Patterns in Data

Nello Cristianini
University of Bristol

“Information Beyond Shannon”
Venice, December 29th, 2008
Information beyond Shannon

.... beyond ....
Something that I care about, and is about information, and about the future...
Venice, 1609
Discovery
SIDEREVS NVNCIVS
MAGNA, LONGEQUE ADMIRABILIA
Speculaculæ pandens, fulpicendiæque proponens
vnicium, præsertim verò
PHILOSOPHIS, atq ASTRONOMIS, que a
GALILEO GALILEO
PATRITIO FLORENTINO
Patarini Gymnasi Publico Mathematico
PERSPICILLI
Nuper à se reperti beneficia sunt observata in LUNÆ FACIE, FIXIS IN-
NUMERIIS, LACTEO CIRCULO, STELLIS NEBULOSIS,
Aprime verò in
QUATVOR PLANETIS
Circa IOVIS Stellam dispaebus interuallis, arque periodis, celebri-
tate mirabili circumvolutis, quos, nemini in hanc vique
diem cognitos, nollirimé Author depræ-
hendit primus; arque
MEDICEA SIDER A
NVNCVPANDOS DECREVIT.
VENETIIS, Apud Thomam Baglionum. M DC X.
Superiorum Permissæ, Æ Privilegio.
Scientific Method

• How do we know what we know? (in science)

• One possible way: experimental science
  • Experiment → Hypothesis → Experiment → Hypothesis → Experiment → Hypothesis → Experiment → Hypothesis → Experiment → Hypothesis
Scientific Method

- Formalising the process of experiment design
- Formalising the representation of information and knowledge
- Formalising the process of inductive inference
- Rigorous (but not perfect) approach to knowledge generation
Scientific Method

- Systematic gathering of information
- Systematic organisation of knowledge
- Systematic sharing of knowledge
- Systematic scrutiny of hypotheses...

- Data acquisition via carefully designed experiments, and hypothesis generation via careful analysis are at the core of scientific method
Scientific Method

- Things are changing fast...
- Biology
- Astronomy
- Chemistry
- Physics
- The process is changing...
SEQUENCING

- Tera bytes
LHC

- 15 petabytes per year
Astronomy

- Automated Sky Surveys

- the Sloan Digital Sky Survey (SDSS) created a 5-wavelength catalogue over 8,000 square degrees of the sky, containing about 200 million objects, described by hundreds of features (data released incrementally to the public).

- Terabytes...
Chemistry

• Combinatorial Chemistry

• Hundreds of thousands of compounds can be generated and tested, either by using robotics, or – increasingly – by computer simulations, in what is essentially a survey of entire regions of chemical space, hunting for compounds with a given set of properties.
To Summarize...

- First we gather GAZILLIONS of data
- Then we let our computers sort it out, and suggest hypotheses...
Many Google DATACENTERs
A Newer Kind of ...?

- High quality experimental information is produced on massive scale, by automated processes.
- And the data is analysed by pattern discovery software, which can generate hypotheses in the form of statistical relations discovered in the data.
Pattern Analysis

- Various disciplines have contributed to this
- Statistics, data mining, machine learning, information theory, ...

- We can MANAGE and ANALYSE massive amounts of data, of any kind...
- Text, DNA, graphs, images, times series, ...

- To what extent can we automate the entire process of scientific discovery?
Pattern Analysis

- The set of principles behind most successful algorithms in PA is actually quite small

- Efficient search and optimisation in vast pattern spaces
- Effective control of multiple testing issues, by use of sharp statistical tools

- Principles that unify all, from suffix trees to itemset mining, from graphical models to kernel methods
Science

• How can this vast toolbox, and this conceptual framework, help us to acquire more scientific knowledge?

• I will show you an example from the most improbable of domains... the SOCIAL sciences...
EXAMPLES FROM BRISTOL

1) Can we make an agent that reads all news in all languages, than makes recommendations?

2) Can we use modern AI tools to map the mediasphere, based on its contents?
Media System

• How do we know what we know? (in society)
Jamie Lynn Spears (18)

Jamie Lynn gives birth
20 Jun 2008 04:16:37 GMT

It’s a girl, y’all! Jamie Lynn Spears gave birth to Maddie Briann this morning at Mississippi Southwest Regional Medical Center in McComb, Miss. It’s the first child for the 17-year old Zoey 101 star and her fiancé, 19-year old Jason Trawick.

Read more...

Jamie Lynn Spears Has a Baby Girl
20 Jun 2008 04:16:11 GMT

There is a new baby in the Spears clan! Jamie Lynn Spears and her fiancé Casey Aldridge welcomed a baby girl Thursday morning. PEOPLE confirms exclusively that the baby was named Maddie Brianna. “Just the family was there,” says a source about the birth around 3:30 a.m. at a hospital in Mississippi, near her Louisiana hometown. “Everyone

Report: Spears Gives Birth To Girl, Maddie
20 Jun 2008 04:16:11 GMT

People magazine reports that Britney Spears’ younger sister, Jamie Lynn Spears, has given birth.
Jamie Lynn Spears (18)

Jamie Lynn gives birth
29 Jun 2008 04:16:57 GMT

It's a girl, y'all! Jamie Lynn Spears gave birth to Maddie Briann this morning at Mississippi Southwest Regional Medical Center in McComb, Miss. It's the first child for the 17-year-old Zoey 101 star and her fiancee, 19-year-old...

