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

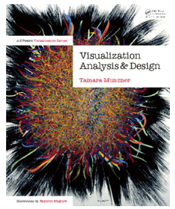
www.cs.ubc.ca/~tmm/talks.html#caulfield17 @tamaramunzner

Visualization (vis) defined & motivated

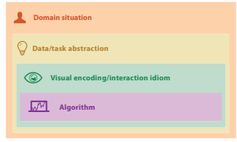
Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

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

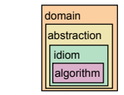
http://www.cs.ubc.ca/labs/imager/tr/2009/NestedModel

Tamara Munzner
@tamaramunzner

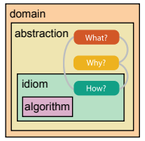
A Nested Model for Visualization Design and Validation.
Munzner. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 09), 15(6):921-928, 2009.

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).*]

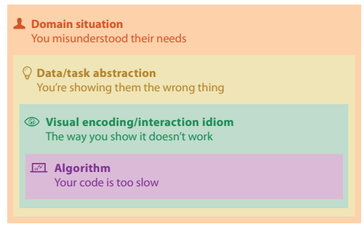


[A Multi-Level Typology of Abstract Visualization Tasks *Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).*]

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

Why is validation difficult?

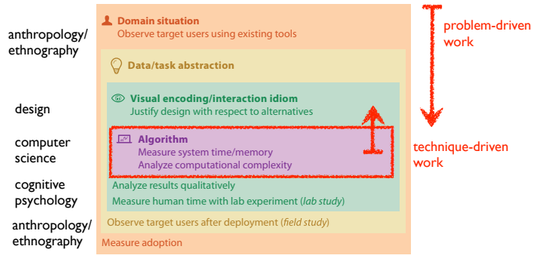
- different ways to get it wrong at each level



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

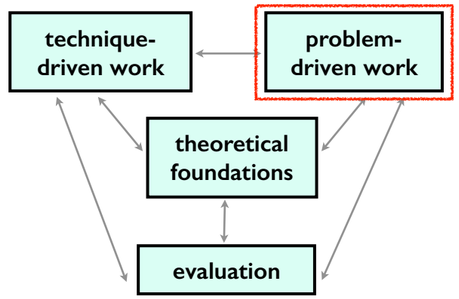
Validation solution: use methods from appropriate fields at each level

- avoid mismatches!



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

Angles of attack: My own research agenda

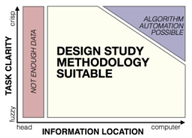


Vis meets bio

- 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

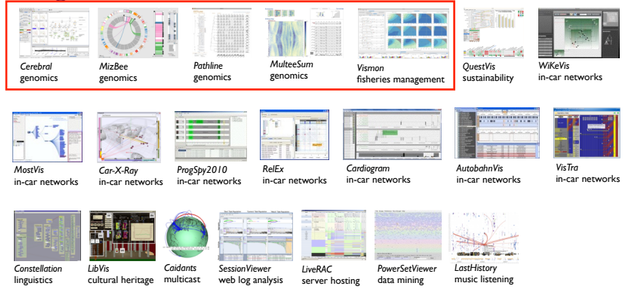
Reflections from the Trenches and from the Stacks

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

Tamara Munzner
@tamaramunzner

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

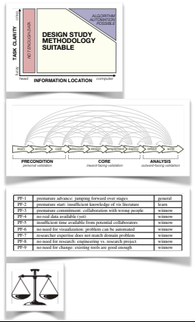
Design Studies: Lessons learned after 21 of them



- commonality of representations cross-cuts domains!

Methodology for problem-driven work

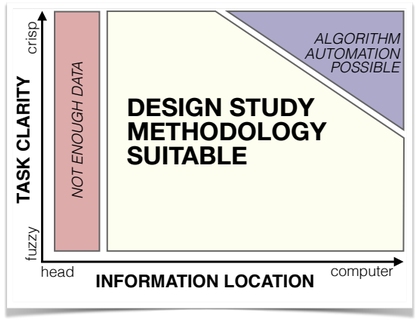
- definitions
- 9-stage framework
- 32 pitfalls & how to avoid them
- comparison to related methodologies



