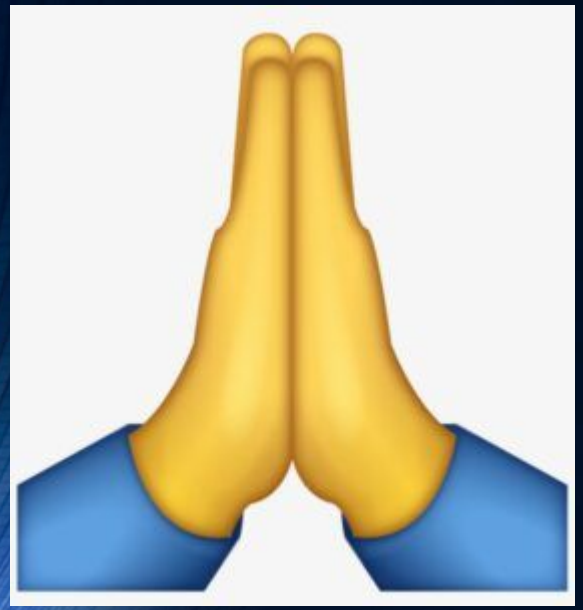
B.E. A.I. & D.S. (2020 Course)

Faculty **O**rientation **P**rogram

Data **M**odelling and **V**isualization

Presentation by

Dr. Araddhana Deshmukh



Dr. Nilesh Uke Sir and BoS, SPPU

Course Name: Data Modelling and Visualization

	Name	College	Email	Contact No
Course Coordinator	Dr. Araddhana Deshmukh	MMCOE, Pune	hodainds@mmcoe.edu.in	9970620939
Course Co-Coordinator	Dr. Sujatha Rao	SRCE, College, Wagholi, Pune		8788803839
Team Members	Shubhangi Suryawanshi	Dr D Y Patil Institute of technology, Pimpri	shubhangi.suryawanshi@dypvp.edu.in	9970082182
	Mrs Devyani Jitendra Bonde	Marathwada Mitra Mandal Institute of Technology, Lohgaon Pune	devyani.bonde@mmit.edu.in	9011076515
Industry Expert	Mr.Nachiket Kulkarni	EYTechnology	nachiketvk@gmail.com	9423529534
Industry Expert	Mr.Sankalp Giridhar	Wissen Technology	sankalpgiridhar@gmail.com	9604603443

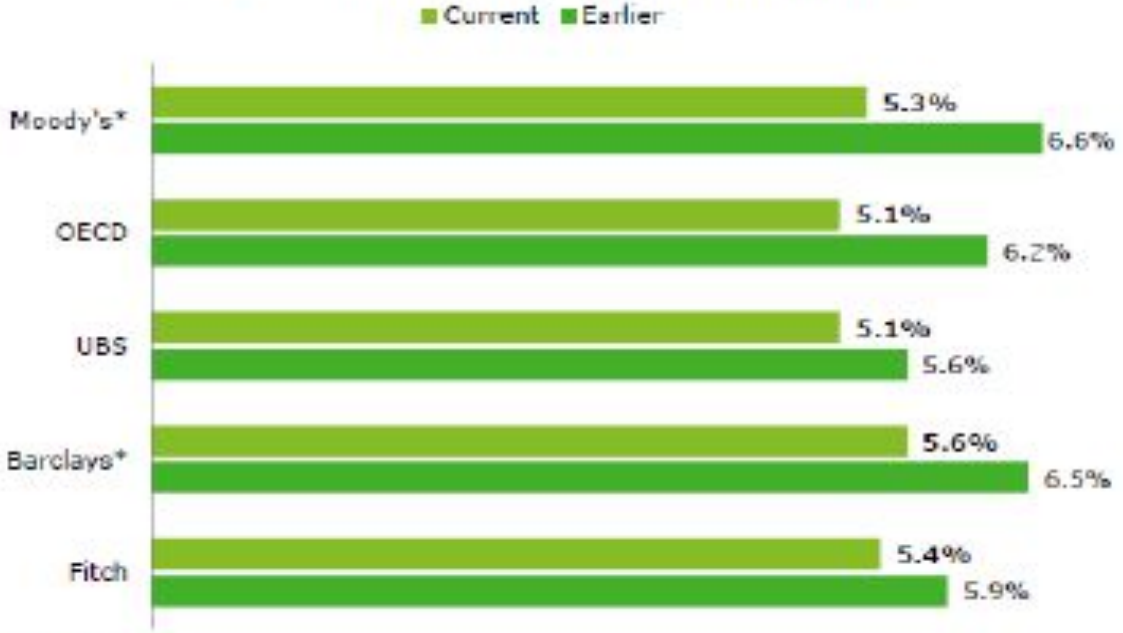
Contents :-

Covid Effect on Economy

The pandemic has set foot in India and is expected to lead the **country towards a major slowdown**

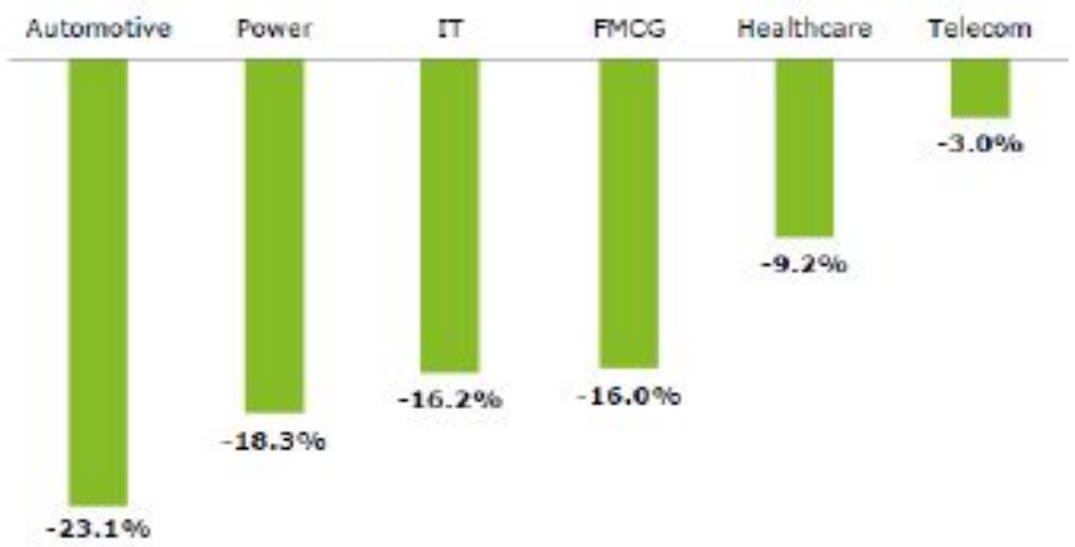
Major financial institutions have lowered growth estimates for India by 0.5–1.5 percent

FY21 growth projections for India



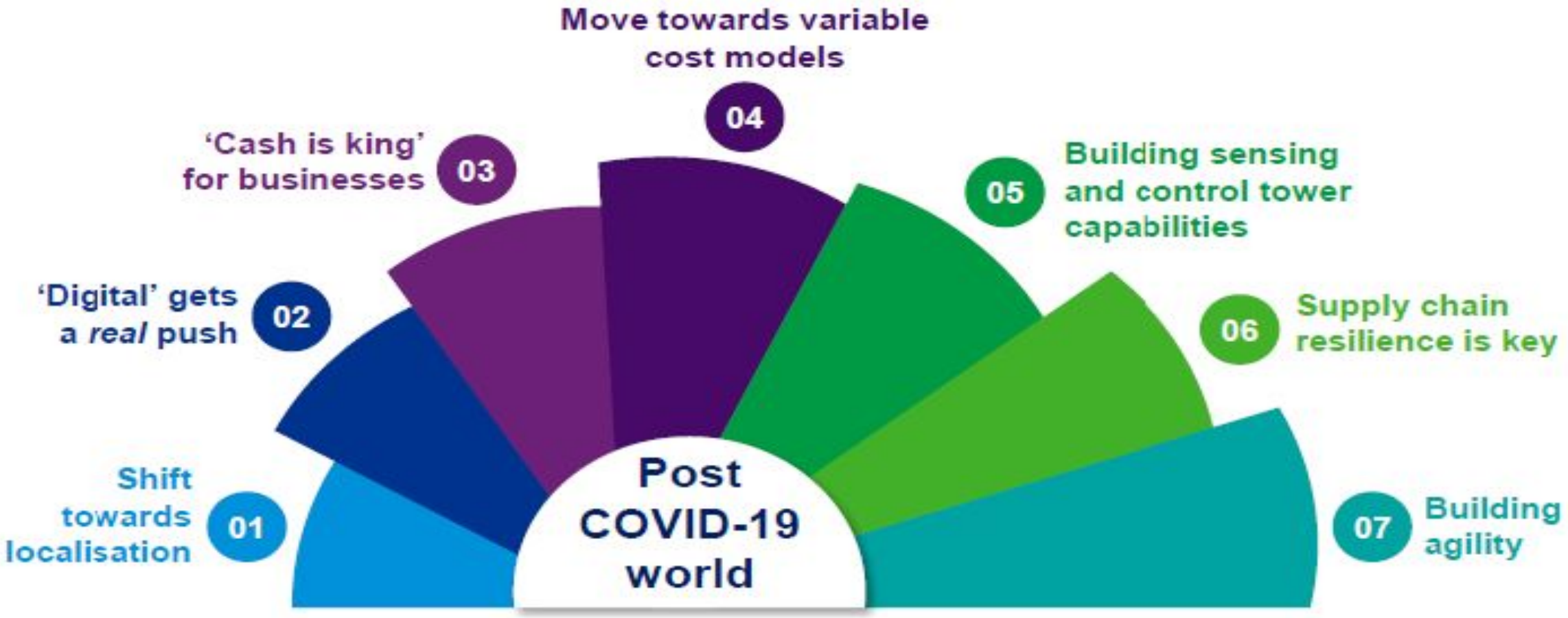
This is likely to put a downward pressure on ~~the~~ markets and industries

Impact on markets



Note*: FY20 projections
Sources: MoSPI, Commerce Ministry, "Indian economy braces for coronavirus-induced shock as curbs set to pull down growth", Mint, 15 March 2020

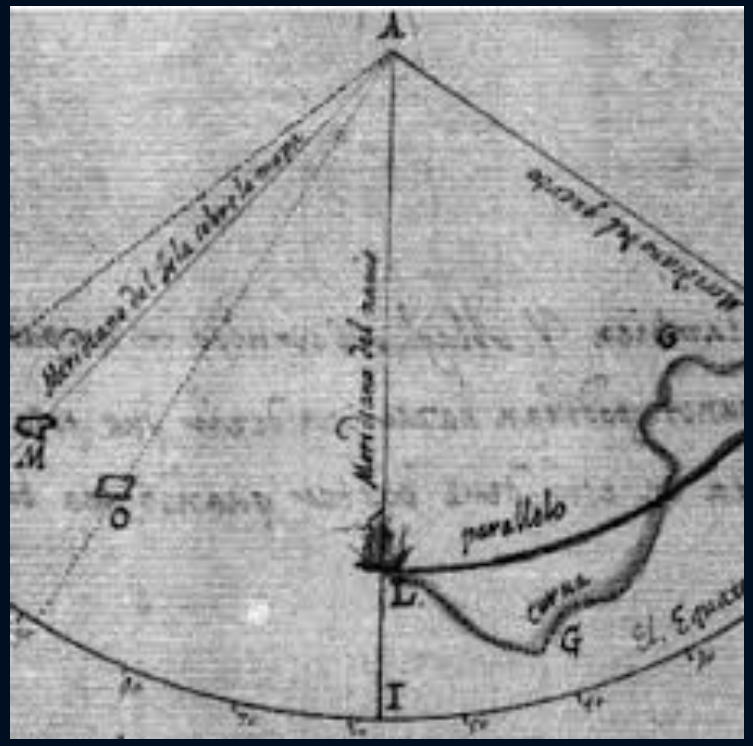
Post COVID -19



Data model concept :- Invention

- The first generation database, called Integrated Data Store (IDS), was designed by Charles Bachman at General Electric. Two famous database models, the network data model and the hierarchical data model, were proposed during this period of time".[12] Towards the end of the 1960s, Edgar F. Codd worked out his theories of data arrangement, and proposed the relational model for database management based on first-order predicate logic.

Who is he????



- 17th Century Michael Florent Van Langren
- In 1644, the idea of statistical data presented in the form of graphical representation was attributed to Flemish astronomer Michael Florent Van Langren.

Savitribai Phule Pune University

Fourth Year of Artificial Intelligence and Data Science (2020 Course) 417522: Data Modeling and Visualization

Teaching Scheme:

TH: 03

Hours/Week

Credit 03

Examination Scheme:

In-Sem (Paper): 30 Marks

End-Sem (Paper): 70 Marks

Prerequisites Courses: Statistics (), Computer Graphics (), Database Management Systems ()

CO Course Objectives and Course Outcomes

Course Objectives:

- Creating an emerging data model for the data to be stored in a database
- Conceptualized representation of Data objects
- Create associations between different data objects, and the rules
- Organize data description, data semantics, and consistency constraints of data
- Identifying data trends
- Incorporate data visualization tools and reap transformative benefits in their critical areas of operations

Course Outcomes:

After completion of the course, learners should be able to-

CO1: Summarize data analysis and visualization in the field of exploratory data science

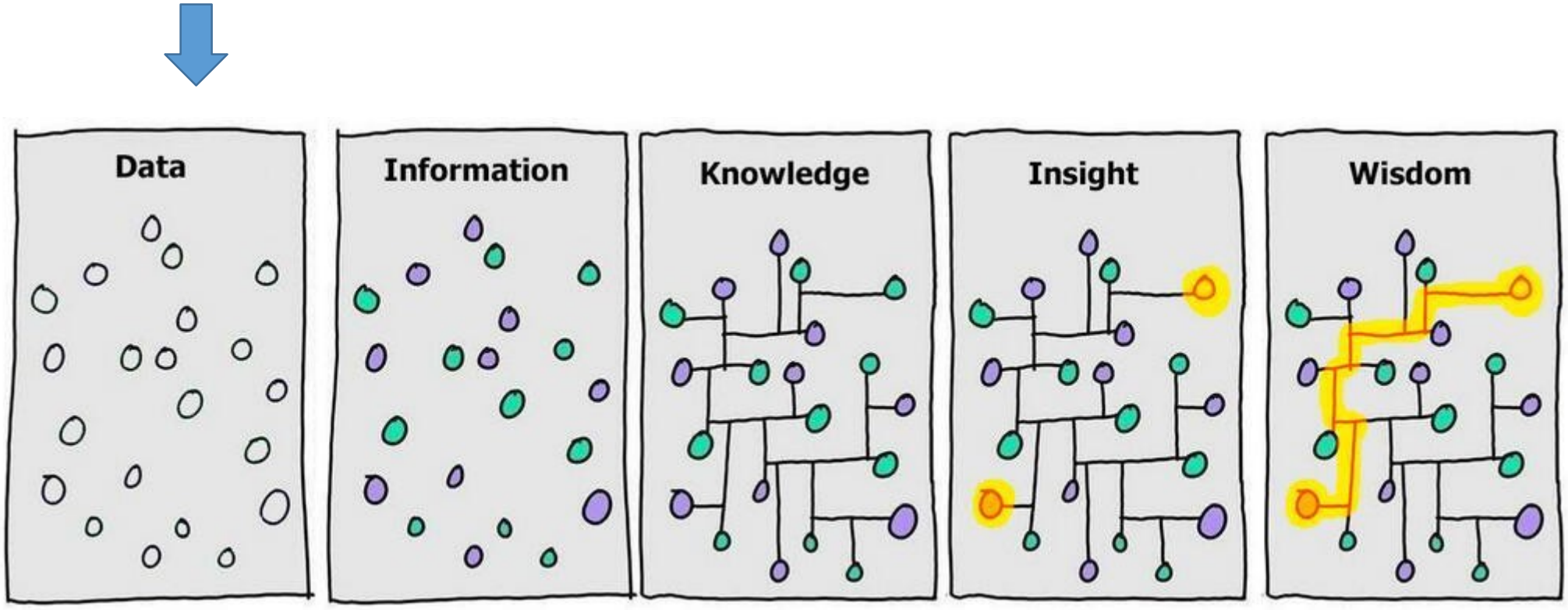
CO2: Analyze the characteristics and requirements of data and select an appropriate data model

CO3: Describe to load, clean, transform, merge and reshape data CO4: Design a probabilistic data modeling, interpretation, and analysis

CO5: Evaluate time series data

CO6: Integrate real world data analysis problems

The Information Continuum



Cartoon by David Somerville, based on a two pane version by Hugh McLeod

Types of Data

Quantitative Data

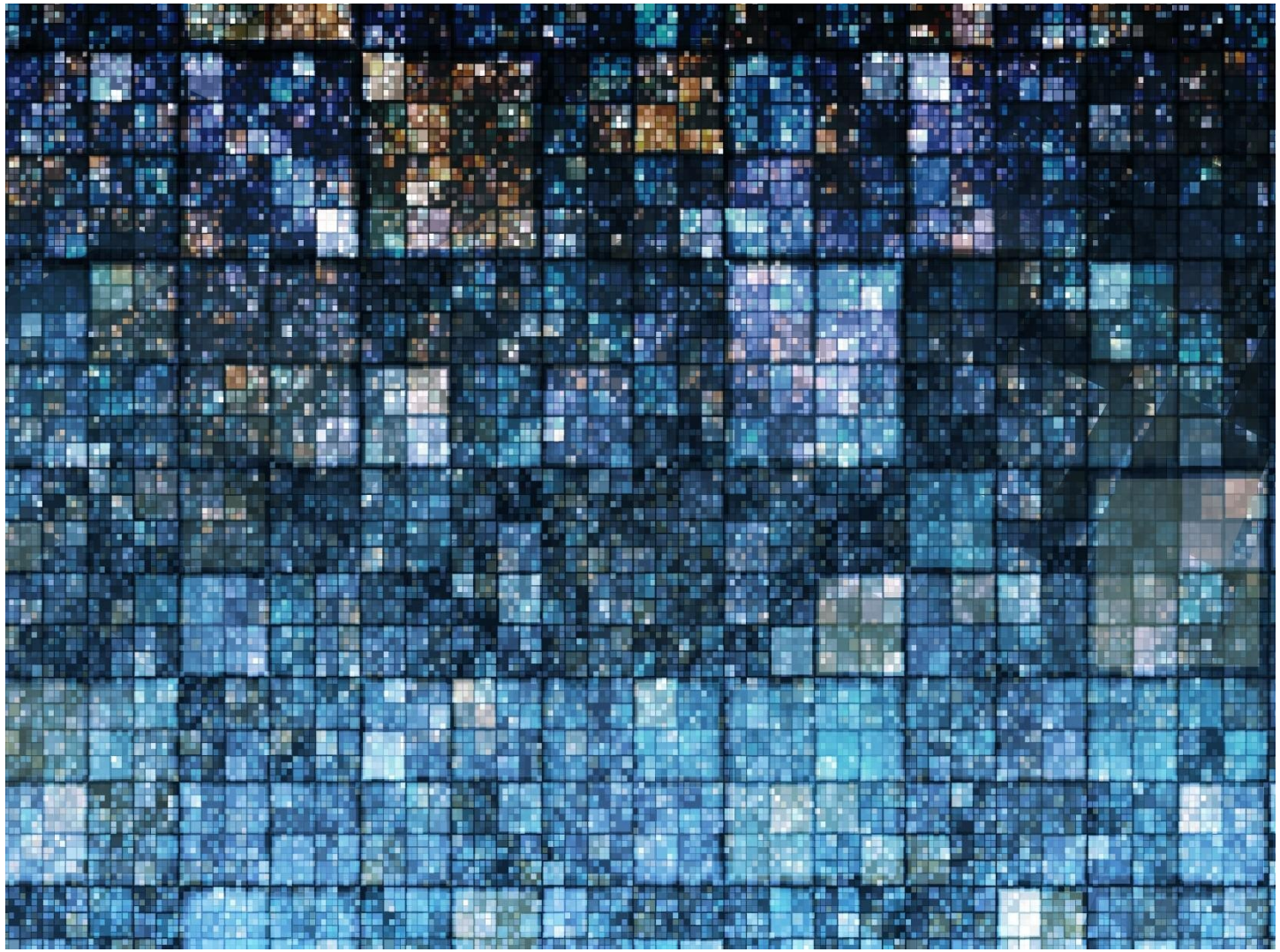
- Measurable
- Collected through measuring things that have a fixed reality
- Close ended

Qualitative Data

- Descriptive
- Collected through observation, field work, focus groups, interviews, recording or filming conversations
- Open ended

Big Data

Data that is too large or too complex to be managed using traditional data processing, analysis, and storage techniques.



What is Big Data?

- Big Data is a collection of large datasets that cannot be adequately processed using traditional processing techniques. Big data is not only data it has become a complete subject, which involves various tools, techniques and frameworks.

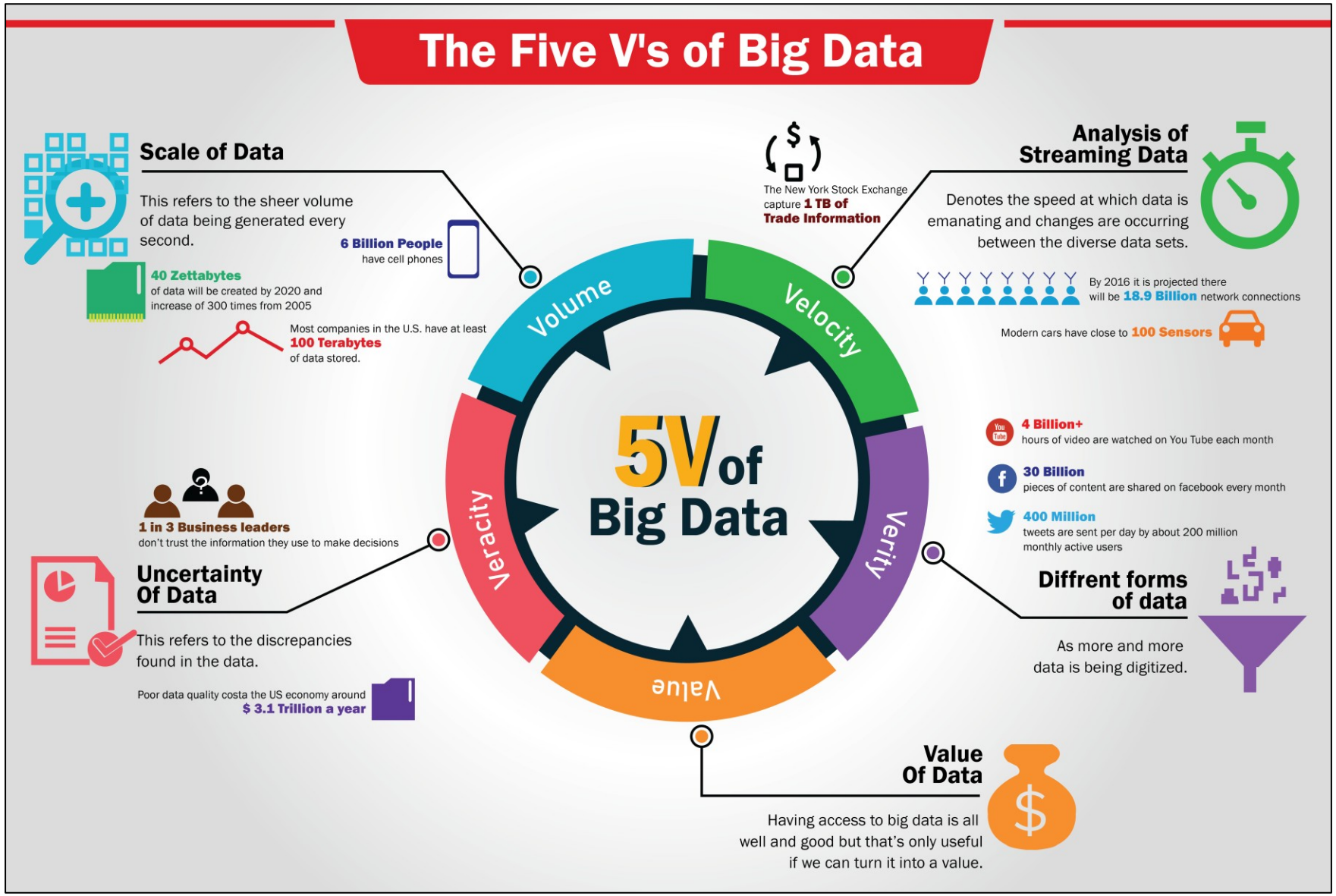
- Big data term describes the volume amount of data both structured and unstructured manner that adapted in day-to-day business environment. It's important that what organizations utilize with these with the data that matters.

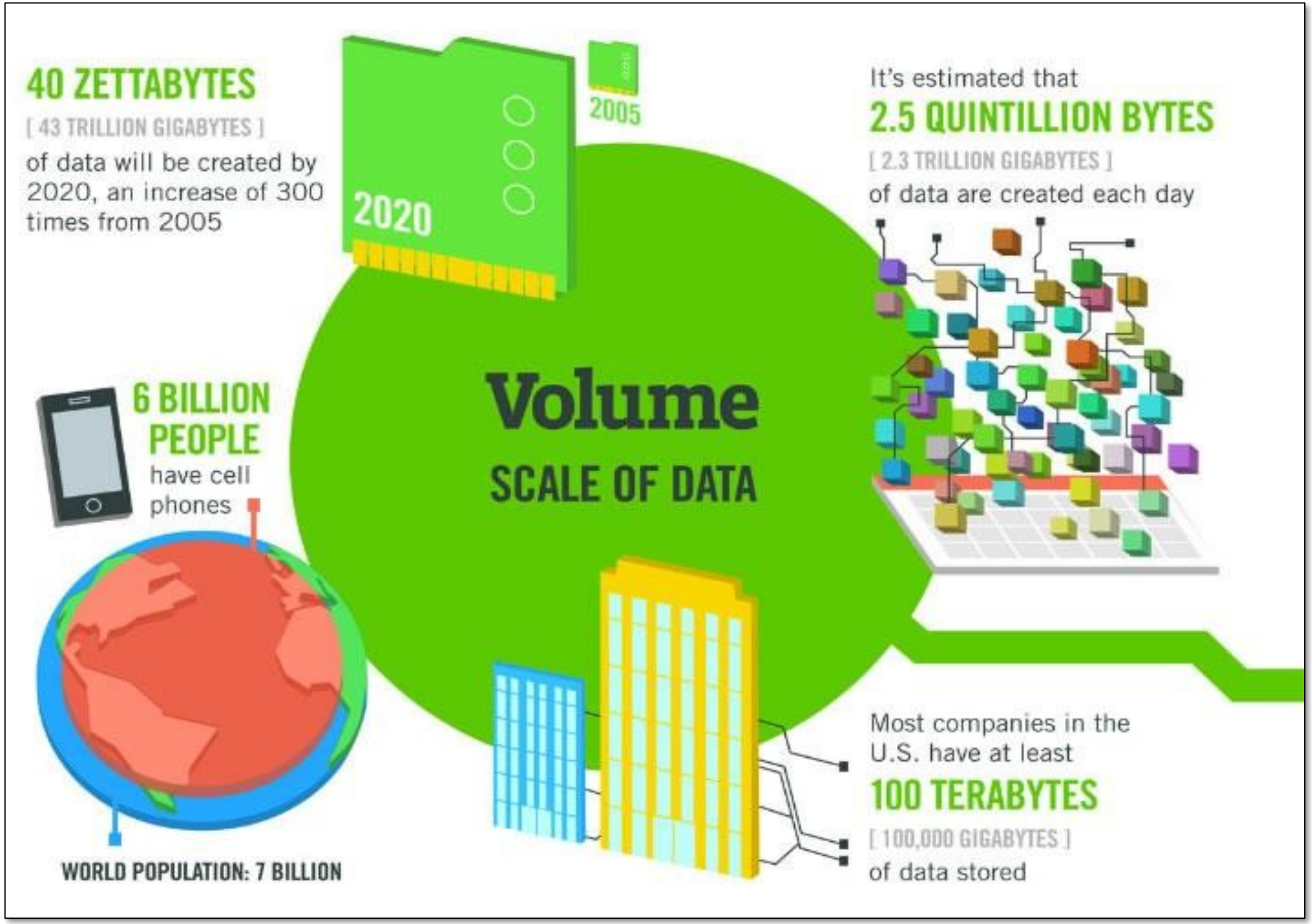
- Big data helps to analyze the in-depth concepts for the better decisions and strategic taken for the development of the organization.

The Evolution of Big Data

The concept of Big Data came into existence in the early 2000s when Industry analyst Doug Laney the definition of big data as the three categories as follows:

- **Volume:** Organizations collects the data from relative sources, which includes business transactions, social media and information from sensor or machine-to-machine data. Before, storage was a big issue but now the advancement of new technologies (such as Hadoop) has reduced the burden.
- **Velocity:** Data streams unparalleled speed of velocity and have improved in timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in real time operations.
- **Variety:** Data comes in all varieties in form of structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.


The Five V's of Big Data
Scale of Data
This refers to the sheer volume of data being generated every second.
6 Billion People have cell phones
40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005
Most companies in the U.S. have at least 100 Terabytes of data stored.
The New York Stock Exchange capture 1 TB of Trade Information
Analysis of Streaming Data
Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.
By 2016 it is projected there will be 18.9 Billion network connections
Modern cars have close to 100 Sensors
Volume
Velocity
5V of Big Data
Veracity
Verity
Value
4 Billion+ hours of video are watched on You Tube each month
30 Billion pieces of content are shared on facebook every month
400 Million tweets are sent per day by about 200 million monthly active users
1 in 3 Business leaders don't trust the information they use to make decisions
Uncertainty Of Data
This refers to the discrepancies found in the data.
Poor data quality costa the US economy around $ 3.1 Trillion a year
Diffrent forms of data
As more and more data is being digitized.
Value Of Data
Having access to big data is all well and good but that's only useful if we can turn it into a value.


