COMP6237 Data Mining

Introduction to Data Mining

Jonathon Hare
jsh2@ecs.soton.ac.uk

Markus Brede
mb8@ecs.soton.ac.uk

Teaching Staff

- **Jonathon Hare**
  - jsh2@ecs.soton.ac.uk
  - 1/2003

- **Markus Brede**
  - mb8@ecs.soton.ac.uk
  - 32/4033

Module Overview

- Fairly new module - this is the second time it has run!
  - See feedback from last year
- Created to fill a gap:
  - Data mining is almost synonymous with applied machine learning
    - Inevitably some overlap in topics with COMP3206/COMP6208
      - Should be complementary & offer different views
    - Much more applied/pragmatic focus
      - How do you work with real world data?
      - How do you solve real problems?

Module Structure

- Around 24 lectures + additional tutorials
  - Wide range of data mining topics

- Assessment:
  - 50% 2 Hour Final Exam
  - 20% Individual Coursework
  - 30% Group Coursework

Coursework Timetable

- Group Coursework
  - Set today; report submission on the 10th May; presentations following that.
  - More info at the end of the lecture!
- Individual coursework
  - Set 24th Feb (week 4); due 22th March (just before Easter break)

Resources

- Course web site (handouts, slides [inc 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.

What is “Data Mining”?

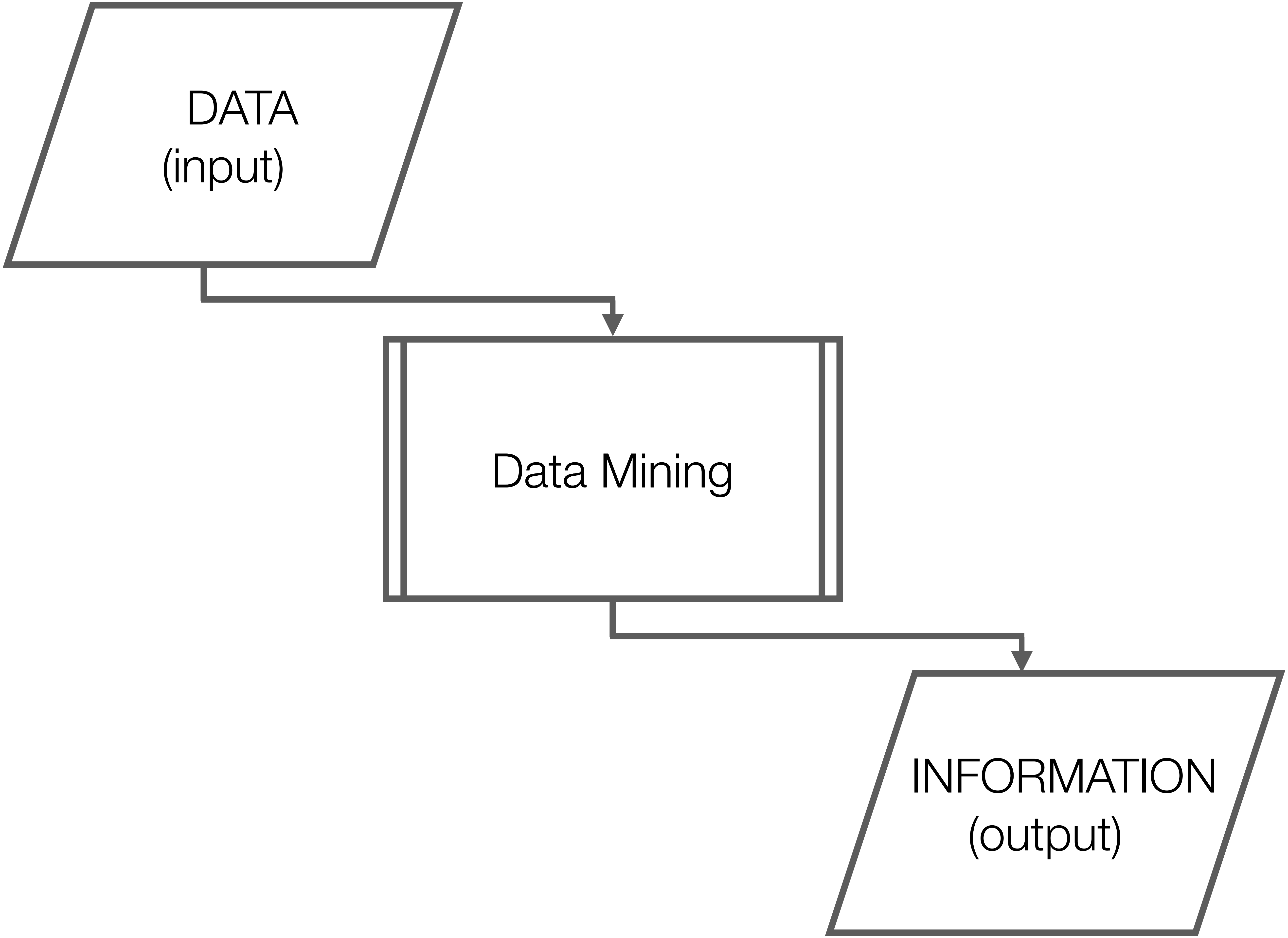
“Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

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

“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?

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?

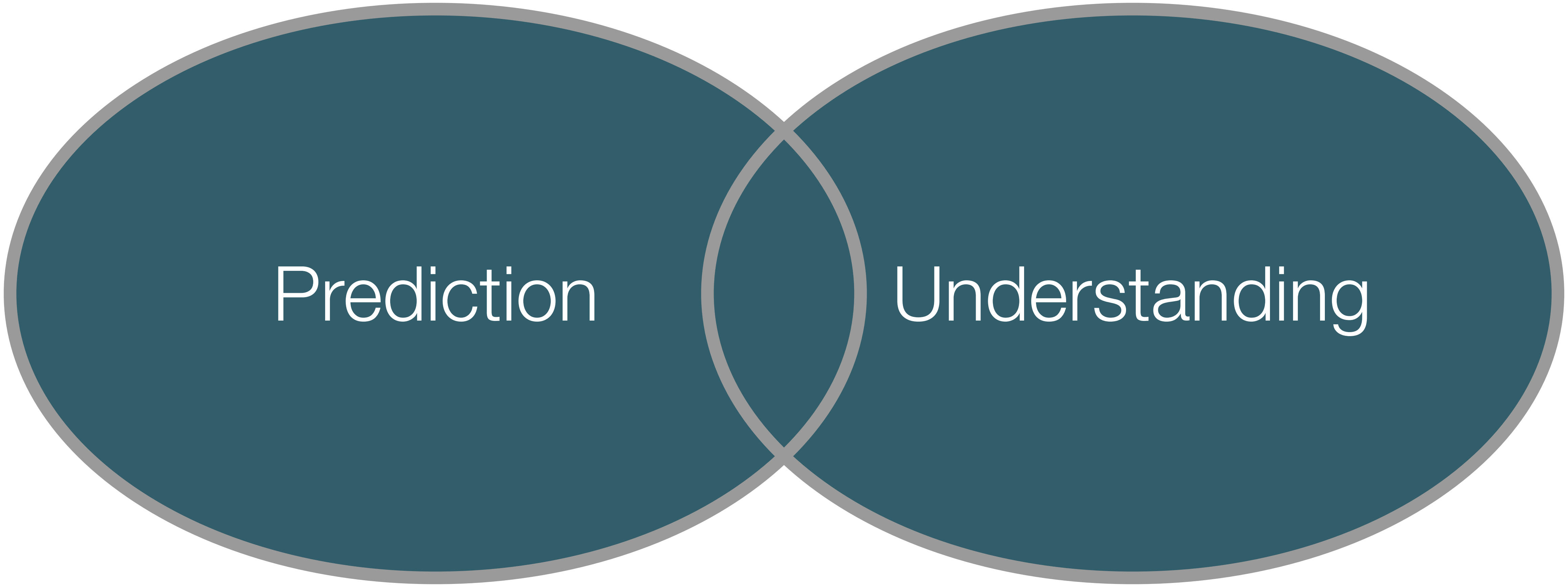
What is information?

