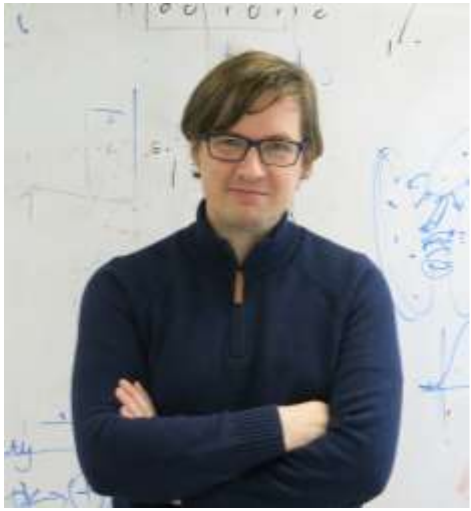
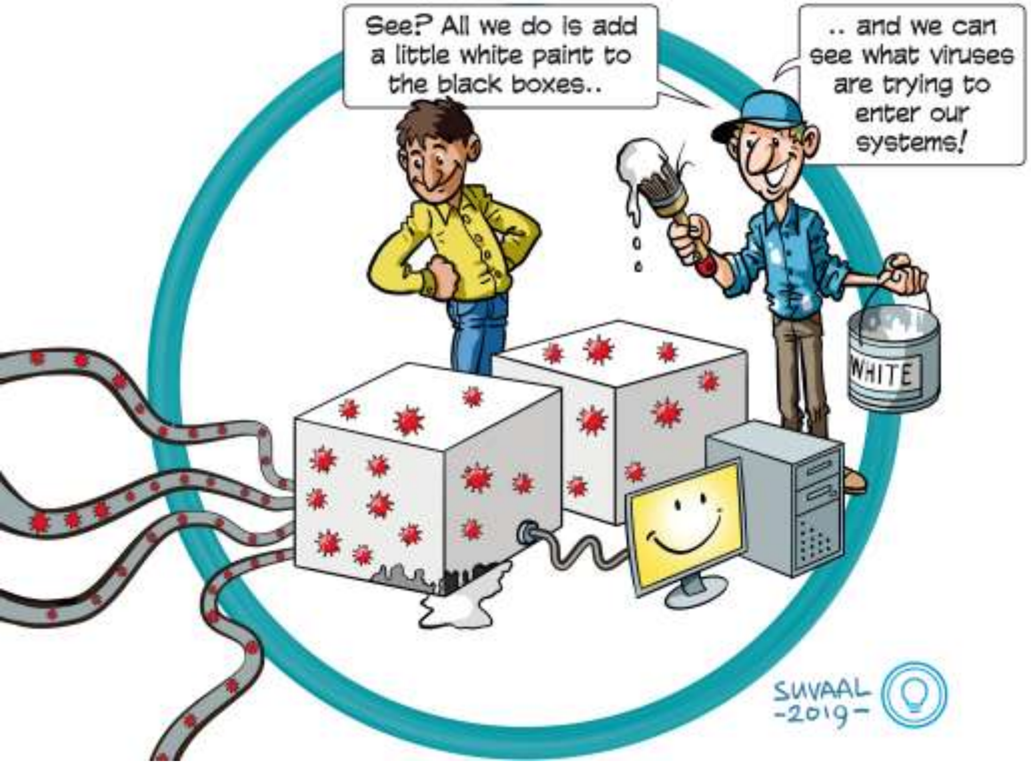
Datamining CSE2525

Nov 11, 2024

Sicco Verwer

Associate professor in Algorithms




See? All we do is add a little white paint to the black boxes..
.. and we can see what viruses are trying to enter our systems!
WHITE
SUVAAL -2019-

https://cyber-analytics.nl/

- Research:
  - algorithm design for learning interpretable models
  - applications in cyber security
  - awards: Veni, Vidi grants, Test-of-Time award

- Teaching:
  - cyber data analytics
  - AI for software reverse engineering
  - data mining
  - Optimization for ML

Avishek Anand

- Associate Professor at the Web Information Systems (ST)
- Topics: Information retrieval, NLP, Explainable AI

- Teaching: Information Retrieval, NLP, Data mining

- Topics covered in this course
  - Text Data Mining
    - How do we mine massive collections of text data ?
    - Word embeddings, indexing text
  - Graph data mining
    - How do we mine large graphs ?
    - Graph embeddings, graph analysis

Nergis Tömen

- Assistant Professor at Intelligent Systems (INSY)
- Topics: Biologically-inspired machine vision, neuromorphic computing
- Labs:

  Computer Vision Lab (member)
  Biomorphic Intelligence Lab (director)
  Biomedical Intervention Optimisation Lab (director)

- Teaching:
  - (MSc) Seminar Computer Vision by Deep Learning
  - (MSc) Machine Learning 2
  - (BSc) Data Mining

- Topics covered in this course:
  - Matrices, PCA, Matrix decomposition, Recommender systems

Today's goals

- Course goals
- Course logistics
- Course content overview
- What data mining is all about
- A word of caution

Course goals

- In this course you learn about Data Mining:
  - Several key algorithms, you must know:
    - How to implement them
    - Their strengths and weaknesses in practice
    - Why and when to use these in practice
  - Core concepts:
    - What they are – including theory –
    - Which concepts exist for different data types
  - Practical skills:
    - What concept to use for a given problem
    - How to succesfully apply algorithms in practice

Three Verticals

Distances

Matrices

Counting

Schedule	Week	Day	Date	Time	Topic	Lab Assignment Deadline
Lecture 1	2.1	Mon	Nov 11	13:45-15:45	Introduction	Lab 1: Anomaly detection
Lecture 2	2.1	Thu	Nov 14	10:45-12:45	Anomaly detection (DTW for Lab 1)	
Lecture 3	2.2	Mon	Nov 18	13:45-15:45	Distances (use case discussion)	
Lecture 4	2.2	Thu	Nov 21	10:45-12:45	Matrices (PCA for Lab 1)	
Lecture 5	2.3	Mon	Nov 25	13:45-15:45	Embeddings	
Lecture 6	2.3	Thu	Nov 28	10:45-12:45	Clustering (for Lab 2)	Lab 1 due date
Lecture 7	2.4	Mon	Dec 2	13:45-15:45	Discrimination (discussion)	Lab 2: Graph Clustering
Lecture 8	2.4	Thu	Dec 5	10:45-12:45	Invited lecture?	
Lecture 9	2.5	Wed	Dec 11	13:45-15:45	Graph Mining	
Lecture 10	2.5	Thu	Dec 12	10:45-12:45	MinHashing (for Lab 3)	
Lecture 11	2.6	Mon	Dec 16	13:45-15:45	Indexing	
Lecture 12	2.6	Thu	Dec 19	10:45-12:45	Sketching	Lab 2 due date
Lecture 13	2.7	Mon	Jan 6	13:45-15:45	NMF (for Lab 3)	Lab 3: Hashing/NMF
Lecture 14	2.7	Thu	Jan 9	10:45-12:45	Recommender systems (for Lab 3)	
Lecture 15	2.8	Mon	Jan 13	13:45-15:45	Manifold learning	
Lecture 16	2.8	Thu	Jan 16	10:45-12:45	Data Visualization (discussion)	
Lecture 17	2.9	Mon	Jan 20	13:45-15:45	Exam summary slides/Q&A	
Lecture 18	2.9	Thu	Jan 23	10:45-12:45	Mock exam answers/Q&A	Lab 3 due date
Exam	2.10	Mon	Jan 27	13:30-16:30	Weblab exam	

Teaching methods

- Lectures: 18
  - 13 content lectures
  - 2 invited lectures
  - 1 Intro
  - 2 Q&A
- Labs: 3
- Homework Assignments: 6

Lecture schedule

- Complete schedule: on Brightspace

- **Older lectures are recorded as backup at Collegerama that will be used sometimes in the flipped classroom**

- **E.g. On 24.11. - The lecture on distances will be flipped**

- *What is a flipped classroom ?*
  - *Please watch the video in Collegerama before come to class*
  - *In the lecture we do case studies – how do you apply what you have learnt in real-world scenarios ?*

Course material

- Required - Brightspace:
  - Lecture slides (after each class)
  - Lab exercises (beginning of the week)
  - Reading materials (book chapters and selected papers)

- Content from 2 books:
  - Mining of Massive Datasets
  - Data Mining

  - *Both are fully available through the TU Delft digital library!*
  - *Selected Chapters will be uploaded to Brightspace*

Lab sessions

- **Mandatory**

- 3 topics
- 9 sessions

- Lab sessions on Friday afternoon
- Assistance and feedback at lab session
  - Queue (https://queue.tudelft.nl/requests)
  - Mattermost (https://mattermost.tudelft.nl/)
  - Answers EWI with tag CSE2525 (https://answers.ewi.tudelft.nl/)
  - Kaggle (https://www.kaggle.com)
  - Weblab (https://weblab.tudelft.nl/)
  - Peer (https://peer.tudelft.nl/)

- Make sure you have a recent version of Python, including Numpy, Scipy, Pandas, Seaborn, Matplotlib on your own computer!

Lab sessions

Please, do not use e-mail!
They will not be answered.

- **Mandatory**

- 3 topics
- 9 sessions

- Lab sessions on Friday afternoon
- Assistance and feedback at lab session
  - Queue (https://queue.tudelft.nl/requests)
  - Mattermost (https://mattermost.tudelft.nl/)
  - Answers EWI with tag CSE2525 (https://answers.ewi.tudelft.nl/)
  - Kaggle (https://www.kaggle.com)
  - Weblab (https://weblab.tudelft.nl/)
  - Peer (https://peer.tudelft.nl/)

- Make sure you have a recent version of Python, including Numpy, Scipy, Pandas, Seaborn, Matplotlib on your own computer!

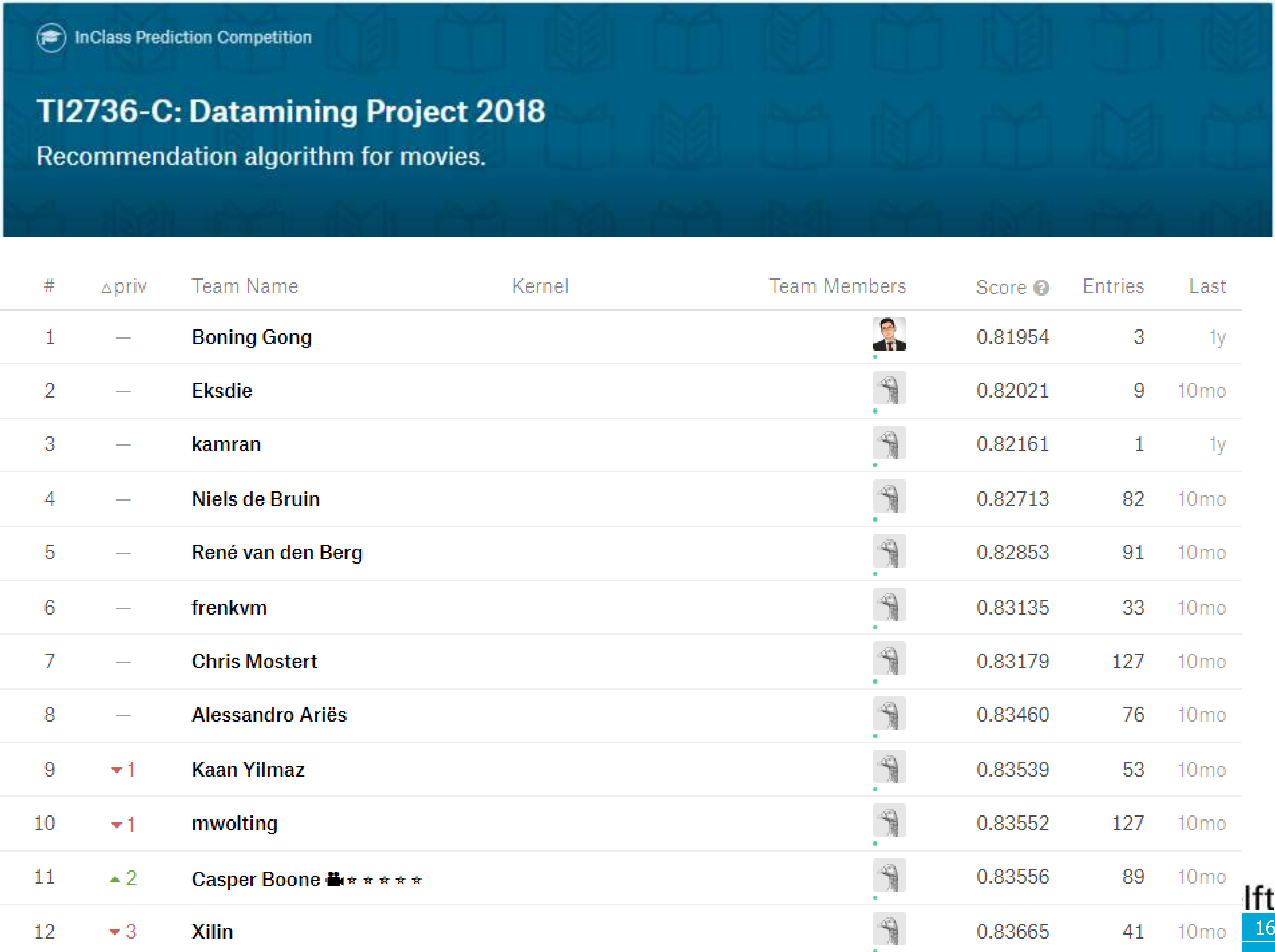
Lab sessions

- 3 topics

- Each lab topic has three components
  - **Algorithm implementation:** Distances, Matrices, Counting
  - **Building a pipeline:** Data transformation, analysis, and visualization
  - **Kaggle competition** [Bonus Points]

- Example - Topic 1 (this Friday): Anomaly detection
  - Algorithms to be implemented – DTW, PCA
  - Build an *anomaly detection pipeline* and evaluate its performance
  - Kaggle competition

- Labs in student pairs! Pair up as soon as possible, and register on Brightspace.