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Volume: scale of

Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8 bits	1 byte
kilobyte (KB)	1000^1 bytes	1,000 bytes
megabyte (MB)	1000^2 bytes	1,000,000 bytes
gigabyte (GB)	1000^3 bytes	1,000,000.000 bytes
terabyte (TB)	1000^4 bytes	1,000,000,000,000 bytes
petabyte (PB)	1000^5 bytes	1,000,000,000,000,000 bytes
exabyte (EB)	1000^6 bytes	1,000,000,000,000,000,000 bytes
zettabyte (ZB)	1000^7 bytes	1,000,000,000,000,000,000,000 bytes
yottabyte (YB)	1000^8 bytes	1,000,000,000,000,000,000,000,000 bytes

Volume: scale of data

- 90% of today's data has been created in just the last 2 years
- Every day we create 2.5 quintillion bytes of data or enough to fill 10 million Blu-ray discs
- 40 zettabytes (4o trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005, and the equivalent of 5,200 gigabytes of data for every man, woman and child on Earth
- Most companies in the US have over 100 terabytes (100,000 gigabytes) of data stored

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

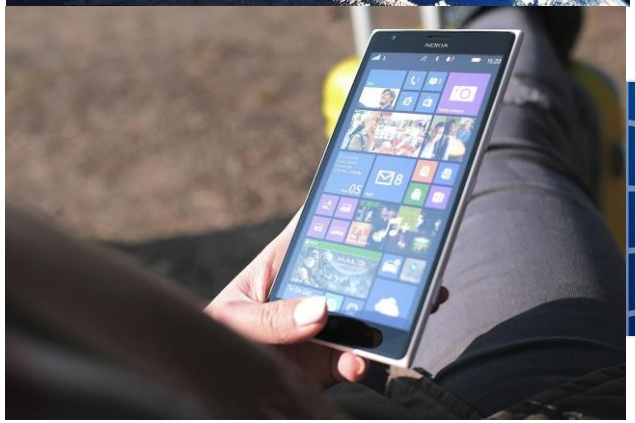
ANALYSIS OF STREAMING DATA

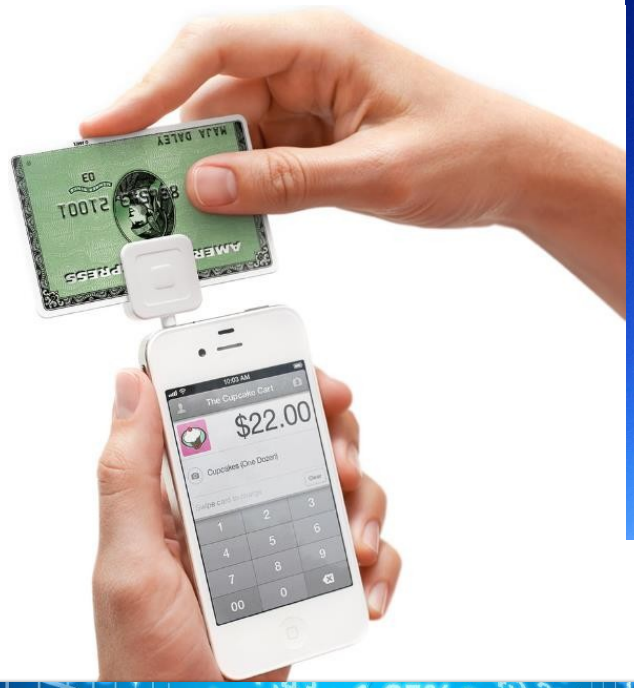
By 2016, it is projected there will be

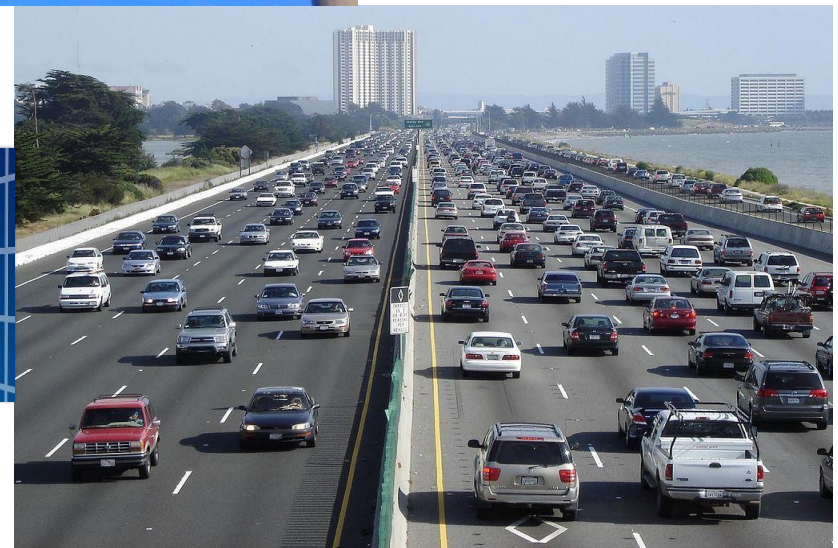
18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Velocity: analysis of streaming data







Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

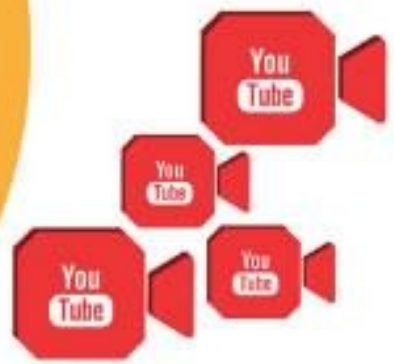
are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

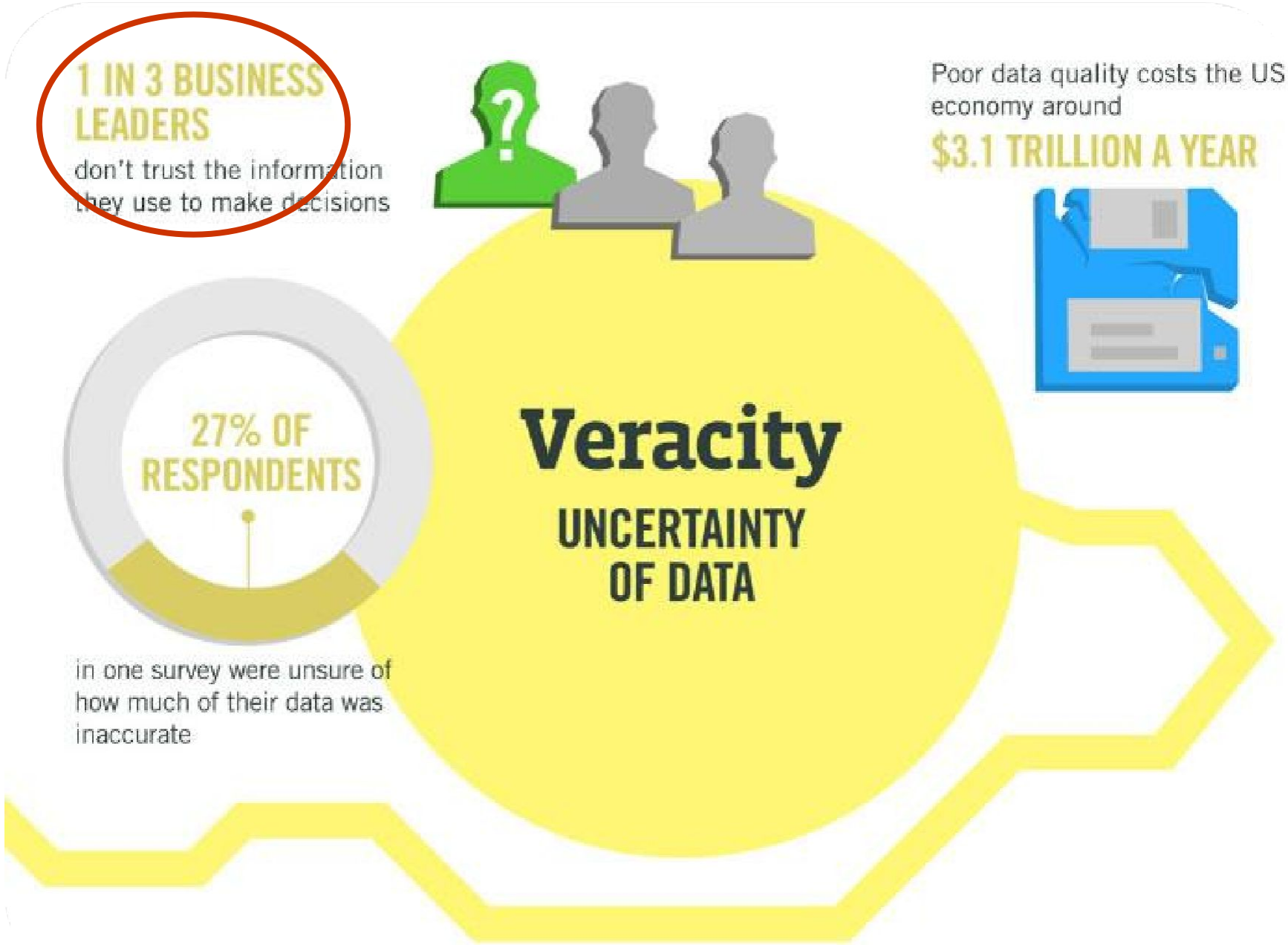
are watched on YouTube each month



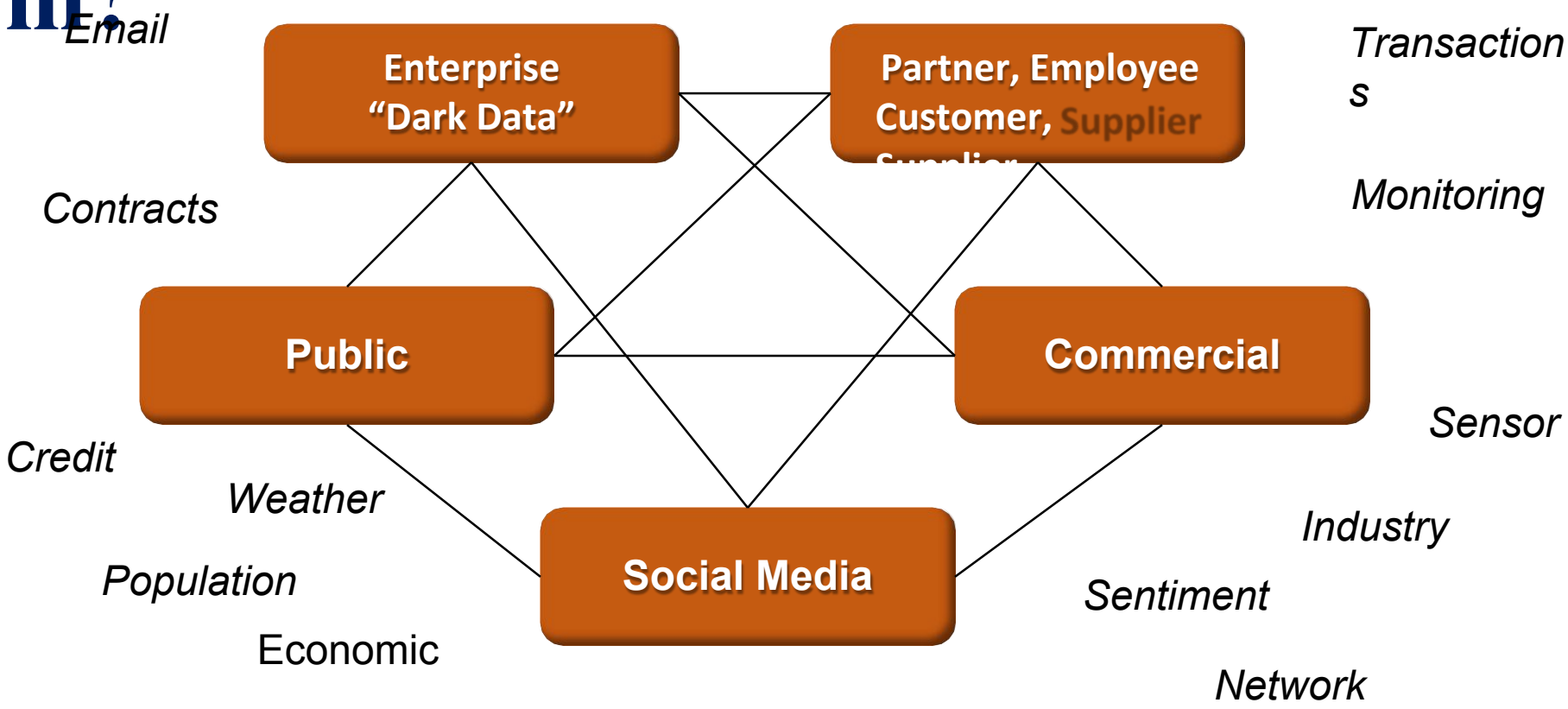
400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Variety: different forms of data

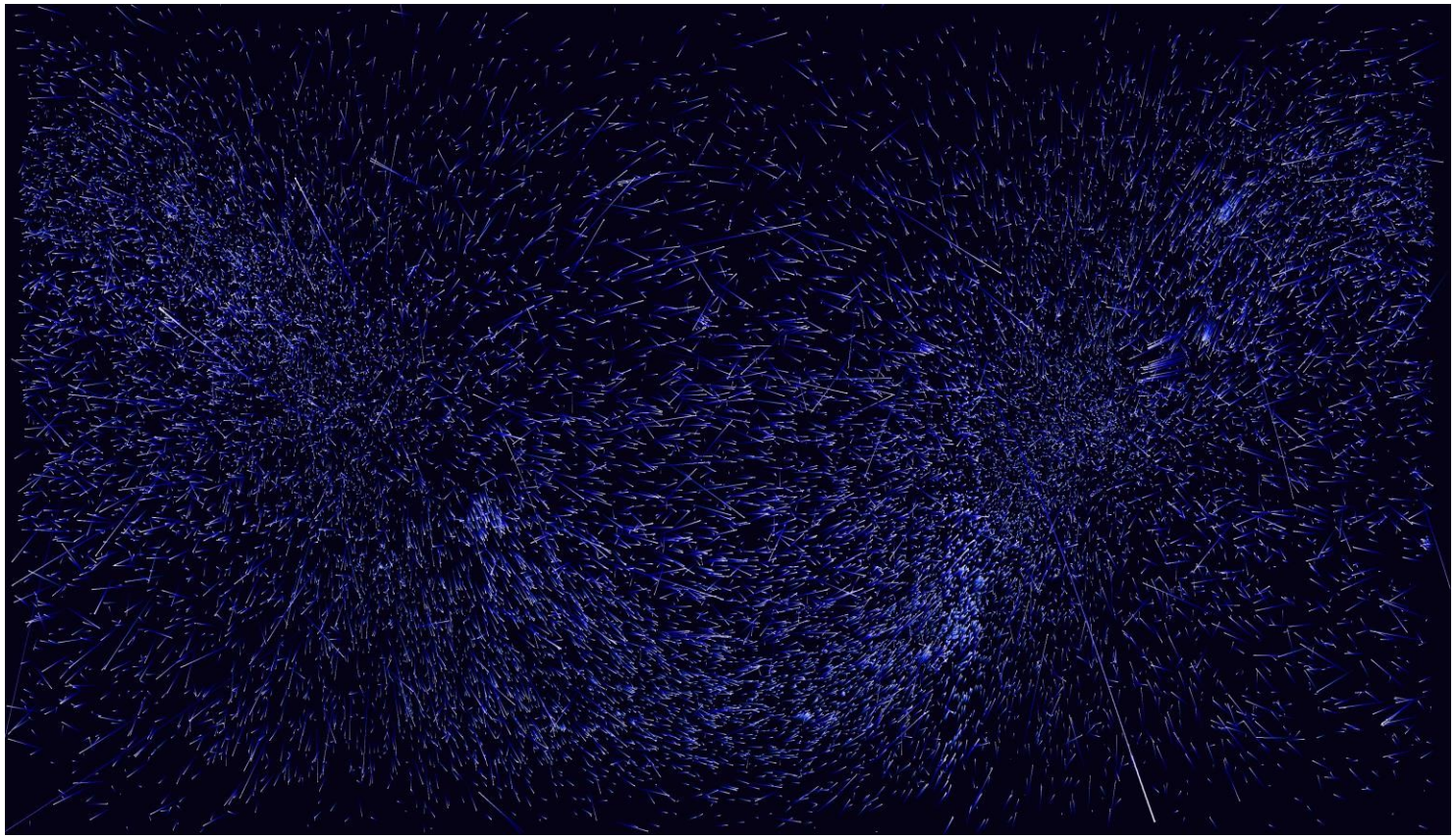

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

Where does Big Data come from?



Veracity: trustworthiness of data

- Origin
- Authenticity
- Trustworthiness
- Completeness
- Integrity



Categories of Big Data - I

Big data works on the data produced by various devices and their applications.

- **Black Box Data:** It includes the conversation between crew members and any other communications (alert messages or any order passed) by the technical grounds duty staff.
- **Social Media Data**: Social networking sites such as Facebook and Twitter contains the information and the views posted by millions of people across the globe.
- **Stock Exchange Data**: It holds information (complete details of in and out of business transactions) about the 'buyer' and 'seller' decisions in terms of share between different companies made by the customers.

Who are the ones who use the Big Data Technology?

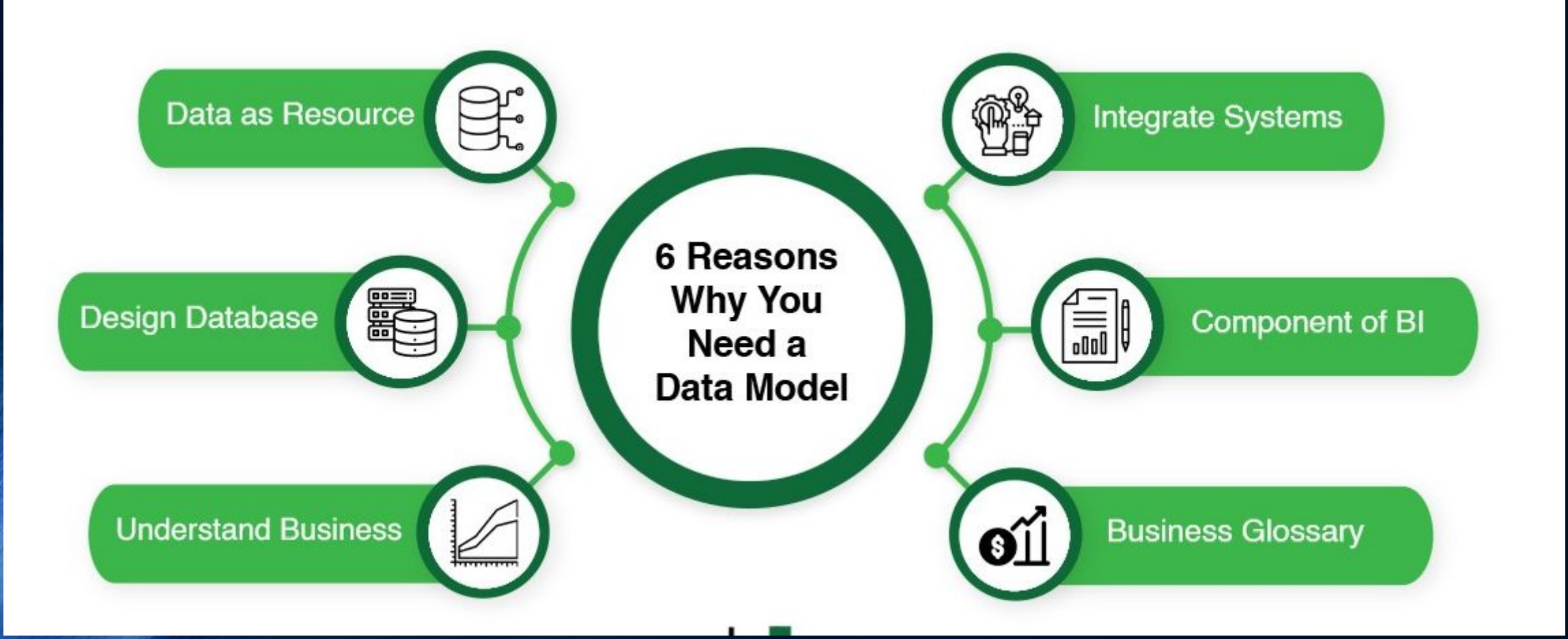
1. Banking
2. Government
3. Education
4. Health Care
5. Manufacturing
6. Retail

Data Modelling Vs Data Visualization

Data Modelling

- Data Modeling refers to the process of creating a visual representation of an entire information system or some of its parts to communicate the relationships between data points and structures.
- The purpose is to show the types of data stored in the system, the relationships among the data types, the formats and attributes of the data, and how the data can be grouped and organized.
- The Data Modeling process begins with the collection of information about business requirements from both stakeholders and end-users. The business requirements are then translated into data structures for the formulation of a concrete Database design.
- Today, Data Modeling finds its application across every sector you could possibly think of, from Financial Institutions to the Healthcare Industry. A study by LinkedIn rates Data Modeling as the fastest-growing profession in the present job market.

Why Data Model?



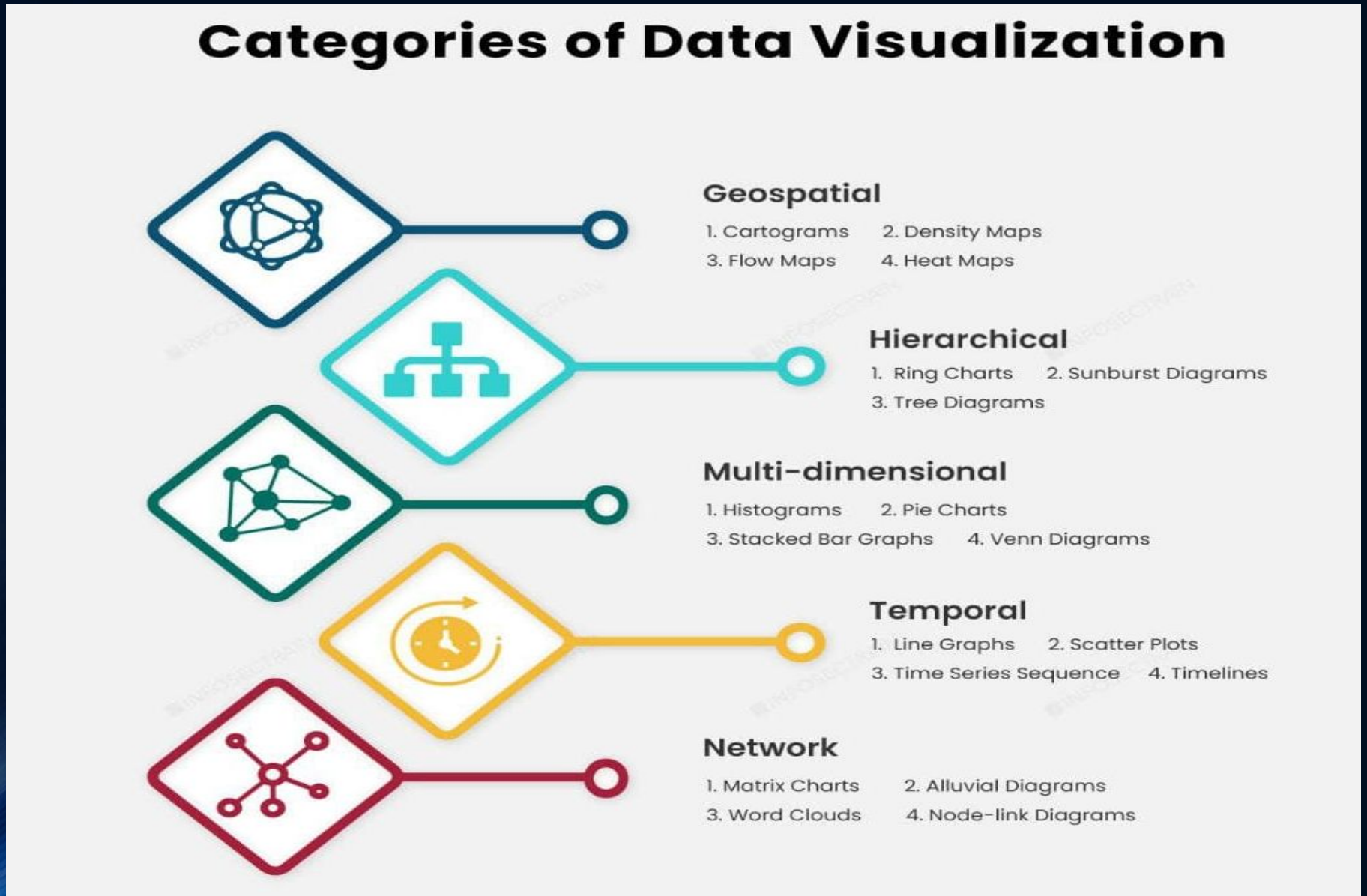
Data Visualization

- Data Visualization refers to the process of representing data and information graphically. By the use of visual elements like graphs, charts, and maps, Data Visualization tools offer an accessible way to view and understand trends and patterns in data.
- Data Visualization helps organizations to analyze huge volumes of data and make data-driven decisions. It also makes it easy for individuals and companies to understand data. Data Visualization is very useful today as companies are generating and collecting huge data volumes. It can help them to unmask hidden gems from data, which are good for growth.
-

key similarities

- The following are the key similarities between Data Modeling and Visualization:
- **They both deal with Data:** Data is at the center of both Data Modeling and Data Visualization. They help users make sense of vague sets of data and get the relevant metrics to help in better decision-making.
- **No need for ML Algorithms:** Both Data Modeling and Visualization don't require the use of Machine Learning algorithms to get the correct results.
- **They both use Visual Elements:** In both Data Modeling and Data Visualization, the answers are in the form of visual elements rather than text or numbers. However, they differ in the types of visual elements that are used.
- **No need for Data Analysis:** Both Data Modeling and Visualization don't require data to be analyzed. Instead, Data Engineers and Data Modelers go straight into working with the data the way it is to discover inconsistencies in the data.

Why Data visulaization?



Difference between DM and DV

Feature	Data Modeling	Data Visualization
Definition	Data Modeling refers to **designing** the Entity-Relationship modeling for Database tables to establish the connections between tables. It also involves designing the schema for Data Warehouses. Thus, it shows how tables are connected in schema terms.	Data Visualization involves **presenting data** in a visual context to show hidden trends and patterns in data. Such trends and patterns may not be explicit in text data. Visualization makes data easy for anyone to understand.

Difference between DM and DV

Feature	Data Modeling	Data Visualization
Techniques	Data Modeling techniques include (ERDs) **Entity Relationship Diagrams** to depict the way data has been stored in the Database. The ERDs show the types of relationships between the different tables in the Database, whether one-to-many, many-to-many, etc. It also uses data dictionaries and (UML) **Unified Modelling Language**	Data Visualization involves the use of graphs, charts, and tables to present data visually. These visual tools show how the different data attributes are related to each other.

Difference between DM and DV

Feature	Data Modeling	Data Visualization
Used For	Data Modeling is used to ensure that data is stored in a database and represented accurately. It shows the inherent structure of data by identifying data identities, attributes, and the relationship between the entities.	Data Visualization is used to communicate information clearly and efficiently to the users by presenting it using visual elements.
Benefits	Facilitate faster access to data across the entire organization. Data Modeling also makes it easy to establish the correct structure of data and enforce compliance standards.	Helps businesses understand their customers, products, and processes better. This is good for sound decision-making and making predictions. TT
Tools	Common data modelling tools include Erwin Data Modeler, ER/Studio, DbSchema, ERBuilder, HeidiSQL, Navicat Data Modeler, Toad Data Modeler, Archi, and others.	Data Visualization is done using tools such as Knowi, Tableau, Dygraphs, QlikView, DataHero, ZingCHhart, Domo, and others. It can also be done in programming languages such as Python and R.
Performed By	Data Architects and Modelers.	Data Engineers.

Course Contents

Unit I	Introduction to Data Modelling	(07 Hours)
Basic probability: Discrete and continuous random variables, independence, covariance, central limit theorem, Chebyshev inequality, diverse continuous and discrete distributions. **Statistics,** Parameter Estimation, and Fitting a Distribution: Descriptive statistics, graphical statistics, method of moments, maximum likelihood estimation **Data Modeling Concepts** • Understand and model subtypes and supertypes • Understand and model hierarchical data • Understand and model recursive relationships • Understand and model historical data		

Basic Probability Concepts :- ref https://web.stanford.edu/class/hrp259/2007/discrete/discrete_259_2007.ppt

- Probability – the chance that an uncertain event will occur (always between 0 and 1)
- Impossible Event – an event that has no chance of occurring (probability = 0)
- Certain Event – an event that is sure to occur (probability = 1)

Assessing Probability

There are three approaches to assessing the probability of an uncertain event:

1. *a priori -- based on prior knowledge of the process*
2. empirical Probability
3. Subjective Probability

Assuming all outcomes are equally likely

probability of occurrence

probability of occurrence

based on a combination of an individual's past experience, personal opinion, and analysis of a particular situation

2. Example of empirical probability

Find the probability of selecting a male taking statistics from the population described in the following table:

	Taking Stats	Not Taking Stats	Total
Male	84	145	229
Female	76	134	210
Total	160	279	439

Subjective probability

- Subjective probability may differ from person to person
  - A media development team assigns a 60% probability of success to its new ad campaign.
  - The chief media officer of the company is less optimistic and assigns a 40% of success to the same campaign
- The assignment of a subjective probability is based on a person's experiences, opinions, and analysis of a particular situation
- Subjective probability is useful in situations when an empirical or a priori probability cannot be computed

Events

Each possible outcome of a variable is an event.

- Simple event
  - An event described by a single characteristic
  - e.g., A day in January from all days in 2015
- Joint event
  - An event described by two or more characteristics
  - e.g. A day in January that is also a Wednesday from all days in 2015
- Complement of an event A (denoted A')
  - All events that are not part of event A
  - e.g., All days from 2015 that are not in January

Random Variable

- A random variable *X* takes on a defined set of values with different probabilities.
  - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
  - For example, if you poll people about their voting preferences, the percentage of the sample that responds "Yes on Proposition 100" is a also a random variable (the percentage will be slightly different every time you poll).

- Roughly, probability is how frequently we expect different outcomes to occur if we repeat the experiment over and over ("frequentist" view)

Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

Probability functions

- A probability function maps the possible values of x against their respective probabilities of occurrence, $p(x)$
- $p(x)$ is a number from 0 to 1.0.
- The area under a probability function is always 1.

Chebyshev's inequality

Let $\boldsymbol{X}$ be a random variable defined over a probability space.

Question: How to capture deviation of $\boldsymbol{X}$ from $\mathbf{E}[\boldsymbol{X}]$?

Define $\boldsymbol{V} = \boldsymbol{X} - \mathbf{E}[\boldsymbol{X}]$

Question: What is $\mathbf{E}[\boldsymbol{V}]$?

Answer: $\mathbf{0}$

Redefine $\boldsymbol{V} = (\boldsymbol{X} - \mathbf{E}[\boldsymbol{X}])^{\mathbf{2}}$

Question: What is $\mathbf{E}[\boldsymbol{V}]$?

Answer: $\mathbf{E}[\boldsymbol{X}^{\mathbf{2}}] - (\mathbf{E}[\mathbf{X}])^{\mathbf{2}}$

Called **variance** of $\boldsymbol{X}$

Chebyshev's inequality

Let X be a random variable defined over a probability space.

$$\mathbf{P}(|X - \mathbf{E}[\mathbf{X}]| \geq t)$$
$$= \mathbf{P}\left(|X - \mathbf{E}[\mathbf{X}]|^2 \geq t^2\right)$$

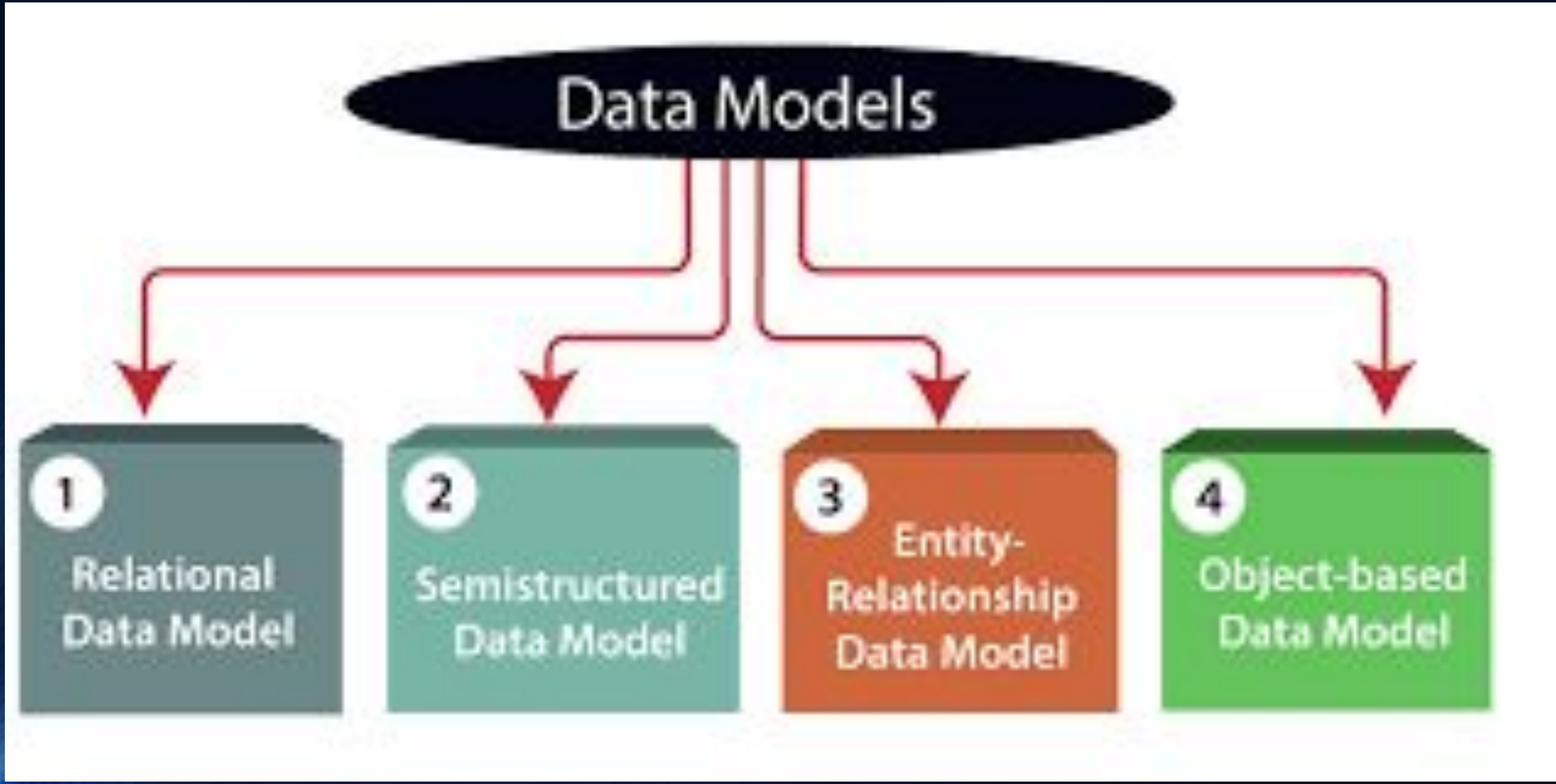
Applying Markov Inequality,

$$\leq \frac{\mathbf{E}[(X - \mathbf{E}[X])^2]}{t^2}$$
$$= \frac{\mathbf{E}[X^2] - (\mathbf{E}[\mathbf{X}])^2}{t^2}$$
$$= \frac{\textbf{variance of } X}{t^2}$$

Limitations:

- Calculating $\mathbf{E}[X^2]$ is sometimes difficult.
- Usually gives bounds that are better than Markov Inequality but inferior to the bound achieved by other methods.
- Simple practice problems will be given to you on the use of **Chebyshev** Inequality.

