Advanced Data Science

Dr. Hamid Mostofi
mostofidarbani@tu-berlin.de

Dr. Hamid Mostofi

Dr. Phill. - Dipl.-Ing – MBA.

Project Manager / Senior Researcher

at TU Berlin

Team member of GECI in German federal ministry for environment

Lecturer of MBA of sustainable mobility management at TU Berlin : Mobility Data Mining

Macroeconomics and business models

Lecturer of System dynamics modeling at Tu Berlin and HTW Berlin

Projects

Dr. Hamid Mostofi

Dr. Phill. - Dipl.-Ing – MBA.

eHaul project: Project Manager: Electrification of long haul heavy-duty commercial vehicles with automated battery swapping stations, founded by **BMWK** (Federal Ministry of Economic Affairs and Climate Actions)

SUMIC 2020 Project Manager: Academic cooperation for developing smart urban mobility considering climate change mitigation solution, founded by **BMBF** (Federal Ministry of Education and Research

GECI project - Developing the academic curricula in the field of renewable energies, smart mobility- founded by **BMU** (Federal Ministry of Environment) "

System dynamic modelling of sustainable urban transportation in

New York, Cairo

Mahnaz Namazizavareh

Dipl.-Ing

Mahnaz is a senior data scientist at OLX group.

She has a hands-on experience applying several ML an statistical algorithms to real-world problems.

Her specialities are Naturally Language Processing, Time-series Analysis, Recommender systems.

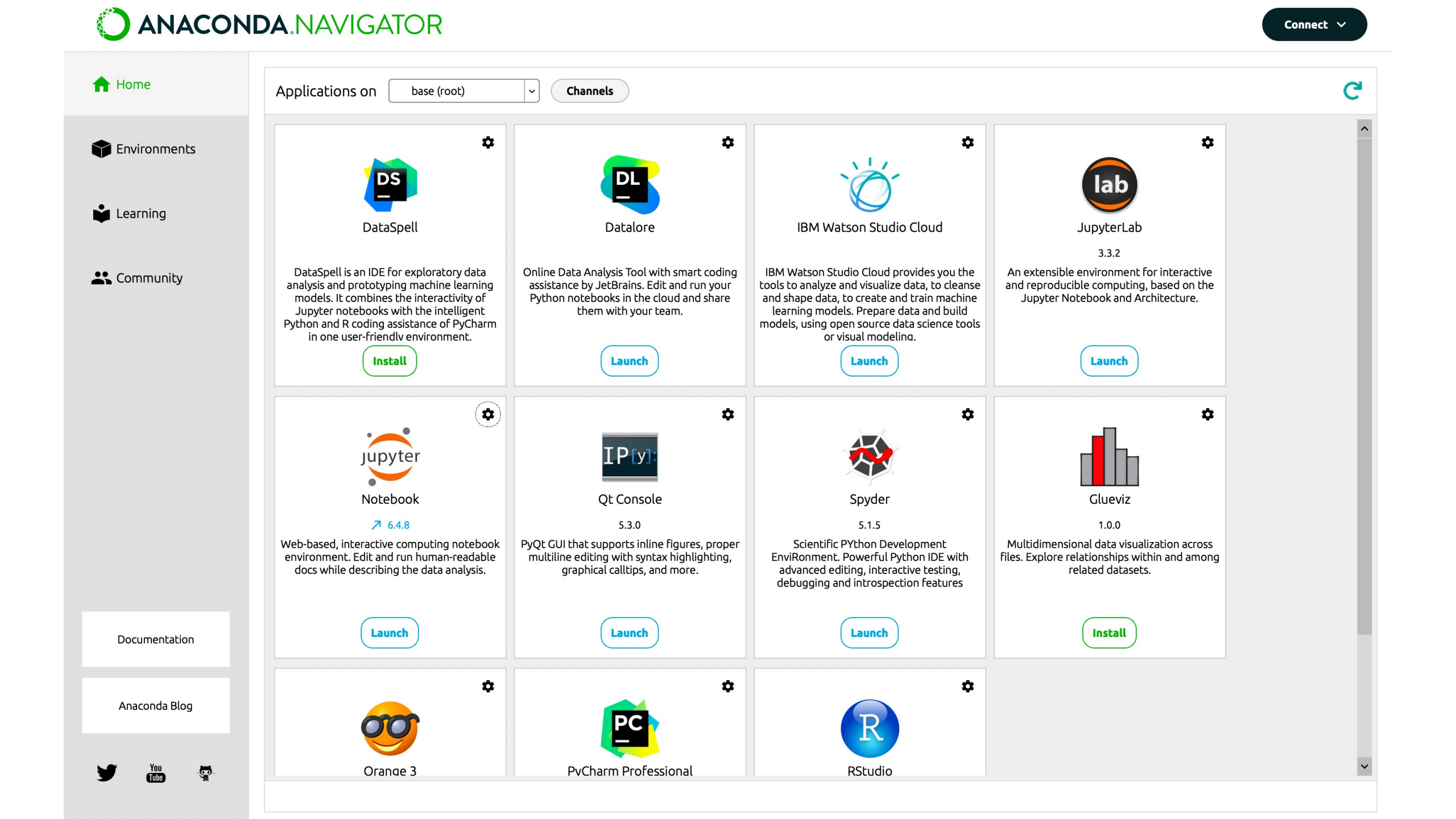
Date	Session Information	Details
Wednesday, June 1st	Course introduction - Data Science principles Descriptive techniques Dr. Hamid Mostofi Time CEST: 10:30 -14 :30 Format: Lecture	• Introduction of the course-agenda, • Review the principles of the data science • Causality vs. Correlations • Causal loop diagrams and Confounding effects • Review the principles of descriptive techniques
Thursday, June 2nd	Statistical inference Linear and Nonlinear Regression Dr. Hamid Mostofi Time CEST: 9:00 -13 :30 Format: Lecture	• Statistical inference, and their applications in the business context, social science and marketing • Review of linear Regression and its applications • Binary logistic regression and its application

Tuesday, June 7th	Nonlinear Regression, Structural Equation Modeling SEM Dr. Hamid Mostofi Time CEST: 9:00 -14 :00	• Ordinal logistic regression and its application • Multinomial logistic regression and its application • Observed and latent variables • Principles of Confirmatory factor analysis (CFA) • Constructs and measurements • Path analysis
Wednesday, June 8th	Time series Mahnaz Namazizavareh Dr. Hamid Mostofi Time CEST: 15:00 - 18:00	• Principles of Time series • Stationarity and differencing of time series data • Autoregressive integrated moving average ARIMA
Thursday, June 9th	Advanced Visualization Techniques for Data Analysis Mahnaz Namazizavareh Time CEST: 15:00 - 18:00	• Advanced visualization techniques as an initial step to solve data analysis problems. • Geo-based visualization • Text visualization • Network visualization

Date	Session Information	Details
Monday, June 13th	Clustering Mahnaz Namazizavareh Dr. Hamid Mostofi Time CEST: 15:00 - 18:00	Overview of Machine Learning (ML) process, supervised vs unsupervised, validation approaches, over/under fitting Introduction to basic Clustering approaches such as K-means in the Social Science context and in real business cases.
Tuesday, June 14th	Classification Mahnaz Namazizavareh Dr. Hamid Mostofi Time CEST: 14:00 - 18:00	• Introduction to basic Classification approaches • Decision Tree • KNN (k-nearest neighbors algorithm) • Text Classification
Wednesday, June 15th	Social Network Analysis Mahnaz Namazizavareh Dr. Hamid Mostofi Time CEST: 15:00 - 18:00	Introduction to Social Network concept and its principles and applications. Techniques of Community Detection such as: • Connected Components • Modularity

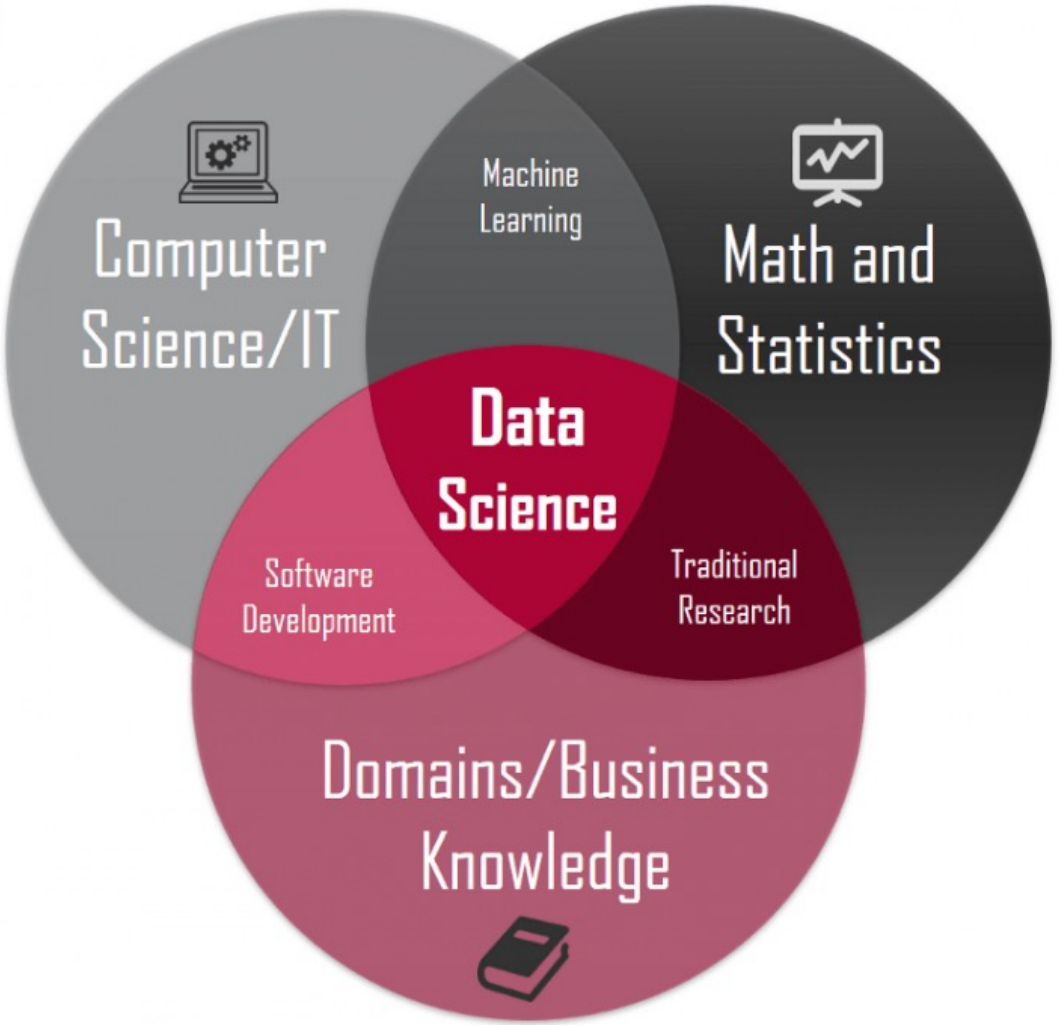
https://github.com/Advanced-Data-Science-TU-Berlin

The easiest way to install Python and its packages like pandas is to install it as part of the Anaconda distribution, a cross platform distribution for data analysis and scientific computing.


ANACONDA.NAVIGATOR
Connect
Home
Environments
Learning
Community
Documentation
Anaconda Blog
Applications on
base (root)
Channels
DataSpell
DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one user-friendly environment.
Install
Datalore
Online Data Analysis Tool with smart coding assistance by JetBrains. Edit and run your Python notebooks in the cloud and share them with your team.
Launch
IBM Watson Studio Cloud
IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling.
Launch
JupyterLab
3.3.2
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch
Notebook
6.4.8
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch
Qt Console
5.3.0
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch
Spyder
5.1.5
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
Launch
Glueviz
1.0.0
Multidimensional data visualization across files. Explore relationships within and among related datasets.
Install
Orange 3
PyCharm Professional
RStudio

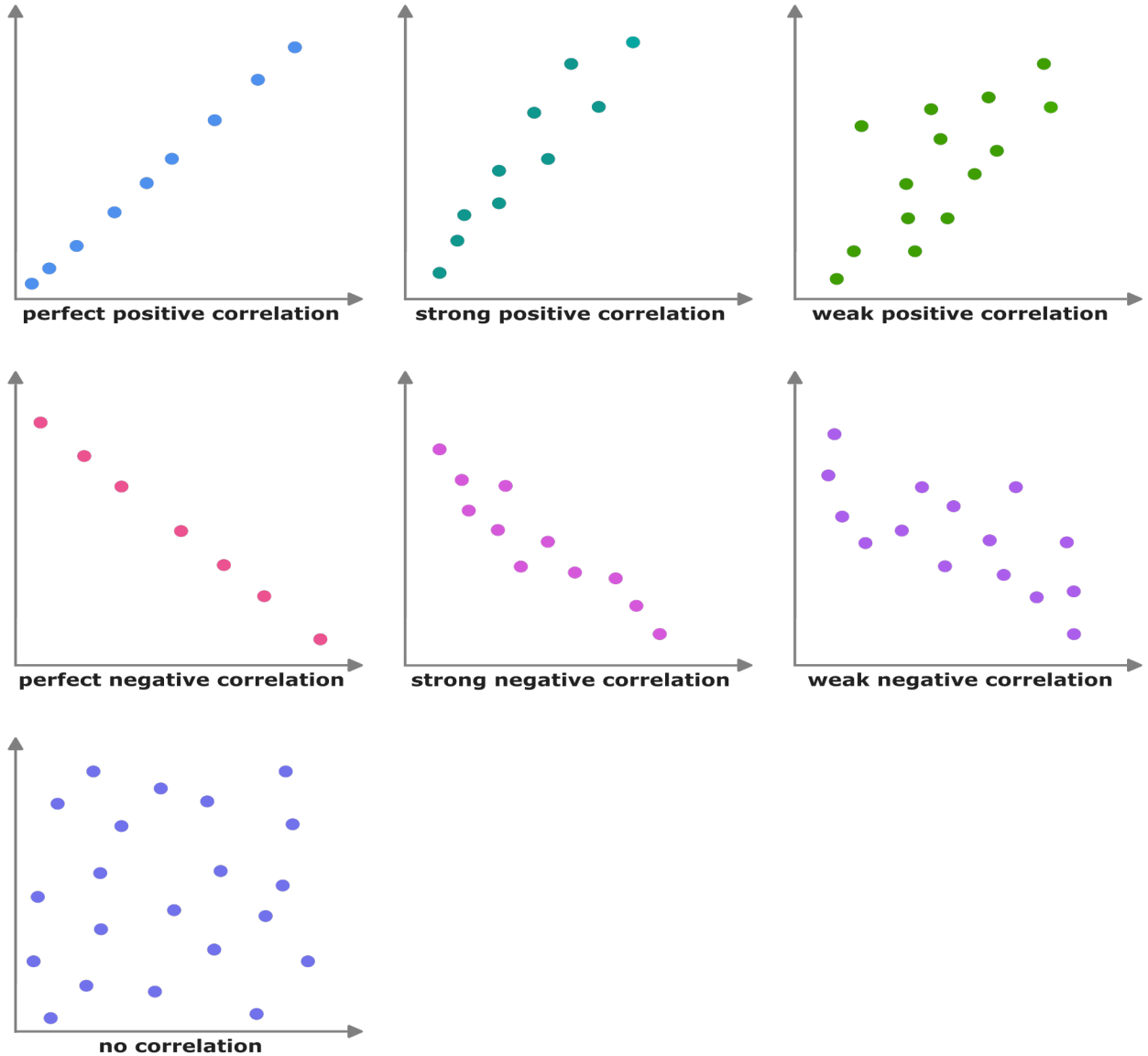
Vensim®

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains.

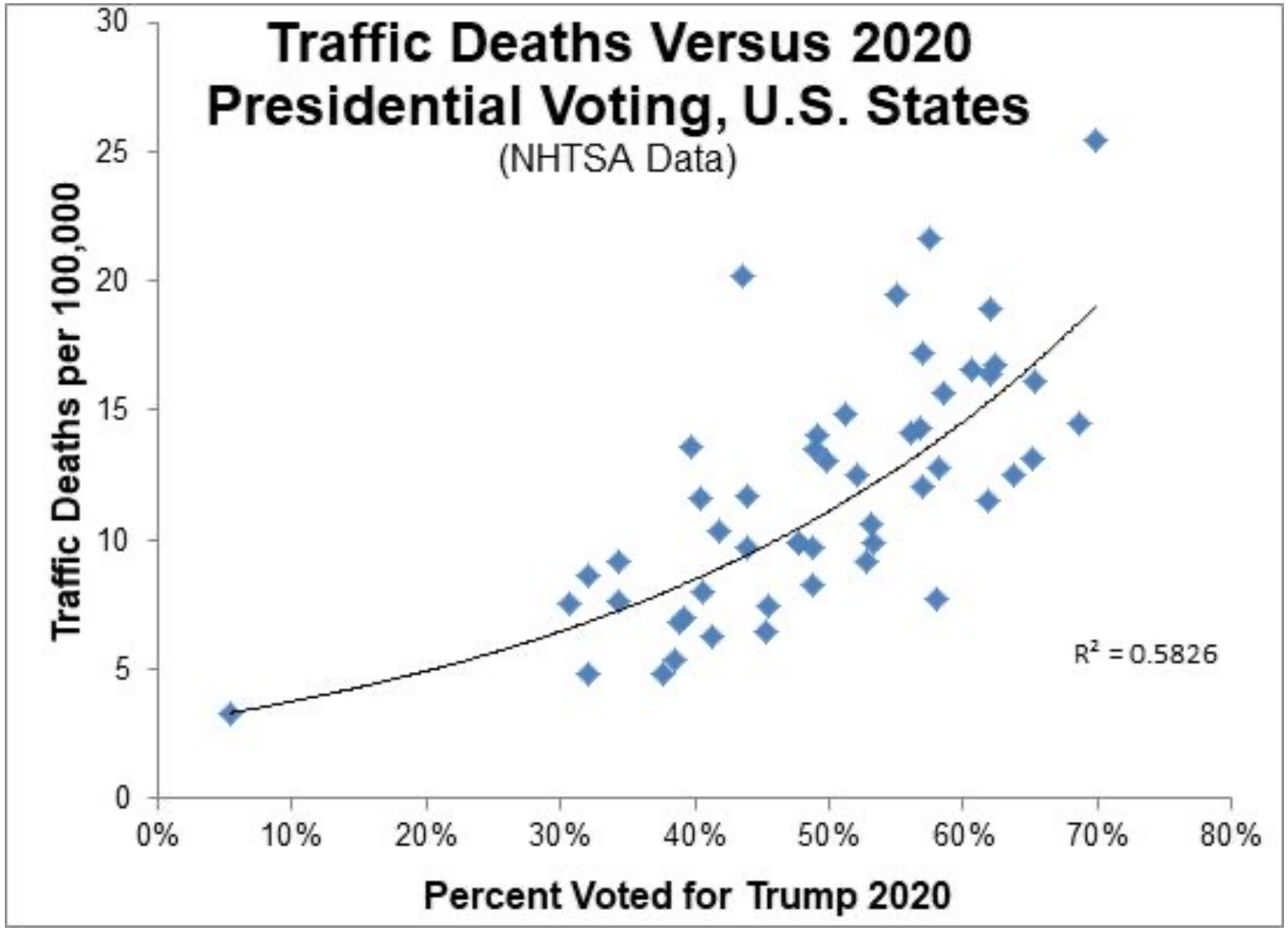


Data science is the field of study that **combines domain expertise**, programming skills, and **knowledge of mathematics and statistics** to extract meaningful insights from data.

