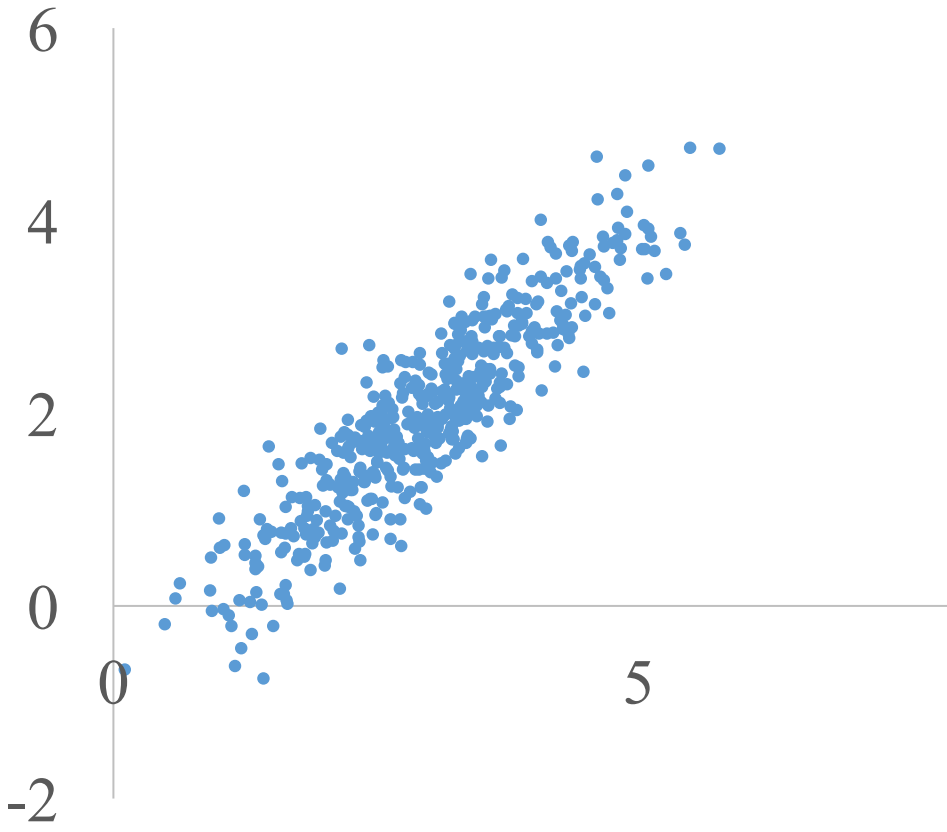
La baila de la Covariance

Spot The Difference

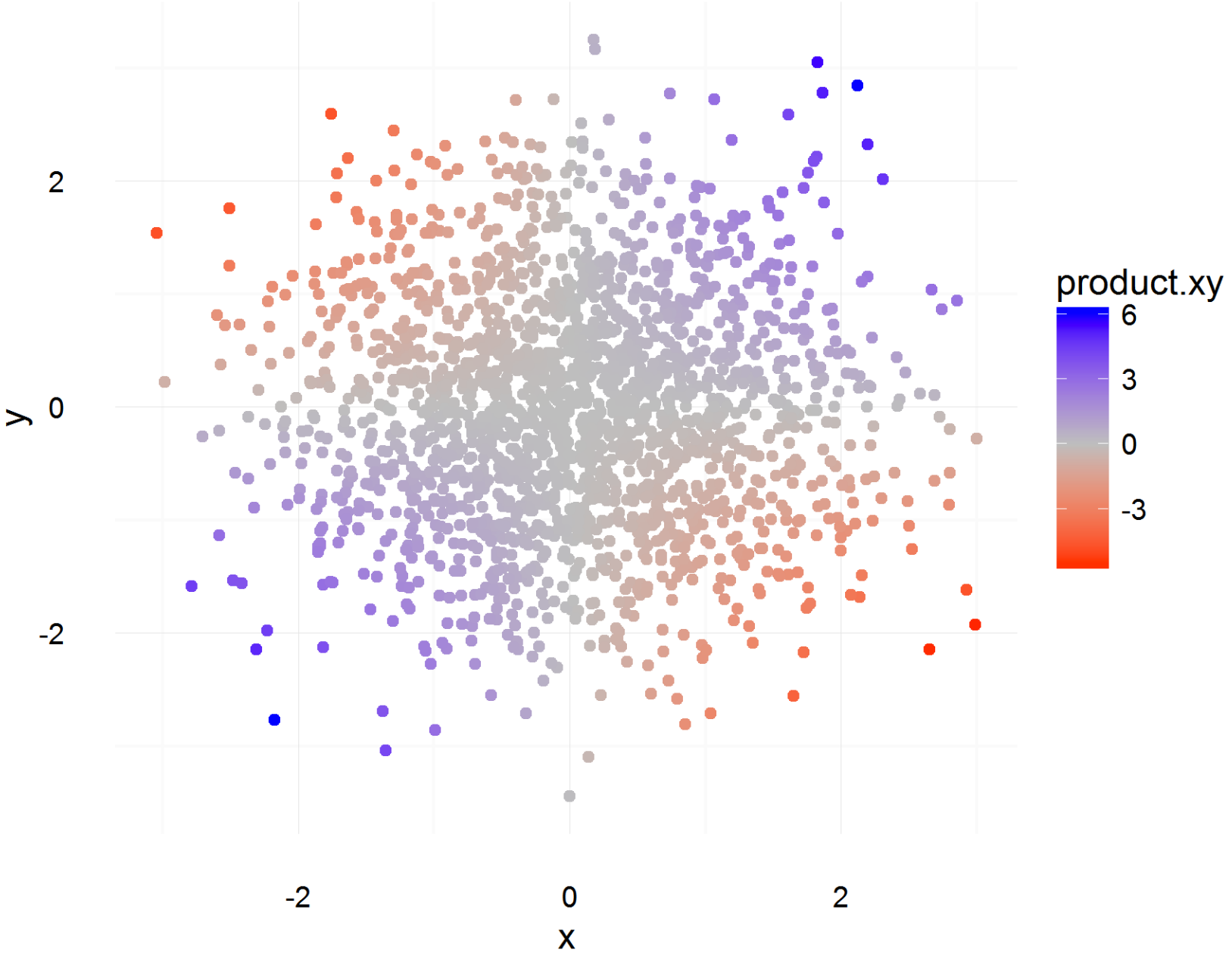


Spot The Difference

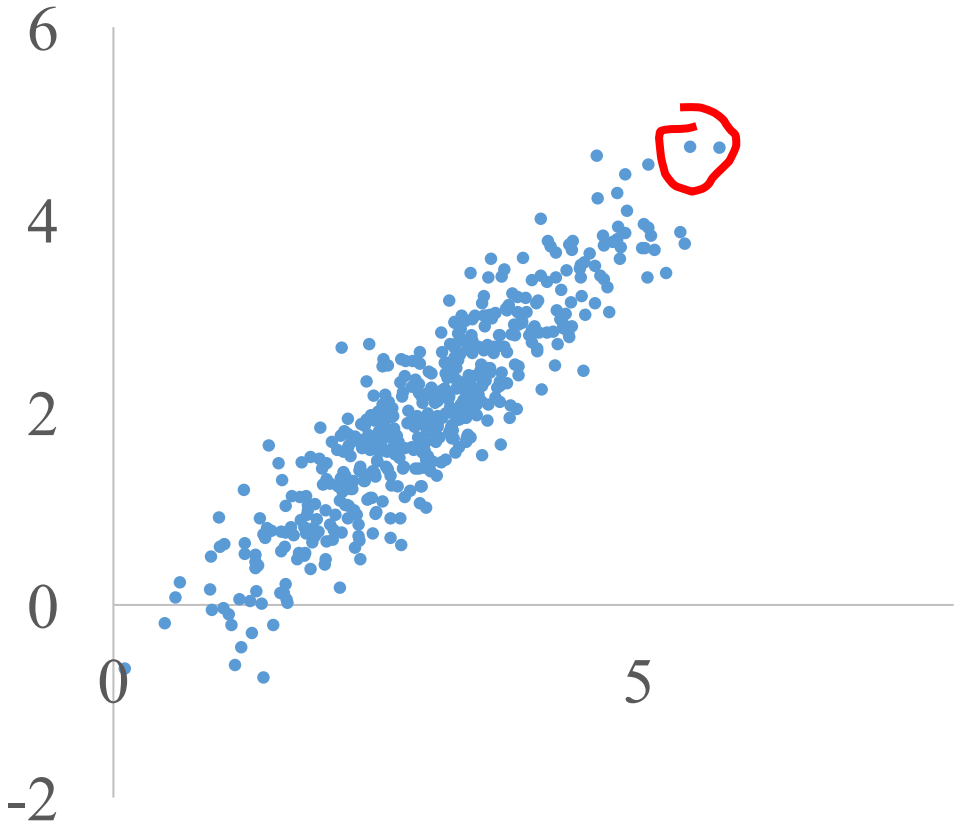




Understanding Covariance



Vary Together

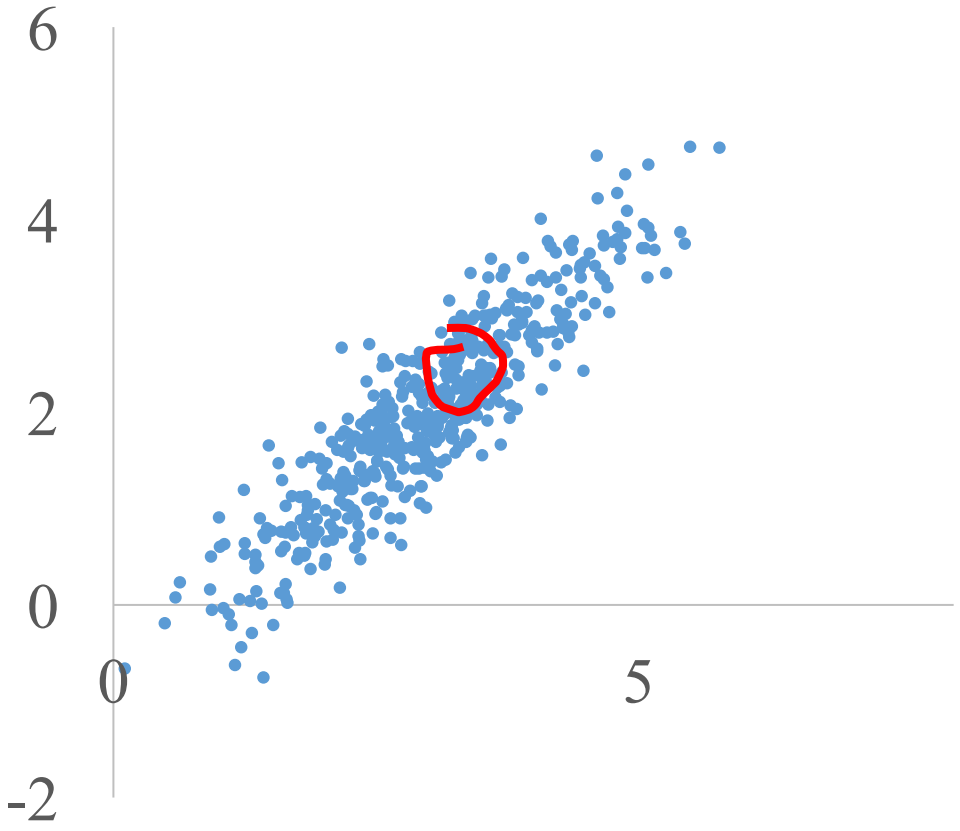


$$x - E[x] = 3$$

$$y - E[y] = 2.6$$

$$(x - E[x])(y - E[y]) = 7.8$$

Vary Together

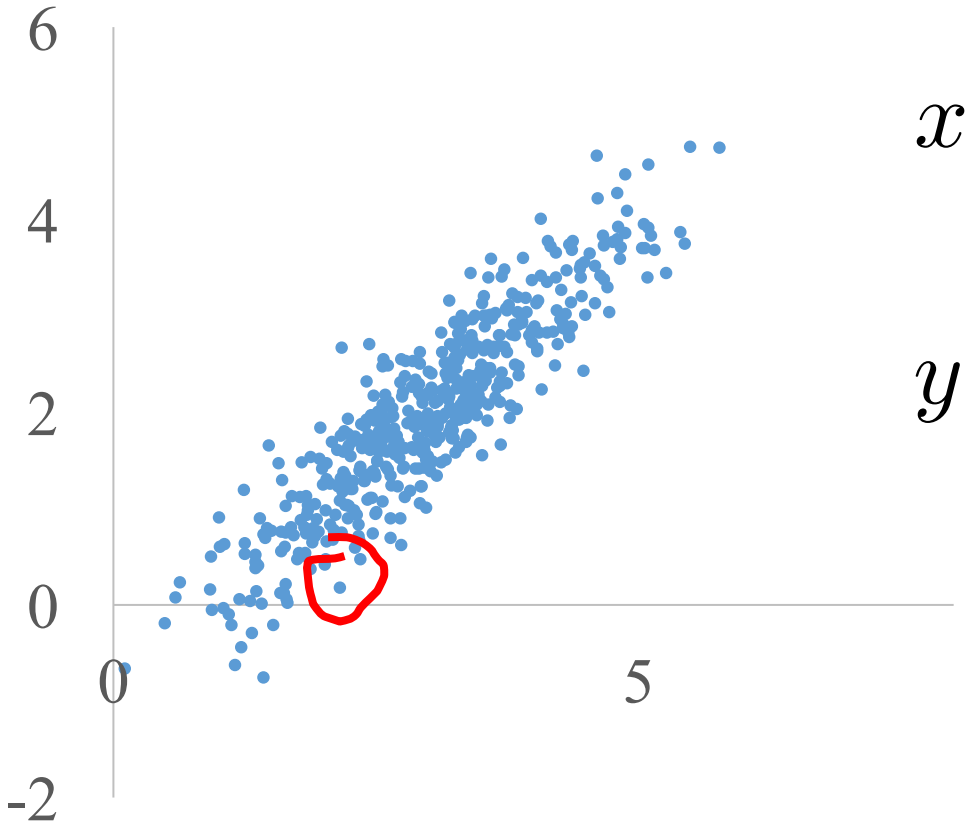


$$x - E[x] \approx 0$$

$$y - E[y] \approx 0$$