Types of Data Models

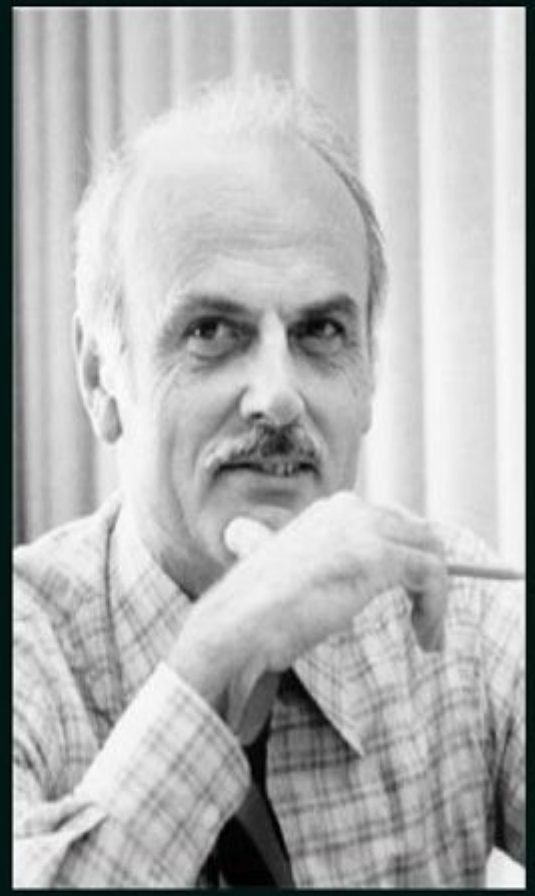


Relational Data Model

Attributes

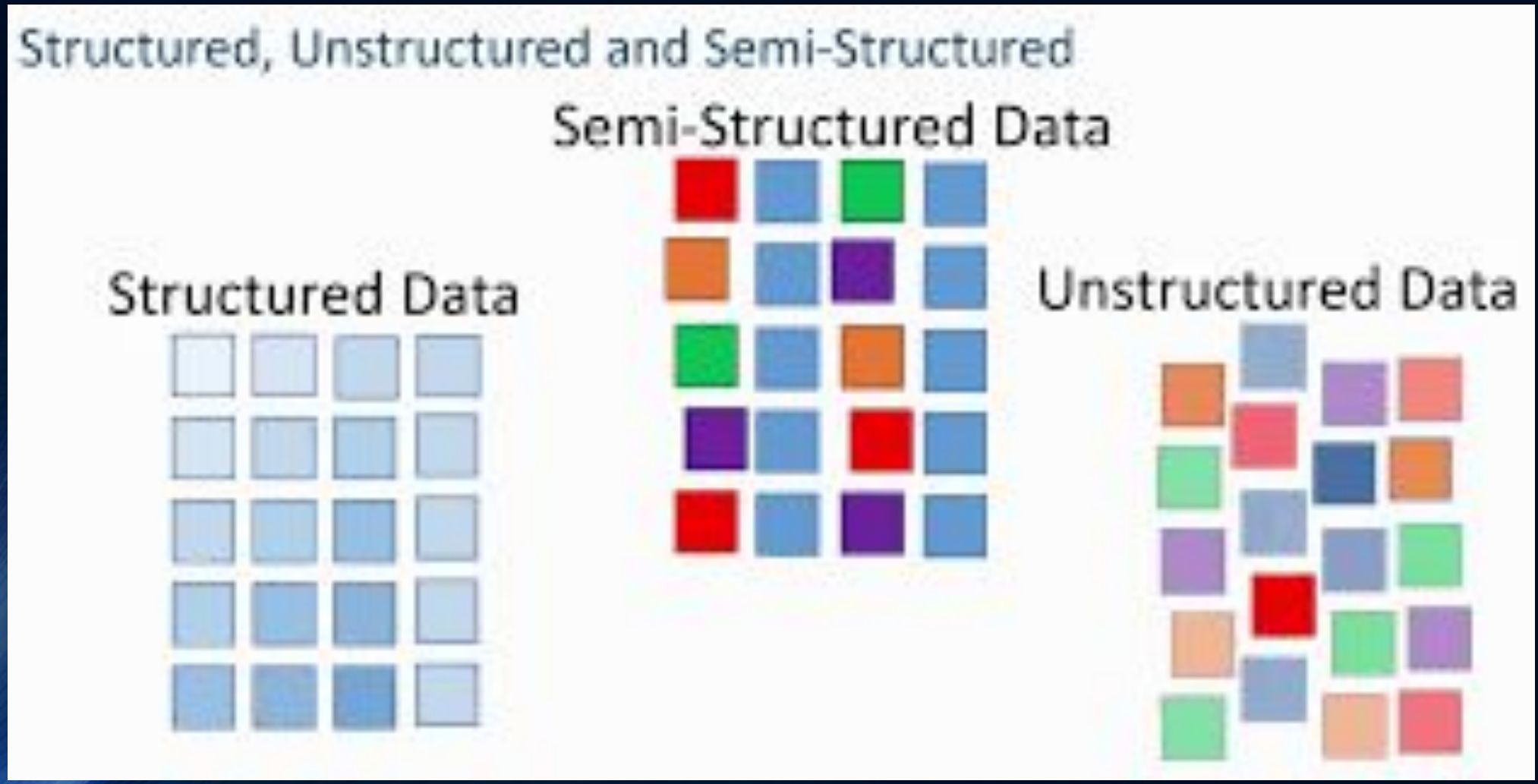
Relation name

STUDENT	ROLL_NO	NAME	AGE	PHONE
Tuples	1	David	18	2525364236
	2	Benjamin	19	3526699663



E. F. Codd in 1970.

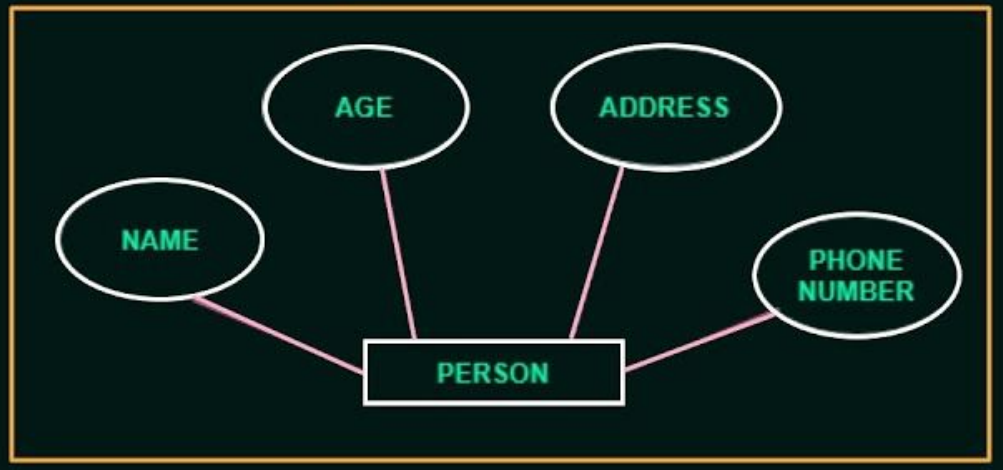
Semi structured Data Model



	Unstructured Data	Semi-Structured Data	Structured Data
Characteristic	No defined data models; difficult to search	Loosely-coupled data models	Clearly-defined data models; easy to search
Example	Image file	Call center log	Spreadsheet
Storage	Data lake	Organized by metatags	Relational Database

Entity-Relationship Model

- Entity
- Attributes
- Complex Attributes
- Null Values
- Key Attribute

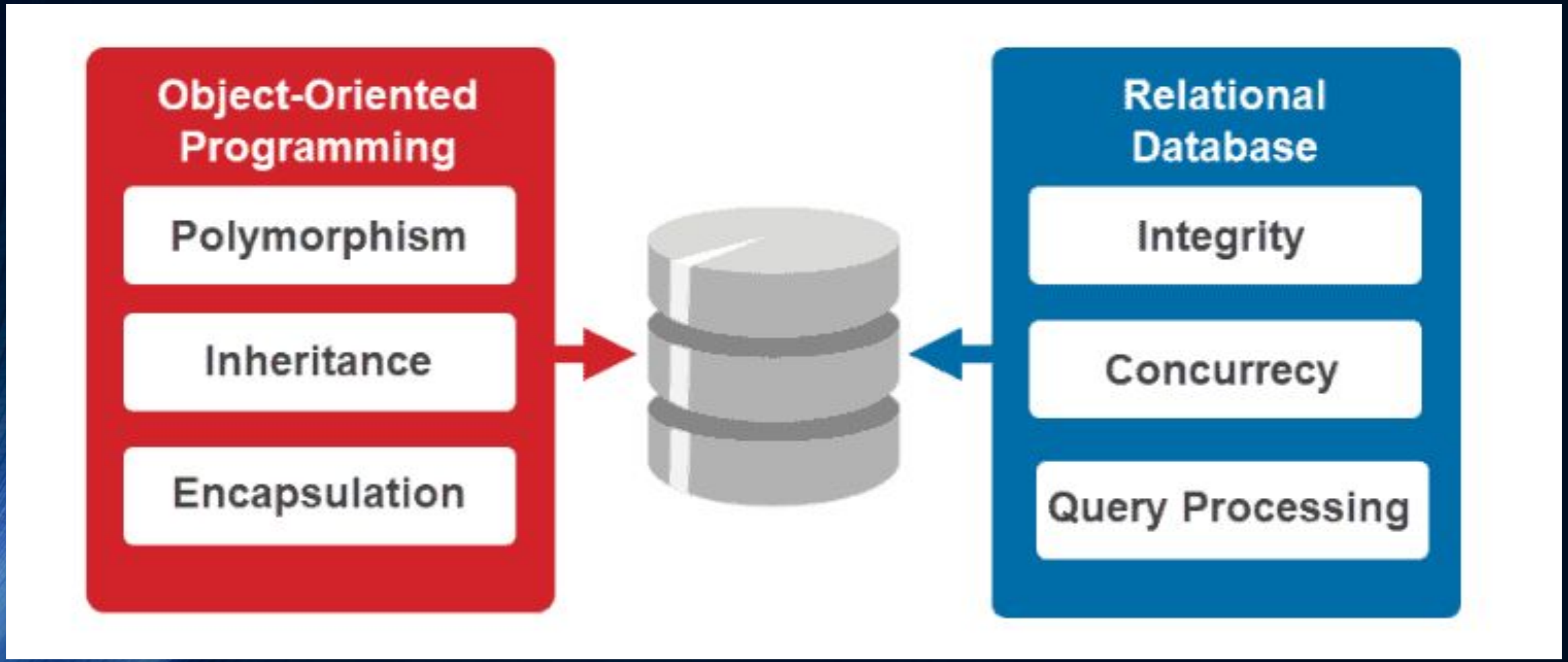


STUDENT

STUDENT_ID	NAME	AGE
1	PRIYANKA	20
2	JEREMY	21
3	PRIYANKA	20

II DBMS

Object based data model



Course Contents

Unit II	Testing and Data Modeling	(07 Hours)
Random Numbers and Simulation: Sampling of continuous distributions, Monte Carlo methods **Hypothesis Testing:** Type I and II errors, rejection regions; Z-test, T-test, F-test, Chi-Square test, Bayesian test **Stochastic Processes and Data Modeling:** Markov process, Hidden Markov Models, Poisson Process, Gaussian Processes, Auto-Regressive and Moving average processes, Bayesian Network, Regression, Queuing systems		

Data analysis

- Statistics of a sample
  - Central tendency
  - Variation
  - Normal distribution
- Inference
  - From sample to population
  - P-value

Data analysis

- Statistics of a sample
  - Central tendency
  - Variation
  - Normal distribution
- Inference
  - From sample to population
  - P-value

Measures of Central Tendency

- Mean … the average score

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Median … the value that lies in the middle after ranking all the scores

$$X_M = \begin{cases} X_{n/2+1} & n\,odd \\ \dfrac{X_{n/2} + X_{n/2+1}}{2} & n\,even \end{cases}$$

- Mode … the most frequently occurring score

Variation or Spread of Distributions

- Range

$$Range = X_{Max} - X_{Min}$$

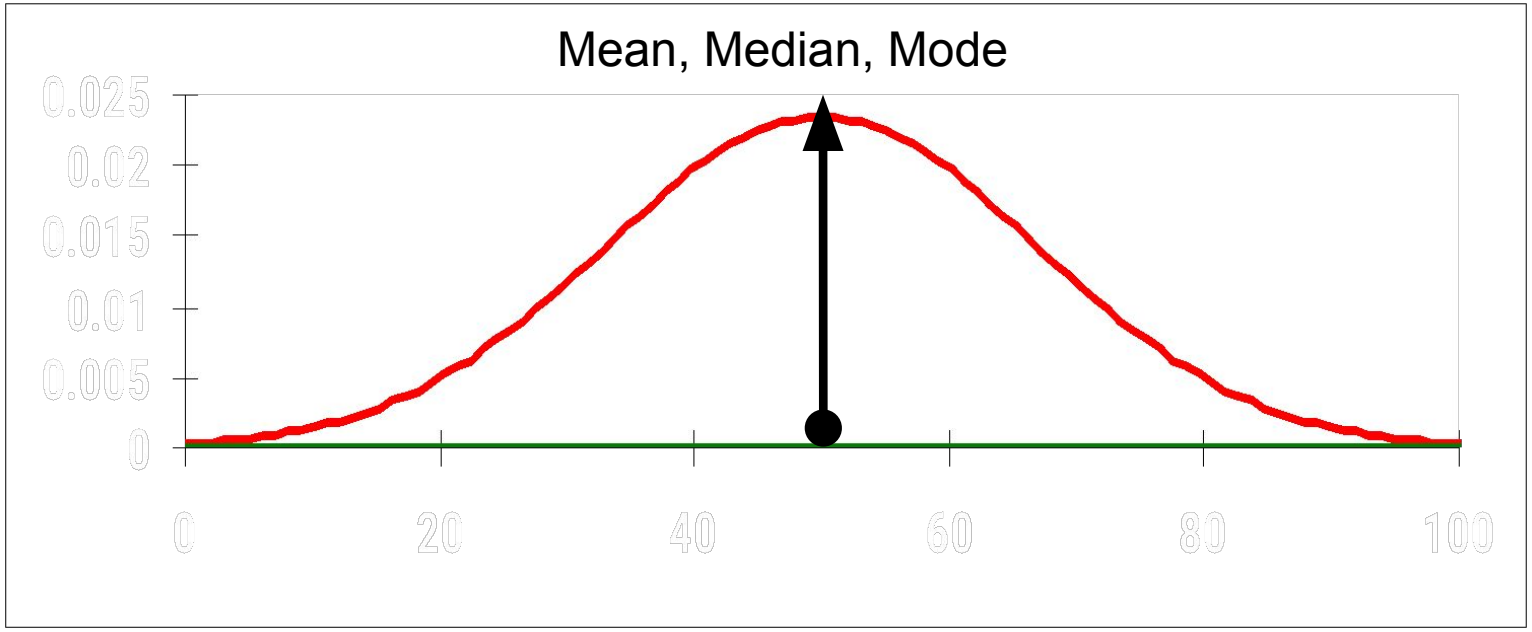
- Variance and Standard Deviation

$$Var(X) = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

$$Std(X) = \sigma = \sqrt{Var(X)}$$

The Normal Distribution Curve

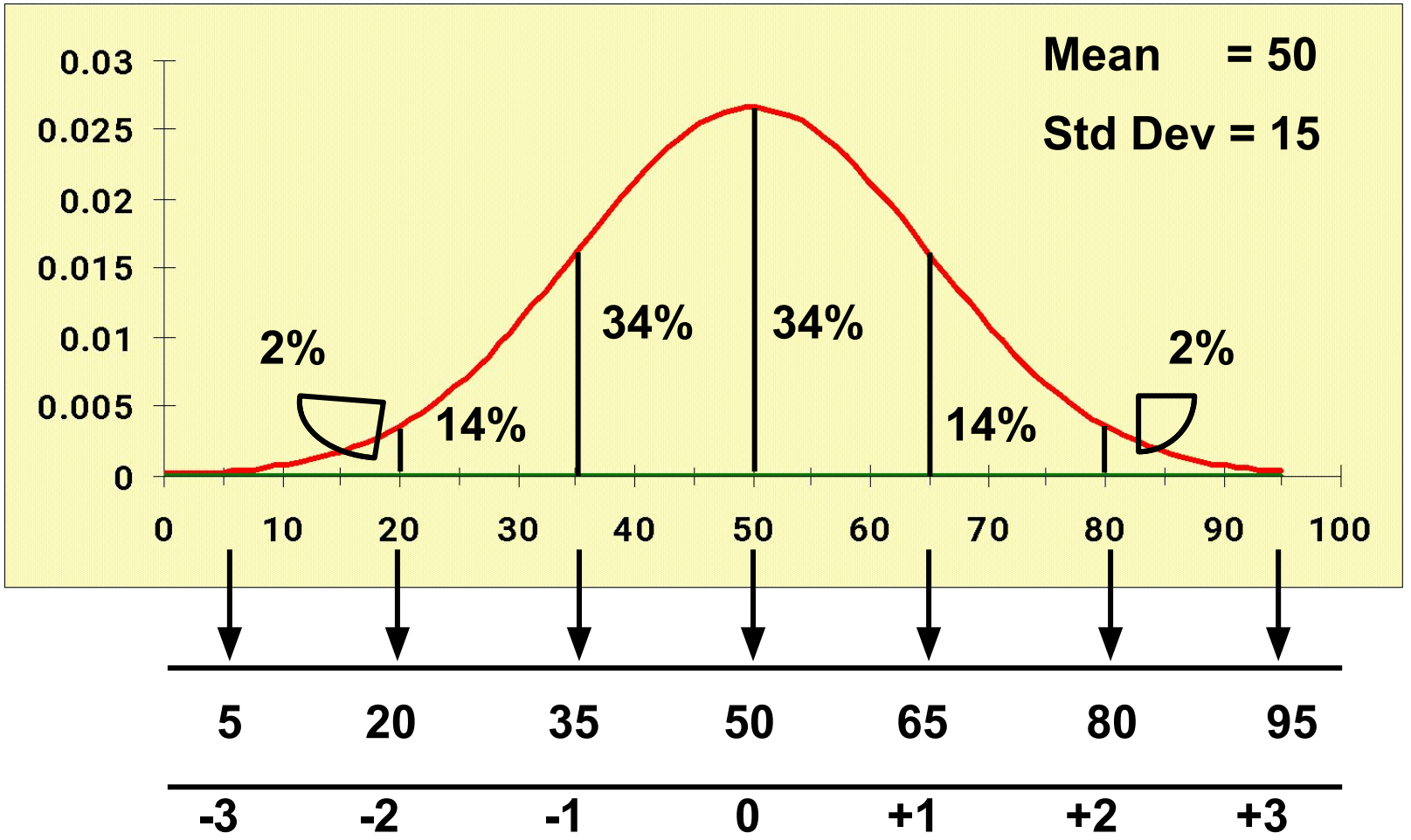
In everyday life many variables such as height, weight, shoe size and exam marks all tend to be normally distributed, that is, they all tend to look like:



It is bell-shaped and symmetrical about the mean

The mean, median and mode are equal

Interpreting a normal distribution



Data analysis

- Statistics of a sample
  - Central tendency
  - Variation
  - Normal distribution

- Inference
  - From sample to population
  - P-value

Statistical Inference

The process of making guesses about the truth from a sample

$$\hat{\mu} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{X}_n\right)^2}{n-1}$$

Truth (not observable)

Population parameters

$$\mu = \frac{\sum_{i=1}^{N} x}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sample (observation)

Make guesses about the whole population

The Central Limit Theorem

If all possible random samples, each of size *n*, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

1. have mean: $\mu_{\bar{x}} = \mu$

2. have standard deviation: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

(standard error)

3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger *n*)

Distribution of the sample mean, computer simulation…

- Specify the underlying distribution of vitamin D in all European men aged 40 to 79.
  - Right-skewed
  - Standard deviation = 33 nmol/L
  - True mean = 62 nmol/L
- Select a random sample of 100 virtual men from the population.
- Calculate the mean vitamin D for the sample.
- Repeat steps (2) and (3) a large number of times (say 1000 times).
- Explore the distribution of the 1000 means.

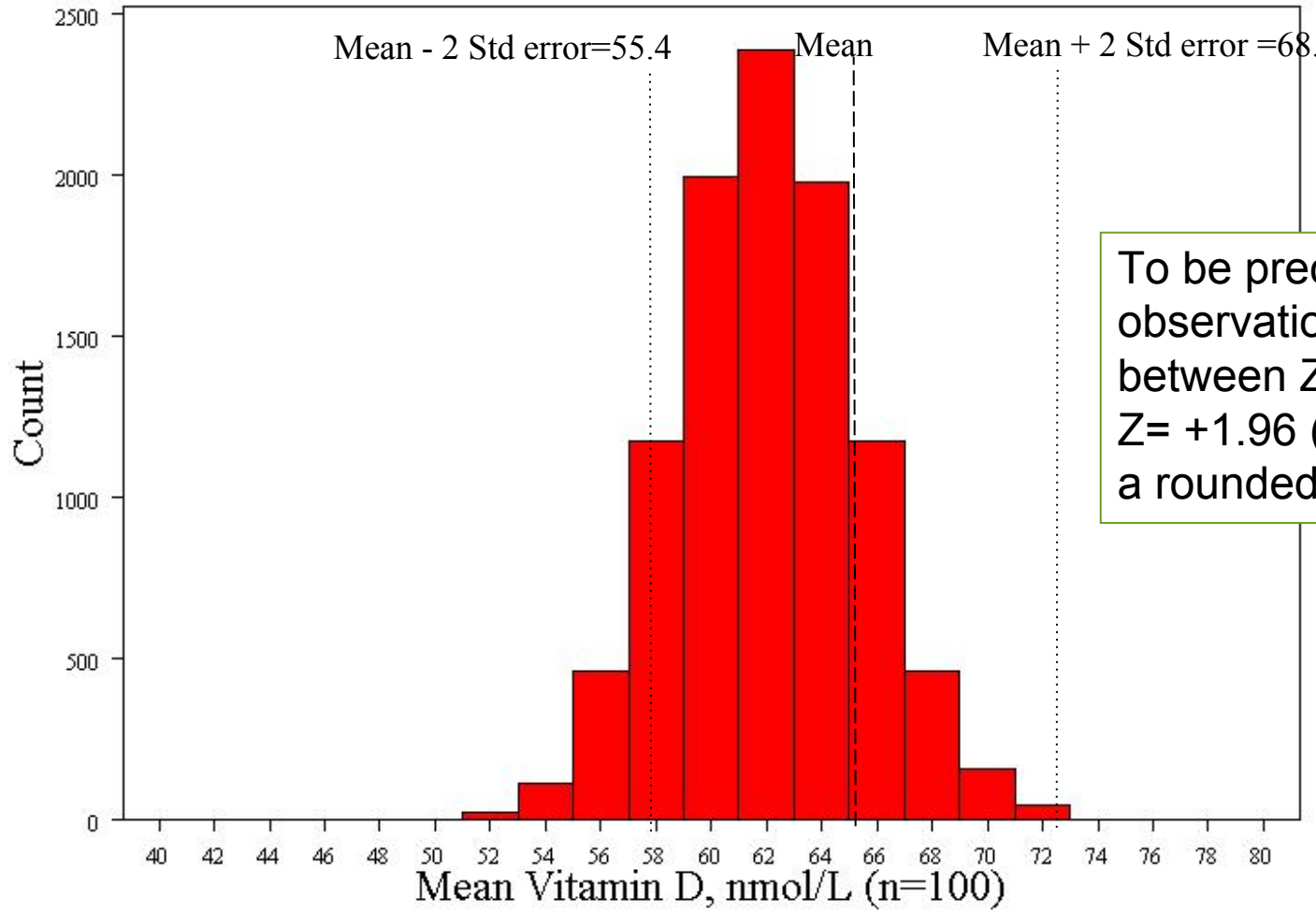
Confidence interval

Given a sample and its statistics (mean and standard deviation), is it possible to get an estimate of the true mean?

The confidence interval is set to capture the true effect "most of the time".

For example, a 95% confidence interval should include the true effect about 95% of the time.

Recall: 68-95-99.7 rule for normal distributions! These is a 95% chance that the sample mean will fall within two standard errors of the true mean= 62 +/- 2*3.3 = 55.4 nmol/L to 68.6 nmol/L



To be precise, 95% of observations fall between Z=-1.96 and Z= +1.96 (so the “2” is a rounded number)…

Confidence interval

The value of the statistic in the sample (mean)

point estimate ± (measure of how confident we want to be) × (standard error)

Standard error of the statistics

From a Z table or a T table, depending on the sampling distribution of the statistic.

Confidence Level	*Z value*
80%	1.28
90%	1.645
95%	1.96
98%	2.33
99%	2.58
99.8%	3.08
99.9%	3.27

Introduction to Monte Carlo Methods

Dr. Araddhana Deshmukh

1. The problems to be solved

- The aim of Monte Carlo methods
  - Problem 1: generating samples from $\{\mathbf{x}^{(r)}\}_{r=1}^{R}$ a given target density $P(\mathbf{x})$.
  - Problem 2: estimating expectations of functions

- The accuracy of $\Phi =< \phi(\mathbf{x}) >\equiv \int d^N\mathbf{x} P(\mathbf{x})\phi(\mathbf{x})$ independent of the dimensionality of the space sampled.
  - The estimate of Φ will be decreased as σ^2/R.
  - Only a few independent R samples are sufficient.

$$\hat{\Phi} = \frac{1}{R}\sum_r \phi(\mathbf{x}^{(r)}), \quad \sigma^2 = \int d^N\mathbf{x} P(\mathbf{x})(\phi(\mathbf{x}) - \Phi)^2$$

1.1 Why is sampling from $P(\mathbf{x})$ Hard?

- Assumption: $P(\mathbf{x})$ can be evaluated.
  - Function $P^*(\mathbf{x})$ can be evaluated, where $P(\mathbf{x})=P^*(\mathbf{x})/Z$.
- But why can problem1 be easily solved?
  - Normalizing term Z

$$Z = \int d^N \mathbf{x} P^*(\mathbf{x})$$

  - Even if knowing Z sampling is still challenging in high-dim.
    - Visiting every location in $\mathbf{x}$ can not be possible.
  - In Gaussian, a sample can be generated by

$$\cos(2\pi u_1)\sqrt{2\log(1/u_2)}$$

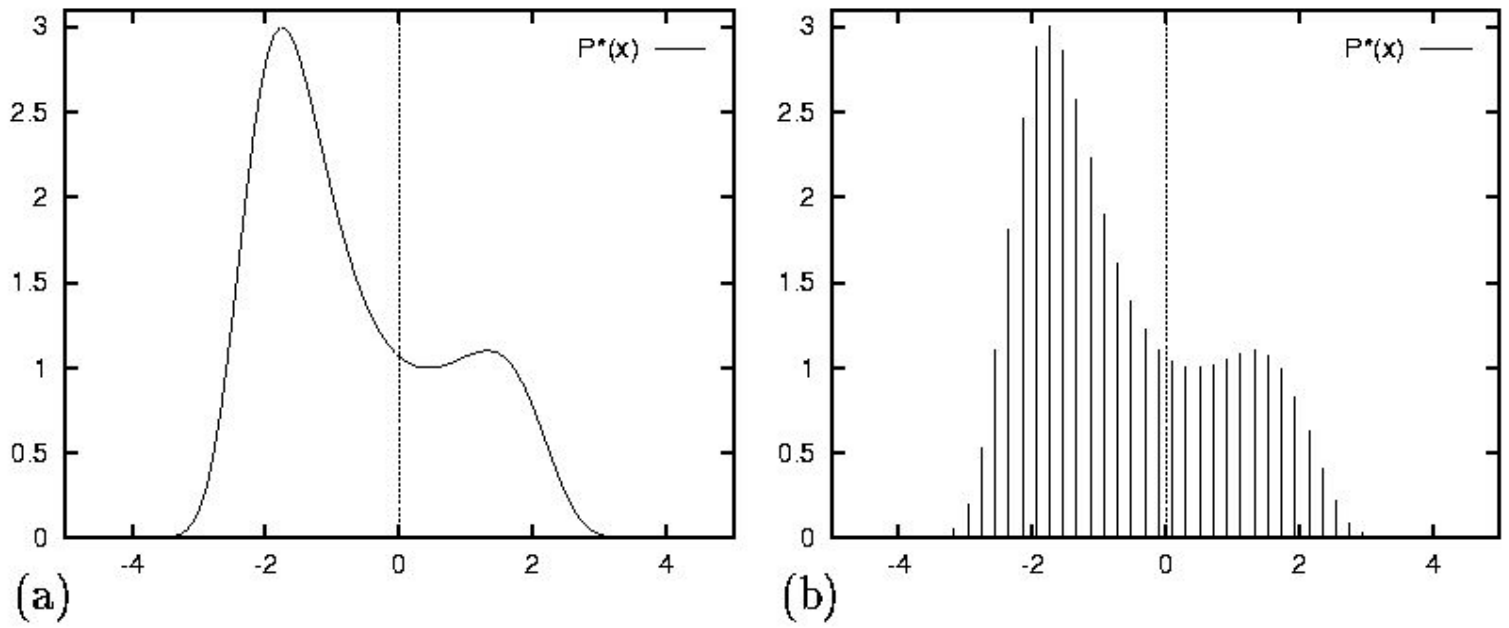


Figure 1. (a) The function $P^*(x) = \exp\left[0.4(x-0.4)^2 - 0.08x^4\right]$. How to draw samples from this density? (b) The function $P^*(x)$ evaluated at a discrete set of uniformly spaced points $\{x_i\}$. How to draw samples from this discrete distribution?

- If 50 discretized point, and dim is 1000 then 50^{1000} evaluation of $P^*(\mathbf{x})$ will be need.

1.2 Uniform Sampling

- Solve problem 2 by drawing random samples uniformly from the state space and evaluating $P^*(\mathbf{x})$.

$$Z_R = \sum_{r=1}^{R} P^*(\mathbf{x}^{(r)}), \hat{\Phi} = \sum_{r=1}^{R} \phi(\mathbf{x}^{(r)}) \frac{P^*(\mathbf{x}^{(r)})}{Z_R}$$

  - Typical set T, whose volume is $|T| \approx 2^{H(\mathbf{X})}$
    - Shannon-Gibbs entropy

$$H(\mathbf{X}) = \sum_{\mathbf{x}} P(\mathbf{x}) \log_2 \frac{1}{P(\mathbf{x})}$$

    - Uniform sampling has meaning only when many samples hit the typical set.
    - In Ising model, $R_{\min} \approx 2^{N-H}$.
    - Our concern is when temperature is intermediate, specially $N/2$.
    - If $N = 1000$, about 10^{150} samples is needed.

