Remaining Data Structures
Strucutres

Structures Hold dissimilar things, consider this to be a data container. Things held are found with ‘fields’. Very nice for packing things together. Like a suitcase!

```matlab
patient.name = 'John Doe';
patient.billing = 127.00;
patient.test = [79 75 73; 180 178 177.5; 220 210 205];

>> patient
patient =
    name: 'John Doe'
    billing: 127
    test: [3x3 double]

>> patient.test(2,2)
ans =
    178
```
The Data Structure Zoo

**Cell Arrays**

**Cells Arrays** Hold dissimilar types of data. Access things being held with numerical arguments, as is done with arrays.

A(1,1) = {{1 4 3; 0 5 8; 7 2 9}};
A(1,2) = {'Anne Smith'};
A(2,1) = {3+7i};
A(2,2) = {-pi:pi/4:pi};
A(3,3) = {5};

```matlab
disp(A{2,1})
ans =
   3.0000 + 7.0000i   A number

disp(A(2,1))
ans =
   [3.0000 + 7.0000i]   A cell containing a number

disp(A{2,2}(3))
ans =
   -1.5708
```
Function Handles

A reference to a function, it can be passed from function to function like any other variable.

```matlab
function y = showFunctHandle(a,b,func)
    y = func(a,b);
end

function [y] = showAdd(a,b)
    y = a + b;
end

function [y] = showMult(a,b)
    y = a * b;
end

>> a = 3;
>> b = 5;
>> showFunctHandle(a,b,@showAdd)
ans =
     8

>> showFunctHandle(a,b,@showMult)
ans =
     15
Case Study: Random Surfer
**Memex.** [Vannevar Bush, 1936] Theoretical hypertext computer system; pioneering concept for world wide web.

- Follow links from book or film to another.
- Tool for establishing links.
World Wide Web

Web Browser

Web browser. Killer application of the 1990s.
Library of Babel

La biblioteca de Babel. [Jorge Luis Borge, 1941]

When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness... There was no personal or world problem whose eloquent solution did not exist in some hexagon.

this inordinate hope was followed by an excessive depression. The certitude that some shelf in some hexagon held precious books and that these precious books were inaccessible seemed almost intolerable.
Web search. Killer application of the 2000s.
Web Search
Web Search

Relevance. Is the document similar to the query term?

Importance. Is the document useful to a variety of users?

Search engine approaches.

- Paid advertisers.
- Manually created classification.
- Feature detection, based on title, text, anchors, ...
- "Popularity."
Disease Dynamics

Relevance. How many contacts does the infected person have?

Importance. What will be the size and nature of the outbreak?

measles  foot-and-mouth  pandemic influenza  small pox

Computational challenge

Science 1 June 2007:
Vol. 316. no. 5829, pp. 1298 - 1301
DOI: 10.1126/science.1134695
Google's PageRank™ algorithm. [Sergey Brin and Larry Page, 1998]

- Measure popularity of pages based on hyperlink structure of Web.
- Revolutionized access to world's information.

Pages and links
90-10 Rule

**Model.** Web surfer chooses next page:
- 90% of the time surfer clicks random hyperlink.
- 10% of the time surfer types a random page.

**Caveat.** Crude, but useful, web surfing model.
- No one chooses links with equal probability.
- No real potential to surf directly to each page on the web.
- The 90-10 breakdown is just a guess.
- It does not take the back button or bookmarks into account.
- We can only afford to work with a small sample of the web.
- ...
Web Graph Input Format

**Input format.**

- N pages numbered 0 through N-1.
- Represent each hyperlink with a pair of integers.
Transition Matrix

**Transition matrix.** \( p(i, j) = \) prob. that surfer moves from page \( i \) to \( j \).

- **Input graph**
- **Link counts**
- **Degrees**
- **Leap probabilities**
- **Link probabilities**
- **Output matrix**

*Surfer on page 1 goes to page 2 next 38% of the time.*
Web Graph to Transition Matrix: Suggestions on Functions

```matlab
function [data]=readfile(filename)
% Import the file

data = % This is for you to do

function [t] = transition(d)

% Form transition matrix.
% Move from zero to one based indexing

% Loop over the data and assign values to transition matrix
% First value skipped because its the number of links
for i=2:2:length(d)

% Apply 90/10 rule
p1 = .9;

% Total links from each page
v = sum(t,2);

