**Final Report**

**Natural Language Processing with Python**

Ban Qu

CDA571 – Project Guidance

Guided by: Dr. Cristian Tiu

SUNY at Buffalo

Fall 2021

**Introduction**

The integration and aggregation of technology has led to information explosion nowadays. Social media channels such as Facebook, Twitter, YouTube, along with forums, blogs, and articles are constantly generating large amount of text data every second. Due to the unstructured nature of data source and complex nature of human language, it is rather difficult and takes a lengthy process to extract valuable information via common approaches. Considering this, Natural Language Processing (NLP), or making natural human language readable by computer programs, serves as an indispensable means to process large volumes of text data at an unprecedented speed. Automation is a critical step to thoroughly analyze text and speech data, and thus understanding the texts and parts of speech. In this study, I am working with Dr. Cristian Tiu to analyze literature articles via NLP with Python in an attempt to count instances of innovation-related words. To be specific, the current study aims to use a text list of innovation-related words and count the instances of each word in any given text articles.

**Data**

The current study uses a list of 500 words as the base list containing all innovation-related words (see Appendix A). This list of words is matched against two pieces of scientific articles (see Appendix B, Appendix C) for different purposes. The first purpose is to find the exact, single matching words between two lists (i.e., the list of 500 words and the first article); and the second purpose is to find the phrase matching words between the two lists (i.e., the list of 500 words and the second article).

**Methods**

The current study uses Python 3 as the only programming tool for analysis. For the two purposes aforementioned, we install and use two main NLP packages in Python - NLTK and Spacy. The first step is to import the list of 500 words and scientific articles into Python. Second, we tokenize the list of 500 words and the article, in order to split up text by word and work with smaller pieces of text. The tokenized words in the article are then filtered against stop words (i.e., words do not add meaning to a text, e.g., ‘in’, ‘is’, ‘an’). After that, we further filter the words to exclude punctuations, special characters, and empty strings. In some cases, we also make lower cases of words in the article and lemmatize the words to reduce words to their core meaning. For the second purpose of the study, a function is used to define ngrams (i.e., the number of words in a word phrase) so that the list could be matched against word phrases rather than single words. In case of using Spacy package, the article is converted to string to be able to find matches. We use Spacy package to define patterns and find matches between lists for both purposes. In the end, the FreqDist method in NLTK packages is used to count word frequencies and instances of matches for both purposes.

**Analysis**

***Single-word matching***

# open the innovation list of 500 words (stored in python directory)

with open('innovation.txt') as f:

n\_words = f.read()

print(n\_words)

# import the NLTK and related packages

import nltk

from nltk.tokenize import sent\_tokenize, word\_tokenize

# tokenize innovation list by words

token\_n\_words = word\_tokenize(n\_words)

token\_n\_words

# open the first article file (stored in python directory)

with open('draft.txt') as f:

draft = f.read()

print(draft)

# tokenize draft article by words

token\_draft = word\_tokenize(draft)

token\_draft

## filter out stop words in draft article ##

# import stopwords package from nltk

from nltk.corpus import stopwords

# focus stopwords in English

stop\_words = set(stopwords.words("english"))

# creat an empty list to store the updated content

filtered\_draft = []

for word in token\_draft:

if word.casefold() not in stop\_words: # casefold() to ignore if words are uppercase or lowercase

filtered\_draft.append(word)

# print draft list without stop words

print(filtered\_draft)

# make filtered draft a string

str\_filtered\_draft = ' '.join([str(ele) for ele in filtered\_draft])

str\_filtered\_draft

# import and load the spacy package

import spacy

nlp = spacy.load('en')

#import phrased-based matching package

from spacy.matcher import PhraseMatcher

matcher = PhraseMatcher(nlp.vocab)

## define patterns and find matches ##

patterns = [nlp(text) for text in token\_n\_words] # make processed Doc for innovation list

matcher.add('TerminologyList', patterns) # add matcher word:'yes'

doc = nlp(str\_filtered\_draft)

matches = matcher(doc)

matched\_list = [doc[start:end] for match\_id, start, end in matches] # print the text string of all matched words

# print matched list

print(matched\_list)

# join all matched words in string, seperated by comma

values = ','.join(str(v) for v in matched\_list

print(values)

# tokenize all words in matched list

new\_list = word\_tokenize(values)

print(new\_list)

## further filtering the matched list ##

# remove punctuations

import string

new\_list = [''.join(c for c in s if c not in string.punctuation) for s in new\_list]

# remove special characters

special\_char = '’'

update\_list = [''.join(x for x in string if not x in special\_char) for string in new\_list]

# remove empty strings

update\_list = list(filter(None, update\_list))

# print updated list

print(update\_list)

## count instances ##

# count word frequency in list

from nltk import FreqDist

frequency\_distribution = FreqDist(update\_list)

print(frequency\_distribution)

***Phrase matching***

# import packages for nlp

import nltk

from nltk.tokenize import sent\_tokenize, word\_tokenize

from nltk.corpus import stopwords

from nltk.stem import WordNetLemmatizer

from nltk import ngrams

from nltk import FreqDist

from collections import Counter

# open the innovation list txt file (in python directory)

with open('innovation.txt') as f:

words500 = f.read()

print(words500)

# open the second article (in python directory)

with open('article.txt') as f:

draft = f.read()

print(draft)

# define a new stopwords list

stop\_words = stopwords.words('english')

