

Understanding the quality of the narratives in corporate filings

Feng Li
University of Michigan

*LSE/LUMS/MBS Conference
London, June 24, 2013*

Roadmap

- Why studying narrative information (using computer programs)
- Framework
- Current research
- Future opportunities

Why narratives?

Analysis of narratives is nothing new

- Bible gospels authorship
- Suicide notes study by psychologists
- Was the rise of fascism tied to the contents in radio broadcasting?
- Spam filtering
- Machine translation
- Bioinformatics
- Searching for terrorists in online forums
- Customer feedback research

Important information source



The Dow Chemical Company
2009 Annual Report

- 207 pages
- 30 pages of tables
- The rest is narrative disclosures
 - ~ Chairman's letter to stockholders
 - ~ MD&A
 - ~ Notes to financial statements
 - ~ ...

Typical 10-K: 300+ numbers, 30,000+ words

Help us understand quantitative data and firm disclosure behavior

- Data generating function
 - ~ Notes to the financial statements
 - ~ E.g., sales revenue increases, but revenue recognition method changed
- Can be forward-looking compared with many quantitative disclosures
- Provide a richer environment to test disclosure theories

“We have incurred significant losses since our inception, and we expect to continue to incur net losses for the foreseeable future.”

Understanding managers' cognitive processes

- How do you measure cognitive/behavioral characteristics using archival data?
- Attributions: concepts, attitudes, beliefs, intentions, emotions, mental states and cognitive processes.
- Social relationships: authority, power.

Example: Manager heuristics (self-attribution bias)

American International Group, Inc. 2006 Annual Report

- “Solid execution of our strategies and the absence of significant catastrophes contributed to our outstanding results in 2006. Around the world and across all of our business segments we are capitalizing on growth opportunities, using our business diversity and matrix management structure to respond quickly to customer needs.”

Example: Manager heuristics (self-attribution bias)

American International Group, Inc. 2008 Annual Report

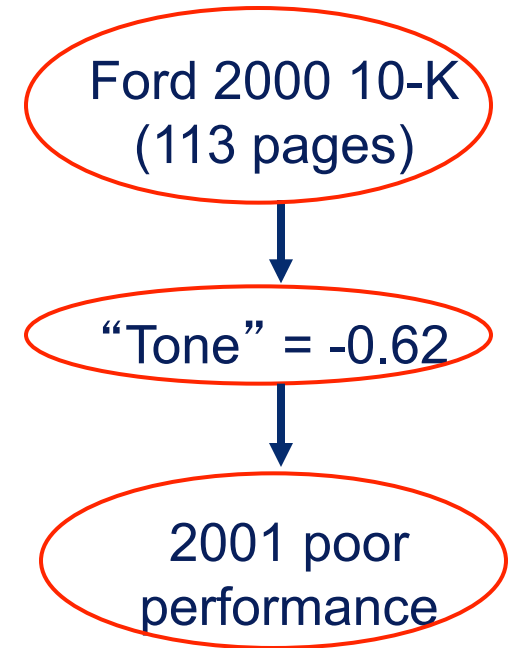
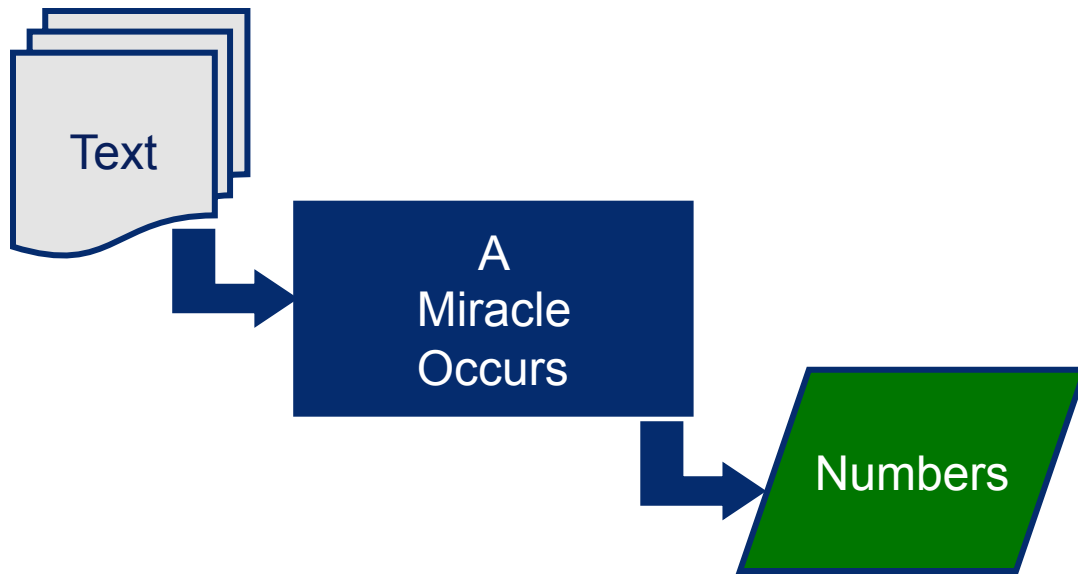
- “AIG reported that the continued severe credit market deterioration, particularly in mortgage-backed securities, and charges related to ongoing restructuring activities, contributed to a record net loss for the fourth quarter of \$61.7 billion, or \$22.95 per diluted share, compared to a 2007 fourth quarter net loss of \$5.3 billion, or \$2.08 per diluted share.”