% Random link p1% of time
% Random page (1-p1)% of time
```
clear;
% specify the file
file = 'tiny.txt';
% read the file
d = readfile(file);
% convert to transition matrix
t = transition(d);
t
>> pageRank

t =

0.0200  0.9200  0.0200  0.0200  0.0200
0.0200  0.0200  0.3800  0.3800  0.2000
0.0200  0.0200  0.0200  0.9200  0.0200
0.9200  0.0200  0.0200  0.0200  0.0200
0.4700  0.0200  0.4700  0.0200  0.0200
Monte Carlo Simulation
Monte Carlo simulation.

- Surfer starts on page 0.
- Repeatedly choose next page, according to transition matrix.
- Calculate how often surfer visits each page.

How? see next slide

\[
\begin{pmatrix}
0.02 & 0.92 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.38 & 0.38 & 0.20 \\
0.02 & 0.02 & 0.02 & 0.92 & 0.02 \\
0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.47 & 0.02 & 0.47 & 0.02 & 0.02 \\
\end{pmatrix}
\]
Random Surfer

Random move. Surfer is on page \( \text{page} \). How to choose next page \( j \)?

- Row \( \text{page} \) of transition matrix gives probabilities.
- Compute cumulative probabilities for row \( \text{page} \).
- Generate random number \( r \) between 0.0 and 1.0.
- Choose page \( j \) corresponding to interval where \( r \) lies.

