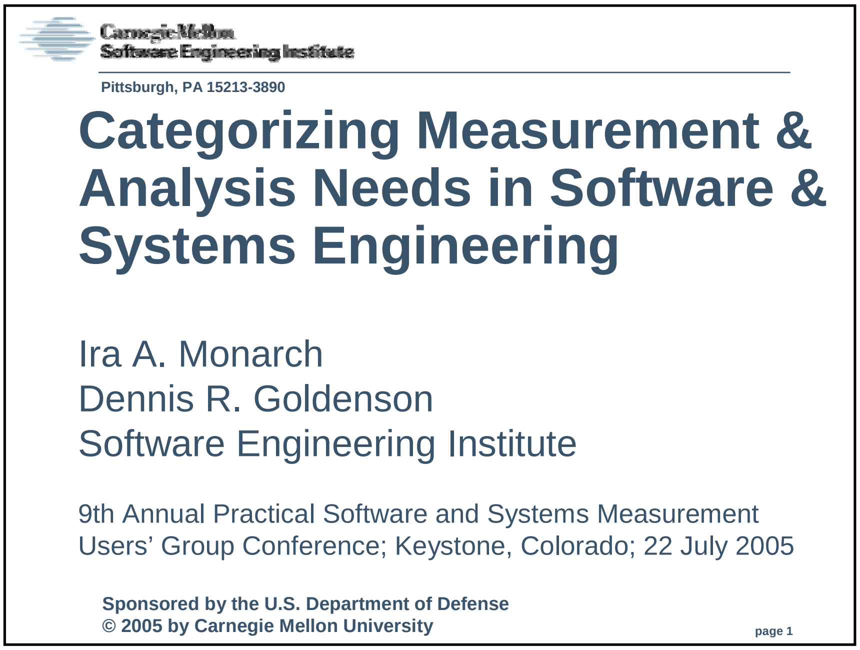


Carnegie Mellon
Software Engineering Institute

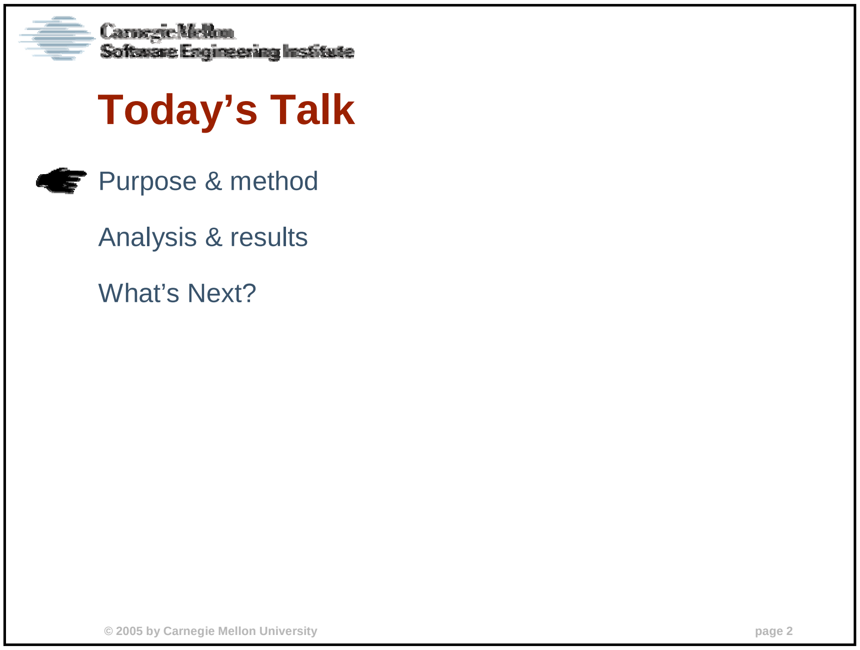
Pittsburgh, PA 15213-3890

Categorizing Measurement & Analysis Needs in Software & Systems Engineering

Ira A. Monarch
Dennis R. Goldenson
Software Engineering Institute

9th Annual Practical Software and Systems Measurement Users' Group Conference; Keystone, Colorado; 22 July 2005

Sponsored by the U.S. Department of Defense



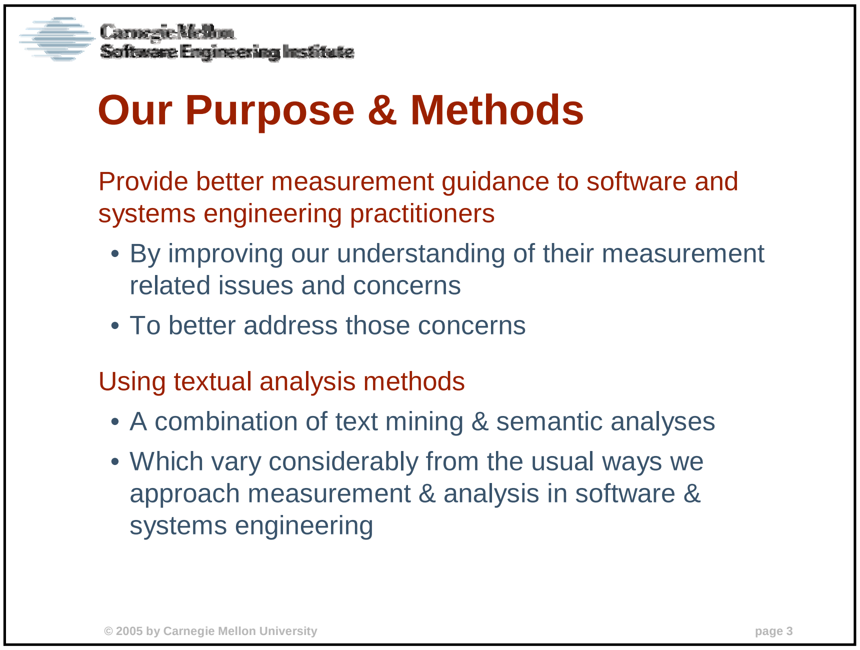
Carnegie Mellon
Software Engineering Institute

Today's Talk

☛ Purpose & method

Analysis & results

What's Next?



Carnegie Mellon
Software Engineering Institute

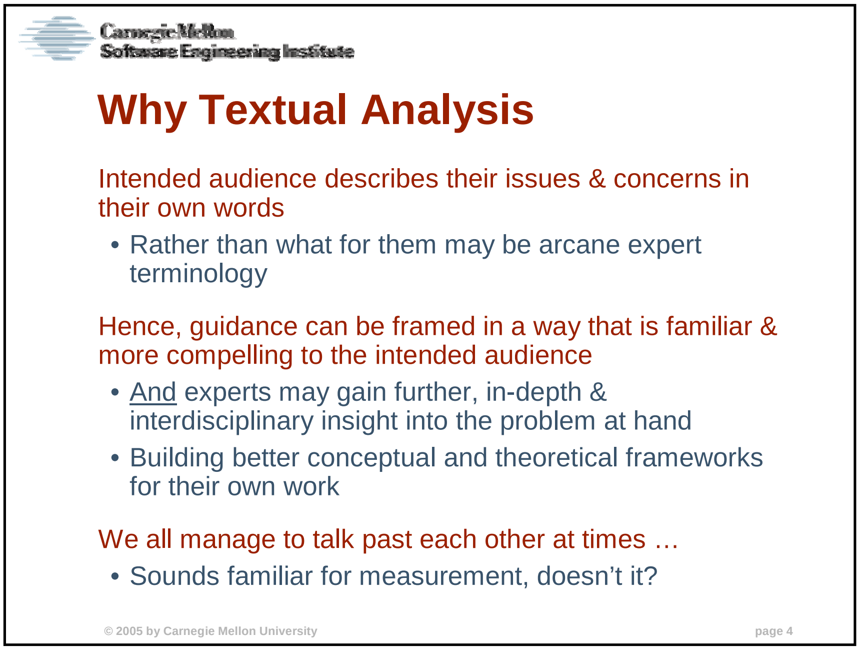
Our Purpose & Methods

Provide better measurement guidance to software and systems engineering practitioners

- By improving our understanding of their measurement related issues and concerns
- To better address those concerns

Using textual analysis methods

- A combination of text mining & semantic analyses
- Which vary considerably from the usual ways we approach measurement & analysis in software & systems engineering



Carnegie Mellon
Software Engineering Institute

Why Textual Analysis

Intended audience describes their issues & concerns in their own words

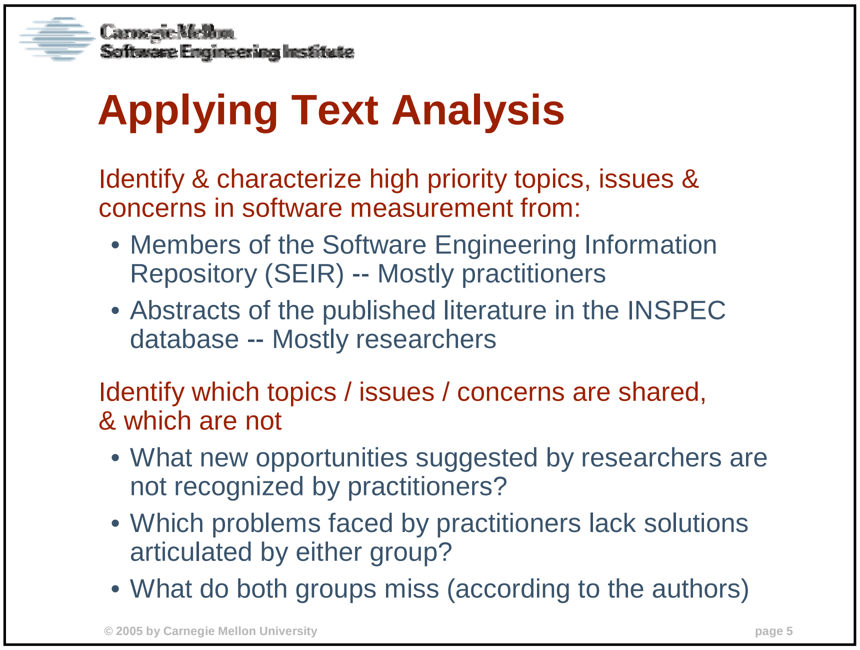
- Rather than what for them may be arcane expert terminology

Hence, guidance can be framed in a way that is familiar & more compelling to the intended audience

- And experts may gain further, in-depth & interdisciplinary insight into the problem at hand
- Building better conceptual and theoretical frameworks for their own work

We all manage to talk past each other at times …

- Sounds familiar for measurement, doesn't it?



