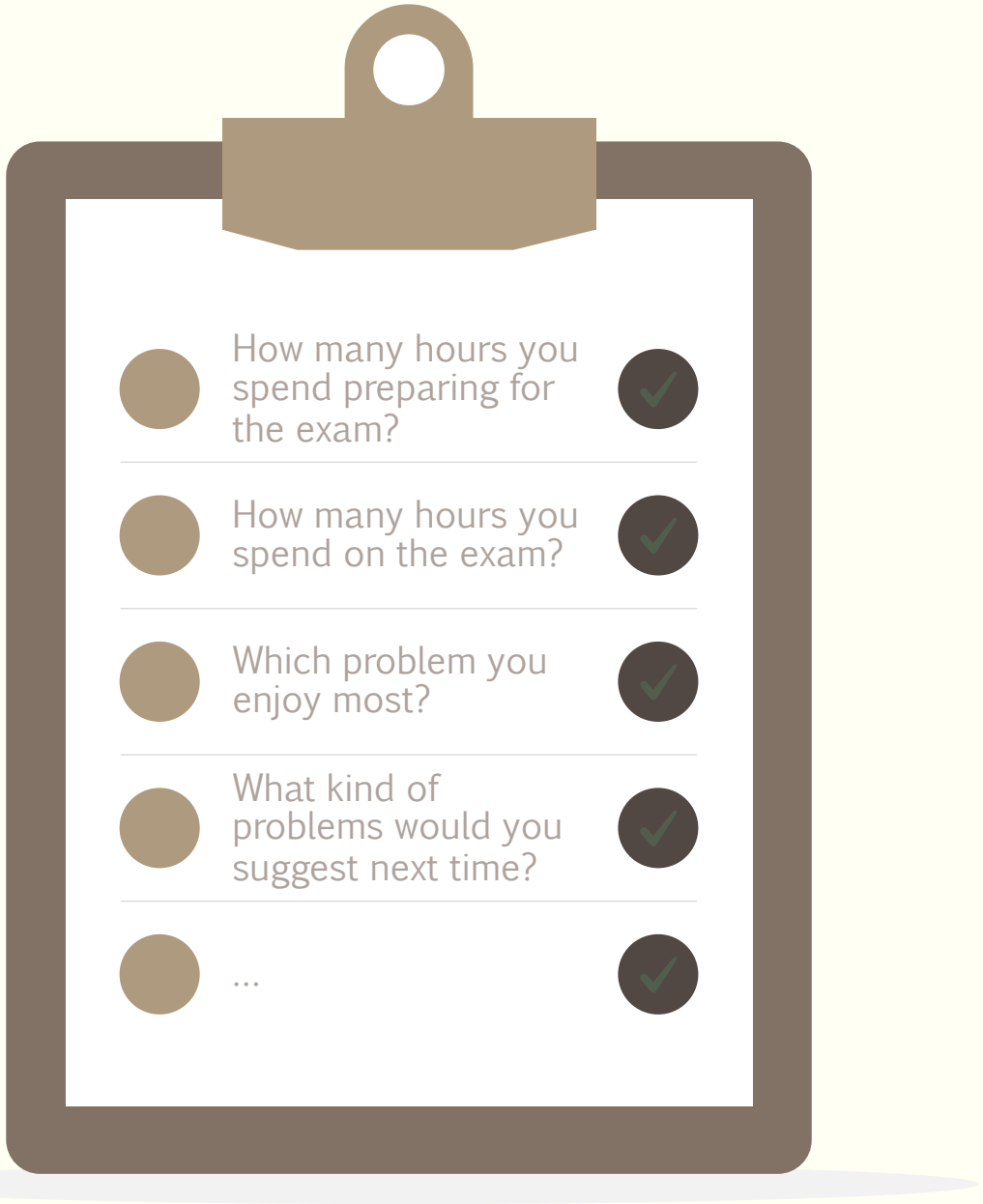
MATH 20: PROBABILITY

Midterm 1

Xingru Chen
xingru.chen.gr@dartmouth.edu

Exam Wrapper

for midterm 1



Problem 1: True or False

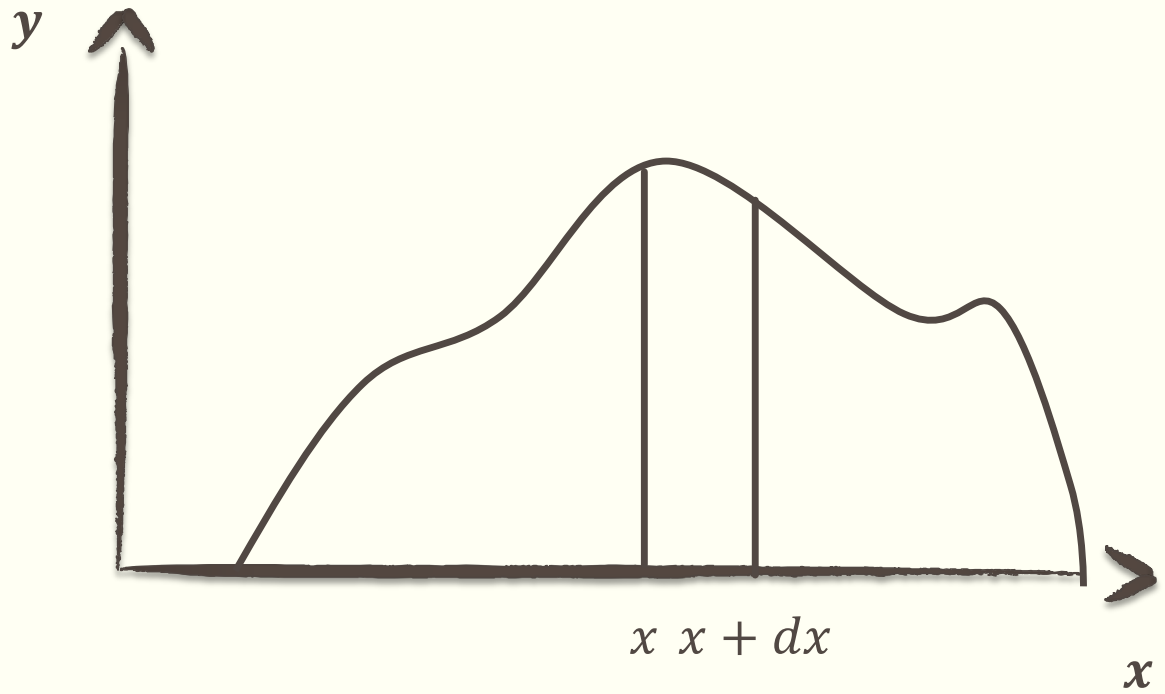
(b) False For a random variable T following an exponential distribution $f(t) = \lambda e^{-\lambda t}$, $P(T = \frac{\ln(2)}{\lambda}) = \frac{\lambda}{2}$.

Density Functions of Continuous Random Variable

- The probability of occurrence of an event of the form $[x, x + dx]$, where dx is small, can be estimated by

$$P([x, x + dx]) \approx f(x)dx.$$

- As $dx \to 0$, the above probability approaches 0, so that the probability of a single point x, $P(\{x\})$ is 0.



Problem 1: True or False

(e) False As $n \to \infty$, both the ratio and the difference of $n!$ and $n^n e^{-n}\sqrt{2\pi n}$ (**Stirling's Formula**) approach 0.

