

COMP6237 Data Mining: Introduction to Data Mining

Shoaib Ehsan (module leader), Zhiwu Huang and
Markus Brede

s.ehsan@soton.ac.uk

zhiwu.huang@soton.ac.uk

Markus.Brede@soton.ac.uk

Teaching Staff

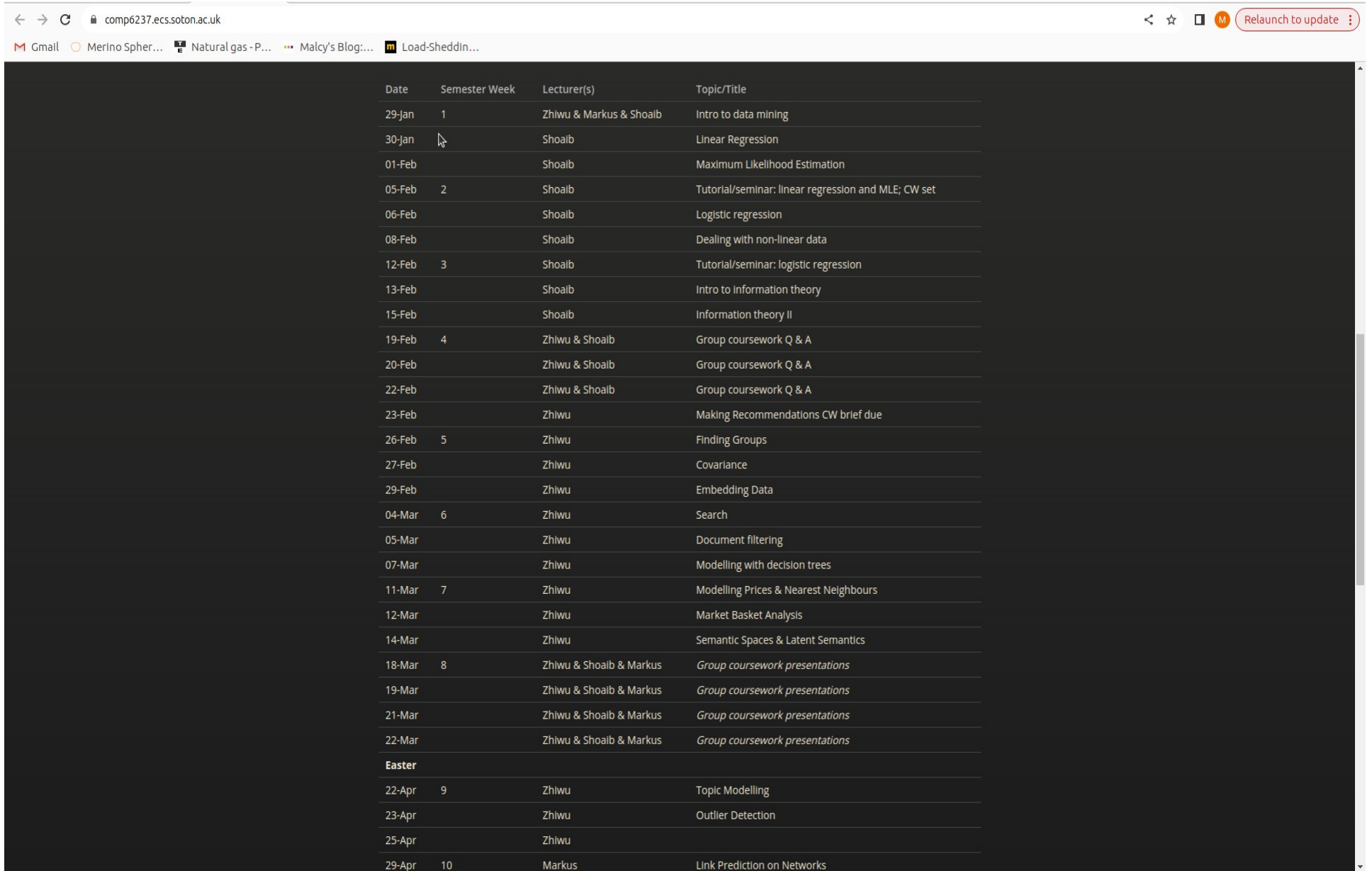
- Credit goes to Jon Hare and Jo Grundy who developed a large part of the module
- Shoaib Ehsan – AIC
 - s.ehsan@soton.ac.uk
 - 32/3001
- Zhiwu Huang – VLC
 - zhiwu.huang@soton.ac.uk
 - 59/4203
- Markus Brede – AIC
 - Markus.Brede@soton.ac.uk
 - 32/4033

Module Overview

- Not quite so new module, run for the 7th time
 - See feedback from last year
- Created to fill a gap
 - Data mining is almost synonymous with advanced machine learning
 - Inevitably some overlaps with COMP3206/COMP6208
 - Should be complementary and offer different views
 - Slightly more applied pragmatic focus
 - How do you work with real world data?
 - How do you solve real problems?