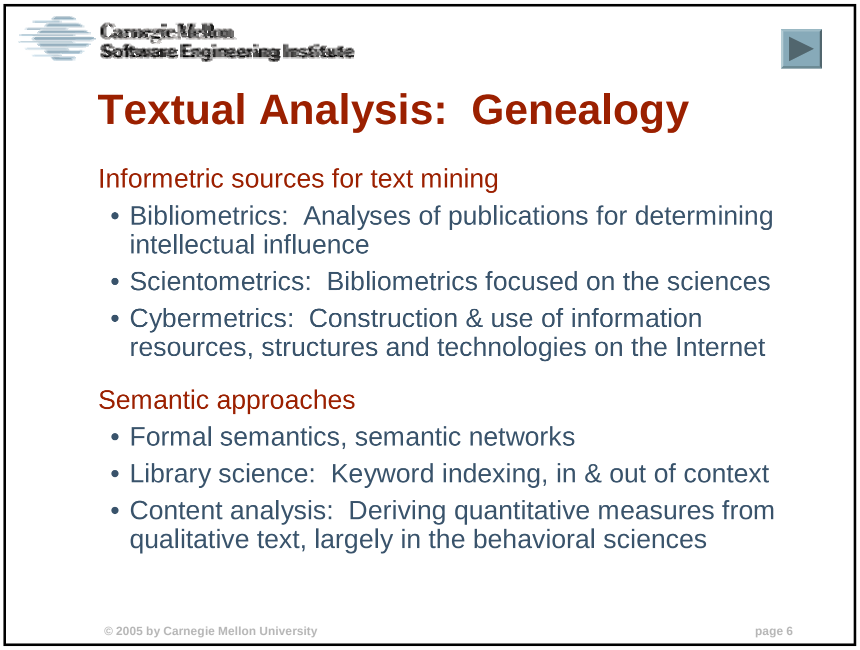
Applying Text Analysis

Identify & characterize high priority topics, issues & concerns in software measurement from:

- Members of the Software Engineering Information Repository (SEIR) -- Mostly practitioners
- Abstracts of the published literature in the INSPEC database -- Mostly researchers

Identify which topics / issues / concerns are shared, & which are not

- What new opportunities suggested by researchers are not recognized by practitioners?
- Which problems faced by practitioners lack solutions articulated by either group?
- What do both groups miss (according to the authors)



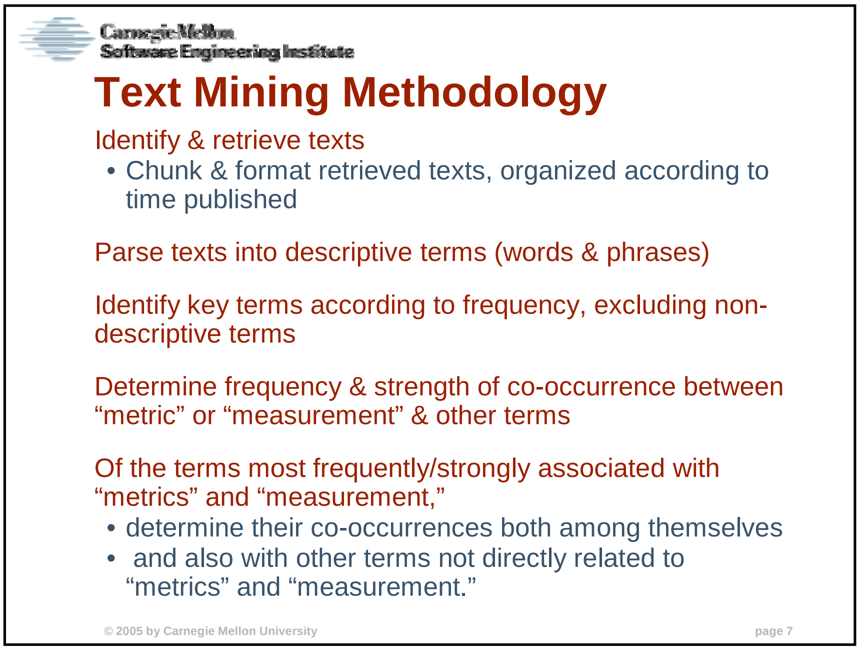
Textual Analysis: Genealogy

Informetric sources for text mining

- Bibliometrics: Analyses of publications for determining intellectual influence
- Scientometrics: Bibliometrics focused on the sciences
- Cybermetrics: Construction & use of information resources, structures and technologies on the Internet

Semantic approaches

- Formal semantics, semantic networks
- Library science: Keyword indexing, in & out of context
- Content analysis: Deriving quantitative measures from qualitative text, largely in the behavioral sciences



Text Mining Methodology

Identify & retrieve texts

- Chunk & format retrieved texts, organized according to time published

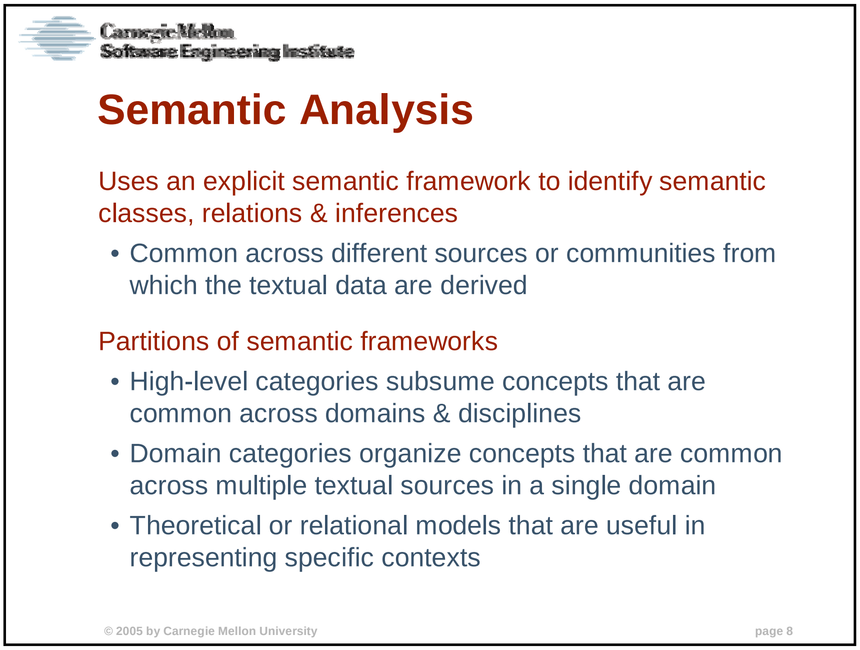
Parse texts into descriptive terms (words & phrases)

Identify key terms according to frequency, excluding non-descriptive terms

Determine frequency & strength of co-occurrence between “metric” or “measurement” & other terms

Of the terms most frequently/strongly associated with “metrics” and “measurement,”

- determine their co-occurrences both among themselves
- and also with other terms not directly related to “metrics” and “measurement.”



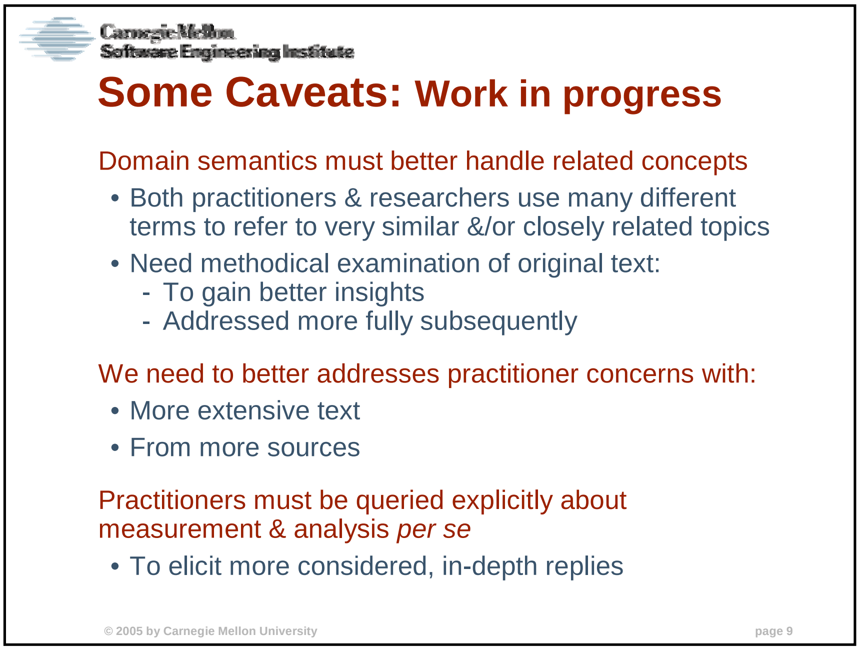
Semantic Analysis

Uses an explicit semantic framework to identify semantic classes, relations & inferences

- Common across different sources or communities from which the textual data are derived

Partitions of semantic frameworks

- High-level categories subsume concepts that are common across domains & disciplines
- Domain categories organize concepts that are common across multiple textual sources in a single domain
- Theoretical or relational models that are useful in representing specific contexts



Some Caveats: Work in progress

Domain semantics must better handle related concepts

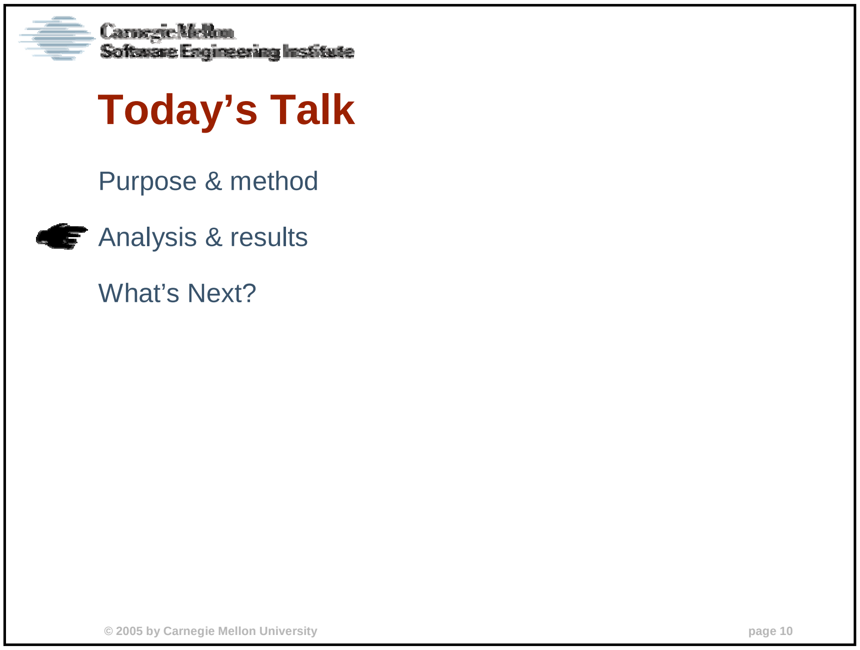
- Both practitioners & researchers use many different terms to refer to very similar &/or closely related topics
- Need methodical examination of original text:
  - To gain better insights
  - Addressed more fully subsequently

We need to better addresses practitioner concerns with:

- More extensive text
- From more sources

Practitioners must be queried explicitly about measurement & analysis *per se*

- To elicit more considered, in-depth replies

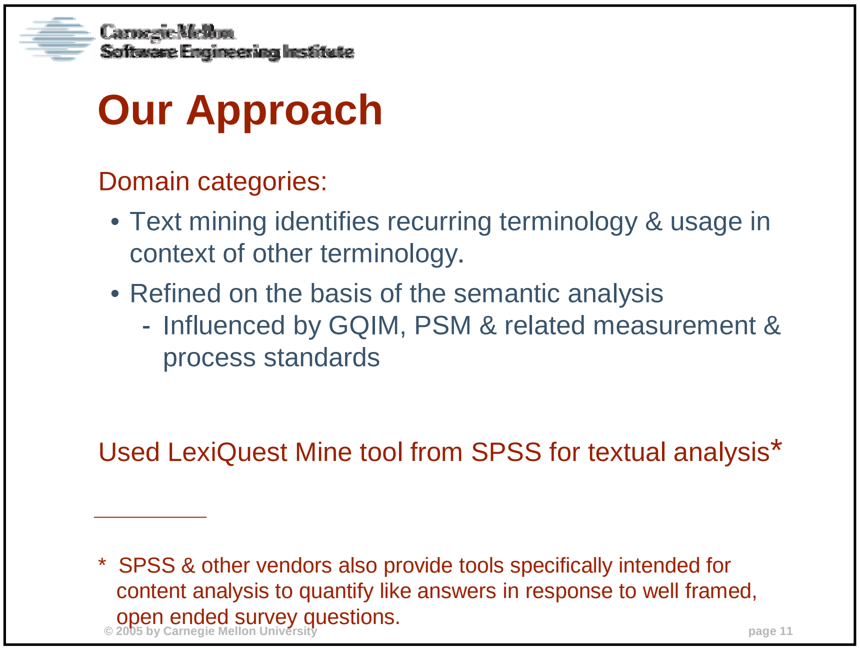


Today’s Talk

Purpose & method

☛ Analysis & results

What’s Next?



