Welcome to General Assembly

- **WiFi** GA Guest
- **Password** yellowpencil

GENERAL ASSEMBLY

DATA SCIENCE

DAT11SYD

Lesson 15: Real World Data Science Skills

-Part1 - Cloud Computing

-Part2 - “Soft Skills”

Course Plan

Date	Class	Lesson	Who
Monday, 19 February 2018	Lesson 1	Introduction to Data Science	Paul
Wednesday, 21 February 2018	Lesson 2	Elements of Data Science	Paul
Monday, 26 February 2018	Lesson 3	Data Visualisation	Paul
Wednesday, 28 February 2018	Lesson 4	Linear Regression	Paul
Monday, 5 March 2018	No Class	* Paul ill	.--
Wednesday, 7 March 2018	Lesson 5	Logistic Regression	Paul
Monday, 12 March 2018	Lesson 6	Model Evaluation	Paul
Wednesday, 14 March 2018	Lesson 7	Regularisation	Paul
Monday, 19 March 2018	Lesson 8	Clustering	Paul
Wednesday, 21 March 2018	Lesson 9	Recommendations	Paul
Monday, 26 March 2018	Lesson 10	SQL + Productivity	Paul
Wednesday, 28 March 2018	Lesson 11	Decision Trees	Greg
Monday, 2 April 2018	No Class	Easter Monday	.--
Wednesday, 4 April 2018	Lesson 12	Ensembles	Greg
Monday, 9 April 2018	Lesson 13	Natural Language Programming	Greg
Wednesday, 11 April 2018	No Class	** Paul ill	.--
Monday, 16 April 2018	Lesson 14	Time Series + R	Paul
Wednesday, 18 April 2018	Lesson 15	Soft Skills + Cloud Computing	Paul
Monday, 23 April 2018	Lesson 16	Network Analysis	Paul
Wednesday, 25 April 2018	No Class	ANZAC Day	.--
Monday, 30 April 2018	Lesson 17	Neural Networks	Paul
Wednesday, 2 May 2018	Lesson 18	GA Data Science Alumni Panel / Final Project Help	Olivia / Paul
Monday, 7 May 2018	Lesson 19	* Final Projects Presentations	Paul
Wednesday, 9 May 2018	Lesson 20	** Final Projects Presentations	Paul

Lesson 15 - Review

FINAL PROJECT

- Final Project split into 4-parts: [Review with James 5mins]
    - (a) Real-world Problem Identification [Lesson 14]
    - (b) Data Cleaning [Lesson 15]
    - (c) Model & Validation [Lesson 16]
    - (d) Presentation & Storytelling [Lesson 17]

Lesson 18 – any final queries
Lesson 19 & 20 – Presenting final project back to class

Git & GitHub – 1 Pager Guide!

(Part B) EVERY CLASS:

At the START of the class, you'll need to sync the latest materials from the COURSE repo:

(1) **Make sure you are in the dat11syd directory:**

```
cd ~/workspace/dat11syd
```

(2) **Make sure to select the "master" branch of your repo:**

```
git checkout master
```

(3) **Fetch the latest changes from the UPSTREAM repo (i.e the course repo)**

```
git fetch upstream
```

(4) **Merge the changes from the upstream repo to your master branch:**

```
git merge upstream/master
```

DURING the class:

(5) **Before editing, either copy files to your "students/" folder, or rename them**

At the END of every class:

(6) **Make sure you are in the dat11syd directory:**

```
cd ~/workspace/dat11syd
```

(7) **Add any files that you've updated to your git registry:**

```
git add -A
```

(8) **Commit the changes with a sensible comment:**

```
git commit -m "my updates for lesson 7"
```

(9) **Push your changes to your PERSONAL repo:**

```
git push origin master
```

DONE!!!!!

AGENDA 6

Part 1 – Big Data & Cloud Computing

1. Real World Data Science .vs. Data Science Course
2. Big Data Environments
3. NO LAB TODAY (Homework! ☺)

--- TEA BREAK! ---

Part 2 - Soft Skills

1. Challenges working in a team (technical & non-technical)
2. Pitching projects to varied audiences
3. Relationship Management
4. Job hunting & Interviewing

PART 1

Real World Data Science 8

1. **BIG DATA! (GB, TB, PB!)**
2. **Data must be extracted & treated (SuperComputer "Cluster") - SQL**
3. **Login via command-line or browser (rather than local)**
4. **Model is complex & consume lots of resource (SuperComputer "Cluster")**
5. **Data Sampling may be required**
6. **Big Data Jobs must be submitted to a QUEUE**
7. **Separate DEVELOPMENT, VALIDATION (testing) and PRODUCTION**
8. **Data Science Project Management**
   a) **Requirements gathering & Project Planning**
   b) **Data discovery → Model Development → Implementation → Monitoring**
   c) **Sales**

Real World Data Science .vs. DAT11 course

1. BIG DATA! (GB, TB, PB!) ✅
2. Data must be extracted & treated (SuperComputer "Cluster") - SQL ✅
3. Login via command-line or browser (rather than local) ✅
4. Model is complex & consume lots of resource (SuperComputer "Cluster")
5. Data Sampling may be required
6. Big Data Jobs must be submitted to a QUEUE
7. Separate DEVELOPMENT, VALIDATION (testing) and PRODUCTION ✅
8. Data Science Project Management
   a) Requirements gathering & Project Planning
   b) Data discovery → Model Development → Implementation → Monitoring
   c) Sales

✅ Final Project

DATA SCIENCE PART TIME COURSE

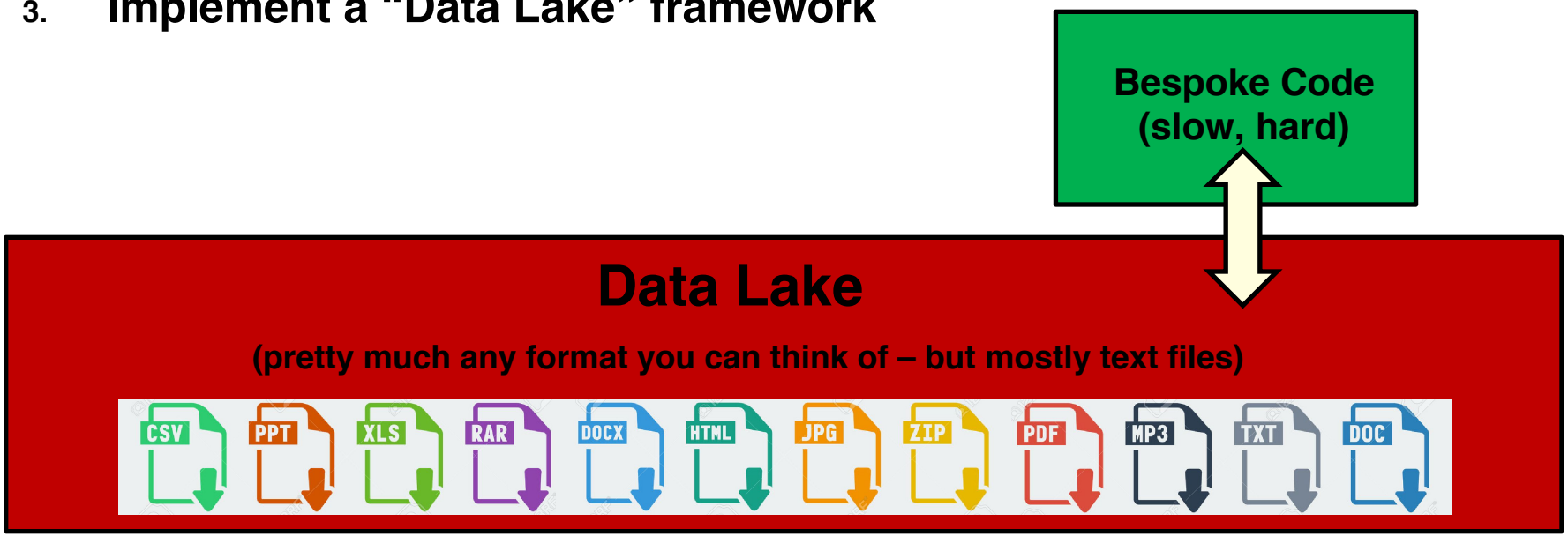
BIG DATA

&

CLOUD COMPUTING

Initial Big Data setup (simplified)

