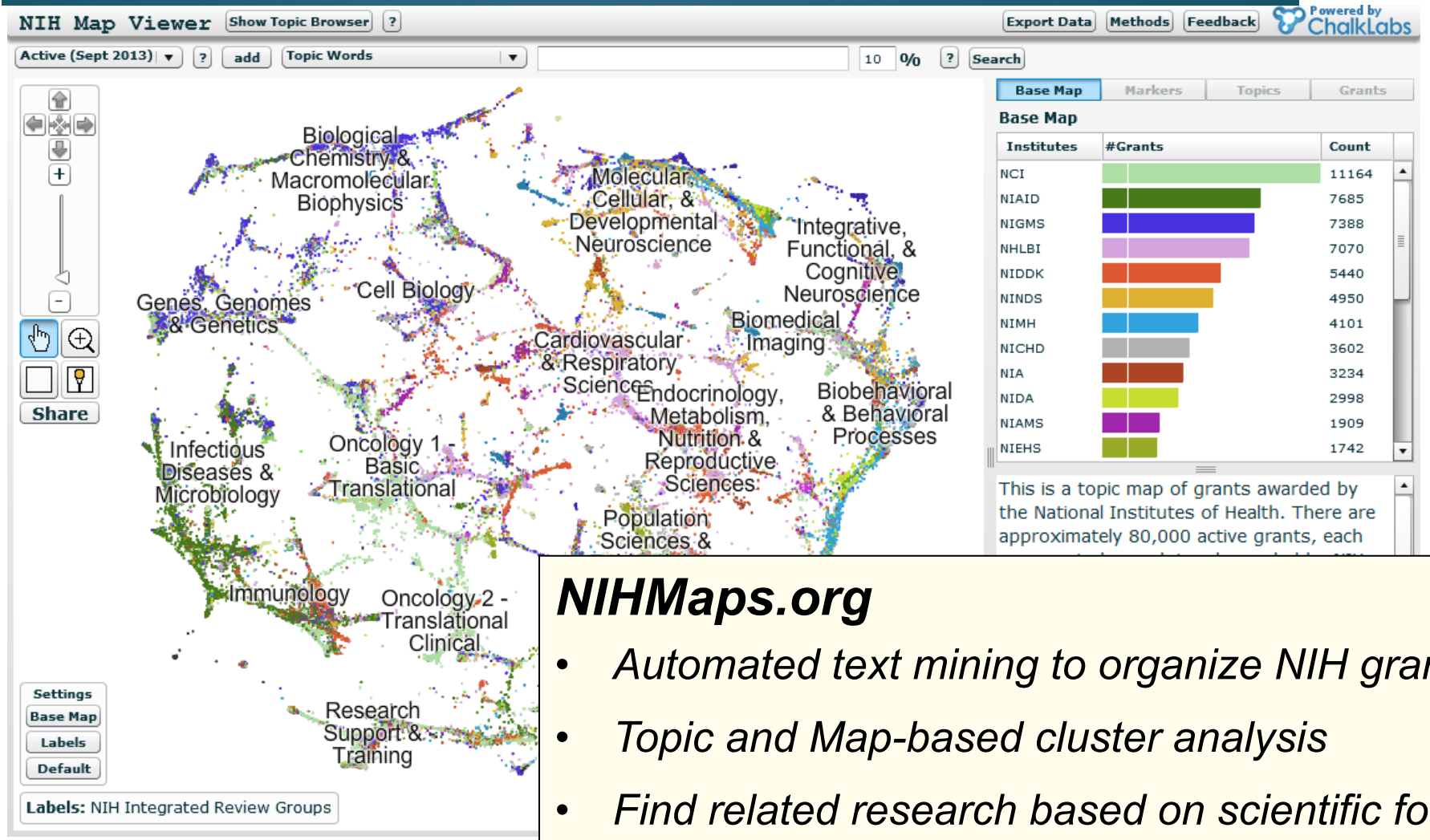
Portfolio Analysis for Basic Biomedical Research Using NIHMaps: Lessons Learned and Future Possibilities

Lisa Dunbar
National Institute of General Medical Sciences (NIGMS)

Ned Talley
National Institute of Neurological Disorders and Stroke (NINDS)

A Topic Database of NIH-Funded Grants



NIHMaps.org

- *Automated text mining to organize NIH grants*
- *Topic and Map-based cluster analysis*
- *Find related research based on scientific focus*
- *Visualize against administrative categories*