$$(x - E[x])(y - E[y]) = 0$$

Vary Together



$$x - E[x] = -1.1$$

$$y - E[y] = -2.8$$

$$(x - E[x])(y - E[y]) \approx 3.1$$

The Dance of the Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y:

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

x	y	$(x - E[X])(y - E[Y])p(x,y)$
Above mean	**Above mean**	**Positive**
Bellow mean	**Bellow mean**	**Positive**
Bellow mean	**Above mean**	**Negative**
Above mean	**Bellow mean**	**Negative**

The Dance of the Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y:

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

- Equivalently:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[XY - E[X]Y - XE[Y] + E[Y]E[X]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

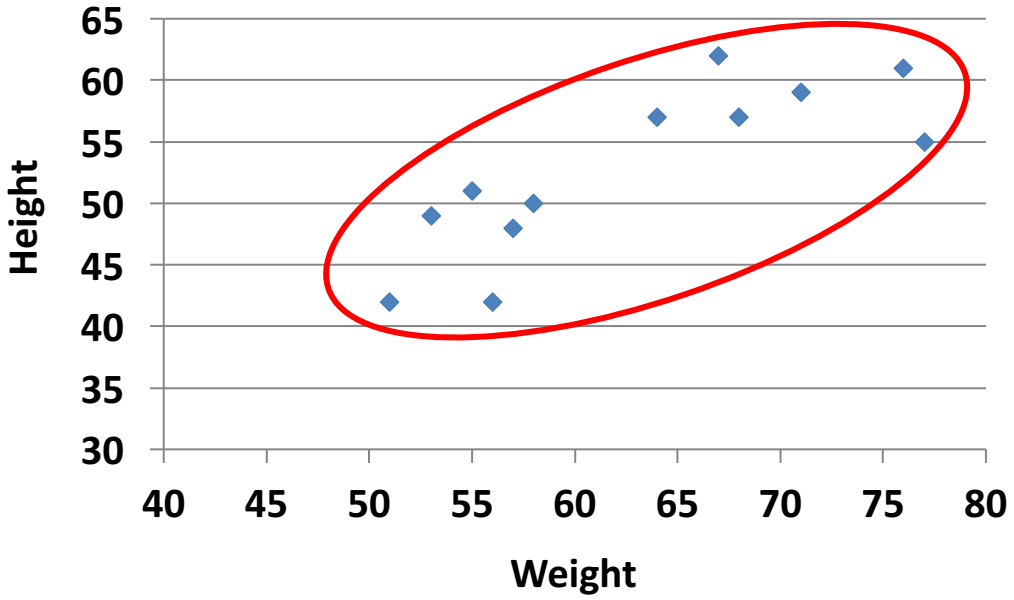
  - X and Y independent, E[XY] = E[X]E[Y] → Cov(X,Y) = 0
  - But Cov(X,Y) = 0 does **not** imply X and Y independent!