Module Structure

- Around 26 lectures + additional tutorials
 - Wide range of data mining topics
- Assessment
 - 70% 2 hour examination
 - 30% Group coursework



Date	Semester Week	Lecturer(s)	Topic/Title
29-Jan	1	Zhiwu & Markus & Shoalb	Intro to data mining
30-Jan		Shoalb	Linear Regression
01-Feb		Shoalb	Maximum Likelihood Estimation
05-Feb	2	Shoalb	Tutorial/seminar: linear regression and MLE; CW set
06-Feb		Shoalb	Logistic regression
08-Feb		Shoalb	Dealing with non-linear data
12-Feb	3	Shoalb	Tutorial/seminar: logistic regression
13-Feb		Shoalb	Intro to information theory
15-Feb		Shoalb	Information theory II
19-Feb	4	Zhiwu & Shoalb	Group coursework Q & A
20-Feb		Zhiwu & Shoalb	Group coursework Q & A
22-Feb		Zhiwu & Shoalb	Group coursework Q & A
23-Feb		Zhiwu	Making Recommendations CW brief due
26-Feb	5	Zhiwu	Finding Groups
27-Feb		Zhiwu	Covariance
29-Feb		Zhiwu	Embedding Data
04-Mar	6	Zhiwu	Search
05-Mar		Zhiwu	Document filtering
07-Mar		Zhiwu	Modelling with decision trees
11-Mar	7	Zhiwu	Modelling Prices & Nearest Neighbours
12-Mar		Zhiwu	Market Basket Analysis
14-Mar		Zhiwu	Semantic Spaces & Latent Semantics
18-Mar	8	Zhiwu & Shoalb & Markus	Group coursework presentations
19-Mar		Zhiwu & Shoalb & Markus	Group coursework presentations
21-Mar		Zhiwu & Shoalb & Markus	Group coursework presentations
22-Mar		Zhiwu & Shoalb & Markus	Group coursework presentations
Easter			
22-Apr	9	Zhiwu	Topic Modelling
23-Apr		Zhiwu	Outlier Detection
25-Apr		Zhiwu	
29-Apr	10	Markus	Link Prediction on Networks

Module Timetable

- We have 4 slots timetabled for every week
 - Mon 9am
 - Tue 9am
 - Thu 10am.
 - Fri 1pm
- **Will not use all slots** every week (some weeks we'll use all of them, in other weeks only 2 of them)
 - Will typically use Mon, Tue, Thu
 - Have a look at the course webpage!
 - This may sometimes also change – we'll update you by email (check ECS module page)
- Roughly the plan is:
Shoai – Zhiwu – Markus -- Revisions

Coursework Timetable

- Group coursework
 - Set next week; report submission at the end of the term (May 16)
 - Will have presentation sessions before Easter
 - More in CW Q & A sessions in week 4; by that time we want you to have formed groups
 - Once you have formed a group, please enter into this wiki:
<https://secure.ecs.soton.ac.uk/student/wiki/w/COMP6237-2023-classlist>

Resources

- Course website [handouts, slides, interactive demos]
 - <http://comp6237.ecs.soton.ac.uk>
- ECS module pages [syllabus, announcements]
 - <https://secure.ecs.soton.ac.uk:/module/comp6237>
- Reading material
 - Toby Segaran. Programming Collective Intelligence: Building Smart Web 2.0 Applications. O'Reilly, 2007
 - Aurélien Géron. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media. March 2017
 - J. Leskovec et al. Mining of Massive Datasets. Third Edition. Cambridge University Press. 2020
 - M. J. Zaki and W. Meira, Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Cambridge University Press. 2020.

What is Data Mining?

“Data mining is an *interdisciplinary* subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of *artificial intelligence, machine learning, statistics, and database systems*. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.”

– wikipedia

What is Data Mining?

“Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.”

– Bill Palace, Anderson Graduate School of Management at UCLA, 1996

DATA
(input)

Data Mining

INFORMATION
(output)



What is Data?

- Data is any sequence of one or more symbols given meaning by specific act(s) of interpretation.
- Data (or datum - a single unit of data) is not information.
 - Data requires interpretation to become information.
 - To translate data to information, there must be several known factors considered. The factors involved are determined by the creator of the data and the desired information.

What is Information?