# contain certain stopwords in list

# because they are in some phrases of the 500 words list

non\_stop = ['and','of','to','your','about']

new\_stop\_words = [word for word in stop\_words if word not in non\_stop]

print(new\_stop\_words)

# tokenize the article

new\_draft = word\_tokenize(draft)

# make all lower cases

new\_draft = [t.lower() for t in new\_draft]

# remove updated stopwords

new\_draft = [t for t in new\_draft if t not in new\_stop\_words]

# only contain alphabet

new\_draft = [t for t in new\_draft if t.isalpha()]

# lemmatizes the words

lemmatizer = WordNetLemmatizer()

new\_draft = [lemmatizer.lemmatize(t) for t in new\_draft]

# print the updated article

print(new\_draft)

# tokenize 500 words list

new\_words500 = word\_tokenize(words500)

print(new\_words500)

# define a function that returns ngrams

def get\_ngrams(text, n ):

n\_grams = ngrams(text, n)

return [ ' '.join(grams) for grams in n\_grams]

# store 4-word grams in 500 list

n4\_words500 = get\_ngrams(new\_words500, 4)

# store 4-word grams in article draft

n4\_new\_draft = get\_ngrams(new\_draft, 4)

# import and load the spaCy package

import spacy

nlp = spacy.load('en')

#import phrased-based matching package

from spacy.matcher import PhraseMatcher

matcher = PhraseMatcher(nlp.vocab)

# make article draft a string again

# to be able to perform spaCy package

str\_filtered\_draft = ' '.join([str(ele) for ele in n4\_new\_draft])

print(str\_filtered\_draft)

## define patterns and find matches ##

# make processed Doc for 500 words list

patterns = [nlp(text) for text in n4\_words500]

# add matcher word

matcher.add('TerminologyList', patterns)

# make processed Doc for article draft

doc = nlp(str\_filtered\_draft)

# make matches

matches = matcher(doc)

# create a new list to print the text string of all matched words

matched\_list = [doc[start:end] for match\_id, start, end in matches]

# count word frequency in list

frequency\_distribution = FreqDist(matched\_list)

print(frequency\_distribution)

**Results**

***Single-word matching***

The result shows that there are a total of 123 samples and therefore we count all 123 instances of the matched words with their frequencies.

# count all 123 instances

frequency\_distribution.most\_common(123)

In this case, we find all the matching words with frequencies in tuples. The results below are the top 5 words appeared in the article:

# find top 5 instances

frequency\_distribution.most\_common(123)[:10]

[('process', 24),

('innovation', 21),

('one', 17),

('two', 13),

('use', 12),

('idea', 12),

('breakthroughs', 10),

('new', 10),

('create', 9),

('question', 9)]

​

***Phrase matching***

The result shows only one match of one instance for the 4-word grams between the two lists,

which is "massachusetts institute of technology".

# count instances of all matches

frequency\_distribution.most\_common(1)

[(massachusetts institute of technology, 1)]

