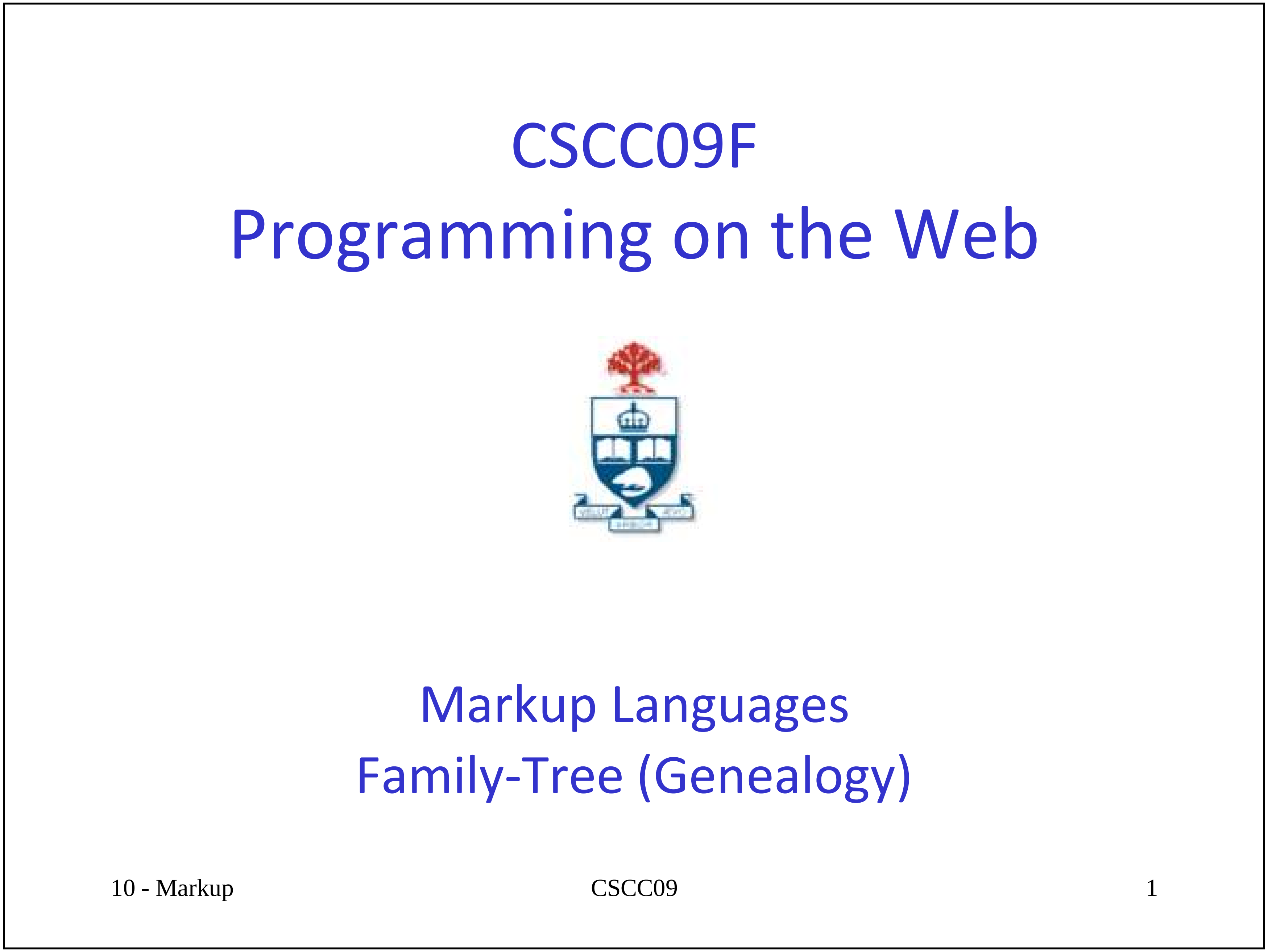
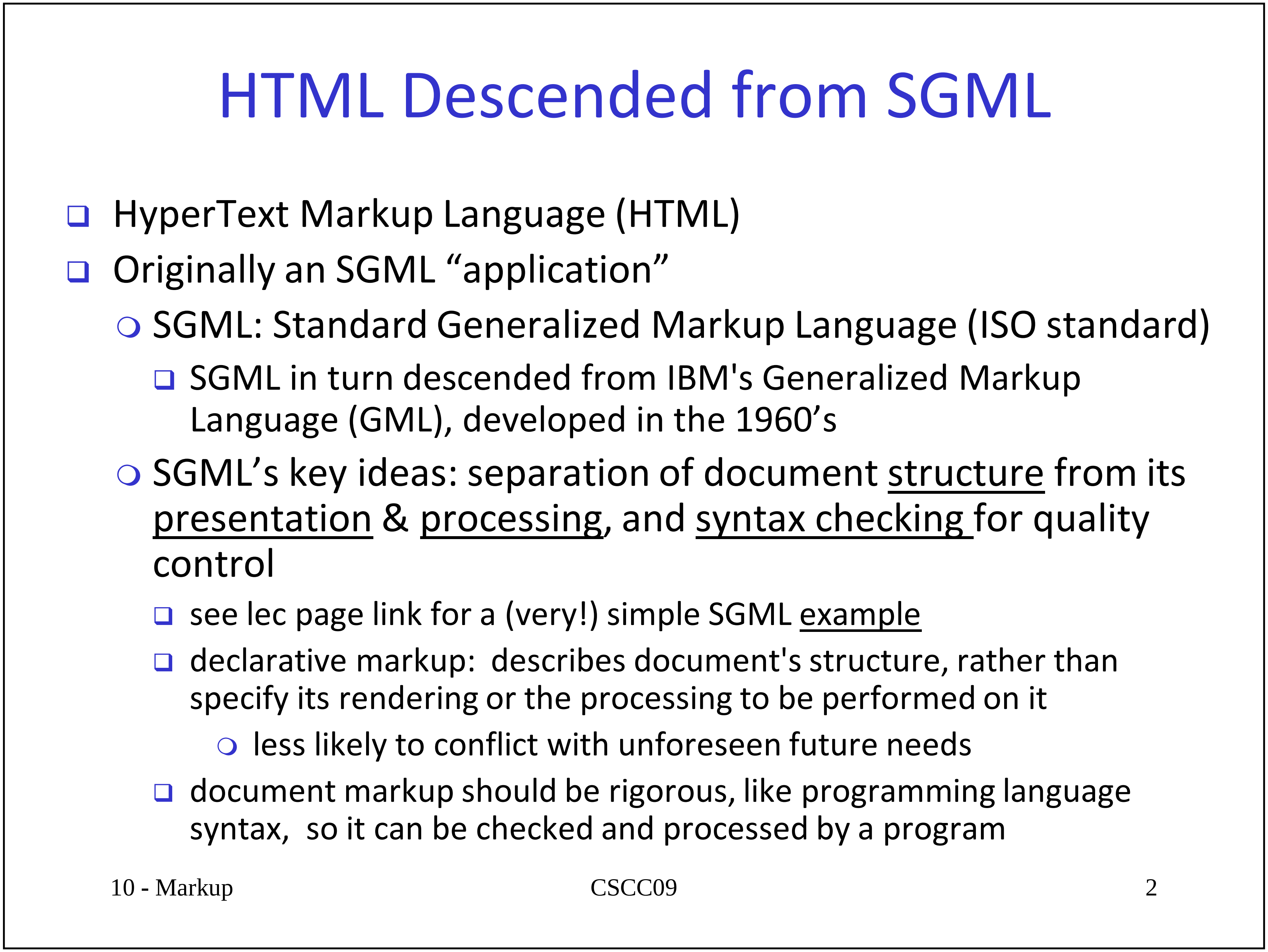