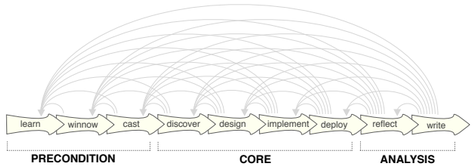
Design study methodology: definitions

- 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

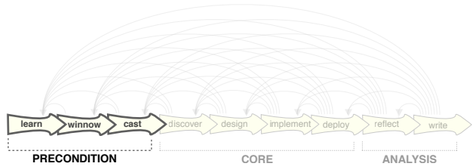


9 stage framework



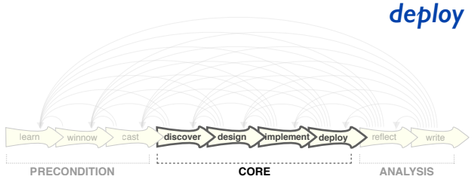
9-stage framework

learn
winnow
cast



9-stage framework

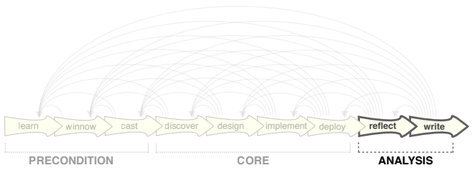
discover
design
implement
deploy



9-stage framework

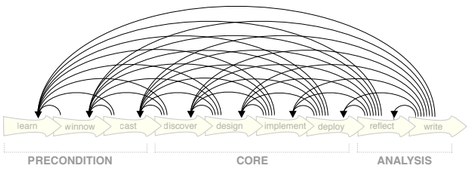
reflect
write

- guidelines: confirm, refine, reject, propose



9-stage framework

iterative



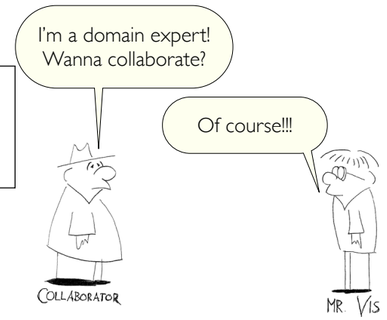
Design study methodology: 32 pitfalls

- 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

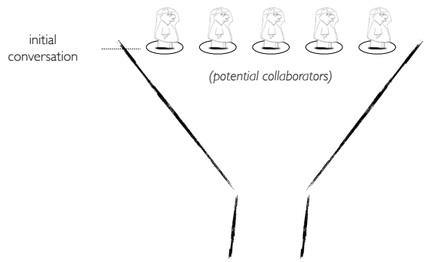
PITFALL

PREMATURE COLLABORATION COMMITMENT

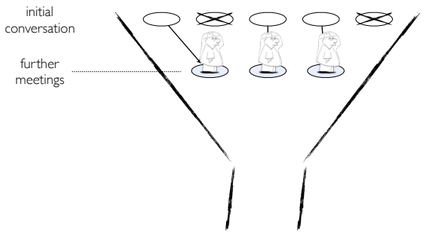


METAPHOR
Winnowing

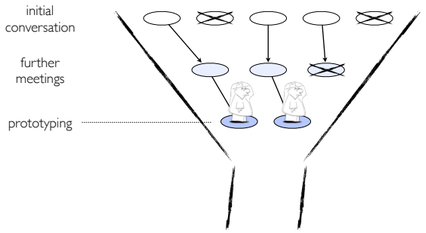
Collaborator winnowing



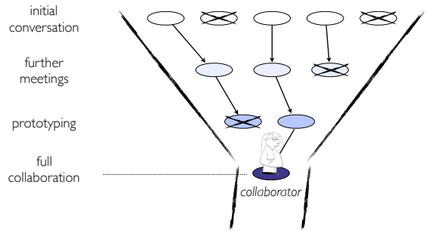
Collaborator winnowing



Collaborator winnowing



Collaborator winnowing



Collaborator winnowing

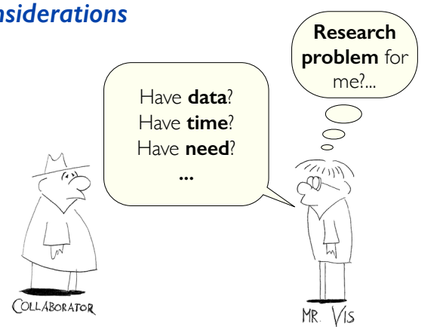
Talk with many, stay with few!