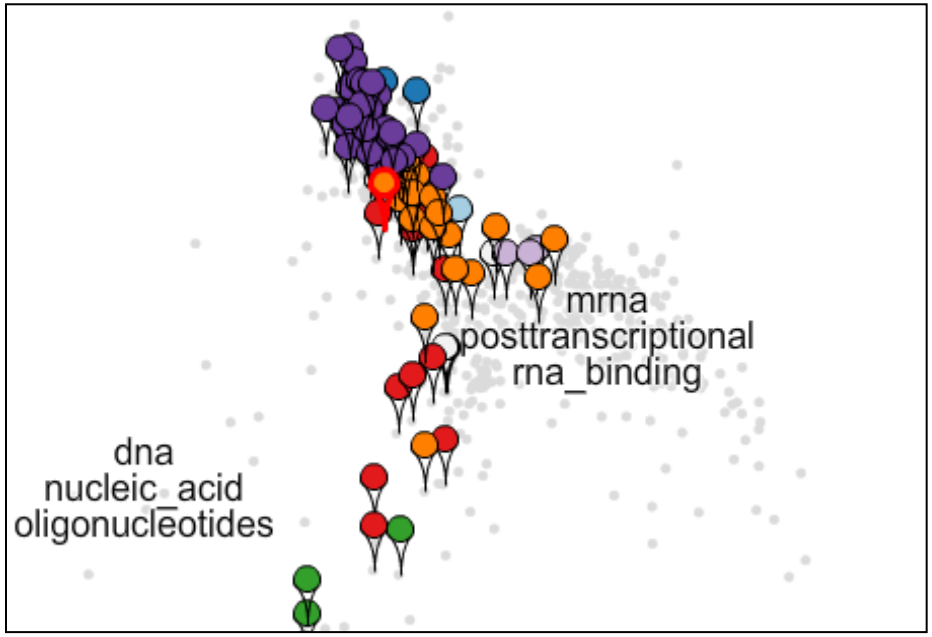
Two Independent Clustering Methods

Map-Based Clustering

- *~1000 clusters, organized in 2D space to reveal local and global relations*

Topic Modeling

- *"Soft" clustering – documents fall in multiple categories*
- *Automatic determination of latent categories from word occurrences in text*
- *Context sensitive – accommodates diverse word meanings*
- *Doesn't rely on biomedical thesaurus – useful for basic research categories*

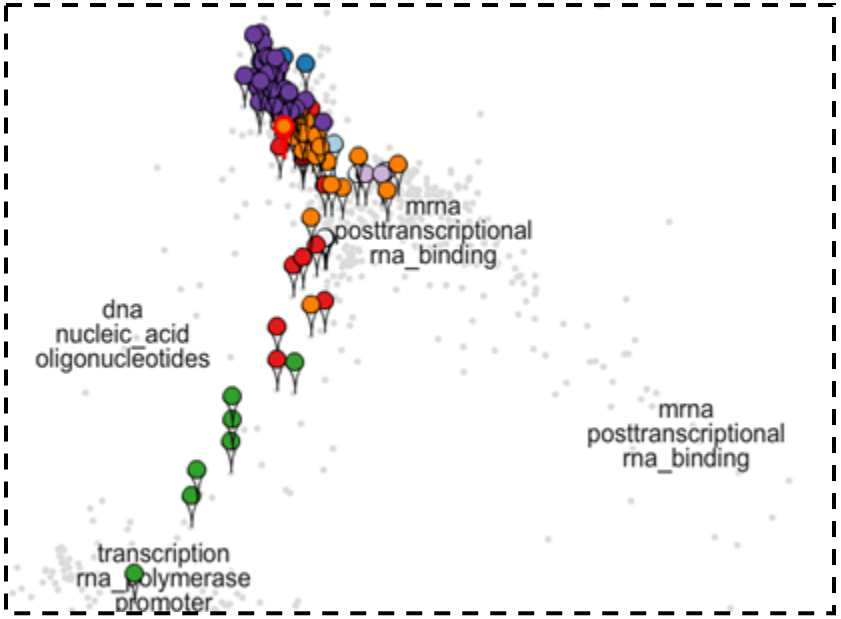


Top Topic	#Grants	Count
trna ribosome codon ribosomal...		33
translation mrna protein_synth...		23
rna_processing rna_binding pro...		16
transcription rna_polymerase pr...		8
selenium selenoprotein selenoc...		3
mrna posttranscriptional rna_bi...		3
binding conformational structur...		2

Two Independent Clustering Methods

Map-Based Clustering

- ~1000 clusters, organized in 2D space to reveal local and global relations



Topic Modeling

- "Soft" clustering – documents fall in multiple categories
- Automatic determination of latent categories from word occurrences in text
- Context sensitive – accommodates diverse word meanings
- Doesn't rely on biomedical thesaurus – useful for basic research categories

Top Topic	#Grants	Count
trna ribosome codon ribosomal...		33
translation mrna protein_synth...		23
rna_processing rna_binding pro...		16
transcription rna_polymerase pr...		8
selenium selenoprotein selenoc...		3
mrna posttranscriptional rna_bi...		3
binding conformational structur...		2