- “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

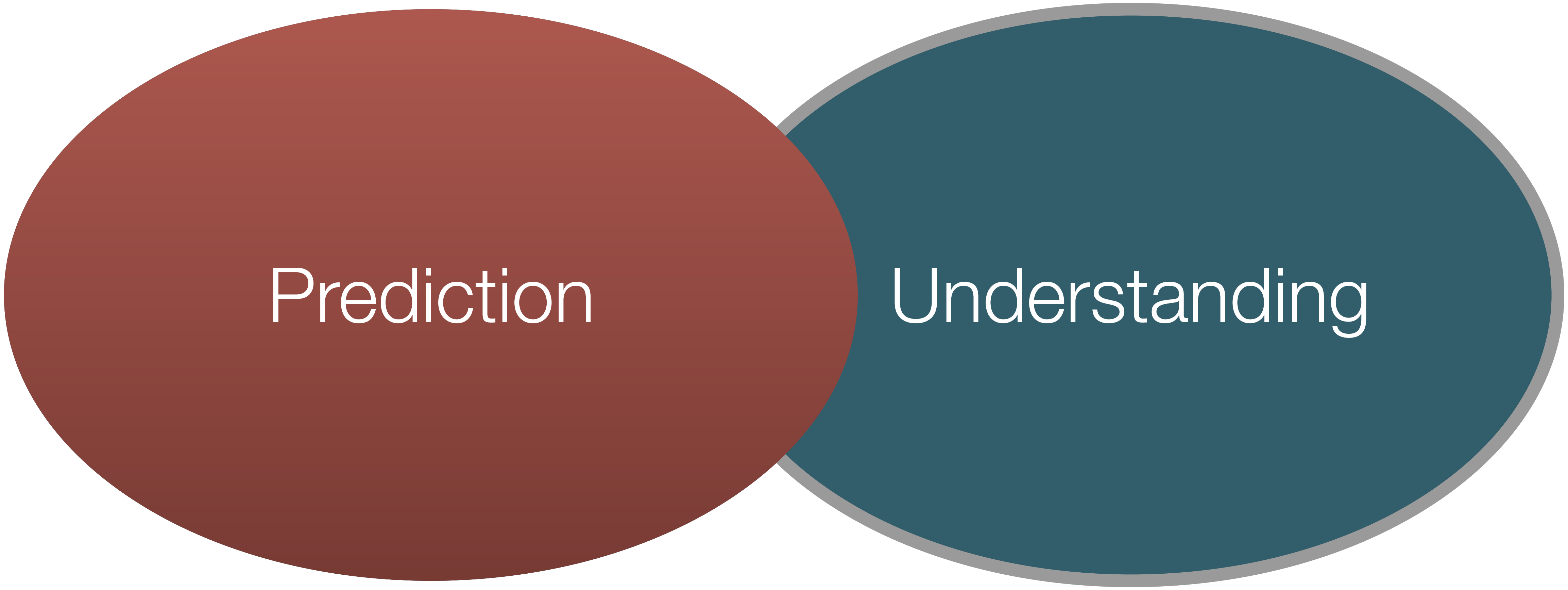
Two complementary goals of data mining



Prediction

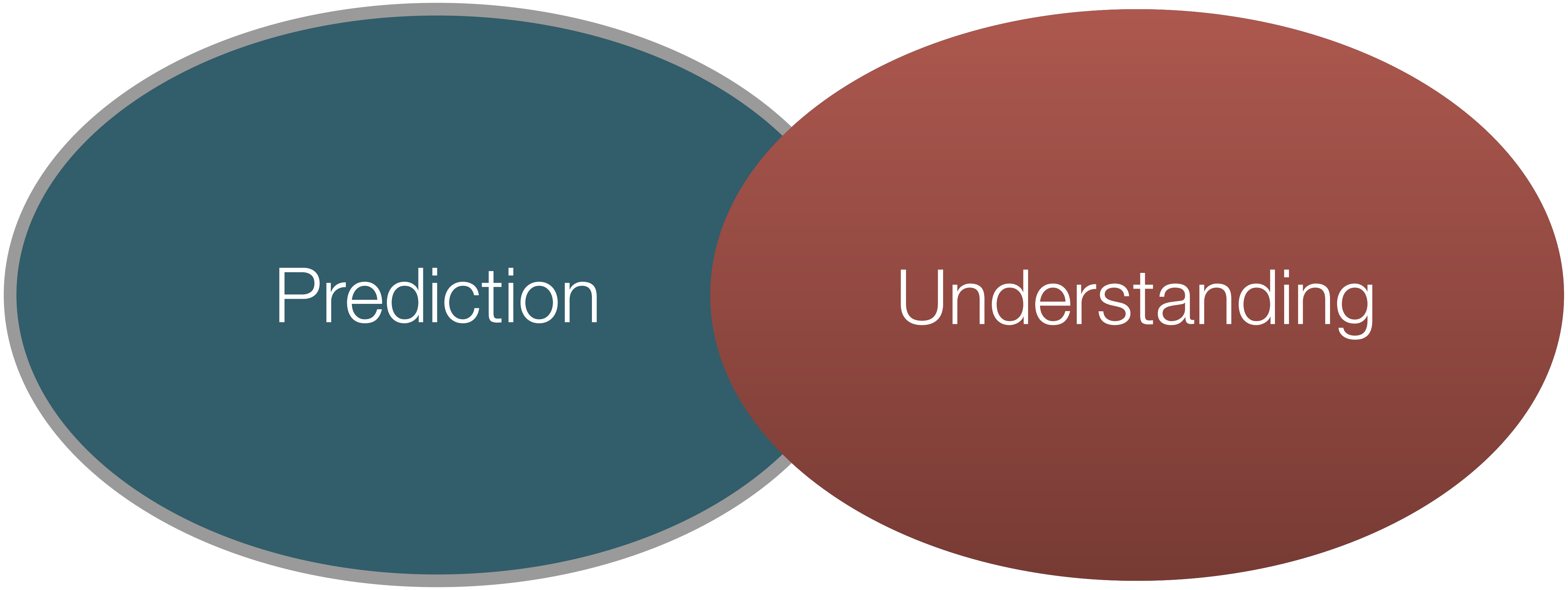
Understanding

Two complementary goals of data mining



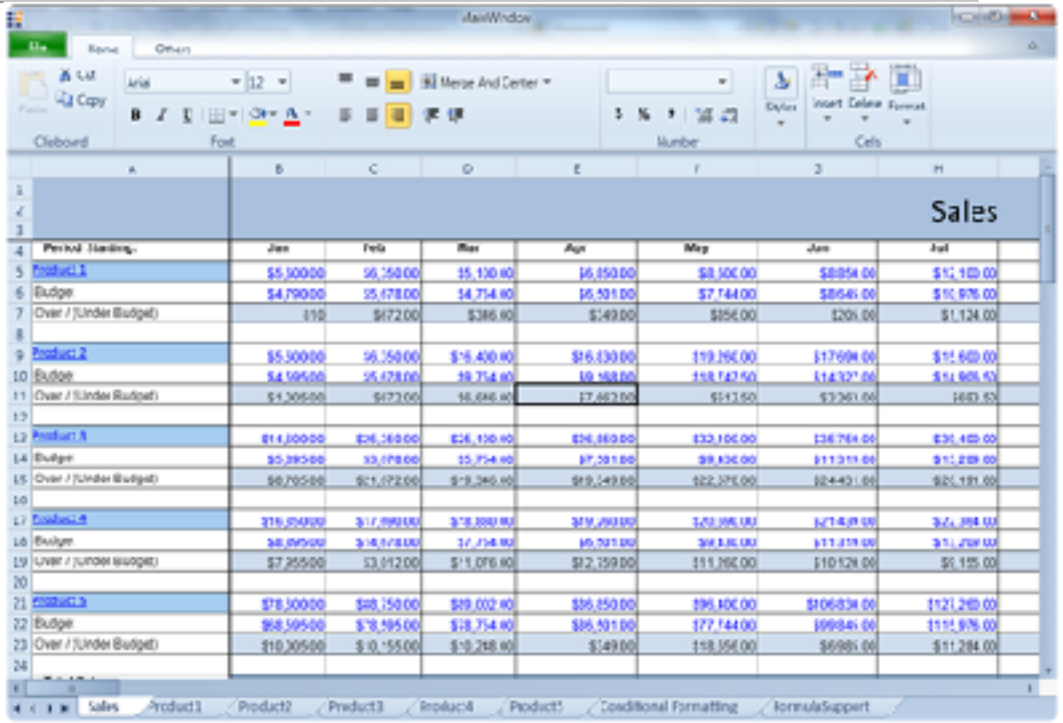
Use some variables to predict unknown
or future values of other variables

Two complementary goals of data mining



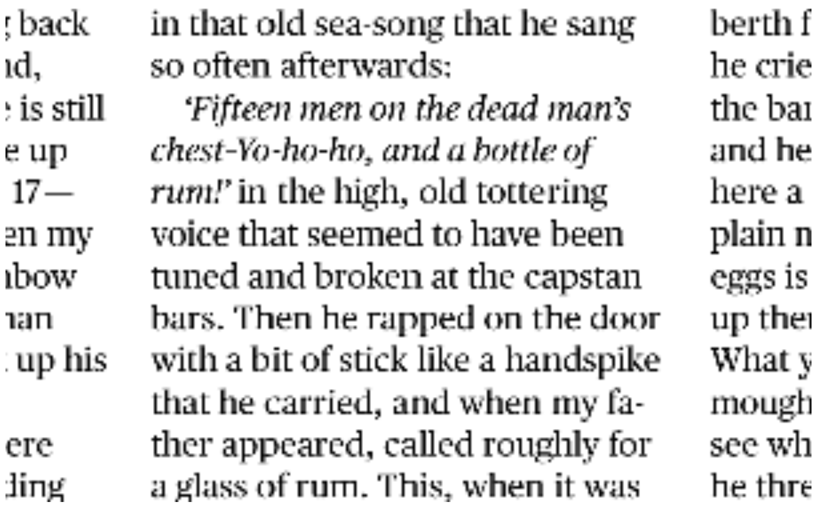
Find human-interpretable patterns that describe the data

What kinds of data are we interested in mining?



	A	B	C	D	E	F	G	H
1								
2								Sales
3								
4	Period Heading	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	$16,830.00	$19,260.00	$17694.00	$11,600.00
10	Budge	$4,595.00	$5,478.00	$9,754.00	$9,168.00	$18,747.50	$14327.00	$11,905.50
11	Over / (Under Budget)	$1,305.00	$672.00	$6,646.00	$7,662.00	$513.50	$3361.00	[illegible]
12								