Human vs. machine

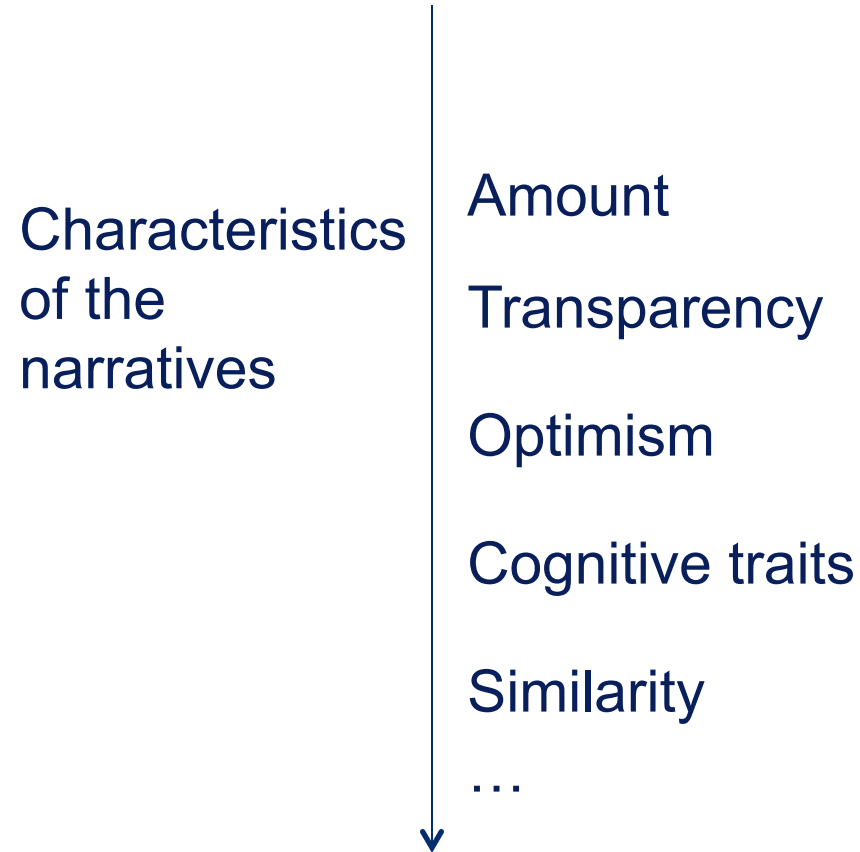
- Human coding
 - ~ Very costly (time and money)
- Recent technology has made quantitative measurement of content easier
 - ~ Digitalization of the record
 - ~ Computing power to analyze and record language