Topic Modeling vs. NIH Research Condition Disease Categorization (RCDC) System

RCDC System

- Supervised classification based on keyword weighting
- Biomedical thesaurus extracts keywords/concepts
- Domain experts set weightings to classify grants for official NIH spending reports

Topic Modeling

- Unsupervised clustering – categories used for discovery of hidden trends
- Doesn’t use thesaurus – good for basic research and for technologies important to biomedicine
- Same words assigned to different concepts depending on context, topics quantified by word count

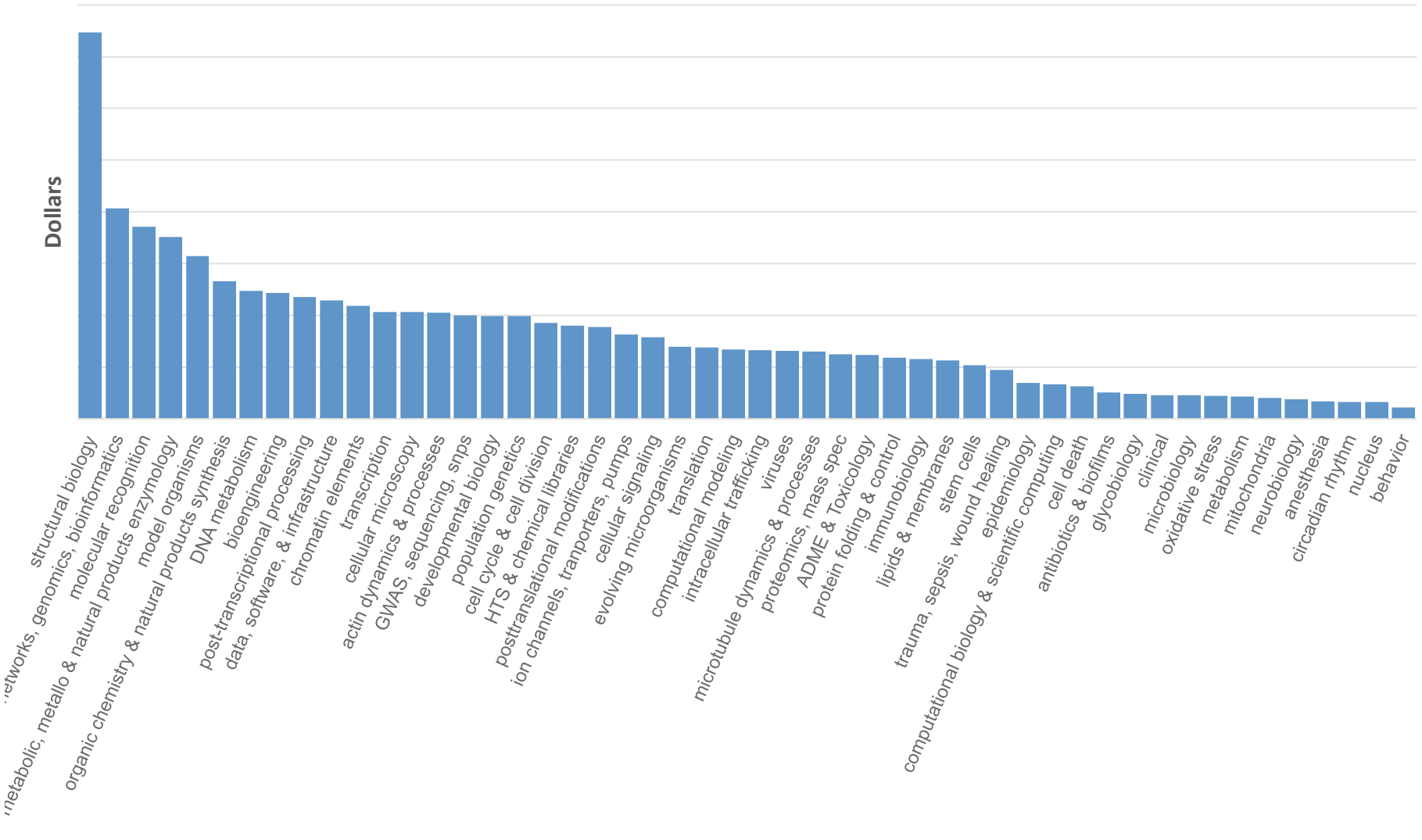
Why Use an Automated Tool for Portfolio Analysis?