2. Importance Sampling

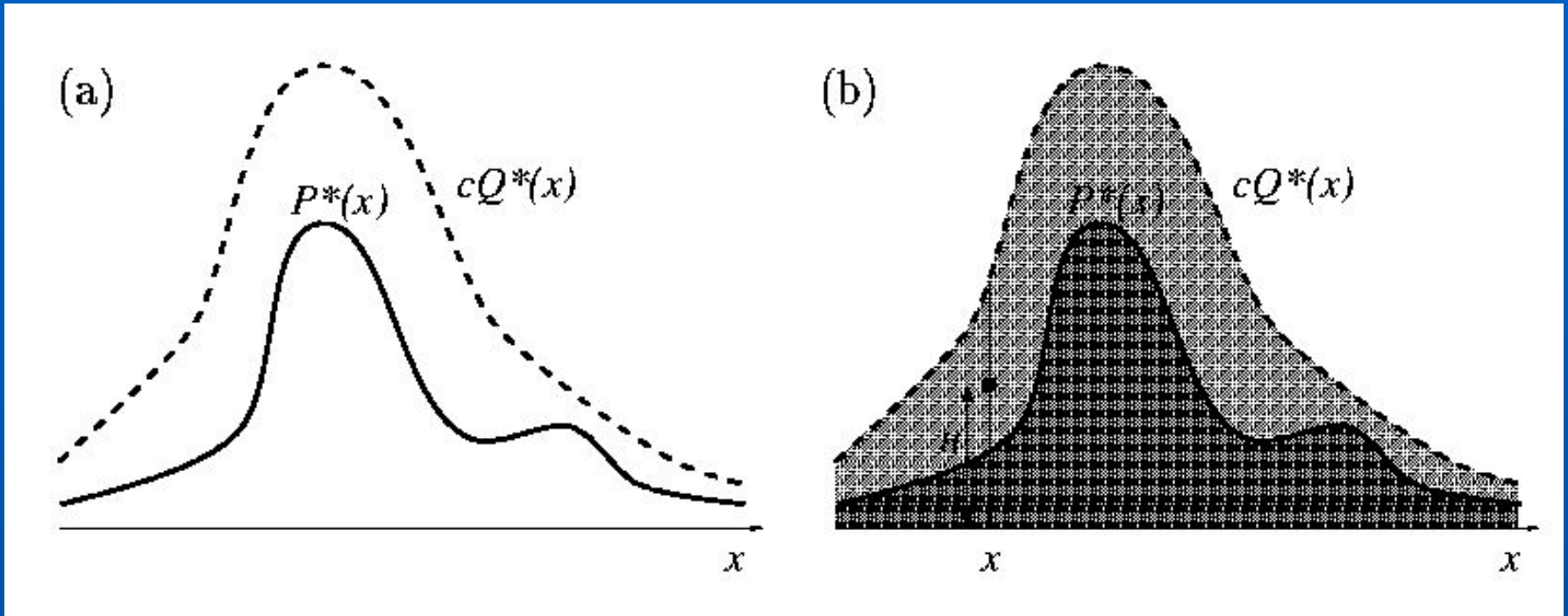
- Not a Problem 1, but a Problem 2.
  - Generalization of the uniform sampling method.
- $P(\mathbf{x})$ is too complicated for us to be able to sample from it directly.
  - Introducing a simpler density $Q(\mathbf{x})$ where $Q(\mathbf{x}) = Q^*(\mathbf{x})/Z_Q$.
  - Over-represented, under-represented.
    - Introducing weights

$$\hat{\Phi} \equiv \frac{\sum_r w_r \phi(\mathbf{x}^{(r)})}{\sum_r w_r} \quad w_r = \frac{P^*(\mathbf{x}^{(r)})}{Q^*(\mathbf{x}^{(r)})}$$

    - If $Q(\mathbf{x})$ is non- ero, the estimate of Φ converges to Φ.
    - Practically, it is hard to estimate how reliable the estimator.

3. Rejection Sampling

- Assumption
  - one-dim. Density $P(x)=P^*(x)/Z$, *proposal density* $Q(x)$,
  - for all x, $cQ^*(x)>P^*(x)$
- Methods
  - 1. Generate two random numbers.
    - x from $Q(x)$
    - u from uniformly distributed in interval $[0, cQ^*(x)]$
  - 2. Accept or reject the sample x.
    - reject if $u > P^*(x)$
    - accept otherwise
    - Acceptance means adding x to samples $\{x^{(r)}\}$
  - This procedure generate samples from $P(x)$.


(a)
P*(x)
cQ*(x)
x
(b)
P*(x)
cQ*(x)
x
x

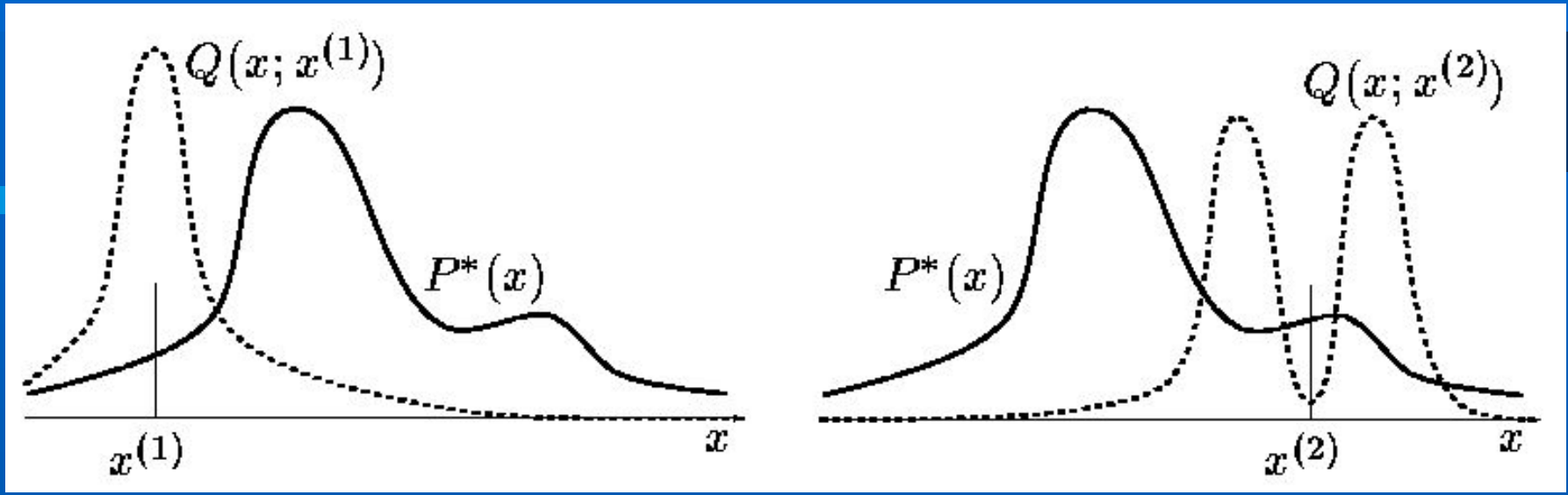
The Metropolis method

- Metropolis algorithm make use of a proposal density Q which depends on the current state $x^{(t)}$.
  - Proposal density $Q(x)$ similar to $P(x)$ is not so easy to be composed.
- Methods

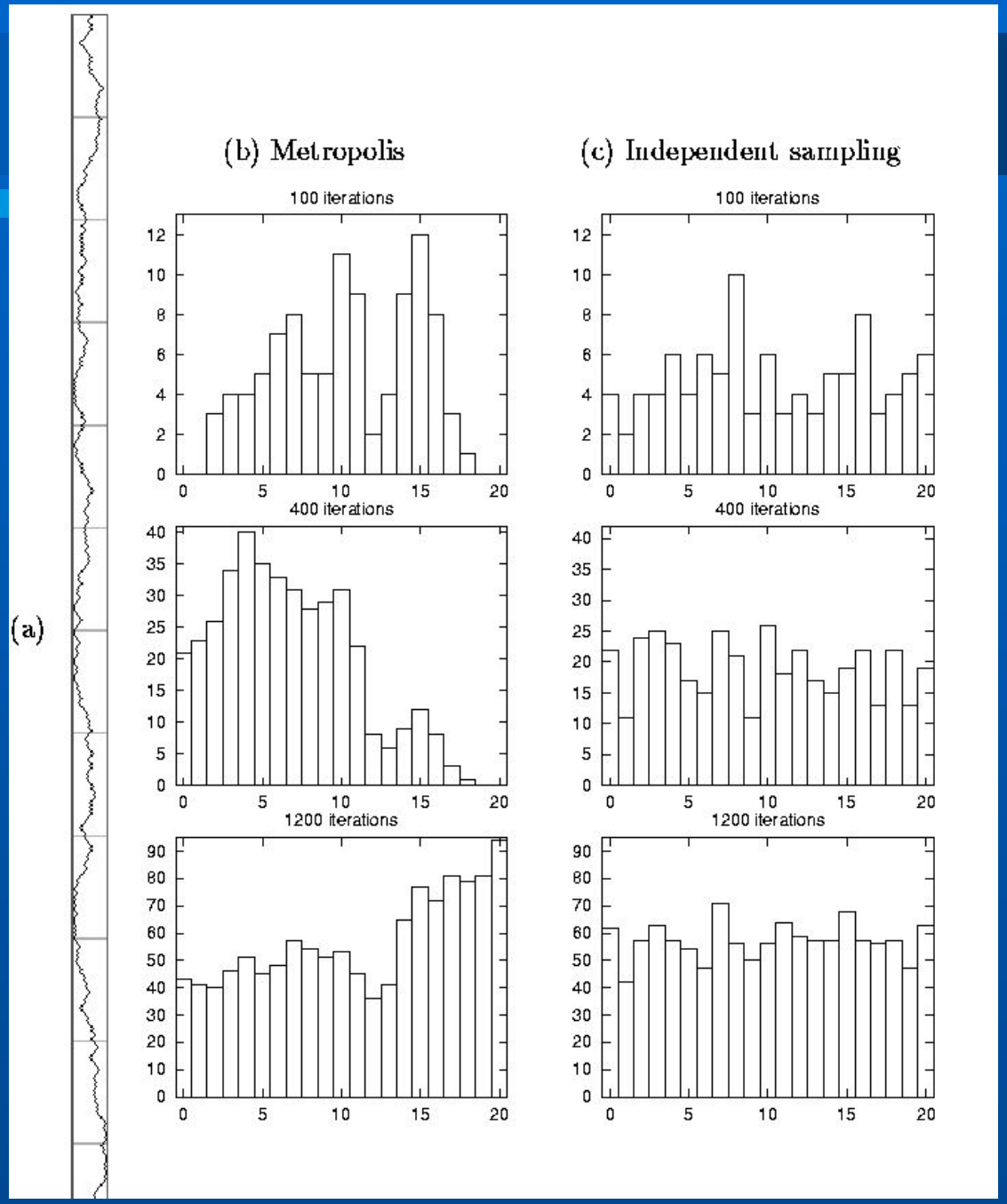
$$a = \frac{P^*(x')}{P^*(x^{(t)})} \frac{Q(x^{(t)}; x')}{Q(x'; x^{(t)})}.$$

If $a \geq 1$ then the new state is accepted.
Otherwise, the new state is accepted with probability a.

  - If accepted $x^{(t+1)}=x'$, if rejected $x^{(t+1)}=x^{(t)}$.
  - Differs from rejection sampling.



- Metropolis's T iterations does not produce T independent samples from P. They are correlated.
- As $t \to \infty$, the probability distribution of $x^{(t)}$ tends to $P(x)=P^*(x)/Z$.
- It is MCMC.
  - $x^{(t)}$ are correlated.
  - *Rejection sampling* is not MCMC, because $x^{(r)}$ are independent samples from the desired distribution.

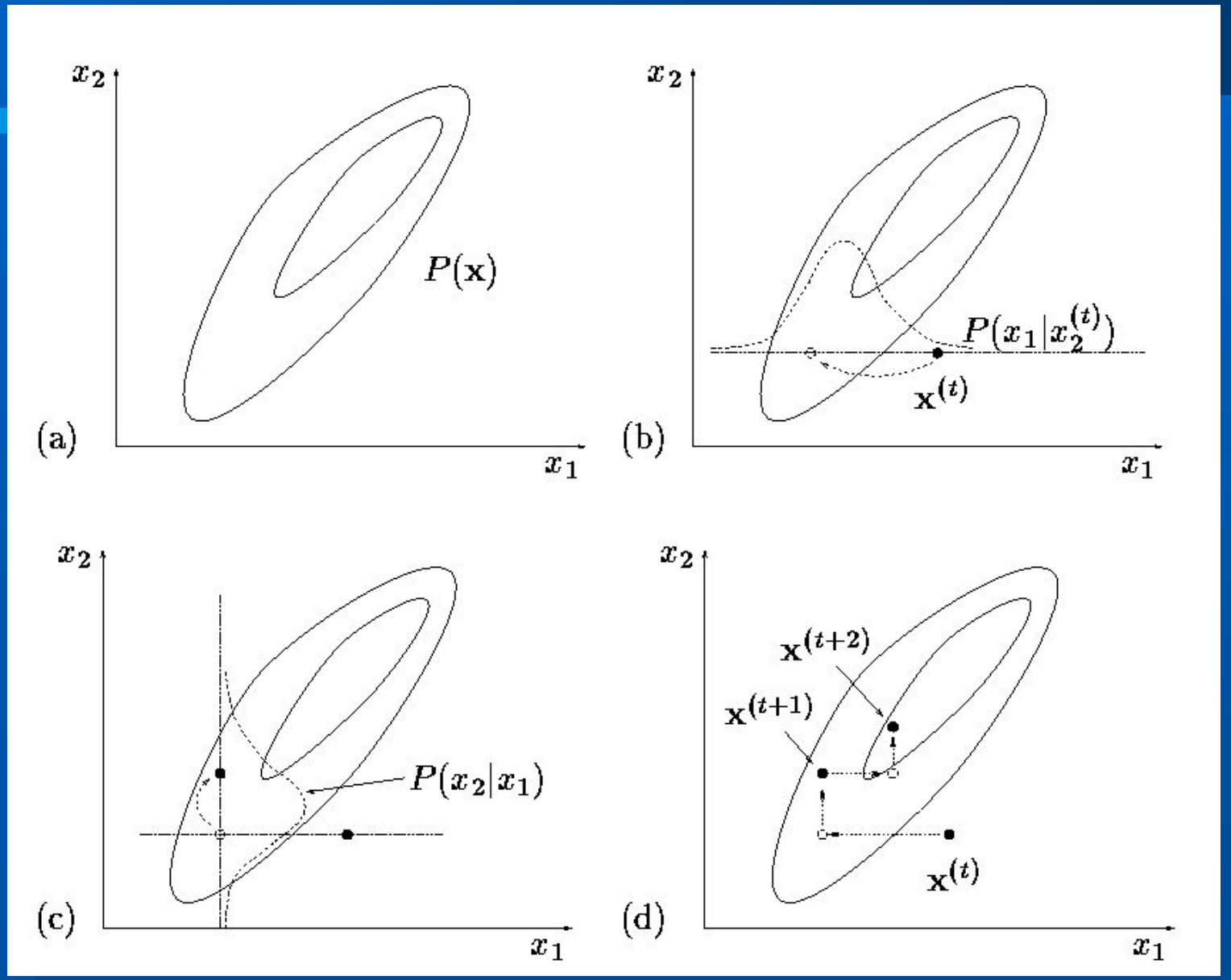

(a)
(b) Metropolis
(c) Independent sampling
100 iterations
400 iterations
1200 iterations
100 iterations
400 iterations
1200 iterations

Gibbs Sampling

- Methods

$$
\begin{aligned}
x_1^{(t+1)} &\sim P(x_1 | x_2^{(t)}, x_3^{(t)}, \ldots x_K^{(t)}) \\
x_2^{(t+1)} &\sim P(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots x_K^{(t)}) \\
x_3^{(t+1)} &\sim P(x_3 | x_1^{(t+1)}, x_2^{(t+1)}, \ldots x_K^{(t)}), \text{etc.}
\end{aligned}
$$

  - Gibbs sampling can be viewed as a Metropolis method which has the property that every proposal is always accepted.


x_2
$P(\mathbf{x})$
(a)
x_1
x_2
$P(x_1|x_2^{(t)})$
$\mathbf{x}^{(t)}$
(b)
x_1
x_2
$P(x_2|x_1)$
(c)
x_1
x_2
$\mathbf{x}^{(t+2)}$
$\mathbf{x}^{(t+1)}$
$\mathbf{x}^{(t)}$
(d)
x_1

Speeding up Monte Carlo Methods

- 7.1.1 Reducing random walk behavior in Metropolis methods.
  - Hybrid Monte Carlo: continuous state spaces which makes use of gradient information to reduce random walk behavior.
    - If gradient is available, there is no reason to use random walk.
    - Hamiltonian c $P(\mathbf{x}) = \frac{e^{-E(\mathbf{x})}}{Z}$
      - momentum variable p
      - $K(\mathbf{p})=\mathbf{p}^T\mathbf{p}$ $H(\mathbf{x},\mathbf{p}) = E(\mathbf{x}) + K(\mathbf{p})$
    - sampling from Joint density

$$P_H(\mathbf{x},\mathbf{p}) = \frac{1}{Z_H}\exp[-H(\mathbf{x},\mathbf{p})] = \frac{1}{Z_H}\exp[-E(\mathbf{x})]\exp[-K(\mathbf{p})]$$

- If the simulation of the Hamiltonian dynamics is numerically perfect then the proposals are accepted every time.
- $H(\mathbf{x},\mathbf{p})$ is a constant of the motion and a is equal to one.
- If the simulation is imperfect, then rejection is made using the change of $H(\mathbf{x},\mathbf{p})$.

7.5. How many samples are needed?

- The variance of estimator Φ depends only on the number of independent samples R and

$$\sigma^2 = \int d^N \mathbf{x} P(\mathbf{x})(\phi(\mathbf{x}) - \Phi)^2$$

- There is little point in knowing Φ to a precision finer than about $\sigma/3$.
- Then, R=12 is sufficient.

7.7. Philosophy

- Monte Carlo methods are all non-Bayesian.
  - In Monte Carlo, computer experiments are used to calculate the *estimators* Φ of quantities of interest.
  - Bayesian approach use the experiments results to infer the properties of the $P(\mathbf{x})$ and generate predictive distribution for quantities of interest such as Φ.
    - It only depends on the computed values $P^*(\mathbf{x}^{(r)})$ at the points $\{\mathbf{x}^{(r)}\}$.

8. Summary

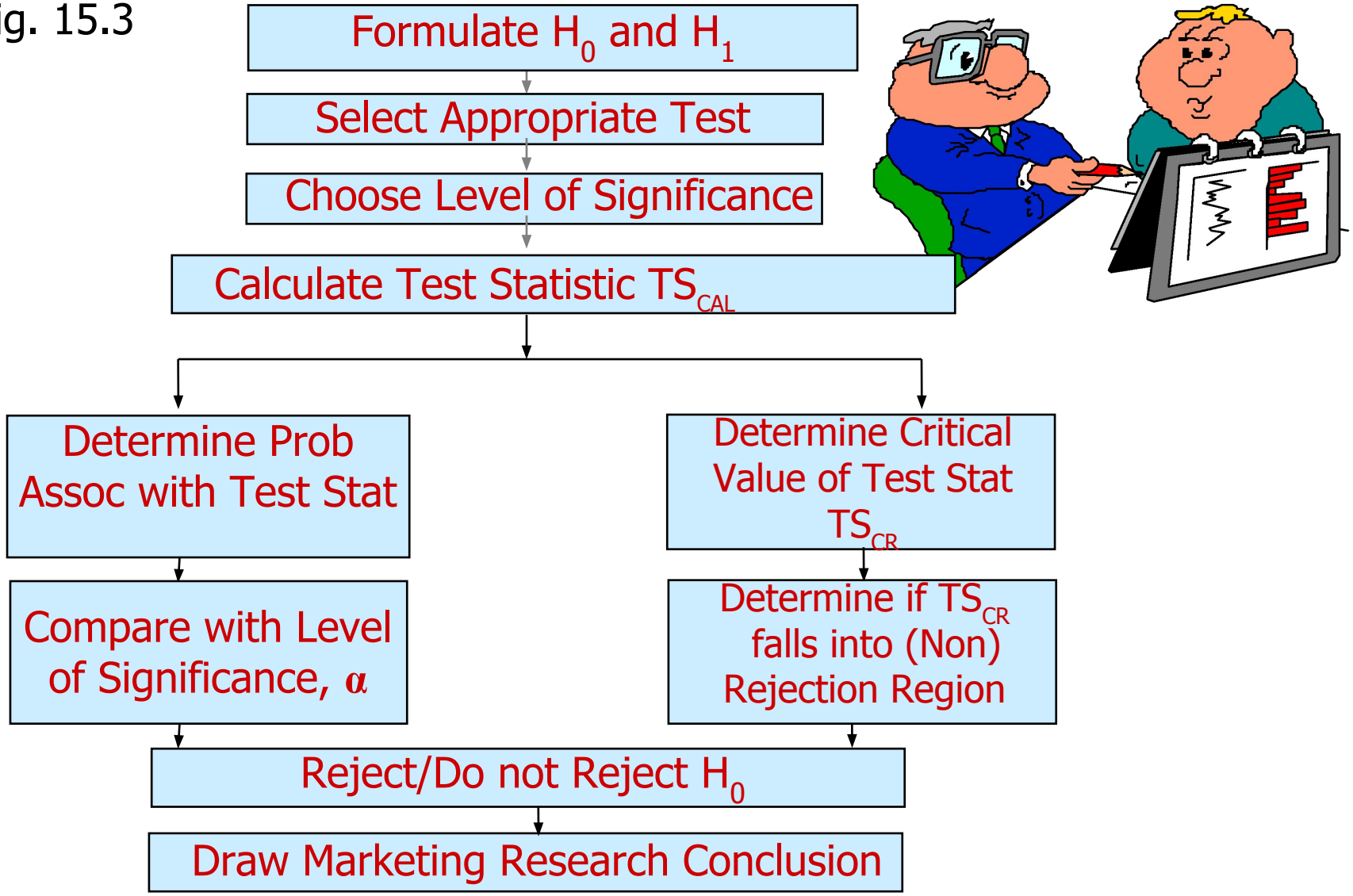
- Monte Carlo methods are a powerful tool that allow one to implement any probability distribution that can be expressed in the form $P(\mathbf{x})=P^*(\mathbf{x})/Z$.
- Monte Carlo methods can answer virtually any query related to $P(\mathbf{x})$ by putting the query in the form

$$\int \Phi(\mathbf{x})P(\mathbf{x}) \approx \frac{1}{R}\sum_r \phi(\mathbf{x}^{(r)})$$

Hypothesis Testing

Steps for Hypothesis Testing

Fig. 15.3



Step 1: Formulate the Hypothesis

- A **null hypothesis** is a statement of the status quo, one of no difference or no effect. If the null hypothesis is not rejected, no changes will be made.
- An **alternative hypothesis** is one in which some difference or effect is expected.
- The null hypothesis refers to a specified value of the population parameter (e.g., μ, σ, π), not a sample statistic (e.g., $\bar{X}$).

Step 1: Formulate the Hypothesis

- A null hypothesis may be rejected, but it can never be accepted based on a single test.
- In marketing research, the null hypothesis is formulated in such a way that its rejection leads to the acceptance of the desired conclusion.
- A new Internet Shopping Service will be introduced if more than 40% people use it:

$$H_0: \pi \leq 0.40$$

$$H_1: \pi > 0.40$$

Step 1: Formulate the Hypothesis

- In eg on previous slide, the null hyp is a **one-tailed test**, because the alternative hypothesis is expressed directionally.
- If not, then a **two-tailed test** would be required as foll:

$$H_0: \pi = 0.40$$

$$H_1: \pi \neq 0.40$$

Step 2: Select an Appropriate Test

- The **test statistic** measures how close the sample has come to the null hypothesis.
- The test statistic often follows a well-known distribution (eg, normal, *t*, or chi-square).
- In our example, the *z* statistic, which follows the standard normal distribution, would be appropriate.

$$z = \frac{p - \pi}{\sigma_p}$$

Where σ_p is standard deviation

Step 3: Choose Level of Significance

Type I Error

- Occurs if the null hypothesis is rejected when it is in fact true.
- The probability of type I error (α) is also called the **level of significance**.

Type II Error

- Occurs if the null hypothesis is not rejected when it is in fact false.
- The probability of type II error is denoted by β .
- Unlike α, which is specified by the researcher, the magnitude of β depends on the actual value of the population parameter (proportion).

It is necessary to balance the two types of errors.

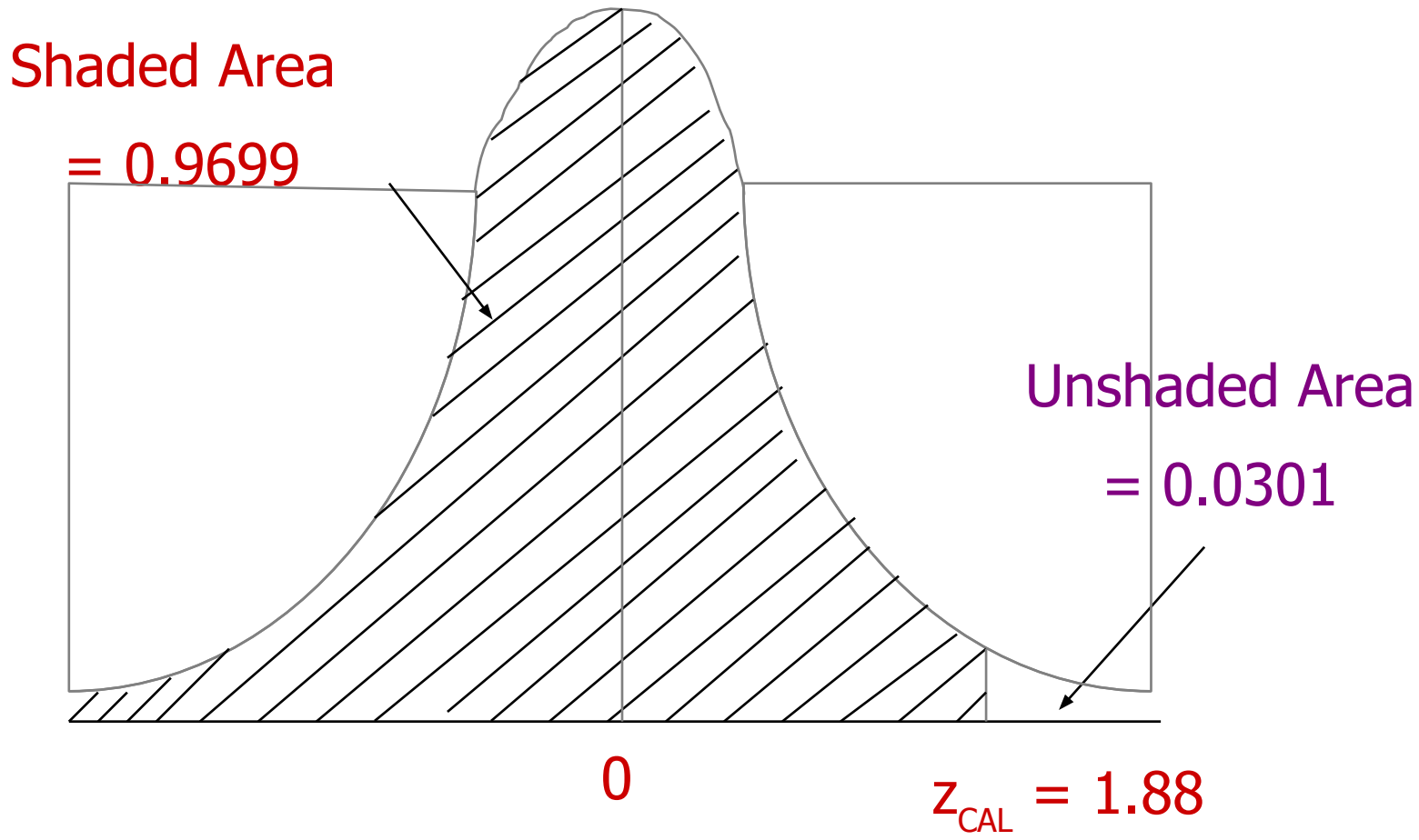
Step 3: Choose Level of Significance

Power of a Test

- The **power of a test** is the probability (1 - β) of rejecting the null hypothesis when it is false and should be rejected.
- Although β is unknown, it is related to α. An extremely low value of α (e.g., = 0.001) will result in intolerably high β errors.

Probability of z with a One-Tailed Test

Fig. 15.5



Step 4: Collect Data and Calculate Test Statistic

- The required data are collected and the value of the test statistic computed.
- In our example, 30 people were surveyed and 17 shopped on the internet. The value of the sample proportion is $\hat{p}$= 17/30 = 0.567.
- The value of σ_p is:

$$\sigma_p = 0.089$$

Step 4: Collect Data and Calculate Test Statistic

The test statistic *z* can be calculated as follows:

$$z_{CAL} = \frac{\hat{p} - \pi}{\sigma_p}$$

$$= \frac{0.567-0.40}{0.089}$$

$$= 1.88$$

Step 5: Determine Probability Value/ Critical Value

- Using standard normal tables (Table 2 of the Statistical Appendix), the area to the right of z_{CAL} is .0301 (z_{CAL} =1.88)
- The shaded area between 0 and 1.88 is 0.4699. Therefore, the area to the right of 1.88 is 0.5 - 0.4699 = 0.0301.
- Thus, the p-value is .0301

- Alternatively, the critical value of *z*, called z_{α}, which will give an area to the right side of the critical value of α=0.05, is between 1.64 and 1.65. Thus z_{α} =1.645.
- Note, in determining the critical value of the test statistic, the area to the right of the critical value is either α or α/2. It is α for a one-tail test and α/2 for a two-tail test.

Steps 6 & 7: Compare Prob and Make the Decision

- If the prob associated with the calculated value of the test statistic (z_{CAL}) is less than the level of significance (α), the null hypothesis is rejected.
- In our case, the p-value is 0.0301.This is less than the level of significance of α =0.05. Hence, the **null hypothesis is rejected**.
- Alternatively, if the calculated value of the test statistic is greater than the critical value of the test statistic (z_{α}), the null hypothesis is rejected.

Steps 6 & 7: Compare Prob and Make the Decision

- The calculated value of the test statistic $z_{CAL} = 1.88$ lies in the rejection region, beyond the value of $z_{\alpha}=1.645$. Again, the same conclusion to reject the null hypothesis is reached.
- Note that the two ways of testing the null hypothesis are equivalent but mathematically opposite in the direction of comparison.
- Writing Test-Statistic as TS:

  If the probability of TS_{CAL} < significance level (α) then reject H_0 but if $TS_{CAL} > TS_{CR}$ then reject H_0.

Step 8: Mkt Research Conclusion

- The conclusion reached by hypothesis testing must be expressed in terms of the marketing research problem.

- In our example, we conclude that there is evidence that the proportion of Internet users who shop via the Internet is significantly greater than 0.40. Hence, the department store should introduce the new Internet shopping service.

Another Test

- Assume that the random variable X is normally dist, with unknown pop variance estimated by the sample variance s^2.
- Then a **t test** is appropriate.
- The t-statistic, $t = (\overline{X} - \mu)/s_{\overline{X}}$ is t distributed with n - 1 df.
- The **t dist** is similar to the normal distribution: bell-shaped and symmetric. As the number of df increases, the t dist approaches the normal dist.

