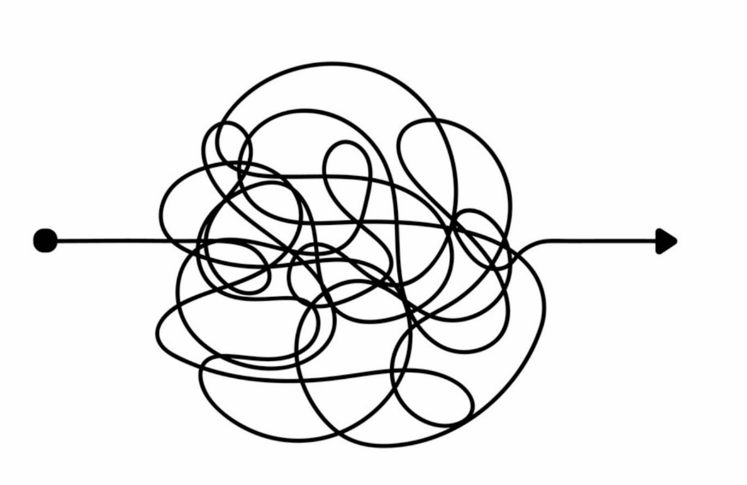
How Data Scientists Use Computational Notebooks for Real-Time Collaboration

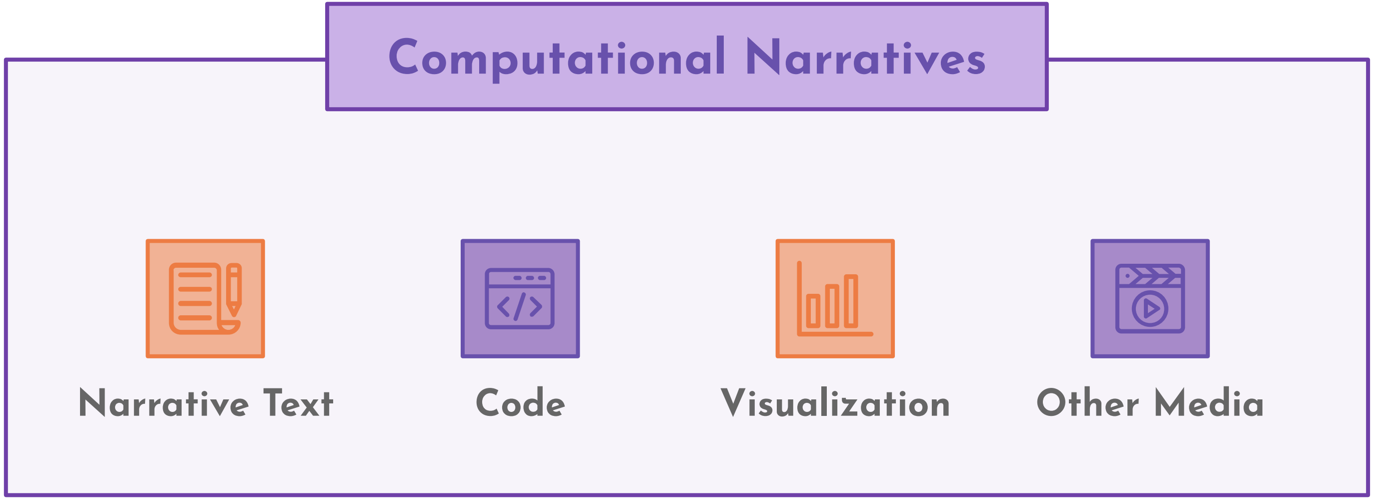
April Yi Wang | Anant Mittal | Chris Brooks | Steve Oney

