CSCI 347: Introduction to Data Mining

Lecture 1

WHAT IS **DATA MINING**?

- "the process of finding valid, novel, useful, and potentially understandable **patterns** in data" — G. Piatetsky-Shapiro, KDnuggets
- "**algorithms** for learning, analysis, data management and visualization of **large datasets**" — C. Faloutsos, CMU
- "Discovery of useful, possibly unexpected, **patterns** in **data**" — J. Ullman, Stanford
- "providing tools to **discover knowledge** from **data**" - J. Han in his textbook, *Data Mining: Concepts and Techniques*
- "the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from **large-scale data**" — M. J. Zaki and W. Meira Jr. in *Data Mining and Analysis: Fundamental Concepts and Algorithms*

DATA MINING APPLICATIONS: INTRUSION DETECTION

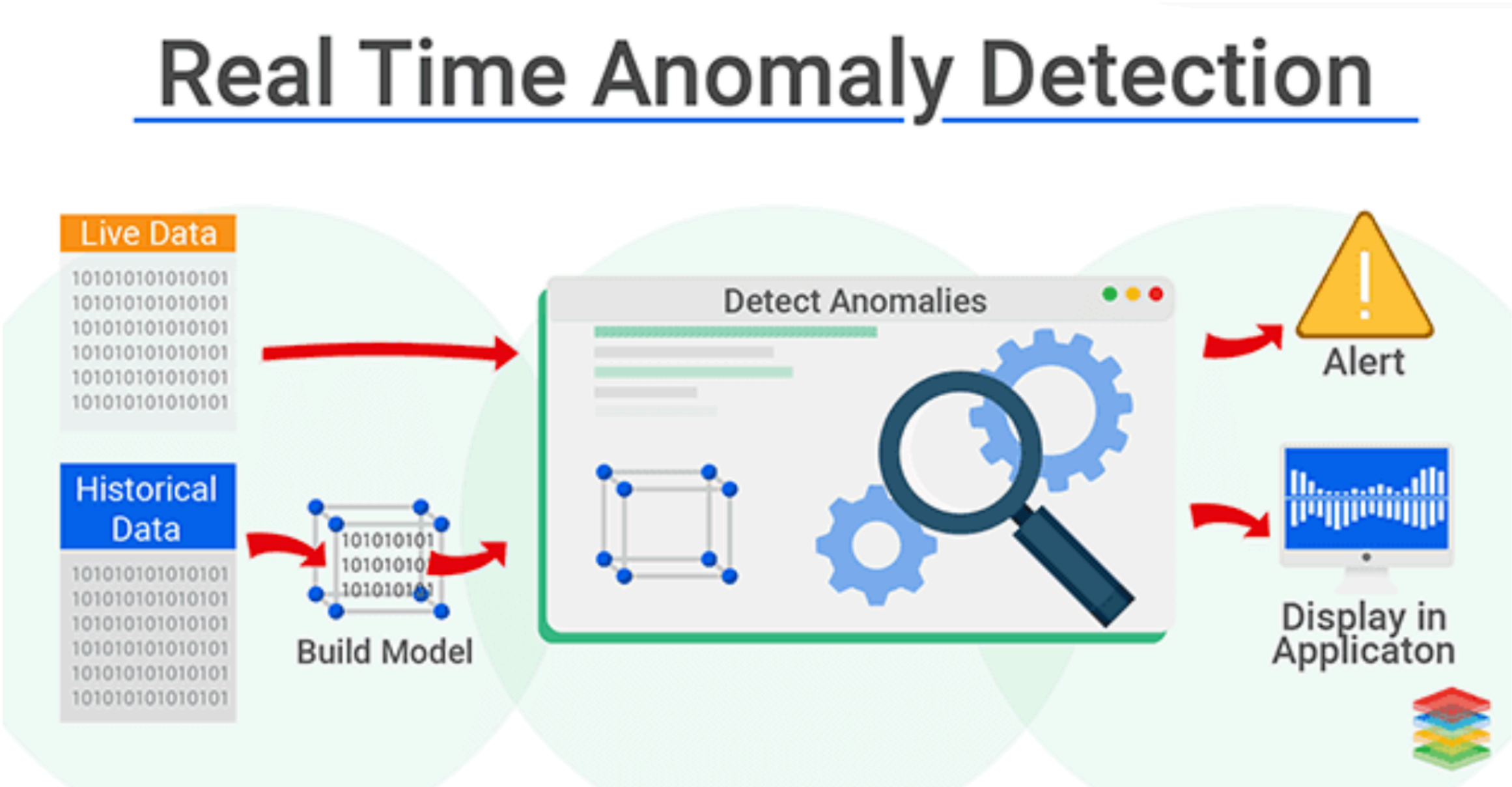


Image taken from: https://www.xenonstack.com/blog/real-time-anomaly-detection/

DATA MINING APPLICATIONS: MARKET BASKET ANALYSIS

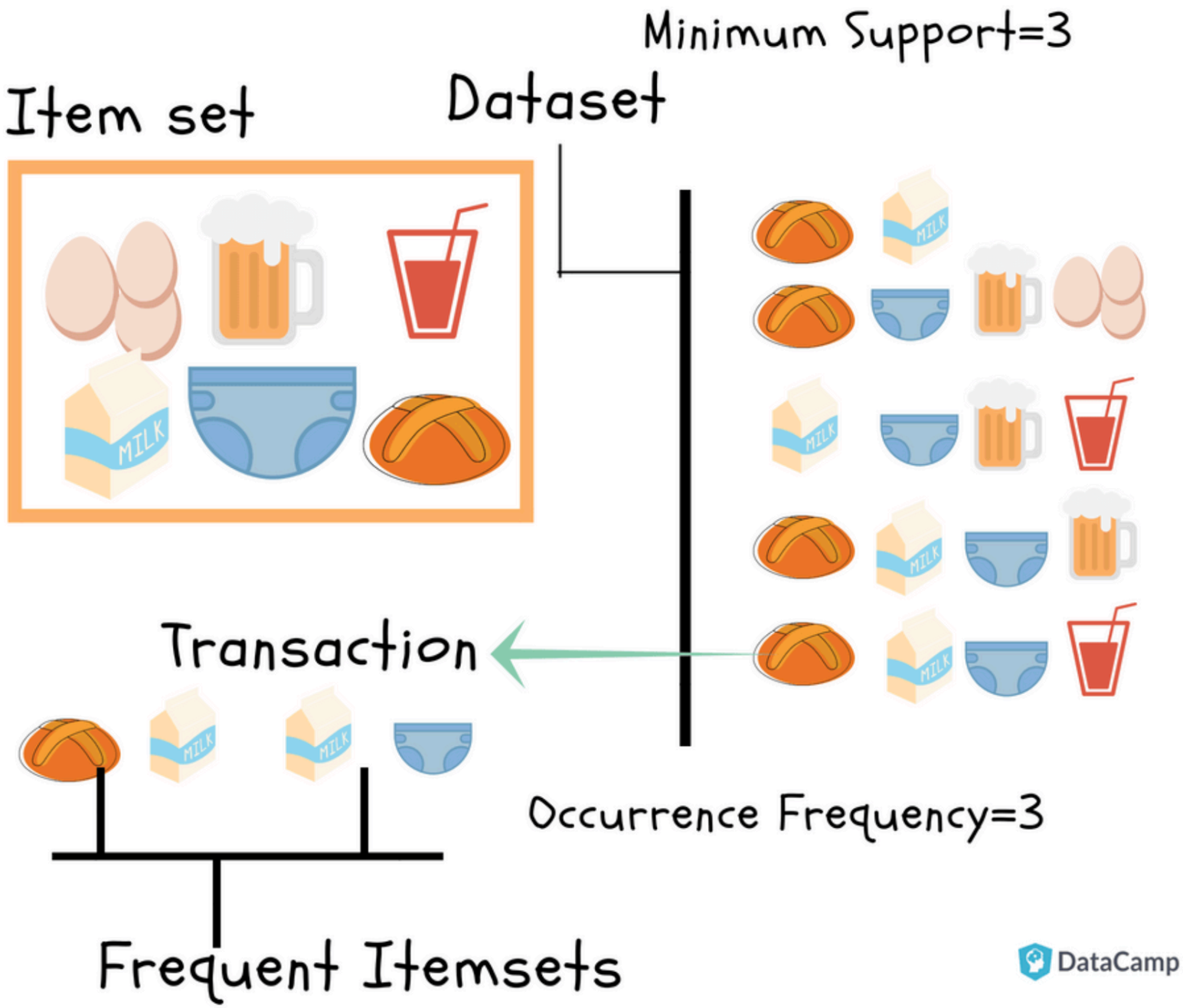


Image taken from: https://www.datacamp.com/community/tutorials/market-basket-analysis-r

