Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: soumyad@cse.iitk.ac.in

Logistics

Course Staff

- Instructor:
  - Soumya Dutta (soumyad@cse.iitk.ac.in)
    - https://soumyadutta-cse.github.io/
- TAs:
  - TA information will be provided to you soon
  - Contact your assigned TA for grading help

- We will use HelloIITK for this course

Class Timings

- Monday & Wednesday
- Time: 2:00pm – 3:15pm
- Location: Rajeev Motwani (RM) Building, Room: 101 (RM - 101)
- Office hours: By email appointment

Course Topics

Index	Module	Topics Covered
1	Fundamentals of Data Visualization	Introduction to Visualization and Visual Analytics
		Foundations of Data Visualization, Visual Abstractions, Visual Variables, Various types of Data
2	Scientific Visualization (SciVis)	Big Data Characteristics, Data Reduction, Various Data Models; Visualization Pipeline
		Scientific Visualization Software such as VTK, ParaView, etc.
		Isosurface Algorithm; Volume Rendering Algorithm
3	Information Visualization (InfoVis)	Fundamentals of Information Visualization, Software for Information Visualization
		High Dimensional Data Analysis and Visualization Techniques
4	Big Data Analysis and Visual Computing Techniques	Big Data Analytics, Statistical Modeling Techniques
		Information Theory Techniques for Visualization
		Time-varying Data, Ensemble Data and Uncertainty Visualization
5	Machine Learning for Visual Computing of Large Data	Machine Learning for Visual Computing and Visual Analytics
		Applications of Machine Learning to Big Data Visual Computing
		Visual Analytics and Explainability of Machine Learning Models
6	Advanced Topics	Extreme-scale Data Analytics, Parallel and High-Performance Visualization
		Exascale Computing, Future Paradigms

Grading/Evaluation Scheme

Category	Split
Attendance	5%
Quiz	10%
Programming Assignments	30%
Mid Sem	25%
Final Sem Project	30%

- Attendance will be taken for a subset of classes and marks will be assigned based on them
- Assignments: Group of 2
- Final semester project: Group of ~7/8 (will be decided later)

Noteworthy Points

1. We might add new, drop existing, or reorder topics depending on the progress and class feedback. Things may be changed by mutual consent after discussion in class.
2. Lectures in the class are the best resources.
3. Grading will be relative.
4. If required, extra classes will also be conducted in weekends.

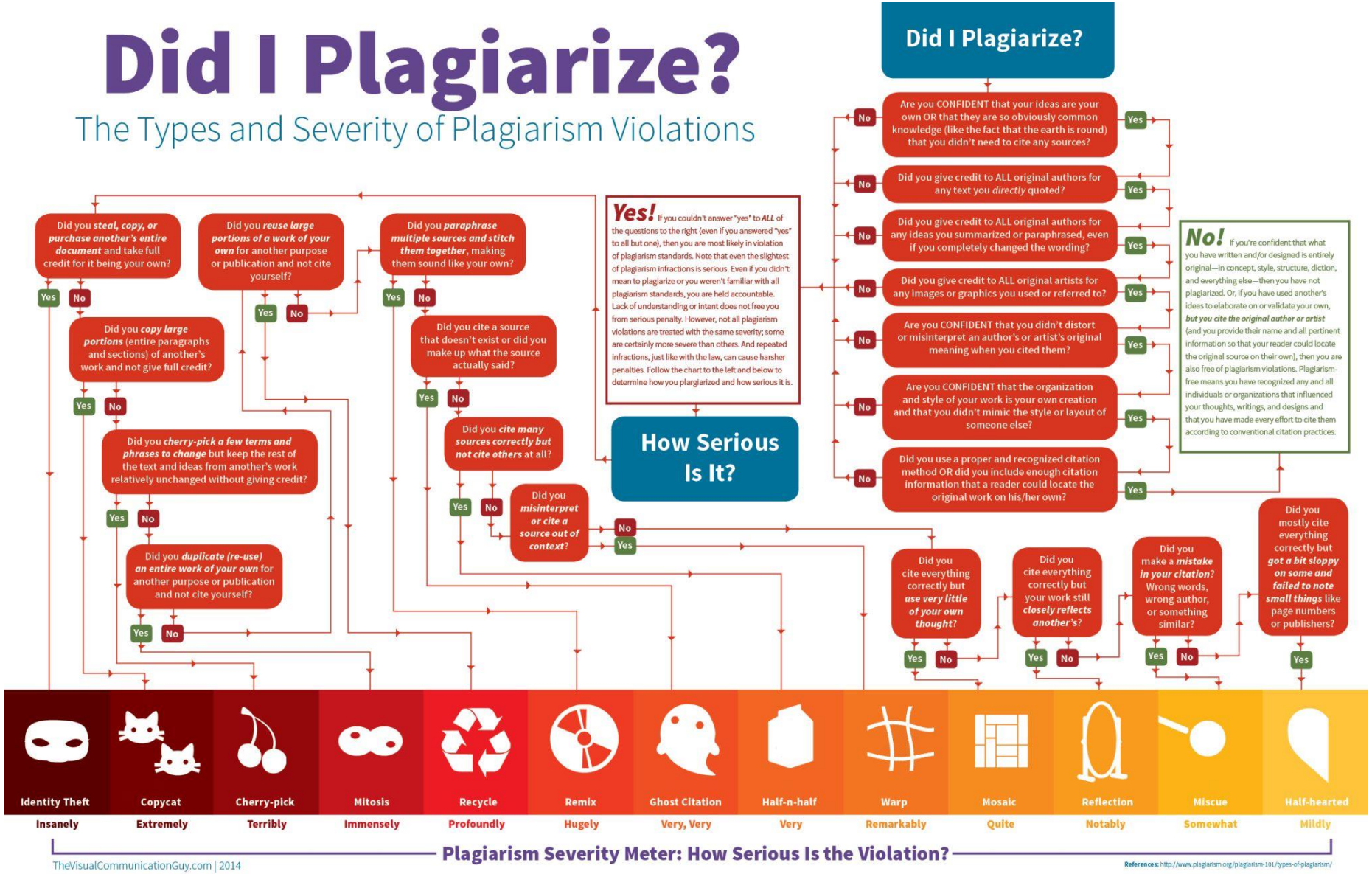
Policies

- Please be on time for the lectures.
- **Attendance will be taken in class from time to time and the marks will be awarded based on it.**
- You are expected to submit your assignments on time.
  - There is a 10% penalty each day after the submission deadline for up to 20% (2 late days). After that, you get zero. This policy will be strictly followed.
- **Students caught cheating or plagiarizing will be dealt with heavy punishment and could automatically fail the course and will be reported to the institute.**
  - Please cite your sources properly in your work.
  - Your assignments should be your own original work.
- If you are unwell, please follow the standard IITK procedure.

Academic Honesty

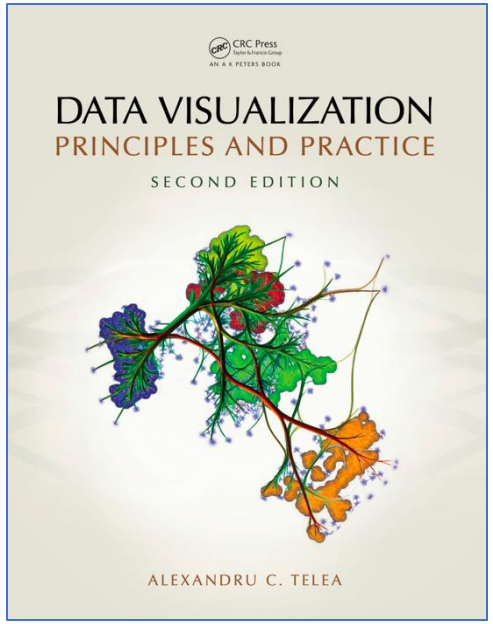
- Please DON'T CHEAT or Plagiarize!!
  - We will do plagiarism check
- Students caught cheating or plagiarizing may fail the course and will be reported to the institute.
- You must cite all sources in your works **including AI tools**.
- Your assignments should be **your own original work**.
- IITK CSE Anti-cheating policy: https://www.cse.iitk.ac.in/pages/AntiCheatingPolicy.html
- The List of Things I Never Want To Hear Again (by Tamara Munzner)
  - https://www.cs.ubc.ca/~tmm/courses/cheat.html

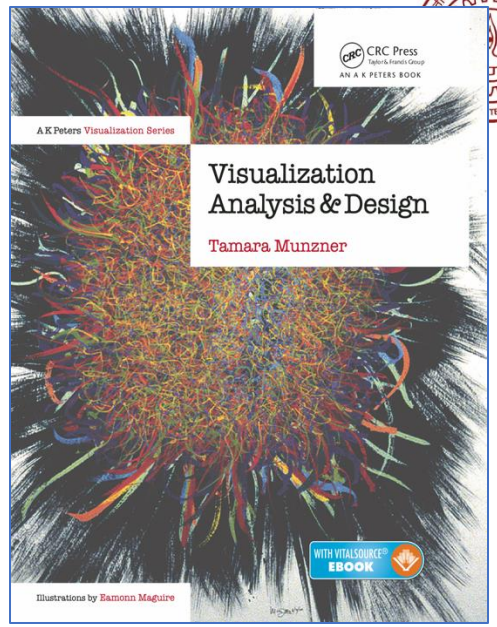
Plagiarism Flow Chart

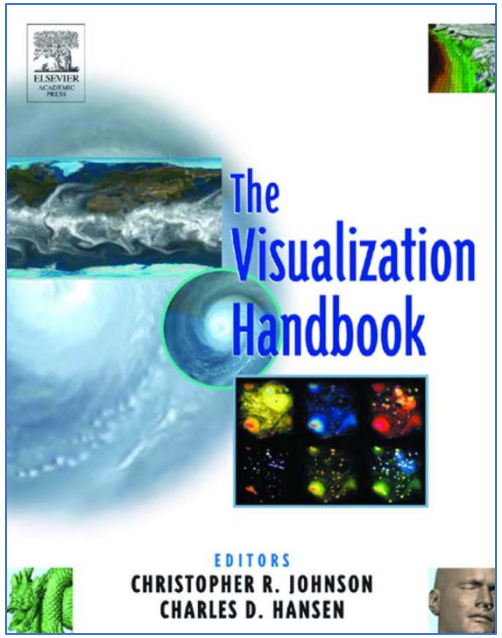


Resources

- Data Visualization: Principles and Practice by Alexandru C. Telea, CRC Press
- Visualization Analysis and Design by Tamara Munzner, A K Peters Visualization Series, CRC Press
- The Visualization Handbook edited by Charles D. Hansen and Chris R. Johnson
- Research papers and other study materials provided during the class to cover selected topics