One Sample : t Test

For the data in Table 15.1, suppose we wanted to test the hypothesis that the mean familiarity rating exceeds 4.0, the neutral value on a 7 point scale. A significance level of $\alpha = 0.05$ is selected. The hypotheses may be formulated as:

$$H_0: \mu \leq 4.0$$

$$H_1: \mu > 4.0$$

$$t = (\bar{X} - \mu)/s_{\bar{X}}$$

$$s_{\bar{X}} = 0.293$$

$$t_{CAL} = (4.724\text{-}4.0)/0.293 = 2.471$$

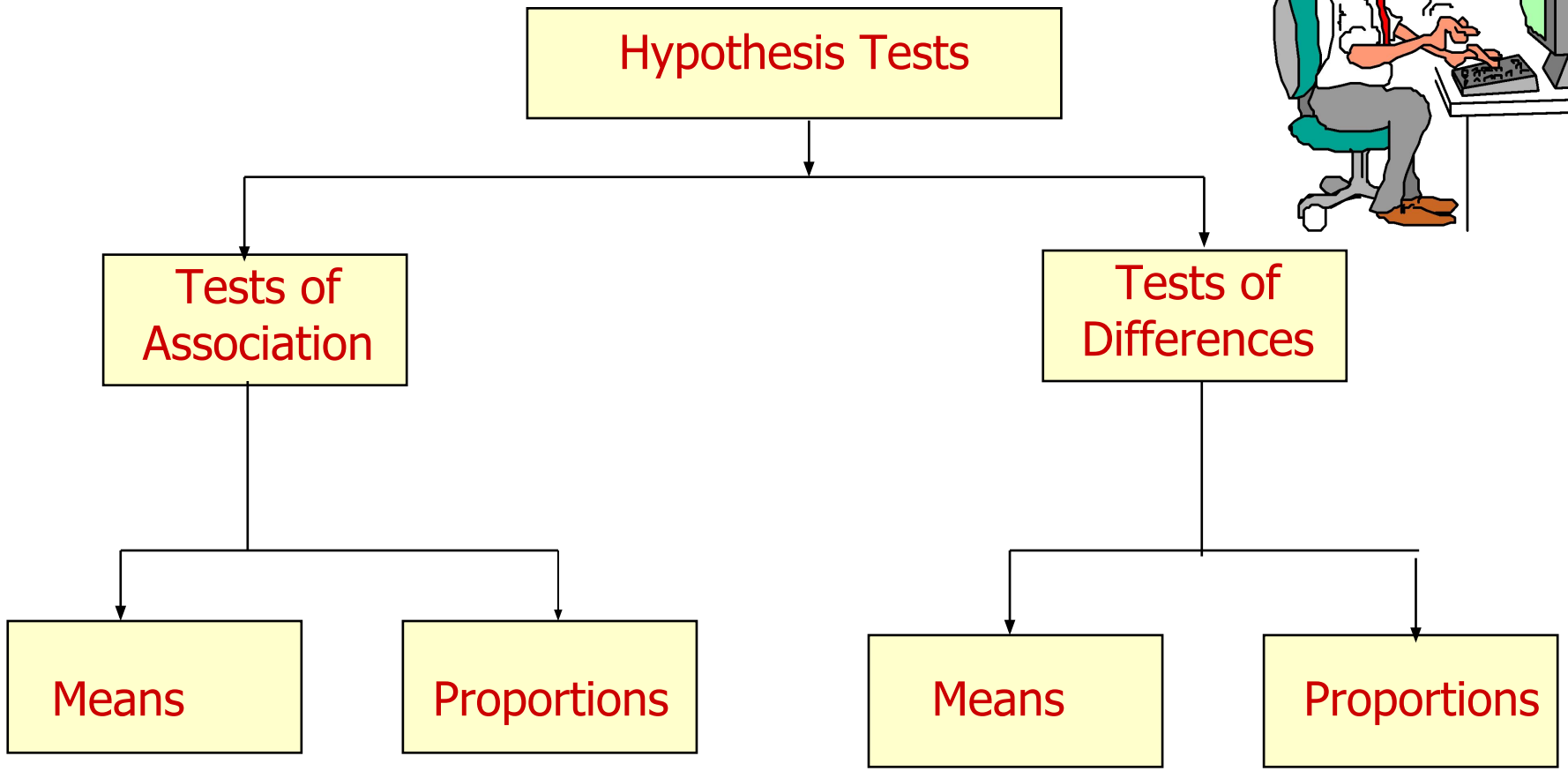
One Sample : t Test

- The df for the *t* stat is *n* - 1. In this case, *n* - 1 = 28.
- The probability assoc with 2.471 is less than 0.05. So the null hypothesis is rejected

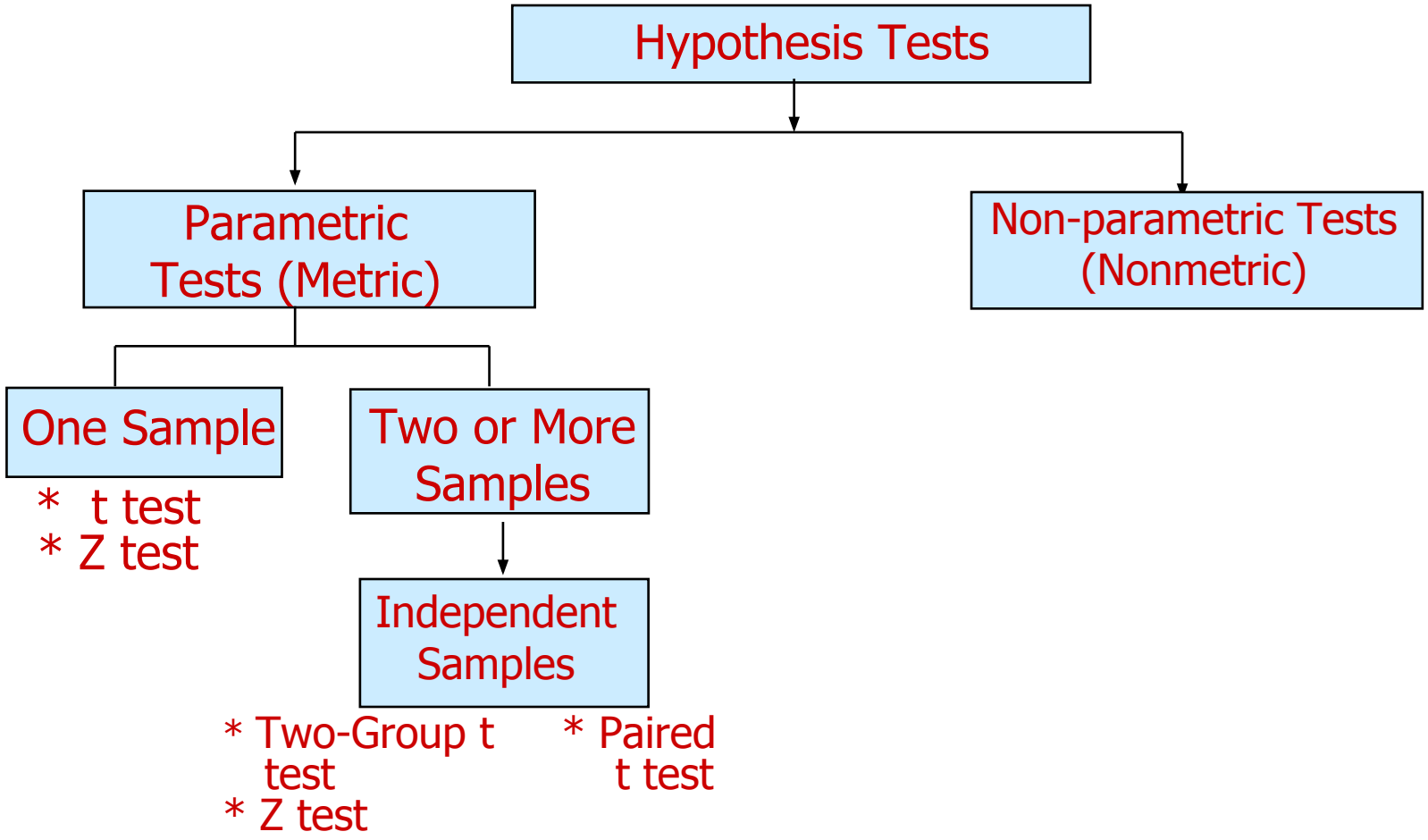
- Alternatively, the critical t_{α} value for a significance level of 0.05 is 1.7011
- Since, 1.7011 <2.471, the null hypothesis is rejected.

- The familiarity level does exceed 4.0.
- Note that if the population standard deviation was **known** to be 1.5, rather than estimated from the sample, a ***z* test** would be appropriate.

Broad Classification of Hyp Tests



Hypothesis Testing for Differences



Two Independent Samples: Means

- In the case of means for two independent samples, the hypotheses take the following form.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

- The two populations are sampled and the means and variances computed based on samples of sizes *n*1 and *n*2.
- The idea behind the test is similar to the test for a single mean, though the formula for standard error is different
- *Suppose we want to determine if internet usage is different for males than for females, using data in Table 15.1*

Internet Usage Data

Table 15.1

Respondent Number	Sex	Familiarity	Internet Usage	Attitude Toward Internet	Attitude Toward Technology	Usage of Internet Shopping	Usage of Internet Banking
1	1.00	7.00	14.00	7.00	6.00	1.00	1.00
2	2.00	2.00	2.00	3.00	3.00	2.00	2.00
3	2.00	3.00	3.00	4.00	3.00	1.00	2.00
4	2.00	3.00	3.00	7.00	5.00	1.00	2.00
5	1.00	7.00	13.00	7.00	7.00	1.00	1.00
6	2.00	4.00	6.00	5.00	4.00	1.00	2.00
7	2.00	2.00	2.00	4.00	5.00	2.00	2.00
8	2.00	3.00	6.00	5.00	4.00	2.00	2.00
9	2.00	3.00	6.00	6.00	4.00	1.00	2.00
10	1.00	9.00	15.00	7.00	5.00	1.00	2.00
11	2.00	4.00	3.00	4.00	3.00	2.00	2.00
12	2.00	5.00	4.00	6.00	4.00	2.00	2.00
13	1.00	6.00	8.00	8.00	5.00	2.00	1.00
14	1.00	6.00	8.00	3.00	2.00	2.00	2.00
15	1.00	6.00	5.00	5.00	4.00	1.00	2.00
16	2.00	4.00	3.00	4.00	3.00	2.00	2.00
17	1.00	6.00	8.00	5.00	3.00	1.00	1.00
18	1.00	4.00	4.00	5.00	4.00	1.00	2.00
19	1.00	7.00	14.00	6.00	6.00	1.00	1.00
20	2.00	6.00	6.00	6.00	4.00	2.00	2.00
21	1.00	6.00	8.00	4.00	2.00	2.00	2.00
22	1.00	5.00	5.00	5.00	4.00	2.00	1.00
23	2.00	3.00	2.00	4.00	2.00	2.00	2.00
24	1.00	7.00	15.00	6.00	6.00	1.00	1.00
25	2.00	6.00	6.00	5.00	3.00	1.00	2.00
26	1.00	6.00	13.00	6.00	6.00	1.00	1.00
27	2.00	5.00	4.00	5.00	5.00	1.00	1.00
28	2.00	4.00	2.00	3.00	2.00	2.00	2.00
29	1.00	4.00	4.00	5.00	3.00	1.00	2.00
30	1.00	3.00	3.00	7.00	5.00	1.00	2.00

Two Independent-Samples: *t* Tests

Table 15.14

Summary Statistics

	Number of Cases	Mean	Standard Deviation
Male	15	9.333	1.137
Female	15	3.867	0.435

***F* Test for Equality of Variances**

F value	2-tail probability
15.507	0.000

***t* Test**

Equal Variances Assumed			Equal Variances Not Assumed		
t value	Degrees of freedom	2-tail probability	*t* value	Degrees of freedom	2-tail probability
-4.492	28	0.000	-4.492	18.014	0.000

Two Independent Samples: Proportions

- Consider data of Table 15.1
- Is the proportion of respondents using the Internet for shopping the same for males and females?

The null and alternative hypotheses are:

$$H_0: \quad \pi_1 = \pi_2$$
$$H_1: \quad \pi_1 \neq \pi_2$$

- The test statistic is similar to the one for difference of means, with a different formula for standard error.

Summary of Hypothesis Tests for Differences

Sample	Application	Level of Scaling	Test/Comments
One Sample	Proportion	Metric	*Z* test
One Sample	Means	Metric	*t* test, if variance is unknown *z* test, if variance is known

Summary of Hypothesis Tests for Differences

Two Indep Samples	Application	Scaling	Test/Comments
Two indep samples	Means	Metric	Tw -group *t* test *F* test for equality of variances
Two indep samples	Proportions	Metric Nonmetric	*z* test Chi -square test

Stochastic Processes

A **stochastic process** is a model that evolves in time or space subject to **probabilistic laws.**

The simplest example is the one-dimensional simple **random walk..** The process starts in state $\mathbf{X_0}$ at time t = 0. Independently, at each time instance, the process takes a jump Z_n:

Prob $\{ Z_n = -1\} = q$, Prob $\{ Z_n = +1\} = p$ and Prob $\{ Z_n = 0 \} = 1 - p - q$.

The state of the process at time n is

$X_n = X_0 + Z_1 + Z_2 + \ldots + Z_n$.

Assume for convenience that $X_0 = 0$. Since $E[Z_n] = p - q$ and $VAR [Z_n] = p + q - (p - q)^2$, then $E[X_n] = n (p - q)$ and $Var [X_n] = n \{ p + q - (p - q)^2 \}$.

A stochastic process, such as the simple random walk, has the **memoryless** or **Markov property** if the conditional distribution of X_n only depends on the most recent information:

Prob $\{ X_n = k \mid X_{n-1} = a, X_{n-2} = b, \ldots \} =$ Prob $\{ X_n = k \mid X_{n-1} = a \}$

We can think of random walks as representing the position of a particle on an infinite line. The position of the particle can be **unrestricted,** or can be restricted by the presence of **barriers**. A barrier is **absorbing** if the process stops once the particle reaches the barrier, or **reflecting** if the particle remains at the barrier until a jump in the appropriate direction causes it to move away. Problems of interest are

What is the expected time to absorption at a barrier, if one exists ?

What is the distribution of time spent at a reflecting barrier, if one exists ?

Examples of Stochastic Processes.

Example [Reservoir Systems] Here Z_n is the inflow of water into a reservoir on day n. Once a particular water threshold a is reached, an amount of water b is released. The system is a random walk on the range [0, a] with a reflecting barrier at a.

Example [Company Cash Flow] X_0 is the initial capital of the company. During trading period i, the company receives revenue r_i and incurs costs c_i, so the change in liquidity is

$$z_i = r_i - c_i.$$

The company will continue to trade profitably as long as its accumulated capital is non zero. The underlying process is defined on the positive real line with an absorbing barrier at zero.

Example [Building Society Funds]. This is similar to the last example, except that the company pays out an amount b if the accumulated funds on a particular day exceeds an amount a. Building societies are designed to provide a steady flow of funds into the housing market and relatively simple models give insight into how the market can be regulated.

Example [Market Share] we are given the original market shares p_i of three companies and the **transition matrix**

$$P = [\, p_{i,j} \,]$$

where $p_{i,j}$ = Prob { that a customer of company i transfers to j over a single trading period}

P (Initial)	Final States 1	2	3
1	$p_{1,1}$	$p_{1,2}$	$p_{1,3}$
2	$p_{2,1}$	$p_{2,2}$	$p_{2,3}$
3	$p_{3,1}$	$p_{3,2}$	$p_{3,3}$

Stochastic processes of this type always reach a **steady state** which is an absorbing barrier and is independent of the **starting distribution**. The rate of convergence to the steady state depends on the values in the transition matrix.

The Infinite Single Server Queue M|M|1

In the simplest queue, customers arrive at an average rate λ to a queue with infinite capacity and one server. Assuming the Markov property holds, by taking very small time slices Δt,

Prob { 1 arrival in the interval $[t, t + \Delta t]$ } = $\lambda \Delta t$,
Prob { 0 arrivals in " $[t, t + \Delta t]$ } = $1 - \lambda \Delta t$,
Prob { More than 1 arrival in $[t, t + \Delta t]$ } = 0.

These are the classical conditions for the Poisson distribution, so

$P_n (t)$ = Prob { n arrivals in the interval [0, t] } = $(\lambda t)^n / n ! \exp (- \lambda t)$.

and Prob {Interarrival time $\leq t$} = Prob {First arrival $\leq t$} = 1-Prob {No arrival in [0, t]}
$= 1 - P_0 (t) = 1 - \exp (- \lambda t)$
so the interarrival time has an exponential distribution with parameter λ .

By the same token, if on average μ customers are served per unit time, then the service times have an exponential distribution with parameter μ. Since both the arrival and service distributions, this single server queue is designated M | M | 1.

The traffic intensity $\rho = \lambda / \mu$ is an important characteristic of queuing networks. Unless $\rho < 1$, the queue is unstable (i.e.) the expected queue size is infinite.

In queuing models, the **system** consists of those in the queue plus those, if any, being served. The main items of interest in queuing models are the means and variances of the

Waiting time for customers and the Queue or System **Sizes.**

Queueing Systems

Note that p_0 = Prob { no customers in the sytem }
so $1 - p_0$ = Prob { server is busy }

If the system size is bounded by n, great care needs to be taken in interpreting the behaviour of customers that arrive and find the queue full. If the excess customers are forced to return at a later time, the arrival rate is no longer Poisson. If the excess customers are lost, the transition diagram is finite and

$$p_n = (1 - \rho) \rho^{i} / (1 - \rho^{n+1}), \qquad \text{for } i = 0, 1, 2, \ldots, n.$$

In banks and other systems with k servers, it is common for customers to form a single queue, and when a server is available to go the relevant service point. If all the servers have a common service rate μ, the queue corresponds to the single server queue M | M | 1, with a service rate of $k \mu$. A similar situation arises if the customers take a ticket on entering the bank.

Queues arise in many applications, such as

The arrival of aircraft in an airport
The arrival of ships in a port
Requests for data within computer memories (e.g.) client-server systems.

Motivated by commercial applications, networks of queues have been studied in great detail. Due to the classification of queuing problems , it has been possible to build sophisticated directories that cover a wide range of practical problems. Equally, queuing theory is the central concept in simulation models. We will briefly review some of the main ideas involved.

Structure of Asynchronnous Simulation Models

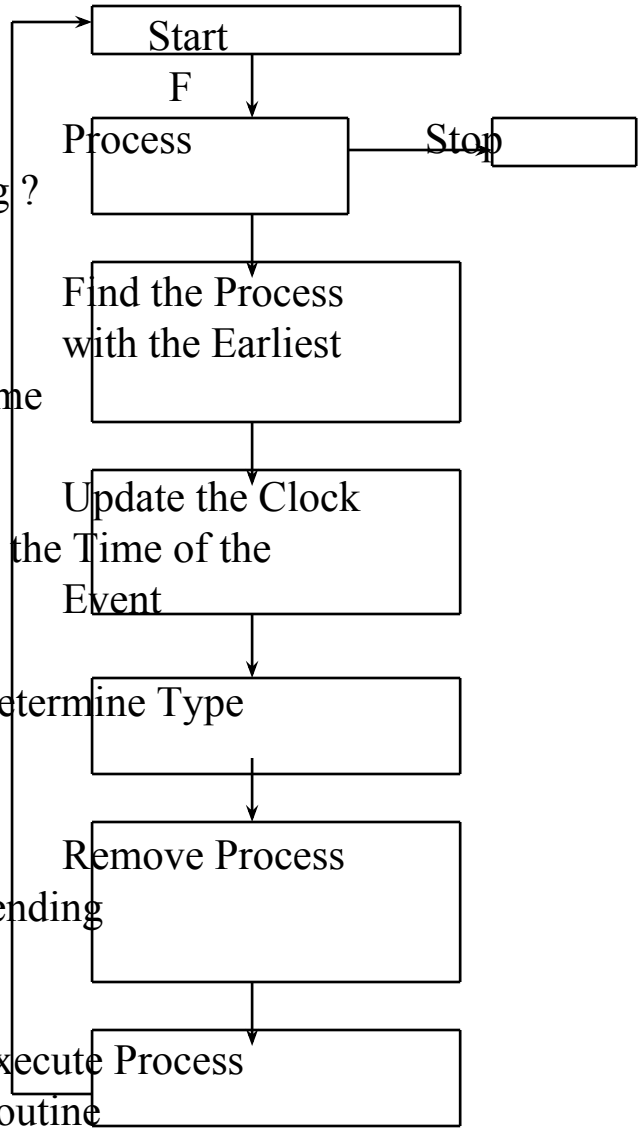
The simulation model **evolves** in a series of **stages**. The **value** of a model is known as its **state.** The **occurrence** of an **event** marks the start of a new stage and causes the model to change state.

It is only necessary to examine the system every time an event occurs. The **time between events** is controlled by a **clock**.

The primary **dynamic object** in a model is a **process**, which represents an **object** and the sequence of **actions** it experiences throughout its life in the model. An object comes into being at **creation time** and becomes active at **activation time**.

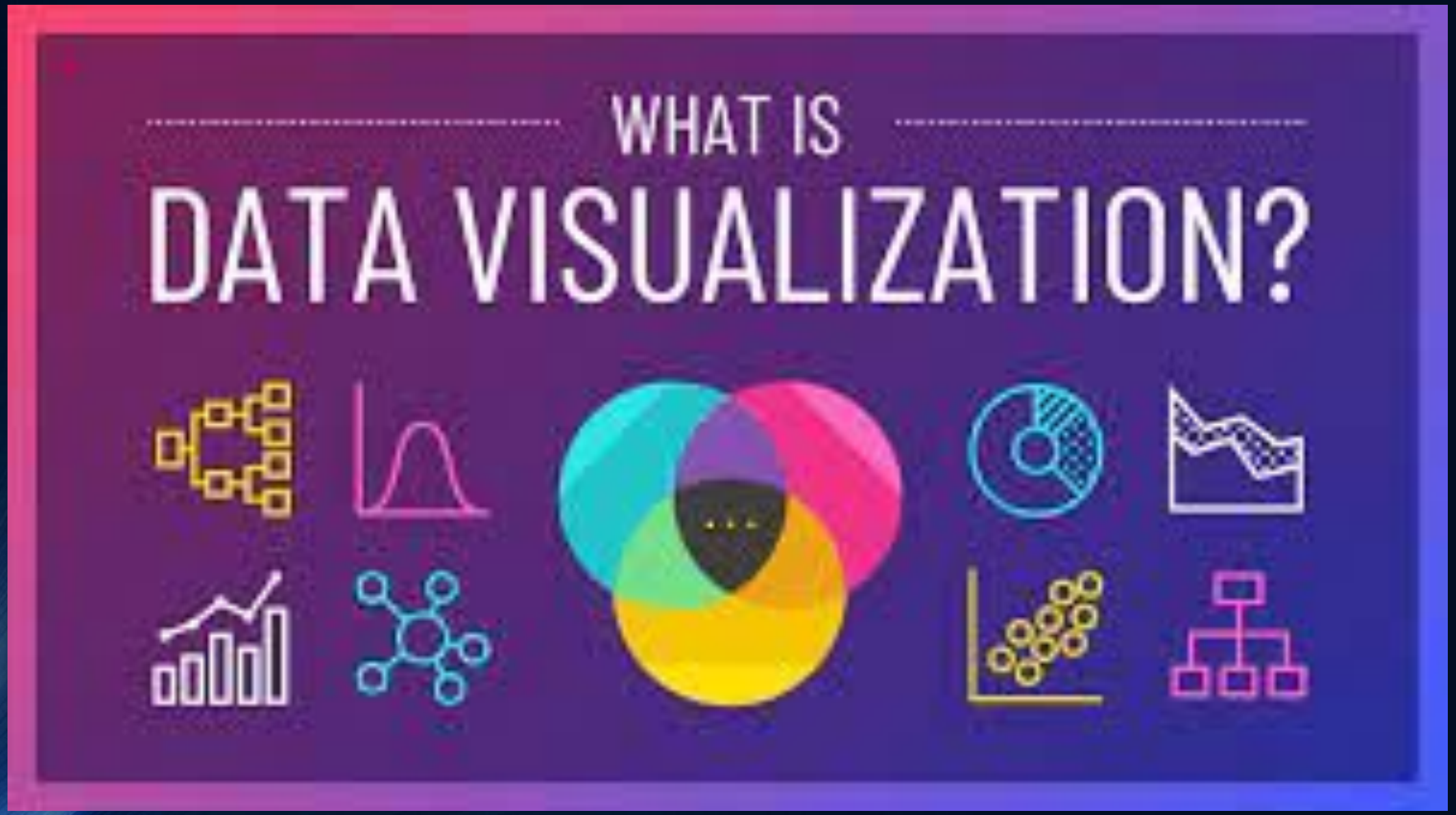
The primary **passive objects** are the **resources** which are shared by competing processes and lead to **internal queues** in the model.

The statistics gathered in simulation models fall into two main categories: **waiting times** are the **difference** or **tally** between service-end **times and arrival times** for customers**,** while **average numbers in the system** are **accumulated** via numerical integration.



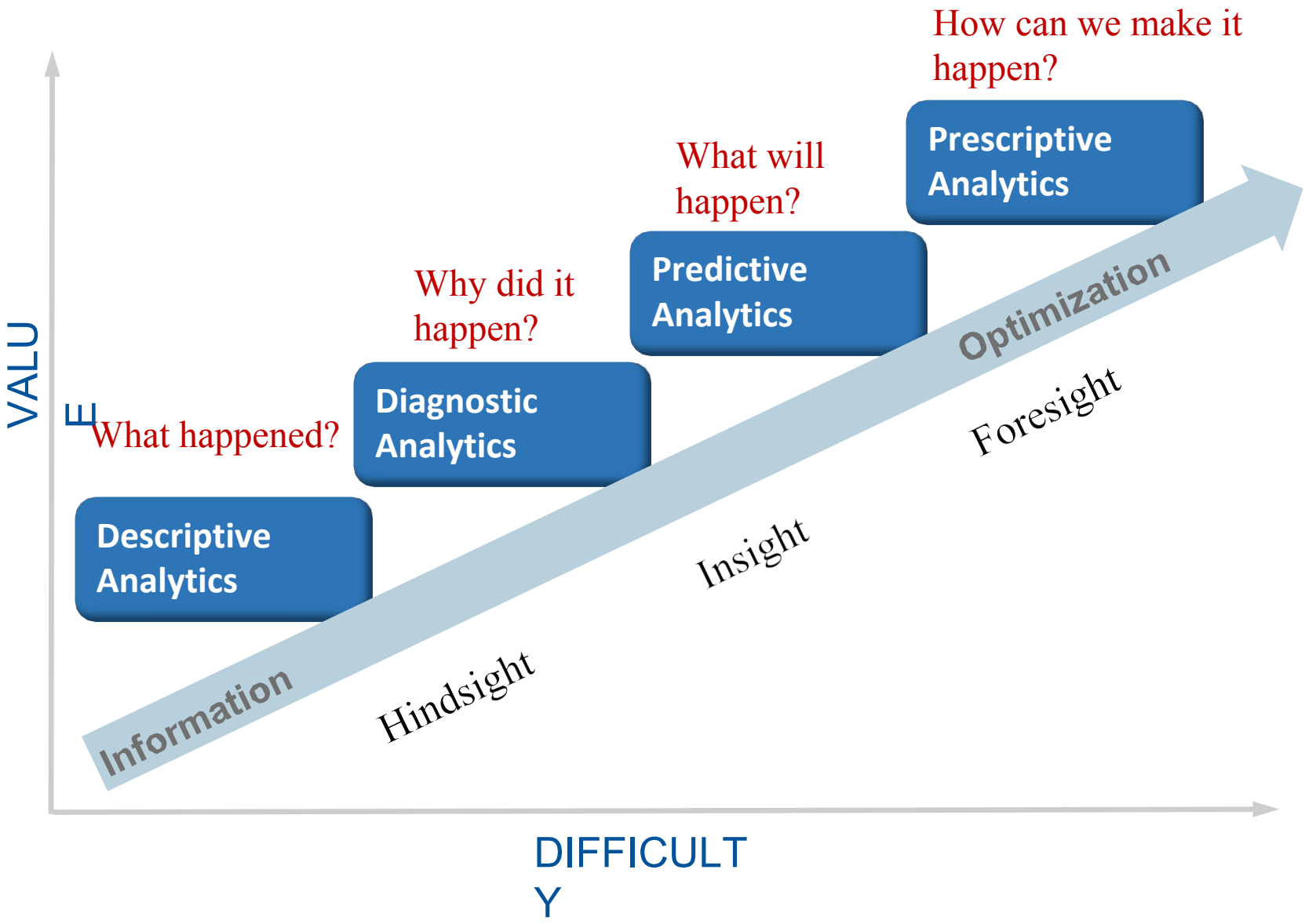
Course Contents

Unit III	Basics of Data Visualization	(07 Hours)
Computational Statistics and Data Visualization, Types of Data Visualization, Presentation and Exploratory Graphics, Graphics and Computing, Statistical Historiography, Scientific **Design Choices in Data Visualization,** Higher-dimensional Displays and Special Structures, **Static Graphics:** Complete Plots, Customization, Extensibility, **Other Issues**: 3-D Plots, Speed, Output Formats, Data Handling		


WHAT IS
DATA VISUALIZATION?

Data visualization:--- Ask students how many types

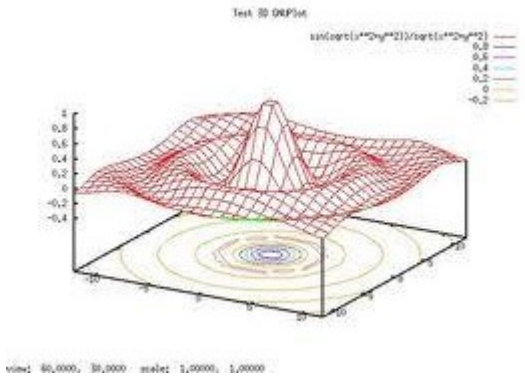
Analytics Models



Descriptive Analytics

- Descriptive analytics, such as reporting/ OLAP, dashboards, and data visualization, have been widely used for some time.
- They are the core of traditional BI.

Year 2000				
	Audio Division		Video Division	
Line Items	Budget	Actual	Budget	Actual
Cost of Goods Sold	$6,851,006.49	$7,132,961.38	$4,322,514.74	$4,526,954.71
Marketing Expense	$750,179.20	$756,596.17	$455,048.05	$462,815.40
Research and Development Expense	$538,243.39	$536,014.73	$329,890.95	$336,808.13
Selling Expense	$1,632,921.64	$1,579,790.18	$986,887.49	$927,970.90
Taxes	$314,659.05	$319,390.19	$202,636.67	$200,205.01
Year 2001				
	Audio Division		Video Division	
Line Items	Budget	Actual	Budget	Actual
Cost of Goods Sold	$2,564,556.31	$2,700,773.16	$1,726,031.16	$1,773,448.08
Marketing Expense	$294,766.22	$290,696.70	$167,757.29	$176,778.55
Research and Development Expense	$200,719.90	$193,236.83	$134,270.95	$125,725.88
Selling Expense	$620,427.30	$611,649.47	$405,092.93	$400,181.91
Taxes	$130,926.70	$122,526.31	$82,450.76	$80,671.87



What has occurred?

Descriptive analytics, such as data visualization, is important in helping users interpret the output from predictive and predictive analytics.

Predictive Analytics

- Algorithms for predictive analytics, such as regression analysis, machine learning, and neural networks, have also been around for some time.

What will occur?

- Marketing is the target for many predictive analytics applications.
- Descriptive analytics, such as data visualization, is important in helping users interpret the output from predictive and prescriptive analytics.

Prescriptive Analytics

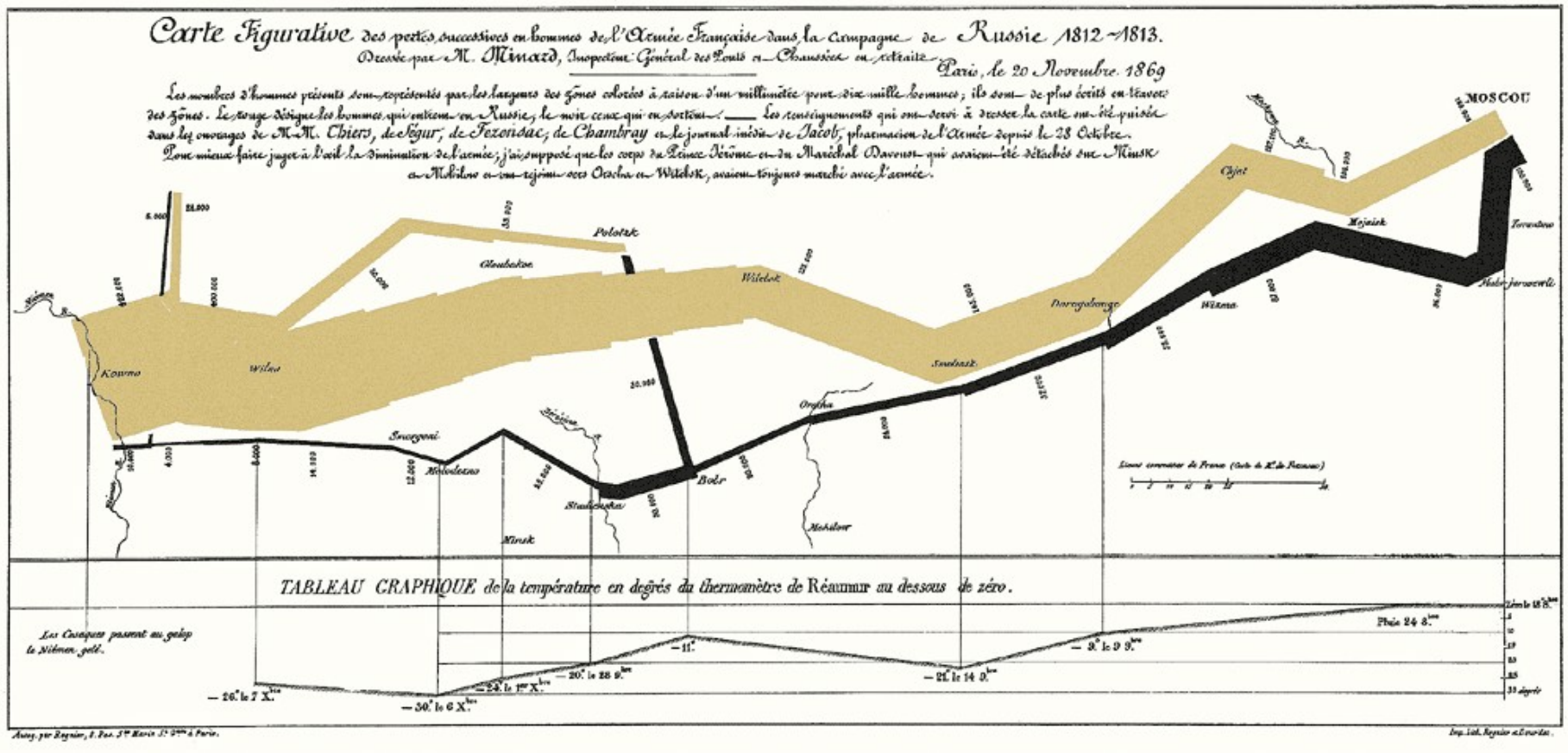
- Prescriptive analytics are often referred to as advanced analytics.
- Often for the allocation of scarce resources
- Optimization

What should occur?