University of Michigan

The Story Behind Data Analysis

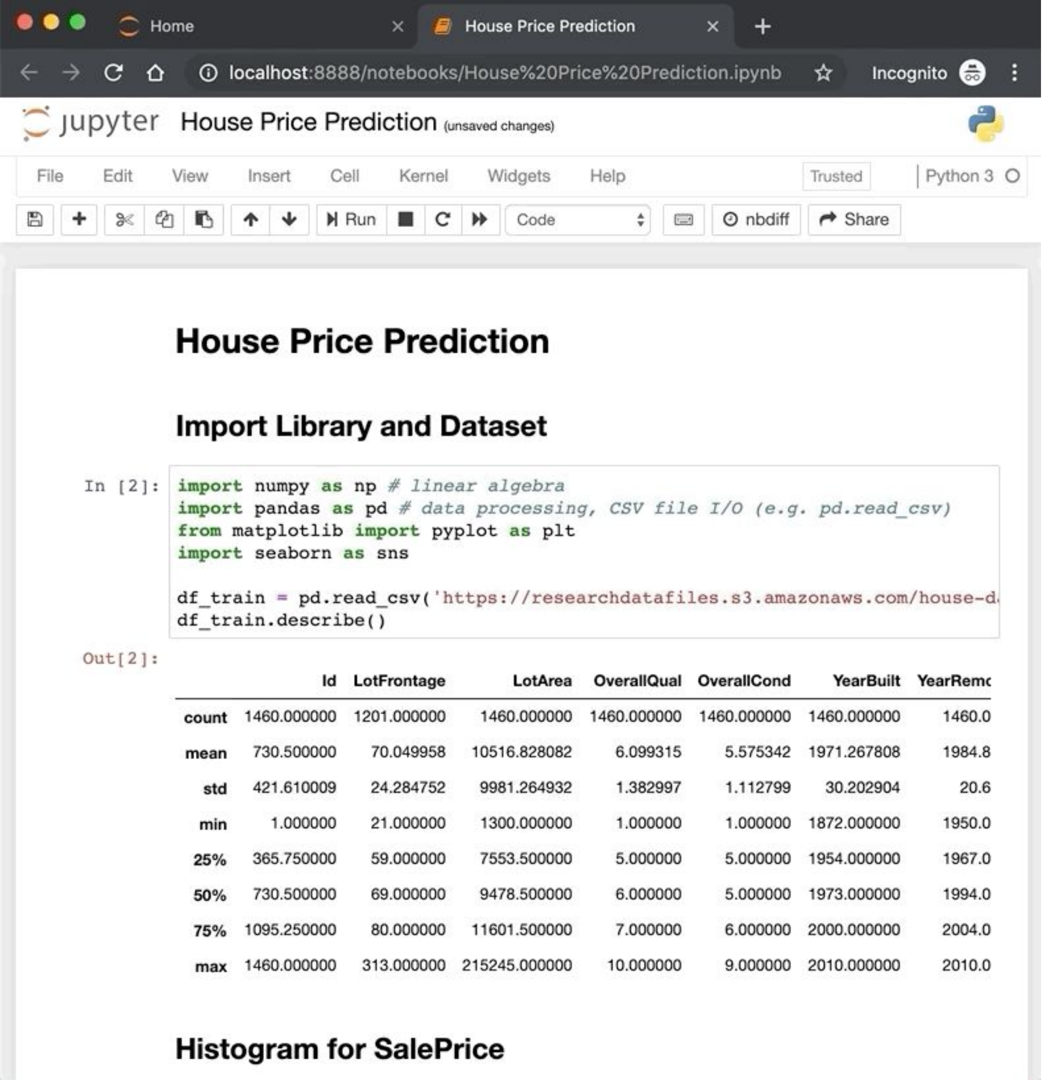


The Story Behind Data Analysis



Computational Narratives

- Narrative Text
- Code
- Visualization
- Other Media



House Price Prediction

Import Library and Dataset

```
In [2]: import numpy as np # linear algebra
        import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
        from matplotlib import pyplot as plt
        import seaborn as sns

        df_train = pd.read_csv('https://researchdatafiles.s3.amazonaws.com/house-d
        df_train.describe()
```

Out[2]:

	Id	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemc
count	1460.000000	1201.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.0
mean	730.500000	70.049958	10516.828082	6.099315	5.575342	1971.267808	1984.8
std	421.610009	24.284752	9981.264932	1.382997	1.112799	30.202904	20.6
min	1.000000	21.000000	1300.000000	1.000000	1.000000	1872.000000	1950.0
25%	365.750000	59.000000	7553.500000	5.000000	5.000000	1954.000000	1967.0
50%	730.500000	69.000000	9478.500000	6.000000	5.000000	1973.000000	1994.0
75%	1095.250000	80.000000	11601.500000	7.000000	6.000000	2000.000000	2004.0
max	1460.000000	313.000000	215245.000000	10.000000	9.000000	2010.000000	2010.0

Histogram for SalePrice

Jupyter Notebook

Jupyter notebooks consist of **"cells"** — typically small chunks of code or narrative text in the Markdown format.

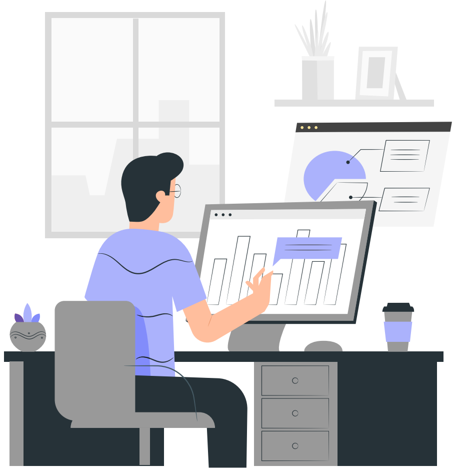
Users can execute cells (typically, but not necessarily, from top to bottom) and observe their outputs.

Writing and Sharing Computational Notebooks in Various Contexts

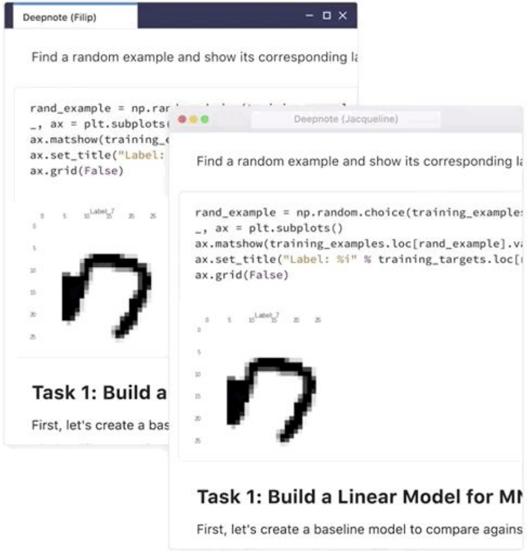
- **Data Science Education**
  Kross and Guo, CHI 19

- **Open Science**
  Randles et al., JCDL 17

- **Professional Data Analytics**
  Kery et al., CHI 18



From Sharing to Synchronous Editing



Find a random example and show its corresponding la

```
rand_example = np.rar
_, ax = plt.subplots(
ax.matshow(training_e
ax.set_title("Label:
ax.grid(False)
```

Task 1: Build a

First, let's create a bas

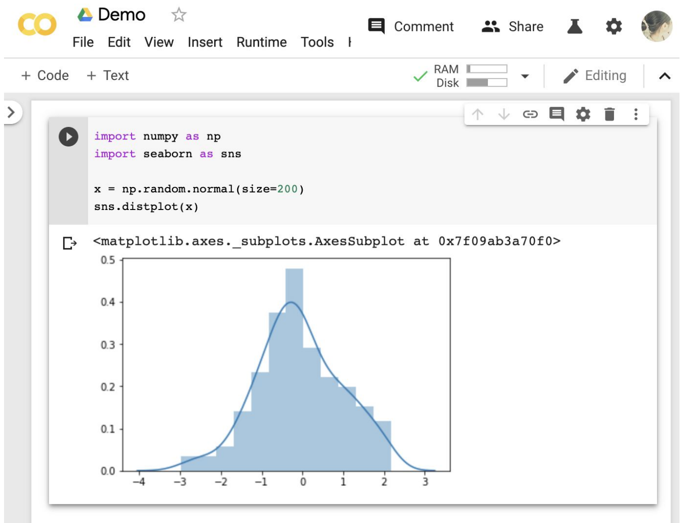
Find a random example and show its corresponding la

```
rand_example = np.random.choice(training_example:
_, ax = plt.subplots()
ax.matshow(training_examples.loc[rand_example].va
ax.set_title("Label: %i" % training_targets.loc[
ax.grid(False)
```

