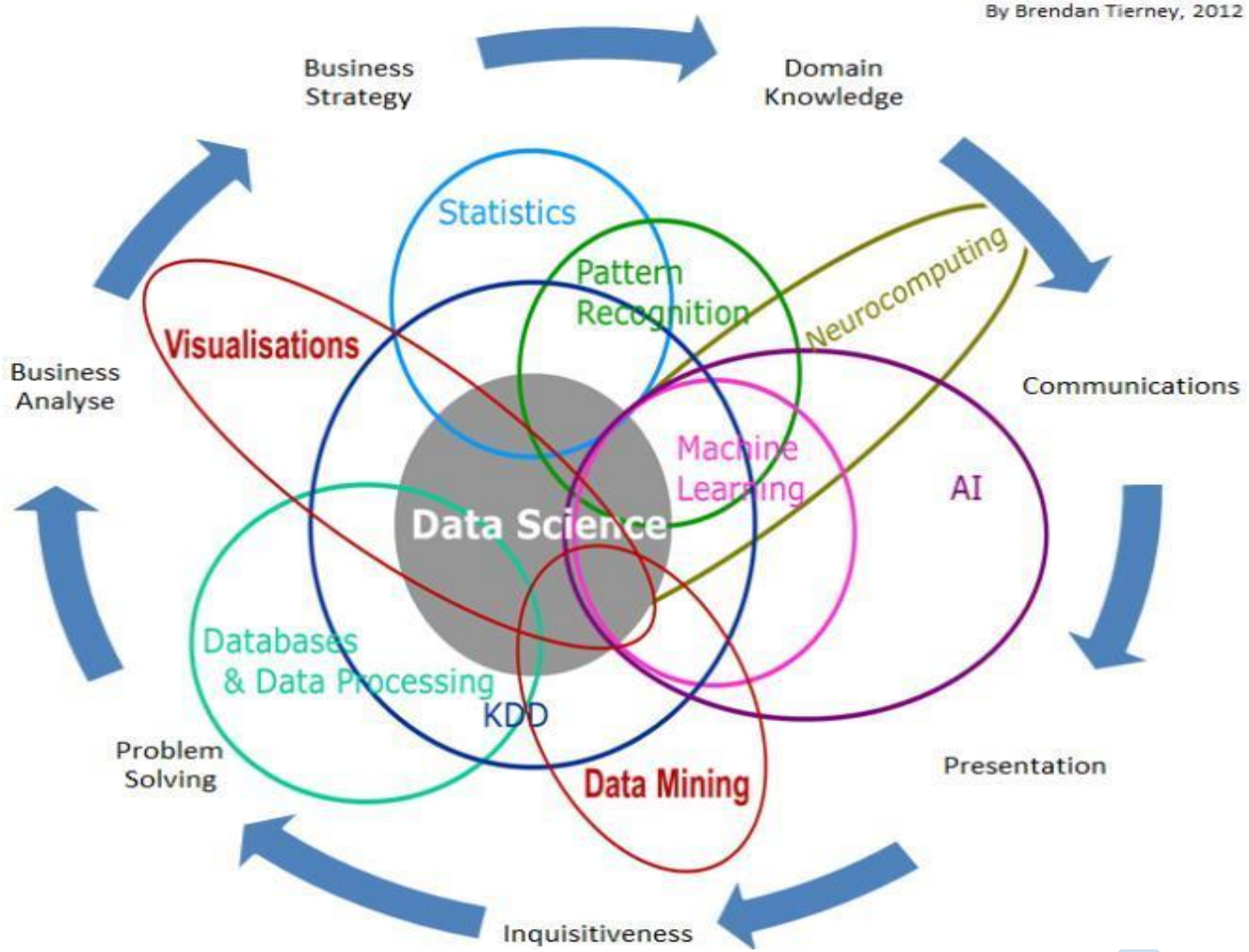
Data Science
Isu-isu Penelitian Sosial

Hendro Margono

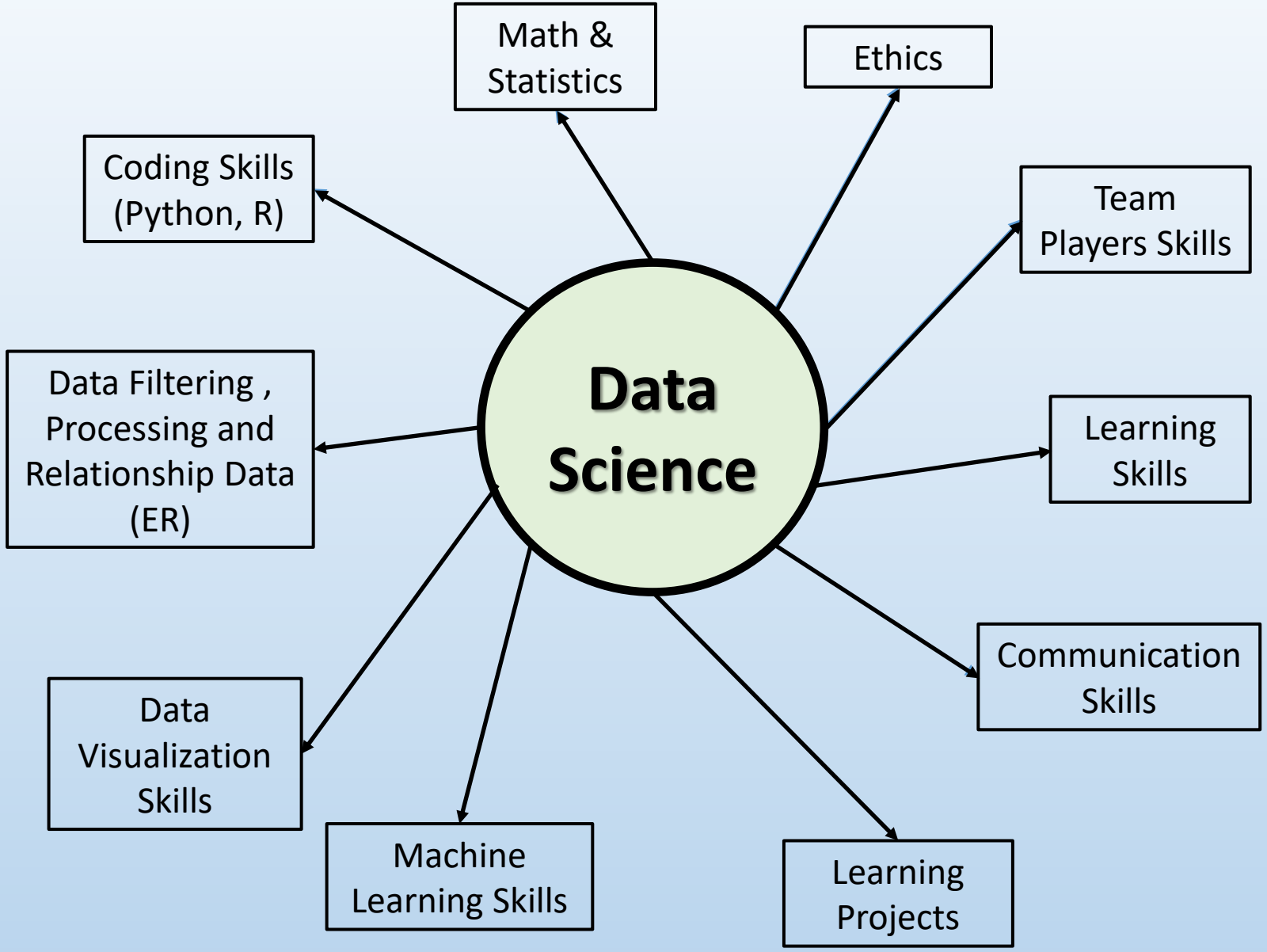
Semarang, 31 August 2022

Data Science Is Multidisciplinary

By Brendan Tierney, 2012



Data Science



Data Science

- Pendekatan ilmiah yang menerapkan ide matematika dan statistik dan alat komputer untuk memproses big data.
- Data Science adalah area yang cukup menantang karena kompleksitas yang terlibat dalam menggabungkan dan menerapkan metode, algoritme, dan teknik pemrograman yang berbeda untuk melakukan analisis cerdas dalam volume data yang besar.

Data Science

- *Capture:* Data Acquisition, Data Entry, Signal Reception, Data Extraction. Tahap ini melibatkan pengumpulan data terstruktur dan tidak terstruktur mentah.
- *Maintain:* Data Warehousing, Data Cleansing, Data Staging, Data Processing, Data Architecture. Tahap ini mencakup pengambilan data mentah dan meletakkannya dalam bentuk yang dapat digunakan.
- *Process:* Data Mining, Clustering/Classification, Data Modeling, Data Summarization. Ilmuwan data mengambil data yang disiapkan dan memeriksa pola, rentang, dan biasnya untuk menentukan seberapa berguna data tersebut dalam analisis prediktif.
- *Analyze:* Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis. Here is the original of the life cycle. Tahap ini melibatkan melakukan berbagai analisis pada data.
- *Communicate:* Data Reporting, Data Visualization, Business Intelligence, Decision Making. Pada langkah terakhir ini, analis menyiapkan analisis dalam bentuk yang mudah dibaca seperti bagan, grafik, dan laporan.

Kebutuhan dalam Data Science

- ***Machine Learning (ML)***

  Machine Learning adalah tulang punggung data sceince. Data Scientist perlu memiliki pemahaman yang kuat tentang ML selain pengetahuan dasar tentang statistic

- ***Modeling***

  Model matematika memungkinkan membuat perhitungan dan prediksi cepat berdasarkan apa yang dapat diketahui tentang data. Pemodelan juga merupakan bagian dari Pembelajaran Mesin (ML) dan melibatkan identifikasi algoritma mana yang paling cocok untuk memecahkan masalah tertentu dan bagaimana melatih model-model ini.

