Introduction to Data Analytics

Dr. Kavi Mahesh

Director, IIIT-Dharwad

06 March 2019

Copyright/Intellectual Property

Note: These slides are meant only for students of IIIT Dharwad in the Data Analytics course.

DO NOT UPLOAD to any web site.

DO NOT share or distribute with others.

Topics in Data Analytics

- Data Analytics and Visualization
- Predictive Analytics
  - Classification
  - Cluster Analysis
  - Recommendation
- Regression and Correlation
  - Logistic Regression
- Time Series Analysis
- Simulation
- Text Mining

- Basics of R
- Statistics

Syllabus: Unit 1

- Data Analytics and Visualization
  - Overview of data science and data analytics
  - Data warehousing, data mining and business intelligence
  - Data collection, curation and modelling
  - Descriptive and predictive analytics
  - Applications in data visualization, forecasting, recommendation, social sciences, etc.
  - Introduction to R language
  - Review of basic statistics

Syllabus: Unit 2

- Predictive Analytics
  - Introduction to prediction
  - Clustering and classification
  - Distance and similarity measures
  - Hierarchical clustering algorithms, K-Means, DBScan clustering method
  - Classification methods
  - Problems with high-dimensional data & Techniques for dimensionality reduction

Syllabus: Unit 3

- Regression and Correlation
  - Correlation and covariance
  - Linear and non-linear regression
  - Logistic regression
  - Prediction using regression

Syllabus: Unit 4

- Time Series Analysis
  - Time as a dimension
  - Trends, seasonality, periodicity, and noise
  - Exponential smoothing and the Holt Winters method
  - Auto correlation and ARIMA
  - Introduction to frequency and spectral analysis

Syllabus: Unit 5

- Simulation
  - Introduction to simulation
  - Monte Carlo simulations
  - Re-sampling methods

Syllabus: Unit 6

- Text Mining and Analytics
  - Introduction to computational linguistics, text analysis and mining
  - Featurization
  - Term frequencies, vectors and matrices
  - Introduction to Social Network Analysis
  - Applications of text mining

Books

- Data Analytics Using R by Seema Acharya, McGraw Hill Education, 2018
- Data Analytics by Anil Maheshwari, McGraw Hill, 2017
- Data Analysis with Open Source Tools by Philipp K. Janert, O'Reilly, 2010
- Predictive Analytics by Eric Siegel, Wiley, 2013
- Any book on the R language

Evaluation Plan

- Quizzes: 5 + 10 = 15
- Mid-term: 20
- Lab assignments (4 or 5): 25
- Final Exam: 40
- **Total** **100**

Do’s and Don’ts

- Attend classes
- Participate productively in classes
- Get really good in R
- Play with data

- Don’t try to find shortcuts
- Don’t procrastinate
- ***Don’t demotivate me!***

What is Data?

What is Data

- Data is an arrangement of known symbols
- Data existed before computers

- Symbols must be known – an alphabet
- Arrangement must also be known
- Arrange for efficiency of storage and/or processing

Data

- Data was always:
  - Collected
  - Analyzed
  - Managed
- Structured, unstructured, semi-structured…
- Computers never understood data…

Data x/y/z/…

- Data structure
- Database
- Data warehouse
- Data mining
- Data analytics
- Data modelling

Data Science

- Science of data?

  Or

- Science using data?

- The Fourth Paradigm
  - Data-intensive scientific discovery

- Epistemology of Data Science

Types of Data

- Structured data
- Unstructured data
  - Semi-structured data, e.g., text, web, …
- Multimedia data
- Metadata
- Open data
- Real-world data
- Real-time data
- Digital data?

Types of Data

- Numerical data
  - Ordinal
  - Cardinal
  - Ratio
  - Dichotomy
- Nominal data

Things to do to Data

- Data collection
- Data curation
- Data loading: ETL – Extract-Transform-Load
- Data analytics
- Data visualization
- Data aggregation
- Data transformation
- Data management
- Data compression
- Data encryption

What about Information and Knowledge?

- Information is data in context
- Knowledge is the ability to do things using information
- Wisdom is known when to do what

What is Analytics?

- Analysis versus Analytics
- Systematic quantitative analysis to (help) find patterns and insights
- Data-driven decision making
  - Examples

What is Analytics?

- Reduction of data to understandable findings
- Analyzing data to find useful insights

Why Analytics?

- To understand a process or phenomenon
- To discover trends, patterns, …
- To predict the future
- To arrive at a conclusion
- To visualize

Why now?

- Lots of data available
- Computing power is adequate
- Cost of not doing analytics is high

Types of Analytics

- Data analytics
- Text analytics
- Web analytics
- Graph analytics
- Social Network Analysis
- Learning analytics
- Knowledge analytics
- …

Types of Analytics

- Descriptive analytics
  - E.g., how much water is in KRS and other dams
- Predictive analytics
  - E.g., how much monsoon rain is expected in 2016
- Prescriptive analytics
  - E.g., Karnataka, thou shalt release 2000 cusecs of water per day to Tamil Nadu for the next 200 days

Applied Analytics

- Big Data Analytics
- Business Analytics
- Financial Analytics
- Econometrics
- Text Analytics
- Visual Analytics
- Graph Analytics
- Social Network Analysis
- Learning Analytics
- Content Analytics / Web Analytics
- Knowledge Analytics, RDF Analytics
- Scientometrics, Informetrics, Bibliometrics, Altmetrics, …

Descriptive Analytics

- What is in the data?
- How is the data distributed?
- Use of statistical and other techniques
- To understand past data
- E.g., analytics of IPL or Indian cricket teams

Predictive Analytics

- Extrapolate trends to the future
- E.g., monsoon rainfall prediction

Prescriptive Analytics

- What action to take based on data
- E.g., Should GoK continue or close Anna Bhagya scheme or modify it? How?

Data Science

- Publishing and sharing data
- Doing research based on data
- The Fourth Paradigm
  - Data-Intensive Scientific Discovery
- Not just data processing…

Who is an Analyst?

- Programmer?
- Computer scientist?
- Statistician?
- Mathematician?
- MBA?
- Linguist?
- Journalist?

Tasks in Analytics

- Design
- Data collection
- Data curation
- Preprocessing
- Analysis
- Presentation
- Reporting

Business Intelligence

- Has nothing to do with AI
- How to get insights from data to make business decisions
- Example: How to manage inventory and distribution of products

Data Warehousing

- Historical data only
- Analyzing trends and other patterns
- E.g., footfalls and sales in a store
- E.g., inventory management

The Problem of Analytics

- Lots of data… and the promise of analytics
  - Unstructured, semi-structured, linked, semantic
- Hardly any knowledge organization
- Poorly defined semantics of data
- Free-form tags
  - Good enough for search and browsing
  - Not for analytics and visualization

The Promise of Analytics

Can Analytics deal with well organized data?

- Analytics is currently not well equipped to handle data with rich semantics or knowledge structures
- Analytical algorithms do not work well on complex data with hierarchy and graph structures

Solution

- New analytical operators defined on knowledge structures by considering semantic constraints in ontology
- A Framework for Knowledge Analytics

The Promise of Analytics

- Data to Insights: find patterns, trends, models…
  - Data mining and machine learning
- Diagnosis, Prediction, Prescription, Data

Source: blog.profoundis.com

State of Art in Analytics

- (Big) Data Analytics
- Graph Analytics
- Social Network Analysis
- Applied analytics:
  - Learning analytics, Web analytics, Scientometrics, …
- Many technologies, e.g., R

Typical Analytical Tasks for Structured Data

- Retrieve Value
- Filter and Sort
- Compute Derived Value
- Find Centers and Extremum
- Determine Range
- Characterize Distribution
- Find Anomalies
- Cluster
- Correlate

Source: Amar, Robert, James Eagan, and John Stasko. "Low-level components of analytic acti vity in information visualization." *Information Visualization, 2005. INFOVIS 2005. IEEE Symp osium on*. IEEE, 2005.

Data Collection, Curation and Modelling

- Example: Historical data about births in Karnataka

Applications of Data Analytics

- Visualization
- Forecasting
- Recommendation
- Decision Making in Business / Governance
- Social Sciences

Introduction to R

What is in Data?

- Element : a piece of data
- Data structures and databases
- Unstructured data

- Kavi Mahesh’s definition of data: an arrangement of known symbols.

Big Data?

When was it small??

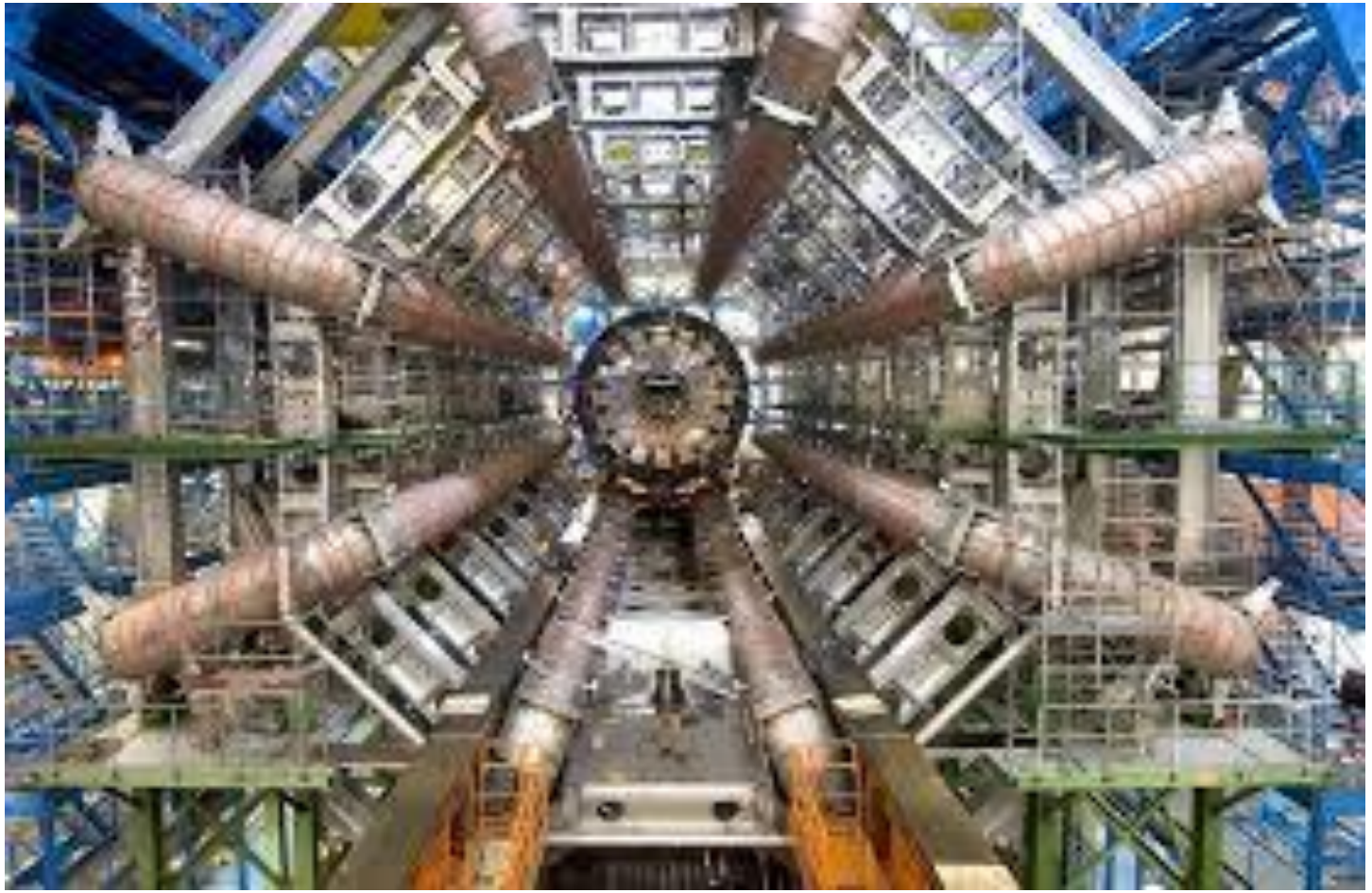
Very Large Data


Proceedings
30th INTERNATIONAL CONFERENCE on
Very Large Data Bases
Editors:
Mario A. Nascimento
M. Tamer Özsu
Donald Kossmann
Renée J. Miller
José A. Blakeley
Berni Schiefer
August 31 - September 3, 2004
Toronto, Canada

The 4 V’s and the I of Things

- Volume
- Variety
- Velocity
- Veracity

Data Science?



GENERAL ASSEMBLY

Data Science: Hype vs Reality

Szilárd Pafka, PhD
Chief Scientist, Epoch

Open Data: Challenges

- How to make data open?
- How to share it?
- How to make it self-contained?
- …

Data Mining, Machine Learning, Analytics, Business Intelligence…

Does it make sense?

- Computers still cannot make sense out of it all…
- Machines can’t really learn
  - Don’t know what is “deep learning”
- Data can’t really be mined
- Images can be processed, but not appreciated…

Where did it all Begin?

A COMPUTER WANTED.

WASHINGTON, May 1.—A civil service examination will be held May 18 in Washington, and, if necessary, in other cities, to secure eligibles for the position of computer in the Nautical Almanac Office, where two vacancies exist—one at $1,000, the other at $1,400.

The examination will include the subjects of algebra, geometry, trigonometry, and astronomy. Application blanks may be obtained of the United States Civil Service Commission.

The New York Times
Published: May 2, 1892

Information Management Using Computers

- Numerical Era
  - Number crunching
  - Scientific computing
- Data Era
  - Database management
  - On-line transactions
  - Analysis (OLAP)
- The Internet Era
  - Information management?
- What next?

How Much Data

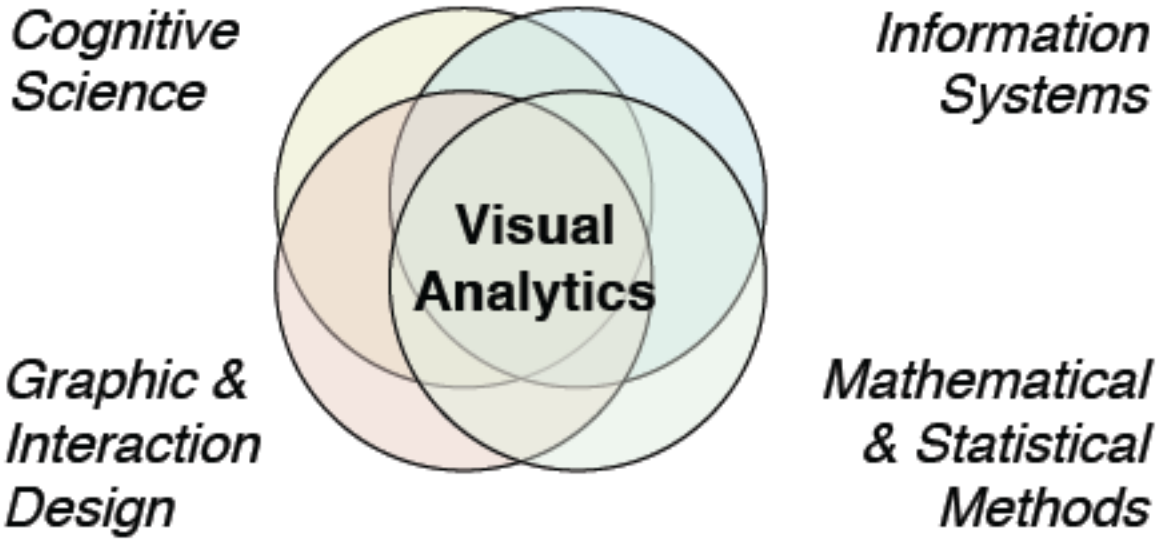
- 667,000 TeraBytes of email per year (?)
- WWW = 170 TeraB (75,000 TeraB)
- Internet moves 21 ExaBytes / month
- Global data size is 800 ExaBytes (2009)
- New data = 5 ExaBytes per year
  - 5,000,000,000,000,000,000 bytes
  - 1 ExaByte/year = 34 GBps
- 800 MegaBytes per human being per year
- Say 99.9% junk
- 0.1% = 5000 Tera of data to be managed
- E.g., 136 TeraB = 17 Mil books in LoC

Text Analytics (Wikipedia)

- The process of deriving high quality information from text.
- 'High quality' in text analytics usually refers to some combination of relevance, novelty, and interestingness.
- High quality information is typically derived through the discovery of patterns and trends through means such as statistical pattern learning.
- Text analytics usually involves
  - The process of structuring the input text
  - Deriving patterns within the structured data
  - Evaluation and interpretation of the output.

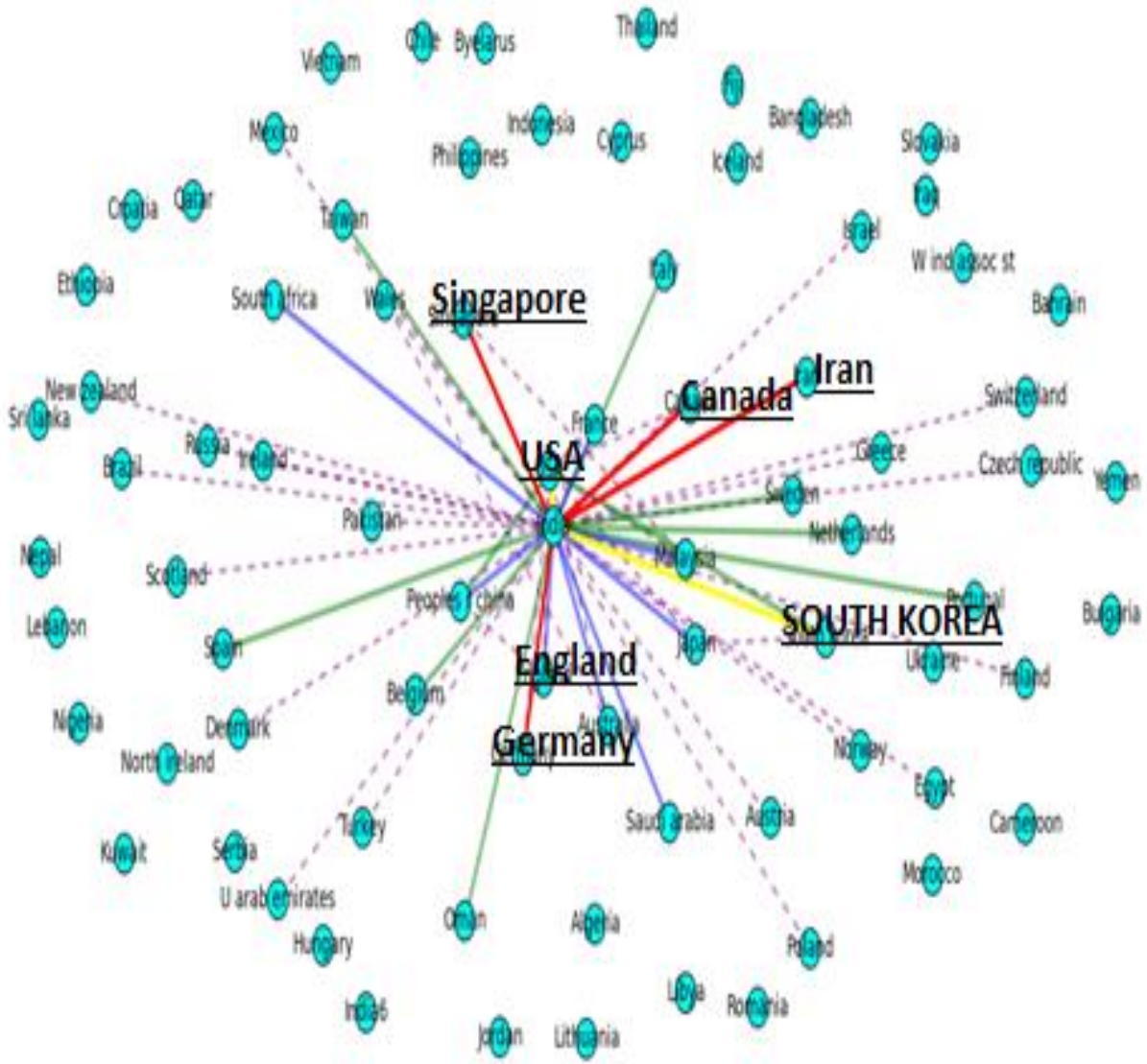
Visual Analytics (Wikipedia)