# CSCC09F
# Programming on the Web

## Markup Languages
## Family-Tree (Genealogy)

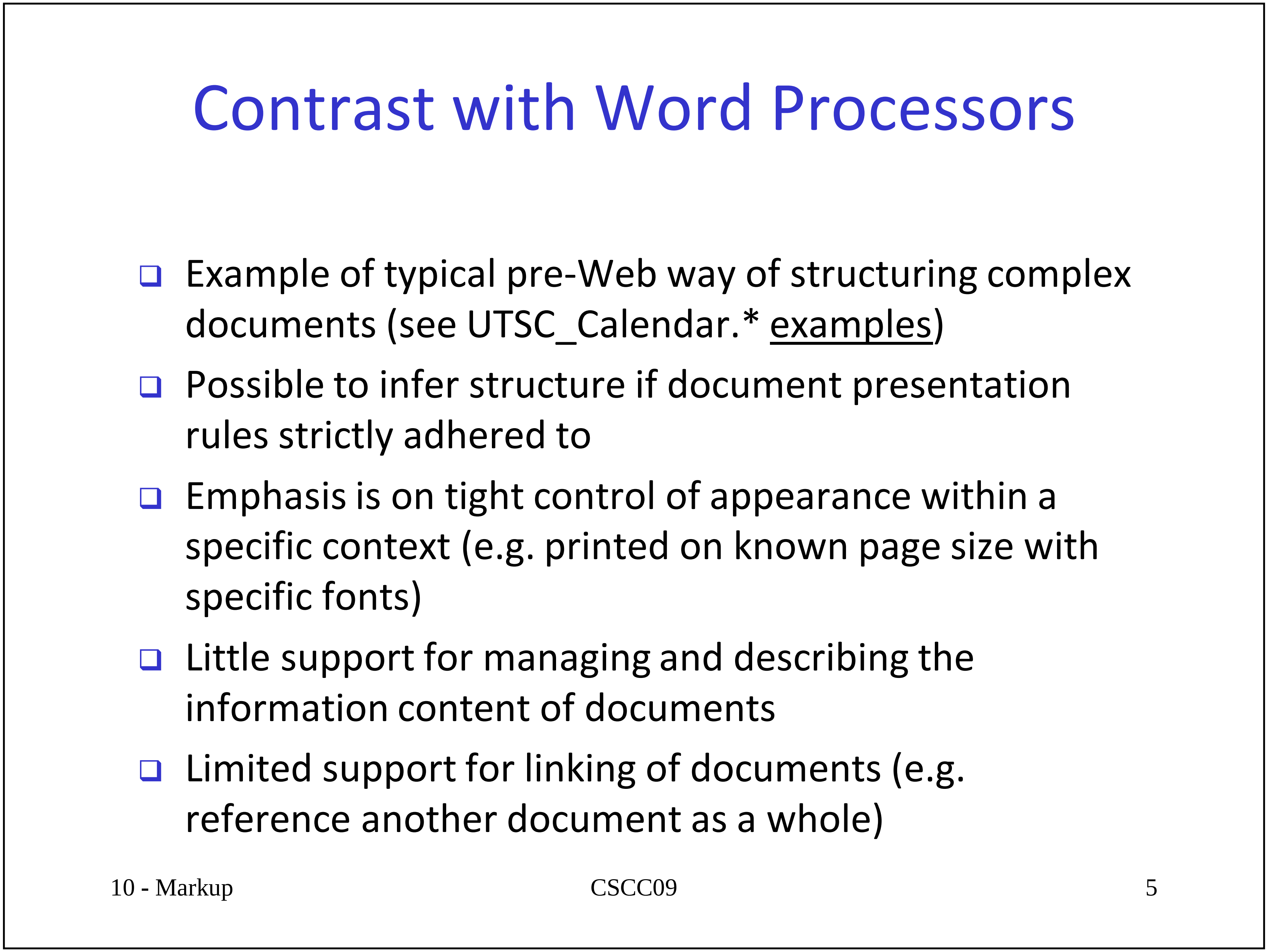


# HTML Descended from SGML

- HyperText Markup Language (HTML)
- Originally an SGML “application”
  - SGML: Standard Generalized Markup Language (ISO standard)
    - SGML in turn descended from IBM's Generalized Markup Language (GML), developed in the 1960’s
  - SGML’s key ideas: separation of document structure from its presentation & processing, and syntax checking for quality control
    - see lec page link for a (very!) simple SGML example
    - declarative markup: describes document's structure, rather than specify its rendering or the processing to be performed on it
      - less likely to conflict with unforeseen future needs
    - document markup should be rigorous, like programming language syntax, so it can be checked and processed by a program

# HTML: Content Structure not Presentation

- HTML was created as an application of SGML specifically intended to markup text documents
  - a compromise in the interest of simplicity
- As originally conceived, HTML was to be a language for the exchange of scientific and technical documents, suitable for use by non-document specialists.
- HTML addressed the problem of SGML complexity by specifying a small set of structural and semantic tags suitable for authoring relatively simple documents
- In addition to simplifying the document structure, HTML crucially added support for hypertext – enabling documents to be “linked” together into a Web

