

Information Visualization Meets Biology: Models and Methods for Collaboration

Tamara Munzner

Department of Computer Science
University of British Columbia

*Keynote, VIZBI 2017 ++ (at Monash Caulfield)
June 21 2017, Caulfield Australia*

www.cs.ubc.ca/~tmm/talks.html#caulfield17

[@tamaramunzner](https://twitter.com/tamaramunzner)

Visualization (vis) defined & motivated

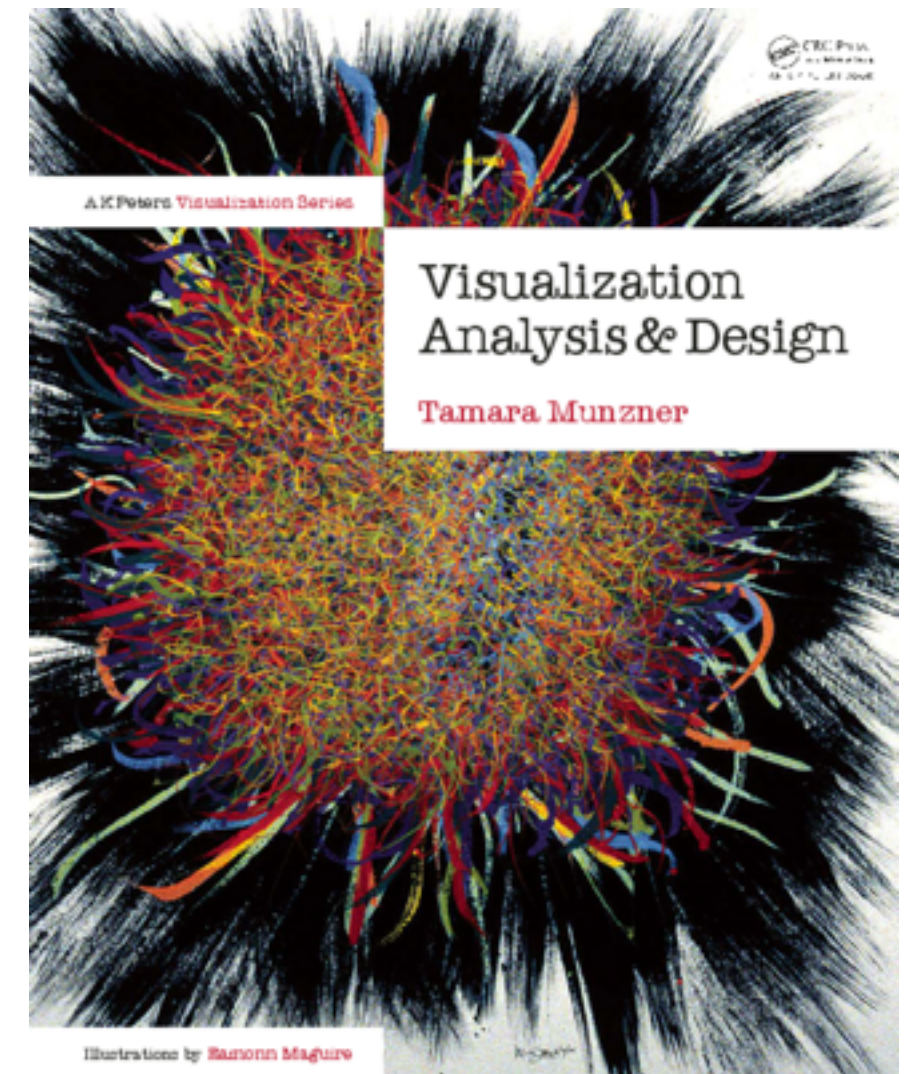
Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

- human in the loop needs the details
 - doesn't know exactly what questions to ask in advance
 - longterm exploratory analysis
 - presentation of known results
 - stepping stone towards automation: refining, trustbuilding
- external representation: perception vs cognition
- intended task, measurable definitions of effectiveness

more at:

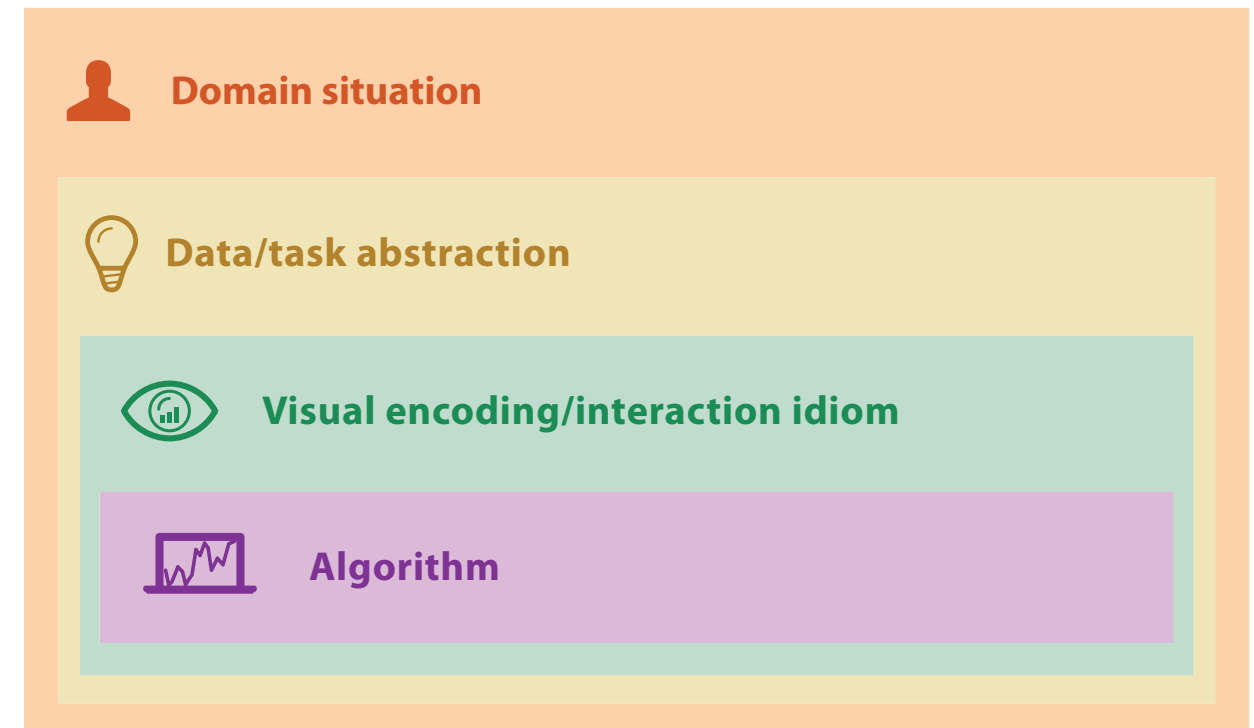
Visualization Analysis and Design, Chapter 1.
Munzner. AK Peters Visualization Series, CRC Press, 2014.



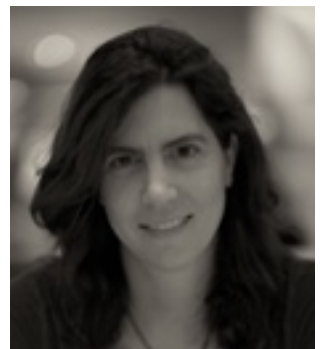
A Nested Model

for Visualization Design and Validation

<http://www.cs.ubc.ca/labs/imager/tr/2009/NestedModel>



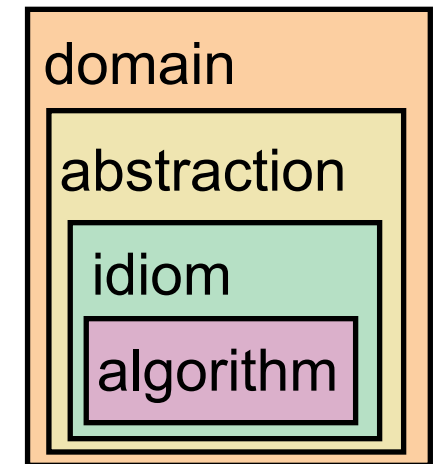
Tamara Munzner
@tamaramunzner



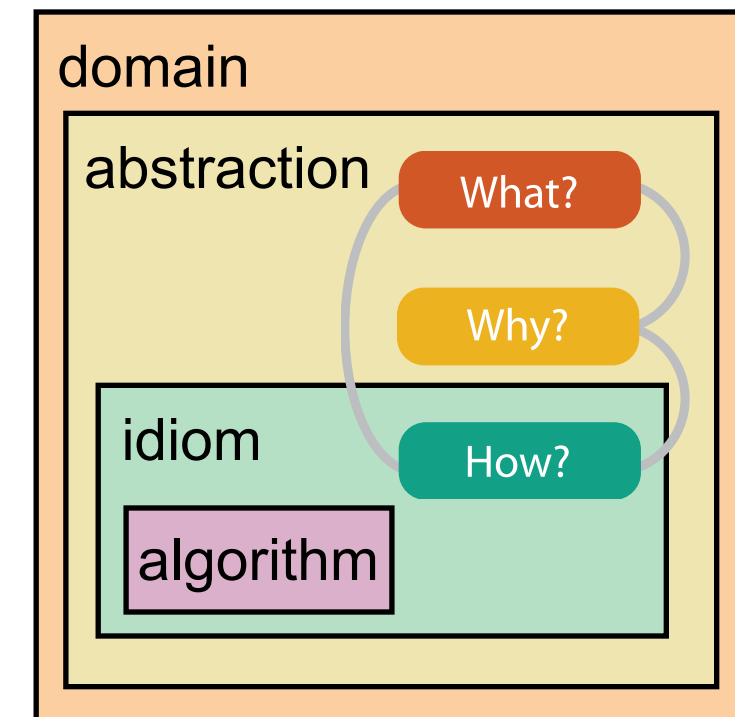
A Nested Model for Visualization Design and Validation.
Munzner. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 09)*, 15(6):921-928, 2009.

Vis analysis framework: Four levels, three questions

- *domain* situation
 - who are the target users? what are their needs & concerns?
- *abstraction*
 - translate from specifics of domain to vocabulary of vis
 - **what** is shown? **data abstraction**
 - often don't just draw what you're given: transform to new form
 - **why** is the user looking at it? **task abstraction**
- *idiom*
 - **how** is it shown?
 - **visual encoding idiom**: how to draw
 - **interaction idiom**: how to manipulate
- *algorithm*
 - efficient computation



[A Nested Model of Visualization Design and Validation.
Munzner. *IEEE TVCG* 15(6):921-928, 2009
(*Proc. InfoVis 2009*).]



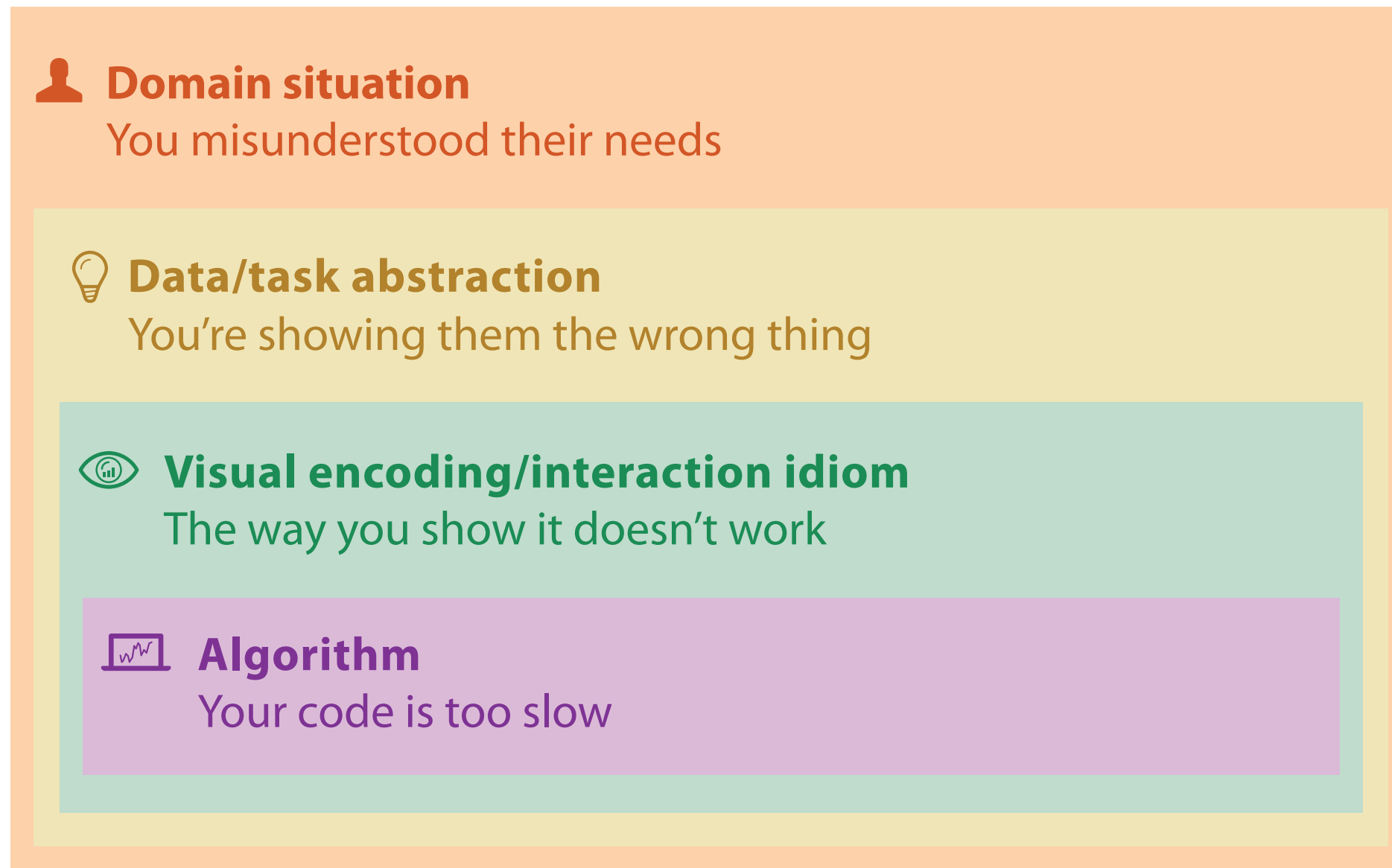
[A Multi-Level Typology of Abstract Visualization Tasks
Brehmer and Munzner. *IEEE TVCG* 19(12):2376-2385, 2013 (*Proc. InfoVis 2013*).]

more at:

Visualization Analysis and Design, Ch 2/3/4. Munzner, CRC Press, 2014.

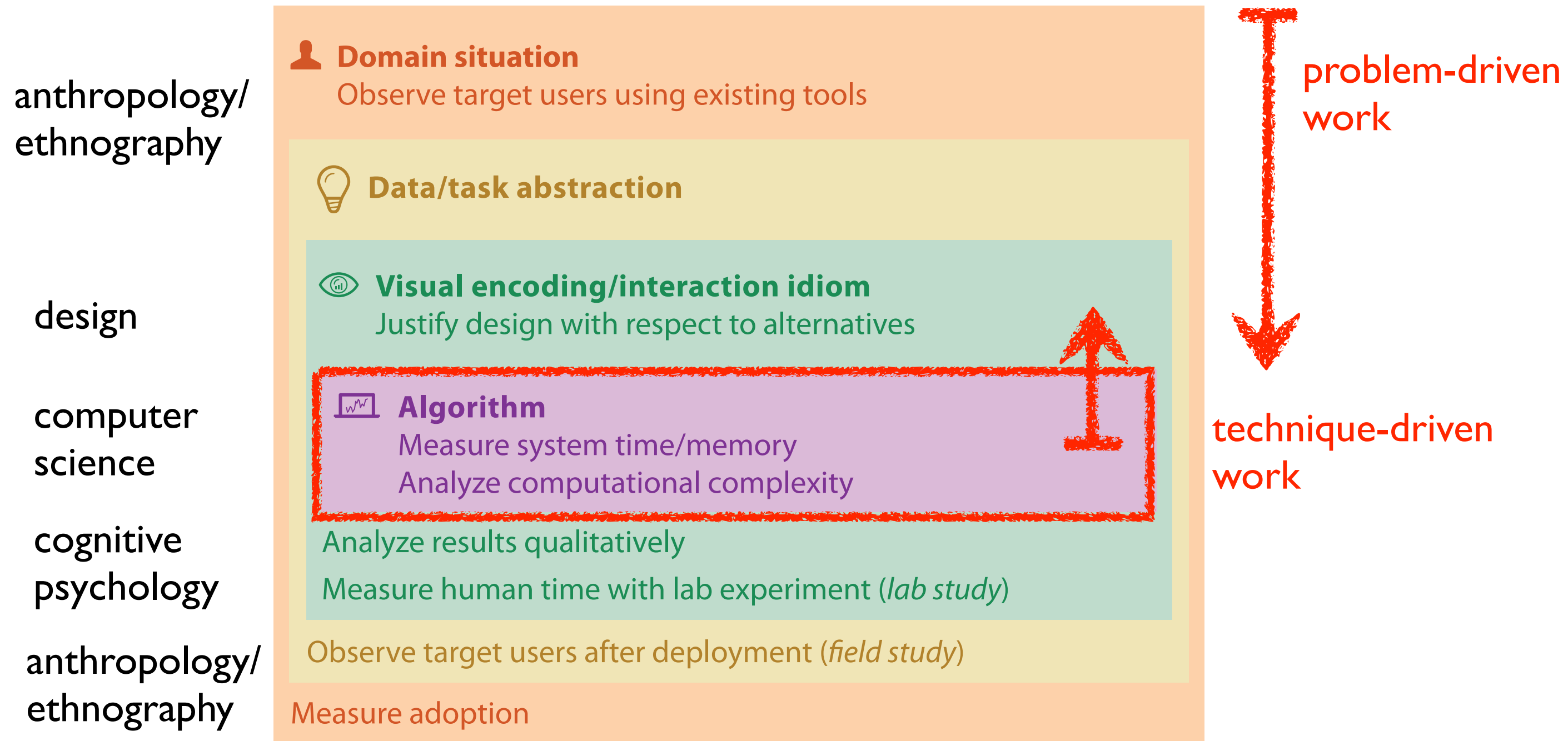
Why is validation difficult?

- different ways to get it wrong at each level

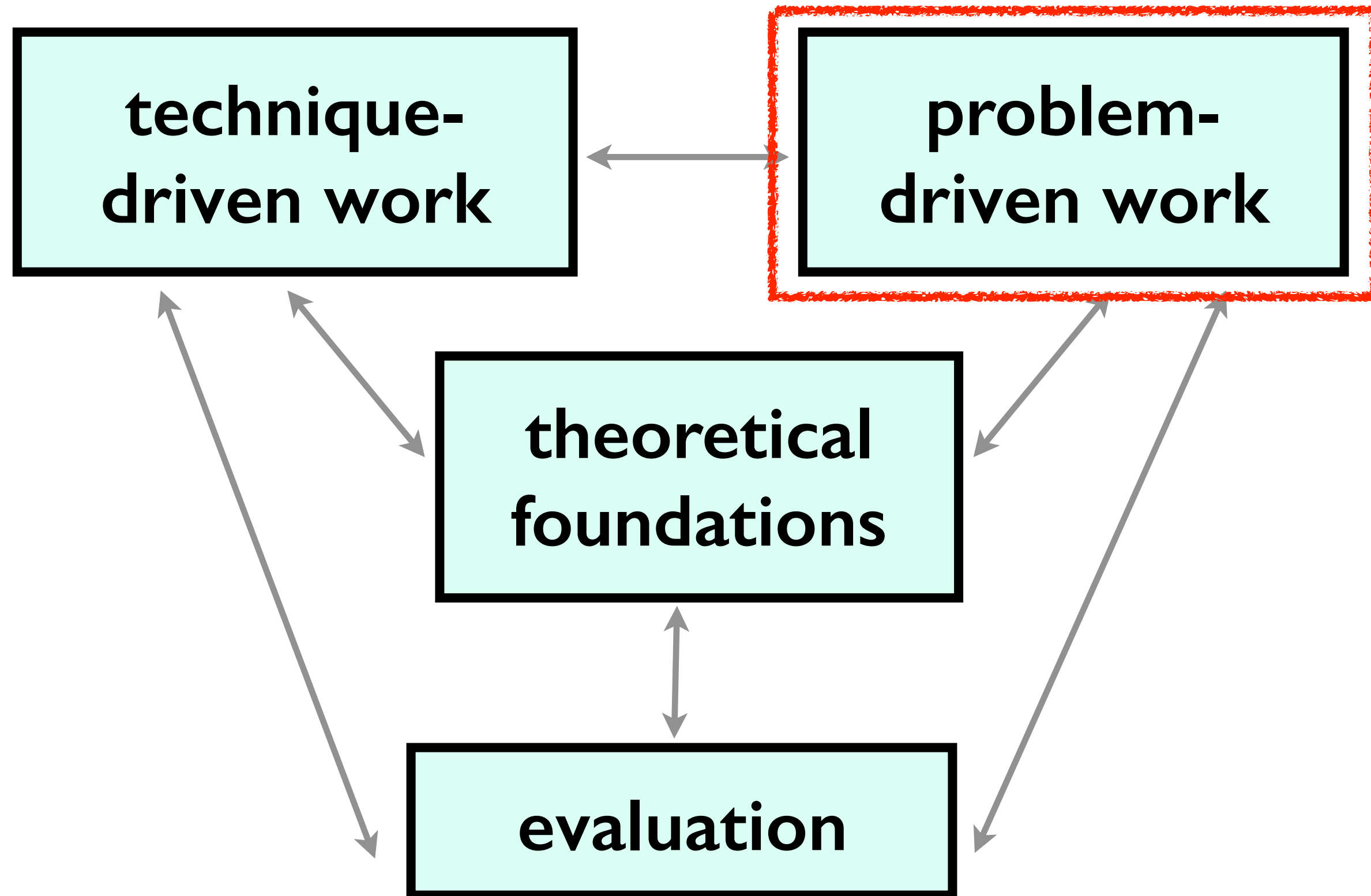


Validation solution: use methods from appropriate fields at each level

- avoid mismatches!



Angles of attack: My own research agenda

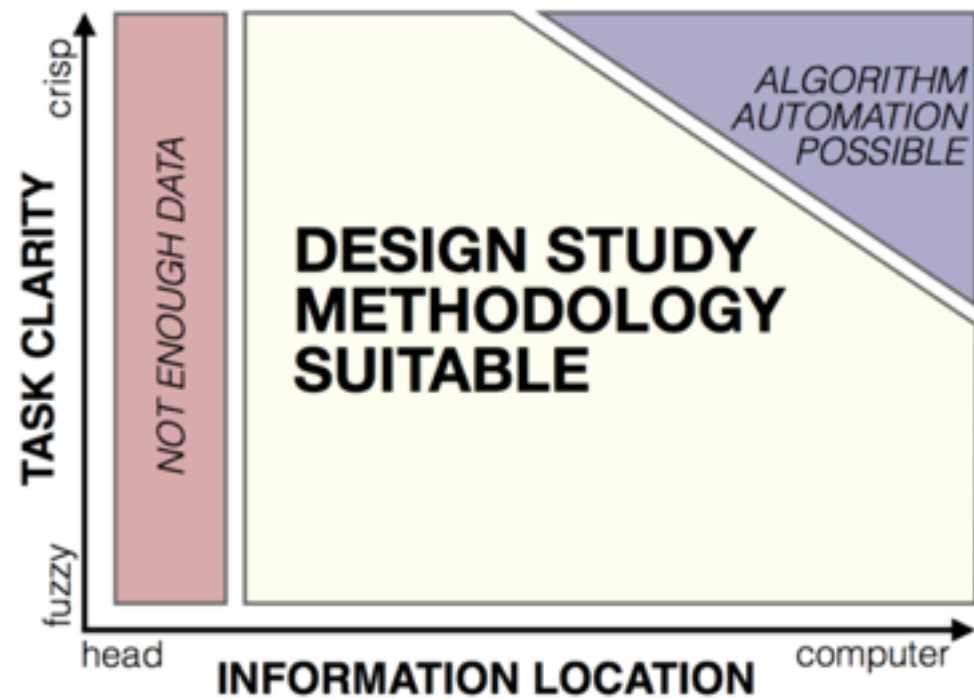


Vis meets bio

- biology encompasses many rich application domain for vis collaboration
 - challenging multi-level problems that won't be automated away any time soon
 - complex tasks, complex datasets
 - often existing infrastructure of computational workflows
 - many points where human-in-the-loop decision-making could bear fruit
- landscape of possible tools
 - axis from eureka to speedup
 - sexy use case: eureka moment
 - enable what was impossible before: vis tools for new insights & discoveries
 - workhorse use case: workflow speedup
 - vis tools to accelerate what you're already doing
 - sometimes enables the previously infeasible
 - axis from targeted to address specific pain points, to general purpose for broad use

Collaboration incentives: Bidirectional

- what's in it for bio?
 - bio win: access to more suitable tools, can do better/faster/cheaper science
 - time spent could pay off with earlier access and/or more customized tools
- what's in it for vis?
 - vis win: access to better understanding of your driving problems
 - crucial element in building effective tools to help
 - opportunities to observe how you use them
 - if they're good enough, vis win: research success stories
 - leads us to develop guidelines on how to build better tools in general
 - vis win: research progress in visualization
 - [The Computer Scientist as Toolsmith II, Fred Brooks, CACM 30(3):61-68 1996]



Michael Sedlmair



Miriah Meyer

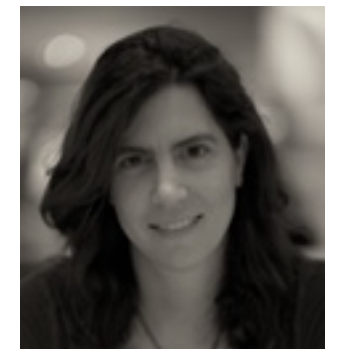


Design Study Methodology

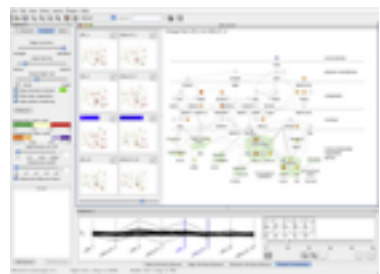
Reflections from the Trenches and from the Stacks

<http://www.cs.ubc.ca/labs/imager/tr/2012/dsm/>

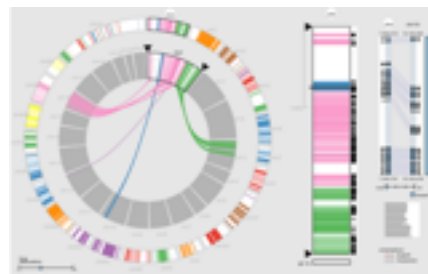
Tamara Munzner
@tamaramunzner



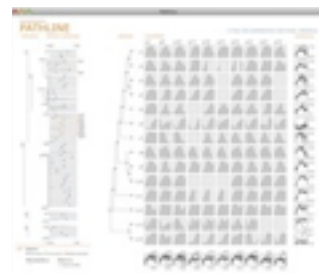
Design Studies: Lessons learned after 21 of them



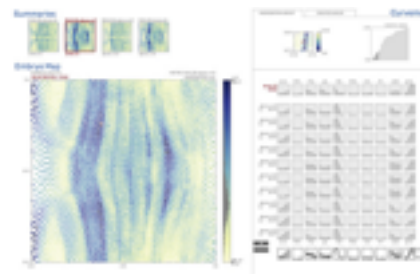
Cerebral
genomics



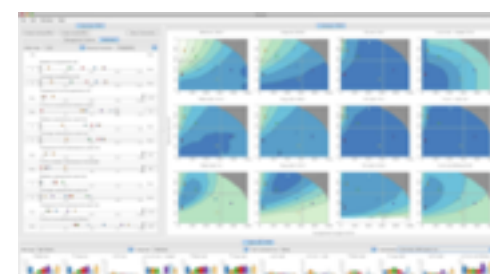
MizBee
genomics



Pathline
genomics



MulteeSum
genomics



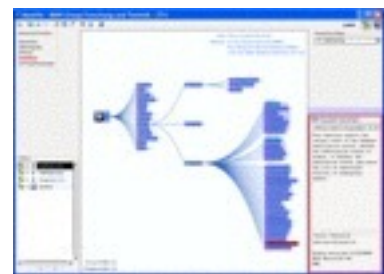
Vismon
fisheries management



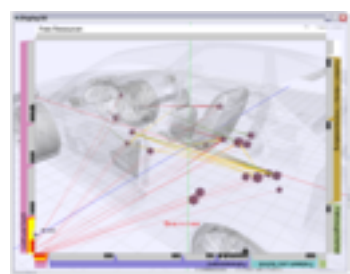
QuestVis
sustainability



WiKeVis
in-car networks



MostVis
in-car networks



Car-X-Ray
in-car networks



ProgSpy2010
in-car networks



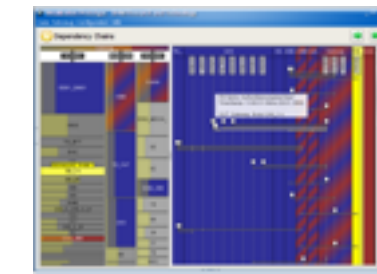
RelEx
in-car networks



Cardiogram
in-car networks



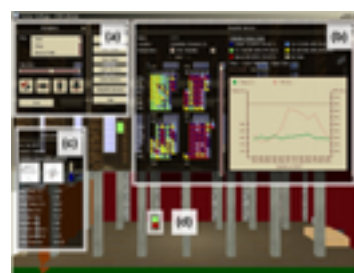
AutobahnVis
in-car networks



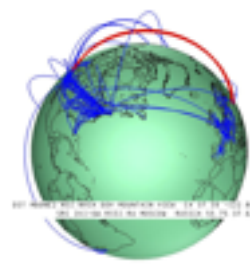
VisTra
in-car networks



Constellation
linguistics



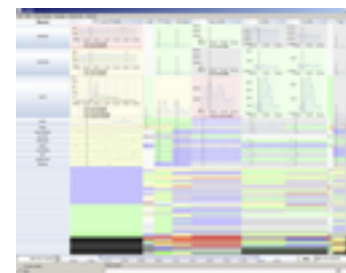
LibVis
cultural heritage



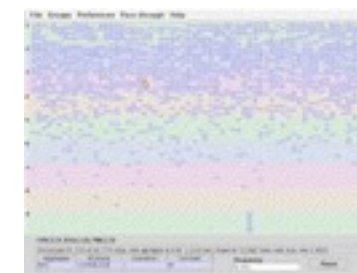
Caidants
multicast



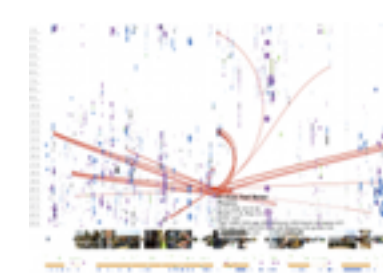
SessionViewer
web log analysis



LiveRAC
server hosting



PowerSetViewer
data mining

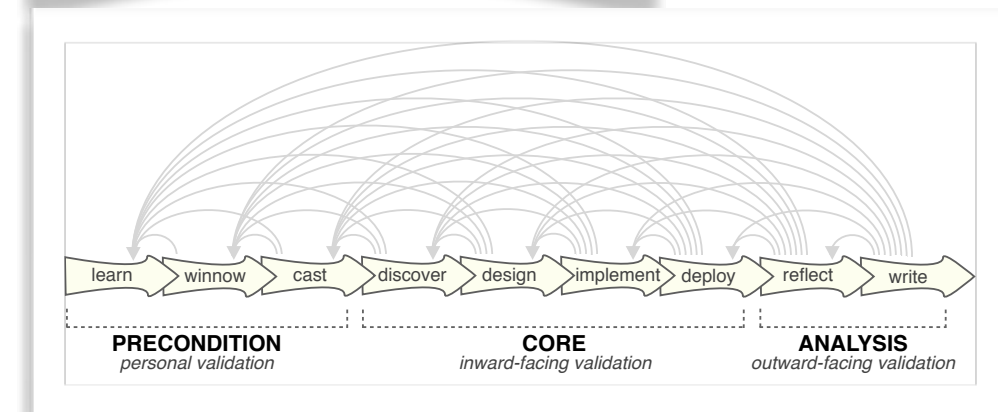
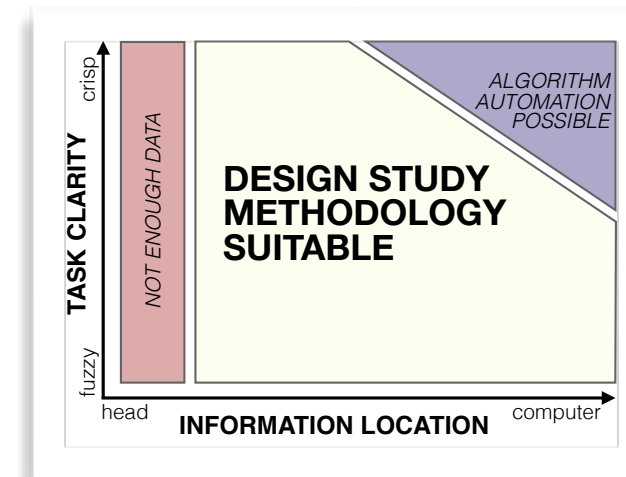


LastHistory
music listening

- commonality of representations cross-cuts domains!

