GTM2015: 5th Global TechMining Conference Atlanta, Georgia 2015.09.16

Classifying Biomedical Text for Mining Keyword Correlations and Technology Opportunities Analysis

Jing Ma[1], Alan Porter[2, 3],Natalie Abrams[4]

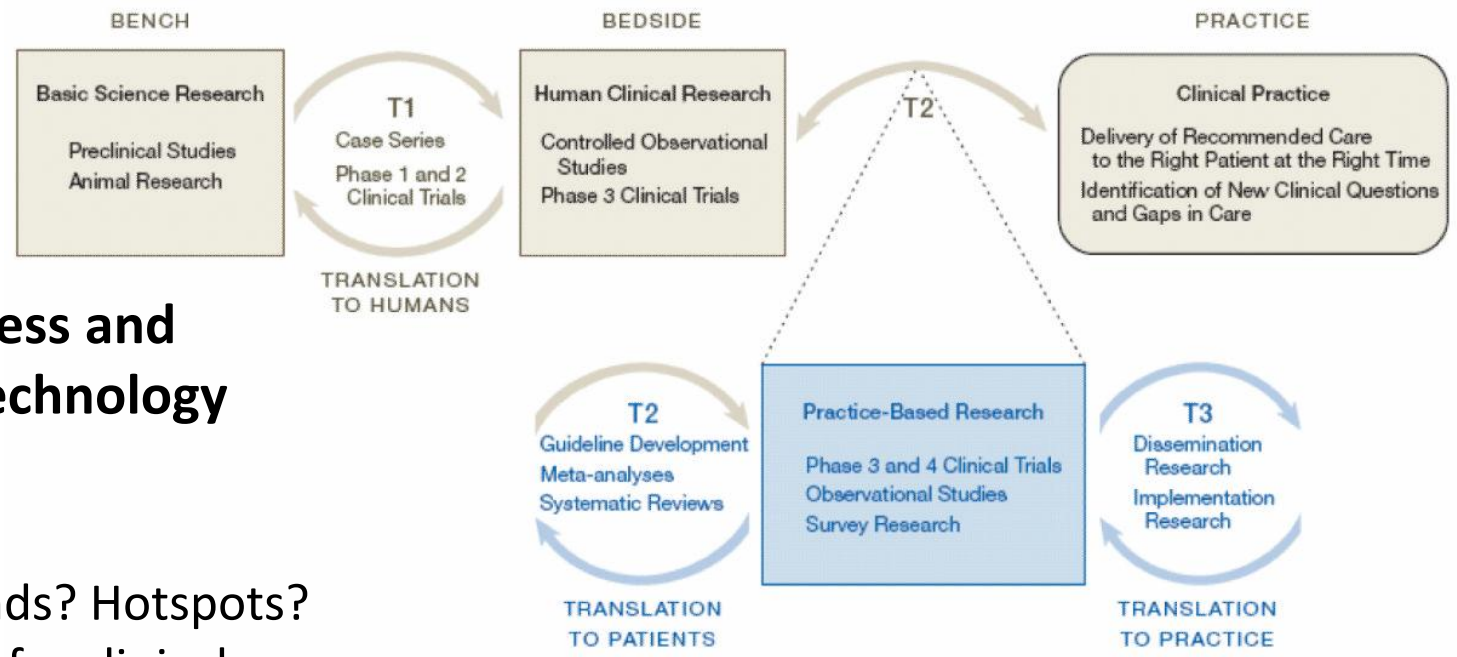
[1.] School of Management and Economics, Beijing Institute of Technology
[2.] School of Public Policy, Georgia Institute of Technology,
[3.] Search Technology, Inc.
[4.] National Cancer Institute

Why study biomedical translation?

- Biomedical research requires strict procedures from formulation development to clinical trial; only a few studies end up leading to marketable products. Application is complex and mixed.
- "It takes an estimated average of 17 years for scientific discoveries to enter day-to-day clinical practice" (and only 14% make it) (Westfall et al., 2007)

Our Perspective:

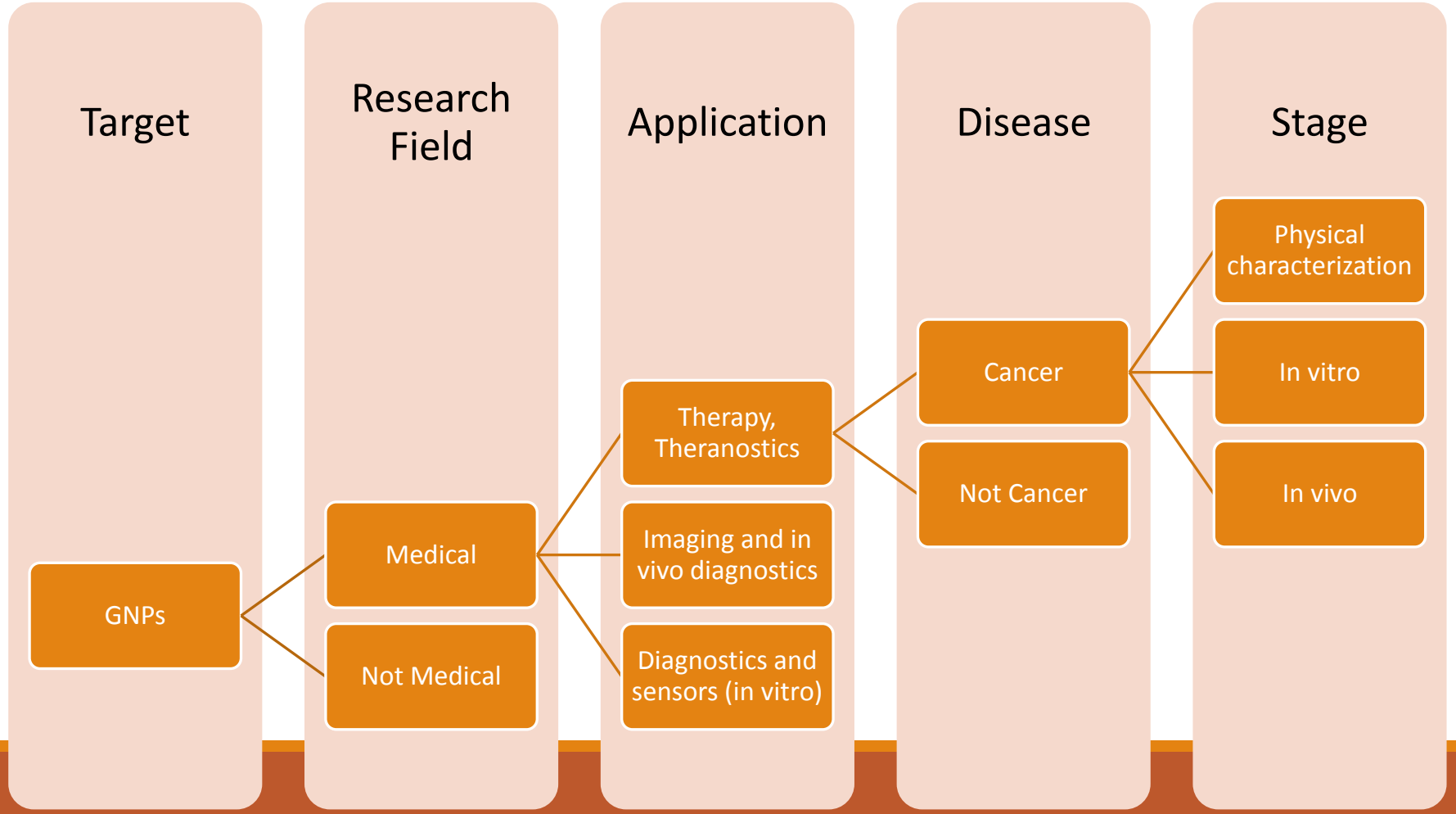
- **Abundant literature resources**
  Since 1990, over 15 million MEDLINE records
  Titles, abstracts and MeSH headings
- **Tracing biomedical translational process and grasping more detailed insights for Technology Opportunities Analysis (TOA)**
- **Tech Mining**:
  "what" questions? – developmental trends? Hotspots?
  "when" questions? – when will be ready for clinical testing?