Framework

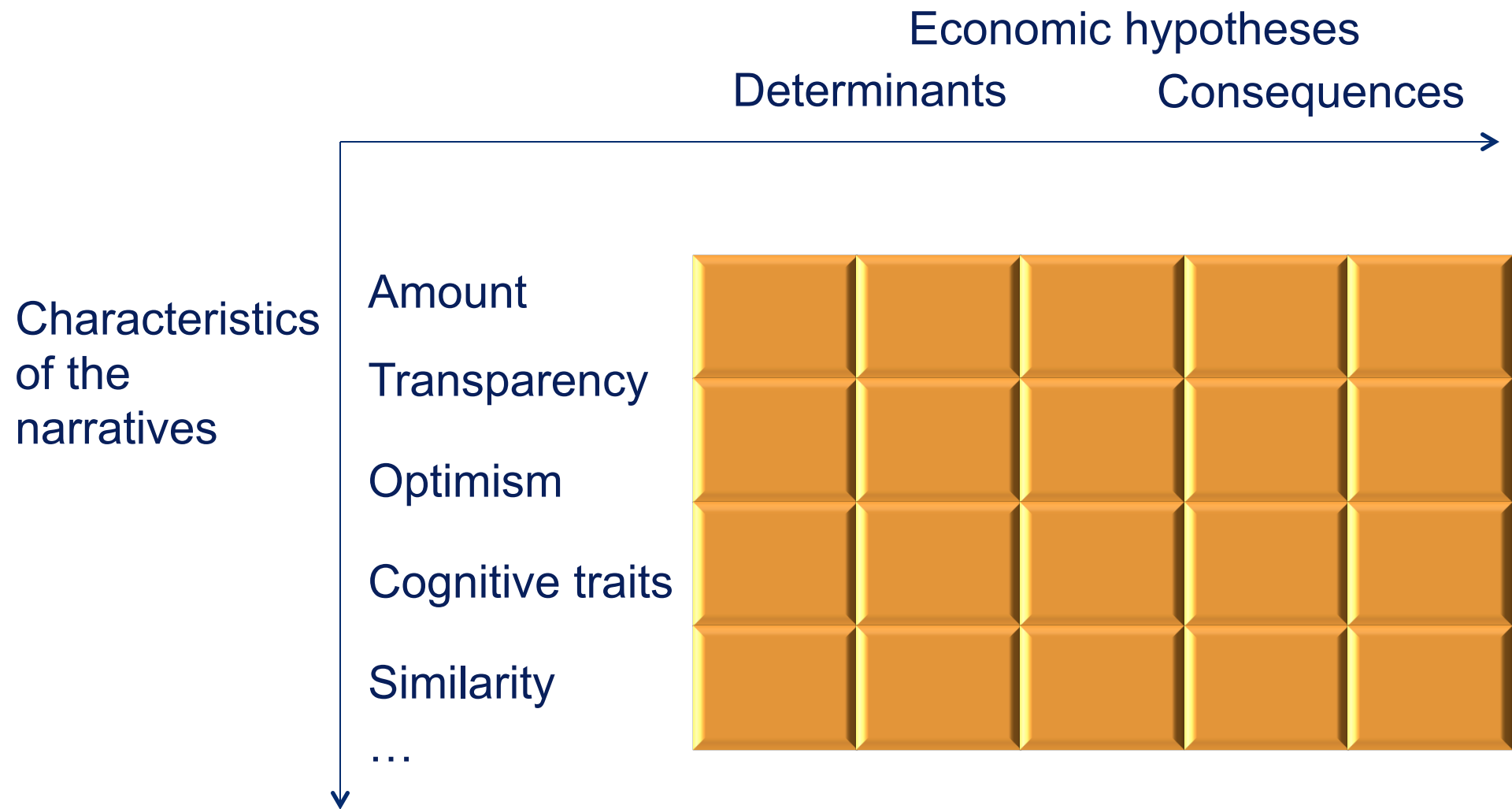
Essence: data reduction



Two dimensions



Two dimensions



Narrative information as main variable

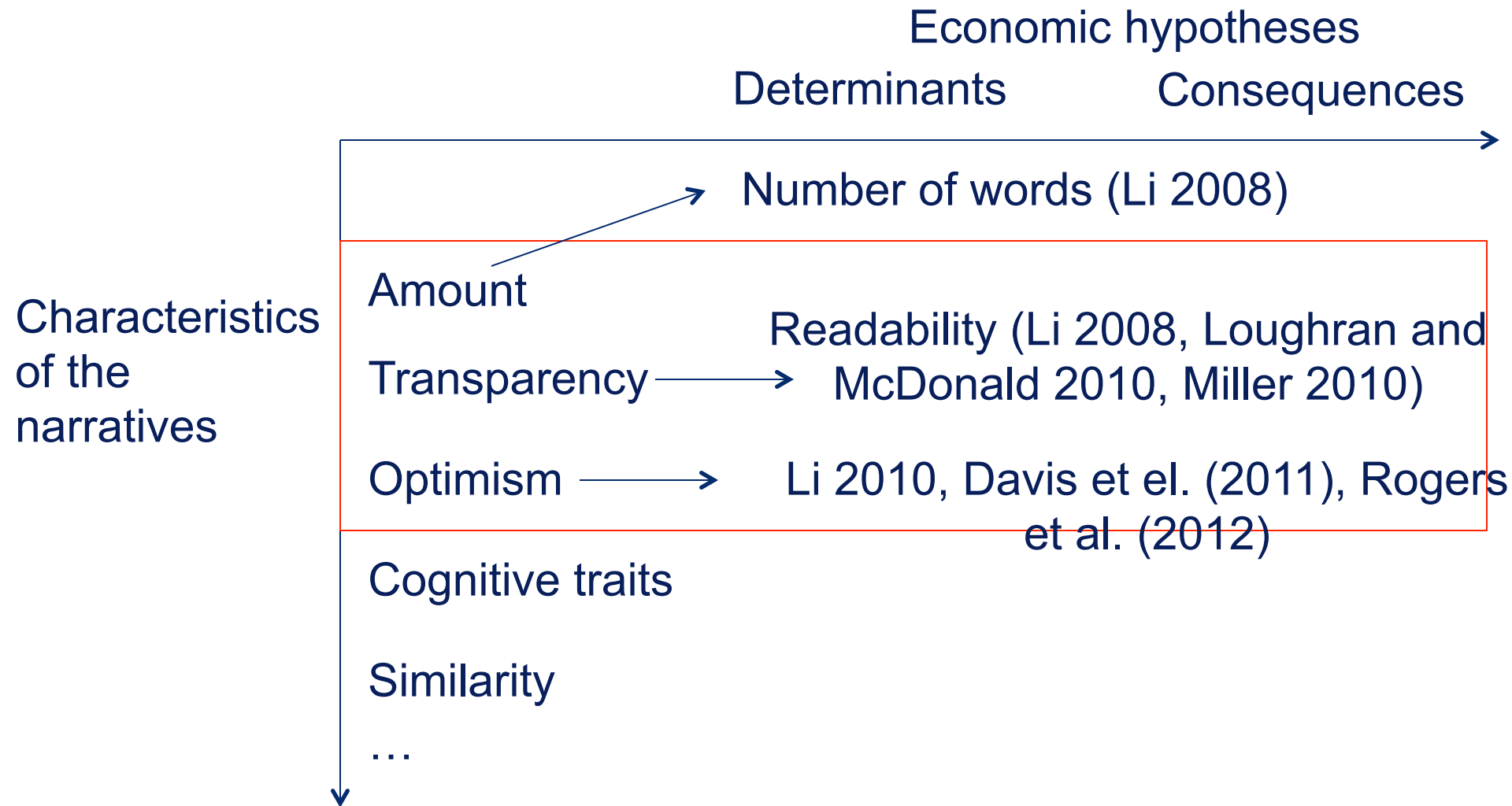
- GE' s “greatometer” (*Scott Davis, Morgan Stanley's lead GE analyst*)
 - ~ In 2002 Q3 call, Messrs. Immelt and Sherin said "great" more than 20 times.
 - ~ 2005 Q2, 70 times (GE shares rose 37%)
 - ~ 2006 Q3, 37 times (GE stock fell 10%)
- But ... 1 number might > 1000 words

