Data Analytics
22 Feb 2022 @ Dr. MCR HRD Institute

Prof. Rajib L. Saha (Rajib_Saha@isb.edu)
Assistant Professor of Information Systems, ISB
https://www.isb.edu/en/research-thought-leadership/faculty/faculty-directory/rajib-saha.html

Hyderabad

Mohali

About me …

Education

- B.Tech. in Computer Science and Engineering, IIT Kharagpur
- MS and PhD in Business Administration, University of Rochester, New York

Academic (2012 – present)

- Faculty at the Information Systems area at ISB since 2012

Industry (2000-2005)

- Software product development
  - Cal2Cal, Novell, Oracle

Research

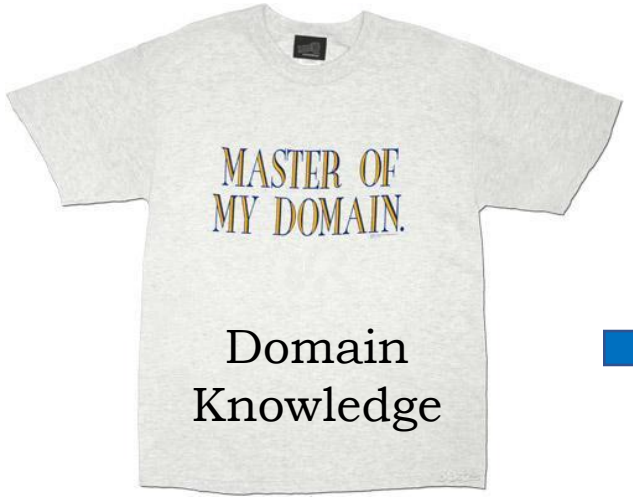
- Applied Economics – Economics of IT and ITeS
- Business Analytics and Data Mining

Teaching

- **Indian School of Business (current)**
  - Economics (Information Technology) and Analytics based courses
    - PGP – Full-time MBA
    - PGPPro – Part-time MBA
    - AMPBA – Business Analytics Masters
    - Custom Programs
- **University of Rochester, NY (during PhD)**
  - Mathematics and Statistics based courses
    - MBA and PhD programs

ISB

Data Analytics in a Nutshell

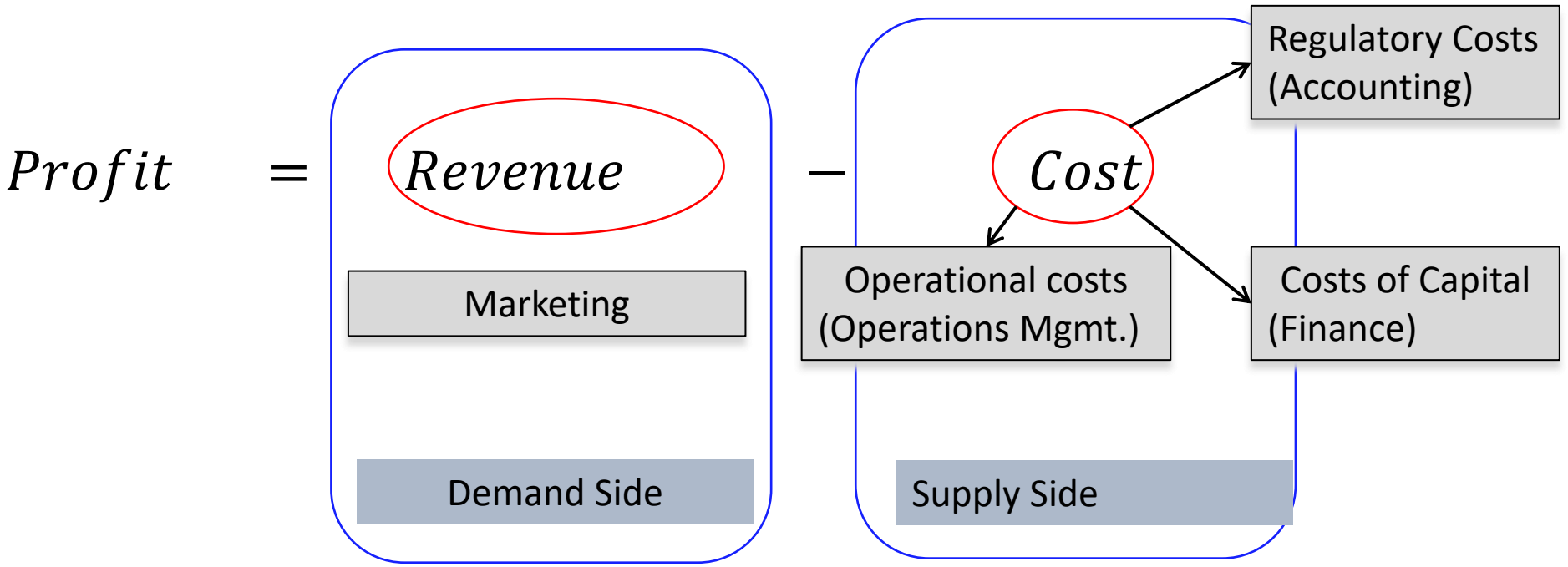


Economics | Statistics | Mathematics | Computer Science | Computer Engineering

Jack of all & master of ~~SOME~~ the DOMAIN

What is your Busines Hat ?

The Objective of a Private Organization:
Maximize (economic) profits



- Market power derives from competencies on the demand and/or the supply side.

The Objectives of the Government:
Maximize Social Welfare

- $Profit\ or\ Producer\ Surplus\ =\ Revenue - Cost$

- $Consumer\ \ Surplus\ =\ Value\ - Price$

$$\begin{pmatrix}\text{Net Societal}\\ \text{Welfare}\end{pmatrix} = \begin{pmatrix}\text{Consumer}\\ \text{Surplus}\end{pmatrix} + \begin{pmatrix}\text{Producer}\\ \text{Surplus}\end{pmatrix}$$

- There is a *tradeoff* between consumer and producer surpluses.
- Extent of control by government gives us different systems.

Technology

Data

Analytics

- **TECHNOLOGY** enables and gives rise to different business and operating models.
  - E-Commerce
  - Gig economy or sharing economy
  - Fintech
  - …
- Digitization enables collecting more footprints
  - More **DATA** with government and firms
- **ANALYTICS** create values across functions in firms and government organizations using Data.

Why do we need Data for better decision-making?

- Human cognition has limitations
- We are affected by our perceptions
  - Difficulty to isolate a problem
  - Tendency to use a narrow range of solutions
- We have bias - we shape responses based on memory, stereotypes
  - So, decision making is prone to serious inaccuracies and errors

Why do we need Analytics to understand Data?

- More data can lead to information overload
  - But not if we understand how to manage and process it.
- To overcome cognitive limitations, perception, and bias
- Because decisions are hard to make
- The number of alternatives keeps increasing
- Decisions must be made under time pressure
- Decisions are more complex

Data Vs. Big Data

What is Big Data?

4 Pillars of Analytics

DESCRIPTIVE

What happened?

exploratory analysis · visualization
BI · dashboards

PREDICTIVE

What will happen next?

data mining · machine learning model fitting · forecasting

CAUSAL

How do inputs impact outcomes?

a/b testing · econometrics

PRESCRIPTIVE

How should we respond?

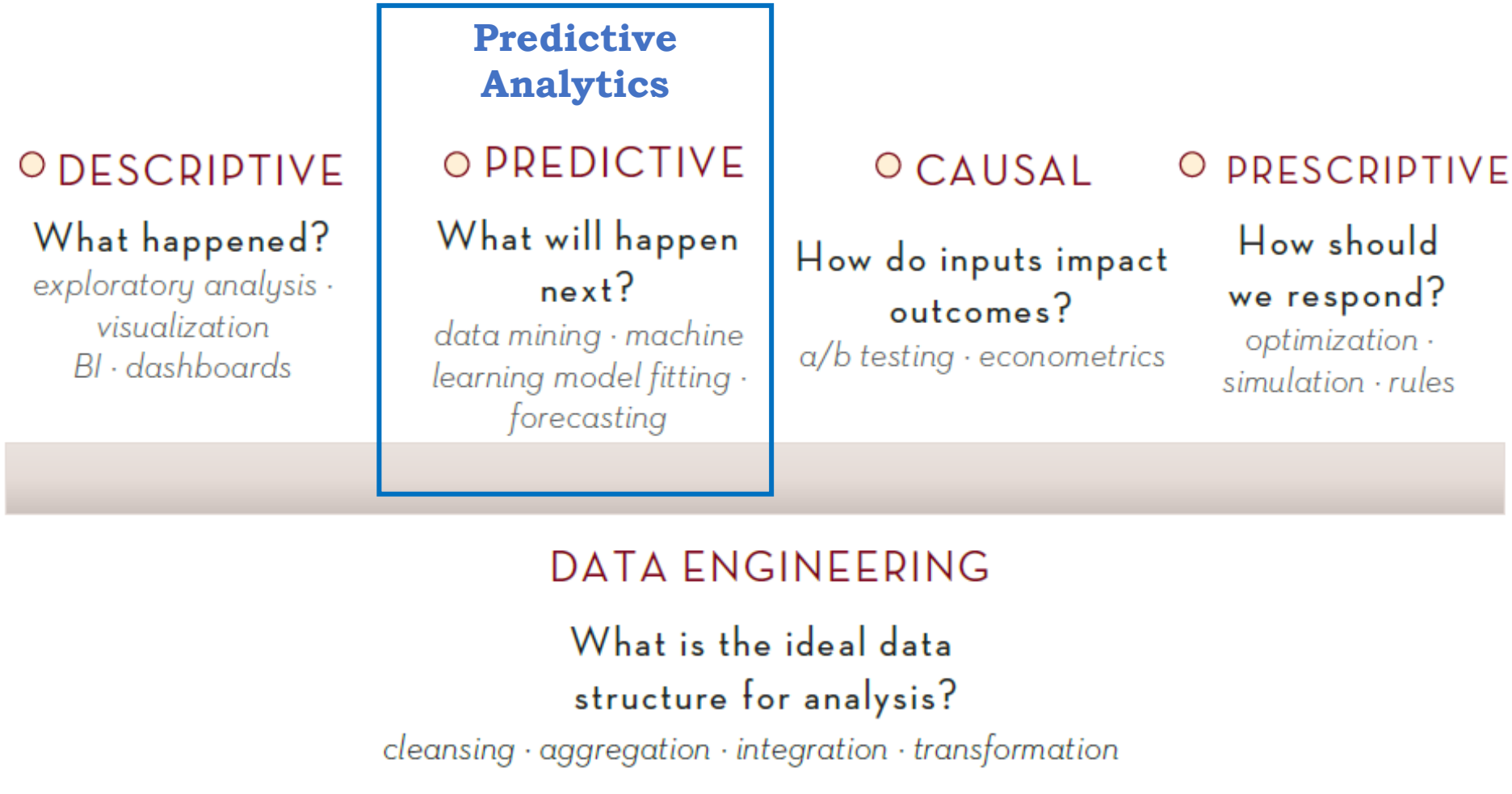
optimization · simulation · rules

DATA ENGINEERING

What is the ideal data structure for analysis?