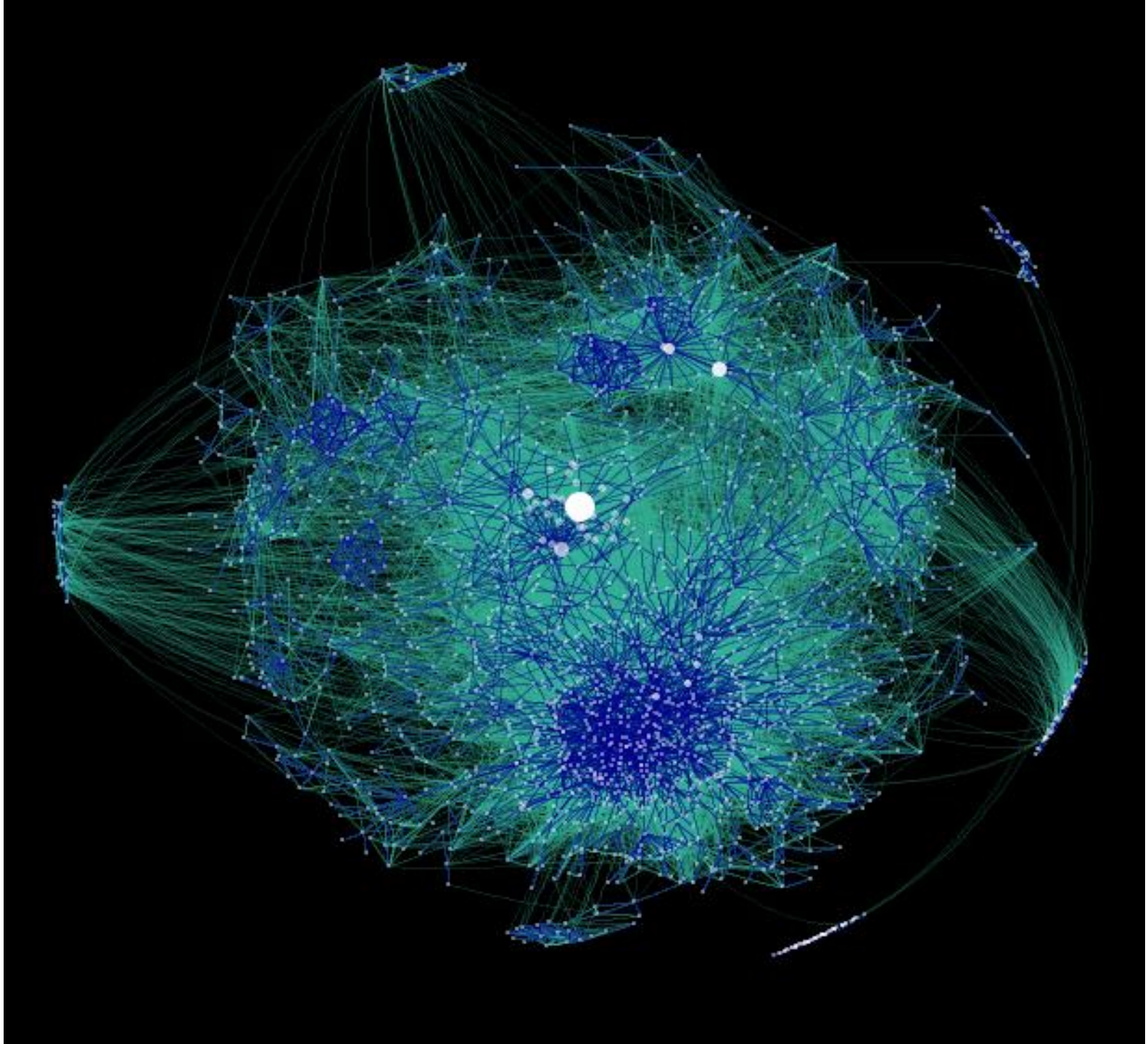
- The science of analytical reasoning facilitated by visual interactive interfaces



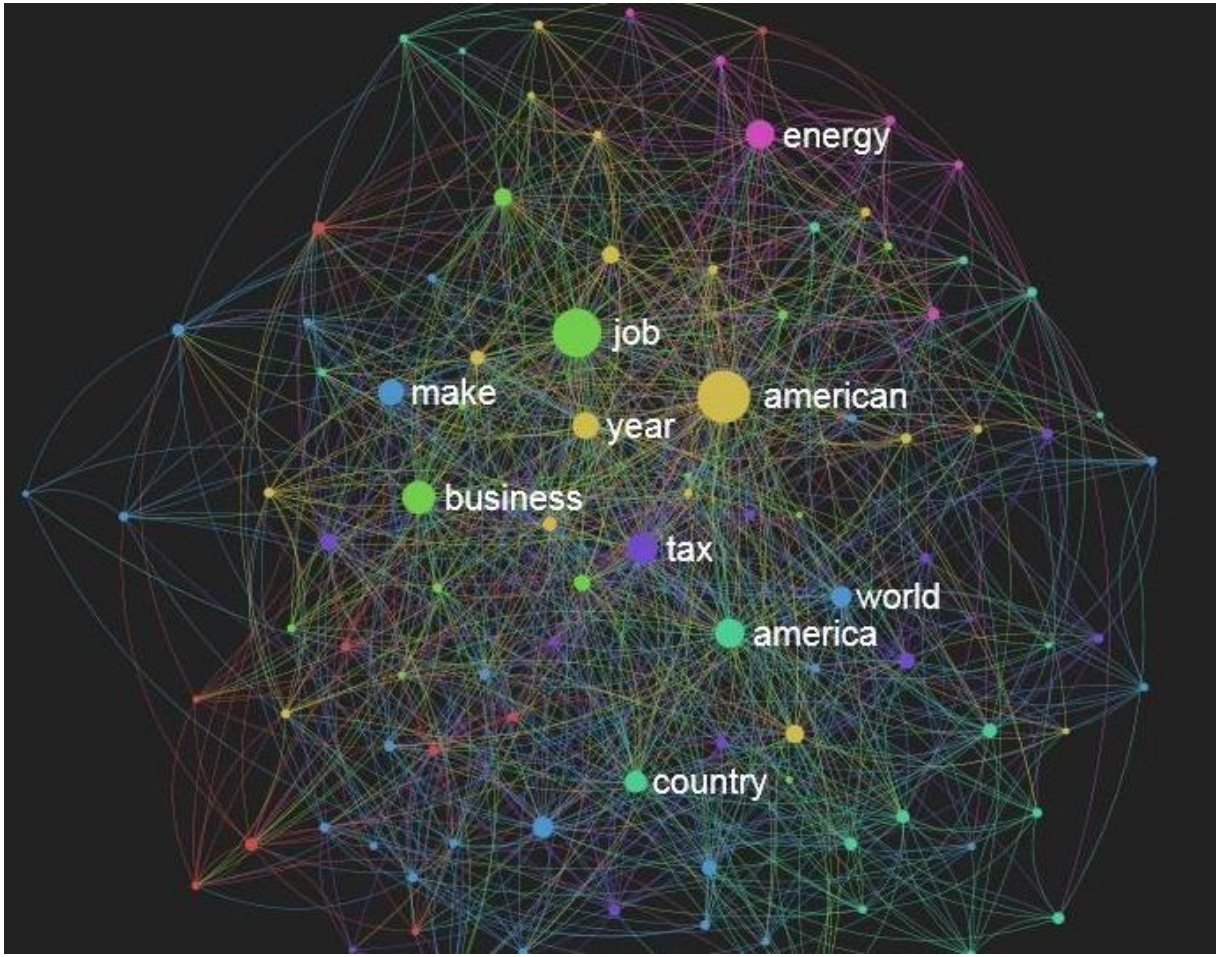
Visual Text Analytics

- Brings together:
  - Text Analytics
    - Information Systems
    - Mathematical and Statistical Methods
  - Visualization
    - Cognitive Science
    - Graphic and Interaction Design