Three-phase translational research model in Westfall et al. (2007).

Research Framework of GNPs

- **Target field: Gold nanoparticles (GNPs) for nano-enabled drug delivery (NEDD)**
  - ✓ seeking articles on therapy, therapy with diagnostics and therapy with imaging



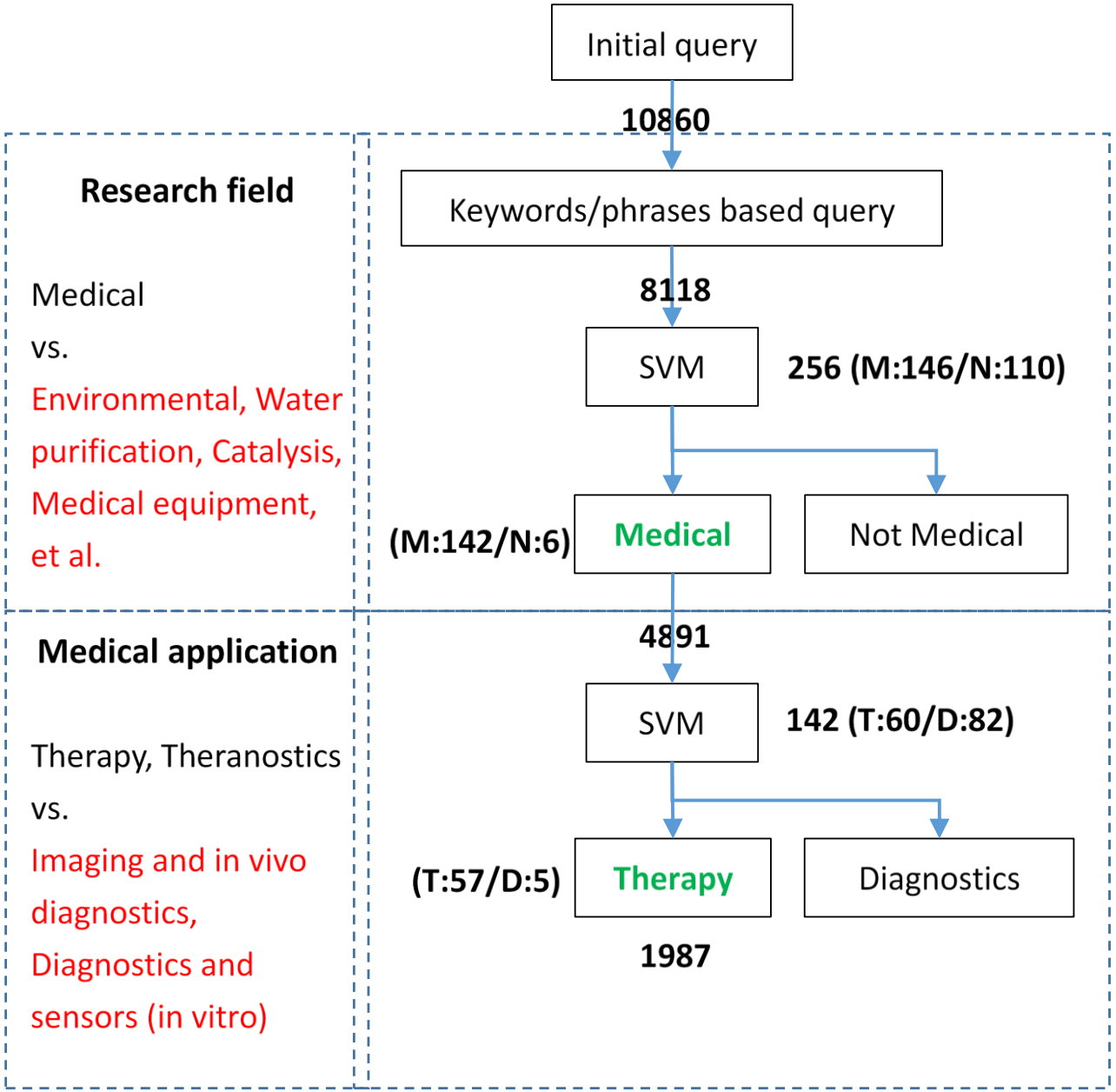
Data Query and Classification

- **Initial Query**
  - ✓ PubMed, 2001-2014
  - ✓ Search all fields for: (gold nano*) – including 600 variations provided by PubMed
  - ✓ Only records with abstracts – with more than 3 sentences
  - ✓ Only research articles – not reviews, comments, evaluation studies, news, etc.
  - ✓ Retrieved ~10,800 records

