Visualization in Data Science: Challenges and Opportunities
I’ve spent a lot of time at ICG - and it was great!

2003 First ICG Class with Franz Leberl and (then new) Horst Bischof; Teaching Assistant: Tom Pock
2006 ICG BS Thesis w. Horst Bischof & Martin Urschler
2008 ICG MS Thesis w. Dieter Schmalstieg
2012 ICG PhD Thesis w. Dieter Schmalstieg
RESEARCH AREAS
<table>
<thead>
<tr>
<th>TECHNICAL CONTRIBUTIONS</th>
<th>DOMAIN DRIVEN TECHNIQUES</th>
<th>EMPIRICAL &amp; THEORETICAL WORK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel Visualization Techniques</td>
<td>Tailored Methods and Systems for High Impact Science Problems</td>
<td>Evaluation Methodology</td>
</tr>
<tr>
<td>Visualization Process Innovations</td>
<td></td>
<td>Qualitative Studies</td>
</tr>
<tr>
<td>Data Wrangling Methods</td>
<td></td>
<td>Design Spaces / Taxonomies</td>
</tr>
<tr>
<td>Accessibility</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VIS IN DATA SCIENCE

CHALLENGES & OPPORTUNITIES
1. Systems are **HARD**

2. Interaction is powerful but **ephemeral**

3. Data is not the **truth**
1. SYSTEMS ARE HARD
Publishing software increases impact of your work

Spent a lot of time on building interactive visualization systems

Adoption is minimal. Why?

SYSTEMS ARE HARD
Hard to build good UX in academic setting

Maintenance != publishing

Analysts don’t want a new & complicated tool unless it is a significant improvement

Limited expressivity

Simplicity - expressivity trade-off
WHAT CAN WE DO?

1. The idea matters most
2. Work on reusable components
3. Meet users where they are
4. Commercialize
<table>
<thead>
<tr>
<th>TITLE</th>
<th>CITED BY</th>
<th>YEAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>UpSetR: An R Package For The Visualization Of Intersecting Sets And Their Properties</td>
<td>1763</td>
<td>2017</td>
</tr>
<tr>
<td>JR Conway, A Lex, N Gehlenborg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bioinformatics 33 (18), 2938-2940</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UpSet: Visualization of Intersecting Sets</td>
<td>1390</td>
<td>2014</td>
</tr>
<tr>
<td>A Lex, N Gehlenborg, H Strobelt, R Vuillermot, H Pfister</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LineUp: Visual Analysis of Multi-Attribute Rankings</td>
<td>344</td>
<td>2013</td>
</tr>
<tr>
<td>S Gratzl, A Lex, N Gehlenborg, H Pfister, M Streit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IEEE Transactions on Visualization and Computer Graphics 19 (12), 2277-2286</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Basic UpSet Idea has been re-implemented in R, Python, Tableau, JavaScript, etc.
Reusable component in common environment

An R package that generates useful figures for inclusion in papers
There is a need for complex visualization systems, but you need resources to make it useful.

datavisyn works with 5 of the 10 biggest pharmaceuticals
OPPORTUNITIES

Publish **good ideas**! [dough]

Publish small, maintainable software **packages that fit into an ecosystem**

- Reusability, flexibility, less maintenance burdens

Develop **DSLs** instead of UI systems (Vega, reVISit, ...)

- UIs are just A LOT OF WORK
- DSLs can do things in well defined ways
- Tooling (including UIs) can follow later
1. SYSTEMS ARE **HARD**

2. **INTERACTION** IS POWERFUL BUT **EPHEMERAL**

3. DATA IS NOT THE **TRUTH**
2. **INTERACTION** is powerful but ephemeral.
WHAT ARE SELECTIONS?
SOME THINGS ARE BEST DONE THROUGH VISUALIZATION + INTERACTION
BUT, unlike code interaction typically leaves no trace

Not reproducible

Not reusable

Interaction is also “silooed”

Typically a dead end
INSPIRATION: LITERATE PROGRAMMING
LITERATE PROGRAMMING

Explain the why and how using any means necessary!

Text
Images / Visualizations
Formulas
Videos
Links
Code

[Donald E. Knuth, 1984]
LITERATE PROGRAMMING IN THE WILD
THERE IS NO STRAIGHTFORWARD WAY TO DO LITERATE DATA VISUALIZATION
Current State: no record of what was done, let alone why

Idea: make the process of an interactive, visual analysis session well reasoned and documented

Gratzl, Lex, Gehlenborg, Cosgrove, Streit; EuroVis 2016
My story to Country Somalia

![Chart showing statistical data for Country Somalia over time]

- Start
- $X=\text{GDP}$
- $Y=\text{Life Expectancy}$
- Size = Population
- Color = Continent
- Year 1800
- $Y=\text{Child Mortality}$
- $\log$ scale($X$)
A web-based provenance library
Easy to integrate in web apps

https://github.com/Trrack/trrackjs
We solved the **WHAT**, but not the **WHY**

No progress towards reusability..

So, what else can we do?
SEMANTIC SELECTIONS
WHY DO WE CARE?

*Speed up complex selections*
WHY DO WE CARE?

**ID Based Selection:**
Selected Elements: 7, 9, 13, 18, 22

**Semantic Selection:**
Elements in K-Means cluster centered at [2, 3]

Meaningful, higher level concept:
*improves reproducibility*

Robust to changes and updates in dataset:
*enables re-usability*
HOW DO WE INFER INTENT?