1. Purchase a supercomputer (a “cluster”) with LOTS of power
2. Purchase LOTS of disk space to connect to the supercomputer
3. Implement a “Data Lake” framework



Hadoop Distributed File System ("HDFS" / "Hadoop")

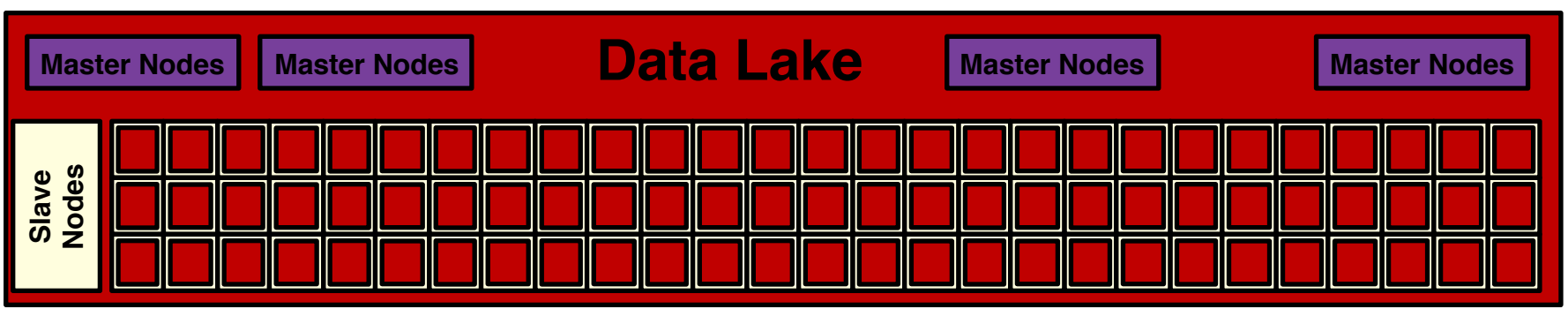
This is also simplified (but you get the idea!)

Typically done using Java

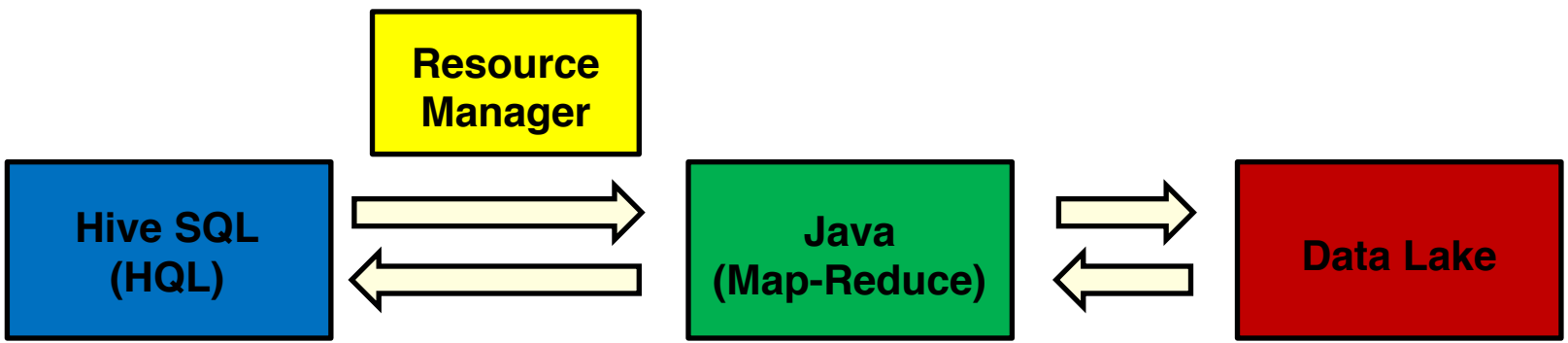
Hadoop "Map-Reduce"

- **Data of the relevant type is mapped by mappers (most of the slaves)**
  (e.g. sorting coins of the same type into 20c, $1, $2 etc)

- **Sorted Data is then collected by the "Reducers" (a few of the slaves) and aggregated as required:**
  (e.g. 20 x $2 → $40, 50 x $1 → $50, 400 x 20c → $80)



- Java = quite specialised in 2010+
- Most people in business know SQL…
- HIVE = Layer over the top of Hadoop → Hive SQL (“HQL”)



Map-Reduce: Pseudo-Code Example from WikiPedia


```
  SELECT age, AVG(contacts)
    FROM social.person
GROUP BY age
ORDER BY age
```

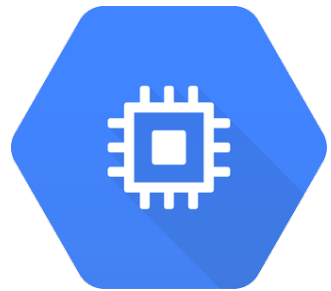
Using MapReduce, the `K1` key values could be the integers 1 through 1100, each representing a batch of 1 million records, the `K2` key value could be a person's age in years, and this computation could be achieved using the following functions:

```
function Map is
    input: integer K1 between 1 and 1100, representing a batch of 1 million social.person records
    for each social.person record in the K1 batch do
        let Y be the person's age
        let N be the number of contacts the person has
        produce one output record (Y,(N,1))
    repeat
end function

function Reduce is
    input: age (in years) Y
    for each input record (Y,(N,C)) do
        Accumulate in S the sum of N*C
        Accumulate in C_new the sum of C
    repeat
    let A be S/C_new
    produce one output record (Y,(A,C_new))
end function
```

WHAT IS IT AND WHY USE IT? (GOOGLE)

- Compute Engine
- Cloud SQL
- Container Engine
- BigQuery



WHAT IS IT AND WHY USE IT? (AWS)

- EC2
- RDS
- Container Service
- Redshift

Amazon EC2

Amazon RDS

Amazon ECS

Amazon **Redshift**

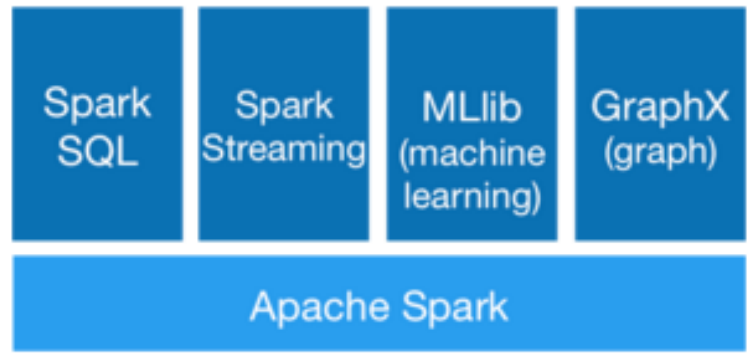

Spark

DATA SCIENCE PART TIME COURSE

SPARK

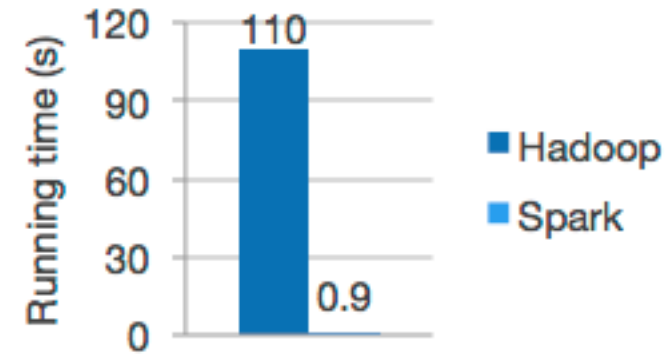
SPARK - WHAT IS IT?