Read more...

Jamie Lynn Spears Has a Baby Girl
29 Jun 2008 04:16:14 GMT

There is a new baby in the Spears clan! Jamie Lynn Spears and her fiancé Casey Aldridge welcomed a baby girl Thursday morning. PEOPLE confirmed exclusively that the baby was named Maddie Briann. A representative for the family said that Carey, a source close to the birth, around 8:30 a.m., at a hospital on Mississippi's...
Jamie Lynn Spears (18)

Jamie Lynn gives birth
21 Jun 2008 04:16:27 GMT

It's a girl, y'all! Jamie Lynn Spears gave birth to Maddie Briann this morning at Mississippi Southwest Regional Medical Center in McComb, Miss. It's the first child for the 17-year-old Zoey 101 star and her fiance, 19-year-old Jamie Lynn Spears.

Read more...

Jamie Lynn Spears Has a Baby Girl
21 Jun 2008 04:16:14 GMT

There is a new baby in the Spears clan! Jamie Lynn Spears and her fiancé Casey Aldridge welcomed a baby girl Thursday morning. PEOPLE confirms exclusively that the baby was named Maddie Briann. "Just the family was there," says a source about the birth around 9:30 a.m. at a hospital in Mississippi, near her Louisiana hometown. "Everyone

Read more...

Reports: Spears Gives Birth To Girl, Maddie
21 Jun 2008 04:15:22 GMT

People magazine reports that Britney Spears' younger sister, Jamie Lynn Spears, has given birth.

Read more...

Jamie Lynn Spears: It's a girl...
Read (73)

China to raise fuel prices
20 Jun 2006 04:13:53 GMT
Read full story for latest details.
Read more...

Thailand braces for anti-government protests
20 Jun 2006 04:13:53 GMT
Read full story for latest details.
Read more...

French economy forecast to halt
20 Jun 2006 04:13:53 GMT
Read full story for latest details.
Read more...
David Beckham (8)

San Francisco crowd goes nuts as David Beckham takes it all off
20 Jun 2008 04:13:55 GMT

Don’t miss the video of a San Francisco crowd shrieking and squealing with delight as a gigantic, almost-nude shot of David Beckham was revealed inch-by-inch at Macy’s in Union Square yesterday. The huge billboard is an ad for Emporio Armani underwear, but the promoters kept it draped until showtime and then dropped the curtain slowly as the crowd went wild. It was better than a strip joint and you don’t have to stick any money in his drawers.

David Beckham’s hot manhood
20 Jun 2008 04:11:04 GMT
EU (286)

EU aid to Bulgaria longer suspended
The European Union has frozen aid to EU Member Bulgaria because of fraud and corruption. The aid will total EUR 560 million and was intended for agriculture and infrastructure.
Read more...

Greenland votes on autonomy
In Greenland on Tuesday, the polls are open for a referendum on extending the autonomy of the capital island to Denmark.
Read more...

Yellow River is heavily polluted
A third of the Yellow River in China is so polluted by industrial waste water not only is inappropriate as drinking water, but also for the industry or agriculture.
Read more...

A woman washes up baby by toilets
A British woman to the judge said Monday that her new-born baby toilets have inadvertently by the year. The woman said she thought she had diarrhoea.
Read more...

Less than 50% support European Union in the Cypriot
Belgium (11)

EU aid to Bulgaria longer suspended

The European Union has frozen aid to EU Member Bulgaria because of fraud and corruption. The aid will total EUR 560 million and was intended for agriculture and infrastructure.

Read more...

Greenland votes on autonomy

In Greenland on Tuesday, the polls are open for a referendum on extending the autonomy of the capital island to Denmark.

Read more...

Yellow River is heavily polluted

A third of the Yellow River in China is so polluted by industrial waste water not only is inappropriate as drinking water, but also for the industry or agriculture.

Read more...

A woman washes up baby by toilets

A British woman told the judge said Monday that her new-born baby toilets have inadvertently by the year. The woman said she thought she had diarrhoea.

Read more...
Mapping the Mediasphere

- We are interested in describing the mediasphere as a single, unique, system
- We are interested in the flow of information through the system
- And in any other patterns, both dynamic and static, that can be detected in its contents
Tracking memes

• We track memes as they travel from outlet to outlet

• We use them to reconstruct the invisible influence structure of the system
The mediasphere is the integrated system of global news media, including tens of thousands of broadcast, press, online outlets, in dozens of languages.

