VISUALIZATION & ANALYTICS CONCENTRATION

UNIVERSITY OF CALGARY COMPUTER SCIENCE

FALL 2017
You want to learn how to use statistics, visualization, and analysis tools to solve data-driven problems.

You want to build new visualization and analysis systems.
TOOLS: TABLEAU
In [41]:

delay_counts_df = delay_counts_df.rename_axis(['ARR_DELAY', 'COUNTS'], axis=1)
trip_counts_df = trip_counts_df.rename_axis(['FL_DATE', 'COUNTS'], axis=1)

Now we divide the delay counts by the total trip counts and perform the same transform we did previous to produce the N by N matrix. In this case, each cell represents the proportion of flights between each (origin → destination) that were delayed.

In [42]:

mat = (delay_counts_df / trip_counts_df).unstack().T.reset_index(level=0, drop=True).T

In this second heatmap plot, the (CA → CA) and (TX → TX) hotspots from the firt visualization no longer stand out. Though there are many in-state delays for these two states, there are even more flights, keeping the percentage of delayed flights for these in-state trips lower than other routes.

To the contrary, we see some cases where all flights from one state to another had arrival delays:

- (AR → UT)
- (MT → NV)
- (CO → RI) and (RI → CO)

We can also see some other moderately hot spots, such as (AK → NJ) and (OK → MN), which seem to have a higher percentage of delays than other state pairs.

One “crosshair” jumps out in the visualization: the row and column representing Illinois are nearly both filled with non-gray cells. On closer inspection, we see Illinois sends flights to and receives flights from every other state except one: TT, the state code abbreviation for U.S. Pacific Trust Territories and Possessions. And though it is difficult to make accurate relative value judgments from this visualization, it appears the run of cells in the row and column for Illinois are darker than most other row or column runs (e.g., GA).

In [43]:

fig, ax = plt.subplots(figsize=(18, 18))
asymmatplot(mat, names=mat.columns, ax=ax, cmap='OrRd', cmap_range=(0., 1.0))
HOW TO...

USE DATA ANALYSIS TOOLS AND STATISTICAL METHODS

MANAGE, ORGANIZE, AND MAKE SENSE OF DATA

BUILD NEW VISUALIZATION AND ANALYSIS TOOLS
POSSIBLE CAREER PATHS

DATA **ANALYST**
DATA **SCIENTIST**
DATA **VISUALIZATION**

DESIGNER
DEVELOPER
ENGINEER
RESEARCHER

IN
LARGE ORGANIZATIONS
SMALL ORGANIZATIONS
STARTUPS
GOVERNMENTS
...

COURSES (6)

one of: CPSC 583 - Intro to Information Visualization
CPSC 599.87 - Visualization & Analytics

one of: CPSC 471 - Databases
CPSC 599.44 - Machine Learning

one of: STAT 321 - Intro to Probability
STAT 327 - Stats for Physical & Environmental Sciences
STAT 423 - Sampling Theory of Surveys

three options (1.5 FCEs) from:
CPSC 453, 481, 502.03, 502.07, 583, 599.87, 599.88, or 599.89
(or 503.03, 503.07)
OPTIONS
3 one-semester courses (or 9 units, 1.5 FCEs)

CPSC 453 - Computer Graphics
CPSC 481 - Human Computer Interaction
CPSC 583 - Information Visualization
CPSC 599.87 - Visualization & Analytics
599.88 - Physical & Tangible HCI
599.89 - Scientific Visualization

CPSC 502.03, 502.07 - Honors Research Project in Graphics / HCI (6 units)
CPSC 503.03, 503.07 - Research Project in Graphics / HCI
CPSC RESEARCHERS IN THIS AREA

DR. SHEELAGH CARPENDALE
DR. WESLEY WILLETT
DR. FRANK MAURER
DR. ANTHONY TANG
DR. LORA OEHLBURG
DR. USMAN ALIM
DR. MARIO COSTA SOUSA
DR. FARAMARZ SAMAVATI
DR. CHRISTIAN JACOB

OPPORTUNITIES FOR 502/503 PROJECTS AND SUMMER RESEARCH
QUESTIONS?

**one** of: CPSC 583 - Intro to Information Visualization
CPSC 599.87 - Visualization & Analytics

**one** of: CPSC 471 - Databases
CPSC 599.44 - Machine Learning

**one** of: STAT 321 - Intro to Probability
STAT 327 - Stats for Physical & Environmental Sciences
STAT 423 - Sampling Theory of Surveys

**three options (1.5 FCEs) from:**
CPSC 453, 481, 502.03, 502.07, 583, 599.87, 599.88, or 599.89
(or 503.03, 503.07)
MINOR IN DATA SCIENCE

UNIVERSITY OF CALGARY FACULTY OF SCIENCE
(STARTING 2018)
WHY MINOR IN DATA SCIENCE?

Tools and skills for data-driven, statistical, and analytic thinking.

Domain knowledge and application areas.
Data Collection - Applying automated, manual, and hybrid approaches to elicit, sense, simulate, or otherwise capture data.

Data Cleaning - Transforming raw, noisy, incomplete, or otherwise dirty data into forms that can support further analysis.

Exploratory Analysis - Using visualization, statistics, and other techniques to examine new data and identify possible opportunities for analysis.

Statistical & Computational Analysis - Applying computational, mathematical, or statistical techniques and models to make sense of and extract knowledge from data.

Presentation - Using visual, oral, and written approaches to communicate analysis processes, observations, and results.
DATA SCIENCE MINOR

**Focus** on learning statistics and data analysis skills and applying them to outside domain areas.
- Using critical thinking and data to solve domain-specific problems.