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

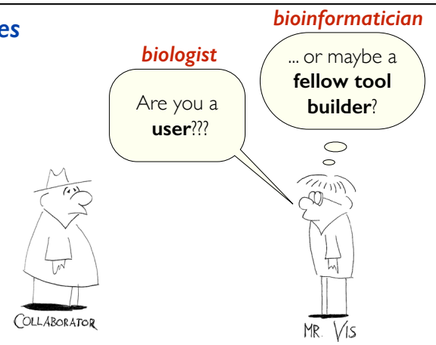
considerations



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 *just talking* or *fly on wall* 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

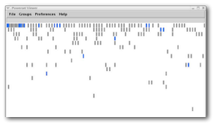
roles



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



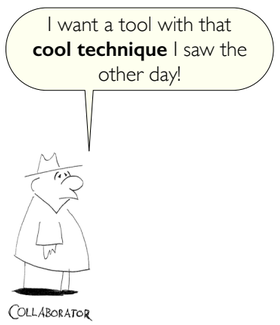
WikeVis
0.5 years / 2 researchers

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 *just talking* or *fly on wall* 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

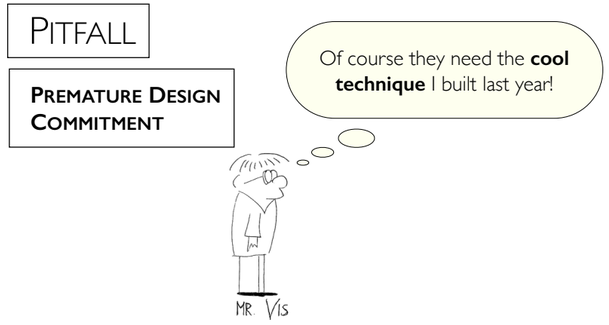
PITFALL

PREMATURE DESIGN COMMITMENT

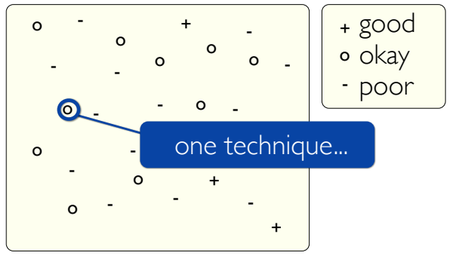


PITFALL

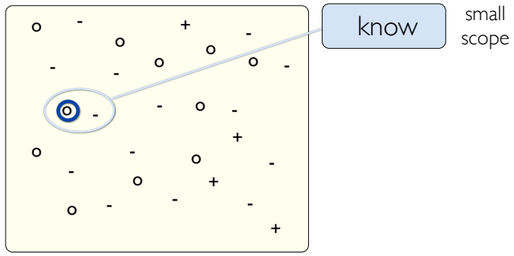
PREMATURE DESIGN COMMITMENT



METAPHOR
Design Space



METAPHOR
Design Space

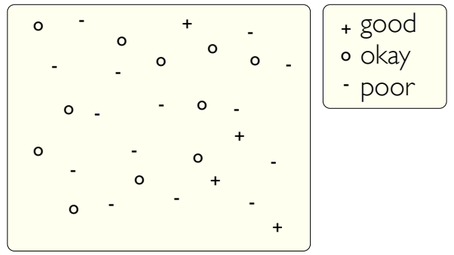


Design study methodology: 32 pitfalls

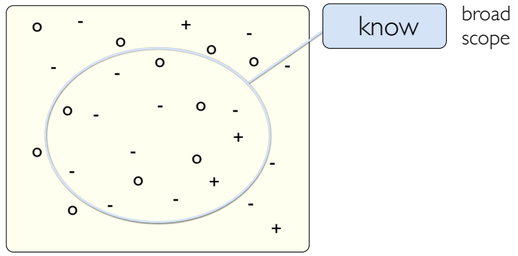
- and how to avoid them

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

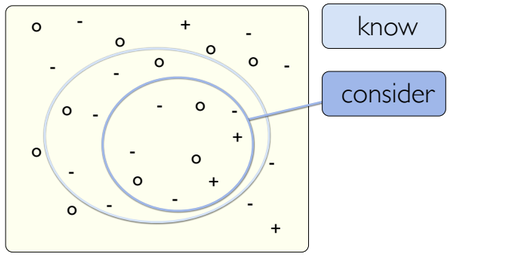
METAPHOR
Design Space



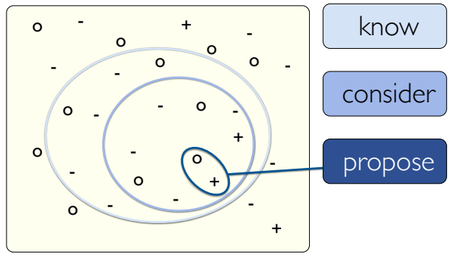
METAPHOR
Design Space



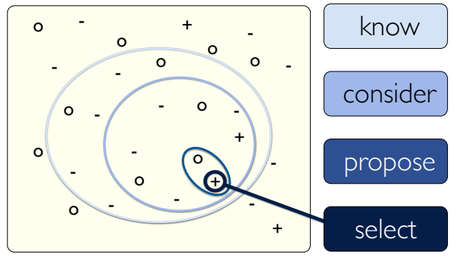
METAPHOR
Design Space



METAPHOR
Design Space



METAPHOR
Design Space



METAPHOR
Design Space

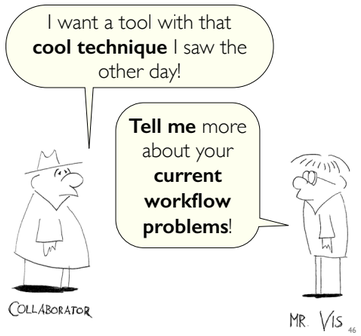
Design study methodology: 32 pitfalls

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 *just talking* or *fly on wall* 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

