Cognitive Systems for Materials Science

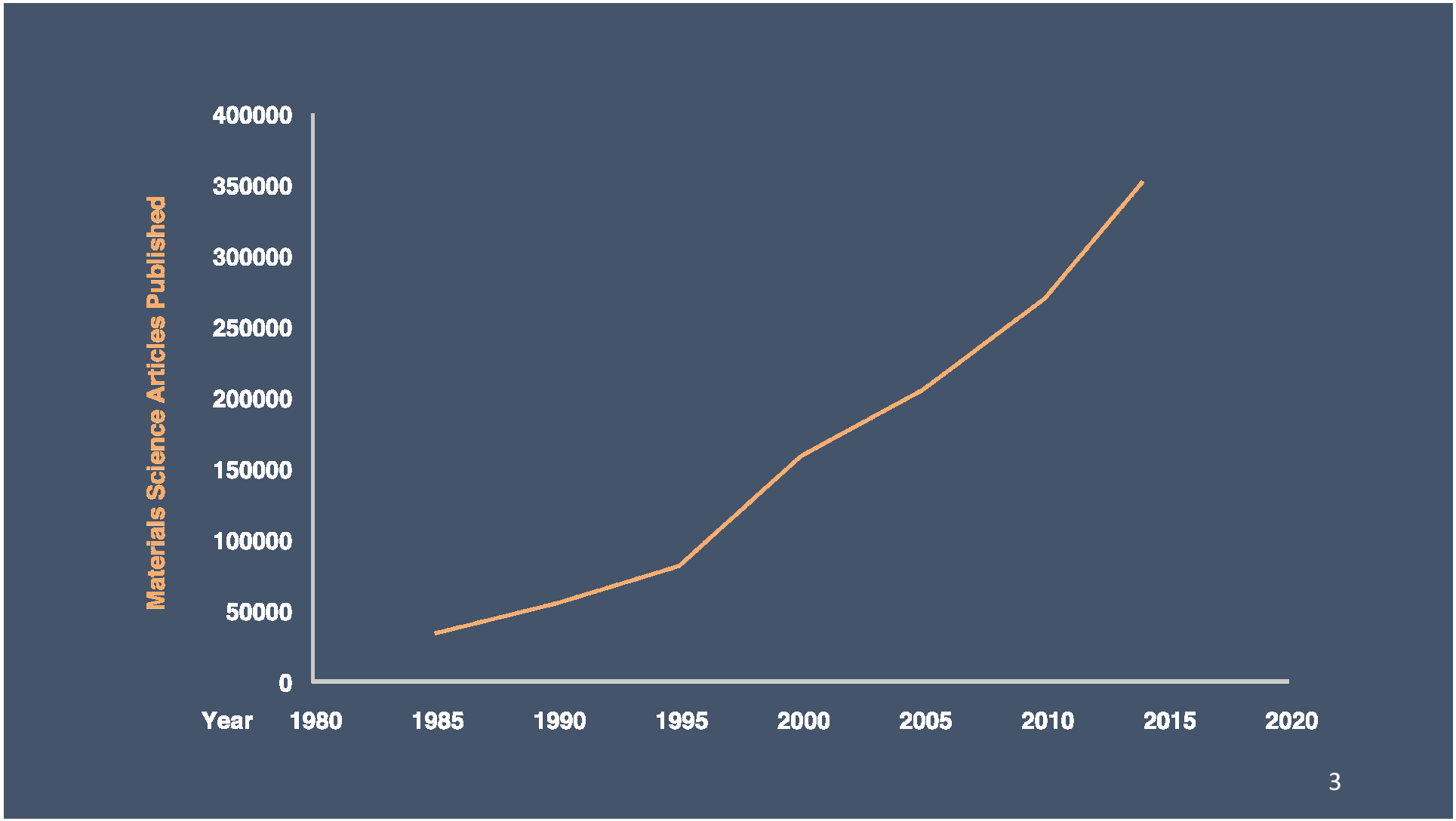
Jed W. Pitera, Ph. D.
Principal Research Staff Member & Manager, Materials Discovery
IBM Research - Almaden

