留学報告書 Part II

UCバークレー統計学科 PhD2年
石原みやび
miyabishihara(at)gmail.com

こんにちは！

カリフォルニアの夏休みはサンタクルーズのビーチやヨセミテ国立公園でバケーションを楽しむ！...というわけではなく、バークレー統計学科のほとんどの学生は研究、Teaching Assitantかインターンシップをします。私は、後期の授業が5月に終わると2つの研究をスタートしました。1つはアメリカの豪雨のデータをもとに時空間解析、もう1つは衛星画像データをもとに植生や夜間光の空間分布を監視するプロジェクトです。というわけで6月〜8月はほとんどキャンパスで過ごしました。そんな中でも7月下旬には、バンクーバーで開催された統計学会に参加し、バークレーに帰る途中、アムトラック列車でシアトルとポートランドに立ち寄りました。シアトルではワシントン大学のキャンパスを散策したり、ポートランドではハニーラベンダー味のアイスクリームを食べながら街歩きを楽しみました。

今回の報告書では、1年目後期の活動の中で特にみなさんとシェアしたいコースワーク、Hadley Wickhamの講演会、キャンパス周辺のお店などについてお伝えします。Hadley Wickhamは、R言語の統合開発環境を提供する RStudio のチーフサイエンティストで、統計分析を行うためのパッケージを開発しています。

まだまだ続くコースワーク！

PhDの1年目は、統計学科の基礎科目である確率論、理論統計、応用統計の全7クラスから年間最低3クラスを履修することになっています。前期は理論統計Aと応用統計Aの2クラスを履修し、課題に悪戦苦闘しました。それにも関わらず後期もまた理論統計Bと応用統計Bを履修し、新たな地獄を見ることになりました。

前述の2クラスに加えて、データサイエンスのプログラム DS421の必修授業、Reproducible and Collaborative Data Scienceも履修しました。異なる分野の人がチームアップして進めていくプロジェクトでは、データ分析の行程を読みやすくプログラムを書くことが大事です。計算スピードは速いけれど短すぎて理解しずらいとか重複だらけのプログラムは “silent failure”と言われることもあります。授業では、tidyverse という可読性を重視したプログラムの概念をもとに、データの分析をするトレーニングをしました。Tidyverse の概念は、Hadley Wickhamが提唱したものです。Tidyは整頓、verseは韻文という意味です。

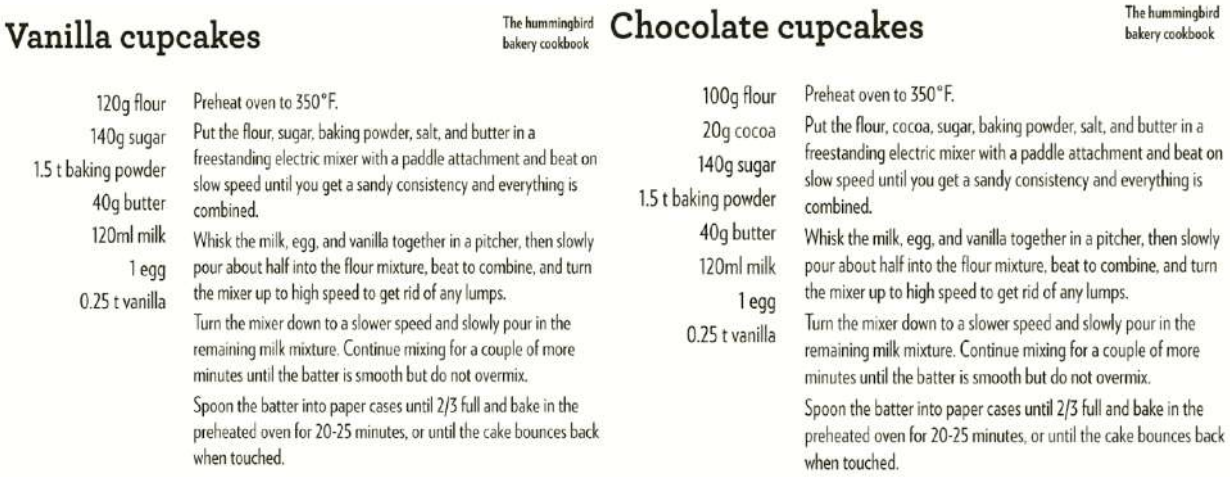
写真の説明　5月のキャンパス．統計学科のあるEvans Hallと時計台のSather Tower．右上は音楽学部の建物の横に咲いているヤマツツジです．

Hadley Wickhamの講演

カップケーキのレシピ本から学ぶ、読みやすいプログラムの書き方

Hadleyはテキサスに在住していますが、今学期はスタンフォード大学で教えるためベイエリアに滞在していました。その間、ベイエリアを中心に講演を行なっていて、私は4月にバークレー、5月にサンフランシスコで開かれた講演会に参加しました。サンフランシスコの講演は、map関数という反復処理を行う関数についてでした。反復処理を行う関数はfor文など他にもあるのですが、その中でもmap関数は特に可読性が高いということについて、カップケーキのレシピを例に説明してくれました。ここでは講演の一部を紹介します。

Hadleyは、まずバニラカップケーキとチョコレートカップケーキのレシピを順にスライドに映し、会場に問いかけます ”What’s the difference?”