Problem 3: Proof

(a) Prove that for any positive integer $n \geq 1$,

$$\binom{2n}{0}+\binom{2n}{2}+\binom{2n}{4}+\cdots\binom{2n}{2n}=\binom{2n}{1}+\binom{2n}{3}+\binom{2n}{5}+\cdots\binom{2n}{2n-1}.$$

This identity is another proof for Question 5 in Quiz 4.

Binomial Theorem

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}.$$

=

Let $a = b = 1$, we have

$$2^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}.$$

=

Let $a = -1$, $b = 1$, we have

$$0 = \binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n \binom{n}{n}.$$

=

5 pts

Consider the **Binomial Theorem** $(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}$.

Let $a = -1$, $b = 1$ and use $2n$ instead of n. We have

$$0 = \binom{2n}{0} - \binom{2n}{1} + \binom{2n}{2} - \binom{2n}{3} + \cdots - \binom{2n}{2n-1} + \binom{2n}{2n},$$

which further leads to the original identity.

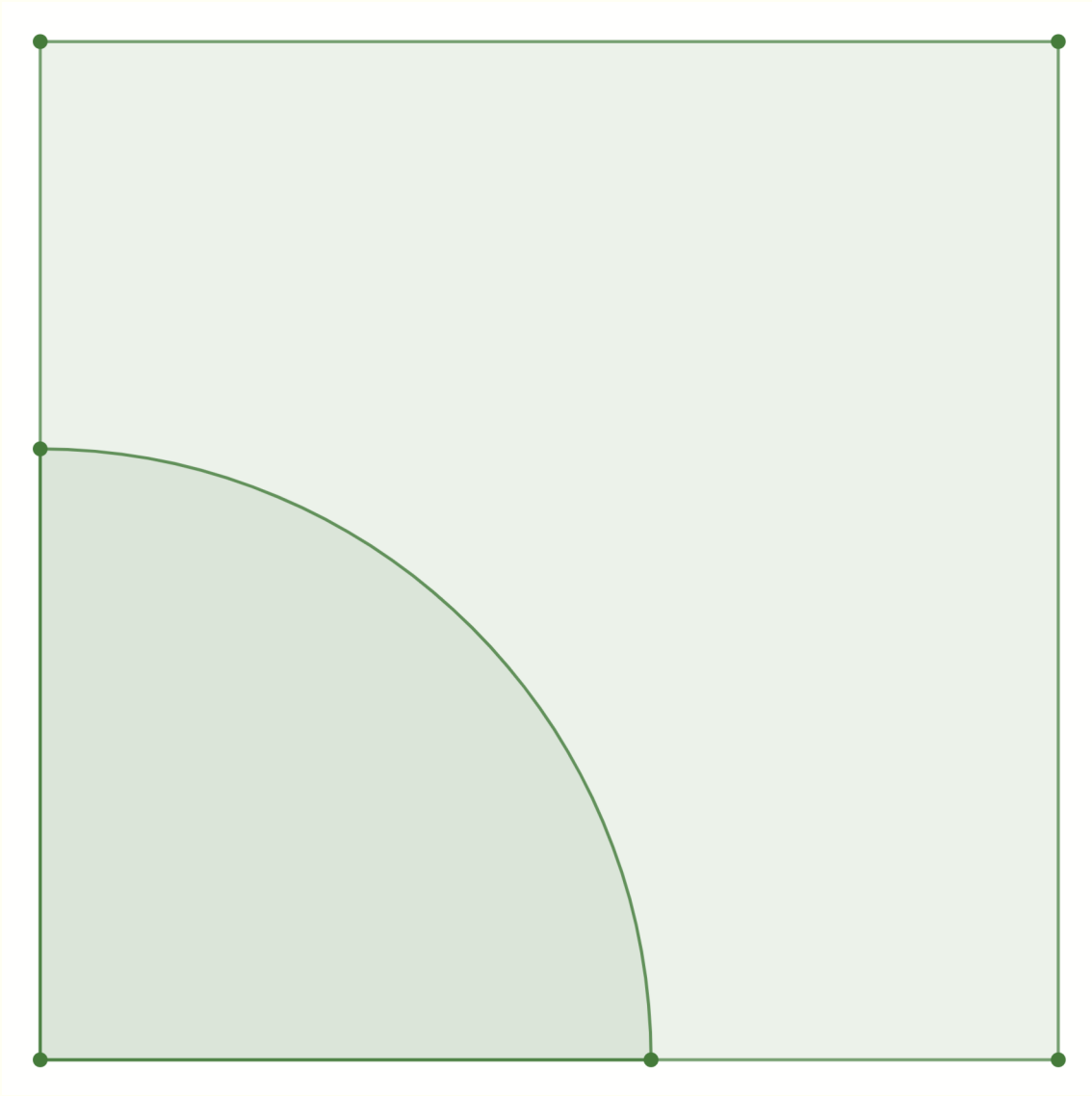
Problem 4: Manipulation

Let X, Y be random numbers chosen independently from the interval $[0, 1]$ with uniform distribution.

(a) Let $Z = X^2 + Y^2$. For $Z \leq 1$, find the cumulative distribution function and the density function of Z.

!

$$0 \leq Z \leq 1$$



!

$$Z > 1$$

Problem 4: Manipulation

Let X, Y be random numbers chosen independently from the interval $[0, 1]$ with uniform distribution.

(a) Let $Z = X^2 + Y^2$. For $Z \leq 1$, find the cumulative distribution function and the density function of Z.