InClass Prediction Competition

TI2736-C: Datamining Project 2018

Recommendation algorithm for movies.

#	∆priv	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	Boning Gong			0.81954	3	1y
2	—	Eksdie			0.82021	9	10mo
3	—	kamran			0.82161	1	1y
4	—	Niels de Bruin			0.82713	82	10mo
5	—	René van den Berg			0.82853	91	10mo
6	—	frenkvm			0.83135	33	10mo
7	—	Chris Mostert			0.83179	127	10mo
8	—	Alessandro Ariës			0.83460	76	10mo
9	▼1	Kaan Yilmaz			0.83539	53	10mo
10	▼1	mwolting			0.83552	127	10mo
11	▲2	Casper Boone 🎥★ ★ ★ ★ ★			0.83556	89	10mo
12	▼3	Xilin			0.83665	41	10mo

Lab Evaluation

- Lab Evaluation – 30% of your final grade

- *Automatic evaluation* of the algorithmic component

- *Peer review* of the pipeline component
  - Please do your peer reviews, penalty if not completed
  - Your submissions will get 4 reviews
  - We will double-check the quality of the reviews

https://peer.tudelft.nl/courses

Lab Evaluation

- Lab Evaluation – 30% of your final grade

- *Automatic evaluation* of the algorithmic component

- *Peer review* of the pipeline component

- Kaggle competition – should beat our baselines to get bonus points

- No solutions! Ask for help during labs.
- Top 3 Kaggle submissions will be shared and asked to present

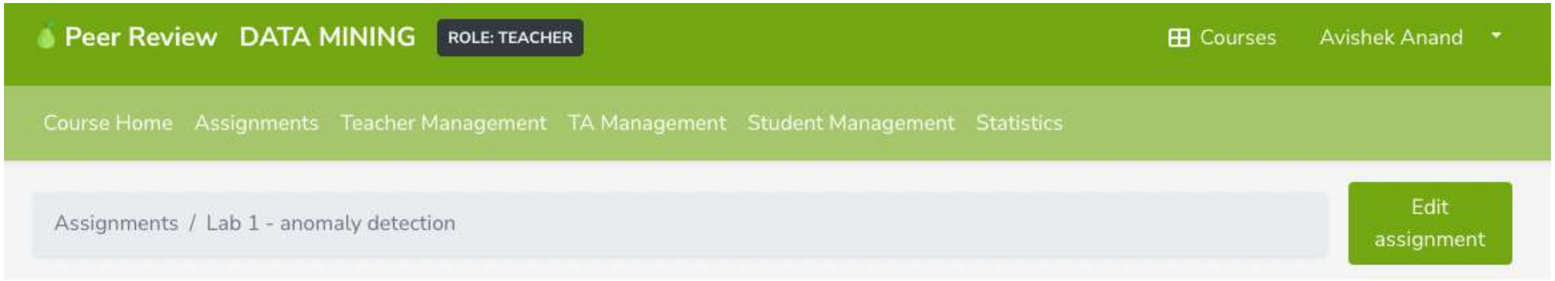
Lab Evaluation

- Lab Evaluation – 30% of your

No scikitlearn or other ML tool will be used, evertyhing is build from scratch!

- *Automatic evaluation* of the algorithmic component
- *Peer review* of the pipeline component
- Kaggle competition – should beat our baselines to get bonus points
- No solutions! Ask for help during labs.
- Top 3 Kaggle submissions will be shared and asked to present

How does Peer Review work ?



Assignments / Lab 1 - anomaly detection

Homework Assignments

- 6 of them
- Idea: Mostly descriptive questions and problems
  - Reflects the type and hardness of questions you can expect in the final exam
- Solutions will be given
- NOT be graded or discussed in the lab

Final Exam

- No (partial) transfer from previous years

- WebLab (https://weblab.tudelft.nl) exam: 70%
  - One resit

- Wednesday Jan 31, 2024
  - Weblab exam (Osiris + weblab registration)

- Open and multiple-choice questions
- No programming questions this year!

- ***Closed book*** – calculator is allowed

*https://mytimetable.tudelft.nl is authoritative

Course changes

- We planned to remove 25% of the older content

- Removed content:
  - Graph cuts
  - Community detection

- New content:
  - High-dim data visualization

Expected prior knowledge

- Discrete mathematics:
  - sets, intersections, and unions
- Linear algebra:
  - matrix multiplication, projections, eigenanalysis
- Probability and statistics:
  - Gaussians, correlation, covariance
- Graph theory:
  - adjacency matrix, degree, clique, bipartite graph, shortest path
- Data structures:
  - hash tables and indexes
- Programming:
  - Python programming skills
- Machine learning:
  - basic algorithms: logistic regression, random forest, svm, …

Prior courses

- CSE1100/TI1206 Object-oriented programming
  CSE1305/TI1316 Algorithms and Data Structures
  CSE1200/TI1106M Calculus
  CSE1205/TI1206M Linear Algebra
  CSE1210/TI2216M Probability Theory and Statistics

- CSE2510 Machine Learning
  CSE2520/TI2736-B Big Data Processing

- Information only - not enforced

- You are responsible for your study success!

Feedback

- When:
  - Any time
- How:
  - E-mail: dm-cs-ewi@tudelft.nl
  - Anonymous evaluations (EvaSys/EvaTool)

Logistics summary

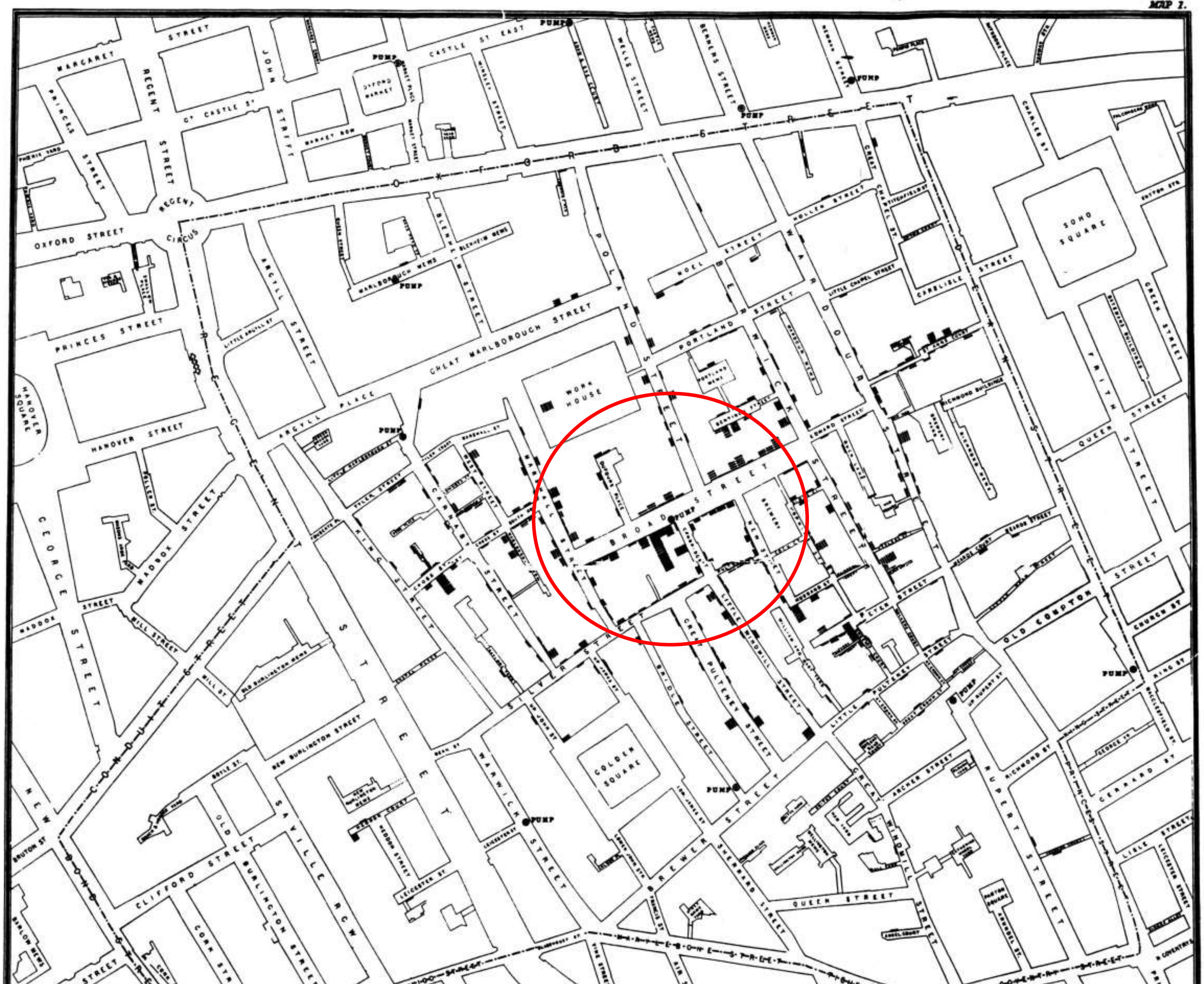
- Optional practicals
  - Questions through online tools and at sessions
- 3 mandatory Labs – peer review + automated tests + kaggle
- Closed-book Exam
- Only for feedback:
  - dm-cs-ewi@tudelft.nl