Vanilla cupcakes

The hummingbird bakery cookbook

120g flour
140g sugar
1.5 t baking powder
40g butter
120ml milk
1 egg
0.25 t vanilla

Preheat oven to 350°F.

Put the flour, sugar, baking powder, salt, and butter in a freestanding electric mixer with a paddle attachment and beat on slow speed until you get a sandy consistency and everything is combined.

Whisk the milk, egg, and vanilla together in a pitcher, then slowly pour about half into the flour mixture, beat to combine, and turn the mixer up to high speed to get rid of any lumps.

Turn the mixer down to a slower speed and slowly pour in the remaining milk mixture. Continue mixing for a couple of more minutes until the batter is smooth but do not overmix.

Spoon the batter into paper cases until 2/3 full and bake in the preheated oven for 20-25 minutes, or until the cake bounces back when touched.

Chocolate cupcakes

The hummingbird bakery cookbook

100g flour
20g cocoa
140g sugar
1.5 t baking powder
40g butter
120ml milk
1 egg
0.25 t vanilla

Preheat oven to 350°F.

Put the flour, cocoa, sugar, baking powder, salt, and butter in a freestanding electric mixer with a paddle attachment and beat on slow speed until you get a sandy consistency and everything is combined.

Whisk the milk, egg, and vanilla together in a pitcher, then slowly pour about half into the flour mixture, beat to combine, and turn the mixer up to high speed to get rid of any lumps.

Turn the mixer down to a slower speed and slowly pour in the remaining milk mixture. Continue mixing for a couple of more minutes until the batter is smooth but do not overmix.

Spoon the batter into paper cases until 2/3 full and bake in the preheated oven for 20-25 minutes, or until the cake bounces back when touched.

この2つのレシピの違いはココアパウダーを入れるか入れないかだけで、それ以外の材料と手順は全く同じです。しかし、この2つの違いを見つけるためにはそれぞれの材料と手順をすべて読む必要があります。ちなみに、このレシピ本には他に15種類ほどのカップケーキのレシピが載っているのですが、どれも材料と手順は基本形のバニラカップケーキとほとんど同じです。このレシピは重複が多すぎて “It is hard to see the theory of the cupcakes”とHadleyは指摘します。

そこで、材料と手順を整理し、可読性が高まったレシピを紹介します。左に共通の手順、右にカップケーキ別の材料を書きます。そのあと、薄力粉、グラニュー糖、ベーキングパウダーを dry ingredients、牛乳、卵、バニラエキス を wet ingredients とグループ分けします。レシピが簡潔になったことで、チョコレートカップケーキは薄力粉20gをココアパウダーで代用できると一目でわかるようになります。

Cupcakes

	Vanilla	Chocolate
Beat dry ingredients + butter until sandy.	120g flour	100g flour
		20g cocoa
Whisk together wet ingredients. Mix half into dry until smooth (use high speed). Beat in remaining half. Mix until smooth.	140g sugar	140g sugar
	1.5t baking powder	1.5t baking powder
	40g butter	40g butter
Bake 20-25 min at 170°C.	120ml milk	120ml milk
	1 egg	1 egg
	0.25 t vanilla	0.25 t vanilla

ここで大切なのは、ムダを省きエッセンスを抽出したレシピによって、カップケーキ作りの基本形の材料と手順が明確になり、甘くてふっくらとしたカップケーキの”本質”（theory）を見極めるために必要な環境が整ったことです。その結果、レシピの応用や展開へのアイデアが出やすくなります。

たとえば、チョコレートの風味を強くするためには、薄力粉を少なくしてより多くのココアパウダーを使うと良いのではないかとか、あるいはもっと極端に、つまり薄力粉を一切使わないで全部ココアパウダーだけでカップケーキを作ったらどうなるかとか、エスプレッソパウダーを牛乳でといて新メニュー”エスプレッソカップケーキ”を作るなど、実験感覚でいろいろ試してみたくなります。もっとも全部ココアパウダーにしたら、カップケーキがうまく膨らまなさそうですが。

