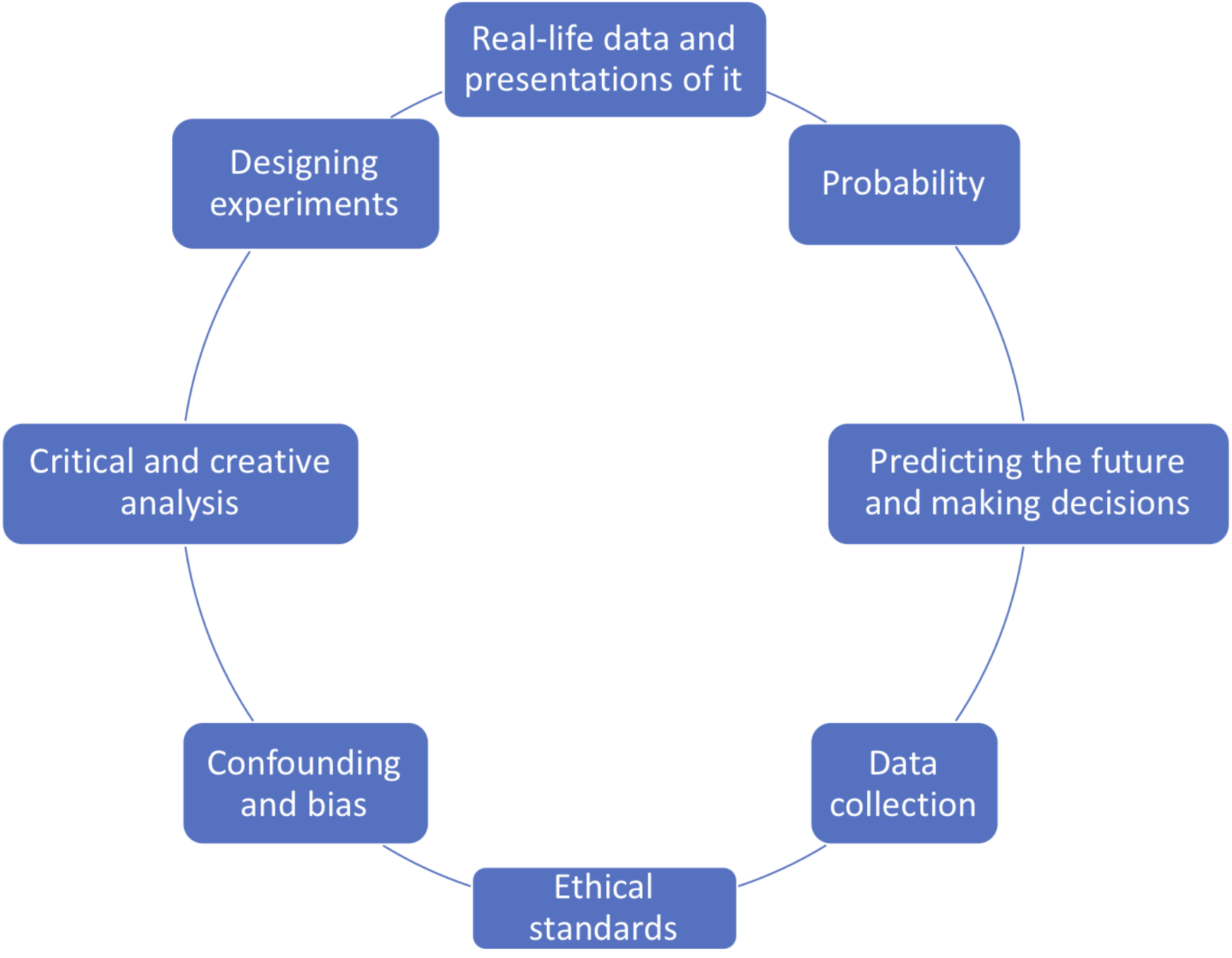
Case Studies

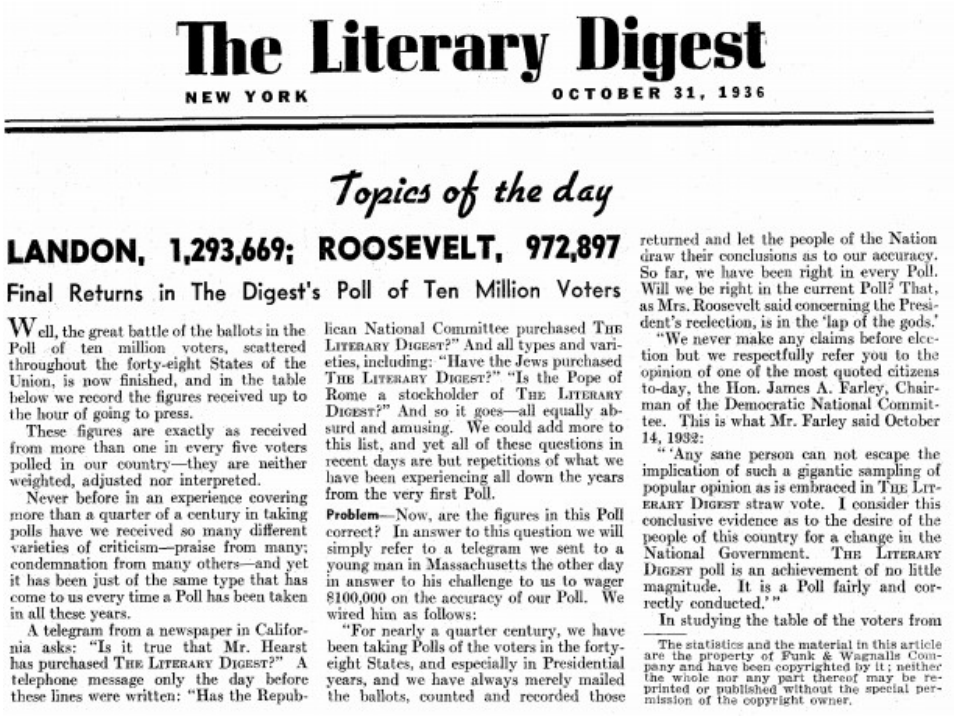


With regard to the 1936 Landon and Roosevelt election Literary Digest poll, which predicted the winner as Landon

a) the sample size was not large enough
b) the sample was not diverse enough
c) Landon would have won—but Roosevelt's win was due to a last minute change in sentiment
d) other

With regard to the 1936 Landon and Roosevelt election Literary Digest poll, which predicted the winner as Landon

a) the sample size was not large enough
b) the sample was not diverse enough
c) Landon would have won—but Roosevelt's win was due to a last minute change in sentiment
d) other



The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased The Literary Digest?" A telephone message only the day before these lines were written: "Has the Republican National Committee purchased The Literary Digest?" And all types and varieties, including: "Have the Jews purchased The Literary Digest?" "Is the Pope of Rome a stockholder of The Literary Digest?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager $100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in The Literary Digest straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. The Literary Digest poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.'"

