Dealing with large volumes in statistical analysis and text mining tools

April 16-17 2012

II-SDV Nice

Laurent Hill
Questel

Questel
Freedom to Operate

Introduction

"What's hot in the semiconductor industry?"

"What's patented in Korea but not in the US?"

"What are the trends in biofuel production?"

To answer those questions you need to analyze 10,000s to millions of patent records in detail

Questel

Freedom to Operate

Presentation Outline

- Problem statement
- Large scale text mining
- Text mining applications in Fampat
- Statistical analysis at work on large data –sets
- Converging scale-out and small corpus techniques

Questel
Freedom to Operate

Scaling out patent analysis: Problem statement

- Some projects deal with up to ~ 10,000 patents
  - Existing tools well adapted
  - Classic cycle: extract, run text mining, vet data, graph results
- For some other problems, the volume is just too high for this approach
  - See introduction questions
    - Last 10 years of semiconductor patents: 854,000 families, 1M+ patents
    - Patented in Korea but not in the US, 2000-2010: 1,134,000 families
    - Trends in biofuel production: 18,800 patent families (51,000 patents)
  - Yet we need fast answers, stakeholders do not care about volume, they just need answers.
  - An entirely different approach is needed, while still having the same user interface, and without giving up on any functionality for smaller data-sets.

Questel
Freedom to Operate

Large scale text mining

- Text mining:
  - Producing a model from text analysis
  - Deriving high quality information / features from the model
- Issues
  - Text mining analysis is resource intensive, cannot be done in real time on a large corpus
  - The model often is a black box to the user, lowering user confidence on derived applications

Questel
Freedom to Operate

Concept extraction

- Main idea: extract concepts rather than words as a semantic model of patent documents, weight concepts according to sentence type
  - Weighted concepts produce a natural semantic vector model of the patent
- To deal with large volume we need to do this upfront when loading the data
  - 3 CPU years to pre-process the whole database

Questel
Freedom to Operate

Key sentences tagging

- Identify key sentences describing patent object, advantages and drawbacks, independent claims
- Uses morpho-syntactic analysis to spot important sentences
  - "patent writer sentiment analysis"
- Good compromise between conciseness of bibliographic abstracts and full text

Questel
Freedom to Operate

Semantic concept tagging

- « noun phrases » identification
  - Part-of-speech + stemming
  - Verb Suppression, some adjectives
  - Suppress patent boilerplate terminology
    - « preferred embodiment », « skilled artisan » ….
  - Syntactic normalization:
    - « surface of screens » → « screen surface »
- Relevance score computation, based on :
  - Field, key sentence morpho-syntactic detection, and number of occurrences

Questel
Freedom to Operate

Concept extraction example

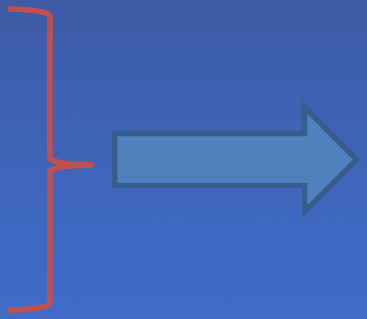
The invention relates generally to molecular level cleaning of parts by vapor degreasing.

More particularly, the invention relates to a solvent mixture comprising n-propyl bromide, a [mixture of low boiling solvents] and, …

The solvent mixture of the invention is non-flammable, non-corrosive, non-hazardous, and has a low ozone depletion potential.