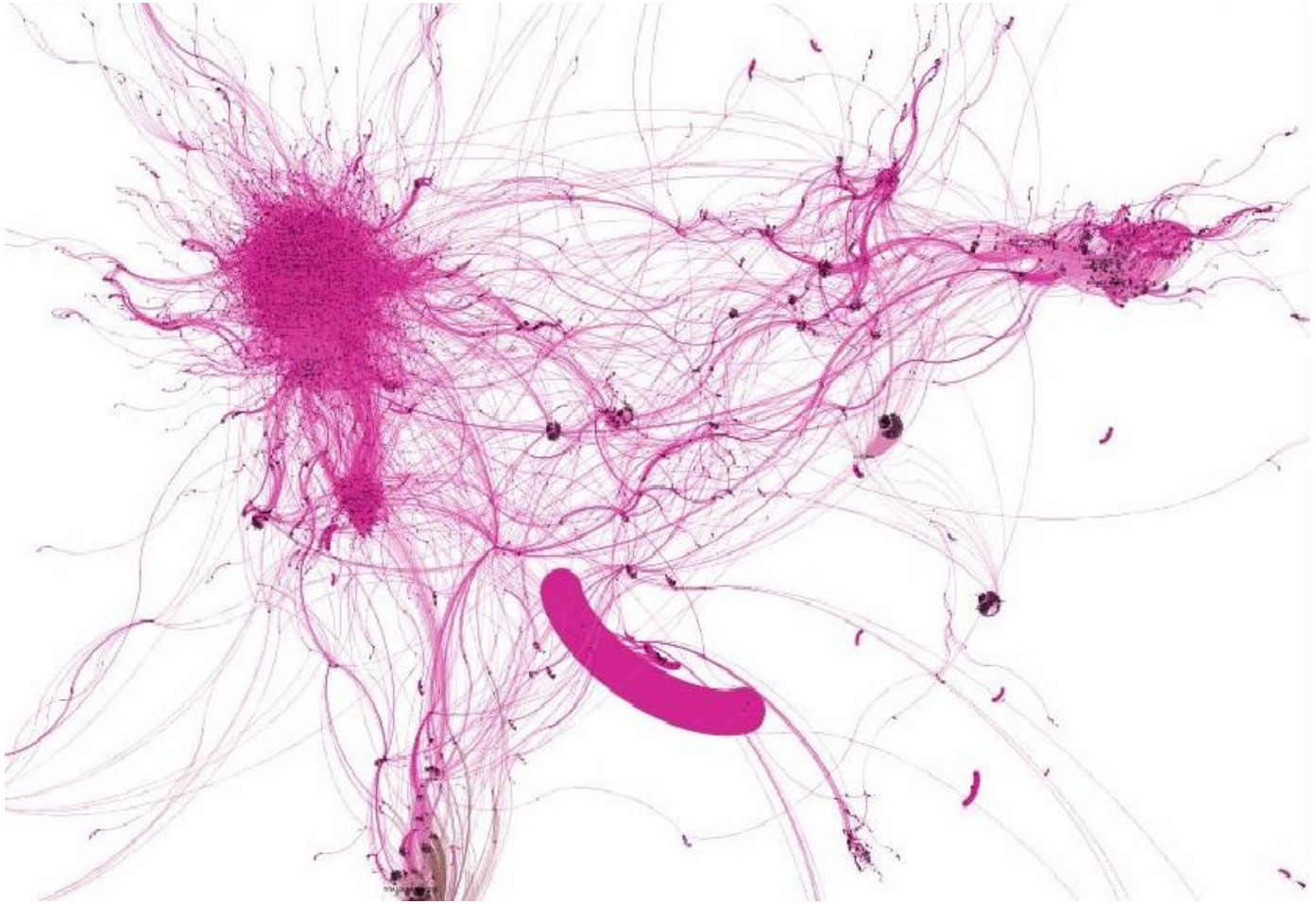




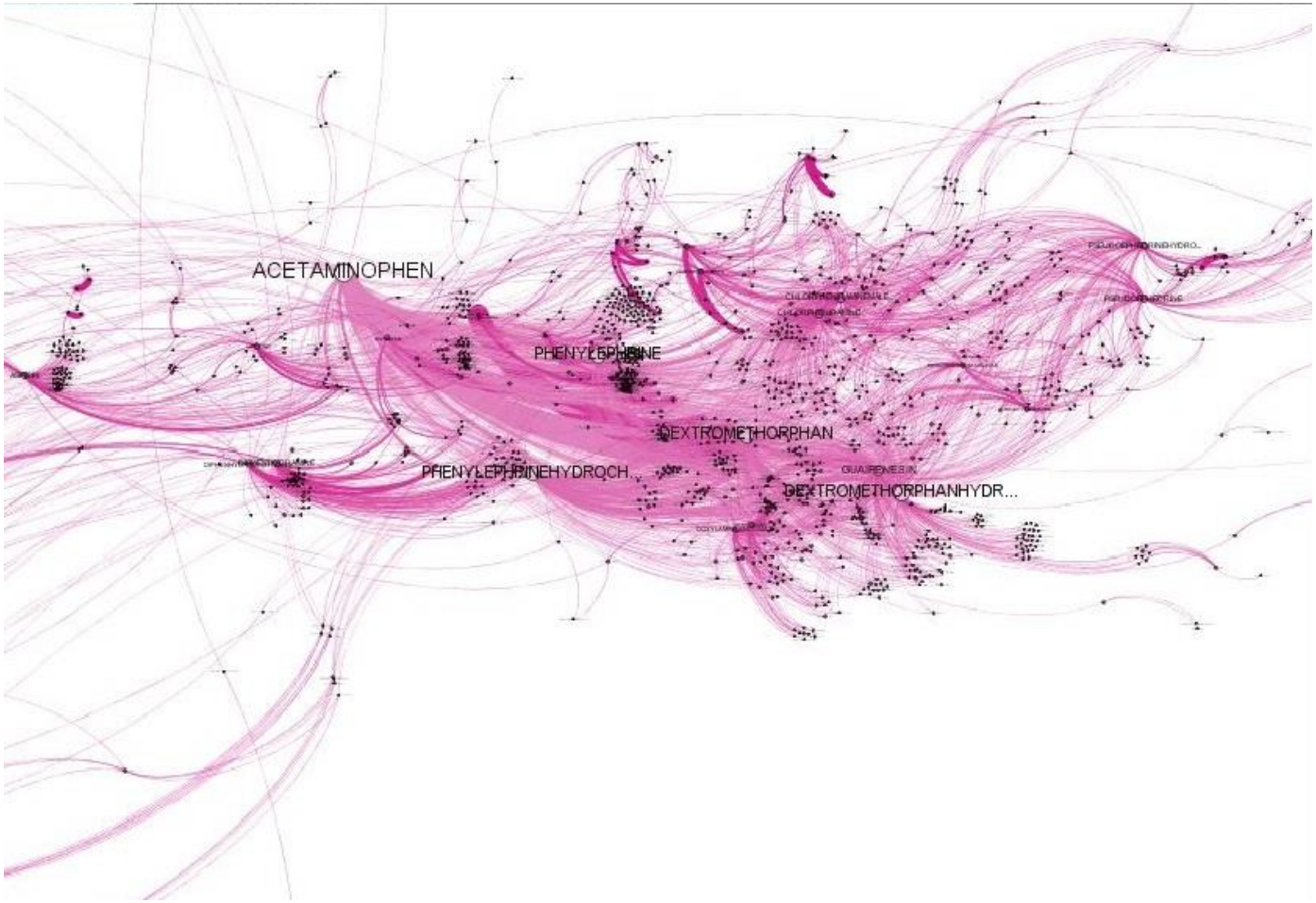
Text Visualization: Obama 2012



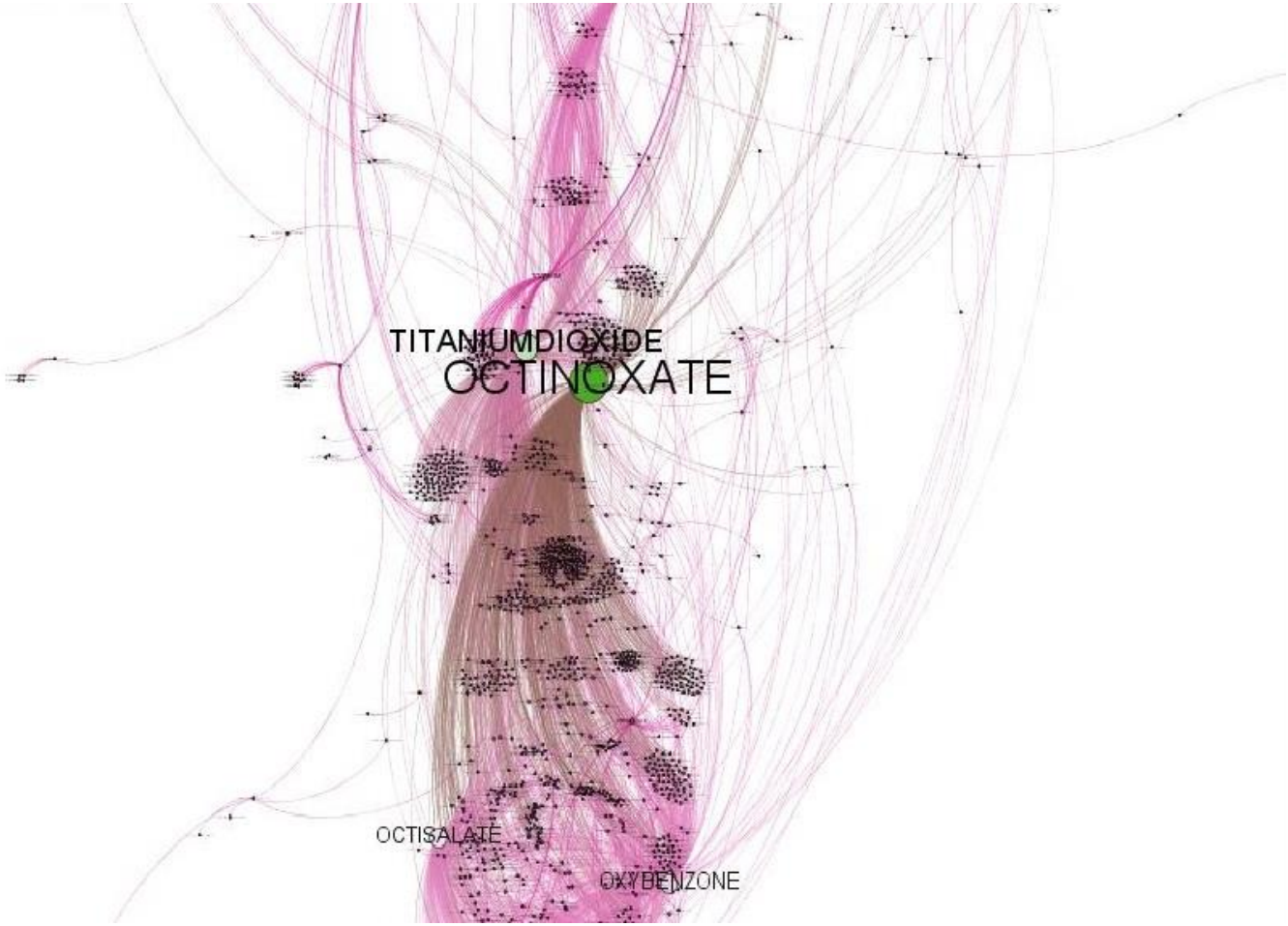
Cluster Visualization: Drugs



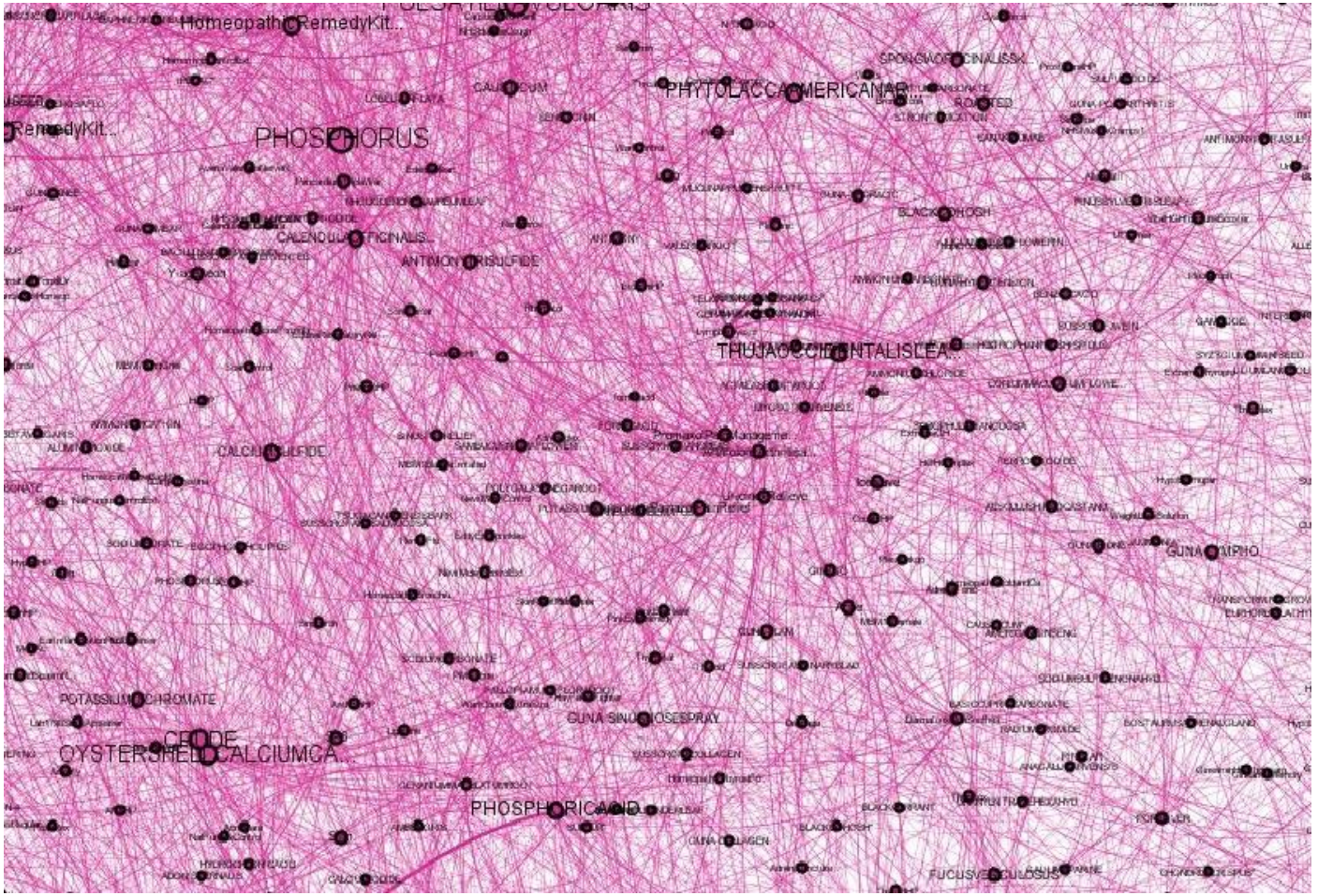
Drug Network 2



Drug Network 3



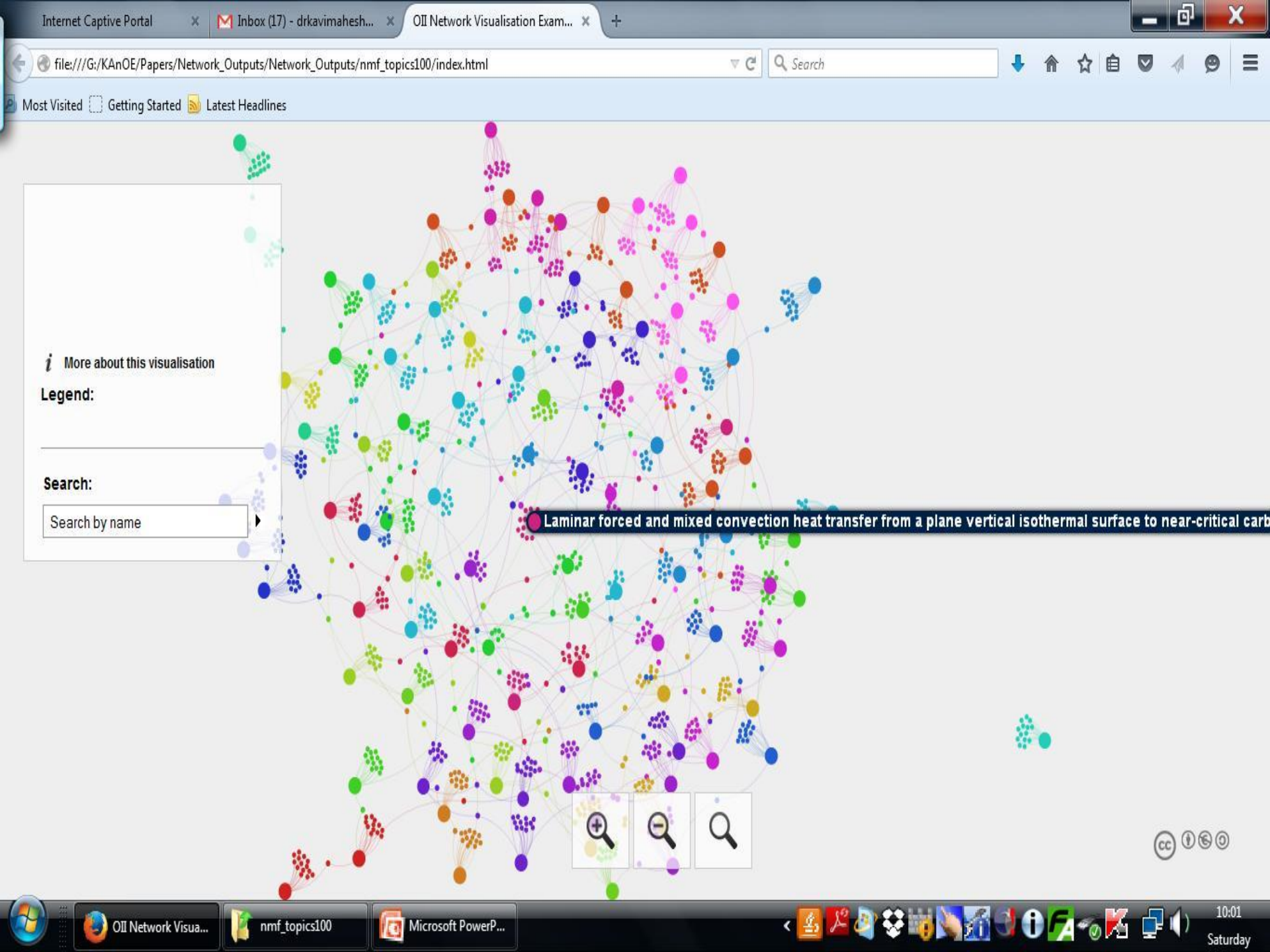
Drug Network 4

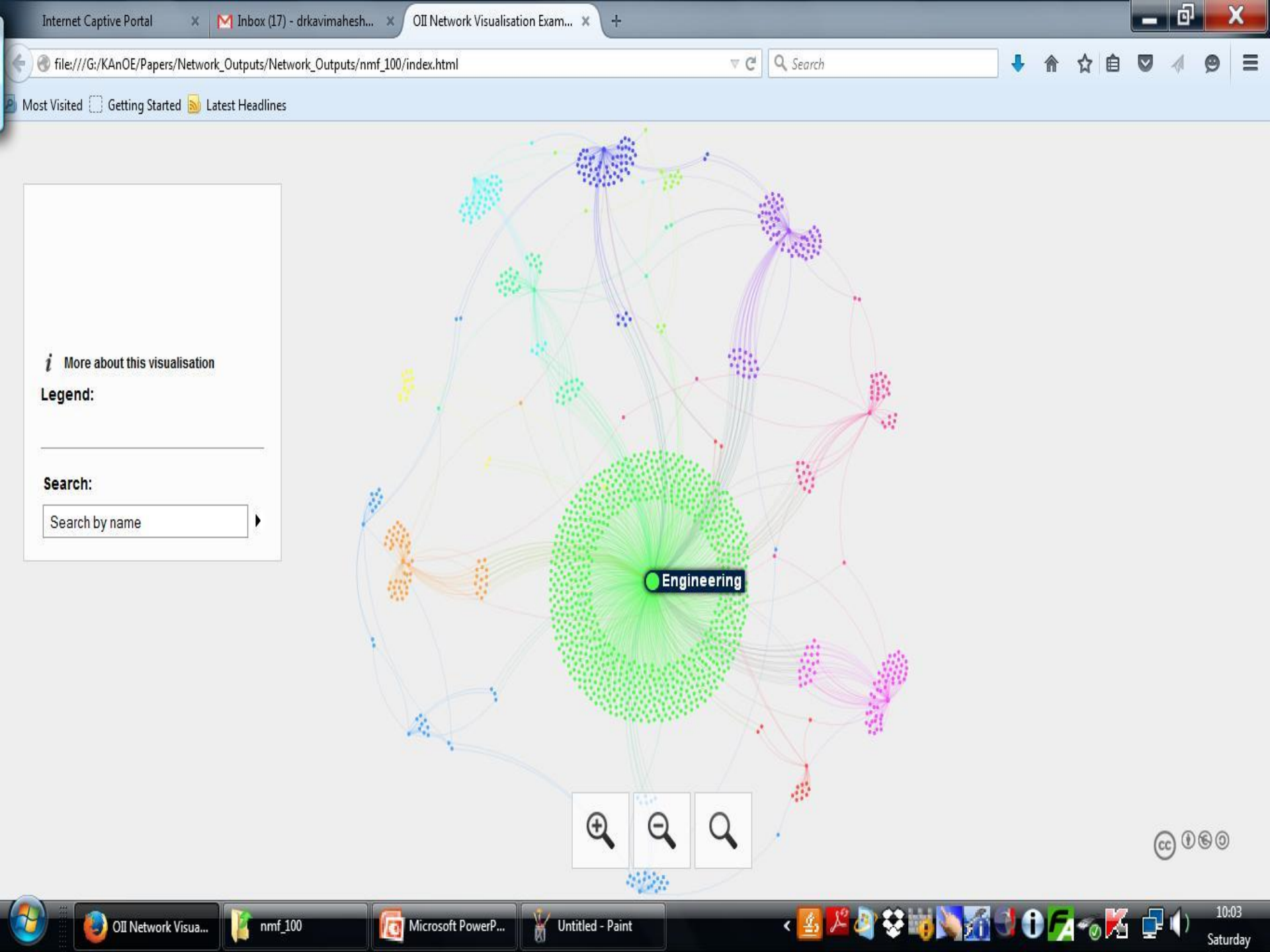


Text analytics: Topic Maps

Analyzing texts as networks:

- Topic maps
- Hypertext graphs
- Synthetic social networks


Internet Captive Portal
Inbox (17) - drkavimahesh...
OII Network Visualisation Exam...
file:///G:/KAnOE/Papers/Network_Outputs/Network_Outputs/nmf_topics100/index.html
Search
Most Visited
Getting Started
Latest Headlines
More about this visualisation
Legend:
Search:
Search by name
Laminar forced and mixed convection heat transfer from a plane vertical isothermal surface to near-critical carb
OII Network Visua...
nmf_topics100
Microsoft PowerP...
10:01
Saturday


Internet Captive Portal
Inbox (17) - drkavimahesh...
OII Network Visualisation Exam...
file:///G:/KAnOE/Papers/Network_Outputs/Network_Outputs/nmf_100/index.html
Search
Most Visited
Getting Started
Latest Headlines
More about this visualisation
Legend:
Search:
Search by name
Engineering
OII Network Visua...
nmf_100
Microsoft PowerP...
Untitled - Paint
10:03
Saturday

Let’s Look at Some Data

ALL-INDIA MONTHLY, SEASONAL AND ANNUAL RAINFALL SERIES FOR THE PERIOD 1813-2006. THE RAINFALL FIGURES ARE IN MM.

YEAR	J	F	M	A	M	J	JU	A	S	O	N	D	ANN	JF	MAM	JJAS	OND
1813	10.0	14.2	20.1	25.6	52.0	190.8	274.1	222.6	153.0	64.2	78.3	13.7	1118.6	24.1	97.7	840.6	156.2
1814	9.0	12.3	14.2	22.5	44.3	133.1	242.4	246.1	215.8	55.8	16.2	9.7	1021.4	21.3	81.1	837.3	81.7
1815	18.5	14.7	16.5	26.1	51.4	149.8	373.0	215.9	156.7	71.4	92.6	15.7	1202.5	33.2	94.1	895.5	179.7
1816	8.5	16.0	14.1	22.2	44.8	134.6	297.5	244.2	217.9	54.6	27.7	6.2	1088.5	24.5	81.2	894.2	88.6
1817	10.5	12.8	14.7	23.2	46.7	242.3	269.9	226.0	238.2	103.9	56.7	15.5	1260.4	23.2	84.6	976.6	176.0
1818	10.0	14.2	16.2	30.4	50.5	187.3	319.1	340.3	189.5	115.4	65.2	20.0	1357.9	24.1	97.1	1036.1	200.5
1819	8.5	12.1	13.5	38.1	42.0	135.8	287.6	244.0	197.5	38.5	28.1	5.8	1051.5	20.7	93.6	864.9	72.3
1820	9.4	13.4	71.3	27.0	134.3	163.1	310.4	275.7	163.0	84.3	14.3	43.9	1310.0	22.7	232.6	912.2	142.5
1821	25.1	13.2	16.5	33.7	43.6	140.5	254.5	286.6	205.8	75.4	27.6	11.7	1134.3	38.3	93.9	887.3	114.7
1822	18.5	12.6	13.9	26.1	43.2	186.7	266.0	310.1	197.6	99.7	46.3	9.3	1230.1	31.1	83.2	960.4	155.3
1823	15.1	12.7	21.3	22.0	44.7	163.8	242.6	254.0	135.1	57.1	4.7	3.9	976.9	27.8	88.0	795.4	65.8
1824	16.9	15.0	16.9	26.7	52.9	115.9	242.7	285.3	126.3	91.7	28.7	12.6	1031.4	31.8	96.4	770.2	132.9
1825	10.3	13.5	15.9	25.1	72.9	186.6	297.5	269.2	167.6	98.0	28.4	17.9	1202.8	23.8	113.9	920.9	144.3

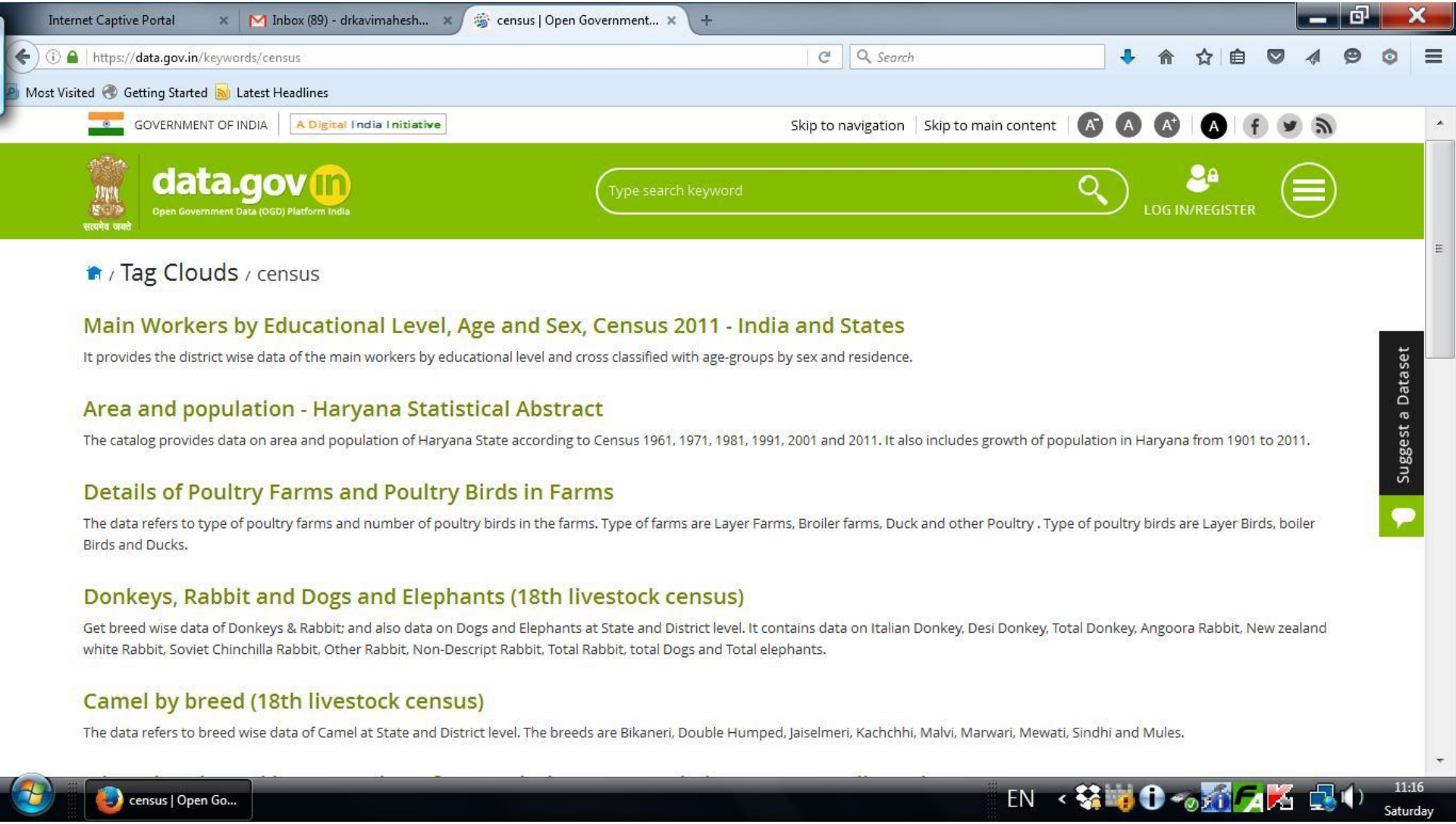
Sample Rainfall Series data – All India, obtained from IITM, Pune

