Microsoft Research Silicon Valley

Data Analytics for Electronic Textbooks

Rakesh Agrawal

Microsoft Technical Fellow

Joint work with Members of Search Labs, Microsoft Research

December 24, 2012

Microsoft Research

International Conference on Big Data Analytics

Thesis

Education generally, and electronic books particularly, provide huge opportunity to expand and grow data mining research

Outline

- Importance of electronic textbooks
- Enriching textbooks through data mining
- Research opportunities
- Concluding thoughts

Microsoft Research

Outline

- Importance of electronic textbooks
- Enriching textbooks through data mining
- Research opportunities
- Concluding thoughts

Three Trends

- Centrality of good educational material for economic development
- Increased adoption of Internet in developing countries
- Emergence of tablets/e-readers

Education and Textbooks

Education: Primary vehicle for improving economic well-being of people

- *World Bank Reports, 1998, 2007*

Textbooks: Most cost-effective means of positively impacting educational quality

- Also indispensable for fostering teacher learning and for their ongoing professional development
- *Works by Clarke, Crossley, Fuller, Hanushek, Lockheed, Murby, Vail, and others*

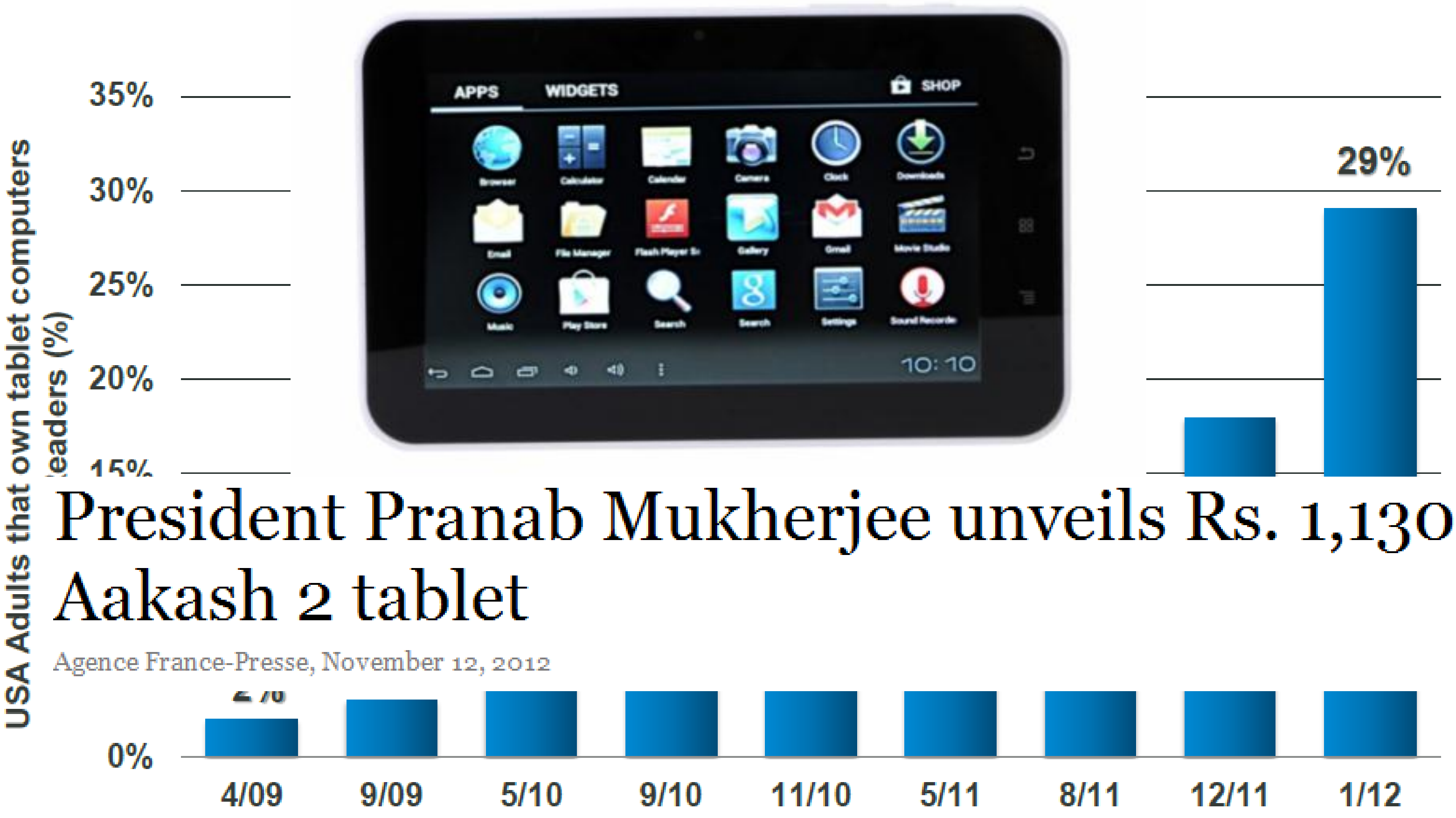
Microsoft Research

2.3B Global Internet Users in 2011* – 8% Growth*, Driven by Emerging Markets

Rank	Country	2008-2011 Internet User Adds (MMs)	2011 Internet Users (MMs)	Y/Y Growth	Populatio Penetratio
1	China	215	513	12%	38%
2	India	69	121	38	10
3	Indonesia	37	55	22	23
4	Philippines	28	34	44	35
5	Nigeria	21	45	--*	28
6	Mexico	19	42	19	37
7	Russia	16	61	3	43
8	USA	15	245	1	79
9	Iran	14	37	--*	48
10	Turkey	11	36	26	49
	Top 10	444	1,189	12%	32%
	World	663	2,250	8%	32%

Growth in Tablets/e-Readers:

29% US adult owners from 2% in less than 3 years



President Pranab Mukherjee unveils Rs. 1,130 Aakash 2 tablet

Agence France-Presse, November 12, 2012

Source: Pew Research Center, 1/12.

Microsoft Research

Reimagining Textbooks

ORGANIZATIONAL COMMUNICATION
STATISTICS for Business and Economics
Words
Basic Chemistry
USED
MILLER
MANAGEMENT ACCOUNTING
Electron Microscopy Principles and Techniques for Biologists
NEUROSCIENCE Exploring the Brain
UNDERSTANDING NUTRITION

Outline

- Importance of electronic textbooks
- Enriching textbooks through data mining
- Research opportunities
- Concluding thoughts

Microsoft® Research

Data Mining for Enriching Textbooks

Textbooks

Diagnostic tool to identify deficiencies

Identify sections needing enrichment

Decision model based on syntactic complexity of writing and dispersion of key concepts in the section [AGK+11]

Quantify comprehension burden on the reader

Assess the extent to which the textbook follows desirable properties and measure the comprehension burden due to non-sequential presentation of concepts [ACG+12]

Augmenting book sections with web content

Add selective links to articles

Determine key concepts in each section of a book and find links to authoritative web articles for these concepts [AGK+11]

Add selective images

Find images most relevant for a section factoring in images in other sections [AGK+11]

Vetted on high school textbooks (NCERT, India)
- Grades IX – XII
- Across subjects

[AGK+11] Data Mining for improving Textbooks, SIGKDD Explorations,13(2), 2011 (summary of 3 earlier papers).
[ACG+12] Empowering Authors to Diagnose Comprehension Burden in Textbooks, KDD 2012.

Data Mining for Enriching Textbooks

Textbooks

Diagnostic tool to identify deficiencies

Identify sections needing enrichment

Decision model based on syntactic complexity of writing and dispersion of key concepts in the section [AGK+11]

Quantify comprehension burden on the reader

Assess the extent to which the textbook follows desirable properties and measure the comprehension burden due to non-sequential presentation of concepts [ACG+12]

Augmenting book sections with web content

Add selective links to articles

Determine key concepts in each section of a book and find links to authoritative web articles for these concepts [AGK+11]

Add selective images

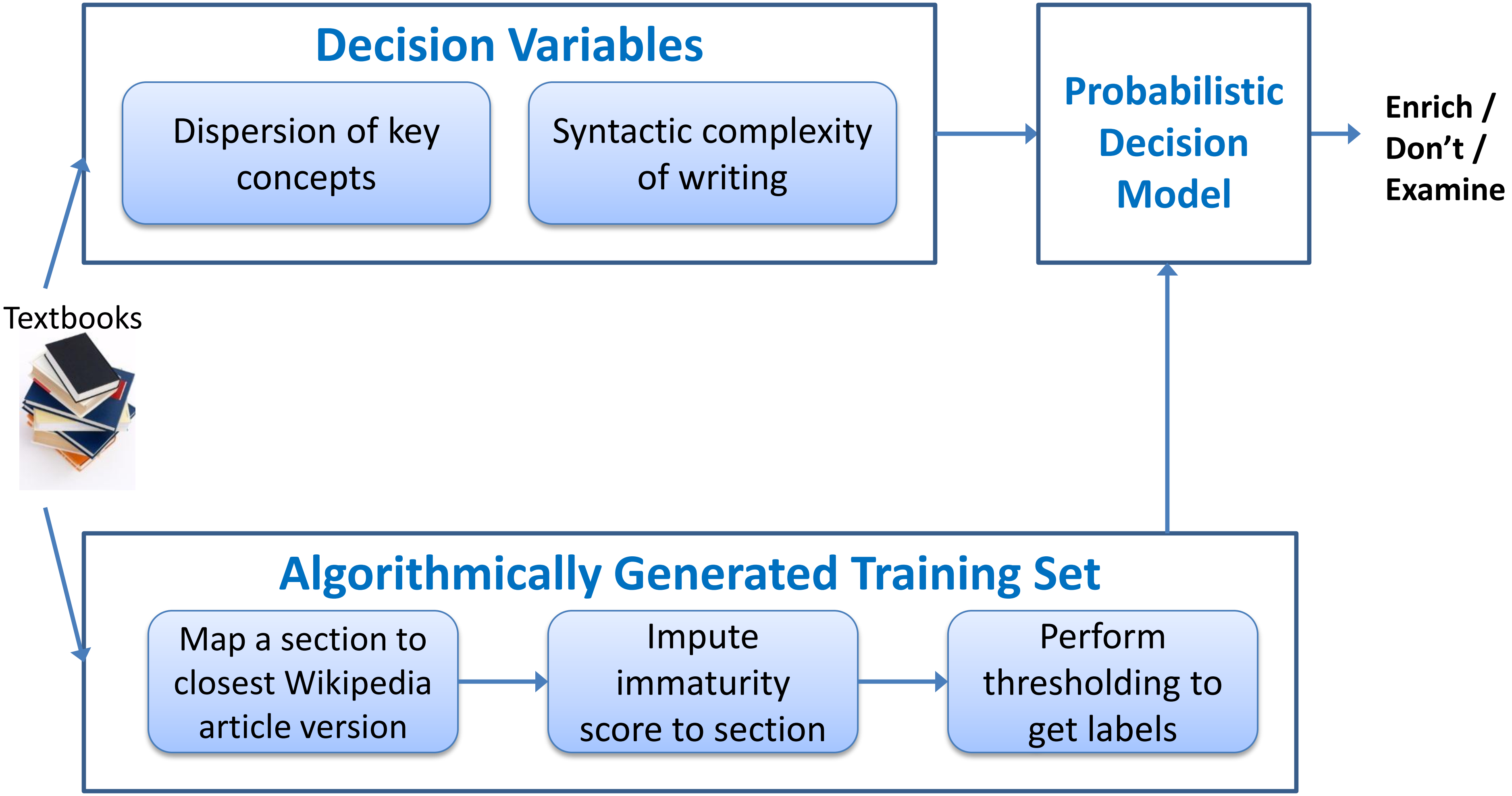
Find images most relevant for a section factoring in images in other sections [AGK+11]

Vetted on high school textbooks (NCERT, India)

- Grades IX – XII
- Across subjects

[AGK+11] Data Mining for improving Textbooks, SIGKDD Explorations,13(2), 2011 (summary of 3 earlier papers).
[ACG+12] Empowering Authors to Diagnose Comprehension Burden in Textbooks, KDD 2012.