Normalization examples

the heat conductivity
the conduction of heat
heat conductivity
heat conduction

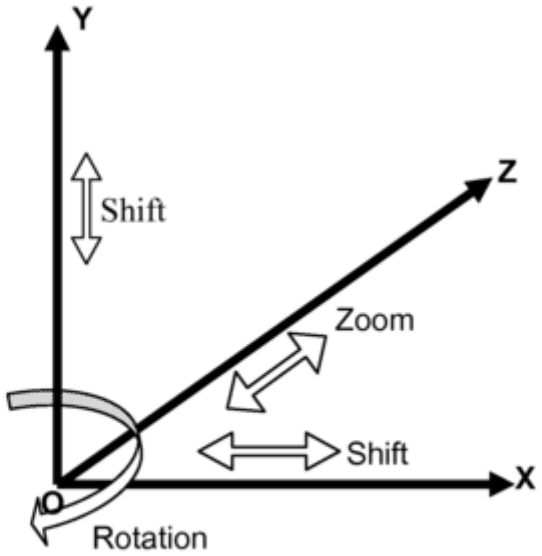


HEAT CONDUCTION

the user of the cellular telephone
cellular phone users
a user of a mobile telephone
any mobile phone users
their user mobile telephones

MOBILE PHONE USER

Questel
Freedom to Operate



ALPHANUMERICAL PUSH BUTTON | APPROACH | BLUETOOTHTRADE MARK | BORDERLINE SITUATION | CAMERA GESTURE | CAMERA MOVEMENT HISTORY | CAMERA ROTATION GESTURE | CAPACITIVE SENSITIVE PAD | CELL PHONE CAMERA | CELLULAR TELEPHONE CAMERA | CLASSIFIER | CLASSIFIER ROUTINE | COMMON MOVEMENT DISTANCE | COMPARING POSITION | COMPUTER READABLE INSTRUCTION | COMPUTER READABLE MEDIUM STORING COMPUTER EXECUTABLE INSTRUCTION | CURRENT GESTURE | CURRENT VIDEO IMAGE | DESKTOP COMPUTER KEYBOARD | DETECTED GESTURE | DIGITAL CAMERA | EARLY IMAGE | FEATURE POINT | FEATURE POINT NUMBER | HANDHELD COMPUTER MOUSE | HISTORICAL GESTURE INFORMATION | HISTORY DURATION | HYPERTEXT MARKUP LANGUAGE | IDENTIFIED FEATURE NUMBER | IDENTIFIED FEATURE POINT | IMAGE SUCCESSIVE FRAME | INTEL CORP | LAST GESTURE INDICATION | LOCAL DESKTOP COMPUTER | LUCAS KANADE PYRAMIDAL IMPLEMENTATION | MICROPROCESSOR RESEARCH LAB | MINIMUM NUMBER | MOBILE CELLULAR TELEPHONE | MOBILE TELEPHONE KEYPAD | MOVEMENT GESTURE LACK | MOVEMENT SPEED | MOVING SPEED | MULTIPURPOSE MICROPROCESSOR | MUSIC PLAYER | NAVIGATION | ONSCREEN CURSOR INTUITIVE MOVEMENT | OUTPUT GESTURE | PATTERN RECOGNITION | PRIOR VIDEO IMAGE | RECORDED GESTURE HISTORY FILE | RELATIVE POSITIONING | REMOVEABLE FLASH MEMORY | REMOVEABLE MEDIA | ROTATION GESTURE | ROTATION GESTURE OUTPUTTING | ROTATION

« tag cloud » style visualization in orbit

Vector model made user-friendly

Text mining applications at database level

- Since text mining has been applied massively, it can be leveraged at the database level (Fampat)
- Similarity searching
  - More like this (one or more patents)
  - Refine by example
- Related concept search
  - “chaussure de ski” yields:
    - Ski boot
    - Skier foot
    - Ski binding

Questel
Freedom to Operate

Large scale statistical analysis

- Nothing special here ?
  - More precisely: that's the goal
  - Be able to get all usual business graphics on patents
  - With close to instant response time on 1 million patent families
  - And *exact results*