```
\begin{pmatrix}
.02 & .92 & .02 & .02 & .02 \\
.02 & .02 & .38 & .38 & .20 \\
.02 & .02 & .02 & .92 & .02 \\
.92 & .02 & .02 & .02 & .02 \\
.47 & .02 & .47 & .02 & .02 \\
\end{pmatrix}
```

Try the Matlab 'cumsum' function!
Random Surfer

Random move. Surfer is on page \textit{page}. How to choose next page \textit{j}?  
\begin{itemize}
  \item Row \textit{page} of transition matrix gives probabilities.
  \item Compute cumulative probabilities for row \textit{page}.
  \item Generate random number \textit{r} between 0.0 and 1.0.
  \item Choose page \textit{j} corresponding to interval where \textit{r} lies.
\end{itemize}

\begin{verbatim}
function [newsite] = randstep(t,site)
    % Given a site (or row) in the transition matrix,
    % randomly select a new site by comparing a random
    % number to the cumulative sum (cumsum) of the row
    % corresponding to the site.
    newsite = find(????,1);
\end{verbatim}

Just find the first index where the logical condition is true.
clear;

file = 'tiny.txt';
% Get data
d = readfile(file);
% Translate data to transition matrix
t = transition(d);

% Do Monte Carlo simulation
trials = 2000;

% The Markov Chain
s = zeros(1,trials+1);
% Starting point
s(1) = 1;

for i = 2:(trials+1)
    s(i) = randstep(t,s(i-1));
end

n = hist(s,size(t,1));hist(s,size(t,1));

fprintf('The normalized page ranks are:
');
disp(n / sum(n));

see previous slide
bin into n and produce histogram
Convergence. For the random surfer model, the fraction of time the surfer spends on each page converges to a unique distribution, independent of the starting page.

"page rank" "stationary distribution" of Markov chain "principal eigenvector" of transition matrix
Mixing a Markov Chain
The Power Method

Q. If the surfer starts on page 0, what is the probability that surfer ends up on page i after one step?

A. First row of transition matrix.

\[
\begin{bmatrix}
\text{rank[\text{]} & p[\text{]}} \\
\text{first move} & \begin{bmatrix}
.02 & .92 & .02 & .02 & .02 \\
.02 & .02 & .38 & .38 & .20 \\
.02 & .02 & .02 & .92 & .02 \\
.92 & .02 & .02 & .02 & .02 \\
.47 & .02 & .47 & .02 & .02
\end{bmatrix}
\text{newRank[\text{}]} = \begin{bmatrix}
.02 & .92 & .02 & .02 & .02
\end{bmatrix}
\]

probabilities of surfing from 0 to i in one move
The Power Method

Q. If the surfer starts on page 0, what is the probability that surfer ends up on page \( i \) after two steps?

A. Matrix-vector multiplication.
The Power Method

Power method. Repeat until page ranks converge.

\[
\text{rank}[] \ \\ \begin{bmatrix}
0.02 & 0.92 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.38 & 0.38 & 0.20 \\
0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.47 & 0.02 & 0.47 & 0.02 & 0.02 \\
\end{bmatrix} \\
\text{p[]}[] \\
\text{new}[] \ \\ \begin{bmatrix}
0.02 & 0.92 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.38 & 0.38 & 0.20 \\
0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.47 & 0.02 & 0.47 & 0.02 & 0.02 \\
\end{bmatrix}
\]

\text{first move} \\
\begin{bmatrix}1.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \times \begin{bmatrix}
0.02 & 0.92 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.38 & 0.38 & 0.20 \\
0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.47 & 0.02 & 0.47 & 0.02 & 0.02 \\
\end{bmatrix} = \begin{bmatrix}
0.02 & 0.92 & 0.02 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.38 & 0.38 & 0.20 \\
0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.47 & 0.02 & 0.47 & 0.02 & 0.02 \\
\end{bmatrix}

\text{probabilities of surfing from 0 to 1 in one move}

\text{second move} \\
\begin{bmatrix}0.02 & 0.92 & 0.02 & 0.02 \end{bmatrix} \times \begin{bmatrix}
0.02 & 0.92 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.38 & 0.38 & 0.20 \\
0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.47 & 0.02 & 0.47 & 0.02 & 0.02 \\
\end{bmatrix} = \begin{bmatrix}
0.05 & 0.04 & 0.36 & 0.37 & 0.19 \\
\end{bmatrix}

\text{probabilities of surfing from 1 to 2 in one move}

\text{third move} \\
\begin{bmatrix}0.05 & 0.04 & 0.36 & 0.37 & 0.19 \end{bmatrix} \times \begin{bmatrix}
0.02 & 0.92 & 0.02 & 0.02 \\
0.02 & 0.02 & 0.38 & 0.38 & 0.20 \\
0.92 & 0.02 & 0.02 & 0.02 & 0.02 \\
0.47 & 0.02 & 0.47 & 0.02 & 0.02 \\
\end{bmatrix} = \begin{bmatrix}
0.44 & 0.06 & 0.12 & 0.36 & 0.03 \\
\end{bmatrix}

\text{probabilities of surfing from 0 to 1 in three moves}
**Mathematical Context**

**Convergence.** For the random surfer model, the power method iterates converge to a unique distribution, independent of the starting page.

"page rank"  
"stationary distribution" of Markov chain  
"principal eigenvector" of transition matrix

20th move

probabilities of surfing from 0 to 1 in 19 moves

\[
\begin{bmatrix}
.27 & .26 & .15 & .25 & .07
\end{bmatrix}
\times
\begin{bmatrix}
.02 & .92 & .02 & .02 & .02 \\
.02 & .02 & .38 & .38 & .20 \\
.02 & .02 & .02 & .92 & .02 \\
.92 & .02 & .02 & .02 & .02 \\
.47 & .02 & .47 & .02 & .02
\end{bmatrix}
= \begin{bmatrix}
.27 & .26 & .15 & .25 & .07
\end{bmatrix}
\]

probabilities of surfing from 0 to 1 in 20 moves (steady state)
Random Surfer: Scientific Challenges

**Google's PageRank™ algorithm.** [Sergey Brin and Larry Page, 1998]
- Rank importance of pages based on hyperlink structure of web, using 90-10 rule.
- Revolutionized access to world's information.

![Diagram of Page ranks]

**Scientific challenges.** Cope with 4 billion-by-4 billion matrix!
- Need **data structures** to enable computation.
- Need **linear algebra** to fully understand computation.
Extra Slides
**Random Surfer and Matrix Multiplication**

**Q.** What is prob that surfer moves from page $i$ to page $j$ in two steps?

**A.** $p^2 = p \times p$. [Matrix multiplication!]

![Diagram of a Markov chain with transition probabilities](image)

$\begin{pmatrix}
0 & 0 & .02 & .02 & .02 \\
.02 & .02 & .38 & .38 & .20 \\
.02 & .02 & .02 & .92 & .02 \\
.92 & .02 & .02 & .02 & .02 \\
.47 & .02 & .47 & .02 & .02 \\
\end{pmatrix}$

$\begin{pmatrix}
.05 & .04 & .36 & .37 & .19 \\
.45 & .04 & .12 & .37 & .02 \\
.86 & .04 & .04 & .05 & .02 \\
.05 & .85 & .04 & .05 & .02 \\
.05 & .44 & .04 & .45 & .02 \\
\end{pmatrix}$

*Squaring a Markov chain*
Random Surfer: Mathematical Context

**Q.** What is prob that surfer moves from page \( i \) to page \( j \) in the limit?

**A.** \( \lim_{k \to \infty} P^k = P \times P \times \ldots \times P. \)

**Mixing theorem.** \( P^k \) converges. Moreover, all rows are equal.

For our random surfer model surfing from 1 to 2 in 8 steps, the fraction of time surfer spends on page \( j \) is independent of starting point!

\[
\begin{array}{cccccccc}
p & .02 & .92 & .02 & .02 & .02 & .02 & .02 \\
p^2 & .05 & .04 & .31 & .31 & .30 & \text{\textcolor{blue}{.31}} & \text{\textcolor{blue}{.41}} & \text{\textcolor{blue}{.05}} & \text{\textcolor{blue}{.19}} & \text{\textcolor{blue}{.04}} \\
p^3 & .20 & .30 & .16 & .19 & .15 & .25 & .32 & .13 & .21 & .10 \\
p^4 & .20 & .30 & .16 & .19 & .15 & .26 & .22 & .17 & .25 & .11 \\
p^5 & .26 & .22 & .17 & .25 & .11 & .30 & .25 & .13 & .24 & .08 \\
p^6 & .30 & .27 & .14 & .22 & .11 & .27 & .26 & .14 & .23 & .10 \\
p^{16} & .27 & .26 & .14 & .23 & .10 & .27 & .26 & .14 & .23 & .10 \\
\end{array}
\]

Surfing from 1 to 2 in 8 steps.