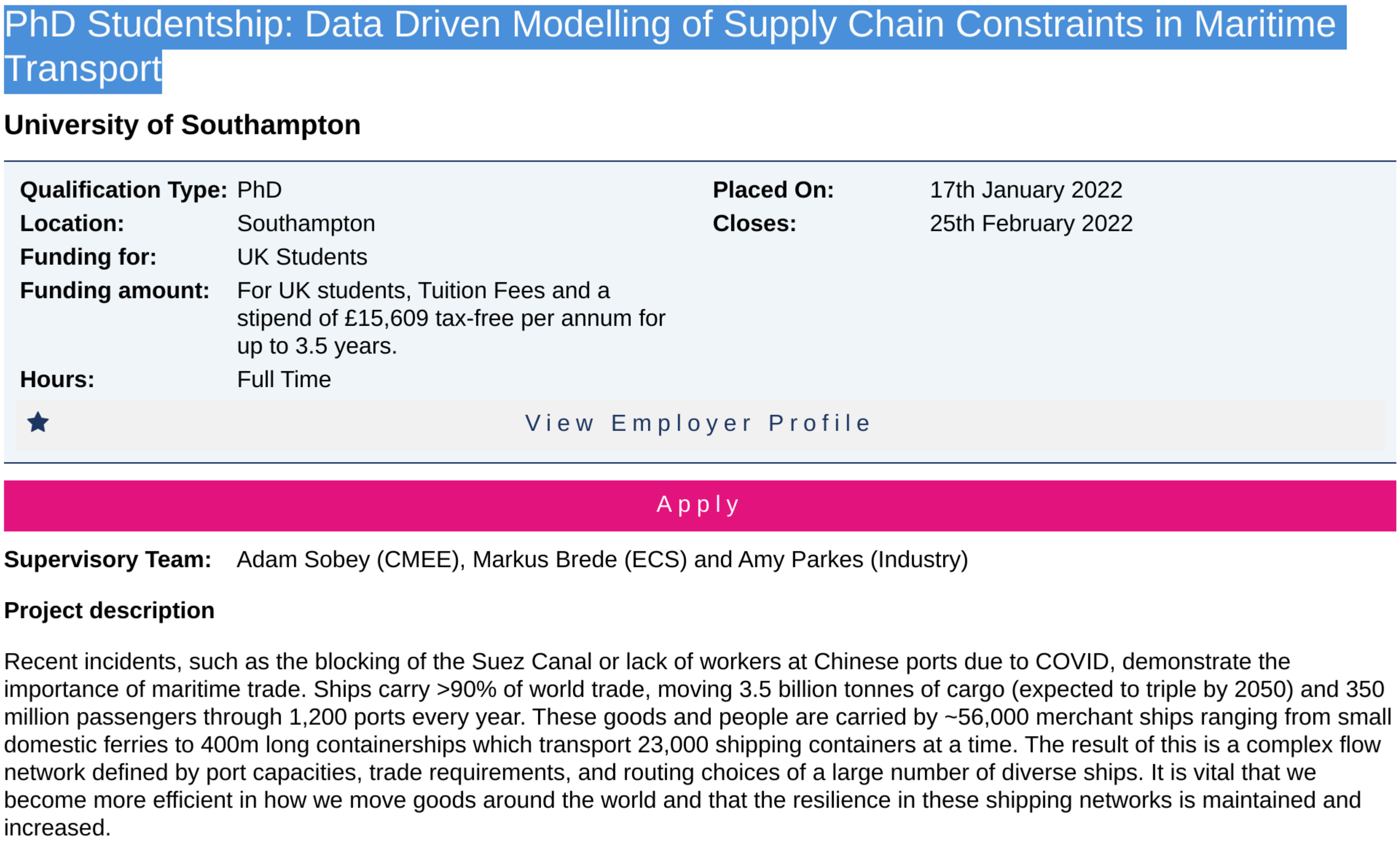
COMP6237 Data Mining:
Introduction to Data Mining

Jo Grundy and
Markus Brede

J.Grundy@soton.ac.uk

Markus.Brede@soton.ac.uk

Funded PhD Position



PhD Studentship: Data Driven Modelling of Supply Chain Constraints in Maritime Transport

University of Southampton

Qualification Type:	PhD	**Placed On:**	17th January 2022
Location:	Southampton	**Closes:**	25th February 2022
Funding for:	UK Students		
Funding amount:	For UK students, Tuition Fees and a stipend of £15,609 tax-free per annum for up to 3.5 years.		
Hours:	Full Time		

★ View Employer Profile

Apply

Supervisory Team: Adam Sobey (CMEE), Markus Brede (ECS) and Amy Parkes (Industry)

Project description

Recent incidents, such as the blocking of the Suez Canal or lack of workers at Chinese ports due to COVID, demonstrate the importance of maritime trade. Ships carry >90% of world trade, moving 3.5 billion tonnes of cargo (expected to triple by 2050) and 350 million passengers through 1,200 ports every year. These goods and people are carried by ~56,000 merchant ships ranging from small domestic ferries to 400m long containerships which transport 23,000 shipping containers at a time. The result of this is a complex flow network defined by port capacities, trade requirements, and routing choices of a large number of diverse ships. It is vital that we become more efficient in how we move goods around the world and that the resilience in these shipping networks is maintained and increased.