Sections Needing Enrichment



Decision Variables

Dispersion of key concepts

Syntactic complexity of writing

Probabilistic Decision Model

Enrich / Don't / Examine

Textbooks

Algorithmically Generated Training Set

Map a section to closest Wikipedia article version

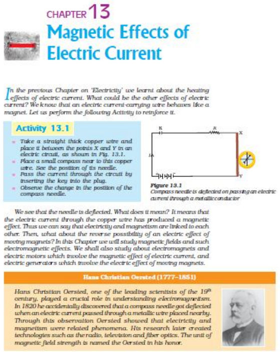
Impute immaturity score to section

Perform thresholding to get labels

Decision Variables

Dispersion of key concepts

Syntactic complexity of writing



CHAPTER 13

Magnetic Effects of Electric Current

In the previous Chapter on 'Electricity' we learnt about the heating effects of electric current. What could be the other effects of electric current? We know that an electric current-carrying wire behaves like a magnet. Let us perform the following Activity to reinforce it.

Activity 13.1

- Take a straight thick copper wire and place it between the points X and Y in an electric circuit, as shown in Fig. 13.1.
- Place a small compass near to this copper wire. See the position of its needle.
- Pass the current through the circuit by inserting the key into the plug.
- Observe the change in the position of the compass needle.

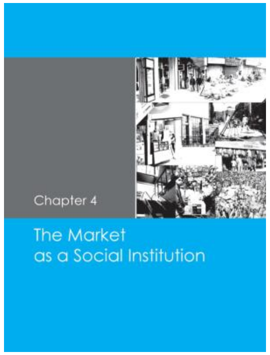
Figure 13.1
Compass needle is deflected on passing an electric current through a metallic conductor

We see that the needle is deflected. What does it mean? It means that the electric current through the copper wire has produced a magnetic effect. Thus we can say that electricity and magnetism are linked to each other. Then, what about the reverse possibility of an electric effect of moving magnets? In this Chapter we will study magnetic fields and such electromagnetic effects. We shall also study about electromagnets and electric motors which involve the magnetic effect of electric current, and electric generators which involve the electric effect of moving magnets.

Hans Christian Oersted (1777–1851)

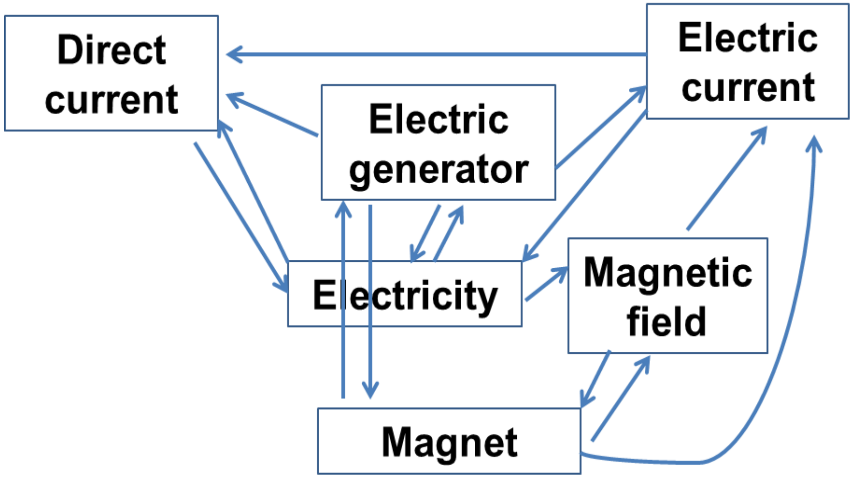
Hans Christian Oersted, one of the leading scientists of the 19th century, played a crucial role in understanding electromagnetism. In 1820 he accidentally discovered that a compass needle got deflected when an electric current passed through a metallic wire placed nearby. Through this observation Oersted showed that electricity and magnetism were related phenomena. His research later created technologies such as the radio, television and fiber optics. The unit of magnetic field strength is named the Oersted in his honor.

Many unrelated concepts in a section → Hard to understand section



Chapter 4

The Market as a Social Institution



Direct current

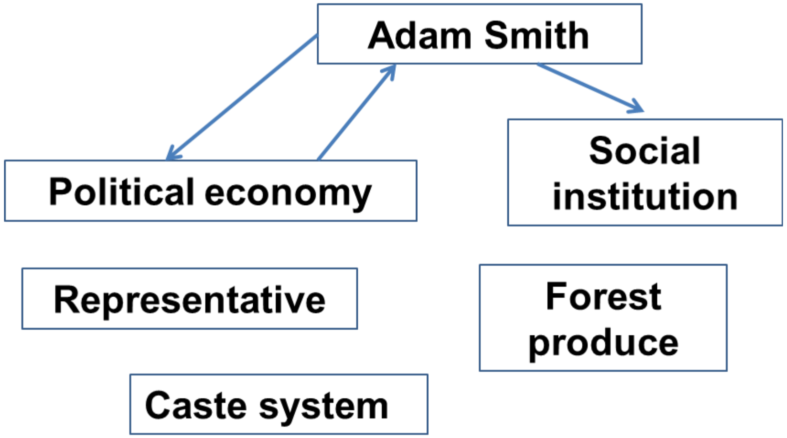
Electric current

Electric generator

Electricity

Magnetic field

Magnet



Adam Smith

Political economy

Social institution

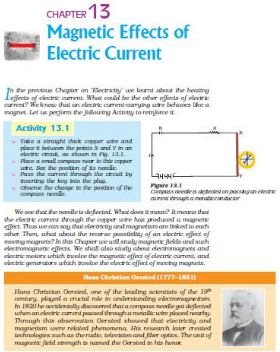
Representative

Forest produce

Caste system

Computing Dispersion

- *V* = set of key concepts discussed in section *s*
  - *Terminological noun phrases:* Linguistic pattern A*N^{+} (A: adjective; N: noun)
  - *"concepti" Wikipedia titles*
- *Related*(*x,y*) = Concept *x* is related to concept *y*
  - *Co-occurrence*
  - *true* if Wikipedia article for *x* links to the article for *y*
- Dispersion(*s*) := Fraction of unrelated concept pairs
  - (1 – Edge Density) of the concept graph



CHAPTER 13

Magnetic Effects of Electric Current

In the previous Chapter on 'Electricity' we learnt about the heating effects of electric current. What could be the other effects of electric current? We know that an electric current-carrying wire behaves like a magnet. Let us perform the following Activity to reinforce it.

Activity 13.1

- Take a straight thick copper wire and place it between the points X and Y in an electric circuit, as shown in Fig. 13.1.
- Place a small compass near to this copper wire. See the position of its needle.
- Pass the current through the circuit by inserting the key into the plug.
- Observe the change in the position of the compass needle.

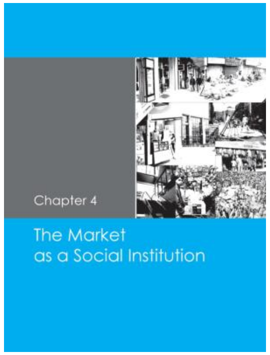
Figure 13.1
Compass needle is deflected on passing an electric current through a metallic conductor

We see that the needle is deflected. What does it mean? It means that the electric current through the copper wire has produced a magnetic effect. Thus we can say that electricity and magnetism are linked to each other. Then, what about the reverse possibility of an electric effect of moving magnets? In this Chapter we will study magnetic fields and such electromagnetic effects. We shall also study about electromagnets and electric motors which involve the magnetic effect of electric current, and electric generators which involve the electric effect of moving magnets.

Hans Christian Oersted (1777–1851)

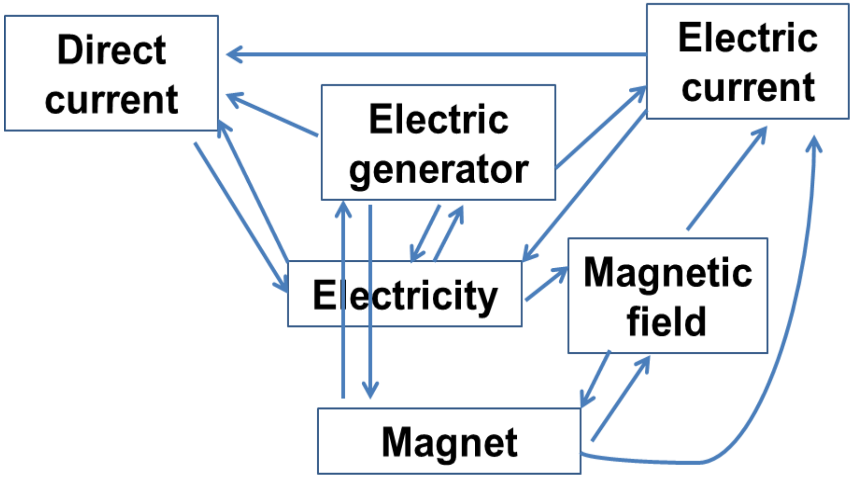
Hans Christian Oersted, one of the leading scientists of the 19th century, played a crucial role in understanding electromagnetism. In 1820 he accidentally discovered that a compass needle got deflected when an electric current passed through a metallic wire placed nearby. Through this observation Oersted showed that electricity and magnetism were related phenomena. His research later created technologies such as the radio, television and fiber optics. The unit of magnetic field strength is named the Oersted in his honor.

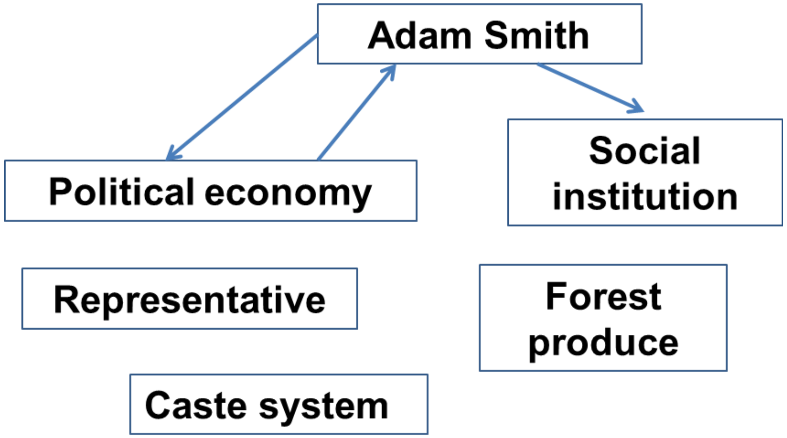
Many unrelated concepts in a section ➔ Hard to understand section



