The Broad Institute Needs A Chief Data Scientist

Michael S. Noble
November 13, 2012

Observation: All Science Is Interdisciplinary

Newton inextricably linked science with math

Turing inextricably linked science with computers

∴ Biologists will write math & code

∴ And generate data. LOTS.

Case In Point: TCGA
gdac.broadinstitute.org

FAQ

Q: **Why does your table of ingested data show that *dis***
A: Our precedence rules for ingesting mutation samples ar

1. Prefer manually-curated MAF from the respective an
2. When no AWG MAF is available, fall back to using w
3. Otherwise Firehose will contain zero mutation sampl

We're in the process of defining a fourth rule, however, to ac
accrue at the DCC (again, automatically submitted by the re

For more information, please consult our provenance table t
will likely support VCFs once they become sufficiently preva

Q: **Why does your table of ingested data show that *dis***
A: We ingest and support both of the major methylation pla
statistical algorithms used by TCGA AWGs to merge both o
higher resolution data.

Q: **What TCGA sample types are Firehose pipelines executed upon?**
A: Since inception Firehose analyses have been executed upon tumor samples and then con
exception is melanoma (SKCM), which we analyze using metastatic tumor samples (code 06)
we will include a larger range of sample types, including normals.

Q: **What do you do when multiple aliquot barcodes exist for a given sample/portion/an**
A: To date GDAC analyses have proceeded upon one single tumor sample per patient, so w
metrics, we use the following rules to make such selections:

1. Prefer R analytes over T, when RNA aliquots of both type exist

2012_08_04 stddata Run

ReleaseNotes	# Datasets	% Processed	Download	
BLCA	20	100%	Open	Protected
BRCA	27	100%	Open	Protected
CESC	11	100%	Open	Protected
COADREAD	21	100%	Open	Protected
DLBC				Protected
GBM				Protected
HNSC				Protected
KIRC				Protected
KIRP				Protected
LAML				Protected
LGG				Protected
LIHC				Protected
LUAD	26	100%	Open	Protected
LUSC	34	100%	Open	Protected
OV	32	100%	Open	Protected
PAAD	6	100%	Open	Protected
PRAD	16			
SKCM	14			
STAD	18			
THCA	18			
UCEC	22			
PANCANCER	48			

Data Dashboard

2012_07_25 analyses Run

AnalysisReport	# Pipelines	% Successful	Download	
BLCA	18	100%	Open	Protected
BRCA	29	100%	Open	Protected
CESC	12	100%	Open	Protected
COADREAD	29	100%	Open	Protected
GBM				Protected
HNSC				Protected
KIRC				Protected
LAML				Protected
LGG				Protected
LIHC				Protected
OV				Protected
PRAD				Protected
SKCM	12	100%	Open	Protected
THCA	15	100%	Open	Protected
UCEC	29	100%	Open	Protected
KIDP	22	96%	Open	Protected
			Open	Protected
			Open	Protected
			Open	Protected
			Open	Protected
			Open	Protected

Analysis Dashboard

Tumor	BCR	Clinical	CN	LowP	Methylation	mRNA	mRNAseq	miR	miRseq	RPPA	MAF
BLCA	126	68	88	0	78	0	85	0	88	0	28
BRCA	914	865	855	0	858	529	805	0	809	408	507
CESC	122	32	102	0	0	0	0	0	42	0	36
COADREAD	592	591	575	104	584	224	83	0	255	399	224
DLBC	27	0	17	0	0	0	0	0	0	0	0
GBM	598	565	563	0	287	542	0	491	0	214	276
HNSC	312	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	212	0
KICH	65									0	0
KIRC	502									454	403
KIRP	135									0	0
LAML	202									0	199
LGG	181	[illegible]	[illegible]	0	[illegible]	[illegible]	[illegible]	0	[illegible]	0	0
LIHC	99	62	82	0	0	0	17	0	54	0	0
LUAD	439	292	303	0	347	32	250	0	100	237	229
LUSC	360	279	289	0	282	154	223	0	202	195	178
OV	592	580	566	0	557	574	297	570	454	412	316
PAAD	48	0	48	0	30	0	0	0	0	0	0
PRAD	174	127	152	0	153	0	53	0	81	0	83
SARC	29	0	20	0	0	0	0	0	0	0	0
SKCM	273	138	252	101	240	0	212	0	240	0	0
STAD	226	159	161	0	133	0	57	0	151	0	133
THCA	353	193	268	81	230	0	158	0	187	0	0
UCEC	512	451	444	106	451	54	266	0	377	200	248
PANCANCER	6881	5671	5810	478	5471	2224	3527	1061	4109	2731	2860

