

제 1회 UNIST-KAIST-POSTECH

데이터 사이언스 경진대회 문제



Contents

- 1. 배경 설명
- 2. 문제 정의
- 3. 데이터

Part 1

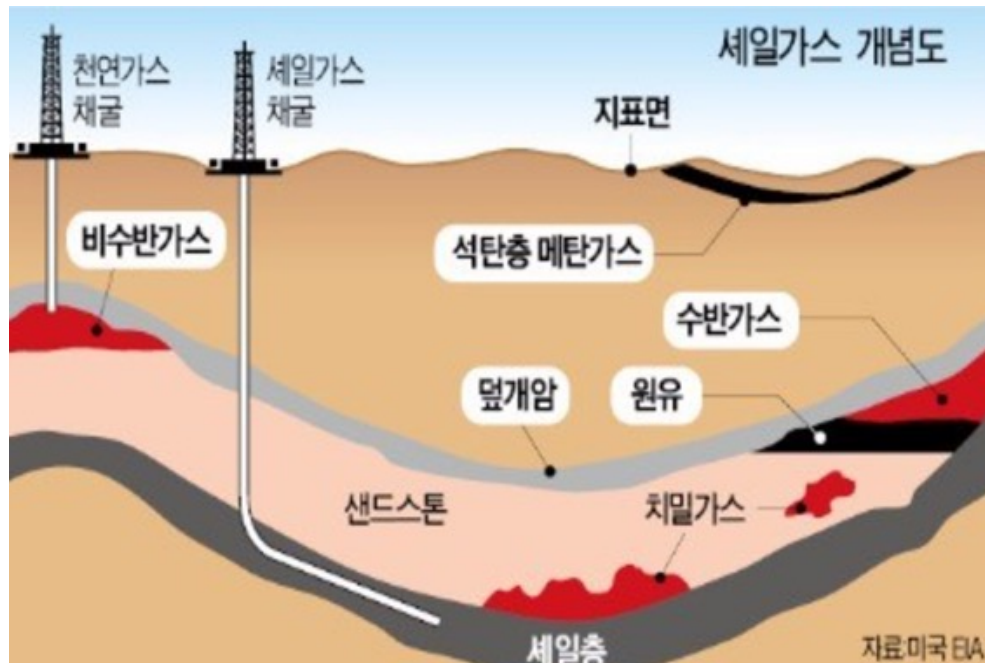


배경 설명



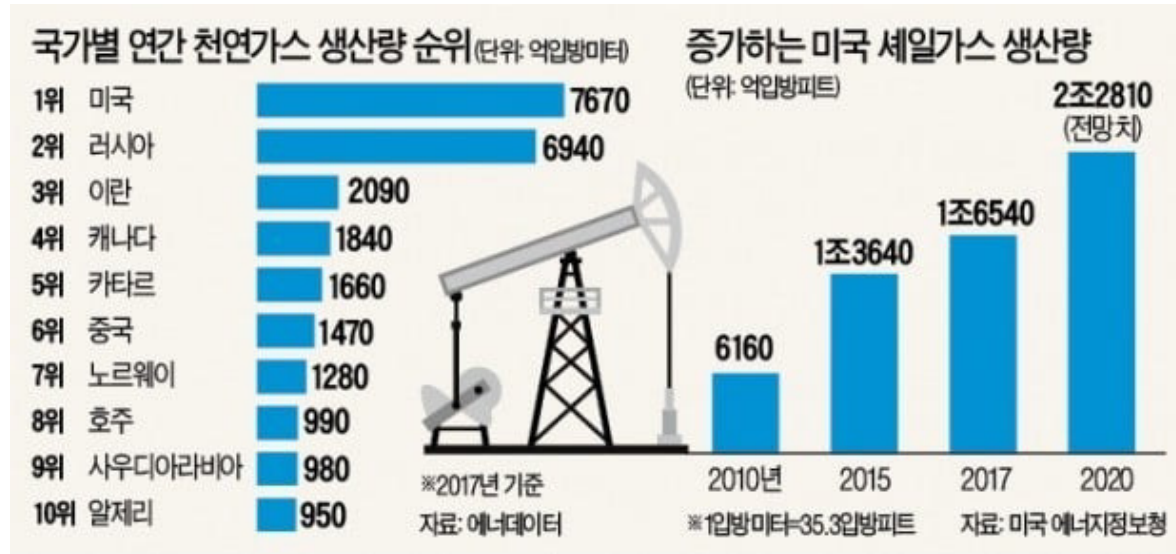