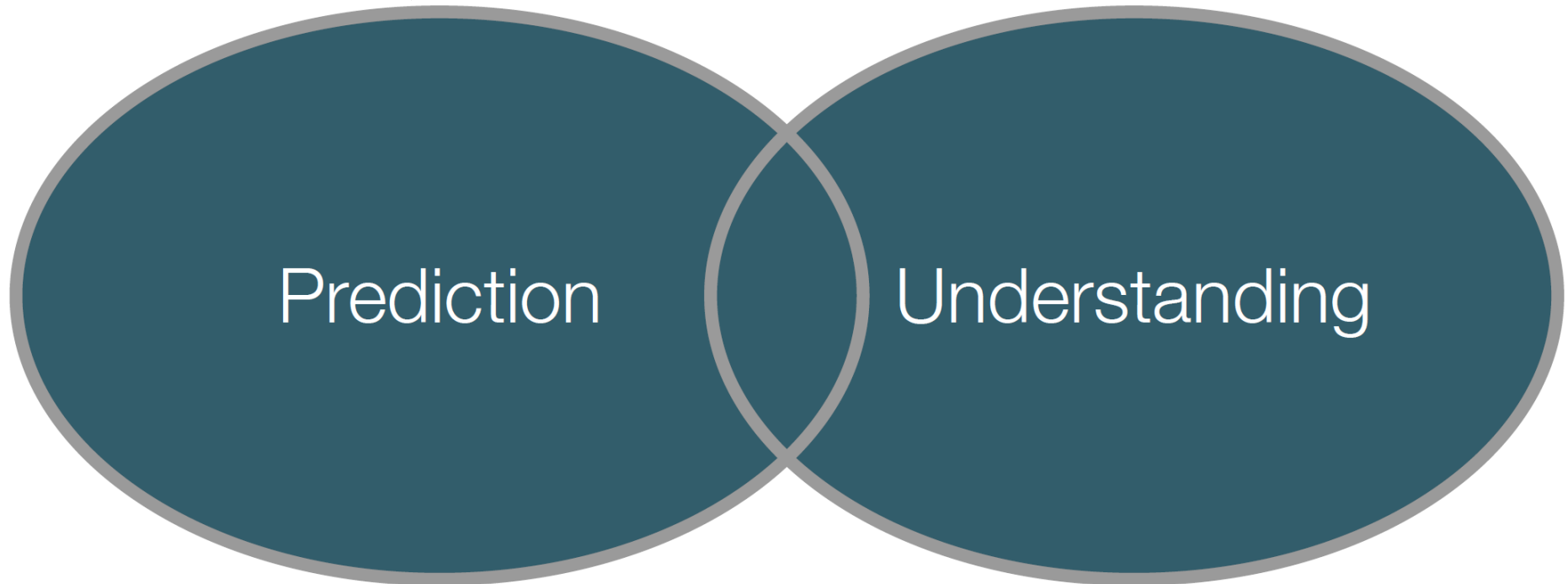
- There is a formal definition → Information theory ... will have a bit of a look at this later.
- “Actionable knowledge”
 - **Prediction**
 - Christoph Adami (Michigan State) defines information as: ‘the ability to make predictions with a likelihood better than chance’.
 - **Understanding**
 - Making sense of the data

What is Data Mining?

- Given lots of data ...
- **Discover patterns and models** that are:
 - **Valid**: hold on new data with some certainty
 - **Useful**: should be possible to act on the item
 - **Unexpected**: non-obvious to the system
 - **Understandable**: humans should be able to interpret the pattern

Two Complementary Goals of Data Mining

Use some variables to predict unknown or future values of other variables



Find human-interpretable patterns that describe the data



What kinds of data are we interested in mining?

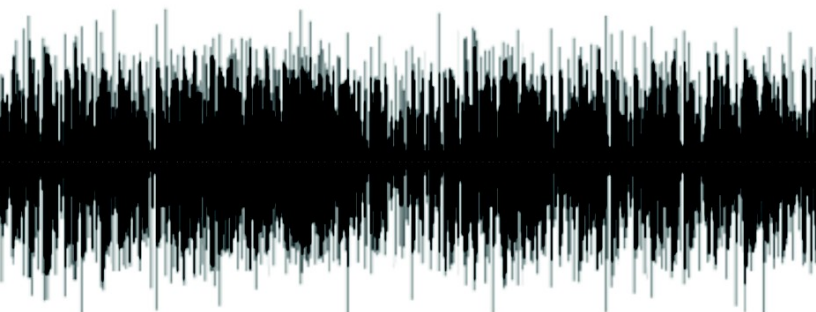
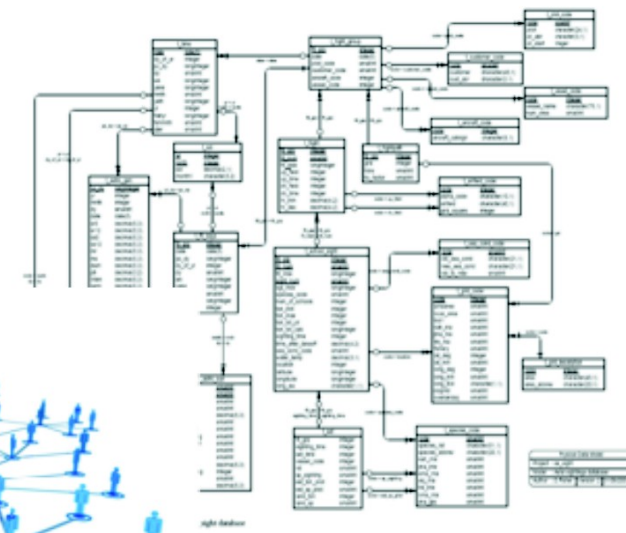