cleansing · aggregation · integration · transformation

4 Pillars of Analytics



How Machine Learning is Being Used to Predict Train Delays in India

Category: Machine Learning

Multiple variables associated with the run of a train can affect the arrival time of a train at the station and passengers are most often left waiting for hours before their train finally arrives. The result; unending anxiety of passengers, many many man hours wasted and unnecessary congestion at all the stations.

RailYatri, a travel start-up, has innovated a unique Estimated Arrival Time (ETA) prediction algorithm using Machine Learning and Statistical Modelling techniques to predict the arrival time of running trains at their upcoming stoppage with much better precision. The algorithm has been trained to analyse historical data of train runs spread over many years and predict the future outcome.

Train delays can safely be considered part and parcel of train travel in India giving the current delay trends, but what bothers the travellers most is the uncertainty around their train travel. Surveys show that while train travellers have submitted to delays being part of their travel, their frustration arises from the inability of the existing systems to correctly guide them on the estimated time of arrival (ETA) of their trains. This leaves them waiting endlessly at platforms without any idea of the exact time of arrival of their trains.

According to **Kapil Raizada, Cofounder of RailYatri** - *"The existing method to predict the ETA of trains in India have not changed over decades and is typically based on the 'distance divided by speed of the train added with some buffer time for safety formula. We believe that a much better technique is to make the ETA prediction based on historical data as it takes proper considerations of ground realities such as increasing traffic, rush, seasonality, etc. Our ETA prediction algorithm is highly adaptive*

Now, IRCTC will predict your wait-listed tickets will get confirmed or not

The refurbished IRCTC website went live at midnight. Know the top 5 facts about the new site that will help commuters struggling to book train tickets

Topics
Irctc

BS Web Team | Agencies | New Delhi
Last Updated at May 29, 2018 09:04 IST

- It lets Indian Railways passengers know the probability of confirmation of wait-listed tickets based on a new algorithm developed by the **Centre for Railway Information Systems (CRIS)**.
- The algorithm used past 13 years' data as of 2018.
- Nearly 1.3 million tickets are booked on the IRCTC website every day against a reserved accommodation of 1.05 million berths.

Which taxpayer warrant an audit?

FEDERAL TIMES
A GANNETT COMPANY

SEARCH

PAY & BENEFITS | PERSONNEL | AGENCY NEWS | TRAVEL | IT | ACQUISITION | FACILITIES, FLEET & ENERGY | CONGRESS | ADVICE & OPINION | DEPARTMENTS | FORUMS

Register for free Federal Times E-Newsletters

- Weekly highlights from print
- Daily round-up of top govt. news
- Monthly topic-specific reports

Software that predicts the future: Does it really help?

By SEAN REILLY | Last Updated: September 18, 2011

Like 4 Tweet 32 SHARE

For the IRS, a type of software known as predictive analytics may offer one means of deciding which taxpayers most likely warrant an audit. The U.S. Postal Service's inspector general has just begun using analytics to identify high-risk contracts. And at least one agency is exploring the technology as a means of predicting which of its employees will be retiring soon.

The approach, heavy on mathematical modeling and long employed by business, is drawing increased attention from government, both to boost efficiency and as a way of making better use of its vast stocks of data to help predict the future.

Shelley Metzenbaum, associate director of

How MassHealth cut Medicaid fraud with predictive analytics

By Rutrell Yasin

FEBRUARY 24, 2014

Predictive analytics system based on NetReveal fraud detection protects MassHealth from fraud and unnecessary expenditures.

Medicaid is the USA's public health insurance program for people with low income.

Vedanta bets on AI to peek into the future

Vedanta resources has identified various business challenges, which could be solved through predictive analytics and AI.

Riya Pahuja • ETCIO • September 30, 2021, 09:24 IST

"In our oil and gas business, we have wells that produce oil. And each of these wells has instrumentation and sensors installed that report how the well is flowing, what is the temperature and pressure at the surface, and a few other parameters. Because of a simple rule of return on investment, we may not have invested in installing sensors in the low oil-producing capacity wells," said Anand Laxshmivarahan, Chief Digital Officer at Vedanta Resources.

"We had two options: either to invest money and then put those sensors downhole so we can get the information needed or to use the data from the other wells. We used that data from the wells which had those sensors to build a supervised machine learning model. We used the information that we had in those wells to train a model that would predict the bottom hole flowing pressure for the rest of the wells which did not have the sensors. With this, we have applied predictive analytics and data to get visibility on how wells are flowing at the bottom."

Source: https://cio.economictimes.indiatimes.com/news/next-gen-technologies/vedanta-bets-on-ai-to-peek-into-the-future/86634142

THE MAGAZINE

October 2012

Buy Reprint »

Big Data: The Management Revolution

by Andrew McAfee and Erik Brynjolfsson

Comments (1)

SPOTLIGHT ON BIG DATA

Expertise from Surprising Sources

Often someone coming from outside an industry can spot a better way to use big data than an insider, just because so many new, unexpected sources of data are available. One of us, Erik, demonstrated this in research he conducted with Lynn Wu, now an assistant professor at Wharton. They used publicly available web search data to predict housing-price changes in metropolitan areas across the United States. They had no special knowledge of the housing market when they began their study, but they reasoned that virtually real-time search data would enable good near-term forecasts about the housing market—and they were right. In fact, their prediction proved more accurate than the official one from the National Association of Realtors, which had developed a far more complex model but relied on relatively slow-changing historical data.

This is hardly the only case in which simple models and big data trump more-elaborate analytics approaches. Researchers at the Johns Hopkins School of Medicine, for example, found that they could use data from Google Flu Trends (a free, publicly available aggregator of relevant search terms) to predict surges in flu-related emergency room visits a week before warnings came from the Centers for Disease Control. Similarly, Twitter updates were as accurate as official reports at tracking the spread of cholera in Haiti after the January 2010 earthquake; they were also two weeks earlier.

Intrusion detection at the border, in the airspace

- Remember Pearl Harbor?
- Classifying radar signal into noise vs. enemy plane
- Video and Image Analytics along the LAC to detect enemy movement
- Big-Data counter part of what CCTV can do.

Defeating Crimes and Cybercrimes

Using Predictive Analytics Software To Head Off Crime

Technology enables British police to proactively stop a crime before it occurs.

May 17, 2010 at 3:24pm

Who will commit juvenile crime?: Florida purchases 'predictive' software to identify at-risk youth

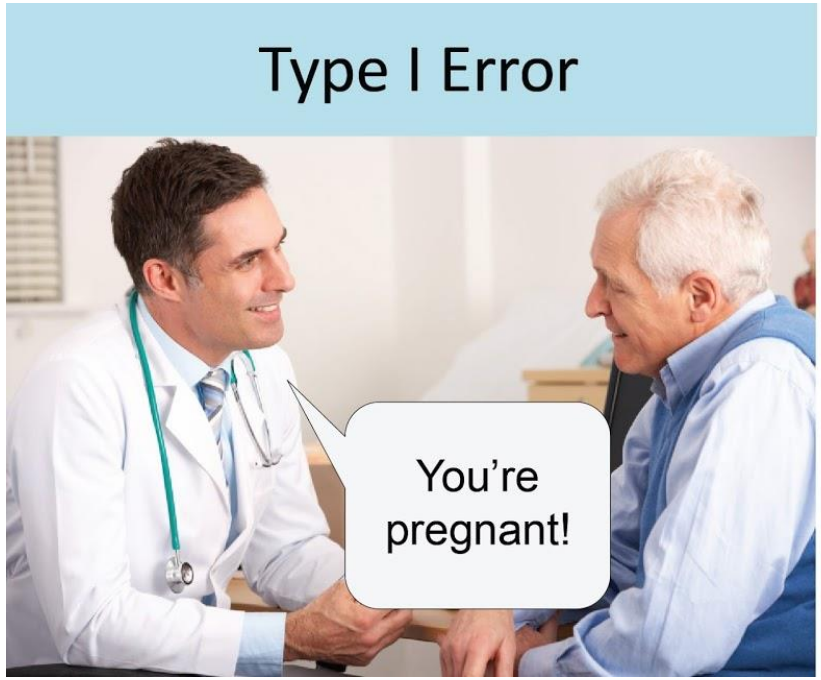
Who will commit juvenile crime?: Florida purchases 'predictive' software to identify at-risk youth
Mike Clary, Sun Sentinel, May 16, 2010

Is it possible to identify young offenders most likely to commit a crime in the future, and then intervene to stop it?