13	Product 3	$14,000.00	$26,350.00	$26,150.00	$26,850.00	$32,100.00	$36764.00	$30,400.00
14	Budge	$5,395.00	$5,078.00	$5,754.00	$7,501.00	$9,836.00	$11311.00	$11,209.00
15	Over / (Under Budget)	$8,705.00	$21,172.00	$19,340.00	$19,349.00	$22,276.00	$24431.00	$20,191.00
16								
17	Product 4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	Budge	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19	Over / (Under Budget)	$7,355.00	$3,012.00	$11,076.00	$12,159.00	$11,260.00	$10124.00	$6,155.00
20								
21	Product 5	$78,500.00	$88,250.00	$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								

Sales
Product1
Product2
Product3
Product4
Product5
Conditional Formatting
FormulaSupport



back
nd,
is still
e up
17—
en my
ıbow
ıan
up his

ere
ding

in that old sea-song that he sang so often afterwards:

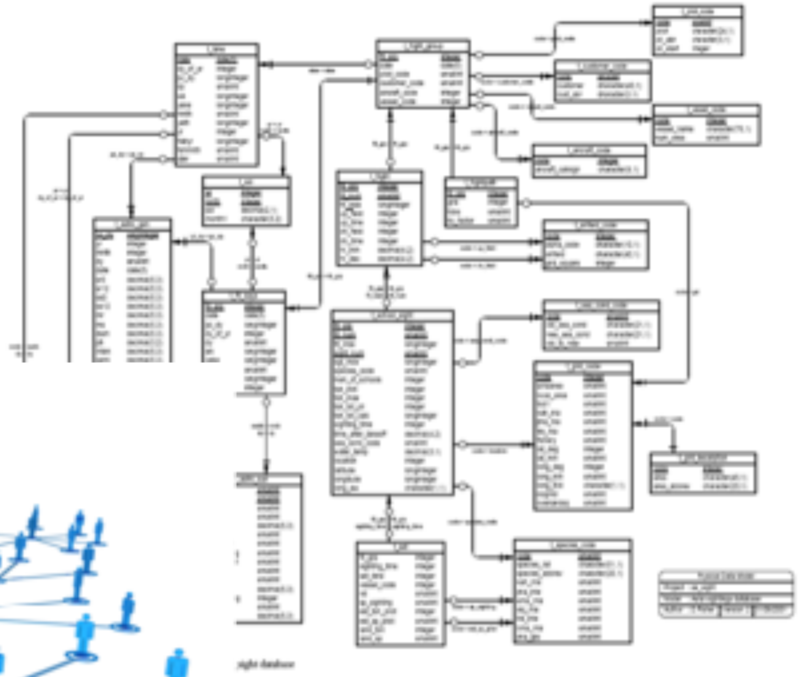
'Fifteen men on the dead man's chest-Yo-ho-ho, and a bottle of rum!' in the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was