Chapter 4

The Market as a Social Institution

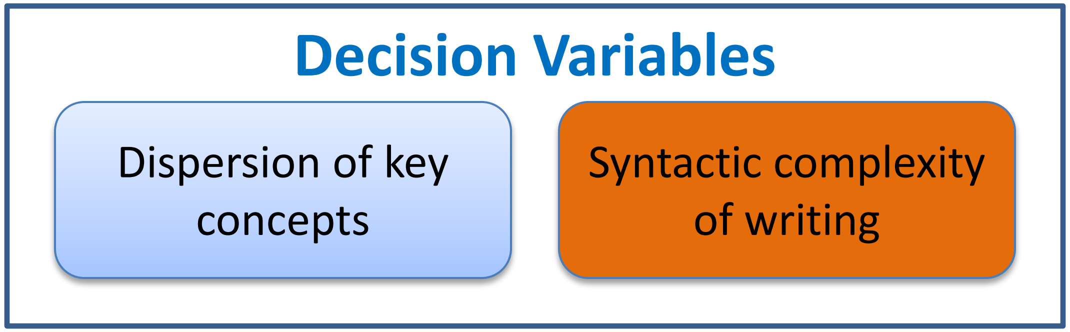




Dispersion = 1 – 15/30 = 0.5

Dispersion = 1 – 3/30 = 0.9

Larger dispersion ➔ greater need for augmentation

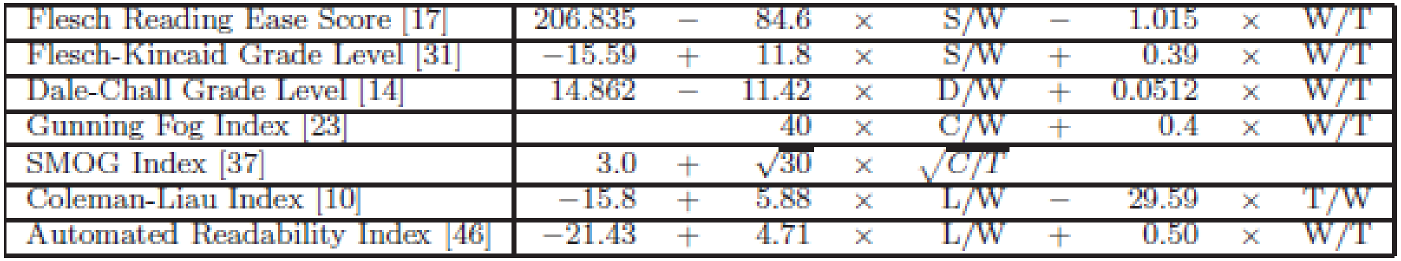


Decision Variables

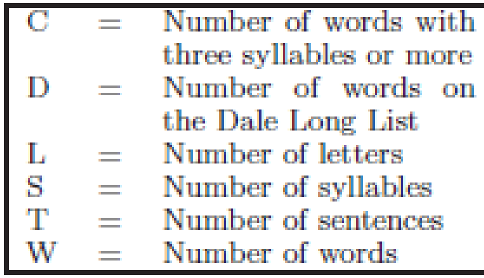
Dispersion of key concepts

Syntactic complexity of writing

- 100+ years of readability research
- 200+ Readability formulas
  - In widespread use (notwithstanding limitations)
- Popular formulas:

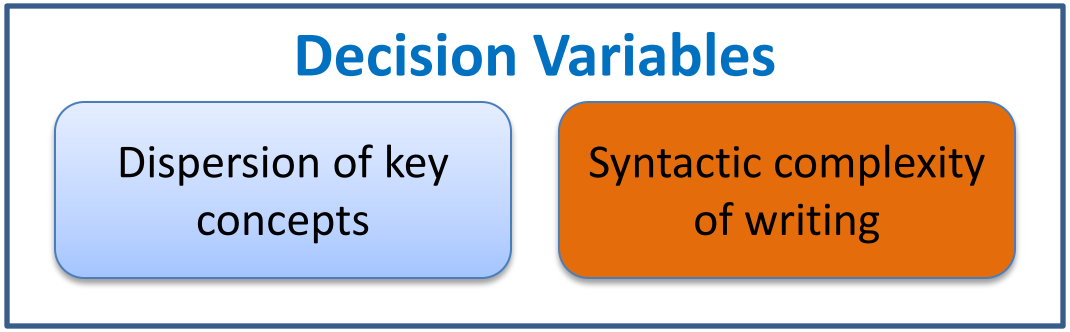


Flesch Reading Ease Score [17]	206.835	−	84.6	×	S/W	−	1.015	×	W/T
Flesch-Kincaid Grade Level [31]	−15.59	+	11.8	×	S/W	+	0.39	×	W/T
Dale-Chall Grade Level [14]	14.862	−	11.42	×	D/W	+	0.0512	×	W/T
Gunning Fog Index [23]			40	×	C/W	+	0.4	×	W/T
SMOG Index [37]	3.0	+	$\sqrt{30}$	×	$\sqrt{C/T}$				
Coleman-Liau Index [10]	−15.8	+	5.88	×	L/W	−	29.59	×	T/W
Automated Readability Index [46]	−21.43	+	4.71	×	L/W	+	0.50	×	W/T



C	=	Number of words with three syllables or more
D	=	Number of words on the Dale Long List
L	=	Number of letters
S	=	Number of syllables
T	=	Number of sentences
W	=	Number of words

- Regression coefficients learned over specific datasets
  - McCall-Crabbs Standard Test Lessons



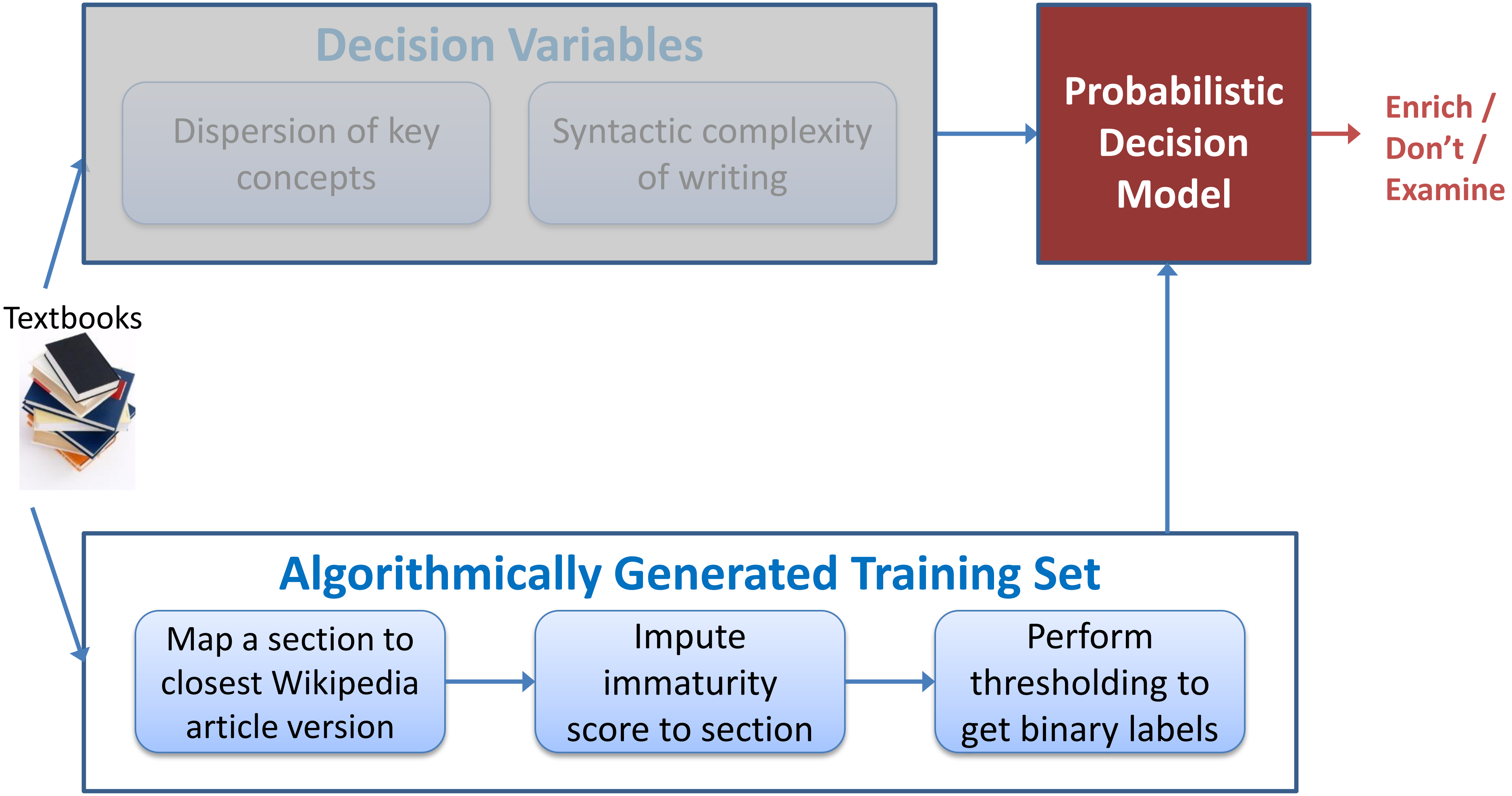
Decision Variables

Dispersion of key concepts

Syntactic complexity of writing

- Direct use of *Readability formulas* yielded poor results
- Variables abstracted from readability formulas:
  - Word length: Average syllables per word (S/W)
  - Sentence length: Average words per sentence (W/T)
- Larger syntactic complexity ➔ greater need for augmentation

System Overview



Decision Variables

Dispersion of key concepts

Syntactic complexity of writing

Probabilistic Decision Model

Enrich / Don't / Examine

Textbooks

Algorithmically Generated Training Set

Map a section to closest Wikipedia article version

Impute immaturity score to section

Perform thresholding to get binary labels

Probabilistic Decision Model

- Probabilistic scoring of a section needing enrichment through logistic regression
- Probability that a section needs enrichment

$$P(y = 1|\mathbf{z}, \mathbf{w}) = \frac{1}{1 + \exp\{-(b + \mathbf{z}^T\mathbf{w})\}}.$$

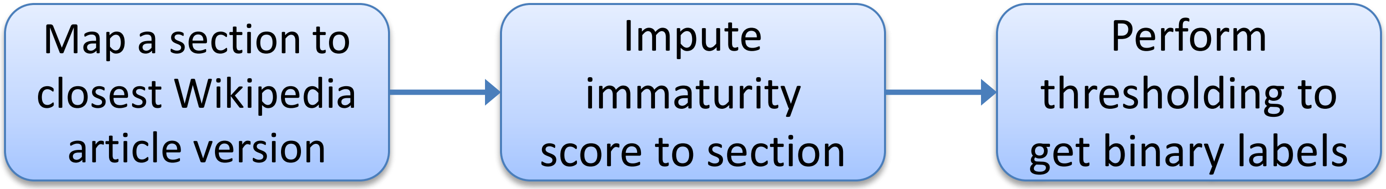
Section needing enrichment

Decision variables

Importance between decision variables

- Optimal weight vector **w** learned from a training set of textbook sections
- Scores binned into
  - "Enrich", "Don't enrich", or "Manually investigate to decide"

