London 12-14th September 2017

Identifying High-Frequency Component Failure using Text-Mining Techniques

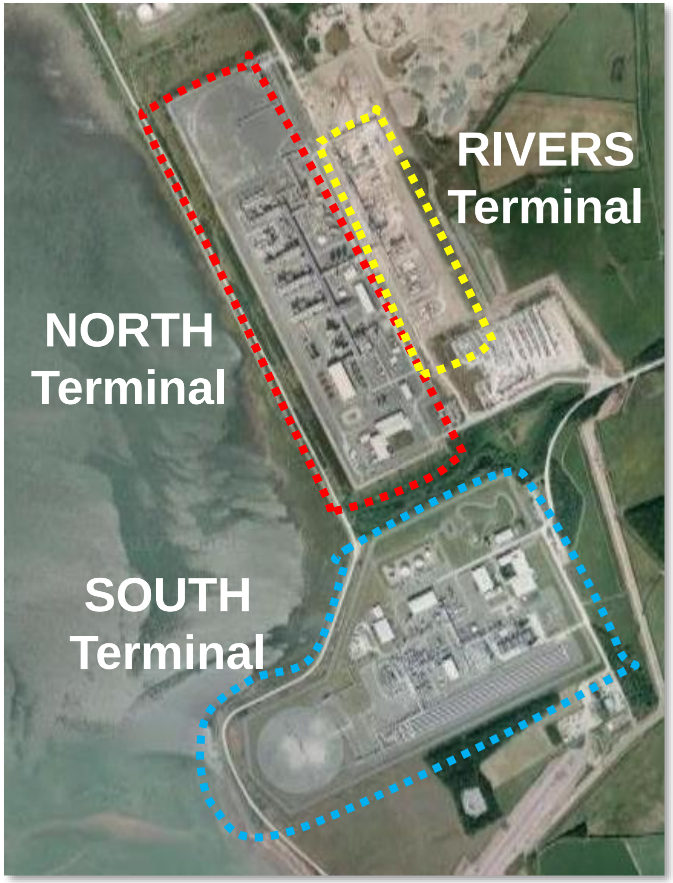
Timothy Wong – Centrica plc

centrica

- We are an energy and services company. Everything we do is focused on satisfying the changing needs of our customers.

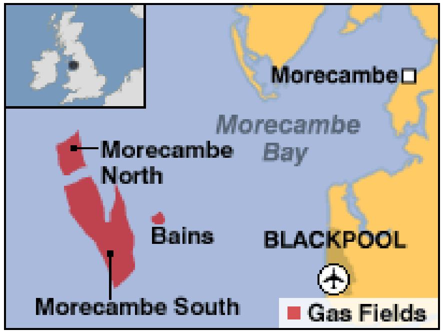
centrica Scottish Gas British Gas Bord Gáis Energy HIVE DYNO Direct Energy Local Heroes

Morecambe Terminals

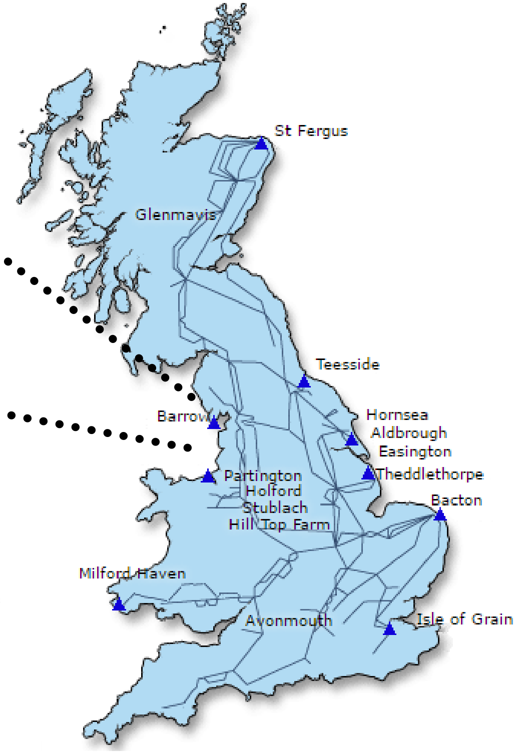


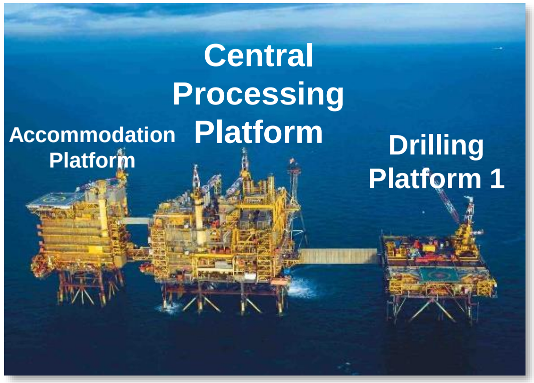
- Morecambe Terminals
  - Maintenance events/incidents are recorded in the repair log
  - Identify recurring vulnerabilities
  - Analytics approach - Text mining algorithms (Natural Language Processing)