Teaching Staff

- Credit goes to Jon Hare who developed a large part of the module

- Jo Grundy – VLC
  - Jo.Grundy@soton.ac.uk
  - 32/4053

- Markus Brede – AIC
  - Markus.Brede@soton.ac.uk
  - 32/4033

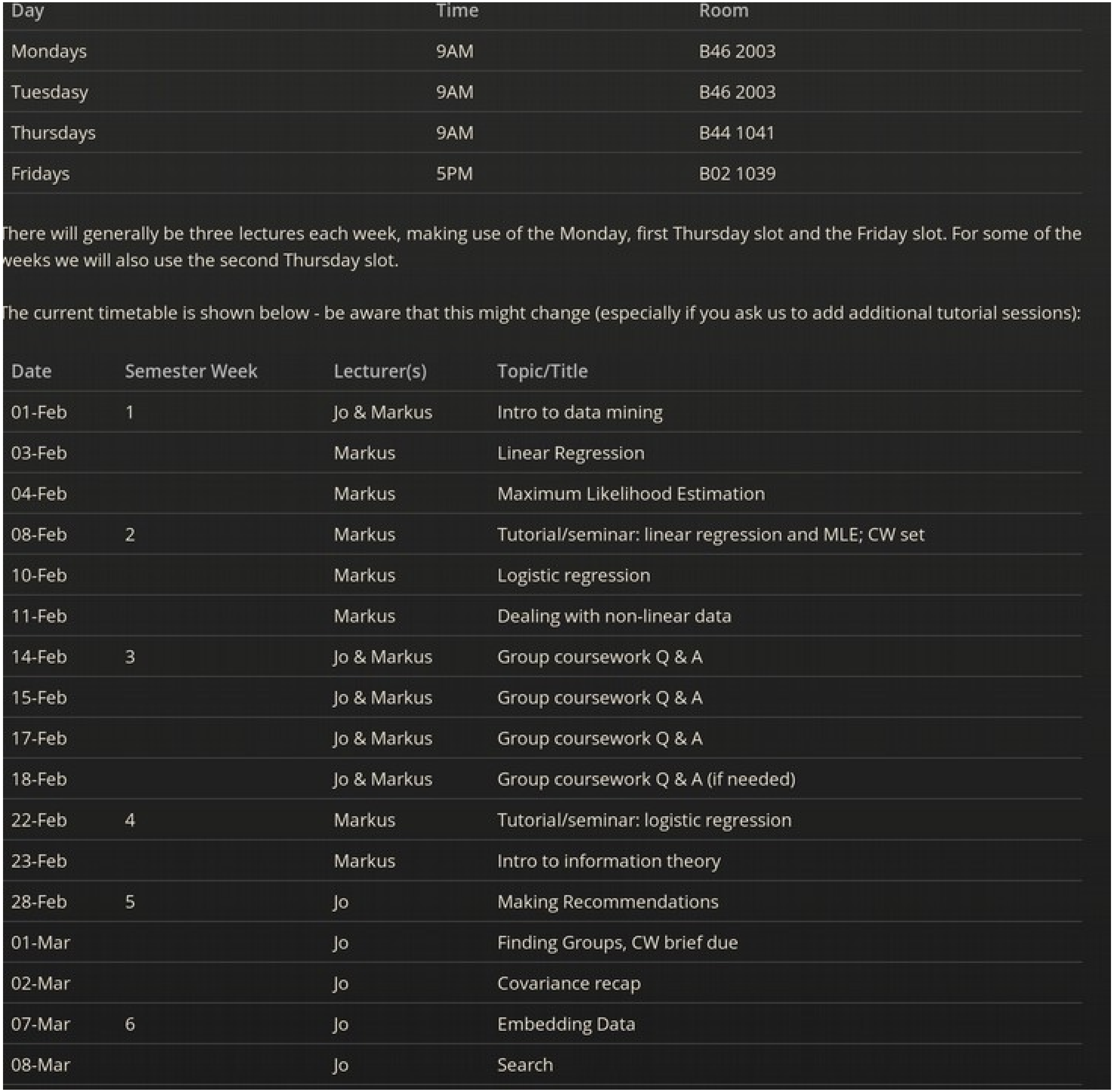
Module Overview

- 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

http://comp6237.ecs.soton.ac.uk/



Day	Time	Room
Mondays	9AM	B46 2003
Tuesdasy	9AM	B46 2003
Thursdays	9AM	B44 1041
Fridays	5PM	B02 1039

There will generally be three lectures each week, making use of the Monday, first Thursday slot and the Friday slot. For some of the weeks we will also use the second Thursday slot.

The current timetable is shown below - be aware that this might change (especially if you ask us to add additional tutorial sessions):

Date	Semester Week	Lecturer(s)	Topic/Title
01-Feb	1	Jo & Markus	Intro to data mining
03-Feb		Markus	Linear Regression
04-Feb		Markus	Maximum Likelihood Estimation
08-Feb	2	Markus	Tutorial/seminar: linear regression and MLE; CW set
10-Feb		Markus	Logistic regression
11-Feb		Markus	Dealing with non-linear data
14-Feb	3	Jo & Markus	Group coursework Q & A
15-Feb		Jo & Markus	Group coursework Q & A
17-Feb		Jo & Markus	Group coursework Q & A
18-Feb		Jo & Markus	Group coursework Q & A (if needed)
22-Feb	4	Markus	Tutorial/seminar: logistic regression
23-Feb		Markus	Intro to information theory
28-Feb	5	Jo	Making Recommendations
01-Mar		Jo	Finding Groups, CW brief due
02-Mar		Jo	Covariance recap
07-Mar	6	Jo	Embedding Data
08-Mar		Jo	Search

Module Timetable

- We have 4 slots timetabled for every week
  - Mon 9am
  - Tue 9am
  - Thu 9am.
  - Fri 5pm
- All sessions will be recorded.
- 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 Tue, Thu, Fri
  - 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:

  Markus – Jo – Markus – Revisions

Coursework Timetable

- Group coursework
  - Set next week; report submission at the end of the term (May 13)
  - Will have presentation sessions after Easter
  - More in CW Q & A sessions in week 3; 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/COMP6237-2022-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: Fundamenal Concepts and Algorithms. Cambridge University Press. 2020.

What is Data Mining?

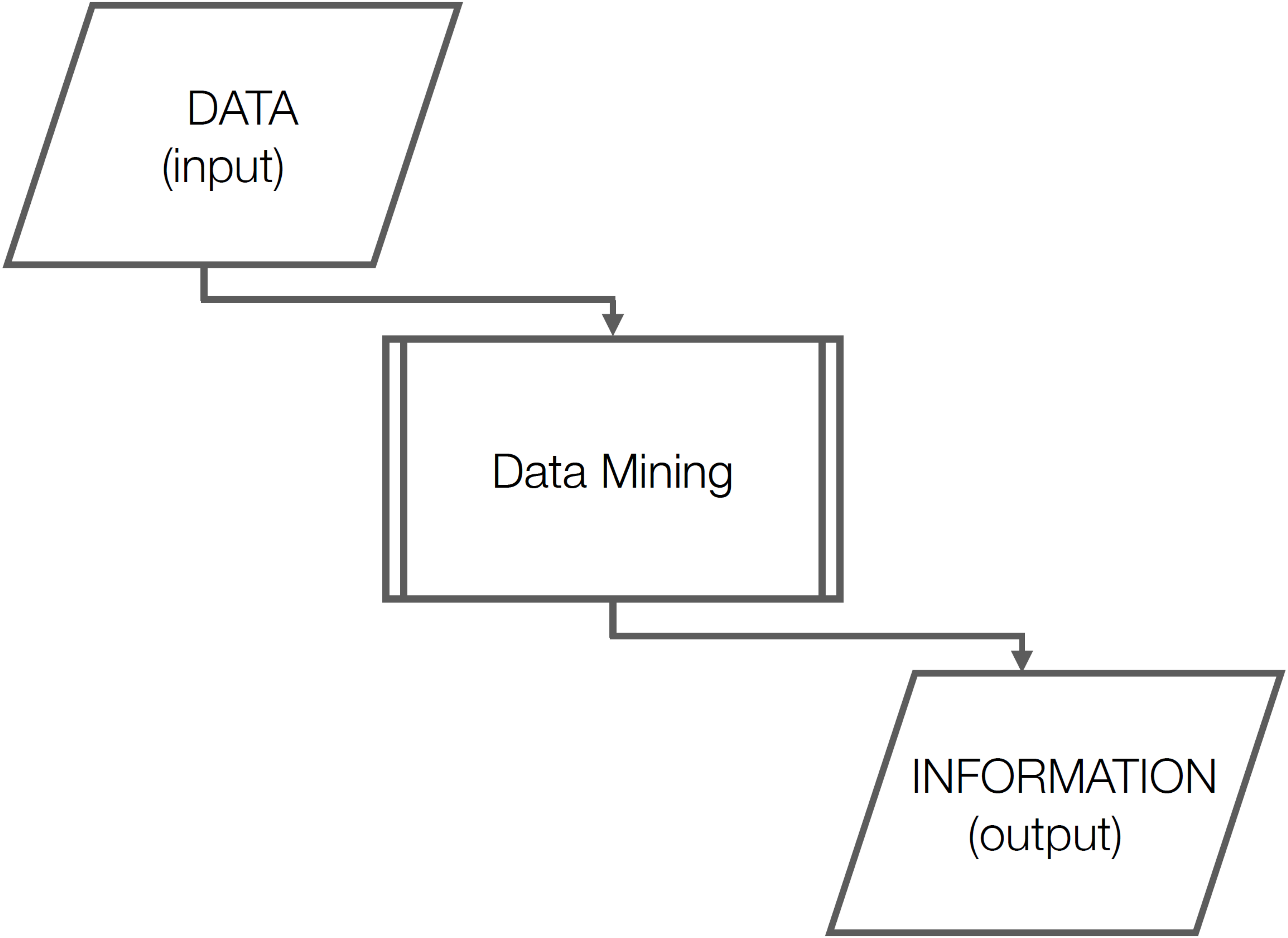
“Data mining is an *interdisciplinary* subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of *artificial intelligence, machine learning, statistics, and database systems*. The overall goal of the data mining process 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.