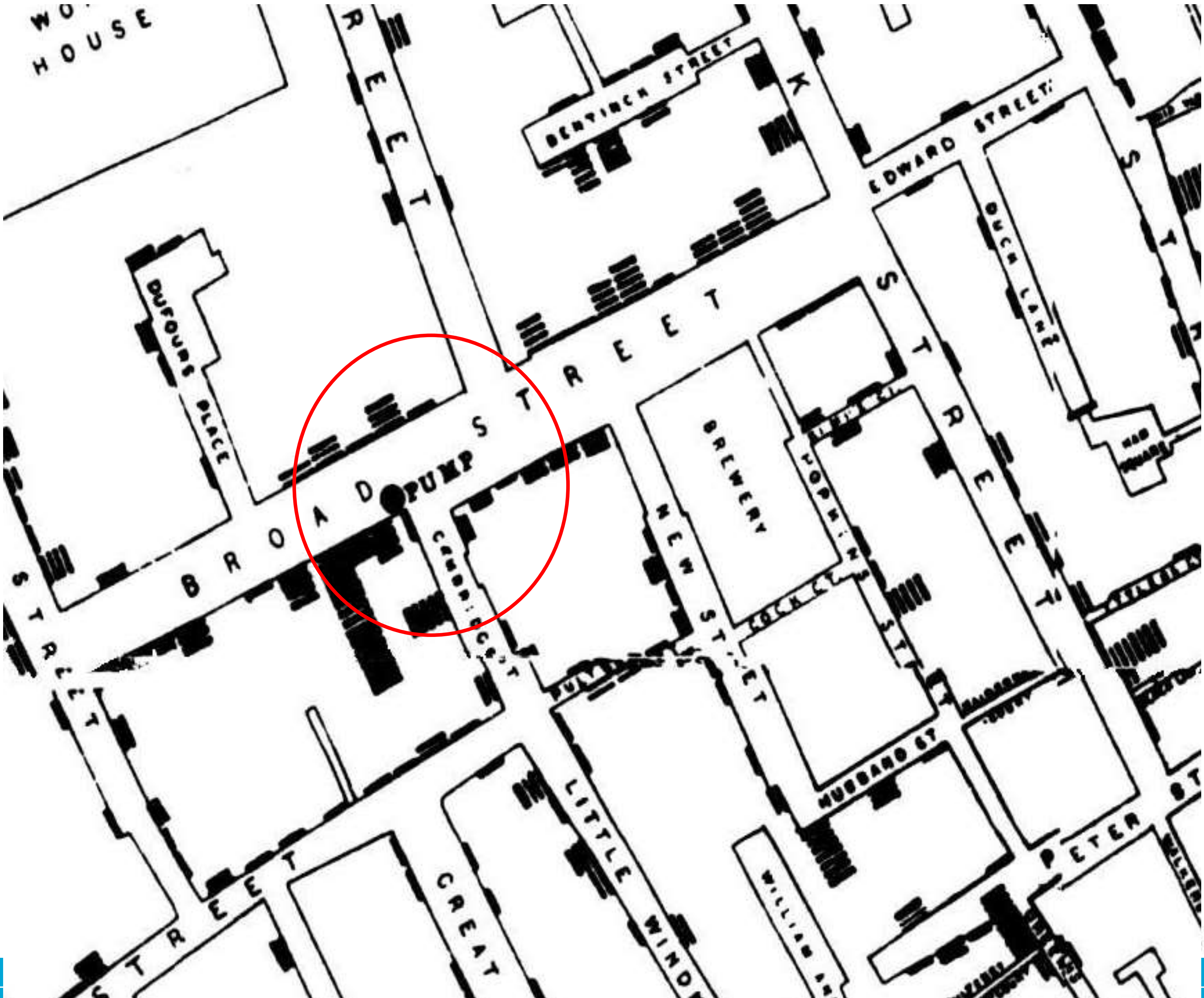
#SCHITTSCREEK
Pop

What is Data Mining?

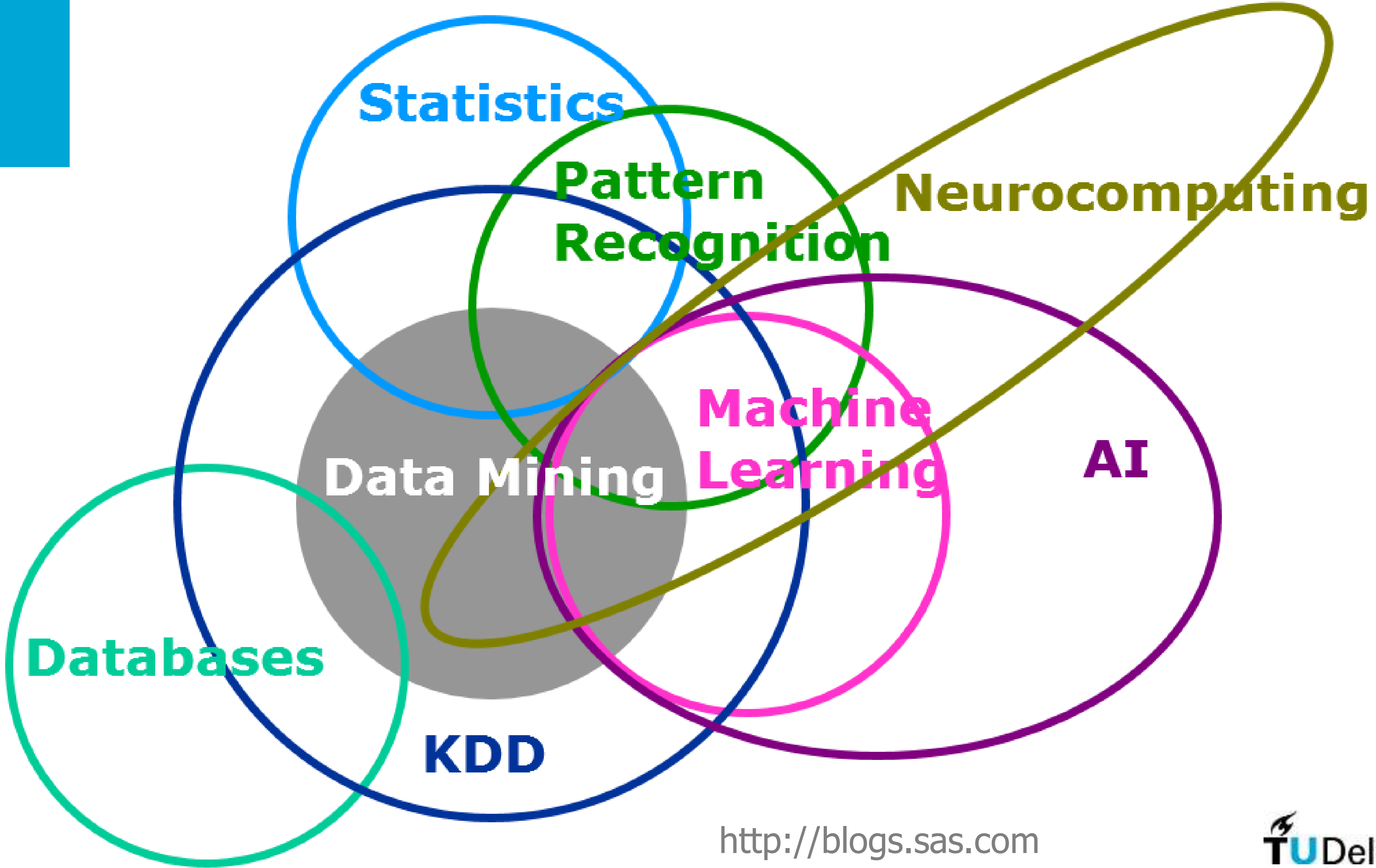
What is data mining?

- The first data miner?
- John Snow (1813-1858)
- Plotted cholera cases on a map of London


MAP 1.
OXFORD STREET
REGENT STREET
REGENT CIRCUS
MARGARET STREET
PRINCES STREET
HANOVER SQUARE
HANOVER STREET
GEORGE STREET
MADDOX STREET
CONDUIT STREET
NEW BOND STREET
BRUTON ST
CLIFFORD STREET
OLD BURLINGTON STREET
SAVILLE ROW
NEW BURLINGTON STREET
BOYLE ST.
MILL STREET
ARGYLL STREET
ARGYLL PLACE
GREAT MARLBOROUGH STREET
MARLBOROUGH MEWS
BLENHEIM STREET
BLENHEIM MEWS
CASTLE ST EAST
GT CASTLE ST
JOHN STREET
OXFORD MARKET
MARKET ROW
WELLS STREET
BERNERS STREET
CHARLES ST
SOHO SQUARE
FRITH STREET
GREEK STREET
DEAN STREET
POLAND STREET
NOEL STREET
BERWICK STREET
WARDOUR STREET
PORTLAND STREET
WORK HOUSE
BROAD STREET
PUMP
MARSHALL STREET
CARNABY STREET
KING STREET
SILVER STREET
GOLDEN SQUARE
BRIDLE STREET
PULTENEY STREET
LITTLE WINDMILL STREET
GREAT WINDMILL STREET
BREWER STREET
WARWICK STREET
RUPERT STREET
RICHMOND BUILDINGS
QUEEN STREET
OLD COMPTON STREET
CHURCH ST
GERRARD STREET
LISLE STREET
LEICESTER STREET
PRINCES STREET
PICCADILLY
ARCHER STREET
HEDDON STREET
VINE STREET
CORK STREET
BURLINGTON STREET
CARLISLE STREET
EDWARD STREET
PETER STREET
HOPKINS STREET
HUNGARY COURT
NEW STREET
SHERRARD STREET
GREAT CHAPEL ST
HOLLEN STREET
QUEEN STREET
RICHMOND ST
KING ST
MACCLESFIELD ST
COVENTRY


WORK HOUSE
BENTINCK STREET
EDWARD STREET
DUFOURS PLACE
BROAD STREET
PUMP
CAMBRIDGE ST
NEW STREET
BREWERY
HOPKINS STREET
DUCK LANE
COCK CT
HUSBAND ST
PETER ST
LITTLE WIND
GREAT
STREET

Data mining in context



http://blogs.sas.com

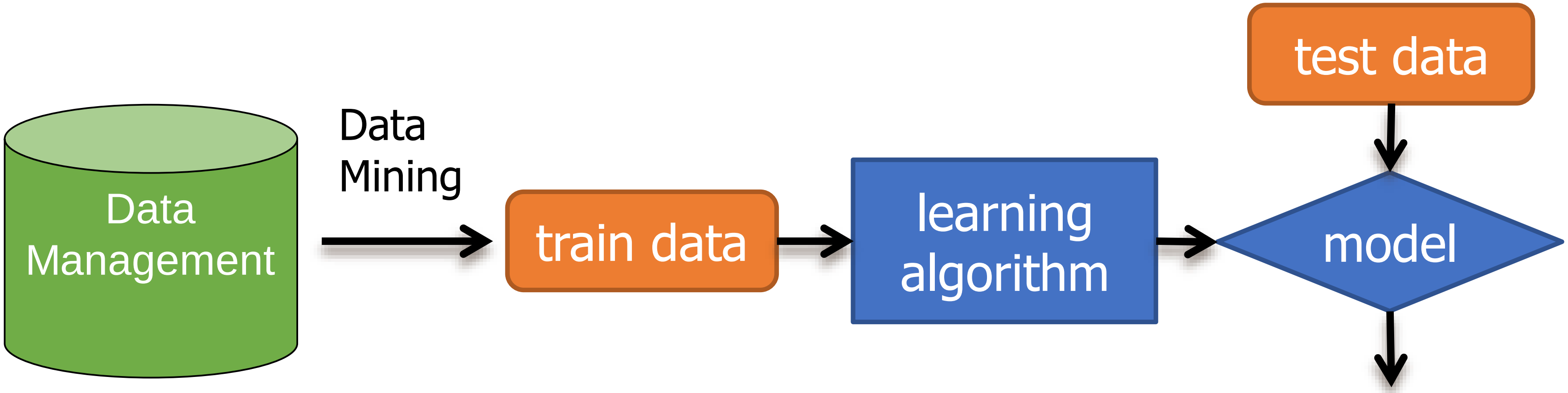
What is data mining?

- “..is the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information”
- "...is the development of models for data in order to extract information from that data.”
- "... is the process of analyzing data from different perspectives and summarizing it into useful information.”
- ".. is done by humans”

What is data mining?

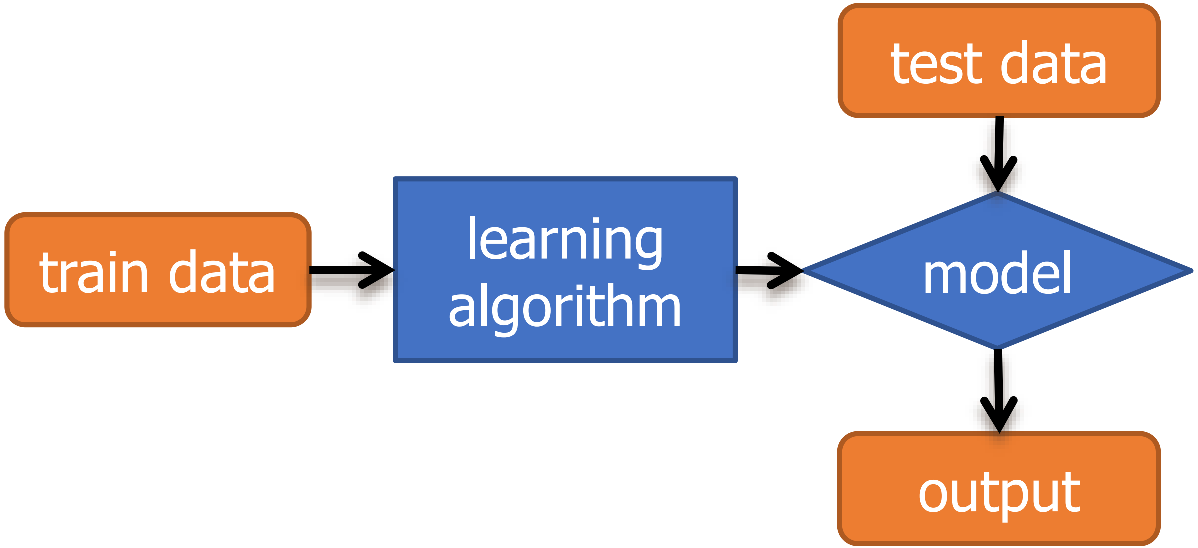
The primary goal of data mining is to extract useful information from a large volume of data and transform it into an understandable structure for further use.