Covariance and Data

- Consider the following data:

Weight	Height	Weight * Height
64	57	3648
71	59	4189
53	49	2597
67	62	4154
55	51	2805
58	50	2900
77	55	4235
57	48	2736
56	42	2352
51	42	2142
76	61	4636
68	57	3876

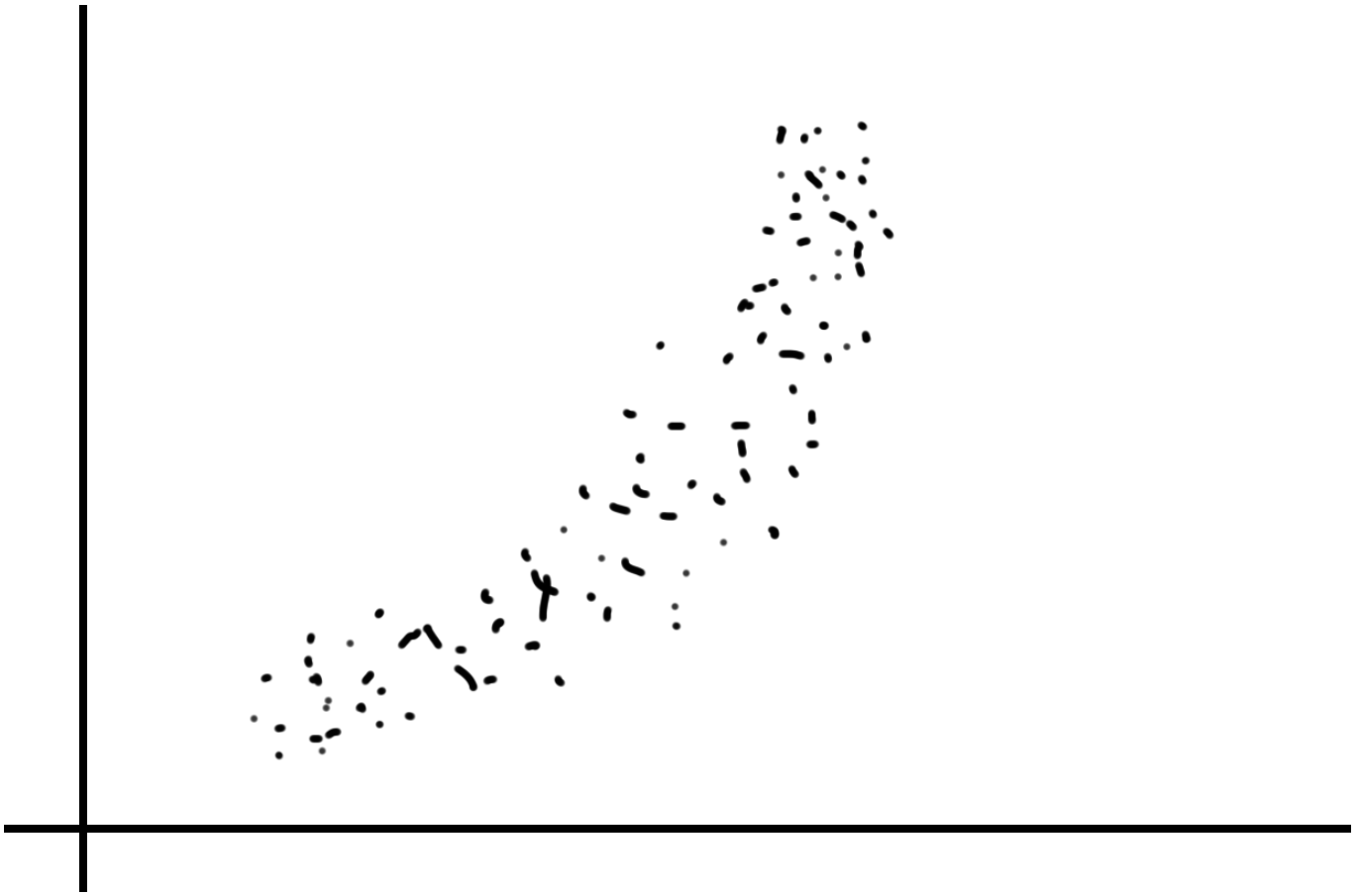
E[W]	E[H]	E[W*H]
= 62.75	= 52.75	= 3355.83



$$\begin{aligned} Cov(W, H) &= E[W*H] - E[W]E[H] \\ &= 3355.83 - (62.75)(52.75) \\ &= 45.77 \end{aligned}$$