Questel
Freedom to Operate

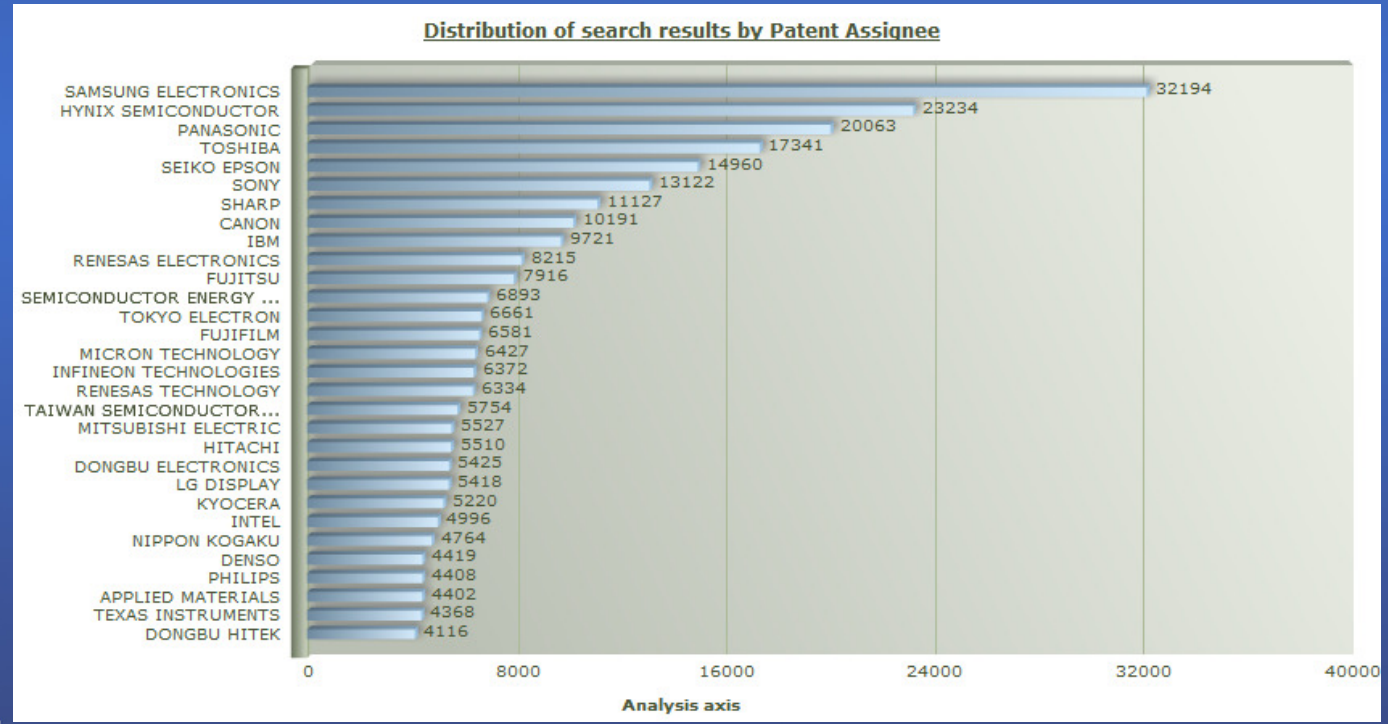
Prerequisites

- All data normalized and clean in the database
  - Normalized assignee names
  - Legal status for all patents and European national phases
  - Normalized text mining concepts
- Everything available at the family level in the same database
- A fast hybrid engine
  - Boolean full-text
  - Semantic
  - Analytic (OLAP)