さて、カップケーキのレシピを簡潔にするための手順とそのメリットを理解したところで、for文について考えてみましょう。Hadleyはスライドに2つのfor文を映し出し、会場に問いかけます “What’s the difference between the two for loops?”

```
out1 <- vector("double", ncol(mtcars))
for(i in seq_along(mtcars)) {
  out1[[i]] <- mean(mtcars[[i]], na.rm = TRUE)
}

out2 <- vector("double", ncol(mtcars))
for(i in seq_along(mtcars)) {
  out2[[i]] <- median(mtcars[[i]], na.rm = TRUE)
}
```

このプログラムの違いは、データの各列の平均値を計算しているか、中央値を計算しているかですが、その他のコードが全く同じなので、前述のバニラとチョコレートカップケーキのレシピと同様、両者の違いがすぐには分かりません。しかしここで、map関数を使うことによって、for文のプログラムと同じ処理を以下のような簡潔なプログラムで行うことができます。

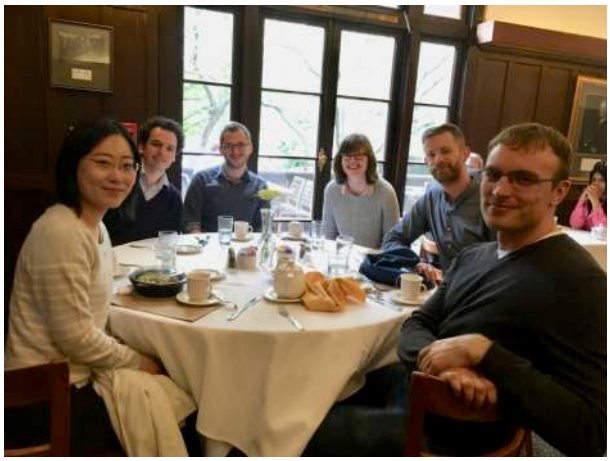
```
out1 <- map_dbl(mtcars, mean)
out2 <- map_dbl(mtcars, median)
```

このプログラムだと重複が減り、可読性の高いプログラムになりました。このようにHadleyはmap関数の他の利点も料理ネタで説明したので、まるで料理教室に参加しているような気分で講演を聞くことができました。

伝えたい情報を短く、わかりやすく伝えるということはプログラムのコンセプトです。しかし、この当たり前と思えるコンセプトは、機械にとってわかりやすい（計算しやすい）のか、人間にとって理解しやすい（行程が理解でき追加修正がしやすい）のか、といった対象によって違ってきます。特にチームプロジェクトでデータ分析を担当するとき、メンバー間で共通認識を持てることや分析の引き継ぎでバトンパスしやすいことを念頭に、プログラムすることが大事だと再認識しました。

map関数の概念・理念は、データビジュアリゼーションや読みやすい文章を書く時、その他さまざまな場面に応用できる考え方だと気づきました。

今回の講演は、専門的な内容を聞くだけでなく、大学内外のデータ分析の現場を知っているHadleyの人となりや人生哲学にも触れることができ、とても充実した時間でした。興味のある方は講演動画をチェックしてみてください。（トークは29:00から、map関数についての説明は1:16:50から始まります）



