

週末サイエンティスト のススメ

2016-09-21 PyConJP2016

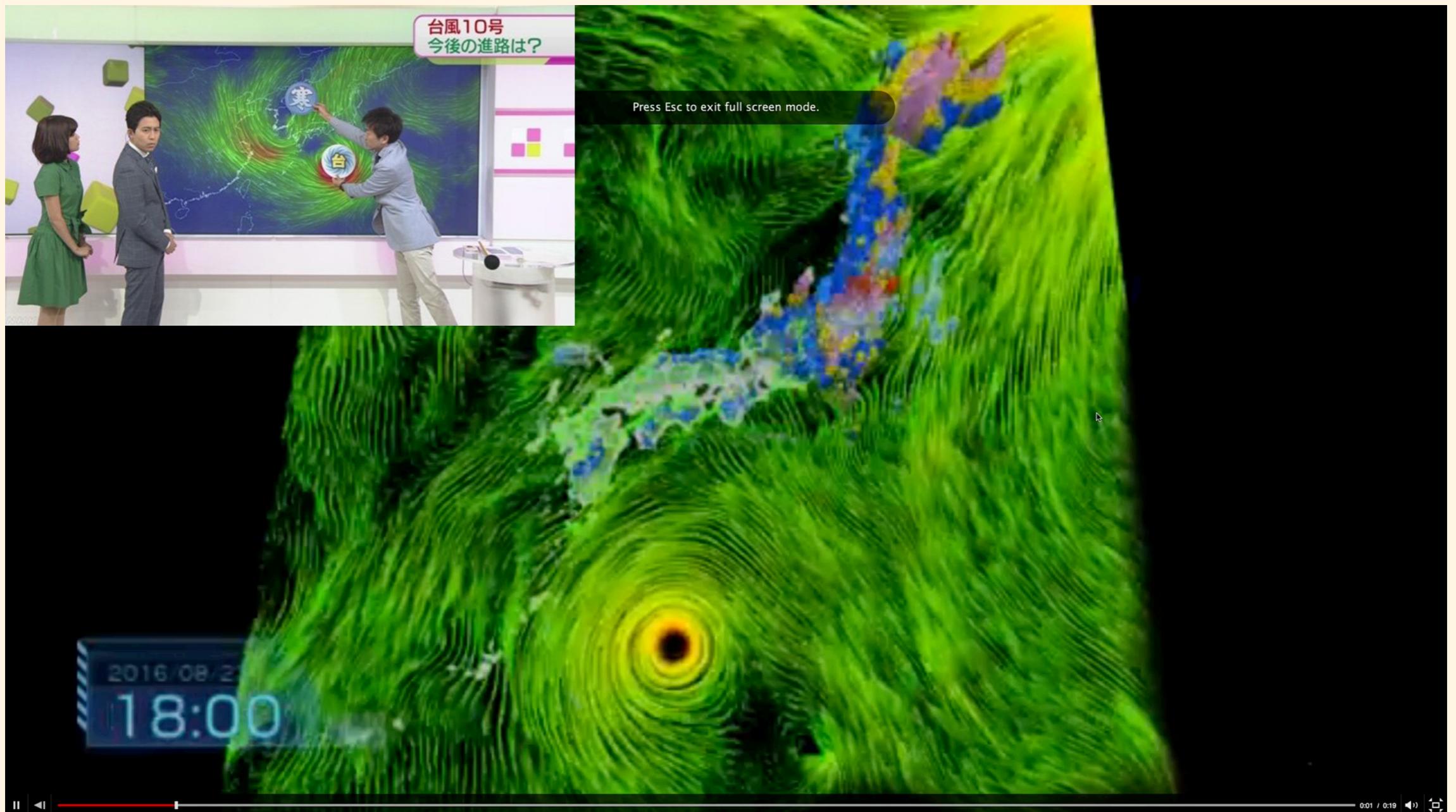
Yuta Kashino

- BakFoo, Inc. CEO
- Astro Physics /Observational Cosmology
- Zope / Python
- Realtime Data Platform for Enterprise



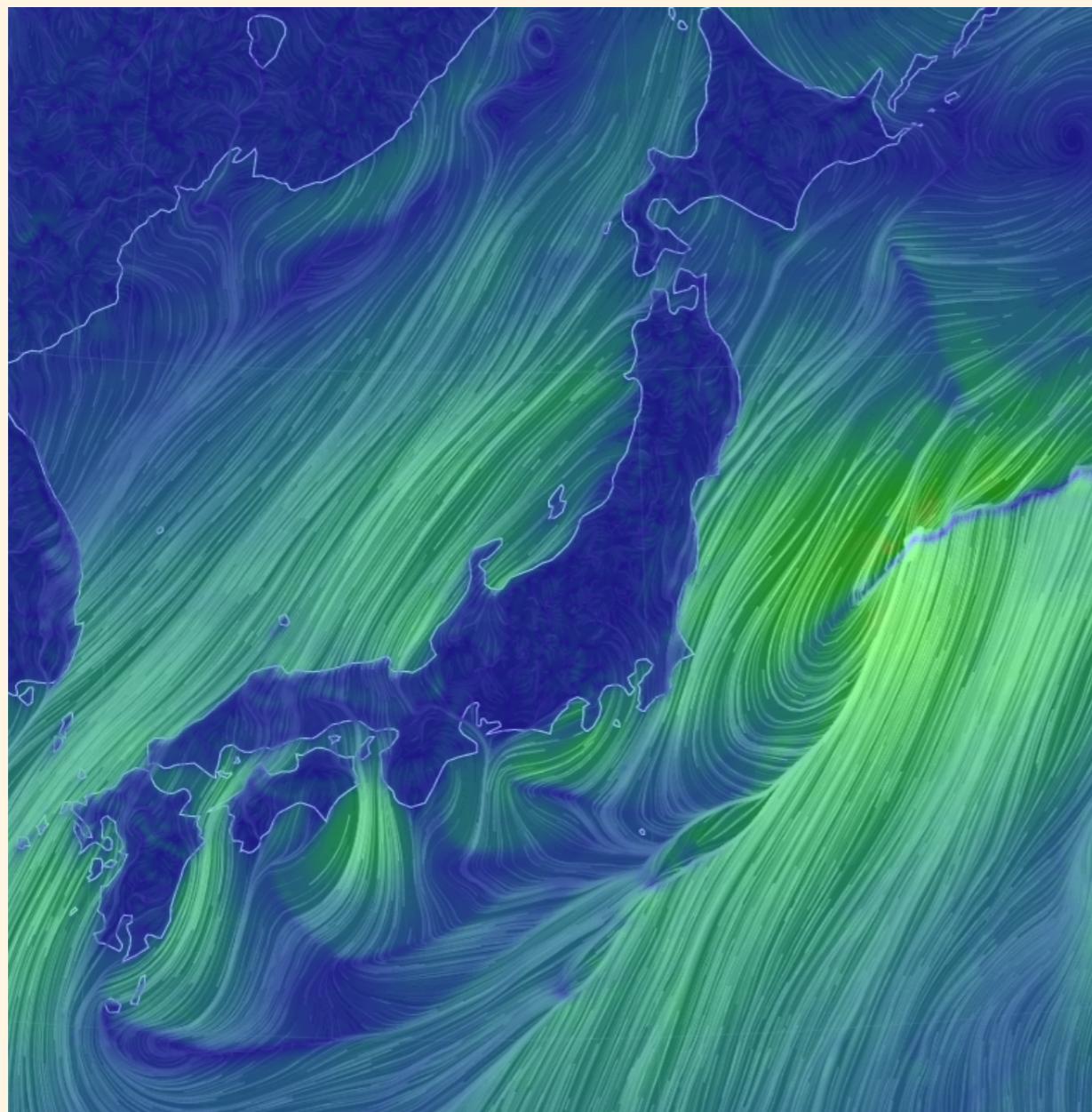
BakFoo, Inc.

