data & information

2011-09-27

- where is the information that is lost in data?

NB: slides for lectures at

www.inf.ed.ac.uk/teaching/courses/il1/slides/

Information in Society

David Hume

The spirit of the age affects all the arts.

- Informatics is the study of systems that store, process and communicate information.
- In the age of information, these systems act, and interact, to affect the ways we live, work and play.
- Information technologies are changing society.
- Informatics is changing the ways we understand the world and ourselves.

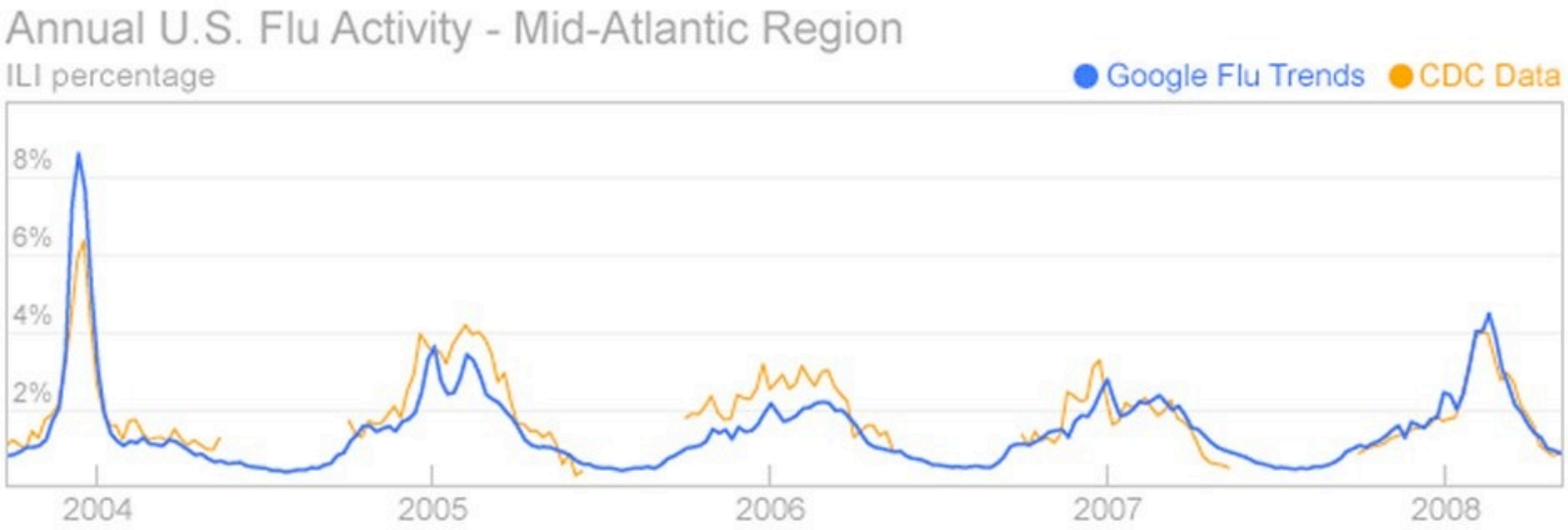
Information is not knowledge
Knowledge is not wisdom
Wisdom is not truth
Truth is not beauty
Beauty is not love
Love is not music
Music is THE BEST...

Frank Zappa - Packard Goose

Saturday, 3 December 2011

Today we'll look at a few examples of data and information

What is information?



Google Flu Trends tracks certain search queries that are directly correlated with the data from traditional flu surveillance systems. These terms are monitored, and used to estimate flu activity around the world.

Data is Google's search logs, which give frequencies of various search terms in various areas.

Information is predicted flu levels in different areas. Traditional methods (reports from clinics) have ~ 2-week lag. Googe Flu Trends is almost real-time.

Look at historical data; find terms that occur frequently when flu occurs – but not otherwise (TFIDF) ; use these for prediction.

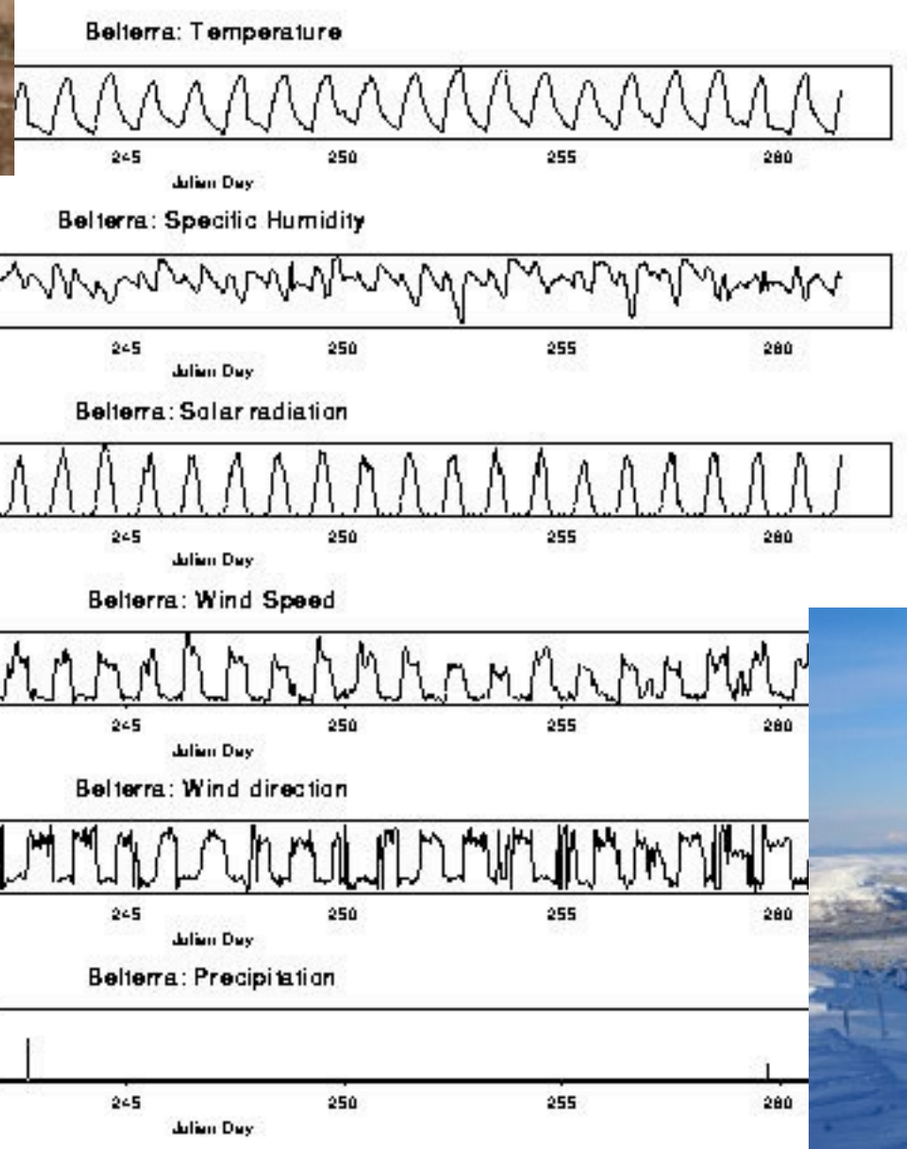
Of course there is lots of noise in the Google search term data – people searching for "swine flu" has nothing to do with whether they have the flu.

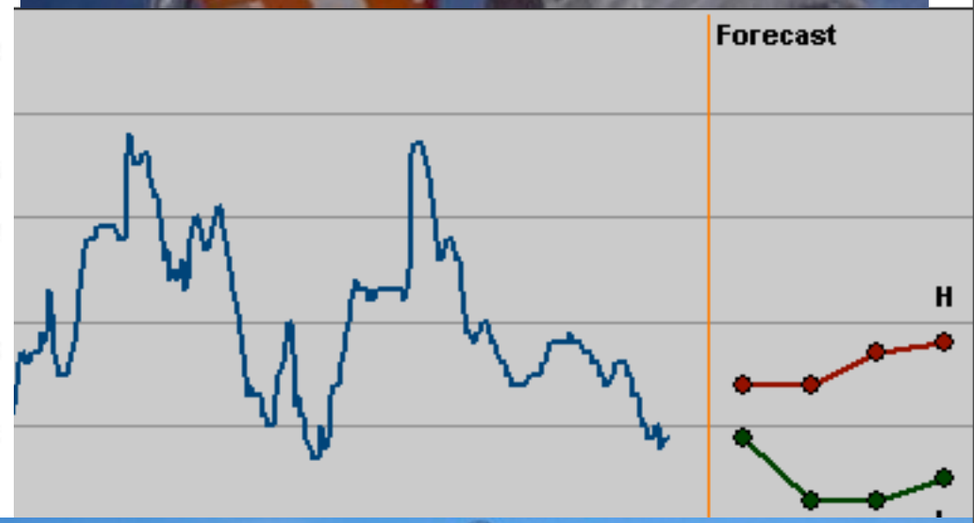
When you have lots of data you can sort out the information from the noise.