Appendices

**Appendix A**

creation

creativity

invention

competitiveness

efficiency

productivity

design

excogitation

technologies

globalization

technology

business

product

market

information technology

excellence

conception

sustainability

advancement

government

development

entrepreneurship

industry

initiatives

technological innovation

entrepreneurial

innovating

innovate

technological innovations

innovation

entrepreneurial spirit

intellectual

technological advancement

service

operational excellence

diversification

research and development

ideation

transportation

advancements

scientific breakthroughs

technological advancements

technological breakthroughs

creativeness

visionary

innovator

continual improvement

commercialization

environmental sustainability

technological advances

cleantech

entrepreneur

customer centric

transformational

key differentiator

collaboration

breakthroughs

pioneering

global

enterprise

expertise

technological developments

innovated

computing

transformation

biosciences

quality

automation

scientific discoveries

solutions

value proposition

innovation awards

environmental stewardship

reinvention

operational efficiency

innovation award

transformative

brightest minds

customer satisfaction

revolutionizing

innovations

nanotech

commoditization

innovative

technological

nanotechnologies

skilled workforce

entrepreneurs

personalization

stifle innovation

interactivity

domain expertise

initiation

performance measurement

origination

institution

introduction

foundation

economics

founding

instauration

procedure

consumer

society

inventions

demand

thinker

rethink

thinkable

communications

start

muse

beginning

commencement

contrivance

paternity

concoction

authorship

capitalism

contemplate

thoughtless

idea

ponder

cogitation

philosophize

think

brainstorm

bethink

nonthinking

transistor

metathinking

cogitate

cerebration

metathought

twithought

misthink

incogitancy

cogitative

ideational

mentality

consideration

contemplation

thoughtful

sustainable

bradyphrenia

ruminatingly

guesser

conceit

gmta

afterthought

thoughtworthy

excogitate

ponderingly

thoughtlet

promote

aims

underthink

mindlike

emphasis

creative

rumination

passthought

promotes

cerebrate

notion

ideate

eclecticism

promoting

focus

subconscious

ruminate

mind

consider

educational

marketing

guess

dianoetic

tool

surmise

integration

emphasizes

pensive

enhance

speculate

forthink

focuses

intellect

improving

transforming

creating

aim

improve

forethink

focusing

preoccupation

extemporaneous

metonym

strategy

collaborative

governance

forefront

intellection

awareness

create

integrating

modernization

daresay

opportunities

doublethink

strategies

brainwave

expanding

initiative

emphasizing

imagine

create idea

focused

dynamic

hospital

social

idealization

forethought

fostering

umbethink

supposition

improvement

engineering

offhand

education

concept

introspection

science

research

achieve

shaping

consolidation

skills

developing

figment

develop

management

learning

contribution

oriented

industrialization

economic

resource

suspect

importance

transform

netherthought

endeavors

project

introducing

opine

integrated

encourage

resources

boost

perspective

conceptional

projects

mindflow

mindset

aftermind

nous

aforethought

psychomotor

industrial economics

conceive

deem

cerebral

ideality

inconstancy

afterthink

speculator

your brain

brownfield

groupthink

citistat

understand problem

aphorism

suppose

free association

unguarded

brainchild

pothole

deliberation

factor endowment

think about

innovators

ai

make right decision

ideogeny

ratiocination

have idea

problem solve

idealism

decide thing

come to conclusion

puzzle over

comparative advantage

commerce

idealist

reason

have brain

joseph schumpeter

brain

computationalism

you get idea

opinion

thoughtfulness

find answer

reflection

innovation economics

train of think

speculation

creative thinking

brainfood

mull over

creative destruction

get headache

high cognitive process

think twice

chew over

reflect on

two cent

inventiveness

give think

ingenuity

consumer demand

analyse something

headwork

use your brain

stanford industrial park

silicon valley

bioscience

shockley semiconductor

differentiation

flight of fancy

growth

nobel laureate

startups

think piece

logical argument

think up

nanotechnology

william shockley

europe

think over

subvocalization

standardization

dynamism

emotionalist

fairchild semiconductor

be aware

visceral

read someone's mind

marketplace

startup company

reach conclusion

inventors

think of

interpretivism

silicon alley

two penny worth

technologists

rack one's brain

new york city

use your mind

light mind

ratiocinate

efficiencies

brainpan

tuppence worth

same mind

give clue

openness

manufacturing

cost-effectiveness

market share

remember phone number

originality

scratch your head

bring to mind

peter drucker

atlassian

come up with

disruptive innovation

patent

brain cell

mode of think

martin o’malley

visionaries

no brainer

food for think

city of baltimore

experimentation

grey matter

solve problem

energy

iron cage

regulation

chew cud

biomedicine

brain cloud

go someplace quiet

addle brain

answer question

entrepreneurialism

mass transit

close my eye

rockport

one track

socle

microelectronics

internationalization

philanthropy

usability

responsiveness

digitization

craftsmanship

biotechnology

evolution

informatics

convergence

profitability

competencies

dedication

products

revolutionary

vitality

prosperity

think balloon

conserver

pioneer

simplification

harnessing

real-time locating system

hybrid vehicle

mobile data terminal

think hard

oecd

electronic medical records

united states department of housing and urban development

determine truth

come to mind

automatic write

hope vi

stream of consciousness

remember past

out of hand

human life

do crossword puzzle

public housing

find quiet place

urban renewal

make decision

impulse purchase

seat of pant

think about something

harlem children’s zone

u.s. environmental protection agency

write thing down

brownfield regulation and development

australian

subject matter

mind body

pharmaceuticals

environmental protection

ibm pollyanna principle

open space reserve

innovativeness

community development

miniaturisation

intrapreneur

meritocracies

adaptiveness

differentiator

artisanship

internationalisation

seawell

nascency

intensiveness

miniaturization

photomicrography

vitalization

technologic

bossies

collaboratory

omnify

biomimetics

peter f. drucker

mental masturbation

linear model of innovation

open your mind

wishful think

make you tire

massachusetts institute of technology

write it down

**Appendix B**

On November 30, 2020, Moderna Therapeutics announced that Phase III clinical trials for its messenger RNA vaccine demonstrated 95% protective efficacy against the SARS-CoV-2 virus that had killed almost 1.5 million people worldwide in the previous 10 months. A relative upstart in the Covid-19 vaccine race and a company that few people had heard of before the pandemic, Moderna looked to be an overnight success. But as its CEO, Stéphane Bancel, has noted, that success was 10 years in the making. Far from a one-and-done stroke of luck, the vaccine was the product of a repeatable process that has been used countless times by the company from which Moderna emerged: Flagship Pioneering, a venture-creation firm based in Cambridge, Massachusetts, whose mission is to conceive, make, and commercialize breakthrough innovations in previously unexplored domains of the life sciences.

The misconception about the Moderna case, as with many other breakthrough innovations, is understandable. Breakthrough innovations are typically seen as the result of chaotic, random, and unmanageable efforts—the product of pure serendipity or the inspiration of a rare visionary. That view, we believe, is deeply flawed. From our different vantage points (Afeyan has spent the past three decades starting ventures based on breakthrough science and technology, and Pisano has studied innovation processes during the same period), we have come to realize that breakthroughs tend to emerge from a relatively well-defined process modeled on the basic principles that drive evolution in nature: variance generation, which creates a variety of life-forms, and selection pressure to select those that can best survive and reproduce in a given environment. The approach, called emergent discovery, is a structured and disciplined process of intellectual leaps, iterative search and experimentation, and selection. And while it relies on exceptionally talented people, it does not require the next Leonardo da Vinci or Steve Jobs to produce a breakthrough innovation.

Emergent discovery starts with prospecting for potentially important ideas in relatively novel scientific, technological, or market spaces with the goal of generating speculative conjectures, or “what if” questions. These serve as the starting point for an intensive Darwinian-style selection process to find and validate better ideas, soliciting critical feedback from outsiders to identify challenges and evolving the concept into a superior and practical solution. Emergent discovery requires a culture in which people, particularly leaders, in an organization are comfortable broaching seemingly infeasible ideas and challenging dogma—a culture that views “flawed” ideas not as dead ends but as building blocks and considers the evolution of ideas to be a collectively shared responsibility.