NHK NMAPS: リアルタイムデータ + 可視化



BakFoo, Inc.

webでの気象データのリアルタイム表示



<http://bluethunder.bakfoo.com:8080/>

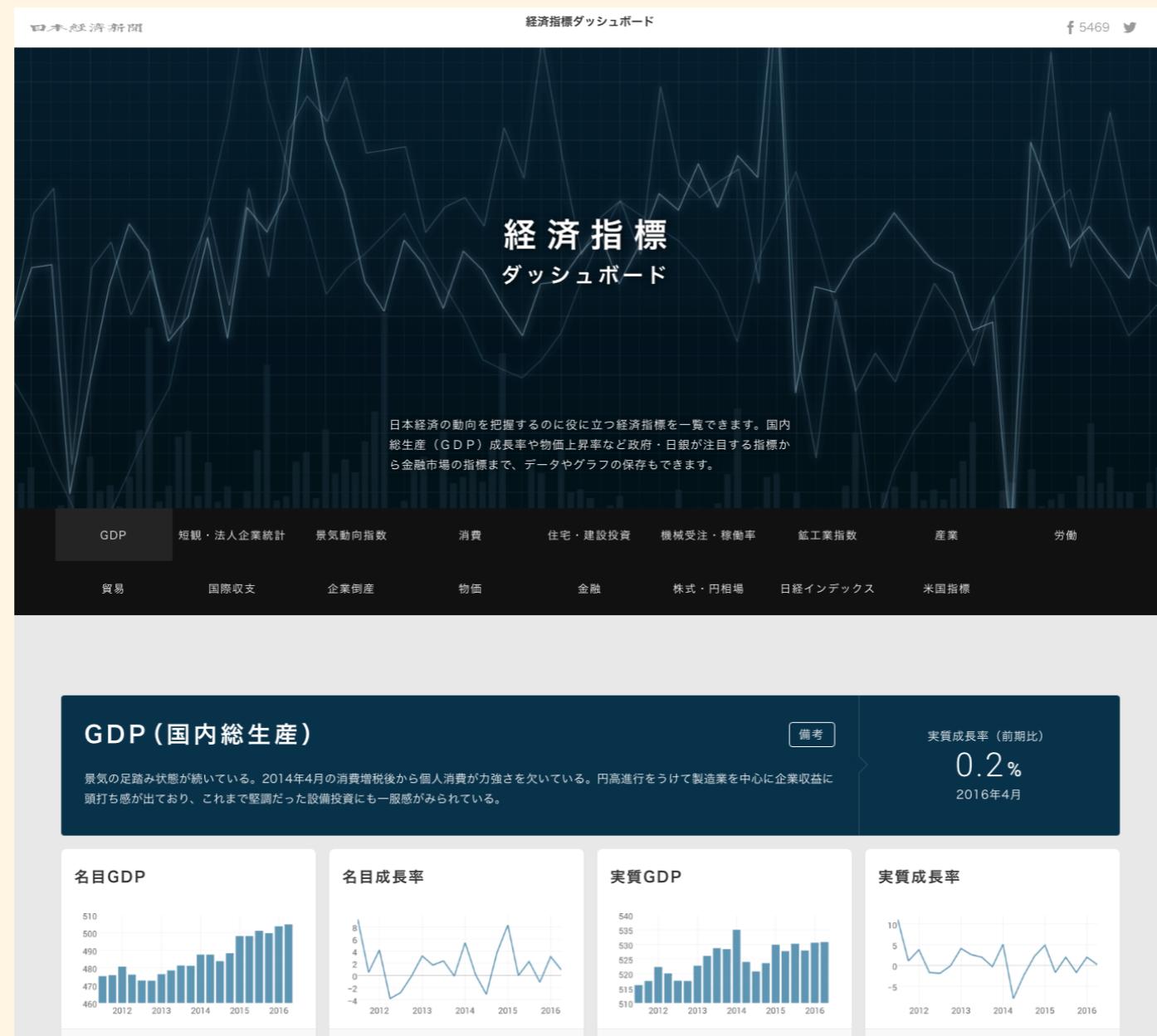
オープンデータ+マイクロサービス



PyConJP 2015
日本のオープンデータ
プラットフォーム
をPythonでつくる

BakFoo, Inc.

日本経済新聞社 経済指標ダッシュボード



BakFoo, Inc.

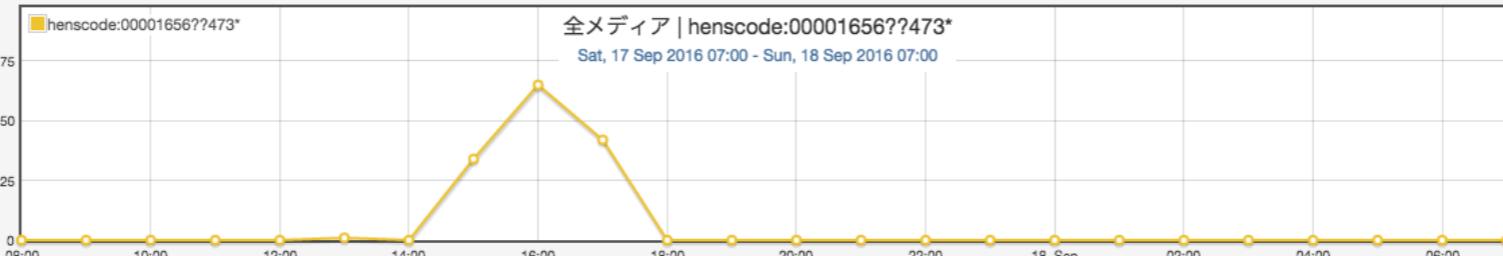
テレビ番組判定: SNS + 自然言語解析

編成コード検索: 1656473 | 472 | 474 |

番組情報		Tweet Counts	Shared Counts
番組 大相撲秋場所 七日目 ▽ゲスト 原沢久喜選手（リオ五輪 柔道 銀メダリスト）	OnAir 106	Twitter 106	
概要 ▽ゲスト 原沢久喜選手（リオ五輪 柔道 100キロ超級 銀メダリスト）（4:10）「幕内取組」【ゲスト】原沢久喜、【解説】正面（幕内）舞の海秀平	Total 145	facebook 5	
放送時間 2016年09月17日 16:00:00 ~ 2016年09月17日 18:00:00	Avg/min 0.87	hatebu 0	
	Max/min 6	PC NA 0 5	
		Mobile NA 0 0	
		OnDemand NA NA NA	