Saturday, 3 December 2011

Google FluTrends movie will play when you reach this page
http://www.youtube.com/watch?v=6111nS66Dpk

Weather Data



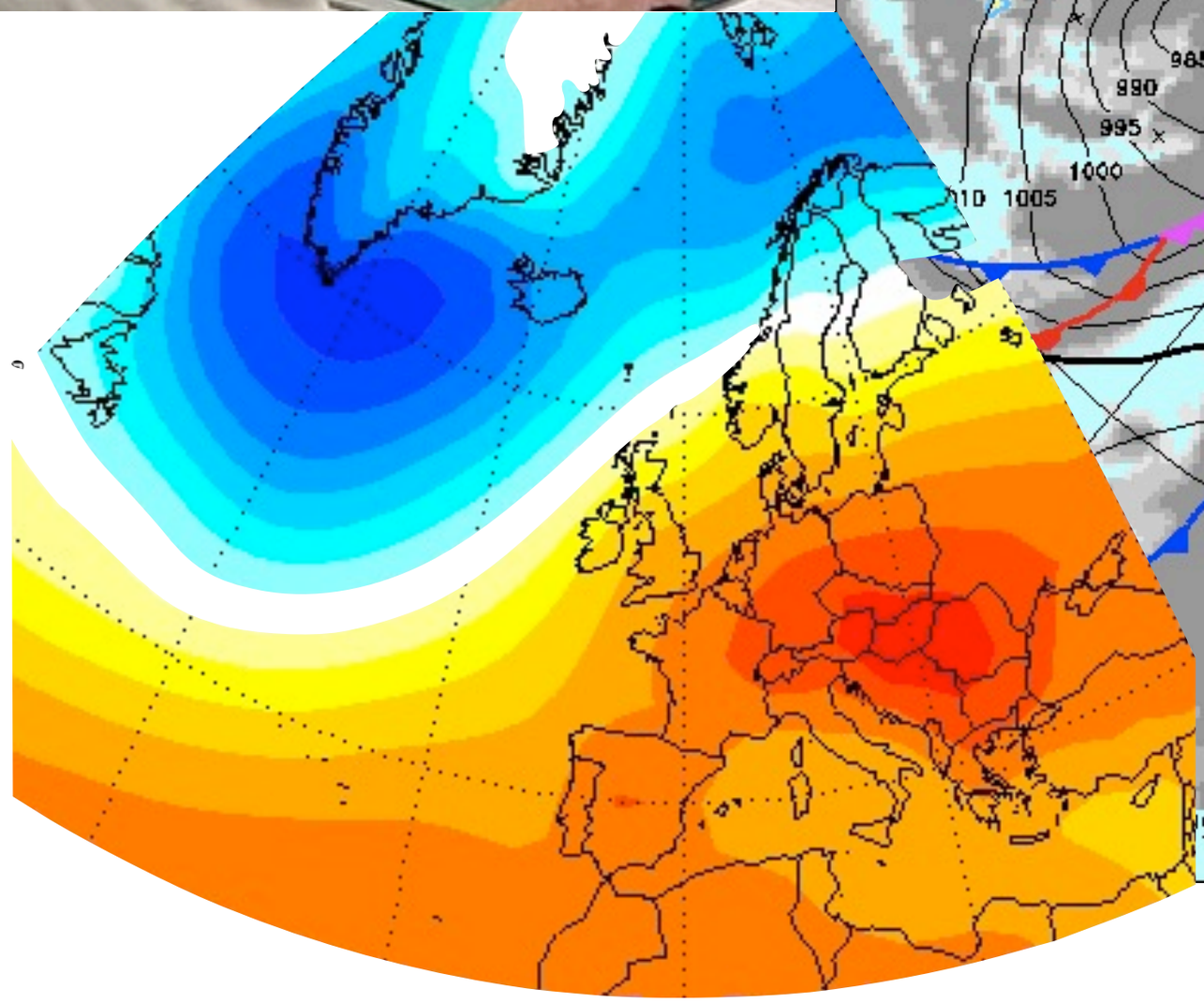


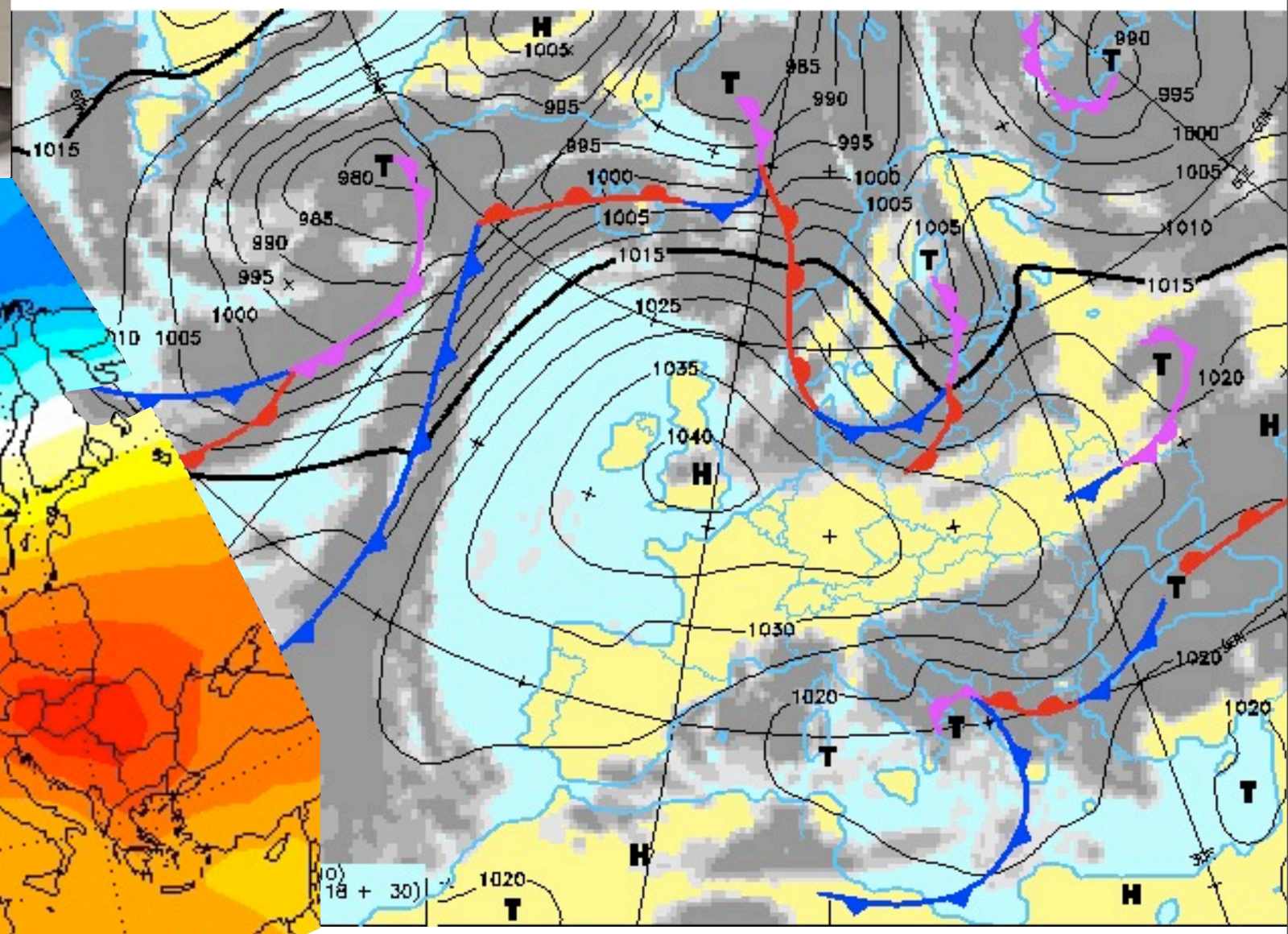
Saturday, 3 December 2011

Weather data comes from tens of thousands of instruments around the world – on land, in the air, and in the oceans.

Lots of different measurements come in, in real-time. Where is the information?

Weather Information





Saturday, 3 December 2011

Some information comes from correlating data from different sources, temporally and spatially – often literally connecting the dots to produce isobars, and isotherms, lows and highs, fronts, storms and calms.

People interpret the information.

Weather Forecast

Saturday, 3 December 2011

They still make mistakes!

What is information?

www.cs.cornell.edu/~crandall/photomap/ **Mapping the World's Photos**

Saturday, 3 December 2011

David Crandall, Lars Backstrom, Dan Huttenlocher, Jon Kleinberg
www.cs.cornell.edu/~crandall/photomap/
Mapping the World's Photos
Flickr contains image data with text tags and lots of them have geospatial data from the GPS now common on phones and cameras

There is no map here, just a plot of the positions of the photos.

Taking all the photos near places where there are many photos and looking at their tags, the authors find tags used in those places - but not elsewhere. These often turn out to be the names of landmarks.