Methodology for problem-driven work

- definitions
- 9-stage framework
- 32 pitfalls & how to avoid them
- comparison to related methodologies



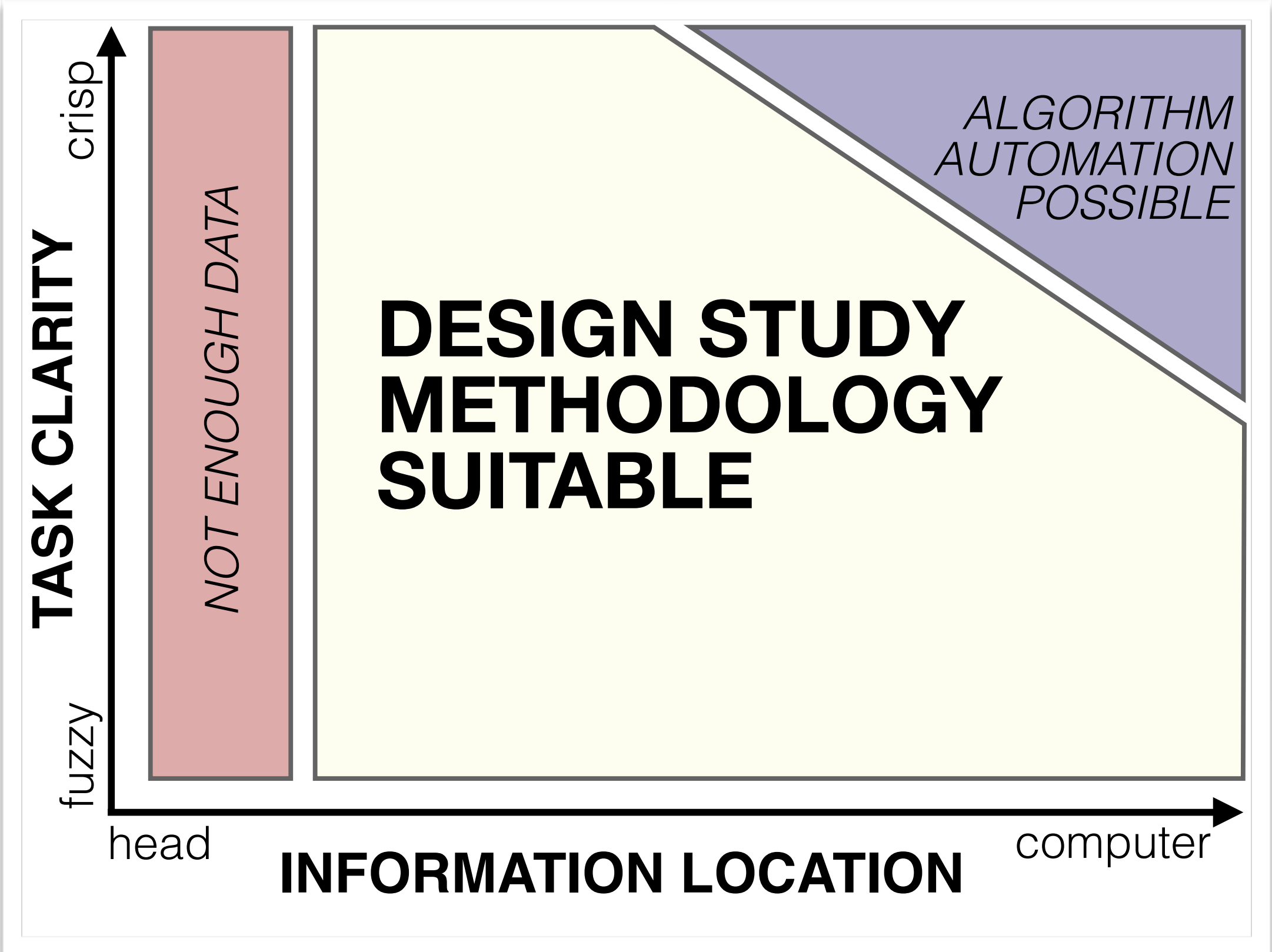
PF-1	premature advance: jumping forward over stages	general
PF-2	premature start: insufficient knowledge of vis literature	learn
PF-3	premature commitment: collaboration with wrong people	winnow
PF-4	no real data available (yet)	winnow
PF-5	insufficient time available from potential collaborators	winnow
PF-6	no need for visualization: problem can be automated	winnow
PF-7	researcher expertise does not match domain problem	winnow
PF-8	no need for research: engineering vs. research project	winnow
PF-9	no need for change: existing tools are good enough	winnow



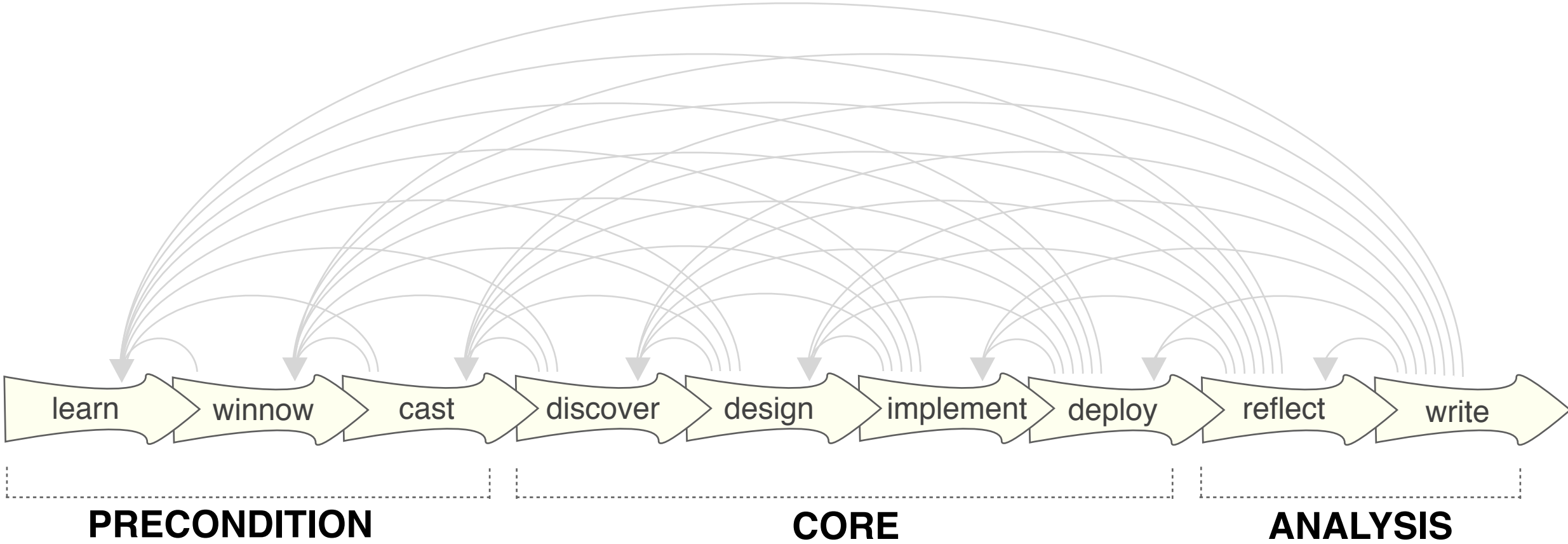
Design study methodology: definitions

- design studies: problem-driven work
 - in collaboration with target users
 - real data, real tasks
 - intensive requirements analysis
 - iterative refinement
 - rapid prototyping
 - deploy tools/systems to target users
 - typical evaluation: field studies
 - case studies provide evidence of utility for target users
 - replicate known results quickly/easily: show workflow speedup
 - examples of new results found using tool

Design study methodology: definitions

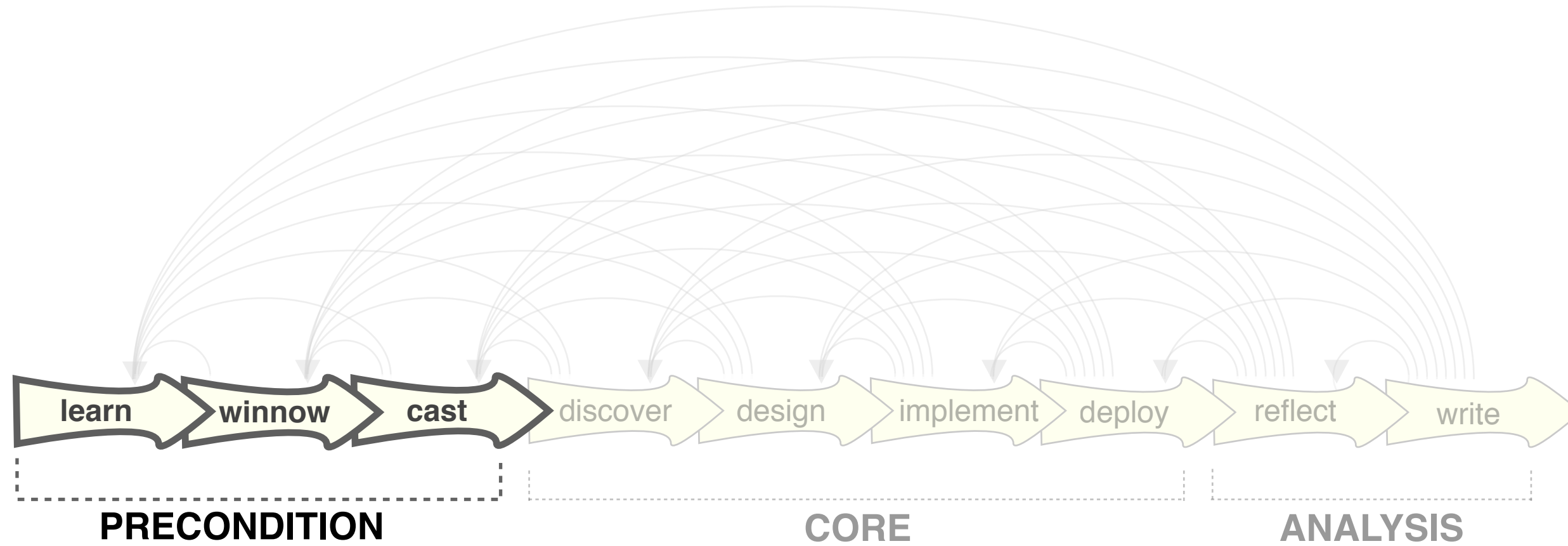


9 stage framework



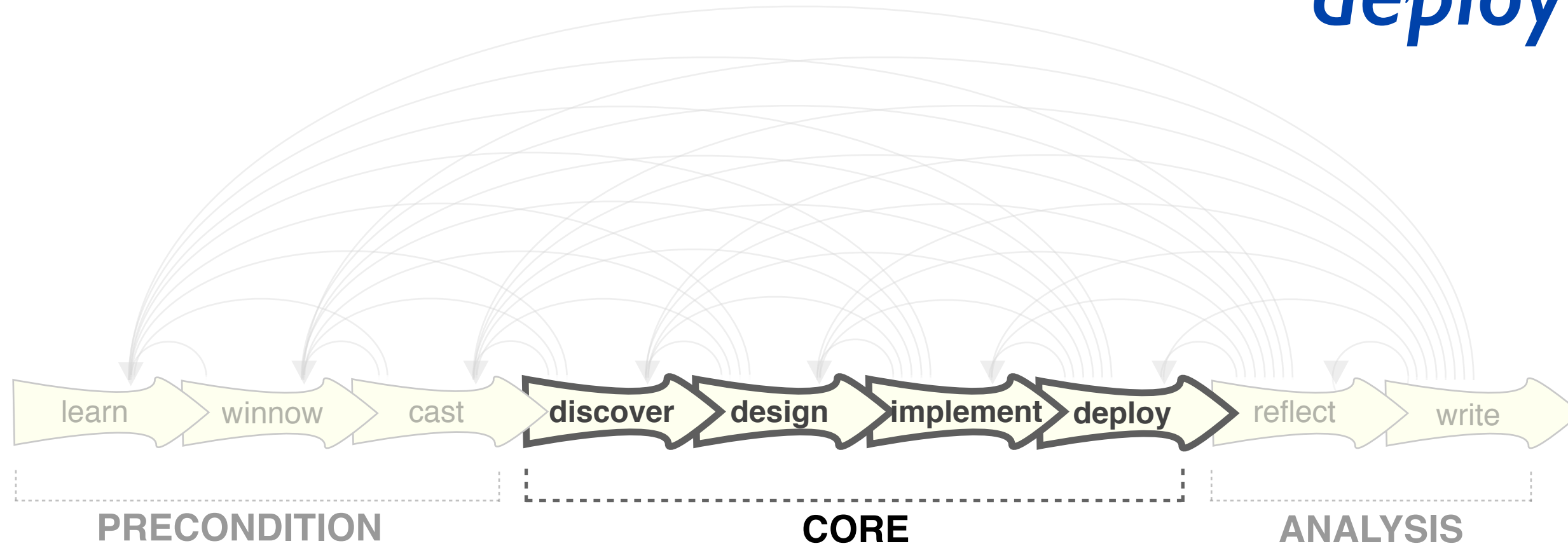
9-stage framework

learn
winnow
cast



9-stage framework

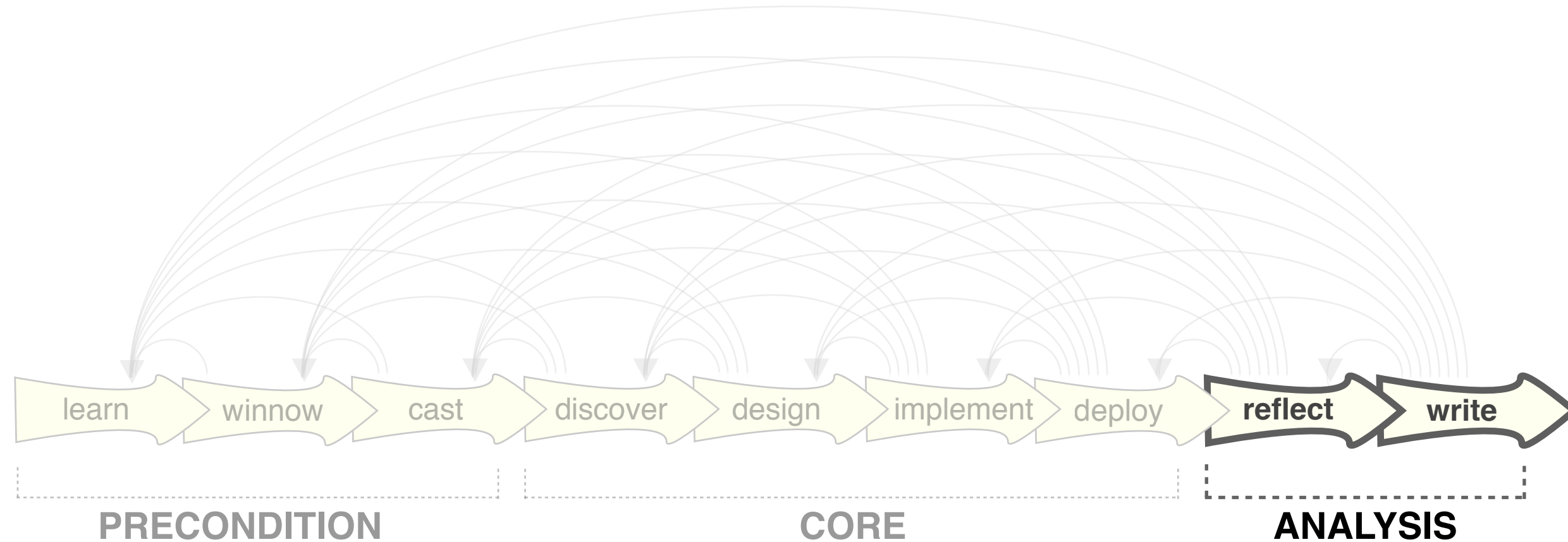
discover
design
implement
deploy



9-stage framework

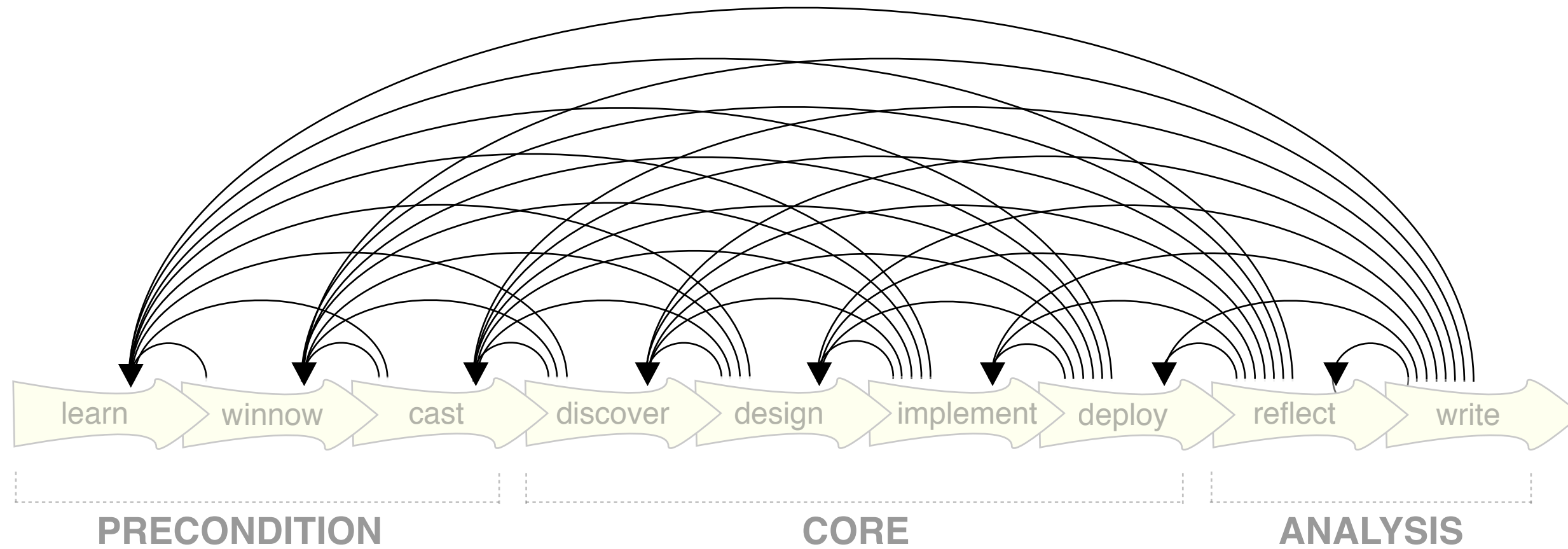
- guidelines: confirm, refine, reject, propose

reflect
write



9-stage framework

iterative



Design study methodology: 32 pitfalls

- and how to avoid them

PF-1	premature advance: jumping forward over stages	general
PF-2	premature start: insufficient knowledge of vis literature	learn
PF-3	premature commitment: collaboration with wrong people	winnow
PF-4	no real data available (yet)	winnow
PF-5	insufficient time available from potential collaborators	winnow
PF-6	no need for visualization: problem can be automated	winnow
PF-7	researcher expertise does not match domain problem	winnow
PF-8	no need for research: engineering vs. research project	winnow
PF-9	no need for change: existing tools are good enough	winnow

PITFALL

**PREMATURE
COLLABORATION
COMMITMENT**

I'm a domain expert!
Wanna collaborate?

Of course!!!