Keywords	
大相撲秋場所 88 , sumo 74 , 原沢久喜 51 , ゲスト 50 , リオ五輪 49 , 銀メダリスト 49 , 中の人 46 , リプライ 46 , 幕内取組 34 , アナウンサー 30 , 大坂敬久 28 , 舞の海秀平 28 , nhk総合 20 , ツイート 12 , 総レス 12 , アナログラジオ 7 , 大相撲 7 , 今日は 7 , 西岩親方 6 , 若の里 6 , 舞の海 5 , 大相撲中継 5 , 豊真将 4 , 手拍子 4 , ラジオ 4 , 松鳳山 3 , 激アツ 3 , 盛り上がり状態 3 , 千代の国 3 , 朝青龍 3 , 隠岐の海 3 , 向正面 3 , 戸部眞輔 2 , 春日野 2 , 固形シャンプー 2 , 日馬富士 2 , 決まり手 2 , シャンプー 2 , リオオリンピック 2 , 伊勢ヶ濱部屋 2	Twitter 106

全メディア | henscode:00001656??473*
Sat, 17 Sep 2016 07:00 - Sun, 18 Sep 2016 07:00



舞の海さんリメールの話しかしないな~(笑) #sumo #nhk
2016年09月17日 17:35:02
かなぶん
@knzwysnr OA

HOT実況 【137人が実況中】大相撲秋場所 七日目 ▽ゲスト 原沢久喜選手（リオ五輪 柔道 銀メダリスト）【勢い:37tw/分】livetter.com/tv/ch1/ 2016年09月17日 17:34:01
#nhk #sumo HOTテレビ実況なう☆Livetter @livetter_hot OA

「稀に勢いの出る大関」の「稀」の時期かな #sumo #nhk 2016年09月17日 17:32:32
さみすたすとりーと@沼津より愛を込めて @sazankuwata OA

お稀勢さん勝ちました。 #nhk #sumo 2016年09月17日 17:32:21
かすみ猫 @ich_kasumi OA

万全ではないが勝った #NHK #sumo 2016年09月17日 17:32:20
たにま @tanimax_GoS OA

BakFoo, Inc.

- ・国政選挙におけるSNSユーザの属性分析
- ・SNSデータのトピックモデル解析
- ・質問票データの機械学習と異常検知.



週末サイエンティスト と週末研究

週末サイエンティスト

- 知的好奇心をもって
- 勤務外や週末に
- 科学的な解析や分析、推測や予測を
- 自分の手で行い
- 研究する。

週末サイエンティスト

- Publish or perishの義務無く
- 隙間時間でできるセットアップで
- 必要に応じて勉強をしながら
- 長期間継続する.

週末研究が可能な背景

- 三つの理由
- 安価な計算リソース
- オープンアクセス
- オープンソース.

背景1：安価な計算リソース

- クラウド・専用サーバの普及
 - クラウド
 - 専用サーバ/GPU
- 安価な高速回線
- 仮想化技術の興隆



- vagrant
- docker



背景2: オープンアクセス

- arXiv・オープンアクセスジャーナル



- オープンデータ 

- 技術ブログ・カンファレンスビデオ 

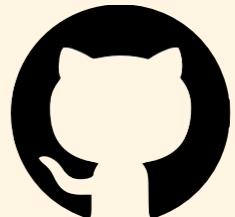
- 勉強会・ミートアップ  dots. 

- MOOC・講義ビデオ  edX

- 計算機競技コンテスト 

背景3: オープンソース

- GitHub : ソフトウェア開発の生態系



- Linux : 生きる術としてのプラットフォーム

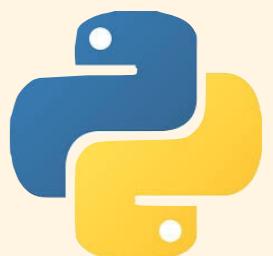


- 多くの専門家や研究者が自分の仕事や教育のためにライブラリを書き始めた

- NeXTの夢 <https://en.wikipedia.org/wiki/NeXT#Background>



- Pythonの力 (障壁の低さ, スケールする言語設計, 豊かなエコシステム)



週末研究とは

- 論文を読んで研究する
- 科学的方法論に沿って研究する
- 研究リテラシーをもって研究する
- 本業でなく勤務外や週末に行う偉大な道楽
- 道楽には惜しみなく時間とお金を注ぐ。