Tags

Edinburgh, Castle, UK,
World Heritage Site

http://www.flickr.com/

Saturday, 3 December 2011

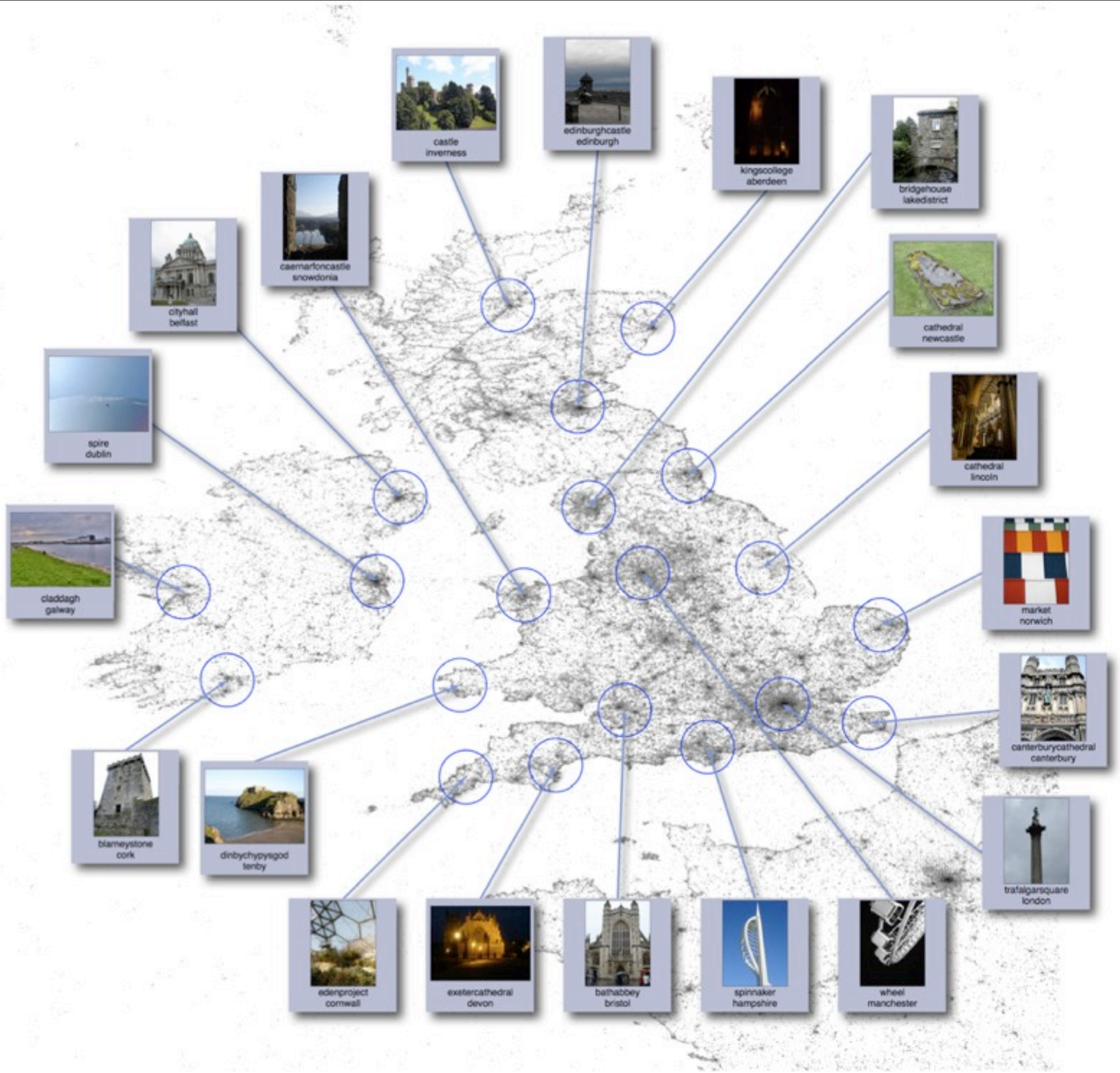
Photo with tags (separated by commas).

Tags

fz, fz30, panasonic, lumix, europe, england, UK, united kingdom, Edinburgh, art, Scotland, 84 Points, JudgmentDay62, 123faves, notpicked, Photoshop, CS2, I500, Explore18Apr06, Interestingness, Interestingness48, Mirea's Realm, accepted@1of100, been@1of100, challenge-you-winner, challenge-you, water, fountain, castle

Saturday, 3 December 2011

Tags often include lots of “noise”. When we have lots of data (in this case lots of photos), we can sort out the information from the noise.



Saturday, 3 December 2011

Information is identification of named landmarks in various cities.

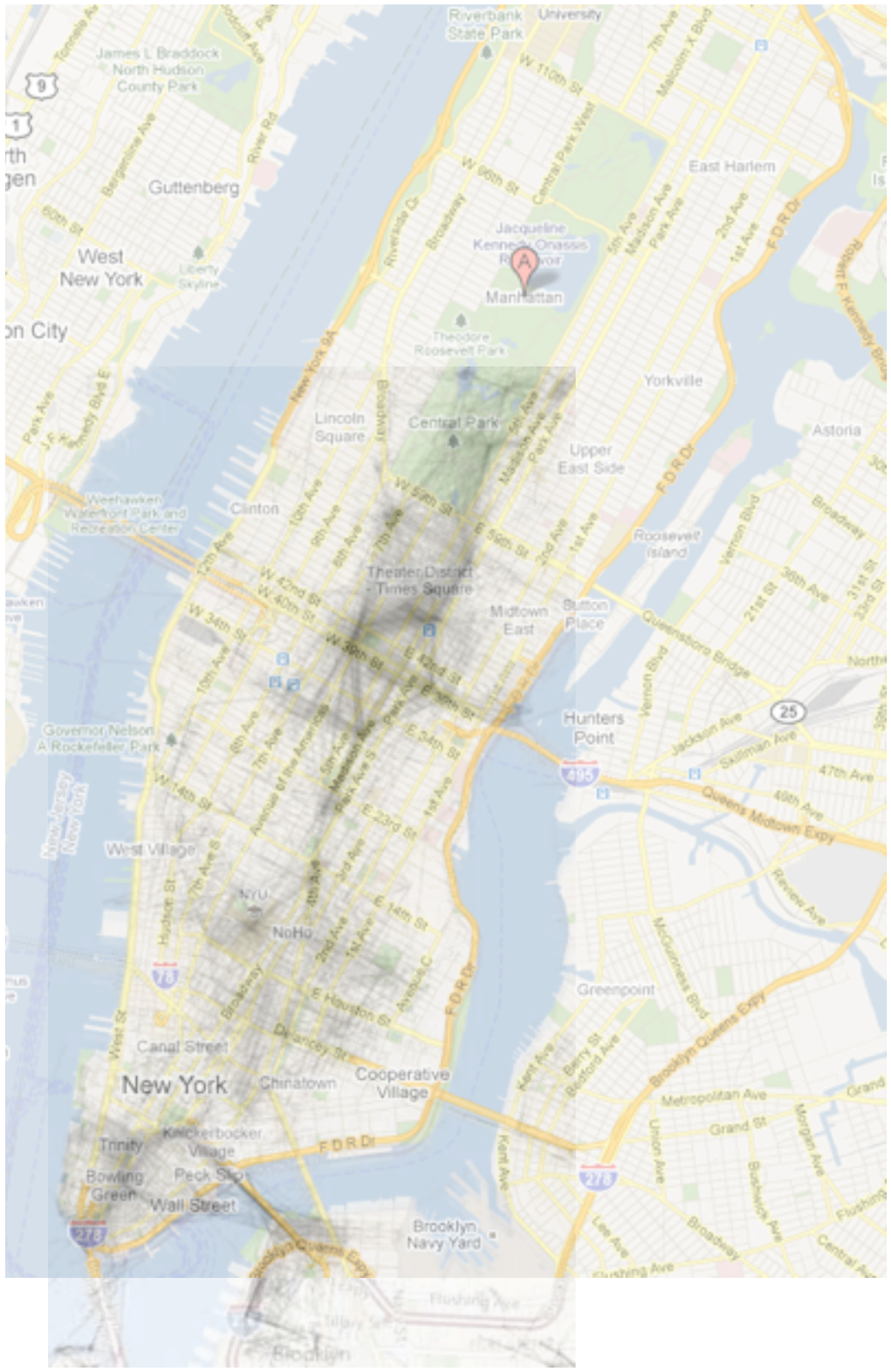
same flickr data

trails of users in Manhattan

Different analyses of the same data can give different information.

By looking at individual photostreams and the times that photos were taken, we can follow people as they travel around taking pictures.

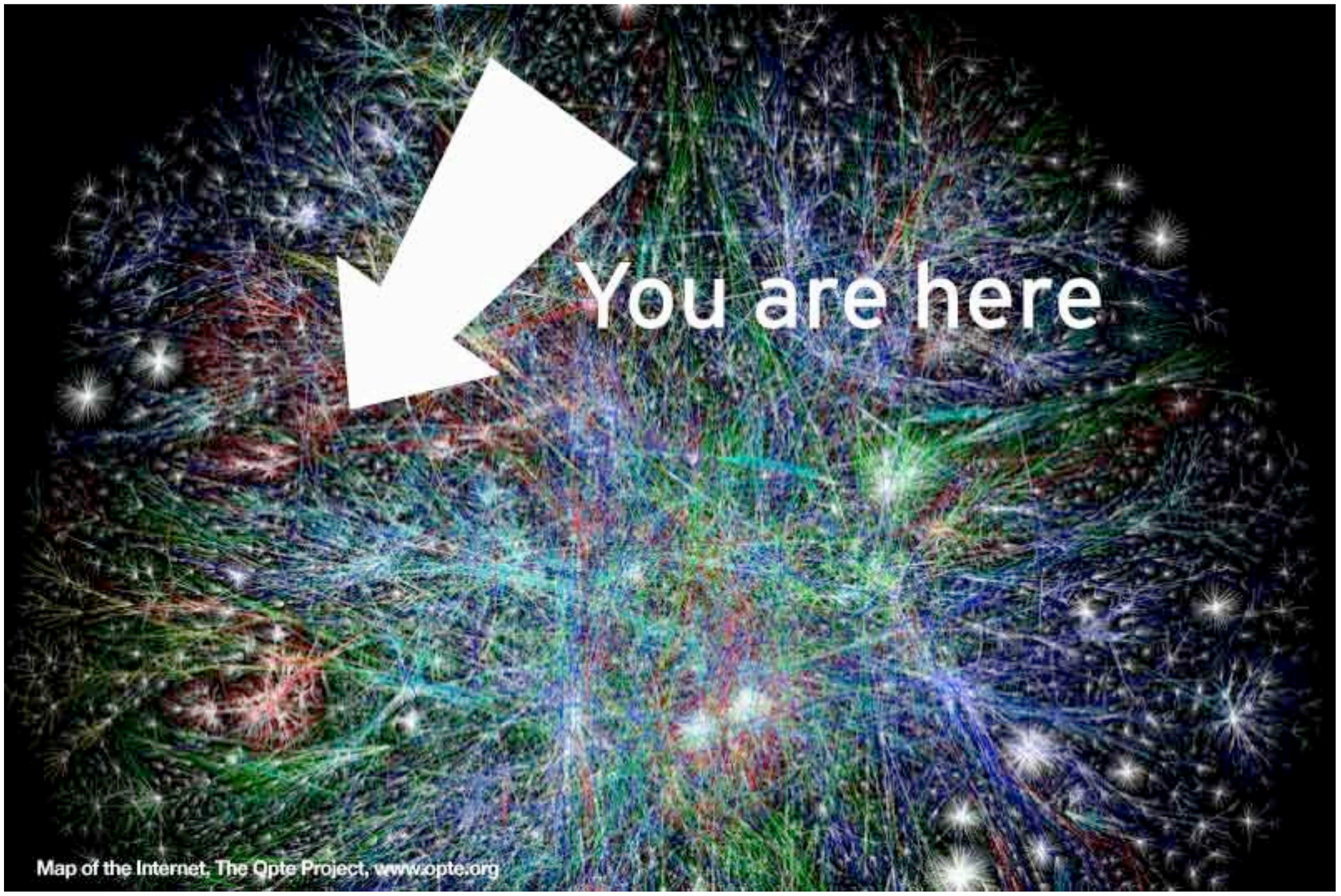
Here we see the network of links people follow when moving from one landmark to the next.