Yes, answer officials at the state Department of Juvenile Justice in touting a new system of "predictive analytics" that would steer at-risk juveniles to specific treatment programs designed to keep them from becoming adult criminals.

But the state's purchase of the $15,000 software package from IBM has alarmed some juvenile-justice experts who fear the program could unfairly label individuals and target minorities. (Read the full article)

Error Types in a Classification Model



What is the cost of False Positive (FP) vis-à-vis False Negative (FN) on Covid test?

Desirability of a test that has at different phases of the pandemic?

Fewer FPs than FNs

vs.

Fewer FNs than FPs

NEWS SITES

ET CIO.com
From The Economic Times

BlackBerry. Intelligent Security. Everywhere. presents ET CISO.in DATA PROTECTION & PRIVACY Bein Data

News | Industries | Tech-Talk | Data

Security • Cloud Computing • Big Data • Tra

3rd edition

The Economic Times SPECTRU
BUSINESS & TECHNOLOGY
VIRTUAL CONCLAVE

BUSINESS | MANAGEMENT

The Algorithm That Tells the Boss Who Might Quit

Wal-Mart, Credit Suisse crunch data to see which workers are likely to leave or stay

Deutsc Global

IT News / Latest IT News / Business Analytics

How this company predicts attrition and turns things around

One of the leading technology suppliers uses HR Analytics in the organization for the full life-cycle of the employees thus increasing employee experience -- predicting the most volatile aspect of HR in the IT sector: Attrition. But how does it work?

Abhishek Sahu • ETHRWorld • November 01, 2021, 14:42 IST

HR Analytics

Now think what if the HR department can predict the second one before candidates decide to bid goodbye to the company? Robert Bosch Engineering and Business Solutions (RBEI), the R&D wing of engineering company Robert Bosch, has done so.

HR Analytics

Resume Screening model

Predict who should be further interviewed?

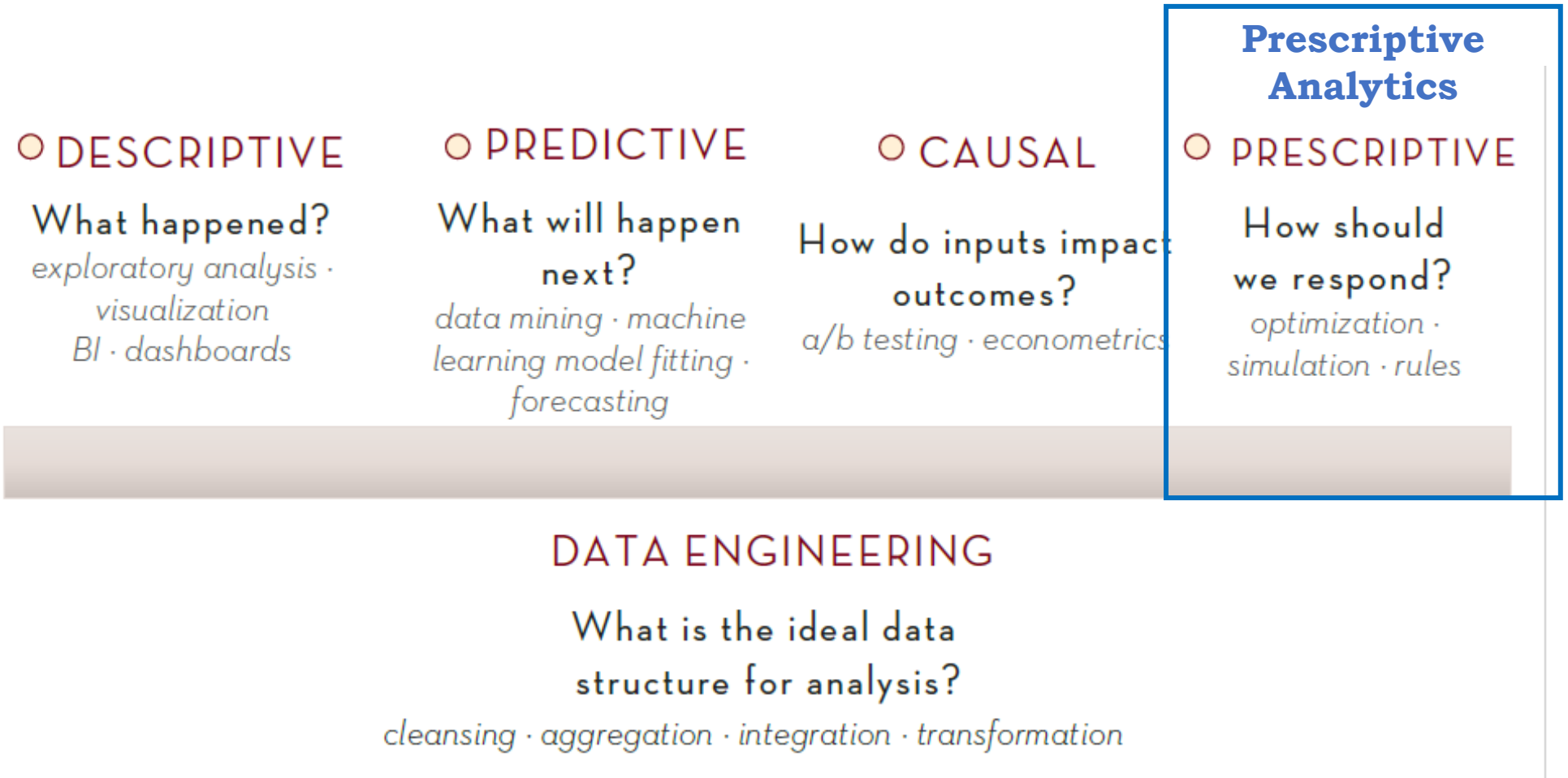
- Use Gender as a predictor?

Employee Attrition model

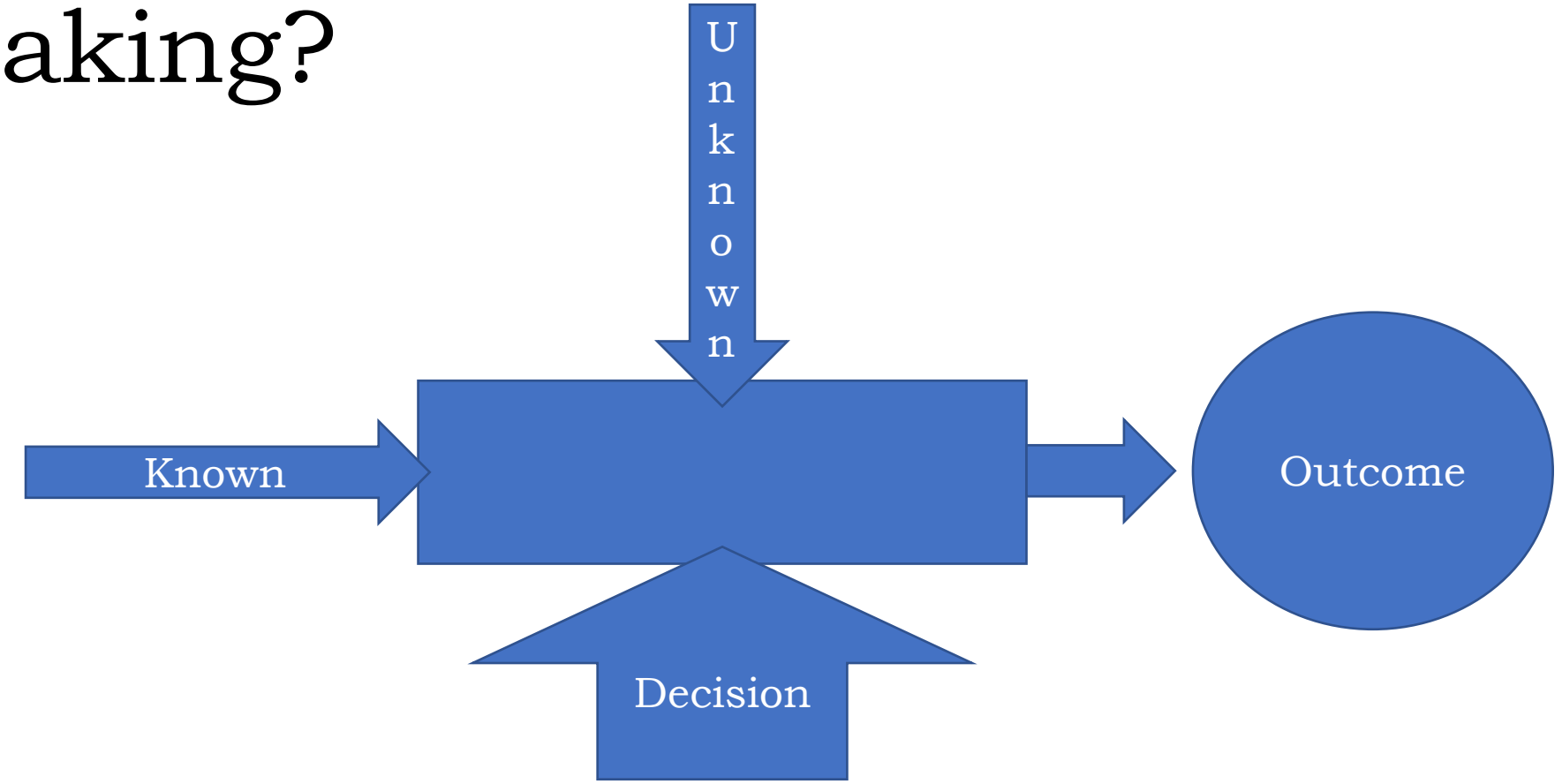
Who is going to leave and why?

- Use Gender as a predictor?

4 Pillars of Analytics



How prediction helps decision making?



