Clustering & Projects

Today

- Projects
  - A taste of unsupervised learning
    - Hierarchical clustering of documents
  - What is a good project?
  - Project discussion

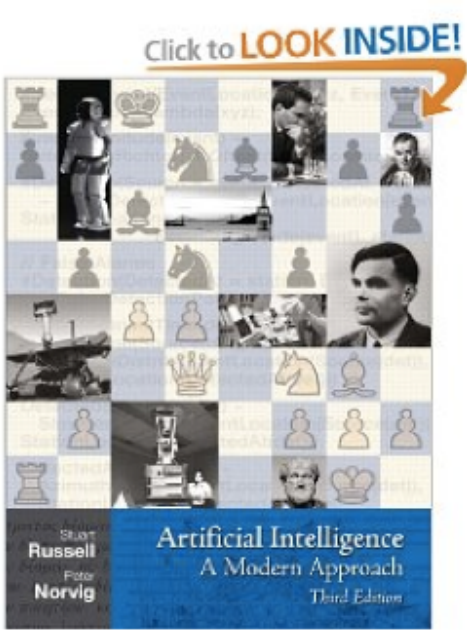
Unsupervised Learning

- Inputs
  - Examples $x_1, x_2, x_3, \ldots, x_N$
  - No labels!

- **Most** common situation in data analysis
  - We are drowning in data:
    - Web documents, gene expression data, medical records, consumer data
  - But few human judgments
    - They are expensive!

- Goal: learn about the data
  - Organization → find subgroups, hierarchies
  - Patterns → if A and B, then C

Product Recommendations



Frequently Bought Together

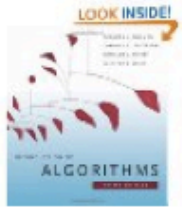
Price For All Three: $270.50

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

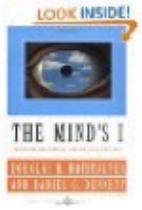
- **This item:** Artificial Intelligence: A Modern Approach (3rd Edition) by Stuart Russell Hardcover
- Introduction to Algorithms by Thomas H. Cormen Hardcover $77.28
- Computer Organization and Design, Revised Fourth Edition, Fourth Edition: The Hardware/S Computer Architecture and Design) by David A. Patterson Paperback $65.72