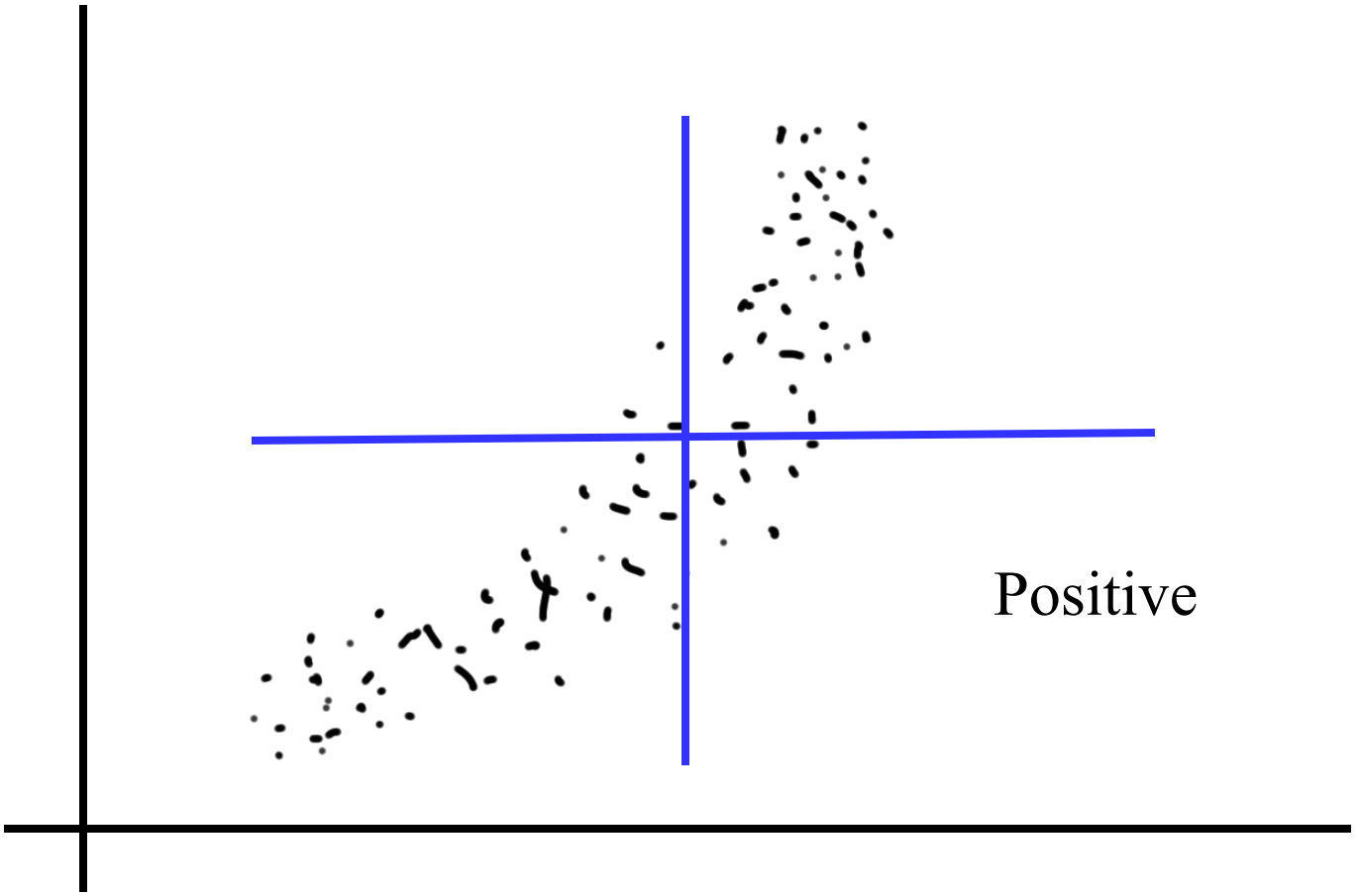
Covariance

Socrative: (a) positive, (b) negative, (c) zero



Covariance

Socrative: (a) positive, (b) negative, (c) zero



Independence and Covariance

- X and Y are random variables with PMF:

Y \ X	-1	0	1	$p_Y(y)$
0	1/3	0	1/3	2/3
1	0	1/3	0	1/3
$p_X(x)$	1/3	1/3	1/3	1

$$Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

  - E[X] = -1(1/3) + 0(1/3) + 1(1/3) = 0
  - E[Y] = 0(2/3) + 1(1/3) = 1/3
  - Since XY = 0, E[XY] = 0
  - Cov(X, Y) = E[XY] – E[X]E[Y] = 0 – 0 = 0
- But, X and Y are clearly dependent!

Properties of Covariance

- Say X and Y are arbitrary random variables
  - $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
  - $\mathrm{Cov}(X,X) = E[X^2] - E[X]E[X] = \mathrm{Var}(X)$
  - $\mathrm{Cov}(aX+b,Y) = a\mathrm{Cov}(X,Y)$
- Covariance of sums of random variables
  - $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are random variables
  - $\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$

Do Indicators Covary?

- Let I_A and I_B be indicators for events A and B

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \qquad I_B = \begin{cases} 1 & \text{if } B \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

  - $E[I_A] = P(A)$, $E[I_B] = P(B)$, $E[I_A I_B] = P(AB)$
  - $\text{Cov}(I_A, I_B) = E[I_A I_B] - E[I_A]\,E[I_B]$
    $= P(AB) - P(A)P(B)$
    $= P(A \mid B)P(B) - P(A)P(B)$
    $= P(B)[P(A \mid B) - P(A)]$
  - $\text{Cov}(I_A, I_B)$ determined by $P(A \mid B) - P(A)$
  - $P(A \mid B) > P(A) \Rightarrow \rho(I_A, I_B) > 0$
  - $P(A \mid B) = P(A) \Rightarrow \rho(I_A, I_B) = 0$ (and $\text{Cov}(I_A, I_B) = 0$)
  - $P(A \mid B) < P(A) \Rightarrow \rho(I_A, I_B) < 0$

Example of Covariance

- Consider rolling a 6-sided die
  - Let indicator variable X = 1 if roll is 1, 2, 3, or 4
  - Let indicator variable Y = 1 if roll is 3, 4, 5, or 6
- What is Cov(X, Y)?
  - E[X] = 2/3 and E[Y] = 2/3
  - E[XY] = $\sum_x \sum_y xy\ p(x,y)$

    = (0 * 0) + (0 * 1/3) + (0 * 1/3) + (1 * 1/3) = 1/3
  - Cov(X, Y) = E[XY] – E[X]E[Y] = 1/3 – 4/9 = -1/9
  - Consider: P(X = 1) = 2/3 and P(X = 1 | Y = 1) = 1/2
    - Observing Y = 1 makes X = 1 *less* likely