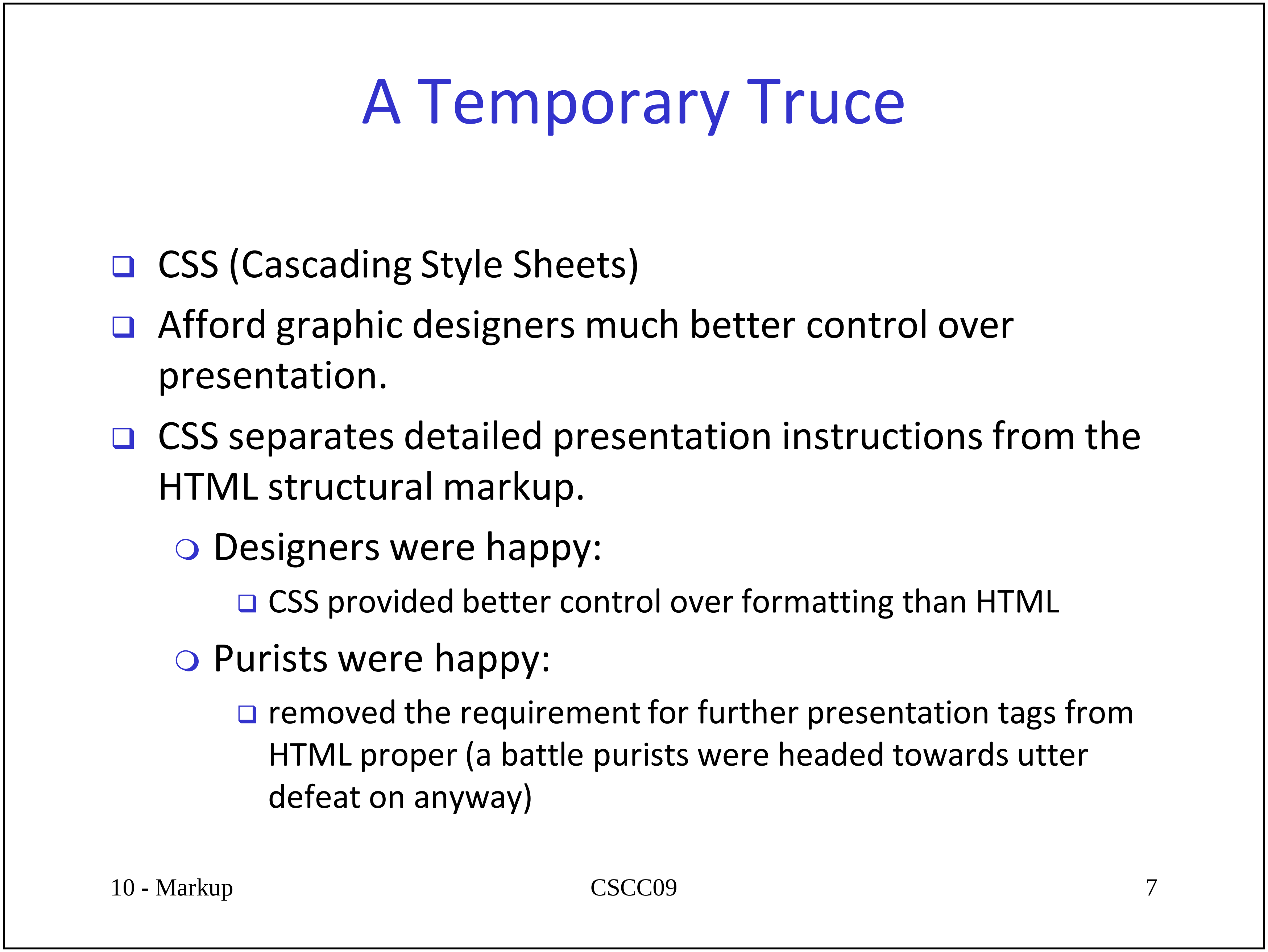


# Contrast with Word Processors

- Example of typical pre-Web way of structuring complex documents (see UTSC_Calendar.* examples)
- Possible to infer structure if document presentation rules strictly adhered to
- Emphasis is on tight control of appearance within a specific context (e.g. printed on known page size with specific fonts)
- Little support for managing and describing the information content of documents
- Limited support for linking of documents (e.g. reference another document as a whole)