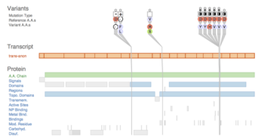
PREMATURE DESIGN COMMITMENT

DOMAIN EXPERTS FOCUSED ON VIS DESIGN VS DOMAIN PROBLEM



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	*liking* 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 ≠ 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



Variant View

Visualizing Sequence Variants in their Gene Context

Joel Ferstay

Cydney Nielsen
@cydneybn

Tamara Munzner
@tamaramunzner

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

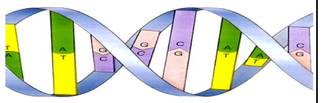
Variant View: Visualizing Sequence Variants in their Gene Context.
Forstay, Nielsen, Munzner. IEEE TVCG 19(12): 2546-2555, 2013 (Proc. InfoVis 2013).

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



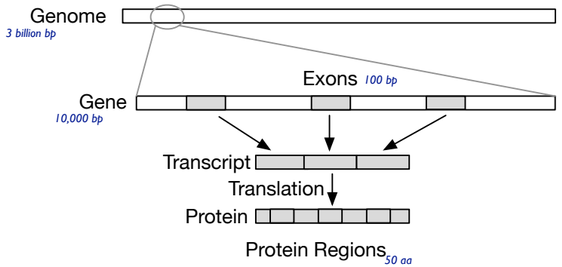
Reference Genome DNA:	ATA TGA TCA ACA CTT	
Sample 1 Genome DNA:	ATA TGG TCA ATA CTT	Harmful?
Sample 2 Genome DNA:	ATA TGA TGA ACA CCT	Harmless?

Cancer Research

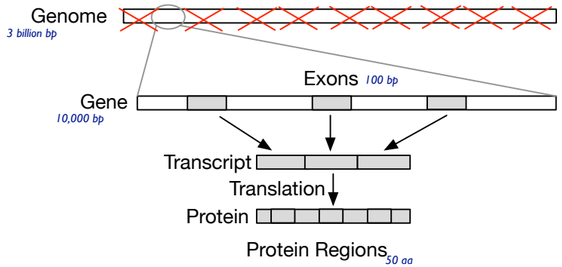
- collaboration with analysts at BC Genome Sciences Center
  - –studying genetic basis of leukemia
- driving task
  - –discover new candidate genes with harmful variants
- two big questions
  - –what to show
    - data abstraction
    - challenge: enormous range of scales in the data
  - –how to show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

Abstractions

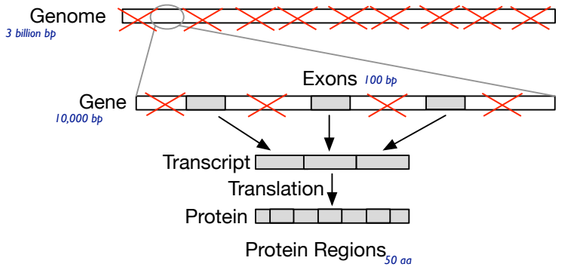
Data: Filtering to relevant biological levels and scales