Questel
Freedom to Operate

Results

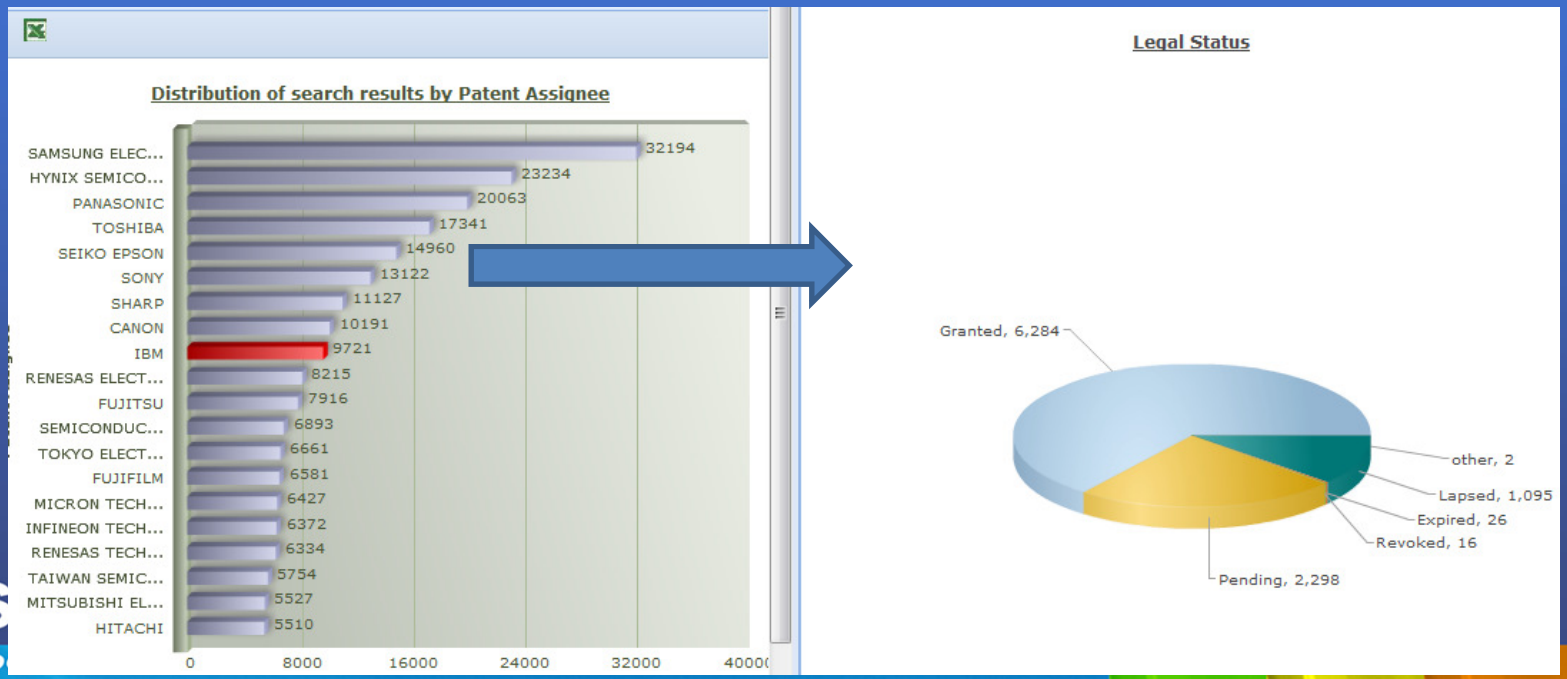
- 1M patent family, less than 10s for top assignees