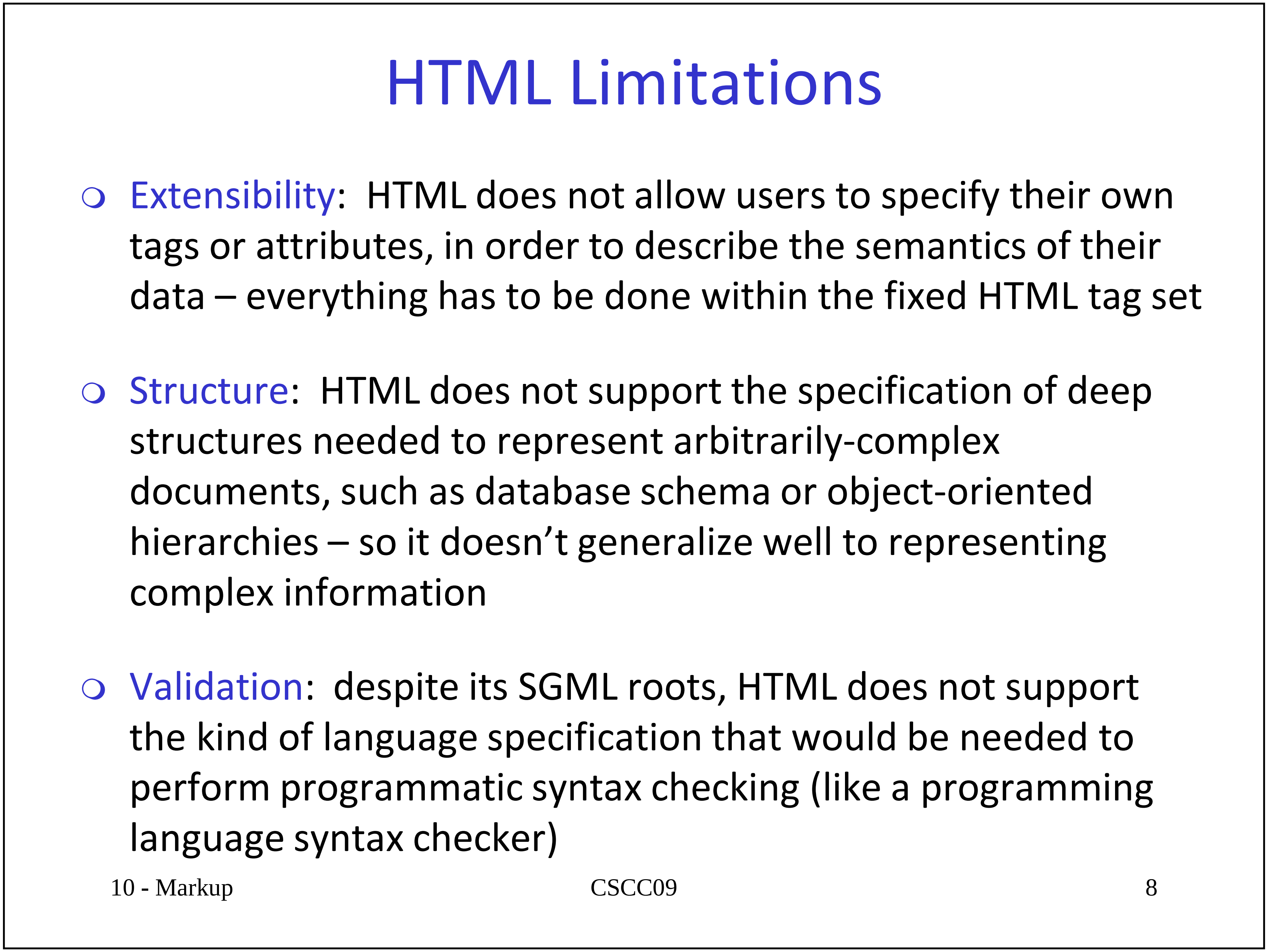
# Presentation not Content

- Once HTML went “viral”, it quickly “degenerated” into a language for expressing both structure and presentation
- Graphic designers were horrified at the lack of expressiveness and introduced numerous hacks, e.g.:
  - "pouring" text into tables for precise paragraph layout
  - using 1-pixel transparent images for precise spacing and alignment
- 2 warring factions: ‘purists’ vs ‘designers’
- Designers won hands-down! Why?



# A Temporary Truce

- CSS (Cascading Style Sheets)
- Afford graphic designers much better control over presentation.