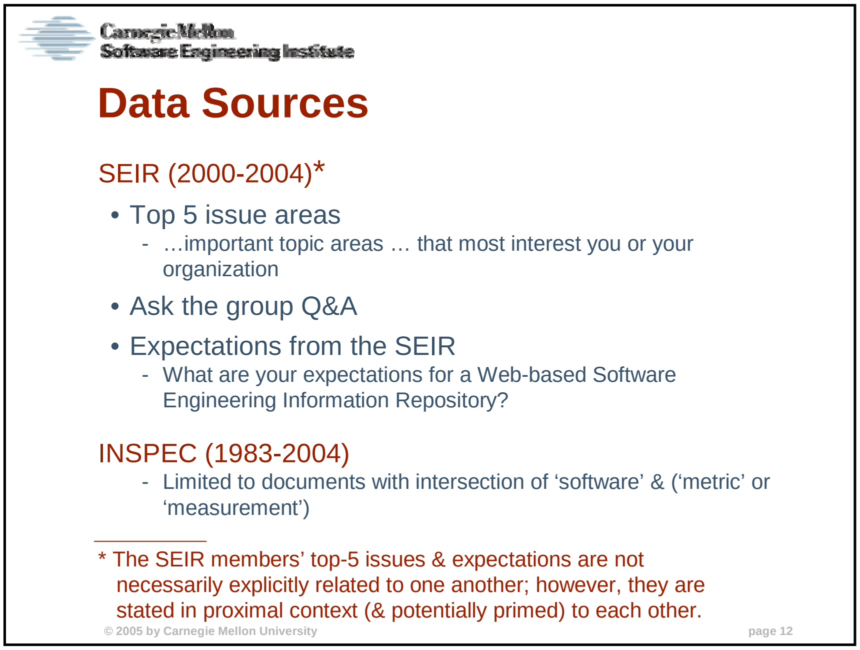
Our Approach

Domain categories:

- Text mining identifies recurring terminology & usage in context of other terminology.
- Refined on the basis of the semantic analysis
  - Influenced by GQIM, PSM & related measurement & process standards

Used LexiQuest Mine tool from SPSS for textual analysis*

* SPSS & other vendors also provide tools specifically intended for content analysis to quantify like answers in response to well framed, open ended survey questions.



Data Sources

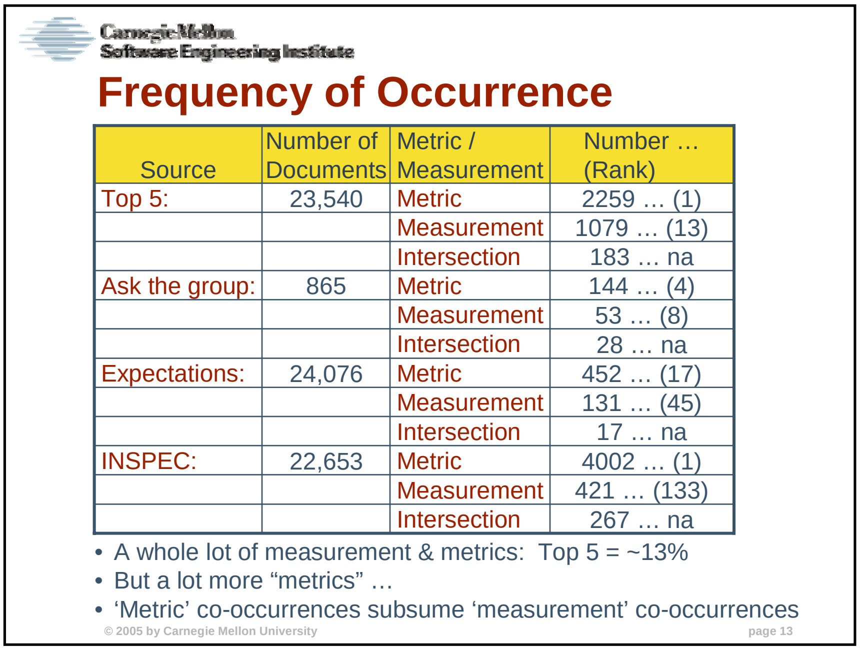
SEIR (2000-2004)*

- Top 5 issue areas
  - …important topic areas … that most interest you or your organization
- Ask the group Q&A
- Expectations from the SEIR
  - What are your expectations for a Web-based Software Engineering Information Repository?

INSPEC (1983-2004)

- Limited to documents with intersection of ‘software’ & (‘metric’ or ‘measurement’)

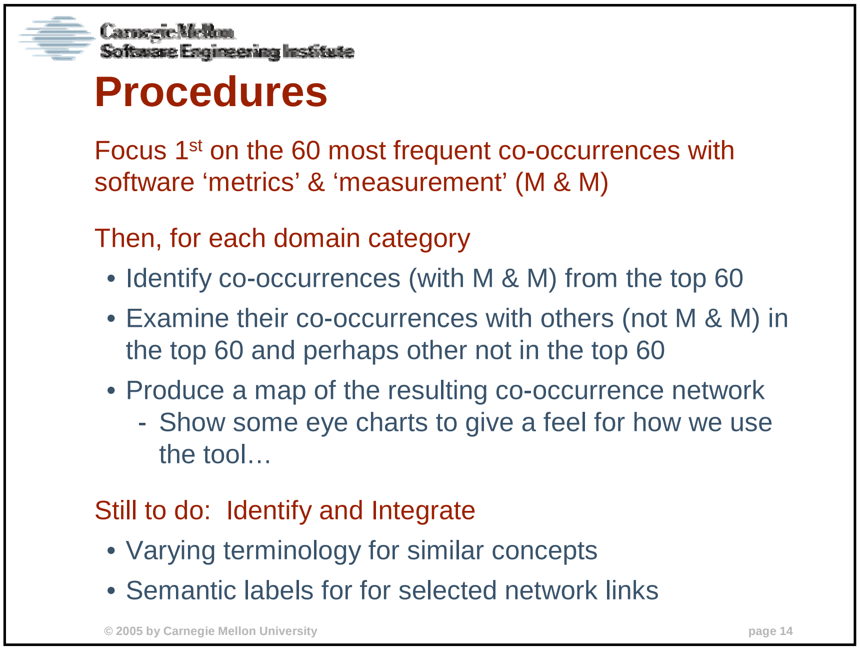
* The SEIR members’ top-5 issues & expectations are not necessarily explicitly related to one another; however, they are stated in proximal context (& potentially primed) to each other.



Frequency of Occurrence

	Number of	Metric /	Number …
Source	Documents	Measurement	(Rank)
Top 5:	23,540	Metric	2259 … (1)
		Measurement	1079 … (13)
		Intersection	183 … na
Ask the group:	865	Metric	144 … (4)
		Measurement	53 … (8)
		Intersection	28 … na
Expectations:	24,076	Metric	452 … (17)
		Measurement	131 … (45)
		Intersection	17 … na
INSPEC:	22,653	Metric	4002 … (1)
		Measurement	421 … (133)
		Intersection	267 … na

- A whole lot of measurement & metrics: Top 5 = ~13%
- But a lot more “metrics” …
- ‘Metric’ co-occurrences subsume ‘measurement’ co-occurrences



Procedures

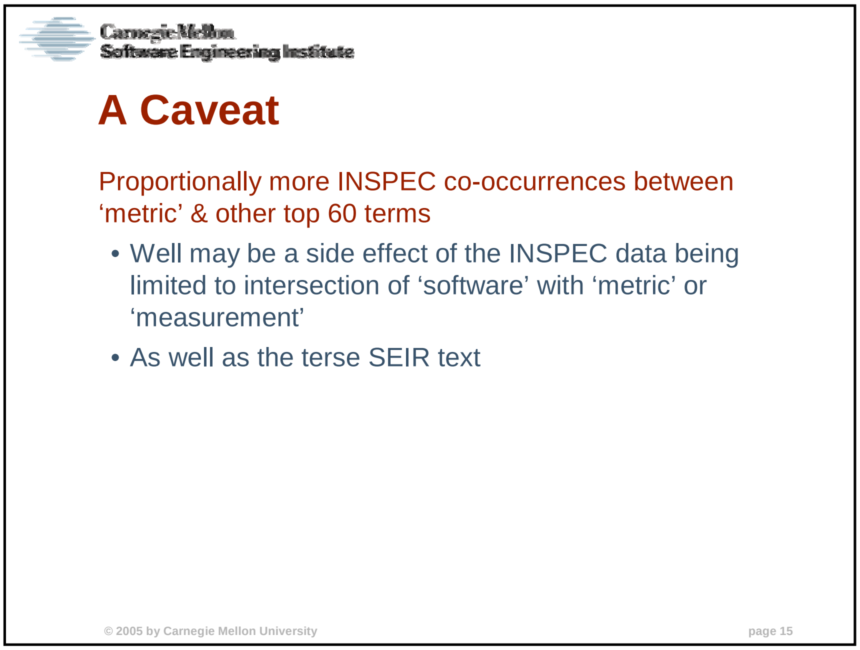
Focus 1st on the 60 most frequent co-occurrences with software ‘metrics’ & ‘measurement’ (M & M)

Then, for each domain category

- Identify co-occurrences (with M & M) from the top 60
- Examine their co-occurrences with others (not M & M) in the top 60 and perhaps other not in the top 60
- Produce a map of the resulting co-occurrence network
  - Show some eye charts to give a feel for how we use the tool…

Still to do: Identify and Integrate

- Varying terminology for similar concepts
- Semantic labels for for selected network links

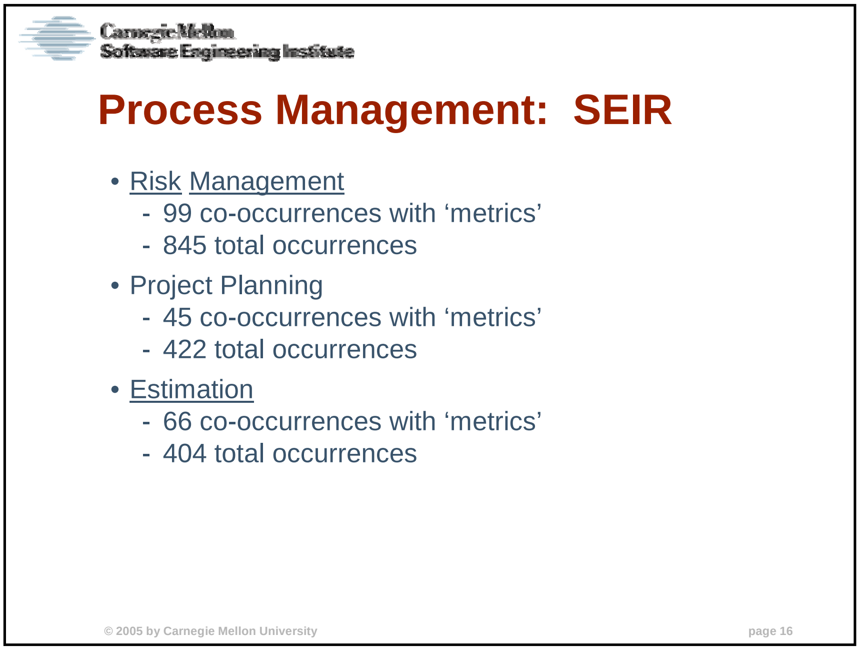


Carnegie Mellon
Software Engineering Institute

A Caveat

Proportionally more INSPEC co-occurrences between ‘metric’ & other top 60 terms

- Well may be a side effect of the INSPEC data being limited to intersection of ‘software’ with ‘metric’ or ‘measurement’
- As well as the terse SEIR text