Applied Analytics: Examples

- Analytics for India
- Where is the data?
- Examples
  - Census data analytics (demographic studies)
  - Econometrics
  - Scientometrics
  - Social and Political analytics
  - E-governance
  - Learning analytics
  - …

Analytics for India

- Social relevance for India
- Difficulty of getting data
- Datasets curated and analyzed:
  - Rainfall
  - Temperature
  - Agriculture: yield, price
  - Census (2011)
  - Terrorism


Internet Captive Portal
Inbox (89) - drkavimahesh...
census | Open Government...
https://data.gov.in/keywords/census
Search
Most Visited
Getting Started
Latest Headlines
GOVERNMENT OF INDIA
A Digital India Initiative
Skip to navigation
Skip to main content
data.gov.in
Open Government Data (OGD) Platform India
Type search keyword
LOG IN/REGISTER
/ Tag Clouds / census
Main Workers by Educational Level, Age and Sex, Census 2011 - India and States
It provides the district wise data of the main workers by educational level and cross classified with age-groups by sex and residence.
Area and population - Haryana Statistical Abstract
The catalog provides data on area and population of Haryana State according to Census 1961, 1971, 1981, 1991, 2001 and 2011. It also includes growth of population in Haryana from 1901 to 2011.
Details of Poultry Farms and Poultry Birds in Farms
The data refers to type of poultry farms and number of poultry birds in the farms. Type of farms are Layer Farms, Broiler farms, Duck and other Poultry . Type of poultry birds are Layer Birds, boiler Birds and Ducks.
Donkeys, Rabbit and Dogs and Elephants (18th livestock census)
Get breed wise data of Donkeys & Rabbit; and also data on Dogs and Elephants at State and District level. It contains data on Italian Donkey, Desi Donkey, Total Donkey, Angoora Rabbit, New zealand white Rabbit, Soviet Chinchilla Rabbit, Other Rabbit, Non-Descript Rabbit, Total Rabbit, total Dogs and Total elephants.
Camel by breed (18th livestock census)
The data refers to breed wise data of Camel at State and District level. The breeds are Bikaneri, Double Humped, Jaiselmeri, Kachchhi, Malvi, Marwari, Mewati, Sindhi and Mules.
Suggest a Dataset
census | Open Go...
EN
11:16
Saturday

Office of the Registrar General & Census Commissioner, India
Ministry of Home Affairs,
Government of India

सत्यमेव जयते

Houselisting and Housing Census Data - 2011

- Percentage of Scheduled Caste/ Scheduled Tribe Households to total Households by Amenities and Assets (India & States/UTs - Sub-District Level)
- Percentage of Households to Total Households by Amenities and Assets (India & States/UTs - Village and Ward Level)
- Percentage of Households to Total Households by Amenities and Assets (India & States/UTs - Sub-District Level)
- Houselisting and Housing Census Data Tables (Total, SC/ST)(India & States/UTs - District Level)
- Tables on Houses, Household Amenities & Assets Among Female Headed Households
- Tables on Housing Stock, Amenities & Assets in Slums - Census 2011

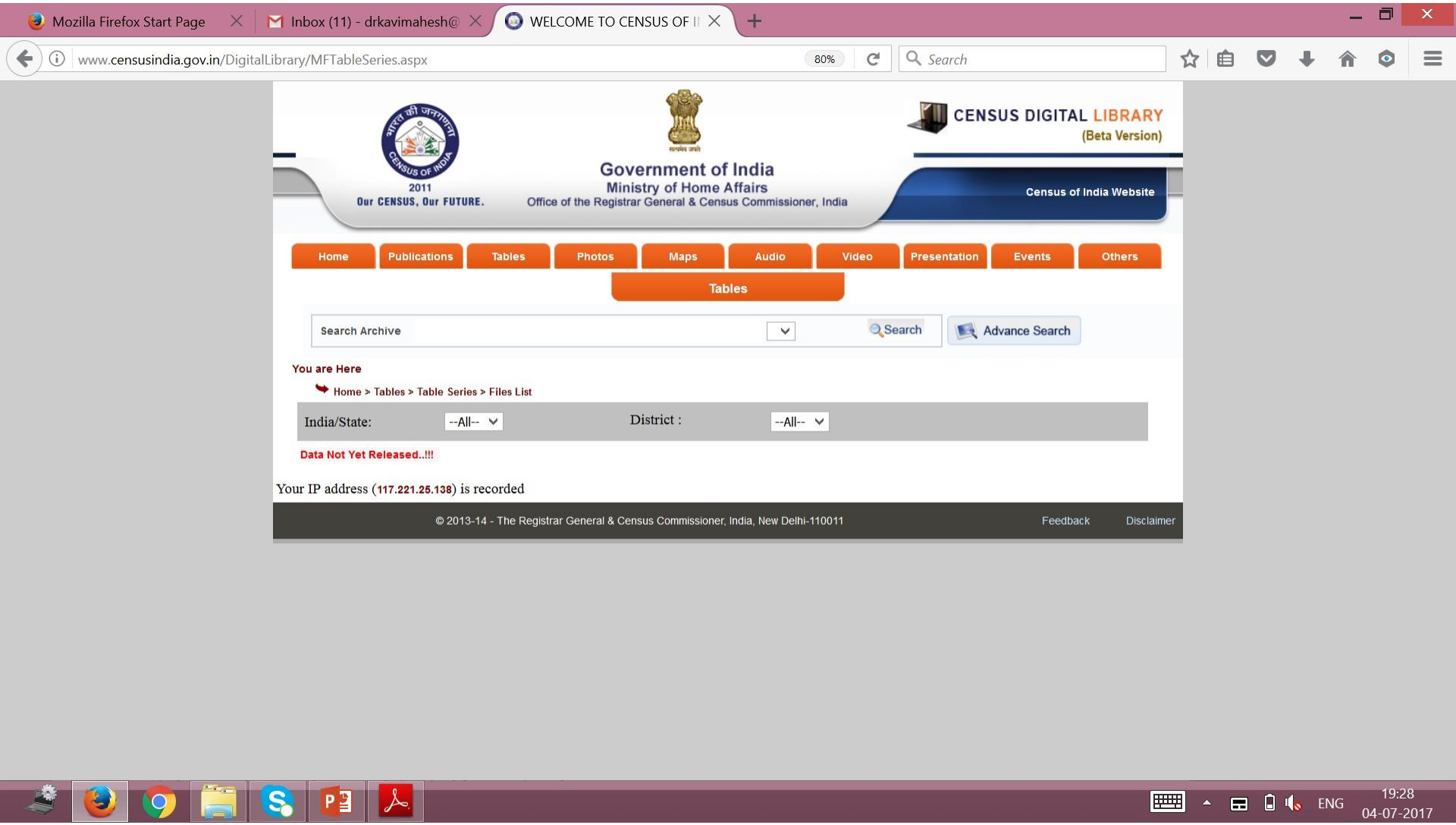
Population Finder 2011
Census Digital Library
District Census Handbook
Medical Certification of Cause of Death
SRS Publications
Data Dissemination Unit
CensusInfo
Media/News
Archive

Contact Us
Jobs @ Census
Tender

Note : For any Data / Customized Tabulation from the level other than above, Please Contact Data Dissemination Unit (DDU)

Last Updated on: 04-07-2017 12:10:00 IST

- **General Economic Tables - (B - Series)**
  - **B-1 to B-10**
    - B-1 : Main Workers,Marginal Workers,Non-Workers and those Marginal Workers,Non-Workers Seeking/ Available for Work Classified by Age and Sex
    - B-1 City : Main Workers,Marginal Workers,Non-Workers and those Marginal Workers,Non-Workers Seeking/ Available for Work Classified by Age and Sex
    - B-1 SC : Main Workers,Marginal Workers,Non-Workers and those Marginal Workers,Non-Workers Seeking/ Available for Work Classified by Age and Sex for Scheduled Castes
    - B-1 ST : Main Workers,Marginal Workers,Non-Workers and those Marginal Workers,Non-Workers Seeking/ Available for Work Classified by Age and Sex for Scheduled Tribes
    - B-2 : Main Workers,Marginal Workers,Non-Workers and Those Marginal Workers,Non-Workers Seeking/ Available for Work Classified By Age,Sex and Religious Community
    - B-3 : Main Workers,Marginal Workers,Non-Workers and those Marginal Workers,Non-Workers Seeking/ Available for Work Classified by Educational Level and Sex
    - B-3 City : Main Workers,Marginal Workers,Non-Workers and those Marginal Workers,Non-Workers Seeking/ Available for Work Classified by Educational Level and Sex
    - B-3 SC : Main Workers,Marginal Workers,Non-Workers and those Marginal Workers,Non-Workers Seeking/ Available for Work Classified by Educational Level and Sex for Scheduled Castes
    - B-3 ST : Main Workers,Marginal Workers,Non-Workers and those Marginal Workers,Non-Workers Seeking/ Available for Work Classified by Educational Level and Sex for Scheduled Tribes
    - B-4 : Main Workers Classified by Age,Industrial Category and Sex
    - B-4 City : Main Workers Classified by Age,Industrial Category and Sex
    - B-4 SC : Main Workers Classified by Age,Industrial Category and Sex for Scheduled Castes
    - B-4 ST : Main Workers Classified by Age,Industrial Category and Sex for Scheduled Tribes


Mozilla Firefox Start Page
Inbox (11) - drkavimahesh@
WELCOME TO CENSUS OF I
www.censusindia.gov.in/DigitalLibrary/MFTableSeries.aspx
80%
Search
CENSUS DIGITAL LIBRARY
(Beta Version)
2011
Our CENSUS, Our FUTURE.
Government of India
Ministry of Home Affairs
Office of the Registrar General & Census Commissioner, India
Census of India Website
Home
Publications
Tables
Photos
Maps
Audio
Video
Presentation
Events
Others
Tables
Search Archive
Search
Advance Search
You are Here
Home > Tables > Table Series > Files List
India/State:
--All--
District :
--All--
Data Not Yet Released..!!!
Your IP address (117.221.25.138) is recorded
© 2013-14 - The Registrar General & Census Commissioner, India, New Delhi-110011
Feedback
Disclaimer
ENG
19:28
04-07-2017

Our Own Data Collection

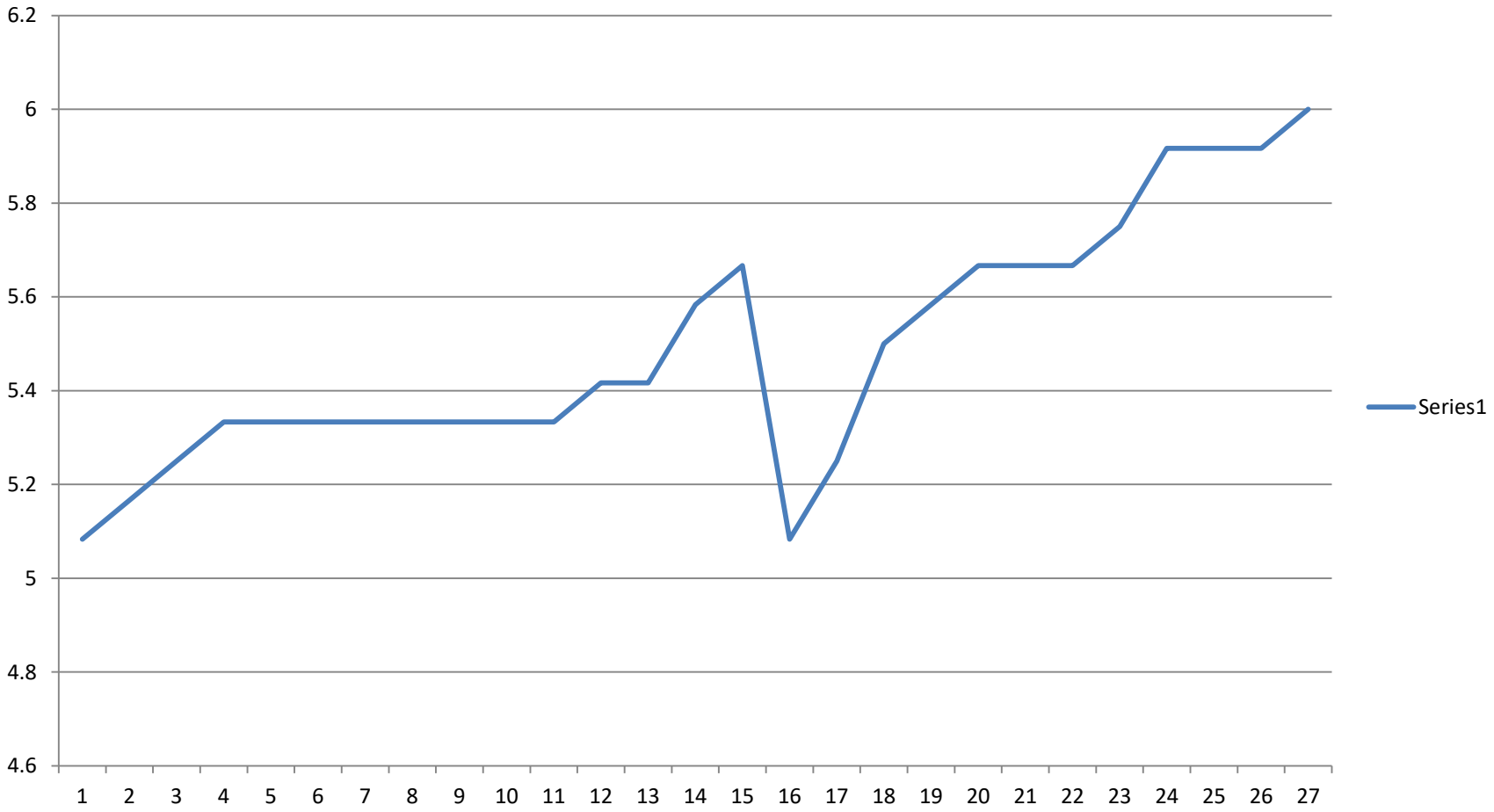
- Height and Weight of all of us in this class
- Add gender
- Min, max, average: descriptive statistics
- Is a plot meaningful? ‘Mode’ from the plot?
- Histogram of height

	A	B	C	D	E	F	G
1	Sl No	Ht_ft	Ht_In	Height	Weight	Gender	BMI
2	3	5	1	5.083333	40	F	16.66222
3	2	5	1	5.083333	45	F	18.745
4	1	5	1	5.083333	65	M	27.07611
5	4	5	2	5.166667	69	F	27.82264
6	6	5	3	5.25	54	F	21.08848
7	5	5	3	5.25	62	M	24.2127
8	9	5	4	5.333333	51	F	19.29936
9	12	5	4	5.333333	51	F	19.29936
10	7	5	4	5.333333	52	F	19.67777
11	14	5	4	5.333333	52	F	19.67777
12	8	5	4	5.333333	55	F	20.81303
13	10	5	4	5.333333	56	F	21.19145
14	11	5	4	5.333333	56	F	21.19145
15	13	5	4	5.333333	62	F	23.46196
16	16	5	5	5.416667	58	F	21.27815
17	15	5	5	5.416667	68	F	24.9468
18	17	5	6	5.5	60	M	21.3499
19	18	5	7	5.583333	60	M	20.71735
20	19	5	7	5.583333	65	F	22.4438
21	23	5	8	5.666667	58	F	19.44208
22	20	5	8	5.666667	60	M	20.1125
23	22	5	8	5.666667	62	M	20.78291
24	21	5	8	5.666667	69	M	23.12937
25	24	5	9	5.75	65	M	21.16156

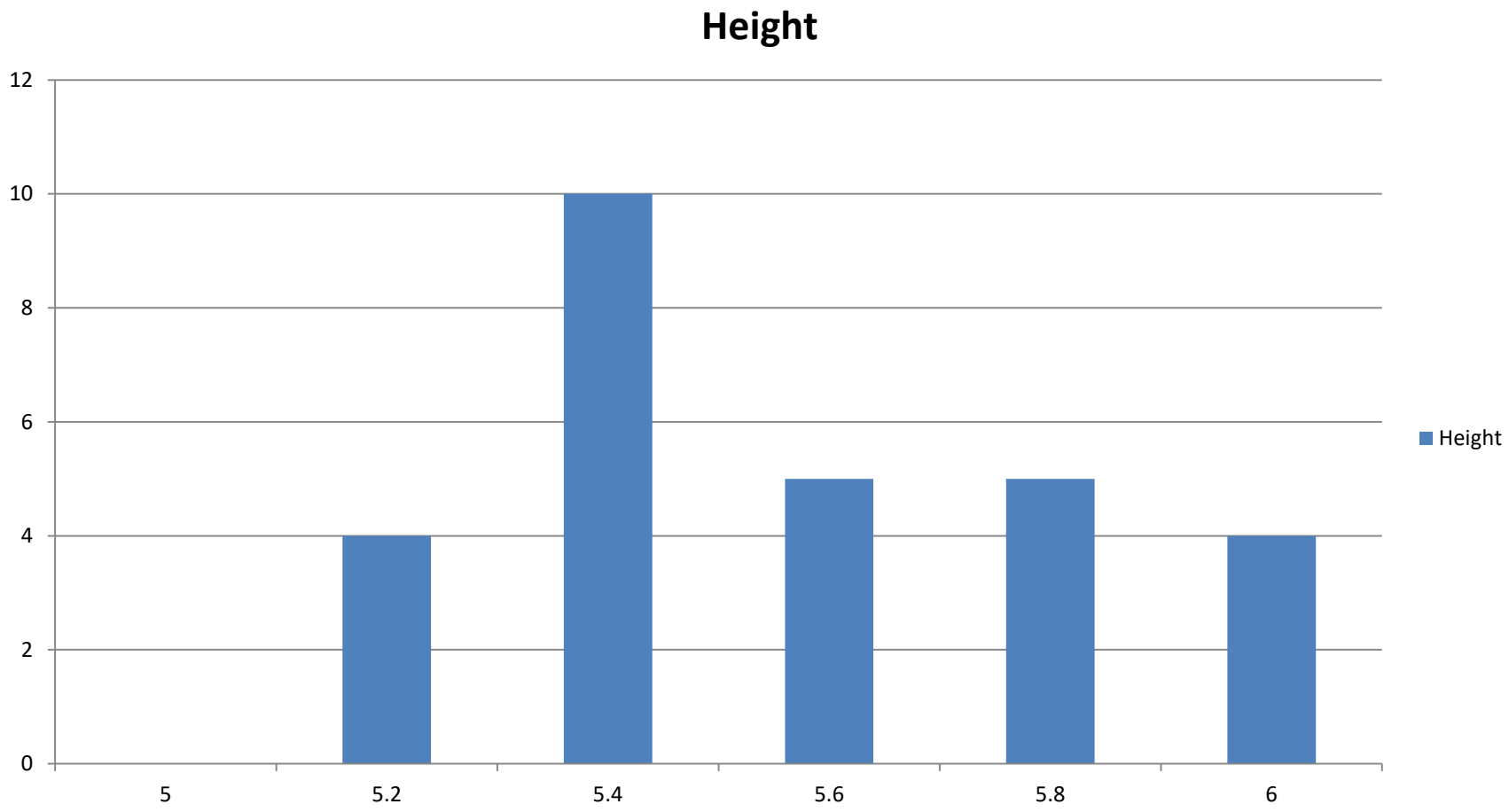
Descriptive Statistics

Number of data points	28
Max height	6
Min height	5.0833
Average height	5.4732
Max weight	75
Min weight	40
Average weight	60.0357
Male average height Female average height Male average weight Female average weight	5.6597 5.3333 65.75 55.75