Task 1: Build a Linear Model for MN

First, let's create a baseline model to compare agains

Deepnote



```
import numpy as np
import seaborn as sns

x = np.random.normal(size=200)
sns.distplot(x)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f09ab3a70f0>
```

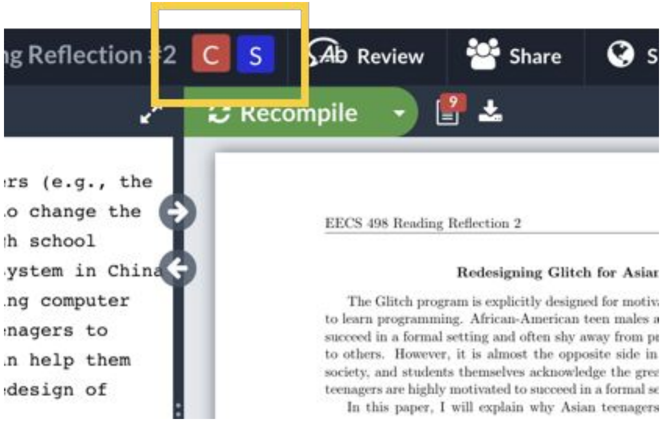
Google colab

Issues with Synchronous Editing

- Reluctant to write together when collaboratively constructing a document

- Social embarrassment to be watched by others when typing

~ Wang et al. CSCW'17



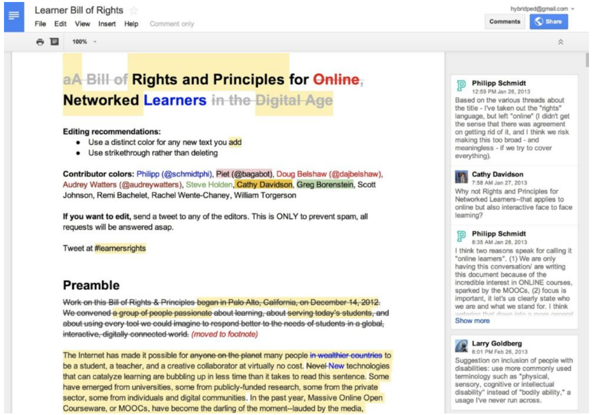
```
rs (e.g., the
o change the
h school
ystem in China
ng computer
nagers to
n help them
design of
```

EECS 498 Reading Reflection 2

Redesigning Glitch for Asiar

The Glitch program is explicitly designed for motiv
to learn programming. African-American teen males a
succeed in a formal setting and often shy away from pe
to others. However, it is almost the opposite side in
society, and students themselves acknowledge the grea
teenagers are highly motivated to succeed in a formal se
In this paper, I will explain why Asian teenagers

Issues with Synchronous Editing



~~aA Bill of~~ Rights and Principles for Online~~,~~ Networked Learners ~~in the Digital Age~~

Editing recommendations:

- Use a distinct color for any new text you add
- Use strikethrough rather than deleting

Contributor colors: Philipp (@schmidtphi), Piet (@bagabot), Doug Belshaw (@dajbelshaw), Audrey Watters (@audreywatters), Steve Holden, Cathy Davidson, Greg Borenstein, Scott Johnson, Remi Bachelet, Rachel Wente-Chaney, William Torgerson

If you want to edit, send a tweet to any of the editors. This is ONLY to prevent spam, all requests will be answered asap.

Tweet at #learnersrights

Preamble

~~Work on this Bill of Rights & Principles began in Palo Alto, California, on December 14, 2012. We convened a group of people passionate about learning, about serving today's students, and about using every tool we could imagine to respond better to the needs of students in a global, interactive, digitally connected world.~~ *(moved to footnote)*

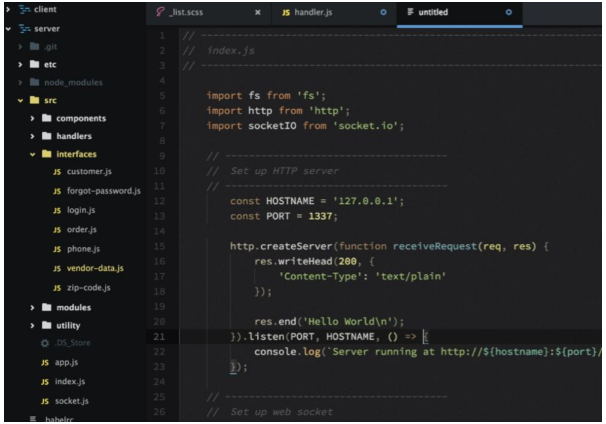
The Internet has made it possible for ~~anyone on the planet~~ many people ~~in wealthier countries~~ to be a student, a teacher, and a creative collaborator at virtually no cost. ~~Novel~~ New technologies that can catalyze learning are bubbling up in less time than it takes to read this sentence. Some have emerged from universities, some from publicly-funded research, some from the private sector, some from individuals and digital communities. In the past year, Massive Online Open Courseware, or MOOCs, have become the darling of the moment--lauded by the media,

Philipp Schmidt
12:59 PM Jan 26, 2013
Based on the various threads about the title - I've taken out the "rights" language, but left "online" (I didn't get the sense that there was agreement on getting rid of it, and I think we risk making this too broad - and meaningless - if we try to cover everything).