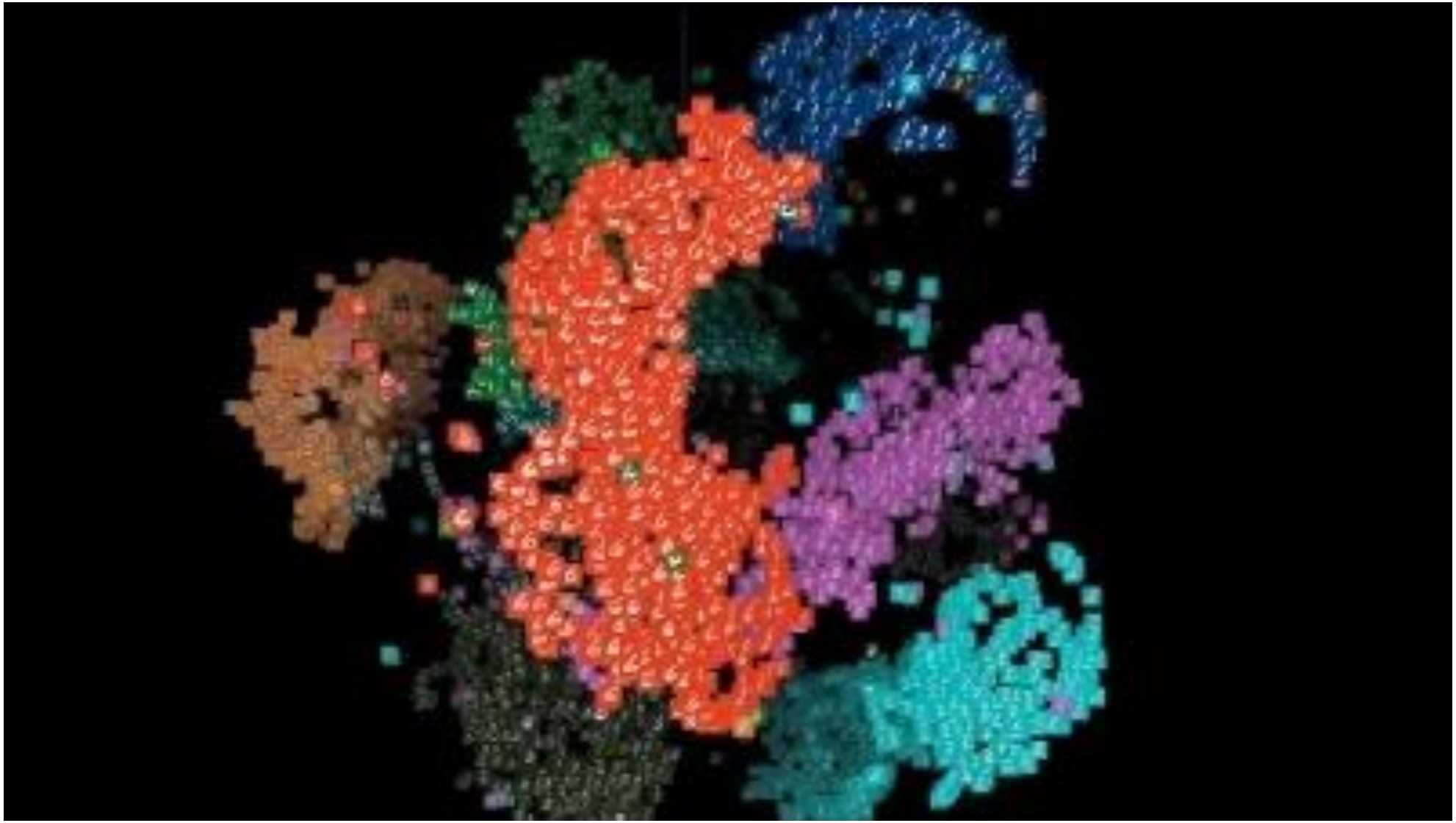






Visualization is Cool! Applications in Science

Visualization is Cool! Applications in ML

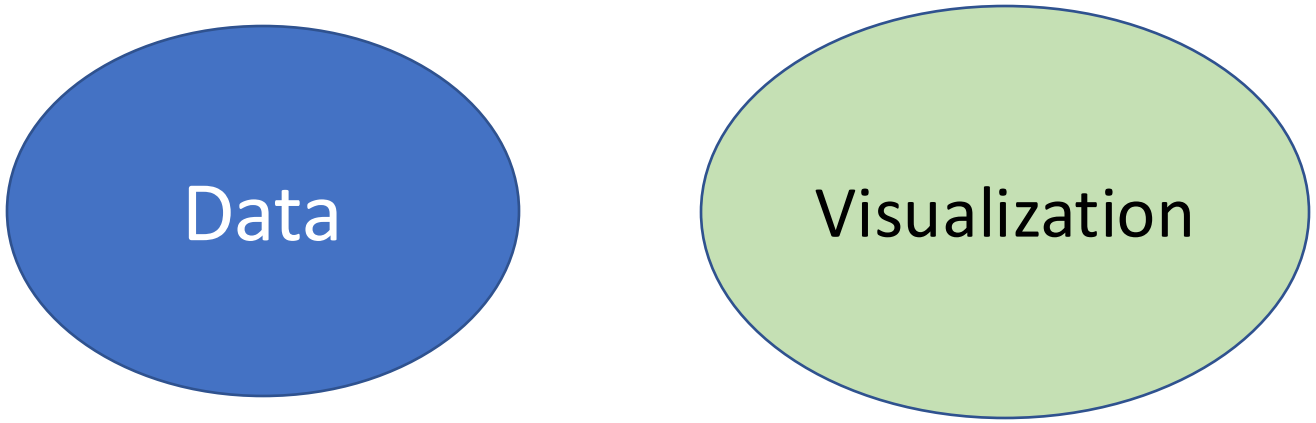


https://www.youtube.com/watch?v=wvsE8jm1GzE&list=RDQMjKx69KjeL5I&index=4

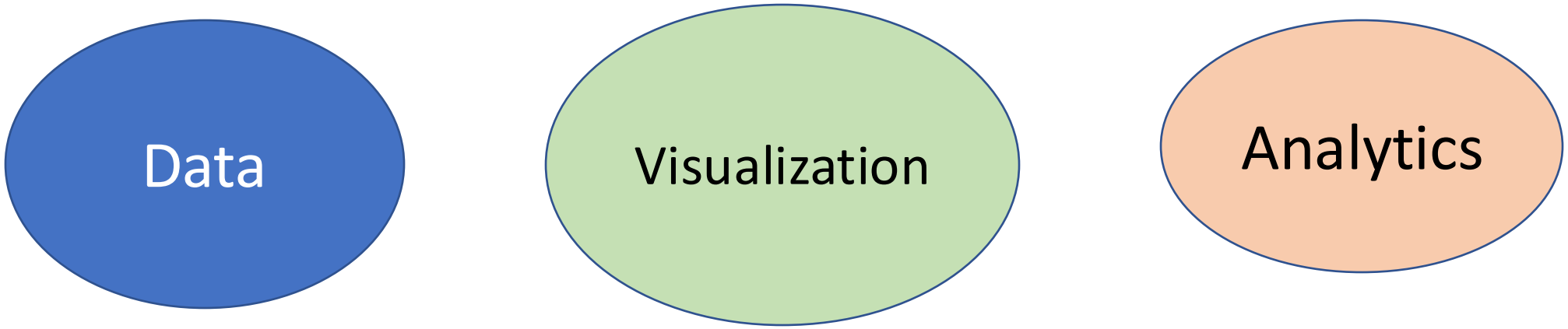
Course Overview: What Are You Going to Learn?

Overview

Overview



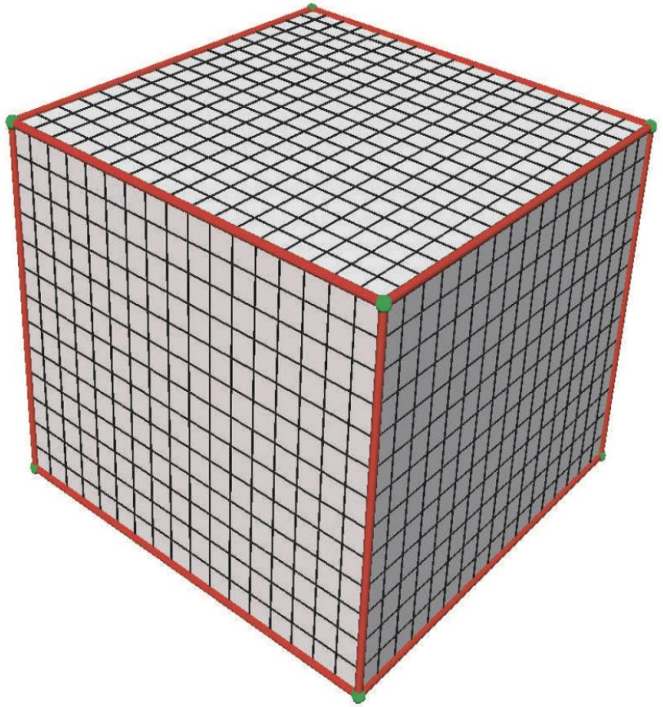
Overview

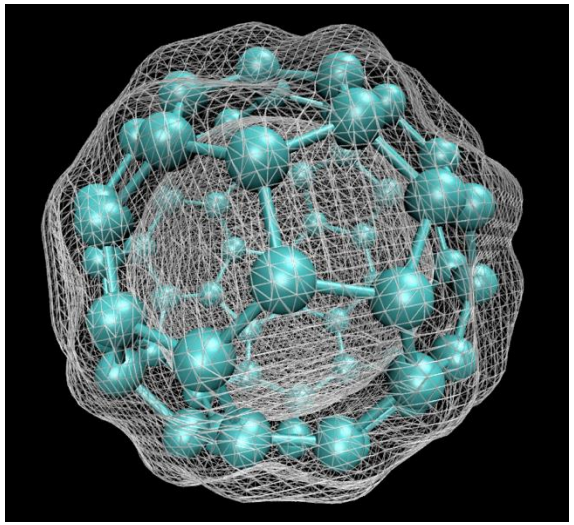


Data

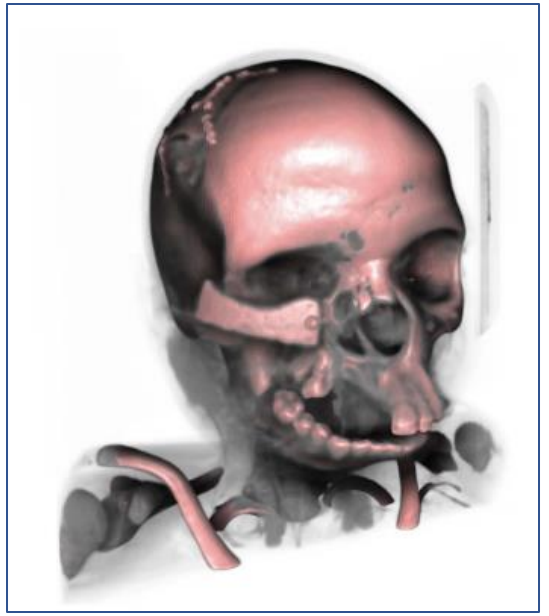
- Various types of data
- How to handle such data?
- How to process and analyze such data?
- How to Visualize such data?
- How to perform interactive analytics with data?
- How to find features/patterns from data?
- How to deal with big data?
- How to intelligently summarize large data?