COLLABORATOR



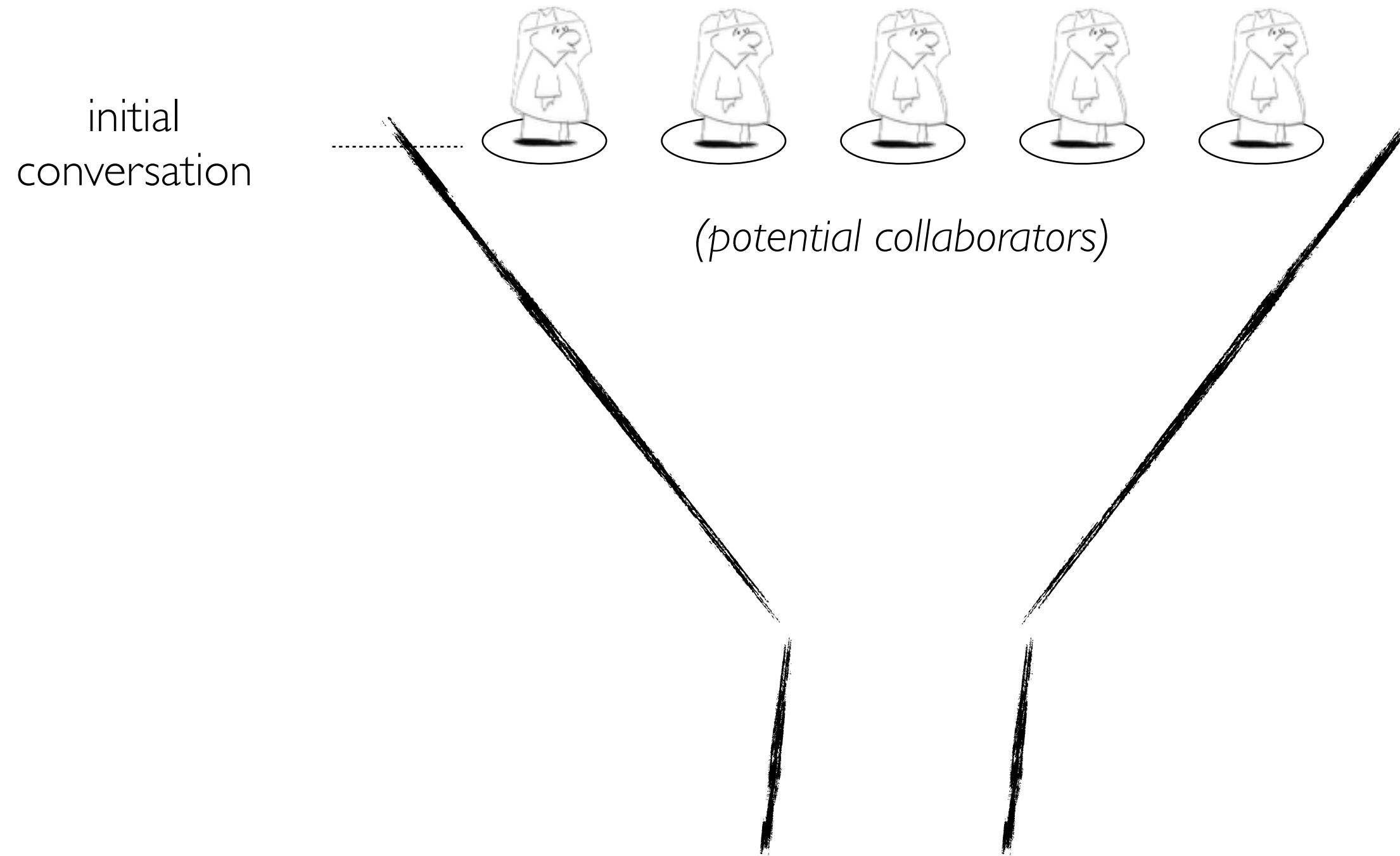
MR. VIS

METAPHOR

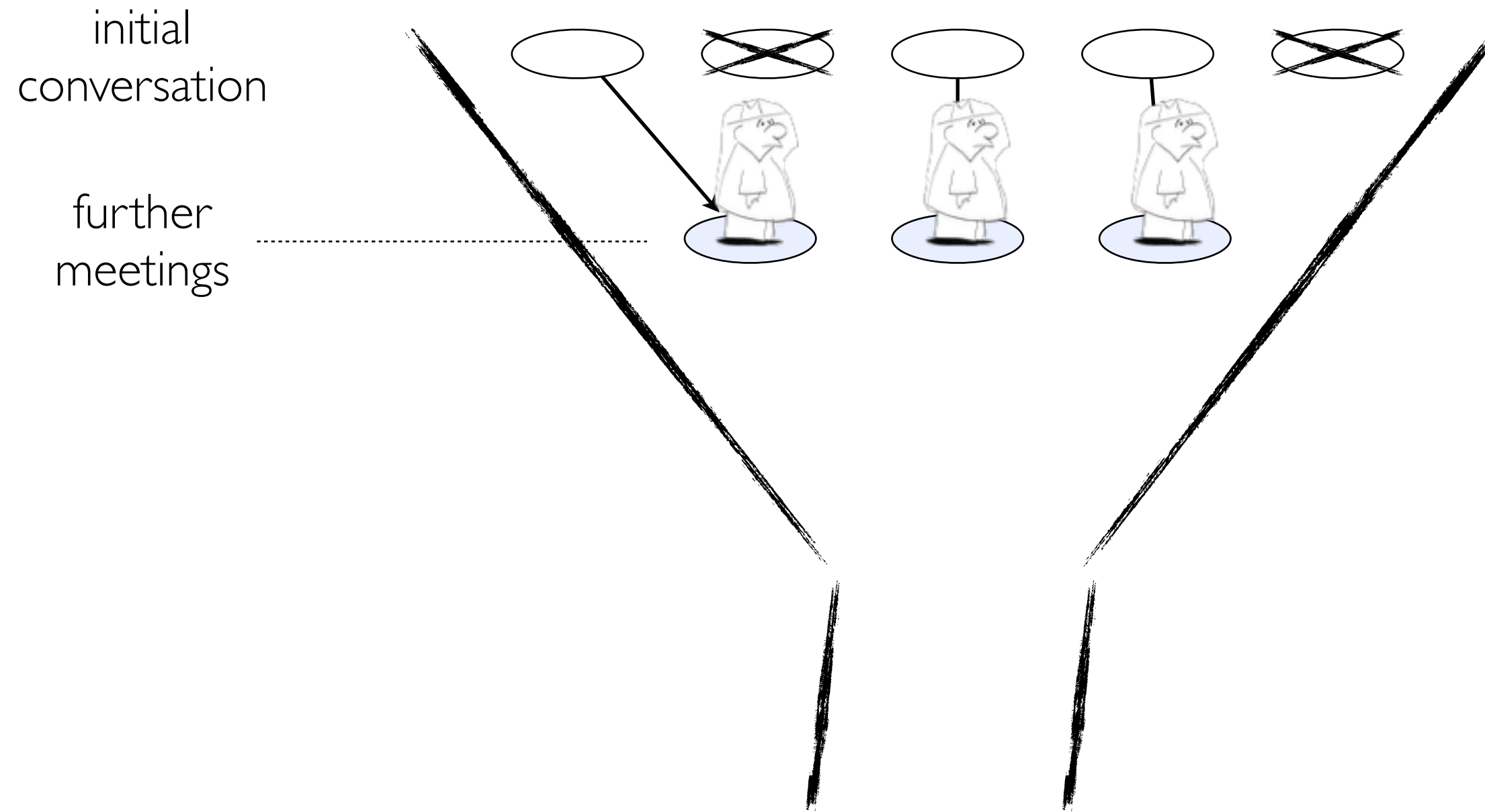
Winnowing



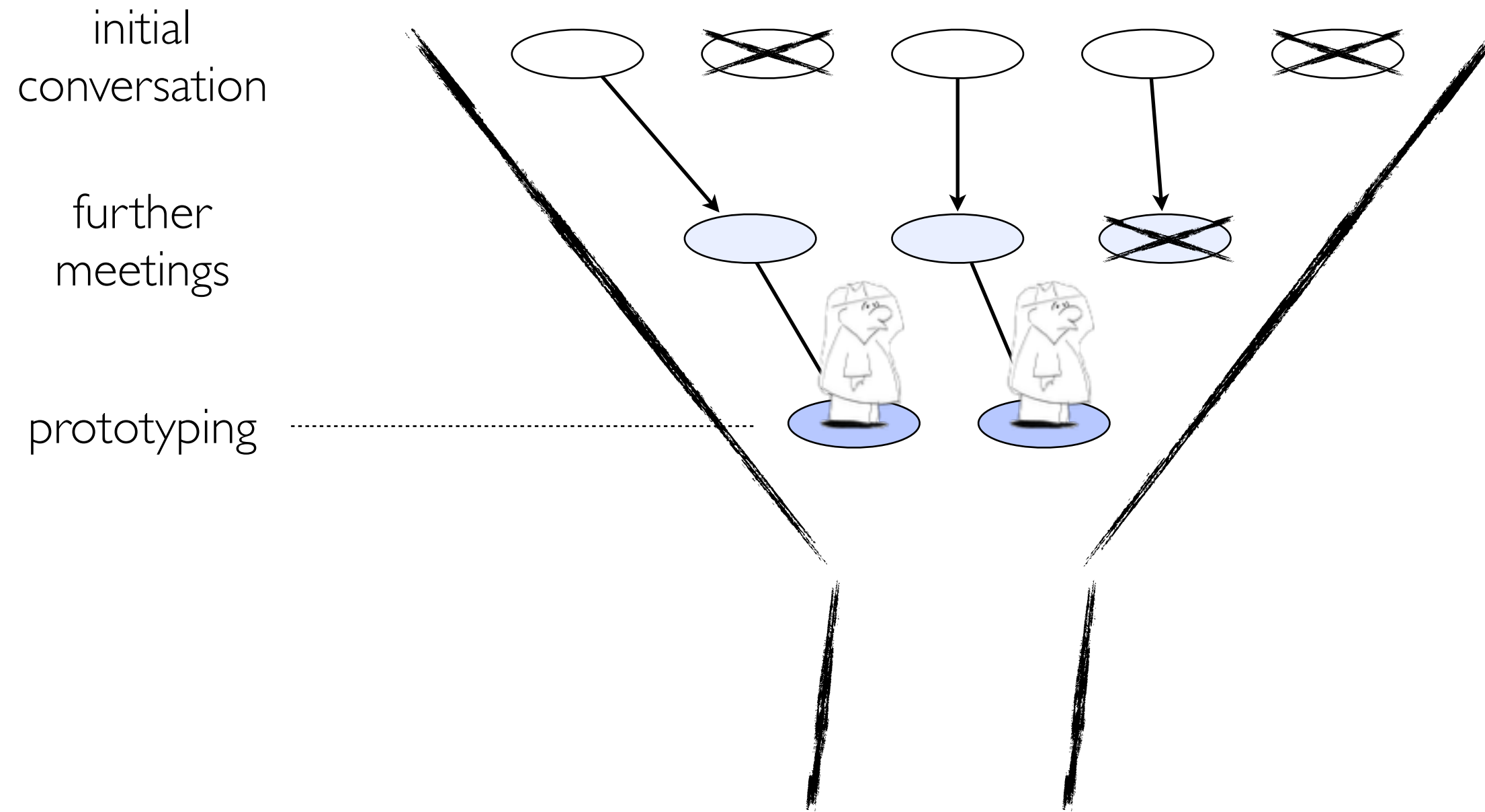
Collaborator winnowing



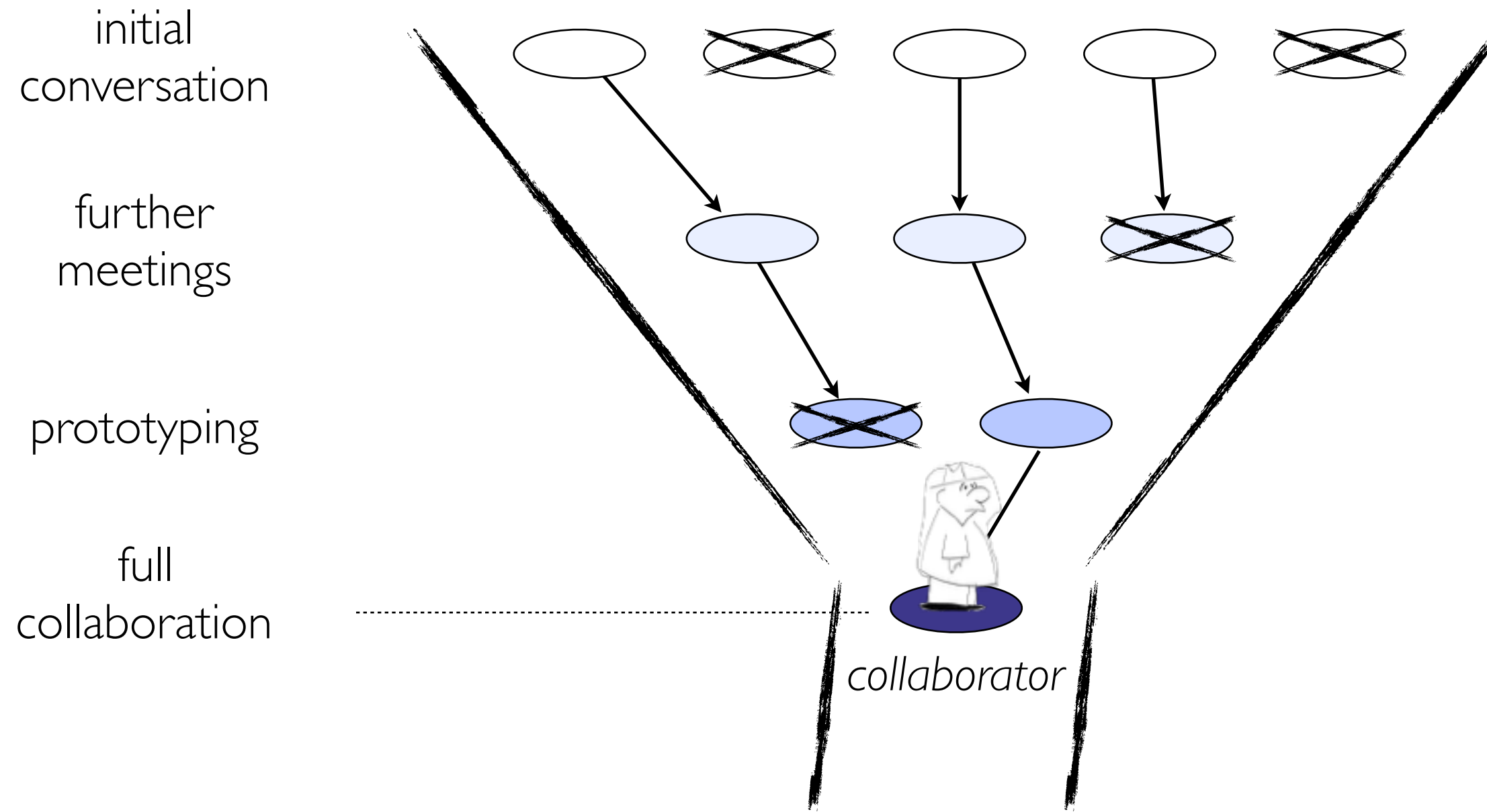
Collaborator winnowing



Collaborator winnowing



Collaborator winnowing



Collaborator winnowing



Design study methodology: 32 pitfalls

- and how to avoid them

PF-1	premature advance: jumping forward over stages	general
PF-2	premature start: insufficient knowledge of vis literature	learn
PF-3	premature commitment: collaboration with wrong people	winnow
PF-4	no real data available (yet)	winnow
PF-5	insufficient time available from potential collaborators	winnow
PF-6	no need for visualization: problem can be automated	winnow
PF-7	researcher expertise does not match domain problem	winnow
PF-8	no need for research: engineering vs. research project	winnow
PF-9	no need for change: existing tools are good enough	winnow

considerations



Have **data**?
Have **time**?
Have **need**?
...



Design study methodology: 32 pitfalls

PF-10	no real/important/recurring task	winnow
PF-11	no rapport with collaborators	winnow
PF-12	not identifying front line analyst and gatekeeper before start	cast
PF-13	assuming every project will have the same role distribution	cast
PF-14	mistaking fellow tool builders for real end users	cast
PF-15	ignoring practices that currently work well	discover
PF-16	expecting <i>just talking</i> or <i>fly on wall</i> to work	discover
PF-17	experts focusing on visualization design vs. domain problem	discover
PF-18	learning their problems/language: too little / too much	discover
PF-19	abstraction: too little	design
PF-20	premature design commitment: consideration space too small	design

roles

bioinformatician

biologist

Are you a
user???

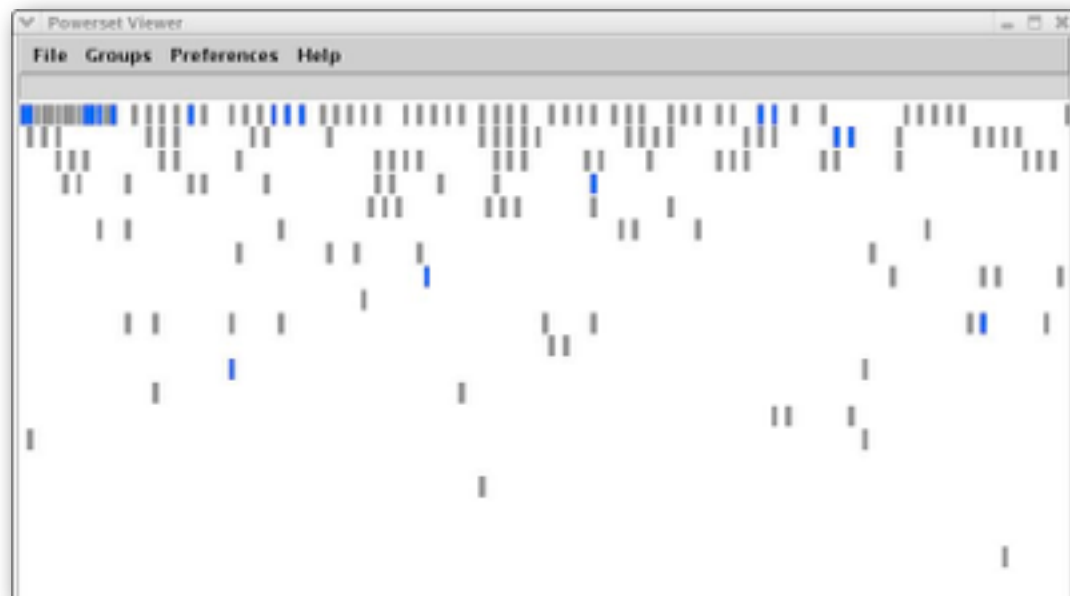
... or maybe a
**fellow tool
builder?**



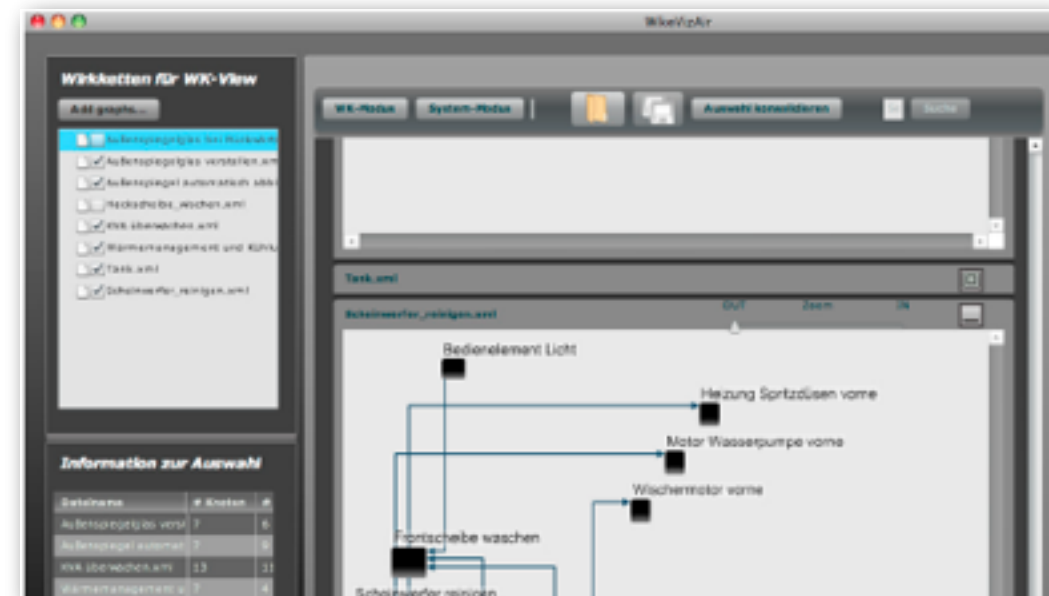
Examples from the trenches

- premature collaboration
- fellow tool builders with inaccurate assumptions about user needs
- data unavailable early so didn't diagnose problems

PowerSet Viewer
2 years / 4 researchers



WikeVis
0.5 years / 2 researchers



Design study methodology: 32 pitfalls

PF-10	no real/important/recurring task	winnow
PF-11	no rapport with collaborators	winnow
PF-12	not identifying front line analyst and gatekeeper before start	cast
PF-13	assuming every project will have the same role distribution	cast
PF-14	mistaking fellow tool builders for real end users	cast
PF-15	ignoring practices that currently work well	discover
PF-16	expecting <i>just talking</i> or <i>fly on wall</i> to work	discover
PF-17	experts focusing on visualization design vs. domain problem	discover
PF-18	learning their problems/language: too little / too much	discover
PF-19	abstraction: too little	design
PF-20	premature design commitment: consideration space too small	design

PITFALL

PREMATURE DESIGN
COMMITMENT

I want a tool with that
cool technique I saw the
other day!



COLLABORATOR

PITFALL

**PREMATURE DESIGN
COMMITMENT**

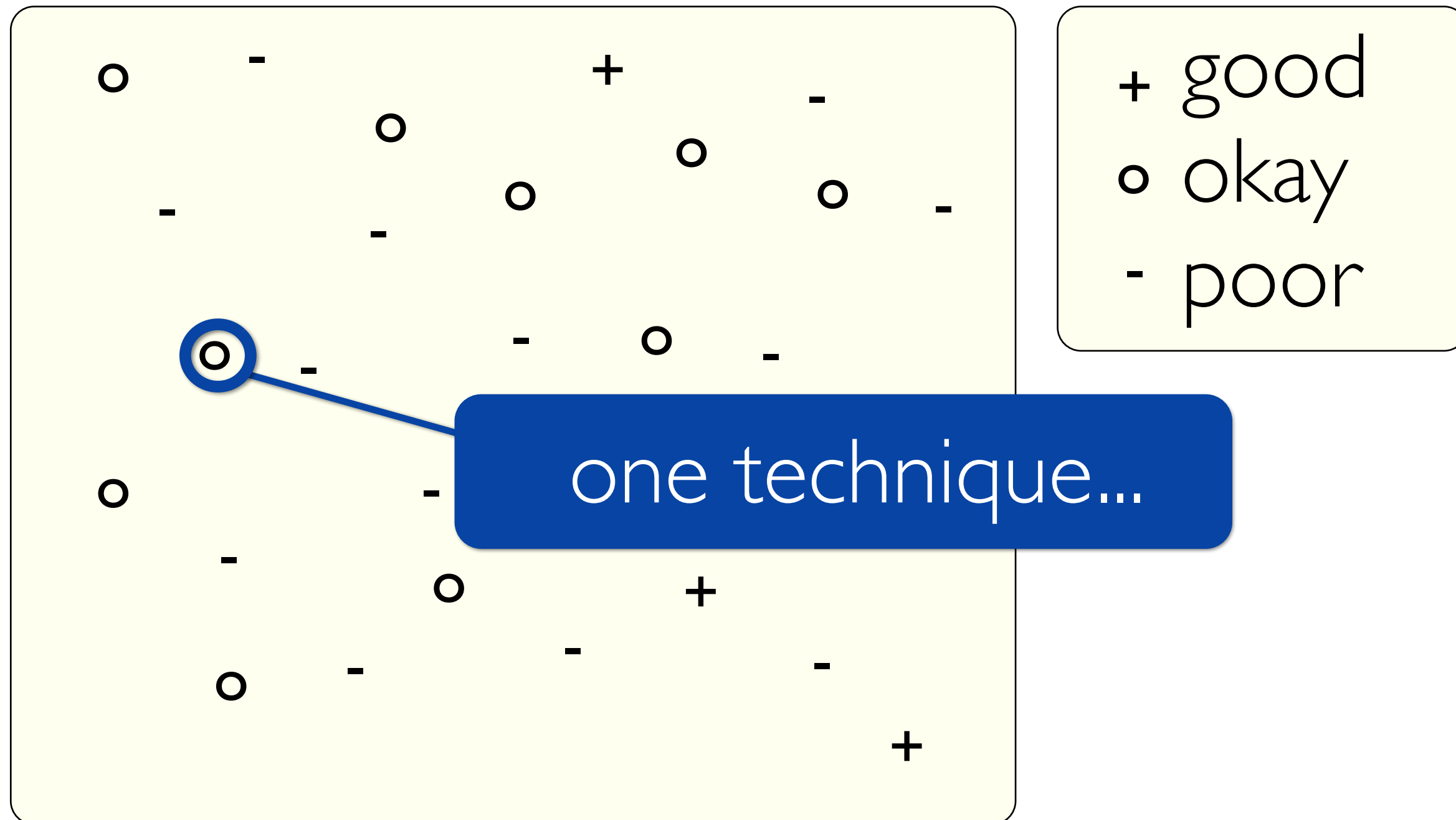
Of course they need the **cool
technique** I built last year!



MR. VIS

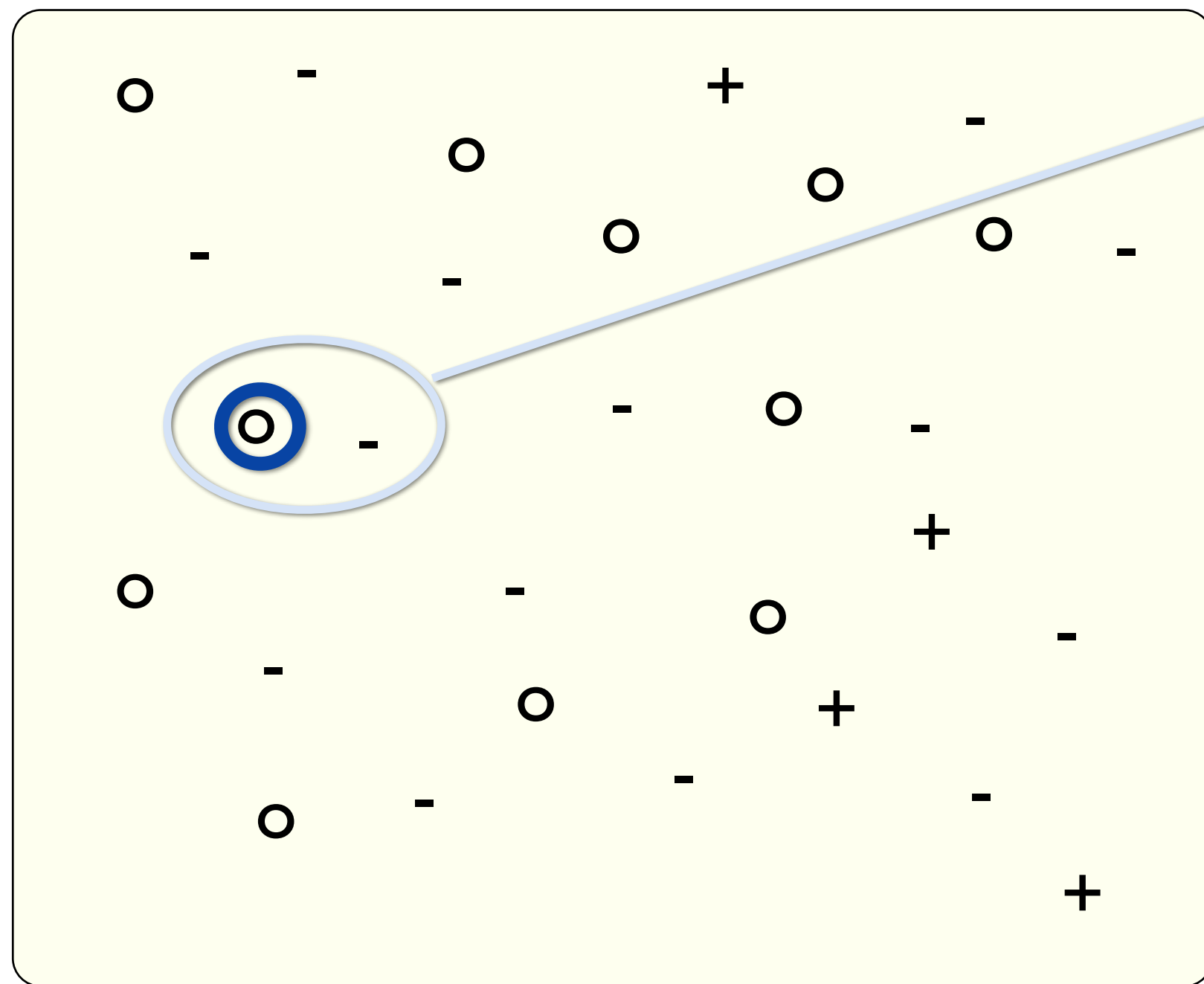
METAPHOR

Design Space



METAPHOR

Design Space



know

small
scope

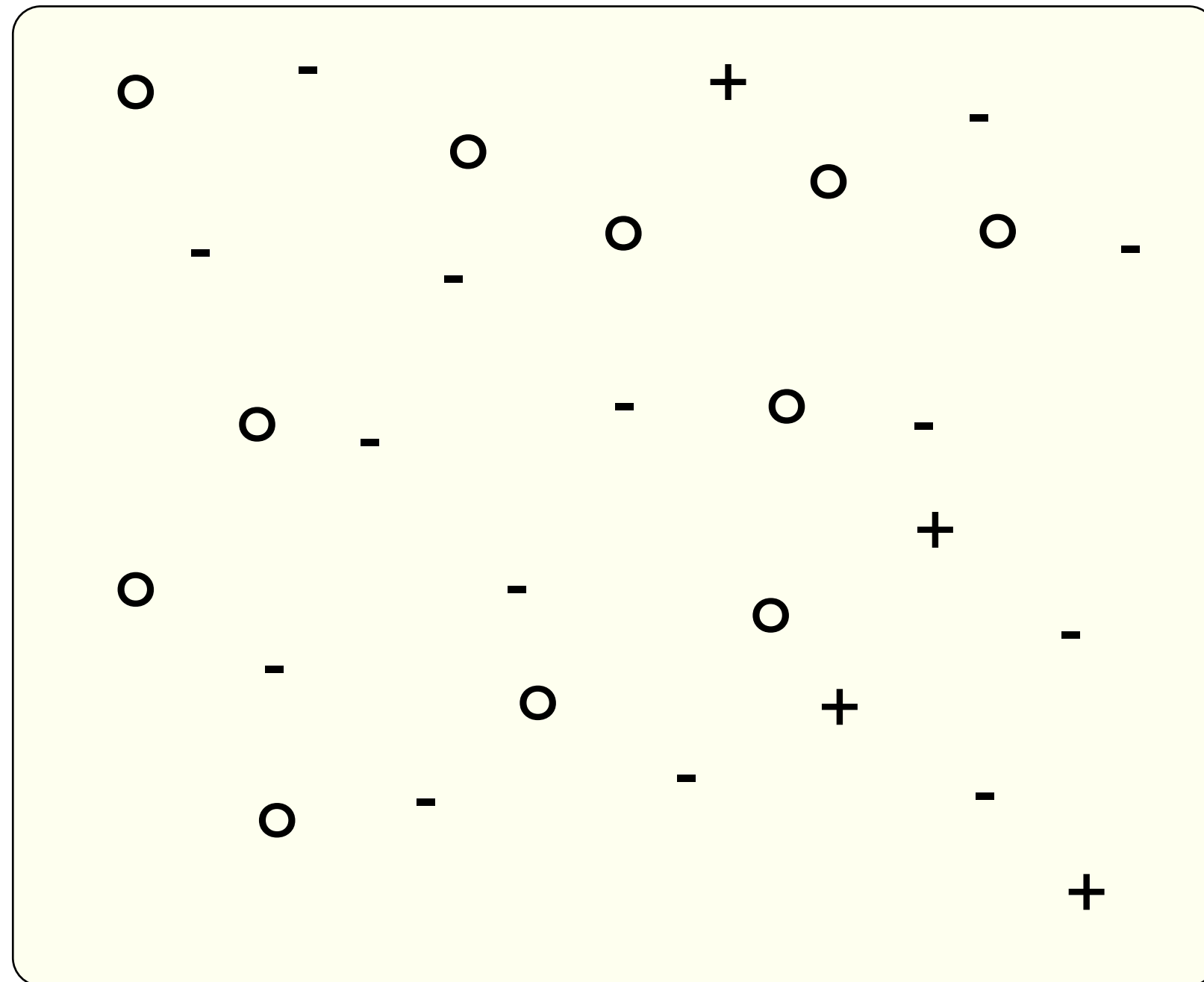
Design study methodology: 32 pitfalls

- and how to avoid them

PF-1	premature advance: jumping forward over stages	general
PF-2	premature start: insufficient knowledge of vis literature	learn
PF-3	premature commitment: collaboration with wrong people	winnow
PF-4	no real data available (yet)	winnow
PF-5	insufficient time available from potential collaborators	winnow
PF-6	no need for visualization: problem can be automated	winnow
PF-7	researcher expertise does not match domain problem	winnow
PF-8	no need for research: engineering vs. research project	winnow
PF-9	no need for change: existing tools are good enough	winnow