Defining Breakthrough Innovation

It’s important to define exactly what we mean by “breakthrough innovation.” We use two criteria. The first is discontinuity. Breakthroughs embody leaps in the principles of science, technology, design, economics, and other knowledge domains and establish new paradigms for future innovation by changing what is expected or considered possible. Honda’s light-jet design was a breakthrough because it was the first to use an engine-over-the-wing configuration, which, until Honda did it, was considered aerodynamically infeasible on a small aircraft. Not all breakthroughs are purely scientific or technological, of course. Google’s search engine was a technological breakthrough, but the company’s cost-per-click pricing method led to a business model innovation that totally upended the economics of the advertising industry.

The second criterion is value. Breakthroughs generate new sources of value by solving important problems or creating demand that did not exist previously. Digital cameras may have killed the film photography business, but today vastly more photos are taken digitally than were ever taken with film. Furthermore, considering that digital images have become an integral part of social media platforms like Facebook and Instagram, digital photography has created tremendous economic value.

The Shots-on-Goal Fallacy

In striving for breakthrough innovation, the predominant strategy today is the “shots on goal” approach—the antithesis of emergent discovery. It entails funding a large portfolio of projects in the hope that the profits from the rare success will more than pay for the cost of the numerous failures. If you invest in enough projects, the theory goes, by the laws of probability (sheer luck) you eventually will “score.” This strategy is common in the life sciences, the tech sector, consumer packaged goods, entertainment, and venture capital. A key element of the approach is strict reviews that can kill what appear to be weak projects quickly. At first glance, that all seems reasonable. Modern financial portfolio theory and practice highlight the benefits of diversifying risks, and a system of strict reviews to ensure that you don’t throw good money after bad would appear to be prudent resource management.

But the shots-on-goal approach ignores the fact that breakthrough concepts are usually riddled with flaws at the outset. Indeed, previous iterations of many celebrated breakthroughs initially looked like duds. We hail the iPhone as a game changer, but most of its predecessors—such as Apple’s Newton—failed. Crixivan represents a breakthrough AIDS drug, but its development program was nearly terminated when early clinical trials yielded disappointing results. Because the mantra of the shots-on-goal approach is to kill early and often, many promising ideas struggle to survive past the embryonic phase.

There was no “aha” moment when the mRNA breakthrough happened. The Moderna platform was built on a constellation of technologies, methods, and know-how that evolved over time.

Another drawback is that the pressure for early results can create a dysfunctional adversarial relationship between project teams and their funders. Funders are eager to see progress, while team members live with the specter of having their projects canceled (with potentially dire consequences for their jobs or reputations) if initial results are poor. This dynamic can lead to situations where teams are reluctant to share bad news with funders—or to share information with other project teams whom they view as competitors for scarce resources. It also means that teams have little incentive to conduct the kind of early experiments that might illuminate important flaws in their concepts.

A superior approach is the emergent discovery process used by Flagship, which is modeled on the basic principles of evolution—the generation and selection of genetic variants—which have proved to be potent engines of innovation in nature. Genetic variance is generated by mutation (random point changes in the DNA code) and recombination (rearrangements of fragments of DNA). Selection pressure refers to elements of the environment—such as competition for food—that affect whether a given trait (say, longer legs) is more or less conducive to survival. Research on innovation and case studies on industries as diverse as chemicals, pharmaceuticals, computers, automobiles, electronics, and aircraft indicate that mechanisms analogous to variance generation and selection pressure play a critical role in innovation. If properly designed and managed, these processes can be harnessed to create breakthroughs. Flagship has developed and applied the principles of emergent discovery to originate more than 100 life science companies over the past two decades. A good case in point is Moderna Therapeutics. (Disclosure: Pisano has consulted to and has a financial interest in Moderna and is on the boards of and has financial interests in two other Flagship-backed firms, Axcella Health and Generate Biomedicines).

A Product of Emergent Discovery: Moderna

Moderna had its origins long before the pandemic. In the spring of 2010, one of us (Afeyan) and MIT’s Robert Langer, a prolific inventor and chemical engineering professor, met to discuss some ideas that Harvard’s Derrick Rossi had been researching about using mRNA—molecules that ferry DNA’s instructions to a cell’s protein-making machinery—to reprogram a certain kind of cell (fibroblasts) to create stem cells that could then be manipulated into many other kinds of cells. Rossi’s research built upon previous work by the University of Pennsylvania’s Katalin Karikó and Drew Weissman, who had used chemically modified mRNA to reduce—but not eliminate—adverse innate immune reactions in animals. In discussions with Langer, Afeyan found the general approach intriguing, but not because of the potential to reprogram adult cells into embryonic-like stem cells. Instead, he wondered whether it might be possible to use mRNA to instruct cells to make drugs—an idea that had been around for decades but had yet to be turned into a reality.

On the basis of this and other discussions, Afeyan and Doug Cole, a managing partner at Flagship, launched a seven-month exploration inside Flagship Labs, the firm’s innovation foundry, to explore the question: “What if we could create engineered mRNA that, when introduced into patients, would turn their own cells into miniature factories that make any biotherapeutic drug we want?” No one had ever successfully engineered mRNA for use as a medicine or proved that it could be done. Afeyan and Cole discussed the idea’s feasibility with scientists from a variety of disciplines, ranging from molecular and cell biology to biological engineering and nanotechnology. They then hired two young researchers from the lab of Nobel laureate Jack Szostak, a pioneer in RNA biology, to tackle the question: “Could mRNA enable patients to make their own therapeutics?”