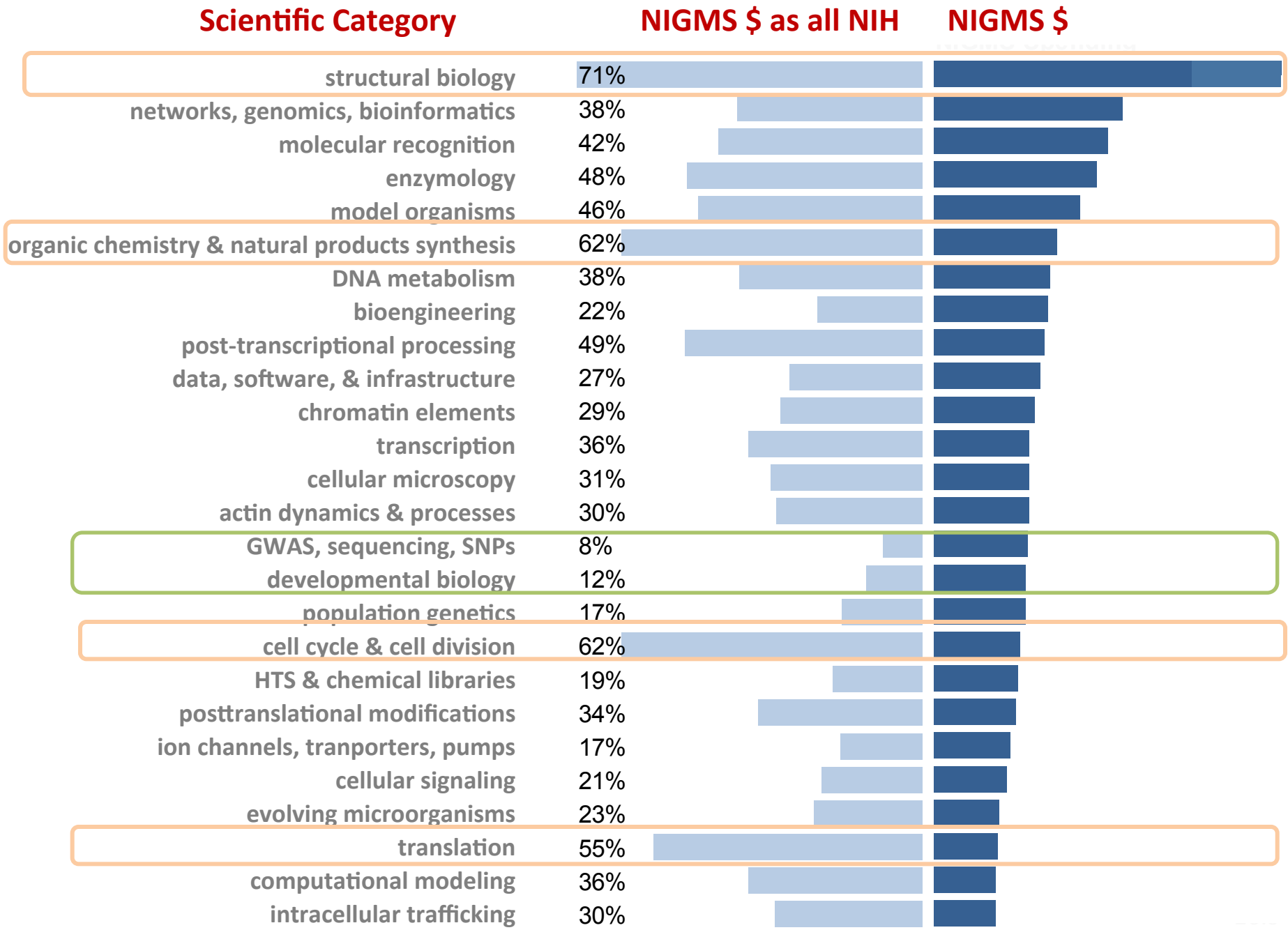
- **Objectivity** – not tied to rater's biases and expertise
- **Consistency** – no inter rater or intra rater variability
- **Transparency** – algorithm can be shared
- **Tunable** – usually quantitative parameters that can be adjusted
- **Scalable** – can be applied to vast numbers of documents and variables