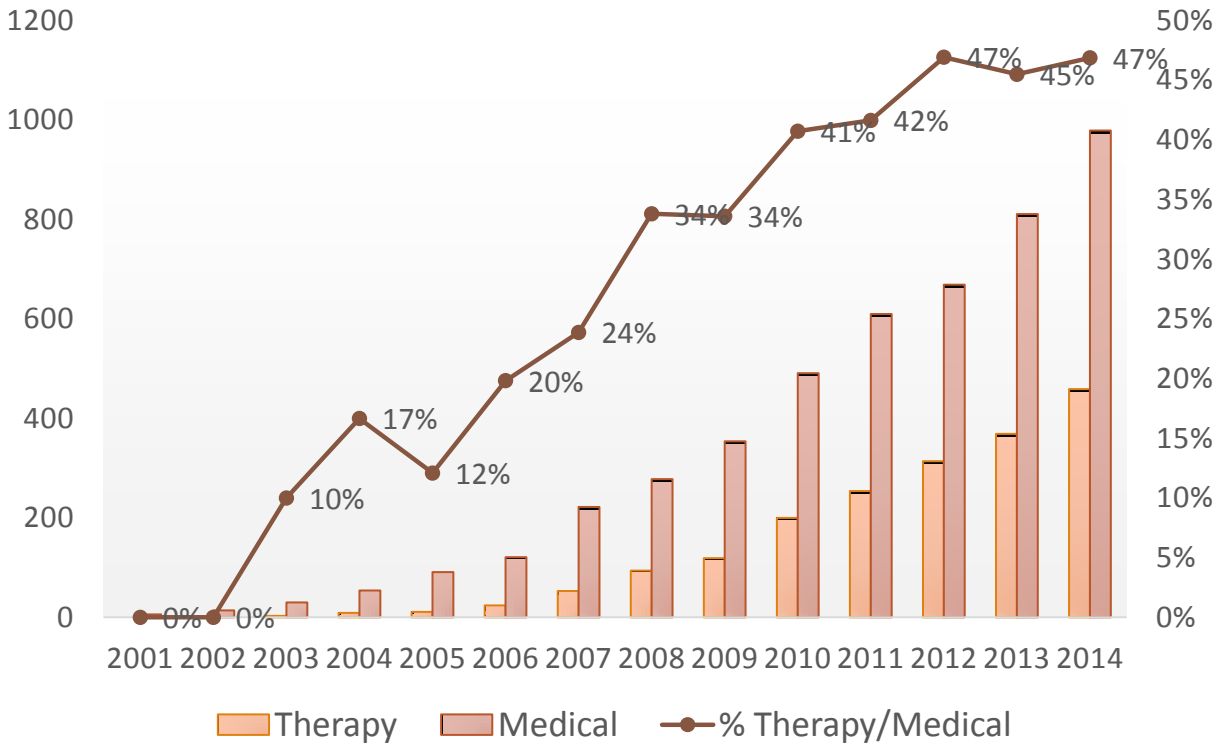
- **Refining**
  - ✓ Keywords/phrases based query
  - ✓ Supervised classification, SVM
  - ✓ A manually annotated sample with ~250 records

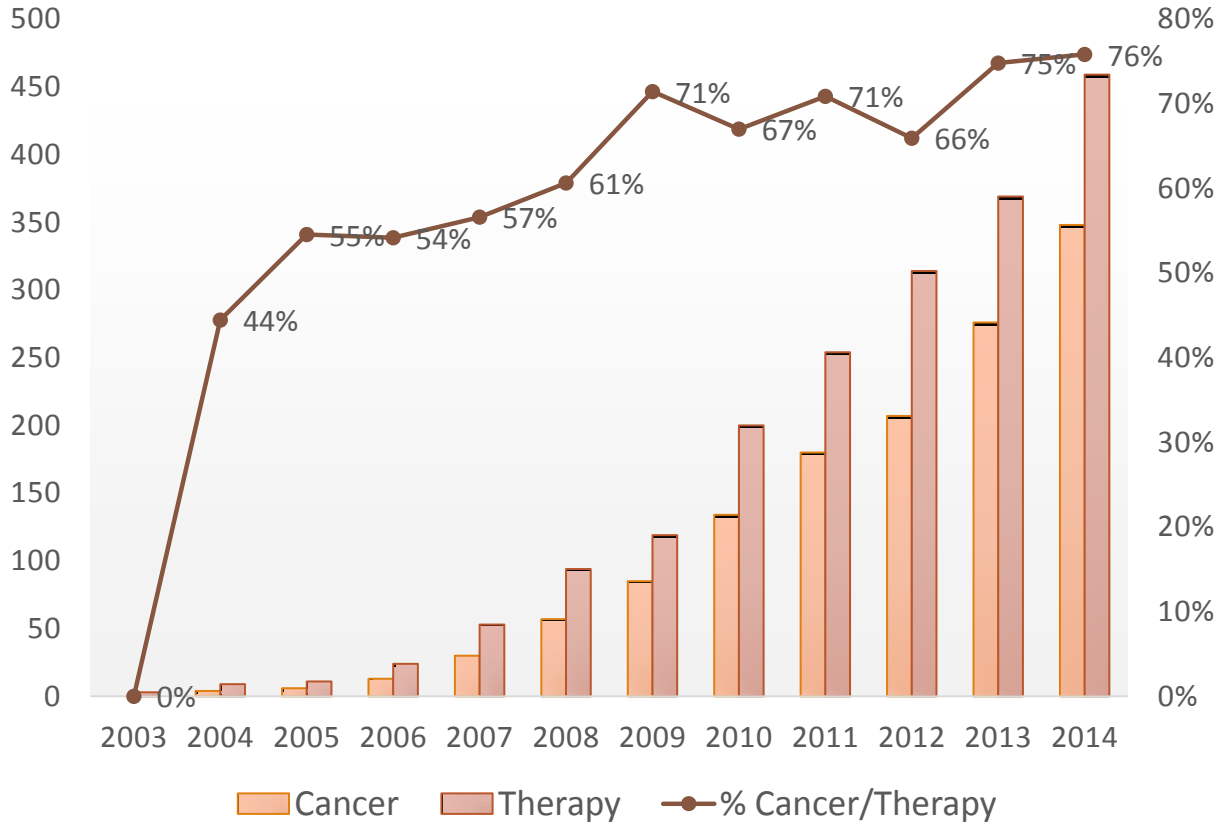
Steps for classifying records by research fields and medical applications