Customers Who Bought This Item Also Bought



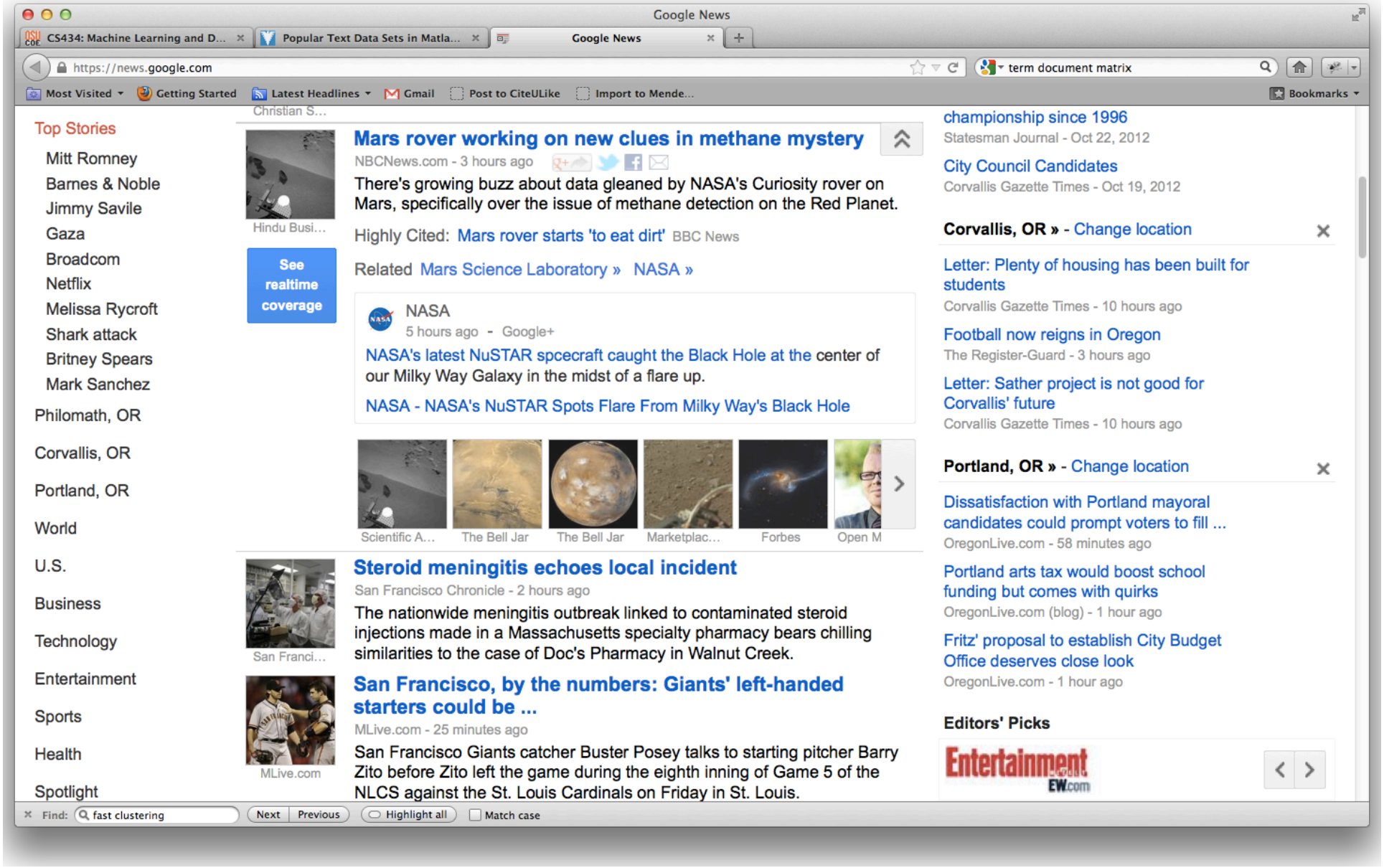
Introduction to Algorithms
› Thomas H. Cormen
(57)
Hardcover
$77.28

Pattern Recognition and Machine Learning …
› Christopher M. Bishop
(65)
Hardcover
$65.59



The Mind's I: Fantasies and Reflections on Self …
› Douglas R. Hofstadter
(34)
Paperback
$13.62

Document Clustering



Clustering

- Examples/instances $x_1, x_2, x_3, \ldots, x_N$
- Partition into groups (clusters) so that
  - Instances in same cluster are similar
  - Instances in different clusters are different
- Similarity / distance
  - "Hard to define, know it when we see it" (Eamon Keogh, UCR)
  - Many mathematical definitions, e.g.