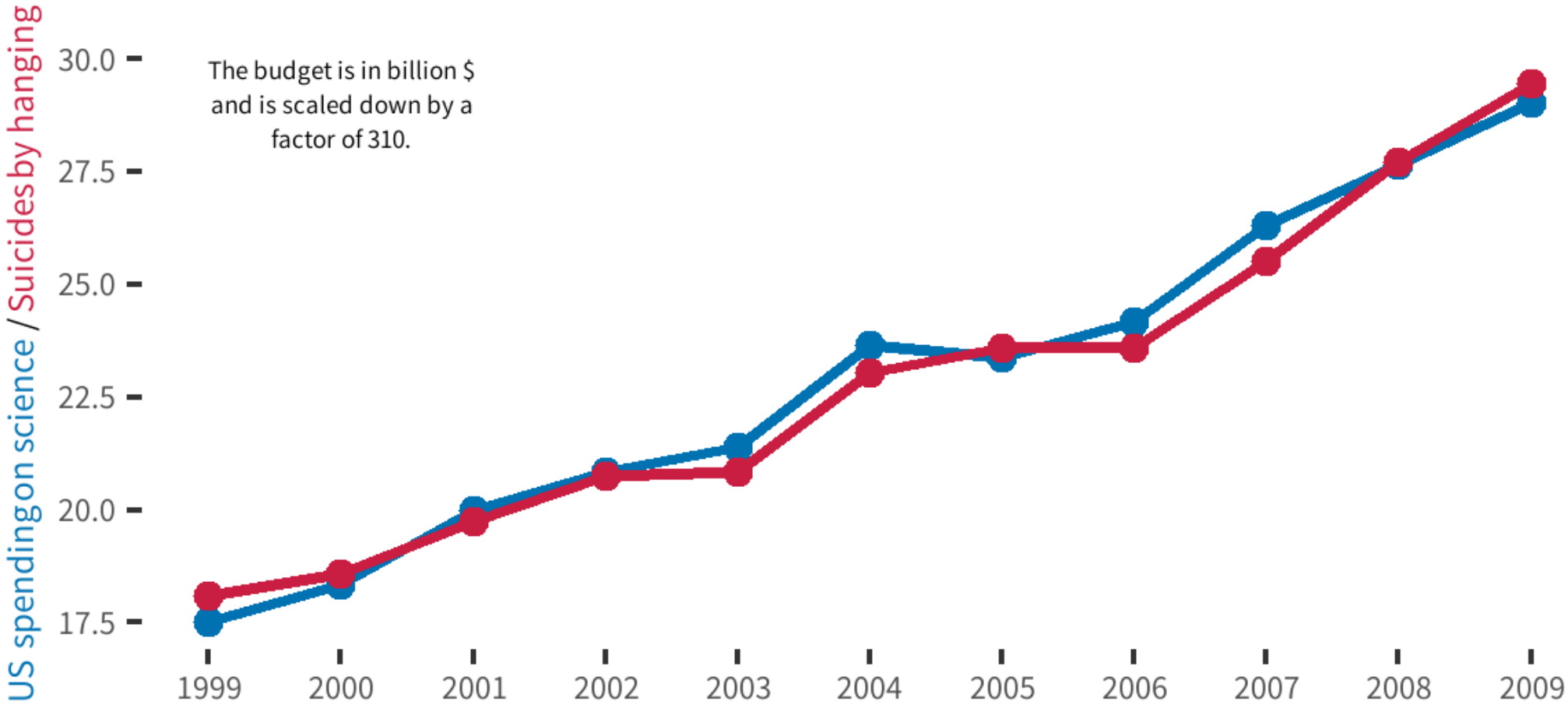
What is correlation?



How do you interpret data ?

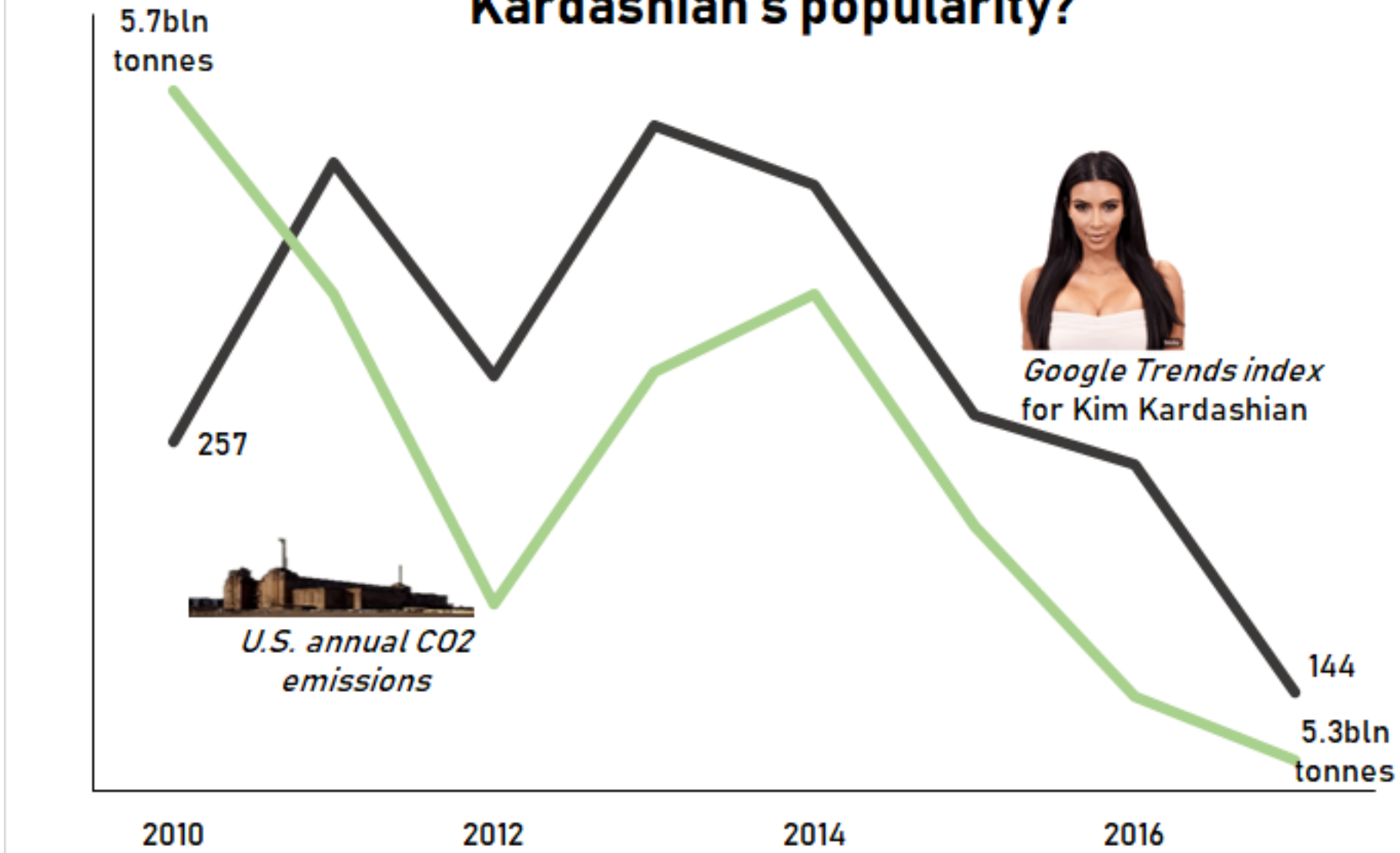


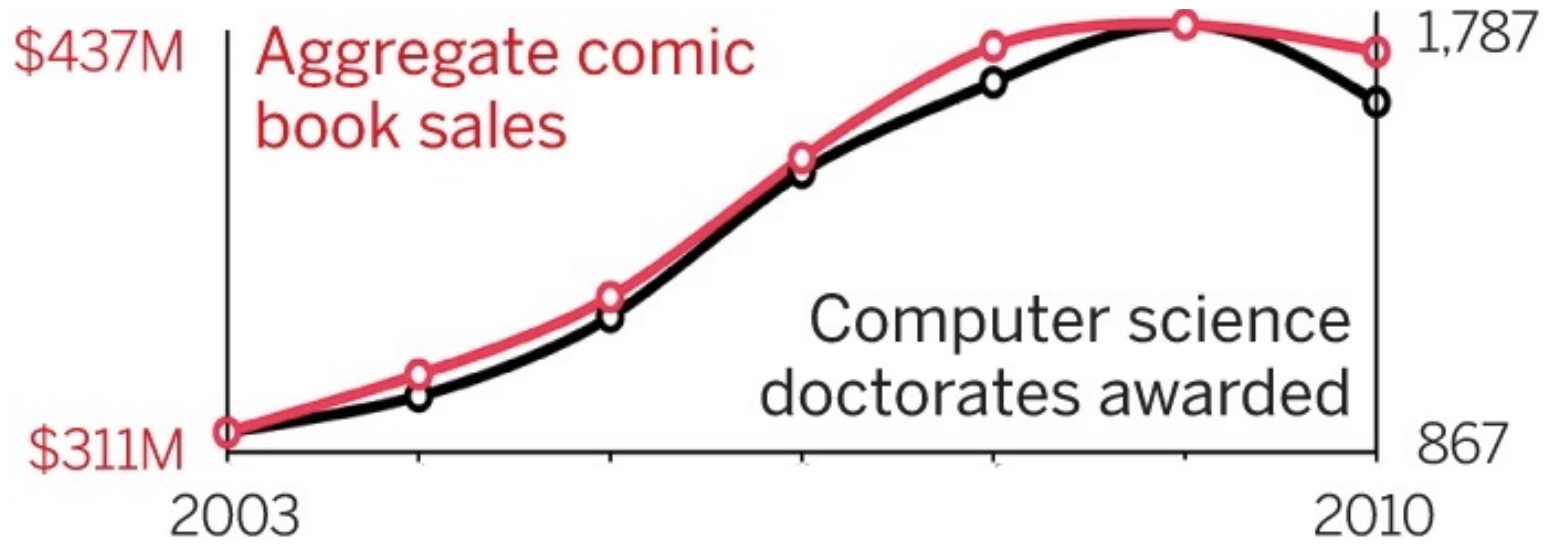
The correlation between US spending on science and suicides by hanging is 99.21%.

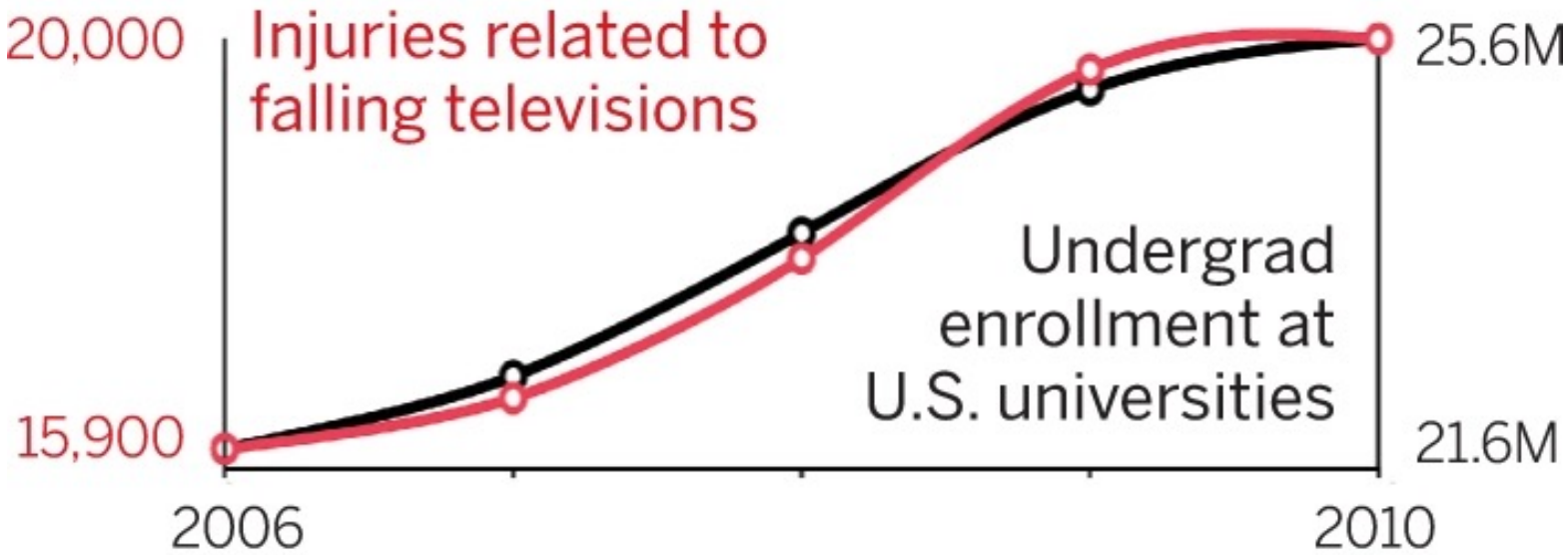


Inspiration: https://www.tylervigen.com
Data: U.S. Office of Management and Budget, Centers for Disease Control & Prevention
Graphic: Abhianv Malasi

Are falling emissions levels impacting Kim Kardashian's popularity?

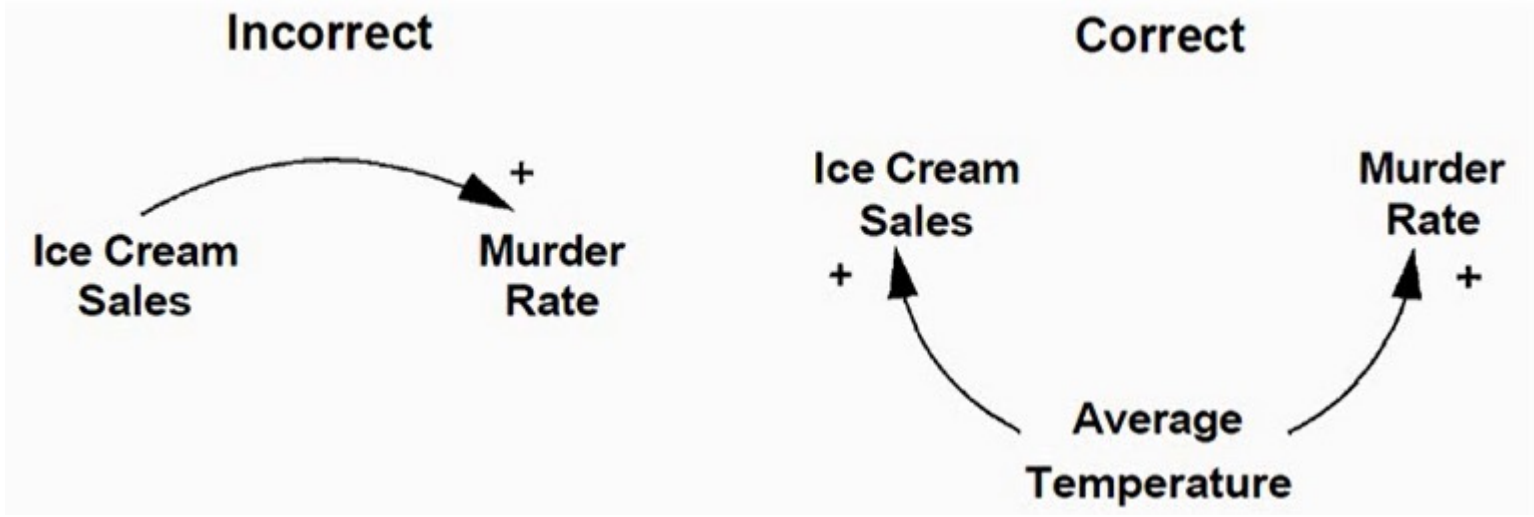



$437M
Aggregate comic book sales
$311M
2003
1,787
Computer science doctorates awarded
867
2010


20,000
Injuries related to falling televisions
15,900
2006
25.6M
Undergrad enrollment at U.S. universities
21.6M
2010

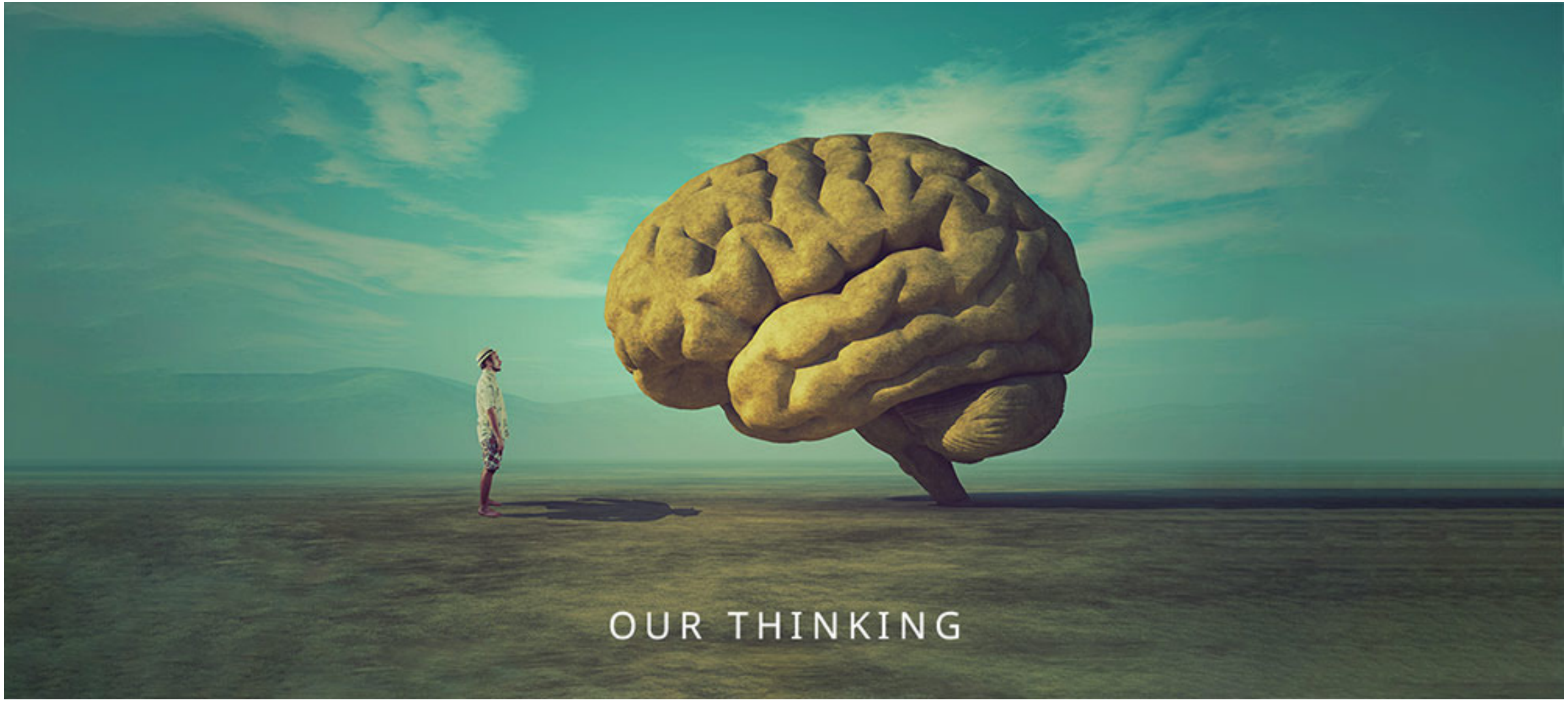
Causality vs. Correlation

- A **correlation** between variables does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.
- **Causation** indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events.