METAPHOR

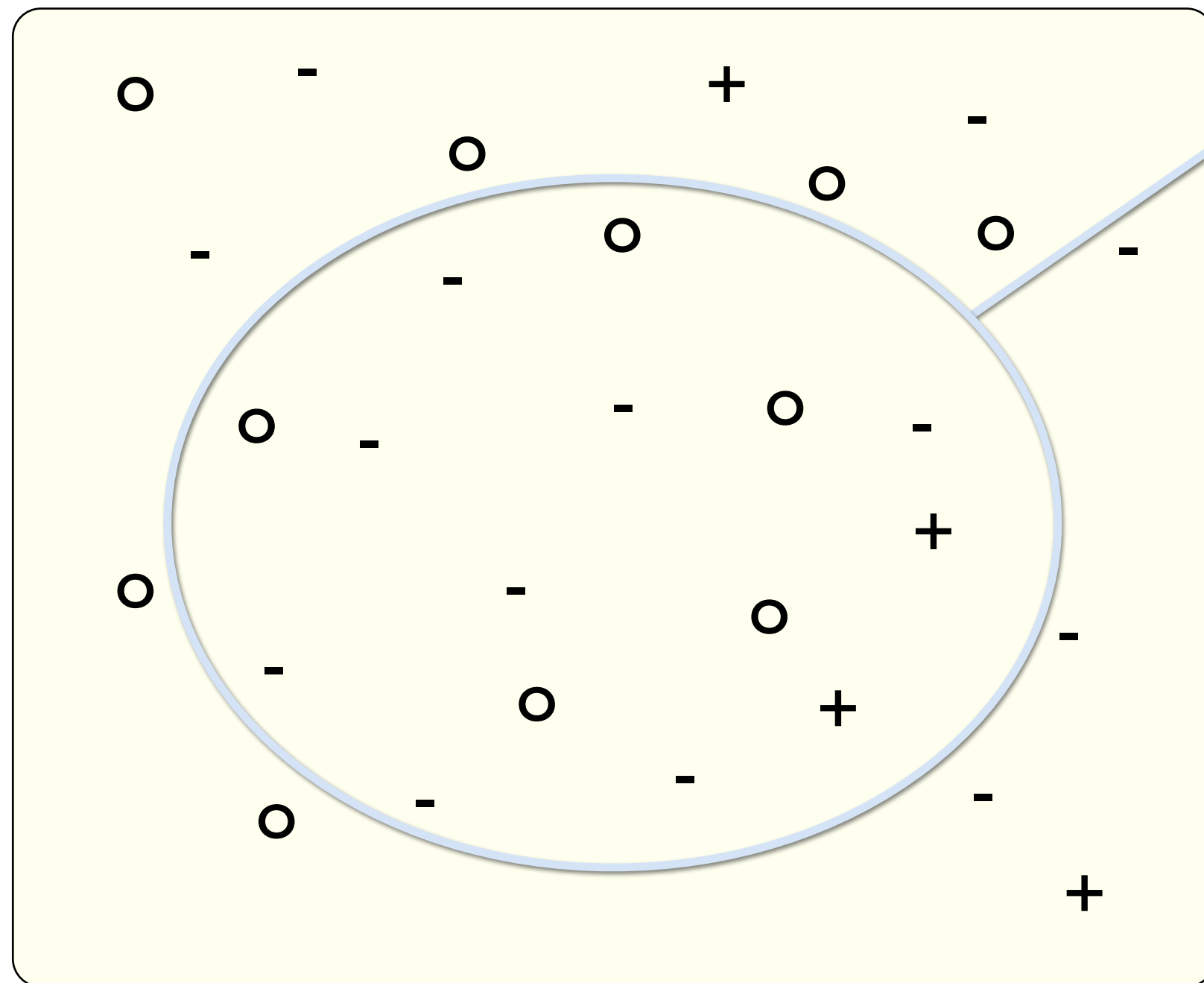
Design Space



+ good
o okay
- poor

METAPHOR

Design Space

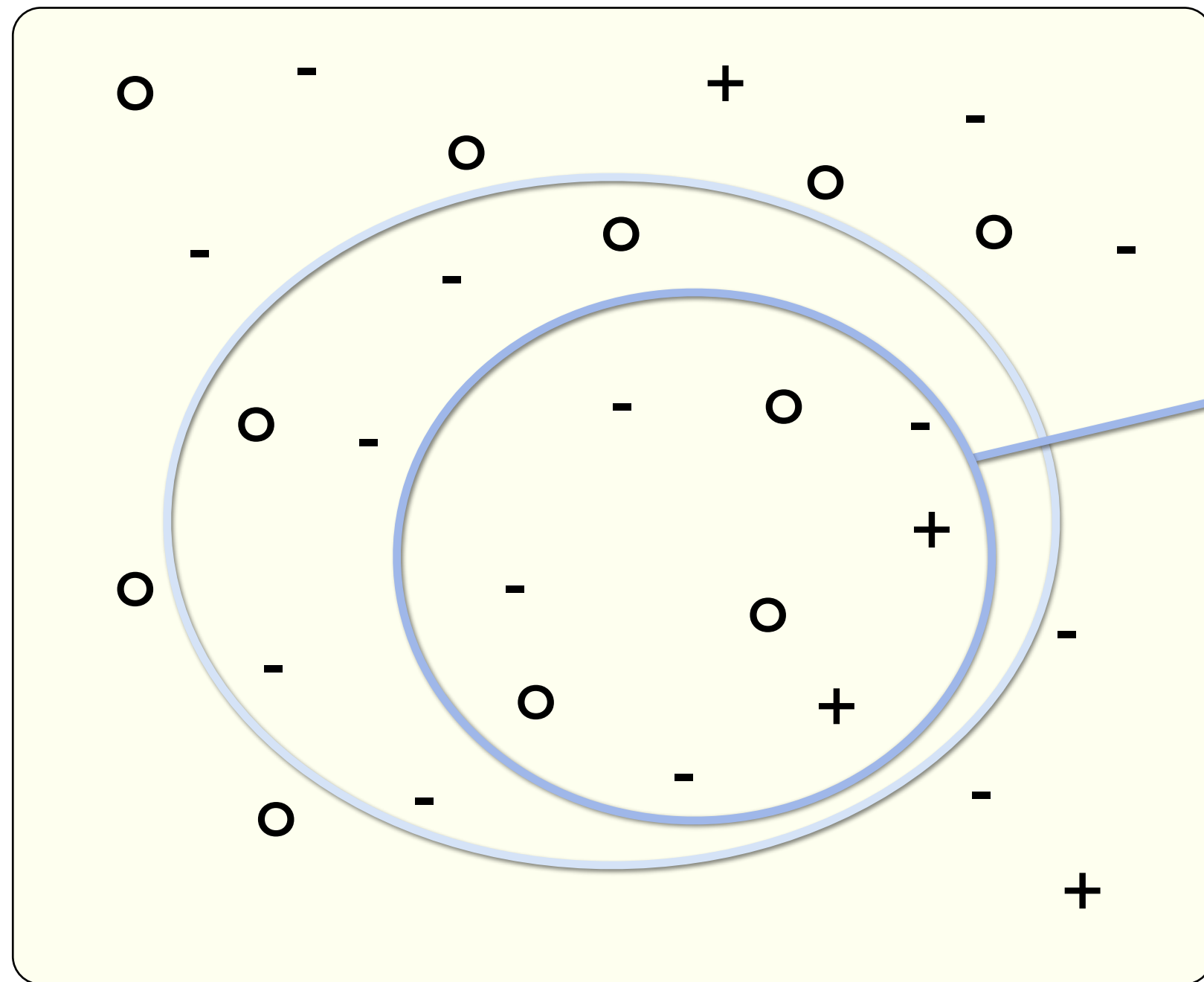


know

broad
scope

METAPHOR

Design Space

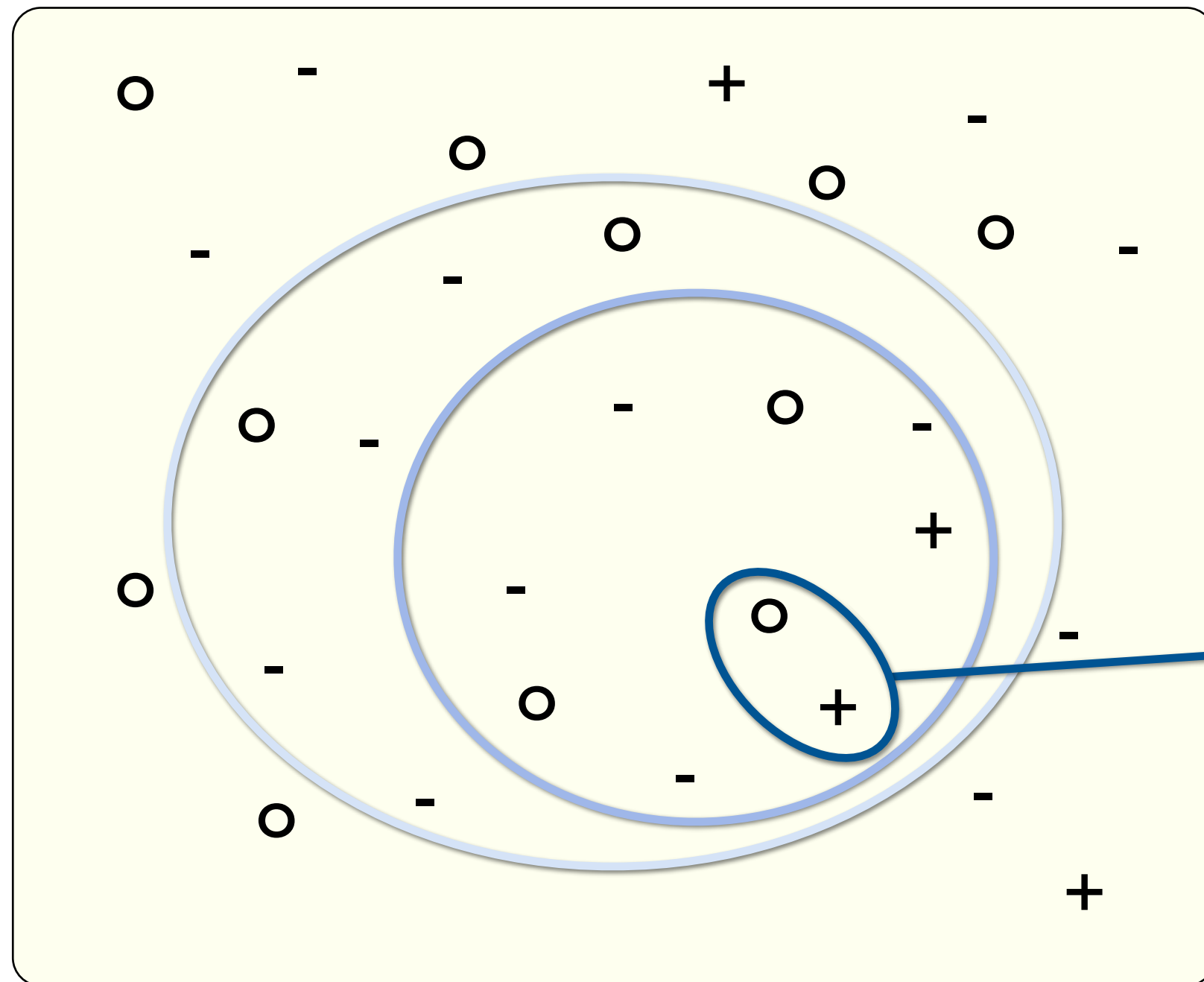


know

consider

METAPHOR

Design Space



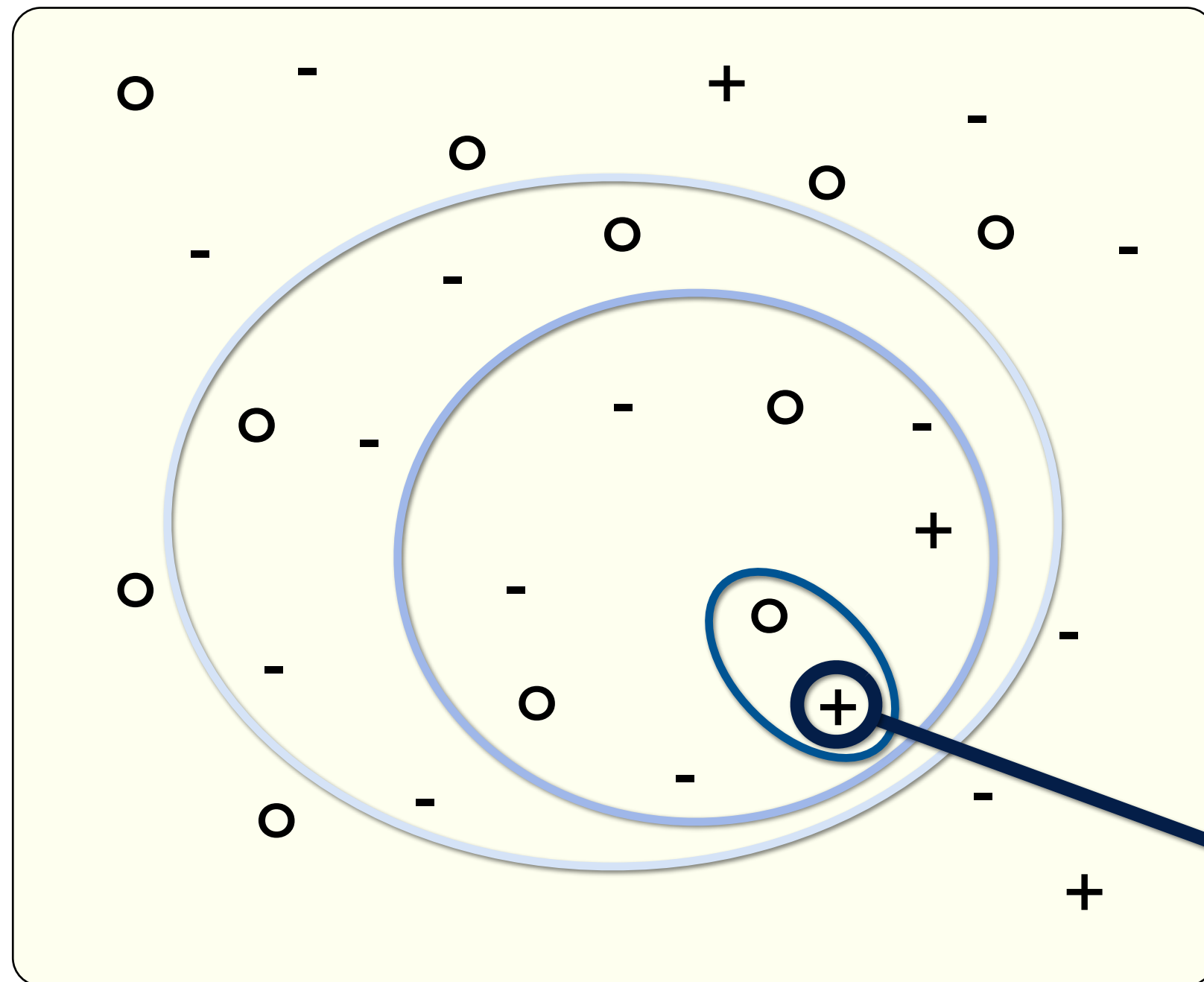
know

consider

propose

METAPHOR

Design Space



know

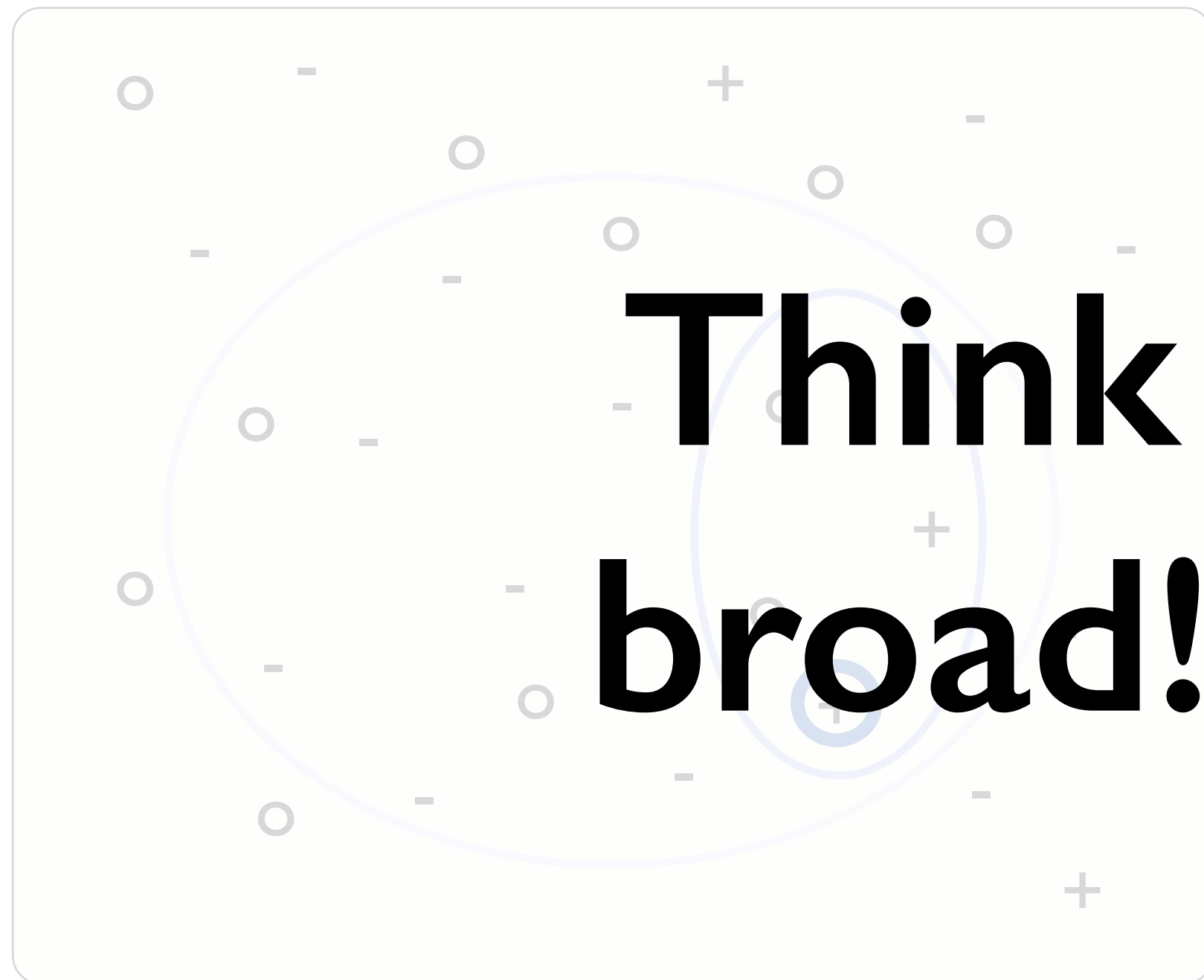
consider

propose

select

METAPHOR

Design Space



+ good
o okay
- poor

consider

propose

select

Design study methodology: 32 pitfalls

PF-10	no real/important/recurring task	winnow
PF-11	no rapport with collaborators	winnow
PF-12	not identifying front line analyst and gatekeeper before start	cast
PF-13	assuming every project will have the same role distribution	cast
PF-14	mistaking fellow tool builders for real end users	cast
PF-15	ignoring practices that currently work well	discover
PF-16	expecting <i>just talking</i> or <i>fly on wall</i> to work	discover
PF-17	experts focusing on visualization design vs. domain problem	discover
PF-18	learning their problems/language: too little / too much	discover
PF-19	abstraction: too little	design
PF-20	premature design commitment: consideration space too small	design

PITFALL

PREMATURE DESIGN
COMMITMENT

DOMAIN EXPERTS
FOCUSED ON VIS
DESIGN VS DOMAIN
PROBLEM

I want a tool with that
cool technique I saw the
other day!

Tell me more
about your
**current
workflow
problems!**



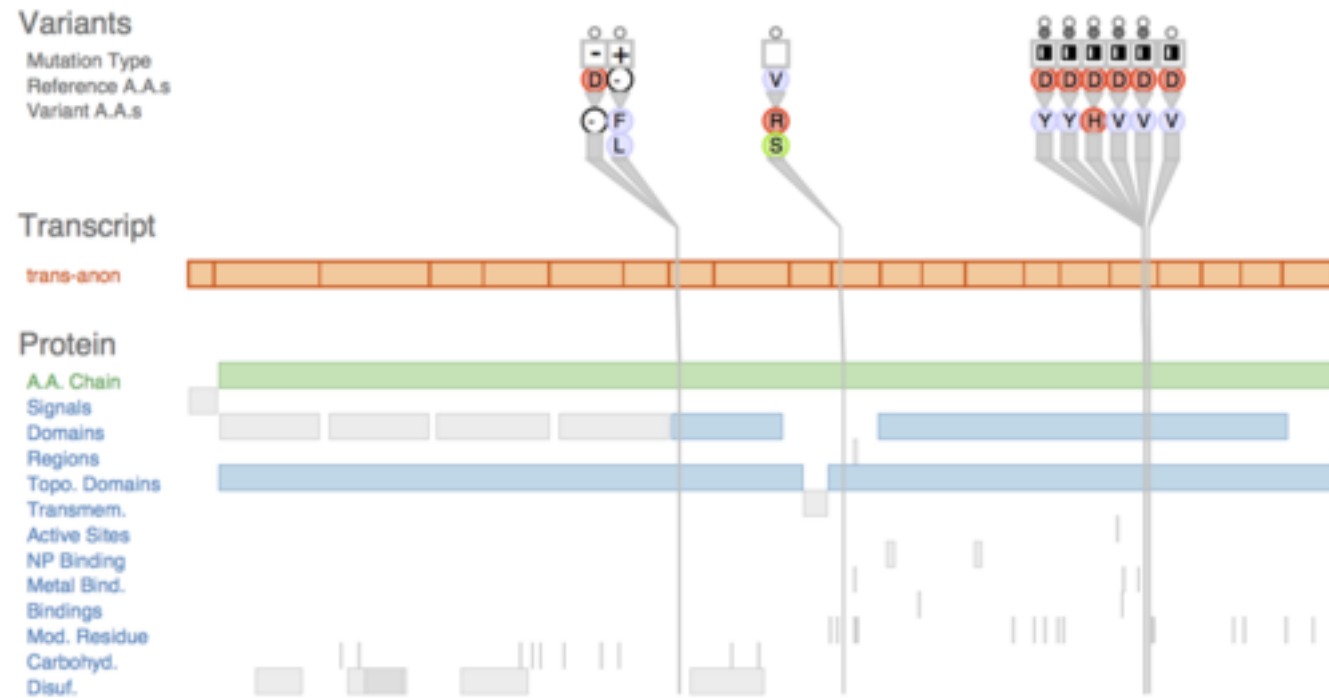
COLLABORATOR



MR. VIS

Design study methodology: 32 pitfalls

PF-21	mistaking technique-driven for problem-driven work	design
PF-22	nonrapid prototyping	implement
PF-23	usability: too little / too much	implement
PF-24	premature end: insufficient deploy time built into schedule	deploy
PF-25	usage study not case study: non-real task/data/user	deploy
PF-26	<i>liking</i> necessary but not sufficient for validation	deploy
PF-27	failing to improve guidelines: confirm, refine, reject, propose	reflect
PF-28	insufficient writing time built into schedule	write
PF-29	no technique contribution \neq good design study	write
PF-30	too much domain background in paper	write
PF-31	story told chronologically vs. focus on final results	write
PF-32	premature end: win race vs. practice music for debut	write



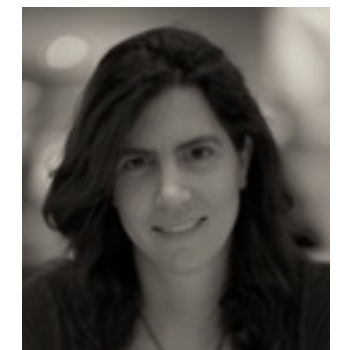
Joel Ferstay



Cydney Nielsen
@cydneybn



Tamara Munzner
@tamaramunzner



Variant View

Visualizing Sequence Variants in their Gene Context

<http://www.cs.ubc.ca/labs/imager/tr/2012/VariantView/>

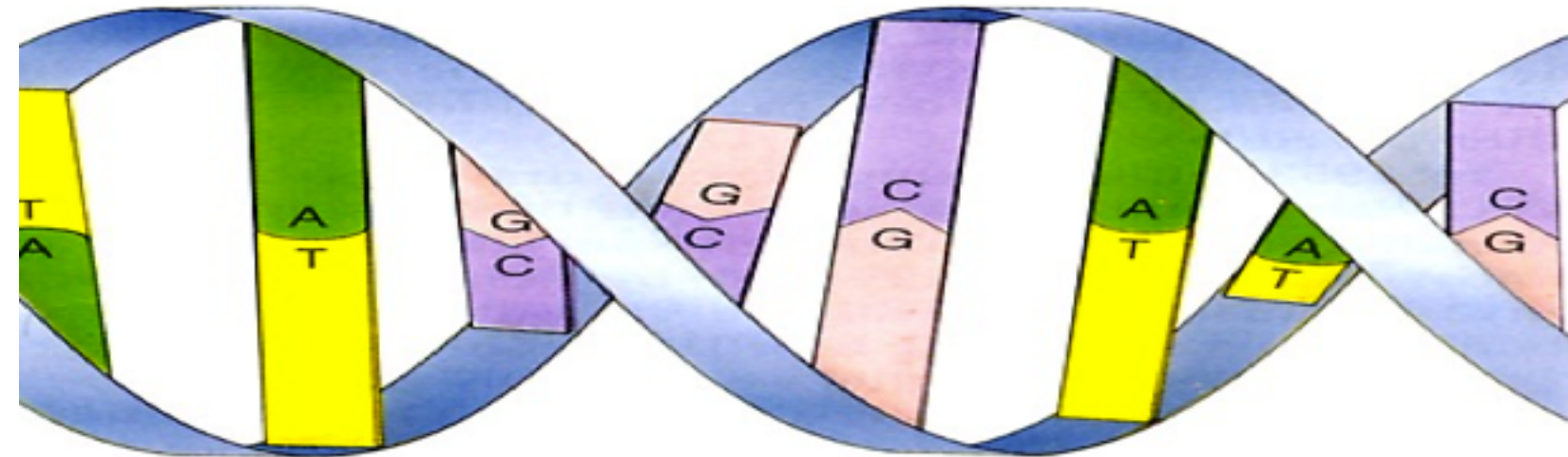
Variant View: Visualizing Sequence Variants in their Gene Context.
Ferstay, Nielsen, Munzner. *IEEE TVCG* 19(12): 2546-2555, 2013 (Proc. InfoVis 2013).

Variant View: Visualization *Design Study*

- first after DSM, tried following guidelines explicitly
- a specific **real-world** problem
 - real users and real data,
 - collaboration is (often) fundamental
- **design** a visualization system
 - implications: requirements, multiple ideas
- **validate** the design
 - at appropriate levels: case studies via deployment
- **reflect** about lessons learned
 - transferable research: improve design guidelines for vis in general
 - confirm, refine, reject, propose

Sequence Variant Definition

- Sequence variants
 - Difference between reference and given genome



Reference Genome DNA: ATA TGA TCA ACA CTT

Sample 1 Genome DNA: ATA TG**G** TCA **ATA** CTT

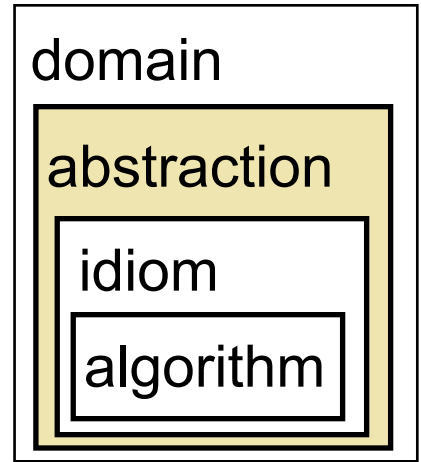
Sample 2 Genome DNA: ATA TGA **TGA** ACA **CCT**

Harmful?

Harmless?

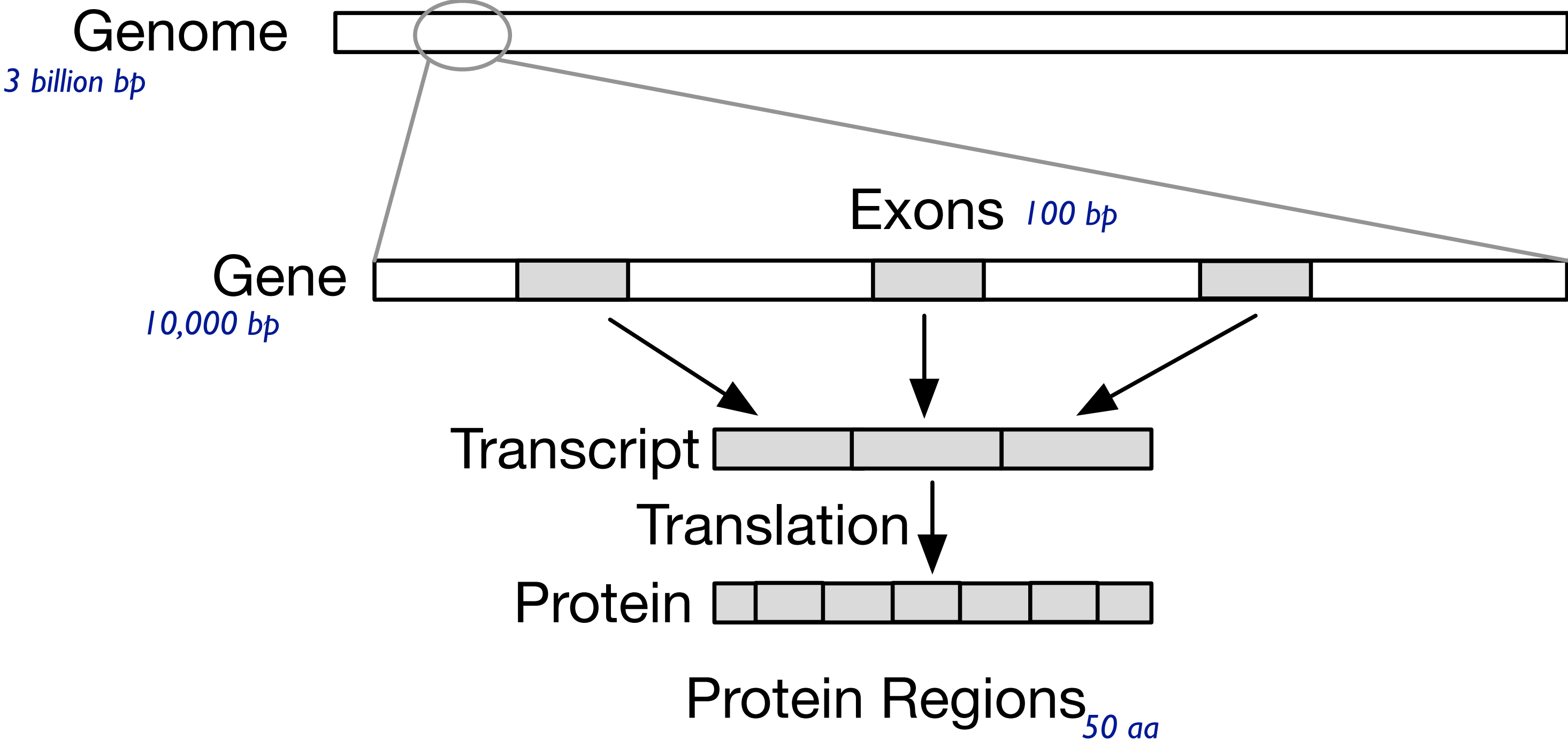
Cancer Research

- collaboration with analysts at BC Genome Sciences Center
 - studying genetic basis of leukemia
- driving task
 - discover new candidate genes with harmful variants
- two big questions
 - what to show
 - data abstraction
 - challenge: enormous range of scales in the data
 - how to show it
 - visual encoding idiom
 - challenge: information density and perceptual considerations

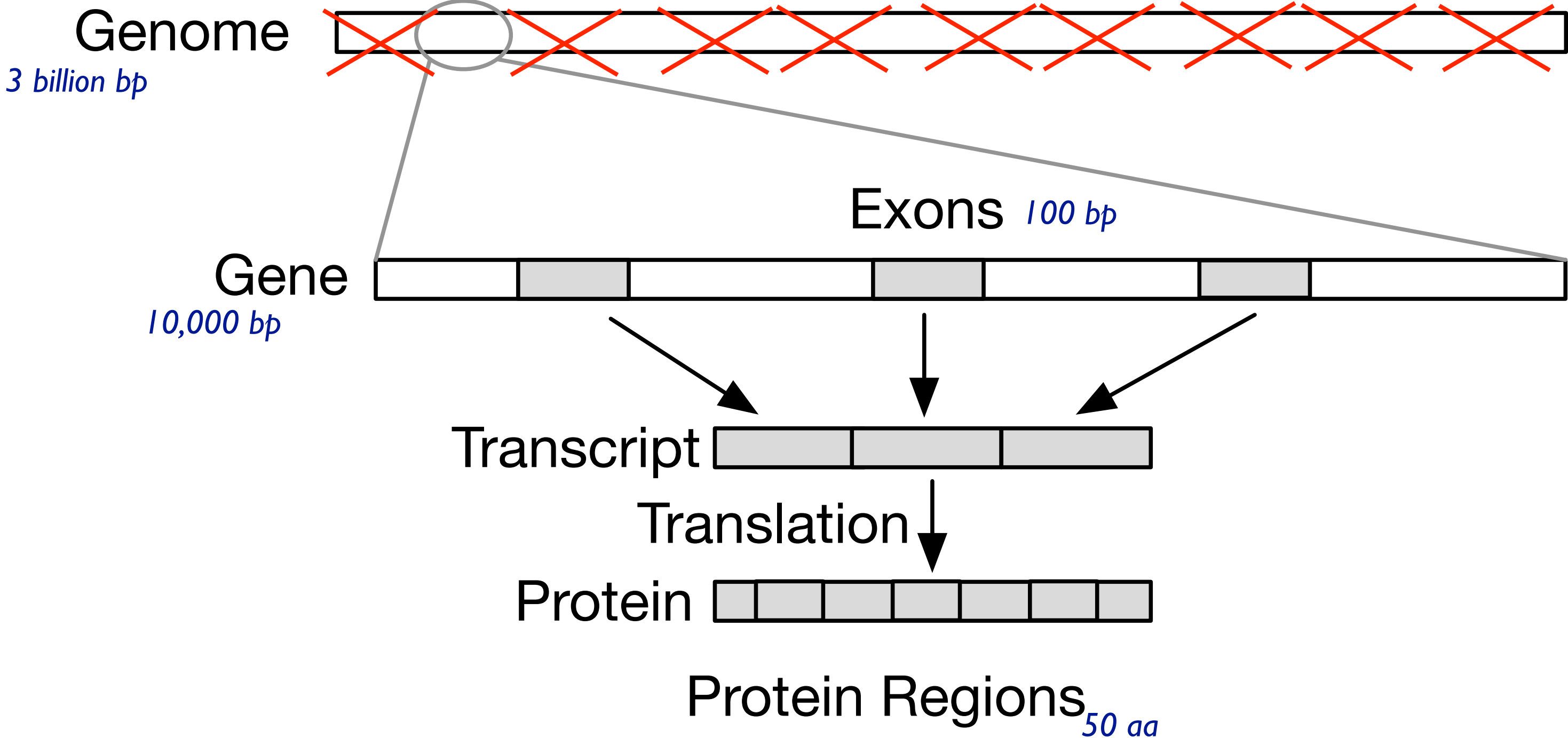


Abstractions

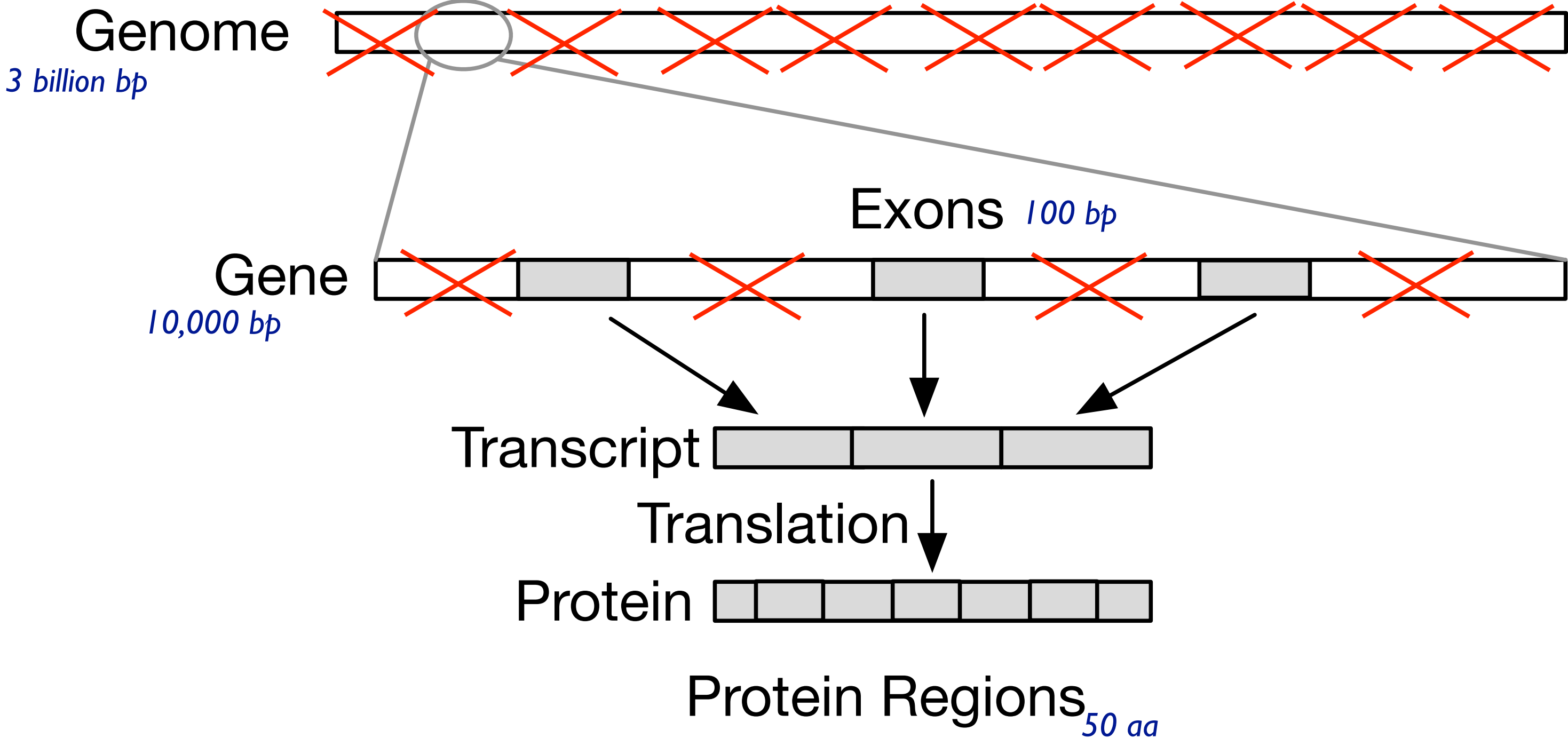
Data: Filtering to relevant biological levels and scales