From Data to Visualization



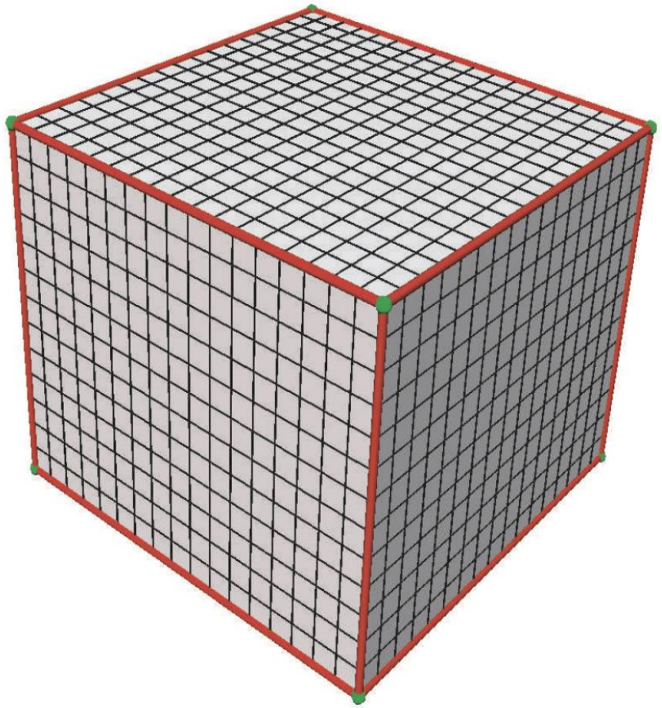


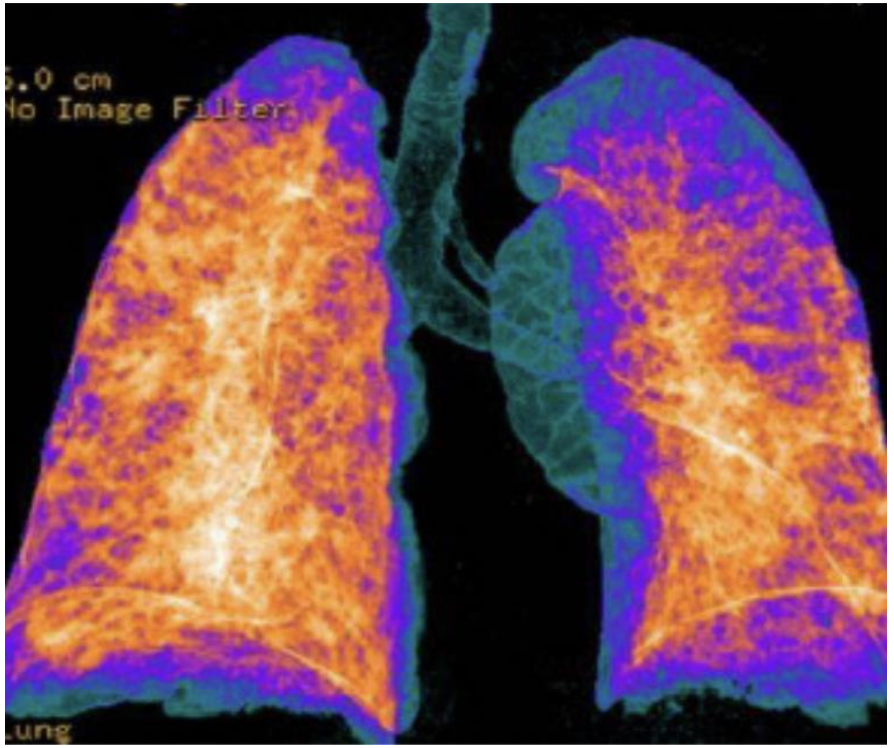
Visualization
in Physical Science



Visualization
in Medical Science

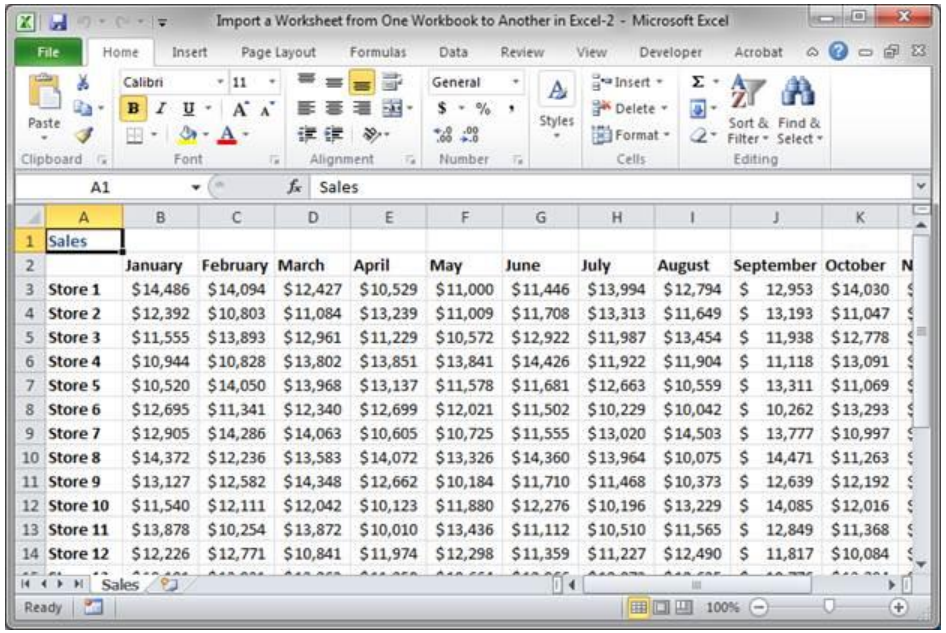
From Data to Visualization



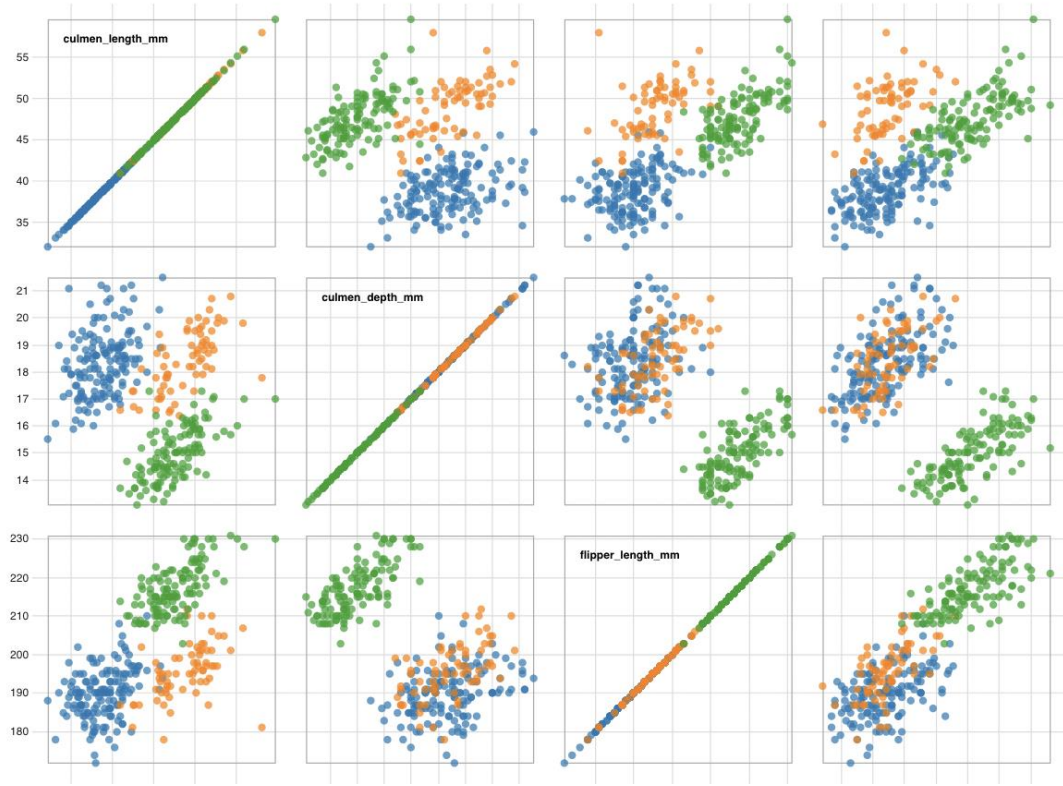


Visualization of Covid 19 Data

From Data to Visualization

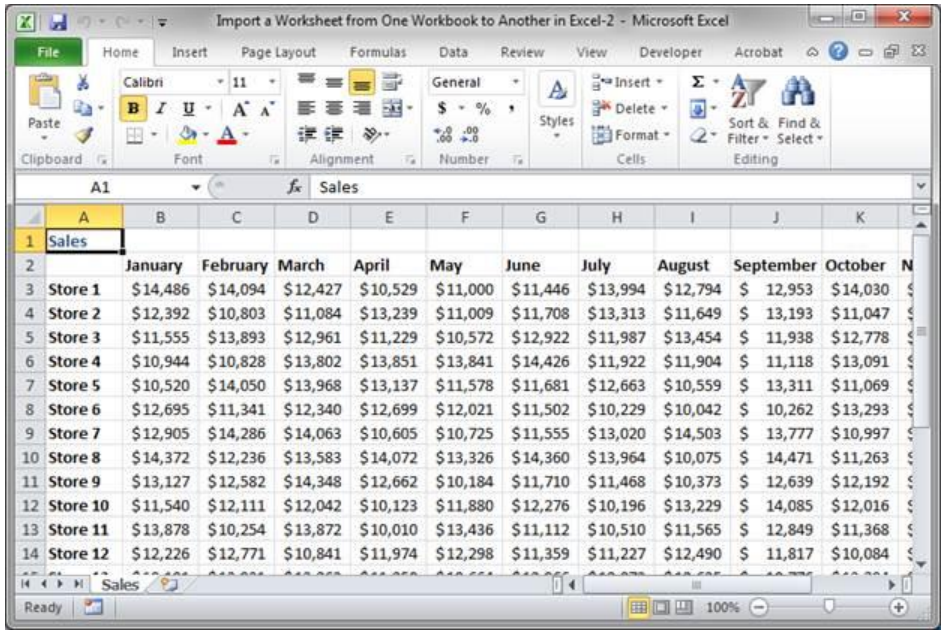


	A	B	C	D	E	F	G	H	I	J	K
1	Sales										
2		January	February	March	April	May	June	July	August	September	October
3	Store 1	$14,486	$14,094	$12,427	$10,529	$11,000	$11,446	$13,994	$12,794	$ 12,953	$14,030
4	Store 2	$12,392	$10,803	$11,084	$13,239	$11,009	$11,708	$13,313	$11,649	$ 13,193	$11,047
5	Store 3	$11,555	$13,893	$12,961	$11,229	$10,572	$12,922	$11,987	$13,454	$ 11,938	$12,778
6	Store 4	$10,944	$10,828	$13,802	$13,851	$13,841	$14,426	$11,922	$11,904	$ 11,118	$13,091
7	Store 5	$10,520	$14,050	$13,968	$13,137	$11,578	$11,681	$12,663	$10,559	$ 13,311	$11,069
8	Store 6	$12,695	$11,341	$12,340	$12,699	$12,021	$11,502	$10,229	$10,042	$ 10,262	$13,293
9	Store 7	$12,905	$14,286	$14,063	$10,605	$10,725	$11,555	$13,020	$14,503	$ 13,777	$10,997
10	Store 8	$14,372	$12,236	$13,583	$14,072	$13,326	$14,360	$13,964	$10,075	$ 14,471	$11,263
11	Store 9	$13,127	$12,582	$14,348	$12,662	$10,184	$11,710	$11,468	$10,373	$ 12,639	$12,192
12	Store 10	$11,540	$12,111	$12,042	$10,123	$11,880	$12,276	$10,196	$13,229	$ 14,085	$12,016
13	Store 11	$13,878	$10,254	$13,872	$10,010	$13,436	$11,112	$10,510	$11,565	$ 12,849	$11,368
14	Store 12	$12,226	$12,771	$10,841	$11,974	$12,298	$11,359	$11,227	$12,490	$ 11,817	$10,084

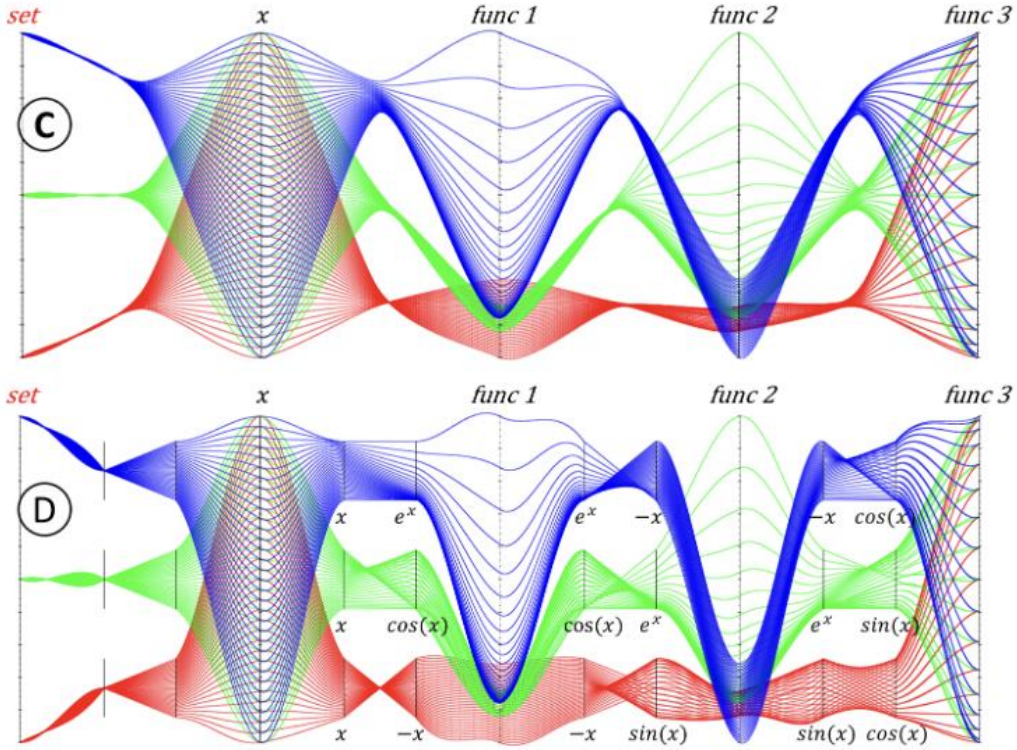


Visualization of correlation

From Data to Visualization

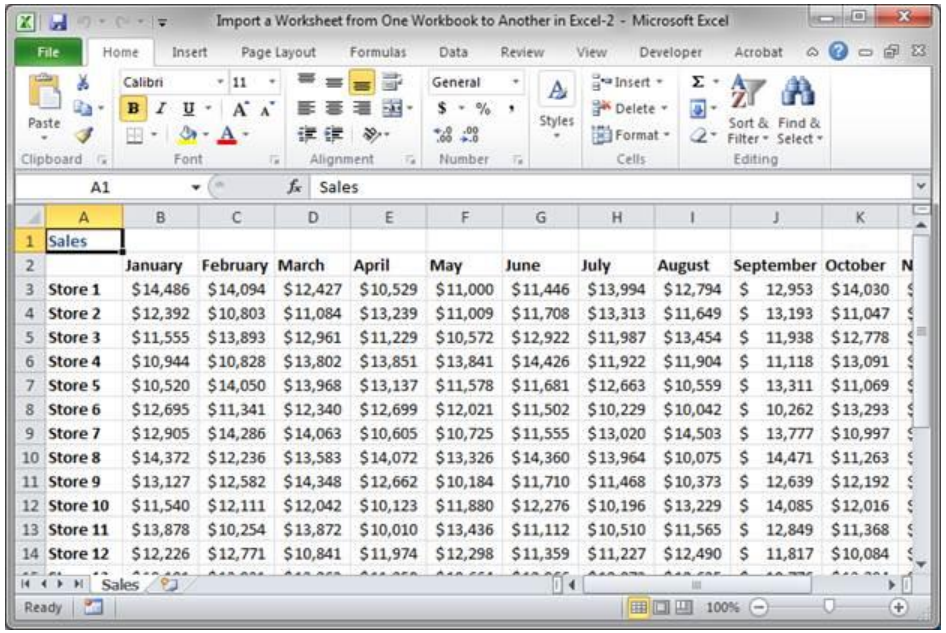


	A	B	C	D	E	F	G	H	I	J	K
1	Sales										
2		January	February	March	April	May	June	July	August	September	October
3	Store 1	$14,486	$14,094	$12,427	$10,529	$11,000	$11,446	$13,994	$12,794	$ 12,953	$14,030
4	Store 2	$12,392	$10,803	$11,084	$13,239	$11,009	$11,708	$13,313	$11,649	$ 13,193	$11,047
5	Store 3	$11,555	$13,893	$12,961	$11,229	$10,572	$12,922	$11,987	$13,454	$ 11,938	$12,778
6	Store 4	$10,944	$10,828	$13,802	$13,851	$13,841	$14,426	$11,922	$11,904	$ 11,118	$13,091
7	Store 5	$10,520	$14,050	$13,968	$13,137	$11,578	$11,681	$12,663	$10,559	$ 13,311	$11,069
8	Store 6	$12,695	$11,341	$12,340	$12,699	$12,021	$11,502	$10,229	$10,042	$ 10,262	$13,293
9	Store 7	$12,905	$14,286	$14,063	$10,605	$10,725	$11,555	$13,020	$14,503	$ 13,777	$10,997
10	Store 8	$14,372	$12,236	$13,583	$14,072	$13,326	$14,360	$13,964	$10,075	$ 14,471	$11,263
11	Store 9	$13,127	$12,582	$14,348	$12,662	$10,184	$11,710	$11,468	$10,373	$ 12,639	$12,192
12	Store 10	$11,540	$12,111	$12,042	$10,123	$11,880	$12,276	$10,196	$13,229	$ 14,085	$12,016
13	Store 11	$13,878	$10,254	$13,872	$10,010	$13,436	$11,112	$10,510	$11,565	$ 12,849	$11,368
14	Store 12	$12,226	$12,771	$10,841	$11,974	$12,298	$11,359	$11,227	$12,490	$ 11,817	$10,084

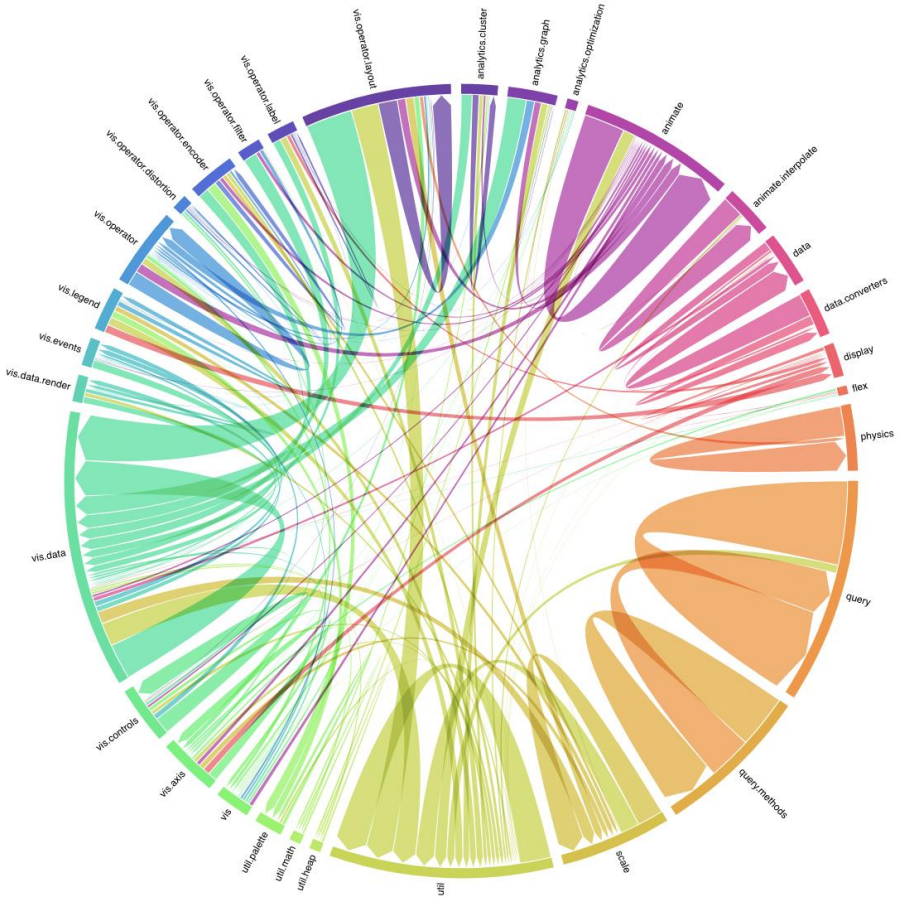


Visualization for finding relationships in variables

From Data to Visualization

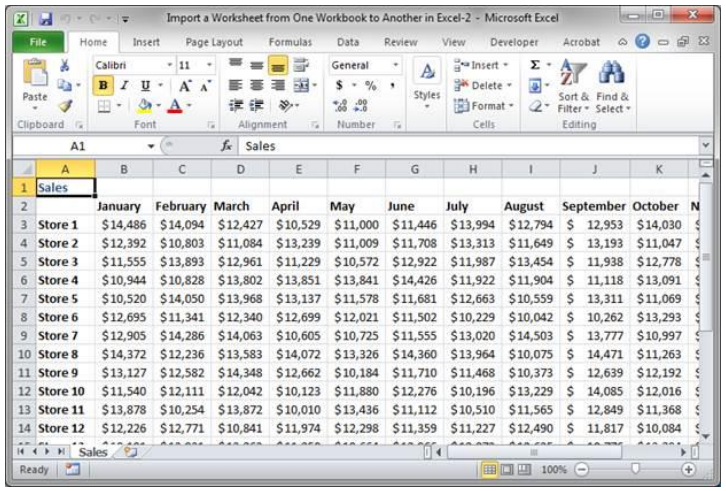


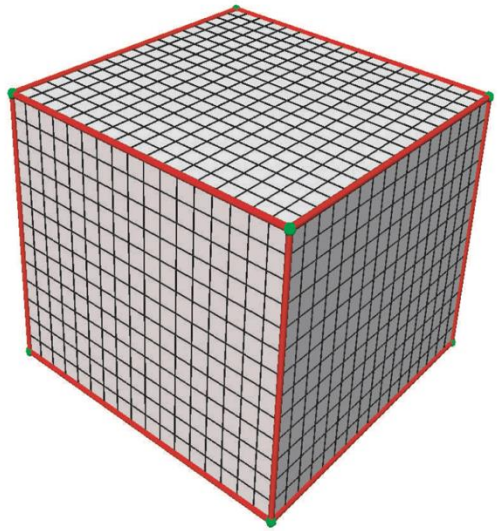
	A	B	C	D	E	F	G	H	I	J	K
1	Sales										
2		January	February	March	April	May	June	July	August	September	October
3	Store 1	$14,486	$14,094	$12,427	$10,529	$11,000	$11,446	$13,994	$12,794	$ 12,953	$14,030
4	Store 2	$12,392	$10,803	$11,084	$13,239	$11,009	$11,708	$13,313	$11,649	$ 13,193	$11,047
5	Store 3	$11,555	$13,893	$12,961	$11,229	$10,572	$12,922	$11,987	$13,454	$ 11,938	$12,778
6	Store 4	$10,944	$10,828	$13,802	$13,851	$13,841	$14,426	$11,922	$11,904	$ 11,118	$13,091
7	Store 5	$10,520	$14,050	$13,968	$13,137	$11,578	$11,681	$12,663	$10,559	$ 13,311	$11,069
8	Store 6	$12,695	$11,341	$12,340	$12,699	$12,021	$11,502	$10,229	$10,042	$ 10,262	$13,293
9	Store 7	$12,905	$14,286	$14,063	$10,605	$10,725	$11,555	$13,020	$14,503	$ 13,777	$10,997
10	Store 8	$14,372	$12,236	$13,583	$14,072	$13,326	$14,360	$13,964	$10,075	$ 14,471	$11,263
11	Store 9	$13,127	$12,582	$14,348	$12,662	$10,184	$11,710	$11,468	$10,373	$ 12,639	$12,192
12	Store 10	$11,540	$12,111	$12,042	$10,123	$11,880	$12,276	$10,196	$13,229	$ 14,085	$12,016
13	Store 11	$13,878	$10,254	$13,872	$10,010	$13,436	$11,112	$10,510	$11,565	$ 12,849	$11,368
14	Store 12	$12,226	$12,771	$10,841	$11,974	$12,298	$11,359	$11,227	$12,490	$ 11,817	$10,084