Narrative information as contextual variable

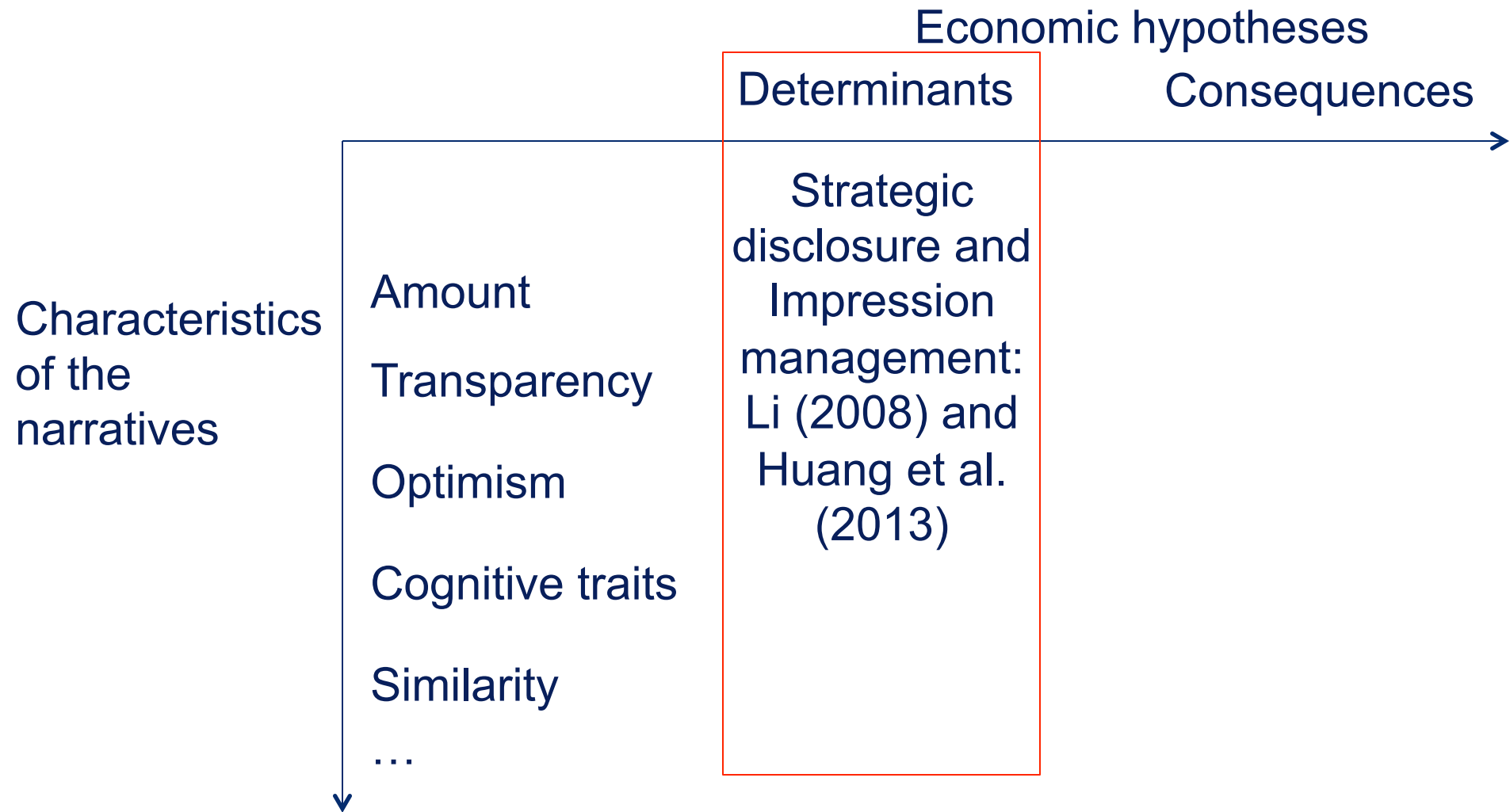
- Combine narratives with quantitative information
- GE' s “greatometer” may help us understand its earnings quality
 - ~ When GE CEO uses “great” more often, its earnings may have higher quality

Current research
(Cole and Jones 2005; Li 2011)

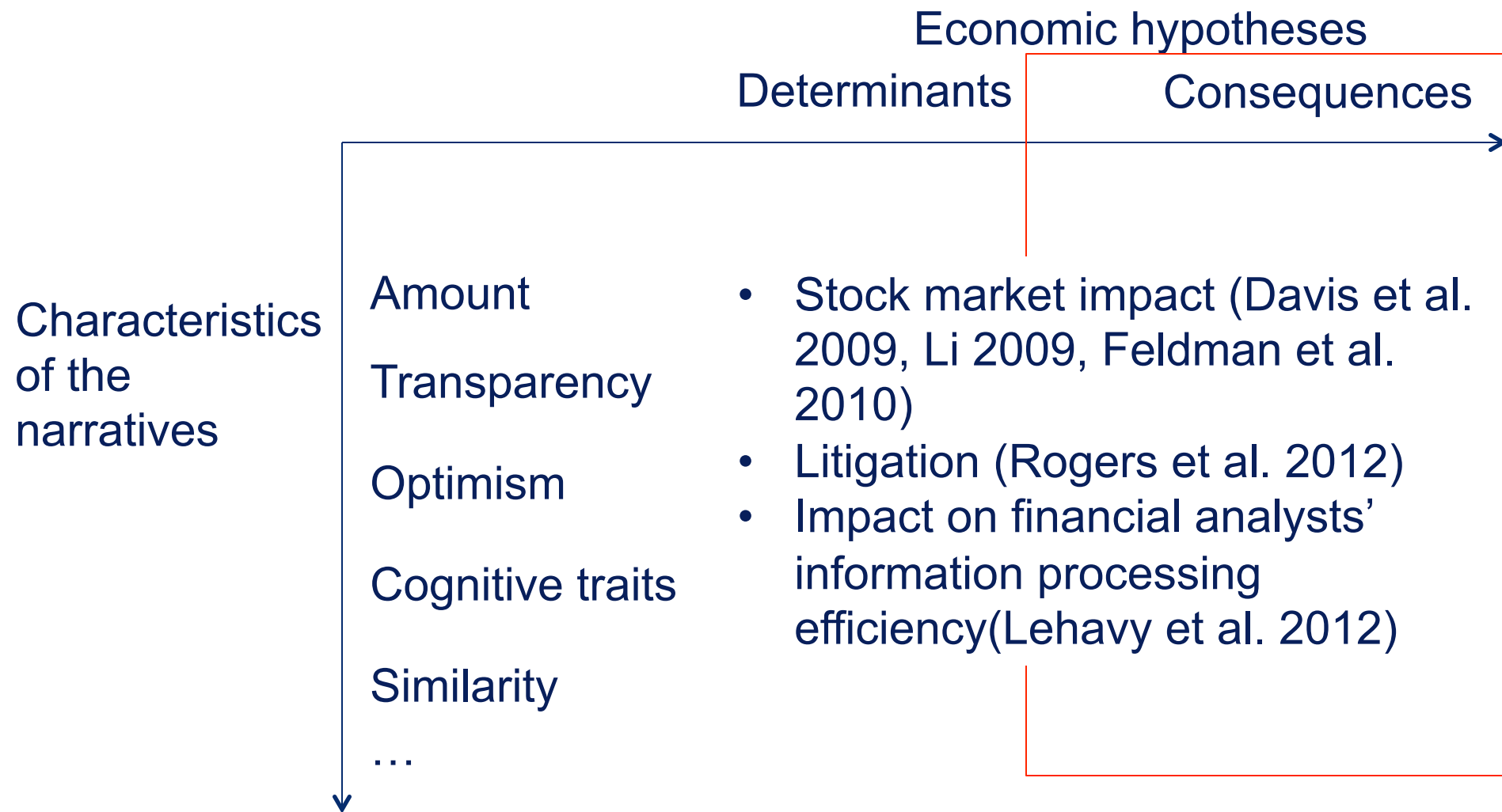
What has been done?