- ✓ Generating NLP words/phrases list from initial dataset (title, abstract, MeSH)
- ✓ Top ~2000 words/phrases are manually checked – if some of them are specialized for a specific research field or an application
- ✓ Keywords based query + supervised model (using selected keywords as properties)
- ✓ Sampling annotated records as training set, and using the others as test set
- ✓ Selecting models with relatively high accuracies to predict unannotated records



GNPs therapy application

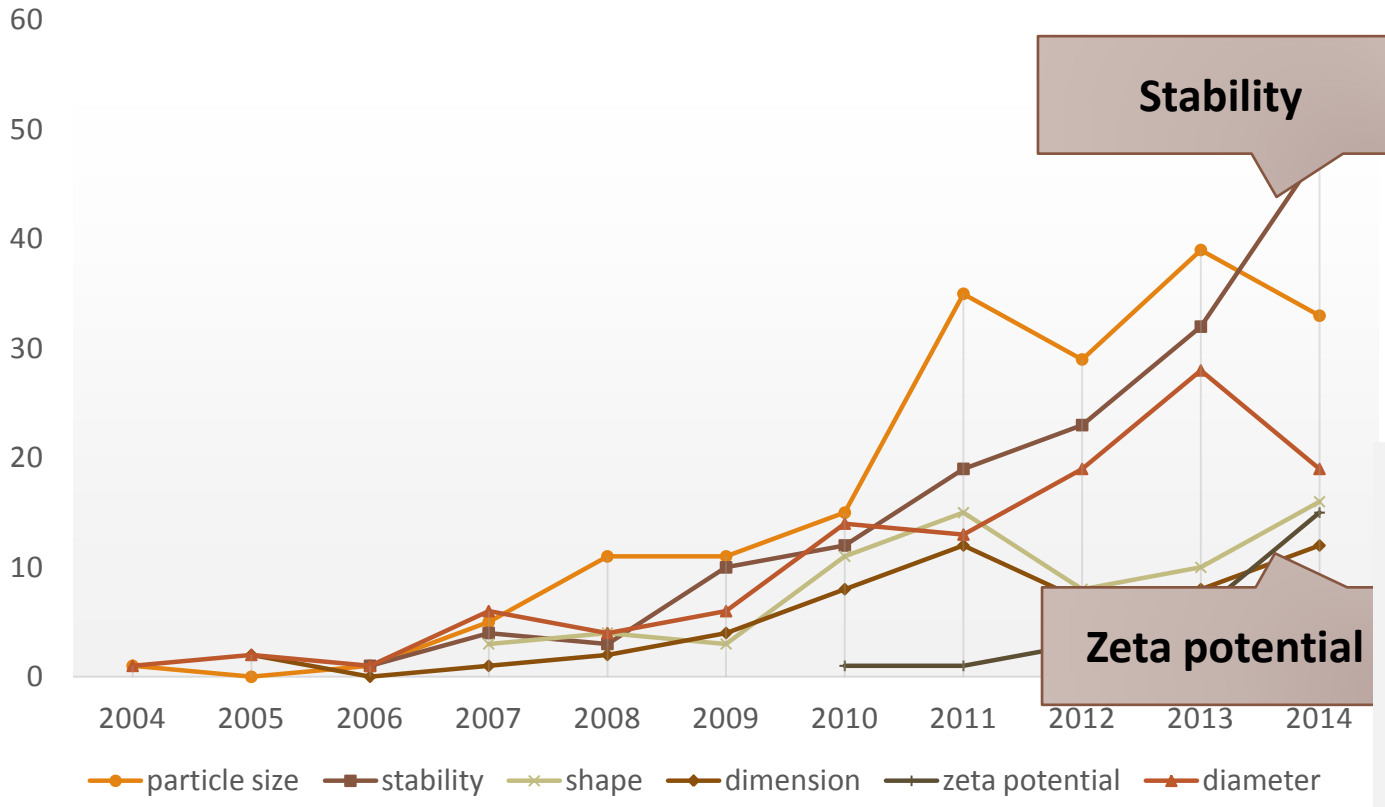


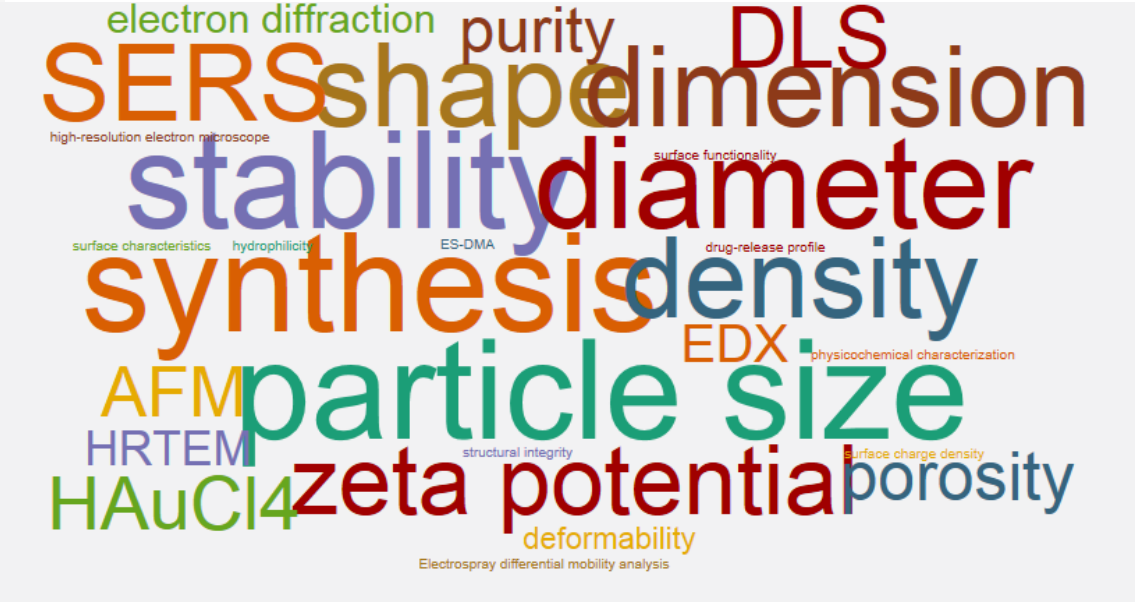


Translational Stages for NEDD

Stage	Descriptions	Sample Keywords
Physical Characterization	**Nanoplatform development** and optimization for specific electric, magnetic, optical and mechanical properties. Fabrication, design, synthesis, optimization.	particle size, size distribution, molecular weight, density, shape, diameter, aspect ratio, surface characteristics, **stability, zeta potential**, loading capacity, purity, HAuCl4, etc.