Period Ending	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,300.00	\$6,250.00	\$5,100.00	\$6,150.00	\$8,100.00	\$8,050.00	\$11,100.00
Budget	\$4,700.00	\$5,078.00	\$4,754.00	\$5,531.00	\$7,744.00	\$8,541.00	\$10,976.00
Over / (Under Budget)	\$100.00	\$1,172.00	\$346.00	\$619.00	\$356.00	\$209.00	\$1,124.00
Product 2	\$5,300.00	\$6,250.00	\$5,400.00	\$6,130.00	\$7,760.00	\$7,699.00	\$11,600.00
Budget	\$4,500.00	\$5,078.00	\$4,754.00	\$5,531.00	\$7,744.00	\$8,541.00	\$10,976.00
Over / (Under Budget)	\$800.00	\$1,172.00	\$646.00	\$600.00	\$1,016.00	\$1,158.00	\$624.00
Product 3	\$14,000.00	\$16,250.00	\$13,100.00	\$16,150.00	\$22,100.00	\$21,764.00	\$31,400.00
Budget	\$12,800.00	\$13,078.00	\$12,754.00	\$15,531.00	\$19,830.00	\$19,811.00	\$27,208.00
Over / (Under Budget)	\$1,200.00	\$3,172.00	\$436.00	\$619.00	\$2,270.00	\$1,953.00	\$4,192.00
Product 4	\$75,000.00	\$77,000.00	\$78,000.00	\$79,000.00	\$80,000.00	\$81,000.00	\$82,000.00
Budget	\$68,000.00	\$74,000.00	\$71,000.00	\$77,000.00	\$79,000.00	\$80,000.00	\$81,000.00
Over / (Under Budget)	\$7,000.00	\$3,000.00	\$7,000.00	\$2,000.00	\$1,000.00	\$1,000.00	\$1,000.00
Product 5	\$75,000.00	\$80,250.00	\$89,000.00	\$86,250.00	\$95,000.00	\$106,000.00	\$121,250.00
Budget	\$68,000.00	\$73,000.00	\$78,000.00	\$83,000.00	\$88,000.00	\$93,000.00	\$98,000.00
Over / (Under Budget)	\$7,000.00	\$7,250.00	\$11,000.00	\$3,250.00	\$7,000.00	\$13,000.00	\$23,250.00

back in that old sea-song that he sang so often afterwards:
 "Fifteen men on the dead man's chest—Yo-ho-ho, and a bottle of rum!"
 in the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was
 berth for the crew to lie on their backs and he cried the bar and here a plain nigger's egg is up there! What you see when he

Tweets follow @twitterapi

- Twitter API** @twitterapi 11 Jun
As part of the retirement plan today, we're also about to discontinue Basic Auth support for unelevated Streaming API roles. Use OAuth 1.0A. Expand
- Twitter API** @twitterapi 11 Jun
The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/api/1.1/750 Expand
- Twitter API** @twitterapi 11 Jun
If you were using a legacy widget that no longer functions after API v1 retirement, we suggest using dev.twitter.com/docs/embedded-... instead. Show Summary



Categorizing data: Structured/ Unstructured



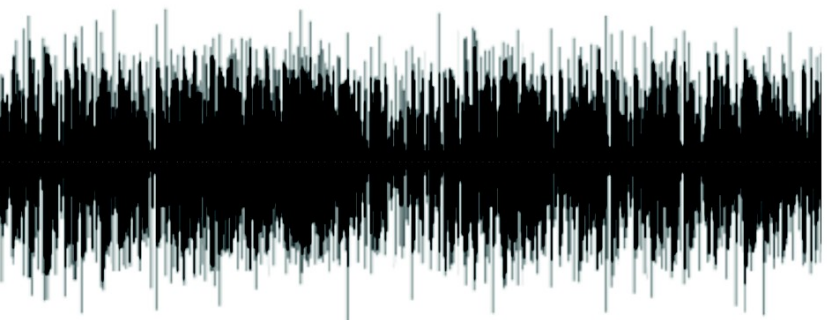
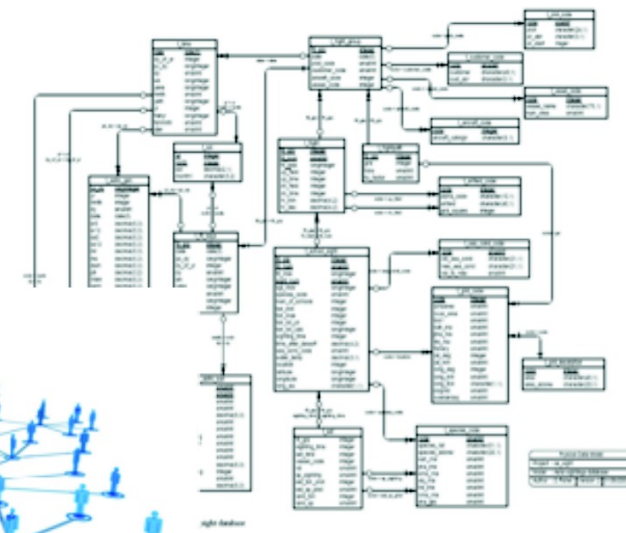
Period	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,300.00	\$6,750.00	\$5,100.00	\$6,150.00	\$8,100.00	\$8,054.00	\$11,100.00
Budget	\$4,700.00	\$5,078.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,054.00	\$10,976.00
Over / Under Budget	\$100.00	\$1,672.00	\$346.00	\$559.00	\$356.00	\$200.00	\$1,124.00
Product 2	\$5,300.00	\$6,750.00	\$5,400.00	\$6,130.00	\$7,790.00	\$7,699.00	\$11,600.00
Budget	\$4,500.00	\$5,078.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,054.00	\$10,976.00
Over / Under Budget	\$800.00	\$1,672.00	\$646.00	\$539.00	\$446.00	\$645.00	\$624.00
Product 3	\$14,000.00	\$18,160.00	\$18,100.00	\$18,880.00	\$32,100.00	\$38,764.00	\$31,400.00
Budget	\$12,800.00	\$13,078.00	\$12,754.00	\$13,591.00	\$19,830.00	\$18,111.00	\$21,200.00
Over / Under Budget	\$1,200.00	\$5,082.00	\$5,346.00	\$5,289.00	\$12,270.00	\$20,653.00	\$10,200.00
Product 4	\$16,000.00	\$17,990.00	\$18,000.00	\$17,900.00	\$32,090.00	\$27,400.00	\$21,000.00
Budget	\$15,000.00	\$16,178.00	\$17,154.00	\$17,714.00	\$29,840.00	\$27,400.00	\$31,200.00
Over / Under Budget	\$1,000.00	\$1,812.00	\$846.00	\$1,186.00	\$2,250.00	\$1,000.00	\$1,800.00
Product 5	\$78,000.00	\$80,750.00	\$89,000.00	\$86,750.00	\$96,400.00	\$106,834.00	\$121,200.00
Budget	\$68,500.00	\$78,595.00	\$78,754.00	\$86,591.00	\$77,744.00	\$99,844.00	\$111,976.00
Over / Under Budget	\$9,500.00	\$2,155.00	\$10,246.00	\$10,159.00	\$18,656.00	\$6,990.00	\$9,224.00