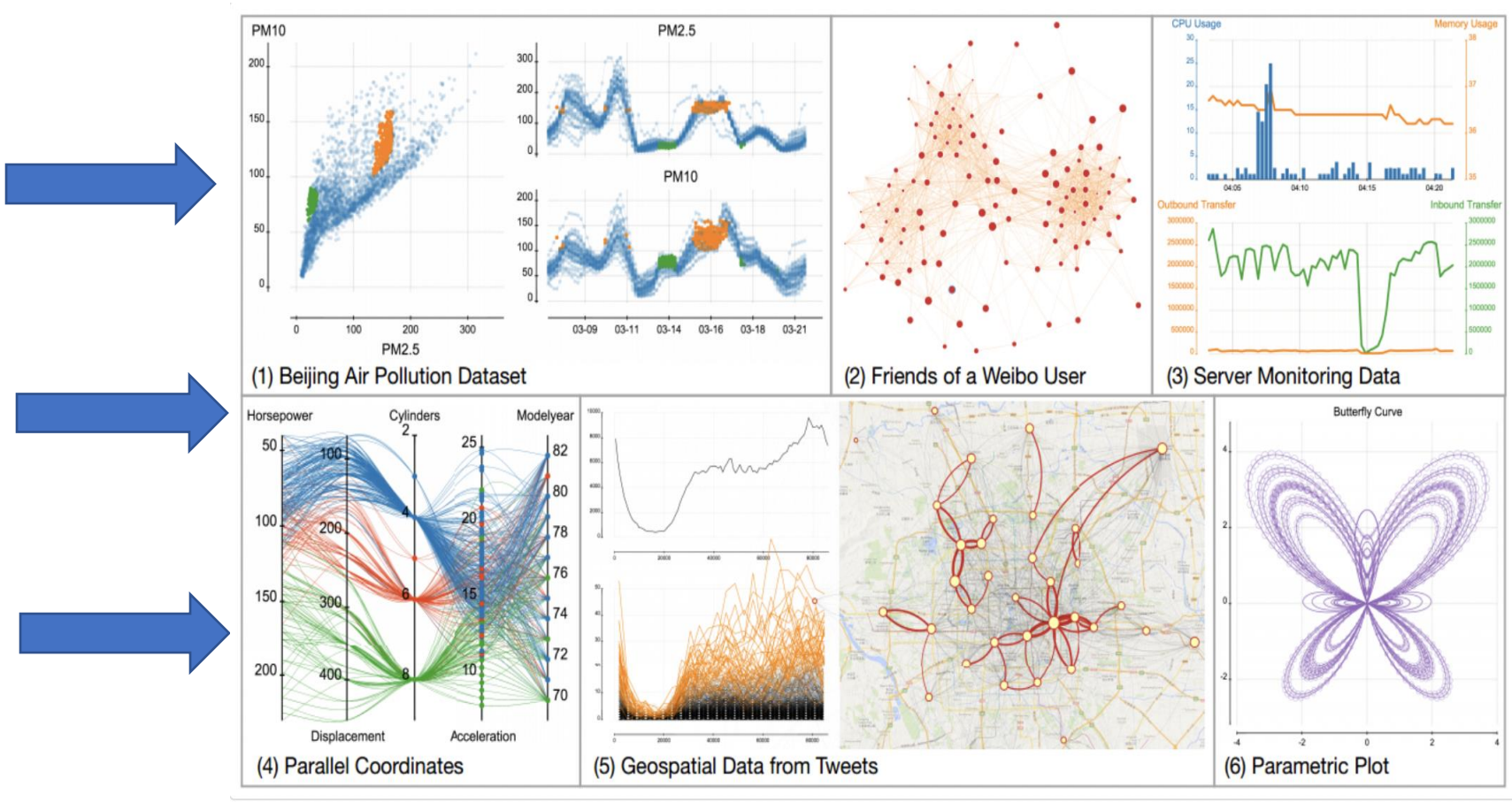


Visualization for finding connections in variables

High Dimensional Data to Visualization Space

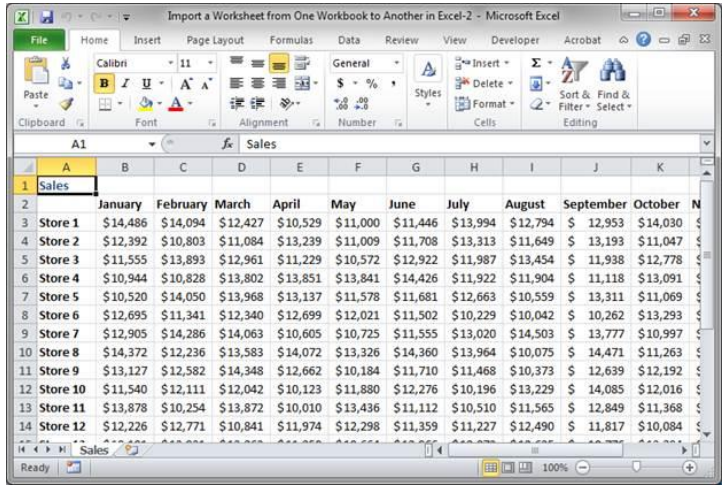




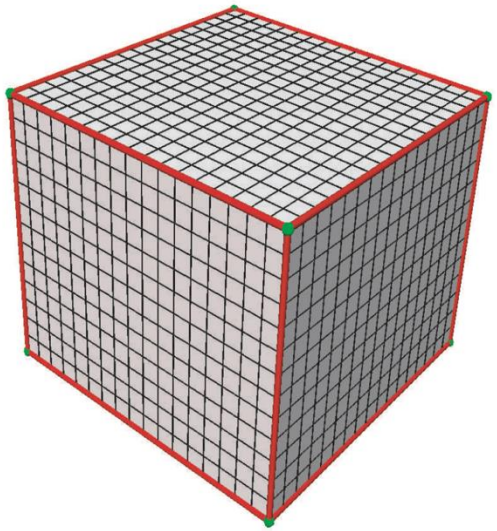


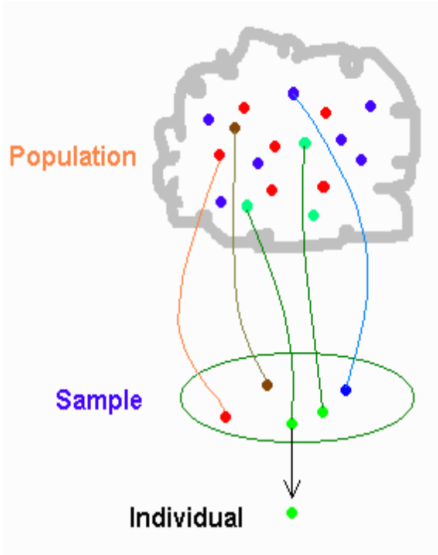
Visualizing High Dimensional Data

From Data to Transformed Representation

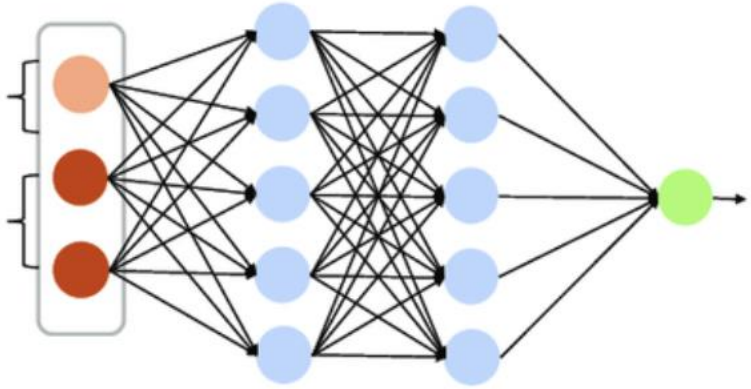


Sales	January	February	March	April	May	June	July	August	September	October
Store 1	$14,486	$14,094	$12,427	$10,529	$11,000	$11,446	$13,994	$12,794	$ 12,953	$14,030
Store 2	$12,392	$10,803	$11,084	$13,239	$11,009	$11,708	$13,313	$11,649	$ 13,193	$11,047
Store 3	$11,555	$13,893	$12,961	$11,229	$10,572	$12,922	$11,987	$13,454	$ 11,938	$12,778
Store 4	$10,944	$10,828	$13,802	$13,851	$13,841	$14,426	$11,922	$11,904	$ 11,118	$13,091
Store 5	$10,520	$14,050	$13,968	$13,137	$11,578	$11,681	$12,663	$10,559	$ 13,311	$11,069
Store 6	$12,695	$11,341	$12,340	$12,699	$12,021	$11,502	$10,229	$10,042	$ 10,262	$13,293
Store 7	$12,905	$14,286	$14,063	$10,605	$10,725	$11,555	$13,020	$14,503	$ 13,777	$10,997
Store 8	$14,372	$12,236	$13,583	$14,072	$13,326	$14,360	$13,964	$10,075	$ 14,471	$11,263
Store 9	$13,127	$12,582	$14,348	$12,662	$10,184	$11,710	$11,468	$10,373	$ 12,639	$12,192
Store 10	$11,540	$12,111	$12,042	$10,123	$11,880	$12,276	$10,196	$13,229	$ 14,085	$12,016
Store 11	$13,878	$10,254	$13,872	$10,010	$13,436	$11,112	$10,510	$11,565	$ 12,849	$11,368
Store 12	$12,226	$12,771	$10,841	$11,974	$12,298	$11,359	$11,227	$12,490	$ 11,817	$10,084