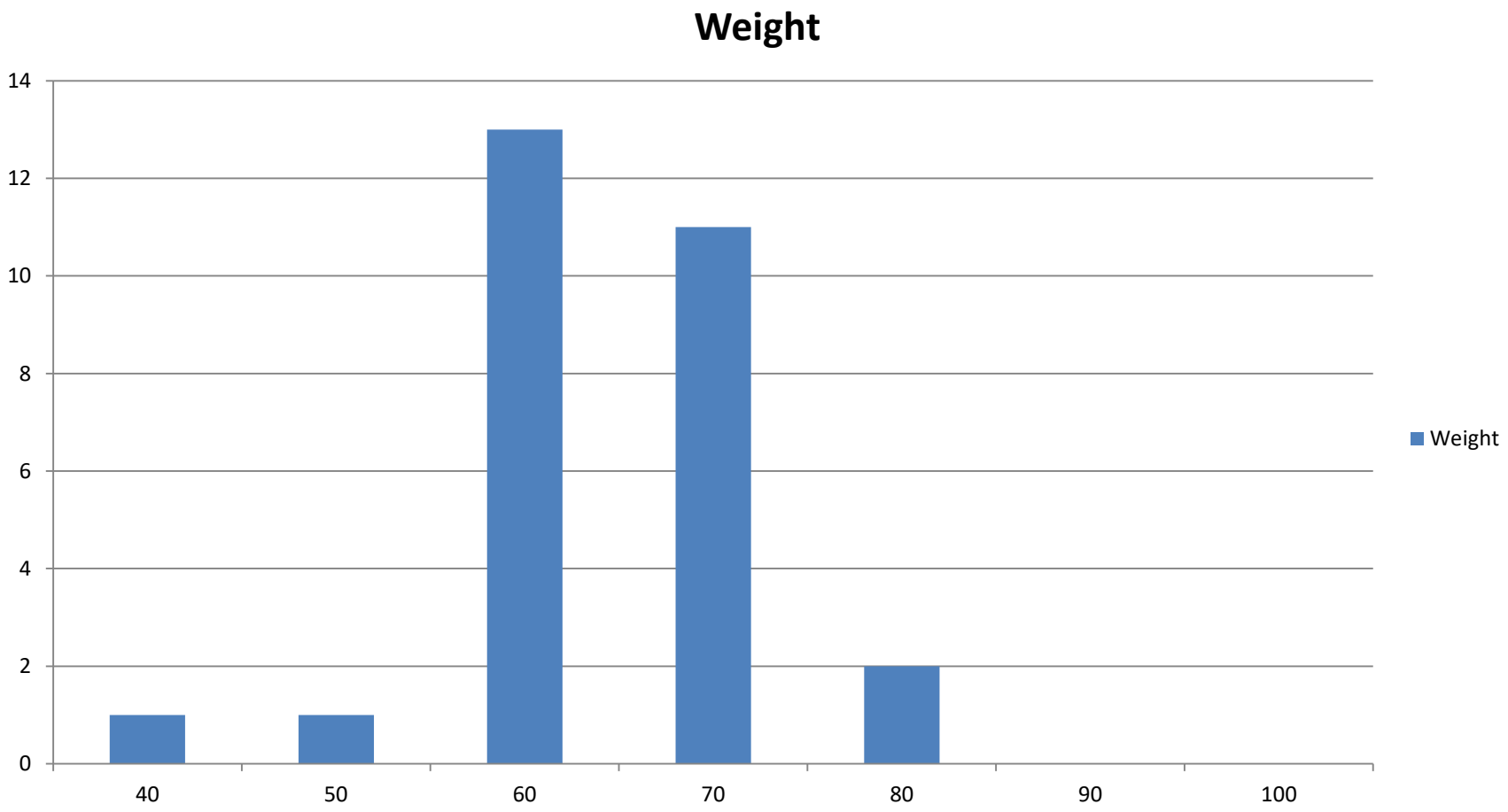
Height Plot (vs. Sl No!!)



Histogram for Height

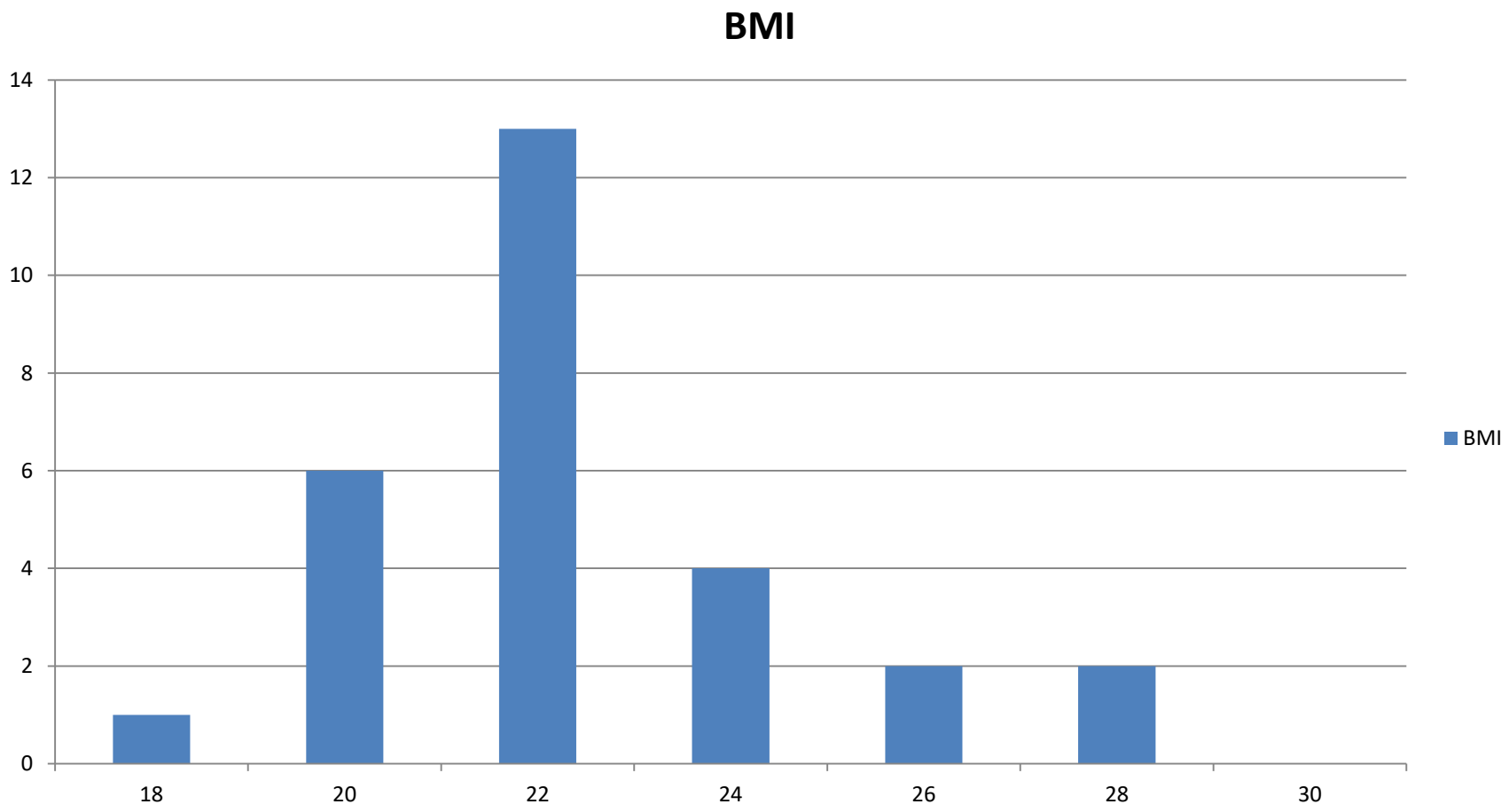


Histogram for Weight

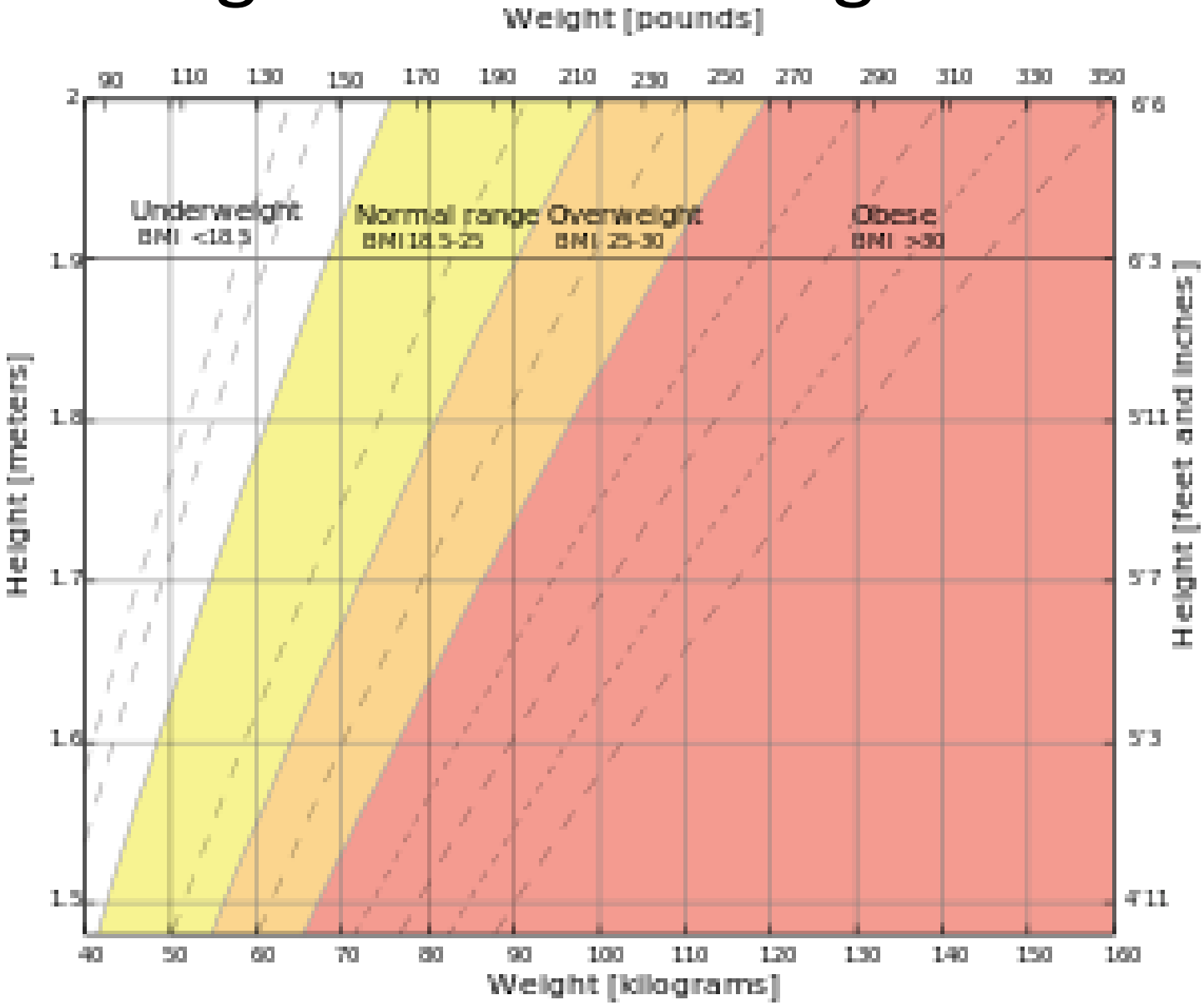


Histogram for BMI

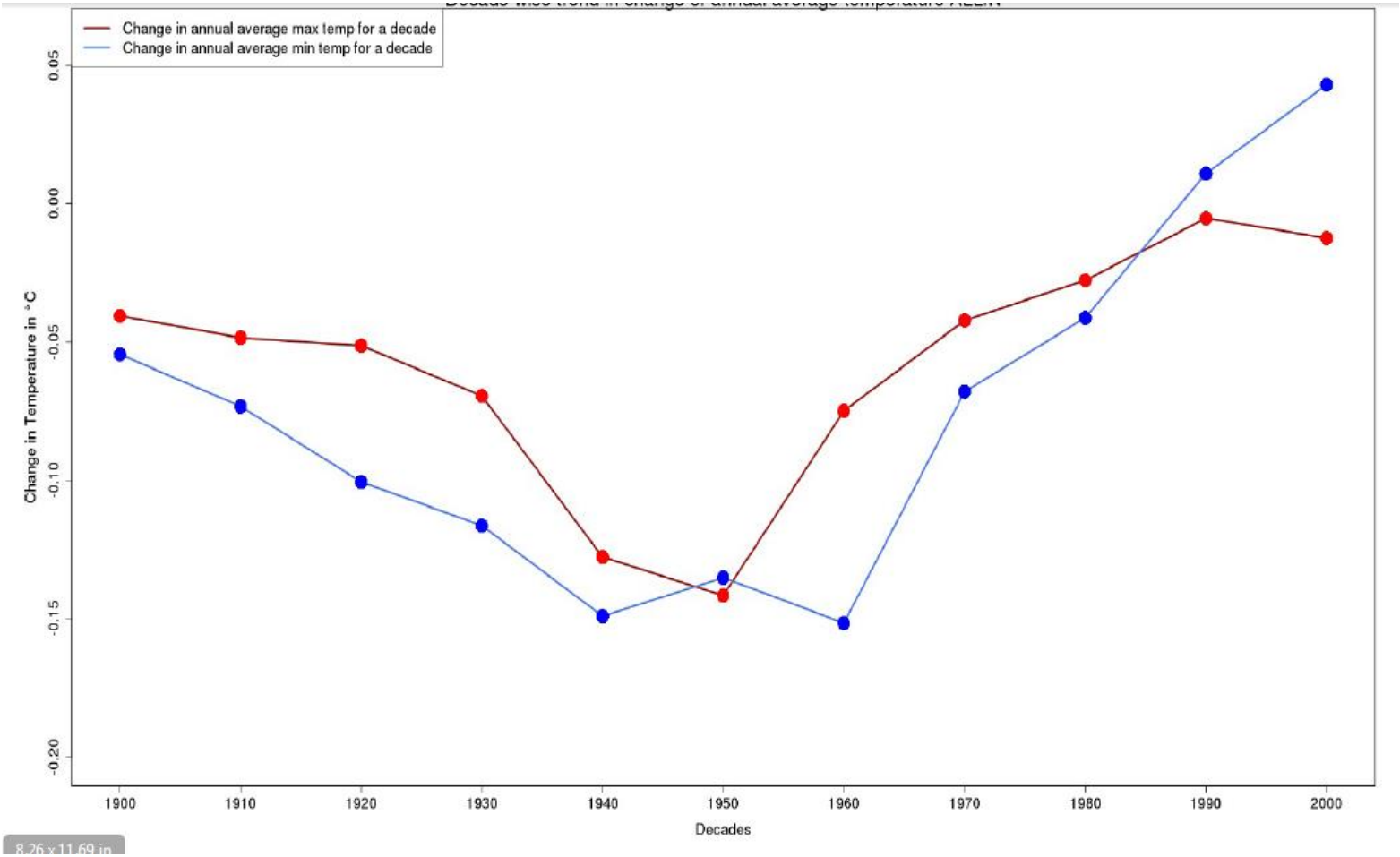
(See the shape!)



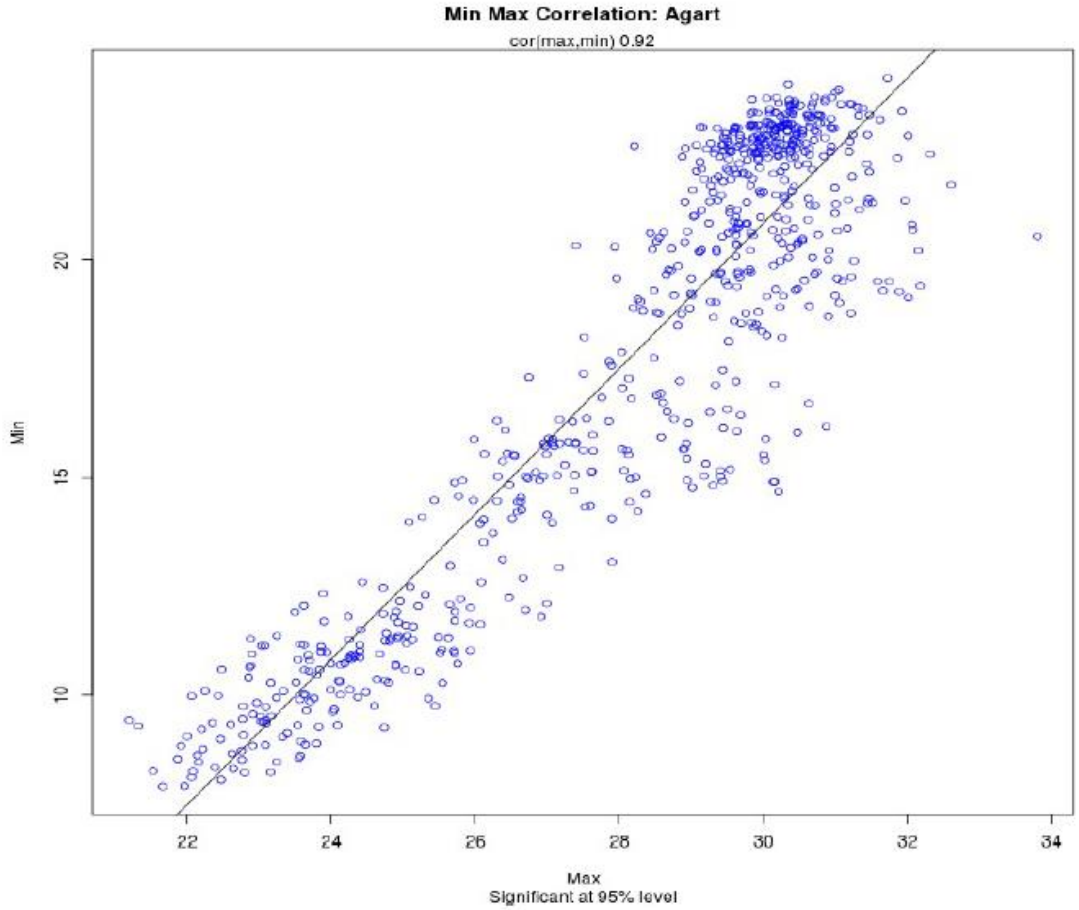
Knowledge of BMI: Making Inferences



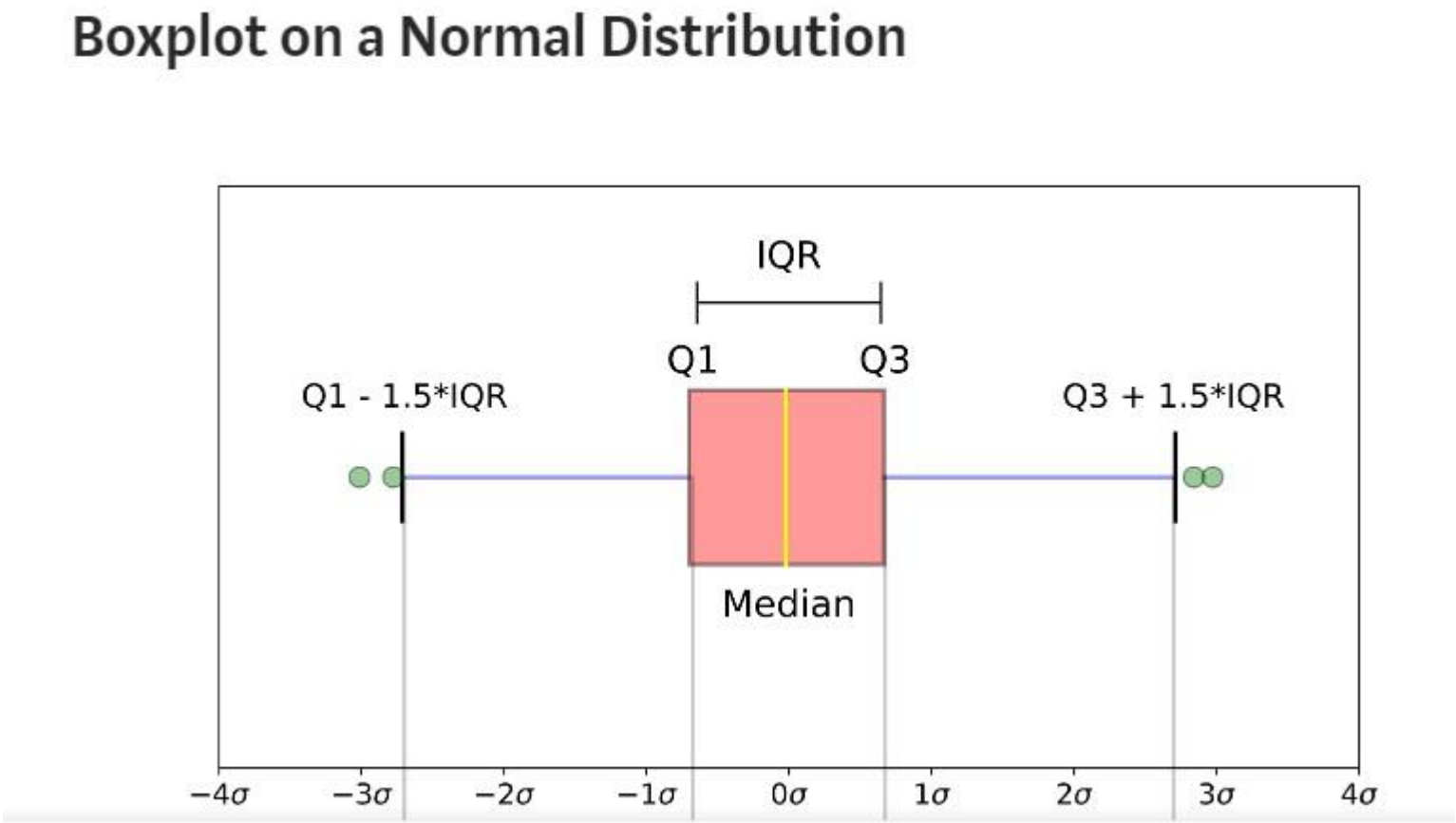
Example: Global Warming?