In studying the table of the voters from

Literary Digest, October 31, 1936

Which of the following in the hw readings did you find most compelling?

a) inconsistencies in scaling can lead to false interpretations
b) the average American or average salary of Lakeside school can be very misleading. Half of the people are not necessarily below average.
c) sampling pitfalls such as convenience sampling, voluntary responses, and asking unclear or misleading questions
d) other

Which of the following in the hw readings did you find most compelling?

a) inconsistencies in scaling can lead to false interpretations
b) the average American or average salary of Lakeside school can be very misleading. Half of the people are not necessarily below average.
c) sampling pitfalls such as convenience sampling, voluntary responses, and asking unclear or misleading questions
d) other

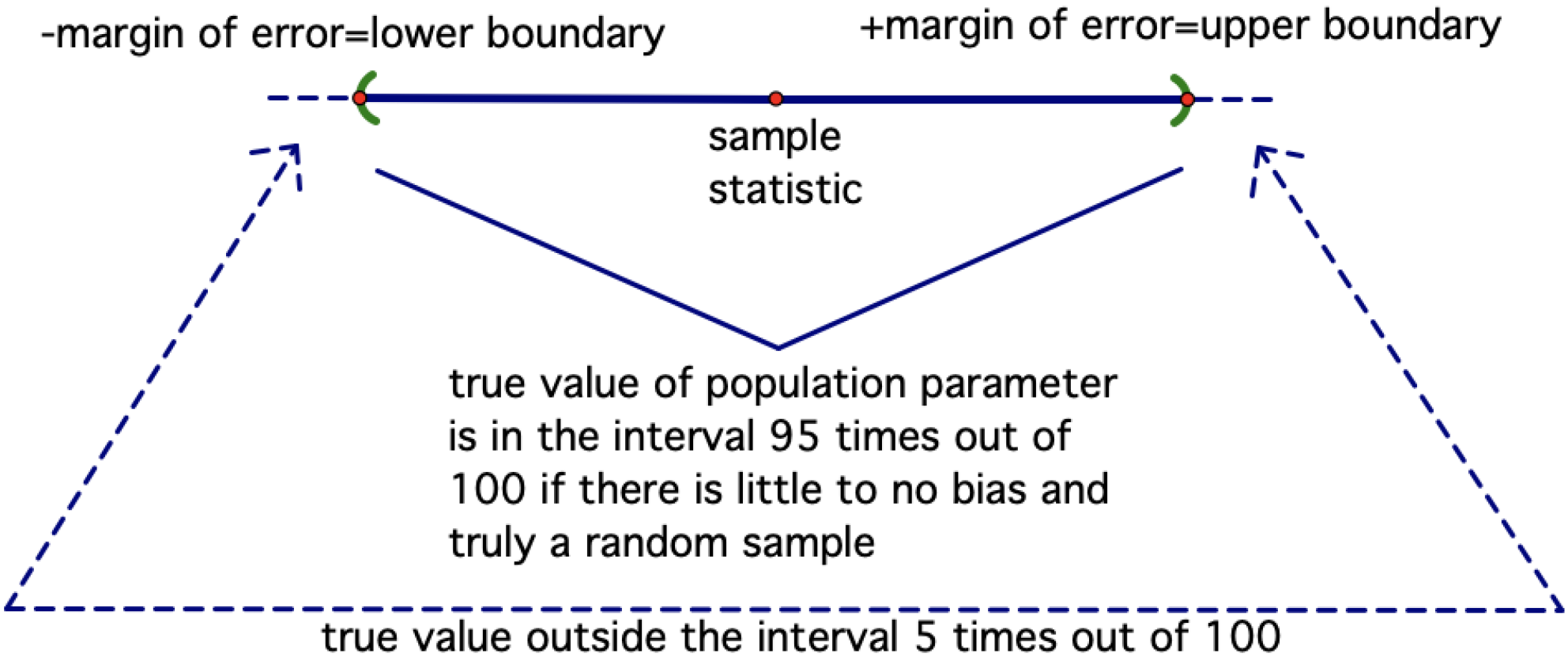
collecting data: reproducibility, consensus, and random sampling if possible

presenting data: entire data set versus numerical or visual snapshots of it

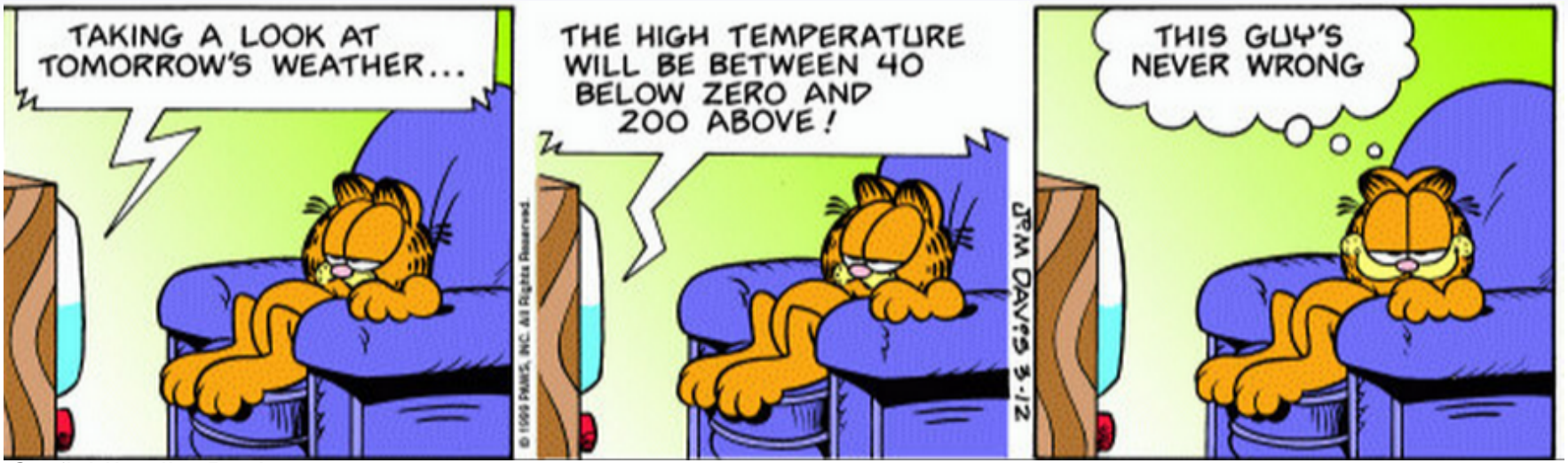
all can be subject to bias and distortion, and are definitely subject to probability and random variations

95% Confidence Level and Margin of Error Intervals

- a sample statistic with its margin of error generates an interval on a number line
- 95% provides a likelihood—no way to know which intervals contain the true percentage and which don't

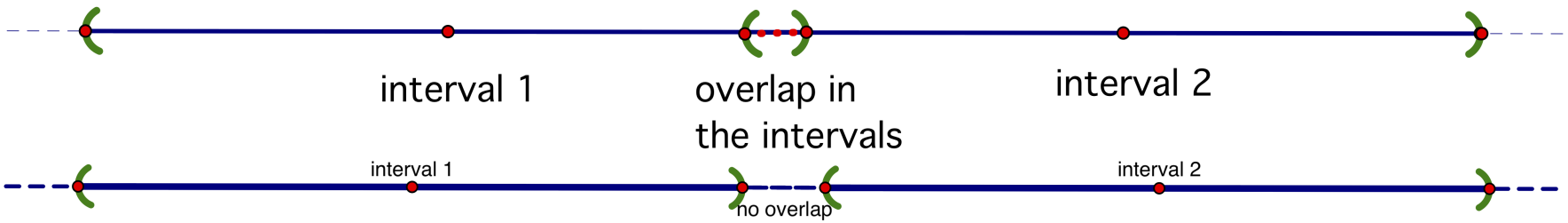


Margin of Error



Garfield by Jim Davis `https://garfield.com/comic/1999/03/12`

- margin of error gives a range the actual percentage is likely to be within. Higher confidence level has a wider interval.
- For a 95% confidence interval, a sample of size n will have margin of error approximately $\frac{1}{\sqrt{n}}$ (conservative estimate).
- We check for overlaps in the intervals in order to evaluate the statistical validity of headlines and statements in polls



Statistically Accurate Claim?