What has been done?



What has been done?

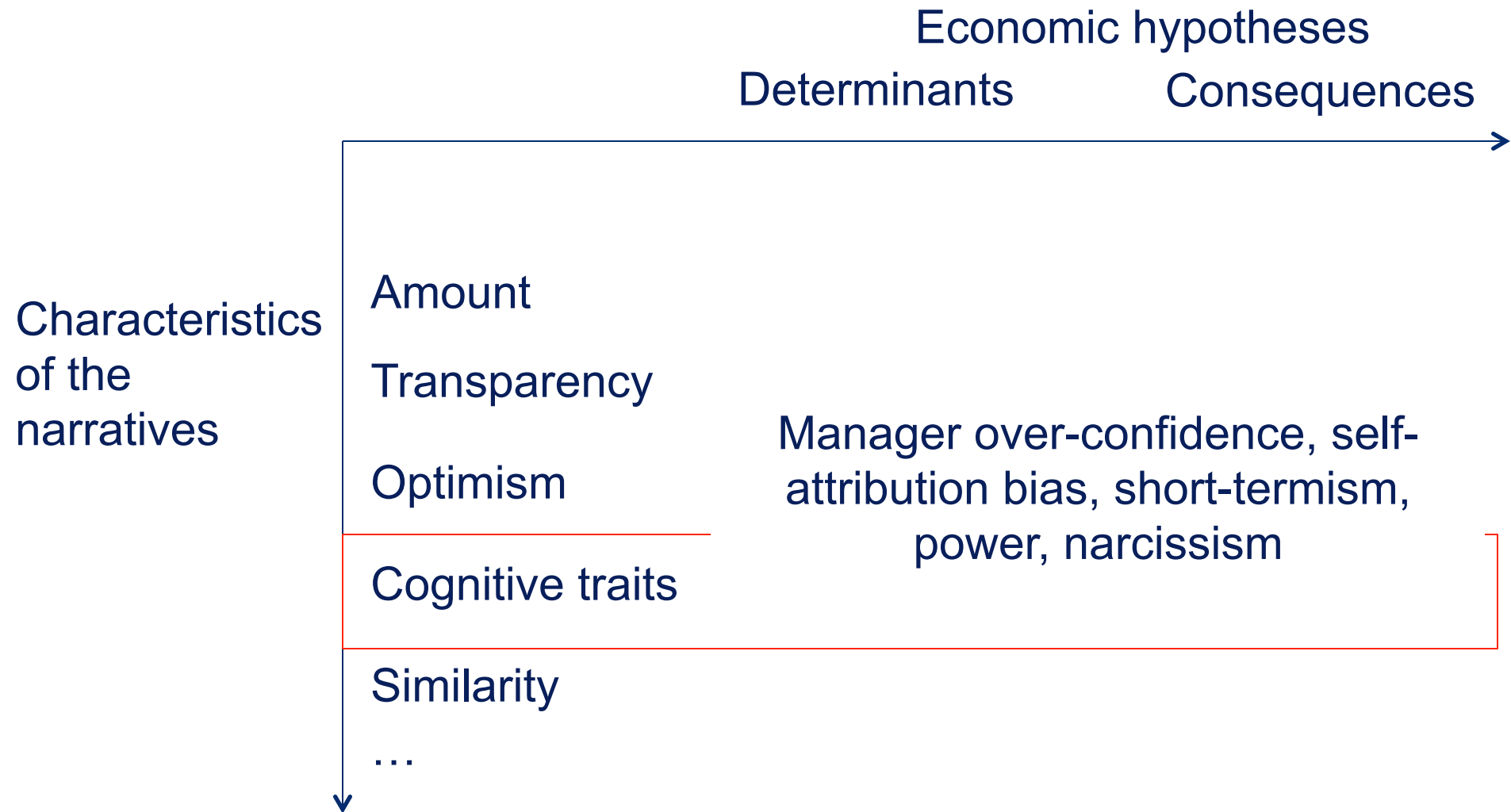


Current methodology: often preliminary

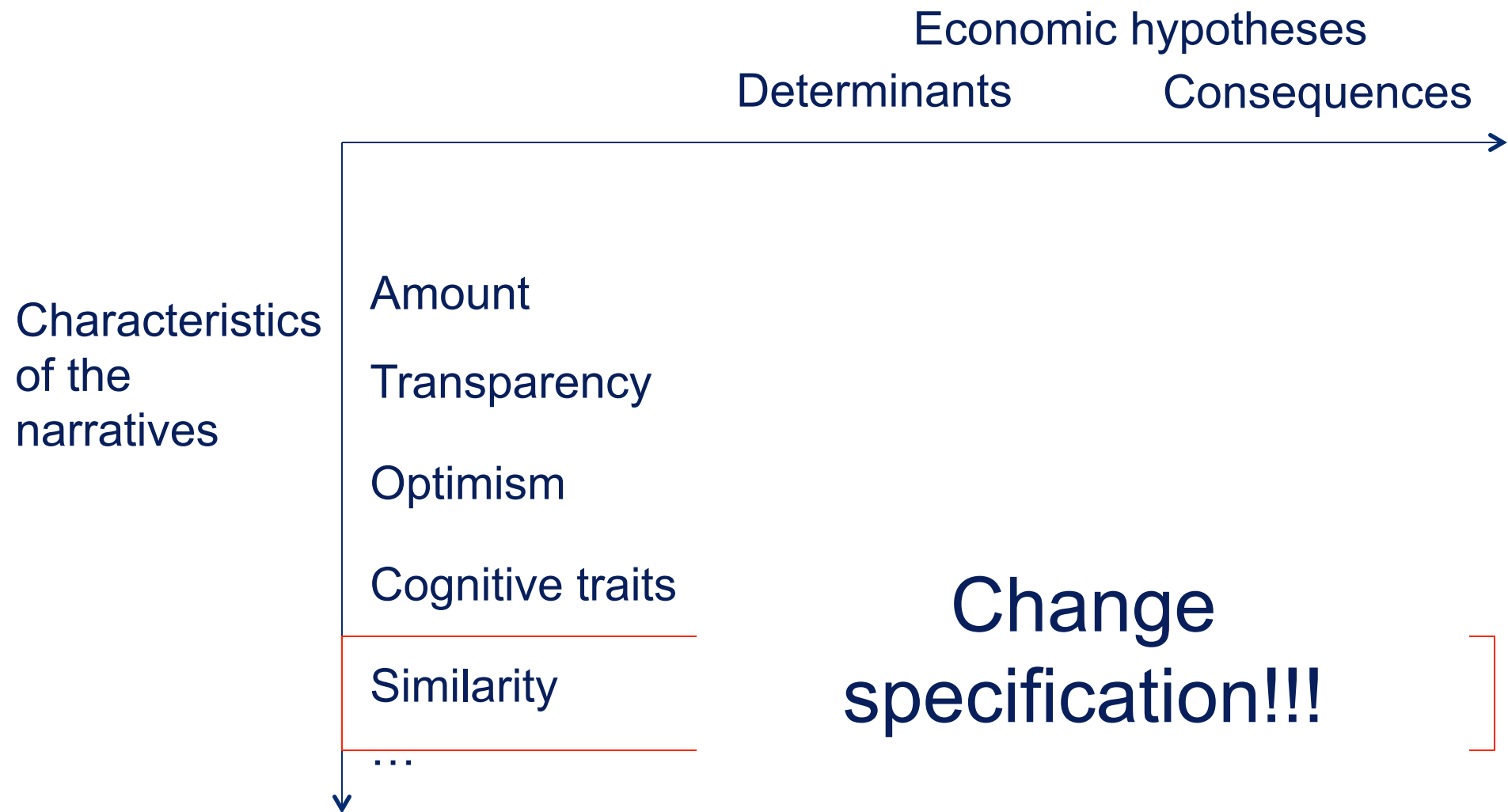
- Bag of words representation
- A sentence or a document is represented as an (unordered) collection of word.
 - ~ disregarding grammar and even word order.
- “Tom ate the wolf” = “The wolf ate Tom”
- Frequency count of specific words based on dictionaries

Future opportunities