In Vitro	***In vitro* assays for efficacy**, activity, functional validation, biocompatibility (not rejected by the body), sterility, off target toxicity, targeting, drug release, bioavailability and internalization to ensure that adequate concentrations of the drug are achieved in the target neoplastic tissue/cells.	biological activity, enzymatic assay, binding affinity, target inhibition, gene silencing, in vitro potency, bioavailability, oxidative stress, hepatocyte assay, macrophage, cytotoxicity, necrosis, **apoptosis**, etc.
In Vivo	**Predictive *in vivo* efficacy models to support the pharmacology** section. *In vivo* toxicity studies to support the toxicology section. *In vivo* assays for ADME (absorption, distribution, metabolism and excretion), and pharmacokinetics to support the toxicology section. Comprehensive studies.	tissue distribution, clearance, half-life, systemic exposure, animal models, mice, rats, dogs, metabolites, biomarkers, optimal timing, target validation, gene silencing, potency, first-in-human, clinical trial, clinical study, etc.

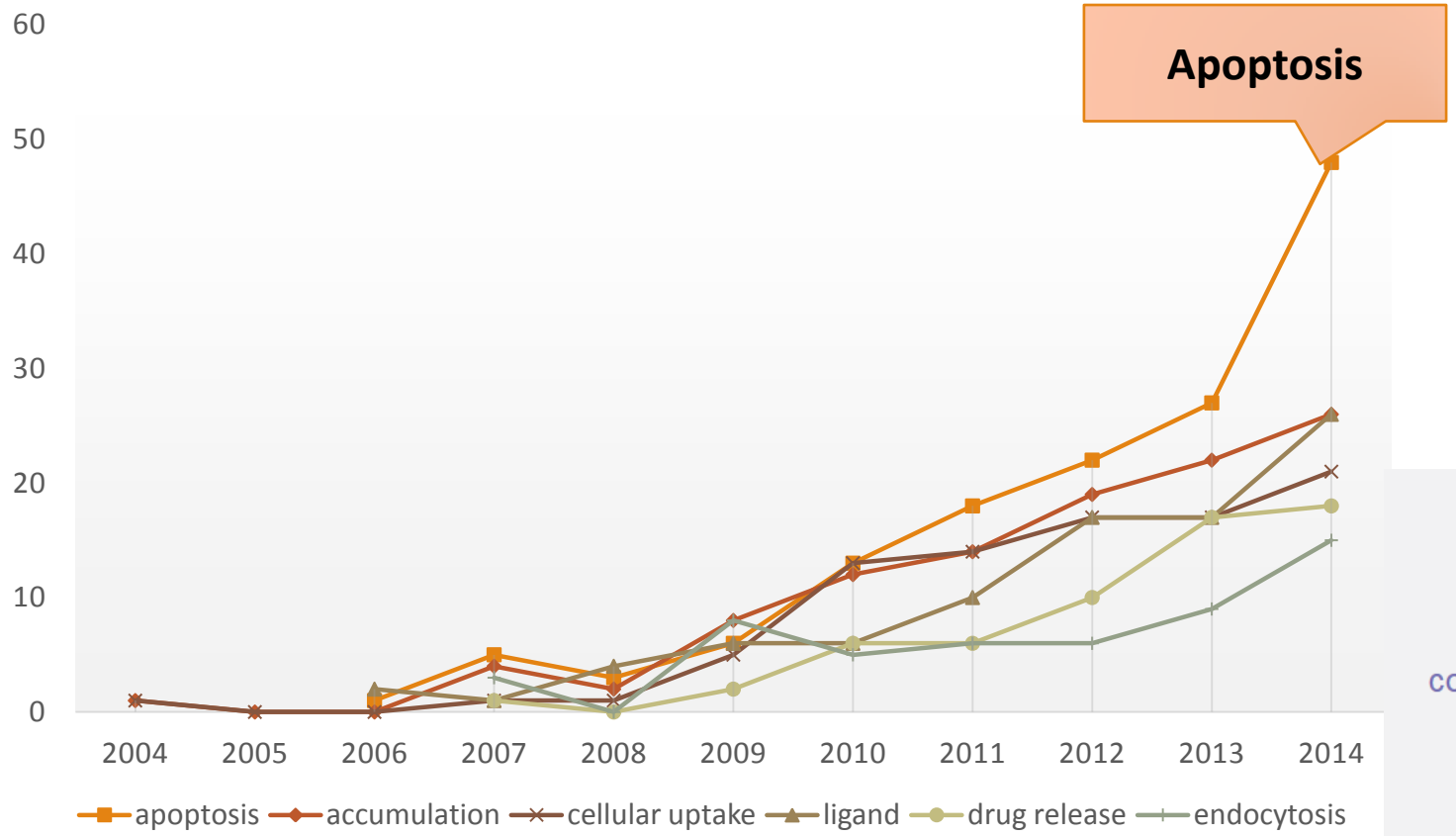
GNPs for Cancer Therapy in Stage 1 – Physical Characterization