"Desire to Migrate Rises in North Africa"

2017 lower boundary: $32 - 2 = 30\%$

2016 upper boundary: $28 + 2 = 30\%$

WORLD APRIL 24, 2018

Desire to Migrate Rises in North Africa

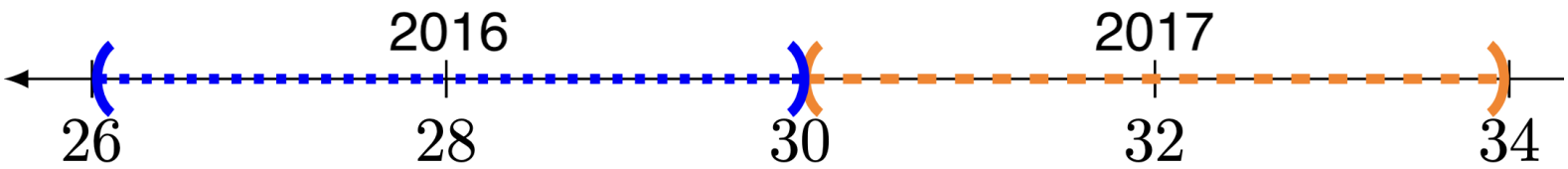
BY IMAN BERRACHED AND **RJ REINHART**

NORTH AFRICANS WHO WOULD LIKE TO MIGRATE TO ANOTHER COUNTRY

2016	2017
28%	32%

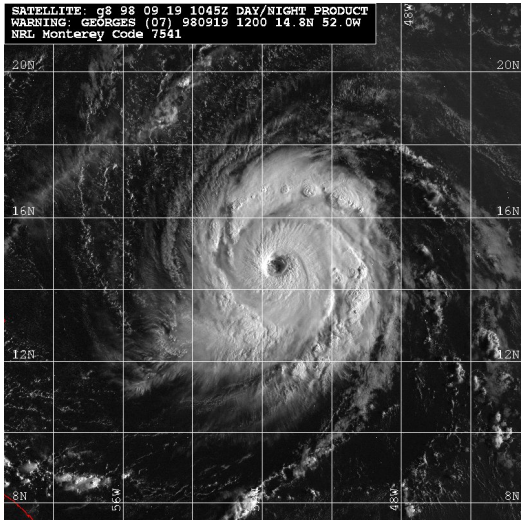
GALLUP WORLD POLL

it could have stayed the same!



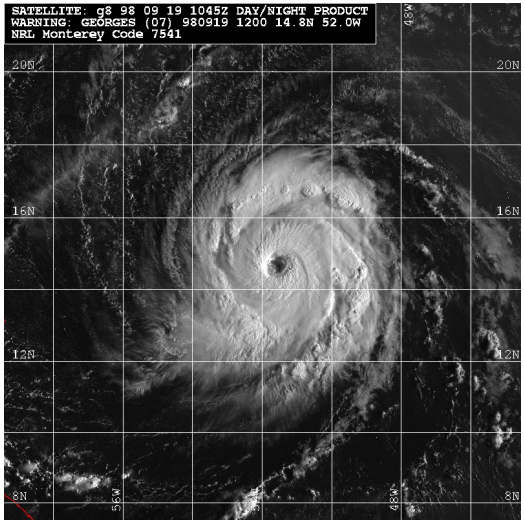
Statistics from Nature

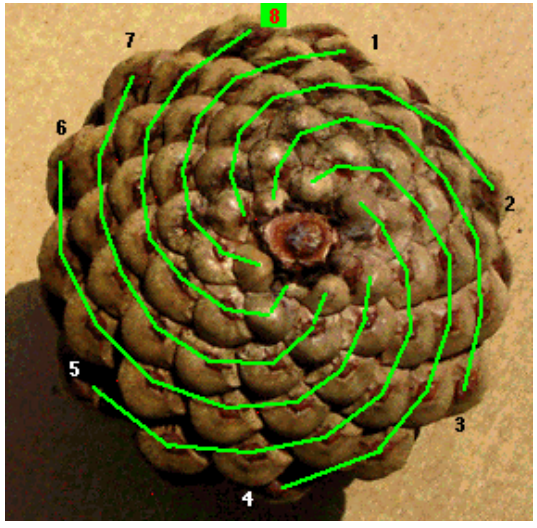
Benford's Law and the likely frequency of the first digit applied to many data sets. Is there a pattern to spirals in nature?

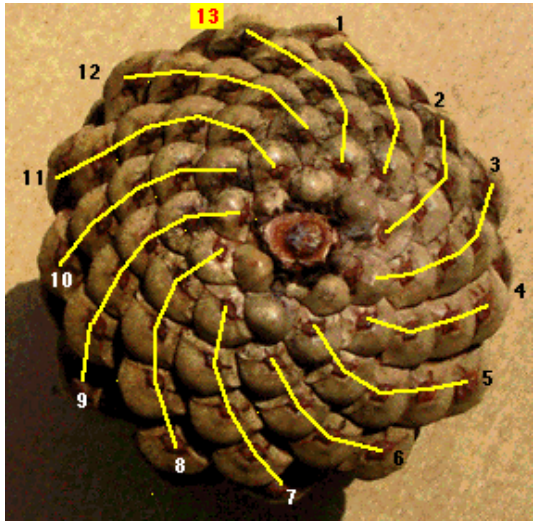


Statistics from Nature

Benford's Law and the likely frequency of the first digit applied to many data sets. Is there a pattern to spirals in nature?

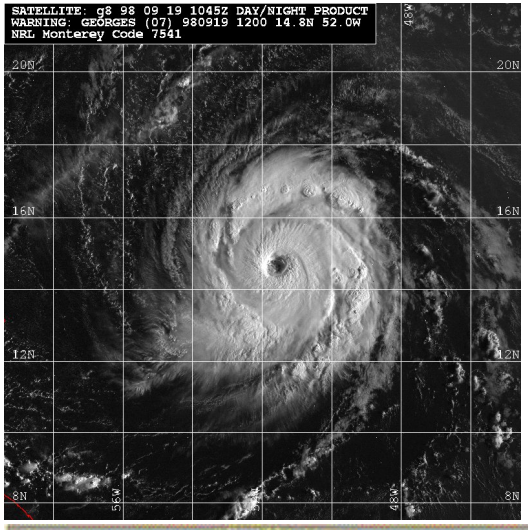


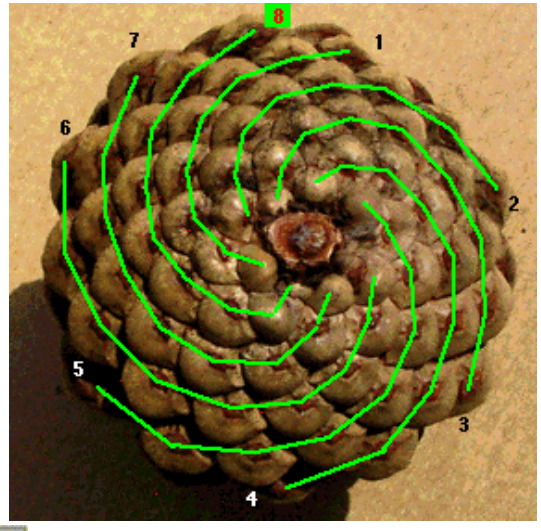


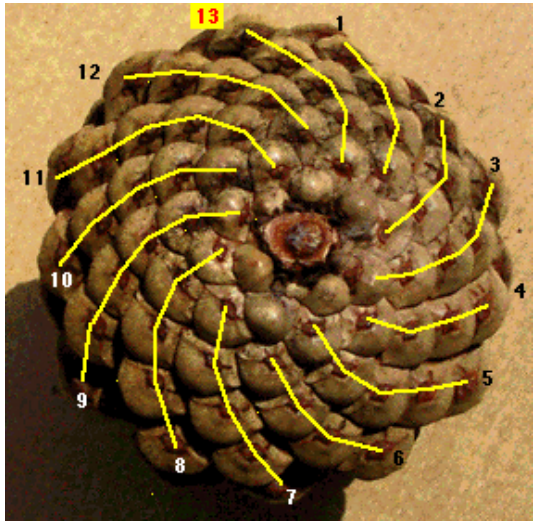


Statistics from Nature

Benford's Law and the likely frequency of the first digit applied to many data sets. Is there a pattern to spirals in nature?