Cathy Davidson
7:58 AM Jan 27, 2013
Why not Rights and Principles for Networked Learners--that applies to online but also interactive face to face learning?

Philipp Schmidt
8:35 AM Jan 28, 2013
I think two reasons speak for calling it "online learners". (1) We are only having this conversation/ are writing this document because of the incredible interest in ONLINE courses, sparked by the MOOCs, (2) focus is important, it let's us clearly state who we are and what we stand for. I think watering that down into a more general
Show more

Larry Goldberg
6:01 PM Feb 26, 2013
Suggestion on inclusion of people with disabilities: use more commonly used terminology such as "physical, sensory, cognitive or intellectual disability" instead of "bodily ability," a usage I've never run across.



```
// ------------------------------------------------------------
//  index.js
// ------------------------------------------------------------

    import fs from 'fs';
    import http from 'http';
    import socketIO from 'socket.io';

    // ------------------------------------
    //  Set up HTTP server
    // ------------------------------------
        const HOSTNAME = '127.0.0.1';
        const PORT = 1337;

        http.createServer(function receiveRequest(req, res) {
            res.writeHead(200, {
                'Content-Type': 'text/plain'
            });

            res.end('Hello World\n');
        }).listen(PORT, HOSTNAME, () => {
            console.log(`Server running at http://${hostname}:${port}/
        });

    // ------------------------------------
    //  Set up web socket
```

Collaborative Writing

Wang et al. CSCW'17

D'Angelo et al. CSCW'18

Collaborative Programming

Goldman et al. UIST'11

Oney et al. CSCW'18

What about collaborative data science?

data science ≠ writing + coding

RQ1 What tools and strategies do data scientists currently use for collaboration?

RQ2 Compared to working on individual notebooks in a collaborative setting, how does synchronous notebook editing change the way data scientists collaborate in computational notebooks?

RQ3 What challenges, if any, do data scientists perceive in synchronous notebook editing?

Study 1 Formative Survey

RQ1 What **tools and strategies** do data scientists currently use for collaboration?

RQ2 Compared to **working on individual notebooks** in a collaborative setting, how does **synchronous notebook editing** change the way data scientists collaborate in computational notebooks?

RQ3 What **challenges**, if any, do data scientists perceive in synchronous notebook editing?

Study 2 Observational Study

Study 1 Formative Survey

Demographic

Data Source

195

Valid Responses

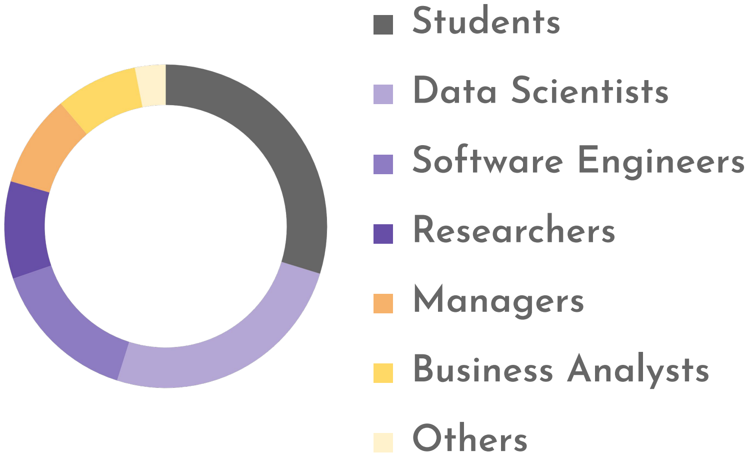
35

Umich

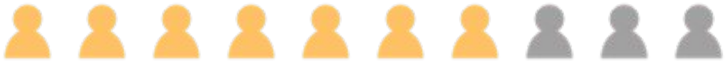
160

Coursera

Job Roles



Experience with Data Science



Choices of Tools

Programming	Jupyter Notebooks (88.72%), IDEs (51.79%), **Google Colab (12.31%)**
Communication	Emails (79.49%), Face-to-face Communication (68.72%)
Project Management	Version Control Tools (49.74%), Task Tracking Tools (21.03%)

Study 1 Formative Survey

Strategies for Keeping a Shared Understanding

Discussions and Meetings	Weekly meeting among team members;
Frequently Check-in	Communicate actively and frequently;
Documentation	Keep notes in Google Docs; ... comments in code;
Organization	Divide up the work into definable parts;
Shared Assets	Common repository for files;
Others	Code review to ensure code matched intent

Study 1 Formative Survey

RQ1 What **tools and strategies** do data scientists currently use for collaboration?

Traditional Collaboration Setting

Working on individual Jupyter notebooks

Emerging Collaboration Setting

Working on notebooks with synchronous editing

Study 1 Formative Survey

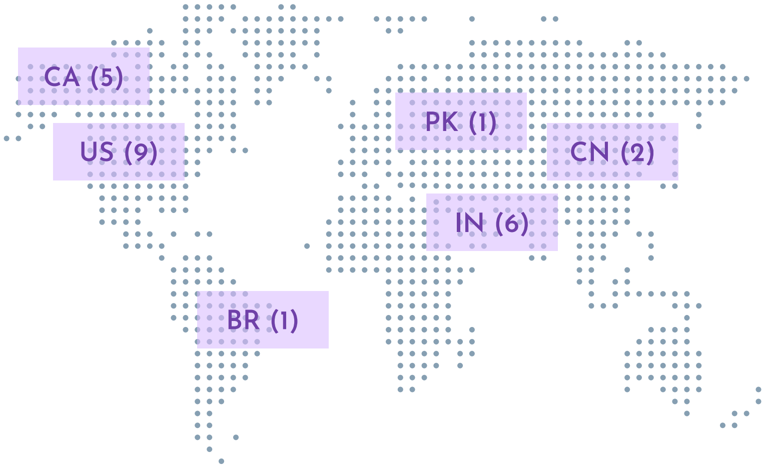
RQ2 Compared to **working on individual notebooks** in a collaborative setting, how does **synchronous notebook editing** change the way data scientists collaborate in computational notebooks?