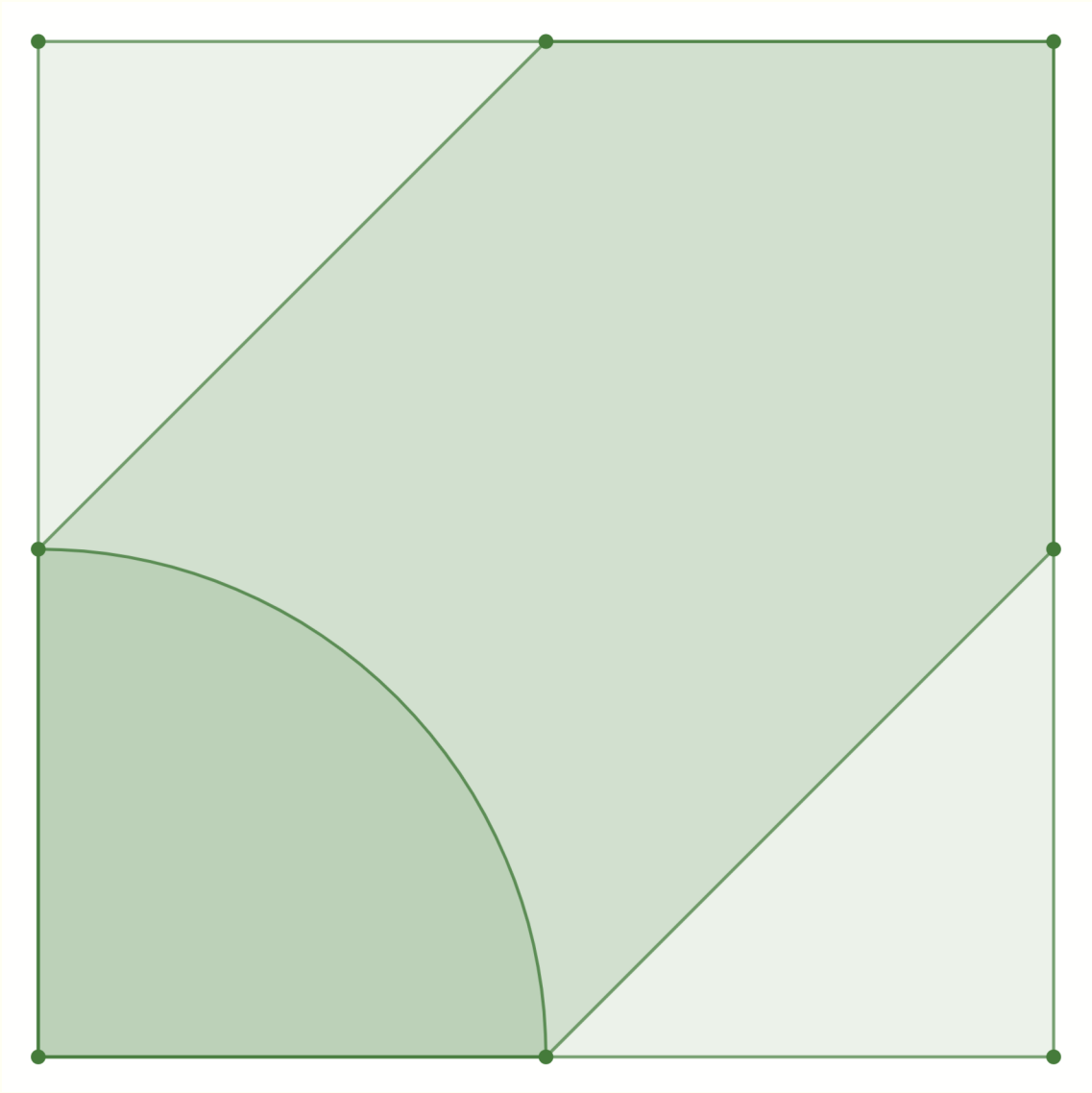
(b) Given that $0 \leq |X - Y| \leq \frac{1}{2}$, find the probability that $Z \leq \frac{1}{4}$.

!

$$0 \leq |X - Y| \leq \frac{1}{2}$$

!

$$0 \leq Z \leq \frac{1}{4}$$



5 pts

$$F_Z(z) = P(Z \le z) = P(X^2 + Y^2 \le z) = \frac{\pi z}{4}.$$

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{\pi}{4}.$$

5 pts

$$P(Z \le \frac{1}{4} | 0 \le |X - Y| \le \frac{1}{2}) = \frac{\frac{\pi}{16}}{\frac{3}{4}} = \frac{\pi}{12}.$$

Problem 5: National Committee of Senators

In the United States, a state is a constituent political entity, of which there are currently 50. Bound together in a political union, each state is represented in the Senate (irrespective of population size) by two senators.

Two senators from the same state

We consider the events that a national committee of 50 senators are chosen at random. Please find the probability, respectively, for:

(a) New Hampshire is represented.

!

$$P(E) = 1 - P(E^c)$$

5 pts

$$p = 1 - \frac{\binom{98}{50}}{\binom{100}{50}} = 1 - \frac{98!50!50!}{100!50!48!} = 1 - \frac{50 \times 49}{100 \times 99} = \frac{149}{198}.$$

Problem 6: Star Trek: Long and Prosper

Duck Musk's company **SpaceD** has sent out 10 starships to search for extraterrestrial life among a target set of n exoplanets. Due to a **'Swan'** program error, however, each of these 10 starships is drifting to one randomly chosen destination out of these n target exoplanets.

Starship officers

(a) Suppose $n = 10$. Derive the probability that these starships each will have landed in different exoplanets.

5 pts

$$p = \frac{10!}{10^{10}}.$$

Problem 6: Star Trek: Long and Prosper

(b) Our universe is vast. Please derive an approximation for the smallest number for n such that these starships each will have landed in different destinations with probability greater than 0.99.

Hint from science officer **Spock**: Stirling's formula $n! \sim n^n e^{-n}\sqrt{2\pi n}$ and Taylor expansion $\log(1+x) \sim x - \frac{x^2}{2}$.

Hint from science officer **Spock**: Stirling's formula $n! \sim n^n e^{-n}\sqrt{2\pi n}$ and Taylor expansion $\log(1+x) \sim x - \frac{x^2}{2}$.

5 pts

$$p = \frac{n!}{n^{10}(n-10)!} \approx \frac{n^n e^{-n}\sqrt{2\pi n}}{n^{10}(n-10)^{n-10}e^{-n+10}\sqrt{2\pi(n-10)}}$$
$$= (\frac{n}{n-10})^{n-10+\frac{1}{2}}e^{-10} = (1+\frac{10}{n-10})^{n-10+\frac{1}{2}}e^{-10}.$$

$$\ln(p) = (n-10+\frac{1}{2})\ln(1+\frac{10}{n-10})-10 \approx (n-10+\frac{1}{2})[\frac{10}{n-10}-\frac{50}{(n-10)^2}]-10$$
$$\approx (10-\frac{45}{n-10})-10 = -\frac{45}{n-10} > \ln(0.99).$$

That is, $n > -\frac{45}{\ln(0.99)} + 10$. The smallest n is 4488.

Problem 7: Role Playing Game (RPG)

The **Duckmouth** is a 2020 role-playing game developed and published by **Math 20: Probability** and is based on the book **Introducktion to Probability**. Players control protagonist **Duck D Random**, a mathematician who is looking for his missing daughter **Duckota C Random**.

Duck D Random

Duckota C Random

Problem 7: Role Playing Game (RPG)

The character has four core attributes: Intelligence (I), Wisdom (W), Charisma (C) and Strength (S). Before the game starts, attribute scores are determined randomly by distributing character points. The values for these four attributes satisfying the three conditions:

- the sum is fixed to 10 points,
- the value of any attribute is no less than 1 point,
- the value of any attribute is no greater than 4 points.

An example of the attribute sequence is given below.

I: 4 points

W: 3 points

C: 2 points

S: 1 points

10 pts

$$10 = 1+1+4+4 = 1+2+3+4 = 1+3+3+3 = 2+2+2+4 = 2+2+3+3.$$

The number of ways is

$$\binom{4}{2} + 4! + \binom{4}{1} + \binom{4}{1} + \binom{4}{2} = 44.$$

An interview question by video game companies

XC 2020

Problem 7: Role Playing Game (RPG)

The character has four core attributes: Intelligence (I), Wisdom (W), Charisma (C) and Strength (S). Before the game starts, attribute scores are determined randomly by distributing character points. The values for these four attributes satisfying the three conditions:

- the sum is fixed to 10 points,
- the value of any attribute is no less than 1 point,
- the value of any attribute is no greater than 4 points.