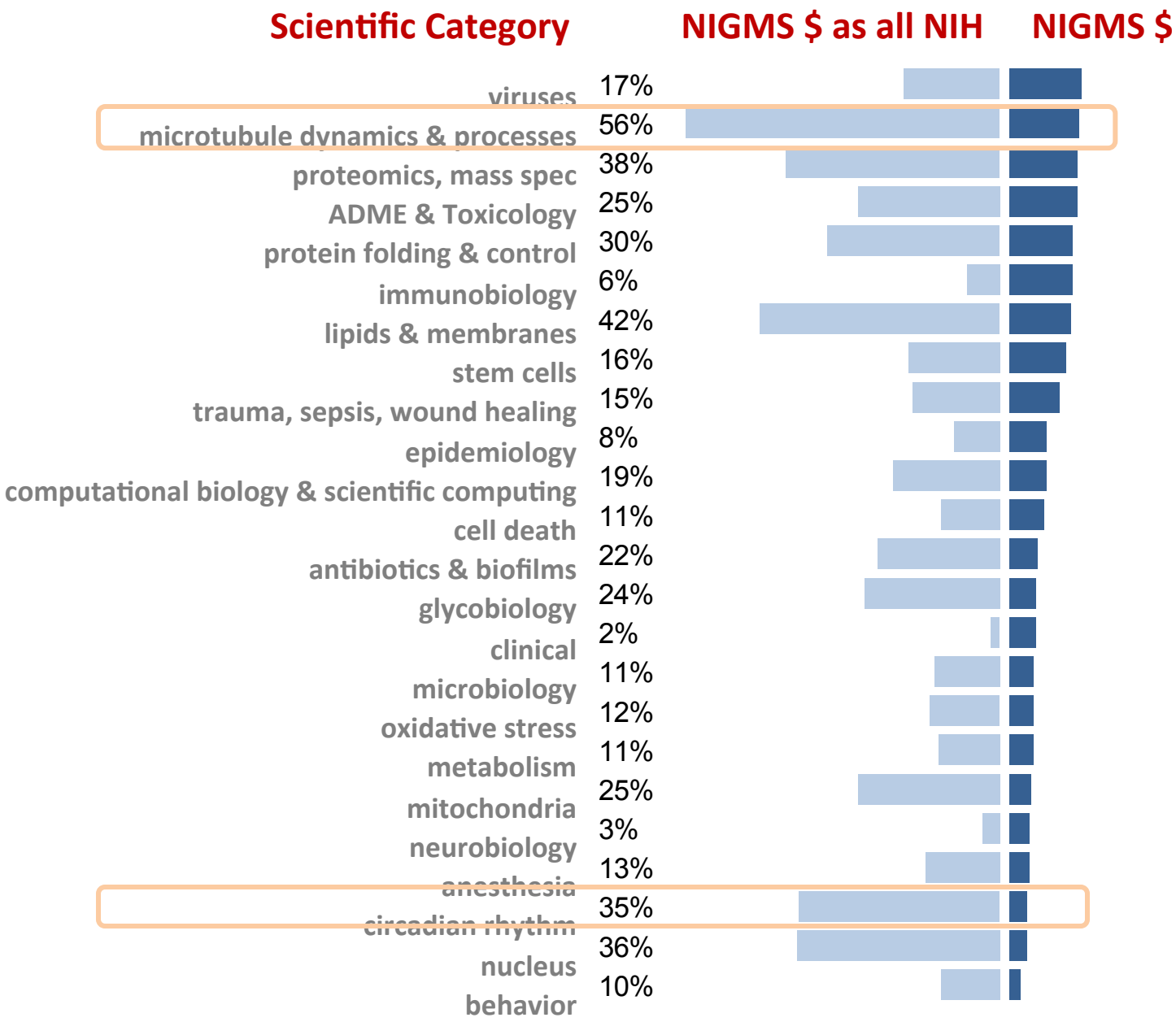
Manual curation and validation still required

- Excluded non-scientific topics; applied threshold weighting; restricted topics to those accounting for 90% of GM $
- Consolidated ~350 topics into 50 categories based on topic co-occurrence and subject matter experts ***labor intensive***
- Validation of grants in appropriate scientific category by subject matter experts ***labor intensive***

NIGMS Investment by Scientific Category




Scientific Category
NIGMS $ as all NIH
NIGMS $
structural biology
71%
networks, genomics, bioinformatics
38%
molecular recognition
42%
enzymology
48%
model organisms
46%
organic chemistry & natural products synthesis
62%
DNA metabolism
38%
bioengineering
22%
post-transcriptional processing
49%
data, software, & infrastructure
27%
chromatin elements
29%
transcription
36%
cellular microscopy
31%
actin dynamics & processes
30%
GWAS, sequencing, SNPs
8%
developmental biology
12%
population genetics
17%
cell cycle & cell division
62%
HTS & chemical libraries
19%
posttranslational modifications
34%
ion channels, tranporters, pumps
17%
cellular signaling
21%
evolving microorganisms
23%
translation
55%
computational modeling
36%
intracellular trafficking
30%


Scientific Category
NIGMS $ as all NIH
NIGMS $
viruses
microtubule dynamics & processes
proteomics, mass spec
ADME & Toxicology
protein folding & control
immunobiology
lipids & membranes
stem cells
trauma, sepsis, wound healing
epidemiology
computational biology & scientific computing
cell death
antibiotics & biofilms
glycobiology
clinical
microbiology
oxidative stress
metabolism
mitochondria
neurobiology
anesthesia
circadian rhythm
nucleus
behavior
17%
56%
38%
25%
30%
6%
42%
16%
15%
8%
19%
11%
22%
24%
2%
11%
12%
11%
25%
3%
13%
35%
36%
10%