RQ3 What **challenges**, if any, do data scientists perceive in synchronous notebook editing?

Study 2 Observational Study

Participants

- 24 participants (12 from the survey)
- Randomly assigned to pairs
- Work collaboratively on a predictive modeling problem remotely



Study Setup

Non-Shared Condition

Participants worked on individual notebooks

- ✓ Exchange the notebook file
- ✓ Set up a git repository
- ✓ Send code snippets through other tools if necessary

Shared Condition

Synchronous editing was supported.

- ✓ Share notebook edits and actions (e.g., moving cursor, adding cells) in real-time
- ✓ Execute code on a single interpreter
- ✓ Update output and runtime variables among collaborators

Study 2 Observational Study

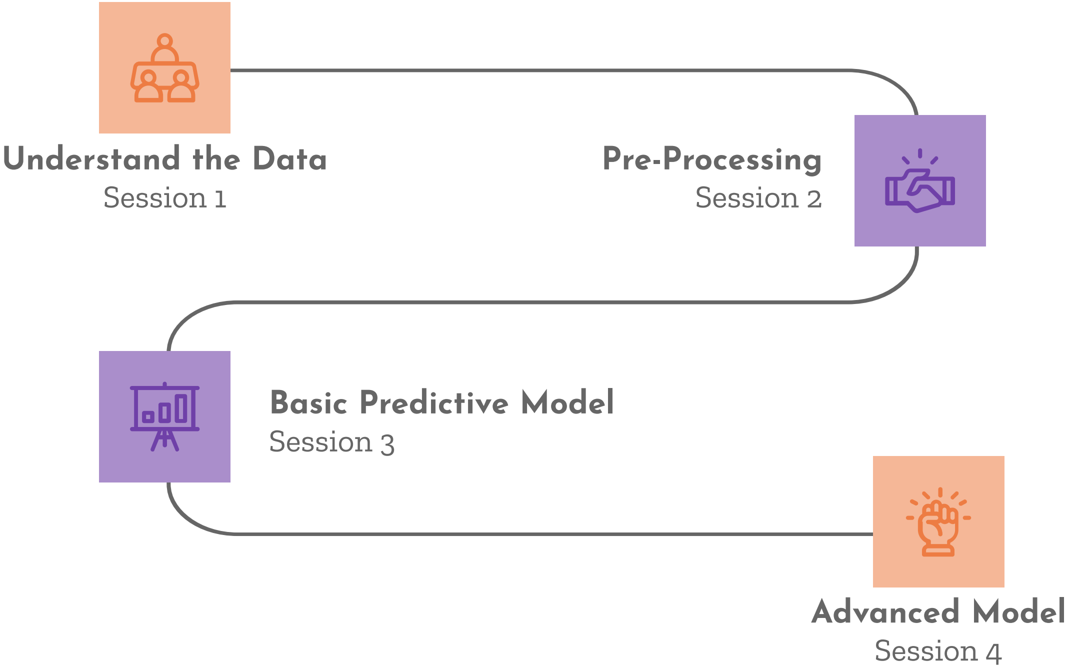
Task

- Predict house sale prices using 80 features (e.g., lot size, year built)
- Additional incentives for the group with the lowest error score
- Submit prediction results as well as one Jupyter notebook report
- Choose from text-messaging (Slack) or video-conferencing (Google Hangouts) for communication

Study 2 Observational Study

Procedure

The study consisted of four sessions, each of which lasted an hour.



Study 2 Observational Study

Collaboration Style

Collaboration Style	GID	Definition
Single Authoring		One team member contributed the majority of ideas and did the majority of the implementation, while the others did not contribute much.
Pair Authoring		
Divide and Conquer		
Competitive Authoring		

Study 2 Observational Study

Collaboration Style

Collaboration Style	GID	Definition
Single Authoring		One team member contributed the majority of ideas and did the majority of the implementation, while the others did not contribute much.
Pair Authoring		One team member did the majority of implementation while the others contributed ideas, engaged in discussions and reviewed the results.
Divide and Conquer		
Competitive Authoring		

Collaboration Style

Collaboration Style	GID	Definition
Single Authoring		One team member contributed the majority of ideas and did the majority of the implementation, while the others did not contribute much.
Pair Authoring		One team member did the majority of implementation while the others contributed ideas, engaged in discussions and reviewed the results.
Divide and Conquer		Members divided the task into subgoals and explored the subgoals independently.
Competitive Authoring		

Collaboration Style

Collaboration Style	GID	Definition
Single Authoring		One team member contributed the majority of ideas and did the majority of the implementation, while the others did not contribute much.
Pair Authoring		One team member did the majority of implementation while the others contributed ideas, engaged in discussions and reviewed the results.
Divide and Conquer		Members divided the task into subgoals and explored the subgoals independently.
Competitive Authoring		Team members wrote the code for the same purpose and reached the consensus to use the code by whomever finished first.

Collaboration Style

Collaboration Style	GID	Definition
Single Authoring	S2, S5	One team member contributed the majority of ideas and did the majority of the implementation, while the others did not contribute much.
Pair Authoring	S6	One team member did the majority of implementation while the others contributed ideas, engaged in discussions and reviewed the results.
Divide and Conquer	N2, N5, S1, S3, S4	Members divided the task into subgoals and explored the subgoals independently.
Competitive Authoring	N1, N3, N4, N6	Team members wrote the code for the same purpose and reached the consensus to use the code by whomever finished first.