- CSS separates detailed presentation instructions from the HTML structural markup.
  - Designers were happy:
    - CSS provided better control over formatting than HTML
  - Purists were happy:
    - removed the requirement for further presentation tags from HTML proper (a battle purists were headed towards utter defeat on anyway)

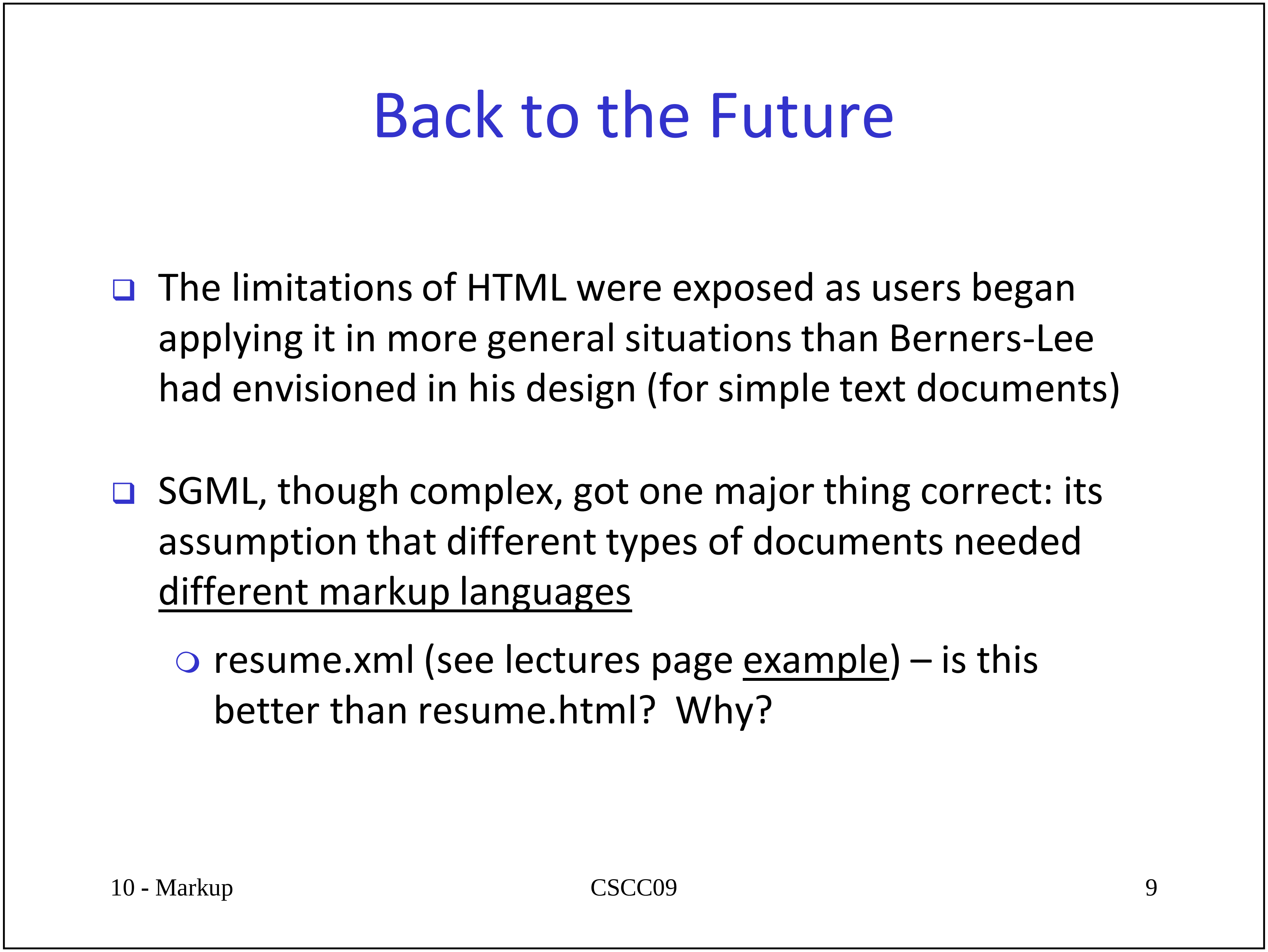


# HTML Limitations

- Extensibility: HTML does not allow users to specify their own tags or attributes, in order to describe the semantics of their data – everything has to be done within the fixed HTML tag set