Which type of code developer would you like to be?

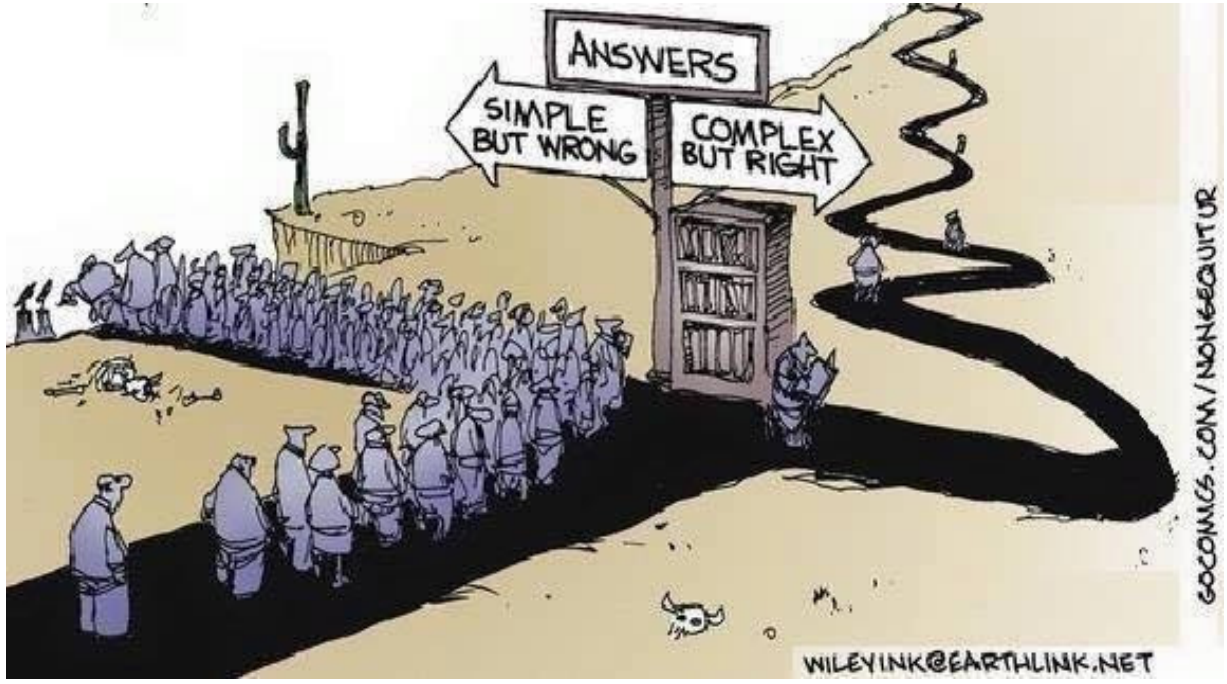
The Key-maker
The Architect
Or Neo



Linear Thinking vs. System Thinking
How to code system

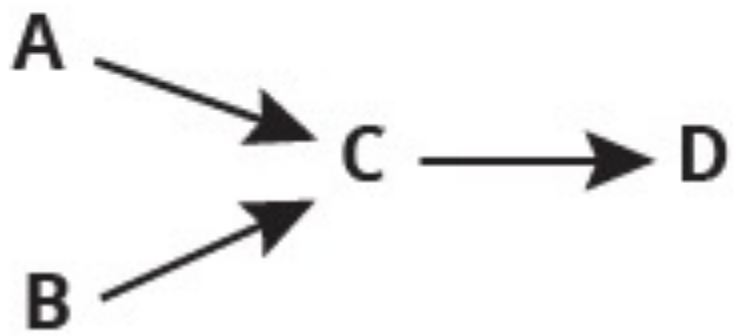
Linear Thinking

- Most people have been taught over their lifetime to see things linearly.
- We are taught that there is one cause and then one effect, a beginning and then an end, a problem and then one solution.
- The reality is that we often simplify things so that they fit into this model of thinking.



Event Oriented Thinking

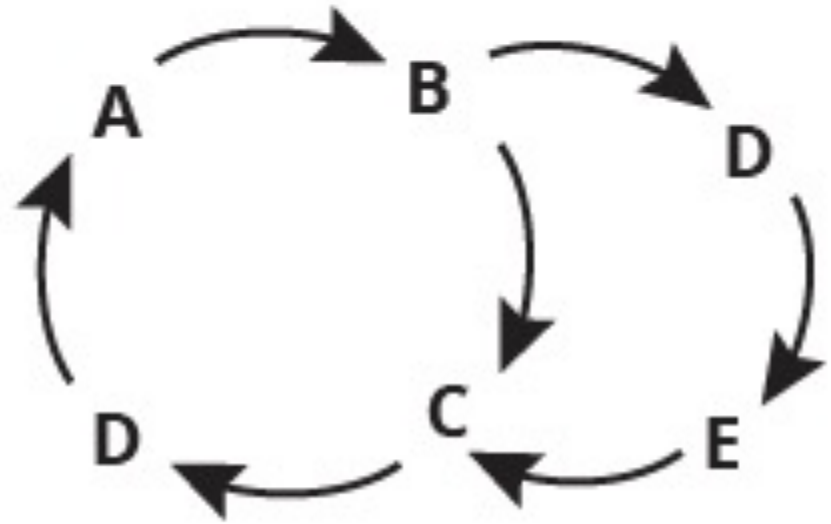
Thinks in straight lines



In event oriented thinking everything can be explained by causal chains of events. From this perspective the **root causes** are the events starting the chains of cause and effect, such as A and B.

Systems Thinking

Thinks in loop structure



In systems thinking a system's behavior emerges from the structure of its feedback loops. **Root causes** are not individual nodes. They are the forces emerging from particular feedback loops.

http://www.thwink.org/

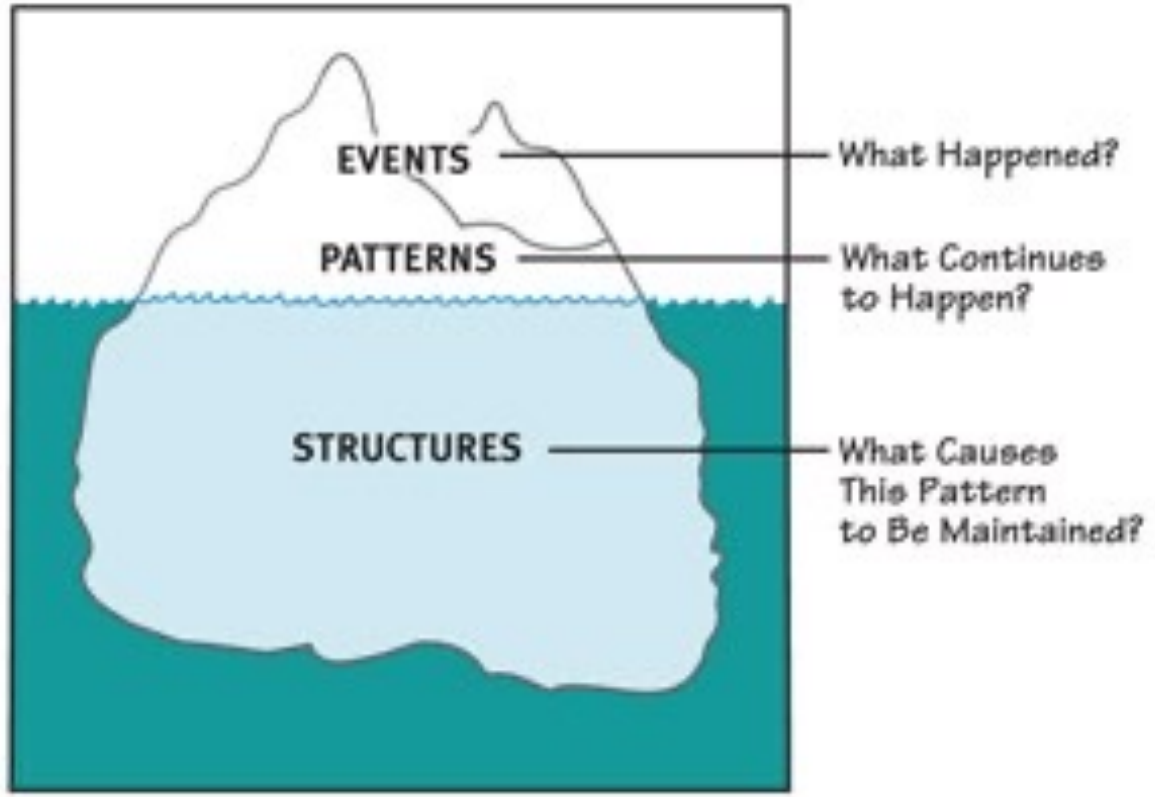
System Thinking (System Definition)

- **System** : An organized entity made up of interrelated and interdependent parts.
- A system is “a group of interacting, interrelated and interdependent components that form a complex and unified whole.
- Key concepts are
- Elements or components that make up the system parts
- Relationships and interactions among each of the elements or components
- Pattern of the behaviors of a system as a whole over time (called its dynamic behavior)

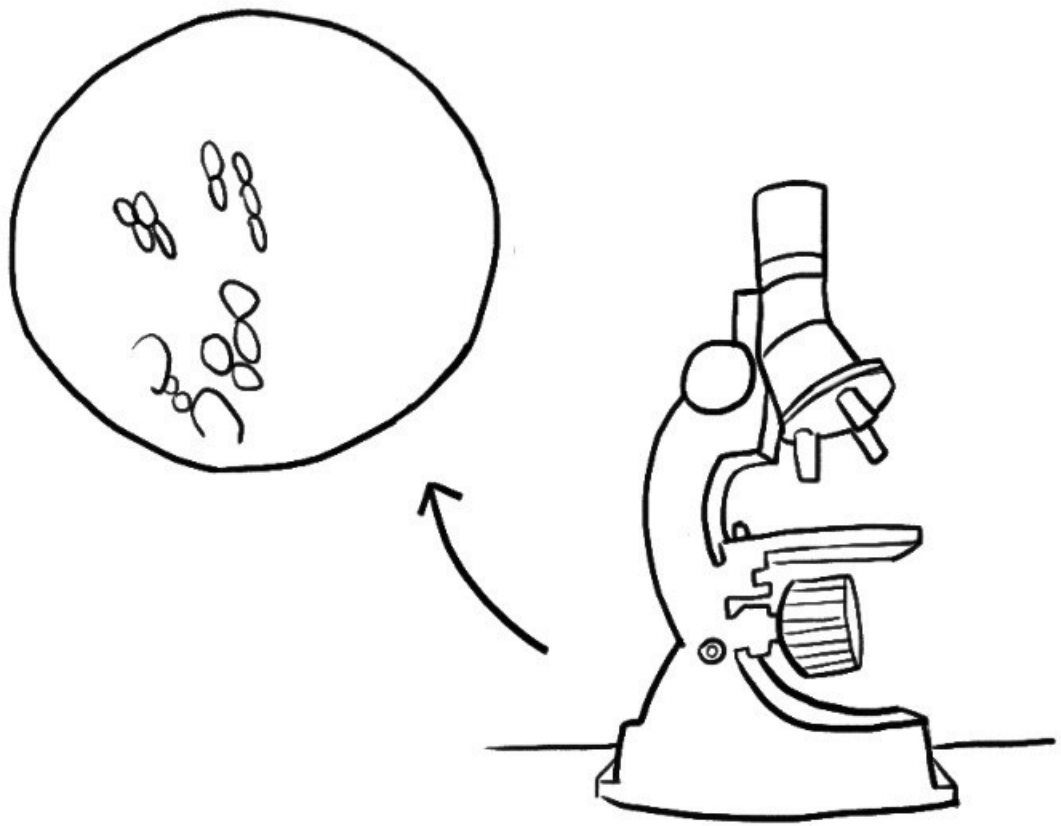
System Thinking

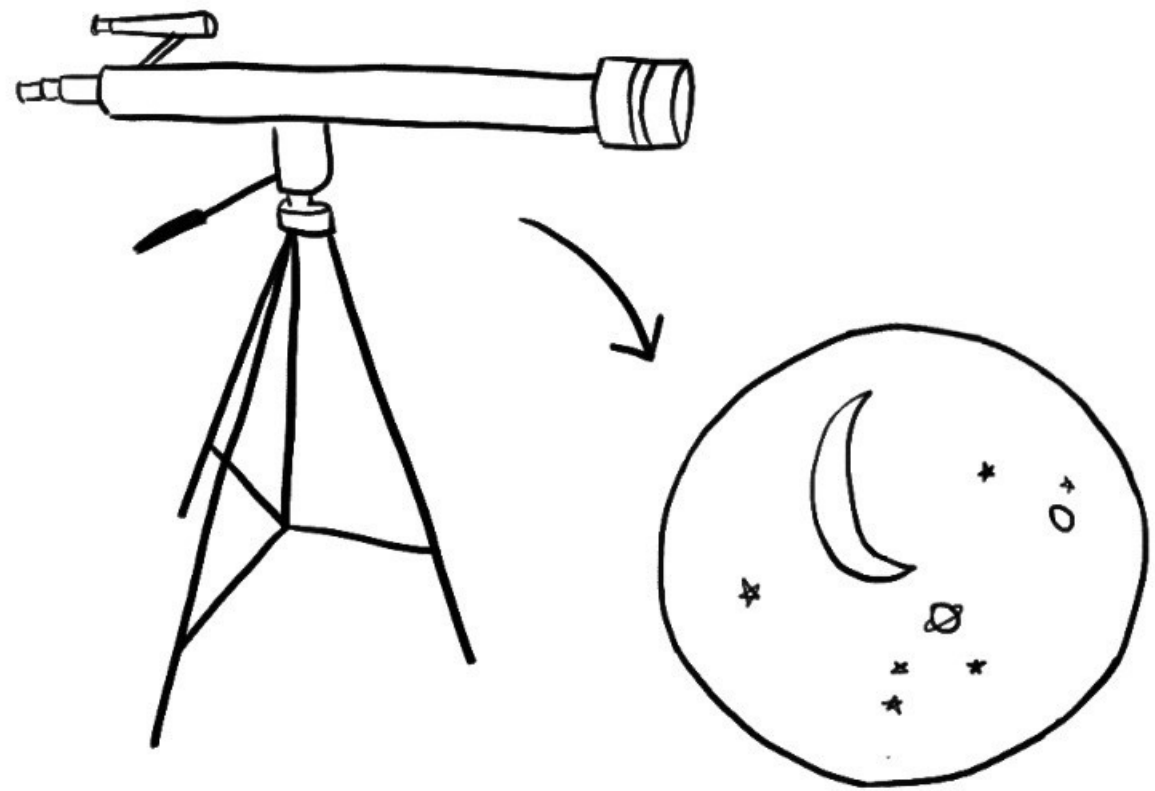
- **Systems thinking** is a management discipline that concerns an understanding of a system by examining the linkages and interactions between the components that comprise the entirety of that defined system.