Questel
Freedom to Operate

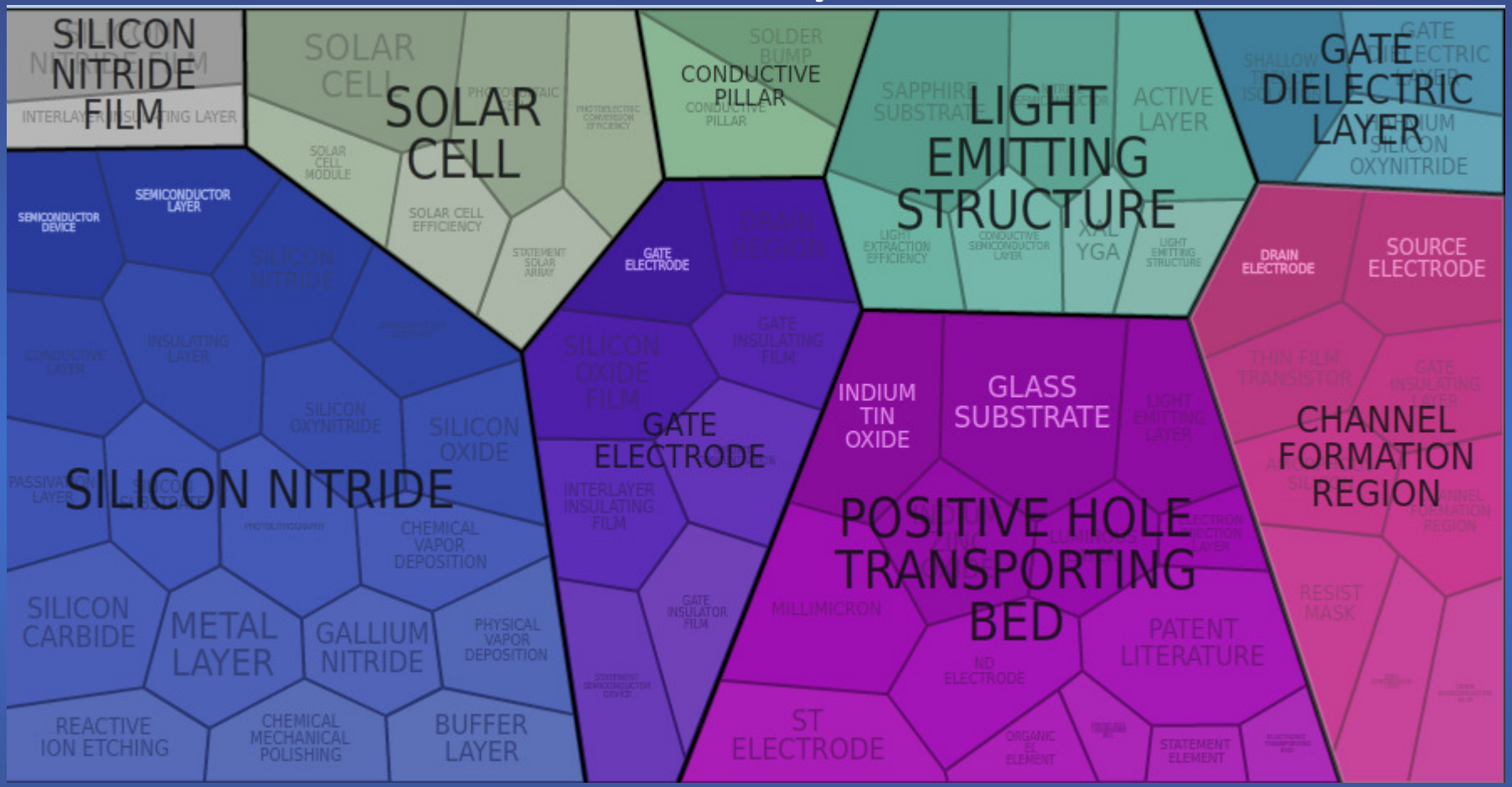
Data exploration

- On large volumes, you often are formulating and testing hypothesis on the spot
  - One click drill down :