$$\text{dist}(\mathbf{x}, \mathbf{y}) = \| \mathbf{x} - \mathbf{y} \|$$

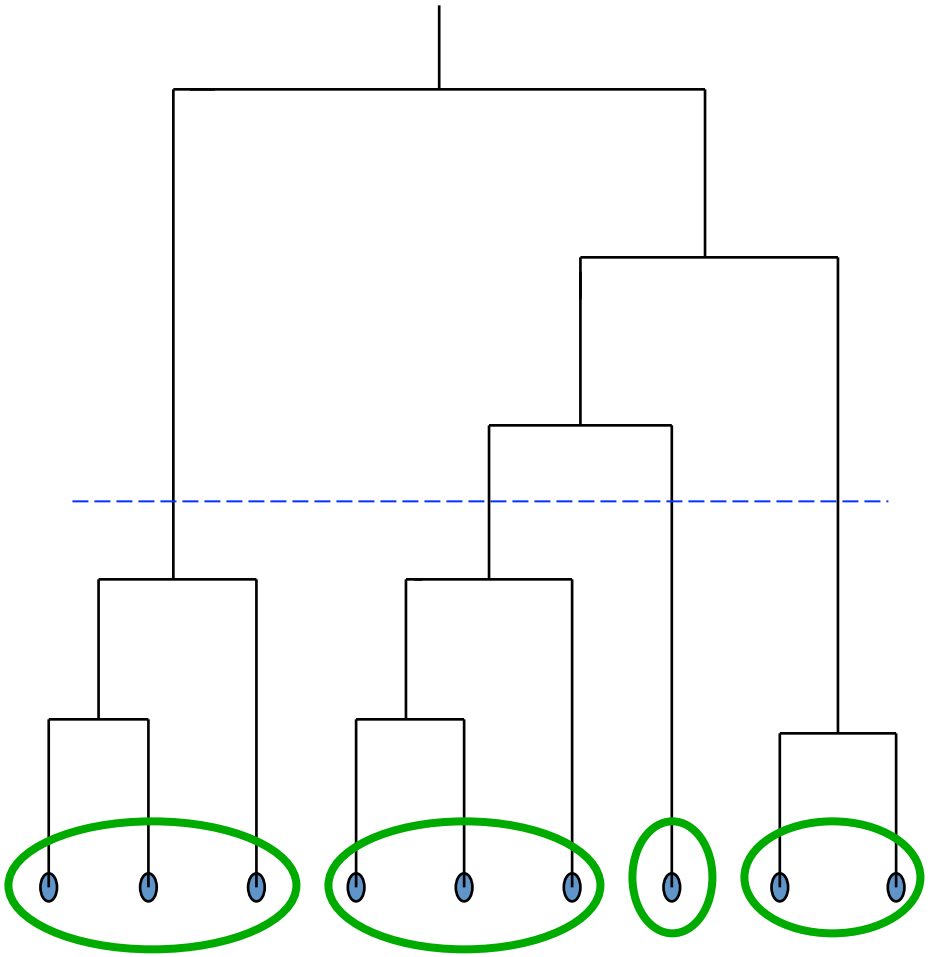
- MATLAB clustering demo

Hierarchical Agglomerative Clustering (HAC)

- Initialize: every instance is a cluster
- Repeat
  - Find *most similar* clusters c_i and c_j
  - Replace them by $c_i \cup c_j$
- Similarity between two clusters?
  - *Single-link*: max. similarity between members
- Demo

Dendrogram

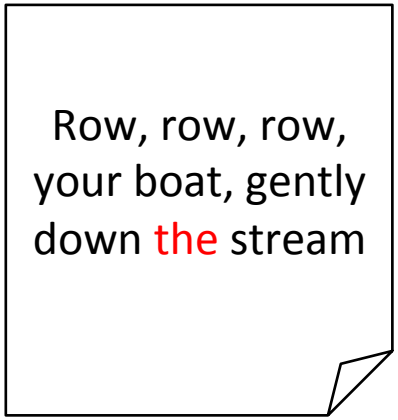
- Upon completion, have a single cluster
- Dendrogram = tree representation of hierarchical clustering
- Cut at any level to get clusters (connected components)



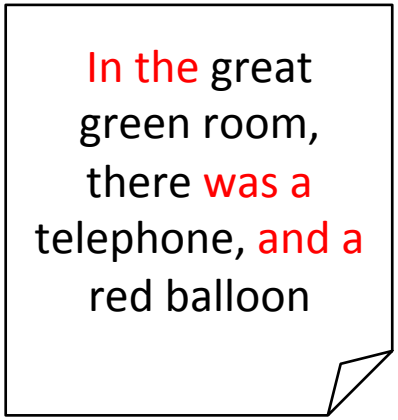
Document Clustering

- How can we cluster documents?
  - **Input**: text documents 1,…,N
  - Need feature representation to compute distance/ similarity

Bag of Words



Document 1



Document 2

	row	your	boat	…	great	green	telephone	red
Doc 1	3	1	1		0	0	0	0
Doc 2	0	0	0		1	1	1	1

Document-Term Matrix

	row	your	boat	…	great	green	telephone	red
x_1	3	1	1		0	0	0	0
x_2	0	0	0		1	1	1	1
x_N	1	0	0		1	0	0	1

- Rows are feature vectors
  - TF vectors ("term-frequency")
- Simple but extremely powerful representation
  - Clustering
  - Classification
  - Dimensionality reduction

Projects

What Is a Project?