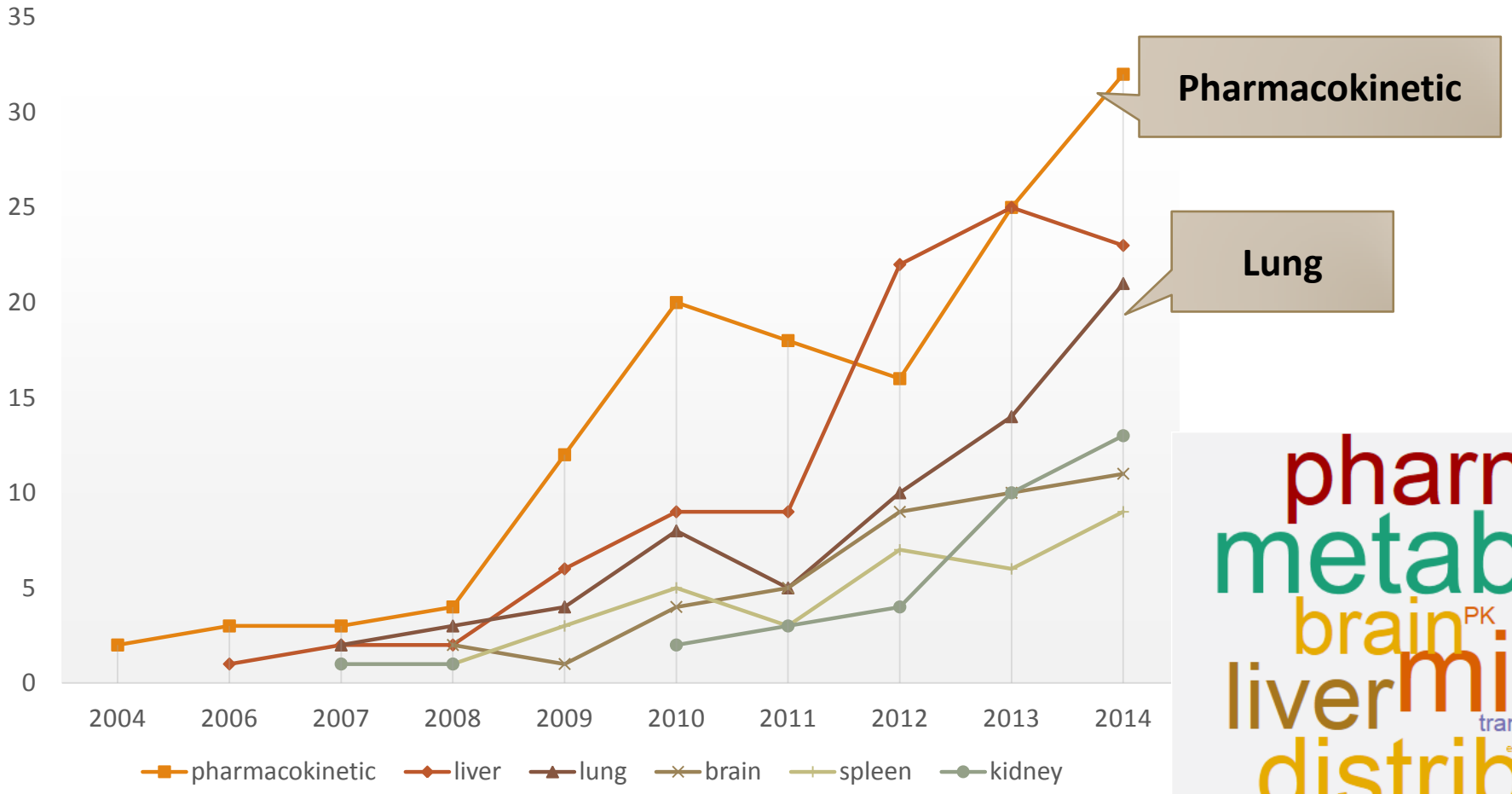
Developmental trends of several stage 1 keywords