The paths (unsurprisingly) correlate with known landmarks.

Networks are of great interest in informatics – they'll keep cropping up.

This is a network of paths taken – another network you've all heard about is the internet.



Saturday, 3 December 2011

The internet is a network of **inter**connected **net**works. We won't get into detail today ...

More on computer networks on Monday 10th October

graphs or networks

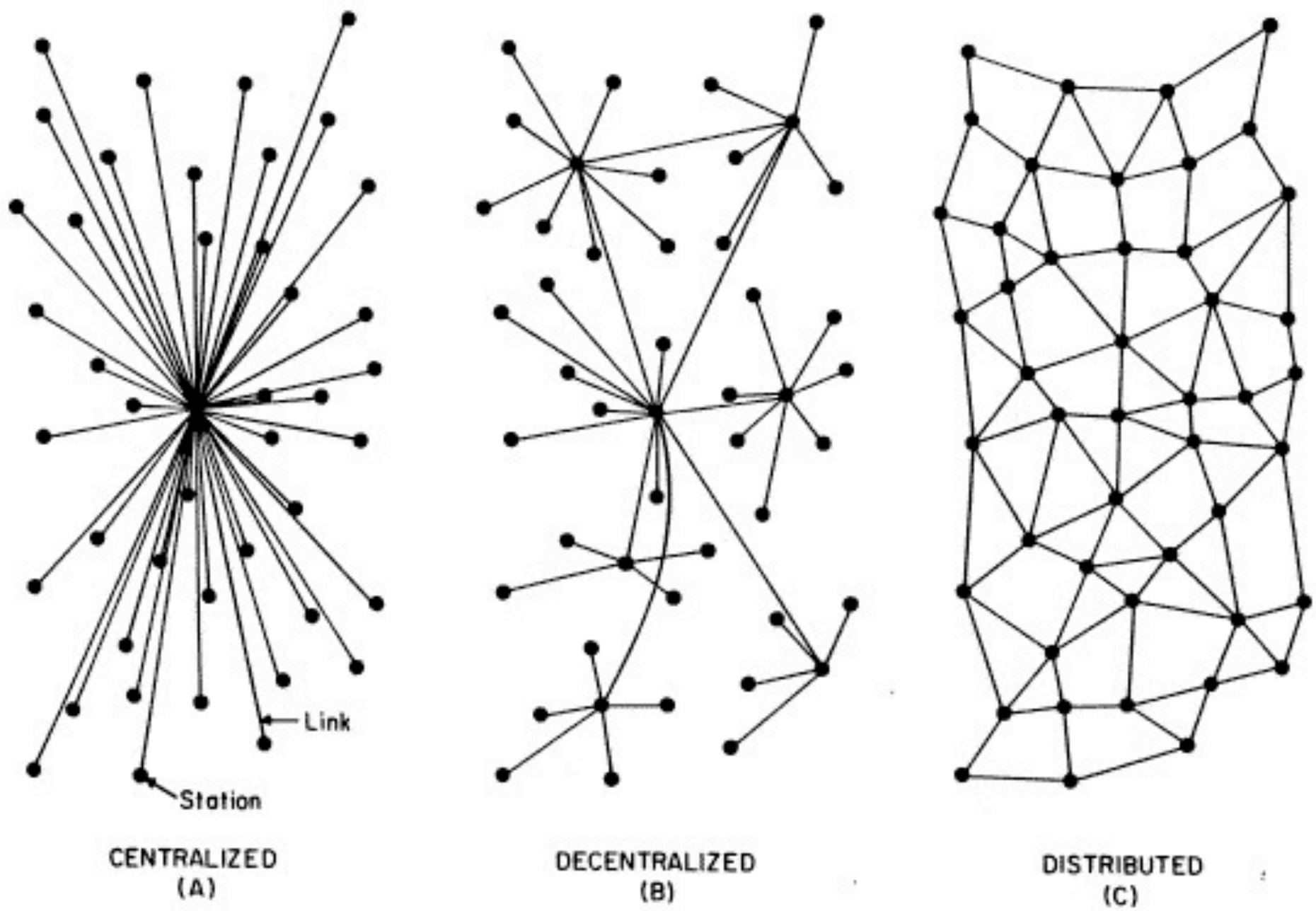
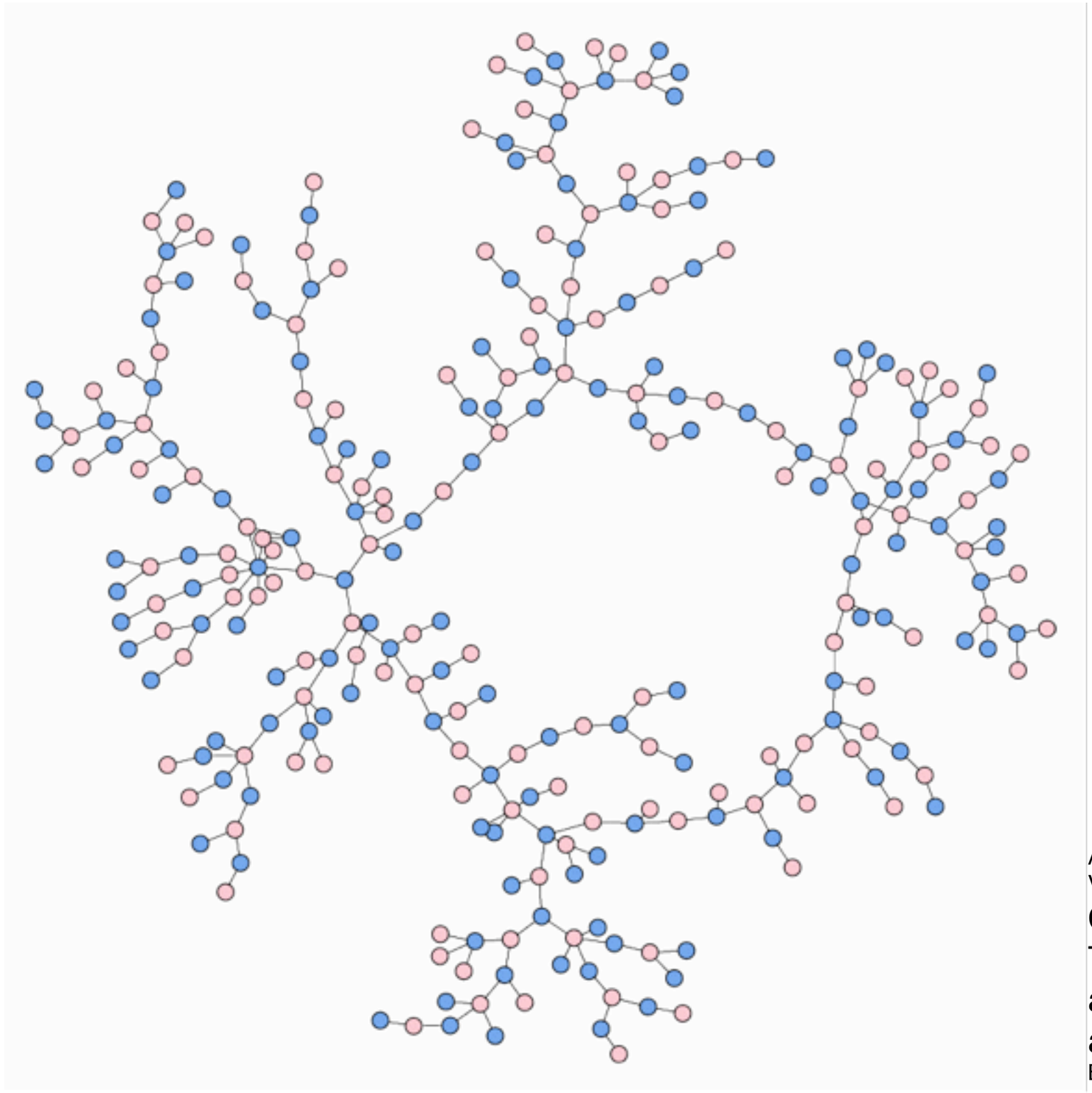


FIG. 1 — Centralized, Decentralized and Distributed Networks

What can we say about networks?
Different kinds of network.
Lots of examples.
Mathematicians call them graphs, and talk about nodes and edges
This diagram talks about stations and links
Other networks might talk about places and paths, or people and friendships.
Lots of examples!