- Two main options
  - Apply a method we learned to a data set of your choosing
  - Explore an ML topic we haven't covered
  - (Or both)

Scope and Purpose

- Scope / effort
  - Similar to ~4–5 homework assignments
    - Multiplier: # of people in group
  - Don't forget time to:
    - Define problem
    - Gather data
    - Decide on questions
    - Design experiments
    - Interpret results…
  - It does not need to be complicated, but…

- It should have ***clear purpose*** and be ***executed*** well
  - Have a reason for applying method X to data set Y
  - Execute a project that fulfills that purpose

Example

- Hand-written digit classification using MNIST dataset
- **Purpose**: practice careful empirical ML methodology on previously-defined problem
  (e.g. you are designing a system for your company)
  - Goal: achieve 95% accuracy
  - Start with baseline method
  - Methodically try different things to improve performance
    - More complex hypothesis space
    - Better fitting algorithms
    - More training data
    - Clever feature design
  - Quantify the improvement due to each

Example

- Text classification with 20 newsgroups data
- **Purpose**: learn basic principles of text classification
  - Do background reading
  - Motivate problem: what is text classification used for? Why are you interested in this?
  - Get hands-on experience
    - Data preparation
    - Feature design
    - Which algorithms work well for text?
    - How are text classification methods evaluated?
    - How can you visualize/interpret the results to learn something about the domain problem?
  - Run an experiment that tests a hypothesis

Example

- Error-correcting output codes for multi-class classification
- **Purpose**: vicarious research
  - Summarize prior work (one-vs-one, one-vs-all)
  - Motivate "new" method
  - Explain it
  - Prove that it works
    - Replicate experiments
    - Design your own
  - Discussion
    - Extensions?
    - Critique?
- **Note**: other "methods" explorations would look very similar. Talk to me for more ideas.

Example

- ML on your own data set
  - Hobby, academic interest, predict Mountain Day
- **Purpose**: learn science/art of applying machine learning to new problems
  - **Why is this problem interesting?**
  - What do we care about?
    - Good predictions?
    - Interpretability?
    - Or both?
  - How to formulate problem?
    - Data gathering, labeling, feature design
  - Which algorithms work well?
  - Design a meaningful experiment
    - Outcome is unknown
    - You learn something afterwards

Methods / concepts

- Linear regression
- Logistic regression
- Decision trees
  - classification/regression
- SVMs
- Kernels
  - Linear regression
  - SVMs
- Clustering
  - Agglomerative
  - K-means

- Methodology
  - Feature design
  - Diagnosing and controlling overfitting
  - Regularization
  - Performance measures
  - Cross-validation

Resources

- Data
  - UCI machine learning repository
- Software
  - MATLAB
    - Implementations of common ML algorithms
    - I can probably help find one or guide you
  - Weka
    - Popular ML toolkit in Java
    - Standalone UI
    - Lots of algorithms
    - Warning: I don't know Weka, proceed at your own risk

(I'll post some links)

Logistics

Groups of up to 3 students

Components

- Proposal (10%) — Wed Oct 31
- Mid-project review (25%) — Mon Nov. 19/26
- Final presentation (25%) — Mon. Dec. 10
- Final report (40%) — Tues. Dec 11

2.5/3.5 weeks

3/2 weeks

Proposal

- 2 pages max: describe project at a high level

- Include following information
  - Title
  - **Purpose**
  - Motivation
  - Logistics
    - Data set (preparation?)
    - Code / software
    - Group members (who will work on what)
    - Milestones / timeline
      - What do you plan to complete by mid-project review?

Etc.

- Mid-project review
  - Two page report & meet with me as a group
  - Graded based on milestones you set forth in proposal
  - Evidence of consistent work to execute your plan
    - If something goes wrong, you will not be penalized
- Final presentation
  - ~15–20 minutes
- Final report
  - 8 pages

Discussion

- Split into groups of 2–3, *(not with your project partner)* and discuss project ideas

- Come back and talk as a group