How big is it?
The Size of the Mediasphere

9138 outlets appeared at some point in Google News
(does this include foreign outlets?) –
they have 22 languages – but different set of languages from us

If we sum together the sizes declared by google for each language
(they are all approx, we get a similar figure – so we know they over-state a bit)

<table>
<thead>
<tr>
<th>Language</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>4500</td>
</tr>
<tr>
<td>China</td>
<td>1000</td>
</tr>
<tr>
<td>Korea</td>
<td>1000</td>
</tr>
<tr>
<td>Spanish</td>
<td>700</td>
</tr>
<tr>
<td>German</td>
<td>700</td>
</tr>
<tr>
<td>Japan</td>
<td>610</td>
</tr>
<tr>
<td>French</td>
<td>500</td>
</tr>
<tr>
<td>Belgie</td>
<td>400</td>
</tr>
<tr>
<td>Russia</td>
<td>400</td>
</tr>
<tr>
<td>cesca rep</td>
<td>400</td>
</tr>
<tr>
<td>Nederland</td>
<td>400</td>
</tr>
<tr>
<td>Norge</td>
<td>400</td>
</tr>
<tr>
<td>Taiwan</td>
<td>350</td>
</tr>
<tr>
<td>Polska</td>
<td>300</td>
</tr>
<tr>
<td>Italian</td>
<td>250</td>
</tr>
<tr>
<td>Turkie</td>
<td>200</td>
</tr>
<tr>
<td>Portugal</td>
<td>200</td>
</tr>
<tr>
<td>Arabic</td>
<td>150</td>
</tr>
<tr>
<td>Sverige</td>
<td>100</td>
</tr>
<tr>
<td>Israel</td>
<td>100</td>
</tr>
<tr>
<td>Greece</td>
<td>100</td>
</tr>
<tr>
<td>IndiaB</td>
<td>20</td>
</tr>
<tr>
<td>SUM</td>
<td>12780</td>
</tr>
</tbody>
</table>

---
Circulation of top 100 global newspapers
Our Data

- For period: April to Nov 2008
  We have 1007 outlets

- For 507 we have complete coverage
  (all top stories for the entire period)

- For 723 we have 80% coverage
  (all top stories for at least 80% of the period).
Our Data – By Media Type

Our Data

By Media Type

0 200 400 600 800 1000 1200
TP>Blog
TP>Broadcast
TP>Magazine
TP>NewsCommunity
TP>Newspaper
TP>Newswire
TP>OnlineOnlyMedia
TP>PRFeed

# Outlets
# feeds
The Contents of the Mediasphere

- NAMED ENTITIES:

- Mean number of appearances per person = 6.0995
  Median number of appearances per person = 1
  Total number of people = 447221
- Mean number of appearances per organisation = 5.3305
  Median number of appearances per organisation = 1
  Total number of organisations = 420637
- Mean number of appearances per location = 61.7462
  Median number of appearances per location = 1
Distribution of Entities

![LogLog plot of the distribution](image)
Distribution of Entities
Distribution of Entities

LogLog plot of the distribution

Appearance Count

Person
Co-Mention Network

- The co-mention network can be used to link significantly associated entities, and hence be used to measure their social CENTRALITY.
Entity Networks

Diagram showing connections between various entities such as 'bc', 'vancouver', 'toronto', 'ottawa', 'calgary', 'eu', 'zimbabwe', 'canada', 'iraq', 'afghanistan', 'russia', 'baghdad', 'taliban', 'al qaeda', 'nato', 'atlanta', 'moscow', 'kabul', 'vladimir putin', 'georgia', 'nepal'.
Detecting Changes

John McCain (153)

The politics of hate
13 Oct 2008 04:28:26 GMT

The ugly desperation of the McCain campaign rampages further out of control. Media Bubble, Oct. 8 -- John McCain's bid for the Oval Office suffered another stunning blow yesterday when
Pope-ularity

- In the US outlets, in the period April to November 2008, we measured the fraction of news items mentioning the Pope (in Content? Just in title-description?)

- The fraction is
  - 0.102% for English speaking outlets
  - 0.377% for Spanish speaking outlets
Pope-ularity

1. 'AFRICA'
2. 'UNITED KINGDOM'
3. 'ASIA'
4. 'CANADA'
5. 'UNITED STATES'
6. 'ITALY'
7. 'FRANCE'
8. 'GERMANY'
9. 'VATICAN'
Readability Index

COM 11.97 2.49
EDU 14.47 2.36
Text re-Use Networks
However...

- As the gathering and analysis of vast amounts of information is enabling a phase transition in the way we do science...

- ...the same force is behind another – not so visible – phase transition.
Privacy in the data-driven age...

Surveillance works exactly in the same way...
Conclusion

- Science is changing (again)
- Social sciences may benefit enormously from AI technology
- Much of science can be done automatically
- Warning: same is happening to other aspects of life
- Data revolution is moving faster than our philosophical and legal frameworks...