Spark is a fast and general processing engine compatible with Hadoop data. It can process data in HDFS, HBase, Cassandra, Hive, and any Hadoop InputFormat. It is designed to perform both batch processing (similar to MapReduce) and new workloads like streaming, interactive queries, and machine learning.

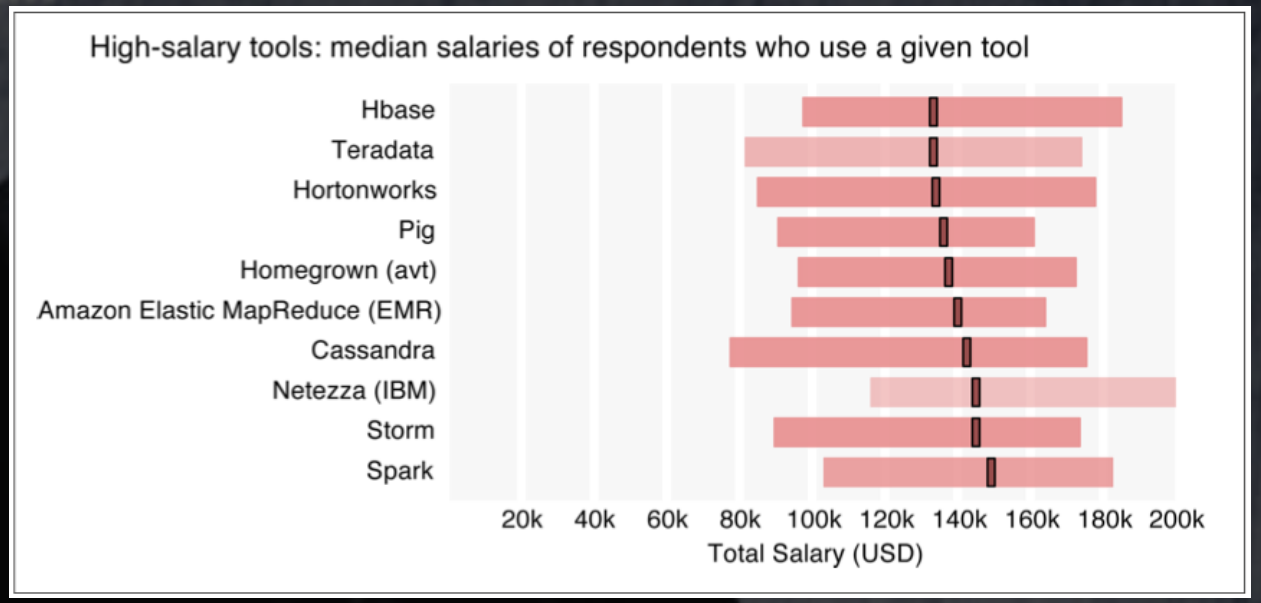


SPARK - HOW DOES IT RELATE TO DATA SCIENCE? 20

- **MLlib is Spark's machine learning library. Its goal is to make practical machine learning scalable and easy. It consists of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as lower-level optimization primitives and higher-level pipeline APIs.**
- **GraphX in Spark for graphs and graph-parallel computation**



Logistic regression in Hadoop and Spark



‘We can talk, but money talks, so talk more bucks’ - Jay-Z (Izzo - The Blueprint)

One use of Spark SQL is to execute SQL queries written using either a basic SQL syntax or HiveQL. Spark SQL can also be used to read data from an existing Hive installation.

Spark SQL provide Spark with more information about the structure of both the data and the computation being performed. Internally, Spark SQL uses this extra information to perform extra optimizations. There are several ways to interact with Spark SQL including SQL, the DataFrames API and the Datasets API. When computing a result the same execution engine is used, independent of which API/language you are using to express the computation.

A DataFrame is a distributed collection of data organized into named columns.

It is conceptually equivalent to a table in a relational database or a data frame in R/Python, but with richer optimizations under the hood.

DataFrames can be constructed from a wide array of sources such as: structured data files, tables in Hive, external databases, or existing RDDs.

Data types

Basic statistics

- summary statistics
- correlations
- stratified sampling
- hypothesis testing
- streaming significance testing
- random data generation

Classification and regression

- linear models (SVMs, logistic regression, linear regression)
- naive Bayes
- decision trees
- ensembles of trees (Random Forests and Gradient-Boosted Trees)
- isotonic regression

Collaborative filtering

- alternating least squares (ALS)

Clustering

- k-means
- Gaussian mixture
- power iteration clustering (PIC)
- latent Dirichlet allocation (LDA)
- bisecting k-means
- streaming k-means

Dimensionality reduction

- singular value decomposition (SVD)
- principal component analysis (PCA)

Feature extraction and transformation

Frequent pattern mining

- FP-growth
- association rules
- PrefixSpan

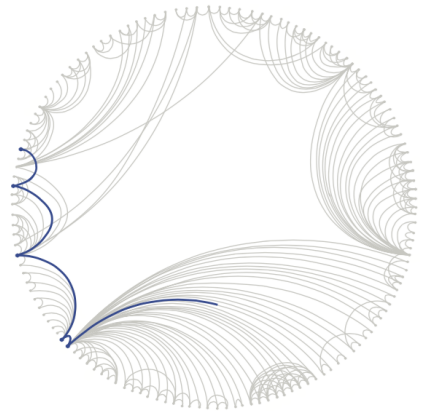
Evaluation metrics

PMML model export

Optimization (developer)

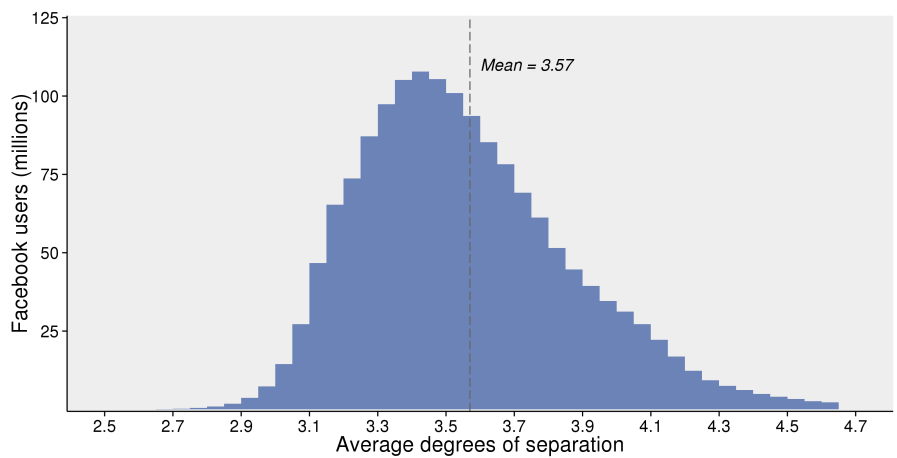
DIGRESSION - GRAPH ANALYSIS - FACEBOOK

How connected is the world?



Each person in the world (at least among the 1.59 billion people active on Facebook) is connected to every other person by an average of three and a half other people.

Rather than calculate it exactly, they estimate distances with statistical algorithms



DATA SCIENCE PART TIME COURSE

BIG DATA WITH...

JSON - WHAT IS IT?