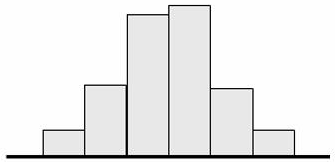


Sampling



ML Model

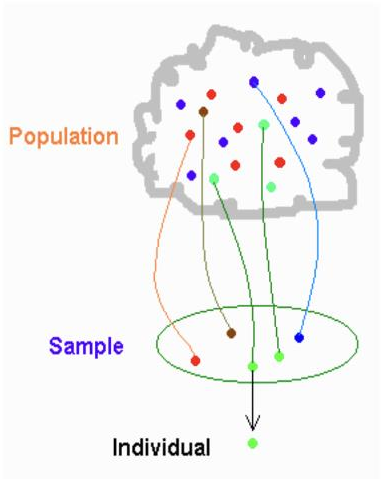
From Transformed Data to Visualization



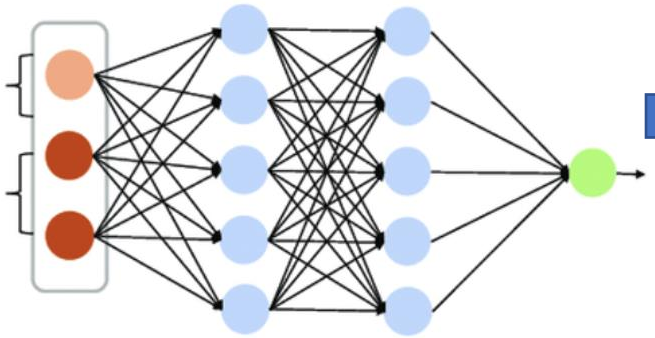
Histogram

Gaussian

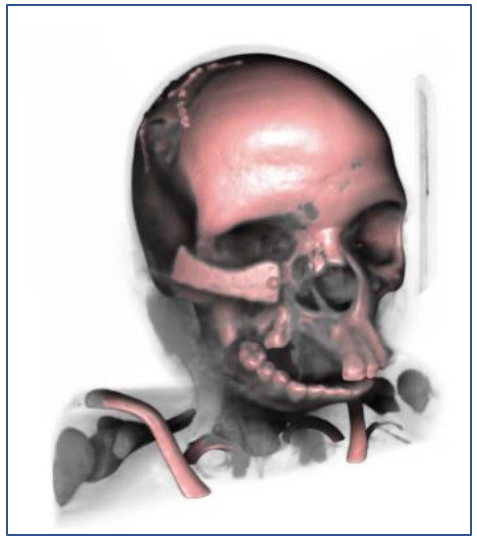
Gaussian Mixture Model

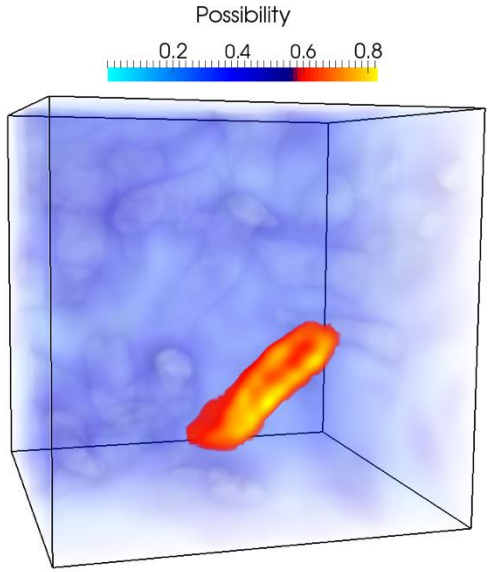


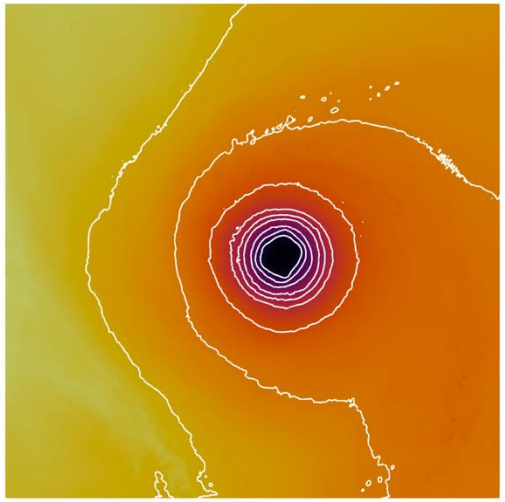
Sampling



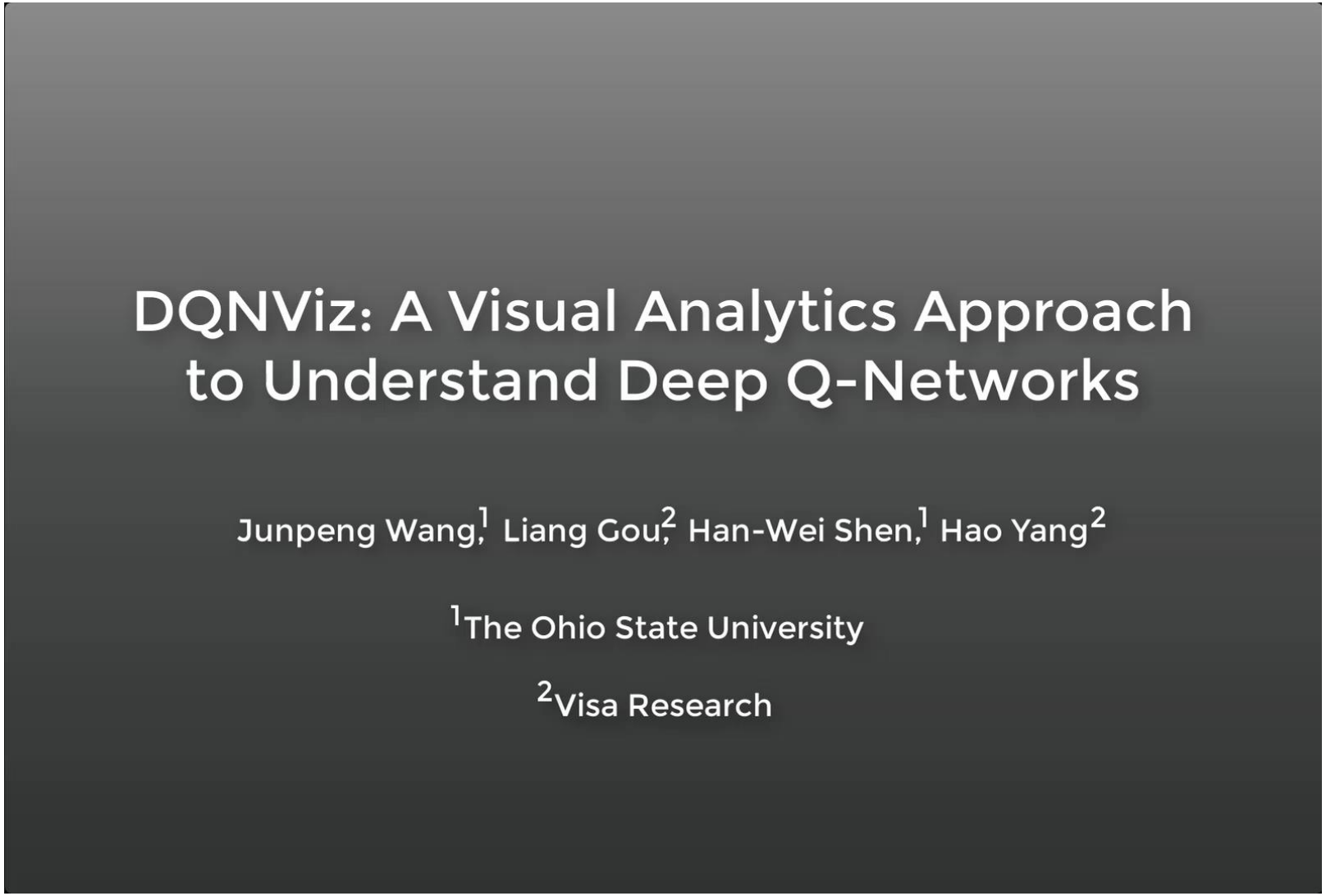
ML Model





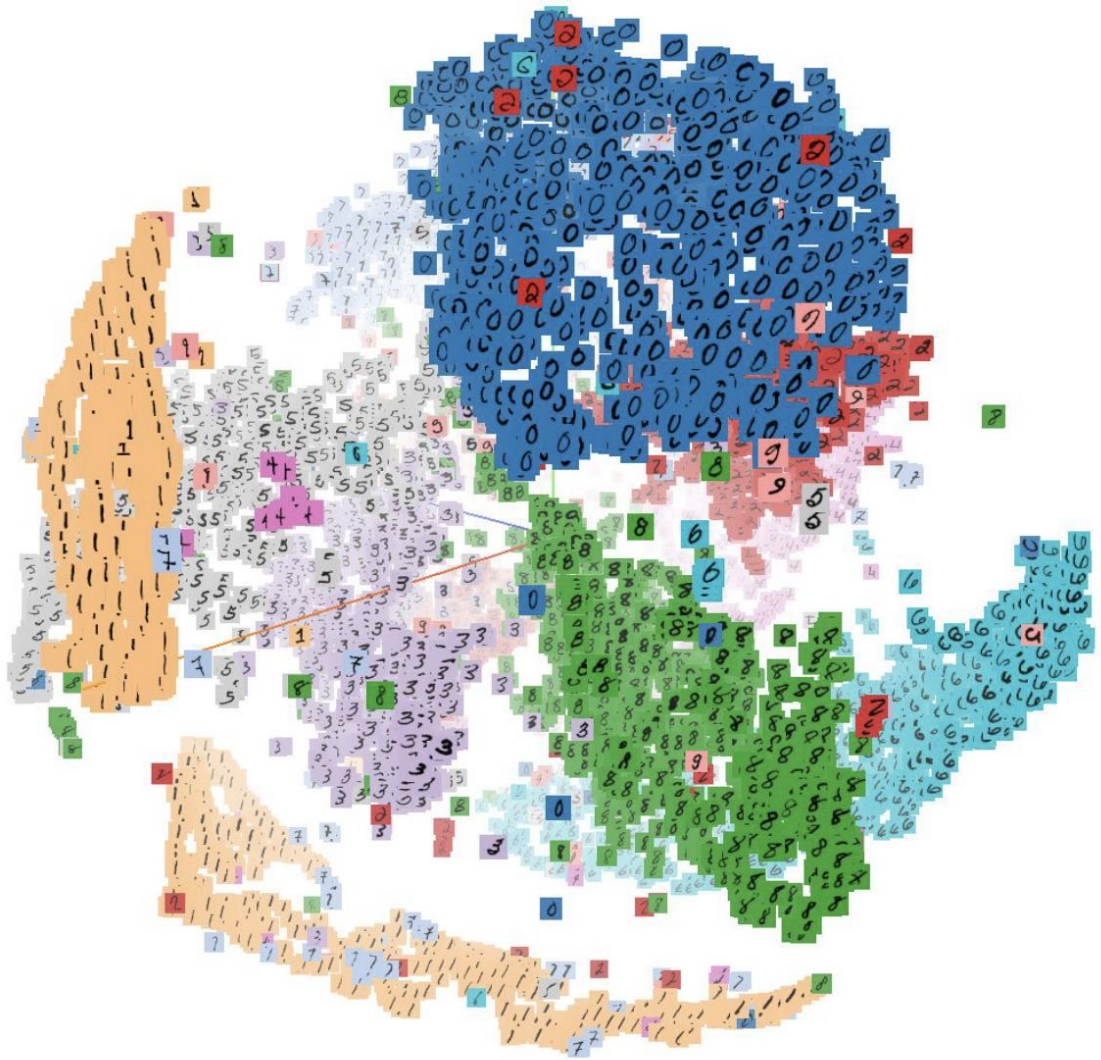


Visualization for Machine Learning



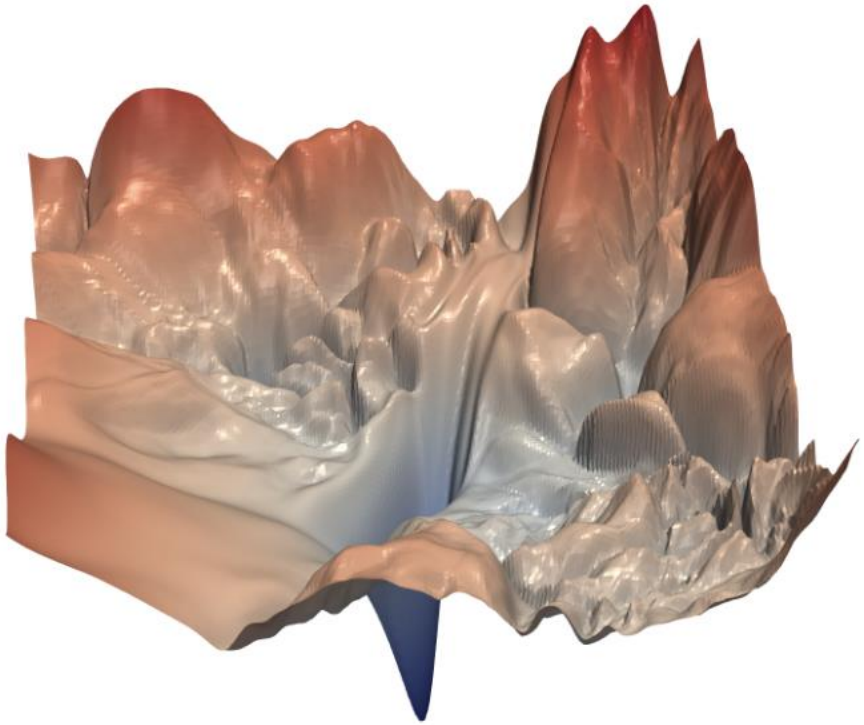
DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks, J. Wang et al. TVCG