back in that old sea-song that he sang so often afterwards:
 "Fifteen men on the dead man's chest—Yo-ho-ho, and a bottle of rum!"
 in the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was
 berth for the crew and he cried the bar and here a plain nigger's up there! What you see when he

Tweets follow @twitterapi

- Twitter API** @twitterapi 11 Jun
As part of the retirement plan today, we're also about to discontinue Basic Auth support for unelevated Streaming API roles. Use OAuth 1.0A.
Expand
- Twitter API** @twitterapi 11 Jun
The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/api/1.7/50
Expand
- Twitter API** @twitterapi 11 Jun
If you were using a legacy widget that no longer functions after API v1 retirement, we suggest using dev.twitter.com/docs/embedded-... instead.
Show Summary

Tweet to @twitterapi



Categorizing data: Dynamic/static/stream



		Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	Budget	\$5,300.00	\$6,250.00	\$5,100.00	\$6,150.00	\$8,100.00	\$8,050.00	\$11,100.00
Product 1	Over / Under Budget	\$4,790.00	\$5,078.00	\$4,754.00	\$5,581.00	\$7,744.00	\$8,040.00	\$11,976.00
Product 2	Budget	\$5,300.00	\$6,250.00	\$5,400.00	\$6,130.00	\$7,790.00	\$7,690.00	\$11,600.00
Product 2	Over / Under Budget	\$1,305.00	\$1,730.00	\$1,646.00	\$1,493.00	\$1,060.00	\$200.00	\$1,124.00
Product 3	Budget	\$14,000.00	\$24,260.00	\$24,100.00	\$26,440.00	\$32,100.00	\$24,760.00	\$31,400.00
Product 3	Over / Under Budget	\$5,285.00	\$3,078.00	\$5,754.00	\$7,591.00	\$9,830.00	\$11,311.00	\$11,289.00
Product 4	Budget	\$7,350.00	\$11,072.00	\$10,200.00	\$12,200.00	\$12,090.00	\$12,740.00	\$12,090.00
Product 4	Over / Under Budget	\$7,355.00	\$3,012.00	\$1,076.00	\$12,159.00	\$11,760.00	\$10,120.00	\$11,190.00
Product 5	Budget	\$78,000.00	\$101,000.00	\$100,000.00	\$110,000.00	\$120,000.00	\$120,000.00	\$120,000.00
Product 5	Over / Under Budget	\$68,500.00	\$3,195.00	\$18,754.00	\$85,501.00	\$77,144.00	\$99,840.00	\$111,976.00

back in that old sea-song that he sang so often afterwards:
 "Fifteen men on the dead man's chest—Yo-ho-ho, and a bottle of rum!"
 in the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was
 berth for the crew and he cried the bar and here a plain nigger's egg is up then What you see when he three

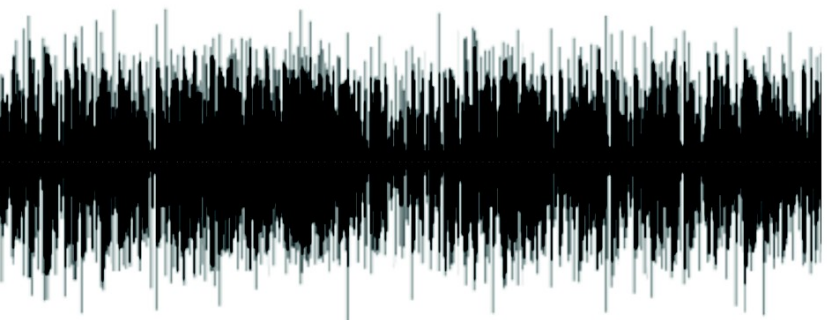
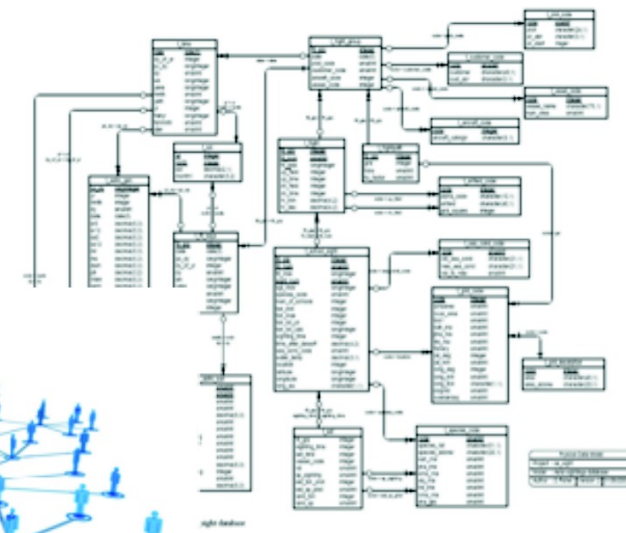
Tweets follow @twitterapi