- Structure: HTML does not support the specification of deep structures needed to represent arbitrarily-complex documents, such as database schema or object-oriented hierarchies – so it doesn’t generalize well to representing complex information

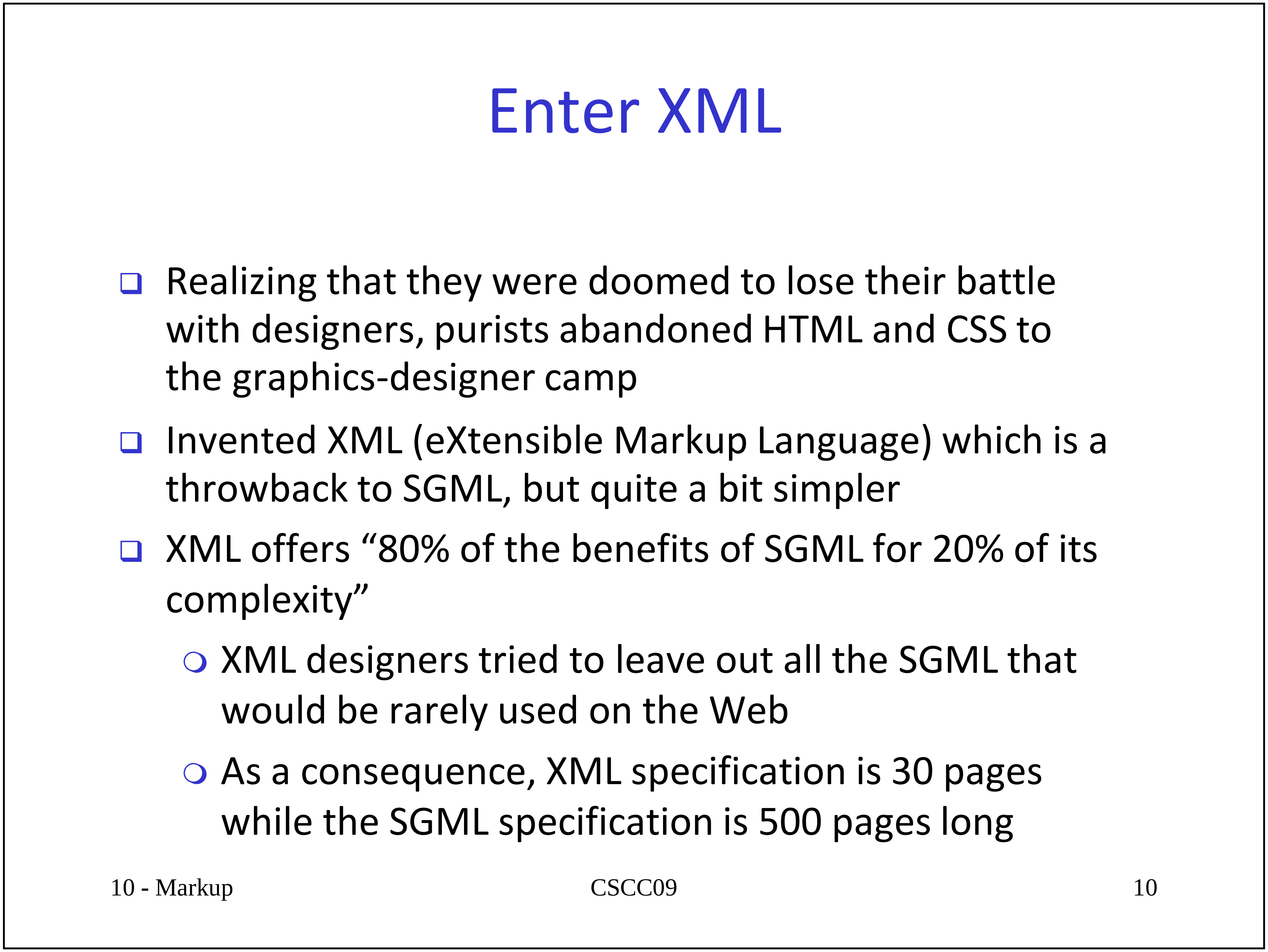
- Validation: despite its SGML roots, HTML does not support the kind of language specification that would be needed to perform programmatic syntax checking (like a programming language syntax checker)