Distribution across scientific divisions

	B	C	D	P
structural biology	15.3%	72.6%	5.1%	7.0%
networks, genomics, bioinformatics	43.4%	32.0%	17.5%	7.1%
molecular recognition	3.9%	58.7%	13.6%	23.9%
enzymology	0.9%	17.4%	8.8%	72.9%
model organisms	2.6%	17.7%	73.5%	6.2%
organic chemistry & natural products synthesis	0.5%	4.0%	1.5%	94.0%
DNA metabolism	0.1%	10.5%	78.7%	10.7%
data, software, & infrastructure	68.4%	16.4%	8.6%	6.6%
post-transcriptional processing	6.2%	20.6%	67.6%	5.6%
bioengineering	42.9%	26.9%	7.1%	23.1%
chromatin elements	1.2%	11.3%	78.9%	8.6%
GWAS, sequencing, snps	7.9%	13.7%	57.3%	21.0%
actin dynamics & processes	5.5%	71.0%	16.9%	6.6%
population genetics	11.5%	6.2%	70.5%	11.8%
transcription	1.4%	20.5%	69.7%	8.4%
developmental biology	3.3%	16.7%	70.3%	9.7%
cellular microscopy	26.5%	61.3%	5.6%	6.7%
cell cycle & cell division	1.7%	38.6%	57.0%	2.6%
posttranslational modifications	6.5%	23.8%	34.2%	35.5%
cellular signaling	4.0%	30.5%	21.6%	43.9%
HTS & chemical libraries	5.8%	32.0%	4.8%	57.5%
ion channels, tranporters, pumps	6.9%	64.0%	6.6%	22.5%
computational modeling	70.7%	11.0%	13.6%	4.7%
evolving microorganisms	8.0%	29.3%	51.7%	11.0%
ADME & Toxicology	2.6%	3.7%	2.6%	91.1%

B: Biomedical Technology, Bioinformatics, & Computational Biology

C: Cell Biology & Biophysics

D: Developmental Biology & Genetics

P: Pharmacology, Physiology, & Biological Chemistry

Detecting shared interests among Institutes

Investment as $ of NIH total for each topic and category

	NIGMS	NIBIB
structural biology	**72.4%**	**3.8%**
crystallization, crystals, x_ray_crystallography, protein, structure_determination, structur	96.4%	1.0%
structural, complexes, biochemical, molecular, structure, x_ray_crystallography, proteins,	57.8%	0.0%
computational, simulations, molecular_dynamics, protein, structures, molecular, structura	89.3%	0.8%
nuclear_magnetic_resonance, spin, structure, nmr_spectroscopy, labeled, _15n, _13c, solu	56.9%	20.7%
shape, structure, size, organization, _3d, architecture, assembly, shaped, arrangement, org	40.6%	1.9%
spectroscopy, raman_spectroscopy, spectral, sers, optical, infrared, scattering, laser, vibra	47.5%	22.0%
nuclear_magnetic_resonance, spectrometer, instrument, mhz, electron_paramagnetic_reso	38.3%	47.9%

- NIBIB (biomedical imaging and bioengineering) focuses on instrumentation and technology development
- NIGMS projects focus on application of these instruments to solve molecular structures

Detecting emerging topics and categories

Categories by percentage of $ invested in projects by age

Newer projects

Category Name	>10 Yrs	< 10 Yrs
networks, genomics, bioinformatics	35%	65%
bioengineering	39%	61%
population genetics	35%	65%
computational modeling	39%	61%
stem cells	15%	85%
computational biology & scientific computing	34%	66%
epidemiology	21%	79%
clinical	27%	73%
microbiology	21%	79%
behavior	18%	82%