2011년 4월 TIME 표지

셰일가스



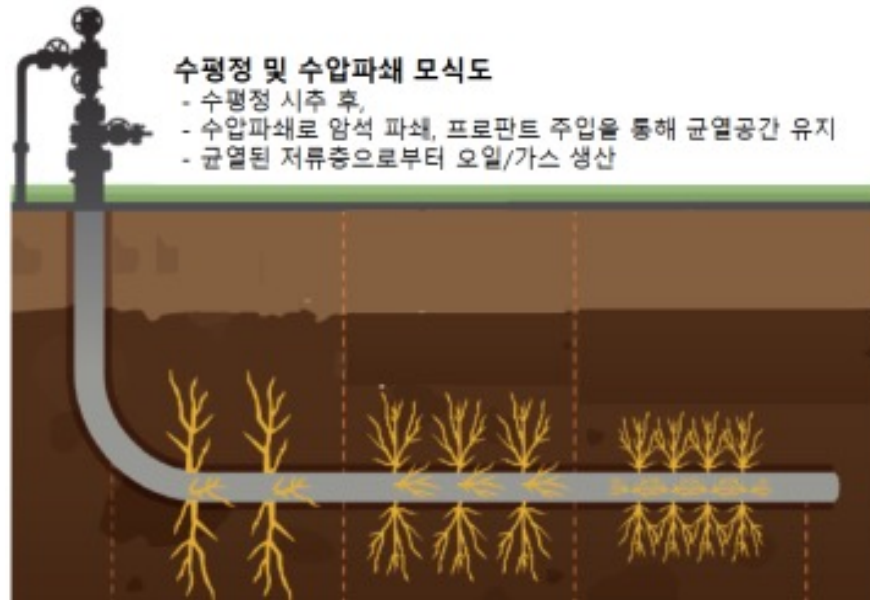
- 셰일가스는 탄화수소가 풍부한 셰일층에 매장되어있는 천연가스로 기술적 제약 때문에 오랫동안 채굴이 이뤄지지 못하다가 2000년대 들어 수평정시추 등이 상용화되며 신에너지원으로 급부상 하게 되었음.