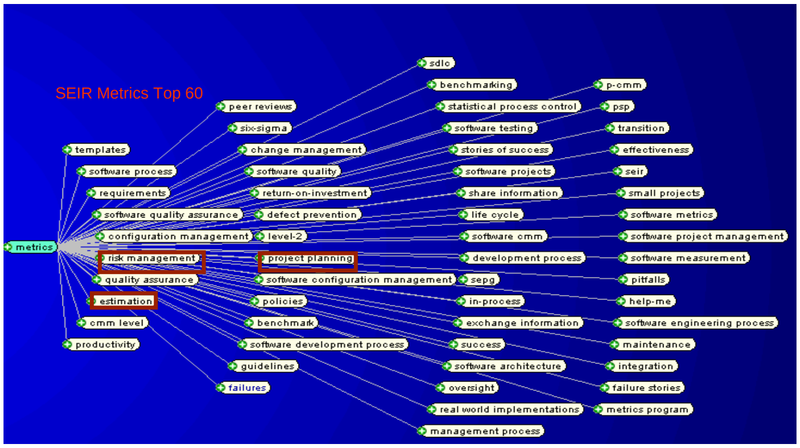
Carnegie Mellon
Software Engineering Institute

Process Management: SEIR

- Risk Management
  - 99 co-occurrences with ‘metrics’
  - 845 total occurrences
- Project Planning
  - 45 co-occurrences with ‘metrics’
  - 422 total occurrences
- Estimation
  - 66 co-occurrences with ‘metrics’
  - 404 total occurrences

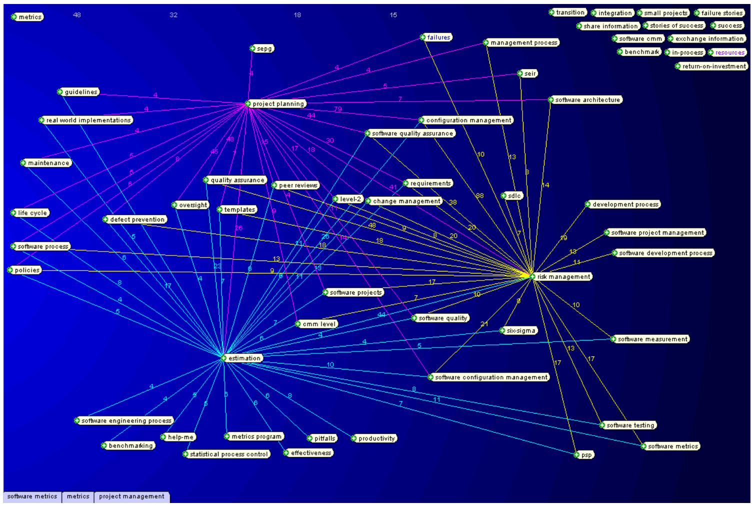
Carnegie Mellon
Software Engineering Institute

SEIR Project Management: Top 60



Carnegie Mellon
Software Engineering Institute

SEIR Project Management Relations



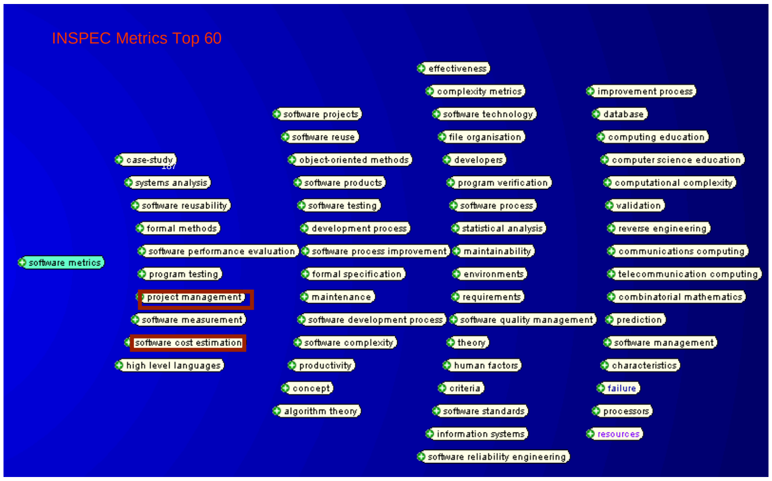
Carnegie Mellon
Software Engineering Institute

Process Management: INSPEC

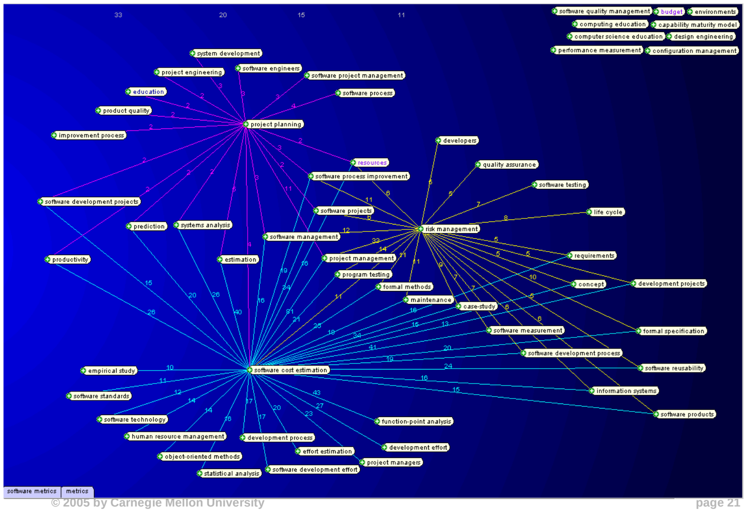
- Project Management
  - 309 co-occurrences with ‘metrics’
  - 447 total occurrences
- Software Cost Estimation
  - 296 co-occurrences with ‘metrics’
  - 357 total occurrences
- Risk Management
  - 81 co-occurrences with ‘metrics’
  - 129 total occurrences
- Project Planning {not in the top 60}
  - 12 co-occurrences with ‘metrics’
  - 22 total occurrences

Carnegie Mellon
Software Engineering Institute

INSPEC Project Management Top 60



INSPEC Project Management Relations



Process Management: Comparison

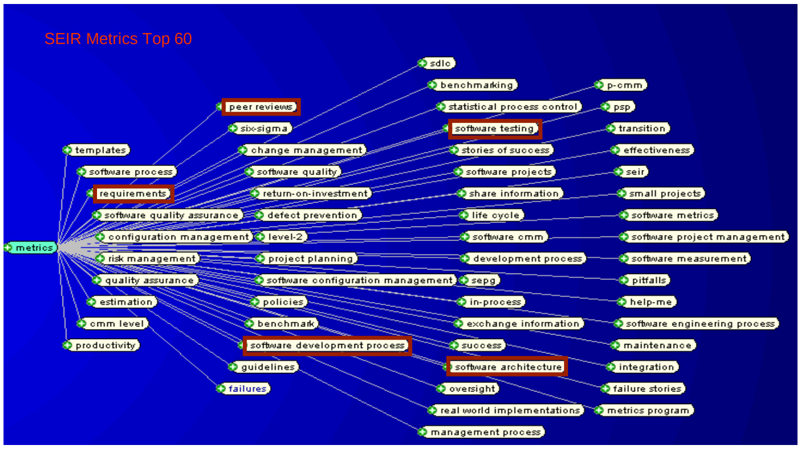
Comparison of co-occurrences

- SEIR
  - Top 60: Project planning, estimation & risk management are frequently associated with each other
  - All 3 also with software project, change management, configuration management, quality assurance, requirements, peer review & defect prevention
- INSPEC
  - Top 60: Software cost estimation is associated with risk management & project planning … but project planning is not associated with risk management
  - Al 3 also with software process improvement

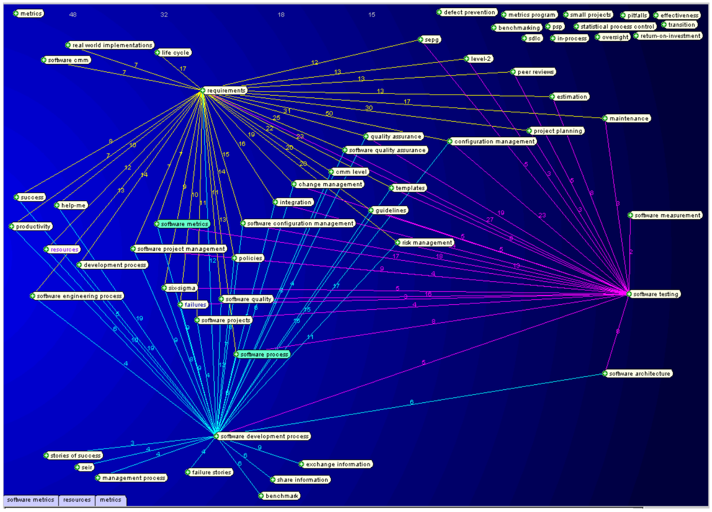
Engineering: SEIR

- Requirements (but not ‘development’ or ‘management’ …)
  - 62 co-occurrences with ‘metrics’
  - 787 total occurrences
- Peer Review (but not ‘validation’ or ‘verification’)
  - 28 co-occurrences with ‘metrics’
  - 206 total occurrences
- Software Testing
  - 20 co-occurrences with ‘metrics’
  - 404 total occurrences
- Software development process (but not ‘technical solution’ or ‘product integration’)
  - 20 co-occurrences with ‘metrics’
  - 287 total occurrences
- Software architecture (20 211)
  - 20 co-occurrences with ‘metrics’
  - 211 total occurrences

SEIR Engineering: Top 60

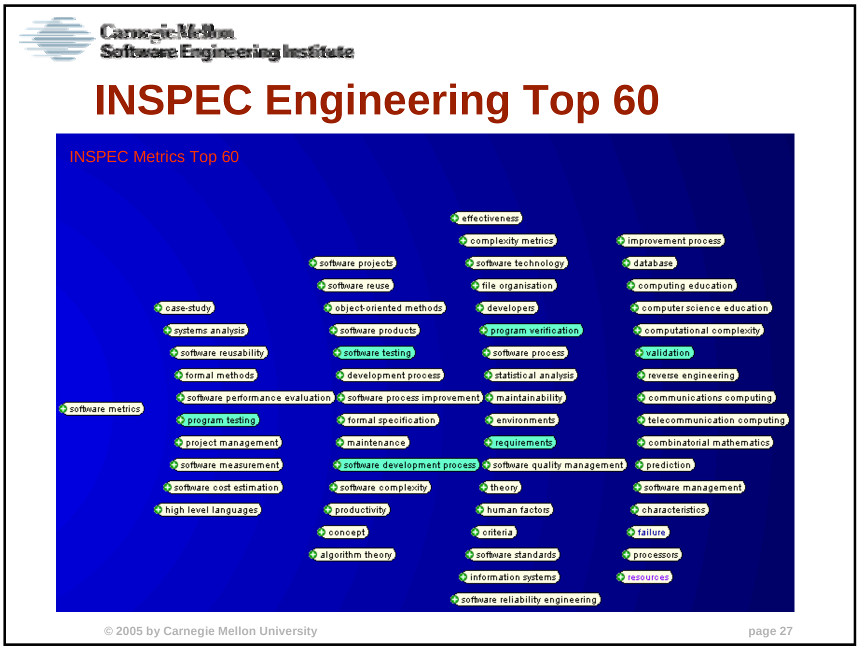


SEIR Engineering Relations



Engineering: INSPEC$_1$

- 'Software' &/or 'Program' Testing
  - 479 co-occurrences with 'metrics'
  - 878 total occurrences
- Software development process (but not 'technical solution' or 'product integration')
  - 168 co-occurrences with 'metrics'
  - 224 total occurrences
- Requirements (but not 'development' or 'management' …)
  - 148 co-occurrences with 'metrics'
  - 750 total occurrences
- Program verification
  - 133 co-occurrences with 'metrics'
  - 240 total occurrences
- Validation
  - 101 co-occurrences with 'metrics'
  - 304 total occurrences