Study 2 Observational Study

Communication Channels

	Non-Shared Condition	Shared Condition
Choices of Tools	Text Messaging (6/6)	Text Messaging (3/6) Video Conferencing (3/6)

Participants in the non-shared condition send files, code snippets, and output more often.

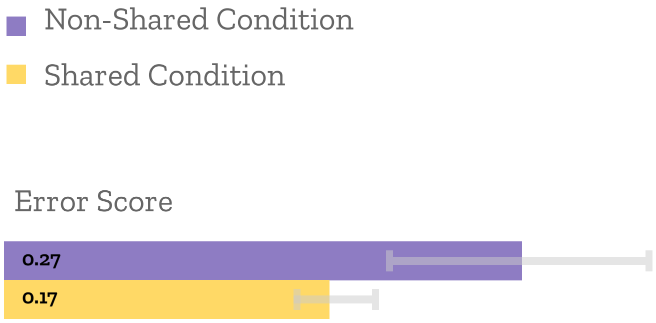
➔ Working in the shared notebook may reduce the communication costs by establishing a shared context.

Study 2 Observational Study

Final Submissions

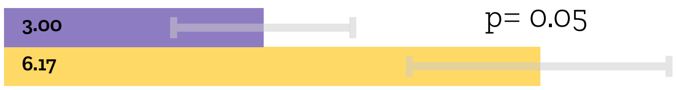
→ Groups in the shared condition achieved a better prediction result.

→ Groups in the shared condition explored more alternative models.

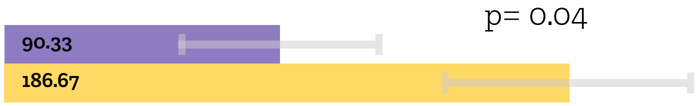


Error Score

Number of Alternative Models*

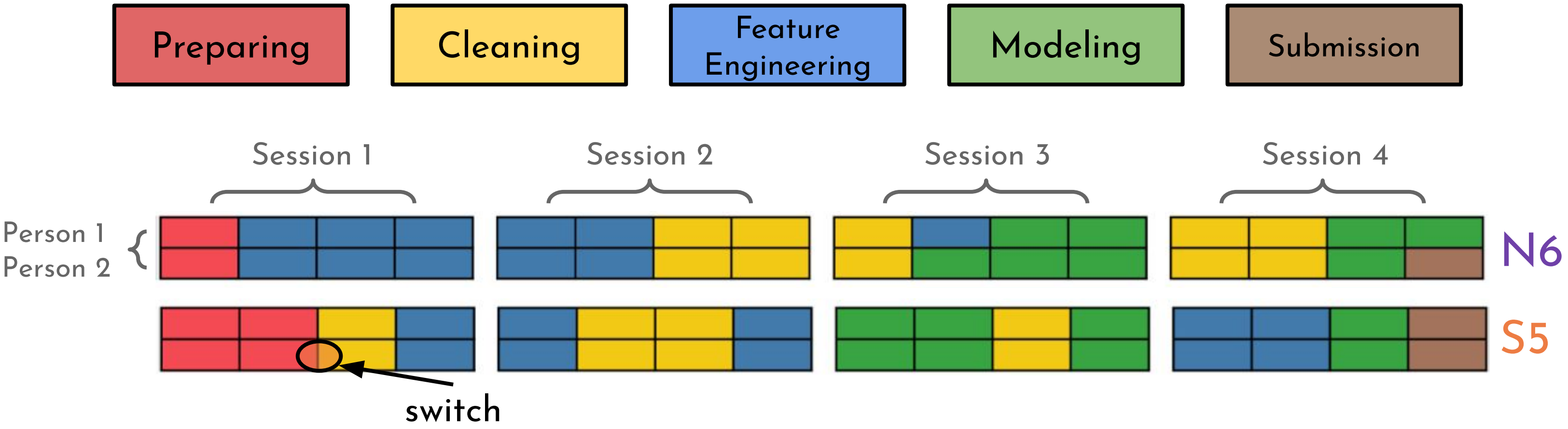


Lines in the Notebook*



Study 2 Observational Study

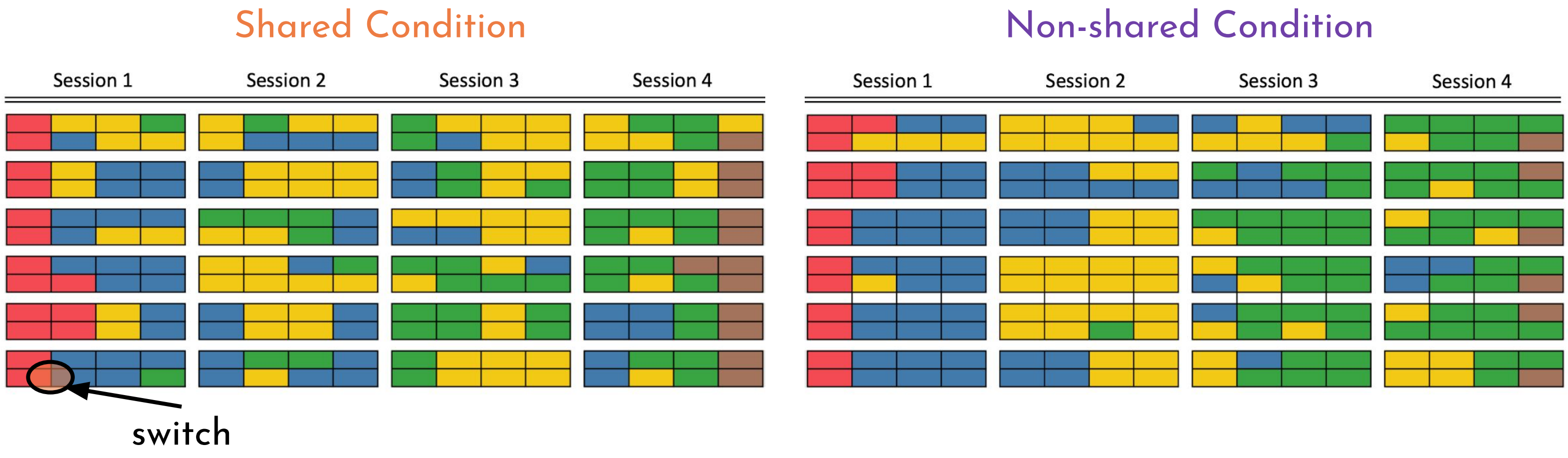
Work Across Phases



Participants in the shared condition switched more frequently ($p<0.001$).