Selection

Predictions
- K-Means
- DBScan
- Regression
- Outlier Detection
- Skyline
- Decision Trees / Ranges
- Categories

Ranking
- Jaccard Distance
- Naive Bayes
- Classifier
- Heuristic
- Measures

Confirming Intent & Annotation

I think this cluster...

1. Range
2. Cluster
3. Outlier
Visualization and Selection

Annotation of Intent and Predictions

http://vdl.sci.utah.edu/predicting-intent/
WORKFLOWS

Based on semantic selections, we can create reusable workflows!

EuroVis 2022. Kiran Gadhave, Zach Cutler, Alexander Lex
CAPTURING SEMANTICS OF WORKFLOWS

Robust “Filter Outliers” Workflow

Original Dataset → Brush selection → Refine selection to select outliers → Filter out → Output

Input

Output
REUSING WORKFLOWS ON UPDATED DATA

Updated Dataset

Apply Workflow

New Data Point

Updated Dataset
REUSING SELECTIONS ON UPDATED DATASETS

Original Selection

Changed Dataset

Tracking A Selected Cluster
Next: Can we all do this with standard Python plots?

- Track interaction in native plots
- Enable data wrangling operations (filter, label, aggregate, etc.)
- Make steps permanent (like code)
- Allow downstream use of modified data
```python
import altair as alt
from vega_datasets import data
import numpy as np
import pandas as pd
import interactivevcd.ide as IDE # Our library
IDE.enable('altair') # Call this to enable integration with altair

source = data.movie_rll()
selected_movies = ['Title', 'IMDB_Rating', 'Rotten_Tomatoes_Rating', 'Major_Centre']

gts = alt.selection_point(name='point_xl', encoding='e')
rect = alt.Chart(source).mark_rect().encode(
    alt.X('IMDB_Rating:Q', bin=True),
    alt.Y('Rotten_Tomatoes_Rating:Q'),
    alt.Color('count:Q').scale(domain='max:q', range='greenblue').title('Total Records'),
)
circ = rect.mark_point().encode(
    alt.ColorValue('grey'),
    alt.Size('count:Q'),
    title='RECORDS in Selection',
    transform_filter=gts
)
bar = alt.Chart(source).mark_bar().encode(
    x='Major_Centre:N',
    y='count:Q',
    color=alt.condition(gts, alt.ColorValue('steelblue'), alt.ColorValue('grey'))
).add_params(gts)

ealt.vconcat(
    rect + circ,
    bar
    ).resolve_legend(
        color='independent',
        size='independent' )

# df_44352_df
# df = df_44352_df;) # groupby_df = df.groupby('''_aggregate''')
# groupby_df_diff = df.groupby('keys')
# groupby_df_get_group('Agg_703123ca')
# groupby_df_reset_axis('reset_axis')
```
Fits perfectly in ecosystem of Python data analysis

No installation burden

Interactive VIS is suddenly a first class operation in notebooks
1. SYSTEMS ARE HARD
2. INTERACTION IS POWERFUL BUT EPHEMERAL
3. DATA IS NOT THE TRUTH
3. DATA IS NOT THE TRUTH
Tracking Coronavirus in Utah: Latest Map and Case Count

Updated Aug. 11, 2022

New reported cases

7–day average

this dip is probably an artifact of the data collection, this should be higher.

DEC. 24, 2021

Day with data anomaly

Daily average: 1,004

New cases: 0
The visualization showed many cases not using blood recycling. But to my knowledge, we almost always turn on the machine for it. Likely, sometimes they were not recorded.
Data is imperfect but experts know how and why

Challenges:

Hardly documented (lab notebooks, methods sections)

Not saliently available to others

Bad tool support

Limited metadata formats
A **DATA HUNCH** is an analyst's knowledge about how and why the data is an imperfect and partial representation of the phenomena of interest.
COVID Cases

New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people, colored by stringency of country's response as of March 01, 2022. Strict Policies stringency indicates a stricter response. If policies vary at the subnational level, the result is shown as the response level of the strictest sub-region. Data shows Mar 01, 2022. Data Source: OurWorldInData

*The categorical labels do not help much*
COVID Cases

New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people, colored by stringency of country's response as of March 01, 2022. Strict Policies stringency indicates a stricter response. If policies vary at the subnational level, the result is shown as the response level of the strictest sub-region. Data shows Mar 01, 2022. Data Source: OurWorldInData

**The categorical labels do not help much**

**RATING, COMMENTING, COLLABORATING**
CHALLENGES

Biases and Trust

Reinforcing preconceived ideas?
Explain away “inconvenient” data points?

Need to provide reasoning and justifications
Intended for trusted teams
Currently a stand-alone tool

Goals:

- integrate with plotting library, such as Vega-Altair (see “systems are hard”)
- preserve hunches across data structures and analysis steps

Many other efforts needed to address the gap between data and truth
1. SYSTEMS ARE HARD
2. INTERACTION IS POWERFUL BUT EPHEMERAL
3. DATA IS NOT THE TRUTH
Thanks to: Kiran Gadhave, Hainan Lin, Zach Cutler, Devin Lange, Max Lisnic, Marc Streit, Jochen Görtler, Oliver Deussen, Miriah Meyer, Jeff Phillips, Samuel Gratzl, Holger Stitz, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, Hanspeter Pfister, and many others!