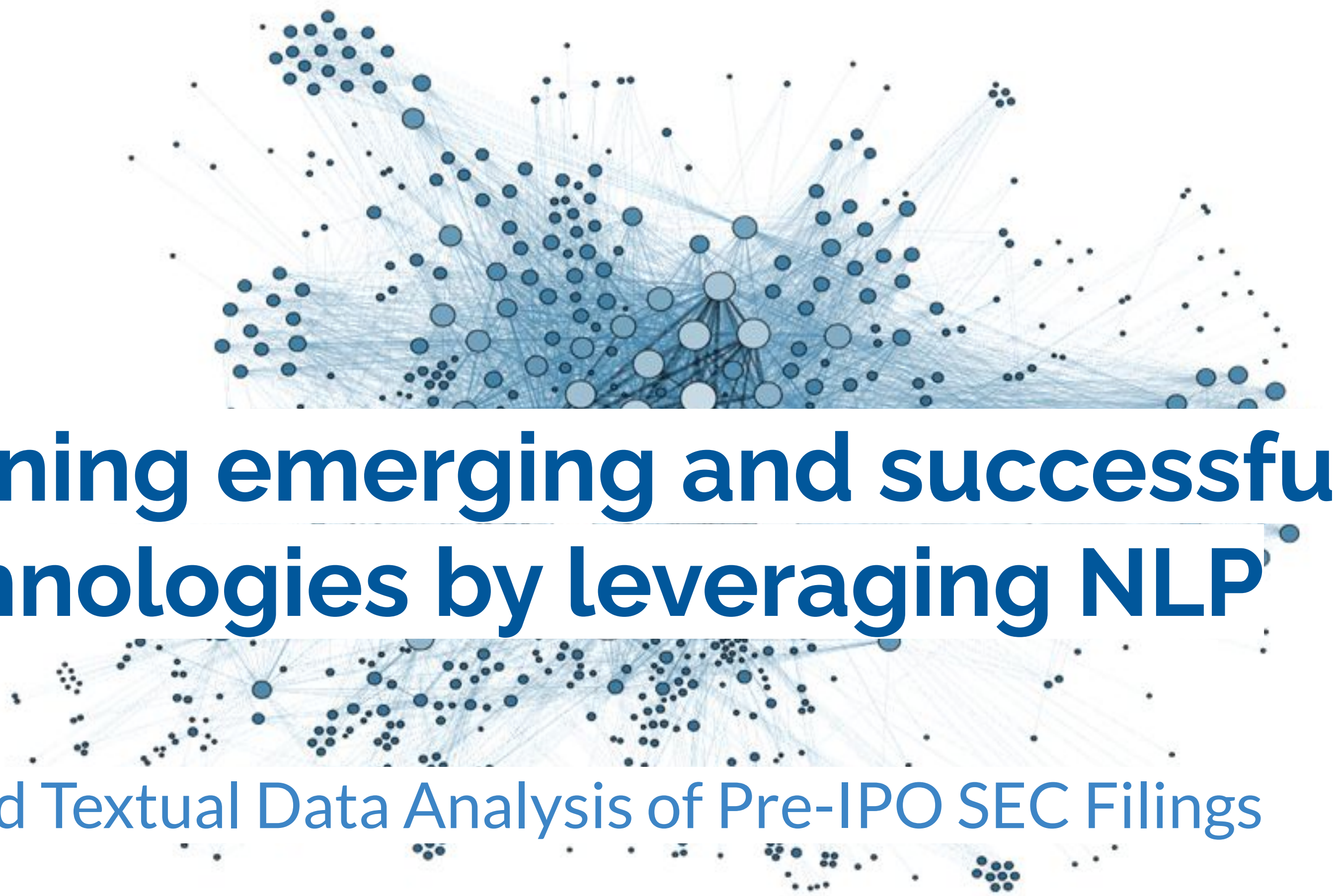


# Defining emerging and successful technologies by leveraging NLP

## Applied Textual Data Analysis of Pre-IPO SEC Filings

Aleksej Hoffärber
19.11.2019

# How to analyse the trends and industries: use of log likelihood for trend analysis and cosine similarity for industries comparison