➔ Working on the same notebook provides collaborators with convenience to branch through tasks

Work Across Phases


Shared Condition

Non-shared Condition

Participants in the shared condition switched more frequently ($p<0.001$).

➔ Working on the same notebook provides collaborators with convenience to branch through tasks

Benefits of Synchronous Editing in Notebook

➔ Reducing communication costs

➔ Flexibility to branch through tasks

➔ Enabling explorations of more alternative models

➔ Leading to a better prediction result

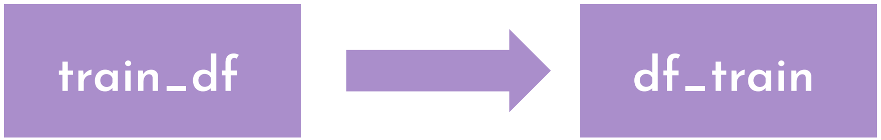
Study 2 Observational Study

Challenges of Synchronous Editing

Study 2 Observational Study

Challenges of Synchronous Editing

1. Interference with each other



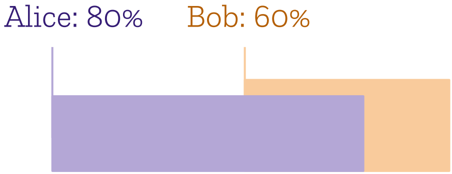
*"… **When using Jupyter Notebook together, it's hard to keep track of variable names.** Everyone might use a different name and may cause issues. For example, my teammate used train_df as name, and later changed it to something else, but I wanted him to keep using the original name…" (P2 from S1)*

Study 2 Observational Study

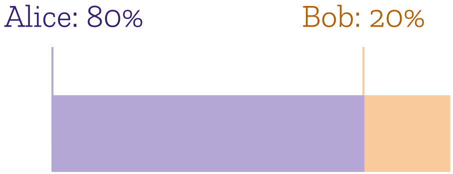
Challenges of Synchronous Editing

2. Lack of Strategic Coordination

Why competitive authoring happens in the non-shared condition?



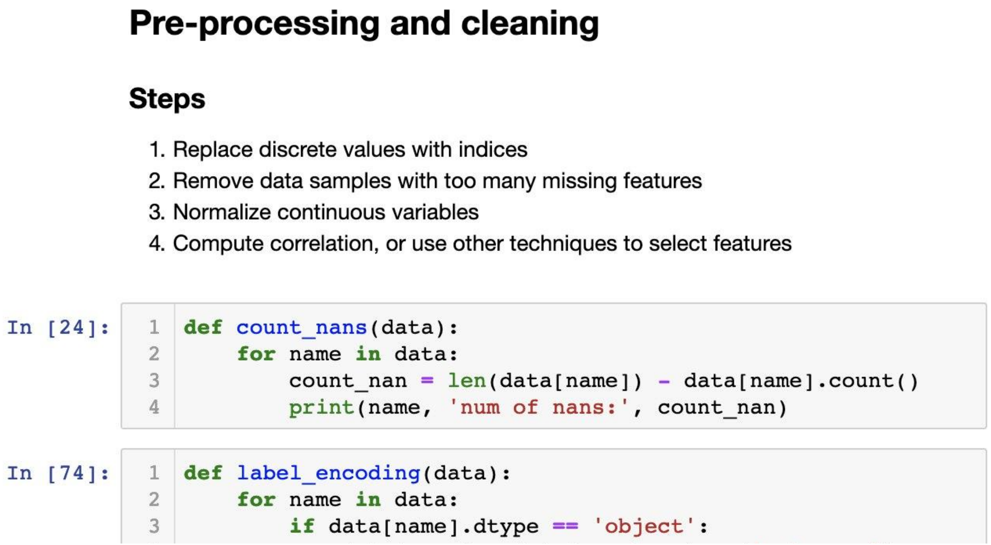
Why single authoring happens in the shared condition?



"... I feel I am not splitting work well enough. ***I was thinking about how to get the work done and just tried the ideas on myself****...." (P11 from S2)*

Challenges of Synchronous Editing

S3 wrote down subtasks in the notebook.



Pre-processing and cleaning

Steps

1. Replace discrete values with indices
2. Remove data samples with too many missing features
3. Normalize continuous variables
4. Compute correlation, or use other techniques to select features

In [24]:

```
def count_nans(data):
    for name in data:
        count_nan = len(data[name]) - data[name].count()
        print(name, 'num of nans:', count_nan)
```

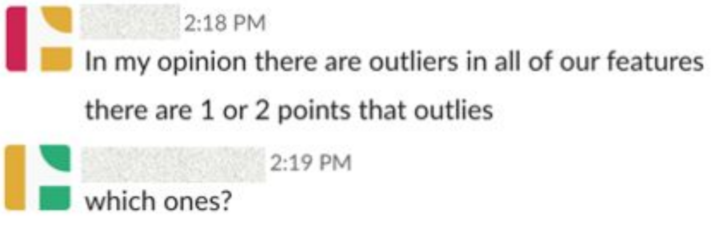
In [74]:

```
def label_encoding(data):
    for name in data:
        if data[name].dtype == 'object':
```