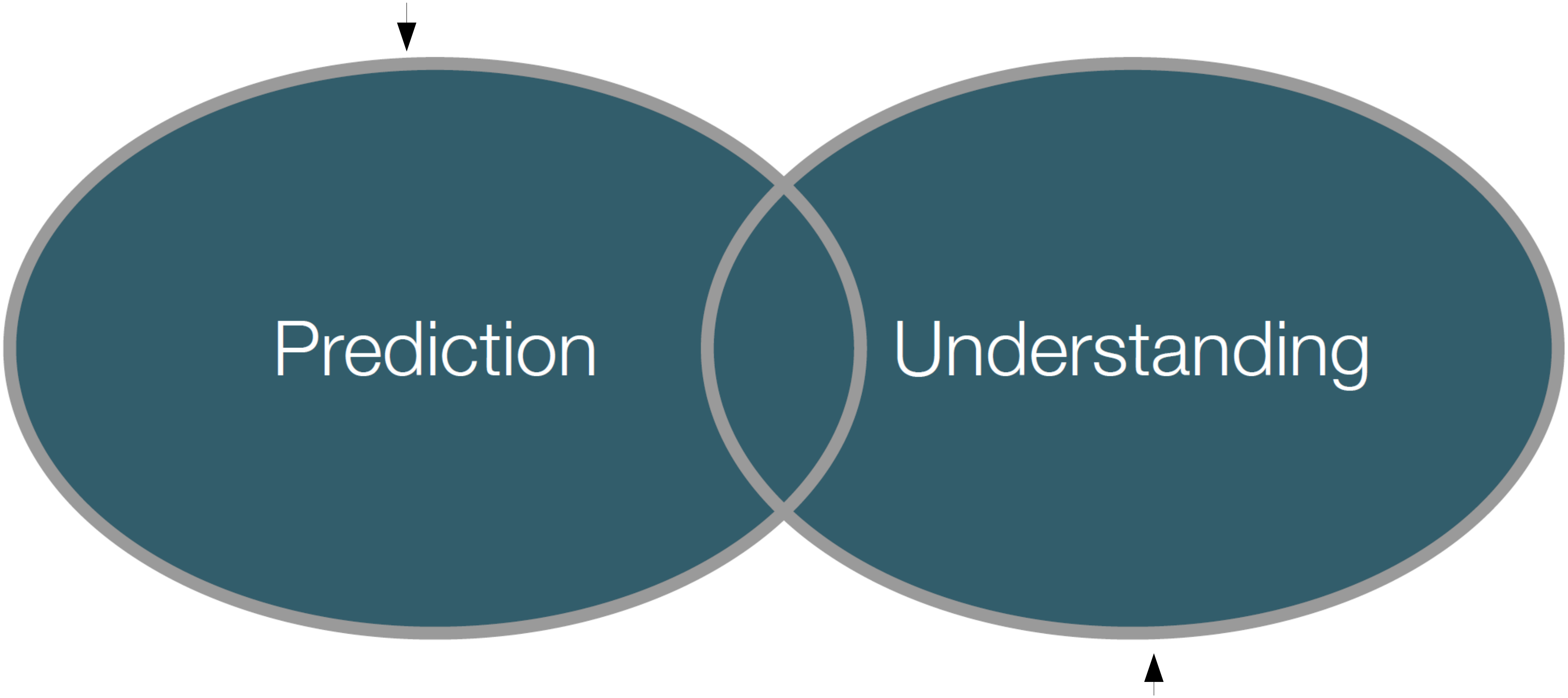
- “Actionable knowledge”
  - **Prediction**
    - Christoph Adami (Michigan State) defines information as: ‘the ability to make predictions with a likelihood better than chance’.
  - **Understanding**
    - Making sense of the data

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

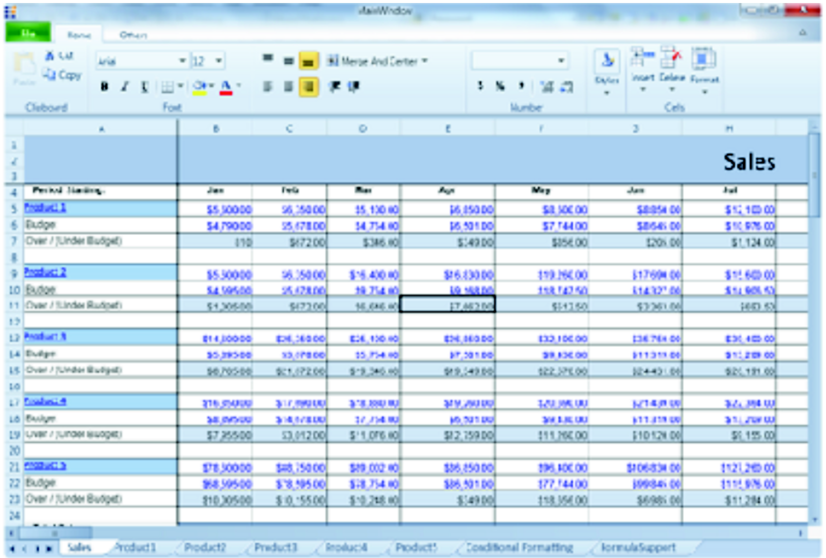


Prediction

Understanding

Find human-interpretable patterns that describe the data

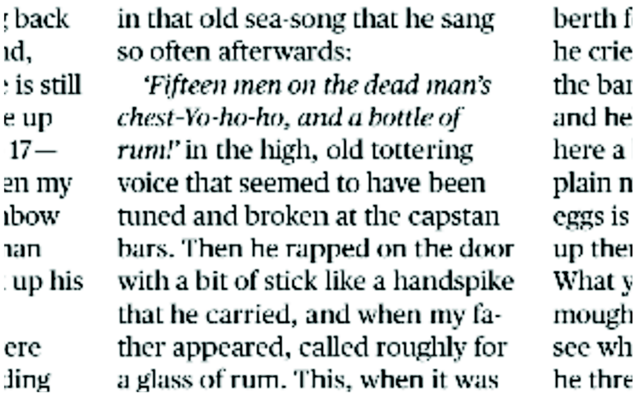
What kinds of data are we interested in mining?