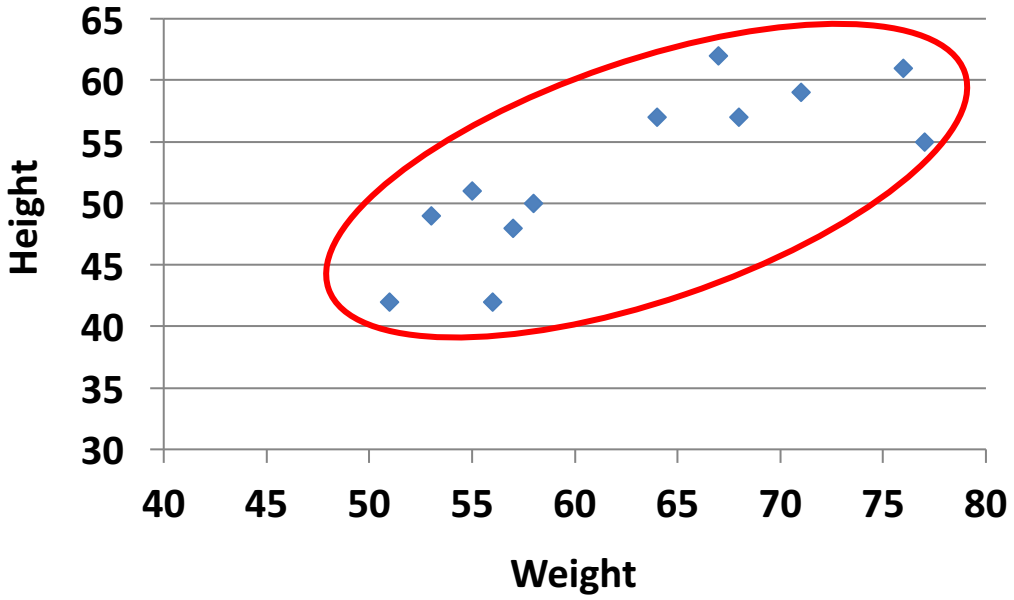
Correlation

What is Wrong With This?

- Consider the following data:

Weight	Height	Weight * Height
64	57	3648
71	59	4189
53	49	2597
67	62	4154
55	51	2805
58	50	2900
77	55	4235
57	48	2736
56	42	2352
51	42	2142
76	61	4636
68	57	3876

E[W]	E[H]	E[W*H]
= 62.75	= 52.75	= 3355.83



Cov(W, H) = E[W*H] – E[W]E[H]

= 3355.83 – (62.75)(52.75)

= 45.77

Viva La Correlatión

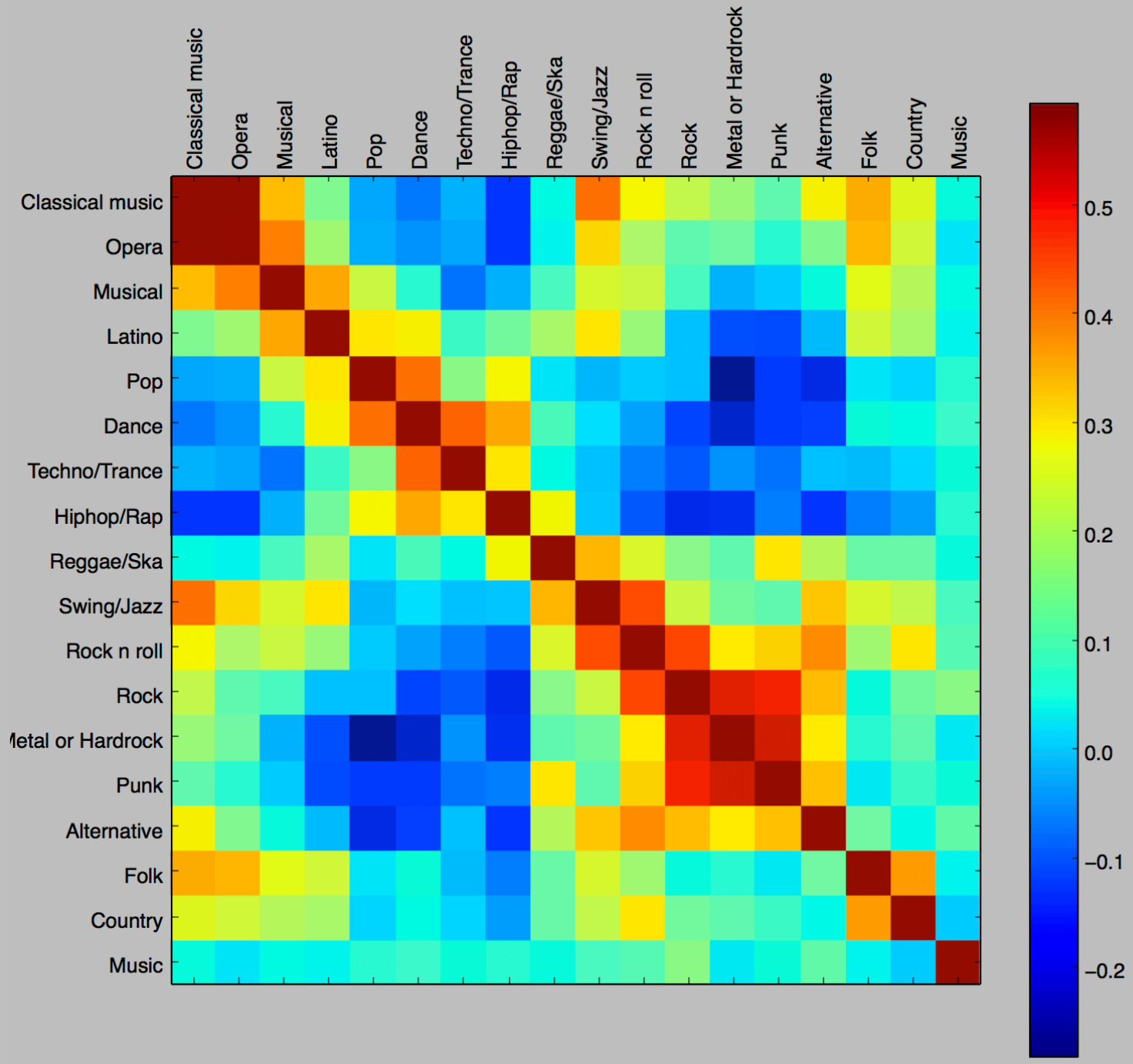
- Say X and Y are arbitrary random variables
  - Correlation of X and Y, denoted $\rho(X, Y)$:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