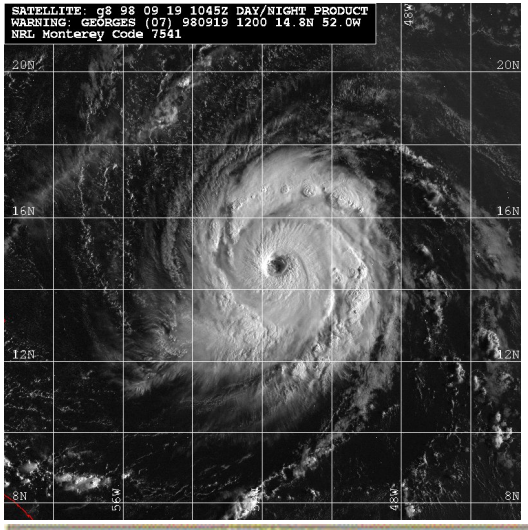


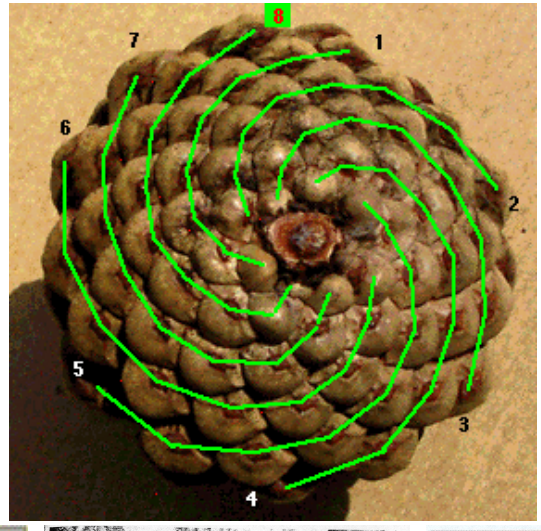


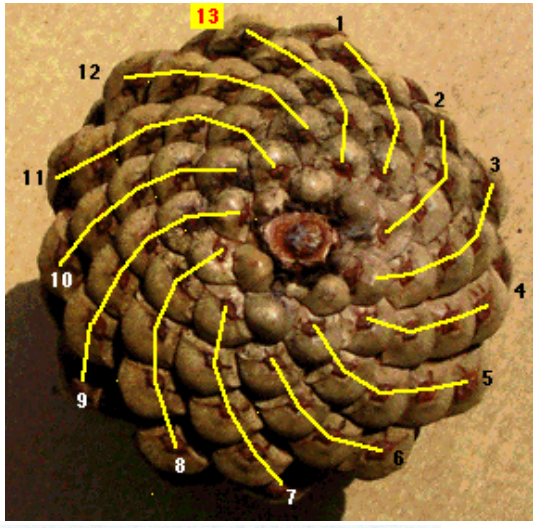


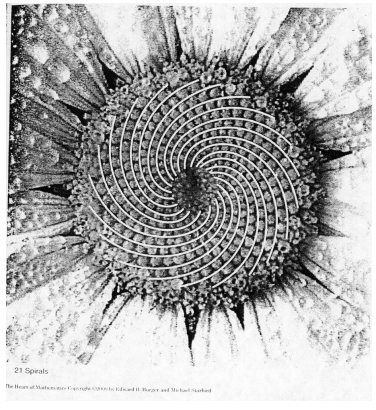
Statistics from Nature

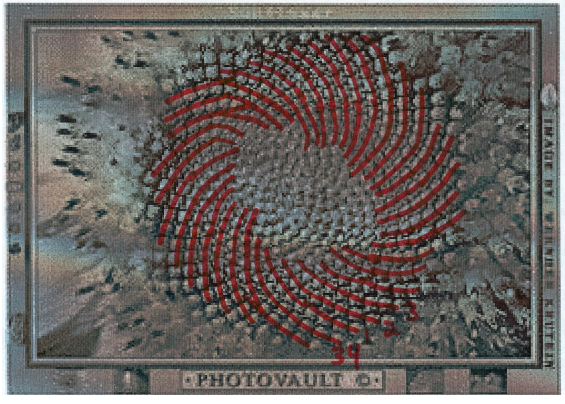
Benford's Law and the likely frequency of the first digit applied to many data sets. Is there a pattern to spirals in nature?











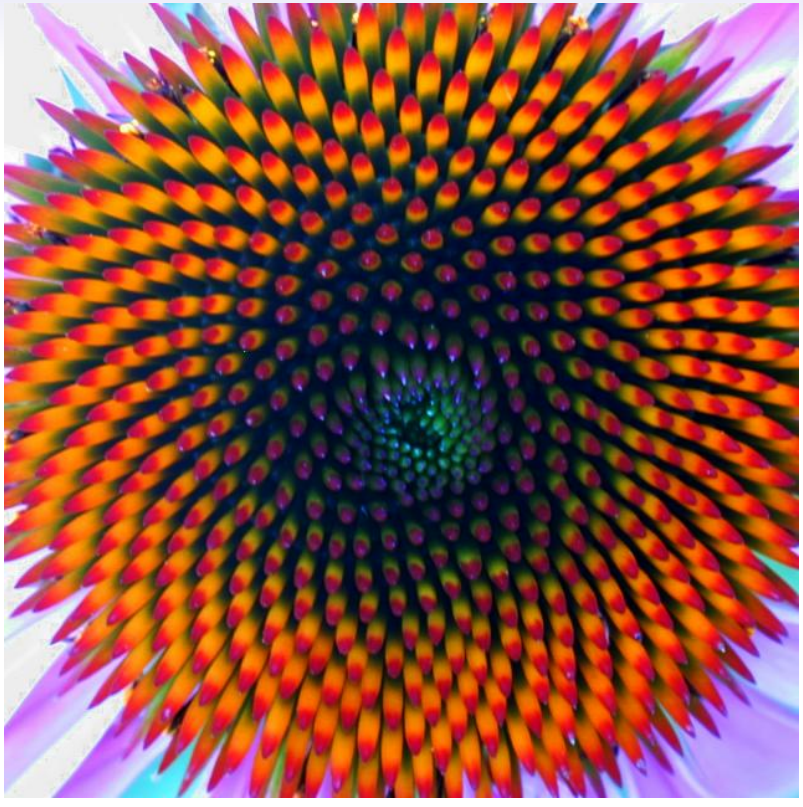
Picture credits:

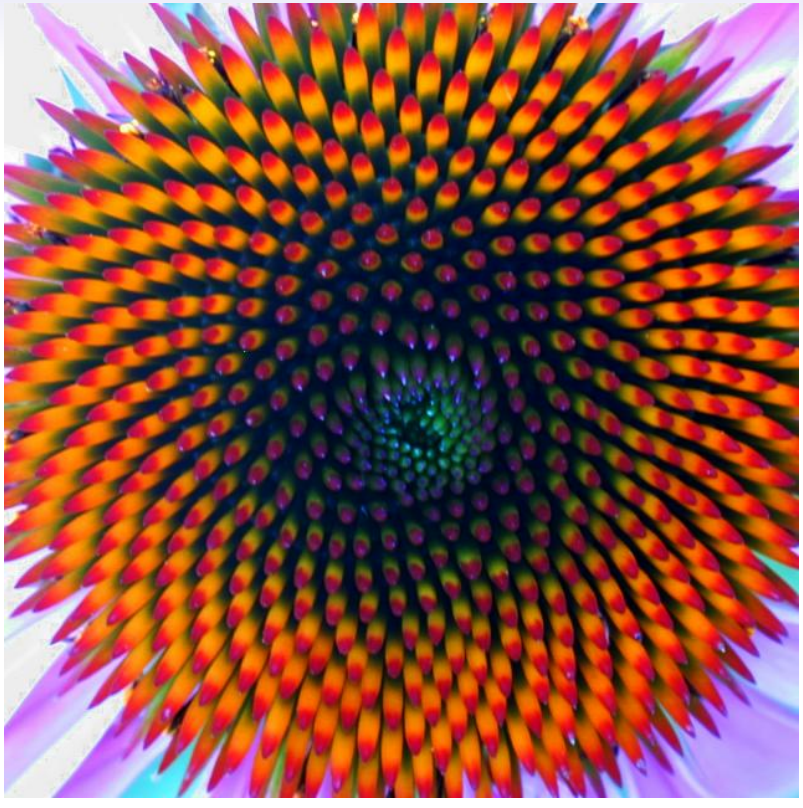
1. Clockwise hurricane Georges originated in southern hemisphere:
`http://www.aoml.noaa.gov/hrd/Storm_pages/georges1998/sat.html`