週末研究とは

本トークのスコープ

論文を読む

研究する（計算機実験・解析）

先行研究

研究リテラシーがある

実験・解析手法の理解

研究動向

科学スタッフの理解

専門知識を持っている

プログラミング技術

計算機環境の整備

基礎的な数学・統計学の知識を運用できる

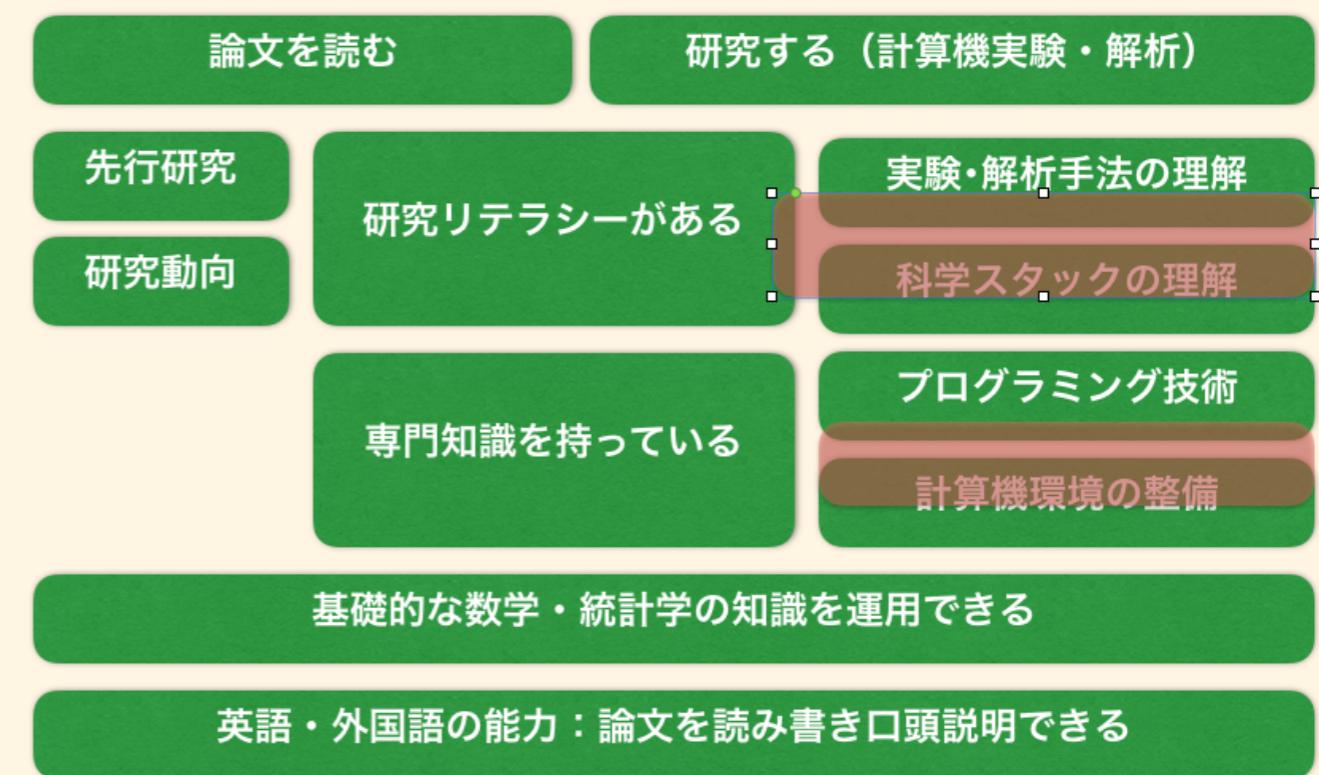
英語・外国語の能力：論文を読み書き口頭説明できる

研究リテラシー

- 研究アイデアを思いつき育てる
- 論文を探すこと
- 理解できないときは足りない知識を自学自習する
- 科学的方法論に沿って研究を進める
- 他者が再現できるように成果の手続きが提示できる
- 捏造や剽窃を行わない。

本トークのトピック

- 計算機環境の整備
- 科学スタックの理解と活用
 - Python科学スタックの生態系
 - 具体的な利用のしかた
- 研究リテラシー
 - 論文の探索
 - 自学自習による知識の積み上げ
 - 科学的方法論に沿う.



週末サイエンティストまとめ

- ・ 週末研究が可能になった
- ・ 論文を読んで科学的方法論で研究する
- ・ 本トークのトピック
 - ・ 計算機環境の整備
 - ・ 科学スタックの理解と活用
 - ・ 研究リテラシー

計算機環境の整備

計算機環境を巡る問題

1. マシンをどうする
2. OSどうする
3. Pythonをどうする
4. 実験環境をどうする
5. コード保全をどうする

1. マシンをどうする

- 選択肢
 - ノートパソコン+仮想環境
 - 自宅デスクトップ+外部アクセス
 - クラウド・専用サーバをホスティング業者に借りる
- 要望
 - 隙間時間にアクセスしたい
 - 計算を止めたくない
 - 最後はスケールさせたい

1. マシンをどうする：回答

- 信頼できる業者からサーバを借りる。
 - 時間と安定稼働をお金で買う大人のソリューション
 - どこからでも隙間時間にアクセスできるようにする。
- GPUの場合はAWS, さくら, Azure, Softlayer
 - さくらの高火力：超安定, コスパが良い.
 - AWS：コスパがよくない, ドライバが古い, しかし劇的なスケーリング

2.OSをどうする

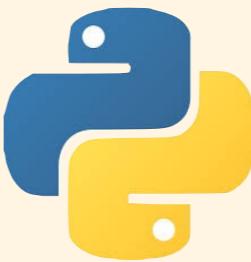
- 選択肢
 - OS X / Windows
 - Ubuntu
 - CentOS, Fedora, Debian, openSUSE
- 要望
 - パッケージ管理を楽したい
 - イケてるライブラリやフレームワークを使いたい
 - GPUドライバなどのデバイス周りで苦労したくない

2.OSをどうする：回答

- Ubuntu LTS一択



- 14.04, 16.04業界標準, ディファクト
- 科学スタックはほとんどUbuntuで開発されている
- ソースからビルドする場合も大体うまくいくことが多い
- aptによりパッケージ管理は困らない
- GPUなどのドライバもまずUbuntuから対応.



3. Pythonをどうする

- Pythonディストリビューション:

- **Python(x,y)** <http://python-xy.github.io/>
- **Scipy Superpack for Homebrew** <http://stronginference.com/ScipySuperpack/>
- **Enthought Canopy** <https://www.enthought.com/canopy-subscriptions/>
- **Continuum Analytics Anaconda** <https://www.continuum.io/downloads>

- Pythonパッケージマネージャ:

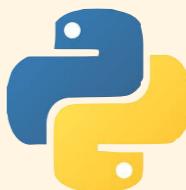
- **easy_install**
- **pip**
- **wheel**
- **Curdling**
- **conda**

- Python仮想環境:

- **virtualenv**
- **conda**

- Pythonバージョン2か3か

3.Pythonをどうする:回答

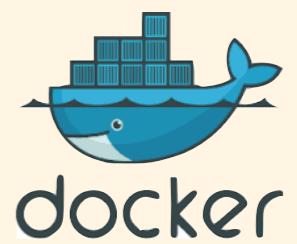


- Anaconda (**conda** + **pip**)にする
 - 現在では事実上のディファクトになった
 - **conda**だけではダメなときも. **pip**と併用
 - **conda**によるバイナリ環境整備があまりに楽
- 2/3問題：将来は3になる
 - 2でしか動かないパッケージがある
 - **conda**ならどちらにするかは大きな問題でない
 - 自分では2で書くとき, **six**や**__future__**をつかって3対応.



4. 実験環境をどうする

- 選択肢:
 - ansible, salt stackなどでホストOSを直接プロビジョニング
 - vagrant
 - docker
- 回答:
 - 状況による
 - docker/docker composeが適している場合が多いかもしれない.