$1 \rightarrow 0$

$4 \rightarrow 5$

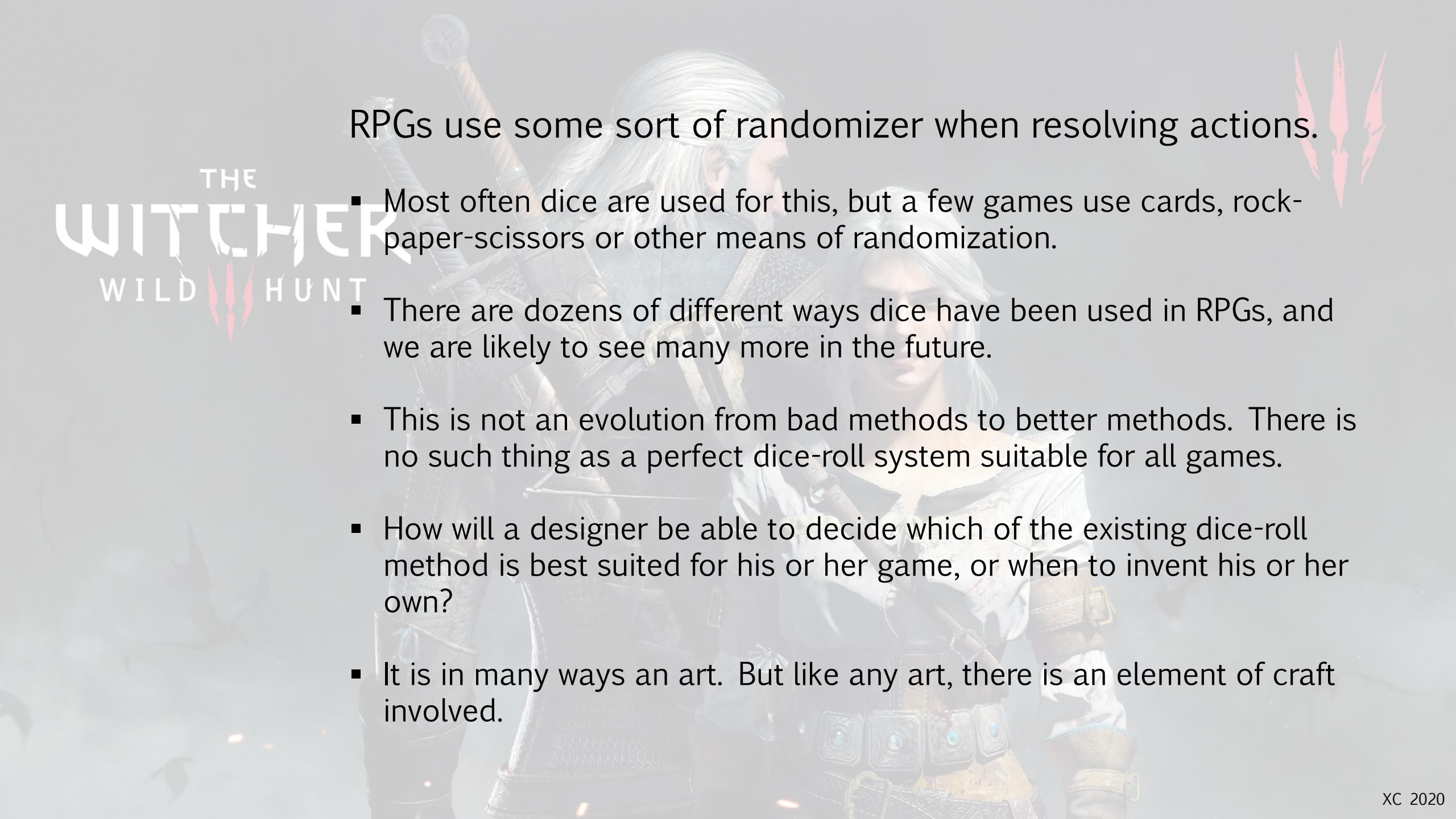
```python
def RPG_points(p_list = [1, 2, 3, 4]):
    '''
    p_list: pool of points
    '''
    attr_list = []

    for a in p_list:
        for b in p_list:
            for c in p_list:
                for d in p_list:
                    if (a + b + c + d == 10) and ((a, b, c, d) not in attr_list):
                        attr_list.append((a, b, c, d))

    print('Number of possible ways: ' + str(len(attr_list)))
```

```python
RPG_points()
RPG_points(p_list = [0, 1, 2, 3, 4, 5])
```

```
Number of possible ways: 44
Number of possible ways: 146
```

DICE ROLLS IN ROLE PLAYING GAMES

As a mathematician, you can find a job in a video game company!



RPGs use some sort of randomizer when resolving actions.

- Most often dice are used for this, but a few games use cards, rock-paper-scissors or other means of randomization.
- There are dozens of different ways dice have been used in RPGs, and we are likely to see many more in the future.
- This is not an evolution from bad methods to better methods. There is no such thing as a perfect dice-roll system suitable for all games.
- How will a designer be able to decide which of the existing dice-roll method is best suited for his or her game, or when to invent his or her own?
- It is in many ways an art. But like any art, there is an element of craft involved.

Problem 8: Role Playing Game (continued)

In the RPG game **Duckmouth**, you will also be assigned another character as your assistant **Quack Random** before the adventure (like **Holmes** and **Watson**).

There are 100 candidates in the backend, vary in rareness and power. Among these characters, 75 are Class R (rare), 20 are Class SR (super rare), and only 5 are Class SSR (specially super rare).

During the storyline, you need to send your assistant to undertake different missions. For R, SR and SSR characters, their chances of accomplishing a mission is shown in the table.

R	SR	SSR
		
25%	50%	100%

Now the server randomly drops a character as your assistant. In the first chapter of the game, your assistant has completed 4 missions in a row. What is the probability that he or she will accomplish the next mission?

Prior probability

The **prior probability** of an event (often simply called **the prior**) is its probability obtained from some prior information.

Posterior probability

The **posterior probability** represents the **updated prior probability** after taking into account some new piece of information.

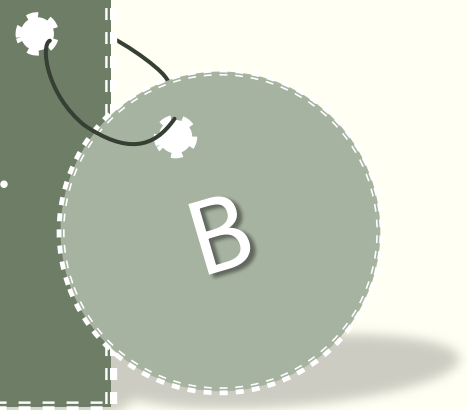
Evidence

The **evidence** term in Bayes' theorem refers to the **overall probability** of this new piece of information.

Likelihood

The **likelihood** represents a conditional probability. It is the degree to which the first event is consistent with the second event.

$$\text{Posterior probability} = \frac{\text{Prior probability} \times \text{Likelihood}}{\text{Evidence}}.$$



Prior probability

The **prior probability** of an event (often simply called **the prior**) is its probability obtained from some prior information.

$R,\ SR,$ or SSR

Posterior probability

The **posterior probability** represents the **updated prior probability** after taking into account some new piece of information.