**Approach A:**
Loglikelihood trend analysis

**Training the LDA** model based on all filings to identify commonalities within documents based on topics

**Granularization** and time-series analysis of respective topics to **define emerging technologies**

**Loglikelihood** analysis paired with KWIC to **determine** one and two-word based **key technologies**

**Approach B:**
Inverse cosine similarity

Topic **modelling for the reference and study** sets - basis for the comparison

**Comparison** and analysis of topics based on the **term frequency**

**KWIC as a support tool** to identify the keywords context and topic perspective

# Defining the change: key differences in the acquired results and model features

## Approach A:
Loglikelihood trend analysis

**Human intervention:** results are based upon manual analysis of rising topics and keyword selection

**Emerging relevance** is captured by combining time-series and loglikelihood across topic subsets

**Handling real-world complexity** by using more flexible key words with more than one word complicated

## Approach B:
Inverse cosine similarity

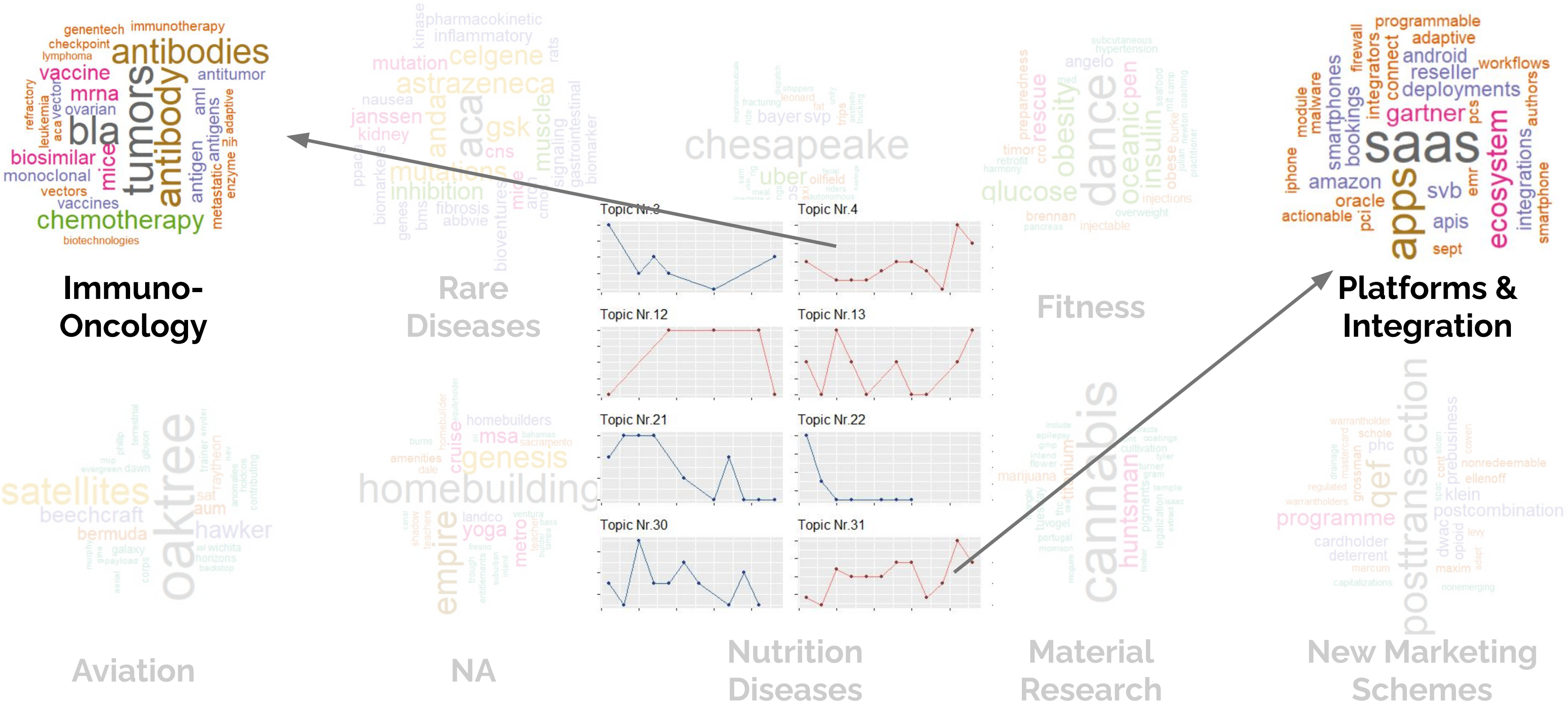
**Human intervention:** solution optimisation is based on the human decisions on the last stages

**Inverse correlation:** the key principle to define the emerging industries

**Keywords exploration:** approach requires extensive analysis of the used key terms

Approach A

# Time-Series analysis indicates likeliness of companies to integrate trend technologies into their business and operating model