JSON - JavaScript Object Notation

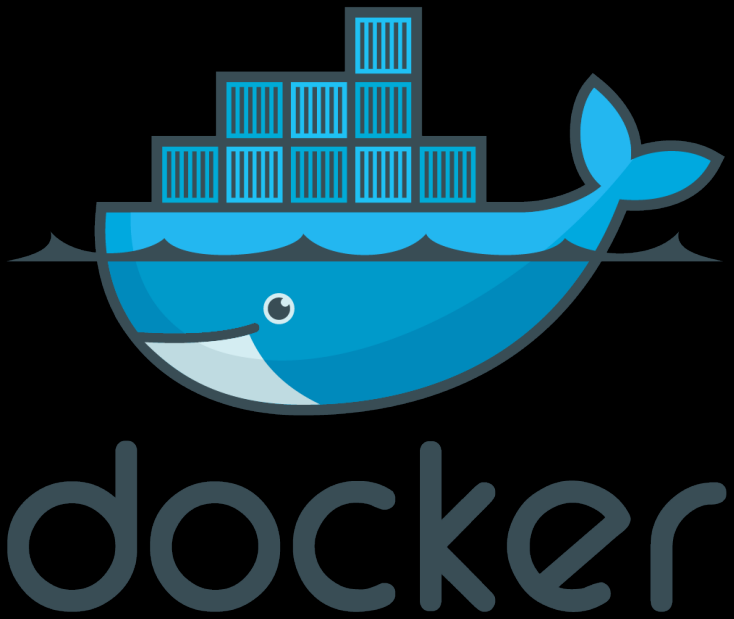
- Human readable data with attribute-value pairs.
- What is inside the curly brackets is an object
- In the object we declare variables with ‘attribute’ : ‘value’ pairs

```
var json = {
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "34 York St",
    "city": "Sydney",
    "state": "NSW",
    "postalCode": "2000"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "02 95999999"
    },
    {
      "type": "office",
      "number": "0431 111 111"
    }
  ],
  "children": [],
  "spouse": null
}
```

JSON - HOW DOES IT RELATE TO DATA SCIENCE?

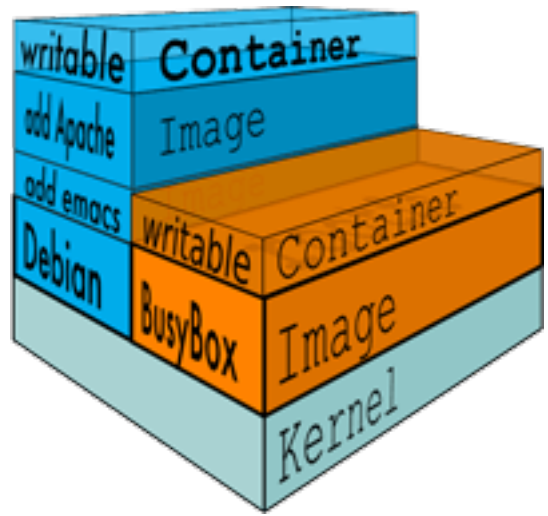
- Webservices provide application programming interfaces (APIs) are now usually transferring data via JSON
- Underlying document databases like MongoDB
- Increasingly common data format

DATA SCIENCE PART TIME COURSE

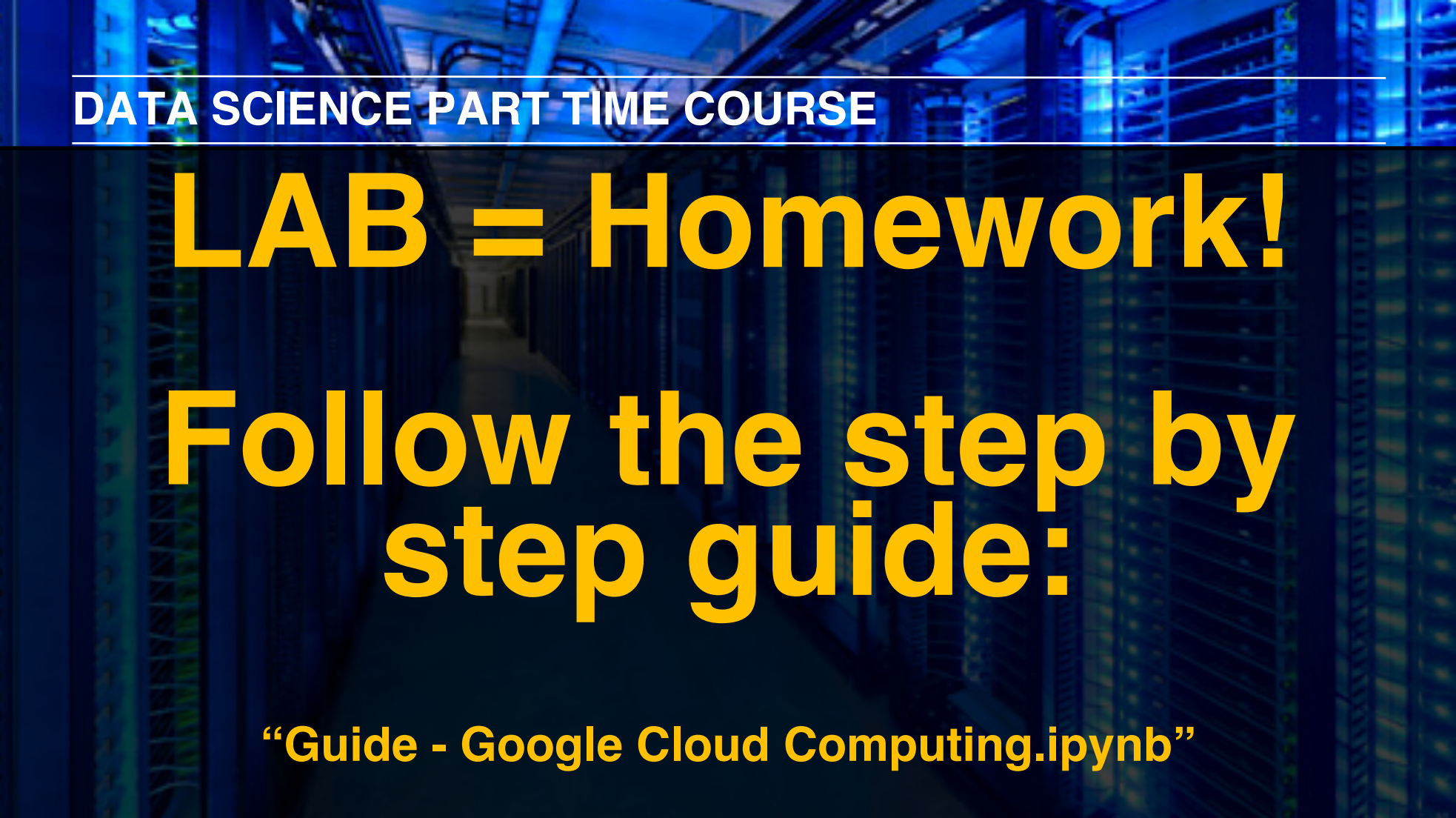


DOCKER - WHAT IS IT?

Docker containers wrap up a piece of software in a complete filesystem that contains everything it needs to run: code, runtime, system tools, system libraries – anything you can install on a server. This guarantees that it will always run the same, regardless of the environment it is running in.



- **Lightweight**
- **Open**
- **Secure**


DATA SCIENCE PART TIME COURSE
LAB = Homework!
Follow the step by step guide:
“Guide - Google Cloud Computing.ipynb”

TEA BREAK

PART 2

Agenda

1. Questions
2. Communication & presentation skills
3. Group exercise
4. Networking & organisational politics
5. Consulting
6. Group exercise
7. Management
8. Group exercise
9. Governance: Data & analytics
10. Job seeking

By the end of this lesson you should be able to ...