Exploration of this question generated dozens more puzzles. Previous in vitro lab studies had shown success in reducing innate immune responses to synthetic mRNA, but even after chemical modifications, when the mRNA was put into cells, it triggered immune reactions that were still too high to allow its use in animals or for repeat doses. The specific biochemical pathways responsible for the immune reactions had not been identified. The team wondered whether different chemical modifications would lead to less-intense innate immune responses. Questions related to stability also came up. mRNA molecules are intrinsically unstable and prone to degradation in the bloodstream. Previous research with other types of RNA had discovered chemical modifications that rendered them more stable. Could mRNA be modified in the same way? (No, it turned out. Unlike the other RNAs, mRNA molecules had to survive two processes—transcription and translation—and the modifications interfered with both.) What alternative modifications might work? Again, there was no animal data on this and other questions. No one, for instance, knew where the mRNA actually went after injection into an animal. No one knew whether the synthetic mRNA would resist degradation or if it did, whether you could get enough of it into cells to manufacture protein. Assuming you could deliver sufficient quantities to the cells to spur production, no one knew whether the proteins would correctly “fold” into the three-dimensional shapes necessary for proper functioning. And assuming that functional proteins could be made, it was far from certain whether therapeutically meaningful quantities could be produced. Not only had these questions never been researched but the tools to address them had not even been created.

Several months in, the team had many hard questions and few answers. Project team members at ProtoCo LS18, as the prototype company was called, believed that there would be enormous commercial value if they could answer those questions. And with so little previous research in the area, many of their advances would be patentable. In the fall of 2010, Flagship began filing patents covering new chemical modifications as well as therapeutic compositions of mRNA. In 2011, the initiative was renamed Moderna, and its scientists moved into a lab on First Street in Cambridge. The team spent the next six months injecting mice with various combinations of chemically modified mRNA. Not unexpectedly, many of the molecules didn’t survive transcription and translation. But a few did. Some of the mice started producing proteins that they wouldn’t otherwise have made—first in tiny amounts and then in larger ones. That was the first real proof of scientific feasibility.

The story of Moderna illuminates several salient aspects of the breakthrough process. First, breakthroughs emerge from the accumulation of numerous advances—some big and many small. There was no precise “aha” moment when the mRNA breakthrough happened. In fact, there was no single mRNA breakthrough: The Moderna mRNA platform was built on a constellation of technologies, methods, techniques, and know-how that evolved over time. For instance, the team realized early on that because the immune system saw the injected mRNA as foreign and hostile, it attacked the molecules and shut down production of the desired proteins. Solving this problem—which involved developing proprietary ways to package mRNA so that it could evade the immune system and deliver it to the right cells in the body—took years.

Second, breakthroughs do not require an initial focus on a specific problem or user need. Flagship’s research started with speculation around a very broad use case: Could mRNA be used as a new drug modality? But there was no particular disease to be cured or user need to be addressed. Although Moderna is best known these days for its Covid-19 vaccine, infectious-disease vaccines were not a major part of the company’s early thinking; nor were cancer therapies or other types of vaccines, which now constitute another major thrust of the company. The search for practical applications coevolved with the deeper understanding of the technology.

Finally, breakthroughs originate with highly speculative, even seemingly unreasonable conjectures. The question “What if messenger RNA could be a drug?” was purely hypothetical in 2010. (As late as the summer of 2020, many experts were skeptical that mRNA-based vaccines against Covid were feasible.) But that was the point. The sole purpose of the what-if question was to frame and aim the exploration. It did not have to be right to be successful. In fact, many initial conjectures about mRNA proved to be wrong, but other important insights were generated along the way. Although no one could predict where the exploration would lead, the process was neither random nor chaotic. The concept evolved and the solution emerged through a highly structured set of activities involving variance generation and selection pressure.

Let’s examine those two elements of the emergent discovery process in greater depth.

Variance Generation: Initiating a “What If?” Hypothesis

With innovation, variance generation is not spontaneous, as it is in the natural world. It must be instigated by people looking for new ways to do things or seeking new understanding. But most of the time, innovation teams limit their thinking to tweaks and refinements of ideas that are known to work. For instance, ideas for making cars more fuel-efficient and less damaging to the environment for many decades consisted of incremental improvements in engine design (adding turbochargers; using electronics to more precisely control combustion), making vehicles lighter, and adding devices such as catalytic converters. All these improvements were predicated on the internal combustion engine as a starting point. Breakthrough innovation requires contemplation of alternatives beyond current scientific, technological, design, or economic horizons. Not until lithium-ion batteries became adequately efficient—thanks to advances generated in the portable electronics business—did electrically powered vehicles become feasible. Such leaps do not come naturally. Indeed, cognitive biases, misaligned incentives, adherence to dogma, and other forces often hinder speculation. Processes are needed to help us overcome these barriers.

Flagship’s variance generation process, which is designed to create breakthroughs in previously unexplored domains, explicitly excludes areas of science in which other companies have already been founded or where prior research is extensive. Small interdisciplinary teams of Flagship’s scientists and senior leaders (all of whom have science backgrounds) are assigned to explore specific domains (say, the application of artificial intelligence to drug discovery). Because the exploration teams are working in areas where little previous scientific work has been done, they cannot follow the typical process of reading the literature, identifying the gaps, and then tackling those gaps. Instead, they start by asking a series of what-if questions derived through rigorous exploration of various strands of science.