- Iceberg Model of system Thinking



SYSTEMS THINKING SCALES





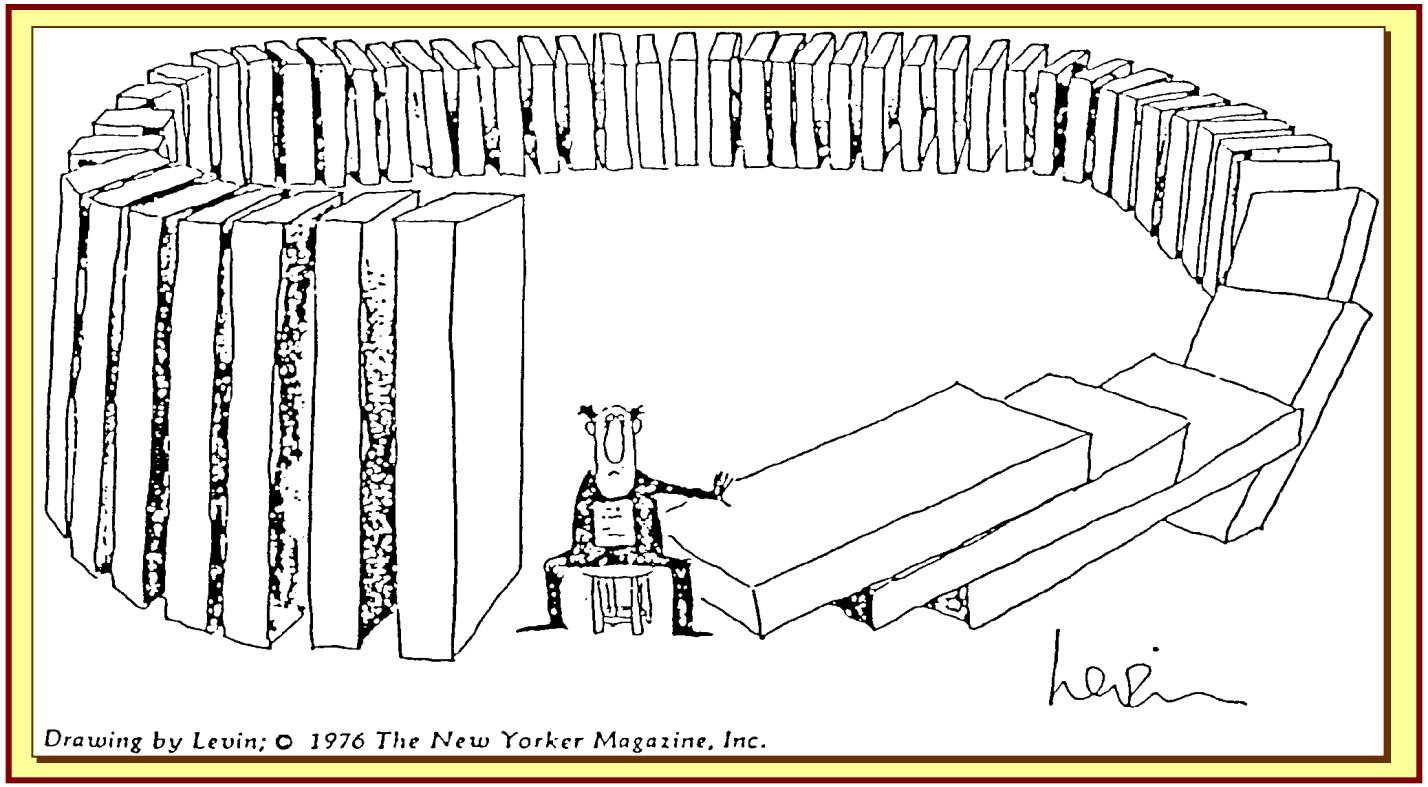
DISRUPT DESIGN

Understanding Complexity

From a Very Particular Distance

*"{System dynamics studies problems} from 'a very particular distance',
not so close as to be concerned with the action of a single individual, but not so far away as to be ignorant of the internal pressures in the system."*

-- George Richardson



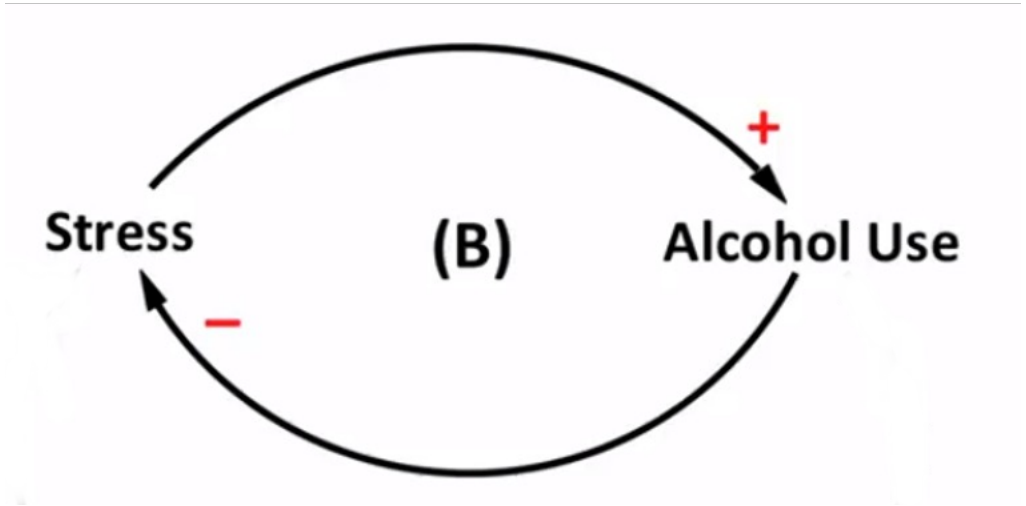
Forrester JW. Counterintuitive behavior of social systems. Technology Review 1971;73(3):53-68.

Meadows DH. Leverage points: places to intervene in a system. Sustainability Institute, 1999.
Available at <http://www.sustainabilityinstitute.org/pubs/Leverage_Points.pdf>.

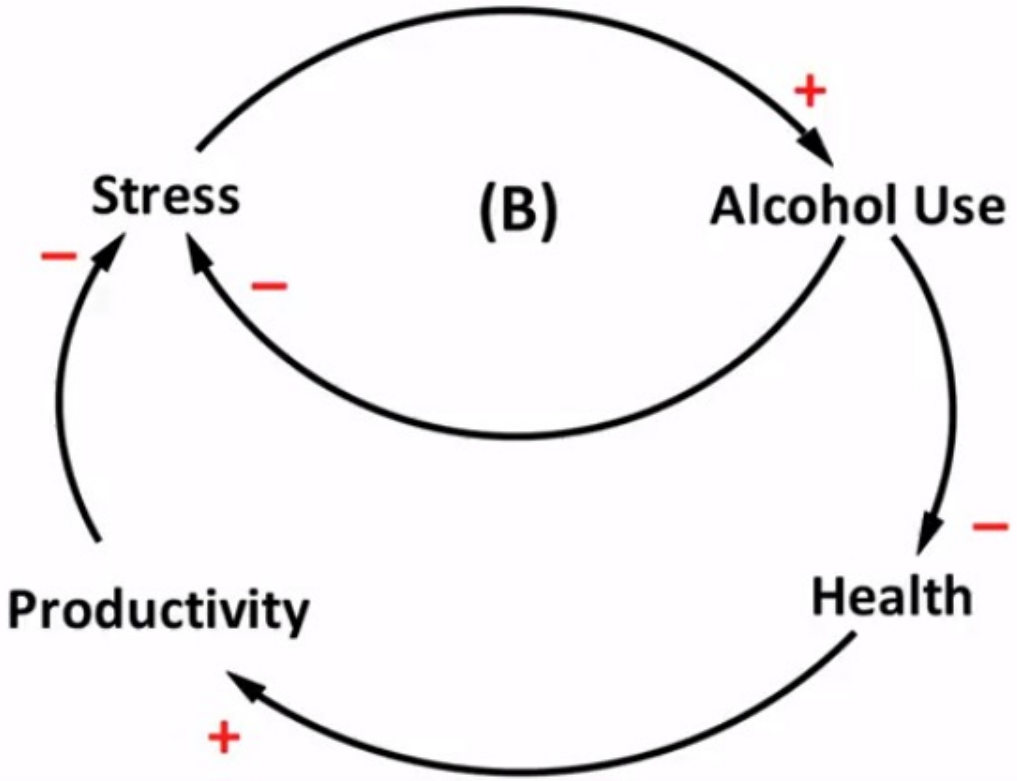
Richardson GP. Feedback thought in social science and systems theory. Philadelphia, PA: University of Pennsylvania Press, 1991.

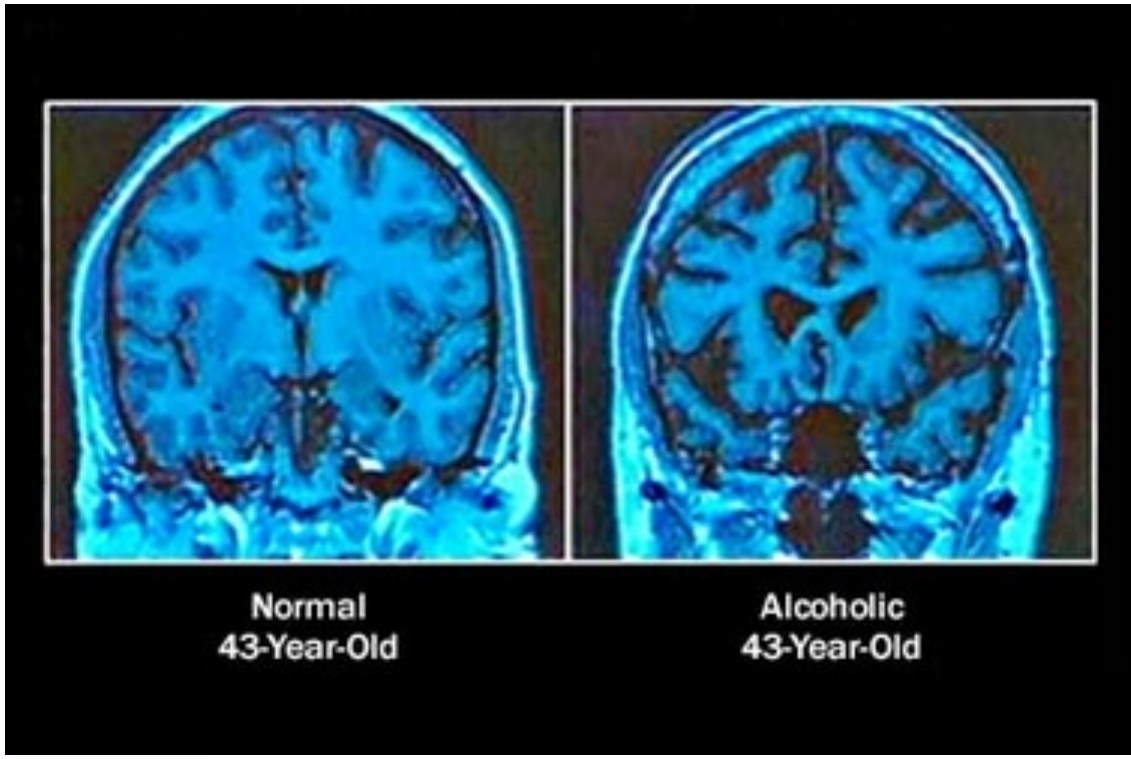
Sterman JD. Business dynamics: systems thinking and modeling for a complex world. Boston, MA: Irwin McGraw-Hill, 2000.

Perfect Drink Pro app: Cocktail maker



Consider balancing and reinforcing loops together





Think in terms of cause-and-effect and feedback loops in time horizon

Some examples: System thinking vs Linear thinking

Question

- If we replace all cars of one city by **high** efficient fuel-consume cars :

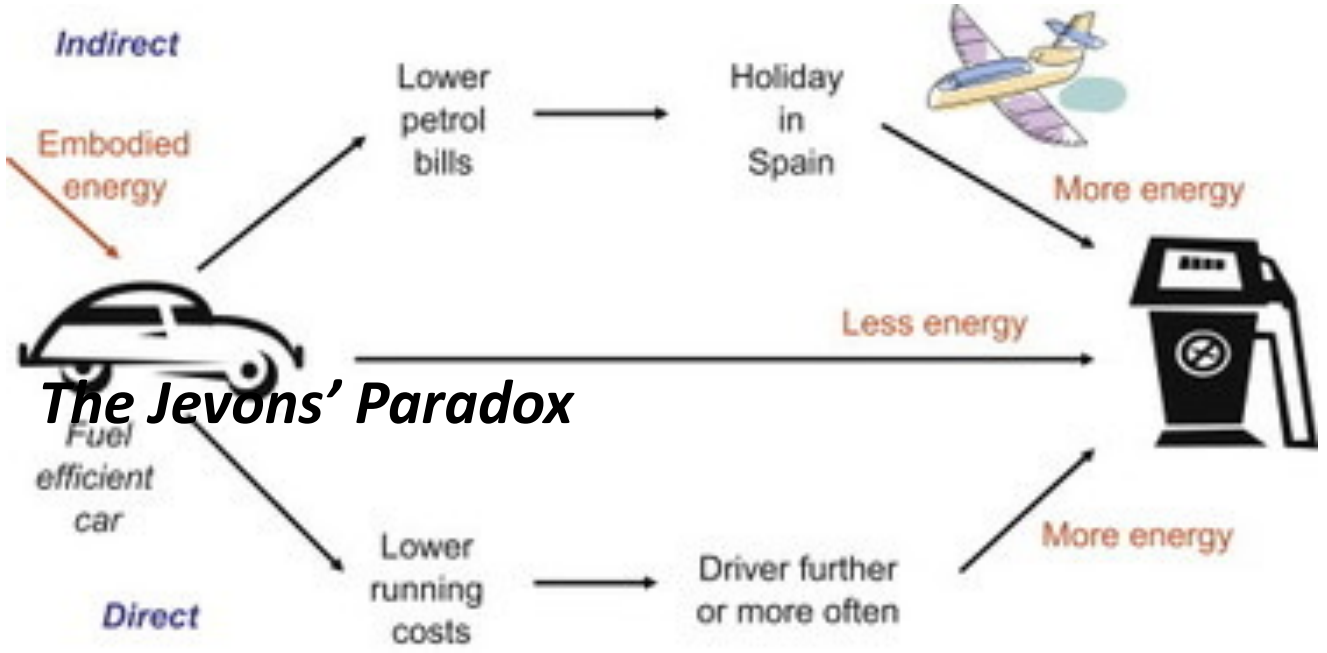
Do you think the fuel consumption of whole of this city will decrease or increase ?

1- Jevons paradox

- In economics, the Jevons paradox occurs when technological progress increases the efficiency in resource consumption the rate of consumption of that resource rises because of increasing demand.
- **Attempts to reduce energy consumption by increasing energy efficiency would simply raise demand for energy in the economy as a whole**

Willliam Stanley Jevons

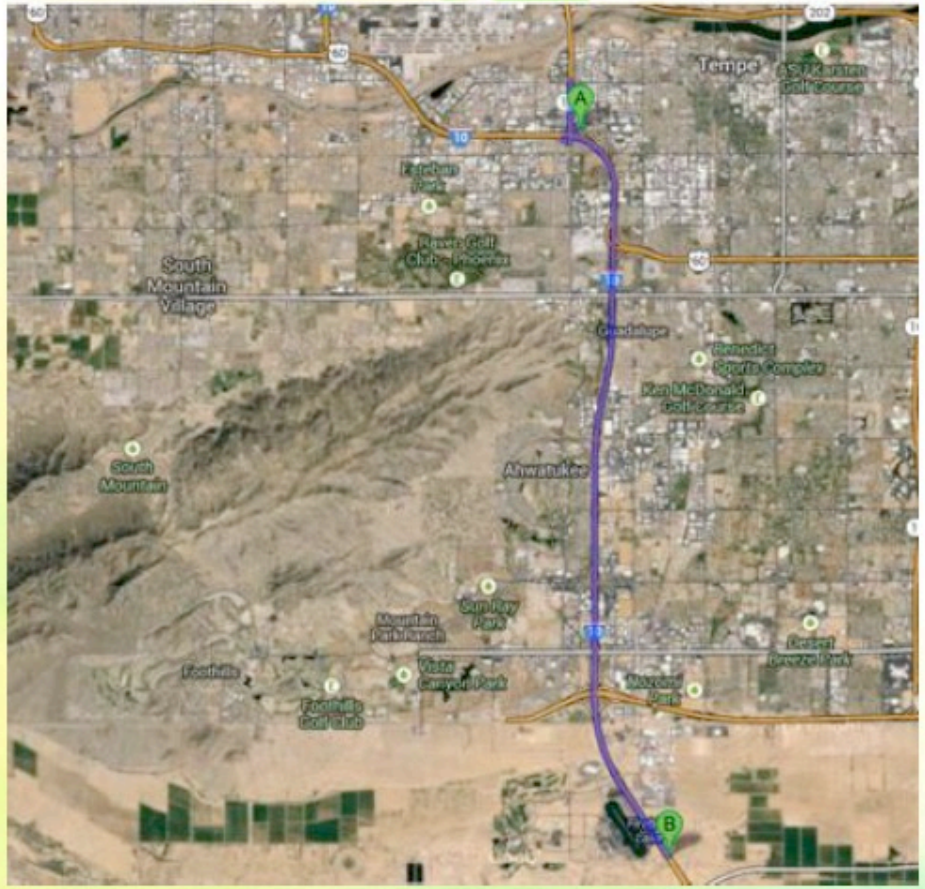
The coal question - 1865



Question:

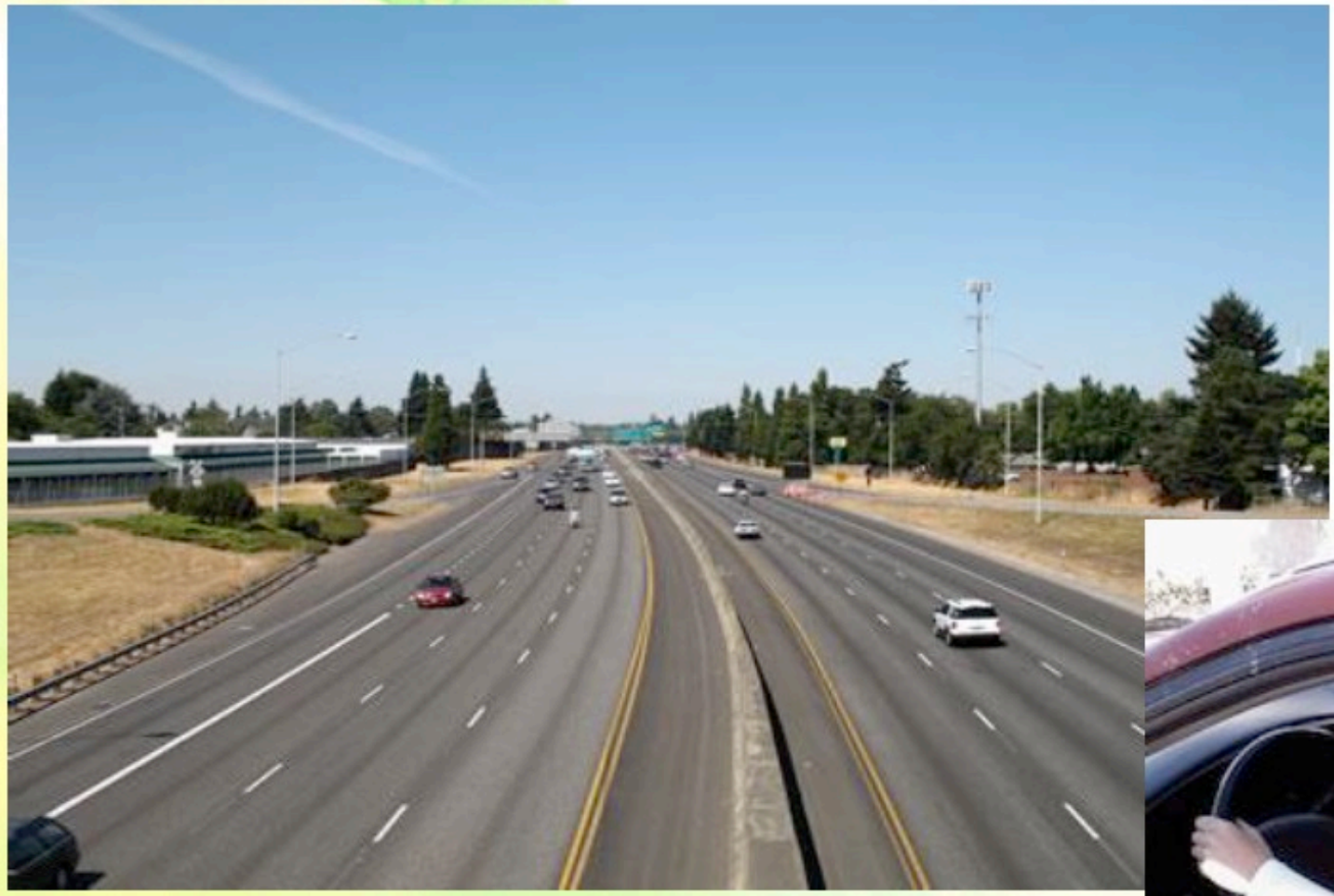
By constructing new roads or increase road capacity of cities , Could we solve traffic congestion ? Or could we reduce the cost of urban transport in cities ?