Established projects

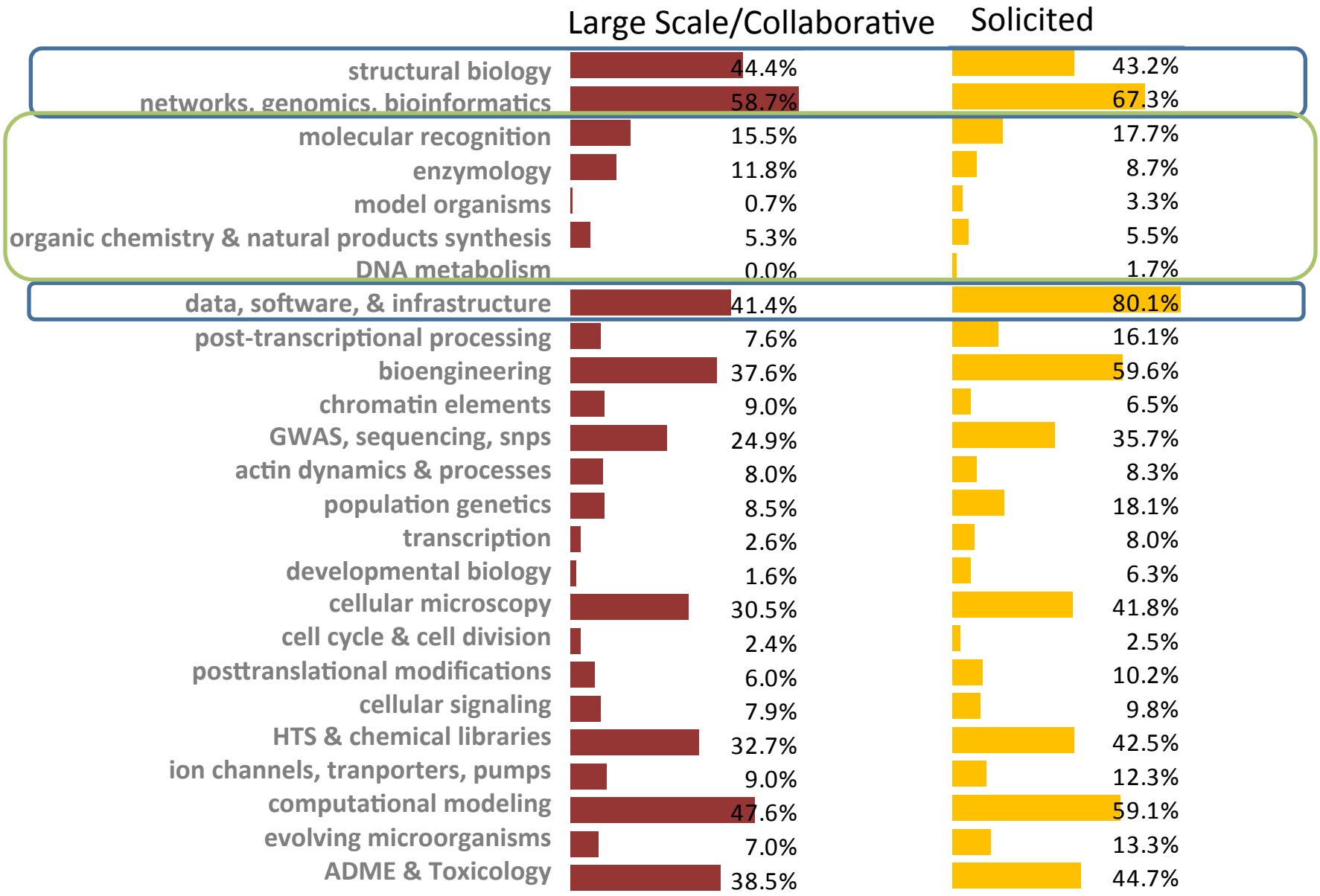
Category Name	>10 Yrs	< 10 Yrs
enzymology	62%	38%
DNA metabolism	66%	34%
transcription	64%	36%
translation	60%	40%
intracellular trafficking	66%	34%
protein folding & control	66%	34%
microtubule dynamics & processes	60%	40%
lipids & membranes	61%	39%
oxidative stress	63%	37%
circadian rhythm	65%	35%

Clusters with >59% invested in either Projects in years 1-10 or Projects in years >10 are highlighted