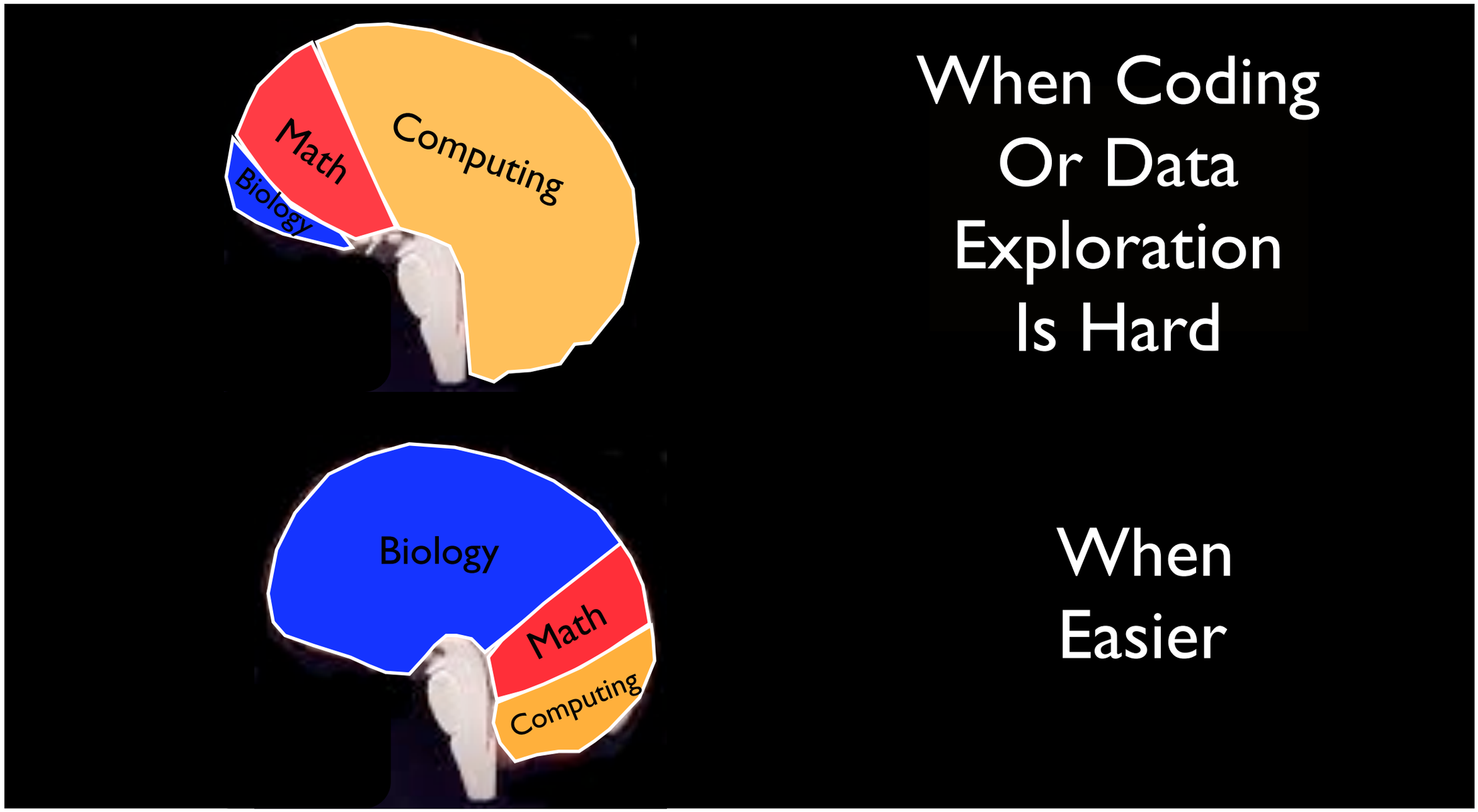
SampleCounts

The Broad is a leading creator, aggregator, and distiller of biological data in the world

But it is a balkanized landscape of projects & funding sources that feel disconnected

Technological, methodological, or interpretational links seem made by good fortune more than intention

This need not be stupidly hard : your brain ...



Civilization advances by extending the number of important operations which we can perform without thought.

A. North Whitehead

Digital Data are now fundamental to the practice of Biological Science

The Genomics-Era is yielding explosive growth in data production & interpretation

Seeing that that all the pieces fit together sensibly

Is thus fundamental to the continued success & growth of Broad projects ...