Ques
Freedom to Op

Semantic exploration



- Cluster of concepts, filtered on year 2010
  - Drill-down on any cell

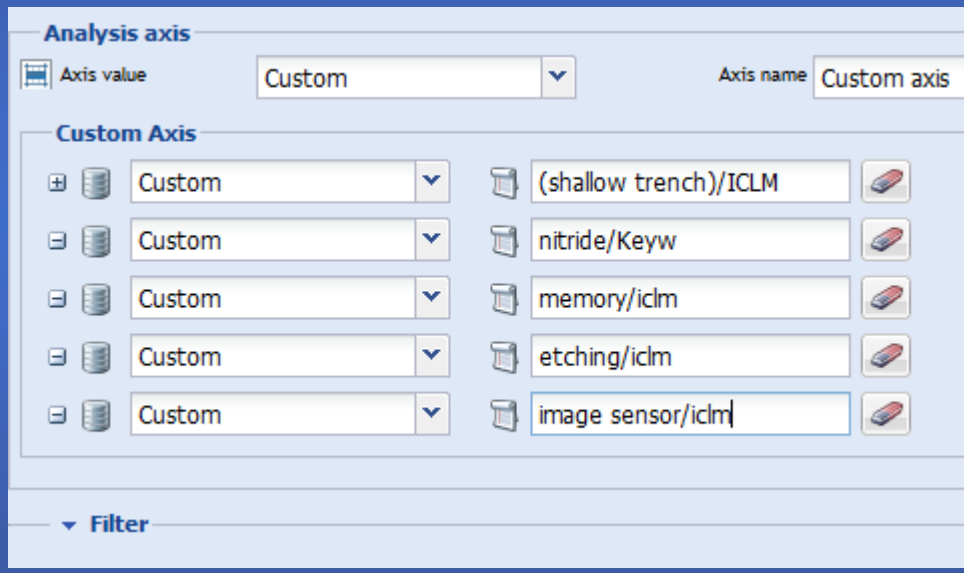
Questel
Freedom to Operate

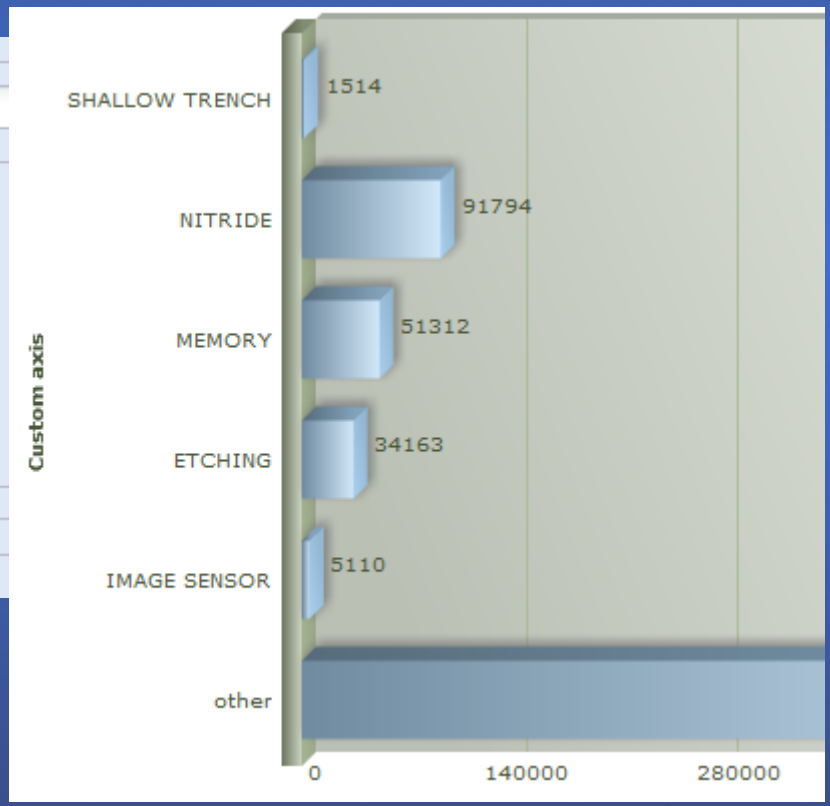
Graphing along Custom axis

- Sometimes you want to produce business graphics on your own queries
  - Combining International classes with concepts
    - A61Q or cosmetic/KEYW
  - Defining precise patent categories by full text queries
    - (bio_diesel P Alga??)/CLMS