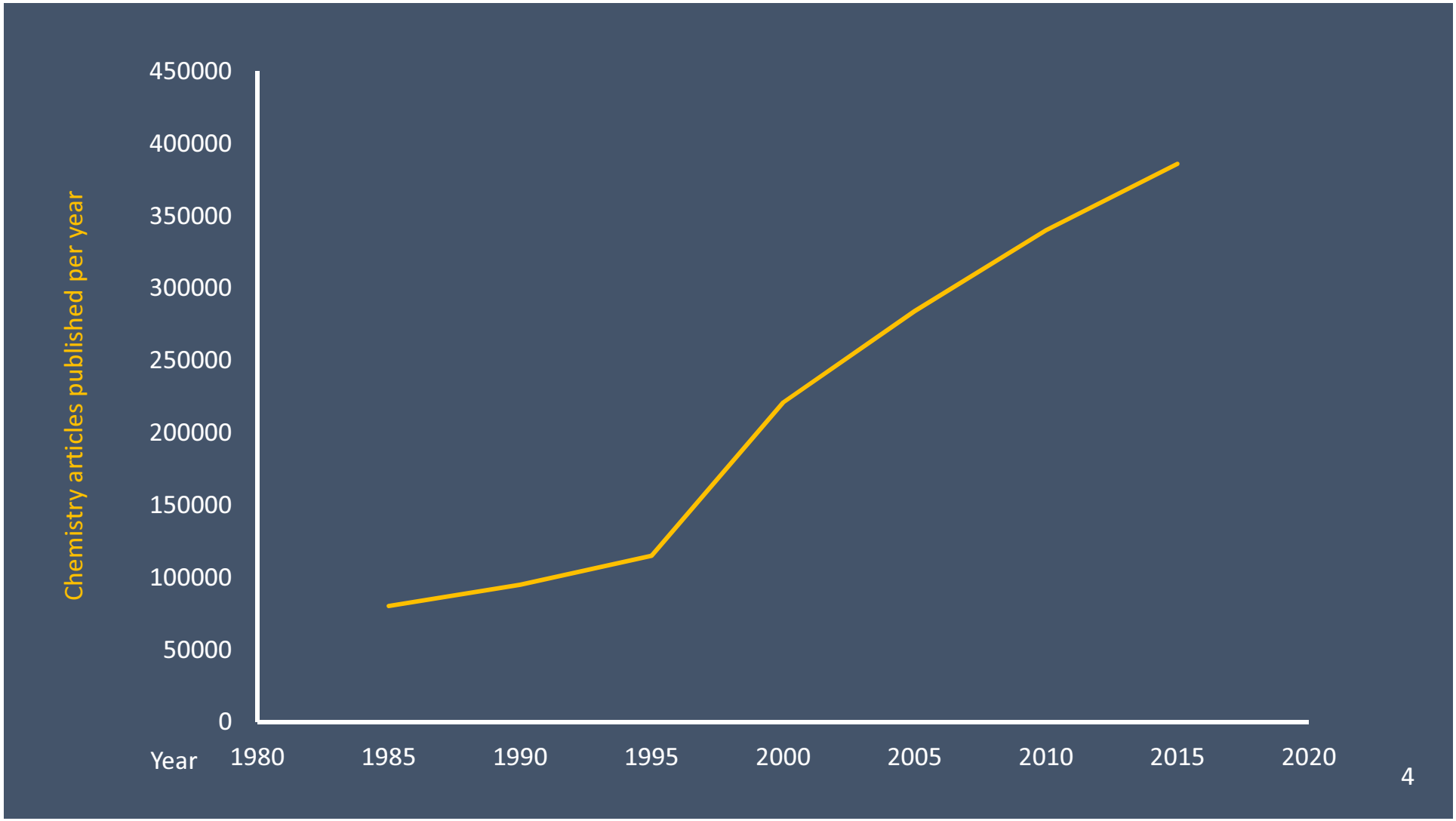
Materials R&D is *too slow*

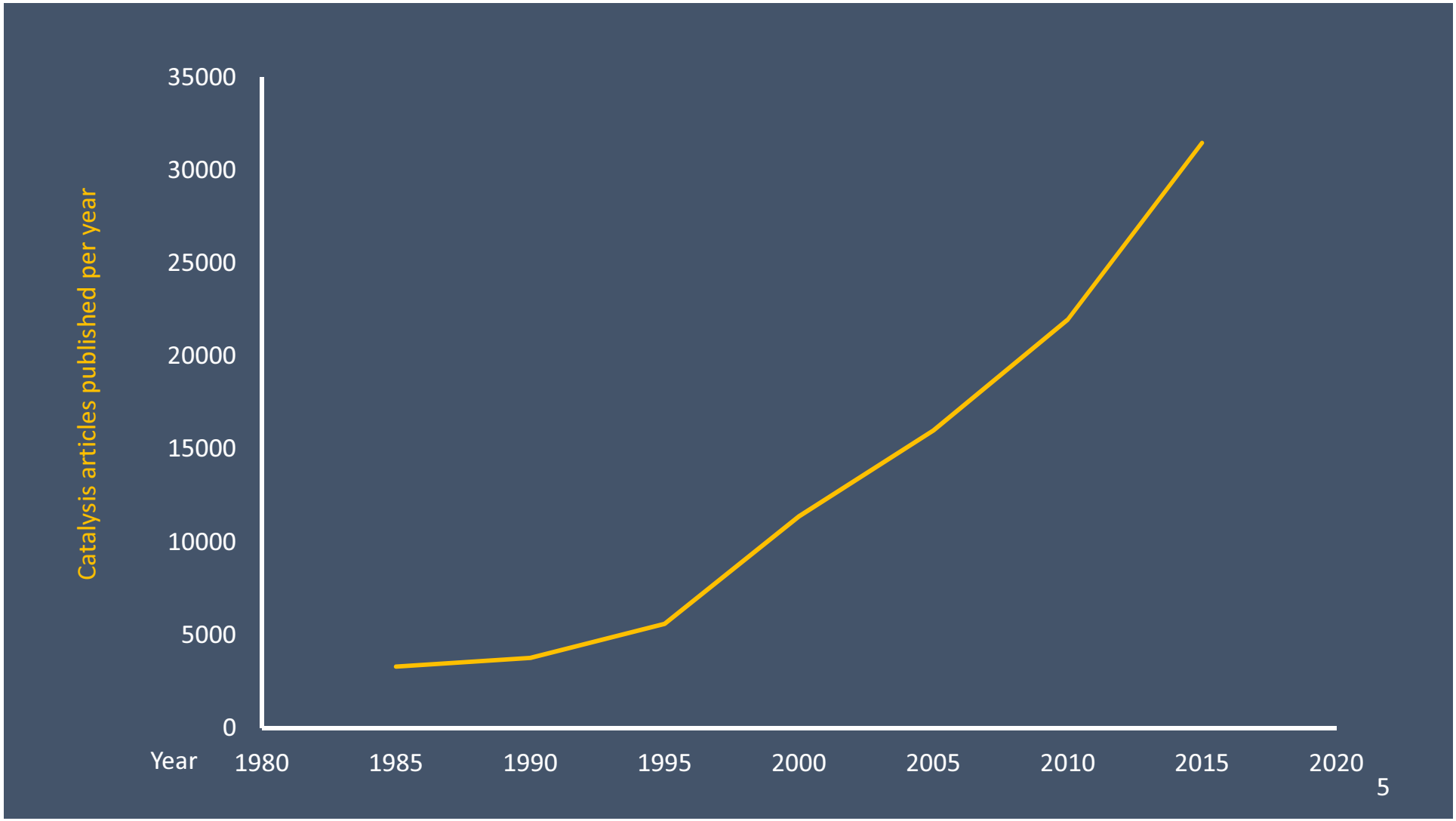
10-20 year materials development cycle

expert-driven trial and error

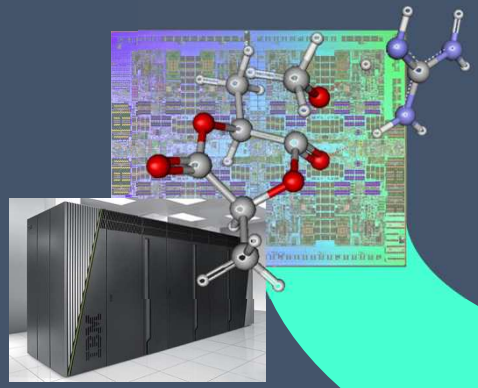
sequential process


Materials Science Articles Published
400000
350000
300000
250000
200000
150000
100000
50000
0
Year
1980
1985
1990
1995
2000
2005
2010
2015
2020


Chemistry articles published per year
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
Year
1980
1985
1990
1995
2000
2005
2010
2015
2020


Catalysis articles published per year
35000
30000
25000
20000
15000
10000
5000
0
Year
1980
1985
1990
1995
2000
2005
2010
2015
2020