We can apply the ideas to social networks, computer networks, genes etc.

high-school sweethearts



American Journal of Sociology, Vol. 100, No. 1.

Chains of affection: The structure of adolescent romantic and sexual networks,

Bearman PS, Moody J, Stovel K.

Saturday, 3 December 2011

the adolescent romantic and sexual network in a population of adolescents residing in a midsized town in the midwestern United States

“the observed structure reveals networks characterized by longer contact chains and fewer cycles than expected”

blue: links <a>

red: tables <table> <tr> <td>

green: structure

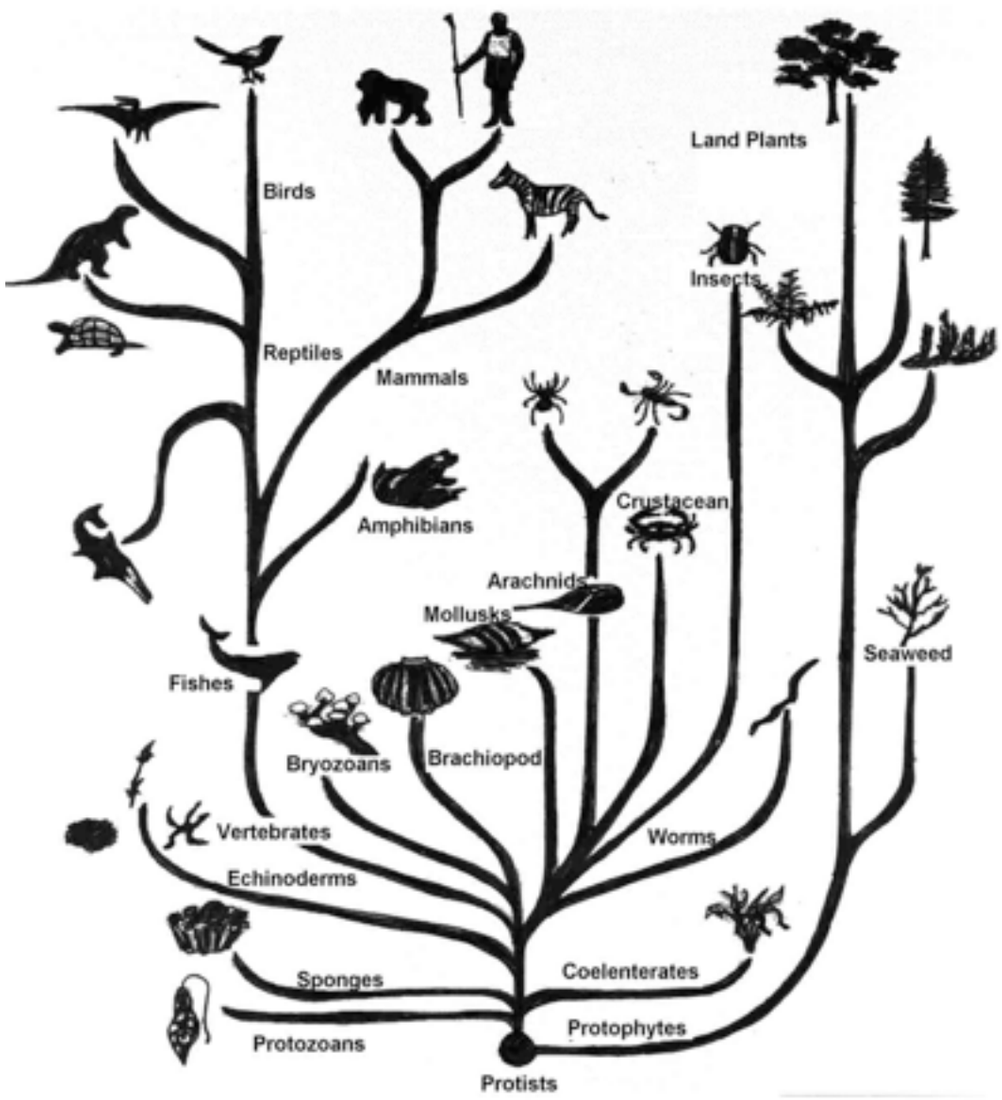
violet: images <img>

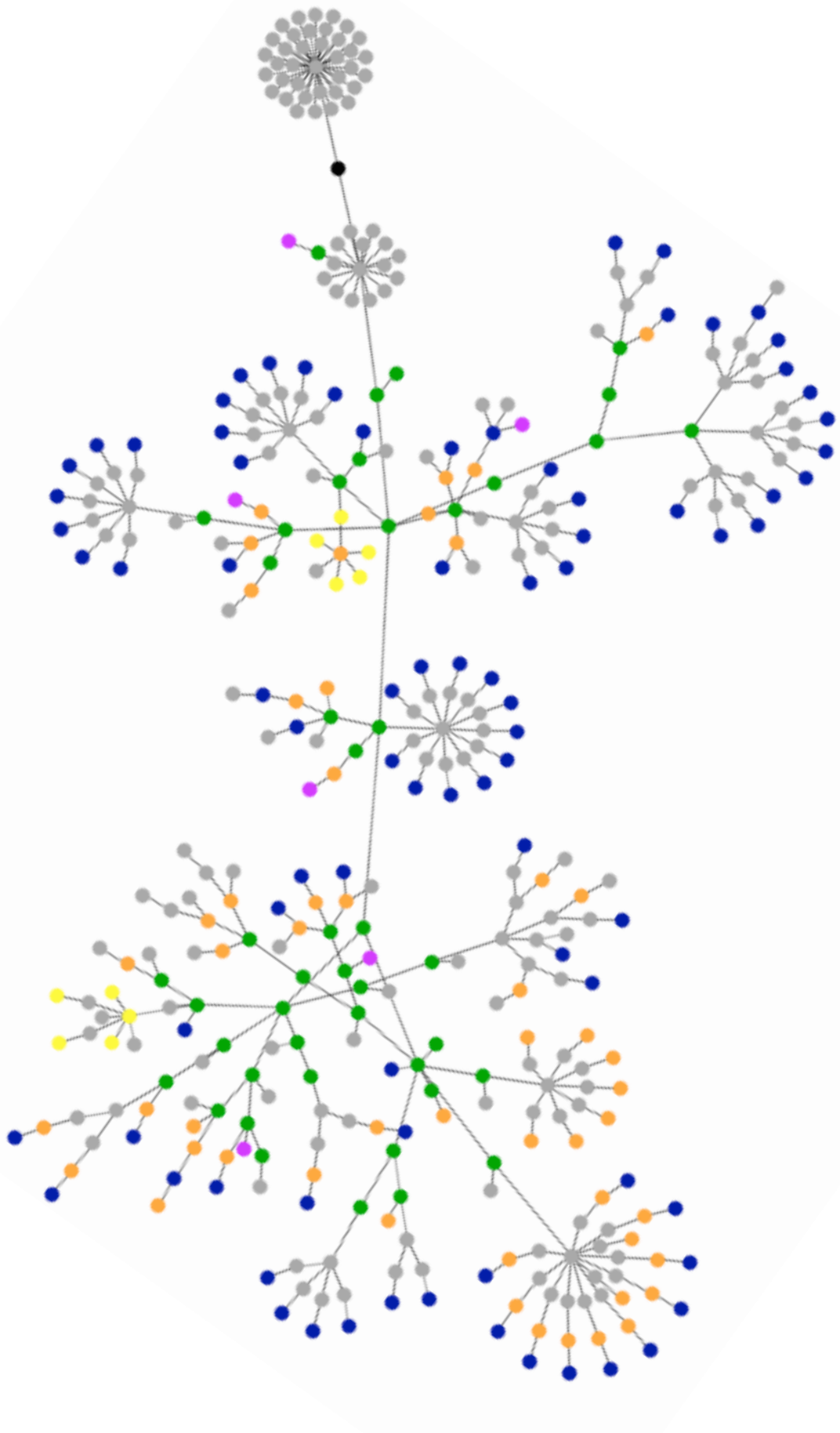
yellow: forms <form> <input> <textarea> <select> <option>