Visualization for Machine Learning

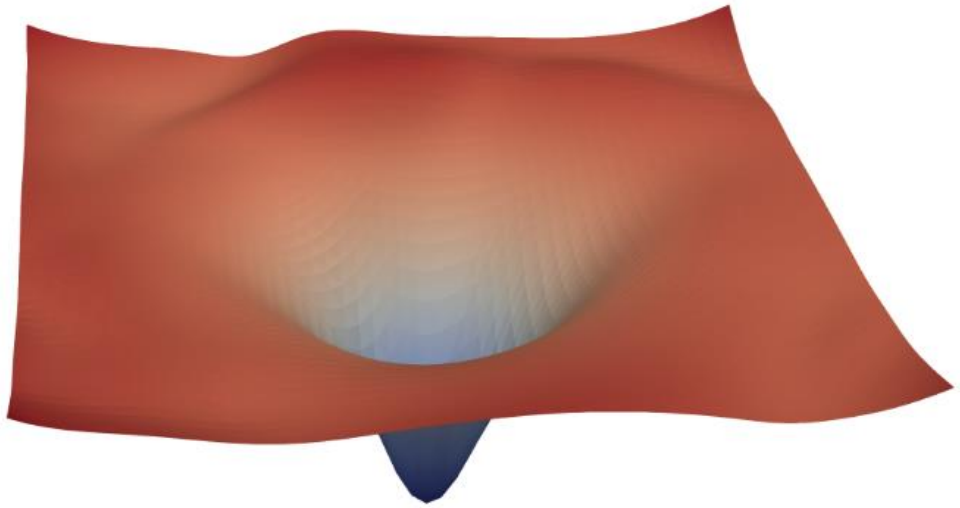


Visualizing Results of a ML Model Prediction

Visualization for Machine Learning: Loss Landscapes



Loss landscape of ResNet-56
without skip-connection

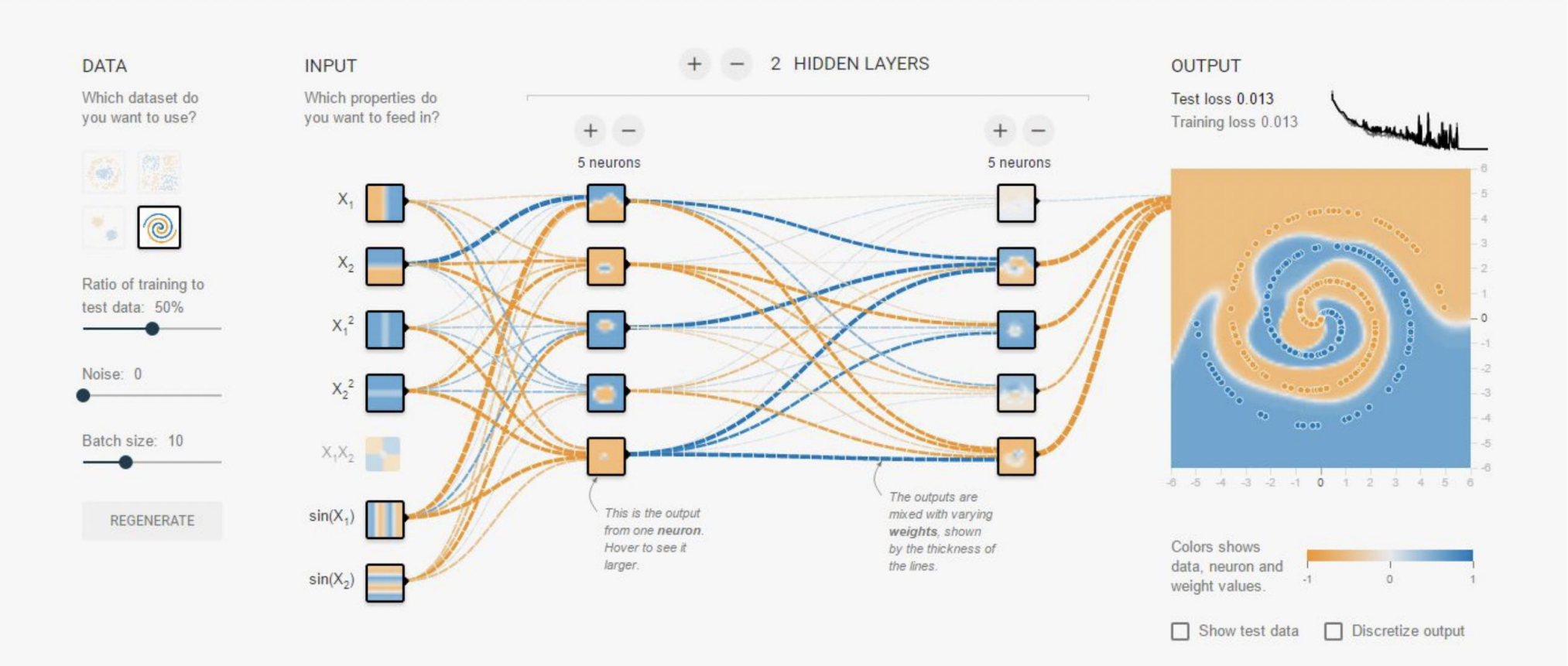


Loss landscape of ResNet-56
with skip-connection

Visualization for Machine Learning

TensorFlow Playground

playground.tensorflow.org



Visualization for Machine Learning

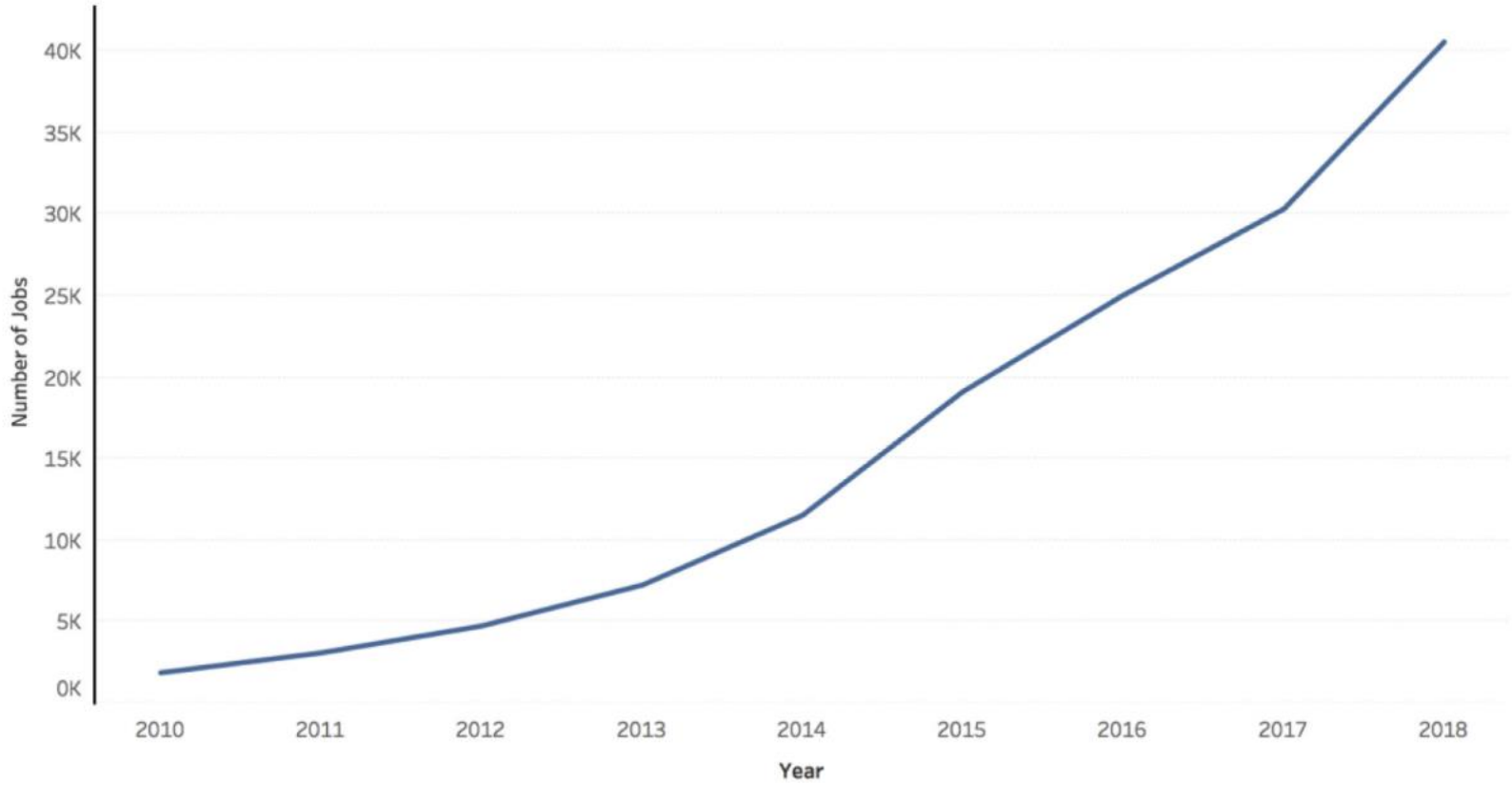
- CNNVis
  - http://shixialiu.com/publications/cnnvis/demo/
- CNN Explainer
  - https://poloclub.github.io/cnn-explainer/
- Understanding DNN:
  - https://distill.pub/2020/grand-tour/
- GAN lab
  - https://poloclub.github.io/ganlab/

Visualization and Data Analysis at Exascale

- High Performance Visualization
- *In Situ* Data Analysis and Visualization

Frontier: World's First Exascale Supercomputer (https://www.olcf.ornl.gov/frontier/)
Exascale: 10^{18} IEEE 754 Double Precision (64-bit) operations (multiplications and/or additions) per second (ExaFlops)

Why Should You Learn Visual Analytics?



The growth of jobs mentioning “data visualization” as a skill from 2010 has steadily increased from only 1,888 jobs in 2010 to 30,327 jobs in 2017 **(16X growth)**

What is Expected From You?

- Basic knowledge of Linear Algebra, Probability, and Statistics
- Strong programming background (C/C++, Python, JavaScript)
- Interest and Motivation to learn new topics (sometimes on your own!)
- Creativity and Imagination

If you do not have the above skills or unsure, talk to me!

- Goal of the Course: Give you a comprehensive view of the Big Data Visual Computing and Analysis Domain
  - Conduct research in these topics
  - Use the learned skills in industry/academia

Introduction

Acknowledgements

- Some of the following slides are adapted from the excellent course materials made available by:
  - Prof. Klaus Mueller (State University of New York at Stony Brook)
  - Prof. Tamara Munzner (University of British Columbia)

Visualization

- Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Visualization: Human in the Loop… Why?

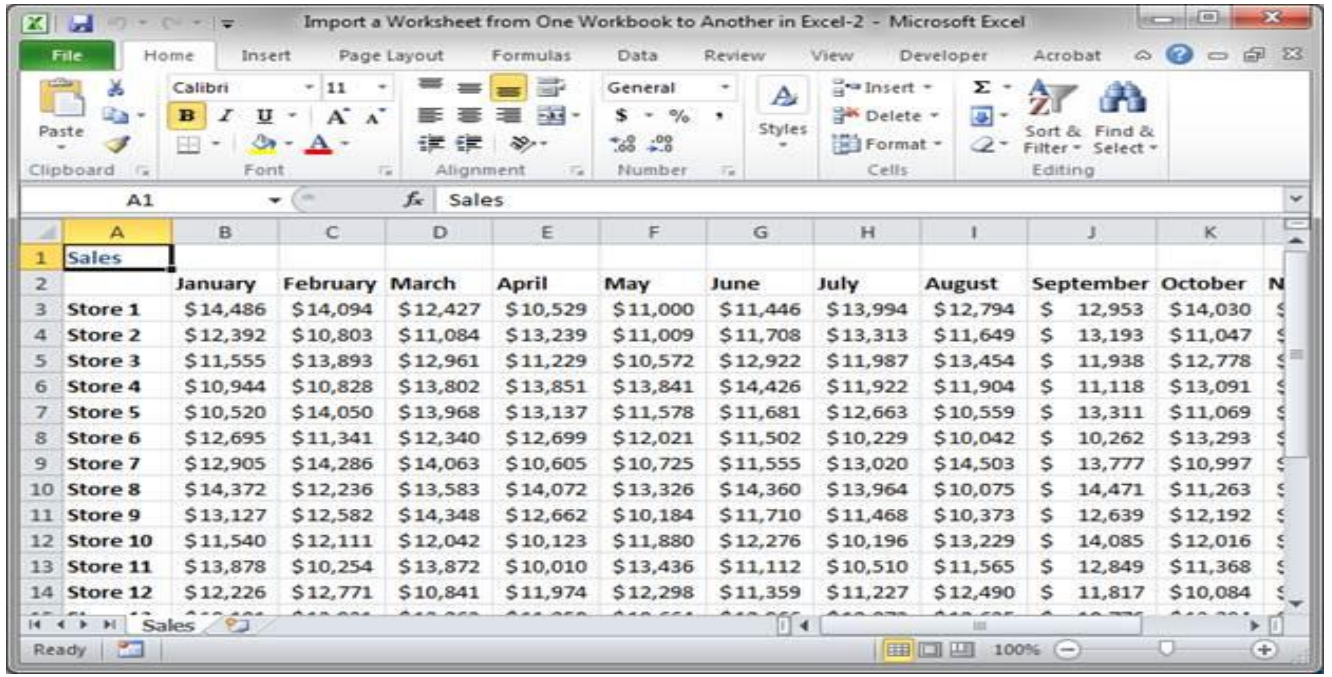
- Computer-based visualization systems provide visual representations of **datasets** designed to help **people** carry out tasks more effectively.
  - Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

Visualization: Human in the Loop… Why?

- Computer-based visualization systems provide visual representations of **datasets** designed to help **people** carry out tasks more effectively.
  - Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.
- Don’t need vis when fully automatic solution exists and is trusted
- Many analysis problems ill-specified
  - don’t know exactly what questions to ask in advance

Representation for Data: Why?

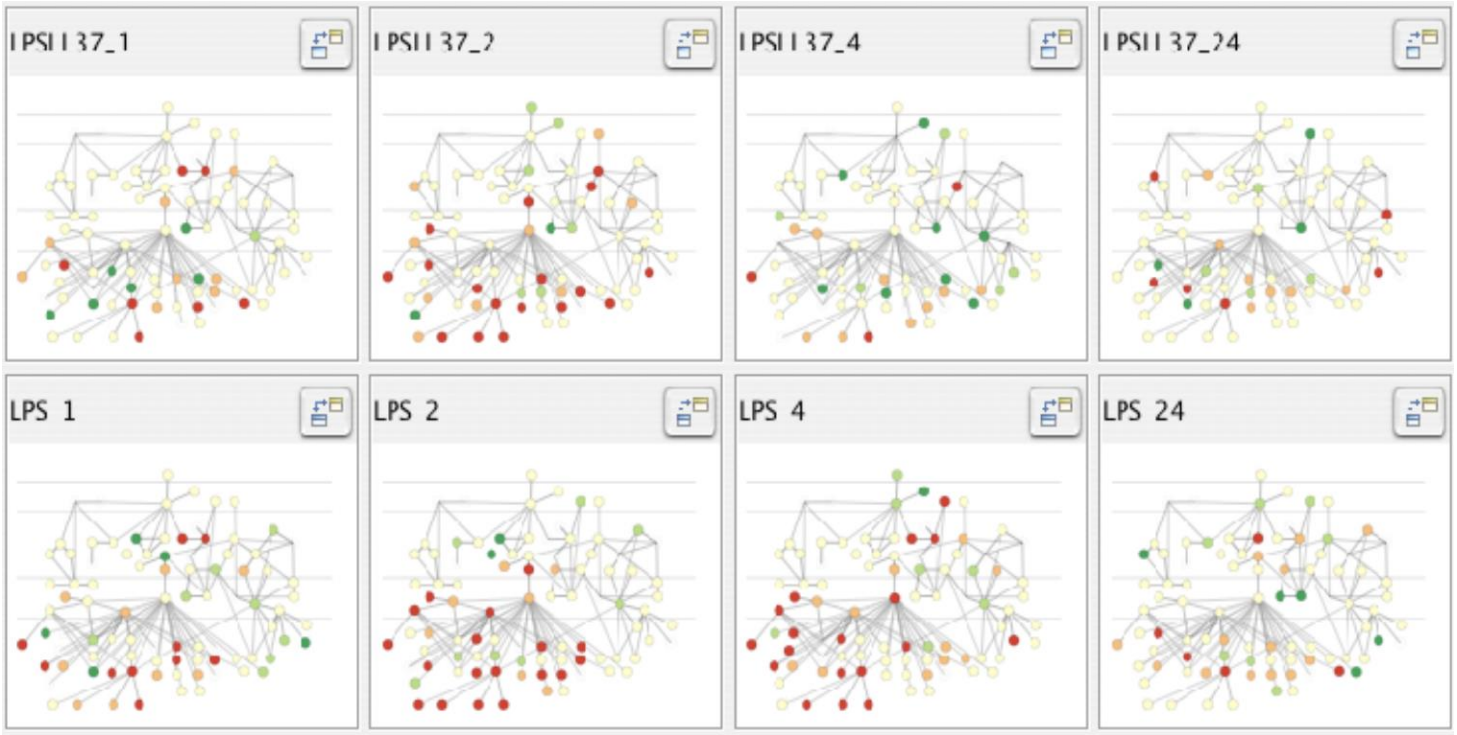
- Computer-based visualization systems provide **visual representations** of datasets designed to help people carry out tasks more effectively.
- Replace cognition with perception



	A	B	C	D	E	F	G	H	I	J	K
1	Sales										
2		January	February	March	April	May	June	July	August	September	October
3	Store 1	$14,486	$14,094	$12,427	$10,529	$11,000	$11,446	$13,994	$12,794	$ 12,953	$14,030
4	Store 2	$12,392	$10,803	$11,084	$13,239	$11,009	$11,708	$13,313	$11,649	$ 13,193	$11,047
5	Store 3	$11,555	$13,893	$12,961	$11,229	$10,572	$12,922	$11,987	$13,454	$ 11,938	$12,778
6	Store 4	$10,944	$10,828	$13,802	$13,851	$13,841	$14,426	$11,922	$11,904	$ 11,118	$13,091
7	Store 5	$10,520	$14,050	$13,968	$13,137	$11,578	$11,681	$12,663	$10,559	$ 13,311	$11,069
8	Store 6	$12,695	$11,341	$12,340	$12,699	$12,021	$11,502	$10,229	$10,042	$ 10,262	$13,293
9	Store 7	$12,905	$14,286	$14,063	$10,605	$10,725	$11,555	$13,020	$14,503	$ 13,777	$10,997
10	Store 8	$14,372	$12,236	$13,583	$14,072	$13,326	$14,360	$13,964	$10,075	$ 14,471	$11,263
11	Store 9	$13,127	$12,582	$14,348	$12,662	$10,184	$11,710	$11,468	$10,373	$ 12,639	$12,192
12	Store 10	$11,540	$12,111	$12,042	$10,123	$11,880	$12,276	$10,196	$13,229	$ 14,085	$12,016
13	Store 11	$13,878	$10,254	$13,872	$10,010	$13,436	$11,112	$10,510	$11,565	$ 12,849	$11,368
14	Store 12	$12,226	$12,771	$10,841	$11,974	$12,298	$11,359	$11,227	$12,490	$ 11,817	$10,084

Representation for Data: Why?

- Computer-based visualization systems provide **visual representations** of datasets designed to help people carry out tasks more effectively.
- Replace cognition with perception



Why Depend on Vision?

- Computer-based visualization systems provide **visual representations** of datasets designed to help people carry out tasks more effectively.

- ~50% (roughly) of our brain is dedicated to vision
- Human visual system is high-bandwidth channel to brain
- Vision is a Massively Parallel Processor dedicated to
  - Detect
  - Analyze
  - Recognize
  - Reason with

Why not Only Show the Summary Data?

- Computer-based visualization systems provide visual **representations of datasets** designed to help people carry out tasks more effectively.

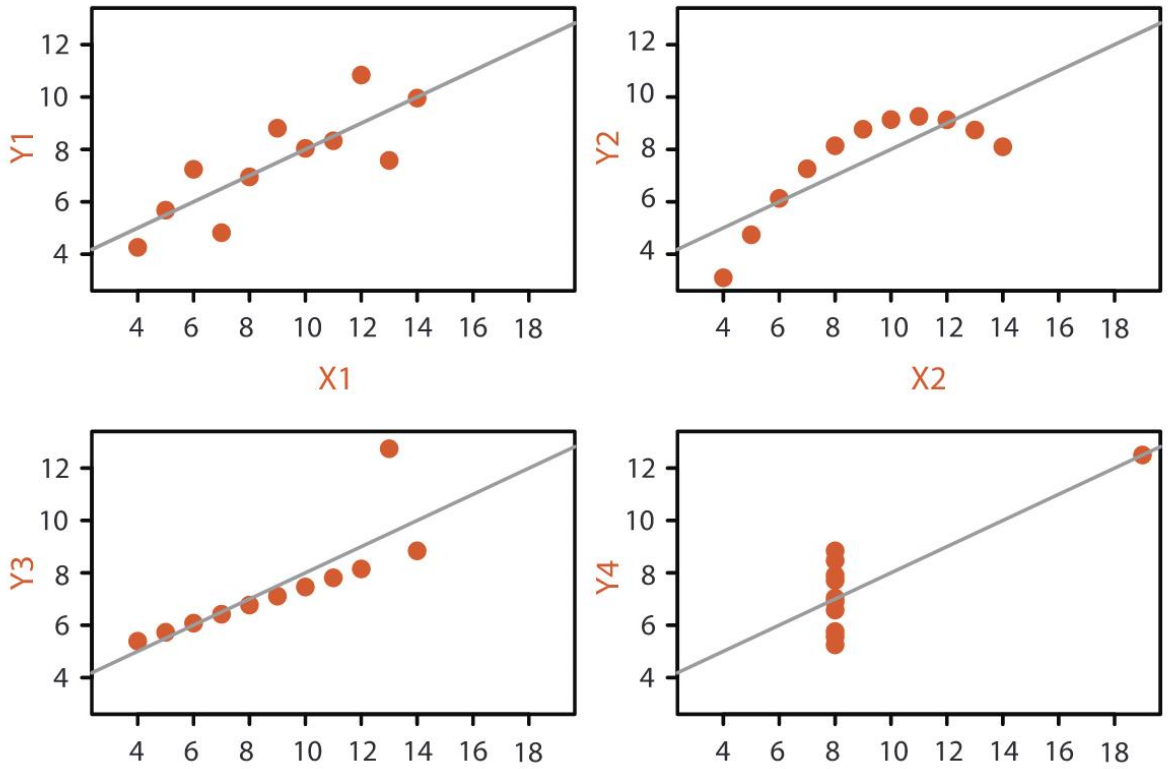
- Summaries can lose information, details matter!
  - Confirm expected
  - Find unexpected patterns
  - Assess validity of data models

Anscombe's Quartet

Identical statistics	
x mean	9
x variance	10
y mean	7.5
y variance	3.75
x/y correlation	0.816

Why not Only Show the Summary Data?