Prescriptive analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demographic trends and population health trends, to more accurately plan for future capital investments such as new facilities and equipment utilization as well as understand the trade-offs between adding additional beds and expanding an existing facility versus building a new one.

A Brief History of Data Visualization

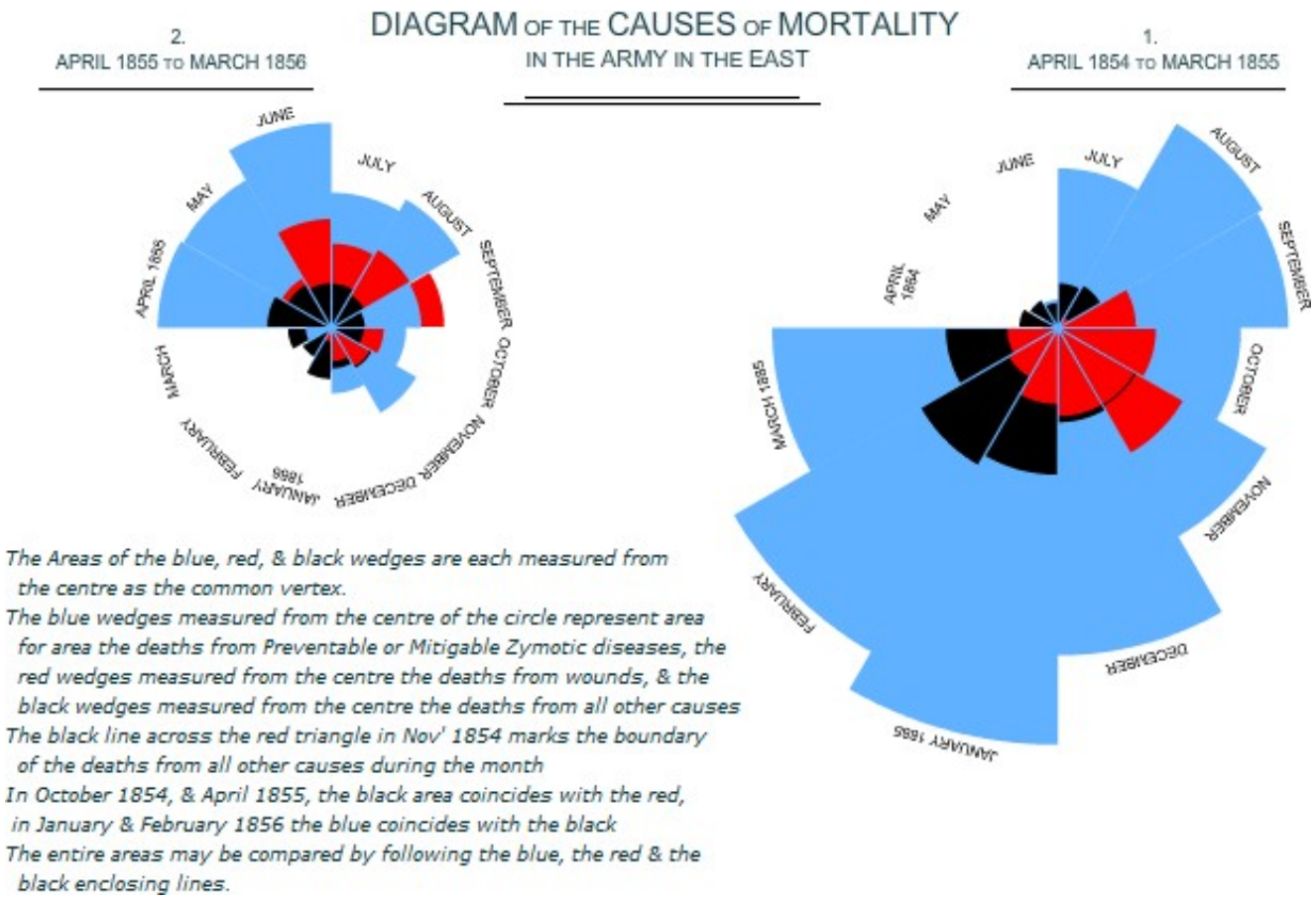
What makes a good chart?



Napoleon's 1812 March by Charles Joseph Minard

Reprinted in Tufte (2009), p. 41

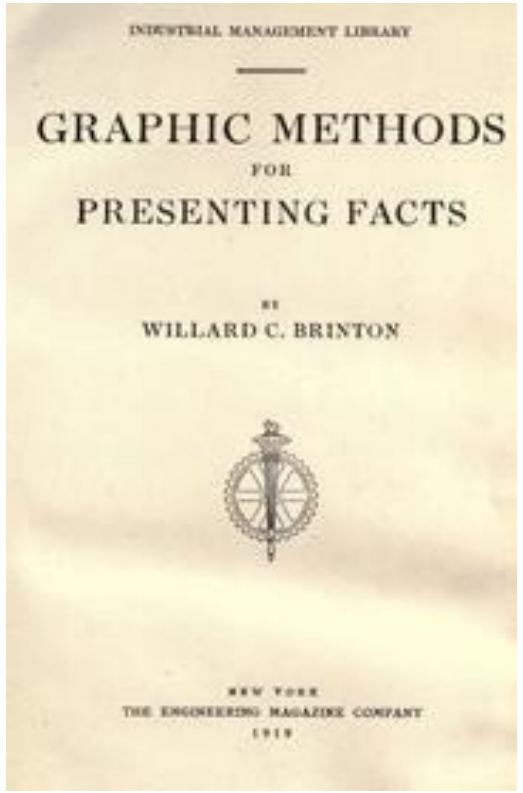
Florence Nightingale's 'Coxcombs‘1858



- Pioneer hospital sanitation
- Meticulously gathered data
- Pioneer in applied statistics and visualization
- Nurse

Willard C. Brinton, 1914

First business book about visualization



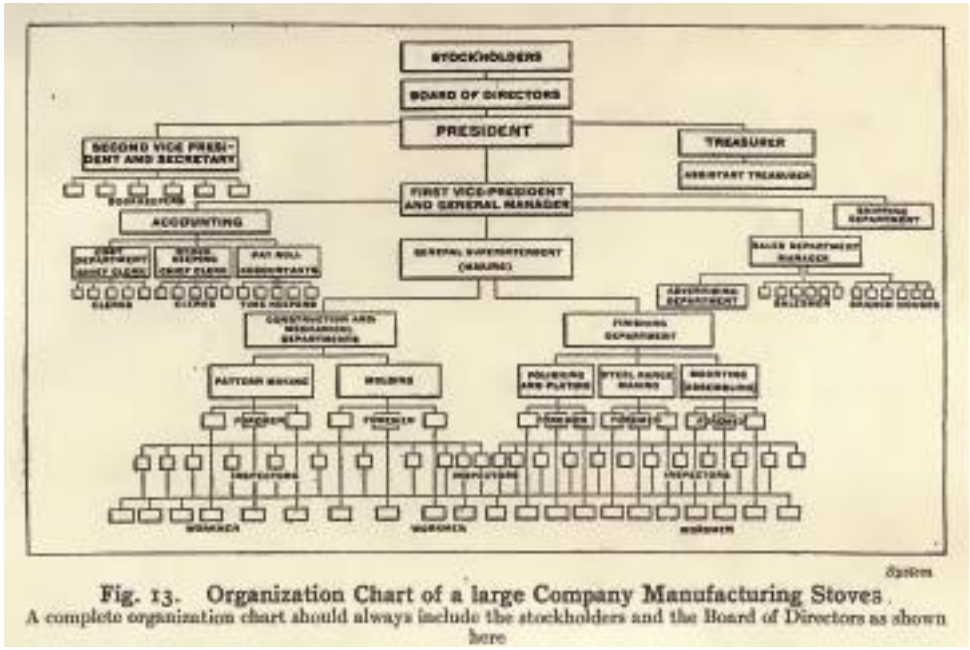
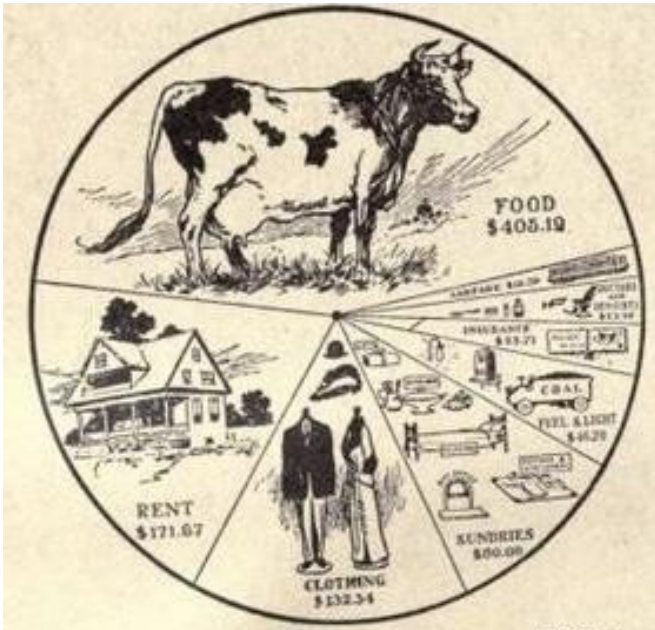


Fig. 13. Organization Chart of a large Company Manufacturing Stoves.
A complete organization chart should always include the stockholders and the Board of Directors as shown here



The Survey

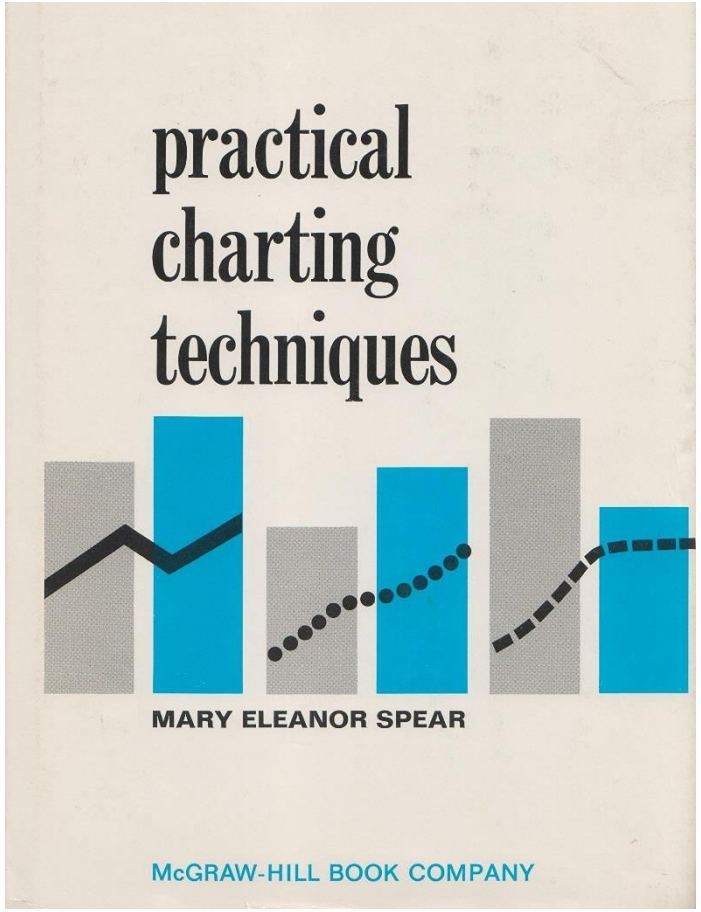
Fig. 3. Disposition of a Family Income of from $900 to $1000

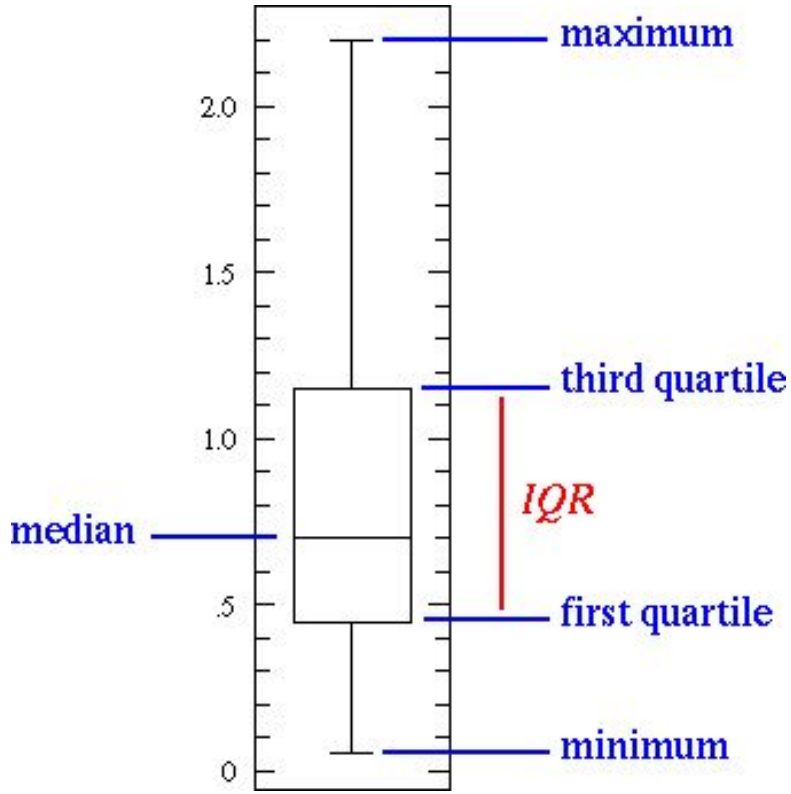
This cut shows an attempt to put figures in popular form. The eye is likely to judge by the size of the pictures rather than by the angles of the sectors

- Rules for presenting data
- American consulting engineer

Mary Eleanor Spear 1952, 1969

- Common-sense advice
- Invented box plot
- Worked for various US government agencies



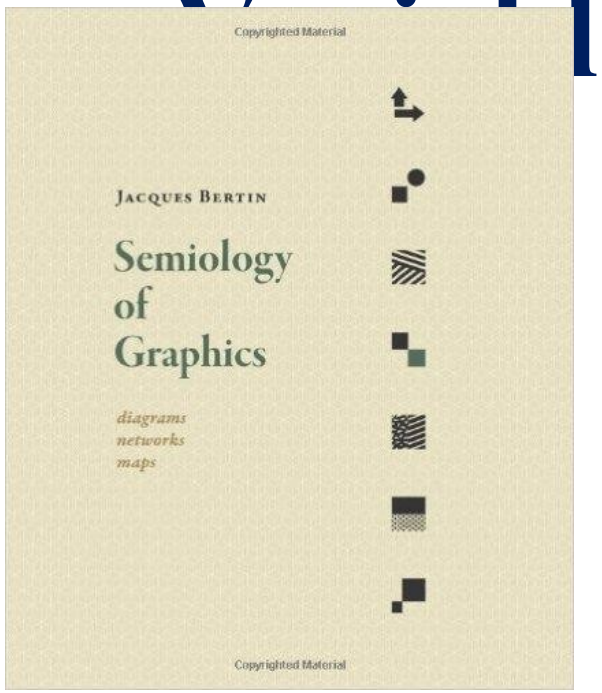


Jacques Bertin1967

- Principle of expressiveness:
  - Say everything you want to say — no more, no less
  - Don’t mislead
- Principle of effectiveness:
  - Use the best method available for showing your data
- Cartographer

Jacques Bertin

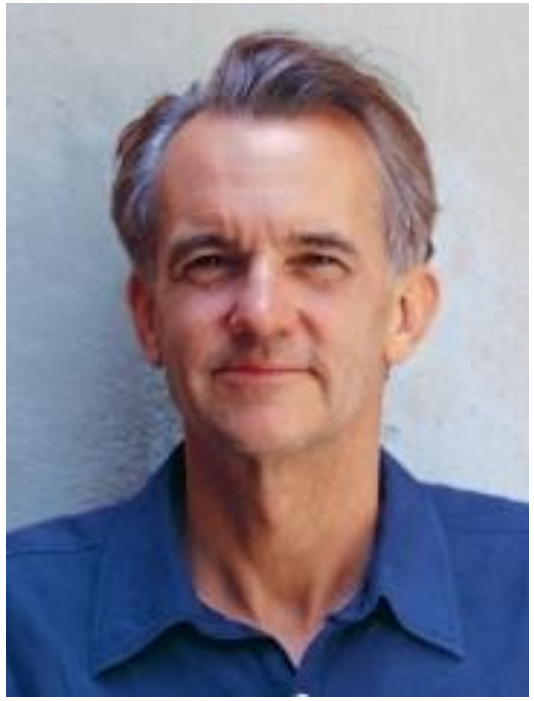
Seven Visual

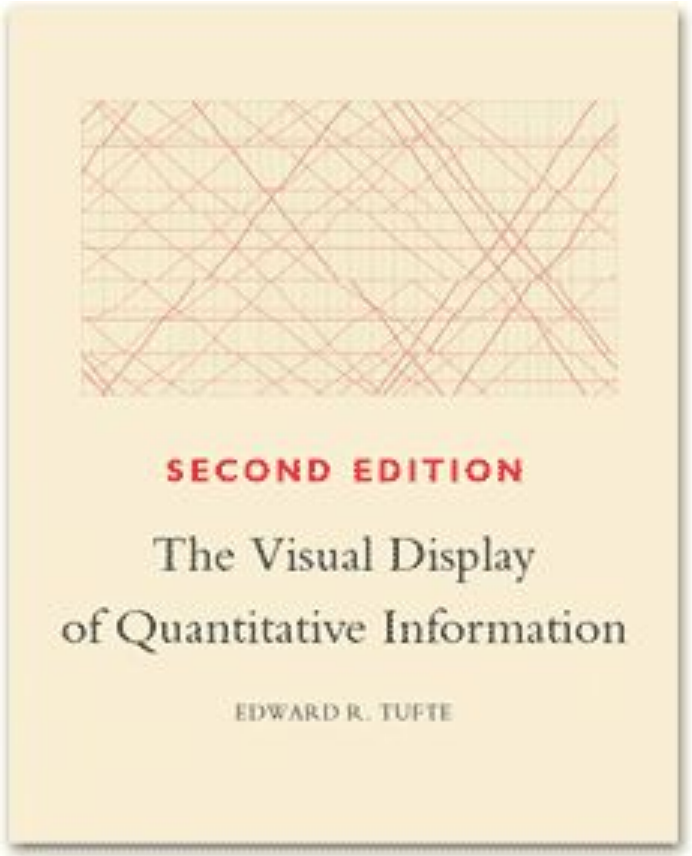


Bertin's Original Visual Variables	
Position changes in the x, y location	
Size change in length, area or repetition	
Shape infinite number of shapes	
Value changes from light to dark	
Colour changes in hue at a given value	
Orientation changes in alignment	
Texture variation in 'grain'	

- Position
- Size
- Shape
- Color
- Brightness
- Orientation
- Texture

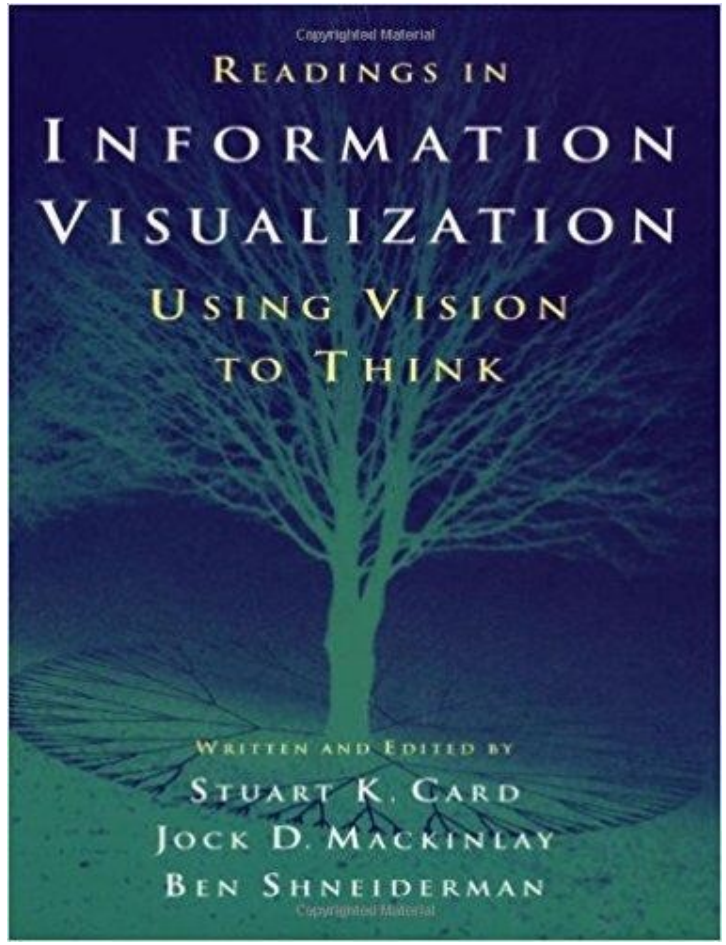
Edward Tufte 1983





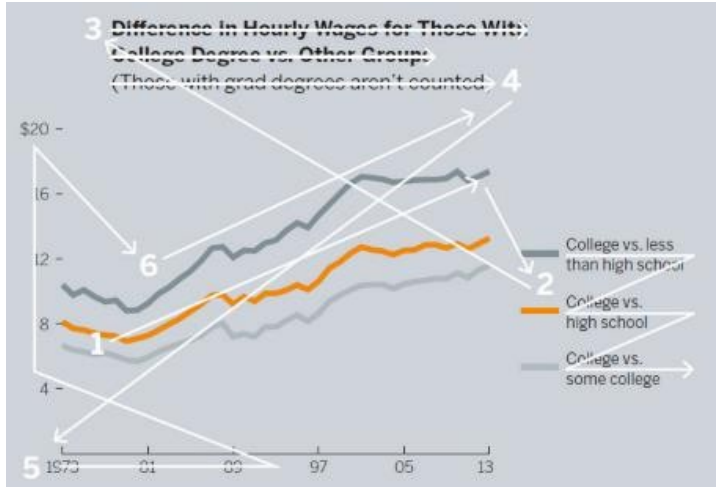
- Disciplined design principles
- Minimalist approach
- Professor emeritus at Yale University

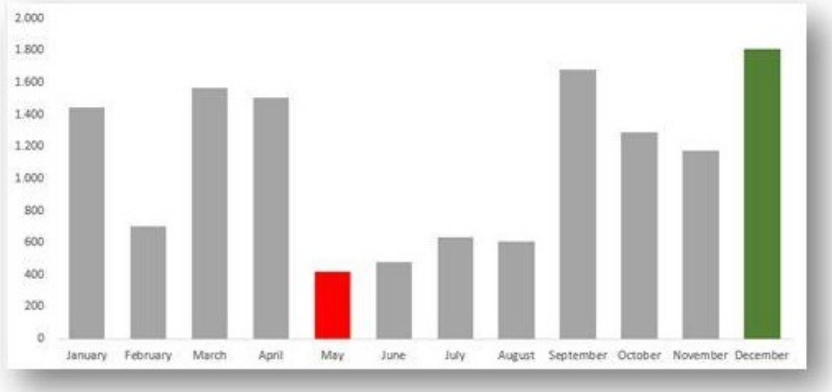
Jock Mackinlay 1986



- Automatically encode data with software
- Enable people to focus on ideas, concepts
- Added eighth variable to Bertin's list: motion
- VP of Research and Design, Tableau Software

When a Chart hits our Eyes





1. Visuals aren't read in a predictable, linear way
   - Create charts spatially, from the visual outward
2. We see first what stands out
   - Whatever stands out should support idea
3. We see only a few visuals at once
   - Plot as few visual elements as possible
4. We seek meaning and make connection
   - Relate visual elements in a meaningful way
5. We rely on conventions and metaphors
   - Embrace deeply ingraine d conventions

Example: USA Energy Resources

massima tra gli animali della categoria

Numero di specie scoperte

Percentuale di specie scoperte sul numero di specie stimate

Gli animali formano un **regno immenso**

Mammiferi

5.501

98%

Uccelli

2.7 m

10.064

96%

8.7 m

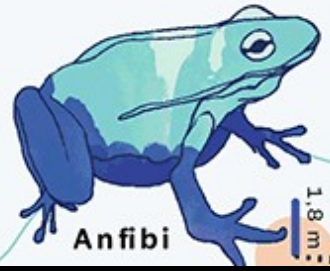
Rettili

9.547

80%

e composto da così tante **specie**,

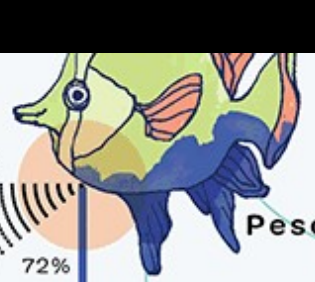
ognuna con le proprie caratteristiche,



Anfibi

1.8 m

6.771



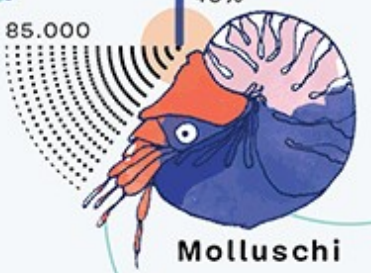
32.400

72%

Pesci

18 m

14 m



43%

85.000

Molluschi

Pare che esistano **20 miliardi** di miliardi di animali nel mondo!

che gli uomini non hanno ancora trovato il tempo di classificarli **tutti!**

Crostacei

4 m

47.000

31%

30 cm

102.246

17%

Aracnidi

Suddivisi **in ben 8,7 milioni** di specie diverse!

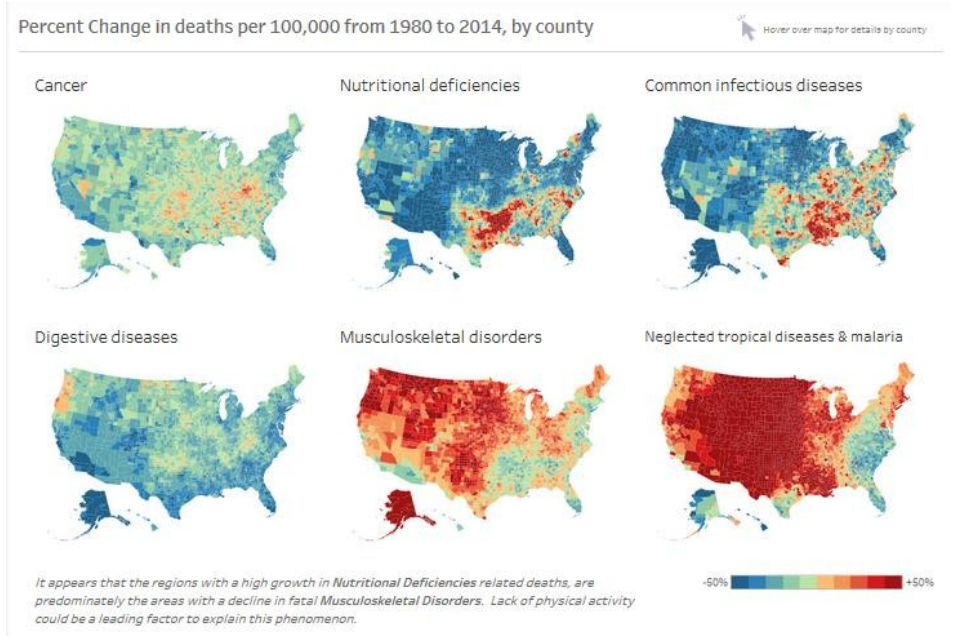
Insetti

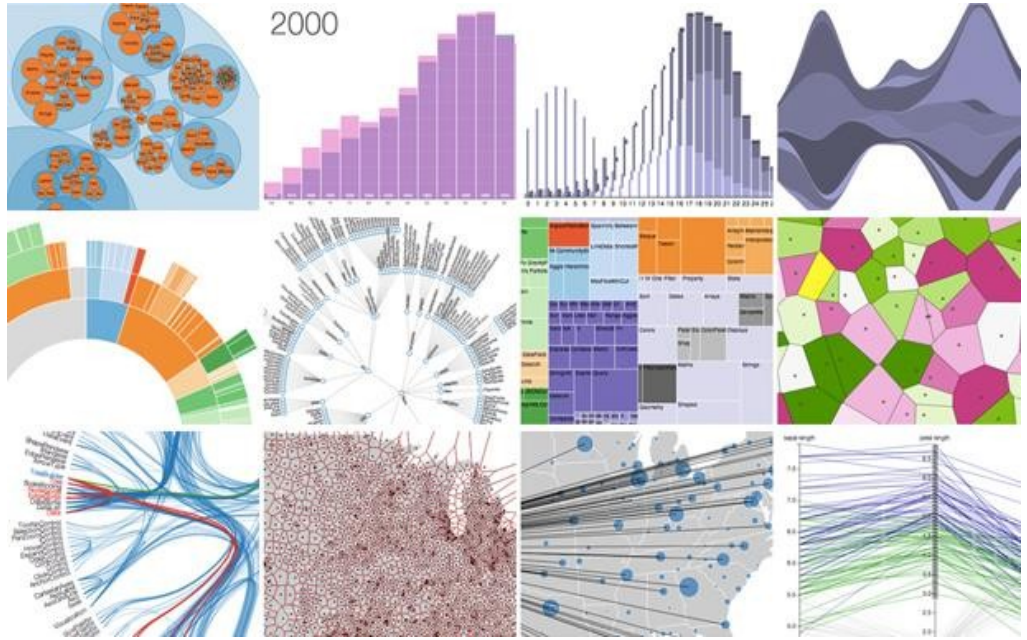
57 cm

20%

1.000.000

Vast Data Visualization Choice





What is data visualization and why is it important?

- **Data visualization** is the act of taking information (**data**) and placing it into a visual context, such as a map or graph. **Data visualizations** make big and small **data** easier for the human brain to understand, and **visualization** also makes it easier to detect patterns, trends, and outliers in groups of **data**.

Is data visualization a part of data science?

- **Data science** and **data visualization** are not two different entities. They are bound to each other. **Data visualization** is a **subset of data science**. **Data science** is not a single process or a method or any workflow.

What are the best data visualization software of 2019?

- WhSisense.
- Looker.
- Periscope Data.
- Zoho Analytics.
- Tableau.
- Domo.
- Microsoft Power BI.
- Qlikview.

What is data discovery and visualization?

- **Data discovery** is the process of breaking complex **data** collections into information that users can understand and manage. It turns incomprehensible mounds of raw **data** into groups, sets, and relationships, making order out of chaos.
- **Data discovery** answers the question, “What does it all mean?”
- Data visualization is its representation.

What are data visualization tools?

- By using visual elements like charts, graphs, and maps, **data visualization tools** provide an accessible way to see and understand trends, outliers, and patterns in **data**.

Is Excel a data visualization tool?

- **Excel** is a spreadsheet **tool**, while Tableau is a **data visualization** one.
- Spreadsheet **tools** are electronic worksheets that display **data** in a tabular format (a table of columns and rows).
- Each **data** point is stored in “cells” and can be manipulated by manually set formulas.