Immuno-Oncology

Rare Diseases

Fitness

Platforms & Integration

Aviation

NA

Nutrition Diseases

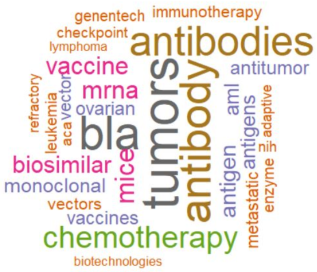
Material Research

New Marketing Schemes

Approach A

# Emerging topics indicate variability - ‘Immuno-Cology’ and ‘Platform & Integration’ being the most exhaustive classifications

Exact Keywords in Backup



**Immuno-Oncology**

kinase
pharmacokinetic
inflammatory
rats
mutation
celgene
astrazeneca
nausea
janssen
kidney
anda
aca
gsk
muscle
cns
signaling
gastrointestinal
biomarker
ppaca
biomarkers
mutations
inhibition
genes
bms
fibrosis
abbvie
bioventures
mice
arch
cmos

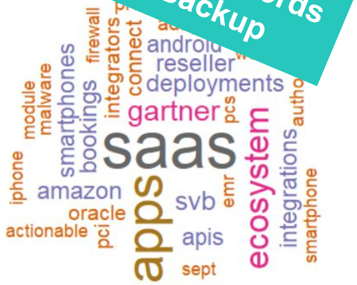
Rare Diseases

dispatch
shippers
leonard
fracturing
fat
trips
bayer
svp
chesapeake
uber
oilfield
cost
taxi
autonomous
passenger
riders
meal
bookings

Trips & Transport

subcutaneous
hypertension
angelo
preparedness
rescue
med
obesity
dance
oceanic
pen
insulin
seafood
burke
julian
obese
timor
retrofit
harmony
glucose
injections
brennan
pancreas
overweight
injectable

Fitness



**Platforms & Integration**

terrestrial
snyder
phillip
gibson
trainer
mip
evergreen
dawn
raytheon
satellites
oaktree
sat
aum
beechcraft
bermuda
hawker
galaxy
payload
wichita
horizons
backstop
corps
murphy
aerial

Aviation

homebuilder
homebuilders
burns
msa
bahamas
sacramento
cruise
amenities
dale
genesis
homebuilding
empire
landco
ventura
yoga
fresno
metro
teachers
shadow
bass
trough
entitlements
suburban
inland

NA

cirrhosis
grey
dietary
biotherapeutics
fibrosis
sanofi
syringe
acids
ldl
serum
dos
psc
hiv
hicks
nash
imagery
alta
amgen
fat
hepatic
bile
shire
muscle
lipid
cholesterol
analogs
potassium
sodium
industrialization
macquarie

Nutrition Diseases

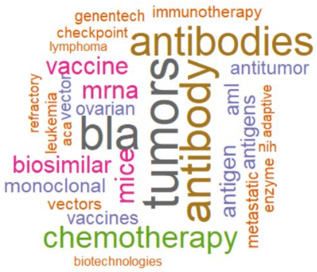
include
epilepsy
gmp
inland
flower
marijuana
titanium
cannabis
extracts
coatings
cultivation
tyler
turner
gram
huntsman
pigments
temple
legalization
isaac
thc
vogel
portugal
morrison
tuesday
triangle

Material Research

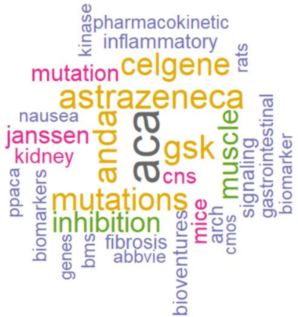
warrantholder
schole
phc
mastercard
grossman
qef
posttransaction
prebusiness
cowen
nonredeemable
ellenoff
regulated
warrantholders
klein
programme
postcombination
cardholder
dwac
opioid
levy
deterrent
marcum
maxim
adapt
capitalizations
nonemerging

New Marketing Schemes

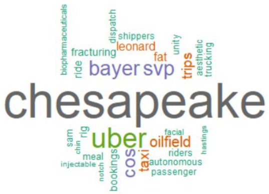
# Emerging topics indicate variability - ‘Immuno-Cology’ and ‘Platform & Integration’ being the most exhaustive classifications