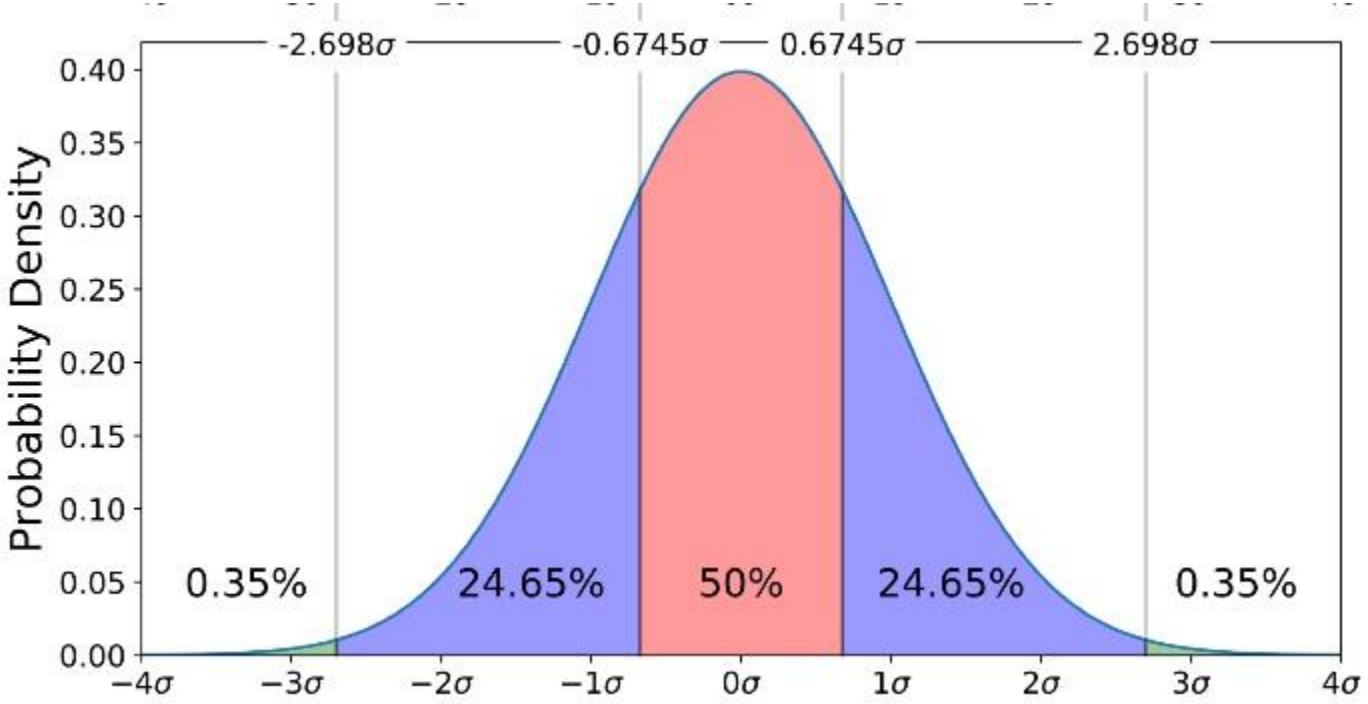
Temperature: Daily Min vs Max



Box Plot


Boxplot on a Normal Distribution
IQR
Q1
Q3
Q1 - 1.5*IQR
Q3 + 1.5*IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ

Normal Distribution



Data – the Basics

- Data sources
  - Measurement
  - Errors
- Authenticity, reliability, provenance
- Confidentiality, privacy, intellectual property, ..
- Security

Data Shape and Distribution

- Location and spread of data points
- Min, max, range
- Even spread or clusters?
- Symmetric distribution or long tail on one side
- Thin or heavy tails? Closeness to centre
- Cluster properties
- Outliers
- Sharp cutoffs, gaps, etc.

Plotting Data

(Refer textbook for details)

- Dot and jitter plots
- Histograms
- Kernel density estimates (KDE)
- Cumulative distribution function
- Rank-order plot and Lift charts
- Summary statistics
- Box-and-whisker plots

Review of Basic Statistics

- Sample, population and census
- Measures of Central Tendency
- Outliers
- Hypothesis Testing

Statistical Techniques

- Plotting …
- Frequency Distribution
- Crosstabulation
- Mean, median and mode
- Standard deviation, variance
- I.Q.V
- Chi-Square Test
- Confidence interval
- T-Test
- ANOVA
- Regression (I.V. and D.V.), reference groups, interaction effects, logistic regression, non-linear regression, etc.

Frequency Distribution Tables

- Value and Frequency
- Percentage frequency
- Cumulative percentage
- Grouped frequency distribution

Exhibit 2.2: A Completed Frequency Distribution

Frequency Distribution of Number of Children

No. of Kids	*Frequency*	*Percentage*	*Cumulative Percentage*
0	3	15%	15%
1	5	25%	40%
2	4	20%	60%
3	4	20%	80%
4	2	10%	90%
5	1	5%	95%
6	1	5%	100%
	n = 20		

Source: Hypothetical data.

With all this information, the viewer can answer a lot of questions really quickly. How many respondents have exactly 3 children? What percentage of the respondents has 3 children? What percentage has 4 children or fewer?

Exhibit 2.4: A Grouped Frequency Distribution

Frequency Distribution of Number of Children

No. of Kids	*Frequency*	*Percentage*	*Cumulative Percentage*
0–2	1	5%	5%
3–5	6	30%	35%
6–8	5	25%	60%
9–11	3	15%	75%
12–14	3	15%	90%
15–17	1	5%	95%
18–21	1	5%	100%
	n = 20		

Source: Hypothetical data.

The benefit of using a grouped frequency distribution is that it limits the number of categories. Had we given each number its own category, we would have had 22 categories, and a very large table. The downside of using a grouped frequency

S

HILDREN

ercent	Valid Percent	Cumulative Percent
27.0	27.0	27.0
15.7	15.8	42.8
25.4	25.5	68.3
16.4	16.4	84.7
8.6	8.6	93.3
3.1	3.1	96.4
1.8	1.8	98.3
1.0	1.0	99.2
.8	.8	100.0
.7	100.0	
.3		
0		

ice that the variable has a catch-
se 35 people have 8, 9, 10, 14,
not much to worry about. Next,
ta for this variable: 13 people

Exhibit 2.6: Dividing the Twelve Responses into Groups

Man: Yes, I own a dog.	*Woman:* Yes, I own a dog.
Man: Yes, I own a dog.	*Woman:* Yes, I own a dog.
Man: Yes, I own a dog.	*Woman:* Yes, I own a dog.
Man: No, I do not own a dog.	*Woman:* No, I do not own a dog.
Man: No, I do not own a dog.	*Woman:* No, I do not own a dog.
	Woman: No, I do not own a dog.
	Woman: No, I do not own a dog.

Next, I'll replace the words with numbers and add up how many of each group I have overall:

Exhibit 2.7: Adding Up the Various Groups

3 men own dogs	3 women own dogs	6 overall own dogs
		6 overall don't own dogs

dog.

dog.

wn a dog.

wn a dog.

wn a dog.

n a dog.

e above lists to
ners:

	Woman: No, I do not own a dog.
	Woman: No, I do not own a dog.

Next, I'll replace the words with numbers and add up how many of each group I have overall:

Exhibit 2.7: Adding Up the Various Groups

3 men own dogs	3 women own dogs	6 overall own dogs
2 men don't own dogs	4 women don't own dogs	6 overall don't own dogs
5 men overall	7 women overall	

Next, I'll remove some of the words and move some other words around:

Exhibit 2.8: Putting Categories above the Columns and by the Rows

	MEN	WOMEN	
OWN DOG	3	3	6
DON'T OWN DOG	2	4	6
	5	7	12

Crosstabulation

- Tables with 2 variables
- E.g., men and women owning dogs
- Crosstab with percentages
- E.g., spanking and race
- More than 2 variables
- E.g., gender, employment and happiness

Next, in order to compare men and women, I need percentages. Right now, if I used the frequencies, I might mistakenly say: oh, the same number of men and women own dogs, so they are equally likely to own a dog. However, there are more women than men in this sample, so this would be an inappropriate comparison. Three out of the 5 men, or 3/5, or 60% own dogs; 40% of the men don't. Three out of 7 of the women, or 3/7, or 43% own dogs; 57% of the women don't. Adding in these percentages—as well as percentages for the overall frequencies: 6/12 (50%) of the people are dog owners, 6/12 (50%) are not dog owners, 5/5 (100%) of the men are men, 7/7 (100%) of the women are women—to the crosstab, as well as a title and a data source, I now have:

Exhibit 2.9: The Completed Crosstab

Crosstabulation of Dog Ownership by Sex

	MEN	WOMEN	
OWN DOG	3	3	6
	60%	43%	50%
DON'T OWN DOG	2	4	6
	40%	57%	50%
	5	7	12
	100%	100%	100%

Source: Hypothetical data.

There. Done. How do I know I'm done? Well, besides carrying out all the steps I think are necessary to create a crosstab, I can answer yes to this question: does the table tell the story? Yes, it does: a higher proportion of men than women have dogs. I have a couple of simple terms to introduce: the numbers within the table are called the **cells** and the numbers outside the boxes are called, appropriately enough, the **marginals**. Also, there are a few rules I want to point out in relation to the above cro[illegible]. The dependent variable in this situation is dog ownership: we are trying t[illegible] ownership is dependent on sex. That mak[illegible]

significant part of one's identity. For both of these reasons, are those who are unemployed more likely to be unhappy? The GSS not only asks about work status, but also asks a question called HAPPY: "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?" Here's a crosstab using GSS 2006 data that allows us to see the relationship between these two variables:

Exhibit 2.17: Crosstab of Happiness by Employment Status

GENERAL HAPPINESS * LABOR FORCE STATUS Crosstabulation

		LABOR FORCE STATUS		
		WORKING FULLTIME	UNEMPL, LAID OFF	Total
GENERAL HAPPINESS	VERY HAPPY	452	13	465
		30.1%	13.7%	29.1%
	PRETTY HAPPY	896	51	947
		59.7%	53.7%	59.3%
	NOT TOO HAPPY	154	31	185
		10.3%	32.6%	11.6%
	Total	1502	95	1597
		100.0%	100.0%	100.0%

Unemployed people are over three times more likely to say that they are not too happy. But then I began to wonder: is holding a job a bigger part of a man's identity than a woman's identity? Might men, who often construct their masculinity around their economic productivity, be more affected by job loss? I elaborated the crosstab using

Exhibit 4.1: A Hypothetical Crosstab of Gun Ownership by Sex

	MEN	WOMEN	
OWN GUN	100 *100%*	0 *0%*	100 *33%*
NO GUN	0 *0%*	200 *100%*	200 *66%*
	100	200	300

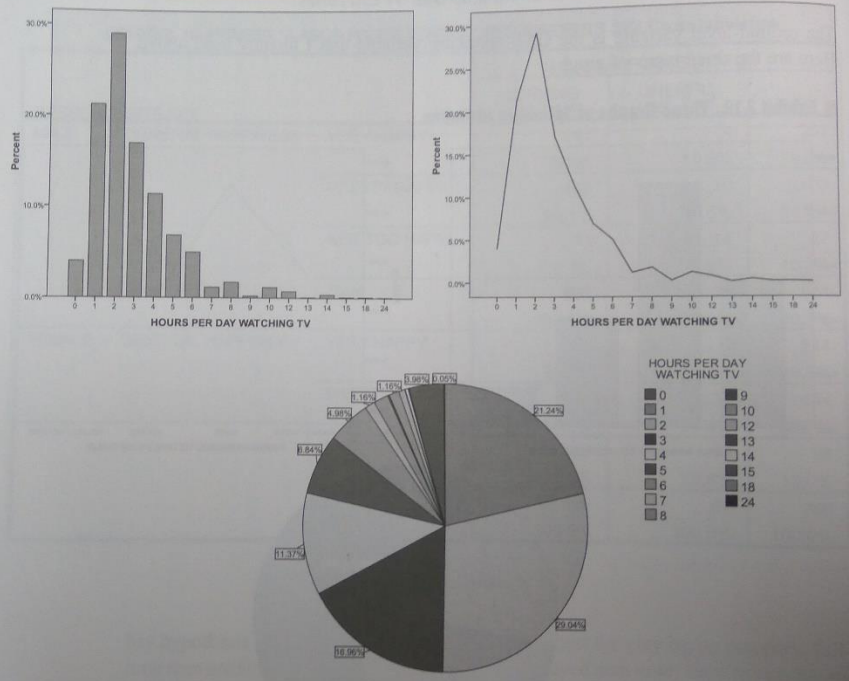
Source: Hypothetical data.

Nick and Sam both examine these results. Nick says: "Based on these sample results, there's no way that there's a relationship between sex and guns in the population!" Sam says: "Based on these sample results, there's definitely a relationship between sex and guns in the population!" What does your gut say? It sides with Sam, right? The differences in the crosstab based on sample results are so very clear that you feel completely confident saying that in the population there is a relationship between sex and guns.

Plotting

- Bar and line graphs
- Pie charts
- E.g., TV watching hours per day
- Stacked bar graph
- 3-D bar graph
- Tufte’s lie factor

Exhibit 2.20: Three Graphs of Television Watching

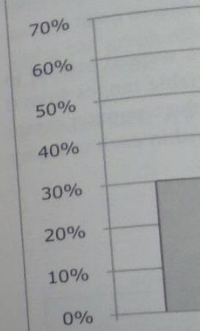


Now it is the pie graph that is clearly out. Too many categories mean too many small slices. The bar graph and the line graph are fairly similar in their ability to get across the shape of how the data are distributed.

GRAPHS WITH TWO VARIABLES

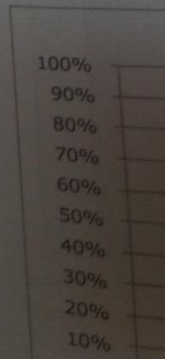
If you want to use a graph to illustrate the relationship between two variables, there are a number of choices. First, there are various types of bar graphs. Let's say you wanted to create a graph to represent the relationship between work sta

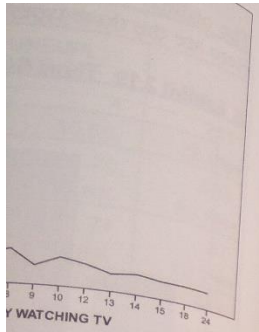
Exhibit 2.21: A Clus



This graph allow
non-working: m
unemployed are

Exhibit 2.22:





OURS PER DAY
WATCHING TV

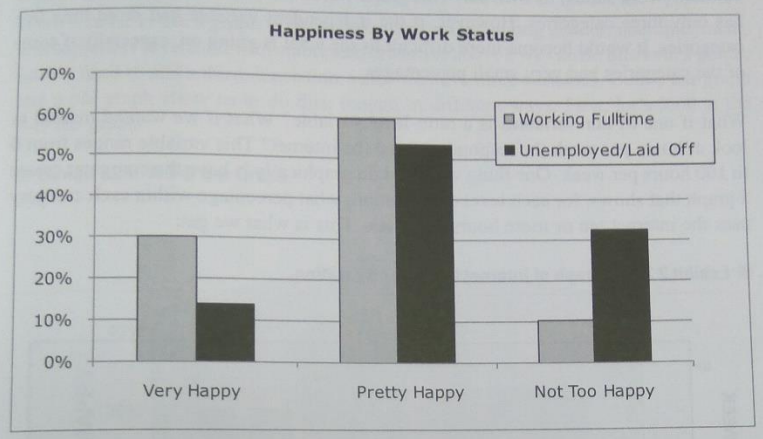
0
1
2
3
4
5

9
10
12
13
14
15
18
24

an too many small
bility to get across

o variables, there
hs. Let's say we
work status and
hapter. Here is a

Exhibit 2.21: A Clustered Bar Graph



This graph allows the eye to quickly take in the differences between the working and non-working: more of the working are very happy, while a larger proportion of the unemployed are not too happy. Another bar graph option is the stacked bar graph:

Exhibit 2.22: A Stacked Bar Graph

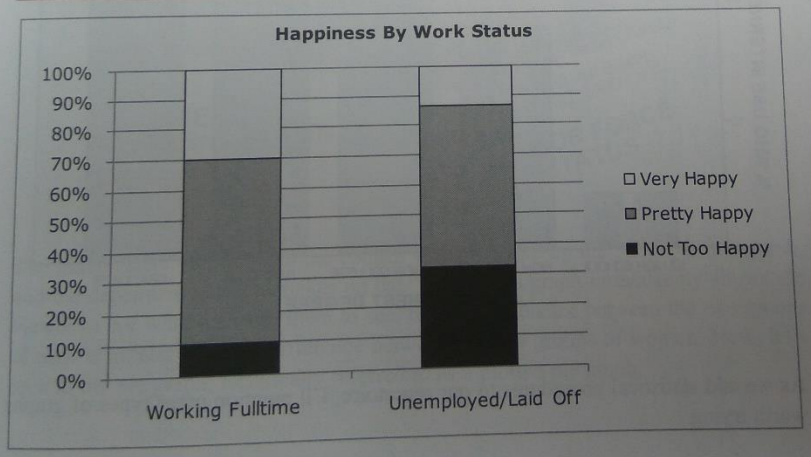
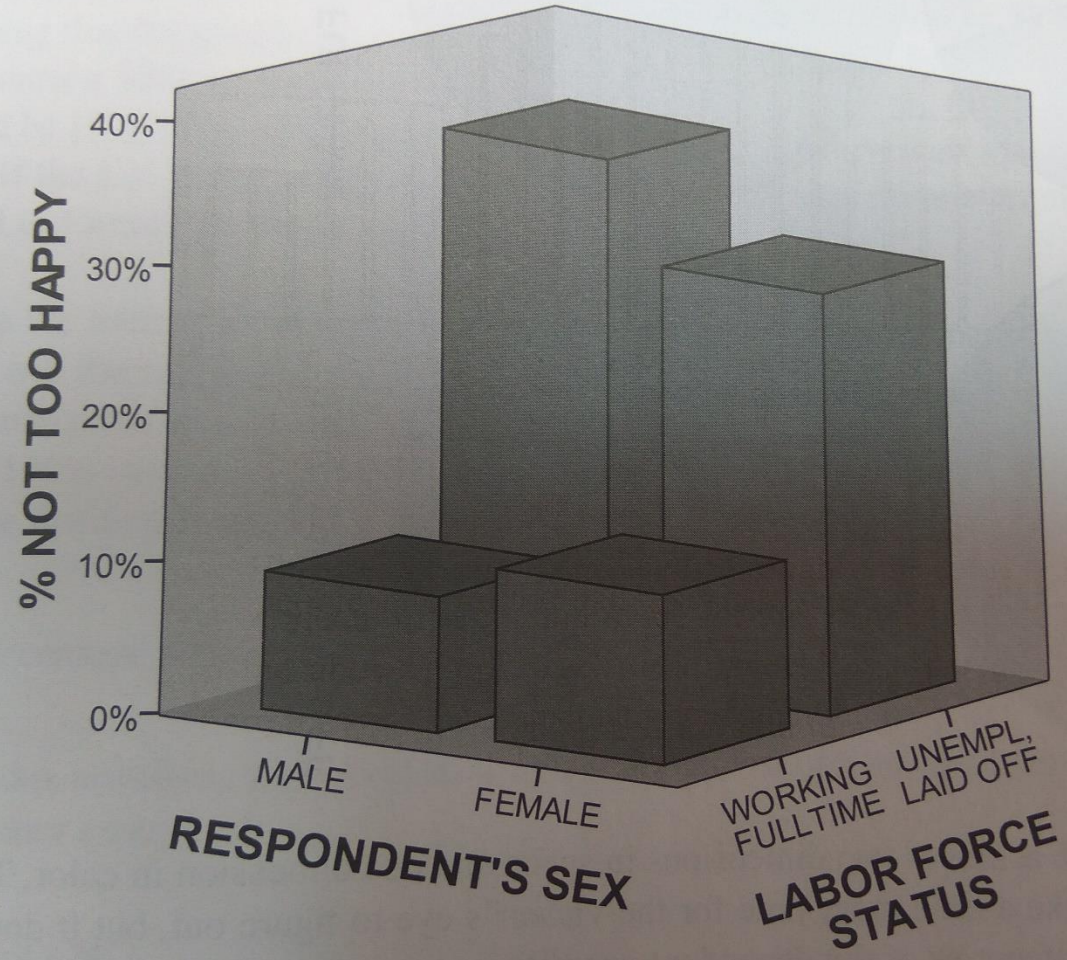


Exhibit 2.24: A 3-D Bar Graph



The independent variable and the control variable make up the "floor" of the graph,

Histograms

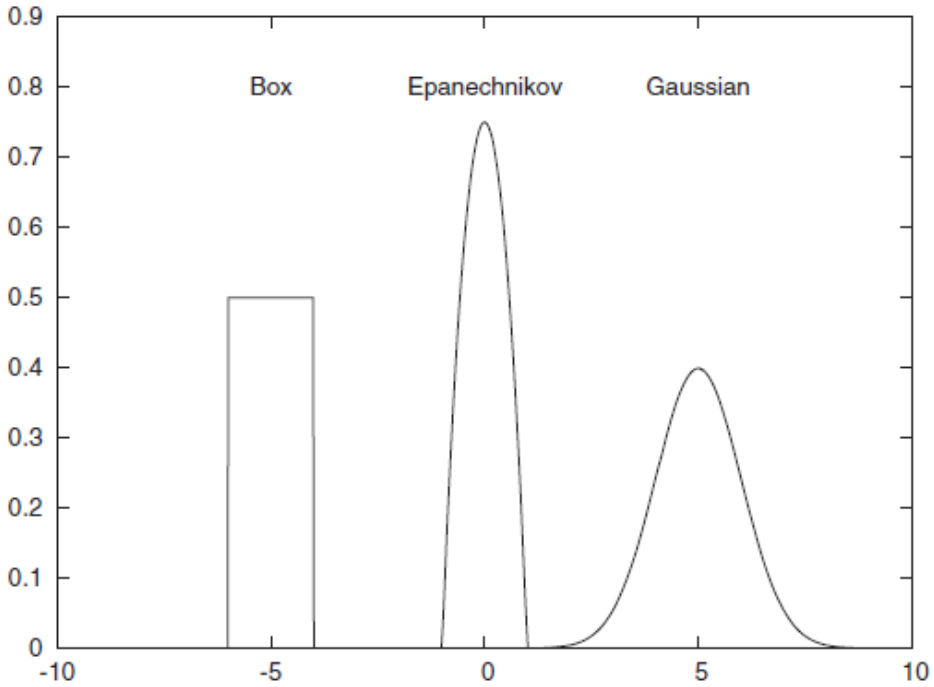
- Raw or normalized
- Equal or unequal width bins
- Disadvantages:
  - Loss of actual information
  - Not unique (different anchoring points for bins)
  - Ragged, not smooth
  - Can’t handle outliers gracefully

Kernel functions

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$ box or boxcar kernel

$$K(x) = \begin{cases} \frac{3}{4}\left(1 - x^2\right) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$ Epanechnikov kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$ Gaussian kernel



Continuous vs. Discrete Scales

- Continuous – real numbers
- Discrete – integers or qualitative/enumerated values

Steven's Levels of Measurement

- Categorical or nominal scales
  - E.g., planet, star, satellite, asteroid
  - Mutually exclusive and Collectively exhaustive
- Ordinal scales (higher but not by how much)
  - E.g., low, medium, high or A/B/C/D/F grades
- Interval scales (comparable but not ratio)
  - E.g., Degrees Celsius
- Ratio (with well-defined zero)
  - E.g., Height, Money, etc.

Attributes

- Natural attributes
- Constructed (derived ?) attributes
- Proxy attributes
  - E.g., GDP

Data Range

- Fixed or dynamic
- Outlier

Measures of Central Tendency

- Mean, median, mode
- Standard deviation and variance
- Shapes of distributions
- Chebyshev's Theorem

do it intuitively first.

MEASURES OF THE CENTER AND LEVELS OF MEASUREMENT

Though I think that ideally one should present the mean, median, and mode together for comparison's sake, sometimes we are limited by the levels at which the variables are measured. Remember that variables can be measured at the nominal, ordinal, or ratio levels. Here are the measures of center that we can use with each level of measurement:

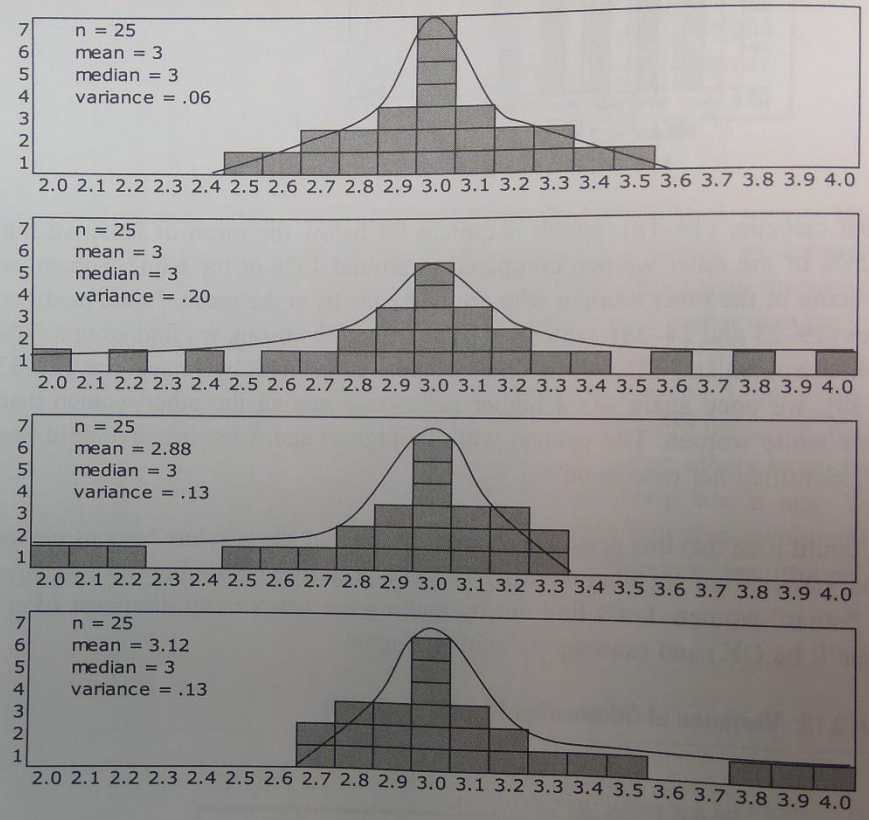
Exhibit 3.6: Levels of Measurement and Measures of the Center

	Nominal	*Ordinal*	*Ratio*
Mode	Yes	Yes	Yes
Median	No	Yes	Yes
Mean	No	No	Yes

If the variable is measured at the nominal level, the only center we can find is the mode. For example, if we were examining the types of schools students attend (public, private secular, private religious), the type of school with the highest frequency of children would be the mode. If the variable is measured at the ordinal level, there is a mode, but we can also find the median. This is because the median requires ordered data, and we can order the categories of an ordinal-level variable. For example, if we are using an ordinal-level measure of education (less than high school, high school, some college, college degree, graduate degree), we would be able to say, hypothetically, that the modal respondent graduated from high school and the median respondent had some college. If we had a ratio-level variable, then we could find all three measures of the center: the mode, the median, and the mean. For example, if we were measuring education in years, we might find that the mode was 12 years, the

It can be helpful to look at descriptive statistics in relation to the shape of the data's distribution. In Exhibit 3.16 have four hypothetical classes, each with 25 students:

Exhibit 3.16 Four Distributions



Carefully examine the descriptive statistics of each distribution. In the first distribution, the mean and median are the same, and the distribution has a symmetrical shape. In the second distribution, the mean and median are again equal, but there is more variation than in the first distribution. This is illustrated by the higher variance, but also by the shape of the distribution, which, though symmetrical, is flatter and has

Elementary Analysis

- Mean (average)
- Median (middle)
- Mode (frequent)
- Frequency distribution
- Histogram
- Various Distributions

Variance

- How scattered around the mean?
- Dispersion
- $s = \sum^{n}_{1} (x_i - \mu)^2 / n$
- Second central moment
- (sometimes n-1 for technical reasons about bias)
- Variance = average of square of x_i minus square of μ

Standard Deviation

- σ = square root of variance
- Std dev = typical distance from mean only if distribution is bell-curved

- Mean absolute deviation (an alternative)

Quintiles and Percentiles

- Generalization of median
- x^{th} percentile = value such that x percent of all values are less than or equal to the value
- Quantile is percentile in decimals: [0.0….1.0]
- Inter-Quartile Range (IQR) = distance between 75^{th} and 25^{th} percentile.
- Used when distribution is not normal
- More “difficult” to compute

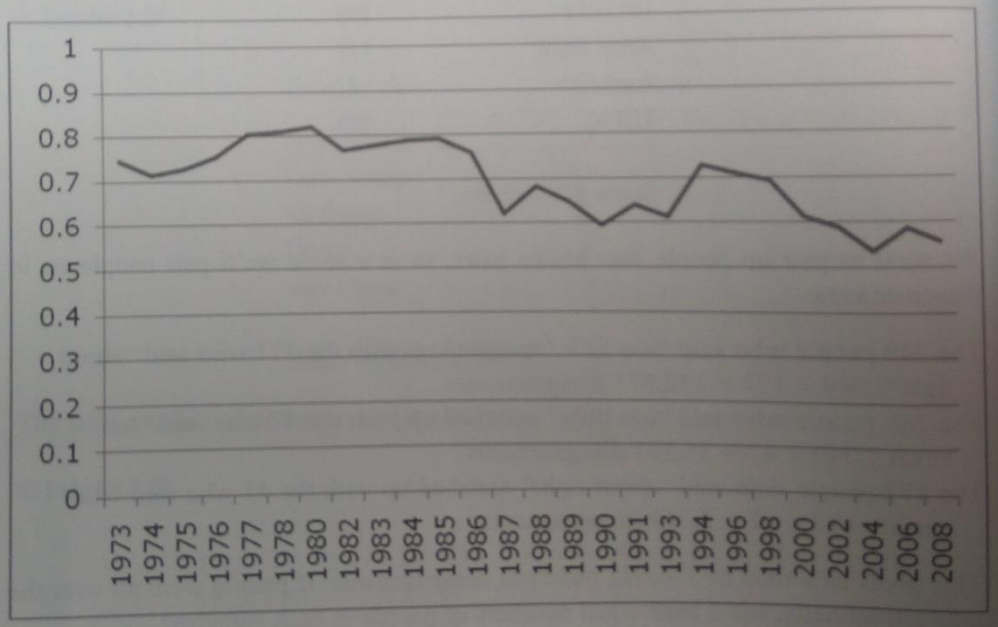
$324.33^2 + 324.33^2 + 324.33^2 = 315{,}576$

This gives us an IQV of 176,095/315,576 = 0.56. Using the formulae we developed above, we would get the same result:

$$\frac{\Sigma f_i f_j}{(n/c)^2 \times ((c \times (c-1))/2)} = \frac{(749 \times 177) + (749 \times 47) + (177 \times 47)}{(973/3)^2 \times ((3 \times (3-1))/2)}$$

It's sometimes hard to attach meaning to the IQV, and a comparative perspective might help. Therefore, I looked at every time this question has been asked by the GSS, and calculated the IQV for each year. Here is a graph of the progression of the IQV over time:

Exhibit 3.31: IQV of Health Care Responses, 1973–2008



Though it's a rather bumpy ride, the overall trend for the IQV is a downward one. Where there used to be a difference of opinion, consensus is forming that *something* needs to be done about our health care system. Notice that consensus is synonymous with a lack of diversity of opinion. Look at the two frequency distributions for the years with the highest and lowest IQVs:

Let us compute for our data..

- Try this for our height-weight data set!

A little bit of probability

- Probability is a ratio of two numbers
- p = ½ for coin toss
- p = 1/6 for fair die
- p = 4 / 52 for getting an A in cards
- p = desired / total events

Probability: Bernoulli Trials

- $P(k, N; p) = {}^{N}C_{k}\, p^{k}\, (1 - p)^{(N - k)}$
- Binomial Distribution
- N = ? How many trials or samples?
- N >= 30
- If N = 100, expected number of heads = 50
- Standard deviation = 5 (i.e., ½ sqrt(100))

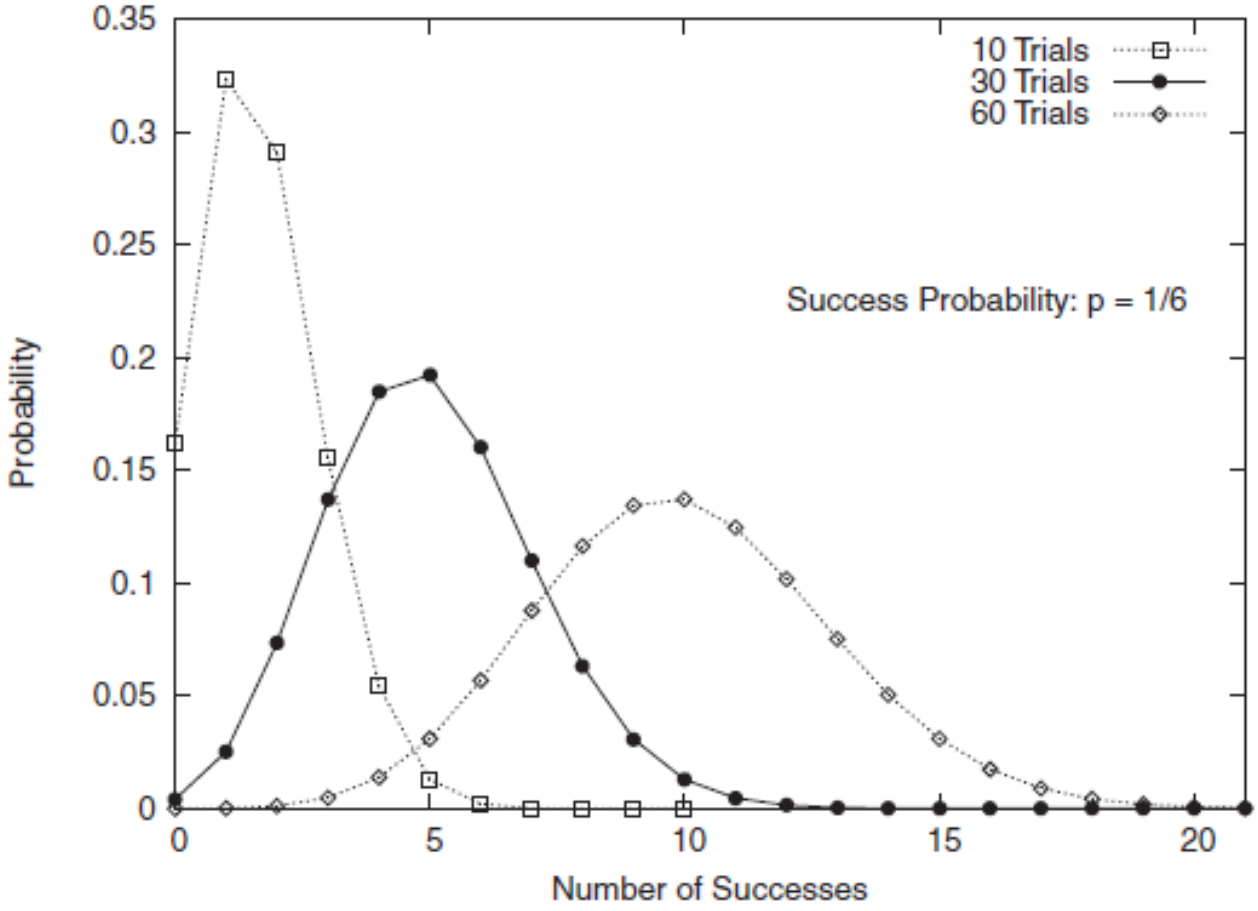


FIGURE 9-1. The Binomial distribution: the probability of obtaining k Successes in N trials with Success probability p.