North Terminal



Approx. 6% of GB's gas supply

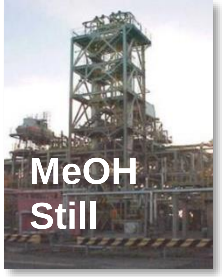




Slugcatcher

Condensate Stabilisation

CO_2 Removal Train

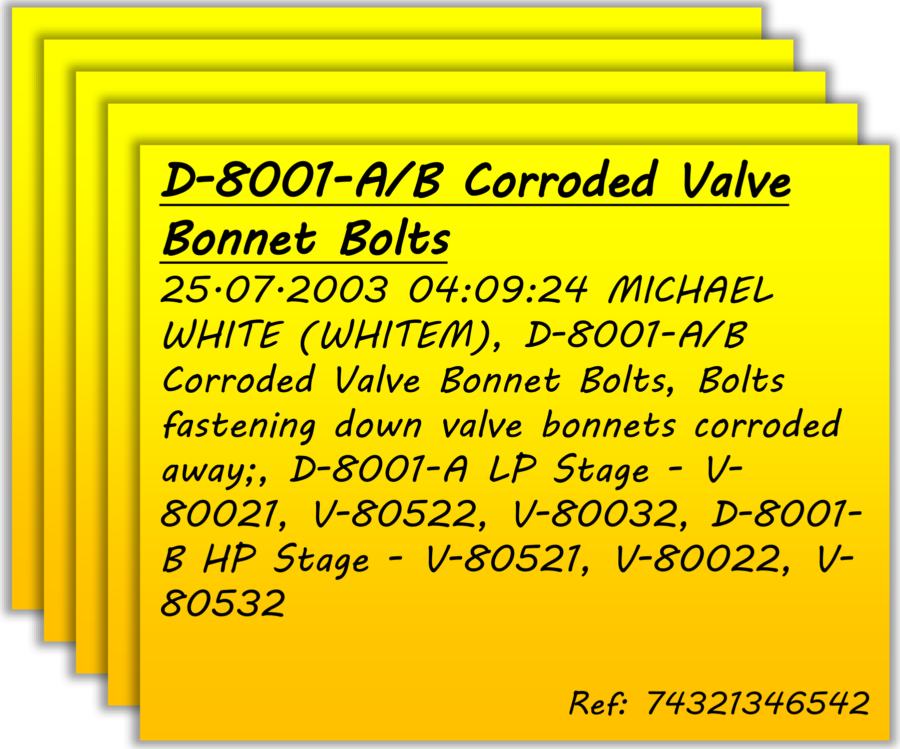

MeOH Still

Product Gas Compressor

Nitrogen Rejection Unit

Dewpoint Control

Source Data



<u>D-8001-A/B Corroded Valve Bonnet Bolts</u>

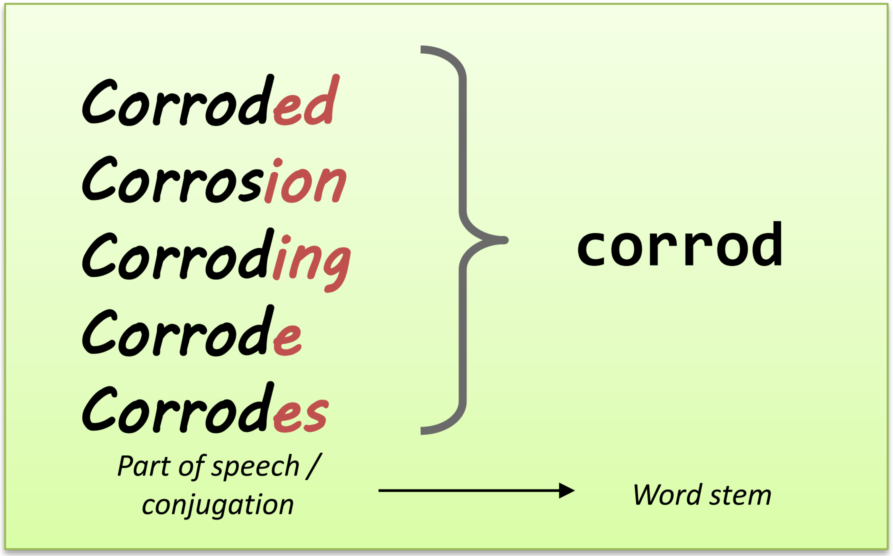
25.07.2003 04:09:24 MICHAEL WHITE (WHITEM), D-8001-A/B Corroded Valve Bonnet Bolts, Bolts fastening down valve bonnets corroded away;, D-8001-A LP Stage - V-80021, V-80522, V-80032, D-8001-B HP Stage - V-80521, V-80022, V-80532

Ref: 74321346542

- Repair log system
  - Unstructured text data
  - Lots of technical abbreviations
    - Component IDs
    - Locations IDs
    - Names, datetime… etc
  - Incomplete syntax
  - Sometimes having typos

Term Extraction (1)

1. Covert to lower case
2. Remove non-alphabets
3. Remove common words
4. Word stemming

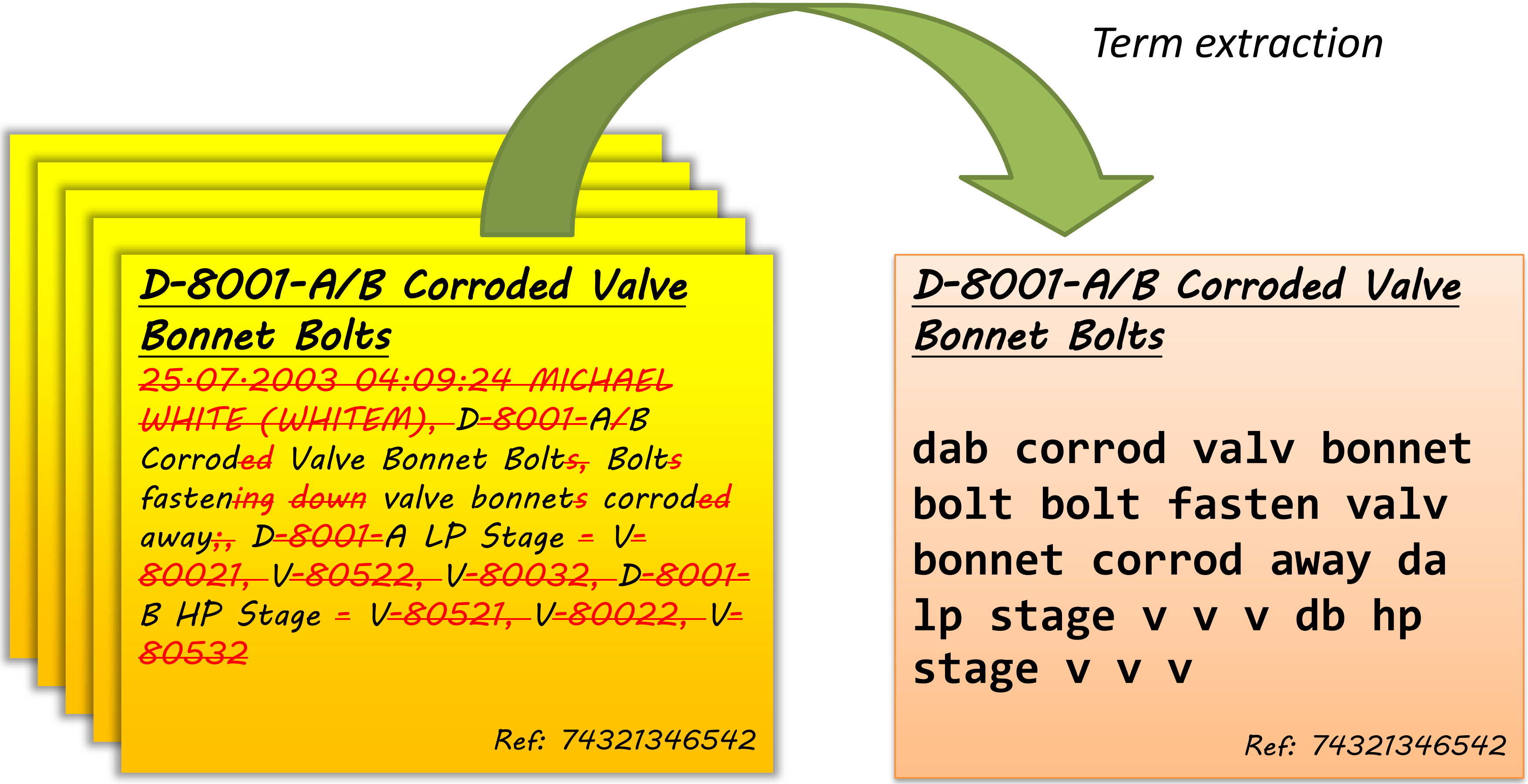


Corroded
Corrosion
Corroding
Corrode
Corrodes

corrod

Part of speech / conjugation → Word stem

Term Extraction (2)



Term extraction

<u>D-8001-A/B Corroded Valve Bonnet Bolts</u>

~~25·07·2003 04:09:24 MICHAEL WHITE (WHITEM),~~ D~~-8001-~~A~~/~~B Corrod~~ed~~ Valve Bonnet Bolt~~s,~~ Bolt~~s~~ fasten~~ing~~ ~~down~~ valve bonnet~~s~~ corrod~~ed~~ away~~;,~~ D~~-8001-~~A LP Stage ~~=~~ V~~-80021,~~ V~~-80522,~~ V~~-80032,~~ D~~-8001-~~B HP Stage ~~=~~ V~~-80521,~~ V~~-80022,~~ V~~-80532~~