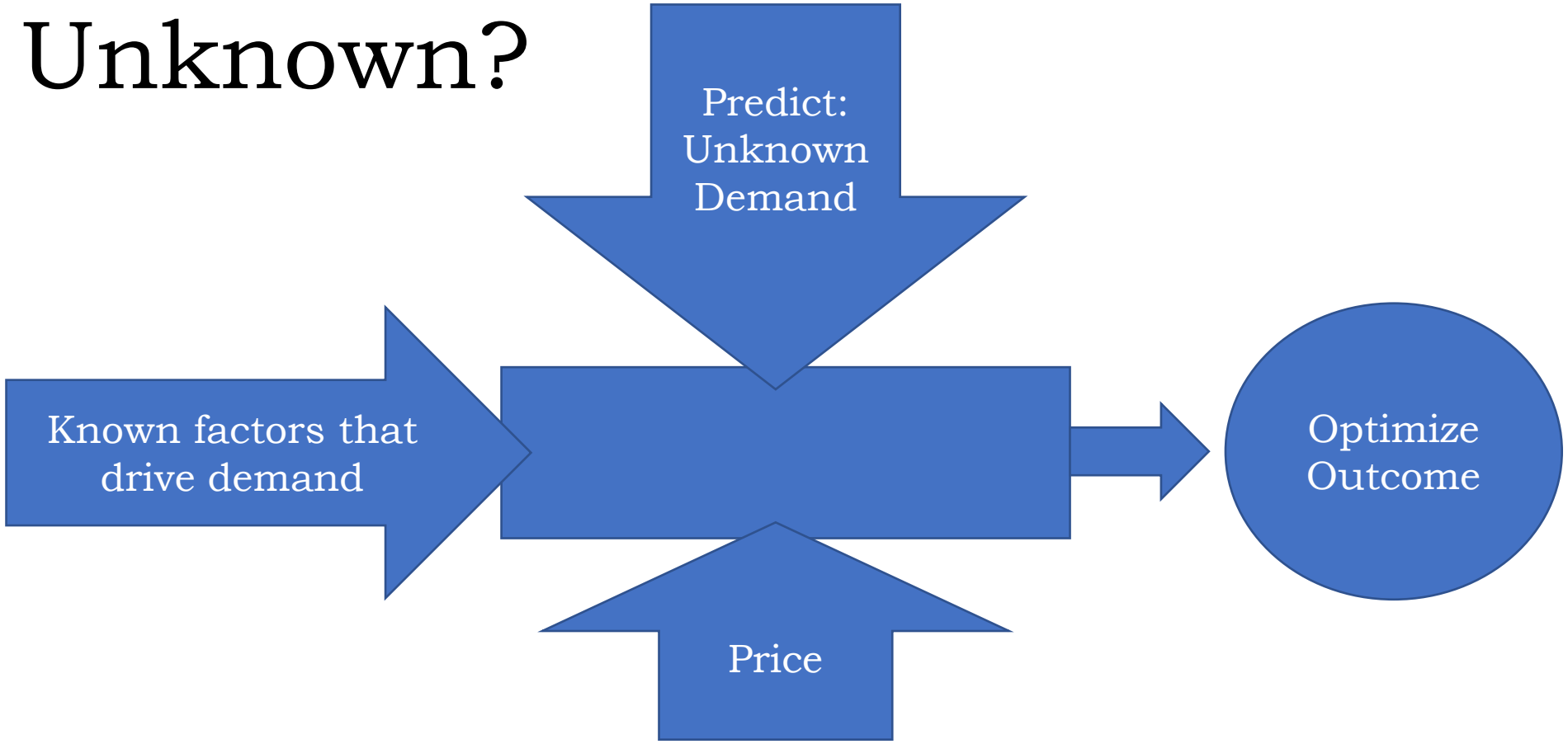
- Now, can you predict the UNKNOWN using past data?

What will you do if you had a data on performance of loans — *defaulted vs. paid in full*?

Demographic info			Loan Attributes			Defaulted	Approved
…	…	…	…	…	…	-	No
…	…	…	…	…	…	Yes	Yes
…	…	…	…	…	…	No	Yes
…	…	…	…	…	…	-	No
…	…	…	…	…	…	Yes	Yes

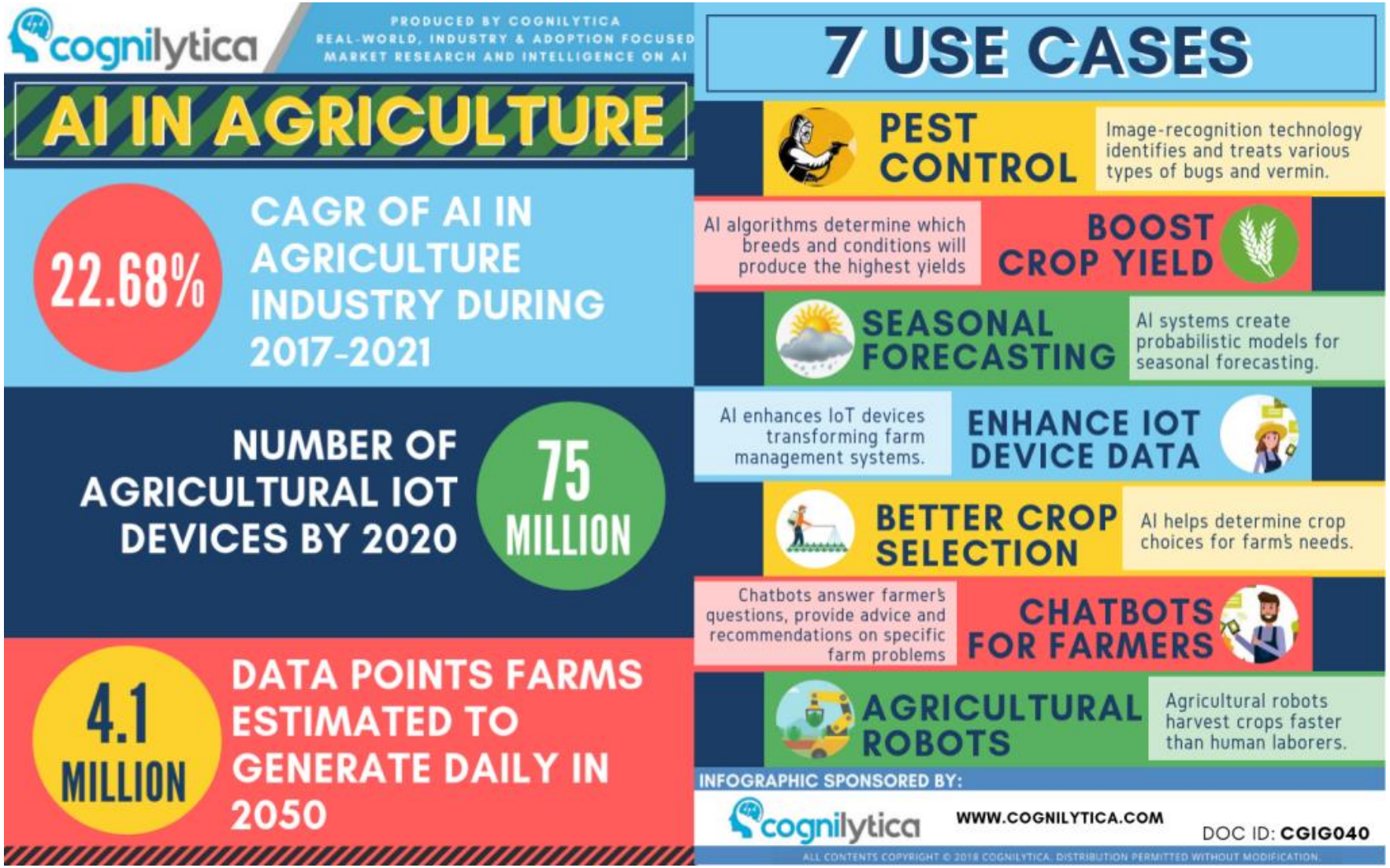
What will you predict? In order to decide what?

What if Decision also drives the Unknown?



- Now it is an **OPTIMIZATION** problem.

Prediction and Optimization together in Agriculture

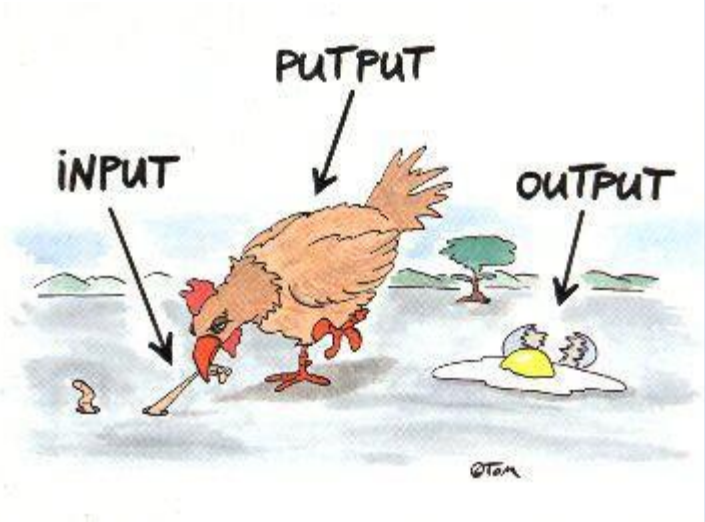


https://www.cognilytica.com/wp-content/uploads/2019/01/Wide_AI-Agri.png

- The previous problems were examples of predictive analytics
  - In some cases, it is part of a larger optimization problem