Sales

	A	B	C	D	E	F	G	H
4	Period Ending	Jan	Feb	Mar	Apr	May	Jun	Jul
5	Product 1	$5,500.00	$6,350.00	$5,100.00	$6,850.00	$8,500.00	$8,854.00	$12,100.00
6	Budge	$4,790.00	$5,678.00	$4,714.00	$6,501.00	$7,744.00	$8,646.00	$10,976.00
7	Over / (Under Budget)	$10	$672.00	$386.00	$349.00	$856.00	$208.00	$1,124.00
8								
9	Product 2	$5,500.00	$6,350.00	$16,400.00	$16,830.00	$19,266.00	$17,694.00	$15,600.00
10	Budge	$4,595.00	$5,678.00	$9,714.00	$9,168.00	$18,742.50	$14,327.06	[illegible]
11	Over / (Under Budget)	$1,305.00	$672.00	$6,686.00	$7,662.00	$513.50	$3,361.00	[illegible]
12								
13	Product 3	[illegible]	[illegible]	[illegible]	[illegible]	$32,100.00	$36,764.00	$30,400.00
14	Budge	$5,395.00	$3,678.00	$5,754.00	$7,501.00	$9,836.00	$11,311.00	$11,209.00
15	Over / (Under Budget)	$8,705.00	$21,672.00	$19,346.00	$19,349.00	$22,275.00	$24,451.00	$20,191.00
16								
17	Product 4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	Budge	[illegible]	[illegible]	[illegible]	$6,501.00	[illegible]	[illegible]	[illegible]
19	Over / (Under Budget)	$7,355.00	$3,012.00	$11,076.00	$12,759.00	$11,266.00	$10,124.00	$9,155.00
20								
21	Product 5	$78,300.00	$88,750.00	$89,002.00	$86,850.00	$96,400.00	$106,834.00	$121,260.00
22	Budge	$68,595.00	$78,595.00	$78,754.00	$86,501.00	$77,744.00	$99,846.00	$111,976.00
23	Over / (Under Budget)	$10,305.00	$10,155.00	$10,248.00	$349.00	$18,356.00	$6,986.00	$11,284.00
24								

Sales
Product1
Product2
Product3
Product4
Product5
Conditional Formatting
FormulaSupport



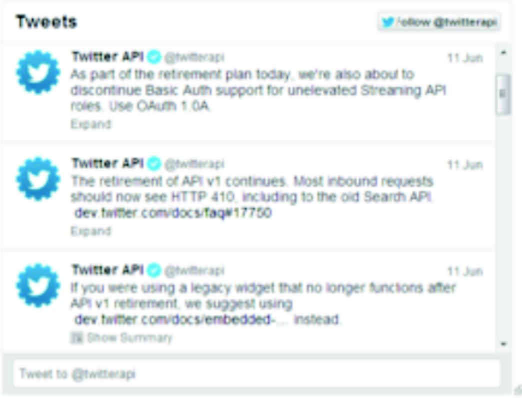
g back
nd,
e is still
e up
17—
en my
nbow
nan
 up his

ere
ding

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 f
he crie
the bar
and he
here a
plain n
eggs is
up the
What y
mough
see wh
he thre

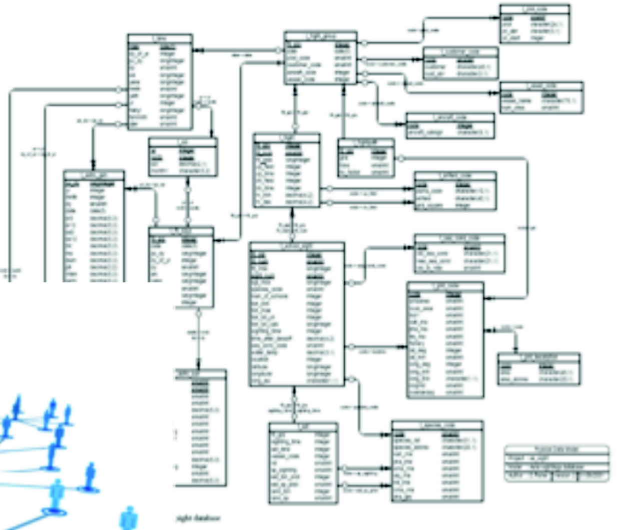


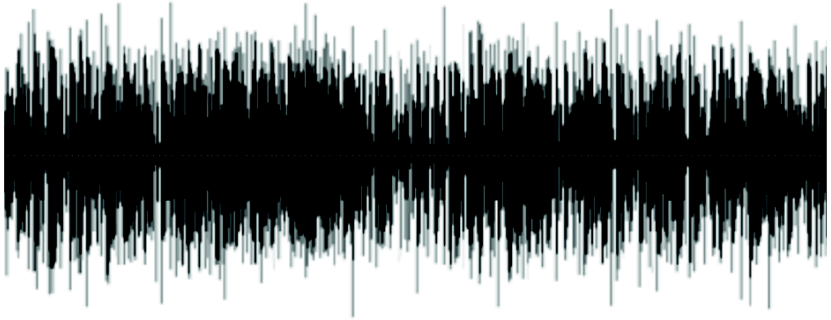
Tweets

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

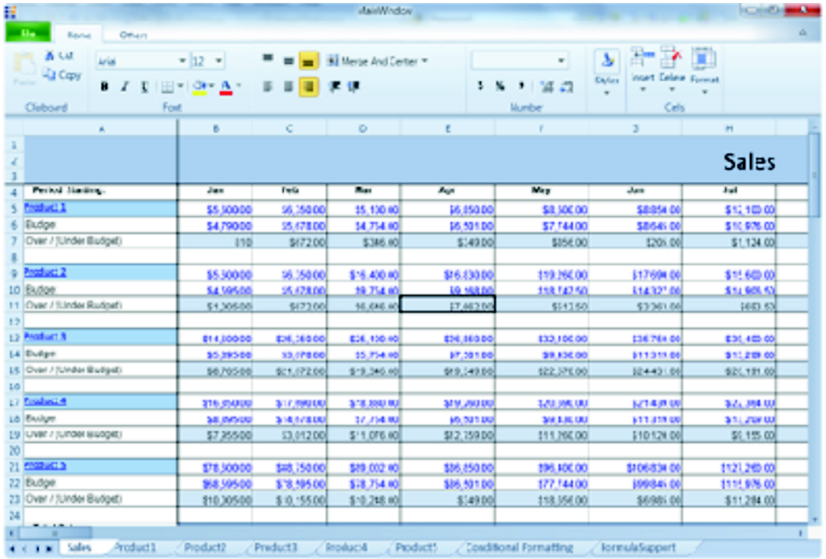
Twitter API @twitterapi 11 Jun
The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/faq#17750
Expand