What can be done?



What can be done?

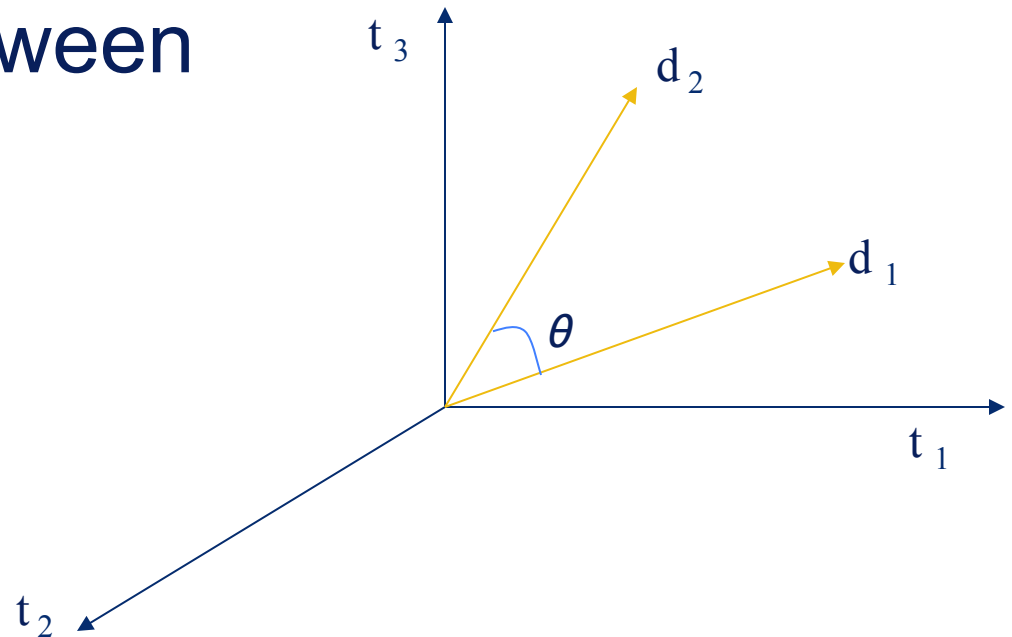


Example

- Time-series comparison
 - ~ BP 2012 annual report compared with its 2011 report
- Cross-sectional comparison
 - ~ BP annual report compared with that of Exxon Mobil (e.g., they have different shale gas reserve booking rules)

Cosine similarity

- Distance between vectors d_1 and d_2 captured by the cosine of the angle θ between them.