- Computer-based visualization systems provide visual **representations of datasets** designed to help people carry out tasks more effectively.



Anscombe's Quartet

Identical statistics	
x mean	9
x variance	10
y mean	7.5
y variance	3.75
x/y correlation	0.816

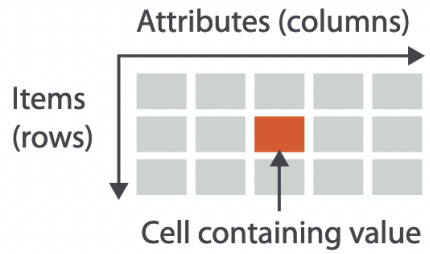
Utilization of Resources

- Three different kinds of resources to think about
- Computational limits
  - Computation time, system memory
- Display limits
  - Limited number of pixels on screen to use
  - Information density: ratio of space used to encode information vs whitespace
    - Tradeoff between clutter and wasting space
- Human limits
  - Human time, human memory, human attention

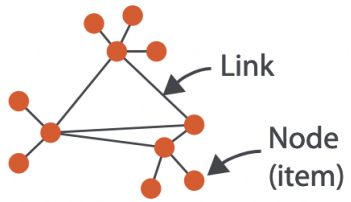
Datasets

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

➔ Tables



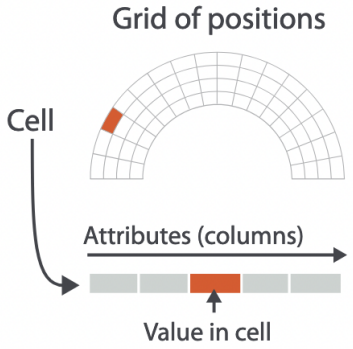
➔ Networks



➔ *Trees*

➔ Spatial

➔ Fields (Continuous)



➔ Geometry (Spatial)

What All Are Needed With Visualization?

- Data (wide variety)

What All Are Needed With Visualization?

- Data (wide variety)
- Algorithms
  - data mining
  - data analytics

What All Are Needed With Visualization?

- Data (wide variety)
- Algorithms
  - data mining
  - data analytics, AI/ML, statistical, ….
- Computer
  - run those algorithms
  - data storage

What All Are Needed With Visualization?

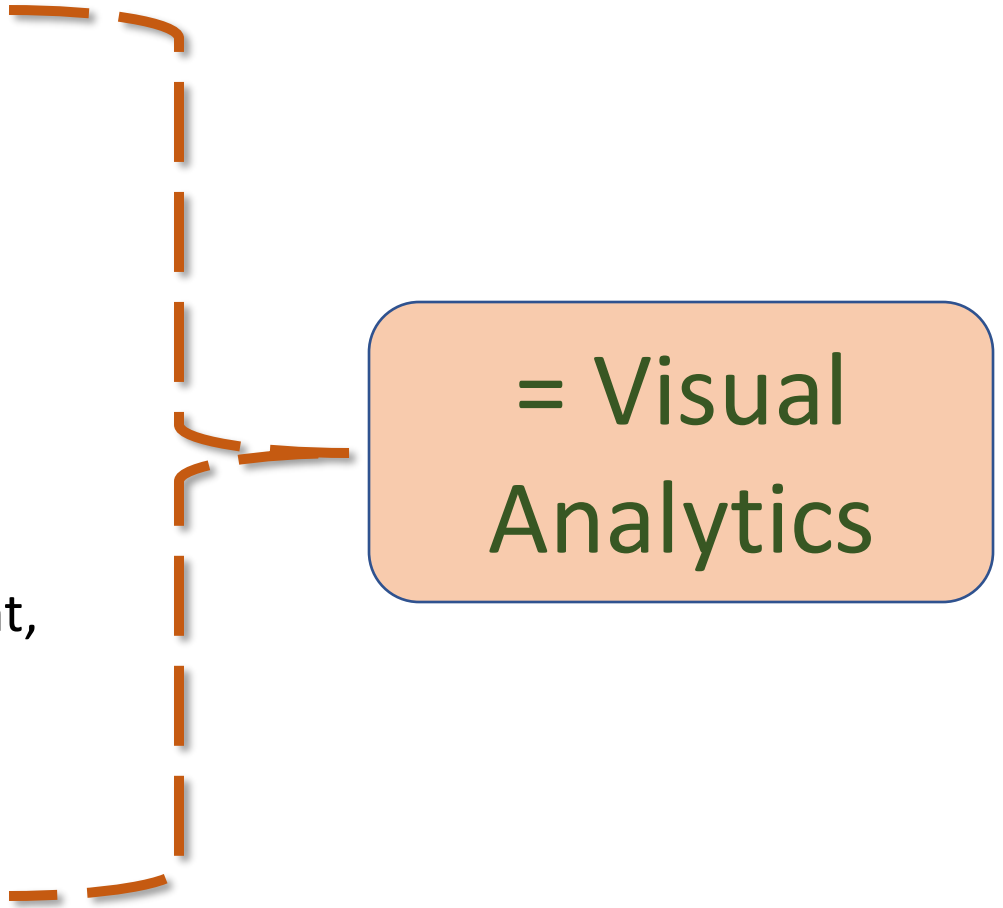
- Data (wide variety)
- Algorithms
  - data mining
  - data analytics, AI/ML, statistical, ….
- Computer
  - run those algorithms
  - data storage
- Humans
  - with a purpose/need to understand their data
  - endowed with cognitive faculties, creative thought, intuition
  - domain expertise

What All Are Needed With Visualization?

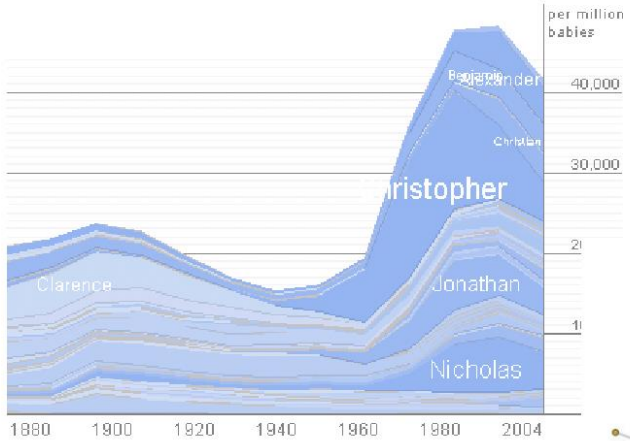
- Data (wide variety)
- Algorithms
  - data mining
  - data analytics, AI/ML, statistical, ….
- Computer
  - run those algorithms
  - data storage
- Humans
  - with a purpose/need to understand their data
  - endowed with cognitive faculties, creative thought, intuition
  - domain expertise
- Understanding of humans
  - perception, cognition, HCI issues
  - we can gain it through experimentation with humans

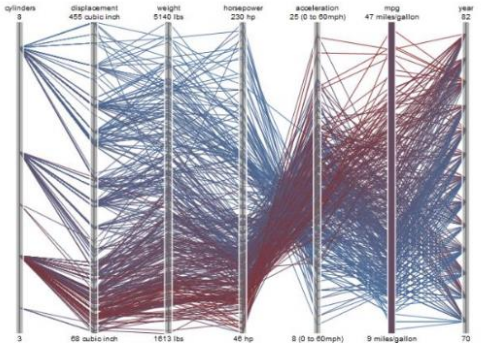
Visual Analytics

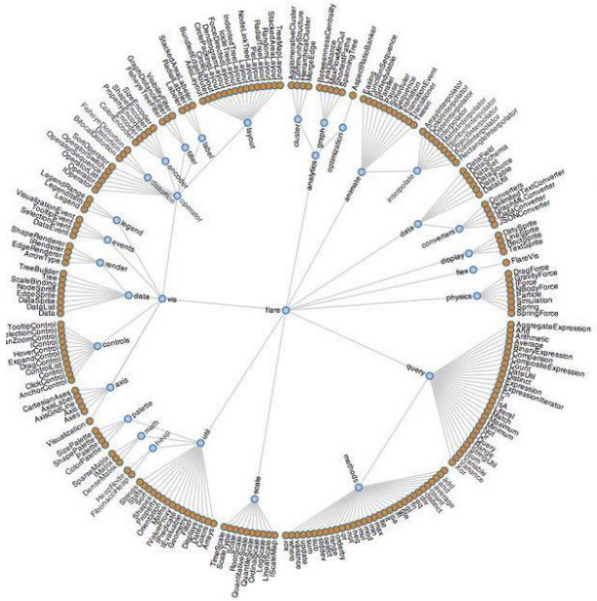
- Data (wide variety)
- Algorithms
  - data mining
  - data analytics, AI/ML, statistical, ….
- Computer
  - run those algorithms
  - data storage
- Humans
  - with a purpose/need to understand their data
  - endowed with cognitive faculties, creative thought, intuition
  - domain expertise
- Understanding of humans
  - perception, cognition, HCI issues
  - we can gain it through experimentation with humans