Data Mining vs. Machine learning vs Data Management



Data Mining vs. Machine learning

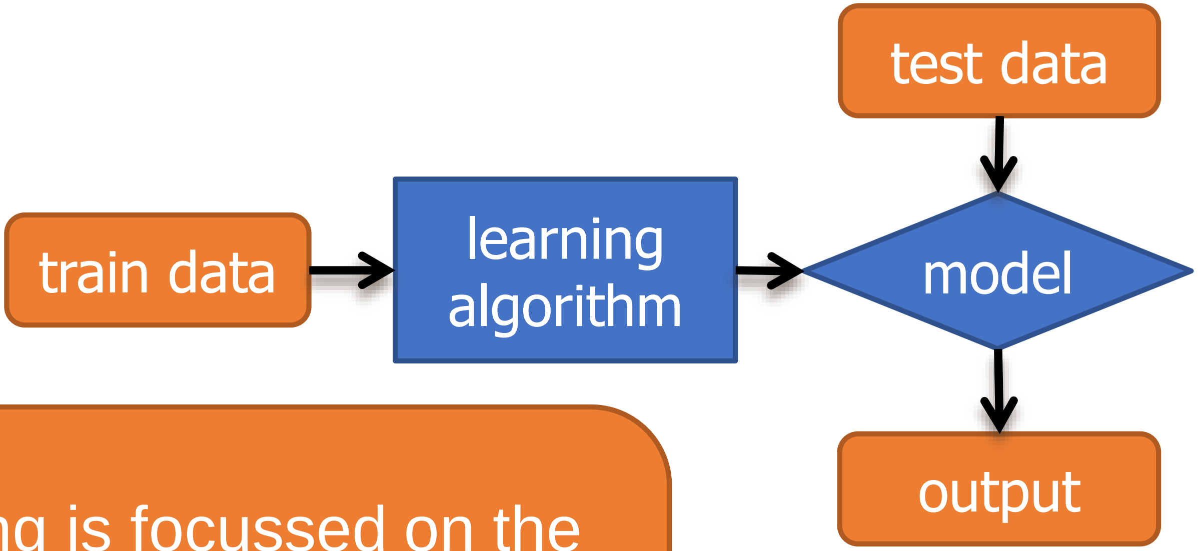
- Classic ML Approach:



  - take a huge data set
  - compute features
  - train a classifier
  - deploy the classifier on test

Data Mining vs. Machine learning

- Classic ML Approach:

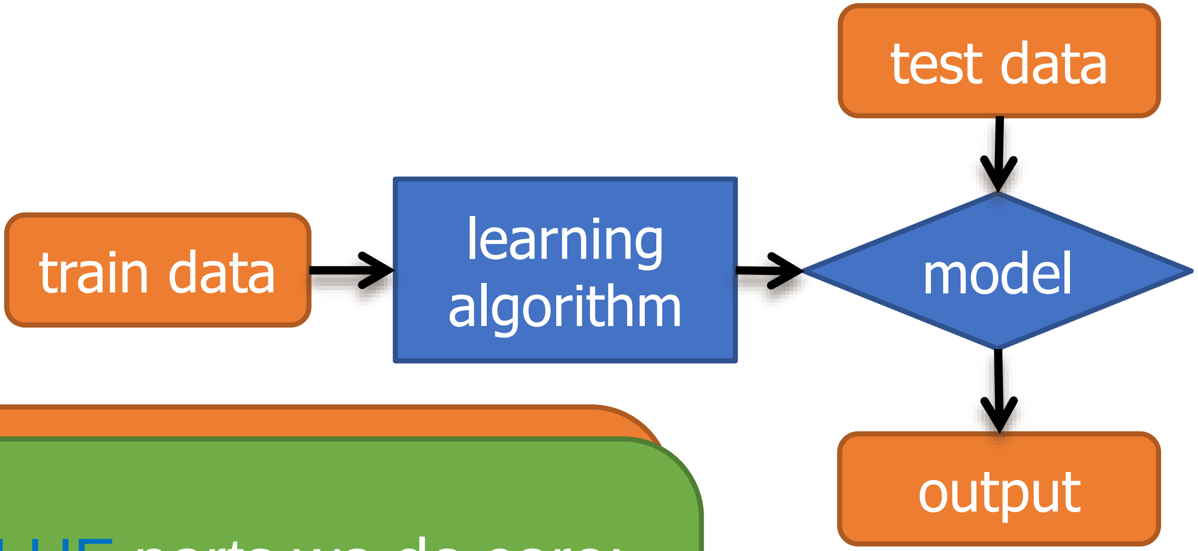


Data Mining is focussed on the ORANGE parts:

What to do with input and output?

Data Mining vs. Machine learning

- Classic ML Approach:



For the BLUE parts we do care:

What algorithm to use and how to run it on our data?

Data Mining: Healthcare system

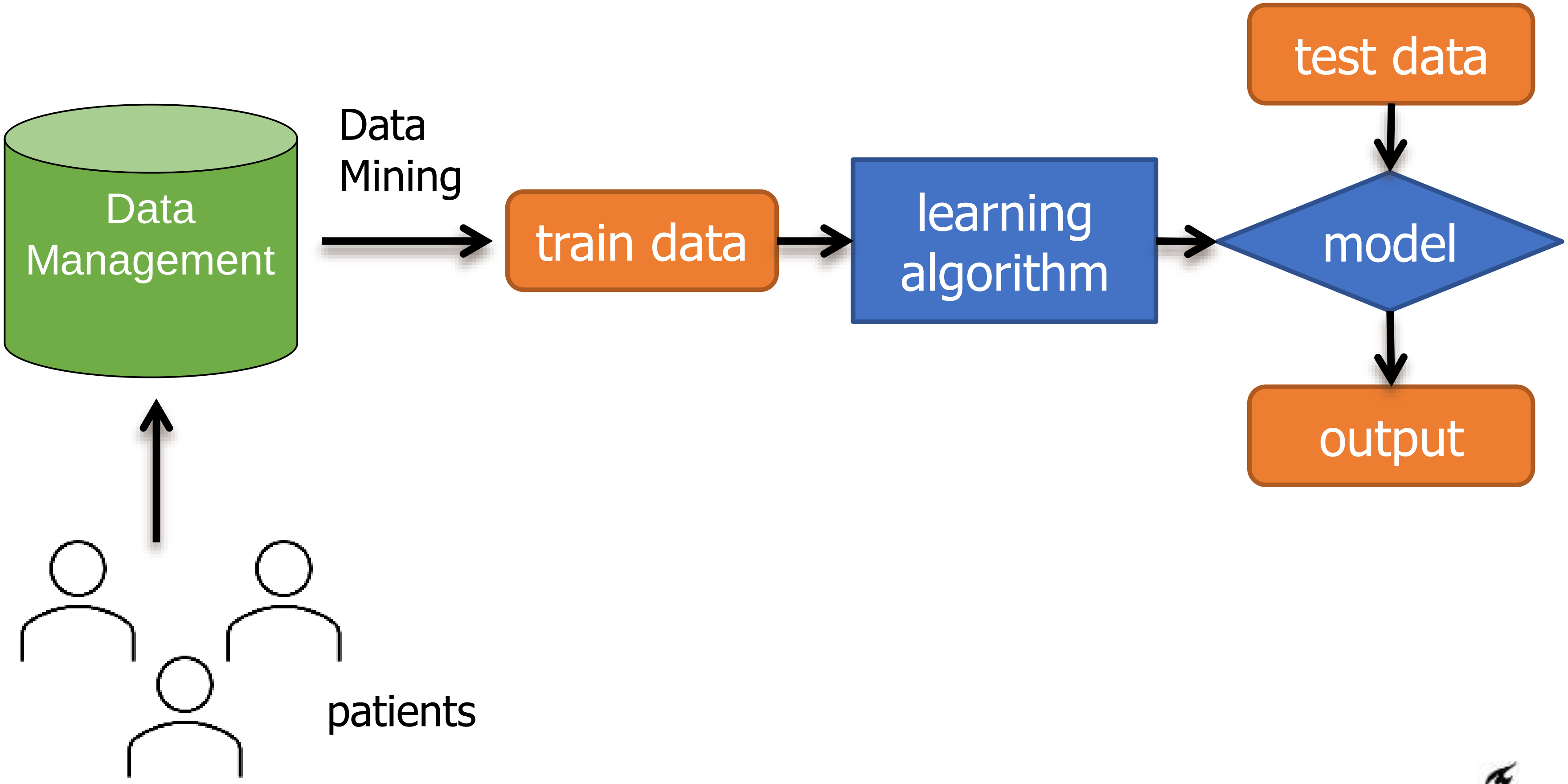
- **Objective**: To find/uncover patterns and correlations in patient data

- **Process**: The system analyzes a vast database of patient records, including symptoms, diagnostics, treatments, and outcomes.
  - Data mining techniques: clustering, association rule mining, and anomaly detection

- **Outcome**: The system identifies patterns such as common symptoms associated with particular diseases, effective treatments for specific conditions, and any anomalies like rare side effects of treatments

Patients with a certain combination of symptoms (e.g., fever, cough, and shortness of breath) often test positive for a specific respiratory illness.