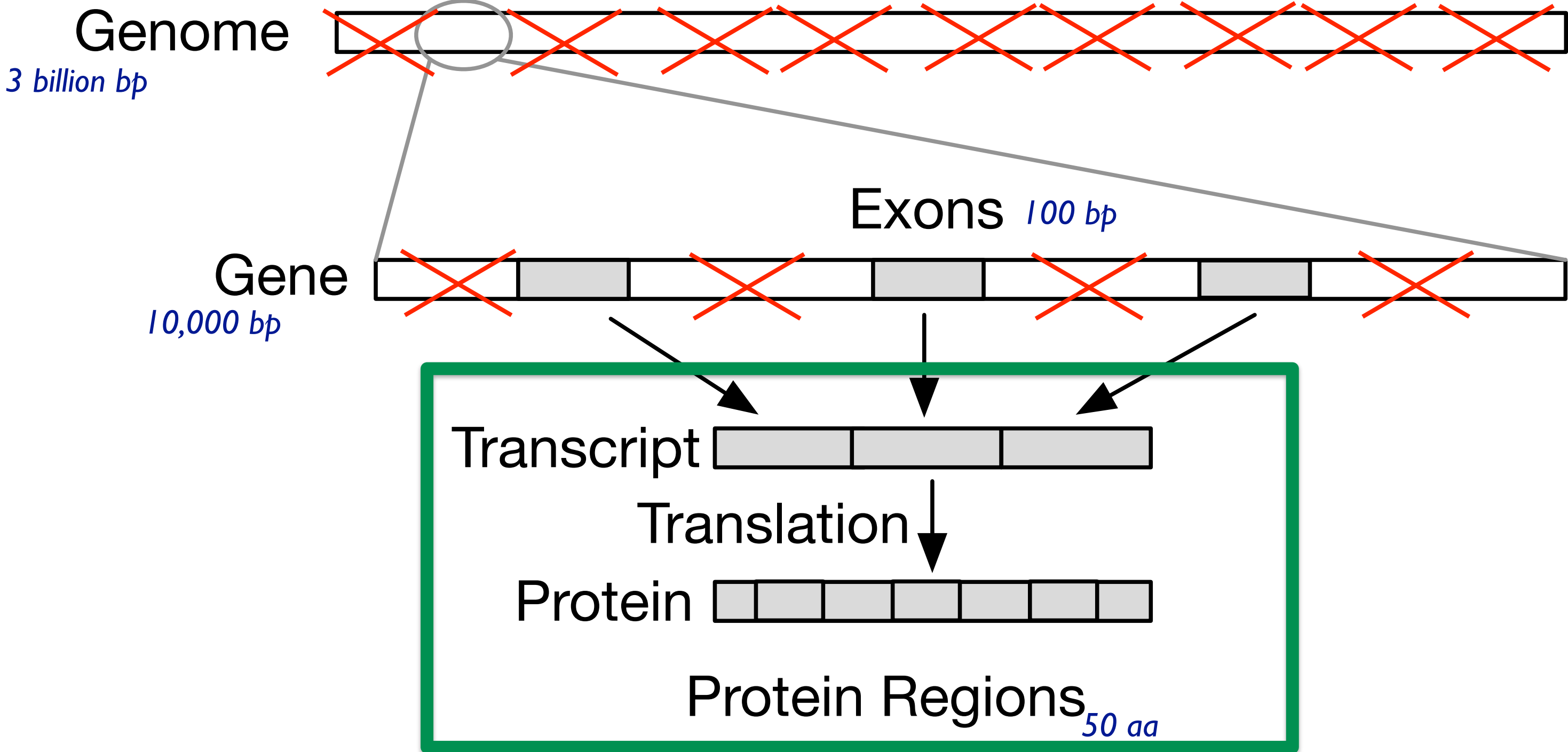
Filter out whole genome; keep genes



Filter out non-exon regions

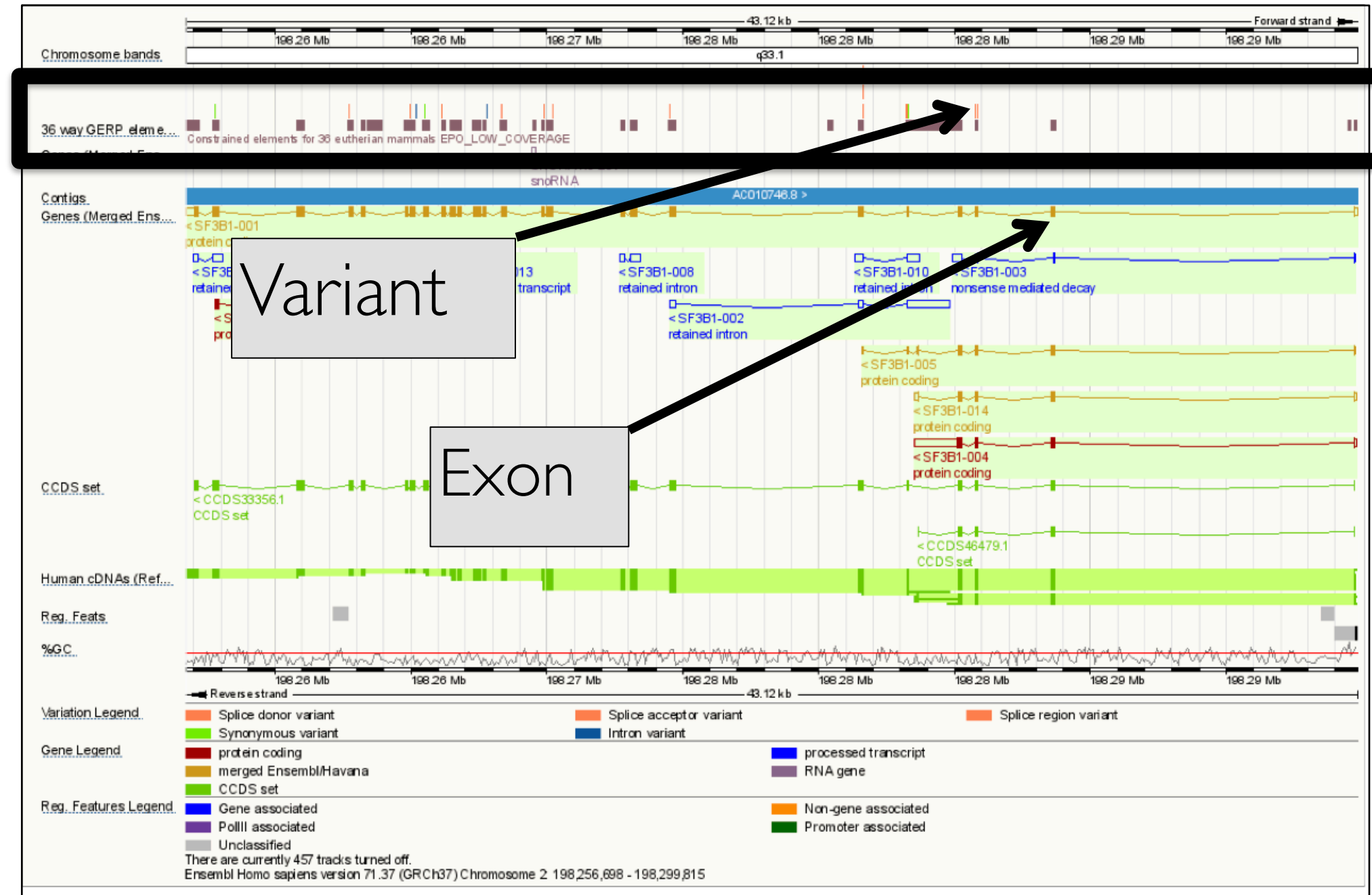


Data abstraction: highly filtered scope of *transcript coordinates*



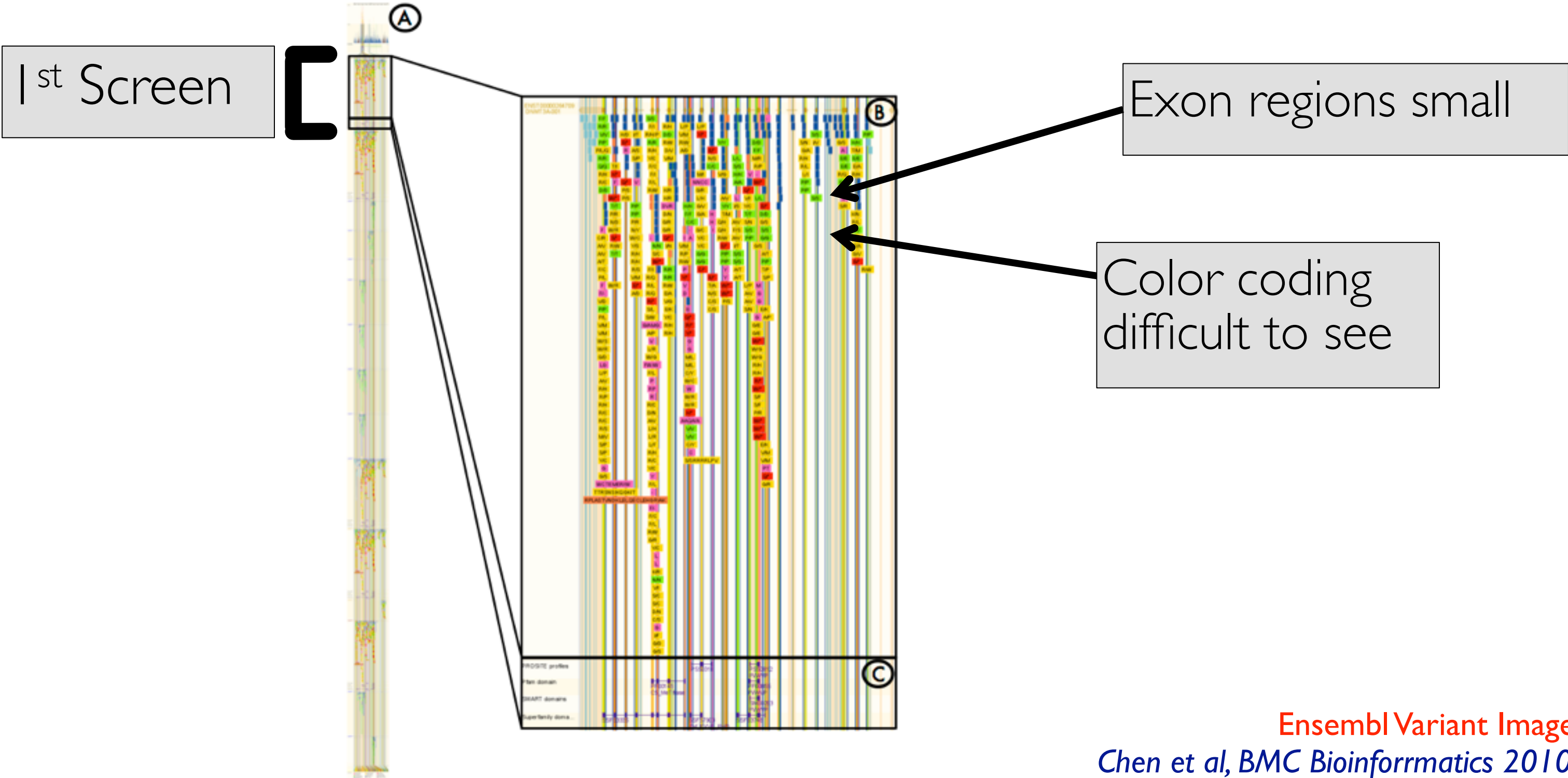
Dominant paradigm: genome browsers

- strengths: flexible and powerful
 - horizontal tracks: user data
 - shared coordinate system: genome coordinates (bp)
- problems
 - tiny features of interest spread out across large extent
 - must zoom far in to inspect known feature, then zoom out and pan to locate next
 - high cognitive load for interaction
 - must already know where to look



representative example: Ensembl
Chen et al, BMC Bioinformatics 2010.

Features of interest small even in variant-specific view



domain

abstraction

idiom

algorithm

Idioms

Variant View

Gene Search:

Alternative Transcripts:

Variants

Mutation Type
Reference A.A.s
Variant A.A.s

Transcript

trans-anon

Protein

A.A. Chain
Domains
Regions
Active Sites
Bindings
Mod. Residue

Sort By Gene:

Alpha Cluster Score **Variant Count**

- DNMT3A (NM_022552)
- IDH2 (NM_002168)
- FLT3 (NM_004119)
- ANKRD36 (NM_001164315)
- ARID1B (NM_017519)
- STAG2 (NM_001042749)
- TNRC18 (NM_001080495)
- WT1 (NM_000378)
- ABCA13 (NM_152701)
- CEBPA (NM_004364)
- TET2 (NM_001127208)
- DNAH10 (NM_207437)
- GPSM1 (NM_015597)
- ASXL1 (NM_015338)
- DNAH1 (NM_015512)
- DNAH6 (NM_001370)
- FAT1 (NM_005245)
- MDN1 (NM_014611)
- PTPN11 (NM_002834)
- SYNE1 (NM_033071)
- ALMS1 (NM_015120)
- C10orf68 (NM_024688)
- CCDC88C (NM_001080414)
- DNAH11 (NM_003777)
- DNAH3 (NM_017539)
- DNAH9 (NM_001372)

Variant Data

Patient ID	Chr. Coord.	Ref Base	Var Base	dbSNP129	dbSNP135	dbSNP137	COSMIC	A.A. Chng.	Gene	Ref. Gene
pid-anon	11288816	G	T	.	.	.	*13028,	G60V	gene-anon	trans-anon
pid-anon	11288816	G	T	.	.	.	*13012,	D61Y	gene-anon	trans-anon
pid-anon	11288819	G	T	.	rs121918	.	13014	A72S	gene-anon	trans-anon
pid-anon	11288819	C	T	.	.	.	*13035,	A72V	gene-anon	trans-anon
pid-anon	11288821	G	C	.	.	.	*13016,	E76Q	gene-anon	trans-anon
pid-anon	11288821	A	G	.	rs121918	.	*13017,	E76G	gene-anon	trans-anon
pid-anon	11288821	G	T	E76D	gene-anon	trans-anon
pid-anon	11292688	T	A	.	rs121918	.	*13020,	S502T	gene-anon	trans-anon
pid-anon	11292688	T	G	.	.	.	*13020,	S502A	gene-anon	trans-anon
pid-anon	11292688	C	T	.	.	.	13023	S502L	gene-anon	trans-anon

Variant View

Information-dense single gene view

Gene Search:

Alternative Transcripts:

Variants
 Mutation Type
 Reference A.A.s
 Variant A.A.s

Transcript
 trans-anon

Protein
 A.A. Chain
 Domains
 Regions
 Active Sites
 Bindings
 Mod. Residue

Variant Data

Patient ID	Chr. Coord.	Ref Base	Var Base	dbSNP129	dbSNP135	dbSNP137	COSMIC	A.A. Chng.	Gene	Ref. Gene
pid-anon	11288816	G	T	.	.	.	*13028,	G60V	gene-anon	trans-anon
pid-anon	11288816	G	T	.	.	.	*13012,	D61Y	gene-anon	trans-anon
pid-anon	11288819	G	T	.	rs121918	.	13014	A72S	gene-anon	trans-anon
pid-anon	11288819	C	T	.	.	.	*13035,	A72V	gene-anon	trans-anon
pid-anon	11288821	G	C	.	.	.	*13016,	E76Q	gene-anon	trans-anon
pid-anon	11288821	A	G	.	rs121918	.	*13017,	E76G	gene-anon	trans-anon
pid-anon	11288821	G	T	E76D	gene-anon	trans-anon
pid-anon	11292688	T	A	.	rs121918	.	*13020,	S502T	gene-anon	trans-anon
pid-anon	11292688	T	G	.	.	.	*13020,	S502A	gene-anon	trans-anon
pid-anon	11292688	C	T	.	.	.	13023	S502L	gene-anon	trans-anon

Sort By Gene:

- DNMT3A (NM_022552)
- IDH2 (NM_002168)
- FLT3 (NM_004119)
- ANKRD36 (NM_001164315)
- ARID1B (NM_017519)
- STAG2 (NM_001042749)
- TNRC18 (NM_001080495)
- WT1 (NM_000378)
- ABCA13 (NM_152701)
- CEBPA (NM_004364)
- TET2 (NM_001127208)
- DNAH10 (NM_207437)
- GPSM1 (NM_015597)
- ASXL1 (NM_015338)
- DNAH1 (NM_015512)
- DNAH6 (NM_001370)
- FAT1 (NM_005245)
- MDN1 (NM_014611)
- PTPN11 (NM_002834)
- SYNE1 (NM_033071)
- ALMS1 (NM_015120)
- C10orf68 (NM_024688)
- CCDC88C (NM_001080414)
- DNAH11 (NM_003777)
- DNAH3 (NM_017539)
- DNAH9 (NM_001372)

Variant View

Gene Search:

Alternative Transcripts:

Information-dense single gene view

Sort By Gene:
Alpha Cluster Score Variant Count

Variant Data

Patient ID	Chr.	Coord.	Ref Base	Var Base	dbSNP129	dbSNP135	dbSNP137	COSMIC	A.A. Ch	Gene	Transcript
pid-anon	11	28816	G	T	.	.	.	*13028,	G60V	gene-anon	trans-anon
pid-anon	11	28816	G	T	.	.	.	*13012,	D61Y	gene-anon	trans-anon
pid-anon	11	28819	G	T	.	rs121918	.	13014	A72S	gene-anon	trans-anon
pid-anon	11	28819	C	T	.	.	.	*13035,	A72V	gene-anon	trans-anon
pid-anon	11	28821	G	C	.	.	.	*13016,	E76Q	gene-anon	trans-anon
pid-anon	11	28821	A	G	.	rs121918	.	*13017,	E76G	gene-anon	trans-anon
pid-anon	11	28821	G	T	E76D	gene-anon	trans-anon
pid-anon	11	292688	T	A	.	rs121918	.	*13020,	S502T	gene-anon	trans-anon
pid-anon	11	292688	T	G	.	.	.	*13020,	S502A	gene-anon	trans-anon
pid-anon	11	292688	C	T	.	.	.	13023	S502L	gene-anon	trans-anon

No need for pan and zoom

DNMT3A (NM_022552) ©
IDH2 (NM_002168)
FLT3 (NM_004119)
ANKRD36 (NM_001164315)
ARID1B (NM_017519)
STAG2 (NM_001042749)
TNRC18 (NM_001080495)
WT1 (NM_000378)
ABCA13 (NM_152701)
CEBPA (NM_004364)
TET2 (NM_001127208)
DNAH10 (NM_207437)
GPSM1 (NM_015597)
ASXL1 (NM_015338)
DNAH1 (NM_015512)
DNAH6 (NM_001270)

SYNET (NM_033071)
ALMS1 (NM_015120)
C10orf68 (NM_024688)
CCDC88C (NM_001080414)
DNAH11 (NM_003777)
DNAH3 (NM_017539)
DNAH9 (NM_001372)

Variant View

*derived data guides
human-in-the-loop analytics*

Sorting metrics guide gene navigation

Alternative Transcripts:

Variants
Mutation Type
Reference A.A.s
Variant A.A.s

Transcript
trans-anon

Protein
A.A. Chain
Domains
Regions
Active Sites
Bindings
Mod. Residue

Variant Data

Patient ID	Chr. Coord.	Ref Base	Var Base	dbSNP129	dbSNP135	dbSNP137	COSMIC	A.A. Chng.	Gene	Ref. Gene
pid-anon	11288816	G	T	.	.	.	*13028,	G60V	gene-anon	trans-anon
pid-anon	11288816	G	T	.	.	.	*13012,	D61Y	gene-anon	trans-anon
pid-anon	11288819	G	T	.	rs121918	.	13014	A72S	gene-anon	trans-anon
pid-anon	11288819	C	T	.	.	.	*13035,	A72V	gene-anon	trans-anon
pid-anon	11288821	G	C	.	.	.	*13016,	E76Q	gene-anon	trans-anon
pid-anon	11288821	A	G	.	rs121918	.	*13017,	E76G	gene-anon	trans-anon
pid-anon	11288821	G	T	E76D	gene-anon	trans-anon
pid-anon	11292688	T	A	.	rs121918	.	*13020,	S502T	gene-anon	trans-anon
pid-anon	11292688	T	G	.	.	.	*13020,	S502A	gene-anon	trans-anon
pid-anon	11292688	C	T	.	.	.	13023	S502L	gene-anon	trans-anon

Sort By Gene:
Alpha

- DNMT3A (NM_022552)
- IDH2 (NM_002168)
- FLT3 (NM_004119)
- ANKRD36 (NM_001164315)
- ARID1B (NM_017519)
- STAG2 (NM_001042749)
- TNRC18 (NM_001080495)
- WT1 (NM_000378)
- ABCA13 (NM_152701)
- CEBPA (NM_004364)
- TET2 (NM_001127208)
- DNAH10 (NM_207437)
- GPSM1 (NM_015597)
- ASXL1 (NM_015338)
- DNAH1 (NM_015512)
- DNAH6 (NM_001370)
- FAT1 (NM_005245)
- MDN1 (NM_014611)
- PTPN11 (NM_002834)
- SYNE1 (NM_033071)
- ALMS1 (NM_015120)
- C10orf68 (NM_024688)
- CCDC88C (NM_001080414)
- DNAH11 (NM_003777)
- DNAH3 (NM_017539)
- DNAH9 (NM_001372)

Callouts: A (Sort By Gene), B (Variant Data table), C (Gene list)

Variant View

Sorting metrics guide gene navigation

The screenshot displays a 'Variant View' interface. At the top, there are sorting options: 'Alpha', 'Cluster Score', and 'Variant Count'. A list of genes is shown on the right, with 'DNMT3A (NM_022552)' at the top. The main area is divided into several sections: 'Alternative Transcripts' (showing 'gene-anon (trans-anon)'), 'Variants' (with mutation types and amino acid changes), 'Transcript' (showing an orange bar for 'trans-anon'), 'Protein' (showing a green bar for 'A.A. Chain' and blue bars for 'Domains'), and 'Variant Data' (a table of variant information). Callouts A, B, and C are present: A points to the sorting options, B points to the 'Variant Data' table, and C points to the gene list.

Variant ID	dbSNP137	COSMIC	A.A. Chng.	Gene	RefSeq
pid-anon	11288819	C	T	gene-anon	trans-anon
pid-anon	11288821	G	C	gene-anon	trans-anon
pid-anon	11288821	A	G	gene-anon	trans-anon
pid-anon	11288821	G	T	gene-anon	trans-anon
pid-anon	11292688	T	A	gene-anon	trans-anon
pid-anon	11292688	T	G	gene-anon	trans-anon
pid-anon	11292688	C	T	gene-anon	trans-anon

Control what shows up here

Variant View

Gene Search:

Alternative Transcripts:

Variants

Mutation Type
Reference A.A.s
Variant A.A.s

Transcript

trans-anon

Protein

A.A. Chain
Domains
Regions
Active Sites
Bindings
Mod. Residue

Variant Data

Patient ID	Chr. Coord.	Ref Base	Var Base	dbSNP129	dbSNP135	dbSNP137	COSMIC	A.A. Chng.	Gene	Ref. Gene
pid-anon	11288816	G	T	.	.	.	*13028,	G60V	gene-anon	trans-anon
pid-anon	11288816	G	T	.	.	.	*13012,	D61Y	gene-anon	trans-anon
pid-anon	11288819	G	T	.	rs121918	.	*13014,	A72S	gene-anon	trans-anon
pid-anon	11288819	C	T	.	.	.	*13035,	E76D	gene-anon	trans-anon
pid-anon	11288821	G	C	.	.	.	*13016,	E76Q	gene-anon	trans-anon
pid-anon	11288821	A	G	.	rs121918	.	*13017,	E76G	gene-anon	trans-anon
pid-anon	11288821	G	T	E76D	gene-anon	trans-anon
pid-anon	11292688	T	A	.	rs121918	.	*13020,	S502T	gene-anon	trans-anon
pid-anon	11292688	T	G	.	.	.	*13020,	S502A	gene-anon	trans-anon
pid-anon	11292688	C	T	.	.	.	13023	S502L	gene-anon	trans-anon

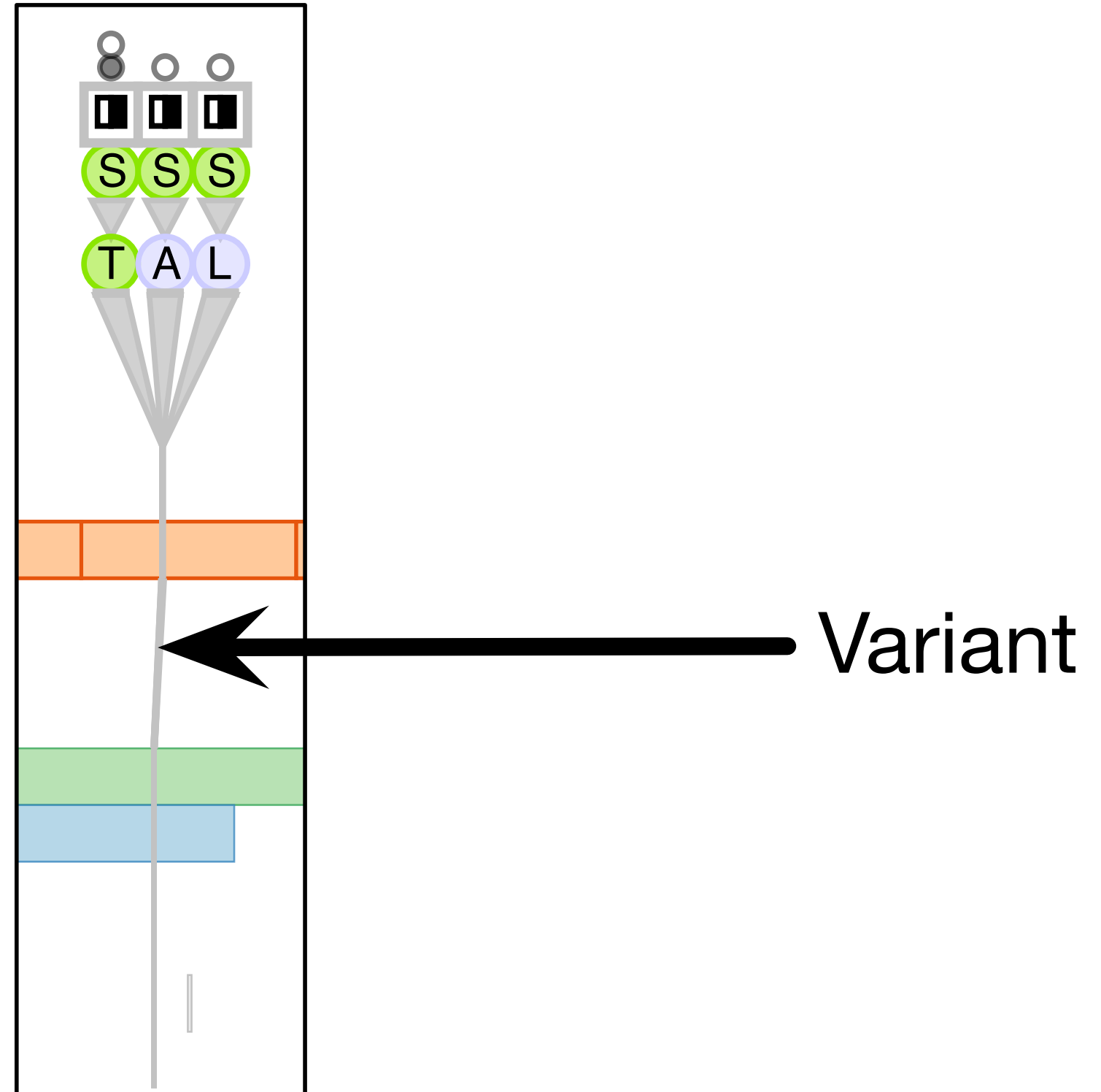
Sort By Gene:
Alpha Cluster Score **Variant Count**

- DNMT3A (NM_022552)
- IDH2 (NM_002168)
- FLT3 (NM_004119)
- ANKRD36 (NM_001164315)
- ARID1B (NM_017519)
- STAG2 (NM_001042749)
- TNRC18 (NM_001080495)
- WT1 (NM_000378)
- ABCA13 (NM_152701)
- CEBPA (NM_004364)
- TET2 (NM_001127208)
- DNAH10 (NM_207437)
- GPSM1 (NM_015597)
- ASXL1 (NM_015338)
- DNAH1 (NM_015512)
- DNAH6 (NM_001370)
- FAT1 (NM_005245)
- MDN1 (NM_014611)
- PTPN11 (NM_002834)
- DNAH11 (NM_003777)
- DNAH3 (NM_017539)
- DNAH9 (NM_001372)