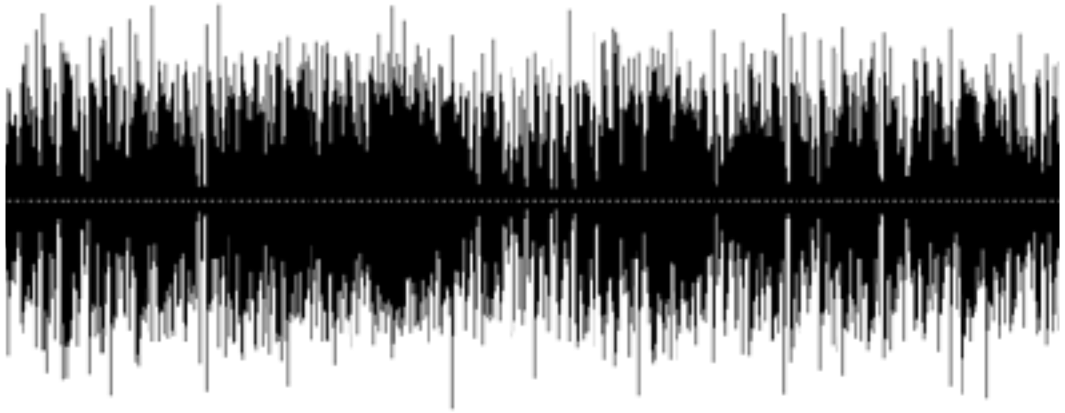
berth f
he crie
the bar
and he
here a
plain n
eggs is
up the
What y
mough
see wh
he thre

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.

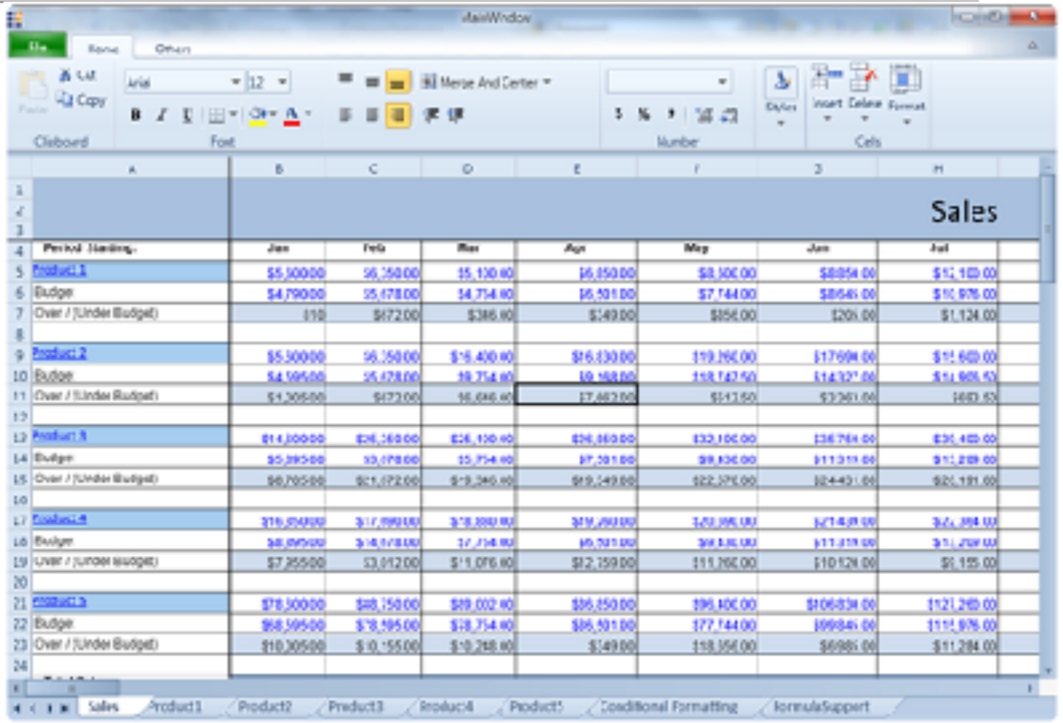
Twitter API @twitterapi 11 Jun
The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/faq#17750

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.