Healthcare System



Data Management

Data Mining

train data

learning algorithm

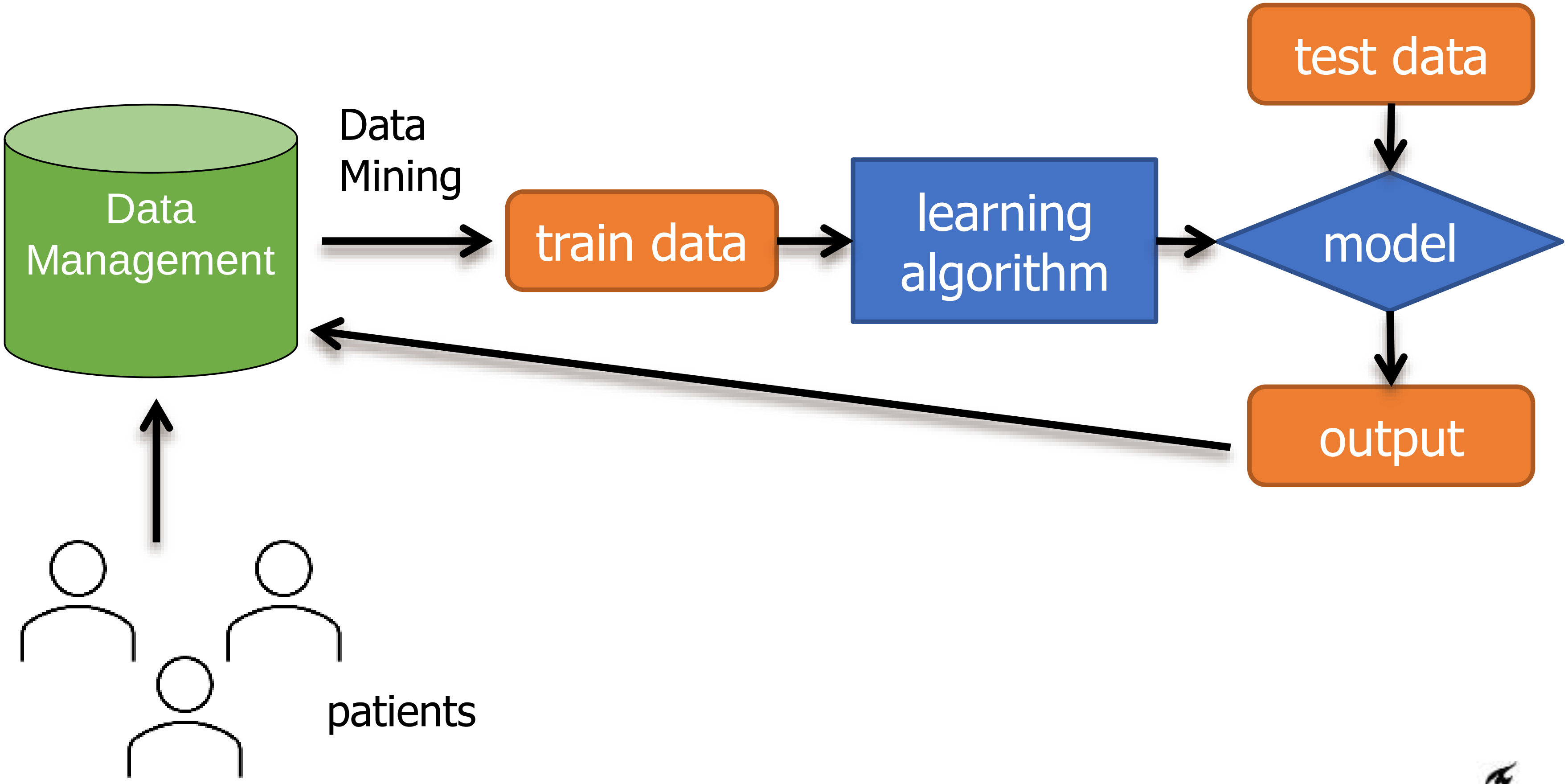
test data

model

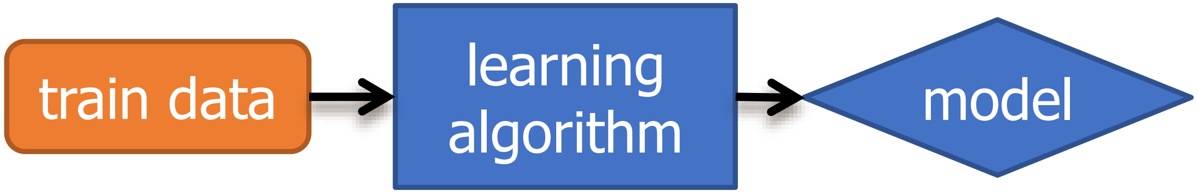
output

patients

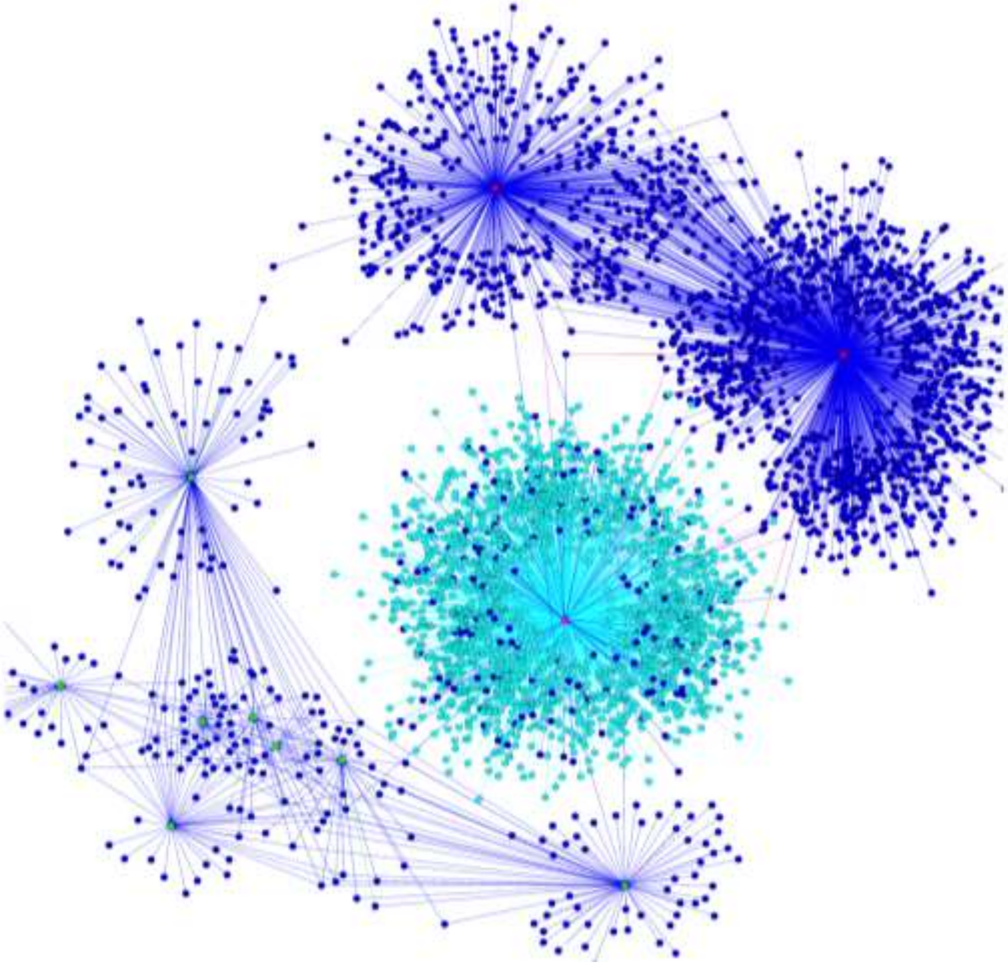
Healthcare System



Often learning is unsupervised



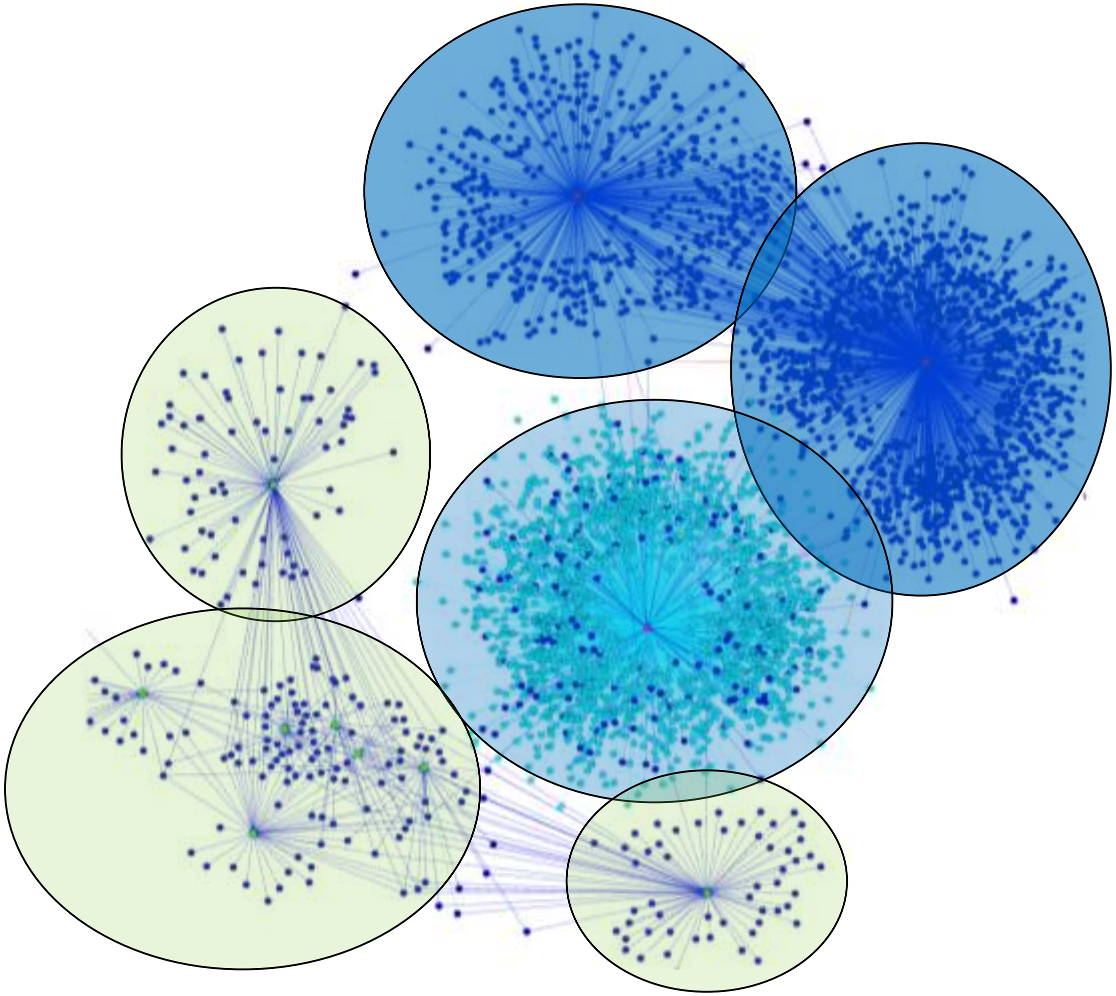
Example: Network/Graph data

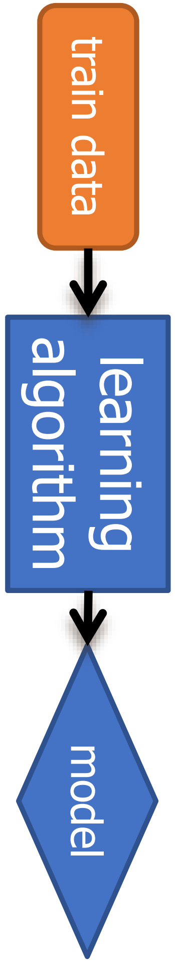


Beyond Labeling: Using Clustering to Build Network Behavioral Profiles of Malware Families. Azqa Nadeem, Christian Hammerschmidt, Carlos H. Ganan, Sicco Verwer. In Malware Analysis using Artificial Intelligence and Deep Learning, Springer, 2020. (Forthcoming)

Hybrid Connection and Host Clustering for Community Detection in Spatial-temporal Network Data. Mark Patrick Roeling, Azqa Nadeem, Sicco Verwer. In Machine Learning for Cybersecurity (MLCS), 2020