5. コード保全をどうする

- ・ 時間のない身であるからこそ版管理は大事
 - test1.py, test2.py, test3.py...は禁止
 - gitはそれなりに使えるように
- ・ githubやbitbucketを使う
 - 公開前提ならgithub一択.



計算環境まとめ

- ホスティング業者からサーバを借りる
- Ubuntu LTS (14.04, 16.04)
- Anaconda: Conda + pip
- dockerを使うと便利
- githubでコード管理



科学スタッフの理解と活用

科学スタックの理解・活用

1. Python科学スタックの生態系
2. 具体的な利用のしかた

1.Python科学スタック

- 非常に豊かなエコシステム
<https://github.com/bakfoo/awesome-pysci>
- **import foovar** すぐに利用
- 研究利用の実績が十分
- ドキュメント・チュートリアル完備のプロジェクト多い
- 専門研究者が書いている.

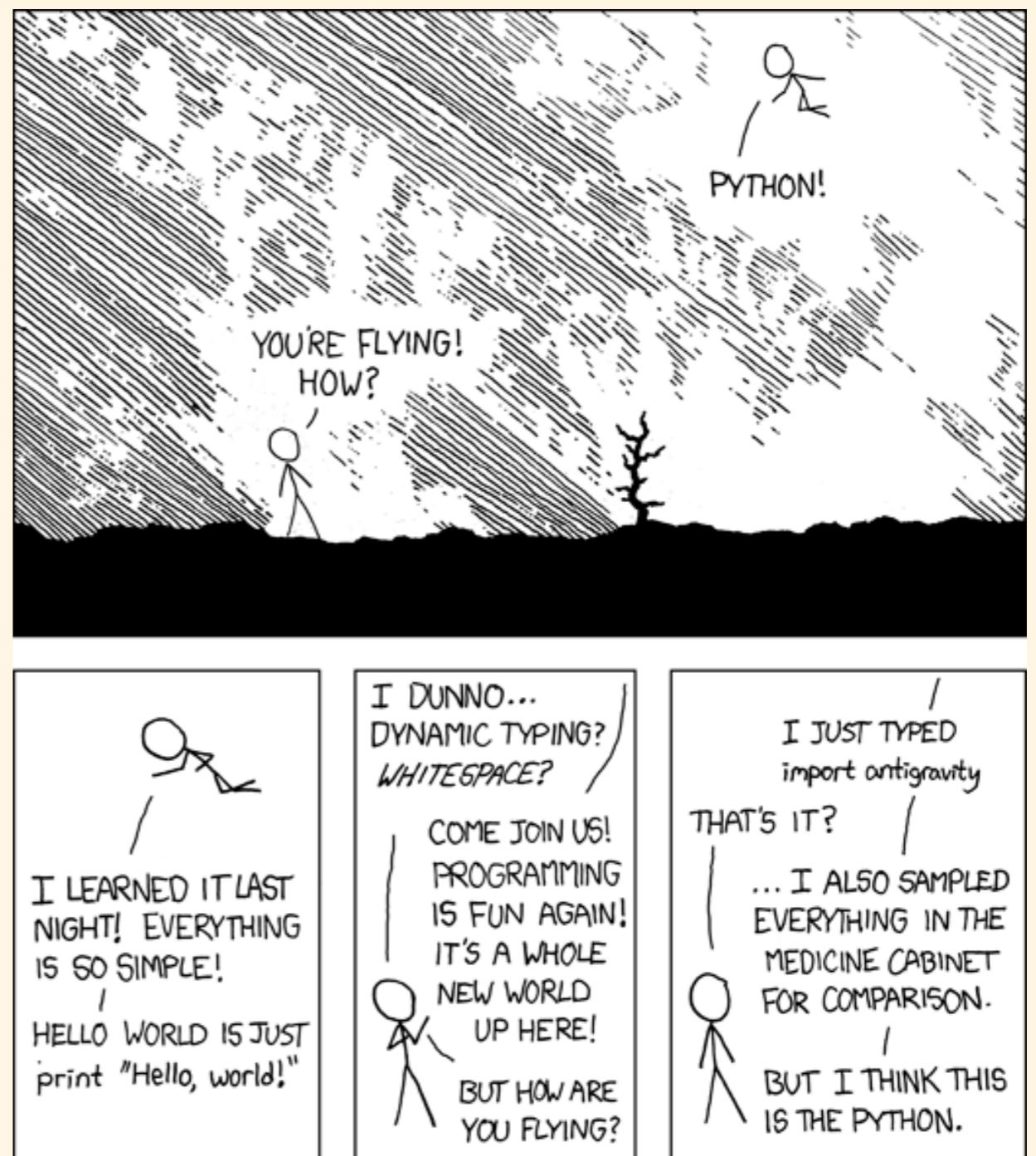
1.科学スタックの生態系

- Awesome Python科学スタック:
 - <https://github.com/bakfoo/awesome-pysci>

2. 具体的な利用方法

- import antigravity

<https://xkcd.com/353/>



事例1：気象データ可視化

事例1: 気象データ可視化

- ulmoという気象学のツールを使う
- オープンデータを手に入れて可視化する.
- デモ: https://github.com/bakfoo/pyconjp2016/blob/master/meteology/ulmo_pyconjp2016.ipynb

事例2：地震波解析

事例2: 地震波解析

- obspyという地震学のツールを使う  ObsPy
A Python Framework for Seismology
- 世界的な地震観測ネットワークからデータを入手する
- デモ: https://github.com/bakfoo/pyconjp2016/blob/master/seismology/obspy_pyconjp2016.ipynb