R |  in a row

SR |  in a row or

SSR |  in a row

Evidence

The **evidence** term in Bayes' theorem refers to the **overall probability** of this new piece of information.

 in a row

Likelihood

The **likelihood** represents a conditional probability. It is the degree to which the first event is consistent with the second event.

 in a row | R

 in a row | SR or

 in a row | SSR

Bayes' formula

$$P(H_i|E) = \frac{P(H_i)P(E|H_i)}{P(E)}$$

$$P(E \cap H_i) = P(E|H_i)P(H_i).$$

=

$$P(E) = \sum_{i=1}^{m} P(E \cap H_i).$$

$$P(E) = \sum_{i=1}^{m} P(E|H_i)P(H_i).$$

Bayes' formula

$$P(H_i|E) = \frac{P(H_i)P(E|H_i)}{\sum_{i=1}^{m} P(E|H_i)P(H_i)}$$

Evidence

The **evidence** term in Bayes' theorem refers to the **overall probability** of this new piece of information.

⚔⚔⚔⚔⚔ in a row

Evidence

$$P(\text{⚔⚔⚔⚔⚔ in a row}) = P(\text{⚔⚔⚔⚔⚔ in a row} \mid R)P(R) +$$
$$P(\text{⚔⚔⚔⚔⚔ in a row} \mid SR)P(SR) +$$
$$P(\text{⚔⚔⚔⚔⚔ in a row} \mid SSR)P(SSR)$$

=

2 or 3 Steps

01

prior probability

R, SR, or SSR

posterior probability

R | ⚔⚔⚔⚔⚔ in a row

SR | ⚔⚔⚔⚔⚔ in a row or

SSR | ⚔⚔⚔⚔⚔ in a row

02

03

new probability

⚔ in the next mission

New probability

⚔ in the next mission

Posterior probability

R | ⚔⚔⚔⚔⚔⚔ in a row

SR | ⚔⚔⚔⚔⚔⚔ in a row or

SSR | ⚔⚔⚔⚔⚔⚔ in a row

New probability

$$P(\text{⚔ next}) = P(\text{⚔ next} \mid R)P(R) + P(\text{⚔ next} \mid SR)P(SR) + P(\text{⚔ next} \mid SSR)P(SSR)$$

=

10 pts

Let E and F be the events that the assistant completed 4 missions in a row and that he or she will complete the next mission. We have

$$P(E) = P(E|R)P(R)+P(E|SR)P(SR)+P(E|SSR)P(SSR) = (\frac{1}{4})^4\frac{75}{100}+(\frac{1}{2})^4\frac{20}{100}+(1)^4\frac{5}{100}.$$

And

$$P(R|E) = \frac{P(R)P(E|R)}{P(E)}, P(SR|E) = \frac{P(SR)P(E|SR)}{P(E)}, P(SSR|E) = \frac{P(SSR)P(E|SSR)}{P(E)}.$$

After some calculation, we get $P(R|E) = \frac{15}{335} = \frac{3}{67}$, $P(SR|E) = \frac{64}{335}$, $P(SSR|E) = \frac{256}{335}$. Therefore,

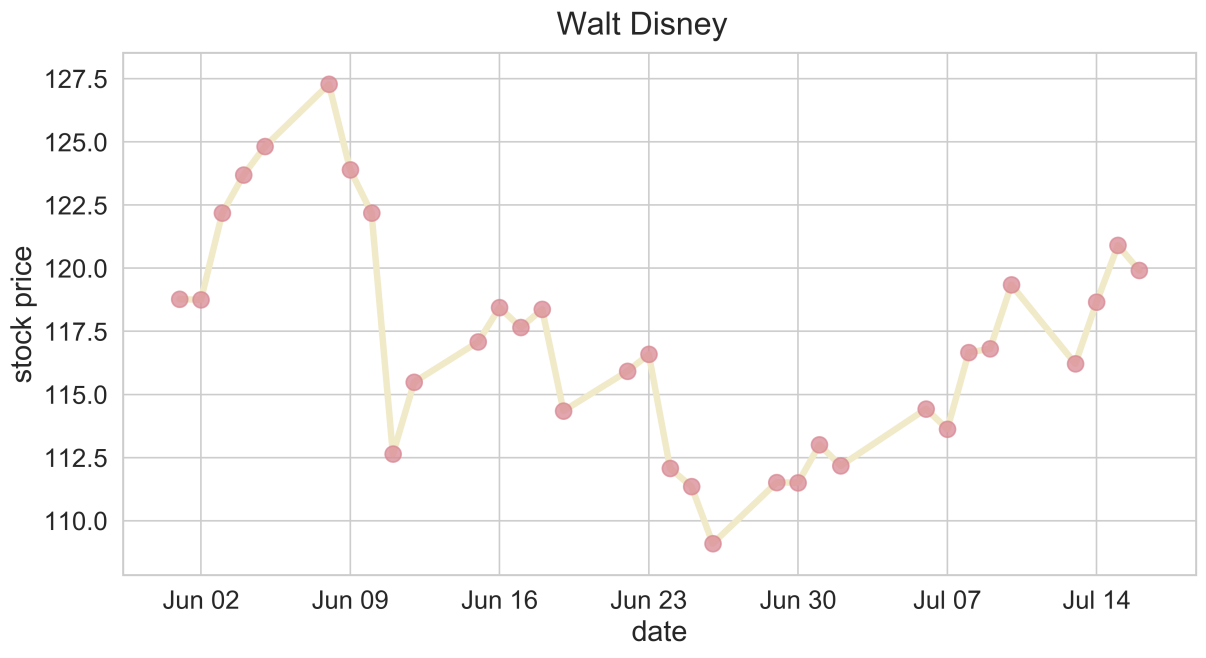
$$\begin{aligned} P(F) &= P(F|R)P(R) + P(F|SR)P(SR) + P(F|SSR)P(SSR) \\ &= \frac{1}{4}\times\frac{15}{335} + \frac{1}{2}\times\frac{64}{335} + 1\times\frac{256}{335} = \frac{1167}{1340}. \end{aligned}$$

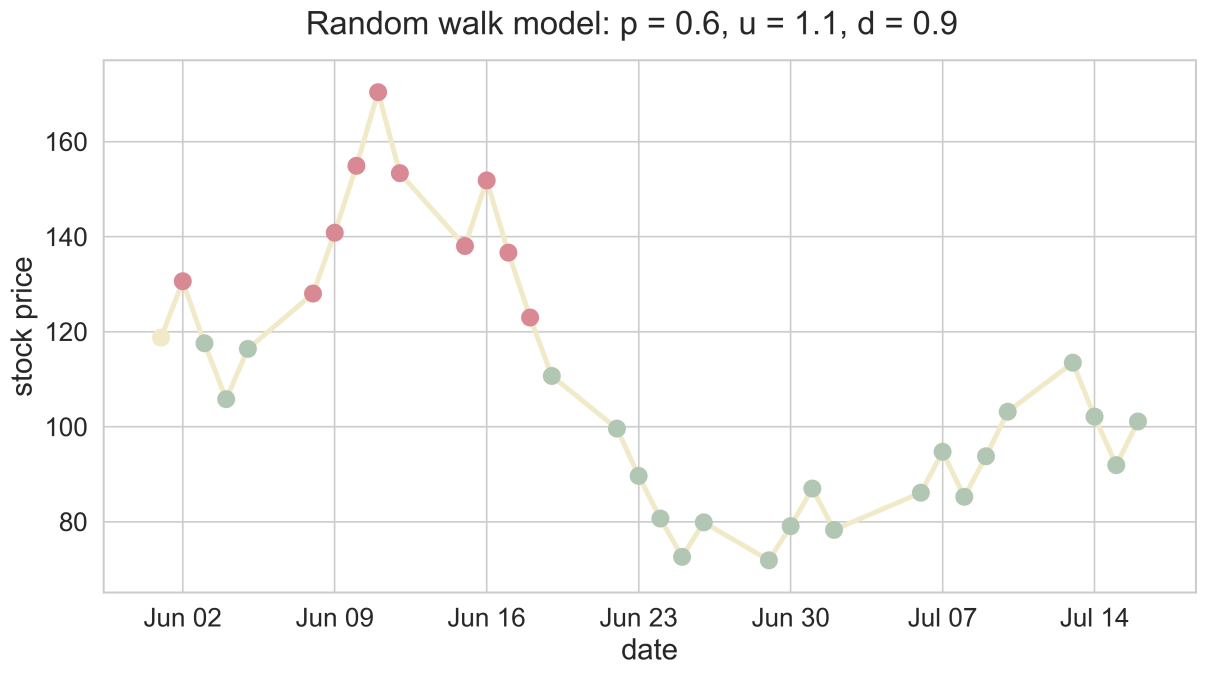
Problem 9: A Random Walk Down Wall Street