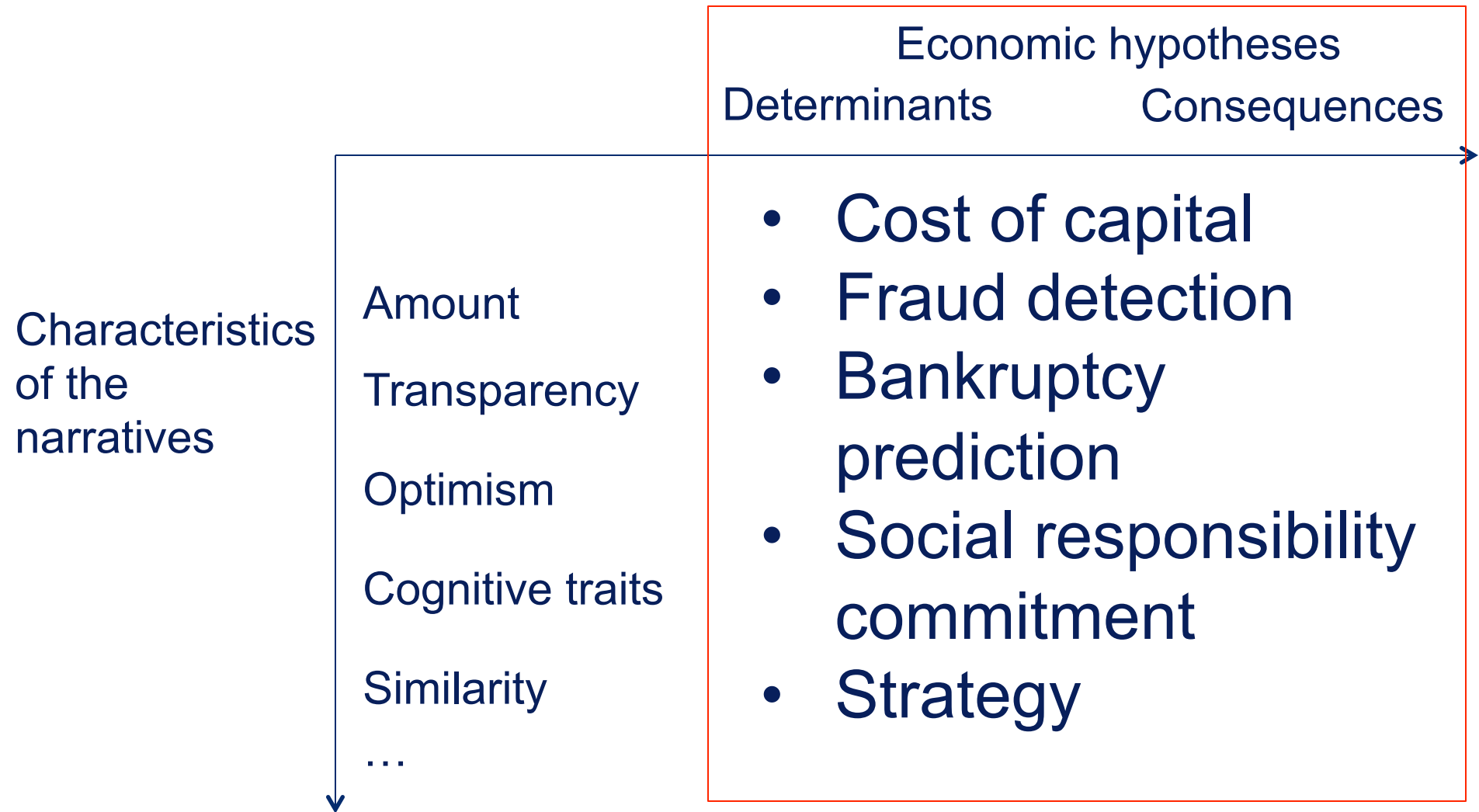


Cosine similarity

$$\text{sim}(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\|\vec{d}_j\| \|\vec{d}_k\|} = \frac{\sum_{i=1}^n w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^n w_{i,j}^2} \sqrt{\sum_{i=1}^n w_{i,k}^2}}$$

- Cosine of angle between two vectors
- The denominator measures the lengths of the vectors.

What can be done?



What can be done? Fundamental analysis

- Can we mimic Warren Buffett using computer programs?

“Other guys read Playboy, I read annual reports.”

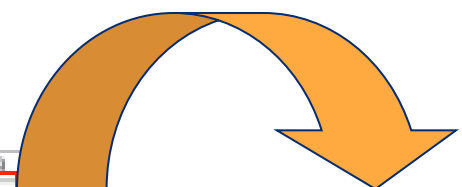


Fundamental analysis: promising direction

Consolidated Statements of Income

(In millions, except per share amounts)