DATA MINING APPLICATIONS: FINDING COMMUNITIES IN A SOCIAL NETWORK

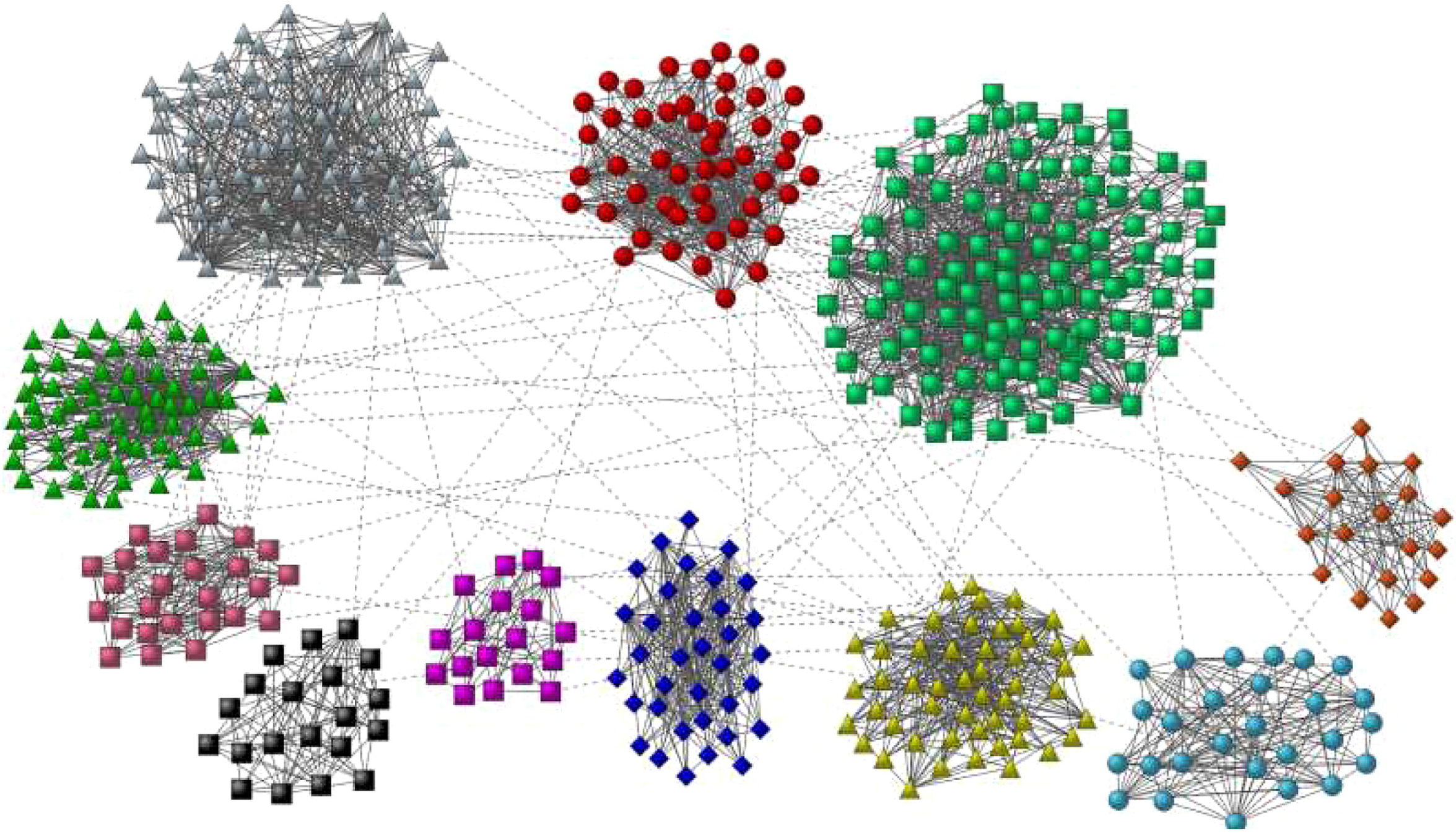
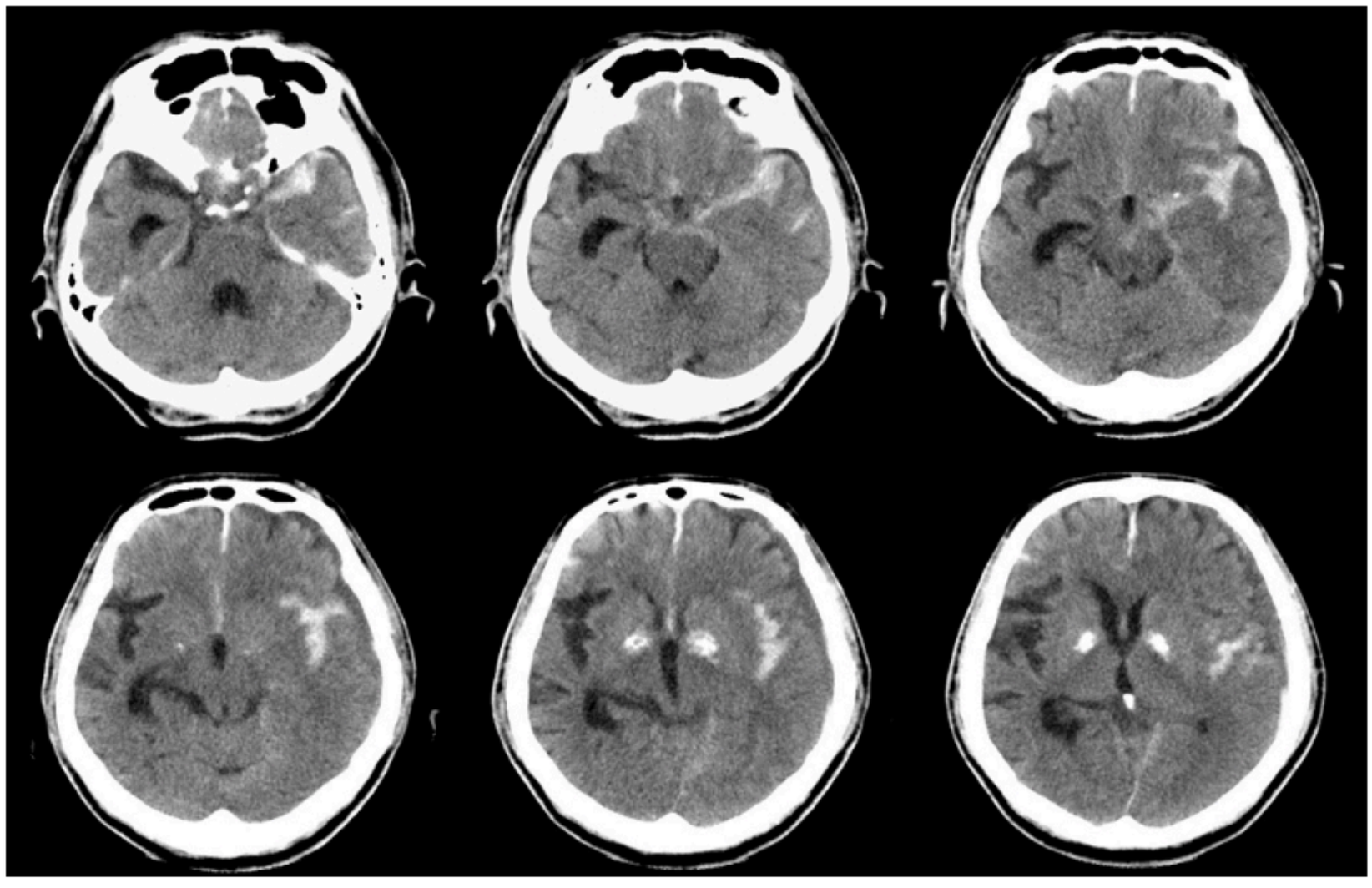