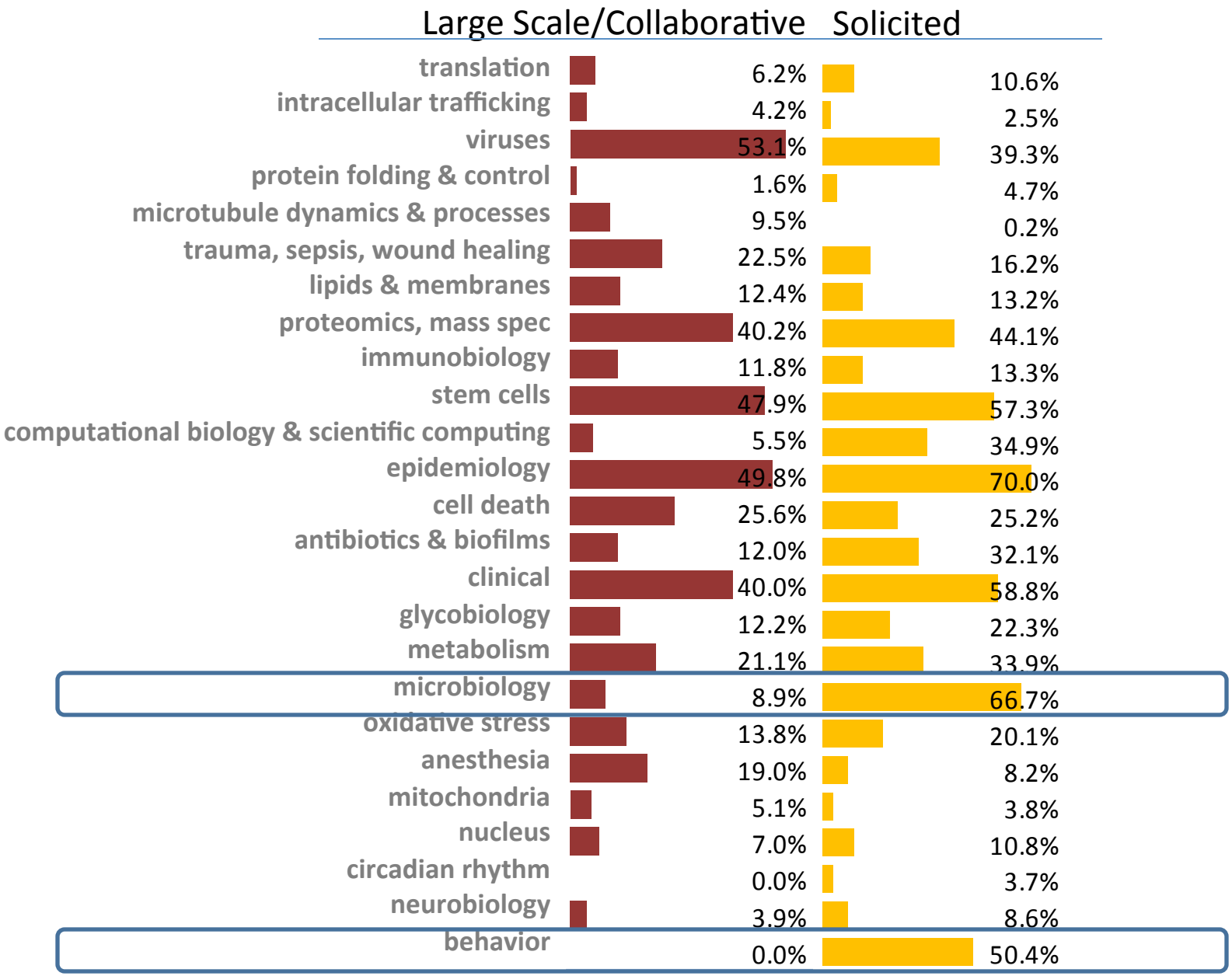
Investment by project age within scientific categories

	Estab.	New
post-transcriptional processing	**49.8%**	**50.2%**
rna_processing, rna_binding, processing, dead_box, rna, mrna, rrna, assembly, proteins, rnase, expor	61.9%	38.1%
rna_splicing, alternative_splicing, pre_mrna, rna, exon, intron, splice, mrna, transcripts, proteins	55.7%	44.3%
noncoding_rna, ncrnas, transcripts, antisense, noncoding, mrna, protein_coding, small_rna, gene_expr	32.7%	67.3%
rna, rna_interference, small_rna, silencing, sirna, gene_silencing, dicer, mirna, dsrna, editing, ar	43.3%	56.7%
microrna, mirna, mir, microrna_mirnas, mrna, targets, genes, noncoding_rna, gene_expression, _3_utr,	37.7%	62.3%
stem cells	**15.0%**	**85.0%**
pluripotent, stem_cells, differentiation, ips_cells, cells, human, reprogramming, ips, es_cells, plu	7.0%	93.0%
regeneration, stem_cells, regenerative, tissue_regeneration, regenerative_medicine, repair, adult, p	33.9%	66.1%
differentiation, stem_cell, tissue_engineering, mesenchymal_stem_cells, _3d, mscs, tissue, cells, mi	11.3%	88.7%
stem_cells, progenitor_cells, differentiation, lineage, population, markers, stem, self_renewal, adu	73.8%	26.2%

Funding mechanisms across categories



Funding mechanisms across categories



Lessons learned

Valuable exploration tool to address portfolio analysis

- Quantitative analysis of the major scientific investments of NIGMS; proportionally alloted dollars to multiple categories per grant.
- Identify areas of shared interest across Institutes
- How scientific categories map to administrative portfolio structures across and within Divisions
- Identify emerging and more established areas of science

Lessons learned

- Manual curation and validation is labor intensive and requires subject matter experts
- Granularity: Algorithms to assist in clustering topics into categories and generating meaningful labels for these categories would be valuable.
- Subsequent analyses expected to be less daunting:
  - Validation carried out – “trust” the data; know where to focus efforts
  - Categories and labels created. Do not expect major changes year-to-year