Discuss the importance of effective communication in data science

Outline the requirements of a successful presentation

Appreciate the complexity and importance of organisational politics

Describe techniques for effective consulting

Explain the role of governance in data and analytics

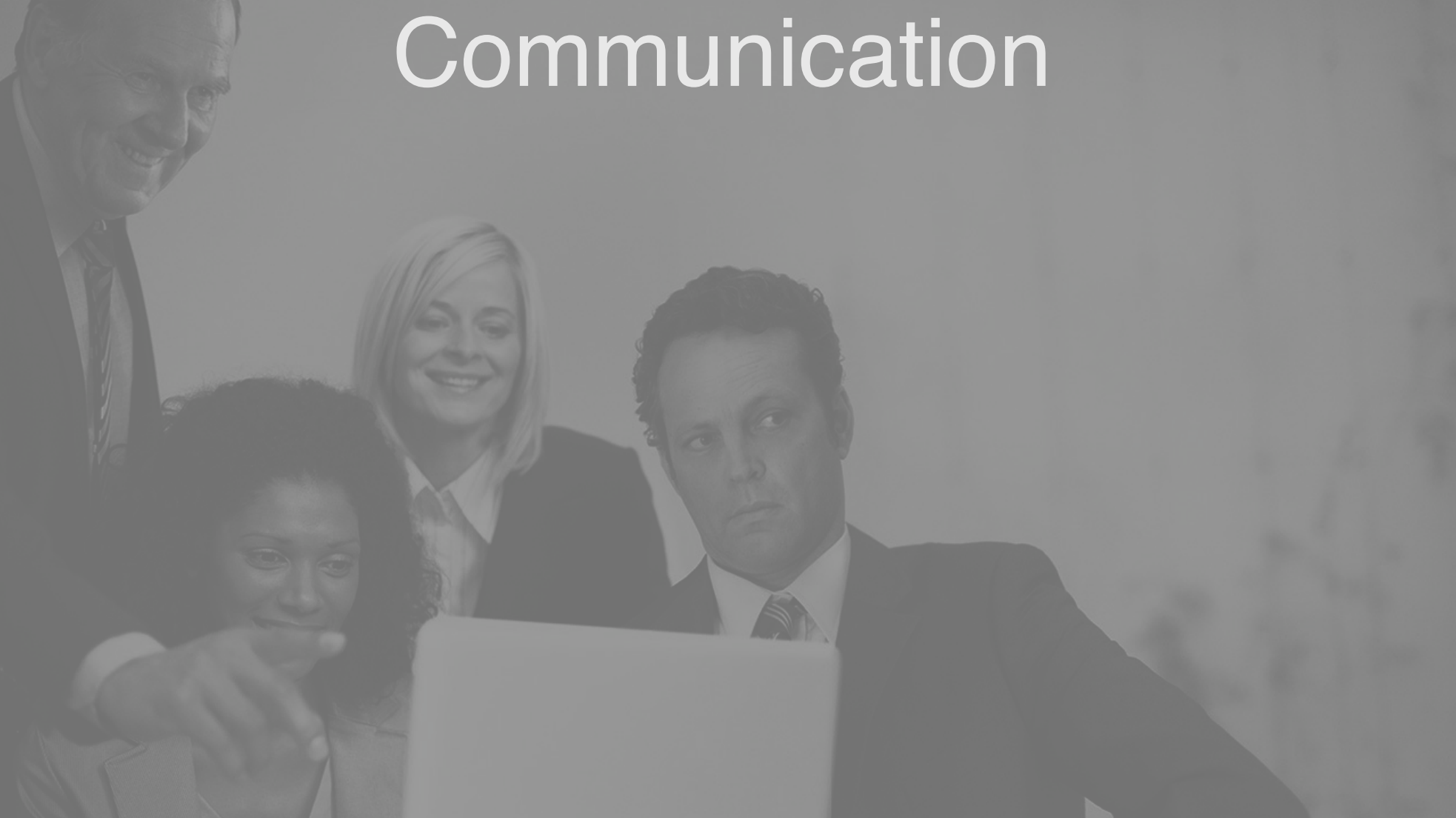
Approach the job market with a more focused strategy

“

Tact is the ability to tell someone to go to hell in such a way that they look forward to the trip.

Winston S. Churchill

Communication



Communication Tips - Verbal business conversations:

know your audience

know the business

know the ‘why’

be clear and concise

over-communication is better than under-communication

listen carefully

ask questions in depth

break bread

be genuine (but not *too* familiar)

Communication Tips – Written business conversations:

make the purpose clear

stick to the topic

if it's a proposal, explain why there is no easy alternative (e.g. Excel)

check spelling

use the language of the business unit

consider who should be cc'd

consider what would happen if your document were forwarded to other people in the organisation?

- comprehension / interpretation issues
- potential for misuse
- political fallout

Communication Tips – Tailor to your audience:

language and complexity should be appropriate to the audience

if you have a mixed audience, explain complex concepts in three stages

- ★ very general level
  - something your parents would understand
- ★ slightly technical, more business-focused
  - something a business analyst would understand
- ★ more technical
  - something a peer would understand

all stages should communicate the key points in appropriate detail

Presentation

Presentation

What are presentations for?

Advice on how to prepare / deliver effective presentations?

Presentation

how you present your work will determine *whether* it gets implemented

- if your work isn't implemented it's worthless

a data science presentation is not organised like a scientific presentation

- 1st: Business problem statement
- 2nd: Results
- 3rd: Recommended action
- 4th: Details
  - how you to got the results
  - appropriate to audience comprehension

considerations

- can your audience understand and appreciate what's in front of them?
- what questions should you prepare for?
  - could there be a surprise expert in the audience?

Presentation

use *graphics* to punctuate your findings

keep charts simple

- highlight the features you consider important
- suppress the features that are distracting
- but, don't compromise *visual integrity*

don't try to say too much on a single slide; build a story

don't save the best for the last

- they'll be bored / disengaged by then
  - *grab their attention!*
- the most important people are the busiest, so use their time carefully
  - shows your respect for being given this opportunity
  - shows them you know what's important

Presentation

present your recommendations

- be concise and confident
- execs want to be able to make a decision *soon*
  - give them what they need
  - multiple options are even better than go/no-go
- rehearse answers to the following:
  - what are our competitors doing?
  - what *you* would do (and why)?
  - what if you're wrong?
- they don't want to hear, "we need to do further study"
  - may come out if the discussion goes deep enough, anyway
  - don't offer it unless you get pinned on something you can't answer

Group Exercise:

Data Science in *Business*

Part 1: Pitch Your Project

Pitch your project

1. Prepare a 2-minute pitch for your project

2. Include:
    a. background (the problem you want to solve)
        - assume the audience is familiar with the issues at some level
    b. strategy (what you will use to solve it and what resources it will consume)
        - don't get too technical
    c. value proposition (what the Business will get for its investment)
        - make this compelling but don't over-promise

3. Pretend you have been given the chance to propose your project to the executives; present it to the class.

10 minutes to prepare; 2 minutes per pitch

How did we go?

Politics

Networking

Why network?

Networking

external

- keep on top of your game
- meet like-minded people
- discover interesting topics & techniques
- assist in finding new roles

internal

- who's good at what
- who shares your vision
- who should you have on your side
- who should you watch out for
- who is supporting whom
- who is influencing whom

Organisational Politics

"While most people are wary of organisational politics, politically savvy leaders know how to manage them. They take initiative, and they forge consensus. Ultimately, the help others maximise their impact, so the organisation can continue to thrive in a highly competitive future."