Peripheral supporting data

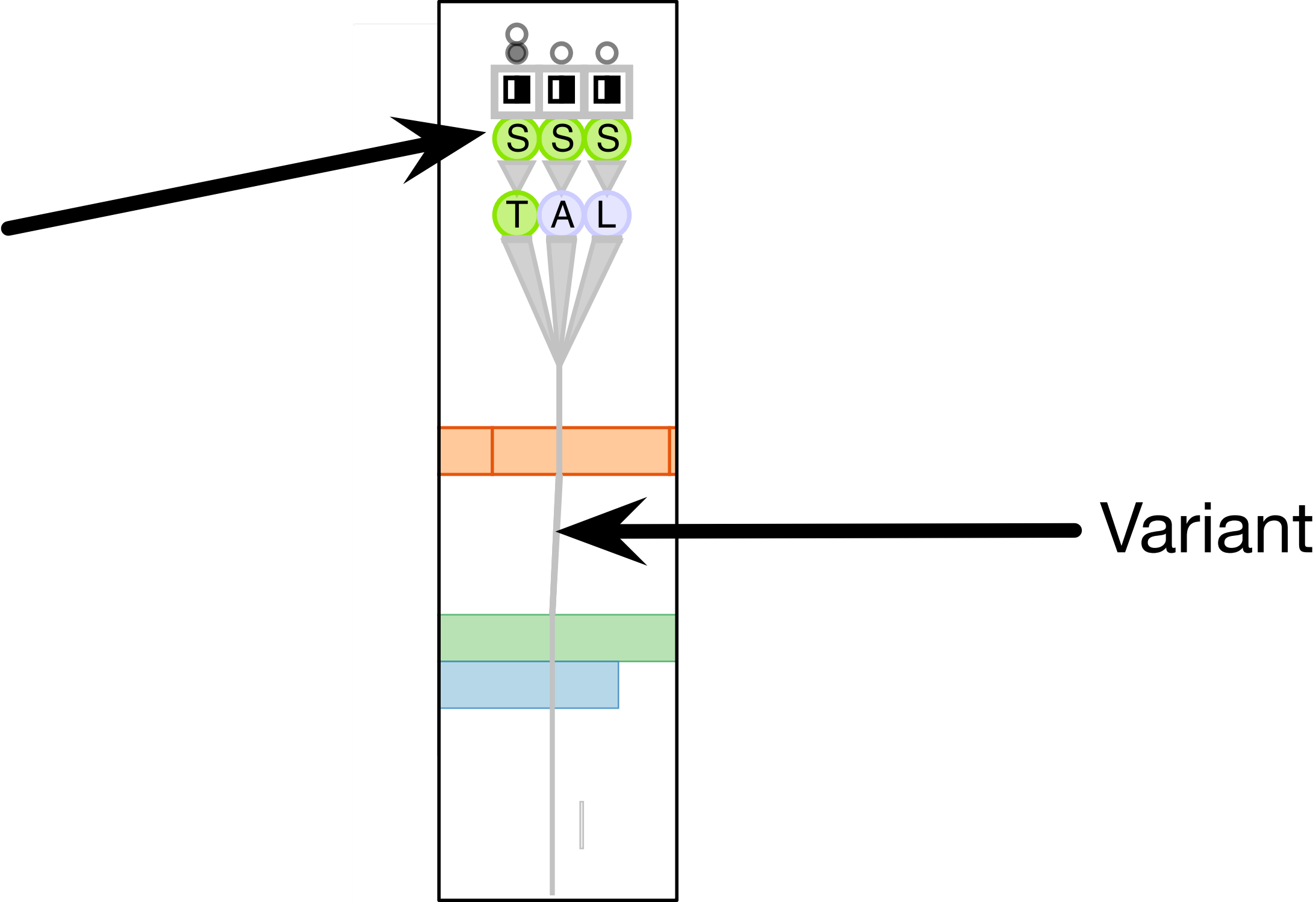
Design information-dense visual encoding

- show all attributes necessary for variant analysis
 - match salience with importance for analysis task
- variant not just a thin line!
- emphasize with high salience
 - collocated variants fan out at top
 - grey variant vertical stroke intersects horizontal colored protein regions



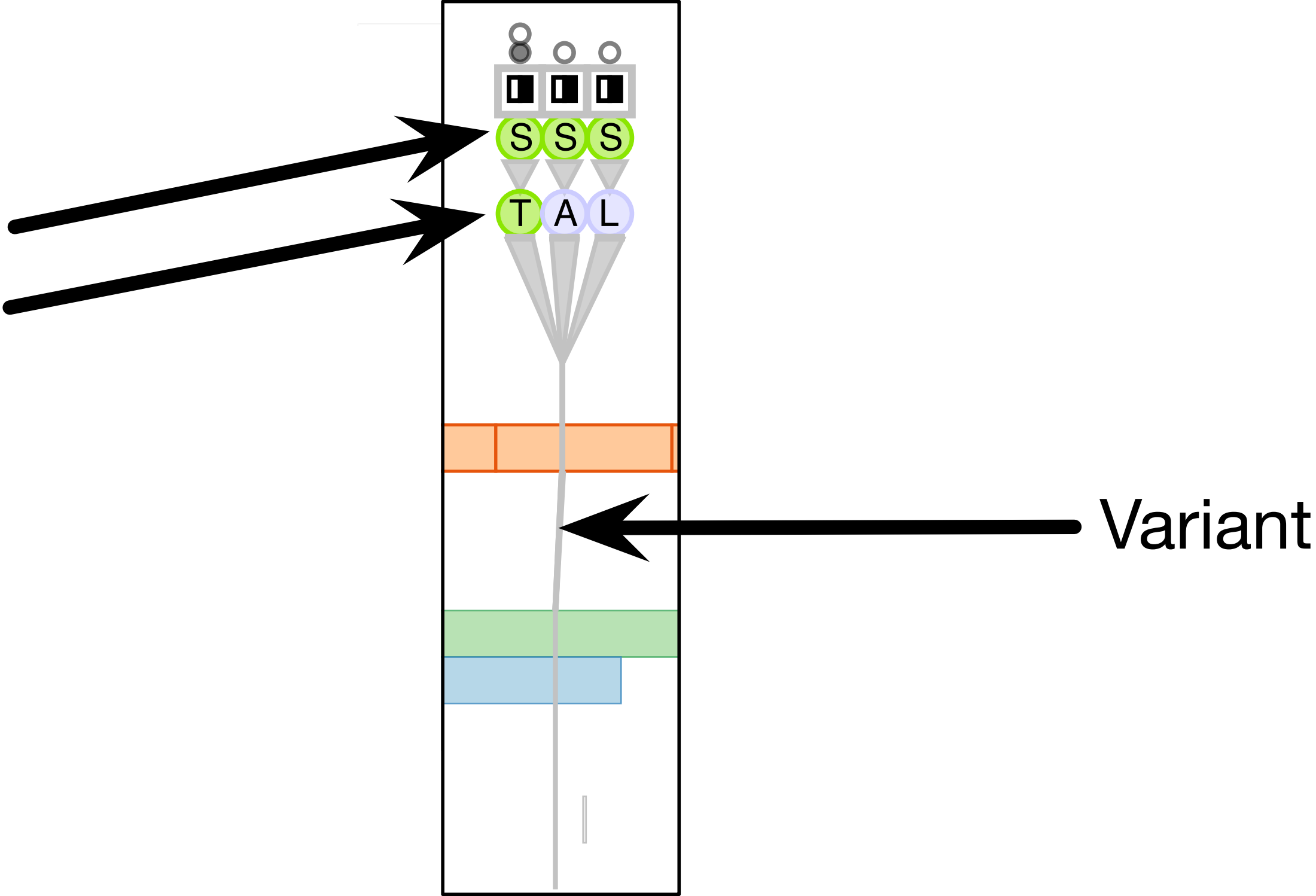
Design information-dense visual encoding

Reference AA



Design information-dense visual encoding

Reference AA
Variant AA



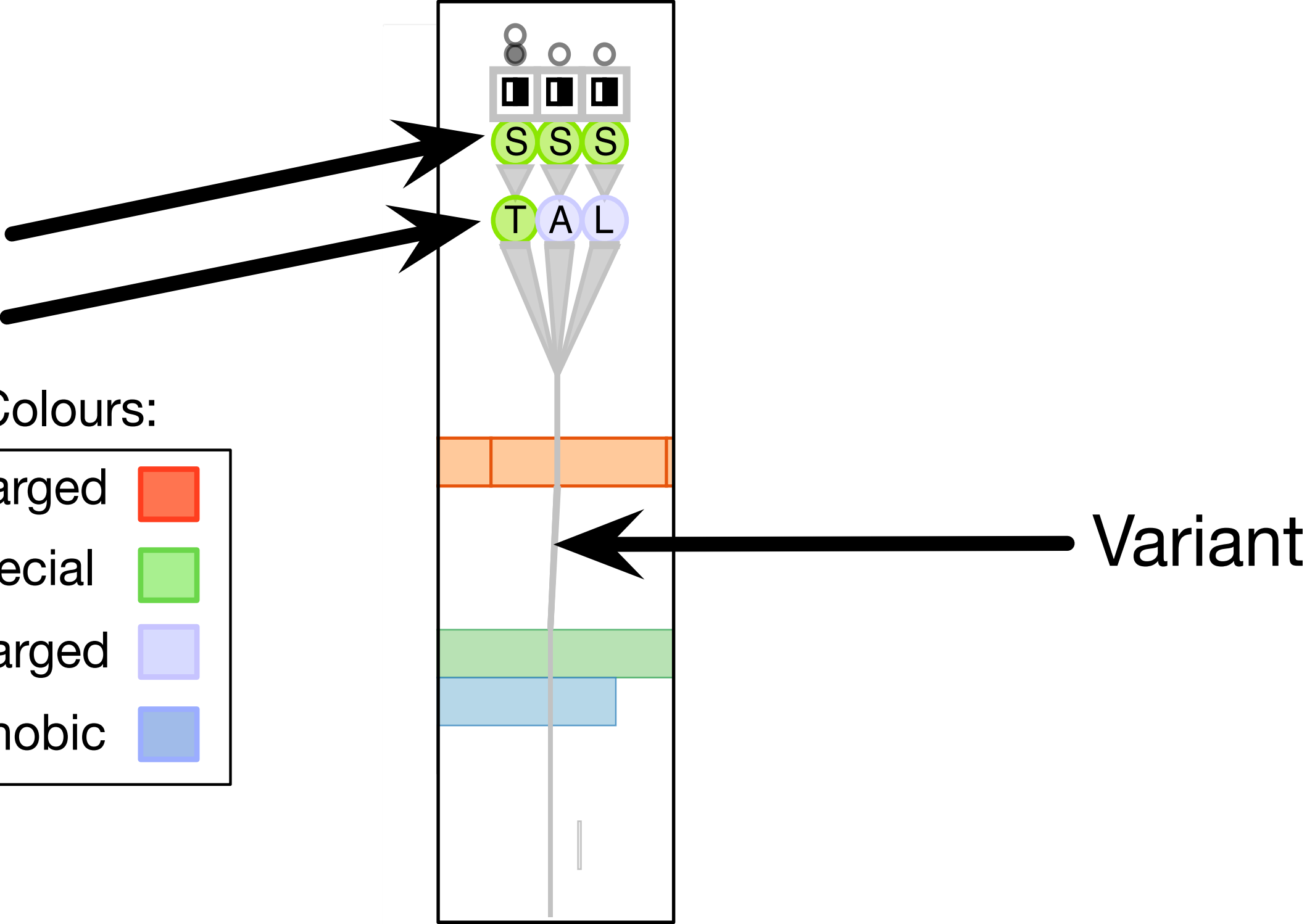
Design information-dense visual encoding

Reference AA

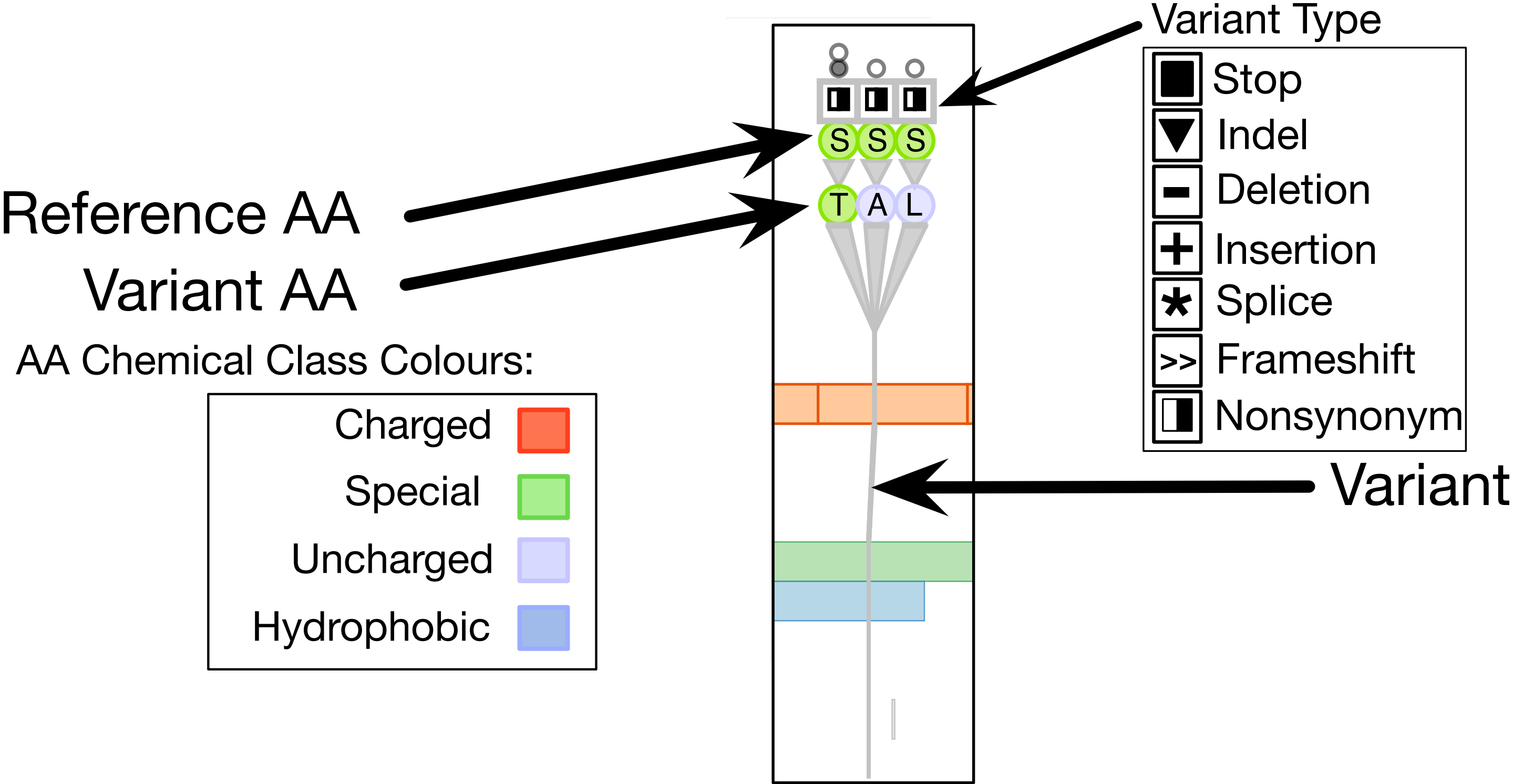
Variant AA

AA Chemical Class Colours:

Charged	■
Special	■
Uncharged	■
Hydrophobic	■



Design information-dense visual encoding



Design information-dense visual encoding

Known Database

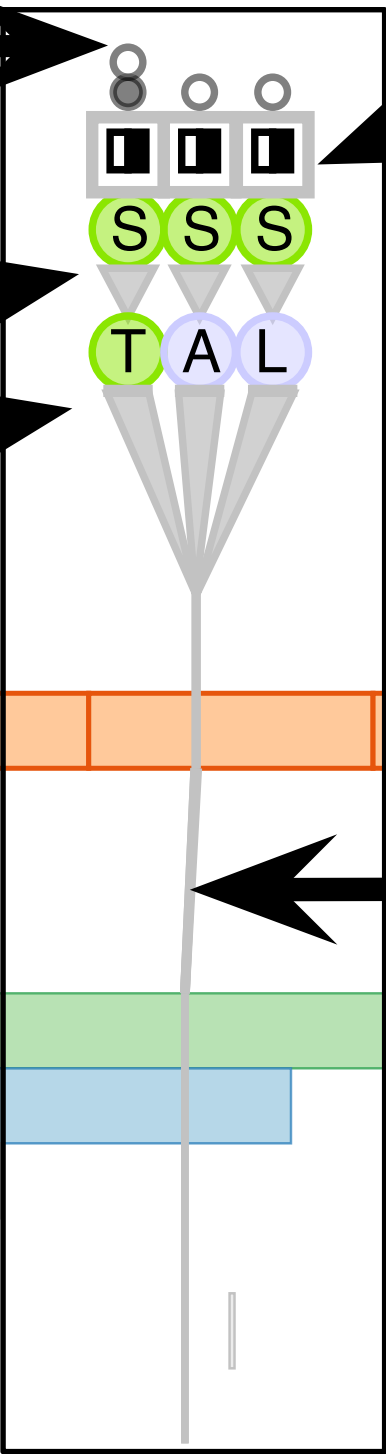
- Known Harmless
- Known Cancer

Reference AA

Variant AA

AA Chemical Class Colours:

- Charged ■
- Special ■
- Uncharged ■
- Hydrophobic ■



Variant Type

- Stop
- ▼ Indel
- ▬ Deletion
- ⊕ Insertion
- * Splice
- >> Frameshift
- ▬ Nonsynonym

Variant

Design information-dense visual encoding

Known Database

- Known Harmless
- Known Cancer

Reference AA

Variant AA

AA Chemical Class Colours:

- Charged ■
- Special ■
- Uncharged ■
- Hydrophobic ■

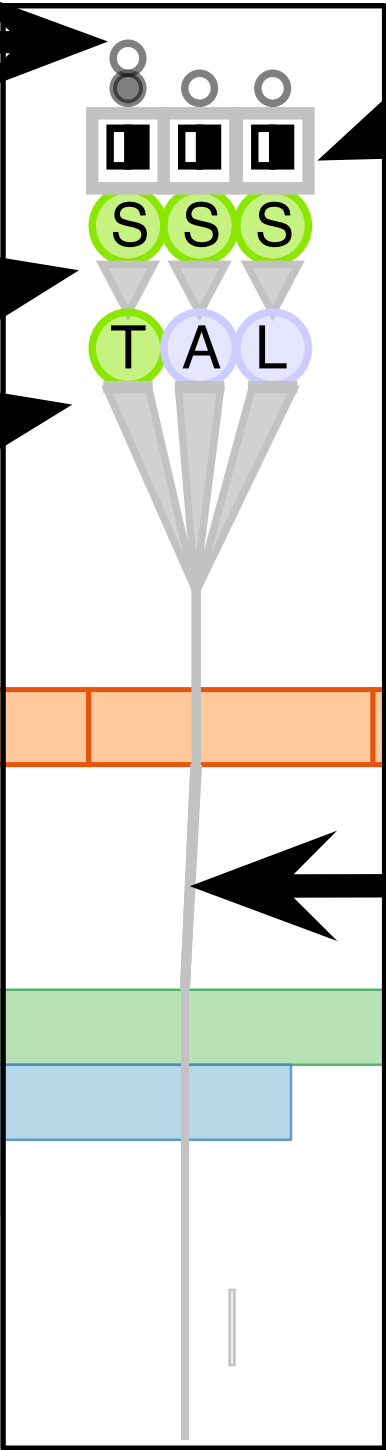
Variant Type

- Stop
- ▼ Indel
- ▬ Deletion
- ⊕ Insertion
- * Splice
- >> Frameshift
- ▬ Nonsynonym

Variant

Transcript/Region Colours:

- Transcript ■
- AA Chain ■
- All Other Regions ■
- Non-Intersected Regions ■

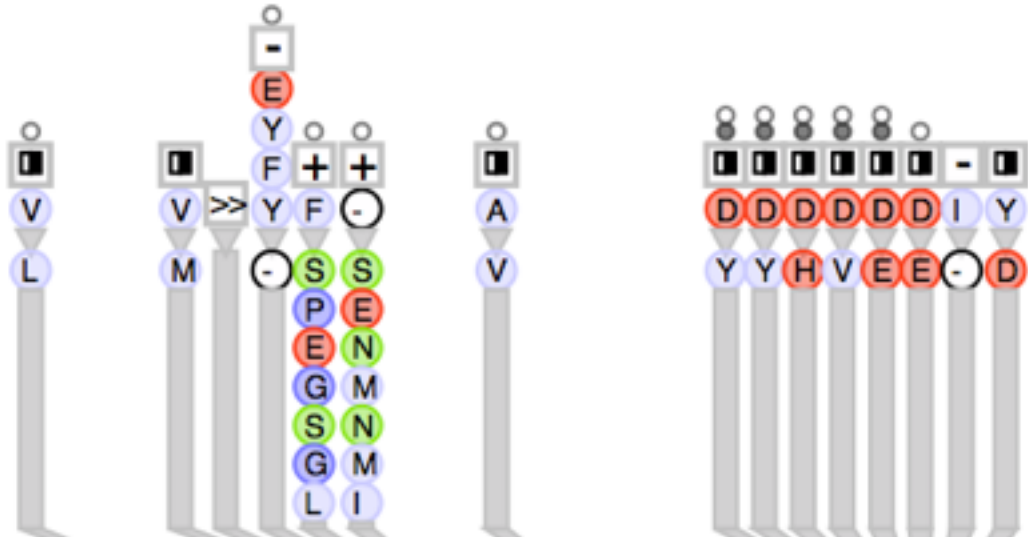


Results

Known leukemia gene: Find fast with sorting metric high score

Variants

Mutation Type
Reference A.A.s
Variant A.A.s



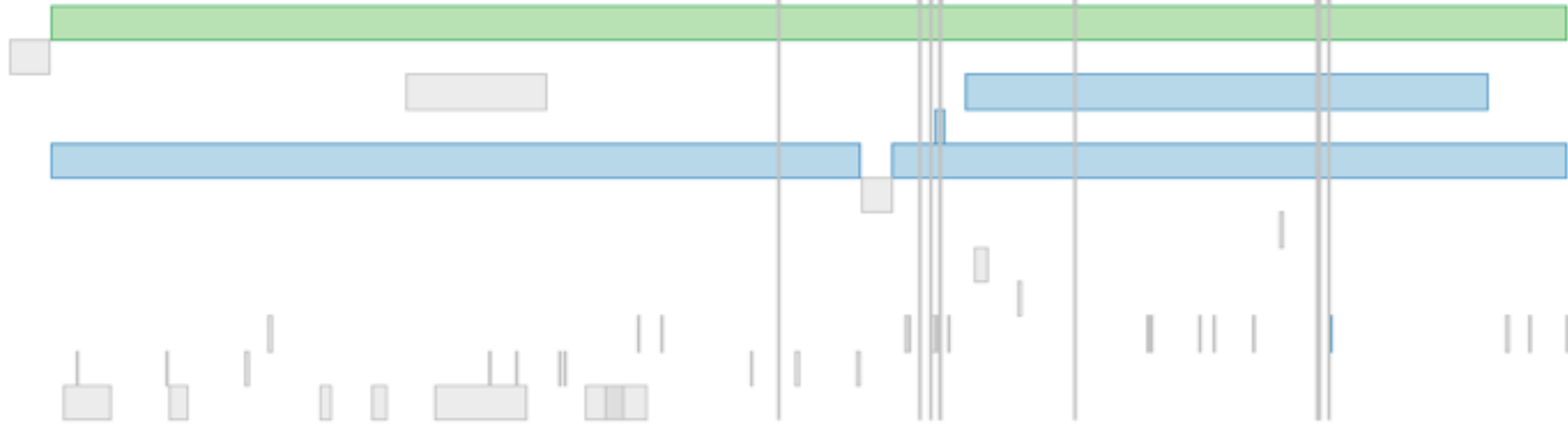
Transcript

trans-anon

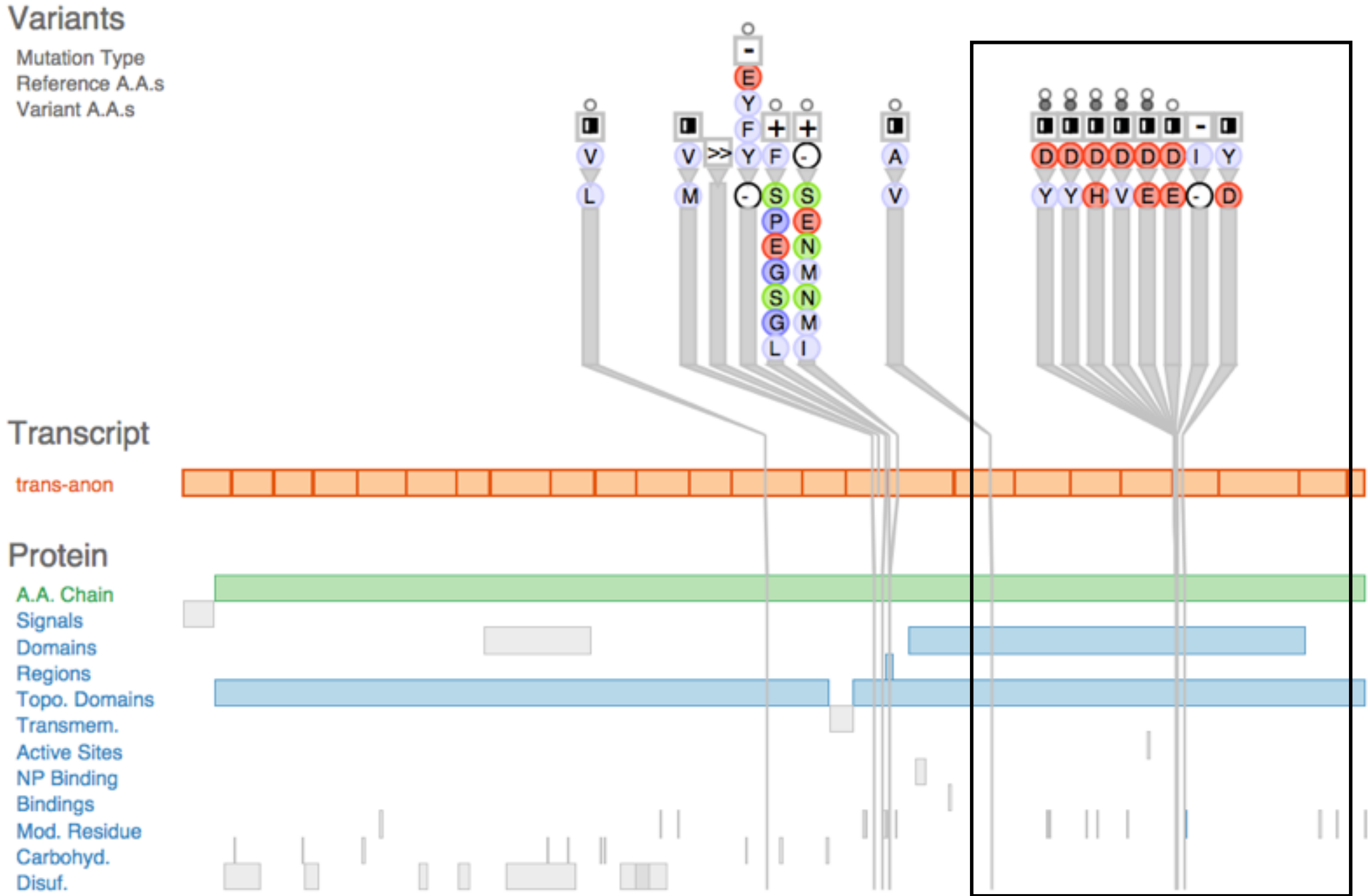


Protein

A.A. Chain
Signals
Domains
Regions
Topo. Domains
Transmem.
Active Sites
NP Binding
Bindings
Mod. Residue
Carbohyd.
Disuf.



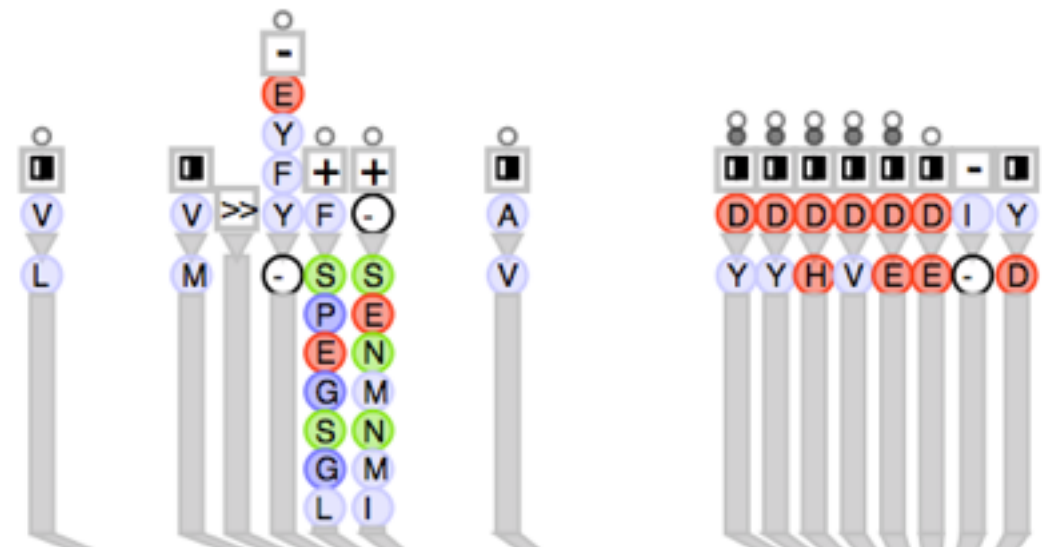
Known leukemia gene: Fanout shows collocation of variants



Known leukemia gene: Several functional protein regions affected

Variants

Mutation Type
Reference A.A.s
Variant A.A.s



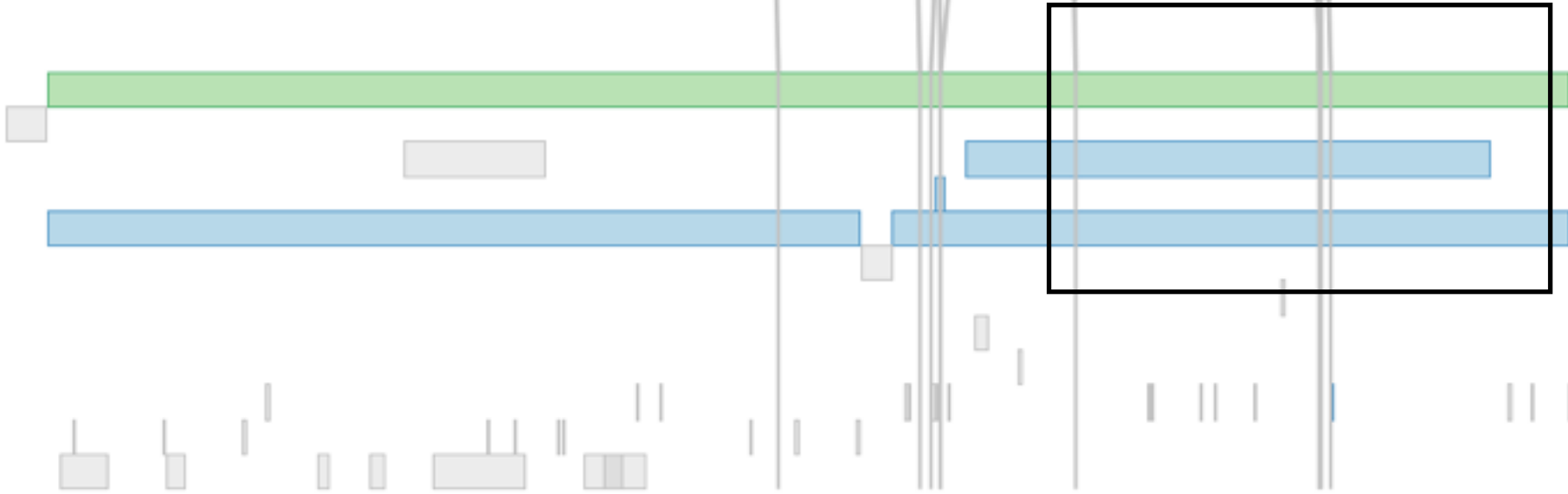
Transcript

trans-anon



Protein

A.A. Chain
Signals
Domains
Regions
Topo. Domains
Transmem.
Active Sites
NP Binding
Bindings
Mod. Residue
Carbohyd.
Disuf.