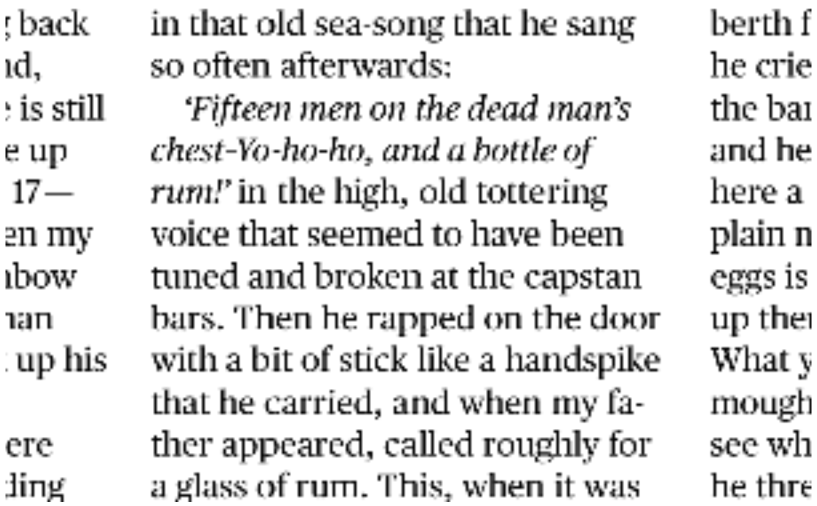


Categorising data: Structured/unstructured



	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	$5,130.40	$6,850.00	$8,500.00	$8854.00	$12,100.00
6	Budge	$4,790.00	$5,478.00	$4,754.40	$6,501.00	$7,744.00	$8646.00	$10,976.00
7	Over / (Under Budget)	$10	$672.00	$386.40	$349.00	$856.00	$208.00	$1,124.00
8								
9	Product 2	$5,500.00	$6,250.00	$16,400.40	$16,830.00	$19,260.00	$17698.00	$15,600.00
10	Budge	$4,595.00	$5,478.00	$9,754.40	$9,168.00	$18,747.50	$14327.00	$14,905.50
11	Over / (Under Budget)	$1,305.00	$672.00	$6,646.40	$7,662.00	$513.50	$3361.00	$693.50
12								
13	Product 3	$14,000.00	$26,360.00	$26,130.40	$26,860.00	$32,100.00	$36,764.00	$30,400.00
14	Budge	$5,295.00	$5,078.00	$5,754.40	$7,501.00	$9,836.00	$11311.00	$12,209.00
15	Over / (Under Budget)	$8,705.00	$21,172.00	$19,340.40	$19,349.00	$22,370.00	$24431.00	$20,191.00
16								
17	Product 4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	Budge	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19	Over / (Under Budget)	$7,355.00	$3,012.00	$11,076.40	$12,159.00	$11,260.00	$10124.00	$9,155.00
20								
21	Product 5	$78,500.00	$88,750.00	$89,002.40	$86,850.00	$96,400.00	$106834.00	$121,260.00
22	Budge	$68,595.00	$78,595.00	$78,754.40	$86,501.00	$77,744.00	$99846.00	$111,976.00
23	Over / (Under Budget)	$10,305.00	$10,155.00	$10,248.40	$349.00	$18,356.00	$6986.00	$11,284.00
24								

Sales
Product1
Product2
Product3
Product4
Product5
Conditional Formatting
FormulaSupport



; back
ıd,
: is still
e up
17—
en my
ıbow
ıan
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

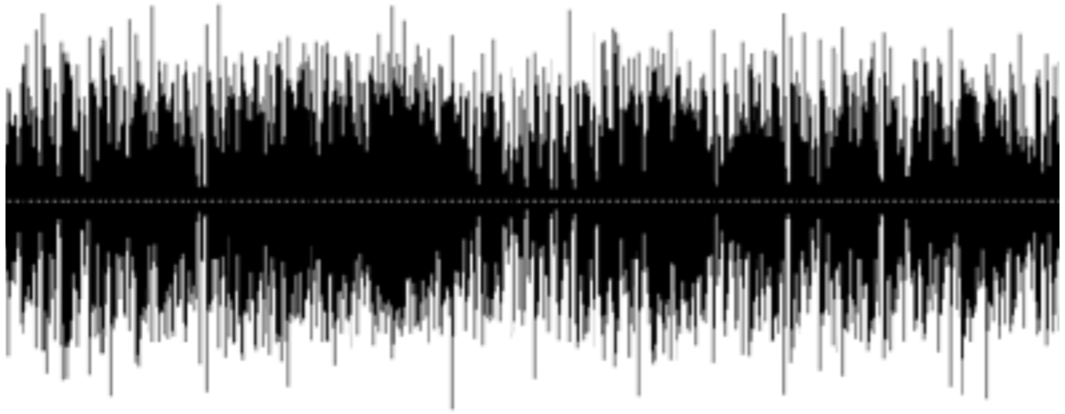
berth f
he crie
the baı
and he
here a
plain n
eggs is
up thei
What y
mough
see wh
he thre

Tweets

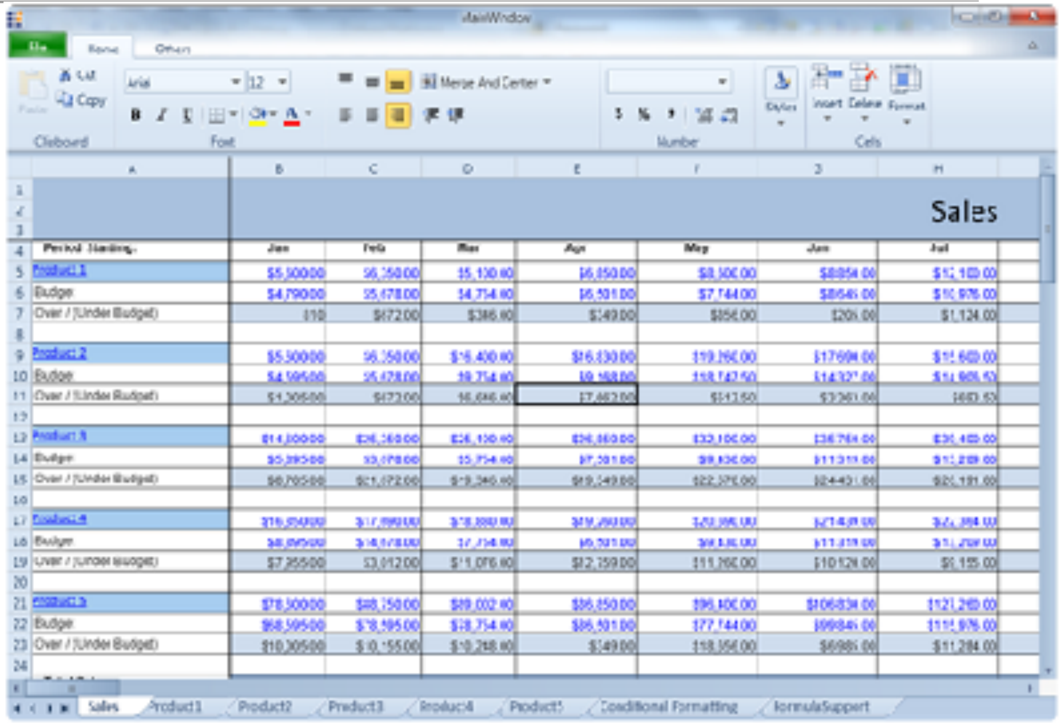
Twitter API @twitterapi 11 Jun
As part of the retirement plan today, we're also about to discontinue Basic Auth support for unelevated Streaming API roles. Use OAuth 1.0A.

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.