- ***Statistics***

  Statistik adalah inti dari ilmu data. Pegangan yang kokoh pada statistik dapat membantu mengekstrak lebih banyak *Intelligence* dan mendapatkan hasil yang lebih bermakna.

- ***Programming***

  Beberapa level pemrograman diperlukan untuk "*data science project*". Bahasa pemrograman yang paling umum adalah Python, dan R. Python sangat populer karena mudah dipelajari, dan mendukung banyak pustaka untuk data science dan Machine Learning

- ***Databases***

  Seorang ilmuwan data yang *capable* perlu memahami cara kerja database, cara mengelolanya, dan cara mengekstrak data

Penggunaan Data Science

- Data Science dapat mendeteksi pola dalam data yang tampaknya tidak terstruktur atau tidak terhubung, memungkinkan kesimpulan dan prediksi dibuat.
- Bisnis teknologi yang memperoleh data pengguna dapat memanfaatkan strategi untuk mengubah data tersebut menjadi informasi yang berharga atau menguntungkan.
- Data Science juga telah membuat terobosan ke dalam industri transportasi, seperti dengan mobil tanpa pengemudi. Sangat mudah untuk menurunkan jumlah kecelakaan dengan menggunakan mobil tanpa pengemudi. Misalnya, dengan mobil tanpa pengemudi, data pelatihan dipasok ke algoritma, dan data diperiksa menggunakan pendekatan Data Science, seperti batas kecepatan di jalan raya, jalan yang sibuk, dll.
- Aplikasi Data Science memberikan tingkat penyesuaian terapeutik yang lebih baik melalui penelitian genetika dan genomik.

Big Data

- Big data memiliki satu atau lebih karakteristik: volume tinggi, kecepatan tinggi, atau variasi tinggi. Kecerdasan buatan (AI), seluler, sosial, dan Internet of Things (IoT) mendorong kompleksitas data melalui bentuk dan sumber data baru.
- Misalnya, big data berasal dari sensor, perangkat, video / audio, jaringan, file log, aplikasi transaksional, web, dan media sosial - sebagian besar dihasilkan secara real time dan dalam skala yang sangat besar.

Big Data

- **Unstructured data** – social networks, emails, blogs, tweets, digital images, digital audio/video feeds, online data sources, mobile data, sensor data, web pages, and so on.
- **Semi-structured** – XML files, system log files, text files, etc.
- **Structured data** – RDBMS (databases), OLTP, transaction data, and other structured data formats.

Data Science and Data Mining

- ***Data Science*** adalah bidang antar-disiplin yang menggunakan metode, proses, algoritma, dan sistem ilmiah untuk mengekstrak pengetahuan dan wawasan dari banyak data struktural dan tidak terstruktur. Data science terkait dengan ***data mining, machine learning, deep learning dan big data***.
-
- ***Data Mining*** adalah proses menemukan pola dalam kumpulan data besar yang melibatkan metode di persimpangan pembelajaran mesin, statistik, dan sistem basis data. Penambangan data adalah sub-bidang ilmu komputer dan statistik antar-disiplin dengan tujuan keseluruhan untuk mengekstrak informasi (dengan metode cerdas) dari kumpulan data dan mengubah informasi menjadi struktur yang dapat dipahami untuk digunakan lebih lanjut.

Data Science for Social Science

- Using “Big Data” systems to project thousands of what-if scenarios to better understand the impact of proposed policies in real-time
- Combining “Big Data” with traditional data sources to produce real-time, cost-effective early warning systems
- Using new techniques to unlock sources of data trapped in various text formats that help facilitate new lines of research and understanding

Penerapan Data Science

- Isu-Isu Sosial
  - Text and structural data mining of influenza mentions in web and social media. https://www.mdpi.com/1660-4601/7/2/596
  - Challenges in mining social network data: processes, privacy, and paradoxes. https://dl.acm.org/doi/pdf/10.1145/1281192.1281195
  - Fake news detection on social media: A data mining perspective. https://dl.acm.org/doi/abs/10.1145/3137597.3137600
  - Detecting and tracking disease outbreaks by mining social media data. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.415.6200&rep=rep1&type=pdf
  - Document Classification Through Data Mining Social Media Networks. https://images.stetson.edu/photos/researchcollection/_PDFs/seniorprojects_2009-003.pdf

Penerapan Data Science

- Medical
  - Predictive data mining for medical diagnosis: An overview of heart disease prediction https://pdfs.semanticscholar.org/fbd6/5a18f6653b56138cd5196d20e2f39de189e3.pdf
  - Medical image classification using an efficient data mining technique. http://dro.deakin.edu.au/view/DU:30005389
  - Combination data mining methods with new medical data to predicting outcome of coronary heart disease. https://ieeexplore.ieee.org/abstract/document/4420369/
  - Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5385648/