New finding: Good candidate with high metric score

Variants

Mutation Type
Reference A.A.s
Variant A.A.s



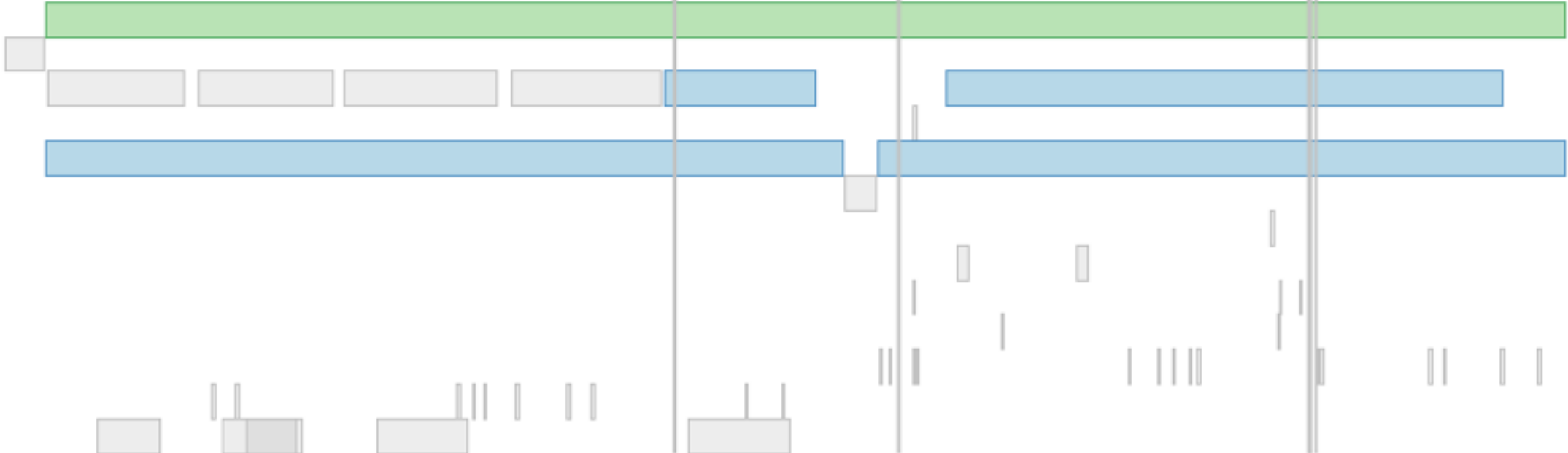
Transcript

trans-anon

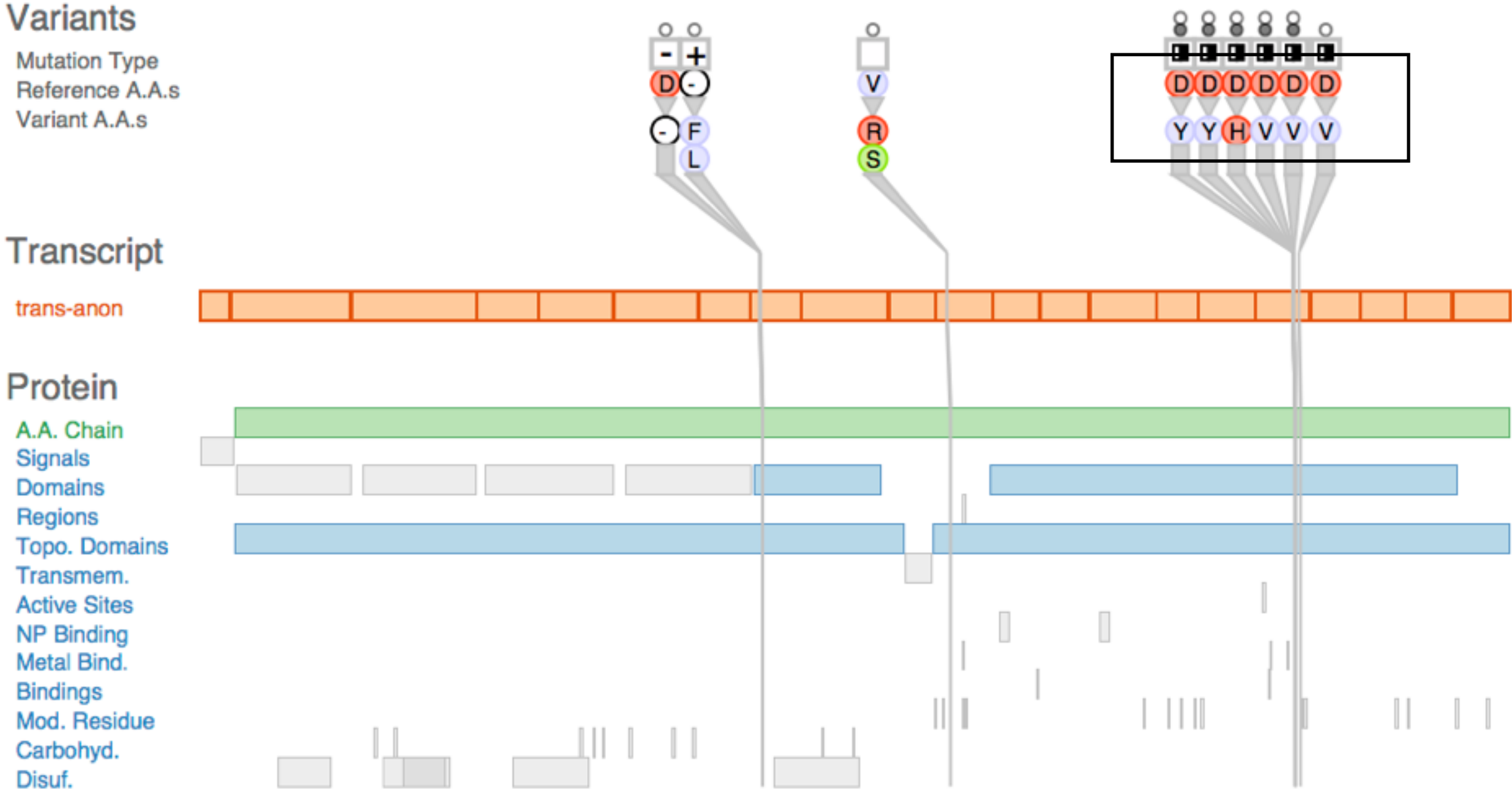


Protein

A.A. Chain
Signals
Domains
Regions
Topo. Domains
Transmem.
Active Sites
NP Binding
Metal Bind.
Bindings
Mod. Residue
Carbohyd.
Disuf.



New finding: Protein chemical class change evident



Low scoring gene: in contrast

Variants

Mutation Type
Reference A.A.s
Variant A.A.s



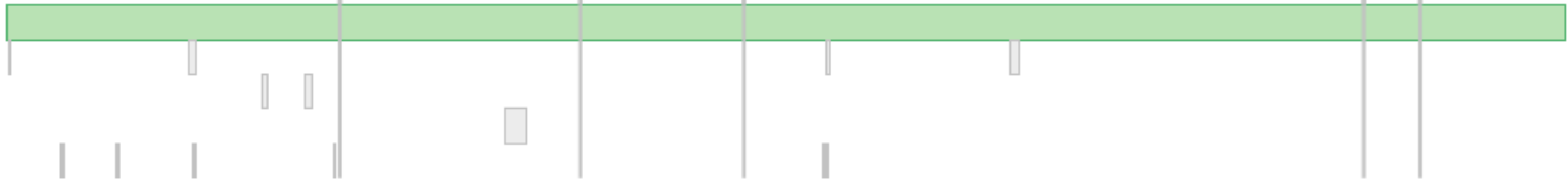
Transcript

trans-anon

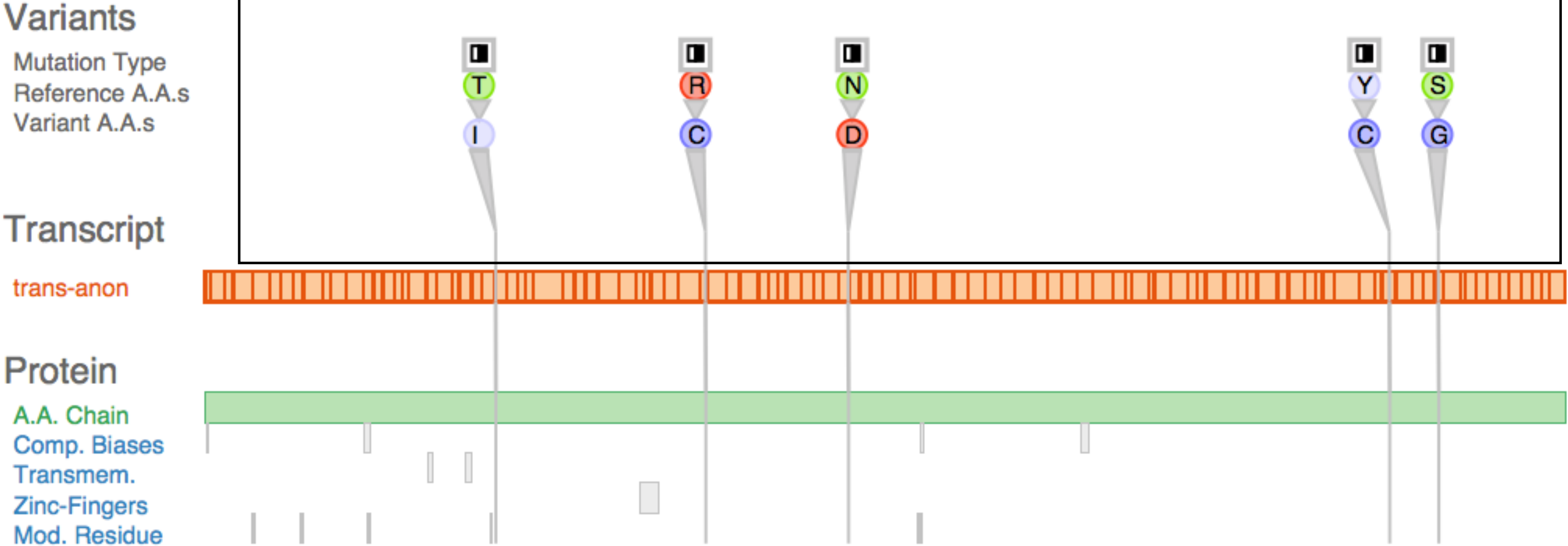


Protein

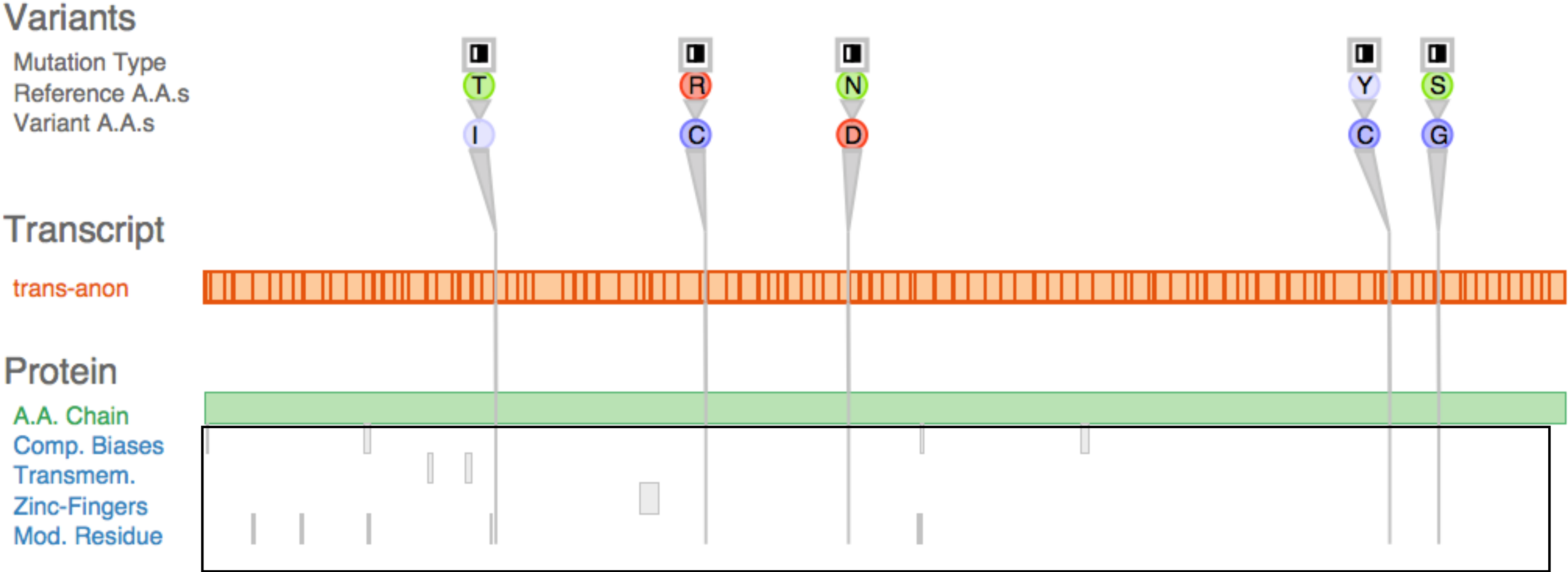
A.A. Chain
Comp. Biases
Transmem.
Zinc-Fingers
Mod. Residue



Low scoring gene: No collocation of variants



Low scoring gene: Mostly unaffected protein regions



Methods

Phase I: Winnow and Cast

5 months



- embedded within GSC for all stages
- winnow stage
 - considered and ruled out many potential collaborators
- cast stage
 - gatekeeper (PI)
 - two front-line analysts (postdocs)



more at:

Design Study Methodology: Reflections from the Trenches and from the Stacks.
Sedlmair, Meyer, Munzner. *IEEE TVCG* 18(12): 2431-2440, 2012 (Proc. InfoVis 2012).

Phase 2: Core Design

5 months



- main task abstraction

- discover gene

- semi-structured interviews

- every week for 1 hr

- iterative refinement

- 8 data sketches deployed

- rapid prototyping to show real data ASAP

- refine utility & usability



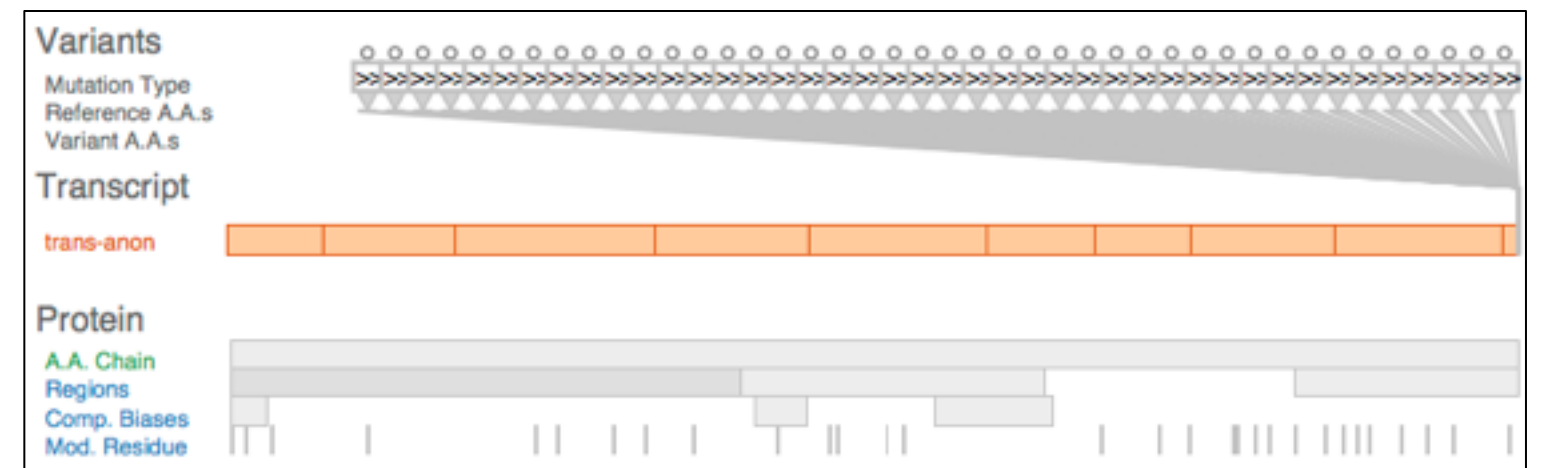
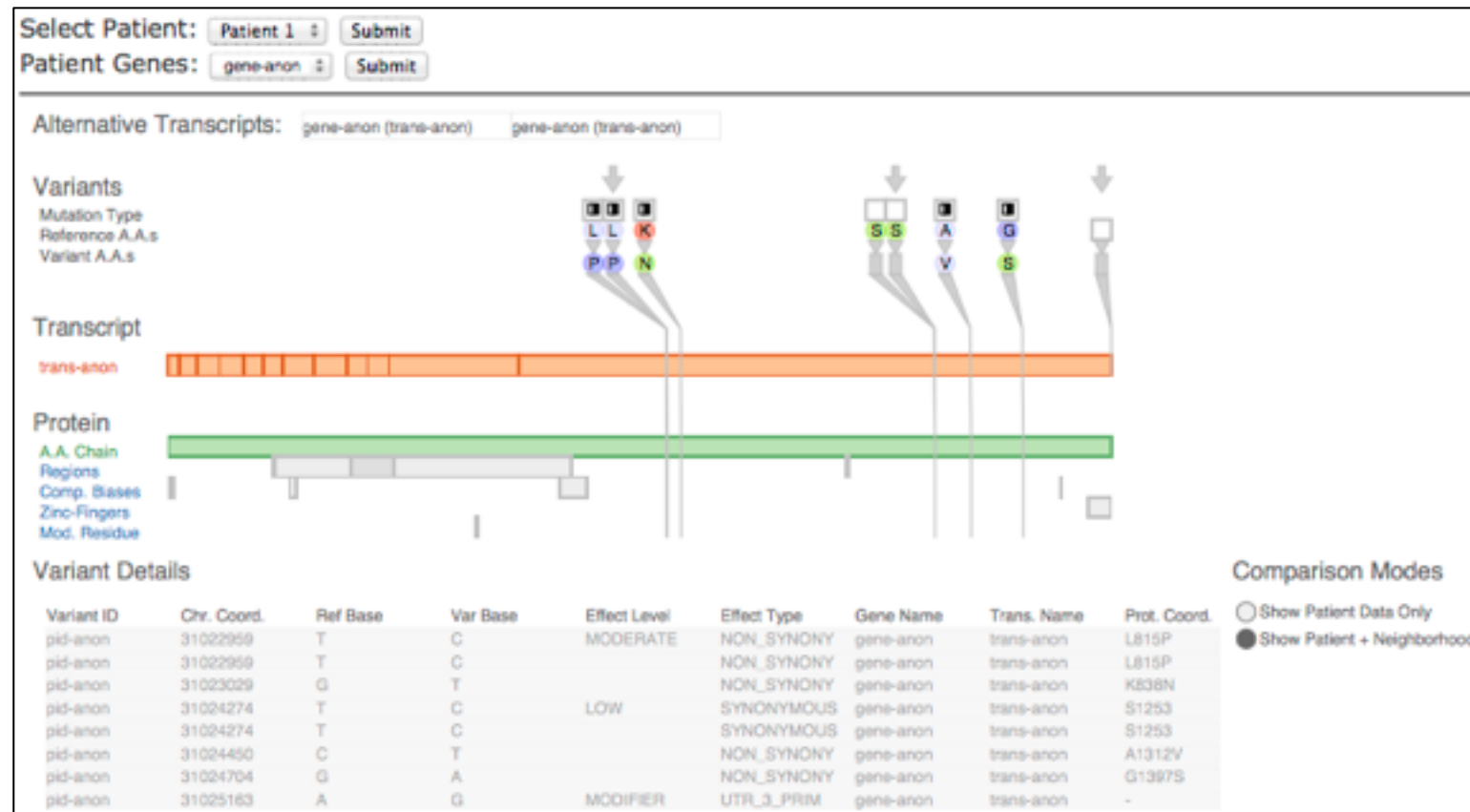
Human-centered approaches in geovisualization design:
investigating multiple methods through a long-term case study.
Lloyd and Dykes. IEEE TVCG (Proc. InfoVis), 17(12):2498–2507, 2011.

Phase 3: Two More Tasks

1 month



- two new analysts
 - connected by enthusiastic gatekeeper
- new task abstractions
 - compare patients
 - debug pipeline
- transferrable with minimal changes



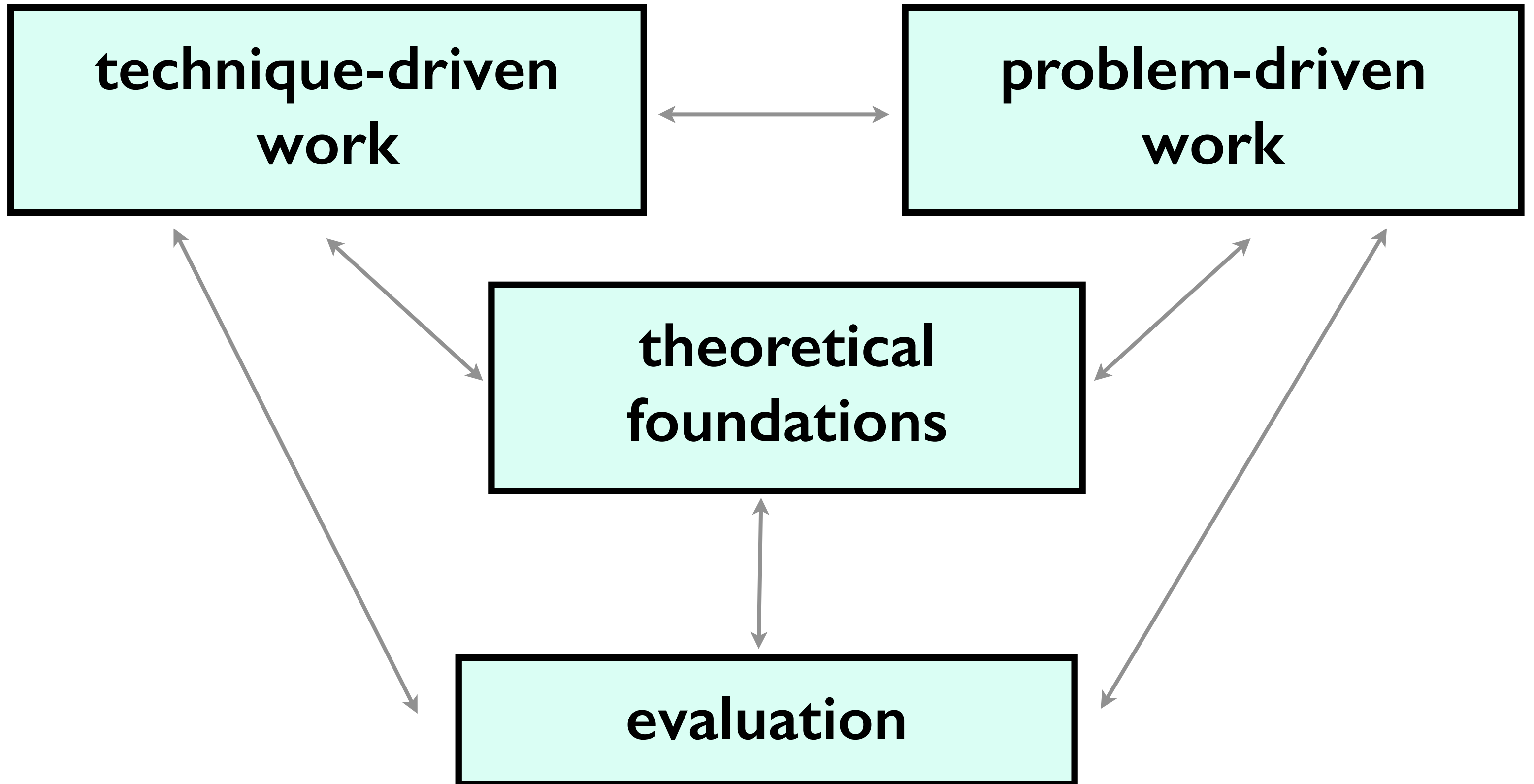
Phase 4: Reflect and write

3 months



- abstraction innovation
 - data abstraction: highly filtered *transcript coordinates* (vs genome coordinates)
- guidelines
 - specialize first, generalize later
 - good for domains with complex data
 - high-level considerations
 - identifying scales of interest
 - what to visually encode directly vs what to support through interaction
 - when (and how) to eliminate navigation

A quick taste of other work!



Problem-driven: Genomics

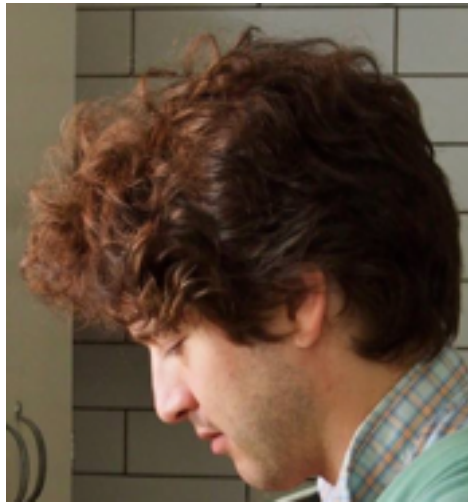
T

P

F

E

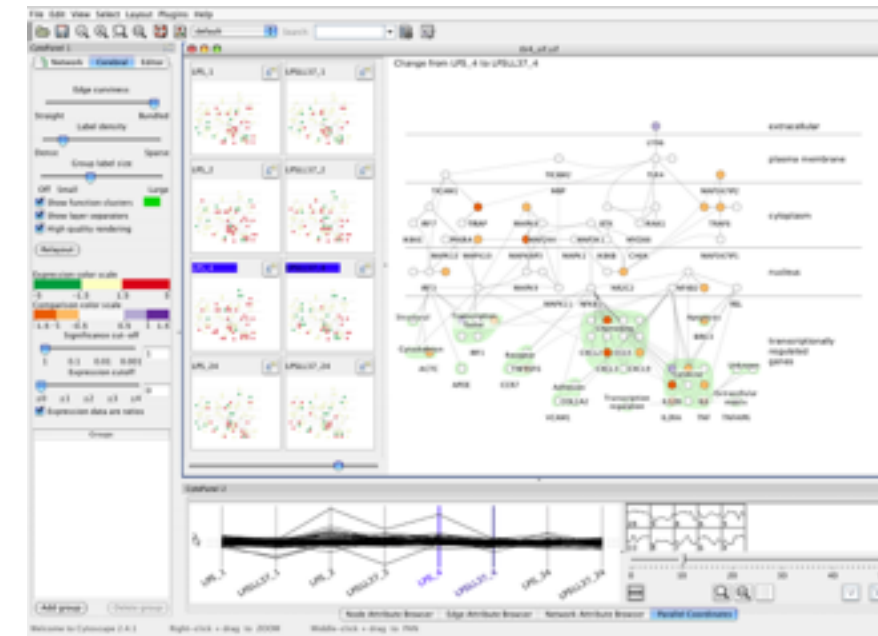
Aaron Barsky



Jenn Gardy
(Microbio)



Robert Kincaid
(Agilent)

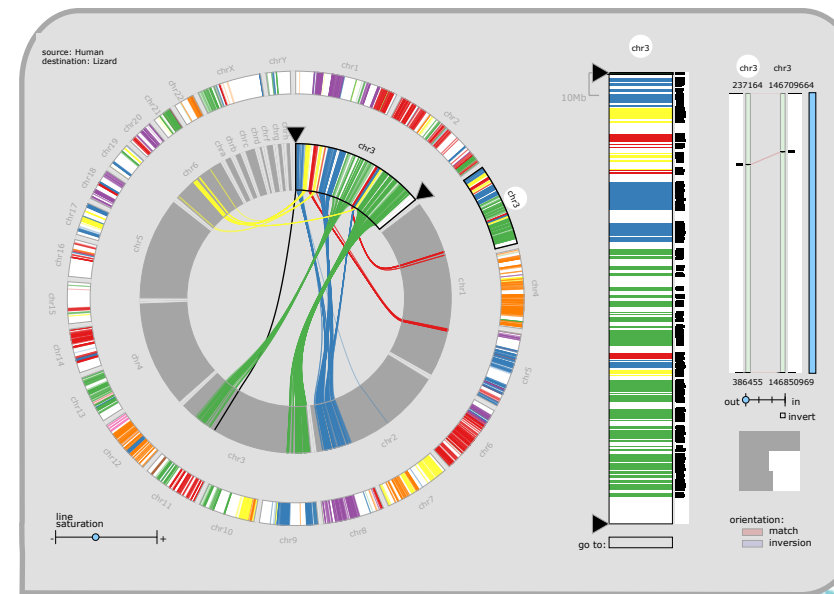


Cerebral

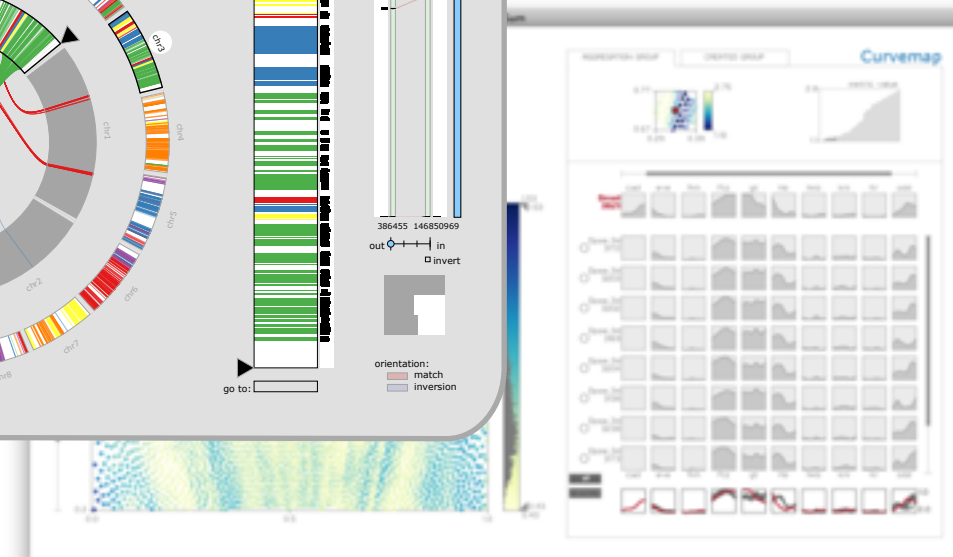
Miriah Meyer



Hanspeter Pfister
(Harvard)