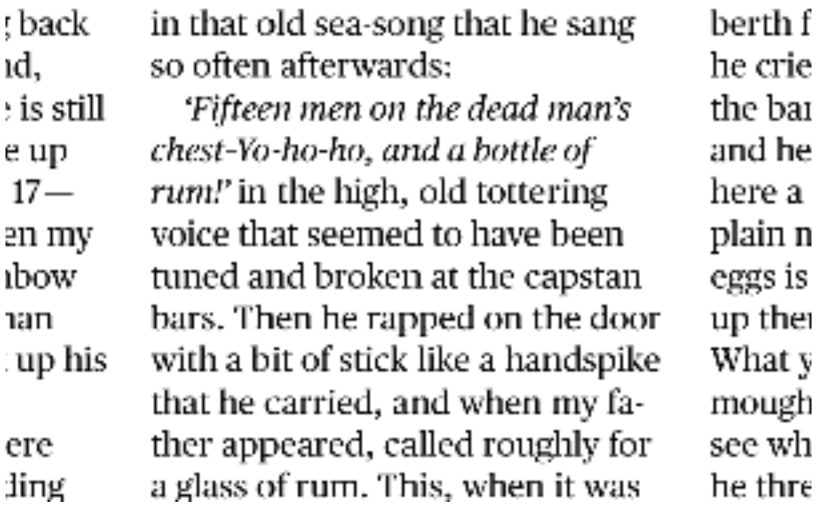


Categorising data: Dynamic/static



	A	B	C	D	E	F	G	H
1								
2								Sales
3								
4	Period Heading	Jan	Feb	Mar	Apr	May	Jun	Jul
5	Product 1	$5,500.00	$6,250.00	$5,100.00	$6,850.00	$8,500.00	$8854.00	$12,100.00
6	Budge	$4,790.00	$5,478.00	$4,714.00	$6,501.00	$7,744.00	$8646.00	$10,976.00
7	Over / (Under Budget)	$10	$672.00	$386.00	$349.00	$856.00	$206.00	$1,124.00
8								
9	Product 2	$5,500.00	$6,750.00	$16,400.00	$16,830.00	$19,260.00	$17694.00	$15,600.00
10	Budge	$4,595.00	$5,478.00	$9,714.00	$9,168.00	$18,747.50	$14327.00	$14,905.50
11	Over / (Under Budget)	$1,305.00	$672.00	$6,686.00	$7,662.00	$513.50	$3361.00	$693.50
12								
13	Product 3	$14,300.00	$26,350.00	$26,100.00	$26,850.00	$33,100.00	$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	$21,272.00	$19,346.00	$19,549.00	$22,276.00	$24431.00	$26,191.00
16								
17	Product 4	$16,050.00	$17,890.00	$18,800.00	$19,260.00	$20,395.00	$21428.00	$22,384.00
18	Budge	$8,695.00	$14,878.00	$7,714.00	$6,501.00	$8,838.00	$11319.00	$11,209.00
19	Over / (Under Budget)	$7,355.00	$3,012.00	$11,076.00	$12,759.00	$11,260.00	$10128.00	$9,155.00
20								
21	Product 5	$78,300.00	$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								

Sales
Product1
Product2
Product3
Product4
Product5
Conditional Formatting
FormulaSupport



; 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 m
eggs is
up ther
What y
mough
see wh
he thre

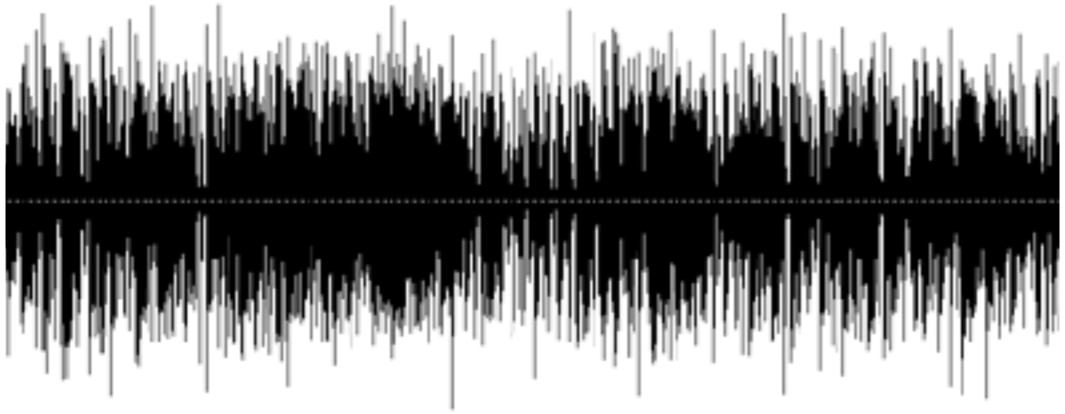
Tweets

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.

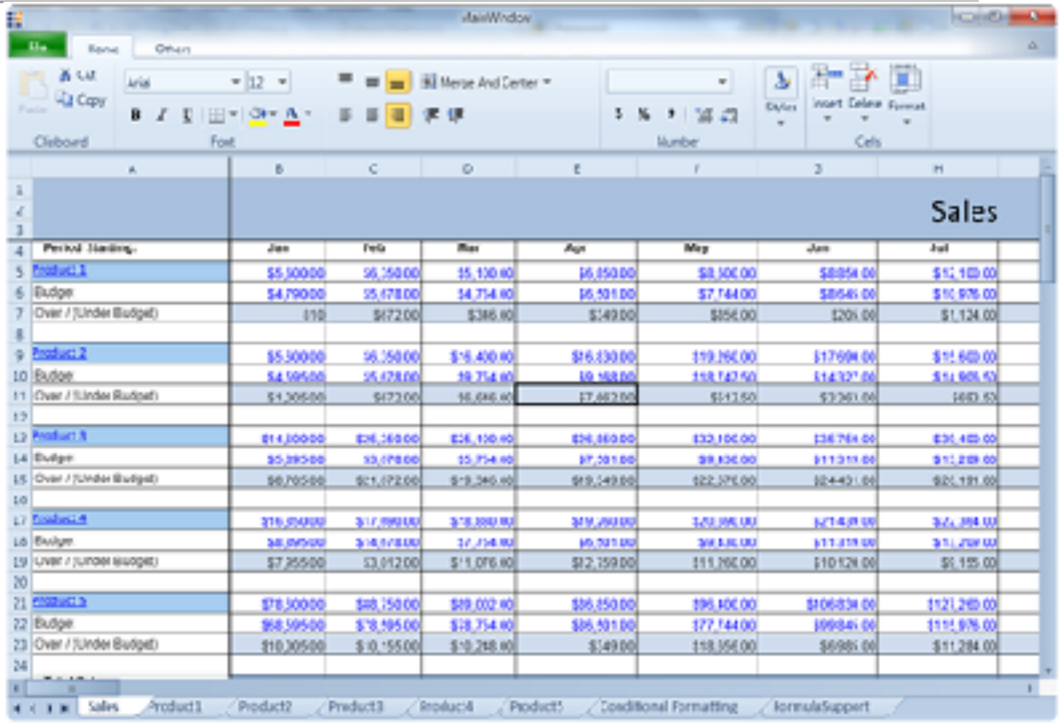
Twitter API @twitterapi 11 Jun
The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/faq#17750

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.