orange: layout
 <p> <blockquote>

black: the root node <html>

gray: all other tags <...>



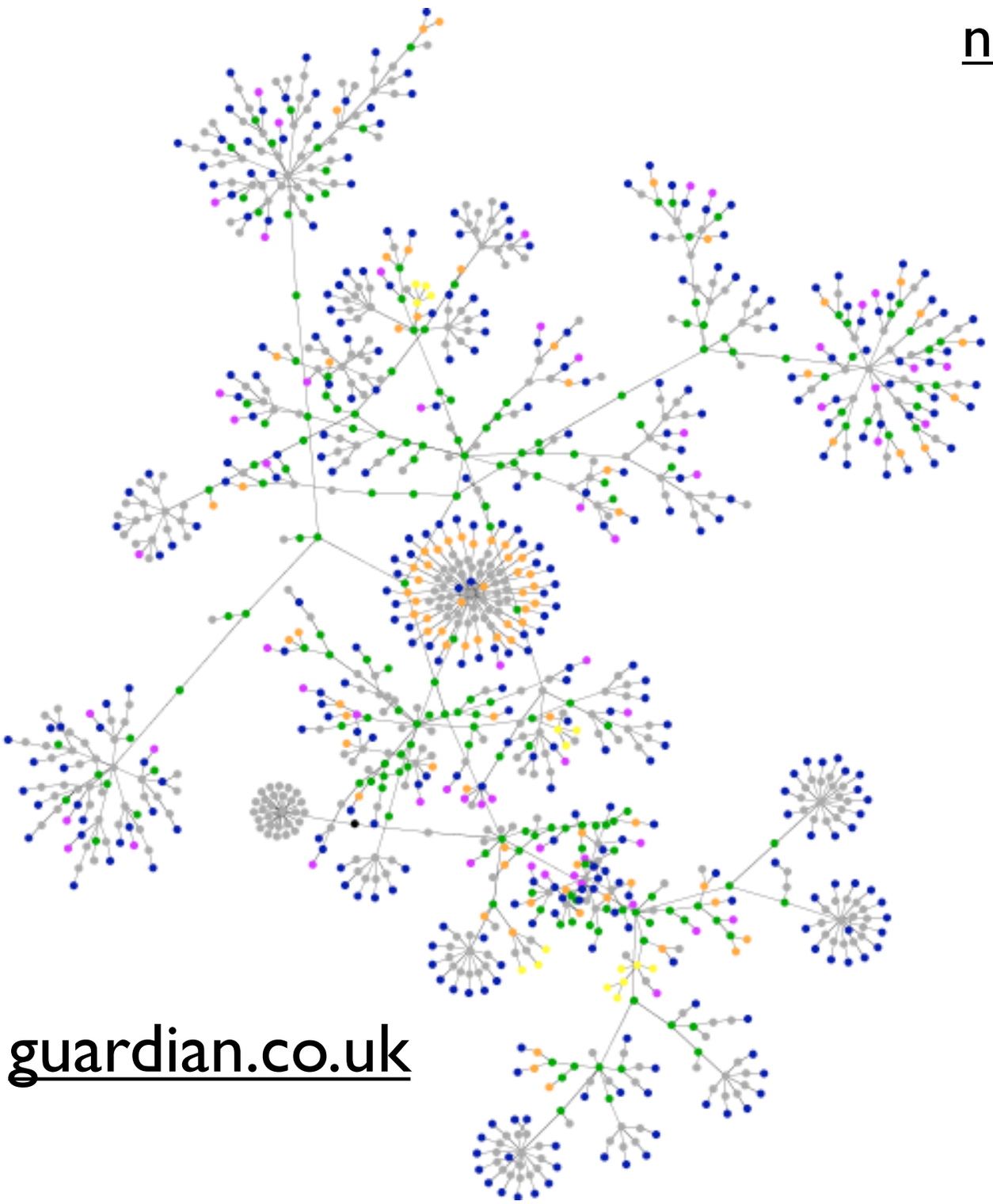


Saturday, 3 December 2011

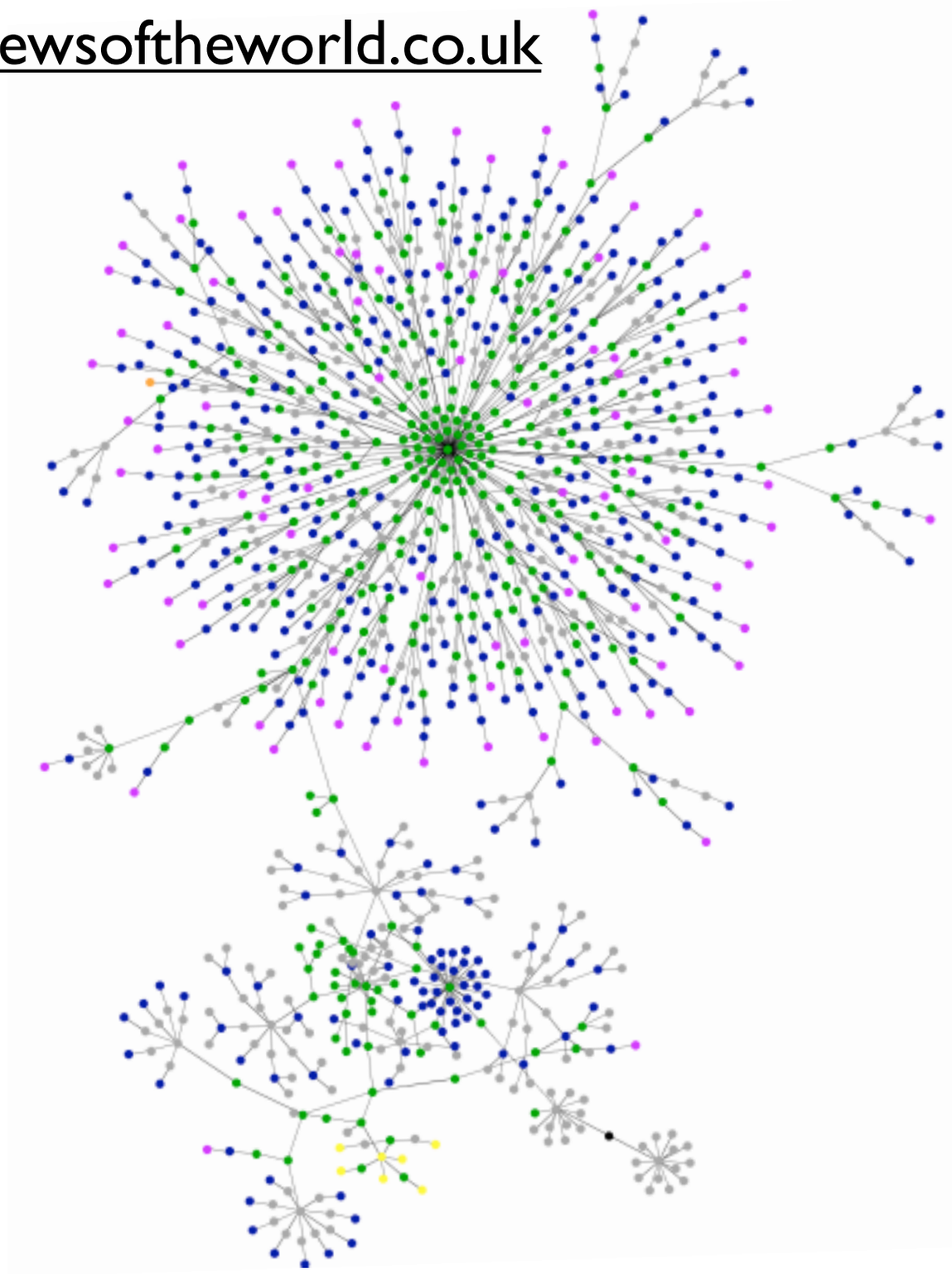
Trees are a special kind of network– networks with exactly one path between any two points.

Trees can be used to represent hierarchical or nested structures like the tree of life with ancestor and descendant species, or a book with chapters, sections, subsections and paragraphs, or a web page ...

Web pages have structure – here we look at the structure of a web page as a tree.
everything is contained in the root node <html> shown in **black** (top right)
pages are divided up into logical sections
text is organised into paragraphs <p>
There are links <a>
tables <table>
forms <form>
– and some parts are contained in others.