Case Study

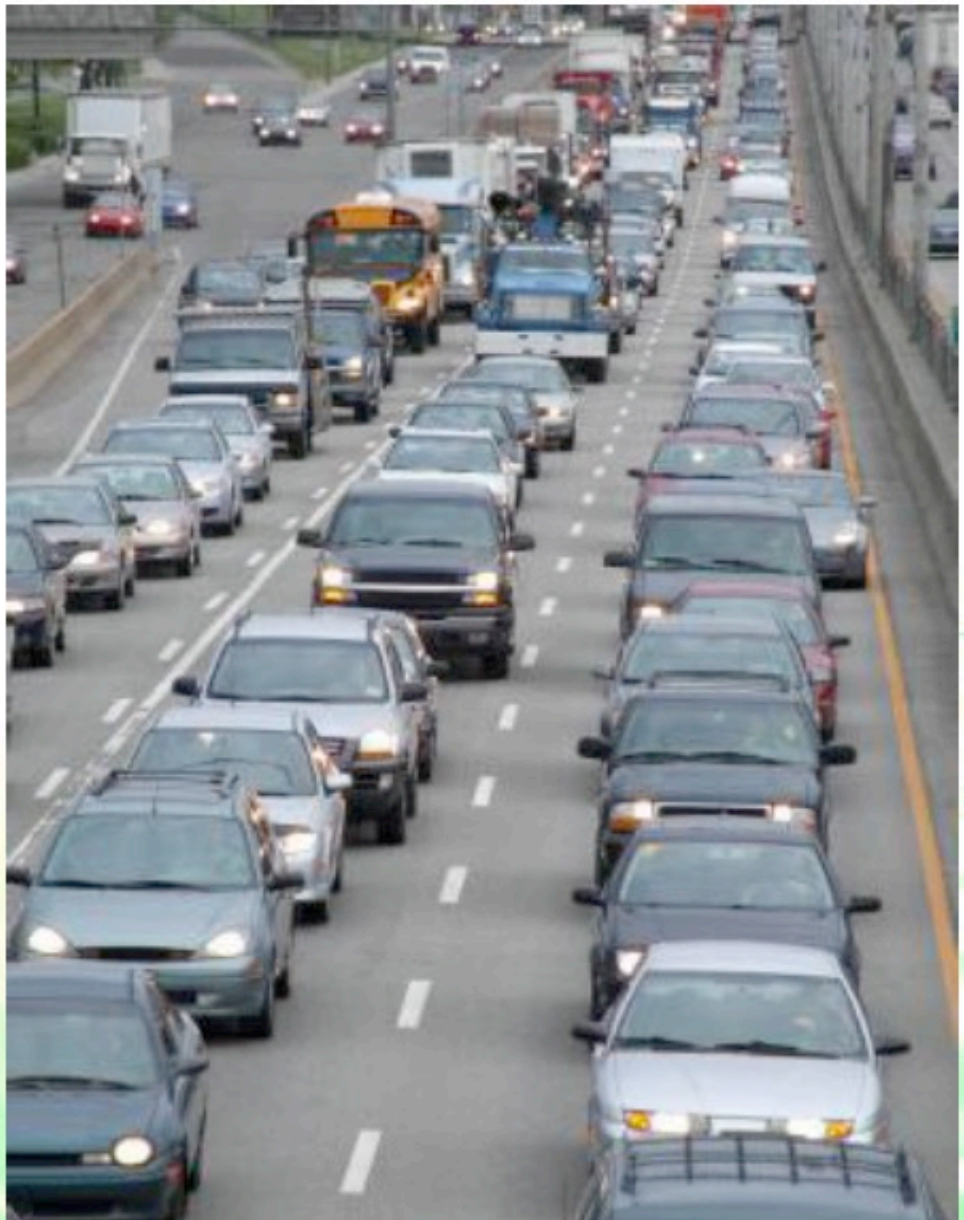


In 1995-1996, I-10 from the Phoenix airport to Queen Creek in Arizona

Reopen with Six Lanes

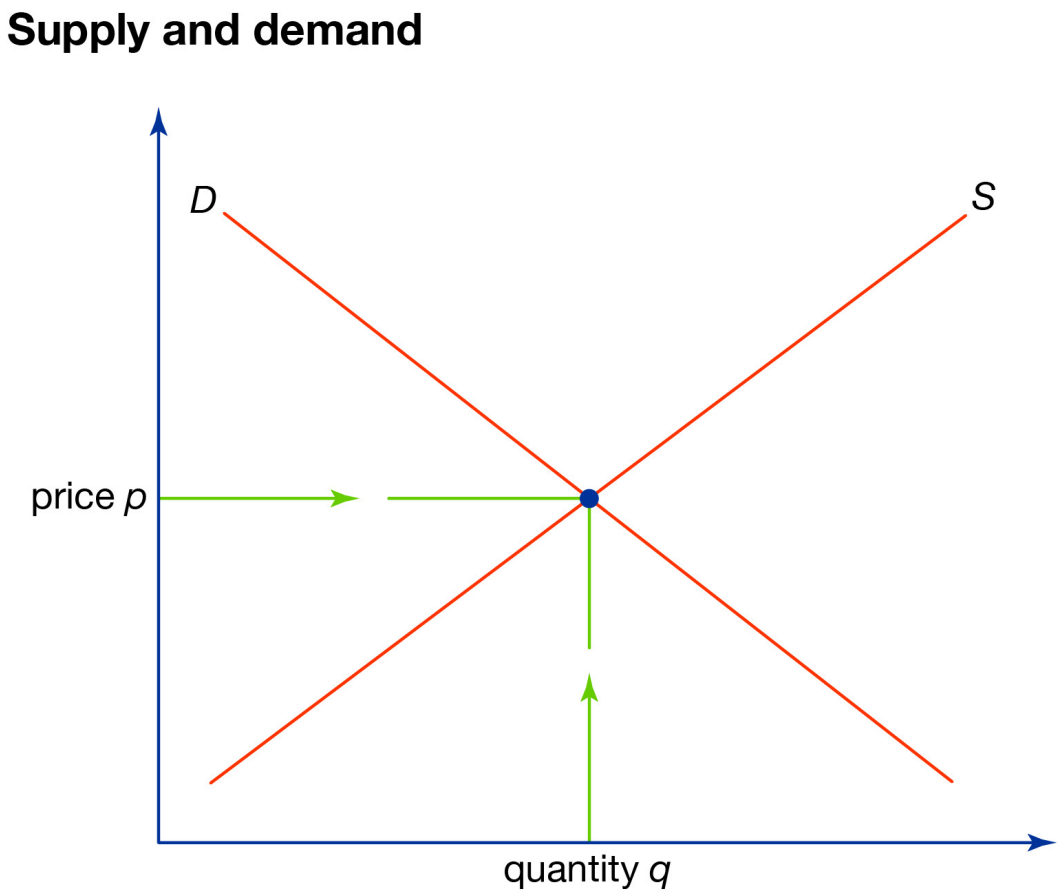


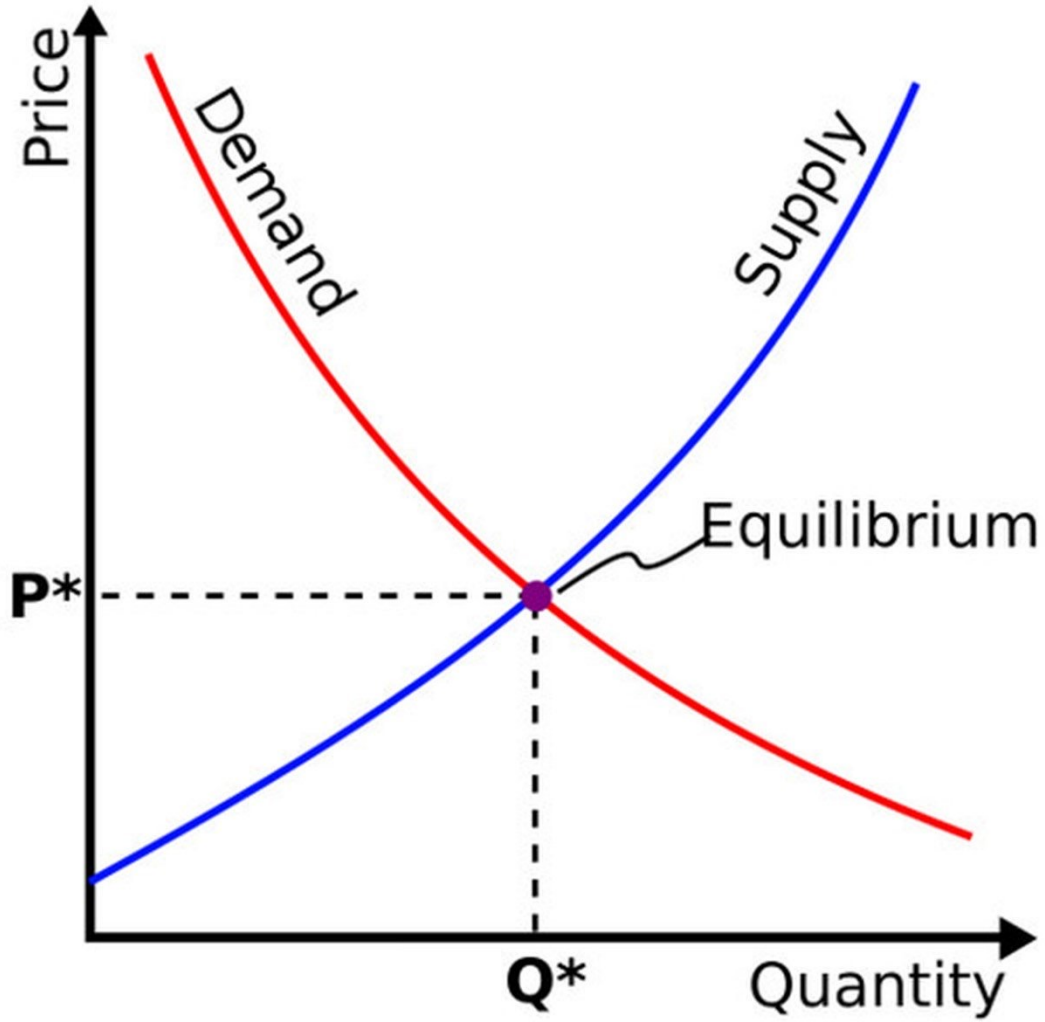
Six Months Later



Congestion returned with a vengeance

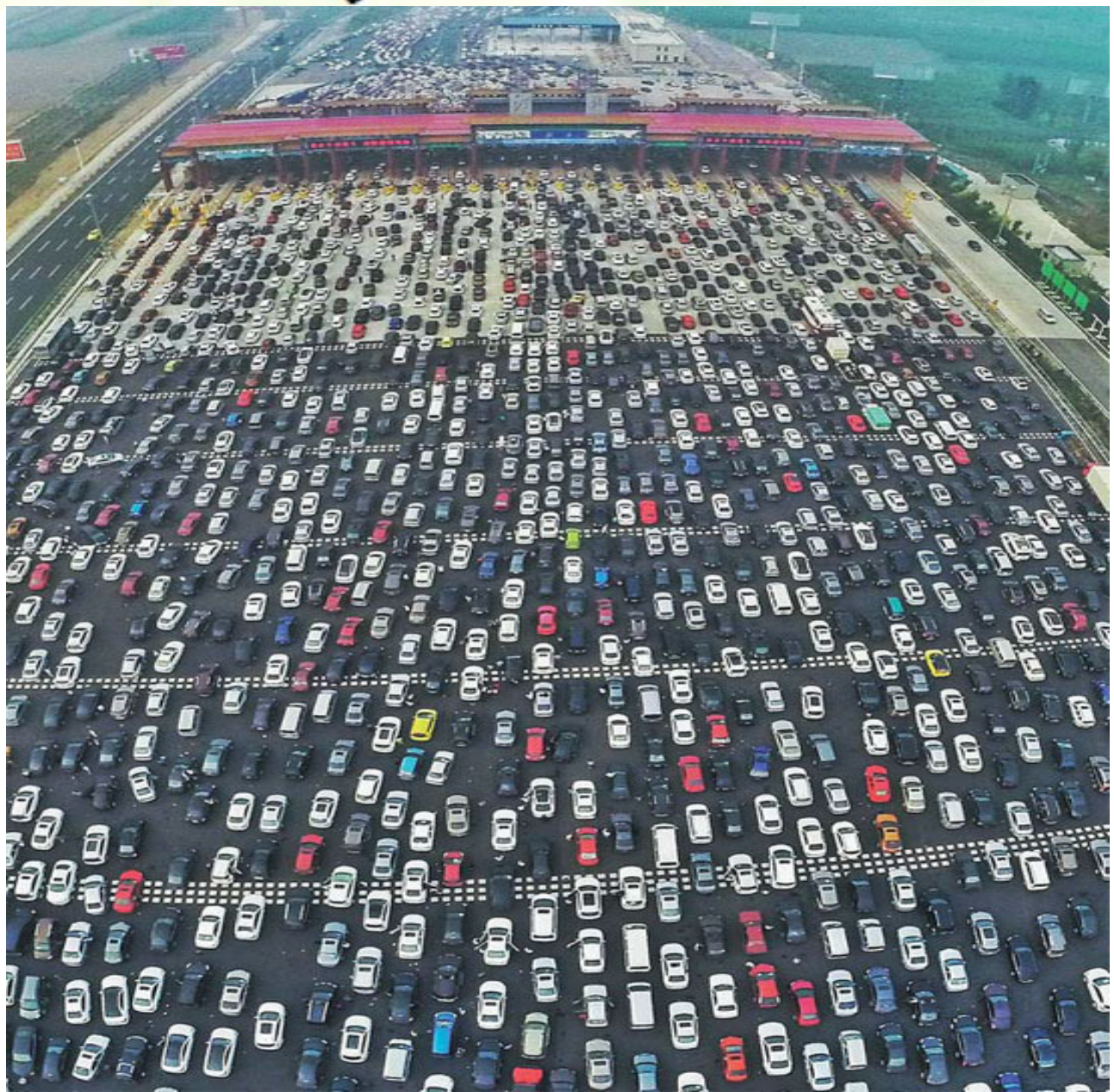
Supply Demand Curves





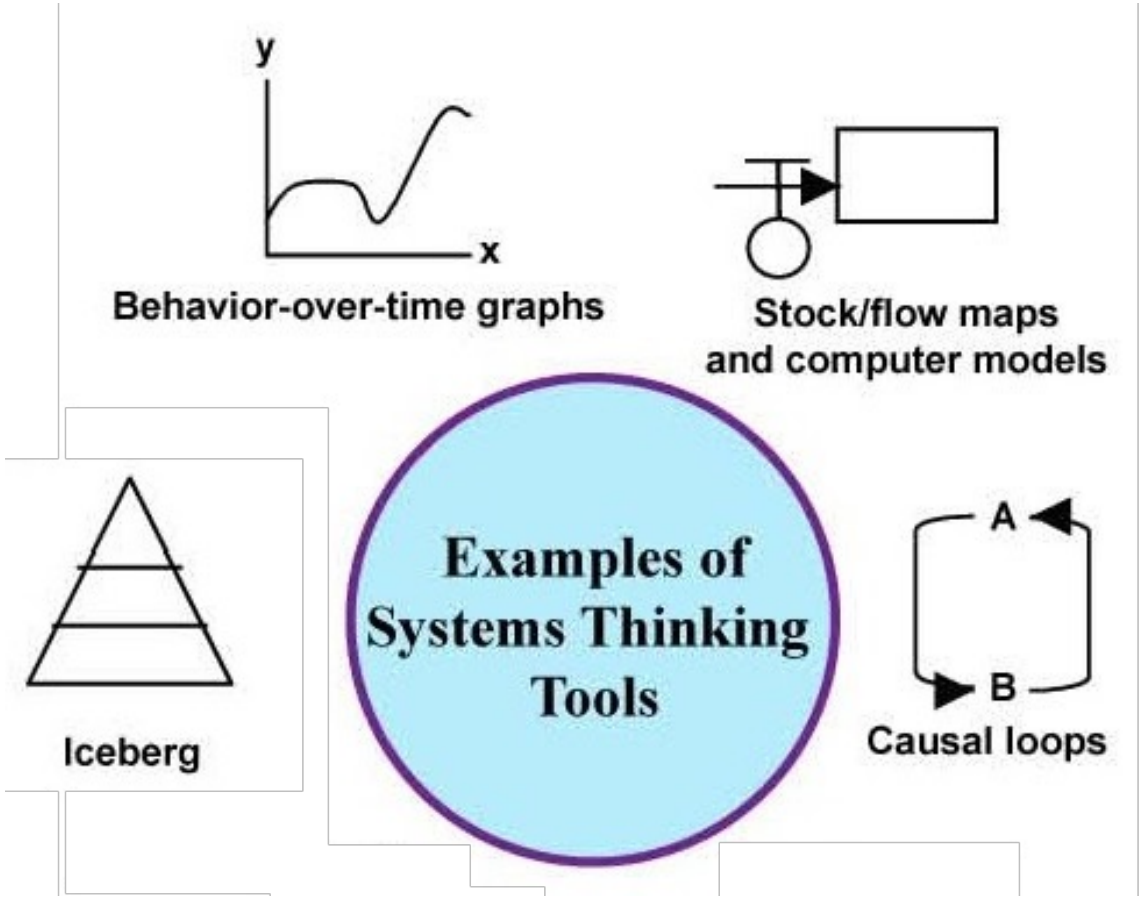
"If you build them, they will come!"