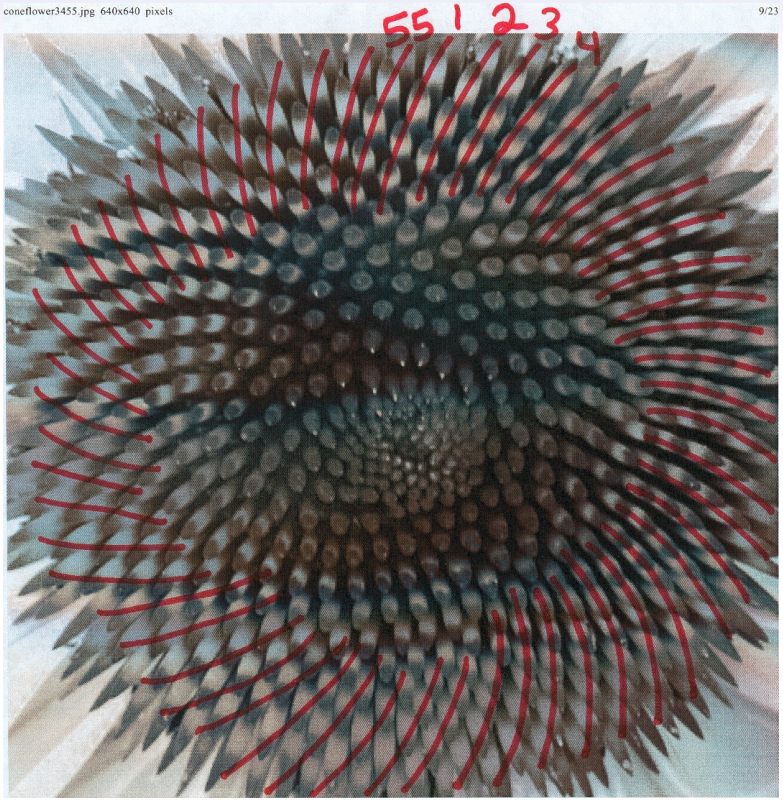
2–3. Ron Knott `http://www.maths.surrey.ac.uk/hosted-sites/R.Knott/Fibonacci/fibnat.html`

4 and 6. Wernher Krutein

5. *The Heart of Mathematics*



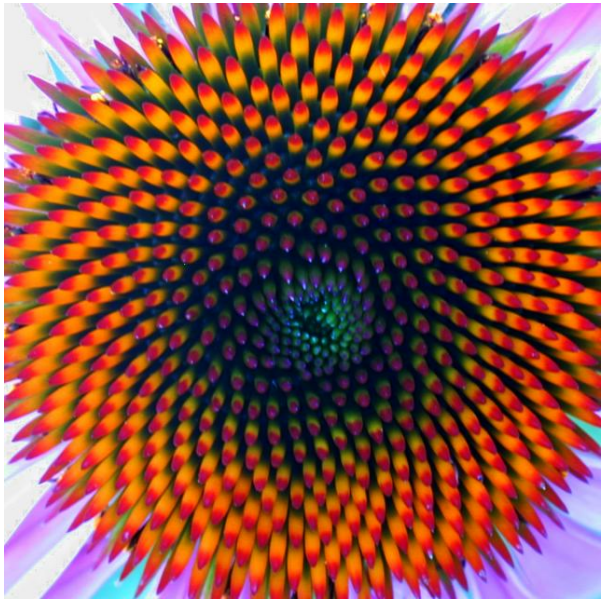




Picture credit:

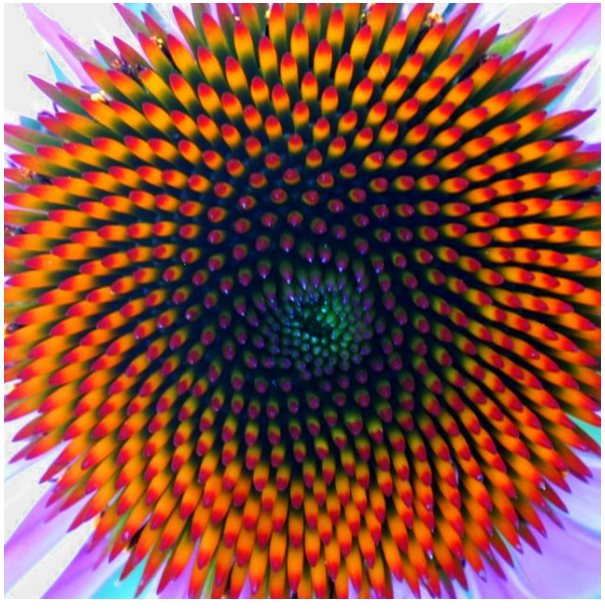
Ron Knott `http://www.maths.surrey.ac.uk/hosted-sites/R.Knott/Fibonacci/fibnat.html`

- Explore ideas systematically
- Look for a pattern
- Create abstract ideas by modeling nature
- Unexpected patterns are often a sign of hidden, underlying structure [and Excel can help us find it]
- Explore the consequences of new ideas



Picture credit:
Ron Knott `http://www.maths.surrey.ac.uk/hosted-sites/R.Knott/Fibonacci/fibnat.html`

- Explore ideas systematically
- Look for a pattern
- Create abstract ideas by modeling nature
- Unexpected patterns are often a sign of hidden, underlying structure [and Excel can help us find it]
- Explore the consequences of new ideas



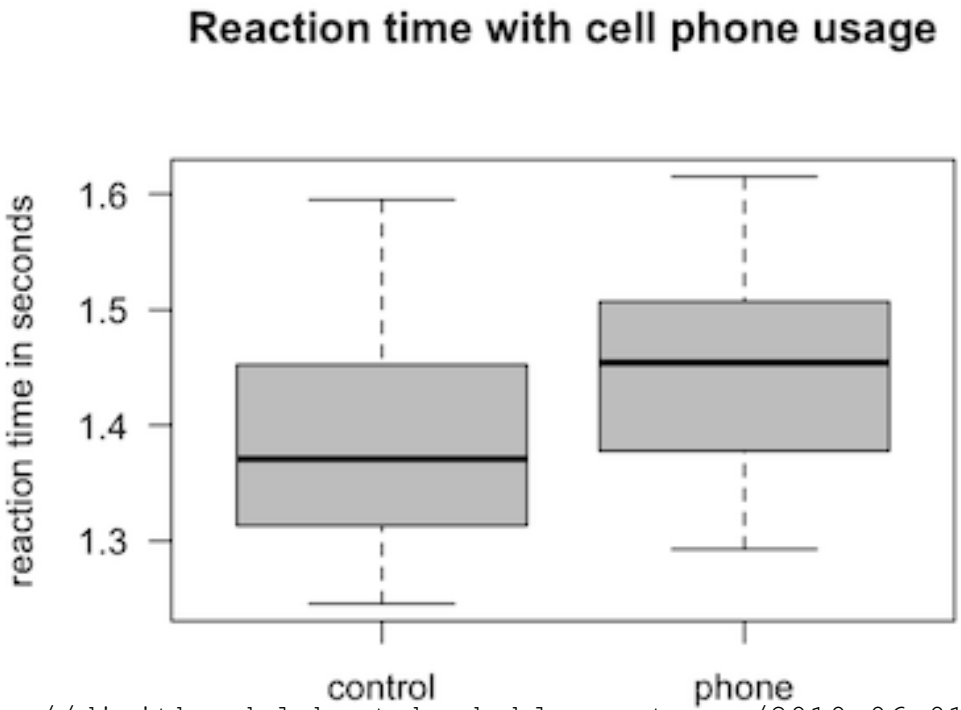
Picture credit:
Ron Knott `http://www.maths.surrey.ac.uk/hosted-sites/R.Knott/Fibonacci/fibnat.html`

$\frac{1+\sqrt{5}}{2} \approx 1.6180339887498948482045868343656381177 20$

Why is it plausible that there is a pattern to spirals in nature? Where else do we find the same pattern?