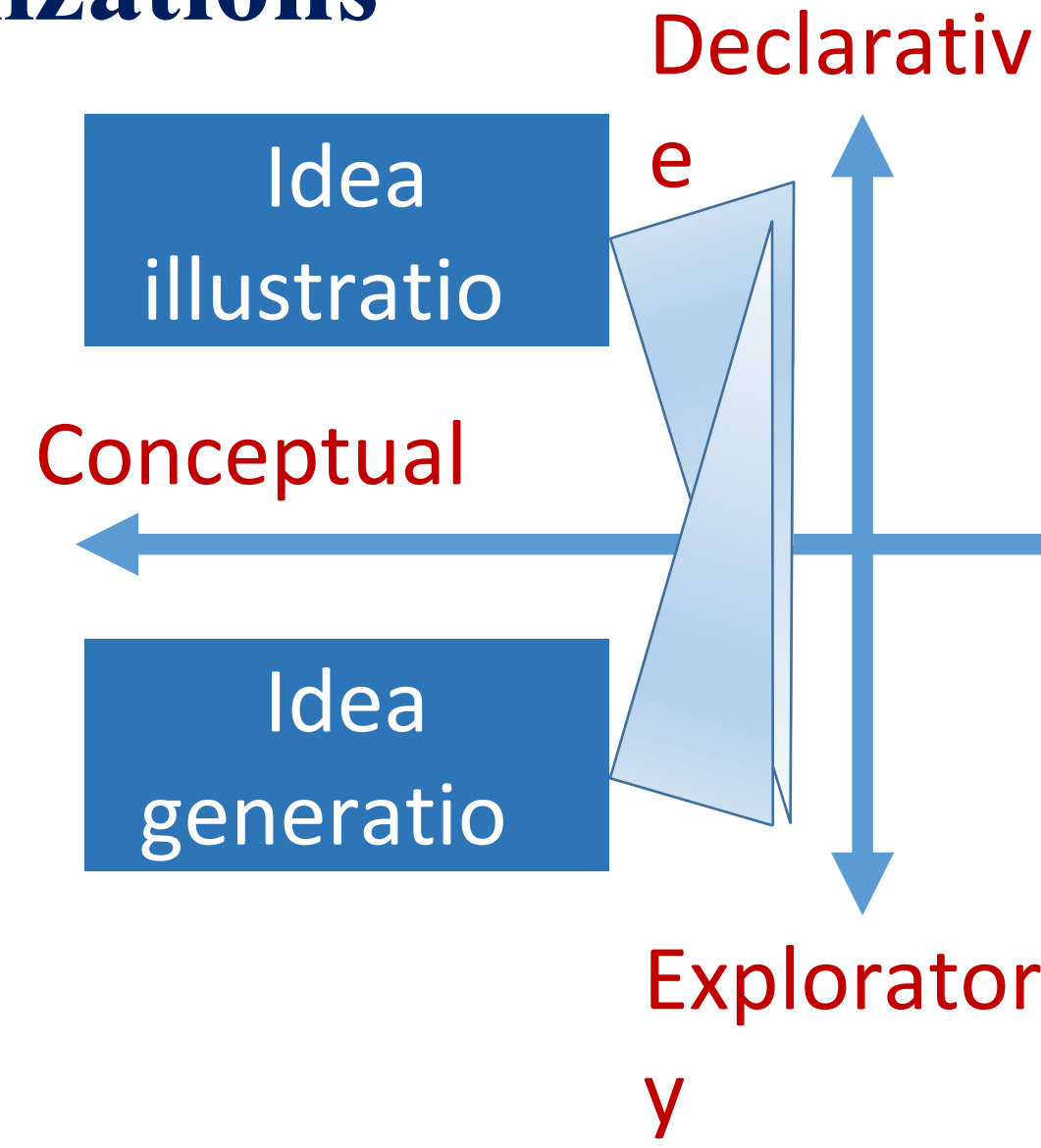
How do you create good data visualization?

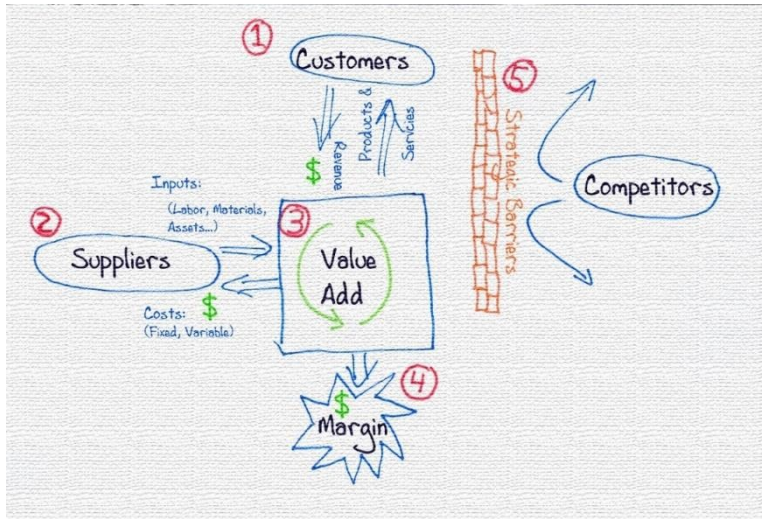
- Use it wisely in your data visualization design.
- Use a single color to represent the same type of data.
- Watch out for positive and negative numbers.
- Make sure there is sufficient contrast between colors.
- Avoid patterns.
- Select colors appropriately.
- Don't use more than 6 colors in a single layout.

What kind of visual communication do you want to create?

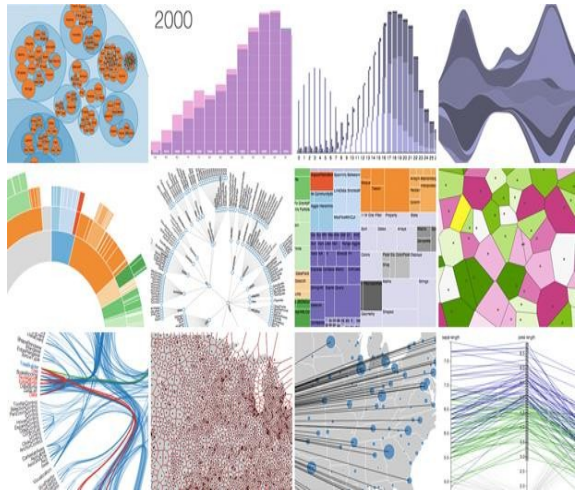
1. Is my information conceptual or data-driven?
   - Conceptual information is qualitative
   - Data-driven information is quantitative
2. Are my visuals meant to be declarative or exploratory?
   - A declarative purpose is to make a statement
   - An exploratory purpose is to look for new ideas

Four Types of Data Visualizations





Four Types of Data Visualizations



Declarativ

Idea illustratio

Everyda y

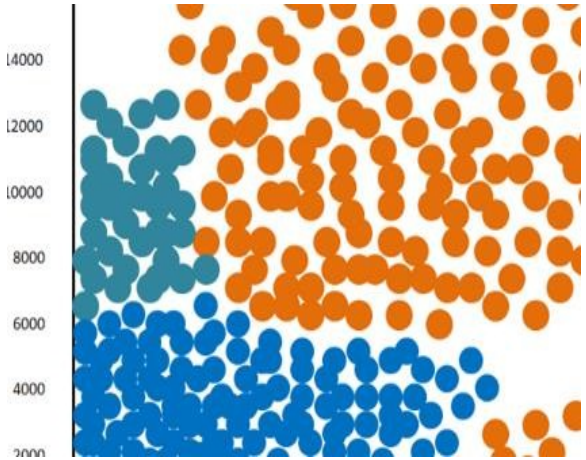
Conceptual

Data-Driven

Idea generatio

Visual discovery

Explorator y



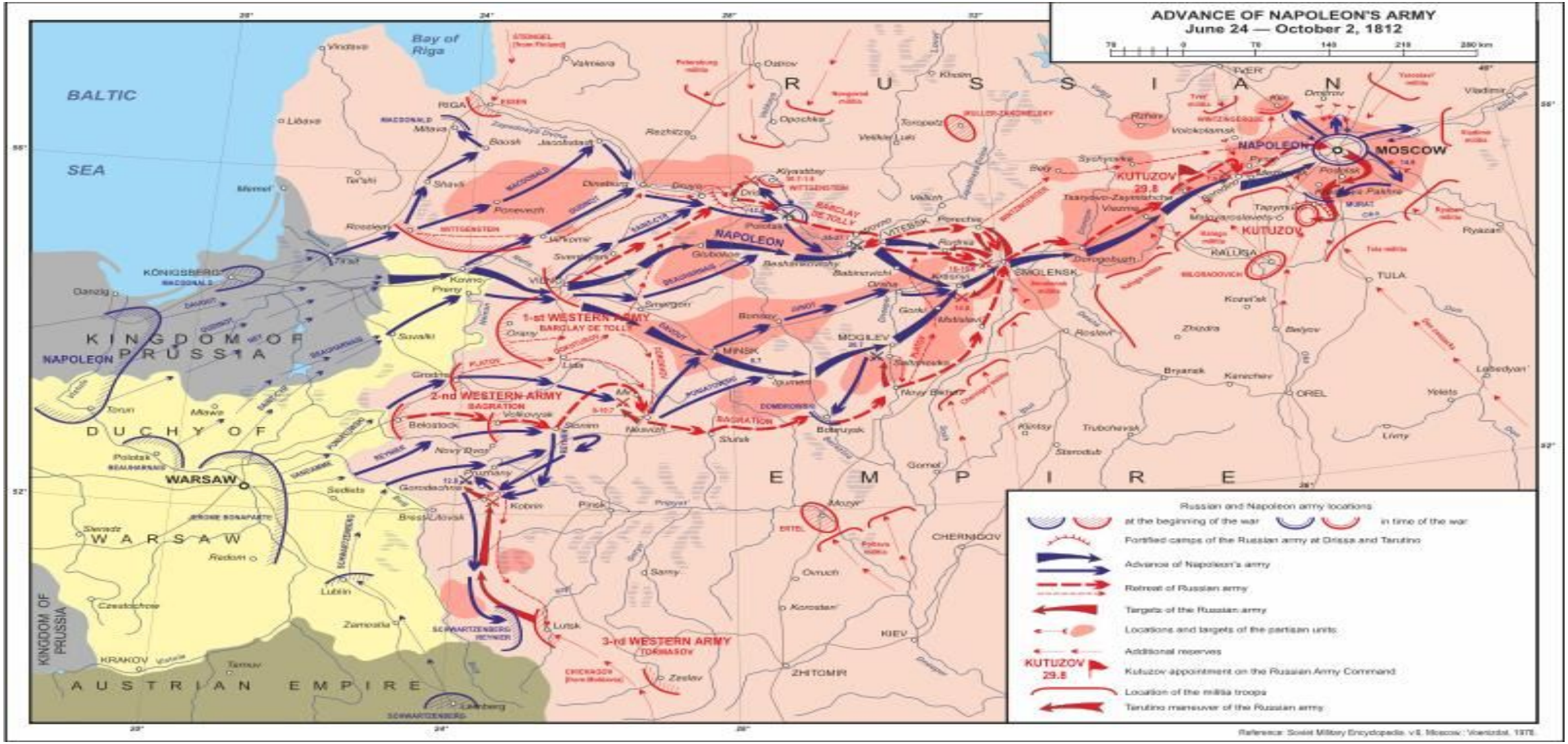
Data Visualization

provide clear understanding of patterns in data

detect hidden structures in data

condense information

What makes a good chart?

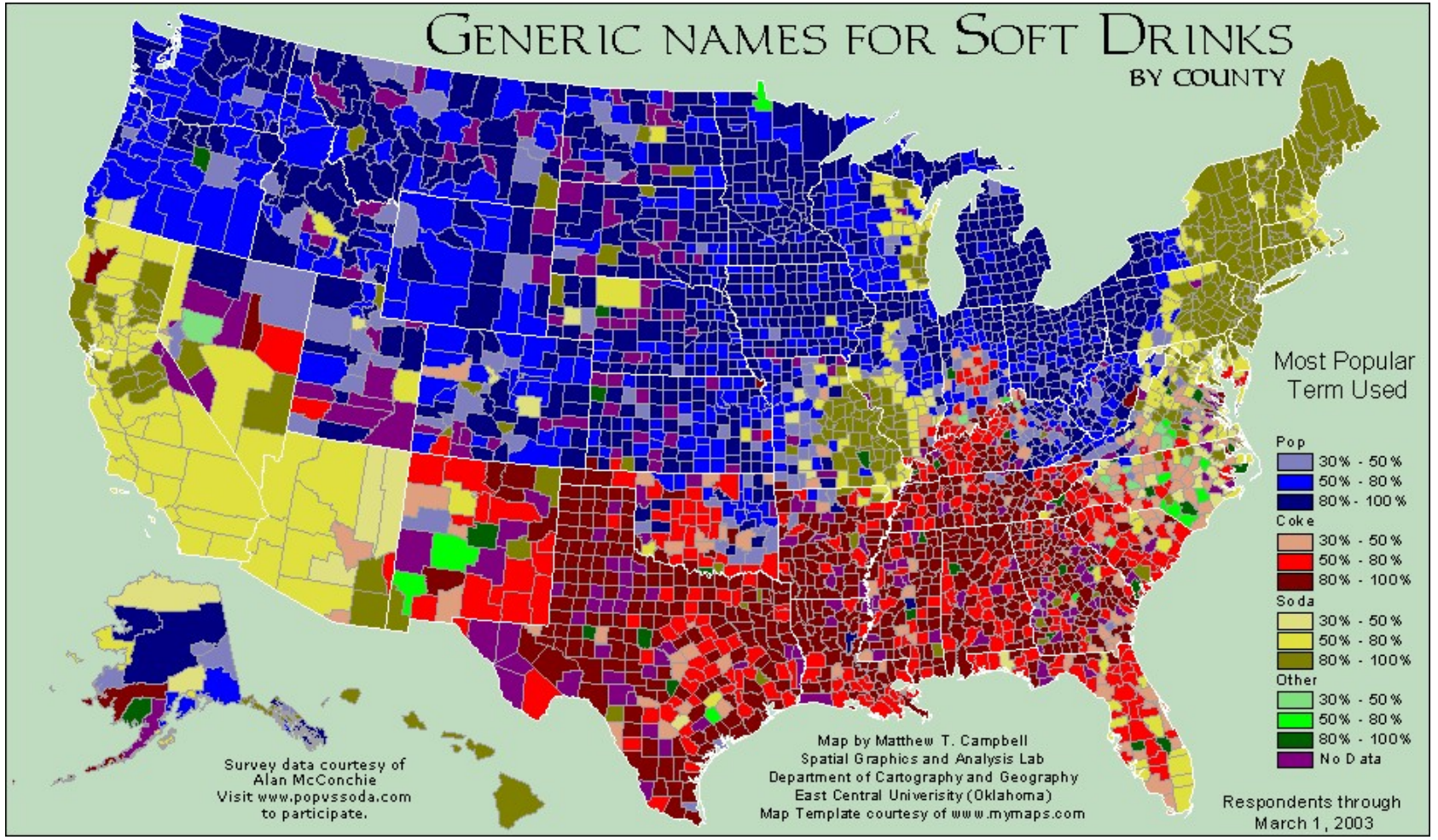


Wikipedia: Patriotic War of 1812

Video: Napoleonic Wars in 8 Minutes

Another video

What can you learn from this map?



http://www.popvssoda.com/countystats/total-county.html

Some basic principles (adapted from Tufte 2009)

1. The chart should tell a story
2. The chart should have graphical integrity
3. The chart should minimize graphical complexity

Tufte's fundamental principle: Above all else show the data

Principle 1: The chart should tell a story

Graphics should be clear on their own
The depictions should enable meaningful comparison

The chart should yield insight beyond the text

"If the statistics are boring, then you've got the wrong numbers." (Tufte 2009)

Principle 2: The chart should have graphical integrity

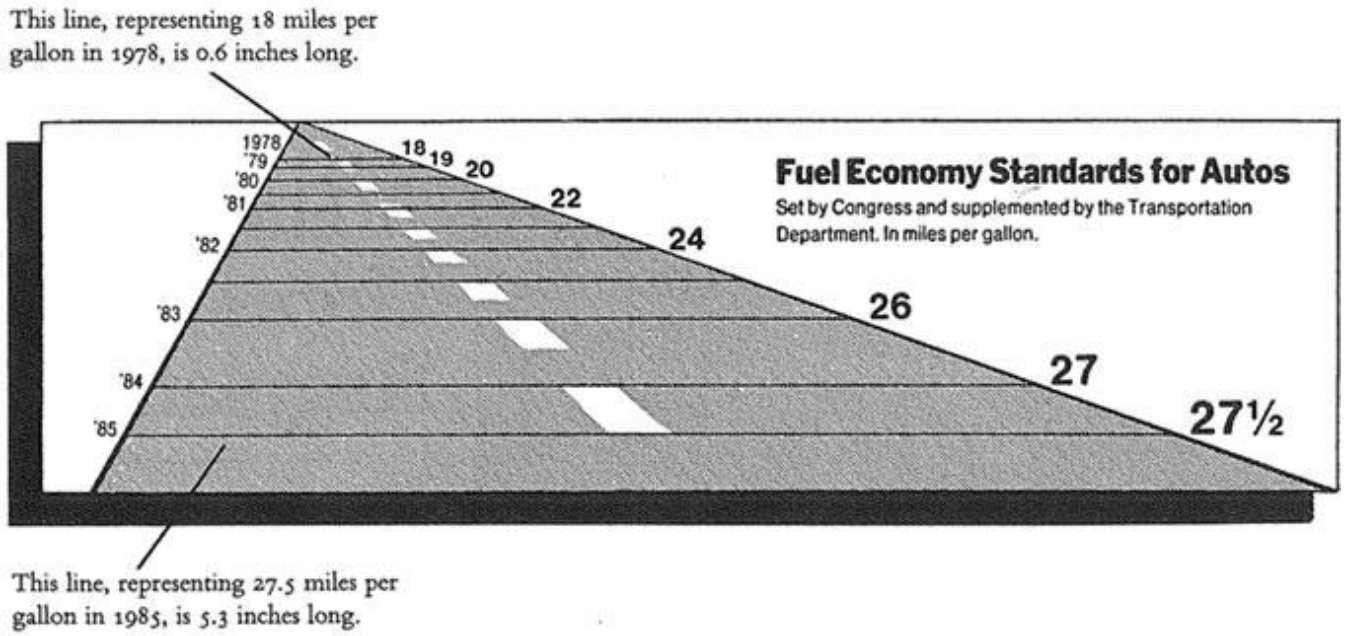
- Basically, it shouldn't "lie" (mislead the reader)
- Tufte's "Lie Factor":
  - $$Lie\ Factor = \frac{size\ of\ effect\ shown\ in\ graphic}{size\ of\ effect\ in\ data}$$

Should be ~ 1

> 1 = exaggerated effect

< 1 = understated effect

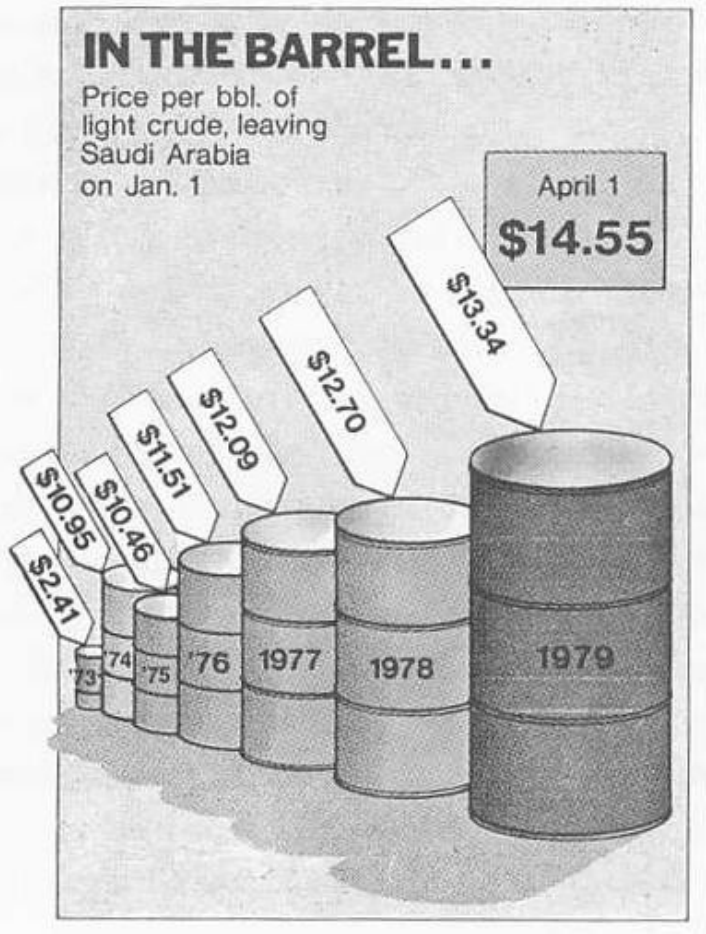
Examples of the "lie factor"



$$LF = \frac{5.3/0.6}{27.5/18} = \frac{8.83}{1.53} = 5.77$$

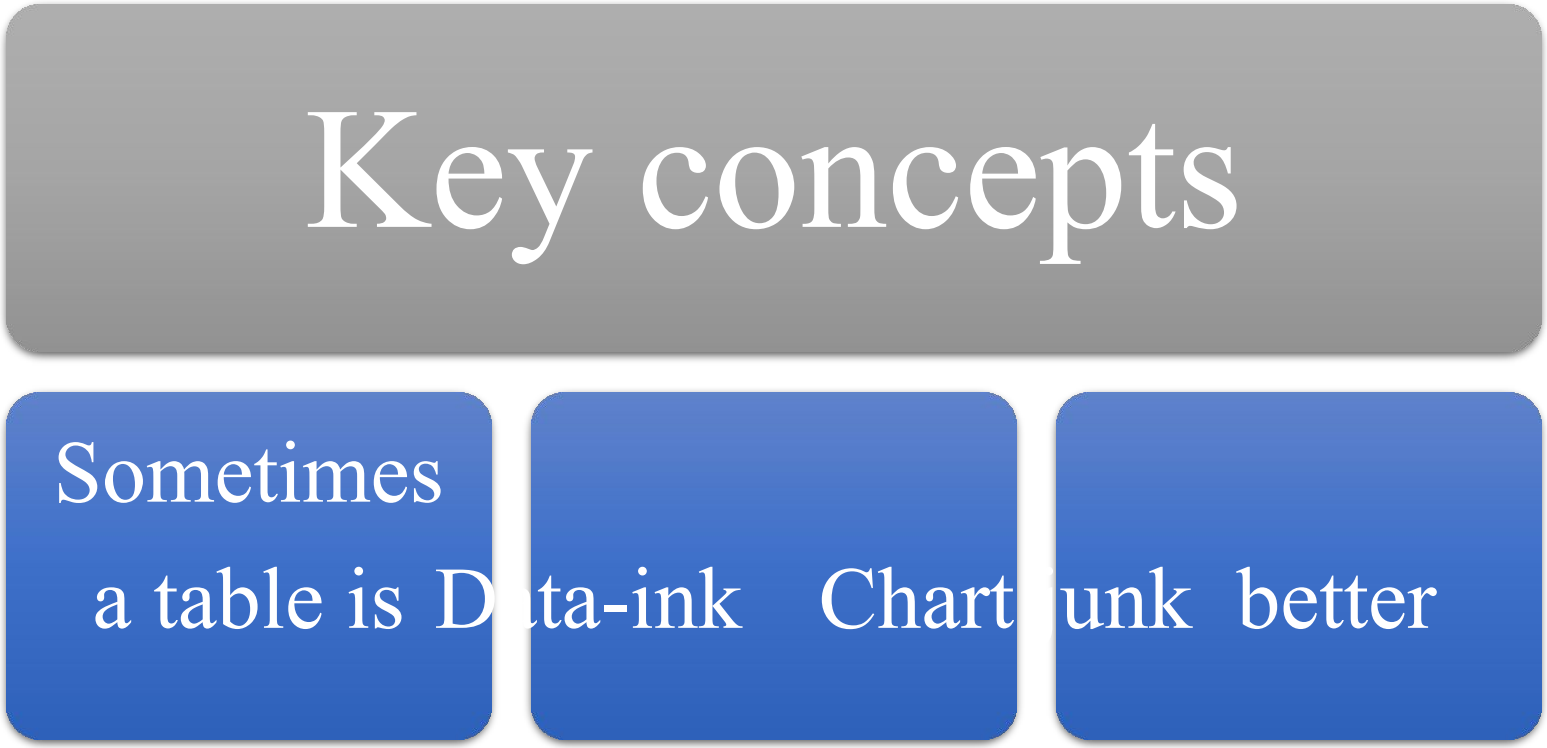
Reprinted from Tufte (2009), p. 57 & p. 62

$$LF = \frac{4280\%\ (change\ in\ volume)}{454\%\ (change\ in\ price)} = 9.4$$



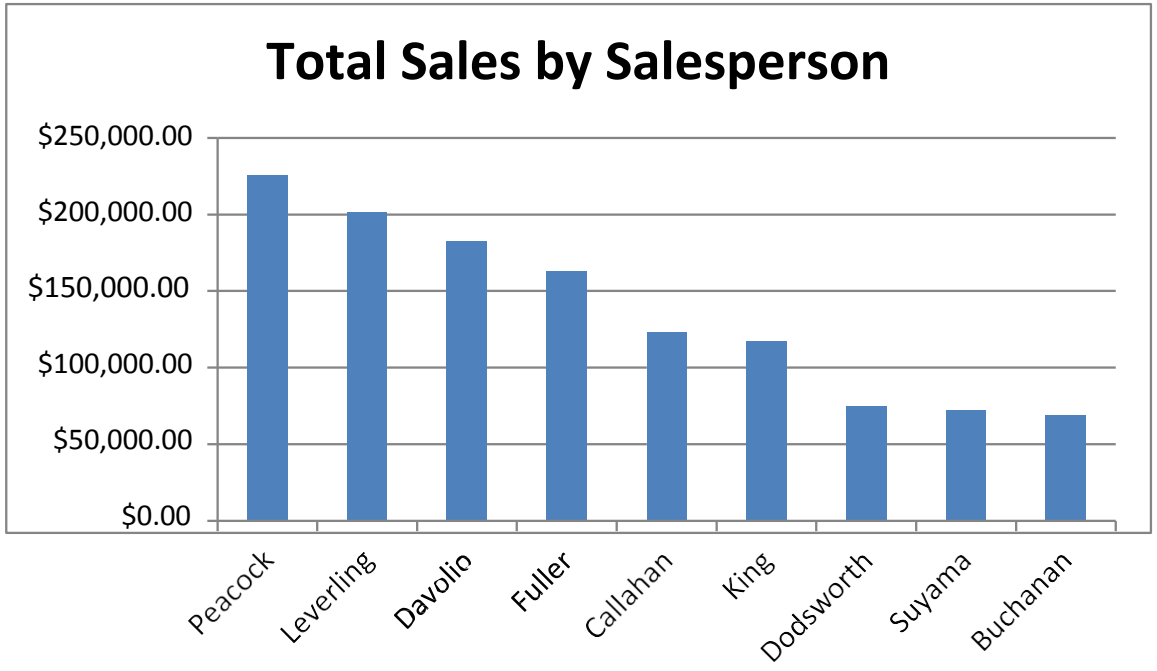
Principle 3: The chart should minimize graphical complexity

Generally, the simpler the better…



When a table is better than a chart

For a few data points, a table can do just as well…



Salesperson	Total Sales
Peacock	$225,763.68
Leverling	$201,196.27
Davolio	$182,500.09
Fuller	$162,503.78
Callahan	$123,032.67
King	$116,962.99
Dodsworth	$75,048.04
Suyama	$72,527.63
Buchanan	$68,792.25

The table carries more information in less space
and is more precise.

The Ultimate Table: The Box Score

- Large amount of information in a very small space

- So why does this work?
  - Depends on the reader's knowledge of the data

Philadelphia Phillies

Hitters	AB	R	H	RBI	BB	SO	#P	AVG	OBP	SLG
S Victorino CF	3	0	0	0	1	0	16	.000	.250	.000
P Polanco 3B	3	1	0	0	1	0	18	.000	.250	.000
J Rollins SS	4	2	2	0	0	0	14	.500	.500	.500
R Howard 1B	3	1	2	1	0	0	15	.667	.500	.667
R Ibanez LF	4	0	0	1	0	0	14	.000	.000	.000
B Francisco RF	3	1	1	1	1	0	17	.333	.500	.333
C Ruiz C	4	0	1	0	0	0	16	.250	.250	.250
W Valdez 2B	4	0	2	1	0	0	7	.500	.500	.750
R Halladay P	1	0	0	0	0	0	2	.000	.000	.000
a-P Orr PH	1	0	0	0	0	0	3	.000	.000	.000
J Romero P	0	0	0	0	0	0	0	.000	.000	.000
D Herndon P	0	0	0	0	0	0	0	.000	.000	.000
R Madson P	0	0	0	0	0	0	0	.000	.000	.000
b-R Gload PH	1	0	1	0	0	0	3	1.000	1.000	1.000
D Baez P	0	0	0	0	0	0	0	.000	.000	.000
c-J Mayberry Jr. PH	1	0	1	1	0	0	5	1.000	1.000	1.000
Totals	**32**	**5**	**10**	**5**	**3**	**0**	**130**			

a-lined out to first for R Halladay in the 6th
b-singled to left center for R Madson in the 8th
c-singled to deep center for D Baez in the 9th

Data Ink

- The amount of "ink" devoted to data in a chart
- Tufte's Data-Ink ratio:
  - $Data - ink\ ratio = \frac{data-ink}{total\ ink\ used\ in\ graphic}$

Should be ~ 1

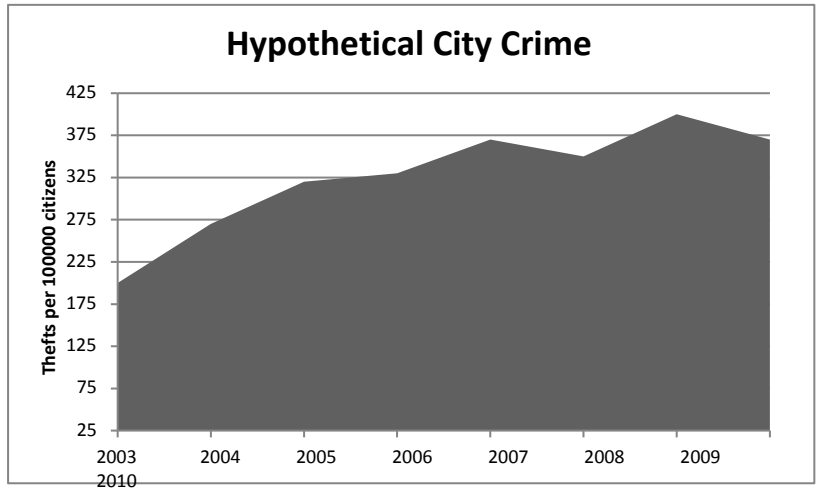
< 1 = more non-data related ink in graphic

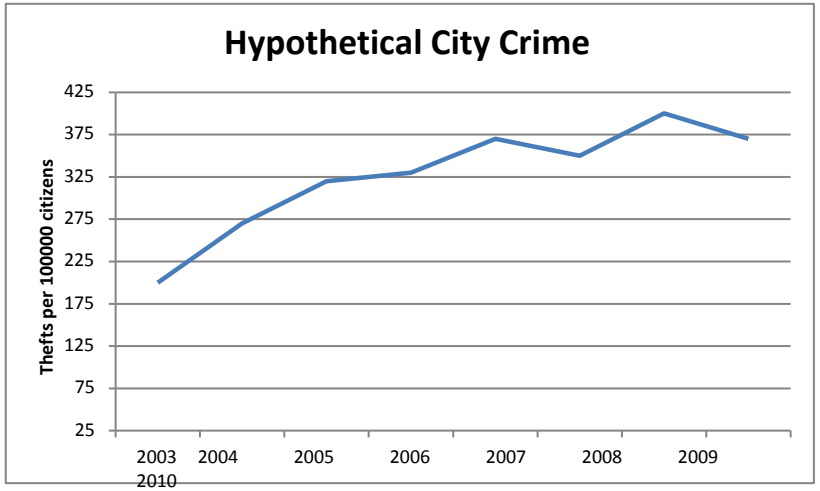
= 1 implies all ink devoted to data

Tufte's principle:
Erase ink whenever possible

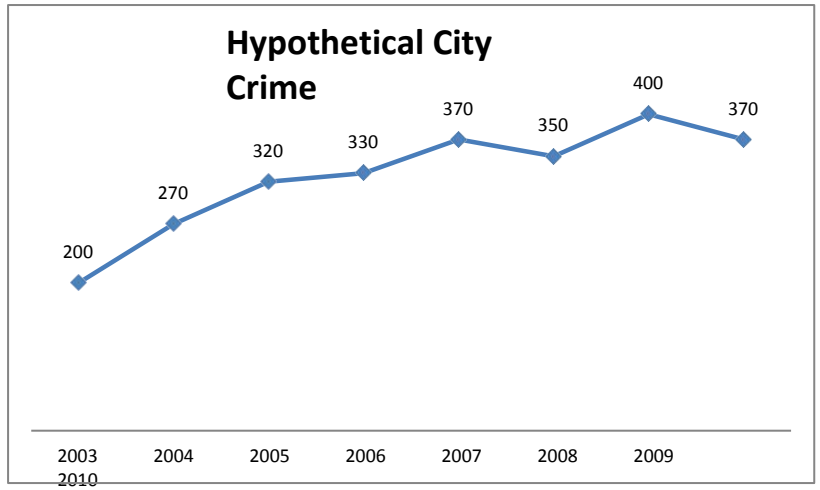
Being conscious of data ink

Lower data-ink ratio
(worse)

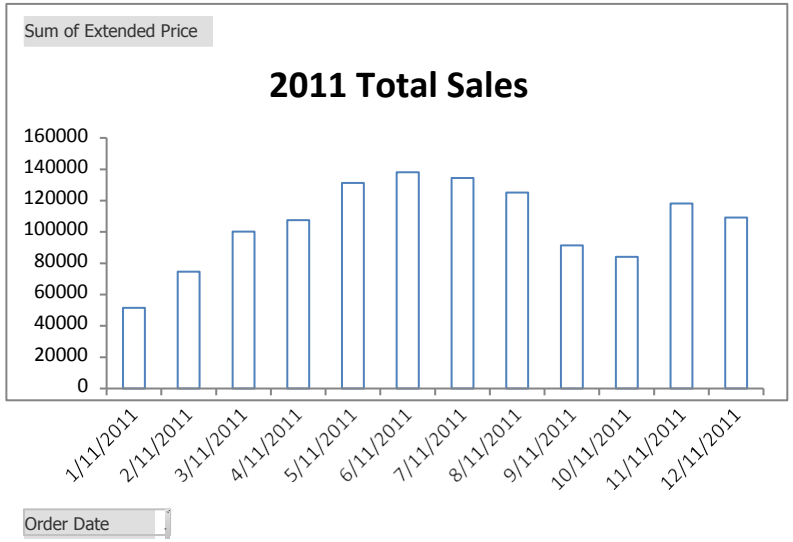




Higher data-ink ratio
(better)



What makes a good chart?