Ref: 74321346542

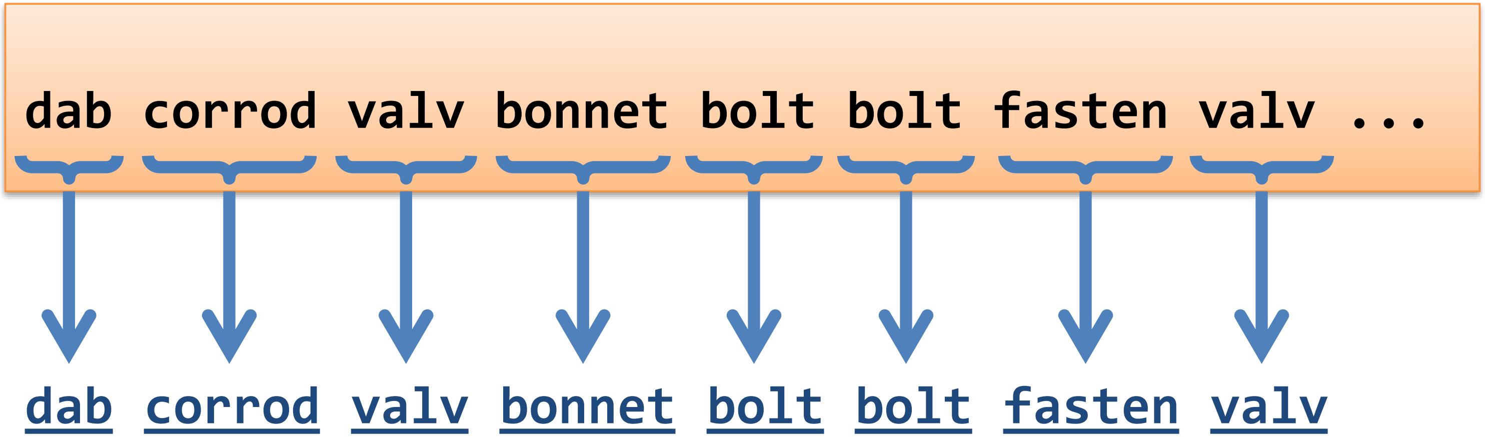
<u>D-8001-A/B Corroded Valve Bonnet Bolts</u>

dab corrod valv bonnet bolt bolt fasten valv bonnet corrod away da lp stage v v v db hp stage v v v

Ref: 74321346542

Term Extraction (3): n-gram model

- Unigram ($n = 1$)



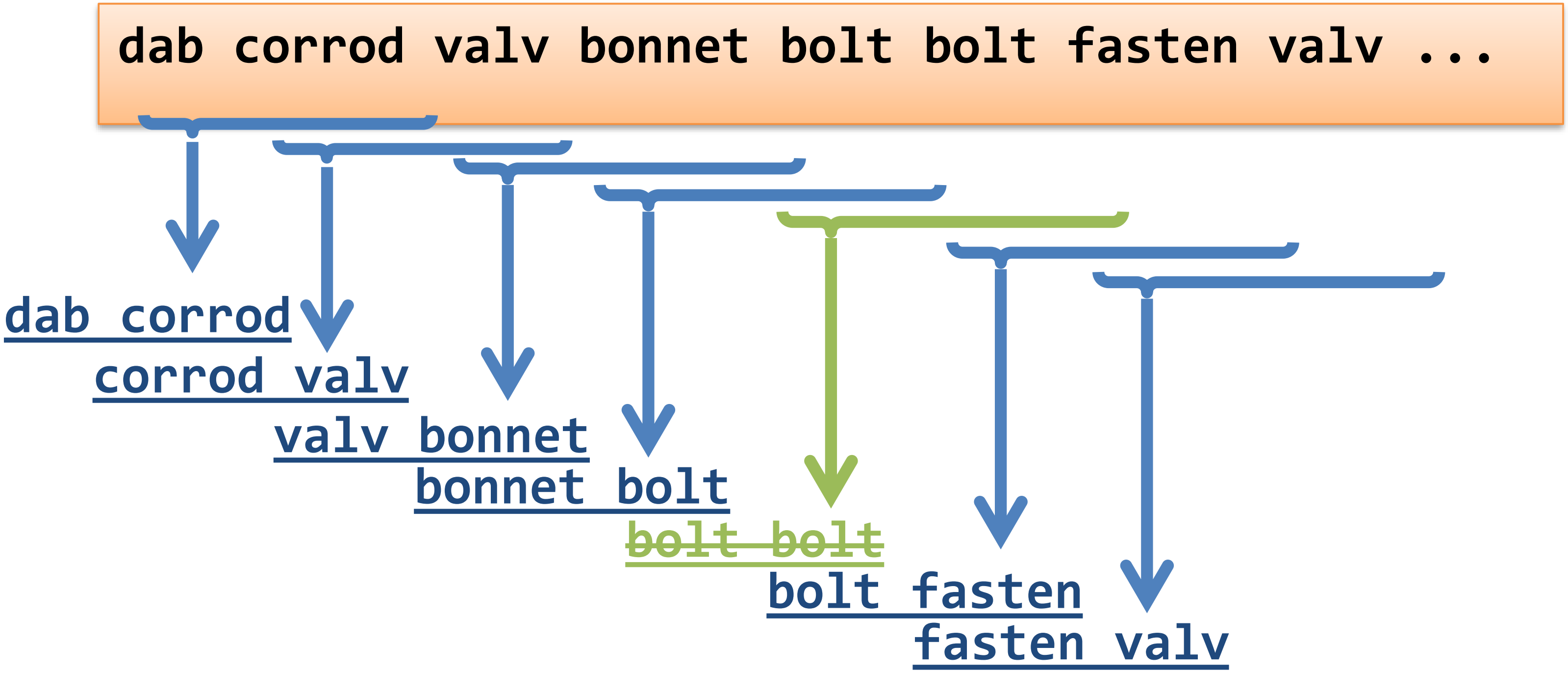
dab corrod valv bonnet bolt bolt fasten valv ...

dab corrod valv bonnet bolt bolt fasten valv

Term Extraction (4): n-gram model

- Bigram ($n = 2$)

  Terms with repetitive words are removed.



dab corrod valv bonnet bolt bolt fasten valv ...