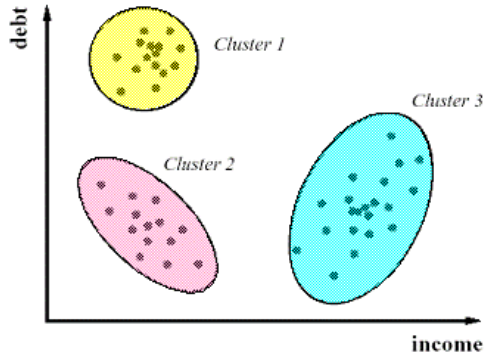
- There is another kind of data mining problems…

Supervised Learning (**Predictive Analytics**)



- Prediction (numerical Y)
- Classification (categorical Y)

Unsupervised Learning (**Descriptive Analytics**)



- Segmentation/Clustering
- Relationship Mining
- Recommender System

4 Pillars of Analytics

Descriptive Analytics

DESCRIPTIVE

What happened?

exploratory analysis · visualization
BI · dashboards

PREDICTIVE

What will happen next?

data mining · machine learning model fitting · forecasting

CAUSAL

How do inputs impact outcomes?

a/b testing · econometrics

PRESCRIPTIVE

How should we respond?

optimization · simulation · rules

DATA ENGINEERING

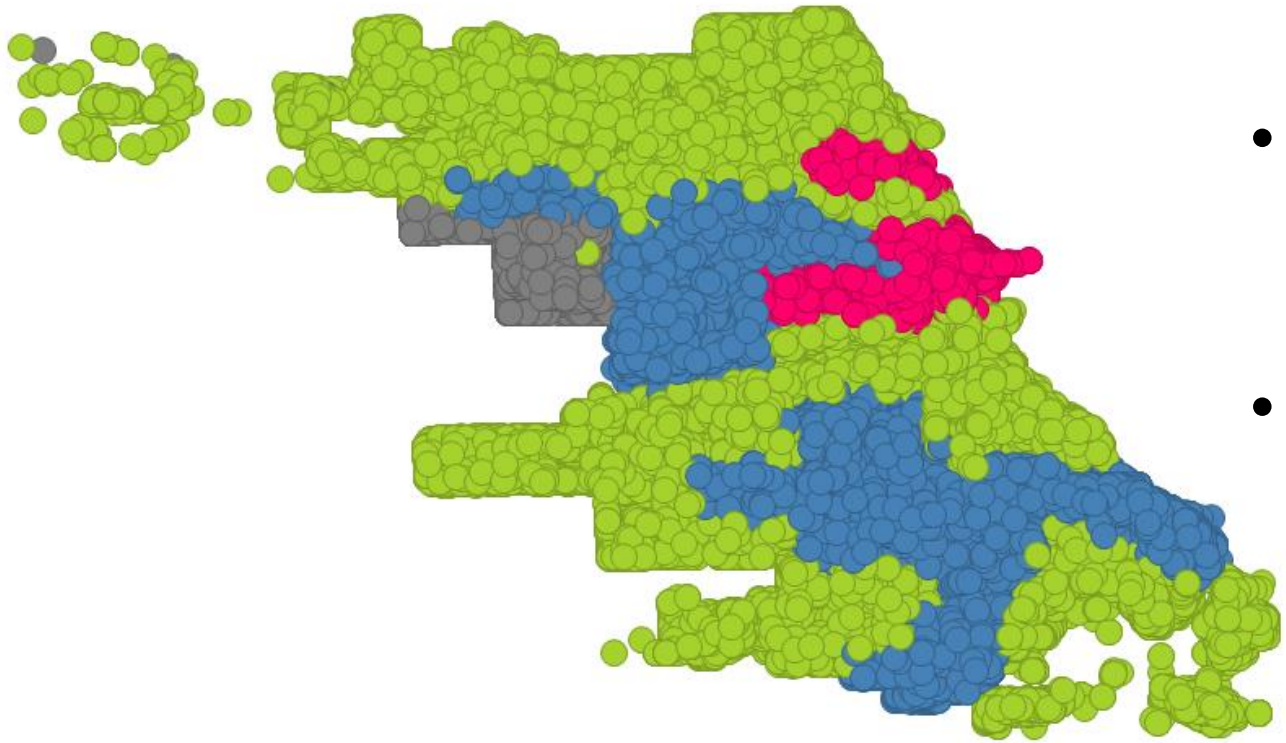
What is the ideal data structure for analysis?

cleansing · aggregation · integration · transformation

Clustering
(grouping similar tiffin boxes)

http://images.indiatvnews.com/mainnational/Mumbai-Dabbawal38721.jpg

Cluster Analysis of 77 Community areas in Chicago based on crimes



- **Violent crimes**: Robbery, Battery, Assault, Homicide, Sexual Assault
- **Property crimes**: Theft, Burglary, Motor Vehicle Theft, Arson
- **Quality-Of-Life crimes**: Criminal damage, Narcotics, Prostitution

Business & Social Values

Law & Order

Real Estate

Insurance

Tourists

Example: Fitting the Troops

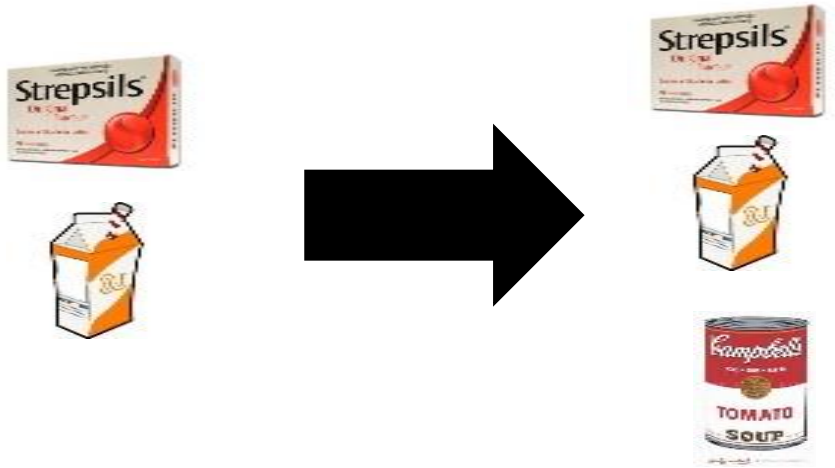
(from *Data Mining Techniques by Berry & Linoff)*

- The US army recently commissioned a study on how to **redesign the uniforms of female soldiers.** The army's goal is to reduce the number of different uniform sizes that have to be kept in inventory while still providing each soldier with well-fitting khakis.

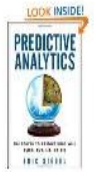
- Researchers Ashdown and Paal @ Cornell University designed a **new set of sizes based on the actual shapes of women** in the army. Unlike traditional clothing size systems, the new sizes are not an ordered set of graduated sizes where all dimensions increase together.

- Instead, they came up with sizes that fit particular **body types** (e.g., short-legged, small-waisted, large-busted women with long torsos, average arms, broad shoulders, and skinny necks).

Finding Associations



Customers Who Bought This Item Also Bought



Predictive Analytics: The Power to Predict ...
› Eric Siegel
(82)
Hardcover
$17.07

Big Data, Big Analytics: Emerging Business ...
› Michael Minelli
(9)
Hardcover
$34.15

Big Data: A Revolution That Will Transform ...
Viktor Mayer-Schonberger
(114)
Hardcover
$20.03

Too Big to Ignore: The Business Case for Big ...
› Phil Simon
(20)
Hardcover
$31.65

DEEP ROOTED .Co

LIVERING TO
Gachibowli, Hyderabad, Tel...

OS

TOP Combo

₹104

DESCRIPTION

Tomato Local(1kg), Onion(1kg) and Potato(1kg)

big basket

Search for Products...

SHOP BY CATEGORY

OFFERS

HOME > GOURMET & WORLD FOOD > PASTA, SOUP & NOODLES > PASTAS & SPAGHETTI > BB COMBO SAN REMO PASTA 500G+DISANO OLIVE

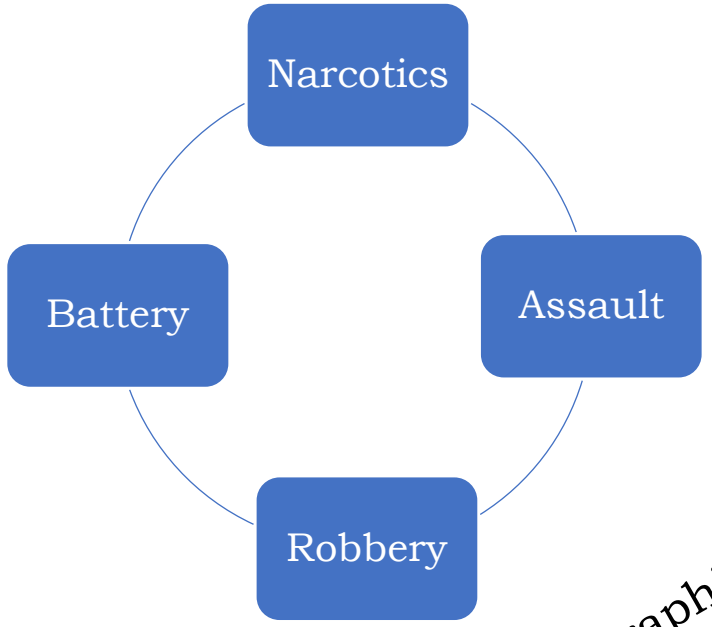
Category

& World Food
oup & Noodles
ted Noodles
ted Soups
he & Sushi Rice
s & Spaghetti

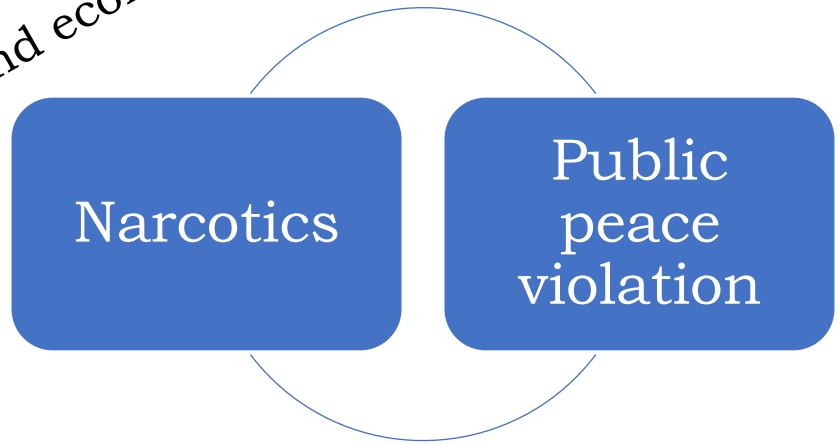
bo
bo Pastas & Spaghetti

Value Pack

Association of Crimes at District Level (Chicago crime data)