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

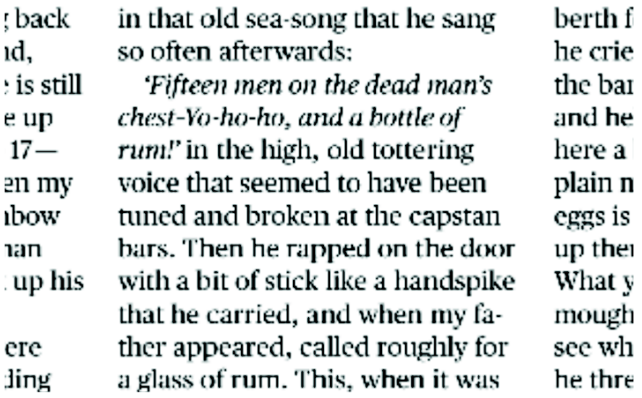




Categorizing data: Structured/ Unstructured



	A	B	C	D	E	F	G	H
1								
2								Sales
3								
4	Period Starting	Jan	Feb	Mar	Apr	May	Jun	Jul
5	Product 1	$5,500.00	$6,250.00	$5,130.00	$6,850.00	$8,500.00	$8854.00	$12,100.00
6	Budge	$4,790.00	$5,478.00	$4,754.00	$6,501.00	$7,744.00	$8646.00	$10,976.00
7	Over / (Under Budget)	$10	$672.00	$386.00	$349.00	$856.00	$208.00	$1,124.00
8								
9	Product 2	$5,500.00	$6,250.00	$16,400.00	$16,030.00	$19,296.00	$17694.00	$15,600.00
10	Budge	$4,595.00	$5,478.00	$9,754.00	$9,168.00	$18,747.50	$14327.00	$14,906.63
11	Over / (Under Budget)	$1,305.00	$672.00	$6,646.00	$7,462.00	$513.50	$3361.00	$693.63
12								
13	Product 3	$14,000.00	$26,250.00	$26,150.00	$26,850.00	$33,100.00	$36,764.00	$36,400.00
14	Budge	$5,295.00	$5,078.00	$5,754.00	$7,501.00	$9,836.00	$11,311.00	$11,209.00
15	Over / (Under Budget)	$8,705.00	$21,172.00	$19,346.00	$19,349.00	$22,376.00	$24453.00	$26,191.00
16								
17	Product 4	$16,050.00	$17,090.00	$18,880.00	$19,260.00	$20,096.00	$21489.00	[illegible]
18	Budget	$8,695.00	$14,078.00	$7,754.00	$6,501.00	$9,836.00	$11,319.00	$11,209.00
19	Over / (Under Budget)	$7,355.00	$3,012.00	$11,076.00	$12,759.00	$11,260.00	$10124.00	$9,155.00
20								
21	Product 5	$78,500.00	$88,750.00	$89,002.00	$86,850.00	$96,400.00	$106834.00	$121,260.00
22	Budge	$68,595.00	$78,595.00	$78,754.00	$86,501.00	$77,744.00	$99846.00	$111,976.00
23	Over / (Under Budget)	$10,305.00	$10,155.00	$10,248.00	$349.00	$18,356.00	$6986.00	$11,284.00
24								



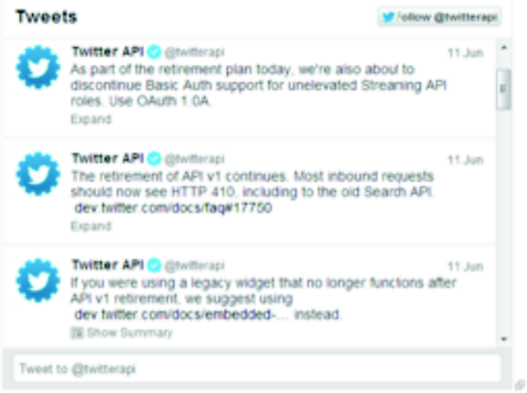
back
id,
is still
e up
17—
en my
ibow
ian
up his

ere
ding

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 f
he crie
the bar
and he
here a
plain n
eggs is
up the
What y
mough
see wh
he thre

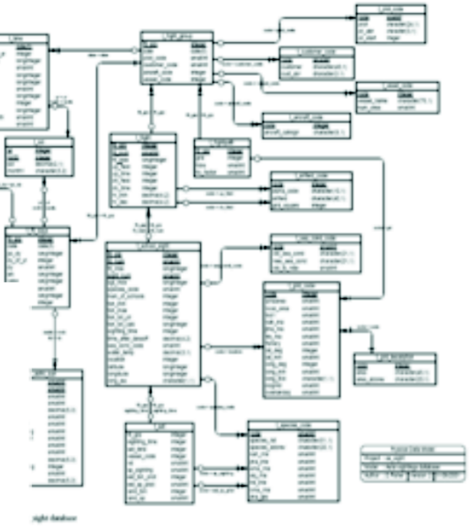


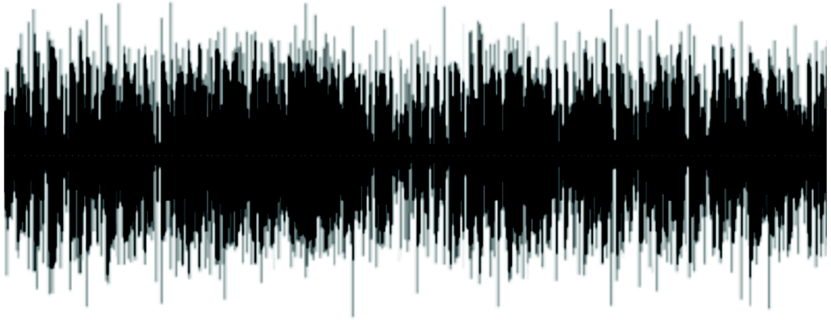
Tweets

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.