dab corrod

corrod valv

valv bonnet

bonnet bolt

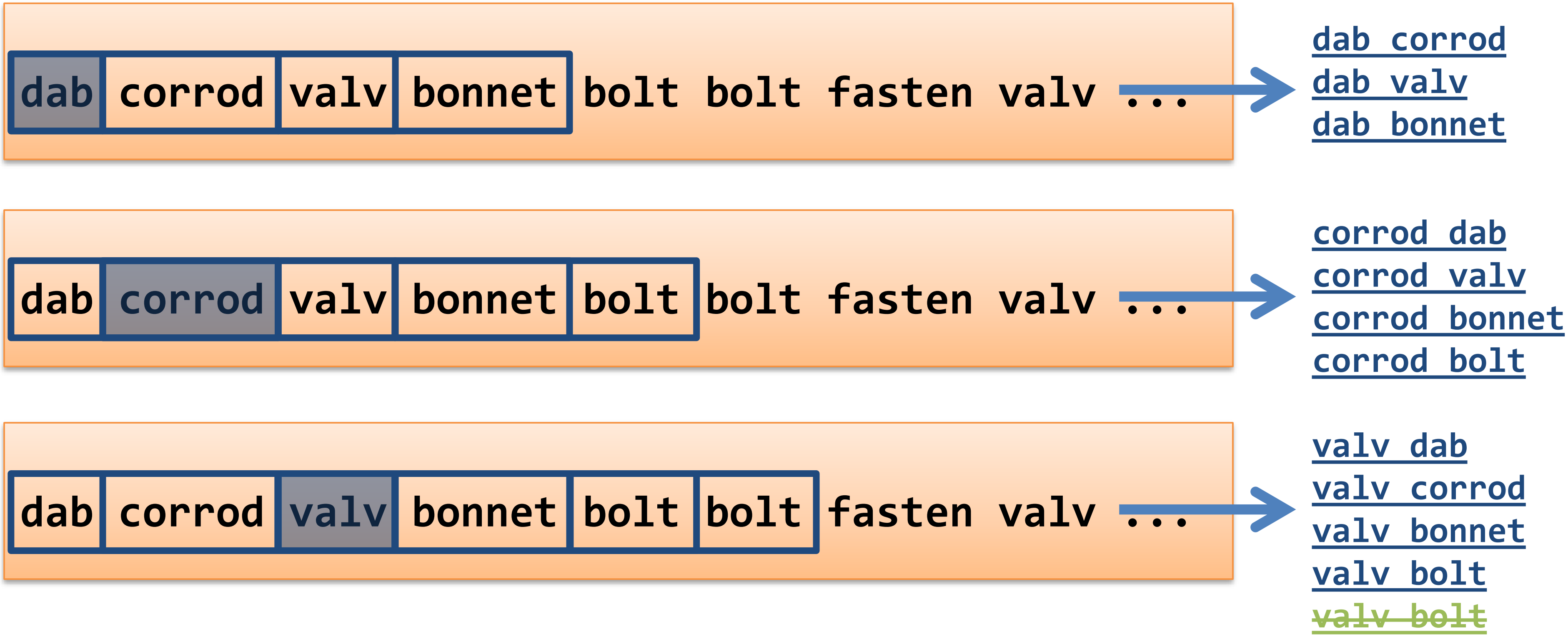
~~bolt bolt~~

bolt fasten

fasten valv

Term Extraction (5): skip-gram model

- Uses a rolling window and takes pair of words



dab corrod valv bonnet bolt bolt fasten valv ...

dab corrod
dab valv
dab bonnet

dab corrod valv bonnet bolt bolt fasten valv ...

corrod dab
corrod valv
corrod bonnet
corrod bolt

dab corrod valv bonnet bolt bolt fasten valv ...

valv dab
valv corrod
valv bonnet
valv bolt
~~valv bolt~~

Information Retrieval: tf-idf scheme (1)

- Term Frequency (tf)

  Reflects occurrence of term t in a given document d

$$tf_{t,d} = \frac{Number\ of\ occurance\ of\ term\ t\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$

- Inverse Document Frequency (idf)

  Reflects occurrence of term in entire corpus D

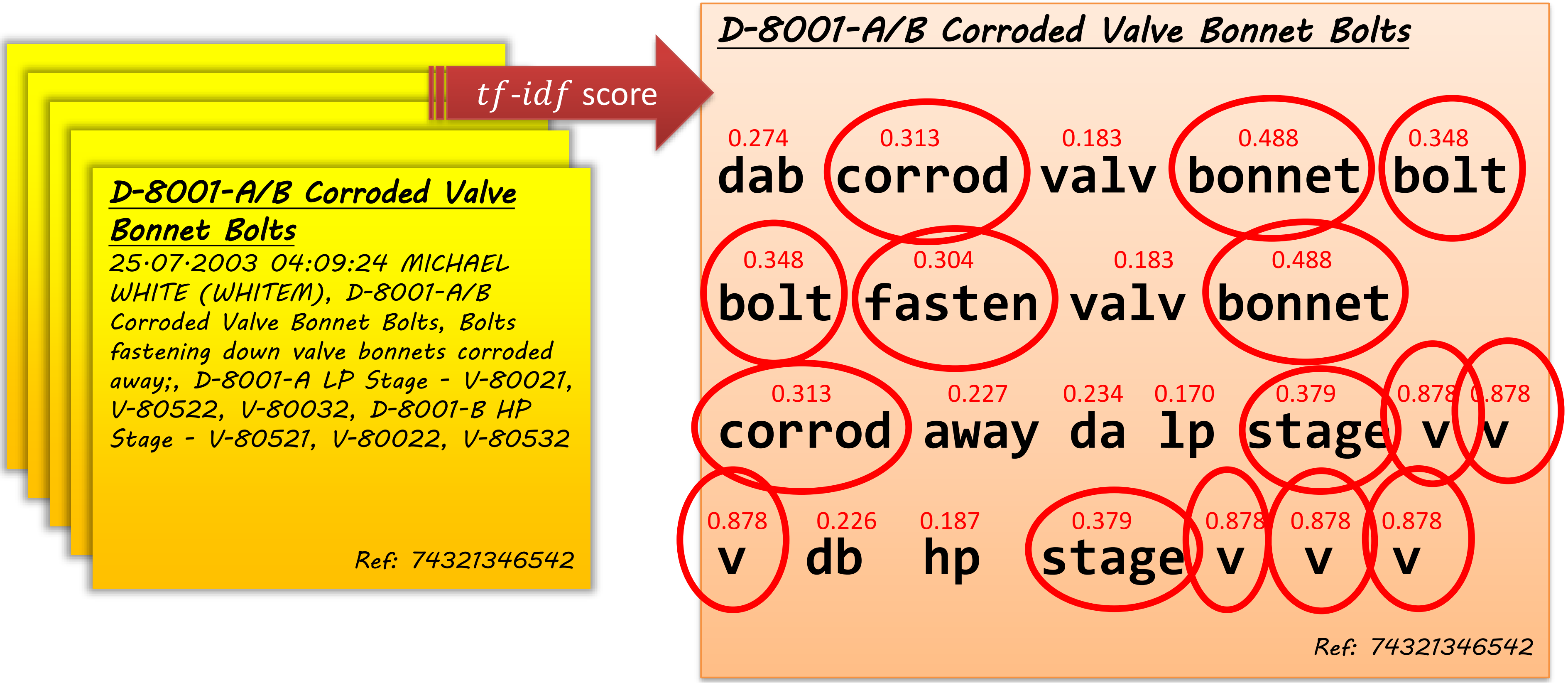
$$idf_{t,D} = \log\left(\frac{Number\ of\ documents\ in\ corpus\ D}{Number\ of\ documents\ having\ term\ t}\right)$$

- Weighted Term Importance ($tf{-}idf$)

$$tfidf = tf_{t,d} \times idf_{t,D}$$

Information Retrieval: tf-idf scheme (2)

- Compute tf-idf for all terms:



D-8001-A/B Corroded Valve Bonnet Bolts
25·07·2003 04:09:24 MICHAEL WHITE (WHITEM), D-8001-A/B Corroded Valve Bonnet Bolts, Bolts fastening down valve bonnets corroded away;, D-8001-A LP Stage - V-80021, V-80522, V-80032, D-8001-B HP Stage - V-80521, V-80022, V-80532