May reflect cultural, demographic, and economic differences across districts

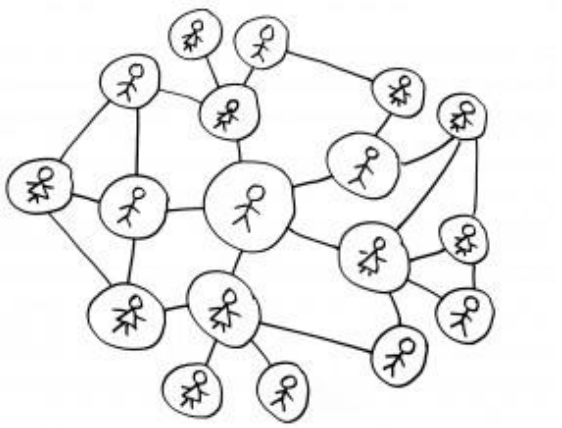


Suicide and associated cause across different sates in India

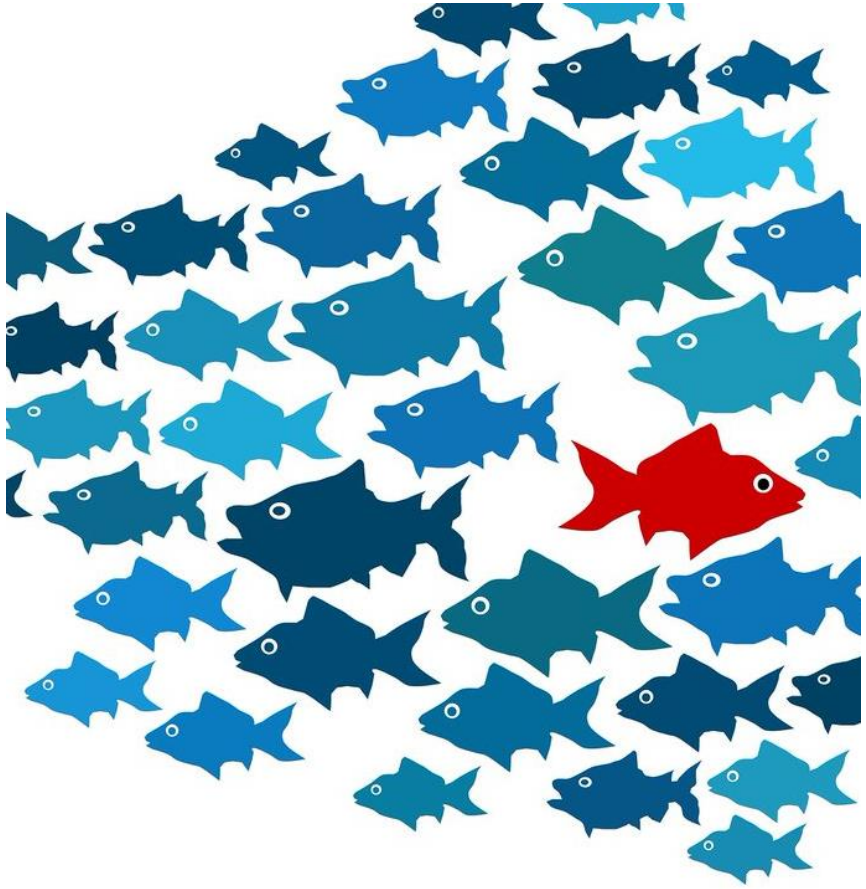
- Education gap between gender
- Climate stress on agriculture
- Mental health

Entity Resolution using transaction network

- Fraud detection by telecom companies
- Identify/track terrorist network
- Track defaulters
  - Detect customer who disappeared after accumulating debt but reappeared by opening a new account.
  - Transactional relationships of these entities with other entities remain stable.



Anomaly/Outlier Detection



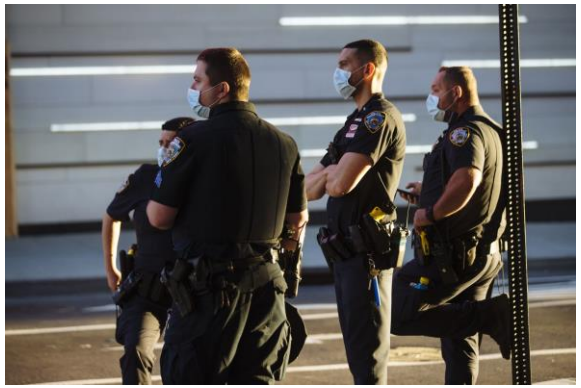
Some Applications of Outlier or Anomaly Detection

Intrusion Detection Systems

Medical Diagnosis

Event Detection – Sensors

Credit Card Fraud



Law Enforcement

Unusual Activity on Social Media

The Power of Anomaly

Burbn tried to be too many things.

- Check in to locations
- Make plans (future check-ins)
- Earn points for hanging out with friends
- Share photos
- *and much more!*

@mblongii @neoinnovate neo

"Most anomalies don't become meaningful trend, but some Do."

The Birth of Instagram

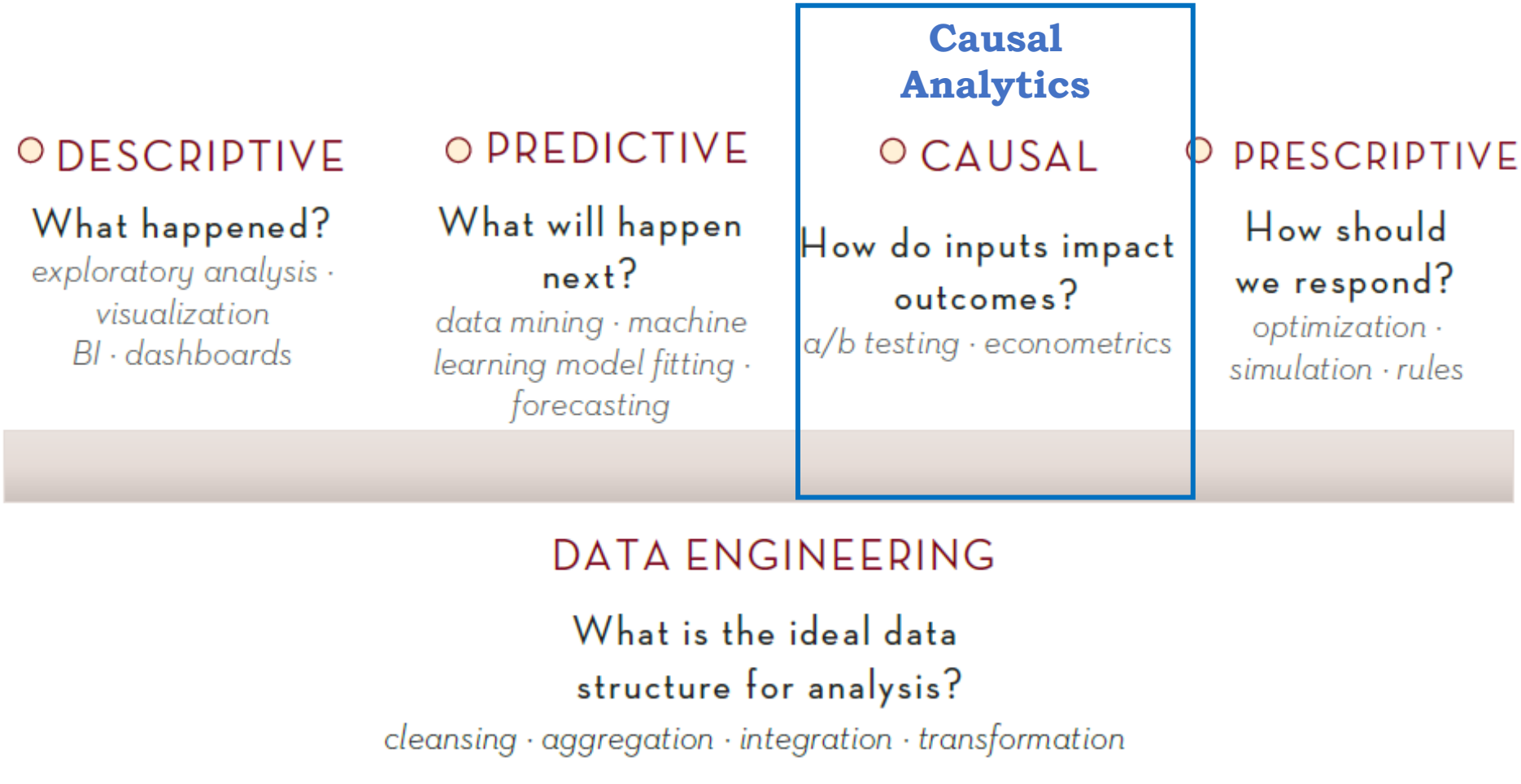
https://www.theatlantic.com/technology/archive/2014/07/instagram-used-to-be-called-brbn/373815/

The connection between Green dress of Jennifer Lopez & Google Image Search

- Searches for the green dress worn by Jennifer Lopez at the Grammy Awards in 2000 broke the record for the most popular query ever.