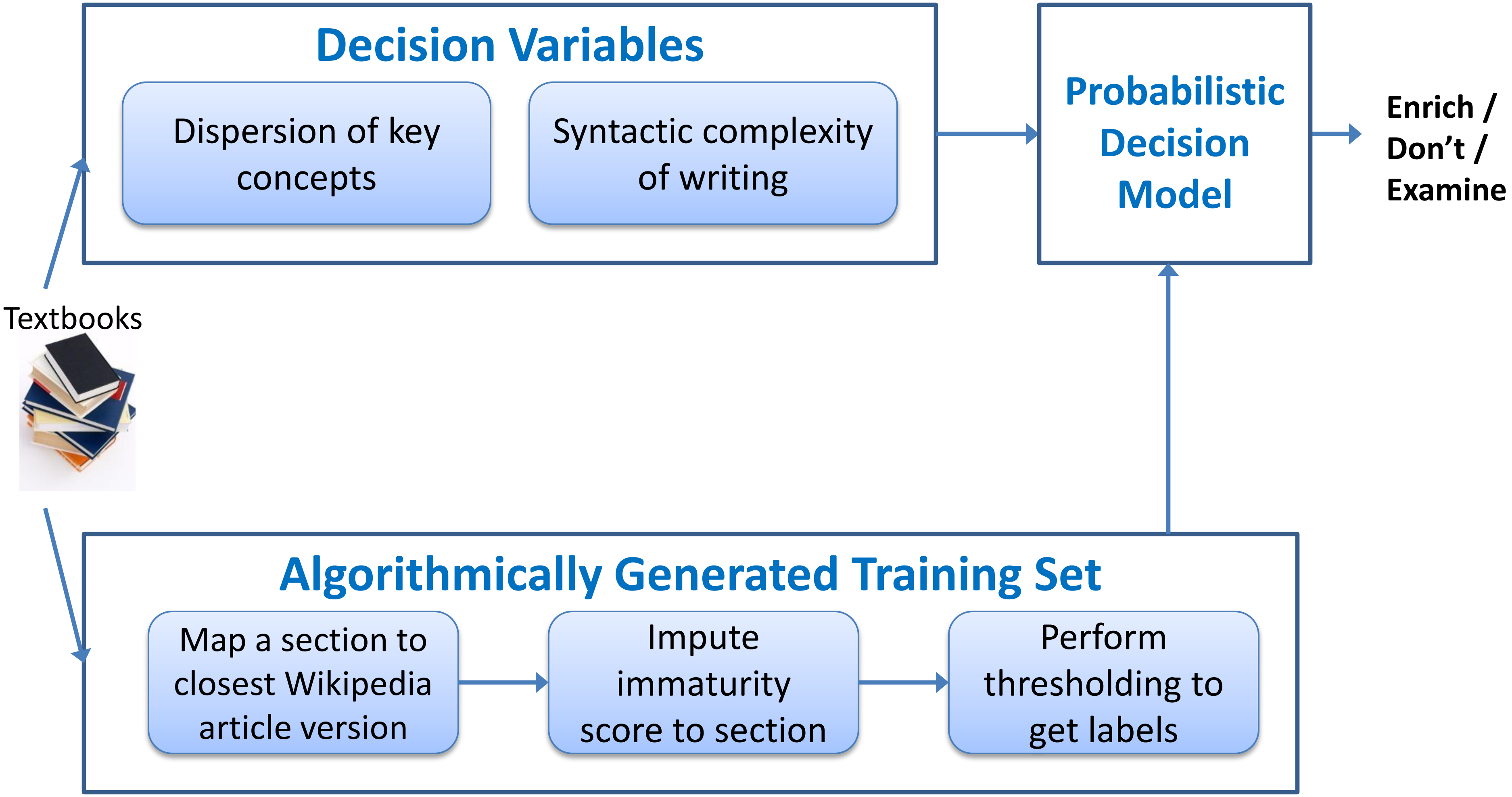
Algorithmically Generated Training Set


Map a section to closest Wikipedia article version
Impute immaturity score to section
Perform thresholding to get binary labels

- Difficult to get qualified judges who would give consistent labels
- Map a textbook section to a most similar version of a similar article in a versioned repository (Wikipedia)
- Compute immaturity of this version as a proxy for that of the section
- Immaturity: function of relative edits on each day and a time window K, with more weight to recent edits (see paper)
- Immaturity computation reliable at only extreme ends
    - But only few quality labels are needed

[AGK+11a] Identifying Enrichment Candidates in Textbooks. WWW 2011.

Sections Needing Enrichment



Decision Variables

Dispersion of key concepts

Syntactic complexity of writing

Probabilistic Decision Model

Enrich / Don’t / Examine

Textbooks

Algorithmically Generated Training Set

Map a section to closest Wikipedia article version

Impute immaturity score to section

Perform thresholding to get labels

Application to Indian Textbooks

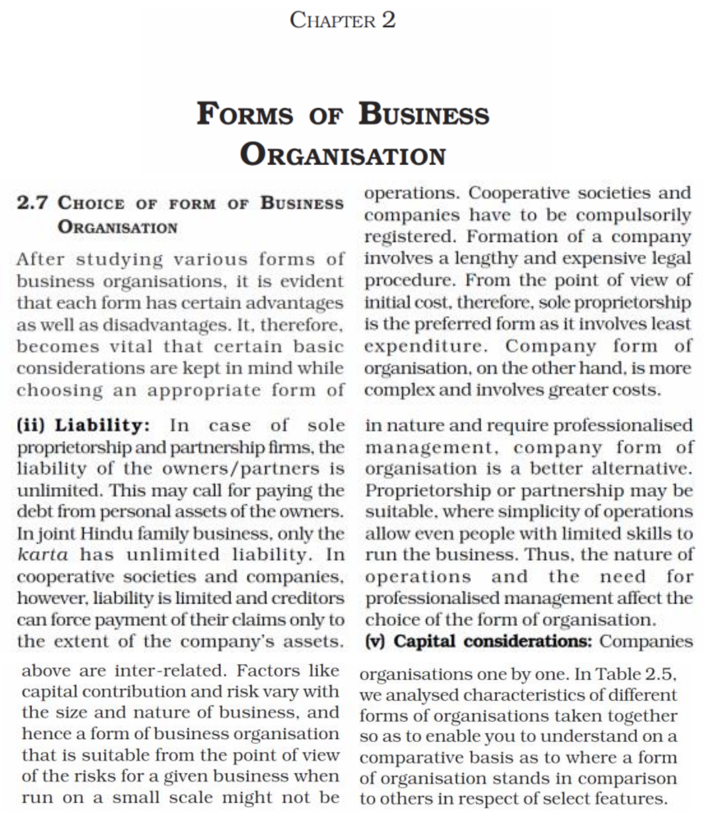
विद्यया ऽमृतमश्नुते

एन सी ई आर टी
NCERT

- Book corpus: 17 high school textbooks published by NCERT*
  - Grades IX – XII
  - Subject areas: Sciences, Social Sciences, Commerce, Math
  - 191 chapters, 1313 sections
- Followed by millions of students
- Available online

* National Council of Educational Research and Training

Results: Sections needing enrichment



CHAPTER 2

FORMS OF BUSINESS ORGANISATION

2.7 CHOICE OF FORM OF BUSINESS ORGANISATION

After studying various forms of business organisations, it is evident that each form has certain advantages as well as disadvantages. It, therefore, becomes vital that certain basic considerations are kept in mind while choosing an appropriate form of

(ii) Liability: In case of sole proprietorship and partnership firms, the liability of the owners/partners is unlimited. This may call for paying the debt from personal assets of the owners. In joint Hindu family business, only the *karta* has unlimited liability. In cooperative societies and companies, however, liability is limited and creditors can force payment of their claims only to the extent of the company's assets.

above are inter-related. Factors like capital contribution and risk vary with the size and nature of business, and hence a form of business organisation that is suitable from the point of view of the risks for a given business when run on a small scale might not be

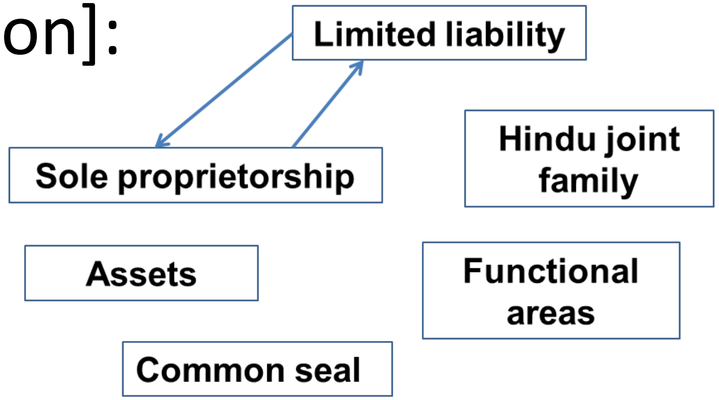
operations. Cooperative societies and companies have to be compulsorily registered. Formation of a company involves a lengthy and expensive legal procedure. From the point of view of initial cost, therefore, sole proprietorship is the preferred form as it involves least expenditure. Company form of organisation, on the other hand, is more complex and involves greater costs.

in nature and require professionalised management, company form of organisation is a better alternative. Proprietorship or partnership may be suitable, where simplicity of operations allow even people with limited skills to run the business. Thus, the nature of operations and the need for professionalised management affect the choice of the form of organisation.

(v) Capital considerations: Companies

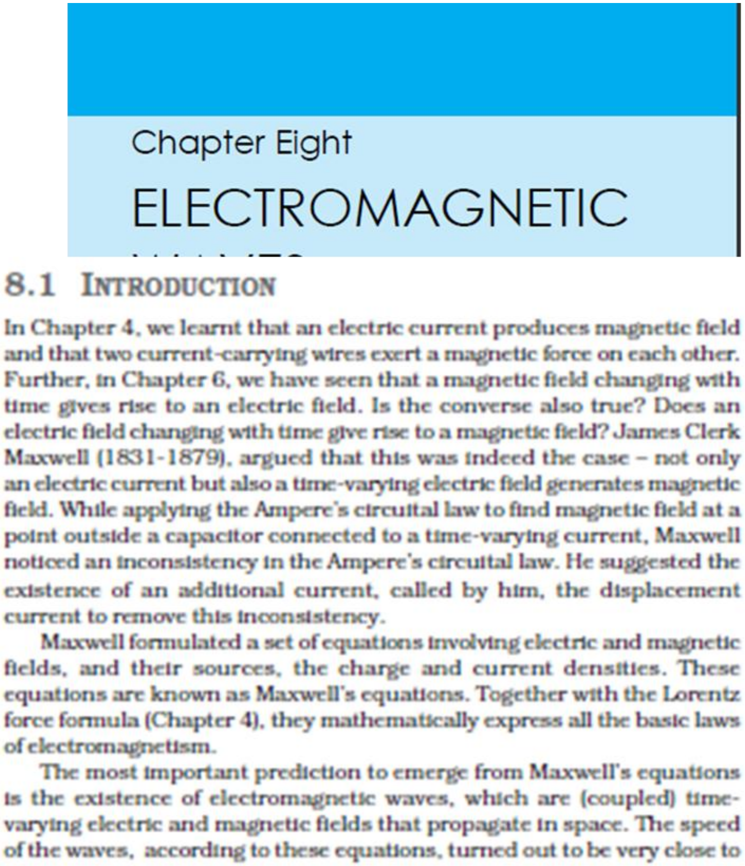
organisations one by one. In Table 2.5, we analysed characteristics of different forms of organisations taken together so as to enable you to understand on a comparative basis as to where a form of organisation stands in comparison to others in respect of select features.

- Many unrelated concepts [high dispersion]:



- Long sentences, e.g.,
  - *Factors like capital contribution and risk vary with the size and nature of business, and hence a form of business organisation that is suitable from the point of view of the risks for a given business when run on a small scale might not be appropriate when the same business is carried on a large scale.*

Results: Sections ***not*** needing enrichment



Chapter Eight

ELECTROMAGNETIC

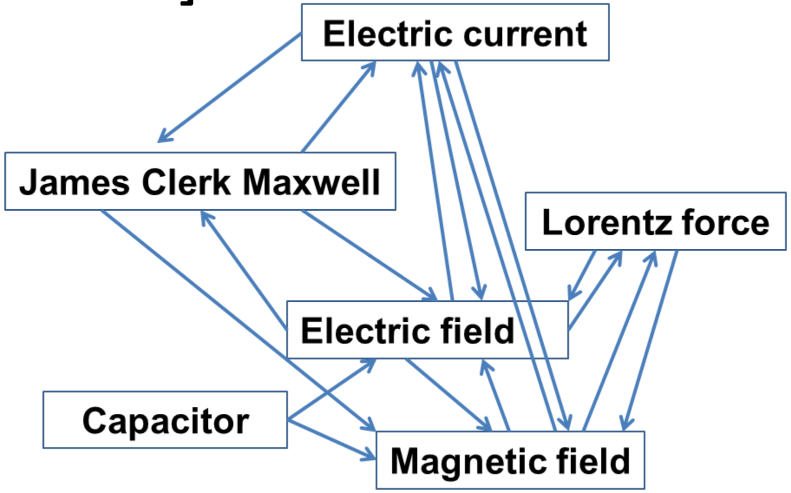
8.1 Introduction

In Chapter 4, we learnt that an electric current produces magnetic field and that two current-carrying wires exert a magnetic force on each other. Further, in Chapter 6, we have seen that a magnetic field changing with time gives rise to an electric field. Is the converse also true? Does an electric field changing with time give rise to a magnetic field? James Clerk Maxwell (1831-1879), argued that this was indeed the case – not only an electric current but also a time-varying electric field generates magnetic field. While applying the Ampere's circuital law to find magnetic field at a point outside a capacitor connected to a time-varying current, Maxwell noticed an inconsistency in the Ampere's circuital law. He suggested the existence of an additional current, called by him, the displacement current to remove this inconsistency.