ACCELERATED MATERIALS DISCOVERY

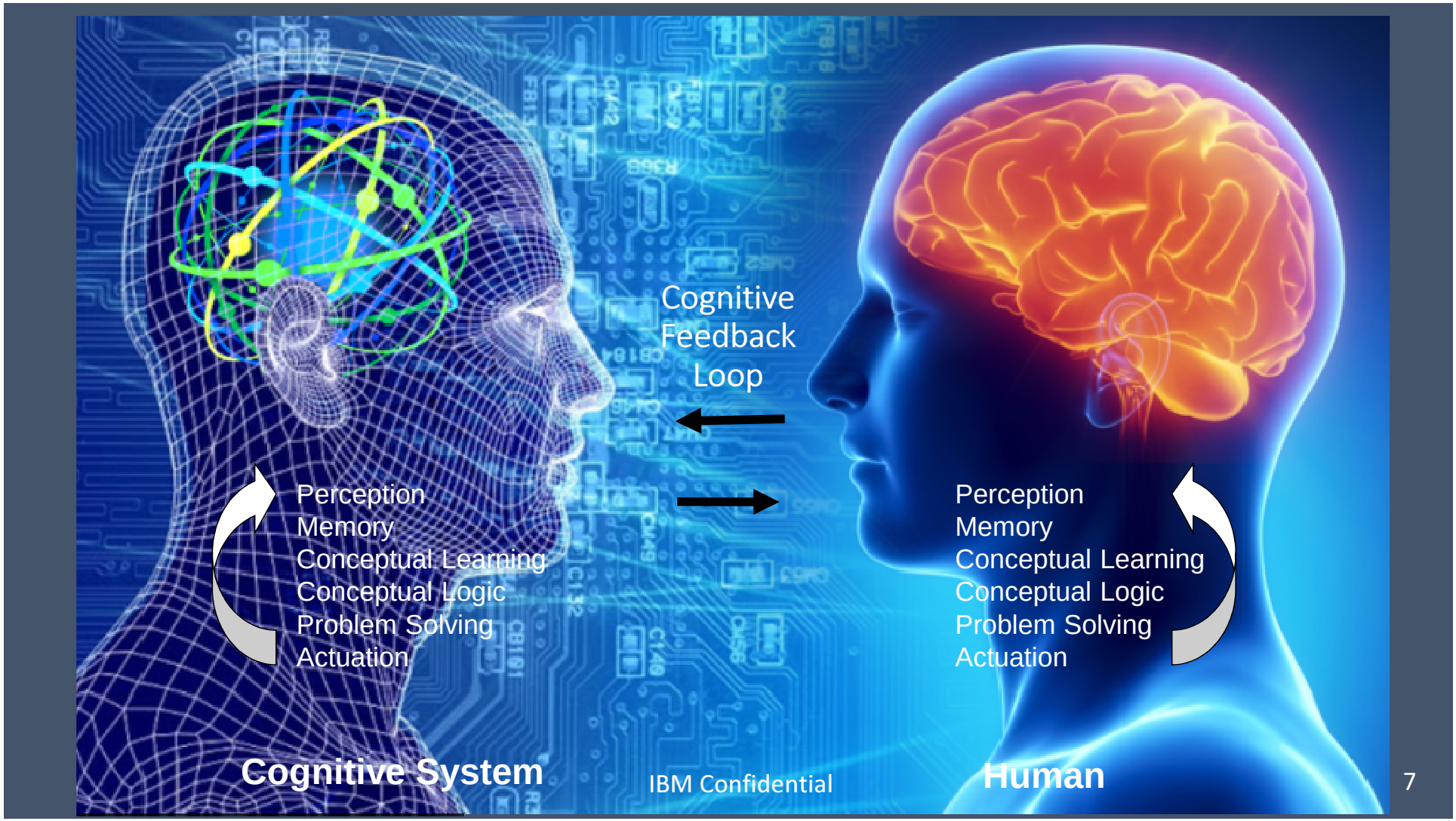


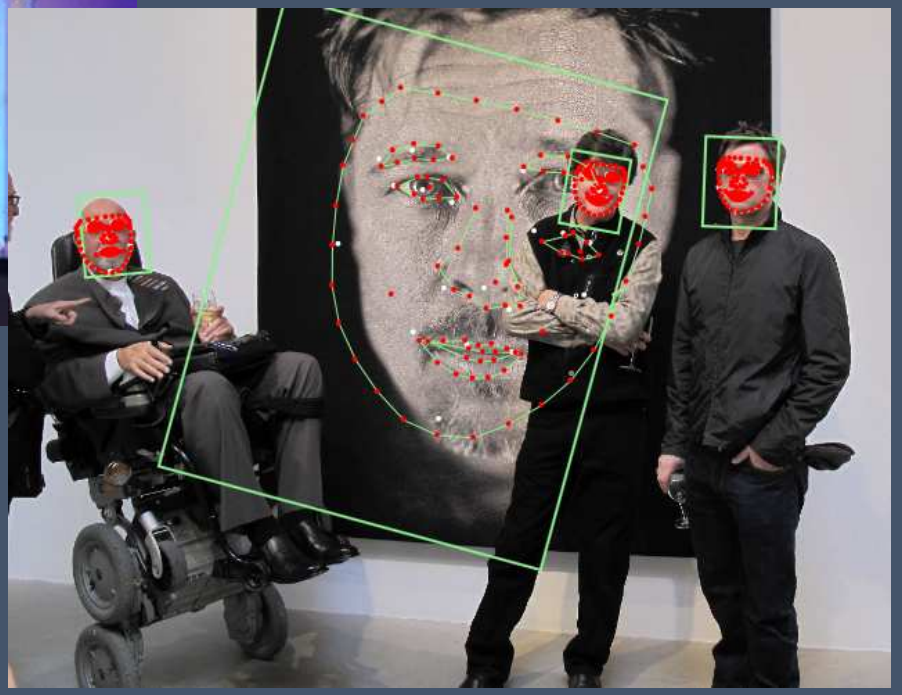
<u>What can be calculated</u>
Predictive models
Simulations, statistics, correlations

<u>Cognitive Systems</u>

New materials
New capabilities
New solutions

<u>What is known</u>
Papers, patents, reports
Experiments, extraction, similarity



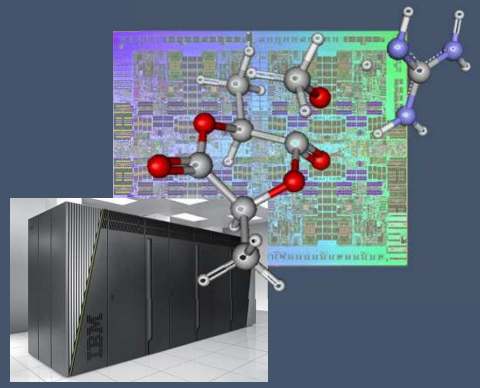


Common feature:
PRACTICE

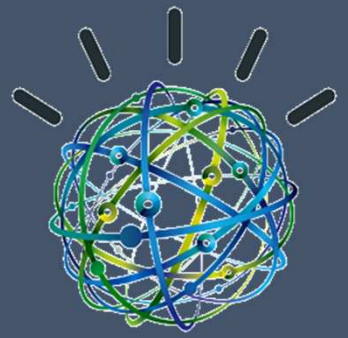
The stages of discovery

Edisonian

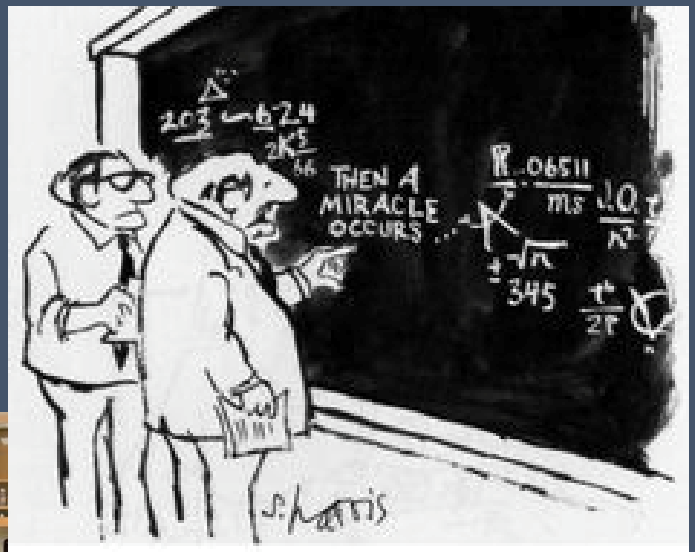
Model-based



Cognitive



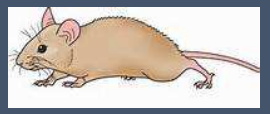
The Cognitive Journey



"I think you should be more explicit here in step two."