`http://www.youtube.com/watch?v=lOIP_Z_-0Hs`

In the following we see two side-by-side boxplots on reaction times of control group vs cell users



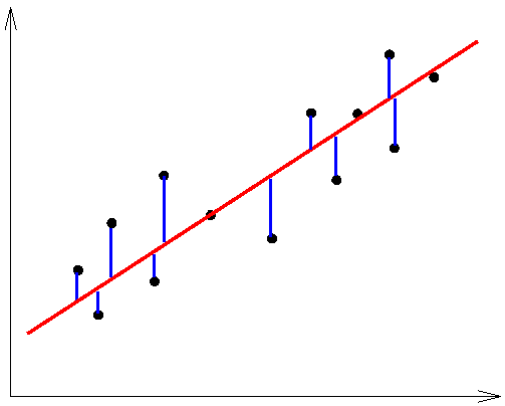
Picture: http://digitheadslabnotebook.blogspot.com/2010_06_01_archive.html

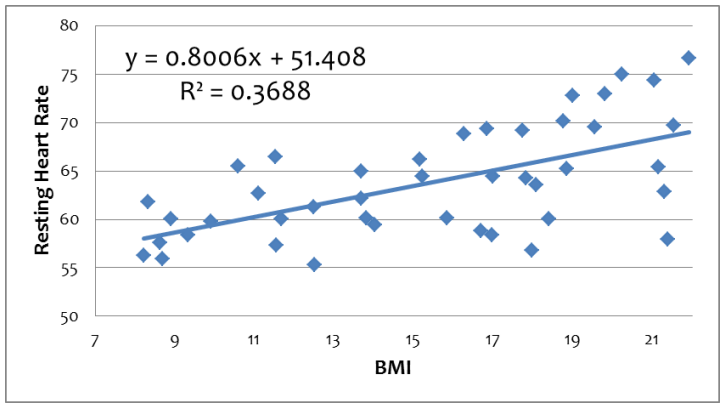
What can you say about the data from the median to Q3, the third quartile, of the reaction times?

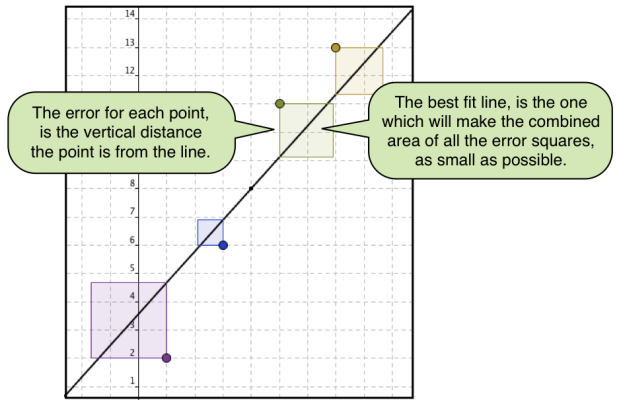
a) cell phone users did better because the data is more tightly clustered together
b) control group did better because the data is lower
c) neither

Strength of the Relationship: r^2 percent

- 0 to 10% no
  10% to 25% weak
  25% to 65% moderate
  above 65% strong
- NOT a probability for correct nor a likelihood of on the line
- measures the y-values distances via sum of squares as variation in the dependent variable explained by linearity







Picture citations:

1. http://cs.wellesley.edu/~cs199/lectures/35-correlation-regression.html
2. http://www2.nau.edu/mat114-c/ch3a.php
3. http://math.maine121.org/welcome/chapter-5/

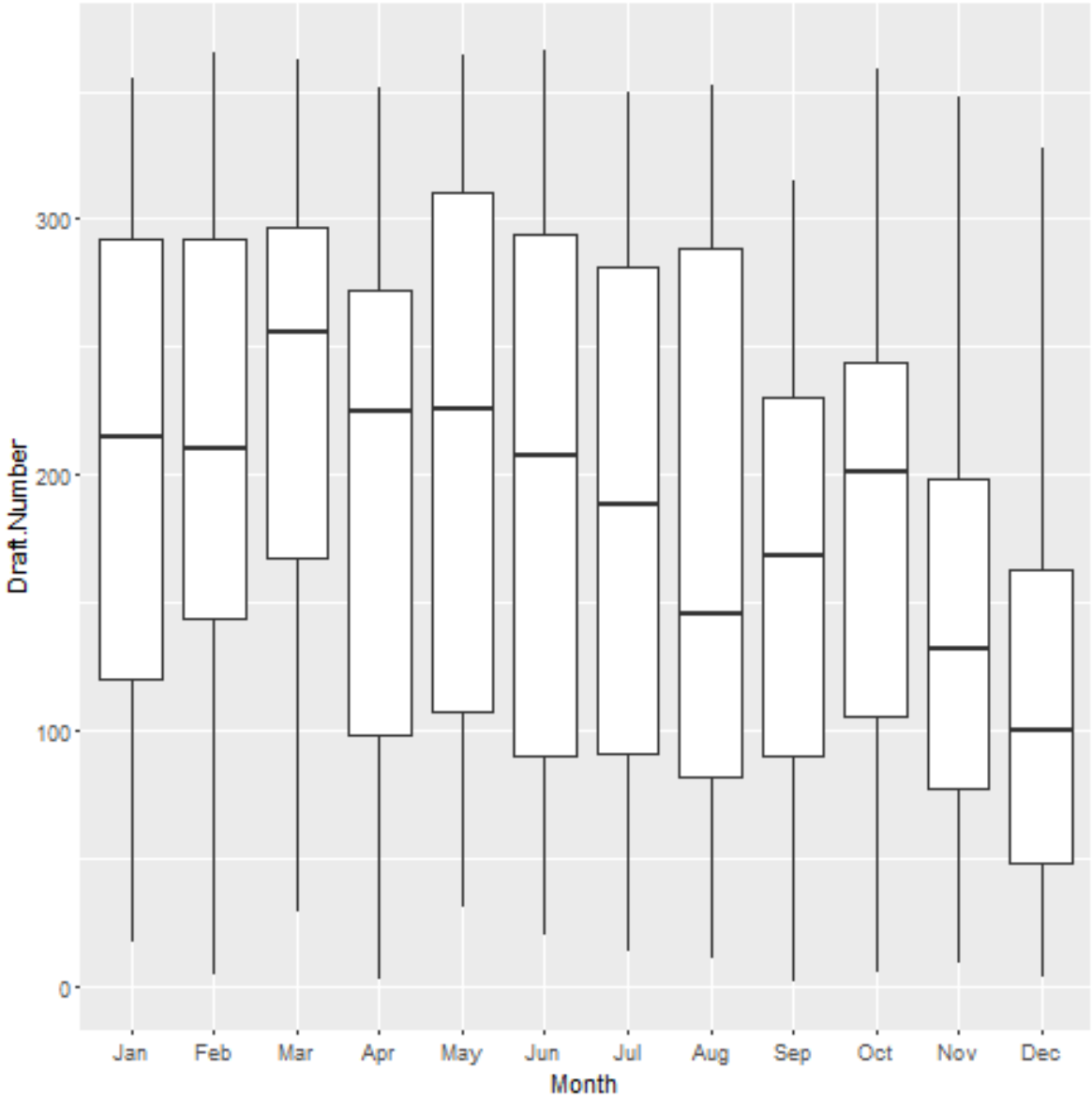
Would you have been drafted for Vietnam in 1969? Is there anyone in the class with the same birthday?