셰일가스



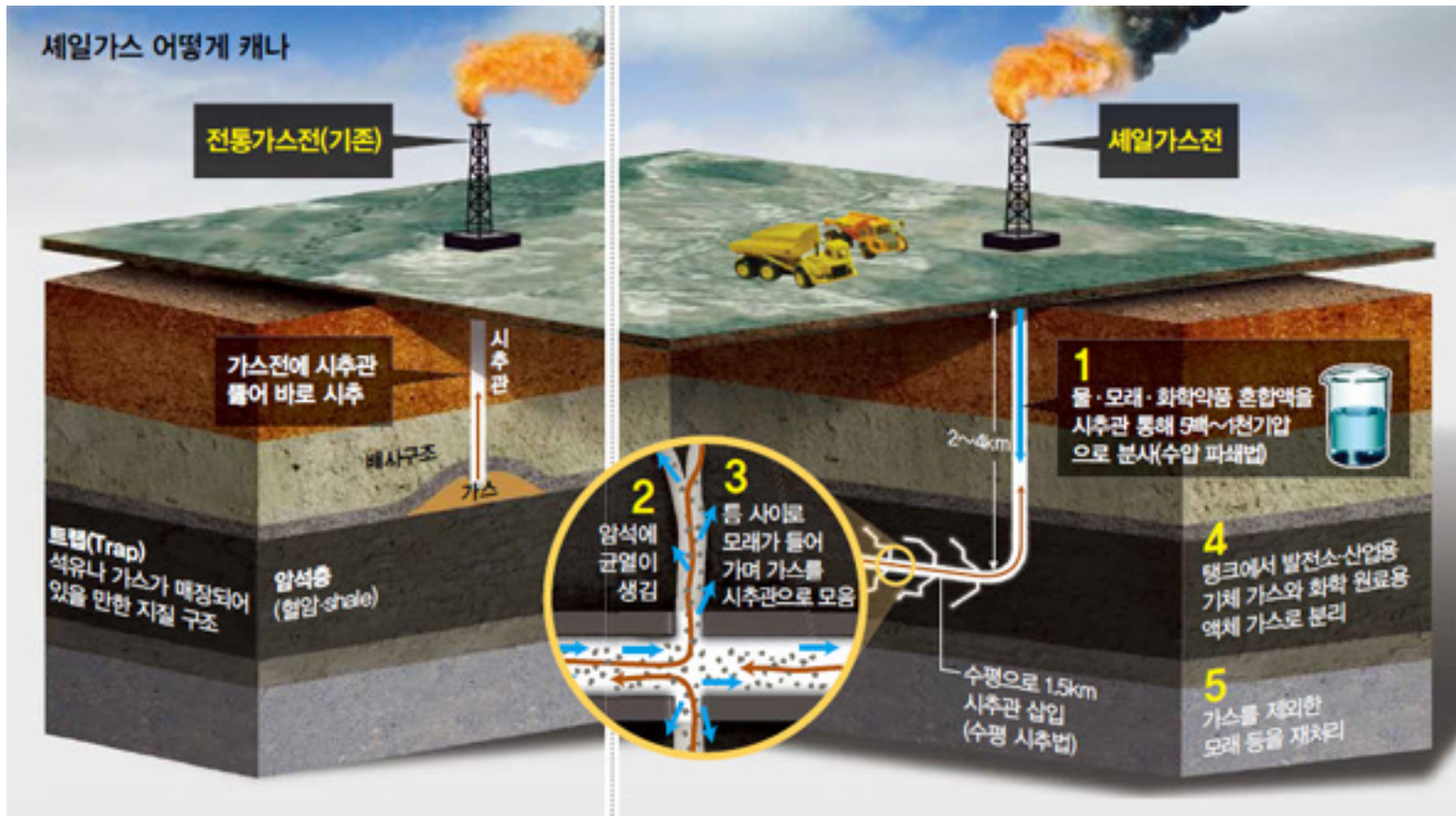
- 특히 미국에 많은 양이 매장되어 있어 2018년 미국은 러시아와 사우디를 제치고 석유 매장량 세계 1위 국가에 오르게 되었음. 이는 곧 세계 최대의 석유 수입국이었던 미국이 더 이상 중동에 의존할 필요가 없음을 의미하기도 함. 세계 에너지 판도를 바꾸는 획기적인 사건이며 세계 정치의 판도도 많이 바뀌게 되었음.

셰일가스 생산



- 하지만 셰일가스를 채취하는 수압파쇄법은 지하수 오염, 지진 등의 공해의 우려가 상당히 높아 법적 규제가 강해지고 있는 추세임. 이로 인해 셰일가스 채취비용도 올라가고 있음. 따라서 앞으로는 데이터를 기반으로 한 정확한 생산량 예측을 바탕으로 보다 정교한 계획 하에 셰일가스를 채취하는 것이 중요해질 것임.

셰일가스 생산



기타 참고자료

- 한국에너지공단 블로그
“에너지 혁명의 중심 ‘세일가스’가 뭐예요?”
(<http://blog.energy.or.kr/?p=4700>)
- 유튜브 세상의모든지식
“당신이 몰랐던 세일가스(천연가스)의 역사”
(<https://youtu.be/A0CT0ZWZ0ME>)
- 유튜브 KBS 다큐
“세일가스가 세계 경제에 미치는 여파(2012) |
KBS 스페셜 - 21세기 골드러쉬 세일가스 (KBS 121125 방송)”
(<https://youtu.be/Svk461QKeHk>)
- 유튜브 슈카월드
“미국 세일가스, 세계를 뒤흔들다. 압도적인 힘으로!”
(<https://youtu.be/ojzem0OfRks>)