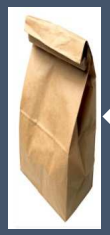
from *What's so Funny about Science?* by Sidney Harris (1977)

Discovering new P53 kinases



Bag of Words Model

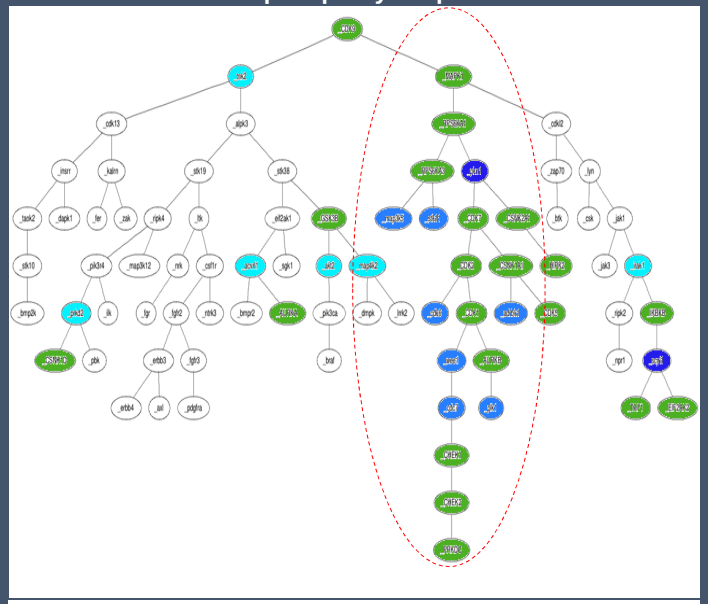
CHEK2 **BMPR2**



1. Quantify the "literature distances" between proteins

$$\mathcal{D} = \begin{pmatrix} d_{11} & d_{12} & \cdots & \cdots & \cdots & d_{1n} \\ \vdots & d_{22} & & & & \vdots \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & \cdots & \cdots & d_{nn} \end{pmatrix}.$$

2. Cluster proteins by their literature distance: known p53 kinases cluster (green) and suggest others that may also phosphorylate p53.