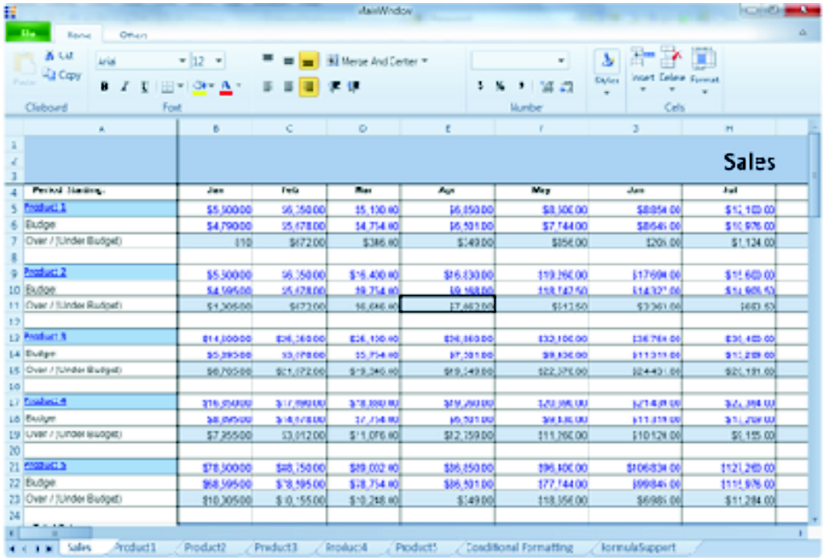
Twitter API @twitterapi 11 Jun
The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/faq#17750

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.

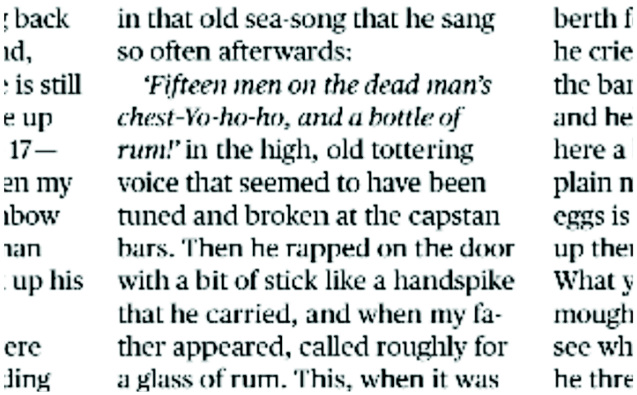




Categorizing data: Dynamic/static/stream



	A	B	C	D	E	F	G	H
1								
2								Sales
3								
4	Period Starting	Jan	Feb	Mar	Apr	May	Jun	Jul
5	Product 1	$5,300.00	$6,250.00	$5,130.00	$6,850.00	$8,300.00	$8854.00	$12,100.00
6	Budge	$4,790.00	$5,478.00	$4,754.00	$6,501.00	$7,744.00	$8646.00	$10,976.00
7	Over / (Under Budget)	$10	$672.00	$386.00	$349.00	$556.00	$208.00	$1,124.00
8								
9	Product 2	$5,300.00	$6,250.00	$16,400.00	$16,030.00	$19,266.00	$17694.00	$15,600.00
10	Budge	$4,595.00	$5,478.00	$9,754.00	$9,168.00	$18,747.50	$14327.00	$14,906.50
11	Over / (Under Budget)	$1,305.00	$672.00	$6,646.00	$7,462.00	$513.50	$3361.00	$693.50
12								
13	Product 3	$14,300.00	$26,250.00	$26,130.00	$26,860.00	$32,100.00	$36764.00	$30,400.00
14	Budge	$5,395.00	$5,078.00	$5,754.00	$7,301.00	$9,836.00	$11311.00	$11,209.00
15	Over / (Under Budget)	$8,705.00	$21,172.00	$19,346.00	$19,549.00	$22,376.00	$24431.00	$20,191.00
16								
17	Product 4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	Budget	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19	Over / (Under Budget)	$7,355.00	$3,012.00	$11,076.00	$12,759.00	$11,260.00	$10129.00	$9,155.00
20								
21	Product 5	$78,300.00	$88,750.00	$89,002.00	$86,850.00	$96,100.00	$106834.00	$121,260.00
22	Budge	$68,595.00	$78,595.00	$78,754.00	$86,501.00	$77,744.00	$99846.00	$111,976.00
23	Over / (Under Budget)	$10,305.00	$10,155.00	$10,248.00	$349.00	$18,356.00	$6986.00	$11,284.00
24								



; back
nd,
: is still
e up
17—
en my
nbow
nan
up his

ere
ding

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

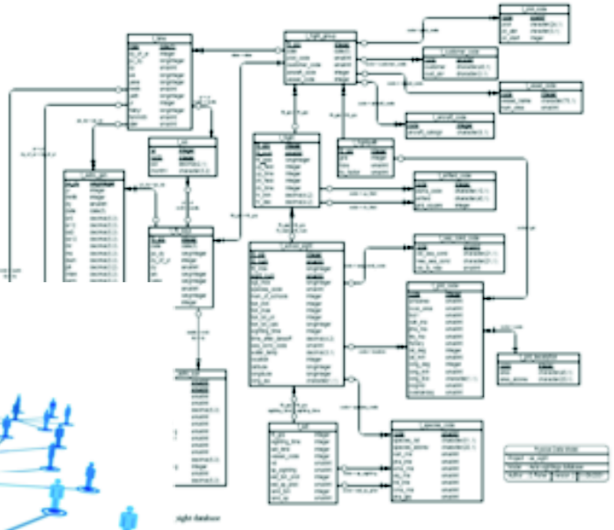
berth f
he crie
the bar
and he
here a
plain n
eggs is
up the
What y
mough
see wh
he thre

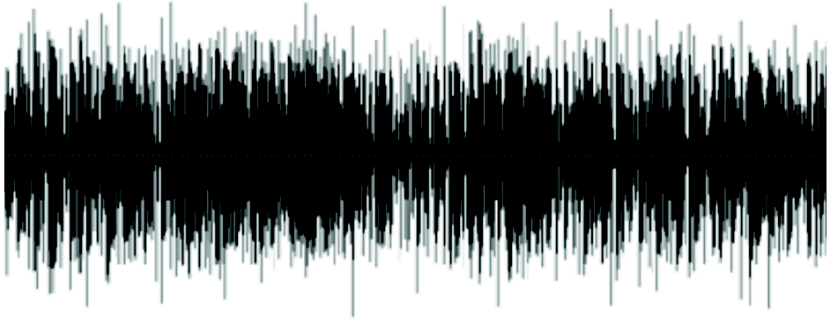
Tweets

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.

