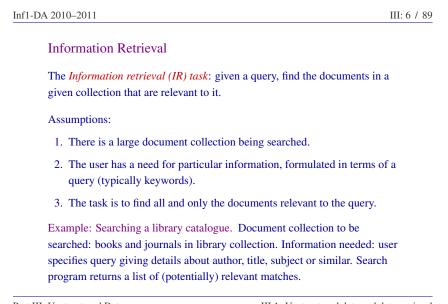
nf1-DA 2010–2011	III: 1 / 89
Informatics 1	
School of Informatics, University of Edinburgh	
Data and Analysis	
Part III	
<b>Unstructured Data</b>	
Ian Stark	
February 2011	
art III: Unstructured Data	

Inf1-DA 2010–2011	III: 2 / 89
Part III — Unstructured Data	
Data Retrieval:	
III.1 Unstructured data and data r	etrieval
Statistical Analysis of Data:	
<b>III.2</b> Data scales and summary statist	tics
<b>III.3</b> Hypothesis testing and correlati	ion
<b>III.4</b> $\chi^2$ and collocations	
Part III: Unstructured Data	III.1: Unstructured data and data retrieval

Inf1-DA 2010–2011	III: 3 / 89
Staff-Student Liaison Meeting	
• Today 1pm	
• Informatics 1 teaching staff and stu	dent reps
• Send mail to the reps at inflreps@l comments you would like them to r	•
Coursework Assignment	
• Three sample exam questions, down	nload from course web page
• Due 4pm Friday 11 March, to box of	outside ITO
• Marked by tutors and returned for a	discussion in week 11 tutorial
• Not for credit; you can discuss and	ask for help (do!)
Part III: Unstructured Data	III.1: Unstructured data and data retrieva

Inf1-DA 2010–2011	III: 4 / 89
Examples of Unstructured Data	I.
• Plain text.	
data, not imposed.	of characters, but this is <i>intrinsic</i> to the by, e.g., annotating (as in Part II).
These again have <i>intrinsic</i> structu	by, e.g., recognising objects, isolating
<ul> <li>Experimental results.</li> <li>Here there may be structure in here points in <i>n</i>-dimensional space).</li> <li>But an important objective is to u confirm or refute an experimental</li> </ul>	
Part III: Unstructured Data	III.1: Unstructured data and data retrieval

f1-DA 2010–2011	III: 5 / 89
Topics	
We consider two topics in dealing with	unstructured data.
1. Information retrieval	
How to find data of interest in with documents.	in a collection of unstructured data
2. Statistical analysis of data	
How to use statistics to identify and	d extract properties from
unstructured data (e.g., general tren components, etc.)	nds, correlations between different
rt III: Unstructured Data	III.1: Unstructured data and data retrieva



III.1: Unstructured data and data retrieval

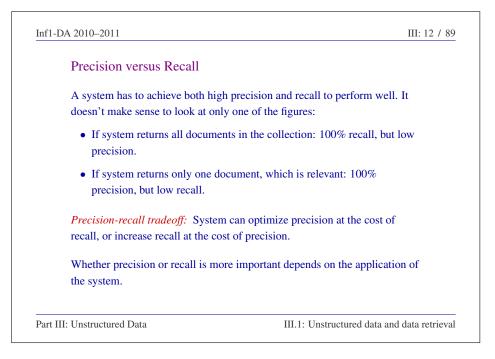
Inf1-DA 2010–2011	III: 7 / 89
Key issues for IR	
Specification issues:	
• Evaluation: How to measure the perfo	rmance of an IR system.
• Query type: How to formulate queries	to an IR system.
• Retrieval model: How to find the best <i>rank</i> them in order of relevance.	matching document, and how to
Implementation issues:	
• Indexing: how to represent the docum that the search can be done efficiently.	ents searched by the system so
The goal of this lecture is to look at the thro detail.	ee specification issues in more
Part III: Unstructured Data	III.1: Unstructured data and data retrieva

1-DA 2010–2011	III: 8 / 89
Evaluation of IR	
The performance of an IR system measures:	n is naturally evaluated in terms of two
• <i>Precision:</i> What proportion match the original objectives	of the documents returned by the system s of the search.
• <i>Recall:</i> What proportion of the search are returned by the	the documents matching the objectives of e system.
We call documents matching the <i>documents</i> .	objectives of the search <i>relevant</i>
rt III: Unstructured Data	III.1: Unstructured data and data retrieva

Inf1-DA 2010-2011				III: 9 / 89
True/false	e positives/neg	atives		
		Relevant	Non-relevant	
	Retrieved	true positives	false positives	
	Not retrieved	false negatives	true negatives	
		umber of non-relev	vant documents tha	t the
	egatives (TN): nu did not retrieve.	mber of non-relev	ant documents that	the
	negatives (FN): n t retrieve.	umber of relevant	documents that the	system

Defining	precision and i	recall		
		Relevant	Non-relevant	
	Retrieved	true positives	false positives	
	Not retrieved	false negatives	true negatives	
Precision				
		$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$		
Recall				
		$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$		

hf1-DA 2010–2011	III: 11 / 89
Comparing 2 IR systems — exam	nple
Document collection with 130 documer 28 documents relevant for a given theor	
System 1: retrieves 25 documents, 16 or $TP_1 = 16$ , $FP_1 = 25 - 16 = 9$ , F	
$P_1 = rac{ ext{TP}_1}{ ext{TP}_1 +  ext{FP}_1} = rac{16}{25} = 0.64$	$R_1 = rac{ ext{TP}_1}{ ext{TP}_1 +  ext{FN}_1} = rac{16}{28} = 0.57$
System 2: retrieves 15 documents, 12 or TP <sub>2</sub> = 12, FP <sub>2</sub> = $15 - 12 = 3$ , F	
$P_2 = rac{ ext{TP}_2}{ ext{TP}_2 +  ext{FP}_2} = rac{12}{15} = 0.80$	$R_2 = rac{ ext{TP}_2}{ ext{TP}_2 +  ext{FN}_2} = rac{12}{28} = 0.43$
N.B. System 2 has higher precision. Sy	stem 1 has higher recall.
art III: Unstructured Data	III.1: Unstructured data and data retrieval



Inf1-DA 2010-2011

III: 13 / 89

## F-score

The *F*-score is an evaluation measure that combines precision and recall.

$$F_{\alpha} = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$

Here  $\alpha$  is a *weighting factor* with  $0 \le \alpha \le 1$ .

High  $\alpha$  means precision more important. Low  $\alpha$  means recall is more important.

Often  $\alpha = 0.5$  is used, giving the *harmonic mean* of *P* and *R*:

$$F_{0.5}=rac{2PR}{P+R}$$

Part III: Unstructured Data

III.1: Unstructured data and data retrieval

nf1-DA 2010–2011	III: 14 / 89
Using F-score to compare — e	xample
We compare the examples on slide II $\alpha = 0.5$ ).	I: 11 using the F-score (with
$F_{0.5}(\text{System}_1) = \frac{2P_1R_1}{P_1 + R_1}$	$=rac{2 imes 0.64 imes 0.57}{0.64+0.57}=0.60$
$F_{0.5}(\text{System}_2) = rac{2P_2R_2}{P_2 + R_2}$	$=rac{2 imes 0.80 imes 0.43}{0.80+0.43}=0.56$
The F-score (with this weighting) rat	es System 1 as better than System 2.
art III: Unstructured Data	III.1: Unstructured data and data retrieval