MizBee



MulteeSum, Pathline

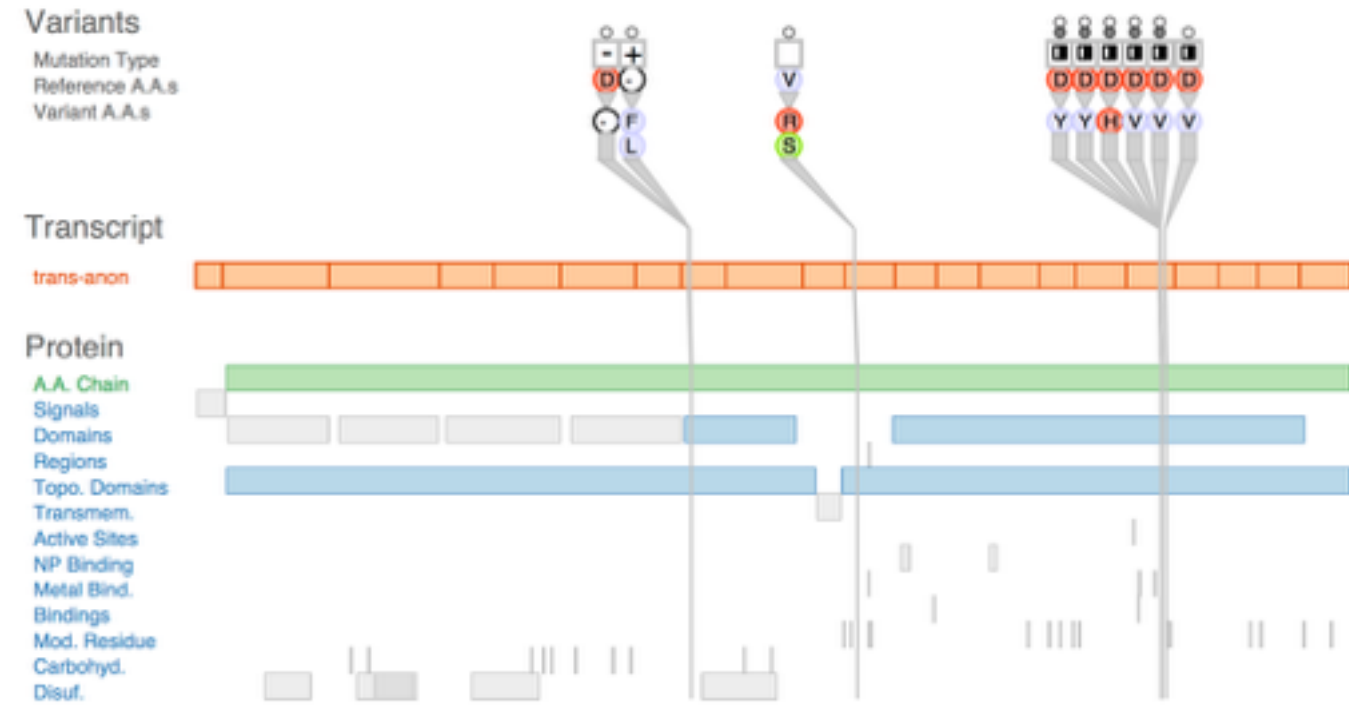
Problem-driven: Genomics, fisheries

T
F
E
P

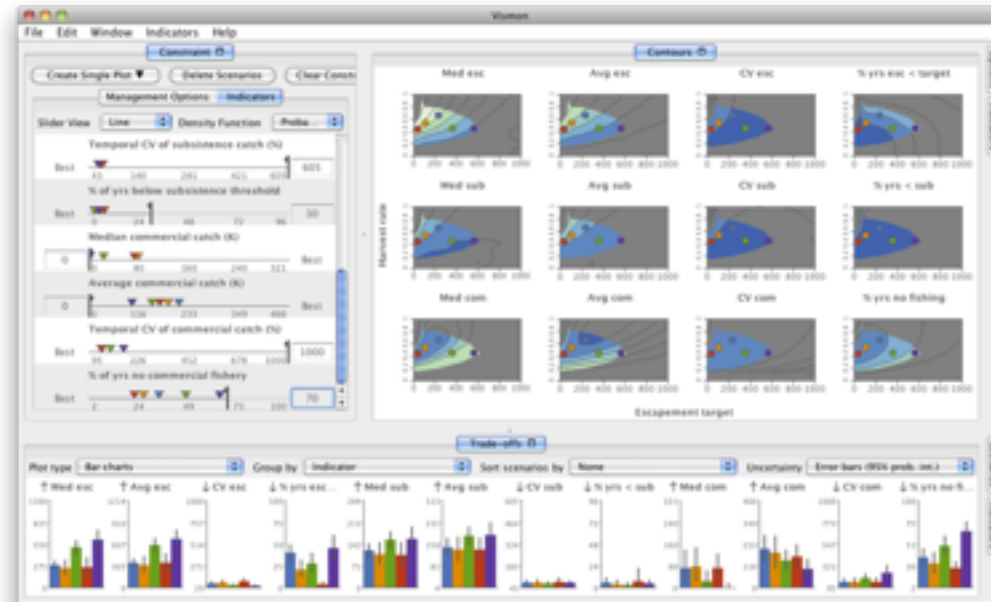
Joel Ferstay



Cydney Nielsen
(BC Cancer)



Variant View



Vismon

Maryam Booshehrian

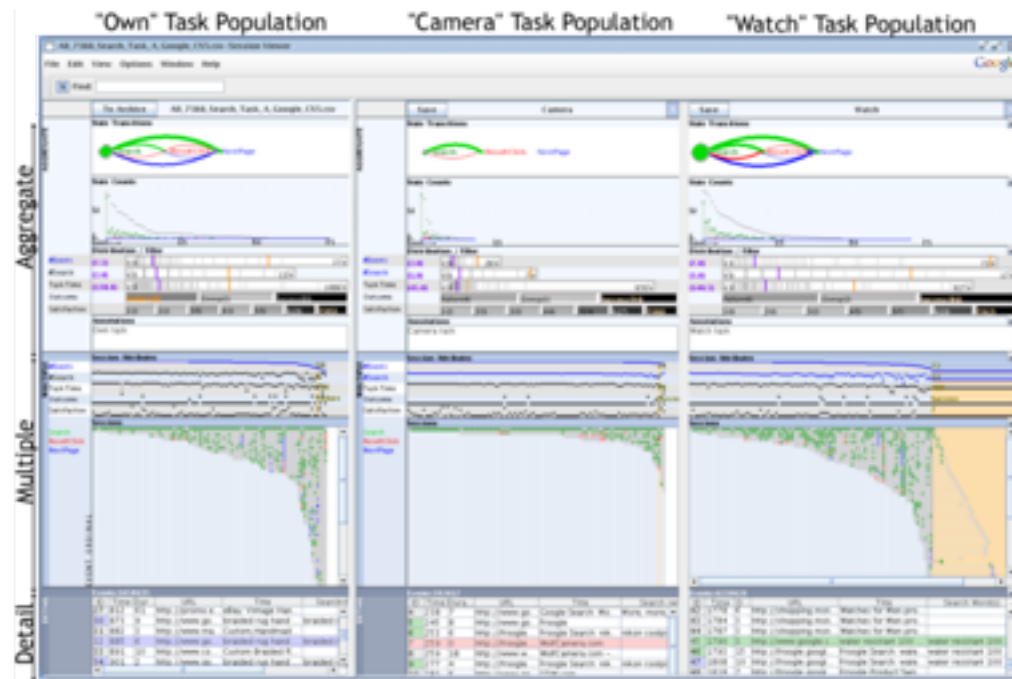


Torsten Moeller
(SFU)



Problem-driven: Tech industry

T F P
F
E



SessionViewer: web log analysis

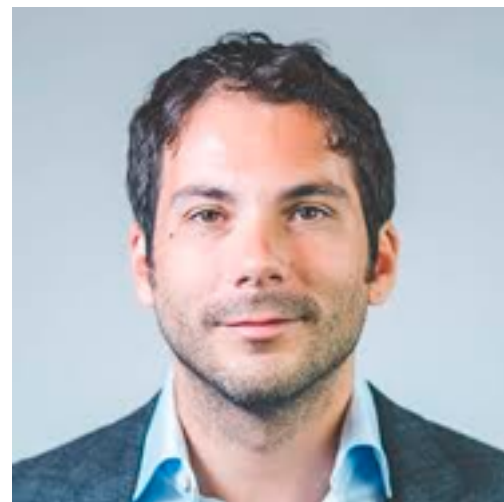
Heidi Lam



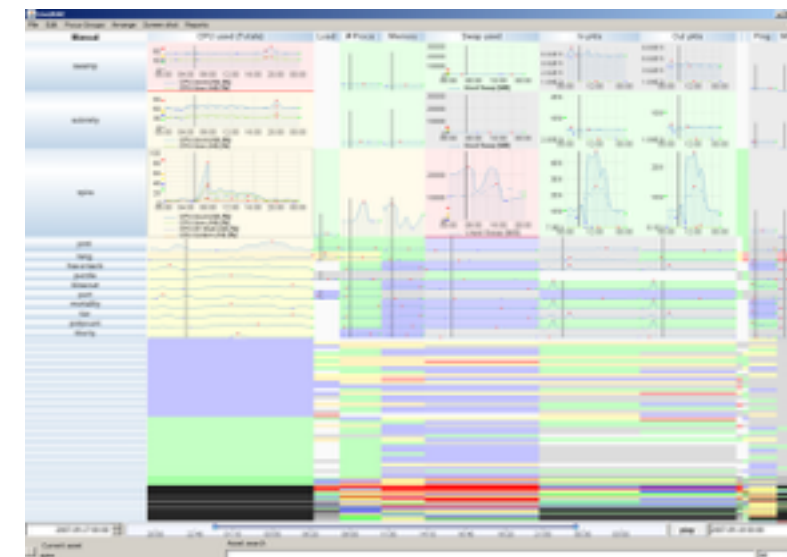
**Diane Tang
(Google)**



Peter McLachlan



**Stephen North
(AT&T Research)**



LiveRAC: systems time-series

Problem-driven: Journalism

T

P

F

E

Matt Brehmer



Stephen Ingram



**Jonathan Stray
(Assoc Press)**



Overview

Technique-driven: Dimensionality reduction

T

P

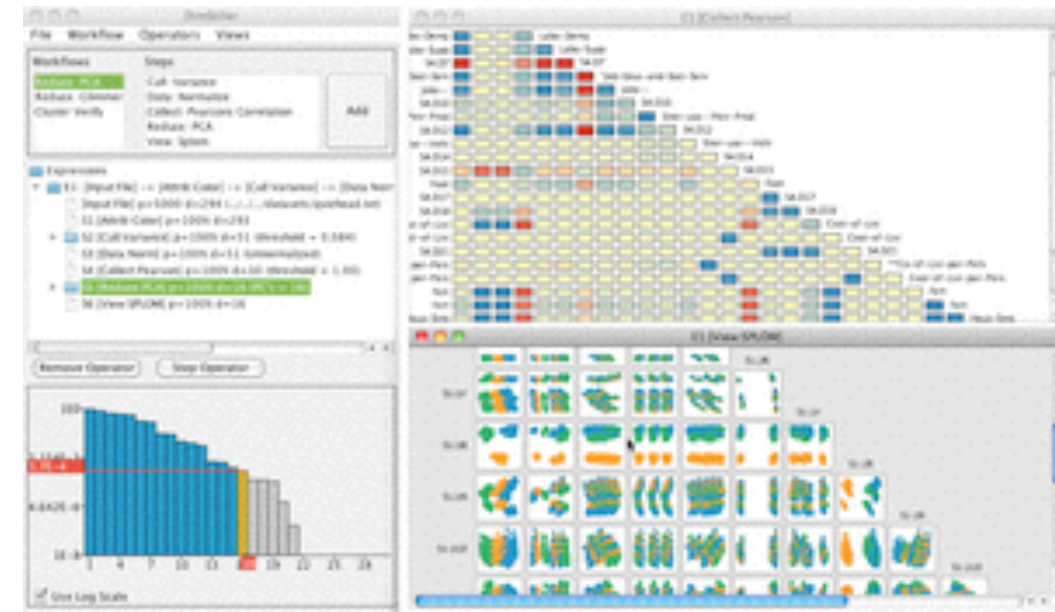
F

E

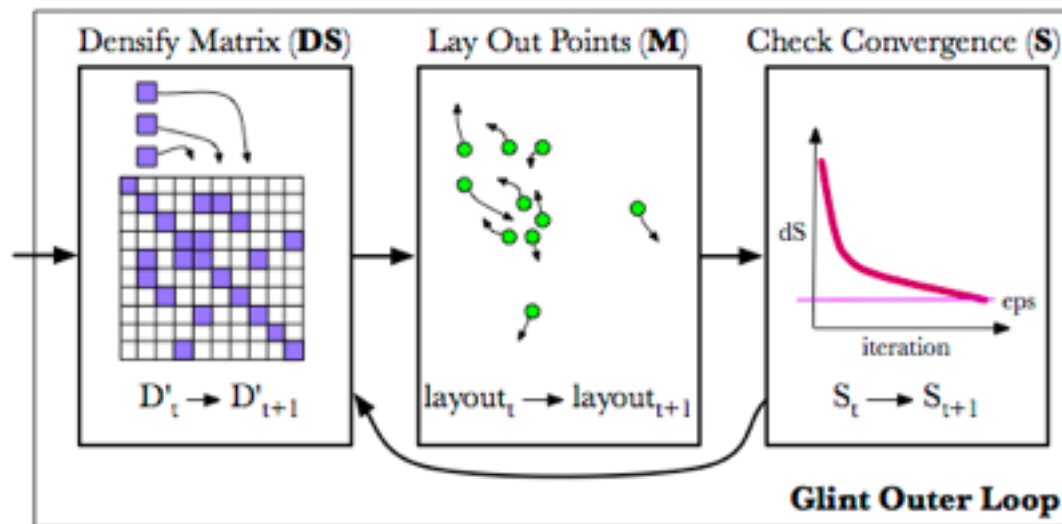
Stephen Ingram



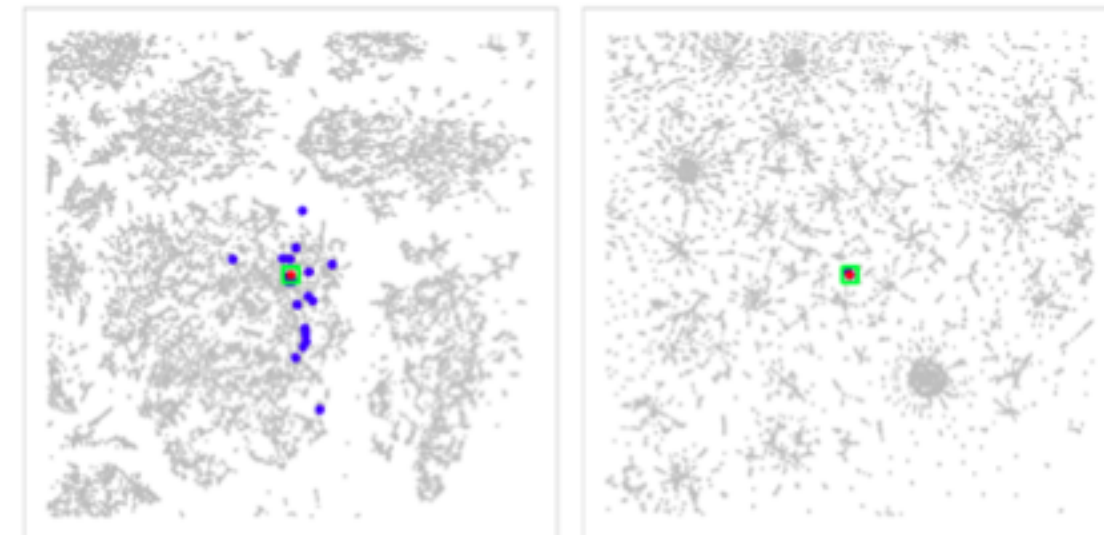
Glimmer



DimStiller



Glint



QSNE

Evaluation: Dimensionality reduction

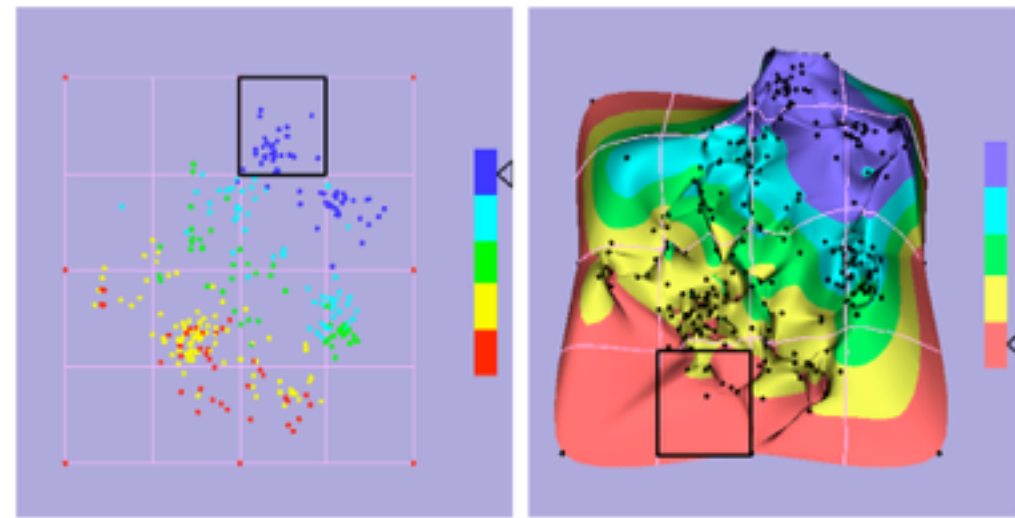
T

P

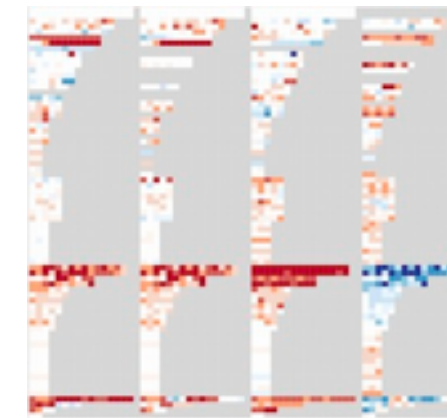
F

E

Melanie Tory



Points vs landscapes for dimensionally reduced data



Guidance on DR & scatterplot choices

Michael Sedlmair



Melanie Tory (UVic)



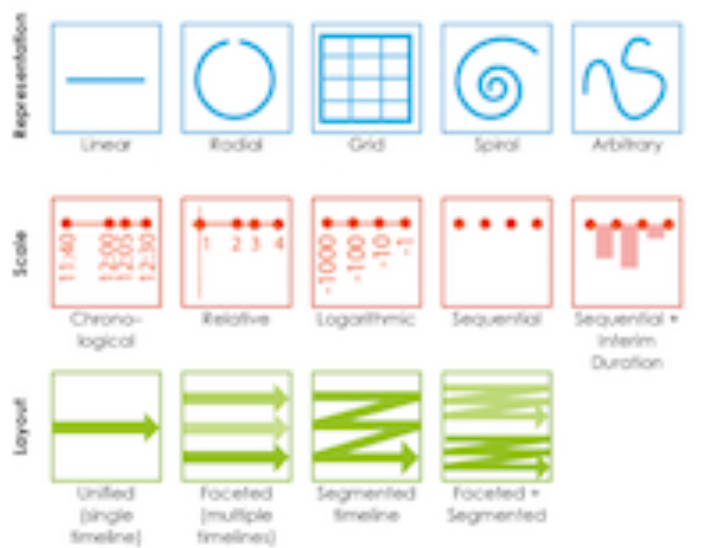
Taxonomy of cluster separation factors

Curation & Presentation: Timelines

T F E P



TimeLineCurator
<https://vimeo.com/123246662>



Timelines Revisited
timelinesrevisited.github.io/

Johanna Fulda
 (Sud. Zeitung)



Matt Brehmer



Bongshin Lee
 (Microsoft)



Benjamin Bach
 (Microsoft)



Nathalie Henry-Riche
 (Microsoft)



Theoretical foundations

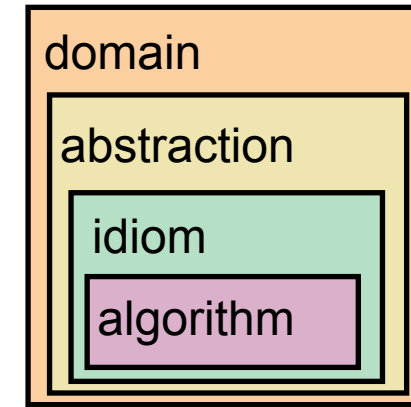
T F P
E

- Visual Encoding Pitfalls

- Unjustified Visual Encoding
- Hammer In Search Of Nail
- 2D Good, 3D Better
- Color Cacophony
- Rainbows Just Like In The Sky

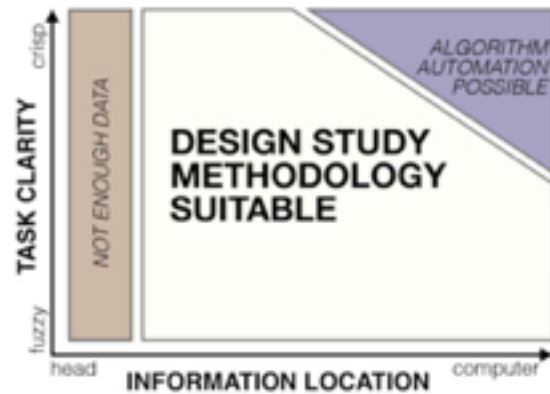
- Strategy Pitfalls

- What I Did Over My Summer
- Least Publishable Unit
- Dense As Plutonium
- Bad Slice and Dice



Nested Model

Papers Process & Pitfalls



Design Study Methodology

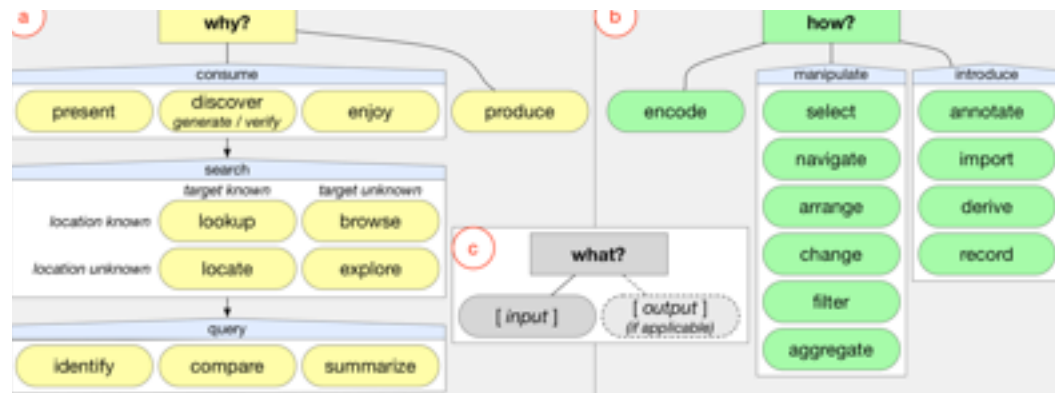
Michael Sedlmair



Miriah Meyer



Matt Brehmer



Abstract Tasks

Geometry Center 1990-1995: Expository videos



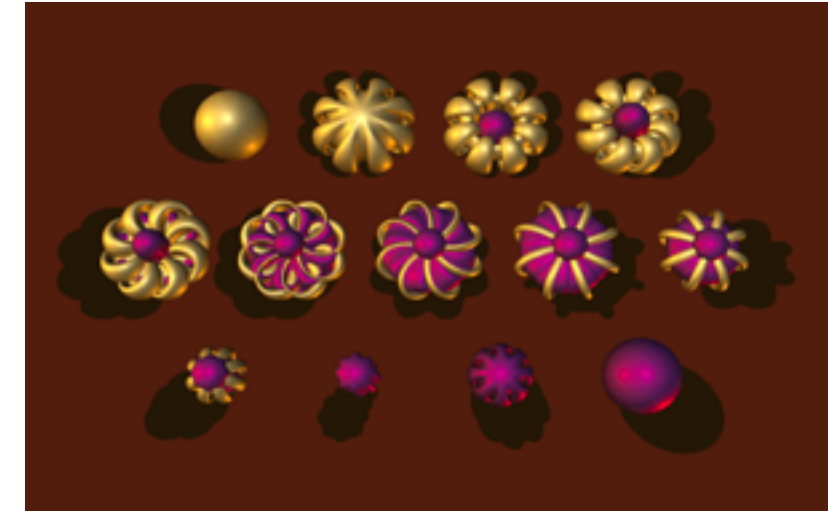
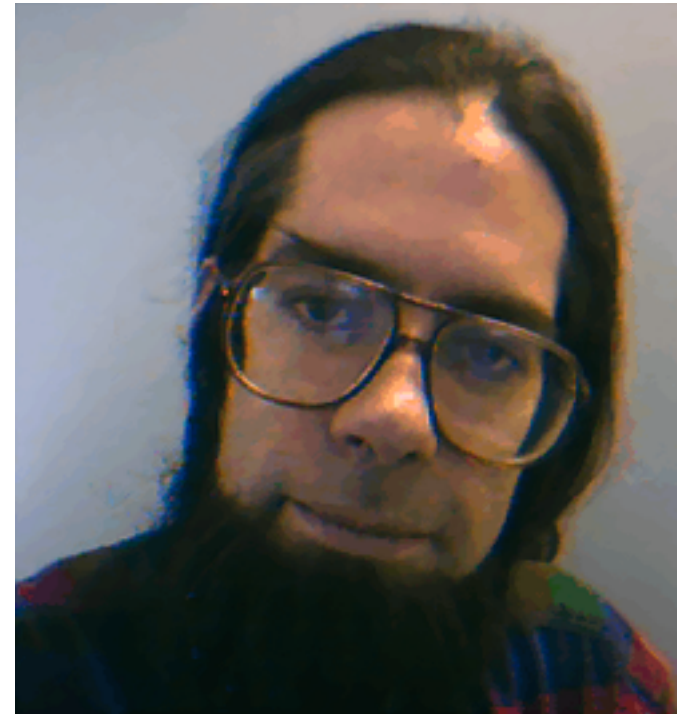
Geomview

Charlie Gunn



The Shape of Space

Stuart Levy



Outside In

Mark Phillips



Delle Maxwell



Wrap-up

- models and methods for design and validation
 - collaboration incentives for vis and bio
- example biovis project
 - Variant View
- methodological dream:
user-centered design spreading from vis to biovis to bioinformatics
 - task/requirements analysis for *all* tools, not just visual ones
 - focus on both utility and usability

More information

[@tamaramunzner](https://twitter.com/tamaramunzner)

- this talk

<http://www.cs.ubc.ca/~tmm/talks.html#caulfield17>

- papers, videos, software, talks, courses

<http://www.cs.ubc.ca/group/infovis>

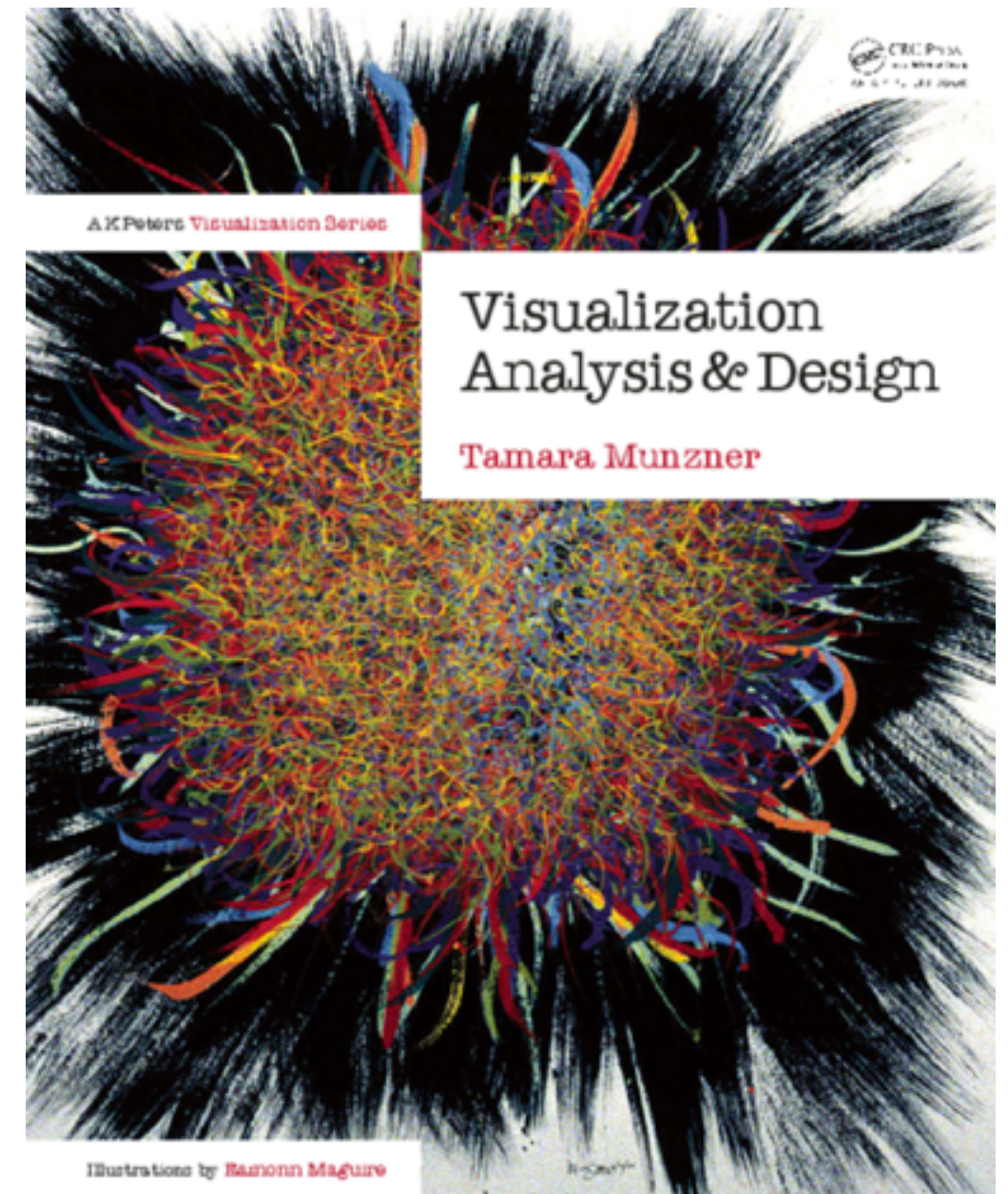
<http://www.cs.ubc.ca/~tmm>

- theoretical foundations: book
(+ free tutorial/course lecture slides)

<http://www.cs.ubc.ca/~tmm/vadbook>

– 20% promo code for book+ebook combo: HVN17

– <http://www.crcpress.com/product/isbn/9781466508910>



Visualization Analysis and Design.
Munzner. A K Peters Visualization Series, CRC Press, Visualization Series, 2014.