Challenges of Synchronous Editing

3. Contextual Chatting

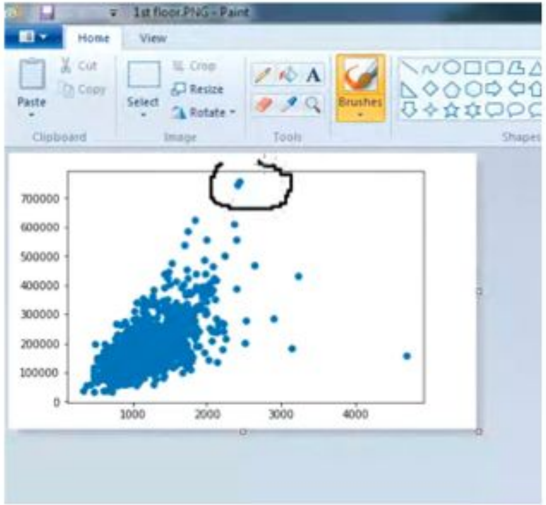
P14 and P15 were looking at the scatterplots of independent variables together.



[redacted] 2:18 PM
In my opinion there are outliers in all of our features
there are 1 or 2 points that outlies

[redacted] 2:19 PM
which ones?

P14 downloaded the graph, opened MS Paint, annotated the graph and sent it back to P15.



Study 2 Observational Study

Challenges of Synchronous Editing

1. Interference with each other
2. Lack of Strategic Coordination
3. Contextual Chatting
4. Lack of Awareness
5. Problems with the Linear Structure
6. Privacy Concerns

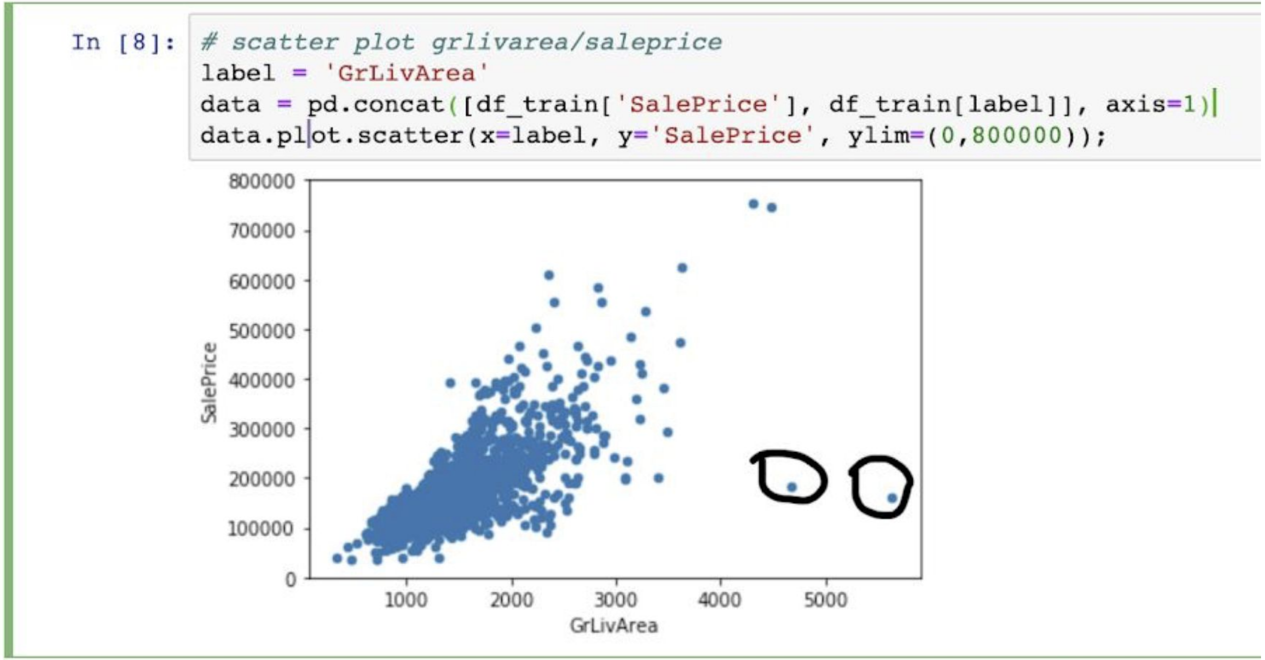
Study 2 Observational Study

- ➔ Working on the same notebook results in different collaboration styles compared to working on individual notebooks.
- ➔ Synchronous editing tools improve collaboration by helping data scientists maintain a shared context and improve work efficiency.
- ➔ However, the current real-time collaborative editing features may lead to several problems (e.g., interference with each others' work, unbalanced contributions).

Extending Our Understanding of Collaborative Editing Across Contexts

- Collaborators may hold different programming backgrounds and domain knowledge
- Different roles in collaborative data science

Example: How to deal with the two dots?



```
In [8]: # scatter plot grlivarea/saleprice
        label = 'GrLivArea'
        data = pd.concat([df_train['SalePrice'], df_train[label]], axis=1)
        data.plot.scatter(x=label, y='SalePrice', ylim=(0,800000));
```

Design Implications

- Improve Awareness of Collaborators' Activity
- Provide Access Control
- Enable Discussions within Notebooks

Limitations

- Generalizability
    - the type of data science problems
    - the expertise of collaborators
    - the size of the team
    - the synchronicity of the collaboration

How Data Scientists Use Computational Notebooks for Real-Time Collaboration

What tools and strategies do data scientists currently use for collaboration?

Study 1 - Formative Survey on Collaborative Data Science

Traditional Collaboration Setting + Emerging Collaboration Setting

How does synchronous notebook editing change the way data scientists collaborate?

What challenges do data scientists perceive in synchronous notebook editing?

Study 2 - Observational Study on Collaborative Data Science

Having synchronous editing is great for collaborative data science, but not perfect!

Presenter: April Yi Wang | aprilww@umich.edu

Co-authors: Anant Mittal, Chris Brooks, Steve Oney