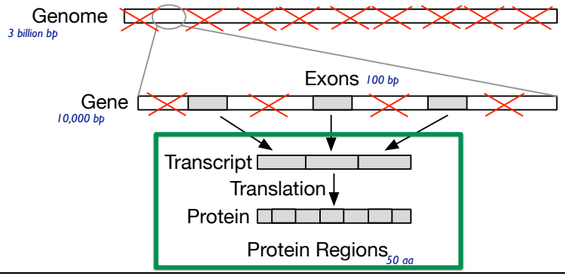
Filter out whole genome; keep genes



Filter out non-exon regions

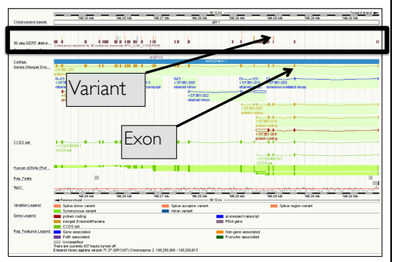


Data abstraction: highly filtered scope of *transcript coordinates*



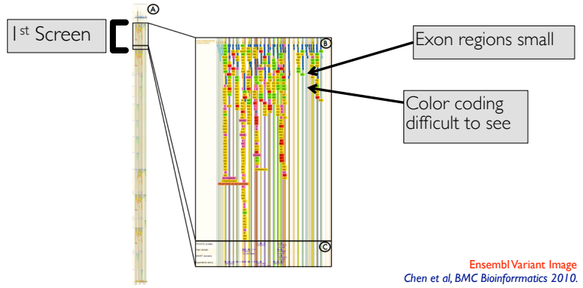
Dominant paradigm: genome browsers

- 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

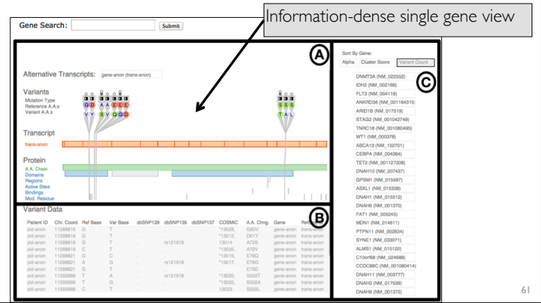


Ensembl Variant Image
Chen et al, BMC Bioinformatics 2010.

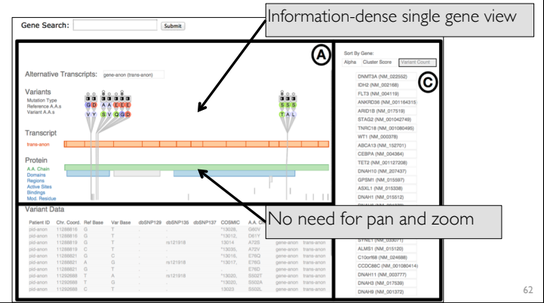
Idioms

Variant View

Variant View

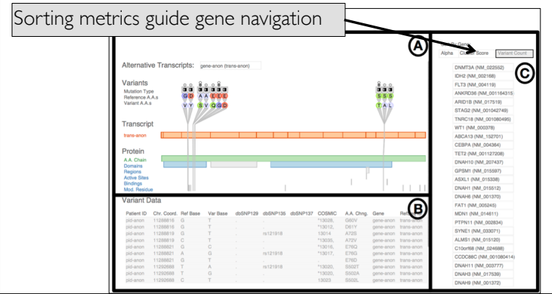


Variant View

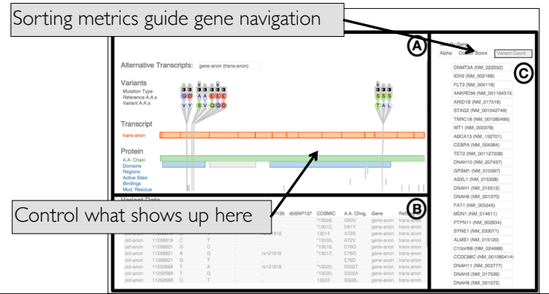


Variant View

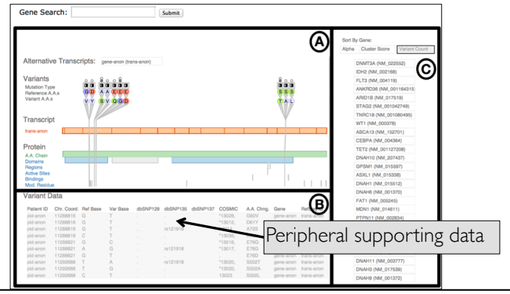
derived data guides human-in-the-loop analytics