# Back to the Future

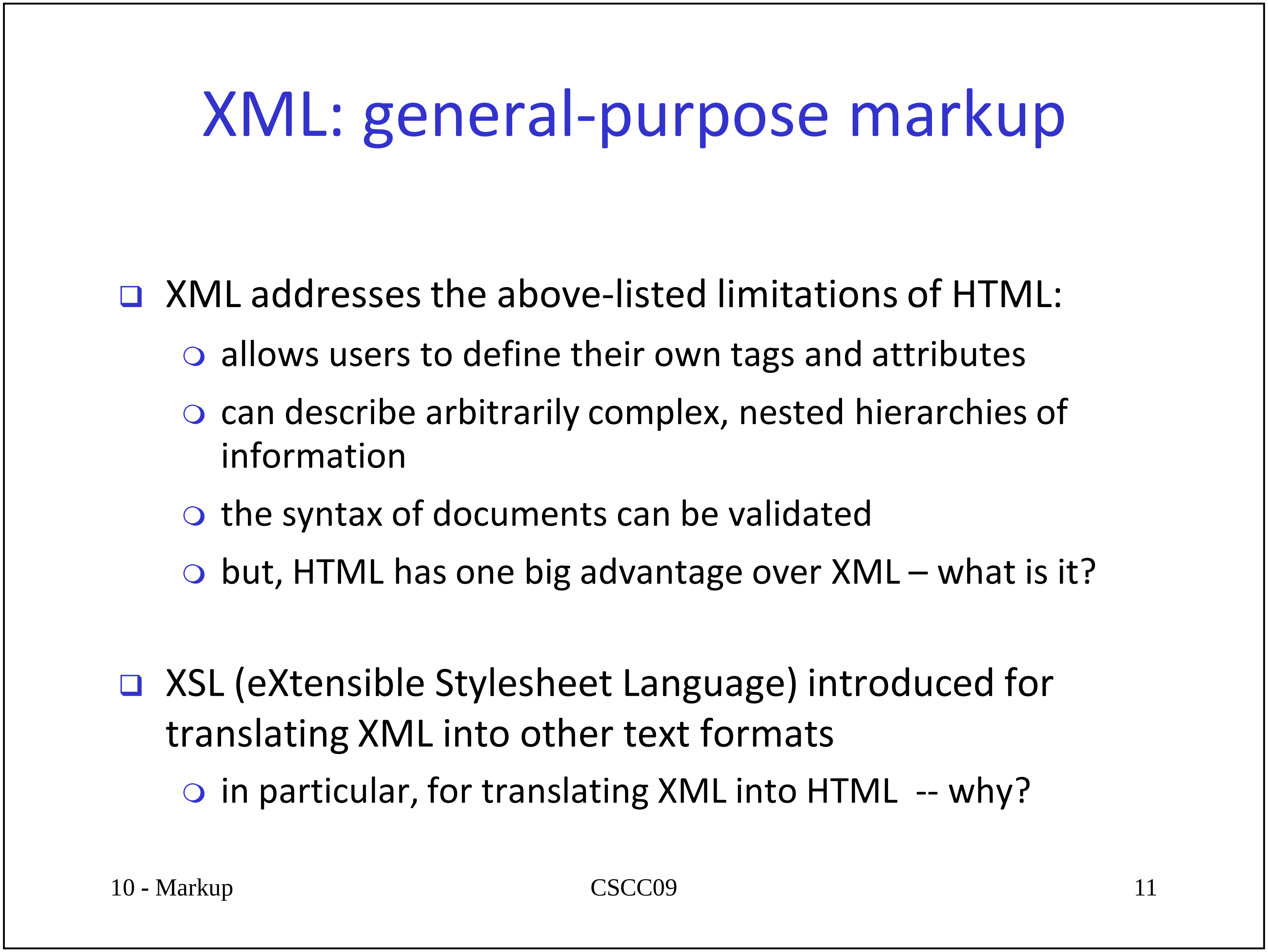
- The limitations of HTML were exposed as users began applying it in more general situations than Berners-Lee had envisioned in his design (for simple text documents)

- SGML, though complex, got one major thing correct: its assumption that different types of documents needed different markup languages
  - resume.xml (see lectures page example) – is this better than resume.html? Why?



# Enter XML

- Realizing that they were doomed to lose their battle with designers, purists abandoned HTML and CSS to the graphics-designer camp
- Invented XML (eXtensible Markup Language) which is a throwback to SGML, but quite a bit simpler
- XML offers “80% of the benefits of SGML for 20% of its complexity”
  - XML designers tried to leave out all the SGML that would be rarely used on the Web
  - As a consequence, XML specification is 30 pages while the SGML specification is 500 pages long