Twitter API @twitterapi 11 Jun
 As part of the retirement plan today, we're also about to discontinue Basic Auth support for unelevated Streaming API roles. Use OAuth 1.0A. Expand

Twitter API @twitterapi 11 Jun
 The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/api/1.1/750 Expand

Twitter API @twitterapi 11 Jun
 If you were using a legacy widget that no longer functions after API v1 retirement, we suggest using dev.twitter.com/docs/embedded-... instead. Show Summary

Tweet to @twitterapi



Categorizing data: Unimodal/multimodal



Period Ending:	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,300.00	\$6,250.00	\$5,100.00	\$6,150.00	\$8,100.00	\$8,054.00	\$11,100.00
Budget	\$4,790.00	\$5,078.00	\$4,754.00	\$5,531.00	\$7,744.00	\$8,054.00	\$10,976.00
Over / (Under Budget)	\$510.00	\$1,172.00	\$346.00	\$619.00	\$356.00	\$0.00	\$1,124.00
Product 2	\$5,300.00	\$6,250.00	\$5,400.00	\$6,130.00	\$7,790.00	\$7,699.00	\$11,600.00
Budget	\$4,595.00	\$5,078.00	\$4,754.00	\$5,531.00	\$7,744.00	\$8,054.00	\$10,976.00
Over / (Under Budget)	\$705.00	\$1,172.00	\$646.00	\$600.00	\$446.00	\$645.00	\$624.00
Product 3	\$14,000.00	\$16,250.00	\$13,100.00	\$16,150.00	\$22,100.00	\$21,764.00	\$31,400.00
Budget	\$12,895.00	\$13,078.00	\$12,754.00	\$13,531.00	\$19,844.00	\$19,844.00	\$28,200.00
Over / (Under Budget)	\$1,105.00	\$3,172.00	\$456.00	\$2,619.00	\$2,256.00	\$1,920.00	\$3,200.00
Product 4	\$16,000.00	\$17,250.00	\$14,000.00	\$17,000.00	\$24,000.00	\$23,400.00	\$34,000.00
Budget	\$14,595.00	\$14,078.00	\$13,754.00	\$14,531.00	\$20,844.00	\$21,174.00	\$31,200.00
Over / (Under Budget)	\$1,405.00	\$3,172.00	\$256.00	\$2,469.00	\$3,156.00	\$2,226.00	\$2,800.00
Product 5	\$78,000.00	\$90,250.00	\$79,000.00	\$90,150.00	\$126,100.00	\$126,034.00	\$181,100.00
Budget	\$68,595.00	\$73,078.00	\$70,754.00	\$80,531.00	\$117,444.00	\$119,844.00	\$171,976.00
Over / (Under Budget)	\$9,405.00	\$17,172.00	\$8,246.00	\$9,619.00	\$8,656.00	\$6,190.00	\$9,124.00

back in that old sea-song that he sang
 id, so often afterwards:
 is still "Fifteen men on the dead man's
 e up chest-Yo-ho-ho, and a bottle of
 17— rum!" in the high, old tottering
 en my voice that seemed to have been
 bow tuned and broken at the capstan
 ran bars. Then he rapped on the door
 up his with a bit of stick like a handspike
 ere that he carried, and when my father
 ding appeared, called roughly for a glass of rum. This, when it was
 berth f he cried the bar and here a plain n
 eggs is up then What y
 ough see wh
 he three

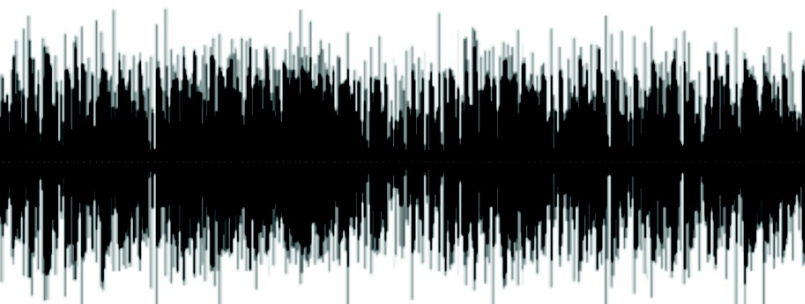
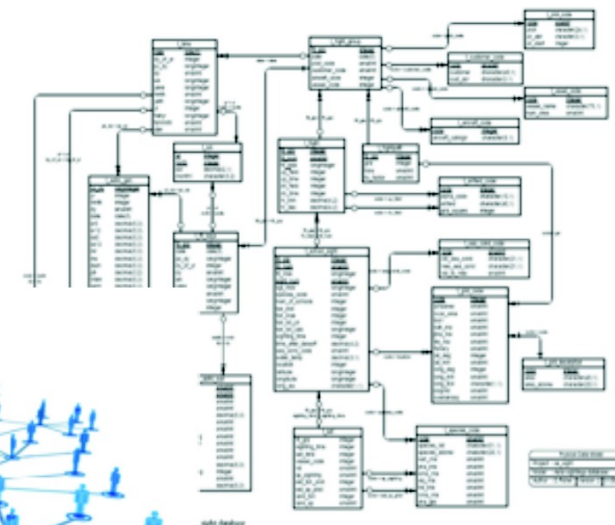
Tweets