Variant View

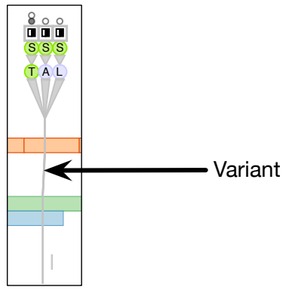


Variant View

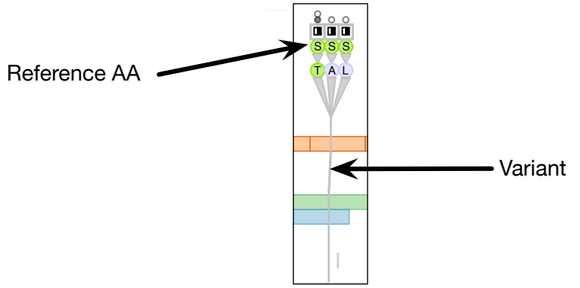


Design information-dense visual encoding

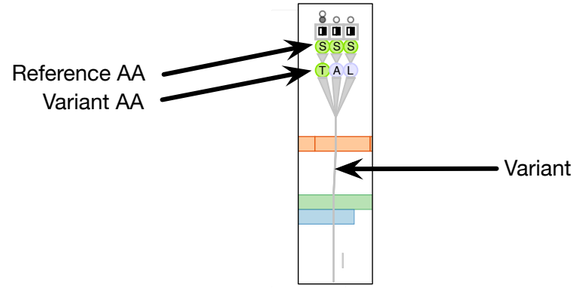
- show all attributes necessary for variant analysis
  - –match salience with importance for analysis task
- variant not just a thin line!
- emphasize with high salience
  - –collocated variants fan out at top
  - –grey variant vertical stroke intersects horizontal colored protein regions



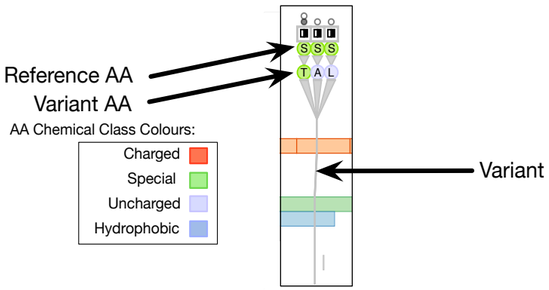
Design information-dense visual encoding



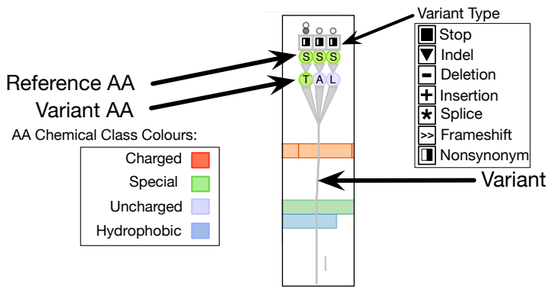
Design information-dense visual encoding



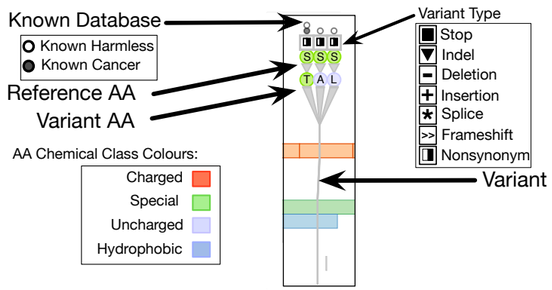
Design information-dense visual encoding



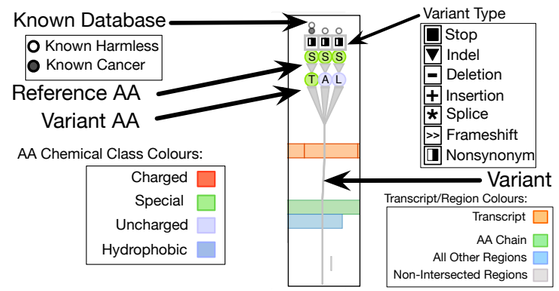
Design information-dense visual encoding