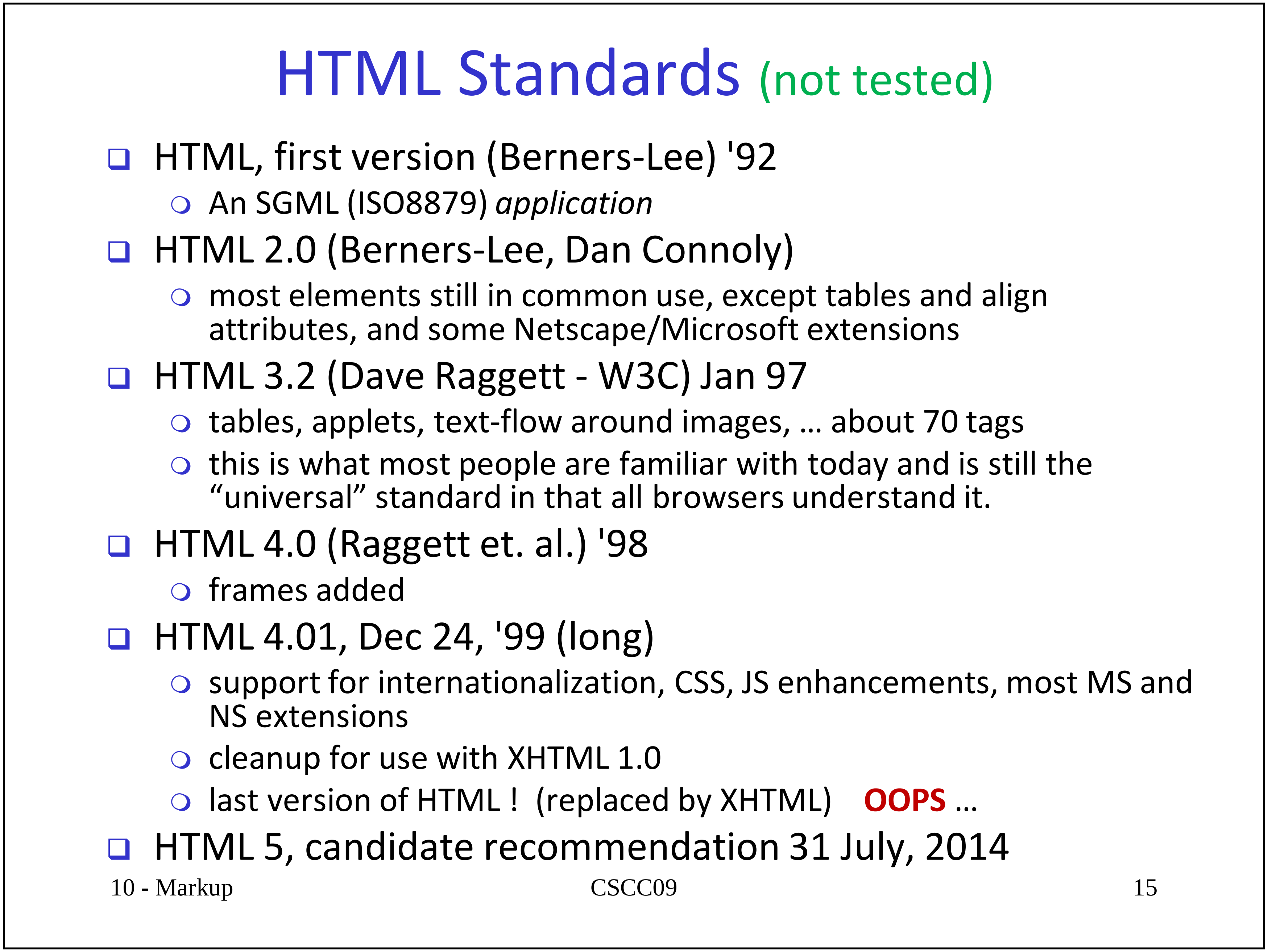


# XML: general-purpose markup

- XML addresses the above-listed limitations of HTML:
  - allows users to define their own tags and attributes
  - can describe arbitrarily complex, nested hierarchies of information
  - the syntax of documents can be validated
  - but, HTML has one big advantage over XML – what is it?
- XSL (eXtensible Stylesheet Language) introduced for translating XML into other text formats
  - in particular, for translating XML into HTML -- why?

# What is the W3C?

- World Wide Web Consortium (W3C)
  - established in October 1994
  - mandate to develop common, open standards and protocols for the Web (e.g. HTML, XML, XSL, CSS, XSchema, WSDL, SOAP)
- Members:
  - companies, governments, standards bodies, …
- the Team:
  - full-time employees, paid for by the Members.
  - Hosted at MIT, ERCIM (Europe), Keio (Japan), and Beihang (China)



# HTML Standards (not tested)

- HTML, first version (Berners-Lee) '92
  - An SGML (ISO8879) *application*
- HTML 2.0 (Berners-Lee, Dan Connoly)
  - most elements still in common use, except tables and align attributes, and some Netscape/Microsoft extensions
- HTML 3.2 (Dave Raggett - W3C) Jan 97
  - tables, applets, text-flow around images, … about 70 tags
  - this is what most people are familiar with today and is still the “universal” standard in that all browsers understand it.
- HTML 4.0 (Raggett et. al.) '98
  - frames added
- HTML 4.01, Dec 24, '99 (long)
  - support for internationalization, CSS, JS enhancements, most MS and NS extensions
  - cleanup for use with XHTML 1.0
  - last version of HTML ! (replaced by XHTML) **OOPS** …
- HTML 5, candidate recommendation 31 July, 2014

# Synopsis

- HTML designed as an application of SGML for display of simple text documents + hyperlinks
- SGML innovations: declarative markup and formal syntax
- As the Web expanded, demand for presentation drove markup evolution
- CSS introduced to separate structure (tags) from presentation (CSS)
- Use of HTML with more general documents revealed serious limitations
- XML introduced to handle the case of fully-general documents, with XSL to bridge between XML and HTML
- XHTML: an XML-compliant version of HTML
- The W3C – what, who, why
- HTML5: evolving standard for developing richer apps