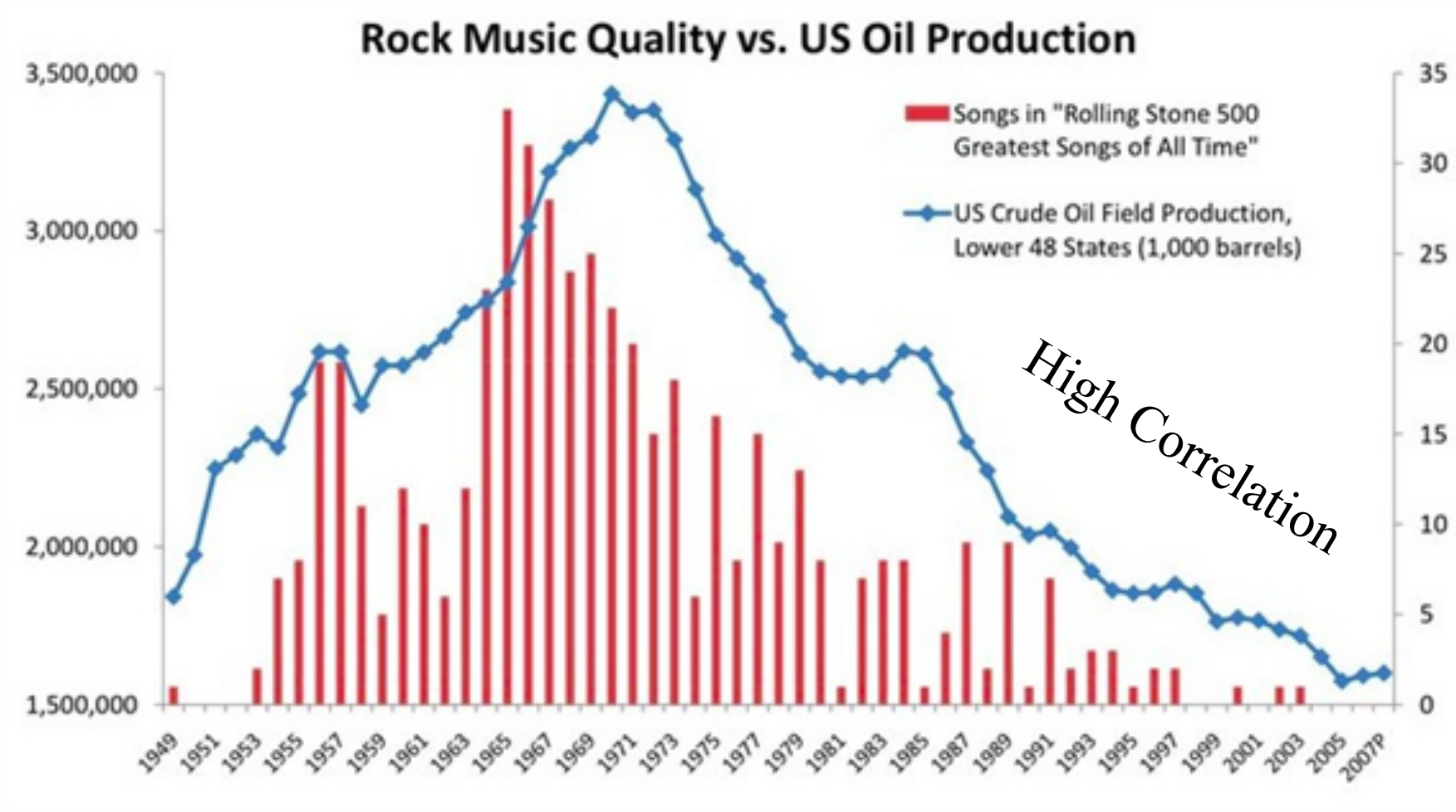
  - Note: $-1 \leq \rho(X, Y) \leq 1$
  - Correlation measures <u>linearity</u> between X and Y
  - $\rho(X, Y) = 1 \quad \Rightarrow$ Y = aX + b where $a = \sigma_y/\sigma_x$
  - $\rho(X, Y) = -1 \quad \Rightarrow$ Y = aX + b where $a = -\sigma_y/\sigma_x$
  - $\rho(X, Y) = 0 \quad \Rightarrow$ absence of <u>linear</u> relationship
    - But, X and Y can still be related in some other way!
  - If $\rho(X, Y) = 0$, we say X and Y are “uncorrelated”
    - Note: Independence implies uncorrelated, but **<u>not</u>** vice versa!