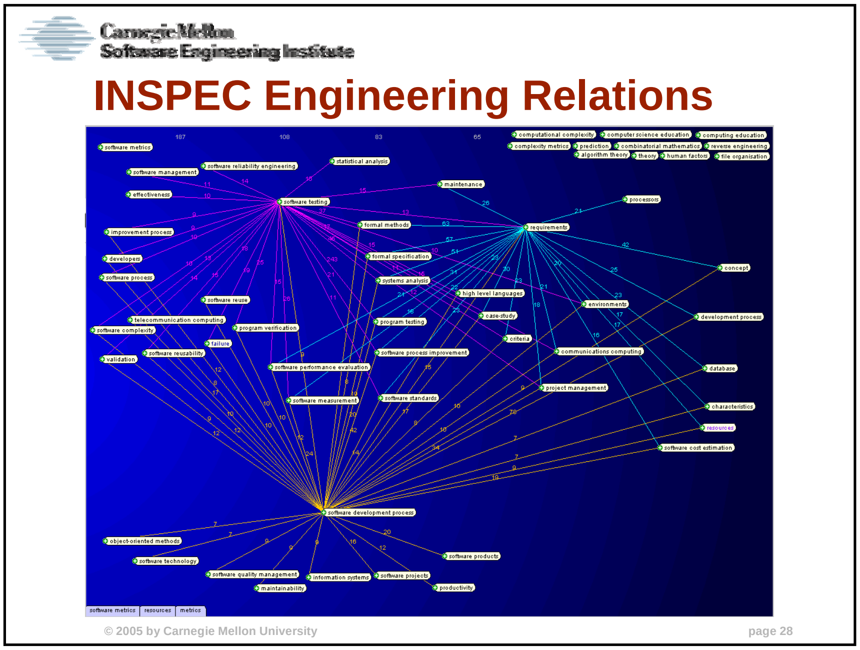


Carnegie Mellon
Software Engineering Institute

INSPEC Engineering Top 60

© 2005 by Carnegie Mellon University

page 27



Carnegie Mellon
Software Engineering Institute

INSPEC Engineering Relations

© 2005 by Carnegie Mellon University

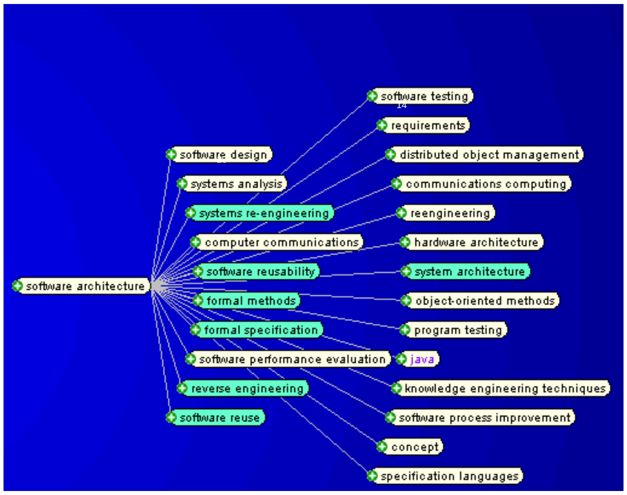
page 28

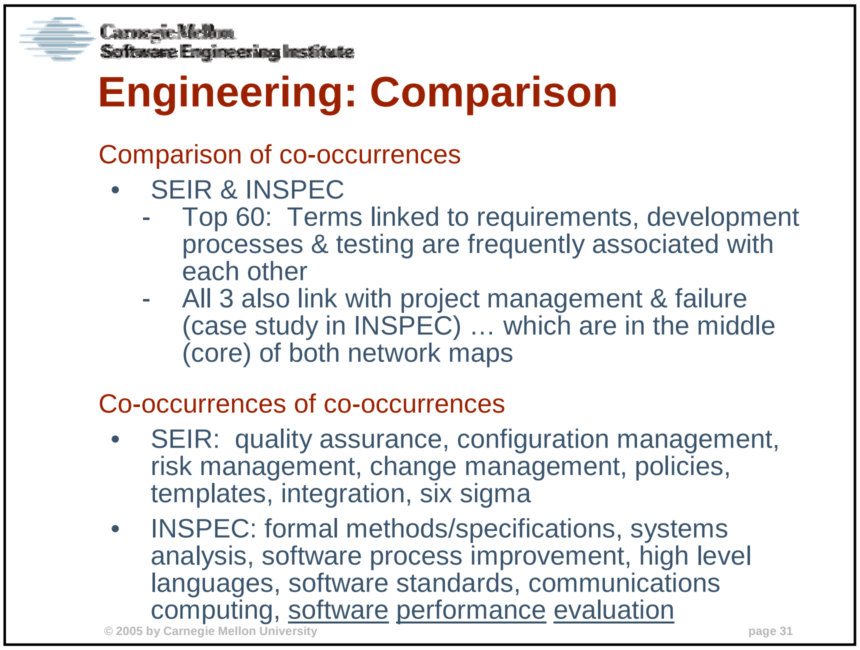
Engineering: INSPEC$_2$

All top 60 & co-occurring with ‘Software Architecture…

- Formal methods &/or specification
  - 455 co-occurrences with ‘metrics’
  - 763 total occurrences
- Software reusability &/or reuse
  - 314 co-occurrences with ‘metrics’
  - 481 total occurrences
- Reverse engineering &/or systems re-engineering
  - 94 co-occurrences with ‘metrics’
  - 138 total occurrences
- Software architecture
  - 135 co-occurrences with ‘metrics’
  - 356 total occurrences

INSPEC Architecture Top 60





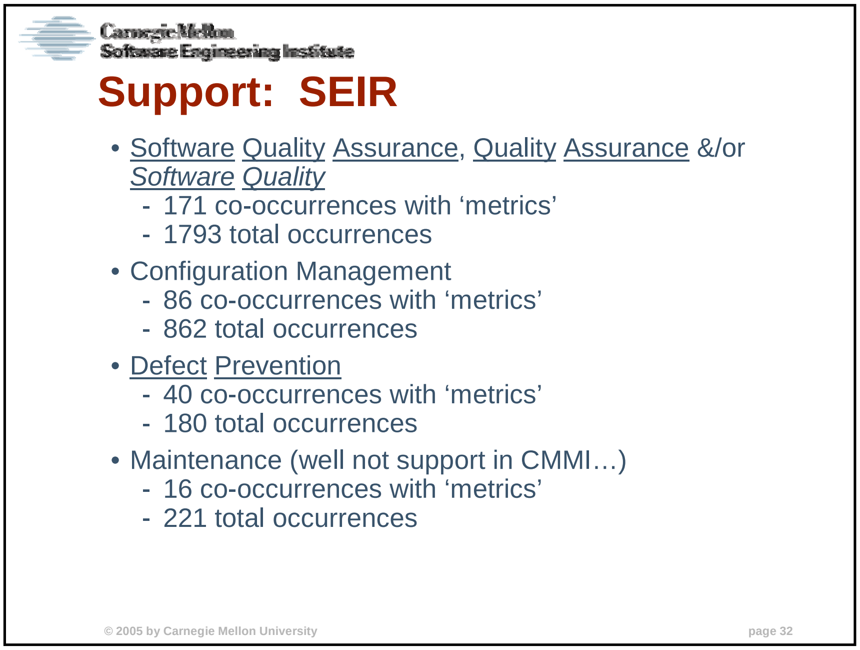
Engineering: Comparison

Comparison of co-occurrences

- SEIR & INSPEC
  - Top 60: Terms linked to requirements, development processes & testing are frequently associated with each other
  - All 3 also link with project management & failure (case study in INSPEC) … which are in the middle (core) of both network maps

Co-occurrences of co-occurrences

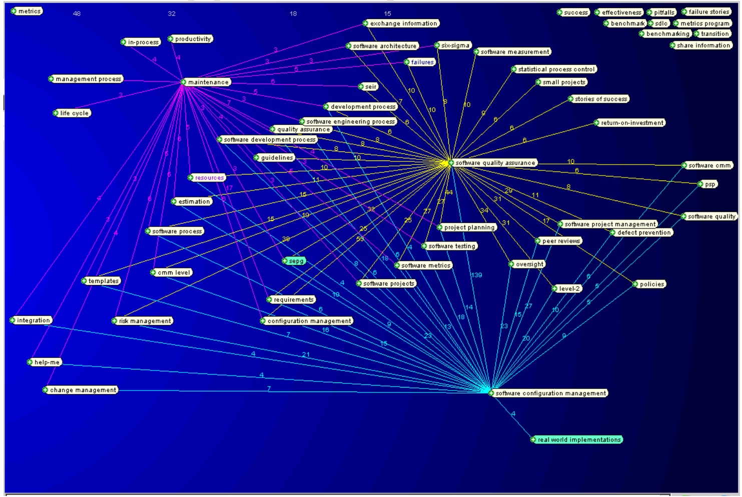
- SEIR: quality assurance, configuration management, risk management, change management, policies, templates, integration, six sigma
- INSPEC: formal methods/specifications, systems analysis, software process improvement, high level languages, software standards, communications computing, software performance evaluation



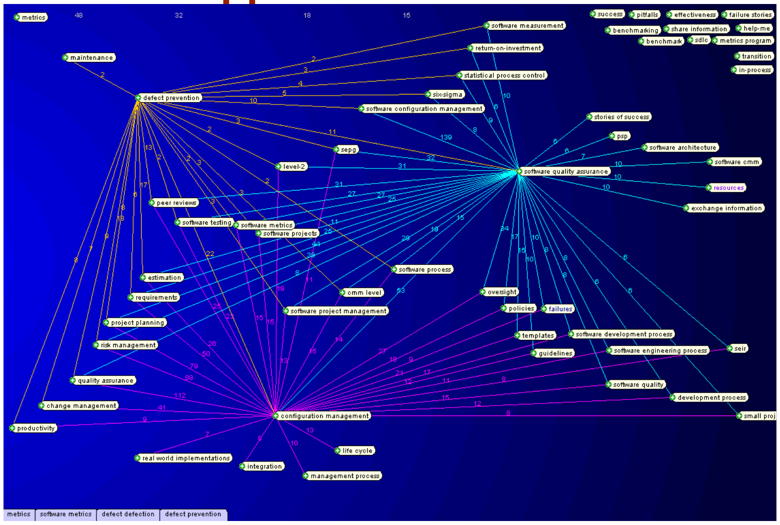
Support: SEIR

- Software Quality Assurance, Quality Assurance &/or *Software Quality*
  - 171 co-occurrences with ‘metrics’
  - 1793 total occurrences
- Configuration Management
  - 86 co-occurrences with ‘metrics’
  - 862 total occurrences
- Defect Prevention
  - 40 co-occurrences with ‘metrics’
  - 180 total occurrences
- Maintenance (well not support in CMMI…)
  - 16 co-occurrences with ‘metrics’
  - 221 total occurrences