Yet we presently have no one with a dedicated, institution-wide eye on that ball

From: Michael Noble
Subject: **Data Officer at Broad?**
Date: October 16, 2012 1:06:33 PM EDT
To: Martin Leach

Hi Martin,

Gaddy and I have been talking lately about roles, and in passing I mentioned to him these recent articles in Harvard Business Review

http://hbr.org/2012/10/big-data-the-management-revolution/are/1
http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/are/1

which crystallize my growing belief that the Broad needs a data officer, like 2 years ago. :) Is there anything in the works along these lines?

Just Wondering,
Mike

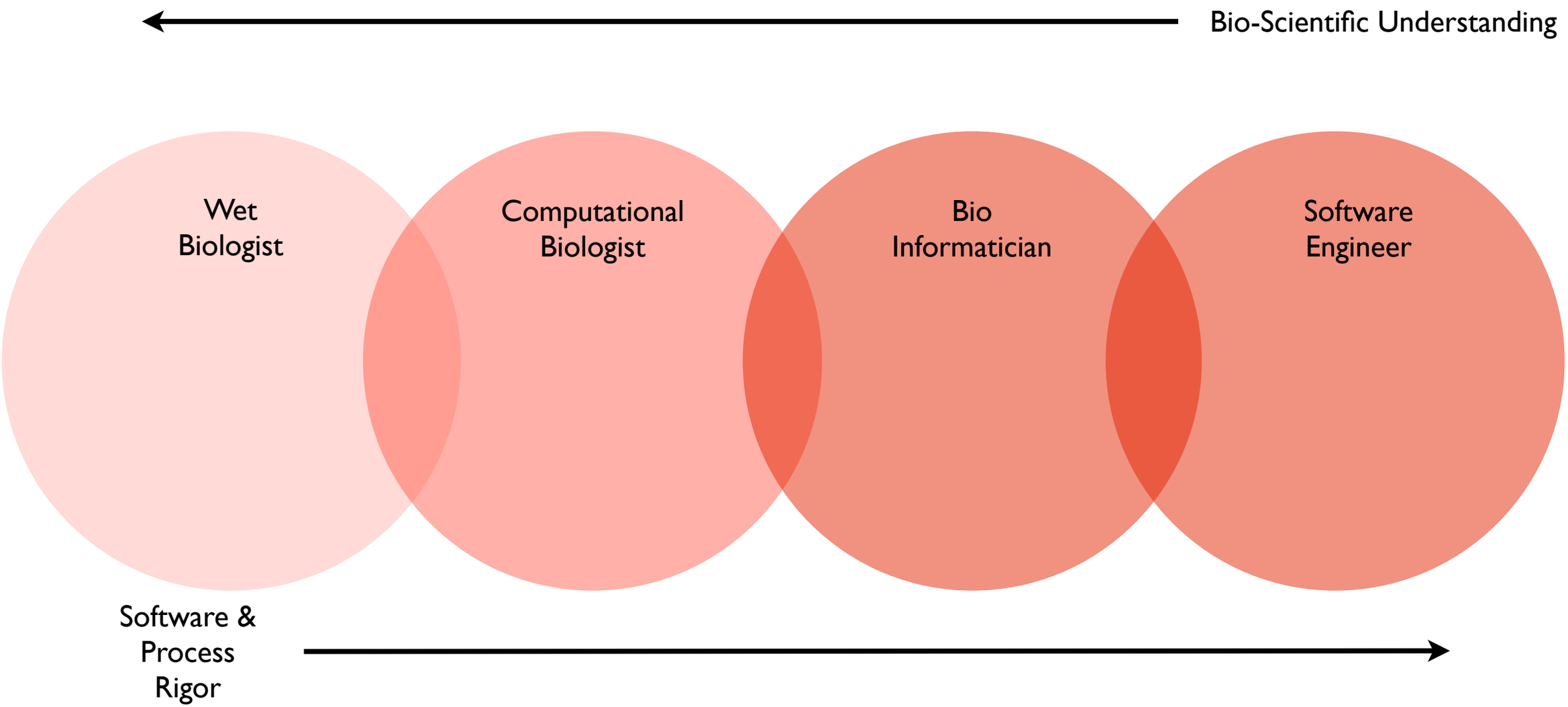
Been working this angle for a while

(2+ years turning TCGA data stream into accessible, transparent results)