Tweet to @twitterapi

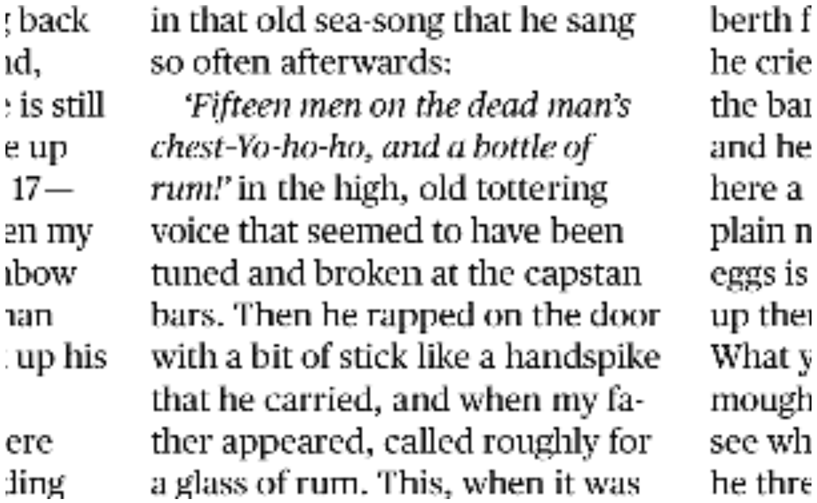


Categorising *data mining*: Unimodal/multimodal



	A	B	C	D	E	F	G	H
1								
2								Sales
3								
4	Period Heading	Jan	Feb	Mar	Apr	May	Jun	Jul
5	Product 1	$5,500.00	$6,250.00	$5,100.00	$6,850.00	$8,500.00	$8854.00	$12,100.00
6	Budge	$4,790.00	$5,478.00	$4,714.00	$6,501.00	$7,744.00	$8646.00	$10,976.00
7	Over / (Under Budget)	$10	$672.00	$386.00	$349.00	$856.00	$206.00	$1,124.00
8								
9	Product 2	$5,500.00	$6,250.00	$16,400.00	$16,830.00	$19,260.00	$17698.00	$15,600.00
10	Budge	$4,595.00	$5,478.00	$9,714.00	$9,168.00	$18,747.50	$14337.00	$14,905.63
11	Over / (Under Budget)	$1,305.00	$672.00	$6,686.00	$7,662.00	$513.50	$3361.00	$693.63
12								
13	Product 3	$14,100.00	$26,150.00	$26,100.00	$26,850.00	$33,100.00	$36764.00	$36,400.00
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	$19,340.00	$19,549.00	$22,376.00	$24451.00	$25,191.00
16								
17	Product 4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	Budge	[illegible]	[illegible]	$7,714.00	$6,501.00	[illegible]	[illegible]	[illegible]
19	Over / (Under Budget)	$7,355.00	$3,012.00	$11,076.00	$12,359.00	$11,266.00	$10124.00	$9,155.00
20								
21	Product 5	$78,300.00	$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								

Sales
Product1
Product2
Product3
Product4
Product5
Conditional Formatting
FormulaSupport



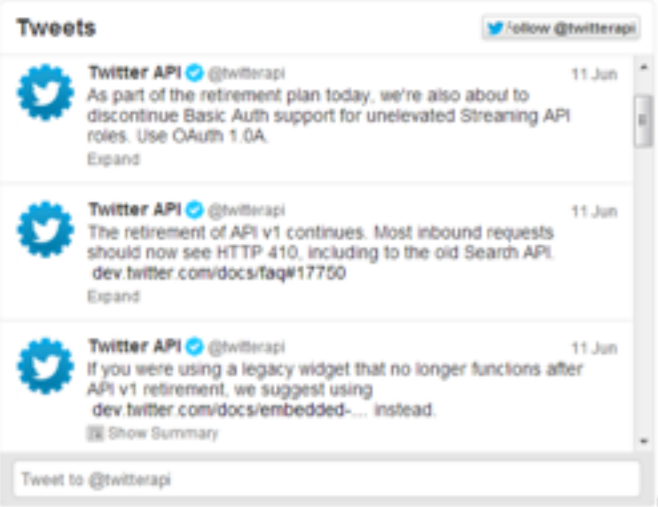
g back
id,
e is still
e up
17—
en my
nbow
nan
up his

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



Tweets

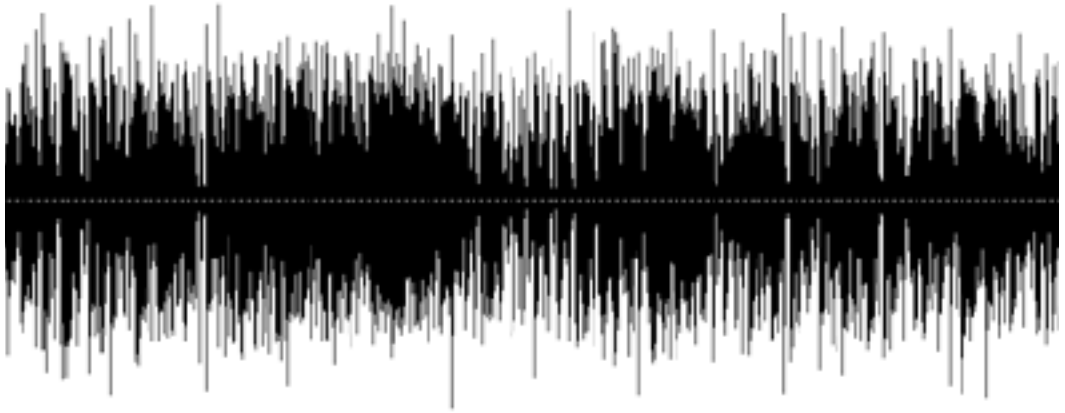
Follow @twitterapi

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

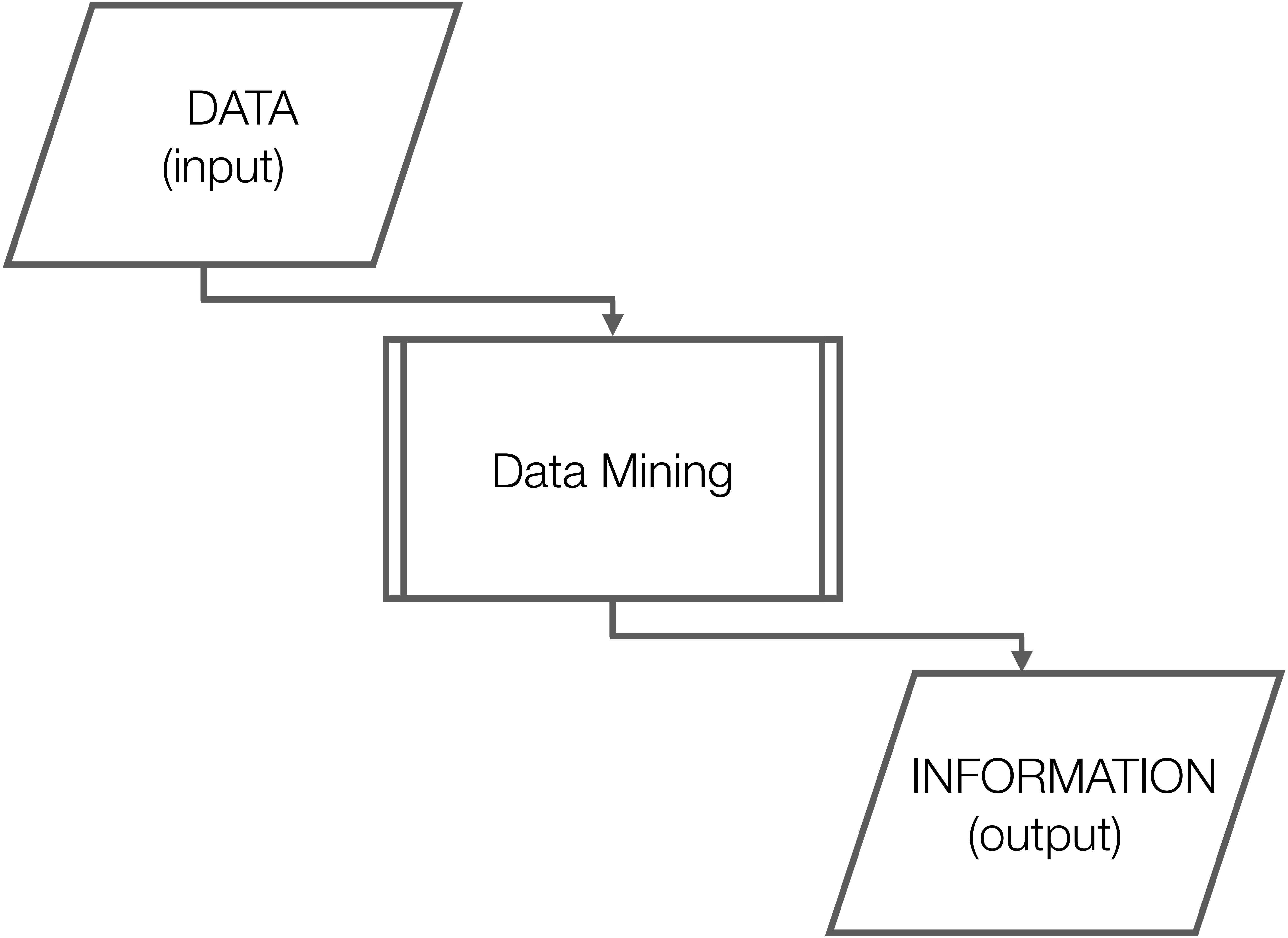
Twitter API @twitterapi 11 Jun
The retirement of API v1 continues. Most inbound requests should now see HTTP 410, including to the old Search API. dev.twitter.com/docs/faq#17750
Expand