SEIR Support: Top 60



SEIR Support Relations

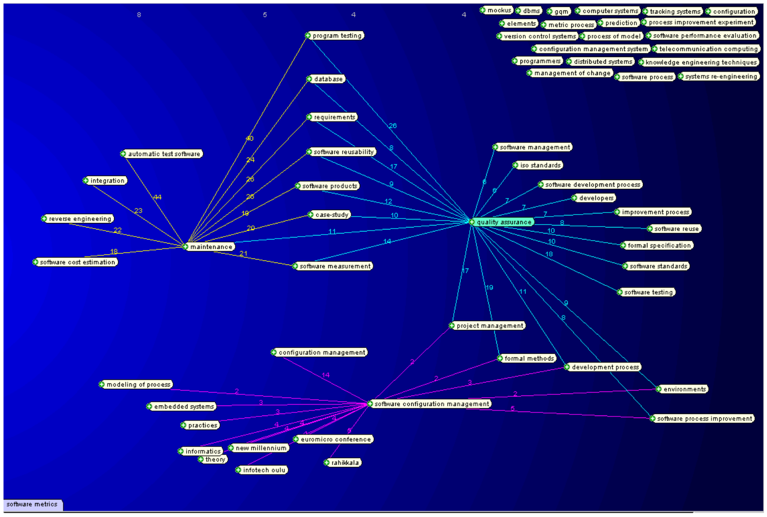


Title
Date

Support: INSPEC

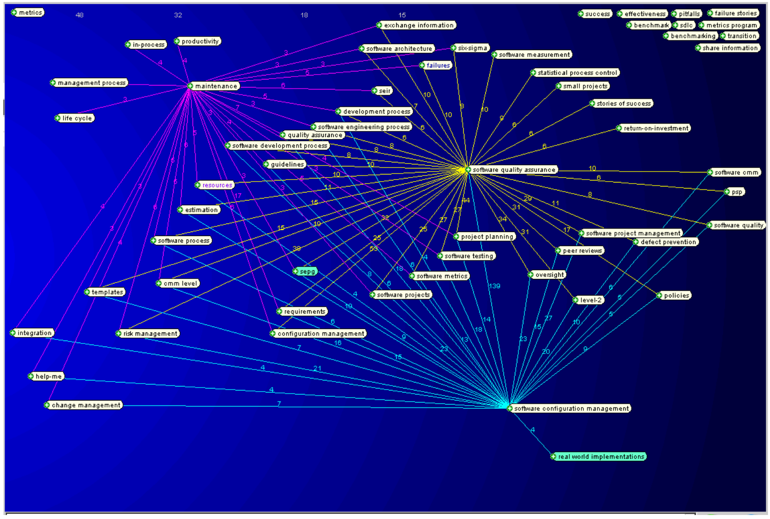
- Software Quality Management, Quality Assurance &/or *Software Quality*
  - 68 co-occurrences with ‘metrics’
  - 910 total occurrences
- Configuration Management (with &/or without ‘software’)
  - 56 co-occurrences with ‘metrics’
  - 104 total occurrences
- Maintenance (well not support in CMMI…)
  - 197 co-occurrences with ‘metrics’
  - 671 total occurrences

INSPEC Support Top 60



Carnegie Mellon
Software Engineering Institute

INSPEC Support Relations

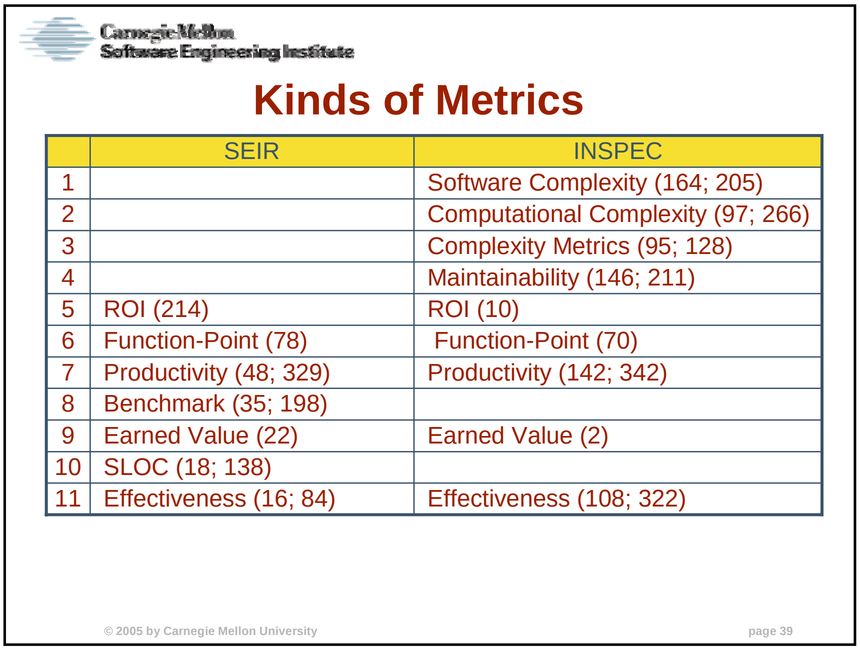


Carnegie Mellon
Software Engineering Institute

Support: Comparison

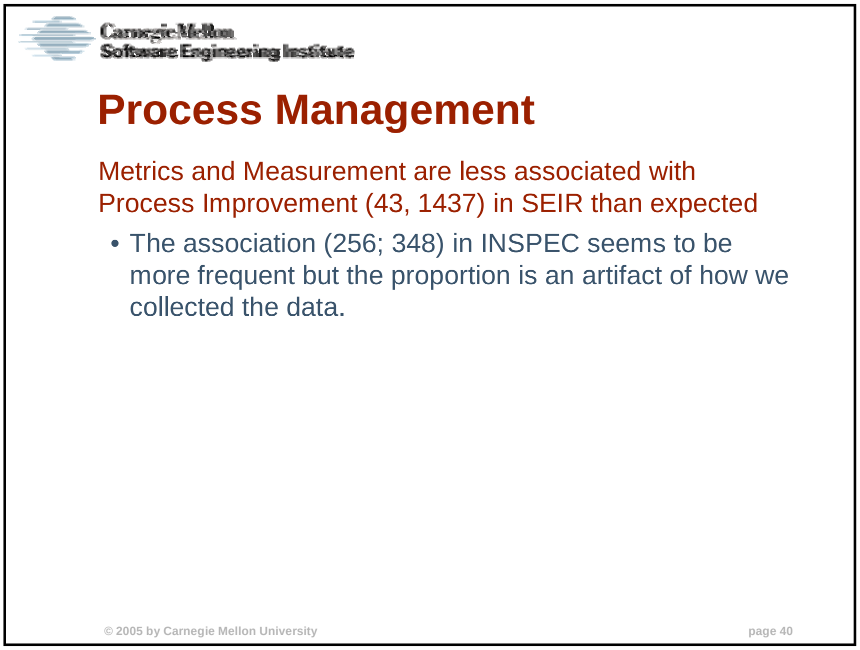
Comparison of co-occurrences

- SEIR & INSPEC
  - Top 60: Terms linked to quality assurance, configuration management & maintenance are frequently associated with each other
  - Although the cluster is more central to SEIR
- SEIR only
  - Defect prevention
- Neither source
  - DAR, OEI, CAR
  - Terms explicitly related to measurement and analysis processes *per se* *
  - (Of course, qualities to be measured and types of metrics are there)



Kinds of Metrics

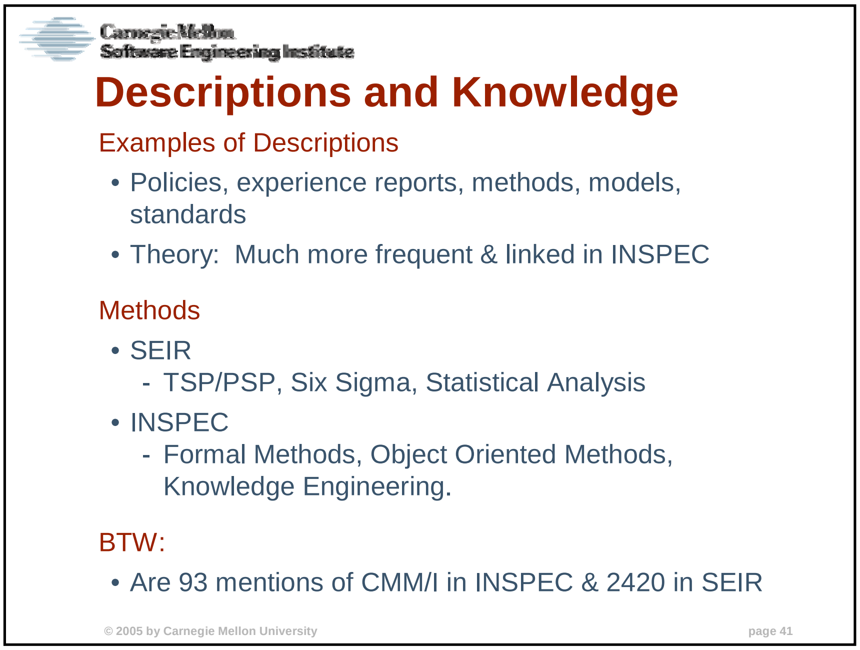
	SEIR	INSPEC
1		Software Complexity (164; 205)
2		Computational Complexity (97; 266)
3		Complexity Metrics (95; 128)
4		Maintainability (146; 211)
5	ROI (214)	ROI (10)
6	Function-Point (78)	Function-Point (70)
7	Productivity (48; 329)	Productivity (142; 342)
8	Benchmark (35; 198)	
9	Earned Value (22)	Earned Value (2)
10	SLOC (18; 138)	
11	Effectiveness (16; 84)	Effectiveness (108; 322)



Process Management

Metrics and Measurement are less associated with Process Improvement (43, 1437) in SEIR than expected

- The association (256; 348) in INSPEC seems to be more frequent but the proportion is an artifact of how we collected the data.



Carnegie Mellon
Software Engineering Institute

Descriptions and Knowledge

Examples of Descriptions

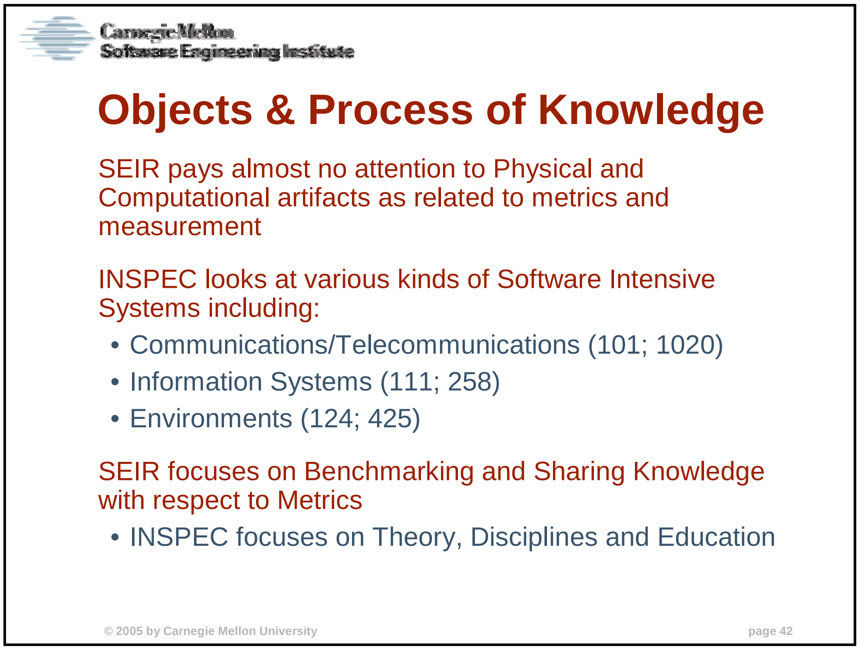
- Policies, experience reports, methods, models, standards
- Theory: Much more frequent & linked in INSPEC

Methods

- SEIR
  - TSP/PSP, Six Sigma, Statistical Analysis
- INSPEC
  - Formal Methods, Object Oriented Methods, Knowledge Engineering.