Image taken from: "Community detection in graphs." Fortunato, Santo. Physics Reports, Vol. 486, Feb. 2010

DATA MINING APPLICATIONS: ANOMALIES IN HEALTHCARE DATA

Deep learning algorithms can identify abnormalities on head computed tomography (CT) scans in patients who present with head trauma or stroke symptoms, according to study results published in the *Lancet*.*



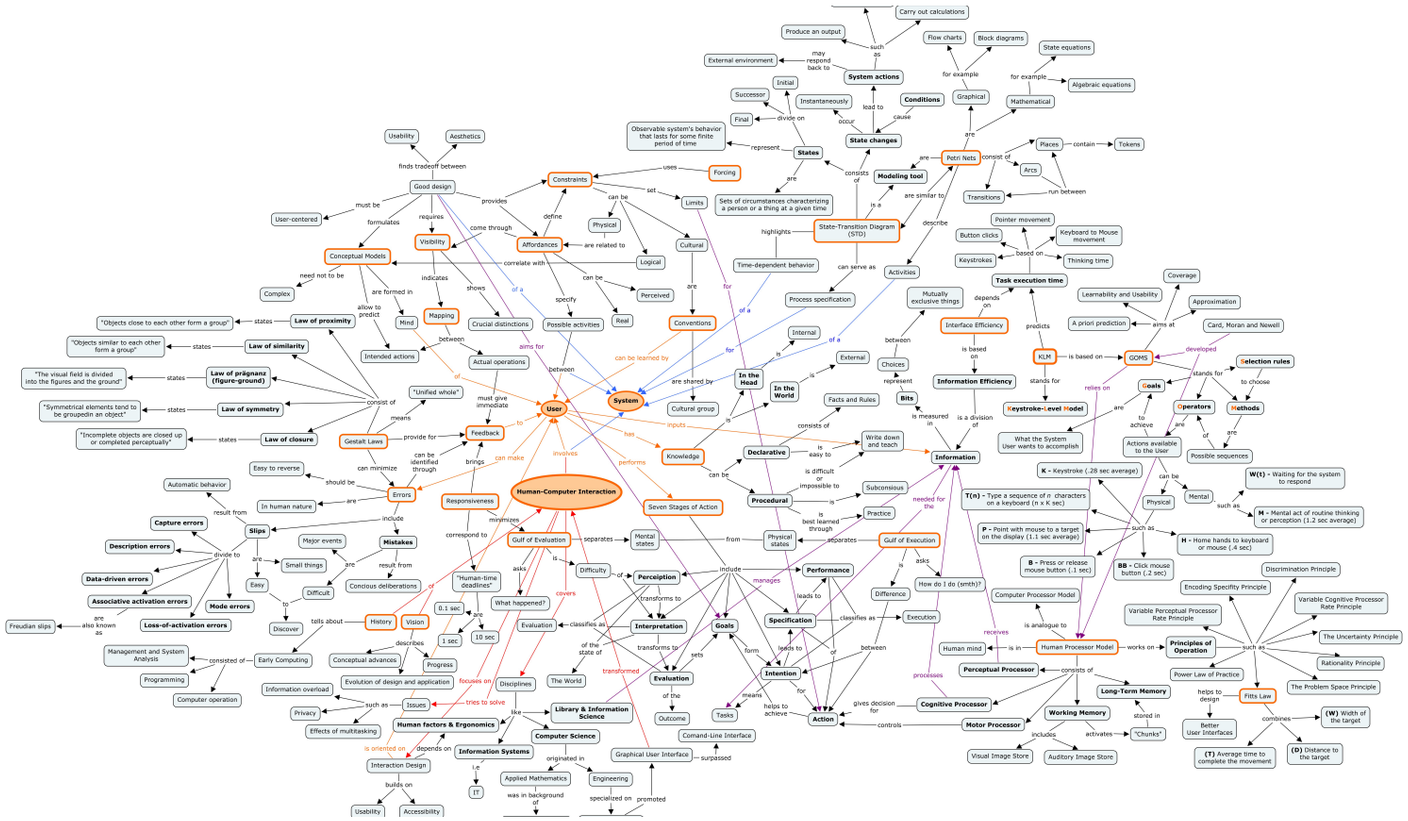
Investigators retrospectively reviewed clinical reports and data from more than 300,000 head computed tomography scans from medical centers in India.