Questel
Freedom to Operate

Custom business graphic

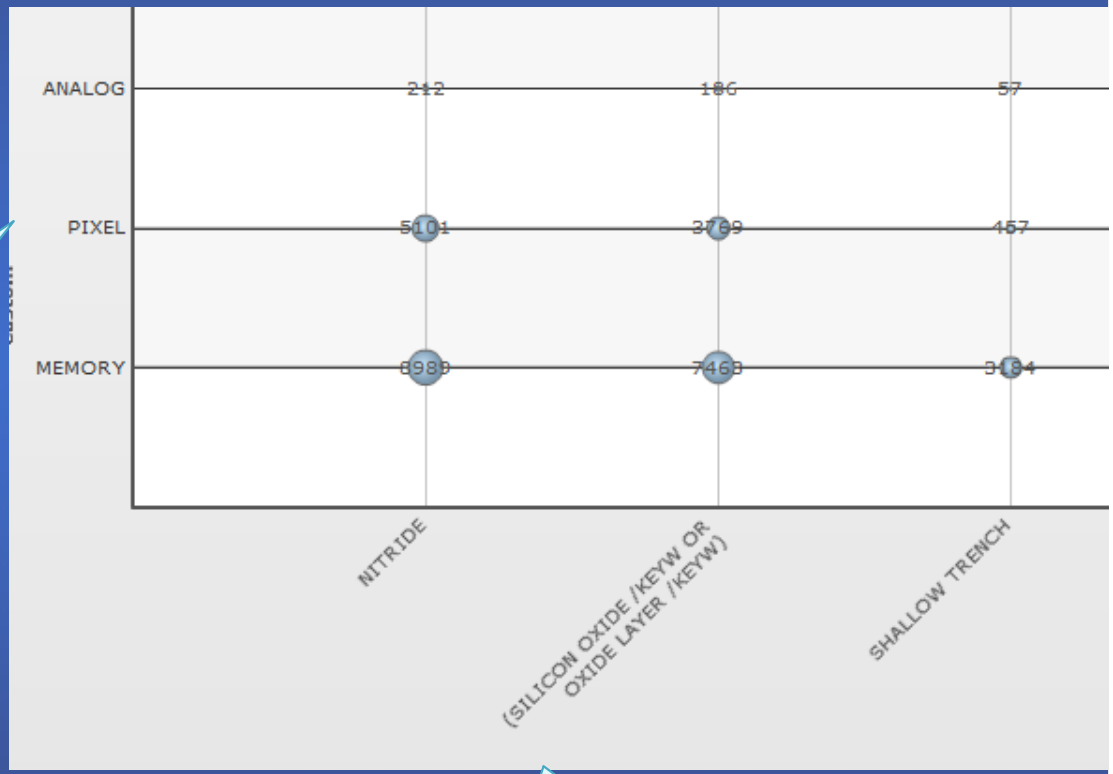




Questel
Freedom to Operate

Crossing custom axis

- Crossing independent claim searches with concepts



Independent Claim content

Concept words

Questel
Freedom to Operate

Comparing scale-out and small corpus techniques

	Traditional /small corpus	Large scale
Pros	• Ability to edit the data • Save edited data • Maps	• Practically unlimited in volume • Full integration with search engine • Full access to text mining features
Cons	• Limited to 60,000 patent families • Full text queries often limited • Have to wait for data to be extracted/saved	• No data vetting / editting • No data set saving