Maxwell formulated a set of equations involving electric and magnetic fields, and their sources, the charge and current densities. These equations are known as Maxwell's equations. Together with the Lorentz force formula (Chapter 4), they mathematically express all the basic laws of electromagnetism.

The most important prediction to emerge from Maxwell's equations is the existence of electromagnetic waves, which are (coupled) time-varying electric and magnetic fields that propagate in space. The speed of the waves, according to these equations, turned out to be very close to

- Highly related concepts [low dispersion]:



Electric current

James Clerk Maxwell

Lorentz force

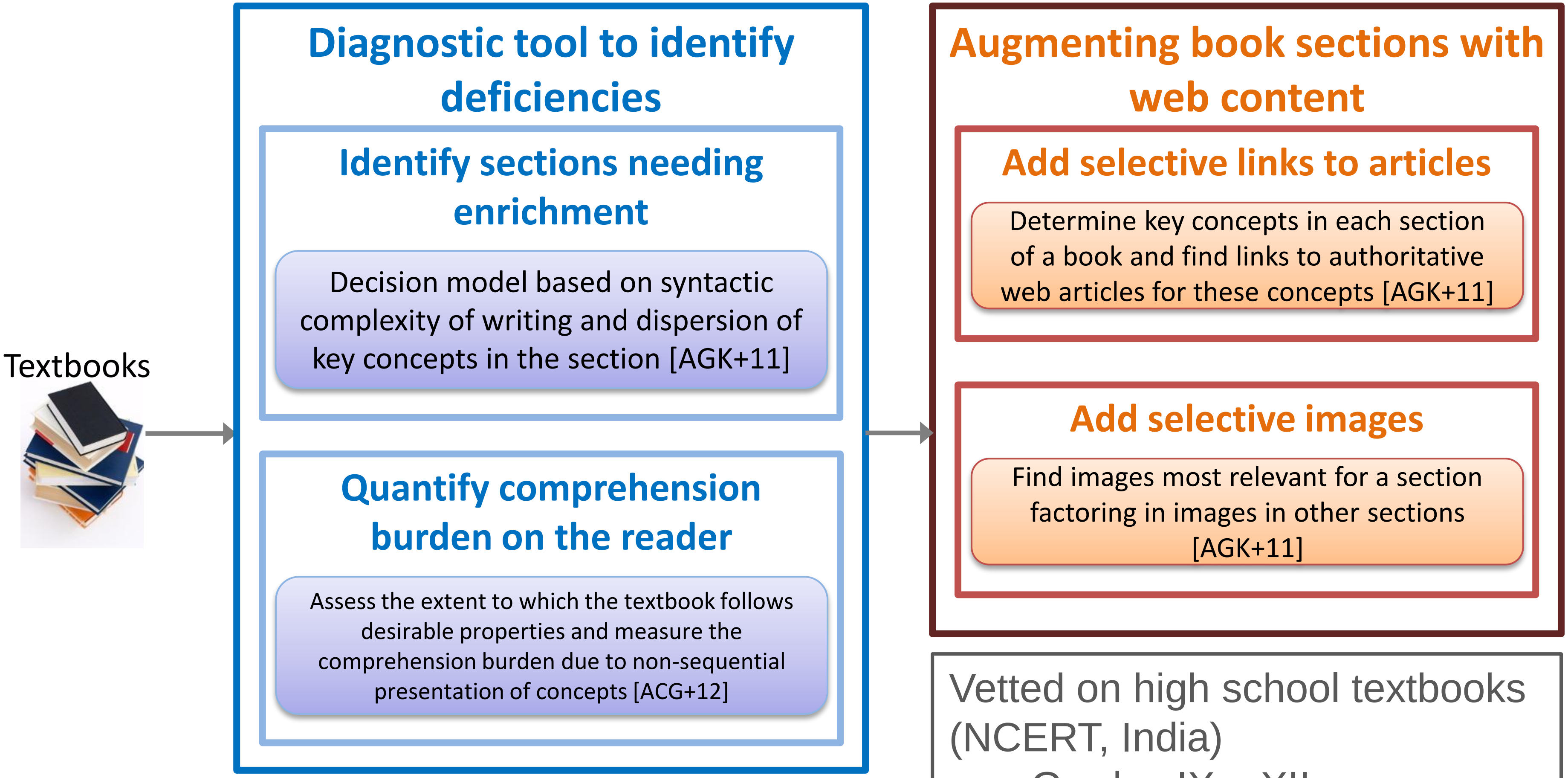
Electric field

Capacitor

Magnetic field

- Written clearly with simple sentences [low syntactic complexity]

Data Mining for Enriching Textbooks



Textbooks

Diagnostic tool to identify deficiencies

Identify sections needing enrichment

Decision model based on syntactic complexity of writing and dispersion of key concepts in the section [AGK+11]

Quantify comprehension burden on the reader

Assess the extent to which the textbook follows desirable properties and measure the comprehension burden due to non-sequential presentation of concepts [ACG+12]

Augmenting book sections with web content

Add selective links to articles

Determine key concepts in each section of a book and find links to authoritative web articles for these concepts [AGK+11]

Add selective images

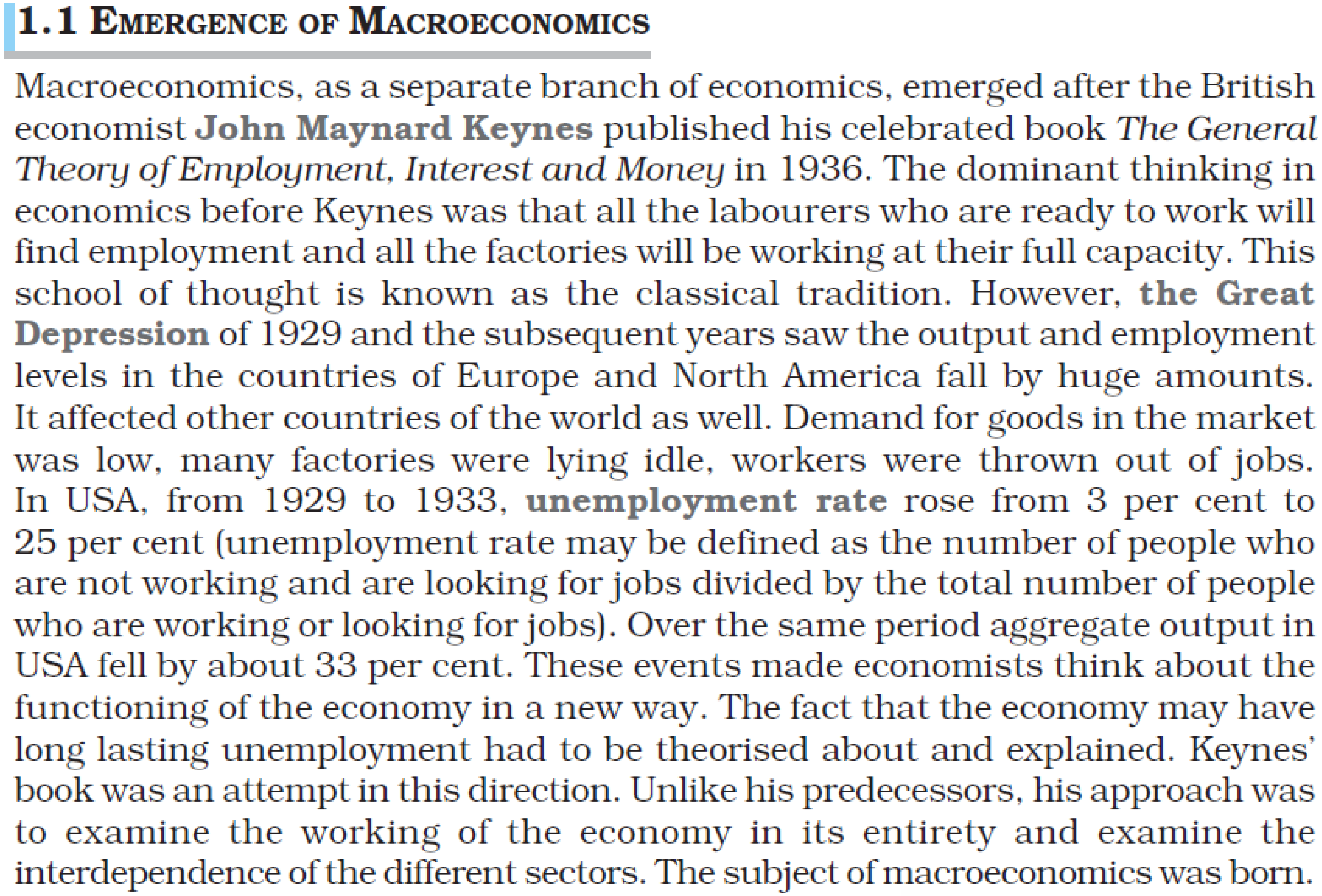
Find images most relevant for a section factoring in images in other sections [AGK+11]

Vetted on high school textbooks (NCERT, India)

- Grades IX – XII
- Across subjects

[AGK+11] Data Mining for improving Textbooks, SIGKDD Explorations,13(2), 2011 (summary of 3 earlier papers).
[ACG+12] Empowering Authors to Diagnose Comprehension Burden in Textbooks, KDD 2012.