- It persuaded Google to introduce image search in 2001, the very next year.

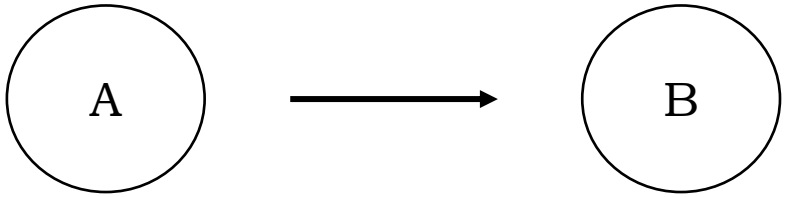
4 Pillars of Analytics



What is **CAUSALITY** (not same as Prediction)

- *The act or process of causing something to happen or exist*
- *The relationship between an event or situation and a possible reason or cause*

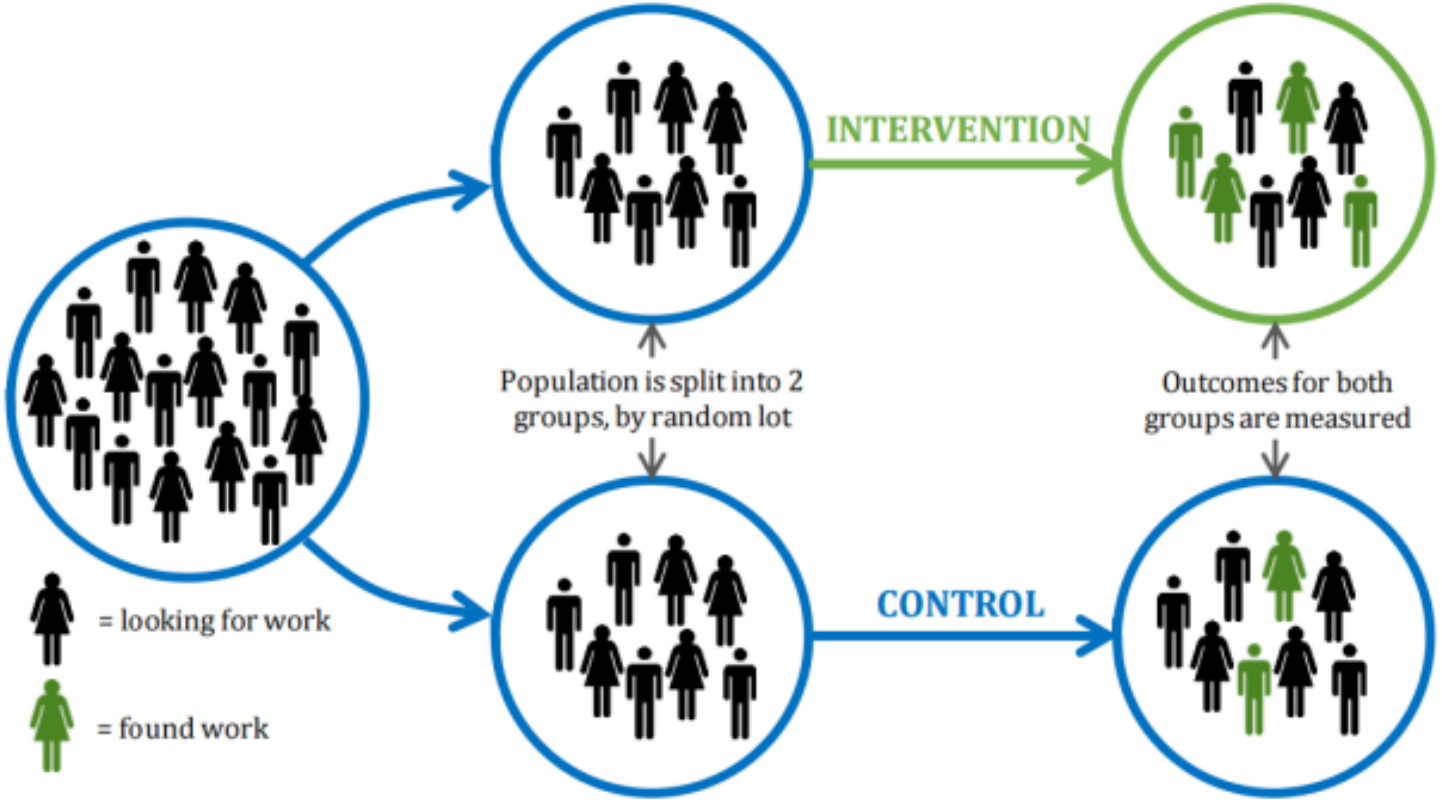
(merriam-webster)



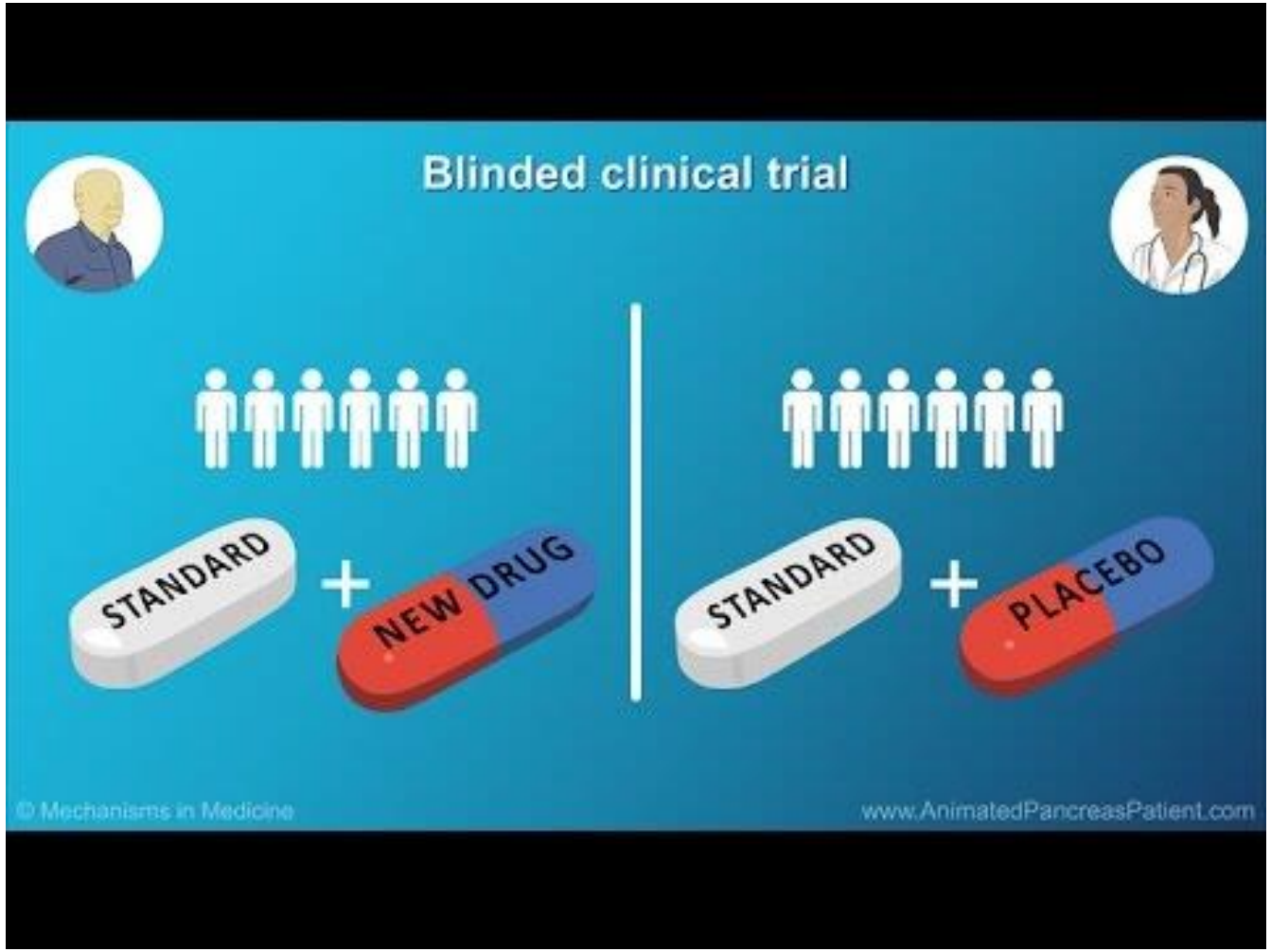
Remember: Correlation does not imply causation!