Community detection



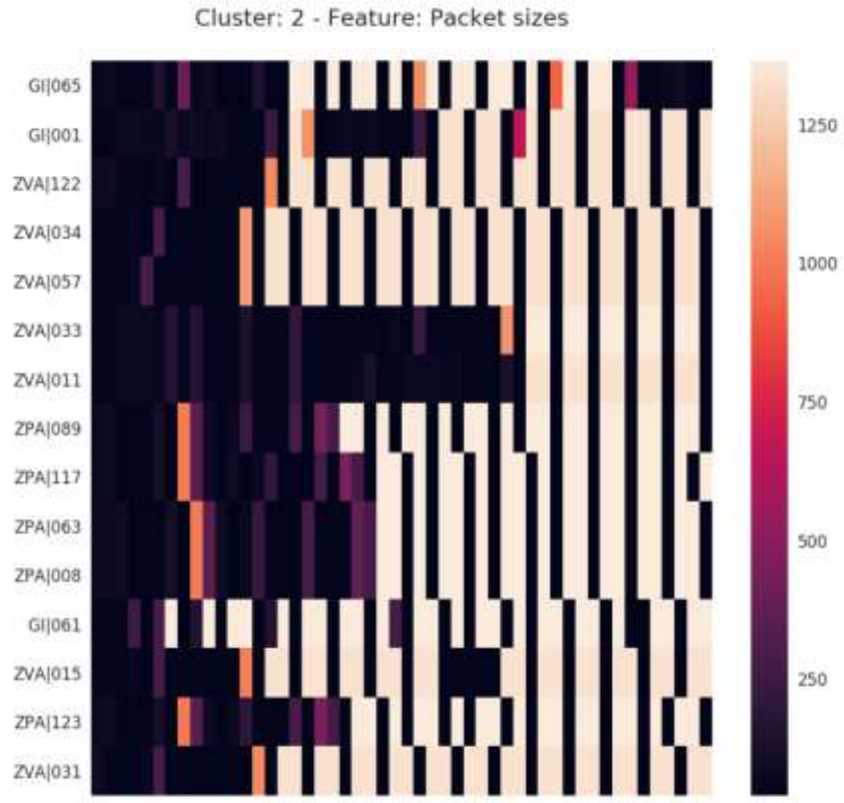


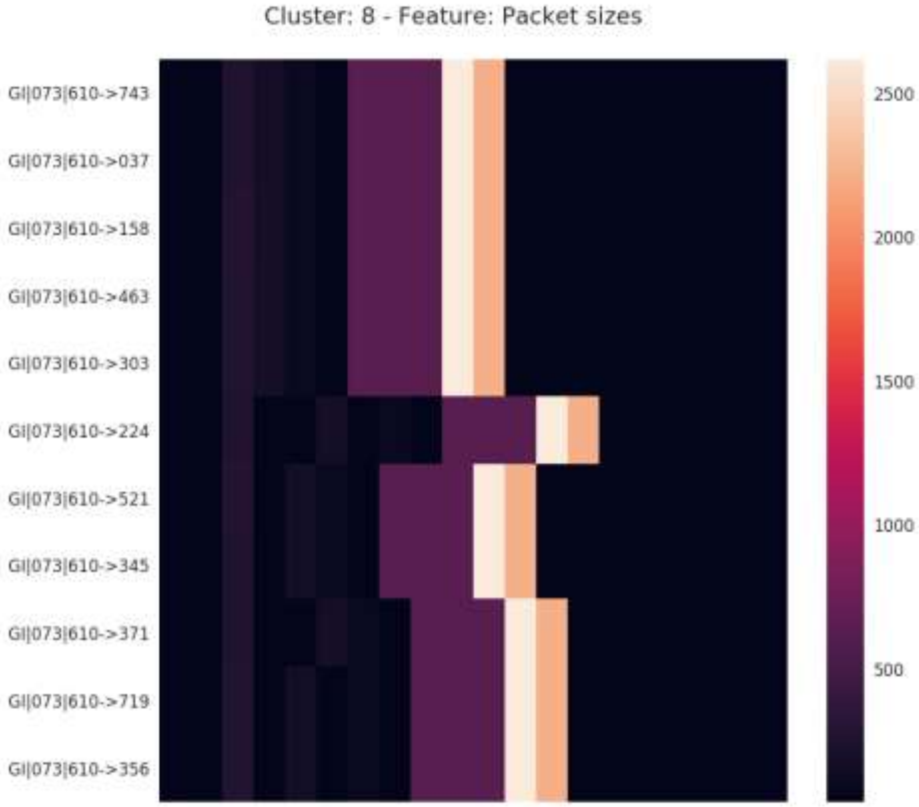
train data

learning algorithm

model

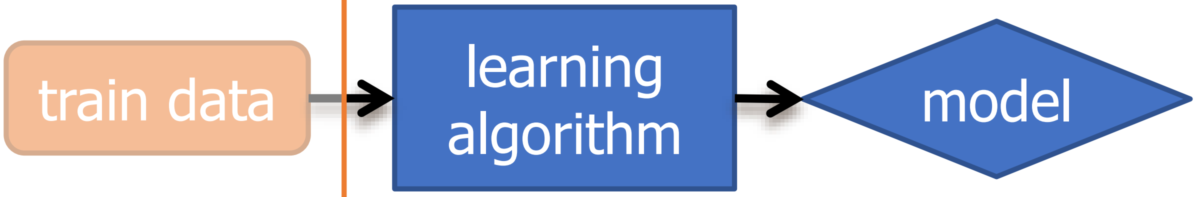
Use heatmaps to visualize groups


Cluster: 2 - Feature: Packet sizes


Cluster: 8 - Feature: Packet sizes

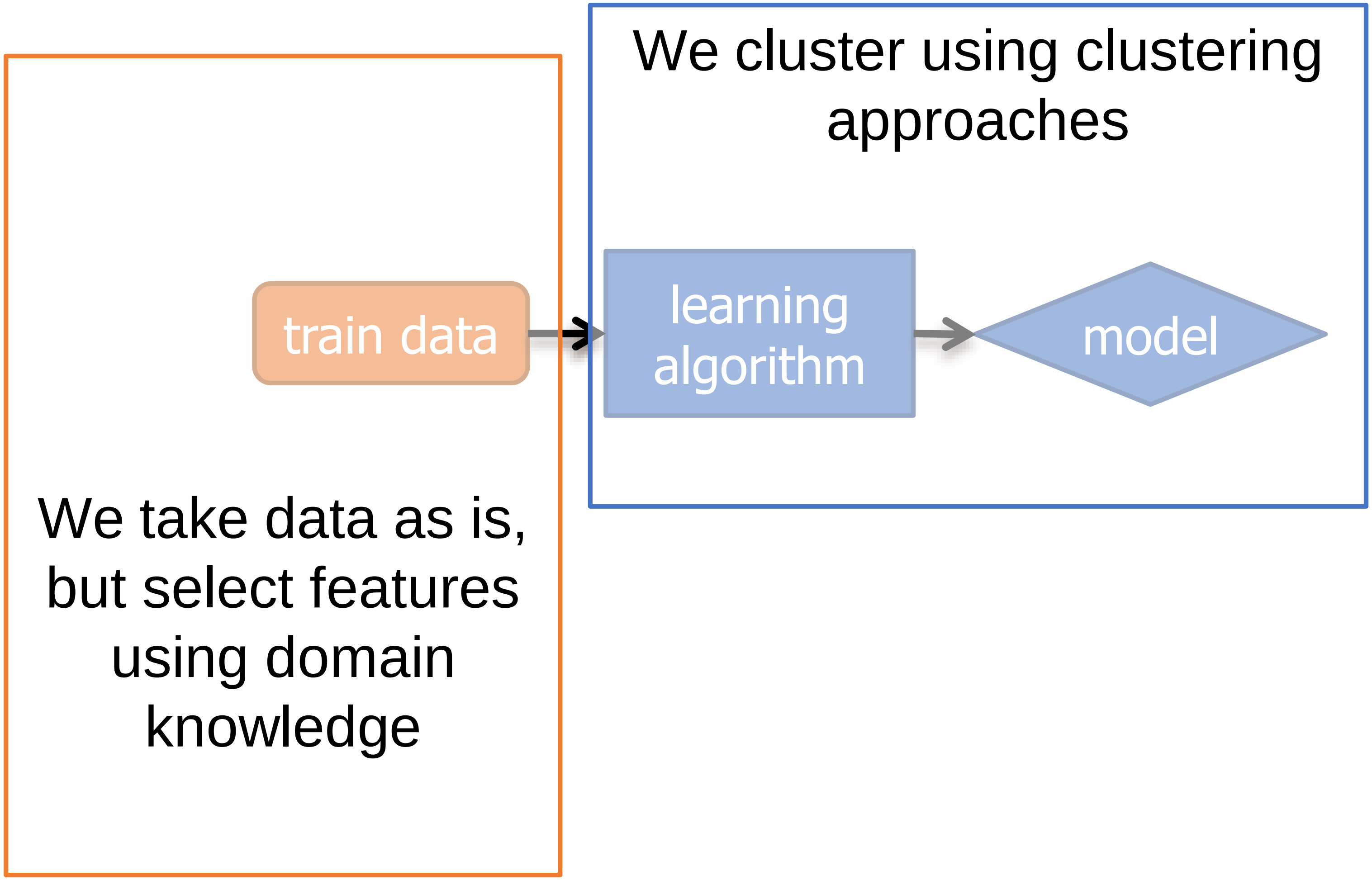
Row = connection
Column = time
Cell = Bytes transfered

These are all Data Mining skills



We take data as is, but select features using domain knowledge

These are all Data Mining skills



We cluster using clustering approaches

train data

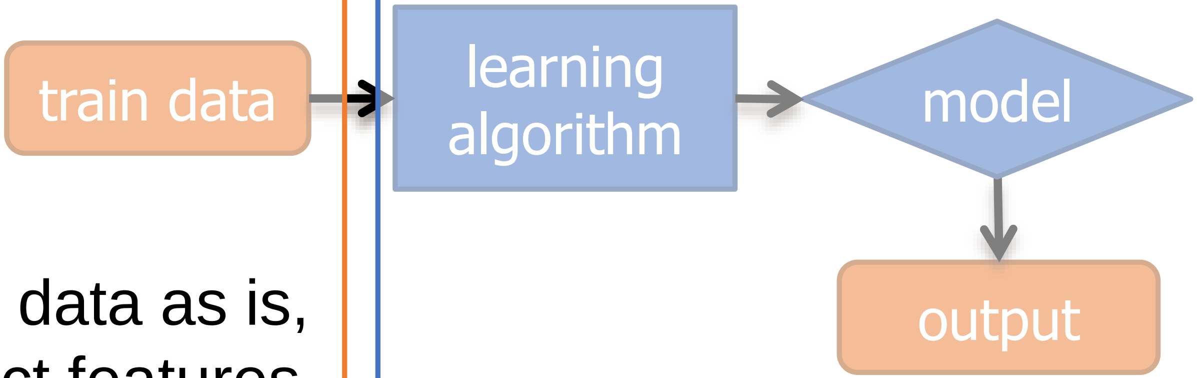
learning algorithm

model

We take data as is, but select features using domain knowledge

These are all Data Mining skills

We cluster using clustering approaches



We take data as is, but select features using domain knowledge

and make the results (gained knowledge) insightful using heatmaps

Why is it hard ?

Volume

Velocity

Variety

Why is it hard ?

Volume

Velocity

Variety

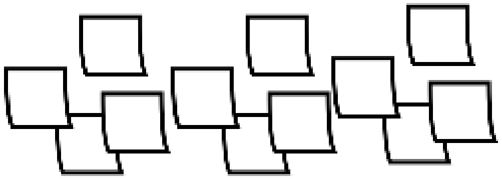
Need
scalable
algorithms

Why is it hard ?

Volume

Velocity

Variety



Need approximate yet accurate estimates

Why is it hard ?

Volume

Velocity

Variety

🖼 📄 📊

Adapt to different datatypes, distributions, ..

COURSE CONTENT

Distances, Matrices, Counting

Distances

- Similarity
- Metrics
- Computation
- DTW
- Text Embeddings
- Graph Embeddings

Matrices

- Representation
- Properties
- Operations
- Factorization
- Decompositions
- Dimensionality red.

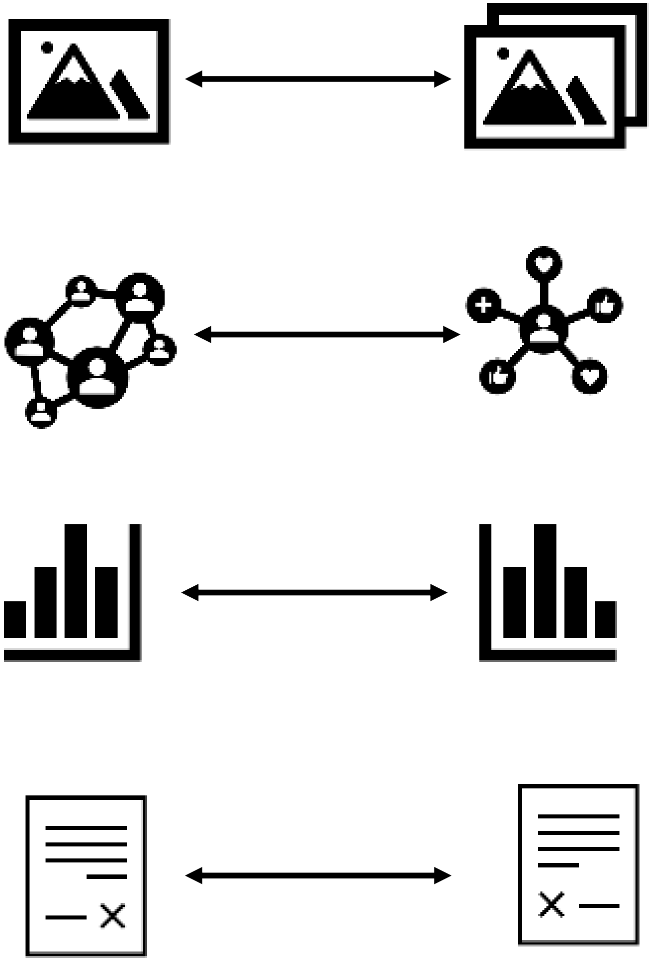
Counting

- Hashing
- Clustering
- Anomaly detection
- Sketching

Distances, Matrices, Counting

Distances

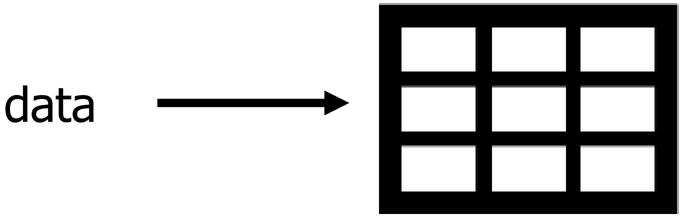
- Similarity
- Metrics
- Computation
- DTW
- Text Embeddings
- Graph Embeddings