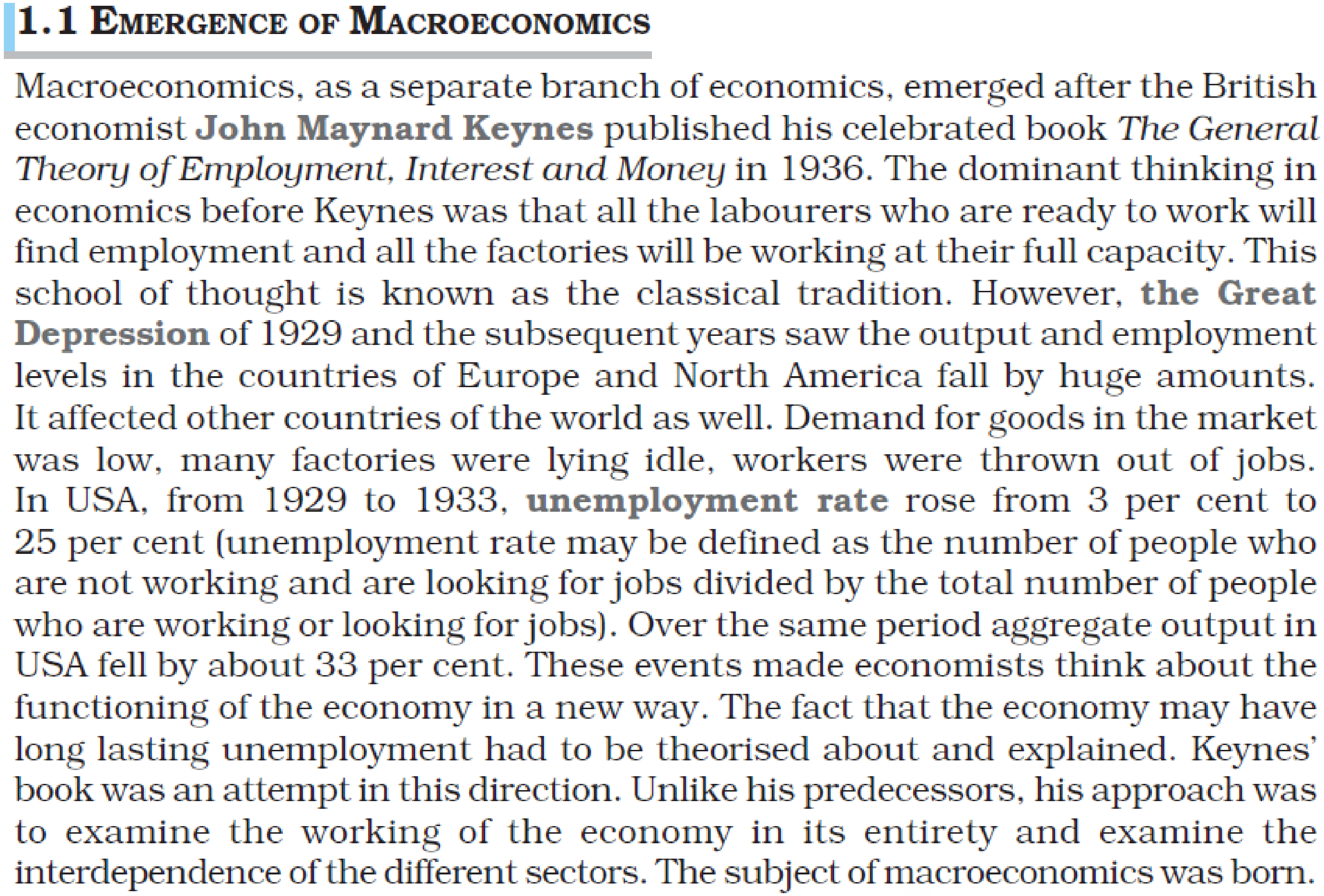
A section from an Economics Textbook



1.1 Emergence of Macroeconomics

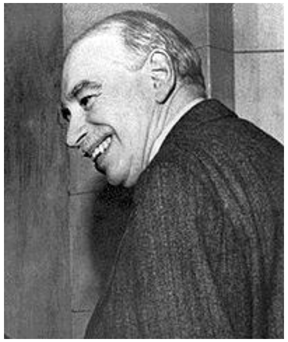
Macroeconomics, as a separate branch of economics, emerged after the British economist **John Maynard Keynes** published his celebrated book *The General Theory of Employment, Interest and Money* in 1936. The dominant thinking in economics before Keynes was that all the labourers who are ready to work will find employment and all the factories will be working at their full capacity. This school of thought is known as the classical tradition. However, **the Great Depression** of 1929 and the subsequent years saw the output and employment levels in the countries of Europe and North America fall by huge amounts. It affected other countries of the world as well. Demand for goods in the market was low, many factories were lying idle, workers were thrown out of jobs. In USA, from 1929 to 1933, **unemployment rate** rose from 3 per cent to 25 per cent (unemployment rate may be defined as the number of people who are not working and are looking for jobs divided by the total number of people who are working or looking for jobs). Over the same period aggregate output in USA fell by about 33 per cent. These events made economists think about the functioning of the economy in a new way. The fact that the economy may have long lasting unemployment had to be theorised about and explained. Keynes' book was an attempt in this direction. Unlike his predecessors, his approach was to examine the working of the economy in its entirety and examine the interdependence of the different sectors. The subject of macroeconomics was born.

Augmented Section

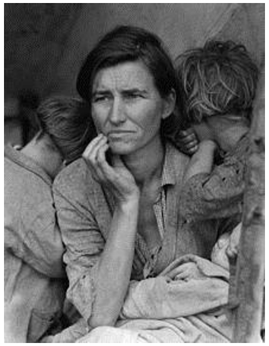


1.1 Emergence of Macroeconomics

Macroeconomics, as a separate branch of economics, emerged after the British economist **John Maynard Keynes** published his celebrated book *The General Theory of Employment, Interest and Money* in 1936. The dominant thinking in economics before Keynes was that all the labourers who are ready to work will find employment and all the factories will be working at their full capacity. This school of thought is known as the classical tradition. However, **the Great Depression** of 1929 and the subsequent years saw the output and employment levels in the countries of Europe and North America fall by huge amounts. It affected other countries of the world as well. Demand for goods in the market was low, many factories were lying idle, workers were thrown out of jobs. In USA, from 1929 to 1933, **unemployment rate** rose from 3 per cent to 25 per cent (unemployment rate may be defined as the number of people who are not working and are looking for jobs divided by the total number of people who are working or looking for jobs). Over the same period aggregate output in USA fell by about 33 per cent. These events made economists think about the functioning of the economy in a new way. The fact that the economy may have long lasting unemployment had to be theorised about and explained. Keynes' book was an attempt in this direction. Unlike his predecessors, his approach was to examine the working of the economy in its entirety and examine the interdependence of the different sectors. The subject of macroeconomics was born.

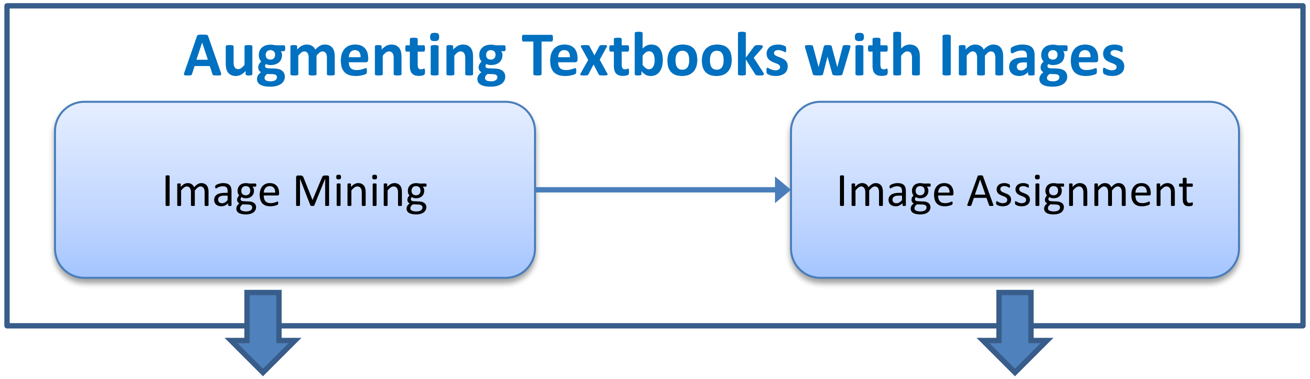


John Maynard Keynes



The Great Depression formed the backdrop against which Keynes's revolution took place. The image is Dorothea Lange's Migrant Mother depiction of destitute pea-pickers in California, taken in March 1936.

Microsoft Research



Augmenting Textbooks with Images

Image Mining → Image Assignment

Obtain images relevant to each section using complementary methods

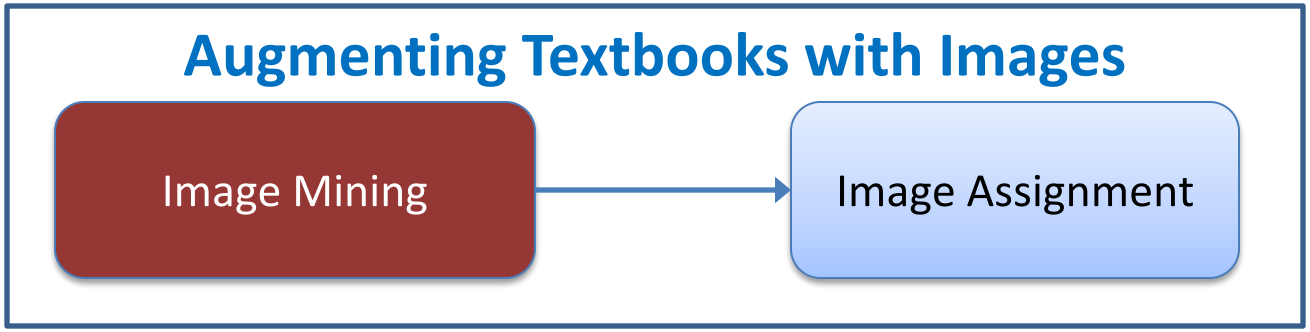
Comity: Leverage image search provided by search engines

Affinity: Leverage image metadata on webpages

Allocate most relevant images to each section such that

- Each section is augmented with at most *k* images
- No image repeats across sections

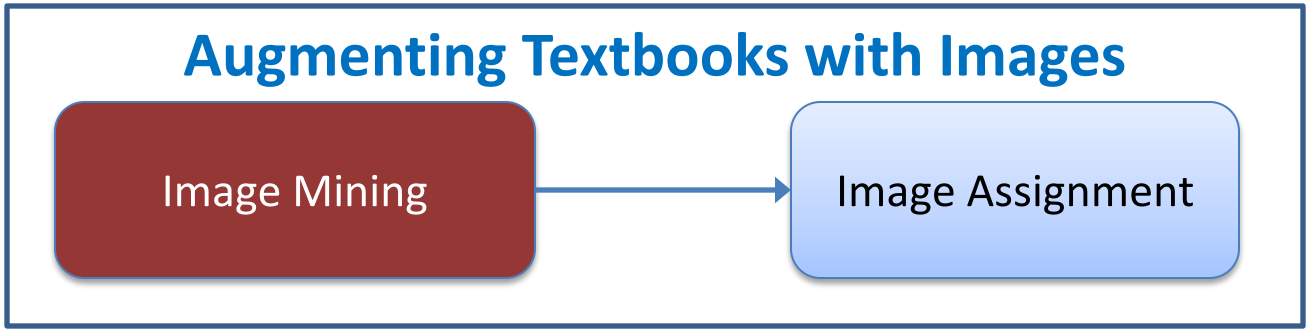
Comity



- Intuition: Combine results of a large number of short, but relevant queries
  - Search engines barf on long queries (such as entire section content)
- Identify key concepts present in a section, C
- Form two-concept and three-concept queries, Q
- For each $q \in Q$, obtain ranked list of images $I(q)$ using image search
- Relevance score(i) of image i =