Political Savvy, Joel R DeLucca
EBG Publications, 1999

Organisational Politics

Thoughts?

Impressions?

Where do you think we might be going with this?

Organisational Politics

Acceptance of a proposal depends on building support

understand the dynamics

- **vested interests**
- **conflicts of interest**
- **potential winners and losers**
- **networks of influence and opinion**

create winners

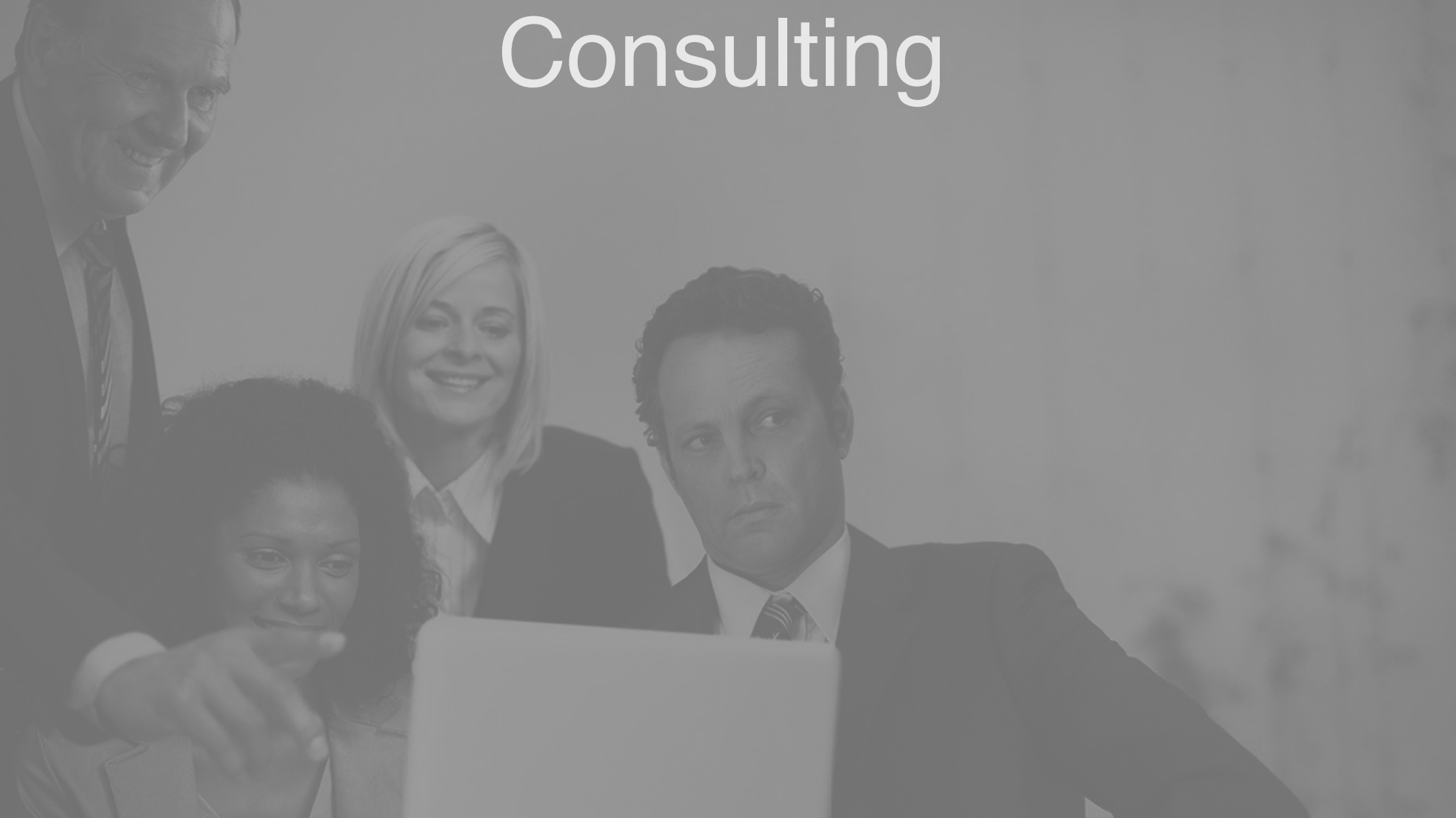
- **show how success will make your champions into heroes**
- **if possible, deal with resistance by offering partnerships**
- **if conflict is inevitable (or encouraged), build a strategic network to get round (or through) the blockers**

Organisational Politics

Success of a project depends on maintaining support

- a successful prototype doesn't guarantee acceptance into production
- expectations diverge
- suspicions propagate
- keep stakeholders on-side throughout project development
- plan to deal with blockers
  - dead wood
  - status quo
  - professional envy; competition for funding, status
  - fear of loss of control or relevance
- prepare end-users for roll-out
  - exploit occupational change management resources if available
  - never stop selling -- even after project delivery

Consulting



Internal and External Consulting

What do these have in common?

How do they differ?

Internal and External Consulting

Internal Consulting

- position is more secure; failure is usually tolerable
- there is time to develop trust, business acumen, and expert networks
- reputations and politics are more visible
- you can negotiate for support and exploit existing alliances

External Consulting

- failure is usually terminal
- you're on your own
- you need to convince *everybody* that they can trust your advice
- your presence will probably engender some hostility and/or anxiety

Stakeholder Cues

Identifying resistance

"I just don't see the point…"

"I still don't understand what you're hoping to achieve…"

- ➢ they are saying they will not be convinced
- ➢ these are confirmed blockers

Eliciting candour

ask 3 times what they concerns/issues are

think about each answer, then ask a follow-up question with more depth

- ➢ the 1st answer will be high-level, vague
- ➢ the 2nd answer will be somewhat indulgent, slightly defensive, and probing
- ➢ the 3rd answer will take you into confidence

Group Exercise:

Data Science in *Business*

Part 2: Networking & Negotiation

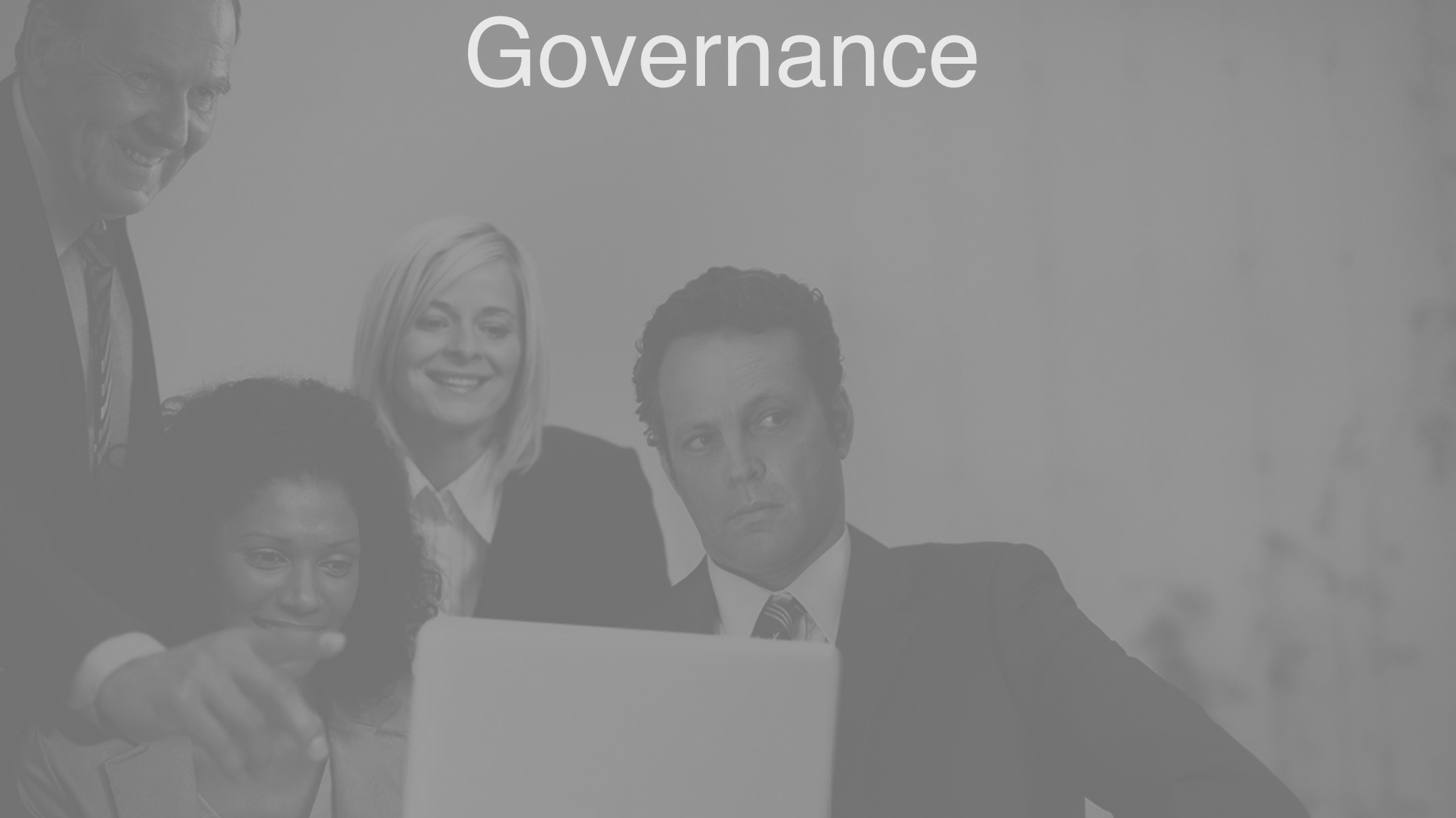
Negotiation Task

1. Form 2 groups.
2. Take turns in the hot seat:
    a. defend your project, try to win support
    b. others: adopt roles of supporter or resistor
        - assume a particular role within the enterprise (on par with the presenter)
        - supporters offer helpful criticism or suggest collaborative angles
        - resistors question the value proposition and try to discourage the proposal
3. Make notes
    a. specifics of criticisms
    b. create a map of the organisational politics for the project
        - you can assume the existence of other people in the organisation (e.g. the decision makers) if you want to make a more realistic map

4 minutes for each presenter

What did we learn?

Governance



Management

Managing Your Client

keep the contract simple

- but make sure it reflects your understanding of the problem unambiguously

keep the scope of any deliverables tightly defined

- it may be better to negotiate, contract, deliver in cycles

don't do any work not agreed to

don’t deliver too little *or* too much

don't encourage trite conclusions

your notice period should be the same as their notice period

read / learn as much as you can

- about contracting in general
- about the client’s business

Management

Managing Your Team

management is hard

- most people leave their manager, not their job

you don’t have to do all the thinking; make sure to keep them challenged

- clever people won’t tolerate micromanagement

let people do what they are good at

- and let them try out their ideas, even if these aren’t what you would try first

keep analysts on a leash

- data scientists like to go down rabbit holes
- sometimes it will be rewarding to chase long shots, but don't lose sight of ROI
- maintain the focus

recognise contributions, share credit

Management

Managing Your Manager

understand their limitations

- many business team managers do not understand statistics or programming
- many data science managers have never worked as an analyst
- many accomplished analysts don’t understand how to win or manage clients
- not everyone can manage teams, even if they do understand the work
  - they may be subject to decision pressure

avoid contradiction

- present alternative perspectives or options

be diplomatic and impartial in presenting your findings and recommendations

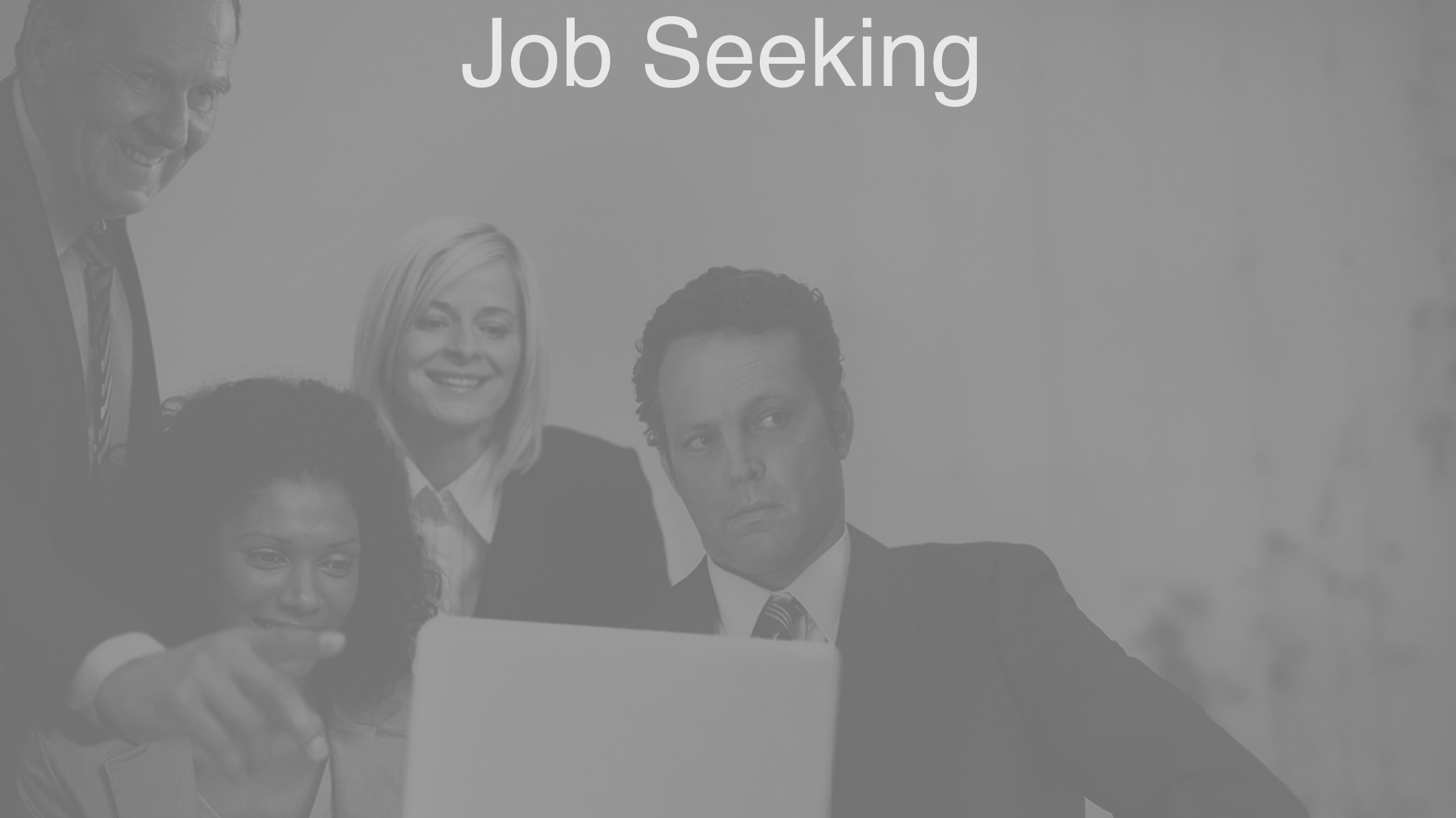
- beware of invitations to extrapolate

guide them tactfully toward optimal, informed decisions

Management

Any other advice?