Part 2

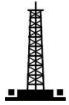


문제 정의

문제 정의



생산정 1



생산정 3



생산정 5



생산정 k



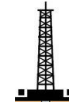
생산정 2



생산정 4



생산정 6



생산정 N



- N개의 셰일가스 생산정

문제 정의

생산정 N

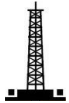
시추공 위치: (401619.6, 6031353.8)
시추 지점 땅높이: 2888 ft
시추 시작 지점 높이: 2908 ft
시추공 곡면 길이: 20325 ft
시추공 수직깊이: 10015 ft
수압파쇄 유체 종류: Slickwater
프로판트 주입 량: 5082.72 ton
균열을 위해 주입된 유체 양: 40553.7
수압파쇄 지점 수: 40
수압파쇄 지점간 평균 간격: 75.76 m
생산 시작 일자: 2018.02.01
운송라인 연결 일자: 2018.02.04
마지막 측정 일자: 2021.7.31

시계열 자료:
최근 36개월 월별 가스 생산량
최근 36개월 월별 조업 시간

생산정 N



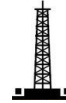
생산정 1



생산정 3



생산정 5



생산정 7



생산정 4



생산정 6

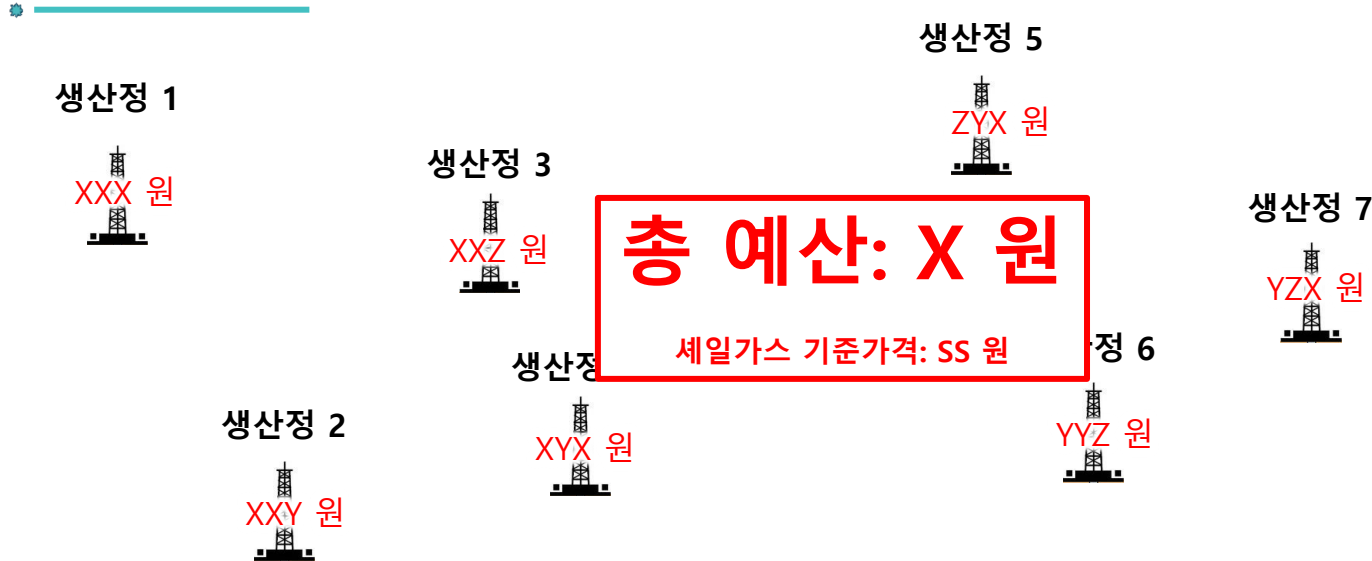


생산정 2



- N개의 셰일가스 생산정
 - 각 생산정에 대한 데이터가 주어짐

문제 정의



생산정 N

시추공 위치: (401619.6, 6031353.8)
시추 지점 땅높이: 2888 ft
시추 시작 지점 높이: 2908 ft
시추공 곡면 길이: 20325 ft
시추공 수직깊이: 10015 ft
수압파쇄 유체 종류: Slickwater
프로판트 주입 량: 5082.72 ton
균열을 위해 주입된 유체 양: 40553.7
수압파쇄 지점 수: 40
수압파쇄 지점간 평균 간격: 75.76 m
생산 시작 일자: 2018.02.01
운송라인 연결 일자: 2018.02.04
마지막 측정 일자: 2021.7.31

시계열 자료:
최근 36개월 월별 가스 생산량
최근 36개월 월별 조업 시간

생산정 N

XYZ 원

■ N개의 세일가스 생산정

- 각 생산정에 대한 데이터가 주어짐
- 각 생산정의 가격은 주어짐

- **Question.** x원의 예산이 주어졌을 때 어떤 생산정을 구매해야 향후 6개월간 생산된 세일가스를 팔아 가장 많은 수익을 낼 수 있는가?