Harvard Business Review kinda agrees

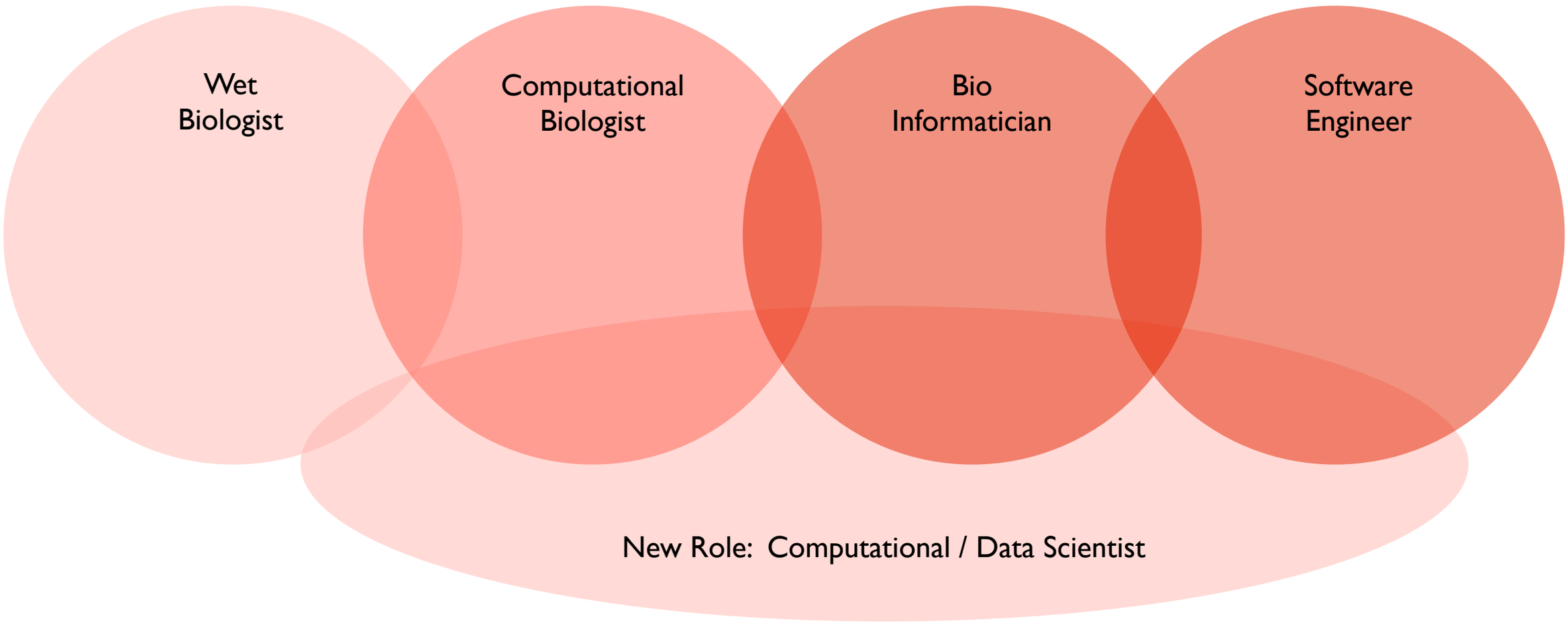
More Related Thoughts ...

Science & Engineering Roles at The Broad Institute



Each group tends to talk within itself, or adjacent boundaries
But little far-reaching cross-talk across entire spectrum
This is to our detriment
As biology continues Genomics-Era transformation from largely qualitative to rigorously quantitative

∴ We need a full-time, techno-savvy, big-picture leader paying attention to how these pieces fit together
Who prevents dogma anywhere along spectrum from impeding progress



Bring much-needed CS methodological rigor to Comp Bios & BInfs
And a credible emphasis on scientific results & timing to SWEs
Never let dogma anywhere along spectrum impede progress

A Tale of Two Coders

Software Engineer	Researcher
Careful, deliberate design Towards production deployment Must be fastidious	Exploratory, open-ended analysis Towards publication Can be messy

Overlapping, But Not Identical, Aims

Profile of A Computational / Data Scientist

Computer-scientist by training / inclination
Generalist tendencies more than specialist
Publication-caliber computational research, engineering, design credentials
Can speak & write prose as well as can code

Not quite SWE, nor computational biologist
Nor pure CS researcher, nor a production engineer
But capable of talking to, and being respected by, all of them

Without (sometimes) provincial / mechanistic rigidity of SWEs
Or (sometimes) messy/just-for-publication-research coding

Bring much-needed CS methodological rigor to Comp Bios & BInfs

Big picture orientation
Cares about scientific result as deeply as biologist does
Result is better end-to-end synergy of all wings of Broad research