Heavy traffic in Beijing



https://youtu.be/yUEHWhO_HdY

How to understand , analyze and predict the reaction of the systems?



understand
structure

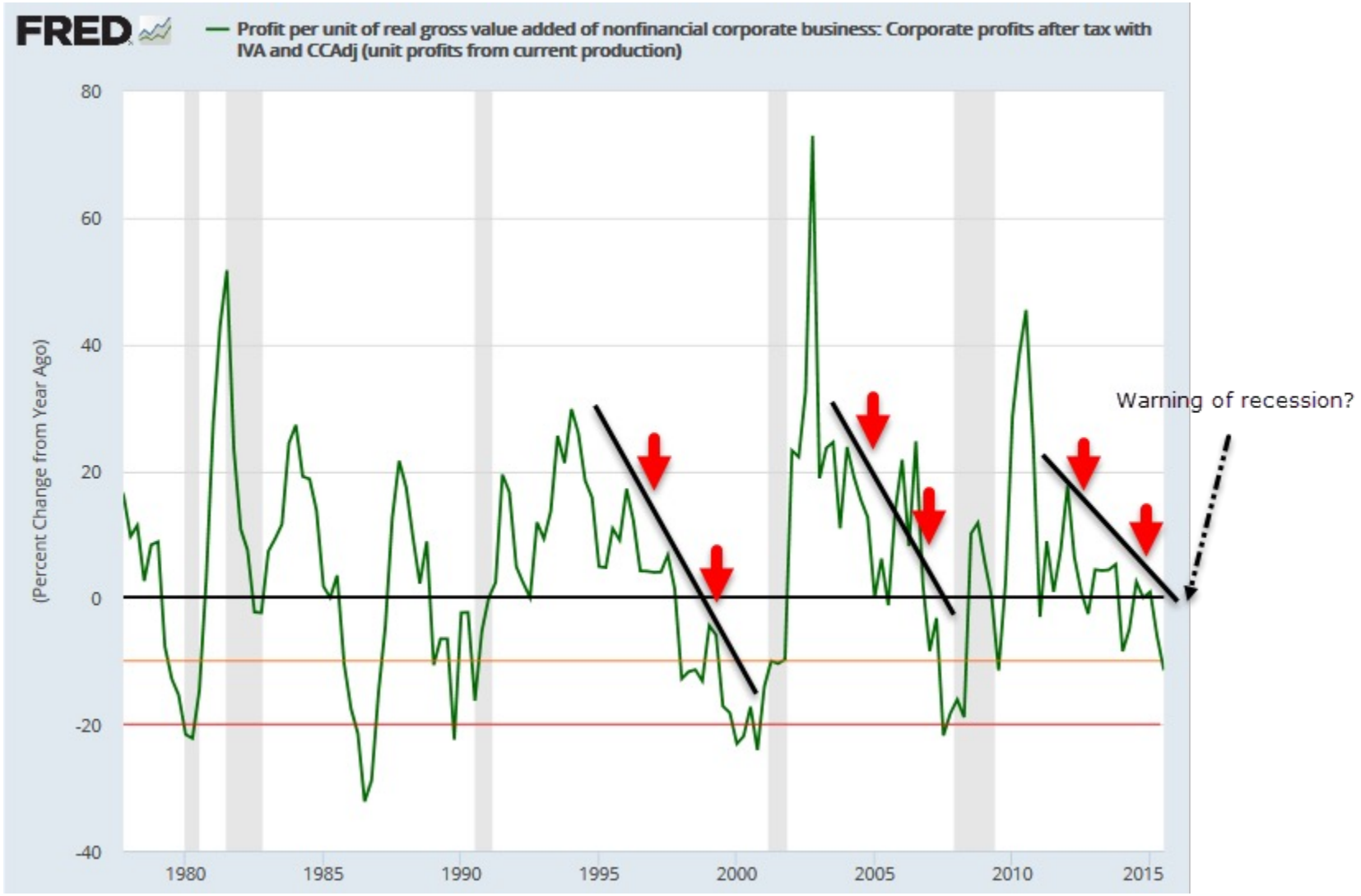
↓

understand
behavior

↓

Code develop
Give solutions

Behavior Patterns



Behavior Patterns in Stock prices

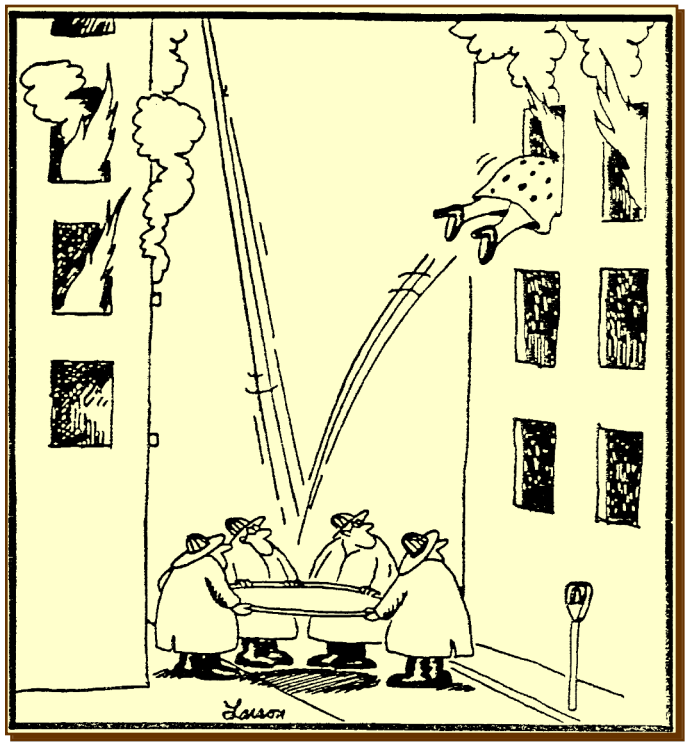
https://www.youtube.com/watch?time_continue=29&v=2FnEGipMUNE

System Dynamics Modeling

- System dynamics modeling was developed from system thinking ideas. It started from the work of Jay Forrester, to study the behavior of various components interrelated each other in the system (Forrester, 1918 - 2016).

- System Dynamics modeling is a computer-aided approach based on cause-and-effect analysis and feedback loop structures, used for to theory building, policy analysis and strategic decision support

System Dynamics Was Designed to Address Problems Marked By Dynamic Complexity



Origins

- Public policy applications starting late 1960s Jay Forrester, MIT

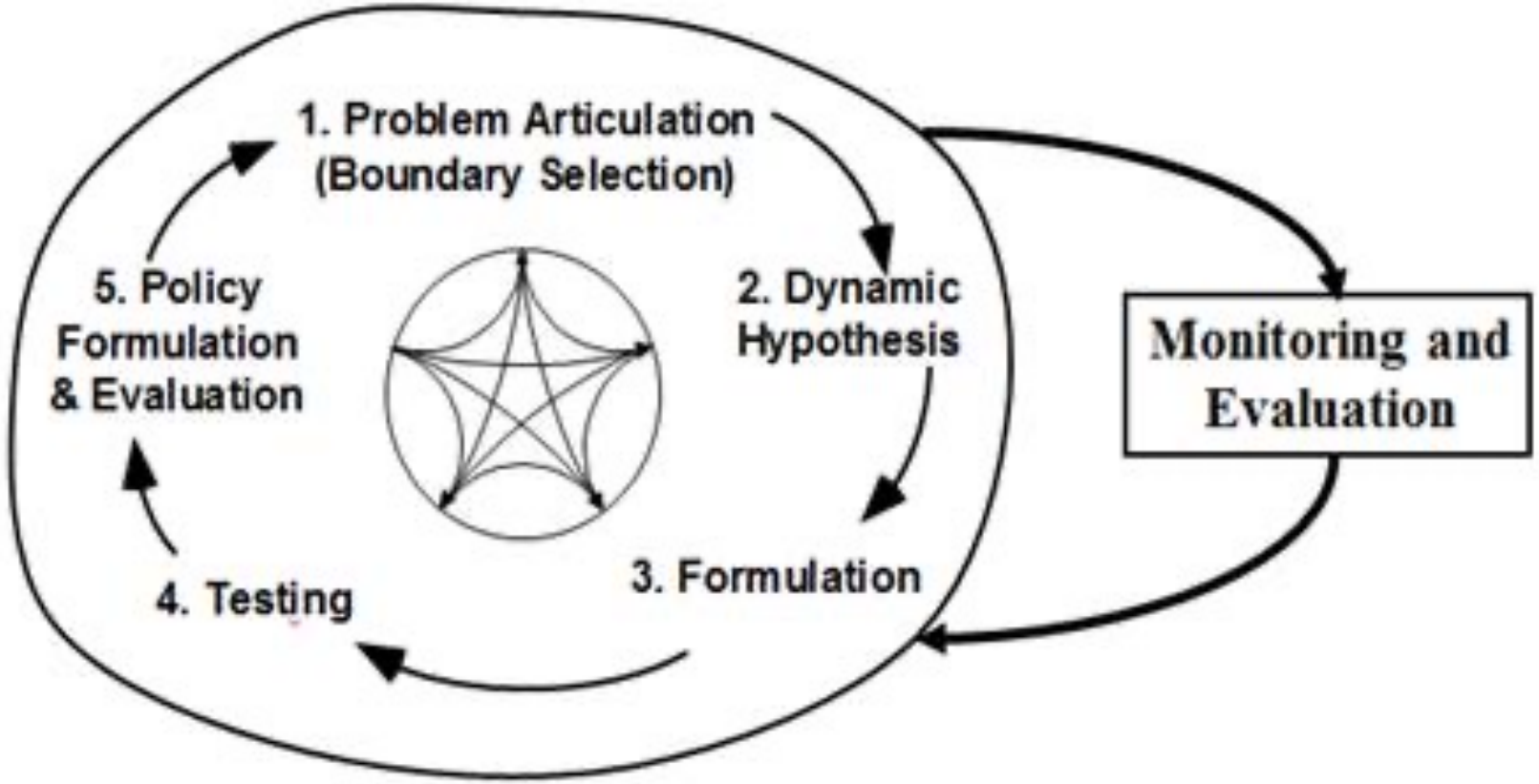
Good at Capturing

- Differences between short- and long-term consequences of an action
- Time delays
  (e.g., transitions, detection, response)
- Accumulations
  (e.g., prevalence, capacity)
- Behavioral feedback
  (e.g., actions trigger reactions)
- Nonlinear causal relationships
  (e.g., effect of *X* on *Y* is not constant)
- Differences or inconsistencies in goals/values among stakeholders

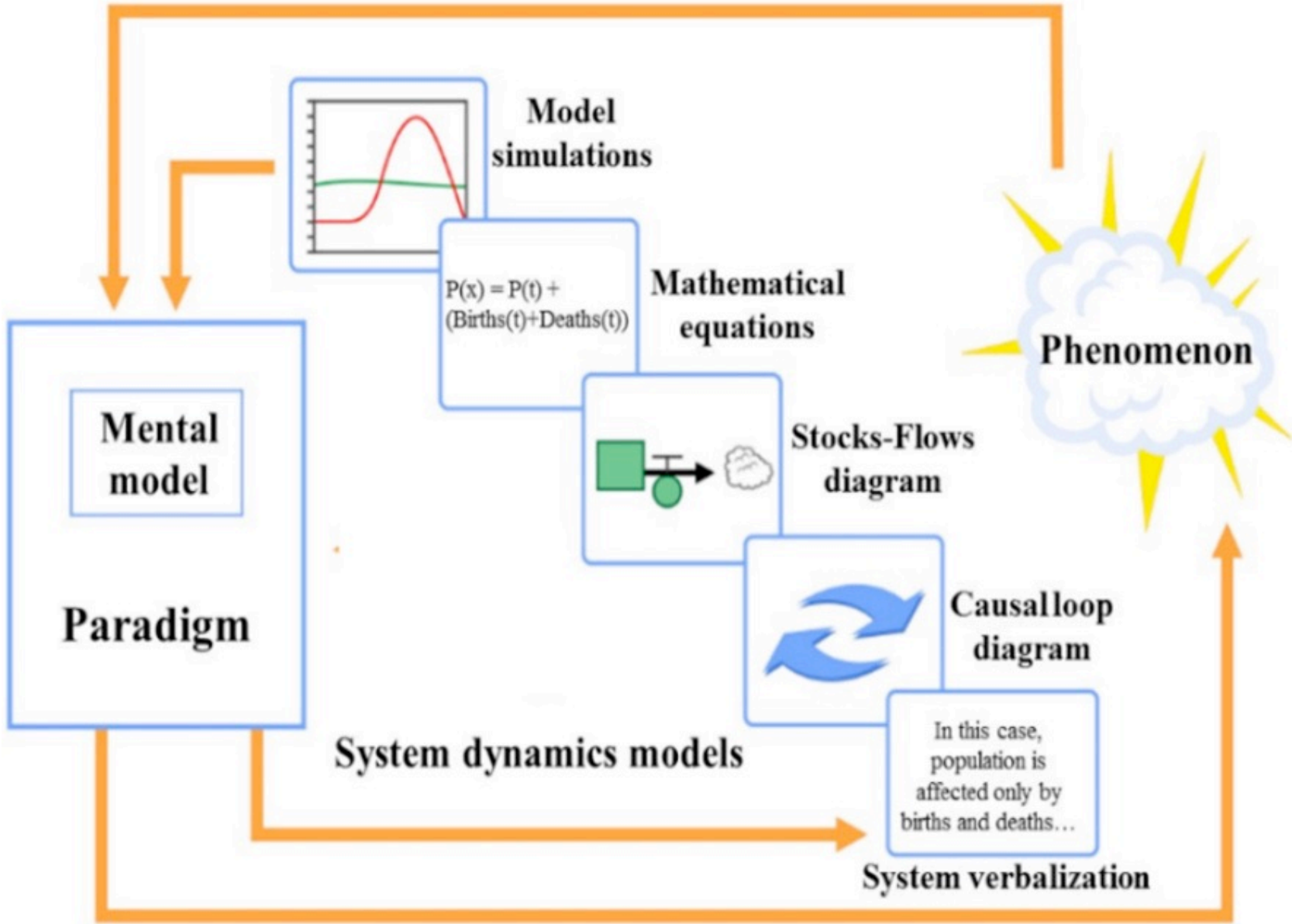
Sterman JD. Business dynamics: systems thinking and modeling for a complex world. Boston, MA: Irwin McGraw-Hill, 2000.

Homer JB, Hirsch GB. System dynamics modeling for public health: background and opportunities. American Journal of Public Health 2006;96(3):452-458.

Steps in System Dynamics Modeling Process



The research methodology adopted is in accordance to the modelling methodology as proposed by Sterman (2000).


Model simulations
P(x) = P(t) + (Births(t)+Deaths(t))
Mathematical equations
Phenomenon
Mental model
Paradigm
Stocks-Flows diagram
Causal loop diagram
In this case, population is affected only by births and deaths…
System dynamics models
System verbalization

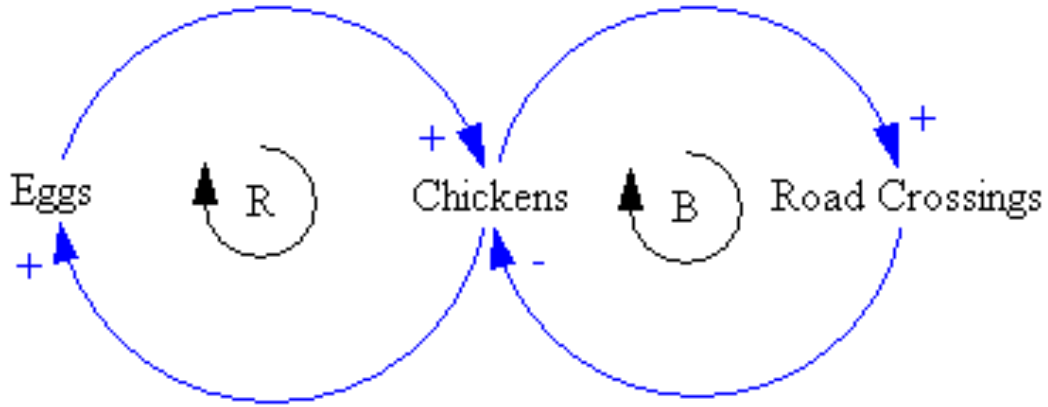
How to model the interrelations among different components of one systems ?

Causal Loop Diagrams

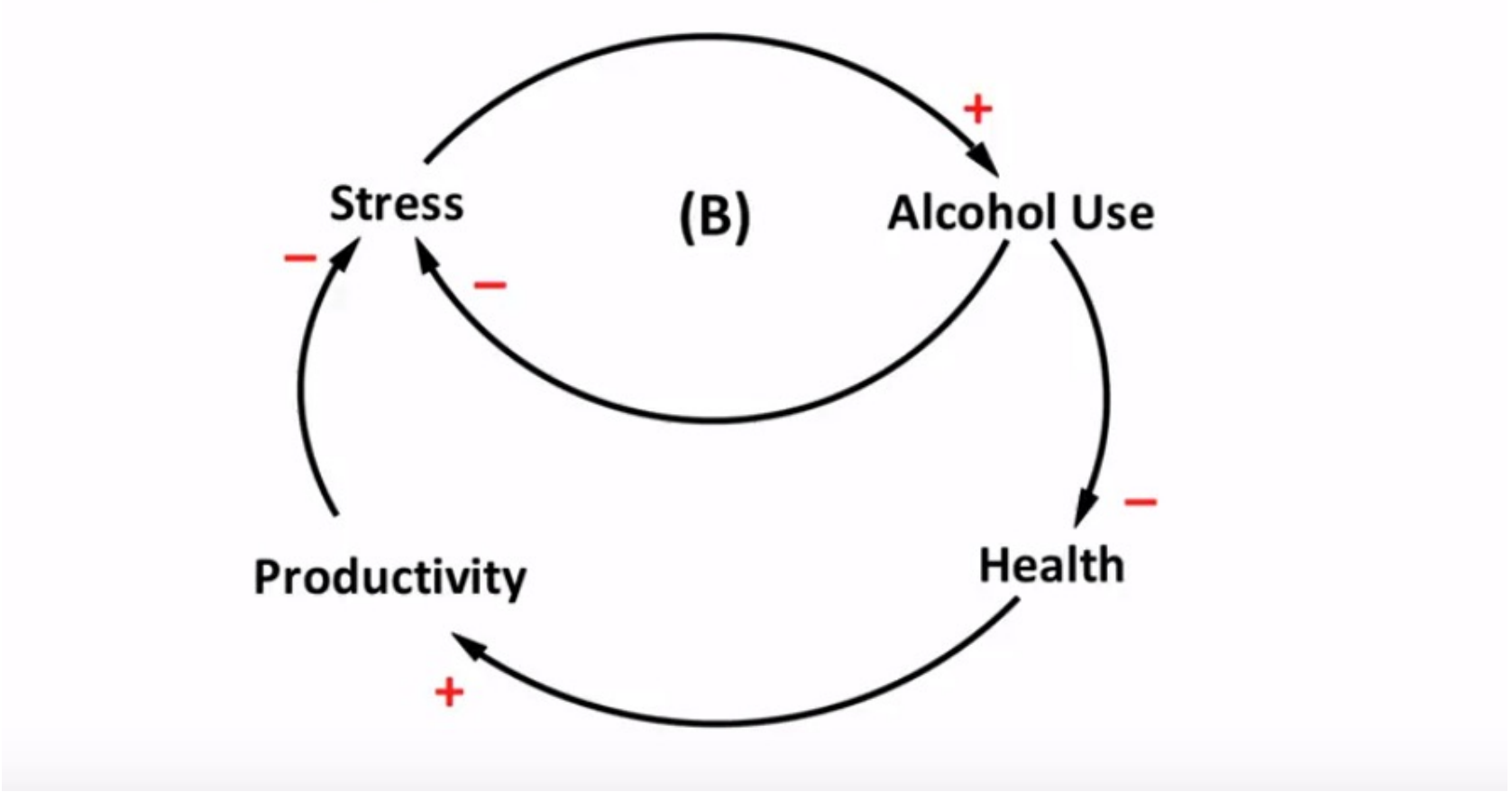
- Causal loop diagrams are a part of systems thinking and are "*a diagram that aids in visualizing how interrelated variables affect each other*". They consist causal links shown graphically by arrows that connect variables. Causal loop diagrams aid in visualizing a system's structure and behavior, and analyzing the system qualitatively.
- These features can change over time. The diagrams show how one variable affects another, because each relation between a variable has a polarity, i.e. positive (+) or negative (-) on the arrow (Monga 2001,Sterman 2000). This indicates whether there is an increasing (+) or decreasing (-) relationship between two variables or there is delay (≠)
- An influence also has a direction, indicated by an arrow, and an indicator as to whether the influenced element is changed in the same (S) or opposite (O) direction as the influencing element.