guardian.co.uk



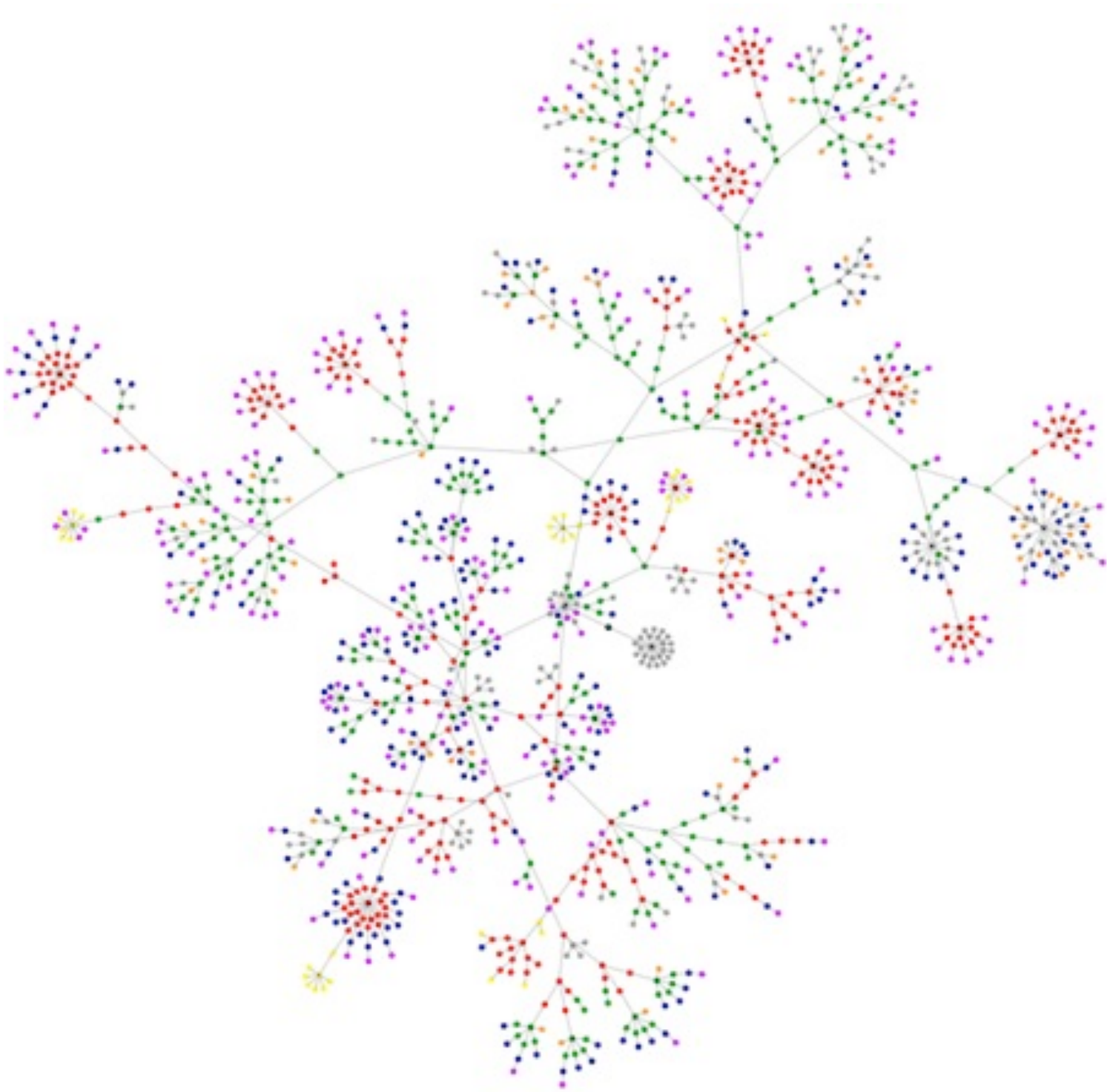
newsoftheworld.co.uk

Saturday, 3 December 2011

Different pages have different structures

Compare the Guardian home page with the News of the World

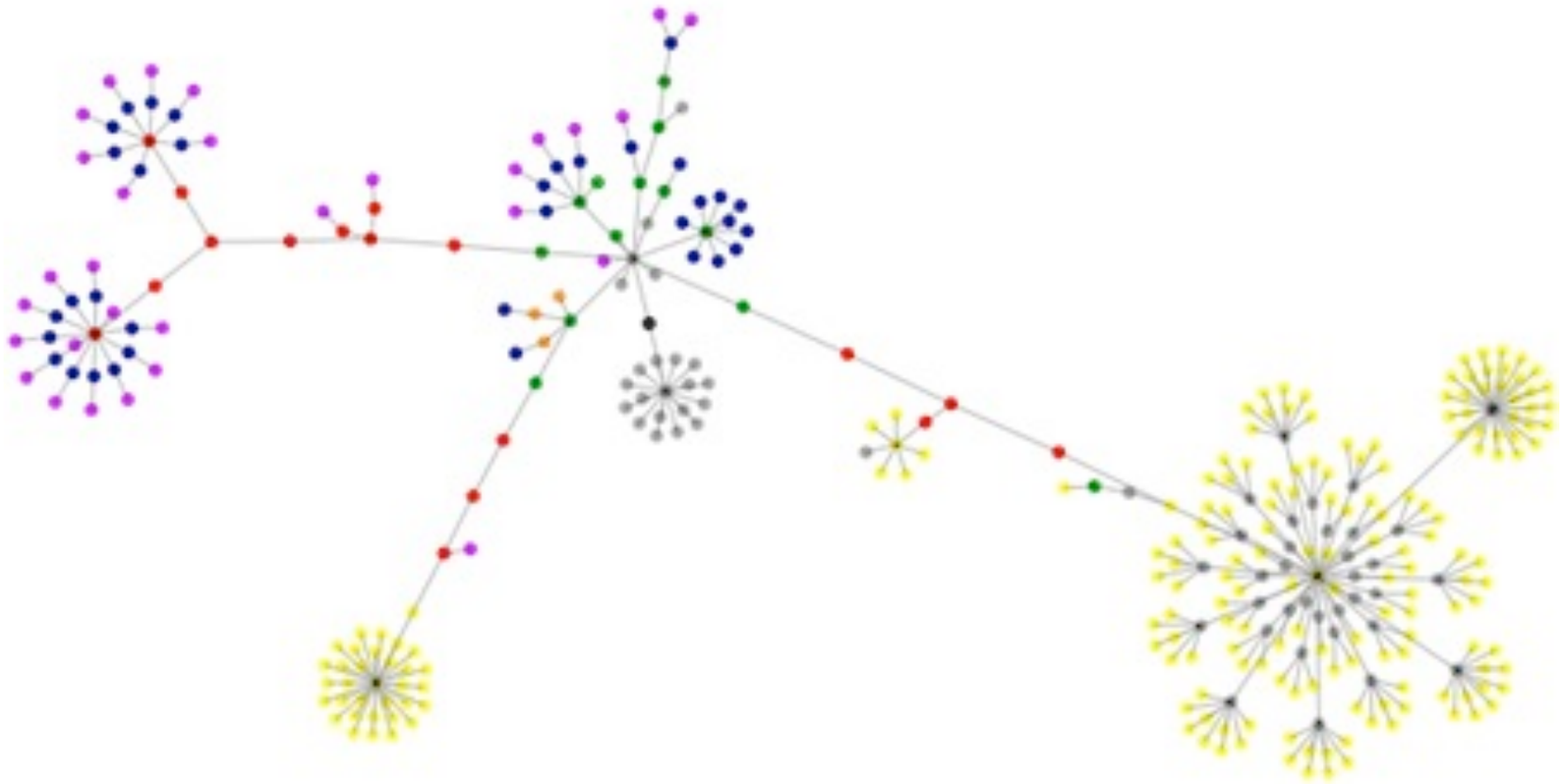
cnn.com



CNN is more like the guardian.

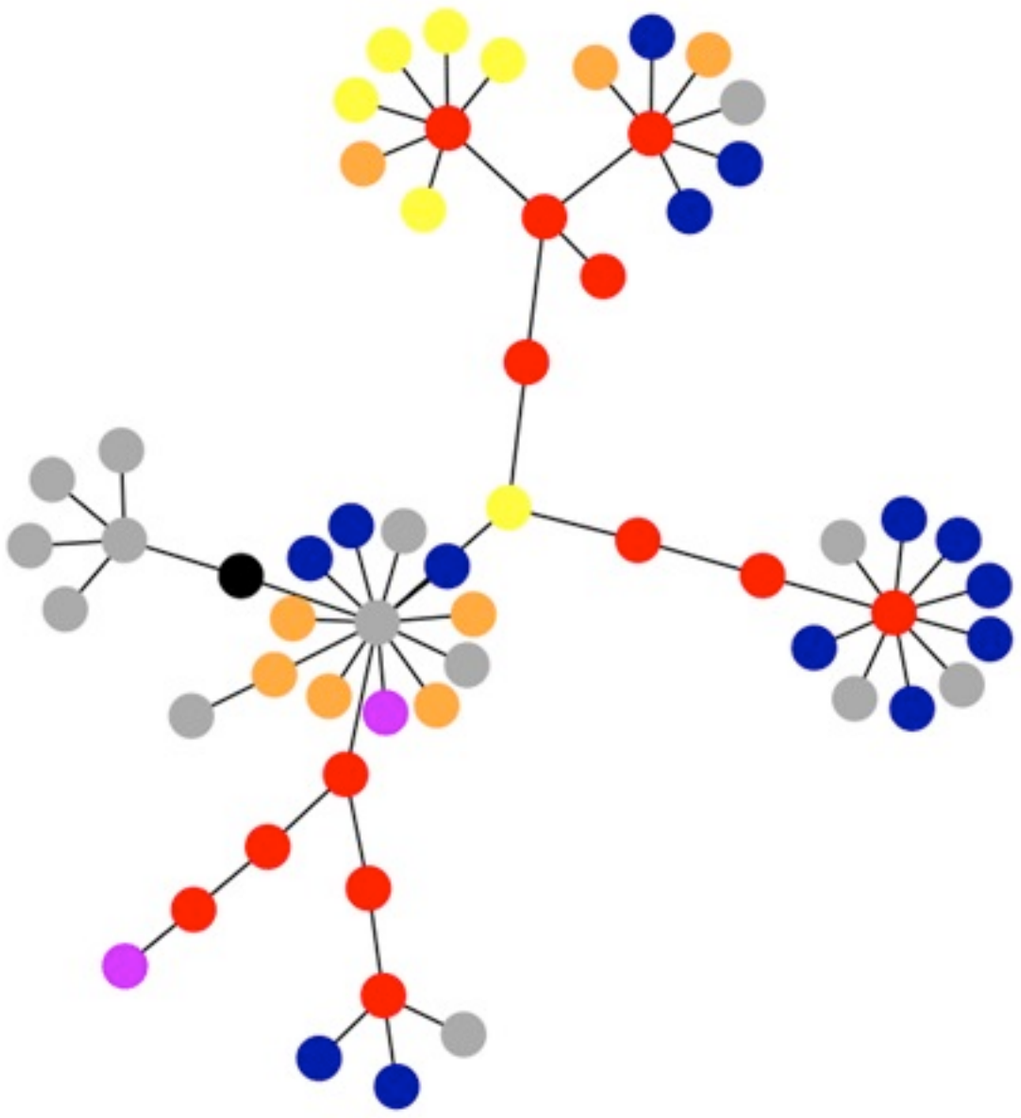
What does this mean? Can we make such statements precise?

apple.com



Apple is simple

google.com



try it!

http://www.aharef.info/

So is Google.

You can try some for yourself.