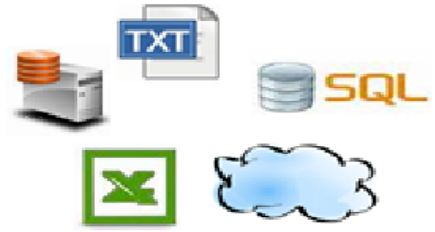
Penerapan Data Science

- Engineering
  - Toward data mining engineering: A software engineering approach. https://www.sciencedirect.com/science/article/abs/pii/S0306437908000355
  - Data mining in an engineering design environment: OR applications from graph matching. https://www.sciencedirect.com/science/article/abs/pii/S0305054805000262
  - A robust data mining approach for formulation of geotechnical engineering systems. https://www.ingentaconnect.com/content/mcb/182/2011/00000028/00000003/art00002
  - Mining software engineering data. https://ieeexplore.ieee.org/abstract/document/4222731

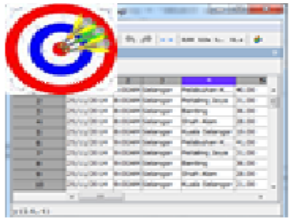
Penerapan Data Science

- Economic
  - Economics in the age of big data. https://science.sciencemag.org/content/346/6210/1243089.abstract
  - Data mining in economic science. https://webspace.science.uu.nl/~feeld101/dmecon.pdf
  - Data mining for financial decision making. https://psycnet.apa.org/record/2004-15263-001
  - Data Pricing--From Economics to Data Science. https://dl.acm.org/doi/abs/10.1145/3394486.3406473

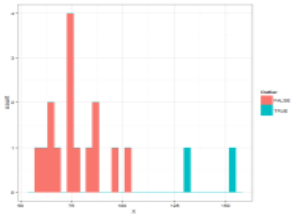
Knowledge Data from Database (KDD)