We need to get the best of both worlds

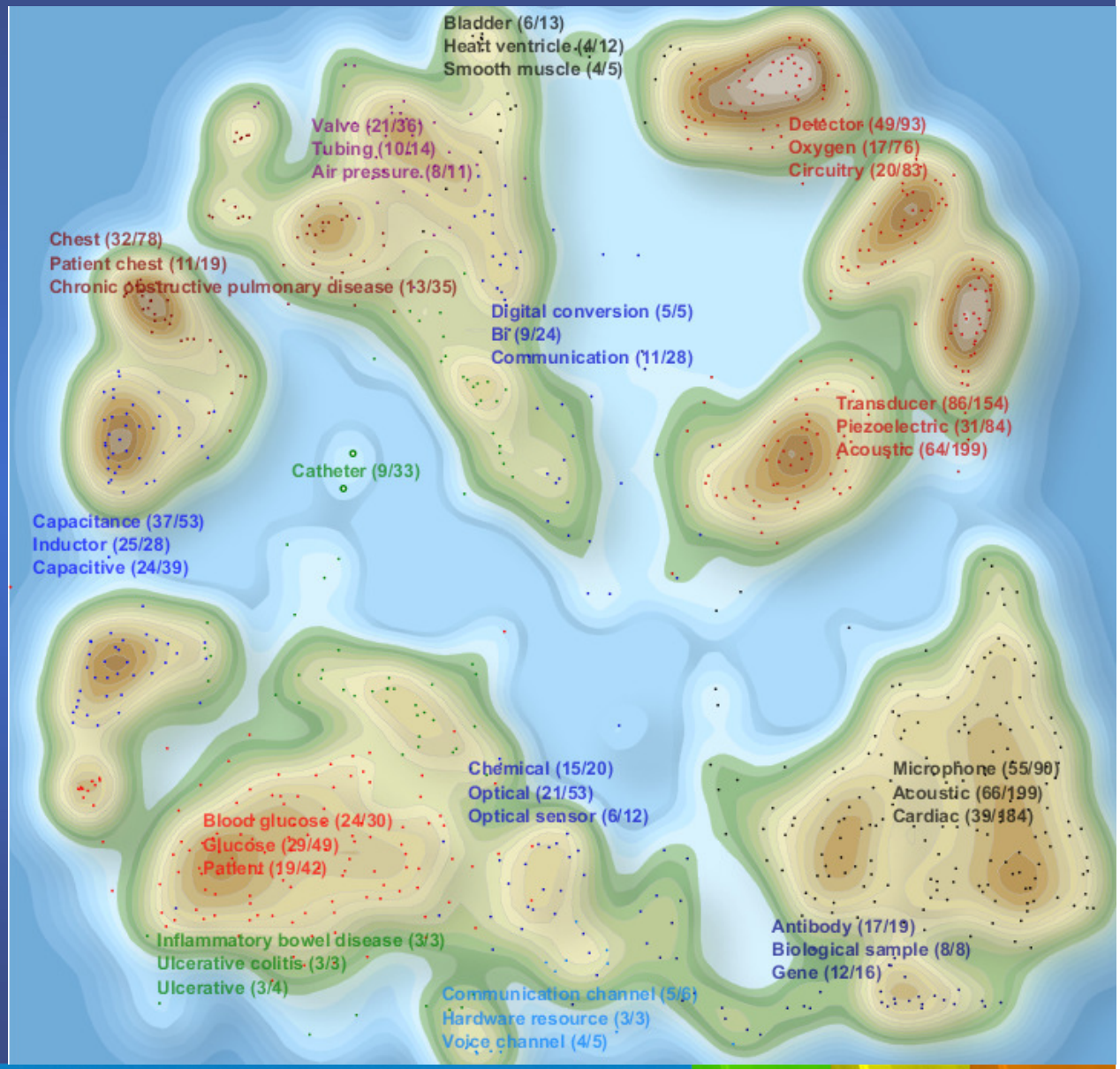
Questel
Freedom to Operate

Converged tool: orbit analysis module

- 3 data source
  - Live from database
  - Workfiles
  - Saved analysis: edited data
- Same features, same user interface

Questel
Freedom to Operate

Leveraging concepts: Mapping module



Questel
Freedom to Operate

Questions

- Feel free to share your horror / success stories in analyzing hundreds of thousand of patents.

Questel
Freedom to Operate

Thank you very much

Questel
Freedom to Operate