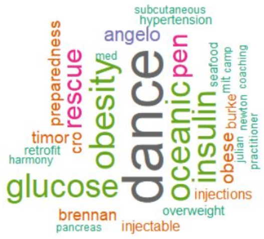
Immuno-Oncology



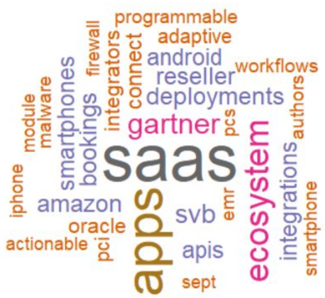
Rare Diseases



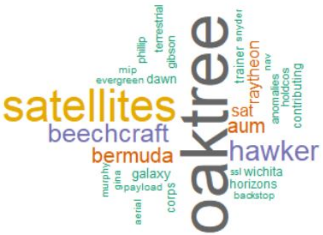
Trips & Transport



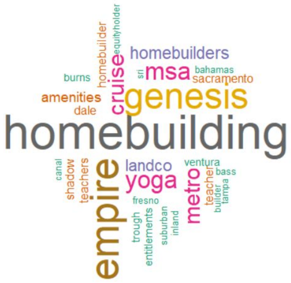
Fitness



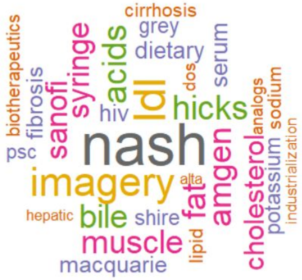
Platforms & Integration



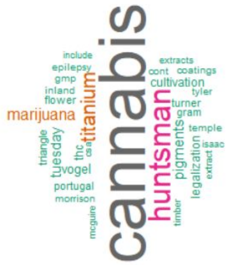
Aviation



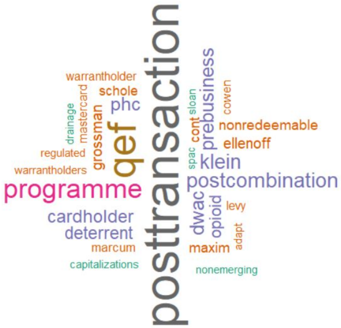
NA



Nutrition Diseases



Material Research



New Marketing Schemes

# Back-Up

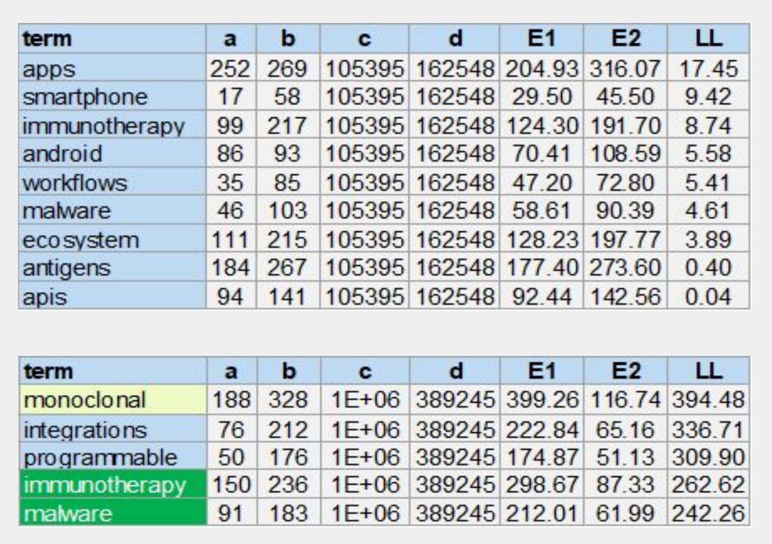
Back-Up

Approach A

# Manual selection of terms considering typology of ‘emerging technology’ - correctly reflecting immunology and platform trends