Ref: 74321346542

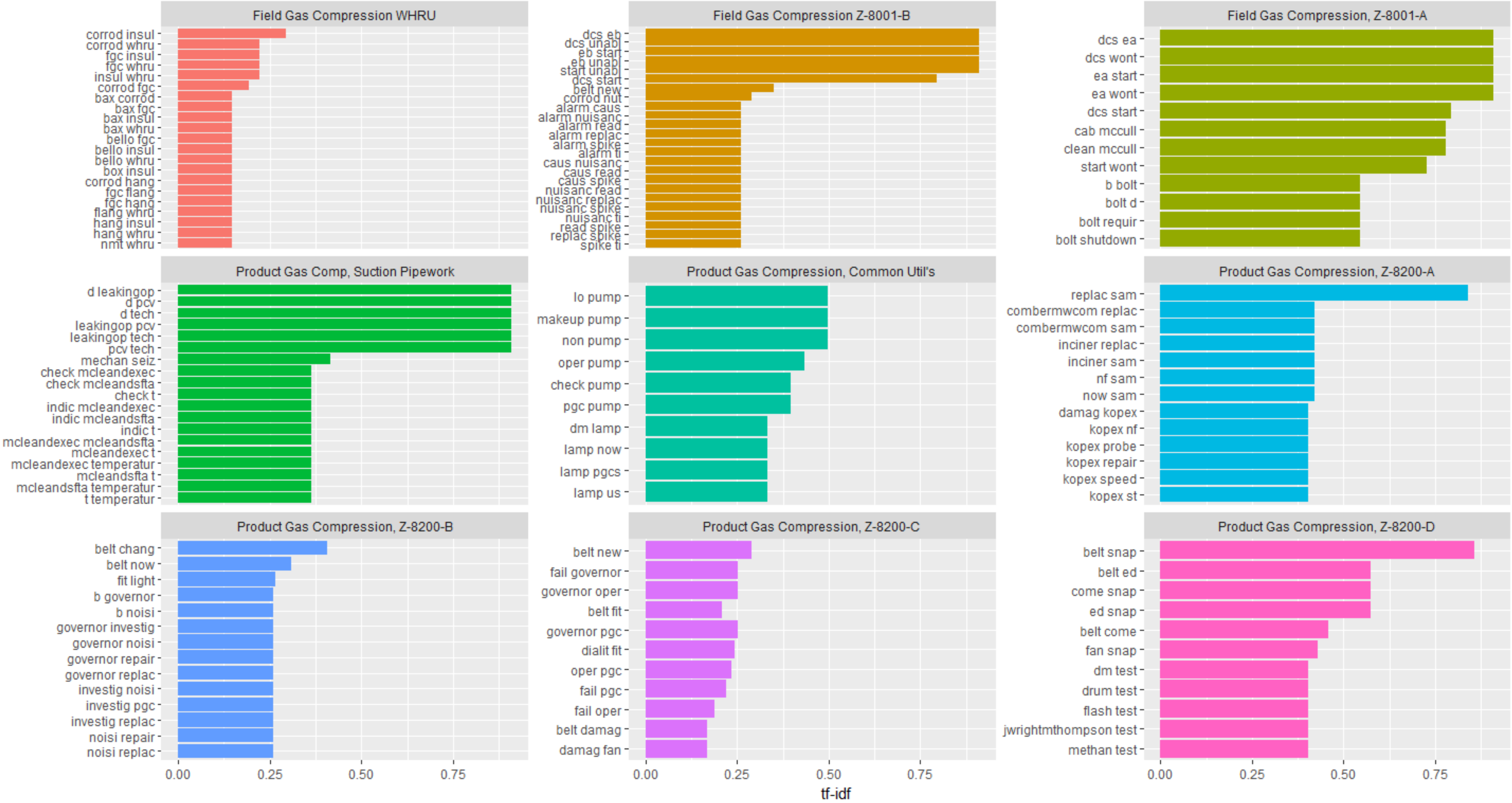
tf-idf score

D-8001-A/B Corroded Valve Bonnet Bolts

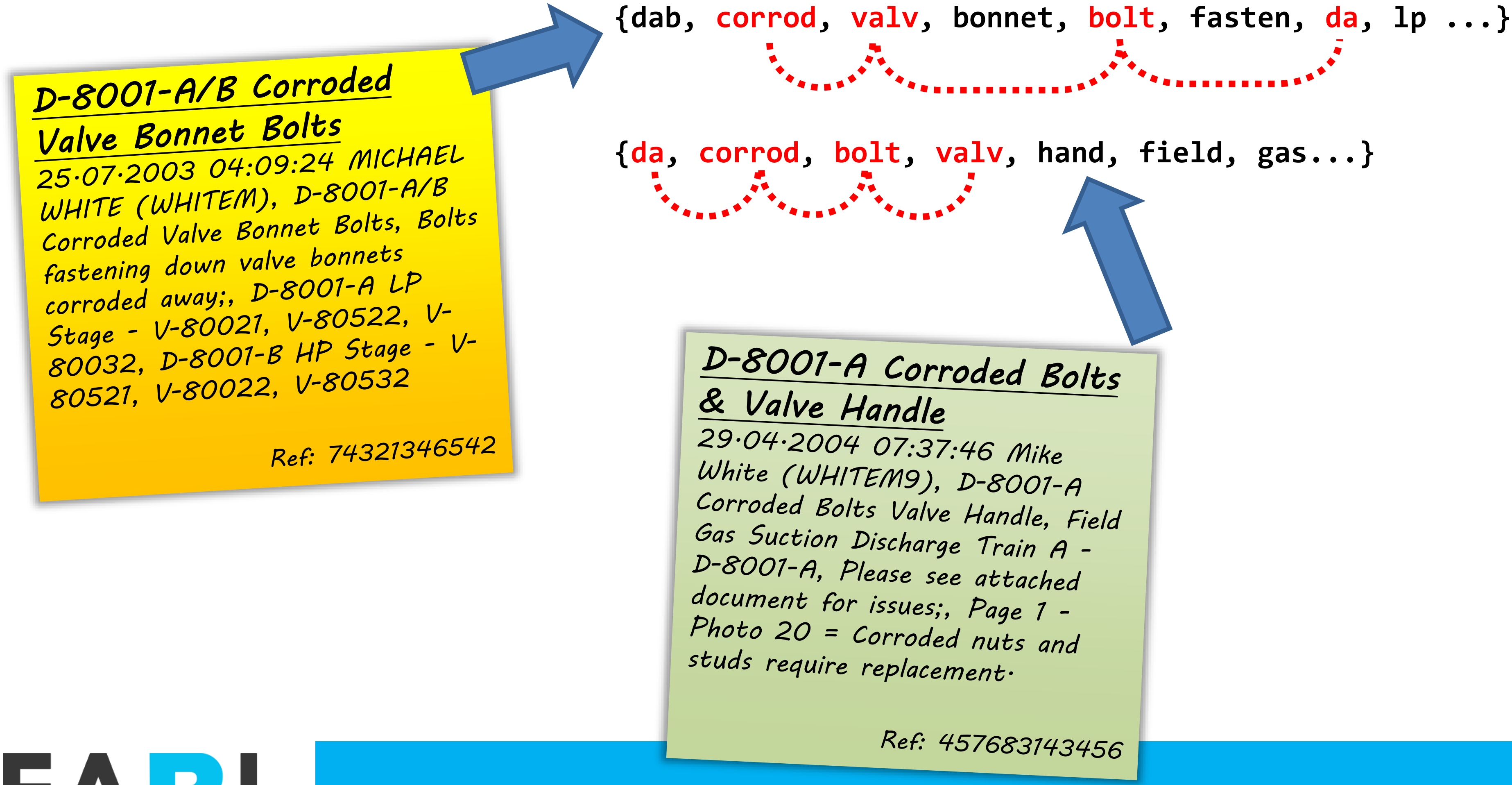
0.274	0.313	0.183	0.488	0.348		
dab	corrod	valv	bonnet	bolt		
0.348	0.304	0.183	0.488			
bolt	fasten	valv	bonnet			
0.313	0.227	0.234	0.170	0.379	0.878	0.878
corrod	away	da	lp	stage	v	v
0.878	0.226	0.187	0.379	0.878	0.878	0.878
v	db	hp	stage	v	v	v