$\sum_q f($position of image in $I(q)$, importance of concepts in $q)$

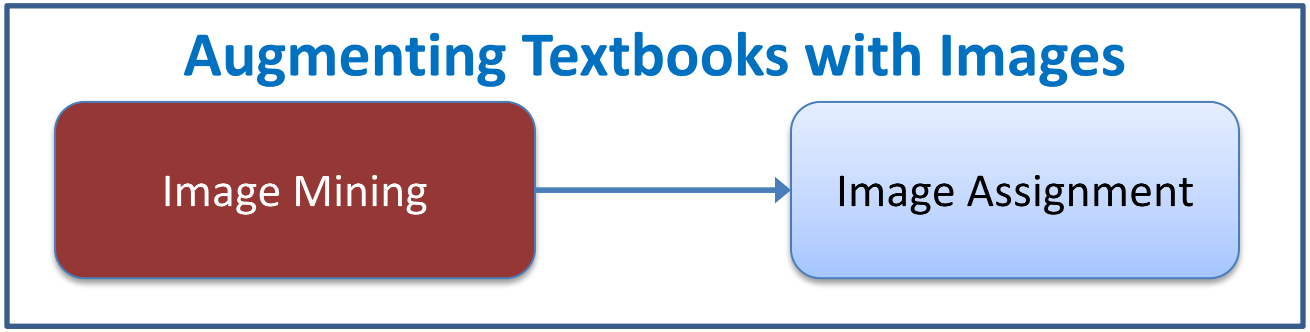
Affinity



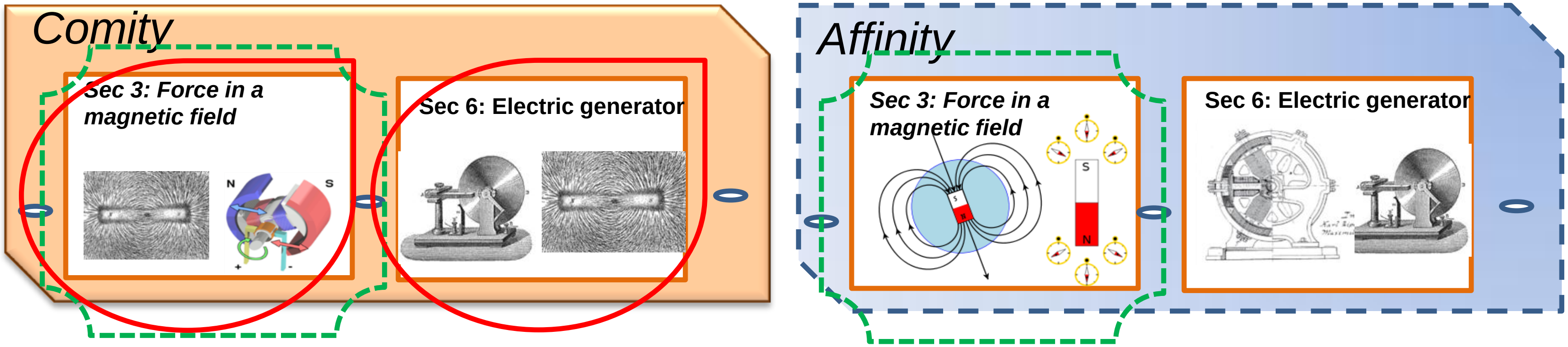
Augmenting Textbooks with Images

- Intuition: Authoritative pages contain authoritative images
- Identify top webpages that have high textual similarity with the given book section *s*
- Score each image *i* in these pages based on a similarity metric
  - Relevance(*i*, *s*) = *f*(metadata associated with the image *i*, key concepts in *s*)
  - Metadata: captions, content relevant to the image, etc.

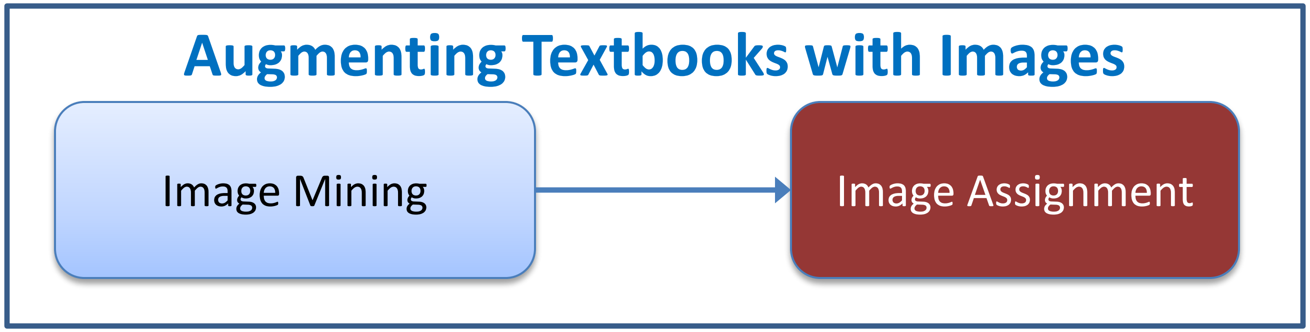
Augmenting Textbooks with Images



1. Complementary algorithms provide a broad selection of images for a section


Comity
Sec 3: Force in a magnetic field
Sec 6: Electric generator
Affinity
Sec 3: Force in a magnetic field
Sec 6: Electric generator

2. But images can repeat across sections because of independent mining at section level



MaxRelevantImageAssignment

$$\max \sum_{i \in I} \sum_{j \in S} x_{ij} \cdot \lambda_{ij}$$

Relevance score of image i to section j

Total relevance score for the chapter: sum of relevance scores of images assigned

s.t.

$$x_{ij} \in \{0,1\} \;\; \forall i \in I \forall j \in S$$

=1 if image i is selected for section j else 0

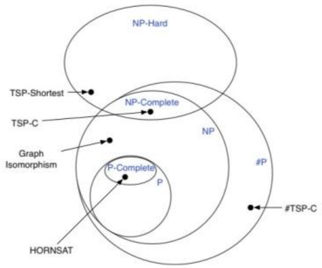
$$\sum_{i \in I} x_{ij} \leq K_j \;\; \forall j \in S$$

Constraint: At most K_j images can be assigned to section j

$$\sum_{j \in S} x_{ij} \leq 1 \;\; \forall i \in I$$

Constraint: An image can belong to at most one section

Solution Technique



MaxRelevantImageAssignment *can be solved optimally in polynomial time*

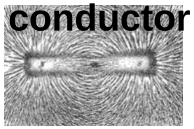
Proof: Follows from an efficient reduction to the Maximum Weighted Bipartite Matching problem

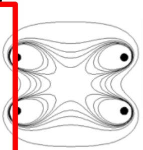
(Algorithm immediate from the proof)

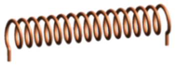
Value of Image Assignment

BEFORE IMAGE ASSIGNMENT

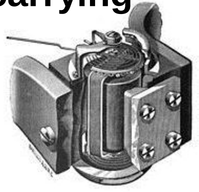
Sec 2: Magnetic field due to a current carrying conductor


Magnetic effect

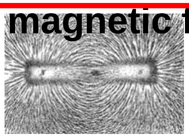

Helmholtz Contour

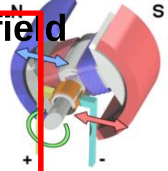

Solenoid

Amperemeter

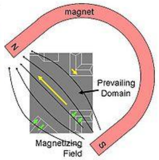

Galvanometer

Sec 3: Force on a current carrying conductor in a magnetic field

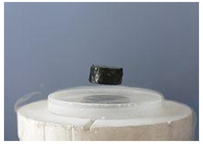

Magnetic effect



Electric motor cycle

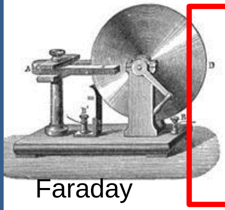


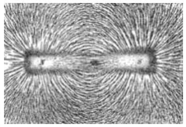
Effect of magnet on domains

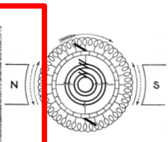

Meissner Effect

Descartes' magnetic field

Sec 6: Electric generator

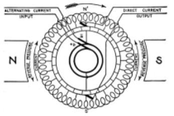

Faraday disk generator


Magnetic effect



Two phase rotary converter

Descartes' magnetic field

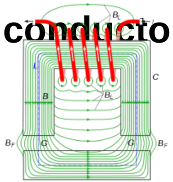


Single phase rotary converter

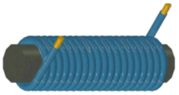
Same images repeat across sections!

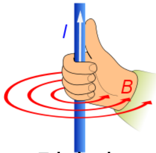
AFTER IMAGE ASSIGNMENT

Sec 2: Magnetic field due to a current carrying conductor

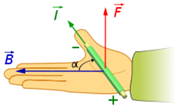


Magnetic field

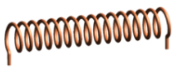

Simple electromagnet



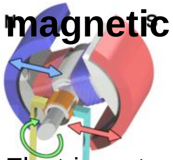
Right hand rule

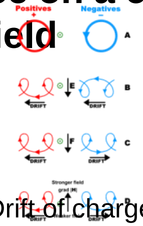


Right hand rule

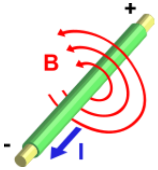

Solenoid

Sec 3: Force on a current carrying conductor in a magnetic field

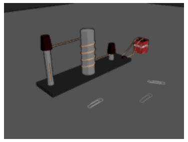

Electric-motor cycle exploits electro magnetism

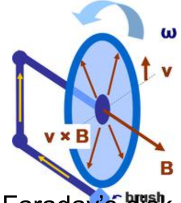


Drift of charged particles



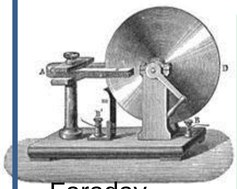
Magnetic field around current

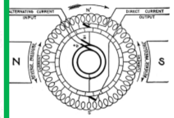

Electromagnets attract paper clips….



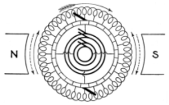
Faraday's disk electric generator

Sec 6: Electric generator

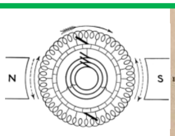

Faraday disk generator



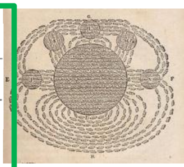
Single phase rotary converter



Two phase rotary converter



Three phase rotary converter


Descartes' magnetic field

Richer set of images to augment the section

Microsoft® Research

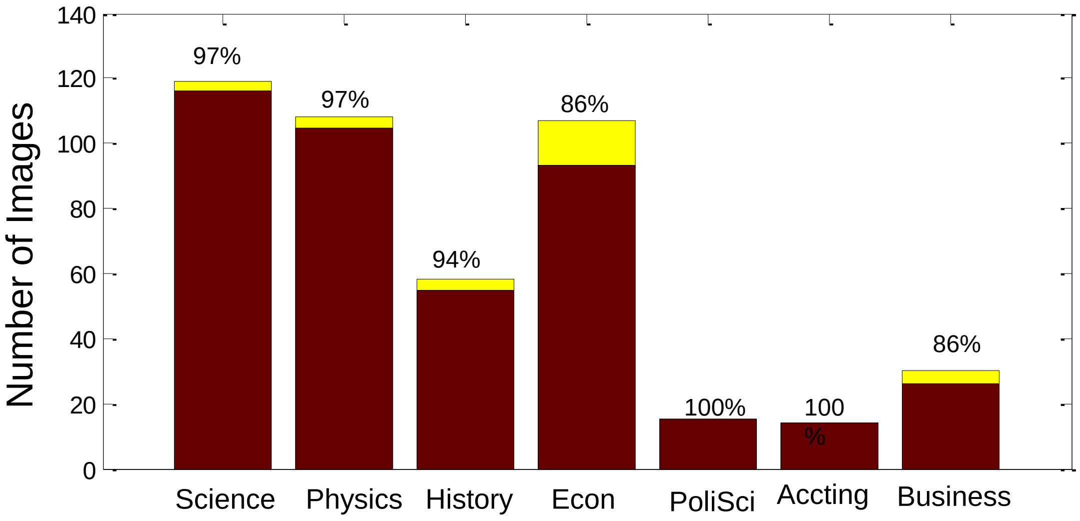
Evaluation on NCERT Textbooks

- User-study employing Amazon Mechanical Turk to judge the quality of results
- HIT (User task): A given image helpful for understanding the section?
- An image deemed helpful if the majority of 7 judges considered it so
- Helpfulness index:
  - Average of helpfulness score of the images over all sections

Agrawal et al. CIKM 2011.