Gaussian Distribution

- The famous “bell curve”
- Guassian or Normal probability density
  - Normal: $\mu = 0$; $\sigma = 1$
- $$p(x;\, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$
- Why: it arises naturally whenever we take averages of almost anything
- The Central Limit Theorem…

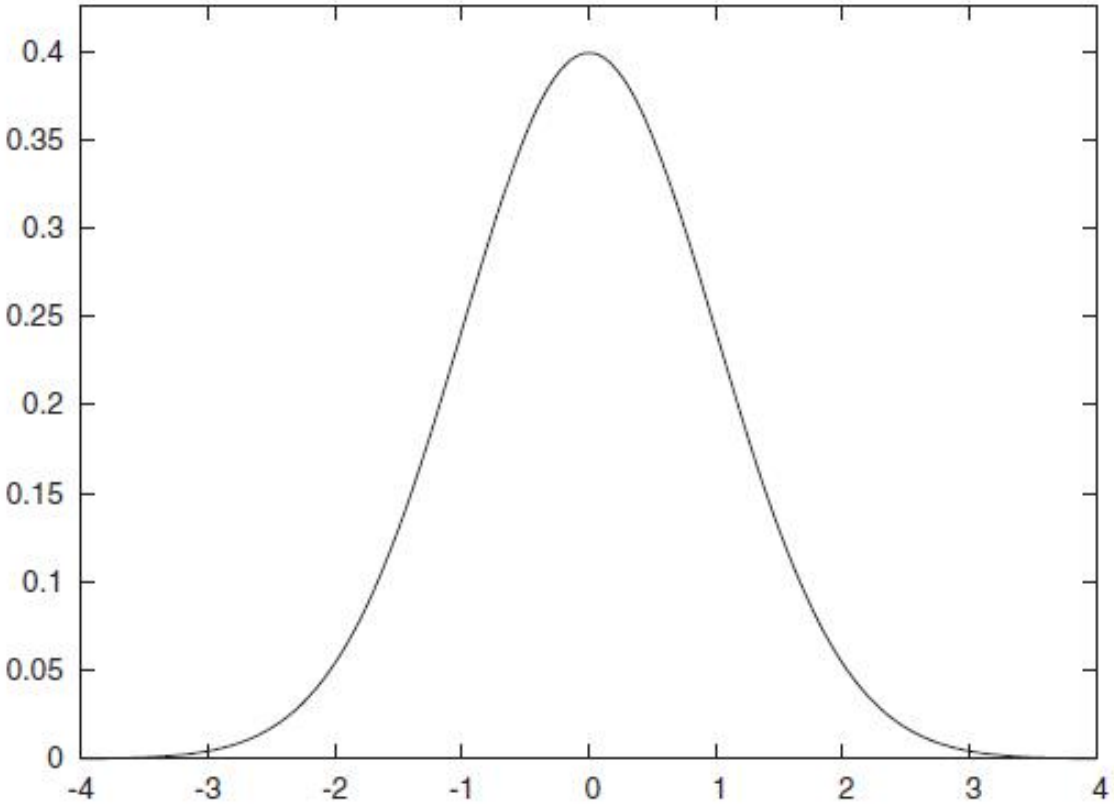


FIGURE 9-2. The Gaussian probability density.

Normal Distribution

- Make mean = 0
- Make standard deviation = 1

The Central Limit Theorem

- We know about the distribution of data based on the distribution of the sum of random data samples
- Theorem: *The sums of random quantities are distributed according to a Gaussian distribution*
- Sums are smooth with a central peak because the highs and the lows cancel out on the average

The Central Limit Theorem (2)

- $\{x_i\}$ sample of size n
- Mutually independent data points
- Drawn from a common distribution
- μ and σ are finite
- Sample average is distributed as a Gaussian distribution with mean μ and std deviation σ /sqrt(n)
- Better approximation with higher n
- Reduction in scatter is in terms of sqrt(n)

Example

- Fair die: 1 to 6
- $\mu = 3.5$
- $\sigma = \text{sqrt}((6^2 - 1)/12) \sim= 1.71$
- 1000 repeats of n throws of the die
- Increase n and plot histograms
- See Fig. 9.3

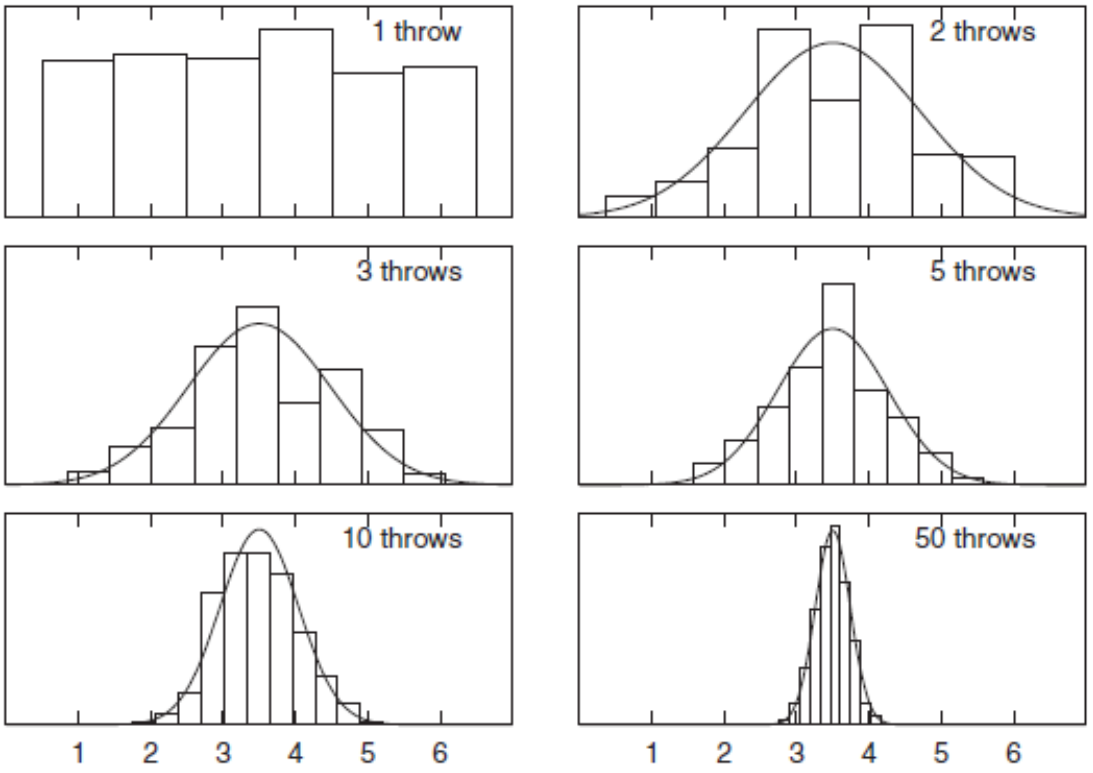


FIGURE 9-3. The Central Limit Theorem in action. Distribution of the average number of points when throwing a fair die several times. The boxes show the histogram of the value obtained; the line shows the distribution according to the Central Limit Theorem.

Caution…

- Not everything is normal
- Does not apply to actual data
- Only to sums or averages
- Power-law distributions are not covered!

Applying Statistics

- Data collection
- Design of experiments
- Parameter estimation
  - Point estimation
  - Interval estimation
- Hypothesis testing
  - Is there a desired effect at all?

Error

- Random error
- Systematic error
- Standard error = σ of an estimated quantity
- If normal distribution, 68% will be within +/- σ of the μ
- σ / sqrt(n) is the standard error of the mean

Standard Error

- Average distance between sample means and population mean
- Std error = std deviation / sqrt(n – 1)
- Thus, n should not be too small

Distribution of Variances: Chi-square

- Distribution of sum of squares of independent random variables
- n is the degrees of freedom

Student-T Distribution

- Ratio T of normally distributed variable Z and chi-square distributed random variable U
- $T = Z / \sqrt{U / n}$
- This is the distribution of the average if the variance is not known
- Bell curve with fatter tails

Fisher's F Distribution

- Ratio of two chi-square random variables
- To compare two variances against each other

Classical Statistics

- A population vs. a sample
- Conclusion about population from sample
  - Statistical inference
- Quantifying “more or less”
- Parameter estimation
  - Assume a distribution (e.g., Gaussian)
  - Point or interval estimation
- Hypothesis testing

Hypothesis Testing

- Is there an effect?
- Is the difference statistically significant?
- Null and alternate hypotheses
- Reject or fail to reject the null hypothesis
- Significance level: 5%

Confidence Intervals

- n observations
- μ / s^2 is t-distributed
- Find interval that has 95% probability of containing true value
- Fig. 10.1

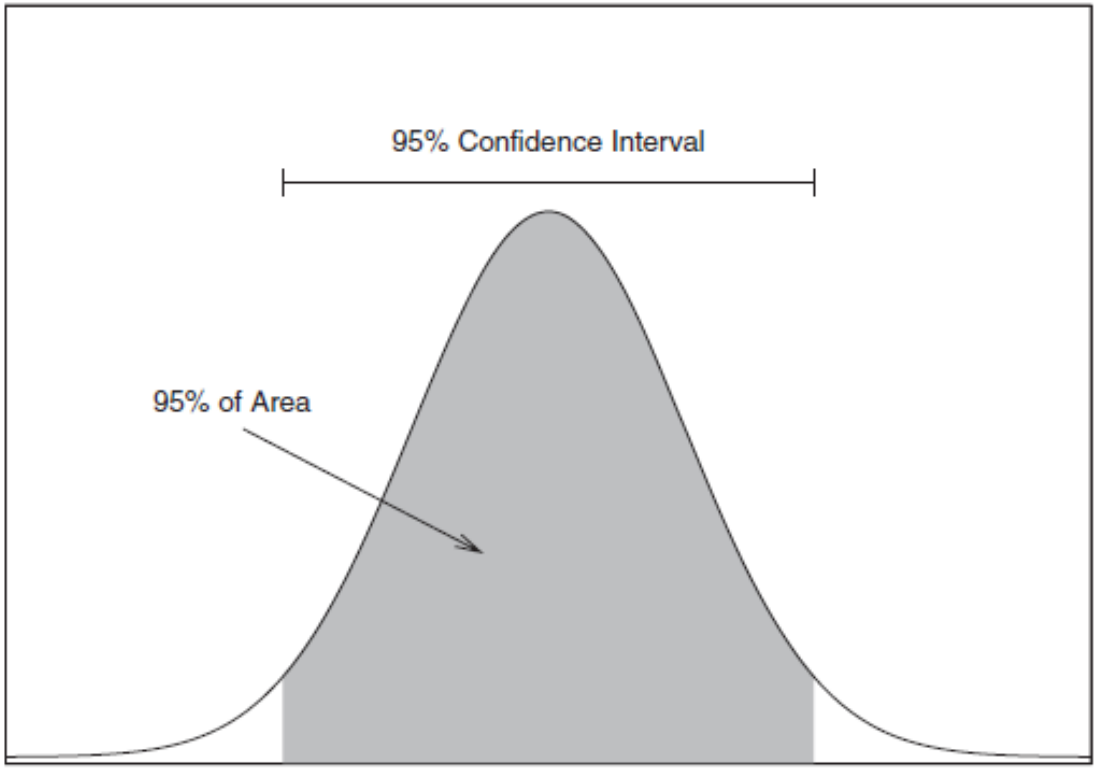


FIGURE 10-1. The shaded area contains 95 percent of the area under the curve; the boundaries of the shaded region are the bounds on the 95 percent confidence interval.

Chebyshev's Theorem

If data are distributed in a relatively normal way (with a few cases at each end and most of the cases in the middle, or in other words the classic "bell curve"), then we know that:

- approximately 68% of the cases will fall within one standard deviation of the mean;
- approximately 95% of the cases will fall within two standard deviations of the mean;
- almost all of the cases will fall within three standard deviations of the mean.

Hypothesis Testing

- Use sampling distributions to calculate p-value
- Distinguish significant and not significant outcomes
- Probability of obtaining a value as (or more) extreme than the one actually observed (assuming that the null hypothesis is true).
- Small p → evidence against null hypothesis
- p != probability of null hypothesis being true

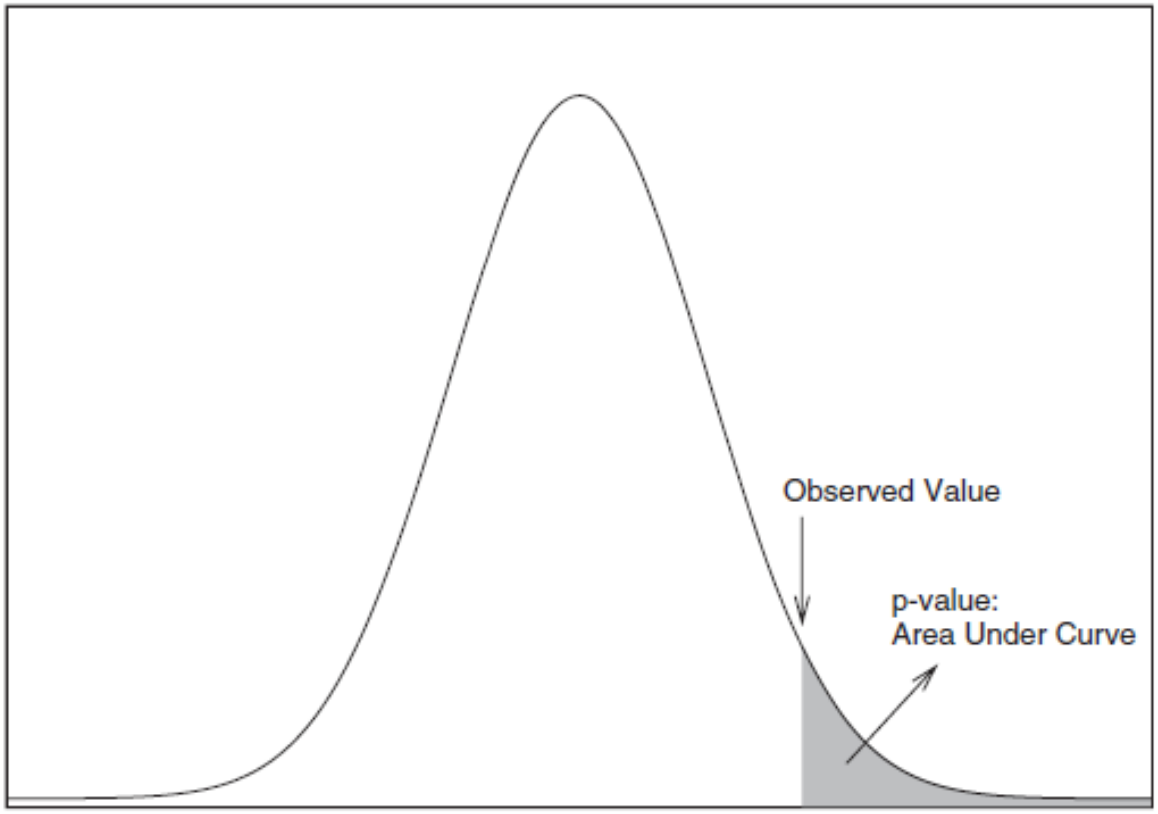


FIGURE 10-2. The p-value is the probability of observing a value as large or larger than the one actually observed if the null hypothesis is true.

Design of Experiments

- Randomization: no bias in data collection
- Replication: repeat to reduce variability
- Blocking: group into blocks of samples
- Factorization: interaction effects

Chi-Square Test

- Is there a relationship in the population?
- Compare observed frequencies to expected

Type 1 and Type 2 Errors

- Type 1 error (false positive): Saying there is a relationship when it is not there
- Type 2 error (false negative): Saying there is no relationship when in fact there is
- p is the probability of making a Type 1 error
- $p < 0.05$ Statistical significance

Example

- 100 men and 200 women
- 58 men are obese and 86 women obese
- Expected percentage of obesity is 50%
- Is gender related to obesity?
- $Chi^2 = 1/fe \sum(fo - fe)^2$
- $=1/50((58–50)^2 + (42–50)^2 + (43–50)^2 + (57–50)^2)$
- $= 1/50\ (64 + 64 + 49 + 49) = 226 / 50$
- $= 4.52$

Chi-Square Table (df = 1)

P-value	Chi-Square Value
0.25	1.32
0.1	2.71
0.05	**3.84**
0.025	5.02
0.010	6.63
0.005	7.88
0.001	10.83

T - Test

- Difference of means test
- t-value = # std errors away from the center of the sampling distribution of sample mean differences
- t = (X1bar – X2bar) / sqrt(S_1^2/n1 + S_2^2/n2)
- E.g., Housewives: mean TV watching = 4.13 (std dev = 2.77, n = 15)
- Working women: mean 2.30 (std dev = 1.63, n = 20)
- t = 2.28 df = 33…. $p < 0.05$

ANOVA and F-Test

- Analysis of variance
- Summarizing collected data
- Compare variance within a group against variances across groups
  - WGSS: within groups sum of squares
  - BGSS: between groups sum of squares
  - F = (BGSS / (#groups – 1)) / (WGSS / (#cases - #groups))
- Are the cross-group variances significant?

When to Use Which Test?

- If both are nominal or ordinal variables
  - Chi-Square Test
- If one ratio variable and one nominal or ordinal
  - If 2 groups: T-test
  - If > 2 groups: ANOVA

ifference among the
this should already
ull and alternative

	Sig.
9	.130

t the population
null hypothesis

Exhibit 6.10: Choosing a Statistical Procedure

	One Nominal or Ordinal Variable and One Nominal or Ordinal Variable	One Nominal or Ordinal Variable and One Ratio
Nominal or Ordinal Variable Has Two Groups	Chi-Square Test	T-Test
Nominal or Ordinal Variable Has More than Two Groups	Chi-Square Test	ANOVA

So, before you run a statistical test, carefully consider what your variables are.

Bayesian Analysis

- Probability as degree of ignorance
  - Not as limiting frequency
- Conditional probability
  - P(A / B) = P(A and B) / P(B)
- Bayes theorem: P(A/B) = P(B/A) P(A) / P(B)
  - P(A/B) – posterior probability
  - P(B/A) – likelihood function
  - P(A) – prior probability
- Diagnostics, troubleshooting
- Getting causes from symptoms

Co-variance

- Measure of how two variables vary together
- Sum of product of the x and y differences from the corresponding means

Linear Regression

- $y = mx + c$
- $m = \sum((X - Xbar)(Y - Ybar)) / \sum((X - Xbar)^2)$
- Then solve for c
- E.g., **(TRY THIS!)**

x	y
0	2
1	6
2	7
3	12
5	16
10	34
20	60

Correlation Coefficient

- $r = \sum((X - Xbar)(Y - Ybar))/sqrt(\sum(X\text{-}Xbar)^2 \sum(Y–Ybar)^2)$
- Ratio of co-variance and the square root of the product of the summed squared distances of each variables
- Effect of outliers

r^2: Coefficient of Determination

- r^2 is the percentage of variation that is explained
- Note that r should be rather high for r^2 to be high enough

r	r^2
1.0	1.0
0.9	0.81
0.8	0.64
0.6	0.36
0.5	0.25
0.3	0.09
0.1	0.01

Classification Algorithms

- Instance-based
- Nearest neighbor (kNN)
- Bayesian: combinatorial, naïve and network
- Regression
- SVM
- Decision trees
- Rule-based

Conclusion

Source: digitalistmag.com

- Analytics is promising
- Can’t simply apply analytics on semantic data
- Must work on Indian data sets
- Need of the hour: pre-digested Indian data sets

THANK YOU!