Ref: 74321346542

Information Retrieval: tf-idf scheme (4)



Measuring Correlation (1)



D-8001-A/B Corroded Valve Bonnet Bolts

25.07.2003 04:09:24 MICHAEL WHITE (WHITEM), D-8001-A/B Corroded Valve Bonnet Bolts, Bolts fastening down valve bonnets corroded away;, D-8001-A LP Stage - V-80021, V-80522, V-80032, D-8001-B HP Stage - V-80521, V-80022, V-80532

Ref: 74321346542

{dab, corrod, valv, bonnet, bolt, fasten, da, lp ...}

{da, corrod, bolt, valv, hand, field, gas...}

D-8001-A Corroded Bolts & Valve Handle

29.04.2004 07:37:46 Mike White (WHITEM9), D-8001-A Corroded Bolts Valve Handle, Field Gas Suction Discharge Train A - D-8001-A, Please see attached document for issues;, Page 1 - Photo 20 = Corroded nuts and studs require replacement.

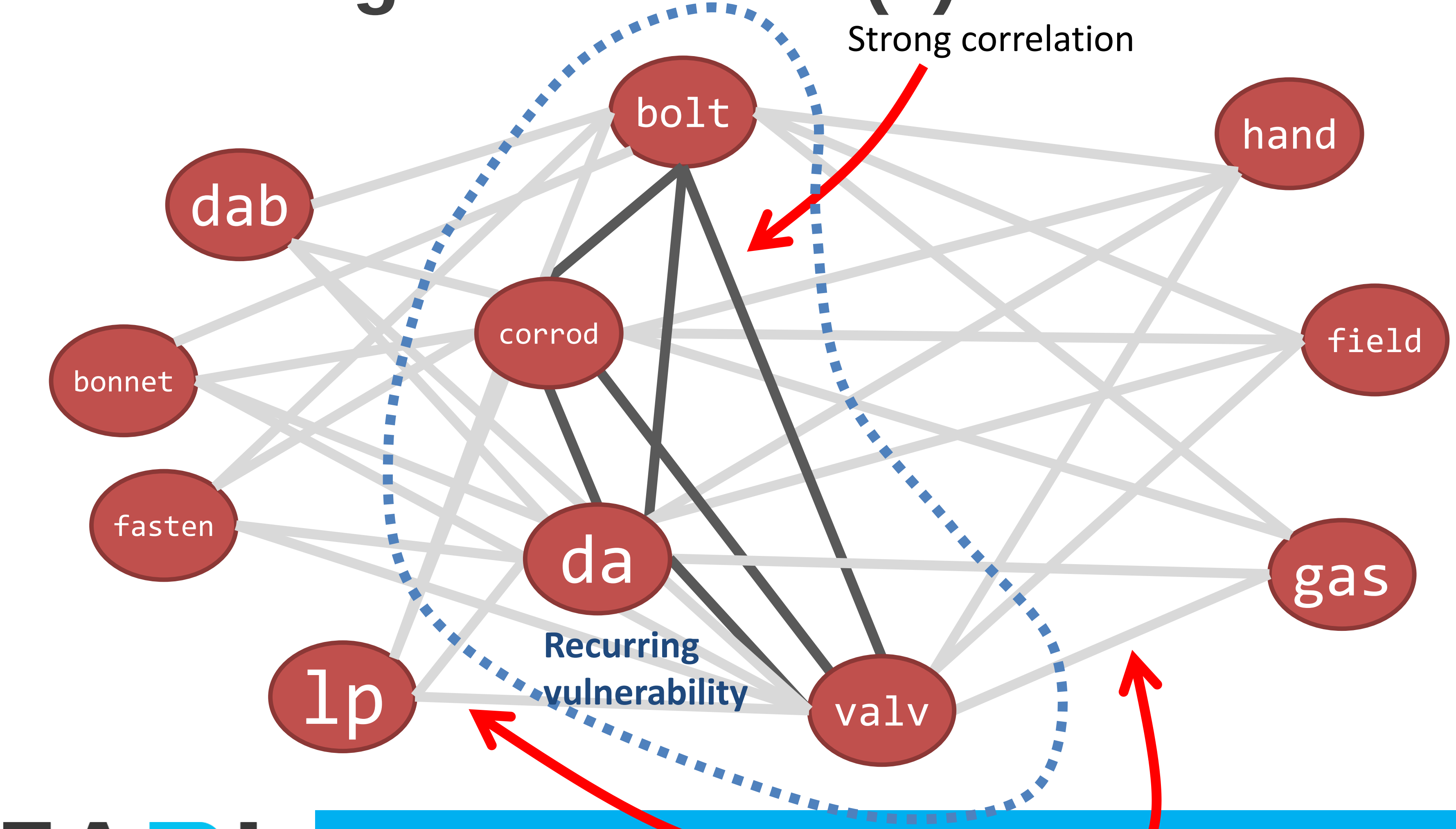
Ref: 457683143456

Measuring Correlation (2)

- Compute pairwise correlation for all terms

	dab	corrod	valv	bonnet	bolt	fasten	da	lp	...
dab									
corrod	0.11								
valv	0.03	0.24							
bonnet	0.05	0.32	0.27						
bolt	0.03	0.39	0.24	0.16					
fasten	0.15	0.13	0.26	0.17	0.35				
da	0.23	0.09	0.16	0.11	0.13	0.10			
lp	0.21	0.12	0.14	0.09	0.12	0.13	0.15		
...	...	...	...	...	...	...	...	...	

Measuring Correlation (3)