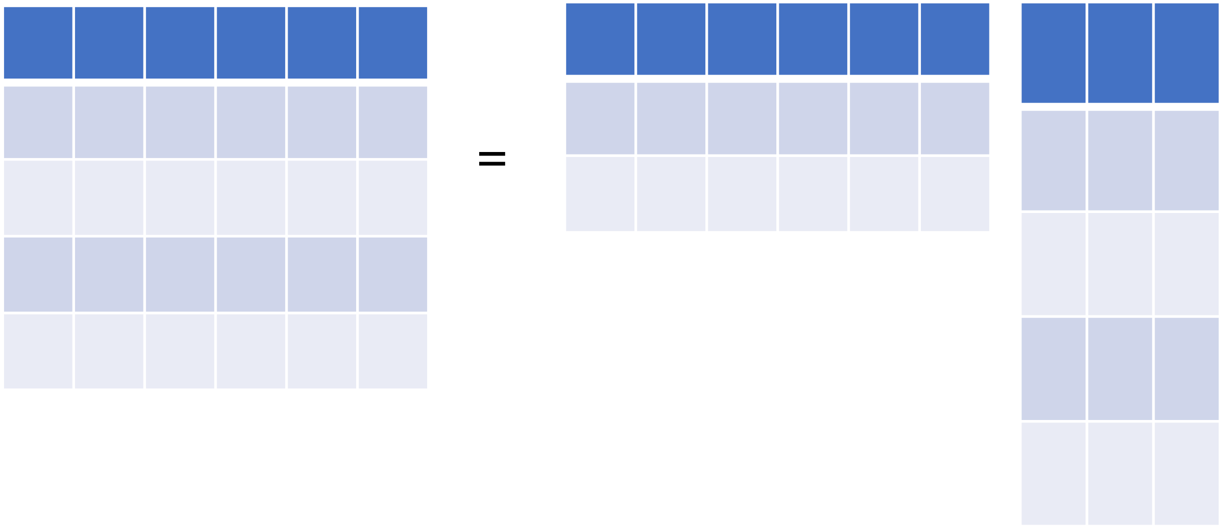


Distances, Matrices, Counting

Matrices

- Representation
- Properties
- Operations
- Factorization
- Decompositions
- Dimensionality red.

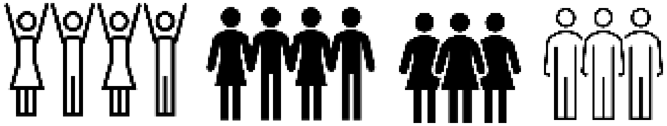


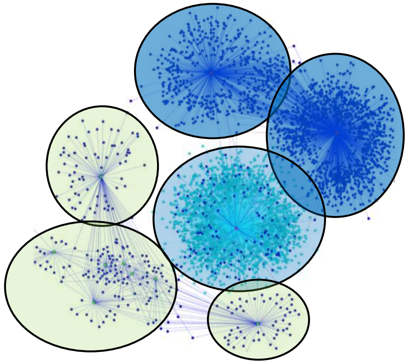


Distances, Matrices, Counting

Counting

- Hashing
- Clustering
- Anomaly detection
- Sketching

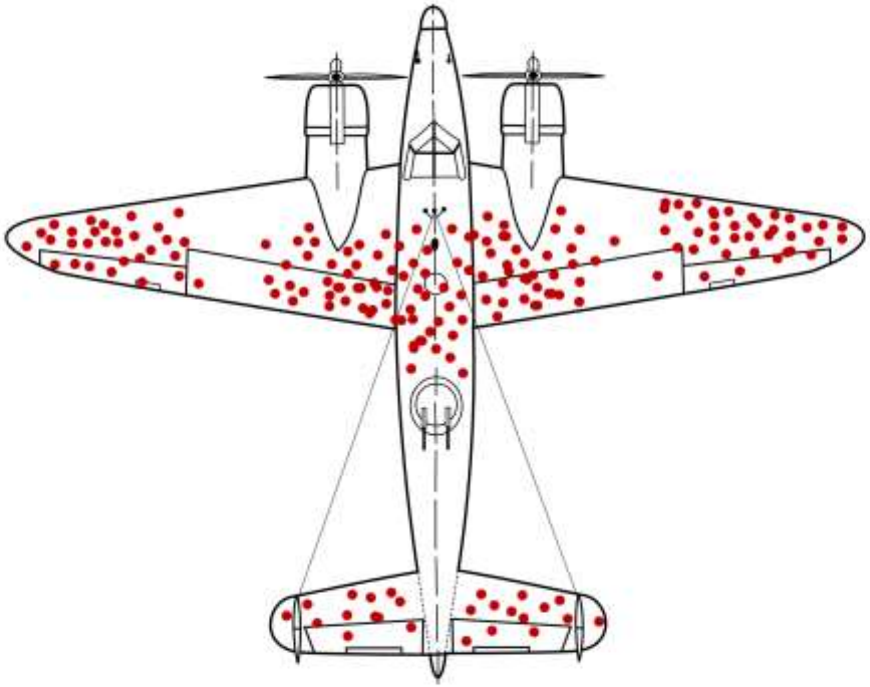




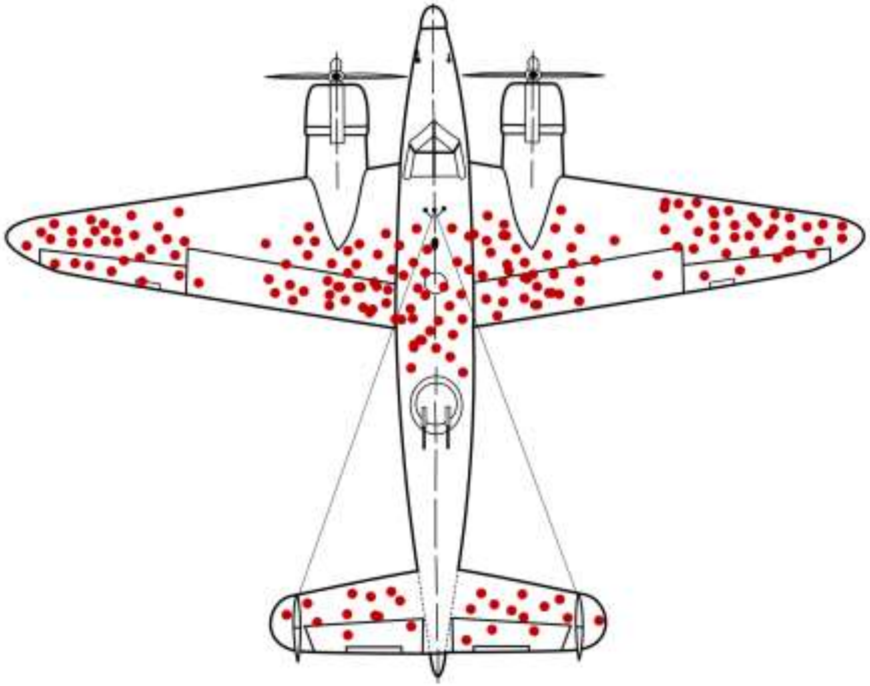
A word of caution

Don’t fool yourself

Where should you reinforce the airplane ?



Survivorship bias

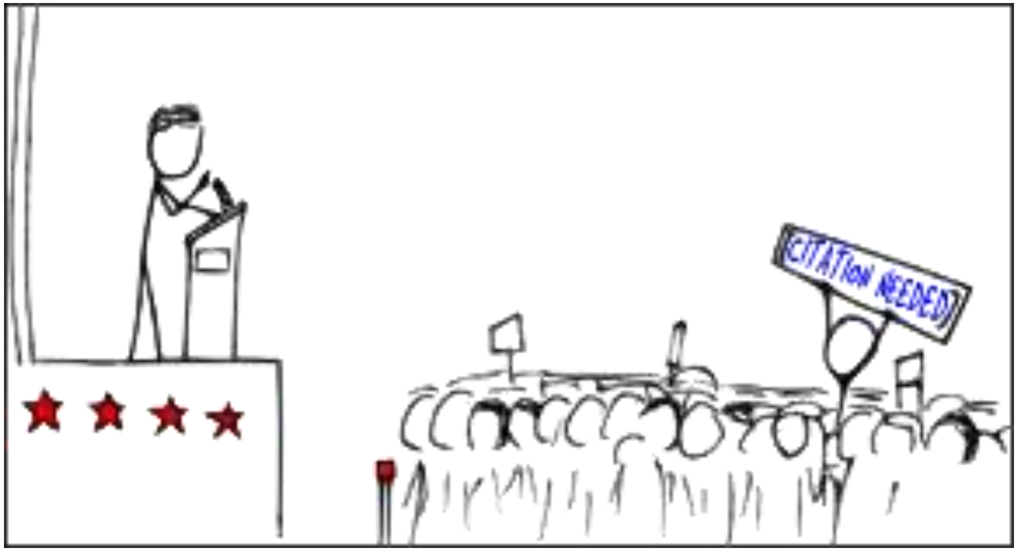


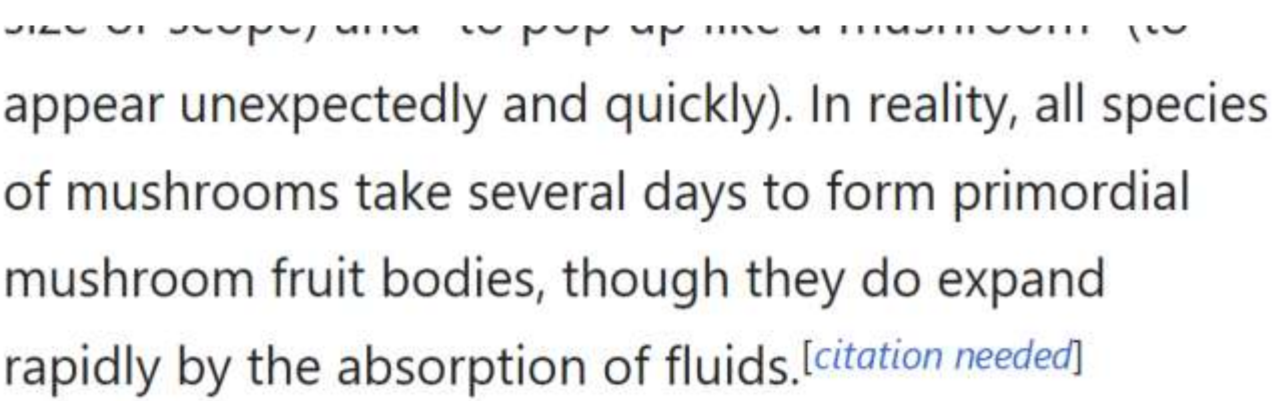
The Clever Hans Effect

Citations in Wikipedia

The gunpowder store (Dutch: Kruithuis) was subsequently re-housed, a 'cannonball's distance away', outside the city, in a new building designed by architect Pieter Post.[14]

Citation needed in Wikipedia





size or scope) and "to pop up like a mushroom" (to appear unexpectedly and quickly). In reality, all species of mushrooms take several days to form primordial mushroom fruit bodies, though they do expand rapidly by the absorption of fluids.[citation needed]

Citations in Wikipedia

The gunpowder store (Dutch: Kruithuis) was subsequently re-housed, a 'cannonball's distance away', outside the city, in a new building designed by architect Pieter Post. [14]

Delft University of Technology (TU Delft) is one of four universities of technology in the Netherlands. [23]

It was founded as an academy for civil engineering in 1842 by King William II.

Today, well over 21,000 students are enrolled. [24]

How can we automatically provide citations for Wikipedia

- The neural network was trained, and was 100% accurate on the test set

- What should you be asking yourself ?

Its too good to be true

Citations in Wikipedia

The gunpowder store (Dutch: Kruithuis) was subsequently re-housed, a 'cannonball's distance away', outside the city, in a new building designed by architect Pieter Post. [14]

→

The gunpowder store (Dutch: Kruithuis) was subsequently re-housed, a 'cannonball's distance away', outside the city, in a new building designed by architect Pieter Post.