Causal loop diagram components

- Causality: A→B
  - Changes in B are *caused* by changes in A
- Polarity: + or –
  - If A and B change in the same direction (either both increase or both decrease), then polarity is +
  - If A and B change in opposite directions, then polarity is –
- Reinforcing feedback loop (vicious or virtuous) – amplifying effect
- Balancing feedback loops – dampening effect

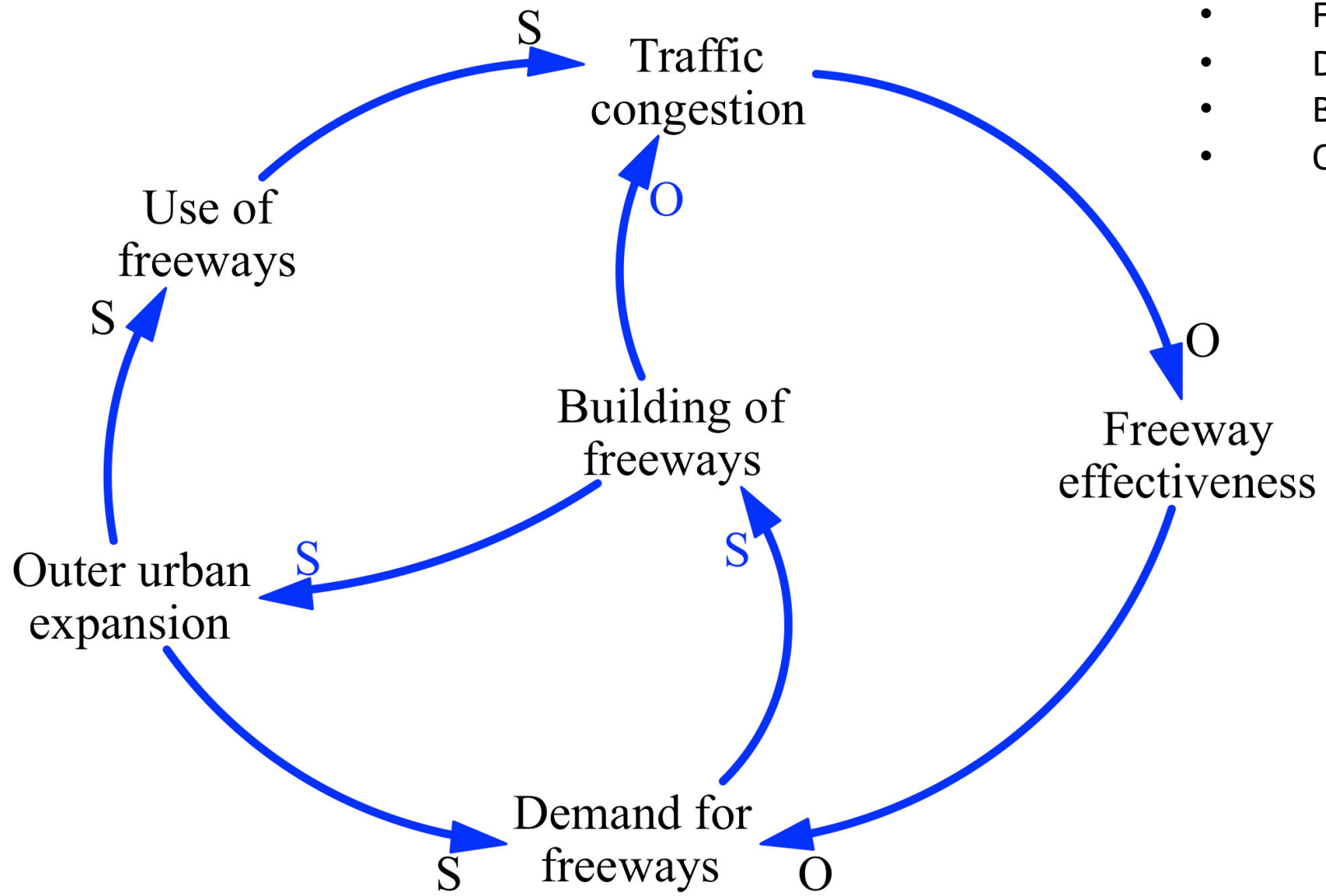


Balancing or Reinforcing loops ?



Think in terms of cause-and-effect and feedback loops

Transport Causal Loop Diagrams

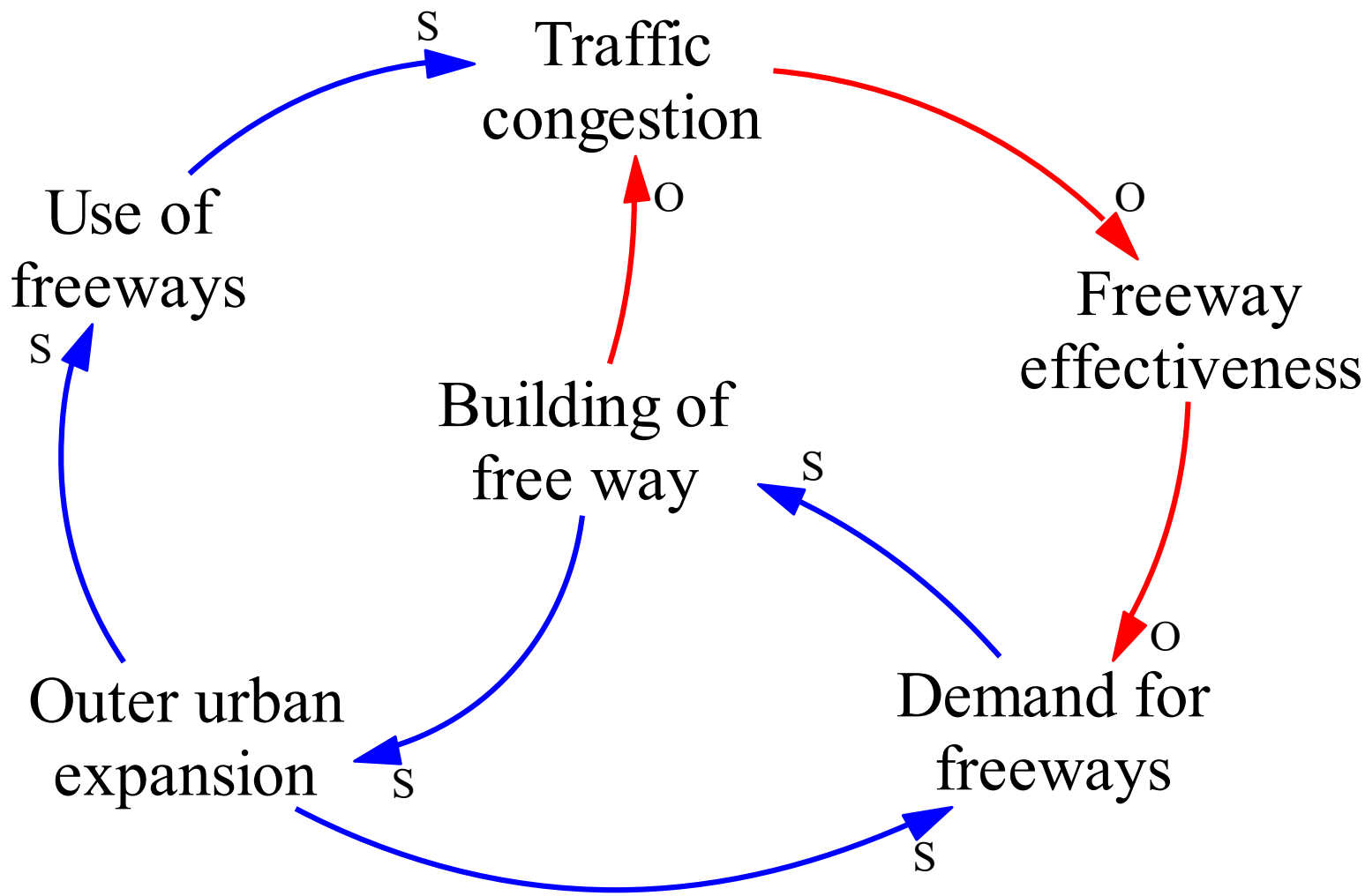


Loop Number 1

- Traffic congestion
- Freeway effectiveness
- Demand for freeways
- Building of free way
- Outer urban expansion

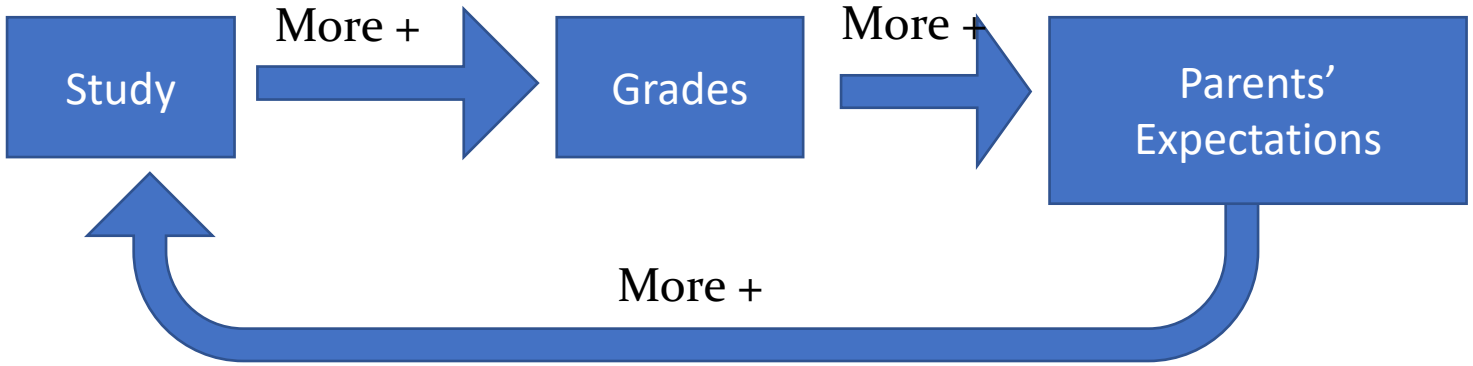
Vensim is simulation software developed by Ventana Systems. It primarily supports continuous simulation (system dynamics), with some discrete event. It is available commercially and as a free "Personal Learning Edition".

Trasport Causual Loop Diagram



Focus on Feedback Loops

situation when output from an event will influence the same event in the future



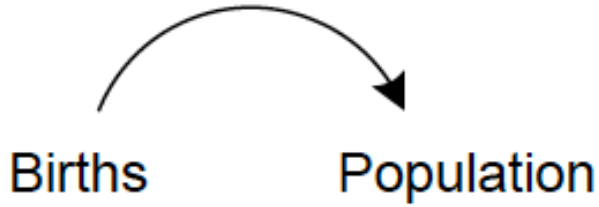
(Step 1)

Place variables

Births Population

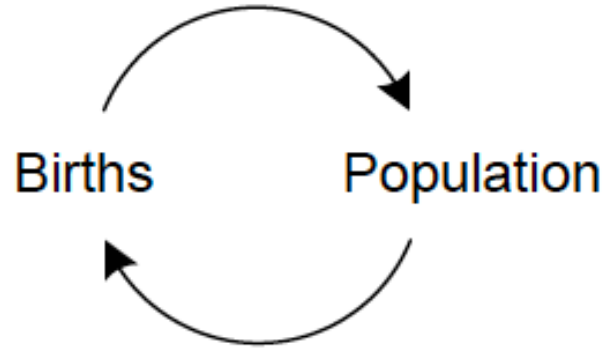
(Step 2)

Determine causality



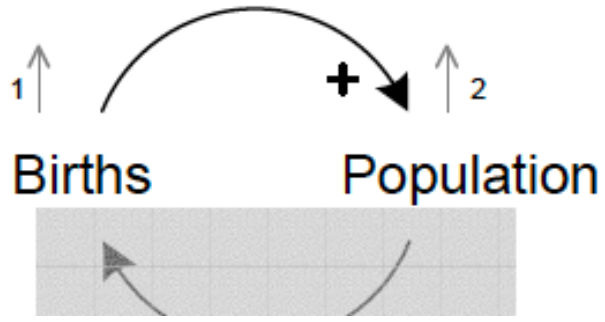
(Step 3)

Is there a link back?



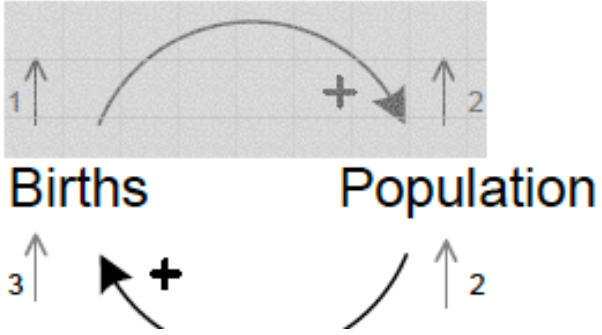
(Step 4)

Write polarity for first link



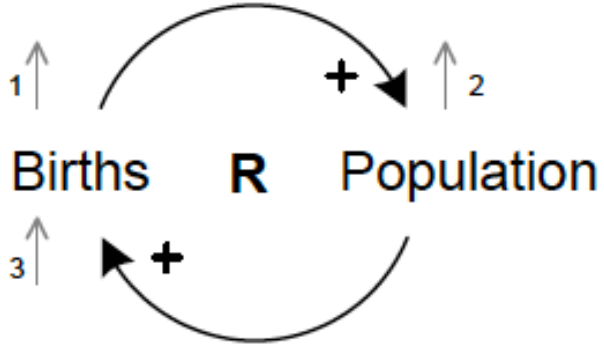
(Step 5)

Write polarity for second link, the feedback



(Step 6)

Write the loop behaviour



(Step 1)
Place variables

Industry

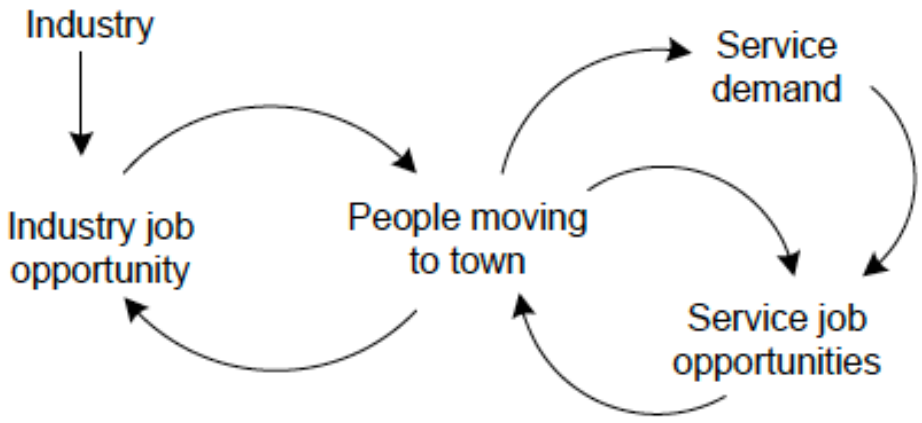
Service demand

Industry job opportunity

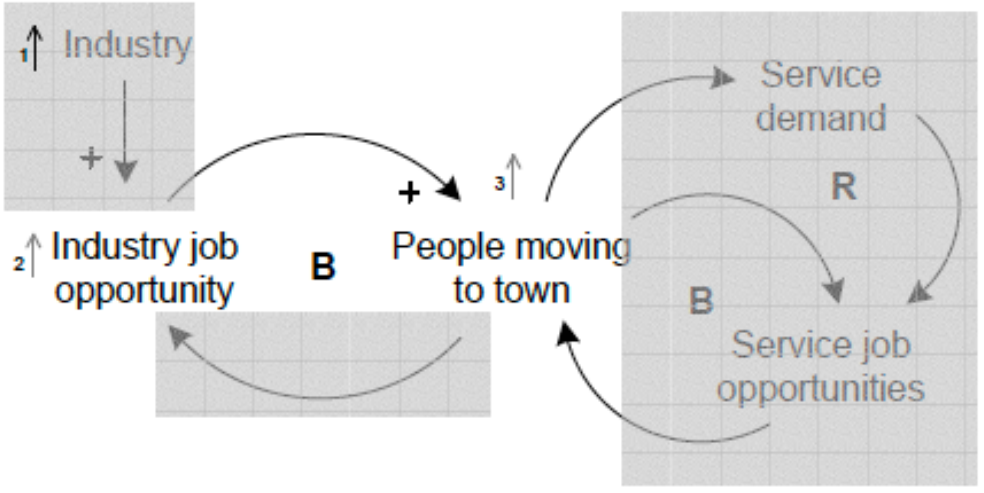
People moving to town

Service job opportunities

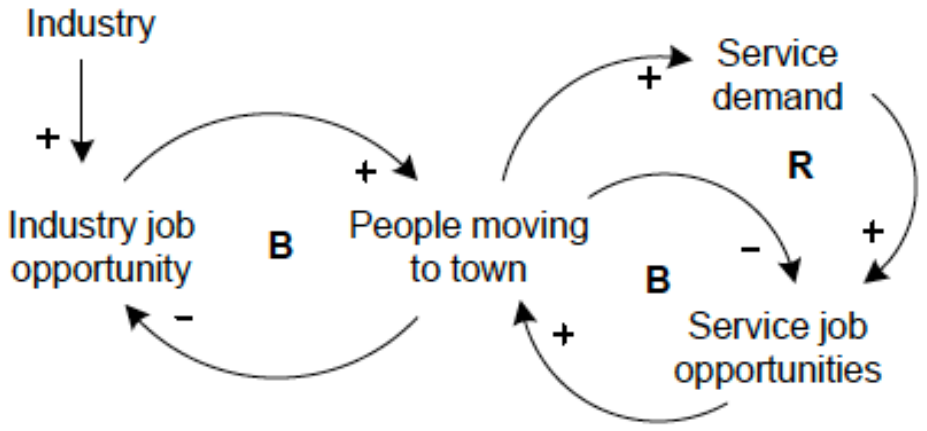
(Step 2
Determine the causality and the feedbacks