研究リテラシー

研究リテラシー

1. 論文の探索
2. 自学自習による知識の積み上げ
3. 科学的方法論に沿う

1.論文を探す

- ArXivと仲良くなる



- @StatMLPapers
- GitXiv / Arxiv Sanity Preserver
- trending_arxiv / 独自クローラ

- カンファレンスをチェックする

- NIPS, ICML, KDD, IVPR...
- 最近は動画を公開 videolectures.net, youtube.com

- SNSでアクティブな研究者をフォローする。

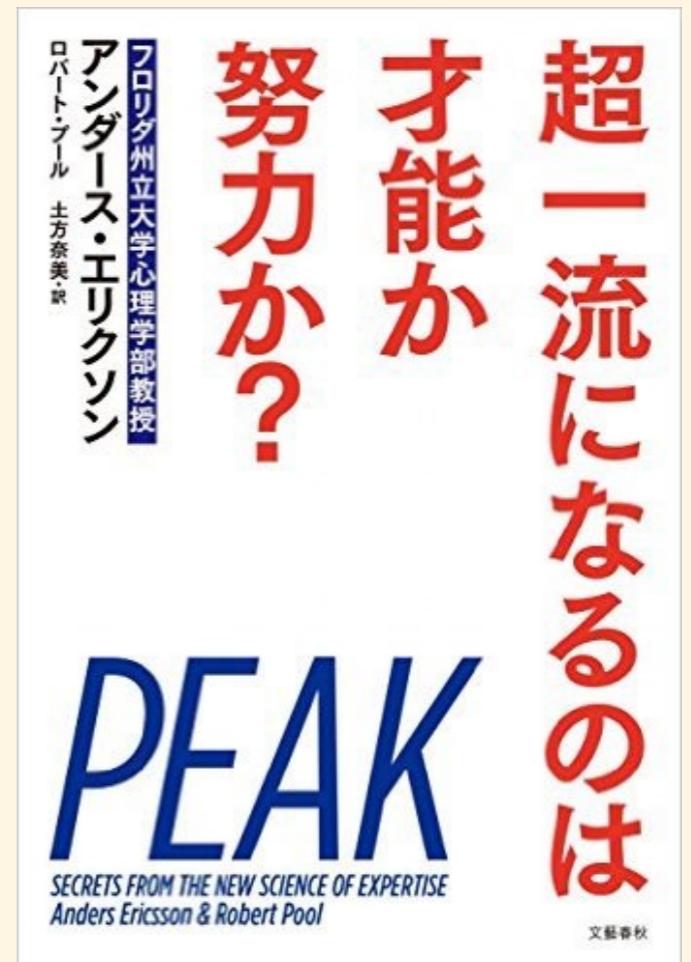


2.自学自習で知識をつける

- 英語の問題: これはどうにか解決する
- 知識の問題: 足りない知識はひたすら学ぶ
 - 勉強会  dots. ATND
 - 教科書
 - MOOC  edX
 - チュートリアルビデオ.  YouTube

2.PEAK問題

- 繙続的なdeliberate practice
 - コンフォートゾーンを越える練習
 - 目的を明確にする
 - 時間は短く常に集中(Focus)
 - フィードバック(Feedback)を受け
 - 間違いを修正(Fix)しながら
 - 飽きないよう工夫して長期間継続する
- 写すだけ、読むだけ、聞くだけは役に立たない。

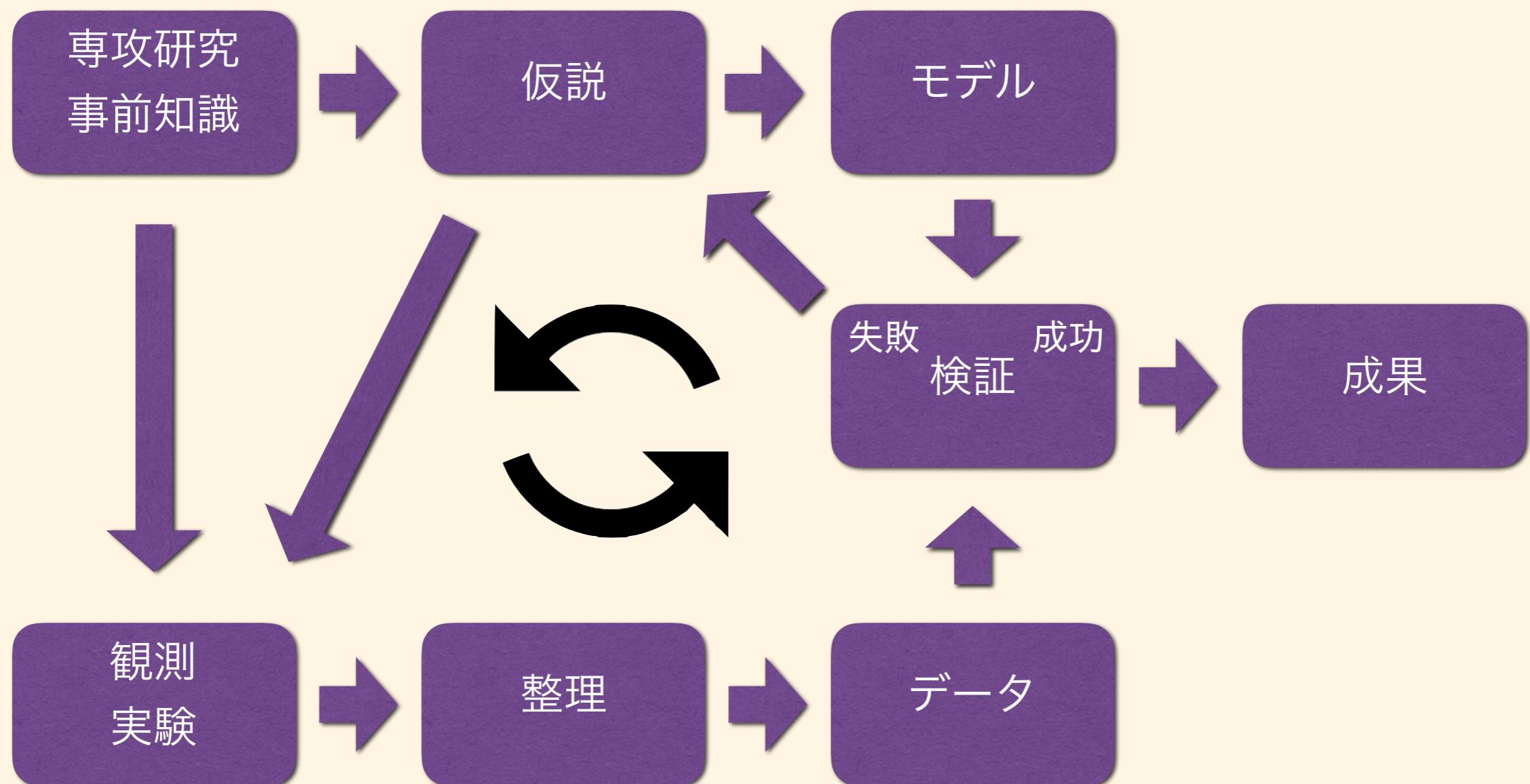


2.自学自習のポイント

- PEAKの原則に沿う
- 想起練習には効果がある
- 自分が発表する以外の勉強会は避ける
- MOOC/教科書は必ず宿題・演習をやる
- チュートリアルビデオは止めながら手を動かす。