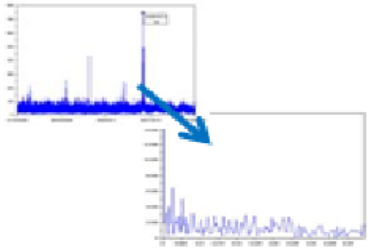
Data Source



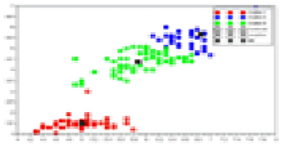
Target Data



Processed Data



Transformed Data



Patterns

Knowledge

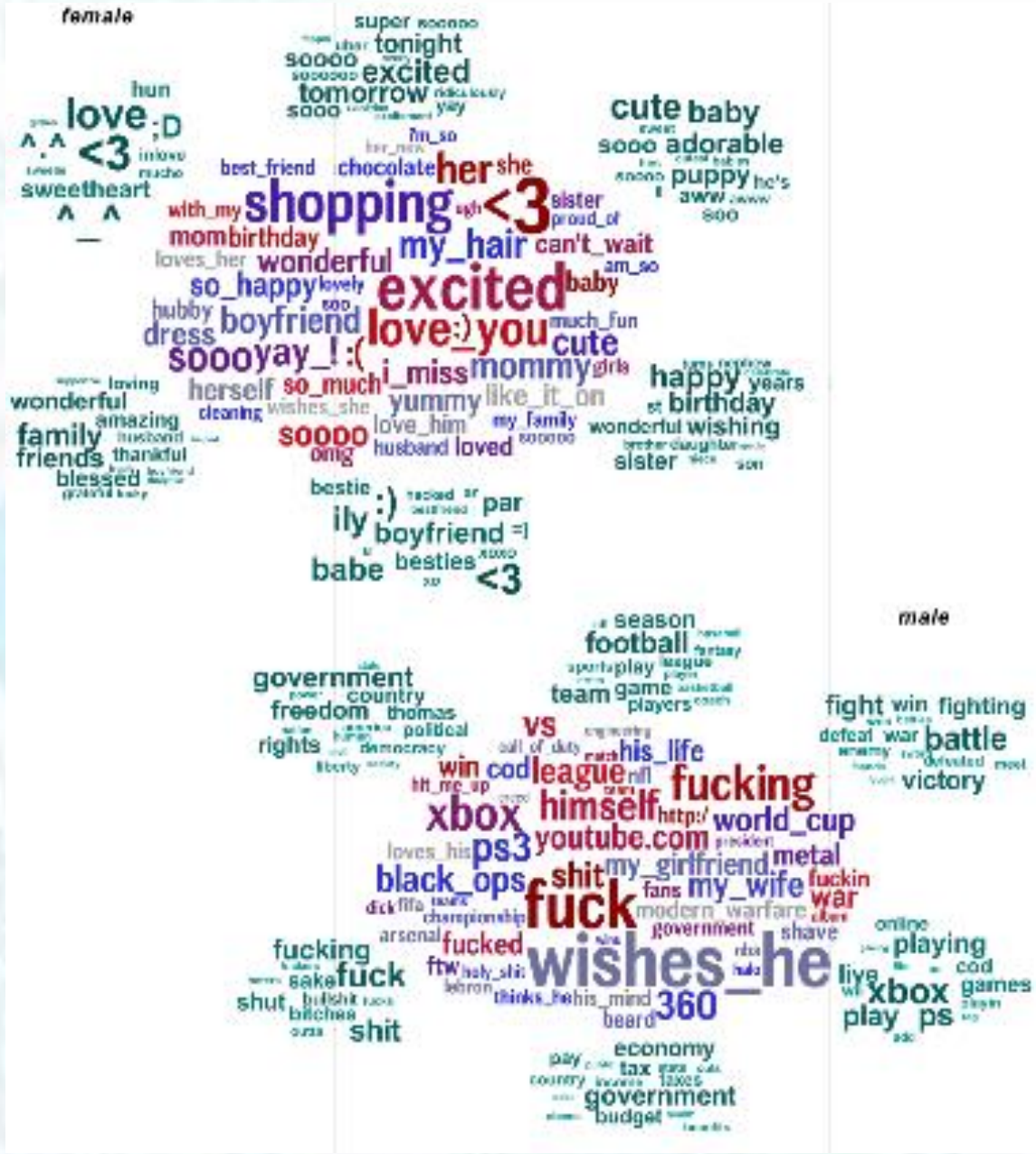
Data Science Tasks

- Prediction Methods
  - Menggunakan beberapa variables untuk memprediksi unknown or future values of other variables. (Classification, Regression, Outlier Detection)
- Description Methods
  - Untuk menemukan human-interpretable patterns yang mendiskripsikan data. (Clustering, Association Rule Mining, Sequential Pattern Discovery)