**Cohort** — All Majors
**Courses** — 7-10
**Electives** — (4) At least 2 non-CPSC
**Capstone** — Required (DATA 501)

VISUALIZATION & ANALYTICS CONCENTRATION

**Focus** on Visualization and Analysis skills with an computer science and human-centered design perspective.
- Designing novel visualizations and exploring/building new analysis tools

**Cohort** — CPSC only
**Courses** — 6 (Most fill existing elective reqs.)
**Electives** — (3) Mostly 400/500-level CPSC
**Capstone** — Optional (CPSC 502/503)
<table>
<thead>
<tr>
<th>Fall</th>
<th>Winter</th>
</tr>
</thead>
</table>
| **DATA 201: “Thinking with Data”** | **DATA 211: “Programing with Data”**  
(students can substitute CPSC 217 or 231) |
| **DATA 311: “Data Processing and Storage”**  
(students can substitute CPSC 471) | **STAT 205**  
(students can substitute STAT 217) |
| **DATA 305: “Computational Statistical Modeling”**  
(students can substitute STAT 323) | |
| **[4x] Data Science Electives**  
(Courses from a variety of departments. CPSC and MATH/STATS majors must take at least 2 electives outside their major field. MATH 211 is strongly suggested.) | |
| **DATA 5xx: Capstone** – Single term experiential learning project course |

Students will work with a faculty/industry sponsor to perform detailed analyses of real world data. Typical projects will require students to integrate skills from the entire program to:

- Acquire data (based on industry, research, social-related data sets, goals and questions).
- Clean, transform, and store data
- Explore data using statistical, computational, and visual methods
- Present findings and results
ELECTIVES

- Biology 315
- **Computer Science** 471, 481, and 583
- English 459
- Geology 597
- Geophysics 517, 549, 557
- Linguistics 560
- Mathematics 211
- **Medical Science** 401, 519
- Sociology 315, 355
- **Statistics** 323, 423, 425, 429, 431, 505, 519, 523, 531, 533, 541

**THIS LIST WILL GROW. YOU CAN PETITION TO HAVE COURSES ADDED IF THEY FIT THE THEME.**
## POSSIBLE STUDENT SCENARIOS

### Visual Studies Major + Data Science Minor
- **DATA 201**
- **DATA 211** *(CPSC 217, 231)*
- **STAT 205** *(STAT 217, 231)*
- **DATA 311** *(CPSC 471)*
- **DATA 305** *(STAT 321/323)*
- **4 DATA SCIENCE ELECTIVES**
- **DATA 501**

= **6-10 HCEs** of new content

### Biology Major + Data Science Minor
- **DATA 201**
- **DATA 211** *(CPSC 217, 231)*
- **STAT 205** *(STAT 217, 231)*
- **DATA 311** *(CPSC 471)*
- **DATA 305** *(STAT 321/323)*
- **4 DATA SCIENCE ELECTIVES**
- **DATA 5xx**

= **5-9 HCEs** of new content

### CPSC Major + Data Science Minor
- **DATA 201**
- **DATA 211** *(CPSC 217, 231)*
- **STAT 205** *(STAT 217, 231)*
- **DATA 311** *(CPSC 471)* [If you take 471]
- **DATA 305** *(STAT 321/323)*
- **4 DATA SCIENCE ELECTIVES**
- **DATA 5xx**

= **7 Courses** of new content

### STAT Major + Data Science Minor
- **DATA 201**
- **DATA 211** *(CPSC 217, 231)*
- **STAT 205** *(STAT 217, 231)*
- **DATA 311** *(CPSC 471)*
- **DATA 305** *(STAT 321/323)*
- **4 DATA SCIENCE ELECTIVES**
- **DATA 5xx**

= **7 HCEs** of new content

*Some Data Science Electives may also fulfill breadth requirements for a non-CPSC or MATH/STATS student’s major program.*
CPSC 217/231
"Programming with Data"
(Cannot be taken for credit after CPSC 217/231)

CPSC 219/233
CPSC 319/331
CPSC 471
Most CPSC courses

DATA 311
"Data Processing & Storage"
(Cannot be taken for credit after CPSC 331)

DATA 211
Small Subset of CPSC courses
ADMISSION

SCIE 201 / DATA 201 “THINKING WITH DATA”
• Offered Now (Including W2018)
• Open to all students / no prerequisites

MINOR (starting 2018)
• Competitive Admission
• Minimum 3.0 GPA over last 15 units (last 5 courses)
• Apply through Undergraduate Science Centre
FOR MORE INFO

Undergraduate Science Centre

Program Website
www.ucalgary.ca/data-science/

Program Directors

Ehud Sharlin
Computer Science

Jim Stallard
Math and Stats

Data science is the application of computing, statistics, and analytic sciences to address data-driven challenges and acquire a better understanding of the world in which we live.

Why Data Science?
Big data is already part of your life. You deal with it every day. Being able to process data will give you power for your studies, your research, your future business and everyday life. No matter what you choose to explore, understanding data is a desirable and essential skill for your future. Programs in Data Science at the University of Calgary will introduce you to the fundamental tools, techniques, and mindsets necessary to succeed in today's data-driven world.

Data Science Programs at Calgary
Calgary offers a number of options for students interested in building expertise in data science.

Take Courses on Data Science Topics.
Courses like SOE 201 (Thinking with Data) offer a practical, hands-on introduction to data-driven thinking.

Earn a Minor in Data Science.
Starting in 2018, the University of Calgary's Data Science Minor will offer undergraduate students a

Pursue a Degree in a Related Program.
Calgary's departments of Mathematics & Statistics and Computer Science offer both