**source: Neurology Advisor, Dec. 2018 (https://www.neurologyadvisor.com/topics/general-neurology/deep-learning-algorithms-identify-ct-scan-abnormalities-in-head-trauma-stroke/)*

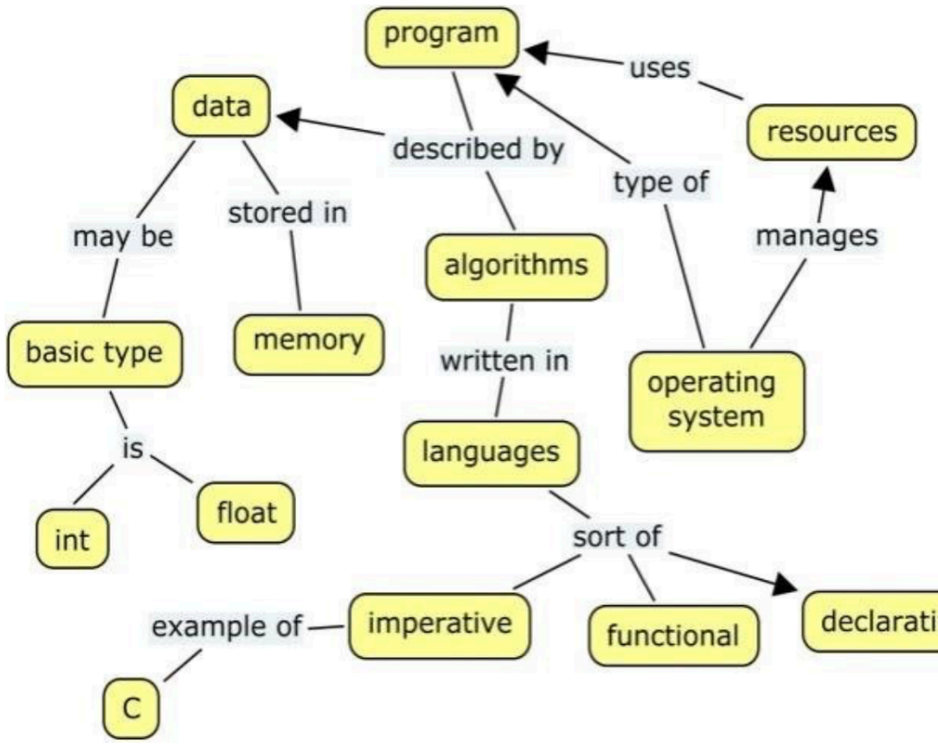
CONCEPT MAPS

"Concept Map (CM) is a graphical tool for organizing and structuring knowledge by depicting concepts as nodes, and relationships between concepts as edges" — Wei et al. in *Concept mapping in computer science education*

EXAMPLE OF CONCEPT MAPS



Concept map of HCI (source: HCI 2013 Study Blog, https://tengira.wordpress.com/category/foundations-of-hci/)