GREEN – p53 Kinases
RED/ORANGE/YELLOW – Predicted Targets

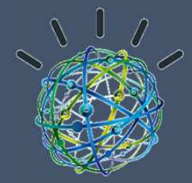
High-level system diagram

unstructured data: patents, papers, technical reports

Data Extraction

Knowledge Graph & Analytics

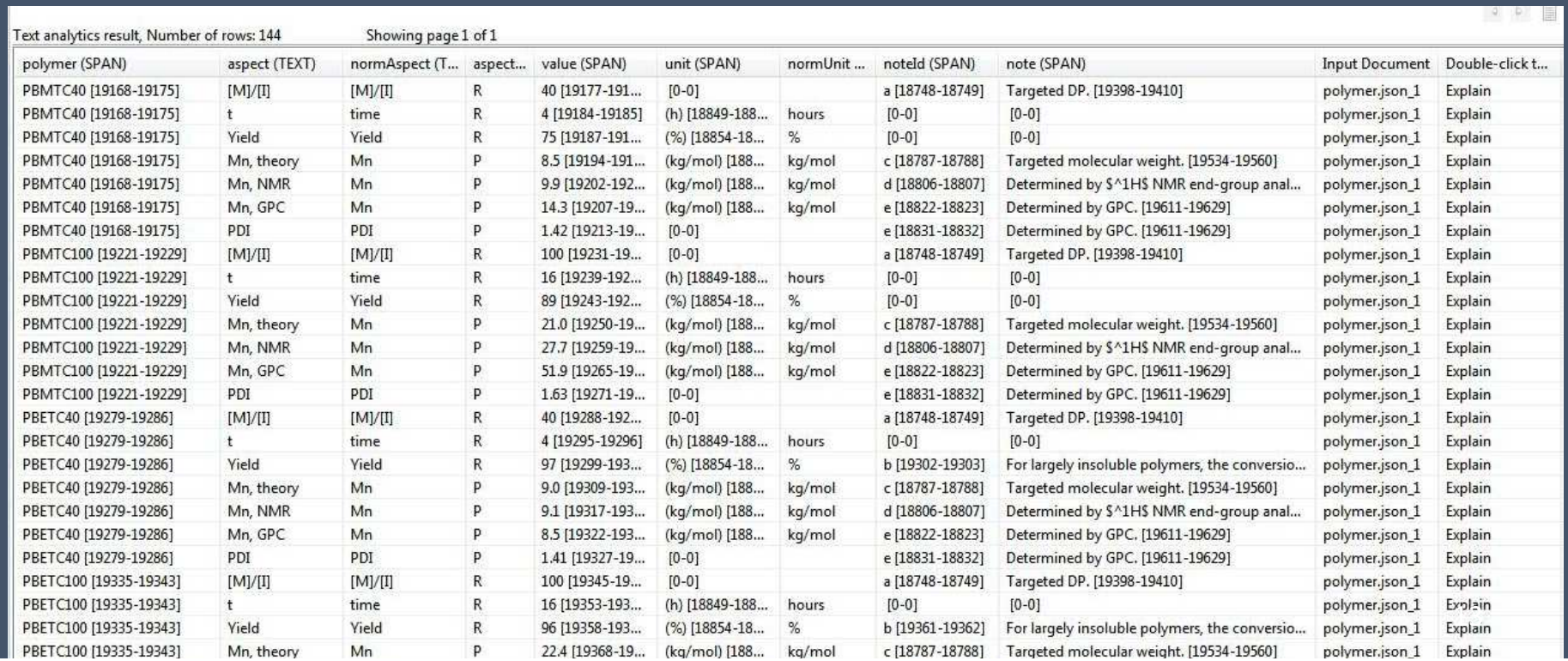
User Interface & Query



Exploration & Discovery

Automated annotation of polymer properties

w/ KK Rao, Ankit Vora, Kristin Schmid & Victoria Piunova

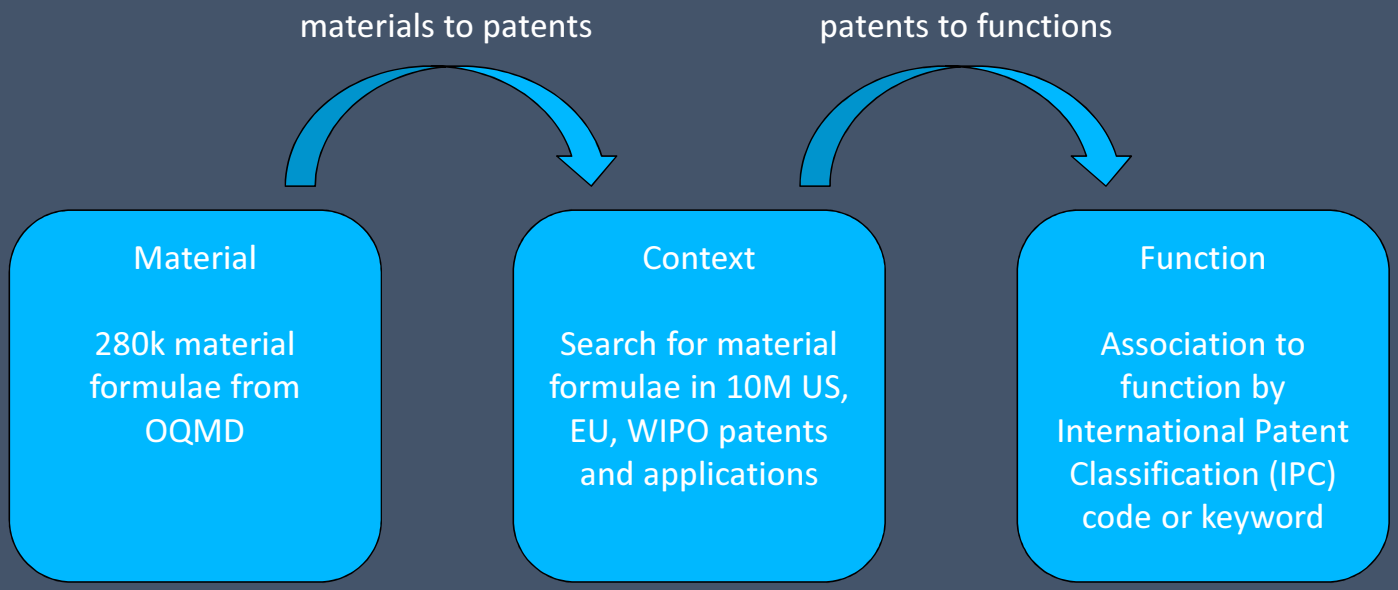


Text analytics result, Number of rows: 144 Showing page 1 of 1

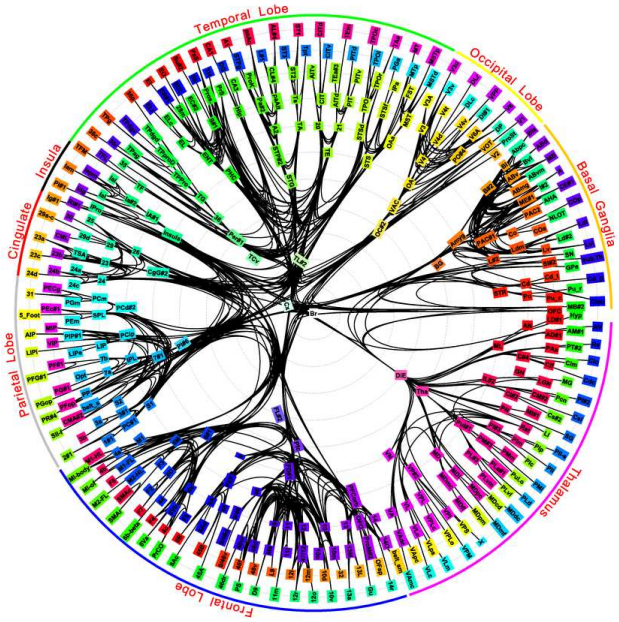
polymer (SPAN)	aspect (TEXT)	normAspect (T...	aspect...	value (SPAN)	unit (SPAN)	normUnit ...	noteId (SPAN)	note (SPAN)	Input Document	Double-click t...
PBMTC40 [19168-19175]	[M]/[I]	[M]/[I]	R	40 [19177-191...	[0-0]		a [18748-18749]	Targeted DP. [19398-19410]	polymer.json_1	Explain
PBMTC40 [19168-19175]	t	time	R	4 [19184-19185]	(h) [18849-188...	hours	[0-0]	[0-0]	polymer.json_1	Explain
PBMTC40 [19168-19175]	Yield	Yield	R	75 [19187-191...	(%) [18854-18...	%	[0-0]	[0-0]	polymer.json_1	Explain
PBMTC40 [19168-19175]	Mn, theory	Mn	P	8.5 [19194-191...	(kg/mol) [188...	kg/mol	c [18787-18788]	Targeted molecular weight. [19534-19560]	polymer.json_1	Explain
PBMTC40 [19168-19175]	Mn, NMR	Mn	P	9.9 [19202-192...	(kg/mol) [188...	kg/mol	d [18806-18807]	Determined by ^1H NMR end-group anal...	polymer.json_1	Explain
PBMTC40 [19168-19175]	Mn, GPC	Mn	P	14.3 [19207-19...	(kg/mol) [188...	kg/mol	e [18822-18823]	Determined by GPC. [19611-19629]	polymer.json_1	Explain
PBMTC40 [19168-19175]	PDI	PDI	P	1.42 [19213-19...	[0-0]		e [18831-18832]	Determined by GPC. [19611-19629]	polymer.json_1	Explain
PBMTC100 [19221-19229]	[M]/[I]	[M]/[I]	R	100 [19231-19...	[0-0]		a [18748-18749]	Targeted DP. [19398-19410]	polymer.json_1	Explain
PBMTC100 [19221-19229]	t	time	R	16 [19239-192...	(h) [18849-188...	hours	[0-0]	[0-0]	polymer.json_1	Explain
PBMTC100 [19221-19229]	Yield	Yield	R	89 [19243-192...	(%) [18854-18...	%	[0-0]	[0-0]	polymer.json_1	Explain