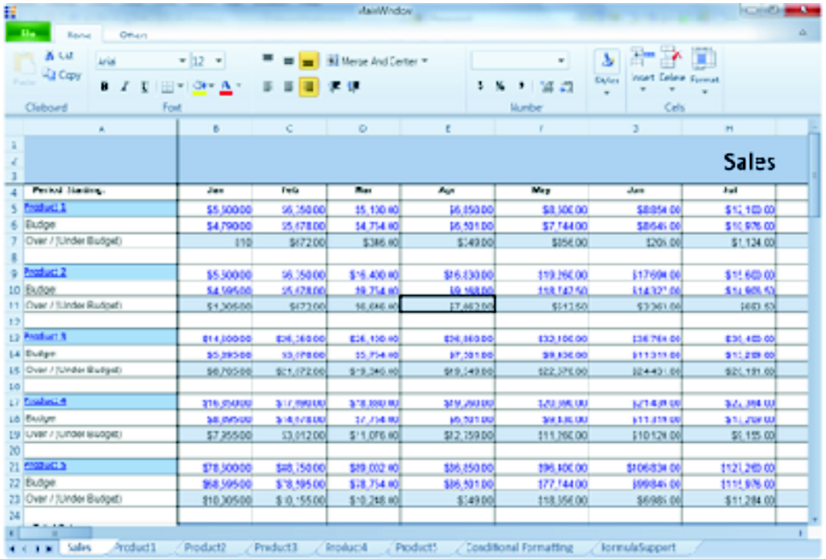
Twitter API @twitterapi 11 Jun
The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/faq#17750

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.





Categorizing data: Unimodal/multimodal



	A	B	C	D	E	F	G	H
1								
2								Sales
3								
4	Period Ending	Jan	Feb	Mar	Apr	May	Jun	Jul
5	Product 1	$5,500.00	$6,250.00	$5,130.00	$6,850.00	$8,500.00	$8,854.00	$12,100.00
6	Budge	$4,790.00	$5,478.00	$4,754.00	$6,501.00	$7,744.00	$8,646.00	$10,976.00
7	Over / (Under Budget)	$10	$772.00	$386.00	$349.00	$856.00	$208.00	$1,124.00
8								
9	Product 2	$5,500.00	$6,250.00	$16,400.00	$16,030.00	$19,266.00	$17,694.00	$15,600.00
10	Budge	$4,595.00	$5,478.00	$9,754.00	$9,168.00	$18,742.50	$14,327.00	$14,905.50
11	Over / (Under Budget)	$1,305.00	$772.00	$6,646.00	$7,462.00	$513.50	$3,361.00	$693.50
12								
13	Product 3	$14,300.00	$26,350.00	$26,100.00	$26,850.00	$32,100.00	$38,764.00	$30,400.00
14	Budge	$5,395.00	$3,678.00	$5,754.00	$7,301.00	$9,836.00	$11,311.00	$11,209.00
15	Over / (Under Budget)	$8,705.00	$21,672.00	$19,346.00	$19,549.00	$22,376.00	$24,431.00	$20,191.00
16								
17	Product 4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	Budge	$8,695.00	$14,678.00	$7,754.00	$6,501.00	[illegible]	$11,319.00	$11,209.00
19	Over / (Under Budget)	$7,355.00	$3,012.00	$11,076.00	$12,759.00	$11,266.00	$10,124.00	$9,155.00
20								
21	Product 5	$78,300.00	$88,750.00	$89,002.00	$86,850.00	$96,400.00	$106,834.00	$121,260.00
22	Budge	$68,595.00	$78,595.00	$78,754.00	$86,501.00	$77,744.00	$99,846.00	$111,976.00
23	Over / (Under Budget)	$10,305.00	$10,155.00	$10,248.00	$349.00	$18,356.00	$6,986.00	$11,284.00
24								

g back
id,
e is still
e up
17—
en my
ibow
ian
up his

ere
ding

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 f
he crie
the bar
and he
here a
plain n
eggs is
up then
What y
mough
see wh
he thre

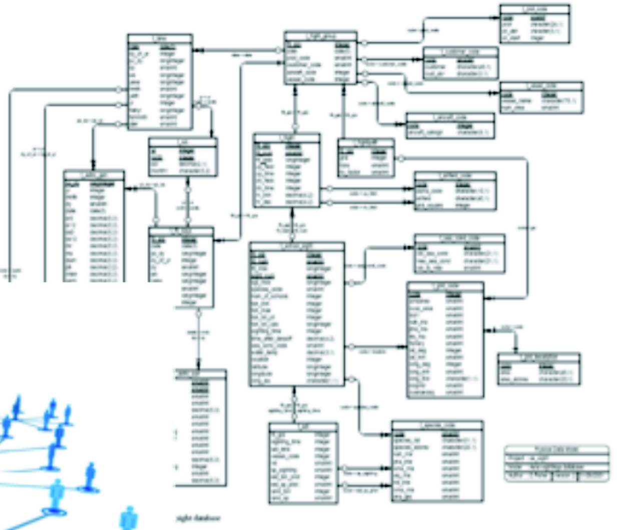
Tweets

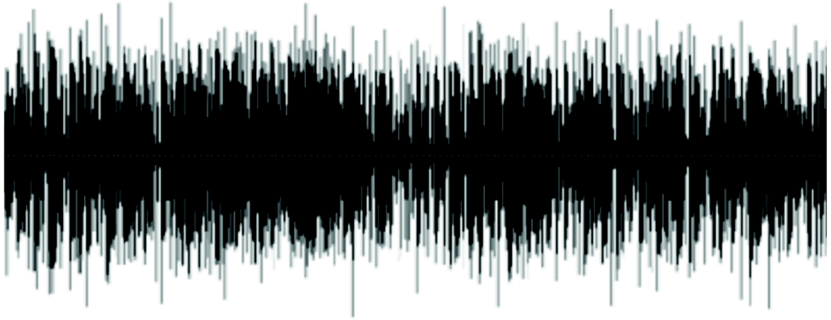
Twitter API @twitterapi 11 Jun
As part of the retirement plan today, we're also about to discontinue Basic Auth support for unelevated Streaning API roles. Use OAuth 1.0A.