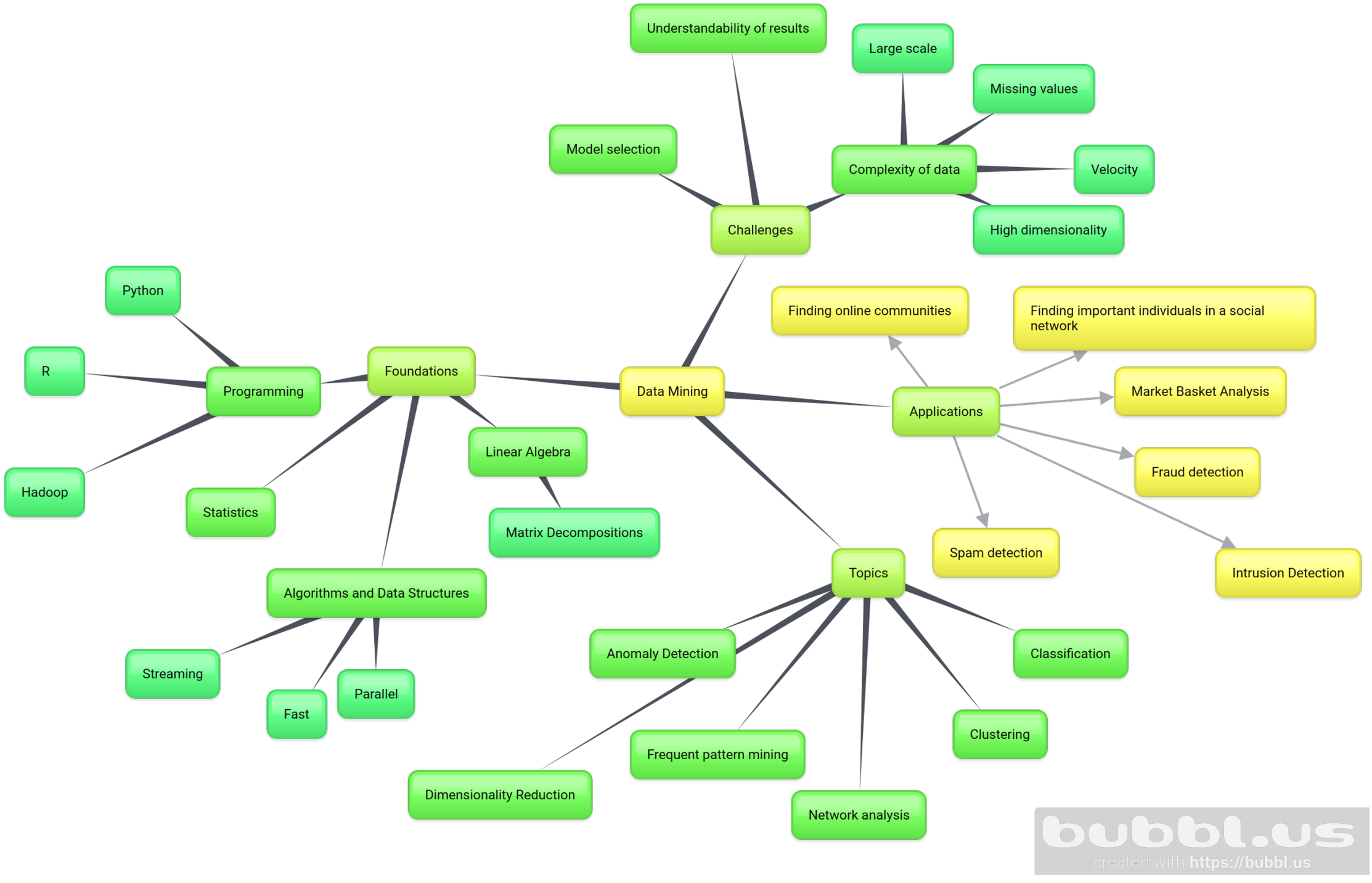


Concept map of an operating system (source: "An Evaluation Methodology for Concept Maps Mined from Lecture Notes: An Educational Perspective." Dec. 2015)

ACTIVITY: DATA MINING CONCEPT MAP

- A concept map helps us visualize important concepts related to a theme
  - Work with a partner to create a data mining concept map
  - Think about the following questions:
    - What are the applications of data mining?
    - What are the skills needed to become an expert in data mining?
    - What are some challenges in data mining?

ACTIVITY: DATA MINING CONCEPT MAP



Example of a data mining concept map (created using bubbl.us)

ABOUT THIS CLASS

Topics:

Common data formats, exploratory data analysis, data preparation, graph analysis, dimensionality reduction, clustering, classification, ethics in data mining and a few advanced topics

Instructor:

- Name: David Millman
- Email: david.millman@montana.edu
- Office: Barnard Hall 359
- Office Hours: See calendly

RESOURCES

- Optional Textbook: *Data Mining and Analysis: Fundamental Concepts and Algorithms* by Mohammed J. Zaki and Wagner Meira Jr. (Free at https://dataminingbook.info/book_html/)
- Additoonal books:
  - Mining of Massive Datasets by Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman
  - Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar
  - Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei
  - …and others listed on the course website
- Online tutorials/resources:
    - R for Data Science: http://r4ds.had.co.nz/
    - KDnuggets: https://www.kdnuggets.com