Viva La Correlatión

1010 students in the UK gave their preferences on everything from music genres to interests


Classical music
Opera
Musical
Latino
Pop
Dance
Techno/Trance
Hiphop/Rap
Reggae/Ska
Swing/Jazz
Rock n roll
Rock
Metal or Hardrock
Punk
Alternative
Folk
Country
Music
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2

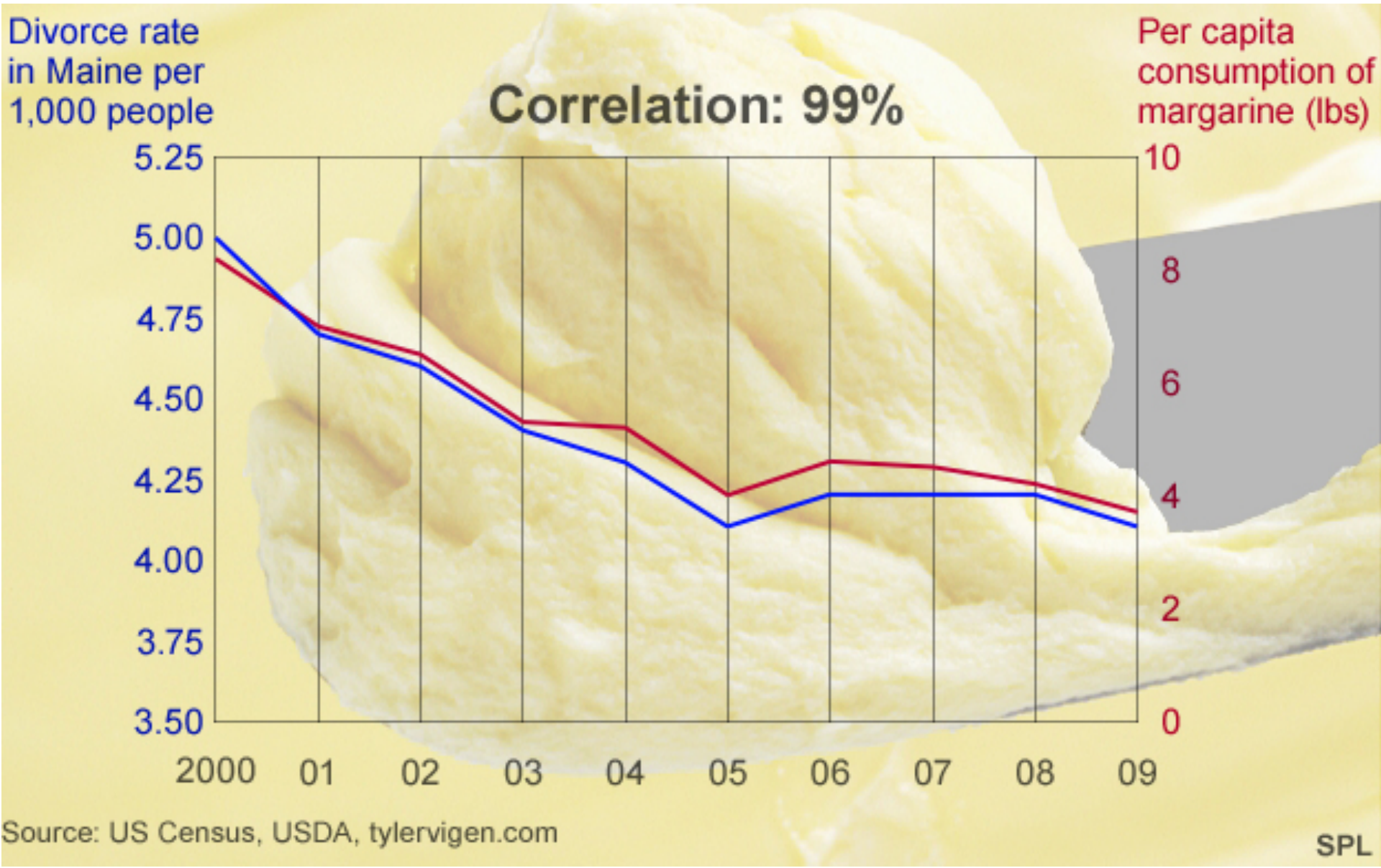
Rock Music Vs Oil?



Hubbert Peak Theory

http://www.aei.org/publication/blog/

Divorce Vs Butter?



http://www.bbc.com/news/magazine-27537142

If we have time

Covariance and the Multinomial

- Computing Cov(X_i, X_j)
  - Indicator $I_i(k)$ = 1 if trial *k* has outcome *i*, 0 otherwise

$$E[I_i(k)] = p_i \qquad X_i = \sum_{k=1}^{n} I_i(k) \qquad X_j = \sum_{k=1}^{n} I_j(k)$$

  - $\mathrm{Cov}(X_i, X_j) = \sum_{a=1}^{n}\sum_{b=1}^{n} \mathrm{Cov}(I_i(b), I_j(a))$
  - When $a \neq b$, trial *a* and *b* independent: $\mathrm{Cov}(I_i(b), I_j(a)) = 0$
  - When $a = b$: $\mathrm{Cov}(I_i(b), I_j(a)) = E[I_i(a)I_j(a)] - E[I_i(a)]E[I_j(a)]$
  - Since trial *a* cannot have outcome *i* and *j*: $E[I_i(a)I_j(a)] = 0$

$$\mathrm{Cov}(X_i, X_j) = \sum_{a=b=1}^{n} \mathrm{Cov}(I_i(b), I_j(a)) = \sum_{a=1}^{n} (-E[I_i(a)]E[I_j(a)])$$

$$= \sum_{a=1}^{n} (-p_i p_j) = -np_i p_j \quad \Rightarrow X_i \text{ and } X_j \text{ negatively correlated}$$

Multinomials All Around

- Multinomial distributions:
  - Count of strings hashed into buckets in hash table
  - Number of server requests across machines in cluster
  - Distribution of words/tokens in an email
  - Etc.
- When *m* (# outcomes) is large, p_i is small
  - For equally likely outcomes: $p_i = 1/m$

$$\mathrm{Cov}(X_i, X_j) = -np_i p_j = -\frac{n}{m^2}$$

  - Large *m* $\Rightarrow$ X_i and X_j very mildly negatively correlated
  - Poisson paradigm applicable

Que te vayas bien