GNPs for Cancer Therapy in Stage 2 – In Vitro

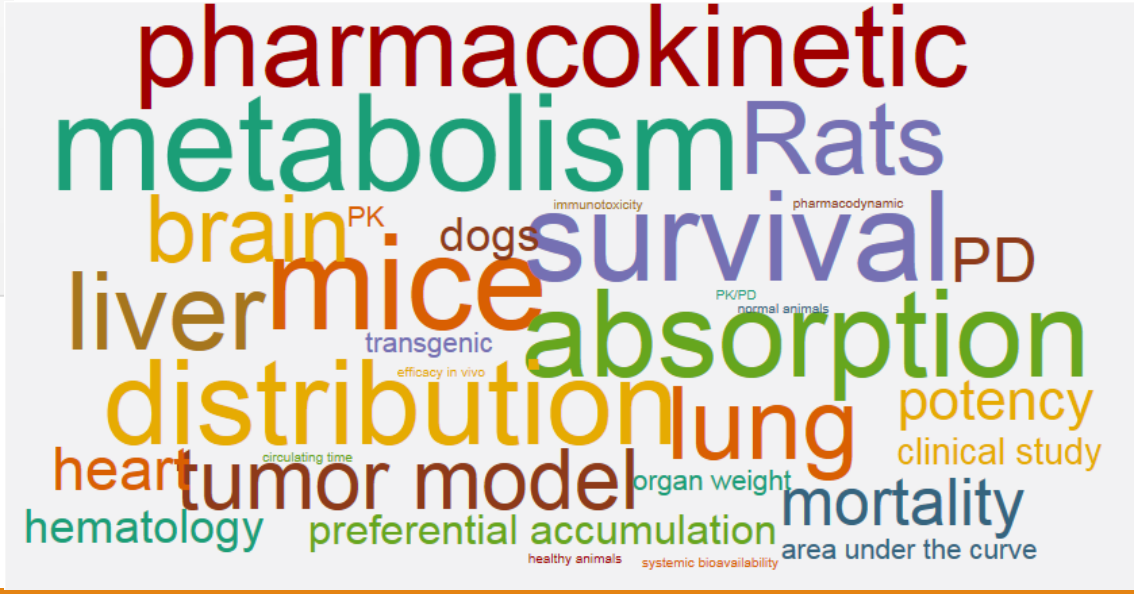


Developmental trends of several stage 2 keywords

GNPs for Cancer Therapy in Stage 3 – In Vivo

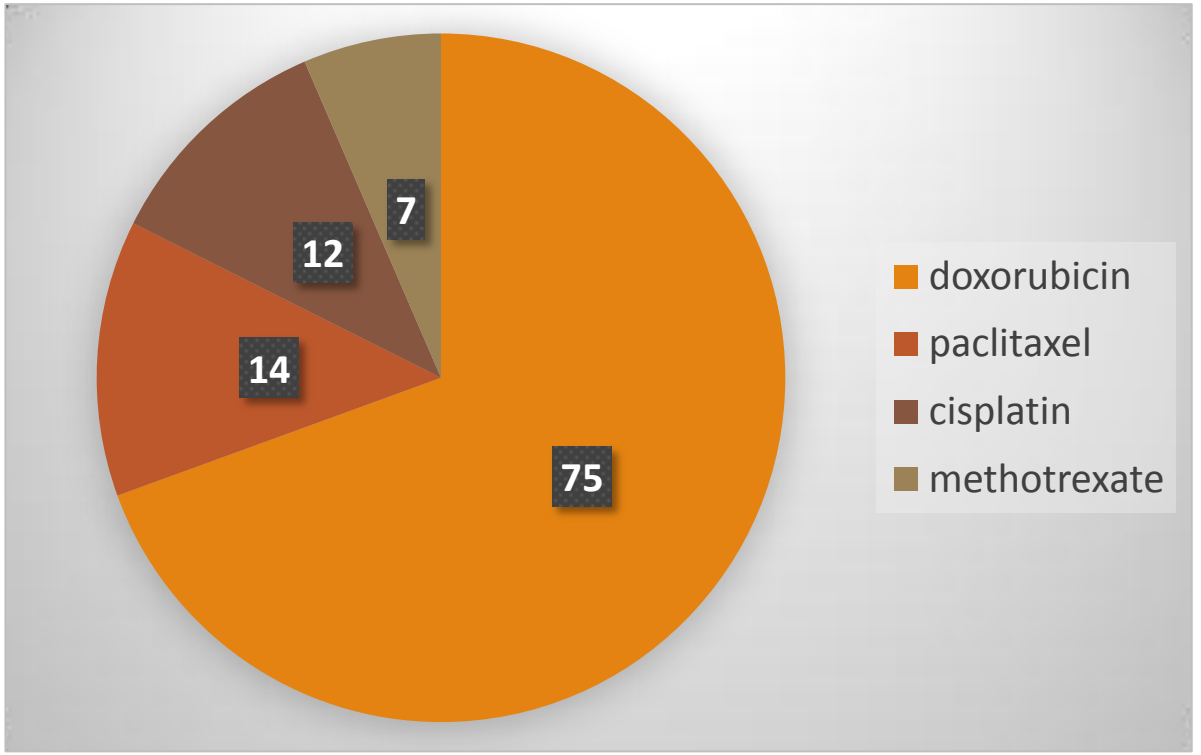


Developmental trends of several stage 3 keywords



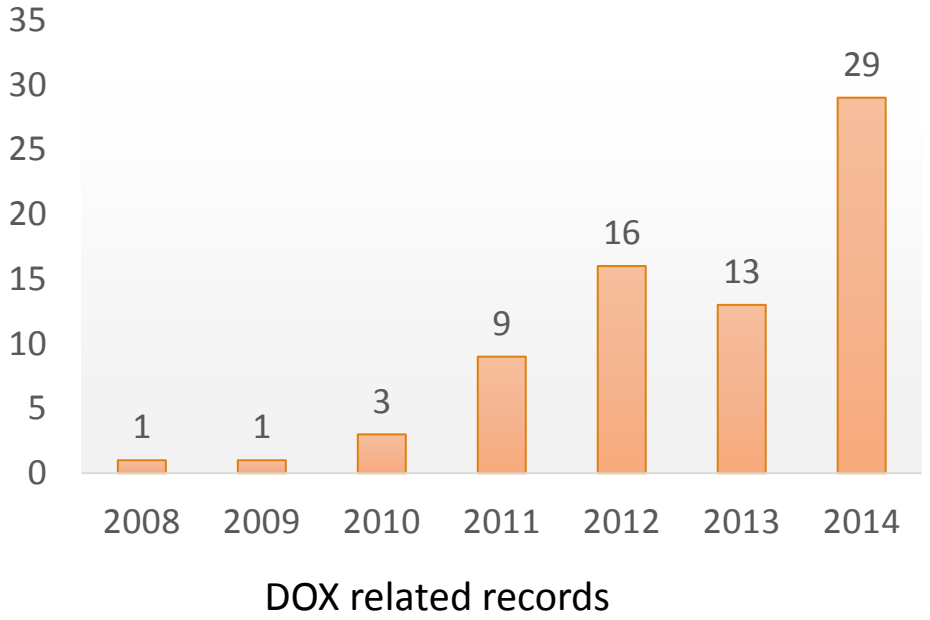
GNPs and Anticancer Drugs

- ✓ GNPs can be used for photothermal therapy, radiotherapy, and drug delivery, etc.
- ✓ How do GNPs work with anticancer drugs? Or how do GNPs facilitate the delivery of anticancer drugs?
- ✓ In the top ~500 NLP words/phrases, there are four kinds of anticancer drugs.
- ✓ #1 is **doxorubicin (DOX)** – 75 records.