3.科学的方法論に沿う



事例3：重力波解析

事例3: 実験装置



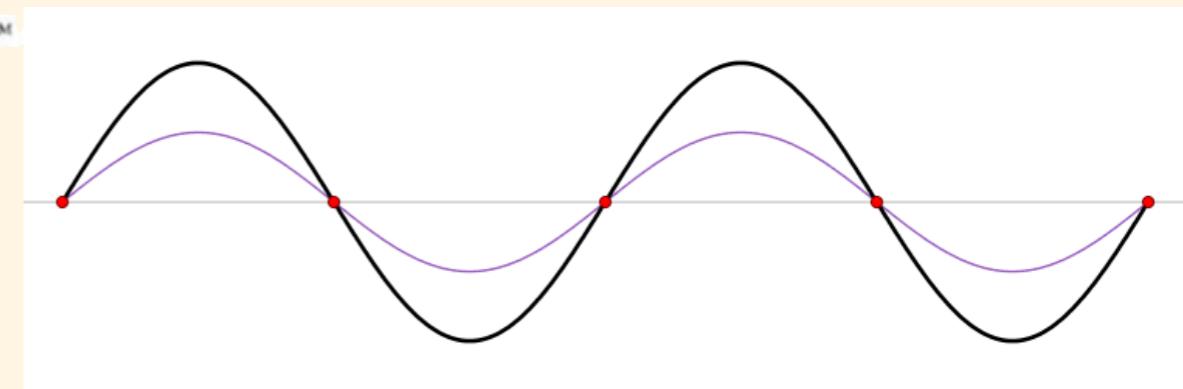
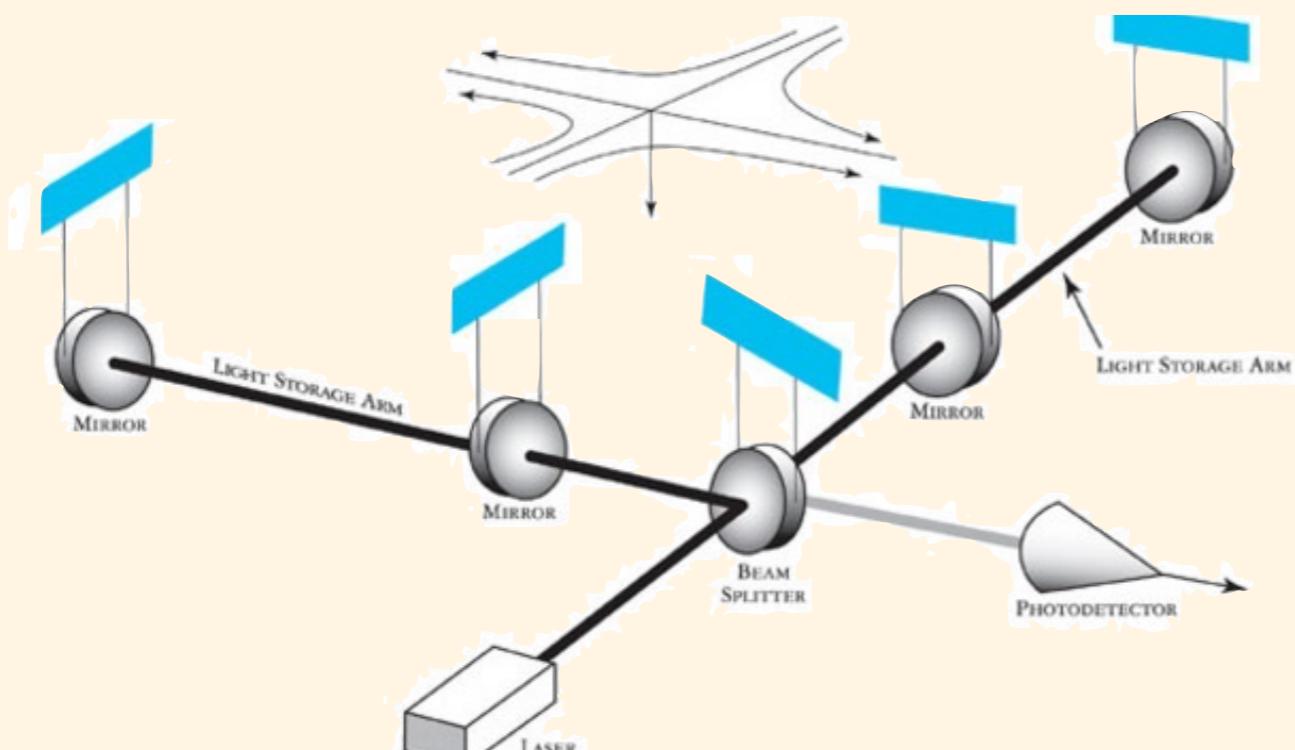
- LIGO The Laser Interferometer Gravitational-Wave Observatory
- 4kmのL字の巨大干渉計
 - Hanford(Washington)とLivingston(Louisiana)