Experiment to Establish Causality Basic Concept

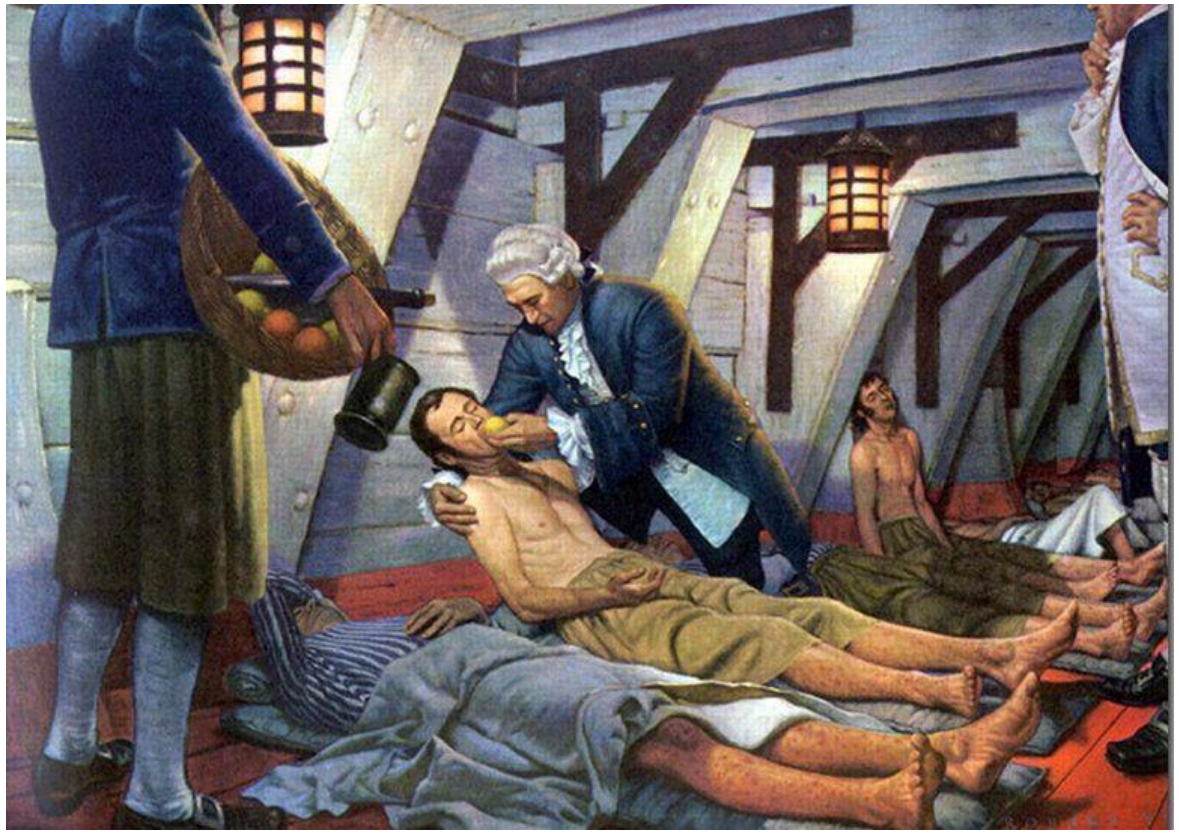
- Make sure there are no other differences between the groups



Clinical Trials



James Lind: The man who helped to cure scurvy with lemons



Impact study:
Do rural roads create pathways out of poverty?

- Roads lead to better integration between rural and urban markets.
- Prices of goods imported from urban areas decline in villages; availability improves.
- Farmers increase the use of modern technologies like fertilizer.
- Teenagers drop out of school to join labor force as urban markets become accessible.

Source: Aggarwal (2018)

School feeding and learning achievement: Evidence from India's midday meal program

- Prolonged exposure to school nutrition improves math and reading test scores.
- The effects are more pronounced when complemented with learning infrastructure.
- All children, irrespective of gender or wealth, benefit equally from the program.
- The learning impact is comparable to those from more direct learning interventions.

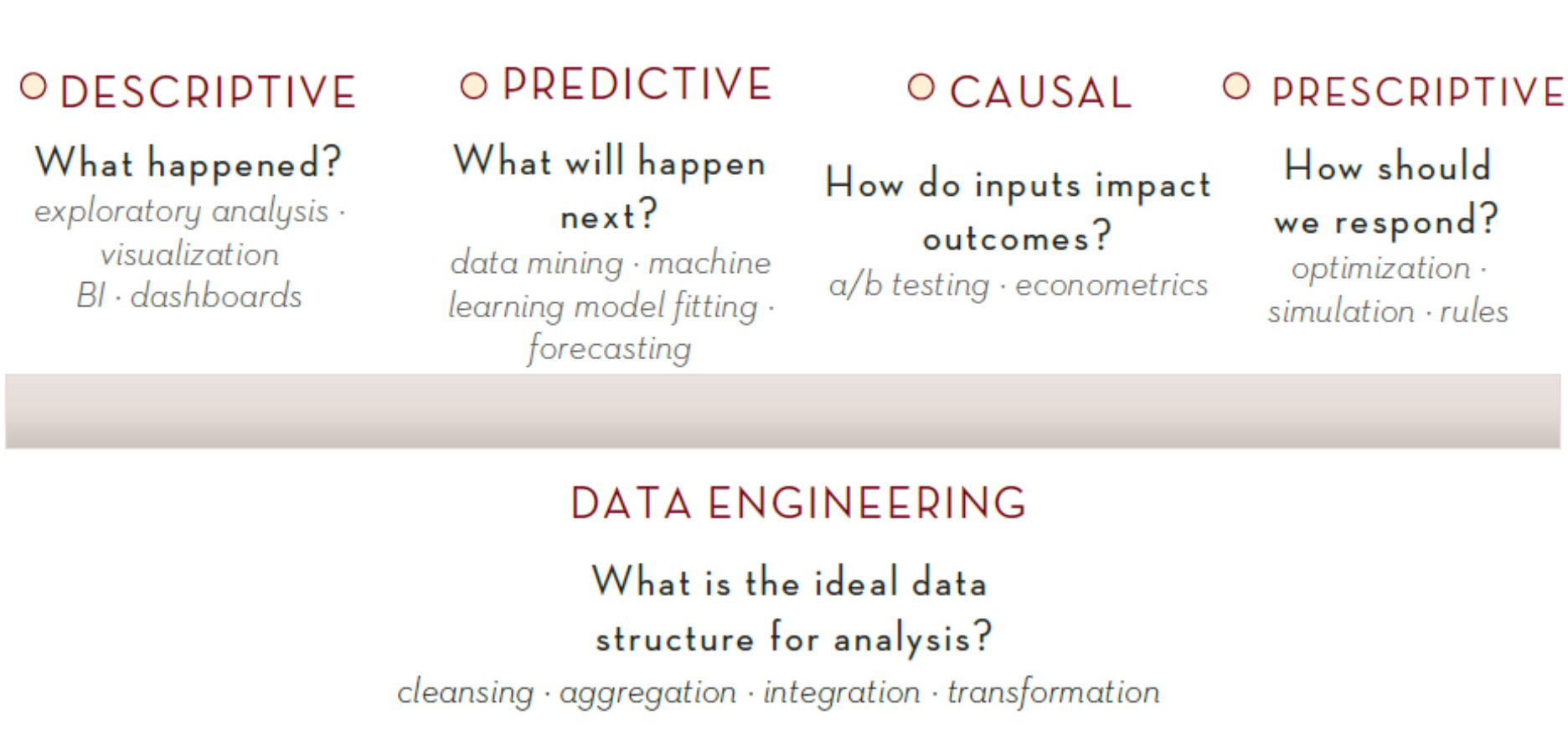
Source: Chakraborty and Jayaraman (2019)

Improving Maternal Health Using Incentives for Mothers and Health Care Workers: Evidence from India

- Role of incentives (cash transfer) for mothers and health care workers in the use of maternal and child health services
  - Increase in overall delivery in healthcare facility
  - Increase in use of pre- and postnatal care services and immunization
  - Reduction on early-neonatal deaths but had no impact on late-neonatal mortality
  - larger incentives to health workers are associated with relatively higher utilization rates

Source: Debnath (2021)

Privacy, Ethics, Bias, and Fairness in Data Science



Can the criminal justice system's artificial intelligence ever be truly fair?

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool used by U.S. courts to assess the likelihood of a defendant committing a crime in the future.
- In 2016, ProPublica reported that the tool was biased against Black defendants.

Bias

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

HR Analytics

Resume Screening model

Predict who should be further interviewed?

- Use Gender as a predictor?

Employee Attrition model

Who is going to leave and why?

- Use Gender as a predictor?

Bias

YOUR MONEY

American Express Kept a (Very) Watchful Eye on Charges

American Express mined its data to spot troubled cardholders. In some insta
lines.

By RON LIEBER
Published: January 30, 2009

Fairness

SHARE

In recent months, American Express has gone far beyond simply checking your credit score and making sure you pay on time. The company has been looking at home prices in your area, the type of mortgage lender you're using and

letters that infuriated many of the cardholders who received them. "Other customers who have used their card at establishments where you recently shopped," one of those letters said, "have a poor repayment history with American Express."

to

In some instances, if it didn't like what it was seeing, the company has cut customer credit lines. It laid out this logic

Could Target Sell Its 'Pregnancy Prediction Score'?

"We may share information with vendors, business partners and other organizations which are not part of the Target family," says the policy.

Thanks to the New
know that Target ha
out which of its sho
As Charles Duhigg, an expert in habit formation, makes clear, this is extremely valuable data as new parents tend to get hooked on certain brands and become loyal customers during the first few years of a child's life — when they're like the cookie monsters of kids' goods.

Target assigns every one of its customers a "pregnancy prediction score," with an estimate of the due date, so that coupons can be timed to the right stage of pregnancy (e.g., maternity ware in the second trimester, placenta teddy bears near the end of the third). Other pregnancy-product companies would surely

Privacy

Privacy

Claims That Google Violates Gmail User Privacy

Complaint: Utilizing a scanning or extraction device, Google intercepts all electronic communications sent to Gmail account holders. Google uses the information and content obtained from the scanning of incoming electronic communications to sell and place advertisements in Gmail account holders' browser windows that are related to the content of intercepted electronic communications.

Trade-off

Google's defense: We are using that same technology that scans for viruses and also scans for spam. It is basically technology that looks for pattern text, and we use that not only for the spam blocking and viruses but also to serve ads within the Gmail user's experience.

Source: http://www.nytimes.com/interactive/2013/10/02/technology/google-email-case.html?_r=0

Responsible AI

Firm should be aware of…

- Who owns the data
- What is legal
- What is ethical
- Whether customers find its practices acceptable

Government should …

- Set legal boundaries to protect individual privacy.
- Use data to create values only when it causes no harm to other citizens.
- Make harmless data available for research community.

Thank You