Job Seeking



Aspiring Data Scientist

Target Industry: *anything ~~cool~~*

Target Salary: *~~$200,000+~~ $100,000*

Worked at:

SAP

ISIS

Job Applications

the Job Description is often a wish list

- could be prepared by somebody who doesn't understand the work
- call recruiter, open a dialog, get the inside scoop
- you should meet ≥ 80% of key requirements of the role

recruiters usually treat blind selection of candidates as a specification-matching exercise

- the client is hiring a person, not a specification,
  but they often don't know what kind of person they need

business experience can be more important than technical experience

- hiring decisions are usually made by someone with more of a business orientation than a technical one:
  they are hiring to solve a business problem!
- emphasise your experience in related industry verticals

Job Applications

Résumé

- 3 pages max (as much white space as possible)
- skills, education, current/recent employment on 1st page
- highlights of projects, not responsibilities
- PAR (Problem addressed, Approach/Action taken, Results/Recommendations)

cover letter

- 2 pages max
- many won't read it unless the résumé already caught their attention, but some will read it first
- demonstrate how your experience and knowledge would fit role and explain why you want to work for this employer
- not every candidate will bother to provide one;
  an opportunity for you to stand out

Job Applications

Follow up

- call the recruiter or advertiser; most people don't, so here is a chance to ensure that your résumé gets read
- ask when they expect to know something
- follow up regularly

Interview

- research the employer
- research the interviewers
- prepare for their questions
- practice your answers
- prepare your own questions
- read advice about interview preparation for your profession
  - data scientists and analysts
  - the employer's industry

Job Applications

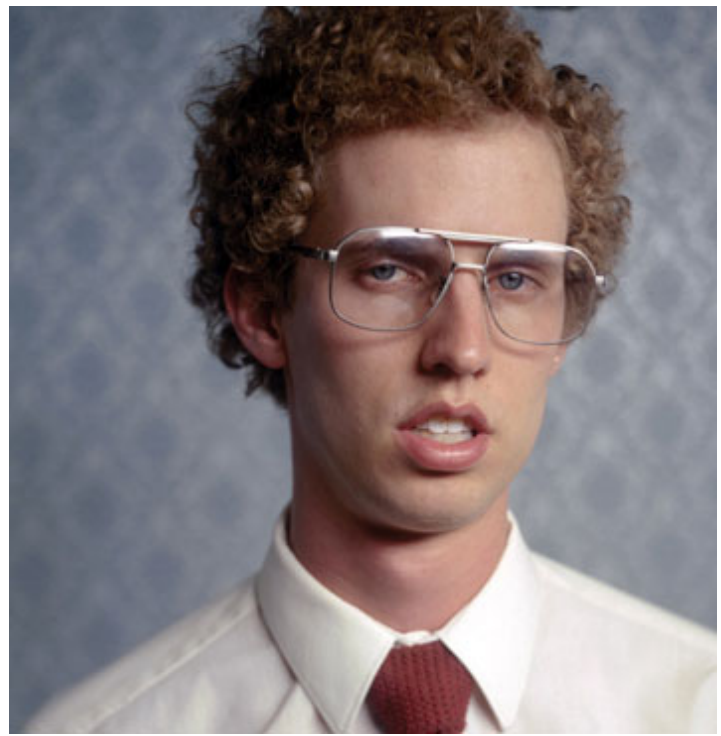
Any more advice?

DATA SCIENCE PART TIME COURSE

INTERVIEWS

SKiLLz

- Problem solving
- Coding ability
- Communication



Knowledge

- Technical knowledge
  - Machine Learning
  - Tools (software)
  - Certifications
- Business Knowledge
  - how the world works
- Statistics
- Data Handling

Personal Traits

- Interest in Technology
- Interest in data and what it represents
- Good Communicator
- Networking
- Willing to share ideas

- **What is R^2** (for linear regression)
- How do you assess whether to include a variable in a linear regression model?
- How would you assess model accuracy?
- Explain what regularisation is and why it is useful.
- Explain what resampling methods are and why they are useful. Also explain their limitations.
- Give an example of how you would use experimental design to answer a question about user behaviour.

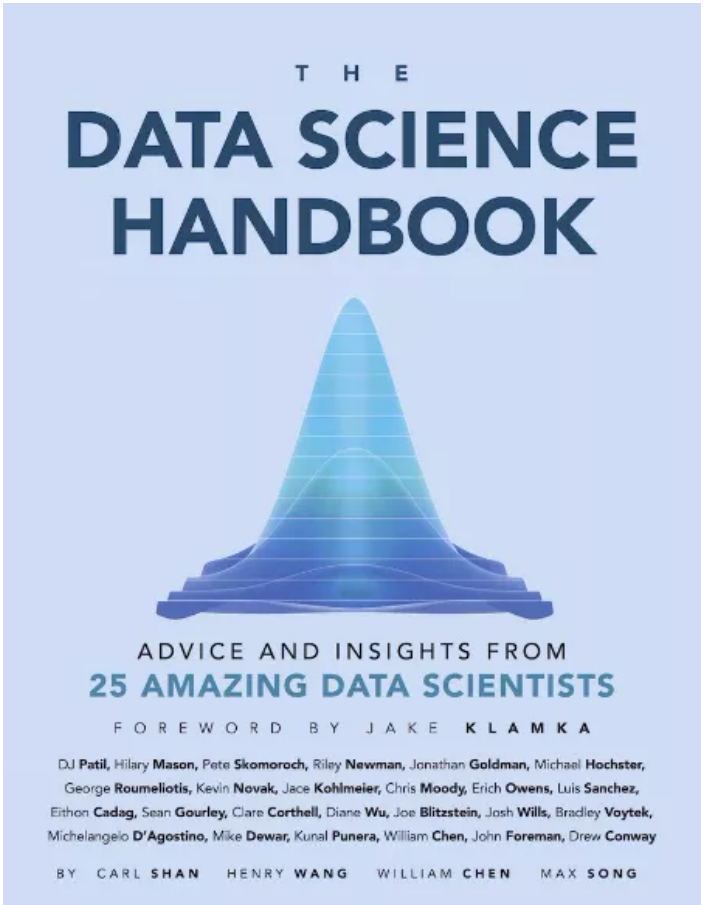
1. What was the last thing that you made for fun?

2. What's your favourite algorithm? Can you explain it to me?

3. Tell me about a data project you've done that was successful. How did you add unique value?

4. Tell me about something that failed. What would you change if you had to do it over again? ...

5. You clearly know a bit about our data and our work. When you look around, what's the first thing that comes to mind as "why haven't you done X"?! ...


THE
DATA SCIENCE HANDBOOK
ADVICE AND INSIGHTS FROM
25 AMAZING DATA SCIENTISTS
FOREWORD BY JAKE KLAMKA
DJ Patil, Hilary Mason, Pete Skomoroch, Riley Newman, Jonathan Goldman, Michael Hochster, George Roumeliotis, Kevin Novak, Jace Kohlmeier, Chris Moody, Erich Owens, Luis Sanchez, Eithon Cadag, Sean Gourley, Clare Corthell, Diane Wu, Joe Blitzstein, Josh Wills, Bradley Voytek, Michelangelo D'Agostino, Mike Dewar, Kunal Punera, William Chen, John Foreman, Drew Conway
BY CARL SHAN HENRY WANG WILLIAM CHEN MAX SONG

Professional Development

You can’t stand still in business. In data science, it isn’t even enough to keep a steady pace.

- join meetups
- attend webinars (use discretion)
- learn and apply new techniques
- learn about novel applications
- follow industry trends
- pick one or two specialities to develop
- conduct projects, share results
- blog, present