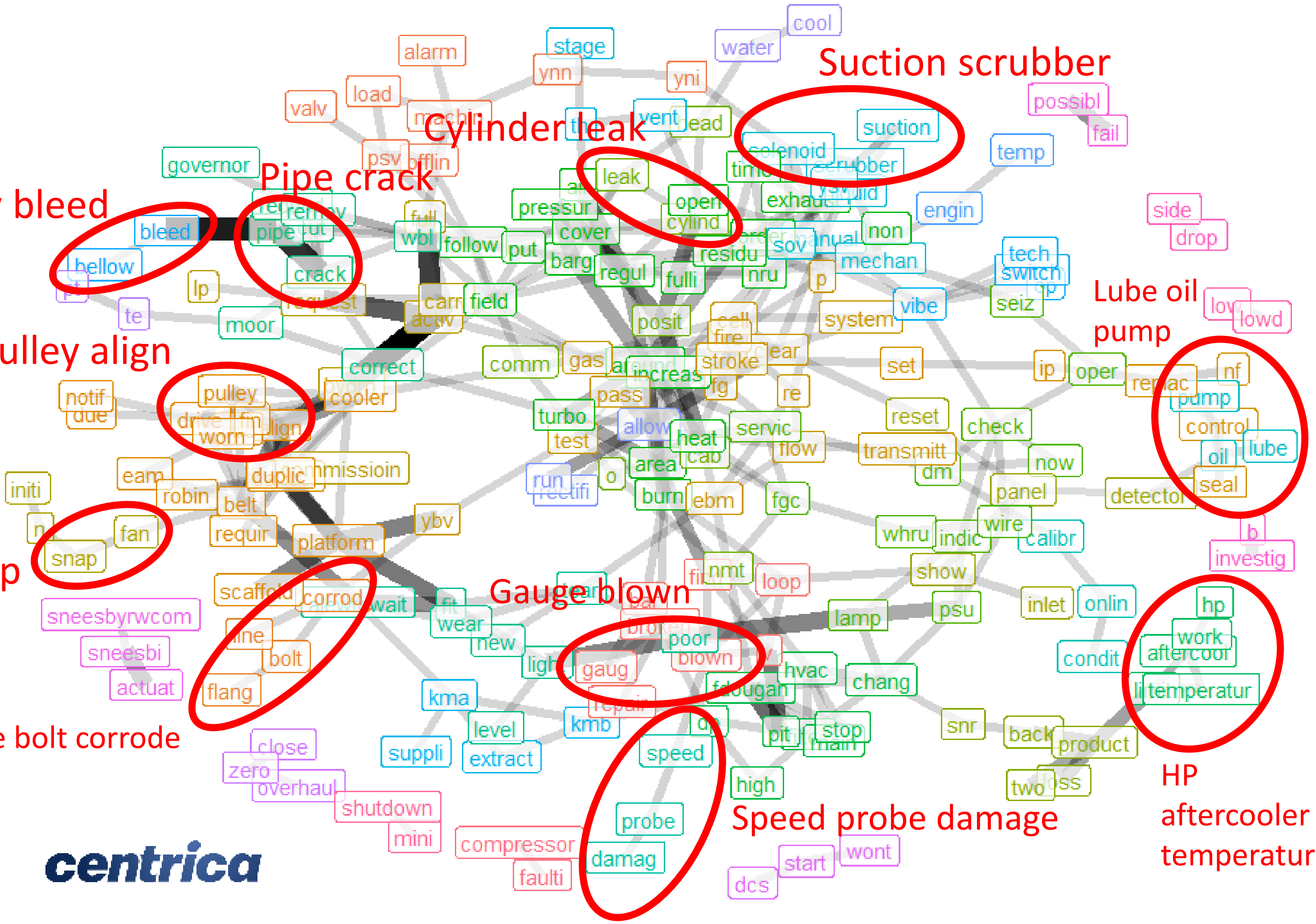
Strong correlation

Recurring vulnerability

Weak correlation

Identifying Vulnerability (1)

Example: Component Breakdown in Gas Compression Subsystem (Unigram)



Suction scrubber

Cylinder leak

Pipe crack

Bellow bleed

Lube oil pump

Pulley align

Fan snap

Gauge blown

Flange bolt corrode

Speed probe damage

HP aftercooler temperature

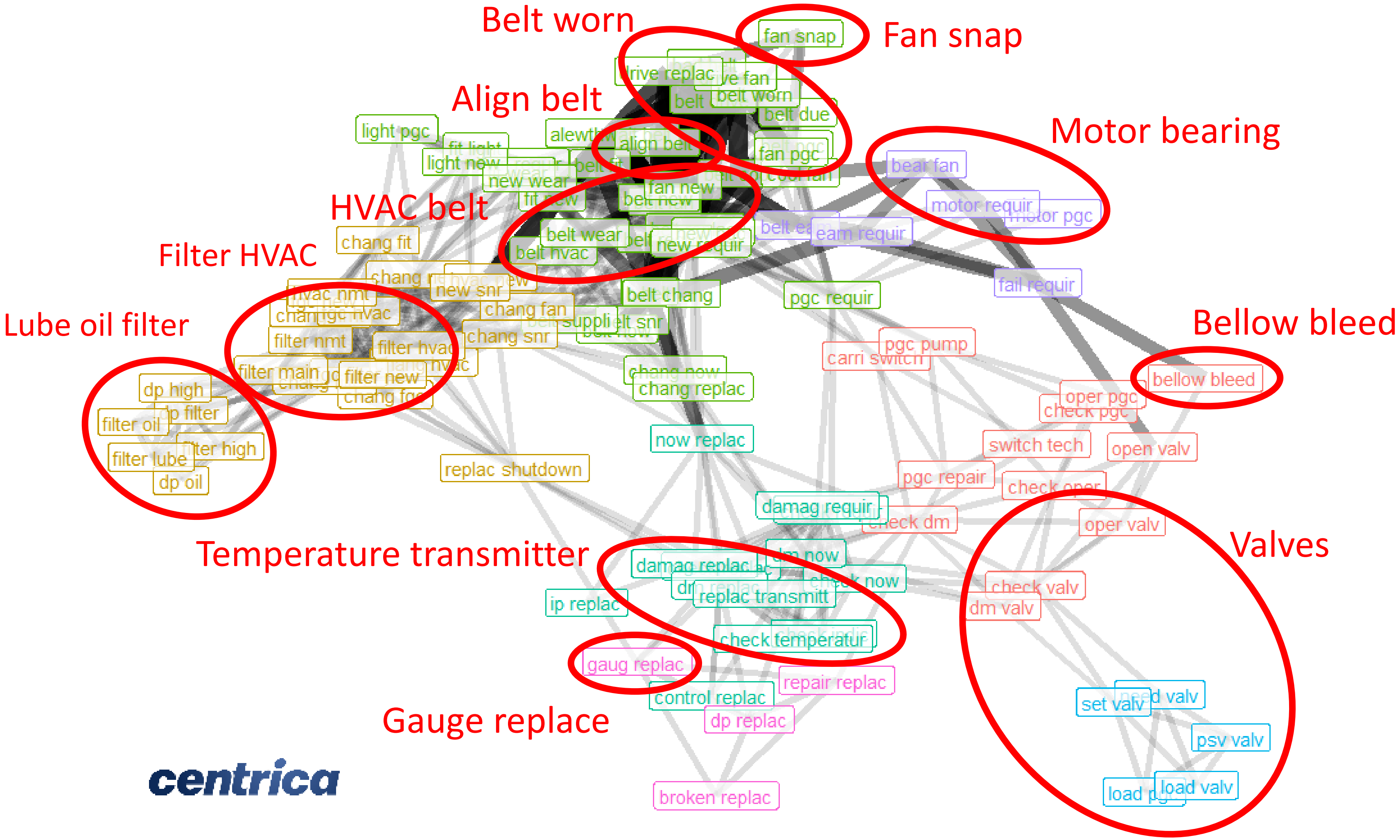
centrica

EARL

ENTERPRISE APPLICATIONS OF THE R LANGUAGE

Identifying Vulnerability (2)

Example: Component Breakdown in Gas Compression Subsystem (Skipgram)