문제 정의

■ 두 가지 유형의 생산정이 존재

- 1. 시추까지만 완료되고 **아직 생산을 시작하지 않은 생산정**
 - 시추공, 유정 완결 관련 데이터만 존재
- 2. **생산 시작 후 최소 30개월이 지난 생산정**
 - 시추공, 유정 완결 관련 데이터
 - + 지난 30개월 생산량, 조업시간 시계열 데이터 존재

문제 정의

■ 두 가지 유형의 생산정이 존재

– 1. 시추까지만 완료되고 아직 생산을 시작하지 않은 생산정

- 시추공, 유정 완결 관련 데이터만 존재
- **학습용 데이터**: 초기 6개월 평균 생산량 데이터 존재

– 2. 생산 시작 후 최소 30개월이 지난 생산정

- 시추공, 유정 완결 관련 데이터
+ 지난 30개월 생산량, 조업시간 시계열 데이터 존재
- **학습용 데이터**: 총 36개월 생산량, 조업시간 시계열 데이터 존재

문제 해결 포인트

- 각 생산정마다 **향후 6개월 세일가스 생산량 예측**
 - 생산이 시작되지 않은 생산정의 경우 static 예측
 - 생산 시작 후 30개월 이상 경과된 생산정의 경우 시계열 예측
- 생산정 가격 대비 기대 수익을 계산하고 이를 바탕으로 예산 내에서 어떠한 생산정을 구입할 것인지 **의사결정**
- 데이터 기반 “**예측**”과 “**의사결정**”이 결합된 전형적인 산업공학 문제

예선 평가 기준

■ 1. 생산량 예측 정확도

- 주어진 생산정에 대한 향후 6개월 생산량 예측 정확도

■ 2. 구입한 생산정에서의 수익

- 구입을 결정한 생산정들에서 향후 6개월 동안 생산된 셰일가스 판매를 통해 얻은 수익

- 각 생산정에 대한 정확한 생산량 예측
+ 이를 바탕으로 한 정확한 의사결정

본선 프리젠테이션

- **문제 풀이 과정에 대한 상세한 프리젠테이션**

- 데이터 전처리, 예측 모델 선택, 의사결정 도출 과정 등에 대한 논리적인 설명

- **추가 문제**

- 생산량 예측, 생산정 구매 의사결정 도출 과정에서 **현실적으로 추가 고려되어야 할 요소들이 어떤 것들이 있을지, 이를 고려한다면 예측, 의사결정 과정에 어떤 변화가 있을지에 대해 자유롭게 가능한 한 구체적으로 서술**

본선 평가 기준

■ 1. 생산량 예측 정확도

- 주어진 생산정에 대한 향후 6개월 생산량 예측 정확도

■ 2. 구입한 생산정에서의 수익

- 구입을 결정한 생산정들에서 향후 6개월 동안 생산된 셰일가스 판매를 통해 얻은 수익

■ 3. 프리젠테이션

- 예측 모델 선택, 의사결정 도출 과정에 대한 논리 전개 과정
- 추가문제에서의 창의성, 논리 전개의 완결성

Part 3



데이터

데이터

- 한국석유공사 제공
- 약 340여개의 세일가스 생산정 데이터
 - 시추공 데이터
 - 시추공 위치 (x, y 좌표)
 - 시추공 길이
 - 대상 저류층 등
 - 유정완결 데이터
 - 파쇄지점 수
 - 파쇄지점 간격
 - Cluster 수 / 파쇄 지점
 - 천공 수 / cluster
 - 주입 프로판트 양
 - Frac 유체 타입 등
 - 생산 시계열 데이터
 - 월별 가스/오일 생산량
 - 월별 운영일수 등
- 데이터는 대회 참여 신청 후 슬랙에 참여하면 주최측에서 제공