The Hedge Fund firm **Renaissance Ducknologies** is developing a stock price forecasting system.

The chief technology officer **Leonardo duck Vinci** would like to modify the original random walk model

$$S(t+1) = \begin{cases} uS(t), & \text{with probability } p \\ dS(t). & \text{with probability } 1-p \end{cases}$$



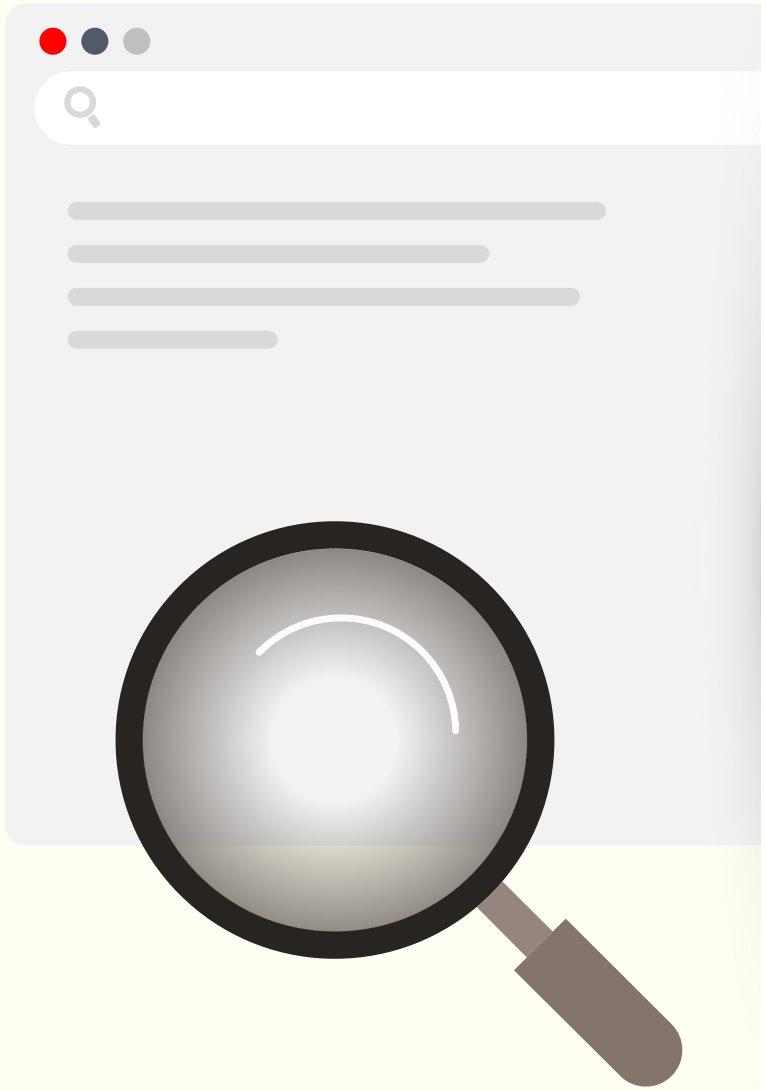


$$S(t+1) = \begin{cases} uS(t), & \text{with prob } p \\ dS(t), & \text{with prob } 1-p \end{cases}$$

Problem 9: A Random Walk Down Wall Street

Academics have not conclusively proved whether the stock market truly operates like a random walk or is based on predictable trends. There have been many published studies that support or undermine both sides of the issue.

As a consultant skilled in probability theory, you are asked to provide constructive suggestions to improve the model and hence the stock price may more closely resemble the simulations. Please present your (no more than one-page) proposal.

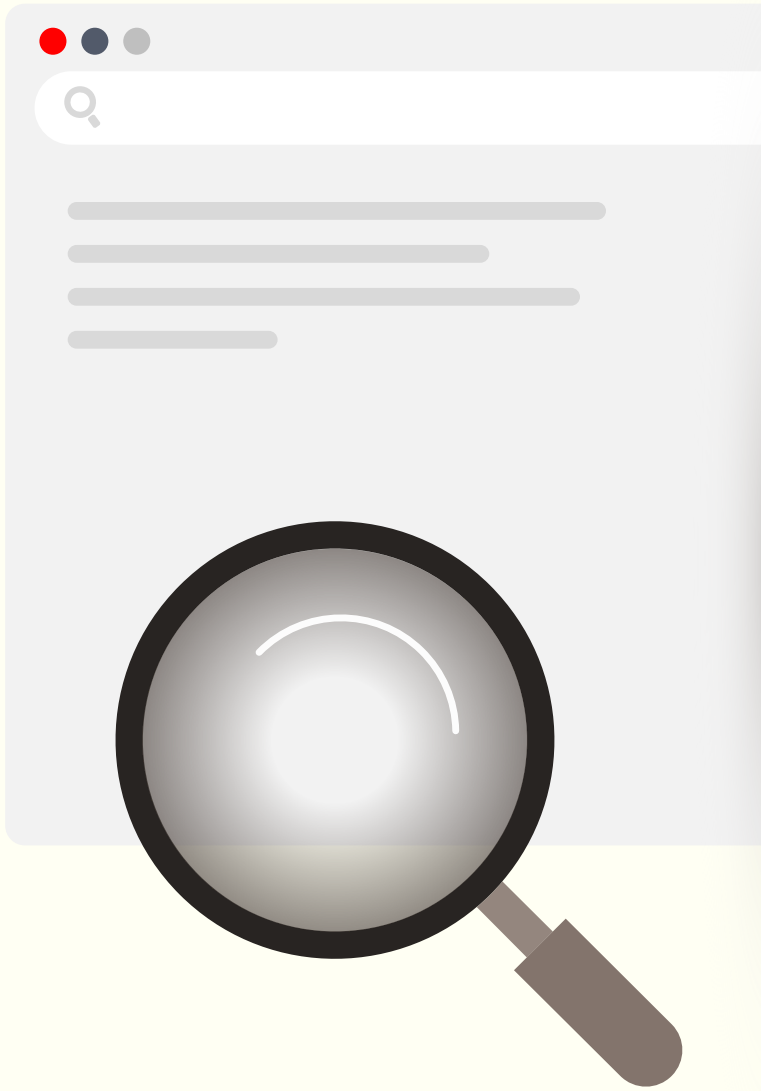


Jonathan Lee

Experiment with both smaller and longer periods of time.