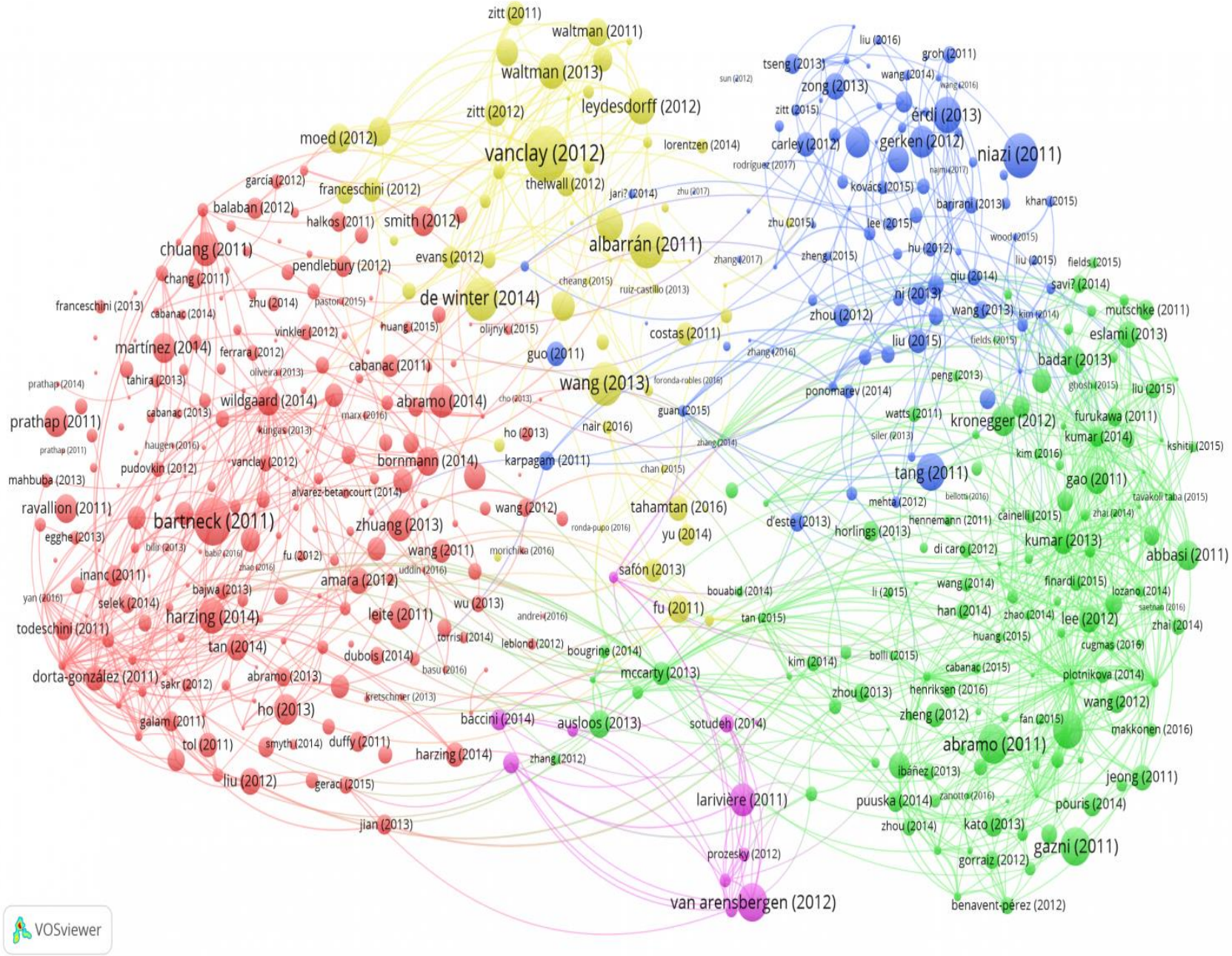
Data Science Techniques

- Linear Regression
- Logistic Regression
- Jackknife Regression
- Density Estimation
- Confidence Interval
- Test of Hypotheses
- Pattern Recognition
- Clustering – (aka Unsupervised Learning)
- Supervised Learning
- Time Series
- Decision Trees
- Random Numbers
- Monte-Carlo Simulation
- Bayesian Statistics
- Naïve Bayes
- Principal Component Analysis – (PCA)
- Ensembles
- Neural Networks
- Support Vector Machine – (SVM)
- Nearest Neighbors – (k-NN)
- Feature Selection – (aka Variable Reduction)
- Indexation / Cataloguing
- (Geo-) Spatial Modeling
- Recommendation Engine
- Search Engine
- Attribution Modeling
- Collaborative Filtering
- Rule System
- Linkage Analysis
- Association Rules
- Scoring Engine
- Segmentation
- Predictive Modeling
- Graphs
- Deep Learning
- Game Theory
- Imputation
- Survival Analysis
- Arbitrage
- Lift Modeling
- Yield Optimization
- Cross-Validation
- Model Fitting
- Relevancy Algorithm
- Experimental Design