데이터 샘플

초기 6개월 셰일가스 생산량

시추공, 유정 완결
데이터

최근 36개월
셰일 가스 생산 데이터

최근 36개월
조업시간 및 CND 생산 데이터

최근 6개월 셰일가스 생산량

데이터 샘플: 생산정 유형 1

초기 6개월 세일가스 생산량

시추공, 유정 완결
데이터

아직 생산이 시작되지 않은 생산정의 경우
시추공, 유정완결 데이터를 통해 초기 6개월 세일가스 생산량을 예측

데이터 샘플: 생산정 유형 2

최근 6개월 세일가스 생산량

시추공, 유정 완결 데이터

30개월 세일 가스 생산 데이터

30개월 조업시간 데이터

30개월 CND 생산 데이터

생산 시작 후 36개월 이상 지난 생산정의 경우
시추공, 유정완결 데이터 + 30개월 시계열 데이터를 통해
최근 6개월 세일가스 생산량을 예측

용어 정리

■ 데이터 각 항목별 설명

– 석유공사 제공

용어 정리

1. Reference (KB) Elev. (ft)

: 지표에서 시추를 시작하는 지점(시추기 높이 고려)의 높이, Ground Elevation + 시추기 높이

2. Ground Elevation (ft)

: 해수면으로부터 시추지점의 땅높이

3. MD (All Wells) (ft)

: Measured Depth. 시추공 곡면을 따라서 켜 총 길이 (저류층 구간에 수평구간을 만들려면, 지표에서 수직으로 들어가다가 방향을 꺾어가면서(곡면을 만들어서) 목표구간에서 수평을 유지하게 되는 개념입니다)

4. TVD (ft)

: True Vertical Depth. 시추공(생산정)의 최종지점의 수직 깊이

5. Bot-Hole direction (N/S)/(E/W)

: 수평구간의 방향을 의미함. 남북길이/동서길이를 구한 값임.

: 수압파쇄가 효과적이라면, 시추공방향이 자연적으로 존재하는 응력방향(2차원 평면으로 생각했을 경우)과 직교할 때 유리합니다. 그래서, 이 방향이 생산량에 영향을 미치는 변수가 됩니다. max-min 노말라이즈를 염두에 두고 만든 변수입니다.

6. Bot-Hole Easting (NAD83), Bot-Hole Northing (NAD83)

: NAD83은 북미지역에서 많이 쓰는 좌표계입니다. 이것도 max-min 노말라이즈를 고려하여 만든 값입니다. Easting은 동서좌표, Northing은 남북좌표입니다. 경도, 위도의 개념으로 보시면 될 겁니다.

: 지역에 따른 생산특성(생산량, Condensate양 등)의 차이가 있을 수 있어서 놔둔 feature입니다.

7. On Prod YYYY/MM/DD, First Prod YYYY/MM, Last Prod. YYYY/MM

: First는 시추직후, 생산한 날짜입니다.

: On은 생산량을 운송라인에 연결해서 제대로 팔기 시작한 날짜입니다.

: Last는 마지막 측정입니다.

: 보통 수압파쇄를 하면, 부러진 암석과 주입한 용액들, 시추 파편들을 지표로 꺼내야 합니다. 이를 clean-up이라고 하는데, 이 기간과 생산라인에 연결하는 시간이 소요되는데, 이 차이가 On과 First의 차이입니다.

8. First 12 mo. Total GAS (Mcf), First 12 mo. Total CND (Bbl), First 12 mo. Total WTR (Bbl)

: 첫 12개월 가스 생산량입니다. M은 10의 3승을 의미합니다. (일반적으로 Mega와 다르

대회 진행

- 대회 진행 관련 세부 사항은 아래 참조
 - 대회 홈페이지 (<https://datascience-contest.com>)
 - 오픈카톡방 (대회 참여 전)
 - 슬랙 (대회 참여 후)
- 질문사항은 오픈카톡방과 슬랙 통해서 편하게 물어봐주세요

Thank you for listening!