Twitter API @twitterapi 11 Jun
 As part of the retirement plan today, we're also about to discontinue Basic Auth support for unelevated Streaming API roles. Use OAuth 1.0A.
 Expand

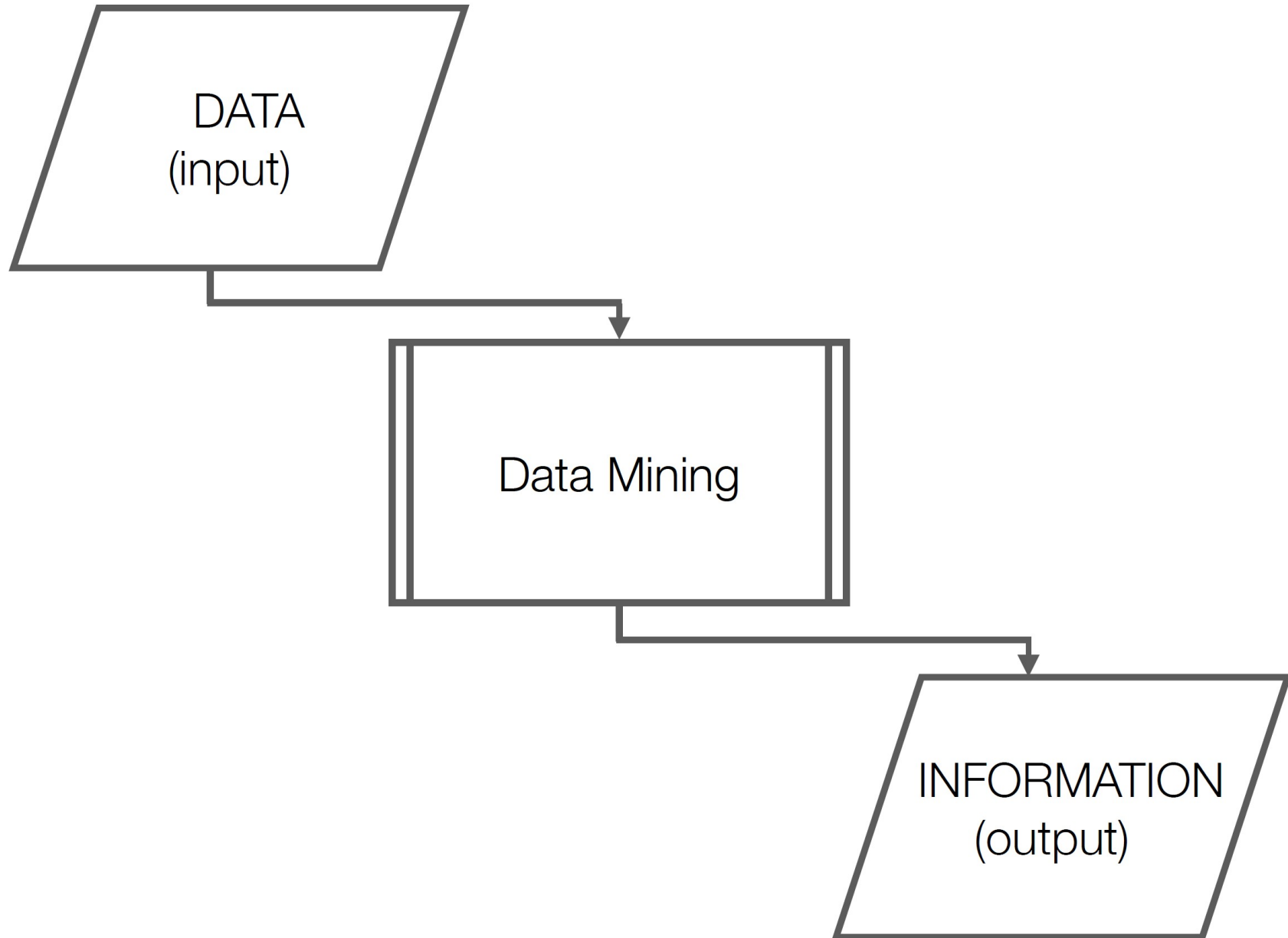
Twitter API @twitterapi 11 Jun
 The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/api/1.1/750
 Expand