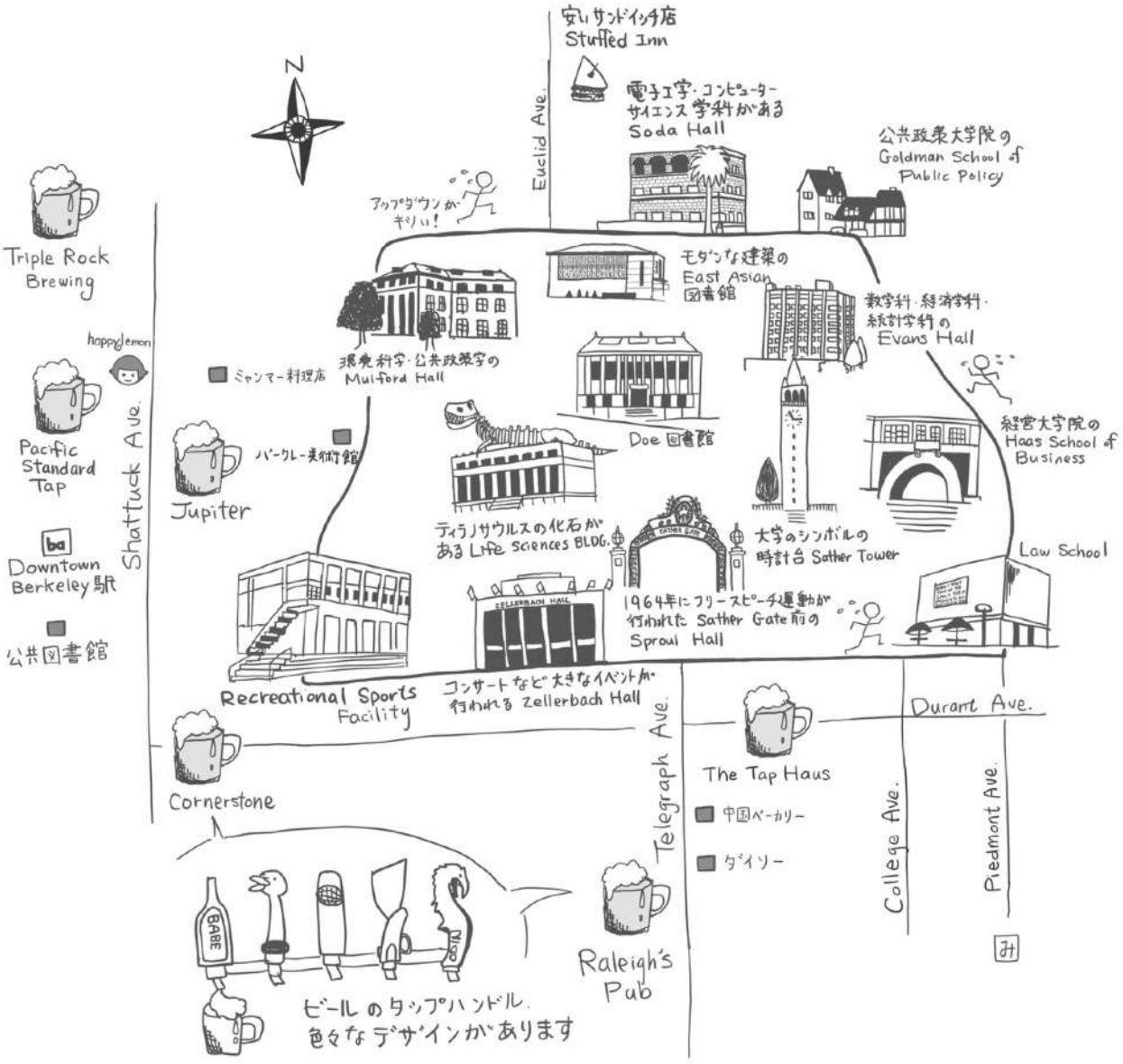
講演のためバークレーに訪れたHadley（右から二番目）と統計学科の学生と学内のレストランでランチをする機会がありました。Hadleyが教えていたスタンフォード大学の輪読授業 Readings in Applied Data Scienceは「データを扱う倫理など、大学の授業ではあまり取り扱われていないけれど大事だと思うトピックスを取り上げるよう心がけた」と話してくれました。ちなみにこの輪読授業の教材とリーディングリストはgithubに公開されています。

学生がよく行くお店

統計学科では毎週金曜日に食事会、通称 wind down を行います。Wind downはネジを緩める、休息するという意味です。毎週、ソーシャル担当の学生がお店をメーリングリストで知らせてくれます。よく行く場所は、食事とビールが楽しめるパブです。バークレーではチェーン店の出店が制限されていて、長年営業している中小の飲食店が多くあります。ビールの醸造所も多く、パブには20種類以上の地元のクラフトビールがあります。時には醸造家がパブにやってきて自家製のビールをプロモートし、一杯目を無料で提供することもあります。私は、ビールはあまり飲まないのですが、醸造家がすすめてくれたウォーターメロンビールは淡い琥珀色でさっぱりとした味わいでした。

サンドイッチ屋さんも多くあります。私のお気に入りはキャンパスの北にある Stuffed Inn です。パンは7種類（hard sour french roll, whole wheat, light or dark ryeなど）、サンドするものやドレッシングも数種類の中から選ぶことができます。中にアボカドやアルファルファがたっぷり入っている Normal Ned's Avocado Sandwich が$5.4ドルというのは物価の高いバークレーでは珍しくコスパが良く、学生のお財布にやさしいお店だと思います。

7月にはダウンタウンにラーメンの「一風堂」がオープンし、行列が絶えない様子ですが、実はMarufukuのラーメンはお手頃の価格で美味しいということを聞きました。近いうちにラーメン好きの統計学科の学生と検証に行く予定です。



UCバークレーの大まかなキャンパスマップと学生の行きつけのパブ．理論統計と応用統計の授業はEvans Hall，データサイエンスの授業はティラノサウルスの化石があるLife Sciences Buildingで行われました．衛星画像の研究はキャンパスでもっとも急勾配な場所に位置する Public Policyの建物で行いました．キャンパスの左はベイでそのさらに左下がサンフランシスコ、右は丘の上にローレンス・バークレー国立研究所があります．

バークレーあるいはベイエリアに来る機会があればぜひご連絡をください。