Delft University of Technology (TU Delft) is one of four universities of technology in the Netherlands. [23]

→

Delft University of Technology (TU Delft) is one of four universities of technology in the Netherlands.

It was founded as an academy for civil engineering in 1842 by King William II.

→

It was founded as an academy for civil engineering in 1842 by King William II.

Today, well over 21,000 students are enrolled. [24]

→

Today, well over 21,000 students are enrolled.

TUDelft

How can we automatically provide citations for Wikipedia

- The neural network was trained, and was 100% accurate on the test set

- When asked humans – they were < 55% accurate

Shortcuts in Learning

The gunpowder store (Dutch: Kruithuis) was subsequently re-housed, a 'cannonball's distance away', outside the city, in a new building designed by architect Pieter Post. [14]

Delft University of Technology (TU Delft) is one of four universities of technology in the Netherlands. [23]

It was founded

civil engineer

William II.

Today, well over 21,000 students are enrolled. [24]

The gunpowder store (Dutch: Kruithuis) was subsequently re-housed, a 'cannonball's distance away', outside the city, in a new building designed by architect Pieter Post.

Delft University of Technology (TU Delft) is one of four universities of technology in the Netherlands.

It was founded as an academy for

ing in 1842 by King

ver 21,000 students are enrolled.

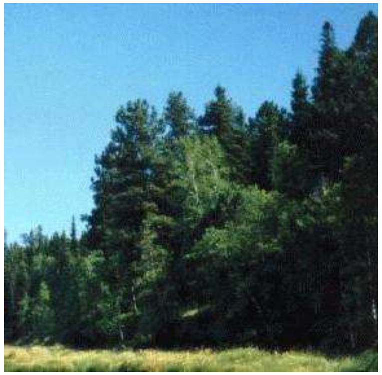
If ." " predict citation needed

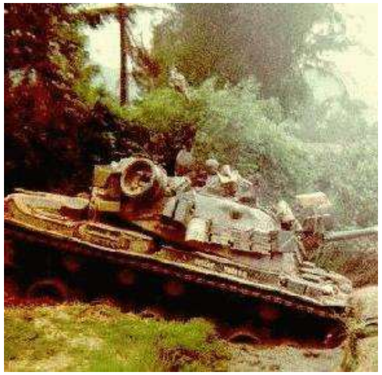
Russian Tanks

- Th on
the
- Wh l
wo

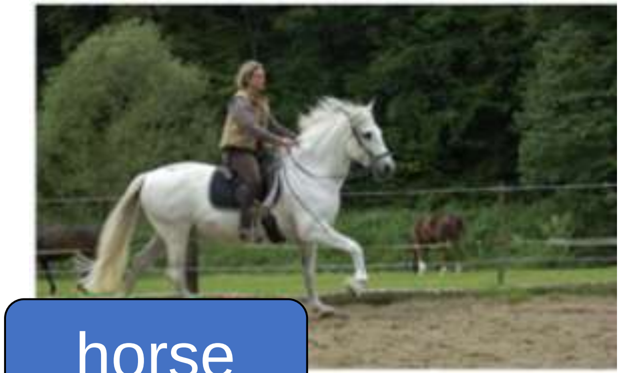
After much analysis, it turned out the network implemented the following:

If the sky is blue, there is no tank otherwise there is



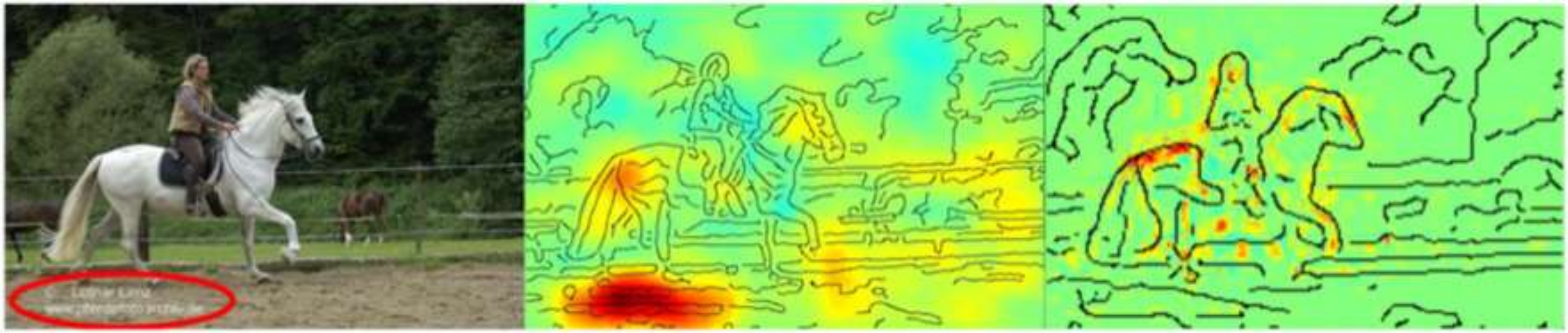


Shortcuts in image models



horse

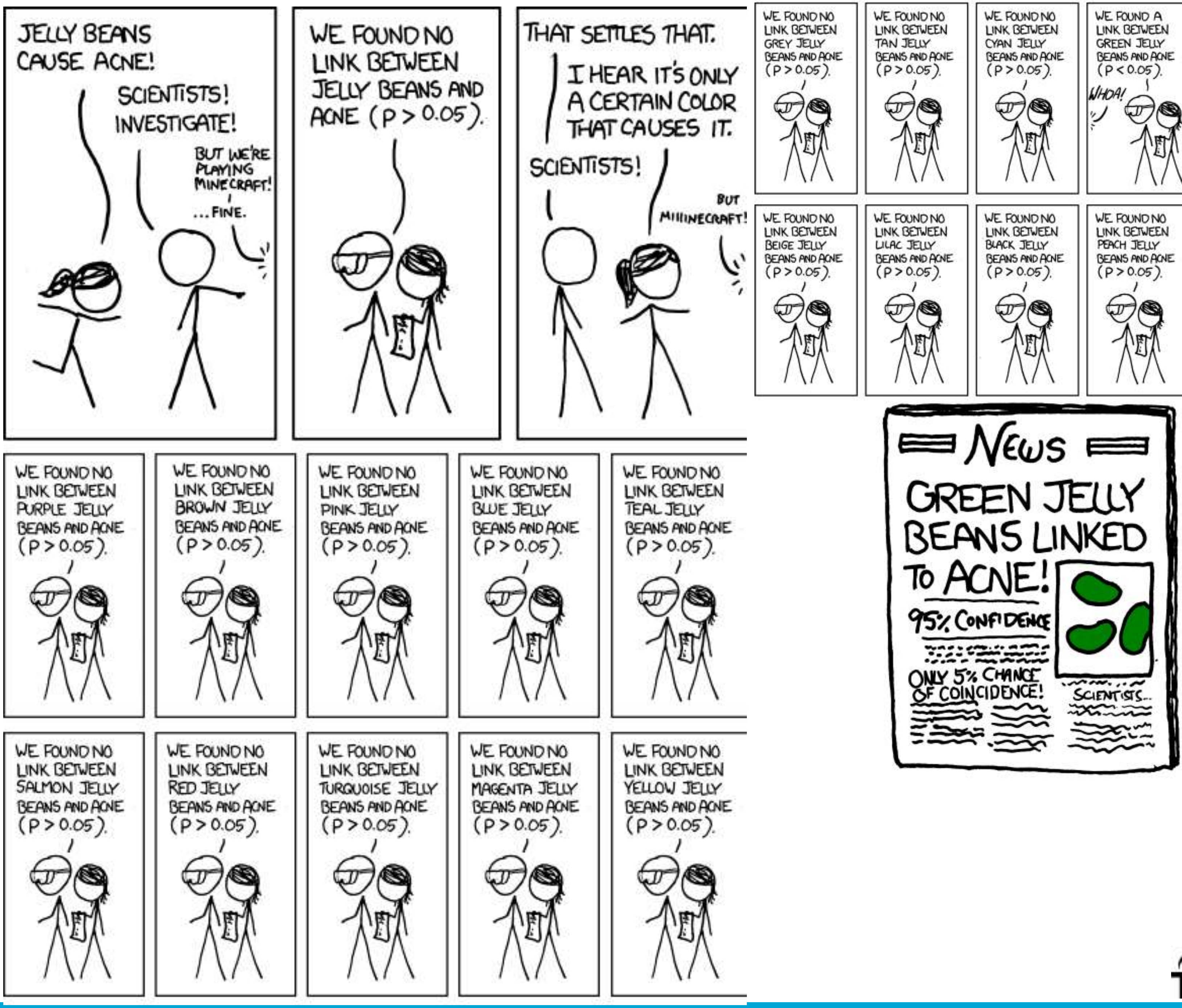
Shortcuts in image models



If "copyright" predict horse

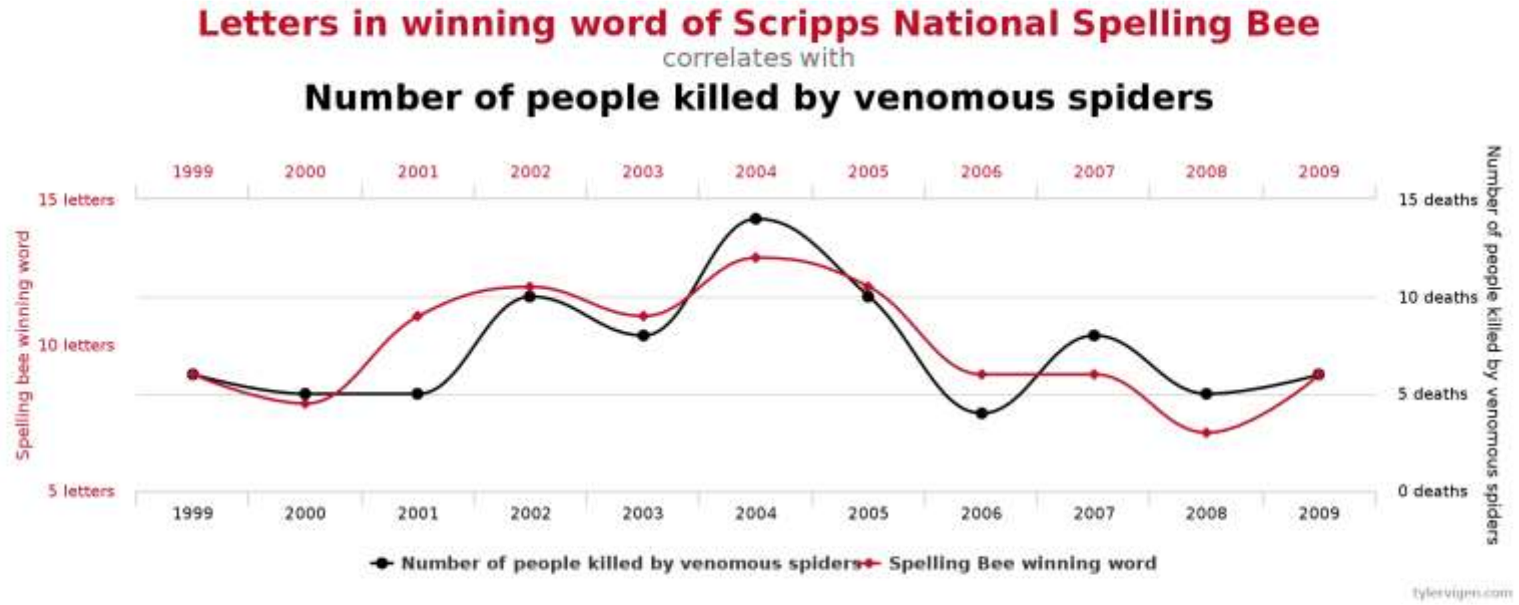
You can always find what you are looking for

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Multiple Comparisons Problem

- *The cause of many errors in data mining…*



https://youtu.be/HpjlcEH4zuY?si=3-KnR06RHipLk_UK

Summary

- **Data Mining:** Extract useful information from a large volume of data and transform it into an understandable structure
- Labs 30% (3 of them), Final exam 70%, Homeworks as examples for finals (6 of them)
- Content organization: Distances, Matrices, Counting
- Data mining as the skill of the 21st century
- But tread with caution