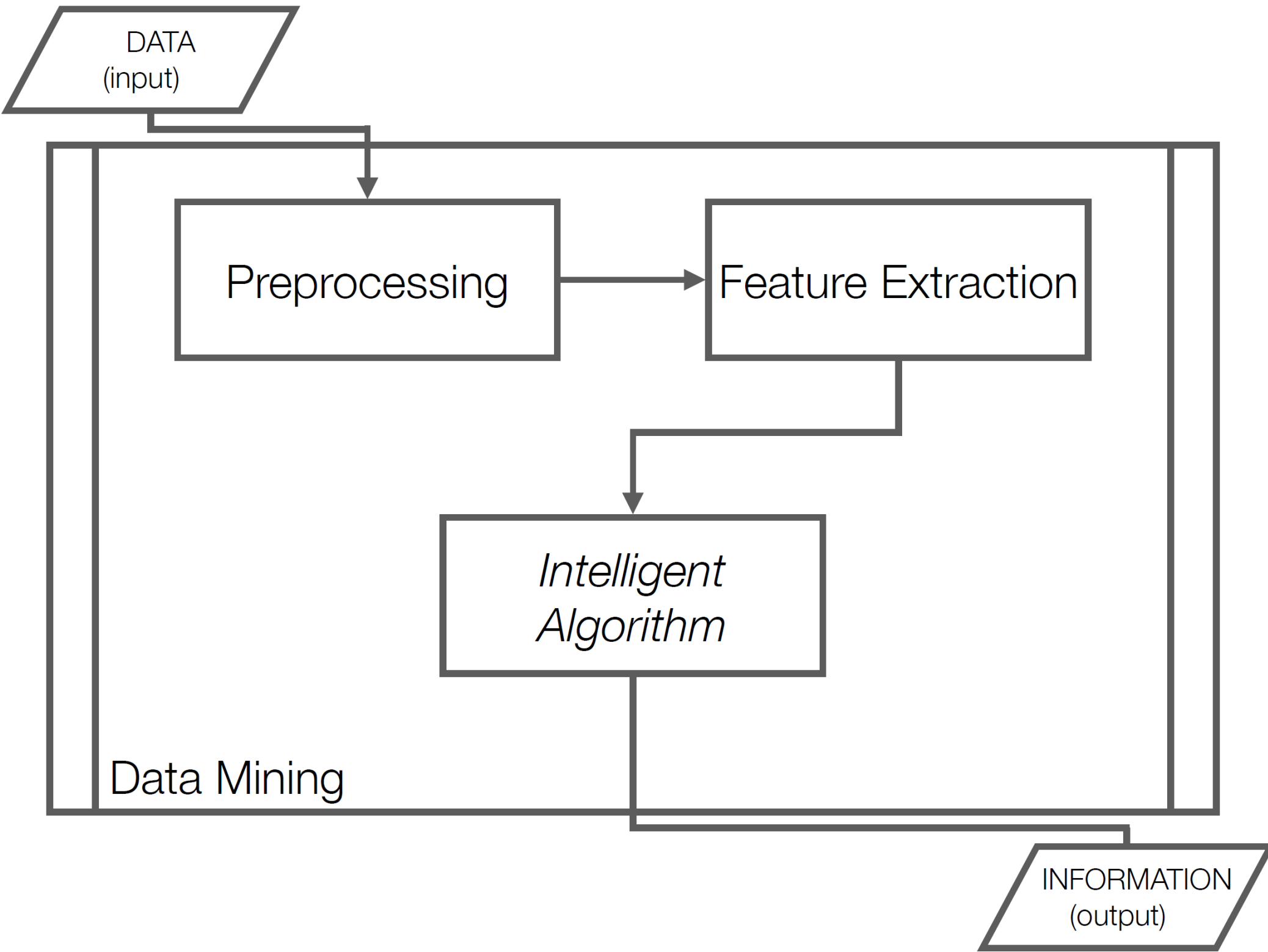
Twitter API @twitterapi 11 Jun
 If you were using a legacy widget that no longer functions after API v1 retirement, we suggest using dev.twitter.com/docs/embedded-... instead.
 Show Summary

Tweet to @twitterapi



Typical Data Mining Pipeline





Descriptive Techniques

PCA

ICA

MDS

Clustering

Anomaly Detection

...

*Intelligent
Algorithm*

Predictive Techniques

Classification

Ranking

Regression

Matrix Completion

...

The Plan for the Next 12 Weeks

- You will learn to solve real-world problems – e.g.:
 - Recommender systems
 - Market Basket Analysis
 - Document filtering and spam detection
 - Duplicate document detection
 - Link prediction
 - Community detection
 - Ranking search results
 - Social network analysis
- You will also learn various tools & techniques - e.g.:
 - Linear algebra (SVD, Eigendecomposition & PCA, NNMF, etc.)
 - Optimisation (e.g. stochastic gradient descent)
 - Dynamic programming (frequent itemsets)
 - Hashing (LSH, Sketching, Bloom Filters)
 - Statistics of regression analysis
 - Information theory
 - Network theory

The Group Coursework

- You need to form groups
 - Target size is 4 (**strictly**)
 - As a group, you need to choose a data mining problem to work on
 - (You'll need to train and evaluate models and compare their performance [possibly against approaches from others])
- Come along to the slots in week 3 to discuss your ideas for problems to work on with us
- Enter your team name and team members on the student wiki:

<https://secure.ecs.soton.ac.uk/student/wiki/w/COMP6237-2023-classlist>

Key Dates

- Each team needs to submit a 1-page project brief by the end of the day of week 4 (23rd of Feb).
- Before Easter groups must present their idea and approaches to the class.
 - Teams should be prepared to present in the first slot; to ensure fairness we will pick teams at random
- Teams must submit a conference paper by 4pm on May 16.