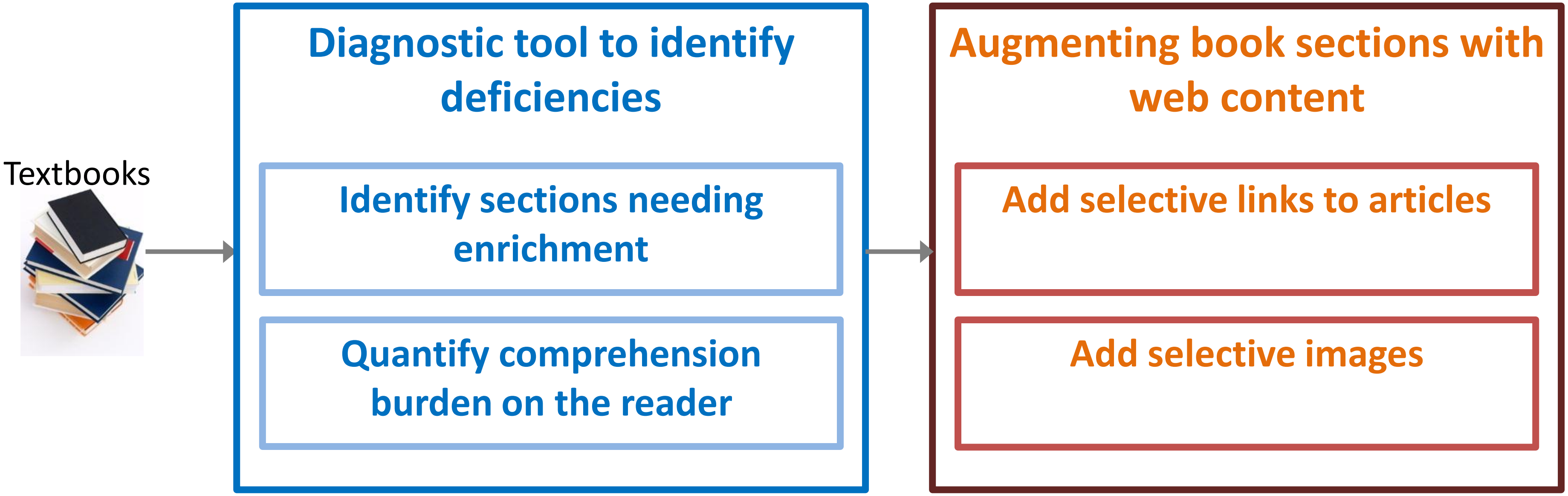
Performance

The number above a bar indicate helpfulness index for the corresponding subject (% of images found helpful)



- 94% of images deemed helpful
- Performance maintained across subjects

Recap



Textbooks

Diagnostic tool to identify deficiencies

Identify sections needing enrichment

Quantify comprehension burden on the reader

Augmenting book sections with web content

Add selective links to articles

Add selective images

- Technological solutions for
  - Diagnosing deficient sections
  - Mining and optimal placement of web objects (images & articles)
- Promising results over High School textbooks across subjects and grades

Outline

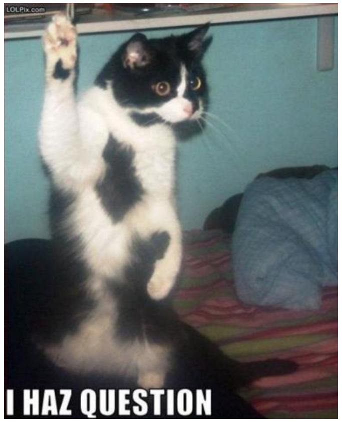
- Importance of electronic textbooks
- Enriching textbooks through data mining
- Research opportunities
- Concluding thoughts

Textbook Augmentation

- Deeper analysis to identify key concepts discussed in a section (Discourse analysis? Formal Concept Analysis?)
- Caption and placement of augmentations
- Extension to other multimedia types (video, audio)
- Modeling for "appropriateness" of augmentations

Microsoft® Research

Broader Questions



- Social networking centered around an electronic textbook
- Complementarity of algorithmic solutions to the crowdsourcing approaches
  - Tools for capturing feedback on textbooks (errors, better explanations, supplementary material, etc.)
  - Trust and ranking
- Deployment issues:
  - Social, behavioral, legal, cultural, policy, and political issues
  - Quantifying impact

Outline

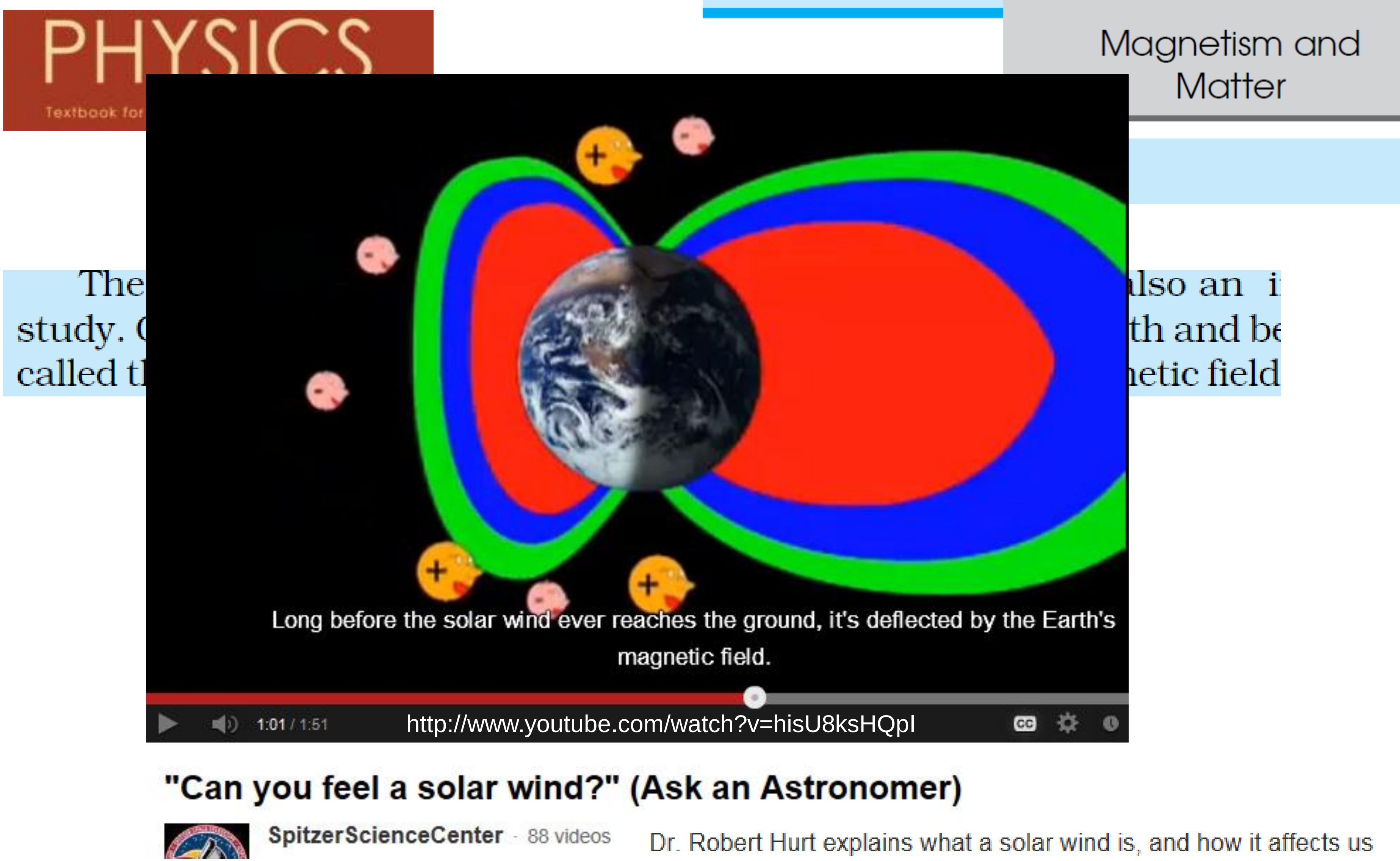
- Importance of electronic textbooks
- Enriching textbooks through data mining
- Research opportunities
- Concluding thoughts

Call to Action

- Thomas Friedman: “Big breakthroughs happen when what is possible meets what is desperately necessary”
- The stage is set for data community to help revolutionize education
- The present audience can (and should) play key role

Thank you!

Augmenting with Videos



PHYSICS

Textbook for

Magnetism and Matter

The … also an i… study. … th and be… called t… netic field

Long before the solar wind ever reaches the ground, it's deflected by the Earth's magnetic field.

1:01 / 1:51

http://www.youtube.com/watch?v=hisU8ksHQpI

"Can you feel a solar wind?" (Ask an Astronomer)

SpitzerScienceCenter · 88 videos

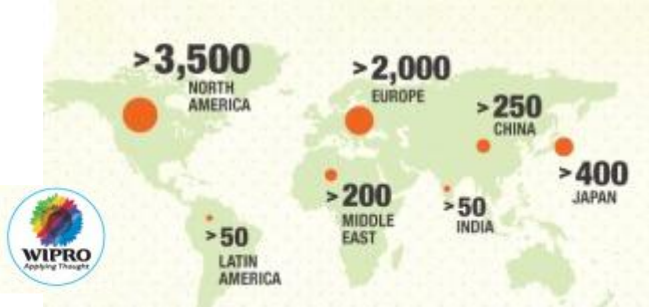
Dr. Robert Hurt explains what a solar wind is, and how it affects us here on Earth.

Improving Education

- Identification of ill-matched material
  - Test score = f (student ability, suitability of material)
  - Learning: Item Response Theory
- Collaborative translation and localization of educational material
- Analysis of new pedagogical approaches

Microsoft® Research

Defining Big Data: 3 V's



Data stored across the world in petabytes

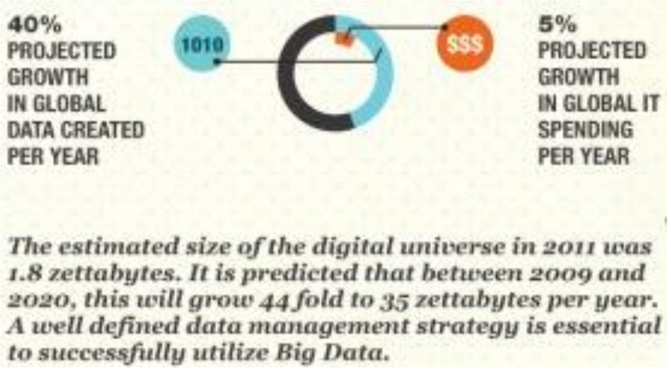
1 petabytes = 1 million gigabytes =
1,000 one-terabyte hard drives

Volume

Big Data

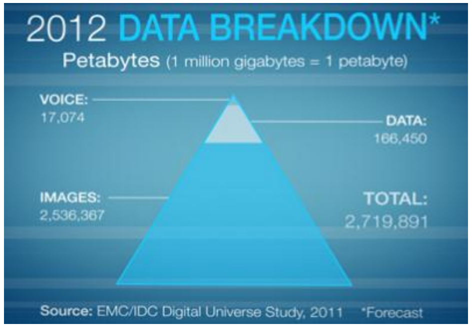
Variety

Velocity

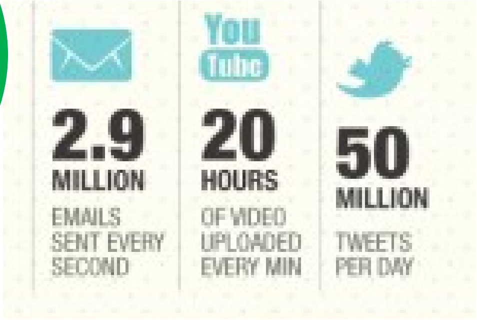


The estimated size of the digital universe in 2011 was 1.8 zettabytes. It is predicted that between 2009 and 2020, this will grow 44 fold to 35 zettabytes per year. A well defined data management strategy is essential to successfully utilize Big Data.

1 zettabytes = 1 million petabytes
(75 billion 16 GB Apple iPads)



Source: EMC/IDC Digital Universe Study, 2011 *Forecast



Doug Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Meta. Feb. 2001.