For instance, we know that the human body contains or, through consumption of food, interacts with many different forms of life: animal cells, fungi, bacteria, plants, viruses, and other single-cell organisms. That fact might lead one to ask questions such as: What do all these life-forms do in our bodies, and how do they interact with one another? Is there molecular communication across these life-forms? Do bacteria in our bodies work with our own cells to carry out metabolic, immunologic, and even neural functions? (It turns out that the answer to all those questions is yes.) Those in turn lead to speculative conjectures: What if we could develop medicines that utilized these networks to improve our health? What-if questions like that were the foundation of Flagship’s Senda Biosciences, a venture focused on health care applications of intersystems biology.

Although they are speculative, good what-if questions are grounded in deep understanding of biological phenomena. (For instance, the human ecosystem contains many life-forms; messenger RNA plays a critical role in the production of proteins inside cells.) They are derived by rigorously probing what is known and unknown about specific biological systems. For example, starting in 2014 a Flagship team launched an exploration of ways to use human red blood cells as therapeutic agents. At that time, other scientists had been developing engineered T cells as a new modality to fight cancer. This led the team to ask, “What if we could produce engineered red blood cells (the most abundant cell type in the body) containing one or more therapeutically active proteins either inside or on their surface as a new type of medicine?” At the time, there was no data suggesting that such cells could be made or that they would be functional. The venture, ProtoCo LS24, later became Rubius Therapeutics.

The variance generation process also requires robust interdisciplinary collaboration. Flagship’s ventures might bring together, say, a chemical engineer, a computational biologist, a cell biologist, and an oncologist—a range of perspectives that is itself a source of variance generation. Research has established that well-functioning interdisciplinary teams expand the scope of exploration by combining previously disparate domains of knowledge.

Florian Sommet/Trunk Archive

Asking what-if questions is a technique as old as Aristotle, yet this seemingly simple creative tool is often quite hard to use in practice. In our experience, three things get in the way.

Mistake #1: The hypothesis must be quickly proved correct. Innovation teams are often under pressure to validate their hypotheses early on. But doing so creates an “intellectual gravitational pull”—a bias against venturing too far from your known knowledge base. The what-if question should be a purely speculative starting point—a conjecture that becomes the focus of intensive iterative experimentation, testing, reevaluation, and evolution. At Flagship, there is explicit acknowledgment that these hypotheses need not be true at the moment they are posed. Everyone assumes that they have flaws; after all, when working in uncharted territory, it is virtually impossible for your initial supposition to be 100% correct.Mistake #2: The what-if question must address a specific problem. Problem-driven discovery—which has been used very successfully by organizations such as the U.S. Defense Advanced Research Projects Agency (DARPA)—starts with a target problem to be solved (say, how to design a hypersonic aircraft capable of flying 20 times the speed of sound). DARPA’s track record demonstrates that this approach can be effective. Flagship’s own successful track record, however, suggests that a narrow focus on a specific problem is not necessary for breakthrough innovation—and that under some circumstances, not having one leads to more creativity. In the early phases of exploration, Flagship considers broad potential realms of applications or use cases rather than specific problems or markets. For example, the process of exploration for another of Flagship’s ventures, Generate Biomedicines, was not motivated by a desire to treat a particular disease; rather, it started as an investigation into whether artificial intelligence could be used to expand the arsenal of possible biologic medicines. This led to the development of a computational platform capable of generating completely novel biotherapeutic proteins.

Neither approach is better than the other, but each is suited to different types of organizations and strategies. DARPA, because of its institutional mission, confronts a very specific set of military problems it seeks to solve. Thus, its problem-driven approach fits its strategy. But organizations seeking breakthrough innovation in unexplored realms need much more freedom for initial inquiry and an expanded scope of exploration. In fact, by starting with hypotheses about solutions and problems, an exploration can toggle between the two to find novel matches.

Mistake #3: The hypothesis may be fuzzy and imprecise. Hypotheses should not be amorphous visions; they should be concrete assertions of how something might be done. Just because they are speculative doesn’t mean they can be vague or thin on specifics. This may seem counterintuitive. After all, why sweat the details at an early stage, when it is highly likely that a proposed solution won’t turn out to be correct? Details are important because they provide a focal point for subsequent inquiry, testing, and evolution. Without them, it is hard to know what questions to ask next and which experiments might be critical to run. Consider the difference between the following two hypotheticals: “What if we could create a car that is self-driving?” as opposed to “What if we could create a fully autonomous driving system using a 360-degree suite of lidar sensors; infrared and ultrasonic sensors; cameras mounted on the front, rear, and sides of the vehicle; an on-board computer capable of 30 trillion floating point operations per second; artificial intelligence; GPS accurate to within one meter; and real-time vehicle-to-vehicle telemetry?” It is hard to know how to react to the first question other than to say something (not particularly helpful) like “Wow, sounds cool.” The second, concrete proposal invites a host of questions such as “Is 30 trillion floating point operations per second enough?” and “What kind of vehicle telemetry is needed?” Obviously, at the very outset of an exploration, one is unlikely to know enough to formulate a very precise hypothesis, but getting to such hypotheses as quickly as possible should be the goal. Think about hypotheses (Flagship often seeds a space with more than one) as alternative destinations. If you are not clear about where you want to reach eventually, then it is hard to choose a direction—and impossible to know if you are making progress.