Belt worn

Fan snap

Align belt

Motor bearing

HVAC belt

Filter HVAC

Lube oil filter

Bellow bleed

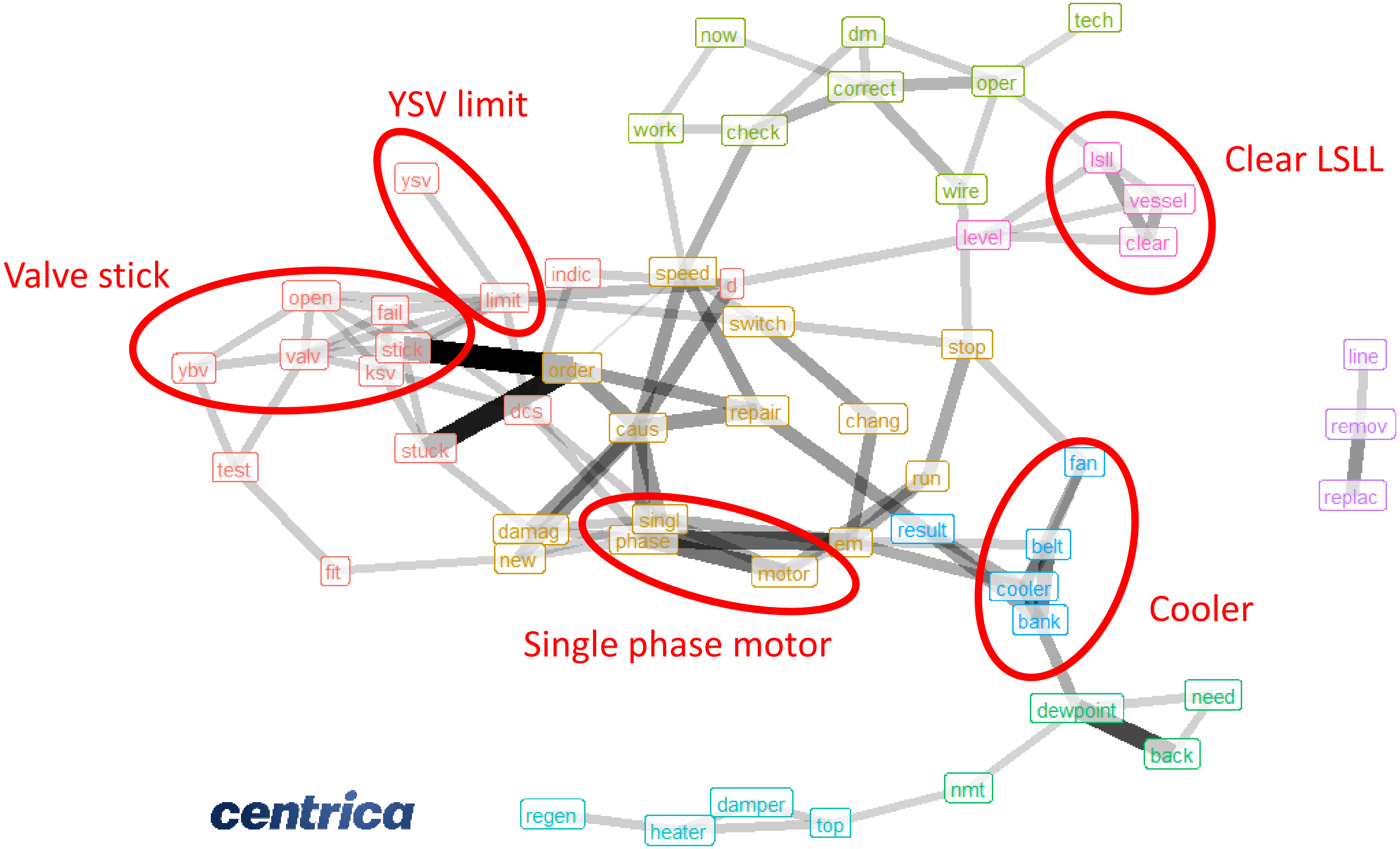
Temperature transmitter

Valves

Gauge replace

Identifying Vulnerability (3)

Example: Component Breakdown in Dewpoint Control Subsystem (Unigram)

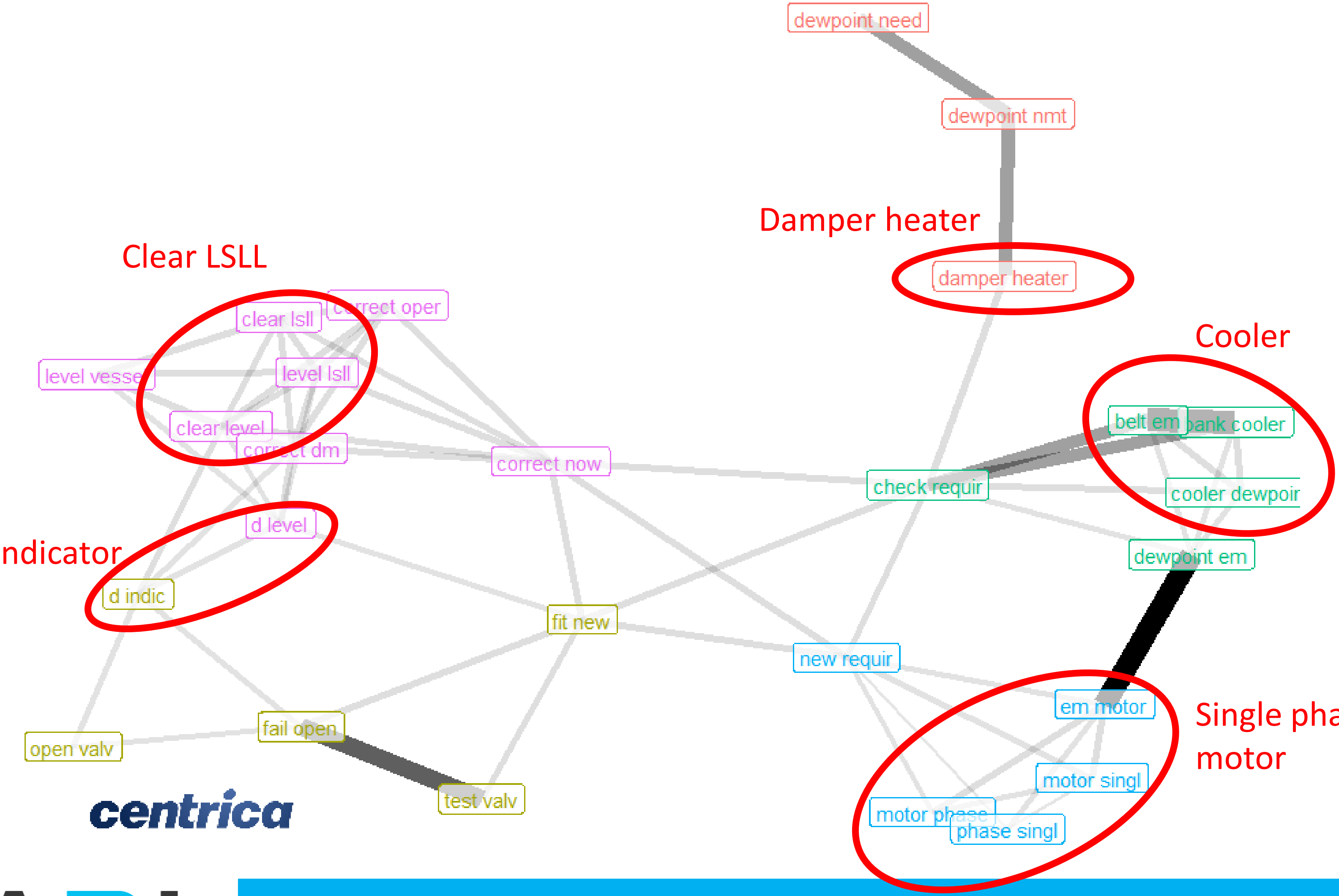


centrica

EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE

Identifying Vulnerability (4)

Example: Component Breakdown in Dewpoint Control Subsystem (Skipgram)



Damper heater

Clear LSLL

Cooler

D indicator

Single phase motor

Identifying Vulnerability (5)

- Brainstorming workshop
  - Technical expertise from Process Engineer
  - Highlighted recurring issues manually
  - Close up investigation

London 12-14th September 2017

Identifying High-Frequency Component Failure using Text-Mining Techniques

Timothy Wong – Centrica plc