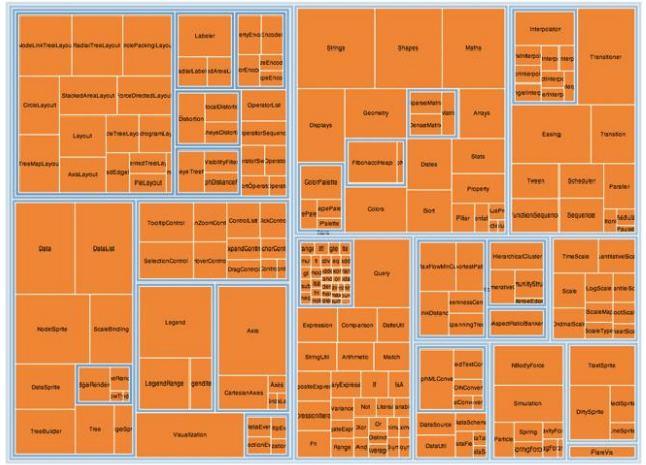


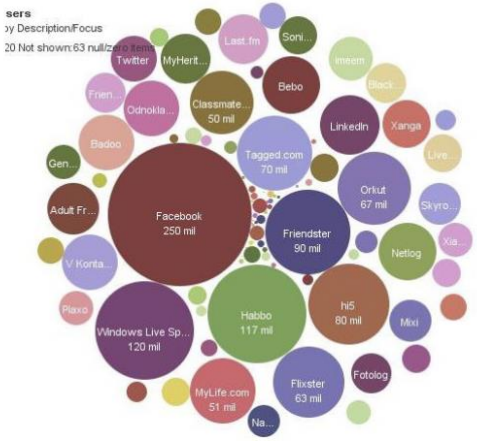
Visualization Can Be Beautiful



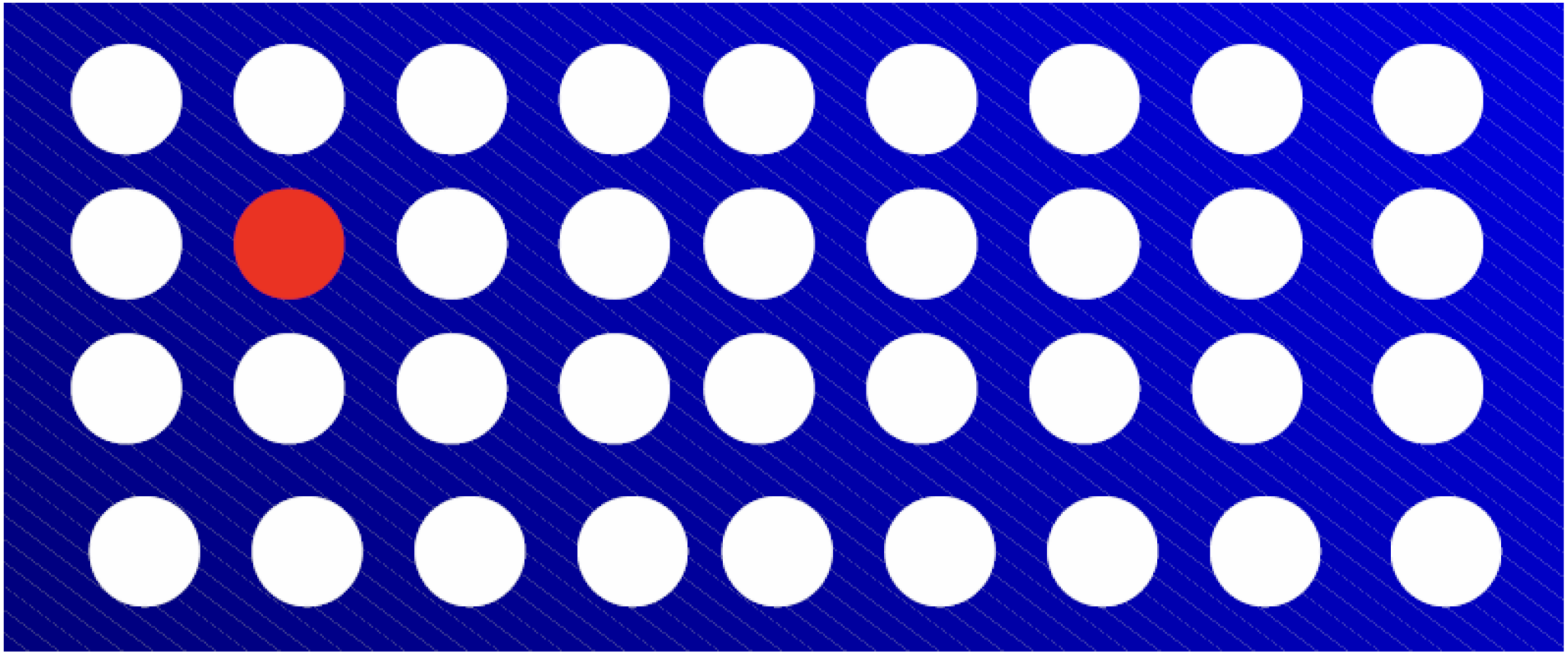






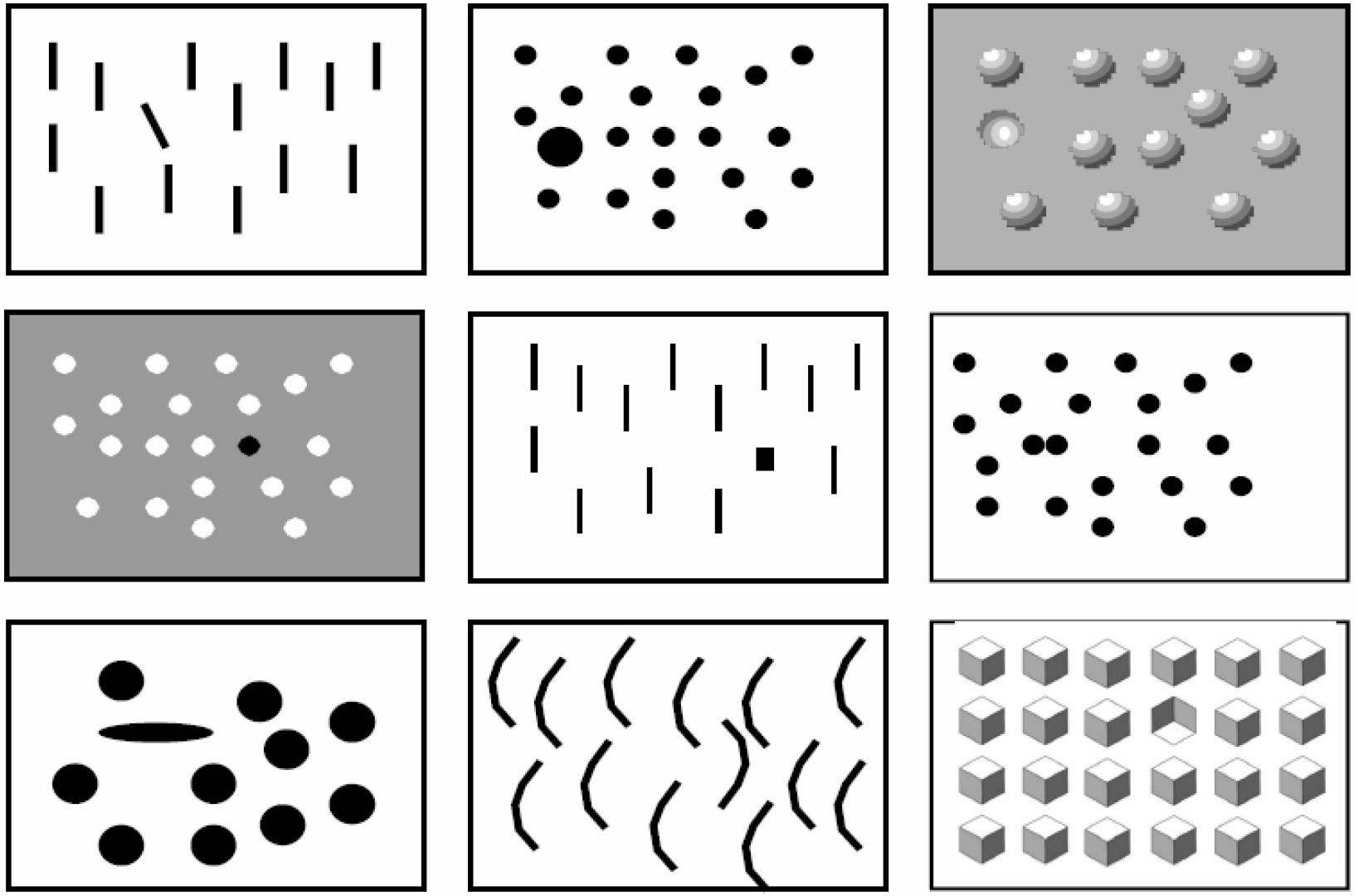


Visualization is Fast



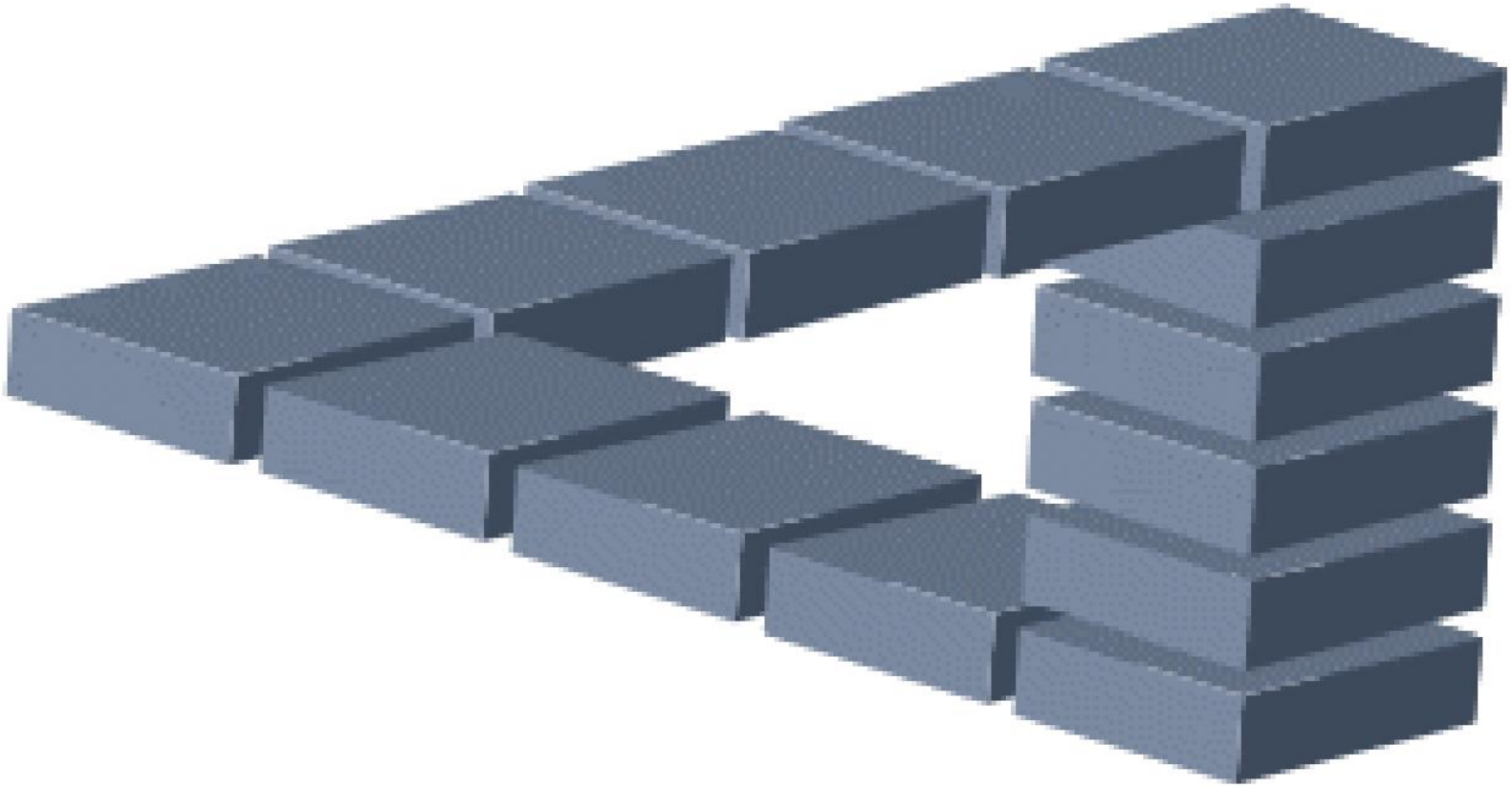
< ~200 ms to recognize the red dot

Visualization is Fast

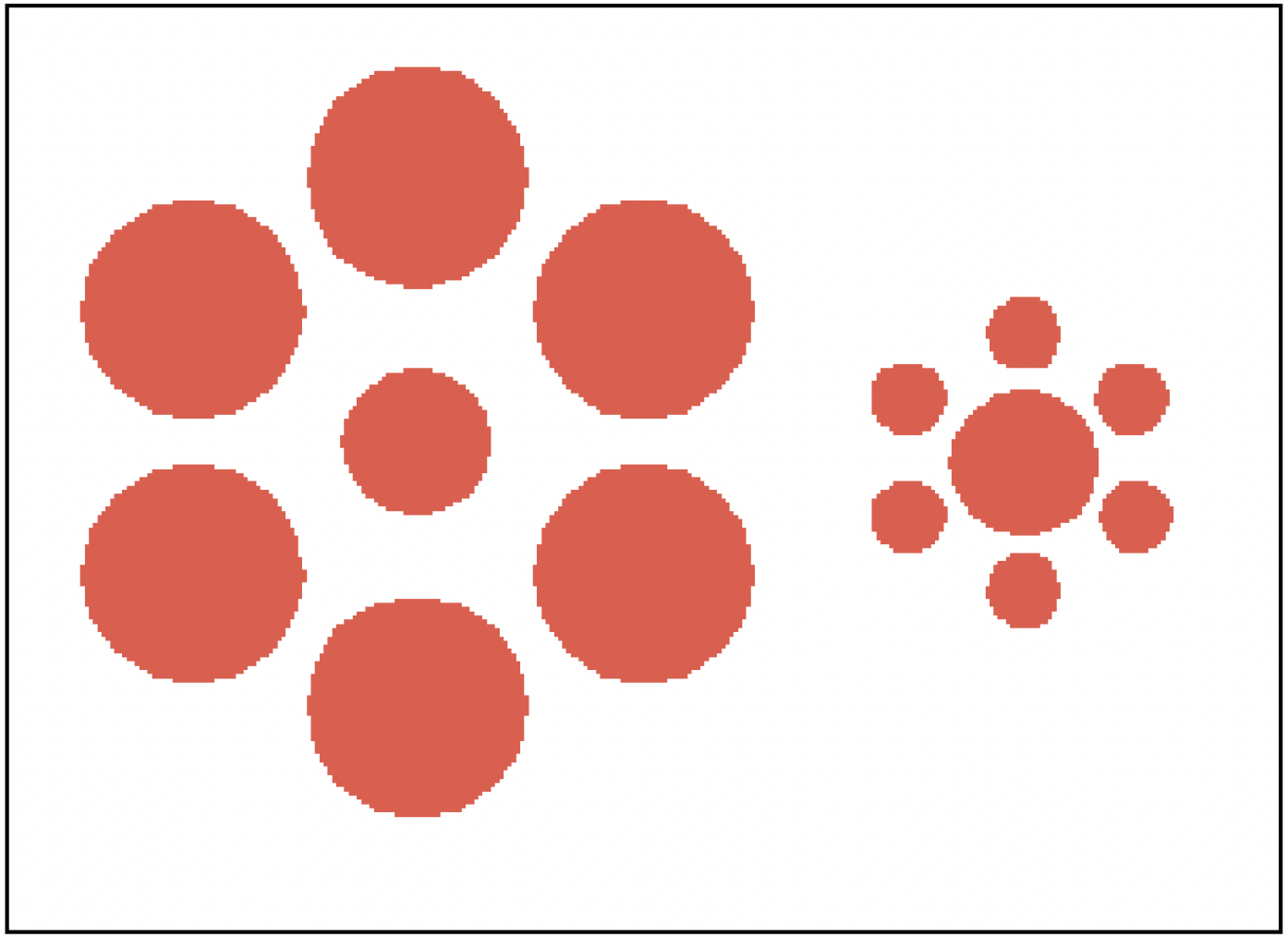


Visualization Can Be Interactive

Visualization Can Be Deceptive

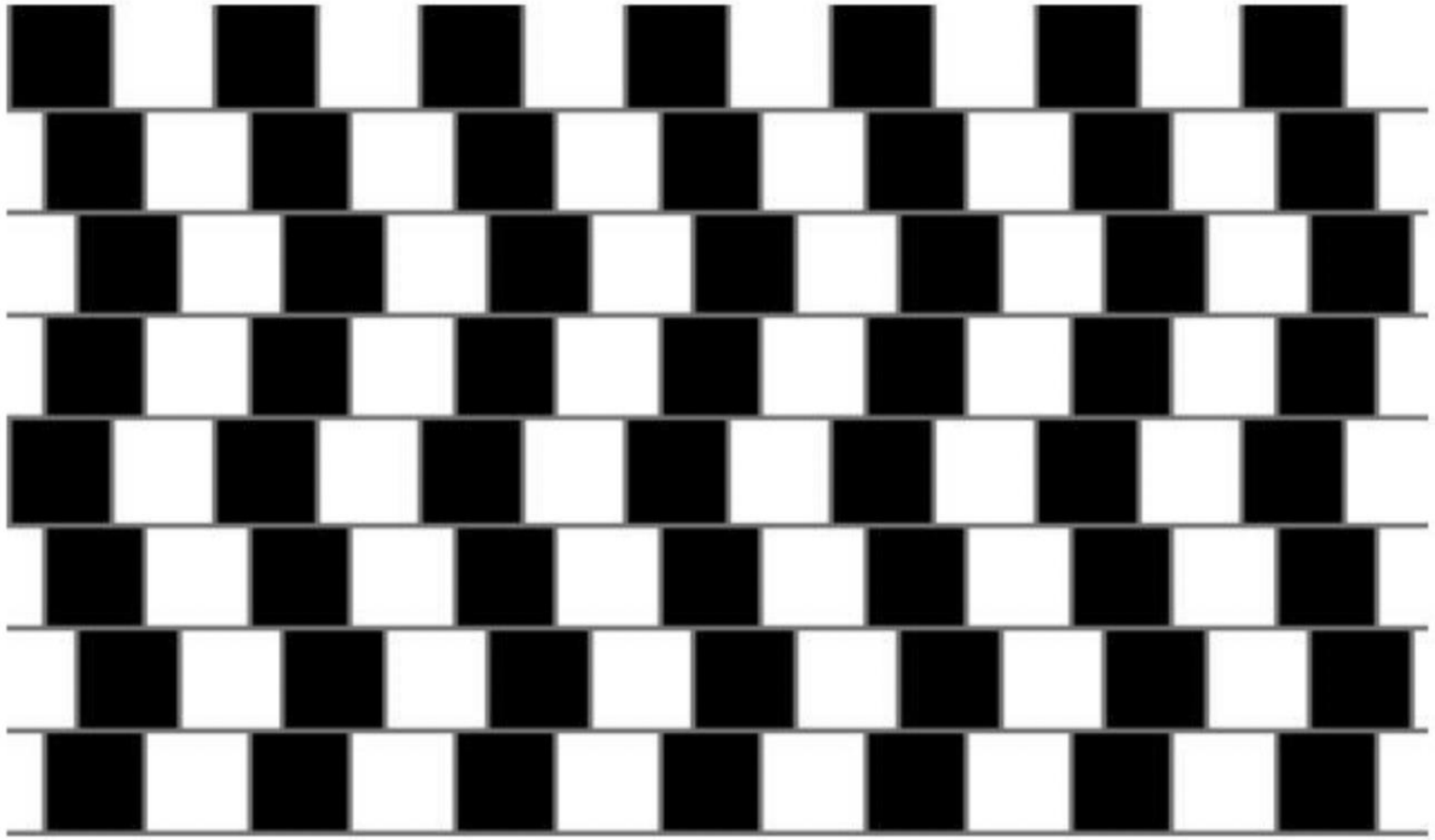


Visualization Can Be Deceptive



Which circle in the middle is larger?

Visualization Can Be Deceptive



Are the horizontal lines parallel or do they slope?