(Step 3
Write polarity for each link, one at a time

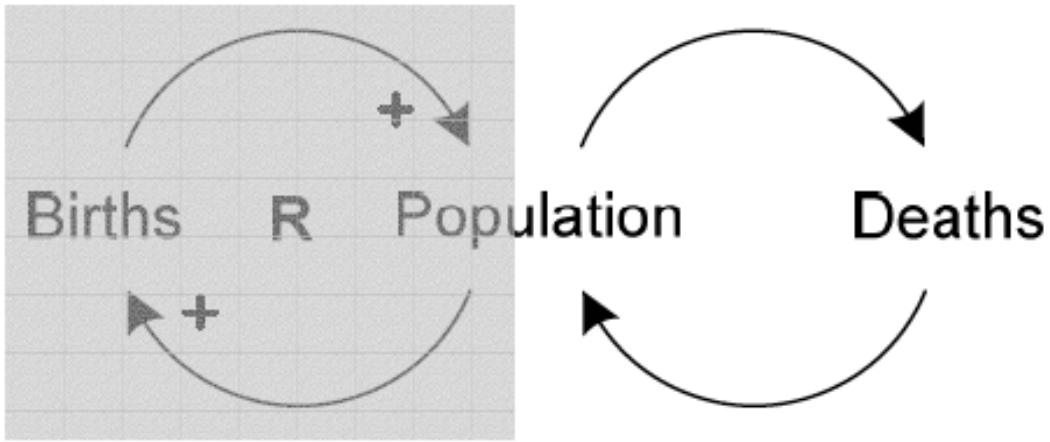


(Step 4
Write the loop behaviour



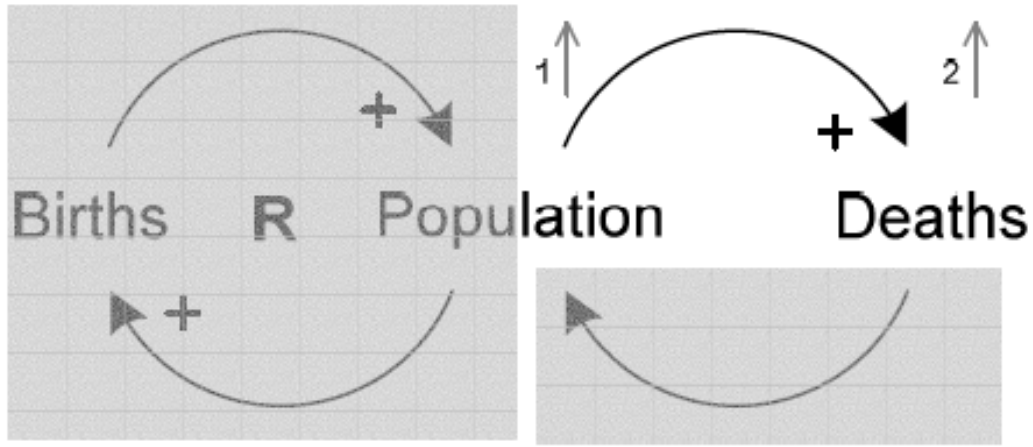
(Step 1)

Determine the causality links of the second loop



(Step 2)

Determine the polarity for the first link

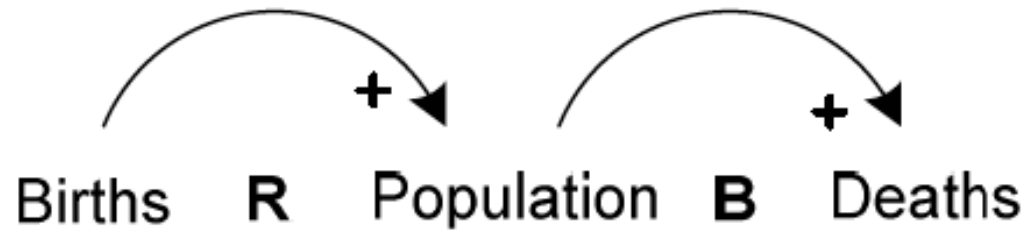


(Step 3)

Determine the polarity for the feedback

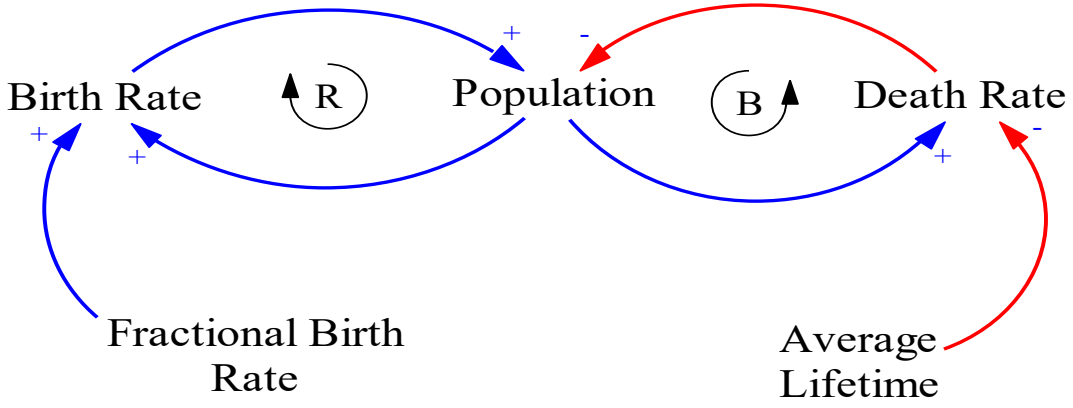
(Step 4)

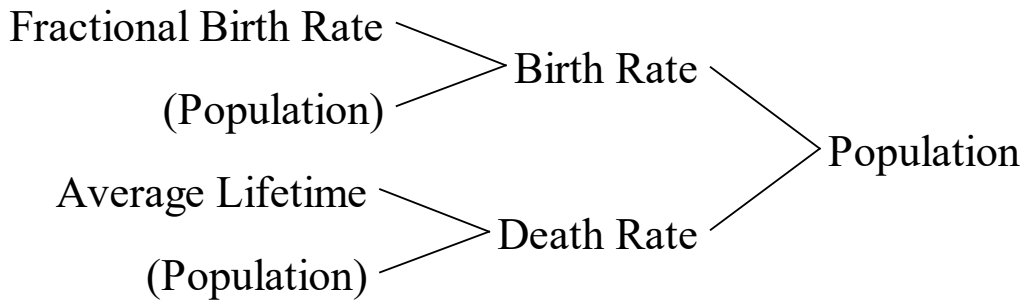
Write the behaviour



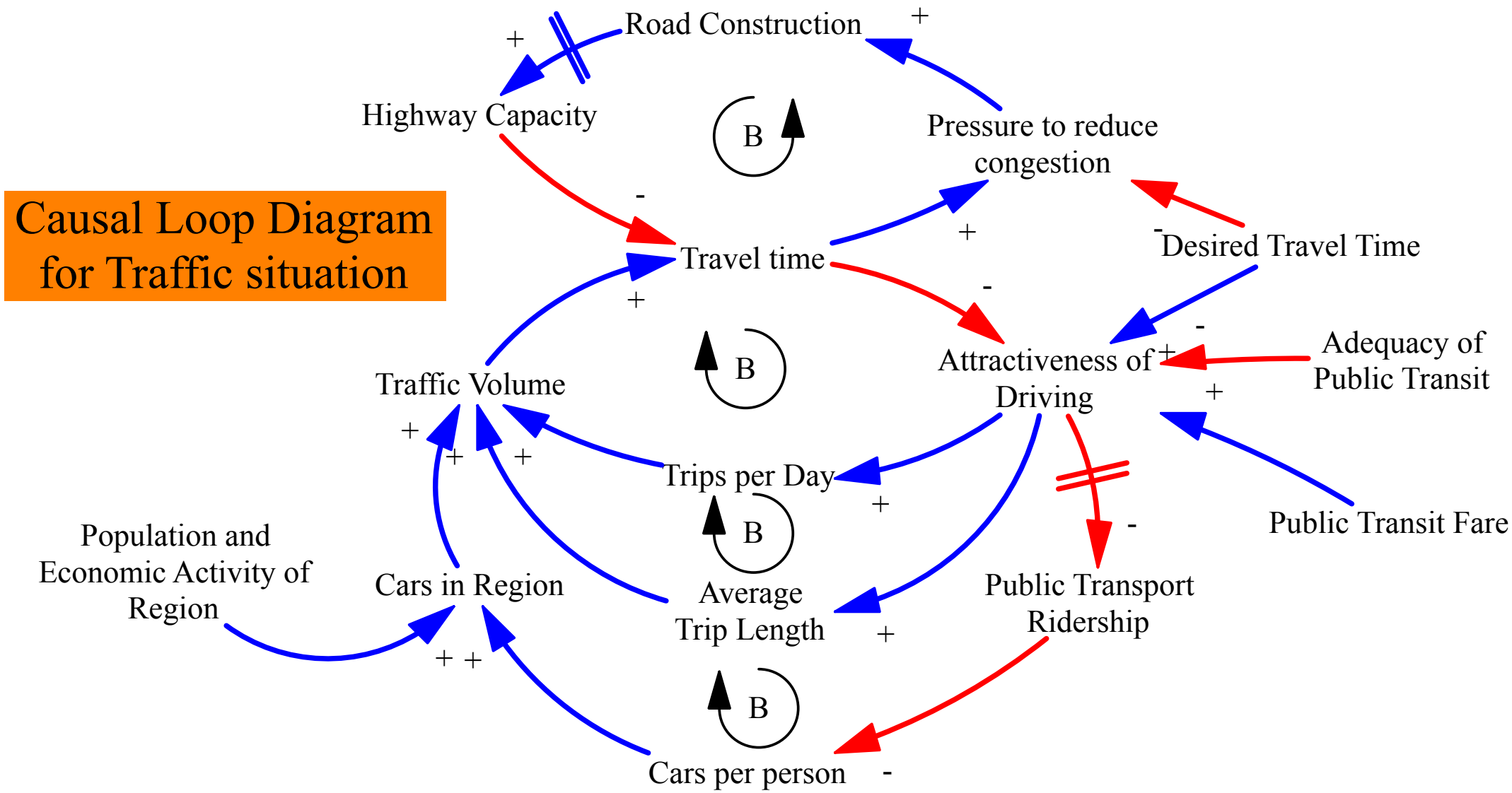
Drawing Causal loop Diagram in Vensim

Causal loop Diagram for population model

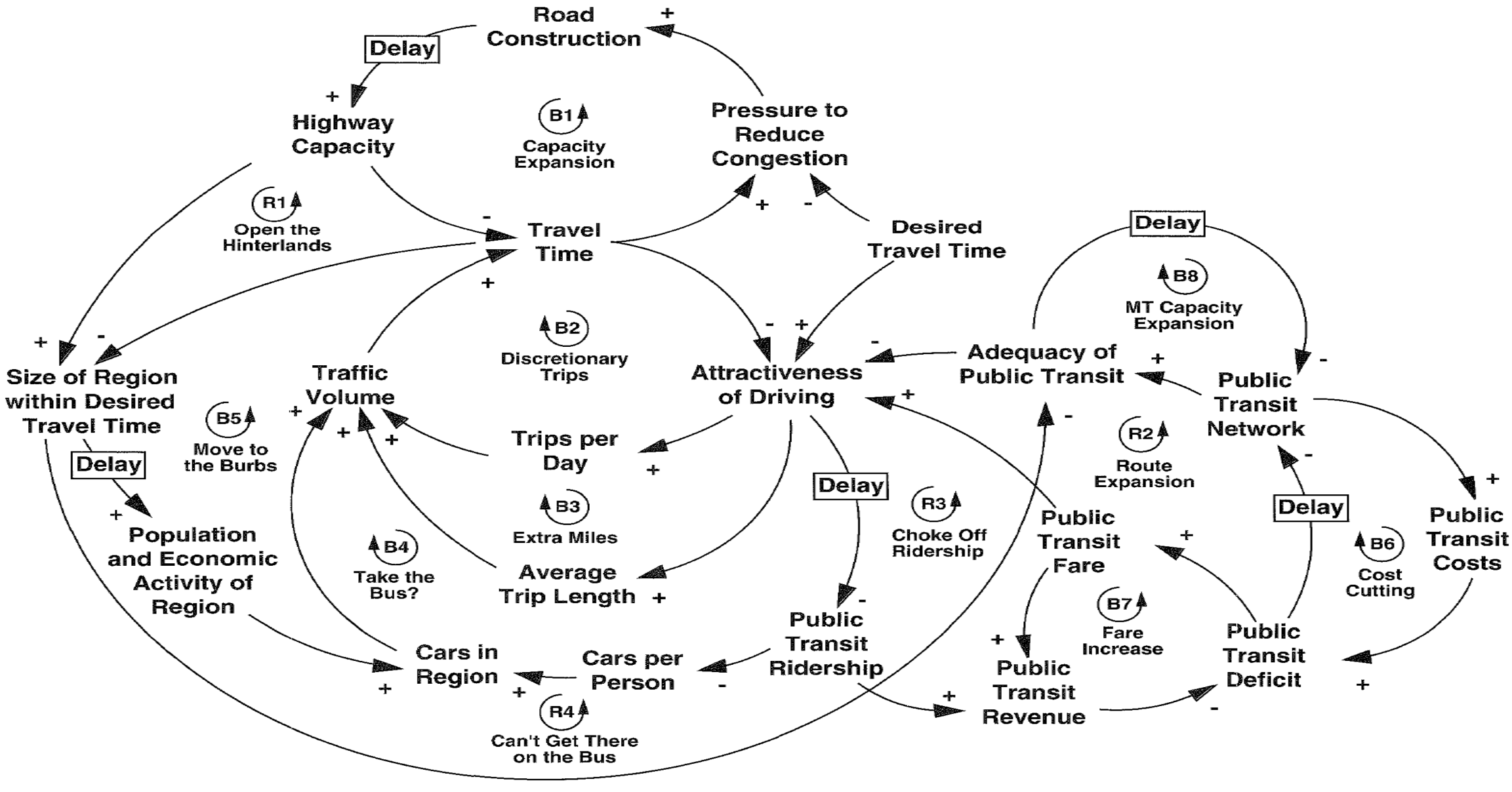




Causal Loop Diagram for Traffic situation



Traffic volume depends on congestion, closing several negative loops that cause traffic to increase whenever new roads are built.
Sterman,2001

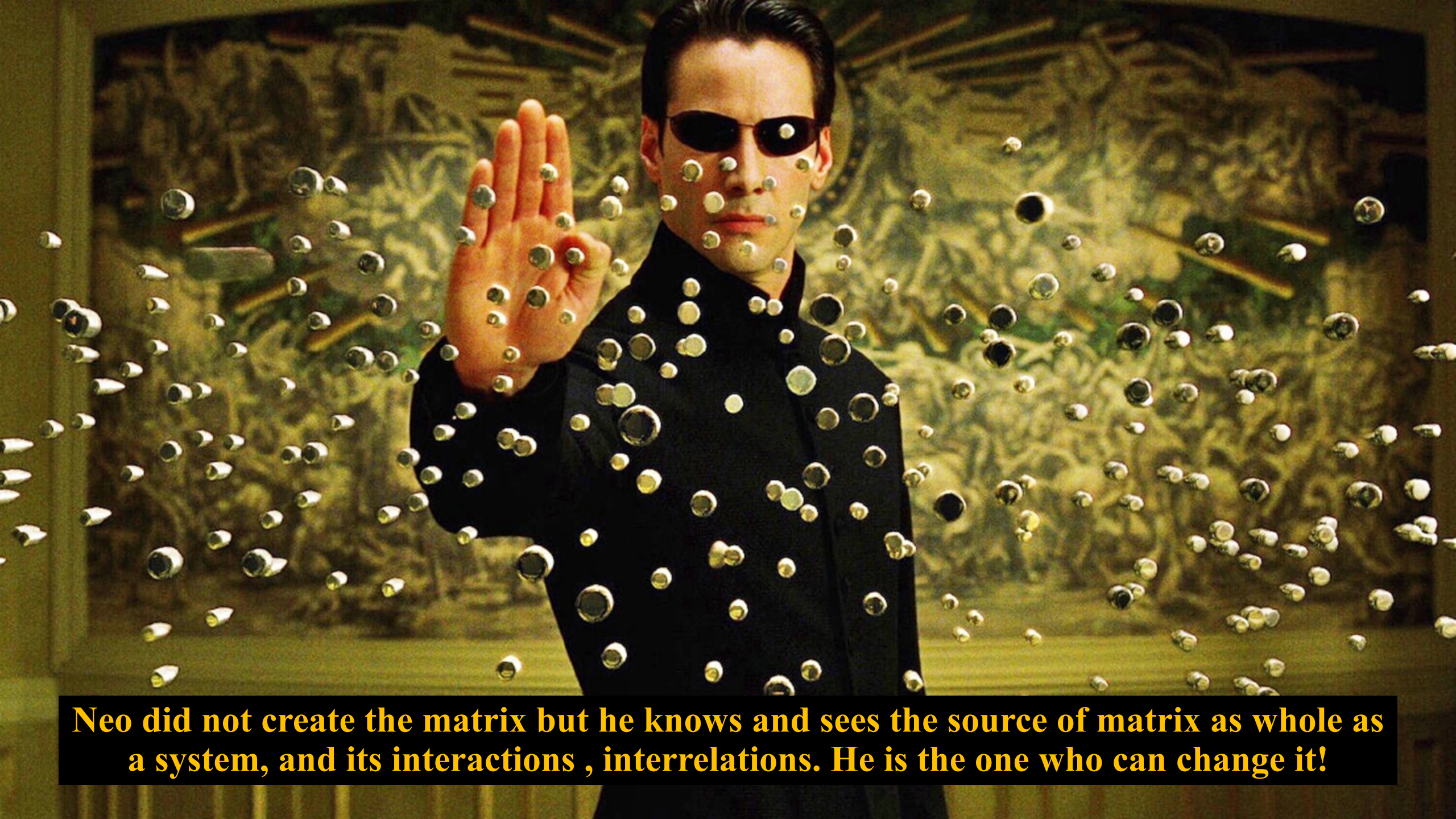


You can't get there on the bus.

As the size of the populated region expands, the adequacy of public transit declines. The result is more driving, more congestion, and still more road construction rather than an expansion of the public transit network.

Architect created matrix therefore he knows about interactions , interrelations, patterns and he predicts the system behaviors

Key-maker: He knows about the key of each door and he can copy the keys very well. He focuses on each part of the system.


Neo did not create the matrix but he knows and sees the source of matrix as whole as a system, and its interactions , interrelations. He is the one who can change it!