Design information-dense visual encoding

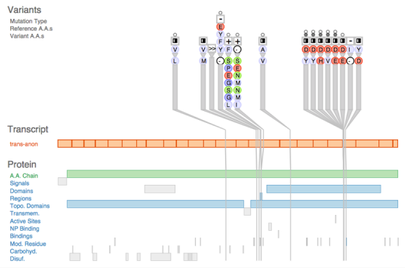


Design information-dense visual encoding

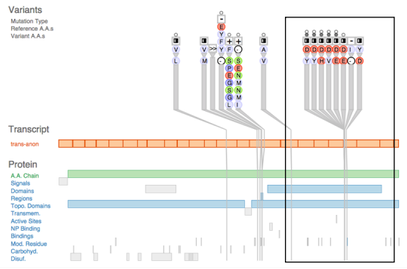


Results

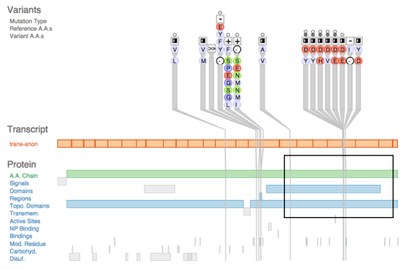
Known leukemia gene: Find fast with sorting metric high score



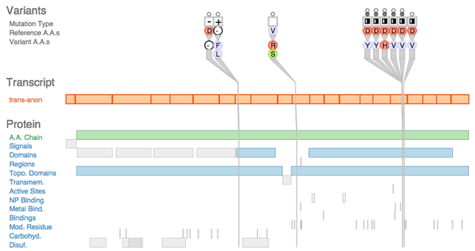
Known leukemia gene: Fanout shows collocation of variants



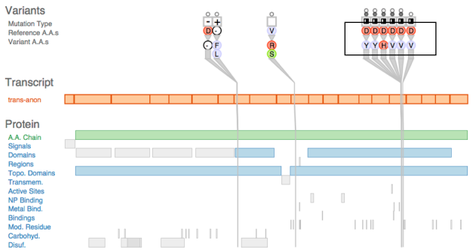
Known leukemia gene: Several functional protein regions affected



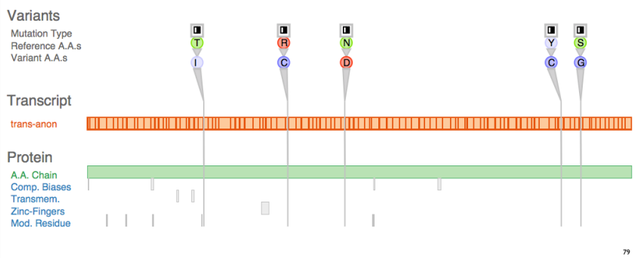
New finding: Good candidate with high metric score



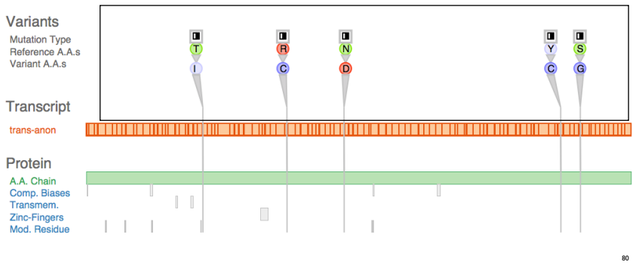
New finding: Protein chemical class change evident



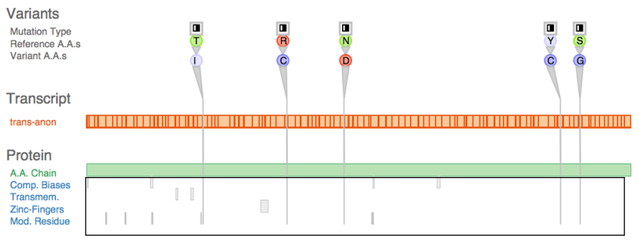
Low scoring gene: in contrast



Low scoring gene: No collocation of variants



Low scoring gene: Mostly unaffected protein regions



Methods

Phase 1: Winnow and Cast

5 months

learn > winnow > cast > discover > design > implement > deploy > reflect > write

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

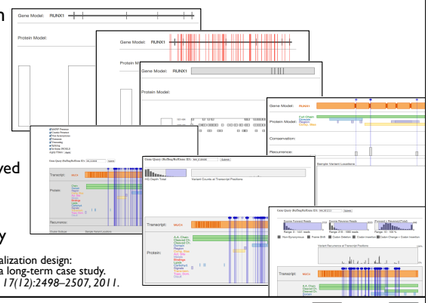
more at:
Design Study Methodology: Reflections from the Trenches and from the Stacks.
Sedlmair, Meyer, Munzner. IEEE TVCG 18(12): 2431-2440, 2012 (Proc. InfoVis 2012).