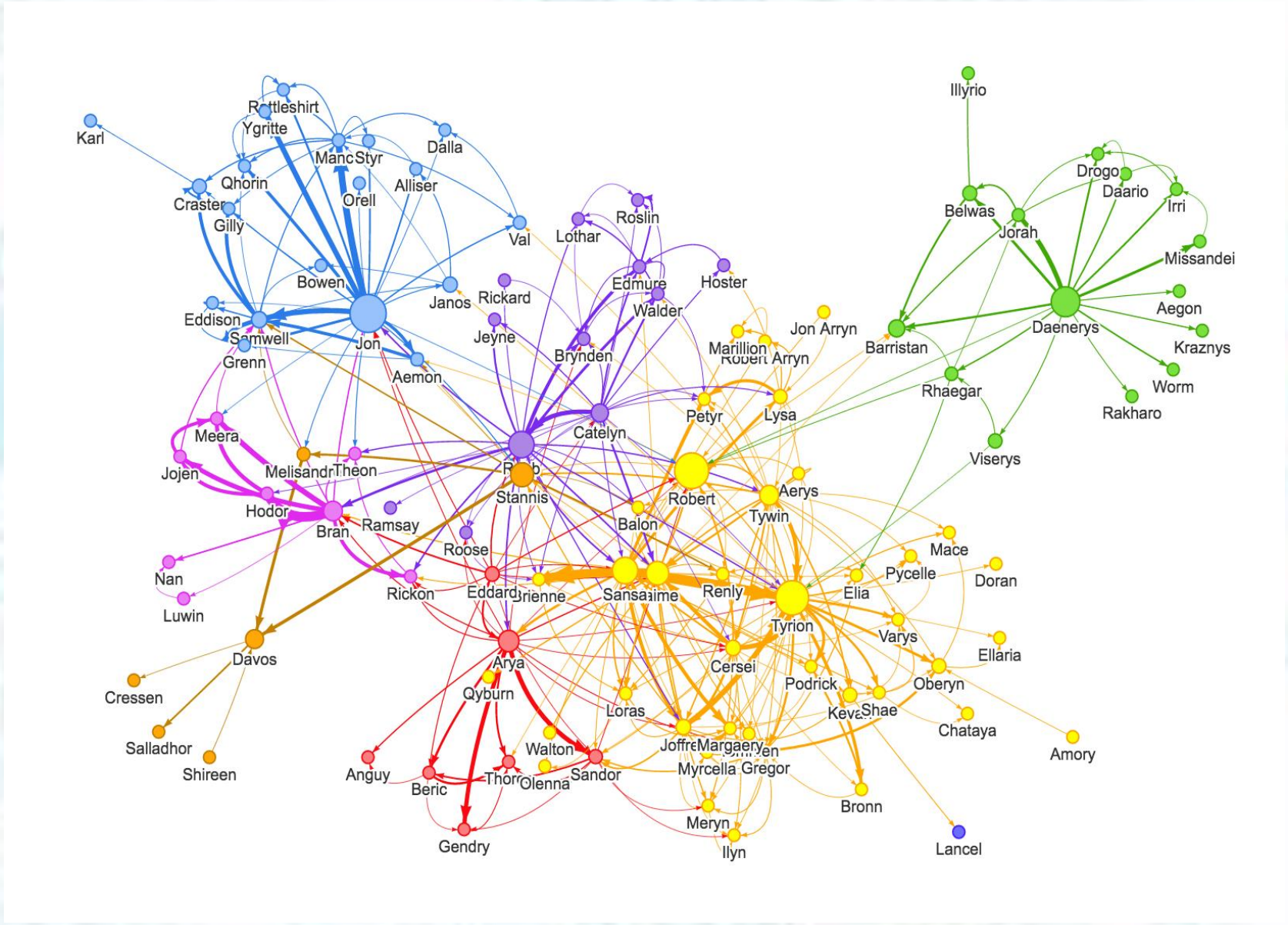
Text Mining in Social Media



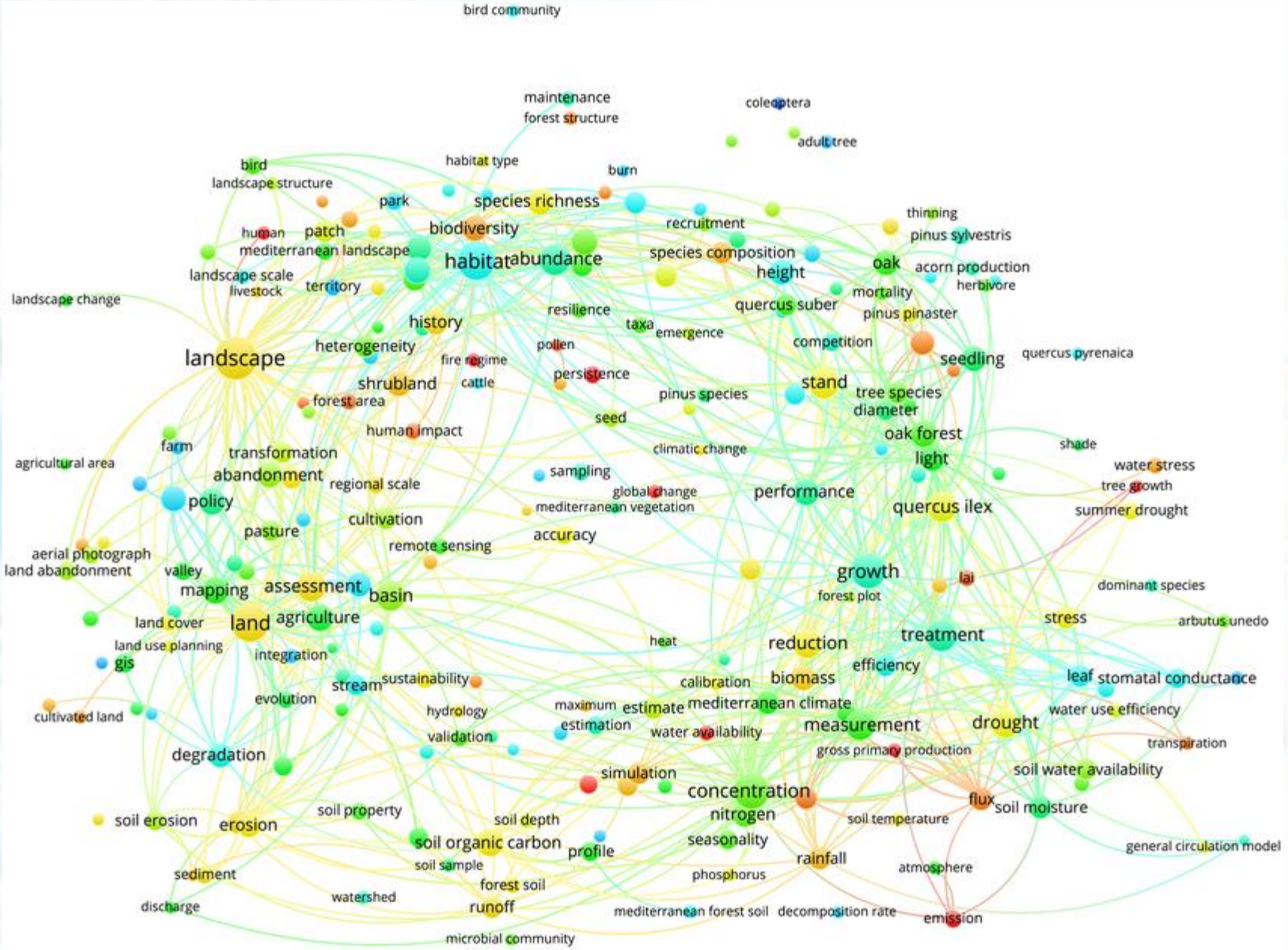
Citation Bibliographics



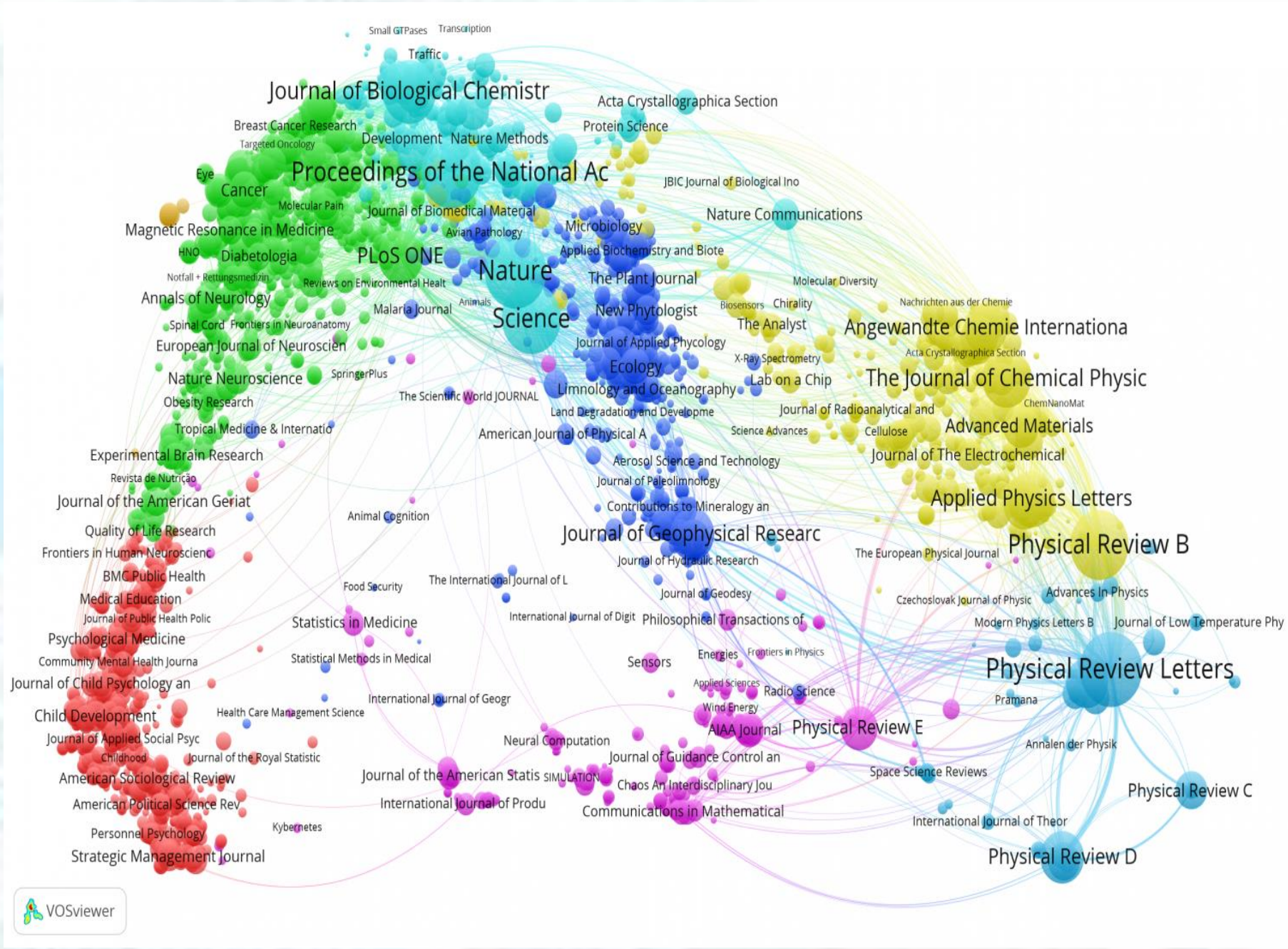
Mining Network in Social Media



Mining Term in Journals



Journals Correlations



Data Science Tools

- **Data Analysis**: RapidMiner, R Studio, MATLAB, Excel, SAS, SPSS Modeller
- **Data Warehousing**: Informatica/ Talend, AWS Redshift
- **Data Visualization**: RapidMiner, Tableau, Cognos, RAW
- **Machine Learning**: RapidMiner, Spark MLib, Mahout, Azure ML studio

Data Science Tools (Open Source)

- Rapidminer Studio Educational https://rapidminer.com/platform/educational/
- Oranges https://orangedatamining.com/
- Weka https://www.cs.waikato.ac.nz/ml/weka/index.html
- Power BI (Free Version) https://powerbi.microsoft.com/en-us/downloads/
- Knime https://www.knime.com/
- Pythons https://www.python.org/downloads/
- R https://cran.r-project.org/bin/windows/base/
- VosViewer https://www.vosviewer.com/download

Data Science Tools (Commercials)

- Tableu https://www.tableau.com/
- IBM SPSS Modeler https://www.ibm.com/support/pages/downloading-ibm-spss-modeler-1822
- Power BI (commercial) https://powerbi.microsoft.com/en-us/downloads/
- MATLAB https://au.mathworks.com/solutions/data-science.html
- SAS https://www.sas.com/en_au/software/stat.html

Terima Kasih

Semarang, 31 August 2022