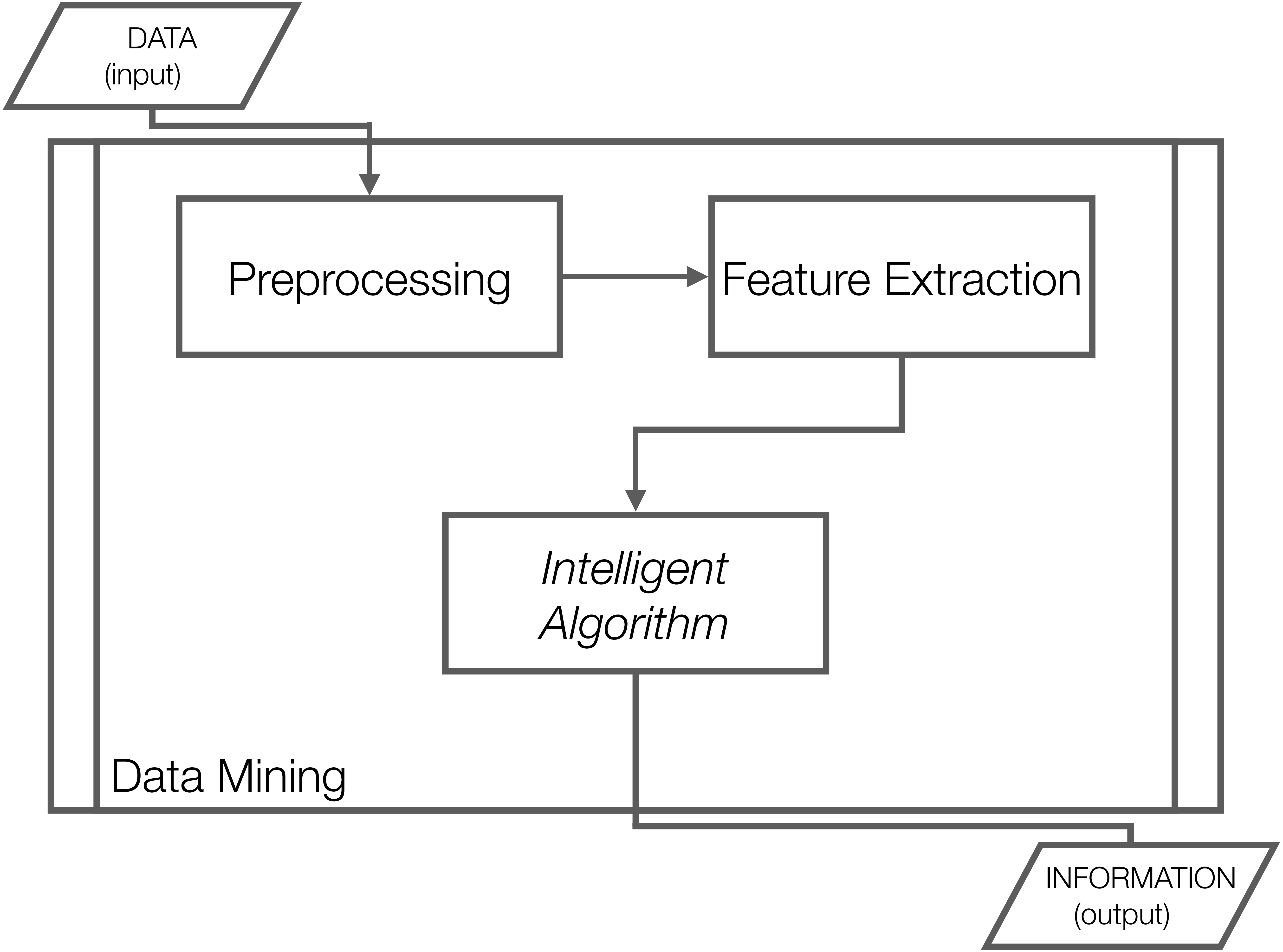
Twitter API @twitterapi 11 Jun
If you were using a legacy widget that no longer functions after API v1 retirement, we suggest using dev.twitter.com/docs/embedded-… instead.
Show Summary

Tweet to @twitterapi



What is the *typical* data mining pipeline?


DATA
(input)
Data Mining
INFORMATION
(output)


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

…we’re going to look at a range of topics…

- **You will learn to solve real-world problems** - e.g.:
  - Recommender systems
  - Market Basket Analysis
  - Document filtering and spam detection
  - Duplicate document detection
- **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)

The Group Coursework

- You need to form groups
    - Target size is 5-6 people
    - As a group you need to chose a **predictive** 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 Friday afternoon slot this week to discuss your ideas for problems to work on with us

- Key dates:
  - Each team must submit a 1-page project brief by the end of the day on the 17th Feb.
  - On the 24th Feb (afternoon) each team must pitch their project to the class (2 minutes to pitch; 3 for Q&A)
  - Teams must submit a conference paper by 4PM on the 10th May
  - Projects will be presented in the lecture/tutorial slots in week 11 & 12
    - Teams should be prepared to present in the first slot; to ensure fairness we’ll pick teams at random.