## Single Keyword Analysis

topic-specific vs. general

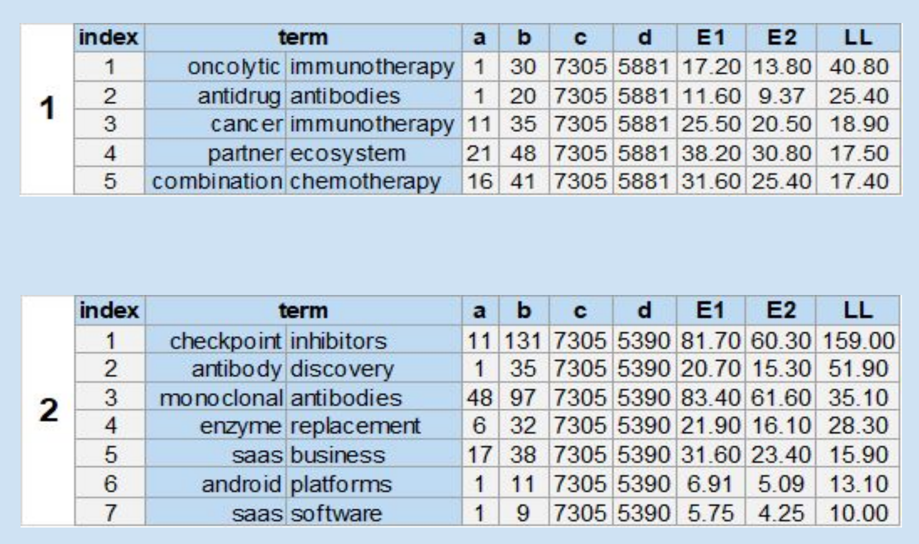


| term | a | b | c | d | E1 | E2 | LL |
|---|---|---|---|---|---|---|---|
| apps | 252 | 269 | 105395 | 162548 | 204.93 | 316.07 | 17.45 |
| smartphone | 17 | 58 | 105395 | 162548 | 29.50 | 45.50 | 9.42 |
| immunotherapy | 99 | 217 | 105395 | 162548 | 124.30 | 191.70 | 8.74 |
| android | 86 | 93 | 105395 | 162548 | 70.41 | 108.59 | 5.58 |
| workflows | 35 | 85 | 105395 | 162548 | 47.20 | 72.80 | 5.41 |
| malware | 46 | 103 | 105395 | 162548 | 58.61 | 90.39 | 4.61 |
| ecosystem | 111 | 215 | 105395 | 162548 | 128.23 | 197.77 | 3.89 |
| antigens | 184 | 267 | 105395 | 162548 | 177.40 | 273.60 | 0.40 |
| apis | 94 | 141 | 105395 | 162548 | 92.44 | 142.56 | 0.04 |

| term | a | b | c | d | E1 | E2 | LL |
|---|---|---|---|---|---|---|---|
| monoclonal | 188 | 328 | 1E+06 | 389245 | 399.26 | 116.74 | 394.48 |
| integrations | 76 | 212 | 1E+06 | 389245 | 222.84 | 65.16 | 336.71 |
| programmable | 50 | 176 | 1E+06 | 389245 | 174.87 | 51.13 | 309.90 |
| immunotherapy | 150 | 236 | 1E+06 | 389245 | 298.67 | 87.33 | 262.62 |
| malware | 91 | 183 | 1E+06 | 389245 | 212.01 | 61.99 | 242.26 |

## Left- and Right KWIC Analysis

topic-specific



| | index | term | | a | b | c | d | E1 | E2 | LL |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | oncolytic | immunotherapy | 1 | 30 | 7305 | 5881 | 17.20 | 13.80 | 40.80 |
| | 2 | antidrug | antibodies | 1 | 20 | 7305 | 5881 | 11.60 | 9.37 | 25.40 |
| | 3 | cancer | immunotherapy | 11 | 35 | 7305 | 5881 | 25.50 | 20.50 | 18.90 |
| | 4 | partner | ecosystem | 21 | 48 | 7305 | 5881 | 38.20 | 30.80 | 17.50 |
| | 5 | combination | chemotherapy | 16 | 41 | 7305 | 5881 | 31.60 | 25.40 | 17.40 |

| | index | term | | a | b | c | d | E1 | E2 | LL |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | checkpoint | inhibitors | 11 | 131 | 7305 | 5390 | 81.70 | 60.30 | 159.00 |
| | 2 | antibody | discovery | 1 | 35 | 7305 | 5390 | 20.70 | 15.30 | 51.90 |
| | 3 | monoclonal | antibodies | 48 | 97 | 7305 | 5390 | 83.40 | 61.60 | 35.10 |
| | 4 | enzyme | replacement | 6 | 32 | 7305 | 5390 | 21.90 | 16.10 | 28.30 |
| | 5 | saas | business | 17 | 38 | 7305 | 5390 | 31.60 | 23.40 | 15.90 |
| | 6 | android | platforms | 1 | 11 | 7305 | 5390 | 6.91 | 5.09 | 13.10 |
| | 7 | saas | software | 1 | 9 | 7305 | 5390 | 5.75 | 4.25 | 10.00 |

🟩 = also appear in topic-specific single keyword overview

▨ = also appear in topic-specific KWIC analysis