Selection Pressure: Getting to “It Turns Out That”

In nature, variation is just the first step in evolution. Selection pressure through competition for resources (say, food) shapes which genetic variations (longer beaks, for example) survive and which do not. Applying selection pressure in the realm of innovation results in relentlessly questioning and refining hypotheses. This can be done by a variety of means, including gathering and analyzing data, formal experimentation, and soliciting outside experts’ input and critiques. Flagship uses all those means. It presents its hypotheses to a broad network of scientists, knowing full well that many will be skeptical and recognizing that even the most skeptical among them (the one who says, “That idea will never work”) will have valuable insights to help evolve the ideas. Through these discussions, Flagship’s team members learn about prior science that may be relevant and about people who may have helpful experience.

If selection pressure is working, flaws in the initial hypotheses will surface. In some instances, the flaws may be so deep that they warrant abandoning or rethinking the basic concept. Early on, “killer experiments” are conducted to determine whether an idea faces an impenetrable roadblock. For instance, in the early days of Moderna’s research program, experiments were aimed at better understanding the immunogenic properties of mRNA and whether an immune response could be avoided—because if it couldn’t, the idea of using mRNA as a drug would be dead.

Unless the team discovers a fatal flaw in the core hypothesis, it continues to evolve the idea, asking, “What did we miss? What should we change? What’s the next experiment?”

In many cases, even a “failed” experiment provides a jumping-off point for further inquiry or development of alternative hypotheses. With each iteration, hypotheses are discarded, confirmed, or refined, and core ideas about what’s possible and useful evolve until an actionable invention is created. Flagship refers to this as the moment when the what-if question is transformed into an “it turns out that” statement.

A key element of selection pressure involves the integration of diverse concepts. Flagship will often conduct parallel efforts to explore a given issue (say, using artificial intelligence to discover novel drugs). The point of parallel efforts is not to foster internal competition that results in killing the “loser” (as is often done in larger enterprises); rather, it is to expand learning and to find paths forward. In some cases when there are two or more parallel efforts, each may have a piece of the puzzle but not the whole thing, and the two would be better off merged. For instance, Flagship launched two efforts in 2013 to explore whether there are strains of bacteria in our guts that can control immune cells—to either activate immune responses or suppress them. The two developed their own research approaches and generated proofs of concept that such bacteria were present and, if developed into monoclonal strains, could be potent immune modulators. The decision to combine the efforts and develop a common platform to discover and produce new oral medicines resulted in Flagship’s Evelo Biosciences.

This approach to experimentation is different from the one employed by many organizations, including venture capital firms and funding agencies. Experiments are commonly used as filtering tools in a shots-on-goal approach to innovation. In the emergent discovery process, experiments are tools of inquiry, designed to find paths forward. When an experiment fails to support a hypothesis, project team members are expected to search for the root causes in order to expand understanding. Unless the team discovers a fatal flaw in the core hypothesis, it continues to evolve the idea, asking, “What did we miss? What is an alternative approach? What should we change? What’s the next experiment?”

For instance, in the early days of the venture that became Axcella Health, the Flagship team focused on producing recombinant proteins for therapeutic benefit composed of amino acids commonly found in the human body. While this was theoretically possible, it turned out to be extraordinarily difficult to manufacture the proteins in the required quantities, at the necessary level of purity, and at a reasonable cost. But that roadblock led to another insight: Why not use carefully designed combinations of the amino acids themselves (which are readily available) as components of the drugs rather than trying to produce a protein that contained them? Further experiments validated this new idea and provided a development path even faster than anticipated under the original approach.

Emergent discovery may sound highly risky and expensive. But iterative processes—if properly designed and managed—can actually be quite efficient. The key is to make each iteration as cheap and fast as possible, rather than going all in on hypotheses. Flagship intentionally keeps the earliest phases of exploration and experimental testing as lean as possible. The goal is to establish feasibility of an idea with an investment of no more than $1 million to $2 million over six to 12 months. Only after this phase has demonstrated a reasonable path forward will a company be formed and larger capital investments made. The objective of this iterative process is to maximize the “learn-to-burn ratio”—that is, to generate the maximum insight from each dollar spent.

Fostering an Emergent Discovery Culture

A disciplined and well-defined process is only part of what it takes to practice emergent discovery. Equally important is having the right mindset, culture, and leadership behaviors. Here are three of the most critical:

Make it acceptable to broach the unreasonable. Almost by definition, breakthroughs in their embryonic stages defy existing theories, principles, and bounds of experience. As such, they should be considered leaps of faith. So to foster emergent discovery in your organization, you need to make it acceptable to consider the seemingly impossible. Early in the process, leaders and team members must be willing to suspend disbelief and to reserve judgment about whether a hypothesis is true or not. Common (and very reasonable) questions such as “Why do you believe that’s true?” and “How do you know that’s the right thing to do?” tend to shut down the process of inquiry. Instead, ask questions like “What experiment could you run to test that hypothesis?” and “If your hypothesis is correct, what are some possible use cases where we might create value?” The way leaders react to early hypotheses heavily influences whether the most creative ideas are snuffed out or have a chance to evolve into something impactful.Leverage your critics’ insights to make your ideas even better. Breakthrough innovations typically challenge prevailing dogma—the set of collectively held beliefs about what is possible and what is acceptable. Challenging dogma also means challenging the people (the “leading authorities”) who have built their reputations around its veracity. History tells us that people who challenge conventional wisdom are often subject to accusations of recklessness, incompetence, or worse.