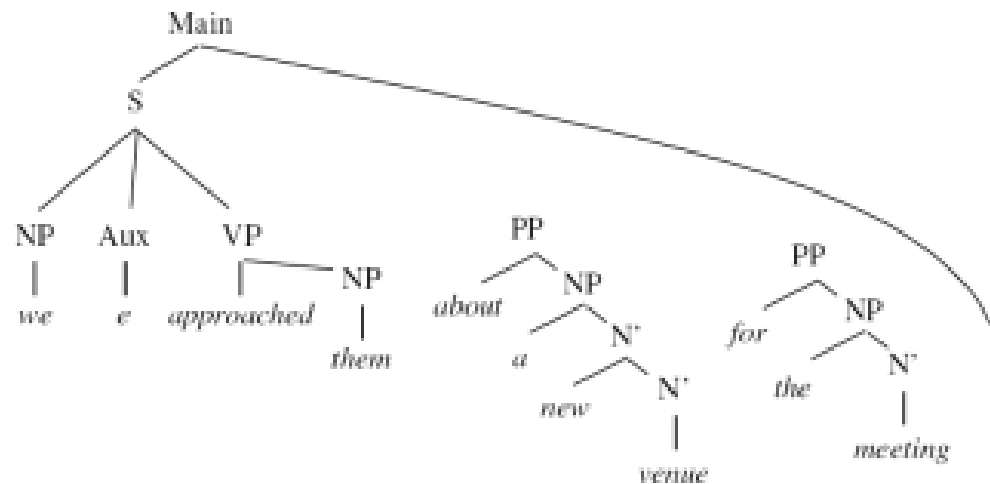
REVENUES	\$35,497
OPERATING EXPENSES:	
Salaries and employee benefits	13,767
Purchased transportation	4,534
Rentals and landing fees	2,429
Depreciation and amortization	1,975
Fuel	3,811
Maintenance and repairs	1,898
Impairment and other charges	1,204
Other	5,132
	34,750
OPERATING INCOME	747



Search for all the sentences that have “sales” or “revenues” as NP, and then extract the VP, ADJP, and ADVP in these sentences for further analysis (e.g., “increase”)

What can be done? more structured approach in terms of algorithms

- Par-of-speech (POS) tagging
 - ~ Rule-Based tagging (Voutilainen 1995)
 - ~ Stochastic (e.g., Hidden Markov Model) tagging (Brants 2000)
 - ~ Transformation-based tagging (Brill 1995)



Stanford NLP parser used in Chen and Li (2013)

<http://nlp.stanford.edu/software/lex-parser.shtml>

9 matches

pcfg Done



The Stanford Natural Language Processing Group

[home](#) · [people](#) · [teaching](#) · [research](#) · [publications](#) · [software](#) · [events](#) · [local](#)

The Stanford Parser: A statistical parser

[About](#) | [Citing](#) | [Questions](#) | [Mailing lists](#) | [Download](#) | [Included Tools](#) | [Extensions](#) | [Release history](#) | [Sample output](#) | [Online](#) | [FAQ](#)

About

A natural language parser is a program that works out the grammatical **structure of sentences**, for instance, which groups of words go together (as "phrases") and which words are the **subject** or **object** of a verb. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the *most likely* analysis of new sentences. These statistical parsers still make some mistakes, but commonly work rather well. Their development was one of the biggest breakthroughs in natural language processing in the 1990s. You can [try out our parser online](#).

This package is a Java implementation of probabilistic natural language parsers, both highly optimized PCFG and lexicalized dependency parsers, and a lexicalized PCFG parser. The original version of this parser was mainly written by Dan Klein, with support code and linguistic grammar development by Christopher Manning. Extensive additional work (internationalization and language-specific modeling, flexible input/output, grammar compaction, lattice parsing, *k*-best parsing, typed dependencies output, user support, etc.) has been done by Roger Levy, Christopher Manning, Teg Grenager, Galen Andrew, Marie-Catherine de Marneffe, Bill MacCartney, Anna Rafferty, Spence Green, Huihsin Tseng, Pi-Chuan Chang, Wolfgang Maier, and Jenny Finkel.

The lexicalized probabilistic parser implements a factored product model, with separate PCFG phrase structure and lexical dependency experts, whose preferences are combined by efficient exact inference, using an A* algorithm. Or the software can be used simply as an accurate unlexicalized stochastic context-free grammar parser. Either of these yields a good performance statistical parsing system. A GUI is provided for viewing the phrase structure tree output of the parser.

As well as providing an **English** parser, the parser can be and has been adapted to work with other languages. A **Chinese** parser based on the Chinese Treebank, a **German** parser based on the Negra corpus and **Arabic** parsers based on the Penn Arabic Treebank are also included. The parser has also been used for other languages, such as Italian, Bulgarian, and Portuguese.

The parser provides [Stanford Dependencies](#) output as well as phrase structure trees. Typed dependencies are otherwise known **grammatical relations**. This style of output is available only for English and Chinese. For more details, please refer to the [Stanford Dependencies webpage](#).

The current version of the parser requires Java 6 (JDK1.6) or later. (You can also download an old version of the parser, version 1.4, which runs under JDK 1.4, or version 2.0 which runs under JDK 1.5, but those distributions are no longer supported.) The parser also requires a reasonable amount of memory (at least 100MB to run as a PCFG parser on sentences up to 40 words in length; typically around 500MB of memory to be able to parse similarly long typical-of-newswire sentences using the factored model).

The parser is available for download, **licensed under the GNU General Public License** (v2 or later). Source is included. The package includes components for command-line invocation, a Java parsing GUI, and a Java API. The parser code is dual licensed (in a similar manner to MySQL, etc.). Open source licensing is under the *full* GPL, which allows many free uses. For distributors of **proprietary software**, **commercial licensing** with a [ready-to-sign agreement](#) is available. If you don't need a commercial license, but would like to support maintenance of these tools, we welcome gift funding.

Chen and Li (2013)

- “PCFG” (Probabilistic context-free grammars)
 - ~ Direct object (verb or object)
 - “Estimate the receivables”
 - ~ Passive nominal subject (verb or object)
 - “Receivables are estimated as”
 - ~ Adjective modifier
 - “Likely loss”
 - ~ Noun compound subjects
 - “Estimation value”
 - ~ Quantifier phrase modifier
 - “Is approximately \$100 million”

Conclusion

- This is an important area with significant research potential!
- E-mail me at feng@umich.edu if you are interested.
- Thank you!