December 1, 1969 Vietnam Draft Lottery

https://www.youtube.com/watch?v=-p5X1FjyD_g

December 1, 1969 Vietnam Draft Lottery

https://www.youtube.com/watch?v=-p5X1FjyD_g



http://academic.uprm.edu/wrolke/esma3102/graphs/cor5.png

Joan Rosenblatt's 1970 Improvements

A truly unbiased military draft

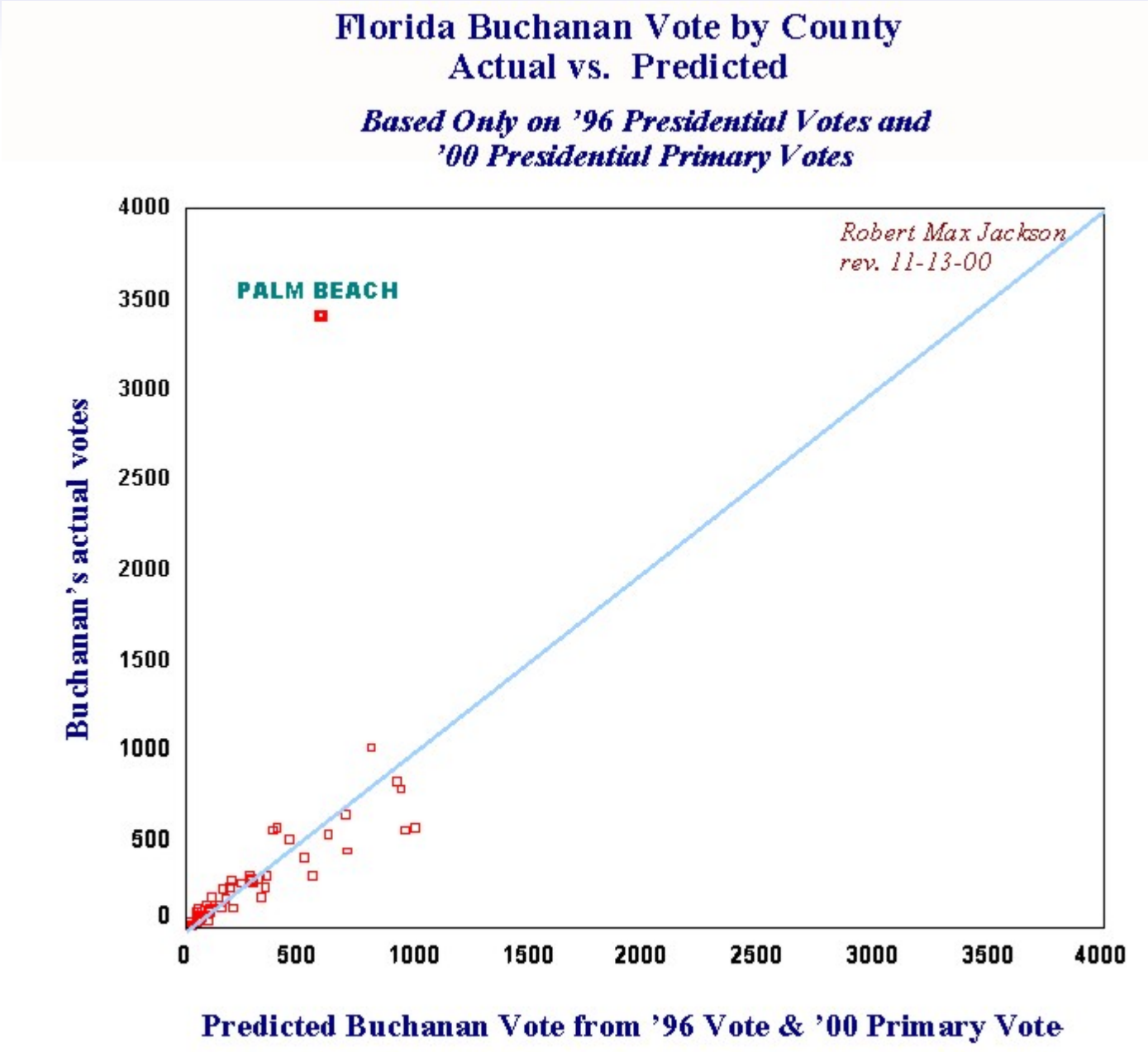
January 01, 1970

Statisticians roundly criticized the 1969 draft lottery as unfair, so the Selective Service System asked NIST to devise an unquestionably random method for the 1970 draft. NIST mathematician Joan Rosenblatt and colleagues developed a method to randomly choose calendars and priority permutations for the draft. The new draft method was praised as fair, and Rosenblatt won the 1971 Federal Women's Award for her efforts on this and other projects.

https://www.nist.gov/node/774336

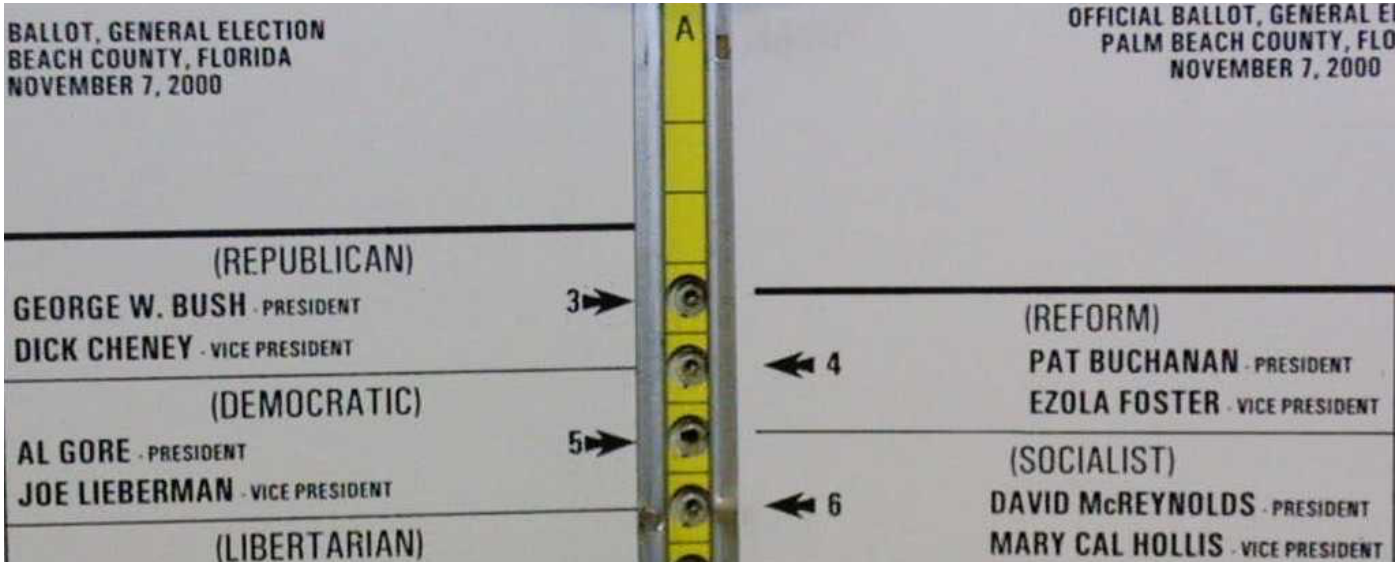
Lottery procedure was improved the next year with a two-drum system. The 365 birthdates (for those born in 1951) were written down, placed in capsules, and put in a drum in the order dictated by random permutations. Similarly, the numbers from 1 to 365 were written down and placed into capsules. One drum was rotated for an hour and the other for a half-hour (its rotating mechanism failed). Pairs of capsules were then drawn, one from each drum, one with a 1951 birthdate and one with a number 1 to 365.