Phase 2: Core Design

5 months

learn > winnow > cast > discover > design > implement > deploy > reflect > write

- 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



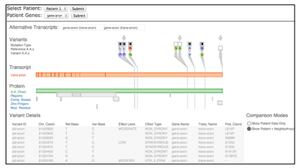
Human-centered approaches in geovisualization design: investigating multiple methods through a long-term case study.
Lloyd and Dykes. IEEE TVCG (Proc. InfoVis), 17(12):2498–2507, 2011.

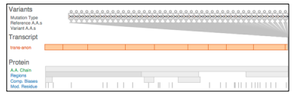
Phase 3: Two More Tasks

1 month

learn > winnow > cast > discover > design > implement > deploy > reflect > write

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





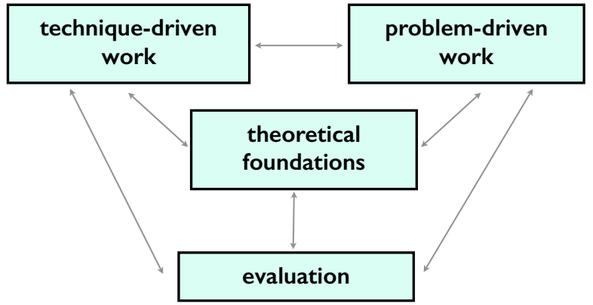
Phase 4: Reflect and write

3 months

learn > winnow > cast > discover > design > implement > deploy > reflect > write

- 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 F E P

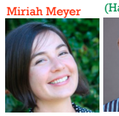
Aaron Barsky

Jenn Gardy (Microbio)

Robert Kincaid (Agilent)

Cerebral

Miriah Meyer



Hanspeter Pfister (Harvard)

MizBee

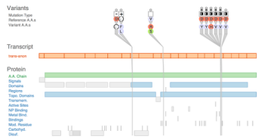
MulteeSum, Pathline

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 E P

SessionViewer: web log analysis

Heidi Lam

Diane Tang (Google)

Peter McLachlan

Stephen North (AT&T Research)

LiveRAC: systems time-series

Problem-driven: Journalism

T F E P

Matt Brehmer

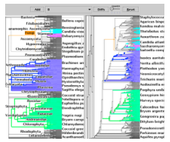
Stephen Ingram

Jonathan Stray (Assoc Press)

Overview

Technique-driven: Graph drawing

T F E P



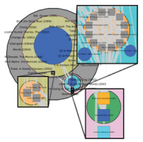
TreeJuxtaposer

James Slack

Kristian Hildebrand

Daniel Archambault

David Auber (Bordeaux)



TopoLayout
SPF
Grouse
GrouseFlocks
TugGraph

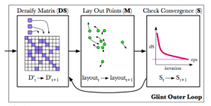
Technique-driven: Dimensionality reduction

T F E P

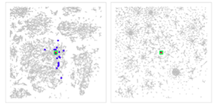
Stephen Ingram

Glimmer

DimStiller



Glint

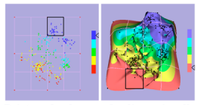


QSNE

Evaluation: Dimensionality reduction

T F E P

Melanie Tory

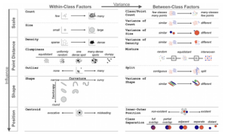


Points vs landscapes for dimensionally reduced data

Guidance on DR & scatterplot choices

Michael Sedlmair

Melanie Tory (UVic)



Taxonomy of cluster separation factors

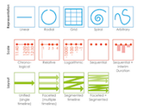
Curation & Presentation: Timelines

T F E P

TimeLineCurator
https://vimeo.com/123246662

Johanna Fulda (Sud. Zeitung)

Matt Brehmer



Timelines Revisited
timelinesrevisited.github.io/

Bongshin Lee (Microsoft)

Benjamin Bach (Microsoft)

Nathalie Henry-Riche (Microsoft)

Theoretical foundations

T F E P

- 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

Papers Process & Pitfalls

Nested Model

Design Study Methodology

Michael Sedlmair

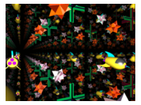
Miriah Meyer

Abstract Tasks

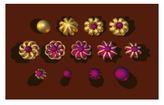
Matt Brehmer

Geometry Center 1990-1995: Expository videos

Geomview



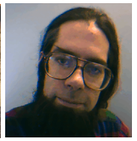
The Shape of Space



Outside In

Charlie Gunn

Stuart Levy



Mark Phillips

Delle Maxwell

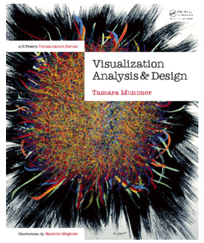
Wrap-up

- 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

- 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.