BTW:

- Are 93 mentions of CMM/I in INSPEC & 2420 in SEIR



Carnegie Mellon
Software Engineering Institute

Objects & Process of Knowledge

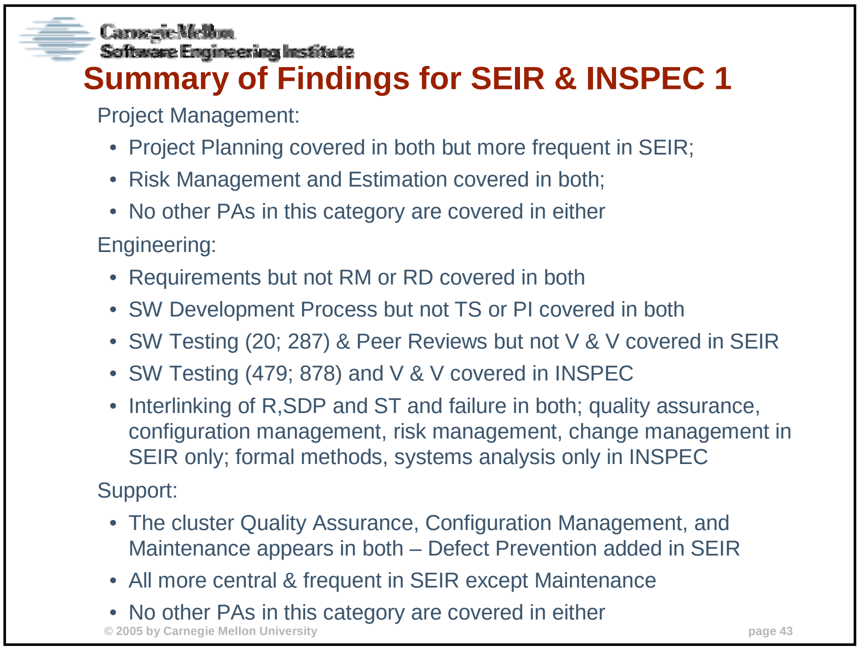
SEIR pays almost no attention to Physical and Computational artifacts as related to metrics and measurement

INSPEC looks at various kinds of Software Intensive Systems including:

- Communications/Telecommunications (101; 1020)
- Information Systems (111; 258)
- Environments (124; 425)

SEIR focuses on Benchmarking and Sharing Knowledge with respect to Metrics

- INSPEC focuses on Theory, Disciplines and Education



Summary of Findings for SEIR & INSPEC 1

Project Management:

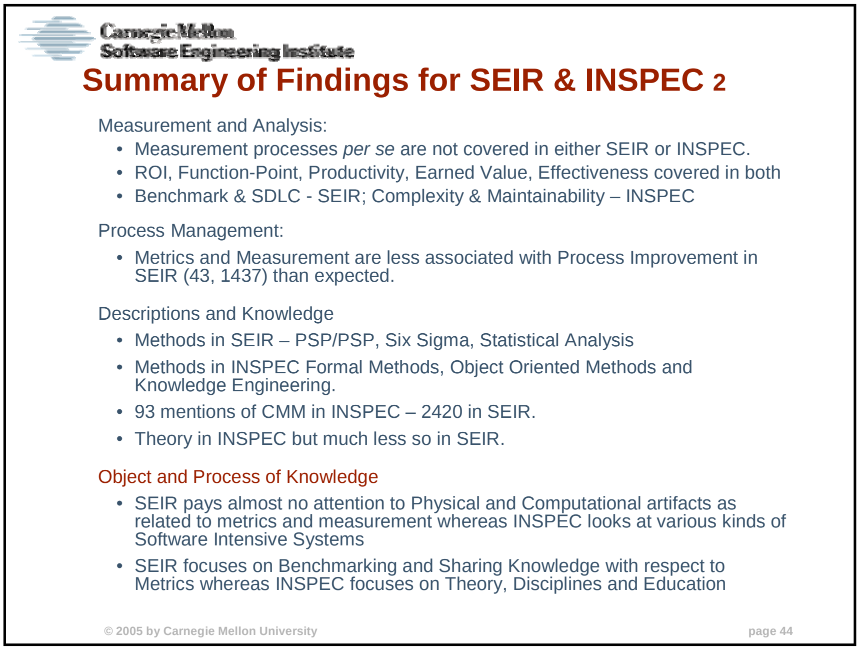
- Project Planning covered in both but more frequent in SEIR;
- Risk Management and Estimation covered in both;
- No other PAs in this category are covered in either

Engineering:

- Requirements but not RM or RD covered in both
- SW Development Process but not TS or PI covered in both
- SW Testing (20; 287) & Peer Reviews but not V & V covered in SEIR
- SW Testing (479; 878) and V & V covered in INSPEC
- Interlinking of R,SDP and ST and failure in both; quality assurance, configuration management, risk management, change management in SEIR only; formal methods, systems analysis only in INSPEC

Support:

- The cluster Quality Assurance, Configuration Management, and Maintenance appears in both – Defect Prevention added in SEIR
- All more central & frequent in SEIR except Maintenance
- No other PAs in this category are covered in either



Summary of Findings for SEIR & INSPEC 2

Measurement and Analysis:

- Measurement processes *per se* are not covered in either SEIR or INSPEC.
- ROI, Function-Point, Productivity, Earned Value, Effectiveness covered in both
- Benchmark & SDLC - SEIR; Complexity & Maintainability – INSPEC

Process Management:

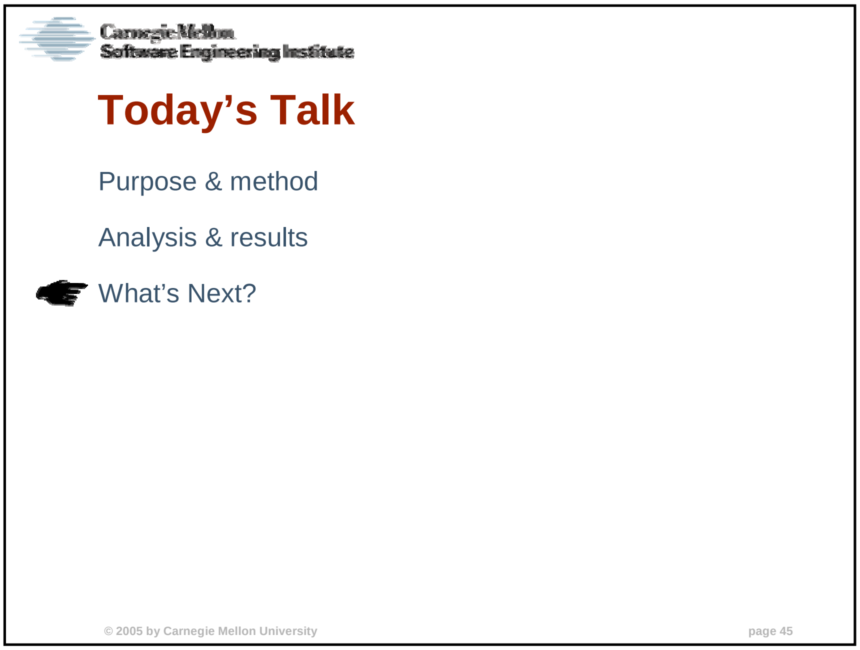
- Metrics and Measurement are less associated with Process Improvement in SEIR (43, 1437) than expected.

Descriptions and Knowledge

- Methods in SEIR – PSP/PSP, Six Sigma, Statistical Analysis
- Methods in INSPEC Formal Methods, Object Oriented Methods and Knowledge Engineering.
- 93 mentions of CMM in INSPEC – 2420 in SEIR.
- Theory in INSPEC but much less so in SEIR.

Object and Process of Knowledge

- SEIR pays almost no attention to Physical and Computational artifacts as related to metrics and measurement whereas INSPEC looks at various kinds of Software Intensive Systems
- SEIR focuses on Benchmarking and Sharing Knowledge with respect to Metrics whereas INSPEC focuses on Theory, Disciplines and Education



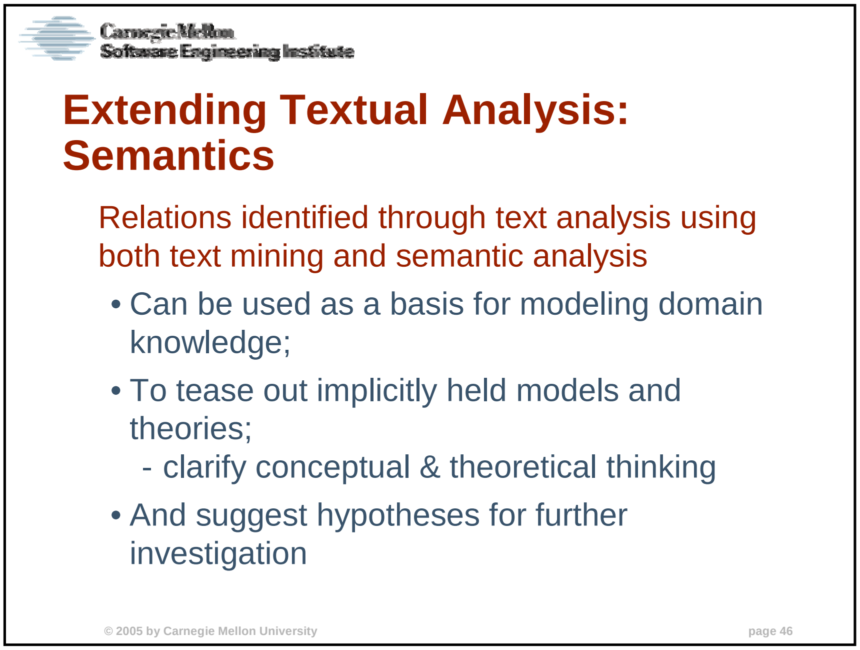
Carnegie Mellon
Software Engineering Institute

Today's Talk

Purpose & method

Analysis & results

☛ What's Next?