2000 Presidential Election—Bush, Gore, Buchanan



http://homepages.nyu.edu/~rmj1/objects/BuchPrio.jpg

2000 Presidential Election—Bush, Gore, Buchanan

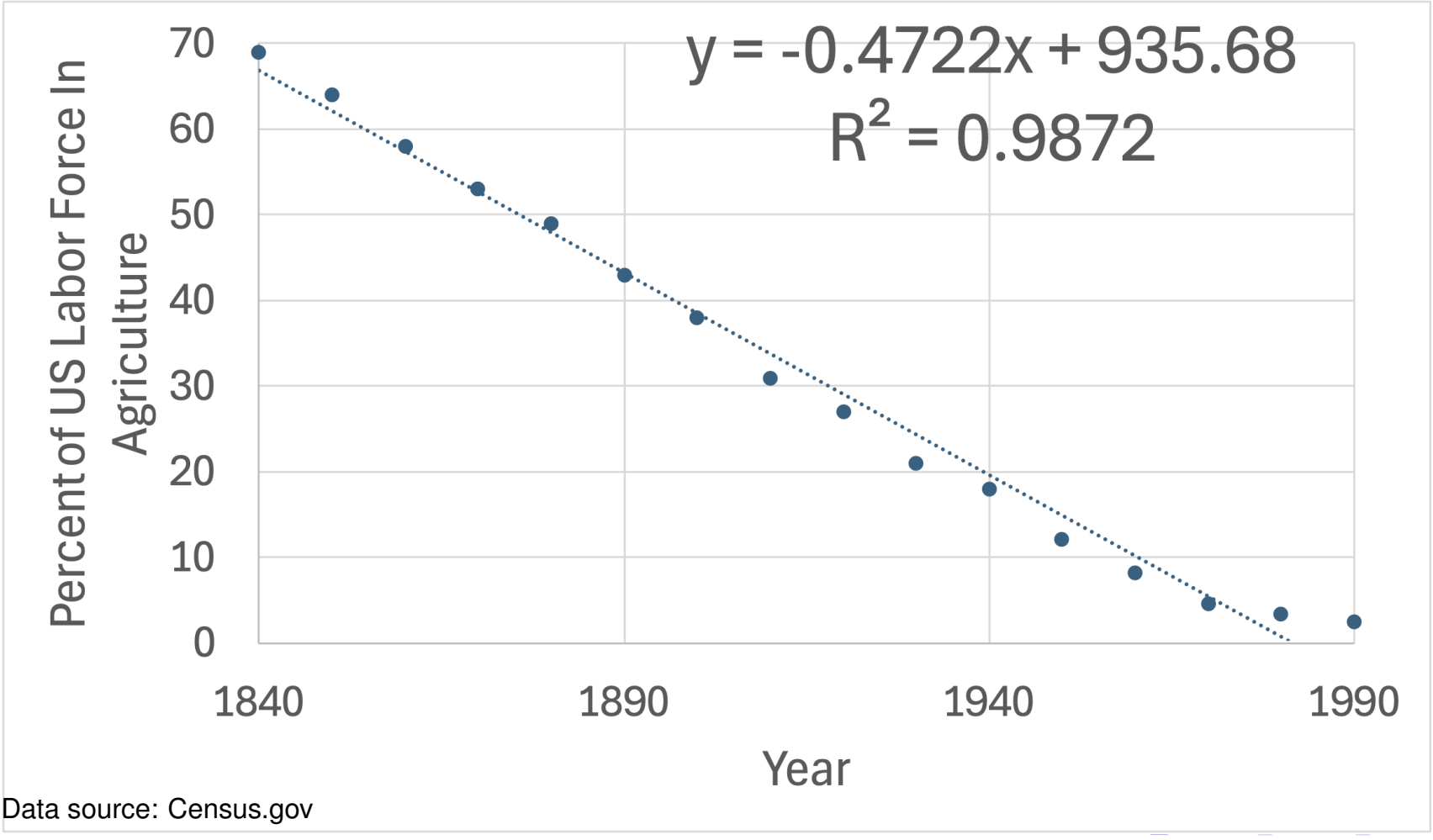


What does the slope mean when x= year and y= percent of US labor force working in agriculture and the best fit line is $y = -0.4722x + 935.68$?

a) at the start of the years the percentage is 935.68
b) percent of labor force is going down by about .5 with every year
c) neither

What does the slope mean when x= year and y= percent of US labor force working in agriculture and the best fit line is $y = -0.4722x + 935.68$?

a) at the start of the years the percentage is 935.68
b) percent of labor force is going down by about .5 with every year
c) neither



Data source: Census.gov

$r^2 = .9872 = 98.72\%$, and this tells us that

a) If we use the line to predict we would get it right $\sim 99\%$ of the time
b) The y-value distances of the data to the best fit line are small so year in this data is a statistically strong predictor of percent of US labor force

$r^2 = .9872 = 98.72\%$, and this tells us that

a) If we use the line to predict we would get it right $\sim 99\%$ of the time
b) The y-value distances of the data to the best fit line are small so year in this data is a statistically strong predictor of percent of US labor force

The r^2 value is strong but the line $y = -0.4722x + 935.68$ predicts a negative value by plugging in any year after 1990 for x. Resolve the apparent conflict.

a) There is a typo—the actual r^2 value should be weak
b) The mathematics of the r^2 value and the prediction are correct
c) There are other reasons why the prediction doesn't hold up like extrapolation

Related data: `https://ourworldindata.org/grapher/share-of-the-labor-force-employed-in-agriculture?time=latest`

How to Get Rich Quick as a Stock Whiz

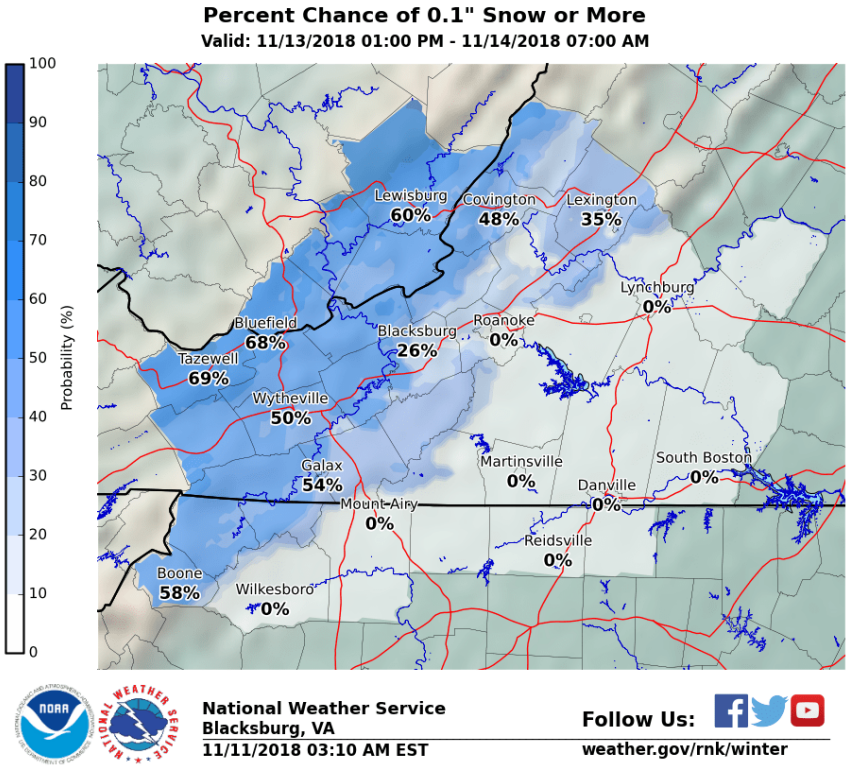
If the r^2 value was 100%, would we be assured to make money by using the best fit line to predict the future performance?

a) yes
b) no

How to Get Rich Quick as a Stock Whiz

If the r^2 value was 100%, would we be assured to make money by using the best fit line to predict the future performance?

a) yes
b) no



Picture and data: `https://www.weather.gov/rnk/winter`

$\geq$ .1"	$\geq$ 1"	$\geq$ 2"	$\geq$ 4"	$\geq$ 6"	$\geq$ 8"	$\geq$ 12"	$\geq$ 18"
58%	3%	0%	0%	0%	0%	0%	0%

The weather, stocks and more are chaotic dynamical systems with uncertainty within expected values

- With r^2, we can categorize correlations
  0 to 10% no
  10% to 25% weak
  25% to 65% moderate
  above 65% strong
- Even without r^2 in front of us, we can visually inspect and categorize relationships.
- NOT a probability for correct nor a likelihood of on the line
- measures the y-values distances via sum of squares as variation in the dependent variable explained by linearity