Try incorporating machine learning to assist with pattern recognition.

Incorporate Bayesian probability. Stock market trends and people's decisions to buy/sell are based on peoples' beliefs.

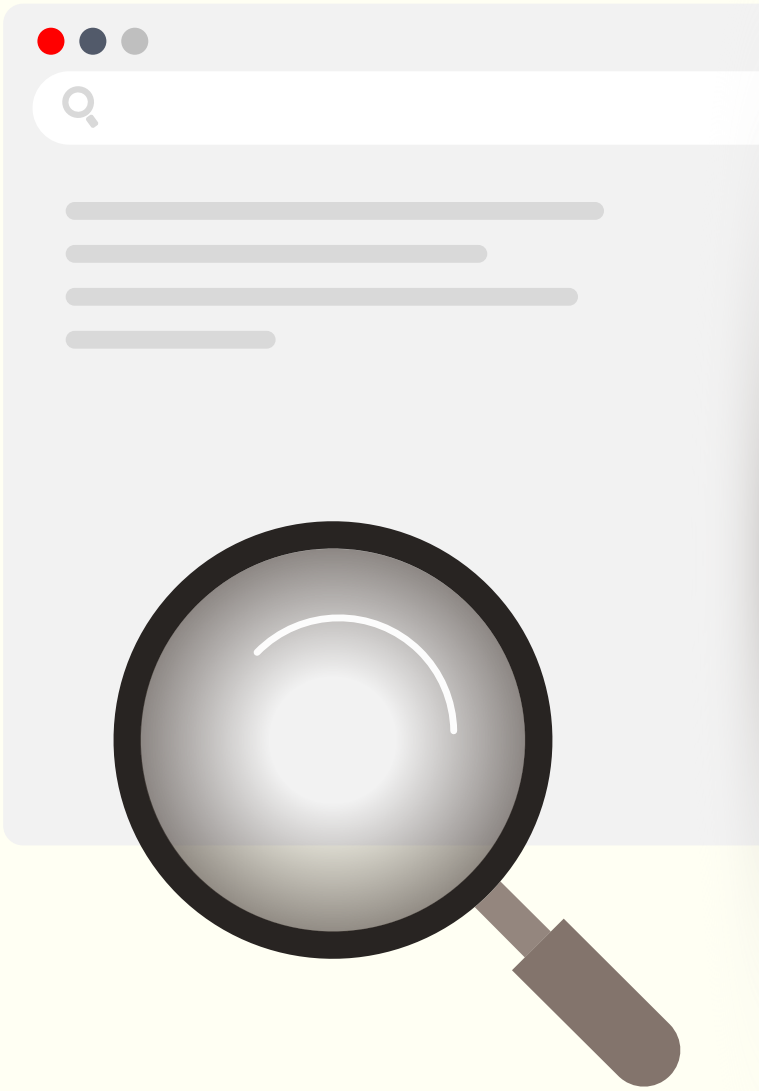


Samuel Baker

Switch the model from bigram to trigram.

Bayesian learning model and Bayesian regression.

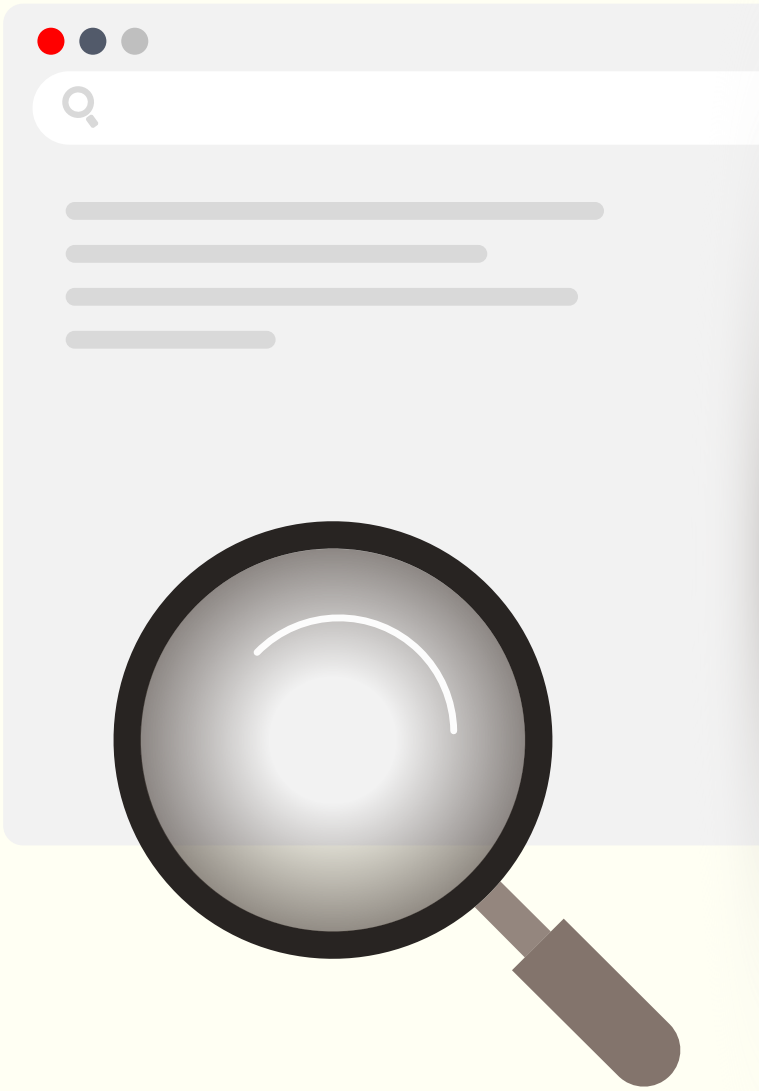
...



Gayeong Song

In order to better predict stock price, I recommend that the company does not use the Random Walk framework in the first place. Stock market prices, or price determined by market forces in any market, depends on factors exogenous to the stock market itself. Rational and profit-maximizing investors sell and buy stocks as a response to new information. Of course, there are some factors that appear random like the coronavirus. However, there are plenty of other exogeneous components that are non-random. One example would be the effect of the predictions on the central bank rate and the certainty of investors. If the Federal Reserve is expected to raise federal funds rates, investors may choose to divert money from long-term maturity bonds and allocate money into the equity market. Another and perhaps the most recent example would be from today's market: the 13% plunge of Moderna's stock prices after JPMorgan devalued Moderna's stock from "overweight" to "neutral."[1] Thus, there is a correlation between exogeneous factors, which are based on educated and calculated predictions of investors, and day-to-day stock prices. Although it is a colossal and controversial task to pinpoint an exact trend that the stock market follows, an elusive trend is not equivalent to no trend, the core argument of the Random Walk model.