PBMTC100 [19221-19229]	Mn, theory	Mn	P	21.0 [19250-19...	(kg/mol) [188...	kg/mol	c [18787-18788]	Targeted molecular weight. [19534-19560]	polymer.json_1	Explain
PBMTC100 [19221-19229]	Mn, NMR	Mn	P	27.7 [19259-19...	(kg/mol) [188...	kg/mol	d [18806-18807]	Determined by ^1H NMR end-group anal...	polymer.json_1	Explain
PBMTC100 [19221-19229]	Mn, GPC	Mn	P	51.9 [19265-19...	(kg/mol) [188...	kg/mol	e [18822-18823]	Determined by GPC. [19611-19629]	polymer.json_1	Explain
PBMTC100 [19221-19229]	PDI	PDI	P	1.63 [19271-19...	[0-0]		e [18831-18832]	Determined by GPC. [19611-19629]	polymer.json_1	Explain
PBETC40 [19279-19286]	[M]/[I]	[M]/[I]	R	40 [19288-192...	[0-0]		a [18748-18749]	Targeted DP. [19398-19410]	polymer.json_1	Explain
PBETC40 [19279-19286]	t	time	R	4 [19295-19296]	(h) [18849-188...	hours	[0-0]	[0-0]	polymer.json_1	Explain
PBETC40 [19279-19286]	Yield	Yield	R	97 [19299-193...	(%) [18854-18...	%	b [19302-19303]	For largely insoluble polymers, the conversio...	polymer.json_1	Explain
PBETC40 [19279-19286]	Mn, theory	Mn	P	9.0 [19309-193...	(kg/mol) [188...	kg/mol	c [18787-18788]	Targeted molecular weight. [19534-19560]	polymer.json_1	Explain
PBETC40 [19279-19286]	Mn, NMR	Mn	P	9.1 [19317-193...	(kg/mol) [188...	kg/mol	d [18806-18807]	Determined by ^1H NMR end-group anal...	polymer.json_1	Explain
PBETC40 [19279-19286]	Mn, GPC	Mn	P	8.5 [19322-193...	(kg/mol) [188...	kg/mol	e [18822-18823]	Determined by GPC. [19611-19629]	polymer.json_1	Explain
PBETC40 [19279-19286]	PDI	PDI	P	1.41 [19327-19...	[0-0]		e [18831-18832]	Determined by GPC. [19611-19629]	polymer.json_1	Explain
PBETC100 [19335-19343]	[M]/[I]	[M]/[I]	R	100 [19345-19...	[0-0]		a [18748-18749]	Targeted DP. [19398-19410]	polymer.json_1	Explain
PBETC100 [19335-19343]	t	time	R	16 [19353-193...	(h) [18849-188...	hours	[0-0]	[0-0]	polymer.json_1	Explain
PBETC100 [19335-19343]	Yield	Yield	R	96 [19358-193...	(%) [18854-18...	%	b [19361-19362]	For largely insoluble polymers, the conversio...	polymer.json_1	Explain
PBETC100 [19335-19343]	Mn, theory	Mn	P	22.4 [19368-19...	(kg/mol) [188...	kg/mol	c [18787-18788]	Targeted molecular weight. [19534-19560]	polymer.json_1	Explain