Leaders must make it acceptable to defy dogma. Consider the common practice of engaging external experts to vet internally generated ideas or to perform due diligence on proposed investments. In principle, having such external input is a good idea. But too often, these experts become defenders of the conventional wisdom. A better approach is to use them to improve the ideas—by identifying a critical assumption that should be tested, for example. If we engage skeptics and can tolerate their sometimes scalding critiques, we can learn a lot about what we need to do to move our ideas forward.

Make it about ideas, not personal ownership. Emergent discovery explicitly recognizes that ideas are built over time with contributions from many people. One person’s ill-formed idea last month might be the essential building block for someone else’s advance this month. The two are equally important to the process. Pursuing emergent discovery in your organization requires a culture where ideas are not “owned” by individuals but are considered part of the intellectual commons of the enterprise. Disconnecting ideas from people also means that a failed idea is not a personal failure. Accordingly, emergent discovery works better if the teams involved in an effort have shared incentives and rewards.

Leading Emergent Discovery

The notion that breakthrough innovation is a random, chaotic process largely dependent on the visionary powers of gifted geniuses makes many organizations hesitant to embrace it as a core element of strategy. That is unfortunate given the massive value that breakthroughs produce for society and the companies that create them. But there is nothing mysterious or magic about the process. Breakthrough innovation can emerge through a rigorous and disciplined process of intellectual leaps, iterative search, experimentation, and selection. Emergent discovery is a repeatable process that can be learned.

Mastering it, however, requires more than understanding the mechanics of the process. It requires an organization in which the people—particularly the leaders—adopt the right mindset and behaviors. They must be willing to consider seemingly unreasonable ideas and suspend judgment early in the discovery process. They must embrace learning through rigorous experimentation and failure and prioritize collective contributions over the personal ownership of ideas.

Ultimately, whether an organization adopts these habits depends critically on the behaviors of its leaders. Pursuing breakthrough innovation is as much a leadership challenge as it is a technical one. If the Covid-19 catastrophe has taught us anything, it is that the world can change dramatically in short order. Looking ahead, all companies must build the capacity to leap beyond existing comfort zones. Now, more than ever, we need leaders who can drive breakthrough innovation.

**Appendix C**

Today, on National STEAM Day, The Village School showcases projects from across grade levels as part of the MIT STEAM Challenge. As one of the only schools in Houston to have this unique partnership with MIT and to offer a program like this, students in grades PreK – 12 are offered the opportunity to work collaboratively with their teachers and peers to create projects focused on extreme weather, something Houstonians know all too well.

The MIT Challenge is a partnership between the Massachusetts Institute of Technology (MIT) and Nord Anglia Education, which serves as the parent organization of The Village School. The goal of the challenge is to enhance the teaching and learning of Science, Technology, Engineering, Arts, and Mathematics (STEAM) by connecting MIT innovation and culture to Nord Anglia schools globally through project-based challenges.

"The STEAM curriculum is integrated into everything we do here at The Village School," said Bill Delbrugge, head of school at Village. "When students are immersed in STEAM education, there are no silos. Teachers take an integrated approach to the curriculum, enabling students, regardless of their affinities for math, science or fine arts, to come together and solve a relevant problem. Sharing STEAM and this cutting-edge collaboration with a renowned institution such as MIT is a truly stimulating experience for us at The Village School."

Students were asked to focus their projects on extreme weather to help answer why, in recent years, we are experiencing an increase in hurricanes, floods, droughts, cyclones, wildfires and other types of severe weather events.

MIT posed the challenge to students at The Village School, citing climate change and warming ocean temperatures for wreaking havoc on Earth. Scientists at MIT are determined to help find ways to save our planet and are researching past and present weather events to explore solutions for changing weather patterns.

"I have really enjoyed working with the teachers and students this year on the extreme weather challenge," said Angel Bradford, Director of Innovative Teaching and Learning and the MIT/Global Campus Lead at Village. "I am so impressed by the creativity and dedication of our students. The cross-functional approach used in the MIT challenge exemplifies how we can help students adapt their skills and make connections across different fields in an ever-changing world. But more importantly, it illustrates the need to properly equip our students for the challenges and opportunities of the future."

Each grade level (PreK-12) across different subjects has been assigned various projects, each requiring a mix of science, technology, engineering, arts and math skills. From first graders creating weather newscasts to older kids using robots to code routes to school during flooding and even making hurricane proof shelters and testing them using various wind speeds, these challenges are a great way to get students interested in the world of STEAM and see how what they design can influence the world of tomorrow.

One of the most ambitious projects came from Village's "Our Engineer Your World" class who built a wind tunnel with the goal of achieving hurricane speeds in a shipping container to test out student projects made in other classes. Students and teachers collaborated with a community expert on the design and were able to apply mathematical and engineering principles to a real-life scenario on a large scale. The project also provides a future learning space for science classrooms from across the elementary to high school divisions as a place to test wind, projectile, and other physics experiments.

The Village School has been participating in MIT challenges each year since 2016. The MIT Challenges provide a unique chance for every teacher and student to experience MIT. Each challenge embodies the teaching and learning culture of MIT, is rooted in the research of MIT faculty and makes that research relevant and accessible to participating students.

"Our mission is to create and share high quality resources to facilitate digital and non-digital learning for PreK-12 and lifelong learners," adds Delbrugge. "By providing STEAM-based instructional materials and an open forum for users to share insights, we aim to inspire a diverse global community of educators, students, and parents to find innovative and humanistic solutions to real-world challenges."