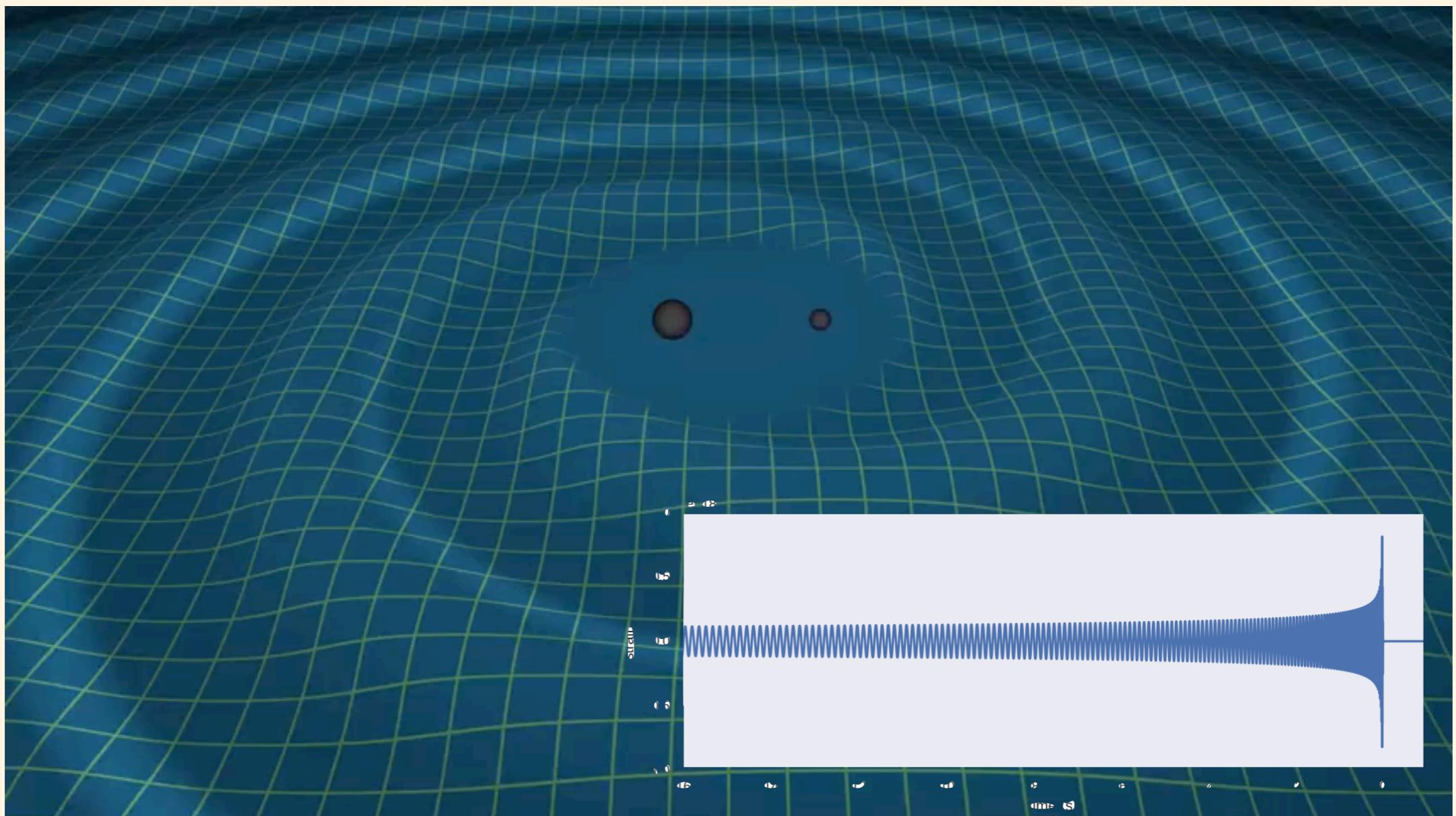
事例3: 重力波干渉計



- 重力波が到達すると光が進む距離が変わる
- 二つの光の干渉の強さの変化を観測



事例3：重力波・連星BH



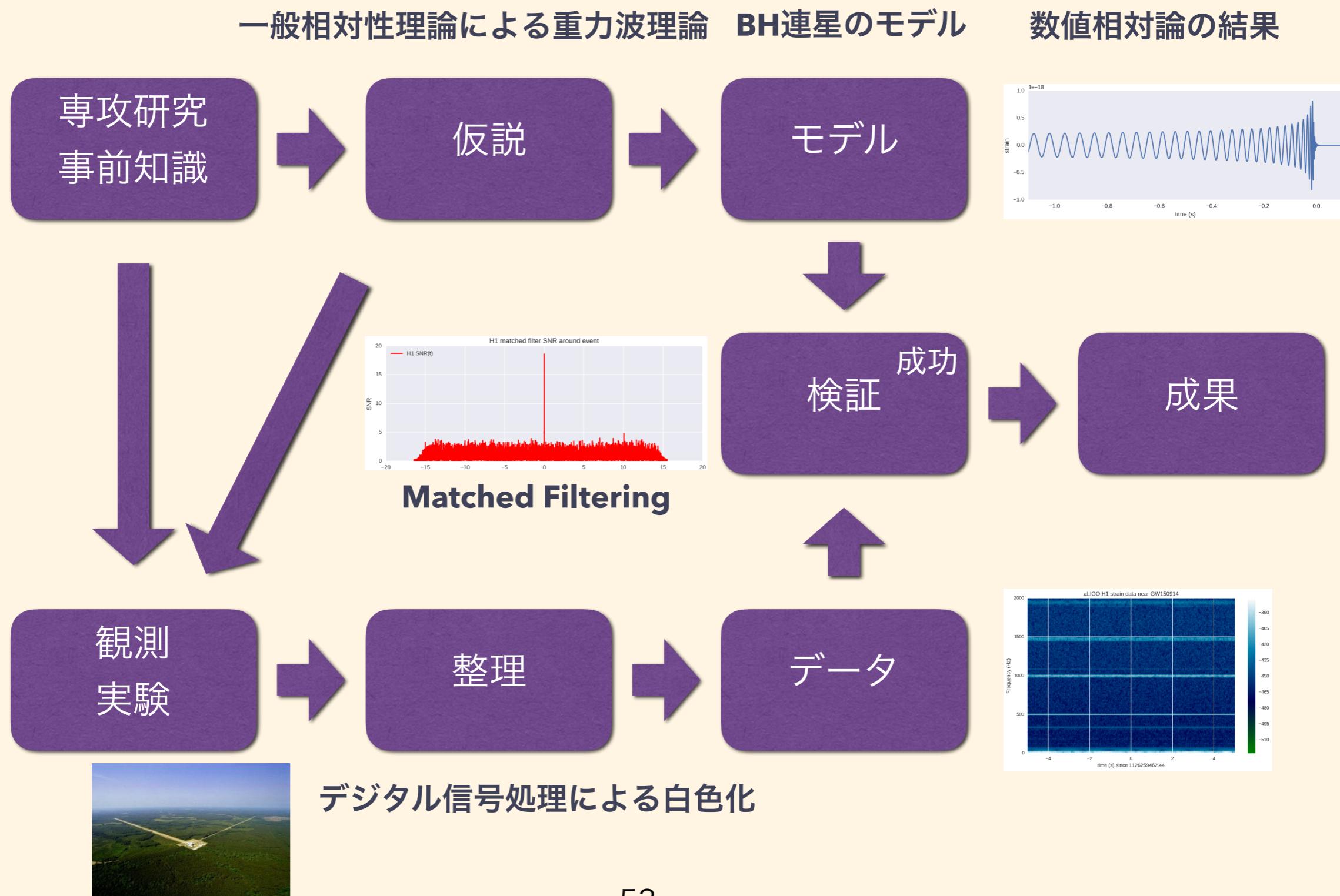
- chirp: 振幅と周波数が徐々に増大する重力波

事例3: デモ



- [https://github.com/bakfoo/pyconjp2016/
blob/master/ligo/gwave_pyconjp2016.ipynb](https://github.com/bakfoo/pyconjp2016/blob/master/ligo/gwave_pyconjp2016.ipynb)

3.事例3：科学的方法論



事例3: PySciの威力

- Python科学スタックだけで重力波解析ができる。
- LIGOではPython科学スタックは準備的・試行錯誤的な用途にもちいていてる。
- 本格的大量のデータを処理するにはLIGOのC++ライブラリを利用している。

事例3: 重力波解析



- 2015年9月14日に人類史上初めて重力波が観測された。
- 太陽質量36と29のBHが合体して62のBHになった。 $36+29 = 65$
- 観測は雑音だらけの環境で4kmの腕を $10^{-18}m$ の精度で測定する神業

まとめ

- ・週末サイエンティストが実現できる時代に
- ・そのための条件を示した。
 - ・計算機環境の整備
 - ・Python科学スタックの理解と活用
 - ・研究リテラシー

Questions

kashino@bakfoo.com
[@yutakashino](https://twitter.com/yutakashino)