In addition, while Random Walk model suggests that the market is completely efficient, there are historic instances of inefficiencies in the stock market, such as mispricing due to barriers to short selling and the discrepancy between earnings growth and price growth. In 1999, eToys' stock value was $8 billion, although the fiscal 1998 eToys' sales were only $30 million with a negative profit of $28.6 million. In comparison, Toys "R" Us's stock value was $6 billion with a sale of $11.2 billion dollars, which was 400 times larger than eToys. Eventually, eToys filed for bankruptcy in 2001. The market absurdly mispriced eToys.[2]



Gayeong Song

If Leonardo duck Vinci would like to keep a Random Walk model, then a compromise could be made by adding some more terms to the original model. It is true that pure chance events and shocks cannot be perfectly examined; however, adding terms for expectations for central bank rate adjustments, outlook on firm outcome like possible scientific discoveries, and firm credit ratings by trustworthy sources can help improve the model. These terms tend to follow a specific historical and rational trend of basic market demand and supply forces. In the current model,

$$S(t+1) = \begin{cases} uS(t), & u > 1 \textit{ with probability } p \\ dS(t), & 0 < d < 1 \textit{ with probability } 1-p \end{cases}$$

I suggest adding a sort of function to p, such as $p = f(A), A \in \{\textit{bank rate expectation}, \textit{firm outcome expectation}, \textit{policy change}, \dots\}$. This can help predictions based on how the market behaved previously for a certain change. To illustrate, Moderna stock's surge before JPMorgan's credit rating change that influenced today's plunge suggests that if a firm is nearing a desirable social outcome, investors can expect p to rise. Therefore, adding such terms would strengthen the current random walk model with no parameters about known human behavior.

---

[1] Flanagan, C. (2020, July 20). Moderna Sinks on Second Downgrade, Data from Rival Vaccines. *Bloomberg*. https://www.bloomberg.com/news/articles/2020-07-20/moderna-s-576-rally-puts-a-cautious-jpmorgan-on-sidelines.

[2] Shiller, R. J. (2005). *Irrational Exuberance*. New York: Broadway Books.

frontiers in Physics

ORIGINAL RESEARCH
published: 10 July 2019
doi: 10.3389/fphy.2019.00098

Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model

Tianyu Ray Li[1], *Anup S. Chamrajnagar*[1], *Xander R. Fong*[1], *Nicholas R. Rizik*[1] *and Feng Fu*[1,2]*

[1] *Department of Mathematics, Dartmouth College, Hanover, NH, United States,* [2] *Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH, United States*

In this paper, we analyze Twitter signals as a medium for user sentiment to predict the price fluctuations of a small-cap alternative cryptocurrency called *ZClassic*. We extracted tweets on an hourly basis for a period of 3.5 weeks, classifying each tweet as positive, neutral, or negative. We then compiled these tweets into an hourly sentiment index, creating an unweighted and weighted index, with the latter giving larger weight to retweets. These two indices, alongside the raw summations of positive, negative, and neutral sentiment were juxtaposed to $\sim$ 400 data points of hourly pricing data to train an Extreme Gradient Boosting Regression Tree Model. Price predictions produced from this model were compared to historical price data, with the resulting predictions having a 0.81 correlation with the testing data. Our model's predictive data yielded statistical significance at the $p < 0.0001$ level. Our model is the first academic proof of concept that social media platforms such as Twitter can serve as powerful social signals for predicting price movements in the highly speculative alternative cryptocurrency, or "alt-coin," market.

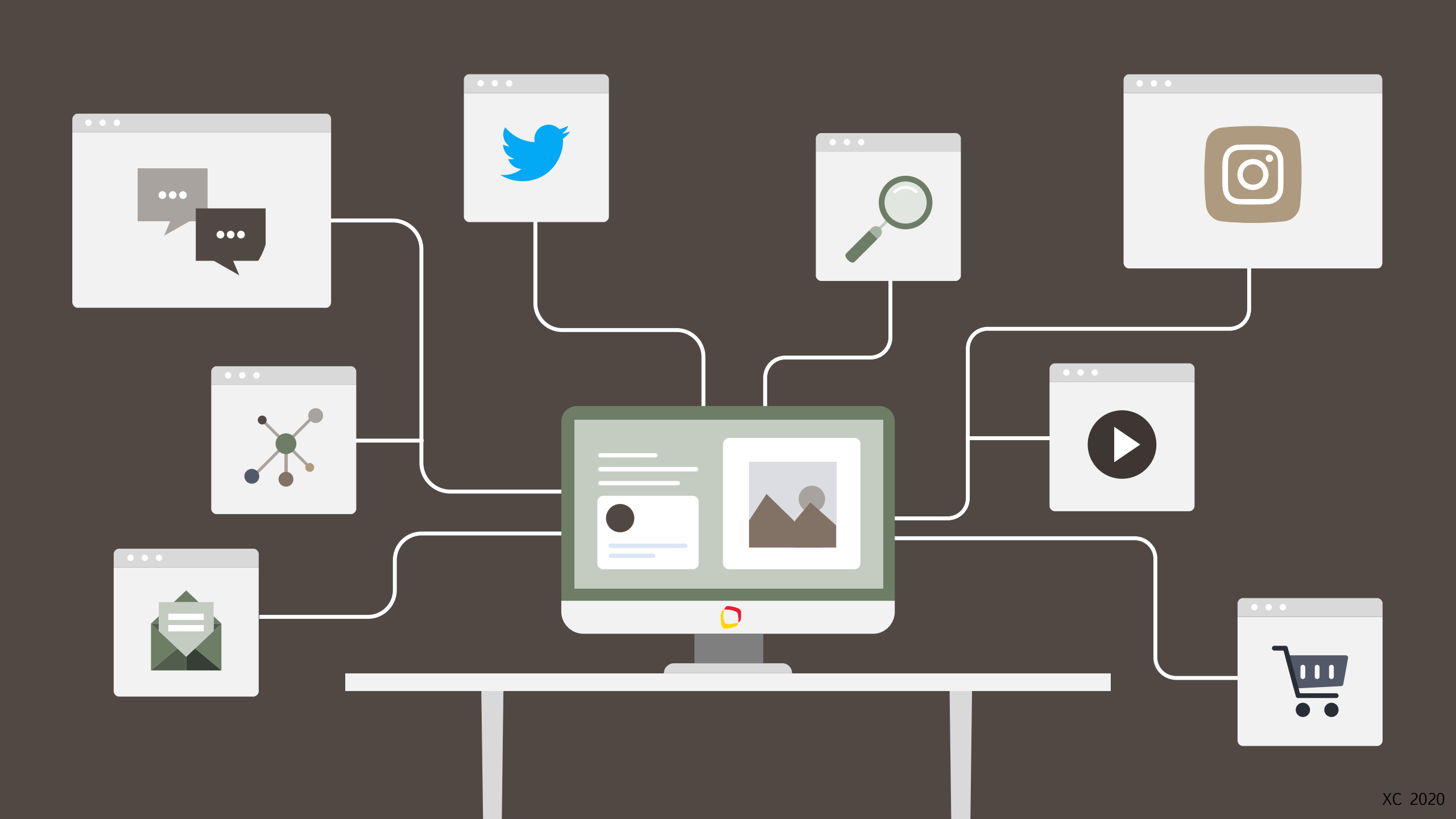
Keywords: data science, cryptocurrency, tree-model, Twitter sentiment, social dynamics, data integration and computational methods

OPEN ACCESS

Edited by:
Matjaž Perc,
University of Maribor, Slovenia

Reviewed by:
Chunyan Zhang,
Nankai University, China
Attila Szolnoki,
Hungarian Academy of Sciences (MTA), Hungary

***Correspondence:**
Feng Fu
feng.fu@dartmouth.edu



XC 2020