    - Mining of Massive Datasets: http://www.mmds.org
- Conference proceedings (research papers):
    - ACM SIGKDD (KDD)
    - IEE ICDM
    - SIAM SDM

GRADING

Point distribution:

- Homework: 30%
- Quizzes: 10%
- Projects: 30%
- Exam: 15%
- Final Project 15%

Work will be **submitted to both Gradescope and Brightspace** unless otherwise noted

HOMEWORK

- Partner work is encouraged
- Short-answer, practice the concepts presented in class
- 30% of final grade

QUIZZES

- Individual work is mandatory
- On Brightspace
- About 5 problems per quiz, based on recent lecture(s)
- One per week (unless announcement stating otherwise)
- 10% of final grade

PROJECTS

- Partner work is mandatory
- Practice with applying data mining algorithms to data sets, programming, and writing up a report
- 30% of final grade

EXAM

- Date: April 21st (expected)
  - Covers approximately the first 3/4 of class material, a broad range of fundamental data mining topics
- In-class, closed-book
- 15% of final grade

CONTENT LOCATIONS

- Course content (syllabus, code, slides, etc.) on Github
- Announcements and discussions on Discord
  - Please see me after class if you are a Discord ninja
- Grade book on Brightspace
- Quizzes on Brightspace

ABOUT ME

- Background
  - Ph.D. from UNC - Chapel Hill in Computer Science
  - M.S. from NYU in Computer Science
  - B.A. from Colgate University in CS
- Research interests:
  - Computational geometry and topology, Geometric and topological data analysis, Large-scale data analysis, Parallel and scientific computing.
  - **Please reach out to me if you are interested in undergraduate research in this area!**