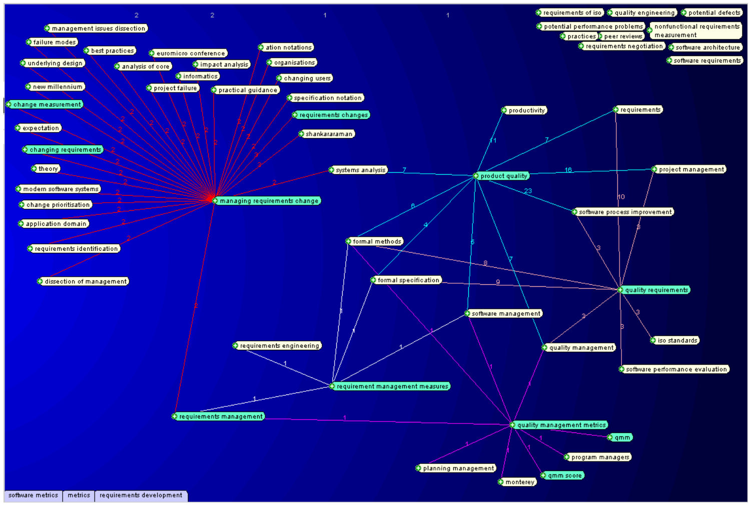
Carnegie Mellon
Software Engineering Institute

Extending Textual Analysis: Semantics

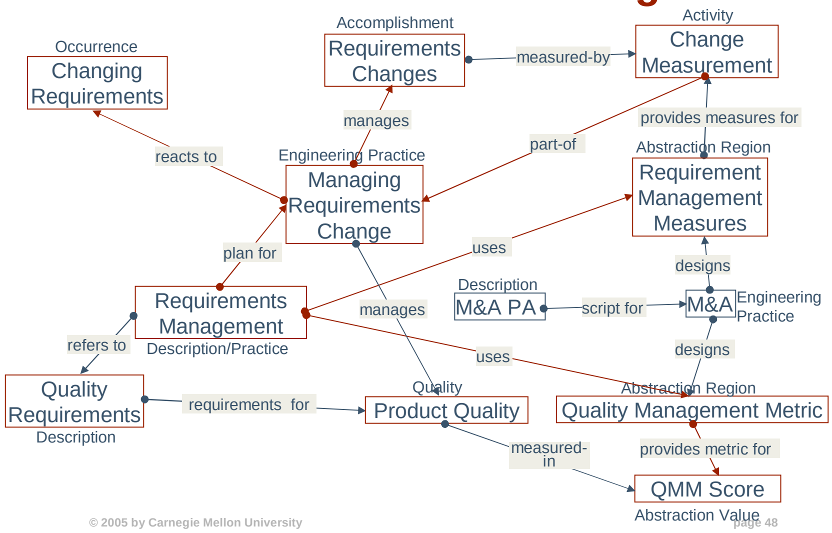
Relations identified through text analysis using both text mining and semantic analysis

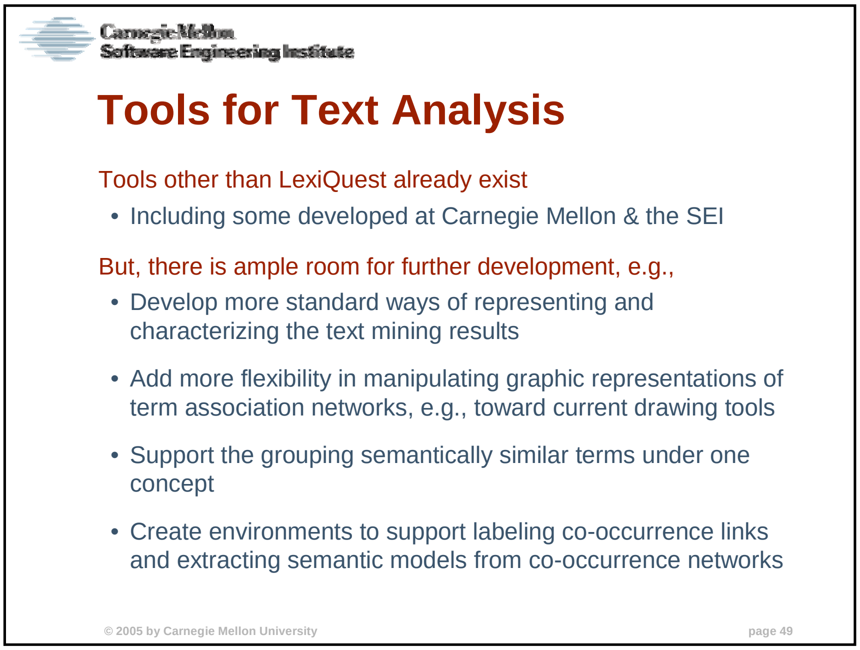
- Can be used as a basis for modeling domain knowledge;
- To tease out implicitly held models and theories;
  - clarify conceptual & theoretical thinking
- And suggest hypotheses for further investigation

Text Mined Relations



A Basis for Domain Modeling





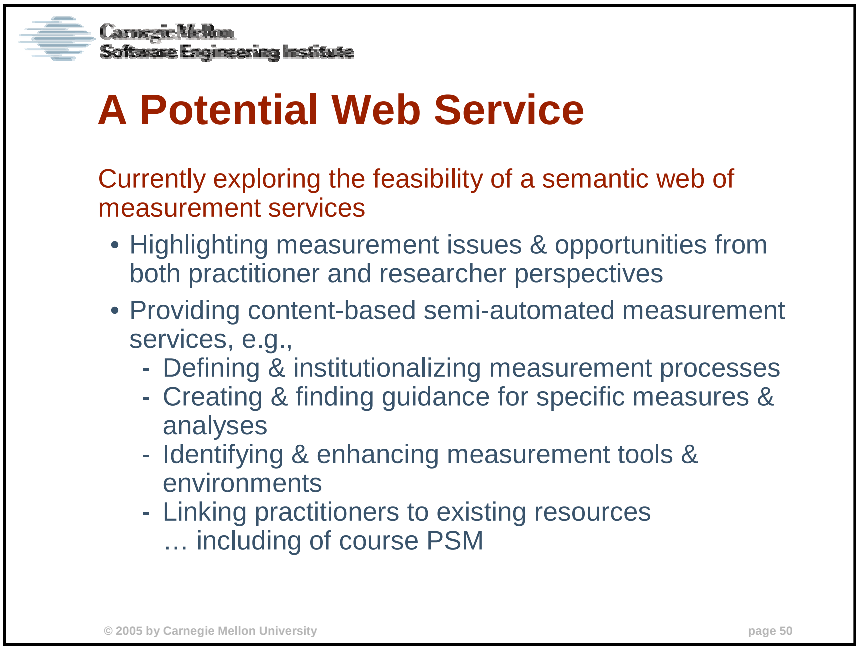
Tools for Text Analysis

Tools other than LexiQuest already exist

- Including some developed at Carnegie Mellon & the SEI

But, there is ample room for further development, e.g.,

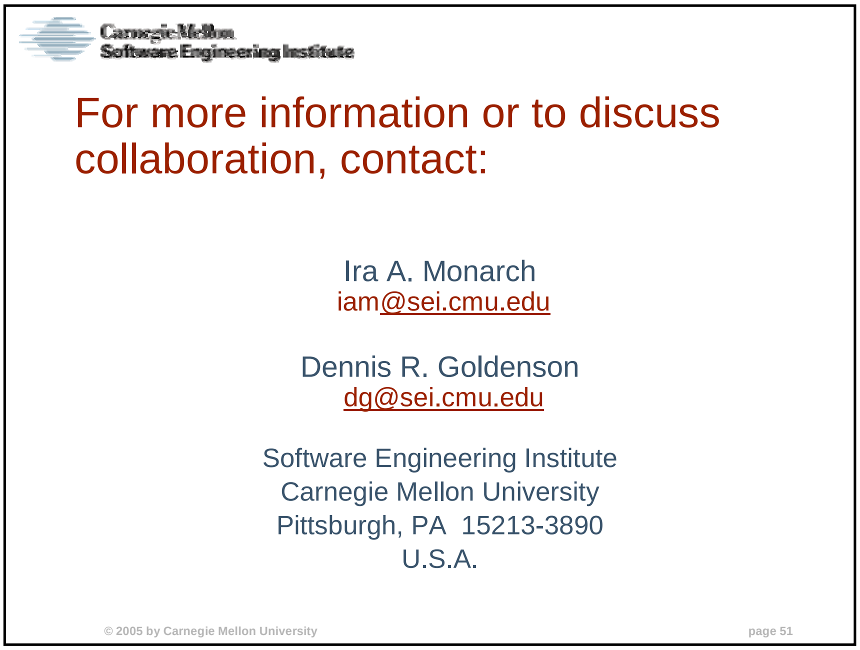
- Develop more standard ways of representing and characterizing the text mining results
- Add more flexibility in manipulating graphic representations of term association networks, e.g., toward current drawing tools
- Support the grouping semantically similar terms under one concept
- Create environments to support labeling co-occurrence links and extracting semantic models from co-occurrence networks



A Potential Web Service

Currently exploring the feasibility of a semantic web of measurement services

- Highlighting measurement issues & opportunities from both practitioner and researcher perspectives
- Providing content-based semi-automated measurement services, e.g.,
  - Defining & institutionalizing measurement processes
  - Creating & finding guidance for specific measures & analyses
  - Identifying & enhancing measurement tools & environments
  - Linking practitioners to existing resources … including of course PSM

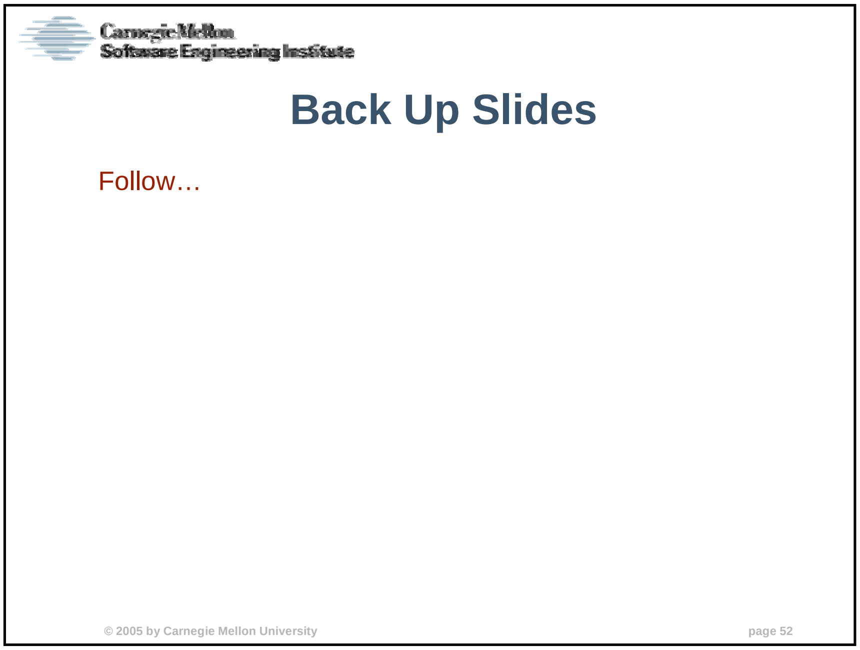


For more information or to discuss collaboration, contact:

Ira A. Monarch
iam@sei.cmu.edu

Dennis R. Goldenson
dg@sei.cmu.edu

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3890
U.S.A.



Back Up Slides

Follow…

Text Mining: An Informetric Technique

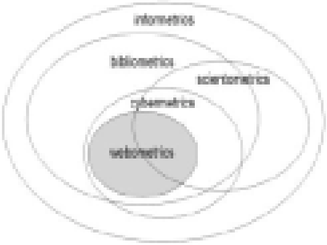
Informetrics: covers Bibliometrics, Scientometrics, Cybermetrics and Webometrics

Bibliometrics: the quantitative analysis of publications for determining intellectual influence, interdisciplinarity, research fronts, trends in subjects pursued, and top producing journals and authors

Scientometrics: bibliometrics focused upon monitoring sciences, both applied and pure, and technology

Cybermetrics: the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the *whole* Internet drawing on informetric approaches

Webometrics: Cybermetrics restricted to the Web



Adopted from Lennart Björneborn and Peter Ingwersen, "Toward a Basic Framework for Webometrics," *JASIS,* December, 2004,

Jean-Pierre V. M. Hérubel, Historical Bibliometrics: Its Purpose and Significance to the History of Disciplines, *Libraries and Culture*, summer, 2004.

Top-Down Upper-Level Categories

drg3

Top-down categories are ones not driven by the results of text-mining.

Particular – aka entity, anything that can be interpreted as an individual in the texts being analyzed.

- Perdurant – aka occurrence, extends in time by accumulating different temporal parts that at any time may not be present
- Endurant – occurs as a whole through time being able to have incompatible properties at different times and still be the same whole
- Quality – what inheres in entities that can be perceived or measured (shapes, colors, weights, lengths)
- Abstraction – aka abstract entities, do not have spatial or temporal parts and may be quality regions (shades of color, measurement units)

Relation – What links one particular to another via such relations as part-of, participant-in, location-of, successor-of, referenced-by or required-by, etc.

Slide 54

drg3 Backup only: This will blow the audience away.

We need to first give them a few high level results, or at least questions to pique their interest.

Dennis R. Goldenson, 7/15/2005