Now we'll look at how this structure is represented for a web page - you'll be doing this in the lab.

blue: links <a>

red: tables <table> <tr> <td>

green: structure

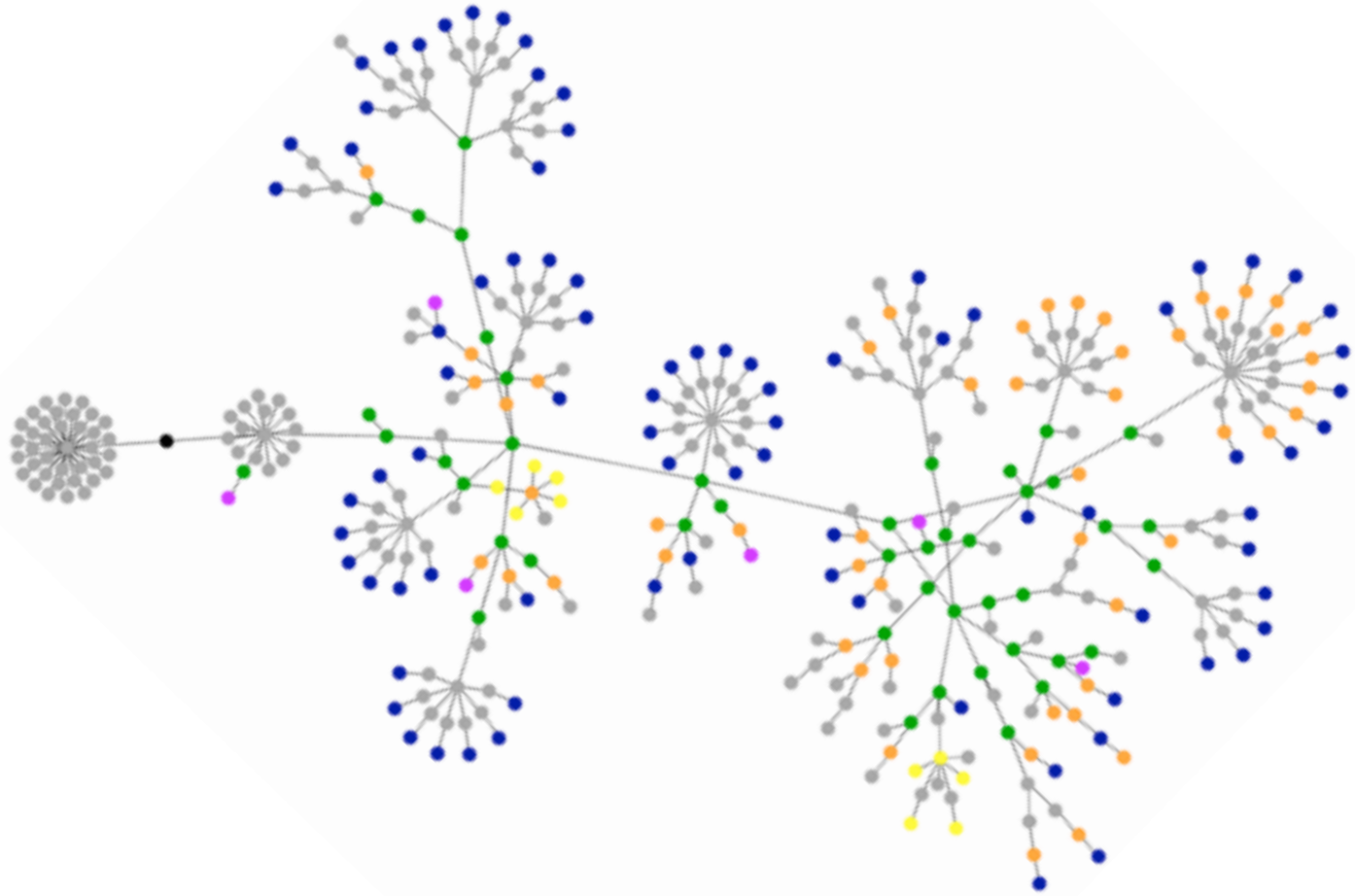
violet: images <img>

yellow: forms <form> <input> <textarea> <select> <option>

orange: layout
 <p> <blockquote>

black: the root node <html>

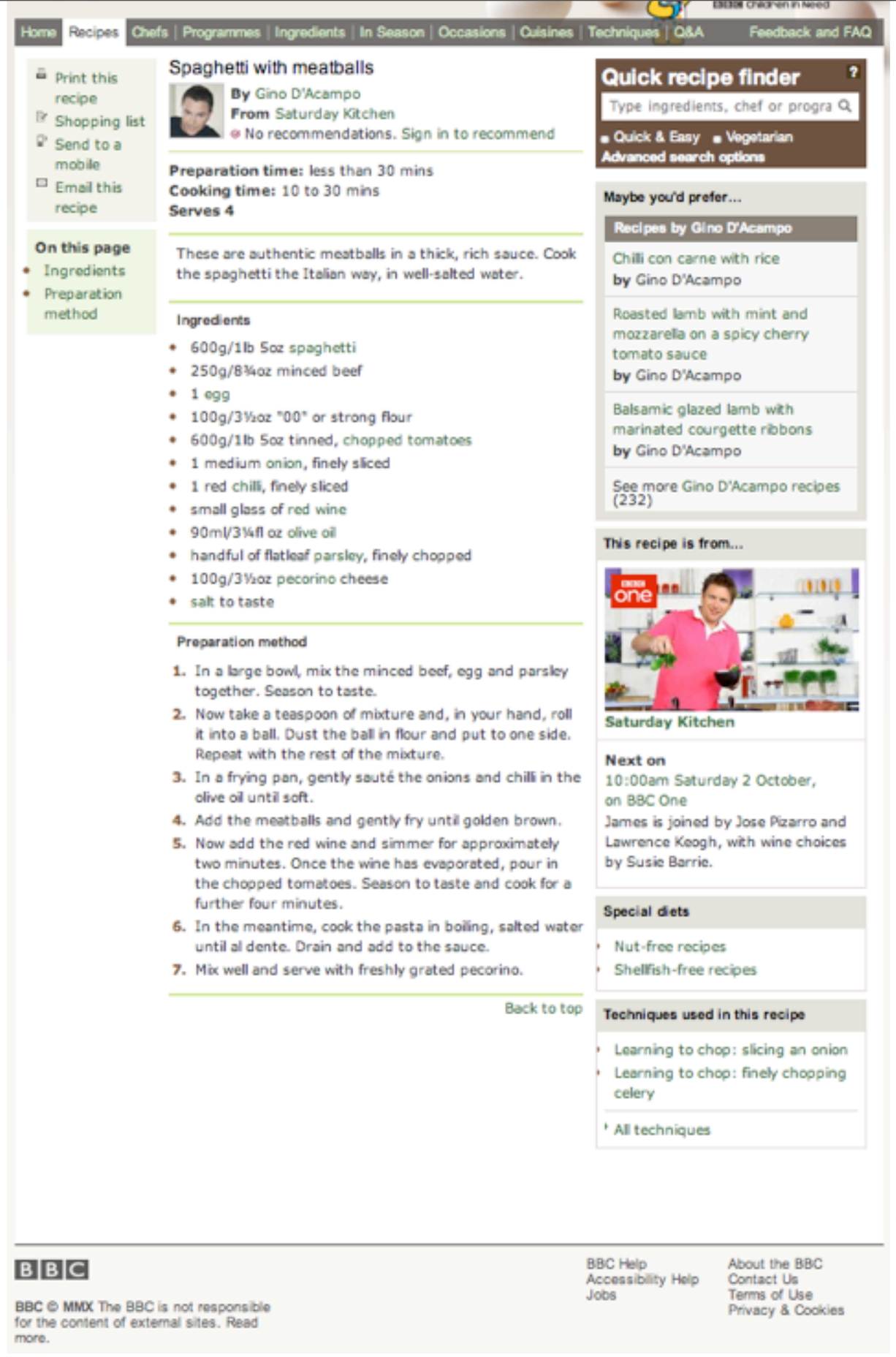
gray: all other tags <...>



Saturday, 3 December 2011

Web pages have structure – a nested, hierarchical structure – here we look at the structure of a web page as a tree.

pages are divided up into logical sections
text is organised into paragraphs <p>
There are links <a>
tables <table>
forms <form>
– and some parts are contained in others.
everything is contained in the root node <html> shown in **black** (top right)




```
<?xml version="1.0" encoding="utf-8"?>
<collection>
  <recipe>
    <title>Spaghetti with meatballs</title>
    <description>
       A classic vegetarian dish.
    </description>

    <ingredient>spaghetti
        <grams>600</grams><ounces>21</ounces>
    </ingredient>
    <ingredient>minced beef
        <grams>250</grams><ounces>8.75</ounces>
    </ingredient>
    <ingredient>egg<number>1</number></ingredient>
    ...
    <method>
      <step>
        In a large bowl, mix the minced beef,
        egg and parsley together. Season to taste.
      </step>
      <step>
         Now take a teaspoon of mixture and, in your
         hand, roll it into a ball. Dust the ball in
         flour and put to one side. Repeat with the
         rest of the mixture.
      </step>
      ...
    </method>
    <time>
      <prep>less than 30 mins</prep>
      <cook>10 to 30 mins</cook>
    </time>
    <servings>4</servings>
    <image>
      <file>1.jpg</file><title>Beef Mince</title>
    </image>
    <by>Gino D'Acampo</by>
    <from>Saturday Kitchen</from>
  </recipe>
  ...
</collection>
```

http://www.bbc.co.uk/food/recipes/database/spaghettiwithmeatbal_72227.shtml

We represent the information in the recipe using XML. XML is a language for representing labelled trees.

Here, we don't use HTML tags – we just make up tags that make sense to us to describe the structure of the recipe – "semantic" tags.

XML is flexible markup language. You can use any set of tags that you want. You just need to match "opening" and "closing" tags.

/

<collection>

<recipe> <recipe> <recipe> <recipe> <recipe> <recipe>

<recipe>

<title> <description> <servings> <time> <image> <ingredient> <by> <from> <method>

<prep> <cook> <file> <title>

<ingredient> <ingredient> <ingredient> <ingredient> <ingredient> <ingredient>

<grams> <ounces> <number>

<step> <step> <step> <step>

XML organises information in labelled trees. We draw them upside-down.

In the next lecture you'll learn how to markup text with these structures using xml and html, and how to control the appearance of web pages using css

Normally web pages are written in html, which uses a standard set of tags that browsers understand.

However, we can also write stylesheets for xml that tell the browser how to present tags that we have invented to describe the structure we are interested in.