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/faq#17750

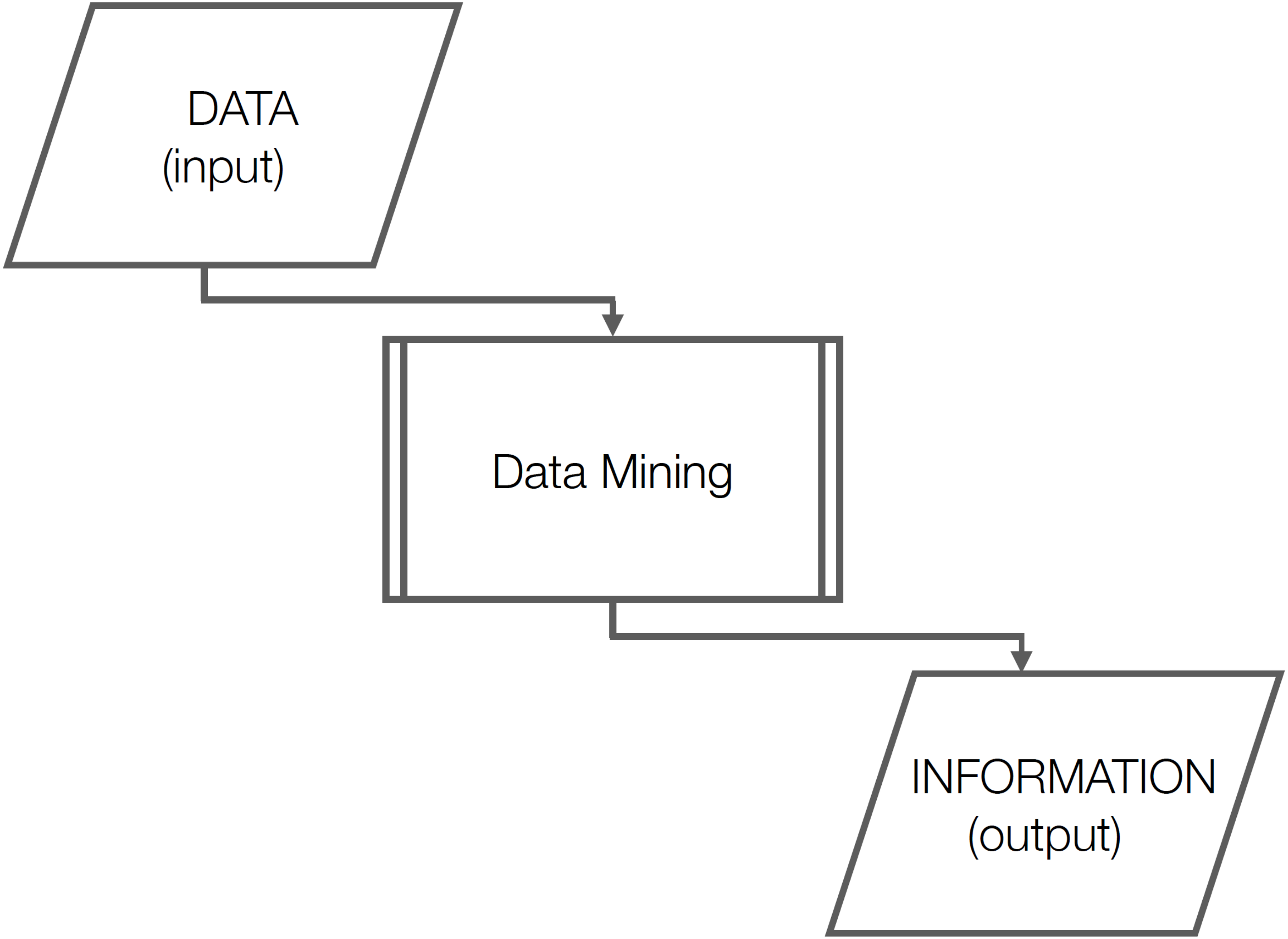
Twitter API @twitterapi 11 Jun
If you were using a legacy widget that no longer functions after API v1 retirement, we suggest using dev.twitter.com/docs/embedded-... instead.

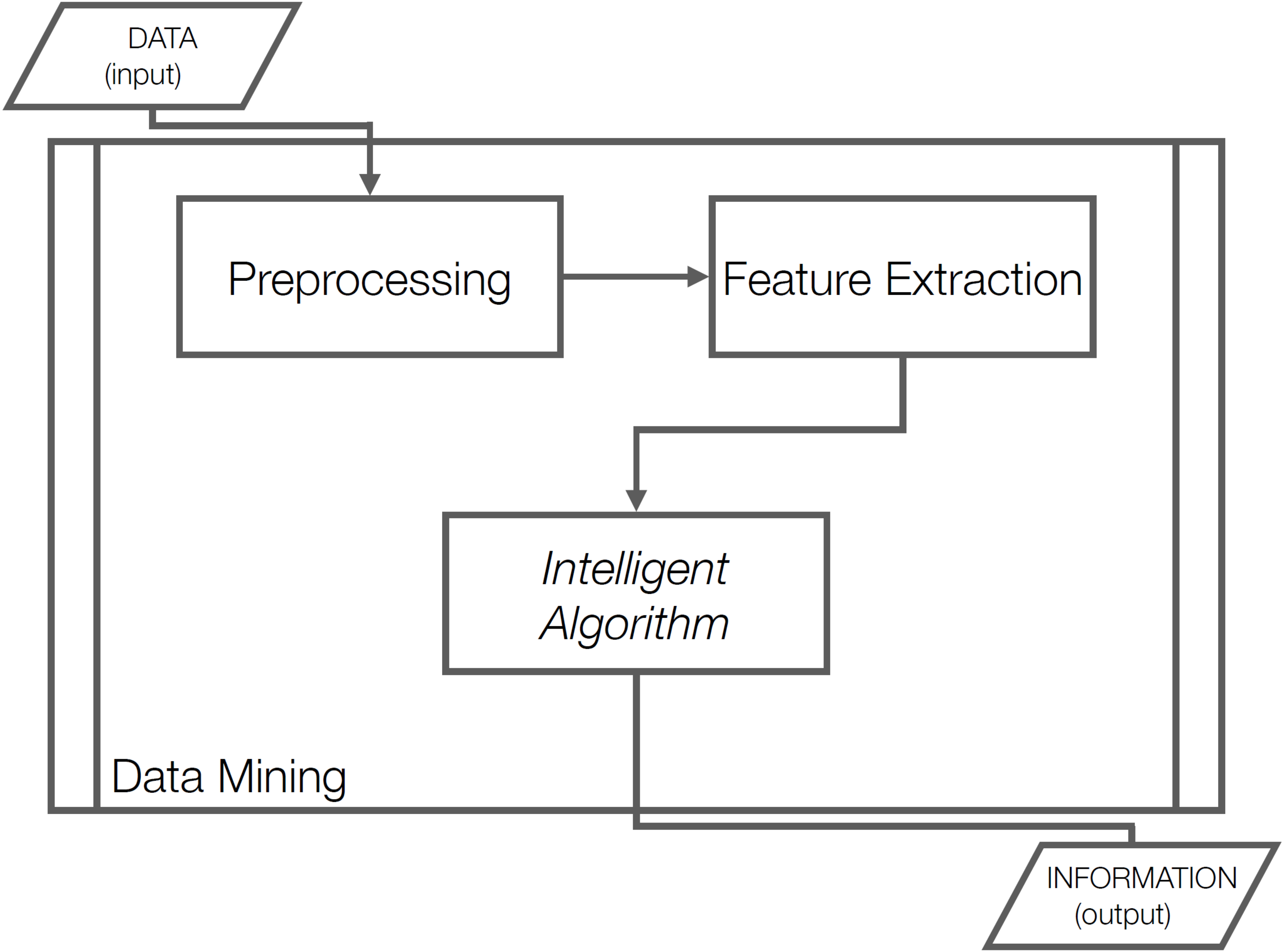
Tweet to @twitterapi





Typical Data Mining Pipeline




DATA
(input)
Preprocessing
Feature Extraction
Intelligent Algorithm
Data Mining
INFORMATION
(output)

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 6 (**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/COMP6237-2022-classlist

Key Dates

- Each team needs to submit a 1-page project brief by the end of the day of week 3 (18th of Feb).
- After 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 13.