Sometimes it's really a matter of preference.

These both minimize data ink.

Why isn't a table better here?

3-D Charts

Evaluate this from a data-ink perspective. How does it affect the clarity of the

One of the golden rules of data visualization is…..

Never use 3D!

Data Integrity/ Lie Factor

- 3D skews numbers, making them difficult to interpret or compare

Graphical Complexity

- Adding 3D to graphs introduces unnecessary chart elements like side and floor panels

Source: Knaflic (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. Chapter

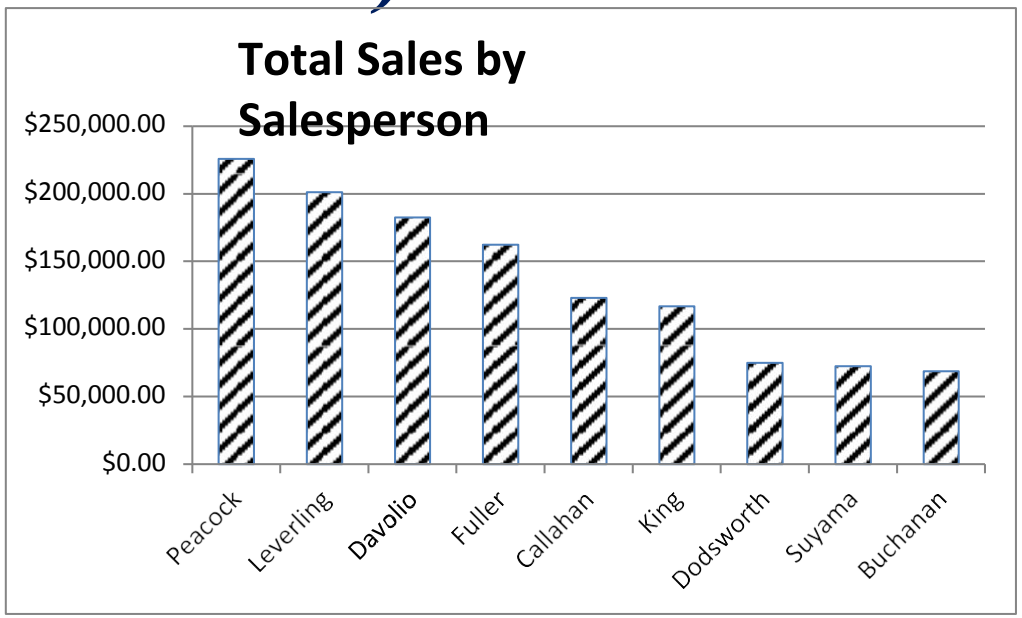
Chartjunk: Data Ink “gone wild”

Unnecessary visual clutter that doesn’t provide additional insight

Distraction from the story the chart is supposed to convey

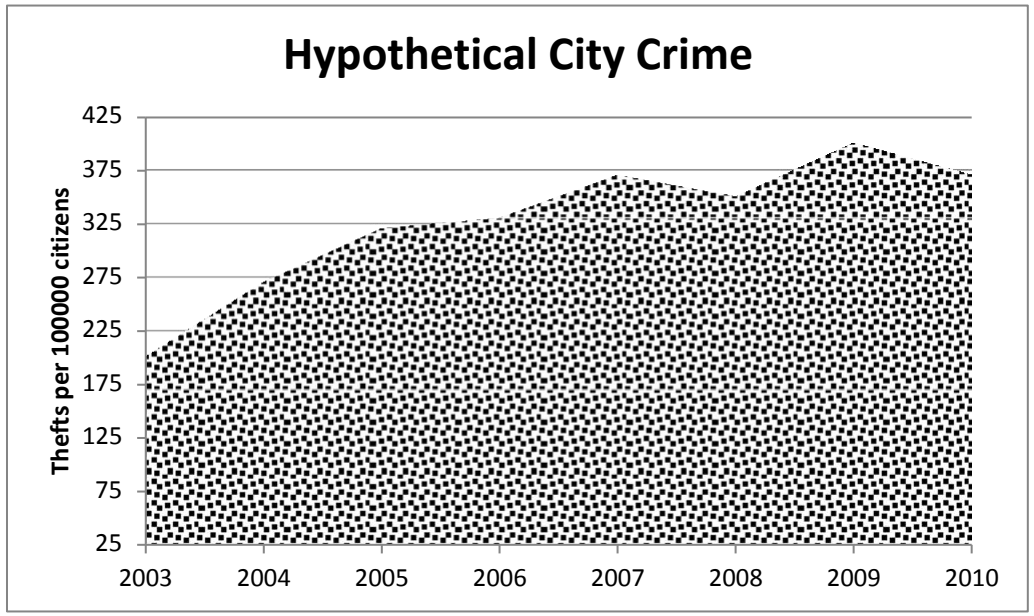
When the data-ink ratio is low, chartjunk

Example: Moiré effects (Tufte 2009)

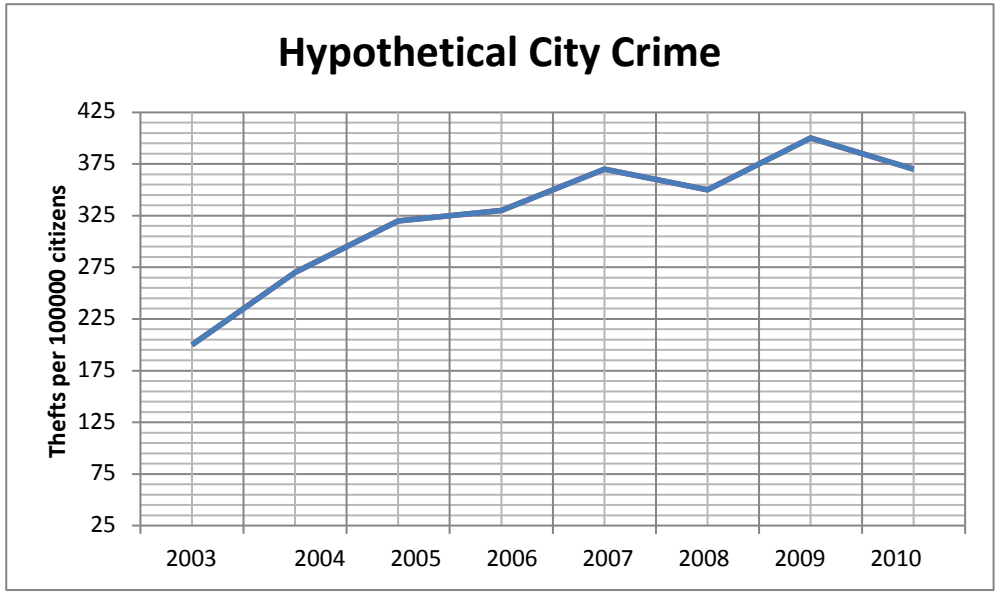


Creates illusion of movement

Stands out, in a bad way

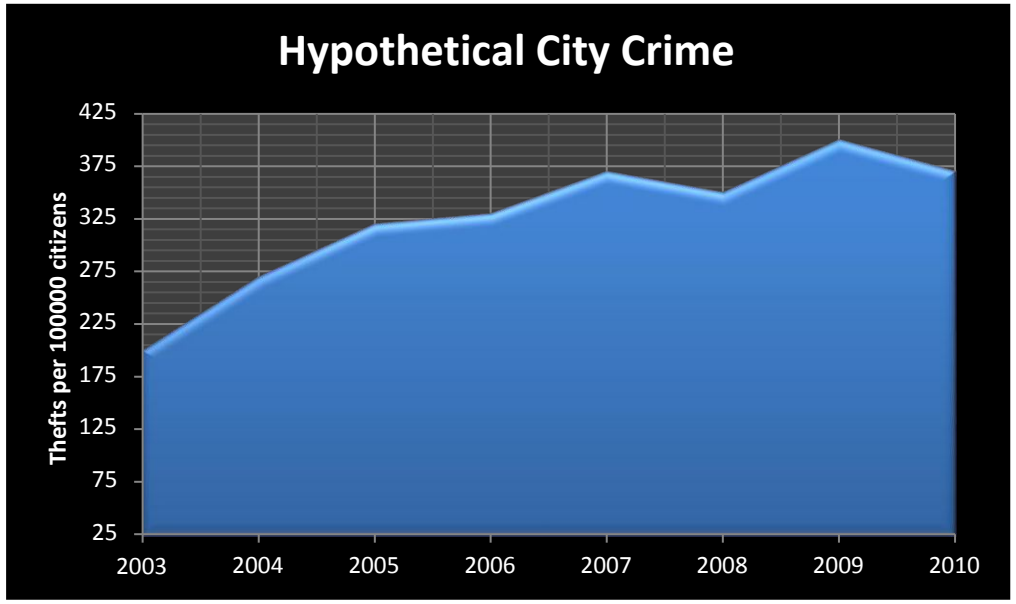


Example: The Grid

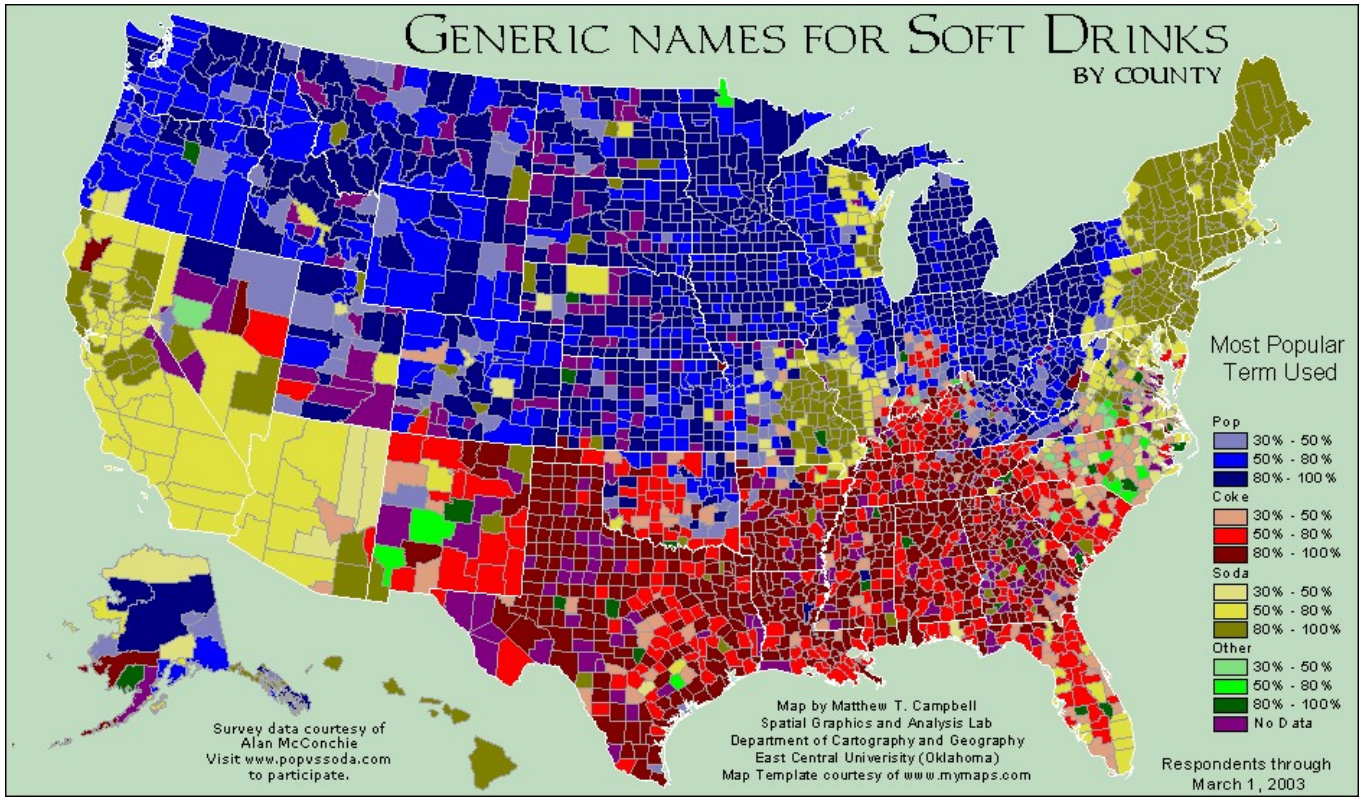


Why are these examples of chartjunk?

What could you do to remedy it?



Data Ink Working For Us



Evaluate this chart in terms of Data Ink.

Imagine this as a bar chart. As a table!!

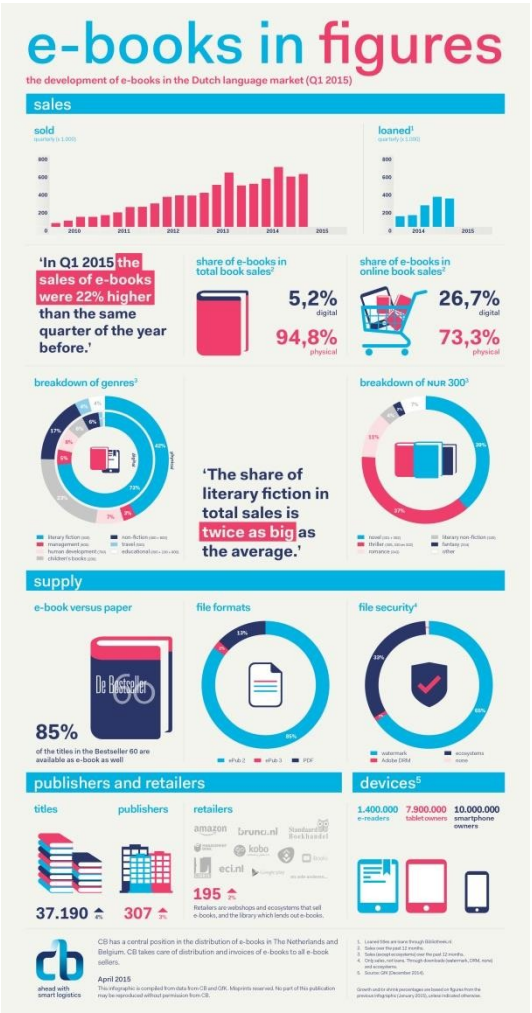
Review: Data principles (adapted from Tufte 2009)

1. The chart should tell a story
2. The chart should have graphical integrity
3. The chart should minimize graphical complexity

Tufte's fundamental principle: Above all else show the data

Infographics

- Information graphics
- Visualization of information, data or knowledge intended to present information quickly and clearly
- We will have an ICA to create infographics using *Piktochart*.



http://the-digital-reader.com/2015/04/13/infographic-ebooks-on-track-to-double-dutch-ebook-market-in-2014/

Summary

- Use data visualization principles to assess a visualization
  - Tell a story
  - Graphical integrity (lie factor)
  - Minimize graphical complexity (data ink, chartjunk)
- Explain how a visualization can be improved based on those principles
- Types of visualization

Resources…

- DataMed https://datamed.org/
- Institute for Health Metrics and Evaluation's Global Health Data Exchange http://ghdx.healthdata.org/
- NNLM RD3: Resources for Data-Driven Discovery https://nnlm.gov/data/
- NNLM's YouTube Channel https://www.youtube.com/channel/UCmZqoegBFKJQF69V8d-05Bw
- OHSU's Big Data to Knowledge https://dmice.ohsu.edu/bd2k/topics.html
- Registry of Research Data Repositories (re3data.org) http://www.re3data.org/

References

- Borgman, Christine L. Big data, little data, no data: Scholarship in the networked world. MIT Press, 2015.
- Federer, Lisa. Beyond the SEA: Data Science 101: An introduction for librarians https://www.youtube.com/watch?v=i78ciP1eGxo&t=3s
- Mayer-Schönberger, Viktor, and Kenneth Cukier. Big data: A revolution that will transform how we live, work and think. Houghton Mifflin Harcourt, 2013.

Bibliography…

- A Good Example of Misleading Visualization
  - http://spatial.ly/2009/09/a-good-example-of-misleading-visualization/
- A quick guide for better data visualizations
  - https://www.tableau.com/good-to-great
- The analysis of visual variables for use in the cartographic design of point symbols for mobile Augmented Reality applications
  - Łukasz Halik, Adam Mickiewicz University Poznan
  - http://www.iag-aig.org/attach/30dee1f85f7bd479367f1f933d48b701/V61N1_2FT.pdf
- The Benefits and Future of Data Visualization
  - StatSilk founder Frank van Cappelle
  - https://www.statsilk.com/blog/benefits-and-future-data-visualization
- Charting Statistics
  - Mary Eleanor Spear
  - https://archive.org/details/ChartingStatistics
- Color Brewer
  - http://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3

Bibliography…

- Data: The language of modern business leaders
  - Steve Proctor, March 17, 2017
  - http://www.itbusiness.ca/sponsored/data-the-language-of-modern-business-leaders
- Data Visualization: The Best Infographic Tools For 2017
  - Bernard Marr, October 10, 2017
  - https://www.huffingtonpost.com/entry/data-visualization-the-best-infographic-tools-for_us_59ca128fe4b0f2df5e83b134
- Data Visualization: The future of data visualization
  - Will Towler, January/February 2015
  - http://analytics-magazine.org/data-visualization-the-future-of-data-visualization
- Data Visualization 101: How to Choose the Right Chart or Graph for Your Data
  - Jami Oetting
  - https://blog.hubspot.com/marketing/types-of-graphs-for-data-visualization
- Data-Driven Design: Dare to Wield the Sword of Data – Part I
  - Brent Dykes, December 4, 2012
  - http://www.analyticshero.com/2012/12/04/data-driven-design-dare-to-wield-the-sword-of-data-part-i/
- Datavis.ca
  - http://www.datavis.ca/index.php

Bibliography

- Diverging color schemes: Showing good data isn't enough; you need to show it well
  - Alberto Cairo, June 26, 2016
  - http://www.thefunctionalart.com/2016/06/diverging-color-schemes-showing-good_26.html
- 8 Horrible Data Visualizations That Make No Sense
  - Eric Limer, September 02, 2013
  - http://gizmodo.com/8-horrible-data-visualizations-that-make-no-sense-1228022038
- 55 Striking Data Visualization and Infographic Poster Designs
  - Igor Ovsyannykov, May 16, 2011
  - http://inspirationfeed.com/inspiration/infographics/55-striking-data-visualization-and-infographic-poster-designs/
- 4 Tips for Promoting Predictive Analytics in Your Organization
  - Fern Halper, September 26, 2017
  - https://tdwi.org/articles/2017/09/26/ADV-ALL-4-Tips-for-Promoting-Predictive-Analytics.aspx

Course Contents

Unit IV	Data Visualization and Data Wrangling	(07 Hours)
Data Wrangling: Hierarchical Indexing, Combining and Merging Data Sets Reshaping and Pivoting. Data Visualization matplotlib: Basics of matplotlib, plotting with pandas and seaborn, other python visualization tools **Data Visualization Through Their Graph Representations:** Data and Graphs Graph Layout Techniques, Force-directed Techniques Multidimensional Scaling, The Pulling Under Constraints Model, Bipartite Graphs		


PUBLISHING
ENRICHMENT
DISCOVERING
TASKS OF
DATA WRANGLING
STRUCTURING
VALIDATING
CLEANING

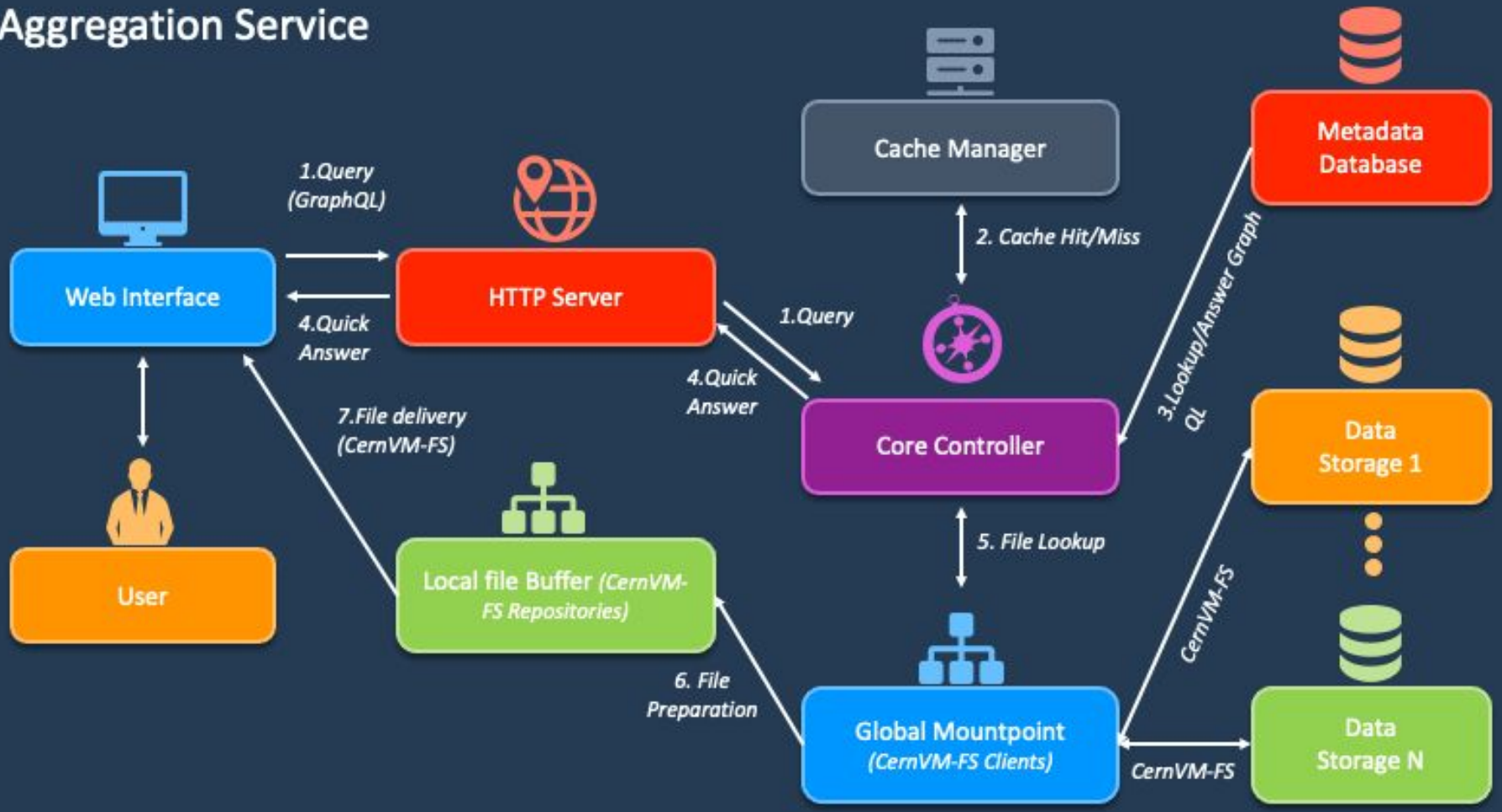

STEPS OF DATA WRANGLING
01 Discovering
02 Structuring
03 Cleaning
04 Enriching
05 Validating
06 Publishing

Course Contents

Unit V	Data Aggregation and Analysis	(07 Hours)
Data Aggregation and Group operations: Group by Mechanics, Data aggregation, General split- apply-combine, Pivot tables and cross tabulation 67 Time Series **Data Analysis:** Date and Time Data Types and Tools, Time series Basics, date Ranges, Frequencies and Shifting, Time Zone Handling, Periods and Periods Arithmetic, Resampling and Frequency conversion, Moving Window Functions.		

DATA AGGREGATION

Data Aggregation Service



Data Analysis: Step to Step Guide

1. Define your questions carefully

2. Establish measurement priorities

3. Collect all the relevant data.

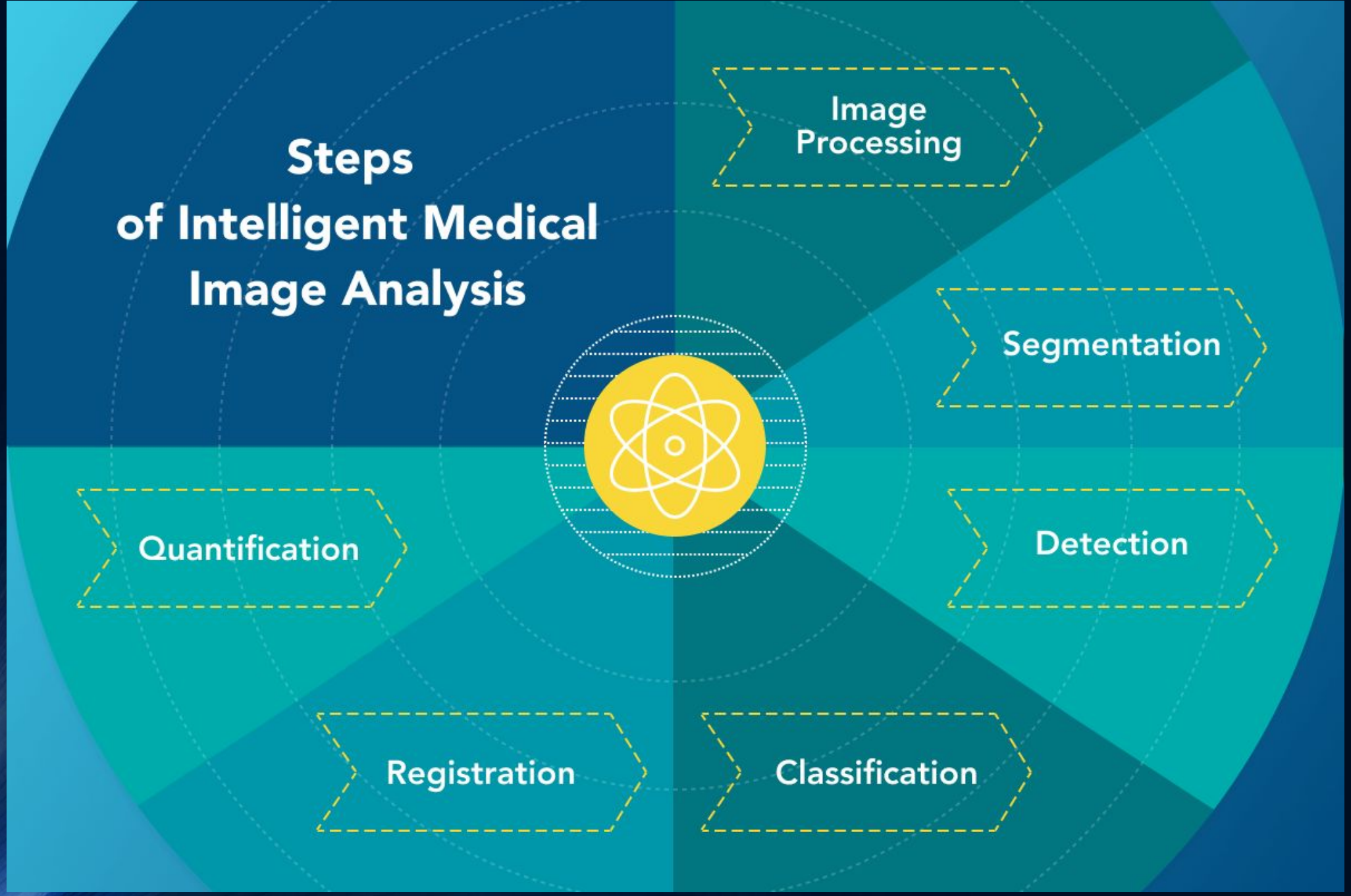
4. Analyze the data you gathered

5. Review and interpret the results.

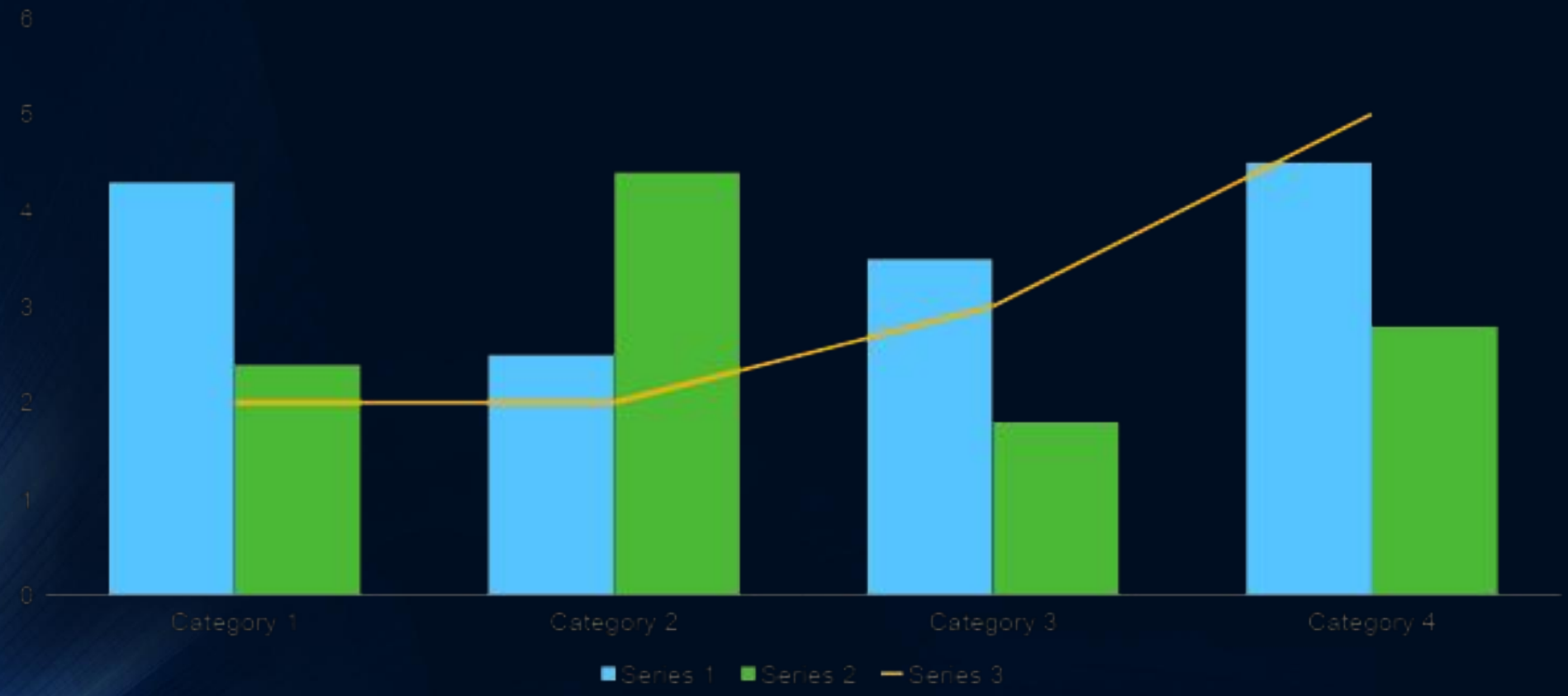
QuestionPro

Course Contents

Unit VI	Data Analysis of Visualization and Modelling	(07 Hours)
Reconstruction, Visualization and Analysis of Medical Images Introduction: - PET Images, Ultrasound Images, Magnetic Resonance Images, Conclusion and Discussion, Case Study: ER/Studio, Erwin data modeler, DbSchema Pro, Archi, SQL Database Modeler, LucidChart, Pgmodeler		


Steps
of Intelligent Medical
Image Analysis
Image
Processing
Segmentation
Detection
Classification
Registration
Quantification

Title and Content Layout with Chart



Two Content Layout with Table

- First bullet point here
- Second bullet point here
- Third bullet point here

Class	Group 1	Group 2
Class 1	82	95
Class 2	76	88
Class 3	84	90

We shape our Data Model , because our mind Visualized us…..