Exploration of materials knowledge graphs

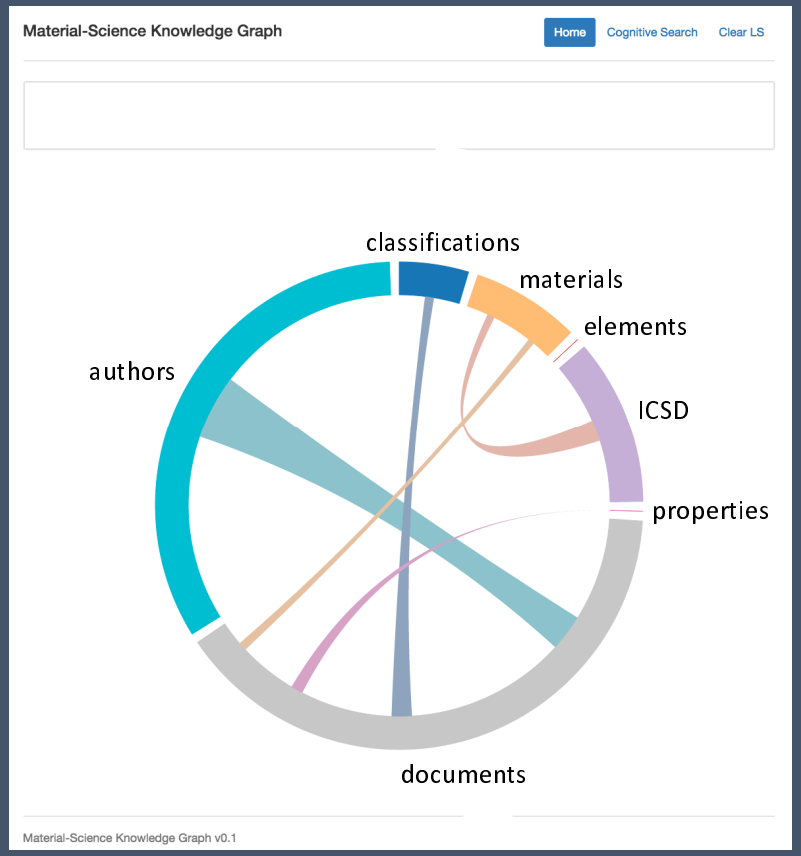
w/ Nitesh Chawla and Pingjie Tang, Notre Dame



macaque brain connectivity network, IBM, PNAS 2010

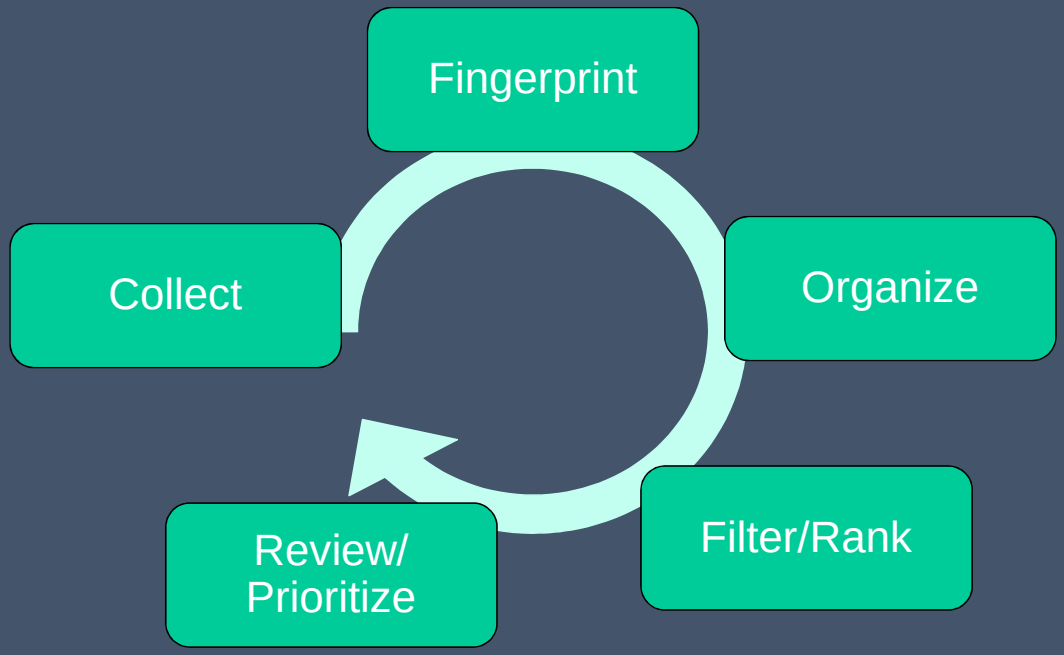


383 nodes, 6602 edges, person-decades



>1M nodes, >100k edges, 600k documents, 1 node-hour
Data from Peter Staar, IBM Research - Zurich

General Informatics Strategy




Questions?
Ideas?
Thank You