GNPs and Doxorubicin (DOX)

- ✓ This topic – DOX loaded GNPs is emerging, started from 2008.
- ✓ Most (63 out of 75) studies have *in vitro assays. In vivo* studies mainly started from 2012.
- ✓ The photothermal effect of GNPs enables controlled release.
- ✓ Potential target diseases – breast cancer, lung cancer, glioma, liver cancer, etc.
- ✓ Co-delivery with siRNA has been introduced.
- ✓ Leading countries – China (27), United States (16), India (7)



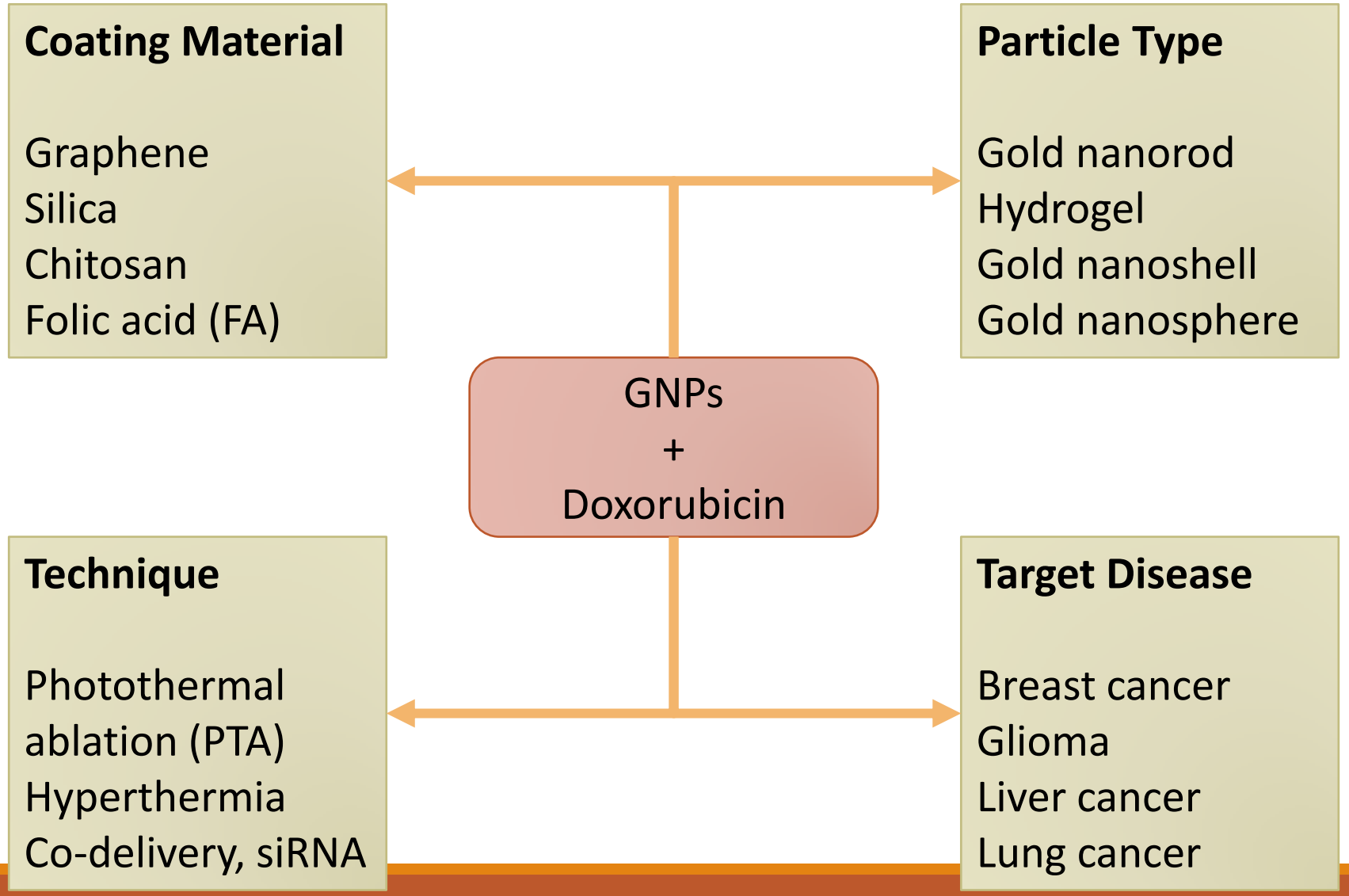
GNPs and Doxorubicin (DOX)

# Records	Stage 1	doxorubicin
186	particle size	13
180	synthesis	15
154	stability	10
150	Transmission electron microscopy	7
143	electron microscopy	8
117	diameter	8
73	shape	3
62	density	6
58	dimension	1
51	aggregation	1
35	scanning electron microscopy	2
33	absorbance	1
32	dynamic light scattering	2
30	zeta potential	6
24	aspect ratio	1
24	surface area	2
23	size distribution	2
23	X-ray diffraction	2

# Records	Stage 2	doxorubicin
490	cancer cell	24
249	toxicity	15
216	Cytotoxicity	23
150	apoptosis	9
121	cell death	8
113	accumulation	5
95	cellular uptake	9
93	ligand	11
91	cell viability	4
68	Cell proliferation	5
65	flow cytometry	7
64	inhibition	4
63	drug release	23
59	cytoplasm	4
56	necrosis	2
54	endocytosis	5
32	cell type	2
32	Confocal microscopy	5
28	MTT assay	4

# Records	Stage 3	doxorubicin
372	metabolism	14
353	mice	24
241	survival	18
196	absorption	7
170	distribution	10
137	pharmacokinetic	12
103	liver	3
80	xenograft	2
73	biodistribution	4
73	lung	4
70	Rats	1
62	tissue distribution	4
47	brain	1
35	kidney	1
33	therapeutic efficacy	4
30	tumor model	2
26	gene silencing	2
21	body weight	1

GNPs and Doxorubicin (DOX)



Conclusions

- Introducing classification algorithms to support lexical query.
  In this case, the research framework of GNPs (or other biomedical related fields) is complex. The lexical query can not be directly used for retrieving a clean dataset for a specific topic.
- Grouping translational stage-oriented keywords to locate translational clues in biomedical research.

Next

- Preliminary results, need more refinement and discussion with domain experts.
- Characterizing records into stage 1, 2 and 3 for more tracking development.
- To analyze text content -- even full text -- for tracing the translational pathways of a specific biomedical topic/technology.

- **Thanks to Dorothy Farrell, Piotr Grodzinski, and Donghua Zhu for their contributions.**

Thank you!

We acknowledge support from the US National Science Foundation (Award #1064146 – “Revealing Innovation Pathways: Hybrid Science Maps for Technology Assessment and Foresight”). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

majing881003@163.com